On the Optimal Recovery of Graph Signals

Simon Foucart, Chunyang Liao, and Nate Veldt — Texas A&M University

***Abstract*—Learning a smooth graph signal from partially observed data is a well-studied task in graph-based machine learning. We consider this task from the perspective of optimal recovery, a mathematical framework for learning a function from observational data that adopts a worst-case perspective tied to model assumptions on the function to be learned. Earlier work in the optimal recovery literature has shown that minimizing a regularized objective produces optimal solutions for a general class of problems, but did not fully identify the regularization parameter. Our main contribution provides a way to compute regularization parameters that are optimal or near-optimal (depending on the setting), specifically for graph signal processing problems. Our results offer a new interpretation for classical optimization techniques in graph-based learning and also come with new insights for hyperparameter selection. We illustrate the potential of our methods in numerical experiments on several semi-synthetic graph signal processing datasets.**

I. Introduction

In graph signal processing, one starts with a dataset defined over an irregular graph domain and the goal is to recover a signal on vertices of the graph (e.g. discrete labels or regression values) when given access to only one part of the signal [Dong et al., 2020, Ortega et al., 2018]. As a concrete example, the graph may encode US counties (nodes) and their physical adjacencies (edges) while the signals may represent voting patterns, birthrates, or any number of other attributes influenced by geographic region [Jia and Benson, 2020]. As other examples, in biology, the graph may represent a gene interaction network while signals may indicate expression levels of individual genes [Dong et al., 2020]; in neuroscience, brain activity signals coming from fMRI data may be analyzed over a graph representing physical connections or co-activations among regions of a brain [Huang et al., 2018], etc.

The task of recovering a graph signal from partial information about it is also known as graph-based semi-supervised learning [Zhou et al., 2003, Belkin et al., 2004, Zhu et al., 2003]. This task has been studied in depth by researchers from many related academic communities including machine learning, statistics, and of course signal processing. In all of these settings, a common assumption is that signals vary smoothly over the graph's edge structure, meaning that adjacent nodes often share similar labels [Zhou et al., 2003, Zhu et al., 2003, Belkin et al., 2004, Xu et al., 2010, Dong et al., 2020]. Many formal objective functions and theoretical results for graph signal processing and semi-supervised learning are justified by assuming that graph signals come from a certain well-behaved probability distributions [Zhu et al., 2003, Dong et al., 2019]. This often leads to objective functions that can be minimized using simple matrix-based methods [Zhu et al., 2003, Zhou et al., 2003, Belkin et al., 2004]. However, the performance is affected by the choice of a regularization parameter in the objective function and it is not always clear how to select such a parameter. In another direction, there has been a recent surge of interest in using graph-neural networks for learning over graphs. This is often successful in practice but typically comes with no mathematical guarantees.

In our work, we address the graph signal processing task from a novel perspective—that of optimal recovery. This perspective does not rely on the assumption that ground truth signals are drawn from a well-behaved distribution. Instead, the goal is to find optimal solutions under worst-case assumptions about graph smoothness and labeling error. This approach comes with several benefits. Primarily, we present new theoretical results on finding best solutions under the optimal recovery framework (locally and globally, see later sections for technical details). Along the way, we highlight the connections between the optimization problems stemming from this framework and the classical techniques encountered in graph signal processing. One significant contribution is to provide rigorous theoretical guarantees for selecting the regularization parameter in the objective function being minimized. Setting this parameter is not entirely free of challenges, as it actually depends on the parameters characterizing graph smoothness and labeling error. Nevertheless, our results offer fresh intuition on how to reasonably choose the regularization parameters intrinsic to objective functions common in graph signal processing. Finally, we provide a proof-of-concept implementation of our approach and illustrate its performance in several empirical graph signal processing experiments.

II. The Perspective from Optimal Recovery

Let $G = (V, E)$ be an undirected graph with $N = |V|$ vertices identified with $1, 2, \ldots, N$. A signal f defined on V is thus identified with a vector $f \in \mathbb{R}^N$. The previously-mentioned common assumption that f varies smoothly over the graph's edge structure qualitatively translates into the fact that the values f_i and f_j do not differ much if the vertices i and j are strongly connected. Quantitatively, putting a weight $w_{i,j} \geq 0$ on the edge connecting i and j, thus defining a (weighted, symmetric) adjacency matrix $W \in \mathbb{R}^{N \times N}$, the assumption takes the form

$$\frac{1}{2} \sum_{i,j=1}^{N} w_{i,j} (f_i - f_j)^2 \leq \varepsilon^2$$

for a small $\varepsilon > 0$ standing for a graph smoothness parameter. Introducing the graph Laplacian $L = D - W \in \mathbb{R}^{N \times N}$, where D is the diagonal matrix with entries $D_{i,i} = \sum_{j=1}^{N} W_{i,j}$, the assumption succinctly reads

$$\langle Lf, f \rangle \leq \varepsilon^2, \qquad \text{or} \quad \|L^{1/2} f\|_2 \leq \varepsilon.$$

We recall that the square-root $L^{1/2}$ of L is well-defined because the graph Laplacian L is positive semidefinite. Note that it is not positive definite, since 0 is always an eigenvalue of L. In fact, its multiplicity equals the number of connected components C of G, with orthogonal eigenvectors provided by the indicator vectors $\mathbf{1}_C \in \{0,1\}^N$ of C. Throughout, we shall assume that the graph G is known, and hence that L is available to the user.

As for the unknown f, it is partially observed—or labeled[1]. In other words, there is a subset V_ℓ, with size $|V_\ell| = n_\ell$, of vertices for which the f_i, $i \in V_\ell$, are known. In reality, they are known up to additive errors, so that the user has access to

$$y_i = f_i + e_i, \qquad i \in V_\ell.$$

To abbreviate, we write $y = \Lambda f + e \in \mathbb{R}^{n_\ell}$, where the linear map $\Lambda : \mathbb{R}^N \to \mathbb{R}^{n_\ell}$ satisfies $\Lambda\Lambda^* = I_{n_\ell}$ here (since, up to a proper ordering of the vertices, (the matrix of) Λ takes the form $\begin{bmatrix} I_{n_\ell} \mid 0 \end{bmatrix}$). We shall assume that an ℓ_2-bound on the error vector e is available, namely

$$\|e\|_2 \le \eta$$

for a small $\eta > 0$ standing for a labeling error parameter.

Our objective is now to estimate the graph signal f on the set of unlabeled vertices, i.e., on $V_u = V \setminus V_\ell$, which has size $|V_u| = n_u = N - n_\ell$. In the framework of optimal recovery, we aim at doing so in a worst-case optimal way given the graph smoothness and labeling error assumptions, expressed as $f \in \mathcal{K}$ and $e \in \mathcal{E}$, where the model set $\mathcal{K}$ and the uncertainty set $\mathcal{E}$ are given as

$$\mathcal{K} = \{f \in \mathbb{R}^N : \|L^{1/2} f\|_2 \le \varepsilon\}, \tag{1}$$

$$\mathcal{E} = \{e \in \mathbb{R}^{n_\ell} : \|e\|_2 \le \eta\}. \tag{2}$$

A scheme to estimate $f_{|V_u} \in \mathbb{R}^{n_u}$ from $y \in \mathbb{R}^{n_\ell}$ is nothing but a map $\Delta : \mathbb{R}^{n_\ell} \to \mathbb{R}^{n_u}$, which we call a recovery map. We are interested in those recovery maps that are optimal

- *in the global setting,* i.e.,
$$\sup\{\|f_{|V_u} - \Delta(\Lambda f + e)\|_2 : f \in \mathcal{K}, e \in \mathcal{E}\}$$
is as small as possible;
- *in the local setting*, i.e., at any given $y \in \mathbb{R}^{n_\ell}$,
$$\sup\{\|f_{|V_u} - z\|_2 : f \in \mathcal{K}, e \in \mathcal{E}, \Lambda f + e = y\}$$
evaluated at $z = \Delta(y)$ is as small as possible. Such a $\Delta(y) \in \mathbb{R}^{n_u}$ is called a Chebyshev center for the set $\mathcal{S} = \{f_{|V_u} : f \in \mathcal{K}, e \in \mathcal{E}, \Lambda f + e = y\}$, as it is easily seen to be a center of a minimal-radius ball containing $\mathcal{S}$.

If we believe that the observed labels need to be adjusted, too, instead of estimating $f_{|V_u}$, we may want to estimate f in full. We may also want to estimate the average of f or its value at a particular vertex $i_0 \in V$. To deal with these situations all at once, we introduce a quantity of interest $Q : \mathbb{R}^N \to \mathbb{R}^n$, which in the examples above is given by, respectively,

$$Q(f) = f_{|V_u}, \;\; Q(f) = f, \;\; Q(f) = \frac{1}{N}\sum_{i \in V} f_i, \;\; Q(f) = f_{i_0}.$$

[1]Labels are real numbers here, not elements of a binary set such as $\{0,1\}$.

