Approximate Message Passing: A Heuristic view

Dheeraj Prasanna

SPC Lab, Department of ECE
Indian Institute of Science

August 17, 2019

Agenda

Table of Contents

Sparse Linear Regression Problem

System Model

$$\mathbf{y} = \mathbf{A}\mathbf{x}_0 + \mathbf{w} \in \mathbb{R}^m \tag{1}$$

where $\mathbf{A} \in \mathbb{R}^{m \times n}$ $(m < n)$ is a known matrix and $\mathbf{w}$ is an unknown disturbance

Recovery algorithms

- Greedy Methods : OMP, CoSAMP, SP, etc.
- Relaxation based methods : BP, Lasso, FOCUSS, etc.
- Iterative methods
- Bayesian methods like MAP estimation, SBL, etc.

Iterative Thresholding

Iterative thresholding

Updates at each iteration given by,

$$\mathbf{z}^t = \mathbf{y} - \mathbf{A}\mathbf{x}^t \tag{2}$$

$$\mathbf{x}^{t+1} = \boldsymbol{\eta}_t(\mathbf{x}^t + \mathbf{A}^T\mathbf{z}^t) \tag{3}$$

where,

$\boldsymbol{\eta}_t$ is scalar component-wise threshold functions
$\mathbf{z}^t$ is the current residue

Simple settings

Case 1: $\boldsymbol{A}$ is orthogonal; Result in one iteration
Case 2: $\boldsymbol{A}$ is invertible; Clever scaling and thresholding

} Both assume $m = n$, hence not under-determined

$m < n$ case

- Recovery when $\mathbf{x}_0$ is sufficiently sparse

Heuristics for Iterative Approaches[1]

A is Gaussian random matrix and $m < n$

Consider $\mathbf{H} = \boldsymbol{A}^T\boldsymbol{A} - \mathbf{I}$, then $\boldsymbol{A}^T\mathbf{y} = \mathbf{x}_0 + \boldsymbol{H}\mathbf{x}_0$

- $\boldsymbol{H}\mathbf{x}_0$ approximated as noisy iid Gaussian vector
- First iteration: Noisy version of sparse vector. Variance $n^{-1}||\mathbf{x}_0||_2^2$
- Second iteration: $\boldsymbol{A}^T(\mathbf{y} - \mathbf{A}\mathbf{x}^{\mathbf{1}}) = \mathbf{x}_0 + \boldsymbol{H}(\mathbf{x_0} - \mathbf{x}^{\mathbf{1}})$, noisy version with variance $n^{-1}||\mathbf{x_0} - \mathbf{x}^{\mathbf{1}}||_2^2$.

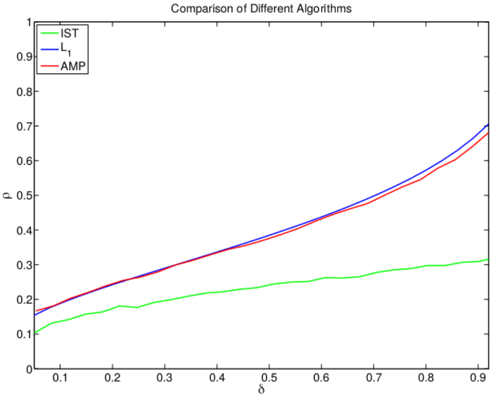
Digital Communication Interpretation

- $\mathbf{w} = \mathbf{H}\mathbf{x}_0$ is cross-channel interpretation (Mutual access interference).
- Thresholding suppresses interference by detecting "silent" channels and setting them a priori to zero
- Remaining interference proportional to estimation error rather than estimand

[1] Donoho,Maleki,Montaneri'09

Why Message Passing?

- Fast iterative methods faster than linear programming
- Phase transition for Iterative methods occur at lower sparsity levels than LP



[2] Figure Source: Donoho,Maleki,Montaneri'09

Table of Contents

Approximate Message passing (AMP)

