11/4

Lecture 17: Intro to Graphical Models, VI

Def: (Undirected) graphical model w/ pairwise interactions:

Let $\{\psi_{ij}\}_{(i,j)\in F}$ be compatibility functions $\{\pm 1\}^2 \to \mathbb{R}_{\geq 0}$ that dictate interactions b/t particles

The Gibbs measure is dist. over $\{\pm 1\}^n$ given by

$$\underbrace{\mu(x)}_{\text{"spins"}} \triangleq \frac{1}{Z} \prod_{(i,j)\in F} \psi_{ij}(x_i, x_j), \quad Z \triangleq \sum_{x\in\{\pm 1\}^n} \prod_{(i,j)} \psi_{ij}(x_i, x_j)$$

where Z is the partition function, i.e. normalizing constant.

We will use the shorthand

$$\mu \propto \prod_{(i,j)} \psi_{ij}(x_i, x_j)$$

Example ("Ising model"):

$$\psi_{ij}(x_i, x_j) = \exp(-\beta A_{ij} x_i x_j), \text{ so}$$

$$\mu(x) \propto \exp\Big(-\frac{\beta}{2} \underbrace{x^T A x}_{\text{"energy"}}\Big)$$

for $A \in \mathbb{R}^{n\times n}$ a symmetric matrix with zero diagonal

β: "inverse temperature"

A: "Hamiltonian"/"interaction matrix"

$$\mathcal{E}(x) \triangleq \frac{\beta}{2} x^T A x = -\lg\Big(\prod_{(i,j)\in F} \psi_{ij}(x_i, x_j)\Big) : \text{"energy"}$$

As $\beta \to 0$, $\mu \to \mathrm{Unif}(\{\pm 1\}^n)$

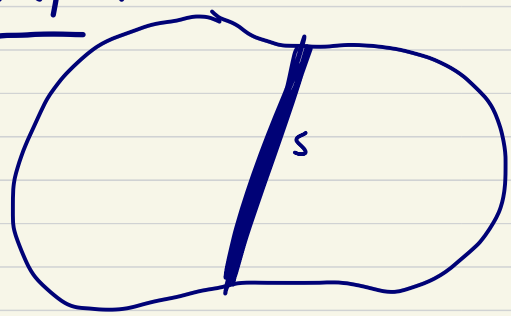
$\beta \to \infty$, $\mu \to \mathrm{Unif}(\{\text{energy minimizers}\})$

Can think of $-A$ as adjacency matrix of weighted graph.* Denote this by G.

$\partial_i \triangleq \{j \text{ s.t. } A_{ij} \neq 0\} = \{j \text{ s.t. } (i,j) \in F\}$,

i.e. the neighbors of node i in G

Markov property:



If $[n] \setminus S$ decomposes into disjoint pieces, then marginal distributions on the pieces are independent, conditioned on any assignment to the spins on S

e.g. if we condition on $x_{\partial i}$, then conditional dist. on x_i is independent of conditional dist on rest of the spins. For Ising model:

$$\Pr[x_i = \sigma \mid x_{\partial i} = s] \propto \exp\Big(\beta \sum_{j \in \partial_i} A_{ij} s_j \sigma\Big)$$

---

* negative sign is b/c of inconvenient culture clash b/t physics and CS: physicists want to minimize energy, in CS we want to maximize $x^T B x$, e.g. in MAXCUT

2 fundamental algorithmic tasks in inference:

1. computing the partition function Z
2. Sampling from Gibbs measure μ

Note alg. for ① $\Rightarrow$ alg. for ② and vice versa ("equivalence of counting + sampling")

Challenge: Z is sum of exponentially many terms, so in many cases we expect it is computationally hard to compute...

e.g. if $\psi_{ij}(x_i, x_j) = \mathbb{1}[x_i \neq x_j]$ for all $(i,j) \in E(G)$,

$Z = \#$ independent sets of G ("#P-complete", i.e. very hard)

So our goal will be to approximate Z/approximately sample from Gibbs measure μ

Some approaches:

This unit

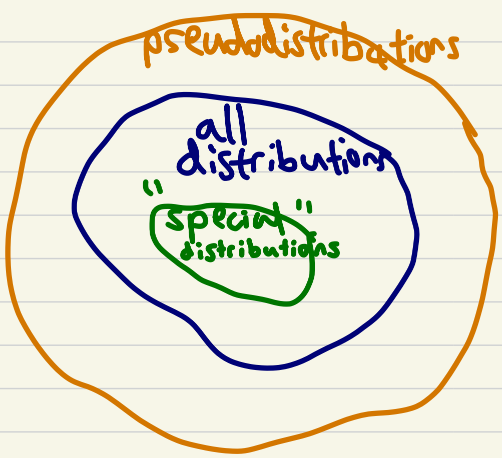
- Markov chain Monte Carlo (MCMC)
- ☆ Variational inference (VI)
- Diffusion models (very recent, more on this at the end of the course)

$\underline{\text{VI}}$: approximate μ by dist. from family $\mathcal{P}$ of "simpler" distributions that are easy to sample from (e.g. product distributions, aka naive mean-field):

$$\min_{\nu \in \mathcal{P}} KL(\nu \,\|\, \mu) \qquad (\star)$$

Note: - if $\mathcal{P}$ is <u>all</u> distributions, minimizer is $\nu^{\star} = \mu$ (Gibbs' inequality)

- "opposite" of SoS relaxation



An issue: can't even evaluate the objective function in $(\star)$, let alone optimize!

Fortunately, this particular issue is not really an issue:

$$KL(\nu \| \mu) = \mathbb{E}_\nu \ln \frac{\nu}{\mu}$$

$$= \mathbb{E}_\nu \ln \frac{\nu}{e^{-\mathcal{E}}/Z}$$

free energy of the Gibbs measure

$$= \underbrace{\mathbb{E}_\nu \ln \nu}_{\substack{\text{negative}\\ \text{entropy}\\ \underbrace{-H(\nu)}_{\text{if } \nu \text{ simple, easy to evaluate}}}} + \underbrace{\mathbb{E}_\nu \mathcal{E}}_{\substack{\text{average}\\ \underbrace{\text{energy}}_{\text{easy to approximate}}}} - \underbrace{\ln(1/Z)}_{\text{independent of } \nu!}$$

$$\boxed{G[\nu] \triangleq -H(\nu) + \mathbb{E}_\nu[\mathcal{E}]}$$

"Gibbs free energy functional" /

$-1 \times$ "evidence lower bound" (ELBO)

(note: convex in ν)

so to minimize $KL(\nu \| \mu)$, suffices to minimize $G[\nu]$ which is easy to evaluate

Interpretation of G as "regularized energy":
for Ising model, recall $\mathcal{E}(x) = \frac{\beta}{2} x^T A x$, so

$$G[\nu] = \frac{\beta}{2} x^T A x \underbrace{- H(\nu)}_{\text{"entropy regularization"}}$$

"hot"

when β small, minimizer prioritizes maximizing entropy

β big, minimizer prioritizes minimizing avg. energy

"cold"

$G[\nu]$ efficiently computable, but computationally intractable to optimize a priori...

Next lecture: powerful heuristic, belief propagation (BP), for solving $\min_{\nu} G[\nu]$.

2 interpretations of the heuristic:

1. dynamic programming

2. finding stationary points of a relaxation of the Gibbs free energy (Bethe free energy) (see supplemental notes)

BP as dynamic programming (see slides for Lecture 18)

Let's first shift focus to easier task than full-blown VI: marginal estimation dist. μ_i over each node is a Bernoulli random variable, goal is to estimate it

Physics motivation:

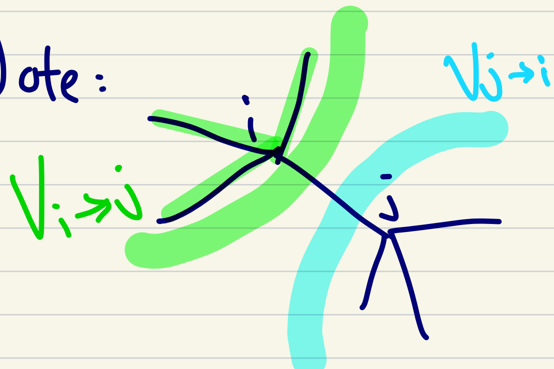
physicists care about limiting objects, and one important that they consider as $n \to \infty$ is the empirical dist. over marginals, i.e.

