Distributionally Robust Structure Learning for Discrete Pairwise Markov Networks

Yeshu Li **Zhan Shi** **Xinhua Zhang** **Brian D. Ziebart**
Department of Computer Science
University of Illinois at Chicago

Abstract

We consider the problem of learning the underlying structure of a general discrete pairwise Markov network. Existing approaches that rely on empirical risk minimization 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. Our approach builds upon distributionally robust optimization (DRO) and maximum conditional log-likelihood. The proposed DRO estimator minimizes the worst-case risk over an ambiguity set of adversarial distributions within bounded transport cost or f-divergence of the empirical data distribution. We show that the primal minimax learning problem 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. Extensive empirical evidence with different corruption models corroborates the effectiveness of the proposed methods.

1 INTRODUCTION

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 empirical risk minimization (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). 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.

To alleviate the above issues, we put forward a distributionally robust optimization (DRO) 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

Proceedings of the 25$^{\text{th}}$ International Conference on Artificial Intelligence and Statistics (AISTATS) 2022, Valencia, Spain. PMLR: Volume 151.

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 Kullback–Leibler (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.

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., 2011). 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.

DRO approaches have been adopted to address many statistical learning problems such as multivariate convex regression (Blanchet et al., 2019), submodular maximization (Staib et al., 2019) and more (Abadeh et al., 2015; Nguyen et al., 2018; Si et al., 2020; Esfahani and Kuhn, 2018; Farnia and Tse, 2016; Lee and Raginsky, 2018; Nguyen et al., 2020a,b).

2 PRELIMINARIES

2.1 Notations

Throughout this manuscript, we denote by $[n]$ the set $\{1, 2, \ldots, n\}$ for $n \in \mathbb{Z}_+$. For a vector $x \in \mathbb{R}^n$, we use x_i for its i-th coordinate, and let $x_{-i} \in \mathbb{R}^{n-1}$ represent $(x_j : j \neq i)$, $x_S \in \mathbb{R}^{|S|}$ represent $(x_i : i \in S)$, and $x_{i=c}$ represent $[x_1, \ldots, x_{i-1}, c, x_{i+1}, \ldots, x_n]^\intercal$ for some $c \in \mathbb{R}$. We follow the column vector convention and write $x^\intercal \in \mathbb{R}^{1\times n}$ as a row vector which is the transpose of the column vector $x \in \mathbb{R}^{n\times 1}$. The ℓ_p-norm of a vector is indicated by $\|\cdot\|_p \coloneqq (\sum_i |x_i|^p)^{1/p}$ with $|\cdot|$

being the modulus. For a matrix $A \in \mathbb{R}^{n\times m}$, we use $A_{i(,)j}$, $A_{i(,)*}$ and $A_{*(,)j}$ to denote its (i,j)-th entry, i-th row and j-th column respectively, where comma is sometimes omitted. The $\ell_{p,q}$ norm of A is computed as $\|A\|_{p,q} \coloneqq \|(\|A_{i*}\|_p : i \in [n])\|_q$. For two matrices $A, B \in \mathbb{R}^{n\times m}$, the inner product of them is designated by $\langle A, B\rangle \triangleq \mathrm{Tr}[A^\intercal B]$, while the Hadamard product is written as $A \odot B$ for element-wise multiplication. By a slight abuse of notation, we use $|A|$ or $\#A$ to indicate the cardinality of a set A. We denote by $\mathbb{T}(x) \in \mathbb{R}^n$ a vector with non-decreasing components as a result of sorting $(x_i : i \in [n])$. We represent a vector with all ones (zeroes) as $\mathbf{1}$ ($\mathbf{0}$). Given a distribution $\mathbb{P}$ on a set Ξ, we denote by $\mathbb{P}^m$ the m-fold product of $\mathbb{P}$ on the Cartesian product Ξ^m. We use $\mathbb{E}^{\mathbb{P}}$ to represent the expectation under $\mathbb{P}$. The i-th standard basis vector is written as $v^{(i)}$ with $v^{(i)}_i = 1$ and $v^{(i)}_j = 0$ for $j \neq i$. Denote $V \coloneqq \{v^{(i)} : i \in [k]\}$ as the set of basis vectors in $\mathbb{R}^k$ and $V^{(n\times k)} \subset \{0,1\}^{n\times k}$ as the set of all $n \times k$ matrices whose rows are k-dimensional standard basis vectors. The least d-Lipschitz constant of a function $f : \Xi \to \mathbb{R}$ is written as $\mathrm{lip}_d(f)$ with a metric d.

2.2 Learning Pairwise Markov Networks over General Alphabet

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

Definition 1. *Let k be the alphabet size. Let $\mathcal{W} = \{W^{(ij)} \in \mathbb{R}^{k\times k} : i \neq j \in [n]\}$ be a collection of symmetric weight matrices and $\Theta = \{\theta^{(i)} \in \mathbb{R}^k : i \in [n]\}$ 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 $\mathcal{E} = \{\{i,j\} \subseteq \mathcal{V} : W^{(ij)} \neq \mathbf{0}\}$. 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*

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

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

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

Assumption 1. *$W^{(ij)}$ has centered rows and columns: $\sum_{a\in[k]} W^{(ij)}_{ab} = \sum_{b\in[k]} W^{(ij)}_{ab} = 0$.*

Assumption 2. *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 1 is made without loss of generality because centering $(\mathcal{W}, \Theta)$ leads to $(\mathcal{W}', \Theta')$ with the same distribution: $\mathcal{D}(\mathcal{W}, \Theta) = \mathcal{D}(\mathcal{W}', \Theta')$. One of the useful properties induced by Assumption 2 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 $\sigma(x) \coloneqq (1+e^{-x})^{-1}$ if the dependent variable is restricted to a pair of values.

Fact 1. *Let $Z \sim \mathcal{D}(\mathcal{W}, \Theta)$ be a discrete pairwise graphical model over $[k]^n$. For any $i \in [n]$ and $\alpha \neq \beta \in [k]$, we have*

$$\begin{aligned}&\mathbb{P}[Z_i = \alpha | Z_i \in \{\alpha, \beta\}, Z_{-i} = z_{-i}]\\ &= \sigma\Big(\sum_{j\neq i}(W^{(ij)}_{\alpha z_j} - W^{(ij)}_{\beta z_j}) + \theta^{(i)}_\alpha - \theta^{(i)}_\beta\Big) \triangleq \sigma(\langle \bar{W}, \bar{Z}\rangle),\end{aligned}$$

where $\bar{W} \in \mathbb{R}^{n\times k}$ is defined as $\bar{W}_{i*} \coloneqq [\theta^{(i)}_\alpha - \theta^{(i)}_\beta, \mathbf{0}^\intercal]$, and $\bar{W}_{j*} \coloneqq W^{(ij)}_{\alpha*} - W^{(ij)}_{\beta*}$ for $j \neq i \in [n]$. $\bar{Z} \coloneqq \mathrm{OneHot}(z_{i=1}) \in V^{(n\times k)}$ encodes $z_{i=1}$ such that $\bar{Z}_{i*} = v^{(1)\intercal}$ and $\bar{Z}_{j*} = v^{(z_j)\intercal}$ for $j \neq i$.

The definition of $\bar{W}$ implies $\|\bar{W}\|_{2,1} \leq 2\lambda\sqrt{k}$. Let $\{\bar{z}^{(1)}, \ldots, \bar{z}^{(m)}\} \overset{iid}{\sim} \mathcal{D}(\mathcal{W}, \Theta)$ be a set of m samples and $\{z^{(1)}, \ldots, z^{(m')}\}$ be its subset with $z^{(j)}_i \in \{\alpha, \beta\}$. Define $y^{(j)} = 1$ if $z^{(j)}_i = \alpha$ and $y^{(j)} = -1$ if $z^{(j)}_i = \beta$. 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:

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

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

$$\begin{aligned}U^{(i\alpha\beta)}_{i*} &\coloneqq \hat{W}^{(i\alpha\beta)}_{i*} + \frac{1}{k}\sum_{j\neq i\in[n],a\in[k]} \hat{W}^{(i\alpha\beta)}_{ja}\mathbf{1}^\intercal \qquad (2)\\ U^{(i\alpha\beta)}_{j*} &\coloneqq \hat{W}^{(i\alpha\beta)}_{j*} - \frac{1}{k}\sum_{a\in[k]} \hat{W}^{(i\alpha\beta)}_{ja}\mathbf{1}^\intercal \quad \forall j \neq i,\end{aligned}$$

yields a minimizer of Eq. (1) due to $\langle \hat{W}^{(i\alpha\beta)}, \bar{Z}\rangle = \langle U^{(i\alpha\beta)}, \bar{Z}\rangle$.

