Distributionally Robust Graph Learning from Smooth Signals under Moment Uncertainty

Xiaolu Wang, Yuen-Man Pun, and Anthony Man-Cho So, *Senior Member, IEEE*

***Abstract*—We consider the problem of learning a graph from a finite set of noisy graph signal observations, the goal of which is to find a smooth representation of the graph signal. Such a problem is motivated by the desire to infer relational structure in large datasets and has been extensively studied in recent years. Most existing approaches focus on learning a graph on which the observed signals are smooth. However, the learned graph is prone to overfitting, as it does not take the unobserved signals into account. To address this issue, we propose a novel graph learning model based on the distributionally robust optimization methodology, which aims to identify a graph that not only provides a smooth representation of but is also robust against uncertainties in the observed signals. On the statistics side, we establish out-of-sample performance guarantees for our proposed model. On the optimization side, we show that under a mild assumption on the graph signal distribution, our proposed model admits a smooth non-convex optimization formulation. We then develop a projected gradient method to tackle this formulation and establish its convergence guarantees. Our formulation provides a new perspective on regularization in the graph learning setting. Moreover, extensive numerical experiments on both synthetic and real-world data show that our model has comparable yet more robust performance across different populations of observed signals than existing models according to various metrics.**

***Index Terms*—Graph learning, network topology inference, graph signal processing, distributionally robust optimization, moment uncertainty**

I. Introduction

A. Background and Motivation

With the widespread availability of large, complex but structured datasets, one fundamental problem in contemporary data processing and analysis is that of inferring relationships among different entities using the data observed from them. Such a problem arises in many different application areas, including road traffic analysis, brain connectivity analysis, and community detection in social networks, just to name a few [1]. To tackle this problem, a common approach is to first model the entities as nodes of an undirected, weighted graph and the data observed from the entities as signals residing on the nodes, and then to learn the edges of the graph together with their weights based on the observed signals. Clearly, in order for the learning task to be well defined, it is necessary to specify how the signals are related to the graph topology. Towards that end, various models have been proposed in the literature; see, e.g., [2], [3] and the references therein. One representative model, which we shall refer to as the *smooth graph signal* model and is motivated by considerations of real-world graph-structured data, postulates that the observed signals vary smoothly on the underlying graph—i.e., signal values at nodes that are adjacent to each other should be similar. In this model, a widely used measure of smoothness is the Laplacian quadratic form. On one hand, the Laplacian quadratic form can be viewed as a discrete analog of the Dirichlet energy—a measure of variability for smooth functions—and has been used early on as a regularizer for learning problems on graphs [4], [5]. On the other hand, by drawing connections to classic signal processing concepts, one can interpret the eigenvectors of the Laplacian as frequency components and the eigenvalues as frequencies of the underlying graph [1], [2]. As such, the Laplacian quadratic form captures the variability of a given signal over the different graph frequency components. There has been a number of works that assume the smooth graph signal model and propose to learn the graph topology by solving a regularized Laplacian quadratic form minimization problem, where the regularizer is used to induce certain structure in the learned graph. For instance, the works [6], [7] use a squared Frobenius norm regularizer to control the distribution of edge weights in the learned graph, while the work [8] combines the squared Frobenius norm with a logarithmic barrier to control both the sparsity and connectivity of the learned graph. Recently, some extensions of these formulations have been proposed, in which additional hard constraints are imposed on the graph topology; see, e.g., [9]–[11].

In order to assess the performance of a graph learning procedure, one possible approach is to first assume that the graph signal follows certain statistical model—for example, the Gaussian Markov random field (GMRF) or the factor analysis model (see, e.g., [2], [3] and the references therein)—and then evaluate the performance of the learned graph as a statistical estimator of the underlying graph. In the context of regularized Laplacian quadratic form minimization, if we assume that the graph signal is generated according to a ground-truth probability distribution (which in general is not known and depends on the underlying graph), then many existing formulations (such as those in [6], [7], [10], [11]) can be viewed as minimizing a regularized empirical risk of the observed signals, where the risk function is given precisely by the Laplacian quadratic form. Such a viewpoint raises

This work was supported in part by the Hong Kong RGC GRF project CUHK 14203920. X. Wang and A. M.-C. So are with the Department of Systems Engineering and Engineering Management, The Chinese University of Hong Kong, Hong Kong SAR, China (e-mails: {xlwang, manchoso}@se.cuhk.edu.hk). Y.-M. Pun is with the CIICADA Lab, School of Engineering, The Australian National University, Canberra, ACT 2601, Australia (e-mail: YuenMan.Pun@anu.edu.au). Most of the work of Y.-M. Pun was done when she was a Ph.D. student at the Department of Systems Engineering and Engineering Management of The Chinese University of Hong Kong. This paper has supplementary downloadable material available at http://ieeexplore.ieee.org, provided by the author. This material is 0.1MB in size.

the interesting possibility of analyzing the performance of these formulations using the vast array of tools developed in the statistical learning community for studying regularized empirical risk minimization (ERM) problems. Nevertheless, to the best of our knowledge, such a possibility has barely been pursued in the graph learning literature. In fact, two important questions concerning the regularized Laplacian quadratic form minimization approach remain unresolved:

- The approach essentially only uses the empirical distribution of the observed signals to learn the graph. As such, it can be prone to overfitting. In other words, the graph learned using the empirical distribution of the observed signals may differ greatly from the one learned if the ground-truth distribution of the graph signal were known. Is it possible to develop an alternative approach that can better exploit the information about the ground-truth distribution contained in the observed signals, so as to alleviate the effect of overfitting in the learned graph?
- The regularizers used in existing formulations are usually ad-hoc in nature, and their impact on the quality of the learned graph is not well understood theoretically. Is there a more principled approach to regularization, so that one can construct regularizers whose effects on the learned graph can be rigorously explained?

B. Our Contributions

In this paper, we propose to take a distributionally robust optimization (DRO) approach to addressing the above questions. Specifically, with the risk function given by the Laplacian quadratic form, instead of minimizing the *empirical risk* (i.e., expected risk with respect to (wrt) the empirical distribution) of the observed signals as in existing formulations, we minimize a *worst-case expected risk*, where the worst-case expectation is taken wrt a set (called the *ambiguity set*) of probability distributions that are consistent with certain information obtained from the observed signals. Intuitively, if the ambiguity set is chosen appropriately, then it contains the ground-truth distribution of the graph signal, which suggests that the graph learned by minimizing the worst-case expected risk not only provides a smooth representation of the graph signal but is also less susceptible to overfitting. Although there is a host of recent works that develop DRO-based techniques to tackle the issue of overfitting in statistical learning tasks (see, e.g., [12]), our work is the first to pursue such an approach in the graph learning setting. Interestingly, our technical developments also lead to novel contributions to both the modeling and algorithmic aspects of DRO. Let us now summarize the main contributions of this paper.

1) Modeling: Using the fact that the risk function is quadratic in the graph signal, it is straightforward to show that for any given Laplacian, the expected risk depends only on the first two moments of the graph signal probability distribution. Based on this, we propose a novel moment-based distributionally robust graph learning model, in which the ambiguity set contains all distributions whose mean vectors and covariance matrices are close to the empirical mean and empirical covariance of the observed signals, respectively, and the goal is to find a Laplacian that minimizes the worst-case expected risk wrt such an ambiguity set. A notable feature of our proposed ambiguity set is that it depends on the decision variable of the model, namely the Laplacian. Such a dependence is crucial in the context of graph learning, as the probability distribution from which the observed signals are generated should depend on the underlying graph. However, the techniques currently available in the DRO literature for tackling decision-dependent ambiguity sets are very limited (see [13] and the references therein) and do not apply to the setting considered in this work. This motivates us to develop new techniques to handle the ambiguity set in our proposed model; see "*Performance Analysis and Reformulation*" below for further elaboration.

We remark that in view of the recent literature on DRO approaches to statistical learning, one may be tempted to consider a distributionally robust graph learning model similar to ours but with a *Wasserstein distance*-based ambiguity set. Such an approach has in fact been explored in the recent work [14], which appeared on arXiv at almost the same time as this work. However, the ambiguity set proposed in [14] does not depend on the Laplacian. Thus, the model in [14] fails to capture the interaction between the graph signal probability distribution and the structure of the underlying graph. In addition, since the expected risk does not distinguish between distributions with the same first and second moments, the Wasserstein distance-based ambiguity set carries much more information than is necessary for the purpose of evaluating the worst-case expected risk. This renders the subsequent analysis of the model more challenging and less direct than our moment-based model.

2) Performance Analysis and Reformulation: Under the assumption that the ground-truth distribution of the graph signal satisfies certain moment growth condition, we construct confidence regions around its mean and covariance by invoking the appropriate concentration inequalities. These regions not only provide a principled way of tuning the size of the ambiguity set but also yield a bound on the expected risk wrt the unknown ground-truth distribution of the graph signal (also known as the *out-of-sample risk*). We then show that the worst-case expected risk minimization problem in our proposed model can be reformulated as a regularized ERM problem, in which the regularizers serve to promote robustness of the learned graph against the uncertainty described in the ambiguity set and the regularization parameters control the size of the ambiguity set. This establishes for the first time a rigorous link between distributional robustness and regularization in the graph learning setting.

3) Algorithm Design and Analysis: Although the aforementioned regularized ERM formulation has nice theoretical interpretations, it is computationally challenging to solve, as its objective function is generally non-smooth and non-convex. Nevertheless, we establish the curious result that as long as the ground-truth distribution of the graph signal has a probability density function, the objective function of the said formulation, though still non-convex, will be smooth almost surely. Consequently, we can apply a projected gradient descent (PGD) method to tackle the formulation. We show that

the iterates generated by the PGD method converge to a stationary point of the regularized ERM formulation. Moreover, we prove that the same convergence result holds for various extensions of the formulation, which could be of independent interest. It is worth mentioning that our work contributes to the emerging area of algorithm design and analysis for DRO; for related works, see, e.g., [15], [16]. To verify the efficacy of our proposed model and algorithm, we conduct extensive numerical experiments on both synthetic and real-world data. The results show that our proposed distributionally robust graph learning model is competitive—as measured by standard performance metrics—with several representative graph learning models in the literature. Moreover, when tested on different populations of observed signals, the former generally achieves a smaller variance in its performance than the latter. This demonstrates the value of incorporating moment-based distributional robustness in graph learning models.

C. Notation

The notation used in this paper is mostly standard. We use $\mathbb{S}^m$, $\mathbb{S}^m_+$, and $\mathbb{S}^m_{++}$ to denote the set of $m\times m$ symmetric, symmetric positive semidefinite, and symmetric positive definite matrices, respectively. We use $\mathbf{1}$ (resp. $\mathbf{0}$) to denote the all-one (resp. all-zero) matrix, whose dimension will be clear from the context, and $\boldsymbol{I}_m$ to denote the $m \times m$ identity matrix. Given a matrix $\boldsymbol{A} \in \mathbb{S}^m$, we use A_{ij} to denote its (i,j)-th element, $\boldsymbol{A}^\dagger$ to denote its Moore–Penrose inverse (or pseudo-inverse), and $\|\boldsymbol{A}\|_F$ (resp. $\|\boldsymbol{A}\|$) to denote its Frobenius (resp. spectral) norm. Given a probability distribution $\mathbb{Q}$, we write $\boldsymbol{x} \sim \mathbb{Q}$ to mean that the random vector $\boldsymbol{x}$ is distributed according to $\mathbb{Q}$, $\mathbb{E}_{\boldsymbol{x}\sim\mathbb{Q}}[\cdot]$ and $\mathrm{Cov}_{\boldsymbol{x}\sim\mathbb{Q}}(\cdot)$ to denote the expectation and covariance wrt $\mathbb{Q}$, respectively. We use $\Pr(\cdot)$ to denote probability, the distribution wrt which it is evaluated will be clear from the context. Given a vector $\boldsymbol{\mu} \in \mathbb{R}^m$ and a matrix $\boldsymbol{\Sigma} \in \mathbb{S}^m_+$, we use $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ to denote the multivariate normal distribution with mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}$.

D. Paper Organization

The rest of this paper is organized as follows: In Section II, we present our proposed moment-based distributionally robust graph learning model and study its statistical and optimization properties. Then, in Section III, we discuss how the PGD method can be used to tackle our proposed model and analyze its convergence behavior under various settings. Next, we describe our experiment setups and report numerical results in Section IV. Finally, we close with some concluding remarks in Section V.

II. Distributionally Robust Graph Learning

A. Problem Formulation

Consider n given signals $\boldsymbol{x}^1, \ldots, \boldsymbol{x}^n \in \mathbb{R}^m$, where $x_i^j \in \mathbb{R}$ $(i = 1, \ldots, m;\ j = 1, \ldots, n)$ denotes the j-th observed value at node i of an unknown m-node, weighted, undirected graph $\mathcal{G}$. We view these signals as independent realizations of a random vector that follows a probability distribution $\mathbb{P}^*$ (called the *ground-truth distribution*) associated with the graph $\mathcal{G}$; cf. the statistical models for graph signals discussed in [3] and the references therein. To identify the graph topology that yields a smooth representation of the graph signal, a popular strategy (see, e.g., [6]–[8]) is to consider the following regularized Laplacian quadratic form minimization problem:

$$\inf_{\boldsymbol{L}\in\mathcal{L}_s} \left\{ \frac{1}{n} \operatorname{tr}(\boldsymbol{X}^\top \boldsymbol{L} \boldsymbol{X}) + h(\boldsymbol{L}) \right\}. \tag{1}$$

Here, $\boldsymbol{X} \coloneqq [\boldsymbol{x}^1\ \cdots\ \boldsymbol{x}^n] \in \mathbb{R}^{m\times n}$ is the data matrix whose columns are the observed signals,

$$\mathcal{L}_s \coloneqq \left\{ \boldsymbol{L} \in \mathbb{S}^m : \begin{array}{r} L_{ij} \le 0 \text{ for } i \ne j, \\ \boldsymbol{L}\mathbf{1} = \mathbf{0}, \\ \operatorname{tr}(\boldsymbol{L}) = 2s \end{array} \right\}$$

is the set of $m\times m$ graph Laplacians whose scale is controlled by the parameter $s > 0$, and $h : \mathbb{S}^m \to \mathbb{R}\cup\{+\infty\}$ is a convex regularizer that aims to promote certain structure in the learned Laplacian. Due to the presence of the scale parameter $s > 0$, an optimal solution to Problem (1), whenever it exists, is a Laplacian that is not identically zero by definition. Thus, it can be used to construct the learned graph.

Although there have been extensive studies on the different choices of regularizer for Problem (1) and their effects on the learned graph, the ramifications of the fact that the graph learned by solving (1) depends on the particular realizations $\{\boldsymbol{x}^j\}_{j=1}^n$ are seldom addressed. To better understand the issue at hand, it is instructive to view Problem (1) through the lens of empirical risk minimization (ERM) in statistical learning. Specifically, let $\widehat{\mathbb{P}}_n$ be the empirical distribution associated with the data $\{\boldsymbol{x}^j\}_{j=1}^n$. Define $\mathcal{L}_s \times \mathbb{R}^n \ni (\boldsymbol{L}, \boldsymbol{x}) \mapsto R(\boldsymbol{L}, \boldsymbol{x}) \coloneqq \boldsymbol{x}^\top \boldsymbol{L} \boldsymbol{x} \in \mathbb{R}_+$ to be the risk function. Since

$$\mathbb{E}_{\boldsymbol{x}\sim\widehat{\mathbb{P}}_n}[R(\boldsymbol{L}, \boldsymbol{x})] = \frac{1}{n}\sum_{j=1}^n {\boldsymbol{x}^j}^\top \boldsymbol{L} \boldsymbol{x}^j = \frac{1}{n} \operatorname{tr}(\boldsymbol{X}^\top \boldsymbol{L} \boldsymbol{X}),$$

we see that Problem (1) is equivalent to the following ERM problem:

$$\inf_{\boldsymbol{L}\in\mathcal{L}_s} \left\{ \mathbb{E}_{\boldsymbol{x}\sim\widehat{\mathbb{P}}_n}[R(\boldsymbol{L}, \boldsymbol{x})] + h(\boldsymbol{L}) \right\}. \tag{2}$$

Such a formulation reveals that an optimal solution $\widehat{\boldsymbol{L}}$ to Problem (1) may suffer from *overfitting*—it yields a graph on which the *observed* signals $\{\boldsymbol{x}^j\}_{j=1}^n$ are smooth but some *unseen* signals generated according to $\mathbb{P}^*$ are not, so that the gap between the *in-sample risk* $\mathbb{E}_{\boldsymbol{x}\sim\widehat{\mathbb{P}}_n}[R(\widehat{\boldsymbol{L}}, \boldsymbol{x})]$ and *out-of-sample risk* $\mathbb{E}_{\boldsymbol{x}\sim\mathbb{P}^*}[R(\widehat{\boldsymbol{L}}, \boldsymbol{x})]$ is large. In principle, one can mitigate the effect of overfitting by choosing a suitable regularizer h. However, it is not easy to rigorously justify how a particular regularizer incorporates the information of the ground-truth distribution $\mathbb{P}^*$. Instead, we consider a DRO approach, in which we replace the *empirical risk* $\mathbb{E}_{\boldsymbol{x}\sim\widehat{\mathbb{P}}_n}[R(\boldsymbol{L}, \boldsymbol{x})]$ in (2) by the *worst-case expected risk*

$$\sup_{\mathbb{Q}\in\mathcal{D}(\widehat{\mathbb{P}}_n)} \mathbb{E}_{\boldsymbol{x}\sim\mathbb{Q}}[R(\boldsymbol{L}, \boldsymbol{x})],$$

where $\mathcal{D}(\widehat{\mathbb{P}}_n)$, the so-called *ambiguity set*, is a set of probability distributions that are "close to" $\widehat{\mathbb{P}}_n$. Intuitively, if the set $\mathcal{D}(\widehat{\mathbb{P}}_n)$ is small and contains $\mathbb{P}^*$, then the solution $\widetilde{\boldsymbol{L}}$ obtained

by minimizing the worst-case expected risk over $\mathcal{L}_s$ will not be too conservative (i.e., it has a small worst-case expected risk) and will be less sensitive to the unseen signals (as $\widetilde{\boldsymbol{L}}$ would have taken the effect of $\mathbb{P}^*$ into account). To construct an ambiguity set with these desiderata, let us make the following simple yet crucial observation:

Proposition 1. *Let $\mathbb{Q}$ be a probability distribution on the Borel σ-algebra in $\mathbb{R}^m$ with mean $\boldsymbol{\mu} \in \mathbb{R}^m$ and covariance $\boldsymbol{\Sigma} \in \mathbb{S}_+^m$; i.e., $\mathbb{E}_{\boldsymbol{x}\sim\mathbb{Q}}[\boldsymbol{x}] = \boldsymbol{\mu}$ and $\mathbb{E}_{\boldsymbol{x}\sim\mathbb{Q}}[(\boldsymbol{x}-\boldsymbol{\mu})(\boldsymbol{x}-\boldsymbol{\mu})^\top] = \boldsymbol{\Sigma}$. Then,*

$$\mathbb{E}_{\boldsymbol{x}\sim\mathbb{Q}}[R(\boldsymbol{L},\boldsymbol{x})] = \operatorname{tr}(\boldsymbol{\Sigma}\boldsymbol{L}) + \boldsymbol{\mu}^\top \boldsymbol{L}\boldsymbol{\mu}.$$

Proof. The result follows from a simple computation:

$$\begin{aligned}
&\mathbb{E}_{\boldsymbol{x}\sim\mathbb{Q}}[R(\boldsymbol{L},\boldsymbol{x})] = \mathbb{E}_{\boldsymbol{x}\sim\mathbb{Q}}[\operatorname{tr}(\boldsymbol{x}\boldsymbol{x}^\top \boldsymbol{L})] \\
&= \mathbb{E}_{\boldsymbol{x}\sim\mathbb{Q}}\left[\operatorname{tr}\left((\boldsymbol{x}-\boldsymbol{\mu})(\boldsymbol{x}-\boldsymbol{\mu})^\top \boldsymbol{L} + 2\boldsymbol{x}\boldsymbol{\mu}^\top \boldsymbol{L} - \boldsymbol{\mu}\boldsymbol{\mu}^\top \boldsymbol{L}\right)\right] \\
&= \operatorname{tr}(\boldsymbol{\Sigma}\boldsymbol{L}) + \boldsymbol{\mu}^\top \boldsymbol{L}\boldsymbol{\mu}.
\end{aligned}$$

□

Proposition 1 shows that the expected risk $\mathbb{E}_{\boldsymbol{x}\sim\mathbb{Q}}[R(\boldsymbol{L},\boldsymbol{x})]$ depends only on the mean and covariance of $\mathbb{Q}$. In particular, we have

$$\mathbb{E}_{\boldsymbol{x}\sim\widehat{\mathbb{P}}_n}[R(\boldsymbol{L},\boldsymbol{x})] = \operatorname{tr}(\widehat{\boldsymbol{\Sigma}}_n\boldsymbol{L}) + \widehat{\boldsymbol{\mu}}_n^\top \boldsymbol{L}\widehat{\boldsymbol{\mu}}_n, \tag{3}$$

where

$$\widehat{\boldsymbol{\mu}}_n \coloneqq \frac{1}{n}\sum_{j=1}^n \boldsymbol{x}^j, \;\; \widehat{\boldsymbol{\Sigma}}_n \coloneqq \frac{1}{n}\sum_{j=1}^n (\boldsymbol{x}^j - \widehat{\boldsymbol{\mu}}_n)(\boldsymbol{x}^j - \widehat{\boldsymbol{\mu}}_n)^\top \tag{4}$$

are the empirical mean and empirical covariance of the observed signals, respectively. By the law of large numbers, we expect that as $n \to \infty$, the empirical mean $\widehat{\boldsymbol{\mu}}_n$ and the mean $\boldsymbol{\mu}^*$ of $\mathbb{P}^*$ will be close to each other, and the same is true for the empirical covariance $\widehat{\boldsymbol{\Sigma}}_n$ and the covariance $\boldsymbol{\Sigma}^*$ of $\mathbb{P}^*$. This suggests we should consider an ambiguity set that contains distributions whose mean vectors and covariance matrices are close to $\widehat{\boldsymbol{\mu}}_n$ and $\widehat{\boldsymbol{\Sigma}}_n$, respectively.

Concretely, let $\mathscr{P}(\boldsymbol{\mu},\boldsymbol{\Sigma})$ denote the set of probability distributions on the Borel σ-algebra in $\mathbb{R}^m$ with mean $\boldsymbol{\mu} \in \mathbb{R}^m$ and covariance $\boldsymbol{\Sigma} \in \mathbb{S}_+^m$. Given a Laplacian $\boldsymbol{L} \in \mathcal{L}_s$ and parameters $\rho_1, \rho_2 > 0$, we define the following ambiguity set:

$$\mathcal{M}(\boldsymbol{L},\rho_1,\rho_2) \coloneqq \left\{ \mathbb{Q} \in \mathscr{P}(\boldsymbol{\mu},\boldsymbol{\Sigma}) : \begin{array}{c} (\boldsymbol{\mu}-\widehat{\boldsymbol{\mu}}_n)^\top \boldsymbol{L}(\boldsymbol{\mu}-\widehat{\boldsymbol{\mu}}_n) \le \rho_1^2, \\ \|\boldsymbol{\Sigma} - \widehat{\boldsymbol{\Sigma}}_n\|_F \le \rho_2, \\ \boldsymbol{\mu} \in \mathbb{R}^m,\ \boldsymbol{\Sigma} \in \mathbb{S}_+^m \end{array} \right\}.$$

In other words, every distribution in $\mathcal{M}(\boldsymbol{L},\rho_1,\rho_2)$ has its mean lying in the ellipsoid $\mathcal{E}(\widehat{\boldsymbol{\mu}}_n,\boldsymbol{L},\rho_1) \coloneqq \{\boldsymbol{\mu} \in \mathbb{R}^m : (\boldsymbol{\mu}-\widehat{\boldsymbol{\mu}}_n)^\top \boldsymbol{L}(\boldsymbol{\mu}-\widehat{\boldsymbol{\mu}}_n) \le \rho_1^2\}$ and its covariance lying in the ball $\mathcal{B}(\widehat{\boldsymbol{\Sigma}}_n,\rho_2) \coloneqq \{\boldsymbol{\Sigma} \in \mathbb{S}^m : \|\boldsymbol{\Sigma}-\widehat{\boldsymbol{\Sigma}}_n\|_F \le \rho_2\}$. On one hand, the use of a Frobenius-norm ball to describe a neighborhood of the empirical covariance $\widehat{\boldsymbol{\Sigma}}_n$ is rather intuitive. On the other hand, the use of an ellipsoid defined by $\boldsymbol{L}$ to describe a neighborhood of the empirical mean $\widehat{\boldsymbol{\mu}}_n$ is motivated by the factor analysis model proposed in [7] for smooth graph signals. Indeed, suppose that the ground-truth Laplacian $\boldsymbol{L}^*$ admits the eigendecomposition $\boldsymbol{L}^* = \boldsymbol{\chi}\boldsymbol{\Lambda}\boldsymbol{\chi}^\top$, where $\boldsymbol{\Lambda} = \operatorname{Diag}(\lambda_1,\ldots,\lambda_m)$ is diagonal and $\boldsymbol{\chi} = [\boldsymbol{u}^1 \;\cdots\; \boldsymbol{u}^m]$ is orthogonal. The factor analysis model in [7] assumes that the graph signal $\boldsymbol{x}$ is generated as

$$\boldsymbol{x} = \boldsymbol{\chi}\boldsymbol{r} + \boldsymbol{\mu}^* + \boldsymbol{\delta}, \tag{5}$$

where $\boldsymbol{r} \sim \mathcal{N}(\boldsymbol{0},\boldsymbol{\Lambda}^\dagger)$ is the latent variable that controls the graph signal $\boldsymbol{x}$, $\boldsymbol{\mu}^* \in \mathbb{R}^m$ is the mean of $\boldsymbol{x}$, and $\boldsymbol{\delta} \sim \mathcal{N}(\boldsymbol{0},\epsilon^2\boldsymbol{I}_m)$ is the noise with power $\epsilon^2 > 0$. Under this model, we have $\boldsymbol{x} \sim \mathbb{P}^* = \mathcal{N}(\boldsymbol{\mu}^*, \boldsymbol{L}^{*\dagger} + \epsilon^2\boldsymbol{I}_m)$; see [7]. As such, it includes another widely-studied graph signal model in the literature, namely the GMRF model with a graph Laplacian precision matrix, as special case (see, e.g., [17] and the references therein). Since the observed signals $\{\boldsymbol{x}^j\}_{j=1}^n$ are assumed to be independent realizations of the Gaussian random vector $\boldsymbol{x}$, we see that $\widehat{\boldsymbol{\mu}}_n = \frac{1}{n}\sum_{j=1}^n \boldsymbol{x}^j$ is also a Gaussian random vector with

$$\begin{aligned}
&\mathbb{E}_{\boldsymbol{x}^1,\ldots,\boldsymbol{x}^n\sim\mathbb{P}^*}[\widehat{\boldsymbol{\mu}}_n] = \frac{1}{n}\cdot n \cdot \mathbb{E}_{\boldsymbol{x}\sim\mathbb{P}^*}[\boldsymbol{x}] = \boldsymbol{\mu}^*, \\
&\operatorname{Cov}_{\boldsymbol{x}^1,\ldots,\boldsymbol{x}^n\sim\mathbb{P}^*}(\widehat{\boldsymbol{\mu}}_n) = \frac{1}{n^2}\sum_{j=1}^n \operatorname{Cov}_{\boldsymbol{x}^j\sim\mathbb{P}^*}(\boldsymbol{x}^j) = \frac{1}{n}(\boldsymbol{L}^{*\dagger} + \epsilon^2\boldsymbol{I}_m),
\end{aligned}$$

so that $\widehat{\boldsymbol{\mu}}_n \sim \mathcal{N}\left(\boldsymbol{\mu}^*, \frac{1}{n}(\boldsymbol{L}^{*\dagger} + \epsilon^2\boldsymbol{I}_m)\right)$. Now, observe that for $i = 1,\ldots,m$, the projection ${\boldsymbol{u}^i}^\top(\widehat{\boldsymbol{\mu}}_n - \boldsymbol{\mu}^*)$ of the deviation $\widehat{\boldsymbol{\mu}}_n - \boldsymbol{\mu}^*$ onto the i-th eigenbasis $\boldsymbol{u}^i$ of $\boldsymbol{L}^*$ is a mean-zero Gaussian random variable with variance

$$\begin{aligned}
\sigma_i^2 &= \mathbb{E}\left[\left({\boldsymbol{u}^i}^\top(\widehat{\boldsymbol{\mu}}_n - \boldsymbol{\mu}^*)\right)^2\right] \\
&= \mathbb{E}\left[\operatorname{tr}\left((\widehat{\boldsymbol{\mu}}_n - \boldsymbol{\mu}^*)(\widehat{\boldsymbol{\mu}}_n - \boldsymbol{\mu}^*)^\top \boldsymbol{u}^i {\boldsymbol{u}^i}^\top\right)\right] \\
&= \frac{1}{n}\operatorname{tr}\left(\left(\boldsymbol{L}^{*\dagger} + \epsilon^2\boldsymbol{I}_m\right)\boldsymbol{u}^i{\boldsymbol{u}^i}^\top\right) = \frac{1}{n}(\nu_i + \epsilon^2)\|\boldsymbol{u}^i\|_2^2 \\
&= \frac{1}{n}(\nu_i + \epsilon^2),
\end{aligned}$$

where $\nu_i = \frac{1}{\lambda_i}$ if $\lambda_i > 0$ and $\nu_i = 0$ if $\lambda_i = 0$. In particular, if $\lambda_i > 0$, then the larger the λ_i, the smaller the σ_i^2 and thus the more concentrated ${\boldsymbol{u}^i}^\top(\widehat{\boldsymbol{\mu}}_n - \boldsymbol{\mu}^*)$ is around 0. Consequently, we expect that

$$(\widehat{\boldsymbol{\mu}}_n - \boldsymbol{\mu}^*)^\top \boldsymbol{L}^*(\widehat{\boldsymbol{\mu}}_n - \boldsymbol{\mu}^*) = \sum_{i=1}^m \lambda_i \left({\boldsymbol{u}^i}^\top(\widehat{\boldsymbol{\mu}}_n - \boldsymbol{\mu}^*)\right)^2$$

will be small with high probability (in fact, the above argument can be not only made rigorous but also extended to more general ground-truth distributions $\mathbb{P}^*$; see Theorem 1). This shows that given a Laplacian $\boldsymbol{L} \in \mathcal{L}_s$, we only need to take into account those distributions $\mathbb{Q}$ whose mean vectors $\boldsymbol{\mu}$ are close to the empirical mean $\widehat{\boldsymbol{\mu}}_n$ under the covariance structure induced by $\boldsymbol{L}$. Such a consideration gives rise to the ellipsoidal constraint $(\boldsymbol{\mu} - \widehat{\boldsymbol{\mu}}_n)^\top \boldsymbol{L}(\boldsymbol{\mu} - \widehat{\boldsymbol{\mu}}_n) \le \rho_1^2$ in the definition of the ambiguity set $\mathcal{M}(\boldsymbol{L},\rho_1,\rho_2)$.

Based on the above development, we propose the following distributionally robust counterpart of the ERM problem (2), which robustifies the learned graph against uncertainties about the ground-truth distribution $\mathbb{P}^*$:

$$\inf_{\boldsymbol{L}\in\mathcal{L}_s} \left\{ \sup_{\mathbb{Q}\in\mathcal{M}(\boldsymbol{L},\rho_1,\rho_2)} \mathbb{E}_{\boldsymbol{x}\sim\mathbb{Q}}[R(\boldsymbol{L},\boldsymbol{x})] + h(\boldsymbol{L}) \right\}. \tag{6}$$

Since we use the first moment $\widehat{\boldsymbol{\mu}}_n$ and second moment $\widehat{\boldsymbol{\Sigma}}_n$ to define the ambiguity set $\mathcal{M}(\boldsymbol{L}, \rho_1, \rho_2)$, we shall refer to (6) as the *Moment-Uncertain Graph Learning* (MUGL) model. Note that we keep the regularizer h in the model, as it offers a way to induce structures beyond those that provide distributional robustness in the learned graph. A notable feature of the formulation (6) is that the ambiguity set $\mathcal{M}(\boldsymbol{L}, \rho_1, \rho_2)$ *depends* on the decision variable $\boldsymbol{L} \in \mathcal{L}_s$. This ensures that the distributions in $\mathcal{M}(\boldsymbol{L}, \rho_1, \rho_2)$ reflect the graph structure encoded in $\boldsymbol{L}$. However, such a feature leads to reformulation and computational challenges that have not been addressed in the DRO literature before. As such, we need to develop new machinery to tackle our proposed MUGL model (6).

B. Bound on Out-of-Sample Risk

Recall from our earlier discussion that it is desirable for an ambiguity set to be small and contain the ground-truth distribution $\mathbb{P}^*$. We now show that the ambiguity set $\mathcal{M}(\boldsymbol{L}, \rho_1, \rho_2)$ will indeed possess such properties with high probability if the ground-truth distribution $\mathbb{P}^*$ satisfies certain moment growth condition and we choose the parameters ρ_1, ρ_2 judiciously. To begin, let us introduce the following definition:

Definition 1 (Moment growth condition; cf. [18])**.** *A probability distribution $\mathbb{Q}$ on the Borel σ-algebra in $\mathbb{R}^m$ with mean $\boldsymbol{\mu} \in \mathbb{R}^m$ is said to satisfy the* moment growth condition *if there exists a constant $c > 0$ such that for all $p \geq 1$,*

$$\mathbb{E}_{\boldsymbol{x}\sim\mathbb{Q}}[\|\boldsymbol{x} - \boldsymbol{\mu}\|_2^p] \leq (cp)^{p/2}.$$

The moment growth condition defined above is rather mild. For instance, it is satisfied by any sub-Gaussian distribution [19]. In the remainder of this subsection, we make the following assumption:

Assumption 1. *The ground-truth distribution $\mathbb{P}^*$ of the graph signal $\boldsymbol{x} \in \mathbb{R}^m$ has mean $\boldsymbol{\mu}^* \in \mathbb{R}^m$, covariance $\boldsymbol{\Sigma}^* \in \mathbb{S}_+^m$ and satisfies the moment growth condition. Let $\{\boldsymbol{x}^j\}_{j=1}^n$ be n independent realizations of $\boldsymbol{x}$ and $\widehat{\boldsymbol{\mu}}_n, \widehat{\boldsymbol{\Sigma}}_n$ be given by* (4)*.*

Now, using the probabilistic techniques developed in [18], we can establish the following confidence region for the mean $\boldsymbol{\mu}^*$ of the ground-truth distribution $\mathbb{P}^*$:

Theorem 1 (Confidence region for the mean)**.** *Suppose that Assumption 1 holds. Let $\xi \in (0, e^{-2})$ be the confidence level, where $e = 2.71828\ldots$ is Euler's number. Then, there exists a constant $c_0 > 0$ such that for any $\boldsymbol{L} \in \mathcal{L}_s$,*

$$(\widehat{\boldsymbol{\mu}}_n - \boldsymbol{\mu}^*)^\top \boldsymbol{L}(\widehat{\boldsymbol{\mu}}_n - \boldsymbol{\mu}^*) \leq \widehat{\rho}_1^2 := \frac{4c_0 e^2 \ln^2(1/\xi)}{n}$$

will hold with probability at least $1 - \xi$.

Proof. Since $\mathcal{L}_s$ is compact, we have $\max_{\boldsymbol{L}\in\mathcal{L}_s} \|\boldsymbol{L}\| < +\infty$. This, together with the assumption that $\mathbb{P}^*$ satisfies the moment growth condition, implies the existence of a constant $c_0 > 0$ such that for all $p \geq 1$,

$$\max_{\boldsymbol{L}\in\mathcal{L}_s} \left\{\mathbb{E}_{\boldsymbol{x}\sim\mathbb{P}^*}\left[\|\boldsymbol{L}^{1/2}(\boldsymbol{x} - \boldsymbol{\mu})\|_2^p\right]\right\} \leq (c_0 p)^{p/2}.$$

The desired result then follows by adapting the proof of [18, Proposition 4]. □

Moreover, we can establish the following confidence region for the covariance $\boldsymbol{\Sigma}^*$ of the ground-truth distribution $\mathbb{P}^*$:

Theorem 2 (Confidence region for the covariance)**.** *Suppose that Assumption 1 holds and $\boldsymbol{\Sigma}^* \in \mathbb{S}_{++}^m$. Let $\xi \in (0, e^{-2})$ be the confidence level, where, as before, $e = 2.71828\ldots$ is Euler's number. Then, there exist constants $c_1, c_2 > 0$ such that*

$$\begin{aligned}&\|\widehat{\boldsymbol{\Sigma}}_n - \boldsymbol{\Sigma}^*\|_F \leq \widehat{\rho}_2\\ &:= \frac{4c_1(2e/3)^{3/2}\ln^{3/2}(2m^{3/2}/\xi)}{n^{1/2}}\|\boldsymbol{\Sigma}^*\| + \frac{4c_2 e^2 \ln^2(2/\xi)}{n}\end{aligned}$$

will hold with probability at least $1 - \xi$.

