CS 57800 – Statistical Machine Learning

Lecture 3 Probabilistic Inference and Statistical Physics

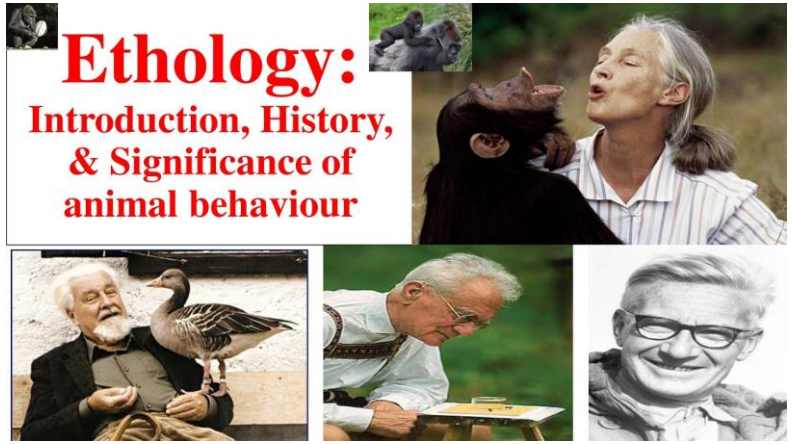
https://ruizhezhang.com/course_fall_2026.html

“Physics” of ML/AI

Math

- Fully rigorous theorems!
- Too idealistic assumptions
- Too simple models (e.g., 2-layer transformer)
- Hard to connect to practice
- Progress very slow

“Ethology”



“Physics” of ML/AI

Math

- Fully rigorous theorems!
- Too idealistic assumptions
- Too simple models (e.g., 1-layer transformer)
- Hard to connect to practice
- Progress very slow

“Ethology”

- Can study large models (GPT 6, Llama 4, …)
- Can be very educational (CoT, ToT, MoE, …)
- Less rigorous, too subjective
- Sometimes lead to false claims

“Physics” of ML/AI

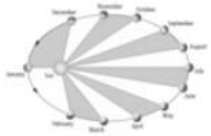
Math

“Ethology”

Newton's laws



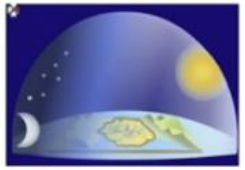
Kepler's laws of planetary motion



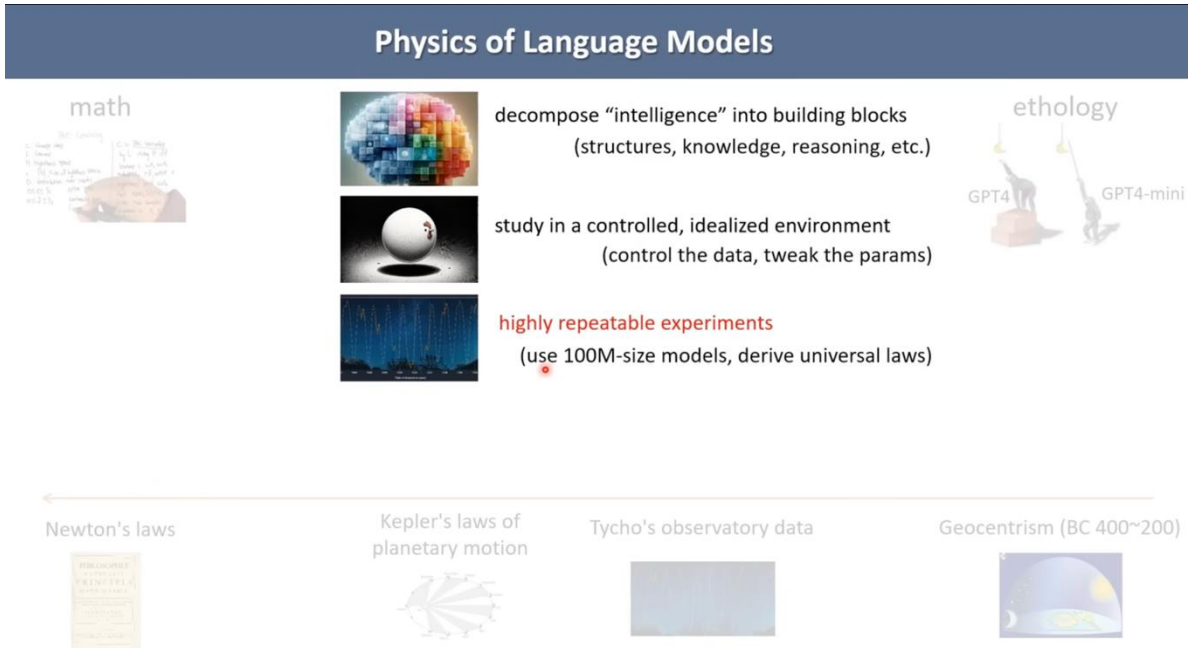
Tycho's observatory data



Geocentrism (BC 400~200)



“Physics” of ML/AI



Physics of Language Models Zeyuan Allen-Zhu

(https://physics.allen-zhu.com/)



ICML 2024
International Conference On Machine Learning

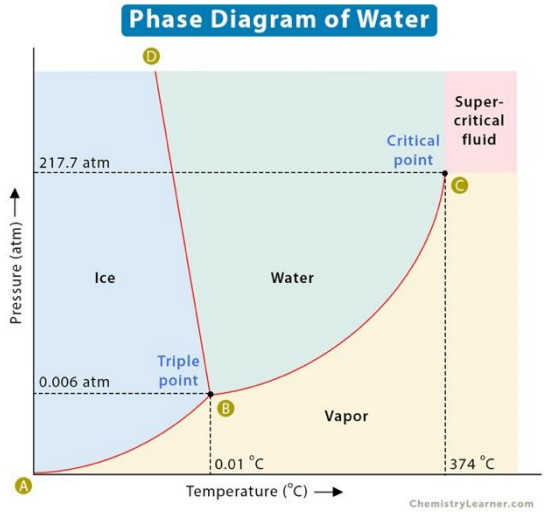
“Physics” of ML/AI

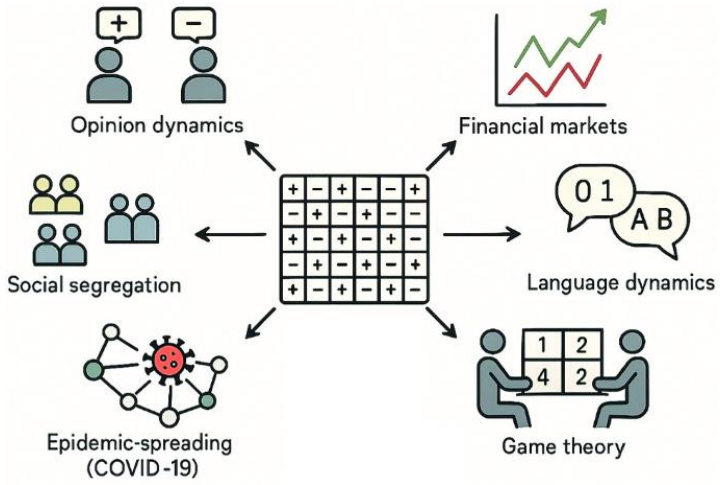
- Our world is super complicated
- Theoretical physicists are very good at working on toy models to understand the world

Example: the Ising model

$$H_N(\boldsymbol{S}) = -J\sum_{\{i,j\}} S_i S_j - h\sum_i S_i\,, \qquad \boldsymbol{S} \in \{-1,1\}^N$$

- Can demonstrate phase transitions (e.g., water freezing into ice, a metal losing its magnetism when heated, etc.)



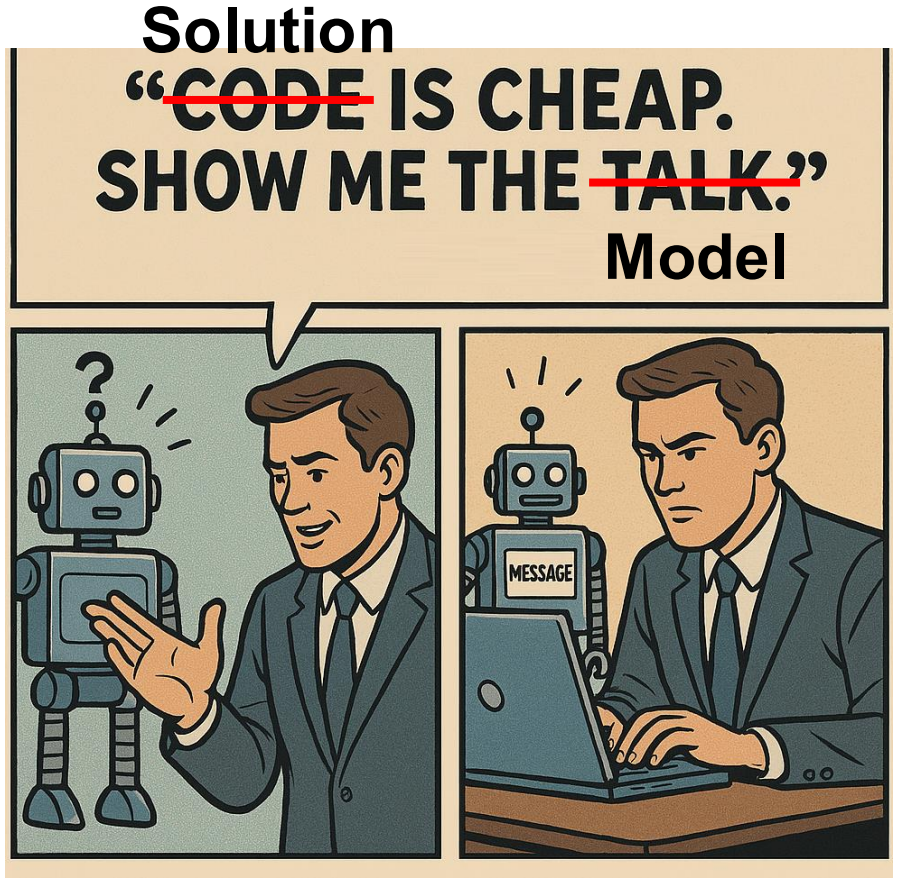


“Physics” of ML/AI

- Our world is super complicated
- Theoretical physicists are very good at working on toy models to understand the world

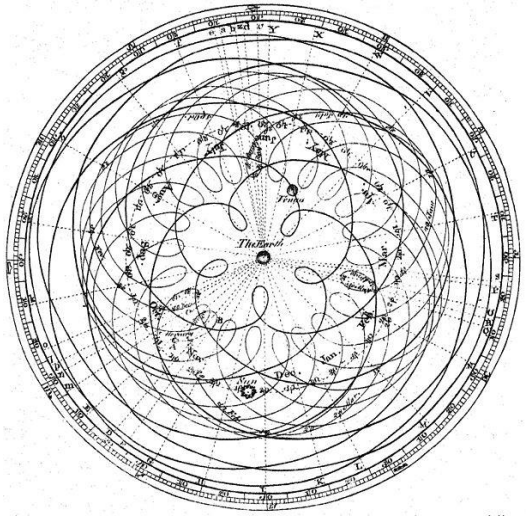
Personal thought:

- In the AGI-era,
- solving a model is much cheaper,
- identifying a useful model (≈problem) is more valuable



ML for Physics

- Physicists have done “machine learning” (inference) over centuries



The Ptolemaic epicycle model

(Overfitting)

Linear regression,
Newton’s method,
etc.

