Analysis of Network Lasso for Semi-Supervised Regression

A. Jung and N. Vesselinova
Department of Computer Science, Aalto University, Finland

Abstract

We apply network Lasso to semi-supervised regression problems involving network-structured data. This approach lends quite naturally to highly scalable learning algorithms in the form of message passing over an empirical graph which represents the network structure of the data. By using a simple non-parametric regression model, which is motivated by a clustering hypothesis, we provide an analysis of the estimation error incurred by network Lasso. This analysis reveals conditions on the network structure and the available training data which guarantee network Lasso to be accurate. Remarkably, the accuracy of network Lasso is related to the existence of sufficiently large network flows over the empirical graph. Thus, our analysis reveals a connection between network Lasso and maximum flow problems.

1 INTRODUCTION

The datasets arising in many applications, ranging from image processing to cyber security carry an intrinsic network structure. In particular, those datasets can be represented conveniently using an empirical graph Chapelle et al. [2006]. The nodes of this empirical graph represent individual data points, which are connected by edges according to some domain-specific notion of similarity.

On top of the network structure, datasets carry additional information in the form of labels for the individual data points. Since the acquisition of label information is often expensive (requiring manual labour), we typically have access to labels of few data points only. Moreover, the available label information will often be noisy due to measurement (labeling) errors.

The available incomplete label information might still suffice to allow for accurate machine learning by exploiting the tendency of labels to conform to the underlying network structure. Indeed, many successful learning methods rely on a clustering hypothesis which requires well-connected data points to have similar labels Bishop [2006], Chapelle et al. [2006].

Various generalisations of the least absolute shrinkage and selection operator (Lasso) from sparse vectors to network-structured data have been proposed recently by Tibshirani et al. [2005], Sharpnack et al. [2012]. In particular, the "network Lasso" (nLasso) Hallac et al. [2015] provides an optimization framework for a wide range of learning problems (regression and classification) involving network-structured datasets. While efficient implementations of nLasso for particular learning problems have been proposed (see M.Yamada et al. [2017]), only little is known about the statistical performance of nLasso methods for general learning problems involving partially labelled network-structure data.

Contribution. In this paper, we apply a generalization of the concept of a compatibility condition, which has been championed by Bühlmann and van de Geer [2011], van de Geer [2007] for characterizing the performance of Lasso methods, to learning problems involving network structured data. Various forms of such "network compatibility conditions" have been studied recently by Jung et al. [2018], Ortelli and van de Geer [2018], Hütter and Rigollet [2016]. Here, we use a particular form of a network compatibility condition to characterize the performance of nLasso for semi-supervised regression problems using squared error loss. The nLasso provides an efficient method for non-parametric regression by leveraging the underlying network structure Kovac and Smith [2012]. Our results give a precise characterization of the statistical performance of such methods and their dependence on the network topology. The closest to our work is Hütter and Rigollet [2016], which studies the statistical properties of nLasso applied to denoising a fully

Proceedings of the 22[nd] International Conference on Artificial Intelligence and Statistics (AISTATS) 2019, Naha, Okinawa, Japan. PMLR: Volume 89.

observed graph signal. In contrast, our analysis allows for nLasso having access only to the signal values of a small subset (the training set) of nodes, which is relevant for semi-supervised learning problems (see Chapelle et al. [2006]).

Outline. This paper is organized as follows: in Section 2, we formalize the problem of semi-supervised learning for network-structured data using a probabilistic model for the observations, which is based on exponential families. Based on this generic probabilistic model, we then show in Section 3 how to apply network Lasso to learn a predictor for all data points based on knowledge of noisy labels for few data points. Our main result is discussed in Section 4, where we present a bound on the estimation error of nLasso. This bound depends on the network compatibility condition which, in turn, relates to the connectivity of sampled nodes.

Notation. We use boldface upper and lower case letters to denote matrices and vectors, respectively. Given a matrix $\mathbf{W}$ we define its supremum norm as $\|\mathbf{W}\|_\infty := \max_{i,j} |W_{i,j}|$. The nullspace (or kernel) of a matrix $\mathbf{L}$ is denoted $\ker\{\mathbf{L}\} := \{\mathbf{x} : \mathbf{L}\mathbf{x} = \mathbf{0}\}$. The pseudo-inverse of a diagonal matrix $\mathbf{A}$ is denoted $\mathbf{A}^\dagger$ and obtained by inverting the non-zero diagonal entries of $\mathbf{A}$ and leaving the zero entries. The pseudo-inverse of an arbitrary matrix $\mathbf{D}$ is obtained via its singular value decomposition $\mathbf{D} = \mathbf{U}\mathbf{\Lambda}\mathbf{V}^T$ as $\mathbf{D}^\dagger = \mathbf{U}\mathbf{\Lambda}^\dagger\mathbf{V}^T$. Given a finite set $\mathcal{V}$, we denote the complement of a subset $\mathcal{M} \subseteq \mathcal{V}$ as $\overline{\mathcal{M}}$.

2 Problem Formulation

We consider network-structured datasets, which are represented by an empirical graph $\mathcal{G}=(\mathcal{V},\mathcal{E},\mathbf{W})$. The nodes $\mathcal{V} = \{1,\ldots,N\}$ of the empirical graph represent individual data points. The undirected edges $\mathcal{E}$ encode domain-specific notions of similarity between data points. The non-negative entries $W_{i,j}$ of the weight matrix $\mathbf{W} \in \mathbb{R}_+^{N\times N}$ quantify the level of similarity between connected nodes. The weight $W_{i,j}$ is non-zero only if nodes $i,j \in \mathcal{V}$ are connected by an edge $\{i,j\} \in \mathcal{E}$.

In what follows, without loss of generality, we assume that the empirical graph is simple (without self loops) and connected. Therefore, since there are no self loops, the weight matrix is such that $W_{i,i} = 0$ for every node $i \in \mathcal{V}$.

2.1 Laplacian and Incidence Matrix

The structure of an empirical graph $\mathcal{G}$ can be characterized using the graph Laplacian matrix

$$\mathbf{L} = \mathbf{\Lambda} - \mathbf{W}, \tag{1}$$

with the weight matrix $\mathbf{W}$ and the diagonal "degree matrix"

$$\mathbf{\Lambda} = \mathrm{diag}\{d_1,\ldots,d_N\} \in \mathbb{R}^{N\times N}.$$

The diagonal elements of $\mathbf{\Lambda}$ are the weighted node degrees $d_i := \sum_{\{j,i\}\in\mathcal{E}} W_{i,j}$.