Algorithm

initialize $\mathbf{x}^0 = \mathbf{0},\ \mathbf{z}^{-1} = \mathbf{0}$

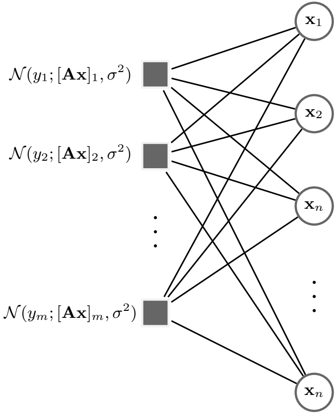
for $t = 0, 1, 2, \ldots$

$$z^t = y - Ax^t + \frac{1}{\delta} z^{t-1} \left\langle \eta_t' \left(A^* z^{t-1} + x^{t-1} \right) \right\rangle \tag{4}$$

$$x^{t+1} = \eta_t \left(A^* z^t + x^t \right) \tag{5}$$

- AMP for linear programming formulation
- "Iterative thresholding" plus "Onsager correction term (MP term)"
- Onsager correction term from theory of belief propagation.
- State Evolution formalism for MSE

Graphical Representation



- Step 1: Construct a joint distribution over $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$ parameterized by β
- Step 2: Write down sum-product algorithm

Messages

- Consider the following joint distribution,

$$\mu(\boldsymbol{x}) = \frac{1}{Z}\prod_{i=1}^{N}\exp\left(-\beta\,|x_i|\right)\prod_{a=1}^{n}\delta_{\{y_a=[\mathbf{Ax}]_a\}}$$

- Update rules for $i \in [n], a \in [m]$

$$\hat{\nu}_{a\to i}^{t}\left(x_i\right) \cong \int \prod_{j\neq i}\nu_{j\to a}^{t}\left(x_i\right)\delta_{\{y_a-[\mathbf{Ax}]_a\}}\mathrm{d}\mathbf{x} \tag{6}$$

$$\nu_{i\to a}^{t+1}\left(x_i\right) \cong e^{-\beta|x_i|}\prod_{b\neq a}\hat{\nu}_{b\to i}^{t}\left(x_i\right) \tag{7}$$

- Step 3: For large system limits, (6) and (7) can be approximated[2]as Gaussian and product of Gaussian and Laplacian densities respectively

$$z_{a\to i}^{t} = y_a - \sum_{j\in[n]\backslash i} A_{aj}x_{j\to a}^{t} \tag{8}$$

$$x_{i\to a}^{t+1} = \eta_t\Big(\sum_{b\in[m]\backslash a} A_{bi}z_{b\to i}^{t}\Big) \tag{9}$$

[2] Donoho,Maleki,Montaneri'10

From Message Passing to AMP

- For the update rules as above, 2mn messages need to be computed
- Weak dependency on i in (8) and a in (9)

$$z_a^t + \delta z_{a\to i}^t = y_a - \sum_{j\in[n]} A_{aj}\left(x_j^t + \delta x_{j\to a}^t\right) + A_{ai}\left(x_i^t + \delta x_{i\to a}^t\right) \tag{10}$$

$$x_i^{t+1} + \delta x_{i\to a}^{t+1} = \eta_t\left(\sum_{b\in[m]} A_{bi}\left(z_b^t + \delta z_{b\to i}^t\right) - A_{ai}\left(z_a^t + \delta z_{a\to i}^t\right)\right) \tag{11}$$

- Single terms of type $A_{ai}\delta z_{a\to i}^t$ are of order $\frac{1}{N}$ and can be neglected. Also, expanding (11) upto linear order, we get,

$$z_a^t + \delta z_{a\to i}^t = y_a - \sum_{j\in[n]} A_{aj}\left(x_j^t + \delta x_{j\to a}^t\right) + A_{ai}x_i^t \tag{12}$$

$$x_i^{t+1} + \delta x_{i\to a}^{t+1} = \eta_t\left(\sum_{b\in[m]} A_{bi}\left(z_b^t + \delta z_{b\to i}^t\right)\right) - \eta_t'\left(\sum_{b\in[m]} A_{bi}\left(z_b^t + \delta z_{b\to i}^t\right)\right) A_{ai}z_a^t \tag{13}$$

Decomposition and Onsager Term

- Decompose into independent sum terms
- Eliminating the weak dependency terms, we get

$$x_i^{t+1} = \eta_t \left(\sum_{b \in [m]} A_{bi} z_b^t + \sum_{b \in [m]} A_{bi}^2 x_i^t \right) \tag{14}$$

$$z_a^t = y_a - \sum_{j \in [n]} A_{aj} x_j^t + \sum_{j \in [n]} A_{aj}^2 \eta_t' \left(x_j^t + (A^* z^t)_j \right) z_a^t \tag{15}$$

