Simplicity and Complexity of Belief-Propagation

Elchanan Mossel\(^1\)

\(^1\)MIT

July 2020
A simple Mathematical model for Phylogenetic reconstruction

- Consider broadcast process on trees for $h$ levels $X_h$ and $d = 2$.
- Unknown permutation $\sigma \in S_{2^d}$.
- Input: i.i.d samples from $Y_s \sim \tilde{X}_h, 1 \leq s \leq m$, where $\tilde{X}_h(i) = X_h(\sigma(i))$.
- Goal: recover $T$, i.e. $\sigma \mod \Gamma$, where $\Gamma =$ ways to draw.
- E.G: 3 possible trees on when $h = 2$ and $7 \times 5 \times 3 \times 3$ when $h = 3$. 

Elchanan Mossel  
Simplicity & Complexity of BP
Estimate the covariance $r_{i,j} = \text{Cov}[\tilde{X}_h(i), \tilde{X}_h(j)]$.

Identify siblings as maximizing correlation.

For each sample $i$, let $Z_i$ be a $2^{d-1}$ dimensional vector where

$$Z_i(w) = \text{maj}(Y_v : v \text{ descendant of } w)$$

Repeat.

Let $p(m, h) := \text{probability of recovering the tree from } m$ samples.

**Exercise:** If $2\theta^2 > 1$, and $m \geq C_\theta h$, then $p(m, h) \geq 0.9$.

**Exercise:** If $2\theta^2 < 1$, then $p(m, h) \leq mc_\theta^h$, where $c_\theta < 1$. 

2\theta^2 < 1 \implies \text{need exp}(Ch) \text{ samples to recover the tree}

Exercise: \[ \|P_T^+ - P_T^-\|_{TV} \leq 2E_T[|M_h|] \leq 2 \times (2\theta^2)^{h/2} \]

\implies \text{If two } h + 2\text{-level trees } T, T' \text{ have the same topology in the last } h \text{ levels then:}

\[ \|X_{h+2} - X'_{h+2}\|_{TV} \leq 8 \times (2\theta^2)^{h/2} \implies \]

\[ \|(X_{h+2})^\otimes m - (X'_{h+2})^\otimes m\|_{TV} \leq 8m \times (2\theta^2)^{h/2} \implies \]

To distinguish between two topologies need at least \( m = \Omega((2\theta^2)^{-h/2}) \) samples.
Application 2: The Block Model

- **Random graph** $G = (V, E)$ on $n$ nodes.
- Half blue / half red ($\pm$).
- Two nodes of the same color are connected with probability $2d\theta/n + d(1 - \theta)/n$.
- Two nodes with different colors are connected with probability $d(1 - \theta)/n$.
- **Note**: average degree is $d$ and if $u \sim v$ then $E[X_uX_v] = \theta$.
- **Inference**: which nodes are likely red/blue?
- **Conjecture** (Decelle, Krzakala, Moore and Zdeborova, 11): "Belief-Propagation" is the optimal algorithm.
- and ... possible to do better than random iff $d\theta^2 > 1$. 

Elchanan Mossel  Simplicity & Complexity of BP
The Block Model in pictures

A sample from the model
The easier direction ...

Theorem (Mossel-Neeman-Sly-12)

If \( d^{\theta} \leq 1 \) then impossible to infer better than random.

Proof uses EKPS non-reconstruction result which holds for random trees of average degree.
The Conjecture is Correct

Theorem (M-Neeman-Sly, Massoulie 14)

If \( d\theta^2 > 1 \) then possible to detect (infer better than random).
BP and a New Type of Random Matrix

- **Thm** If $d\theta^2 > 1$ then possible to detect.

