Binary Graph-Signal Recovery from Noisy Samples

Gita Babazadeh Eslamlou
Institute of Telecommunications
TU Wien, Vienna, Austria
Email: gita.babazadeh@nt.tuwien.ac.at

Norbert Goertz
Institute of Telecommunications
TU Wien, Vienna, Austria
Email: Norbert.Goertz@nt.tuwien.ac.at

***Abstract*—We study the problem of recovering a smooth graph signal from incomplete noisy measurements, using random sampling to choose from a subset of graph nodes. The signal recovery is formulated as a convex optimization problem. We reformulate the optimization problem in a way that the optimality conditions form a system of linear equations which is solvable via Laplacian solvers. We use an incomplete Cholesky factorization conjugate gradient (ICCG) method for graph signal recovery. Numerical experiments validate the performance of the recovery method over real-world blog-data of 2004 US election.**

I. Introduction

We consider an undirected symmetric weighted graph $\mathcal{G}_0 = (\mathcal{V}, \mathcal{E}, \mathbf{W})$ with a node set $\mathcal{V}$, an edge set $\mathcal{E}$ and the weighted adjacency matrix $\mathbf{W} \in R^{N \times N}$. Each non-zero component $W_{i,j}$ represents the strength of the connection between signal values x_i and x_j. For a given graph $\mathcal{G}_0$, a graph signal $\mathbf{x}$ is a labeling of the graph nodes with real numbers. A graph signal can be represented as a vector $\mathbf{x} \in R^N$ by having a vector component x_i to be the graph-signal value at node $i \in \mathcal{V}$.

Many applications of graph signal processing rely on smoothness: a graph signal is smooth when neighboring nodes have similar signal values. Smoothness can be calculated via the notion of the Laplacian quadratic function [1]:

$$\| \bigtriangledown_i \mathbf{x} \|_2^2 = \frac{1}{2} \sum_{j \in \mathcal{N}_i} W_{i,j} \left(x_j - x_i \right)^2 = \mathbf{x}^T \mathbf{L} \mathbf{x}, \tag{1}$$

where $\bigtriangledown_i \mathbf{x}$ is the graph gradient of $\mathbf{x}$ at vertex i, with $\mathcal{N}_i := \{ j \in \mathcal{V} \mid W_{i,j} \neq 0 \}$ the neighborhood of the node $i \in \mathcal{V}$ and $\mathbf{L} := \mathbf{D} - \mathbf{W}$ the graph Laplacian matrix with diagonal degree matrix $\mathbf{D}$ that has the diagonal elements $D_{i,i} = \sum_{j \in \mathcal{V}} W_{i,j}$.

II. Problem formulation and recovery solution

We deal with the problem of recovering a smooth graph signal $\mathbf{x} = \{x_i, i = 1, 2, ..., N\}$ from a subset of noisy graph-signal samples. The observation vector is given by $\mathbf{y} = \mathbf{A}\mathbf{x} + \mathbf{n}$. In this work, the noise vector $\mathbf{n} \sim N(0, \sigma^2)$ represents the effect of modeling and measurement errors, and the sampling process is represented by the measurement matrix $\mathbf{A} \in R^{M \times N}$, which mostly consists of zeros, with at most one non-zero entry in each column and exactly one non-zero entry in each row. Our recovery method is based on balancing the graph signal's smoothness (cf. (1)) with the empirical error, which yields the following convex optimization problem:

$$\hat{\mathbf{x}} = \arg\min_{\mathbf{x}} \frac{1}{2} ||\mathbf{A}\mathbf{x} - \mathbf{y}||_2^2 + \alpha \mathbf{x}^T \mathbf{L} \mathbf{x}. \tag{2}$$

The tuning parameter $\alpha > 0$ trades-off the two components. For recovery of a noisy graph signal $\mathbf{x}$, we need to solve the optimization problem (2), whose optimal points $\hat{\mathbf{x}}$ are given by the linear equation

$$\underbrace{\left(\mathbf{A}^T \mathbf{A} + \alpha \mathbf{L}\right)}_{\mathbf{C}} \hat{\mathbf{x}} = \underbrace{\mathbf{A}^T \mathbf{y}}_{\mathbf{b}} \tag{3}$$

The graph $\mathcal{G}_0$ with K connected components C_l, has a single unique optimal point $\hat{\mathbf{x}}$ of the optimization problem (2) if and only if the sampling set $\mathcal{S} \subseteq \mathcal{V}$ contains at least one node $r_l, l = 1, ..., K$, from each connected component C_l of the graph $\mathcal{G}_0$.