- Using LLN and assumption that A is column normalized,

$$\sum_{j \in [n]} A_{aj}^2 \eta_t' \left(x_j^t + (A^* z^t)_j \right) \approx \frac{1}{m} \sum_{j \in [n]} \eta_t' \left(x_j^t + (A^* z^t)_j \right) \tag{16}$$

$$\to \frac{1}{\delta} \left\langle \eta_t' \left(x_j^t + (A^* z^t)_j \right) \right\rangle \tag{17}$$

Table of Contents

Importance of Onsager Correction term

- Onsager term approximates combined effect of reconstruction of passing of mn messages in MP

Is AMP optimal?

- Sparse denoising heuristic agrees qualitatively
- Assume the "interference" is Gaussian and independent from iteration to iteration
- Recursive equation for formal MSE called State evolution
- SE does not predict observed properties of iterative thresholding algorithms!

Onsager term makes algorithm amenable to analysis through SE!

Universality of State Evolution

- State Evolution:

$$\tau_r^{(t)} = \Psi(\tau_r^{(t-1)}) = \delta^{-1}\mathcal{E}^{(t)} + \tau_w \tag{18}$$

$$\mathcal{E}^{(t+1)} = \mathbb{E}\left\{\left[\eta^{(t)}\left(X + \mathcal{N}\left(0, \tau_r^{(t)}\right)\right) - X\right]^2\right\} \tag{19}$$

where, $X \sim p(x) = \lim_{n\to\infty} \frac{1}{n}\sum_{l=1}^{n} \delta\left(x - x_{0_l}\right)$ and
$\tau_w = \lim_{m\to\infty} \frac{1}{m}\sum_{i=1}^{m} w_i^2$,
Ψ is called MSE map

- The state evolution holds when A is drawn from i.i.d. A_{ij} such that

$$\begin{aligned}\mathbb{E}\left\{A_{ij}\right\} &= 0\\ \mathbb{E}\left\{A_{ij}^2\right\} &= 1/m\\ \mathbb{E}\left\{A_{ij}^6\right\} &= C/m \text{ for some fixed } C > 0\end{aligned}$$

often abbreviated as "sub-Gaussian A_{ij}"

Optimality Results

Exponential convergence of the algorithm

- If "Highest Fixed point" of the MSE map is stable, then State evolution converges exponentially fast to its limiting value[3]

MSE optimality of AMP

- If State Evolution has a fixed unique point, then MSE converges to the replica prediction of the MMSE as $t \to \infty$[4]

Achievability Analysis via AMP SE

- Closed form for sparsity/undersampling region or Phase transition[5]

$$\rho(\delta) = \max_{c>0} \frac{1 - 2\delta^{-1}\left[\left(1+c^2\right)\Phi(-c) - c\phi(c)\right]}{1 + c^2 - 2\left[\left(1+c^2\right)\Phi(-c) - c\phi(c)\right]}$$

- If $\rho < \rho(\delta)$, then the formal MSE of optimally-tuned AMP evolves to zero under SE[6].

[3] Donoho,Maleki,Montaneri'10
[4] Donoho,Maleki,Montaneri'11
[5] [6] Donoho,Maleki,Montaneri'09

Table of Contents

Algorithm Recap

$$\boldsymbol{v}^{(t)} = \boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}^{(t)} + \boldsymbol{\mu}^{(t)}$$

$$\boldsymbol{x}^{(t+1)} = \eta^{(t)}\left(\underbrace{\boldsymbol{x}^{(t)} + \boldsymbol{A}^\top \boldsymbol{v}^{(t)}}_{\triangleq \boldsymbol{r}^{(t)}}\right)$$

$$\boldsymbol{\mu}^{(t)} = \frac{1}{m}\boldsymbol{v}^{(t-1)} \sum_{j=1}^{n} \eta^{(t-1)\prime}\left(r_j^{(t-1)}\right)$$

Original AMP Assumptions

- $\mathbf{A} \in \mathbb{R}^{m\times n}$ drawn from i.i.d subgaussian
- $m, n \to \infty$ s.t. $\frac{m}{n} \to \delta \in (0, \infty)$ (large-system limit)

Additional assumption for proof sketch[7]

