Graphical Models

Powerful model for describing high-dimensional distributions...

... via their conditional independence structure

def: The Ising model is a distribution on $\{\pm 1\}^n$ described as

$$\mathbb{P}[\sigma] = \frac{e^{-\beta H(\sigma)}}{Z_\beta}$$

where ① $H(\sigma) = -\sum_{ij} J_{ij}\sigma_i\sigma_j - \sum_j h_j \sigma_j$

is the Hamiltonian

② $Z_\beta = \sum_\sigma e^{-\beta H(\sigma)}$

is the partition function

③ and β is the inverse temperature

Properties

The case where $J_{ij} \geq 0$ is called ferromagnetic

i.e. the σ_i's want to point in the same dir.

Similarly $J_{ij} \leq 0$ is called antiferromagnetic one of the original motivations was to give a microscopic explanation of spontaneous magnetization:

At what temperature is the average magnetization $\frac{\sum \sigma_i}{n}$ typically far from zero?

Claim: As $\beta \to \infty$ the distribution is concentrated on the ground states i.e minimizers of $H(\sigma)$

Key property: Let $G = (\overset{V=[n]}{[n]}, E)$ where

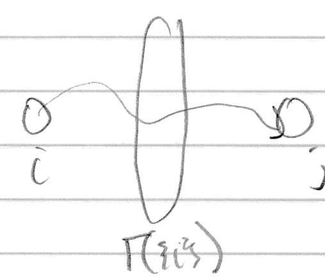
$$E = \{(i,j) \mid J_{ij} \neq 0\}$$

Proposition: Let $T = \underbrace{\Gamma(\{i\})}_{\text{neighbors}} \cup \{i\}$. Then

$$\sigma_i \perp\!\!\!\perp \sigma_{V \setminus T} \mid \sigma_{\Gamma(\{i\})}$$

"once you know the spins of the neighbors of i, its spin is independent of everything else"

Thus independence relations follow from graph separation



all paths are blocked

<u>Proof</u>: If we fix $\sigma_{T(\{i\})}$ the new Hamiltonian is

$$-[\sigma_i, \underbrace{\cdots\cdots}_{\sigma_j\text{'s for } j \in V\setminus T}] \left[\begin{array}{c|c} J_{11} & 0 \\ \hline 0 & \tilde{J} \end{array}\right] \begin{bmatrix} \sigma_i \\ \vdots \end{bmatrix} - [\sigma_i, \cdots] \begin{bmatrix} h_i \\ \vdots \\ \end{bmatrix} + C$$

$$= H_i(\sigma_i) + H_{V\setminus T}(\sigma_{V\setminus T})$$

Hence the distribution factorizes $\boxed{}$

<u>Takeaway</u>: If you know the independence structure, can often write the distribution more concisely than specifying $2^n - 1$ values

<u>Algorithmic Questions</u>

1. Given an Ising model, can you sample from it?

2. Can you compute the posterior, given a partial assignment?

   i.e. inference

3. Can you learn it from samples?

In many situations ① $\approx$ ②

Hardcore model

Given a bounded degree graph $G=(V,E)$
$\Delta = $ max degree

define

$$\mathbb{P}[I] = \frac{\lambda^{|I|}}{Z_\lambda}$$

independent sets

Thm [Jerrum, Sinclair] [Weitz] If $\Delta \geq 3$ then for any

$$\lambda < \lambda_c(\Delta) \triangleq \frac{(\Delta-1)^{\Delta-1}}{(\Delta-2)^{\Delta}}$$

there is an efficient algorithm for sampling

Thm [Sly, Sun] If $\Delta \geq 3$ then for any

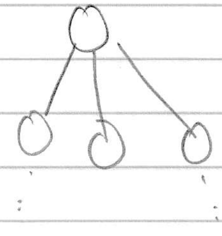
$$\lambda > \lambda_c(\Delta)$$

It is NP-hard to approximately sample

What is happening at $\lambda_c(\Delta)$?