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

$$\hat{W}^{(ij)}_{\alpha*} \coloneqq \frac{1}{k}\sum_{\beta\in[k]} U^{(\alpha\beta)}_{j*} \quad \forall j \neq i \in [n], \alpha \in [k]. \tag{3}$$

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

$$\hat{\mathcal{E}} \coloneqq \{\{i,j\} : \|\hat{W}^{(ij)}\|_\infty \geq \eta/2, i < j \in [n]\}. \tag{4}$$

2.3 Distributionally Robust Learning

In classical statistical learning, 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 $f : \mathcal{Z} \to \mathbb{R}_+$, sometimes considered as the hypothesis space, where each $f \in \mathcal{F}$ assigns a scalar cost value to each instance $\xi \in \Xi$. Given a distribution $\mathbb{P} \in \mathcal{P}(\Xi)$, the goal is to infer a hypothesis f^* whose risk is minimum or nearly optimal with high confidence, which is equivalent to a stochastic optimization problem:

$$\inf_{f\in\mathcal{F}} \int_\Xi f(\xi)\mathbb{P}(\mathrm{d}\xi). \tag{5}$$

In practical terms, only a finite set of in-sample data $\{\xi^{(1)}, \ldots, \xi^{(m)}\}$ drawn i.i.d. from the unknown $\mathbb{P}$ is accessible. On account of this, regularized ERM is usually adopted to find a hypothesis function $\hat{f}$ that faithfully minimizes an approximate expected risk:

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

where $\hat{\mathbb{P}}_m = \frac{1}{m}\sum_{i=1}^m \delta_{\xi^{(i)}}$ with $\delta_{\xi^{(i)}}$ being the Dirac point measure at $\xi^{(i)}$, $\Omega(\cdot)$ represents a function quantifying hypothesis complexities and λ' is a trade-off coefficient to combat overfitting.

Distributionally robust optimization provides an alternative perspective to ERM. Because of limited information about the true data-generating distribution, the DRO framework explicitly models uncertainty by constructing an ambiguity set that contains the unknown distribution with high confidence, based on a nominal distribution. DRO seeks to minimize the worst-case expected risk instead of the empirical risk:

$$\inf_{f\in\mathcal{F}} \sup_{\mathbb{Q}\in\mathcal{A}} \int_\Xi f(\xi)\mathbb{Q}(\mathrm{d}\xi), \tag{6}$$

where $\mathcal{A} \subseteq \mathcal{P}(\Xi)$ is an ambiguity set. Intuitively, an appropriate ambiguity set is expected to yield an efficiently solvable optimization problem, a near-optimal and asymptotically consistent estimator.

The ambiguity set $\mathcal{A}$ is typically taken as a ball of radius ε centered at a nominal measure: $B_\varepsilon(\mathbb{P}) \coloneqq \{\mathbb{Q} \in \mathcal{P}(\Xi) : \mathrm{div}(\mathbb{Q}, \mathbb{P}) \leq \varepsilon\}$, where $\mathrm{div}(\cdot,\cdot)$ measures the discrepancy between two distributions. We consider two popular choices of $\mathrm{div}(\cdot,\cdot)$, based on the Wasserstein metric and relative entropy.

Definition 2. *Assume that Ξ is a Polish space equipped with a metric $d : \Xi \times \Xi \to \mathbb{R}_+$. 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 $M(\Xi^2)$ 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*

$$W_p(\mathbb{P},\mathbb{Q}) \coloneqq \inf_{\Pi\in M(\Xi^2)} \Bigg\{ \Big[\int_{\Xi^2} d^p(\xi,\xi')\Pi(\mathrm{d}\xi,\mathrm{d}\xi')\Big]^{\frac{1}{p}} : \Pi(\mathrm{d}\xi,\Xi) = \mathbb{P}(\mathrm{d}\xi), \Pi(\Xi,\mathrm{d}\xi') = \mathbb{Q}(\mathrm{d}\xi') \Bigg\}.$$

Wasserstein distances arise in the problem of optimal transport and can be interpreted as the minimum cost of moving the probability measure $\mathbb{P}$ to $\mathbb{Q}$ with unit transport costs specified by $d(\xi,\xi')$.

Definition 3. *Let $\mathbb{Q} \in \mathcal{P}(\Xi)$ be absolutely continuous with respect to $\mathbb{P} \in \mathcal{P}(\Xi)$. Let $\frac{\mathrm{d}\mathbb{Q}(\xi)}{\mathrm{d}\mathbb{P}(\xi)}$ be the Radon-Nikodym derivative. The Kullback-Leibler divergence from $\mathbb{Q}$ to $\mathbb{P}$ is defined as*

$$D_{\mathrm{KL}}(\mathbb{Q} \,\|\, \mathbb{P}) \triangleq \int_\Xi \ln \frac{\mathrm{d}\mathbb{Q}(\xi)}{\mathrm{d}\mathbb{P}(\xi)} \mathrm{d}\mathbb{Q}(\xi).$$

The relative entropy, or KL divergence, arises in information theory and is a well-known asymmetric measure of discrepancy between distributions.

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. We defer all proofs to the supplementary material.

3.1 Learning Discrete Pairwise Markov Networks with DRO

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 4. *Let $\Xi = \mathcal{X} \times \mathcal{Y} = V^{((n-1)\times k)} \times \{-1, 1\}$. Given m samples $\{\bar{z}^{(1)}, \ldots, \bar{z}^{(m)}\} \overset{iid}{\sim} \mathcal{D}(\mathcal{W}, \Theta)$, the goal of learning discrete pairwise MRFs with distributionally robust logistic regression is to find the optimal $\hat{W}^{(i\alpha\beta)}$ for each $i \in [n]$ and $\alpha \neq \beta \in [k]$ via minimax statistical learning, formally,*

$$\hat{W}^{(i\alpha\beta)} \in \arg \inf_{W \in \mathbb{R}^{n\times k}} \sup_{\mathbb{Q} \in B_\varepsilon(\hat{\mathbb{P}}_{m'})} \int_\Xi \ell(y\langle W, X\rangle) \mathbb{Q}(\mathrm{d}(x, y)), \tag{7}$$

where $X \coloneqq [x^\intercal_{1\ldots i-1,*}, v^{(1)\intercal}, x^\intercal_{i\ldots n-1,*}]^\intercal$ inserts the first standard basis vector into the i-th row of x. $\hat{\mathbb{P}}_{m'}$ is the empirical distribution for a set of transformed m' samples $\{\xi^{(1)}, \ldots, \xi^{(m')}\}$ such that, for any $\xi^{(j')} = (x^{(j')}, y^{(j')}) \in \Xi, j' \in [m']$ and its corresponding original sample $\bar{z}^j, j \in [m]$, 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 $x^{(j')} = \mathrm{OneHot}(\bar{z}_{-i}^{(j)})$.

Note that if ε is set to zero, Eq. (7) reduces to an unconstrained version of Eq. (1). More importantly, the DRO formulation in Eq. (7) is an infinite-dimensional optimization problem, which is generally impossible to solve directly.

3.2 Tractable Reformulations

We show that the DRO problem in Definition 4 can be solved efficiently via its dual formulations. The following theorem presents a tractable convex reformulation for the primal problem in Eq. (7) if a Wasserstein ball is adopted as the ambiguity set.

