How Hard is Inference for Structured Prediction?

David Sontag

Joint work with Amir Globerson, Tim Roughgarden, and Cafer Yildirim
Structured Prediction

Computer vision
*Image segmentation*

*Stereopsis*

Natural language processing
*Parsing*

*input: image*  
*output: segmentation*

*input: two images*  
*output: disparity*

*input: sentence*  
*output: dependency parse*
Structured Prediction

- **Input:** \( x \in \mathcal{X} \)

  **Output:** labeling \( y \in \mathcal{Y} \)

- Given an input \( x \), the “goodness” of a prediction \( y \) is characterized by a score function \( s(x, y) \) such that

  \[
  s(x, y) = \begin{cases} 
  \text{High if } y \text{ is a good labeling for } x \\
  \text{Low if } y \text{ is a bad labeling for } x 
  \end{cases}
  \]

- Pairwise models have a score that decomposes over edges of a graph, e.g.

  \[
  s(x, y) = \sum_{ij \in E} s_{ij}(x, y_i, y_j) + \sum_{i \in V} s_i(x, y_i)
  \]
Structured Prediction

- Input: \( x \in \mathcal{X} \)
  
  Output: labeling \( y \in \mathcal{Y} \)

- Given an input \( x \), the “goodness” of a prediction \( y \) is characterized by a score function \( s(x, y) \) such that

\[
s(x, y) = \begin{cases} 
  \text{High if } y \text{ is a good labeling for } x \\
  \text{Low if } y \text{ is a bad labeling for } x
\end{cases}
\]

- Consider the following distribution over labelings:

\[
\Pr(y \mid x) = \frac{1}{Z(x)} \exp \left\{ \sum_{i,j \in E} s_{ij}(x, y_i, y_j) + \sum_{i \in V} s_i(x, y_i) \right\}
\]

- Conditional random fields (Lafferty et al. ’01) use maximum likelihood learning, and predict using \textit{marginal inference}

\[
\arg \max_{y_i} \Pr(y_i \mid x) \text{ for all } i
\]
Structured Prediction

- **Input:** \( x \in \mathcal{X} \)
- **Output:** labeling \( y \in \mathcal{Y} \)

- Given an input \( x \), the “goodness” of a prediction \( y \) is characterized by a score function \( s(x,y) \) such that
  \[
  s(x,y) = \begin{cases} 
  \text{High if } y \text{ is a good labeling for } x \\
  \text{Low if } y \text{ is a bad labeling for } x 
  \end{cases}
  \]

- Consider the following distribution over labelings:
  \[
  \Pr(y \mid x) = \frac{1}{Z(x)} \exp \left\{ \sum_{ij \in E} s_{ij}(x, y_i, y_j) + \sum_{i \in V} s_i(x, y_i) \right\}
  \]

- Max-margin learning (Collins ‘02, Taskar et al. ’03, Tsochantaridis et al. ‘05) seeks large margin, and predicts using **MAP inference**
  \[
  \arg \max_y \Pr(y \mid x)
  \]
Inference is NP-hard. So why does approximate inference work so well?

- Both marginal and MAP inference are in general NP-hard
- Nonetheless, heuristic inference algorithms can get state-of-the-art results for structured prediction

![Stereo vision](Image)

Inference is NP-hard. So why does approximate inference work so well?

- Both marginal and MAP inference are in general NP-hard
- Nonetheless, heuristic inference algorithms can get state-of-the-art results for structured prediction

**Foreground-background segmentation**

(Borenstein & Ullman ‘02, Domke ‘13)
Inference is NP-hard. So why does approximate inference work so well?