- Components of $\mathbf{A}$ i.i.d Bernoulli (i.e $A_{i,j} \in \pm\frac{1}{\sqrt{m}}$)
- $A_{i,j}$ independent of $\left\{r_{il}^{(t-1)}\right\}_{l=1}^{n}$, $\{x_l\}_{l=1}^{n}$, and $\{w_k\}_{k=1}^{m}$

[3] Schniter'19

Proof - I (Input error analysis)

- Analyze error $\mathbf{e}^{(t)}$ on input to denoiser

$$\mathbf{e}^{(t)} = \mathbf{r}^{(t)} - \mathbf{x}$$

- Re-writing,

$$\mathbf{e}^{(t)} = \left(\boldsymbol{I} - \boldsymbol{A}^\top \boldsymbol{A}\right) \boldsymbol{x}^{(t)} - \left(\boldsymbol{I} - \boldsymbol{A}^\top \boldsymbol{A}\right) \boldsymbol{x} + \boldsymbol{A}^\top \left(\boldsymbol{w} + \boldsymbol{\mu}^{(t)}\right) \qquad (20)$$

- Examine jth component of first term of $\mathbf{e}^{(t)}$

$$\begin{aligned}\left[\left(\boldsymbol{I} - \boldsymbol{A}^\top \boldsymbol{A}\right) \mathbf{x}^{(t)}\right]_j &= \left(1 - \sum_{i=1}^{m} a_{ij}^2\right) x_j^{(t)} - \sum_i a_{ij} \sum_{l \neq j} a_{il} x_l^{(t)} \\ &= -\sum_i a_{ij} \sum_{l \neq j} a_{il} \eta^{(t-1)} \left(r_l^{(t-1)}\right) \\ &= -\sum_i a_{ij} \sum_{l \neq j} a_{il} \eta^{(t-1)} \underbrace{(x_l^{(t-1)} + \sum_{k \neq i} a_{kl} v_k^{(t-1)}}_{\triangleq r_{il}^{(t-1)}} + a_{il} v_i^{(t-1)})\end{aligned}$$

Proof - II (Important Lemma's

- **Lemma 1.** Consider the quantity $z_i = \sum_{j=1}^{n} a_{ij} u_j$, where a_{ij} are realizations of i.i.d. random variables A_{ij} with zero mean and $\mathbb{E}\left\{A_{ij}^2\right\} = 1/m$.
  If $\{A_{ij}\}$ are drawn independently of $\{u_j\}$, and $\{u_j\}$ scale as $O(1)$ in the large-system limit, then z_i also scales as $O(1)$

- **Lemma 2.** Under the additional assumptions and Onsager choice for $\boldsymbol{\mu}^{(t)}$, the elements of $\mathbf{v}^{(t)}, \mathbf{r}^{(t)}, \mathbf{x}^{(t)}$ and $\boldsymbol{\mu}^{(t)}$ scale as $O(1)$ in the large-system limit for all iterations t.

$$\eta^{(t-1)}\left(r_{il}^{(t-1)} + a_{il} v_i^{(t-1)}\right)$$
$$= \eta^{(t-1)}\left(r_{il}^{(t-1)}\right) + a_{il} v_i^{(t-1)} \eta^{(t-1)\prime}\left(r_{il}^{(t-1)}\right) + \underbrace{\frac{1}{2} a_{il}^2 \left(v_i^{(t-1)}\right)^2 \eta^{(t-1)\prime\prime}\left(r_{il}^{(t-1)}\right)}_{O(1/m)}$$

Proof - III (Error characterization)

- Applying Taylor series expansion

$$[(\boldsymbol{I}-\boldsymbol{A}^\top\boldsymbol{A})\,\mathbf{x}^{(t)}]_j \approx -\sum_i a_{ij}\sum_{l\neq j} a_{il}\eta^{(t-1)}\left(r_{il}^{(t-1)}\right) - \frac{1}{m}\sum_i a_{ij}v_i^{(t-1)}\sum_{l\neq j}\eta^{(t-1)\prime}\left(r_{il}^{(t-1)}\right) \tag{21}$$

- Similarly,

$$\left[\left(\boldsymbol{I}-\boldsymbol{A}^\top\boldsymbol{A}\right)\boldsymbol{x}\right]_j = -\sum_i a_{ij}\sum_{l\neq j} a_{il}x_l \tag{22}$$

