

Distributionally Robust Structural Learning

by

Yeshe Li

B.E., Beihang University, 2016

THESIS

Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in Computer Science
in the Graduate College of the
University of Illinois at Chicago, 2023

Chicago, Illinois

Defense Committee:

Brian D. Ziebart, Chair and Advisor

Xinhua Zhang

Ian Kash

Xiaorui Sun

Will Perkins, Georgia Institute of Technology

Copyright by

Yeshu Li

2023

To my family

ACKNOWLEDGMENT

I wish to express my foremost and deepest gratitude towards my advisor, Prof. Brian D. Ziebart, for his constant support, perpetual guidance, scientific enthusiasm, selfless dedication and endless encouragement, during my PhD study. I have been extremely fortunate to work with him on fundamental machine learning research. Brian is the ideal supervisor I could imagine before I pursue a PhD. He has created a flexible laboratory atmosphere and given me the freedom to pursue the research projects where I get to exert my strength. He has exposed me to the cutting-edge work and various opportunities to communicate with experts in my field. He has helped me develop my own flavor and skills as a researcher. I could still remember the toy projects on Markov logic networks and adversarial tree prediction Brian assigned to me when I started doing machine learning research. They have served as very great starting points for my follow-up work and taught me how to learn from examples. I cannot overstate my appreciation for his continual support in the past few years when I was unable to conduct my research on campus. My deep thankfulness goes to Prof. Xinhua Zhang who serves as one of my defense committee members for his guidance on mathematical optimization and theoretical machine learning research. Without his constant availability and passionate dedication, most of my work is impossible.

I wish to express my appreciation to the rest of my defense committee members, Prof. Ian Kash, Prof. Xiaorui Sun, Prof. Will Perkins and a close collaborator, Prof. Kevin Gimpel, who contributed to my work and served as a committee member in my preliminary exam. I am

ACKNOWLEDGMENT (Continued)

indebted to them for their valuable comments on my thesis and gracious commitment. I am also deeply thankful to Prof. John Lillis for many conversations about classic computer science.

I would like to acknowledge the National Science Foundation (NSF) sponsoring this work.

I am lucky to collaborate with many excellent researchers: Zhan Shi, Danyal Saeed, Omid Memarrast, Ashkan Rezaei and Rizal Fathony, for their contributions to our works. I want to personally thank Shanshan Wu and Brandon Amos for communications on my queries. I would like to thank my other labmates: Wei Xing, Sanket Gaurav, Zainab Al-Qurashi, Sima Behpour, Kaiser Asif, Mohammad Ali Bashiri, Jurat Shayidin, Rushit N. Shah, Nikolaos Agadakos and George Maratos for all the discussions, encouragement, support and feedback throughout my PhD study. My life in UIC is certainly less colorful without the accompany and support of my friends: Zhiwei Liu, Yingtong Dou, Yingjie Li, Xiaohan Li, Zheng Liu, Lichao Sun, Ye Liu, Zhou Yu, Ruizhe Chen, Christopher Tran, Hai Tran, Thep Siwathep Singh Khanderpor.

Finally, none of this would have been possible without my parents, Jianwu Li and Silin Li, for their unconditional love and unlimited support through all the ups and downs of my PhD study. They have sacrificed a lot of their desires to bring me up with unreserved support. I would not have been able to proceed in this long journey without them. My appreciation also goes to my love and my family, the Li family, that originates from the Tang Dynasty more than two thousand years ago.

YL

CONTRIBUTIONS OF AUTHORS

Chapter 2 presents a published paper (Li et al., 2022b) in which I was the primary author. I came up with the idea, formulated the method, proved most of the theorems, designed and implemented the algorithms, and conducted the experiments. Zhan Shi and Prof. Xinhua Zhang provided guidance on proof of the main theorem in the paper. Prof. Brian D. Ziebart participated in designing and conducting the experiments.

Chapter 3 presents a paper working in progress in which I was the primary author. I came up with the idea, formulated the method, proved all the theorems, designed and implemented the algorithms, and conducted the experiments. Prof. Brian D. Ziebart provided guidance on ideas about classic structure learning problems and participated in designing the experiments.

Chapter 4 presents a published paper (Li et al., 2022a) in which I was the primary author. I proved all the theorems, designed the algorithms, and conducted the experiments. Danyal Saeed contributed to implementations of the algorithms. Prof. Kevin Gimpel helped design the experiments. Prof. Xinhua Zhang provided the formulation of an arborescence projection algorithm. Prof. Brian D. Ziebart provided the idea and general formulation of adversarial prediction.

Chapter 5 presents a published manuscript (Li and Ziebart, 2023) in which I was the primary author. I proved most of the theorems, designed most of the algorithms, and conducted most of the experiments. Prof. Brian D. Ziebart came up with the idea, formulated the method, proved the main theoretical result and contributed to the discussion about experimental design.

TABLE OF CONTENTS

<u>CHAPTER</u>		<u>PAGE</u>
1	INTRODUCTION	1
1.1	Structural Learning	1
1.2	Challenges	4
1.3	Distributionally Robust Optimization	8
1.4	Overview of Distributionally Robust Structural Learning . . .	14
1.5	Contribution and Outline	16
1.6	Notation	18
2	DISTRIBUTIONALLY ROBUST STRUCTURE LEARNING OF UNDIRECTED GRAPHICAL MODELS	20
2.1	Introduction and Related Work	21
2.1.1	Related Work	23
2.2	Preliminaries	24
2.3	Distributionally Robust Structure Learning	27
2.3.1	Distributionally Robust Discrete Pairwise Markov Network Learning	27
2.3.2	Tractable Reformulations	28
2.4	Theoretical Guarantees	37
2.5	Experiments	48
2.6	Concluding Remarks	53
3	DISTRIBUTIONALLY ROBUST STRUCTURE LEARNING OF DIRECTED GRAPHICAL MODELS	54
3.1	Introduction	54
3.1.1	Related Work	58
3.2	Preliminaries	59
3.3	Method	63
3.3.1	Basic Formulation	64
3.3.2	Wasserstein Formulation	65
3.3.2.1	Lemmas for Non-asymptotic Analysis	69
3.3.2.2	Main Results	84
3.3.3	Kullback-Leibler Formulation	96
3.4	Experiments	102
3.5	Concluding Remarks	103
4	MOMENT DISTRIBUTIONALLY ROBUST TREE STRUC- TURED PREDICTION	105

TABLE OF CONTENTS (Continued)

<u>CHAPTER</u>		<u>PAGE</u>
4.1	Introduction	106
4.2	Background and Related Work	108
4.2.1	Tree Structured Prediction	108
4.2.2	Maximum Likelihood	109
4.2.3	Maximum Margin	110
4.2.4	Minimum Risk	111
4.3	Method	111
4.3.1	Formulation	112
4.3.2	Constraint Generation Solution	115
4.3.3	Marginal Distribution Formulation	117
4.3.4	Inference	118
4.3.5	Statistical Properties	118
4.4	Projection onto Arborescence Polytopes	125
4.4.1	Frank-Wolfe Algorithm	126
4.4.2	Martin's Polytope	126
4.5	Extensions	130
4.5.1	Undirected Spanning Trees	130
4.5.2	Dependency Trees	131
4.5.3	Higher-order Polytope	133
4.6	Experiments	133
4.7	Concluding Remarks	140
5	MOMENT DISTRIBUTIONALLY ROBUST PROBABILISTIC SUPERVISED LEARNING	142
5.1	Introduction	143
5.1.1	Related Work	145
5.2	Preliminaries	146
5.2.1	Probabilistic Loss Functionals	146
5.2.2	Probabilistic Supervised Learning	148
5.3	Method	149
5.3.1	Formulation	149
5.3.2	Statistical Properties	152
5.3.3	Algorithm	154
5.3.4	Differentiable Learning	161
5.4	Experiments	162
5.5	Concluding Remarks	165
6	CONCLUSION AND DISCUSSION	166
6.1	Structure Learning	166
6.1.1	Structured Prediction	169
6.2	Potential Societal Impacts	171

TABLE OF CONTENTS (Continued)

<u>CHAPTER</u>	<u>PAGE</u>
CITED LITERATURE	172
APPENDIX	199
VITA	201

LIST OF TABLES

<u>TABLE</u>		<u>PAGE</u>
I	Comparisons of F1 scores for benchmark datasets and BIC for real-world datasets (backache, voting).	103
II	Comparison of mean UAS and execution time under different training set sizes. Time refers to the CPU time taken to finish one gradient descent step. Statistically significant differences compared to <i>BiAF</i> are marked with † (paired t-test, $p < 0.05$). The best UAS are highlighted in bold.	136
III	Comparison of mean UAS, LAS, UCM and LCM under different training set sizes. Statistically significant differences compared to BiAF are marked with † (paired t-test, $p < 0.05$). We highlight in bold the best results among the four methods.	137
IV	Dataset statistics and normalized generalization losses with 95% confidence intervals on each dataset. The best results are indicated in bold. † indicates statistical significance with paired t-test ($p < 0.05$).	164

LIST OF FIGURES

<u>FIGURE</u>		<u>PAGE</u>
1	The adopted underlying graphs. Two nodes are connected to the others in the diamond graph. The grid graph has d^2 nodes. Each edge weight matrix is centered with random values $\pm\theta$	49
2	Plots of the probability of successfully estimating the structure versus the number of samples for Wasserstein DRO structure learning (WDRSL), KL DRO (KLDRL) and sparse logistic regression (SLR). Top, from left to right: (a) diamond, 4 classes, noiseless, $\theta = 0.2$, varying nodes; (b) diamond, 6 nodes, 4 classes, noiseless, varying θ ; (c) diamond, 6 nodes, noiseless, $\theta = 0.2$, varying classes. Bottom, from left to right: (d) grid, 9 nodes, 4 classes, $\theta = 0.2$, varying noise models with $\zeta = 0.2$; (e) grid, 9 nodes, 4 classes, $\theta = 0.2$, independent failure model, varying probability of noise; (f) grid, 9 nodes, 4 classes, $\theta = 0.2$, Huber's contamination model, varying noise level.	51
3	Comparisons of the execution time of one run. $\theta = 0.2$ and noiseless model are adopted in all settings. $\kappa = 1$, $\varepsilon_0 = 33$ for KLDRL and $\varepsilon_0 = 1.5$ for WDRSL. From left to right: (a) grid, 9 nodes, 4 classes, varying samples; (b) diamond, 4 classes, varying nodes; (c) diamond, 3 nodes, varying classes.	52
4	Convergence of ADMM and FW for random points with 95% confidence intervals.	138
5	The best UAS with the Marginal algorithm as μ and λ vary in logarithmic scales.	139
6	The expected value of four loss functions for three classes with $\mathbb{Q}_Y(1) = 0.6$ and $\mathbb{Q}_Y(2) = \mathbb{Q}_Y(3) = 0.2$. $\mathbb{P}_Y(2) = \mathbb{P}_Y(3)$ as $\mathbb{P}_Y(1)$ varies. Each loss is normalized to cross $(1, 0)$ and $(0.5, 0.5)$ according to the binary case with a hard label. Best viewed in color.	148
7	Normalized generalization losses with different coefficients or noise levels. Left: varying β in $[0.001, 0.1]$. Right: varying probability of contamination in $[0, 0.5]$. The X axes of the left subfigure is in logarithmic scale. Best viewed in color.	165

SUMMARY

Decision-making under uncertainty is common in various areas of study. Structural learning is a decision problem that involves seeking the optimal structure typically from an exponential number of structures. The task is usually performed on a finite set of samples observed from uncertain environments, which may be subject to unexpected contamination thus unreliable. The combinatorial nature and uncertainty pose challenges to relevant algorithms, particularly in the large-scale setting. We suggest that a successful structural learning method should have low time complexity, high sample efficiency, estimator consistency and robustness at the same time. In this thesis, we propose a statistical learning framework that fulfills these requirements to tackle several structural learning problems based on techniques in the emerging fields of distributionally robust optimization (DRO). Our models hedge against a set of distributions consistent with data in terms of certain a priori assumptions. The set constitutes our uncertainty about the underlying data-generating mechanism and can be constructed in a flexible way. We establish desirable theoretical guarantees and put forward practical algorithms for specific learning problems with judiciously chosen uncertainty sets.

In the first two parts of the thesis, we study structure learning problems whose goal is to recover the graphical structure of a probabilistic graphical model from samples. The only assumptions we make are bounded model weights for undirected graphs and restricted eigenvalue as well as mutual incoherence for directed graphs without faithfulness. Motivated by neighborhood selection methods, we propose to optimize the worst-case expected regression loss

SUMMARY (Continued)

over all distributions within bounded Wasserstein distances or Kullback-Leibler divergences. First, we present iterative algorithms that find the optimal estimator with per-iteration polynomial time complexity. Second, we illustrate equivalence between our Wasserstein DRO method and baseline regularization methods. Third, we derive near-optimal sample complexities for the proposed methods that match the state-of-the-art results. Experiments are conducted on simulated and real-world data.

In the last two parts of the thesis, we consider structured prediction problems which are supervised learning problems whose goal is to learn a mapping from features to structures. Optimizing typical performance metrics with training data is usually intractable and elicits a surrogate loss for efficiency. Fisher consistency is highly desirable in this setting which requires the surrogate to yield Bayes optimal prediction given population distribution. Inspired by the success of existing DRO methods with moment-based ambiguity sets, we propose similar formulations for tree structured prediction and discrete probabilistic supervised learning. We generalize existing theoretical results. Specifically, we show that such DRO problems are exactly equivalent to a regularized empirical risk minimization problem with strong duality. Fisher consistency is established by relating them to Fenchel-Young losses. Novel excess true risk bounds are derived based on uniform convergence. For such class of structured prediction methods, we illustrate their practicability by showing how to incorporate deep learning techniques into the framework for end-to-end representation learning.

CHAPTER 1

INTRODUCTION

1.1 Structural Learning

In the big data era, machine learning approaches are widely adopted to extract information and patterns from a massive amount of data. These data-driven methods enable automatic decision making and provide valuable insights into data. Out of many applications, an important class of tasks aims at understanding structural relationships among objects with respect to data distribution. Producing an interpretable and compact representation for complicated data is desirable especially when there are a large number of variables of interest. The learning tasks that involve complex structures empower a lot of applications in scientific areas and daily life such as protein-protein interaction networks (Jaimovich et al., 2006), gene regulatory networks (Werhli et al., 2006), medical decision making (Kyrimi et al., 2020), spam filtering (Manjusha and Kumar, 2010), protein structure prediction (Kuhlman and Bradley, 2019), logic theorem proving (Bansal et al., 2019), source code generation (Svyatkovskiy et al., 2020), machine translation (Stahlberg, 2020), visual object localization (Yu et al., 2020), recommendation systems (He et al., 2018), search engines (Grbovic and Cheng, 2018), to name a few.

The *structural learning* tasks discussed here refer to several different tasks in the literature. We list those we study in the thesis for disambiguation but it is worth mentioning that there are dubious meanings for the term structural learning in related work. The task of *structure*

learning (Drton and Maathuis, 2017; Heinze-Deml et al., 2018) usually stands for the problem of recovering the structure of a probabilistic graphical model (PGM). PGMs (Koller and Friedman, 2009) as well as deep probabilistic models (Poon and Domingos, 2011) are useful mathematical frameworks for modeling a high-dimensional distribution equipped with a graphical structure. A structure learning algorithm yields a structure possibly associated with parameters such that the learned structure together with the parameters approximates the underlying data generating process as well as possible. In contrast, a *parameter learning* (Jordan, 1999) task assumes that the true structure is given. *Structured prediction* (Taskar et al., 2005), or *structured (output) learning*, appears unambiguously as a supervised learning task (Hastie et al., 2009) where, unlike structure learning, data comes with labels defined or designated by humans. Each label is a possible structure encoding a relationship among a subset of random variables. A structured prediction algorithm learns a hypothesis from a set of data called training data and the goal is to make the learned hypothesis a good mapping from features to labels (structures). Another supervised learning task seeks a single structure that constitutes the underlying structure of the output space and reuses it for subsequent structured prediction tasks (Meshi et al., 2013). In other words, the learned structure acts as a convenient tool for predicting the joint state of output variables. Our focus in the manuscript is the tasks of structure learning and structured prediction discussed above and we refer to both of them as structural learning.

Data-driven structural learning problems can be typically formulated as a mathematical optimization problem:

$$\inf_{f \in \mathcal{F}} -\text{Score}(f; \mathcal{D}),$$

where \mathcal{D} is a set of samples drawn from the underlying distribution \mathbb{P} we are interested in. The goal is to find a decision to maximize a given scoring function that measures the goodness of fit of the decision with respect to the actual distribution or a set of observations. *Combinatorial algorithms* normally pick one optimal structure/sub-structure out of the set of all possible structures \mathcal{F} . There are usually an exponential number of candidate structures. This includes exact search (Parviainen and Koivisto, 2009; De Campos et al., 2009), greedy search (Jalali et al., 2011a; Chickering, 2002), neighborhood selection (Bresler, 2015), integer programming (Martins et al., 2009; Bartlett and Cussens, 2017) and dynamic programming (Silander and Myllymäki, 2006). Note that algorithms such as PC (Spirtes and Glymour, 1991) could also be viewed as an optimization problem that decides whether to remove one edge at a time based on independence tests. *Continuous optimization algorithms*, alternatively, acquire an optimal solution over continuous variables, which can be transformed to a discrete structure afterwards. Note that the transformation, or a so-called inference method, is itself a combinatorial algorithm. The optimal parameters learned with continuous optimization play a role in helping determine a structure. For example, the learned weights of a Markov network are filtered by a threshold value to get the final structure (Wu et al., 2019) while the learned parameters in

structured prediction problems constitute a scoring function in parametric form that assigns a score to each sub-structure for inference (Smith and Smith, 2007). Algorithms completely or partially based on continuous optimization include score matching (Hyvärinen and Dayan, 2005; Zheng et al., 2018), neighborhood selection (Ravikumar et al., 2010; Wu et al., 2019), linear regression (Park et al., 2021), graphical lasso (Friedman et al., 2008; Loh and Bühlmann, 2014), maximum likelihood (McDonald and Satta, 2007), maximum margin (Martins et al., 2010), risk minimization (Stoyanov and Eisner, 2012). On account of the generality and combinatorial nature of such problems, numerous methods have been proposed for specific function classes \mathcal{F} , scoring functions $\text{Score}(\cdot)$ and distributions \mathbb{P} under various assumptions. Since there is a large body of works on structure learning and structured prediction, we refer the interested readers to survey papers (Drton and Maathuis, 2017; Heinze-Deml et al., 2018), books (Spirtes et al., 2000; Pearl, 2009; Nowozin et al., 2014; Peters et al., 2017) and the follow-up thesis chapters that study specific structural learning problems for a more detailed discussion on related work.

1.2 Challenges

The design of a structural learning method still poses several major challenges for researchers and practitioners. We argue that a good structural learning algorithm should at least encapsulate the following characteristics.

- **Computational efficiency:** The first and foremost property of a successful algorithm for data-driven problems is that it should be tractable in a reasonable amount of time. Scalability is especially crucial in the era of big data nowadays where the amount and dimension of available data for processing could be overwhelming. Therefore the minimum

requirement is for an algorithm to have polynomial time complexity or polynomial time per-iteration cost for iterative methods.

- **Sample efficiency:** Sample complexity informs us of how many samples needed in order to approximate the true optimal solution with a specified error tolerance. In the context of structural learning, this is the magnitude of the amount of samples required to exactly recover the correct structure. If the structural learning problem is cast as a statistical learning problem, sample efficiency becomes closely related to the complexity of the considered hypotheses space that could be measured by the Vapnik-Chervonenkis (VC) dimension or Rademacher complexity. Since there are information-theoretic lower bounds for the studied problems, we believe it is important for the sample complexity of a proposed algorithm to draw near to the optimal bound or at least match the state-of-the-art results.
- **Asymptotic consistency or Fisher consistency:** Unlike non-asymptotic guarantees given by a sample complexity bound, consistency refers to an asymptotic result that the estimator converges to the true parameter as the number of samples goes to infinity. Sometimes we are concerned about some statistics of the proposed estimators so that no bias is introduced. For instance, given a performance metric such as the Hamming distance or F-measure, we are interested in whether the expected metric of an estimator under the true distribution is equal to that of a global optimal solution, which is a concept called Fisher consistency in decision theory. It might be trivial for structure learning tasks but non-trivial for structured prediction tasks because the target structures vary according to input features and typical performance metrics for structures are intractable to optimize

directly. As a result, another good characteristic is that an algorithm should yield an asymptotically consistent estimator or a Fisher consistent estimator with respect to some performance metric.

- **Robustness:** The uncertainty governing a data-driven structural learning problem arises from the fact that we only have a limited knowledge of the unknown true distribution through access to a finite set of observations. A commonly used traditional way of accounting for the uncertainty is to construct a nominal distribution from observable samples by assuming that it is an estimate that faithfully represents the underlying distribution. Nevertheless, the nominal distribution may not be reliable thus not representative of the true distribution. For example, observed data may be acquired from noisy environments and experiments where data contamination happens due to measurement error, sensor failure, transmission error, missing value or a large number of unobserved uncertainties. A realization with very low probability may be absent in data but critical for applications such cost-sensitive classification. Moreover, an attacker would leverage carefully crafted adversarial examples that totally deviate from the true distribution to fool a machine learning model. Ideally, even if we have an unbiased estimator for the distribution, the uncertainty will generally be amplified in the optimization process because of the optimizer’s curse (Smith and Winkler, 2006). All of the above unexpected factors suggest that an algorithm be robust to noises in data or possess guarantees for the worst-case model performance with the input data.

To the best of our knowledge, it is hard for existing methods to tick all four boxes at the same time. For example, the exact learning approaches in (Silander and Myllymäki, 2006; Jaakkola et al., 2010) find an optimal directed acyclic graph (DAG) in an exponential worst-case running time, thus not applicable in large-scale settings. The greedy equivalence search (GES) algorithm (Chickering, 2002) may be computationally efficient in practice, but without finite-sample guarantees. In this case, we are ignorant of how well the algorithm can do with a certain set of data. Structured prediction algorithms based on log-likelihood (Koller and Friedman, 2009) or large margin learning (Tsochantaridis et al., 2005) are known to be inconsistent with respect to a prescribed loss metric in general (Nowak-Vila et al., 2019). Losing consistency causes discrepancy between learning and prediction objectives and is likely to deteriorate prediction performance. Recent advanced methods (Wu et al., 2019; Bank and Honorio, 2020) are both computationally and sample efficient but rely on instinctive regularization to combat overfitting. It is also unclear what kind of data uncertainty a regularization method is able to handle.

This dissertation pursues the goal of developing structural learning methods that take into account all the above design considerations simultaneously. We achieve this goal mainly by taking advantage of techniques in the emerging field of distributionally robust optimization (DRO). We show that the proposed methods are not only robust to different types of data uncertainties, but also enjoying desirable computational and statistical properties.

To proceed with the rest of this introduction chapter, we provide an introduction of the distributionally robust optimization framework in Section 1.3 and a brief discussion on how the above design concerns are addressed with appropriate DRO formulations in Section 1.4. The

contributions as well as an outline of the thesis are summarized in Section 1.5. We attach a description of notation conventions at the end of this chapter.

1.3 Distributionally Robust Optimization

Machine learning usually deals with decision-making problems under uncertainty, which are closely related to mathematical optimization. Mathematical optimization not only provides solutions to many machine learning problems but also inspires design of some learning models. A few modeling approaches have been proposed to tackle optimization under uncertainty including stochastic optimization, robust optimization and distributionally robust optimization, etc.

In a classical machine learning problem, we are given a class $\mathcal{P}(\Xi)$ of probability measures supported on a measurable instance space Ξ as well as a class \mathcal{F} of measurable functions $f : \Xi \rightarrow \mathbb{R}_+$, sometimes considered as a hypothesis space, where each $f \in \mathcal{F}$ assigns a scalar cost value to each instance $\xi \in \Xi$. A *stochastic optimization* approach (Shapiro et al., 2021) infers a hypothesis f^* whose expectation under a known distribution $\mathbb{P} \in \mathcal{P}(\Xi)$, is minimum or nearly optimal with high confidence:

$$\inf_{f \in \mathcal{F}} \int_{\Xi} f(\xi) \mathbb{P}(\mathrm{d}\xi). \quad (1.1)$$

In practical terms, the distribution governing uncertainty is often not accessible and computing a multivariate integral is not easy. Instead, only a finite set of in-sample data $\{\xi^{(1)}, \dots, \xi^{(m)}\}$ drawn i.i.d. from the unknown \mathbb{P} is given. On account of this, regularized *empirical risk mini-*

mization (ERM) could be adopted to construct a nominal distribution to approximate \mathbb{P} , which is the learning framework adopted by a lot of machine learning problems:

$$\inf_{f \in \mathcal{F}} \int_{\Xi} f(\xi) \tilde{\mathbb{P}}_m(d\xi) + \tilde{\lambda} \Omega(f),$$

where $\tilde{\mathbb{P}}_m = \frac{1}{m} \sum_{i=1}^m \delta_{\xi^{(i)}}$ is the uniform distribution on data with $\delta_{\xi^{(i)}}$ being the Dirac point measure at $\xi^{(i)}$, $\Omega(\cdot)$ represents a function quantifying hypothesis complexities and $\tilde{\lambda}$ is a trade-off coefficient. The regularization term $\tilde{\lambda} \Omega(f)$ is usually added to the vanilla ERM objective to combat overfitting and outlier data, which has been shown to be an implicit way of restricting the hypothesis space (Bartlett and Mendelson, 2002). A norm is a common choice while adopting different norms leads to different regularization effects. For instance, the ℓ_1 norm imposes a strong prior assumption of sparsity and results in a non-smooth problem, while the ℓ_2 norm may not be effective in feature selection or high-dimensional settings (Ng, 2004). In addition, the regularizer is instinctively added without sound probabilistic interpretation in most cases.

Another approach from modern robust optimization (Ben-Tal et al., 2009; Bertsimas et al., 2011) proposes to optimize the following objective:

$$\inf_{f \in \mathcal{F}} \sup_{\xi \in \Xi} f(\xi),$$

which does not require distributional information but only an uncertainty region Ξ consisting of possible realizations of ξ . A carefully chosen set Ξ would lead to computationally tractable problems (Trafalis and Gilbert, 2006; Yang and Xu, 2013; Bertsimas and Copenhaver, 2018).

However, the optimal decision could be very conservative because only a single cost value is considered regardless of the statistics in the samples.

Distributionally robust optimization is an intermediate remedy that combines the advantages of stochastic optimization and robust optimization. Because of the limited information about the true data-generating distribution, the DRO framework explicitly models the uncertainty by constructing an ambiguity set that possibly contains the unknown distribution based on a nominal distribution in a probabilistic way. DRO seeks to minimize the worst-case risk instead of the empirical risk:

$$\inf_{f \in \mathcal{F}} \sup_{\mathbb{Q} \in \mathcal{A}} \int_{\Xi} f(\xi) \mathbb{Q}(\mathrm{d}\xi), \quad (1.2)$$

where $\mathcal{A} \subseteq \mathcal{P}(\Xi)$ is an ambiguity set. This formulation has its origin from John von Neumann's game theory (von Neumann and Morgenstern, 1944). DRO has attracted attention recently in operations research and machine learning communities by virtue of its several advantages: (1) it admits distributional uncertainty by explicitly modeling it; (2) as long as the true distribution falls within the constructed ambiguity set, the out-of-sample performance is guaranteed to be no worse than the worst-case performance; (3) equivalence or an alternative to regularization with a theoretically sound interpretation; (4) an appropriate ambiguity set gives rise to an efficiently solvable reformulation; (5) it yields desirable statistical properties with judiciously chosen ambiguity sets; (6) possible realizations that are absent from in-sample data could be taken into account.

The ambiguity set \mathcal{A} is typically defined by a radius ε and a nominal probability measure: $\mathcal{A}_\varepsilon(\mathbb{P}) := \{\mathbb{Q} \in \mathcal{P}(\Xi) : \text{div}(\mathbb{Q}, \mathbb{P}) \leq \varepsilon\}$, where $\text{div}(\cdot, \cdot)$ measures the discrepancy between two distributions. A desirable ambiguity set incorporates characteristics of specific applications and ensures tractability. Throughout the thesis, we consider three popular choices of $\text{div}(\cdot, \cdot)$, based on feature moments, the relative entropy and the Wasserstein metric.

Definition 1. Let $\mathbb{P}, \mathbb{Q} \in \mathcal{P}(\Xi)$ be two distributions. Given a mapping $\phi : \Xi \rightarrow \mathbb{R}^d$ and a norm $\|\cdot\|$, the moment difference between \mathbb{P} and \mathbb{Q} is defined as

$$D_M(\mathbb{P}, \mathbb{Q}) := \left\| \int_{\Xi} \phi(\xi) \mathbb{P}(\mathrm{d}\xi) - \int_{\Xi} \phi(\xi) \mathbb{Q}(\mathrm{d}\xi) \right\|.$$

This divergence is able to take advantage of application-specific features and sometimes restricts the input marginal to be identical, e.g., $\mathbb{P}_{\mathbf{X}} = \mathbb{Q}_{\mathbf{X}}$ for $\Xi = \mathcal{X} \times \mathcal{Y}$. All moment-based ambiguity sets are defined through certain moment conditions and support information. DRO formulations based on moments have been shown to induce tractable reformulations (Scarf, 1958; Popescu, 2007; Delage and Ye, 2010; Goh and Sim, 2010; Zymmler et al., 2013; Wiesemann et al., 2014; Mehrotra and Zhang, 2014; Chen et al., 2019). Despite their tractability, this type of ambiguity sets promotes conservative decisions and fails to converge to a singleton with low-order moments (Shafieezadeh-Abadeh et al., 2019).

Definition 2. Let $\mathbb{Q} \in \mathcal{P}(\Xi)$ be absolutely continuous with respect to $\mathbb{P} \in \mathcal{P}(\Xi)$. Let $\frac{\mathbb{Q}(\mathrm{d}\xi)}{\mathbb{P}(\mathrm{d}\xi)}$ be the Radon-Nikodym derivative. The Kullback-Leibler (KL) divergence from \mathbb{P} to \mathbb{Q} is defined as

$$D_{KL}(\mathbb{Q} \parallel \mathbb{P}) := \int_{\Xi} \ln \frac{\mathbb{Q}(\mathrm{d}\xi)}{\mathbb{P}(\mathrm{d}\xi)} \mathbb{Q}(\mathrm{d}\xi).$$

The relative entropy, or KL divergence, arises in information theory and is a well-known asymmetric measure of difference between distributions. Unlike moments, it is a statistical distance and a special case of ϕ -divergences. Such ambiguity sets are sometimes called likelihood-based ambiguity sets. DRO formulations based on these divergences have also been shown to be tractable (Calafiore and El Ghaoui, 2006; Ben-Tal et al., 2013; Hu and Hong, 2013; Bayraksan and Love, 2015; Jiang and Guan, 2016; Wang et al., 2016; Sun and Xu, 2016; Lam, 2019) and first-order equivalent to variance regularization (Lam, 2016; Duchi and Namkoong, 2019). An obvious drawback is that all distributions in such ambiguity sets are required to be absolutely continuous with respect to the nominal distribution. In this way, the support is constrained by empirical data, which is harmful to generalization ability of the learned model. Furthermore, (Gao and Kleywegt, 2022) illustrate with an image retrieval example that divergence measures result in pathological worst-case distributions that are excessively conservative.

Definition 3. Assume that Ξ is a Polish space equipped with a metric $c : \Xi \times \Xi \rightarrow \mathbb{R}_+$. Denote by $\mathcal{P}(\Xi)$ the space of all Borel probability measures on Ξ , and by $\mathcal{P}_p(\Xi)$ the space of all $\mathbb{P} \in \mathcal{P}(\Xi)$

with finite p -th moments for $p \geq 1$. Let $\mathcal{M}(\Xi^2)$ be the set of probability measures on the product space $\Xi \times \Xi$. The p -Wasserstein distance between two distributions $\mathbb{P}, \mathbb{Q} \in \mathcal{P}_p(\Xi)$ is defined as

$$W_p(\mathbb{P}, \mathbb{Q}) := \inf_{\Pi \in \mathcal{M}(\Xi^2)} \left\{ \left[\int_{\Xi^2} c^p(\xi, \xi') \Pi(d\xi, d\xi') \right]^{\frac{1}{p}} : \Pi(d\xi, \Xi) = \mathbb{P}(d\xi), \Pi(\Xi, d\xi') = \mathbb{Q}(d\xi') \right\}.$$

Wasserstein distances originate from the study of optimal transport theory (Villani and others, 2009) and can be interpreted as the minimum cost of moving the probability measure \mathbb{P} to \mathbb{Q} with the unit transport cost quantified by $c(\xi, \xi')$. In contrast to ϕ -divergences, a Wasserstein ambiguity set includes both discrete and continuous distributions whose support is outside samples. This leads to stronger generalizability, which is further strengthened by the measure concentration results in (Bolley et al., 2007; Boissard and others, 2011; Fournier and Guillin, 2015; Singh and Póczos, 2018; Weed et al., 2019; Lei and others, 2020). What's more, it allows us to make use of custom metrics to measure a notion of closeness between points, which could be useful when paired with application-specific loss functions. Wasserstein DRO has given rise to a number of learning frameworks (Wozabal and others, 2012; Wozabal, 2014; Esfahani and Kuhn, 2018; Zhao and Guan, 2018; Chen and Paschalidis, 2018; Sinha et al., 2018; Luo and Mehrotra, 2019; Blanchet et al., 2019a; Blanchet et al., 2019b; Blanchet and Murthy, 2019; Shafieezadeh-Abadeh et al., 2019; Gao and Kleywegt, 2022; Gao et al., 2022) and has been shown to be equivalent to Lipschitz regularization (Cranko et al., 2021).

DRO has also been adopted to tackle structural learning problems or closely related problems including sub-modular maximization (Staib and Jegelka, 2019), inverse covariance estimation

(Nguyen et al., 2022), graphical lasso learning (Cisneros-Velarde et al., 2020), graph Laplacian learning (Wang et al., 2021), causal inference (Bertsimas et al., 2022), contextual bandit (Si et al., 2020) and structured prediction (Fathony et al., 2018; Fathony et al., 2018). A thorough review can be found in (Rahimian and Mehrotra, 2019; Lin et al., 2022).

1.4 Overview of Distributionally Robust Structural Learning

We propose to solve structural learning problems with the DRO framework and judiciously chosen ambiguity sets. We advocate continuous optimization methods that can be recast as a DRO problem as in Equation 1.2 since most combinatorial optimization algorithms for structural learning problems with complex structures easily become intractable in the large-scale setting. Continuous optimization methods also allow us to establish finite-sample guarantees and leverage advanced optimization techniques.

We consider the Wasserstein distances and KL divergences for structure learning problems.

Following a data-driven Wasserstein DRO framework, we find that the dual problem of a structure learning problem can be written in the form of

$$\inf_{f \in \mathcal{F}, \gamma \geq 0} \gamma \varepsilon + \frac{1}{m} \sum_{i=1}^m \sup_{\xi \in \Xi} f(\xi) - \gamma c(\xi, \xi^{(i)}),$$

which is valid and strong duality holds as long as $c(\cdot, \cdot)$ is non-negative and lower semi-continuous (Kuhn et al., 2019). In practice, a twice differentiable function f_θ parameterized by θ is usually optimized instead. A key challenge lies in solving the inner supremum problem. In typical structure learning problems, the uncertain random variable ξ has an exponential number of

states, which makes Ξ non-convex. In contrast to a classical regression setting with $\Xi = \mathbb{R}^d$, a Fenchel conjugate is not applicable here to simplify the supremum problem into a closed form. Nevertheless, we show that it can be solved with greedy algorithms in polynomial time exactly or approximately for undirected and directed graphical models with certain hypotheses classes. Based on uniform convergence (Shalev-Shwartz et al., 2010), Rademacher complexities (Bartlett and Mendelson, 2002), Lipschitz regularization (Cranko et al., 2021) and a primal-dual witness construction method (Wainwright, 2009), we derive out-of-sample performance guarantees that match the state-of-the-art sample complexities.

Structure learning problems can be tackled with a DRO method based on KL divergences as well. Established results in (Hu and Hong, 2013) allow us to reformulate such problems as a single minimization problem. Similar sample complexity bounds can be computed by noting that the worst-case risk over a KL divergence ambiguity set is equivalent to variance regularization (Lam, 2019). Although requiring absolute continuity and apparently losing modeling power for generalization, as supported by our experimental results, KL DRO is able to account for distributional uncertainty to some degree and comparable to classic regularized ERM problems in terms of efficiency, which is an advantage over the Wasserstein DRO counterpart.

Now we turn to structured prediction problems. A highly desirable property in structured prediction is Fisher consistency (Liu, 2007) of the loss function used for training. Motivated by success of moment-based ambiguity sets in DRO formulations for structured prediction problems (Fathony et al., 2018), we develop a framework with more general theoretical results. Specifically, with the help of Fenchel duality, we prove that our moment-based DRO formulation leads to a

ERM problem regularized by the dual norm associated with the norm that defines the ambiguity set. Such problems are closely connected to the Fenchel-Young loss framework (Blondel et al., 2020) that leads to Fisher consistency.

Robustness of the proposed methods is justified by the constructed ambiguity sets and empirical study that explicitly considers several data contamination models.

1.5 Contribution and Outline

The follow-up of this thesis is divided into several self-contained chapters that consider two structure learning problems and two structured prediction problems respectively.

In Chapter 2, we study the problem of learning the structure of a general discrete pairwise undirected graphical model. Building on a constrained logistic regression method, we propose two DRO approaches with tractable reformulations. The only assumptions we make are lower and upper bounds on the model weights. The contributions in this work can be summarized as follows:

- We propose the first computationally efficient and robust structure learning approach for discrete pairwise Markov random fields.
- We prove that it subsumes constrained and regularized logistic regression methods as special cases.
- We provide near-optimal sample complexities that induce robustness at little cost.
- We conduct extensive experiments on synthetic data, comparing our methods against the state-of-the-art baseline.

In Chapter 3, we revisit a structure learning problem for discrete directed graphical models. We develop DRO methods based on a group norm regularized linear regression approach. The proposed learning methods are valid under mild conditions without the faithfulness assumption. Specifically, our contributions are

- We propose the first computationally efficient and distributionally robust method for Bayesian network structure learning over purely categorical random variables.
- We illustrate the connection between the DRO formulation and group norm regularization.
- For skeleton learning, we derive its sample complexities that are polynomial for general graphs and logarithmic for bounded-degree graphs.
- Empirical study on benchmark and real-world datasets verify the effectiveness of our methods.

In Chapter 4, a structured prediction problem of tree-shaped objects is considered. We present a fresh perspective to Fisher consistent structured prediction in terms of DRO with general theoretical results. The main contributions in this work are listed below.

- We propose a distributionally robust tree structured prediction method and show its equivalence to regularized surrogate loss minimization.
- We derive its generalization bounds and Fisher consistency.
- We propose efficient algorithms based on projection oracles for arborescence polytopes.
- We perform empirical study on real-world datasets.

In Chapter 5, we tackle a probabilistic supervised learning problem that can be regarded as structured prediction of objects in a simplex. This problem traces back to consistent probability estimation in statistics. A Fisher consistent loss naturally acts as a proper scoring rule. Based on a moment-based ambiguity set, the proposed DRO approach yields consistent conditional probability distribution prediction and can be easily incorporated in an end-to-end representation learning framework. A summary of the contributions is

- We propose a distributionally robust probabilistic supervised learning method, show its Fisher consistency and derive its generalization bounds.
- We characterize the solutions to the proposed method and present an efficient algorithm for specific losses.
- We incorporate our method into neural networks and perform extensive empirical study on real-world data.

In Chapter 6, we make some discussions and conclude the thesis.

1.6 Notation

The following notation conventions are adopted throughout this thesis. We refer to $[n]$ as the index set $\{1, 2, \dots, n\}$ for a positive integer n . For a vector $\mathbf{x} \in \mathbb{R}^n$, we use x_i for its i -th element, $\mathbf{x}_{\bar{i}}$ or \mathbf{x}_{-i} for all elements excluding the i -th element and $\mathbf{x}_{\mathcal{S}}$ for the subset of elements indexed by $\mathcal{S} \subseteq [n]$. $\mathbf{x}_{i=c}$ represents $[x_1, \dots, x_{i-1}, c, x_{i+1}, \dots, x_n]^\top$ for some $c \in \mathbb{R}$. For a matrix $\mathbf{A} \in \mathbb{R}^{n \times m}$, we use A_{ij} , $\mathbf{A}_{i\cdot}$ (\mathbf{A}_{i*}) and $\mathbf{A}_{\cdot j}$ (\mathbf{A}_{*j}) to denote its (i, j) -th entry, i -th row and j -th column respectively. $\mathbf{A}_{\mathcal{ST}}$ represents the sub-matrix of \mathbf{A} with rows restricted to \mathcal{S} and columns

restricted to \mathcal{T} . We define a row-partitioned block matrix as $\mathbf{A} \triangleq [\mathbf{A}_1 \mathbf{A}_2 \cdots \mathbf{A}_k]^\top \in \mathbb{R}^{\sum_i n_i \times m}$ where $\mathbf{A}_i \in \mathbb{R}^{n_i \times m}$. The ℓ_p -norm of a vector \mathbf{x} is defined as $\|\mathbf{x}\|_p := (\sum_i |x_i|^p)^{1/p}$ with $|\cdot|$ being the absolute value function. The $\ell_{p,q}$ norm of a matrix \mathbf{A} is defined as $\|\mathbf{A}\|_{p,q} := (\sum_j \|\mathbf{A}_{\cdot j}\|_p^q)^{1/q}$. When $p = q = 2$, the $\ell_{p,q}$ norm becomes the Frobenius norm or the Hilbert–Schmidt norm denoted by $\|\cdot\|_F$. We define the operator norm of a matrix as $\|\mathbf{A}\|_{p,q} := \sup_{\|\mathbf{v}\|_p=1} \|\mathbf{A}\mathbf{v}\|_q$. The block matrix norm is defined similarly: $\|\mathbf{A}\|_{B,p,q} := (\sum_{i=1}^k \|\mathbf{A}_i\|_p^q)^{1/q}$. The inner product of two matrices is designated by $\langle \mathbf{A}, \mathbf{B} \rangle \triangleq \text{Tr}[\mathbf{A}^\top \mathbf{B}]$ where \mathbf{A}^\top is the transpose of \mathbf{A} . The Hadamard product is written as $\mathbf{A} \odot \mathbf{B}$ for element-wise multiplication. Denote by \otimes the tensor product operation. With a slight abuse of notation, $|\mathcal{S}|$ or $\#\mathcal{S}$ stands for the cardinality of a set \mathcal{S} . We denote by $\mathbb{T}(\mathbf{x}) \in \mathbb{R}^n$ a vector with non-decreasing components as a result of sorting $(x_i : i \in [n])$. We denote by $\mathbf{1}$ ($\mathbf{0}$) a vector or matrix of all ones (zeros). Given a distribution \mathbb{P} on Ξ , we denote by \mathbb{P}^m the m -fold product of \mathbb{P} on the Cartesian product Ξ^m and by $\mathbb{E}_{\mathbb{P}}$ the expectation under \mathbb{P} . The least c -Lipschitz constant of a function $f : \Xi \rightarrow \mathbb{R}$ with a metric $c : \Xi \times \Xi \rightarrow \mathbb{R}_+$ is written as $\text{lip}_c(f) := \inf \Lambda_c(f)$ where $\Lambda_c(f) := \{\lambda > 0 : \forall \xi_1, \xi_2 \in \Xi \quad |f(\xi_1) - f(\xi_2)| \leq \lambda c(\xi_1, \xi_2)\}$.

The i -th standard basis vector is written as $\mathbf{b}^{(i)}$ with $b_i^{(i)} = 1$ and $b_j^{(i)} = 0$ for $j \neq i$. Denote $\mathcal{B} := \{\mathbf{b}^{(i)} : i \in [k]\}$ as the set of basis vectors in \mathbb{R}^k and $\mathbf{B}^{(n \times k)} \subset \{0, 1\}^{n \times k}$ as the set of all $n \times k$ matrices whose rows are k -dimensional standard basis vectors.

CHAPTER 2

DISTRIBUTIONALLY ROBUST STRUCTURE LEARNING OF UNDIRECTED GRAPHICAL MODELS

(Parts of this chapter were previously published as “Distributionally Robust Structure Learning for Discrete Pairwise Markov Networks” in Proceedings of the 25th International Conference on Artificial Intelligence and Statistics (AISTATS) 2022 (Li et al., 2022b).)

In this chapter, we consider the problem of learning the underlying structure of a general discrete pairwise Markov network. Existing approaches that rely on ERM may perform poorly in settings with noisy or scarce data. To overcome these limitations, we propose a computationally efficient and robust learning method for this problem with near-optimal sample complexities based on DRO and maximum conditional log-likelihood. We describe the motivation and related work in Section 2.1, followed by preliminaries with a baseline approach in Section 2.2. In Section 2.3, we propose a minimax learning formulation and show that it can be efficiently solved by leveraging sufficient statistics and greedy maximization in the ostensibly intractable dual formulation. Based on DRO’s approximation to Lipschitz and variance regularization, we derive near-optimal sample complexities matching existing results in Section 2.4. Extensive empirical evidence in Section 2.5 with different corruption models corroborates the effectiveness of the proposed methods. This chapter is concluded with a few discussions in Section 2.6.

2.1 Introduction and Related Work

Undirected graphical models, also known as Markov random fields (MRFs) or Markov networks, are an influential framework for modeling structured high-dimensional probability distributions. The underlying graphical structure specifying the distribution encodes conditional independencies among a set of random variables and provides valuable information about their correlations. One of the core problems in graphical models is structure learning, whose goal is to recover the dependency graph with high confidence given i.i.d. samples drawn from the distribution. A flurry of work focuses on developing efficient algorithms for structure learning of discrete pairwise and higher-order MRFs (Vuffray et al., 2016; Klivans and Meka, 2017; Hamilton et al., 2017; Wu et al., 2019; Vuffray et al., 2020). These methods have almost exclusively made the assumption that the samples are not contaminated. In practice, however, noisy data is prevalent due to sensor failure, decentralized collection, or even adversarial perturbation (Nikolakakis et al., 2019a).

Existing algorithms based on neighborhood selection typically optimize a convex objective for each node to find its adjacent nodes. This essentially becomes a standard ERM problem in statistical learning. Regularization is usually added to the vanilla ERM objective to combat overfitting and outlier data, which has been shown to be an implicit way of restricting the hypothesis space (Bartlett and Mendelson, 2002).

To alleviate the above issues, we put forward a distributionally robust optimization approach for solving a node-wise maximum log-likelihood problem for structure learning of pairwise MRFs over a general alphabet. The presence of data corruption and limited sample sizes are of

particular interest for our approach. In contrast to regularized ERM that suppresses hypothesis complexity, the DRO method makes no restriction on parameters to be optimized. To account for uncertainty about the true distribution due to noisy finite samples, it explicitly constructs an ambiguity set of distributions consistent with the true distribution pertaining to certain a priori properties. The optimal decision rule is then found by minimizing the worst-case expected cost over the ambiguity set so that it has the best performance evaluated by all adversarial distributions in the set. If the true distribution is included in the uncertainty set, it has implicitly optimized the estimator on it. The worst-case risk thus serves as an upper confidence bound on the true expected loss. An exponential number of outcomes in the discrete probability space of MRFs makes the naïve dual formulation based on the Wasserstein distance NP-hard thus intractable. By exploiting the greedy property of finding the worst-case risk, we reformulate the primal DRO problems based on the Wasserstein distance and KL divergence into efficiently solvable convex optimization problems. Furthermore, the DRO approach has better probabilistic elucidation than standard regularization. We show that it encompasses both the $\ell_{2,1}$ -constrained and $\ell_{2,1}$ -regularized logistic regression as special cases. It is inherently robust due to explicitly modeling distributional uncertainty. Based on Lipschitz and variance regularization, we derive near-optimal sample complexities with an additional linear term with ambiguity radius as its coefficient. Extensive experiments in different settings including three contamination models are conducted to validate our method against the state-of-the-art baseline (Wu et al., 2019), which is hardly done in related work.

Contribution. Our contributions can be summarized as follows: (1) We propose the first computationally efficient and robust structure learning approach for discrete pairwise MRFs and prove that it subsumes existing methods as special cases. (2) We provide near-optimal sample complexities that induce robustness at little cost. (3) We conduct extensive experiments on synthetic data, comparing our methods against the state-of-the-art baseline.

2.1.1 Related Work

The MRF structure learning task plays an essential role in applications in a number of areas such as statistical mechanics (Chayes et al., 1984), computer vision (Szeliski et al., 2006), sociology (Eagle et al., 2009) and neuroscience (Schneidman et al., 2006).

There has been a rich body of work on structure learning of Ising models as well as non-binary higher-order MRFs. The study of this problem was initiated by the seminal work of (Chow and Liu, 1968) on the maximum likelihood estimator of a tree-structured MRF. Early attempts include hypothesis testing (Spirtes et al., 2000), exhaustive neighborhood search (Bresler et al., 2013) and regularized pseudo-likelihood (Ravikumar et al., 2010; Jalali et al., 2011b). (Bresler, 2015) put forward a simple greedy algorithm that learns the structure of any sparse bounded-degree Ising models, which was improved to near-optimal sample complexity (Vuffray et al., 2016; Lokhov et al., 2018) and generalized to arbitrary MRFs (Hamilton et al., 2017; Vuffray et al., 2020). A multiplicative weight update approach called Sparsitron, achieving near-optimal run-time and near-optimal sample efficiency, was introduced by (Klivans and Meka, 2017). (Wu et al., 2019) revisited the classical regularized likelihood method (Ravikumar et al., 2010) and

made a slight improvement over the sample complexity of Sparsitron with respect to dependence on model width.

The Ising model structure learning problem under the missing data setting was raised as an open problem by (Chen, 2010). Preliminary unidentifiability results on robust learning of Ising models were derived by (Lindgren et al., 2019). Provably robust binary Ising model structure learning algorithms were developed for independent failure corruption (Goel et al., 2019), tree-structured Ising model (Nikolakakis et al., 2019a; Katiyar et al., 2020), Huber’s contamination model (Prasad et al., 2020) and total variation contamination (Diakonikolas et al., 2021). Robust structure learning methods for non-binary MRFs were studied in (Nikolakakis et al., 2019b) and (Katiyar et al., 2021) by assuming a tree-shaped underlying graph. To the best of our knowledge, there has been no robust structure learning algorithms for non-binary MRFs without structural constraints on the true graph.

2.2 Preliminaries

To begin with, we consider the definition of a general discrete pairwise MRF.

Definition 4. Let k be the alphabet size. Let $\mathcal{W} = \{\mathbf{W}^{(ij)} \in \mathbb{R}^{k \times k} : i \neq j \in [n]\}$ be a collection of symmetric weight matrices and $\Theta = \{\boldsymbol{\theta}^{(i)} \in \mathbb{R}^k : i \in [n]\}$ be a collection of external field vectors. Let $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ be an undirected graph with nodes $\mathcal{V} = [n]$ and edges $\mathcal{E} = \{\{i, j\} \subseteq \mathcal{V} : \mathbf{W}^{(ij)} \neq \mathbf{0}\}$. Then the n -variable pairwise undirected graphical model with underlying dependency graph \mathcal{G} is a distribution $\mathcal{D} \equiv \mathcal{D}(\mathcal{W}, \Theta)$ over $[k]^n$ such that

$$\mathbb{P}_{\mathbf{Z} \sim \mathcal{D}(\mathcal{W}, \Theta)}[\mathbf{Z} = \mathbf{z}] \propto \exp \left(\sum_{i < j \in [n]} W_{z_i z_j}^{(ij)} + \sum_{i \in [n]} \theta_{z_i}^{(i)} \right).$$

Define the width of the model as $\lambda(\mathcal{D}) := \sup_{i \in [n], a \in [k]} \left(\sum_{j \neq i \in [n]} \sup_{b \in [k]} |W_{ab}^{(ij)}| + |\theta_a^{(i)}| \right)$ and the minimum edge weight as $\eta(\mathcal{D}) := \inf_{\{i,j\} \in \mathcal{E}} \sup_{a,b \in [k]} |W_{ab}^{(ij)}|$.

We make the following assumptions on $\mathcal{D}(\mathcal{W}, \Theta)$.

Assumption 5. $\mathbf{W}^{(ij)}$ has centered rows and columns: $\sum_{a \in [k]} W_{ab}^{(ij)} = \sum_{b \in [k]} W_{ab}^{(ij)} = 0$.

Assumption 6. The model width is upper bounded by a positive constant λ : $\lambda(\mathcal{D}) \leq \lambda$. The minimum edge weight is lower bounded by a positive constant η : $\eta(\mathcal{D}) \geq \eta$.

According to Fact 8.2 in (Klivans and Meka, 2017), Assumption 5 is made without loss of generality because centering (\mathcal{W}, Θ) leads to (\mathcal{W}', Θ') with the same distribution: $\mathcal{D}(\mathcal{W}, \Theta) = \mathcal{D}(\mathcal{W}', \Theta')$. One of the useful properties induced by Assumption 6 is that the node-wise conditional distributions are bounded away from 0 and 1. Although η is usually assumed to be known, in practice it can be determined based on the tail of the learned weights distribution in the vicinity of zero.

We note the following fact that the conditional distributions of a pairwise MRF can be written as a logistic function $\sigma(x) := (1 + e^{-x})^{-1}$ if the dependent variable is restricted to a pair of values.

Fact 7. Let $\mathbf{Z} \sim \mathcal{D}(\mathcal{W}, \Theta)$ be a discrete pairwise graphical model over $[k]^n$. For any $i \in [n]$ and $\alpha \neq \beta \in [k]$, we have

$$\mathbb{P}[Z_i = \alpha | Z_i \in \{\alpha, \beta\}, \mathbf{Z}_{-i} = \mathbf{z}_{-i}] = \sigma\left(\sum_{j \neq i} (W_{\alpha z_j}^{(ij)} - W_{\beta z_j}^{(ij)}) + \theta_\alpha^{(i)} - \theta_\beta^{(i)}\right) \triangleq \sigma(\langle \bar{\mathbf{W}}, \bar{\mathbf{Z}} \rangle),$$

where $\bar{\mathbf{W}} \in \mathbb{R}^{n \times k}$ is defined as $\bar{\mathbf{W}}_{i*} := [\theta_\alpha^{(i)} - \theta_\beta^{(i)}, \mathbf{0}^\top]$, and $\bar{\mathbf{W}}_{j*} := \mathbf{W}_{\alpha*}^{(ij)} - \mathbf{W}_{\beta*}^{(ij)}$ for $j \neq i \in [n]$.

$\bar{\mathbf{Z}} := \text{OneHot}(\mathbf{z}_{i=1}) \in \mathbf{B}^{(n \times k)}$ encodes $\mathbf{z}_{i=1}$ such that $\bar{\mathbf{Z}}_{i*} = \mathbf{b}^{(1)\top}$ and $\bar{\mathbf{Z}}_{j*} = \mathbf{b}^{(z_j)\top}$ for $j \neq i$.

The definition of $\bar{\mathbf{W}}$ implies $\|\bar{\mathbf{W}}^\top\|_{2,1} \leq 2\lambda\sqrt{k}$. Let $\{\bar{\mathbf{z}}^{(1)}, \dots, \bar{\mathbf{z}}^{(m)}\} \stackrel{iid}{\sim} \mathcal{D}(\mathcal{W}, \Theta)$ be a set of m samples and $\{\mathbf{z}^{(1)}, \dots, \mathbf{z}^{(m')}\}$ be its subset with $z_i^{(j)} \in \{\alpha, \beta\}$. Define $y^{(j)} = 1$ if $z_i^{(j)} = \alpha$ and $y^{(j)} = -1$ if $z_i^{(j)} = \beta$. In order to estimate the graph parameters \mathcal{W} , it is thus natural to solve an $\ell_{2,1}$ -constrained logistic regression problem by minimizing the negative conditional log-likelihood for each $i \in [n]$ and $\alpha \neq \beta \in [k]$ as follows:

$$\hat{\mathbf{W}}^{(i\alpha\beta)} \in \arg \inf_{\substack{\mathbf{W} \in \mathbb{R}^{n \times k}, \\ \|\mathbf{W}^\top\|_{2,1} \leq 2\lambda\sqrt{k}}} \frac{1}{m'} \sum_{j=1}^{m'} \ell(y^{(j)} \langle \mathbf{W}, \text{OneHot}(\mathbf{z}_{i=1}^{(j)}) \rangle), \quad (2.1)$$

where $\ell(x) := \ln(1 + e^{-x}) \triangleq -\ln \sigma(x)$ represents the logistic loss function. Centering $\hat{\mathbf{W}}^{(i\alpha\beta)}$ as

$$\begin{aligned} \mathbf{U}_{i*}^{(i\alpha\beta)} &:= \hat{\mathbf{W}}_{i*}^{(i\alpha\beta)} + \frac{1}{k} \sum_{j \neq i \in [n], a \in [k]} \hat{\mathbf{W}}_{ja}^{(i\alpha\beta)} \mathbf{1}^\top \\ \mathbf{U}_{j*}^{(i\alpha\beta)} &:= \hat{\mathbf{W}}_{j*}^{(i\alpha\beta)} - \frac{1}{k} \sum_{a \in [k]} \hat{\mathbf{W}}_{ja}^{(i\alpha\beta)} \mathbf{1}^\top \quad \forall j \neq i, \end{aligned} \quad (2.2)$$

yields a minimizer of Equation 2.1 due to $\langle \hat{\mathbf{W}}^{(i\alpha\beta)}, \bar{\mathbf{Z}} \rangle = \langle \mathbf{U}^{(i\alpha\beta)}, \bar{\mathbf{Z}} \rangle$.

Finally, we can estimate the weight matrices $\mathbf{W}^{(ij)}$ via

$$\hat{\mathbf{W}}_{\alpha*}^{(ij)} := \frac{1}{k} \sum_{\beta \in [k]} \mathbf{U}_{j*}^{(\alpha\beta)} \quad \forall j \neq i \in [n], \alpha \in [k]. \quad (2.3)$$

The edge set of the estimated dependency graph can be formed by thresholding (Ravikumar et al., 2010; Wu et al., 2019):

$$\hat{\mathcal{E}} := \{\{i, j\} : \|\hat{\mathbf{W}}^{(ij)}\|_{\infty} \geq \eta/2, i < j \in [n]\}. \quad (2.4)$$

2.3 Distributionally Robust Structure Learning

We propose to reconstruct the structure of a discrete pairwise undirected graphical model with a distributionally robust learning framework, inspired by the $\ell_{2,1}$ -constrained logistic regression approach and the DRO framework. In this section, we present our DRO formulation and its dual formulations that give rise to tractable convex programs. We additionally show connections of our method to regularized ERM as well as $\ell_{2,1}$ -constrained logistic regression.

2.3.1 Distributionally Robust Discrete Pairwise Markov Network Learning

In the setting where the in-sample data is sparse or noisy, directly applying the sparse logistic regression approach usually results in a problematic dependency graph with missing or spiky edges due to overfitting. In consideration of uncertainty about the unknown true distribution, based on the logistic objective, we propose to learn pairwise MRFs by minimizing the worst-case risk taken over an ambiguity set centered at the empirical probability measure:

Definition 8. Let $\Xi = \mathcal{X} \times \mathcal{Y} = \mathbf{B}^{((n-1) \times k)} \times \{-1, 1\}$. Given m samples $\{\bar{\mathbf{z}}^{(1)}, \dots, \bar{\mathbf{z}}^{(m)}\} \stackrel{iid}{\sim} \mathcal{D}(\mathcal{W}, \Theta)$, the goal of learning discrete pairwise MRFs with distributionally robust logistic

regression is to find the optimal $\hat{\mathbf{W}}^{(i\alpha\beta)}$ for each $i \in [n]$ and $\alpha \neq \beta \in [k]$ via minimax statistical learning, formally,

$$\hat{\mathbf{W}}^{(i\alpha\beta)} \in \arg \inf_{\mathbf{W} \in \mathbb{R}^{n \times k}} \sup_{\mathbb{Q} \in \mathcal{A}_\varepsilon(\hat{\mathbb{P}}_{m'})} \int_{\Xi} \ell(y \langle \mathbf{W}, \mathbf{X} \rangle) \mathbb{Q}(\mathrm{d}(\mathbf{x}, y)), \quad (2.5)$$

where $\mathbf{X} := [\mathbf{x}_{1 \dots i-1, *}^\top, \mathbf{b}^{(1)\top}, \mathbf{x}_{i \dots n-1, *}^\top]^\top$ inserts the first standard basis vector into the i -th row of \mathbf{x} . $\hat{\mathbb{P}}_{m'}$ is the empirical distribution for a set of transformed m' samples $\{\boldsymbol{\xi}^{(1)}, \dots, \boldsymbol{\xi}^{(m')}\}$ such that, for any $\boldsymbol{\xi}^{(j')} = (\mathbf{x}^{(j')}, y^{(j')}) \in \Xi, j' \in [m']$ and its corresponding original sample $\bar{\mathbf{z}}^j, j \in [m]$, we have $\bar{z}_i^{(j)} \in \{\alpha, \beta\}$, $y^{(j')} = 1$ if $\bar{z}_i^{(j)} = \alpha$ and $y^{(j')} = -1$ if $\bar{z}_i^{(j)} = \beta$, with $\mathbf{x}^{(j')} = \text{OneHot}(\bar{\mathbf{z}}_{-i}^{(j)})$.

Note that if ε is set to zero, Equation 2.5 reduces to an unconstrained version of Equation 2.1. More importantly, the DRO formulation in Equation 2.5 is an infinite-dimensional optimization problem, which is generally impossible to solve directly.