In this generality, the global and local worst-case errors for the estimation of Q are defined, for $\Delta : \mathbb{R}^{n_\ell} \to \mathbb{R}^n$ and for $y \in \mathbb{R}^{n_\ell}$, $z \in \mathbb{R}^n$, by

$$\mathrm{gwce}_Q(\Delta) = \sup_{f \in \mathcal{K}, e \in \mathcal{E}} \{\|Q(f) - \Delta(\Lambda f + e)\|_2\},$$

$$\mathrm{lwce}_Q(y,z) = \sup_{f \in \mathcal{K}, e \in \mathcal{E}} \{\|Q(f) - z\|_2 : \Lambda f + e = y\}.$$

We call a recovery map $\Delta : \mathbb{R}^{n_\ell} \to \mathbb{R}^n$ globally optimal if it minimizes $\mathrm{gwce}_Q(\Delta)$ and locally optimal if $\Delta(y)$ minimizes $\mathrm{lwce}_Q(y,z)$ for any given $y \in \mathbb{R}^{n_\ell}$. Of course, locally optimal recovery maps are automatically globally optimal, but they are typically harder to produce (as the current work will also illustrate). We may therefore relax the aspiration of genuine optimality to one of near optimality by merely requiring that $\mathrm{lwce}_Q(y, \Delta(y)) \le C \inf\{\mathrm{lwce}_Q(y,z), z \in \mathbb{R}^n\}$ for some absolute constant $C > 1$.

III. Selection of the Regularization Parameter

We now show that (near-)optimal recovery maps can be obtained through Tikhonov-style regularization and we uncover a principled way to choose the regularization parameter based on the graph smoothness and labeling error parameters. We start with some preparatory information about regularization before presenting our genuine optimality result in the global setting and our near optimality result in the local setting.

A. Rundown on regularization

When searching for the signal $f \in \mathbb{R}^N$ that produced the observation vector $y \in \mathbb{R}^{n_\ell}$, it is natural to try and make the data-fidelity term $\|\Lambda f - y\|_2^2$ small. Furthermore, to enforce the graph smoothness condition that $\|L^{1/2} f\|_2^2$ is small, one can incorporate this condition as a constraint in a miminization problem or add the regularization term $\gamma \|L^{1/2} f\|_2^2$ to the objective function, as done e.g. in [Belkin et al., 2004]. Instead of parametrizing by $\gamma > 0$, it will be more convenient for our purpose to parametrize by some $\tau \in (0,1)$, thus leading to the regularization map $\Delta_\tau : \mathbb{R}^{n_\ell} \to \mathbb{R}^N$ given by

$$\Delta_\tau : y \mapsto \operatorname*{argmin}_{f \in \mathbb{R}^N} \; (1-\tau)\|L^{1/2} f\|_2^2 + \tau \|\Lambda f - y\|_2^2.$$

This map is well defined under the assumption that at least one vertex is observed in each connected component of the graph, which translates into $\ker(L) \cap \ker(\Lambda) = \{0\}$ or equivalently into the invertibility of $(1-\tau)L + \tau\Lambda^*\Lambda$ (see §1 in the supplementary material). Indeed, as the minimizers f_τ of the above objective function are characterized by the normal equation $(1-\tau)Lf_\tau + \tau\Lambda^*(\Lambda f_\tau - y) = 0$, this invertibity shows that f_τ is unique and is equal to

$$\Delta_\tau(y) = \big((1-\tau)L + \tau\Lambda^*\Lambda\big)^{-1}(\tau\Lambda^* y). \tag{3}$$

This expression reveals in particular that Δ_τ is a linear map.

The extreme case $\tau \to 0$ is interpreted as the minimizer of $\|\Lambda f - y\|_2^2$ subject to $L^{1/2} f = 0$, which is not very interesting, see §2 for explanation. The extreme case $\tau \to 1$ is interpreted as the minimizer of $\|L^{1/2} f\|_2^2$ subject to $\Lambda f = y$, which appears commonly in graph signal processing under the names

of harmonic method [Zhu et al., 2003] or interpolatory method [Belkin et al., 2004].

B. Genuine optimality in the global setting

It was recognized already in [Melkman and Micchelli, 1979, Micchelli, 1993] that the regularization maps Δ_τ produce a globally optimal recovery map for *some* parameter $\tau \in (0,1)$, but the choice of this parameter was not made explicit. Theorem 1 below shows that this parameter can be obtained by solving a semidefinite program. Such a result was established in [Foucart and Liao, To appear] in a slightly more restrictive setting, namely the place of $L^{1/2}$ was taken by an orthogonal projector and only full recovery (i.e., $Q = I_N$) was considered.

Theorem 1. Given a linear quantity of interest $Q : \mathbb{R}^N \to \mathbb{R}^n$, let $\mathcal{K}$ and $\mathcal{E}$ be the model and uncertainty sets from (1)-(2). Defining $\tau_\flat = d_\flat/(c_\flat + d_\flat)$ where $c_\flat, d_\flat \geq 0$ are solutions to

$$\underset{c,d\geq 0}{\text{minimize}} \; c\varepsilon^2 + d\eta^2 \quad \text{s.to} \;\; cL + d\Lambda^*\Lambda \succeq Q^*Q, \qquad (4)$$

the linear map $Q \circ \Delta_{\tau_\flat} : \mathbb{R}^{n_\ell} \to \mathbb{R}^n$ is a globally optimal recovery map relative to $\mathcal{K}$ and $\mathcal{E}$, meaning that

$$\mathrm{gwce}_Q(Q \circ \Delta_{\tau_\flat}) = \inf_{\Delta : \mathbb{R}^{n_\ell} \to \mathbb{R}^n} \mathrm{gwce}_Q(\Delta).$$

The justification of this theorem relies on the two lemmas below, whose proofs appear in §3 and §4 of the supplementary material. Lemma 2 relies on a version of the S-procedure due to Polyak [1998] and follows [Foucart and Liao, To appear] closely, but the argument for Lemma 3 follows a different route to deal with a quantity of interest $Q \neq I_N$.

Lemma 2. For an arbitrary recovery map $\Delta : \mathbb{R}^{n_\ell} \to \mathbb{R}^n$, the squared global worst-case error is lower-bounded as

$$\begin{aligned}\mathrm{gwce}_Q(\Delta)^2 &\geq \sup_{h \in \mathbb{R}^N} \{\|Q(h)\|_2^2 : \|L^{1/2}h\|_2^2 \leq \varepsilon^2, \|\Lambda h\|_2^2 \leq \eta^2\} \\ &= \inf_{c,d\geq 0} \{c\varepsilon^2 + d\eta^2 : cL + d\Lambda^*\Lambda \succeq Q^*Q\}.\end{aligned}$$

Lemma 3. If $c, d \geq 0$ satisfy $cL + d\Lambda\Lambda^* \succeq Q^*Q$, then, setting $\tau = d/(c+d)$, one has, for all $f \in \mathbb{R}^N$ and all $e \in \mathbb{R}^{n_\ell}$,

$$\|Q(I - \Delta_\tau\Lambda)f - Q\Delta_\tau e\|_2^2 \leq c\|L^{1/2}f\|_2^2 + d\|e\|_2^2.$$

Proof of Theorem 1. Let $c_\flat, d_\flat \geq 0$ be minimizers of (4). On the one hand, according to Lemma 2, the squared global worst-case error of any recovery map $\Delta : \mathbb{R}^{n_\ell} \to \mathbb{R}^n$ satisfies

$$\mathrm{gwce}_Q(\Delta)^2 \geq c_\flat\varepsilon^2 + d_\flat\eta^2.$$

On the other hand, the linearity of $Q \circ \Delta_{\tau_\flat}$, $\tau_\flat := d_\flat/(c_\flat + d_\flat)$, implies that its squared global worst-case error becomes

$$\mathrm{gwce}_Q(Q \circ \Delta_{\tau_\flat})^2 = \sup_{f \in \mathcal{K}, e \in \mathcal{E}} \|Q(I - \Delta_\tau\Lambda)f - Q\Delta_\tau e\|_2^2,$$

which, according to Lemma 3, does not exceed

$$\sup_{f \in \mathcal{K}, e \in \mathcal{E}} c\|L^{1/2}f\|_2^2 + d\|e\|_2^2 = c_\flat\varepsilon^2 + d_\flat\eta^2.$$

All in all, we have shown that $\mathrm{gwce}_Q(Q \circ \Delta_{\tau_\flat}) \leq \mathrm{gwce}_Q(\Delta)$ for any map $\Delta : \mathbb{R}^{n_\ell} \to \mathbb{R}^n$, which is the desired result. □

Remark. To achieve genuine optimality, exact knowledge of the parameters ε and η is needed, but near optimality is achievable when these are overestimated by some $\bar{\varepsilon}$ and $\bar{\eta}$ satisfying $\bar{\varepsilon} \leq C\varepsilon$ and $\bar{\eta} \leq C\eta$, see §5.