We denote the (ordered) non-negative eigenvalues of the Laplacian matrix $\mathbf{L}$ by $\lambda_1 = 0 \leq \ldots \leq \lambda_N$. The eigenvalues of $\mathbf{L}$ provide insight into the connectivity structure of the graph $\mathcal{G}$. A graph $\mathcal{G}$ is connected if and only if $\lambda_2 > 0$ in which case the nullspace of $\mathbf{L}$ is a one-dimensional subspace spanned by the constant graph signal with value $x_i = 1$ for every node $i \in \mathcal{V}$. The spectral gap $\rho(\mathcal{G}) := \lambda_2$ quantifies the connectivity of the graph $\mathcal{G}$. If $\rho(\mathcal{G})$ is close to zero, the graph $\mathcal{G}$ can be cut into two disconnected subgraphs without removing too many edges Spielman [2012].

Another important matrix assigned to an empirical graph is the incidence matrix. To this end, we (arbitrarily) orient the empirical graph $\mathcal{G} = (\mathcal{E},\mathcal{V},\mathbf{W})$ by specifying for each edge $e = \{i,j\}$ one node as the head e^+ and the other node as the tail e^-. We define the incidence matrix $\mathbf{D} \in \mathbb{R}^{\mathcal{E}\times\mathcal{V}}$ element-wise as

$$D_{e,i} = \begin{cases} \sqrt{W_e} & \text{if } i = e^+ \\ -\sqrt{W_e} & \text{if } i = e^- \\ 0 & \text{else.} \end{cases} \tag{2}$$

We highlight that the exact choice of orientation for the undirected edges in the empirical graph $\mathcal{G}$ has no effect on our results. The use of an orientation only serves a notational convenience provided by the incidence matrix $\mathbf{D}$.

The incidence matrix $\mathbf{D}$ is closely related to the graph Laplacian $\mathbf{L}$. Indeed, both matrices have the same nullspace $\ker\{\mathbf{D}\} = \ker\{\mathbf{L}\}$. Moreover, the spectrum of $\mathbf{D}\mathbf{D}^T$ coincides with the spectrum of $\mathbf{L}^{(\mathcal{G})}$. The columns $\mathbf{s}_j$ of the pseudo-inverse $\mathbf{D}^\dagger = (\mathbf{s}_1,\ldots,\mathbf{s}_{|\mathcal{E}|})$ of $\mathbf{D}$ satisfy

$$\|\mathbf{s}_j\| \leq \sqrt{2\|\mathbf{W}\|_\infty}/\rho(\mathcal{G}). \tag{3}$$

This bound can be verified using the identity $\mathbf{D}^\dagger = (\mathbf{D}\mathbf{D}^T)^\dagger\mathbf{D}^T$ and well-known vector norm inequalities (see, e.g., Horn and Johnson [1985]).

2.2 Linear Regression

In addition to the network structure, which is encoded by the empirical graph $\mathcal{G}$, datasets typically convey

additional information. This additional information comes in the form of labels y_i associated with individual data points $i \in \mathcal{V}$.

We model the labels y_i of data points $i \in \mathcal{V}$ as random variables whose probability distribution is parametrized by a graph signal $\bar{\mathbf{x}} : \mathcal{V} \to \mathbb{R}$. In particular, we use the linear model

$$y_i = \bar{x}_i + \varepsilon_i, \tag{4}$$

with some unknown underlying graph signal $\bar{\mathbf{x}}$. The noise terms ε_i in (4), which are modelled as i.i.d. Gaussian random variables with zero-mean and variance σ^2, cover any modelling or measurement (labeling) errors. We will use the following tail bound

$$\mathrm{P}\{|y - \mathrm{E}\{y\}| \geq \eta\} \leq 2\exp\left(-N^2\eta^2 / \left(2\sigma^2 \sum_{i=1}^{N} w_i^2\right)\right), \tag{5}$$

for the weighted sum $y = (1/N)\sum_{i=1}^{N} y_i w_i$ with arbitrary but fixed weights $w_i \in \mathbb{R}$.

The graph signal $\bar{\mathbf{x}}$ in (4) assigns a real number $\bar{x}_i \in \mathbb{R}$ to each node $i \in \mathcal{V}$. We can think of a graph signal also as a vector whose entries are indexed by the nodes $i \in \mathcal{V}$. The space of all graph signals constitutes an Euclidean space $\mathbb{R}^{\mathcal{V}}$. It will be convenient to define, for a subset $\mathcal{M} \subseteq \mathcal{V}$ of size $M := |\mathcal{M}|$, the norm

$$\|\mathbf{x}\|_{\mathcal{M}} := \sqrt{(1/M)\sum_{i \in \mathcal{M}} x_i^2}. \tag{6}$$

Since acquiring labels is costly, we consider having access to the (noisy) labels y_i (see (4)) only for the nodes in a (small) training set

$$\mathcal{M} = \{i_1, \ldots, i_M\} \text{ with } M \ll N. \tag{7}$$

2.3 Clustering Hypothesis

Our approach to learning the graph signal $\bar{\mathbf{x}}$ in (4) from the labels y_i of the nodes in the training set $\mathcal{M}$, is based on the assumption that the graph signal $\bar{\mathbf{x}}$ is clustered in the sense of being constant over well-connected subsets (clusters) of nodes. This clustering hypothesis conforms to the finding that the labels of data points arising in many application domains, such as signal or image processing as well as social networks, are similar if the data points are well-connected in the empirical graph (see Chapelle et al. [2006]).

We measure the amount by which a graph signal $\mathbf{x}$ conforms with the cluster structure of the empirical graph $\mathcal{G}$ using the (weighted) total variation (TV)

$$\|\mathbf{x}\|_{\mathrm{TV}} := \sum_{\{i,j\} \in \mathcal{E}} \sqrt{W_{ij}}\,|x_j - x_i|\,. \tag{8}$$

Indeed, a graph signal $\mathbf{x}$ has a small TV only if the signal values x_i are approximately constant over well connected subsets (clusters) of nodes. Such a "clustering hypothesis" (or variations thereof) motivates many methods for (semi-) supervised learning Chapelle et al. [2006].

