CS 2243 Fall 2024 Scribes: <names>
11/11 Based on notes by: Oliver Cheng

Lecture 19: Intro to Approximate Message Passing

1 Recap

We talked about Belief Propagation (BP) last time. We managed to do well on trees and get exact marginals, but passing from trees to general non-cyclic graphs is quite challenging.

These notes will on the study of Approximate Message Passing (APM). This is a BP-inspired algorithm, which is conjectured to be optimal for a wide range of inference problems. We will focus on a specific application to *denoising low-rank matrices*.

2 $\mathbb{Z}_2$ Synchronization

Consider a regime with a hidden Boolean vector $X \sim \{\pm 1\}^n$ where we receive a noisy rank-1 matrix $Y = \sqrt{\frac{\lambda}{n}} XX^\top + W$ where the noise W is defined as

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

We choose λ such that Y has operator norm approximately $\sqrt{n}$. Our goal is to estimate $\mathbb{E}[X|Y]$[1]. Another way to say this is as follows: given Y, find a a denoiser $\hat{X}(Y)$ minimizing

$$\text{MSE}(\hat{X}) \triangleq \frac{1}{n^2} \mathbb{E}_{X,Y} \|\hat{X}(Y)\hat{X}(Y)^\top - XX^\top\|_F^2.$$

As a baseline, the trivial estimate ($\hat{X}(Y) = 0$) achieves MSE = 1.

2.1 Spectral Method Baseline

Naively, we could consider taking the top eigenvector v of Y.

Theorem 1 (Baik-Arous-Peche '04). *"BBP transition" - the top eigenvalue of Y escapes from "bulk" when $\lambda > 1$ [BBAP05].*

When $\lambda = 0$, the histogram of eigenvalues of $\frac{1}{\sqrt{n}} Y$ form a semi-circle from Wigner semicircle law, but as $\lambda \gg 1$ the "top eigenvalue" escapes this semi-circular bulk. Indeed, analytically one can find that

$$\frac{1}{\sqrt{n}} \lambda_1(Y) \to \begin{cases} \lambda + \frac{1}{\lambda} & \lambda > 0 \\ 2 & \text{if } \lambda \leq 1 \end{cases}$$

$$\frac{1}{\sqrt{n}} \cos \angle(X, v) \to \begin{cases} .\sqrt{1 - 1/\lambda^2} & \text{if } \lambda > 1 \\ 0 & \text{if } \lambda \leq 1 \end{cases}$$

The issue here is that the algorithm does not incorporate the prior on X. e.g. when the prior is Gaussian, this algorithm is optimal, but this does not give the right answer when X has the discrete structure as in our problem.

[1] By symmetry, $\mathbb{E}[X|Y] = 0$, but we can break symmetry e.g. by conditioning on $X_1 = 1$, and the algorithms we consider naturally break this symmetry.

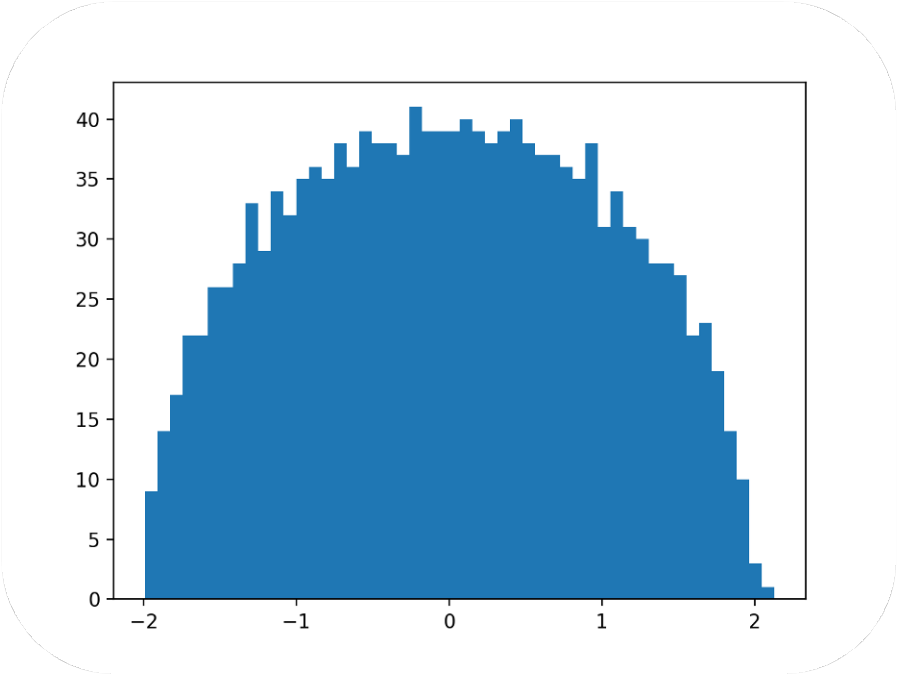

Figure 1: (Wigner semi-circle law) When $\lambda \gg 1$ The majority of eigenvalues are within this semicircular bulk, but the top eigenvalue of Y on the right side begins to escape this semi-circle.Caption