2.3.2 Tractable Reformulations

We show that the DRO problem in Definition 8 can be solved efficiently via its dual formulations. The following theorem presents a tractable convex reformulation for the primal problem in Equation 2.5 if a Wasserstein ball is adopted as the ambiguity set.

Theorem 9. *Let $W_1(\cdot, \cdot)$ be the type-1 Wasserstein distance with $p = 1$ and metric $c(\boldsymbol{\xi}, \boldsymbol{\xi}') \triangleq c((\mathbf{x}, y), (\mathbf{x}', y')) := \|\mathbf{x} - \mathbf{x}'\|_{1,1} + \frac{\kappa}{2}|y - y'|$ for $\boldsymbol{\xi}, \boldsymbol{\xi}' \in \Xi$, $\kappa \in \mathbb{R}_+$. Let $\mathcal{A}_\varepsilon^{W_1}(\hat{\mathbb{P}}_{m'}) := \{\mathbb{Q} \in \mathcal{P}(\Xi) :$*

$W_1(\hat{\mathbb{P}}_{m'}, \mathbb{Q}) \leq \varepsilon = \frac{\varepsilon_0}{\sqrt{m'}}$ be the ambiguity set. Then the primal problem in Equation 2.5 is equivalent to

$$\inf_{\substack{\mathbf{W} \in \mathbb{R}^{n \times k}, \\ \gamma \geq 0}} \gamma \varepsilon + \frac{1}{m'} \sum_{j=1}^{m'} \sup_{\substack{0 \leq r \leq n-1 \in \mathbb{Z}, \\ g \in \{-1, 1\}}} \left[-\frac{1}{2} \gamma \kappa(1 + g y^{(j)}) - 2r\gamma + \ln(1 + e^{\langle g\mathbf{W}, \mathbf{X}^{(j)} \rangle + \langle \mathbb{T}(\boldsymbol{\delta})_{1 \dots r}, \mathbf{1} \rangle}) \right], \quad (2.6)$$

where $\mathbf{X}^{(j)} := [\mathbf{x}_{1 \dots i-1, *}^{(j)\top}, \mathbf{b}^{(1)\top}, \mathbf{x}_{i \dots n-1, *}^{(j)\top}]^\top$, $\boldsymbol{\delta} := [\sup_{l \in [k]} (g\mathbf{W})_{jl} : j \neq i \in [n]]^\top - (g\mathbf{W}_{-i, *} \odot \mathbf{x}^{(j)}) \mathbf{1}$, and $\mathbb{T}(\mathbf{x})$ is defined as a vector with non-decreasing components as a result of sorting x , introduced in Section 1.6.

Proof. Recall that $\Xi = \mathbf{B}^{((n-1) \times k)} \times \{-1, 1\}$ where $\mathbf{B}^{((n-1) \times k)}$ is the set of matrices with rows of basis vectors. To avoid clutter of notations, we define

$$\ell_{\mathbf{W}}(\boldsymbol{\xi}) := \ell(y \langle \mathbf{W}, [\mathbf{x}_{1 \dots i-1, *}^\top, \mathbf{b}^{(1)\top}, \mathbf{x}_{i \dots n-1, *}^\top]^\top \rangle).$$

Similar to (Abadeh et al., 2015), we rewrite the worst-case risk in Equation 2.5 as

$$\sup_{\mathbb{Q} \in \mathcal{A}_\varepsilon(\hat{\mathbb{P}}_{m'})} \int_{\Xi} \ell_{\mathbf{W}}(\boldsymbol{\xi}') \mathbb{Q}(d\boldsymbol{\xi}') = \begin{cases} \sup_{\Pi \in \mathcal{M}(\Xi^2)} & \int_{\Xi} \ell_{\mathbf{W}}(\boldsymbol{\xi}') \Pi(d\boldsymbol{\xi}') \\ \text{s.t.} & \int_{\Xi^2} d(\boldsymbol{\xi}, \boldsymbol{\xi}') \Pi(d\boldsymbol{\xi}, d\boldsymbol{\xi}') \leq \varepsilon \\ & \Pi(d\boldsymbol{\xi}, \Xi) = \hat{\mathbb{P}}_{m'}(d\boldsymbol{\xi}). \end{cases}$$

Plugging $\Pi(d\boldsymbol{\xi}, d\boldsymbol{\xi}') = \frac{1}{m'} \sum_{j=1}^{m'} \delta_{\boldsymbol{\xi}^{(j)}}(d\boldsymbol{\xi}) \mathbb{Q}^{(j)}(d\boldsymbol{\xi}')$ into the above expression yields

$$\left\{ \begin{array}{l} \sup_{\mathbb{Q}^{(j)}} \quad \frac{1}{m'} \sum_{j=1}^{m'} \int_{\Xi} \ell_{\mathbf{W}}(\boldsymbol{\xi}') \mathbb{Q}^{(j)}(d\boldsymbol{\xi}') \\ \text{s.t.} \quad \frac{1}{m'} \sum_{j=1}^{m'} \int_{\Xi} d(\boldsymbol{\xi}^{(j)}, \boldsymbol{\xi}') \mathbb{Q}^{(j)}(d\boldsymbol{\xi}') \leq \varepsilon \\ \int_{\Xi} \mathbb{Q}^{(j)}(d\boldsymbol{\xi}') = 1, \forall j \in [m']. \end{array} \right. \quad (2.7)$$

By defining $\mathbb{Q}_{\pm 1}^{(j)}(d\mathbf{x}) := \mathbb{Q}^{(j)}(d(\mathbf{x}, \pm 1))$, we are able to decompose $\mathbb{Q}^{(j)}(d\boldsymbol{\xi})$ based on the value of y as

$$\mathbb{Q}^{(j)}(d\boldsymbol{\xi}) = \mathbb{Q}_{-1}^{(j)}(d\mathbf{x}) + \mathbb{Q}_{+1}^{(j)}(d\mathbf{x}),$$

which can simplify Equation 2.7 to

$$\left\{ \begin{array}{l} \sup_{\mathbb{Q}_{\pm 1}^{(j)}} \quad \frac{1}{m'} \sum_{j=1}^{m'} \int_{\mathbf{B}^{((n-1) \times k)}} \ell_{\mathbf{W}}(\mathbf{x}', -1) \mathbb{Q}_{-1}^{(j)}(d\mathbf{x}') + \ell_{\mathbf{W}}(\mathbf{x}', +1) \mathbb{Q}_{+1}^{(j)}(d\mathbf{x}') \\ \text{s.t.} \quad \frac{1}{m'} \sum_{j=1}^{m'} \int_{\mathbf{B}^{((n-1) \times k)}} d(\boldsymbol{\xi}^{(j)}, (\mathbf{x}', -1)) \mathbb{Q}_{-1}^{(j)}(d\mathbf{x}') + d(\boldsymbol{\xi}^{(j)}, (\mathbf{x}', +1)) \mathbb{Q}_{+1}^{(j)}(d\mathbf{x}') \leq \varepsilon \\ \int_{\mathbf{B}^{((n-1) \times k)}} \mathbb{Q}_{-1}^{(j)}(d\mathbf{x}') + \mathbb{Q}_{+1}^{(j)}(d\mathbf{x}') = 1, \forall j \in [m']. \end{array} \right.$$

By substituting the metric definition into the above expressions, we rewrite them as

$$\left\{ \begin{array}{l} \sup_{\mathbb{Q}_{\pm 1}^{(j)}} \frac{1}{m'} \sum_{j=1}^{m'} \int_{\mathbf{B}^{((n-1) \times k)}} \ell_{\mathbf{W}}((\mathbf{x}', -1)) \mathbb{Q}_{-1}^{(j)}(d\mathbf{x}') + \ell_{\mathbf{W}}((\mathbf{x}', +1)) \mathbb{Q}_{+1}^{(j)}(d\mathbf{x}') \\ \text{s.t.} \quad \frac{1}{m'} \int_{\mathbf{B}^{((n-1) \times k)}} \kappa \sum_{j:y^{(j)}=-1} \mathbb{Q}_{+1}^{(j)}(d\mathbf{x}') + \kappa \sum_{j:y^{(j)}=+1} \mathbb{Q}_{-1}^{(j)}(d\mathbf{x}') \\ \quad + \sum_{j=1}^{m'} \|\mathbf{x}^{(j)} - \mathbf{x}'\|_1 (\mathbb{Q}_{-1}^{(j)}(d\mathbf{x}') + \mathbb{Q}_{+1}^{(j)}(d\mathbf{x}')) \leq \varepsilon \\ \int_{\mathbf{B}^{((n-1) \times k)}} \mathbb{Q}_{-1}^{(j)}(d\mathbf{x}') + \mathbb{Q}_{+1}^{(j)}(d\mathbf{x}') = 1, \forall j \in [m']. \end{array} \right.$$

Its Lagrange dual problem is as follows:

$$\left\{ \begin{array}{l} \inf_{\gamma, s^{(j)}} \quad \gamma \varepsilon + \frac{1}{m'} \sum_{j=1}^{m'} s^{(j)} \\ \text{s.t.} \quad \sup_{\mathbf{x}' \in \mathbf{B}^{((n-1) \times k)}} \ell_{\mathbf{W}}((\mathbf{x}', -1)) - \gamma \|\mathbf{x}^{(j)} - \mathbf{x}'\|_1 - \frac{1}{2} \gamma \kappa (1 + y^{(j)}) \leq s^{(j)} \quad \forall j \in [m'] \\ \sup_{\mathbf{x}' \in \mathbf{B}^{((n-1) \times k)}} \ell_{\mathbf{W}}((\mathbf{x}', +1)) - \gamma \|\mathbf{x}^{(j)} - \mathbf{x}'\|_1 - \frac{1}{2} \gamma \kappa (1 - y^{(j)}) \leq s^{(j)} \quad \forall j \in [m'] \\ \gamma \geq 0. \end{array} \right.$$

Strong duality holds according to Theorem 1 in (Gao and Kleywegt, 2022). By incorporating the outer minimization of Equation 2.5, plugging in the expression of $\ell_{\mathbf{W}}(\cdot)$ and rearranging the terms in the above expressions, we have

$$\inf_{\substack{\mathbf{W} \in \mathbb{R}^{n \times k}, \\ \gamma \geq 0}} \gamma \varepsilon + \frac{1}{m'} \sum_{j=1}^{m'} \sup_{\substack{\mathbf{x} \in \mathbf{B}^{((n-1) \times k)}, \\ g \in \{-1, 1\}}} \ln(1 + e^{g \langle \mathbf{W}, \mathbf{x} \rangle}) - \gamma \|\mathbf{x}^{(j)} - \mathbf{x}\|_1 - \frac{1}{2} \gamma \kappa (1 + g y^{(j)}),$$

where $\mathbf{X} = [\mathbf{x}_{1\dots i-1,*}^\top, \mathbf{b}^{(1)\top}, \mathbf{x}_{i\dots n-1,*}^\top]^\top$. The objective of the above convex program is the sum of m' point-wise maximum functions of $2k^{n-1}$ convex functions. We now consider the following function of \mathbf{x} :

$$h(\mathbf{x}) = \ln(1 + e^{\langle g\mathbf{W}, \mathbf{X} \rangle}) - \gamma \|\mathbf{x}^{(j)} - \mathbf{x}\|_1 - \frac{1}{2} \gamma \kappa(1 + gy^{(j)}).$$

Let $\mathbf{X}^{(j)} := [\mathbf{x}_{1\dots i-1,*}^{(j)\top}, \mathbf{b}^{(1)\top}, \mathbf{x}_{i\dots n-1,*}^{(j)\top}]^\top$ and $\boldsymbol{\delta} := [\sup_{l \in [k]} (g\mathbf{W})_{jl} : j \neq i \in [n]]^\top - (g\mathbf{W}_{-i,*} \odot \mathbf{x}^{(j)})\mathbf{1}$. As a result, $\boldsymbol{\delta} \in \mathbb{R}^{n-1}$ is a vector of differences between the maximum and the selected element according to $\mathbf{x}^{(j)}$ for each row of $\mathbf{W}_{-i,*}$. Denote by (b_1, \dots, b_{n-1}) a permutation of $[n-1]$ satisfying $\delta_{b_1} \geq \delta_{b_2} \geq \dots \geq \delta_{b_{n-1}}$. It is thus not hard to show that, for any integer $0 \leq r \leq n-1$, and $\mathbf{x} \in \mathbf{B}^{((n-1) \times k)}$ that satisfies $\|\mathbf{x}^{(j)} - \mathbf{x}\|_1 = 2r$, we have

$$\sup_{\substack{\mathbf{x} \in \mathbf{B}^{((n-1) \times k)}, \\ \|\mathbf{x}^{(j)} - \mathbf{x}\|_1 = 2r}} h(\mathbf{x}) = \ln(1 + e^{\langle g\mathbf{W}, \mathbf{X}^{(j)} \rangle + \sum_{u=1}^r \delta_{b_u}}) - 2r\gamma - \frac{1}{2} \gamma \kappa(1 + gy^{(j)}),$$

where $\sum_{u=1}^r \delta_{b_u}$ is simply the sum of the first r largest elements of $\boldsymbol{\delta}$. Note that if $\delta_{b_r} \leq 0$ for some $r \in [n-1]$, we always have

$$\ln(1 + e^{\langle g\mathbf{W}, \mathbf{X}^{(j)} \rangle + \sum_{u=1}^r \delta_{b_u}}) - 2r\gamma \geq \ln(1 + e^{\langle g\mathbf{W}, \mathbf{X}^{(j)} \rangle + \sum_{u=1}^{r'} \delta_{b_u}}) - 2r'\gamma, \forall r \leq r'.$$

So only the positive elements in δ is of interest to finding the supremum. As a consequence, the objective of the dual problem can be rewritten as the point-wise maximum over $2n$ convex functions as follows:

$$\inf_{\substack{\mathbf{W} \in \mathbb{R}^{n \times k}, \\ \gamma \geq 0}} \gamma \varepsilon + \frac{1}{m'} \sum_{j=1}^{m'} \sup_{\substack{0 \leq r \leq n-1, \\ g \in \{-1, 1\}}} -2r\gamma - \frac{1}{2}\gamma\kappa(1 + gy^{(j)}) + \ln(1 + e^{\langle g\mathbf{W}, \mathbf{X}^{(j)} \rangle + \sum_{u=1}^r \delta_{bu}}).$$

To characterize the sequence of the sorted indices more formally, we have defined $\mathbb{T}(\mathbf{x})$ as a vector of sorted components of \mathbf{x} . The sorting operation required to evaluate $\mathbb{T}(\cdot)$ can be accomplished in $\Theta(n \log n)$ for sub-derivative evaluation. In such matter, we can reformulate the above convex program as

$$\inf_{\substack{\mathbf{W} \in \mathbb{R}^{n \times k}, \\ \gamma \geq 0}} \gamma \varepsilon + \frac{1}{m'} \sum_{j=1}^{m'} \sup_{\substack{0 \leq r \leq n-1, \\ g \in \{-1, 1\}}} -2r\gamma - \frac{1}{2}\gamma\kappa(1 + gy^{(j)}) + \ln(1 + e^{\langle g\mathbf{W}, \mathbf{X}^{(j)} \rangle + \langle \mathbb{T}(\boldsymbol{\delta})_{1 \dots r}, \mathbf{1} \rangle}).$$

□

One of the benefits brought by the Wasserstein DRO formulation is that it subsumes the $\ell_{2,1}$ -constrained (Wu et al., 2019) as well as regularized logistic regression approaches (Ravikumar et al., 2010) as special cases, as shown by the following theorem, which implies that minimizing the classic objectives is not enough to ensure distributional robustness:

Theorem 10. *If $\kappa = \infty$, $\|\mathbf{W}^\top\|_{2,1} \leq 2\lambda\sqrt{k}$ and $\gamma \geq (n+2)\lambda\sqrt{k}$, the convex program in Equation 2.6 subsumes the standard $\ell_{2,1}$ -constrained logistic regression approach in Equation 2.1*

as a special case. If $\kappa = \infty$ and $\gamma \geq \frac{n+2}{2} \|\mathbf{W}^\top\|_{2,1}$, it subsumes the $\ell_{2,1}$ -regularized logistic regression approach as a special case.

Proof. To begin with, we rewrite Equation 2.6 based on the cases where $g = y^{(j)}$ and $g = -y^{(j)}$:

$$\inf_{\substack{\mathbf{W} \in \mathbb{R}^{n \times k}, \\ \gamma \geq 0}} \gamma \varepsilon + \frac{1}{m'} \sum_{j=1}^{m'} \sup_{0 \leq r \leq n-1} \{-2r\gamma + \ln(1 + e^{\langle -y^{(j)} \mathbf{W}, \mathbf{X}^{(j)} \rangle + \langle \mathbb{T}(\boldsymbol{\delta})_{1 \dots r}, \mathbf{1} \rangle}), \\ -2r\gamma - \gamma \kappa + \ln(1 + e^{\langle y^{(j)} \mathbf{W}, \mathbf{X}^{(j)} \rangle + \langle \mathbb{T}(\boldsymbol{\delta})_{1 \dots r}, \mathbf{1} \rangle})\}.$$

Assume that $\gamma > 0$. Since $\kappa = \infty$, the second expression in the supremum makes the entire objective goes to $-\infty$, thus dominated by the first expression. Hence it can be simplified as

$$\inf_{\substack{\mathbf{W} \in \mathbb{R}^{n \times k}, \\ \gamma > 0}} \gamma \varepsilon + \frac{1}{m'} \sum_{j=1}^{m'} \sup_{0 \leq r \leq n-1} -2r\gamma + \ln(1 + e^{\langle -y^{(j)} \mathbf{W}, \mathbf{X}^{(j)} \rangle + \langle \mathbb{T}(\boldsymbol{\delta})_{1 \dots r}, \mathbf{1} \rangle}). \quad (2.8)$$

If $\gamma \geq (n+2)\lambda\sqrt{k} > 0$, $\|\mathbf{W}^\top\|_{2,1} \leq 2\lambda\sqrt{k}$ and $n, k \in \mathbb{Z}_+$, then for any $\mathbf{X} \in V^{(n \times k)}$, we have

$$\begin{aligned}
& \|\mathbf{W}\|_\infty \triangleq \|\mathbf{W}\|_{\infty,\infty} \leq \|\mathbf{W}^\top\|_{2,\infty} \leq \|\mathbf{W}^\top\|_{2,1} \leq 2\lambda\sqrt{k} \leq 2\gamma/(n+2) \\
& \implies e^{\|\mathbf{W}\|_\infty(n+2)} \leq e^{2\gamma} \\
& \implies e^{2\|\mathbf{W}\|_\infty(n+2)} \leq e^{2\gamma + \|\mathbf{W}\|_\infty(n+2)} \\
& \implies e^{2\|\mathbf{W}\|_\infty(n+2)} - (e^{2\gamma} - 1)e^{\|\mathbf{W}\|_\infty(n+2)} - e^{2\gamma} \leq 0 \\
& \implies e^{\|\mathbf{W}\|_\infty(n+2)} - e^{2\gamma - \|\mathbf{W}\|_\infty(n+2)} \leq e^{2\gamma} - 1 \\
& \implies e^{\|\mathbf{W}\|_\infty(n+2)} - e^{2\gamma - \|\mathbf{W}\|_\infty n} \leq e^{2\gamma} - 1 \\
& \implies e^{\|\mathbf{W}\|_\infty(n+2)} \leq e^{2\gamma} + e^{2\gamma - \|\mathbf{W}\|_\infty n} - 1 \\
& \implies \|\mathbf{W}\|_\infty \leq \frac{1}{2} [\ln(e^{2\gamma} + e^{2\gamma - \|\mathbf{W}\|_\infty n} - 1) - \|\mathbf{W}\|_\infty n] \\
& \leq \frac{1}{2} [\ln(e^{2\gamma} + e^{2\gamma + \langle \mathbf{W}, \mathbf{X} \rangle} - 1) - \langle \mathbf{W}, \mathbf{X} \rangle] \\
& \implies e^{\langle \mathbf{W}, \mathbf{X} \rangle + 2\|\mathbf{W}\|_\infty} \leq e^{2\gamma} + e^{\langle \mathbf{W}, \mathbf{X} \rangle + 2\gamma} - 1 \\
& \implies \frac{1 + e^{\langle \mathbf{W}, \mathbf{X} \rangle + 2\|\mathbf{W}\|_\infty}}{1 + e^{\langle \mathbf{W}, \mathbf{X} \rangle}} \leq e^{2\gamma} \\
& \implies \ln(1 + e^{\langle \mathbf{W}, \mathbf{X} \rangle + 2\|\mathbf{W}\|_\infty}) - \ln(1 + e^{\langle \mathbf{W}, \mathbf{X} \rangle}) \leq 2\gamma.
\end{aligned}$$

Therefore, the supremum in Equation 2.8 is achieved only when $r = 0$. Finally, Equation 2.8 can be rewritten as the following convex program:

$$\begin{cases} \inf_{\mathbf{W}} & \frac{1}{m'} \sum_{j=1}^{m'} \ln(1 + e^{-y^{(j)} \langle \mathbf{W}, \mathbf{X}^{(j)} \rangle}) \\ \text{s.t.} & \|\mathbf{W}^\top\|_{2,1} \leq 2\lambda\sqrt{k}, \end{cases}$$

which coincides with the $\ell_{2,1}$ -constrained logistic regression problem in Equation 2.1.

On the other hand, if $\gamma \geq \frac{n+2}{2} \|\mathbf{W}^\top\|_{2,1}$, by following the above same process, the supremum in Equation 2.8 is achieved only when $r = 0$. Note that only the first term in Equation 2.8 is related to γ . After minimizing over γ , we can rewrite Equation 2.8 as

$$\inf_{\mathbf{W} \in \mathbb{R}^{n \times k}} \frac{(n+2)\varepsilon}{2} \|\mathbf{W}^\top\|_{2,1} + \frac{1}{m'} \sum_{j=1}^{m'} \ln(1 + e^{-y^{(j)} \langle \mathbf{W}, \mathbf{x}^{(j)} \rangle}),$$

which is a standard $\ell_{2,1}$ -regularized logistic regression problem with $\tilde{\lambda} = \frac{(n+2)\varepsilon}{2}$. \square

Intuitively, when $\kappa = \infty$, flipping $y^{(j)}$ causes infinite transport cost. In this case, it is assumed that the realization of each $y^{(j)}$ given $\mathbf{x}^{(j)}$ is deterministic. Instead of taking into account the ambiguity only in the covariate measure $\mathbb{Q}(\mathrm{d}\mathbf{x})$, the Wasserstein DRO structure learning formulation grants flexibility to the joint measure $\mathbb{Q}(\mathrm{d}\boldsymbol{\xi})$. Modeling joint measure uncertainty is non-trivial here because all the random variables are involved in the node-alphabet-wise distributionally robust logistic regression problem in Equation 2.5.

If KL divergence is adopted to construct the ambiguity set, a tractable convex program can be derived as a corollary from Theorem 4 in (Hu and Hong, 2013):

Corollary 11. *Let $\mathcal{A}_\varepsilon^{\text{KL}}(\hat{\mathbb{P}}_{m'}) := \{\mathbb{Q} \in \mathcal{P}(\Xi) : D_{\text{KL}}(\mathbb{Q}, \hat{\mathbb{P}}_{m'}) \leq \varepsilon = \frac{\varepsilon_0}{m'}\}$ be a KL divergence ball.*

The primal problem in Equation 2.5 with $B_\varepsilon(\hat{\mathbb{P}}_{m'}) = B_\varepsilon^{\text{KL}}(\hat{\mathbb{P}}_{m'})$ is equivalent to

$$\inf_{\substack{\mathbf{W} \in \mathbb{R}^{n \times k}, \\ \gamma \geq 0}} \gamma \ln \left[\frac{1}{m'} \sum_{j=1}^{m'} (1 + e^{-y^{(j)} \langle \mathbf{W}, \mathbf{x}^{(j)} \rangle})^{\frac{1}{\gamma}} \right] + \gamma \varepsilon. \quad (2.9)$$

Proof. The problem we study satisfies Assumption 1 in (Hu and Hong, 2013) because $\ell(\boldsymbol{\xi})$ has finite support on Ξ . Substituting $P_0 = \hat{\mathbb{P}}_{m'}$ and $H(x, \xi) = \ell(y\langle \mathbf{W}, \mathbf{X} \rangle)$ into Theorem 4 in (Hu and Hong, 2013) leads to our result. \square

In contrast to the convex program with inner maximization in Equation 2.6, the direct minimization problem in Equation 2.9 based on KL divergence balls can be solved more efficiently. This class of problems have been shown to recover adversarial reweighting (Li and Dunson, 2020) and variance regularization (Duchi and Namkoong, 2019).

Algorithm 1 Structure Learning of Discrete Pairwise Graphical Models

Input: Alphabet size k ; number of variables n ; sample data $\{\bar{\mathbf{z}}^{(1)}, \dots, \bar{\mathbf{z}}^{(m)}\}$; model width λ ; minimum edge weight η
Output: Recovered edge set $\hat{\mathcal{E}}$
for all $(i, \alpha, \beta) \in [n] \times [k] \times [k]$ **do**
 Form a subset $\{\mathbf{z}^{(1)}, \dots, \mathbf{z}^{(m')}\}$ with $z_i^j \in \{\alpha, \beta\} \forall j \in [m']$
 Compute $\hat{\mathbf{W}}^{(i\alpha\beta)}$ by Equation 2.1 or Equation 2.6
 Centering $\hat{\mathbf{W}}^{(i\alpha\beta)}$ by Equation 2.2
 Estimate the weight matrices $\mathbf{W}^{(ij)}$ by Equation 2.3
 Estimate the edge set $\hat{\mathcal{E}}$ by Equation 2.4
end for

We illustrate the algorithmic details in Algorithm 1.

2.4 Theoretical Guarantees

In this section, we study statistical properties of the proposed estimators. More specifically, we derive generalization bounds, excess true risk bounds and sample complexities of our methods.

It is non-trivial to quantify the number of samples needed to recover the dependency graph with high probability in a structure learning problem. An initial attempt we made is to leverage a 0-concentration bound under 1-Wasserstein distance in the form of $\mathbb{P}^m[W_1(\mathbb{P}, \hat{\mathbb{P}}_m) \geq \varepsilon] \leq f(d, n, k, m, \varepsilon)$ to get a uniform upper confidence bound on the generalization error. It turns out that even the most advanced mean-concentration bounds $O(m^{-\frac{1}{n}})$ (Lei and others, 2020; Weed et al., 2019) with essentially optimal dependence on data dimensionality n lead to a sample complexity $O(C^{\frac{nk}{2}})$ with exponential dependence on n . The cause of the issue might be that convergence of $\hat{\mathbb{P}}_m$ to \mathbb{P} is much slower than convergence of $W_1(\hat{\mathbb{P}}_m, \mathbb{P})$ to its mean $\mathbb{E}_{\mathbb{P}^m} W_1(\hat{\mathbb{P}}_m, \mathbb{P})$ in high dimensional settings ($p = 1$ with large n). Hence the generalization bounds obtained via measure concentration are too conservative to be useful in our case.

Instead, we consider the following lemma about a uniform generalization bound based on bounded Lipschitz loss functions (Shalev-Shwartz and Ben-David, 2014) and Rademacher complexities (Bartlett and Mendelson, 2002).

Lemma 12 (Lemma 11 in (Wu et al., 2019)). *Let \mathcal{D} be a distribution on $\mathcal{X} \times \mathcal{Y}$, where $\mathcal{X} := \{\mathbf{x} \in \mathbb{R}^{n \times k} : \|\mathbf{x}^\top\|_{2,\infty} \leq X_{2,\infty}\}$ and $\mathcal{Y} := \{-1, 1\}$. Let $\ell : \mathbb{R} \rightarrow \mathbb{R}$ be a loss function with Lipschitz constant L_ℓ . Define the expected loss as $\mathcal{L}(\mathbf{w}) := \mathbb{E}_{\mathcal{D}} \ell(y \langle \mathbf{w}, \mathbf{x} \rangle)$ and the empirical loss as $\hat{\mathcal{L}}(\mathbf{w}) := \frac{1}{m} \sum_{i=1}^m \ell(y^{(i)} \langle \mathbf{w}, \mathbf{x}^{(i)} \rangle)$, where $\{\mathbf{x}^{(i)}, y^{(i)}\}_{i=1}^m \stackrel{iid}{\sim} \mathcal{D}$. Define $\mathcal{W} := \{\mathbf{w} \in \mathbb{R}^{n \times k} : \|\mathbf{w}^\top\|_{2,1} \leq W_{2,1}\}$. Then with probability at least $1 - \rho$ over the draw of m samples, we have that for all $\mathbf{w} \in \mathcal{W}$, $0 < \rho \leq 1$,*

$$\mathcal{L}(\mathbf{w}) - \hat{\mathcal{L}}(\mathbf{w}) \leq C \sqrt{\frac{24 \ln(n)}{m}} + C \sqrt{\frac{2 \ln(2/\rho)}{m}},$$

where $C = L_\ell X_{2,\infty} W_{2,1}$.

Proof. Please refer to Lemma 11 in (Wu et al., 2019) for the proof. \square

In order to get a sample complexity bound, we derive an excess true risk bound for transport-based DRO estimators, in terms of generalization errors, which may be of independent interest.

Proposition 13. *Assume that (Ξ, c) is a Banach space, $\mathcal{P}_p(\Xi)$ is the space of Borel probability measures on Ξ with finite p -th moment for $p \geq 1$, $\hat{\mathbb{P}}_m \in \mathcal{P}_p(\Xi)$ is the empirical measure for some $\mathbb{P} \in \mathcal{P}_p(\Xi)$, $\mathcal{A} = \mathcal{A}_\varepsilon^{W_p}(\hat{\mathbb{P}}_m)$ is a type- p Wasserstein ball centered at $\hat{\mathbb{P}}_m$ with radius ε , \mathcal{F} is a space of closed convex functions $f : \Xi \rightarrow \mathbb{R}_+$ with $\text{lip}_c(f) < \infty$. Let \hat{f} be a minimizer of the DRO problem in Equation 2.5 and f^* be a minimizer of the stochastic optimization problem in Equation 1.1, we have*

$$\int_{\Xi} \hat{f}(\xi) \mathbb{P}(d\xi) - \int_{\Xi} f^*(\xi) \mathbb{P}(d\xi) \leq \varepsilon \text{lip}_c(f^*) + 2 \sup_{f \in \mathcal{F}} \left| \int_{\Xi} f(\xi) \mathbb{P}(d\xi) - \int_{\Xi} f(\xi) \hat{\mathbb{P}}_m(d\xi) \right|.$$

Proof. To avoid clutter of notations, we define $\mathcal{A}(\mathbb{P}) := \mathcal{A}_\varepsilon^{W_p}(\mathbb{P})$.

According to Theorem 1 in (Cranko et al., 2021), the following relation holds for any $f \in \mathcal{F}$ and a fixed $\mathbb{P} \in \mathcal{P}_p(\Xi)$:

$$\int_{\Xi} f(\xi) \mathbb{P}(d\xi) \leq \sup_{\mathbb{Q} \in \mathcal{A}(\mathbb{P})} \int_{\Xi} f(\xi) \mathbb{Q}(d\xi) \leq \int_{\Xi} f(\xi) \mathbb{P}(c\xi) + \varepsilon \text{lip}_c(f).$$

Note that we are given a worst-case risk minimizer \hat{f} defined as

$$\hat{f} \in \arg \inf_{f \in \mathcal{F}} \sup_{\mathbb{Q} \in \mathcal{A}(\hat{\mathbb{P}}_m)} \int_{\Xi} f(\xi) \mathbb{Q}(\mathrm{d}\xi),$$

and a true risk minimizer f^* defined as

$$f^* \in \arg \inf_{f \in \mathcal{F}} \int_{\Xi} f(\xi) \mathbb{P}(\mathrm{d}\xi).$$

As a result of uniform boundedness, we have

$$\begin{aligned} & \left| \int_{\Xi} \hat{f}(\xi) \mathbb{P}(\mathrm{d}\xi) - \int_{\Xi} f^*(\xi) \mathbb{P}(\mathrm{d}\xi) \right| \\ &= \int_{\Xi} \hat{f}(\xi) \mathbb{P}(\mathrm{d}\xi) - \int_{\Xi} f^*(\xi) \mathbb{P}(\mathrm{d}\xi) \\ &= \int_{\Xi} \hat{f}(\xi) \mathbb{P}(\mathrm{d}\xi) - \sup_{\mathbb{Q} \in \mathcal{A}(\hat{\mathbb{P}}_m)} \int_{\Xi} \hat{f}(\xi) \mathbb{Q}(\mathrm{d}\xi) + \sup_{\mathbb{Q} \in \mathcal{A}(\hat{\mathbb{P}}_m)} \int_{\Xi} \hat{f}(\xi) \mathbb{Q}(\mathrm{d}\xi) \\ &\quad - \sup_{\mathbb{Q} \in \mathcal{A}(\hat{\mathbb{P}}_m)} \int_{\Xi} f^*(\xi) \mathbb{Q}(\mathrm{d}\xi) + \sup_{\mathbb{Q} \in \mathcal{A}(\hat{\mathbb{P}}_m)} \int_{\Xi} f^*(\xi) \mathbb{Q}(\mathrm{d}\xi) - \int_{\Xi} f^*(\xi) \mathbb{P}(\mathrm{d}\xi) \\ &\leq \int_{\Xi} \hat{f}(\xi) \mathbb{P}(\mathrm{d}\xi) - \sup_{\mathbb{Q} \in \mathcal{A}(\hat{\mathbb{P}}_m)} \int_{\Xi} \hat{f}(\xi) \mathbb{Q}(\mathrm{d}\xi) + \sup_{\mathbb{Q} \in \mathcal{A}(\hat{\mathbb{P}}_m)} \int_{\Xi} f^*(\xi) \mathbb{Q}(\mathrm{d}\xi) - \int_{\Xi} f^*(\xi) \mathbb{P}(\mathrm{d}\xi) \\ &\leq \int_{\Xi} \hat{f}(\xi) \mathbb{P}(\mathrm{d}\xi) - \int_{\Xi} \hat{f}(\xi) \hat{\mathbb{P}}_m(\mathrm{d}\xi) + \int_{\Xi} f^*(\xi) \hat{\mathbb{P}}_m(\mathrm{d}\xi) + \varepsilon \mathrm{lip}_c(f^*) - \int_{\Xi} f^*(\xi) \mathbb{P}(\mathrm{d}\xi) \\ &\leq \varepsilon \mathrm{lip}_c(f^*) + 2 \sup_{f \in \mathcal{F}} \left| \int_{\Xi} f(\xi) \mathbb{P}(\mathrm{d}\xi) - \int_{\Xi} f(\xi) \hat{\mathbb{P}}_m(\mathrm{d}\xi) \right|. \end{aligned}$$

□

Hereupon, following the proofs of Lemma 2 and Theorem 2 in (Wu et al., 2019), we derive a sample complexity bound for our Wasserstein DRO structure learning method by upper bounding $\|\mathbf{W}_{\alpha*}^{(ij)} - \mathbf{W}_{\beta*}^{(ij)} - \mathbf{U}_{j*}^{(i\alpha\beta)}\|_1$ based on the excess risk bound in Proposition 13.

Theorem 14. *Given that: $\mathcal{D}(\mathcal{W}, \Theta)$ is an unknown pairwise Markov network with n variables, alphabet size k , dependency graph \mathcal{G} ; that Assumption 5 and Assumption 6 hold; that $\|\mathbf{W}^\top\|_{2,1} \leq 2\lambda\sqrt{k}$ in Equation 2.5; that $\mathbf{W}^{(ij)} \in \mathcal{W}$ is the true weight matrix; and that $\hat{\mathbf{W}}^{(ij)}$ is the estimated weight matrix from Equation 2.6 with the Wasserstein ambiguity set and properly centered, then, for any $\rho \in (0, 1]$, $\omega > 0$, $n \in \mathbb{Z}_+$ and $i \neq j \in [n]$, if the number of i.i.d. samples satisfies $m = O(\frac{\lambda^2 k^4 e^{14\lambda} (\varepsilon_0^2 + \ln \frac{nk}{\rho})}{\omega^4})$, with probability at least $1 - \rho$, the following bound holds:*

$$\|\mathbf{W}^{(ij)} - \hat{\mathbf{W}}^{(ij)}\|_{\infty, \infty} \leq \omega.$$

Let $\omega < \frac{\eta}{2}$ and $\hat{\mathcal{G}}$ be reconstructed via thresholding in Equation 2.4. Now if

$$m = O(\frac{\lambda^2 k^4 e^{14\lambda} (\varepsilon_0^2 + \ln \frac{nk}{\rho})}{\eta^4}),$$

with probability $1 - \rho$, we have $\mathcal{G} = \hat{\mathcal{G}}$.

Proof. We use \mathbb{P} to denote the true distribution and $\hat{\mathbb{P}}_{m'}$ to represent the empirical distribution.

Define $\ell_{\mathbf{W}}(\boldsymbol{\xi}) := \ell(y \langle \mathbf{W}, [\mathbf{x}_{1 \dots i-1, *}^\top, \mathbf{b}^{(1)\top}, \mathbf{x}_{i \dots n-1, *}^\top]^\top \rangle)$.

We follow the proof of Theorem 2 in (Wu et al., 2019) by starting with upper bounding the excess true risk.

By Assumption 6, we have $\|\bar{\mathbf{W}}^\top\|_{2,1} \leq 2\lambda\sqrt{k}$ for all $i \in [n]$, $\alpha \neq \beta \in [k]$, where $\bar{\mathbf{W}}$ is defined in Fact 7 based on the true weight matrices \mathcal{W} . By the assumptions stated in this theorem, $\hat{\mathbf{W}}^{(i\alpha\beta)}$ in Equation 2.5 should also satisfy $\|\hat{\mathbf{W}}^{(i\alpha\beta)\top}\|_{2,1} \leq 2\lambda\sqrt{k}$. The one-hot matrices $\bar{\mathbf{Z}}$ in Fact 7 and \mathbf{X} in Equation 2.5 satisfy $\|\bar{\mathbf{Z}}^\top\|_{2,\infty} \leq 1$, $\|\mathbf{X}^\top\|_{2,\infty} \leq 1$ by definition. The logistic loss function $\ell(\cdot)$ has a Lipschitz constant of 1.

According to Lemma 12, for all $\mathbf{W} \in \mathbb{R}^{n \times k}$ that satisfy $\|\mathbf{W}^\top\|_{2,1} \leq 2\lambda\sqrt{k}$,

$$\mathbb{P}^{m'} \left\{ \mathbb{E}_{\mathbb{P}}[\ell_{\mathbf{W}}(\boldsymbol{\xi})] - \mathbb{E}_{\hat{\mathbb{P}}_{m'}}[\ell_{\mathbf{W}}(\boldsymbol{\xi})] \leq 2\lambda\sqrt{k} \left(2\sqrt{\frac{6 \ln(n)}{m'}} + \sqrt{\frac{2 \ln(2/\rho)}{m'}} \right) \right\} \geq 1 - \rho. \quad (2.10)$$

Define $\mathbf{W}^{(i\alpha\beta)} \in \mathbb{R}^{n \times k}$ as $\mathbf{W}_{i*}^{(i\alpha\beta)} := [\theta_\alpha^{(i)} - \theta_\beta^{(i)}, \mathbf{0}^\top]$, and $\mathbf{W}_{j*}^{(i\alpha\beta)} := \mathbf{W}_{\alpha*}^{(ij)} - \mathbf{W}_{\beta*}^{(ij)}$ for $j \neq i \in [n]$. Recall that $\hat{\mathbf{W}}^{(i\alpha\beta)}$ is a minimizer of Equation 2.5 with a Wasserstein ball:

$$\hat{\mathbf{W}}^{(i\alpha\beta)} \in \arg \inf_{\mathbf{W} \in \mathbb{R}^{n \times k}} \sup_{\mathbb{Q} \in \mathcal{A}_\varepsilon^{W_1}(\hat{\mathbb{P}}_{m'})} \mathbb{E}_{\mathbb{Q}}[\ell_{\mathbf{W}}(\boldsymbol{\xi})].$$

By Proposition 13,

$$\mathbb{E}_{\mathbb{P}}[\ell_{\hat{\mathbf{W}}^{(i\alpha\beta)}}(\boldsymbol{\xi})] - \mathbb{E}_{\mathbb{P}}[\ell_{\mathbf{W}^{(i\alpha\beta)}}(\boldsymbol{\xi})] \leq 2\lambda\sqrt{k}\varepsilon + 2 \sup_{\mathbf{W}: \|\mathbf{W}^\top\|_{2,1} \leq 2\lambda\sqrt{k}} |\mathbb{E}_{\mathbb{P}}[\ell_{\mathbf{W}}(\boldsymbol{\xi})] - \mathbb{E}_{\hat{\mathbb{P}}_{m'}}[\ell_{\mathbf{W}}(\boldsymbol{\xi})]|,$$

which can be combined with Equation 2.10 and the definition $\varepsilon = \varepsilon_0/\sqrt{m'}$, yielding

$$\begin{aligned} & \mathbb{P}^{m'} \left\{ \mathbb{E}_{\mathbb{P}}[\ell_{\hat{\mathbf{W}}^{(i\alpha\beta)}}(\boldsymbol{\xi})] - \mathbb{E}_{\mathbb{P}}[\ell_{\mathbf{W}^{(i\alpha\beta)}}(\boldsymbol{\xi})] \leq 2\lambda\sqrt{k} \left(\frac{\varepsilon_0}{\sqrt{m'}} + 4\sqrt{\frac{6 \ln(n)}{m'}} + 2\sqrt{\frac{2 \ln(2/\rho)}{m'}} \right) \right\} \\ & \geq 1 - \rho. \end{aligned}$$

Therefore, there exists a global constant $C > 0$ such that if $m' = \frac{C\lambda^2 k(\varepsilon_0^2 + \ln \frac{2n}{\rho})}{4\omega^2}$, with probability at least $1 - \rho$,

$$\mathbb{E}_{\mathbb{P}}[\ell_{\hat{\mathbf{W}}^{(i\alpha\beta)}}(\boldsymbol{\xi})] - \mathbb{E}_{\mathbb{P}}[\ell_{\mathbf{W}^{(i\alpha\beta)}}(\boldsymbol{\xi})] \leq 2\omega.$$

Using Lemma 9 and Lemma 10 in (Wu et al., 2019), if the number of samples satisfies $m' = O(\frac{\lambda^2 k(\varepsilon_0^2 + \ln \frac{n}{\rho})}{\omega^2})$, with probability at least $1 - \rho$,

$$\begin{aligned} & \mathbb{E}_{\mathbb{P}}[\sigma(\langle \mathbf{W}^{(i\alpha\beta)}, \mathbf{x} \rangle) - \sigma(\langle \hat{\mathbf{W}}^{(i\alpha\beta)}, \mathbf{x} \rangle)]^2 \\ & \leq \mathbb{E}_{\mathbb{P}} D_{\text{KL}}(\sigma(\langle \mathbf{W}^{(i\alpha\beta)}, \mathbf{x} \rangle) \parallel \sigma(\langle \hat{\mathbf{W}}^{(i\alpha\beta)}, \mathbf{x} \rangle))/2 \\ & \leq \frac{1}{2} (\mathbb{E}_{\mathbb{P}}[\ell_{\hat{\mathbf{W}}^{(i\alpha\beta)}}(\boldsymbol{\xi})] - \mathbb{E}_{\mathbb{P}}[\ell_{\mathbf{W}^{(i\alpha\beta)}}(\boldsymbol{\xi})]) \\ & \leq \omega. \end{aligned}$$

Now fix some $i \in [n]$, $\alpha \neq \beta \in [k]$. Denote by $m^{(i\alpha\beta)}$ the number of samples in which $\tilde{z}_i^j \in \{\alpha, \beta\}$. Recall that $\mathbf{U}^{(i\alpha\beta)}$, the centered version of $\hat{\mathbf{W}}^{(i\alpha\beta)}$, satisfies $\langle \hat{\mathbf{W}}^{(i\alpha\beta)}, \mathbf{x} \rangle = \langle \mathbf{U}^{(i\alpha\beta)}, \mathbf{x} \rangle$. As a result, if $m^{(i\alpha\beta)} = O(\frac{\lambda^2 k(\varepsilon_0^2 + \ln \frac{n}{\rho})}{\omega^2})$, with probability at least $1 - \rho$,

$$\mathbb{E}_{\mathbb{P}}[\sigma(\langle \mathbf{W}^{(i\alpha\beta)}, \mathbf{x} \rangle) - \sigma(\langle \mathbf{U}^{(i\alpha\beta)}, \mathbf{x} \rangle)]^2 \leq \omega.$$

By Definition 3 in (Wu et al., 2019), a distribution \mathcal{D} is δ -unbiased if its conditional probability of a variable given the others is bounded away from 0 by at least δ .

By Lemma 4 and Lemma 7 in (Wu et al., 2019), we know that $\mathbf{Z} \sim \mathcal{D}$ is δ -unbiased with $\delta = e^{-2\lambda(\mathcal{D})/k}$, and so is \mathbf{Z}_{-i} conditioned on $Z_i \in \{\alpha, \beta\}$. Applying Lemma 6 in (Wu et al., 2019), if $m^{(i\alpha\beta)} = O(\frac{\lambda^2 k^3 e^{12\lambda}(\varepsilon_0^2 + \ln \frac{n}{\rho'})}{\omega^4})$ the following inequality holds with probability at least $1 - \rho'$:

$$\begin{aligned} & \|\mathbf{W}^{(i\alpha\beta)} - \mathbf{U}^{(i\alpha\beta)}\|_{\infty, \infty} \leq \omega \\ \implies & |W_{\alpha b}^{(ij)} - W_{\beta b}^{(ij)} - U_{jb}^{(i\alpha\beta)}| \leq \omega, \forall j \neq i \in [n], b \in [k]. \end{aligned}$$

Since $\mathbf{Z} \sim \mathcal{D}$ is δ -unbiased, we have $\mathbb{P}[Z_i \in \{\alpha, \beta\}] \geq 2\delta$. By the Chernoff bound, if the total number of samples satisfies $m = O(\frac{1}{\delta}(m^{(i\alpha\beta)} + \log(\frac{1}{\rho'})))$, with probability at least $1 - \rho''$, we have $m^{(i\alpha\beta)}$ samples for the fixed $i \in [n]$, $\alpha \neq \beta \in [k]$.

Now set $\rho' = \rho'' = \frac{\rho}{2nk^2}$ and take a union bound over all $\alpha \neq \beta \in [k]$, then with probability at least $1 - \frac{\rho}{n}$ and $m = O(\frac{\lambda^2 k^4 e^{14\lambda}(\varepsilon_0^2 + \ln \frac{nk}{\rho})}{\omega^4})$, we have

$$|W_{\alpha b}^{(ij)} - W_{\beta b}^{(ij)} - U_{jb}^{(i\alpha\beta)}| \leq \omega, \forall j \neq i \in [n], b \in [k], \alpha \neq \beta \in [k].$$

Because $\mathbf{W}^{(ij)}$ are centered, summing the above equalities for all $\beta \in [k]$ leads to

$$\begin{aligned} & |W_{\alpha b}^{(ij)} - \frac{1}{k} \sum_{\beta \in [k]} U_{jb}^{(i\alpha\beta)}| \leq \omega, \forall j \neq i \in [n], b, \alpha \in [k] \\ \implies & |W_{\alpha b}^{(ij)} - \hat{W}_{\alpha b}^{(ij)}| \leq \omega, \forall j \neq i \in [n], b, \alpha \in [k] \\ \implies & \|\mathbf{W}^{(ij)} - \hat{\mathbf{W}}^{(ij)}\|_{\infty, \infty} \leq \omega, \forall j \neq i \in [n], \end{aligned}$$

which holds with probability at least $1 - \frac{\rho}{n}$ and $m = O(\frac{\lambda^2 k^4 e^{14\lambda}(\varepsilon_0^2 + \ln \frac{nk}{\rho})}{\omega^4})$, for fixed $i \in [n]$.

We conclude by taking a union bound for all $i \in [n]$, so that with probability at least $1 - \rho$ and $m = O(\frac{\lambda^2 k^4 e^{14\lambda} (\varepsilon_0^2 + \ln \frac{nk}{\rho})}{\omega^4})$,

$$\|\mathbf{W}^{(ij)} - \hat{\mathbf{W}}^{(ij)}\|_{\infty, \infty} \leq \omega, \forall i, j \in [n], i \neq j.$$

□

The sample complexity is in terms of an $\ell_{\infty, \infty}$ error bound to ensure that every true edge is recovered. It shows that theoretically the number of samples needed to recover the true graph is polynomial in $\frac{1}{\omega}$, k , ε_0 , $\ln \frac{nk}{\rho}$, but exponential in model width λ . Similarly, we derive a sample complexity for the KL divergence-based DRO estimator via variance regularization (Lam, 2019) instead of Lipschitz regularization (Cranko et al., 2021).

Theorem 15. *Given assumptions in Theorem 14, except that $\hat{\mathbf{W}}^{(ij)}$ is the estimated weight matrix from Equation 2.9 with the KL ambiguity set. Let $\hat{\mathcal{G}}$ be constructed via thresholding in Equation 2.4. Then, for any $\rho \in (0, 1]$, $\eta > 0$, $\varepsilon < 1$, $n \in \mathbb{Z}_+$ and $i \neq j \in [n]$, if the number of i.i.d. samples satisfies $m = O(\frac{\lambda^2 k^4 e^{14\lambda} (\varepsilon_0^2 + \ln \frac{nk}{\rho})}{\eta^4})$, with probability at least $1 - \rho$, the following bound holds:*

$$\|\mathbf{W}^{(ij)} - \hat{\mathbf{W}}^{(ij)}\|_{\infty, \infty} < \frac{\eta}{2} \implies \mathcal{G} = \hat{\mathcal{G}}.$$

Proof. According to Theorem 7 in (Lam, 2019), for any \mathbf{W} ,

$$\begin{aligned}
& \mathbb{E}_{\hat{\mathbb{P}}_m}[\ell_{\mathbf{W}}(\boldsymbol{\xi})] \\
& \leq \sup_{\mathbb{Q} \in \mathcal{A}_\varepsilon^{\text{KL}}(\hat{\mathbb{P}}_m)} \mathbb{E}_{\mathbb{Q}}[\ell_{\mathbf{W}}(\boldsymbol{\xi})] \\
& \leq \mathbb{E}_{\hat{\mathbb{P}}_m}[\ell_{\mathbf{W}}(\boldsymbol{\xi})] + \sqrt{2\varepsilon \text{Var}_{\hat{\mathbb{P}}_m}(\ell_{\mathbf{W}}(\boldsymbol{\xi}))} + 2\varepsilon C \frac{1}{m'} \frac{\sum_i (\ell_{\mathbf{W}}(\boldsymbol{\xi}_i) - \overline{\ell_{\mathbf{W}}}(\boldsymbol{\xi}))^3}{\sum_i (\ell_{\mathbf{W}}(\boldsymbol{\xi}_i) - \overline{\ell_{\mathbf{W}}}(\boldsymbol{\xi}))^2} \\
& \leq \mathbb{E}_{\hat{\mathbb{P}}_m}[\ell_{\mathbf{W}}(\boldsymbol{\xi})] + \sqrt{2\varepsilon \text{Var}_{\hat{\mathbb{P}}_m}(\ell_{\mathbf{W}}(\boldsymbol{\xi}))} + 2\varepsilon C \frac{1}{m'} \sum_i |\ell_{\mathbf{W}}(\boldsymbol{\xi}_i) - \overline{\ell_{\mathbf{W}}}(\boldsymbol{\xi})|,
\end{aligned}$$

where $\overline{\ell_{\mathbf{W}}} = \frac{1}{m'} \sum_i \ell_{\mathbf{W}}(\boldsymbol{\xi}_i)$ and $C > 0$ is a constant independent of n .

Note that

$$\text{Var}_{\hat{\mathbb{P}}_m}(\ell_{\mathbf{W}}(\boldsymbol{\xi})) \leq \sup_{\mathbf{W}, \mathbf{W}', \boldsymbol{\xi}, \boldsymbol{\xi}'} |\ell_{\mathbf{W}}(\boldsymbol{\xi}) - \ell_{\mathbf{W}'}(\boldsymbol{\xi}')|^2 \leq (4\lambda\sqrt{k})^2,$$

yielding

$$\begin{aligned}
& \sup_{\mathbb{Q} \in \mathcal{A}_\varepsilon^{\text{KL}}(\hat{\mathbb{P}}_m)} \mathbb{E}_{\mathbb{Q}}[\ell_{\mathbf{W}}(\boldsymbol{\xi})] \\
& \leq \mathbb{E}_{\hat{\mathbb{P}}_m}[\ell_{\mathbf{W}}(\boldsymbol{\xi})] + 4\lambda\sqrt{k}(\sqrt{2\varepsilon} + 2\varepsilon C) \\
& \leq \mathbb{E}_{\hat{\mathbb{P}}_m}[\ell_{\mathbf{W}}(\boldsymbol{\xi})] + 4\lambda\sqrt{k}(2\sqrt{\varepsilon} + 2C\sqrt{\varepsilon})
\end{aligned}$$

for $\varepsilon < 1$.

Therefore,

$$\mathbb{P}^{m'} \left\{ \mathbb{E}_{\mathbb{P}}[\ell_{\hat{\mathbf{W}}^{(i\alpha\beta)}}(\boldsymbol{\xi})] - \mathbb{E}_{\mathbb{P}}[\ell_{\mathbf{W}^{(i\alpha\beta)}}(\boldsymbol{\xi})] \leq 2\lambda\sqrt{k}((4C+4)\sqrt{\frac{\varepsilon_0}{m'}} + 4\sqrt{\frac{6\ln(n)}{m'}} + 2\sqrt{\frac{2\ln(2/\rho)}{m'}}) \right\} \geq 1 - \rho.$$

Following the same procedure in the proof in Theorem 14, we get the conclusion that with probability at least $1 - \rho$ and $m = O(\frac{\lambda^2 k^4 e^{14\lambda}(\varepsilon_0 + \ln \frac{nk}{\rho})}{\omega^4})$,

$$\|\mathbf{W}^{(ij)} - \hat{\mathbf{W}}^{(ij)}\|_{\infty, \infty} \leq \omega, \forall i, j \in [n], i \neq j.$$

□

The two sample complexity bounds differ by a factor of ε_0 because the Wasserstein ball radius is chosen in the square root order $\frac{1}{\sqrt{m'}}$ while the KL ball radius decays in a non-asymptotic $\frac{1}{m'}$ -rate. In practice, $\varepsilon_0^2 \ll \ln \frac{nk}{\rho}$ for Wasserstein DRO whereas ε_0 for KL DRO is not too larger than $\ln \frac{nk}{\rho}$. Compared to the state-of-the-art result $O(\frac{\lambda^2 k^4 e^{14\lambda} \ln \frac{nk}{\rho}}{\eta^4})$ (Wu et al., 2019), our complexities have an additional term that scales as $O(\frac{\lambda^2 k^4 e^{14\lambda}}{\eta^4})$, weighted by ε_0 or ε_0^2 . The result in (Wu et al., 2019) is slightly better than that in (Vuffray et al., 2020) in the pairwise setting, even though the latter is applicable to higher-order models. If the radius is set to zero, we recover the non-robust near-optimal bound (Wu et al., 2019) but the learned graphical structure will be vulnerable to perturbation. On the contrary, a larger radius corresponds to more robustness at the risk of underfitting. On that account, with a similar number of samples,

the proposed estimators have the statistical property of distributional robustness at almost no cost. In the noisy-data setting, the benefit with a little extra sample complexity is obvious since non-robust methods may fail.

Remark 16. The derived sample complexities are with respect to clean data since we do not assume any specific contamination models. Our approach can be considered as regularization with better probabilistic and robust interpretation. Given recoverability and noisy data, a contamination model usually has to be assumed in order to obtain a sample complexity for this kind of noise.

The radius ε_0 should be judiciously chosen with expectation that the ambiguity set encompasses true distribution with high confidence while excluding pathological distributions (Gao and Kleywegt, 2022). There are two approaches to choosing the radius. One of them is to select the best value based on empirical cross-validation errors. The other one is to determine the radius defining an ambiguity set that encompasses the true distribution with a given confidence (e.g., $1 - \rho = 0.95$) based on concentration bounds of the corresponding measures. The latter approach is more theoretically sound but likely yielding a pessimistic radius.

2.5 Experiments

We conduct a simulated study of synthetic data perturbed by the following contamination models:

Noiseless Model. The common setting with no contamination to samples drawn from $\mathcal{D}(\mathcal{W}, \Theta)$.

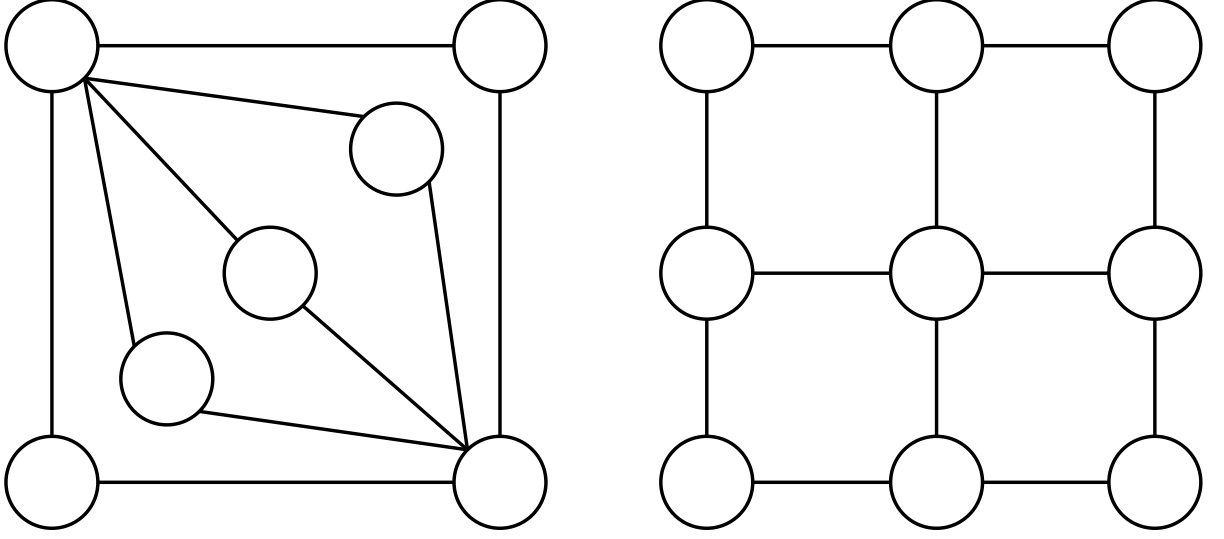


Figure 1: The adopted underlying graphs. Two nodes are connected to the others in the diamond graph. The grid graph has d^2 nodes. Each edge weight matrix is centered with random values $\pm\theta$.

Huber's Contamination Model. Let \mathcal{D}_e be an arbitrary probability measure on $[k]^n$. Each sample is drawn i.i.d. from $(1 - \zeta)\mathcal{D} + \zeta\mathcal{D}_e$. We adopt the uniform distribution $\mathcal{U}([k]^n)$ for \mathcal{D}_e .

Independent Failure Model. Each entry is independently randomly corrupted during sampling. We consider a special case in our experiments where each component $z_i \in [k]$ of $\mathbf{z} \sim \mathcal{D}$ is randomly replaced with a different value with probability ζ .

We adopt a diamond and a grid underlying graph, illustrated in Figure 1, where each edge has a centered weight matrix of random values $\pm\theta$. Since we compute the true distribution exactly, it is impossible to generate samples for large graph without approximate methods such as Gibbs

sampling. This due to the memory and precision limit of modern computers. Gibbs sampling and other Markov chain Monte Carlo (MCMC) algorithms require very long mixing time for good samples. Quantum computers yield good-quality real-world samples but are inaccessible for the authors at the time of writing. We form different setups by varying graph size $n \in \{6, 9, 12\}$, alphabet size $k \in \{2, 4, 6\}$, edge weight $\theta \in \{0.1, 0.2, 0.3\}$, noise rate $\zeta \in \{0, 0.1, 0.2, 0.3, 0.5\}$ and contamination models. In each setup, we record the probability of success among 100 runs, in which success means the estimated graph is identical to the true graph. This corresponds to a zero-one loss evaluating complete matching. However, there are feasible soft evaluation metrics including the Hamming distance, measuring the fraction of correctly recovered edges, and a statistical distance between distributions. At the beginning of each run, we draw m i.i.d. samples from $\mathcal{D}(\mathcal{W}, \Theta)$ with exact distribution, where $m \in [1000, 10000]$. Afterwards, the samples are corrupted accordingly and provided as input to each algorithm.

We compare our methods against sparse logistic regression with parameters suggested by (Wu et al., 2019), where the number of mirror descent iterations is 50000. We tune our model hyperparameters $\varepsilon_0, \kappa \in [0.01, 100]$ using a logarithmic scale on random graphs of same size as the target graph. We adopt L-BFGS-B (Byrd et al., 1995) in SciPy (Virtanen et al., 2020) as the optimizer. Default values are adopted for unmentioned parameters. We conduct all experiments on a laptop with an Intel Core i7 2.7 GHz processor.

