Message Passing Algorithms for Compressed Sensing

by D. Donoho, A. Maleki, and A. Montanari

presented by

Nate Strawn

November 4th, 2011

Outline

- Review of sparse reconstruction and compressed sensing
- Introduction to approximate message passing (AMP)
- Justifying derivation of AMP using belief propagation
- Derivation of AMP: Four successive approximations
- Analysis of AMP: State Evolution and Minimax Thresholding

Sparse reconstruction

A is $n \times N$, $y = Ax^0$, solve

$$\hat{x}_{\ell 0} = \arg \min_{Ax=y} \|x\|_0$$

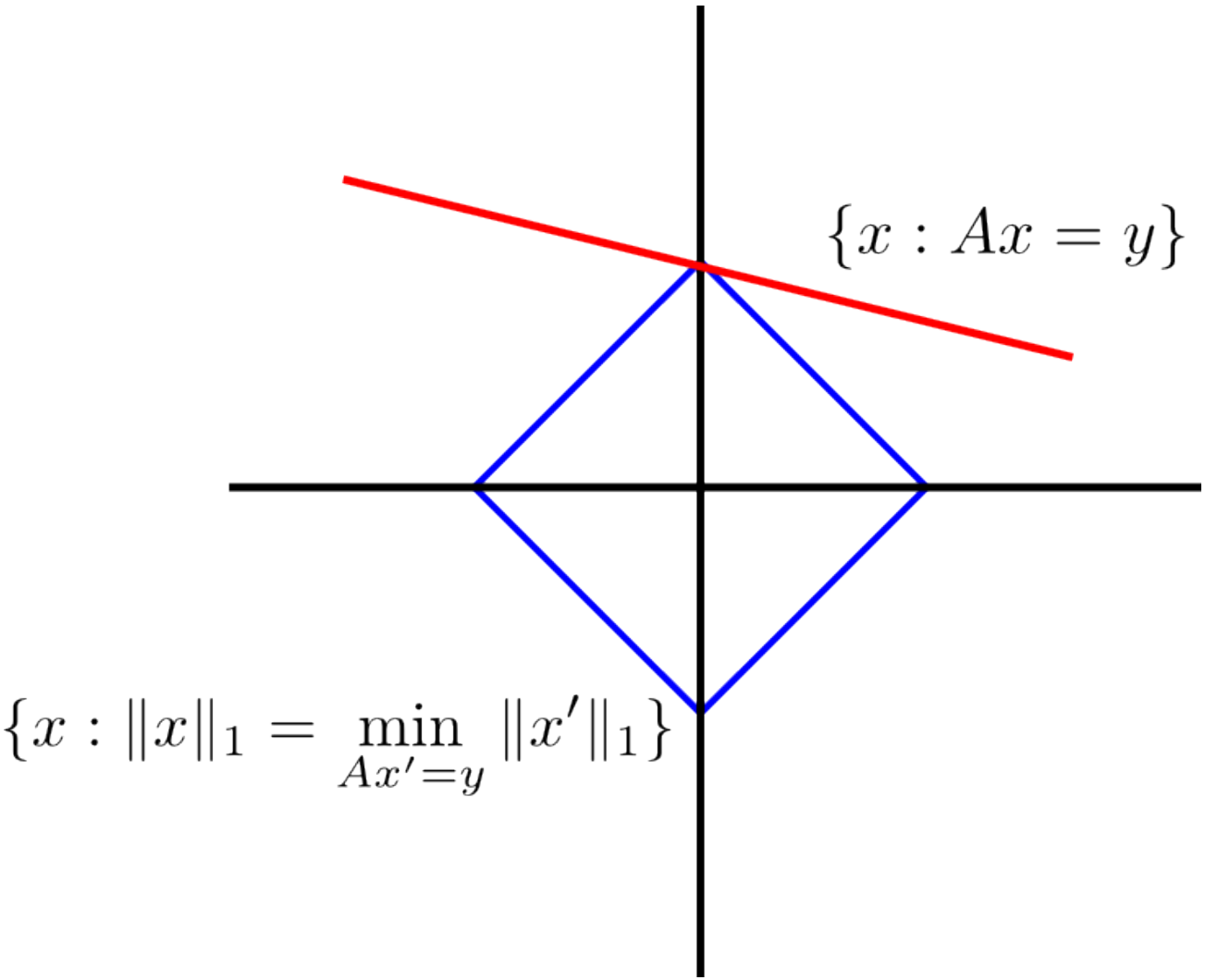
where $\|x\|_0 = \#\text{Nonzero entries of } x$. This NP-Hard combinatorial optimization problem is relaxed to linear programming Basis Pursuit problem:

$$\hat{x}_{\ell 1} = \arg \min_{Ax=y} \|x\|_1$$

where $\|x\|_1 = \sum |x_i|$.

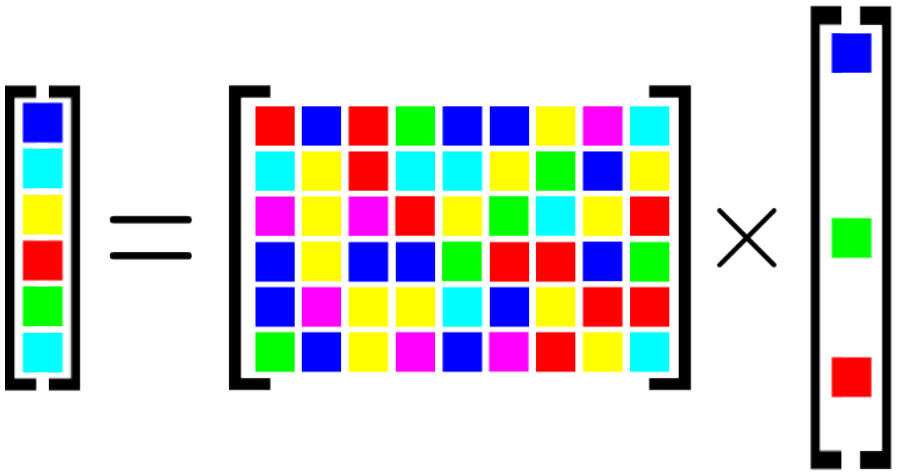
Heuristic for the success of Basis Pursuit

Success ($\hat{x}_{\ell 1} = \hat{x}_{\ell 0}$) depends on the kernel of A!



Compressed Sensing

Observation: $n \times N$ matrix A with a "good" kernel can have $n << N$ if sparsity k is sufficiently low

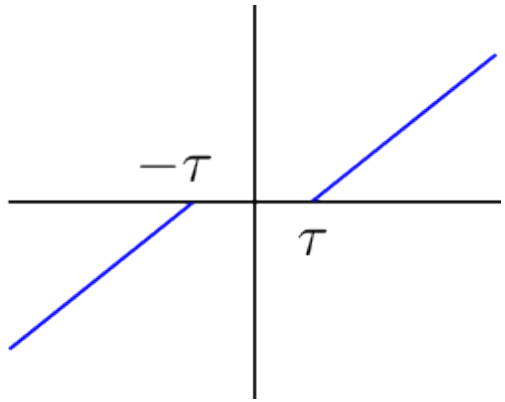


"Goodness" is NP-Hard to check, but many families of random matrices are good with high probability (A with i.i.d. $\mathcal{N}(0, \frac{1}{n})$ entries assumed for the rest of the talk).

Iterative soft thresholding algorithm

Can we get something more efficient than Basis Pursuit (faster than linear programming)?

$$\eta(x;\tau) = \begin{cases} x+\tau, & x < -\tau \\ 0, & -\tau \le x \le \tau \\ x-\tau, & \tau < x \end{cases}$$



