How Hard Is Inference for Structured Prediction?

Tim Roughgarden (Stanford University)

joint work with Amir Globerson (Tel Aviv), David Sontag (NYU), and Cafer Yildirum (NYU)
Structured Prediction

- **structured prediction**: predict labels of many objects at once, given information about relationships between objects

- **applications**:
  - computer vision (objects = pixels, prediction = image segmentation)

(Borkar et al., 2013)
Structured Prediction

- **structured prediction**: predict labels of many objects at once, given information about relationships between objects
- **applications**:  
  - computer vision (objects = pixels, prediction = image segmentation)  
  - NLP (objects = words, prediction = parse tree)  
  - etc.
- **today’s focus**: complexity of inference (given a model), not of learning a model
Recovery From Exact Parities

Setup: (noiseless)
• known graph $G = (V,E)$
• unknown labeling $X:V \rightarrow \{0,1\}$
• given parity of each edge
  – “+” if $X(u) = X(v)$, “-” otherwise

Goal: recover $X$.

Solution: (for connected $G$) label some vertex arbitrarily and propagate.
Recovery From Noisy Parities

Setup: (with noise)

• known graph $G = (V,E)$
• unknown labeling $X: V \rightarrow \{0,1\}$
• given noisy parity of each edge
  – flipped with probability $p$

Goal: (approximately) recover $X$.

Formally: want algorithm $A: \{+,-\}^E \rightarrow \{0,1\}^V$ that minimizes worst-case expected Hamming error:

$$\max_X \{ E_{L \sim D(X)}[\text{error}(A(L), X)] \}$$
Research Agenda

Formally: want estimator $A: \{+,-\}^E \rightarrow \{0,1\}^V$ that minimizes worst-case expected Hamming error:

$$\max_X \{E_{L \sim D(X)}[error(A(L),X)]\}$$

- **(Info-theoretic)** What is the minimum expected error possible? How does it depend on $p$? Or on the structure of the graph?
- **(Computational)** When can approximate recovery be done efficiently? How does the answer depend on $p$ and the graph?
Analogy: Stochastic Block Model

Stochastic Block Model Setup:
[Boppana 87], [Bui/Chaudhuri/Leighton/Sipser 92] [Feige/Killian 01], [McSherry 01], [Mossel/Neeman/Sly 13,14], [Massoulié 14], [Abbe/Bandeira/Hall 15], [Makarychev/Makarychev/Vijayaraghavan 15,16], [Moitra/Perry/Wein 16] ...

• known vertices V, unknown labeling X:V -> \{0,1\}
• for *every pair* v,w, get noisy signal if X(u)=X(v)
  – no noise => get two disjoint cliques
  – noise = 1-a/n if X(u)=X(v), = b/n if X(u)≠X(v)  [a > b]

Goal: (approximately) recover X.
The Graph Matters

Example: $G = \text{path graph with } n \text{ vertices.}$

- ground truth = $1^{\text{st}} i$ vertices are 0, last $n-i$ vertices are 1 ($i$ is unknown)

- $p = \Theta(\log n/n)$ [very small]

- w.h.p., input has $\Theta(\log n)$ “-” edges; only one is consistent with the data

- no algorithm can reliably guess the true “-” edge; $\Theta(n)$ expected error unavoidable
The Grid: Empirical Evidence

Node noise = 0.45, Edge noise = 0.08 (1000 samples)

20x20 grid
Approximate Recovery

Definition: a family of graphs allows approximate recovery if there exists an algorithm with expected error $f(p)n$, where $f(p) \to 0$ as $p \to 0$ [for n large]

• non-example: paths.
• potential example: grids.

Questions:
• which graphs admit approximate recovery?
• in poly-time? with what functions $f(.)$?
Our algorithm: compute labeling minimizing:

- number of “+” edges with bichromatic endpoints
  + number of “-” edges with monochromatic endpoints

Fact: Can be implemented in polynomial time in planar (and bounded genus) graphs.