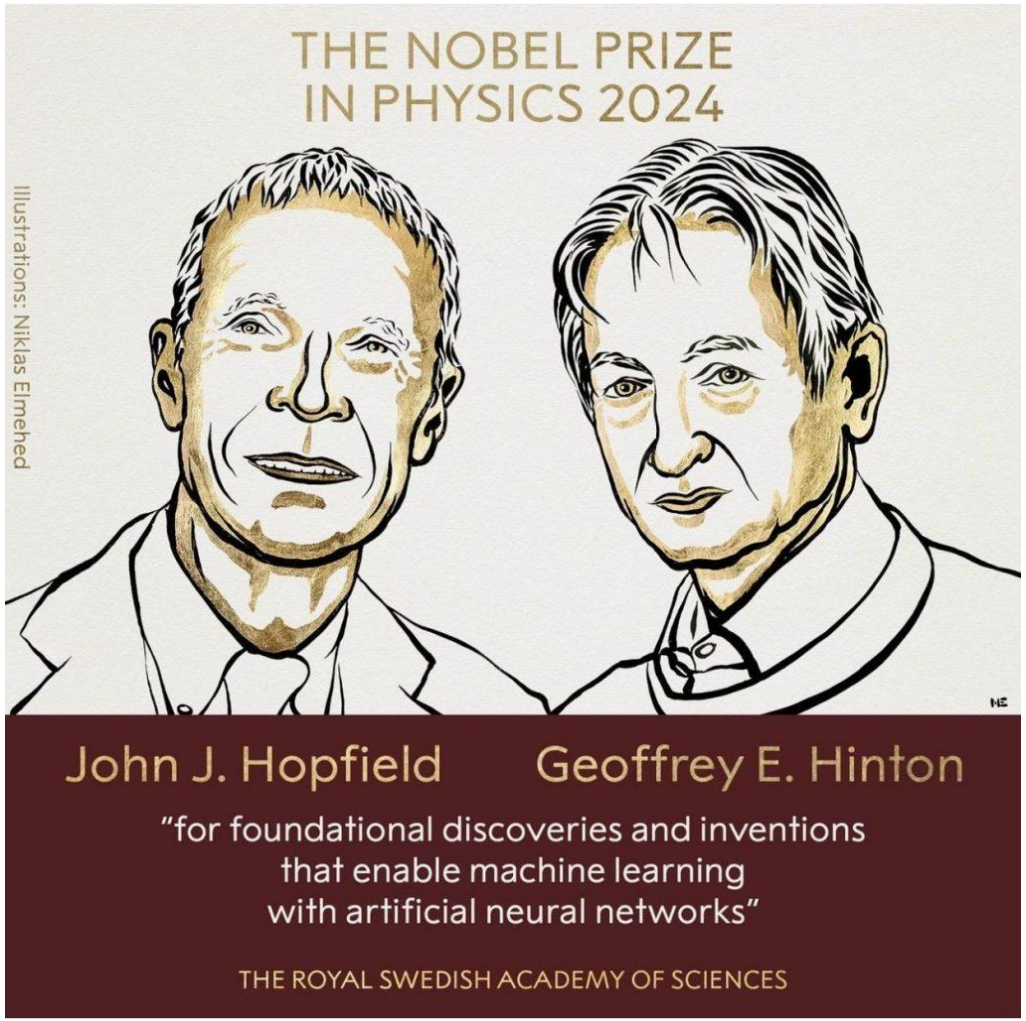
New data

$$F = G\frac{m_1 m_2}{r^2}$$

Newton's law of universal gravitation

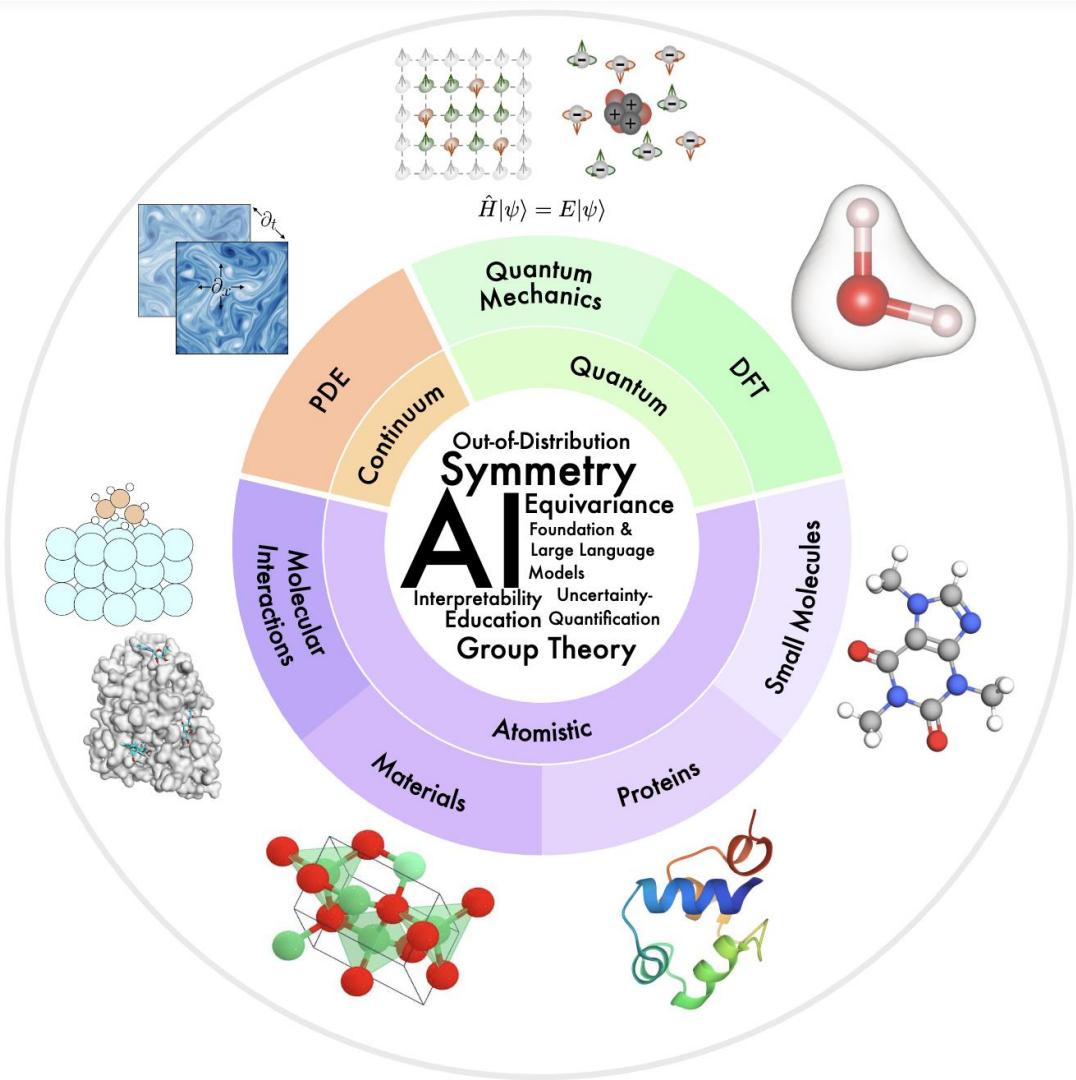
ML for Physics

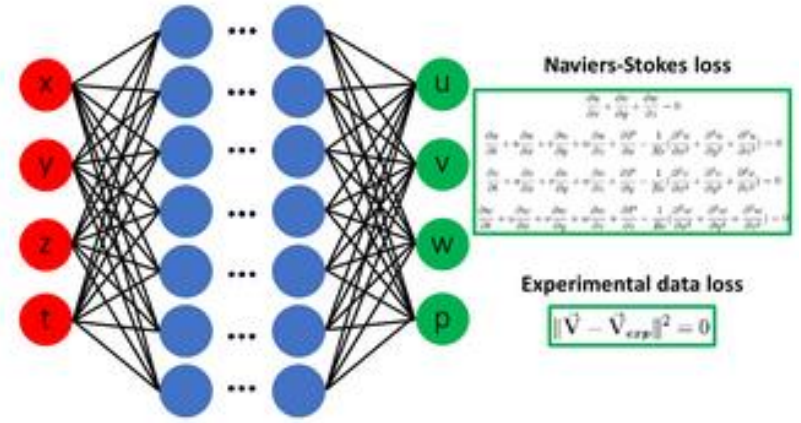
Hopfield network

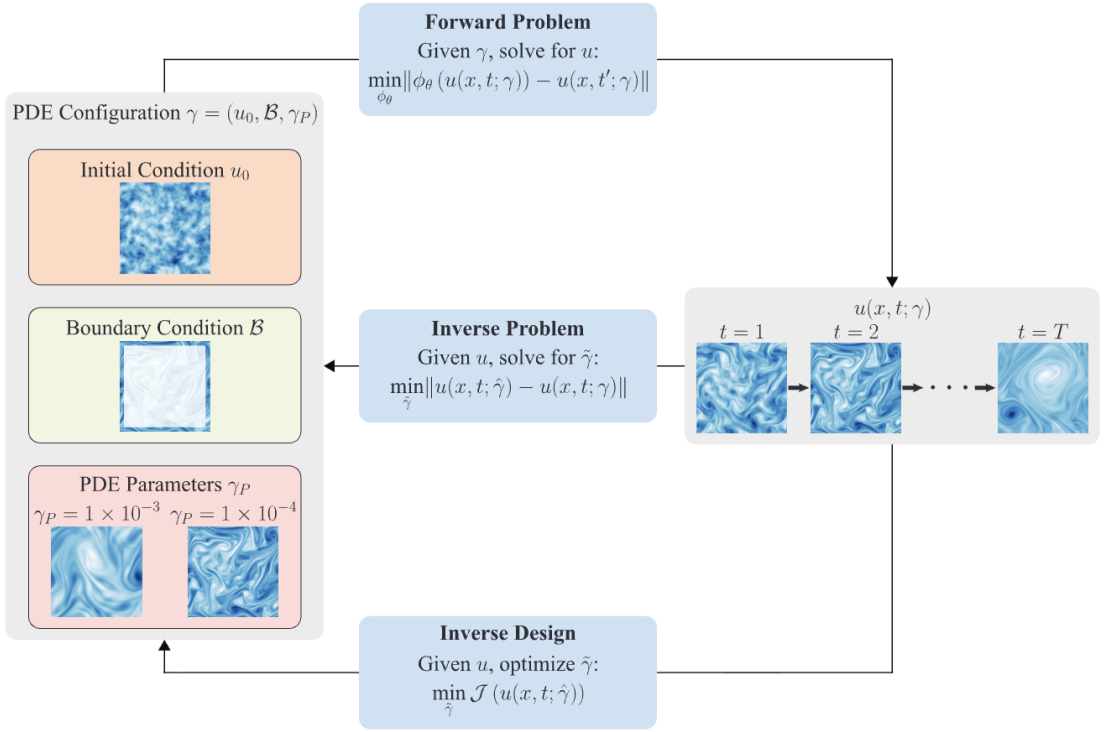


Boltzmann machine

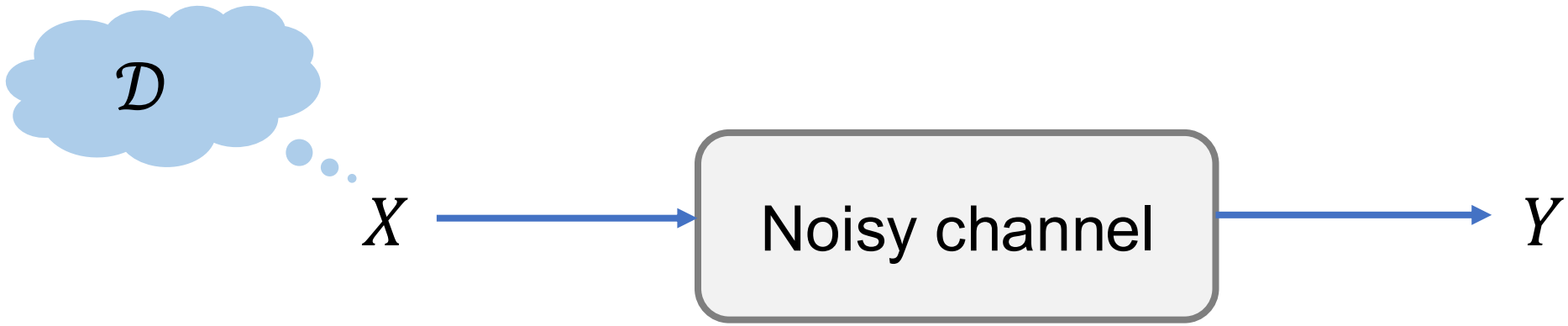
ML for Physics







Physics & Inference



Observing Y, how can you infer X?

- **Teacher-student model for neural network learning:** X is the weight of a neural network F_X. Giving data $\{(x_i, y_i) : x_i \sim \mathcal{N}(0, I_d), y_i = F_X(x_i)\}$, can you learn the neural network weight X?
- **Image denoising:** X is an image from a distribution $\mathcal{D}$, Y is the image with added noise to pixels. Can you recover the original image?

Physics & Inference

The most widely used inference method is the Bayes rule:

$$\Pr[X = x|Y = y] = \frac{\Pr[Y = y|X = x]\Pr[X = x]}{\Pr[Y = y]}$$

- $\Pr[X = x]$: our prior on X
- $\Pr[Y = y|X = x]$: likelihood of the noisy channel
- $\Pr[Y = y]$: normalization constant (or called partition function in physics)

$$Z \coloneqq \Pr[Y = y] = \int \Pr[Y = y|X = x]\Pr[X = x]\,\mathrm{d}x$$

- **Boltzmann energy distribution:**

$$\Pr[X = x|Y = y] = \frac{1}{Z}e^{-\mathcal{E}(x)}$$

$$\mathcal{E}(x) \coloneqq -\log(\Pr[Y = y|X = x]\Pr[X = x])$$

References

- Marc Mezard and Andrea Montanari. *Information, physics, and computation.* 2009.
- Florent Krzakala and Lenka Zdeborová. *Statistical Physics Methods in Optimization and Machine Learning: An Introduction to Replica, Cavity & Message-Passing techniques.* 2022.
- Martin J. Wainwright and Michael Irwin Jordan. *Graphical models, exponential families, and variational inference.* 2008.
- Andrea Montanari and Subhabrata Sen. *A short tutorial on mean-field spin glass techniques for non-physicists.* 2022.

Undirected Graphical Models

- Let $\{\psi_{ij}\}_{(ij)\in F}$ be the compatibility function $\{-1,1\}^2 \to \mathbb{R}_{\geq 0}$ that dictate interactions between particles
- F is a family of edges
- The Gibbs measure μ is a distribution over $\{-1,1\}^n$ given by

$$\mathcal{E}(\boldsymbol{x}) = -\sum_{(ij)\in F} \log\left(\psi_{ij}\left(x_i, x_j\right)\right),$$

$$\mu(\boldsymbol{x}) = \frac{1}{Z} e^{-\mathcal{E}(\boldsymbol{x})} = \frac{1}{Z} \prod_{(ij)\in F} \psi_{ij}\left(x_i, x_j\right),$$

$$Z = \sum_{\boldsymbol{x}\in\{-1,1\}^n} e^{-\mathcal{E}(\boldsymbol{x})}.$$

- $\boldsymbol{x} \in \{-1,1\}^n$ is a configuration of spins

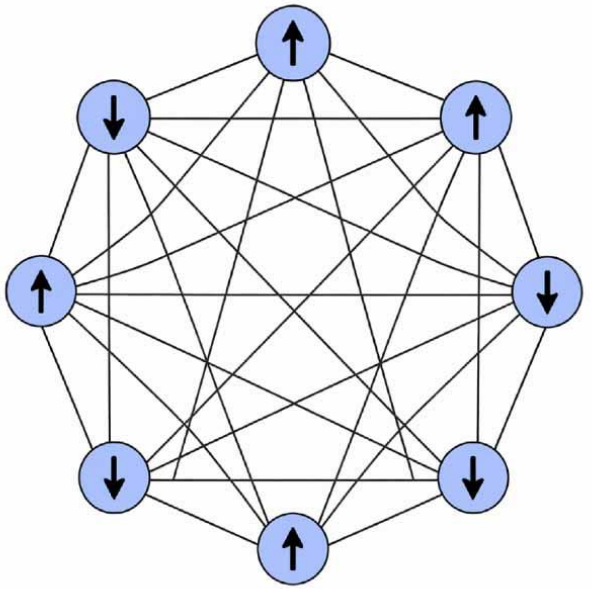
Undirected Graphical Models

Ising models

- $\psi_{ij}(x_i, x_j) = \exp(\beta x_i A_{ij} x_j)$ for some matrix A such that $A_{ii} = 0 \; \forall i \in [n]$ (why?) and A is symmetric

$$\mu(\boldsymbol{x}) \propto \exp\left(+\frac{\beta}{2}\boldsymbol{x}^\top A\boldsymbol{x}\right)$$

- β is the inverse temperature
- A is the Hamiltonian or the interaction matrix of the system
- $\mathcal{E}(\boldsymbol{x}) = -\frac{\beta}{2}\boldsymbol{x}^\top A\boldsymbol{x}$ is the energy of the system
- When $\beta \to 0$ (high-temperature limit), $\mu \to \mathrm{Unif}(\{-1,1\}^n)$
- When $\beta \to \infty$ (low-temperature limit), $\mu \to \mathrm{Unif}(\arg\min \mathcal{E}(\boldsymbol{x}))$

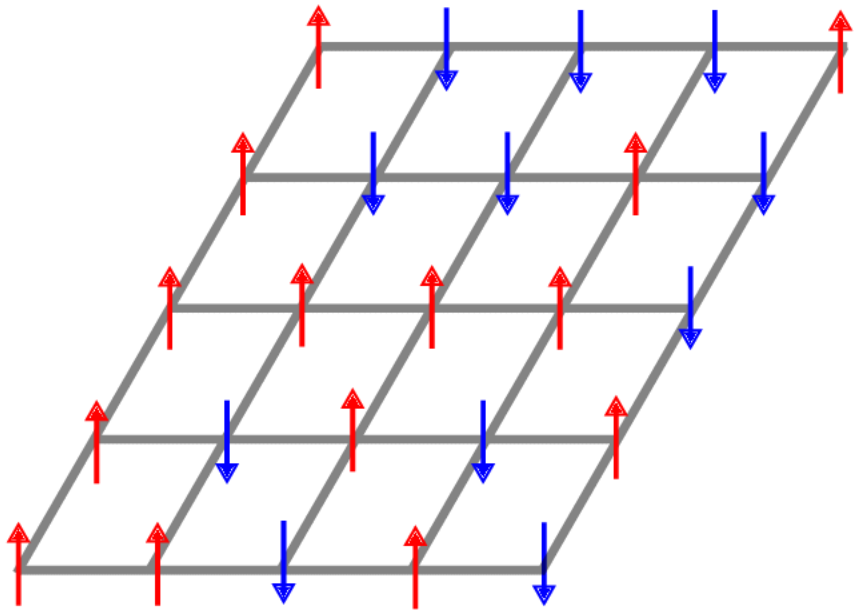


A defines a weighted graph G

Undirected Graphical Models

$$\mu(\boldsymbol{x}) = \frac{\exp\left(\frac{\beta}{2}\boldsymbol{x}^\top A\boldsymbol{x}\right)}{Z}, \qquad Z = \sum_{\boldsymbol{x}\in\{-1,1\}^n} \exp\left(\frac{\beta}{2}\boldsymbol{x}^\top A\boldsymbol{x}\right)$$