2.2 Belief Propagation

For BP, we need some Gibbs measure. Let this be the posterior $\mu(x) \triangleq \Pr[X = x|Y]$ by Bayes and the fact X is uniform over the Boolean hypercube, we have

$$\begin{aligned}\mu(x) &\propto \Pr[Y|X = x] \\ &\propto \exp\left(-\frac{1}{4}\left\|Y - \sqrt{\frac{\lambda}{n}}xx^\top\right\|_F^2\right) \\ &= \exp\left(\frac{1}{2}\sqrt{\frac{\lambda}{n}}x^\top Yx\right)\end{aligned}$$

This form looks familiar–it is simply an Ising model with a random interaction matrix. As λ goes to ∞ (temperature goes to 0), the maximizer is simply the underlying spike that arises when Y becomes close to $xx^\top$.

Recall the compatibility functions in this regime; for $i < j$

$$\Psi_{ij}(x_i, x_j) = \exp(A_{ij}x_ix_j) \quad A \triangleq \sqrt{\frac{\lambda}{n}}Y$$

Every entry of A is $O(\frac{1}{\sqrt{n}})$. Belief propagation update rules:

$$m_\sigma^{(i)\to k}[t+1] \propto \prod_{j\neq k}\sum_{s\in\{\pm1\}} m_s^{(j)\to i}[t] \cdot \exp(A_{ij}\sigma s)$$

See the previous lecture on BP for a review on what the messages mean. Note that generally we take this product over $j \in \partial j \backslash k$, but since A has values for every entry, this is simply all nodes except k.

Instead of marginal probabilities, we can equivalently parameterize this value in terms of marginal expectations:

$$-m_-^{(j)\to i}[t] \implies m_s^{(j)\to i}[t] = \frac{1}{2} + s \cdot \frac{\hat{x}_t^{(j)\to i}}{2}$$

Thus our update rule can be modified as follows

$$\begin{aligned} m_{\sigma}^{\textcircled{i}\to k}[t+1] &\propto \sum_{s\in\{\pm 1\}} m_{s}^{\textcircled{j}\to i}[t]\cdot \exp(A_{ij}\sigma_s) \\ &= \frac{e^{A_{ij}\sigma}+e^{-Aij\sigma}}{2}\left(\exp(A_{ij}\sigma)-\exp(A_{ij}\sigma)\right) \\ &= \frac{e^{A_{ij}\sigma}+e^{-Aij\sigma}}{2}\left(1+\tanh(A_{ij}\sigma)\hat{x}_t^{\textcircled{j}\to i}\right) \end{aligned}$$

In order to compute the marginal expectations, we have

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

Where the second to last step comes from a a linear approximation since tanh is linear close to 0. Define

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

Finally, from here we can substitute in the definition for A.

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

We are using tanh since this appeared naturally as we derived the formulae, but our algorithm will be valid for any nonlinearity. Let this nonlinearity be denoted $f_t(x_t^{\textcircled{j}\to i}) \triangleq \tanh(\sqrt{\lambda}x_t^{\textcircled{j}\to i})$.

We have a total of $O(n^2)$ total messages. The main question is whether we can 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 messages, one per vertex. Let's also replace the $\sum_{j\neq k}\frac{1}{\sqrt{n}}Y_{ij}f_t(x_t^{\textcircled{j}\to i})x_{t+1}^{\textcircled{i}\to k}$ with $\sum_{j=1}^{k}\frac{1}{\sqrt{n}}Y_{ij}f_t(x_t^{\textcircled{j}\to i})x_{t+1}^{\textcircled{i}\to k}$ to make life easier. i.e.

$$x_{t+1}^i \triangleq \sum_{j=1}^{n}\frac{1}{\sqrt{n}}Y_{ij}f_t(x_t^j) \implies x_{t+1} \triangleq \frac{1}{\sqrt{n}}Yf_t(x_t)$$

where $f_t(x_t)$ applies f_t to each entry of x_t.