The proof of the above result can be found in Appendix A.

Theorems 1 and 2 imply that given a confidence level $\xi \in (0, e^{-2})$ and a Laplacian $\boldsymbol{L} \in \mathcal{L}_s$, the ambiguity set $\mathcal{M}(\boldsymbol{L}, \widehat{\rho}_1, \widehat{\rho}_2)$ will contain the ground-truth distribution $\mathbb{P}^*$ with probability at least $1 - 2\xi$. Consequently, the out-of-sample risk bound

$$\mathbb{E}_{\boldsymbol{x}\sim\mathbb{P}^*}[R(\boldsymbol{L}, \boldsymbol{x})] \leq \sup_{\mathbb{Q}\in\mathcal{M}(\boldsymbol{L},\widehat{\rho}_1,\widehat{\rho}_2)} \mathbb{E}_{\boldsymbol{x}\sim\mathbb{Q}}[R(\boldsymbol{L}, \boldsymbol{x})] \tag{7}$$

will also hold with probability at least $1 - 2\xi$. In particular, suppose that $\widehat{\boldsymbol{L}}_n^{\text{MUGL}}$ is an optimal solution to Problem (6). Then, whenever the bound (7) holds, we have

$$\mathbb{E}_{\boldsymbol{x}\sim\mathbb{P}^*}\left[R(\widehat{\boldsymbol{L}}_n^{\text{MUGL}}, \boldsymbol{x})\right] \leq \sup_{\mathbb{Q}\in\mathcal{M}(\boldsymbol{L},\widehat{\rho}_1,\widehat{\rho}_2)} \mathbb{E}_{\boldsymbol{x}\sim\mathbb{Q}}\left[R(\widehat{\boldsymbol{L}}_n^{\text{MUGL}}, \boldsymbol{x})\right],$$

which indicates that the out-of-sample risk achieved by $\widehat{\boldsymbol{L}}_n^{\text{MUGL}}$ is lower than the optimal worst-case expected risk.

C. Reformulation of the MUGL Model

Since $\mathcal{M}(\boldsymbol{L}, \rho_1, \rho_2)$ is a set of probability distributions, it may seem at first sight that the inner supremum in the MUGL model (6) is an infinite-dimensional optimization problem. Nevertheless, using Proposition 1 and the definition of $\mathcal{M}(\boldsymbol{L}, \rho_1, \rho_2)$, we can reformulate (6) as the following finite-dimensional optimization problem:

$$\begin{array}{ll}\inf\limits_{\boldsymbol{L}} & \left\{\sup\limits_{\boldsymbol{\mu},\boldsymbol{\Sigma}} \left\{\mathrm{tr}(\boldsymbol{\Sigma}\boldsymbol{L}) + \boldsymbol{\mu}^\top\boldsymbol{L}\boldsymbol{\mu}\right\} + h(\boldsymbol{L})\right\}\\ \text{s.t.} & (\boldsymbol{\mu} - \widehat{\boldsymbol{\mu}}_n)^\top\boldsymbol{L}(\boldsymbol{\mu} - \widehat{\boldsymbol{\mu}}_n) \leq \rho_1^2,\\ & \|\boldsymbol{\Sigma} - \widehat{\boldsymbol{\Sigma}}_n\|_F \leq \rho_2,\\ & \boldsymbol{L} \in \mathcal{L}_s,\ \boldsymbol{\mu} \in \mathbb{R}^m,\ \boldsymbol{\Sigma} \in \mathbb{S}_+^m.\end{array} \tag{8}$$

Observe that for any given $\boldsymbol{L} \in \mathcal{L}_s$, the inner supremum in (8) is separable in the variables $\boldsymbol{\mu} \in \mathbb{R}^m$ and $\boldsymbol{\Sigma} \in \mathbb{S}_+^m$. Hence, we can express Problem (8) as

$$\inf_{\boldsymbol{L}\in\mathcal{L}_s} \{\varphi_1(\boldsymbol{L}) + \varphi_2(\boldsymbol{L}) + h(\boldsymbol{L})\},$$

where

$$\begin{array}{rl}\varphi_1(\boldsymbol{L}) := & \sup\limits_{\boldsymbol{\mu}\in\mathbb{R}^m} \boldsymbol{\mu}^\top\boldsymbol{L}\boldsymbol{\mu}\\ & \text{s.t. } (\boldsymbol{\mu} - \widehat{\boldsymbol{\mu}}_n)^\top\boldsymbol{L}(\boldsymbol{\mu} - \widehat{\boldsymbol{\mu}}_n) \leq \rho_1^2\end{array} \tag{9}$$

and

$$\varphi_2(\boldsymbol{L}) \coloneqq \sup_{\boldsymbol{\Sigma}\in\mathbb{S}_+^m} \operatorname{tr}(\boldsymbol{\Sigma}\boldsymbol{L}) \quad \text{s.t.} \quad \|\boldsymbol{\Sigma}-\widehat{\boldsymbol{\Sigma}}_n\|_F \leq \rho_2. \tag{10}$$

As it turns out, both optimal value functions φ_1 and φ_2 have closed-form expressions.

Proposition 2. *For any given $\boldsymbol{L}\in\mathcal{L}_s$, the optimal values of Problems* (9) *and* (10) *are given by*

$$\varphi_1(\boldsymbol{L}) = \left(\|\boldsymbol{L}^{1/2}\widehat{\boldsymbol{\mu}}_n\|_2+\rho_1\right)^2,$$
$$\varphi_2(\boldsymbol{L}) = \operatorname{tr}(\widehat{\boldsymbol{\Sigma}}_n\boldsymbol{L})+\rho_2\|\boldsymbol{L}\|_F,$$

respectively.

The proof of Proposition 2 can be found in Appendix B.

Using Proposition 2 and identity (3), we obtain the following reformulation of the MUGL model (6):

$$\inf_{\boldsymbol{L}\in\mathcal{L}_s}\left\{\frac{1}{n}\operatorname{tr}(\boldsymbol{X}^\top\boldsymbol{L}\boldsymbol{X})+2\rho_1\|\boldsymbol{L}^{\frac{1}{2}}\widehat{\boldsymbol{\mu}}_n\|_2+\rho_2\|\boldsymbol{L}\|_F+h(\boldsymbol{L})\right\}. \tag{11}$$

Compared with the non-robust graph learning model (1), the distributionally robust MUGL model (11) has two additional regularizers $h_1(\boldsymbol{L}) \coloneqq 2\|\boldsymbol{L}^{1/2}\widehat{\boldsymbol{\mu}}_n\|_2$ and $h_2(\boldsymbol{L}) \coloneqq \|\boldsymbol{L}\|_F$. The regularizer h_1 (resp. h_2) can be understood as promoting robustness of the learned graph against uncertainty about the mean (resp. covariance) of the ground-truth distribution $\mathbb{P}^*$, with the size ρ_1 (resp. ρ_2) of the uncertainty region around the empirical mean $\widehat{\boldsymbol{\mu}}_n$ (resp. empirical covariance $\widehat{\boldsymbol{\Sigma}}_n$) serving as the regularization parameter. Although it is known that various distributionally robust risk minimization problems with ϕ-divergence-based or Wasserstein distance-based ambiguity sets can be reformulated as regularized ERM problems [20], [21], the reformulation of the distributionally robust MUGL model (6) as the regularized ERM problem (11) does not follow from existing results and is new. Moreover, our development above suggests that the (non-squared) Frobenius norm regularizer $\boldsymbol{L}\mapsto h_2(\boldsymbol{L}) = \|\boldsymbol{L}\|_F$ is more interpretable than the commonly used squared Frobenius norm regularizer $\boldsymbol{L}\mapsto\|\boldsymbol{L}\|_F^2$. Indeed, even though it is often argued that the latter is used to control the sparsity in the learned graph (see, e.g., [2], [7], [11]), such an argument has not been rigorously justified. In addition, as pointed out in [8], it is not easy to interpret the squared Frobenius norm regularizer, as the elements of $\boldsymbol{L}$ are not only of different scales but also linearly dependent.

Since the scale parameter s appears in the MUGL model (11), a natural question is how it affects the learned Laplacian. To address this, let us consider the following parametrized version of the objective function of Problem (11):

$$f_{(\rho_1,\rho_2,\rho_3)}(\boldsymbol{L})\coloneqq\frac{1}{n}\operatorname{tr}(\boldsymbol{X}^\top\boldsymbol{L}\boldsymbol{X})+2\rho_1\|\boldsymbol{L}^{\frac{1}{2}}\widehat{\boldsymbol{\mu}}_n\|_2+\rho_2\|\boldsymbol{L}\|_F+\rho_3 h(\boldsymbol{L}).$$

Clearly, the objective function of Problem (11) corresponds to $f_{(\rho_1,\rho_2,1)}$, while the quantities ρ_1,ρ_2,ρ_3 correspond to the regularization parameters associated with the regularizers h_1,h_2,h, respectively. Now, we have the following proposition:

Proposition 3. *Suppose that for any $\kappa>0$, the regularizer h satisfies*

$$h(\kappa\boldsymbol{L}) = h(\boldsymbol{L})+C_\kappa \tag{12}$$

for some constant $C_\kappa\in\mathbb{R}$. Suppose further that for some $\bar{s},\bar{\rho}_1,\bar{\rho}_2>0$, the problem

$$\inf_{\boldsymbol{L}\in\mathcal{L}_{\bar{s}}} f_{(\bar{\rho}_1,\bar{\rho}_2,1)}(\boldsymbol{L})$$

has an optimal solution $\boldsymbol{L}^*$. Then, $\frac{1}{\bar{s}}\boldsymbol{L}^*$ is an optimal solution to the problem

$$\inf_{\boldsymbol{L}\in\mathcal{L}_1} f_{(\bar{\rho}_1/\sqrt{\bar{s}},\bar{\rho}_2,1/\bar{s})}(\boldsymbol{L}).$$

Proposition 3 shows that if the regularizer h satisfies condition (12), then any Laplacian learned by the MUGL model (11) with scale parameter $\bar{s}$ and regularization parameters $(\bar{\rho}_1,\bar{\rho}_2,1)$ is a multiple of certain Laplacian learned by the same model with scale parameter 1 and regularization parameters $\left(\frac{\bar{\rho}_1}{\sqrt{\bar{s}}},\bar{\rho}_2,\frac{1}{\bar{s}}\right)$. In particular, these Laplacians yield graphs with the same set of edges. Hence, we do not need to tune the parameter s and can simply set it to any positive value.

Now, let $s>0$ be arbitrary. It can be readily seen that the zero regularizer $\mathcal{L}_s\ni\boldsymbol{L}\mapsto h(\boldsymbol{L})=0$ satisfies condition (12). Moreover, the logarithmic barrier regularizer $\mathcal{L}_s\ni\boldsymbol{L}\mapsto h_{\log}(\boldsymbol{L}) \coloneqq -\alpha\sum_{i=1}^m\ln(L_{ii})\in\mathbb{R}$ with parameter $\alpha>0$, which is introduced in [8] to improve the overall connectivity of the learned graph, also satisfies condition (12). In what follows, we refer to the MUGL models that arise from these two choices of h as MUGL-*o* and MUGL-*l*, respectively.

III. Solving the MUGL Model

Now, let us turn to the algorithmic aspects of the MUGL model (11). Observe that since $\boldsymbol{L}\mapsto\|\boldsymbol{L}^{1/2}\widehat{\boldsymbol{\mu}}_n\|_2=\sqrt{\widehat{\boldsymbol{\mu}}_n^\top\boldsymbol{L}\widehat{\boldsymbol{\mu}}_n}$ is non-convex and even non-Lipschitz at any $\boldsymbol{L}$ satisfying $\boldsymbol{L}^{1/2}\widehat{\boldsymbol{\mu}}_n=\boldsymbol{0}$, the MUGL model (11) gives rise to a challenging non-smooth non-convex optimization problem. Our goal in this section is to develop a PGD method that can efficiently tackle Problem (11) and establish its convergence guarantee.

A. Vectorized MUGL Formulation

Since every Laplacian $\boldsymbol{L}\in\mathcal{L}_s$ is symmetric and satisfies $\boldsymbol{L}\boldsymbol{1}=\boldsymbol{0}$, it can be completely specified by, say, the entries below the main diagonal. This motivates us to vectorize Problem (11) to get a more compact formulation. Specifically, let $\bar{m}\coloneqq\frac{m(m-1)}{2}$ be the number of entries below the main diagonal of an $m\times m$ matrix and define the linear operator $\mathcal{F}:\mathbb{R}^{\bar{m}}\to\mathbb{S}^m$ that maps a vector to a Laplacian matrix [22], which is given by

$$[\mathcal{F}(\boldsymbol{w})]_{ij}\coloneqq\begin{cases}-w_{i-j+\frac{j-1}{2}(2m-j)}, & \text{if } i>j,\\ [\mathcal{F}(\boldsymbol{w})]_{ji}, & \text{if } i<j,\\ -\sum_{k\neq i}[\mathcal{F}(\boldsymbol{w})]_{ik}, & \text{if } i=j.\end{cases}$$

More explicitly, given a vector $\boldsymbol{w} \in \mathbb{R}^{\bar{m}}$, the entries below the main diagonal of the matrix $\mathcal{F}(\boldsymbol{w})$ are given by

$$\begin{bmatrix} * & * & * & \cdots & * \\ -w_1 & * & * & \cdots & * \\ -w_2 & -w_m & * & \cdots & * \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ -w_{m-1} & -w_{2m-3} & -w_{3m-6} & \cdots & * \end{bmatrix}.$$

Furthermore, define

$$\Delta_s \coloneqq \left\{\boldsymbol{w} \in \mathbb{R}^{\bar{m}} : \mathbf{1}^\top \boldsymbol{w} = s,\ \boldsymbol{w} \geq \mathbf{0}\right\}.$$

It is not hard to show that $\boldsymbol{w} \in \Delta_s$ if and only if $\mathcal{F}(\boldsymbol{w}) \in \mathcal{L}_s$. Now, let $\mathcal{F}^* : \mathbb{S}^m \to \mathbb{R}^{\bar{m}}$ be the adjoint operator of $\mathcal{F}$; i.e., $\mathcal{F}^*$ satisfies $\operatorname{tr}(\mathcal{F}(\boldsymbol{w})\boldsymbol{M}) = \boldsymbol{w}^\top \mathcal{F}^*(\boldsymbol{M})$ for all $\boldsymbol{w} \in \mathbb{R}^{\bar{m}}$ and $\boldsymbol{M} \in \mathbb{S}^m$. It can be verified that for $1 \leq j < i \leq m$,

$$[\mathcal{F}^*(\boldsymbol{M})]_{i-j+\frac{j-1}{2}(2m-j)} = M_{ii} - M_{ij} - M_{ji} + M_{jj}. \quad (13)$$

With the above preparations, we deduce that for any $\boldsymbol{L} \in \mathcal{L}_s$, we can find a $\boldsymbol{w} \in \Delta_s$ such that

$$\operatorname{tr}(\boldsymbol{X}^\top \boldsymbol{L} \boldsymbol{X}) = \operatorname{tr}(\mathcal{F}(\boldsymbol{w})\boldsymbol{X}\boldsymbol{X}^\top) = \boldsymbol{w}^\top \mathcal{F}^*\left(\widehat{\boldsymbol{\Sigma}}_n + \widehat{\boldsymbol{\mu}}_n \widehat{\boldsymbol{\mu}}_n^\top\right),$$
$$\|\boldsymbol{L}^{1/2}\widehat{\boldsymbol{\mu}}_n\|_2 = \sqrt{\operatorname{tr}\left(\mathcal{F}(\boldsymbol{w})\widehat{\boldsymbol{\mu}}_n\widehat{\boldsymbol{\mu}}_n^\top\right)} = \sqrt{\boldsymbol{w}^\top \mathcal{F}^*(\widehat{\boldsymbol{\mu}}_n\widehat{\boldsymbol{\mu}}_n^\top)}.$$

This, together with the fact that $\mathcal{F}(\boldsymbol{w}) \in \mathcal{L}_s$ for any $\boldsymbol{w} \in \Delta_s$, implies that Problem (11) admits the vectorized formulation

$$\inf_{\boldsymbol{w} \in \Delta_s} g(\boldsymbol{w}) \coloneqq \phi_1(\boldsymbol{w}) + \phi_2(\boldsymbol{w}) + h(\mathcal{F}(\boldsymbol{w})), \quad (14)$$

where

$$\phi_1(\boldsymbol{w}) \coloneqq \frac{1}{n}\boldsymbol{w}^\top \mathcal{F}^*\left(\widehat{\boldsymbol{\Sigma}}_n + \widehat{\boldsymbol{\mu}}_n\widehat{\boldsymbol{\mu}}_n^\top\right) + \rho_2\|\mathcal{F}(\boldsymbol{w})\|_F,$$
$$\phi_2(\boldsymbol{w}) \coloneqq \sqrt{\boldsymbol{a}^\top \boldsymbol{w}}, \quad \boldsymbol{a} \coloneqq 4\rho_1^2 \mathcal{F}^*\left(\widehat{\boldsymbol{\mu}}_n\widehat{\boldsymbol{\mu}}_n^\top\right).$$