- Using (21) and (22) on (20),

$$e_j^{(t)} = \sum_i a_{ij}\sum_{l\neq j} a_{il}\left[x_l-\eta^{(t-1)}\left(r_{il}^{(t-1)}\right)\right] + \sum_i a_{ij}w_i + \sum_i a_{ij}\left[\mu_i^{(t)} - v_i^{(t-1)}\frac{1}{m}\sum_{l\neq j}\eta^{(t-1)\prime}\left(r_{il}^{(t-1)}\right)\right] \tag{23}$$

- First and Second terms behave like realizations of Gaussians in large-system limit.
- In the third term $v_i^{(t-1)}$ is strongly coupled to a_{ij}, hence difficult to characterize for general $\mu_i^{(t)}$

Proof - IV (Onsager term)

- For Onsager choice, 3rd term of (23) takes the form,

$$\sum_i a_{ij} \left[\frac{1}{m} v_i^{(t-1)} \sum_l \eta^{(t-1)\prime} \left(r_l^{(t-1)} \right) - \frac{1}{m} v_i^{(t-1)} \sum_{l \neq j} \eta^{(t-1)\prime} \left(r_{il}^{(t-1)} \right) \right]$$
$$= \frac{1}{m} \sum_i a_{ij} v_i^{(t-1)} \left[\eta^{(t-1)\prime} \left(r_j^{(t-1)} \right) + \sum_{l \neq j} \left(\eta^{(t-1)\prime} \left(r_l^{(t-1)} \right) - \eta^{(t-1)\prime} \left(r_{il}^{(t-1)} \right) \right) \right]$$
$$\approx \frac{1}{m} \sum_i a_{ij} v_i^{(t-1)} \left[\eta^{(t-1)\prime} \left(r_j^{(t-1)} \right) + \sum_{l \neq j} a_{il} v_i^{(t-1)} \eta^{(t-1)\prime\prime} \left(r_{il}^{(t-1)} \right) \right]$$

- First term,

$$\frac{1}{m} \sum_{i=1}^{m} \underbrace{a_{ij} v_i^{(t-1)} \eta^{(t-1)\prime} \left(r_j^{(t-1)} \right)}_{O(1/\sqrt{m})} = O(1/\sqrt{m})$$

- Second term,

$$\underbrace{\frac{1}{m} \sum_{i=1}^{m} a_{ij} \left(v_i^{(t-1)} \right)^2 \underbrace{\sum_{l \neq j} a_{il} \eta^{(t-1)\prime\prime} \left(r_{il}^{(t-1)} \right)}_{O(1)}}_{O(\frac{1}{\sqrt{m}})} = O(1/\sqrt{m})$$

Proof - V (Mean Squared error)

- For large m and Onsager term choice, (20) becomes,

$$e_j^{(t)} \approx \sum_i a_{ij} \sum_{l \neq j} a_{il} \underbrace{[x_l - \eta^{(t-1)}\left(r_{il}^{(t-1)}\right)]}_{\triangleq \epsilon_{il}^{(t)}} + \sum_i a_{ij} w_i \tag{24}$$

- First term converges to Gaussian with mean and variance,

$$\mathbb{E}\left\{\sum_i A_{ij} \sum_{l \neq j} A_{il} \epsilon_{il}^{(t)}\right\} = \sum_i \mathbb{E}\{A_{ij}\} \sum_{l \neq j} \mathbb{E}\{A_{il}\} \epsilon_{il}^{(t)} = 0 \tag{25}$$

$$\begin{aligned}\mathbb{E}\left\{\left(\sum_i A_{ij} \sum_{l \neq j} A_{il} \epsilon_{il}^{(t)}\right)^2\right\} &= \sum_i \mathbb{E}\{A_{ij}^2\} \sum_{l \neq j} \mathbb{E}\{A_{il}^2\} \left(\epsilon_{il}^{(t)}\right)^2 \\ &= \frac{1}{m^2} \sum_i \sum_{l \neq j} \left(\epsilon_{il}^{(t)}\right)^2 \end{aligned} \tag{26}$$

- From Taylor expansion, $\epsilon_{il}^{(t)}$ can be approximated to $\epsilon_l^{(t)}$ in large system limit
- Dependence on i is removed in (26), and can be approximated as $\delta^{-1}\mathcal{E}^{(t)}$
- $\mathcal{E}^{(t)} \triangleq \lim_{n \to \infty} \frac{1}{n} \sum_{l=1}^{n} \left(\epsilon_l^{(t)}\right)^2$ is the average squared error