If we orient the empirical graph, we can represent the TV using the incidence matrix $\mathbf{D}$ (see (2) and (8)) as

$$\|\mathbf{x}\|_{\mathrm{TV}} = \|\mathbf{D}\mathbf{x}\|_1. \tag{9}$$

It will be convenient to define a shorthand for the TV over a subset $\mathcal{S} \subseteq \mathcal{E}$ of edges as

$$\|\mathbf{x}\|_{\mathcal{S}} := \sum_{\{i,j\} \in \mathcal{S}} \sqrt{W_{ij}}\,|x_j - x_i|\,. \tag{10}$$

One of our main contributions (see Section 4) is a precise analysis of the ability of nLasso to learn clustered graph signals. In particular, our analysis is based on the following simple model for clustered (piece-wise constant) graph signals (see Chen et al. [2017]):

$$x_i = \sum_{\mathcal{C} \in \mathcal{F}} a_{\mathcal{C}} \mathcal{I}_{\mathcal{C}}[i]. \tag{11}$$

Here, $a_{\mathcal{C}} \in \mathbb{R}$ is the signal value of cluster $\mathcal{C}$ and we used the indicator signal

$$\mathcal{I}_{\mathcal{C}}[i] = \begin{cases} 1 & \text{for } i \in \mathcal{C} \\ 0 & \text{otherwise.} \end{cases}$$

The model (11) involves a partitioning $\mathcal{F} = \{\mathcal{C}_1, \ldots, \mathcal{C}_{|\mathcal{F}|}\}$ of the nodes $\mathcal{V}$ into disjoint subsets $\mathcal{C}_l$. We assume that the subgraph induced by any cluster $\mathcal{C}_l$ is connected.

We emphasize that our analysis allows for an arbitrary partitioning $\mathcal{F} = \{\mathcal{C}_1, \ldots, \mathcal{C}_{|\mathcal{F}|}\}$ used to define the model (11). However, our results are most useful (i.e., the error bound (18) will be tighter) if the partition conforms with the "intrinsic (cluster) structure" of the empirical graph $\mathcal{G}$. In particular, we focus on partitions $\mathcal{F}$ such that the cluster boundaries

$$\partial\mathcal{F} := \{\{i,j\} \in \mathcal{E} : i \in \mathcal{C}, j \in \mathcal{C}' (\neq \mathcal{C})\}$$

satisfy $\sum_{\{i,j\} \in \partial\mathcal{F}} \sqrt{W_{i,j}} \ll \sum_{\{i,j\} \in \overline{\partial\mathcal{F}}} \sqrt{W_{i,j}}$.

It will be useful to define the spectral gap of a partitioning $\mathcal{F} = \{\mathcal{C}_1, \ldots, \mathcal{C}_{|\mathcal{F}|}\}$ as

$$\rho(\mathcal{F}) := \min_{\mathcal{C}_l \in \mathcal{F}} \rho(\mathcal{C}_l). \tag{12}$$

Here, $\rho(\mathcal{C}_l)$ denotes the spectral gap of the subgraph induced by the cluster $\mathcal{C}_l$.

3 THE NETWORK LASSO

It is sensible to learn a graph signal $\hat{\mathbf{x}} \in \mathbb{R}^{\mathcal{V}}$ based on (few) labels $\{y_i\}_{i\in\mathcal{M}}$ by maximizing the probability ("evidence") $\mathrm{P}\{\{y_i\}_{i\in\mathcal{M}}; \mathbf{x}\}$ of observing them under the probabilistic model (4) for the labels. This is equivalent to minimizing the empirical error:

$$\widehat{E}(\mathbf{x}) := (1/M) \sum_{i\in\mathcal{M}} (y_i - x_i)^2. \tag{13}$$

The criterion (13) by itself is not sufficient for guiding the learning of a graph signal based on few labels $\{y_i\}_{i\in\mathcal{M}}$, since it ignores the signal values x_i for $i \in \overline{\mathcal{M}}$.

In order to learn an entire graph signal $\bar{\mathbf{x}}$ from the incomplete information provided by the initial labels $\{y_i\}_{i\in\mathcal{M}}$, we need to impose some structure on the true underlying graph signal $\bar{\mathbf{x}}$ as well as any learnt graph signal $\hat{\mathbf{x}}$ which should accurately resemble $\bar{\mathbf{x}}$. This additional structure is provided by the empirical graph $\mathcal{G}$. In particular, we assume that any reasonable graph signal $\hat{\mathbf{x}}$ needs to conform with the *cluster structure* of $\mathcal{G}$ (see Newman [2010]).

We are led quite naturally to learning a graph signal $\hat{\mathbf{x}}$ by balancing a small empirical error (risk) $\widehat{E}(\hat{\mathbf{x}})$ (see (13)) with a small TV $\|\hat{\mathbf{x}}\|_{\mathrm{TV}}$ (see (8)). Thus, we arrive at the following *regularized empirical risk minimization*

$$\hat{\mathbf{x}} \in \operatorname*{arg\,min}_{\mathbf{x}\in\mathbb{R}^{\mathcal{V}}} \widehat{E}(\mathbf{x}) + \lambda \|\mathbf{x}\|_{\mathrm{TV}}. \tag{14}$$

The parameter $\lambda > 0$ in (14) allows to trade off a small TV $\|\hat{\mathbf{x}}\|_{\mathrm{TV}}$ against a small empirical error. Choosing a small value of λ will result in a graph signal $\hat{\mathbf{x}}$ with small empirical error $\widehat{E}(\hat{\mathbf{x}})$ (see (13)), while choosing a large value of λ favours $\hat{\mathbf{x}}$ with small TV $\|\hat{\mathbf{x}}\|_{\mathrm{TV}}$ (being more clustered).