Theorem 1. *Let $W_1(\cdot, \cdot)$ be the type-1 Wasserstein distance with $p = 1$ and metric $d(\xi, \xi') \triangleq d((x, y), (x', y')) \coloneqq \|x - x'\|_{1,1} + \frac{\kappa}{2}|y - y'|$ for $\xi, \xi' \in \Xi$, $\kappa \in \mathbb{R}_+$. Let $B_\varepsilon^{W_1}(\hat{\mathbb{P}}_{m'}) \coloneqq \{\mathbb{Q} \in \mathcal{P}_1(\Xi) : W_1(\hat{\mathbb{P}}_{m'}, \mathbb{Q}) \leq \varepsilon = \frac{\varepsilon_0}{\sqrt{m'}}\}$ be the ambiguity set. Then the primal problem in Eq. (7) is equivalent to*

$$\inf_{\substack{W\in\mathbb{R}^{n\times k},\\ \gamma\geq 0}} \gamma\varepsilon + \frac{1}{m'}\sum_{j=1}^{m'} \sup_{\substack{0\leq r\leq n-1\in\mathbb{Z},\\ g\in\{-1,1\}}} [-\frac{1}{2}\gamma\kappa(1 + gy^{(j)}) - 2r\gamma + \ln(1 + e^{\langle gW, X^{(j)}\rangle + \langle \mathbb{T}(\delta)_{1\ldots r}, \mathbf{1}\rangle})], \tag{8}$$

where $X^{(j)} \coloneqq [x^{(j)\intercal}_{1\ldots i-1,*}, v^{(1)\intercal}, x^{(j)\intercal}_{i\ldots n-1,*}]^\intercal$, $\delta \coloneqq [\sup_{l\in[k]}(gW)_{jl} : j \neq i \in [n]]^\intercal - (gW_{-i,*} \odot x^{(j)})\mathbf{1}$, and $\mathbb{T}(x)$ is defined as a vector with non-decreasing components as a result of sorting x, introduced in Section 2.1.

We give a proof sketch here. After decomposing the Wasserstein constraint based on the label domain, the Lagrangian dual problem of Eq. (7) includes an inner maximization over a set of $k^{\Theta(n)}$ discrete values: $\sup_{g\in\{-1,1\}, x\in V^{((n-1)\times k)}} \ln(1 + e^{g\langle W, X\rangle}) - \gamma\|x^{(j)} - x\|_1 - \frac{1}{2}\gamma\kappa(1 + gy^{(j)})$. This cannot be further simplified via the Fenchel conjugate trick in Esfahani and Kuhn (2018) due to the non-convexity of our distribution support $V^{((n-1)\times k)}$. Instead, by noting the monotonicity of $e^{\langle gW, X^{(j)}\rangle + \langle \mathbb{T}(\delta)_{1\ldots r}, \mathbf{1}\rangle}$ by fixing $\|x^{(j)} - x\|_1 = 2r$ satisfied by some x's, we obtain the above convex objective (8) with a sum of m' point-wise maximum functions of $\Theta(n)$ convex functions of the primal variable W and the dual variable γ. The sorting operation required to evaluate $\mathbb{T}(\cdot)$ can be accomplished in $\Theta(n \log n)$ for sub-derivative evaluation.

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 2. *If $\kappa = \infty$, $\|W\|_{2,1} \leq 2\lambda\sqrt{k}$ and $\gamma \geq (n + 2)\lambda\sqrt{k}$, the convex program in (8) subsumes the standard $\ell_{2,1}$-constrained logistic regression approach in (1) as a special case. If $\kappa = \infty$ and $\gamma \geq \frac{n+2}{2}\|W\|_{2,1}$, it subsumes the $\ell_{2,1}$-regularized logistic regression approach as a special case.*

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 $x^{(j)}$ is deterministic. Instead of taking into account the ambiguity only in the covariate measure $\mathbb{Q}(\mathrm{d}x)$, the Wasserstein DRO structure learning formulation grants flexibility to the joint measure $\mathbb{Q}(\mathrm{d}\xi)$. 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 Eq. (7).

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 1. *Let $B_\varepsilon^{\mathrm{KL}}(\hat{\mathbb{P}}_{m'}) \coloneqq \{\mathbb{Q} \in \mathcal{P}(\Xi) : D_{\mathrm{KL}}(\mathbb{Q}, \hat{\mathbb{P}}_{m'}) \leq \varepsilon = \frac{\varepsilon_0}{m'}\}$ be a KL divergence ball. The primal problem in Eq. (7) with $B_\varepsilon(\hat{\mathbb{P}}_{m'}) = B_\varepsilon^{\mathrm{KL}}(\hat{\mathbb{P}}_{m'})$ is equivalent to*

$$\inf_{\substack{W\in\mathbb{R}^{n\times k},\\ \gamma\geq 0}} \gamma \ln[\frac{1}{m'}\sum_{j=1}^{m'}(1 + e^{-y^{(j)}\langle W, X^{(j)}\rangle})^{\frac{1}{\gamma}}] + \gamma\varepsilon. \tag{9}$$

In contrast to the convex program with inner maximization in Eq. (8), the direct minimization problem in Eq. (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).

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

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

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

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 1. *Assume that (Ξ, d) 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 $\mathbb{P} \in \mathcal{P}_p(\Xi)$, $\mathcal{A} = B_\varepsilon^{W_p}(\hat{\mathbb{P}}_m)$ is a type-p Wasserstein ball centered at $\hat{\mathbb{P}}_m$ with radius ε, $\mathcal{F}$ is a space of closed convex functions $f : \Xi \to \mathbb{R}_+$ with $\mathrm{lip}_d(f) < \infty$. Let $\hat{f}$ be a minimizer of the DRO problem in Eq. (6) and f^* be a minimizer of the stochastic optimization problem in Eq. (5), we have*

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

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 $\|W_{\alpha*}^{(ij)} - W_{\beta*}^{(ij)} - U_{j*}^{(i\alpha\beta)}\|_1$ based on the excess risk bound in Proposition 1.

Theorem 3. *Given that: $\mathcal{D}(\mathcal{W}, \Theta)$ is an unknown pairwise Markov network with n variables, alphabet size k, dependency graph $\mathcal{G}$; that Assumptions 1 and 2 hold; that $\|W\|_{2,1} \leq 2\lambda\sqrt{k}$ in Eq. (7); that $W^{(ij)} \in \mathcal{W}$ is the true weight matrix; and that $\hat{W}^{(ij)}$ is the estimated weight matrix from Eq. (8) with the Wasserstein ambiguity set and properly centered (Section 2.2), 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:*

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

Let $\omega < \frac{\eta}{2}$ and $\hat{\mathcal{G}}$ be reconstructed via thresholding in Eq. (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 $\mathcal{G} = \hat{\mathcal{G}}$.

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., 2020).

Theorem 4. *Given assumptions in Theorem 3, except that $\hat{W}^{(ij)}$ is the estimated weight matrix from Eq. (9) with the KL ambiguity set. Let $\hat{\mathcal{G}}$ be constructed via thresholding in Eq. (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 $m = O(\frac{\lambda^2 k^4 e^{14\lambda}(\varepsilon_0 + \ln\frac{nk}{\rho})}{\eta^4})$, with probability at least $1-\rho$, the following bound holds:*

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

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[1].

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, 2016). 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.

5 EXPERIMENTS

We conduct a simulated study of synthetic data[2] perturbed by the following contamination models:

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

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 $z \sim \mathcal{D}$ is randomly replaced with a different value with probability ζ.

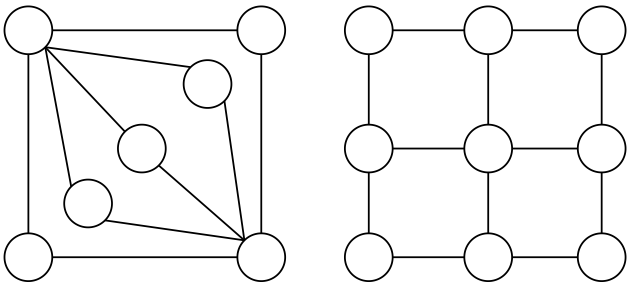

Figure 1: The adopted underlying graphs. Two nodes are connected to the others in the diamond graph. The grid graph has d^2 nodes. Each edge weight matrix is centered with random values $\pm\theta$.

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[3]. 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[4]. 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

[1]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.

[2]Our code and data generator are publicly available at `https://github.com/DanielLeee/drslmarkov`.

[3]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.

[4]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.