Proof - VI (AMP State evolution)

- Similar to average squared error, second term in (24) converges to Gaussian with mean 0 and variance τ_w which is the second moment of noise.

$$\tau_w \triangleq \lim_{m\to\infty} \frac{1}{m}\sum_{i=1}^{m} w_i^2$$

- With AMP's choice of $\boldsymbol{\mu}^{(t)}$, jth component of denoiser input error,

$$e_j^{(t)} \sim \mathcal{N}(0, \underbrace{\delta^{-1}\mathcal{E}^{(t)} + \tau_w}_{\triangleq \tau_r^{(t)}})$$

- Using definition of $\mathcal{E}^{(t)}$,

$$\frac{1}{n}\sum_{l=1}^{n}\left(\epsilon_l^{(t)}\right)^2 \approx \frac{1}{n}\sum_{l=1}^{n}\left[\eta^{(t-1)}\left(x_l + \mathcal{N}\left(0, \tau_r^{(t-1)}\right)\right) - x_l\right]^2$$
$$= \mathbb{E}\left[\eta^{(t-1)}\left(X + \mathcal{N}\left(0, \tau_r^{(t-1)}\right)\right) - X\right]^2$$

Table of Contents

AMP with Prior Information

- So far, distribution of signal not known
- If input distribution can be estimated, it can improve the recovery algorithms and also can be used as a benchmark
- Consider distribution,

$$\mu(\mathbf{x}) = \frac{1}{Z} \prod_{a=1}^{n} \delta_{\{y_a = [\mathbf{A}\mathbf{x}]_a\}} \prod_{i=1}^{N} \alpha_i\left(x_i\right)$$

- Sum-product update rules can be simplified similar to AMP updates
- Consider a family of densities,

$$f_i(\mathrm{d}s; x, b) \equiv \frac{1}{z_\beta(x, b)} \exp\left\{-\frac{\beta}{2b}(s - x)^2\right\} \alpha_i(\mathrm{d}s)$$

  F denote its mean.
- AMP updates obtained

$$x^t = \mathrm{F}\left(x^t + A^* z^t; \tau^t\right)$$

$$z^{t+1} = y - Ax^t + \frac{1}{\delta} z^t \left\langle \mathrm{F}'\left(x^{t-1} + A^* z^{t-1}\right)\right\rangle$$

AMP Extensions

Few algorithms

- Vector AMP
- DCS-AMP
- AMP for SBL
- AMP with Side Information

Future ideas

- AMP for structured sparse signal like block sparsity, hierarchical sparsity, etc.
- Effect of correlation in signals

P. Schniter, "A simple derivation of amp and its state evolution via first-order cancellation," *arXiv preprint arXiv:1907.04235*, 2019.

D. L. Donoho, A. Maleki, and A. Montanari, "Message-passing algorithms for compressed sensing," *Proceedings of the National Academy of Sciences*, vol. 106, no. 45, pp. 18 914–18 919, 2009.

——, "Message passing algorithms for compressed sensing: I. motivation and construction," in *2010 IEEE information theory workshop on information theory (ITW 2010, Cairo)*. IEEE, 2010, pp. 1–5.

D. L. Donoho, A. Maleki, and A. Montanari, "Message passing algorithms for compressed sensing: Ii. analysis and validation," in *2010 IEEE Information Theory Workshop on Information Theory (ITW 2010, Cairo)*, Jan 2010, pp. 1–5.

M. Bayati and A. Montanari, "The dynamics of message passing on dense graphs, with applications to compressed sensing," *IEEE Transactions on Information Theory*, vol. 57, no. 2, pp. 764–785, Feb 2011.

M. Al-Shoukairi and B. Rao, "Sparse bayesian learning using approximate message passing," in *2014 48th Asilomar Conference on Signals, Systems and Computers*, Nov 2014, pp. 1957–1961.

M. Bayati, M. Lelarge, A. Montanari *et al.*, "Universality in polytope phase transitions and message passing algorithms," *The Annals of Applied Probability*, vol. 25, no. 2, pp. 753–822, 2015.

A. Montanari and D. Tse, "Analysis of belief propagation for non-linear problems: The example of cdma (or: How to prove tanaka's formula)," in *2006 IEEE Information Theory Workshop - ITW '06 Punta del Este*, March 2006, pp. 160–164.