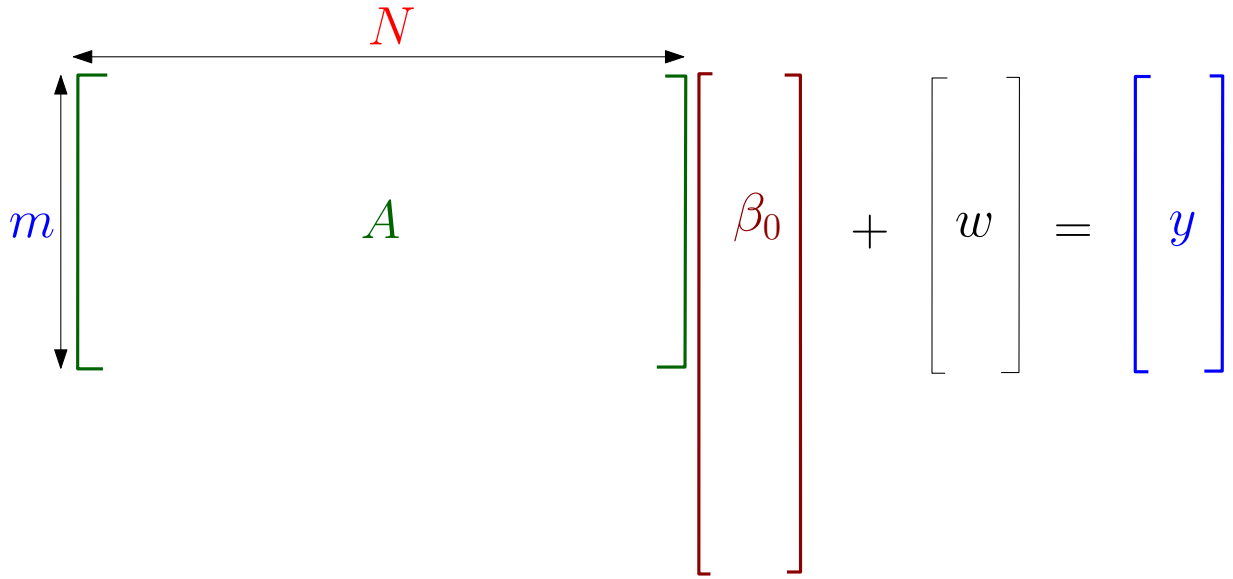
Recent Applications of Approximate Message Passing Algorithms for High-dimensional Statistical Estimation

Cynthia Rush, Columbia University

Joint work with
Ramji Venkataramanan (University of Cambridge)

February 12, 2018

High-dimensional Linear Regression



Want to reconstruct β_0 from $y = A\beta_0 + w$

- y: measurement vector in $\mathbb{R}^m$
- w: measurement noise in $\mathbb{R}^m$
- A: $m \times N$ design matrix
- Number of measurements $m < N$
- β_0 has $k < N$ non-zero elements, i.e. it is k-**sparse**

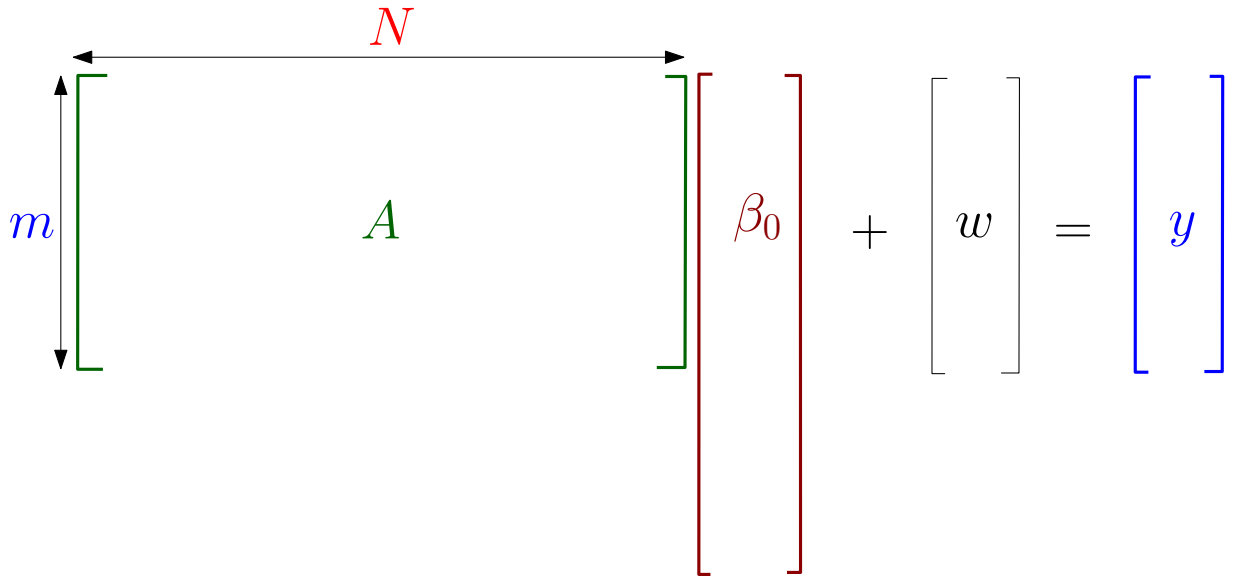
High-dimensional Linear Regression

Many Applications

- **Channel Coding in Communications**
  y = received sample w = noise/interference
  A = coding dictionary β_0 = message
- **Imaging: Medical, Seismic, Compressive Sensing...**
  y = measurements w = sensor noise
  A = basis representation β_0 = sparse image/signal
- **Statistics/Machine Learning**
  y = experimental outcome w = model error
  A = feature data β_0 = prediction coefficients

Problem sizes are large, computational complexity of reconstruction algorithm is a concern.