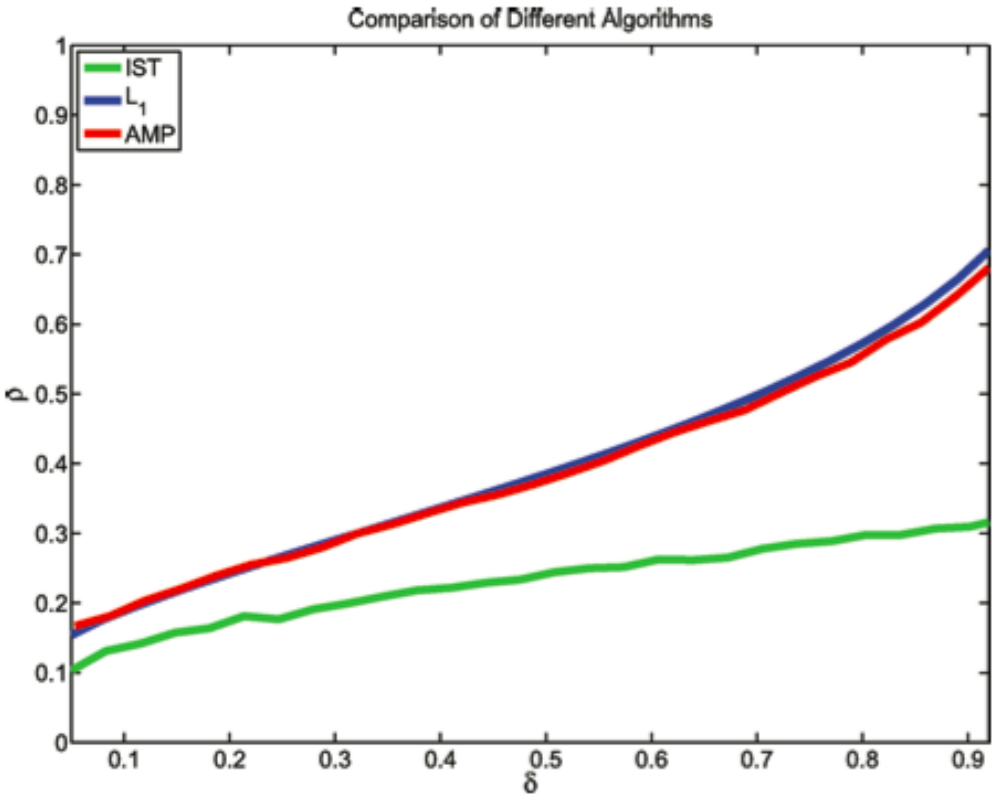
ISTA:

- $x^{t+1} = \eta(A^* z^t + x^t; \tau^t)$
- $z^t = y - Ax^t$

Heurisitic: switch between thresholding insignificant contributions and reinforcing $Ax = y$

Sparse recovery phase transition of ISTA vs BP

$\delta = n/N$ is the undersampling rate, $\rho = k/n$ is the sparsity-undersampling ratio



Above PTB, success probability declines rapidly, below the PTB, success is almost certain.

Approximate message passing algorithm

The AMP algorithm:

- $x^{t+1} = \eta(A^* z^t + x^t; \tau^t)$
- $z^t = y - Ax^t + \frac{z^{t-1}}{\delta} \langle \eta'(A^* z^{t-1} + x^{t-1}; \tau^{t-1}) \rangle \leftarrow$ The Onsager reaction term
- $\tau^t = \frac{\tau^{t-1}}{\delta} \langle \eta'(A^* z^{t-1} + x^t; \tau^{t-1}) \rangle$

Properties of this algorithm:

- Fastest game in town
- Admits a complete analysis in the large system limit:
  - Explicit form for the phase transition boundary of CS
  - Elegant relationship with minimax optimality
  - Linear convergence under the PTB

Motivating the derivation of AMP: phase transition phenomenon in compressed sensing

- An explicit form for the phase transition boundary provides the sharpest asymptotic undersampling conditions for compressed sensing.
- Explicit expression for the Basis Pursuit phase transition boundary (Donoho, Tanner '05).
- Traditionally (in statistical mechanics), graphical models are our best tool for analyzing phase transitions.
- We shall derive AMP from Belief Propagation, but first we want to understand why this might help uncover a phase transition.