- Both marginal and MAP inference are in general NP-hard
- Nonetheless, heuristic inference algorithms can get state-of-the-art results for structured prediction Why?
- These instances do not correspond to any known tractable family (they are not tree-structured, submodular, ...)
- Intuitively, however, they are close to something tractable
- **This paper**: We demonstrate a setting in which approximate inference algorithms provably obtain small Hamming error,

\[
H(Y, \hat{Y}) = \sum_{i=1}^{N} 1[\hat{Y}_i \neq Y_i]
\]

\(Y\): Ground truth
\(\hat{Y}\): Prediction by approx inf
Key questions for theoretical analysis

• What are the information theoretic limits?
• What are the computational & statistical trade-offs?
  – How much worse is MAP inference compared to marginal inference?
  – What is the best prediction accuracy attainable in polynomial time?
  – Provable guarantees for linear programming relaxations?
Generative process

- Goal is to predict a set of labels $Y_1, ..., Y_N$, $Y_i \in \{-1, 1\}$, from observations $X$
- Our analysis assumes observations $X$ generated from $Y$ by the following process on graph $G=(V,E)$:
  - $X_i = -Y_i$ with probability $q$, and $X_i = Y_i$ otherwise
  - For $i,j \in E$, $X_{ij} = -Y_i Y_j$ with probability $p$, and $X_{ij} = Y_i Y_j$ otherwise

$q = \text{node noise}$
$p = \text{edge noise}$
Generative process

- Goal is to predict a set of labels $Y_1, \ldots, Y_N$, $Y_i \in \{-1, 1\}$, from observations $X$
- Our analysis assumes observations $X$ generated from $Y$ by the following process on graph $G=(V,E)$:
  - $X_i = -Y_i$ with probability $q$, and $X_i = Y_i$ otherwise
  - For $i,j \in E$, $X_{ij} = -Y_i Y_j$ with probability $p$, and $X_{ij} = Y_i Y_j$ otherwise

We focus on setting where the node noise $q$ is close to $\frac{1}{2}$, i.e. there is no correlation decay and global inference is essential.
Generative process

- Goal is to predict a set of labels $Y_1, \ldots, Y_N$, $Y_i \in \{-1, 1\}$, from observations $X$.
- Our analysis assumes observations $X$ generated from $Y$ by the following process on graph $G=(V,E)$:
  - $X_i = -Y_i$ with probability $q$, and $X_i = Y_i$ otherwise.
  - For $i,j \in E$, $X_{ij} = -Y_i Y_j$ with probability $p$, and $X_{ij} = Y_i Y_j$ otherwise.
- The maximum likelihood (ML) estimator is:
  $$\max_Y \sum_{uv \in E} \frac{1}{2} X_{uv} Y_u Y_v \log \frac{1 - p}{p} + \sum_{v \in V} \frac{1}{2} X_{uv} Y_u \log \frac{1 - q}{q}$$
- Even when $G$ is a planar graph, this maximization problem is NP-hard (reduction from max-cut)
Generative process

- Goal is to predict a set of labels $Y_1, \ldots, Y_N$, $Y_i \in \{-1, 1\}$, from observations $X$
- Our analysis assumes observations $X$ generated from $Y$ by the following process on graph $G=(V,E)$:
  - $X_i = -Y_i$ with probability $q$, and $X_i = Y_i$ otherwise
  - For $i,j \in E$, $X_{ij} = -Y_i Y_j$ with probability $p$, and $X_{ij} = Y_i Y_j$ otherwise
- The maximum likelihood (ML) estimator is:
  $$\max_Y \sum_{uv \in E} \frac{1}{2} X_{uv} Y_u Y_v \log \frac{1-p}{p} + \sum_{v \in V} \frac{1}{2} X_v Y_v \log \frac{1-q}{q}$$
- Even when $G$ is a planar graph, this maximization problem is NP-hard (reduction from max-cut)
Relating the generative process to CRFs

Conditional random field for foreground-background segmentation

\[
\Pr(\hat{Y} | Z) \propto \exp( \sum_{uv \in E} \beta_{uv} \hat{Y}_u \hat{Y}_v + \sum_{u \in V} \beta_u \hat{Y}_u )
\]

with image-dependent weights

\[
\beta_{uv} = f_{uv}(Z; \theta) \quad f \text{ is a linear function of features of } Z \text{ and parameters } \theta
\]

\[
\beta_u = f_u(Z; \theta)
\]

\[
\beta_{uv} \approx X_{uv} \frac{1}{2} \log \frac{1-p}{p}
\]

\[
\beta_u \approx X_u \frac{1}{2} \log \frac{1-q}{q}
\]