To have a better understanding of the structure of Problem (14), we first observe that $\boldsymbol{w} \mapsto \|\mathcal{F}(\boldsymbol{w})\|_F$ is smooth over Δ_s. This follows since $\boldsymbol{w} \in \Delta_s$ implies that $\mathcal{F}(\boldsymbol{w}) \in \mathcal{L}_s$, which in turn implies that $\|\mathcal{F}(\boldsymbol{w})\|_F > 0$. Next, consider the function $\boldsymbol{w} \mapsto \phi_2(\boldsymbol{w}) = \sqrt{\boldsymbol{a}^\top \boldsymbol{w}}$. Using (13), we deduce that for $1 \leq j < i \leq m$,

$$\begin{aligned} &a_{i-j+\frac{j-1}{2}(2m-j)} \\ &= 4\rho_1^2\left((\widehat{\boldsymbol{\mu}}_n)_i^2 - (\widehat{\boldsymbol{\mu}}_n)_i(\widehat{\boldsymbol{\mu}}_n)_j - (\widehat{\boldsymbol{\mu}}_n)_j(\widehat{\boldsymbol{\mu}}_n)_i + (\widehat{\boldsymbol{\mu}}_n)_j^2\right) \\ &= 4\rho_1^2\left((\widehat{\boldsymbol{\mu}}_n)_i - (\widehat{\boldsymbol{\mu}}_n)_j\right)^2 \geq 0; \end{aligned} \quad (15)$$

i.e., $\boldsymbol{a} \geq \mathbf{0}$. Thus, the function ϕ_2 is well defined on Δ_s and is non-smooth at any $\boldsymbol{w} \in \mathcal{Z}_s \coloneqq \Delta_s \cap \{\boldsymbol{w} \in \mathbb{R}^{\bar{m}} : \boldsymbol{a}^\top \boldsymbol{w} = 0\}$. Now, observe that $\mathcal{Z}_s \neq \emptyset$ if and only if there exists a $k \in \{1, \ldots, \bar{m}\}$ such that $a_k = 0$, as $\boldsymbol{w} \in \Delta_s$ implies that $\boldsymbol{w} \geq \mathbf{0}$ and $\boldsymbol{w} \neq \mathbf{0}$. By (15), such an event occurs when at least two coordinates of $\widehat{\boldsymbol{\mu}}_n$ are equal. Intuitively, however, if the ground-truth distribution $\mathbb{P}^*$ of the graph signal $\boldsymbol{x} \in \mathbb{R}^m$ is continuous and $\{\boldsymbol{x}^j\}_{j=1}^n$ are n independent realizations of $\boldsymbol{x}$, then the probability that $\widehat{\boldsymbol{\mu}}_n$ has at least two equal coordinates should be zero. In other words, the function ϕ_2 should be smooth on Δ_s almost surely (i.e., with probability 1). It turns out that such an intuition is almost correct and can be made precise as follows:

Theorem 3 (Smoothness of MUGL objective)**.** *Let $\mathbb{P}^*$ be the ground-truth distribution of the graph signal $\boldsymbol{x} \in \mathbb{R}^m$ and $\{\boldsymbol{x}^j\}_{j=1}^n$ be n independent realizations of $\boldsymbol{x}$. Suppose that $\mathbb{P}^*$ is absolutely continuous wrt the m-dimensional Lebesgue measure ν (denoted by $\mathbb{P}^* \ll \nu$); i.e., for any measurable set $\mathcal{A} \subseteq \mathbb{R}^m$, $\mathbb{P}^*(\mathcal{A}) = 0$ whenever $\nu(\mathcal{A}) = 0$. Then, the event $\mathcal{Z}_s = \emptyset$ will occur almost surely. Consequently, the function ϕ_2 will be smooth on Δ_s almost surely.*

Proof. For $k = 1, \ldots, \bar{m}$, define $\mathcal{V}_k \coloneqq \{\boldsymbol{u} \in \mathbb{R}^{\bar{m}} : u_k = 0\}$. Recall that $\mathcal{Z}_s \neq \emptyset$ if and only if there exists a $k \in \{1, \ldots, \bar{m}\}$ such that $a_k = 0$. Moreover, note that $\boldsymbol{a}$ depends on $\boldsymbol{x}^1, \ldots, \boldsymbol{x}^n$, which are identical and independently distributed according to $\mathbb{P}^*$. Thus, we have

$$\Pr(\mathcal{Z}_s \neq \emptyset) = \Pr\left(\boldsymbol{a} \in \bigcup_{k=1}^{\bar{m}} \mathcal{V}_k\right) \leq \sum_{k=1}^{\bar{m}} \Pr(\boldsymbol{a} \in \mathcal{V}_k), \quad (16)$$

where the probability is evaluated wrt the product measure $\mathbb{P}^{*n}$.

Now, consider a fixed $k \in \{1, \ldots, \bar{m}\}$. Then, we can find $i, j \in \{1, \ldots, m\}$ with $1 \leq j < i \leq m$ such that $k = i - j + \frac{j-1}{2}(2m - j)$. Using (15), we deduce that

$$\Pr(\boldsymbol{a} \in \mathcal{V}_k) = \Pr(n\widehat{\boldsymbol{\mu}}_n \in \Pi_{ij}), \quad (17)$$

where $\Pi_{ij} \coloneqq \{\boldsymbol{v} \in \mathbb{R}^m : v_i = v_j\}$ is an $(m-1)$-dimensional linear subspace in $\mathbb{R}^m$. Since $n\widehat{\boldsymbol{\mu}}_n = \boldsymbol{x}^1 + \cdots + \boldsymbol{x}^n$ is the sum of n independent random vectors that are identically distributed according to $\mathbb{P}^*$, its distribution is given by the n-fold convolution of $\mathbb{P}^*$, denoted by $\mathbb{P}^{*\circledast n}$ (cf. [23, Section 20]). Moreover, since $\mathbb{P}^* \ll \nu$, we have $\mathbb{P}^{*\circledast n} \ll \nu$ (cf. [23, Exercise 31.14(b)]). This, together with the well-known fact that $\nu(\Pi_{ij}) = 0$, implies that

$$\Pr(n\widehat{\boldsymbol{\mu}}_n \in \Pi_{ij}) = \mathbb{P}^{*\circledast n}(\Pi_{ij}) = 0. \quad (18)$$

Upon noting that the above argument holds for arbitrary $k \in \{1, \ldots, \bar{m}\}$ and combining (16)–(18), we conclude that

$$\Pr(\mathcal{Z}_s \neq \emptyset) \leq \sum_{k=1}^{\bar{m}} \Pr(\boldsymbol{a} \in \mathcal{V}_k) = \sum_{1 \leq j < i \leq m} \mathbb{P}^{*\circledast n}(\Pi_{ij}) = 0.$$

This completes the proof. □

By the Radon–Nikodym theorem [23, Section 32], the probability distributions that are absolutely continuous wrt the Lebesgue measure are precisely those that have probability density functions wrt the Lebesgue measure. Thus, Theorem 3 applies to a wide range of ground-truth distributions. In what follows, we assume that $\mathbb{P}^* \ll \nu$.

B. PGD Method for MUGL and Its Convergence Analysis

Theorem 3 implies that the gradient of ϕ_2 at any $\boldsymbol{w} \in \Delta_s$ will be well defined almost surely. Thus, for any given $\boldsymbol{w} \in \Delta_s$, as long as the gradient of the regularizer h at $\mathcal{F}(\boldsymbol{w}) \in \mathcal{L}_s$ is well defined, we can compute the gradient of the objective function g of Problem (14) at $\boldsymbol{w}$ as follows:

$$\nabla g(\boldsymbol{w}) = \frac{1}{n}\mathcal{F}^*\left(\widehat{\boldsymbol{\Sigma}}_n + \widehat{\boldsymbol{\mu}}_n\widehat{\boldsymbol{\mu}}_n^\top\right) + \rho_1 \frac{\mathcal{F}^*\left(\widehat{\boldsymbol{\mu}}_n\widehat{\boldsymbol{\mu}}_n^\top\right)}{\sqrt{\boldsymbol{w}^\top \mathcal{F}^*\left(\widehat{\boldsymbol{\mu}}_n\widehat{\boldsymbol{\mu}}_n^\top\right)}} + \rho_2 \frac{\mathcal{F}^*(\mathcal{F}(\boldsymbol{w}))}{\|\mathcal{F}(\boldsymbol{w})\|_F} + \mathcal{F}^*(\nabla h(\mathcal{F}(\boldsymbol{w}))).$$

This suggests that Problem (14) can be tackled by the PGD method, whose update formula is given by

$$\boldsymbol{w}^{k+1} \longleftarrow \Pi_{\Delta_s}(\boldsymbol{w}^k - \eta_k \nabla g(\boldsymbol{w}^k)), \quad k = 0, 1, \ldots. \tag{19}$$

Here, $\eta_k > 0$ is the step size and $\Pi_{\Delta_s}(\boldsymbol{w}) := \operatorname{argmin}_{\boldsymbol{v} \in \Delta_s} \|\boldsymbol{v} - \boldsymbol{w}\|_2$ is the projection of $\boldsymbol{w}$ onto Δ_s. It is well known that the projection of a vector $\boldsymbol{w} \in \mathbb{R}^{\bar{m}}$ onto the simplex Δ_s can be computed in $\mathcal{O}(\bar{m} + \operatorname{nnz}(\Pi_{\Delta_s}(\boldsymbol{w})) \cdot \log \bar{m})$ time, where $\operatorname{nnz}(\boldsymbol{u})$ denotes the number of non-zero elements in $\boldsymbol{u}$; see, e.g., [24] and the references therein. Moreover, the gradient $\nabla g(\boldsymbol{w})$ can be computed in $\mathcal{O}(\bar{m})$ time. Thus, the update (19) can be implemented efficiently.

Since Problem (14) is non-convex in general, one does not expect that the PGD method (19) will find an optimal solution to the problem. Nevertheless, under some mild assumptions, it is possible to establish the convergence of the PGD method (19) to a stationary point of Problem (14). Recall that a point $\bar{\boldsymbol{w}} \in \mathbb{R}^{\bar{m}}$ at which the function g is continuously differentiable is a stationary point of Problem (14) if there exists a vector of dual multipliers $(\bar{d}_0, \bar{\boldsymbol{d}}) \in \mathbb{R}^{\bar{m}+1}$ such that $(\bar{\boldsymbol{w}}; (\bar{d}_0, \bar{\boldsymbol{d}}))$ satisfies the following Karush–Kuhn–Tucker (KKT) conditions:

$$\begin{aligned} &\nabla g(\bar{\boldsymbol{w}}) + \bar{d}_0 \mathbf{1} \geq \mathbf{0}, \\ &\mathbf{1}^\top \bar{\boldsymbol{w}} = s, \;\; \bar{\boldsymbol{w}} \geq \mathbf{0}, \\ &\bar{\boldsymbol{w}}^\top (\nabla g(\bar{\boldsymbol{w}}) + \bar{d}_0 \mathbf{1}) = 0. \end{aligned}$$

For notational simplicity, define $\widetilde{g} := g + \mathbb{I}_{\Delta_s}$, where $\mathbb{I}_{\Delta_s}$ is the indicator function associated with Δ_s; i.e., $\mathbb{I}_{\Delta_s}(\boldsymbol{w}) = 0$ if $\boldsymbol{w} \in \Delta_s$ and $\mathbb{I}_{\Delta_s}(\boldsymbol{w}) = +\infty$ otherwise. We can now state and prove our first convergence result.

Theorem 4 (Global convergence of PGD under Lipschitz continuous gradient)**.** *Suppose that the regularizer $h : \mathbb{S}^m \to \mathbb{R}$ is continuously differentiable and its gradient is Lipschitz continuous on $\mathcal{L}_s$. Then, the following will hold almost surely:*

(a) *The function g is continuously differentiable and its gradient is Lipschitz continuous for some parameter $\ell_g > 0$ on Δ_s.*

(b) *If the function $\widetilde{g}$ possesses the* Kurdyka-Łojasiewicz *(KŁ) property (see [25, Section 2] for the definition and a brief discussion of its significance) and the step sizes $\{\eta_k\}_{k \geq 0}$ satisfy $\eta_k \in (0, 1/\ell_g)$ for all $k \geq 0$, then for any initial point $\boldsymbol{w}^0 \in \Delta_s$, the sequence $\{\boldsymbol{w}^k\}_{k \geq 0}$ generated by the PGD method* (19) *converges to a stationary point of Problem* (14)*.*

Proof. Since $\|\mathcal{F}(\boldsymbol{w})\|_F > 0$ for all $\boldsymbol{w} \in \Delta_s$ and Δ_s is compact, it is straightforward to show that ϕ_1 is continuously differentiable and its gradient is Lipschitz continuous on Δ_s. On the other hand, we have $\mathbb{P}^* \ll \nu$ by assumption, so that Theorem 3 applies. In particular, we will have $\phi_2(\boldsymbol{w}) > 0$ for all $\boldsymbol{w} \in \Delta_s$ almost surely, which, together with the compactness of Δ_s, implies that ϕ_2 will be continuously differentiable and its gradient will be Lipschitz continuous on Δ_s almost surely. These results and the assumption on h immediately yield the result in (a). The result in (b) then follows from a direct application of [25, Theorem 5.3]. □

As discussed in [25], [26], the assumption that $\widetilde{g}$ possesses the KŁ property is a rather mild one. In particular, since Δ_s is polyhedral and both ϕ_1, ϕ_2 are actually analytic on Δ_s,[1] if h is also analytic on Δ_s (which implies that it satisfies the assumption on h in Theorem 4), then $\widetilde{g}$ possesses the KŁ property; see, e.g., the discussion in [26, Section 4.3]. An important consequence of the KŁ property of $\widetilde{g}$ is that it ensures the *convergence* and not just *subsequential convergence* of the sequence $\{\boldsymbol{w}^k\}_{k \geq 0}$ generated by the PGD method (19).

It is worth noting that h does not have to be convex in order for Theorem 4 to hold. On the other hand, if h is convex and $\rho_1 = 0$ (i.e., there is essentially no uncertainty about the mean of the ground-truth distribution), then (14) is a convex optimization problem. In this case, the KKT conditions associated with Problem (14) are necessary and sufficient for optimality. Thus, if in addition the assumptions on h and $\widetilde{g}$ in Theorem 4 hold, then the iterates generated by the PGD method (19) converge to an optimal solution to Problem (14).

The convergence result in Theorem 4 relies crucially on the Lipschitz continuity of ∇g on Δ_s. However, for certain choices of the regularizer h, the resulting function g may not have such a property. A case in point is the logarithmic barrier regularizer $\boldsymbol{L} \mapsto h_{\log}(\boldsymbol{L}) = -\alpha \sum_{i=1}^m \ln(L_{ii})$ with parameter $\alpha > 0$. Indeed, for any sequence $\{\boldsymbol{w}^k\}_{k \geq 0}$ in Δ_s such that $[\mathcal{F}(\boldsymbol{w}^k)]_{ii} \to 0$ for some $i \in \{1, \ldots, m\}$, we have $\|\nabla h_{\log}(\mathcal{F}(\boldsymbol{w}^k))\|_2 \to +\infty$. This implies that the function g cannot have a Lipschitz continuous gradient on Δ_s.

As it turns out, it is possible to ensure the convergence of the PGD method (19) under weaker smoothness assumptions on g if the step sizes $\{\eta_k\}_{k \geq 0}$ are chosen via an appropriate line search strategy. Specifically, let $0 < \eta_{\min} \leq \eta_{\max} < +\infty$ and $\beta, \gamma \in (0, 1)$ be given parameters. Given a sequence $\{\eta_k\}_{k \geq 0}$ satisfying $\eta_k \in [\eta_{\min}, \eta_{\max}]$, consider a line search-based PGD method (LS-PGD) with the following update scheme:

For $k = 0, 1, \ldots$, do the following:

1) (Projected gradient step). Compute

$$\widetilde{\boldsymbol{w}}^k \longleftarrow \Pi_{\Delta_s}(\boldsymbol{w}^k - \eta_k \nabla g(\boldsymbol{w}^k)), \tag{20a}$$

$$\boldsymbol{v}^k \longleftarrow \widetilde{\boldsymbol{w}}^k - \boldsymbol{w}^k. \tag{20b}$$

2) (Armijo-type line search). Compute the least non-negative integer t such that

$$\widetilde{g}(\boldsymbol{w}^k + \gamma^t \boldsymbol{v}^k) \leq \widetilde{g}(\boldsymbol{w}^k) + \beta \gamma^t \Gamma_k, \tag{20c}$$

where $\Gamma_k := \nabla g(\boldsymbol{w}^k)^\top \boldsymbol{v}^k + \frac{1}{2\eta_k} \|\boldsymbol{v}^k\|_2^2$.

3) (Update). Set

$$\boldsymbol{w}^{k+1} \longleftarrow \boldsymbol{w}^k + \gamma^t \boldsymbol{v}^k. \tag{20d}$$

The method described above is a particular instantiation of the one studied in [27]. However, the convergence guarantees established in [27] for the method assume that the function g is continuously differentiable on an open set $\Omega_g \subseteq \mathbb{R}^{\bar{m}}$ containing Δ_s. As such, they cannot be directly applied to the setting where the logarithmic barrier regularizer $h_{\log}$ is

[1] A real-valued function $f : \mathbb{R}^p \to \mathbb{R}$ is said to be *analytic* on a set $S \subseteq \mathbb{R}^p$ if it is infinitely differentiable at and agrees with its Taylor series in a neighborhood of every point in S.

used. Nevertheless, a closer inspection of [27] reveals that the convergence results therein are still valid if the open set Ω_g merely satisfies $\Omega_g \cap \Delta_s \neq \emptyset$ and the initial point $\boldsymbol{w}^0$ is chosen to lie in $\Omega_g \cap \Delta_s$. This leads to our second convergence result.