Remark. The semidefinite program (4), featuring an $N \times N$ matrix, does not run when N is in the thousands. Nonetheless, it is expected that the computational burden could be alleviated if Q maps into a low-dimensional space $\mathbb{R}^n$. This is certainly the case in the extreme case $n = 1$, see §6.

C. Near optimality in the local setting

In contrast to the global setting, we are unaware of a general result stating that the regularization maps Δ_τ produce a locally optimal recovery map for *some* parameter $\tau \in (0,1)$. For full recovery ($Q = I_N$), such a statement is true in at least two situations, though. The first situation requires $\Lambda\Lambda^* = I_{n_\ell}$ (which is the case here) and an orthogonal projector P in place of $L^{1/2}$ (up to normalization, $L^{1/2}$ happens to be an orthogonal projector if G is an unweighted graph made of disconnected complete subgraphs of identical sizes, see §7): it was established in [Foucart and Liao, To appear] that $\Delta_{\tau_\sharp}$ is a locally optimal recovery map when $\tau_\sharp$ is the unique parameter τ between $1/2$ and $\varepsilon/(\varepsilon+\eta)$ satisfying the eigenvalue equation

$$\begin{aligned}&\lambda_{\min}((1-\tau)P + \tau\Lambda^*\Lambda) \\ &\qquad = \frac{(1-\tau)^2\varepsilon^2 - \tau^2\eta^2}{(1-\tau)\varepsilon^2 - \tau\eta^2 + (1-\tau)\tau(1-2\tau)\delta^2},\end{aligned}$$

where $\delta = \min\{\|Pf\|_2 : \Lambda f = y\} = \min\{\|\Lambda f - y\|_2 : Pf = 0\}$. The second situation requires working in the complex setting: it was established in [Beck and Eldar, 2007] that $\Delta_{\tau_\flat}$ is a locally optimal recovery map when $\tau_\flat = d_\flat/(c_\flat + d_\flat)$, with $c_\flat, d_\flat$ solving the semidefinite program

$$\begin{aligned}\underset{c,d,t\geq 0}{\text{minimize}} \; & c\varepsilon^2 + d(\eta^2 - \|y\|_2^2) + t \quad \text{s.to} \; cL + d\Lambda^*\Lambda \succeq I_N \\ & \text{and} \; \left[\begin{array}{c|c} cL + d\Lambda^*\Lambda & d\Lambda^*y \\ \hline dy^*\Lambda & t \end{array}\right] \succeq 0.\end{aligned}$$

In the real setting, although the value of $\mathrm{lwce}_{I_N}(y, \Delta_{\tau_\flat}(y))$ is only guaranteed to provide an upper bound for the minimal local worst-case error, it is not unlikely that $\Delta_{\tau_\flat}$ is genuinely a locally optimal recovery map—this is the case in the first situation (result not published yet).

Regardless of the above considerations, relaxing genuine optimality to near optimality, it can always be guaranteed that some regularization map Δ_τ produces a locally near optimal recovery map for a parameter $\tau \in (0,1)$ that can be computed. This is the gist of the following result, see §8 for a proof.

Theorem 4. Given a linear quantity of interest $Q : \mathbb{R}^N \to \mathbb{R}^n$, let $\mathcal{K}$ and $\mathcal{E}$ be the model and uncertainty sets from (1)-(2). For $y \in \mathbb{R}^{n_\ell}$, let $\widehat{f} \in \mathbb{R}^N$ be the solution to

$$\underset{f \in \mathbb{R}^N}{\text{minimize}} \; \max\left\{\|L^{1/2}f\|_2^2, \frac{\varepsilon^2}{\eta^2}\|\Lambda f - y\|_2^2\right\}.$$

Then $\widehat{f}$ agrees with $\Delta_{\tau_\natural}(y)$, where $\tau_\natural$ is the unique parameter $\tau \in (0,1)$ satisfying

$$\|L^{1/2}\Delta_\tau(y)\|_2 = \frac{\varepsilon}{\eta}\|\Lambda\Delta_\tau(y) - y\|_2.$$

Moreover, one has $\mathrm{lwce}_Q(y, Q(\widehat{f})) \leq 2\inf_{z \in \mathbb{R}^N} \mathrm{lwce}_Q(y, z)$.

IV. Numerical Validation

In this section, we illustrate the performance of optimal recovery methods on several semi-synthetic regression datasets and verify the near optimality of a regularization parameter in the local setting. We implement our algorithms in MATLAB and use CVX [Grant and Boyd, 2014] for solving semidefinite programs. All numerical experiments are available on the GitHub repository https://github.com/liaochunyang/ORofGraphSignals.

Let us recall that all our optimal recovery maps correspond to solving a regularized optimization program with a specific choice of hyperparameter. This type of regularized objective is already standard in graph-based machine learning [Belkin et al., 2004], but it is often unclear how this parameter should be chosen. The primary goal of our numerical experiments is to illustrate how our techniques for hyperparameter selection work in a controlled environment where we have access to a ground truth graph signal f and we can estimate the true smoothness and error parameters ϵ and η. In practical settings, we do not actually have access to f, nor can we estimate ε and η exactly, but we shall confirm that near optimality is achievable under mild overestimations of ϵ and η (see §5).

We consider several real-world graphs and, using a standard approach [Dong et al., 2019] (see §9 for more details), we generate synthetic signals f whose values at the nodes are normalizing to be between 0 and 1 for simplicity. In our first experiment, we show how the prediction error changes as the number of labeled vertices grows. We begin by running all methods for $n_\ell = 5$ and we keep adding 5 new labeled nodes at a time—by the end, we have run experiments for n_ℓ ranging from 5 to $N/2$ in increments of 5. For each choice of n_ℓ, our goal is to recover labels at unlabeled nodes, i.e., $Q(f) = f|_{V_u}$, where f is the vector of true node labels. The prediction error for any estimator $\hat{f}$ of $f|_{V_u}$ is defined as $\|f|_{V_u} - \hat{f}\|_2$. For the smoothness parameter, we set $\epsilon = 2\|L^{1/2}f\|_2^2$. Next, we introduce noise artificially by generating a uniform random vector e and subtract the mean before scaling so that $\|e\|_2 \leq \eta$, where η is chosen as $\eta = 2$ (see §9 for details on other types of noise). We then create the corrupted labels $y = f_{|V_\ell} + e$. In the supplementary material, we also consider a mild overestimation on η by setting $\bar{\eta} = 2\eta$.

In Figure 1, we display the prediction errors produced by locally/globally optimal recovery maps for the graph Adjnoun (see §9 for results on other graphs). The constructions of the globally optimal recovery map and the locally near optimal recovery map were presented in Theorem 1 and Theorem 4, respectively. We also use a grid-search approach to find the smallest prediction error over all regularization parameters in order to display a curve of the lowest possible prediction error. This is neither computationally efficient nor realistic, as it assumes that we can always check the prediction error for any estimator $\hat{f}$, but it provides a bound on the best case scenario for solving the regularized objective. Comparing the magenta and the red lines and the black and the green lines, we observe that an overestimation of η does not lead to large differences in prediction error for both local and global optimal recovery maps, which suggests that one can safely use a mild overestimation of η when we cannot access the true η. We also notice that the prediction errors produced by locally/globally optimal recovery maps are close to the prediction error produced by the best Tikhonov regularization method (blue line).

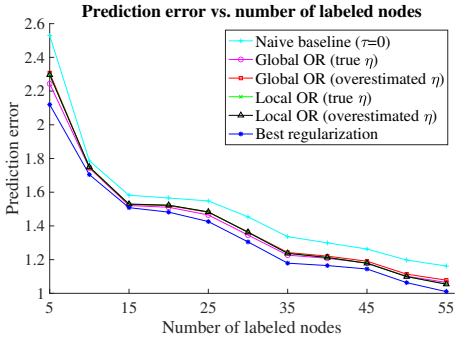


Fig. 1. Prediction error vs. number of labeled nodes on graph Adjnoun with added uniform noise.

In the second experiment (see Figure 2), we confirm the near optimality of the locally optimal recovery map described in Theorem 4. The setup of this experiment is similar to the first experiment. The parameter $\tau_\sharp$—the unique $\tau \in (0,1)$ such that $\|L^{1/2}\Delta_\tau(y)\|_2 = (\varepsilon/\eta)\|\Lambda\Delta_\tau(y) - y\|_2$— is found by the bisection method and is displayed in Figure 2 by the dashed vertical line. The blue curve represents an upper bound for the local worst-case error $\mathrm{lwce}_Q(y, Q \circ \Delta_\tau(y))$ as a function of the regularization parameter. For each τ in a grid of $[0,1]$, this upper bound was computed by solving a semidefinite relaxation for the local worst-case error, see §10 for details. Figure 2 not only supports the local near optimality of the recovery map $Q \circ \Delta_{\tau_\sharp}$, but it also hints that $\tau_\sharp$ is not far away the best regularization parameter.