High-dimensional Linear Regression



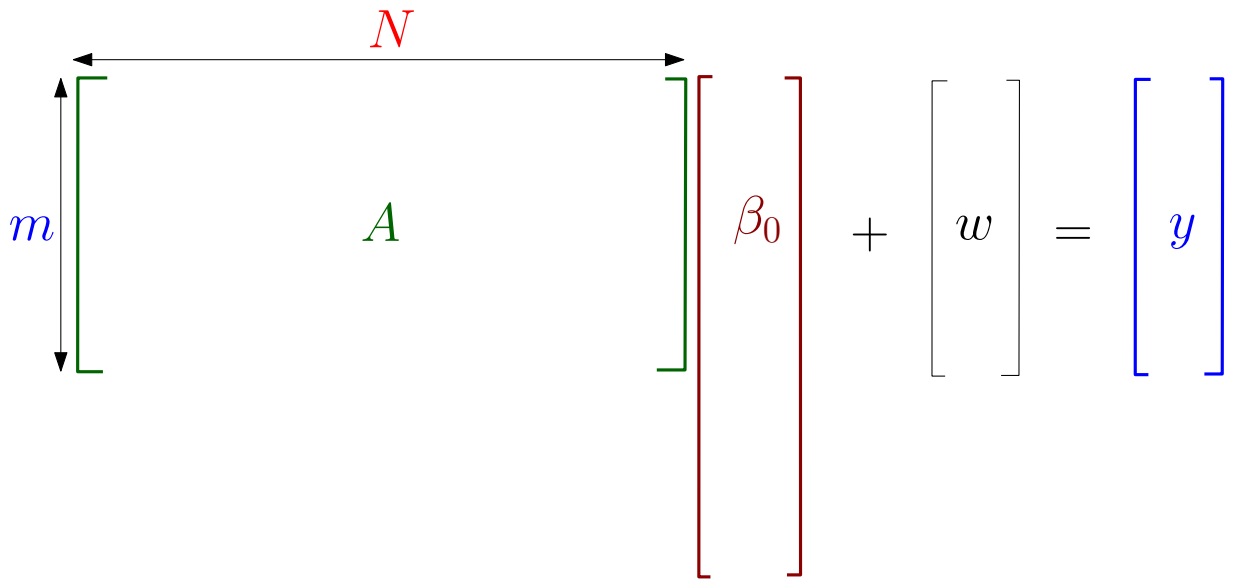
Goal: reconstruct k-sparse β_0 from $y = A\beta_0 + w$

Want to solve:

$$\hat{\beta} = \arg\min_{\beta \in \mathbb{R}^N} \|y - A\beta\|^2 \text{ subject to } \sum_{j=1}^{N} \mathbb{I}\{\beta_j \neq 0\} \leq k.$$

Unfortunately, a very hard combinatorial problem.

High-dimensional Linear Regression



Goal: reconstruct k-sparse β_0 from $y = A\beta_0 + w$

Want to solve:

$$\hat{\beta} = \underset{\beta \in \mathbb{R}^N}{\arg\min} \|y - A\beta\|^2 \text{ subject to } \|\beta\|_0 \leq k.$$

Unfortunately, a very hard combinatorial problem.

High-dimensional Linear Regression

Want to solve:

$$\hat{\beta} = \underset{\beta \in \mathbb{R}^N}{\arg\min} \ \|y - A\beta\|^2 \text{ subject to } \|\beta\|_0 \leq k.$$

Unfortunately, a very hard combinatorial problem.

Instead, a convex relaxation:

$$\hat{\beta} = \underset{\beta \in \mathbb{R}^N}{\arg\min} \ \|y - A\beta\|^2 \text{ subject to } \|\beta\|_1 \leq \lambda.$$

High-dimensional Linear Regression

Want to solve:

$$\hat{\beta} = \arg\min_{\beta \in \mathbb{R}^N} \|y - A\beta\|^2 \text{ subject to } \|\beta\|_0 \leq k.$$

Unfortunately, a very hard combinatorial problem.

Instead, a convex relaxation:

$$\hat{\beta} = \arg\min_{\beta \in \mathbb{R}^N} \|y - A\beta\|^2 \text{ subject to } \|\beta\|_1 \leq \lambda.$$

If A satisfies certain conditions (e.g., RIP) then can get a good estimate of *sparse* β_0 by solving a convex program (LASSO):

$$\hat{\beta} = \arg\min_{\beta \in \mathbb{R}^N} \|y - A\beta\|^2 + \tilde{\lambda}\|\beta\|_1$$

[Donoho '06, Candes-Romberg-Tao'06, . . .]

Approximate Message Passing (AMP)

AMP: low complexity, scalable algorithm studied to solve the high-dimensional linear regression task for **compressed sensing**.

Outline

1. AMP algorithm for the LASSO.
   - Derivation from message passing
   - Comparison to other LASSO solvers
2. General AMP algorithms.
3. State evolution and performance guarantees.
4. Generalizations and extensions of AMP.