The results for comparing probabilities of success are shown in Figure 2. Generally speaking, the proposed two DRO approaches outperform $\ell_{2,1}$ -constrained logistic regression (SLR) across all the experimental settings by a large margin whereas the Wasserstein DRO approach (WDRSL)

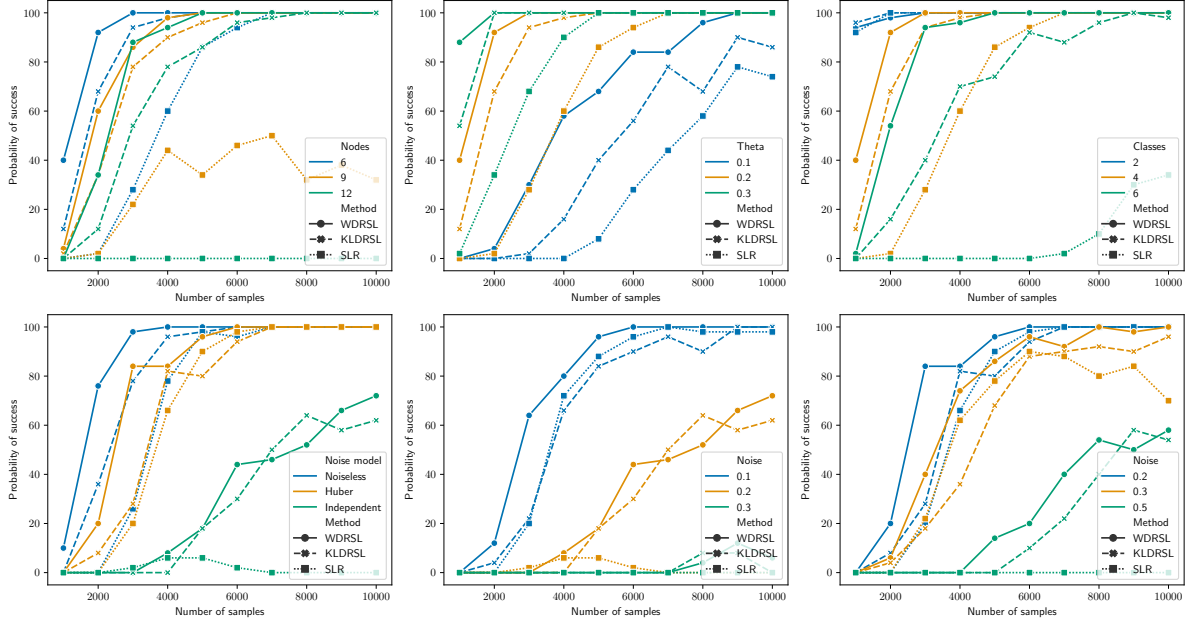


Figure 2: Plots of the probability of successfully estimating the structure versus the number of samples for Wasserstein DRO structure learning (WDRSL), KL DRO (KLDRSL) and sparse logistic regression (SLR). Top, from left to right: (a) diamond, 4 classes, noiseless, $\theta = 0.2$, varying nodes; (b) diamond, 6 nodes, 4 classes, noiseless, varying θ ; (c) diamond, 6 nodes, noiseless, $\theta = 0.2$, varying classes. Bottom, from left to right: (d) grid, 9 nodes, 4 classes, $\theta = 0.2$, varying noise models with $\zeta = 0.2$; (e) grid, 9 nodes, 4 classes, $\theta = 0.2$, independent failure model, varying probability of noise; (f) grid, 9 nodes, 4 classes, $\theta = 0.2$, Huber's contamination model, varying noise level.

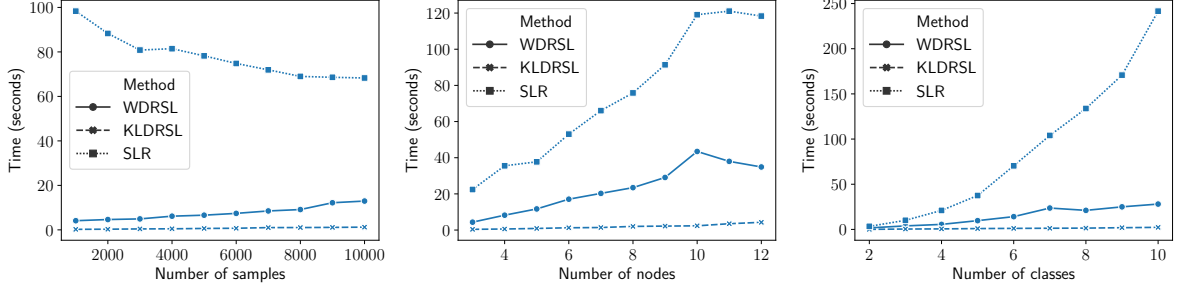


Figure 3: Comparisons of the execution time of one run. $\theta = 0.2$ and noiseless model are adopted in all settings. $\kappa = 1$, $\varepsilon_0 = 33$ for KLDRSL and $\varepsilon_0 = 1.5$ for WDRSL. From left to right: (a) grid, 9 nodes, 4 classes, varying samples; (b) diamond, 4 classes, varying nodes; (c) diamond, 3 nodes, varying classes.

further outperforms the KL DRO approach (KLDRSL) significantly. Our method has better scalability according to the upper part of Figure 2, where we vary the number of nodes, the model width and the number of classes on the diamond graph. For example, in the top right plot, for 6 classes, given about 3000 samples, WDRSL is already able to recover the graph with probability 90% while SLR cannot achieve that even with more samples. The advantage can also be observed in the upper center plot when $\theta = 0.3$ with only 1000 samples. The results on noiseless data are thus consistent with our analysis on the probabilistic interpretation of DRO as a more general alternative to standard regularization. The results in the bottom left plot of Figure 2 imply that, with a similar perturbation budget, the independent failure model is more powerful at corrupting data in the structure learning setting. As we vary the probability of contaminating each entry independently (bottom center plot), it becomes significantly more difficult to learn the underlying graph. For example, even our DRO methods that are inherently

robust can hardly succeed when $\zeta = 0.3$. That being said, we still expect there to be a large margin of performance comparison between our method and SLR as more samples are accessed. Under Huber’s contamination model with 50% data being noisy, we are still able to exactly reconstruct the structure with about a 50% chance. It is noteworthy that in some cases such as 10% independent failure, SLR outperforms KLDRSL probably because of the equivalence of KL DRO to adversarial reweighting and domination of pathological distributions. Despite not being comparable to WDRSL in terms of success rate, KLDRSL is the most efficient one according to Figure 3, whereas WDRSL provides a trade-off between computational efficiency and structure learning ability.

2.6 Concluding Remarks

In this chapter, we develop distributionally robust approaches based on two ambiguity sets for structure learning of pairwise MRFs with general alphabet from sample data. We provide tractable dual reformulations for the primal problems and showed their connections to regularization schemes. We derive near-optimal sample complexities and demonstrated consistent benefits over sparse logistic regression. We conduct empirical study which is lacking in the literature since most of the related work are purely theoretical.

CHAPTER 3

DISTRIBUTIONALLY ROBUST STRUCTURE LEARNING OF DIRECTED GRAPHICAL MODELS

In this chapter, we consider the problem of learning the structure of general discrete Bayesian networks from potentially corrupted data. Building on DRO and a linear regression approach, we propose a method that optimizes the most adverse risk over a family of distributions. The proposed approach applies for general categorical random variables without assuming faithfulness, an ordinal relationship or a specific form of conditional distribution. We provide necessary background in Section 3.1 and Section 3.2. Under mild assumptions, we present efficient algorithms and non-asymptotic guarantees for successful structure learning with logarithmic sample complexities for bounded-degree graphs for a Wasserstein DRO method Section 3.3.2 and a KL DRO method Section 3.3.3. Numerical study on synthetic and real datasets is provided in Section 3.4 with concluding remarks in Section 3.5.

3.1 Introduction

A Bayesian network is a prominent class of probabilistic graphical models that encodes the conditional dependencies among variables with a directed acyclic graph (DAG). It provides a mathematical framework for formally understanding the interaction among variables of interest, together with computationally attractive factorization for modeling multivariate distributions. If we impose causal relationships on the edges between variables, the model becomes a causal

Bayesian network that encodes the more informative causation. Without such interpretation, a Bayesian network serves as a dependency graph for factorization of a multivariate distribution. We focus on discrete Bayesian networks with purely categorical random variables that are not ordinal, but will discuss related work on both discrete and continuous Bayesian networks for completeness.

The associated DAG structure of a Bayesian network is usually unknown. Structure learning is therefore an important task that infers the structure from data. The *score-based* approach defines a scoring function that measures the goodness-of-fit of each structure and aims to find an optimal DAG that maximizes the score. Unfortunately, the resulting combinatorial optimization problem is known to be NP-hard (Chickering et al., 2004) without distributional assumptions. Representative approaches include those based on heuristic search (Chickering, 2002), dynamic programming (Silander and Myllymäki, 2006), integer linear programming (Jaakkola et al., 2010) or continuous optimization (Zheng et al., 2018), which either yields an approximate solution or an exact solution in worst-case exponential time. The *constraint-based* approach (Spirtes and Glymour, 1991; Spirtes et al., 1999; Colombo et al., 2014) performs conditional independence tests to determine the existence and directionality of edges. The time complexity is, however, exponential with the maximum in-degree. Furthermore, the independence test results may be unreliable or inconsistent with the true distribution because of finite samples or even corrupted samples. In general, without interventional data or assumptions on the underlying distribution, we can only identify a Markov equivalence class (MEC) the true DAG

belongs to from observational data where DAGs in the MEC are Markov equivalent, that is, encoding the same set of conditional independencies.

A super-structure is an undirected graph that contains as sub-graphs the skeleton which removes directionality from the true DAG. It has been shown that a given super-structure possibly reduces the search space or the number of independence tests to be performed. For example, exact structure learning of Bayesian networks may be (fixed-parameter) tractable if the super-structure satisfies certain graph-theoretic properties such as bounded tree-width (Korhonen and Parviainen, 2013; Loh and Bühlmann, 2014), bounded maximum degree (Ordyniak and Szeider, 2013) and the feedback edge number (Ganian and Korchemna, 2021). An incomplete super-structure with missing edges also helps improve the learned DAG with a post-processing hill-climbing method (Tsamardinos et al., 2006; Perrier et al., 2008). Furthermore, a combination of a skeleton and a variable ordering determines a unique DAG structure. Learning the exact skeleton rather than a rough super-structure is desirable in Bayesian network structure learning.

(Spirtes and Glymour, 1991; Tsamardinos et al., 2006) make use of independence tests to estimate the skeleton. (Loh and Bühlmann, 2014) learn a super-structure called moralized graph via graphical lasso (Friedman et al., 2008). (Shojaie and Michailidis, 2010) learn the skeleton assuming an ordering of variables. (Bank and Honorio, 2020) leverage linear regression for skeleton recovery in polynomial time. These methods either rely on independence test results, which are unstable, or a regularized ERM problem, where regularization is usually heuristically chosen to combat overfitting. In practice, the observational data is commonly contaminated by sensor failure, transmission error or adversarial perturbation. Sometimes only a small amount

of data is available for learning. As a result, the existing algorithms are vulnerable to such distributional uncertainty and may produce false edges in the estimated skeleton.

In this chapter, we propose a DRO method (Rahimian and Mehrotra, 2019) that solves a node-wise multivariate regression problem (Bank and Honorio, 2020) for structure learning of general discrete Bayesian networks to overcome the above limitations. We focus on *skeleton learning* based on the above arguments. We do not assume any specific form of conditional distributions. We take into account the settings with only a small amount of samples (high-dimensional) and potential perturbations, which makes the true data generating distribution highly uncertain. Our method explicitly models the uncertainty by constructing an ambiguity set of distributions characterized by certain a priori properties of the true distribution. The optimal parameter is learned by minimizing the worst-case expected loss over all the distributions within the ambiguity set so that it performs uniformly well on all the considered distributions. The ambiguity set is usually defined in such a way that it includes all the distributions close to the empirical distribution in terms of some divergence. With an appropriately chosen divergence measure, the set contains the true distribution with high probability. Hence the worst-case risk can be interpreted as an upper confidence bound of the true risk. The fact that a discrete Bayesian network encompasses an exponential number of states may pose a challenge to solve the DRO problem. We develop efficient algorithms for problems with ambiguity sets defined by Wasserstein distances and KL divergences. We show that a group regularized regression method is a special case of our approach. We study statistical guarantees of the proposed estimators such

as sample complexities. Experimental results on synthetic and real-world datasets contaminated by various perturbations validate the superior performance of the proposed methods.

3.1.1 Related Work

In addition to the score-based structure learning methods and constraint-based methods discussed in the introduction section, there are a third class of hybrid algorithms leveraging constraint-based methods to restrict the search space of a score-based method (Tsamardinos et al., 2006; Gasse et al., 2014; Nandy et al., 2018). Due to space limitation, it is quite likely that the related work is not covered thoroughly, and we refer the interested readers to survey papers (Drton and Maathuis, 2017; Heinze-Deml et al., 2018; Constantinou et al., 2021) for more details.

Recently, there is a emerging line of work proposing polynomial-time algorithms for DAG learning (Park and Raskutti, 2017; Ghoshal and Honorio, 2017; Ghoshal and Honorio, 2018; Chen et al., 2019; Bank and Honorio, 2020; Gao et al., 2020; Rajendran et al., 2021), among which (Bank and Honorio, 2020) particularly focuses on general discrete Bayesian networks without resorting to independence tests. There is also a flurry of work on score-based methods based on neural networks and continuous optimization (Zheng et al., 2018; Wei et al., 2020; Ng et al., 2020; Yu et al., 2021b; Ng et al., 2022; Gao et al., 2022), motivated by differentiable characterization of acyclicity without rigorous theoretical guarantees.

Learning a super-structure can be done by independence tests, graphical lasso or regression, as discussed in introduction. Given a super-structure, how to determine the orientation has

been studied by (Perrier et al., 2008; Ordyniak and Szeider, 2013; Korhonen and Parviainen, 2013; Loh and Bühlmann, 2014; Ng et al., 2021; Ganian and Korchemna, 2021).

3.2 Preliminaries

We introduce necessary background for Bayesian networks, a baseline method and a few assumptions in this section.

Let \mathbb{P} be a discrete joint probability distribution on n categorical random variables $\mathcal{V} := \{X_1, X_2, \dots, X_n\}$. Let $\mathcal{G} := (\mathcal{V}, \mathcal{E}_{\text{true}})$ be a DAG with edge set $\mathcal{E}_{\text{true}}$. We use X_i to represent the i -th random variable or node interchangeably. We call $(\mathcal{G}, \mathbb{P})$ a Bayesian network if it satisfies the Markov condition, i.e., each variable X_r is independent of any subset of its non-descendants conditioned on its parents \mathbf{Pa}_r . We denote the children of X_r by \mathbf{Ch}_r , its neighbors by $\mathbf{Ne}_r := \mathbf{Pa}_r \cup \mathbf{Ch}_r$ and the complement by $\mathbf{Co}_r := [n] - \mathbf{Ne}_r - \{r\}$. The joint probability distribution can thus be factorized in terms of local conditional distributions:

$$\mathbb{P}(\mathbf{X}) = \mathbb{P}(X_1, X_2, \dots, X_n) \triangleq \prod_{i=1}^n \mathbb{P}(X_i | \mathbf{Pa}_i).$$

Let the skeleton $\mathcal{G}_{\text{skel}} := (\mathcal{V}, \mathcal{E}_{\text{skel}})$ be the undirected graph that removes directionality from \mathcal{G} . Given a set of m samples $\{\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \dots, \mathbf{x}^{(m)}\}$ drawn i.i.d. from \mathbb{P} , the goal of skeleton learning is to estimate $\mathcal{G}_{\text{skel}}$ from the samples.

We do not assume faithfulness (Spirtes et al., 2000) or any specific parametric form for the conditional probability distributions. The unavailability of a true model entails a substitute

model. (Bank and Honorio, 2020) propose such a model based on encoding schemes and surrogate parameters as follows.

Assume that each variable X_r takes values from a finite set \mathcal{C}_r with cardinality $|\mathcal{C}_r| > 1$. For an indexing set $\mathcal{S} \subseteq [n]$, define $\rho_{\mathcal{S}} := \sum_{i \in \mathcal{S}} |\mathcal{C}_i| - 1$ and $\rho_{\mathcal{S}}^+ := \sum_{i \in \mathcal{S}} |\mathcal{C}_i|$. The maximum cardinality minus one is defined as $\rho_{\max} := \max_{i \in [n]} |\mathcal{C}_i| - 1$. Let $\mathcal{S}_r := \bigcup_{i \in \mathbf{Ne}_r} \{\rho_{[i-1]} + 1, \dots, \rho_{[i]}\}$ be indices for \mathbf{Ne}_r in $\rho_{[n]}$ and its complement by $\mathcal{S}_r^c := [\rho_{[n]}] - \mathcal{S}_r - \{\rho_{[r-1]} + 1, \dots, \rho_{[r]}\}$. Let $\mathcal{E} : \mathcal{C}_r \rightarrow \mathcal{B}^{\rho_r}$ be an encoding mapping for a bounded and countable set $\mathcal{B} \subset \mathbb{R}$. We adopt encoding schemes with $\mathcal{B} = \{-1, 0, 1\}$ such as dummy encoding and unweighted effects encoding which satisfy a linear independence condition. With a little abuse of notation, we reuse \mathcal{E} for encoding any X_r and denote by $\mathcal{E}(\mathbf{X}_{\mathcal{S}}) \in \mathcal{B}^{\rho_{\mathcal{S}}}$ the concatenation of the encoded vectors $\{\mathcal{E}(X_i)\}_{i \in \mathcal{S}}$. Consider a linear structural equation model for each X_r : $\mathcal{E}(X_r) = \mathbf{W}^{*\top} \mathcal{E}(\mathbf{X}_{\bar{r}}) + \mathbf{e}$, where $\mathbf{W}^* \triangleq [\mathbf{W}_1^* \cdots \mathbf{W}_{r-1}^* \mathbf{W}_{r+1}^* \cdots \mathbf{W}_n^*]^\top \in \mathbb{R}^{\rho_{\bar{r}} \times \rho_r}$ with $\mathbf{W}_i^* \in \mathbb{R}^{\rho_i \times \rho_r}$ is a surrogate parameter matrix and $\mathbf{e} \in \mathbb{R}^{\rho_r}$ is a vector of errors not necessarily independent of other quantities. A natural choice of a fixed \mathbf{W}^* is the solution to the following least-square problem given knowledge of the true distribution and the true DAG:

$$\begin{aligned} \mathbf{W}^* &\in \arg \inf_{\mathbf{W}} \frac{1}{2} \mathbb{E}_{\mathbb{P}} \|\mathcal{E}(X_r) - \mathbf{W}^\top \mathcal{E}(\mathbf{X}_{\bar{r}})\|_2^2 \\ \text{s.t. } \quad &\mathbf{W}_i = \mathbf{0} \quad \forall i \in \mathbf{Co}_r. \end{aligned} \tag{3.1}$$

Therefore

$$\mathbf{W}^* = (\mathbf{W}_{\mathcal{S}_r}^*; \mathbf{0})$$

$$\mathbf{W}_{\mathcal{S}_r}^* = \mathbb{E}_{\mathbb{P}}[\mathcal{E}(\mathbf{X}_{\bar{r}})_{\mathcal{S}_r} \mathcal{E}(\mathbf{X}_{\bar{r}})_{\mathcal{S}_r}^{\top}]^{-1} \mathbb{E}_{\mathbb{P}}[\mathcal{E}(\mathbf{X}_{\bar{r}})_{\mathcal{S}_r} \mathcal{E}(X_r)^{\top}]$$

is the optimal solution by the first-order optimality condition assuming that $\mathbb{E}_{\mathbb{P}}[\mathcal{E}(\mathbf{X}_{\bar{r}})_{\mathcal{S}_r} \mathcal{E}(\mathbf{X}_{\bar{r}})_{\mathcal{S}_r}^{\top}]$ is invertible. The expression of $\mathbf{W}_{\mathcal{S}_r}^*$ captures the intuitions that neighbor nodes should be highly related to the current node r while the interaction among neighbor nodes should be weak for them to be distinguishable. We further assume that the errors are bounded:

Assumption 17 (Bounded error). For the error vector, $\|\mathbf{e}\|_{\infty} \leq \sigma$ and $\|\mathbb{E}_{\mathbb{P}}[\mathbf{e}]\|_{\infty} \leq \mu$.

Note that the true distribution does not have to follow a linear structural equation model. Equation 3.1 only serves as a surrogate model to find technical conditions for successful skeleton learning, which will be discussed in a moment.

The surrogate model under the true distribution indicates that $\|\mathbf{W}_i^*\|_{2,2} > 0 \implies X_i \in \mathbf{Ne}_r$.

This suggests a regularized empirical risk minimization problem to estimate \mathbf{W}^* :

$$\tilde{\mathbf{W}} \in \arg \inf_{\mathbf{W}} \tilde{L}(\mathbf{W}) := \frac{1}{2} \mathbb{E}_{\tilde{\mathbb{P}}_m} \|\mathcal{E}(X_r) - \mathbf{W}^{\top} \mathcal{E}(\mathbf{X}_{\bar{r}})\|_2^2 + \tilde{\lambda} \|\mathbf{W}\|_{B,2,1}, \quad (3.2)$$

where $\tilde{\lambda} > 0$ is a regularization coefficient, the block $\ell_{2,1}$ norm is adopted to induce sparsity and $\tilde{\mathbb{P}}_m := \frac{1}{m} \sum_{i=1}^m \delta_{\mathbf{x}^{(i)}}$ stands for the empirical distribution with $\delta_{\mathbf{x}^{(i)}}$ being the Dirac point

measure at $\mathbf{x}^{(i)}$. This approach is expected to succeed as long as only neighbor nodes have a non-trivial impact on the current node, namely, $\|\mathbf{W}_i^*\|_{2,2} > 0 \iff X_i \in \mathbf{Ne}_r$.

Define the risk of some \mathbf{W} under a distribution $\tilde{\mathbb{P}}$ as

$$R^{\tilde{\mathbb{P}}}(\mathbf{W}) := \mathbb{E}_{\tilde{\mathbb{P}}} \ell_{\mathbf{W}}(\mathbf{X}) := \mathbb{E}_{\tilde{\mathbb{P}}} \frac{1}{2} \|\mathcal{E}(X_r) - \mathbf{W}^\top \mathcal{E}(\mathbf{X}_{\bar{r}})\|_2^2,$$

where $\ell_{\mathbf{W}}(\cdot)$ is the squared loss function. The Hessian of the empirical risk $R^{\tilde{\mathbb{P}}_m}(\mathbf{W})$ is a block diagonal matrix:

$$\nabla^2 R^{\tilde{\mathbb{P}}_m}(\mathbf{W}) \triangleq \tilde{\mathbf{H}} \otimes \mathbf{I}_{\rho_r} \in \mathbb{R}^{\rho_r \rho_{\bar{r}} \times \rho_r \rho_{\bar{r}}},$$

where $\tilde{\mathbf{H}} := \mathbb{E}_{\tilde{\mathbb{P}}_m} [\mathcal{E}(\mathbf{X}_{\bar{r}}) \mathcal{E}(\mathbf{X}_{\bar{r}})^\top] \in \mathbb{R}^{\rho_{\bar{r}} \times \rho_{\bar{r}}}$ and $\mathbf{I}_{\rho_r} \in \mathbb{R}^{\rho_r \times \rho_r}$ is the identity matrix of dimension ρ_r . Similarly under the true distribution, $\mathbf{H} := \mathbb{E}_{\mathbb{P}} [\mathcal{E}(\mathbf{X}_{\bar{r}}) \mathcal{E}(\mathbf{X}_{\bar{r}})^\top]$. As a result, \mathbf{H} is independent of the surrogate parameters \mathbf{W}^* thus conditions on the Hessian translates to conditions on a matrix of cross-moments of encodings, which only depend on the encoding function \mathcal{E} and \mathbb{P} .

In order for this baseline method to work, we make the following assumptions.

Assumption 18 (Minimum weight). For each node r , the minimum norm of the true weight matrix \mathbf{W}^* for neighbor nodes is lower bounded: $\min_{i \in \mathbf{Ne}_r} \|\mathbf{W}_i\|_F \geq \beta > 0$.

Assumption 19 (Positive definiteness of the Hessian). For each node r , $\mathbf{H}_{\mathcal{S}_r \mathcal{S}_r} > 0$, or equivalently, $\Lambda_{\min}(\mathbf{H}_{\mathcal{S}_r \mathcal{S}_r}) \geq \Lambda > 0$ where $\Lambda_{\min}(\cdot)$ denotes the minimum eigenvalue.

Assumption 20 (Mutual incoherence). For each node r , $\|\mathbf{H}_{\mathcal{S}_r^c \mathcal{S}_r} \mathbf{H}_{\mathcal{S}_r \mathcal{S}_r}^{-1}\|_{B,1,\infty} \leq 1 - \alpha$ for some $0 < \alpha \leq 1$.

Assumption 19 ensures that Equation 3.2 yields a unique solution. Assumption 20 is a widely adopted assumption that controls the impact of non-neighbor nodes on r (Wainwright, 2009; Ravikumar et al., 2010; Daneshmand et al., 2014). One interpretation is that the rows of $\mathbf{H}_{\mathcal{S}_r^c \mathcal{S}_r}$ should be nearly orthogonal to the rows of $\mathbf{H}_{\mathcal{S}_r \mathcal{S}_r}$. (Bank and Honorio, 2020) show that these assumptions hold for common encoding schemes and finite-sample settings with high probability under mild conditions. They also show that incoherence is more commonly satisfied for the neighbors than the Markov blanket, which justifies the significance of skeleton learning.

Finally, we take the union of all the learned neighbor nodes for each $r \in [n]$ by solving Equation 3.2 to get the estimated skeleton $\tilde{\mathcal{G}} := (\mathcal{V}, \tilde{\mathcal{E}}_{\text{skel}})$. The directions can be inferred based on the learned skeleton to obtain a DAG by applying existing methods introduced in Section 3.1.1.

3.3 Method

As noted in (Bank and Honorio, 2020), due to model mis-specification, even in the infinite sample setting, there is possible discrepancy between the ERM minimizer $\tilde{\mathbf{W}}$ and the true solution \mathbf{W}^* , resulting in false or missing edges. In the high-dimensional setting ($m < n$) or the adversarial setting, this issue becomes more serious due to a limited knowledge about the data-generating mechanism \mathbb{P} .

In this section, we attempt to leverage a DRO framework to incorporate distributional uncertainty into the estimation process. We adopt two types of ambiguity sets and present

efficient algorithms to solve the specific problems. We derive theoretical guarantees, together with a connection between Equation 3.2 and our methods.

3.3.1 Basic Formulation

Let \mathcal{X} be a measurable space of all states of the Bayesian network $(\mathcal{G}, \mathbb{P})$, i.e., $\mathbf{X} \in \mathcal{X}$. Let $\mathcal{P}(\mathcal{X})$ be the space of all Borel probability measures on \mathcal{X} . Denote by $\mathcal{X}^{\mathcal{E}} := \{\mathcal{E}(\mathbf{X}) : \forall \mathbf{X} \in \mathcal{X}\}$ the space of all the allowed encodings.

Instead of minimizing the empirical risk and relying on regularization, we seek a distributionally robust estimator that optimizes the worst-case risk over an ambiguity set of distributions:

$$\hat{\mathbf{W}} \in \arg \inf_{\mathbf{W}} \sup_{\mathbb{Q} \in \mathcal{A}} \frac{1}{2} \mathbb{E}_{\mathbb{Q}} \|\mathcal{E}(X_r) - \mathbf{W}^{\top} \mathcal{E}(\mathbf{X}_{\bar{r}})\|_2^2. \quad (3.3)$$

This way of uncertainty quantification can be interpreted as an adversary that captures out-of-sample effect by making perturbations on samples within some budget ε . Some common statistical distances satisfy $\text{div}(\mathbb{Q}, \mathbb{P}) = 0 \iff \mathbb{Q} = \mathbb{P}$. In this case, if ε is set to zero, Equation 3.3 reduces to Equation 3.2 without regularization. We will show that the DRO estimator $\hat{\mathbf{W}}$ can be found efficiently and encompasses attractive statistical properties with a judicious choice of \mathcal{A} .

3.3.2 Wasserstein Formulation

We adopt the Wasserstein distance of order $p = 1$ as the discrepancy measure, the empirical distribution as the nominal distribution, and cost function $c(\mathbf{x}, \mathbf{x}') = \|\mathcal{E}(\mathbf{x}) - \mathcal{E}(\mathbf{x}')\|$ for some norm $\|\cdot\|$. The primal DRO formulation becomes

$$\hat{\mathbf{W}} \in \arg \inf_{\mathbf{W}} \sup_{\mathbb{Q} \in \mathcal{A}_\varepsilon^{W_p}(\tilde{\mathbb{P}}_m)} \frac{1}{2} \mathbb{E}_{\mathbb{Q}} \|\mathcal{E}(X_r) - \mathbf{W}^\top \mathcal{E}(\mathbf{X}_{\bar{r}})\|_2^2. \quad (3.4)$$

The dual problem of Equation 3.4 can be written as

$$\inf_{\mathbf{W}, \gamma \geq 0} \gamma \varepsilon + \frac{1}{m} \sum_{i=1}^m \sup_{\mathbf{x} \in \mathcal{X}} \frac{1}{2} \|\mathcal{E}(x_r) - \mathbf{W}^\top \mathcal{E}(\mathbf{x}_{\bar{r}})\|_2^2 - \gamma \|\mathcal{E}(\mathbf{x}) - \mathcal{E}(\mathbf{x}^{(i)})\|. \quad (3.5)$$

Strong duality holds according to Theorem 1 in (Gao and Kleywegt, 2022). The inner supremum problems can be solved independently for each data sample $\mathbf{x}^{(i)}$. Henceforth, we focus on solving it for some $i \in [m]$:

$$\sup_{\mathbf{x} \in \mathcal{X}} \frac{1}{2} \|\mathcal{E}(x_r) - \mathbf{W}^\top \mathcal{E}(\mathbf{x}_{\bar{r}})\|_2^2 - \gamma \|\mathcal{E}(\mathbf{x}) - \mathcal{E}(\mathbf{x}^{(i)})\|. \quad (3.6)$$

Equation 3.6 is a supremum of $|\mathcal{X}|$ convex functions of \mathbf{W} , thus convex. Since \mathcal{X}^ε is a discrete set consisting of a factorial number of points $(\Pi_{i \in [n]} \rho_i)$, unlike (Chen and Paschalidis, 2018), we may not simplify Equation 3.6 into a regularization form by leveraging convex conjugate functions because \mathcal{X}^ε is non-convex and not equal to $\mathbb{R}^{\rho[n]}$. Moreover, since changing the value of x_j for some $j \in \bar{r}$ is equivalent to changing $\mathbf{W}^\top \mathcal{E}(\mathbf{x}_{\bar{r}})$ by a vector, unlike (Li et al., 2022b)

where only a set of scalar values are dealt with, there may not be a greedy algorithm based on sufficient statistics to find the optimal solution to Equation 3.6. In fact, let the norm be the ℓ_1 norm, we can rewrite Equation 3.6 by fixing the values of $\|\mathcal{E}(\mathbf{x}) - \mathcal{E}(\mathbf{x}^{(i)})\|_1$:

$$\sup_{\substack{\mathbf{x} \in \mathcal{X}, 0 \leq k \leq \rho_{[n]}^+, \\ \|\mathcal{E}(\mathbf{x}) - \mathcal{E}(\mathbf{x}^{(i)})\|_1 = k}} \frac{1}{2} \|\mathcal{E}(x_r) - \mathbf{W}^\top \mathcal{E}(\mathbf{x}_{\bar{r}})\|_2^2 - \gamma k. \quad (3.7)$$

If we fix k , Equation 3.7 is a generalization of the 0-1 quadratic programming problem, which can be transformed into a maximizing quadratic programming (MAXQP) problem. As a result, Equation 3.6 is an NP-hard problem. (Charikar and Wirth, 2004) develop an algorithm to find an $\Omega(1/\log n)$ solution based on semi-definite programming (SDP) and sampling for the MAXQP problem. Instead of adopting a similar SDP algorithm with quadratic constraints, we propose a greedy algorithm to approximate the optimal solution, which is illustrated in Algorithm 2. It follows a simple idea that for a random node order $\boldsymbol{\pi}$, we select a partial optimal solution sequentially from π_1 to π_n . We enumerate the possible states of the first node to reduce uncertainty. In practice, we find that this algorithm always finds the exact solution that is NP-hard to find for random data with $n \leq 10$ and $\rho_{\max} \leq 5$.

Since $\mathcal{X}^\mathcal{E}$ is non-convex and not equal to $\mathbb{R}^{\rho[n]}$, using convex conjugate functions will not yield exact equivalence between Equation 3.5 and a regularized ERM problem. However, we can draw such a connection by imposing constraints on the dual variables as shown by the following proposition:

Algorithm 2 Greedy Algorithm for the Wasserstein Worst-case Risk

Input: \mathbf{W} , γ , $\mathbf{x}^{(i)}$
Output: a solution $\hat{\mathbf{x}}$ to Equation 3.6
 Initialize $\hat{\mathbf{x}} = \mathbf{x}^{(i)}$
for all $(j, x_j^t) \in [n] \times \mathcal{C}_j$ **do**
 Get a random permutation π over $[n]$ with $\pi_1 = j$
 for $k := 2$ **to** n **do**
 $x_{\pi_j}^t \leftarrow \arg \sup_{x_{\pi_k}^t} \ell_{\mathbf{W}}(\mathbf{x}_{\pi_{[k]}}^t) - \gamma \|\mathcal{E}(\mathbf{x}_{\pi_{[k]}}^t) - \mathcal{E}(\mathbf{x}_{\pi_{[k]}}^{(i)})\|$
 end for
 if \mathbf{x}^t yields a greater objective than $\hat{\mathbf{x}}$ **then**
 $\hat{\mathbf{x}} \leftarrow \mathbf{x}^t$
 end if
end for

Proposition 21 (Regularization Equivalence). *Let $\ddot{\mathbf{W}} := [\mathbf{W}; -\mathbf{I}_{\rho_r}]^\top \in \mathbb{R}^{\rho_{[n]} \times \rho_r}$ with $\mathbf{W}_r = -\mathbf{I}_{\rho_r}$. If $\gamma \geq \rho_{[n]} \|\ddot{\mathbf{W}}\|_F^2$, the Wasserstein distributionally robust regression problem in Equation 3.5 is equivalent to*

$$\inf_{\mathbf{W}} \mathbb{E}_{\tilde{\mathbb{P}}_m} \frac{1}{2} \|\mathcal{E}(X_r) - \mathbf{W}^\top \mathcal{E}(\mathbf{X}_{\bar{r}})\|_2^2 + \varepsilon \rho_{[n]} \|\ddot{\mathbf{W}}\|_F^2,$$

which subsumes a linear regression approach regularized by the Frobenius norm as a special case.

Proof. Recapitulating on Equation 3.6:

$$\sup_{\mathbf{x} \in \mathcal{X}} \frac{1}{2} \|\mathcal{E}(x_r) - \mathbf{W}^\top \mathcal{E}(\mathbf{x}_{\bar{r}})\|_2^2 - \gamma \|\mathcal{E}(\mathbf{x}) - \mathcal{E}(\mathbf{x}^{(i)})\|_1.$$

Observe that

$$\begin{aligned}
\|\mathcal{E}(x_r) - \mathbf{W}^\top \mathcal{E}(\mathbf{x}_{\bar{r}})\|_2^2 &\triangleq \|\ddot{\mathbf{W}}^\top \mathcal{E}(\mathbf{x}_{[n]})\|_2^2 \\
&\leq \|\ddot{\mathbf{W}}^\top\|_{\infty,2}^2 \\
&\leq \|\ddot{\mathbf{W}}\|_{1,2}^2 \\
&\leq \rho_{[n]} \|\ddot{\mathbf{W}}\|_F^2 \\
&\leq \gamma
\end{aligned}$$

Therefore, for any $\mathbf{x} \neq \mathbf{x}^{(i)}$,

$$\begin{aligned}
&\frac{1}{2} \|\mathcal{E}(x_r) - \mathbf{W}^\top \mathcal{E}(\mathbf{x}_{\bar{r}})\|_2^2 - \gamma \|\mathcal{E}(\mathbf{x}) - \mathcal{E}(\mathbf{x}^{(i)})\|_1 \\
&\quad - \left(\frac{1}{2} \|\mathcal{E}(x_r^{(i)}) - \mathbf{W}^\top \mathcal{E}(\mathbf{x}_{\bar{r}}^{(i)})\|_2^2 - \gamma \|\mathcal{E}(\mathbf{x}^{(i)}) - \mathcal{E}(\mathbf{x}^{(i)})\|_1 \right) \\
&\leq \frac{1}{2} (\|\mathcal{E}(x_r) - \mathbf{W}^\top \mathcal{E}(\mathbf{x}_{\bar{r}})\|_2^2 - \|\mathcal{E}(x_r^{(i)}) - \mathbf{W}^\top \mathcal{E}(\mathbf{x}_{\bar{r}}^{(i)})\|_2^2) - \gamma \|\mathcal{E}(\mathbf{x}) - \mathcal{E}(\mathbf{x}^{(i)})\|_1 \\
&\leq \frac{1}{2} (2\gamma) - \gamma \|\mathcal{E}(\mathbf{x}) - \mathcal{E}(\mathbf{x}^{(i)})\|_1 \\
&\leq \gamma - \gamma \\
&= 0,
\end{aligned}$$

which implies that the supremum can always be achieved at $\mathbf{x} = \mathbf{x}^{(i)}$. Minimizing over γ leads to

$$\inf_{\mathbf{W}} \mathbb{E}_{\tilde{\mathbb{P}}_m} \frac{1}{2} \|\mathcal{E}(X_r) - \mathbf{W}^\top \mathcal{E}(\mathbf{X}_{\bar{r}})\|_2^2 + \varepsilon \rho_{[n]} \|\ddot{\mathbf{W}}\|_F^2.$$

□

This suggests that minimizing a regularized empirical risk may not be enough to achieve distributional robustness. Note that the exact equivalence result in (Chen and Paschalidis, 2018) requires $\mathcal{X}^{\mathcal{E}} = \mathbb{R}^d$ for some d .

Now we perform non-asymptotic analysis on the proposed DRO estimator $\hat{\mathbf{W}}$. First, we would like to show that the solution to the Wasserstein DRO estimator in Equation 3.4 is unique so that we refer to an estimator unambiguously. Note that Equation 3.4 is a convex optimization problem but not necessarily strictly convex, and actually never convex in the high-dimensional setting. However, given a sufficient number of samples, the problem becomes strictly convex and yields a unique solution with high probability. Second, we show that the correct skeleton $\mathcal{E}_{\text{skel}}$ can be recovered with high probability given enough samples. This is achieved by showing that, for each node X_r , the estimator has zero weights for non-neighbor nodes \mathbf{Co}_r and has non-zero weights for its neighbors \mathbf{Ne}_r with high confidence.

3.3.2.1 Lemmas for Non-asymptotic Analysis

Before presenting the main results, we note that they are based on several important lemmas.

Lemma 22. *Suppose Ξ is separable Banach space and fix $\mathbb{P}_0 \in \mathcal{P}(\Xi')$ for some $\Xi' \subseteq \Xi$. Suppose $c : \Xi \rightarrow \mathbb{R}_{\geq 0}$ is closed convex, k -positively homogeneous. Suppose $f : \Xi \rightarrow \mathcal{Y}$ is a mapping in the Lebesgue space of functions with finite first-order moment under \mathbb{P}_0 and upper semi-continuous*

with finite Lipschitz constant $\text{lip}_c(f)$. Then for all $\varepsilon \geq 0$, the following inequality holds with probability 1:

$$\sup_{\substack{\mathbb{Q} \in \mathcal{A}_\varepsilon^{W_p}(\mathbb{P}_0), \\ \mathbb{Q} \in \mathcal{P}(\Xi')}} \int f(\xi') \mathbb{Q}(d\xi') \leq \varepsilon \text{lip}_c(f) + \int f(\xi') \mathbb{P}_0(d\xi').$$

Proof. The result follows directly from Theorem 1 in (Cranko et al., 2021):

$$\sup_{\mathbb{Q} \in \mathcal{A}_\varepsilon^{W_p}(\mathbb{P}_0), \mathbb{Q} \in \mathcal{P}(\Xi)} \int f(\xi) \mathbb{Q}(d\xi) \leq \varepsilon \text{lip}_c(f) + \int f(\xi') \mathbb{P}_0(d\xi').$$

Since $\Xi' \subseteq \Xi$, observe

$$\sup_{\mathbb{Q} \in \mathcal{A}_\varepsilon^{W_p}(\mathbb{P}_0), \mathbb{Q} \in \mathcal{P}(\Xi')} \int f(\xi') \mathbb{Q}(d\xi') \leq \sup_{\mathbb{Q} \in \mathcal{A}_\varepsilon^{W_p}(\mathbb{P}_0), \mathbb{Q} \in \mathcal{P}(\Xi)} \int f(\xi) \mathbb{Q}(d\xi).$$

□

Lemma 22 follows directly from (Cranko et al., 2021) and allows us to obtain an upper bound between the worst-case risk and empirical risk. It is crucial for the following finite-sample guarantees.

Lemma 23. *If Assumption 19 holds, for any $\mathbb{Q} \in \mathcal{A}_\varepsilon^{W_p}(\tilde{\mathbb{P}}_m)$, with high probability, $\mathbf{H}_{S_r S_r}^\mathbb{Q}$ is positive definite.*

Proof. The minimum eigenvalue of the true covariance matrix $\mathbf{H}_{\mathcal{S}_r\mathcal{S}_r}$ satisfies

$$\begin{aligned}\Lambda_{\min}(\mathbf{H}_{\mathcal{S}_r\mathcal{S}_r}) &\triangleq \min_{\|\mathbf{v}\|_2=1} \mathbf{v}^\top \mathbf{H}_{\mathcal{S}_r\mathcal{S}_r} \mathbf{v} \\ &= \min_{\|\mathbf{v}\|_2=1} \mathbf{v}^\top \mathbf{H}_{\mathcal{S}_r\mathcal{S}_r}^{\mathbb{Q}} \mathbf{v} + \mathbf{v}^\top (\tilde{\mathbf{H}}_{\mathcal{S}_r\mathcal{S}_r} - \mathbf{H}_{\mathcal{S}_r\mathcal{S}_r}^{\mathbb{Q}}) \mathbf{v} + \mathbf{v}^\top (\mathbf{H}_{\mathcal{S}_r\mathcal{S}_r} - \tilde{\mathbf{H}}_{\mathcal{S}_r\mathcal{S}_r}) \mathbf{v} \\ &\leq \Lambda_{\min}(\mathbf{H}_{\mathcal{S}_r\mathcal{S}_r}^{\mathbb{Q}}) + \mathbf{u}^\top (\tilde{\mathbf{H}}_{\mathcal{S}_r\mathcal{S}_r} - \mathbf{H}_{\mathcal{S}_r\mathcal{S}_r}^{\mathbb{Q}}) \mathbf{u} + \mathbf{u}^\top (\mathbf{H}_{\mathcal{S}_r\mathcal{S}_r} - \tilde{\mathbf{H}}_{\mathcal{S}_r\mathcal{S}_r}) \mathbf{u},\end{aligned}$$

where $\|\mathbf{u}\|_2 = 1$ is an eigenvector of $\mathbf{H}_{\mathcal{S}_r\mathcal{S}_r}^{\mathbb{Q}}$ with minimum eigenvalue.

Therefore, $\Lambda_{\min}(\mathbf{H}_{\mathcal{S}_r\mathcal{S}_r}^{\mathbb{Q}})$ can be lower bounded as follows:

$$\begin{aligned}\Lambda_{\min}(\mathbf{H}_{\mathcal{S}_r\mathcal{S}_r}^{\mathbb{Q}}) &\geq \Lambda_{\min}(\mathbf{H}_{\mathcal{S}_r\mathcal{S}_r}) - \mathbf{u}^\top (\tilde{\mathbf{H}}_{\mathcal{S}_r\mathcal{S}_r} - \mathbf{H}_{\mathcal{S}_r\mathcal{S}_r}^{\mathbb{Q}}) \mathbf{u} - \mathbf{u}^\top (\mathbf{H}_{\mathcal{S}_r\mathcal{S}_r} - \tilde{\mathbf{H}}_{\mathcal{S}_r\mathcal{S}_r}) \mathbf{u} \\ &\geq \Lambda_{\min}(\mathbf{H}_{\mathcal{S}_r\mathcal{S}_r}) - |\mathbf{u}^\top (\tilde{\mathbf{H}}_{\mathcal{S}_r\mathcal{S}_r} - \mathbf{H}_{\mathcal{S}_r\mathcal{S}_r}^{\mathbb{Q}}) \mathbf{u}| - \|(\mathbf{H}_{\mathcal{S}_r\mathcal{S}_r} - \tilde{\mathbf{H}}_{\mathcal{S}_r\mathcal{S}_r})\|_F,\end{aligned}$$

due to the fact that

$$\mathbf{u}^\top \mathbf{H} \mathbf{u} \leq \Lambda_{\max}(\mathbf{H}) \leq \sqrt{\sum_i (\Lambda_i(\mathbf{H}))^2} \leq \|\mathbf{H}\|_{2,2}.$$

We can obtain an upper bound on $|\mathbf{u}^\top (\tilde{\mathbf{H}}_{\mathcal{S}_r\mathcal{S}_r} - \mathbf{H}_{\mathcal{S}_r\mathcal{S}_r}^{\mathbb{Q}}) \mathbf{u}|$ based on Lemma 22:

$$|\mathbf{u}^\top (\tilde{\mathbf{H}}_{\mathcal{S}_r\mathcal{S}_r} - \mathbf{H}_{\mathcal{S}_r\mathcal{S}_r}^{\mathbb{Q}}) \mathbf{u}| \leq 4|\mathcal{S}_r|^{\frac{1}{2}}\varepsilon,$$

because for function $g(\mathcal{E}(\mathbf{x})) := \mathbf{u}^\top \mathbf{H}_{\mathcal{S}_r \mathcal{S}_r} \mathbf{u}$, it can be shown that for any $\|\mathcal{E}(\mathbf{x}) - \mathcal{E}(\mathbf{x}')\|_1 = k$ and some $|\mathcal{S}| = k$,

$$|g(\mathcal{E}(\mathbf{x})) - g(\mathcal{E}(\mathbf{x}'))| \leq \sum_{k \in \mathcal{S}} \sum_{i \in \mathcal{S}_r} |H_{ik} - H'_{ik}| u_i u_k + |H_{ki} - H'_{ki}| u_k u_i \leq 4k |\mathcal{S}_r|^{\frac{1}{2}}.$$

Recall that we assume that the encoding schemes take values in $\mathcal{B} = \{-1, 0, 1\}$. Therefore $\text{lip}_c(g) = 4|\mathcal{S}_r|^{\frac{1}{2}}$.

We derive an upper bound of $\|(\mathbf{H}_{\mathcal{S}_r \mathcal{S}_r} - \tilde{\mathbf{H}}_{\mathcal{S}_r \mathcal{S}_r})\|_F$ as follows. Consider a random variable

$$Z_{ij} := (\tilde{\mathbf{H}}_{\mathcal{S}_r \mathcal{S}_r})_{ij} = \frac{1}{m} \sum_{l=1}^m \mathcal{E}(\mathbf{x}_{\tilde{r}}^{(l)})_i \mathcal{E}(\mathbf{x}_{\tilde{r}}^{(l)})_j \in [-1/m, 1/m]$$

$$\mathbb{E}_{\mathbb{P}} Z_{ij} = (\mathbf{H}_{\mathcal{S}_r \mathcal{S}_r})_{ij}.$$

By Hoeffding's inequality, we observe

$$\text{Prob}(|(\tilde{\mathbf{H}}_{\mathcal{S}_r \mathcal{S}_r})_{ij} - (\mathbf{H}_{\mathcal{S}_r \mathcal{S}_r})_{ij}| \geq t) \leq 2 \exp\left(-\frac{mt^2}{2}\right),$$

for $t > 0$. Setting $t = \frac{t}{|\mathcal{S}_r|}$ for all $i, j \in \mathcal{S}_r$ and applying the union bound,

$$\text{Prob}(\|(\tilde{\mathbf{H}}_{\mathcal{S}_r \mathcal{S}_r}) - (\mathbf{H}_{\mathcal{S}_r \mathcal{S}_r})\|_F \geq t) \leq 2|\mathcal{S}_r|^2 \exp\left(-\frac{mt^2}{2|\mathcal{S}_r|^2}\right). \quad (3.8)$$

To conclude, with probability at least $1 - 2|\mathcal{S}_r|^2 \exp(-\frac{mt^2}{2|\mathcal{S}_r|^2})$, we have

$$\Lambda_{\min}(\mathbf{H}_{\mathcal{S}_r \mathcal{S}_r}^{\mathbb{Q}}) \geq \Lambda_{\min}(\mathbf{H}_{\mathcal{S}_r \mathcal{S}_r}) - 4\varepsilon|\mathcal{S}_r|^{\frac{1}{2}} - t.$$

□

Lemma 24. *If Assumption 19 and Assumption 20 hold, for any $\mathbb{Q} \in \mathcal{A}_\varepsilon^{W_p}(\tilde{\mathbb{P}}_m)$ and $\alpha \in (0, 1]$, with high probability,*

$$\|\mathbf{H}_{\mathcal{S}_r^c \mathcal{S}_r}^{\mathbb{Q}} (\mathbf{H}_{\mathcal{S}_r \mathcal{S}_r}^{\mathbb{Q}})^{-1}\|_{B,1,\infty} \leq 1 - \frac{\alpha}{2}.$$

Proof. We would like to obtain an upper bound for $\|\mathbf{H}_{\mathcal{S}_r^c \mathcal{S}_r}^{\mathbb{Q}} (\mathbf{H}_{\mathcal{S}_r \mathcal{S}_r}^{\mathbb{Q}})^{-1}\|_{B,1,\infty}$. We may write

$$\begin{aligned} \mathbf{H}_{\mathcal{S}_r^c \mathcal{S}_r}^{\mathbb{Q}} (\mathbf{H}_{\mathcal{S}_r \mathcal{S}_r}^{\mathbb{Q}})^{-1} &= \mathbf{H}_{\mathcal{S}_r^c \mathcal{S}_r} [\mathbf{H}_{\mathcal{S}_r \mathcal{S}_r}^{\mathbb{Q}}]^{-1} - (\mathbf{H}_{\mathcal{S}_r \mathcal{S}_r})^{-1} \\ &\quad + [\mathbf{H}_{\mathcal{S}_r^c \mathcal{S}_r}^{\mathbb{Q}} - \mathbf{H}_{\mathcal{S}_r^c \mathcal{S}_r}] (\mathbf{H}_{\mathcal{S}_r \mathcal{S}_r})^{-1} \\ &\quad + [\mathbf{H}_{\mathcal{S}_r^c \mathcal{S}_r}^{\mathbb{Q}} - \mathbf{H}_{\mathcal{S}_r^c \mathcal{S}_r}] [(\mathbf{H}_{\mathcal{S}_r \mathcal{S}_r}^{\mathbb{Q}})^{-1} - (\mathbf{H}_{\mathcal{S}_r \mathcal{S}_r})^{-1}] \\ &\quad + \mathbf{H}_{\mathcal{S}_r^c \mathcal{S}_r} (\mathbf{H}_{\mathcal{S}_r \mathcal{S}_r})^{-1} \end{aligned}$$

$$\implies$$

$$\begin{aligned} \|\mathbf{H}_{\mathcal{S}_r^c \mathcal{S}_r}^{\mathbb{Q}} (\mathbf{H}_{\mathcal{S}_r \mathcal{S}_r}^{\mathbb{Q}})^{-1}\|_{B,1,\infty} &\leq \|\mathbf{H}_{\mathcal{S}_r^c \mathcal{S}_r} [(\mathbf{H}_{\mathcal{S}_r \mathcal{S}_r}^{\mathbb{Q}})^{-1} - (\mathbf{H}_{\mathcal{S}_r \mathcal{S}_r})^{-1}]\|_{B,1,\infty} \\ &\quad + \|[\mathbf{H}_{\mathcal{S}_r^c \mathcal{S}_r}^{\mathbb{Q}} - \mathbf{H}_{\mathcal{S}_r^c \mathcal{S}_r}] (\mathbf{H}_{\mathcal{S}_r \mathcal{S}_r})^{-1}\|_{B,1,\infty} \\ &\quad + \|[\mathbf{H}_{\mathcal{S}_r^c \mathcal{S}_r}^{\mathbb{Q}} - \mathbf{H}_{\mathcal{S}_r^c \mathcal{S}_r}] [(\mathbf{H}_{\mathcal{S}_r \mathcal{S}_r}^{\mathbb{Q}})^{-1} - (\mathbf{H}_{\mathcal{S}_r \mathcal{S}_r})^{-1}]\|_{B,1,\infty} \\ &\quad + \|\mathbf{H}_{\mathcal{S}_r^c \mathcal{S}_r} (\mathbf{H}_{\mathcal{S}_r \mathcal{S}_r})^{-1}\|_{B,1,\infty}. \end{aligned}$$

By Hoeffding's inequality,

$$\text{Prob}(|(\tilde{\mathbf{H}}_{\mathcal{S}_r^c \mathcal{S}_r})_{ij} - (\mathbf{H}_{\mathcal{S}_r^c \mathcal{S}_r})_{ij}| \geq t) \leq 2 \exp\left(-\frac{mt^2}{2}\right),$$

for $t > 0$. Taking $t = \frac{t}{\rho_i |\mathcal{S}_r|}$ and applying the union bound over $i \in \mathbf{Co}_r$, we observe that

$$\begin{aligned} \text{Prob}(\|\tilde{\mathbf{H}}_{\mathcal{S}_r^c \mathcal{S}_r} - \mathbf{H}_{\mathcal{S}_r^c \mathcal{S}_r}\|_{B,1,\infty} \geq t) &\leq \sum_{i \in \mathbf{Co}_r} 2\rho_i |\mathcal{S}_r| \exp\left(-\frac{mt^2}{2\rho_i^2 |\mathcal{S}_r|^2}\right) \\ &\leq 2|\mathcal{S}_r^c| |\mathcal{S}_r| \exp\left(-\frac{mt^2}{2\rho_{\max}^2 |\mathcal{S}_r|^2}\right). \end{aligned}$$

Similarly, taking $t = \frac{t}{|\mathcal{S}_r|}$,

$$\begin{aligned} \text{Prob}(\|\tilde{\mathbf{H}}_{\mathcal{S}_r \mathcal{S}_r} - \mathbf{H}_{\mathcal{S}_r \mathcal{S}_r}\|_{\infty,\infty} \geq t) &\leq \sum_{i \in \mathcal{S}_r} \sum_{j \in \mathcal{S}_r} 2 \exp\left(-\frac{mt^2}{2|\mathcal{S}_r|^2}\right) \\ &= 2|\mathcal{S}_r|^2 \exp\left(-\frac{mt^2}{2|\mathcal{S}_r|^2}\right). \end{aligned}$$

In order to bound $\|\mathbf{H}_{\mathcal{S}_r^c \mathcal{S}_r}^{\mathbb{Q}} - \mathbf{H}_{\mathcal{S}_r^c \mathcal{S}_r}\|_{B,1,\infty}$, for $\mathbb{Q} \neq \tilde{\mathbb{P}}$, consider

$$\begin{aligned} &\|\mathbf{H}_{\mathcal{S}_r^c \mathcal{S}_r}^{\mathbb{Q}} - \tilde{\mathbf{H}}_{\mathcal{S}_r^c \mathcal{S}_r}\|_{B,1,\infty} \\ &\leq \|\mathbf{H}_{\mathcal{S}_r^c \mathcal{S}_r}^{\mathbb{Q}}\|_{B,1,\infty} + \|\tilde{\mathbf{H}}_{\mathcal{S}_r^c \mathcal{S}_r}\|_{B,1,\infty} \\ &\leq \mathbb{E}_{\mathbb{Q}} \|\mathcal{E}(\mathbf{X}_{\bar{r}})_{\mathcal{S}_r^c} \mathcal{E}(\mathbf{X}_{\bar{r}})_{\mathcal{S}_r}^{\mathbf{T}}\|_{B,1,\infty} + \mathbb{E}_{\tilde{\mathbb{P}}_m} \|\mathcal{E}(\mathbf{X}_{\bar{r}})_{\mathcal{S}_r^c} \mathcal{E}(\mathbf{X}_{\bar{r}})_{\mathcal{S}_r}^{\mathbf{T}}\|_{B,1,\infty} \\ &= \sup_{\tilde{\mathbb{P}}'_m, \mathbb{Q}' \in \mathcal{A}_{\varepsilon}^{Wp}(\tilde{\mathbb{P}}'_m)} |\mathbb{E}_{\mathbb{Q}'} \xi_1 \|\mathcal{E}(\mathbf{X}_{\bar{r}})_{\mathcal{S}_r^c} \mathcal{E}(\mathbf{X}_{\bar{r}})_{\mathcal{S}_r}^{\mathbf{T}}\|_{B,1,\infty} - \mathbb{E}_{\tilde{\mathbb{P}}'_m} \xi_2 \|\mathcal{E}(\mathbf{X}_{\bar{r}})_{\mathcal{S}_r^c} \mathcal{E}(\mathbf{X}_{\bar{r}})_{\mathcal{S}_r}^{\mathbf{T}}\|_{B,1,\infty}|, \end{aligned}$$

where \mathbb{Q}' and $\tilde{\mathbb{P}}'_m$ are probability measures on $\mathcal{X} \times \Xi$ with $\Xi = \{-1, +1\}$ and identical marginals as \mathbb{Q} and $\tilde{\mathbb{P}}_m$ respectively. We assume that $\mathbb{Q} \neq \tilde{\mathbb{P}}$ because otherwise $\|\mathbf{H}_{\mathcal{S}_r^c \mathcal{S}_r}^{\mathbb{Q}} - \tilde{\mathbf{H}}_{\mathcal{S}_r^c \mathcal{S}_r}\|_{B,1,\infty} = 0$ holds trivially. In this way, the equality is always achieved by some $\mathbb{Q}', \tilde{\mathbb{P}}'_m$, i.e., setting $\mathbb{Q}'(\mathcal{X}, \xi = 1) = 1$ and $\tilde{\mathbb{P}}'_m(\mathcal{X}, \xi = -1) = 1$.

Define the transport cost function in the ambiguity set $\mathcal{A}_\varepsilon^{W_p}(\tilde{\mathbb{P}}'_m)$ to be $c'((\mathbf{X}_1, \xi_1), (\mathbf{X}_2, \xi_2)) := \|\mathcal{E}(\mathbf{X}_1) - \mathcal{E}(\mathbf{X}_2)\|_1$ with zero cost for ξ . Let $g(\mathbf{X}, \xi) := \xi_1 \|\mathcal{E}(\mathbf{X}_{\bar{r}})_{\mathcal{S}_r^c} \mathcal{E}(\mathbf{X}_{\bar{r}})_{\mathcal{S}_r}^\top\|_{B,1,\infty}$. Consider the Lipschitz constants of g :

$$\begin{aligned} \text{lip}_{c'}(g) &\leq \sup_{\mathbf{X}, \xi, \mathbf{X}', \xi'} \frac{|g(\mathbf{X}, \xi) - g(\mathbf{X}', \xi')|}{c'((\mathbf{X}, \xi), (\mathbf{X}', \xi'))} \\ &\leq \sup_{\mathbf{X}, \mathbf{X}'} \frac{\|\mathcal{E}(\mathbf{X}_{\bar{r}})_{\mathcal{S}_r^c} \mathcal{E}(\mathbf{X}_{\bar{r}})_{\mathcal{S}_r}^\top\|_{B,1,\infty} + \|\mathcal{E}(\mathbf{X}'_{\bar{r}})_{\mathcal{S}_r^c} \mathcal{E}(\mathbf{X}'_{\bar{r}})_{\mathcal{S}_r}^\top\|_{B,1,\infty}}{\|\mathcal{E}(\mathbf{X}) - \mathcal{E}(\mathbf{X}')\|_1} \\ &\leq 2\rho_{\max} |\mathcal{S}_r|. \end{aligned} \tag{3.9}$$

Therefore, by the Kantorovich-Rubinstein theorem (Kantorovich and Rubinshtein, 1958),

$$\begin{aligned} \|\mathbf{H}_{\mathcal{S}_r^c \mathcal{S}_r}^{\mathbb{Q}} - \tilde{\mathbf{H}}_{\mathcal{S}_r^c \mathcal{S}_r}\|_{B,1,\infty} &\leq \sup_{\tilde{\mathbb{P}}'_m, \mathbb{Q}' \in \mathcal{A}_\varepsilon^{W_p}(\tilde{\mathbb{P}}'_m)} |\mathbb{E}_{\mathbb{Q}'} g(\mathbf{X}, \xi) - \mathbb{E}_{\tilde{\mathbb{P}}'_m} g(\mathbf{X}, \xi)| \\ &\leq \sup_{\tilde{\mathbb{P}}'_m, \mathbb{Q}' \in \mathcal{A}_\varepsilon^{W_p}(\tilde{\mathbb{P}}'_m)} \text{lip}_{c'}(g) |\mathbb{E}_{\mathbb{Q}'} g(\mathbf{X}, \xi) / \text{lip}_{c'}(g) - \mathbb{E}_{\tilde{\mathbb{P}}'_m} g(\mathbf{X}, \xi) / \text{lip}_{c'}(g)| \\ &\leq \sup_{\tilde{\mathbb{P}}'_m, \mathbb{Q}' \in \mathcal{A}_\varepsilon^{W_p}(\tilde{\mathbb{P}}'_m)} \text{lip}_{c'}(g) W_1(\mathbb{Q}', \tilde{\mathbb{P}}'_m) \\ &\leq \text{lip}_{c'}(g) \varepsilon \\ &\leq 2\varepsilon \rho_{\max} |\mathcal{S}_r|. \end{aligned}$$

Similarly,

$$\|\mathbf{H}_{\mathcal{S}_r\mathcal{S}_r}^{\mathbb{Q}} - \tilde{\mathbf{H}}_{\mathcal{S}_r\mathcal{S}_r}\|_{\infty,\infty} \leq 2\varepsilon|\mathcal{S}_r|.$$

Based on the above two inequalities, we find that

$$\begin{aligned} \|\mathbf{H}_{\mathcal{S}_r^c\mathcal{S}_r}^{\mathbb{Q}} - \mathbf{H}_{\mathcal{S}_r^c\mathcal{S}_r}\|_{B,1,\infty} &\leq \|\mathbf{H}_{\mathcal{S}_r^c\mathcal{S}_r}^{\mathbb{Q}} - \tilde{\mathbf{H}}_{\mathcal{S}_r^c\mathcal{S}_r}\|_{B,1,\infty} + \|\tilde{\mathbf{H}}_{\mathcal{S}_r^c\mathcal{S}_r} - \mathbf{H}_{\mathcal{S}_r^c\mathcal{S}_r}\|_{B,1,\infty} \\ &\leq 2\varepsilon\rho_{\max}|\mathcal{S}_r| + t, \end{aligned} \quad (3.10)$$

with probability at least $1 - 2|\mathcal{S}_r^c||\mathcal{S}_r| \exp(-\frac{mt^2}{2\rho_{\max}^2|\mathcal{S}_r|^2})$, and

$$\|\mathbf{H}_{\mathcal{S}_r\mathcal{S}_r}^{\mathbb{Q}} - \mathbf{H}_{\mathcal{S}_r\mathcal{S}_r}\|_{\infty,\infty} \leq 2\varepsilon|\mathcal{S}_r| + t, \quad (3.11)$$

with probability at least $1 - 2|\mathcal{S}_r|^2 \exp(-\frac{mt^2}{2|\mathcal{S}_r|^2})$.