The learning problem (14) is a particular instance of the nLasso introduced in Hallac et al. [2015] which allows for efficient implementations using modern convex optimization methods Parikh and Boyd [2013], Boyd et al. [2010]. In particular, we obtain Algorithm 1 by applying the primal-dual method proposed by Pock and Chambolle [2011] to

$$\begin{aligned}\hat{\mathbf{x}} &\in \operatorname*{arg\,min}_{\mathbf{x}\in\mathbb{R}^{\mathcal{V}}} \widehat{E}(\mathbf{x}) + \lambda \|\mathbf{D}\mathbf{x}\|_1 \\ &= \operatorname*{arg\,min}_{\mathbf{x}\in\mathbb{R}^{\mathcal{V}}} \max_{\|\mathbf{u}\|_\infty \leq 1} \widehat{E}(\mathbf{x}) + \lambda \mathbf{u}^T \mathbf{D}\mathbf{x}\end{aligned}$$

which, due to (9), is equivalent to (14).

Algorithm 1

input: $\mathbf{D} \in \mathbb{R}^{\mathcal{E}\times\mathcal{V}}$, $\mathcal{M}$, $\{y_i\}_{i\in\mathcal{M}}$, λ

initialize: $k{:=}0, \bar{\mathbf{x}}^{(0)}{=}\hat{\mathbf{x}}^{(-1)}{=}\hat{\mathbf{x}}^{(0)}{=}\hat{\mathbf{y}}^{(0)}{:=}\mathbf{0}$,
$\nu{:=}1/(\lambda M), \gamma_i{:=}\sum_{j\in\mathcal{V}} \sqrt{W_{i,j}}$
$\mathbf{\Gamma}{:=}\mathrm{diag}\{1/\gamma_i\}{\in}\mathbb{R}^{\mathcal{V}\times\mathcal{V}}$
$\mathbf{\Lambda}{:=}\mathrm{diag}\{1/(2\sqrt{W_{i,j}})\}{\in}\mathbb{R}^{\mathcal{E}\times\mathcal{E}}$,

repeat:
1: $\mathbf{x} := 2\hat{\mathbf{x}}^{(k)} - \hat{\mathbf{x}}^{(k-1)}$
2: $\hat{\mathbf{z}} := \hat{\mathbf{y}}^{(k)} + \mathbf{\Lambda}\mathbf{D}\mathbf{x}$
3: $\hat{y}_e^{(k+1)} := \hat{z}_e / \max\{1, |\hat{z}_e|\}$ for all $e{\in}\mathcal{E}$
4: $\hat{\mathbf{x}}^{(k+1)} := \hat{\mathbf{x}}^{(k)} - \mathbf{\Gamma}\mathbf{D}^T\hat{\mathbf{y}}^{(k+1)}$
5: $\hat{x}_i^{(k+1)} := \frac{2\nu y_i + \gamma_i \hat{x}_i^{(k+1)}}{2\nu+\gamma_i}$ for all $i \in \mathcal{M}$
6: $k := k+1$
7: $\bar{\mathbf{x}}^{(k)} := (1{-}1/k)\bar{\mathbf{x}}^{(k-1)}{+}(1/k)\hat{\mathbf{x}}^{(k)}$
until stopping criterion is satisfied
output: labels $\hat{x}_i := \bar{x}_i^{(k)}$ for all $i \in \mathcal{V}$

4 STATISTICAL PROPERTIES OF NETWORK LASSO

The accuracy of nLasso methods depends on how close the solutions $\hat{\mathbf{x}}$ of (14) are to the true underlying clustered graph signal $\bar{\mathbf{x}} \in \mathbb{R}^{\mathcal{V}}$ (see (4) and (11)).

In what follows, we derive a condition on the cluster structure $\mathcal{F}$ and training set $\mathcal{M}$, which guarantee any solution $\hat{\mathbf{x}}$ of (14) to be close to the underlying graph signal $\bar{\mathbf{x}}$. This condition, which we refer to as network compatibility condition (NCC) extends the concept of compatibility conditions used for analyzing Lasso methods for learning sparse vectors van de Geer and Bühlmann [2009], to network-structured data. We then show that this network compatibility condition is related to the existence of a sufficiently large network flow. The existence of such network flows indirectly characterizes the connectivity of sampled nodes $\mathcal{M}$ in different clusters via the cluster boundaries $\partial\mathcal{F}$.

4.1 Flows over the Empirical Graph

The main conceptual contribution of this paper is the insight that the accuracy of nLasso methods, aiming at solving (14), depends on the topology of the underlying empirical graph via the existence of certain *flows with demands* Kleinberg and Tardos [2006].

A flow over the empirical graph $\mathcal{G}$ is a mapping $h : \mathcal{V}\times\mathcal{V} \to \mathbb{R}$ which assigns each directed edge (i,j) the value $h(i,j)$, which can be interpreted as the amount of some quantity flowing through the edge (i,j) (see Kleinberg and Tardos [2006]).

A flow with demands has to satisfy the conservation law

$$\sum_{j \in \mathcal{N}(i)} h(i,j) = f_i, \text{ for any } i \in \mathcal{V} \tag{15}$$

with a prescribed demand f_i for each node $i \in \mathcal{V}$. Moreover, we require flows to satisfy the capacity constraints

$$|h(i,j)| \leq \sqrt{W_{i,j}} \text{ for any } (i,j) \in \overline{\partial \mathcal{F}}. \tag{16}$$

Note that the capacity constraint (16) applies only to intra-cluster edges and does not involve the boundary edges $\partial\mathcal{F}$. The flow values $h(i,j)$ at the boundary edges $(i,j) \in \partial\mathcal{F}$ take a special role in the following definition of the notion of resolving training sets.