Algorithmic Challenges

Want to solve:

$$\hat{\beta} = \underset{\beta \in \mathbb{R}^N}{\arg\min} \ \|y - A\beta\|^2 + \lambda \|\beta\|_1$$

Convex Optimization Tools for Solving the LASSO

1. Classic Interior Point Method:
   - Usually matrix-matrix multiplication or matrix decomposition
   - Appropriate for, say, $N < 5000$
2. Homotopy Methods (e.g. LARS)
   - Use the structure of the LASSO cost
   - Appropriate for, say, $N < 50000$
3. First-order Methods
   - Low computational complexity per iteration
   - Require many iterations

Solving the LASSO

$$\hat{\beta} = \arg\min_{\beta} \|y - A\beta\|^2 + \lambda\|\beta\|_1$$

First-order methods: Iteratively generate estimates of β_0

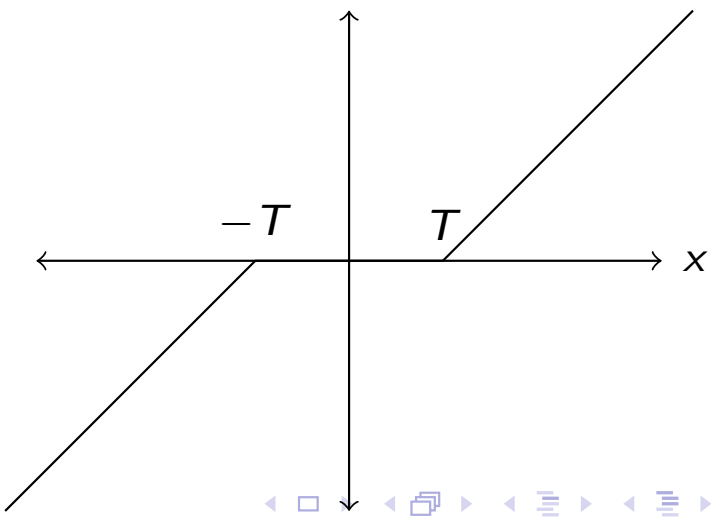
$$\beta^1, \beta^2, \ldots$$

1. Proximal Gradient (aka Iterative Soft-Thresholding)

$$r^t = y - A\beta^t$$

$$\beta^{t+1} = \eta(\beta^t + sA^T r^t;\ s\lambda)$$

$$\eta(x; T) = \begin{cases} x - T, & x \geq T \\ 0, & -T < x < T \\ x + T, & x \leq -T \end{cases}$$



Solving the LASSO

2. Proximal Gradient + Momentum (FISTA/Nesterov)

$$\text{momentum term} \qquad \tilde{\beta}^t = \beta^t + \frac{t-1}{t+2}(\beta^t - \beta^{t-1})$$

$$\text{same as IST} \qquad r^t = y - A\tilde{\beta}^t$$

$$\text{same as IST} \qquad \beta^{t+1} = \eta(\tilde{\beta}^t + sA^T r^t;\, s\lambda)$$

FISTA is good, but want faster convergence as N grows large

Can we use a *message passing* algorithm?

Assumptions

$$y = A\beta_0 + w$$

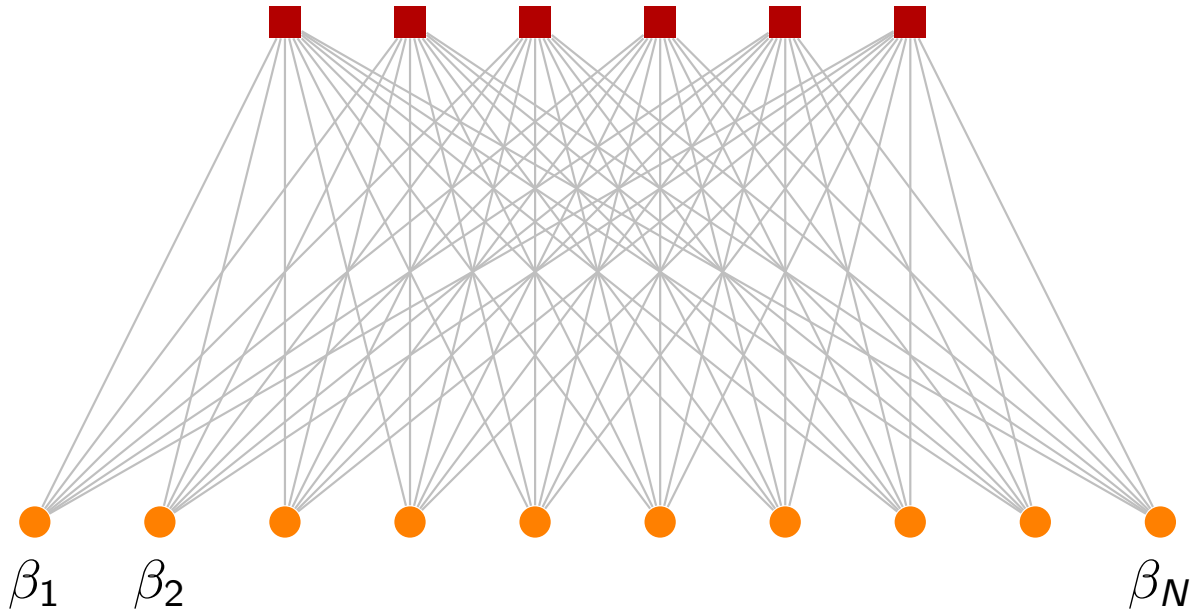
- Let us first assume that the entries of A are iid $\mathcal{N}(0, \frac{1}{m})$
- Dimensions of A: m, N large, $\quad \frac{m}{N} \to \delta \quad (\delta \text{ is } \Theta(1))$

AMP derived as approximation of loopy belief propagation for dense graphs

[Donoho-Maleki-Montanari '09], [Rangan '11], [Krzakala et al '12], [Schniter '11], . . .