Based on Equation 3.8, we also have

$$\|[\mathbf{H}_{\mathcal{S}_r\mathcal{S}_r} - \mathbf{H}_{\mathcal{S}_r\mathcal{S}_r}^{\mathbb{Q}}]\|_F \leq 2\varepsilon|\mathcal{S}_r| + t, \quad (3.12)$$

with probability at least $1 - 2|\mathcal{S}_r|^2 \exp(-\frac{mt^2}{2|\mathcal{S}_r|^2})$.

Next we look at the upper bound on the difference between the inverses of $\mathbf{H}_{\mathcal{S}_r\mathcal{S}_r}^{\mathbb{Q}}$ and $\mathbf{H}_{\mathcal{S}_r\mathcal{S}_r}$.

Observe that

$$\begin{aligned}
\|(\mathbf{H}_{\mathcal{S}_r\mathcal{S}_r}^{\mathbb{Q}})^{-1} - (\mathbf{H}_{\mathcal{S}_r\mathcal{S}_r})^{-1}\|_{\infty,\infty} &= \|(\mathbf{H}_{\mathcal{S}_r\mathcal{S}_r})^{-1}[\mathbf{H}_{\mathcal{S}_r\mathcal{S}_r} - \mathbf{H}_{\mathcal{S}_r\mathcal{S}_r}^{\mathbb{Q}}](\mathbf{H}_{\mathcal{S}_r\mathcal{S}_r}^{\mathbb{Q}})^{-1}\|_{\infty,\infty} \\
&\leq \sqrt{|\mathcal{S}_r|} \|(\mathbf{H}_{\mathcal{S}_r\mathcal{S}_r})^{-1}[\mathbf{H}_{\mathcal{S}_r\mathcal{S}_r} - \mathbf{H}_{\mathcal{S}_r\mathcal{S}_r}^{\mathbb{Q}}](\mathbf{H}_{\mathcal{S}_r\mathcal{S}_r}^{\mathbb{Q}})^{-1}\|_{2,2} \\
&\leq \sqrt{|\mathcal{S}_r|} \|(\mathbf{H}_{\mathcal{S}_r\mathcal{S}_r})^{-1}\|_{2,2} \|[\mathbf{H}_{\mathcal{S}_r\mathcal{S}_r} - \mathbf{H}_{\mathcal{S}_r\mathcal{S}_r}^{\mathbb{Q}}]\|_{2,2} \|(\mathbf{H}_{\mathcal{S}_r\mathcal{S}_r}^{\mathbb{Q}})^{-1}\|_{2,2} \\
&\leq \sqrt{\frac{|\mathcal{S}_r|}{\Lambda_{\min}(\mathbf{H}_{\mathcal{S}_r\mathcal{S}_r})}} \|[\mathbf{H}_{\mathcal{S}_r\mathcal{S}_r} - \mathbf{H}_{\mathcal{S}_r\mathcal{S}_r}^{\mathbb{Q}}]\|_{2,2} \|(\mathbf{H}_{\mathcal{S}_r\mathcal{S}_r}^{\mathbb{Q}})^{-1}\|_{2,2}.
\end{aligned}$$

According to Lemma 23, with probability at least $1 - 2|\mathcal{S}_r|^2 \exp(-\frac{mt^2}{2|\mathcal{S}_r|^2})$, we have

$$\Lambda_{\min}(\mathbf{H}_{\mathcal{S}_r\mathcal{S}_r}^{\mathbb{Q}}) \geq \Lambda_{\min}(\mathbf{H}_{\mathcal{S}_r\mathcal{S}_r}) - 4\varepsilon|\mathcal{S}_r|^{\frac{1}{2}} - t.$$

Let $t = \frac{1}{2}\Lambda_{\min}(\mathbf{H}_{\mathcal{S}_r\mathcal{S}_r})$ and $\varepsilon \leq \frac{\Lambda_{\min}(\mathbf{H}_{\mathcal{S}_r\mathcal{S}_r})}{16|\mathcal{S}_r|^{\frac{1}{2}}}$. We get that, with probability at least $1 - 2|\mathcal{S}_r|^2 \exp(-\frac{m(\Lambda_{\min}(\mathbf{H}_{\mathcal{S}_r\mathcal{S}_r}))^2}{8|\mathcal{S}_r|^2})$,

$$\begin{aligned}
\Lambda_{\min}(\mathbf{H}_{\mathcal{S}_r\mathcal{S}_r}^{\mathbb{Q}}) &\geq \frac{1}{4}\Lambda_{\min}(\mathbf{H}_{\mathcal{S}_r\mathcal{S}_r}) \\
\implies \|(\mathbf{H}_{\mathcal{S}_r\mathcal{S}_r}^{\mathbb{Q}})^{-1}\|_{2,2} &\leq \sqrt{\frac{4}{\Lambda_{\min}(\mathbf{H}_{\mathcal{S}_r\mathcal{S}_r})}}.
\end{aligned} \tag{3.13}$$

Set $t = \frac{t\Lambda_{\min}(\mathbf{H}_{\mathcal{S}_r\mathcal{S}_r})}{4\sqrt{|\mathcal{S}_r|}}$ and $\varepsilon \leq \frac{t\Lambda_{\min}(\mathbf{H}_{\mathcal{S}_r\mathcal{S}_r})}{8|\mathcal{S}_r|\sqrt{|\mathcal{S}_r|}}$ in Equation 3.12, we get that, with probability at least $1 - 2|\mathcal{S}_r|^2 \exp\left(-\frac{mt^2(\Lambda_{\min}(\mathbf{H}_{\mathcal{S}_r\mathcal{S}_r}))^2}{32|\mathcal{S}_r|^3}\right)$,

$$\|[\mathbf{H}_{\mathcal{S}_r\mathcal{S}_r} - \mathbf{H}_{\mathcal{S}_r\mathcal{S}_r}^{\mathbb{Q}}]\|_{2,2} \leq \|[\mathbf{H}_{\mathcal{S}_r\mathcal{S}_r} - \mathbf{H}_{\mathcal{S}_r\mathcal{S}_r}^{\mathbb{Q}}]\|_F \leq \frac{t\Lambda_{\min}(\mathbf{H}_{\mathcal{S}_r\mathcal{S}_r})}{2\sqrt{|\mathcal{S}_r|}}.$$

Therefore, with probability at least

$$1 - 2|\mathcal{S}_r|^2 \exp\left(-\frac{mt^2(\Lambda_{\min}(\mathbf{H}_{\mathcal{S}_r\mathcal{S}_r}))^2}{32|\mathcal{S}_r|^3}\right) - 2|\mathcal{S}_r|^2 \exp\left(-\frac{m(\Lambda_{\min}(\mathbf{H}_{\mathcal{S}_r\mathcal{S}_r}))^2}{8|\mathcal{S}_r|^2}\right)$$

and $\varepsilon \leq \min\left(\frac{t\Lambda_{\min}(\mathbf{H}_{\mathcal{S}_r\mathcal{S}_r})}{8|\mathcal{S}_r|\sqrt{|\mathcal{S}_r|}}, \frac{\Lambda_{\min}(\mathbf{H}_{\mathcal{S}_r\mathcal{S}_r})}{16|\mathcal{S}_r|^{\frac{1}{2}}}\right)$,

$$\|(\mathbf{H}_{\mathcal{S}_r\mathcal{S}_r}^{\mathbb{Q}})^{-1} - (\mathbf{H}_{\mathcal{S}_r\mathcal{S}_r})^{-1}\|_{\infty,\infty} \leq t. \quad (3.14)$$

Now we are ready to obtain upper bounds for the four terms recapitulated here:

$$\begin{aligned} \|\mathbf{H}_{\mathcal{S}_r^c\mathcal{S}_r}^{\mathbb{Q}}(\mathbf{H}_{\mathcal{S}_r\mathcal{S}_r}^{\mathbb{Q}})^{-1}\|_{B,1,\infty} &\leq \|\mathbf{H}_{\mathcal{S}_r^c\mathcal{S}_r}[(\mathbf{H}_{\mathcal{S}_r\mathcal{S}_r}^{\mathbb{Q}})^{-1} - (\mathbf{H}_{\mathcal{S}_r\mathcal{S}_r})^{-1}]\|_{B,1,\infty} \\ &\quad + \|[\mathbf{H}_{\mathcal{S}_r^c\mathcal{S}_r}^{\mathbb{Q}} - \mathbf{H}_{\mathcal{S}_r^c\mathcal{S}_r}](\mathbf{H}_{\mathcal{S}_r\mathcal{S}_r})^{-1}\|_{B,1,\infty} \\ &\quad + \|[\mathbf{H}_{\mathcal{S}_r^c\mathcal{S}_r}^{\mathbb{Q}} - \mathbf{H}_{\mathcal{S}_r^c\mathcal{S}_r}][(\mathbf{H}_{\mathcal{S}_r\mathcal{S}_r}^{\mathbb{Q}})^{-1} - (\mathbf{H}_{\mathcal{S}_r\mathcal{S}_r})^{-1}]\|_{B,1,\infty} \\ &\quad + \|\mathbf{H}_{\mathcal{S}_r^c\mathcal{S}_r}(\mathbf{H}_{\mathcal{S}_r\mathcal{S}_r})^{-1}\|_{B,1,\infty}. \end{aligned}$$

We derive the bounds separately.

For the first term, based on Assumption 20, consider

$$\begin{aligned}
& \| \mathbf{H}_{\mathcal{S}_r^c \mathcal{S}_r} [(\mathbf{H}_{\mathcal{S}_r \mathcal{S}_r}^{\mathbb{Q}})^{-1} - (\mathbf{H}_{\mathcal{S}_r \mathcal{S}_r})^{-1}] \|_{B,1,\infty} \\
&= \| \mathbf{H}_{\mathcal{S}_r^c \mathcal{S}_r} (\mathbf{H}_{\mathcal{S}_r \mathcal{S}_r})^{-1} [\mathbf{H}_{\mathcal{S}_r \mathcal{S}_r} - \mathbf{H}_{\mathcal{S}_r \mathcal{S}_r}^{\mathbb{Q}}] (\mathbf{H}_{\mathcal{S}_r \mathcal{S}_r}^{\mathbb{Q}})^{-1} \|_{B,1,\infty} \\
&\leq \| \mathbf{H}_{\mathcal{S}_r^c \mathcal{S}_r} (\mathbf{H}_{\mathcal{S}_r \mathcal{S}_r})^{-1} \|_{B,1,\infty} \| \mathbf{H}_{\mathcal{S}_r \mathcal{S}_r} - \mathbf{H}_{\mathcal{S}_r \mathcal{S}_r}^{\mathbb{Q}} \|_{\infty,\infty} \| (\mathbf{H}_{\mathcal{S}_r \mathcal{S}_r}^{\mathbb{Q}})^{-1} \|_{\infty,\infty} \\
&\leq (1 - \alpha) \| \mathbf{H}_{\mathcal{S}_r \mathcal{S}_r} - \mathbf{H}_{\mathcal{S}_r \mathcal{S}_r}^{\mathbb{Q}} \|_{\infty,\infty} \sqrt{|\mathcal{S}_r|} \| (\mathbf{H}_{\mathcal{S}_r \mathcal{S}_r}^{\mathbb{Q}})^{-1} \|_{2,2}.
\end{aligned}$$

Taking $t = \frac{\alpha}{24(1-\alpha)} \sqrt{\frac{\Lambda_{\min}(\mathbf{H}_{\mathcal{S}_r \mathcal{S}_r})}{|\mathcal{S}_r|}}$ and $\varepsilon \leq \frac{\alpha}{48(1-\alpha)|\mathcal{S}_r|} \sqrt{\frac{\Lambda_{\min}(\mathbf{H}_{\mathcal{S}_r \mathcal{S}_r})}{|\mathcal{S}_r|}}$ in Equation 3.11 and adopting Equation 3.13, we conclude that, with probability at least $1 - 2|\mathcal{S}_r|^2 \exp(-\frac{m\alpha^2 \Lambda_{\min}(\mathbf{H}_{\mathcal{S}_r \mathcal{S}_r})}{1152(1-\alpha)^2 |\mathcal{S}_r|^3}) - 2|\mathcal{S}_r|^2 \exp(-\frac{m(\Lambda_{\min}(\mathbf{H}_{\mathcal{S}_r \mathcal{S}_r}))^2}{8|\mathcal{S}_r|^2})$ and $\varepsilon \leq \min(\frac{\alpha}{48(1-\alpha)|\mathcal{S}_r|} \sqrt{\frac{\Lambda_{\min}(\mathbf{H}_{\mathcal{S}_r \mathcal{S}_r})}{|\mathcal{S}_r|}}, \frac{\Lambda_{\min}(\mathbf{H}_{\mathcal{S}_r \mathcal{S}_r})}{16|\mathcal{S}_r|^{\frac{1}{2}}})$,

$$\| \mathbf{H}_{\mathcal{S}_r^c \mathcal{S}_r} [(\mathbf{H}_{\mathcal{S}_r \mathcal{S}_r}^{\mathbb{Q}})^{-1} - (\mathbf{H}_{\mathcal{S}_r \mathcal{S}_r})^{-1}] \|_{B,1,\infty} \leq \frac{\alpha}{6}.$$

For the second term, we rewrite it as

$$\begin{aligned}
& \| [\mathbf{H}_{\mathcal{S}_r^c \mathcal{S}_r}^{\mathbb{Q}} - \mathbf{H}_{\mathcal{S}_r^c \mathcal{S}_r}] (\mathbf{H}_{\mathcal{S}_r \mathcal{S}_r})^{-1} \|_{B,1,\infty} \\
&\leq \| [\mathbf{H}_{\mathcal{S}_r^c \mathcal{S}_r}^{\mathbb{Q}} - \mathbf{H}_{\mathcal{S}_r^c \mathcal{S}_r}] \|_{B,1,\infty} \| (\mathbf{H}_{\mathcal{S}_r \mathcal{S}_r})^{-1} \|_{\infty,\infty} \\
&\leq \| [\mathbf{H}_{\mathcal{S}_r^c \mathcal{S}_r}^{\mathbb{Q}} - \mathbf{H}_{\mathcal{S}_r^c \mathcal{S}_r}] \|_{B,1,\infty} \sqrt{|\mathcal{S}_r|} \| (\mathbf{H}_{\mathcal{S}_r \mathcal{S}_r})^{-1} \|_{2,2} \\
&\leq \| [\mathbf{H}_{\mathcal{S}_r^c \mathcal{S}_r}^{\mathbb{Q}} - \mathbf{H}_{\mathcal{S}_r^c \mathcal{S}_r}] \|_{B,1,\infty} \sqrt{\frac{|\mathcal{S}_r|}{\Lambda_{\min}(\mathbf{H}_{\mathcal{S}_r \mathcal{S}_r})}}.
\end{aligned}$$

Using Equation 3.10 by setting $t = \frac{\alpha}{12} \sqrt{\frac{\Lambda_{\min}(\mathbf{H}_{\mathcal{S}_r \mathcal{S}_r})}{|\mathcal{S}_r|}}$ and $\varepsilon \leq \frac{\alpha}{24\rho_{\max}|\mathcal{S}_r|} \sqrt{\frac{\Lambda_{\min}(\mathbf{H}_{\mathcal{S}_r \mathcal{S}_r})}{|\mathcal{S}_r|}}$, we have, with probability at least $1 - 2|\mathcal{S}_r^c||\mathcal{S}_r| \exp(-\frac{m\alpha^2\Lambda_{\min}(\mathbf{H}_{\mathcal{S}_r \mathcal{S}_r})}{288\rho_{\max}^2|\mathcal{S}_r|^3})$ and $\varepsilon \leq \frac{\alpha}{24\rho_{\max}|\mathcal{S}_r|} \sqrt{\frac{\Lambda_{\min}(\mathbf{H}_{\mathcal{S}_r \mathcal{S}_r})}{|\mathcal{S}_r|}}$,

$$\|[\mathbf{H}_{\mathcal{S}_r^c \mathcal{S}_r}^{\mathbb{Q}} - \mathbf{H}_{\mathcal{S}_r^c \mathcal{S}_r}](\mathbf{H}_{\mathcal{S}_r \mathcal{S}_r})^{-1}\|_{B,1,\infty} \leq \frac{\alpha}{6}.$$

For the third term, we obtain the upper bound

$$\begin{aligned} & \|[\mathbf{H}_{\mathcal{S}_r^c \mathcal{S}_r}^{\mathbb{Q}} - \mathbf{H}_{\mathcal{S}_r^c \mathcal{S}_r}][(\mathbf{H}_{\mathcal{S}_r \mathcal{S}_r}^{\mathbb{Q}})^{-1} - (\mathbf{H}_{\mathcal{S}_r \mathcal{S}_r})^{-1}]\|_{B,1,\infty} \\ & \leq \|[\mathbf{H}_{\mathcal{S}_r^c \mathcal{S}_r}^{\mathbb{Q}} - \mathbf{H}_{\mathcal{S}_r^c \mathcal{S}_r}]\|_{B,1,\infty} \|[(\mathbf{H}_{\mathcal{S}_r \mathcal{S}_r}^{\mathbb{Q}})^{-1} - (\mathbf{H}_{\mathcal{S}_r \mathcal{S}_r})^{-1}]\|_{\infty,\infty}. \end{aligned}$$

Taking $t = \sqrt{\frac{\alpha}{6}}$ in Equation 3.14. Taking $t = \frac{1}{2}\sqrt{\frac{\alpha}{6}}$ and $2\varepsilon\rho_{\max}|\mathcal{S}_r| \leq \frac{1}{2}\sqrt{\frac{\alpha}{6}}$ in Equation 3.10.

We establish the upper bound that, with probability at least $1 - 2|\mathcal{S}_r^c||\mathcal{S}_r| \exp(-\frac{m\alpha}{48\rho_{\max}^2|\mathcal{S}_r|^2}) - 2|\mathcal{S}_r|^2 \exp(-\frac{m\alpha(\Lambda_{\min}(\mathbf{H}_{\mathcal{S}_r \mathcal{S}_r}))^2}{192|\mathcal{S}_r|^3}) - 2|\mathcal{S}_r|^2 \exp(-\frac{m(\Lambda_{\min}(\mathbf{H}_{\mathcal{S}_r \mathcal{S}_r}))^2}{8|\mathcal{S}_r|^2})$ and

$$\varepsilon \leq \min\left(\frac{1}{4\rho_{\max}|\mathcal{S}_r|} \sqrt{\frac{\alpha}{6}}, \frac{\Lambda_{\min}(\mathbf{H}_{\mathcal{S}_r \mathcal{S}_r})}{8|\mathcal{S}_r|} \sqrt{\frac{\alpha}{6|\mathcal{S}_r|}}, \frac{\Lambda_{\min}(\mathbf{H}_{\mathcal{S}_r \mathcal{S}_r})}{16|\mathcal{S}_r|^{\frac{1}{2}}}\right),$$

we have

$$\|[\mathbf{H}_{\mathcal{S}_r^c \mathcal{S}_r}^{\mathbb{Q}} - \mathbf{H}_{\mathcal{S}_r^c \mathcal{S}_r}][(\mathbf{H}_{\mathcal{S}_r \mathcal{S}_r}^{\mathbb{Q}})^{-1} - (\mathbf{H}_{\mathcal{S}_r \mathcal{S}_r})^{-1}]\|_{B,1,\infty} \leq \frac{\alpha}{6}.$$

For the fourth term, in accordance with Assumption 20,

$$\|\mathbf{H}_{\mathcal{S}_r^c \mathcal{S}_r}(\mathbf{H}_{\mathcal{S}_r \mathcal{S}_r})^{-1}\|_{B,1,\infty} \leq 1 - \alpha.$$

In conclusion, we have shown that, with probability at least $1 - 2|\mathcal{S}_r|^2 \exp(-\frac{m\alpha^2 \Lambda_{\min}(\mathbf{H}_{\mathcal{S}_r \mathcal{S}_r})}{1152(1-\alpha)^2 |\mathcal{S}_r|^3}) - 2|\mathcal{S}_r|^2 \exp(-\frac{m(\Lambda_{\min}(\mathbf{H}_{\mathcal{S}_r \mathcal{S}_r}))^2}{8|\mathcal{S}_r|^2}) - 2|\mathcal{S}_r^c||\mathcal{S}_r| \exp(-\frac{m\alpha^2 \Lambda_{\min}(\mathbf{H}_{\mathcal{S}_r \mathcal{S}_r})}{288\rho_{\max}^2 |\mathcal{S}_r|^3}) - 2|\mathcal{S}_r^c||\mathcal{S}_r| \exp(-\frac{m\alpha}{48\rho_{\max}^2 |\mathcal{S}_r|^2}) - 2|\mathcal{S}_r|^2 \exp(-\frac{m\alpha(\Lambda_{\min}(\mathbf{H}_{\mathcal{S}_r \mathcal{S}_r}))^2}{192|\mathcal{S}_r|^3}) - 2|\mathcal{S}_r|^2 \exp(-\frac{m(\Lambda_{\min}(\mathbf{H}_{\mathcal{S}_r \mathcal{S}_r}))^2}{8|\mathcal{S}_r|^2})$ and

$$\varepsilon \leq \min\left(\frac{\alpha}{48(1-\alpha)|\mathcal{S}_r|} \sqrt{\frac{\Lambda_{\min}(\mathbf{H}_{\mathcal{S}_r \mathcal{S}_r})}{|\mathcal{S}_r|}}, \frac{\Lambda_{\min}(\mathbf{H}_{\mathcal{S}_r \mathcal{S}_r})}{16|\mathcal{S}_r|^{\frac{1}{2}}}, \frac{\alpha}{24\rho_{\max}|\mathcal{S}_r|} \sqrt{\frac{\Lambda_{\min}(\mathbf{H}_{\mathcal{S}_r \mathcal{S}_r})}{|\mathcal{S}_r|}}, \frac{1}{4\rho_{\max}|\mathcal{S}_r|} \sqrt{\frac{\alpha}{6}}, \frac{\Lambda_{\min}(\mathbf{H}_{\mathcal{S}_r \mathcal{S}_r})}{8|\mathcal{S}_r|} \sqrt{\frac{\alpha}{6|\mathcal{S}_r|}}, \frac{\Lambda_{\min}(\mathbf{H}_{\mathcal{S}_r \mathcal{S}_r})}{16|\mathcal{S}_r|^{\frac{1}{2}}}\right),$$

the mutual incoherence condition holds for any worst-case distribution:

$$\|\mathbf{H}_{\mathcal{S}_r^c \mathcal{S}_r}^{\mathbb{Q}}(\mathbf{H}_{\mathcal{S}_r \mathcal{S}_r}^{\mathbb{Q}})^{-1}\|_{B,1,\infty} \leq 1 - \frac{\alpha}{2}.$$

Simplifying the above expressions, with probability at least

$$1 - \mathcal{O}\left(\exp\left(-\frac{Cm}{\rho_{\max}^2 |\mathcal{S}_r|^3} + \log |\mathcal{S}_r^c| + \log |\mathcal{S}_r|\right)\right)$$

$$\text{and } \varepsilon \leq \frac{C}{\rho_{\max} |\mathcal{S}_r|^{3/2}},$$

$$\|\mathbf{H}_{\mathcal{S}_r^c \mathcal{S}_r}^{\mathbb{Q}}(\mathbf{H}_{\mathcal{S}_r \mathcal{S}_r}^{\mathbb{Q}})^{-1}\|_{B,1,\infty} \leq 1 - \frac{\alpha}{2},$$

where C only depends on α , $\Lambda_{\min}(\mathbf{H}_{\mathcal{S}_r \mathcal{S}_r})$. \square

Lemma 25. *If Assumption 17 holds, then for any $\mathbb{Q} \in \mathcal{A}_\varepsilon^{W_p}(\tilde{\mathbb{P}}_m)$ and $\alpha \in (0, 1]$, with probability at least $1 - |\mathcal{S}_r| \rho_r \exp(-\frac{m\mu^2}{2\sigma^2})$, $\varepsilon \leq \frac{\mu}{\sigma}$ and $\lambda_B^* > \frac{32\mu\sqrt{\rho_r}(1-\alpha/2)}{\alpha}$, we have*

$$\|\mathbb{E}_{\mathbb{Q}} \mathcal{E}(\mathbf{X}_{\bar{r}})_{\mathcal{S}_r} \mathbf{e}^\top\|_{2,\infty} \leq \frac{\lambda_B^* \alpha}{8(1-\alpha/2)}.$$

With probability at least $1 - |\mathcal{C}\mathcal{O}_r| \rho_r \exp(-\frac{m\mu^2}{2\sigma^2})$, $\varepsilon \leq \frac{\mu}{\sigma}$ and $\lambda_B^* > \frac{32\mu\sqrt{\rho_{\max}\rho_r}}{\alpha}$, we have

$$\|\mathbb{E}_{\mathbb{Q}} \mathcal{E}(\mathbf{X}_{\bar{r}})_{\mathcal{S}_r^c} \mathbf{e}^\top\|_{B,2,\infty} \leq \frac{\lambda_B^* \alpha}{8}.$$

Proof. We start with $\|\mathbb{E}_{\mathbb{Q}} \mathcal{E}(\mathbf{X}_{\bar{r}})_{\mathcal{S}_r} \mathbf{e}^\top\|_{2,\infty}$. After some algebraic manipulation, we find that

$$\begin{aligned} \|\mathbb{E}_{\mathbb{Q}} \mathcal{E}(\mathbf{X}_{\bar{r}})_{\mathcal{S}_r} \mathbf{e}^\top\|_{2,\infty} &\leq \max_{i \in \mathcal{S}_r} \|\mathbb{E}_{\mathbb{Q}} \mathcal{E}(\mathbf{X}_{\bar{r}})_i \mathbf{e}\|_2 \\ &\leq \max_{i \in \mathcal{S}_r} \sqrt{\rho_r} \max_{j \in \rho_r} |\mathbb{E}_{\mathbb{Q}} \mathcal{E}(\mathbf{X}_{\bar{r}})_i e_j| \\ &\leq \max_{i \in \mathcal{S}_r} \sqrt{\rho_r} \max_{j \in \rho_r} \mathbb{E}_{\mathbb{Q}} |\mathcal{E}(\mathbf{X}_{\bar{r}})_i e_j| \\ &\leq \max_{i \in \mathcal{S}_r} \sqrt{\rho_r} \max_{j \in \rho_r} \mathbb{E}_{\mathbb{Q}} |e_j|. \end{aligned}$$

Since $|e_j|$ is a bounded random variable according to Assumption 17, we apply Hoeffding's inequality to get

$$\text{Prob}(\mathbb{E}_{\tilde{\mathbb{P}}_m} |e_j| \geq \mu + t) \leq \exp(-\frac{mt^2}{2\sigma^2}).$$

Base on a similar argument as Equation 3.9, we can derive

$$\mathbb{E}_{\mathbb{Q}}|e_j| - \mathbb{E}_{\tilde{\mathbb{P}}_m}|e_j| \leq 2\varepsilon\sigma,$$

which leads to

$$\text{Prob}(\mathbb{E}_{\mathbb{Q}}|e_j| \geq 2\varepsilon\sigma + \mu + t) \leq \exp\left(-\frac{mt^2}{2\sigma^2}\right).$$

Taking the union bound over all $i \in \mathcal{S}_r$ and $j \in \rho_r$, we find that

$$\text{Prob}(\|\mathbb{E}_{\mathbb{Q}}\mathcal{E}(\mathbf{X}_{\bar{r}})_{\mathcal{S}_r}\mathbf{e}^{\top}\|_{2,\infty} \geq \sqrt{\rho_r}(2\varepsilon\sigma + \mu + t)) \leq |\mathcal{S}_r|\rho_r \exp\left(-\frac{mt^2}{2\sigma^2}\right).$$

Setting $t = \mu$ and $\varepsilon \leq \frac{\mu}{\sigma}$ while requiring $\lambda_B^* > \frac{32\mu\sqrt{\rho_r}(1-\alpha/2)}{\alpha}$. With probability at least $1 - |\mathcal{S}_r|\rho_r \exp\left(-\frac{m\mu^2}{2\sigma^2}\right)$, we have

$$\|\mathbb{E}_{\mathbb{Q}}\mathcal{E}(\mathbf{X}_{\bar{r}})_{\mathcal{S}_r}\mathbf{e}^{\top}\|_{2,\infty} \leq \frac{\lambda_B^*\alpha}{8(1-\alpha/2)}. \quad (3.15)$$

Then we consider $\|\mathbb{E}_{\mathbb{Q}}\mathcal{E}(\mathbf{X}_{\bar{r}})_{\mathcal{S}_r^c}\mathbf{e}^{\top}\|_{B,2,\infty}$:

$$\begin{aligned} \|\mathbb{E}_{\mathbb{Q}}\mathcal{E}(\mathbf{X}_{\bar{r}})_{\mathcal{S}_r^c}\mathbf{e}^{\top}\|_{B,2,\infty} &\leq \max_{i \in \mathbf{Co}_r} \|\mathbb{E}_{\mathbb{Q}}\mathcal{E}(X_i)\mathbf{e}^{\top}\|_{2,2} \\ &\leq \max_{i \in \mathbf{Co}_r} \sqrt{\rho_i\rho_r} \max_{j \in \rho_i, k \in \rho_r} |\mathbb{E}_{\mathbb{Q}}\mathcal{E}(X_i)_j e_k| \\ &\leq \max_{i \in \mathbf{Co}_r} \sqrt{\rho_i\rho_r} \max_{k \in \rho_r} \mathbb{E}_{\mathbb{Q}}|e_k|. \end{aligned}$$

Similarly, applying Hoeffding's inequality and the Kantorovich-Rubinstein theorem gives us

$$\text{Prob}(\|\mathbb{E}_{\mathbb{Q}}\mathcal{E}(\mathbf{X}_{\bar{r}})_{\mathcal{S}_r^c}\mathbf{e}^{\top}\|_{B,2,\infty} \geq \sqrt{\rho_{\max}\rho_r}(2\varepsilon\sigma + \mu + t)) \leq |\mathbf{Co}_r|\rho_r \exp\left(-\frac{mt^2}{2\sigma^2}\right).$$

Let $t = \mu$, $\varepsilon \leq \frac{\mu}{\sigma}$ and $\lambda_B^* > \frac{32\mu\sqrt{\rho_{\max}\rho_r}}{\alpha}$ hold, we have, with probability at least $1 - |\mathbf{Co}_r|\rho_r \exp\left(-\frac{m\mu^2}{2\sigma^2}\right)$,

$$\|\mathbb{E}_{\mathbb{Q}}\mathcal{E}(\mathbf{X}_{\bar{r}})_{\mathcal{S}_r^c}\mathbf{e}^{\top}\|_{B,2,\infty} \leq \frac{\lambda_B^*\alpha}{8}.$$

□

3.3.2.2 Main Results

The above lemmas illustrate that Assumption 19 and Assumption 20 hold in the finite-sample setting. Let the estimated skeleton, neighbor nodes and the complement be $\hat{\mathcal{G}} := (\mathcal{V}, \hat{\mathcal{E}}_{\text{skel}})$, $\hat{\mathbf{N}}_{\mathbf{e}_r}$ and $\hat{\mathbf{Co}}_r$, respectively. We derive the following guarantees for the proposed Wasserstein DRO estimator.

Theorem 26. *Given a Bayesian network $(\mathcal{G}, \mathbb{P})$ of n categorical random variables and its skeleton $\mathcal{G}_{\text{skel}} := (\mathcal{V}, \mathcal{E}_{\text{skel}})$. Assume that the condition $\|\mathbf{W}^*\|_{B,2,1} \leq \bar{B}$ holds for some $\bar{B} > 0$ associated with an optimal Lagrange multiplier $\lambda_B^* > 0$ for \mathbf{W}^* defined in Equation 3.1. Suppose that $\hat{\mathbf{W}}$ is a DRO risk minimizer of Equation 3.4 with a Wasserstein distance of order 1 and*

an ambiguity radius $\varepsilon = \varepsilon_0/m$ where m is the number of samples drawn i.i.d. from \mathbb{P} . Under Assumptions 17, 18, 19, 20, if the number of samples satisfies

$$m = \mathcal{O}\left(\frac{C(\varepsilon_0 + \log(n/\delta) + \log \rho_{[n]})\sigma^2 \rho_{\max}^4 \rho_{[n]}^3}{\min(\mu^2, 1)}\right),$$

where C only depends on α , Λ , and if the Lagrange multiplier satisfies

$$\frac{32\mu\rho_{\max}}{\alpha} < \lambda_B^* < \frac{\beta}{(\alpha/(4-2\alpha) + 2)\rho_{\max}\sqrt{\rho_{[n]}}}\sqrt{\frac{\Lambda}{4}},$$

then for any $\delta \in (0, 1]$, $r \in [n]$, with probability at least $1 - \delta$, the following properties hold:

- (a) The optimal estimator $\hat{\mathbf{W}}$ is unique.
- (b) All the non-neighbor nodes are excluded: $\mathbf{Co}_r \subseteq \hat{\mathbf{Co}}_r$.
- (c) All the neighbor nodes are identified: $\mathbf{Ne}_r \subseteq \hat{\mathbf{Ne}}_r$.
- (d) The true skeleton is successfully reconstructed: $\mathcal{G}_{\text{skel}} = \hat{\mathcal{G}}_{\text{skel}}$.

Proof. We prove the statements in this theorem in several steps. In order to prove (a) and (b), we will show that the DRO problem is strictly convex if true non-neighbors are known so that there is an optimal solution. Next we would like to demonstrate that this solution with a non-neighbor constraint is indeed unique for all the solutions without constraints. The proof for uniqueness comes with a conclusion that we do not accidentally include any edge between the current node and its non-neighbors. Next, to prove (c), we present a generalization bound for the DRO estimator in terms of its true risk, which leads to a ℓ_∞ bound of the difference

between the estimator $\hat{\mathbf{W}}$ and the true weight matrix \mathbf{W}^* . Combined with the assumption on the minimum weight, it implies that we include all the neighbor nodes successfully. Finally, by taking a union bound for all the nodes, we could conclude that the correct skeleton is recovered with high probability, which proves (d).

(i) Given the true non-neighbors, there is a unique solution.

We start with the Wasserstein DRO problem, which we recapitulate here for convenience:

$$\hat{\mathbf{W}} \in \arg \inf_{\mathbf{W}} \sup_{\mathbb{Q} \in \mathcal{A}_\varepsilon^{W_p}(\tilde{\mathbb{P}}_m)} \frac{1}{2} \mathbb{E}_{\mathbb{Q}} \|\mathcal{E}(X_r) - \mathbf{W}^\top \mathcal{E}(\mathbf{X}_{\bar{r}})\|_2^2.$$

The objective is convex because it is a supremum of convex functions.

For now, we assume that the non-neighbor nodes \mathbf{Co}_r are given. We can then explicitly restrict $\mathbf{W}_i = \mathbf{0}$ for all $i \in \mathbf{Co}_r$. The Hessian of $\mathbf{W}_{\mathcal{S}_r}$ is a block diagonal matrix reads

$$\nabla^2 R^{\mathbb{Q}}(\mathbf{W}_{\mathcal{S}_r}) = \begin{bmatrix} \mathbf{H}_{\mathcal{S}_r \mathcal{S}_r}^{\mathbb{Q}} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{H}_{\mathcal{S}_r \mathcal{S}_r}^{\mathbb{Q}} & \cdots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{H}_{\mathcal{S}_r \mathcal{S}_r}^{\mathbb{Q}} \end{bmatrix} \in \mathbb{R}^{\rho_r \rho_{\mathbf{Ne}_r} \times \rho_r \rho_{\mathbf{Ne}_r}},$$

where

$$\mathbf{H}^{\mathbb{Q}} := \mathbb{E}_{\mathbb{Q}}[\mathcal{E}(\mathbf{X}_{\bar{r}})\mathcal{E}(\mathbf{X}_{\bar{r}})^\top] \in \mathbb{R}^{\rho_{\bar{r}} \times \rho_{\bar{r}}}$$

is the covariance matrix of encodings of $\mathbf{X}_{\bar{r}}$ under some distribution $\mathbb{Q} \in \mathcal{A}_\varepsilon^{W_p}(\tilde{\mathbb{P}}_m)$.

Since $\mathbf{W}_{\mathcal{S}_r^c}$ is fixed to be zero and $\nabla^2 R^{\mathbb{Q}}(\mathbf{W}_{\mathcal{S}_r})$ is a block diagonal matrix, we focus on showing that $\mathbf{H}_{\mathcal{S}_r \mathcal{S}_r}^{\mathbb{Q}} > \mathbf{0}$.

We apply Lemma 23 to get the bound

$$\Lambda_{\min}(\mathbf{H}_{\mathcal{S}_r \mathcal{S}_r}^{\mathbb{Q}}) \geq \Lambda_{\min}(\mathbf{H}_{\mathcal{S}_r \mathcal{S}_r}) - 4\varepsilon |\mathcal{S}_r|^{\frac{1}{2}} - t,$$

with probability at least $1 - 2|\mathcal{S}_r|^2 \exp(-\frac{mt^2}{2|\mathcal{S}_r|^2})$. $\Lambda_{\min}(\mathbf{H}_{\mathcal{S}_r \mathcal{S}_r}) - 4\varepsilon |\mathcal{S}_r|^{\frac{1}{2}} - t > 0$ will guarantee that the DRO problem in Equation 3.4 has a unique solution when the $\mathbf{W}_i = \mathbf{0}$ is satisfied for non-neighbor nodes.

(ii) Given the true non-neighbors, the solution is optimal.

We would like to show that the solution to Equation 3.4 with true non-neighbor constraints is optimal. In this way, we do not recover any non-neighbor nodes in the skeleton. We adopt the primal-dual witness (PDW) (Wainwright, 2009) method to show optimality for the constrained unique solution.

Recall that we assume $\|\mathbf{W}\|_{B,2,1} \leq \bar{B}$. To begin with, we write the dual problem as

$$\hat{\mathbf{W}} \in \arg \inf_{\mathbf{W}} \sup_{\mathbb{Q} \in \mathcal{A}_{\varepsilon}^{Wp}(\hat{\mathbb{P}}_m), \|\mathbf{Z}\|_{B,2,\infty} \leq 1, \lambda_B \geq 0} \frac{1}{2} \mathbb{E}_{\mathbb{Q}} \|\mathcal{E}(X_r) - \mathbf{W}^{\top} \mathcal{E}(\mathbf{X}_{\bar{r}})\|_2^2 + \lambda_B (\langle \mathbf{Z}, \mathbf{W} \rangle - \bar{B}) \quad (3.16)$$

$$\text{s.t. } \forall i \in \mathbf{Co}_r \quad \mathbf{W}_i = \mathbf{0},$$

where λ_B is the Lagrange multiplier for the norm constraint on \mathbf{W} .

$\hat{\mathbf{W}}$ is optimal if and only if there exists $(\mathbf{Q}^*, \mathbf{Z}^*, \lambda_B^*)$ that satisfies the KKT condition:

$$\mathbb{E}_{\mathbf{Q}^*} \mathcal{E}(\mathbf{X}_{\bar{r}}) \mathcal{E}(\mathbf{X}_{\bar{r}})^\top \hat{\mathbf{W}} - \mathbb{E}_{\mathbf{Q}^*} \mathcal{E}(\mathbf{X}_{\bar{r}}) \mathcal{E}(\mathbf{X}_r)^\top + \lambda_B^* \mathbf{Z}^* = \mathbf{0}$$

$$\mathbf{Q}^* \in \mathcal{A}_\varepsilon^{W_p}(\tilde{\mathbb{P}}_m), \|\mathbf{Z}^*\|_{B,2,\infty} \leq 1, \lambda_B^* \geq 0, \|\hat{\mathbf{W}}\|_{B,2,1} \leq \bar{B}$$

$$\langle \mathbf{Z}^*, \hat{\mathbf{W}} \rangle = \|\hat{\mathbf{W}}\|_{B,2,1}, \lambda_B^* (\|\hat{\mathbf{W}}\|_{B,2,1} - \bar{B}) = 0.$$

Note that we assume that the constraint $\|\mathbf{W}\|_{B,2,1} \leq \bar{B}$ is active such that $\lambda_B^* > 0$. This assumption is only for convenience of theoretical analysis and not restrictive. If it is not active, we have $\|\hat{\mathbf{W}}\|_{B,2,1} = \check{B} < \bar{B}$ for some \check{B} and $\lambda_B^* = 0$, which leads to an unconstrained problem similar to the ordinary least square problem, which is known to suffer from overfitting. Instead, we are usually interested in solutions that have finite norms so we can always find $\bar{B} = \check{B} - \epsilon < \check{B}$ for some small positive constant $\epsilon > 0$ to make the constraint active and thus $\lambda_B^* > 0$.

Substituting $\mathcal{E}(\mathbf{X}_r) = \mathbf{W}^{*\top} \mathcal{E}(\mathbf{X}_{\bar{r}}) + \mathbf{e}$ into the first-order optimality condition yields

$$\begin{aligned} & \mathbb{E}_{\mathbf{Q}^*} \mathcal{E}(\mathbf{X}_{\bar{r}}) \mathcal{E}(\mathbf{X}_{\bar{r}})^\top (\hat{\mathbf{W}} - \mathbf{W}^*) - \mathbb{E}_{\mathbf{Q}^*} \mathcal{E}(\mathbf{X}_{\bar{r}}) \mathbf{e}^\top + \lambda_B^* \mathbf{Z}^* = \mathbf{0} \\ \iff & \begin{bmatrix} \mathbf{H}_{\mathcal{S}_r \mathcal{S}_r}^{\mathbf{Q}^*} & \mathbf{H}_{\mathcal{S}_r \mathcal{S}_r^c}^{\mathbf{Q}^*} \\ \mathbf{H}_{\mathcal{S}_r^c \mathcal{S}_r}^{\mathbf{Q}^*} & \mathbf{H}_{\mathcal{S}_r^c \mathcal{S}_r^c}^{\mathbf{Q}^*} \end{bmatrix} \begin{bmatrix} \hat{\mathbf{W}}_{\mathcal{S}_{r\cdot}} - \mathbf{W}_{\mathcal{S}_{r\cdot}}^* \\ \mathbf{0} \end{bmatrix} - \begin{bmatrix} \mathbb{E}_{\mathbf{Q}^*} \mathcal{E}(\mathbf{X}_{\bar{r}})_{\mathcal{S}_r} \mathbf{e}^\top \\ \mathbb{E}_{\mathbf{Q}^*} \mathcal{E}(\mathbf{X}_{\bar{r}})_{\mathcal{S}_r^c} \mathbf{e}^\top \end{bmatrix} + \lambda_B^* \begin{bmatrix} \mathbf{Z}_{\mathcal{S}_{r\cdot}}^* \\ \mathbf{Z}_{\mathcal{S}_r^c}^* \end{bmatrix} = \begin{bmatrix} \mathbf{0} \\ \mathbf{0} \end{bmatrix}. \quad (3.17) \end{aligned}$$

Solving for $\mathbf{Z}_{\mathcal{S}_r^c}^*$, we find that

$$\lambda_B^* \mathbf{Z}_{\mathcal{S}_r^c}^* = \lambda_B^* \mathbf{H}_{\mathcal{S}_r^c \mathcal{S}_r}^{\mathbf{Q}^*} (\mathbf{H}_{\mathcal{S}_r \mathcal{S}_r}^{\mathbf{Q}^*})^{-1} \mathbf{Z}_{\mathcal{S}_{r\cdot}}^* - \mathbf{H}_{\mathcal{S}_r^c \mathcal{S}_r}^{\mathbf{Q}^*} (\mathbf{H}_{\mathcal{S}_r \mathcal{S}_r}^{\mathbf{Q}^*})^{-1} \mathbb{E}_{\mathbf{Q}^*} \mathcal{E}(\mathbf{X}_{\bar{r}})_{\mathcal{S}_r} \mathbf{e}^\top + \mathbb{E}_{\mathbf{Q}^*} \mathcal{E}(\mathbf{X}_{\bar{r}})_{\mathcal{S}_r^c} \mathbf{e}^\top,$$

which can be bounded such that

$$\begin{aligned}
& \lambda_B^* \| \mathbf{Z}_{\mathcal{S}_r^c}^* \|_{B,2,\infty} \\
&= \| \lambda_B^* \mathbf{H}_{\mathcal{S}_r^c \mathcal{S}_r}^{\mathbb{Q}^*} (\mathbf{H}_{\mathcal{S}_r \mathcal{S}_r}^{\mathbb{Q}^*})^{-1} \mathbf{Z}_{\mathcal{S}_r^c}^* - \mathbf{H}_{\mathcal{S}_r^c \mathcal{S}_r}^{\mathbb{Q}^*} (\mathbf{H}_{\mathcal{S}_r \mathcal{S}_r}^{\mathbb{Q}^*})^{-1} \mathbb{E}_{\mathbb{Q}^*} \mathcal{E}(\mathbf{X}_{\bar{r}})_{\mathcal{S}_r} \mathbf{e}^\top + \mathbb{E}_{\mathbb{Q}^*} \mathcal{E}(\mathbf{X}_{\bar{r}})_{\mathcal{S}_r^c} \mathbf{e}^\top \|_{B,2,\infty} \\
&\leq \lambda_B^* \| \mathbf{H}_{\mathcal{S}_r^c \mathcal{S}_r}^{\mathbb{Q}^*} (\mathbf{H}_{\mathcal{S}_r \mathcal{S}_r}^{\mathbb{Q}^*})^{-1} \mathbf{Z}_{\mathcal{S}_r^c}^* \|_{B,2,\infty} + \| \mathbf{H}_{\mathcal{S}_r^c \mathcal{S}_r}^{\mathbb{Q}^*} (\mathbf{H}_{\mathcal{S}_r \mathcal{S}_r}^{\mathbb{Q}^*})^{-1} \mathbb{E}_{\mathbb{Q}^*} \mathcal{E}(\mathbf{X}_{\bar{r}})_{\mathcal{S}_r} \mathbf{e}^\top \|_{B,2,\infty} \\
&\quad + \| \mathbb{E}_{\mathbb{Q}^*} \mathcal{E}(\mathbf{X}_{\bar{r}})_{\mathcal{S}_r^c} \mathbf{e}^\top \|_{B,2,\infty} \\
&\leq \lambda_B^* \| \mathbf{H}_{\mathcal{S}_r^c \mathcal{S}_r}^{\mathbb{Q}^*} (\mathbf{H}_{\mathcal{S}_r \mathcal{S}_r}^{\mathbb{Q}^*})^{-1} \|_{B,1,\infty} \| \mathbf{Z}_{\mathcal{S}_r^c}^* \|_{2,\infty} + \| \mathbf{H}_{\mathcal{S}_r^c \mathcal{S}_r}^{\mathbb{Q}^*} (\mathbf{H}_{\mathcal{S}_r \mathcal{S}_r}^{\mathbb{Q}^*})^{-1} \|_{B,1,\infty} \| \mathbb{E}_{\mathbb{Q}^*} \mathcal{E}(\mathbf{X}_{\bar{r}})_{\mathcal{S}_r} \mathbf{e}^\top \|_{2,\infty} \\
&\quad + \| \mathbb{E}_{\mathbb{Q}^*} \mathcal{E}(\mathbf{X}_{\bar{r}})_{\mathcal{S}_r^c} \mathbf{e}^\top \|_{B,2,\infty}.
\end{aligned}$$

Note that

$$\| \mathbf{Z}_{\mathcal{S}_r^c}^* \|_{2,\infty} \leq \| \mathbf{Z}^* \|_{B,2,\infty} \leq 1.$$

Recall that $0 < \alpha \leq 1$ in Assumption 20. Based on Lemma 24 and Lemma 25, we may write

$$\begin{aligned}
& \lambda_B^* \| \mathbf{Z}_{\mathcal{S}_r^c}^* \|_{B,2,\infty} \\
&\leq \lambda_B^* \| \mathbf{H}_{\mathcal{S}_r^c \mathcal{S}_r}^{\mathbb{Q}^*} (\mathbf{H}_{\mathcal{S}_r \mathcal{S}_r}^{\mathbb{Q}^*})^{-1} \|_{B,1,\infty} \| \mathbf{Z}_{\mathcal{S}_r^c}^* \|_{2,\infty} + \| \mathbf{H}_{\mathcal{S}_r^c \mathcal{S}_r}^{\mathbb{Q}^*} (\mathbf{H}_{\mathcal{S}_r \mathcal{S}_r}^{\mathbb{Q}^*})^{-1} \|_{B,1,\infty} \| \mathbb{E}_{\mathbb{Q}^*} \mathcal{E}(\mathbf{X}_{\bar{r}})_{\mathcal{S}_r} \mathbf{e}^\top \|_{2,\infty} \\
&\quad + \| \mathbb{E}_{\mathbb{Q}^*} \mathcal{E}(\mathbf{X}_{\bar{r}})_{\mathcal{S}_r^c} \mathbf{e}^\top \|_{B,2,\infty} \\
&\leq \lambda_B^* (1 - \frac{\alpha}{2}) + (1 - \frac{\alpha}{2}) (\frac{\lambda_B^* \alpha}{8(1 - \alpha/2)}) + \frac{\lambda_B^* \alpha}{8} \\
&\leq \lambda_B^* (1 - \frac{\alpha}{4}) \\
&< \lambda_B^*,
\end{aligned}$$

with high probability and certain conditions on λ_B^* and ε .

Henceforth, $\|\mathbf{Z}_{\mathcal{S}_r^c}^*\|_{B,2,\infty} < 1$ satisfies strict dual feasibility and we must have $\|\hat{\mathbf{W}}_{\mathcal{S}_r^c}^*\|_{B,2,1} = 0$ according to complementary slackness: $\langle \mathbf{Z}^*, \hat{\mathbf{W}} \rangle = \|\hat{\mathbf{W}}\|_{B,2,1}$. In other words, we have

$$\forall i \in \mathbf{Co}_r \quad \hat{\mathbf{W}}_i = \mathbf{0},$$

with high probability. This guarantees that we do not recover any node that is not a neighbor of r with high probability.

(iii) Without information about the true skeleton, we have a unique and optimal solution.

We follow the proof of Lemma 11.2 in (Hastie et al., 2015).

We have shown that $\hat{\mathbf{W}}$ satisfying $\hat{\mathbf{W}}_i = \mathbf{0} \quad \forall i \in \mathbf{Co}_r$ is an optimal solution with optimal dual variables $\|\mathbf{Z}_{\mathcal{S}_r^c}^*\|_{B,2,\infty} < 1$.

To avoid clutter of notations, we define

$$L^{\text{DRO}}(\mathbf{W}) := \sup_{\mathbb{Q} \in \mathcal{A}_\varepsilon^{W_P}(\tilde{\mathbb{P}}_m)} \frac{1}{2} \mathbb{E}_{\mathbb{Q}} \|\mathcal{E}(X_r) - \mathbf{W}^\top \mathcal{E}(\mathbf{X}_{\bar{r}})\|_2^2.$$

Let $(\check{\mathbf{W}}, \check{\lambda})$ be any other optimal solution to $\inf_{\mathbf{W}} \sup_{\lambda} L^{\text{DRO}}(\mathbf{W}) + \lambda(\|\mathbf{W}\|_{B,2,1} - \bar{B})$. By definition,

$$\begin{aligned} L^{\text{DRO}}(\check{\mathbf{W}}) + \check{\lambda}(\|\check{\mathbf{W}}\|_{B,2,1} - \bar{B}) &= L^{\text{DRO}}(\hat{\mathbf{W}}) + \lambda_B^*(\langle \mathbf{Z}^*, \hat{\mathbf{W}} \rangle - \bar{B}) \\ \iff L^{\text{DRO}}(\check{\mathbf{W}}) + \check{\lambda}(\|\check{\mathbf{W}}\|_{B,2,1} - \bar{B}) - \lambda_B^* \langle \mathbf{Z}^*, \check{\mathbf{W}} \rangle &= L^{\text{DRO}}(\hat{\mathbf{W}}) + \lambda_B^*(\langle \mathbf{Z}^*, \hat{\mathbf{W}} - \check{\mathbf{W}} \rangle - \bar{B}). \end{aligned}$$

The first-order optimality condition for $\hat{\mathbf{W}}$ says

$$\nabla L^{\text{DRO}}(\hat{\mathbf{W}}) + \lambda_B^* \mathbf{Z}^* = \mathbf{0},$$

which implies

$$\check{\lambda}(\|\check{\mathbf{W}}\|_{B,2,1} - \bar{B}) + \lambda_B^*(\bar{B} - \langle \mathbf{Z}^*, \check{\mathbf{W}} \rangle) = L^{\text{DRO}}(\hat{\mathbf{W}}) + \langle \nabla L^{\text{DRO}}(\hat{\mathbf{W}}), \check{\mathbf{W}} - \hat{\mathbf{W}} \rangle - L^{\text{DRO}}(\check{\mathbf{W}}).$$

By definition, $\|\check{\mathbf{W}}\|_{B,2,1} - \bar{B} = 0$ and $\lambda_B^* > 0$. Since $L^{\text{DRO}}(\cdot)$ is convex, the RHS of the above equation should be non-positive, or equivalently,

$$\|\check{\mathbf{W}}\|_{B,2,1} \leq \langle \mathbf{Z}^*, \check{\mathbf{W}} \rangle.$$

On the other hand,

$$\langle \mathbf{Z}^*, \check{\mathbf{W}} \rangle \leq \|\mathbf{Z}^*\|_{B,2,\infty} \|\check{\mathbf{W}}\|_{B,2,1} \leq \|\check{\mathbf{W}}\|_{B,2,1}.$$

Therefore, the equality holds for the above inequalities, which leads to

$$\|\check{\mathbf{W}}\|_{B,2,1} = \langle \mathbf{Z}^*, \check{\mathbf{W}} \rangle.$$

Recall that $\|\mathbf{Z}_{\mathcal{S}_r^c}^*\|_{B,2,\infty} < 1$. In order for $\|\check{\mathbf{W}}\|_{B,2,1} = \langle \mathbf{Z}^*, \check{\mathbf{W}} \rangle$ to hold, we must have

$$\check{\mathbf{W}}_{\mathcal{S}_r^c} = \mathbf{0}.$$

In that wise, all the optimal solutions $\check{\mathbf{W}}$ have

$$\check{\mathbf{W}}_i = \mathbf{0} \quad \forall i \in \mathbf{Co}_r.$$

This implies that we have a unique solution that excludes all the non-neighbor nodes without information about the true skeleton. Until now, we have proven properties (a) and (b).

(iv) The set of correct neighbors is recovered.

Consider again the first-order optimality condition in Equation 3.17,

$$\begin{aligned} \hat{\mathbf{W}}_{\mathcal{S}_r} - \mathbf{W}_{\mathcal{S}_r}^* &= (\mathbf{H}_{\mathcal{S}_r \mathcal{S}_r}^{\mathbb{Q}*})^{-1} (\mathbb{E}_{\mathbb{Q}*} \mathcal{E}(\mathbf{X}_{\bar{r}})_{\mathcal{S}_r} \mathbf{e}^\top - \lambda_B^* \mathbf{Z}_{\mathcal{S}_r}^*) \\ \implies \|\hat{\mathbf{W}}_{\mathcal{S}_r} - \mathbf{W}_{\mathcal{S}_r}^*\|_{B,2,\infty} &= \|(\mathbf{H}_{\mathcal{S}_r \mathcal{S}_r}^{\mathbb{Q}*})^{-1} (\mathbb{E}_{\mathbb{Q}*} \mathcal{E}(\mathbf{X}_{\bar{r}})_{\mathcal{S}_r} \mathbf{e}^\top - \lambda_B^* \mathbf{Z}_{\mathcal{S}_r}^*)\|_{B,2,\infty} \\ &\leq \|(\mathbf{H}_{\mathcal{S}_r \mathcal{S}_r}^{\mathbb{Q}*})^{-1}\|_{B,1,\infty} \|\mathbb{E}_{\mathbb{Q}*} \mathcal{E}(\mathbf{X}_{\bar{r}})_{\mathcal{S}_r} \mathbf{e}^\top - \lambda_B^* \mathbf{Z}_{\mathcal{S}_r}^*\|_{2,\infty} \\ &\leq \|(\mathbf{H}_{\mathcal{S}_r \mathcal{S}_r}^{\mathbb{Q}*})^{-1}\|_{B,1,\infty} (\|\mathbb{E}_{\mathbb{Q}*} \mathcal{E}(\mathbf{X}_{\bar{r}})_{\mathcal{S}_r} \mathbf{e}^\top\|_{2,\infty} + \|\lambda_B^* \mathbf{Z}_{\mathcal{S}_r}^*\|_{2,\infty}) \\ &\leq \rho_{\max} \|(\mathbf{H}_{\mathcal{S}_r \mathcal{S}_r}^{\mathbb{Q}*})^{-1}\|_{\infty,\infty} (\|\mathbb{E}_{\mathbb{Q}*} \mathcal{E}(\mathbf{X}_{\bar{r}})_{\mathcal{S}_r} \mathbf{e}^\top\|_{2,\infty} + \lambda_B^*) \\ &\leq \rho_{\max} \sqrt{|\mathcal{S}_r|} \|(\mathbf{H}_{\mathcal{S}_r \mathcal{S}_r}^{\mathbb{Q}*})^{-1}\|_{2,2} (\|\mathbb{E}_{\mathbb{Q}*} \mathcal{E}(\mathbf{X}_{\bar{r}})_{\mathcal{S}_r} \mathbf{e}^\top\|_{2,\infty} + \lambda_B^*) \end{aligned}$$

According to Equation 3.13, with probability at least $1 - 2|\mathcal{S}_r|^2 \exp(-\frac{m(\Lambda_{\min}(\mathbf{H}_{\mathcal{S}_r \mathcal{S}_r}))^2}{8|\mathcal{S}_r|^2})$ and $\varepsilon \leq \frac{\Lambda_{\min}(\mathbf{H}_{\mathcal{S}_r \mathcal{S}_r})}{16|\mathcal{S}_r|^{\frac{1}{2}}}$,

$$\|(\mathbf{H}_{\mathcal{S}_r \mathcal{S}_r}^{\mathbb{Q}})^{-1}\|_{2,2} \leq \sqrt{\frac{4}{\Lambda_{\min}(\mathbf{H}_{\mathcal{S}_r \mathcal{S}_r})}}.$$

According to Equation 3.15, with probability at least $1 - |\mathcal{S}_r|\rho_r \exp(-\frac{m\mu^2}{2\sigma^2})$, $\varepsilon \leq \frac{\mu}{\sigma}$ and $\lambda_B^* > \frac{32\mu\sqrt{\rho_r}(1-\alpha/2)}{\alpha}$, we have

$$\|\mathbb{E}_{\mathbb{Q}} \mathcal{E}(\mathbf{X}_{\bar{r}})_{\mathcal{S}_r} \mathbf{e}^{\top}\|_{2,\infty} \leq \frac{\lambda_B^* \alpha}{8(1-\alpha/2)}.$$

On that account, with probability at least

$$1 - 2|\mathcal{S}_r|^2 \exp(-\frac{m(\Lambda_{\min}(\mathbf{H}_{\mathcal{S}_r \mathcal{S}_r}))^2}{8|\mathcal{S}_r|^2}) - |\mathcal{S}_r|\rho_r \exp(-\frac{m\mu^2}{2\sigma^2})$$

and $\varepsilon \leq \min(\frac{\Lambda_{\min}(\mathbf{H}_{\mathcal{S}_r \mathcal{S}_r})}{16|\mathcal{S}_r|^{\frac{1}{2}}}, \frac{\mu}{\sigma})$ while requiring $\lambda_B^* > \frac{32\mu\sqrt{\rho_r}(1-\alpha/2)}{\alpha}$,

$$\|\hat{\mathbf{W}}_{\mathcal{S}_r \cdot} - \mathbf{W}_{\mathcal{S}_r \cdot}^*\|_{B,2,\infty} \leq \rho_{\max} \sqrt{|\mathcal{S}_r|} \sqrt{\frac{4}{\Lambda_{\min}(\mathbf{H}_{\mathcal{S}_r \mathcal{S}_r})}} \lambda_B^* (\frac{\alpha}{8(1-\alpha/2)} + 1).$$

By Assumption 18, if the condition $\lambda_B^* < \frac{\beta}{2(\frac{\alpha}{8(1-\alpha/2)} + 1)\rho_{\max} \sqrt{|\mathcal{S}_r|}} \sqrt{\frac{\Lambda_{\min}(\mathbf{H}_{\mathcal{S}_r \mathcal{S}_r})}{4}}$ is satisfied, the following inequality holds:

$$\|\hat{\mathbf{W}}_{\mathcal{S}_r \cdot} - \mathbf{W}_{\mathcal{S}_r \cdot}^*\|_{B,2,\infty} < \beta/2.$$

In this way, we are able to recover all the neighbor nodes with a threshold $\beta/2$. This proves (c).