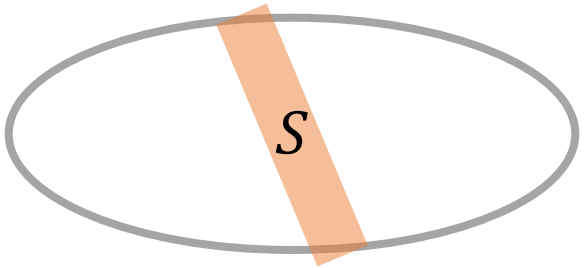
- Suppose A is the adjacent matrix of a 2D lattice (unit weight)
- $\mathcal{E}(\boldsymbol{x}) = -\frac{\beta}{2}\boldsymbol{x}^\top A\boldsymbol{x} = -\frac{\beta}{2}\sum_{(i,j)\in F} x_i x_j$
- $\mathcal{E}(\boldsymbol{x})$ is minimized when $x_1 = x_2 = \cdots = x_n = \pm 1$



Undirected Graphical Models

Markov property

- For $S \subseteq [n]$ such that $[n] \backslash S$ decomposes into disjoint pieces in G, then conditioned on any assignment to the spins in S, the marginal distributions on the pieces are independent



- For Ising models, for any $i \in [n]$, let $S = \partial_i$ (the neighbors of i in G)

$$\Pr\left[x_i = \sigma \middle| \boldsymbol{x}_{\partial_i} = \boldsymbol{s}\right] \propto \exp\left(-\beta\sigma \sum_{j \in \partial_i} A_{ij} s_j\right)$$

Undirected Graphical Models

Markov property

- For $S \subseteq [n]$ such that $[n] \backslash S$ decomposes into disjoint pieces in G, then conditioned on any assignment to the spins in S, the marginal distributions on the pieces are independent

- Global Markov property: the above condition holds for any $S \subseteq [n]$
- Local Markov property: the above condition holds for $S = \partial_i \; \forall i \in [n]$
- Pairwise Markov property: the above condition holds for $S = [n] \backslash \{i, j\}$ for any non-adjacent $i, j \in [n]$

Global Markov $\Longrightarrow$ Local Markov $\Longrightarrow$ Pairwise Markov

Hammersley–Clifford theorem: if $\mu(\boldsymbol{x}) > 0$ for any $\boldsymbol{x} \in \{-1,1\}^n$ (i.e., a Gibbs distribution),
then Pairwise Markov $\Longrightarrow$ Global Markov

Fundamental Algorithmic Problems in Inference

1. Draw samples from the Gibbs distribution μ
2. Compute the partition function Z

Jerrum, Valiant, and Vazirani '86: the equivalence of sampling and counting

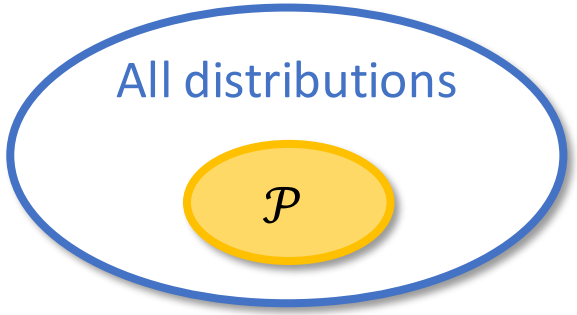
- Computing the partition function exactly is very hard: for $\psi_{ij}(x_i, x_j) = \mathbf{1}_{(x_i,x_j)\neq(1,1)}$, computing Z is counting the number of independent sets of the interaction graph, which is a **#P**-complete problem
- Our goal is to estimate Z approximately. Some approaches:
  - Markov chain Monte Carlo (MCMC)
  - Variational inference (VI)
  - Diffusion models

Variational Inference

Approximate μ by distributions from some family $\mathcal{P}$ of "simpler" distributions that are easy to sample from (e.g., product distributions a.k.a. naïve mean-field):

$$\min_{\nu \in \mathcal{P}} \mathrm{KL}(\nu \| \mu) \geq 0$$

- If $\mathcal{P} = \{\text{all distributions}\}$, then the minimizer is $\nu = \mu$ (Gibbs' inequality)



- **Issue:** the KL optimizer is hard to evaluate

$$\mathrm{KL}(\nu \| \mu) = \mathbb{E}_{\boldsymbol{x} \sim \nu}\left[\log \frac{\nu(\boldsymbol{x})}{\mu(\boldsymbol{x})}\right] = \mathbb{E}_{\boldsymbol{x} \sim \nu}\left[\log \frac{\nu(\boldsymbol{x})}{e^{-\mathcal{E}(\boldsymbol{x})}/Z}\right] = \underbrace{\mathbb{E}_{\boldsymbol{x} \sim \nu}[\log \nu(\boldsymbol{x})]}_{-H(\nu)} + \underbrace{\mathbb{E}_{\boldsymbol{x} \sim \nu}[\mathcal{E}(\boldsymbol{x})]}_{\text{Averaged energy}} + \log Z$$

- Define the Gibbs free energy $G[\nu] \coloneqq -H(\nu) + \mathbb{E}_{\boldsymbol{x} \sim \nu}[\mathcal{E}(\boldsymbol{x})]$ $-1\times$ "evidence lower bound" (ELBO)
- To minimize the KL divergence, it suffices to minimize the free energy $G[\nu]$

Detour: Evidence Lower Bound

$$G[q] \coloneqq -H(q) + \mathbb{E}_{\boldsymbol{x}\sim q}[\mathcal{E}(\boldsymbol{x})], \qquad \mathcal{E}(\boldsymbol{x}) = -\log \Pr\big[[X = \boldsymbol{x}, Y = \boldsymbol{y}]\big]$$

$$\mathrm{KL}\big(q(X)\big\|\Pr[X|Y = \boldsymbol{y}]\big) = \mathbb{E}_{\boldsymbol{x}\sim q}[\log q(\boldsymbol{x})] - \mathbb{E}_{\boldsymbol{x}\sim q}[\log \Pr[X = \boldsymbol{x}, Y = \boldsymbol{y}]] + \log \Pr[Y = \boldsymbol{y}]$$

$$\log \Pr[Y = \boldsymbol{y}] = \mathbb{E}_{\boldsymbol{x}\sim q}[\log \Pr[X = \boldsymbol{x}, Y = \boldsymbol{y}]] - \mathbb{E}_{\boldsymbol{x}\sim q}[\log q(\boldsymbol{x})] + \mathrm{KL}\big(q(X)\big\|\Pr[X|Y = y]\big)$$

$$\mathrm{ELBO}(q) \coloneqq \mathbb{E}_{\boldsymbol{x}\sim q}[\log \Pr[X = \boldsymbol{x}, Y = \boldsymbol{y}]] - \mathbb{E}_{\boldsymbol{x}\sim q}[\log q(\boldsymbol{x})] = \mathbb{E}_{\boldsymbol{x}\sim q}\left[\frac{\log \Pr[X = \boldsymbol{x}, Y = \boldsymbol{y}]}{\log q(\boldsymbol{x})}\right]$$

- $\mathrm{ELBO}(q)$ is a lower bound for the log-evidence $\log \Pr[Y]$
- Frequently used in ML practice (e.g., variational autoencoder, diffusion models, EM algorithms, …)

Variational Inference

$$G[\nu] \coloneqq -H(\nu) + \mathbb{E}_{\boldsymbol{x}\sim\nu}[\mathcal{E}(\boldsymbol{x})]$$

Consider the Ising model:

$$\mu(\boldsymbol{x}) \propto \exp\left(\frac{\beta}{2}\boldsymbol{x}^\top A\boldsymbol{x}\right)$$

$$G[\nu] = -\mathbb{E}_{\boldsymbol{x}\sim\nu}\left[\frac{\beta}{2}\boldsymbol{x}^\top A\boldsymbol{x}\right] - H(\nu)$$

Entropy regularization

- When β is large, minimization prioritizes minimizing the averaged energy
- When β is small, minimization prioritizes maximizing the entropy
- For a fixed $\nu \in \mathcal{P}$, $G[\nu]$ is easy to evaluate. But it is hard to optimize a priori
- We will introduce a general approach: the belief propagation (BP), to minimize $G[\nu]$

Belief Propagation

- Belief propagation (BP) is a powerful, general heuristic to solve $\min_{\nu} G[\nu]$
- When the interaction graph is a tree, BP is exact

Two interpretations of BP:

1. Dynamic programming on a tree
2. Nonconvex optimization for finding a stationary point of a relaxation of the Gibbs free energy (Bethe free energy)

In the following discussions, we choose a canonical example: random independent sets

- Let $G = (V, E)$ be a graph (in particular, a tree)
- The compatibility function is $\psi_{ij}(x_i x_j) = \mathbf{1}_{(x_i, x_j) \neq (1,1)}$ for every edge $(i, j) \in E$

Belief Propagation

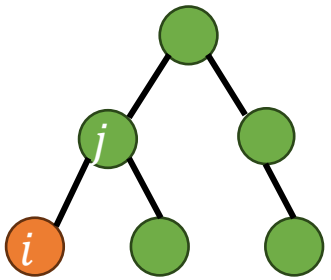
- The Gibbs free energy on a tree only depends on 1- and 2-wise marginals of μ

Lemma. If the interaction graph $G = (V, E)$ is a tree, then for any distribution μ on G,

$$\mu(\boldsymbol{x}) = \prod_{(i,j)\in E} \mu_{ij}(x_i, x_j) \prod_{i\in V} \mu_i(x_i)^{1-|\partial i|}$$

Proof.

- We prove by induction: $|V| = 1$ is trivially true
- For $|V| > 1$, let i be any leaf node of the tree connected by (i, j)



$$\begin{aligned}
\Pr_{\boldsymbol{x}\sim\mu}[\boldsymbol{x} = s] &= \Pr_{\mu}\left[\boldsymbol{x}_{V\setminus\{i\}} = \boldsymbol{s}_{V\setminus\{i\}}\right] \Pr_{\mu}\left[x_i = s_i \middle| \boldsymbol{x}_{V\setminus\{i\}} = \boldsymbol{s}_{V\setminus\{i\}}\right] \\
&= \Pr_{\mu}\left[\boldsymbol{x}_{V\setminus\{i\}} = \boldsymbol{s}_{V\setminus\{i\}}\right] \Pr_{\mu}\left[x_i = s_i \middle| x_j = s_j\right] \qquad \text{(Markov property)} \\
&= \Pr_{\mu}\left[\boldsymbol{x}_{V\setminus\{i\}} = \boldsymbol{s}_{V\setminus\{i\}}\right] \mu_{ij}(s_i, s_j)/\mu_j(s_j)
\end{aligned}$$

Belief Propagation

Proof.

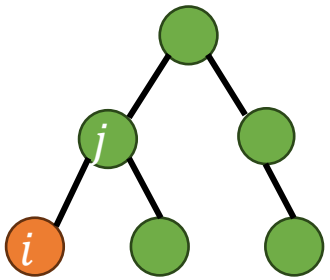
- By the induction hypothesis,

$$\Pr_{\mu}\left[\boldsymbol{x}_{[n]\setminus\{i\}} = \boldsymbol{s}_{[n]\setminus\{i\}}\right] = \prod_{(k,\ell)\in E\setminus(i,j)} \mu_{k\ell}(s_k, s_\ell) \prod_{k\in V\setminus\{i\}} \mu_k(s_k)^{1-|\partial_{V\setminus\{i\}}k|}$$

- Thus,

$$\Pr_{\boldsymbol{x}\sim\mu}[\boldsymbol{x} = s] = \prod_{(k,\ell)\in E\setminus(i,j)} \mu_{k\ell}(s_k, s_\ell) \prod_{k\in V\setminus\{i\}} \mu_k(s_k)^{1-|\partial_{V\setminus\{i\}}k|} \cdot \frac{\mu_{ij}(s_i, s_j)}{\mu_j(s_j)}$$

$$= \prod_{(k,\ell)\in E} \mu_{k\ell}(s_k, s_\ell) \prod_{k\in V} \mu_k(s_k)^{1-|\partial_V k|}$$



■

Belief Propagation

- The Gibbs free energy on a tree only depends on 1- and 2-wise marginals of μ

Lemma. If the interaction graph $G = (V, E)$ is a tree, then then for any distribution μ on G,

$$\mu(\boldsymbol{x}) = \prod_{(i,j)\in E} \mu_{ij}(x_i, x_j) \prod_{i\in V} \mu_i(x_i)^{1-|\partial i|}$$

- The Gibbs free energy is defined as

$$G[\mu] \coloneqq -H(\mu) + \mathbb{E}_{\boldsymbol{x}\sim\mu}[\mathcal{E}(\boldsymbol{x})]$$

$$\mathbb{E}_{\boldsymbol{x}\sim\mu}[\mathcal{E}(\boldsymbol{x})] = \sum_{(i,j)\in E} -\mathbb{E}_{\mu_{ij}}[\log \psi_{ij}(x_i, x_j)]$$

$$-H(\mu) = \mathbb{E}_{\boldsymbol{x}\sim\mu}[\log \mu(\boldsymbol{x})] = \sum_{(i,j)\in E} \mathbb{E}_{\mu_{ij}}[\log \mu_{ij}(x_i, x_j)] + \sum_{i\in V} (1 - |\partial i|) \cdot \mathbb{E}_{\mu_i}[\log \mu_i(x_i)]$$

Belief Propagation

- The Gibbs free energy on a tree only depends on 1- and 2-wise marginals of μ

Lemma. If the interaction graph $G = (V, E)$ is a tree, then then for any distribution μ on G,

$$\mu(\boldsymbol{x}) = \prod_{(i,j)\in E} \mu_{ij}(x_i, x_j) \prod_{i\in V} \mu_i(x_i)^{1-|\partial i|}$$

- Thus, we have

$$G[\mu] = \sum_{(i,j)\in E} \mathbb{E}_{\mu_{ij}}\left[\log \frac{\mu_{ij}(x_i, x_j)}{\psi_{ij}(x_i, x_j)}\right] + \sum_{i\in V} (1 - |\partial i|) \cdot \mathbb{E}_{\mu_i}[\log \mu_i(x_i)]$$

- It suffices to compute the marginals μ_i and μ_{ij}

Remark: The marginal expectation $\{\mathbb{E}_{\mu_i}[x_i]\}_{i\in[n]}$ is also the MMSE estimator:

$$\arg\min_{\hat{x}(\cdot)} \mathbb{E}_{X,Y}[\|X - \hat{x}(Y)\|^2] = \mathbb{E}_Y\left[\mathbb{E}_{X|Y}[\|X - \hat{x}(Y)\|^2]\right]$$

Belief Propagation

Random Independent Sets

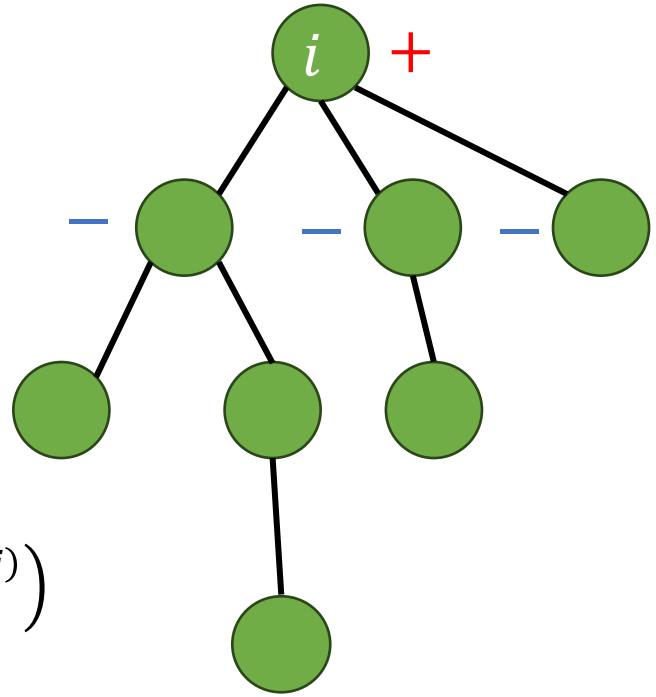
- Let $G = (V, E)$ be a graph (in particular, a tree)
- The compatibility function is $\psi_{ij}(x_i x_j) = \mathbf{1}_{(x_i, x_j) \neq (1,1)}$ for every edge $(i, j) \in E$

How to compute $\mu_i(x_i)$ on a tree?