- **Conj:** (Krzakala, Moore, M, Neeman, Sly, Zdebrovoa, Zhang 13): If $A$ is the adjacency matrix, then w.h.p the second eigenvector of

  $$N = \begin{pmatrix} 0 & D - I \\ -I & A \end{pmatrix}, \quad D = \text{diag}(d_{v_1}, \ldots, d_{v_n}),$$

  is correlated with the partition and the second eigenvalue is $d(1 - 2\varepsilon) + o_n(1)$.

- No orthogonal structure! $N$ is not symmetric or normal. Singular vector of $N$ are useless.

- KMMNSZZ derived $N$ by **Linearizing** Belief Propagation and applying a number-theory identity by Hashimoto (89).

- Note: conjectured linear algebra algorithm is deterministic.

- Conjecture established by Bordenave-Lelarge-Massoulie 15.
The Eigenvalues of $N$

\[ d = 3, \quad d(1 - 2\varepsilon) = 2, \quad \sqrt{d} = 1.732... \]
The spectrum on real networks

Polblogs q=2
Overlap: 0.8533

Dolphins q=2
Overlap: 0.7419

Football q=10

Adjnoun q=2
Overlap: 0.6250

Karate q=2
Overlap: 1

Polbooks q=2
Overlap: 0.7571
Part 2: NON-LINEAR THEORY

Large $q$
Claim: For all $q$ if $d\theta^2 > 1$ then:
- For the tree broadcast model, can distinguish.
- Can detect the in the block model.
- Recover phylogenies from sequences of length $O(\log n)$.

Pf (for $q$ even): Divide $q$ colors to two sets of size $q/2$. Call one $+$ and the other $-$.  

More generally, this is true for broadcast process with Markov chains $M$ on edges where

$$\theta = \max(|s| : s \in \sigma(A) \setminus \{1\})$$

Pfs:
- For tree broadcast models: Kesten-Stigum 66.
- For phylogeny, M-Roch-Sly-15.
Thm: For large $q$, $\exists \theta_q$ with $d\theta_q^2 < 1$ and such that for $\theta > \theta_q$:

- For the tree broadcast model, can distinguish (M-01, Sly-09 ...)
- But not using linear or robust estimators (M-Peres-03, Janson-M-04)
- Can detect the in the block model.
- But believed to have computational/statistical gap (Abbe-Sandon-15, Banks-Moore-Neeman-Netrapalli-16)
- Recover phylogenies from sequences of length $O(\log n)$.
- Not written (Conjecture: cannot be done robustly).
Linear reconstruction for large $q$

Theorem (Count Reconstruction, Robust Reconstruction (Mossel-Peres, Janson-Peres))

For all $q$ and $d$-ary tree, $d\theta^2 = 1$ is the threshold for:

- **Count reconstruction**: inference of root better than random, based only on the census of $c_h \in \mathbb{Z}^q$.

  $$c_h(a) = \left| \{ v \in L_h : X_v = a \} \right|, \quad \text{Var}[\mathbb{E}[X_0|c_h]] \to 0 \iff d\theta^2 \leq 1$$

- **Robust Reconstruction**: inference given noisy versions of the leaves ($Y_v : v \in L_h$), where $Y_v = X_v$ with probability $\eta$ and $Y_v \sim U[q]$ with probability $1 - \eta$ for some fixed $\eta > 0$.

  $$\text{Var}[\mathbb{E}[X_0|Y_{L_h}]] \to 0 \iff d\theta^2 \leq 1$$
A Double phase transition for large $q$

**Theorem (Count Reconstruction, Robust Reconstruction (Mossel-Peres, Janson-Peres))**

For all $q$ and $d$-ary tree, $d\theta^2 = 1$ is the threshold for: census and robust reconstruction.

**Theorem (Reconstruction for large $q$ (Mossel 00))**

If $d\theta > 1$ then for $q > q_\theta$ can distinguish the root better than random:

$$\lim_{h \to \infty} Var[\mathbb{E}[X_0 | X_{L_h}]] > 0$$

$\implies$ Non-linear estimators are superior.

**Pf:** Shows fractal nature of information.