(v) The true skeleton is recovered with high probability.

The above arguments tell us that with high probability and certain conditions for ε and λ_B^* satisfied, for each node r , we do not recover any non-neighbor and we do recover all the neighbor nodes. The correct \mathbf{Ne}_r and \mathbf{Co}_r are thus identified. Now we are ready to prove (d).

Putting everything together and taking the the union bound for all nodes $r \in [n]$, with probability at least $1 - \mathcal{O}(n \exp(-\frac{Cm\mu^2}{\sigma^2\rho_{\max}^4\rho_{[n]}^3} + 2 \log \rho_{[n]}))$, $\varepsilon \leq \frac{C\mu}{\sigma\rho_{\max}\rho_{[n]}^{3/2}}$ and $\frac{32\mu\rho_{\max}}{\alpha} < \lambda_B^* < \frac{\beta}{2(\frac{\alpha}{8(1-\alpha/2)}+1)\rho_{\max}\sqrt{\rho_{[n]}}}\sqrt{\frac{\Lambda}{4}}$, where C only depends on α, Λ , we have

$$\hat{\mathcal{G}}_{\text{skel}} = \mathcal{G}_{\text{skel}}.$$

Setting $\varepsilon = \frac{\varepsilon_0}{m}$ and making the dependence on the sample size more explicit. We draw the conclusion that, if the number of samples satisfies

$$m = \mathcal{O}\left(\frac{C(\varepsilon_0 + \log(n/\delta) + \log \rho_{[n]})\sigma^2\rho_{\max}^4\rho_{[n]}^3}{\min(\mu^2, 1)}\right),$$

where C only depends on α, Λ , and if λ_B^* satisfies

$$\frac{32\mu\rho_{\max}}{\alpha} < \lambda_B^* < \frac{\beta}{(\alpha/(4-2\alpha) + 2)\rho_{\max}\sqrt{\rho_{[n]}}}\sqrt{\frac{\Lambda}{4}},$$

then with probability at least $1 - \delta$ for $\delta \in (0, 1]$:

$$\hat{\mathcal{G}}_{\text{skel}} = \mathcal{G}_{\text{skel}}.$$

Moreover, if we assume that the target graph has a bounded degree of d , the sample complexity becomes logarithmic in n :

$$m = \mathcal{O}\left(\frac{C(\varepsilon_0 + \log(n/\delta) + \log n + \log \rho_{\max})\sigma^2 \rho_{\max}^7 d^3}{\min(\mu^2, 1)}\right).$$

□

The results in Theorem 26 encompass some intuitive interpretations. Compared to Theorem 1 in (Bank and Honorio, 2020), we make more explicit the relationship among m , λ_B^* and δ . On one hand, the lower bound of λ_B^* ensures that a sparse solution excluding non-neighbor nodes is obtained. A large error magnitude expectation μ therefore elicits stronger regularization. On the other hand, the upper bound λ_B^* is imposed to guarantee that all the neighbor nodes are identified with less restriction on \mathbf{W} . There is naturally a trade-off when choosing \bar{B} in order to learn the exact skeleton. The sample complexity depends on cardinalities $\rho_{[n]}$, confidence level δ , the number of nodes n , the ambiguity level ε_0 and assumptions on errors. The dependence on σ indicates that higher uncertainty caused by larger error norms demands more samples whereas the dependence on μ^{-2} results from the lower bound condition on λ_B^* with respect to μ . The ambiguity level is set to ε_0/m based on the observation that obtaining more samples reduces ambiguity of the true distribution. In practice, we find that ε_0 is usually small thus

negligible. Note that the sample complexity is polynomial in n . Furthermore, if we assume that the true graph has a bounded degree of d , we find that

$$m = \mathcal{O}\left(\frac{C(\varepsilon_0 + \log(n/\delta) + \log n + \log \rho_{\max})\sigma^2\rho_{\max}^7 d^3}{\min(\mu^2, 1)}\right)$$

is logarithmic with respect to n , consistent with the results in (Wainwright, 2009).

We introduce constants \bar{B} and λ_B^* in order to find a condition for the statements in Theorem 26 to hold. If there exists a \mathbf{W} incurring a finite loss, we can always find a solution $\hat{\mathbf{W}}$ satisfying $\|\hat{\mathbf{W}}\|_{B,2,1} < +\infty$ and let \bar{B} be the maximum norm. Imposing $\|\mathbf{W}\|_{B,2,1} \leq \bar{B}$ is equivalent to the original problem. By Lagrange duality and similar argument for the lasso estimator, there exists a λ_B^* that finds all the solutions with $\|\hat{\mathbf{W}}\|_{B,2,1} = \bar{B}$. Therefore we have a mapping between ε and λ_B^* .

3.3.3 Kullback-Leibler Formulation

In addition to optimal transport, ϕ -divergence is also widely used to construct an ambiguity set for DRO problems. We consider the KL divergence in this sub-section. Note that any other point outside the support of the nominal distribution remains to have zero probability in an ambiguity set constructed by the KL divergence. However, we argue that adopting the KL divergence may bring advantages over the Wasserstein distance since the Bayesian network distribution we study is a discrete distribution over purely categorical random variables. Moreover, as illustrated below, adopting the KL divergence leads to better computational efficiency.

Let $\mathcal{A} \triangleq \mathcal{A}_\varepsilon^D(\tilde{\mathbb{P}}_m)$ be the ambiguity set, the dual formulation of Equation 3.3 follows directly from Theorem 4 in (Hu and Hong, 2013):

$$\inf_{\mathbf{W}, \gamma \geq 0} \gamma \ln \left[\frac{1}{m} \sum_{i \in [m]} e^{\frac{1}{2} \|\mathcal{E}(x_r^{(i)}) - \mathbf{W}^\top \mathcal{E}(\mathbf{x}_r^{(i)})\|_2^2 / \gamma} \right] + \gamma \varepsilon,$$

which directly minimizes a convex objective. In contrast to the approximate Wasserstein estimator, this KL DRO estimator finds the exact solution to the primal problem by strong duality.

The worst-case risk over a KL divergence ball can be bounded by variance (Lam, 2019), similar to Lipschitz regularization in Lemma 22. Based on this observation, we derive the following results:

Theorem 27. *Suppose that $\hat{\mathbf{W}}$ is a DRO risk minimizer of Equation 3.4 with the KL divergence and an ambiguity radius $\varepsilon = \varepsilon_0/m$. Given the same definitions of $(\mathcal{G}, \mathbb{P})$, \mathcal{G}_{skel} , \bar{B} , λ_B^* , m in Theorem 26. Under Assumptions 17, 18, 19, 20, if the number of samples satisfies*

$$m = \mathcal{O}\left(\frac{C(\varepsilon_0 + \log(n/\delta) + \log \rho_{[n]})\sigma^2 \rho_{max}^4 \rho_{[n]}^3}{\min(\mu^2, 1)}\right).$$

where C depends on α , Λ while independent of n , and if the Lagrange multiplier satisfies the same condition as in Theorem 26, then for any $\delta \in (0, 1]$, $r \in [n]$, with probability at least $1 - \delta$, the properties (a)-(d) in Theorem 26 hold.

Proof. Define

$$\ell_{\mathbf{W}}(\mathbf{X}) := \frac{1}{2} \|\mathcal{E}(X_r) - \mathbf{W}^\top \mathcal{E}(\mathbf{X}_{\bar{r}})\|_2^2.$$

According to Theorem 7 in (Lam, 2019), the worst-case risk with a KL divergence ambiguity set can be bounded as follows:

$$\begin{aligned} \sup_{\mathbb{Q} \in \mathcal{A}_\varepsilon^D(\tilde{\mathbb{P}}_m)} \mathbb{E}_{\mathbb{Q}} \ell_{\mathbf{W}}(\mathbf{X}) &\leq \mathbb{E}_{\tilde{\mathbb{P}}_m} \ell_{\mathbf{W}}(\mathbf{X}) + \sqrt{\varepsilon} \sqrt{\frac{1}{m} \sum_{i \in [m]} (\ell_{\mathbf{W}}(\mathbf{x}^{(i)}) - \bar{\ell}_{\mathbf{W}})^2} \\ &\quad + C\varepsilon \frac{\sum_{i \in [m]} |\ell_{\mathbf{W}}(\mathbf{x}^{(i)}) - \bar{\ell}_{\mathbf{W}}|^3}{\sum_{i \in [m]} (\ell_{\mathbf{W}}(\mathbf{x}^{(i)}) - \bar{\ell}_{\mathbf{W}})^2} \\ &\leq \mathbb{E}_{\tilde{\mathbb{P}}_m} \ell_{\mathbf{W}}(\mathbf{X}) + \sqrt{\varepsilon} \max_{i \in [m]} |\ell_{\mathbf{W}}(\mathbf{x}^{(i)}) - \bar{\ell}_{\mathbf{W}}| + C\varepsilon \max_{i \in [m]} |\ell_{\mathbf{W}}(\mathbf{x}^{(i)}) - \bar{\ell}_{\mathbf{W}}|, \end{aligned}$$

where $\bar{\ell}_{\mathbf{W}} = \frac{1}{m} \sum_{i \in [m]} \ell_{\mathbf{W}}(\mathbf{x}^{(i)})$ and $C > 0$ is constant independent of n .

Consider

$$\begin{aligned}
\max_{i \in [m]} |\ell_{\mathbf{W}}(\mathbf{x}^{(i)}) - \ell_{\mathbf{W}}^-| &\leq \max_{\mathbf{W}, \mathbf{W}', \mathbf{x}, \mathbf{x}'} |\ell_{\mathbf{W}}(\mathbf{x}) - \ell_{\mathbf{W}'}(\mathbf{x}')| \\
&\leq \max_{\mathbf{W}, \mathbf{x}} |\ell_{\mathbf{W}}(\mathbf{x})| \\
&\leq \frac{1}{2} \max_{\mathbf{W}, \mathbf{x}} (\|\mathcal{E}(X_r)\|_2 + \|\mathbf{W}^\top \mathcal{E}(\mathbf{X}_{\bar{r}})\|_2)^2 \\
&\leq \frac{1}{2} \max_{\mathbf{W}, \mathbf{x}} (\sqrt{\rho_{\max}} + \|\mathbf{W}^\top\|_{\infty, 2})^2 \\
&\leq \frac{1}{2} \max_{\mathbf{W}, \mathbf{x}} (\sqrt{\rho_{\max}} + \|\mathbf{W}\|_{1, 2})^2 \\
&\leq \frac{1}{2} \max_{\mathbf{W}, \mathbf{x}} (\sqrt{\rho_{\max}} + \sqrt{\rho_{[n]}} \|\mathbf{W}\|_F)^2 \\
&\leq \frac{1}{2} \max_{\mathbf{W}, \mathbf{x}} (\sqrt{\rho_{\max}} + \sqrt{\rho_{[n]}} \|\mathbf{W}\|_{B, 2, 1})^2 \\
&\leq \frac{1}{2} (\sqrt{\rho_{\max}} + \sqrt{\rho_{[n]}} \bar{B})^2 \\
&:= B_\rho.
\end{aligned}$$

Define $\varepsilon_{\max} := \max(\sqrt{\varepsilon}, \varepsilon)$. Therefore, we find that

$$\sup_{\mathbb{Q} \in \mathcal{A}_\varepsilon^D(\tilde{\mathbb{P}}_m)} \mathbb{E}_{\mathbb{Q}} \ell_{\mathbf{W}}(\mathbf{X}) \leq \mathbb{E}_{\tilde{\mathbb{P}}_m} \ell_{\mathbf{W}}(\mathbf{X}) + C \varepsilon_{\max} B_\rho.$$

Similar to the Wasserstein robust risk, we observe that the following results hold for any

$$\mathbb{Q} \in \mathcal{A}_\varepsilon^D(\tilde{\mathbb{P}}_m).$$

With probability at least $1 - 2|\mathcal{S}_r|^2 \exp(-\frac{mt^2}{2|\mathcal{S}_r|^2})$, we have

$$\Lambda_{\min}(\mathbf{H}_{\mathcal{S}_r \mathcal{S}_r}^{\mathbb{Q}}) \geq \Lambda_{\min}(\mathbf{H}_{\mathcal{S}_r \mathcal{S}_r}) - C\varepsilon_{\max}|\mathcal{S}_r|^{\frac{1}{2}} - t.$$

With probability at least $1 - 2|\mathcal{S}_r^c||\mathcal{S}_r| \exp(-\frac{mt^2}{2\rho_{\max}^2|\mathcal{S}_r|^2})$,

$$\|\mathbf{H}_{\mathcal{S}_r^c \mathcal{S}_r}^{\mathbb{Q}} - \mathbf{H}_{\mathcal{S}_r^c \mathcal{S}_r}\|_{B,1,\infty} \leq C\varepsilon_{\max}\rho_{\max}|\mathcal{S}_r| + t.$$

With probability at least $1 - 2|\mathcal{S}_r|^2 \exp(-\frac{mt^2}{2|\mathcal{S}_r|^2})$,

$$\|\|\mathbf{H}_{\mathcal{S}_r \mathcal{S}_r}^{\mathbb{Q}} - \mathbf{H}_{\mathcal{S}_r \mathcal{S}_r}\|\|_{\infty,\infty} \leq C\varepsilon_{\max}|\mathcal{S}_r| + t.$$

With probability at least $1 - 2|\mathcal{S}_r|^2 \exp(-\frac{mt^2(\Lambda_{\min}(\mathbf{H}_{\mathcal{S}_r \mathcal{S}_r}))^2}{32|\mathcal{S}_r|^3}) - 2|\mathcal{S}_r|^2 \exp(-\frac{m(\Lambda_{\min}(\mathbf{H}_{\mathcal{S}_r \mathcal{S}_r}))^2}{8|\mathcal{S}_r|^2})$ and

$$\varepsilon_{\max} \leq C \min\left(\frac{t\Lambda_{\min}(\mathbf{H}_{\mathcal{S}_r \mathcal{S}_r})}{8|\mathcal{S}_r|\sqrt{|\mathcal{S}_r|}}, \frac{\Lambda_{\min}(\mathbf{H}_{\mathcal{S}_r \mathcal{S}_r})}{16|\mathcal{S}_r|^{\frac{1}{2}}}\right),$$

$$\|\|(\mathbf{H}_{\mathcal{S}_r \mathcal{S}_r}^{\mathbb{Q}})^{-1} - (\mathbf{H}_{\mathcal{S}_r \mathcal{S}_r})^{-1}\|\|_{\infty,\infty} \leq t.$$

With probability at least $1 - \mathcal{O}(\exp(-\frac{Cm}{\rho_{\max}^2|\mathcal{S}_r|^3} + \log|\mathcal{S}_r^c| + \log|\mathcal{S}_r|))$ and $\varepsilon_{\max} \leq \frac{C}{\rho_{\max}|\mathcal{S}_r|^{3/2}}$,

$$\|\mathbf{H}_{\mathcal{S}_r^c \mathcal{S}_r}^{\mathbb{Q}}(\mathbf{H}_{\mathcal{S}_r \mathcal{S}_r}^{\mathbb{Q}})^{-1}\|_{B,1,\infty} \leq 1 - \frac{\alpha}{2},$$

where C only depends on α , $\Lambda_{\min}(\mathbf{H}_{\mathcal{S}_r \mathcal{S}_r})$.

Thanks to the boundedness of the error term \mathbf{e} , we have similar conclusions to Lemma 25 if $\varepsilon_{\max} \leq \frac{\mu}{\sigma}$ holds.

In such wise, the properties in Theorem 26 hold with the same condition on λ_B^* and the condition on ε_{\max} that $\varepsilon_{\max} \leq \frac{C\mu}{\sigma\rho_{\max}\rho_{[n]}^{3/2}}$. Since we set $\varepsilon = \frac{\varepsilon_0}{m}$ and define $\varepsilon_{\max} := \max(\sqrt{\varepsilon}, \varepsilon)$, the condition on ε_{\max} implies that

$$m \geq \max\left(\frac{\varepsilon_0 C^2 \sigma^2 \rho_{\max}^2 \rho_{[n]}^3}{\mu^2}, \frac{\varepsilon_0 C \sigma \rho_{\max} \rho_{[n]}^{3/2}}{\mu}\right).$$

The final sample complexity becomes

$$m = \mathcal{O}\left(\frac{C(\varepsilon_0 + \log(n/\delta) + \log \rho_{[n]})\sigma^2 \rho_{\max}^4 \rho_{[n]}^3}{\min(\mu^2, 1)}\right).$$

□

The sample complexities in Theorem 26 and Theorem 27 differ in the constant C due to the difference between the two probability metrics. Note that C is independent of n in both methods. The dependency on $1/(\lambda_B^*)^2$ is absorbed in the denominator because we require that $\lambda_B^* - 16\mu\rho_{\max}/\alpha > 0$. The sample complexities provide a perspective of our confidence on upper bounding the true risk in terms of the ambiguity radius. ε_0 serves as our initial guess on distributional uncertainty and increases the sample complexity only slightly because it is usually dominated by other terms in practice: $\varepsilon \ll \log(n/\delta)$. Even though the samples are drawn from an adversarial distribution with a proportion of noises, the proposed methods may still succeed as long as the true distribution can be made close to an upper confidence bound.

3.4 Experiments

We conduct experiments on benchmark datasets (Scutari, 2010) and real-world datasets (Malone et al., 2015) perturbed by the following contamination models:

- **Noisefree model.** This is the baseline model without any noises.
- **Huber’s contamination model.** In this model, each sample has a fixed probability of ζ to be replaced by a sample drawn from an arbitrary distribution. We adopt the uniform distribution.
- **Independent failure model.** Each entry in each sample is independently corrupted with probability ζ . We consider the model that replaces it with a different value uniformly in the experiments.

We conduct all experiments on a laptop with an Intel Core i7 2.7 GHz processor. We adopt the proposed approaches based on Wasserstein DRO and KL DRO as well as the group norm regularization method (Bank and Honorio, 2020) and the PC algorithm (Spirtes et al., 2000) for skeleton learning. Based on the learned skeletons, we infer a DAG with the hill-climbing (HC) algorithm (Tsamardinos et al., 2006). For the Wasserstein-based method, we leverage Adam (Kingma and Ba, 2014) to optimize the overall objective with $\beta_1 = 0.9$, $\beta_2 = 0.990$, a learning rate of 0.1, a batch size of 200, and a maximum of 100 iterations. For the KL-based and standard regularization methods, we use the L-BFGS-B (Byrd et al., 1995) optimization method with default parameters. We adopt the original version of the PC algorithm and set the cardinality of the maximum conditional set to 2. The Bayesian information criterion (BIC)

TABLE I: Comparisons of F1 scores for benchmark datasets and BIC for real-world datasets (backache, voting).

Dataset	asia	asia	asia	asia	asia	child	alarm	hailfinder	backache	voting
n	8	8	8	8	8	20	37	56	32	17
Noise	Noisefree	Huber	Indep	Indep	Indep	Indep	Indep	Indep	Indep	Indep
ζ	0	0.5	0.1	0.3	0.5	0.5	0.5	0.5	0.5	0.5
Wass	0.6606	0.5965	0.6740	0.5237	0.2190	0.3261	0.2065	0.1773	N/A	N/A
KL	0.6591	0.6655	0.6952	0.3285	0.4212	0.3679	0.1557	0.1629	N/A	N/A
Reg	0.7374	0.6655	0.6857	0.3285	0.0000	0.3417	0.1525	0.1551	N/A	N/A
PC	0.7062	0.5421	0.6292	0.1778	0.0000	0.1690	0.1132	0.1446	N/A	N/A
Wass+HC	0.7318	0.3732	0.3436	0.0444	0.0444	0.0891	N/A	N/A	-1793.2164	-3106.1863
KL+HC	0.7153	0.2702	0.3846	0.0000	0.1164	0.0874	N/A	N/A	-1793.2164	-3106.1863
Reg+HC	0.6589	0.2702	0.3846	0.0000	0.0000	0.1241	N/A	N/A	-1793.2164	-3106.1863
PC+HC	0.4675	0.2368	0.4195	0.0444	0.0000	0.0385	N/A	N/A	-1795.4472	-3106.1863

(Neath and Cavanaugh, 2012) score is adopted in the HC algorithm. Each experimental result is taken as an average over 5 independent runs where a random set of 1000 samples is obtained at the beginning. When dealing with real-world datasets, we split the data into two halves for training and testing.

We use the F1-score to evaluate performance on benchmark datasets and BIC for real-world datasets. The results are reported in Table I. We can observe that the proposed DRO methods either find the best skeleton or the best DAG with the help of HC across different datasets and different data contamination settings. For the alarm and hailfinder datasets, HC could not find a DAG in a reasonable amount of time. For the backache and voting datasets, BIC is only valid for DAGs but not for skeletons thus some results are not applicable.

3.5 Concluding Remarks

In this chapter, we put forward a distributionally robust optimization method to recover the skeleton of a general discrete Bayesian network. We discuss two specific probability metrics,

developed tractable algorithms to compute the estimators. We establish the connection between the proposed method and regularization. We derive non-asymptotic bounds polynomial in the number of nodes for successful identification of the true skeleton. The sample complexities become logarithmic for bounded-degree graphs. Empirical results showcase the effectiveness and practicability of our methods.

CHAPTER 4

MOMENT DISTRIBUTIONALLY ROBUST TREE STRUCTURED PREDICTION

(Parts of this chapter were previously published as “Moment Distributionally Robust Tree Structured Prediction” in the 36th Conference on Neural Information Processing Systems (NeurIPS 2022) (Li et al., 2022a).)

Structured prediction of tree-shaped objects is heavily studied under the name of syntactic dependency parsing. Current practice based on maximum likelihood or margin is either agnostic to or inconsistent with the evaluation loss. Risk minimization alleviates the discrepancy between training and test objectives but typically induces a non-convex problem. These approaches adopt explicit regularization to combat overfitting without probabilistic interpretation.

In this chapter, we propose a moment-based distributionally robust optimization approach for tree structured prediction, where the worst-case expected loss over a set of distributions within bounded moment divergence from the empirical distribution is minimized. We begin with an introduction in Section 4.1 and problem setup together with related work in Section 4.2. We develop efficient algorithms with theoretical analysis in Section 4.3, which includes Fisher consistency, convergence rates and generalization bounds. Section 4.4 proposes efficient projection oracles. Section 4.5 discusses extensions beyond first-order directed trees. Experimental results of comparing our method with a competitive baseline on dependency parsing benchmarks are given in Section 4.6. We conclude the chapter in Section 4.7.

4.1 Introduction

Structured prediction is an important learning setting for joint prediction of interdependent variables. The output space typically consists of an exponential number of structured objects whose inherent relations can be exploited to develop efficient learning algorithms and capture key properties of data (Ciliberto et al., 2019). Trees are widely used structures that offer expressiveness and simplicity. We distinguish between two different tree structured prediction tasks in the literature. The first task is a structure learning problem in graphical models (Bradley and Guestrin, 2010), aimed at constructing trees underlying a predictive model from training data. The optimal tree is found easily with greedy algorithms for generative models (Chow and Liu, 1968), while it is NP-hard for the discriminative max-margin setting (Meshi et al., 2013). The second task requires prediction itself to be a tree-shaped object (e.g., an incidence vector). Dependency parsing is a crucial application of this problem that has inspired a flurry of work in natural language processing. The first-order spanning tree prediction assuming factorization over arcs can be done in $\mathcal{O}(n^2)$ (Stanojević and Cohen, 2021), whereas exact inference is NP-hard for certain (non-projective) higher-order trees (e.g., considering siblings) (McDonald and Satta, 2007). We study the latter in this chapter.

A common evaluation criterion in dependency parsing is the attachment score, namely, the score we would like to maximize on test data. It is cost-sensitive to allow partially correct prediction. Ideally, the training objective should be aligned with the test objective. An early attempt to directly mimic test conditions leads to a non-convex piece-wise constant objective (Och, 2003). Risk minimization in appropriate parametric form has a non-convex smooth

objective, solvable with gradient descent, but still losing global convergence and generalization guarantees. Maximum likelihood approaches formulate a convex smooth problem minimizing a logistic loss, consistent with conditional probability estimates but oblivious to test losses. Maximum margin methods have convex objectives able to implicitly incorporate custom losses by scaling margins, but are known to be inconsistent with test losses generally (Nowak et al., 2022). Unfortunately, none of these approaches yield a Bayes optimal estimator for test losses with global convergence and finite-sample generalization guarantees.

Consistent structured prediction methods include (Ciliberto et al., 2016; Blondel, 2019; Nowak-Vila et al., 2020), the latter two of which are based on Fenchel-Young losses (Blondel et al., 2020). However, none of them have addressed the tree structured prediction problem explicitly. For instance, (Blondel, 2019) calls for Euclidean or Kullback-Leibler projection oracles, which do not exist in an efficient sense from what we know for arborescence (directed tree) polytopes. In addition, the Frank-Wolfe type algorithm adopted by (Nowak-Vila et al., 2020) requires a max-min oracle and converges in a rate of $\mathcal{O}(\frac{1}{\epsilon})$. Furthermore, all of the above methods belong to empirical risk minimization that requires explicit regularization to combat overfitting, which can be quite vulnerable in high-dimensional settings (e.g., scarce data).

To address the above issues, we propose an estimator from first principles in distributionally robust optimization. It minimizes the worst-case risk over an ambiguity set of distributions within bounded moment divergence from the empirical distribution. We seek probabilistic prediction by assuming non-deterministic groundtruth labels, which, together with the ambiguity set, models uncertainty about the unknown true distribution. We interpret the primal problem as

a dual-norm-regularized surrogate loss minimization problem. Note that prior art applying moment-based DRO to tree-structured graphical models (Fathony et al., 2018) and bipartite matching (Fathony et al., 2018) adopts a special case of our ambiguity set in which the empirical feature moments are matched exactly and regularization has to be imposed manually. This moment-based DRO also allows us to derive generalization bounds regarding true worst-case risks. When the ambiguity radius is zero, the DRO estimator is shown to be consistent. We develop two practical algorithms, one based on game theory and the other based on marginal probabilities of tree parts. We further propose efficient Euclidean projection oracles onto the arborescence polytope with linearly convergent guarantees. We conduct experiments on three common dependency parsing datasets, suggesting that our method is particularly effective with little training data.

Contributions. Our contributions are summarized as follows. (1) We propose a distributionally robust tree structured prediction method and show its equivalence to regularized surrogate minimization. (2) We derive its generalization bounds and consistency. (3) We propose efficient algorithms based on projection oracles for arborescence polytopes. (4) We perform empirical study on real-world datasets.

4.2 Background and Related Work

4.2.1 Tree Structured Prediction

Consider a weighted directed multi-graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ where each arc $(i, j, l) \in \mathcal{E}$ from node i to j has a label l . By designating a root node $r \in \mathcal{V}$, we say that $\mathcal{A} \subseteq \mathcal{E}$ is an r -arborescence of \mathcal{G} if $(\mathcal{V}, \mathcal{A})$ is a directed spanning tree rooted at r . For any $v \in \mathcal{V}$, denote by

$\delta^-(v) := \{(i, j, l) \in \mathcal{E} : j = v\}$ the set of its incoming arcs, and $\delta^+(v) := \{(i, j, l) \in \mathcal{E} : i = v\}$ the set of its outgoing arcs.

Let \mathcal{X} be the input space and $\mathcal{Y} \triangleq \bigcup_{\mathbf{x} \in \mathcal{X}} \mathcal{Y}(\mathbf{x})$ be the output space where $\mathcal{Y}(\mathbf{x})$ represents the set of r -arborescences of a graph $\mathcal{G}(\mathbf{x})$ formed by \mathbf{x} . Dependence on \mathbf{x} is suppressed when context is clear. Let $\mathcal{R} \subseteq 2^{\mathcal{E}}$ be a set of parts with $\mathcal{E} \subseteq \mathcal{R}$. Each part $s \in \mathcal{R}$ is a subset of arcs. It is convenient to represent $\mathbf{y} \in \mathcal{Y}$ as a binary vector with $y_s = 1$ iff part s appears in \mathbf{y} . Let $w_{\boldsymbol{\theta}}(\mathbf{x}, \mathbf{y}) \triangleq \sum_{s \in \mathcal{R}} w_{\boldsymbol{\theta}}(\mathbf{x}, y_s)$ be a score function decomposing over parts, parameterized by $\boldsymbol{\theta}$. Let $\{(\mathbf{x}^{(i)}, \mathbf{y}^{(i)})\}_{i=1}^m$ be a set of m training examples drawn i.i.d. from a distribution $\mathbb{P} \in \mathcal{P}(\mathcal{X} \times \mathcal{Y})$, where each $\mathbf{y}^{(i)}$ is an r -arborescence. The goal of tree structured prediction is to learn a function $h : \mathcal{X} \rightarrow \mathcal{Y}$ from training data. Assume that the evaluation criterion is a loss function $\ell : \mathcal{Y} \times \mathcal{Y} \rightarrow \mathbb{R}_{\geq 0}$.

We introduce existing methods in the setting of (graph-based, non-projective, syntactic) dependency parsing where \mathbf{x} is a sequence of tokens and $\mathcal{G}(\mathbf{x})$ encodes dependencies among tokens.

4.2.2 Maximum Likelihood

A probabilistic modeling approach based on exponential family distributions maximizes the conditional log-likelihood of the training data:

$$\min_{\boldsymbol{\theta}} - \sum_{i=1}^m \log p_{\boldsymbol{\theta}}(\mathbf{y}^{(i)} | \mathbf{x}^{(i)}) := - \sum_{i=1}^m \log [\exp(w_{\boldsymbol{\theta}}(\mathbf{x}^{(i)}, \mathbf{y}^{(i)})) / Z(\mathbf{x}^{(i)})],$$

where $Z(\mathbf{x}) \triangleq \sum_{\mathbf{y} \in \mathcal{Y}(\mathbf{x})} \exp(w_{\boldsymbol{\theta}}(\mathbf{x}, \mathbf{y}))$. This problem is convex for log-linear models, but intractable for general \mathcal{R} (Koller and Friedman, 2009). The first-order arc-factored model ($\mathcal{R} = \mathcal{E}$) is equivalent to a loop-free factor graph, rendering it tractable via the matrix-tree theorem (Kirchhoff, 1847; William, 1984; Koo et al., 2007; McDonald and Satta, 2007; Smith and Smith, 2007). Neural parsers either leverage the same theorem to compute the partition function (Ma and Hovy, 2017) or consider the parent node distribution independently for each node by local normalization (Dozat and Manning, 2017; Zhang et al., 2017). Higher-order models require approximate algorithms such as loopy belief propagation (Murphy et al., 1999) and Markov chain Monte Carlo (Brooks, 1998). This approach does not incorporate task-specific losses. In fact, with maximum a posteriori (MAP) decoding, it is not consistent with any specific loss in general (Nowak-Vila et al., 2019).

4.2.3 Maximum Margin

An alternative approach based on maximum margin Markov networks (Taskar et al., 2003) or structured support vector machines (Tsochantaridis et al., 2005) optimizes a hinge-type surrogate:

$$\min_{\boldsymbol{\theta}} \sum_{i=1}^m -w_{\boldsymbol{\theta}}(\mathbf{x}^{(i)}, \mathbf{y}^{(i)}) + \max_{\mathbf{y}} \ell(\mathbf{y}^{(i)}, \mathbf{y}) + w_{\boldsymbol{\theta}}(\mathbf{x}^{(i)}, \mathbf{y}),$$

which inspires a rich line of work based on MAP inference with manual features (Taskar et al., 2004; McDonald et al., 2005; McDonald and Pereira, 2006; Martins et al., 2009; Martins et al., 2010; Martins et al., 2015; Zhang et al., 2014) or deep learning (Kiperwasser and Goldberg,

2016; Wang and Chang, 2016). Approximate MAP inference is required for models beyond first-order. A smooth variant called softmax-margin (Gimpel and Smith, 2010) incorporates the task-specific loss ℓ but still implicitly minimizes it. Margin-based objectives are known to be consistent only under very restrictive conditions (Liu, 2007; Nowak et al., 2022) (i.e., data with majority label, loss being a distance).

4.2.4 Minimum Risk

Empirical risk minimization suggests directly optimizing the expected target loss on training data:

$$\min_{\theta} \sum_{i=1}^m \sum_{\mathbf{y}} p_{\theta}(\mathbf{y}|\mathbf{x}^{(i)}) \ell(\mathbf{y}^{(i)}, \mathbf{y}),$$

which is commonly non-convex due to normalization of p_{θ} . There are a few parsers optimizing this objective via back-propagation (Stoyanov and Eisner, 2012), k -best lists (Smith and Eisner, 2006), semirings (Li and Eisner, 2009; Zmigrod et al., 2021) and other differentiable approximations (Gormley et al., 2015; Mensch and Blondel, 2018). Local optima found by these algorithms do not satisfy the premise of Fisher consistency and make it difficult to quantify generalization errors.

4.3 Method

We introduce the formulation, followed by practical algorithms for learning and inference. Afterwards, we present the theoretical guarantees.

4.3.1 Formulation

We assume that the evaluation criterion is the Hamming loss $\ell(\mathbf{y}, \mathbf{y}') := \sum_i \mathbb{1}(y_i \neq y'_i)$ with $\mathbb{1}(\cdot)$ being the 0-1 indicator function, but the results in this chapter generalize to losses with affine decomposition (Ramaswamy et al., 2013) easily.

Let \mathbb{P}^{true} be the true distribution and \mathbb{P}^{emp} be the empirical distribution. Our approach builds upon a probabilistic predictor that non-parametrically minimizes the expected loss with regard to the most adverse distribution in an uncertainty set where the distributions are ε away from the empirical distribution in terms of feature moment difference:

$$\min_{\mathbb{P}} \max_{\mathbb{Q} \in \mathcal{B}(\mathbb{P}^{\text{emp}})} \mathbb{E}_{\mathbb{Q}_{\mathbf{X}}, \mathbb{P}_{\hat{\mathbf{Y}}|\mathbf{X}}} \ell(\hat{\mathbf{Y}}, \check{\mathbf{Y}}), \quad (4.1)$$

where $\mathcal{B}(\mathbb{P}^{\text{emp}}) := \{\mathbb{Q} : \mathbb{Q}_{\mathbf{X}} = \mathbb{P}_{\mathbf{X}}^{\text{emp}} \wedge \|\mathbb{E}_{\mathbb{P}^{\text{emp}}} \phi(\cdot) - \mathbb{E}_{\mathbb{Q}} \phi(\cdot)\| \leq \varepsilon\}$ with $\varepsilon \geq 0$ and $\phi : \mathcal{X} \times \mathcal{Y} \rightarrow \mathbb{R}^d$ is a joint feature mapping decomposable over parts: $\phi(\mathbf{x}, \mathbf{y}) \triangleq \sum_s \phi(\mathbf{x}, y_s)$. In (Farnia and Tse, 2016), cross-moments are adopted: $\phi(\mathbf{x}, \mathbf{y}) := \phi_{\mathbf{X}}(\mathbf{x}) \otimes \phi_{\mathbf{Y}}(\mathbf{y})$ where \otimes is the tensor product.

By Fenchel duality (Altun and Smola, 2006) and strong duality (von Neumann and Morgenstern, 1944), we show that Equation 4.1 is analogous to dual-norm-regularized surrogate loss minimization:

Proposition 28. *The distributionally robust tree structured prediction problem based on moment divergence in Equation 4.1 can be rewritten as*

$$\min_{\boldsymbol{\theta}} \mathbb{E}_{\mathbb{P}_{\mathbf{X}, \mathbf{Y}}^{\text{emp}}} \underbrace{\min_{\mathbb{P}} \max_{\mathbb{Q}} \mathbb{E}_{\mathbb{P}_{\hat{\mathbf{Y}}|\mathbf{X}}, \mathbb{Q}_{\check{\mathbf{Y}}|\mathbf{X}}} \ell(\hat{\mathbf{Y}}, \check{\mathbf{Y}}) + \boldsymbol{\theta}^\top (\phi(\mathbf{X}, \check{\mathbf{Y}}) - \phi(\mathbf{X}, \mathbf{Y})) + \varepsilon \|\boldsymbol{\theta}\|_*}_{\ell_{\text{adv}}(\boldsymbol{\theta}, (\mathbf{X}, \mathbf{Y}))}, \quad (4.2)$$

where $\boldsymbol{\theta} \in \mathbb{R}^d$ is the vector of Lagrangian multipliers and $\|\cdot\|_*$ is the dual norm of $\|\cdot\|$.

Proof. Recall the primal problem

$$\min_{\mathbb{P}} \max_{\mathbb{Q} \in \mathcal{B}(\mathbb{P}^{\text{emp}})} \mathbb{E}_{\mathbb{Q}_{\mathbf{X}, \check{\mathbf{Y}}} \mathbb{P}_{\hat{\mathbf{Y}}|\mathbf{X}}} \ell(\hat{\mathbf{Y}}, \check{\mathbf{Y}}),$$

where $\mathcal{B}(\mathbb{P}^{\text{emp}}) := \{\mathbb{Q} : \mathbb{Q}_{\mathbf{X}} = \mathbb{P}_{\mathbf{X}}^{\text{emp}} \wedge \|\mathbb{E}_{\mathbb{P}^{\text{emp}}} \phi(\cdot) - \mathbb{E}_{\mathbb{Q}} \phi(\cdot)\| \leq \varepsilon\}$ with $\varepsilon \geq 0$.

Note the feature function $\phi(\cdot)$ is fixed and given. Since $\mathbb{P}_{\hat{\mathbf{Y}}|\mathbf{X}} \in \Delta$ and $\mathbb{Q}_{\mathbf{X}, \check{\mathbf{Y}}} \in \Delta \cap \mathcal{B}(\mathbb{P}^{\text{emp}})$ where Δ is the probability simplex with dimension omitted, the constraint sets are convex. The objective function is convex in \mathbb{P} and concave in \mathbb{Q} because it is affine in both. Therefore strong duality holds:

$$\max_{\mathbb{Q} \in \mathcal{B}(\mathbb{P}^{\text{emp}})} \min_{\mathbb{P}} \mathbb{E}_{\mathbb{Q}_{\mathbf{X}, \check{\mathbf{Y}}} \mathbb{P}_{\hat{\mathbf{Y}}|\mathbf{X}}} \ell(\hat{\mathbf{Y}}, \check{\mathbf{Y}}).$$

Let $\mathcal{C} := \{\mathbf{u} : \|\mathbf{u} - \mathbb{E}_{\mathbb{P}^{\text{emp}}} \phi(\cdot)\| \leq \varepsilon\}$. Rewrite the problem with this constraint:

$$\begin{aligned} & \sup_{\mathbf{Q}, \mathbf{u}} \min_{\mathbb{P}} \mathbb{E}_{\mathbb{P}_{\mathbf{X}}^{\text{emp}} \mathbb{Q}_{\check{\mathbf{Y}}|\mathbf{X}} \mathbb{P}_{\check{\mathbf{Y}}|\mathbf{X}}} \ell(\hat{\mathbf{Y}}, \check{\mathbf{Y}}) - I_{\mathcal{C}}(\mathbf{u}) \\ \text{s.t. } & \mathbf{u} = \mathbb{E}_{\mathbb{P}_{\mathbf{X}}^{\text{emp}} \mathbb{Q}_{\check{\mathbf{Y}}|\mathbf{X}}} \phi(\mathbf{X}, \check{\mathbf{Y}}), \end{aligned}$$

where $I_{\mathcal{C}}(\cdot)$ is the indicator function with $I_{\mathcal{C}}(\mathbf{x}) = 0$ if $\mathbf{x} \in \mathcal{C}$ and $+\infty$ otherwise. The simplex constraints are omitted.

The dual problem by relaxing the equality constraint is

$$\sup_{\mathbf{Q}, \mathbf{u}} \min_{\boldsymbol{\theta}} \min_{\mathbb{P}} \mathbb{E}_{\mathbb{P}_{\mathbf{X}}^{\text{emp}} \mathbb{Q}_{\check{\mathbf{Y}}|\mathbf{X}} \mathbb{P}_{\check{\mathbf{Y}}|\mathbf{X}}} \ell(\hat{\mathbf{Y}}, \check{\mathbf{Y}}) - I_{\mathcal{C}}(\mathbf{u}) + \boldsymbol{\theta}^{\top} \mathbb{E}_{\mathbb{P}_{\mathbf{X}}^{\text{emp}} \mathbb{Q}_{\check{\mathbf{Y}}|\mathbf{X}}} \phi(\mathbf{X}, \check{\mathbf{Y}}) - \boldsymbol{\theta}^{\top} \mathbf{u},$$

where $\boldsymbol{\theta}$ is the vector of Lagrange multipliers.

Given $\mathbf{X} = \mathbf{x}$, optimization of $\mathbb{Q}_{\check{\mathbf{Y}}|\mathbf{x}}$ and $\mathbb{P}_{\check{\mathbf{Y}}|\mathbf{x}}$ can be done independently. Again by strong duality, we can rearrange the terms:

$$\min_{\boldsymbol{\theta}} \mathbb{E}_{\mathbb{P}_{\mathbf{X}}^{\text{emp}}} \min_{\mathbb{P}} \max_{\mathbb{Q}} \mathbb{E}_{\mathbb{Q}_{\check{\mathbf{Y}}|\mathbf{X}} \mathbb{P}_{\check{\mathbf{Y}}|\mathbf{X}}} \ell(\hat{\mathbf{Y}}, \check{\mathbf{Y}}) + \boldsymbol{\theta}^{\top} \phi(\mathbf{X}, \check{\mathbf{Y}}) + \sup_{\mathbf{u}} -I_{\mathcal{C}}(\mathbf{u}) - \boldsymbol{\theta}^{\top} \mathbf{u}.$$

The associated dual norm $\|\cdot\|_*$ of the norm $\|\cdot\|$ is defined as

$$\|\mathbf{z}\|_* := \sup\{\mathbf{z}^{\top} \mathbf{x} : \|\mathbf{x}\| \leq 1\},$$

based on which we are able to simplify the optimization over \mathbf{u} as

$$\sup_{\mathbf{u}} -I_{\mathcal{C}}(\mathbf{u}) - \boldsymbol{\theta}^\top \mathbf{u} = \sup_{\mathbf{u} \in \mathcal{C}} -\boldsymbol{\theta}^\top \mathbf{u} = \sup_{\mathbf{e}: \|\mathbf{e}\| \leq 1} -\boldsymbol{\theta}^\top (\mathbb{E}_{\mathbb{P}^{\text{emp}}} \boldsymbol{\phi}(\cdot) - \varepsilon \mathbf{e}) = -\boldsymbol{\theta}^\top \mathbb{E}_{\mathbb{P}^{\text{emp}}} \boldsymbol{\phi}(\cdot) + \varepsilon \|\boldsymbol{\theta}\|_*.$$

Plugging it back to the dual problem, we have

$$\min_{\boldsymbol{\theta}} \mathbb{E}_{\mathbb{P}_{\mathbf{X}, \mathbf{Y}}^{\text{emp}}} \min_{\mathbb{P}} \max_{\mathbb{Q}} \mathbb{E}_{\mathbb{Q}_{\hat{\mathbf{Y}}|\mathbf{X}} \mathbb{P}_{\check{\mathbf{Y}}|\mathbf{X}}} \ell(\hat{\mathbf{Y}}, \check{\mathbf{Y}}) + \boldsymbol{\theta}^\top (\boldsymbol{\phi}(\mathbf{X}, \check{\mathbf{Y}}) - \boldsymbol{\phi}(\mathbf{X}, \mathbf{Y})) + \varepsilon \|\boldsymbol{\theta}\|_*.$$

□

4.3.2 Constraint Generation Solution

From a game-theoretic rationale (Topsøe, 1979; Grünwald and Dawid, 2004), Equation 4.1 is considered as an adversary-constrained zero-sum game. A prediction player chooses a set of stochastic strategies (conditional distributions over arborescences) in order to minimize the expected payoff whereas an adversarial player chooses constrained strategies to maximize it. The payoff for a pair of pure strategies is the incurred loss, $\ell(\hat{\mathbf{y}}, \check{\mathbf{y}})$. The constrained game is transformed to a set of unconstrained ones in Equation 4.2 whose payoffs are parameterized by $\boldsymbol{\theta}$: $\text{payoff}(\hat{\mathbf{y}}, \check{\mathbf{y}}) \triangleq \ell(\hat{\mathbf{y}}, \check{\mathbf{y}}) + \boldsymbol{\theta}^\top \boldsymbol{\phi}(\mathbf{x}, \check{\mathbf{y}})$. Note that the games in Equation 4.1 are jointly constrained for all \mathbf{x} 's in the support of $\mathbb{P}_{\mathbf{X}}^{\text{emp}}$ while the ones in Equation 4.2 are conditionally independent given \mathbf{x} . The unconstrained game can be solved by a linear program (von Neumann and Morgenstern, 1944). However, there are $\mathcal{O}(n^n)$ spanning trees in a complete graph, thus making explicit construction of the full payoff matrix impractical.

Algorithm 3 Double Oracle Game Solver

Input: Lagrange multipliers θ ; feature function $\phi(\cdot, \cdot)$; initial set of trees $\{\mathbf{y}_{\text{initial}}\}$
Output: A sparse Nash equilibrium $(\hat{\mathcal{T}}, \check{\mathcal{T}}, \mathbb{P}, \mathbb{Q})$
 Initialize $\hat{\mathcal{T}} \leftarrow \check{\mathcal{T}} \leftarrow \{\mathbf{y}_{\text{initial}}\}$
repeat
 $(\mathbb{P}, \hat{v}_{\text{Nash}}) \leftarrow \text{SolveZeroSumGame}_{\hat{\mathcal{T}}}(\ell, \theta^\top \phi, \hat{\mathcal{T}}, \check{\mathcal{T}})$
 $(\check{\mathbf{y}}_{\text{BR}}, \check{v}_{\text{BR}}) \leftarrow \text{FindBestResponse}(\ell, \theta^\top \phi, \mathbb{P}, \hat{\mathcal{T}})$
 if $\hat{v}_{\text{Nash}} \neq \check{v}_{\text{BR}}$ **then**
 $\check{\mathcal{T}} \leftarrow \check{\mathcal{T}} \cup \{\check{\mathbf{y}}_{\text{BR}}\}$
 end if
 $(\mathbb{Q}, \check{v}_{\text{Nash}}) \leftarrow \text{SolveZeroSumGame}_{\check{\mathcal{T}}}(\ell, \theta^\top \phi, \hat{\mathcal{T}}, \check{\mathcal{T}})$
 $(\hat{\mathbf{y}}_{\text{BR}}, \hat{v}_{\text{BR}}) \leftarrow \text{FindBestResponse}(\ell, \theta^\top \phi, \mathbb{Q}, \check{\mathcal{T}})$
 if $\check{v}_{\text{Nash}} \neq \hat{v}_{\text{BR}}$ **then**
 $\hat{\mathcal{T}} \leftarrow \hat{\mathcal{T}} \cup \{\hat{\mathbf{y}}_{\text{BR}}\}$
 end if
until $\hat{v}_{\text{Nash}} = \check{v}_{\text{BR}} = \check{v}_{\text{Nash}} = \hat{v}_{\text{BR}}$
return $(\hat{\mathcal{T}}, \check{\mathcal{T}}, \mathbb{P}, \mathbb{Q})$

We adopt a constraint generation algorithm named double oracle (McMahan et al., 2003), with the pseudo-code illustrated in Algorithm 3. It builds a payoff sub-matrix starting from small initial sets of strategies. In each iteration, each player takes their turn based on the game payoff sub-matrix by finding the best response among all possible strategies to the opponent's optimal mixture strategies. The response is added to a player's strategy set if it improves the value of the game, with the sub-matrix updated. The algorithm terminates and converges to a Nash equilibrium of the original game when the strategy sets no longer grow. The size of the final sub-matrix is usually small in practice but there are no known theoretical guarantees, thus no way to analyze the convergence behavior. Finding the best response requires an oracle, equivalent to finding the minimum weight arborescence. The objective in Equation 4.2 is a

convex function of $\boldsymbol{\theta}$, so we can optimize it with sub-gradients based on solutions of the inner zero-sum games. Although lacking convergence guarantees, this algorithm is flexible with custom losses and provides a game-theoretic perspective to a typical DRO problem.

4.3.3 Marginal Distribution Formulation

The r -arborescence polytope is defined as the convex hull of all vectors representing r -arborescences: $\mathcal{A}_{\text{arb}}(\mathbf{x}) := \text{Conv}(\{\mathbf{y} \in \mathbb{R}^{|\mathcal{R}|} : \mathbf{y} \in \mathcal{Y}(\mathbf{x})\})$. Note that each $\mathbf{p} \in \mathcal{A}_{\text{arb}}$ is a convex combination of all r -arborescences: $\mathbf{p} \triangleq \sum_{\mathbf{y}} \text{Prob}(\mathbf{y})\mathbf{y}$, where p_s denotes the marginal probability of part s . Here we adopt the squared ℓ_2 norm as the dual norm and an ambiguity radius of $\varepsilon = \lambda/2$. By substituting the marginal probability vectors and switching min-max optimization orders, we simplify Equation 4.2 into

$$\max_{\mathbf{q}^{(i)} \in \mathcal{A}_{\text{arb}}} \min_{\boldsymbol{\theta}} \frac{1}{m} \sum_{i=1}^m \min_{\mathbf{p} \in \mathcal{A}_{\text{arb}}} (\mathbf{q}^{(i)} - \mathbf{p}_{\text{emp}}^{(i)})^\top \boldsymbol{\Phi}^{(i)} \boldsymbol{\theta} - \langle \mathbf{p}, \mathbf{q}^{(i)} \rangle + \frac{\mu}{2} \|\mathbf{p}\|_2^2 - \frac{\mu}{2} \|\mathbf{q}^{(i)}\|_2^2 + \frac{\lambda}{2} \|\boldsymbol{\theta}\|_2^2, \quad (4.3)$$

where $\boldsymbol{\Phi}^{(i)} \in \mathbb{R}^{|\mathcal{R}| \times d}$ denotes the feature matrix of the i -th training data, $\mu \in \mathbb{R}_{\geq 0}$ is a smoothing parameter to induce strong convexity. We push the maximization over \mathbf{q} to the outermost level because of its large computational cost. If $\mu = 0$, the solution to Equation 4.3 is also optimal to Equation 4.2 by strong duality but the problem becomes non-smooth. Therefore we expect $\boldsymbol{\theta}^*$ obtained with a very small positive μ to be a good approximation of $\boldsymbol{\theta}^*$ obtained with $\mu = 0$.

To optimize it, with fixed \mathbf{q} , due to strong convexity, the unconstrained minimization over $\boldsymbol{\theta}$ yields $\boldsymbol{\theta}^* = -\frac{1}{m\lambda} \sum_{i=1}^m (\boldsymbol{\Phi}^{(i)})^\top (\mathbf{q}^{(i)} - \mathbf{p}_{\text{emp}}^{(i)})$. In contrast, the constrained minimization over \mathbf{p} admits no closed-form solution but can be cast as Euclidean projection onto \mathcal{A}_{arb} instead,

independently for each $i \in [m]$: $\mathbf{p}^* = \min_{\mathbf{p} \in \mathcal{A}_{\text{arb}}} \|\mathbf{p} - \frac{1}{\mu} \mathbf{q}^{(i)}\|_2^2 \triangleq \text{Proj}_{\mathcal{A}_{\text{arb}}}(\frac{1}{\mu} \mathbf{q}^{(i)})$. Given $\boldsymbol{\theta}^*$ and \mathbf{p}^* , the outermost maximization can be solved by a projected quasi-Newton algorithm (Schmidt et al., 2009) that also requires the projection oracle $\text{Proj}_{\mathcal{A}_{\text{arb}}}(\cdot)$, elaborated in Section 4.4.

4.3.4 Inference

We propose two algorithms to make inference with given $\boldsymbol{\theta}^*$.

Weight construction. Construct the part weights as $\boldsymbol{\Phi}\boldsymbol{\theta}^* \in \mathbb{R}^{|\mathcal{R}|}$ and find the maximum weight arborescence: $\mathbf{y}^* \in \arg \max_{\mathbf{y}} \mathbf{y}^\top \boldsymbol{\Phi}\boldsymbol{\theta}^*$ by the Gabow-Tarjan (GT) algorithm (Gabow et al., 1986; Zmigrod et al., 2020) or approximate methods for higher-order trees.

Minimum Bayes risk decoding. The optimal probabilistic prediction \mathbb{P}^* or \mathbf{p}^* can be obtained from Equation 4.2 or Equation 4.3. The marginal probabilities enable minimum Bayes risk decoding: $\mathbf{y}^* \in \arg \min_{\mathbf{y}} \mathbb{E}_{\mathbb{P}_{\hat{\mathbf{Y}}|x}^*} \ell(\mathbf{y}, \hat{\mathbf{Y}}) \triangleq \arg \max_{\mathbf{y}} \sum_{s: y_s=1} \mathbf{p}_s^*$, a maximum weight arborescence problem.

4.3.5 Statistical Properties

Basic generalization bounds of DRO methods derived from measure concentration are not appropriate for an ambiguity set defined by low-order moments since it fails to converge (Shafieezadeh-Abadeh et al., 2019). We take an alternate approach following (Farnia and Tse, 2016) to obtain excess out-of-sample risk bounds by assuming boundedness on features and losses.

Theorem 29. *Given m samples, a non-negative loss $\ell(\cdot, \cdot)$ such that $|\ell(\cdot, \cdot)| \leq K$, a feature function $\phi(\cdot, \cdot)$ such that $\|\phi(\cdot, \cdot)\| \leq B$, a positive ambiguity level $\varepsilon > 0$, then, for any $\rho \in (0, 1]$, with a probability at least $1 - \rho$, the following excess true worst-case risk bound holds:*

$$\max_{\mathbb{Q} \in \mathcal{B}(\mathbb{P}^{true})} R_{\mathbb{Q}}^L(\boldsymbol{\theta}_{emp}^*) - \max_{\mathbb{Q} \in \mathcal{B}(\mathbb{P}^{true})} R_{\mathbb{Q}}^L(\boldsymbol{\theta}_{true}^*) \leq \frac{4KB}{\varepsilon\sqrt{m}} \left(1 + \frac{3}{2} \sqrt{\frac{\ln(4/\rho)}{2}} \right),$$

where $\boldsymbol{\theta}_{emp}^*$ and $\boldsymbol{\theta}_{true}^*$ are the optimal parameters learned in Equation 4.2 under \mathbb{P}^{emp} and \mathbb{P}^{true} respectively. The original risk of $\boldsymbol{\theta}$ under \mathbb{Q} is $R_{\mathbb{Q}}^L(\boldsymbol{\theta}) := \mathbb{E}_{\mathbb{Q}_{\mathbf{X}, \mathbf{Y}}, \mathbb{P}_{\hat{\mathbf{Y}}|\mathbf{X}}^{\boldsymbol{\theta}}} \ell(\hat{\mathbf{Y}}, \mathbf{Y})$ with Bayes prediction $\mathbb{P}_{\mathbf{Y}|\mathbf{x}}^{\boldsymbol{\theta}} \in \arg \min_{\mathbb{P}} \max_{\mathbb{Q}} \mathbb{E}_{\mathbb{Q}_{\hat{\mathbf{Y}}|\mathbf{x}}, \mathbb{P}_{\hat{\mathbf{Y}}|\mathbf{x}}} \ell(\hat{\mathbf{Y}}, \check{\mathbf{Y}}) + \boldsymbol{\theta}^\top \phi(\mathbf{x}, \check{\mathbf{Y}})$.

Proof. Define the adversarial surrogate risk of $\boldsymbol{\theta}$ with respect to $\tilde{\mathbb{P}}$ as

$$\begin{aligned} R_{\tilde{\mathbb{P}}}^S(\boldsymbol{\theta}) &:= \mathbb{E}_{\tilde{\mathbb{P}}_{\mathbf{X}, \mathbf{Y}}} \ell_{adv}(\boldsymbol{\theta}, (\mathbf{X}, \mathbf{Y})) := \mathbb{E}_{\tilde{\mathbb{P}}_{\mathbf{X}, \mathbf{Y}}} \min_{\mathbb{P}} \max_{\mathbb{Q}} \mathbb{E}_{\mathbb{Q}_{\hat{\mathbf{Y}}|\mathbf{x}}, \mathbb{P}_{\hat{\mathbf{Y}}|\mathbf{x}}} \ell(\hat{\mathbf{Y}}, \check{\mathbf{Y}}) \\ &\quad + \boldsymbol{\theta}^\top (\phi(\mathbf{X}, \check{\mathbf{Y}}) - \phi(\mathbf{X}, \mathbf{Y})) + \varepsilon \|\boldsymbol{\theta}\|_*. \end{aligned}$$

Let $\boldsymbol{\theta}_{true}^* \in \arg \min_{\boldsymbol{\theta}} R_{\mathbb{P}^{true}}^S(\boldsymbol{\theta})$ and $\boldsymbol{\theta}_{emp}^* \in \arg \min_{\boldsymbol{\theta}} R_{\mathbb{P}^{emp}}^S(\boldsymbol{\theta})$ be the optimal parameters learned with $\mathbb{P}_{\mathbf{X}, \mathbf{Y}}^{true}$ and $\mathbb{P}_{\mathbf{X}, \mathbf{Y}}^{emp}$ respectively.