Min-Sum Message Passing

Want to compute $\hat{\beta} = \arg\min_\beta \sum_{a=1}^{m} (y_a - [A\beta]_a)^2 + \lambda \sum_{i=1}^{N} |\beta_i|$

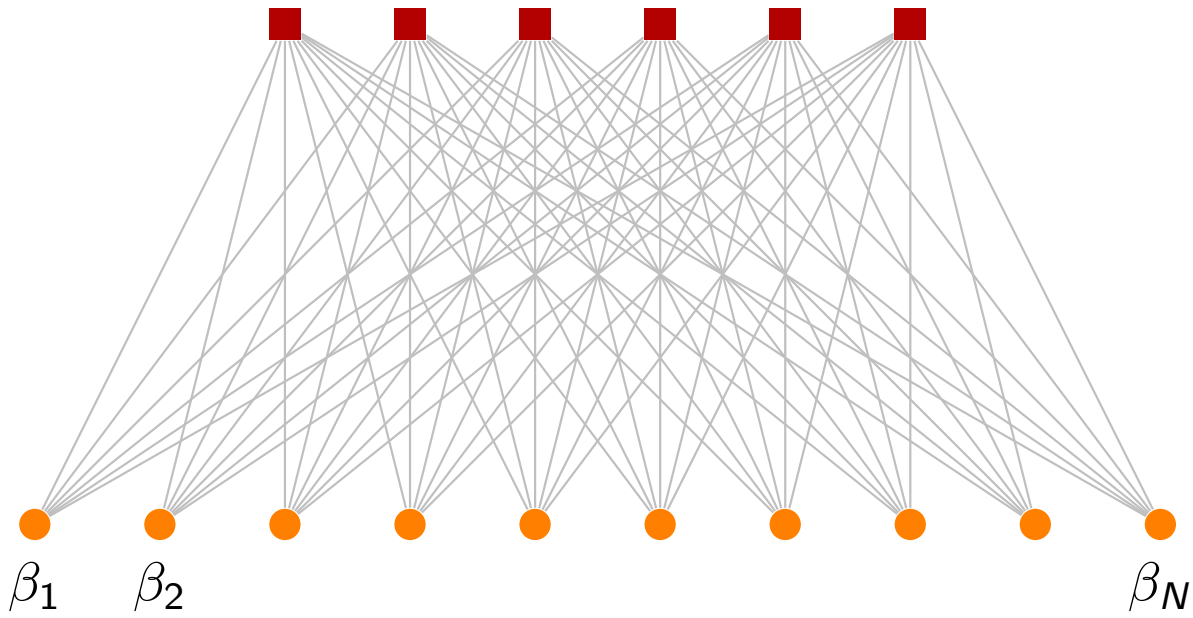


Factor Graph Representation fo LASSO Cost

- m factor nodes corresponding to $(y_1 - [A\beta]_1)^2, (y_2 - [A\beta]_2)^2, \ldots, (y_m - [A\beta]_m)^2$
- N variable nodes corresponding to $\beta_1, \beta_2, \ldots, \beta_N$
- Edge between factor node a and variable node i if $A_{a,i} \neq 0$.

Min-Sum Message Passing

Want to compute $\hat{\beta} = \arg\min_\beta \sum_{a=1}^m (y_a - [A\beta]_a)^2 + \lambda \sum_{i=1}^N |\beta_i|$

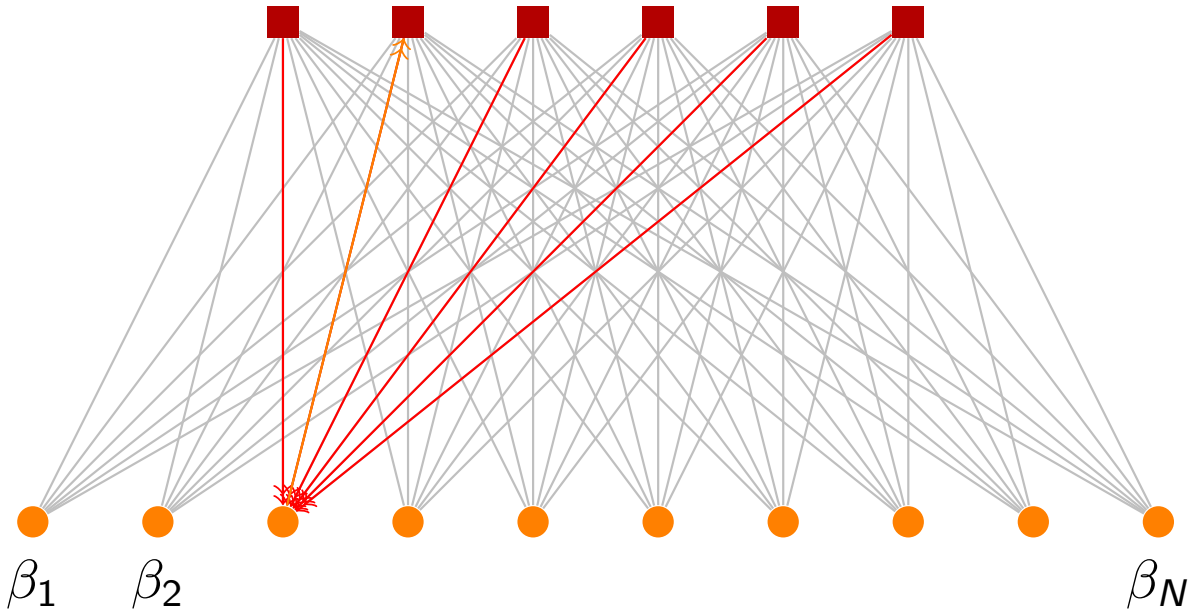


Factor Graph Representation fo LASSO Cost

- Min-sum – popular optimization algorithm for graph-structured cost
- AMP derived from min-sum on above graph

Min-Sum Message Passing

Want to compute $\hat{\beta} = \arg\min_\beta \sum_{a=1}^{m} (y_a - [A\beta]_a)^2 + \lambda \sum_{i=1}^{N} |\beta_i|$