Dynamic programming!

- We can take i to be the root
- Let $Z_\sigma^{(i)} :=$ the number of independent sets in the (sub-)tree with root i being labelled $\sigma \in \{-1, 1\}$
- We have the DP recursion:

$$Z_+^{(i)} = \prod_{j \in \partial i} Z_-^{(j)}, \qquad Z_-^{(i)} = \prod_{j \in \partial i} \left(Z_-^{(j)} + Z_+^{(j)} \right)$$



Belief Propagation

Random Independent Sets

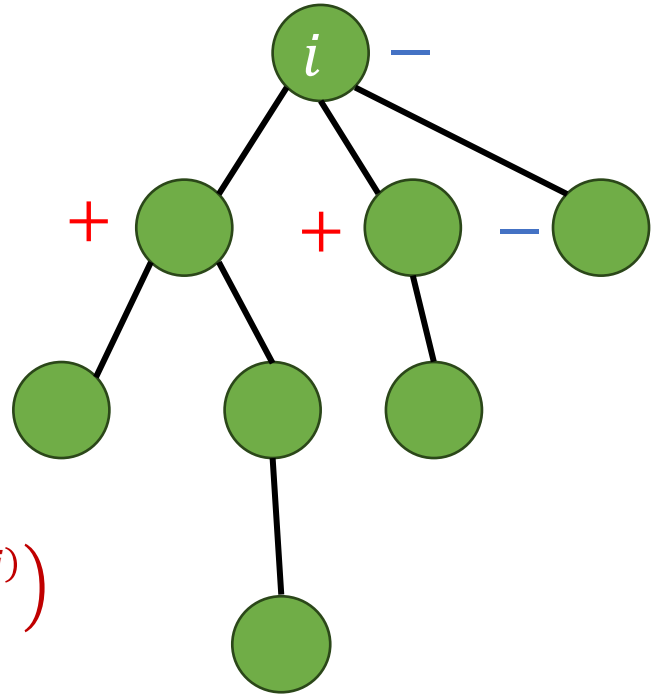
- Let $G = (V, E)$ be a graph (in particular, a tree)
- The compatibility function is $\psi_{ij}(x_i x_j) = \mathbf{1}_{(x_i, x_j) \neq (1,1)}$ for every edge $(i, j) \in E$

How to compute $\mu_i(x_i)$ on a tree?

Dynamic programming!

- We can take i to be the root
- Let $Z_\sigma^{(i)} :=$ the number of independent sets in the (sub-)tree with root i being labelled $\sigma \in \{-1, 1\}$
- We have the DP recursion:

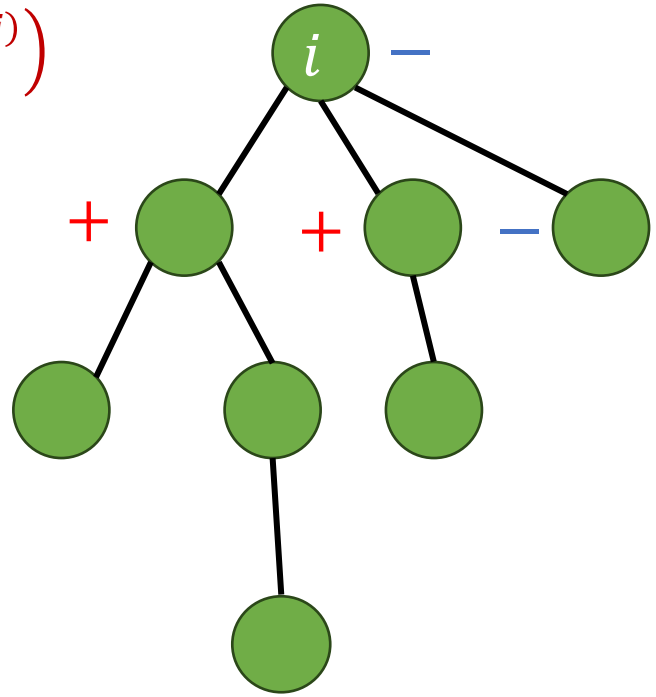
$$Z_+^{(i)} = \prod_{j \in \partial i} Z_-^{(j)}, \qquad Z_-^{(i)} = \prod_{j \in \partial i} \left(Z_-^{(j)} + Z_+^{(j)} \right)$$



Belief Propagation

- We can take i to be the root
- Let $Z_\sigma^{(i)} :=$ the number of independent sets in the (sub-)tree with root i being labelled $\sigma \in \{-1,1\}$
- We have the DP recursion:

$$Z_+^{(i)} = \prod_{j \in \partial i} Z_-^{(j)}, \qquad Z_-^{(i)} = \prod_{j \in \partial i} \left(Z_-^{(j)} + Z_+^{(j)} \right)$$



- The marginal is

$$\Pr[x_i = +1] = \frac{Z_+^{(i)}}{Z_+^{(i)} + Z_-^{(i)}} \propto Z_+^{(i)}$$

$$\Pr[x_i = -1] \propto Z_-^{(i)}$$

- This DP relies on a global top-down tree structure
  BP takes a more local view (which makes it work for general graphs)

Belief Propagation

We need to define several notations

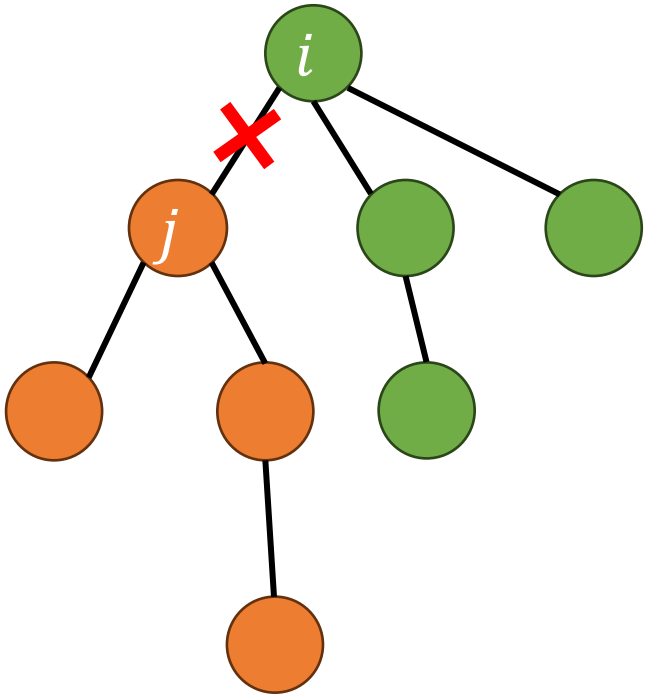
- Let $V^{j\to i}$ denote the sub-tree by cutting the edge (i,j) and keeping the part containing j
- Let $m_\sigma^{j\to i}$ denote the probability that $x_j=\sigma$ in $V^{j\to i}$

a "message" that node j passes to node i

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

$$Z_+^{(i)}=\prod_{j\in\partial i} Z_-^{(j)}=\prod_{j\in\partial i}\Pr_{\mu_{Vj\to i}}[x_j=-]\left(Z_-^{(j)}+Z_+^{(j)}\right)$$

$$\propto \prod_{j\in\partial i}\Pr_{\mu_{Vj\to i}}[x_j=-]=\prod_{j\in\partial i} m_-^{j\to i}$$



$$Z_+^{(i)}=\prod_{j\in\partial i} Z_-^{(j)}, \qquad Z_-^{(i)}=\prod_{j\in\partial i}\left(Z_-^{(j)}+Z_+^{(j)}\right)$$

$$\Pr_{\mu_V}[x_i=\sigma]\propto Z_\sigma^{(i)}$$

Belief Propagation

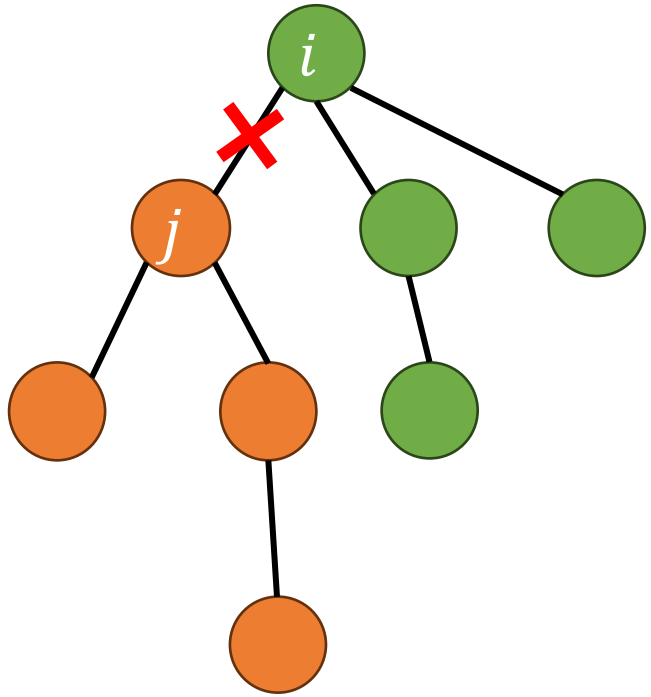
$$\Pr_{\mu}[x_i = \sigma] \propto \sum_{s \in \{\pm 1\}^n : s_i = \sigma} \prod_{(k,\ell) \in E} \psi_{k\ell}(s_k, s_\ell)$$

$$= \sum_{s \in \{\pm 1\}^n : s_i = \sigma} \prod_{j \in \partial i} \left(\psi_{ij}(\sigma, s_j) \prod_{(k,\ell) \in E^{j \to i}} \psi_{k\ell}(s_k, s_\ell) \right)$$

$$= \prod_{j \in \partial i} \sum_{s_{V^{j \to i}}} \left(\psi_{ij}(\sigma, s_j) \prod_{(k,\ell) \in E^{j \to i}} \psi_{k\ell}(s_k, s_\ell) \right)$$

$$= \prod_{j \in \partial i} \sum_{s_j \in \{\pm 1\}} \psi_{ij}(\sigma, s_j) \underbrace{\left(\sum_{s_{V^{j \to i}_{\setminus \{j\}}}} \prod_{(k,\ell) \in E^{j \to i}} \psi_{k\ell}(s_k, s_\ell) \right)}_{\propto m_{s_j}^{j \to i}}$$

$$\boxed{\Pr_{\mu}[x_i = \sigma] \propto \prod_{j \in \partial i} \sum_{s \in \{\pm 1\}} \psi_{ij}(\sigma, s) m_s^{j \to i}}$$



$$\Pr_{\mu}[x_i = \sigma] = \frac{\Pr_{\mu}[x_i = \sigma]}{\Pr_{\mu}[x_i = \sigma] + \Pr_{\mu}[x_i = -\sigma]}$$

Belief Propagation

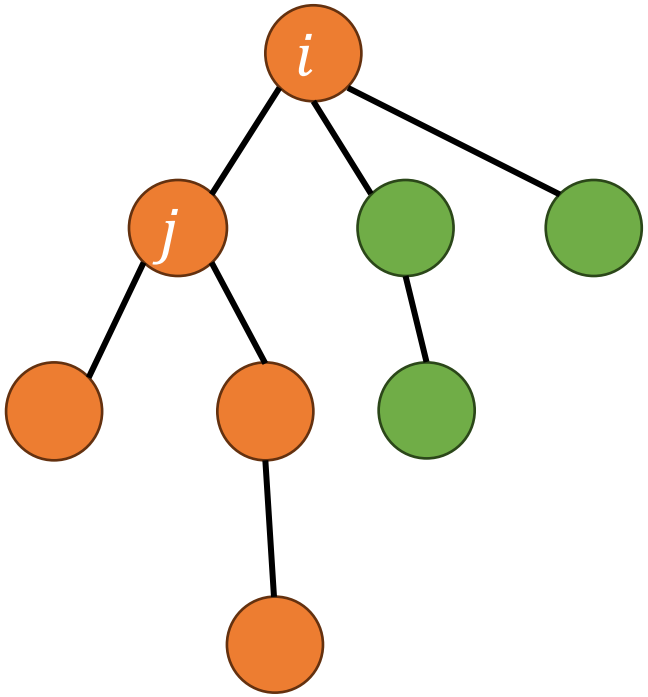
We need to define several notations

- Let $\overline{V}^{j\to i}$ denote the sub-tree by cutting the edge (i,j) and keeping the part containing j and connecting (i,j)
- Let $\overline{m}_{\sigma}^{j\to i}$ denote the probability that $x_i = \sigma$ in $\overline{V}^{j\to i}$

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

$$\overline{m}_{\sigma}^{j\to i} \propto \sum_{s\in\{\pm 1\}} \psi_{ij}(\sigma, s) m_s^{j\to i}$$

$$\Pr_{\mu}[x_i = \sigma] \propto \prod_{j\in\partial i} \sum_{s\in\{\pm 1\}} \psi_{ij}(\sigma, s) m_s^{j\to i}$$

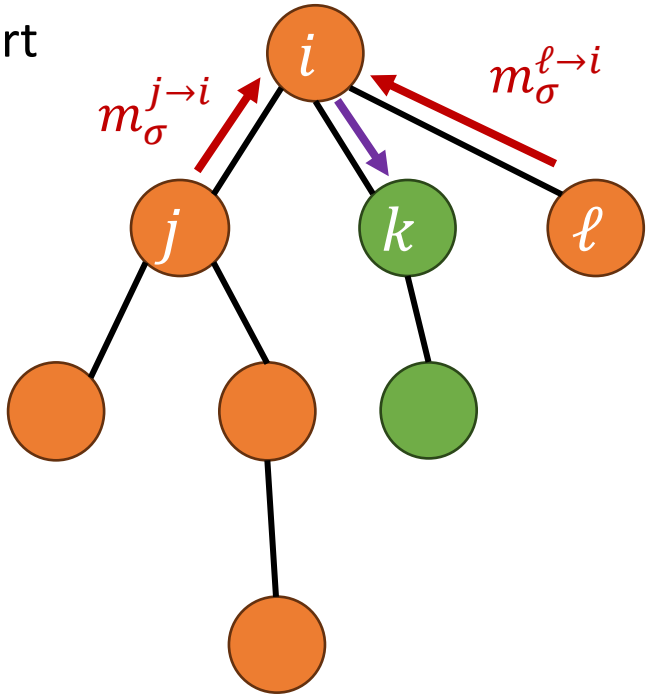


Belief Propagation

We need to define several notations

- Let $\overline{V}^{j\to i}$ denote the sub-tree by cutting the edge (i,j) and keeping the part containing j and connecting (i,j)
- Let $\overline{m}_\sigma^{j\to i}$ denote the probability that $x_i=\sigma$ in $\overline{V}^{j\to i}$

$$\overline{m}_\sigma^{j\to i} \propto \sum_{s\in\{\pm 1\}} \psi_{ij}(\sigma,s) m_s^{j\to i}$$



Given $\left\{m_\sigma^{j\to i} : j\in\partial i, \sigma\in\{\pm 1\}\right\}$, how to compute $m_\sigma^{i\to k}$ for some $k\in\partial i$?