Theorem 5 (Global convergence of LS-PGD under locally Lipschitz continuous gradient)**.** *Let $h : \mathbb{S}^m \to \mathbb{R} \cup \{+\infty\}$ be a regularizer whose domain $\mathrm{dom}(h) := \{\boldsymbol{L} \in \mathbb{S}^m : h(\boldsymbol{L}) < +\infty\}$ is open and satisfies $\mathrm{dom}(h) \cap \mathcal{L}_s \neq \emptyset$. In addition, suppose that h is continuously differentiable on $\mathrm{dom}(h)$ and its gradient ∇h is* locally Lipschitz continuous *on $\mathrm{dom}(h)$ (i.e., for every compact set $\mathcal{B} \subseteq \mathrm{dom}(h)$, there exists a constant $\ell_{\mathcal{B}} > 0$ such that $\|\nabla h(\boldsymbol{L}) - \nabla h(\boldsymbol{L}')\|_F \leq \ell_{\mathcal{B}} \|\boldsymbol{L} - \boldsymbol{L}'\|_F$ for all $\boldsymbol{L}, \boldsymbol{L}' \in \mathcal{B}$). Then, the following will hold almost surely:*

(a) *The function g is continuously differentiable on an open set $\Omega_g \subseteq \mathbb{R}^{\bar{m}}$ with $\Omega_g \cap \Delta_s \neq \emptyset$ and its gradient ∇g is locally Lipschitz continuous on Ω_g.*

(b) *If the function $\widetilde{g}$ possesses the KŁ property, then for any initial point $\boldsymbol{w}^0 \in \Omega_g \cap \Delta_s$, the sequence $\{\boldsymbol{w}^k\}_{k \geq 0}$ generated by the LS-PGD method* (20) *converges to a stationary point of Problem* (14)*.*

Proof. For any $\mathcal{U} \subseteq \mathbb{S}^m$, define $\mathcal{F}^{-1}(\mathcal{U}) := \{\boldsymbol{w} \in \mathbb{R}^{\bar{m}} : \mathcal{F}(\boldsymbol{w}) \in \mathcal{U}\}$. Observe that

$$\begin{aligned}\mathcal{F}^{-1}(\mathrm{dom}(h) \cap \mathcal{L}_s) &= \mathcal{F}^{-1}(\mathrm{dom}(h)) \cap \mathcal{F}^{-1}(\mathcal{L}_s) \\ &= \mathcal{F}^{-1}(\mathrm{dom}(h)) \cap \Delta_s.\end{aligned}$$

This, together with the assumption on h and the continuity of $\mathcal{F}$, implies that $\boldsymbol{w} \mapsto h(\mathcal{F}(\boldsymbol{w}))$ is continuously differentiable on the open set $\mathcal{F}^{-1}(\mathrm{dom}(h))$ with $\mathcal{F}^{-1}(\mathrm{dom}(h)) \cap \Delta_s \neq \emptyset$. On the other hand, note that the function ϕ_1 is continuously differentiable on a bounded open set Ξ_1 that contains Δ_s. Moreover, since $\mathbb{P}^* \ll \nu$ by assumption, Theorem 3 implies that almost surely, the function ϕ_2 will be continuously differentiable on a bounded open set Ξ_2 that contains Δ_s. By taking $\Omega_g = \mathcal{F}^{-1}(\mathrm{dom}(h)) \cap \Xi_1 \cap \Xi_2$, we see that Ω_g is open and g is continuously differentiable on Ω_g with $\Omega_g \cap \Delta_s = \mathcal{F}^{-1}(\mathrm{dom}(h)) \cap \Delta_s \neq \emptyset$. Lastly, the assumption that ∇h is locally Lipschitz continuous on $\mathrm{dom}(h)$, together with the fact that $\nabla \phi_1$ and $\nabla \phi_2$ are Lipschitz continuous on Ξ_1 and Ξ_2, respectively due to the boundedness of Ξ_1 and Ξ_2, implies that ∇g is locally Lipschitz continuous on Ω_g. This establishes the result in (a).

Now, let us prove by induction that $\boldsymbol{w}^k \in \Omega_g \cap \Delta_s$ for all $k \geq 0$. The base case follows from our assumption. For the inductive step, we first note that by the convexity of Δ_s, we have $\boldsymbol{w}^k + \gamma^t \boldsymbol{v}^k = (1-\gamma^t)\boldsymbol{w}^k + \gamma^t \widetilde{\boldsymbol{w}}^k \in \Delta_s$ for all $t \geq 0$. This, together with the fact that $\Delta_s \subset \Xi_i$ for $i = 1, 2$, implies that if $\boldsymbol{w}^k + \gamma^{t'} \boldsymbol{v}^k \notin \Omega_g$ for some $t' \geq 0$, then $\widetilde{g}(\boldsymbol{w}^k + \gamma^{t'} \boldsymbol{v}^k) = +\infty$; i.e., condition (20c) is not satisfied. Next, note that since Ω_g is open with $\boldsymbol{w}^k \in \Omega_g$ and $\gamma \in (0, 1)$, there exists an integer $T \geq 0$ such that $\boldsymbol{w}^k + \gamma^t \boldsymbol{v}^k \in \Omega_g$ for all $t \geq T$. Since the line search in the LS-PGD method (20) terminates in a finite number of steps (see the discussion in [27, Section 3.1]), we conclude that $\boldsymbol{w}^{k+1} \in \Omega_g \cap \Delta_s$, which completes the inductive step. The result in (b) can then be obtained by following the development in [27, Section 3.3]. □

The assumption on the regularizer h in Theorem 5 is much milder than that in Theorem 4. In particular, Theorem 5 applies to the setting where the logarithmic barrier regularizer $h_{\log}$ is used. Under such a setting, the KŁ property of $\widetilde{g}$ follows from the analyticity of $\boldsymbol{w} \mapsto h_{\log}(\mathcal{F}(\boldsymbol{w}))$ on the open set $\mathcal{F}^{-1}(\mathrm{dom}(h_{\log})) = \{\boldsymbol{w} \in \mathbb{R}^{\bar{m}} : [\mathcal{F}(\boldsymbol{w})]_{ii} > 0 \text{for } i = 1, \ldots, m\}$, the analyticity of ϕ_1, ϕ_2 on a bounded open set containing Δ_s, and the polyhedrality of Δ_s. As an aside, let us point out that the LS-PGD method (20) can also be used to solve the non-robust graph learning model proposed in [8], which is an instance of Problem (1) with $h(\boldsymbol{L}) = -\alpha \sum_{i=1}^m \ln(L_{ii}) + \frac{\beta}{2} \sum_{1 \leq i \neq j \leq m} L_{ij}^2$ and can be equivalently written as

$$\inf_{\boldsymbol{w} \in \Delta_s} \{\phi(\boldsymbol{w}) + h_{\log}(\mathcal{F}(\boldsymbol{w}))\} \tag{21}$$

with $\phi(\boldsymbol{w}) := \boldsymbol{w}^\top \mathcal{F}^* \left(\widehat{\boldsymbol{\Sigma}}_n + \widehat{\boldsymbol{\mu}}_n \widehat{\boldsymbol{\mu}}_n^\top\right) + \frac{\beta}{2} \sum_{1 \leq i \neq j \leq m} [\mathcal{F}(\boldsymbol{w})]_{ij}^2$. Since Problem (21) has a strongly convex objective function and a convex feasible set, Theorem 5 guarantees that the iterates generated by the method converge globally to its unique optimal solution. Interestingly, in the context of solving the non-robust graph learning model (21), both the proposed LS-PGD method (20) and its convergence guarantee given in Theorem 5 are new and can be of independent interest; cf. [28].

IV. Numerical Results

In this section, we study the performance of two distributionally robust and several representative non-robust graph learning models via numerical experiments on both synthetic and real-world data. Specifically, for the distributionally robust models, we consider our proposed MUGL formulation (11) with $s = m$, $h(\boldsymbol{L}) = 0$ (i.e., MUGL-*o*) and $s = m$, $h(\boldsymbol{L}) = -\alpha \sum_{i=1}^m \ln(L_{ii})$ (i.e., MUGL-*l*), as well as the following formulation:

- Wasserstein robust graph learning (WRGL) model [14] with parameters $\beta, \gamma > 0$:

$$\inf_{\boldsymbol{L} \in \mathcal{L}_m} \left\{ \frac{1}{n} \mathrm{tr}(\boldsymbol{X}^\top \boldsymbol{L} \boldsymbol{X}) + \beta \|\boldsymbol{L}\|_F + \gamma \|\boldsymbol{L}\|_F^2 \right\}.$$

For the non-robust graph learning models, we consider the following formulations:

- Vanilla smooth graph learning (VSGL) model:

$$\inf_{\boldsymbol{L} \in \mathcal{L}_m} \mathrm{tr}(\boldsymbol{X}^\top \boldsymbol{L} \boldsymbol{X}).$$

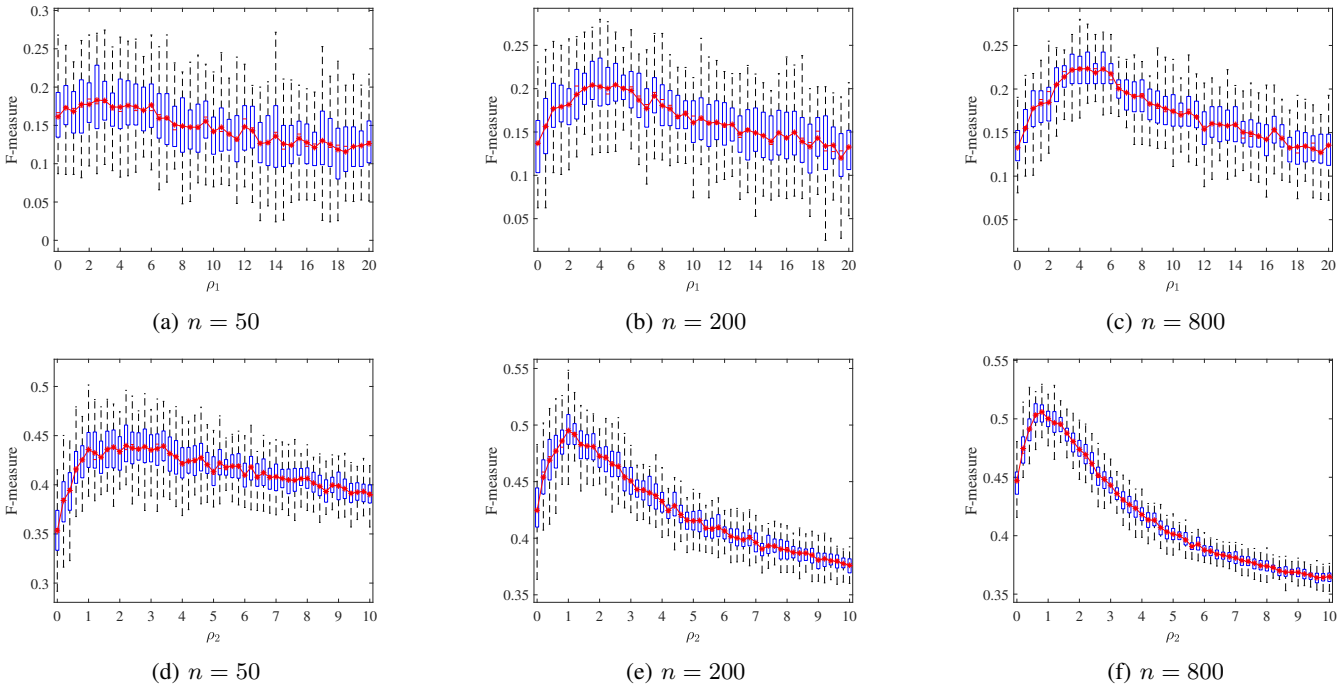
- GL-SigRep model [7] with parameters $\beta, \gamma > 0$:

$$\inf_{\boldsymbol{L} \in \mathcal{L}_m, \boldsymbol{Y} \in \mathbb{R}^{m \times n}} \left\{ \|\boldsymbol{X} - \boldsymbol{Y}\|_F^2 + \beta \, \mathrm{tr}(\boldsymbol{Y}^\top \boldsymbol{L} \boldsymbol{Y}) + \gamma \|\boldsymbol{L}\|_F^2 \right\}.$$

- Log-barrier model [8], [10] with parameters $\beta, \gamma > 0$:

$$\inf_{\boldsymbol{W} \in \mathbb{S}^m} \left\{ \frac{1}{2} \mathrm{tr}(\boldsymbol{Z} \boldsymbol{W}) - \beta \sum_{i=1}^m \ln \left(\sum_{j=1}^m W_{ij} \right) + \frac{\gamma}{2} \|\boldsymbol{W}\|_F^2 \right\}$$
$$\text{s.t.} \quad \boldsymbol{W} \geq \boldsymbol{0}, \, W_{ii} = 0 \text{ for } i = 1, \ldots, m,$$

where $\boldsymbol{Z} \in \mathbb{S}^m$ is the pairwise distance matrix given by $Z_{jk} = \|\boldsymbol{x}^j - \boldsymbol{x}^k\|_2^2$ for $j, k = 1, \ldots, m$.



(a) $n = 50$ (b) $n = 200$ (c) $n = 800$

(d) $n = 50$ (e) $n = 200$ (f) $n = 800$

Fig. 1: Prediction performance of MUGL-o with different values of ρ_1, ρ_2, and n

We use the following metrics to evaluate the performance of the different models (see, e.g., [29]):

$$\text{F-measure} = \frac{2\mathsf{TP}}{2\mathsf{TP}+\mathsf{FN}+\mathsf{FP}},$$
$$\text{Precision} = \frac{\mathsf{TP}}{\mathsf{TP}+\mathsf{FP}}, \quad \text{Recall} = \frac{\mathsf{TP}}{\mathsf{TP}+\mathsf{FN}},$$
$$\text{Normalized Mutual Information (NMI)} = \frac{2 \times I(\mathsf{TP}+\mathsf{FN};\mathsf{TP}+\mathsf{FP})}{H(\mathsf{TP}+\mathsf{FN})+H(\mathsf{TP}+\mathsf{FP})}.$$

Here, TP, FP, and FN denote the number of true positives, false positives, and false negatives, respectively; $H(\mathsf{TP}+\mathsf{FN})$ and $H(\mathsf{TP}+\mathsf{FP})$ denote the entropy of the edges in the underlying graph and in the learned graph, respectively; $I(\mathsf{TP}+\mathsf{FN};\mathsf{TP}+\mathsf{FP})$ denotes the mutual information between the edges in the underlying graph and those in the learned graph. The metrics *precision* and *recall* measure the fraction of correctly retrieved edges among all the edges in the learned graph and in the true graph, respectively. It should be noted that a high value in just one of these two metrics does not imply the graph is accurately learned. This motivates the metric *F-measure*, which is the harmonic mean of precision and recall. The metric *NMI* measures the mutual dependence between the learned graph and the ground-truth graph from an information-theoretic viewpoint, normalized by their entropy. A learning algorithm is deemed good if it achieves a high F-measure or NMI value.

Both the MUGL-o and VSGL models can be solved using the PGD method (19), while the MUGL-l model can be solved using the LS-PGD method (20). We initialize both the PGD (19) and LS-PGD (20) methods by the centroid of Δ_m. Our code is available at https://github.com/xwangcu/mugl. We use the code provided by the authors of [7] at http://web.media.mit.edu/~xdong/code/graphlearning.zip, which implements an alternating minimization method, to solve the GL-SigRep model. We use the code provided in the Graph Signal Processing Toolbox [30] at https://epfl-lts2.github.io/gspbox-html/doc/demos/gsp_demo_learn_graph_large.html, which implements a primal-dual method with the tricks given in [10], to solve the Log-barrier model. Lastly, we use the projected gradient method with backtracking line search [14, Algorithm 4] to solve the reformulation [14, Eq. (44)] of the WRGL model. All reported results are obtained using the best-tuned model parameters, so that the learned graphs have the highest quality in terms of F-measure or Rand index. The hyperparameter values used in our MUGL models for generating the results in Tables II–IV are provided in the supplementary material.

n	F-measure	MU	CU	ρ_1	ρ_2
50	0.542±7.62%	0.356	5.039	0.12	3.0
200	0.573±6.74%	0.093	4.713	0.09	2.7
800	0.610±3.66%	0.022	4.650	0.02	2.2

TABLE I: Prediction performance of MUGL-o with best-tuned values of ρ_1 and ρ_2 under different sample sizes n (MU (resp. CU) denotes the average value of the mean uncertainty $(\widehat{\boldsymbol{\mu}}_n - \boldsymbol{\mu}^*)^\top \boldsymbol{L}^*(\widehat{\boldsymbol{\mu}}_n - \boldsymbol{\mu}^*)$ (resp. covariance uncertainty $\|\widehat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}^*\|_F$))

A. Experiments on Synthetic Data

We conduct experiments on three types of synthetic graphs, namely, the Gaussian graph, the Erdős-Rényi (ER) graph, and the preferential attachment (PA) graph. The Gaussian graphs used in our experiments are generated as follows: First, the nodes are placed uniformly at random in a unit square. Then,

		F-measure	Precision	Recall	NMI
Gaussian	VSGL	0.411±9.87%	**0.963±5.07%**	0.262±12.46%	0.212±17.57%
	GL-SigRep	0.740±4.99%	0.593±8.00%	**0.988±1.30%**	0.401±13.39%
	Log-barrier	0.756±5.53%	0.850±5.14%	0.683±8.01%	0.396±16.05%
	WRGL	0.739±6.76%	0.861±6.55%	0.649±8.67%	0.344±22.96%
	MUGL-*o*	0.779±5.21%	0.831±5.38%	0.734±7.05%	0.415±16.26%
	MUGL-*l*	**0.785±5.70%**	0.816±6.46%	0.758±6.91%	**0.418±19.08%**
ER	VSGL	0.413±14.95%	**0.663±13.77%**	0.302±17.33%	0.146±33.57%
	GL-SigRep	0.580±7.26%	0.462±8.84%	0.784±7.99%	0.213±22.77%
	Log-barrier	**0.620±8.28%**	0.550±9.21%	0.712±9.23%	**0.248±23.77%**
	WRGL	0.553±6.43%	0.394±8.80%	**0.931±3.90%**	0.209±21.17%
	MUGL-*o*	0.554±6.45%	0.416±8.34%	0.832±5.99%	0.195±20.63%
	MUGL-*l*	0.570±6.49%	0.424±8.35%	0.876±5.34%	0.223±20.73%
PA	VSGL	0.546±7.46%	0.386±9.76%	0.938±5.16%	0.294±17.61%
	GL-SigRep	0.556±7.75%	0.414±10.02%	0.851±7.43%	0.271±18.11%
	Log-barrier	**0.654±7.86%**	**0.679±10.09%**	0.636±9.87%	**0.359±17.06%**
	WRGL	0.572±5.19%	0.401±7.20%	**1.000±0.01%**	0.354±8.11%
	MUGL-*o*	0.567±6.68%	0.400±9.28%	0.981±2.82%	0.337±13.00%
	MUGL-*l*	0.571±6.50%	0.405±9.15%	0.972±4.13%	0.336±13.63%