Phase transitions, Gibbs free energy, and belief propagation

Boltzmann's law for statistical mechanical system is

$$P(\alpha) = \frac{\exp\{-\beta E_\alpha\}}{\sum_\alpha \exp\{-\beta E_\alpha\}}$$

with $\beta = 1/T$, E_α energy of state α. Partition function is then

$$Z(\beta) = \sum_\alpha \exp\{-\beta E_\alpha\} = \exp\{-\beta F(\beta)\}$$

where $F(\beta)$ is the free energy. All thermodynamic quantities of interest can be computed if we know

$$F(\beta) = -\frac{1}{\beta} \ln Z(\beta)$$

Phase transitions, Gibbs free energy, and belief propagation

- Phase transitions occur if there are singularities in the free energy or any of its derivatives
- Phase transitions in physics first analytically uncovered via Onsager's solution to the partition function of the Ising model
- Computing the partition function is generally VERY HARD, so we must be satisfied with approximations

Phase transitions, Gibbs free energy, and belief propagation

The Gibbs free energy is a variational gadget that returns Boltzmann's law at the minimum:

$$G(P) = \sum_{\alpha} P(\alpha)E(\alpha) - T\sum_{\alpha} P(\alpha)\ln P(\alpha)$$

we then assume a simplified form for P and minimize G_{approx} to obtain an approximate free energy:

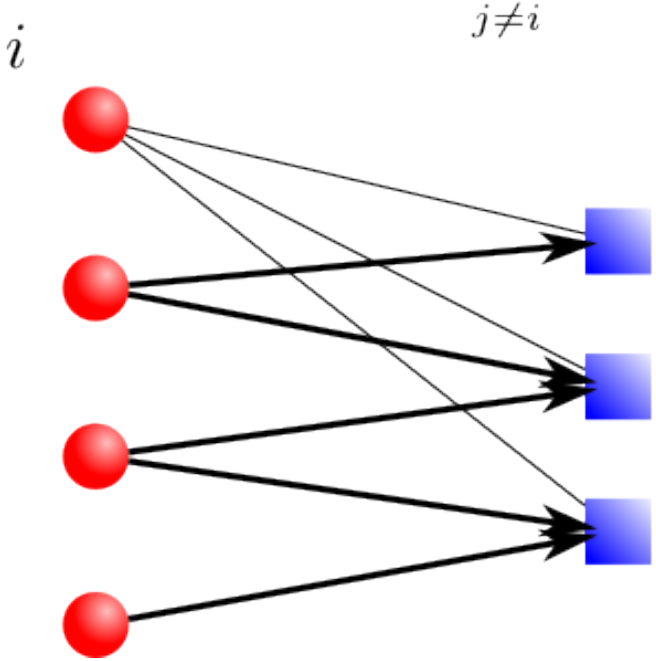
- Mean field approximation: $P(\alpha) = \prod p_i(\alpha_i)$
- Bethe approximation: $P(\alpha) = \prod_a p(\alpha_{A_a}) \prod p_i(\alpha_i)^{1-\deg_i}$

There is a deep connection between belief propagation and the Bethe approximation.

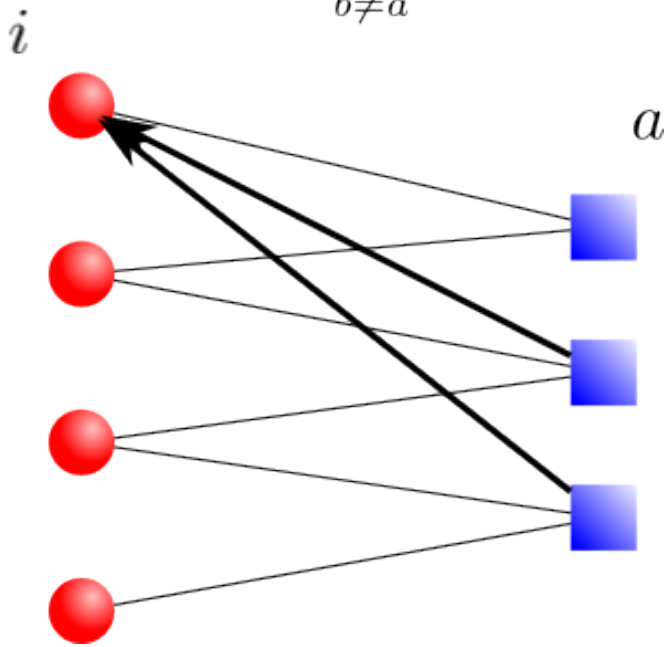
Belief propagation

Belief propagation computes the marginals of $p(x) = \prod_a f_a(\mathbf{x}_a)$. Messages are passed between variables and factors:

$$m^t_{a\to i}(x_i) = \sum f_a(x) \prod_{j\neq i} n^t_{j\to a}(x_j)$$



$$n^{t+1}_{i\to a}(x_i) = \prod_{b\neq a} m^t_{b\to i}(x_i)$$



Beliefs are $b_i(x_i) \propto \prod_a m_{a\to i}$, $b_a(\mathbf{x}_a) \propto f_a(\mathbf{x}_a) \prod n_{i\to a}$.

Connection between the Bethe approximation and belief propagation

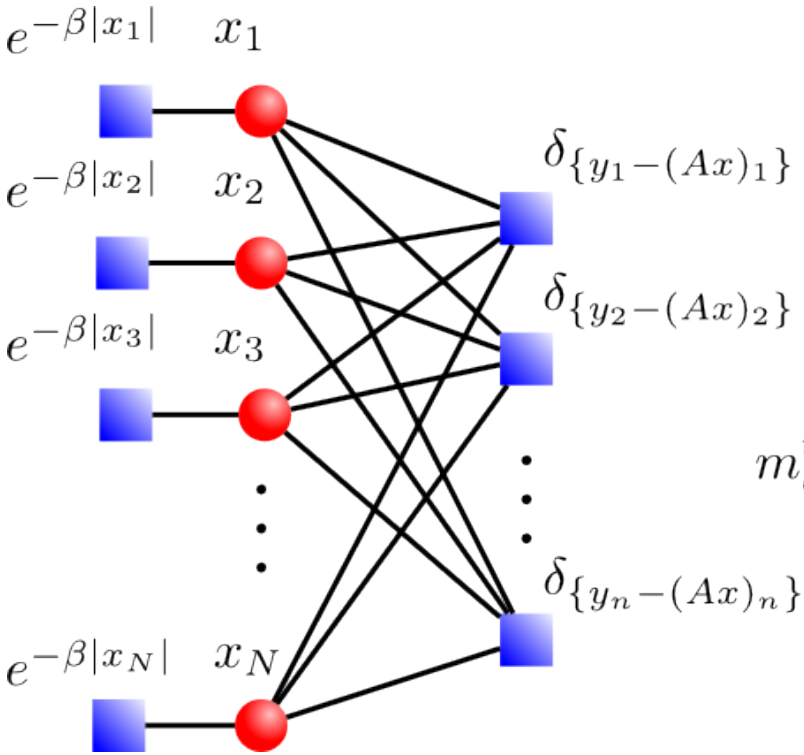
Theorem (Yedida, Freeman, Weiss '02)

Let $\{m_{a\to i}(x_i), n_{i\to a}(x_i)\}$ be a set of BP messages and let $\{b_a(\mathbf{x}_a), b_i(x_i)\}$ be the corresponding beliefs. Then the beliefs are fixed points of the BP algorithm if and only if they are zero gradient points of the Bethe free energy subject to the constraint that all the beliefs are normalized and consistent.

Deriving AMP from belief propagation

The following density concentrates on the basis pursuit solution as $T \to 0$:

$$p(x) = \frac{1}{Z} \exp\{-\beta \|x\|_1\} \delta_{y=Ax}$$



Messages:

$$m^t_{a\to i}(x_i) \propto \int \delta_{\{y_a-(Ax)_a\}} \prod_{j\neq i} n^t_{j\to a}(x_j)dx_j$$

$$n^{t+1}_{i\to a}(x_i) \propto e^{-\beta|x_1|} \prod_{b\neq a} m^t_{b\to i}(x_i)$$

Deriving AMP from belief propagation

Lemma (Normal approximation to $m^t_{a\to i}$)

For large N, suppose $n^t_{j\to a}$ has mean $x^t_{j\to a}$ and variance $\tau^t_{j\to a}/\beta$. If the third moments of the $n^t_{j\to a}$'s are uniformly bounded, then

$$m^t_{a\to i}(x_i) \approx \sqrt{\frac{\beta A^2_{ai}}{2\pi\hat{\tau}^t_{a\to i}}}\, e^{-\beta(A_{ai}x_i - z^t_{a\to i})^2/2\hat{\tau}^t_{a\to i}}$$

where $z^t_{a\to i} = y_a - \sum_{j\neq i} A_{aj}x^t_{j\to a}$ and $\hat{\tau}^t_{a\to i} = \sum_{j\neq i} A^2_{aj}\tau^t_{j\to a}$.