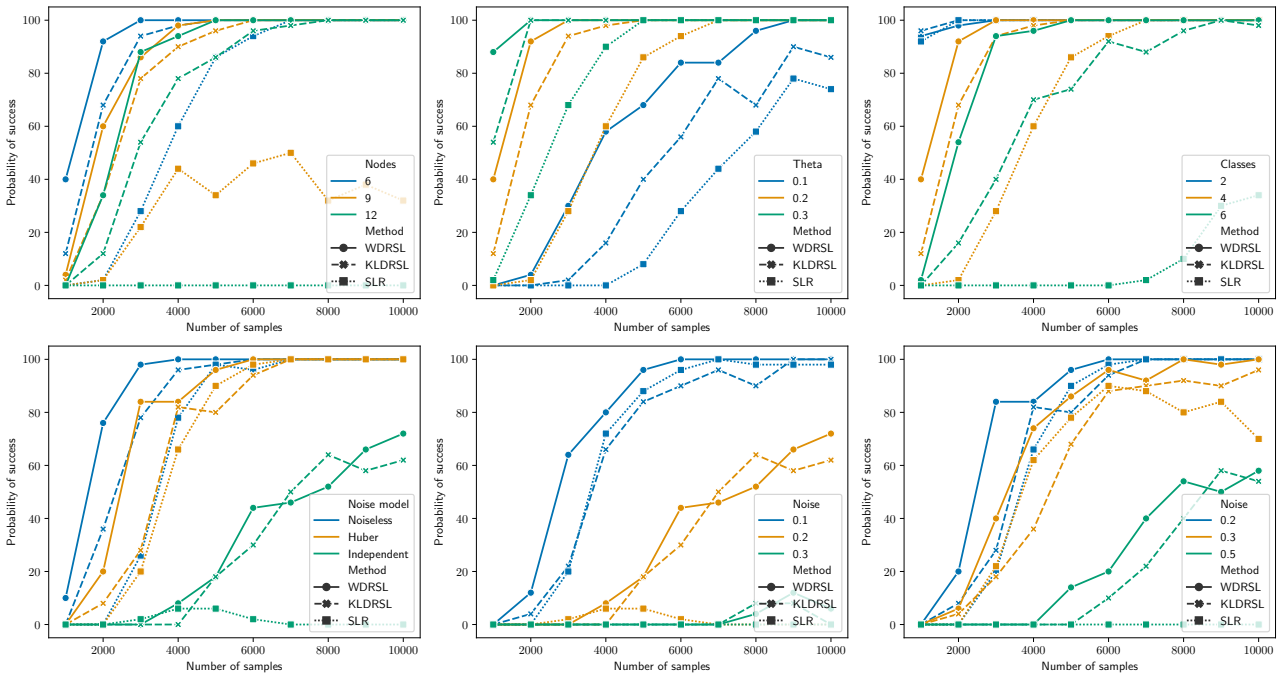

Figure 2: Plots of the probability of successfully estimating the structure versus the number of samples for Wasserstein DRO structure learning (WDRSL), KL DRO (KLDRSL) and sparse logistic regression (SLR). Top, from left to right: (a) diamond, 4 classes, noiseless, $\theta = 0.2$, varying nodes; (b) 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.

graphs of same size as the target graph. The chosen hyperparameters can be found in Appendix B. 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) 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

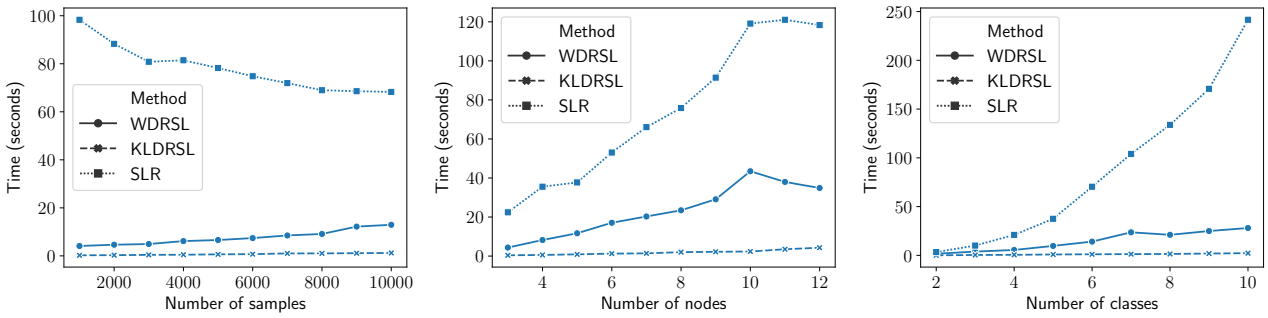

Figure 3: Comparisons of the execution time of one run. $\theta = 0.2$ and noiseless model are adopted in all settings. $\kappa = 1$, $\varepsilon_0 = 33$ for KLDRSL and $\varepsilon_0 = 1.5$ for WDRSL. From left to right: (a) grid, 9 nodes, 4 classes, varying samples; (b) diamond, 4 classes, varying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

one according to Figure 3, whereas WDRSL provides a trade-off between computational efficiency and structure learning ability.

6 DISCUSSION AND CONCLUSION

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

The per-iteration costs $\widetilde{O}(nk + n \log n)$ and $\widetilde{O}(nk)$ in terms of n and k to optimize our objectives may not be further improved. However, faster overall convergence rates (e.g., better than $\widetilde{O}(n^2k^2)$) are possible if we replace L-BFGS-B with advanced optimization methods designed for DRO (Yu et al., 2021; Namkoong and Duchi, 2016). Although robust to a set of adversarial distributions, our estimators may not be superior to robust estimators tailored to a certain contamination model if they can be generalized to non-binary MRFs. Despite absolute continuity, KL divergence usually allows a DRO problem to have a simple dual problem and good statistical guarantees, which was shown in this work. It would be interesting to extend our approaches to learning continuous or higher-order MRFs.

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.

Acknowledgements

The authors would like to thank Shanshan Wu for the helpful comments on experimental design. This material is based upon work supported by the National Science Foundation under Grant No. 1652530.

References

Abadeh, S. S., Esfahani, P. M. M., and Kuhn, D. (2015). Distributionally robust logistic regression. In *Advances in Neural Information Processing Systems*, pages 1576–1584.

Bartlett, P. L. and Mendelson, S. (2002). Rademacher and Gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research*, 3(Nov):463–482.

Blanchet, J., Glynn, P. W., Yan, J., and Zhou, Z. (2019). Multivariate distributionally robust convex regression under absolute error loss. In *Advances in Neural Information Processing Systems*, pages 11794–11803.

Bresler, G. (2015). Efficiently learning Ising models on arbitrary graphs. In *Proceedings of the forty-seventh annual ACM symposium on Theory of computing*, pages 771–782.

Bresler, G., Mossel, E., and Sly, A. (2013). Reconstruction of Markov random fields from samples: Some observations and algorithms. *SIAM Journal on Computing*, 42(2):563–578.

Byrd, R. H., Lu, P., Nocedal, J., and Zhu, C. (1995). A limited memory algorithm for bound constrained optimization. *SIAM Journal on scientific computing*, 16(5):1190–1208.

Chayes, J., Chayes, L., and Lieb, E. H. (1984). The inverse problem in classical statistical mechanics. *Communications in Mathematical Physics*, 93(1):57–121.

Chen, Y. (2010). Learning sparse Ising models with missing data.

Chow, C. and Liu, C. (1968). Approximating discrete probability distributions with dependence trees. *IEEE transactions on Information Theory*, 14(3):462–467.

Cranko, Z., Shi, Z., Zhang, X., Nock, R., and Kornblith, S. (2020). Generalised lipschitz regularisation equals distributional robustness. *arXiv preprint arXiv:2002.04197*.

Diakonikolas, I., Kane, D. M., Stewart, A., and Sun, Y. (2021). Outlier-robust learning of ising models under dobrushin's condition. In Belkin, M. and Kpotufe, S., editors, *Proceedings of Thirty Fourth Conference on Learning Theory*, volume 134 of *Proceedings of Machine Learning Research*, pages 1645–1682. PMLR.

Duchi, J. and Namkoong, H. (2019). Variance-based regularization with convex objectives. *The Journal of Machine Learning Research*, 20(1):2450–2504.

Eagle, N., Pentland, A. S., and Lazer, D. (2009). Inferring friendship network structure by using mobile phone data. *Proceedings of the national academy of sciences*, 106(36):15274–15278.

Esfahani, P. M. and Kuhn, D. (2018). Data-driven distributionally robust optimization using the Wasserstein metric: Performance guarantees and tractable reformulations. *Mathematical Programming*, 171(1-2):115–166.

Farnia, F. and Tse, D. (2016). A minimax approach to supervised learning. In Lee, D., Sugiyama, M., Luxburg, U., Guyon, I., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc.

Gao, R. and Kleywegt, A. J. (2016). Distributionally robust stochastic optimization with Wasserstein distance. *arXiv preprint arXiv:1604.02199*.

Goel, S., Kane, D. M., and Klivans, A. R. (2019). Learning Ising models with independent failures. In *Conference on Learning Theory*, pages 1449–1469.