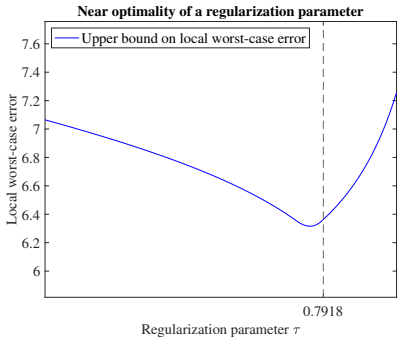


Fig. 2. Local worst-case error vs. regularization parameter.

Acknowledgment

S. F. is partially supported by grants from the NSF (DMS-2053172) and from the ONR (N00014-20-1-2787).

References

A. Beck and Y. C. Eldar. Regularization in regression with bounded noise: a Chebyshev center approach. *SIAM Journal on Matrix Analysis and Applications*, 29(2):606–625, 2007.

M. Belkin, I. Matveeva, and P. Niyogi. Regularization and semi-supervised learning on large graphs. In *Learning Theory: 17th Annual Conference on Learning Theory*, pages 624–638. Springer, 2004.

T. A. Davis and Y. Hu. The University of Florida Sparse Matrix Collection. *ACM Trans. Math. Softw.*, 38(1), 2011. URL https://doi.org/10.1145/2049662.2049663.

X. Dong, D. Thanou, M. Rabbat, and P. Frossard. Learning graphs from data: A signal representation perspective. *IEEE Signal Processing Magazine*, 36(3):44–63, 2019. doi: 10.1109/MSP.2018.2887284.

X. Dong, D. Thanou, L. Toni, Mi. Bronstein, and P. Frossard. Graph signal processing for machine learning: A review and new perspectives. *IEEE Signal Processing Magazine*, 37(6): 117–127, 2020.

S. Foucart and C. Liao. Optimal recovery from inaccurate data in Hilbert spaces: Regularize, but what of the parameter? *Constructive Approximation*, To appear.

A. L. Garkavi. On the optimal net and best cross-section of a set in a normed space. *Izvestiya Rossiiskoi Akademii Nauk. Seriya Matematicheskaya*, 26(1):87–106, 1962.

M Girvan and M. Newman. Community structure in social and biological networks. *Proceedings of the National Academy of Sciences*, 99(12):7821–7826, 2002.

M. Grant and S. Boyd. CVX: Matlab software for disciplined convex programming, version 2.1. http://cvxr.com/cvx, March 2014.

W. Huang, T. Bolton, J. Medaglia, D. Bassett, A. Ribeiro, and D. Van De Ville. A graph signal processing perspective on functional brain imaging. *Proceedings of the IEEE*, 106(5): 868–885, 2018.

J. Jia and A. Benson. Residual correlation in graph neural network regression. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 588–598, 2020.

V. Krebs. Books about US Politics. URL http://networkdata.ics.uci.edu/data.php?d=polbooks.

D. Lusseau, K. Schneider, O. Boisseau, P. Haase, E. Slooten, and S. Dawson. The bottlenose dolphin community of doubtful sound features a large proportion of long-lasting associations: can geographic isolation explain this unique trait? *Behavioral Ecology and Sociobiology*, 54:396–405, 2003.

A. A. Melkman and C. A. Micchelli. Optimal estimation of linear operators in Hilbert spaces from inaccurate data. *SIAM Journal on Numerical Analysis*, 16(1):87–105, 1979.

C. A. Micchelli. Optimal estimation of linear operators from inaccurate data: a second look. *Numerical Algorithms*, 5(8): 375–390, 1993.

M. Newman. Finding community structure in networks using the eigenvectors of matrices. *Physical Review E*, 74(3): 036104, 2006.

A. Ortega, P. Frossard, J. Kovačević, J. Moura, and P. Vandergheynst. Graph signal processing: Overview, challenges, and applications. *Proc. IEEE*, 106(5):808–828, 2018.

B. T. Polyak. Convexity of quadratic transformations and its use in control and optimization. *Journal of Optimization Theory and Applications*, 99(3):553–583, 1998.

Y. Xu, J. Dyer, and A. Owen. Empirical stationary correlations for semi-supervised learning on graphs. *The Annals of Applied Statistics*, 4(2):589–614, 2010.

D. Zhou, O. Bousquet, T. Lal, J. Weston, and B. Schölkopf. Learning with local and global consistency. *Advances in Neural Information Processing Systems*, 16, 2003.

X. Zhu, Z. Ghahramani, and J. Lafferty. Semi-supervised learning using Gaussian fields and harmonic functions. In *Proceedings of the 20th International Conference on Machine Learning*, pages 912–919, 2003.

SUPPLEMENTARY MATERIAL

In this section, we supply the theoretical justifications that are missing from the main text.

§1. *The regularization map is well defined for $\tau \in (0,1)$:* The regularization problem can be written as

$$\underset{f\in\mathbb{R}^N}{\text{minimize}}\ \|Af-b\|_2^2, \quad A := \begin{bmatrix}\sqrt{1-\tau}L^{1/2}\\ \sqrt{\tau}\Lambda\end{bmatrix}, \ \ b := \begin{bmatrix}0\\ \sqrt{\tau}y\end{bmatrix}.$$

Its solutions $\bar{f}$ are characterized by the normal equation $A^*A\bar{f} = A^*b$, i.e., by $\big((1-\tau)L+\tau\Lambda^*\Lambda\big)\bar{f} = \tau\Lambda^*y$. Note that we always make the assumption $\ker(L)\cap\ker(\Lambda)=\{0\}$, otherwise, fixing $f_0\in\mathcal{K}$ and $e_0\in\mathcal{E}$ with $\Lambda f_0+e_0=y$, the existence of $h\in\mathbb{R}^N\setminus\{0\}$ such that $Lh=0$ and $\Lambda h=0$ implies that $f_t := f_0+th\in\mathcal{K}$, $e_t:=e_0\in\mathcal{E}$, and $\Lambda f_t+e_t=y$ for all $t\in\mathbb{R}$, yielding an infinite local (in turn global) worst-case error for the recovery of $Q=I_N$. This assumption ensures that $(1-\tau)L+\tau\Lambda^*\Lambda$ is positive definite—hence invertible—for any $\tau\in(0,1)$, since

$$\big\langle\big((1-\tau)L+\tau\Lambda^*\Lambda\big)h,h\big\rangle = (1-\tau)\|L^{1/2}h\|_2^2+\tau\|\Lambda h\|_2^2\geq 0$$

for all $h\in\mathbb{R}^N$, with equality possible when and only when $h\in\ker(L)\cap\ker(\Lambda)=\{0\}$. This shows that $\bar{f}$ is unique and given by $\bar{f}=\big((1-\tau)L+\tau\Lambda^*\Lambda\big)^{-1}(\tau\Lambda^*y)$. Finally, if the graph G is made of K connected components $C_1,\dots,C_K$, we observe that

$$\begin{aligned}\ker(L)\cap\ker(\Lambda) &= \left\{h=\sum_{k=1}^K a_k\mathbf{1}_{C_k},\, a\in\mathbb{R}^K,\, h_{|V_\ell}=0\right\}\\ &= \left\{\sum_{k=1}^K a_k\mathbf{1}_{C_k},\, a\in\mathbb{R}^K,\, a_k=0 \text{ when } C_k\cap V_\ell\neq\emptyset\right\},\end{aligned}$$

so that $\ker(L)\cap\ker(\Lambda)=\{0\}$ if and only if $C_k\cap V_\ell\neq\emptyset$ for all $k=1,\dots,K$, which means that at least one vertex is observed in each connected component.