(a) $m = 20, n = 30, \epsilon = 0.1$

		F-measure	Precision	Recall	NMI
Gaussian	VSGL	0.473±8.51%	**0.987±2.63%**	0.312±11.24%	0.264±12.55%
	GL-SigRep	0.791±5.78%	0.788±3.65%	0.799±10.66%	0.423±16.21%
	Log-barrier	0.773±5.23%	0.879±4.03%	0.692±7.98%	0.430±14.16%
	WRGL	0.788±3.33%	0.671±5.35%	**0.958±1.64%**	0.390±13.40%
	MUGL-*o*	0.830±3.08%	0.856±2.46%	0.807±4.45%	0.497±9.75%
	MUGL-*l*	**0.840±3.84%**	0.842±3.91%	0.842±7.33%	**0.517±12.20%**
ER	VSGL	0.493±11.27%	0.763±10.22%	0.366±13.88%	0.216±25.59%
	GL-SigRep	0.512±9.19%	**0.767±11.55%**	0.387±11.54%	0.228±23.01%
	Log-barrier	0.560±8.83%	0.591±9.13%	0.535±10.85%	0.201±23.68%
	WRGL	0.512±4.37%	0.344±5.87%	**0.999±0.38%**	0.218±10.58%
	MUGL-*o*	0.545±9.49%	0.752±8.07%	0.429±12.39%	**0.241±20.14%**
	MUGL-*l*	**0.584±8.90%**	0.560±9.45%	0.613±11.01%	0.214±26.01%
PA	VSGL	0.715±7.39%	**0.994±2.27%**	0.561±10.99%	0.541±10.98%
	GL-SigRep	0.557±5.19%	0.561±11.25%	0.562±10.09%	0.249±12.06%
	Log-barrier	0.747±4.72%	0.765±6.27%	0.732±5.86%	0.477±10.57%
	WRGL	0.783±4.66%	0.645±7.59%	**0.998±1.04%**	0.591±8.71%
	MUGL-*o*	0.826±5.35%	0.905±6.17%	0.764±8.83%	0.620±11.64%
	MUGL-*l*	**0.893±5.44%**	0.963±3.71%	0.834±7.80%	**0.747±12.50%**

(b) $m = 20, n = 80, \epsilon = 0.1$

		F-measure	Precision	Recall	NMI
Gaussian	VSGL	0.222±21.30%	**0.867±14.30%**	0.128±23.33%	0.092±44.54%
	GL-SigRep	0.653±18.30%	0.521±25.83%	**0.919±6.98%**	0.241±59.03%
	Log-barrier	0.626±8.37%	0.730±8.76%	0.550±10.32%	0.225±28.22%
	WRGL	0.643±6.66%	0.524±8.00%	0.834±6.50%	0.135±44.85%
	MUGL-*o*	0.661±8.20%	0.670±10.62%	0.656±8.03%	0.231±30.33%
	MUGL-*l*	**0.700±7.09%**	0.608±9.86%	0.831±7.11%	**0.272±27.34%**
ER	VSGL	0.182±34.18%	**0.461±29.80%**	0.115±37.22%	0.037±68.13%
	GL-SigRep	0.354±11.66%	0.257±16.71%	0.645±29.52%	0.035±55.10%
	Log-barrier	0.490±8.50%	0.342±9.68%	**0.863±8.06%**	**0.148±31.94%**
	WRGL	0.432±12.31%	0.328±13.59%	0.634±11.75%	0.063±54.91%
	MUGL-*o*	0.438±5.50%	0.355±14.98%	0.576±11.51%	0.091±46.28%
	MUGL-*l*	**0.510±9.67%**	0.408±13.21%	0.692±10.04%	0.139±33.37%
PA	VSGL	0.297±40.12%	**0.558±37.85%**	0.205±43.42%	0.120±72.23%
	GL-SigRep	0.365±22.03%	0.418±25.05%	0.328±22.33%	0.109±51.04%
	Log-barrier	0.424±21.77%	0.426±20.12%	0.424±20.26%	0.139±42.90%
	WRGL	0.418±11.06%	0.274±12.99%	**0.886±7.72%**	0.168±31.69%
	MUGL-*o*	0.436±15.82%	0.331±17.56%	0.643±15.09%	0.141±40.08%
	MUGL-*l*	**0.460±25.93%**	0.490±19.72%	0.434±21.59%	**0.180±54.72%**

(c) $m = 20, n = 30, \epsilon = 1$

		F-measure	Precision	Recall	NMI
Gaussian	VSGL	0.279±20.63%	**0.929±8.62%**	0.165±23.85%	0.133±34.70%
	GL-SigRep	0.772±8.21%	0.757±7.46%	0.794±12.25%	0.393±23.90%
	Log-barrier	0.718±7.02%	0.817±6.66%	0.641±8.85%	0.340±21.29%
	WRGL	0.712±5.35%	0.590±6.96%	0.899±4.69%	0.240±28.23%
	MUGL-*o*	0.756±4.98%	0.637±8.67%	**0.938±4.38%**	0.387±16.27%
	MUGL-*l*	**0.856±2.46%**	0.807±4.45%	0.830±3.08%	**0.497±9.75%**
ER	VSGL	0.265±27.30%	**0.573±23.03%**	0.174±30.03%	0.074±56.41%
	GL-SigRep	0.373±6.95%	0.252±12.70%	**0.752±14.49%**	0.042±34.87%
	Log-barrier	0.446±18.05%	0.554±18.68%	0.375±19.05%	0.133±46.00%
	WRGL	0.495±7.67%	0.351±8.92%	0.495±7.67%	0.128±34.24%
	MUGL-*o*	0.441±15.8%	0.541±15.32%	0.376±18.36%	0.125±39.25%
	MUGL-*l*	**0.508±9.18%**	0.410±13.30%	0.676±8.80%	**0.135±31.58%**
PA	VSGL	0.454±23.57%	**0.827±15.89%**	0.320±29.69%	0.261±37.10%
	GL-SigRep	0.494±15.17%	0.762±16.97%	0.369±17.62%	0.263±31.47%
	Log-barrier	0.534±12.43%	0.579±12.88%	0.497±13.05%	0.236±27.16%
	WRGL	0.489±7.21%	0.329±9.09%	**0.962±4.69%**	0.256±17.52%
	MUGL-*o*	0.537±18.57%	0.789±17.31%	0.413±22.07%	**0.304±34.17%**
	MUGL-*l*	**0.557±11.09%**	0.476±13.41%	0.678±12.71%	0.247±26.52%

(d) $m = 20, n = 80, \epsilon = 1$

TABLE II: Prediction performance on synthetic data with different values of m, n, and ϵ

an edge is placed between nodes i and j ($i \neq j$) if the weight determined by the radial basis function $\exp\left(-d(i,j)^2/2\sigma^2\right)$, where $d(i,j)$ is the Euclidean distance between nodes i and j and $\sigma = 0.5$ is the kernel width parameter, is at least 0.75. The ER graphs are generated by placing an edge between each pair of nodes independently with probability $p = 0.2$. The PA graphs are generated by having $\theta_0 = 2$ connected nodes initially and then adding new nodes one at a time, where each new node is connected to exactly $\theta = 1$ previous node that is randomly chosen with a probability proportional to its degree at the time. The edges in the Gaussian graph have weights given by the radial basis function, while those in the ER and PA graphs are set to 1. After obtaining the synthetic graphs, we use the factor analysis model introduced in [7] to generate the graph signals on them. Specifically, given a graph, let $\boldsymbol{L}^*$ be its Laplacian whose eigen-decomposition is given by $\boldsymbol{L}^* = \boldsymbol{\chi}\boldsymbol{\Lambda}\boldsymbol{\chi}^\top$. The graph signal $\boldsymbol{x} \in \mathbb{R}^m$ is then generated according to (5), where the entries of $\boldsymbol{\mu}^*$ are independently generated according to the uniform distribution on the interval $(0, 1)$.

1) Ablation Study: We first conduct an ablation study on how the parameters ρ_1 and ρ_2 in the MUGL-*o* model, which control the size of the ambiguity set $\mathcal{M}(\boldsymbol{L}, \rho_1, \rho_2)$, affect its performance. We take an ER graph with $m = 20$ and consider the noise level $\epsilon = 0.01$ in (5). We use MUGL-*o* with different choices of ρ_1 and ρ_2 to recover the graph. The experiment results are reported in Figure 1, where each red asterisk (resp. blue box) represents the average F-measure (resp. interval between the 25th and 75th percentiles) over 50 runs with $n = 50, 200,$ or 800 independently generated noisy graph signals.

Figures 1a–1c show how fixing $\rho_2 = 0$ and changing ρ_1 affects the F-measure of the learned graphs. Similarly, Figures 1d–1f show how fixing $\rho_1 = 0$ and changing ρ_2 affects

the F-measure of the learned graphs. We observe from the sub-figures that as ρ_1 or ρ_2 increases (i.e., the ambiguity set $\mathcal{M}(\boldsymbol{L}, \rho_1, \rho_2)$ becomes bigger), the percentile box becomes smaller. Moreover, viewing Figure 1 row-by-row, we see that the larger the number of observed graph signals (which we shall refer to as the *sample size*) n, the smaller the percentile box. These results suggest that the prediction performance becomes more consistent across different populations of graph signals when the ambiguity set enlarges or the sample size increases. However, it should be noted that the average F-measure of the learned graphs eventually decreases as the size of the ambiguity set increases. This is due in part to the fact that MUGL-*o* becomes more conservative when the ambiguity set becomes larger. We can also observe from Figure 1 that the larger the sample size n, the smaller the optimal ρ_1 and ρ_2 that achieve the highest average F-measure. This can be attributed to the fact that when the sample size n increases, there is less uncertainty about the ground-truth distribution and thus smaller ρ_1 and ρ_2 suffice for the ambiguity set to contain the ground-truth distribution; see Theorems 1 and 2. In summary, Figure 1 demonstrates that with suitable choices of ρ_1 and ρ_2, the MUGL model can yield graphs that have a high prediction performance not just on a single population but consistently across different populations of observed signals.

Next, we examine how the sample size n affects the performance of MUGL-*o*. For $n = 50, 200,$ or 800, we simultaneously tune the hyperparameters ρ_1 and ρ_2 of MUGL-*o* so that it yields the graph with the highest average F-measure. Table I presents the average F-measure and the associated normalized standard deviations, the average values of the mean uncertainty $(\widehat{\boldsymbol{\mu}}_n - \boldsymbol{\mu}^*)^\top \boldsymbol{L}^* (\widehat{\boldsymbol{\mu}}_n - \boldsymbol{\mu}^*)$ and covariance uncertainty $\|\widehat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}^*\|_F$, and the best-tuned values of ρ_1 and ρ_2. As the sample size n increases, the average F-measure increases and the normalized standard deviations decreases. Analogous to the pattern shown in Figure 1, the larger the n, the smaller the mean and covariance uncertainties, thus smaller ρ_1 and ρ_2 are required.

2) Model Performance: Let us now compare the efficacy of different graph learning models. We consider two noise levels $\epsilon = 0.1$ and 1 in (5). We examine the performance of the graphs learned from $n = 30$ or 80 independently generated noisy graph signals, averaged over 50 runs. The results are reported in Table II, where the entries give the average performance as measured by the stated metrics and the associated normalized standard deviations. We observe that VSGL always yields the highest precision values, since this vanilla model usually produces very sparse graphs. Moreover, MUGL-*o* exhibits significant performance gain over VSGL in terms of F-measure and NMI. This demonstrates the advantage of robustifying the vanilla graph learning model against moment uncertainties about the ground-truth distribution.

As can also be observed from the results, when the noise level is low ($\epsilon = 0.1$) and the sample size is small ($n = 30$), the distributionally robust models do not have a clear advantage over the non-robust GL-SigRep and Log-barrier models. On one hand, the low noise level results in less uncertainty about the ground-truth distribution of the graph signal, so that the performance of non-robust models is less affected by the use of the empirical distribution as a surrogate of the ground-truth distribution. On the other hand, the distributionally robust models, which need to take into account the ambiguity in the ground-truth distribution caused by both the graph signal noise and limited sample size, tend to produce more conservative solutions. Still, MUGL-*l* achieves the highest F-measure and NMI values in most cases. Moreover, the normalized standard deviation of MUGL is generally lower than those of the other three non-robust methods, particularly in the experiments involving Gaussian graphs with relatively noisy ($\epsilon = 1$) graph signals. This indicates that our proposed MUGL model is able to achieve its main aim of attaining a more consistent performance across different populations of observed signals.

	F-measure	Precision	Recall	NMI
VSGL	0.771	0.637	**0.976**	0.418
GL-SigRep	0.791	**0.828**	0.757	0.415
Log-barrier	0.812	0.749	0.887	0.441
WRGL	0.619	0.527	0.749	0.145
MUGL-*o*	0.827	0.736	0.943	0.488
MUGL-*l*	**0.837**	0.802	0.875	**0.504**

TABLE III: Prediction performance on real temperature data

B. Experiments on Real Temperature Data

We conduct experiments on the real-world temperature data provided in [7]. The dataset consists of monthly temperature data from 1981 to 2010 collected by 89 measuring stations in Switzerland. We construct a graph in which the nodes correspond to the measuring stations and the edges correspond to two stations whose altitude difference is less than 300 meters. Furthermore, we assign a weight of 1 to each edge. Such a construction is motivated by the fact that temperature difference is highly related to altitude difference. For each station, we compute the average temperature of each month over the 30-year period. Thus, each month yields a graph signal, and we have 12 graph signals in total. Given these signals, we aim to recover the graph that reflects the altitude relationships between the stations. We then evaluate the learned graph using the same metrics as those in the previous subsection. The results are reported in Table III. VSGL performs fairly well in this scenario, but MUGL-*o* performs even better in terms of F-measure and NMI. Among the six considered approaches, MUGL-*l* achieves the highest F-measure and NMI values.

C. Experiments on Real Image Data

We further evaluate the efficacy of the different models by applying them to learn the similarity graph of real images and using the learned graph to perform spectral clustering [31]. We consider two different image datasets, namely, USPS [32] and COIL-20 [33]. The former consists of 7291 training images and 2007 test images, each of which is a 16×16 grayscale handwritten digit from 0 to 9. The latter consists of 1440 images of 20 different objects, each of which is downsampled to a size of 32×32. In the context of clustering, these two image datasets contain 10 and 20 clusters, respectively.

	RI	JC	FMI
VSGL	0.512±20.56%	0.138±21.78%	0.321±13.99%
Log-barrier	0.476±12.48%	0.100±3.99%	0.243±6.30%
GL-SigRep	0.782±2.03%	0.066±6.89%	0.124±7.30%
WRGL	0.803±9.48%	0.299±33.80%	0.487±20.51%
k-NNG	0.689±9.31%	0.156±19.19%	0.307±15.66%
MUGL-*o*	0.892±1.67%	0.333±15.18%	0.498±11.33%
MUGL-*l*	**0.896±1.64%**	**0.341±16.45%**	**0.507±12.25%**

(a) USPS

	RI	JC	FMI
VSGL	0.412±11.97%	0.881±3.66%	0.226±19.22
Log-barrier	0.616±3.44%	0.113±30.69%	0.221±19.75%
GL-SigRep	0.578±5.73%	0.167±25.04 %	0.290±19.32%
WRGL	0.955±0.55%	0.412±10.56%	0.584±7.44%
k-NNG	0.761±9.35%	0.107±27.21%	0.252±16.35%
MUGL-*o*	0.964±0.67%	0.485±12.65%	0.651±8.83%
MUGL-*l*	**0.964±0.65%**	**0.490±11.93%**	**0.657±8.36%**

(b) COIL-20

TABLE IV: Spectral clustering performance on image datasets

To perform spectral clustering, we use the algorithm in [34], which involves (i) constructing a graph based on the *similarity matrix* $\boldsymbol{W}$ of the dataset, (ii) finding the Laplacian $\boldsymbol{L} = \mathrm{Diag}(\boldsymbol{W}\mathbf{1}) - \boldsymbol{W}$ and computing the normalized random-walk Laplacian $\boldsymbol{L}_{\mathrm{rw}} = (\mathrm{Diag}(\boldsymbol{W}\mathbf{1}))^{-1}\boldsymbol{L}$, (iii) finding the k leading eigenvectors of $\boldsymbol{L}_{\mathrm{rw}}$, and (iv) splitting the graph by applying the standard k-means algorithm on the rows of the k eigenvectors. The similarity matrix captures the relative similarity of each pair of points in the dataset in a quantitative manner, and its quality will influence the performance of spectral clustering. Given an image dataset, we postulate that there is an unknown complete m-node graph in which (i) each node corresponds to an image in the dataset and (ii) the weight of an edge between two nodes represents the similarity between the two corresponding images. Each pixel of an image gives an observed value at the node corresponding to that image. The collection of all such values constitutes our graph signals. In particular, if there are m images of the same size $n_r \times n_c$ in a dataset, then there are $n = n_r \times n_c$ graph signals, each of which has dimension m. Our goal then is to learn a similarity matrix from these graph signals, so as to facilitate the subsequent clustering task.

In each run of the experiment, we randomly pick 100 images from the USPS dataset and 200 images from the COIL-20 dataset, which give rise to graphs with $m = 100$ and $m = 200$ nodes, respectively. According to the sizes of the images in the USPS and COIL-20 datasets, we obtain $n = 16 \times 16 = 256$ graph signals from the former and $n = 32 \times 32 = 1024$ graph signals from the latter. As a baseline, we construct a k-nearest neighbor graph (k-NNG) associated with the images as follows. For each node i (which corresponds to an image), we connect it node j ($i \neq j$) and assign the weight $\exp\left(-d(i,j)^2/0.5\right)$ to the edge if the Euclidean distance $d(i,j)$ is among the k smallest distances from i to all other nodes. By ignoring the directions of the edges, we obtain an undirected graph, which yields a symmetric similarity matrix. We choose $k = 8$ in our experiments since it exhibits the best empirical clustering performance on both datasets. Given the graphs learned by different models and the k-NNG, we run the spectral clustering algorithm in [34] and evaluate the results using the following common clustering performance metrics: Rand index (RI), Jaccard coefficient (JC), and Fowlkes and Mallows index (FMI) [29], [35]. These metrics measure the similarity between the partition returned by a clustering algorithm and the true data partition. Suppose that the clustering algorithm partitions a set of m images $\mathcal{C} = \{\boldsymbol{a}_1, \ldots, \boldsymbol{a}_m\}$ into k disjoint subsets $\mathcal{C}_1, \ldots, \mathcal{C}_k$, and the true partition of $\mathcal{C}$ is $\mathcal{C}_1^*, \ldots, \mathcal{C}_s^*$. Let $\ell, \ell^* \in \mathbb{R}^m$ be the data label vectors associated with $\mathcal{C}$ and $\mathcal{C}^*$, respectively. If we define the sets $\mathsf{SS} \coloneqq \{(\boldsymbol{a}_i, \boldsymbol{a}_j) : \ell_i = \ell_j, \ell_i^* = \ell_j^*, i < j\}$, $\mathsf{SD} \coloneqq \{(\boldsymbol{a}_i, \boldsymbol{a}_j) : \ell_i = \ell_j, \ell_i^* \neq \ell_j^*, i < j\}$, $\mathsf{DS} \coloneqq \{(\boldsymbol{a}_i, \boldsymbol{a}_j) : \ell_i \neq \ell_j, \ell_i^* = \ell_j^*, i < j\}$, and $\mathsf{DD} \coloneqq \{(\boldsymbol{a}_i, \boldsymbol{a}_j) : \ell_i \neq \ell_j, \ell_i^* \neq \ell_j^*, i < j\}$, then the metrics RI, JC, and FMI can be defined in terms of the cardinalities of the these sets as follows:

$$\mathrm{RI} = \frac{|\mathsf{SS}| + |\mathsf{DD}|}{m(m-1)/2},\ \mathrm{JC} = \frac{|\mathsf{SS}|}{|\mathsf{SS}| + |\mathsf{SD}| + |\mathsf{DS}|},$$
$$\mathrm{FMI} = \sqrt{\frac{|\mathsf{SS}|}{|\mathsf{SS}| + |\mathsf{SD}|} \cdot \frac{|\mathsf{SS}|}{|\mathsf{SS}| + |\mathsf{DS}|}}.$$

In particular, RI measures the percentage of correct decisions made by the algorithm, which is simply the clustering accuracy. All the above metrics yield values that lie in $[0, 1]$, and a higher value indicates better performance in principle. The results are presented in Table IV, where the entries give the average performance as measured by the stated metrics and the associated normalized standard deviations over 10 runs. Among the compared approaches, MUGL-*o* and MUGL-*l* achieve better and more consistent performance in terms of all three aforementioned metrics in both the USPS and COIL-20 datasets. This indicates the high quality and robustness of the learned graphs produced by the MUGL model.