Hamilton, L., Koehler, F., and Moitra, A. (2017). Information theoretic properties of Markov random fields, and their algorithmic applications. In *Advances in Neural Information Processing Systems*, pages 2463–2472.

Hu, Z. and Hong, L. J. (2013). Kullback-leibler divergence constrained distributionally robust optimization. *Available at Optimization Online*.

Jalali, A., Ravikumar, P., Vasuki, V., and Sanghavi, S. (2011). 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.

Katiyar, A., Basu, S., Shah, V., and Caramanis, C. (2021). Robust estimation of tree structured markov random fields. *arXiv preprint arXiv:2102.08554*.

Katiyar, A., Shah, V., and Caramanis, C. (2020). Robust estimation of tree structured ising models. *arXiv preprint arXiv:2006.05601*.

Klivans, A. and Meka, R. (2017). Learning graphical models using multiplicative weights. In *2017 IEEE 58th Annual Symposium on Foundations of Computer Science*, pages 343–354. IEEE.

Lam, H. (2019). Recovering best statistical guarantees via the empirical divergence-based distributionally robust optimization. *Operations Research*, 67(4):1090–1105.

Lee, J. and Raginsky, M. (2018). Minimax statistical learning with wasserstein distances. *Advances in Neural Information Processing Systems*, 2018:2687–2696.

Lei, J. et al. (2020). Convergence and concentration of empirical measures under wasserstein distance in unbounded functional spaces. *Bernoulli*, 26(1):767–798.

Li, M. and Dunson, D. B. (2020). Comparing and weighting imperfect models using d-probabilities. *Journal of the American Statistical Association*, 115(531):1349–1360.

Lindgren, E. M., Shah, V., Shen, Y., Dimakis, A. G., and Klivans, A. (2019). On robust learning of ising models. In *NeurIPS Workshop on Relational Representation Learning*.

Lokhov, A. Y., Vuffray, M., Misra, S., and Chertkov, M. (2018). Optimal structure and parameter learning of ising models. *Science advances*, 4(3):e1700791.

Namkoong, H. and Duchi, J. C. (2016). Stochastic gradient methods for distributionally robust optimization with f-divergences. In *NIPS*, volume 29, pages 2208–2216.

Ng, A. Y. (2004). Feature selection, l 1 vs. l 2 regularization, and rotational invariance. In *Proceedings of*

the twenty-first international conference on Machine learning, page 78.

Nguyen, V. A., Kuhn, D., and Esfahani, P. M. (2018). Distributionally robust inverse covariance estimation: The Wasserstein shrinkage estimator. *arXiv preprint arXiv:1805.07194*.

Nguyen, V. A., Zhang, F., Blanchet, J., Delage, E., and Ye, Y. (2020a). Distributionally robust local non-parametric conditional estimation. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M. F., and Lin, H., editors, *Advances in Neural Information Processing Systems*, volume 33, pages 15232–15242. Curran Associates, Inc.

Nguyen, V. A., Zhang, X., Blanchet, J., and Georghiou, A. (2020b). Distributionally robust parametric maximum likelihood estimation. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M. F., and Lin, H., editors, *Advances in Neural Information Processing Systems*, volume 33, pages 7922–7932. Curran Associates, Inc.

Nikolakakis, K. E., Kalogerias, D. S., and Sarwate, A. D. (2019a). Learning tree structures from noisy data. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 1771–1782.

Nikolakakis, K. E., Kalogerias, D. S., and Sarwate, A. D. (2019b). Non-parametric structure learning on hidden tree-shaped distributions. *arXiv preprint arXiv:1909.09596*.

Prasad, A., Srinivasan, V., Balakrishnan, S., and Ravikumar, P. (2020). On learning ising models under huber's contamination model. *Advances in Neural Information Processing Systems*, 33.

Ravikumar, P., Wainwright, M. J., and Lafferty, J. D. (2010). High-dimensional Ising model selection using ℓ_1-regularized logistic regression. *The Annals of Statistics*, 38(3):1287–1319.

Schneidman, E., Berry, M. J., Segev, R., and Bialek, W. (2006). Weak pairwise correlations imply strongly correlated network states in a neural population. *Nature*, 440(7087):1007–1012.

Shalev-Shwartz, S. and Ben-David, S. (2014). *Understanding machine learning: From theory to algorithms*. Cambridge university press.

Si, N., Zhang, F., Zhou, Z., and Blanchet, J. (2020). Distributionally robust policy evaluation and learning in offline contextual bandits. In *International Conference on Machine Learning (ICML'20)*.

Spirtes, P., Glymour, C. N., Scheines, R., and Heckerman, D. (2000). *Causation, prediction, and search*. MIT press.

Staib, M., Wilder, B., and Jegelka, S. (2019). Distributionally robust submodular maximization. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 506–516.

Szeliski, R., Zabih, R., Scharstein, D., Veksler, O., Kolmogorov, V., Agarwala, A., Tappen, M., and Rother, C. (2006). A comparative study of energy minimization methods for Markov random fields. In *European conference on computer vision*, pages 16–29. Springer.

Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., et al. (2020). Scipy 1.0: fundamental algorithms for scientific computing in python. *Nature methods*, 17(3):261–272.

Vuffray, M., Misra, S., and Lokhov, A. (2020). Efficient learning of discrete graphical models. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M. F., and Lin, H., editors, *Advances in Neural Information Processing Systems*, volume 33, pages 13575–13585. Curran Associates, Inc.

Vuffray, M., Misra, S., Lokhov, A., and Chertkov, M. (2016). Interaction screening: Efficient and sample-optimal learning of Ising models. In *Advances in Neural Information Processing Systems*, pages 2595–2603.

Weed, J., Bach, F., et al. (2019). Sharp asymptotic and finite-sample rates of convergence of empirical measures in wasserstein distance. *Bernoulli*, 25(4A):2620–2648.

Wu, S., Sanghavi, S., and Dimakis, A. G. (2019). Sparse logistic regression learns all discrete pairwise graphical models. In *Advances in Neural Information Processing Systems*, pages 8069–8079.

Yu, Y., Lin, T., Mazumdar, E., and Jordan, M. I. (2021). Fast distributionally robust learning with variance reduced min-max optimization. *arXiv preprint arXiv:2104.13326*.

Supplementary Material: Distributionally Robust Structure Learning for Discrete Pairwise Markov Networks

A OPTIMIZATION DETAILS

This section describes the sub-gradients of the DRO objective functions in the paper for readers who are interested in implementation details. The algorithmic details are illustrated in Algorithm 1.

For the DRO dual problem based on the Wasserstein ambiguity set, we have the objective function

$$f(W,\gamma) \coloneqq \gamma\varepsilon + \frac{1}{m'}\sum_{j=1}^{m'} \sup_{\substack{0\le r\le n-1,\\ g\in\{-1,1\}}} -2r\gamma - \frac{1}{2}\gamma\kappa(1+gy^{(j)}) + \ln\left(1+e^{\langle gW, X^{(j)}\rangle + \langle \mathbb{T}(\delta)_{1\dots r}, \mathbf{1}\rangle}\right).$$

Assume that the inner supremum is achieved at $r^{(j*)}$ and $g^{(j*)}$ for $j\in[m']$. A sub-derivative is

$$\frac{1}{m'}\sum_{j=1}^{m'} \sigma(g^{(j*)}\langle W, X^{(j)}\rangle + \langle \mathbb{T}(\delta)_{1\dots r^{(j*)}}, \mathbf{1}\rangle) g^{(j*)} X^{(j)} \in \frac{\partial}{\partial W} f$$

$$\varepsilon + \frac{1}{m'}\sum_{j=1}^{m'} -2r^{(j*)} - \frac{1}{2}\kappa(1+g^{(j*)}y^{(j)}) \in \frac{\partial}{\partial \gamma} f,$$

where, for each $g\in\{-1,1\}$, $r^{(j*)}$ can be found in $\Theta(n\log n)$ by replacing the rows of $X^{(j)}$ with the basis vector for the row-wise maximum element in gW according to δ.