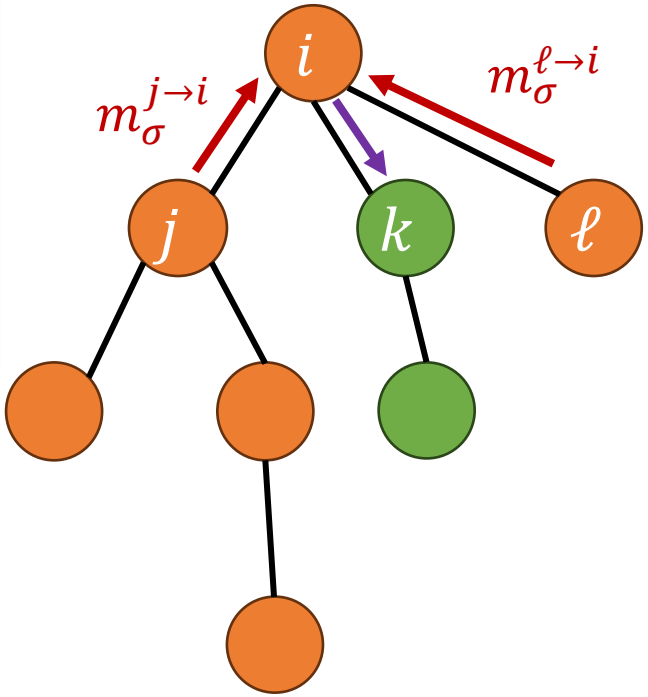
$$m_\sigma^{i\to k} = \Pr_{\mu_{V^{i\to k}}}[x_i=\sigma] \propto \prod_{j\in\partial_{V^{i\to k}} i} \sum_{s\in\{\pm 1\}} \psi_{ij}(\sigma,s) m_s^{j\to i}$$

$$\propto \prod_{j\in\partial i\setminus\{k\}} \overline{m}_\sigma^{j\to i}$$

Belief Propagation

$$m_\sigma^{i\to k} \propto \prod_{j\in\partial i\setminus\{k\}} \overline{m}_\sigma^{j\to i}, \qquad \overline{m}_\sigma^{j\to i} \propto \sum_{s\in\{\pm 1\}} \psi_{ij}(\sigma,s) m_s^{j\to i}$$

$$\Pr_\mu[x_i=\sigma] \propto \prod_{j\in\partial i} \overline{m}_\sigma^{j\to i} \propto m_\sigma^{i\to k}\cdot \overline{m}_\sigma^{k\to i}$$

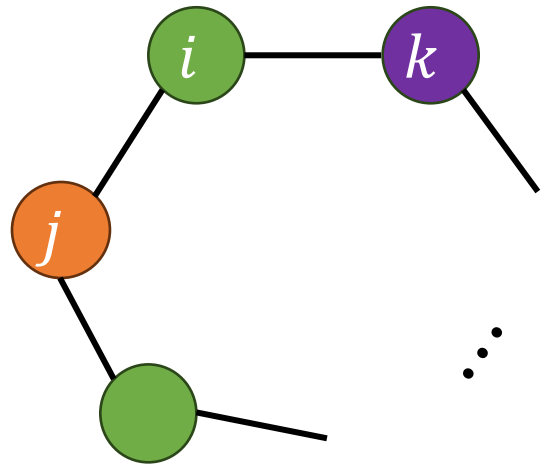


BP algorithm:

- Initialize $m_\sigma^{j\to i} = 1/2$ for all $(i,j) \in E$
- Iteratively update $\left\{m_\sigma^{j\to i}, \overline{m}_\sigma^{j\to i} : (i,j)\in E, \sigma\in\{\pm 1\}\right\}$ in parallel

Example: Belief Propagation on a circle

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



- $m_{+}^{i\to k} \propto m_{-}^{j\to i}$
- $m_{-}^{i\to k} \propto m_{+}^{j\to i} + m_{-}^{j\to i}$
- If $\left(m_{+}^{j\to i}, m_{-}^{j\to i}\right) = (p, 1-p)$ before the update, then

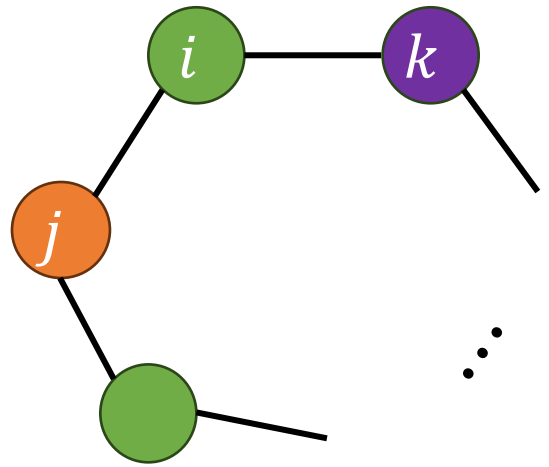
$$m_{+}^{i\to k} = \frac{1-p}{1-p+1} = \frac{1-p}{2-p}, \qquad m_{-}^{i\to k} = \frac{1}{2-p}$$

- We initialize the messages with $p = 1/2$ at the beginning
- After one round of BP updates, $p \longmapsto \frac{1-p}{2-p}$
- $\frac{1}{2} \longrightarrow \frac{1}{3} \longrightarrow \frac{2}{5} \longrightarrow \frac{3}{8} \longrightarrow \frac{5}{13} \longrightarrow \cdots \longrightarrow \frac{1}{\varphi^2} \approx 0.382$

$\varphi = \frac{\sqrt{5}+1}{2}$ is the golden ratio

Example: Belief Propagation on a circle

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



- $m_{+}^{i\to k} \propto m_{-}^{j\to i}$
- $m_{-}^{i\to k} \propto m_{+}^{j\to i} + m_{-}^{j\to i}$
- If $\left(m_{+}^{j\to i}, m_{-}^{j\to i}\right) = (p, 1-p)$ before the update, then

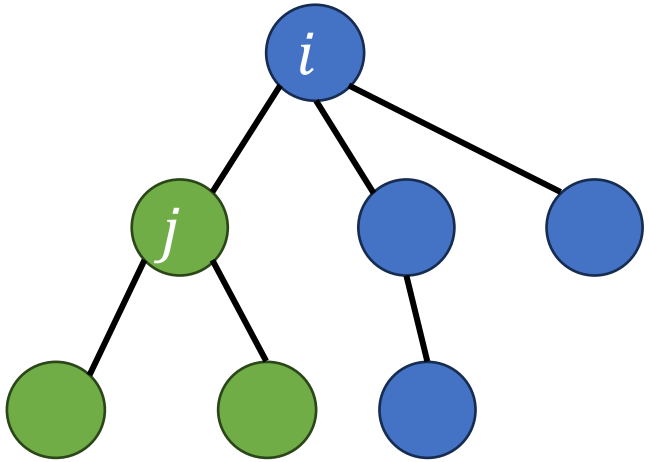
$$m_{+}^{i\to k} = \frac{1-p}{1-p+1} = \frac{1-p}{2-p}, \qquad m_{-}^{i\to k} = \frac{1}{2-p}$$

- We initialize the messages with $p = 1/2$ at the beginning
- After one round of BP updates, $p \longmapsto \frac{1-p}{2-p}$
- $\frac{1}{2} \longrightarrow \frac{1}{3} \longrightarrow \frac{2}{5} \longrightarrow \frac{3}{8} \longrightarrow \frac{5}{13} \longrightarrow \cdots \longrightarrow \frac{1}{\varphi^2} \approx 0.382$

$\varphi = \frac{\sqrt{5}+1}{2}$ is the golden ratio

Belief Propagation

$$\Pr_{\mu}[x_i = \sigma] \propto \prod_{j \in \partial i} \overline{m}_{\sigma}^{j \to i} \propto m_{\sigma}^{i \to k} \cdot \overline{m}_{\sigma}^{k \to i}$$



The messages can also compute 2-wise marginals

- For any $(i,j) \in E$, let $\partial(i,j) \coloneqq (\partial i \cup \partial j) \backslash \{i,j\}$
- For $k \in \partial(i,j)$, let $i(k)$ denote the node such that $\big(k, i(k)\big) \in E$

$$\Pr_{\mu}\big[(x_i, x_j) = (\sigma_i, \sigma_j)\big] \propto \psi_{ij}(\sigma_i, \sigma_j) \left(\sum_{\boldsymbol{s}_V j \to i : s_j = \sigma_j} \prod_{(k,\ell) \in E^{j \to i}} \psi_{k\ell}(s_k, s_\ell) \right) \left(\sum_{\boldsymbol{s}_V i \to j : s_i = \sigma_i} \prod_{(k,\ell) \in E^{i \to j}} \psi_{k\ell}(s_k, s_\ell) \right)$$

$$\propto \psi_{ij}(\sigma_i, \sigma_j) m_{\sigma_j}^{j \to i} m_{\sigma_i}^{i \to j}$$

Belief Propagation

Theorem (BP is exact on trees**).** If G is a tree, belief propagation exactly recovers the marginals of the Gibbs measure

- If G is a general graph, BP is very hard to analyze rigorously
- BP can work when the graph "locally" looks like a tree
- Suppose G is a random graph where every edge is included with probability c/n for $c = O(1)$
- For any vertex v, it has $\approx c$ descendants; and $\approx c^d$ descendants of distance $\leq d$
- Since each descendant is a uniformly random vertex, the probability that there is no path from v to itself of length $\leq d$ is

$$\left(1 - \frac{1}{n}\right)^{c^d} \approx e^{-c^d/n} \to 1 \quad \text{when} \quad c^d \ll n, \qquad \text{i.e.,} \qquad d \ll \log n$$

Belief Propagation

Two interpretations of BP:

1. Dynamic programming on a tree
2. Nonconvex optimization for finding a stationary point of a relaxation of the Gibbs free energy (Bethe free energy)

Bethe free energy

- Recall that

$$H(\mu) = -\sum_{(i,j)\in E} \mathbb{E}_{\mu_{ij}}\left[\log \mu_{ij}(x_i, x_j)\right] + \sum_{i\in V}(|\partial i| - 1)\cdot \mathbb{E}_{\mu_i}[\log \mu_i(x_i)]$$

$$= \sum_{(i,j)\in E} H(\mu_{ij}) - \sum_{i\in V}(|\partial i| - 1)\cdot H(\mu_i)$$

- Gibbs free energy: $G[\mu] \coloneqq -H(\mu) + \mathbb{E}_{\boldsymbol{x}\sim\mu}[\mathcal{E}(\boldsymbol{x})]$

Belief Propagation

Bethe free energy

$$G_\beta[\nu] \coloneqq -\sum_{(i,j)\in E} H(\nu_{ij}) + \sum_{i\in V}(|\partial i| - 1)\cdot H(\nu_i) + \mathbb{E}_{\boldsymbol{x}\sim\nu}[\mathcal{E}(\boldsymbol{x})]$$

- ν is given by the marginal data $\{\nu_i, \nu_{ij}\}$ that is guaranteed to satisfy "local consistency":

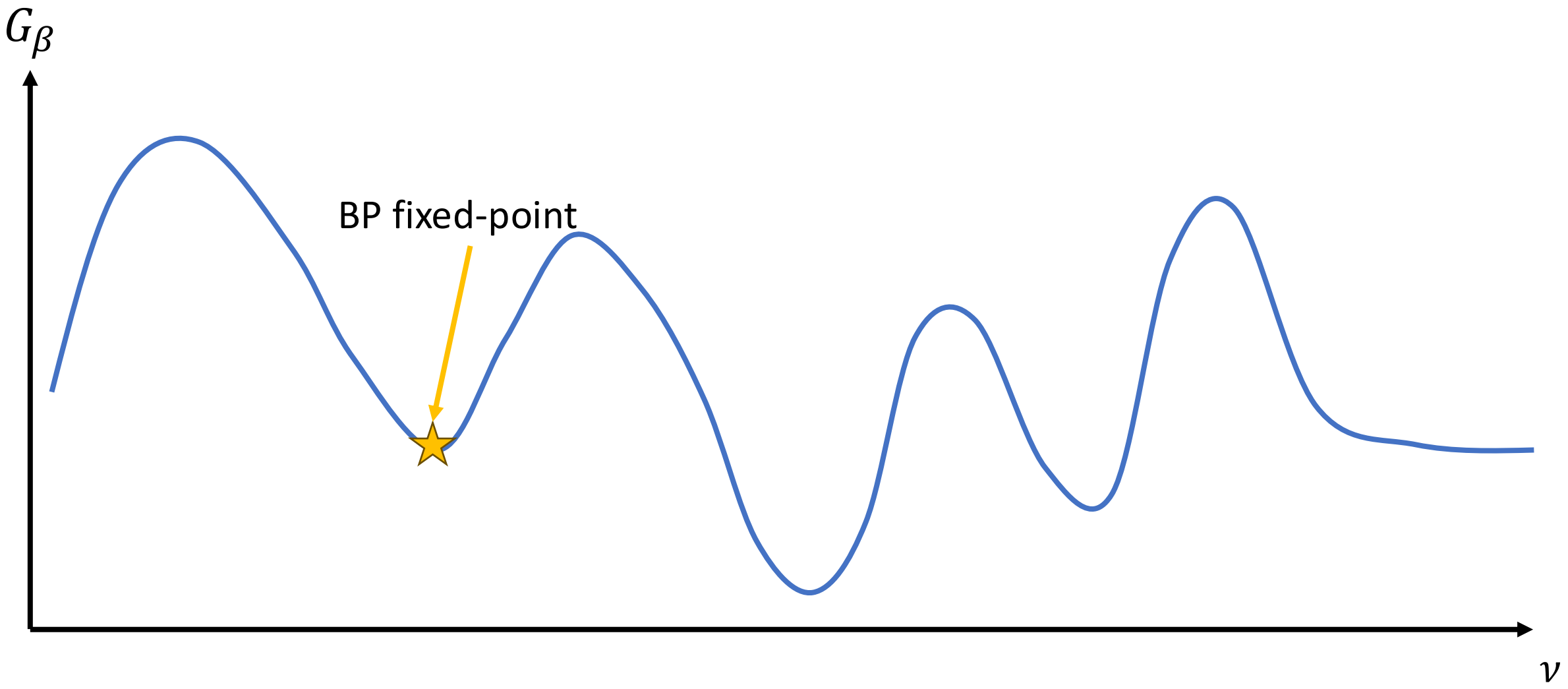
$$\sum_{x_j\in\{\pm 1\}} \nu_{ij}(x_i, x_j) = \nu_i(x_i), \qquad \forall i \in V, (i,j) \in E, x_i \in \{\pm 1\}$$

"degree-2 Sherali-Adams constraints"

- ν is called a pseudo-distribution
- On a tree,

Bethe free energy = Gibbs free energy

Belief Propagation



Belief Propagation

BP solution:

$$\nu_i(\sigma) \propto m_\sigma^{i\to j} \cdot \overline{m}_\sigma^{j\to i}, \qquad \nu_{ij}(\sigma_i,\sigma_j) \propto \psi_{ij}(\sigma_i,\sigma_j) m_{\sigma_j}^{j\to i} m_{\sigma_i}^{i\to j}$$

Lemma. Fixed points of BP satisfy local consistency.