§2. *The limiting case $\tau\to 0$:* Writing $f_\tau=\Delta_\tau(y)$, if we divide the objective function that f_τ minimizes by $\tau>0$, we see that

$$f_\tau = \underset{f\in\mathbb{R}^N}{\text{argmin}}\ \frac{1-\tau}{\tau}\|L^{1/2}f\|_2^2+\|\Lambda f_\tau-y\|_2^2.$$

Intuitively, the limit f_0 of f_τ as $\tau\to 0$ should satisfy $L^{1/2}f_0=0$, otherwise $\|L^{1/2}f_\tau\|_2^2\geq\kappa$ for some $\kappa>0$ when τ is sufficiently small, and then $\big((1-\tau)/\tau\big)\|L^{1/2}f_\tau\|_2^2$ blows up as $\tau\to 0$, preventing f_τ to be a minimizer of the divided objective function. It suggests—and this can be made precise—that

$$f_0 = \underset{f\in\mathbb{R}^N}{\text{argmin}}\ \|\Lambda f-y\|_2^2 \quad \text{s.to} \ \ L^{1/2}f=0.$$

The constraint $L^{1/2}f=0$ is equivalent to f taking the form $f=\sum_{k=1}^K a_k\mathbf{1}_{C_k}$ for some $a\in\mathbb{R}^K$, where $C_1,\dots,C_K$ denote the connected components of the graph G. Under this constraint, we then have

$$\Lambda f-y = f_{|V_\ell}-y = \sum_{k=1}^K\big(a_k\mathbf{1}_{C_k\cap V_\ell}-y_{C_k}\big),$$

and hence, since the summands have disjoint supports,

$$\begin{aligned}\|\Lambda f-y\|_2^2 &= \sum_{k=1}^K\|a_k\mathbf{1}_{C_k\cap V_\ell}-y_{C_k}\|_2^2\\ &= \sum_{k=1}^K\big(a_k^2|C_k\cap V_\ell|-2a_k\langle\mathbf{1}_{C_k\cap V_\ell},y_{C_k}\rangle+\|y_{C_k}\|_2^2\big).\end{aligned}$$

This quantity is easily seen to attain its minimal value when $a_k=\langle\mathbf{1}_{C_k\cap V_\ell},y_{C_k}\rangle/|C_k\cap V_\ell|$ for each $k=1,\dots,K$. All in all, this signifies that the component of f_0 on each C_k is equal to the average of y_{C_k}—as announced, the case $\tau\to 0$ is not very interesting!

§3. *Proof of Lemma 2:* To prove the inequality, consider $h\in\mathbb{R}^N$ with $\|L^{1/2}h\|_2^2\leq\varepsilon^2$ and $\|\Lambda h\|_2^2\leq\eta$. Then define $f_\pm=\pm h\in\mathcal{K}$ and $e_\pm=\mp\Lambda h\in\mathcal{E}$, so that

$$\begin{aligned}\mathrm{gwce}_Q(\Delta) &\geq \max_\pm\|Q(f_\pm)-\Delta(\Lambda f_\pm+e_\pm)\|_2\\ &= \max_\pm\|Q(\pm h)-\Delta(0)\|_2\\ &\geq \frac{1}{2}\|Q(h)-\Delta(0)\|_2+\frac{1}{2}\|Q(-h)-\Delta(0)\|_2\\ &\geq \frac{1}{2}\|(Q(h)-\Delta(0))-(Q(-h)-\Delta(0))\|_2\\ &= \frac{1}{2}\|2Q(h)\|_2=\|Q(h)\|_2.\end{aligned}$$

In remains to take the supremum over admissible h to obtain the announced lower bound.

Next, the transformation of the lower bound for the global worst-case error relies on a case of validity of the S-procedure due to Polyak, see Polyak [1998]. We start by writing this (squared) lower bound as

$$\inf_{\gamma\in\mathbb{R}}\ \gamma \ \ \text{s.to } \|Q(h)\|_2^2\leq\gamma \text{ whenever } \|L^{1/2}h\|_2^2\leq\varepsilon^2,\ \|\Lambda h\|_2^2\leq\eta^2.$$

Using the S-procedure of Polyak, the above constraint is equivalent to the existence of $c,d\geq 0$ such that, for all $h\in\mathbb{R}^N$,

$$\|Q(h)\|_2^2-\gamma\leq c(\|L^{1/2}h\|_2^2-\varepsilon^2)+d(\|\Lambda h\|_2^2-\eta^2), \quad (5)$$

under the proviso that there exist $\widetilde{h}\in\mathbb{R}^N$ and $\alpha,\beta\in\mathbb{R}$ such that $\|L^{1/2}\widetilde{h}\|_2^2<\varepsilon^2$, $\|\Lambda\widetilde{h}\|_2^2<\eta^2$, and $\alpha L+\beta\Lambda^*\Lambda\succ 0$. This proviso is met by taking $\widetilde{h}=0$ and $(\alpha,\beta)=(1-\tau,\tau)$ for any $\tau\in(0,1)$, see §1 above. Now, the constraint (5) can be written as

$$\langle(cL+d\Lambda^*\Lambda-Q^*Q)h,h\rangle+\gamma-c\varepsilon^2-d\eta^2\geq 0$$

for all $h\in\mathbb{R}^N$, which in fact decouples as the two constraints $cL+d\Lambda^*\Lambda-Q^*Q\succeq 0$ and $\gamma-c\varepsilon^2-d\eta^2\geq 0$. Under the latter constraint, the minimal value of γ is $c\varepsilon^2+d\eta^2$ and we arrive at the desired expression. □

§4. *Proof of Lemma 3:* The two additional lemmas below are needed.

Lemma 5. If A, B, C are square matrices of similar size and if $C \succeq 0$, then

$$\left[\begin{array}{c|c} A & 0 \\ \hline 0 & B \end{array}\right] \succeq \left[\begin{array}{c|c} A-C & C \\ \hline C & B-C \end{array}\right].$$

Proof. To prove that the difference of these two matrices is positive semidefinite, we simply write, for any vectors x, y,

$$\begin{aligned} &\left\langle \left[\begin{array}{c|c} C & -C \\ \hline -C & C \end{array}\right] \begin{bmatrix} x \\ y \end{bmatrix}, \begin{bmatrix} x \\ y \end{bmatrix} \right\rangle \\ &= \langle Cx, x\rangle - \langle Cy, x\rangle - \langle Cx, y\rangle + \langle Cy, y\rangle \\ &= \langle C^{1/2}x, C^{1/2}x\rangle - 2\langle C^{1/2}x, C^{1/2}y\rangle + \langle C^{1/2}y, C^{1/2}y\rangle \\ &= \|C^{1/2}x - C^{1/2}y\|_2^2, \end{aligned}$$

which is obviously nonnegative. □

Lemma 6. If A and B are positive semidefinite matrices of similar size such that $A + B \succ 0$, then $C := A(A+B)^{-1}B$ is positive semidefinite.

Proof. Writing C as $C = A(A+B)^{-1}(A+B-A)$, i.e., $C = A - A(A+B)^{-1}A$, shows that C is self-adjoint and reveals that we in fact have to prove that $A(A+B)^{-1}A \preceq A$. To see why this is so, we start from $A \preceq A + B$, so that $M := (A+B)^{-1/2}A(A+B)^{-1/2}$ satisfies $M \preceq I$. This implies that $M^2 \preceq M$, which reads

$$(A+B)^{-1/2}A(A+B)^{-1}A(A+B)^{-1/2} \preceq (A+B)^{-1/2}A(A+B)^{-1/2}.$$

Multiplying on the left and on the right by $(A+B)^{1/2}$ yields the desired result. □

Focusing now on the proof of Lemma 3, let us consider $c, d \geq 0$ such that

$$cL + d\Lambda^*\Lambda \succeq Q^*Q \tag{6}$$

and let us set $\tau = d/(c+d)$. From (3), we notice that

$$\begin{aligned} \Delta_\tau\Lambda &= (cL + d\Lambda^*\Lambda)^{-1}d\Lambda^*\Lambda, \\ I - \Delta_\tau\Lambda &= (cL + d\Lambda^*\Lambda)^{-1}cL. \end{aligned}$$

Multiplying (6) on the right by $\left[I - \Delta_\tau\Lambda \,\middle|\, \Delta_\tau\Lambda\right]$, which equals $(cL + d\Lambda^*\Lambda)^{-1}\left[cL \,\middle|\, d\Lambda^*\Lambda\right]$, and on the left by its adjoint, we arrive at

$$\begin{aligned} &\left[\begin{array}{c} cL \\ \hline d\Lambda^*\Lambda \end{array}\right](cL + d\Lambda^*\Lambda)^{-1}\left[cL \,\middle|\, d\Lambda^*\Lambda\right] \\ &\qquad \succeq \left[\begin{array}{c} (I-\Delta_\tau\Lambda)^* \\ \hline (\Delta_\tau\Lambda)^* \end{array}\right] Q^*Q\left[I - \Delta_\tau\Lambda \,\middle|\, \Delta_\tau\Lambda\right]. \end{aligned} \tag{7}$$