V. Conclusion

We have developed a novel DRO-based approach to graph learning, which provides a way to identify a graph that not only yields a smooth representation of the observed signals but is also robust against uncertainties about the ground-truth distribution of the graph signal. We have demonstrated how to construct the ambiguity set in our distributionally robust graph learning model by exploiting the structure of the Laplacian quadratic form and establishing confidence regions for the mean and covariance of the ground-truth distribution. We have also shown that whenever the ground-truth distribution has a probability density function, our proposed model admits a smooth non-convex optimization formulation. Interestingly, such a formulation provides a new perspective on regularization in the graph learning setting. Then, we have presented a PGD method to numerically tackle the formulation and established its convergence guarantees. Through extensive numerical experiments, we have shown that our proposed model improves the quality of the learned graphs and robustifies the performance across different populations of observed signals.

One promising future direction is to extend our proposed approach to tackle more general graph learning scenarios.

APPENDIX

A. Proof of Theorem 2

To set the stage, let us introduce some additional notation. Given a real number $p \geq 1$ and a $q_1 \times q_2$ matrix $\boldsymbol{A}$, we use $\|\boldsymbol{A}\|_{S_p}$ to denote the Schatten p-norm of $\boldsymbol{A}$; i.e., $\|\boldsymbol{A}\|_{S_p} \coloneqq \|\sigma(\boldsymbol{A})\|_p$, where $\sigma(\boldsymbol{A}) \in \mathbb{R}_+^{\min\{q_1,q_2\}}$ is the vector of singular values of $\boldsymbol{A}$ and $\|\cdot\|_p$ is the usual vector p-norm. By definition, we have $\|\boldsymbol{A}\|_{S_2} = \|\boldsymbol{A}\|_F$.

We begin by establishing a relationship between the matrix

$$\widetilde{\boldsymbol{\Sigma}}_n \coloneqq \frac{1}{n}\sum_{j=1}^n (\boldsymbol{x}^j - \boldsymbol{\mu}^*)(\boldsymbol{x}^j - \boldsymbol{\mu}^*)^\top$$

and the covariance matrix $\boldsymbol{\Sigma}^*$ of the ground-truth distribution $\mathbb{P}^*$. Note that the matrix $\widetilde{\boldsymbol{\Sigma}}_n$ is not the same as the empirical covariance matrix $\widehat{\boldsymbol{\Sigma}}_n$ defined in (4), as the former is defined using $\boldsymbol{\mu}^*$ and not $\widehat{\boldsymbol{\mu}}_n$. Nevertheless, as we shall see, we can use the relationship between $\widetilde{\boldsymbol{\Sigma}}_n$ and $\boldsymbol{\Sigma}^*$ to establish the desired relationship between $\widehat{\boldsymbol{\Sigma}}_n$ and $\boldsymbol{\Sigma}^*$.

Proposition 4. *Under the setting of Theorem 2, there exists a constant $c_1 > 0$ such that*

$$\|\widetilde{\boldsymbol{\Sigma}}_n - \boldsymbol{\Sigma}^*\|_F \leq \frac{4c_1(2e/3)^{3/2}\ln^{3/2}(4m^{3/2}/\xi)}{n^{1/2}}\|\boldsymbol{\Sigma}^*\|$$

will hold with probability at least $1 - \xi/2$.

Proof. For $j = 1, \ldots, n$, define

$$\boldsymbol{Q}_j \coloneqq \boldsymbol{\Sigma}^{*-1/2}(\boldsymbol{x}^j - \boldsymbol{\mu}^*)(\boldsymbol{x}^j - \boldsymbol{\mu}^*)^\top \boldsymbol{\Sigma}^{*-1/2} - \boldsymbol{I}_m.$$

A straightforward calculation shows that $\mathbb{E}_{\boldsymbol{x}^j \sim \mathbb{P}^*}[\boldsymbol{Q}_j] = \boldsymbol{0}$ for $j = 1, \ldots, n$. Moreover, since $\mathbb{P}^*$ satisfies the moment growth condition, there exists a constant $c' > 0$ such that for all $p \geq 1$,

$$\mathbb{E}_{\boldsymbol{x} \sim \mathbb{P}^*}\left[\|\boldsymbol{\Sigma}^{*-1/2}(\boldsymbol{x} - \boldsymbol{\mu})\|_2^p\right] \leq (c'p)^{p/2}. \tag{22}$$

By combining (22) with the argument in the proof of [18, Proposition 5], we deduce that for any $p \geq 1$,

$$\mathbb{E}_{\boldsymbol{x}^1, \ldots, \boldsymbol{x}^n \sim \mathbb{P}^*}\left[\left\|\sum_{j=1}^n \boldsymbol{Q}_j\right\|_{S_p}^p\right] \leq 2^p n^{p/2} p^{p/2}(m + (2c'p)^p).$$

Now, using the fact that $\|\boldsymbol{v}\|_2 \leq q^{1/2}\|\boldsymbol{v}\|_p$ for any $\boldsymbol{v} \in \mathbb{R}^q$ and $p \in [2, +\infty]$ [2] and applying Markov's inequality, we have, for any $p \geq 2$ and $t > 0$, that

$$\begin{aligned}
&\Pr\left(\left\|\frac{1}{n}\sum_{j=1}^n \boldsymbol{Q}_j\right\|_F > t\right) = \Pr\left(\left\|\frac{1}{n}\sum_{j=1}^n \boldsymbol{Q}_j\right\|_{S_2}^p > t^p\right) \\
&\leq \frac{2^p p^{p/2} m^{1/2}(m + (2c'p)^p)}{t^p n^{p/2}}.
\end{aligned}$$

In particular, by setting $c_1 = \max\{c', 1/4\}$,

$$t = \frac{4c_1(2e/3)^{3/2}\ln^{3/2}(4m^{3/2}/\xi)}{n^{1/2}}, \quad p = \left(\frac{tn^{1/2}}{4c_1 e^{3/2}}\right)^{2/3}$$

and noting that $\xi \leq e^{-2}$, we have $p = 2\ln(4m^{3/2}/\xi)/3 \geq 2$ and

$$\frac{2^p p^{p/2} m^{1/2}(m + (2c'p)^p)}{t^p n^{p/2}} = \frac{m^{3/2} + m^{1/2}(2c'p)^p}{e^{3p/2}(2c_1 p)^p} \leq \frac{\xi}{2}.$$

This, together with

$$\|\widetilde{\boldsymbol{\Sigma}}_n - \boldsymbol{\Sigma}^*\|_F = \left\|\boldsymbol{\Sigma}^{*1/2}\left(\frac{1}{n}\sum_{j=1}^n \boldsymbol{Q}_j\right)\boldsymbol{\Sigma}^{*1/2}\right\|_F,$$

implies the desired result. □

To proceed, observe that

$$\begin{aligned}
\widetilde{\boldsymbol{\Sigma}}_n &= \frac{1}{n}\sum_{j=1}^n (\boldsymbol{x}^j - \widehat{\boldsymbol{\mu}}_n + \widehat{\boldsymbol{\mu}}_n - \boldsymbol{\mu}^*)(\boldsymbol{x}^j - \widehat{\boldsymbol{\mu}}_n + \widehat{\boldsymbol{\mu}}_n - \boldsymbol{\mu}^*)^\top \\
&= \widehat{\boldsymbol{\Sigma}}_n + \frac{1}{n}\sum_{j=1}^n (\boldsymbol{x}^j - \widehat{\boldsymbol{\mu}}_n)(\widehat{\boldsymbol{\mu}}_n - \boldsymbol{\mu}^*)^\top \\
&\quad + \frac{1}{n}\sum_{j=1}^n (\widehat{\boldsymbol{\mu}}_n - \boldsymbol{\mu}^*)(\boldsymbol{x}^j - \widehat{\boldsymbol{\mu}}_n)^\top + (\widehat{\boldsymbol{\mu}}_n - \boldsymbol{\mu}^*)(\widehat{\boldsymbol{\mu}}_n - \boldsymbol{\mu}^*)^\top \\
&= \widehat{\boldsymbol{\Sigma}}_n + (\widehat{\boldsymbol{\mu}}_n - \boldsymbol{\mu}^*)(\widehat{\boldsymbol{\mu}}_n - \boldsymbol{\mu}^*)^\top.
\end{aligned}$$

Hence, we have

$$\begin{aligned}
\|\widehat{\boldsymbol{\Sigma}}_n - \boldsymbol{\Sigma}^*\|_F &\leq \|\widehat{\boldsymbol{\Sigma}}_n - \widetilde{\boldsymbol{\Sigma}}_n\|_F + \|\widetilde{\boldsymbol{\Sigma}}_n - \boldsymbol{\Sigma}^*\|_F \\
&= (\widehat{\boldsymbol{\mu}}_n - \boldsymbol{\mu}^*)^\top(\widehat{\boldsymbol{\mu}}_n - \boldsymbol{\mu}^*) + \|\widetilde{\boldsymbol{\Sigma}}_n - \boldsymbol{\Sigma}^*\|_F.
\end{aligned}$$

Since $\mathbb{P}^*$ satisfies the moment growth condition, by taking c_2 to be the constant c in Definition 1 and adapting the proof of [18, Proposition 4], we deduce that with probability at least $1 - \xi/2$,

$$(\widehat{\boldsymbol{\mu}}_n - \boldsymbol{\mu}^*)^\top(\widehat{\boldsymbol{\mu}}_n - \boldsymbol{\mu}^*) \leq \frac{4c_2 e^2 \ln^2(2/\xi)}{n}.$$

This, together with Proposition 4, implies that $\|\widehat{\boldsymbol{\Sigma}}_n - \boldsymbol{\Sigma}^*\|_F \leq \widehat{\rho}_2$ will hold with probability at least $1 - \xi$, as desired.

B. Proof of Proposition 2

We first consider Problem (9). Upon letting $\widetilde{\boldsymbol{\mu}} = \boldsymbol{L}^{1/2}\boldsymbol{\mu}$, we can rewrite Problem (9) as

$$\begin{aligned}
\sup_{\widetilde{\boldsymbol{\mu}} \in \mathbb{R}^m} \ & \|\widetilde{\boldsymbol{\mu}}\|_2^2 \\
\text{s.t.} \ & \|\widetilde{\boldsymbol{\mu}} - \boldsymbol{L}^{1/2}\widehat{\boldsymbol{\mu}}_n\|_2^2 \leq \rho_1^2,
\end{aligned} \tag{23}$$

whose associated Lagrangian function can be written as $\mathscr{L}(\widetilde{\boldsymbol{\mu}}, \lambda) = -\|\widetilde{\boldsymbol{\mu}}\|_2^2 + \lambda(\|\widetilde{\boldsymbol{\mu}} - \boldsymbol{L}^{1/2}\widehat{\boldsymbol{\mu}}_n\|_2^2 - \rho_1^2)$. Since Problem (23) satisfies the linear independence constraint qualification, its associated KKT conditions, which are given by

$$\begin{aligned}
-\widetilde{\boldsymbol{\mu}} + \lambda(\widetilde{\boldsymbol{\mu}} - \boldsymbol{L}^{1/2}\widehat{\boldsymbol{\mu}}_n) &= \boldsymbol{0}, \\
\lambda\left(\|\widetilde{\boldsymbol{\mu}} - \boldsymbol{L}^{1/2}\widehat{\boldsymbol{\mu}}_n\|_2^2 - \rho_1^2\right) &= 0, \\
\lambda &\geq 0, \\
\|\widetilde{\boldsymbol{\mu}} - \boldsymbol{L}^{1/2}\widehat{\boldsymbol{\mu}}_n\|_2^2 &\leq \rho_1^2,
\end{aligned}$$

are necessary for optimality. We consider the following two possibilities for the dual multiplier λ:

Case I: $\lambda = 0$. The KKT conditions reduce to

$$\widetilde{\boldsymbol{\mu}} = \boldsymbol{0}, \quad \|\boldsymbol{L}^{1/2}\widehat{\boldsymbol{\mu}}_n\|_2^2 \leq \rho_1^2.$$

[2] The stated bound is not sharp but is sufficient for our purposes. Readers who are interested in the sharp bound can refer to, e.g., [36, Lemma 1].

The objective value of Problem (23) associated with the solution $\widetilde{\boldsymbol{\mu}} = \mathbf{0}$ is 0, which is obviously not the maximum.

Case II: $\lambda > 0$. The KKT conditions become

$$-\widetilde{\boldsymbol{\mu}} + \lambda(\widetilde{\boldsymbol{\mu}} - \boldsymbol{L}^{1/2}\widehat{\boldsymbol{\mu}}_n) = \mathbf{0}, \tag{24a}$$

$$\|\widetilde{\boldsymbol{\mu}} - \boldsymbol{L}^{1/2}\widehat{\boldsymbol{\mu}}_n\|_2^2 = \rho_1^2, \tag{24b}$$

$$\lambda > 0. \tag{24c}$$

Let $\mathrm{null}(\boldsymbol{L}^{1/2})$ denote the nullspace of $\boldsymbol{L}^{1/2}$. Consider the following two subcases:

(i) If $\widehat{\boldsymbol{\mu}}_n \in \mathrm{null}(\boldsymbol{L}^{1/2})$, then $\widetilde{\boldsymbol{\mu}}$ satisfies $\|\widetilde{\boldsymbol{\mu}}\|_2^2 = \rho_1^2 > 0$ by (24b), which yields the objective value ρ_1^2.

(ii) If $\widehat{\boldsymbol{\mu}}_n \notin \mathrm{null}(\boldsymbol{L}^{1/2})$, then $\lambda \neq 1$. By (24a), we have $\widetilde{\boldsymbol{\mu}} = \frac{\lambda}{\lambda-1}\boldsymbol{L}^{1/2}\widehat{\boldsymbol{\mu}}_n$. Substituting this into (24b) yields

$$\lambda = 1 \pm \frac{\|\boldsymbol{L}^{1/2}\widehat{\boldsymbol{\mu}}_n\|_2}{\rho_1}.$$

If $\lambda = 1 - \frac{\|\boldsymbol{L}^{1/2}\widehat{\boldsymbol{\mu}}_n\|_2}{\rho_1}$ and $\lambda > 0$, then $\rho_1 > \|\boldsymbol{L}^{1/2}\widehat{\boldsymbol{\mu}}_n\|_2$ and

$$\widetilde{\boldsymbol{\mu}} = \widetilde{\boldsymbol{\mu}}_- := \left(1 - \frac{\rho_1}{\|\boldsymbol{L}^{1/2}\widehat{\boldsymbol{\mu}}_n\|_2}\right)\boldsymbol{L}^{1/2}\widehat{\boldsymbol{\mu}}_n.$$

The objective value of Problem (23) associated with the solution $\widetilde{\boldsymbol{\mu}}_-$ is

$$\begin{aligned}\|\widetilde{\boldsymbol{\mu}}_-\|_2^2 &= \left(1 - \frac{\rho_1}{\|\boldsymbol{L}^{1/2}\widehat{\boldsymbol{\mu}}_n\|_2}\right)^2 \|\boldsymbol{L}^{1/2}\widehat{\boldsymbol{\mu}}_n\|_2^2 \\ &= \left(\|\boldsymbol{L}^{1/2}\widehat{\boldsymbol{\mu}}_n\|_2 - \rho_1\right)^2.\end{aligned}$$

On the other hand, if $\lambda = 1 + \frac{\|\boldsymbol{L}^{1/2}\widehat{\boldsymbol{\mu}}_n\|_2}{\rho_1}$, then

$$\widetilde{\boldsymbol{\mu}} = \widetilde{\boldsymbol{\mu}}_+ := \left(1 + \frac{\rho_1}{\|\boldsymbol{L}^{1/2}\widehat{\boldsymbol{\mu}}_n\|_2}\right)\boldsymbol{L}^{1/2}\widehat{\boldsymbol{\mu}}_n,$$

which yields the objective value

$$\|\widetilde{\boldsymbol{\mu}}_+\|_2^2 = \left(\|\boldsymbol{L}^{1/2}\widehat{\boldsymbol{\mu}}_n\|_2 + \rho_1\right)^2.$$

Summarizing the above cases, we conclude that $\widetilde{\boldsymbol{\mu}}_+$ is an optimal solution to Problem (23). This implies that $\varphi_1(\boldsymbol{L}) = \left(\|\boldsymbol{L}^{1/2}\widehat{\boldsymbol{\mu}}_n\|_2 + \rho_1\right)^2$, as desired.