This is unfortunately an insufficient attempt. This approximation is analagous to recursively applying the linear transformation $\frac{1}{\sqrt{n}}Y$ to $f_t(x_t)$. This can be thought of as a "mean-field approximation" because we are replacing the complex interaction that happens at each node with simply an average of the messages that are coming out of the given node.

Attempt 2: All of the messages out of a given vertex i are close to each other, but these fluctuations get amplified nontrivially after BP iteration, leading to a crucial correction to the mean field approximation. Our goal

is still similar though, where we want to simplify the n^2 messages into one for each vertex. Consider rewriting the sum

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

where

$$x_{t+1}^i \triangleq \sum_{j=1}^n \frac{1}{\sqrt{n}} Y_{ij} f_t(x_t^{\textcircled{j}\to i}), \quad \delta_{t+1}^{\textcircled{i}\to k} \triangleq \frac{1}{\sqrt{n}} U_{ik} f_t(x_t^{\textcircled{k}\to i}) = O(\frac{1}{\sqrt{n}})$$

Immediately, we have

$$x_{t+1}^i = \sum_{j=1}^n \frac{1}{\sqrt{n}} Y_{ij} f_t(x_t^{\textcircled{j}\to i}) \pm O(1/\sqrt{n})$$

Furthermore, since f_t is tanh and hence Lipschitz, we can perturb $x_t^{\textcircled{k}\to i}$ to be the average message x_t^k and obtain

$$\begin{aligned}\delta_{t+1}^{\textcircled{i}\to k} &= \frac{1}{\sqrt{n}} Y_{ik} f_t(x_t^{\textcircled{k}\to i}) \\ &= \frac{1}{\sqrt{n}} Y_{ik} f_t(x_t^k) \pm O(1/\sqrt{n})\end{aligned}$$

This allows us to write x_{t+1}^i as follows.

$$x_{t+1}^i = \sum_{j=1}^n \frac{1}{\sqrt{n}} Y_{ij} f_t(x_t^j - \delta_t^{\textcircled{j}\to i}) \pm O(1/\sqrt{n})$$

If we Taylor Expand the inside of f_t around x_t^j, we get

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

Since δ is equal to the value of the previous message $f_{t-1}(x_{t-1}^i)$, we can write this by what we wrote above. And we get

$$= \sum_{j=1}^n \frac{1}{\sqrt{n}} Y_{ij} f_t(x_t^j) - f_{t-1}(x_{t-1}^i) \cdot \sum \frac{1}{n} \sum_{j=1}^n (Y_{ij})^2 \cdot f_t'(x_t^j) \pm O(1/\sqrt{n})$$

We note that $Y_{ij}^2 \approx 1$ since Y_{ij} was roughly sampled from a Gaussian. Let $b_t \triangleq \frac{1}{n} \sum_{j=1}^n f_t'(x_t^j)$. We can vectorize this and see that

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

Where f_t applies the nonlinearity entrywise. We can interpret the first term as the mean-field approximation and the second term is similar to a kind of memory term called an *Onsager correction*. Finally, plugging in our definition for $\hat{x}_{t+1}^{\textcircled{i}\to k}$, we have

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

which is the estimate for the marginal distribution. One can consider random or spectral initialization.

3 Analyzing AMP, State Evolution

As $n \to \infty$, the behavior of the iterates of AMP is precisely captured by a certain distributional recursion, state evolution.

Consider the following thought experiment. Suppose inductively that the t-th iterate of AMP has normally distributed coordinates

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

i.e. some noisy estimate of our signal X. If we can get that $\mu_t \to 1$ and $\sigma_t \to 0$, then we are done. But this is not quite possible. In general though, we should be able to bound some of these terms.