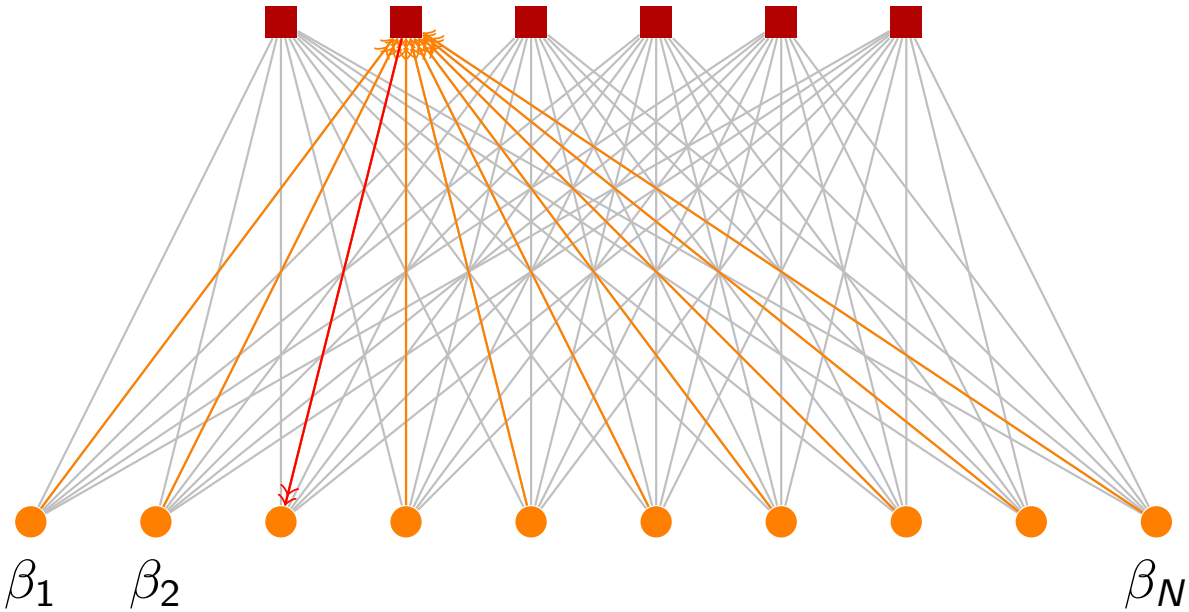


For $i = 1, \ldots, N$ and $a = 1, 2, \ldots, m$:

$$M_{i\to a}^{t}(\beta_i) = \lambda|\beta_i| + \sum_{b\in[m]\setminus a} \hat{M}_{b\to i}^{t-1}(\beta_i)$$

Min-Sum Message Passing

Want to compute $\hat{\beta} = \arg\min_\beta \sum_{a=1}^{m} (y_a - [A\beta]_a)^2 + \lambda \sum_{i=1}^{N} |\beta_i|$



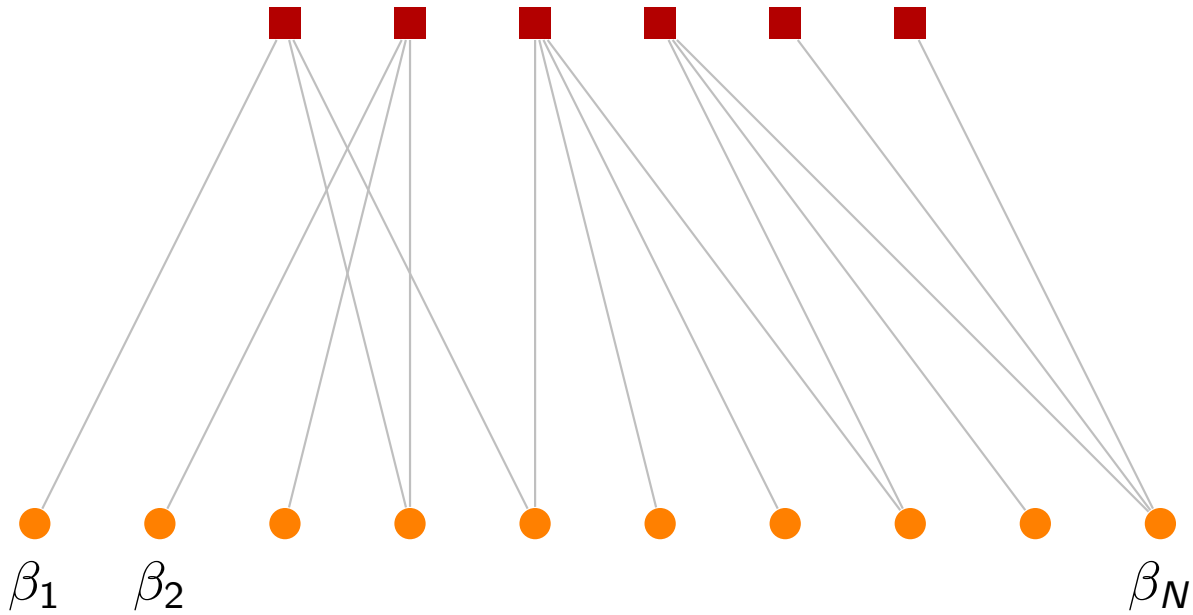
For $i = 1, \ldots, N$ and $a = 1, 2, \ldots, m$:

$$M^t_{i \to a}(\beta_i) = \lambda |\beta_i| + \sum_{b \in [m] \setminus a} \hat{M}^{t-1}_{b \to i}(\beta_i)$$

$$\hat{M}^t_{a \to i}(\beta_i) = \min_{\beta \setminus \beta_i} \left[(y_a - [A\beta]_a)^2 + \sum_{j \in [N] \setminus i} M^t_{j \to a}(\beta_j) \right]$$

Why Min-sum?