Given \mathbf{x} , define the decoded prediction by $\boldsymbol{\theta}$ as

$$\mathbb{P}_{\mathbf{Y}|\mathbf{x}}^{\boldsymbol{\theta}} \in \arg \min_{\mathbb{P}} \max_{\mathbb{Q}} \mathbb{E}_{\mathbb{Q}_{\hat{\mathbf{Y}}|\mathbf{x}}, \mathbb{P}_{\hat{\mathbf{Y}}|\mathbf{x}}} \ell(\hat{\mathbf{Y}}, \check{\mathbf{Y}}) + \boldsymbol{\theta}^\top \phi(\mathbf{x}, \check{\mathbf{Y}}).$$

Let the original risk of loss ℓ under some distribution \mathbb{Q} be

$$R_{\mathbb{Q}}^L(\boldsymbol{\theta}) := \mathbb{E}_{\mathbb{Q}_{\mathbf{X}, \mathbf{Y}}, \mathbb{P}_{\hat{\mathbf{Y}}|\mathbf{X}}^{\boldsymbol{\theta}}} \ell(\hat{\mathbf{Y}}, \mathbf{Y}).$$

According to Proposition 28, for any fixed \mathbb{P} , we have similarly

$$\begin{aligned} & \max_{\mathbb{Q} \in \mathcal{B}(\mathbb{P}^{\text{emp}})} \mathbb{E}_{\mathbb{Q}_{\mathbf{X}, \check{\mathbf{Y}}}, \mathbb{P}_{\hat{\mathbf{Y}}|\mathbf{X}}} \ell(\hat{\mathbf{Y}}, \check{\mathbf{Y}}) \\ & \triangleq \min_{\boldsymbol{\theta}} \mathbb{E}_{\mathbb{P}_{\mathbf{X}, \mathbf{Y}}^{\text{emp}}} \max_{\mathbb{Q}} \mathbb{E}_{\mathbb{Q}_{\check{\mathbf{Y}}|\mathbf{X}}, \mathbb{P}_{\hat{\mathbf{Y}}|\mathbf{X}}} \ell(\hat{\mathbf{Y}}, \check{\mathbf{Y}}) + \boldsymbol{\theta}^\top (\phi(\mathbf{X}, \check{\mathbf{Y}}) - \phi(\mathbf{X}, \mathbf{Y})) + \varepsilon \|\boldsymbol{\theta}\|_*. \end{aligned}$$

We start by looking at the worst-case risk of $\boldsymbol{\theta}_{\text{true}}^*$ and $\boldsymbol{\theta}_{\text{emp}}^*$.

$$\begin{aligned} & \max_{\mathbb{Q} \in \mathcal{B}(\mathbb{P}^{\text{true}})} R_{\mathbb{Q}}^L(\boldsymbol{\theta}_{\text{emp}}^*) \\ & = \min_{\boldsymbol{\theta}} \mathbb{E}_{\mathbb{P}_{\mathbf{X}, \mathbf{Y}}^{\text{true}}} \max_{\mathbb{Q}} \mathbb{E}_{\mathbb{Q}_{\check{\mathbf{Y}}|\mathbf{X}}, \mathbb{P}_{\hat{\mathbf{Y}}|\mathbf{X}}^{\boldsymbol{\theta}_{\text{emp}}^*}} \ell(\hat{\mathbf{Y}}, \check{\mathbf{Y}}) + \boldsymbol{\theta}^\top (\phi(\mathbf{X}, \check{\mathbf{Y}}) - \phi(\mathbf{X}, \mathbf{Y})) + \varepsilon \|\boldsymbol{\theta}\|_* \\ & \leq \mathbb{E}_{\mathbb{P}_{\mathbf{X}, \mathbf{Y}}^{\text{true}}} \max_{\mathbb{Q}} \mathbb{E}_{\mathbb{Q}_{\check{\mathbf{Y}}|\mathbf{X}}, \mathbb{P}_{\hat{\mathbf{Y}}|\mathbf{X}}^{\boldsymbol{\theta}_{\text{emp}}^*}} \ell(\hat{\mathbf{Y}}, \check{\mathbf{Y}}) + \boldsymbol{\theta}_{\text{emp}}^* \cdot (\phi(\mathbf{X}, \check{\mathbf{Y}}) - \phi(\mathbf{X}, \mathbf{Y})) + \varepsilon \|\boldsymbol{\theta}_{\text{emp}}^*\|_*, \end{aligned}$$

where the last inequality holds because $\boldsymbol{\theta}_{\text{emp}}^*$ is not necessarily a minimizer. Similarly for $\boldsymbol{\theta}_{\text{true}}^*$,

$$\begin{aligned} & \max_{\mathbb{Q} \in \mathcal{B}(\mathbb{P}^{\text{true}})} R_{\mathbb{Q}}^L(\boldsymbol{\theta}_{\text{true}}^*) \leq \mathbb{E}_{\mathbb{P}_{\mathbf{X}, \mathbf{Y}}^{\text{true}}} \max_{\mathbb{Q}} \mathbb{E}_{\mathbb{Q}_{\check{\mathbf{Y}}|\mathbf{X}}, \mathbb{P}_{\hat{\mathbf{Y}}|\mathbf{X}}^{\boldsymbol{\theta}_{\text{true}}^*}} \ell(\hat{\mathbf{Y}}, \check{\mathbf{Y}}) \\ & \quad + \boldsymbol{\theta}_{\text{true}}^* \cdot (\phi(\mathbf{X}, \check{\mathbf{Y}}) - \phi(\mathbf{X}, \mathbf{Y})) + \varepsilon \|\boldsymbol{\theta}_{\text{true}}^*\|_*. \end{aligned}$$

On the other hand,

$$\begin{aligned}
& \mathbb{E}_{\mathbb{P}_{\mathbf{X}, \mathbf{Y}}^{\text{true}}} \max_{\mathbb{Q}} \mathbb{E}_{\mathbb{Q}_{\check{\mathbf{Y}}|\mathbf{X}} \mathbb{P}_{\check{\mathbf{Y}}|\mathbf{X}}^{\boldsymbol{\theta}_{\text{true}}^*}} \ell(\hat{\mathbf{Y}}, \check{\mathbf{Y}}) + \boldsymbol{\theta}_{\text{true}}^* \cdot (\phi(\mathbf{X}, \check{\mathbf{Y}}) - \phi(\mathbf{X}, \mathbf{Y})) + \varepsilon \|\boldsymbol{\theta}_{\text{true}}^*\|_* \\
&= \min_{\boldsymbol{\theta}} \mathbb{E}_{\mathbb{P}_{\mathbf{X}, \mathbf{Y}}^{\text{true}}} \min_{\mathbb{P}} \max_{\mathbb{Q}} \mathbb{E}_{\mathbb{Q}_{\check{\mathbf{Y}}|\mathbf{X}} \mathbb{P}_{\check{\mathbf{Y}}|\mathbf{X}}} \ell(\hat{\mathbf{Y}}, \check{\mathbf{Y}}) + \boldsymbol{\theta}^\top (\phi(\mathbf{X}, \check{\mathbf{Y}}) - \phi(\mathbf{X}, \mathbf{Y})) + \varepsilon \|\boldsymbol{\theta}\|_* \\
&= \min_{\mathbb{P}} \min_{\boldsymbol{\theta}} \mathbb{E}_{\mathbb{P}_{\mathbf{X}, \mathbf{Y}}^{\text{true}}} \max_{\mathbb{Q}} \mathbb{E}_{\mathbb{Q}_{\check{\mathbf{Y}}|\mathbf{X}} \mathbb{P}_{\check{\mathbf{Y}}|\mathbf{X}}} \ell(\hat{\mathbf{Y}}, \check{\mathbf{Y}}) + \boldsymbol{\theta}^\top (\phi(\mathbf{X}, \check{\mathbf{Y}}) - \phi(\mathbf{X}, \mathbf{Y})) + \varepsilon \|\boldsymbol{\theta}\|_* \\
&\leq \min_{\boldsymbol{\theta}} \mathbb{E}_{\mathbb{P}_{\mathbf{X}, \mathbf{Y}}^{\text{true}}} \max_{\mathbb{Q}} \mathbb{E}_{\mathbb{Q}_{\check{\mathbf{Y}}|\mathbf{X}} \mathbb{P}_{\check{\mathbf{Y}}|\mathbf{X}}^{\boldsymbol{\theta}_{\text{true}}^*}} \ell(\hat{\mathbf{Y}}, \check{\mathbf{Y}}) + \boldsymbol{\theta}^\top (\phi(\mathbf{X}, \check{\mathbf{Y}}) - \phi(\mathbf{X}, \mathbf{Y})) + \varepsilon \|\boldsymbol{\theta}\|_* \\
&= \max_{\mathbb{Q} \in \mathcal{B}(\mathbb{P}^{\text{true}})} R_{\mathbb{Q}}^L(\boldsymbol{\theta}_{\text{true}}^*),
\end{aligned}$$

where the first equality holds according to the definition of $\boldsymbol{\theta}_{\text{true}}^*$. The above two inequalities imply the equality:

$$\begin{aligned}
\max_{\mathbb{Q} \in \mathcal{B}(\mathbb{P}^{\text{true}})} R_{\mathbb{Q}}^L(\boldsymbol{\theta}_{\text{true}}^*) &= \mathbb{E}_{\mathbb{P}_{\mathbf{X}, \mathbf{Y}}^{\text{true}}} \max_{\mathbb{Q}} \mathbb{E}_{\mathbb{Q}_{\check{\mathbf{Y}}|\mathbf{X}} \mathbb{P}_{\check{\mathbf{Y}}|\mathbf{X}}^{\boldsymbol{\theta}_{\text{true}}^*}} \ell(\hat{\mathbf{Y}}, \check{\mathbf{Y}}) \\
&\quad + \boldsymbol{\theta}_{\text{true}}^* \cdot (\phi(\mathbf{X}, \check{\mathbf{Y}}) - \phi(\mathbf{X}, \mathbf{Y})) + \varepsilon \|\boldsymbol{\theta}_{\text{true}}^*\|_*.
\end{aligned}$$

Therefore,

$$\begin{aligned}
& \max_{\mathbb{Q} \in \mathcal{B}(\mathbb{P}^{\text{true}})} R_{\mathbb{Q}}^L(\boldsymbol{\theta}_{\text{emp}}^*) - \max_{\mathbb{Q} \in \mathcal{B}(\mathbb{P}^{\text{true}})} R_{\mathbb{Q}}^L(\boldsymbol{\theta}_{\text{true}}^*) \\
&\leq \mathbb{E}_{\mathbb{P}_{\mathbf{X}, \mathbf{Y}}^{\text{true}}} \max_{\mathbb{Q}} \mathbb{E}_{\mathbb{Q}_{\check{\mathbf{Y}}|\mathbf{X}} \mathbb{P}_{\check{\mathbf{Y}}|\mathbf{X}}^{\boldsymbol{\theta}_{\text{emp}}^*}} \ell(\hat{\mathbf{Y}}, \check{\mathbf{Y}}) + \boldsymbol{\theta}_{\text{emp}}^* \cdot (\phi(\mathbf{X}, \check{\mathbf{Y}}) - \phi(\mathbf{X}, \mathbf{Y})) + \varepsilon \|\boldsymbol{\theta}_{\text{emp}}^*\|_* \\
&\quad - (\mathbb{E}_{\mathbb{P}_{\mathbf{X}, \mathbf{Y}}^{\text{true}}} \max_{\mathbb{Q}} \mathbb{E}_{\mathbb{Q}_{\check{\mathbf{Y}}|\mathbf{X}} \mathbb{P}_{\check{\mathbf{Y}}|\mathbf{X}}^{\boldsymbol{\theta}_{\text{true}}^*}} \ell(\hat{\mathbf{Y}}, \check{\mathbf{Y}}) + \boldsymbol{\theta}_{\text{true}}^* \cdot (\phi(\mathbf{X}, \check{\mathbf{Y}}) - \phi(\mathbf{X}, \mathbf{Y})) + \varepsilon \|\boldsymbol{\theta}_{\text{true}}^*\|_*). \tag{4.4}
\end{aligned}$$

The main idea is thus to use uniform convergence bounds. Firstly, by substituting $\mathbb{Q} = \mathbb{P}^{\text{true}}$, note that

$$\min_{\mathbb{P}} \max_{\mathbb{Q}} \mathbb{E}_{\mathbb{Q}_{\check{Y}|\mathbf{X}} \mathbb{P}_{\check{Y}|\mathbf{X}}} \ell(\hat{\mathbf{Y}}, \check{\mathbf{Y}}) + \boldsymbol{\theta}^\top (\phi(\mathbf{X}, \check{\mathbf{Y}}) - \phi(\mathbf{X}, \mathbf{Y})) \geq \min_{\mathbb{P}} \mathbb{E}_{\mathbb{P}_{\check{Y}|\mathbf{X}}^{\text{true}} \mathbb{P}_{\check{Y}|\mathbf{X}}} \ell(\hat{\mathbf{Y}}, \mathbf{Y}) \geq 0.$$

We can get an upper bound of the norm of any optimal solution $\boldsymbol{\theta}_{\text{true}}^*$ or $\boldsymbol{\theta}_{\text{emp}}^*$ as follows:

$$\begin{aligned} 0 + \varepsilon \|\boldsymbol{\theta}_{\text{true}}^*\|_* &\leq R_{\mathbb{P}^{\text{true}}}^S(\boldsymbol{\theta}_{\text{true}}^*) \leq R_{\mathbb{P}^{\text{true}}}^S(\mathbf{0}) \leq \mathbb{E}_{\mathbb{P}_{\mathbf{X}, \mathbf{Y}}^{\text{true}}} \min_{\mathbb{P}} \max_{\mathbb{Q}} \mathbb{E}_{\mathbb{Q}_{\check{Y}|\mathbf{X}} \mathbb{P}_{\check{Y}|\mathbf{X}}} \ell(\hat{\mathbf{Y}}, \check{\mathbf{Y}}) \leq K \\ &\implies \|\boldsymbol{\theta}_{\text{true}}^*\|_* \leq \frac{K}{\varepsilon}. \end{aligned}$$

Let $\psi(\mathbf{X}, \mathbf{Y}) := \boldsymbol{\theta}^\top \phi(\mathbf{X}, \mathbf{Y})$ and $\boldsymbol{\psi}_x := (\psi(\mathbf{x}, \mathbf{y}))_{\mathbf{y} \in \mathcal{Y}}$. Define

$$\begin{aligned} f(\boldsymbol{\theta}, \tilde{\mathbb{P}}) &:= \mathbb{E}_{\tilde{\mathbb{P}}_{\mathbf{X}, \mathbf{Y}}} \min_{\mathbb{P}} \max_{\mathbb{Q}} \mathbb{E}_{\mathbb{Q}_{\check{Y}|\mathbf{X}} \mathbb{P}_{\check{Y}|\mathbf{X}}} \ell(\hat{\mathbf{Y}}, \check{\mathbf{Y}}) + \boldsymbol{\theta}^\top (\phi(\mathbf{X}, \check{\mathbf{Y}}) - \phi(\mathbf{X}, \mathbf{Y})) \\ &\triangleq \mathbb{E}_{\tilde{\mathbb{P}}_{\mathbf{X}, \mathbf{Y}}} \max_{\mathbb{Q}} \mathbb{E}_{\mathbb{Q}_{\check{Y}|\mathbf{X}} \mathbb{P}_{\check{Y}|\mathbf{X}}^{\boldsymbol{\theta}}} \ell(\hat{\mathbf{Y}}, \check{\mathbf{Y}}) + \boldsymbol{\theta}^\top (\phi(\mathbf{X}, \check{\mathbf{Y}}) - \phi(\mathbf{X}, \mathbf{Y})) \\ &\triangleq \mathbb{E}_{\tilde{\mathbb{P}}_{\mathbf{X}, \mathbf{Y}}} \max_{\mathbb{Q}} \mathbb{E}_{\mathbb{Q}_{\check{Y}|\mathbf{X}} \mathbb{P}_{\check{Y}|\mathbf{X}}^{\boldsymbol{\theta}}} \ell(\hat{\mathbf{Y}}, \check{\mathbf{Y}}) + \psi(\mathbf{X}, \check{\mathbf{Y}}) - \psi(\mathbf{X}, \mathbf{Y}) \\ &\triangleq g(\boldsymbol{\psi}, \tilde{\mathbb{P}}). \end{aligned}$$

Let $\mathbf{q}_x \in \Delta$ be the probability vector of $\mathbb{Q}_{\check{Y}|\mathbf{x}}$ and \mathbf{e}_y be the standard basis vector with y -th entry equal to 1. We have that for any (\mathbf{x}, \mathbf{y}) ,

$$\frac{\partial}{\partial \boldsymbol{\psi}_x} g(\boldsymbol{\psi}, \delta_{(\mathbf{x}, \mathbf{y})}) \subseteq \text{Conv}(\{\mathbf{q}_x - \mathbf{e}_y : \mathbf{q}_x \in \Delta\}) \implies \left\| \frac{\partial}{\partial \boldsymbol{\psi}_x} g(\boldsymbol{\psi}, \delta_{(\mathbf{x}, \mathbf{y})}) \right\|_1 \leq \max_{\mathbf{q}_x \in \Delta} \|\mathbf{q}_x - \mathbf{e}_y\|_1 \leq 2,$$

where $\delta_{(\mathbf{x}, \mathbf{y})}$ is the Dirac point measure. $g(\cdot, \tilde{\mathbb{P}})$ is therefore 2-Lipschitz with respect to the ℓ_1 norm. As per the assumption, $\|\phi(\cdot, \cdot)\| \leq B$. This further implies that

$$f(\boldsymbol{\theta}_1, \delta_{(\mathbf{x}_1, \mathbf{y}_1)}) - f(\boldsymbol{\theta}_2, \delta_{(\mathbf{x}_2, \mathbf{y}_2)}) \leq \frac{4KB}{\varepsilon} \quad \forall \boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \mathbf{x}_1, \mathbf{x}_2, \mathbf{y}_1, \mathbf{y}_2 \quad \text{s.t.} \quad \|\boldsymbol{\theta}_i\|_* \leq \frac{K}{\varepsilon} \quad \forall i = 1, 2.$$

We then follow the proof of Theorem 3 in (Farnia and Tse, 2016). According to Theorem 26.12 in (Shalev-Shwartz and Ben-David, 2014), by uniform convergence, for any $\rho \in (0, 2]$, with a probability at least $1 - \frac{\rho}{2}$,

$$f(\boldsymbol{\theta}_{\text{emp}}^*, \mathbb{P}^{\text{true}}) - f(\boldsymbol{\theta}_{\text{emp}}^*, \mathbb{P}^{\text{emp}}) \leq \frac{4KB}{\varepsilon\sqrt{m}} \left(1 + \sqrt{\frac{\ln(4/\rho)}{2}} \right).$$

According to the definition of $\boldsymbol{\theta}_{\text{true}}^*$, the following inequality holds:

$$f(\boldsymbol{\theta}_{\text{emp}}^*, \mathbb{P}^{\text{emp}}) + \varepsilon\|\boldsymbol{\theta}_{\text{emp}}^*\|_* - f(\boldsymbol{\theta}_{\text{true}}^*, \mathbb{P}^{\text{emp}}) - \varepsilon\|\boldsymbol{\theta}_{\text{true}}^*\|_* \leq 0.$$

Since $\boldsymbol{\theta}_{\text{true}}^*$ do not depend on samples, according to the Hoeffding's inequality, with a probability $1 - \rho/2$,

$$f(\boldsymbol{\theta}_{\text{true}}^*, \mathbb{P}^{\text{emp}}) - f(\boldsymbol{\theta}_{\text{true}}^*, \mathbb{P}^{\text{true}}) \leq \frac{2KB}{\varepsilon\sqrt{m}} \sqrt{\frac{\ln(4/\rho)}{2}}.$$

Applying the union bound to the above three inequations, with a probability $1 - \rho$, we have

$$f(\boldsymbol{\theta}_{\text{emp}}^*, \mathbb{P}^{\text{true}}) + \varepsilon \|\boldsymbol{\theta}_{\text{emp}}^*\|_* - f(\boldsymbol{\theta}_{\text{true}}^*, \mathbb{P}^{\text{true}}) - \varepsilon \|\boldsymbol{\theta}_{\text{true}}^*\|_* \leq \frac{4KB}{\varepsilon\sqrt{m}} \left(1 + \frac{3}{2} \sqrt{\frac{\ln(4/\rho)}{2}} \right).$$

As stated by Equation 4.4, we conclude with the following excess risk bound:

$$\max_{\mathbb{Q} \in \mathcal{B}(\mathbb{P}^{\text{true}})} R_{\mathbb{Q}}^L(\boldsymbol{\theta}_{\text{emp}}^*) - \max_{\mathbb{Q} \in \mathcal{B}(\mathbb{P}^{\text{true}})} R_{\mathbb{Q}}^L(\boldsymbol{\theta}_{\text{true}}^*) \leq \frac{4KB}{\varepsilon\sqrt{m}} \left(1 + \frac{3}{2} \sqrt{\frac{\ln(4/\rho)}{2}} \right).$$

□

Theorem 29 presents a bound based on uniform convergence and Rademacher complexities (Bartlett and Mendelson, 2002), which improves the results in (Asif et al., 2015), who merely show that the worst-case risk upper bounds the risk under any distribution in the ambiguity set.

The dual problem in Equation 4.2 suggests an adversarial surrogate loss $\ell_{\text{adv}}(\boldsymbol{\theta}, (\mathbf{x}, \mathbf{y}))$ in a ERM form. The special case of $\varepsilon = 0$ in our DRO estimator has a similar form to the max-min surrogate loss in (Nowak-Vila et al., 2020) except that we assume probabilistic prediction. A conclusion of its Fisher consistency can thus be drawn based on (Fathony et al., 2018; Nowak-Vila et al., 2020).

Corollary 30. *When $\varepsilon = 0$, ℓ_{adv} is Fisher consistent with respect to ℓ . Namely, $\mathbb{P}_{\hat{\mathbf{Y}}|\mathbf{X}}^{\boldsymbol{\theta}_{\text{true}}^*}$ is the probabilistic prediction made by the Bayes optimal decision rule, where $\boldsymbol{\theta}_{\text{true}}^*$ is defined in Theorem 29.*

Proof. Our formulation differs from (Nowak-Vila et al., 2020) in the fact that we allow probabilistic prediction to be ground truth. By defining $y^*(\mu)$ as the gold standard probabilistic prediction and \mathcal{Y} as the set of all possible probabilistic predictions in Proposition C.2 in (Nowak-Vila et al., 2020), we have

$$\mathbb{P}_{\hat{\mathbf{Y}}|\mathbf{x}}^{\theta_{\text{true}}^*} \in \text{Conv}(\arg \min_{\mathbb{P}_{\hat{\mathbf{Y}}|\mathbf{x}}} \mathbb{E}_{\mathbb{P}_{\mathbf{Y}|\mathbf{x}}^{\text{true}}, \mathbb{P}_{\hat{\mathbf{Y}}|\mathbf{x}}} \ell(\hat{\mathbf{Y}}, \mathbf{Y})).$$

Therefore,

$$\mathbb{P}_{\hat{\mathbf{Y}}|\mathbf{x}}^{\theta_{\text{true}}^*} \in \arg \min_{\mathbb{P}_{\hat{\mathbf{Y}}|\mathbf{x}}} \mathbb{E}_{\mathbb{P}_{\mathbf{Y}|\mathbf{x}}^{\text{true}}, \mathbb{P}_{\hat{\mathbf{Y}}|\mathbf{x}}} \ell(\hat{\mathbf{Y}}, \mathbf{Y}).$$

□

If $\varepsilon > 0$, the decoded prediction for each \mathbf{x} will not belong to the convex hull of true conditional distributions, thus not a minimizer of ℓ . On the other hand, if ε is chosen as $m^{-\alpha}$ for $0 < \alpha < 1/2$, ℓ_{adv} will be universally consistent according to the comparison inequality in (Nowak-Vila et al., 2020).

4.4 Projection onto Arborescence Polytopes

The Euclidean projection onto an r -arborescence polytope is a quadratic programming problem. This is a well-defined convex optimization problem, different from that in differentiable

structured prediction methods (Peng et al., 2018; Mihaylova et al., 2020) which elicit gradients with respect to inputs.

$$\min_{\mathbf{x} \in \mathcal{A}_{\text{arb}}} f(\mathbf{x}) := \|\mathbf{x} - \mathbf{w}\|_2^2.$$

We focus on first-order models and discuss the extensions to other classes of trees in Section 4.5.

4.4.1 Frank-Wolfe Algorithm

The Frank-Wolfe (FW) method (Frank et al., 1956) is an iterative first-order algorithm that enforces constraints by optimizing a linear objective over the feasible set at each iteration t :

$$\mathbf{s}^t \in \arg \min_{\mathbf{s} \in \mathcal{A}_{\text{arb}}} \mathbf{s}^\top \nabla f(\mathbf{x}^t), \quad (4.5)$$

which is a minimum weight arborescence problem with weights $\nabla f(\mathbf{x}^t)$ in our case. The solution is updated and stays feasible: $\mathbf{x}^{t+1} \leftarrow \mathbf{x}^t + \gamma_t(\mathbf{s}^t - \mathbf{x}^t)$, where γ_t is a step size typically set to $\frac{2}{t+2}$. FW style algorithms are known to have a convergence rate of $\mathcal{O}(\frac{1}{\epsilon})$ (Jaggi, 2013).

4.4.2 Martin's Polytope

A compact representation of \mathcal{A}_{arb} with a polynomial number of linear constraints is attractive to lead to efficient algorithms. To the best of our knowledge, there is no existing projection method exploiting special structures of this polytope. An extended formulation of the arborescence polytope (Friesen, 2019; Martin, 1991) follows a lift-and-project approach. It relates each element

to existence of k -arborescences of the underlying undirected graph for all $k \in \mathcal{V}$. We extend it to multi-graphs:

$$\mathcal{A}_{\text{marb}} := \{\mathbf{z}^r : \exists \mathbf{z}^k \geq \mathbf{0} \sum_{a \in \delta^-(j)} z_a^k = \mathbf{1}(j \neq k) \forall k, j \in \mathcal{V} \wedge \sum_{a \in \mathcal{E}'_{ij}} z_a^k = \sum_{a \in \mathcal{E}_{ij}} z_a^r \forall k \neq r, i, j \in \mathcal{V} \wedge \mathbf{z}^r \geq \mathbf{0}\},$$

where $\mathbf{z}^r \in \mathbb{R}^{|\mathcal{E}|}$ is associated with the original arcs \mathcal{E} , $\mathbf{z}^k \in \mathbb{R}^{|\mathcal{E}'|}$ for $k \neq r$ is associated with a simple directed graph $(\mathcal{V}, \mathcal{E}')$ formed by removing directions and splitting each edge $\{i, j\}$ into two directed ones, $\mathcal{E}_{ij} := \{a \in \mathcal{E} : \bar{a} = \{i, j\}\}$ is the set of arcs connecting i and j with $\bar{a} \triangleq \overline{(i, j, l)} := \{i, j\}$ denoting the underlying undirected edge. We show exact correspondence between $\mathcal{A}_{\text{marb}}$ and \mathcal{A}_{arb} based on a similar argument for simple graphs (Friesen, 2019):

Proposition 31. *Let \mathcal{G} be a multi-graph. $\mathcal{A}_{\text{marb}} \triangleq \mathcal{A}_{\text{arb}}$.*

Proof. We follow the proof of (Friesen, 2019) for simple graphs. Recall the definition of $\mathcal{A}_{\text{marb}}$:

$$\mathcal{A}_{\text{marb}} := \{\mathbf{z}^r : \exists \mathbf{z} \geq \mathbf{0}$$

$$\sum_{a \in \delta^-(j)} z_a^k = \mathbf{1}(j \neq k) \forall k, j \in \mathcal{V} \wedge \quad (4.6)$$

$$\sum_{a \in \mathcal{E}'_{ij}} z_a^k = \sum_{a \in \mathcal{E}_{ij}} z_a^r \quad \forall k \neq r, i, j \in \mathcal{V}\}. \quad (4.7)$$

On one hand, given a legal r -arborescence with characteristic vector \mathbf{z}^r , Equation 4.6 and Equation 4.7 hold by the definition of arborescences. The equality also holds for a convex combination of the characteristic vectors of r -arborescences.

On the other hand, given $\mathbf{z} \in \mathcal{A}_{\text{marb}}$. Consider Edmond's definition of r -arborescence polytope based on rank constraints:

$$\sum_{a \in S} x_a \leq |S| - 1 \quad \forall S \subset \mathcal{V} \text{ with } S \neq \emptyset \quad (4.8)$$

$$\sum_{a \in \delta^-(j)} x_a = \mathbf{1}(j \neq r) \quad \forall j \in \mathcal{V} \quad (4.9)$$

$$\mathbf{x} \geq \mathbf{0}.$$

We have Equation 4.6 directly implies Equation 4.9. According to Equation 4.7,

$$\sum_{a \in S} z_a^r = \sum_{a \in S} z_a^u \quad \forall S \subseteq \mathcal{V} \wedge u \in \mathcal{V}.$$

Therefore,

$$\sum_{a \in S} z_a^r = \sum_{a \in S} z_a^u \leq \sum_{j \in S} \sum_{a \in \delta^-(j)} z_a^u = |S| - 1 \quad \forall S \subseteq \mathcal{V} \wedge u \in S,$$

which is exactly Equation 4.8. □

To solve $\min_{\mathbf{x} \in \mathcal{A}_{\text{marb}}} \|\mathbf{x} - \mathbf{w}\|_2^2$, we propose to adopt the alternating direction method of multipliers (ADMM) and rewrite it into the following separable form:

$$\begin{aligned} \min_{\mathbf{u}} g(\mathbf{u}) &:= \sum_{k \in \mathcal{V}} \frac{1}{|\mathcal{V}|} \|\mathbf{u}_k - \mathbf{w}\|_2^2 + I_{\mathcal{U}_k}(\mathbf{u}_k) \\ \text{s.t. } \mathcal{U}_k &:= \{\mathbf{x} \in \mathbb{R}^{|\mathcal{E}|} : \exists \mathbf{z} \in \mathbb{R}_{\geq 0}^{|\mathcal{E}'|} \sum_{a \in \delta^-(j)} z_a = \mathbb{1}(j \neq k) \wedge \sum_{a \in \mathcal{E}'_{ij}} z_a = \sum_{a \in \mathcal{E}_{ij}} x_a \ \forall i, j \in \mathcal{V}\} \\ \mathbf{u}_r &= \mathbf{u}_k \quad \forall k \in \mathcal{V} \setminus r, \quad \mathcal{U}_r := \{\mathbf{x} \in \mathbb{R}_{\geq 0}^{|\mathcal{E}|} : \sum_{a \in \delta^-(j)} x_a = \mathbb{1}(j \neq r) \ \forall j \in \mathcal{V}\}, \end{aligned}$$

where $I_{\mathcal{U}}(\cdot)$ is the characteristic function with $I_{\mathcal{U}}(\mathbf{x}) = 0$ if $\mathbf{x} \in \mathcal{U}$ and ∞ otherwise.

Let $\boldsymbol{\lambda}'_k$ be the dual variables and $\boldsymbol{\lambda}_k := \frac{1}{\rho_k} \boldsymbol{\lambda}'_k$. The scaled augmented Lagrangian function is $L_\rho(\mathbf{u}, \boldsymbol{\lambda}) = g(\mathbf{u}) + \sum_{k \neq r} \frac{\rho_k}{2} \|\mathbf{u}_r - \mathbf{u}_k + \boldsymbol{\lambda}_k\|_2^2 - \frac{\rho_k}{2} \|\boldsymbol{\lambda}_k\|_2^2$.

The ADMM algorithm updates the parameters as follows:

$$\begin{aligned} \mathbf{u}_k^{t+1} &:= \arg \min_{\mathbf{u}_k \in \mathcal{U}_k} L_\rho((\mathbf{u}_r^t, \mathbf{u}_k^t), \boldsymbol{\lambda}^t) \triangleq \text{Proj}_{\mathcal{U}_k} \left(\frac{2\mathbf{w} + \rho_k |\mathcal{V}| (\mathbf{u}_r^t + \boldsymbol{\lambda}_k^t)}{2 + \rho_k |\mathcal{V}|} \right) \quad \forall k \neq r \\ \mathbf{u}_r^{t+1} &:= \arg \min_{\mathbf{u}_r \in \mathcal{U}_r} L_\rho((\mathbf{u}_r^t, \mathbf{u}_k^{t+1}), \boldsymbol{\lambda}^t) \triangleq \text{Proj}_{\mathcal{U}_r} \left(\frac{2\mathbf{w} + |\mathcal{V}| \sum_{k \neq r} \rho_k (\mathbf{u}_k^{t+1} - \boldsymbol{\lambda}_k^t)}{2 + |\mathcal{V}| \sum_{k \neq r} \rho_k} \right) \\ \boldsymbol{\lambda}_k^{t+1} &:= \boldsymbol{\lambda}_k^t + (\mathbf{u}_r^{t+1} - \mathbf{u}_k^{t+1}) \quad \forall k \neq r. \end{aligned}$$

This decomposes the original projection problem into simpler projection problems. Projection onto \mathcal{U}_k for $k = r$ decomposes over $j \in \mathcal{V}$ into $|\mathcal{V}|$ projections onto simplex, solvable as fast as

$\mathcal{O}(n)$ in the worst case (Condat, 2016). For $k \neq r$, computation of \mathbf{u}_k^{t+1} can be done in parallel.

The Lagrange dual problem of $\text{Proj}_{\mathcal{U}_k}(\cdot)$ can be written as

$$\max_{\boldsymbol{\alpha} \in \mathbb{R}^{|\mathcal{V}|}} \sum_{\{i,j\} \in \bar{\mathcal{E}}} h_{ij}(\boldsymbol{\alpha}) - \sum_{j \neq k} \alpha_j \quad \text{s.t. } h_{ij}(\boldsymbol{\alpha}) = \begin{cases} w_{ij}^2/n_{ij} & \text{if } \alpha_{ij} > 2w_{ij}/n_{ij}, \\ -n_{ij}\alpha_{ij}^2/4 + \alpha_{ij}w_{ij} & \text{if } \alpha_{ij} \leq 2w_{ij}/n_{ij}, \end{cases}$$

where $w_{ij} := \sum_{a \in \mathcal{E}_{ij}} w_a$, $n_{ij} := |\mathcal{E}_{ij}|$, $\alpha_{ij} := \min(\alpha_i, \alpha_j)$ and $\alpha_k := +\infty$. Strong duality holds by linear constraint qualification. Primal solutions are recovered by $x_a^* = w_a - \min(\alpha_a^*/2, w_a/n_a)$.

Convergence. The dual objective of $\text{Proj}_{\mathcal{U}_k}(\cdot)$ is strongly concave on $\{\boldsymbol{\alpha} \in \mathbb{R}^{|\mathcal{V}|} : \forall i \exists j \{i, j\} \in \bar{\mathcal{E}} \wedge \alpha_i \leq \alpha_j \wedge \alpha_i \leq 2w_{ij}/n_{ij}\}$, with a unique global maximizer. This implies fast convergence in practice given good initialization. The negative Lagrange dual function has restricted strong convexity with $\nu = \min_{ij}(n_{ij}/2)$, near the optimum, suggesting linear convergence (Zhang and Cheng, 2015). Alternatively, exact solutions can be found by enumerating rankings (with duplicates) of $\boldsymbol{\alpha}$ in $\mathcal{O}(|\mathcal{V}|^{|\mathcal{V}|})$. In this manner, the ADMM algorithm with a strongly convex objective has a linear convergence rate $\mathcal{O}(\log \frac{1}{\epsilon})$ with either exact (Deng and Yin, 2016) or linearly convergent approximate solution (Hager and Zhang, 2020) of $\text{Proj}_{\mathcal{U}_k}(\cdot)$. Using Nesterov's accelerated gradient algorithm (Nesterov, 2003) to optimize Equation 4.3 leads to iteration complexity $\mathcal{O}(C \log \frac{1}{\epsilon})$ with constant C dependent on Lipschitz constants of gradients and μ .

4.5 Extensions

4.5.1 Undirected Spanning Trees

A straight-forward way of extending to undirected spanning trees is to split $\{i, j\}$ into two arcs (i, j) , (j, i) and make the feature mapping direction-invariant, i.e., $\phi(\mathbf{x}, y_s) = \phi(\mathbf{x}, y_{s'})$

for s and s' having the same underlying undirected graph. We post-process the prediction by removing directions.

Alternatively, we seek projection oracles for undirected graphs. Projection via FW is done by using any minimum spanning tree algorithm in Equation 4.5. For ADMM, the formulation in (Martin, 1991) is originally for undirected trees: $\mathcal{A}_{\text{mund}} := \{\mathbf{x} : \exists \mathbf{z} \geq \mathbf{0} \sum_{a \in \delta^-(j)} z_a^k = \mathbf{1}(j \neq k) \wedge z_{ij}^k + z_{ji}^k = x_{\{i,j\}} \forall k, i, j \in \mathcal{V}\}$. ADMM is easily adapted to this case with $\sum_{a \in \mathcal{E}_{ij}} x_a$ replaced by $x_{\{i,j\}}$.

4.5.2 Dependency Trees

The spanning tree structure in dependency parsing is a special one where the outdegree of root is restricted to be one. We can use the GT algorithm for inference with either the same training objective or an aligned objective where a dependency tree polytope is considered: $\mathcal{A}_{\text{dep}}(\mathbf{x}) := \text{Conv}(\{\mathbf{y} \in \mathcal{Y}(\mathbf{x}) : |\delta^+(r)| = 1\})$. A straightforward extension of $\mathcal{A}_{\text{marb}}$ to characterizing dependency trees is $\mathcal{A}_{\text{mdep}} := \{\mathbf{z}^r : \mathbf{z}^r \in \mathcal{A}_{\text{marb}} \wedge \sum_{a \in \delta^+(r)} z_a^r = 1\}$, equivalent to \mathcal{A}_{dep} by the following proposition:

Proposition 32. *Let \mathcal{G} be a multi-graph. $\mathcal{A}_{\text{mdep}} \triangleq \mathcal{A}_{\text{dep}}$.*

Proof. Recall the definition of $\mathcal{A}_{\text{mdep}}$:

$$\mathcal{A}_{\text{mdep}} := \{\mathbf{z}^r : \mathbf{z}^r \in \mathcal{A}_{\text{marb}} \wedge \sum_{a \in \delta^+(r)} z_a^r = 1\}. \quad (4.10)$$

On one hand, given a legal dependency tree $\mathbf{z}^r \in \mathcal{A}_{\text{dep}}$, it satisfies Equation 4.6 and Equation 4.7 by Proposition 31. It also satisfies Equation 4.10 by the definition of \mathcal{A}_{dep} .

On the other hand, given $\mathbf{z}^r \in \mathcal{A}_{\text{mdep}}$, firstly, \mathbf{z}^r must be in \mathcal{A}_{arb} by Proposition 31, which implies that we can write it as a convex combination of k r -arborescences vectors: $\mathbf{z}^r \triangleq \alpha_1 \mathbf{t}^1 + \alpha_2 \mathbf{t}^2 + \dots + \alpha_k \mathbf{t}^k$. All of them are legal r -arborescences, so $\sum_{a \in \delta^+(r)} t_a^i \geq 1$ for all $i \in [k]$. Now if $\sum_{a \in \delta^+(r)} t_a^i > 1$ for some i , we would have a contradiction, $\sum_{a \in \delta^+(r)} z_a^r > 1$. \square

FW methods leverage the GT algorithm in Equation 4.5. As for ADMM, the dual problem of projection onto $\mathcal{U}'_r := \{\mathbf{x} : \mathbf{x} \in \mathcal{U}_r \wedge \sum_{a \in \delta^+(r)} x_a = 1\}$ becomes

$$\max_{\alpha, \beta} \sum_{a \in \mathcal{E}} h_a(\alpha, \beta) - \sum_{j \neq r} \alpha_j - \beta \quad \text{s.t.} \quad h_a(\alpha, \beta) = \begin{cases} w_a^2 & \gamma_a > 2w_a, \\ w_a \gamma_a - \gamma_a^2/4 & \gamma_a \leq 2w_a, \end{cases}$$

where $\gamma_{(i,j,l)} := \alpha_j + \mathbf{1}(i = r)\beta$. This can be solved in $\mathcal{O}(|\mathcal{E}| \log |\mathcal{E}|)$ (Zhang et al., 2010). Recall that the dual problem of projection onto $\mathcal{U}'_r := \{\mathbf{x} : \mathbf{x} \in \mathcal{U}_r \wedge \sum_{a \in \delta^+(r)} x_a = 1\}$ is

$$\max_{\alpha, \beta} \sum_{a \in \mathcal{E}} h_a(\alpha, \beta) - \sum_{j \neq r} \alpha_j - \beta \quad \text{s.t.} \quad h_a(\alpha, \beta) = \begin{cases} w_a^2 & \gamma_a > 2w_a, \\ w_a \gamma_a - \gamma_a^2/4 & \gamma_a \leq 2w_a, \end{cases}$$

where $\gamma_{(i,j,l)} := \alpha_j + \mathbf{1}(i = r)\beta$. Following (Zhang et al., 2010) similarly, we sort $2w_{(i,j,l)}$ for each j and compute the optimal α_j^* with $\beta = 0$. Let the sorted w 's be $(w_1^{(j)}, \dots, w_n^{(j)})$ for each j . We blend create a set $\{w_x^{(j)} - \alpha_j^*\}$ for all j and x . Let the sorted sequence be $-\infty = t_1 < t_2 < \dots < t_{n_t} = \infty$. The derivative with respect to β is piecewise-linear in each

interval $[t_k, t_{k+1}]$. Since the objective is concave in β , we can iterate over all the intervals or find the optimal β^* with binary search.

4.5.3 Higher-order Polytope

Compact higher-order polytope descriptions exist for undirected spanning trees but are still unknown for arborescences with even one monomial (Friesen, 2019). FW requires a linear oracle that is NP-hard to solve exactly in higher-order settings (McDonald and Pereira, 2006).

Instead, we can approximate it with a local polytope where the marginal probabilities of each part s is required to be locally consistent with that of each arc a . For simplicity, we consider only features for the all-true assignments, i.e., all arcs exist in part s . The resulting polytope can be written as $\mathcal{A}_{\text{mloc}} := \{\mathbf{x} : \mathbf{x}_{\mathcal{E}} \in \mathcal{A}_{\text{marb}} \wedge \forall s \in \mathcal{R}, a \in s \quad p_s \leq p_a\}$, which suggests an ADMM algorithm with additional constraint sets for each part: $\mathcal{U}_s := \{\mathbf{x} \in \mathbb{R}_{\geq 0}^{|\mathcal{R}|} : x_s \leq x_a \quad \forall a \in s\}$, the projection onto which can be done in $\mathcal{O}(|s| \log |s|)$. The central problem is the projection onto

$$\mathcal{U}_s := \{\mathbf{x} \in \mathbb{R}_{\geq 0}^{|\mathcal{R}|} : x_s \leq x_a \quad \forall a \in s\}.$$

The only variables of interest are x_a and x_s , given x_s , the optimal x_a is simply $x_a^* = \max(w_a, x_s)$.

We can sort $(w_a, w_s)_{a \in s}$ and enumerate the range x_s takes over this set.

4.6 Experiments

We evaluate our proposed method on dependency parsing tasks and compare its ability to *BiAF* (Dozat and Manning, 2017), arguably the state-of-the-art neural dependency parser. We implement our methods in Python and C. Our code is publicly available (<https://github.com/>

DanielLee/drtreesp). We leverage the implementations in SuPar (<https://github.com/yzhangcs/parser>) (Zhang et al., 2020) for the baseline. All experiments are conducted on a computer with an Intel Core i7 CPU (2.7 GHz) and an NVIDIA Tesla P100 GPU (16 GB).

We adopt three public datasets, the English Penn Treebank (PTB v3.0) (Marcus et al., 1993), the Penn Chinese Treebank (CTB v5.1) (Xue et al., 2002), the Dutch Lassy Small Treebank and the Turkish Treebank in Universal Dependencies (UD v2.3) (Nivre et al., 2016). We follow conventions in (Chen and Manning, 2014; Dyer et al., 2015) to prepare our data. We make standard train/validation/test splits. We use Stanford Dependencies (SD v3.3.0) (De Marneffe and Manning, 2008) to convert dependencies in PTB and CTB. The predicted POS tags with Stanford POS tagger (Toutanova et al., 2003) are adopted for PTB whereas gold POS tags are adopted for CTB and UD. Punctuation is excluded during evaluation. A token is a punctuation if its gold POS tag is space, semi-colon, comma or period for English and PU for Chinese.

Representation learning is not the focus of this work. We follow (Levy et al., 2020) and compare our method with the last biaffine classification layer in *BiAF* on top of pretrained features preceding this layer (backbone’s output). The pretrained embeddings produced by complicated non-linear models make Fisher consistency’s assumption of optimizing over all measurable functions less violated. To featurize the data, for each dataset, we train a *BiAF* network with the whole training set to obtain a pretrained model. Note that this may create unfair advantages for the baseline because the last layer was optimized together with the backbone network in an end-to-end manner during pretraining. Moreover, pretraining uses a standard ERM objective with the cross-entropy loss and local normalization over head nodes. The

pretrained features are thus more adequate for the ERM objective than for our DRO objective. The pretrained models are trained with the suggested hyperparameters in SuPar. The pretrained models achieve 97.25%, 91.91% and 94.78% UAS on PTB, CTB and UD Dutch respectively, where RoBERTa (Liu et al., 2019), ELECTRA (Cui et al., 2020) and XLM-RoBERTa (Conneau et al., 2020) are adopted as encoders. No BERT embeddings are adopted for the UD Turkish dataset. To make use of the features as inputs in our method, we take the outer product of the embedding vectors for two nodes as the arc feature vector. Our method and the biaffine layer therefore share the same number of parameters (501×501 , including bias terms). We focus on predicting the unlabeled dependency tree while relying on pretrained models for relation label prediction. The evaluation criteria are the labeled/unlabeled attachment scores (LAS/UAS) and labeled/unlabeled complete matches (LCM/UCM). The attachment score can be transformed to the Hamming loss with linear mapping: $AS(\mathbf{y}, \mathbf{y}') \triangleq |\mathcal{V}| - 1 - \ell(\mathbf{y}, \mathbf{y}')/2$.

Full batch learning is adopted for *Marginal* (Equation 4.3). Mini-batch training is adopted for *Game*, the game-theoretic algorithm, and *Stochastic*, which solves the inner min-max problem in Equation 4.2 using Equation 4.3 with fixed $\boldsymbol{\theta}$. All models are trained with the training set only. The optimal hyperparameters and parameters are chosen based on the validation set. For our ADMM algorithm, we adopt the adaptive scheme of varying penalty parameters ($\tau_{\text{incr}} = \tau_{\text{decr}} = 1.1$, $\mu = 1$) in (Boyd et al., 2011) and the stopping criterion ($\epsilon_{\text{tol}} = 10^{-2}$) for consensus ADMM in (Xu et al., 2017). In FW, the learning rate is set to $\frac{2}{t+2}$. The smoothness weight μ and ambiguity radius $\lambda = 2\varepsilon$ are tuned using a logarithmic scale on $[10^{-7}, 1]$. The batch size for the game-theoretic algorithm is 10. The batch size for *Stochastic* is 200. The error

TABLE II: Comparison of mean UAS and execution time under different training set sizes. Time refers to the CPU time taken to finish one gradient descent step. Statistically significant differences compared to *BiAF* are marked with † (paired t-test, $p < 0.05$). The best UAS are highlighted in bold.

Method	Time (s)	PTB				CTB				UD Dutch				UD Turkish (low resource)			
		m = 10	50	100	1000	m = 10	50	100	1000	m = 10	50	100	1000	m = 10	50	100	1000
BiAF	0.34	93.48	96.87	96.95	97.16	88.45	90.89	91.15	91.70	90.86	93.80	94.15	94.98	17.64	26.59	30.75	42.82
Marginal	0.28	94.51†	96.81†	96.92	97.12	89.19†	91.03†	91.27	91.67	92.41 †	94.22†	94.50†	95.15 †	24.85†	32.83 †	33.75 †	43.18
Stochastic	2.72	94.62 †	96.81	96.93	97.14	89.27 †	91.03†	91.27	91.66	92.40†	94.23†	94.47	95.14†	25.06 †	31.35†	33.62†	41.20†
Game	7.25	94.51†	96.86	96.92	97.08†	89.22†	91.06 †	91.22	91.57†	92.32†	94.34 †	94.59 †	95.01	19.85	23.18†	27.12†	36.30†

tolerance in *Game* is set to 10^{-2} . In stochastic gradient training, we use Adam with $lr = 10^{-2}$, $\beta_1 = 0.9$, $\beta_2 = 0.999$, $\epsilon = 10^{-8}$. In our experiments, for efficiency, we again adopt the FW algorithm for the outer maximization in *Marginal*.

To showcase the ability of DRO methods tackling scarce data, in each run, we randomly draw $m \in \{10, 50, 100, 1000\}$ samples without replacement from the training set and keep the original validation and test sets. All the models are trained on the same set of sampled data. The process is repeated 5 times for each m . The main UAS results on the PTB, CTB and UD Dutch Lassy Small datasets are reported in Table II with complete results provided in Table III. Our methods consistently deliver higher UAS than *BiAF* especially with a small amount of data (The UAS is high with 10 training samples possibly because (1) the backbone sub-network and linear layer were trained together with the whole training set; (2) BERT embeddings yield data representation that is easily linearly separable; (3) 10 samples result in as many as $10 \times 20 \times 20$ balanced head-selection instances for *BiAF*). With little training data, DRO approaches minimize the worst-case risk to avoid overfitting. With more training data available, our method is still comparable to *BiAF* which is not significantly better than our methods by

TABLE III: Comparison of mean UAS, LAS, UCM and LCM under different training set sizes. Statistically significant differences compared to BiAF are marked with † (paired t-test, $p < 0.05$). We highlight in bold the best results among the four methods.

Dataset	# train	Metric	BiAF	Marginal	Stochastic	Game
PTB	10	UAS	93.48 ± 2.30	94.51 ± 1.71†	94.62 ± 1.60†	94.51 ± 1.75†
		LAS	92.02 ± 2.26	93.04 ± 1.69†	93.14 ± 1.58†	93.04 ± 1.73†
		UCM	47.17 ± 10.28	52.30 ± 8.71†	52.62 ± 8.18†	52.50 ± 8.60†
		LCM	39.73 ± 7.96	43.63 ± 6.71†	43.97 ± 6.39†	43.86 ± 6.58†
	50	UAS	96.87 ± 0.06	96.81 ± 0.05†	96.81 ± 0.05	96.86 ± 0.05
		LAS	95.34 ± 0.06	95.28 ± 0.05†	95.28 ± 0.05	95.33 ± 0.05
		UCM	67.65 ± 0.81	67.38 ± 0.62	67.18 ± 0.79	67.73 ± 0.64
		LCM	55.46 ± 0.59	54.93 ± 0.56†	54.79 ± 0.59†	55.17 ± 0.49
	100	UAS	96.95 ± 0.05	96.92 ± 0.06	96.93 ± 0.05	96.92 ± 0.03
		LAS	95.42 ± 0.05	95.39 ± 0.06	95.40 ± 0.04	95.39 ± 0.02
		UCM	68.79 ± 0.42	68.27 ± 0.72	68.36 ± 0.41	68.29 ± 0.34
		LCM	56.21 ± 0.14	55.68 ± 0.56	55.67 ± 0.45	55.66 ± 0.33
	1000	UAS	97.16 ± 0.02	97.12 ± 0.03	97.14 ± 0.02	97.08 ± 0.03†
		LAS	95.63 ± 0.03	95.59 ± 0.02	95.60 ± 0.02	95.55 ± 0.03†
		UCM	70.99 ± 0.23	70.59 ± 0.49	70.61 ± 0.32	69.94 ± 0.34†
		LCM	57.57 ± 0.09	57.18 ± 0.28†	57.24 ± 0.28†	56.80 ± 0.23†
CTB	10	UAS	88.45 ± 0.67	89.19 ± 0.38†	89.27 ± 0.33†	89.22 ± 0.39†
		LAS	84.79 ± 0.62	85.50 ± 0.35†	85.58 ± 0.30†	85.53 ± 0.36†
		UCM	35.21 ± 1.67	36.83 ± 1.20	37.14 ± 0.94†	36.95 ± 1.23†
		LCM	25.86 ± 0.87	26.82 ± 0.62	26.95 ± 0.59†	26.95 ± 0.63†
	50	UAS	90.89 ± 0.10	91.03 ± 0.05†	91.03 ± 0.05†	91.06 ± 0.05†
		LAS	87.08 ± 0.10	87.20 ± 0.05†	87.20 ± 0.05†	87.23 ± 0.06†
		UCM	42.54 ± 0.24	42.92 ± 0.24†	42.86 ± 0.12†	42.99 ± 0.30
		LCM	29.70 ± 0.23	29.69 ± 0.36	29.72 ± 0.38	29.79 ± 0.23
	100	UAS	91.15 ± 0.16	91.27 ± 0.08	91.27 ± 0.10	91.22 ± 0.05
		LAS	87.32 ± 0.14	87.42 ± 0.06	87.42 ± 0.08	87.37 ± 0.05
		UCM	43.41 ± 0.35	43.91 ± 0.27†	43.86 ± 0.43†	43.81 ± 0.22
		LCM	30.02 ± 0.22	30.27 ± 0.25	30.23 ± 0.28	30.26 ± 0.26
	1000	UAS	91.70 ± 0.04	91.67 ± 0.03	91.66 ± 0.03	91.57 ± 0.03†
		LAS	87.84 ± 0.04	87.80 ± 0.03	87.79 ± 0.03	87.70 ± 0.03†
		UCM	45.80 ± 0.27	45.43 ± 0.11†	45.41 ± 0.12†	45.36 ± 0.27†
		LCM	31.14 ± 0.19	31.11 ± 0.18	31.08 ± 0.17	31.20 ± 0.11
UD Dutch	10	UAS	90.86 ± 1.23	92.41 ± 0.94†	92.40 ± 0.91†	92.32 ± 1.03†
		LAS	86.54 ± 1.26	88.10 ± 0.95†	88.08 ± 0.91†	87.99 ± 1.00†
		UCM	64.11 ± 2.18	67.26 ± 2.16†	67.21 ± 1.91†	67.26 ± 1.97†
		LCM	48.33 ± 1.88	50.32 ± 1.75†	50.48 ± 1.45†	50.46 ± 1.30†
	50	UAS	93.80 ± 0.43	94.22 ± 0.26†	94.23 ± 0.18†	94.34 ± 0.24†
		LAS	89.36 ± 0.33	89.79 ± 0.21†	89.79 ± 0.12†	89.89 ± 0.18†
		UCM	70.57 ± 1.52	72.42 ± 0.90†	72.05 ± 0.99	72.60 ± 1.39
		LCM	52.40 ± 0.61	53.47 ± 0.62†	53.40 ± 0.59	53.58 ± 0.76
	100	UAS	94.15 ± 0.18	94.50 ± 0.18†	94.47 ± 0.13	94.59 ± 0.12†
		LAS	89.69 ± 0.18	90.04 ± 0.15†	90.01 ± 0.12	90.12 ± 0.10†
		UCM	71.71 ± 0.92	73.24 ± 0.88†	73.01 ± 0.99	73.63 ± 0.75†
		LCM	53.01 ± 0.81	53.79 ± 0.40	53.70 ± 0.55	54.13 ± 0.44†
	1000	UAS	94.98 ± 0.07	95.15 ± 0.10†	95.14 ± 0.11†	95.01 ± 0.05
		LAS	90.44 ± 0.06	90.59 ± 0.08†	90.59 ± 0.08†	90.44 ± 0.06
		UCM	74.73 ± 0.33	75.87 ± 0.63†	75.64 ± 0.57†	75.41 ± 0.56
		LCM	54.59 ± 0.13	55.21 ± 0.17†	55.16 ± 0.21†	54.70 ± 0.22
UD Turkish	10	UAS	17.64 ± 2.45	24.85 ± 2.35†	25.06 ± 0.58†	19.85 ± 0.46
		LAS	4.86 ± 2.74	5.33 ± 2.97	5.40 ± 2.85	5.02 ± 3.04
		UCM	7.69 ± 1.72	9.03 ± 1.33	7.88 ± 2.27	10.03 ± 0.54
		LCM	1.46 ± 1.03	1.50 ± 1.07	1.50 ± 1.07	1.74 ± 1.38
	50	UAS	26.59 ± 2.37	32.83 ± 1.50†	31.35 ± 1.10†	23.18 ± 2.03†
		LAS	10.14 ± 0.57	10.73 ± 0.86	10.74 ± 0.54	10.10 ± 0.69
		UCM	10.03 ± 1.31	10.63 ± 0.50	10.81 ± 0.50	10.34 ± 0.36
		LCM	3.24 ± 0.31	3.26 ± 0.24	3.38 ± 0.27	3.43 ± 0.27
	100	UAS	30.75 ± 1.13	33.75 ± 0.86†	33.62 ± 1.49†	27.12 ± 1.25†
		LAS	10.84 ± 0.80	11.48 ± 0.75	11.69 ± 0.67†	10.48 ± 0.70†
		UCM	11.61 ± 1.22	11.30 ± 0.29	11.34 ± 0.26	11.08 ± 0.44
		LCM	3.53 ± 0.60	3.61 ± 0.31	3.57 ± 0.23	3.55 ± 0.23
	1000	UAS	42.82 ± 1.82	43.18 ± 1.73	41.20 ± 2.17†	36.30 ± 2.79†
		LAS	18.44 ± 1.00	18.24 ± 1.62	18.13 ± 1.13	16.38 ± 1.20†
		UCM	15.86 ± 0.40	15.18 ± 0.81	13.78 ± 0.30†	13.52 ± 0.43†
		LCM	4.49 ± 0.47	4.37 ± 0.46	4.31 ± 0.41†	4.29 ± 0.38†

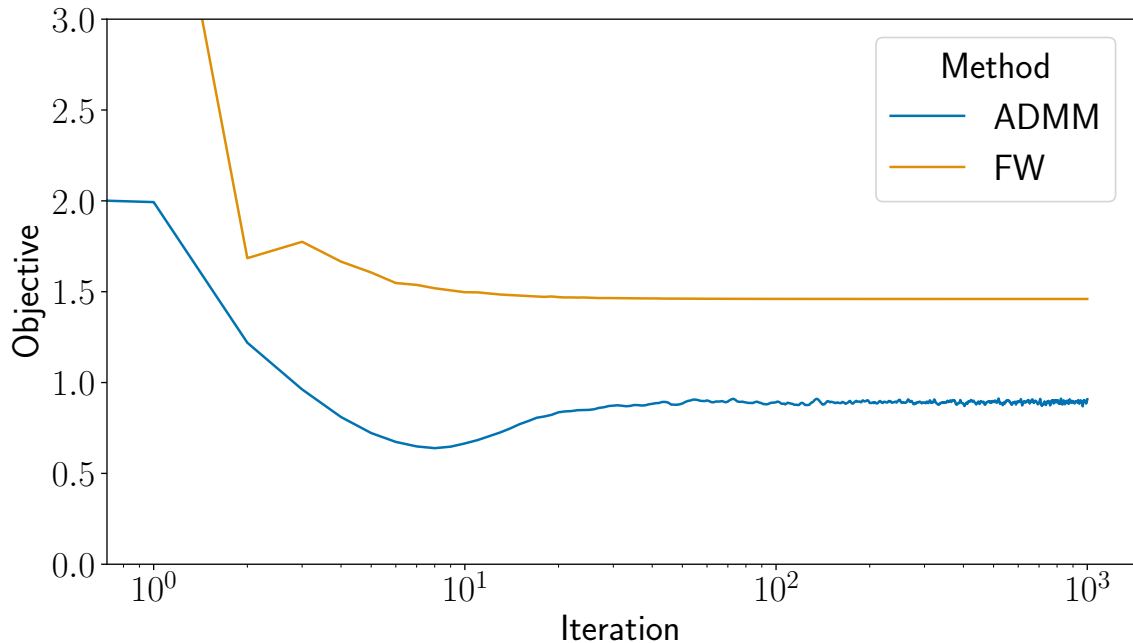


Figure 4: Convergence of ADMM and FW for random points with 95% confidence intervals.

statistical tests. This illustrates the advantages of replacing conditional log-likelihood with our Fisher consistent surrogate loss without changing the number of model parameters. Moreover, we study a low-resource setting with the UD Turkish dataset in which only the sampled data is used for pretraining without BERT embeddings. The binary cross-entropy loss (single normalization) is adopted during pretraining in this setting to avoid pretrained features biased towards the multi-class cross-entropy loss (local normalization) adopted by *BiAF*. We observe consistently competitive performance of our methods in the low-resource setting in Table II as well.

We report computational time of one gradient descent step in the second column of Table II, averaged across 10 runs. For fair comparisons, all the models are run with CPU only, with a

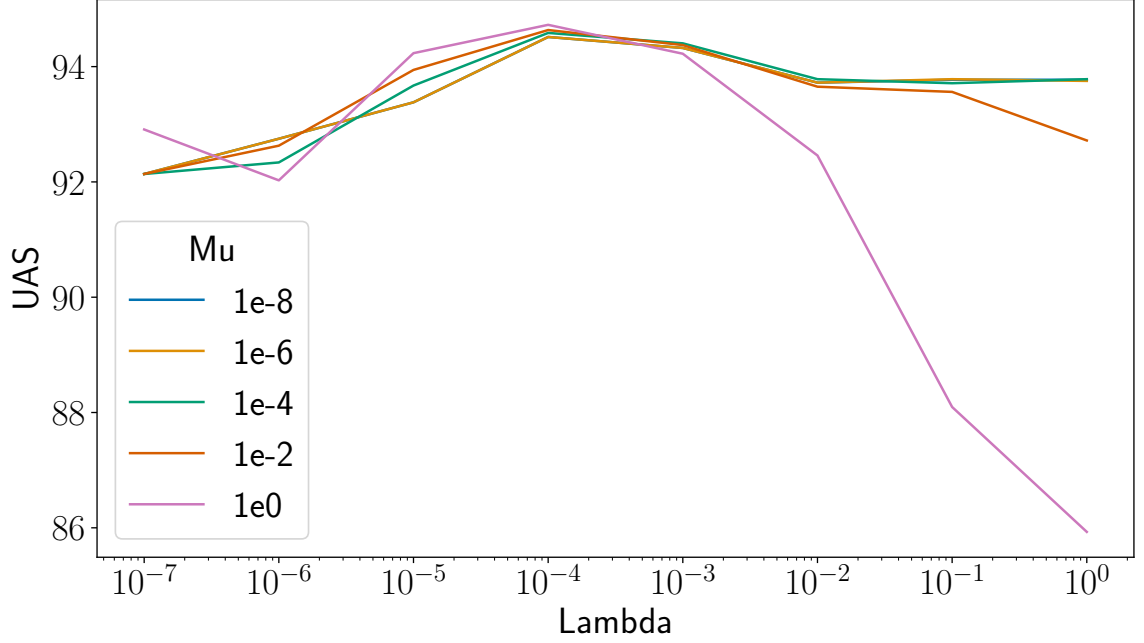


Figure 5: The best UAS with the Marginal algorithm as μ and λ vary in logarithmic scales.

batch size of 200. All the methods achieve their optimal validation set performance in 150-300 steps. *BiAF* and *Marginal* are the fastest because the most time-consuming step of computing dot products of features and parameters is only performed once whereas the other two methods perform it multiple times. However, since *Marginal* is unable to leverage stochastic gradients, its execution time grows linearly in the full batch size. Henceforth, there is a trade-off between *Marginal* and *Stochastic/Game* for computational efficiency. The extra cost compared to *BiAF* with cross entropy is expected because distributional robustness against a set of adversarial distributions is guaranteed.

We compare ADMM and FW by performing for 100 times projection of random points in $[-5, 5]^{75}$ on a graph with 5 nodes and 3 parallel arcs between each (i, j) . We subtract the integral part of the observed minimum values in each run for better illustration. As shown in Figure 4, ADMM usually finds a better solution in the arborescence polytope than FW does within 1000 iterations (One explanation is that FW relies on first-order approximations while there are exponential number of facets in the arborescence polytope). That being said, the per-iteration cost of ADMM is about $8n$ times higher than that of FW due to consensus optimization of n subproblems. In practice, the solution computed with FW usually leads to an approximately good sub-derivative to optimize the DRO objective. We have verified that the solutions suggested by ADMM satisfy the polytope constraints for graphs of up to 10 nodes.

We conduct sensitivity analysis by varying μ and λ on UD Dutch with 100 training samples. Figure 5 implies that moderate smoothing is beneficial to generalization. The ambiguity radius should be judiciously chosen because a small λ causes overfitting while a large λ leads to conservative models.

4.7 Concluding Remarks

We propose a distributionally robust and consistent tree structured prediction method. We show its equivalence to regularized surrogate loss minimization. We put forward a provably convergent algorithm based on efficient projection oracles for arborescence polytopes. Our proposed method enjoys Fisher consistency and robustness against noise in conditional distributions in terms of feature moments. Theoretical and empirical results validate its effectiveness.

Representation learning. Our method can be easily adapted to a representation learning framework with automatic differentiation. Although this may lead to a non-convex problem without the theoretical guarantees derived in this chapter, it is highly desired in practice if feature mappings are optimized as well. We discuss a possible approach as follows. Modern neural networks for supervised learning typically have a linear layer in the end without activation. Assume the penultimate layer outputs $\Phi(\mathbf{x})$ for input \mathbf{x} , the last layer with parameters θ will typically output $\psi(\mathbf{x}) := \Phi(\mathbf{x})\theta \in \mathbb{R}^k$, sometimes called logits, with $k = n^2$ labels for all arcs when parsing a sentence of n tokens. Note that θ in our formulation naturally serves as the parameters of this linear layer. Moreover, knowing $\psi(\mathbf{x})$ is sufficient for us to solve the inner minimax problem in Equation 4.2 to get $\mathbb{P}_{\hat{\mathbf{Y}}|\mathbf{x}}^*$ and $\mathbb{Q}_{\hat{\mathbf{Y}}|\mathbf{x}}^*$. In this way, our DRO method can be considered a loss layer without learnable parameters, which backpropagates the sub-derivative of the objective with respect to $\psi(\mathbf{x})$:

$$\frac{\partial}{\partial \psi(\mathbf{x})} \ell_{\text{adv}} \in \frac{1}{B} \sum_{i=1}^B (\mathbf{q}^{(i)*} - \mathbf{p}_{\text{emp}}^{(i)*}),$$

where B is the batch size. The sub-derivative of the regularization term with respect to θ should be added to the linear layer. Now we are able to take advantage of automatic differentiation and focus on solving the inner adversarial problem given $\psi(\mathbf{x})$ and \mathbf{y} . Since the computational bottleneck lies in computing $\psi(\mathbf{x})$ and backward passes, the overhead of computing the adversarial loss may be dominated and not significant compared to the cross-entropy loss.

