MCMC Learning

Varun Kanade
UC Berkeley

Elchanan Mossel
UC Berkeley

August 30, 2013

Outline

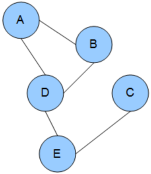
Uniform Distribution Learning

- Unknown target function $f : \{-1, 1\}^n \to \{-1, 1\}$ from some class C
- Uniform distribution over $\{-1, 1\}^n$
  - Random Examples: Monotone Decision Trees [OS06]
  - Random Walk: DNF expressions [BMOS03]
  - Membership Query: DNF, TOP [J95]
- Main Tool: Discrete Fourier Analysis

$$f(\mathbf{x}) = \sum_{S \subseteq [n]} \hat{f}(S) \chi_S(\mathbf{x}); \quad \chi_S(\mathbf{x}) = \prod_{i \in S} x_i$$

- Can utilize sophisticated results: hypercontractivity, invariance, etc.
- Connections to cryptography, hardness, de-randomization etc.
- Unfortunately, too much of an idealization. In practice, variables are correlated.

Markov Random Fields

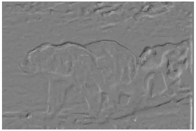


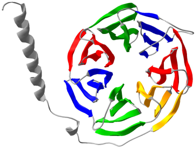
- Graph $G = ([n], E)$. Each node takes some value in finite set A.
- Distribution over A^n: (for ϕ_C non-negative, Z normalization constant)

$$\Pr((\sigma_v)_{v\in[n]}) = \frac{1}{Z} \prod_{\text{clique } C} \phi_C((\sigma_v)_{v\in C})$$

Markov Random Fields

- MRFs widely used in vision, computational biology, biostatistics etc.
- Extensive Algorithmic Theory for sampling from MRFs, recovering parameters and structures
- Learning Question: Given $f : A^n \to \{-1, 1\}$. (How) Can we learn with respect to MRF distribution?
  - Can we utilize the structure of the MRF to aid in learning?





Learning Model

- Let M be a MRF with distribution π and $f : A^n \to \{-1, 1\}$ the target function
- Learning algorithm gets i.i.d. examples $(\mathbf{x}, f(\mathbf{x}))$ where $\mathbf{x} \sim \pi$
- Learning algorithm "knows" MRF

Gibbs Sampling (MCMC Algorithm)

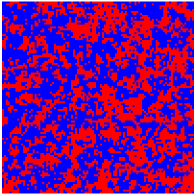
- **Sampling Algorithm**
  Starting from $\mathbf{x}^{(0)} = (x_1^{(0)}, \ldots, x_n^{(0)}) \in A^n$
  1. Pick $i \in [n]$ uniformly at random
  2. Pick $x_i^{(t+1)} \sim p(x_i \mid x_1^{(t)}, \ldots, x_{i-1}^{(t)}, x_{i+1}^{(t)}, \ldots, x_n^{(t)})$
  3. Set $x_j^{(t+1)} = x_j^{(t)}$ for $j \neq i$.
- Stationary distribution is MRF distribution
- For constant degree MRF graphs, conditional distribution has constant number of parameters
- We are interested in cases when Gibbs MC is rapidly mixing

Ising Model

- Let $G = ([n], E)$ be some degree-Δ graph
- For each $(i, j) \in E$, β_{ij} (bounded) interaction energy
- Configuration $\sigma \in \{-1, 1\}^n$; Hamiltonian

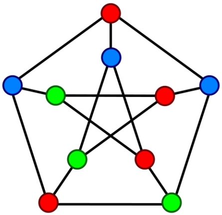
$$H(\sigma) = -\sum_{(i,j)\in E} \beta_{ij}\sigma_i\sigma_j - B\sum_{i\in[n]} \sigma_i$$

- Probability distribution: $p(\sigma) \propto \exp(-H(\sigma))$
- If $0 \leq \beta_{ij} \leq \beta(\Delta)$, Gibbs MC is rapidly mixing



Graph Colouring

- $G = ([n], E)$ be some degree-Δ graph
- For $q \geq 3\Delta$, a q-colouring is $C : [n] \to [q]$
- Probability distribution: uniform over valid colourings
- Gibbs MC is rapidly mixing



Harmonic Analysis Using Eigenvectors

- Let $\Omega = A^n$ be the statespace (MRF graph $G = ([n], E)$)
- Gibbs Markov Chain over Ω is reversible
  - Let P be the transition matrix and π the stationary distribution
  - Reversibility: $\pi_i P_{ij} = \pi_j P_{ji}$
- An eigenvector of P is a function $\nu : \Omega \to \mathbb{R}$
- Set of all eigenvectors forms orthonormal basis w.r.t. stationary distribution π
- Can we perform "Fourier" analysis using this basis?

Harmonic Analysis Using Eigenvectors

- The approach seems naïve:
  - Each eigenvector is of size $|A|^n$
  - How do we find these eigenvectors?
  - How do we find the expansion of an arbitrary function using eigenvectors?