For the DRO dual problem based on the KL ambiguity set, we have the objective function:

$$f(W,\gamma) \coloneqq \gamma \ln\Big(\frac{1}{m'}\sum_{j=1}^{m'} (1+e^{-y^{(j)}\langle W, X^{(j)}\rangle})^{\frac{1}{\gamma}}\Big) + \gamma\varepsilon,$$

one of whose sub-derivatives is

$$\frac{\sum_{j=1}^{m'} e^{\ell_W(\xi^{(j)})/\gamma}\cdot \sigma(-y^{(j)}\langle W, X^{(j)}\rangle)\cdot(-y^{(j)}X^{(j)})}{\sum_{j=1}^{m'} e^{\ell_W(\xi^{(j)})/\gamma}} \in \frac{\partial}{\partial W} f$$

$$\ln\Big(\frac{1}{m'}\sum_{j=1}^{m'} e^{\ell_W(\xi^{(j)})/\gamma}\Big) - \frac{\sum_{j=1}^{m'} e^{\ell_W(\xi^{(j)})/\gamma}\cdot \ell_W(\xi^{(j)})}{\gamma\sum_{j=1}^{m'} e^{\ell_W(\xi^{(j)})/\gamma}} + \varepsilon \in \frac{\partial}{\partial \gamma} f,$$

where $\ell_W(\xi) \coloneqq \ell(y\langle W, [x_{1\dots i-1,*}^\intercal, v^{(1)\intercal}, x_{i\dots n-1,*}^\intercal]^\intercal\rangle)$.

We can use any optimization algorithm able to leverage sub-gradients to solve these two problems.

B EXPERIMENTAL DETAILS

For reproducibility, the optimal hyperparameters, specifically the ambiguity radius values ε, chosen by grid search on a random graph of the same size as the target graph, are shown in Table 1.

C TECHNICAL PROOFS IN SECTION 3

Theorem 1. *Let $W_1(\cdot,\cdot)$ be the type-1 Wasserstein distance with $p=1$ and metric $d(\xi,\xi') \triangleq d((x,y),(x',y')) \coloneqq \|x-x'\|_{1,1} + \frac{\kappa}{2}|y-y'|$ for $\xi,\xi'\in\Xi$, $\kappa\in\mathbb{R}_+$. Let $B_\varepsilon^{W_1}(\hat{\mathbb{P}}_{m'}) \coloneqq \{\mathbb{Q}\in\mathcal{P}_1(\Xi) : W_1(\hat{\mathbb{P}}_{m'},\mathbb{Q}) \le \varepsilon = \frac{\varepsilon_0}{\sqrt{m'}}\}$ be the*

Algorithm 1 Structure Learning of Discrete Pairwise Graphical Models

Input: Alphabet size k; number of variables n; sample data $\{\bar{z}^{(1)}, \ldots, \bar{z}^{(m)}\}$; 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 $\{z^{(1)}, \ldots, z^{(m')}\}$ with $z_i^j \in \{\alpha, \beta\} \forall j \in [m']$
  Compute $\hat{W}^{(i\alpha\beta)}$ by Eq. (1) or Eq. (8)
  Centering $\hat{W}^{(i\alpha\beta)}$ by Eq. (2)
  Estimate the weight matrices $W^{(ij)}$ by Eq. (3)
  Estimate the edge set $\hat{\mathcal{E}}$ by Eq. (4)
end for

ambiguity set. Then the primal problem in Eq. (7) *is equivalent to*

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

where $X^{(j)} \coloneqq [x^{(j)\intercal}_{1\ldots i-1,*}, v^{(1)\intercal}, x^{(j)\intercal}_{i\ldots n-1,*}]^\intercal$, $\delta \coloneqq [\sup_{l \in [k]} (gW)_{jl} : j \neq i \in [n]]^\intercal - (gW_{-i,*} \odot x^{(j)})\mathbf{1}$, *and* $\mathbb{T}(x)$ *is defined as a vector with non-decreasing components as a result of sorting* x*, introduced in Section 2.1.*

Proof. Recall that $\Xi = V^{((n-1)\times k)} \times \{-1, 1\}$ where $V^{((n-1)\times k)}$ is the set of matrices with rows of basis vectors. To avoid clutter of notations, we define $\ell_W(\xi) \coloneqq \ell(y\langle W, [x^\intercal_{1\ldots i-1,*}, v^{(1)\intercal}, x^\intercal_{i\ldots n-1,*}]^\intercal \rangle)$. Similar to Abadeh et al. (2015), we rewrite the worst-case risk in Equation (7) as

$$\sup_{\mathbb{Q} \in B_\varepsilon(\hat{\mathbb{P}}_{m'})} \int_\Xi \ell_W(\xi')\mathbb{Q}(\mathrm{d}\xi') = \begin{cases} \sup\limits_{\Pi \in M(\Xi^2)} & \int_\Xi \ell_W(\xi')\Pi(\Xi, \mathrm{d}\xi') \\ \text{s.t.} & \int_{\Xi^2} d(\xi, \xi')\Pi(\mathrm{d}\xi, \mathrm{d}\xi') \leq \varepsilon \\ & \Pi(\mathrm{d}\xi, \Xi) = \hat{\mathbb{P}}_{m'}(\mathrm{d}\xi). \end{cases}$$

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

$$\begin{cases} \sup\limits_{\mathbb{Q}^{(j)}} & \frac{1}{m'} \sum_{j=1}^{m'} \int_\Xi \ell_W(\xi')\mathbb{Q}^{(j)}(\mathrm{d}\xi') \\ \text{s.t.} & \frac{1}{m'} \sum_{j=1}^{m'} \int_\Xi d(\xi^{(j)}, \xi')\mathbb{Q}^{(j)}(\mathrm{d}\xi') \leq \varepsilon \\ & \int_\Xi \mathbb{Q}^{(j)}(\mathrm{d}\xi') = 1, \forall j \in [m']. \end{cases} \tag{10}$$

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

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

which can simplify (10) to

$$\begin{cases} \sup\limits_{\mathbb{Q}^{(j)}_{\pm 1}} & \frac{1}{m'} \sum_{j=1}^{m'} \int_{V^{((n-1)\times k)}} \ell_W((x', -1))\mathbb{Q}^{(j)}_{-1}(\mathrm{d}x') + \ell_W((x', +1))\mathbb{Q}^{(j)}_{+1}(\mathrm{d}x') \\ \text{s.t.} & \frac{1}{m'} \sum_{j=1}^{m'} \int_{V^{((n-1)\times k)}} d(\xi^{(j)}, (x', -1))\mathbb{Q}^{(j)}_{-1}(\mathrm{d}x') + d(\xi^{(j)}, (x', +1))\mathbb{Q}^{(j)}_{+1}(\mathrm{d}x') \leq \varepsilon \\ & \int_{V^{((n-1)\times k)}} \mathbb{Q}^{(j)}_{-1}(\mathrm{d}x') + \mathbb{Q}^{(j)}_{+1}(\mathrm{d}x') = 1, \forall j \in [m']. \end{cases}$$

Table 1: Optimal radii of ambiguity sets in all settings of our experiments.

METHOD	GRAPH	n	k	θ	ζ	NOISE MODEL	ε_0^*
KLDRSL	Diamond	6	4	0.2	0.2	Noiseless	20
WDRSL	Diamond	6	4	0.2	0.2	Noiseless	1.2
KLDRSL	Diamond	9	4	0.2	0.2	Noiseless	26
WDRSL	Diamond	9	4	0.2	0.2	Noiseless	1.2
KLDRSL	Diamond	12	4	0.2	0.2	Noiseless	28
WDRSL	Diamond	12	4	0.2	0.2	Noiseless	1.5
KLDRSL	Diamond	6	4	0.1	0.2	Noiseless	18
WDRSL	Diamond	6	4	0.1	0.2	Noiseless	1.2
KLDRSL	Diamond	6	4	0.3	0.2	Noiseless	26
WDRSL	Diamond	6	4	0.3	0.2	Noiseless	1.2
KLDRSL	Diamond	6	2	0.2	0.2	Noiseless	1
WDRSL	Diamond	6	2	0.2	0.2	Noiseless	0.1
KLDRSL	Diamond	6	6	0.2	0.2	Noiseless	55
WDRSL	Diamond	6	6	0.2	0.2	Noiseless	2.0
KLDRSL	Grid	9	4	0.2	0.2	Noiseless	33
WDRSL	Grid	9	4	0.2	0.2	Noiseless	1.5
KLDRSL	Grid	9	4	0.2	0.2	Huber	17
WDRSL	Grid	9	4	0.2	0.2	Huber	1.1
KLDRSL	Grid	9	4	0.2	0.3	Huber	12
WDRSL	Grid	9	4	0.2	0.3	Huber	0.9
KLDRSL	Grid	9	4	0.2	0.5	Huber	1
WDRSL	Grid	9	4	0.2	0.5	Huber	0.3
KLDRSL	Grid	9	4	0.2	0.1	Independent	14
WDRSL	Grid	9	4	0.2	0.1	Independent	1.0
KLDRSL	Grid	9	4	0.2	0.2	Independent	3
WDRSL	Grid	9	4	0.2	0.2	Independent	0.5
KLDRSL	Grid	9	4	0.2	0.3	Independent	0.02
WDRSL	Grid	9	4	0.2	0.3	Independent	0.04