Definition 1 *Consider an empirical graph $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \mathbf{W})$ and a partition $\mathcal{F} = \{\mathcal{C}_1, \ldots, \mathcal{C}_{|\mathcal{F}|}\}$. A (training) set $\mathcal{M} = \{i_1, \ldots, i_M\} \subseteq \mathcal{V}$ resolves $\mathcal{F}$ with constants $K, L > 0$ if, for any $b_{i,j} \in \{-1, 1\}^{\partial\mathcal{F}}$, there is a flow $h[\cdot]$ on $\mathcal{G}$ (cf. (15), (16)) with $h(i,j) = b_{i,j} L \sqrt{W_{i,j}}$ for $\{i,j\} \in \partial\mathcal{F}$ and demands (cf. (15)) $|f_i| \leq K/M$ for $i \in \mathcal{M}$ and $f_i = 0$ for $i \in \overline{\mathcal{M}}$.*

This definition requires nodes of a resolving training set to be sufficiently well connected with each boundary edge $\{i,j\} \in \partial\mathcal{F}$. In particular, we could think of injecting (absorbing) certain amounts of flow into (from) the empirical graph at the sampled nodes. At each sampled node $i \in \mathcal{M}$, we can inject (absorb) a flow of value at most K/M. The injected (absorbed) flow has to be routed from (to) the sampled nodes $\mathcal{M}$ via the intra-cluster edges $\overline{\partial\mathcal{F}}$ to (from) each boundary edge $\{i,j\} \in \partial\mathcal{F}$ such that it carries a flow value $L \cdot \sqrt{W_{i,j}}$.

Note that Definition 1 is quantitive as it involves the numerical constants K and L. Our main result stated below is an upper bound on the estimation error of nLasso methods, which depends on the value of these constants. It will turn out that resolving sampling sets with a small values of K and large values of L are beneficial for the ability of nLasso to recover the entire graph signal from noisy samples $\{y_i\}_{i \in \mathcal{M}}$ observed on the training set $\mathcal{M}$.

4.2 Linear Regression with nLasso

For the analysis of the nLasso problem (14), we will make use of the network compatibility condition (NCC) defined as follows.

Definition 2 *Consider an empirical graph $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \mathbf{W})$ with a particular partition $\mathcal{F}$ of its nodes $\mathcal{V}$. A sampling set $\mathcal{M} \subseteq \mathcal{V}$ is said to satisfy NCC with constants $K, L > 1$, if*

$$L\|\mathbf{z}\|_{\partial\mathcal{F}} \leq K\|\mathbf{z}\|_{\mathcal{M}} + \|\mathbf{z}\|_{\overline{\partial\mathcal{F}}} \tag{17}$$

for any graph signal $\mathbf{z} \in \mathbb{R}^{\mathcal{V}}$.

The NCC guarantees nLasso (14) to accurately recover graph signals of the form (11). Note that the NCC involves the partition $\mathcal{F}$ underlying the signal model (11). However, the partition is not required for the implementation of nLasso (14).

It turns out that the resolving sets (see Definition 1) satisfy the NCC.

Lemma 1 *Consider an empirical graph $\mathcal{G}$ whose nodes are partitioned as $\mathcal{F} = \{\mathcal{C}_1, \ldots, \mathcal{C}_{|\mathcal{F}|}\}$. If a set $\mathcal{M}$ resolves $\mathcal{F}$, it satisfies NCC with the same parameters K, L.*

The statement follows easily from [Jung et al., 2018, Lemma 6] and the Cauchy-Schwarz inequality, which implies $\sum_{i \in \mathcal{M}} |z_i| \leq \sqrt{M \sum_{i \in \mathcal{M}} z_i^2}$.

Our main result is that the NCC, with suitable constants L and K, implies that solutions of the nLasso problem (14) are close to the true underlying clustered graph signal $\bar{\mathbf{x}}$ (cf. (11)).

Theorem 1 *Consider an empirical graph $\mathcal{G}$, whose nodes have labels y_i distributed according to (4) with underlying clustered graph signal $\bar{\mathbf{x}}$ (11). We observe the labels on a training set $\mathcal{M}$ which satisfies the NCC with parameters $L > 3$, and $K \in (1, L-2)$ and condition number $\kappa := \frac{K+3}{L-3}$ (see Definition 2). Based on the observed noisy labels y_i, we estimate the underlying graph signal $\bar{\mathbf{x}}$ by a solution $\hat{\mathbf{x}}$ to the nLasso problem (14) for the choice $\lambda := \eta/(308\kappa^2)$ with some pre-specified error level $\eta > 0$. The probability of the nLasso error exceeding η is upper bounded as*

$$\mathrm{P}\{\|\hat{\mathbf{x}} - \bar{\mathbf{x}}\|_{\mathrm{TV}} \geq \eta\} \leq 2|\mathcal{F}| \exp\left(-\frac{|\mathcal{C}_l|\eta^2}{8 \cdot 308^2 \kappa^2 \sigma^2}\right) + 2M \exp\left(-\frac{M^2 \rho_{\mathcal{F}}^2 \eta^2}{64 \cdot 308^2 \kappa^4 \sigma^2 \|\mathbf{D}\|_{\infty}^2}\right). \tag{18}$$

The bound (18) indicates that, for a prescribed accuracy level η, the training set size M has to scale according to $\kappa^2\sigma/\rho_{\mathcal{F}}$. Thus, the sample size required by Algorithm 1 scales with the square of the condition number $\kappa = \frac{K+3}{L-3}$ (see Definition 2) and inversely with the spectral gap $\rho_{\mathcal{F}}$ of the partitioning $\mathcal{F}$. Thus, nLasso methods (14) (such as Algorithm 1) require less training data if the condition number κ is small and the spectral gap $\rho_{\mathcal{F}}$ is large. This is reasonable, since according to Lemma 1, a small condition number (NCC parameter L is large compared to K) requires the edges within clusters to have larger weights on average than the weights of the boundary edges. Moreover, it is reasonable that nLasso tends to be

more accurate for a larger spectral gap $\rho_{\mathcal{F}}$, which requires the nodes within each cluster $\mathcal{C}_l$ to be well connected. Indeed, an empirical graph $\mathcal{G}$ consisting of well-connected clusters $\mathcal{C}_l$ favours clustered graph signals, such as the true underlying graph signal $\bar{\mathbf{x}}$ in (4), to be solutions of the nLasso (14).

4.3 Proof of Theorem 1

By following the reasoning pattern in Bach [2010] and Bühlmann and van de Geer [2011], we organize the proof in two parts. The first part is to verify that, with high probability, the estimation error $\tilde{\mathbf{x}}:=\hat{\mathbf{x}}-\bar{\mathbf{x}}$ incurred by nLasso (14) is approximately clustered according to (11). The second part is to upper bound the nLasso error $\tilde{\mathbf{x}}$ using the NCC (17).