First, we claim that the left-hand side of (7) takes the form $\left[\begin{array}{c|c} A-C & C \\ \hline C & B-C \end{array}\right]$ with $A = cL$ and $B = d\Lambda^*\Lambda$. To see this, it suffices to observe, e.g., that $A = cL$ is indeed the sum of its upper two blocks, which is clear since these blocks are $cL(cL + d\Lambda^*\Lambda)^{-1}cL$ and $cL(cL + d\Lambda^*\Lambda)^{-1}d\Lambda\Lambda^*$. Second, we claim that C can be written as $C = A(A+B)^{-1}B$, which is also clear—the relation $C = cL(cL + d\Lambda^*\Lambda)^{-1}d\Lambda\Lambda^*$ was just pointed out. Therefore, according to our two additional lemmas, the left-hand side of (7) does not exceed, in the positive semidefinite sense, $\left[\begin{array}{c|c} cL & 0 \\ \hline 0 & d\Lambda\Lambda^* \end{array}\right]$. At this point, we have shown that

$$\left[\begin{array}{c} (I-\Delta_\tau\Lambda)^* \\ \hline (\Delta_\tau\Lambda)^* \end{array}\right] Q^*Q\left[I - \Delta_\tau\Lambda \,\middle|\, \Delta_\tau\Lambda\right] \preceq \left[\begin{array}{c|c} cL & 0 \\ \hline 0 & d\Lambda^*\Lambda \end{array}\right],$$

which is equivalent to

$$\|Q(I - \Delta_\tau\Lambda)f + Q(\Delta_\tau\Lambda)g\|_2^2 \leq c\|L^{1/2}f\|_2^2 + d\|\Lambda g\|_2^2$$

for all $f, g \in \mathbb{R}^N$. The observation map Λ is obviously surjective in the present situation[2], so that any $e \in \mathbb{R}^{n_\ell}$ can be written as $e = \Lambda g$ for some $g \in \mathbb{R}^N$. From here, the desired result follows. □

§5. *Near optimality under mild overstimation of ε and η:* According to Theorem 1 (and its proof) and using the same notation, we have

$$\inf_{\Delta:\mathbb{R}^{n_\ell}\to\mathbb{R}^n} \mathrm{gwce}_Q(\Delta)^2 = c_\flat\varepsilon^2 + d_\flat\eta^2$$

while $c_\flat L + d_\flat\Lambda^*\Lambda \succeq Q^*Q$. Now suppose that ε and η are not exactly known but overestimated by $\bar{\varepsilon}$ and $\bar{\eta}$. Solving the semidefinite program (4) with $\bar{\varepsilon}$ and $\bar{\eta}$ provides a parameter $\bar{\tau} = \bar{d}/(\bar{c}+\bar{d})$ such that

$$\begin{aligned} \sup_{\substack{\|L^{1/2}f\|_2\leq\bar{\varepsilon} \\ \|e\|_2\leq\bar{\eta}}} &\|Q(f) - Q\circ\Delta_{\bar{\tau}}(\Lambda f + e)\|_2^2 \\ &= \min\left\{c\bar{\varepsilon}^2 + d\bar{\eta}^2 :\ cL + d\Lambda^*\Lambda \succeq Q^*Q\right\}. \end{aligned}$$

Since $\bar{\varepsilon} \geq \varepsilon$ and $\bar{\eta} \geq \eta$, we deduce in particular that

$$\sup_{\substack{\|L^{1/2}f\|_2\leq\varepsilon \\ \|e\|_2\leq\eta}} \|Q(f) - Q\circ\Delta_{\bar{\tau}}(\Lambda f + e)\|_2^2 \leq c_\flat\bar{\varepsilon}^2 + d_\flat\bar{\eta}^2.$$

Under the mild overestimations $\bar{\varepsilon} \leq C\varepsilon$ and $\bar{\eta} \leq C\eta$, this implies that

$$\begin{aligned} \mathrm{gwce}_Q(Q\circ\Delta_{\bar{\tau}})^2 &\leq C^2\left[c_\flat\varepsilon^2 + d_\flat\eta^2\right] \\ &= C^2 \inf_{\Delta:\mathbb{R}^{n_\ell}\to\mathbb{R}^n} \mathrm{gwce}_Q(\Delta)^2, \end{aligned}$$

proving that the recovery map $Q\circ\Delta_{\bar{\tau}}$ is globally near optimal.

§6. *No SDPs to optimal estimate a linear functional:* If $Q = \langle q, \cdot\rangle : \mathbb{R}^N \to \mathbb{R}$ is a linear functional, then solving the semidefinite program (4) and composing the resulting regularization map $\Delta_{\tau_\flat}$ with Q to obtain a globally optimal recovery map is quite wasteful. In such a situation, a globally optimal recovery map can be more directly obtained as $\langle a_\flat, \cdot\rangle$, where $a_\flat \in \mathbb{R}^{n_\ell}$ is a solution to

$$\underset{a\in\mathbb{R}^{n_\ell}}{\text{minimize}} \left[\sup_{\|L^{1/2}f\|_2\leq\varepsilon} |\langle q - \Lambda^*a, f\rangle| \times \varepsilon + \|a\|_2 \times \eta\right]. \tag{8}$$

[2]In general, it is always assumed that $\Lambda : \mathbb{R}^N \to \mathbb{R}^m$ is surjective, as it does not make sense to collect an observation that can be deduced from the others.

This laborious-looking optimization program can be turned into a more manageable one. For instance, if the graph G has connected components $C_1, \ldots, C_K$, then the eigenvalues of L are $0 = \lambda_1 = \cdots = \lambda_K < \lambda_{K+1} \leq \cdots \leq \lambda_N$. Denoting by $(v_1, \ldots, v_N)$ an orthonormal basis associated with these eigenvalues (so that $v_k = \mathbf{1}_{C_k}/\sqrt{|C_k|}$, $k = 1, \ldots, K$), the problem (8) reduces to

$$\begin{array}{l} \underset{a \in \mathbb{R}^{n_\ell}}{\text{minimize}} \left[\displaystyle\sum_{k=K+1}^{N} \frac{\langle q - \Lambda^* a, v_k \rangle^2}{\lambda_k} \right]^{1/2} \times \varepsilon + \|a\|_2 \times \eta \\ \text{s.to} \ \ \langle q - \Lambda^* a, v_k \rangle = 0, \quad k = 1, \ldots, K. \end{array}$$

Note that the vector $\Lambda^* a \in \mathbb{R}^N$ appearing above is just the vector $a \in \mathbb{R}^{n_\ell}$ padded with zeros on the unlabeled vertices.

§7. *A graph whose Laplacian is a scaled orthogonal projector:* Suppose that G is an unweighted graph (so that $w_{i,j} \in \{0, 1\}$ for all $i, j = 1, \ldots, N$) made of connected components $C_1, \ldots, C_K$ which are all complete graphs on an equal number n of vertices. The adjacency matrix W_k and graph Laplacian L_k of each C_k are

$$W_k = \begin{bmatrix} 0 & 1 & 1 & \cdots & 1 \\ 1 & 0 & 1 & \cdots & 1 \\ 1 & 1 & 0 & \ddots & \vdots \\ \vdots & \vdots & \ddots & \ddots & 1 \\ 1 & 1 & \cdots & 1 & 0 \end{bmatrix},$$

$$L_k = \begin{bmatrix} n-1 & -1 & -1 & \cdots & -1 \\ -1 & n-1 & -1 & \cdots & -1 \\ -1 & -1 & n-1 & \ddots & \vdots \\ \vdots & \vdots & \ddots & \ddots & -1 \\ -1 & -1 & \cdots & -1 & n-1 \end{bmatrix}.$$

Note that L_k has eigenvalue 0 of multiplicity 1 and eigenvalue n of multiplicity $n - 1$. Therefore, the whole graph Laplacian

$$L = \left[\begin{array}{c|c|c|c} L_1 & 0 & \cdots & 0 \\ \hline 0 & L_2 & \ddots & \vdots \\ \hline \vdots & \ddots & \ddots & 0 \\ \hline 0 & \cdots & 0 & L_K \end{array}\right]$$

has eigenvalue 0 of multiplicity K and eigenvalue n of multiplicity $(n-1)K$. This means that the renormalized Laplacian $(1/n)L$ is an orthogonal projector.

§8. *Proof of Theorem 4:* The argument is divided into three parts, namely:
a) there is a parameter $\tau_\natural \in (0, 1)$ yielding

$$\|L^{1/2} \Delta_{\tau_\natural}(y)\|_2 = \frac{\varepsilon}{\eta} \|\Lambda \Delta_{\tau_\natural}(y) - y\|_2, \tag{9}$$

and the corresponding reguralizer $\Delta_{\tau_\natural}(y)$ is a solution to

$$\underset{f \in \mathbb{R}^N}{\text{minimize}} \ \max \left\{ \|L^{1/2} f\|_2^2, \frac{\varepsilon^2}{\eta^2} \|\Lambda f - y\|_2^2 \right\}; \tag{10}$$

b) the optimization program (10) admits a unique solution $f_\natural$ (hence equal to $\Delta_{\tau_\natural}(y)$);
c) the solution $f_\natural$ to (10) does provide a near optimal local worst-case error.