$$q_n(z) \triangleq \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}[z = \mu_i]$$

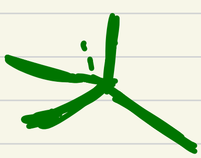
and, given a sequence of Gibbs measures $(\mu^{(n)})$, want to understand $\lim_{n\to\infty} q_n$

To motivate the algorithm, assume G is a tree

Note:



removing (i,j) from tree splits G into two subtrees



: $\bar{V}_{i\to j}$ $(V_{j\to i} + \text{edge }(i,j))$

: $\bar{V}_{j\to i}$ $(V_{i\to j} + \text{edge }(i,j))$

To sample from μ_i,

1). Sample spins on subtrees $V_{j\to i}$ for $j \in \partial i$, yields assignment $s \in \{\pm 1\}^{\partial i}$ to ∂i

2). Sample from conditional dist. on x_i, i.e.

$$\mathbb{P}[x_i = \sigma \mid x_{\partial i} = s] \propto \prod_{j\in\partial i} \psi_{ij}(\sigma, s_j)$$

By law of total probability,

$$
\begin{aligned}
\Pr[x_i=\sigma] &\propto \sum_{s\in\{\pm1\}^{\partial i}} \prod_{j\in\partial i} \Pr[x_j=s_j]\,\psi_{ij}(\sigma,s_j) \\
&= \prod_{j\in\partial i} \underbrace{\sum_{s_j\in\{\pm1\}} \Pr_{\mu_{V_{j\to i}}}[x_j=s_j]\,\psi_{ij}(\sigma,s_j)}_{(\star\star\star)}
\end{aligned}
\tag{$\star\star$}
$$

b/c marginal dist's on spins in $V_{j\to i}$ are independent across j's

$(\star\star\star)$ proportional to $\Pr_{\mu_{\bar{V}_{j\to i}}}[x_i=\sigma]$

(i.e. can express marginals of $\mu_{\bar{V}_{j\to i}}$ in terms of marginals of $\mu_{V_{j\to i}}$)

Unsatisfying b/c we've gone from $\Pr_{\mu}[x_i=\sigma]$ to $\Pr_{\mu_{\bar{V}_{j\to i}}}[x_i=\sigma]$, but we're very close.

Define messages:

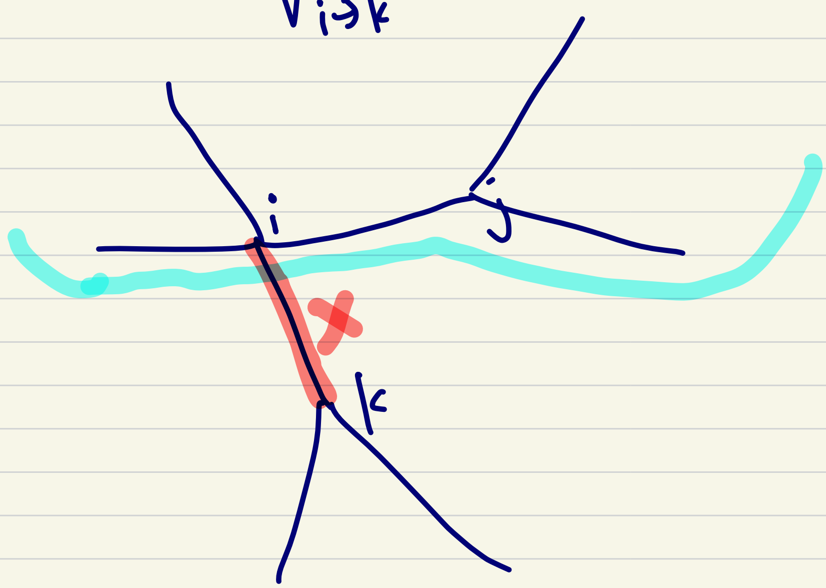
$$m_\sigma^{j \to i} \triangleq \Pr_{\mu_{V_{j\to i}}} [x_j = \sigma]$$

$$\bar{m}_\sigma^{j \to i} \triangleq \Pr_{\mu_{\bar{V}_{j\to i}}} [x_i = \sigma]$$