Lemma (Laplace times Normal approximation to $n^t_{i\to a}$)

$$n^{t+1}_{i\to a}(x_i) \approx f_\beta(x_i; \sum_{b\neq a} A_{bi}z^t_{b\to i}, \hat{\tau}^t) = \frac{e^{-\beta|x_i| - \frac{\beta}{2\hat{\tau}^t}(x_i - \sum_{b\neq a} A_{bi}z^t_{b\to i})^2}}{z_\beta(x_i, \hat{\tau}^t)}$$

Message behavior as $\beta \to \infty$

As $\beta \to \infty$, $f_\beta(x_i; x, b)$ sharply concentrates at

$$\arg\min_s |s| + \frac{1}{2b}(s-x)^2 = \eta(x; b)$$

Lemma (Large β limit of means and variances)

In the limit $\beta \to \infty$,

$$x_{i\to a}^{t+1} = \eta\Big(\sum_{b\neq a} A_{bi} z_{b\to i}^t; \hat{\tau}^t\Big)$$

$$\hat{\tau}^{t+1} = \frac{\hat{\tau}^t}{N\delta} \sum_{i=1}^{N} \eta'\Big(\sum_b A_{bi} z_{b\to i}^t; \hat{\tau}^t\Big)$$

From message passing to AMP

Now, reduce from nN messages to $N + n$ using one more approximation. Assume

$$x_{i\to a}^t \approx x_i^t + \Delta x_{i\to a}^t \text{ and } z_{a\to i}^t \approx z_a^t + \Delta z_{a\to i}^t$$

with the Δ terms $O(1/\sqrt{N})$.

- Remove $O(1/N)$ terms to get $z_{a\to i}^t \approx y_a - \sum_j A_{aj} x_{j\to a}^t + A_{ai} x_i^t$, so $\Delta z_{a\to i}^t = A_{ai} x_i^t$
- Taylor expand to get

$$x_{i\to a}^{t+1} = \eta(\sum A_{bi}(z_b^t + \Delta z_{b\to i}^t); \tau^t) - \eta'(\sum A_{bi}(z_b^t + \Delta z_{b\to i}^t); \tau^t) A_{ai} z_a^t$$

- For large N, $\sum A_{bi}^2 x_i^t \approx x_i^t$, so $x^{t+1} = \eta(A^* z^t + x^t; \tau^t)$ and $\Delta x_{i\to a}^t = -\eta'(\sum A_{bi}(z_b^{t-1} + \Delta z_{b\to i}^{t-1}); \tau^t) A_{ai} z_a^t$
- Sub into $z_a^t = y_a - \sum A_{aj} x_j^{t-1} + \sum A_{aj}^2 \eta'(x_j^{t-1} + (A^* z^{t-1})_j); \tau^t) z_a^{t-1}$ and follows from LLN
- Also get expression for τ^t

Analysis of AMP: State Evolution

State evolution assumes that, for this iteration,

$$\lim_{N\to\infty} \frac{1}{N}\|x^t - x^0\|_2^2 = \mathbb{E}[\eta(X_0 + \frac{\sigma_t}{\sqrt{\delta}}Z; \lambda\sigma_t) - X_0]^2 = \sigma_{t+1}^2 = \psi(\sigma_t^2)$$

Where X_0 is the weak limit of empirical distributions on x_0's entries, and Z is standard normal. If this is true, then the asymptotic mean square error is TRACTABLE:

$$\left.\frac{d\psi}{d\sigma^2}\right|_{\sigma\downarrow 0} = (\frac{1}{\delta} + \lambda^2)\rho\delta + (\frac{1}{\delta} + \lambda^2)(1-\rho\delta)2\Phi(-\lambda\sqrt{\delta}) - \frac{\lambda}{\sqrt{\delta}}(1-\rho\delta)2\phi(-\lambda\sqrt{\delta})$$

The MSE map only has one fixed point at zero if and only if it is convex (true in this case) and this derivative is less than 1. Setting this to 1, we obtain the phase transition boundary.