Justification of **a)**. For any $\tau \in [0, 1]$, let f_τ denote $\Delta_\tau(y)$. Recalling that f_0 and f_1 are interpreted as

$$\begin{aligned} f_0 &= \underset{f \in \mathbb{R}^N}{\text{argmin}} \|\Lambda f - y\|_2 \quad \text{s.to } L^{1/2} f = 0, \\ f_1 &= \underset{f \in \mathbb{R}^N}{\text{argmin}} \|L^{1/2} f\|_2 \quad \text{s.to } \Lambda f = y. \end{aligned}$$

we have

$$\begin{aligned} \|L^{1/2} f_0\|_2 - \frac{\varepsilon}{\eta} \|\Lambda f_0 - y\|_2 &= -\frac{\varepsilon}{\eta} \|\Lambda f_0 - y\|_2 < 0, \\ \|L^{1/2} f_1\|_2 - \frac{\varepsilon}{\eta} \|\Lambda f_1 - y\|_2 &= \|L^{1/2} f_1\|_2 > 0. \end{aligned}$$

The continuity of $\tau \mapsto f_\tau = ((1-\tau)L + \tau \Lambda^* \Lambda)^{-1} (\tau \Lambda^* y)$ guarantees that there exists some $\tau_\natural \in (0, 1)$ satisfying $\|L^{1/2} f_{\tau_\natural}\|_2 - (\varepsilon/\eta) \|\Lambda f_{\tau_\natural} - y\|_2 = 0$, as announced in (9). We additionally point out that this $\tau_\natural$ is unique, which is a consequence of the facts that $\tau \mapsto \|L^{1/2} f_\tau\|_2$ is strictly increasing and that $\tau \mapsto \|\Lambda f_\tau - y\|_2$ is strictly decreasing. To see the former, say, recall that f_τ is the unique minimizer of $((1-\tau)/\tau) \|L^{1/2} f\|_2^2 + \|\Lambda f - y\|_2^2$. Therefore, given $\sigma < \tau$,

$$\begin{aligned} &\left(\frac{1}{\sigma} - 1\right) \|L^{1/2} f_\sigma\|_2^2 + \|\Lambda f_\sigma - y\|_2^2 \\ &\quad < \left(\frac{1}{\sigma} - 1\right) \|L^{1/2} f_\tau\|_2^2 + \|\Lambda f_\tau - y\|_2^2 \\ &\quad = \left(\frac{1}{\tau} - 1\right) \|L^{1/2} f_\tau\|_2^2 + \|\Lambda f_\tau - y\|_2^2 \\ &\qquad + \left(\frac{1}{\sigma} - \frac{1}{\tau}\right) \|L^{1/2} f_\tau\|_2^2 \\ &\quad < \left(\frac{1}{\tau} - 1\right) \|L^{1/2} f_\sigma\|_2^2 + \|\Lambda f_\sigma - y\|_2^2 \\ &\qquad + \left(\frac{1}{\sigma} - \frac{1}{\tau}\right) \|L^{1/2} f_\tau\|_2^2. \end{aligned}$$

Rearranging this inequality reads

$$\left(\frac{1}{\sigma} - \frac{1}{\tau}\right) \|L^{1/2} f_\sigma\|_2^2 < \left(\frac{1}{\sigma} - \frac{1}{\tau}\right) \|L^{1/2} f_\tau\|_2^2,$$

i.e., $\|L^{1/2} f_\sigma\|_2 < \|L^{1/2} f_\tau\|_2$, as expected. To finish, we now need to show that $f_{\tau_\natural}$ is a solution to (10). To this end, we remark on the one hand that the objective function of (10) evaluated at $f_{\tau_\natural}$ is

$$\max \left\{ \|L^{1/2} f_{\tau_\natural}\|_2^2, \frac{\varepsilon^2}{\eta^2} \|\Lambda f_{\tau_\natural} - y\|_2^2 \right\} = \gamma^2,$$

where γ is the common value of both terms in (9). On the other hand,

$$\begin{aligned} &\text{setting} && \tau_\natural' = \frac{(\eta^2/\varepsilon^2)\tau_\natural}{1 - \tau_\natural + (\eta^2/\varepsilon^2)\tau_\natural} \in [0, 1], \\ &\text{so that} && 1 - \tau_\natural' = \frac{1 - \tau_\natural}{1 - \tau_\natural + (\eta^2/\varepsilon^2)\tau_\natural} \in [0, 1], \end{aligned}$$

the objective function of (10) evaluated at any $f \in \mathbb{R}^N$ satisfies

$$\begin{aligned}
&\max\left\{\|L^{1/2}f\|_2^2, \frac{\varepsilon^2}{\eta^2}\|\Lambda f - y\|_2^2\right\}\\
&\geq (1-\tau_\natural')\|L^{1/2}f\|_2^2 + \tau_\natural'\frac{\varepsilon^2}{\eta^2}\|\Lambda f - y\|_2^2\\
&= \frac{1}{1-\tau_\natural+(\eta^2/\varepsilon^2)\tau_\natural}\left((1-\tau_\natural)\|L^{1/2}f\|_2^2 + \tau_\natural\|\Lambda f - y\|_2^2\right)\\
&\geq \frac{1}{1-\tau_\natural+(\eta^2/\varepsilon^2)\tau_\natural}\left((1-\tau_\natural)\|L^{1/2}f_{\tau_\natural}\|_2^2 + \tau_\natural\|\Lambda f_{\tau_\natural} - y\|_2^2\right)\\
&= \frac{1}{1-\tau_\natural+(\eta^2/\varepsilon^2)\tau_\natural}\left((1-\tau_\natural)\gamma^2 + \tau_\natural(\eta^2/\varepsilon^2)\gamma^2\right) = \gamma^2.
\end{aligned}$$

This justifies that $f_{\tau_\natural}$ is a solution to (10). □

Justification of **b)**. Here, we aim at showing that (10) admits a unique minimizer. Let $\widehat{f}$ and μ^2 denote a minimizer and the minimal value of (10), respectively. We first claim that

$$\|L^{1/2}\widehat{f}\|_2 = \frac{\varepsilon}{\eta}\|\Lambda\widehat{f} - y\|_2 = \mu. \tag{11}$$

Indeed, suppose e.g. that $\|L^{1/2}\widehat{f}\|_2 < (\varepsilon/\eta)\|\Lambda\widehat{f} - y\|_2 = \mu$. Pick an $h \in \mathbb{R}^N$ such that $\langle \Lambda\widehat{f} - y, \Lambda h\rangle \neq 0$ (which exists, for otherwise $\Lambda^*(\Lambda\widehat{f} - y) = 0$, hence $\Lambda\widehat{f} - y = \Lambda\Lambda^*(\Lambda\widehat{f} - y) = 0$, and so $\mu = 0$, in which case $\|L^{1/2}\widehat{f}\|_2 < \mu$ cannot occur). Then, considering $\widehat{f}_t := \widehat{f} + th$ for a small enough t in absolute value, we see that

$$\frac{\varepsilon}{\eta}\|\Lambda\widehat{f}_t - y\|_2 = \frac{\varepsilon}{\eta}\left(\|\Lambda\widehat{f} - y\|_2 + t\langle\Lambda\widehat{f} - y, \Lambda h\rangle + o(t)\right)$$

can be made smaller that μ, while $\|L^{1/2}\widehat{f}_t\|_2$ can remain smaller than μ. This contradicts the defining property of $\widehat{f}$ and establishes (11).

Now let $\widehat{f}$ and $\widetilde{f}$ be two minimizers of (10). Applying (11) to $\widehat{f}$, $\widetilde{f}$, and $(\widehat{f} + \widetilde{f})/2$, which is also a minimizer of (10), yields

$$\begin{aligned}
\left\|\frac{1}{2}L^{1/2}\widehat{f} + \frac{1}{2}L^{1/2}\widetilde{f}\right\|_2 &= \left\|L^{1/2}\widehat{f}\right\|_2 = \left\|L^{1/2}\widetilde{f}\right\|_2 = \mu,\\
\left\|\frac{1}{2}\Lambda(\widetilde{f} - y) + \frac{1}{2}\Lambda(\widehat{f} - y)\right\|_2 &= \left\|\Lambda\widehat{f} - y\right\|_2 = \left\|\Lambda\widetilde{f} - y\right\|_2 = \frac{\eta}{\varepsilon}\mu,
\end{aligned}$$

which forces $L^{1/2}\widehat{f} = L^{1/2}\widetilde{f}$ and $\Lambda\widehat{f} = \Lambda\widetilde{f}$, implying that i.e. $\widehat{f} - \widetilde{f} \in \ker(L^{1/2}) \cap \ker(\Lambda) = \ker(L) \cap \ker(\Lambda) = \{0\}$, i.e., that $\widehat{f} = \widetilde{f}$ is a unique minimizer. □