Input image \(Z\)

Compare to:

\[
\max_Y \sum_{uv \in E} \frac{1}{2} X_{uv} Y_u Y_v \log \frac{1-p}{p} + \sum_{v \in V} \frac{1}{2} X_u Y_u \log \frac{1-q}{q}
\]
Empirical study of inference

- Ground truth = all -1’s
- Node noise $q=0.4$
- Results averaged over 100 trials

Pairwise LP relaxation of MAP inference
- Does poorly for large edge noise!
- LP solution is $(\frac{1}{2}, \frac{1}{2})$ fractional

Marginal inference
- Information theoretically optimal
- NP-hard, but for 20x20 grid can compute exactly
Empirical study of inference

- Ground truth = all -1’s
- Node noise $q=0.4$
- Results averaged over 100 trials

- Pairwise LP relaxation of MAP inference
  - Does poorly for large edge noise!
  - LP solution is $(\frac{1}{2}, \frac{1}{2})$ fractional

- Cycle LP relaxation of MAP inference
  - Sontag et al., UAI 2012

- Marginal inference
  - Information theoretically optimal
  - NP-hard, but for 20x20 grid can compute exactly

20 x 20 grid graph

(a) Ground truth
(b) Observed evidence
(c) Approximate recovery
What are the information theoretic limits?

- **Theorem (lower bound):** Every algorithm must have error $\Omega(p^2N)$, where $N$ is the number of nodes.

- **Proof sketch:**
  
  (a) Consider the following distribution over $Y$ (ground truth)
  
  Shaded nodes fixed to -1.
  
  White nodes sampled uniformly, +1 with prob. ½, -1, otherwise.

  (b) Call a node *ambiguous* if exactly two of its edge observations are $\neq$ (i.e., -1) and two are $=$ (i.e. +1)
  
  How many?
  
  $$\frac{N}{2} \binom{4}{2} p^2 (1-p)^2 \approx \frac{5N}{2} p^2$$

  (c) Best is to predict according to node observation. Will be wrong with probability $q$

  (d) $E[H] \geq \frac{5N}{2} p^2 q$, i.e. $\Omega(p^2N)$

  $q = \text{node noise}$

  $p = \text{edge noise}$
Two-stage approximate inference

- **We analyze the following approximate inference algorithm:**

  **Require:** Edge and node observations $X$

  1. $\hat{Y} \leftarrow \arg \max_Y \sum_{uv \in E} X_{uv} Y_u Y_v$
  2. if $\sum_{v \in V} X_u \hat{Y}_v < 0$ then
  3. $\hat{Y} \leftarrow -\hat{Y}$
  4. end if

  **Stage 1** (uses only edge observations)

  **Stage 2**

- **MAP inference for Stage 1** is polynomial time using matching (Fisher ’66) or solving cycle LP (Barahona ’82)

- **Intuition:** after stage 1, either $\hat{Y}$ or its flip $-\hat{Y}$ is close in Hamming distance to the ground truth:

  ![Diagram showing two possible configurations for stage 2]

  We choose one of these by looking at the node observations (stage 2)
Two-stage approximate inference

- Ground truth = all -1’s
- Node noise $q=0.4$
- Results averaged over 100 trials

Pairwise LP relaxation of MAP inference

Two-stage approximate inference

Cycle LP relaxation of MAP inference
- Sontag et al., UAI 2012

Marginal inference
- Information theoretically optimal
- NP-hard, but for 20x20 grid can compute exactly
Two-stage algorithm is optimal for grids

- **Theorem (upper bound):** The two-stage algorithm obtains error $O(p^2N)$ when $p < 0.017$
Key structural lemma