Proof.

- For any $\sigma_i \in \{\pm 1\}$,

$$\sum_{\sigma_j\in\{\pm1\}} \nu_{ij}(\sigma_i,\sigma_j) \propto m_{\sigma_i}^{i\to j} \cdot \sum_{\sigma_j\in\{\pm1\}} \psi_{ij}(\sigma_i,\sigma_j) m_{\sigma_j}^{j\to i} \propto m_{\sigma_i}^{i\to j} \cdot \overline{m}_\sigma^{j\to i} \propto \nu_i(\sigma_i)$$

- If we sum over $\sigma_i \in \{\pm 1\}$, $\mathrm{LHS} = \mathrm{RHS} = 1$. So "$\propto$" is actually "$=$"

■

Belief Propagation

We define more notations:

$$Z_i \coloneqq \sum_{\sigma \in \{\pm 1\}} \prod_{j \in \partial i} \overline{m}_{\sigma}^{j \to i}$$

$$Z_{ij} \coloneqq \sum_{\sigma_i, \sigma_j \in \{\pm 1\}} \psi_{ij}(\sigma_i, \sigma_j) m_{\sigma_j}^{j \to i} m_{\sigma_i}^{i \to j}$$

$$Z_{i;j} \coloneqq \sum_{\sigma \in \{\pm 1\}} m_{\sigma}^{i \to j} \cdot \overline{m}_{\sigma}^{j \to i}$$

Lemma. Bethe free energy can be expressed as

$$G_{\beta}[\nu] = -\sum_{i \in [n]} \log Z_i - \sum_{(i,j) \in E} \log Z_{ij} + \sum_{i \in [n], j \in \partial i} \log Z_{i;j}$$

Belief Propagation

Lemma. Bethe free energy can be expressed as

$$G_\beta[\nu] = -\sum_{i\in[n]} \log Z_i - \sum_{(i,j)\in E} \log Z_{ij} + \sum_{i\in[n], j\in\partial i} \log Z_{i;j}$$

Proof.

$$G_\beta[\nu] \coloneqq -\sum_{(i,j)\in E} H(\nu_{ij}) + \sum_{i\in V} (|\partial i| - 1)\cdot H(\nu_i) + \mathbb{E}_{\boldsymbol{x}\sim\nu}[\mathcal{E}(\boldsymbol{x})]$$

$$-\sum_{(i,j)\in E} H(\nu_{ij}) + \mathbb{E}_\nu[\mathcal{E}(x)] = -\sum_{(i,j)\in E} \mathbb{E}_{\nu_{ij}}\left[\log \frac{\psi_{ij}(x_i, x_j)}{\nu_{ij}(x_i, x_j)}\right] = -\sum_{(i,j)\in E} \mathbb{E}_{\nu_{ij}}\left[\log \frac{Z_{ij}}{m_{x_j}^{j\to i} m_{x_i}^{i\to j}}\right]$$

$$= -\sum_{(i,j)\in E} \log Z_{ij} + \sum_{(i,j)\in E} \left(\mathbb{E}_{\nu_{ij}}\left[\log m_{x_j}^{j\to i}\right] + \mathbb{E}_{\nu_{ij}}\left[\log m_{x_i}^{i\to j}\right]\right)$$

Belief Propagation

Lemma. Bethe free energy can be expressed as

$$G_\beta[\nu] = -\sum_{i\in[n]} \log Z_i - \sum_{(i,j)\in E} \log Z_{ij} + \sum_{i\in[n], j\in\partial i} \log Z_{i;j}$$

Proof.

$$G_\beta[\nu] \coloneqq -\sum_{(i,j)\in E} H(\nu_{ij}) + \sum_{i\in V} (|\partial i| - 1)\cdot H(\nu_i) + \mathbb{E}_{\boldsymbol{x}\sim\nu}[\mathcal{E}(\boldsymbol{x})]$$

$$-\sum_{i\in[n]} H(\nu_i) = -\sum_{i\in[n]} \mathbb{E}_{\nu_i}\left[\log \frac{1}{\nu_i(x_i)}\right] = -\sum_{i\in[n]} \mathbb{E}_{\nu_i}\left[\log \frac{Z_i}{\prod_{j\in\partial i} \overline{m}_{x_i}^{j\to i}}\right]$$

$$= -\sum_{i\in[n]} \log Z_i + \sum_{i\in[n], j\in\partial i} \mathbb{E}_{\nu_i}\left[\log \overline{m}_{x_i}^{j\to i}\right]$$

Belief Propagation

Lemma. Bethe free energy can be expressed as

$$G_\beta[\nu] = -\sum_{i\in[n]} \log Z_i - \sum_{(i,j)\in E} \log Z_{ij} + \sum_{i\in[n], j\in\partial i} \log Z_{i;j}$$

Proof.

$$G_\beta[\nu] \coloneqq -\sum_{(i,j)\in E} H(\nu_{ij}) + \sum_{i\in[n]} (|\partial i| - 1)\cdot H(\nu_i) + \mathbb{E}_{\boldsymbol{x}\sim\nu}[\mathcal{E}(\boldsymbol{x})]$$

$$\sum_{i\in[n]}\sum_{j\in\partial i} H(\nu_i) = \sum_{i\in[n]}\sum_{j\in\partial i} \mathbb{E}_{\nu_i}\left[\log\frac{1}{\nu_i(x_i)}\right] = \sum_{i\in[n]}\sum_{j\in\partial i} \mathbb{E}_{\nu_i}\left[\log\frac{Z_{i;j}}{m_{x_i}^{i\to j}\cdot \overline{m}_{x_i}^{j\to i}}\right]$$

$$= \sum_{i\in[n]}\sum_{j\in\partial i} \log Z_{i;j} - \sum_{i\in[n]}\sum_{j\in\partial i}\left(\mathbb{E}_{\nu_i}\left[\log m_{x_i}^{i\to j}\right] + \mathbb{E}_{\nu_i}\left[\log \overline{m}_{x_i}^{j\to i}\right]\right)$$

Belief Propagation

Lemma. Bethe free energy can be expressed as

$$G_\beta[\nu] = -\sum_{i\in[n]} \log Z_i - \sum_{(i,j)\in E} \log Z_{ij} + \sum_{i\in[n], j\in\partial i} \log Z_{i;j}$$

Proof.

$$\sum_{(i,j)\in E} \left(\mathbb{E}_{\nu_{ij}}\left[\log m_{x_j}^{j\to i}\right] + \mathbb{E}_{\nu_{ij}}\left[\log m_{x_i}^{i\to j}\right]\right) + \sum_{i\in[n], j\in\partial i} \mathbb{E}_{\nu_i}\left[\log \overline{m}_{x_i}^{j\to i}\right]$$

$$- \sum_{i\in[n]} \sum_{j\in\partial i} \left(\mathbb{E}_{\nu_i}\left[\log m_{x_i}^{i\to j}\right] + \mathbb{E}_{\nu_i}\left[\log \overline{m}_{x_i}^{j\to i}\right]\right)$$

$$= 0$$

■

- This lemma and its proof hold for any graph

Belief Propagation

Theorem. For any graph $G = (V, E)$, there is a 1—1 correspondence between

Fixed points of BP $\Leftrightarrow$ stationary points of G_β

Proof.

$$G_\beta[\nu] = -\sum_{i\in[n]} \log Z_i - \sum_{(i,j)\in E} \log Z_{ij} + \sum_{i\in[n], j\in\partial i} \log Z_{i;j}$$

$$Z_i := \sum_{\sigma\in\{\pm1\}} \prod_{j\in\partial i} \overline{m}_\sigma^{j\to i}, \qquad Z_{ij} := \sum_{\sigma_i,\sigma_j\in\{\pm1\}} \psi_{ij}(\sigma_i,\sigma_j) m_{\sigma_j}^{j\to i} m_{\sigma_i}^{i\to j}, \qquad Z_{i;j} := \sum_{\sigma\in\{\pm1\}} m_\sigma^{i\to j} \cdot \overline{m}_\sigma^{j\to i}$$

$$\frac{\partial G_\beta[\nu]}{\partial m_{\sigma_i}^{i\to j}} = -\sum_{(i,j)\in E} \frac{\sum_{\sigma_j\in\{\pm1\}} \psi_{ij}(\sigma_i,\sigma_j) m_{\sigma_j}^{j\to i}}{Z_{ij}} + \sum_{i\in[n]} \frac{\overline{m}_\sigma^{j\to i}}{Z_{i;j}} = 0 \iff \overline{m}_\sigma^{j\to i} \propto \sum_{\sigma_j\in\{\pm1\}} \psi_{ij}(\sigma_i,\sigma_j) m_{\sigma_j}^{j\to i}$$

Belief Propagation

Theorem. For any graph $G = (V, E)$, there is a 1—1 correspondence between

Fixed points of BP $\Leftrightarrow$ stationary points of G_β

Proof.

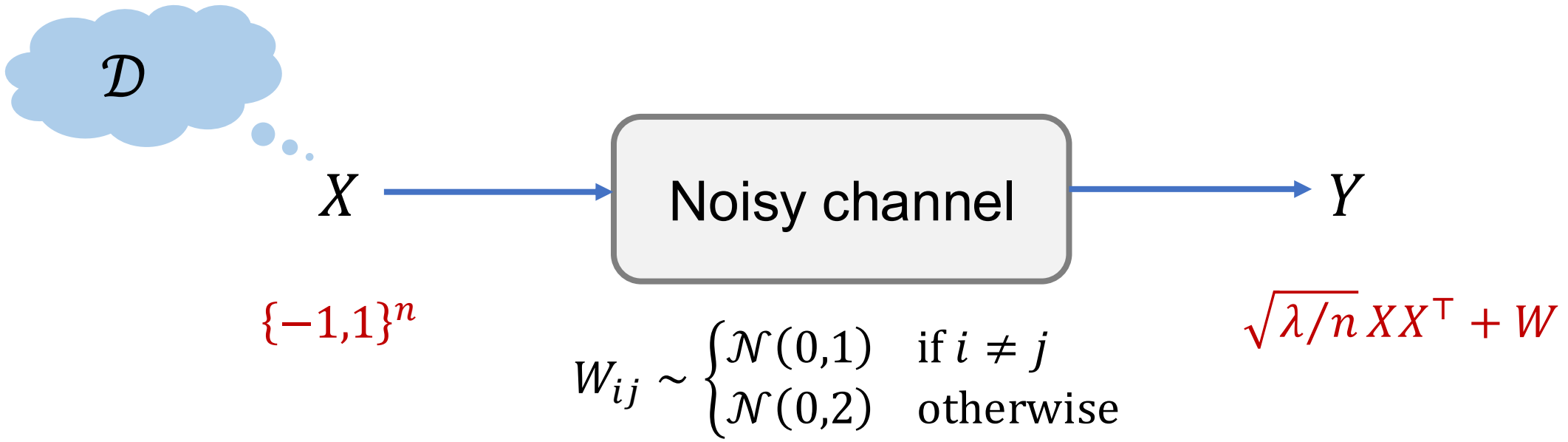
$$G_\beta[\nu] = -\sum_{i\in[n]} \log Z_i - \sum_{(i,j)\in E} \log Z_{ij} + \sum_{i\in[n], j\in\partial i} \log Z_{i;j}$$

$$Z_i := \sum_{\sigma\in\{\pm1\}} \prod_{j\in\partial i} \overline{m}_\sigma^{j\to i}, \qquad Z_{ij} := \sum_{\sigma_i,\sigma_j\in\{\pm1\}} \psi_{ij}(\sigma_i,\sigma_j) m_{\sigma_j}^{j\to i} m_{\sigma_i}^{i\to j}, \qquad Z_{i;j} := \sum_{\sigma\in\{\pm1\}} m_\sigma^{i\to j} \cdot \overline{m}_\sigma^{j\to i}$$

$$\frac{\partial G_\beta[\nu]}{\partial \overline{m}_\sigma^{j\to i}} = -\sum_{i\in[n]} \frac{\prod_{k\in\partial i\setminus\{j\}} \overline{m}_\sigma^{k\to i}}{Z_i} + \sum_{i\in[n]} \frac{m_\sigma^{i\to j}}{Z_{i;j}} = 0 \iff m_\sigma^{i\to j} \propto \prod_{k\in\partial i\setminus\{j\}} \overline{m}_\sigma^{k\to i}$$

■

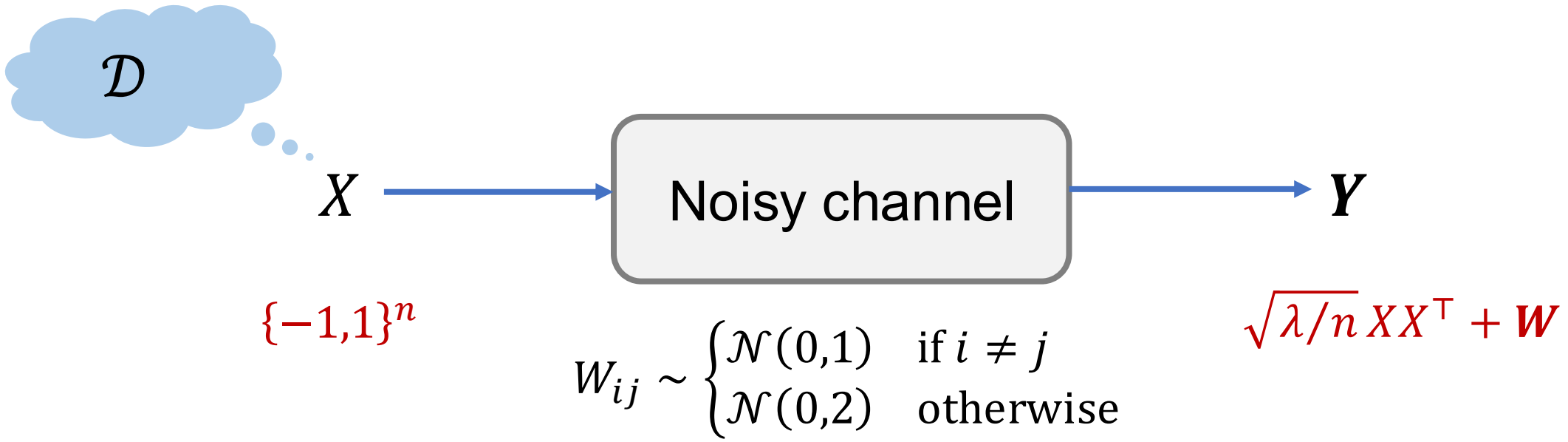
Application: $\mathbb{Z}_2$ Synchronization



Observing $\boldsymbol{Y}$, how can you infer X?