Justification of **c)**. Since the original signal $f \in \mathbb{R}^N$ that we try to recover satisfies $\|L^{1/2}f\|_2 \leq \varepsilon$ and $\|\Lambda f - y\|_2 \leq \eta$, it is clear that the minimizer $\widehat{f}$ of (10) satisfies $\|L^{1/2}\widehat{f}\|_2 \leq \varepsilon$ and $\|\Lambda\widehat{f} - y\|_2 \leq \eta$, too. In other words, it is model- and data-consistent, which always leads to near optimality of the local worst case error with a factor 2. Indeed, considering the set $\{Q(f)\colon f \in \mathcal{K}, e \in \mathcal{E}, \Lambda f + e = y\}$, let $f^\star$ denote its Chebyshev center (in our situation, it exists and is unique, see Garkavi [1962]). Then, for any $f \in \mathcal{K}$ and $e \in \mathcal{E}$ with $\Lambda f + e = y$, we have

$$\begin{aligned}
\|Q(f) - Q(\widehat{f})\|_2 &\leq \|Q(f) - Q(f^\star)\|_2 + \|Q(\widehat{f}) - Q(f^\star)\|_2\\
&\leq \mathrm{lwce}_Q(y, f^\star) + \mathrm{lwce}_Q(y, f^\star)\\
&= 2\inf_{z\in\mathbb{R}^N}\mathrm{lwce}_Q(y, z).
\end{aligned}$$

Taking the supremun over the admissable $f \in \mathcal{K}$ and $e \in \mathcal{E}$ gives $\mathrm{lwce}_Q(y, \widehat{f}) \leq 2\inf_{z\in\mathbb{R}^N}\mathrm{lwce}_Q(y, z)$, as desired. □

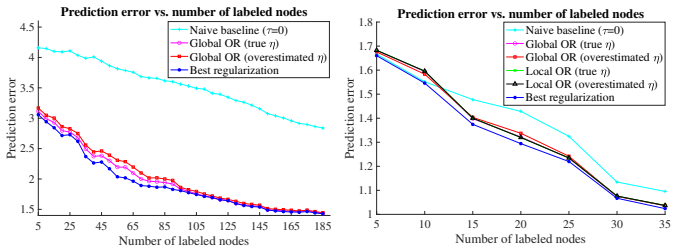
§9. *Implementation details and additional experiments:* We consider the following well-known graph datasets: adjnoun (112 nodes, 425 edges) [Newman, 2006], Netscience (379 nodes, 914 edges) [Girvan and Newman, 2002], polbooks (105 nodes, 441 edges) [Krebs], lesmis (77 nodes, 254 edges) [], and dolphins (62 nodes, 159 edges) [Lusseau et al., 2003]. All of these can be downloaded from the Suitesparse Matrix Collection [Davis and Hu, 2011]. When generating synthetic signals, we follow an approach similar to Equation (15) in [Dong et al., 2019]. Let $L = \chi D\chi^T$ be an eigendecomposition of the graph Laplacian, let $D^\dagger$ be the pseudoinverse of D, and let $c \sim \mathcal{N}(0, D^\dagger)$ be a Gaussian vector. The ground truth labels are then given by $f = \chi c$. The main difference with [Dong et al., 2019] is that f is not assumed to be corrupted simply by Gaussian noise, but we consider different additive noise vectors satisfying $\|e\|_2 \leq \eta$. In the main text, the plots were shown for a noise vector that is generated by taking a uniform random noise vector and subtracting the mean, and before scaling to ensure that $\|e\|_2 \leq \eta$. Here, to illustrate results of a more deterministic flavor, we show results for noise of magnitude proportional to the node degree (Figure 4) and to the inverse degree (Figure 5).

Keeping the same parameters as those used in the main text, we test several optimal recovery methods on different graphs and with different error models. Figures 3-5 support our conclusions that a mild overestimation of η does not lead to bad prediction error and that the prediction errors attached to the locally/globally optimal recovery maps are close to the smallest prediction error possible for any choice of regularization parameter. This confirms that these methods provide a suitable way to choose regularization parameters.

It is worth pointing out that the globally optimal recovery map is linear since the regularization parameter does not depend on the observation vector y. In contrast, the locally near optimal recovery map is nonlinear since the unique parameter $\tau_\natural$ satisfying

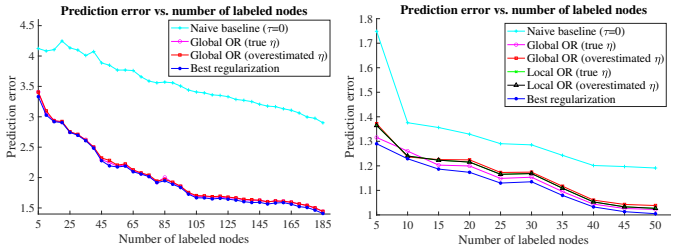
$$\|L^{1/2}\Delta_\tau(y)\|_2 = \frac{\varepsilon}{\eta}\|\Lambda\Delta_\tau(y) - y\|_2$$

does depend on the observation vector y. When implementing globally optimal recovery maps, we compute the globally optimal regularization parameter $\tau_\flat$ for each n_ℓ once and make a prediction when receiving different observation vectors y. For locally optimal recovery maps, we have to recompute the locally near optimal parameter $\tau_\natural$ when receiving new observation vectors y. Therefore, it is recommended to opt for the globally optimal recovery map in order to reduce computational complexity, see e.g. Figures 3(a), 4(a), and 5(a) where the locally near optimal recovery map is not executed. However, dealing with large graphs may result in semidefinite programs that cannot be run, so it can be better to implement the locally near optimal recovery map by using the bisection method to find the near optimal parameter $\tau_\natural$.



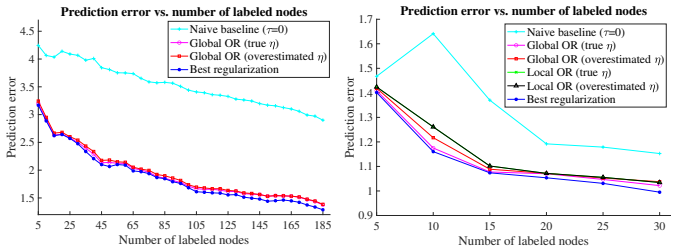
(a) (b)

Fig. 3. Prediction errors vs. number of labeled nodes on two different graphs with additive noise generated uniformly: (a) Netscience (b) lesmis.



(a) (b)

Fig. 4. Prediction errors vs. number of labeled nodes on two different graphs with additive noise proportional to degree: (a) Netscience (b) polbooks.



(a) (b)

Fig. 5. Prediction errors vs. number of labeled nodes on two different graphs with additive noise proportional to the inverse of degree: (a) Netscience (b) dolphins.

§10. *Numerical computation on the upper bound:* Given $y \in \mathbb{R}^{n_\ell}$, the square of the local worst-case error $\mathrm{lwce}_Q(y,z)$ for the estimation of Q by $z \in \mathbb{R}^n$ is

$$\sup_{f \in \mathbb{R}^N} \|Q(f)-z\|_2^2 \quad \text{s.to } \|L^{1/2}f\|_2^2 \le \epsilon^2, \|\Lambda f - y\|_2^2 \le \eta^2.$$

Introducing a slack variable γ, we write the above optimization program as

$$\begin{aligned} \inf_{\gamma} \gamma \quad & \text{s.to } \|Q(f)-z\|_2^2 \le \gamma \\ & \text{whenever } \|L^{1/2}f\|_2^2 \le \epsilon^2, \|\Lambda f - y\|_2^2 \le \eta^2. \end{aligned}$$

The constraint is a consequence of (but is not equivalent to) the existence of $c, d \ge 0$ such that

$$\|Q(f)-z\|_2^2 - \gamma \le c(\|L^{1/2}f\|_2^2 - \epsilon^2) + d(\|\Lambda f - y\|_2^2 - \eta^2)$$

for all $f \in \mathbb{R}^N$. The latter can be also reformulated as the condition that, for all $f \in \mathbb{R}^N$,

$$\begin{aligned} \langle (cL + d\Lambda^*\Lambda - Q^*Q)f, f\rangle - 2\langle Q^*z - \Lambda^* y, f\rangle \\ + \gamma - \|z\|_2^2 - c\epsilon^2 + d(\|y\|_2^2 - \eta^2) \ge 0, \end{aligned}$$

or, more succinctly, that

$$\left[\begin{array}{c|c} cL + d\Lambda^*\Lambda - Q^*Q & Q^*z - d\Lambda^* y \\ \hline (Q^*z - d\Lambda^* y)^* & \gamma - \|z\|_2^2 - c\epsilon^2 + d(\|y\|_2^2 - \eta^2) \end{array}\right] \succeq 0. \quad (12)$$

We conclude from the above considerations that the squared local worst-case error is upper-bounded by the optimal value of a semidefinite program, namely

$$\mathrm{lwce}_Q(y,z)^2 \le \inf_{\substack{\gamma \\ c,d \ge 0}} \gamma \quad \text{s.to (12).}$$