Harmonic Analysis Using Eigenvectors

- We want to extract eigenvectors using power-iteration method
- Let $g : \Omega \to \{-1, 1\}$ (may be $\mathbb{R}$) be some function:

$$g = \alpha_1 \nu_1 + \alpha_2 \nu_2 + \cdots + \alpha_k \nu_k + \cdots .$$

- ν_i is eivenvector with eigenvalue λ_i and $\lambda_1 > \lambda_2 > \cdots$
- Then, (suppose g satisfies all the nice properties that we want)

$$\begin{aligned} P^t g &= \alpha_1 \lambda_1^t \nu_1 + \alpha_2 \lambda_2^t \nu_2 + \cdots \\ \mathbb{1}_{\mathbf{x}}^{\dagger} P^t g &= \alpha_1 \lambda_1^t \nu_1(\mathbf{x}) + \alpha_2 \lambda_2^t \nu_2(\mathbf{x}) + \cdots \\ \alpha_1^{-1} \lambda_1^{-t} \mathbb{1}_{\mathbf{x}}^{\dagger} P^t g &= \nu_1(\mathbf{x}) + \alpha_1^{-1} \alpha_2^{-1} (\lambda_1^{-1} \lambda_2)^t \nu_2(\mathbf{x}) + \cdots \end{aligned}$$

Harmonic Analysis Using Eigenvectors

So, we have:

$$\mathbb{1}_{\mathbf{x}}^{\dagger} P^t g = \alpha_1 \lambda_1^t \nu_1(\mathbf{x}) + \alpha_2 \lambda_2^t \nu_2(\mathbf{x}) + \cdots$$

$\mathbb{1}_{\mathbf{x}}^{\dagger} P^t$ is the distribution obtained by running Gibbs MC for t steps starting from $\mathbf{x}$

$$\mathbb{1}_{\mathbf{x}}^{\dagger} P^t g = \mathbb{E}_{\mathbf{x}' \sim \mathbb{1}_{\mathbf{x}}^{\dagger} P^t}[g(\mathbf{x}')]$$

LHS can be estimated by sampling from Gibbs MC

Summarizing ...

Given compact representation of function $g : \Omega \to \{-1, 1\}$ and access to Gibbs MC of MRF

- For any $\mathbf{x} \in \Omega$, we can output $\nu(\mathbf{x})$ (approximately), where ν is largest eigenvector in g
- By subtracting off previously found eigenvectors can extract top (constant number of) eigenvectors of g
  - need technical conditions that eigenvectors need to satisfy
  - errors add up due to sampling (cannot extract more than constant number)

Useful auxilliary functions

- Let $S \subseteq [n]$ and $b : S \to A$ be some assignment to variables in S. Then, define

$$g_{S,b}(\mathbf{x}) = \prod_{i \in S} (\mathbb{1}(x_i = b(i)) - \Pr(x_i = b(i)))$$

Learning Algorithm

- Let $\mathcal{V} = \{\nu_1, \dots, \nu_m\}$ be set of extracted eigenfunctions
- Let $\langle \mathbf{x}^i, f(\mathbf{x}^i) \rangle_{i=1}^s$ be a sample from π
- Set $\hat{\alpha}_j = (1/s) \sum f(\mathbf{x}^i) \nu_j(\mathbf{x}^i)$
- Output: $h(\mathbf{x}) = \sum_{j=1}^m \hat{\alpha}_j \nu_j(\mathbf{x}^i)$

- "Low-degree Algorithm"
- Part of spectrum used is that with high eigenvalues
  - Easier to access
  - More likely to capture "signal" rather than "noise"

Main Result

Theorem (Informal)

Let M be a markov random field with statespace A^n and suppose that the corresponding Gibbs MC is rapidly mixing. Suppose that $\mathcal{G}$ is a class of functions satisfying certain technical conditions (boundedness, low "L1" mass, appropriate gaps in eigenvalues). Then,

- *It is possible to extract a constant number of eigenvectors of P, the transition matrix of Gibbs MC, for every $g \in \mathcal{G}$. Let $\mathcal{V}$ denote the set of all eigenvectors obtained in this way.*
- *If $\mathcal{F}$ is a class of that are well-approximated using eigenvectors in $\mathcal{V}$, then the class $\mathcal{F}$ is learnable using the algorithm described on previous slide.*

- The natural MRF corresponding to the uniform distribution satisfies the conditions
- Thus, the "low-degree" algorithm could be obtained in this manner

Some Experiments

- For each $\mathbf{x} \in \Omega$: feature set $\Phi(\mathbf{x}) = (\nu_1(\mathbf{x}), \nu_2(\mathbf{x}), \ldots, \nu_m(\mathbf{x}))$
- Can consider higher order features (degree d "eigenfeatures"):

$$\Phi(\mathbf{x}) = \left(\prod_{i \in S} \nu_i(\mathbf{x}) \right)_{S \subseteq [m], |S| \leq d}$$

- Degree 2 regression performs much better in very basic experiments
- Products of eigenfunctions are often "close" to eigenfunctions

Open Questions

- For a simple model (MRF) with a non-product distribution and for a simple class of functions $\mathcal{F}$, is it possible to show that $\mathcal{F}$ is well-approximated by higher eigenvectors?
- The auxilliary function g we used, depended on a small number of variables. Thus, the highest eigenvectors in g are likely to be localized? This may be why (many of the) products of eigenvectors are close to eigenvectors. Can we understand these connections better?

Open Questions

- Can access to a labelled random walk from Gibbs MC help?

$$(\mathbf{x}^0, f(\mathbf{x}^0)), (\mathbf{x}^1, f(\mathbf{x}^1)), \cdots,$$

- Under some conditions on the MRF can learn k-juntas by a very simple algorithm
- Is rapid mixing of Gibbs MC enough for learning k-juntas?