First, any solution $\hat{\mathbf{x}}$ of (14) trivially satisfies

$$(1/M)\sum_{i\in\mathcal{M}}\big[-y_i(\hat{x}_i-\bar{x}_i)+(1/2)(\hat{x}_i^2-\bar{x}_i^2)\big] \leq (\lambda/2)(\|\bar{\mathbf{x}}\|_{\rm TV}-\|\hat{\mathbf{x}}\|_{\rm TV}). \quad (19)$$

Inserting (4) into (19) yields the inequality

$$(1/M)\sum_{i\in\mathcal{M}}-2\varepsilon_i\tilde{x}_i+\lambda\|\hat{\mathbf{x}}\|_{\rm TV}\leq\lambda\|\bar{\mathbf{x}}\|_{\rm TV}, \quad (20)$$

which is satisfied by any solution $\hat{\mathbf{x}}$ of nLasso (14).

Let us, for the time being, assume the label noise ε_i (see (4)) is sufficiently small such that

$$\big|(1/M)\sum_{i\in\mathcal{M}}\varepsilon_i\tilde{x}_i\big|\leq(\lambda/2)\kappa\|\tilde{\mathbf{x}}\|_{\mathcal{M}}+(\lambda/4)\|\tilde{\mathbf{x}}\|_{\rm TV} \quad (21)$$

holds for every graph signal $\tilde{\mathbf{x}}\in\mathbb{R}^{\mathcal{V}}$.

Combining (21) with (20),

$$\|\hat{\mathbf{x}}\|_{\rm TV}\leq(1/2)\|\tilde{\mathbf{x}}\|_{\rm TV}+\|\bar{\mathbf{x}}\|_{\rm TV}+\kappa\|\tilde{\mathbf{x}}\|_{\mathcal{M}}$$

and, in turn, via the decomposition property $\|\mathbf{x}\|_{\rm TV}=\|\mathbf{x}\|_{\partial\mathcal{F}}+\|\mathbf{x}\|_{\overline{\partial\mathcal{F}}}$ (see (10)),

$$\begin{aligned}\|\hat{\mathbf{x}}\|_{\overline{\partial\mathcal{F}}}&\leq(1/2)\|\tilde{\mathbf{x}}\|_{\rm TV}+\|\bar{\mathbf{x}}\|_{\rm TV}-\|\hat{\mathbf{x}}\|_{\partial\mathcal{F}}+\kappa\|\tilde{\mathbf{x}}\|_{\mathcal{M}}\\ &\overset{(a)}{\leq}(1/2)\|\tilde{\mathbf{x}}\|_{\rm TV}+\|\bar{\mathbf{x}}\|_{\partial\mathcal{F}}-\|\hat{\mathbf{x}}\|_{\partial\mathcal{F}}+\kappa\|\tilde{\mathbf{x}}\|_{\mathcal{M}}\\ &\overset{(b)}{\leq}(1/2)\|\tilde{\mathbf{x}}\|_{\rm TV}+\|\bar{\mathbf{x}}-\hat{\mathbf{x}}\|_{\partial\mathcal{F}}+\kappa\|\tilde{\mathbf{x}}\|_{\mathcal{M}}, \quad (22)\end{aligned}$$

where step (a) is valid since the true underlying graph signal $\bar{\mathbf{x}}$ is assumed to be clustered (see (11)) which implies $\|\bar{\mathbf{x}}\|_{\rm TV}=\|\bar{\mathbf{x}}\|_{\partial\mathcal{F}}$. In step (b) we used the (reverse) triangle inequality for the semi-norm $\|\cdot\|_{\partial\mathcal{F}}$.

Inserting $\|\hat{\mathbf{x}}\|_{\overline{\partial\mathcal{F}}}=\|\tilde{\mathbf{x}}\|_{\overline{\partial\mathcal{F}}}$ into (22) yields

$$\|\tilde{\mathbf{x}}\|_{\overline{\partial\mathcal{F}}}\leq3\|\tilde{\mathbf{x}}\|_{\partial\mathcal{F}}+2\kappa\|\tilde{\mathbf{x}}\|_{\mathcal{M}}. \quad (23)$$

Thus, for sufficiently small observation noise ε_i (such that (21) is valid), the nLasso error $\tilde{\mathbf{x}}=\hat{\mathbf{x}}-\bar{\mathbf{x}}$ is approximately clustered according to (11).

Next, we control the nLasso error $\tilde{\mathbf{x}}=\hat{\mathbf{x}}-\bar{\mathbf{x}}$ (see (14)). According to (19),

$$(1/M)\sum_{i\in\mathcal{M}}\big[-2\varepsilon_i\tilde{x}_i+\tilde{x}_i^2\big]+\lambda\|\hat{\mathbf{x}}\|_{\rm TV}\leq\lambda\|\bar{\mathbf{x}}\|_{\rm TV}. \quad (24)$$

Using the (reverse) triangle inequality for the TV seminorm $\|\cdot\|_{\partial\mathcal{F}}$ (see (10)), (24) becomes

$$(1/M)\sum_{i\in\mathcal{M}}\big[-2\varepsilon_i\tilde{x}_i+\tilde{x}_i^2\big]\leq\lambda\|\tilde{\mathbf{x}}\|_{\rm TV}. \quad (25)$$

Inserting (21) into (25),

$$\begin{aligned}\|\tilde{\mathbf{x}}\|_{\mathcal{M}}^2&\leq(3/2)\lambda\|\tilde{\mathbf{x}}\|_{\rm TV}+\kappa\lambda\|\tilde{\mathbf{x}}\|_{\mathcal{M}}\\ &\overset{(23)}{\leq}6\lambda\|\tilde{\mathbf{x}}\|_{\partial\mathcal{F}}+4\kappa\lambda\|\tilde{\mathbf{x}}\|_{\mathcal{M}}. \quad (26)\end{aligned}$$

Combining (23) with (17) yields

$$\|\tilde{\mathbf{x}}\|_{\partial\mathcal{F}}\leq\frac{K+2\kappa}{L-3}\|\tilde{\mathbf{x}}\|_{\mathcal{M}}\overset{(b)}{\leq}3\kappa\|\tilde{\mathbf{x}}\|_{\mathcal{M}}, \quad (27)$$

where (b) is due to $L>3$. Inserting (27) into (26),

$$\|\tilde{\mathbf{x}}\|_{\mathcal{M}}^2\leq22\lambda\kappa\|\tilde{\mathbf{x}}\|_{\mathcal{M}}, \quad (28)$$

and, in turn,

$$\|\tilde{\mathbf{x}}\|_{\mathcal{M}}\leq22\lambda\kappa. \quad (29)$$

Putting together the pieces, by combining (29), (27) and (23), we arrive at

$$\|\tilde{\mathbf{x}}\|_{\rm TV}\leq308\lambda\kappa^2. \quad (30)$$

Note that (30) implies the bound $\|\tilde{\mathbf{x}}\|_{\rm TV}\leq\eta$ for the choice $\lambda:=\eta/(308\kappa^2)$.