CHAPTER 5

MOMENT DISTRIBUTIONALLY ROBUST PROBABILISTIC SUPERVISED LEARNING

(Parts of this chapter were previously public online as “Moment Distributionally Robust Probabilistic Supervised Learning” in the OpenReview preprint (Li and Ziebart, 2023).)

Probabilistic supervised learning assumes the groundtruth itself is a distribution instead of a single label, as in classic settings. It is equivalent to predicting a structured object from the simplex such that the object represents a label distribution. Common approaches learn with a proper composite loss and obtain probability estimates via an invertible link function. Typical links such as the softmax yield restrictive and problematic uncertainty certificates.

In this chapter, we propose to make direct prediction of conditional label distributions from first principles in DRO based on an ambiguity set defined by moments. A brief introduction with related work is given in Section 5.1. We equip the readers with problem setup in Section 5.2. The formulation is presented in Section 5.3.1. We derive its generalization bounds and Fisher consistency under mild assumptions in Section 5.3.2. We illustrate how to manipulate penalties for underestimation and overestimation with specific losses and algorithms in Section 5.3.3. As shown in Section 5.3.4, our method can be easily incorporated into neural networks for end-to-end representation learning. Experimental results in Section 5.4 on datasets with probabilistic labels illustrate the flexibility, effectiveness, and efficiency of this learning paradigm. We conclude this chapter in Section 5.5.

5.1 Introduction

The goal of classical supervised learning is point estimation—predicting a single target from the label domain given features—usually without justifying the confidence. The outcome distribution of an event can be inherently uncertain and more desirable than point predictions in some scenarios. For example, weather predictions that express the uncertainty of events such as rain occurring are more sensible than binary-valued predictions, while a uniform distribution prediction for the outcome of a fair dice roll is more sensible than speculating an integral value randomly. On one hand, the predicted distribution quantifies label uncertainty and is thus more informative than a point prediction, which is widely studied in weakly supervised learning (Yoshida et al., 2021), boosting (Friedman et al., 2000) and optimal treatment (Leibovici et al., 2000). On the other hand, the ground truth naturally comes with multiple targets, possibly with different importances. For instance, there can be multiple emotions in a human face image, there are different gene expression levels over a period of time in biological experiments, and many annotators might disagree over a highly ambiguous instance. In the above settings, each predefined label is part of the ground truth as long as it has a positive probability in the true distribution. Hence, it is natural to use probabilistic labels in both training and inference when the ground truth is no longer a point. In the literature, the task of predicting full distributions from features is called probabilistic supervised learning (Gressmann et al., 2018).

A probabilistic supervised learning task comes with a probabilistic loss functional quantitatively measuring the utility of the prediction (Bickel, 2007). (Williamson et al., 2016) propose a composite multiclass loss that separates properness and convexity. They illuminate

the connection between classification calibration (Tewari and Bartlett, 2007) and properness (Gneiting and Raftery, 2007; Dawid, 2007), representing Fisher consistency for classification and probability estimation respectively. A proper loss is minimized when predictions match the true underlying probability, which implies classification calibration, but not vice versa. Among proper losses, the logarithmic loss (Good, 1952) severely penalizes underestimation of rare outcomes and assessing the “surprise” of the predictor in an information-theoretic sense, the Brier score—originally proposed for evaluating weather forecasts (Brier, 1950)—is useful for assessing prediction calibration, and the spherical scoring rule (Bickel, 2007) is used when a distribution with lower entropy is desired. A single proper loss is sometimes not sufficient for scenarios that elicit optimistic or pessimistic predictions for decision making with practical concerns (Elsberry, 2002; Chapman, 2012). For example, underestimating disastrous events may provide very low utility, motivating more pessimistic predictions. Therefore it is desirable for a proper loss to be flexible in its penalties for deviated predictions that combine statistical properties of multiple losses.

Deep neural networks typically adopt the softmax function to predict a legal distribution. However, softmax intentionally renormalizes the logits and therefore assumes that it follows a logistic distribution (Bendale and Boulton, 2016). It is poor at calibration, uncertainty quantification and robustness against overfitting (Joo et al., 2020). The inverse of the canonical link function in (Williamson et al., 2016) can be used to recover probabilities but commonly resembles softmax (Zou et al., 2008).

We propose a probabilistic supervised learning method from first principles in distributionally robust optimization for general proper losses that realize desired prediction properties. Instead of specifying a parametric distribution, it starts with a minimax learning problem in which the predictor non-parametrically minimizes the the most adverse risk among all distributions in an ambiguity set defined by empirical feature moments. The ambiguity set represents our uncertainty about the underlying distribution. By strong duality, we show that the primal DRO problem is equivalent to a regularized empirical risk minimization problem. The regularization results naturally from the ambiguity set instead of being explicitly imposed. The ERM form also allows us to derive generalization bounds and make inferences from unseen data. We illustrate a set of solutions for general proper losses satisfying certain mild conditions and an efficient algorithm for a weighted sum of two common strictly proper losses. We conduct experiments on real-world datasets by adapting our method to end-to-end differentiable learning.

Contributions. Our contributions are summarized as follows. (1) We propose a distributionally robust probabilistic supervised learning method. (2) We characterize the solutions to the proposed method and present an efficient algorithm for specific losses. (3) We incorporate our method into neural networks and perform extensive empirical study on real-world data.

5.1.1 Related Work

Model assessment of probabilistic models via predictive likelihood has been studied in Bayesian models (Gelman et al., 2014), probabilistic forecasting (Gneiting and Raftery, 2007), machine learning (Masnadi-Shirazi and Vasconcelos, 2009), conditional density estimation (Sugiyama et al., 2010), information theory (Reid and Williamson, 2011) and representation

learning (Dubois et al., 2020). A comprehensive framework for probabilistic supervised learning can be found in (Gressmann et al., 2018).

Techniques developed to explicitly tackle multiclass probabilistic classification include multiclass logistic regression (Collins et al., 2002), support vector machines (Lyu et al., 2019; Wang et al., 2019), learning from noisy labels (Zhang et al., 2021), weakly supervised learning (Yoshida et al., 2021), and neural networks (Papadopoulos, 2013; Gast and Roth, 2018). Multi-label classification, aimed at predicting multiple classes with equal importance, has been analyzed by (Cheng et al., 2010) and (Geng, 2016) in a general probabilistic setting. Note that confidence calibration (Guo et al., 2017) has a different objective from probabilistic supervised learning.

Fisher consistency results have been established for classification losses (Tewari and Bartlett, 2007), structured losses (Ciliberto et al., 2016; Nowak-Vila et al., 2020), proper losses (Williamson et al., 2016) and Fenchel-Young losses (Blondel et al., 2020).

The moment-based ambiguity set adopted in this chapter originates from maximum entropy (Cortes et al., 2015; Mazuelas et al., 2022).

5.2 Preliminaries

5.2.1 Probabilistic Loss Functionals

A loss function measures the quality of a prediction associated with an event. Scoring rules are widely adopted to assess probabilistic predictions, but can be naturally translated to loss functions by appropriate negation and normalization. To illustrate some examples, we consider a decision problem in which $y \in \mathcal{Y}$ is an outcome and $\mathbb{P}_Y \in \mathcal{P}(\mathcal{Y})$ is a predicted distribution over

\mathcal{Y} where we denote by $\mathcal{P}(\mathcal{Y})$ the set of all probability distributions on a set \mathcal{Y} . We denote by $\mathbf{p}_Y \triangleq (\mathbb{P}_Y(y))_{y \in \mathcal{Y}}^T$ a vector of probabilities.

The **zero-one loss** is defined for deterministic prediction so that a penalty of 1 is incurred whenever y' and y differ: $\ell_{01}(y', y) \triangleq \mathbb{I}(y' \neq y)$ where $\mathbb{I}(\cdot)$ is the indicator function. It extends to probabilistic predictions as $\ell_{01}(\mathbb{P}_Y, y) \triangleq 1 - \mathbb{P}_Y(y)$. In the literature, the zero-one loss is sometimes defined as $\ell_{01}(\mathbb{P}_Y, y) := \mathbb{I}(y \notin \arg \max_{y'} \mathbb{P}_Y(y'))$, which is proper, but discontinuous and not strictly proper. The **cost-sensitive loss** for multiclass classification is similarly defined with a confusion cost matrix $\mathbf{C} \in \mathbb{R}_+^{|\mathcal{Y}| \times |\mathcal{Y}|}$: $\ell_{cs}(\mathbb{P}_Y, y) \triangleq \sum_{i \in \mathcal{Y}} \mathbb{P}_Y(i) C_{iy}$.

The multiclass **Brier loss**, based on the Brier score or quadratic scoring rule, measures the mean squared difference between \mathbb{P}_Y and y : $\ell_{br}(\mathbb{P}_Y, y) \triangleq \sum_{y'} (\mathbb{P}_Y(y') - \mathbb{I}(y' = y))^2$.

The **logarithmic loss**, also called log-likelihood loss, incurs a rapidly increasing penalty as the predicted probability of the target event approaches zero: $L_{\log}(\mathbb{P}_Y, y) \triangleq -\ln \mathbb{P}_Y(y)$.

The **spherical scoring rule** can be interpreted as the spherical projection of the true belief onto the prediction vector. To use it as a loss function, we define $\ell_{sp}(\mathbb{P}_Y, y) \triangleq 1 - \mathbb{P}_Y(y) / \|\mathbf{p}_Y\|_2$.

For ease of exposition, we define $L(\mathbb{P}, \mathbb{Q}) := \sum_y \mathbb{Q}_Y(y) \ell(\mathbb{P}_Y, y)$ where $\ell(\cdot, \cdot) : \mathcal{P}(\mathcal{Y}) \times \mathcal{Y} \rightarrow \mathbb{R}_+$ is a probabilistic loss function as illustrated above. A loss L is called **proper** if $L(\mathbb{Q}, \mathbb{Q}) \leq L(\mathbb{P}, \mathbb{Q})$ for all \mathbb{P}, \mathbb{Q} , and called **strictly proper** if \mathbb{Q} is the unique minimizer of $L(\cdot, \mathbb{Q})$. Figure 6 provides a graphical comparison of the above losses for prediction with three classes. We can infer that the zero-one loss is an improper loss.

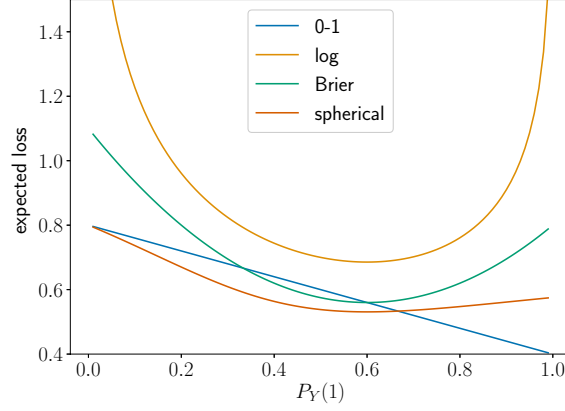


Figure 6: The expected value of four loss functions for three classes with $\mathbb{Q}_Y(1) = 0.6$ and $\mathbb{Q}_Y(2) = \mathbb{Q}_Y(3) = 0.2$. $\mathbb{P}_Y(2) = \mathbb{P}_Y(3)$ as $\mathbb{P}_Y(1)$ varies. Each loss is normalized to cross $(1, 0)$ and $(0.5, 0.5)$ according to the binary case with a hard label. Best viewed in color.

5.2.2 Probabilistic Supervised Learning

We study the probabilistic supervised learning task where we are given n training samples $\{(\mathbf{x}^{(1)}, y^{(1)}), (\mathbf{x}^{(2)}, y^{(2)}), \dots, (\mathbf{x}^{(n)}, y^{(n)})\}$ drawn i.i.d. from a distribution \mathbb{P} on the joint space $\mathcal{X} \times \mathcal{Y}$, in which \mathcal{X} is a feature space and \mathcal{Y} is a univariate finite discrete label space. A probabilistic multiclass loss function $L : \mathcal{P}(\mathcal{Y}) \times \mathcal{P}(\mathcal{Y}) \rightarrow \mathbb{R}_+$ is given. The goal of ERM is to learn from the samples a mapping $h : \mathcal{X} \rightarrow \mathcal{P}(\mathcal{Y})$ to minimize the empirical L -risk of h :

$$h^* \in \arg \min_{h \in \mathcal{H}} R_{\mathbb{P}^{\text{emp}}}^L(h) := \mathbb{E}_{\mathbb{P}_{\mathbf{X}}^{\text{emp}}} \left[L(h(\mathbf{X}), \mathbb{P}_{Y|\mathbf{X}}^{\text{emp}}) \right], \quad (5.1)$$

where $\mathbb{P}_{\mathbf{X}, Y}^{\text{emp}}$ represents the empirical distribution and \mathcal{H} is a hypothesis space. Here we assume \mathbf{x} may be accompanied with a probabilistic label by aggregating instances with the same $\mathbf{x}^{(i)}$. In

this way, both learning and inference are accomplished in the general setting subsuming classical supervised learning.

5.3 Method

We now present our formulation for learning with general multiclass probabilistic losses. We provide theoretical results of consistency and generalization. We study the solution for general proper losses in our formulation and develop an efficient algorithm for two typical proper losses.

5.3.1 Formulation

We consider a continuous proper loss L to be optimized under the unknown distribution \mathbb{P}^{true} . We assume that a class-sensitive feature function $\phi : \mathcal{X} \times \mathcal{Y} \rightarrow \mathbb{R}^d$ that maps a data point to a d -dimensional feature vector is given. Examples include the multi-vector representation and class-dependent TF-IDF scores. Choosing a good ϕ is a representation learning problem, but as we will discuss in Section 5.3.4, it is not a concern once our method is incorporated into neural networks as a layer. Intuitively, the elements of the vector $\phi(\mathbf{x}, y)$ can be regarded as scores indicating how well the label y matches with the feature \mathbf{x} . For example, with a linear hypothesis $h_{\mathbf{w}}(\mathbf{x}, y) = \langle \mathbf{w}, \phi(\mathbf{x}, y) \rangle$, a good parameter vector \mathbf{w}^* should yield

$$\langle \mathbf{w}^*, \phi(\mathbf{x}, y) \rangle > \langle \mathbf{w}^*, \phi(\mathbf{x}, y') \rangle \implies \mathbb{P}(\mathbf{x}, y) > \mathbb{P}(\mathbf{x}, y').$$

Instead of specifying a parametric form of predictions, we adopt a minimax statistical learning formulation:

$$\min_{\mathbb{P}_{Y|\mathbf{X}} \in \mathcal{P}(\mathcal{Y})} \max_{\mathbb{Q} \in \mathcal{A}_\varepsilon^M(\mathbb{P}^{\text{emp}})} \mathbb{E}_{\mathbb{Q}_{\mathbf{X}}} [L(\mathbb{P}_{Y|\mathbf{X}}, \mathbb{Q}_{Y|\mathbf{X}})], \quad (5.2)$$

where $\mathcal{A}_\varepsilon^M(\mathbb{P}^{\text{emp}}) := \{\mathbb{Q} : \mathbb{Q} \in \mathcal{P}(\mathcal{X} \times \mathcal{Y}) \wedge \mathbb{P}_{\mathbf{X}}^{\text{emp}} = \mathbb{Q}_{\mathbf{X}} \wedge \|\mathbb{E}_{\mathbb{P}^{\text{emp}}}[\phi(\cdot, \cdot)] - \mathbb{E}_{\mathbb{Q}}[\phi(\cdot, \cdot)]\| \leq \varepsilon\}$. The ambiguity set is different from that in (Wiesemann et al., 2014) and (Farnia and Tse, 2016) due to the inequality and feature mapping respectively. The minimization over the function space \mathcal{H} is replaced by directly minimizing over $\mathcal{P}(\mathcal{Y})$ for each $\mathbf{x} \in \mathcal{X}$. The probabilistic predictions are chosen to minimize the worst-case risk evaluated on a set of distributions in an ambiguity set defined by the empirical distribution \mathbb{P}^{emp} and feature mapping ϕ . The ambiguity set $\mathcal{A}_\varepsilon^M(\mathbb{P}^{\text{emp}})$ includes distributions that share the same marginal on \mathcal{X} and are no more than ε away from \mathbb{P}^{emp} in terms of feature moment divergence. Note that given any feature function ϕ , the ambiguity set is a compact convex set. Conceptually, we restrict the support of \mathbb{Q} on \mathcal{X} to be the same as the empirical distribution for convenience in both algorithm design and theoretical analysis.

Minimizing the worst-case risk by allowing a certain amount of label uncertainty makes this method inherently robust. It can also be shown to be equivalent to a dual-norm regularized ERM problem:

Proposition 33. *The distributionally robust probabilistic supervised learning problem based on moment divergence in Equation 5.2 can be rewritten as*

$$\min_{\boldsymbol{\theta}} \mathbb{E}_{\mathbb{P}_{\mathbf{X}}^{emp}} \underbrace{\min_{\mathbb{P}} \max_{\mathbb{Q}} L(\mathbb{P}_{Y|\mathbf{X}}, \mathbb{Q}_{Y|\mathbf{X}}) + \boldsymbol{\theta}^\top (\mathbb{E}_{\mathbb{Q}_{\tilde{Y}|\mathbf{X}}} \phi(\mathbf{X}, \tilde{Y}) - \mathbb{E}_{\mathbb{P}_{\tilde{Y}|\mathbf{X}}^{emp}} \phi(\mathbf{X}, \tilde{Y})) + \varepsilon \|\boldsymbol{\theta}\|_*}_{L_{adv}(\boldsymbol{\theta}, \mathbb{P}_{\tilde{Y}|\mathbf{X}}^{emp})}, \quad (5.3)$$

where $\boldsymbol{\theta} \in \mathbb{R}^D$ is the vector of Lagrangian multipliers and $\|\cdot\|_*$ is the dual norm of $\|\cdot\|$.

Proof. The proof follows similarly Proposition 28. Both $\mathcal{P}(\mathcal{Y})$ and $\mathcal{A}_\varepsilon^M(\tilde{\mathbb{P}})$ are non-empty closed convex sets. Since we assume L is continuous and proper, we know that $L(\cdot, \mathbb{Q})$ is quasi-convex for every \mathbb{Q} and $L(\mathbb{P}, \cdot)$ is concave for every \mathbb{P} by definition. Equation 5.2 is therefore a quasi-convex-concave problem and strong duality holds (Sion, 1958). The regularization is obtained via Lagrangian and Fenchel conjugate. \square

It is well-known that continuous proper losses are quasi-convex, such as the Brier score, the logarithmic score, the spherical score, the Winkler's score, the ranked probability score, etc. However, some improper (possibly discrete and non-convex) losses can be quasi-convex in the predicted distribution (e.g., the zero-one loss). In contrast, surrogate classification losses are usually convex in a parameter space that is easy to work with, for example, the multiclass hinge loss (Weston and Watkins, 1998), $\ell_{\text{ww}}(\boldsymbol{\psi}, y) = \sum_{y' \neq y} \max\{0, 1 + \psi_{y'} - \psi_y\}$, and the multiclass logistic loss (Nelder and Wedderburn, 1972), $\ell_{\log}(\boldsymbol{\psi}, y) = \ln(\sum_{y'} \exp(\psi_{y'})) - \psi_y$, where $\boldsymbol{\psi} \in \mathbb{R}^{|\mathcal{Y}|}$ is a vector of class scores.

From a game theoretic point of view, our formulation in Equation 5.2 is equivalent to a two-player zero-sum game in which the predictor player chooses a distribution to minimize the expected game payoff while the adversary player chooses one to maximize the game value while constrained to satisfy certain statistical properties of training data (Grünwald and Dawid, 2004). In the dual problem (Equation 5.3), the Lagrange multipliers parameterize the payoff function for an augmented game and provide a new payoff function for unseen data to construct predictors.

5.3.2 Statistical Properties

It well known that minimizing strictly proper losses leads to Fisher consistent probability estimation (Williamson et al., 2016). However, minimization of the surrogate risk in Equation 5.3 may induce a sub-optimal classifier because of misalignment between the surrogate loss L_{adv} and the original loss L . Fisher consistency provides desirable statistical implications for a surrogate loss such that minimizing it yields an estimator that also minimizes the original loss.

The adversarial surrogate loss L_{adv} is endowed with an additional regularization term. It reduces to a Fenchel-Young loss (Blondel et al., 2020) when the ambiguity radius ε is zero. A conclusion of consistency can drawn based on (Nowak-Vila et al., 2020; Blondel et al., 2020) and our assumption that the groundtruth is probabilistic:

Corollary 34. *When $\varepsilon = 0$, L_{adv} is Fisher consistent with respect to L . Namely, for any \mathbf{x} ,*

$$\mathbb{P}_{Y|\mathbf{x}}^{\theta^*} \in \arg \min_{\mathbb{P}_{Y|\mathbf{x}}} L(\mathbb{P}_{Y|\mathbf{x}}, \mathbb{P}_{Y|\mathbf{x}}^{true})$$

is the Bayes optimal probabilistic prediction made by θ_{true}^* , the solution in Equation 5.3 under \mathbb{P}^{true} . The prediction made by θ is $\mathbb{P}_{Y|\mathbf{X}}^\theta \in \arg \min_{\mathbb{P}} \max_{\mathbb{Q}} L(\mathbb{P}_{Y|\mathbf{X}}, \mathbb{Q}_{Y|\mathbf{X}}) + \mathbb{E}_{\mathbb{Q}_{Y|\mathbf{X}}} \theta^\top \phi(\mathbf{X}, \check{Y})$.

Proof. The proof follows similarly Corollary 30. \square

The consistency result guarantees that the learned probabilistic prediction rules yield Bayes optimal risk as ERM with proper losses in the ideal setting with true distributions and all measurable functions. Also note that the conclusion holds for all quasi-convex losses.

Basic generalization bounds related to true risk for DRO methods can be derived from measure concentration. This approach depends on the choice of ambiguity sets and may have a dimensionality issue. It is also not appropriate for ambiguity sets defined by low-order moments. Thus, we take an alternate approach following (Farnia and Tse, 2016) to prove excess out-of-sample risk bounds. We assume $\varepsilon > 0$ to ensure boundedness of $\|\theta\|_*$. We establish the following theorem by making mild assumptions on boundedness on features and losses:

Theorem 35. *Given n samples, a non-negative multiclass probabilistic loss $L(\cdot, \cdot)$ such that $|L(\cdot, \cdot)| \leq K$, a feature function $\phi(\cdot, \cdot)$ such that $\|\phi(\cdot, \cdot)\| \leq B$ and a positive ambiguity level $\varepsilon > 0$, then, for any $0 < \delta \leq 1$, with a probability at least $1 - \delta$, the following excess true worst-case risk bound holds:*

$$\max_{\mathbb{Q} \in \mathcal{A}_\varepsilon^M(\mathbb{P}^{true})} R_{\mathbb{Q}}^L(\theta_{emp}^*) - \max_{\mathbb{Q} \in \mathcal{A}_\varepsilon^M(\mathbb{P}^{true})} R_{\mathbb{Q}}^L(\theta_{true}^*) \leq \frac{4KB}{\varepsilon\sqrt{n}} \left(1 + \frac{3}{2} \sqrt{\frac{\ln(4/\delta)}{2}} \right), \quad (5.4)$$

where $\boldsymbol{\theta}_{emp}^*$ and $\boldsymbol{\theta}_{true}^*$ are the optimal parameters learned in Equation 5.3 under the empirical distribution \mathbb{P}^{emp} and true distribution \mathbb{P}^{true} , respectively. The original risk of $\boldsymbol{\theta}$ under \mathbb{Q} is $R_{\mathbb{Q}}^L(\boldsymbol{\theta}) := \mathbb{E}_{\mathbb{Q}_{\mathbf{X}, Y}, \mathbb{P}_{Y|\mathbf{X}}^{\boldsymbol{\theta}}} L(\mathbb{P}_{Y|\mathbf{X}}, \mathbb{Q}_{Y|\mathbf{X}})$.

Proof. The proof follows Theorem 29. □

Theorem 35 improves the results of (Asif et al., 2015) and (Fathony et al., 2016) that only show qualitative bounds. Under positive regularization, this bound explains the rate of uniform convergence of the true worst-case risk of the estimator $\boldsymbol{\theta}_{emp}^*$ learned through the empirical distribution \mathbb{P}^{emp} to the true worst-case risk of the ideal estimator $\boldsymbol{\theta}_{true}^*$ learned under \mathbb{P}^{true} . Although the empirical estimator is obtained based on a finite set of samples, Theorem 35 justifies the roles which the ambiguity set $\mathcal{A}_{\varepsilon}^M(\cdot)$, the feature function $\phi(\cdot, \cdot)$, the loss function $L(\cdot, \cdot)$ and the ambiguity parameter ε play in upper bounding the excess out-of-sample worst-case risk. Intuitively, a larger ε will reject more hypotheses that are sensitive with larger dual norms, whereas the worst-case risk scales with the range of loss and feature functions.

5.3.3 Algorithm

Since $L(\cdot, \cdot)$ is a continuous quasiconvex-concave function, a saddle point in Equation 5.3 given $\boldsymbol{\theta}$ must have a zero derivative with respect to \mathbb{P} and \mathbb{Q} :

$$\sum_y \mathbb{Q}_{Y|\mathbf{x}}(y) \partial \ell(\mathbb{P}_{Y|\mathbf{x}}, y) / \partial \mathbb{P}_{Y|\mathbf{x}}(y') + Z_{\mathbb{P}_{Y|\mathbf{x}}} = 0 \quad (5.5)$$

$$\ell(\mathbb{P}_{Y|\mathbf{x}}, y) + \boldsymbol{\theta}^\top \phi(\mathbf{x}, y) + Z_{\mathbb{Q}_{Y|\mathbf{x}}} = 0, \quad (5.6)$$

where $Z_{\mathbb{P}_{Y|\mathbf{x}}}$ is the Lagrange multipliers for the simplex constraint $\sum_y \mathbb{P}_{Y|\mathbf{x}}(y) = 1$, similarly for $Z_{\mathbb{Q}_{Y|\mathbf{x}}}$. Note that $Z_{\mathbb{Q}_{Y|\mathbf{x}}}$ is constant for all y given \mathbf{x} . If ℓ is local, e.g., $\ell(\mathbb{P}_{Y|\mathbf{x}}, y)$ is independent of $\mathbb{P}_{Y|\mathbf{x}}(y')$ for $y' \neq y$ and if $\ell(\cdot, y)$ is monotone in $\mathbb{P}_{Y|\mathbf{x}}(y) > 0$ (without simplex constraints) with range \mathbb{R} , which is the case for the logarithmic loss, Equation 5.6 always has a solution and the system of equations for all y along with the simplex constraint $\sum_y \mathbb{P}_{Y|\mathbf{x}}(y)$ has a unique solution. With few assumptions on the boundedness of ℓ and $\boldsymbol{\theta}^\top \boldsymbol{\phi}$, Equation 5.6 is ill-posed. Given $\mathbb{P}_{Y|\mathbf{x}}^*$ from Equation 5.6, the solution $\mathbb{Q}_{Y|\mathbf{x}}^*$ to Equation 5.5 exists iff

$$\begin{pmatrix} \partial \ell(\mathbb{P}_{Y|\mathbf{x}}, 1) / \partial \mathbb{P}_{Y|\mathbf{x}}(1) & \dots & \partial \ell(\mathbb{P}_{Y|\mathbf{x}}, |\mathcal{Y}|) / \partial \mathbb{P}_{Y|\mathbf{x}}(1) & 1 \\ \dots & & & \\ \partial \ell(\mathbb{P}_{Y|\mathbf{x}}, 1) / \partial \mathbb{P}_{Y|\mathbf{x}}(|\mathcal{Y}|) & \dots & \partial \ell(\mathbb{P}_{Y|\mathbf{x}}, |\mathcal{Y}|) / \partial \mathbb{P}_{Y|\mathbf{x}}(|\mathcal{Y}|) & 1 \\ 1 & \dots & 1 & 0 \end{pmatrix}$$

is singular. By assuming locality and positiveness, there exists a unique solution $\mathbb{Q}_{Y|\mathbf{x}}^*$. One benefit of the proposed method is that users only need to focus on solve Equation 5.6 and Equation 5.5 for proper losses while (Williamson et al., 2016) additionally require a canonical link function for convexity.

Next we show how the system of equations can always be solved with specific losses. We consider an additive combination of the multiclass Brier loss and the logarithmic loss, both of which are continuous strictly proper losses. As indicated by Figure 6, these losses differ primarily in how they penalize the ground truth label's prediction probability as it goes to zero and one. The Brier loss exhibits quadratic growth. The logarithmic loss has a vertical asymptote

for labels considered increasingly unlikely to the point of impossibility by the predictor. They have different penalties for underestimation and overestimation of the desired prediction. A trade-off between the log loss and the Brier loss thus provides flexibility to control the cost for misalignment between the prediction and the observation.

We employ this kind of loss in our DRO method and present an efficient algorithm that can be implemented in practice. With only slight loss of generality and for computational consideration, we assume a fixed positive weight on the log loss. To begin with, the mixture loss is

$$\ell_{\text{mix}}(\mathbb{P}_{Y|\mathbf{x}}, y) = -\ln \mathbb{P}_{Y|\mathbf{x}}(y) + \beta(1 - 2\mathbb{P}_{Y|\mathbf{x}}(y) + \sum_{y'} \mathbb{P}_{Y|\mathbf{x}}^2(y')),$$

with derivative

$$\partial \ell_{\text{mix}}(\mathbb{P}_{Y|\mathbf{x}}, y) / \partial \mathbb{P}_{Y|\mathbf{x}}(y) = -1/\mathbb{P}_{Y|\mathbf{x}}(y) - 2\beta + 2\beta \mathbb{P}_{Y|\mathbf{x}}(y).$$

Scalar β weights the contribution of the Brier loss, to this additive combination, controlling the sensitivity of the predictor to underestimation. The adversarial surrogate of this mixture loss is Fisher consistent as a direct corollary. Methods that solely mix the predictions of classifiers designed for logarithmic loss minimization and Brier loss optimization, may be appealing for their simplicity, but are demonstrably sub-optimal. For example, with the logistic loss, logistic regression provides a natural parametric form for the predictor, that equates loss minimization with data likelihood maximization.

Although the Brier loss is not local, the additional sum of quadratic terms $\sum_{y'} \mathbb{P}_{Y|\mathbf{x}}^2(y')$ is constant across all y . Therefore Equation 5.6 has a closed form expression in terms of the Lambert W function. Furthermore, the sum over y for all $\mathbb{Q}_{Y|\mathbf{x}}(y)$ will cancel out, leaving terms only dependent on the same y . So Equation 5.5 is simplified into an expression of \mathbb{Q} in terms of \mathbb{P} . Normalizing \mathbb{Q} solves $Z_{\mathbb{P}}$, yielding the following proposition:

Proposition 36. *The DRO method for a probabilistic loss based on logarithmic loss, and β Brier loss has a solution $\mathbb{P}_{Y|\mathbf{x}}^*$ for the predictor parameterized by $\boldsymbol{\theta}$ defined by the following systems of equations:*

$$\forall \mathbf{x} \in \mathcal{X}, \exists C \in \mathbb{R}, \forall y \in \mathcal{Y} \quad \mathbb{P}_{Y|\mathbf{x}}^*(y) = \exp(C + \boldsymbol{\theta}^T \boldsymbol{\phi}(\mathbf{x}, y) - W_0(2\beta e^{C + \boldsymbol{\theta}^T \boldsymbol{\phi}(\mathbf{x}, y)})), \quad (5.7)$$

where C is a constant dependent on $\boldsymbol{\theta}$ and \mathbf{x} but independent of y , $W(\cdot)$ is the principal branch of the Lambert W function. The corresponding adversary $\mathbb{Q}_{Y|\mathbf{x}}^*$ is defined as

$$\mathbb{Q}_{Y|\mathbf{x}}^*(y) = \frac{2\beta \mathbb{P}_{Y|\mathbf{x}}^{*2}(y) + Z_{\mathbb{P}_{Y|\mathbf{x}}} \mathbb{P}_{Y|\mathbf{x}}^*(y)}{1 + 2\beta \mathbb{P}_{Y|\mathbf{x}}^*(y)} \text{ and } Z_{\mathbb{P}_{Y|\mathbf{x}}} = \frac{1 - \sum_y 2\beta \mathbb{P}_{Y|\mathbf{x}}^{*2}(y)/(1 + 2\beta \hat{\mathbb{P}}_{Y|\mathbf{x}}^*(y))}{\sum_y \mathbb{P}_{Y|\mathbf{x}}^*(y)/(1 + 2\beta \hat{\mathbb{P}}_{Y|\mathbf{x}}^*(y))}. \quad (5.8)$$

Proof. Recall the saddle-point optimality condition:

$$\sum_y \mathbb{Q}_Y(y) \partial \ell(\mathbb{P}_Y, y) / \partial \mathbb{P}_Y(y') + Z_{\mathbb{P}_Y} = 0$$

$$\ell(\mathbb{P}_Y, y) + \boldsymbol{\theta}^T \boldsymbol{\phi}(\mathbf{x}, y) + Z_{\mathbb{Q}_Y} = 0.$$

Dependence on \mathbf{x} is omitted when context is clear. Substituting ℓ_{mix} yields:

$$\begin{aligned} \mathbb{Q}_Y(y) \left(-\frac{1}{\mathbb{P}_Y(y)} - 2\beta \right) + 2\beta \mathbb{P}_Y(y) + Z_{\mathbb{P}_Y} &= 0 \\ -\ln \mathbb{P}_Y(y) + \beta(1 - 2\mathbb{P}_Y(y) + \sum_{y'} \mathbb{P}_Y^2(y')) + \boldsymbol{\theta}^\top \boldsymbol{\phi}(\mathbf{x}, y) + Z_{\mathbb{Q}_Y} &= 0. \end{aligned}$$

Note that $C := \beta + \beta \sum_{y'} \mathbb{P}_Y^2(y') + Z_{\mathbb{Q}_Y}$ is constant across all y 's given $\boldsymbol{\theta}, \mathbf{x}$. Thus for fixed $\boldsymbol{\theta}, \mathbf{x}$, we have for some $C_{\boldsymbol{\theta}, \mathbf{x}}^*$,

$$C_{\boldsymbol{\theta}, \mathbf{x}}^* + \boldsymbol{\theta} \cdot \boldsymbol{\phi}(\mathbf{x}, y) = \ln \mathbb{P}_Y(y) + 2\beta \mathbb{P}_Y(y) \quad \forall y \in \mathcal{Y},$$

which is equivalent to

$$2\beta \mathbb{P}_Y(y) e^{2\beta \mathbb{P}_Y(y)} = 2\beta e^{\boldsymbol{\theta} \cdot \boldsymbol{\phi}(\mathbf{x}, y) + C_{\boldsymbol{\theta}, \mathbf{x}}^*}.$$

By the definition of the Lambert W function,

$$2\beta \mathbb{P}_Y(y) = W(2\beta e^{\boldsymbol{\theta} \cdot \boldsymbol{\phi}(\mathbf{x}, y) + C_{\boldsymbol{\theta}, \mathbf{x}}^*}).$$

Since $2\beta e^{\boldsymbol{\theta} \cdot \boldsymbol{\phi}(\mathbf{x}, y) + C_{\boldsymbol{\theta}, \mathbf{x}}^*} \geq 0$, the principal branch W_0 of the Lamber W function is always applicable. Also by the formula $e^{-W(x)} = \frac{W(x)}{x}$, we have

$$\mathbb{P}_Y(y) = \exp(C_{\boldsymbol{\theta}, \mathbf{x}}^* + \boldsymbol{\theta}^\top \boldsymbol{\phi}(\mathbf{x}, y) - W_0(2\beta e^{C_{\boldsymbol{\theta}, \mathbf{x}}^* + \boldsymbol{\theta}^\top \boldsymbol{\phi}(\mathbf{x}, y)})) \quad \forall y.$$

Let \mathbb{P}_Y^* (for a given θ) be a solution to this set of equations that also satisfies $\sum_y \mathbb{P}_Y^*(y) = 1$. By Equation 5.5, the optimal \mathbb{Q} satisfies

$$\mathbb{Q}_Y^*(y) = \frac{2\beta\mathbb{P}_Y^*(y) + Z_{\mathbb{P}_Y}}{\frac{1}{\mathbb{P}_Y^*(y)} + 2\beta} = \frac{2\beta\mathbb{P}_Y^{*2}(y) + Z_{\mathbb{P}_Y}\mathbb{P}_Y^*(y)}{1 + 2\beta\mathbb{P}_Y^*(y)}.$$

$Z_{\mathbb{P}_Y}$ must be chosen to properly normalize $\mathbb{Q}_Y^*(y)$:

$$\begin{aligned} \sum_y \mathbb{Q}_Y^*(y) &= Z_{\mathbb{P}_Y} \sum_y \frac{1}{\frac{1}{\mathbb{P}_Y^*(y)} + \alpha + 2\beta} + \sum_y \frac{2\beta\mathbb{P}_Y^*(y)}{\frac{1}{\mathbb{P}_Y^*(y)} + \alpha + 2\beta} = 1 \\ \implies Z_{\mathbb{P}_Y}^* &= \frac{1 - \sum_y \frac{2\beta\mathbb{P}_Y^*(y)}{\frac{1}{\mathbb{P}_Y^*(y)} + \alpha + 2\beta}}{\sum_y \frac{1}{\frac{1}{\mathbb{P}_Y^*(y)} + \alpha + 2\beta}} = \frac{1 - \sum_y \frac{2\beta\mathbb{P}_Y^{*2}(y)}{1 + (\alpha + 2\beta)\mathbb{P}_Y^*(y)}}{\sum_y \frac{\mathbb{P}_Y^*(y)}{1 + (\alpha + 2\beta)\mathbb{P}_Y^*(y)}}. \end{aligned}$$

Both $Z_{\mathbb{P}_Y}^*$ and $\mathbb{Q}_Y^*(y)$ are positive because $\mathbb{P}_Y^* \in \mathcal{P}(\mathcal{Y})$ is a solution. □

Algorithm 4 Distributionally robust learning for probabilistic supervised learning with a mixture of logistic and Brier losses

Input: ϕ , $\mathbb{P}_{\mathbf{X},Y}^{\text{emp}}$, β , learning rate γ

Output: θ^*

Initialize θ to be a random vector

repeat

for all $\mathbf{x} \in \mathcal{X}$ **do**

$C, \mathbb{P}_{Y|\mathbf{X}}^*(\cdot|\mathbf{x}) \leftarrow \text{Bisection}(\mathbf{x}, \phi, \theta, \beta)$ by Equation 5.7

 Compute $\mathbb{Q}_{Y|\mathbf{X}}^*(\cdot|\mathbf{x})$ by Equation 5.8

end for

 Compute $\partial L_{\text{adv}}/\partial \theta$ by Equation 5.9

$\theta \leftarrow \theta - \gamma \partial L_{\text{adv}}/\partial \theta$

until convergence

Now we show how to solve Equation 5.7 with simplex constraints to obtain $\mathbb{P}_{Y|\mathbf{x}}^*$ given θ for any $\mathbf{x} \in \mathcal{X}$. Let $C = f_y(t) = \theta^T \phi(\mathbf{x}, y) - \ln t - 2\beta t$ be a function of $t = \mathbb{P}_{Y|\mathbf{x}}^*(y)$. By definition, $f(\cdot)$ is a monotonically decreasing function with domain \mathbb{R}_{++} and range \mathbb{R} . Its inverse mapping $f^{-1}(\cdot)$ is monotonically decreasing with domain \mathbb{R} and range \mathbb{R}_{++} . Therefore, let $g(C) = \sum_y f_y^{-1}(C) = \sum_y \mathbb{P}_{Y|\mathbf{x}}^*(y)$, according to the intermediate value theorem, there exists $C^* \in \mathbb{R}$ such that $g(C^*) = \sum_y \mathbb{P}_{Y|\mathbf{x}}^*(y) = 1$. Because of their monotonicity, we can find C^* and

$\mathbb{P}_{Y|\mathbf{x}}^*(\cdot)$ as a solution to Equation 5.7 via bisection method. Once $\mathbb{P}_{Y|\mathbf{x}}^*$ is obtained, we can find $\mathbb{Q}_{Y|\mathbf{x}}^*$ simply by substitution. After that, the sub-gradient,

$$\partial L_{\text{adv}}/\partial \boldsymbol{\theta} \triangleq \mathbb{E}_{\mathbb{P}_{\mathbf{X}}^{\text{emp}}}(\mathbb{E}_{\mathbb{Q}_{Y|\mathbf{x}}^*}[\phi(\mathbf{X}, Y)] - \mathbb{E}_{\mathbb{P}_{Y|\mathbf{x}}^{\text{emp}}}[\phi(\mathbf{X}, Y)]) + \partial \varepsilon \|\boldsymbol{\theta}\|_*/\partial \boldsymbol{\theta}, \quad (5.9)$$

can be leveraged to optimize $\boldsymbol{\theta}$. The above steps are summarized in Algorithm 4.

5.3.4 Differentiable Learning

By taking advantage of deep neural networks, our method will be able to jointly optimize data representation and the Lagrange multipliers:

$$\min_{\boldsymbol{\theta}, \phi} \mathbb{E}_{\mathbb{P}_{\mathbf{X}}^{\text{emp}}} L_{\text{adv}}(\boldsymbol{\theta}, \mathbb{P}_{\hat{Y}|\mathbf{X}}^{\text{emp}}),$$

enjoying the benefits of end-to-end representation learning without manually looking for a good feature mapping ϕ . More off-the-shelf mini-batch training tools could be leveraged as well.

We show how to make use of our DRO method as a loss layer in neural network training. A network for supervised learning typically has a linear classification layer in the end without activation. Assume the penultimate layer outputs $\phi(\mathbf{x})$, the last layer will output a $|\mathcal{Y}|$ -dimensional vector $\boldsymbol{\psi}(\mathbf{x}) = [(\boldsymbol{\theta}^{(1)})^\top \phi(\mathbf{x}), \dots, (\boldsymbol{\theta}^{(|\mathcal{Y}|)})^\top \phi(\mathbf{x})]$. This is essentially equivalent to adopting a multi-vector representation to construct ϕ . Specifically, given $\mathbf{x} \in \mathbb{R}^d$ and $y \in [|\mathcal{Y}|]$, the resulting feature vector $\mathbf{v} = \phi(\mathbf{x}, y) \in \mathbb{R}^{d|\mathcal{Y}|}$ satisfies $v_{yd-d+i} = x_i$ for $i \in [d]$ and $v_j = 0$ otherwise. Therefore taking $\boldsymbol{\psi}(\mathbf{x})$ as the input is sufficient for us to compute $\mathbb{P}_{Y|\mathbf{x}}^*$ and $\mathbb{Q}_{Y|\mathbf{x}}^*$. In

this way, our method is the loss layer without learnable parameters, which backpropagates the sub-derivative of loss with respect to $\psi(\mathbf{x})$ to the linear classification layer:

$$\mathbb{E}_{\mathbb{P}_{\mathbf{X}}^{\text{emp}}}(\mathbf{q}_{Y|\mathbf{X}} - \mathbf{p}_{Y|\mathbf{X}}^{\text{emp}}) \in \partial L_{\text{adv}} / \partial \psi(\mathbf{x}).$$

Recall \mathbf{q} and \mathbf{p}^{emp} are the probability vectors for \mathbb{Q} and \mathbb{P}^{emp} . The sub-gradient with respect to θ is added to the classification layer.

5.4 Experiments

In the experiments, we consider as the performance measure the L -risk $R_{\mathbb{P}}^L(h)$, also called the expected generalization loss. The mixture loss ℓ_{mix} of the log loss and Brier loss is adopted. The normalized generalization loss $\frac{1}{(1+\alpha+\beta)} R_{\mathbb{P}^{\text{test}}}^L(h)$ is estimated based on the test set distribution $\mathbb{P}_{\mathbf{X},Y}^{\text{test}}$.

We compare our adversarial learning approach against an uninformed baseline (UNINF) (Gressmann et al., 2018), multi-layer perceptron (MLP) (Hinton, 1990) and k-nearest neighbor (KNN) (Beygelzimer et al., 2006). All the baseline methods are able to make use of probabilistic labels in both training and testing. The uninformed baseline simply outputs the marginal label distribution $\tilde{\mathbb{P}}_Y$ based on training data as inference. We adopt a three-layer neural network for MLP and our method, who share the same number of parameters. To make a more fair comparison, we set $\varepsilon = 0$ such that the final classification layer is unregularized. MLP computes the target loss L_{mix} with an additional softmax layer applied to the logits.

We implement MLP and our method using PyTorch (Paszke et al., 2019). We adopt the KNN implementation from the scikit-learn library (Pedregosa et al., 2011). The uninformed baseline is implemented in Python. For optimization, we use Adam (Kingma and Ba, 2014) for MLP and our method. The number of hidden units is set to 100. The number of training steps is set to 500 with a batch size of 64. The number of neighbors is 11 for KNN. We set $\beta = 1$. Default values are used for unmentioned hyperparameters.

We conduct experiments on several real-world datasets, including `core15k` (Duygulu et al., 2002), `Emotion6` (Peng et al., 2015), `flags` (Gonçalves et al., 2013), `Stackex_chess` (Charte et al., 2015), `GpositivePseAAC`, `PlantPseAAC`, `GnegativePseAAC` and `VirusPseAAC` (Xu et al., 2016), having statistics reported in Table IV. The ground truth labels in these dataset are either originally probabilistic or converted to a uniform distribution for multi-label classification datasets. At the beginning of each run, we randomly choose 80% of the dataset as the training set and the remaining 20% for evaluation. We further take 20% of the training set as the validation set to determine the best parameter for final testing.

We repeat the above process 100 times for each dataset on a laptop with a 2.7 GHz Quad-Core Intel Core i7 CPU. All the methods take less than 1 minute per run in wall time. The results in Table IV show that our proposed method outperforms the baselines in most of the adopted datasets or achieves similar performance to the best method with no statistical significance.

For sensitivity analysis, we fix a random split of the `Stackex_chess` dataset and vary β with other settings unchanged. The experiments are repeated 10 times. As shown in Figure 7, the expected loss of our method on the test set is slightly better than MLP when β is small but

TABLE IV: Dataset statistics and normalized generalization losses with 95% confidence intervals on each dataset. The best results are indicated in bold. † indicates statistical significance with paired t-test ($p < 0.05$).

Dataset	core15k	GnegativePseAAC	Emotion6	flags
n	5000	1392	1980	194
$ \mathcal{Y} $	374	8	7	7
Features	499	440	300	19
UNINF	$2.912 \pm 0.002^\dagger$	$0.367 \pm 0.003^\dagger$	$1.347 \pm 0.001^\dagger$	$1.315 \pm 0.004^\dagger$
MLP	$2.700 \pm 0.004^\dagger$	0.308 ± 0.003	1.343 ± 0.001	1.306 ± 0.007
KNN	$3.783 \pm 0.013^\dagger$	$0.324 \pm 0.004^\dagger$	$1.374 \pm 0.002^\dagger$	$1.353 \pm 0.012^\dagger$
Ours	2.696 ± 0.004	0.308 ± 0.003	$1.344 \pm 0.001^\dagger$	1.306 ± 0.007

Dataset	GpositivePseAAC	PlantPseAAC	Stackex_chess	VirusPseAAC
n	519	978	1672	207
$ \mathcal{Y} $	4	12	227	6
Features	440	440	585	440
UNINF	$0.385 \pm 0.004^\dagger$	$0.724 \pm 0.003^\dagger$	$2.720 \pm 0.005^\dagger$	$0.707 \pm 0.007^\dagger$
MLP	0.336 ± 0.005	0.668 ± 0.003	$2.522 \pm 0.009^\dagger$	0.684 ± 0.008
KNN	$0.344 \pm 0.005^\dagger$	$0.730 \pm 0.005^\dagger$	$3.448 \pm 0.014^\dagger$	$0.733 \pm 0.011^\dagger$
Ours	0.336 ± 0.005	0.668 ± 0.003	2.504 ± 0.008	0.686 ± 0.008

has large variance as β increases. In contrast, baselines including UNINF and KNN are trained obliviously to the final metric, thus not comparable to our method and MLP that minimize the target loss directly.

Additionally, we study the robustness of our approach by introducing noise to the training set of the **Stackex_chess** dataset, repeated 10 times. To this end, for each instance \mathbf{x} , with a probability p_{noise} , we replace the ground truth by a random distribution from $\mathcal{P}(\mathcal{Y})$. We vary p_{noise} from 0 to 0.5. As seen in Figure 7, our method is slightly better when $p_{\text{noise}} < 0.3$ and becomes vulnerable for large p_{noise} possibly because of the backbone neural network model.

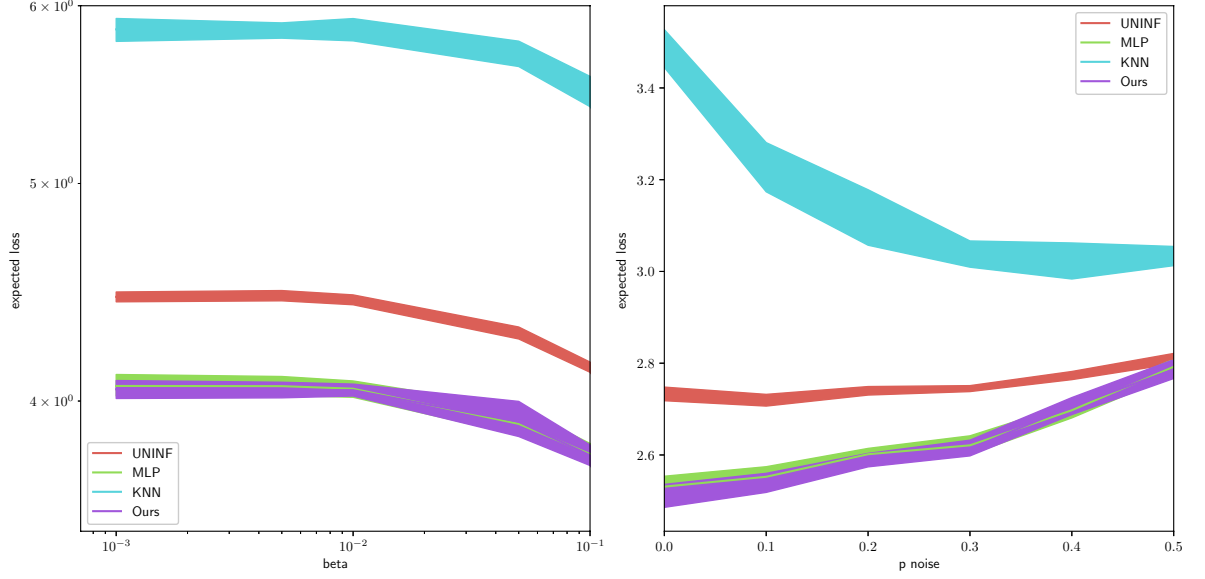


Figure 7: Normalized generalization losses with different coefficients or noise levels. Left: varying β in $[0.001, 0.1]$. Right: varying probability of contamination in $[0, 0.5]$. The X axes of the left subfigure is in logarithmic scale. Best viewed in color.

5.5 Concluding Remarks

We propose a moment-based distributionally robust learning framework for probabilistic supervised learning under mild assumptions, show its equivalence to dual-norm regularization for a surrogate loss, present its out-of-sample guarantees, develop efficient algorithms for typical continuous proper losses, incorporate the proposed method into differentiable learning and conduct experiments on several real-world datasets.

CHAPTER 6

CONCLUSION AND DISCUSSION

In this thesis, we study several structural learning problems from the perspective of distributionally robust optimization. Specifically, we propose a statistical learning framework for learning the structure of a discrete pairwise Markov network and a Bayesian network, as well as learning a structure mapping for tree-shaped objects and objects in a simplex. Based on the Wasserstein distances, KL divergences and feature moments, we show that the proposed methods are computationally efficient, sample efficient, Fisher consistent and robust at the same time. Extensive experimental results showcase their generalization ability and robustness under varying data contamination. This thesis illustrates a powerful framework for data-driven structural problems under high uncertainty. We expect our work to inspire similar or complementary structural learning paradigms and practical machine learning algorithms in the era of big data.

In the following sections, we point out limitations, future work and potential societal impacts of our methods.

6.1 Structure Learning

Formulating the complete DAG learning problem as one optimization problem may lead to a non-convex problem. A crucial challenge that leads to such non-convexity is the acyclicity constraint on the output graph. Existing methods either characterize the acyclicity constraint by matrix exponentials (Zheng et al., 2018) or simply optimize over the space of permutations

of nodes (Park and Klabjan, 2017), both of which lead to a highly non-convex problem whose global optima are difficult to find exactly. In Chapter 3, we focus on skeleton learning with provable guarantees and rely on existing orientation determination methods to produce the final DAG. We argue that recovering the exact skeleton with polynomial time complexities and sample complexities is perhaps the best we can expect. Figuring out the directionalities is closely related to the graph theoretic nature of the problem, which is dependent on fundamental results in computer science. Nonetheless, leveraging a principled adversarial training approach (Sinha et al., 2018) with advanced representation learning models is a promising future direction to pursue for practical use. For example, we may seek the following DRO estimator:

$$\inf_{\mathcal{E} \in \mathcal{F}, \mathbf{W} \in \mathbb{R}^{n \times n}, \text{tr}(e^{\mathbf{W} \circ \mathbf{W}}) - n = 0} \sup_{\mathbb{Q} \in \mathcal{A}} \mathbb{E}_{\mathbb{Q}} \|\mathcal{E}(\mathbf{X}) - \mathbf{W}\mathcal{E}(\mathbf{X})\|,$$

where the minimization is taken over a highly non-convex set, thus in fundamental contrast to the convex optimization problems considered in this dissertation. Since continuous optimization approaches with neural networks may lead to a trivial solution (Wei et al., 2020), a natural question is to what extent can a representation learning model $\mathcal{E}(\cdot)$ take advantage of DRO to alleviate these issues given the fact that a globally optimal solution cannot be found in usual. It is also questionable whether distributional robustness is harmful to DAG learning when the performance metric is on structures rather than statistical distances.

Despite absolute continuity, KL divergence usually allows a DRO problem to have a simple dual problem and good statistical guarantees, as shown in this thesis. These formulations have

been shown to recover adversarial reweighting (Li and Dunson, 2020), which is intuitive given that likelihood ratios are explicitly used. KL DRO is known to lead to pathological distributions compared to Wasserstein DRO that incorporates a notion of closeness and encompasses sound measure concentration guarantees. However, the empirical results illustrated in the thesis cast doubt on why a pathological distribution absolutely continuous with respect to the empirical distribution whose support is much sparse in the space of all possible states is in any way beneficial to successful learning. What kind of graphical models are they capable of dealing with? For example, a distribution in structure learning that makes KL DRO fail but Wasserstein DRO succeed. Be that as it may, a method would be highly desirable if it combines the efficiency of KL DRO and the stronger generalization ability of Wasserstein DRO.

We observe superior performance of Wasserstein DRO methods in structure learning problems. A noteworthy drawback is the more expensive computational cost. For undirected graphical models, the per-iteration costs $\tilde{O}(nk + n \log n)$ and $\tilde{O}(nk)$ in terms of n and k to optimize our objectives may not be improved further unless approximate gradient computation is acceptable. However, faster overall convergence rates (e.g., better than $\tilde{O}(n^2k^2)$) are possible if we replace L-BFGS-B with advanced optimization methods designed for DRO (Yu et al., 2021a; Namkoong and Duchi, 2016). Similar approaches based on stochastic gradient descents and more efficient approximate algorithms would work for Bayesian networks as well. In practice, especially in large-scale settings, it is often desirable to sacrifice optimality for a much more efficient algorithm that yields a sub-optimal but reasonably feasible solution.

Although robust to a set of adversarial distributions, our structure learning estimators may not be superior to robust estimators tailored to a certain class of contamination models or parametric distributions, for instance, (Goel et al., 2019; Prasad et al., 2020; Diakonikolas et al., 2021). In order to match these approaches in this case, one may consider incorporating prior distributional information to reshape ambiguity sets. For example, one can construct the set by adding all possible noises to the nominal distribution or by including all parametric distributions with the assumption on a parametric form of noises.

Furthermore, we are also curious about a general characterization of the conditions under which a structural learning problem has a tractable exact reformulation for Wasserstein DRO. Namely, what are the sufficient conditions in terms of functional analysis for tractability with general transport-based ambiguity sets? The metric that defines a Wasserstein ambiguity set is also crucial for a tractable reformulation (Nguyen et al., 2020). Formulating such conditions is beneficial to understanding both distributionally robust structure learning problems and DRO problems that study discrete distributions.

6.1.1 Structured Prediction

We address structured prediction problems with moment-based ambiguity sets. It is unclear if other types of ambiguity sets lead to Fisher consistency as well. Intuitively, a Wasserstein ambiguity set is expected to induce consistency due to its measure concentration results. However, the computational difficulty may be similar to a ERM approach with an identical loss function. Moreover, the Fenchel-Young loss framework (Blondel et al., 2020) is very similar to a

moment DRO framework. Specifically, the dual problem of moment-based DRO is equivalent to minimizing a Fenchel-Young loss with zero regularization and shared feature parameters.

In addition to consistency, the finite-sample guarantees and algorithms derived in the thesis should be easily generalizable to other structured prediction problems. An important challenge lies in developing efficient projection oracles for a polytope of specific structures of interest. We introduce a few quadratic terms to induce strong convexity, which, however, elicits such projection oracles. In tasks with more complicated structures and high-order structured prediction, computing a Euclidean projection may not be tractable. The max-min oracle proposed in (Nowak-Vila et al., 2020) is a Frank-Wolfe algorithm with an $\mathcal{O}(\frac{1}{\epsilon})$ convergence rate. The next step to improve on this is to propose a unified DRO framework for general structured prediction possibly without projection or with better projection oracles whose convergence rate is better than $\mathcal{O}(\frac{1}{\epsilon})$.

We assume that an expressive feature mapping is given such that a sufficiently good linear discriminant rule can be learned. This is aligned with the assumption of Fisher consistency that all the measurable functions are available. The class-sensitive form $\phi(\mathbf{x}, \mathbf{y})$ is general but consumes more memory than the decomposable form $\phi_{\mathbf{X}}(\mathbf{x}) \otimes \phi_{\mathbf{Y}}(\mathbf{y})$. A deep learning model typically transforms the original feature $\phi_{\mathbf{X}}(\mathbf{x})$ and maps the transformed features to logits $\phi_{\mathbf{Y}}(\phi_{\mathbf{X}}(\mathbf{x}))$ that encode its belief on conditional labels. What effect is this subtle difference in computational and statistical perspectives?

In the probabilistic supervised learning problem, a drawback of our method is that solving the saddle-point problem can be difficult for complicated losses. Neural networks equipped with

a soft-max layer makes use of automatic differentiation to avoid facing this issue though local optima are returned. Harnessing powerful representation learning tools to jointly learn data representation and parameters in our methods is worth further exploring. We are wondering how well can the learning framework enhance the performance of a deep learning model on consistent uncertainty estimation regardless of its lacking theoretical guarantees. What data corruption is the learned representation robust to? Are they strong against adversarial attacks on the feature space?

For some other future work, it would be interesting to extend the proposed DRO approaches to continuous higher-order graphical models and conditional density estimation. Reducing computational costs is expected to benefit all of the proposed methods. Another direction to consider is to adopt ambiguity sets based on higher-order moments (de Klerk et al., 2020).

6.2 Potential Societal Impacts

Potential negative societal impacts of our work depend on applications. For example, the structure of a private network could be revealed if the underlying graph satisfies certain assumptions. For voting network analysis, our method can help understand relation between voters. However, without appropriate tuning, the recovered structure could mislead specific decisions. Its robustness could also filter out outlier data that are possibly representative of minority groups. Moreover, using the prediction for decision-making in crucial clinical scenarios without verification may be harmful to subjects. Therefore, users should be careful to apply our methods to guide human-centered design.

CITED LITERATURE

- [Abadeh et al. , 2015]Abadeh, S. S., Esfahani, P. M. M., and Kuhn, D.: Distributionally robust logistic regression. In Advances in Neural Information Processing Systems, pages 1576–1584, 2015.
- [Altun and Smola, 2006]Altun, Y. and Smola, A.: Unifying divergence minimization and statistical inference via convex duality. In International Conference on Computational Learning Theory, pages 139–153. Springer, 2006.
- [Asif et al. , 2015]Asif, K., Xing, W., Behpour, S., and Ziebart, B. D.: Adversarial cost-sensitive classification. In Proceedings of the Conference on Uncertainty in Artificial Intelligence, pages 92–101, 2015.
- [Bank and Honorio, 2020]Bank, A. and Honorio, J.: Provable efficient skeleton learning of encodable discrete bayes nets in poly-time and sample complexity. In 2020 IEEE International Symposium on Information Theory (ISIT), pages 2486–2491. IEEE, 2020.
- [Bansal et al. , 2019]Bansal, K., Loos, S., Rabe, M., Szegedy, C., and Wilcox, S.: Holist: An environment for machine learning of higher order logic theorem proving. In International Conference on Machine Learning, pages 454–463. PMLR, 2019.
- [Bartlett and Cussens, 2017]Bartlett, M. and Cussens, J.: Integer linear programming for the bayesian network structure learning problem. Artificial Intelligence, 244:258–271, 2017.
- [Bartlett and Mendelson, 2002]Bartlett, P. L. and Mendelson, S.: Rademacher and gaussian complexities: Risk bounds and structural results. Journal of Machine Learning Research, 3(Nov):463–482, 2002.
- [Bayraksan and Love, 2015]Bayraksan, G. and Love, D. K.: Data-driven stochastic programming using phi-divergences. In The operations research revolution, pages 1–19. INFORMS, 2015.
- [Ben-Tal et al. , 2013]Ben-Tal, A., Den Hertog, D., De Waegenaere, A., Melenberg, B., and Rennen, G.: Robust solutions of optimization problems affected by uncertain probabilities. Management Science, 59(2):341–357, 2013.

