Exact Inference in Structured Prediction

Kevin Bello and Jean Honorio
{kbellome, jhonorio}@purdue.edu
Department of Computer Science, Purdue University, West Lafayette, IN, USA

Motivation

Structured prediction has several applications, to name a few:

- Image segmentation and matching in computer vision.
- Dependency parsing, part-of-speech tagging and named entity recognition in natural language processing.
- Protein folding in bioinformatics, among others.

Figure 1: Example of image segmentation from Nowozin et al. (2011)

Inference Problem

We focus in inference and assume parameters of the model are already learned. In the context of MRFs we have:

$$\max_{\boldsymbol{y}\in\mathcal{M}^{|\mathcal{V}|}} \sum_{v\in\mathcal{V},m\in\mathcal{M}} c_v(m)\mathbb{1}\Big[y_v = m\Big] + \sum_{(u,v)\in\mathcal{E},m_1,m_2\in\mathcal{M}} c_{u,v}(m_1,m_2)\mathbb{1}\Big[y_u = m_1, y_v = m_2\Big], \quad (1)$$

for an undirected graph $\mathcal{G} = (\mathcal{V},\mathcal{E})$, where $\mathcal{M}$ is the set of possible labels, $c_u(m)$ is the cost of assigning label m to node v and $c_{u,v}(m_1,m_2)$ is the cost of assigning m_1 and m_2 to the neighbors u,v respectively.

Eq.(1) is NP-Hard to solve in general. Motivated by the setting in (Globerson et al., 2015), we next define a generative process:

Definition 1 (Generative process). *Let us consider a n-node undirected connected graph $\mathcal{G} = (\mathcal{V},\mathcal{E})$, an edge noise $p \in (0, 0.5)$, a node noise $q \in (0, 0.5)$, and a ground truth node labeling $\boldsymbol{y}^*$ where $y_i^* \in \{+1,-1\}$. We then observe edge observations defined as $\mathrm{X}_{u,v} = y_u^* y_v^* \mathrm{z}_p^{(u,v)} \mathbb{1}\Big[(u,v)\in\mathcal{E}\Big]$ and node observations defined as $c_u = y_u^* \mathrm{z}_q^{(u)}$, where $\mathrm{z}_p^{(u,v)}$ and $\mathrm{z}_q^{(u)}$ are "biased" Rademacher variables with parameters p and q respectively, e.g., $P(\mathrm{z}_p = +1) = 1-p$, and $P(\mathrm{z}_p = -1) = p$.*

Given the process in Definition 1 we aim to solve the following combinatorial problem:

$$\max_{\boldsymbol{y}} \quad \frac{1}{2}\boldsymbol{y}^\top \mathbf{X}\boldsymbol{y} + \alpha \boldsymbol{c}^\top \boldsymbol{y} \quad \text{subject to} \quad y_i = \pm 1, \quad (2)$$

where $\alpha = \log\frac{1-q}{q}/\log\frac{1-p}{p}$. Note that Eq.(2) is still NP-Hard in general (Barahona, 1982).

Could there be any structural properties of the graph $\mathcal{G}$ that suffice to achieve, with high probability, exact recovery in polynomial time?

Exact Inference through SDP relaxation

A few more concepts used for proof of main result. Let $|\mathcal{E}(\mathcal{S},\mathcal{S}^C)|$ denote the number of edges between $\mathcal{S}$ and $\mathcal{S}^C$.

Definition 2 (Cheeger constant). *The edge expansion or Cheeger constant of a graph $\mathcal{G} = (\mathcal{V},\mathcal{E})$ is defined as: $\phi_{\mathcal{G}} = \min_{\mathcal{S}\subset\mathcal{V},|\mathcal{S}|\le n/2} |\mathcal{E}(\mathcal{S},\mathcal{S}^C)|/|\mathcal{S}|$.*

Intuitively, the Cheeger constant measures the "bottleneckedness" of a graph.

Definition 3 (Signed Laplacian). *For a graph $\mathcal{G}$, a signed Laplacian matrix, $\boldsymbol{M}$, is a symmetric matrix that satisfies $\boldsymbol{x}^\top \boldsymbol{M}\boldsymbol{x} = \sum_{(i,j)\in\mathcal{E}} (y_i x_i - y_j x_j)^2$, where $\boldsymbol{y}$ is an eigenvector of $\boldsymbol{M}$ with eigenvalue 0, and $y_i \in \{+1,-1\}$.*

Theorem 1. *Let $\mathcal{G}$ be a n-node undirected graph, let $\boldsymbol{M}$ be a signed Laplacian, and let $\lambda_1 \le \lambda_2 \le \cdots \le \lambda_n$ be the eigenvalues of $\boldsymbol{M}$. Then, we have that:*

$$\frac{\phi_{\mathcal{G}}^2}{4\Delta_{\max}} \le \lambda_2.$$

Note that Theorem 1 is more general than Cheeger (1969) classical result as it applies to signed Laplacian matrices.

Approach

We consider a similar 2-stage approach as in (Globerson et al., 2015):

1. Solve Eq.(2) by ignoring the linear term and obtain 2 optimal assignments $\boldsymbol{y}$ and $-\boldsymbol{y}$.
2. Use node observations to choose between $\boldsymbol{y}$ and $-\boldsymbol{y}$.