State Evolution: the Heuristic

- Instead, $A(t)$ are iid with $A_{ij}(t) \sim \mathcal{N}(0, 1/n)$, $x^{t+1} = \eta_t(A(t)^* z^t + x^t)$, $z^t = y^t - A(t)x^t$
- Eliminate z^t: $x^{t+1} = \eta_t(x_0 + A(t)^* w + B(t)(x^t - x_0))$, $B(t) = I - A(t)^* A(t)$
- $B(t)(x^t - x_0)$ iid normal entries with 0 mean and variance $\hat{\tau}_t^2/\delta$
- Conditional on w, $A(t)^* w$ has iid normal entries with 0 and variance $\|w\|^2/n \to \sigma^2$
- Now, $A(t)^* w$ approximately independent from $B(t)(x^t - x_0)$
- So random entries of $x_0 + A(t)^* w + B(t)(x^t - x_0)$ approach $X_0 + \tau_t Z$ where $\tau_t^2 = \sigma^2 + \frac{1}{\sigma}\hat{\tau}_t^2$
- Conclude that $\lim_{N\to\infty} \frac{1}{N}\|x^{t+1} - x_0\|^2 = \mathbb{E}[\eta_t(X_0 + \tau_t Z) - X_0]^2$

State Evolution: result of Bayati and Montanari

State evolution has been rigorously shown in a general context (Bayati, Montanari '10).

- Proof uses a conditioning technique similar to Bolthausen '09: an asymptotic expression is derived for each term of the iteration by knowing how to take conditional expectations with respect to Gaussian matrices
- The proof uses an induction argument to ensure the asymptotic expressions for each t, and then uses the expression to show SE at each t

Relationship with minimax thresholding

- Define the minimax thresholding policy

$$M^*(\varepsilon) = \inf_{\lambda} \sup_{F \in \mathcal{F}_\varepsilon} \mathbb{E}_F \left[\eta(X + Z; \lambda) - X\right]^2$$

  where $\mathcal{F}_\varepsilon$ consists of all distributions with $\varepsilon = \rho\delta$ sparsity fraction

- The best possible threshold is the minimax threshold at nonzero fraction ε appropriately scaled by the effective noise level, $\lambda^t = \lambda^*(\varepsilon)\sigma/\sqrt{\delta}$. This only depends on the iteration through the effective noise level at the iteration.

- The guarantee is then

$$\text{MSE} \leq M^*(\varepsilon)\tau^2 = M^*(\varepsilon)\frac{\sigma^2}{\delta},$$

  which ensures a reduction in the MSE if and only if $M^* < \delta$.

Relationship with minimax thresholding

- Define $\rho_{MM}(\delta) = \sup\{\rho : M^*(\rho\delta) < \delta\}$ the minimax thresholding phase transition such that below the transition, state evolution with minimax thresholding converges.
- Theorem: $M^*(\rho\delta) = \delta$ if and only if $\rho = \rho_{\mathsf{SE}}(\delta)$
- Below the SE phase transition, both SE and Minimax converge. In general, the thresholds are different. As a consequence, the convergence rates may be different; in general minimax will converge more quickly
- Can also demonstrate that no other nonlinear thresholding function is superior when $\delta \to 0$

References

Bayati, Montanari (2010)

The dynamics of message passing on dense graphs, with applications to compressed sensing.

IEEE Trans. Inform. Theory

Bolthausen (2009)

On the high-temperature phase of the Sherrington-Kirkpatrick model.

Seminar at EURANDOM

Yedida, Freeman, Weiss (2002)

Constructing Free Energy Approximations and Generalized Belief Propagation Algorithms.

Mitsubishi Research Technical Report