Tree Uniqueness

Consider a $(\Delta-1)$-ary tree



<u>weak spatial mixing</u>: For what values of λ does the marginal at the root become almost independent of the leaves?

$$\lim_{\text{height} \to \infty} \left| \mathbb{P}[\underset{\substack{\uparrow \\ \text{root}}}{\rho} \in I \mid \underset{\substack{\uparrow \\ \text{boundary condition on leaves}}}{\tau}] - \mathbb{P}[\rho \in I \mid \sigma] \right| = 0$$

Need this to be true for the distribution on an infinite tree to be well defined

<u>Proposition</u>: WSM holds iff $\lambda < \lambda_c(\Delta)$

Main ingredients

1. marginal at root can be computed recursively
2. Can show it's contractive if $\lambda < \lambda_c(\Delta)$

<u>Takeaway</u>: Sampling exhibits a computational phase transition

---

In 2015, a big surprise

<u>Thm [Bresler]</u> There is a polynomial time algorithm for learning any bounded

degree Ising model

<u>Note</u>: The constants depend on upper/lower bounds on nonzero $|J_{ij}|$'s an upper bounds on $|h_i|$'s, which is necessary

<u>Takeaway</u>: You can learn even when you can't sample from it

---

Warm-Up: Hardcore Model

<u>def</u>: A distribution D on S^n is δ-unbiased
(S: alphabet)

if for any i, and assignment $x \in S^{n-1}$ we have

$$\mathbb{P}[X_i = \alpha \mid X_{-i} = x] \geq \delta \quad \forall \alpha \in S$$

i.e. all variables have some randomness

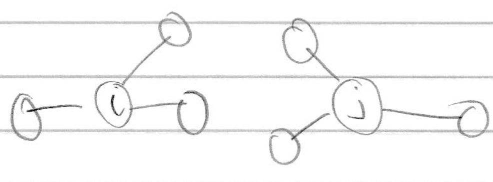
closed under taking marginals

<u>Proposition</u>: If the hardcore model is δ-unbiased, can learn its edges

<u>Proof</u>: (sketch) Consider any pair i, j

① If ever $X_i = X_j = 1 \Rightarrow (i,j) \notin E$

② Conversely if $(i,j) \notin E$ consider



Fix a partial assignment where none of the neighbors of i or j are in I. Then

$$\mathbb{P}[X_i = X_j = 1 \mid X_{-i,j} = x] \geq \delta^2 \qquad \boxtimes$$

Finally note we can prove a lower bound on δ in terms of Δ and an upperbound on λ

Now consider general Ising models, define

$$\text{width } \lambda = \max_i \left(\sum_j |A_{ij}| + |h_i| \right)$$

and let $\eta = \min_{ij \text{ s.t. } A_{ij} \neq 0} |A_{ij}|$

We'll follow [Wu, Sanghavi, Dimakis]

<u>def</u>: The logistic function

$$\sigma(z) = \frac{1}{1+e^{-z}}$$

<u>Key Fact</u>: The conditional distribution in an Ising model is logistic, i.e.

$$\mathbb{P}[X_i = 1 \mid X_{-i} = x] = \sigma(\langle w, x' \rangle)$$

↑ padded

$[x, 1]$

Moreover $\|w\|_1 \leq 2\lambda$

<u>Proof:</u> By definition

$$\mathbb{P}[X_i=1 \mid X_{-i}=x] =$$

$$\frac{\exp\left(\sum_{j\neq i} A_{ij}x_j + h_i\right)}{\exp\left(\sum_{j\neq i} A_{ij}x_j + h_i\right) + \exp\left(-\sum_{j\neq i} A_{ij}x_j - h_i\right)}$$

Thus we find

$$w = 2[A_{i1}, A_{i2}, \dots A_{i,(i-1)}, A_{i(i+1)}, \dots A_{in}, h_i]$$

and so $\|w\|_1 \leq 2\lambda$. $\boxed{\text{E}}$

Thus we can try to learn the parameters by logistic regression

$$\hat{w} = \underset{\hat{w}}{\operatorname{argmin}} \underbrace{\frac{1}{N}}_{\#\text{samples}} \sum_{s=1}^{N} \ell(\underbrace{y_s}_{\text{label of } i} \langle \hat{w}, x_s' \rangle) \quad (*)$$

where $\ell \equiv$ negative log-likelihood

$$= \ln\left(1 + e^{-y_s\langle \hat{w}, x_s' \rangle}\right) = \begin{cases} -\ln \sigma(\langle \hat{w}, x_s' \rangle) & \text{if } y_s = 1 \\ -\ln(1-\sigma(\langle \hat{w}, x_s' \rangle)) & \text{else} \end{cases}$$