- Let $\delta(S)$ denote the outer boundary of a set of vertices $S$
- An edge is bad if $X_{uv} = -Y_u Y_v$
- **Lemma 1 (Flipping Lemma):** Let $S$ denote a maximal connected subgraph of $G$ with every node of $S$ mispredicted by $\hat{Y}$. Then, at least half the edges of $\delta(S)$ are bad

**Example:**

- Suppose ground truth $Y$ is all -1, and we mispredicted the middle node $\hat{Y}_1$
- Suppose for contradiction that all four edges of $\delta(S)$ are “=” (i.e., *not* bad)
- Flipping $\hat{Y}_1$ to -1 strictly improves the objective, contradicting optimality of $\hat{Y}$

$$\hat{Y} \leftarrow \arg \max_Y \sum_{uv \in E} X_{uv} Y_u Y_v$$
Bounding number and size of maximally connected mispredicted sets

- Let $\delta(S)$ denote the outer boundary of a set of vertices $S$

- A set $S$ is bad if at least half its outer boundary $\delta(S)$ is bad
- **Lemma 2:** For every set $S$ with $|\delta(S)| = k$, $\Pr[S \text{ is bad}] \leq (9p)^{k/2}$
- **Lemma 3:** For every set $S$, $|S| \leq c|\delta(S)|^2$
- **Lemma 4:** There are at most $4N3^{k-2}/(2k)$ sets with $|\delta(S)| = k$ for even length $k$ (and zero for odd $k$)
- Many large sets (Lemma 3+4), but unlikely to be bad (Lemma 2)
  
  Result is then shown using a Union Bound.
Discussion & Conclusions

• Results extend to other generative processes, planar graphs and d-regular expander graphs

• **Take away 1:** Think about approximate inference for structured prediction in terms of *recovering ground truth*

• **Take away 2:** When using dual decomposition or LP relaxations, look for tractable *and accurate* components

• Many open problems
  
  – Non-binary models (e.g., for stereo vision), and other prediction tasks such as dependency parsing
  
  – Analysis of cycle LP relaxation: might need new proof techniques
Extra slides
Error of an algorithm

• The error of an algorithm $A$ is defined to be the worst-case (over $Y$) expected Hamming error:

$$err(A) = \max_y \mathbb{E}_{X|Y=y} [H(y, A(X))]$$

• Marginal inference using a uniform prior for $Y$ can be shown to be minimax optimal
  • Statistically efficient, but not computationally efficient
Theorem (upper bound): The two-stage algorithm obtains error $O(p^2N)$

$$H = \sum_{\text{cycles } C : \delta(S) = C} \sum_{S : |S| = |\delta(S)| = k} |S| \mathbb{1}[S \text{ is maximally connected mispredicted set}]$$

$$\leq \sum_{k=4,6,8,...} \left( \max_{S : |\delta(S)| = k} |S| \right) \sum_{\text{cycles } C : |C| = k} 1 \left[ \text{at least half of edges in } C \text{ are bad} \right]$$

Lemma 1

$$\leq \sum_{k=4,6,8,...} k^2 \sum_{\text{cycles } C : |C| = k} 1 \left[ \text{at least half of edges in } C \text{ are bad} \right]$$

Lemma 3

$$E[H] \leq \sum_{k=4,6,8,...} k^2 \cdot (9p)^{k/2} \cdot 4N3^{k-2}/(2k)$$

Lemma 2

$$\approx N \sum_{k=4,6,8,...} k \cdot (9p)^{k/2}3^k = N \sum_{l=2}^{\infty} 2l \cdot (9p)^l9^l \approx N \sum_{l=2}^{\infty} l(81p)^l = O(p^2N)$$

Lemma 4

(Using results from percolation, can substantially improve constants)
Generalizations

• Planar graphs
  – Use two-step algorithm: still polynomial time
  – Need two properties
    • Weak expansion: $|F| \leq c_1|\delta(F)|^{c_2}$, for every set $F$
    • Bounded dual degree
      (used in bounding the number of sets)

• d-regular expander graphs
  – Use two-step algorithm: not computationally efficient
  – Expected Hamming error $O(Np)$: different analysis