- [Ben-Tal et al. , 2009]Ben-Tal, A., El Ghaoui, L., and Nemirovski, A.: Robust optimization. Princeton university press, 2009.
- [Bendale and Boulton, 2016]Bendale, A. and Boulton, T. E.: Towards open set deep networks. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 1563–1572, 2016.
- [Bertsimas et al. , 2011]Bertsimas, D., Brown, D. B., and Caramanis, C.: Theory and applications of robust optimization. SIAM review, 53(3):464–501, 2011.
- [Bertsimas and Copenhaver, 2018]Bertsimas, D. and Copenhaver, M. S.: Characterization of the equivalence of robustification and regularization in linear and matrix regression. European Journal of Operational Research, 270(3):931–942, 2018.
- [Bertsimas et al. , 2022]Bertsimas, D., Imai, K., and Li, M. L.: Distributionally robust causal inference with observational data. arXiv preprint arXiv:2210.08326, 2022.
- [Beygelzimer et al. , 2006]Beygelzimer, A., Kakade, S., and Langford, J.: Cover trees for nearest neighbor. In Proceedings of the 23rd international conference on Machine learning, pages 97–104, 2006.
- [Bickel, 2007]Bickel, J. E.: Some comparisons among quadratic, spherical, and logarithmic scoring rules. Decision Analysis, 4(2):49–65, 2007.
- [Blanchet et al. , 2019a]Blanchet, J., Glynn, P. W., Yan, J., and Zhou, Z.: Multivariate distributionally robust convex regression under absolute error loss. In Advances in Neural Information Processing Systems, pages 11794–11803, 2019.
- [Blanchet et al. , 2019b]Blanchet, J., Kang, Y., and Murthy, K.: Robust wasserstein profile inference and applications to machine learning. Journal of Applied Probability, 56(3):830–857, 2019.
- [Blanchet and Murthy, 2019]Blanchet, J. and Murthy, K.: Quantifying distributional model risk via optimal transport. Mathematics of Operations Research, 44(2):565–600, 2019.
- [Blondel, 2019]Blondel, M.: Structured prediction with projection oracles. Advances in Neural Information Processing Systems, 32:12145–12156, 2019.
- [Blondel et al. , 2020]Blondel, M., Martins, A. F., and Niculae, V.: Learning with Fenchel-Young losses. J. Mach. Learn. Res., 21(35):1–69, 2020.

- [Boissard and others, 2011]Boissard, E. et al.: Simple bounds for the convergence of empirical and occupation measures in 1-wasserstein distance. Electronic Journal of Probability, 16:2296–2333, 2011.
- [Bolley et al. , 2007]Bolley, F., Guillin, A., and Villani, C.: Quantitative concentration inequalities for empirical measures on non-compact spaces. Probability Theory and Related Fields, 137:541–593, 2007.
- [Boyd et al. , 2011]Boyd, S., Parikh, N., and Chu, E.: Distributed optimization and statistical learning via the alternating direction method of multipliers. Now Publishers Inc, 2011.
- [Bradley and Guestrin, 2010]Bradley, J. K. and Guestrin, C.: Learning tree conditional random fields. In Proceedings of the 27th International Conference on International Conference on Machine Learning, pages 127–134, 2010.
- [Bresler, 2015]Bresler, G.: Efficiently learning ising models on arbitrary graphs. In Proceedings of the forty-seventh annual ACM symposium on Theory of computing, pages 771–782, 2015.
- [Bresler et al. , 2013]Bresler, G., Mossel, E., and Sly, A.: Reconstruction of Markov random fields from samples: Some observations and algorithms. SIAM Journal on Computing, 42(2):563–578, 2013.
- [Brier, 1950]Brier, G. W.: Verification of forecasts expressed in terms of probability. Monthly weather review, 78(1):1–3, 1950.
- [Brooks, 1998]Brooks, S.: Markov chain Monte Carlo method and its application. Journal of the Royal Statistical Society: Series D (the Statistician), 47(1):69–100, 1998.
- [Byrd et al. , 1995]Byrd, R. H., Lu, P., Nocedal, J., and Zhu, C.: A limited memory algorithm for bound constrained optimization. SIAM Journal on scientific computing, 16(5):1190–1208, 1995.
- [Calafiore and El Ghaoui, 2006]Calafiore, G. C. and El Ghaoui, L.: On distributionally robust chance-constrained linear programs. Journal of Optimization Theory and Applications, 130(1):1–22, 2006.
- [Chapman, 2012]Chapman, L.: Probabilistic road weather forecasting. In Proceedings of the 16th SIRWEC Conference, Helsinki, Finland, May 2012, 2012.

- [Charikar and Wirth, 2004]Charikar, M. and Wirth, A.: Maximizing quadratic programs: Extending grothendieck’s inequality. In 45th Annual IEEE Symposium on Foundations of Computer Science, pages 54–60. IEEE, 2004.
- [Charte et al. , 2015]Charte, F., Rivera, A. J., del Jesus, M. J., and Herrera, F.: Quinta: A question tagging assistant to improve the answering ratio in electronic forums. In Ieee eurocon 2015-international conference on computer as a tool (eurocon), pages 1–6. IEEE, 2015.
- [Chayes et al. , 1984]Chayes, J., Chayes, L., and Lieb, E. H.: The inverse problem in classical statistical mechanics. Communications in Mathematical Physics, 93(1):57–121, 1984.
- [Chen and Manning, 2014]Chen, D. and Manning, C. D.: A fast and accurate dependency parser using neural networks. In Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP), pages 740–750, 2014.
- [Chen and Paschalidis, 2018]Chen, R. and Paschalidis, I. C.: A robust learning approach for regression models based on distributionally robust optimization. Journal of Machine Learning Research, 19(13), 2018.
- [Chen et al. , 2019]Chen, W., Drton, M., and Wang, Y. S.: On causal discovery with an equal-variance assumption. Biometrika, 106(4):973–980, 2019.
- [Chen, 2010]Chen, Y.: Learning sparse Ising models with missing data. Stanford University, 2010.
- [Chen et al. , 2019]Chen, Z., Sim, M., and Xu, H.: Distributionally robust optimization with infinitely constrained ambiguity sets. Operations Research, 67(5):1328–1344, 2019.
- [Cheng et al. , 2010]Cheng, W., Hüllermeier, E., and Dembczynski, K. J.: Bayes optimal multilabel classification via probabilistic classifier chains. In Proceedings of the 27th international conference on machine learning (ICML-10), pages 279–286, 2010.
- [Chickering, 2002]Chickering, D. M.: Optimal structure identification with greedy search. Journal of machine learning research, 3(Nov):507–554, 2002.
- [Chickering et al. , 2004]Chickering, M., Heckerman, D., and Meek, C.: Large-sample learning of bayesian networks is np-hard. Journal of Machine Learning Research, 5:1287–1330, 2004.
- [Chow and Liu, 1968]Chow, C. and Liu, C.: Approximating discrete probability distributions with dependence trees. IEEE transactions on Information Theory, 14(3):462–467, 1968.

- [Ciliberto et al. , 2019]Ciliberto, C., Bach, F., and Rudi, A.: Localized structured prediction. Advances in Neural Information Processing Systems, 32, 2019.
- [Ciliberto et al. , 2016]Ciliberto, C., Rosasco, L., and Rudi, A.: A consistent regularization approach for structured prediction. Advances in neural information processing systems, 29:4412–4420, 2016.
- [Cisneros-Velarde et al. , 2020]Cisneros-Velarde, P., Petersen, A., and Oh, S.-Y.: Distributionally robust formulation and model selection for the graphical lasso. In International Conference on Artificial Intelligence and Statistics, pages 756–765. PMLR, 2020.
- [Collins et al. , 2002]Collins, M., Schapire, R. E., and Singer, Y.: Logistic regression, adaboost and bregman distances. Machine Learning, 48(1-3):253–285, 2002.
- [Colombo et al. , 2014]Colombo, D., Maathuis, M. H., et al.: Order-independent constraint-based causal structure learning. J. Mach. Learn. Res., 15(1):3741–3782, 2014.
- [Condat, 2016]Condat, L.: Fast projection onto the simplex and the l_1 ball. Mathematical Programming, 158(1-2):575, 2016.
- [Conneau et al. , 2020]Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, É., Ott, M., Zettlemoyer, L., and Stoyanov, V.: Unsupervised cross-lingual representation learning at scale. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 8440–8451, 2020.
- [Constantinou et al. , 2021]Constantinou, A. C., Liu, Y., Chobtham, K., Guo, Z., and Kitson, N. K.: Large-scale empirical validation of bayesian network structure learning algorithms with noisy data. International Journal of Approximate Reasoning, 131:151–188, 2021.
- [Cortes et al. , 2015]Cortes, C., Kuznetsov, V., Mohri, M., and Syed, U.: Structural maxent models. In International Conference on Machine Learning, pages 391–399. PMLR, 2015.
- [Cranko et al. , 2021]Cranko, Z., Shi, Z., Zhang, X., Nock, R., and Kornblith, S.: Generalised lipschitz regularisation equals distributional robustness. In International Conference on Machine Learning, pages 2178–2188. PMLR, 2021.
- [Cui et al. , 2020]Cui, Y., Che, W., Liu, T., Qin, B., Wang, S., and Hu, G.: Revisiting pre-trained models for chinese natural language processing. In Findings of the Association for Computational Linguistics: EMNLP 2020, pages 657–668, 2020.

- [Daneshmand et al. , 2014]Daneshmand, H., Gomez-Rodriguez, M., Song, L., and Schoelkopf, B.: Estimating diffusion network structures: Recovery conditions, sample complexity & soft-thresholding algorithm. In International conference on machine learning, pages 793–801. PMLR, 2014.
- [Dawid, 2007]Dawid, A. P.: The geometry of proper scoring rules. Annals of the Institute of Statistical Mathematics, 59(1):77–93, 2007.
- [De Campos et al. , 2009]De Campos, C. P., Zeng, Z., and Ji, Q.: Structure learning of bayesian networks using constraints. In Proceedings of the 26th Annual International Conference on Machine Learning, pages 113–120, 2009.
- [de Klerk et al. , 2020]de Klerk, E., Kuhn, D., and Postek, K.: Distributionally robust optimization with polynomial densities: theory, models and algorithms. Mathematical Programming, 181:265–296, 2020.
- [De Marneffe and Manning, 2008]De Marneffe, M.-C. and Manning, C. D.: The Stanford typed dependencies representation. In Coling 2008: proceedings of the workshop on cross-framework and cross-domain parser evaluation, pages 1–8, 2008.
- [Delage and Ye, 2010]Delage, E. and Ye, Y.: Distributionally robust optimization under moment uncertainty with application to data-driven problems. Operations research, 58(3):595–612, 2010.
- [Deng and Yin, 2016]Deng, W. and Yin, W.: On the global and linear convergence of the generalized alternating direction method of multipliers. Journal of Scientific Computing, 66(3):889–916, 2016.
- [Diakonikolas et al. , 2021]Diakonikolas, I., Kane, D. M., Stewart, A., and Sun, Y.: Outlier-robust learning of ising models under dobrushin’s condition. In Proceedings of Thirty Fourth Conference on Learning Theory, eds. M. Belkin and S. Kpotufe, volume 134 of Proceedings of Machine Learning Research, pages 1645–1682. PMLR, 15–19 Aug 2021.
- [Dozat and Manning, 2017]Dozat, T. and Manning, C. D.: Deep biaffine attention for neural dependency parsing. In 5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings. OpenReview.net, 2017.
- [Drton and Maathuis, 2017]Drton, M. and Maathuis, M. H.: Structure learning in graphical modeling. Annual Review of Statistics and Its Application, 4:365–393, 2017.

- [Dubois et al. , 2020]Dubois, Y., Kiela, D., Schwab, D. J., and Vedantam, R.: Learning optimal representations with the decodable information bottleneck. Advances in Neural Information Processing Systems, 33:18674–18690, 2020.
- [Duchi and Namkoong, 2019]Duchi, J. and Namkoong, H.: Variance-based regularization with convex objectives. The Journal of Machine Learning Research, 20(1):2450–2504, 2019.
- [Duygulu et al. , 2002]Duygulu, P., Barnard, K., de Freitas, J. F., and Forsyth, D. A.: Object recognition as machine translation: Learning a lexicon for a fixed image vocabulary. In European conference on computer vision, pages 97–112. Springer, 2002.
- [Dyer et al. , 2015]Dyer, C., Ballesteros, M., Ling, W., Matthews, A., and Smith, N. A.: Transition-based dependency parsing with stack long short-term memory. In Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), pages 334–343, 2015.
- [Eagle et al. , 2009]Eagle, N., Pentland, A. S., and Lazer, D.: Inferring friendship network structure by using mobile phone data. Proceedings of the national academy of sciences, 106(36):15274–15278, 2009.
- [Elsberry, 2002]Elsberry, R. L.: Predicting hurricane landfall precipitation: Optimistic and pessimistic views from the symposium on precipitation extremes. Bulletin of the American Meteorological Society, 83(9):1333–1339, 2002.
- [Esfahani and Kuhn, 2018]Esfahani, P. M. and Kuhn, D.: Data-driven distributionally robust optimization using the Wasserstein metric: Performance guarantees and tractable reformulations. Mathematical Programming, 171(1-2):115–166, 2018.
- [Farnia and Tse, 2016]Farnia, F. and Tse, D.: A minimax approach to supervised learning. In Advances in Neural Information Processing Systems, pages 4240–4248, 2016.
- [Fathony et al. , 2018]Fathony, R., Behpour, S., Zhang, X., and Ziebart, B.: Efficient and consistent adversarial bipartite matching. In International Conference on Machine Learning, pages 1457–1466, 2018.
- [Fathony et al. , 2016]Fathony, R., Liu, A., Asif, K., and Ziebart, B.: Adversarial multiclass classification: A risk minimization perspective. Advances in Neural Information Processing Systems, 29:559–567, 2016.

- [Fathony et al. , 2018]Fathony, R., Rezaei, A., Bashiri, M. A., Zhang, X., and Ziebart, B.: Distributionally robust graphical models. Advances in Neural Information Processing Systems, 31, 2018.
- [Fournier and Guillin, 2015]Fournier, N. and Guillin, A.: On the rate of convergence in wasserstein distance of the empirical measure. Probability theory and related fields, 162(3-4):707–738, 2015.
- [Frank et al. , 1956]Frank, M., Wolfe, P., et al.: An algorithm for quadratic programming. Naval research logistics quarterly, 3(1-2):95–110, 1956.
- [Friedman et al. , 2000]Friedman, J., Hastie, T., and Tibshirani, R.: Additive logistic regression: a statistical view of boosting (with discussion and a rejoinder by the authors). The annals of statistics, 28(2):337–407, 2000.
- [Friedman et al. , 2008]Friedman, J., Hastie, T., and Tibshirani, R.: Sparse inverse covariance estimation with the graphical lasso. Biostatistics, 9(3):432–441, 2008.
- [Friesen, 2019]Friesen, M.: Extended formulations for higher order polytopes in combinatorial optimization. Doctoral dissertation, Otto von Guericke University Magdeburg, 2019.
- [Gabow et al. , 1986]Gabow, H. N., Galil, Z., Spencer, T., and Tarjan, R. E.: Efficient algorithms for finding minimum spanning trees in undirected and directed graphs. Combinatorica, 6(2):109–122, 1986.
- [Ganian and Korchemna, 2021]Ganian, R. and Korchemna, V.: The complexity of bayesian network learning: Revisiting the superstructure. Advances in Neural Information Processing Systems, 34:430–442, 2021.
- [Gao et al. , 2020]Gao, M., Ding, Y., and Aragam, B.: A polynomial-time algorithm for learning non-parametric causal graphs. Advances in Neural Information Processing Systems, 33:11599–11611, 2020.
- [Gao et al. , 2022]Gao, R., Chen, X., and Kleywegt, A. J.: Wasserstein distributionally robust optimization and variation regularization. Operations Research, 2022.
- [Gao and Kleywegt, 2022]Gao, R. and Kleywegt, A.: Distributionally robust stochastic optimization with wasserstein distance. Mathematics of Operations Research, 2022.

- [Gao et al. , 2022]Gao, T., Bhattacharjya, D., Nelson, E., Liu, M., and Yu, Y.: Idyno: Learning nonparametric dags from interventional dynamic data. In International Conference on Machine Learning, pages 6988–7001. PMLR, 2022.
- [Gasse et al. , 2014]Gasse, M., Aussem, A., and Elghazel, H.: A hybrid algorithm for bayesian network structure learning with application to multi-label learning. Expert Systems with Applications, 41(15):6755–6772, 2014.
- [Gast and Roth, 2018]Gast, J. and Roth, S.: Lightweight probabilistic deep networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 3369–3378, 2018.
- [Gelman et al. , 2014]Gelman, A., Hwang, J., and Vehtari, A.: Understanding predictive information criteria for bayesian models. Statistics and computing, 24(6):997–1016, 2014.
- [Geng, 2016]Geng, X.: Label distribution learning. IEEE Transactions on Knowledge and Data Engineering, 28(7):1734–1748, 2016.
- [Ghoshal and Honorio, 2017]Ghoshal, A. and Honorio, J.: Learning identifiable gaussian bayesian networks in polynomial time and sample complexity. Advances in Neural Information Processing Systems, 30, 2017.
- [Ghoshal and Honorio, 2018]Ghoshal, A. and Honorio, J.: Learning linear structural equation models in polynomial time and sample complexity. In International Conference on Artificial Intelligence and Statistics, pages 1466–1475. PMLR, 2018.
- [Gimpel and Smith, 2010]Gimpel, K. and Smith, N. A.: Softmax-margin crfs: Training log-linear models with cost functions. In Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics, pages 733–736, 2010.
- [Gneiting and Raftery, 2007]Gneiting, T. and Raftery, A. E.: Strictly proper scoring rules, prediction, and estimation. Journal of the American statistical Association, 102(477):359–378, 2007.
- [Goel et al. , 2019]Goel, S., Kane, D. M., and Klivans, A. R.: Learning Ising models with independent failures. In Conference on Learning Theory, pages 1449–1469, 2019.
- [Goh and Sim, 2010]Goh, J. and Sim, M.: Distributionally robust optimization and its tractable approximations. Operations research, 58(4-part-1):902–917, 2010.

- [Gonçalves et al. , 2013]Gonçalves, E. C., Plastino, A., and Freitas, A. A.: A genetic algorithm for optimizing the label ordering in multi-label classifier chains. In 2013 IEEE 25th International Conference on Tools with Artificial Intelligence, pages 469–476. IEEE, 2013.
- [Good, 1952]Good, I.: Rational decisions. Journal of the Royal Statistical Society. Series B (Methodological), 14(1):107–114, 1952.
- [Gormley et al. , 2015]Gormley, M. R., Dredze, M., and Eisner, J.: Approximation-aware dependency parsing by belief propagation. Transactions of the Association for Computational Linguistics, 3:489–501, 2015.
- [Grbovic and Cheng, 2018]Grbovic, M. and Cheng, H.: Real-time personalization using embeddings for search ranking at airbnb. In Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining, pages 311–320, 2018.
- [Gressmann et al. , 2018]Gressmann, F., Király, F. J., Mateen, B., and Oberhauser, H.: Probabilistic supervised learning. arXiv preprint arXiv:1801.00753, 2018.
- [Grünwald and Dawid, 2004]Grünwald, P. D. and Dawid, A. P.: Game theory, maximum entropy, minimum discrepancy and robust Bayesian decision theory. the Annals of Statistics, 32(4):1367–1433, 2004.
- [Guo et al. , 2017]Guo, C., Pleiss, G., Sun, Y., and Weinberger, K. Q.: On calibration of modern neural networks. In International Conference on Machine Learning, pages 1321–1330. PMLR, 2017.
- [Hager and Zhang, 2020]Hager, W. W. and Zhang, H.: Convergence rates for an inexact admm applied to separable convex optimization. Computational Optimization and Applications, 77(3):729–754, 2020.
- [Hamilton et al. , 2017]Hamilton, L., Koehler, F., and Moitra, A.: Information theoretic properties of Markov random fields, and their algorithmic applications. In Advances in Neural Information Processing Systems, pages 2463–2472, 2017.
- [Hastie et al. , 2009]Hastie, T., Tibshirani, R., Friedman, J., Hastie, T., Tibshirani, R., and Friedman, J.: Overview of supervised learning. The elements of statistical learning: Data mining, inference, and prediction, pages 9–41, 2009.
- [Hastie et al. , 2015]Hastie, T., Tibshirani, R., and Wainwright, M.: Statistical learning with sparsity. Monographs on statistics and applied probability, 143:143, 2015.

- [He et al. , 2018]He, X., He, Z., Du, X., and Chua, T.-S.: Adversarial personalized ranking for recommendation. In The 41st International ACM SIGIR conference on research & development in information retrieval, pages 355–364, 2018.
- [Heinze-Deml et al. , 2018]Heinze-Deml, C., Maathuis, M. H., and Meinshausen, N.: Causal structure learning. Annual Review of Statistics and Its Application, 5:371–391, 2018.
- [Hinton, 1990]Hinton, G. E.: Connectionist learning procedures. In Machine learning, pages 555–610. Elsevier, 1990.
- [Hu and Hong, 2013]Hu, Z. and Hong, L. J.: Kullback-leibler divergence constrained distributionally robust optimization. Available at Optimization Online, 2013.
- [Hyvärinen and Dayan, 2005]Hyvärinen, A. and Dayan, P.: Estimation of non-normalized statistical models by score matching. Journal of Machine Learning Research, 6(4), 2005.
- [Jaakkola et al. , 2010]Jaakkola, T., Sontag, D., Globerson, A., and Meila, M.: Learning bayesian network structure using lp relaxations. In Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics, pages 358–365. JMLR Workshop and Conference Proceedings, 2010.
- [Jaggi, 2013]Jaggi, M.: Revisiting Frank-Wolfe: Projection-free sparse convex optimization. In International Conference on Machine Learning, pages 427–435. PMLR, 2013.
- [Jaimovich et al. , 2006]Jaimovich, A., Elidan, G., Margalit, H., and Friedman, N.: Towards an integrated protein–protein interaction network: A relational markov network approach. Journal of Computational Biology, 13(2):145–164, 2006.
- [Jalali et al. , 2011a]Jalali, A., Johnson, C., and Ravikumar, P.: On learning discrete graphical models using greedy methods. Advances in Neural Information Processing Systems, 24, 2011.
- [Jalali et al. , 2011b]Jalali, A., Ravikumar, P., Vasuki, V., and Sanghavi, S.: On learning discrete graphical models using group-sparse regularization. In Proceedings of the fourteenth international conference on artificial intelligence and statistics, pages 378–387. JMLR Workshop and Conference Proceedings, 2011.
- [Jiang and Guan, 2016]Jiang, R. and Guan, Y.: Data-driven chance constrained stochastic program. Mathematical Programming, 158(1-2):291–327, 2016.

- [Joo et al. , 2020]Joo, T., Chung, U., and Seo, M.-G.: Being bayesian about categorical probability. In International Conference on Machine Learning, pages 4950–4961. PMLR, 2020.
- [Jordan, 1999]Jordan, M. I.: Learning in graphical models. MIT press, 1999.
- [Kantorovich and Rubinshtein, 1958]Kantorovich, L. V. and Rubinshtein, S.: On a space of totally additive functions. Vestnik of the St. Petersburg University: Mathematics, 13(7):52–59, 1958.
- [Katiyar et al. , 2021]Katiyar, A., Basu, S., Shah, V., and Caramanis, C.: Robust estimation of tree structured markov random fields. arXiv preprint arXiv:2102.08554, 2021.
- [Katiyar et al. , 2020]Katiyar, A., Shah, V., and Caramanis, C.: Robust estimation of tree structured ising models. arXiv preprint arXiv:2006.05601, 2020.
- [Kingma and Ba, 2014]Kingma, D. P. and Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980, 2014.
- [Kiperwasser and Goldberg, 2016]Kiperwasser, E. and Goldberg, Y.: Simple and accurate dependency parsing using bidirectional lstm feature representations. Transactions of the Association for Computational Linguistics, 4:313–327, 2016.
- [Kirchhoff, 1847]Kirchhoff, G.: Ueber die auflösung der gleichungen, auf welche man bei der untersuchung der linearen vertheilung galvanischer ströme geführt wird. Annalen der Physik, 148(12):497–508, 1847.
- [Klivans and Meka, 2017]Klivans, A. and Meka, R.: Learning graphical models using multiplicative weights. In 2017 IEEE 58th Annual Symposium on Foundations of Computer Science, pages 343–354. IEEE, 2017.
- [Koller and Friedman, 2009]Koller, D. and Friedman, N.: Probabilistic graphical models: principles and techniques. MIT press, 2009.
- [Koo et al. , 2007]Koo, T., Globerson, A., Carreras Pérez, X., and Collins, M.: Structured prediction models via the matrix-tree theorem. In Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL), pages 141–150, 2007.

- [Korhonen and Parviainen, 2013]Korhonen, J. and Parviainen, P.: Exact learning of bounded tree-width bayesian networks. In Artificial Intelligence and Statistics, pages 370–378. PMLR, 2013.
- [Kuhlman and Bradley, 2019]Kuhlman, B. and Bradley, P.: Advances in protein structure prediction and design. Nature Reviews Molecular Cell Biology, 20(11):681–697, 2019.
- [Kuhn et al. , 2019]Kuhn, D., Esfahani, P. M., Nguyen, V. A., and Shafieezadeh-Abadeh, S.: Wasserstein distributionally robust optimization: Theory and applications in machine learning. In Operations research & management science in the age of analytics, pages 130–166. Informa, 2019.
- [Kyrimi et al. , 2020]Kyrimi, E., Neves, M. R., McLachlan, S., Neil, M., Marsh, W., and Fenton, N.: Medical idioms for clinical bayesian network development. Journal of Biomedical Informatics, 108:103495, 2020.
- [Lam, 2016]Lam, H.: Robust sensitivity analysis for stochastic systems. Mathematics of Operations Research, 41(4):1248–1275, 2016.
- [Lam, 2019]Lam, H.: Recovering best statistical guarantees via the empirical divergence-based distributionally robust optimization. Operations Research, 67(4):1090–1105, 2019.
- [Lei and others, 2020]Lei, J. et al.: Convergence and concentration of empirical measures under wasserstein distance in unbounded functional spaces. Bernoulli, 26(1):767–798, 2020.
- [Leibovici et al. , 2000]Leibovici, L., Fishman, M., Schonheyder, H. C., Riekehr, C., Kristensen, B., Shraga, I., and Andreassen, S.: A causal probabilistic network for optimal treatment of bacterial infections. IEEE Transactions on Knowledge and Data Engineering, 12(4):517–528, 2000.
- [Levy et al. , 2020]Levy, D., Carmon, Y., Duchi, J. C., and Sidford, A.: Large-scale methods for distributionally robust optimization. Advances in Neural Information Processing Systems, 33:8847–8860, 2020.
- [Li and Dunson, 2020]Li, M. and Dunson, D. B.: Comparing and weighting imperfect models using d-probabilities. Journal of the American Statistical Association, 115(531):1349–1360, 2020.
- [Li et al. , 2022a]Li, Y., Saeed, D., Zhang, X., Ziebart, B. D., and Gimpel, K.: Moment distributionally robust tree structured prediction. Advances in Neural Information Processing Systems, 35, 2022.

- [Li et al. , 2022b]Li, Y., Shi, Z., Zhang, X., and Ziebart, B.: Distributionally robust structure learning for discrete pairwise markov networks. In International Conference on Artificial Intelligence and Statistics, pages 8997–9016. PMLR, 2022.
- [Li and Ziebart, 2023]Li, Y. and Ziebart, B. D.: Moment distributionally robust probabilistic supervised learning, 2023.
- [Li and Eisner, 2009]Li, Z. and Eisner, J.: First-and second-order expectation semirings with applications to minimum-risk training on translation forests. In Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing, pages 40–51, 2009.
- [Lin et al. , 2022]Lin, F., Fang, X., and Gao, Z.: Distributionally robust optimization: A review on theory and applications. Numerical Algebra, Control and Optimization, 12(1):159–212, 2022.
- [Lindgren et al. , 2019]Lindgren, E. M., Shah, V., Shen, Y., Dimakis, A. G., and Klivans, A.: On robust learning of ising models. In NeurIPS Workshop on Relational Representation Learning, 2019.
- [Liu et al. , 2019]Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., and Stoyanov, V.: Roberta: A robustly optimized BERT pretraining approach. CoRR, abs/1907.11692, 2019.
- [Liu, 2007]Liu, Y.: Fisher consistency of multicategory support vector machines. In Artificial Intelligence and Statistics, pages 291–298. PMLR, 2007.
- [Loh and Bühlmann, 2014]Loh, P.-L. and Bühlmann, P.: High-dimensional learning of linear causal networks via inverse covariance estimation. The Journal of Machine Learning Research, 15(1):3065–3105, 2014.
- [Lokhov et al. , 2018]Lokhov, A. Y., Vuffray, M., Misra, S., and Chertkov, M.: Optimal structure and parameter learning of ising models. Science advances, 4(3):e1700791, 2018.
- [Luo and Mehrotra, 2019]Luo, F. and Mehrotra, S.: Decomposition algorithm for distributionally robust optimization using wasserstein metric with an application to a class of regression models. European Journal of Operational Research, 278(1):20–35, 2019.
- [Lyu et al. , 2019]Lyu, S., Tian, X., Li, Y., Jiang, B., and Chen, H.: Multiclass probabilistic classification vector machine. IEEE Transactions on Neural Networks and Learning Systems, 2019.

- [Ma and Hovy, 2017]Ma, X. and Hovy, E.: Neural probabilistic model for non-projective MST parsing. In Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers), pages 59–69, 2017.
- [Malone et al. , 2015]Malone, B. M., Järvisalo, M., and Myllymäki, P.: Impact of learning strategies on the quality of bayesian networks: An empirical evaluation. In UAI, pages 562–571. Citeseer, 2015.
- [Manjusha and Kumar, 2010]Manjusha, K. and Kumar, R.: Spam mail classification using combined approach of bayesian and neural network. In 2010 International Conference on Computational Intelligence and Communication Networks, pages 145–149. IEEE, 2010.
- [Marcus et al. , 1993]Marcus, M., Santorini, B., and Marcinkiewicz, M. A.: Building a large annotated corpus of english: The penn treebank. Computational Linguistics, 19(2):313–330, 1993.
- [Martin, 1991]Martin, R. K.: Using separation algorithms to generate mixed integer model reformulations. Operations Research Letters, 10(3):119–128, 1991.
- [Martins et al. , 2015]Martins, A. F., Figueiredo, M. A., Aguiar, P. M., Smith, N. A., and Xing, E. P.: Ad3: Alternating directions dual decomposition for map inference in graphical models. The Journal of Machine Learning Research, 16(1):495–545, 2015.
- [Martins et al. , 2009]Martins, A. F., Smith, N. A., and Xing, E.: Concise integer linear programming formulations for dependency parsing. In Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP, pages 342–350, 2009.
- [Martins et al. , 2010]Martins, A. F., Smith, N. A., Xing, E., Aguiar, P., and Figueiredo, M.: Turbo parsers: Dependency parsing by approximate variational inference. In Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing, pages 34–44, 2010.
- [Masnadi-Shirazi and Vasconcelos, 2009]Masnadi-Shirazi, H. and Vasconcelos, N.: On the design of loss functions for classification: theory, robustness to outliers, and savageboost. In Advances in neural information processing systems, pages 1049–1056, 2009.
- [Mazuelas et al. , 2022]Mazuelas, S., Shen, Y., and Pérez, A.: Generalized maximum entropy for supervised classification. IEEE Transactions on Information Theory, 2022.

- [McDonald and Pereira, 2006]McDonald, R. and Pereira, F.: Online learning of approximate dependency parsing algorithms. In 11th Conference of the European Chapter of the Association for Computational Linguistics, 2006.
- [McDonald et al. , 2005]McDonald, R., Pereira, F., Ribarov, K., and Hajic, J.: Non-projective dependency parsing using spanning tree algorithms. In Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing, pages 523–530, 2005.
- [McDonald and Satta, 2007]McDonald, R. and Satta, G.: On the complexity of non-projective data-driven dependency parsing. In Proceedings of the Tenth International Conference on Parsing Technologies, pages 121–132, 2007.
- [McMahan et al. , 2003]McMahan, H. B., Gordon, G. J., and Blum, A.: Planning in the presence of cost functions controlled by an adversary. In Proceedings of the 20th International Conference on Machine Learning (ICML-03), pages 536–543, 2003.
- [Mehrotra and Zhang, 2014]Mehrotra, S. and Zhang, H.: Models and algorithms for distributionally robust least squares problems. Mathematical Programming, 146(1-2):123–141, 2014.
- [Mensch and Blondel, 2018]Mensch, A. and Blondel, M.: Differentiable dynamic programming for structured prediction and attention. In International Conference on Machine Learning, pages 3462–3471. PMLR, 2018.
- [Meshi et al. , 2013]Meshi, O., Eban, E., Elidan, G., and Globerson, A.: Learning max-margin tree predictors. In Proceedings of the Twenty-Ninth Conference on Uncertainty in Artificial Intelligence, pages 411–420, 2013.
- [Mihaylova et al. , 2020]Mihaylova, T., Niculae, V., and Martins, A. F.: Understanding the mechanics of spigot: Surrogate gradients for latent structure learning. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 2186–2202, 2020.
- [Murphy et al. , 1999]Murphy, K. P., Weiss, Y., and Jordan, M. I.: Loopy belief propagation for approximate inference: an empirical study. In Proceedings of the Fifteenth Conference on Uncertainty in Artificial Intelligence, pages 467–475, 1999.
- [Namkoong and Duchi, 2016]Namkoong, H. and Duchi, J. C.: Stochastic gradient methods for distributionally robust optimization with f-divergences. In NIPS, volume 29, pages 2208–2216, 2016.

- [Nandy et al. , 2018]Nandy, P., Hauser, A., and Maathuis, M. H.: High-dimensional consistency in score-based and hybrid structure learning. The Annals of Statistics, 46(6A):3151–3183, 2018.
- [Neath and Cavanaugh, 2012]Neath, A. A. and Cavanaugh, J. E.: The bayesian information criterion: background, derivation, and applications. Wiley Interdisciplinary Reviews: Computational Statistics, 4(2):199–203, 2012.
- [Nelder and Wedderburn, 1972]Nelder, J. A. and Wedderburn, R. W.: Generalized linear models. Journal of the Royal Statistical Society: Series A (General), 135(3):370–384, 1972.
- [Nesterov, 2003]Nesterov, Y.: Introductory lectures on convex optimization: A basic course, volume 87. Springer Science & Business Media, 2003.
- [Ng, 2004]Ng, A. Y.: Feature selection, l_1 vs. l_2 regularization, and rotational invariance. In Proceedings of the twenty-first international conference on Machine learning, page 78, 2004.
- [Ng et al. , 2020]Ng, I., Ghassami, A., and Zhang, K.: On the role of sparsity and dag constraints for learning linear dags. Advances in Neural Information Processing Systems, 33:17943–17954, 2020.
- [Ng et al. , 2021]Ng, I., Zheng, Y., Zhang, J., and Zhang, K.: Reliable causal discovery with improved exact search and weaker assumptions. Advances in Neural Information Processing Systems, 34:20308–20320, 2021.
- [Ng et al. , 2022]Ng, I., Zhu, S., Fang, Z., Li, H., Chen, Z., and Wang, J.: Masked gradient-based causal structure learning. In Proceedings of the 2022 SIAM International Conference on Data Mining (SDM), pages 424–432. SIAM, 2022.
- [Nguyen et al. , 2022]Nguyen, V. A., Kuhn, D., and Mohajerin Esfahani, P.: Distributionally robust inverse covariance estimation: The wasserstein shrinkage estimator. Operations Research, 70(1):490–515, 2022.
- [Nguyen et al. , 2020]Nguyen, V. A., Zhang, F., Blanchet, J., Delage, E., and Ye, Y.: Distributionally robust local non-parametric conditional estimation. In Advances in Neural Information Processing Systems, eds. H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, volume 33, pages 15232–15242. Curran Associates, Inc., 2020.

- [Nikolakakis et al. , 2019a]Nikolakakis, K. E., Kalogierias, D. S., and Sarwate, A. D.: Learning tree structures from noisy data. In The 22nd International Conference on Artificial Intelligence and Statistics, pages 1771–1782, 2019.
- [Nikolakakis et al. , 2019b]Nikolakakis, K. E., Kalogierias, D. S., and Sarwate, A. D.: Non-parametric structure learning on hidden tree-shaped distributions. arXiv preprint arXiv:1909.09596, 2019.
- [Nivre et al. , 2016]Nivre, J., De Marneffe, M.-C., Ginter, F., Goldberg, Y., Hajic, J., Manning, C. D., McDonald, R., Petrov, S., Pyysalo, S., Silveira, N., et al.: Universal dependencies v1: A multilingual treebank collection. In Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC’16), pages 1659–1666, 2016.
- [Nowak et al. , 2022]Nowak, A., Rudi, A., and Bach, F.: On the consistency of max-margin losses. In International Conference on Artificial Intelligence and Statistics, pages 4612–4633. PMLR, 2022.
- [Nowak-Vila et al. , 2019]Nowak-Vila, A., Bach, F., and Rudi, A.: A general theory for structured prediction with smooth convex surrogates. arXiv preprint arXiv:1902.01958, 2019.
- [Nowak-Vila et al. , 2020]Nowak-Vila, A., Bach, F., and Rudi, A.: Consistent structured prediction with max-min margin Markov networks. In Proceedings of the International Conference on Machine Learning (ICML), 2020.
- [Nowozin et al. , 2014]Nowozin, S., Gehler, P. V., Jancsary, J., and Lampert, C. H.: Advanced structured prediction. MIT Press, 2014.
- [Och, 2003]Och, F. J.: Minimum error rate training in statistical machine translation. In Proceedings of the 41st annual meeting of the Association for Computational Linguistics, pages 160–167, 2003.
- [Ordyniak and Szeider, 2013]Ordyniak, S. and Szeider, S.: Parameterized complexity results for exact bayesian network structure learning. Journal of Artificial Intelligence Research, 46:263–302, 2013.
- [Papadopoulos, 2013]Papadopoulos, H.: Reliable probabilistic classification with neural networks. Neurocomputing, 107:59–68, 2013.

- [Park et al. , 2021]Park, G., Moon, S. J., Park, S., and Jeon, J.-J.: Learning a high-dimensional linear structural equation model via ℓ_1 -regularized regression. The Journal of Machine Learning Research, 22(1):4607–4647, 2021.
- [Park and Raskutti, 2017]Park, G. and Raskutti, G.: Learning quadratic variance function (qvf) dag models via overdispersion scoring (ods). J. Mach. Learn. Res., 18:224–1, 2017.
- [Park and Klabjan, 2017]Park, Y. W. and Klabjan, D.: Bayesian network learning via topological order. The Journal of Machine Learning Research, 18(1):3451–3482, 2017.
- [Parviainen and Koivisto, 2009]Parviainen, P. and Koivisto, M.: Exact structure discovery in bayesian networks with less space. In Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence, pages 436–443, 2009.
- [Paszke et al. , 2019]Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al.: Pytorch: An imperative style, high-performance deep learning library. Advances in neural information processing systems, 32:8026–8037, 2019.
- [Pearl, 2009]Pearl, J.: Causality. Cambridge university press, 2009.
- [Pedregosa et al. , 2011]Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., et al.: Scikit-learn: Machine learning in python. the Journal of machine Learning research, 12:2825–2830, 2011.
- [Peng et al. , 2018]Peng, H., Thomson, S., and Smith, N. A.: Backpropagating through structured argmax using a spigot. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 1863–1873, 2018.
- [Peng et al. , 2015]Peng, K.-C., Chen, T., Sadovnik, A., and Gallagher, A. C.: A mixed bag of emotions: Model, predict, and transfer emotion distributions. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 860–868, 2015.
- [Perrier et al. , 2008]Perrier, E., Imoto, S., and Miyano, S.: Finding optimal bayesian network given a super-structure. Journal of Machine Learning Research, 9(10), 2008.
- [Peters et al. , 2017]Peters, J., Janzing, D., and Schölkopf, B.: Elements of causal inference: foundations and learning algorithms. The MIT Press, 2017.

- [Poon and Domingos, 2011]Poon, H. and Domingos, P.: Sum-product networks: A new deep architecture. In 2011 IEEE International Conference on Computer Vision Workshops (ICCV Workshops), pages 689–690. IEEE, 2011.
- [Popescu, 2007]Popescu, I.: Robust mean-covariance solutions for stochastic optimization. Operations Research, 55(1):98–112, 2007.
- [Prasad et al. , 2020]Prasad, A., Srinivasan, V., Balakrishnan, S., and Ravikumar, P.: On learning models under huber’s contamination model. Advances in Neural Information Processing Systems, 33, 2020.
- [Rahimian and Mehrotra, 2019]Rahimian, H. and Mehrotra, S.: Distributionally robust optimization: A review. arXiv preprint arXiv:1908.05659, 2019.
- [Rajendran et al. , 2021]Rajendran, G., Kivva, B., Gao, M., and Aragam, B.: Structure learning in polynomial time: Greedy algorithms, bregman information, and exponential families. Advances in Neural Information Processing Systems, 34:18660–18672, 2021.
- [Ramaswamy et al. , 2013]Ramaswamy, H. G., Agarwal, S., and Tewari, A.: Convex calibrated surrogates for low-rank loss matrices with applications to subset ranking losses. In Advances in Neural Information Processing Systems, pages 1475–1483, 2013.
- [Ravikumar et al. , 2010]Ravikumar, P., Wainwright, M. J., and Lafferty, J. D.: High-dimensional learning model selection using ℓ_1 -regularized logistic regression. The Annals of Statistics, 38(3):1287–1319, 2010.
- [Reid and Williamson, 2011]Reid, M. D. and Williamson, R. C.: Information, divergence and risk for binary experiments. The Journal of Machine Learning Research, 12:731–817, 2011.
- [Scarf, 1958]Scarf, H.: A min max solution of an inventory problem. Studies in the mathematical theory of inventory and production, 1958.
- [Schmidt et al. , 2009]Schmidt, M., Berg, E., Friedlander, M., and Murphy, K.: Optimizing costly functions with simple constraints: A limited-memory projected quasi-newton algorithm. In Artificial Intelligence and Statistics, pages 456–463. PMLR, 2009.
- [Schneidman et al. , 2006]Schneidman, E., Berry, M. J., Segev, R., and Bialek, W.: Weak pairwise correlations imply strongly correlated network states in a neural population. Nature, 440(7087):1007–1012, 2006.

- [Scutari, 2010]Scutari, M.: Learning bayesian networks with the bnlearn r package. Journal of Statistical Software, 35:1–22, 2010.
- [Shafieezadeh-Abadeh et al. , 2019]Shafieezadeh-Abadeh, S., Kuhn, D., and Esfahani, P. M.: Regularization via mass transportation. Journal of Machine Learning Research, 20(103):1–68, 2019.
- [Shalev-Shwartz and Ben-David, 2014]Shalev-Shwartz, S. and Ben-David, S.: Understanding machine learning: From theory to algorithms. Cambridge University Press, 2014.
- [Shalev-Shwartz et al. , 2010]Shalev-Shwartz, S., Shamir, O., Srebro, N., and Sridharan, K.: Learnability, stability and uniform convergence. The Journal of Machine Learning Research, 11:2635–2670, 2010.
- [Shapiro et al. , 2021]Shapiro, A., Dentcheva, D., and Ruszczyński, A.: Lectures on stochastic programming: modeling and theory. SIAM, 2021.
- [Shojaie and Michailidis, 2010]Shojaie, A. and Michailidis, G.: Penalized likelihood methods for estimation of sparse high-dimensional directed acyclic graphs. Biometrika, 97(3):519–538, 2010.
- [Si et al. , 2020]Si, N., Zhang, F., Zhou, Z., and Blanchet, J.: Distributionally robust policy evaluation and learning in offline contextual bandits. In International Conference on Machine Learning (ICML’20), 2020.
- [Silander and Myllymäki, 2006]Silander, T. and Myllymäki, P.: A simple approach for finding the globally optimal bayesian network structure. In Proceedings of the Twenty-Second Conference on Uncertainty in Artificial Intelligence, pages 445–452, 2006.
- [Singh and Póczos, 2018]Singh, S. and Póczos, B.: Minimax distribution estimation in wasserstein distance. arXiv preprint arXiv:1802.08855, 2018.
- [Sinha et al. , 2018]Sinha, A., Namkoong, H., and Duchi, J.: Certifiable distributional robustness with principled adversarial training. In International Conference on Learning Representations, 2018.
- [Sion, 1958]Sion, M.: On general minimax theorems. Pacific Journal of mathematics, 8(1):171–176, 1958.

- [Smith and Eisner, 2006]Smith, D. A. and Eisner, J.: Minimum risk annealing for training log-linear models. In Proceedings of the COLING/ACL 2006 Main Conference Poster Sessions, pages 787–794, 2006.
- [Smith and Smith, 2007]Smith, D. A. and Smith, N. A.: Probabilistic models of nonprojective dependency trees. In Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL), pages 132–140, 2007.
- [Smith and Winkler, 2006]Smith, J. E. and Winkler, R. L.: The optimizer’s curse: Skepticism and postdecision surprise in decision analysis. Management Science, 52(3):311–322, 2006.
- [Spirtes and Glymour, 1991]Spirtes, P. and Glymour, C.: An algorithm for fast recovery of sparse causal graphs. Social science computer review, 9(1):62–72, 1991.
- [Spirtes et al. , 2000]Spirtes, P., Glymour, C. N., Scheines, R., and Heckerman, D.: Causation, prediction, and search. MIT press, 2000.
- [Spirtes et al. , 1999]Spirtes, P., Meek, C., and Richardson, T.: An algorithm for causal inference in the presence of latent variables and selection bias. Computation, causation, and discovery, 21:211–252, 1999.
- [Stahlberg, 2020]Stahlberg, F.: Neural machine translation: A review. Journal of Artificial Intelligence Research, 69:343–418, 2020.
- [Staib and Jegelka, 2019]Staib, M. and Jegelka, S.: Distributionally robust optimization and generalization in kernel methods. Advances in Neural Information Processing Systems, 32:9134–9144, 2019.
- [Stanojević and Cohen, 2021]Stanojević, M. and Cohen, S. B.: A root of a problem: Optimizing single-root dependency parsing. In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pages 10540–10557, 2021.
- [Stoyanov and Eisner, 2012]Stoyanov, V. and Eisner, J.: Minimum-risk training of approximate CRF-based NLP systems. In Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 120–130, 2012.
- [Sugiyama et al. , 2010]Sugiyama, M., Takeuchi, I., Suzuki, T., Kanamori, T., Hachiya, H., and Okanohara, D.: Conditional density estimation via least-squares density ratio estima-

- tion. In Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics, pages 781–788, 2010.
- [Sun and Xu, 2016]Sun, H. and Xu, H.: Convergence analysis for distributionally robust optimization and equilibrium problems. Mathematics of Operations Research, 41(2):377–401, 2016.
- [Svyatkovskiy et al. , 2020]Svyatkovskiy, A., Deng, S. K., Fu, S., and Sundaresan, N.: Intelli-code compose: Code generation using transformer. In Proceedings of the 28th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering, pages 1433–1443, 2020.
- [Szeliski et al. , 2006]Szeliski, R., Zabih, R., Scharstein, D., Veksler, O., Kolmogorov, V., Agarwala, A., Tappen, M., and Rother, C.: A comparative study of energy minimization methods for Markov random fields. In European conference on computer vision, pages 16–29. Springer, 2006.
- [Taskar et al. , 2005]Taskar, B., Chatalbashev, V., Koller, D., and Guestrin, C.: Learning structured prediction models: A large margin approach. In Proceedings of the 22nd international conference on Machine learning, pages 896–903, 2005.
- [Taskar et al. , 2003]Taskar, B., Guestrin, C., and Koller, D.: Max-margin Markov networks. Advances in Neural Information Processing Systems, 16, 2003.
- [Taskar et al. , 2004]Taskar, B., Klein, D., Collins, M., Koller, D., and Manning, C. D.: Max-margin parsing. In Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing, pages 1–8, 2004.
- [Tewari and Bartlett, 2007]Tewari, A. and Bartlett, P. L.: On the consistency of multiclass classification methods. Journal of Machine Learning Research, 8(5), 2007.
- [Topsøe, 1979]Topsøe, F.: Information-theoretical optimization techniques. Kybernetika, 15(1):8–27, 1979.
- [Toutanova et al. , 2003]Toutanova, K., Klein, D., Manning, C. D., and Singer, Y.: Feature-rich part-of-speech tagging with a cyclic dependency network. In Proceedings of the 2003 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics, pages 252–259, 2003.

- [Trafalis and Gilbert, 2006]Trafalis, T. B. and Gilbert, R. C.: Robust classification and regression using support vector machines. European Journal of Operational Research, 173(3):893–909, 2006.
- [Tsamardinos et al. , 2006]Tsamardinos, I., Brown, L. E., and Aliferis, C. F.: The max-min hill-climbing bayesian network structure learning algorithm. Machine learning, 65(1):31–78, 2006.
- [Tsochantaridis et al. , 2005]Tsochantaridis, I., Joachims, T., Hofmann, T., Altun, Y., and Singer, Y.: Large margin methods for structured and interdependent output variables. Journal of Machine Learning Research, 6(9), 2005.
- [Villani and others, 2009]Villani, C. et al.: Optimal transport: old and new, volume 338. Springer, 2009.
- [Virtanen et al. , 2020]Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., et al.: Scipy 1.0: fundamental algorithms for scientific computing in python. Nature methods, 17(3):261–272, 2020.
- [von Neumann and Morgenstern, 1944]von Neumann, J. and Morgenstern, O.: Theory of games and economic behavior. Science and Society, 9(4), 1944.
- [Vuffray et al. , 2020]Vuffray, M., Misra, S., and Lokhov, A.: Efficient learning of discrete graphical models. In Advances in Neural Information Processing Systems, eds. H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, volume 33, pages 13575–13585. Curran Associates, Inc., 2020.
- [Vuffray et al. , 2016]Vuffray, M., Misra, S., Lokhov, A., and Chertkov, M.: Interaction screening: Efficient and sample-optimal learning of Ising models. In Advances in Neural Information Processing Systems, pages 2595–2603, 2016.
- [Wainwright, 2009]Wainwright, M. J.: Sharp thresholds for high-dimensional and noisy sparsity recovery using ℓ_1 -constrained quadratic programming (lasso). IEEE transactions on information theory, 55(5):2183–2202, 2009.
- [Wang and Chang, 2016]Wang, W. and Chang, B.: Graph-based dependency parsing with bidirectional LSTM. In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 2306–2315, 2016.

- [Wang et al. , 2021]Wang, X., Pun, Y.-M., and Man-Cho So, A.: Distributionally robust graph learning from smooth signals under moment uncertainty. arXiv e-prints, pages arXiv-2105, 2021.
- [Wang et al. , 2019]Wang, X., Zhang, H. H., and Wu, Y.: Multiclass probability estimation with support vector machines. Journal of Computational and Graphical Statistics, 28(3):586–595, 2019.
- [Wang et al. , 2016]Wang, Z., Glynn, P. W., and Ye, Y.: Likelihood robust optimization for data-driven problems. Computational Management Science, 13:241–261, 2016.
- [Weed et al. , 2019]Weed, J., Bach, F., et al.: Sharp asymptotic and finite-sample rates of convergence of empirical measures in wasserstein distance. Bernoulli, 25(4A):2620–2648, 2019.
- [Wei et al. , 2020]Wei, D., Gao, T., and Yu, Y.: Dags with no fears: A closer look at continuous optimization for learning bayesian networks. Advances in Neural Information Processing Systems, 33:3895–3906, 2020.
- [Werhli et al. , 2006]Werhli, A. V., Grzegorzcyk, M., and Husmeier, D.: Comparative evaluation of reverse engineering gene regulatory networks with relevance networks, graphical gaussian models and bayesian networks. Bioinformatics, 22(20):2523–2531, 2006.
- [Weston and Watkins, 1998]Weston, J. and Watkins, C.: Multi-class support vector machines. Technical report, Citeseer, 1998.
- [Wiesemann et al. , 2014]Wiesemann, W., Kuhn, D., and Sim, M.: Distributionally robust convex optimization. Operations Research, 62(6):1358–1376, 2014.
- [William, 1984]William, T.: Tutte. graph theory. Encyclopedia of Mathematics and its Applications, 21, 1984.
- [Williamson et al. , 2016]Williamson, R. C., Vernet, E., and Reid, M. D.: Composite multiclass losses. Journal of Machine Learning Research, 17:1–52, 2016.
- [Wozabal, 2014]Wozabal, D.: Robustifying convex risk measures for linear portfolios: A nonparametric approach. Operations Research, 62(6):1302–1315, 2014.
- [Wozabal and others, 2012]Wozabal, D. et al.: A framework for optimization under ambiguity. Annals of Operations Research, 193(1):21–47, 2012.

- [Wu et al. , 2019]Wu, S., Sanghavi, S., and Dimakis, A. G.: Sparse logistic regression learns all discrete pairwise graphical models. Advances in Neural Information Processing Systems, 32, 2019.
- [Xu et al. , 2016]Xu, J., Liu, J., Yin, J., and Sun, C.: A multi-label feature extraction algorithm via maximizing feature variance and feature-label dependence simultaneously. Knowledge-Based Systems, 98:172–184, 2016.
- [Xu et al. , 2017]Xu, Z., Taylor, G., Li, H., Figueiredo, M. A., Yuan, X., and Goldstein, T.: Adaptive consensus ADMM for distributed optimization. In International Conference on Machine Learning, pages 3841–3850. PMLR, 2017.
- [Xue et al. , 2002]Xue, N., Chiou, F.-D., and Palmer, M.: Building a large-scale annotated Chinese corpus. In COLING 2002: The 19th International Conference on Computational Linguistics, 2002.
- [Yang and Xu, 2013]Yang, W. and Xu, H.: A unified robust regression model for lasso-like algorithms. In International Conference on Machine Learning, pages 585–593. PMLR, 2013.
- [Yoshida et al. , 2021]Yoshida, S. M., Takenouchi, T., and Sugiyama, M.: Lower-bounded proper losses for weakly supervised classification. In International Conference on Machine Learning, pages 12110–12120. PMLR, 2021.
- [Yu et al. , 2020]Yu, H., Ye, W., Feng, Y., Bao, H., and Zhang, G.: Learning bipartite graph matching for robust visual localization. In 2020 IEEE International Symposium on Mixed and Augmented Reality (ISMAR), pages 146–155. IEEE, 2020.
- [Yu et al. , 2021a]Yu, Y., Lin, T., Mazumdar, E., and Jordan, M. I.: Fast distributionally robust learning with variance reduced min-max optimization. arXiv preprint arXiv:2104.13326, 2021.
- [Yu et al. , 2021b]Yu, Y., Gao, T., Yin, N., and Ji, Q.: Dags with no curl: An efficient dag structure learning approach. In International Conference on Machine Learning, pages 12156–12166. PMLR, 2021.
- [Zhang and Cheng, 2015]Zhang, H. and Cheng, L.: Restricted strong convexity and its applications to convergence analysis of gradient-type methods in convex optimization. Optimization Letters, 9(5):961–979, 2015.

- [Zhang et al. , 2021]Zhang, M., Lee, J., and Agarwal, S.: Learning from noisy labels with no change to the training process. In International Conference on Machine Learning, pages 12468–12478. PMLR, 2021.
- [Zhang et al. , 2017]Zhang, X., Cheng, J., and Lapata, M.: Dependency parsing as head selection. In Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers, pages 665–676, 2017.
- [Zhang et al. , 2010]Zhang, X., Saha, A., and Vishwanathan, S.: Regularized risk minimization by Nesterov’s accelerated gradient methods: Algorithmic extensions and empirical studies. arXiv preprint arXiv:1011.0472, 2010.
- [Zhang et al. , 2020]Zhang, Y., Li, Z., and Zhang, M.: Efficient second-order TreeCRF for neural dependency parsing. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 3295–3305, 2020.
- [Zhang et al. , 2014]Zhang, Y., Lei, T., Barzilay, R., and Jaakkola, T.: Greed is good if randomized: New inference for dependency parsing. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 1013–1024, 2014.
- [Zhao and Guan, 2018]Zhao, C. and Guan, Y.: Data-driven risk-averse stochastic optimization with wasserstein metric. Operations Research Letters, 46(2):262–267, 2018.
- [Zheng et al. , 2018]Zheng, X., Aragam, B., Ravikumar, P. K., and Xing, E. P.: Dags with no tears: Continuous optimization for structure learning. Advances in Neural Information Processing Systems, 31, 2018.
- [Zmigrod et al. , 2020]Zmigrod, R., Vieira, T., and Cotterell, R.: Please mind the root: Decoding arborescences for dependency parsing. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 4809–4819, 2020.
- [Zmigrod et al. , 2021]Zmigrod, R., Vieira, T., and Cotterell, R.: Efficient computation of expectations under spanning tree distributions. Transactions of the Association for Computational Linguistics, 9:675–690, 2021.
- [Zou et al. , 2008]Zou, H., Zhu, J., and Hastie, T.: New multiclass boosting algorithms based on multiclass fisher-consistent losses. The Annals of Applied Statistics, 2(4):1290, 2008.
- [Zymler et al. , 2013]Zymler, S., Kuhn, D., and Rustem, B.: Distributionally robust joint chance constraints with second-order moment information. Mathematical Programming, 137, 2013.

APPENDIX

COPYRIGHT POLICIES

A.1 Copyright Policy of Neural Information Processing Systems

According to U.S. Copyright Office's page What is a Copyright¹. When you create an original work you are the author and the owner and hold the copyright, unless you have an agreement to transfer the copyright to a third party such as the company or school you work for.

Authors do not transfer the copyright of their paper to NeurIPS, instead they grant NeurIPS a non-exclusive, perpetual, royalty-free, fully-paid, fully-assignable license to copy, distribute and publicly display all or part of the paper.

A.2 Copyright Policy of Artificial Intelligence and Statistics

The International Conference on Artificial Intelligence and Statistics' (AISTATS) proceeding is published by the Proceedings of Machine Learning Research (PMLR).

The Proceedings of Machine Learning Research (formerly JMLR Workshop and Conference Proceedings) is a series aimed specifically at publishing machine learning research presented at workshops and conferences. Each volume is separately titled and associated with a particular workshop or conference and will be published online on the PMLR website. Authors retain

¹<https://www.copyright.gov/what-is-copyright/>

APPENDIX (Continued)

ownership of all rights under copyright in all versions of the article, and all rights not expressly granted in this agreement.

A.3 Copyright Policy of International Conference on Learning Representations

Our paper (Li and Ziebart, 2023) was rejected and made public by the International Conference on Learning Representations (ICLR) on OpenReview.

According to U.S. Copyright Office’s page What is a Copyright¹. When you create an original work you are the author and the owner and hold the copyright, unless you have an agreement to transfered the copyright to a third party such as the company or school you work for.

Authors do not tranfer the copyright of their paper to ICLR, instead they grant ICLR a non-exclusive, perpetual, royalty-free, fully-paid, fully-assignable license to copy, distribute and publicly display all or part of the paper.

¹<https://www.copyright.gov/what-is-copyright/>

VITA

NAME	Yeshu Li
EDUCATION	Ph.D., Computer Science, University of Illinois at Chicago, 2023 (expected) M.S., Computer Science, University of Illinois at Chicago, 2022 B.E., Computer Science and Engineering, Beihang University
CAREER	Research Assistant, University of Illinois at Chicago, 2019-2022 Research Intern, Microsoft Research Asia, 2019
PUBLICATIONS	Yeshu Li , Brian D. Ziebart. “Distributionally Robust Skeleton Learning of Discrete Bayesian Networks”. <i>Work in progress</i> Yeshu Li , Brian D. Ziebart. “Moment Distributionally Robust Probabilistic Supervised Learning”. In <i>OpenReview preprint</i> Yeshu Li , Danyal Saeed, Xinhua Zhang, Brian D. Ziebart, Kevin Gimpel. “Moment Distributionally Robust Tree Structured Prediction”. In <i>Proceedings of Neural Information Processing Systems (NeurIPS)</i> , 2022 Yeshu Li , Zhan Shi, Xinhua Zhang, Brian D. Ziebart. “Distributionally Robust Structure Learning for Discrete Pairwise Markov Networks”. In <i>Proceedings of International Conference on Artificial Intelligence and Statistics (AISTATS)</i> , 2022 Yeshu Li , Ziming Qiu, Xingyu Fan, Xianglong Liu, Eric I-Chao Chang, Yan Xu. “Integrated 3D Flow-based Multi-atlas Brain Structure Segmentation”. In <i>PloS one</i> , 2022 Yeshu Li , Jonathan Cui, Yilun Sheng, Xiao Liang, Jingdong Wang, Eric I-Chao Chang, Yan Xu. “Whole Brain Segmentation with Full Volume Neural Network”. In <i>Computerized Medical Imaging and Graphics (CMIG)</i> , 2021

Yan Xu, **Yeshu Li**, Zhengyang Shen, Ziwei Wu, Teng Gao, Yubo Fan, Maode Lai, Eric I-Chao Chang. “Parallel Multiple Instance Learning for Extremely Large Histopathology Image Analysis”. In *BMC Bioinformatics*, 2017