<u>Claim</u> $(*)$ is a convex optimization problem

Note: This would work, but would require

$m \gtrsim n$ samples

and we can do much better, i.e. $m \gtrsim C_{\lambda,\Delta} \log n$

New approach, use ℓ_1-regularization

$$\hat{w} = \underset{\hat{w}}{\operatorname{argmin}} \frac{1}{N} \sum_{s=1}^{N} \ell(y_s \langle \hat{w}, x'_s \rangle)$$

$$\text{s.t. } \|\hat{w}\|_1 \leq 2\lambda$$

Main ingredients

(1) Generalization bounds for ℓ_1:

Lemma Suppose ℓ has Lipschitz constant L and $\mathcal{D}$ is a distribution on $X \times Y$ with $X = \{x \mid \|x\|_\infty \leq R\}$. Then for any $w \in W$

n-dimensional

$$L(w) \leq \hat{L}(w) + 2LRD\sqrt{\frac{2\ln(2n)}{N}}$$

true loss

empirical loss

$$+ LRD\sqrt{\frac{2\ln(2/\delta)}{N}}$$

failure probability

and $W = \{w \mid \|w\|_1 \leq W\}$

Intuition: Instead of needing an ε-net over the unit ball in $\mathbb{R}^n$, just need a ε-net over the W-dilation of the ℓ_1 ball, which has size $\sim \binom{n}{\hat{W}}$

(2) Strong convexity: For any $\hat{w}$, want to show

$$L(\hat{w}) - L(w) \geq \mathbb{E}_{x'}\left[\left(\sigma(\langle \hat{w}, x\rangle) - \sigma(\langle w, x\rangle)\right)^2\right]$$

Intuition: Consider two Bernoulli distributions with parameters a, b respectively. Then

$$d_{TV}(\mathrm{Ber}(a), \mathrm{Ber}(b)) = |a-b|$$

and the KL-divergence btwn general p, q is

$$d_{KL}(p \| q) = \sum_x p(x) \ln \frac{p(x)}{q(x)}$$

which turns into

$$d_{KL}(\mathrm{Ber}(a) \| \mathrm{Ber}(b)) = a \ln\left(\frac{a}{b}\right) + (1-a)\ln\left(\frac{1-a}{1-b}\right)$$

Fact [Pinsker] $d_{TV}^2 \leq \frac{1}{2} d_{KL}$

Finally since our loss function is the negative log-likelihood we have

$$L(\hat{w}) - L(w) = \mathbb{E}_{x'}\left[d_{KL}(\sigma(\langle \hat{w}, x\rangle) \| \sigma(\langle w, x\rangle))\right]$$

③ unbiasedness

<u>Proposition</u>: Suppose D is δ-unbiased. Then

$$\underset{X\sim D}{\mathbb{E}}\Big[\big(\underbrace{\sigma(\langle w,X\rangle+b)}_{\sigma}-\underbrace{\sigma(\langle w',X\rangle+b')}_{\sigma'}\big)^2\Big]\leq \varepsilon \quad (\square)$$

where $\varepsilon \lesssim \delta e^{-2\|w\|_1-2|b'|}$ then

$$\|w-w'\|_\infty \lesssim e^{\|w\|_1+|b'|}\sqrt{\frac{\varepsilon}{\delta}}$$

<u>Proof</u>: By subconditioning $(\square) \Rightarrow$

$$\varepsilon \geq \underset{X_{-i}}{\mathbb{E}}\Big[\underset{X_i|X_{-i}}{\mathbb{E}}\big[(\sigma-\sigma')^2\big]\Big]$$

$$= \underset{X_{-i}}{\mathbb{E}}\Big[\big(\sigma|_{X_i=+1}-\sigma'|_{X_i=+1}\big)^2\,\mathbb{P}[X_i=1|X_{-i}]$$

$$+\big(\sigma|_{X_i=-1}-\sigma'|_{X_i=-1}\big)^2\,\mathbb{P}[X_i=-1|X_{-i}]\Big]$$