- **Goal:** Given $\boldsymbol{Y}$, estimate $\mathbb{E}[X|\boldsymbol{Y}]$ (?)
- By the symmetry between X and $-X$, $\mathbb{E}[X|\boldsymbol{Y}] = 0$. We can break the symmetry by forcing e.g., $X_1 = 1$

Application: $\mathbb{Z}_2$ Synchronization



$$W_{ij} \sim \begin{cases} \mathcal{N}(0,1) & \text{if } i \neq j \\ \mathcal{N}(0,2) & \text{otherwise} \end{cases}$$

Observing $\boldsymbol{Y}$, how can you infer X?

- **Goal:** Given $\boldsymbol{Y}$, find a denoiser $\hat{X}(\boldsymbol{Y})$ that minimizes

$$\mathrm{MSE}(\widehat{\mathrm{X}}) \coloneqq \frac{1}{n^2}\,\mathbb{E}_{X,\mathbf{Y}}\left[\left\|\hat{X}(\boldsymbol{Y})\hat{X}(\boldsymbol{Y})^\top - XX^\top\right\|_F^2\right]$$

- **Baseline:** $\mathrm{MSE}(\mathbf{0}) = 1$

Application: $\mathbb{Z}_2$ Synchronization

$$\boldsymbol{Y} = \sqrt{\lambda/n}\, XX^\top + \boldsymbol{W}, \qquad W_{ij} \sim \begin{cases} \mathcal{N}(0,1) & \text{if } i \neq j \\ \mathcal{N}(0,2) & \text{otherwise} \end{cases}$$

Spectral method

- Compute the top eigenvector v of $\boldsymbol{Y}$ and hope that v is strongly correlated with X
- Let $u \coloneqq X/\sqrt{n}$ so $\|u\| = 1$, and let $\boldsymbol{A} \coloneqq \boldsymbol{Y}/\sqrt{n} = \sqrt{\lambda} uu^\top + \boldsymbol{W}/\sqrt{n}$
- If there is no noise, then $\boldsymbol{A}$ is rank-1 with the only eigenvalue $\sqrt{\lambda}$ and eigenvector $u \propto X$
- How much can this eigenvector survive after adding $\boldsymbol{W}/\sqrt{n}$

Application: $\mathbb{Z}_2$ Synchronization

$$\boldsymbol{Y} = \sqrt{\lambda/n}\, XX^\top + \boldsymbol{W}, \qquad W_{ij} \sim \begin{cases} \mathcal{N}(0,1) & \text{if } i \neq j \\ \mathcal{N}(0,2) & \text{otherwise} \end{cases}$$

Spectral method

- Compute the top eigenvector v of $\boldsymbol{Y}$ and hope that v is strongly correlated with X
- Let $u \coloneqq X/\sqrt{n}$ so $\|u\| = 1$, and let $\boldsymbol{A} \coloneqq \boldsymbol{Y}/\sqrt{n} = \sqrt{\lambda} uu^\top + \boldsymbol{W}/\sqrt{n}$

Theorem (Baik-Ben Arous-Péché (BBP) transition).

The top eigenvalue and the top eigenvector of A satisfy

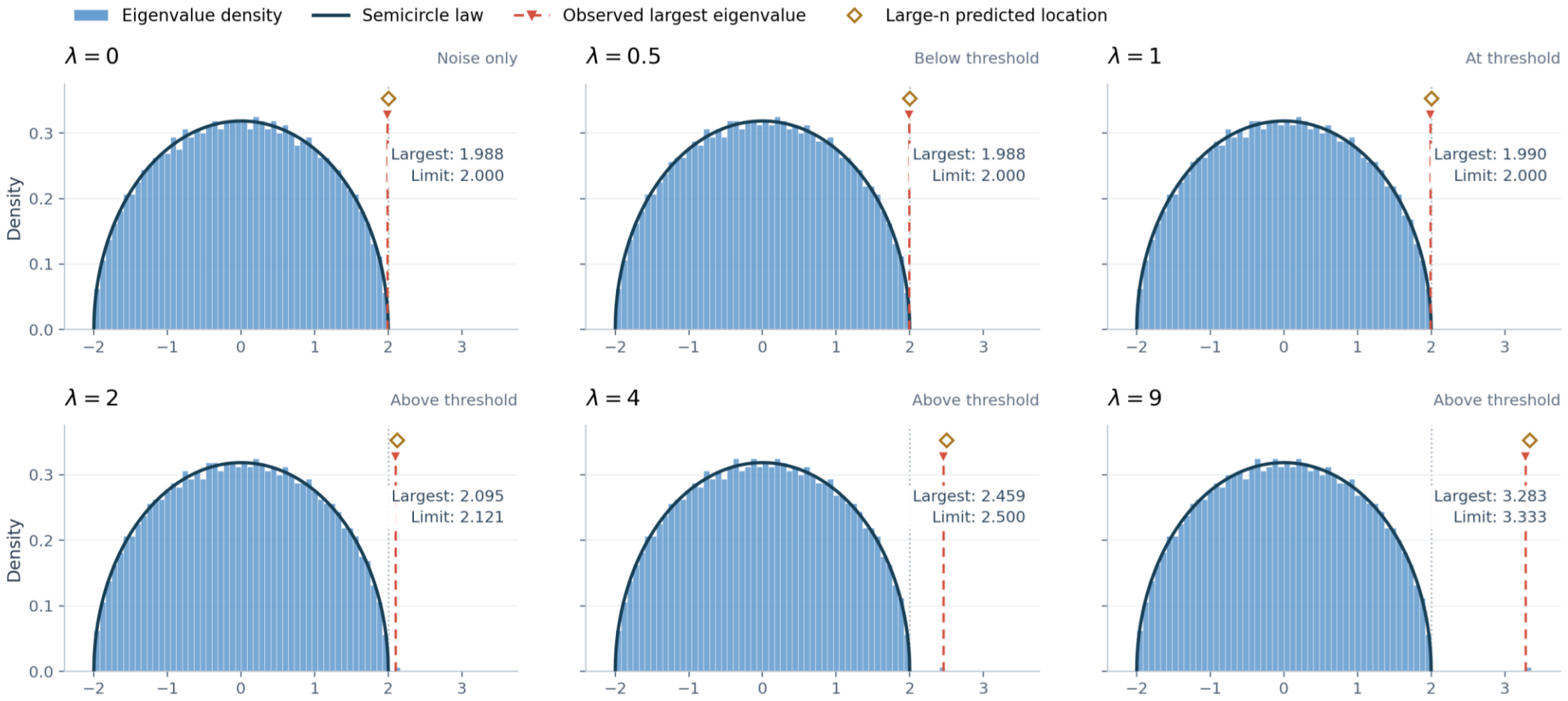
$$\lambda_1(\boldsymbol{A}) \to \begin{cases} 2 & \text{if } \lambda \leq \lambda_{\mathrm{BBP}} \coloneqq 1 \\ \sqrt{\lambda} + 1/\sqrt{\lambda} & \text{if } \lambda > \lambda_{\mathrm{BBP}} \end{cases}, \qquad |\langle u, v \rangle|^2 \to \begin{cases} 0 & \text{if } \lambda \leq \lambda_{\mathrm{BBP}} \\ 1 - 1/\lambda & \text{if } \lambda > \lambda_{\mathrm{BBP}} \end{cases}$$

as $n \to \infty$

Application: $\mathbb{Z}_2$ Synchronization

Histograms of eigenvalues of $A = \sqrt{\lambda}uu^\top + \boldsymbol{W}/\sqrt{n}$ with different λ

Wigner's semi-circle law



Application: $\mathbb{Z}_2$ Synchronization

Stronger baseline, but does not incorporate the prior on X

Spectral method

Theorem (Baik-Ben Arous-Péché (BBP) transition).

The top eigenvalue and the top eigenvector of A satisfy

$$\lambda_1(\boldsymbol{A}) \to \begin{cases} 2 & \text{if } \lambda \le \lambda_{\mathrm{BBP}} \coloneqq 1 \\ \sqrt{\lambda} + 1/\sqrt{\lambda} & \text{if } \lambda > \lambda_{\mathrm{BBP}} \end{cases}, \qquad |\langle u, v\rangle|^2 \to \begin{cases} 0 & \text{if } \lambda \le \lambda_{\mathrm{BBP}} \\ 1 - 1/\lambda & \text{if } \lambda > \lambda_{\mathrm{BBP}} \end{cases}$$

as $n \to \infty$

- We can take $\hat{X}(Y) \coloneqq \sqrt{n\alpha}v$ for some $\alpha > 0$ to be determined

$$\frac{1}{n^2}\left\|\hat{X}(\boldsymbol{Y})\hat{X}(\boldsymbol{Y})^{\mathsf{T}} - XX^{\mathsf{T}}\right\|_F^2 = \|\alpha vv^{\mathsf{T}} - uu^{\mathsf{T}}\|_F^2 = 1 + \alpha^2 - 2\alpha|\langle u, v\rangle|^2$$

- If $\lambda > 1$, we can take $\alpha \coloneqq 1 - 1/\lambda$, so that $\mathrm{MSE}(\widehat{\mathrm{X}}) \to 1 - (1 - 1/\lambda)^2 < 1$
- If $\lambda \le 1$, the best choice is to take $\alpha \coloneqq 0$

Application: $\mathbb{Z}_2$ Synchronization

$$\boldsymbol{Y} = \sqrt{\lambda/n}\, XX^\top + \boldsymbol{W}, \qquad W_{ij} \sim \begin{cases} \mathcal{N}(0,1) & \text{if } i \neq j \\ \mathcal{N}(0,2) & \text{otherwise} \end{cases}$$

- Suppose X is sampled uniformly from $\{-1,1\}^n$
- Compute the Gibbs measure of the posterior $\mu(x) \coloneqq \Pr[X = x|\boldsymbol{Y}]$:

$$\mu(x) = \frac{\Pr[\boldsymbol{Y}|X = x]\Pr[X = x]}{\Pr[\boldsymbol{Y}]} \propto \Pr[\boldsymbol{Y}|X = x] \qquad \text{Bayes rule}$$

$$\propto \exp\left(-\frac{1}{4}\left\|\boldsymbol{Y} - \sqrt{\lambda/n}\, xx^\top\right\|_F^2\right)$$

$$= \exp\left(-\frac{1}{4}\|\boldsymbol{Y}\|_F^2 + \frac{\sqrt{\lambda}}{2\sqrt{n}} x^\top \boldsymbol{Y} x - \frac{\lambda}{4} n\right)$$

$$\propto \exp\left(\frac{\sqrt{\lambda}}{2\sqrt{n}} x^\top \boldsymbol{Y} x\right) \qquad \text{An Ising model!}$$

Application: $\mathbb{Z}_2$ Synchronization

$$\boldsymbol{Y} = \sqrt{\lambda/n}\, XX^\top + \boldsymbol{W}, \qquad W_{ij} \sim \begin{cases} \mathcal{N}(0,1) & \text{if } i \neq j \\ \mathcal{N}(0,2) & \text{otherwise} \end{cases}$$

- Suppose X is sampled uniformly from $\{-1,1\}^n$
- Compute the Gibbs measure of the posterior $\mu(x) \coloneqq \Pr[X = x|\boldsymbol{Y}]$:

$$\mu(x) \propto \exp\left(\frac{1}{2} x^\top \boldsymbol{A} x\right), \qquad \boldsymbol{A} \coloneqq \sqrt{\lambda/n}\, \boldsymbol{Y}$$

- Compatibility function: $\psi_{ij}(x_i, x_j) = \exp(A_{ij} x_i x_j)$ for any $i < j$
- Belief propagation update rule:

$$m_\sigma^{i\to k}[t+1] \propto \prod_{j \in \partial i \setminus \{k\}} \sum_{s \in \{\pm 1\}} \psi_{ij}(\sigma, s) m_s^{j\to i}[t] = \prod_{j \neq k} \sum_{s \in \{\pm 1\}} \exp(A_{ij}\sigma s)\, m_s^{j\to i}[t]$$

Application: $\mathbb{Z}_2$ Synchronization

$$m_\sigma^{i\to k}[t+1] \propto \prod_{j\neq k} \sum_{s\in\{\pm 1\}} \exp\left(A_{ij}\sigma s\right) m_s^{j\to i}[t]$$

- We can use the messages to express the marginal expectations:

$$\hat{x}_t^{j\to i} := m_+^{j\to i}[t] - m_-^{j\to i}[t]$$

$$m_s^{j\to i}[t] = \frac{1 + s\hat{x}_t^{j\to i}}{2}$$

$$\sum_{s\in\{\pm 1\}} \exp\left(A_{ij}\sigma s\right) m_s^{j\to i}[t] = \frac{\exp\left(A_{ij}\sigma\right) + \exp\left(-A_{ij}\sigma\right)}{2} + \frac{\hat{x}_t^{j\to i}}{2}\left(\exp\left(A_{ij}\sigma\right) - \exp\left(-A_{ij}\sigma\right)\right)$$

$$= \frac{\exp\left(A_{ij}\sigma\right) + \exp\left(-A_{ij}\sigma\right)}{2}\left(1 + \tanh\left(A_{ij}\sigma\right)\hat{x}_t^{j\to i}\right)$$

$$\propto 1 + \tanh\left(A_{ij}\sigma\right)\hat{x}_t^{j\to i} = 1 + \sigma\tanh\left(A_{ij}\right)\hat{x}_t^{j\to i}$$

$$\tanh(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}}$$

Application: $\mathbb{Z}_2$ Synchronization

$$m_\sigma^{i\to k}[t+1] \propto \prod_{j\neq k}\left(1+\sigma\tanh\left(A_{ij}\right)\hat{x}_t^{j\to i}\right)$$

$$\begin{aligned}\hat{x}_{t+1}^{i\to k} &= m_+^{i\to k}[t+1] - m_-^{i\to k}[t+1] \\ &= \frac{\prod_{j\neq k}\left(1+\tanh\left(A_{ij}\right)\hat{x}_t^{j\to i}\right) - \prod_{j\neq k}\left(1-\tanh\left(A_{ij}\right)\hat{x}_t^{j\to i}\right)}{\prod_{j\neq k}\left(1+\tanh\left(A_{ij}\right)\hat{x}_t^{j\to i}\right) + \prod_{j\neq k}\left(1-\tanh\left(A_{ij}\right)\hat{x}_t^{j\to i}\right)} \\ &\approx \frac{\prod_{j\neq k}\left(1+\tanh\left(A_{ij}\right)\hat{x}_t^{j\to i}\right) - \prod_{j\neq k}\left(1-\tanh\left(A_{ij}\right)\hat{x}_t^{j\to i}\right)}{\exp\left(\sum_{j\neq k}\tanh\left(A_{ij}\right)\hat{x}_t^{j\to i}\right) + \exp\left(-\sum_{j\neq k}\tanh\left(A_{ij}\right)\hat{x}_t^{j\to i}\right)}\end{aligned}$$