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

$$
\begin{cases}
\sup\limits_{\mathbb{Q}_{\pm 1}^{(j)}} & \dfrac{1}{m'}\displaystyle\sum_{j=1}^{m'}\int_{V^{((n-1)\times k)}} \ell_W((x',-1))\mathbb{Q}_{-1}^{(j)}(\mathrm{d}x') + \ell_W((x',+1))\mathbb{Q}_{+1}^{(j)}(\mathrm{d}x') \\
\text{s.t.} & \dfrac{1}{m'}\displaystyle\int_{V^{((n-1)\times k)}} \kappa \sum_{j:y^{(j)}=-1} \mathbb{Q}_{+1}^{(j)}(\mathrm{d}x') + \kappa \sum_{j:y^{(j)}=+1} \mathbb{Q}_{-1}^{(j)}(\mathrm{d}x') \\
& \quad + \displaystyle\sum_{j=1}^{m'} \|x^{(j)} - x'\|_1 (\mathbb{Q}_{-1}^{(j)}(\mathrm{d}x') + \mathbb{Q}_{+1}^{(j)}(\mathrm{d}x')) \le \varepsilon \\
& \quad \displaystyle\int_{V^{((n-1)\times k)}} \mathbb{Q}_{-1}^{(j)}(\mathrm{d}x') + \mathbb{Q}_{+1}^{(j)}(\mathrm{d}x') = 1, \forall j \in [m'].
\end{cases}
$$

Its Lagrange dual problem is as follows:

$$\begin{cases} \inf\limits_{\gamma, s^{(j)}} & \gamma\varepsilon + \dfrac{1}{m'}\sum\limits_{j=1}^{m'} s^{(j)} \\ \text{s.t.} & \sup\limits_{x' \in V^{((n-1)\times k)}} \ell_W((x',-1)) - \gamma\|x^{(j)} - x'\|_1 - \dfrac{1}{2}\gamma\kappa(1+y^{(j)}) \le s^{(j)} \quad \forall j \in [m'] \\ & \sup\limits_{x' \in V^{((n-1)\times k)}} \ell_W((x',+1)) - \gamma\|x^{(j)} - x'\|_1 - \dfrac{1}{2}\gamma\kappa(1-y^{(j)}) \le s^{(j)} \quad \forall j \in [m'] \\ & \gamma \ge 0. \end{cases}$$

Strong duality holds according to Theorem 1 in Gao and Kleywegt (2016). By incorporating the outer minimization of Equation (7), plugging in the expression of $\ell_W(\cdot)$ and rearranging the terms in the above expressions, we have

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

where $X = [x_{1\dots i-1,*}^{\intercal}, v^{(1)\intercal}, x_{i\dots n-1,*}^{\intercal}]^{\intercal}$. 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 x:

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

Let $X^{(j)} \coloneqq [x_{1\dots i-1,*}^{(j)\intercal}, v^{(1)\intercal}, x_{i\dots n-1,*}^{(j)\intercal}]^{\intercal}$ and $\delta \coloneqq [\sup_{l\in[k]}(gW)_{jl} : j \ne i \in [n]]^{\intercal} - (gW_{-i,*} \odot x^{(j)})\mathbf{1}$. As a result, $\delta \in \mathbb{R}^{n-1}$ is a vector of differences between the maximum and the selected element according to $x^{(j)}$ for each row of $W_{-i,*}$. Denote by $B = (b_1, \dots, b_{n-1})$ a permutation of $[n-1]$ satisfying $\delta_{b_1} \ge \delta_{b_2} \ge \cdots \ge \delta_{b_{n-1}}$. It is thus not hard to show that, for any integer $0 \le r \le n-1$, and $x \in V^{((n-1)\times k)}$ that satisfies $\|x^{(j)} - x\|_1 = 2r$, we have

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

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

$$\ln\left(1+e^{\langle gW,X^{(j)}\rangle + \sum_{u=1}^{r}\delta_{b_u}}\right) - 2r\gamma \ge \ln\left(1+e^{\langle gW,X^{(j)}\rangle + \sum_{u=1}^{r'}\delta_{b_u}}\right) - 2r'\gamma, \forall r \le r'.$$

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:

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

To characterize the sequence of the sorted indices in B more formally, we have defined $\mathbb{T}(x)$ as a vector of sorted components of x in Section 2. In such matter, we can reformulate the above convex program as

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

□

Theorem 2. *If $\kappa = \infty$, $\|W\|_{2,1} \le 2\lambda\sqrt{k}$ and $\gamma \ge (n+2)\lambda\sqrt{k}$, the convex program in (8) subsumes the standard $\ell_{2,1}$-constrained logistic regression approach in (1) as a special case. If $\kappa = \infty$ and $\gamma \ge \frac{n+2}{2}\|W\|_{2,1}$, it subsumes the $\ell_{2,1}$-regularized logistic regression approach as a special case.*

Proof. To begin with, we rewrite Eq. (8) based on the cases where $g = y^{(j)}$ and $g = -y^{(j)}$:

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

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

$$\inf_{\substack{W\in\mathbb{R}^{n\times k},\\ \gamma>0}} \gamma\varepsilon + \frac{1}{m'}\sum_{j=1}^{m'} \sup_{0\leq r\leq n-1} -2r\gamma + \ln\left(1+e^{\langle -y^{(j)}W, X^{(j)}\rangle + \langle \mathbb{T}(\delta)_{1\ldots r}, \mathbf{1}\rangle}\right). \tag{11}$$

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

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

Therefore, the supremum in Eq. (11) is achieved only when $r = 0$. Finally, Eq. (11) can be rewritten as the following convex program:

$$\begin{cases} \inf\limits_{W} & \dfrac{1}{m'}\displaystyle\sum_{j=1}^{m'} \ln\left(1+e^{-y^{(j)}\langle W, X^{(j)}\rangle}\right)\\ \text{s.t.} & \|W\|_{2,1} \leq 2\lambda\sqrt{k}, \end{cases}$$

which coincides with the $\ell_{2,1}$-constrained logistic regression problem in Eq. (1).

On the other hand, if $\gamma \geq \frac{n+2}{2}\|W\|_{2,1}$, by following the above same process, the supremum in Eq. (11) is achieved only when $r = 0$. Note that only the first term in Eq.(11) is related to γ. After minimizing over γ, we can rewrite Eq.(11) as

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

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

Corollary 1. *Let $B_\varepsilon^{\mathrm{KL}}(\hat{\mathbb{P}}_{m'}) \coloneqq \{\mathbb{Q} \in \mathcal{P}(\Xi) : D_{\mathrm{KL}}(\mathbb{Q}, \hat{\mathbb{P}}_{m'}) \leq \varepsilon = \frac{\varepsilon_0}{m'}\}$ be a KL divergence ball. The primal problem in Eq. (7) with $B_\varepsilon(\hat{\mathbb{P}}_{m'}) = B_\varepsilon^{\mathrm{KL}}(\hat{\mathbb{P}}_{m'})$ is equivalent to*

$$\inf_{,\gamma\geq 0} \gamma\ln\Big(\frac{1}{m'}\sum_{j=1}^{m'}(1+e^{-y^{(j)}\langle W, X^{(j)}\rangle})^{\frac{1}{\gamma}}\Big) + \gamma\varepsilon.$$

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

D TECHNICAL PROOFS IN SECTION 4

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

$$\mathcal{L}(w) - \hat{\mathcal{L}}(w) \leq 2L_\ell X_{2,\infty} W_{2,1}\sqrt{\frac{6\ln(n)}{m}} + L_\ell X_{2,\infty} W_{2,1}\sqrt{\frac{2\ln(2/\rho)}{m}}.$$

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

Proposition 1. *Assume that (Ξ, d) 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 $\mathbb{P} \in \mathcal{P}_p(\Xi)$, $\mathcal{A} = B_\varepsilon^{W_p}(\hat{\mathbb{P}}_m)$ is a type-p Wasserstein ball centered at $\hat{\mathbb{P}}_m$ with radius ε, $\mathcal{F}$ is a space of closed convex functions $f : \Xi \to \mathbb{R}_+$ with $\mathrm{lip}_d(f) < \infty$. Let $\hat{f}$ be a minimizer of the DRO problem in Eq. (6) and f^* be a minimizer of the stochastic optimization problem in Eq. (5), we have*

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

Proof. To avoid clutter of notations, we define $B(\mathbb{P}) \coloneqq B_\varepsilon^{W_p}(\mathbb{P})$.