By δ-unbiasedness

$$\geq \delta \underset{X_{-i}}{\mathbb{E}}\Big[\big(\sigma|_{X_i=+1}-\sigma'|_{X_i=-1}\big)^2+\big(\sigma|_{X_i=-1}-\sigma'|_{X_i=-1}\big)^2\Big]$$

Now by the Fact we have

Fact: [Klivans, Meka]

$$|\sigma(a)-\sigma(b)| \geq e^{-|a|-3} \min(1, |a-b|)$$

$$\gtrsim \delta e^{-2\|w\|_1 - 2|h'|} \underset{X_{-i}}{\mathbb{E}}\Big[\min\big(1, (\underbrace{\langle w, X^{i,+}\rangle + h - \langle w', X^{i,+}\rangle - h'}_{a})^2\big)$$

$X^{i,+}$: X_i set to $+1$

$$+ \underbrace{\ldots}_{b}\Big]$$

b: same, but X_i set to -1

$$\geq \delta e^{-2\|w\|_1 - 2|h'|} \underset{X_{-i}}{\mathbb{E}}\left[\min(1, 2(w_i - w'_i)^2)\right]$$

which follows b/c

$$\min(1, a^2) + \min(1, b^2) \geq \min\left(1, \frac{(a-b)^2}{2}\right)$$

This bound holds for all i, and rearranging algebraically completes the proof. ☒

<u>Remark</u>: Similar algorithms work for Markov random fields

$$\mathbb{P}[X = x] = \frac{e^{\beta p(x)}}{Z_\beta}$$

$p(x)$: degree $\leq d$ polynomial

There are $C_{\lambda,\Delta}\, n^{O(d)}$ time algorithms

Moreover there are $n^{\Omega(d)}$ lower bounds based on sparse parity with noise

Sufficient Statistics

Suppose we are given samples

$$X_1, \dots, X_N \sim P_\theta(x)$$

Is there a <u>sufficient statistic</u>

— i.e. we can compress to $T(X_1, \dots, X_N)$ w/o losing any information?

Factorization Theorem [Neyman, ...]
A statistic is sufficient iff

$$P_\theta(X_1, \dots, X_N) = u(X_1, \dots X_N)\, v(T(X_1, \dots X_N), \theta)$$

Graphically this means

$$\boxed{\theta} \longrightarrow \boxed{T(X_1, \dots X_N) = t} \longrightarrow \boxed{X_1 \dots X_N}$$

i.e. $X_1 \dots X_N \perp\!\!\!\perp \theta \mid T(X_1 \dots X_N) = t$

There is a canonical way to satisfy this condition

<u>def:</u> An exponential family has the form

$$P_\theta(x) = \frac{h(x)\, e^{\langle \theta, T(x) \rangle}}{Z(\theta)}$$

e.g. for the Ising model, we can take

$$T(x) = [\mathrm{vec}(xx^T), x]$$

$$\theta = [\mathrm{vec}(A), h]$$

<u>Corollary</u>: For an exponential family, any estimator

$$\hat{\theta}(X_1 \ldots X_N)$$

can be turned into another one $\tilde{\hat{\theta}}(T(X_1 \ldots X_N) = t)$

<u>Problem</u>: Looking inside

$$\boxed{\tilde{\theta}: \quad \boxed{T(X_1 \ldots X_N) = t} \rightarrow X_1 \ldots X_N \rightarrow \hat{\theta}}$$

But sampling can be hard

<u>Thm</u> [Montanari] [Bresler et al] There are graphical models that can be efficiently learned, but not if you reduce to sufficient statistics

<u>Open</u>: Are there computational-vs-statistical tradeoffs for learning exponential families?