- Vast literature justifying and studying use of min-sum. For example [Murphy, Weiss, Jordan '99].
- Computes exact minimum when graph is tree (no cycles)
- Generally not guaranteed to converge on 'loopy' graphs
- Nonetheless works well in some 'loopy' applications (coding, machine vision, compressed sensing, ...)



Why Min-sum?

- Vast literature justifying and studying use of min-sum. For example [Murphy, Weiss, Jordan '99].
- Computes exact minimum when graph is tree (no cycles)
- Generally not guaranteed to converge on 'loopy' graphs
- Nonetheless works well in some 'loopy' applications (coding, machine vision, compressed sensing, ...)

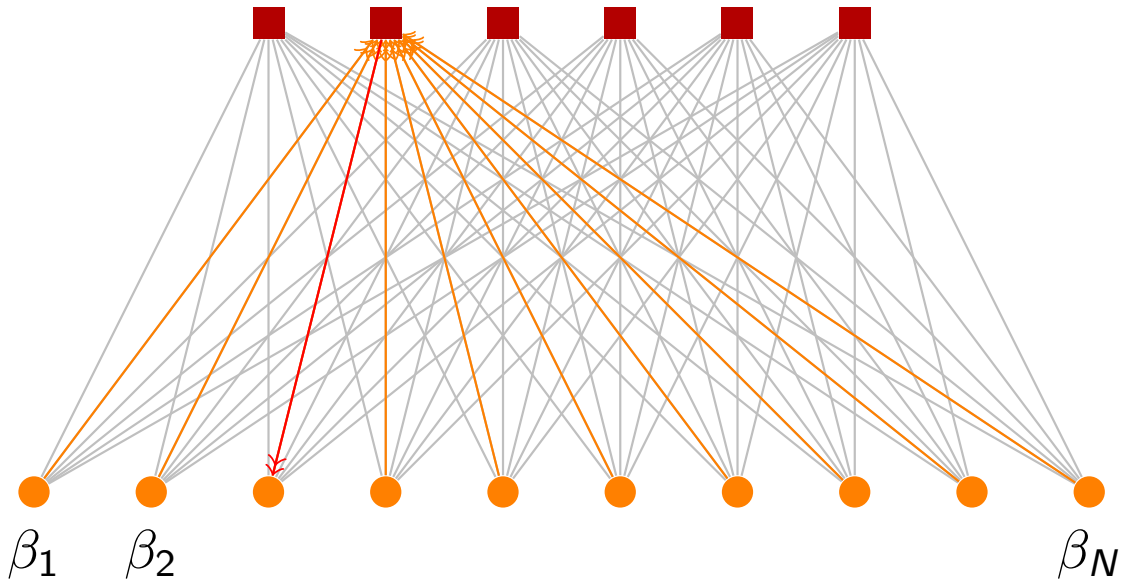
Further Limitations

But computing these messages is infeasible:

— Each message needs to be computed for all $\beta_i \in \mathbb{R}$

— There are mN such messages

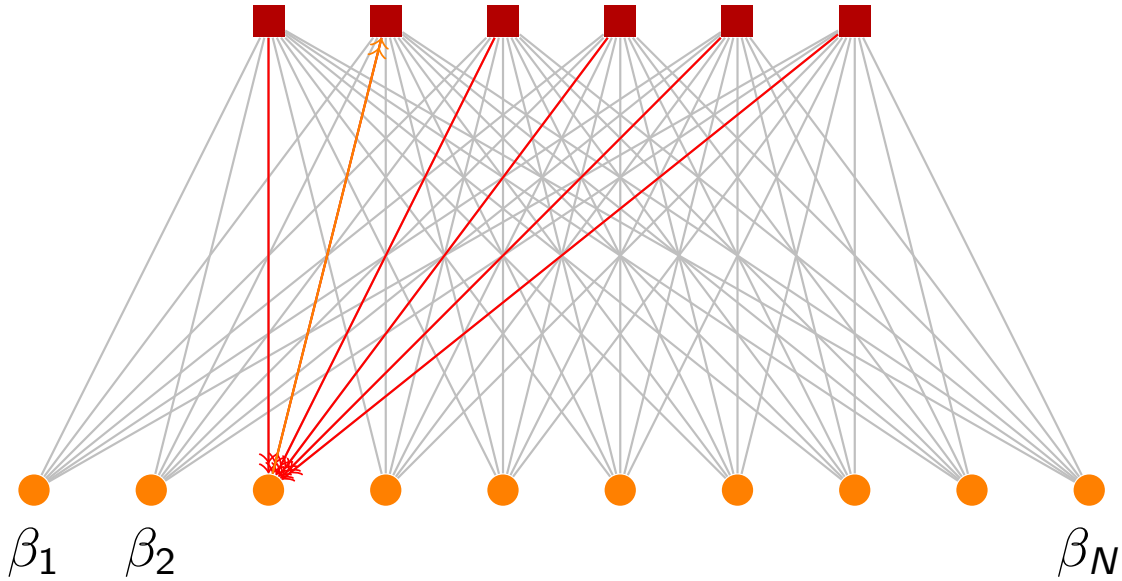
Quadratic Approximation of Messages



Messages approximated by two *numbers* (via quadratic approximation):

$$r^t_{a\to i} = y_a - \sum_{j\in[N]\setminus i} A_{aj}\,\beta^t_{j\to a}$$

Quadratic Approximation of Messages



Messages approximated by two *numbers* (via quadratic approximation):

$$r^t_{a\to i} = y_a - \sum_{j\in[N]\setminus i} A_{aj}\,\beta^t_{j\to a} \qquad \beta^{t+1}_{i\to a} = \eta\left(\sum_{b\in[m]\setminus a} A_{bi}\,r^t_{b\to i}\,;\,\theta_t\right)$$

We still have mN messages in each step . . .

$$r^t_{a\to i} = y_a - \sum_{j\in[N]} A_{aj}\beta^t_{j\to a} + A_{ai}\beta^t_{i\to a}$$

$$\beta^{t+1}_{i\to a} = \eta\left(\sum_{b\in[k]} A_{bi} r^t_{b\to i} - A_{ai} r^t_{a\to i}\,;\,\theta_t\right)$$

- Weak dependence between messages and target indices
- Neglecting dependence altogether gives IST
- A more careful analysis, using Taylor approximations ...