In order to obtain the recovered signal $\hat{\mathbf{x}}$, we have to solve a system of linear equations. An iterative method to solve large systems of linear equations such as (3) is the incomplete Cholesky factorization conjugate gradient (ICCG) method [2]. The CG algorithm requires a preconditioner matrix $\mathbf{M}$, which is usually an approximation of $\mathbf{C}^{-1}$. To improve the convergence of CG, the ideal preconditioner is the exact inverse of $\mathbf{C}$, hence a suitable preconditioner can be an approximation of the Cholesky factor. To find such a preconditioner we first calculate the lower triangular matrix $\mathbf{H}$ using the exact Cholesky decomposition algorithm; except here for each zero entry in $\mathbf{C}$, the corresponding entry in $\mathbf{H}$ should also be set to zero. This gives an incomplete Cholesky factorization of matrix $\mathbf{C}$, which is as sparse as the matrix $\mathbf{C}$. Matrix $\mathbf{H}$ is required to set $\mathbf{M} = \mathbf{H}\mathbf{H}^T$ as preconditioner of the CG. The preconditioner $\mathbf{M}$ attempts to improve the spectral properties of the coefficient matrix $\mathbf{C}$, so the matrix $\mathbf{M}^{-1}\mathbf{C}$ will be better conditioned and CG converges faster [3].

III. Numerical results and Conclusion

In order to asses the accuracy of the iterative recovery algorithm, we applied it to a real-world political blogs data-set [4]. This data-set consists information about left-leaning and right-leaning political blogs. Blogs are represented by the nodes of the graph and nodes representing the same political party are connected by an edge. Each blog is assigned a value, i.e., -1 for the right- and +1 for left-leaning blogs. The graph of the raw data contained 266 isolated nodes. We selected the largest connected subgraph $\mathcal{G}$ for our numerical experiments. In this graph there are $N = 1224$ nodes (588 left-leaning and 636 right-leaning blogs) and $|E| = 16661$ edges. Edge directions were omitted to obtain an undirected graph. We randomly selected M noise-contaminated signal samples x_i and this way obtained the measurement vector $\mathbf{y}$. For sufficient statistical significance of the results we ran the recovery method for 1000 times with $\alpha = 0.1$ (and each time with different noise realizations), and set the stopping criterion to a maximum of 100 iterations. The final result is averaged over the outcomes of the individual runs of the recovery scheme.

We analyzed the effect of different noise levels σ^2 and varying sampling rates (SR) M/N on the normalized mean squared error (NMSE), NMSE = $\frac{\|\mathbf{x}-\hat{\mathbf{x}}\|_2}{\|\mathbf{x}\|_2}$. The obtained results are shown in Table I: ICCG shows very good recovery performance especially for sampling rates (SR) $M/N > 0.1$.

Another metric considered besides the NMSE is the recovery ratio (RR) defined as the fraction of nodes $i \in \mathcal{V}$, for which the blogs' political leanings are correctly detected. To compute this metric we rounded the value of recovered signals to the nearest signal value i.e., $x_i \in \{-1, 1\}$. The obtained results for recovery ratio are shown in Table II: even for high noise variance σ^2 and low sampling rates M/N such as $\sigma^2 = 0.65$ and $M/N = 0.1$, the algorithm can recover correctly more than 80% of the binary graph signal values.

TABLE I: NMSE of the ICCG algorithm.

NMSE	$\sigma^2 = 0$	$\sigma^2 = 0.25$	$\sigma^2 = 0.5$	$\sigma^2 = 0.65$
SR=0.1	0.6107	0.7092	0.7629	0.7971
SR=0.2	0.1920	0.2129	0.2395	0.2623
SR=0.3	0.1448	0.1642	0.1957	0.2200
SR=0.4	0.1217	0.1397	0.1765	0.2038
SR=0.5	0.1022	0.1195	0.1592	0.1925
SR=0.6	0.0839	0.1009	0.1449	0.1799
SR=0.7	0.0678	0.0848	0.1339	0.1720
SR=0.8	0.0519	0.0690	0.1226	0.1652
SR=0.9	0.0353	0.0548	0.1164	0.1640

TABLE II: Recovery ratio (percentage) of the ICCG algorithm.

Recovery ratio	$\sigma^2 = 0$	$\sigma^2 = 0.25$	$\sigma^2 = 0.5$	$\sigma^2 = 0.65$
SR=0.1	84.7	82.3	80.9	80.1
SR=0.3	96.4	95.9	95.1	94.5
SR=0.5	97.5	96.8	96.0	95.1
SR=0.7	98.2	97.6	96.6	95.6
SR=0.9	99.1	98.6	97.0	95.9

References

[1] D. I. Shuman, S. K. Narang, P. Frossard, A. Ortega, and P. Vandergheynst, "The emerging field of signal processing on graphs: Extending high-dimensional data analysis to networks and other irregular domains," *IEEE Signal Processing Magazine*, vol. 30, no. 3, pp. 83–98, 2013.

[2] G. H. Golub and C. F. Van Loan, *Matrix Computations*, 3rd ed. Johns Hopkins University Press, 1996.

[3] M. Benzi, "Preconditioning Techniques for Large Linear Systems: A Survey," *Journal of Computational Physics*, vol. 182, no. 2, pp. 418–477, nov 2002.

[4] L. A. Adamic and N. Glance, "The political blogosphere and the 2004 US election: divided they blog," in *Proceedings of the 3rd international workshop on Link discovery*. ACM, 2005, pp. 36–43.