Next, we consider Problem (10). By dropping the constraint $\boldsymbol{\Sigma} \in \mathbb{S}_+^m$ from Problem (10), we obtain the following relaxation:

$$\begin{aligned}\sup_{\boldsymbol{\Sigma}\in\mathbb{S}^m}\ & \mathrm{tr}(\boldsymbol{\Sigma}\boldsymbol{L}) \\ \text{s.t.}\ & \|\boldsymbol{\Sigma} - \widehat{\boldsymbol{\Sigma}}_n\|_F^2 \leq \rho_2^2.\end{aligned} \tag{25}$$

The Lagrangian function associated with Problem (25) is given by $\mathscr{L}(\boldsymbol{\Sigma}, \tau) = -\mathrm{tr}(\boldsymbol{\Sigma}\boldsymbol{L}) + \tau(\|\boldsymbol{\Sigma} - \widehat{\boldsymbol{\Sigma}}_n\|_F^2 - \rho_2^2)$. Since Problem (25) is convex and satisfies the Slater condition, its associated KKT conditions, which are given by

$$-\boldsymbol{L} + 2\tau(\boldsymbol{\Sigma} - \widehat{\boldsymbol{\Sigma}}_n) = \mathbf{0}, \tag{26a}$$

$$\|\boldsymbol{\Sigma} - \widehat{\boldsymbol{\Sigma}}_n\|_F^2 \leq \rho_2^2, \tag{26b}$$

$$\tau\left(\|\boldsymbol{\Sigma} - \widehat{\boldsymbol{\Sigma}}_n\|_F^2 - \rho_2^2\right) = 0,\ \ \tau \geq 0, \tag{26c}$$

are necessary and sufficient for optimality. Now, observe that we must have $\tau > 0$, for otherwise $\boldsymbol{L} = \mathbf{0}$ by (26a), which contradicts the fact that $\boldsymbol{L} \in \mathcal{L}_s$ satisfies $\mathrm{tr}(\boldsymbol{L}) = 2s > 0$. Consequently, we have $\|\boldsymbol{\Sigma} - \widehat{\boldsymbol{\Sigma}}_n\|_F^2 = \rho_2^2$ by (26c). This, together with (26a), implies that

$$\boldsymbol{\Sigma} = \boldsymbol{\Sigma}^* := \widehat{\boldsymbol{\Sigma}}_n + \frac{\rho_2}{\|\boldsymbol{L}\|_F}\boldsymbol{L}$$

is an optimal solution to Problem (25). Since $\widehat{\boldsymbol{\Sigma}}_n, \boldsymbol{L} \in \mathbb{S}_+^m$ and $\rho_2 > 0$, we have $\boldsymbol{\Sigma}^* \in \mathbb{S}_+^m$. It follows that $\boldsymbol{\Sigma}^*$ is also optimal for Problem (10) and

$$\varphi_2(\boldsymbol{L}) = \mathrm{tr}(\boldsymbol{\Sigma}^*\boldsymbol{L}) = \mathrm{tr}(\widehat{\boldsymbol{\Sigma}}_n\boldsymbol{L}) + \rho_2\|\boldsymbol{L}\|_F.$$

C. Proof of Proposition 3

Using the definition of $\boldsymbol{L}^*$, for any $\kappa > 0$, we compute

$$\begin{aligned}\boldsymbol{L}^* &= \operatorname*{argmin}_{\boldsymbol{L}\in\mathcal{L}_{\bar{s}}} \left\{\frac{1}{n}\mathrm{tr}(\boldsymbol{X}^\top\boldsymbol{L}\boldsymbol{X}) + 2\bar{\rho}_1\|\boldsymbol{L}^{1/2}\widehat{\boldsymbol{\mu}}_n\|_2 \right. \\ &\qquad \left. + \bar{\rho}_2\|\boldsymbol{L}\|_F + h(\boldsymbol{L})\right\} \\ &= \kappa \cdot \operatorname*{argmin}_{\boldsymbol{L}'\in\frac{1}{\kappa}\mathcal{L}_{\bar{s}}} \left\{\frac{1}{n}\mathrm{tr}(\boldsymbol{X}^\top(\kappa\boldsymbol{L}')\boldsymbol{X}) + 2\bar{\rho}_1\|(\kappa\boldsymbol{L}')^{1/2}\widehat{\boldsymbol{\mu}}_n\|_2 \right. \\ &\qquad \left. + \bar{\rho}_2\|\kappa\boldsymbol{L}'\|_F + h(\kappa\boldsymbol{L}')\right\} \\ &= \kappa \cdot \operatorname*{argmin}_{\boldsymbol{L}'\in\mathcal{L}_{\bar{s}/\kappa}} \left\{\frac{\kappa}{n}\mathrm{tr}(\boldsymbol{X}^\top\boldsymbol{L}'\boldsymbol{X}) + 2\sqrt{\kappa}\bar{\rho}_1\|(\boldsymbol{L}')^{1/2}\widehat{\boldsymbol{\mu}}_n\|_2 \right. \\ &\qquad \left. + \kappa\bar{\rho}_2\|\boldsymbol{L}'\|_F + h(\boldsymbol{L}')\right\} \qquad (27) \\ &= \kappa \cdot \operatorname*{argmin}_{\boldsymbol{L}'\in\mathcal{L}_{\bar{s}/\kappa}} \left\{\frac{1}{n}\mathrm{tr}(\boldsymbol{X}^\top\boldsymbol{L}'\boldsymbol{X}) + 2\frac{\bar{\rho}_1}{\sqrt{\kappa}}\|(\boldsymbol{L}')^{1/2}\widehat{\boldsymbol{\mu}}_n\|_2 \right. \\ &\qquad \left. + \bar{\rho}_2\|\boldsymbol{L}'\|_F + \frac{1}{\kappa}h(\boldsymbol{L}')\right\},\end{aligned}$$

where (27) follows from (12) and the fact that

$$\begin{aligned}\frac{1}{\kappa}\mathcal{L}_{\bar{s}} &= \left\{\begin{array}{c}\boldsymbol{L}_{ij} \leq 0 \text{ for } i \neq j, \\ \frac{1}{\kappa}\boldsymbol{L} \in \mathbb{S}^m : \boldsymbol{L}\mathbf{1} = \mathbf{0}, \\ \mathrm{tr}(\boldsymbol{L}) = 2\bar{s}\end{array}\right\} \\ &= \left\{\begin{array}{c}\boldsymbol{L}'_{ij} \leq 0 \text{ for } i \neq j, \\ \boldsymbol{L}' \in \mathbb{S}^m : \boldsymbol{L}'\mathbf{1} = \mathbf{0}, \\ \mathrm{tr}(\boldsymbol{L}') = 2\bar{s}/\kappa\end{array}\right\} = \mathcal{L}_{\bar{s}/\kappa}.\end{aligned}$$

Upon setting $\kappa = \bar{s}$, the desired result follows.

REFERENCES

[1] A. Ortega, P. Frossard, J. Kovačević, J. M. Moura, and P. Vandergheynst, "Graph signal processing: Overview, challenges, and applications," *Proc. IEEE*, vol. 106, no. 5, pp. 808–828, 2018.

[2] G. Mateos, S. Segarra, A. G. Marques, and A. Ribeiro, "Connecting the dots: Identifying network structure via graph signal processing," *IEEE Signal Process. Mag.*, vol. 36, no. 3, pp. 16–43, 2019.

[3] X. Dong, D. Thanou, M. Rabbat, and P. Frossard, "Learning graphs from data: A signal representation perspective," *IEEE Signal Process. Mag.*, vol. 36, no. 3, pp. 44–63, 2019.

[4] A. J. Smola and R. Kondor, "Kernels and regularization on graphs," in *Learning Theory and Kernel Machines*, ser. Lecture Notes in Artificial Intelligence, B. Schölkopf and M. K. Warmuth, Eds., vol. 2777, 2003, pp. 144–158.

[5] D. Zhou and B. Schölkopf, "A regularization framework for learning from graph data," in *Proc. ICML-2004 Workshop on Statistical Relational Learning and Its Connections to Other Fields*, 2004.

[6] C. Hu, L. Cheng, J. Sepulcre, G. El Fakhri, Y. M. Lu, and Q. Li, "A graph theoretical regression model for brain connectivity learning of Alzheimer's disease," in *Proc. 2013 IEEE 10th ISBI*, 2013, pp. 616–619.

[7] X. Dong, D. Thanou, P. Frossard, and P. Vandergheynst, "Learning Laplacian matrix in smooth graph signal representations," *IEEE Trans. Signal Process.*, vol. 64, no. 23, pp. 6160–6173, 2016.

[8] V. Kalofolias, "How to learn a graph from smooth signals," in *Proc. 19th AISTATS*, 2016, pp. 920–929.

[9] S. P. Chepuri, S. Liu, G. Leus, and A. O. Hero, "Learning sparse graphs under smoothness prior," in *Proc. 2017 IEEE ICASSP*, 2017, pp. 6508–6512.

[10] V. Kalofolias and N. Perraudin, "Large scale graph learning from smooth signals," in *ICLR*, 2019.

[11] P. Berger, G. Hannak, and G. Matz, "Efficient graph learning from noisy and incomplete data," *IEEE Trans. Signal Inf. Process. Netw.*, vol. 6, pp. 105–119, 2020.

[12] D. Kuhn, P. Mohajerin Esfahani, V. A. Nguyen, and S. Shafieezadeh-Abadeh, "Wasserstein distributionally robust optimization: Theory and applications in machine learning," in *INFORMS TutORials in Operations Research*, 2019, pp. 130–166.

[13] F. Luo and S. Mehrotra, "Distributionally robust optimization with decision dependent ambiguity sets," *Optim. Lett.*, vol. 14, no. 8, pp. 2565–2594, 2020.

[14] X. Zhang, Y. Xu, Q. Liu, Z. Liu, J. Lu, and Q. Wang, "Robust graph learning under Wasserstein uncertainty," *arXiv preprint arXiv:2105.04210*, 2021.

[15] J. Li, S. Huang, and A. M.-C. So, "A first-order algorithmic framework for Wasserstein distributionally robust logistic regression," in *Adv. Neural Inf. Process. Syst.*, vol. 32, 2019, pp. 3937–3947.

[16] J. Li, C. Chen, and A. M.-C. So, "Fast epigraphical projection-based incremental algorithms for Wasserstein distributionally robust support vector machine," in *Adv. Neural Inf. Process. Syst.*, vol. 33, 2020, pp. 4029–4039.

[17] H. E. Egilmez, E. Pavez, and A. Ortega, "Graph learning from data under Laplacian and structural constraints," *IEEE J. Sel. Top. Signal Process.*, vol. 11, no. 6, pp. 825–841, 2017.

[18] A. M.-C. So, "Moment inequalities for sums of random matrices and their applications in optimization," *Math. Program.*, vol. 130, no. 1, pp. 125–151, 2011.

[19] C. Jin, P. Netrapalli, R. Ge, S. M. Kakade, and M. I. Jordan, "A short note on concentration inequalities for random vectors with subGaussian norm," *arXiv preprint arXiv:1902.03736*, 2019.

[20] J. Duchi and H. Namkoong, "Variance-based regularization with convex objectives," *J. Mach. Learn. Res.*, vol. 20, no. 68, pp. 1–55, 2019.

[21] S. Shafieezadeh-Abadeh, D. Kuhn, and P. Mohajerin Esfahani, "Regularization via mass transportation," *J. Mach. Learn. Res.*, vol. 20, no. 103, pp. 1–68, 2019.

[22] S. Kumar, J. Ying, J. V. de Miranda Cardoso, and D. P. Palomar, "A unified framework for structured graph learning via spectral constraints." *J. Mach. Learn. Res.*, vol. 21, no. 22, pp. 1–60, 2020.

[23] P. Billingsley, *Probability and Measure*, 3rd ed., ser. Wiley Series in Probability and Mathematical Statistics. John Wiley & Sons, Inc., 1995.

[24] L. Condat, "Fast projection onto the simplex and the l_1 ball," *Math. Program.*, vol. 158, no. 1-2, pp. 575–585, 2016.

[25] H. Attouch, J. Bolte, and B. F. Svaiter, "Convergence of descent methods for semi-algebraic and tame problems: Proximal algorithms, forward-backward splitting, and regularized Gauss–Seidel methods," *Math. Program.*, vol. 137, no. 1-2, pp. 91–129, 2013.

[26] H. Attouch, J. Bolte, P. Redont, and A. Soubeyran, "Proximal alternating minimization and projection methods for nonconvex problems: An approach based on the Kurdyka–Łojasiewicz inequality," *Math. Oper. Res.*, vol. 35, no. 2, pp. 438–457, 2010.

[27] S. Bonettini, M. Prato, and S. Rebegoldi, "A block coordinate variable metric linesearch based proximal gradient method," *Comput. Optim. Appl.*, vol. 71, no. 1, pp. 5–52, 2018.

[28] X. Wang, C. Yao, H. Lei, and A. M.-C. So, "An efficient alternating direction method for graph learning from smooth signals," in *Proc. 2021 IEEE ICASSP*, 2021, pp. 5380–5384.

[29] C. D. Manning, P. Raghavan, and H. Schütze, *Introduction to Information Retrieval*. Cambridge University Press, 2008.

[30] N. Perraudin, J. Paratte, D. Shuman, L. Martin, V. Kalofolias, P. Vandergheynst, and D. K. Hammond, "GSPBOX: A toolbox for signal processing on graphs," *arXiv preprint arXiv:1408.5781*, 2014.

[31] U. von Luxburg, "A tutorial on spectral clustering," *Stat. Comput.*, vol. 17, no. 4, pp. 395–416, 2007.

[32] J. J. Hull, "A database for handwritten text recognition research," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 16, no. 5, pp. 550–554, 1994.

[33] S. Nene, S. Nayar, and H. Murase, "Columbia University image library (COIL-20)," *Technical Report CUCS-005-96*, 1996.

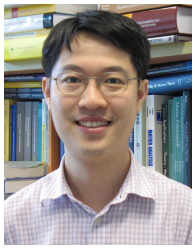
[34] M. Meila and J. Shi, "Learning segmentation by random walks," in *Adv. Neural Inf. Process. Syst.*, vol. 13, 2000, pp. 837–843.

[35] D. Xu and Y. Tian, "A comprehensive survey of clustering algorithms," *Ann. Data Sci.*, vol. 2, no. 2, pp. 165–193, 2015.

[36] M. Goldberg, "Equivalence constants for l_p norms of matrices," *Linear Multilinear Algebra*, vol. 21, no. 2, pp. 173–179, 1987.

Xiaolu Wang received the BEng degree in Telecommunications Engineering and the MSE degree in Communication and Information Systems, both from Xidian University, Xi'an, China, and the PhD degree in Systems Engineering and Engineering Management (SEEM) from The Chinese University of Hong Kong (CUHK), Hong Kong SAR, China. He is currently a research associate with the Department of SEEM, CUHK. His research interests include data science, networks, and optimization.

Yuen-Man Pun is currently a postdoctoral fellow with the CIICADA Lab, School of Engineering at Australian National University. Prior to that, she received a PhD degree in Systems Engineering and Engineering Management (SEEM) in 2022, an MPhil degree in SEEM in 2018, and a BSc degree in Mathematics in 2016, all from The Chinese University of Hong Kong, Hong Kong SAR, China. Her research focuses on algorithmic design and analysis and its applications in data science and signal processing.



Anthony Man-Cho So (Senior Member, IEEE) received the B.S.E. degree in computer science from Princeton University, Princeton, NJ, USA, with minors in applied and computational mathematics, engineering and management systems, and German language and culture; the M.Sc. degree in computer science and the Ph.D. degree in computer science with a Ph.D. minor in mathematics from Stanford University, Stanford, CA, USA. He is currently Dean of the Graduate School, Deputy Master of Morningside College, and a Professor in the Department of Systems Engineering and Engineering Management at The Chinese University of Hong Kong (CUHK), Hong Kong SAR, China. His research interests include optimization theory and its applications in various areas of science and engineering, including computational geometry, machine learning, signal processing, and statistics.

Dr. So was appointed as an Outstanding Fellow of the Faculty of Engineering at CUHK in 2019. He is the recipient of a number of research and teaching awards, including the 2022 University Grants Committee (UGC) Teaching Award, 2018 IEEE Signal Processing Society Best Paper Award, 2015 IEEE Signal Processing Society Signal Processing Magazine Best Paper Award, 2014 IEEE Communications Society Asia-Pacific Outstanding Paper Award, 2013 CUHK Vice-Chancellor's Exemplary Teaching Award, and 2010 Institute for Operations Research and the Management Sciences (INFORMS) Optimization Society Optimization Prize for Young Researchers. He currently serves on the Editorial Boards of *Journal of Global Optimization*, *Mathematical Programming*, *Optimization Methods and Software*, and *SIAM Journal on Optimization*. He was also the Lead Guest Editor of the Special Issue on Non-Convex Optimization for Signal Processing and Machine Learning of the IEEE SIGNAL PROCESSING MAGAZINE.

Supplementary Material for *Distributionally Robust Graph Learning from Smooth Signals under Moment Uncertainty*

Xiaolu Wang, Yuen-Man Pun, and Anthony Man-Cho So

Tables 1, 2, and 3 below contain the hyperparameter values used in our MUGL models for generating the results in Tables V, VI, and VII, respectively.

		ρ_1	ρ_2	α
Gaussian	MUGL-o	0.1	3.0	–
	MUGL-l	0.1	2.9	1.0
ER	MUGL-o	0.05	0.2	–
	MUGL-l	0.05	0.8	0.15
PA	MUGL-o	0.04	0.1	–
	MUGL-l	0.6	0.1	0.01

(a) $m = 20, n = 30, \epsilon = 0.1$

		ρ_1	ρ_2	α
Gaussian	MUGL-o	0.02	1.5	–
	MUGL-l	0.01	1.4	2.0
ER	MUGL-o	0.01	0.5	–
	MUGL-l	0.01	0.5	0.1
PA	MUGL-o	0.01	1.0	–
	MUGL-l	0.01	0.05	0.1

(b) $m = 20, n = 80, \epsilon = 0.1$

		ρ_1	ρ_2	α
Gaussian	MUGL-o	0.2	6.0	–
	MUGL-l	0.24	6.9	6.5
ER	MUGL-o	0.05	5.0	–
	MUGL-l	0.02	3.5	0.8
PA	MUGL-o	0.01	4.0	–
	MUGL-l	0.01	0.05	0.1

(c) $m = 20, n = 30, \epsilon = 1$

		ρ_1	ρ_2	α
Gaussian	MUGL-o	0.15	5.6	–
	MUGL-l	0.2	6.6	7.0
ER	MUGL-o	0.01	1.2	–
	MUGL-l	0.1	1.3	3.0
PA	MUGL-o	0.01	0.4	–
	MUGL-l	0.1	0.5	2.0

(d) $m = 20, n = 80, \epsilon = 1$

Table 1: MUGL parameters used for Table II

	ρ_1	ρ_2	α
MUGL-o	20	600	–
MUGL-l	1.7	630	360

Table 2: MUGL parameters used for Table III

	ρ_1	ρ_2	α
MUGL-o	0.1	0.2	–
MUGL-l	1.0	20	110

(a) USPS

	ρ_1	ρ_2	α
MUGL-o	0.1	0.2	–
MUGL-l	0.11	0.18	0.001

(b) COIL-20

Table 3: MUGL parameters used for Table IV