The final step of the proof is to control the probability of (21) to hold. By Corollary 1, (21) holds if

$$\max_{\mathcal{C}_l\in\mathcal{F}}(1/|\mathcal{C}_l|)\sum_{i\in\mathcal{C}_l}\varepsilon_i\leq(\lambda/2)\kappa, \quad (31)$$

and simultaneously

$$\max_{\mathcal{C}_l\in\mathcal{F}}\big\|\big(\mathbf{D}_{\mathcal{C}_l}^{\dagger}\big)^T\boldsymbol{\varepsilon}_{\mathcal{C}_l}\big\|_{\infty}\leq M\lambda/4. \quad (32)$$

We first bound the probability that (31) fails to hold. For a particular cluster $\mathcal{C}_l$, (5) yields

$$\mathrm{P}\{(1/|\mathcal{C}_l|)\sum_{i\in\mathcal{C}_l}\varepsilon_i\geq\lambda\kappa\}\leq2\exp\left(-\frac{|\mathcal{C}_l|\lambda^2\kappa^2}{8\sigma^2}\right). \quad (33)$$

Applying a union bound to (33) yields

$$\mathrm{P}\{\text{“(31) invalid”}\}\leq2|\mathcal{F}|\exp\left(-\frac{|\mathcal{C}_l|\lambda^2\kappa^2}{8\sigma^2}\right). \quad (34)$$

For controlling the probability of (32) failing to hold, we note that the entries of $\left(\mathbf{D}_{\mathcal{C}_l}^{\dagger}\right)^T \boldsymbol{\varepsilon}_{\mathcal{C}_l}$ are normally distributed with zero-mean and some variance which is upper bounded by $2\sigma^2\|\mathbf{W}\|_\infty/\rho^2(\mathcal{C}_l)$ (see (3)). Therefore, (5) and a union bound yields

$$\mathrm{P}\{\text{“(32) invalid”}\} \leq 2M \exp\left(-\frac{M^2\rho_{\mathcal{F}}^2\lambda^2}{64\sigma^2\|\mathbf{D}\|_\infty^2}\right). \quad (35)$$

A union bound yields (18) by summing the bounds (34) and (35) for the choice $\lambda := \eta/(308\kappa^2)$.

Lemma 2 *Consider an empirical graph $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \mathbf{W})$. For any two graph signals $\mathbf{u}, \mathbf{v} \in \mathbb{R}^{\mathcal{V}}$,*

$$\sum_{i\in\mathcal{V}} u_i v_i \leq (1/|\mathcal{V}|)\sum_{i\in\mathcal{V}} v_i \sum_{j\in\mathcal{V}} u_j + \left\|\left(\mathbf{D}^{\dagger}\right)^T \mathbf{v}\right\|_\infty \|\mathbf{u}\|_{\mathrm{TV}}. \quad (36)$$

Here, $\mathbf{D} \in \mathbb{R}^{\mathcal{E}\times\mathcal{V}}$ denotes the incidence matrix of the graph $\mathcal{G}$ under an arbitrary orientation of its edges $\mathcal{E}$.

Proof: Any graph signal $\mathbf{u}$ can be decomposed as

$$\mathbf{u} = \mathbf{P}\mathbf{u} + (\mathbf{I}-\mathbf{P})\mathbf{u}, \quad (37)$$

with $\mathbf{P}$ denoting the orthogonal projection matrix on the nullspace of the graph Laplacian matrix $\mathbf{L}$ (1).

For a connected graph, the nullspace $\mathcal{K}(\mathbf{L})$ is the one-dimensional subspace of constant graph signals (see von Luxburg [2007]). In this case

$$\mathbf{P} = (1/(\mathbf{1}^T\mathbf{1}))\mathbf{1}\mathbf{1}^T = (1/|\mathcal{V}|)\mathbf{1}\mathbf{1}^T \quad (38)$$

with the constant graph signal $\mathbf{1}$ assigning all nodes the same signal value 1. Therefore,

$$\mathbf{P}\mathbf{u} \overset{(38)}{=} (1/|\mathcal{V}|)\mathbf{1}(\mathbf{1}^T\mathbf{u}) = (1/|\mathcal{V}|)\sum_{i\in\mathcal{V}} u_i\mathbf{1}. \quad (39)$$

The projection on the orthogonal complement of the nullspace $\mathcal{K}(\mathbf{L}) \subseteq \mathbb{R}^{\mathcal{V}}$ is given by $\mathbf{I}-\mathbf{P}$. We can represent this projection conveniently using the incidence matrix $\mathbf{D}$ (2) (see Hütter and Rigollet [2016])

$$\mathbf{I}-\mathbf{P} = \mathbf{D}^{\dagger}\mathbf{D}. \quad (40)$$

Combining (39) and (40) with (37),

$$\sum_{i\in\mathcal{V}} u_i v_i = (1/|\mathcal{V}|)\sum_{i\in\mathcal{V}} u_i \sum_{j\in\mathcal{V}} v_j + \mathbf{v}^T\mathbf{D}^{\dagger}\mathbf{D}\mathbf{u}. \quad (41)$$

Combining (41) with the inequality $\mathbf{a}^T\mathbf{b} \leq \|\mathbf{a}\|_\infty\|\mathbf{b}\|_1$,

$$\sum_{i\in\mathcal{V}} u_i v_i \leq (1/|\mathcal{V}|)\sum_{i\in\mathcal{V}} u_i \sum_{j\in\mathcal{V}} v_j + \left\|\left(\mathbf{D}^{\dagger}\right)^T \mathbf{v}\right\|_\infty \|\mathbf{D}\mathbf{u}\|_1. \quad (42)$$

The result (36) follows from (42) by using (9). □

Applying Lemma 2 to the subgraphs induced by a partition $\mathcal{F} = \{\mathcal{C}_1, \ldots, \mathcal{C}_{|\mathcal{F}|}\}$, yields the following result.