According to Theorem 1 in Cranko et al. (2020), the following relation holds for any $f \in \mathcal{F}$ and a fixed $\mathbb{P} \in \mathcal{P}_p(\Xi)$:

$$\int_\Xi f(\xi)\mathbb{P}(\mathrm{d}\xi) \leq \sup_{\mathbb{Q}\in B(\mathbb{P})} \int_\Xi f(\xi)\mathbb{Q}(\mathrm{d}\xi) \leq \int_\Xi f(\xi)\mathbb{P}(\mathrm{d}\xi) + \varepsilon \mathrm{lip}_d(f).$$

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

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

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

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

□

Theorem 3. *Given that: $\mathcal{D}(\mathcal{W},\Theta)$ is an unknown pairwise Markov network with n variables, alphabet size k, dependency graph $\mathcal{G}$; that Assumptions 1 and 2 hold; that $\|W\|_{2,1}\leq 2\lambda\sqrt{k}$ in Eq. (7); that $W^{(ij)}\in\mathcal{W}$ is the true weight matrix; and that $\hat{W}^{(ij)}$ is the estimated weight matrix from Eq. (8) with the Wasserstein ambiguity set and properly centered (Section 2.2), 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 $m=O(\frac{\lambda^2k^4e^{14\lambda}(\varepsilon_0^2+\ln\frac{nk}{\rho})}{\omega^4})$, with probability at least $1-\rho$, the following bound holds:*

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

Let $\omega<\frac{\eta}{2}$ and $\hat{\mathcal{G}}$ be reconstructed via thresholding in Eq. (4). Now if $m=O(\frac{\lambda^2k^4e^{14\lambda}(\varepsilon_0^2+\ln\frac{nk}{\rho})}{\eta^4})$, 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 $\ell_W(\xi)\coloneqq\ell(y\langle W,[x_{1\dots i-1,*}^{\intercal},v^{(1)\intercal},x_{i\dots n-1,*}^{\intercal}]^{\intercal}\rangle)$.

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

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

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

$$\mathbb{P}^{m'}\left\{\mathbb{E}^{\mathbb{P}}[\ell_W(\xi)]-\mathbb{E}^{\hat{\mathbb{P}}_{m'}}[\ell_W(\xi)]\leq 2\lambda\sqrt{k}(2\sqrt{\frac{6\ln(n)}{m'}}+\sqrt{\frac{2\ln(2/\rho)}{m'}})\right\}\geq 1-\rho. \tag{12}$$

Define $W^{(i\alpha\beta)}\in\mathbb{R}^{n\times k}$ as $W_{i*}^{(i\alpha\beta)}\coloneqq[\theta_{\alpha}^{(i)}-\theta_{\beta}^{(i)},\mathbf{0}^{\intercal}]$, and $W_{j*}^{(i\alpha\beta)}\coloneqq W_{\alpha*}^{(ij)}-W_{\beta*}^{(ij)}$ for $j\neq i\in[n]$. Recall that $\hat{W}^{(i\alpha\beta)}$ is a minimizer of Eq. (7) with a Wasserstein ball:

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

By Proposition 1,

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

which can be combined with Eq. (12) and the definition $\varepsilon = \varepsilon_0/\sqrt{m'}$, yielding

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

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 $1-\rho$,

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

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 $1-\rho$,

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

Now fix some $i \in [n]$, $\alpha \neq \beta \in [k]$. Denote by $m^{(i\alpha\beta)}$ the number of samples in which $\bar{z}_i^j \in \{\alpha, \beta\}$. Recall that $U^{(i\alpha\beta)}$, the centered version of $\hat{W}^{(i\alpha\beta)}$, satisfies $\langle \hat{W}^{(i\alpha\beta)}, x\rangle = \langle U^{(i\alpha\beta)}, x\rangle$. 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 $1-\rho$,

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

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 $Z \sim \mathcal{D}$ is δ-unbiased with $\delta = e^{-2\lambda(\mathcal{D})/k}$, and so is 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 $1-\rho'$:

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

Since $Z \sim \mathcal{D}$ is δ-unbiased, we have $\mathbb{P}[Z_i \in \{\alpha, \beta\}] \ge 2\delta$. 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

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

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

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

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

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

□

Theorem 4. *Given assumptions in Theorem 3, except that $\hat{W}^{(ij)}$ is the estimated weight matrix from Eq. (9) with the KL ambiguity set. Let $\hat{\mathcal{G}}$ be constructed via thresholding in Eq. (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 $m=O(\frac{\lambda^2k^4e^{14\lambda}(\varepsilon_0+\ln\frac{nk}{\rho})}{\eta^4})$, with probability at least $1-\rho$, the following bound holds:*

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

Proof. According to Theorem 7 in Lam (2019), for any W,

$$\begin{aligned}
&\mathbb{E}^{\hat{\mathbb{P}}_m}[\ell_W(\xi)]\\
\leq&\sup_{\mathbb{Q}\in B_\varepsilon^{\mathrm{KL}}(\hat{\mathbb{P}}_m)}\mathbb{E}^{\mathbb{Q}}[\ell_W(\xi)]\\
\leq&\mathbb{E}^{\hat{\mathbb{P}}_m}[\ell_W(\xi)]+\sqrt{2\varepsilon\mathrm{Var}_{\hat{\mathbb{P}}_m}(\ell_W(\xi))}+2\varepsilon C\frac{1}{m'}\frac{\sum_i(\ell_W(\xi_i)-\overline{\ell_W}(\xi))^3}{\sum_i(\ell_W(\xi_i)-\overline{\ell_W}(\xi))^2}\\
\leq&\mathbb{E}^{\hat{\mathbb{P}}_m}[\ell_W(\xi)]+\sqrt{2\varepsilon\mathrm{Var}_{\hat{\mathbb{P}}_m}(\ell_W(\xi))}+2\varepsilon C\frac{1}{m'}\sum_i|\ell_W(\xi_i)-\overline{\ell_W}(\xi)|,
\end{aligned}$$

where $\overline{\ell_W}=\frac{1}{m'}\sum_i\ell_W(\xi_i)$ and $C>0$ is a constant independent of n.

Note that

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

yielding

$$\begin{aligned}
&\sup_{\mathbb{Q}\in B_\varepsilon^{\mathrm{KL}}(\hat{\mathbb{P}}_m)}\mathbb{E}^{\mathbb{Q}}[\ell_W(\xi)]\\
\leq&\mathbb{E}^{\hat{\mathbb{P}}_m}[\ell_W(\xi)]+4\lambda\sqrt{k}(\sqrt{2\varepsilon}+2\varepsilon C)\\
\leq&\mathbb{E}^{\hat{\mathbb{P}}_m}[\ell_W(\xi)]+4\lambda\sqrt{k}(2\sqrt{\varepsilon}+2C\sqrt{\varepsilon})
\end{aligned}$$

for $\varepsilon<1$.

Therefore,

$$\begin{aligned}
&\mathbb{P}^{m'}\left\{\mathbb{E}^{\mathbb{P}}[\ell_{\hat{W}^{(i\alpha\beta)}}(\xi)]-\mathbb{E}^{\mathbb{P}}[\ell_{W^{(i\alpha\beta)}}(\xi)]\leq2\lambda\sqrt{k}((4C+4)\sqrt{\frac{\varepsilon_0}{m'}}+4\sqrt{\frac{6\ln(n)}{m'}}+2\sqrt{\frac{2\ln(2/\rho)}{m'}})\right\}\\
\geq&1-\rho.
\end{aligned}$$

Following the same procedure in the proof of Theorem 3, we get the conclusion that with probability at least $1-\rho$ and $m=O(\frac{\lambda^2k^4e^{14\lambda}(\varepsilon_0+\ln\frac{nk}{\rho})}{\omega^4})$,

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

□