Fact: Not information-theoretically optimal.
  - optimal: marginal inference
Flipping Lemma

Definition: A bad set $S$ is a maximal connected subgraph of mislabeled nodes. (w.r.t. $X$, $A(L)$)
**Flipping Lemma**

**Definition:** A *bad set* $S$ is a maximal connected subgraph of mislabeled nodes. (w.r.t. $X, A(L)$)

**Lemma:** $S$ bad $\Rightarrow$ at least half the edges of $\delta(S)$ were corrupted.

**Proof idea:** (i) Our algorithm gets $\geq$ half the edges of $\delta(S)$ correct w.r.t. input (else flipping $S$ improves alleged optimal solution). (ii) But we get all the edges of $\delta(S)$ wrong w.r.t. the ground truth.
Warm-Up #1: Expanders

Graph family: d-regular expander graphs, with $|\delta(S)| \geq c \cdot d \cdot |S|$ (for constant $c$, for all $|S| \leq n/2$)

Analysis: let bad sets = $C_1, ..., C_k$.

- maximal connected subgraphs of mislabeled nodes

- note: error = $\Sigma_i |C_i|$

- flipping lemma => at least half of the $\geq c \cdot d \cdot |C_i|$ edges of $\delta(C_i)$ were corrupted

- $E[error] \leq 4 \cdot E[\# \text{ corrupted edges}]/(c \cdot d)$
  \[ \leq 2pn/c \] [so $f(p) = 2p/c$]
Open Questions

Open Question #1: polynomial-time recovery.
- correlation clustering NP-hard for expanders
- roughly equivalent to min multicut [Demaine/Emanuel/Fiat/Immorlica 05]
- does semidefinite programming help?

Open Question #2: determine optimal error rate.
- upper bound works even in a bounded adversary model (budget of $|E|$ edges to corrupt)
- expected error $O(pn)$ optimal for adversarial case
- conjecture: $O(p^{d/2}n)$ is tight for random errors
Warm-Up #2: Large Min Cut

Graph family: graphs with global min cut $\Omega(\log n)$.

Easy (Chernoff): for a connected subgraph $S$ with $|\delta(S)| = i$, $\Pr[S$ is bad$] \approx p^{i/2}$.

Key fact: (e.g., [Karger 93]) for every $\alpha \geq 1$, number of $\alpha$-approximate minimum cuts is at most $n^{2\alpha}$.

Result:

$$E[error] \leq \sum_{i=c^*}^{\infty} \sum_{S: |\delta(S)| = i} |S| \cdot \Pr[S \text{ bad}]$$

$$\leq n \sum_{i=c^*}^{\infty} n^{2(i/c^*)} p^{i/2} = o(1)$$
Grid-Like Graphs

Graph family: $\sqrt{n} \times \sqrt{n}$ grids.

Key properties:

- planar, each face has $O(1)$ sides
- weak expansion: for every $S$ with $|S| \leq n/2$, $|\delta(S)| \geq |S|^c$ (some $c > 0$)

Theorem: computationally efficient recovery with expected Hamming error $O(p^2 n)$.
- information-theoretic lower bound: $\Omega(p^2 n)$
  - 4-regular $=>$ each node ambiguous with prob $\approx p^2$
Grid-Like Graphs: Analysis

Lemma: number of connected subgraphs $S$ with $|\delta(S)| = i$ is at most $\approx n \cdot 3^i$.

Proof sketch: suffices to count cycles in the dual graph (also a grid). $n$ choices for first vertex; $\leq 3$ choices for each successive vertex.

Result:

$$E[error] \leq \sum_{i=4}^{\infty} \sum_{S:|\delta(S)|=i} |S| \cdot \Pr[S \text{ bad}]$$

$$\leq n \sum_{i=4}^{\infty} 3^i p^{i/2} = O(np^2)$$
Bug: forgot about connected sets $S$ like

![Diagram of connected sets $S$ and $G$]
A Subtlety and a Fix

**Generalized Flipping Lemma:**
for connected sets $S$ “with holes,” at least half of edges of outer boundary corrupted.
[else, flip all labels inside outer boundary]
=> charge errors in $S$ to its “filled-in version” $F(S)$

$$E[error] \leq \sum_{i=4}^{\infty} \sum_{|S|:|\delta(S)|=i} |S| \cdot \Pr[S \text{ bad}]$$

sum only over filled-in sets

$$\leq n \sum_{i=4}^{\infty} 3^i p^{i/2} = O(np^2)$$
More Open Questions

1. Characterize graphs where good approximate recovery is possible (as noise -> 0).
   - is “weak expansion” sufficient?

2. Computationally efficient recovery beyond planar graphs. (or hardness results)
   - does semidefinite programming help?

3. Take advantage of noisy node labels.
   - major progress: [Foster/Reichman/Sridharan 16]

4. More than two labels.