Corollary 1 *Consider an empirical graph $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \mathbf{W})$ whose nodes are partitioned into disjoint clusters $\mathcal{F} = \{\mathcal{C}_1, \ldots, \mathcal{C}_{|\mathcal{F}|}\}$. We overload notation and denote by $\mathcal{C}_l$ also the subgraph induced by the nodes in $\mathcal{C}_l$ and assume that these subgraphs are connected. Then, for any two graph signals $\mathbf{u}, \mathbf{v} \in \mathbb{R}^{\mathcal{V}}$,*

$$\sum_{i\in\mathcal{M}} v_i u_i \leq \max_{l=1,\ldots,|\mathcal{F}|} |(1/|\mathcal{C}_l|)\sum_{i\in\mathcal{C}_l} v_i| \sum_{j\in\mathcal{M}} |u_j| + \max_{l=1,\ldots,|\mathcal{F}|} \left\|\left(\mathbf{D}_{\mathcal{C}_l}^{\dagger}\right)^T \mathbf{v}_{\mathcal{C}_j}\right\|_\infty \|\mathbf{u}\|_{\mathrm{TV}}. \quad (43)$$

Here, $\mathbf{D}_{\mathcal{C}_l} \in \mathbb{R}^{\mathcal{E}\times\mathcal{V}}$ denotes the incidence matrix of the subgraph $\mathcal{C}_l$ under an arbitrary orientation of its edges.

5 Conclusion

Using a simple non-parametric regression model for network-structured datasets, we have derived an upper bound on the probability of the nLasso error to exceed a given threshold. This bound applies if the training set satisfies the NCC with respect to a partitioning of the empirical graph into clusters of data points with similar labels. The NCC is related to the existence of a sufficiently large flow between nodes of the training set and the boundaries between clusters in the dataset. Our analysis reveals how the accuracy of nLasso depends on the empirical graph structure and identifies two key quantities which determine the required size of the training set. These quantities are the condition number associated with the NCC and the spectral gap of the cluster structure. A promising avenue for future work is the extension of our analysis of nLasso to more general probabilistic models for networked data. In particular we plan to extend our analysis to probabilistic models which form exponential families. This larger class of probabilistic models would allow to cover multi-class and multi-label classification problems.

References

F. Bach. Self-concordant analysis for logistic regression. *Electronic Journal of Statistics*, 4:384 – 414, 2010.

C. M. Bishop. *Pattern Recognition and Machine Learning*. Springer, 2006.

S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein. *Distributed Optimization and Statistical Learning via the Alternating Direction Method of Multipliers*, volume 3 of *Foundations and Trends in*

Machine Learning. Now Publishers, Hanover, MA, 2010.

P. Bühlmann and S. van de Geer. *Statistics for High-Dimensional Data*. Springer, New York, 2011.

O. Chapelle, B. Schölkopf, and A. Zien, editors. *Semi-Supervised Learning*. The MIT Press, Cambridge, Massachusetts, 2006.

S. Chen, Y. Yang, J. M. F. Moura, and J. Kovačević. Signal localization, decomposition and dictionary learning on graphs. *arxiv:1607.01100*, 2017.

D. Hallac, J. Leskovec, and S. Boyd. Network lasso: Clustering and optimization in large graphs. In *Proc. SIGKDD*, pages 387–396, 2015.

R. A. Horn and C. R. Johnson. *Matrix Analysis*. Cambridge Univ. Press, Cambridge, UK, 1985.

J.-C. Hütter and P. Rigollet. Optimal rates for total variation denoising. In *29th Annual Conference on Learning Theory*, pages 1115–1146, Columbia University, New York, New York, USA, 23–26 Jun 2016. PMLR.

A. Jung, N. Quang, and A. Mara. When is network lasso accurate? *Frontiers in Appl. Math. and Stat.*, 3, 2018.

J. Kleinberg and E. Tardos. *Algorithm Design*. Addison Wesley, 2006.

A. Kovac and A. Smith. Nonparametric regression on a graph. *Graphs and Graphical Models*, pages 432–447, Jan. 2012.

M.Yamada, T. Koh, T. Iwata, J. Shawe-Taylor, and S. Kaski. Localized Lasso for High-Dimensional Regression. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, volume 54, pages 325–333, Fort Lauderdale, FL, USA, Apr. 2017. PMLR.

M. E. J. Newman. *Networks: An Introduction*. Oxford Univ. Press, 2010.

F. Ortelli and S. van de Geer. On the total variation regularized estimator over the branched path graph. *arXiv preprint*, 2018.

N. Parikh and S. Boyd. Proximal algorithms. *Foundations and Trends in Optimization*, 1(3):123–231, 2013.

T. Pock and A. Chambolle. Diagonal preconditioning for first order primal-dual algorithms in convex optimization. In *IEEE International Conference on Computer Vision (ICCV)*, 2011.

J. Sharpnack, A. Rinaldo, and A. Singh. Sparsistency of the edge lasso over graphs. *AIStats (JMLR WCP)*, 2012.

D. Spielman. Spectral graph theory. In U. Naumann and O. Schenk, editors, *Combinatorial Scientific Computing*. Chapman and Hall/CRC, 2012.

R. Tibshirani, M. Saunders, S. Rosset, J. Zhu, and K. Knight. Sparsity and smoothness via the fused lasso. *J. R. Statist. Soc. B*, 67:91–108, 2005.

S. van de Geer. The deterministic lasso. *JSM proceedings*, 2007.

S. van de Geer and P. Bühlmann. On the conditions used to prove oracle results for the Lasso. *Electron. J. Statist.*, 3:1360 – 1392, 2009.

U. von Luxburg. A tutorial on spectral clustering. *Statistics and Computing*, 17(4):395–416, Dec. 2007.