The AMP algorithm

AMP iteratively produces estimates $\beta^0 = 0, \beta^1, \ldots, \beta^t, \ldots$

$$r^t = y - A\beta^t + \frac{r^{t-1}}{m}\|\beta^t\|_0$$
$$\beta^{t+1} = \eta(\beta^t + A^T r^t\,;\, \theta_t)$$

- r^t is the 'modified residual' after step t
- η denoises the *effective observation* to produce β^{t+1}

The AMP algorithm

AMP iteratively produces estimates $\beta^0 = 0, \beta^1, \ldots, \beta^t, \ldots$

$$r^t = y - A\beta^t + \frac{r^{t-1}}{m}\|\beta^t\|_0$$
$$\beta^{t+1} = \eta(\beta^t + A^T r^t; \theta_t)$$

- r^t is the 'modified residual' after step t
- η denoises the *effective observation* to produce β^{t+1}

Compare to Iterative Soft-Thresholding

$$r^t = y - A\beta^t$$
$$\beta^{t+1} = \eta(\beta^t + sA^T r^t; s\lambda)$$

The AMP algorithm

AMP iteratively produces estimates $\beta^0 = 0, \beta^1, \ldots, \beta^t, \ldots$

$$r^t = y - A\beta^t + \frac{r^{t-1}}{m}\|\beta^t\|_0$$
$$\beta^{t+1} = \eta(A^T r^t + \beta^t ; \theta_t)$$

With the assumptions:

- Entries of A are iid $\mathcal{N}(0, \frac{1}{m})$
- Dimensions of A: m, N large, $\quad \frac{m}{N} \to \delta \quad$ (δ is constant)

The *momentum* term in r^t ensures that asymptotically

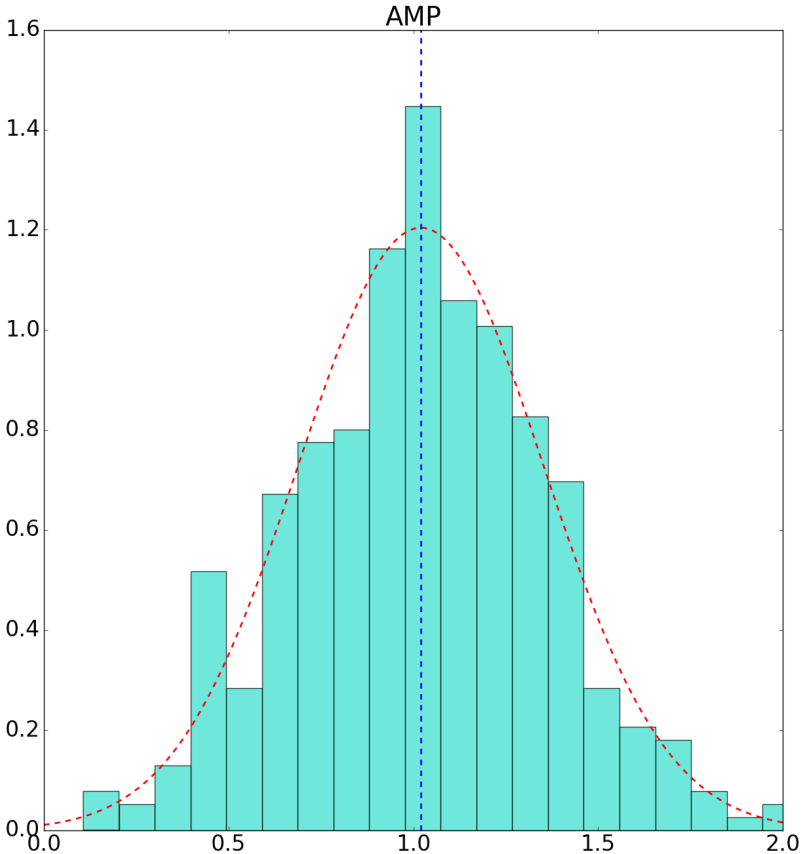
$$A^T r^t + \beta^t \approx \beta_0 + \tau_t Z \quad \text{where } Z \text{ is } \mathcal{N}(0,1)$$