Stage 1

Ignoring the linear term in Eq.(2), here we want to solve:

$$\max_{\boldsymbol{y}} \quad \frac{1}{2}\boldsymbol{y}^\top \mathbf{X}\boldsymbol{y} \quad \text{subject to} \quad y_i = \pm 1, \quad (3)$$

The SDP relaxation of Eq.(3) is then defined as follows:

$$\max_{\boldsymbol{Y}} \quad \langle \mathbf{X}, \boldsymbol{Y}\rangle \quad \text{subject to} \quad Y_{ii} = 1,\ \boldsymbol{Y} \succeq 0. \quad (4)$$

Theorem 2. *Let $\mathcal{G} = (\mathcal{V},\mathcal{E})$ be an undirected connected graph with n nodes, Cheeger constant $\phi_{\mathcal{G}}$, and maximum node degree $\Delta_{\max}$. Then, for the combinatorial problem (3), a solution $\boldsymbol{y} \in \{\boldsymbol{y}^*, -\boldsymbol{y}^*\}$ is achievable in polynomial time by solving the SDP based relaxation (4), with probability at least $1-\epsilon_1(n,T,\Delta_{\max})$, where*

$$\epsilon_1(n,T,\Delta_{\max}) = 2n\cdot e^{\frac{-k_1 T^2}{k_2\Delta_{\max}+k_3 T}}$$

for some constants $k_1,k_2,k_3 > 0$, and $T = \phi_{\mathcal{G}}^2/\Delta_{\max}$.

For the proof of Theorem 2, we first use KKT optimality conditions to find sufficient conditions for when $Y = (y^*)(y^*)^\top$ is the unique optimal solution to Eq.(4). Second, we focus on showing when these conditions are fulfilled with high probability by using Theorem 1 and Bernstein matrix inequalities.

Stage 2

We output the vector $\boldsymbol{y}$ that maximizes the score $\boldsymbol{c}^\top\boldsymbol{y}$.

Theorem 3. *Let $\boldsymbol{y} \in \{\boldsymbol{y}^*, -\boldsymbol{y}^*\}$. Then, with probability at least $1-\epsilon_2(n,q)$, we have that: $\boldsymbol{c}^\top\boldsymbol{y}^* = \max_{\boldsymbol{y}\in\{\boldsymbol{y}^*,-\boldsymbol{y}^*\}} \boldsymbol{c}^\top\boldsymbol{y}$, where $\epsilon_2(n,q) = e^{-\frac{n}{2}(1-2q)^2}$ and q is the node noise.*

Combining Theorems 2 and 3, we obtain exact recovery with high probability for sufficiently smalls ϵ_1 and ϵ_2.

Examples

It is known that graphs with bad edge expansion are not suited for exact inference, e.g., grid graphs. (See Abbe et al. (2014).) Therefore, we provide the following examples of graphs that can achieve exact inference, where perhaps the most important is the following example related to smoothed analysis on connected graphs (Krivelevich et al., 2015).

Corollary 1 (Bad expanders + perturbations). *Let $\mathcal{G} = (\mathcal{V},\mathcal{E})$ be any connected graph (possibly a bad expander), choose $\widetilde{\mathcal{E}} \sim \mathsf{ER}(n, \log^8 n/n)$, let $\widetilde{\mathcal{G}} = (\mathcal{V}, \mathcal{E}\cup\widetilde{\mathcal{E}})$ and let $\Delta_{\max}^{\widetilde{\mathcal{G}}}$ be the maximum node degree of $\widetilde{\mathcal{G}}$. Then, we have that $\phi_{\widetilde{\mathcal{G}}}^2/\Delta_{\max}^{\widetilde{\mathcal{G}}} \in \Omega(\log^5 n)$ and $\Delta_{\max}^{\widetilde{\mathcal{G}}} \in \mathcal{O}(\log^9 n)$ with high probability. Therefore, exact recovery in polynomial time is achievable with high probability.*

Definition 4 (d-regular expander). *A d-regular graph with n nodes is an expander with constant $c > 0$ if, for every set $\mathcal{S} \subset \mathcal{V}$ with $|\mathcal{S}| \le n/2$, $|\mathcal{E}(\mathcal{S},\mathcal{S}^C)| \ge c\cdot d\cdot|\mathcal{S}|$.*

Corollary 2 (Expanders graphs). *Let $\mathcal{G}$ be a d-regular expander with constant c. Then, we have that $\phi_{\mathcal{G}}^2/\Delta_{\max} \in \Omega(d)$. If $d \in \Omega(\log n)$ then exact recovery in polynomial time is achievable with high probability.*

Corollary 3 (Complete graphs). *Let $\mathcal{G} = \mathcal{K}_n$, where $\mathcal{K}_n$ denotes a complete graph of n nodes. Then, we have that $\phi_{\mathcal{G}}^2/\Delta_{\max} \in \Omega(n)$. Therefore, exact recovery in polynomial time is achievable with high probability.*

References

Abbe, E., A. S. Bandeira, A. Bracher, and A. Singer
2014. Decoding binary node labels from censored edge measurements: Phase transition and efficient recovery. *IEEE Transactions on Network Science and Engineering*, 1(1):10–22.

Barahona, F.
1982. On the computational complexity of ising spin glass models. *Journal of Physics A: Mathematical and General*, 15(10):3241.

Cheeger, J.
1969. A lower bound for the smallest eigenvalue of the laplacian. In *Proceedings of the Princeton conference in honor of Professor S. Bochner*.

Globerson, A., T. Roughgarden, D. Sontag, and C. Yildirim
2015. How hard is inference for structured prediction? In *International Conference on Machine Learning*, Pp. 2181–2190.

Krivelevich, M., D. Reichman, and W. Samotij
2015. Smoothed analysis on connected graphs. *SIAM Journal on Discrete Mathematics*, 29(3):1654–1669.

Nowozin, S., C. H. Lampert, et al.
2011. Structured learning and prediction in computer vision. *Foundations and Trends® in Computer Graphics and Vision*, 6(3–4):185–365.

Work supported by the National Science Foundation under Grant No. 1716609-IIS.