Then (☆☆☆) can be written as

$$\text{(I)} \quad \boxed{\bar{m}_\sigma^{j\to i} \propto \sum_{s \in \{\pm 1\}} m_s^{j \to i} \cdot \psi_{ij}(\sigma, s)}$$

Also note that (☆☆☆) can be modified to apply to $\mu_{V_{i\to k}}$ instead of μ, i.e.



previously, $(\star\star\star)$ gave

$$\mathbb{P}_{\mu}\left[x_i=\sigma\right] \propto \prod_{j\in\partial i} \bar{m}_\sigma^{j\to i}$$

after removing edge (i,k), we get

$$\mathrm{II} \qquad \boxed{m_\sigma^{i\to k} \propto \prod_{j\in\partial i\setminus k} \bar{m}_\sigma^{j\to i}}$$

$$\left(= \mathbb{P}_{\mu_V^{i\to k}}\left[x_i=\sigma\right]\right)$$

we can then write marginals succinctly in terms of the messages:

$$(\star\star\star\star) \qquad \mathbb{P}_{\mu}\left[x_i=\sigma\right] = m_\sigma^{i\to j}\, m_\sigma^{j\to i}$$

Combining I and II yields:

$$(\text{rec}) \qquad \boxed{m_\sigma^{i\to k} \propto \prod_{j\in\partial i\setminus k} \sum_{s\in\{\pm 1\}} m_s^{j\to i} \cdot \psi_{ij}(\sigma,s)}$$

BP on trees:

1). Pick arbitrary root vertex
2). For every leaf j and parent i, initialize $m_\sigma^{j\to i} = 1/2 \quad \forall \sigma \in \{\pm 1\}$
3). Use (rec) to compute $\bar{m}$'s via dynamic programming, starting from leaves
4) Use (I) to compute m's
5) Use (~~$\ast\ast\ast$~~) to compute marginals

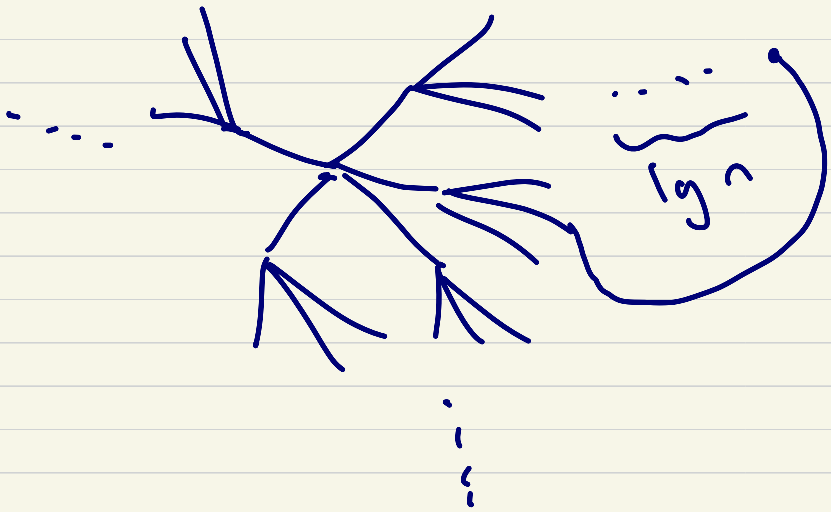
What if G is not a tree? Then sub-tree marginals $\{\mu_{\nu}^{j\to i}\}_{j\in\partial i}$ are not independent...

Nevertheless, can still run the above algorithm* and hope it gives something interesting!

* As stated, the algorithm is stated w/ a tree structure in mind. Without this, we can still apply update rules for $\bar{m}$ and m in parallel

over many rounds.

Intuition for why this is a good idea:
if the graph is a random sparse graph,
then locally it looks like a tree



If every edge appears w.p. $\frac{c}{n}$ for $c = O(1)$,
then probability that some "descendant" at depth d
"returns" to ancestor i is

$$1 - \left(1 - \frac{1}{n}\right)^{c^d}$$

so as long as $c^d \ll n$, this is $o(1)$.

Next lecture we will see a natural setting where
such a sparse random graph arises.

Even in such cases, BP is notoriously hard to analyze. We will instead see 2 rigorous alg's inspired by BP:

1). Spectral methods on nonbacktracking operators

2). approximate message passing.