- $\boldsymbol{A} \coloneqq \sqrt{\lambda/n}\,\boldsymbol{Y} = (\lambda/n)XX^\top + \sqrt{\lambda/n}\,\boldsymbol{W}$, so that $A_{ij} = O(1/\sqrt{n})$ and $\tanh\left(A_{ij}\right) \approx A_{ij}$

Application: $\mathbb{Z}_2$ Synchronization

$$m_\sigma^{i\to k}[t+1] \propto \prod_{j\neq k}\left(1+\sigma\tanh\left(A_{ij}\right)\hat{x}_t^{j\to i}\right)$$

$$\begin{aligned}
\hat{x}_{t+1}^{i\to k} &= m_+^{i\to k}[t+1] - m_-^{i\to k}[t+1] \\
&\approx \frac{\prod_{j\neq k}\left(1+\tanh\left(A_{ij}\right)\hat{x}_t^{j\to i}\right) - \prod_{j\neq k}\left(1-\tanh\left(A_{ij}\right)\hat{x}_t^{j\to i}\right)}{\exp\left(\sum_{j\neq k}\tanh\left(A_{ij}\right)\hat{x}_t^{j\to i}\right) + \exp\left(-\sum_{j\neq k}\tanh\left(A_{ij}\right)\hat{x}_t^{j\to i}\right)} \\
&\approx \frac{\prod_{j\neq k}\left(1+\tanh\left(A_{ij}\right)\hat{x}_t^{j\to i}\right) - \prod_{j\neq k}\left(1-\tanh\left(A_{ij}\right)\hat{x}_t^{j\to i}\right)}{\exp\left(\sum_{j\neq k}A_{ij}\hat{x}_t^{j\to i}\right) + \exp\left(-\sum_{j\neq k}A_{ij}\hat{x}_t^{j\to i}\right)} \\
&\approx \tanh\left(\sum_{j\neq k}A_{ij}\hat{x}_t^{j\to i}\right)
\end{aligned}$$

Application: $\mathbb{Z}_2$ Synchronization

$$\hat{x}_{t+1}^{i\to k} \approx \tanh\left(\sum_{j\neq k} A_{ij}\hat{x}_t^{j\to i}\right) = \tanh\left(\sqrt{\lambda}\underbrace{\sum_{j\neq k}\frac{1}{\sqrt{n}}Y_{ij}\hat{x}_t^{j\to i}}_{x_{t+1}^{i\to k}}\right)$$

$$\hat{x}_{t+1}^{i\to k} := \tanh\left(\sqrt{\lambda}x_{t+1}^{i\to k}\right)$$

$$x_{t+1}^{i\to k} := \sum_{j\neq k}\frac{1}{\sqrt{n}}Y_{ij}\hat{x}_t^{j\to i} = \sum_{j\neq k}\frac{1}{\sqrt{n}}Y_{ij}\underbrace{\tanh\left(\sqrt{\lambda}x_t^{j\to i}\right)}_{f_t\left(x_t^{j\to i}\right)}$$

Approximate Message Passing:

$$x_{t+1}^{i\to k} := \sum_{j\neq k}\frac{1}{\sqrt{n}}Y_{ij}f_t\left(x_t^{j\to i}\right)$$

Approximate Message Passing

$$x_{t+1}^{i\to k} := \sum_{j\neq k} \frac{1}{\sqrt{n}} Y_{ij} f_t\left(x_t^{j\to i}\right), \qquad \hat{x}_{t+1}^{i\to k} := f_{t+1}\left(x_{t+1}^{i\to k}\right)$$

- Intuitively, $x_t^{i\to k}$ is a score of evidence about X_i collected from the neighbors of i except k
- $\hat{x}_t^{i\to k}$ is a soft estimate of $X_i \in \{-1,1\}$

$$\text{evidence score} \xrightarrow{f_t} \text{soft estimate}$$

$O(n^2)$ messages in total. Can we reduce the number of messages we need to track?

Attempt 1: All of the message out of a given vertex i are close to each other, so we could try just tracking n message

Mean-field approximation:

$$x_{t+1}^{i} := \sum_{j\in[n]} \frac{1}{\sqrt{n}} Y_{ij} f_t\left(x_t^{j}\right), \qquad x_{t+1} := \frac{1}{\sqrt{n}} Y f_t(x_t)$$

f_t applied entrywise

Detour: Mean-Field Approximation

$$\mu(x) \propto \exp\left(\frac{1}{2} x^\top \boldsymbol{A} x\right)$$

- Fix any $i \in [n]$. The terms in the exponent involving x_i is

Local field on i

$$x_i \sum_{j \neq i} A_{ij} x_j = h_i x_i, \qquad h_i \coloneqq \sum_{j \neq i} A_{ij} x_j$$

$$\Pr[X_i = x_i | x_{-i}, \boldsymbol{Y}] = \frac{e^{h_i x_i}}{e^{h_i} + e^{-h_i}}, \qquad \mathbb{E}[X_i | x_{-i}, \boldsymbol{Y}] = \tanh(h_i)$$

- In the mean-field approximation, we replace x_j by the estimated mean m_j in the local field:

$$X_j \rightsquigarrow m_j, \qquad h_i \rightsquigarrow \sum_{j \neq i} A_{ij} m_j$$

$$m_j = \tanh(h_i) = \tanh\left(\sum_{j \neq i} A_{ij} m_j\right)$$

Detour: Mean-Field Approximation

$$\mu(x) \propto \exp\left(\frac{1}{2} x^\top \boldsymbol{A} x\right)$$

- Another way to understand the mean-field approximation is via variational inference:

$$\mu(x) \rightsquigarrow \nu_{\text{mean}}(x) \coloneqq \prod_{i \in [n]} \nu_i(x_i)\,, \qquad \nu_i(x_i) = \frac{1 + m_i x_i}{2}$$

- We choose m_i by minimizing $\mathrm{KL}(\nu_{\text{mean}} \| \mu)$:

$$\mathrm{KL}(\nu_{\text{mean}} \| \mu) = -\sum_{i \in [n]} H(\nu_i) - \sum_{i<j} A_{ij} m_i m_j + \log Z$$

$$\frac{\partial}{\partial m_i} \mathrm{KL}(\nu_{\text{mean}} \| \mu) = \arctan(m_i) - \sum_{j \neq i} A_{ij} m_j = 0$$

$$m_j = \tanh\left(\sum_{j \neq i} A_{ij} m_j\right)$$

Approximate Message Passing

$$x_{t+1}^{i\to k} := \sum_{j\neq k} \frac{1}{\sqrt{n}} Y_{ij} f_t\left(x_t^{j\to i}\right), \qquad \hat{x}_{t+1}^{i\to k} := f_{t+1}\left(x_{t+1}^{i\to k}\right)$$

$O(n^2)$ messages in total. Can we reduce the number of messages we need to track?

Attempt 2: All the messages out of a given vertex i are close to each other, but these fluctuations are amplified non-trivially after BP iterations, leading to a crucial correction to the mean-field approximation.

$$x_{t+1}^{i\to k} := \sum_{j\neq k} \frac{1}{\sqrt{n}} Y_{ij} f_t\left(x_t^{j\to i}\right) = \sum_{j\in[n]} \frac{1}{\sqrt{n}} Y_{ij} f_t\left(x_t^{j\to i}\right) - \frac{1}{\sqrt{n}} Y_{ik} f_t\left(x_t^{k\to i}\right)$$

x_{t+1}^{i} $\qquad$ $\delta_{t+1}^{i\to k} = O(1/\sqrt{n})$



Approximate Message Passing

$$x_{t+1}^{i\to k} = x_{t+1}^{i} - \delta_{t+1}^{i\to k}, \qquad x_{t+1}^{i} := \sum_{j\in[n]} \frac{1}{\sqrt{n}} Y_{ij} f_t\left(x_t^{j\to i}\right), \qquad \delta_{t+1}^{i\to k} := \frac{1}{\sqrt{n}} Y_{ik} f_t\left(x_t^{k\to i}\right)$$

$$x_{t+1}^{i\to k} = x_{t+1}^{i} \pm O(1/\sqrt{n})$$

$$\delta_{t+1}^{i\to k} = \frac{1}{\sqrt{n}} Y_{ik} f_t\left(x_t^{k} \pm O(1/\sqrt{n})\right)$$

$$= \frac{1}{\sqrt{n}} Y_{ik} f_t\left(x_t^{k}\right) \pm O(1/n)$$

$f_t(x) = \tanh\left(\sqrt{\lambda}x\right)$ is Lipschitz

$$x_{t+1}^{i} := \sum_{j\in[n]} \frac{1}{\sqrt{n}} Y_{ij} f_t\left(x_t^{j} - \delta_t^{j\to i}\right) = \sum_{j\in[n]} \frac{1}{\sqrt{n}} Y_{ij}\left(f_t\left(x_t^{j}\right) - f_t'\left(x_t^{j}\right)\delta_t^{j\to i}\right) \pm O(1/\sqrt{n})$$

Approximate Message Passing

$$\delta_{t+1}^{i\to k} = \frac{1}{\sqrt{n}} Y_{ik} f_t\left(x_t^k\right) \pm O(1/n)$$

$$x_{t+1}^i = \sum_{j\in[n]} \frac{1}{\sqrt{n}} Y_{ij}\left(f_t\left(x_t^j\right) - f_t'\left(x_t^j\right)\delta_t^{j\to i}\right) \pm O(1/\sqrt{n})$$

$$= \sum_{j\in[n]} \frac{1}{\sqrt{n}} Y_{ij} f_t\left(x_t^j\right) - f_{t-1}\left(x_{t-1}^i\right) \underbrace{\sum_{j\in[n]} \frac{1}{n} Y_{ij}^2 f_t'\left(x_t^j\right)}_{\approx \frac{1}{n}\sum_{j\in[n]} f_t'\left(x_t^j\right) =: b_t} \pm O(1/\sqrt{n})$$

$$x_{t+1}^i = \sum_{j\in[n]} \frac{1}{\sqrt{n}} Y_{ij} f_t\left(x_t^j\right) - f_{t-1}\left(x_{t-1}^i\right) b_t, \qquad x_{t+1} = \frac{1}{\sqrt{n}} Y f_t(x_t) - f_{t-1}(x_{t-1}) b_t$$

Approximate Message Passing

mean-field Onsager correction

$$x_{t+1} = \frac{1}{\sqrt{n}} Y f_t(x_t) - f_{t-1}(x_{t-1}) b_t$$

$$b_t = \frac{1}{n} \sum_{j \in [n]} f_t'(x_t^j)$$

$$\hat{x}_{t+1} = f_{t+1}(x_{t+1})$$

State evolution: as $n \to \infty$, the behavior of the iterates of AMP is precisely captured by a certain distributional recursion

Approximate Message Passing

State evolution: as $n \to \infty$, the behavior of the iterates of AMP is precisely captured by a certain distributional recursion

- Consider a very simplified setting:

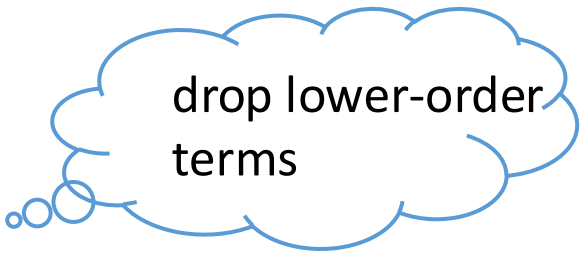
$$x_t \sim \mathcal{N}(\mu_t X, \sigma_t^2 I)$$

  Gaussian is not important here; we focus on tracking the means and variances

- We ignore the Onsager correction term, and assume that Y is resampled in each iteration

$$x_{t+1} = \frac{1}{\sqrt{n}}\left(\frac{\sqrt{\lambda}}{\sqrt{n}} XX^\top + W\right) f_t(x_t)$$

$$\sim \mathcal{N}\left(\frac{\sqrt{\lambda}}{n}\langle X, f_t(x_t)\rangle X, \frac{\|f_t(x_t)\|^2}{n} I\right)$$



Approximate Message Passing

$$x_{t+1} = \frac{1}{\sqrt{n}}\left(\frac{\sqrt{\lambda}}{\sqrt{n}}XX^\top + W\right) f_t(x_t)$$

$$\sim \mathcal{N}\left(\frac{\sqrt{\lambda}}{n}\langle X, f_t(x_t)\rangle X, \frac{\|f_t(x_t)\|^2}{n} I\right)$$

$$\frac{\sqrt{\lambda}}{n}\langle X, f_t(x_t)\rangle = \sqrt{\lambda}\frac{1}{n}\sum_{i=1}^{n} X_i f_t(\mu_t X_i + \sigma_t g_i), \qquad g_i \sim \mathcal{N}(0,1)$$

$$\approx \underbrace{\sqrt{\lambda}\mathbb{E}[x f_t(\mu_t x + \sigma_t g_i)]}_{\mu_{t+1}}, \qquad x \sim \{\pm 1\}, g_i \sim \mathcal{N}(0,1)$$

$$\frac{\|f_t(x_t)\|^2}{n} = \frac{1}{n}\sum_{i=1}^{n} f_t(\mu_t X_i + \sigma_t g_i)^2 \approx \underbrace{\mathbb{E}[f_t(\mu_t x + \sigma_t g_i)^2]}_{\sigma_{t+1}^2}, \qquad x \sim \{\pm 1\}, g_i \sim \mathcal{N}(0,1)$$

Approximate Message Passing

State evolution recursion:

$$x_t \sim \mathcal{N}(\mu_t X, \sigma_t^2 I)$$

$$\mu_{t+1} = \sqrt{\lambda}\mathbb{E}[x f_t(\mu_t x + \sigma_t g_i)], \qquad \sigma_{t+1}^2 = \mathbb{E}[f_t(\mu_t x + \sigma_t g_i)^2]$$

$$\mu_0 = \sqrt{\lambda}\mathbb{E}\left[x f_0\left(x^{(0)}\right)\right], \qquad \sigma_0^2 = \mathbb{E}\left[f_0\left(x^{(0)}\right)^2\right]$$

where $x^{(0)}$ is the random variable whose law is the limit of the empirical distribution of x_0 as $n \to \infty$

- The Onsager term makes this heuristic derivation rigorous!

Theorem (Bayati-Montanari '11**).** If the f_t's are Lipschitz, then for any "nice" test function $\psi : \mathbb{R}^2 \to \mathbb{R}$ and any t,

$$\lim_{n\to\infty} \frac{1}{n}\sum_i \psi\left(x_t^i, X_i\right) = \mathbb{E}[\psi(\mu_t x + \sigma_t g, x)]$$