Let us apply one step of AMP except we ignore the Onsager term but in exchange we pretend Y gets resampled from scratch and try to apply some inductive argument.

$$\begin{aligned} x_{t+1} &= \frac{1}{\sqrt{n}}\left(\sqrt{\frac{\lambda}{n}}XX^\top + W\right) f_t(x_t) \\ &\sim \left(\frac{\sqrt{\lambda}}{n}\langle X, f_t(x_t)\rangle X, \frac{1}{n}\|f_t(x_t)\|^2 \cdot \mathrm{Id}\right) \end{aligned}$$

This inner product is simply

$$\begin{aligned} \frac{\sqrt{\lambda}}{n}\langle X, f_t(x_t)\rangle &= \sqrt{\lambda}\frac{1}{n}\sum_{i=1}^n X_i \cdot f_t(\mu_t X_i + \sigma_t g_i), && g_i \sim \mathcal{N}(0,1) \\ &\approx \sqrt{\lambda}\mathbb{E}[x \cdot f_t(\mu_t x + \sigma_t g)], x \sim \{\pm 1\}, && g \sim \mathcal{N}(0,1) \end{aligned}$$

Hence we can approximate this inner product as

$$\mu_{t+1} \triangleq \sqrt{\lambda}\,\mathbb{E}[x \cdot f_t(\mu_t x + \sigma_t g)]$$

For the variance, we apply similar logic to obtain

$$\begin{aligned} \sigma_{t+1}^2 &\triangleq \mathbb{E}[f_t(\mu_t x + \sigma_t g)^2] \\ &\approx \frac{1}{n}\sum f_t(\mu_t X_i + \sigma_t^2 g_i)^2 \\ &= \frac{1}{n}\|f_t(x_t)\|^2 \end{aligned}$$

The crucial reason why this works is because we are resampling W every time in this thought experiment. We can initialize by setting

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

We arrived upon this recursion by dropping the Onsager term and pretending $Y's$ randomness is fresh at every iteration of AMP. In reality, the randomness of Y is fixed at the outset, and the *the Onsager term makes this heuristic thought experiment rigorous!*

Theorem 2 (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,*

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

[BM11]

The heuristic derivation becomes rigorous from this theorem where our RHS is the random sampling at each iteration compared to the LHS which corresponds to the AMP evolution of our iterates. In the asymptotic limit, we only need to track this two dimensional iteration $\mu_t\ \sigma_t$.

We have a two dimensional recursion

$$\mu_{t=1} = \mathbb{E}[x f_t(\mu_t x + \sigma_t g)], \sigma_{t+1}^2 = \mathbb{E}[f_t(\mu_t x + \sigma_t g)^2], x \sim \{\pm 1\}, g \sim \mathcal{N}(0,1)$$

and we choose a clever nonlinearity

$$f_t(y) = \mathbb{E}[\sqrt{\lambda}x | \mu_t x + \sigma_t g = y]$$

This simplifies our values as follows

$$\begin{aligned}\mu_{t+1} &= \mathbb{E}[\sqrt{\lambda}x \cdot \mathbb{E}[x|\mu_t x + \sigma_t g = y]] \\ &= \sqrt{\lambda}\mathbb{E}[\mathbb{E}[x|\mu_t x + \sigma_t g]^2] \\ &= \sqrt{\lambda}\mathbb{E}[f_t(\mu_t x + \sigma g)^2] \\ &= \sqrt{\lambda}\sigma_{t+1}^2\end{aligned}$$

Hence we have reduced the entire distributional recursion into

$$\mu_{t=1} = \sqrt{\lambda}\sigma[t+1]^2$$

We know that for the MMSE for the scalar denoising problem (MMSE for estimating x given noisy observation $\mu_t x + \sigma_t g$ where $g \sim N(0,1)$) we have:

$$\begin{aligned}\text{MMSE} &= \mathbb{E}[(X - \mathbb{E}[X|\mu_t x + \sigma_t g])^2] \\ &= \mathbb{E}[X^2] - \mathbb{E}[\mathbb{E}[X|\mu_t x + \sigma g]^2] \\ &= 1 - \lambda\sigma_{t+1}^2\end{aligned}$$

where the second equality is simply expanding out the cross term and combining.

Define the $\text{mmse}(\gamma)$ to be the MMSE for estimating X given $\sqrt{\gamma}x + \xi$ where $\xi \sim N(0,1)$. Hence if we plot $\text{mmse}(\gamma)$ as a function of the the signal to noise ratio, we see a curve that approaches 0. This means that we can simplify the distributional recursion to be

$$\sigma_{t+1}^2 = 1 - \text{mmse}(\lambda\sigma_t^2)$$

The remainder of the derivation will be completed in the next lecture.

References

[BBAP05] Jinho Baik, Gérard Ben Arous, and Sandrine Péché. Phase transition of the largest eigenvalue for nonnull complex sample covariance matrices. 2005.

[BM11] Mohsen Bayati and Andrea Montanari. The dynamics of message passing on dense graphs, with applications to compressed sensing. *IEEE Transactions on Information Theory*, 57(2):764–785, 2011.