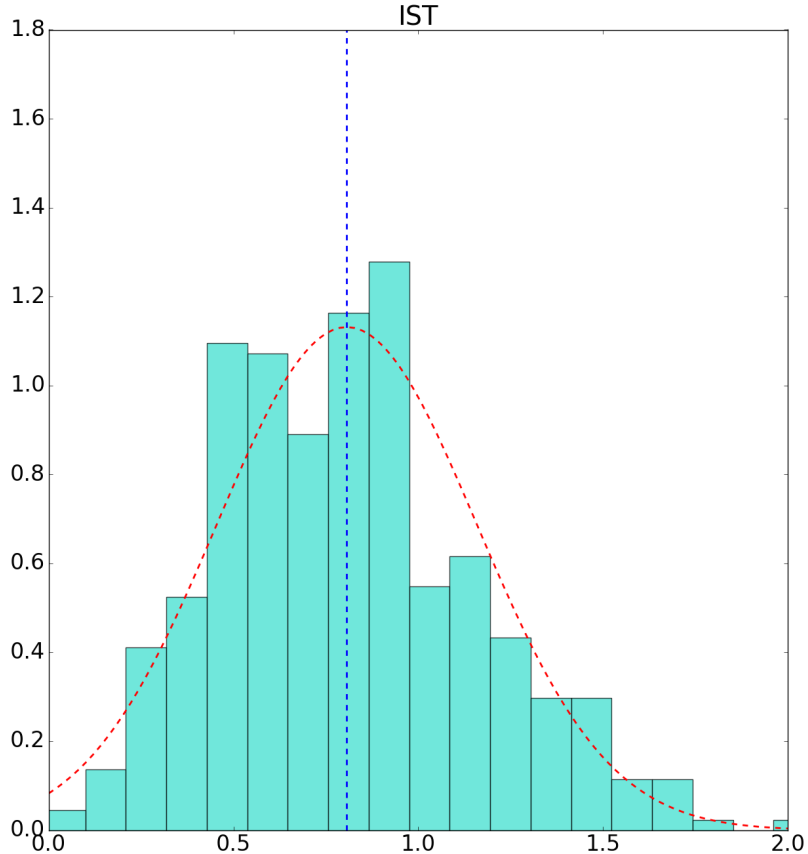
$\Rightarrow$ The *effective observation* $A^T r^t + \beta^t$ is true signal observed in independent Gaussian noise.

Example: $y = A\beta_0$

$A : m \times N = 2000 \times 4000$; β_0 has 500 non-zeros $\sim$ iid unif ± 1

Histogram of $A^T r^t + \beta^t$ at $t = 10$





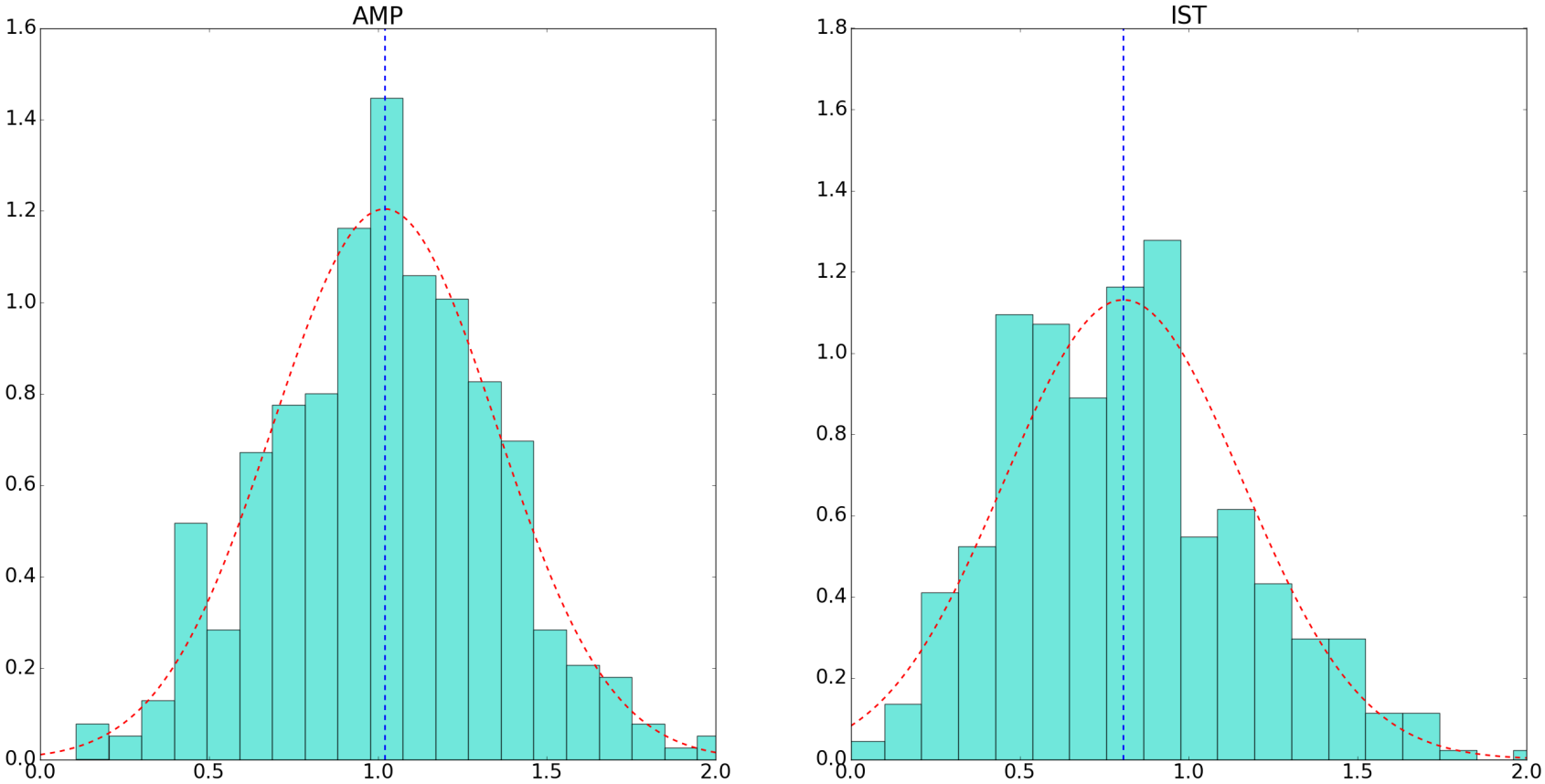
with $r^t = y - A^T \beta^t + r^{t-1} \frac{\|\beta^t\|_0}{m}$

with $r^t = y - A^T \beta^t$

Example: $y = A\beta_0$

$A : m \times N = 2000 \times 4000$; β_0 has 500 non-zeros $\sim$ iid unif ± 1

Histogram of $A^T r^t + \beta^t$ at $t = 10$

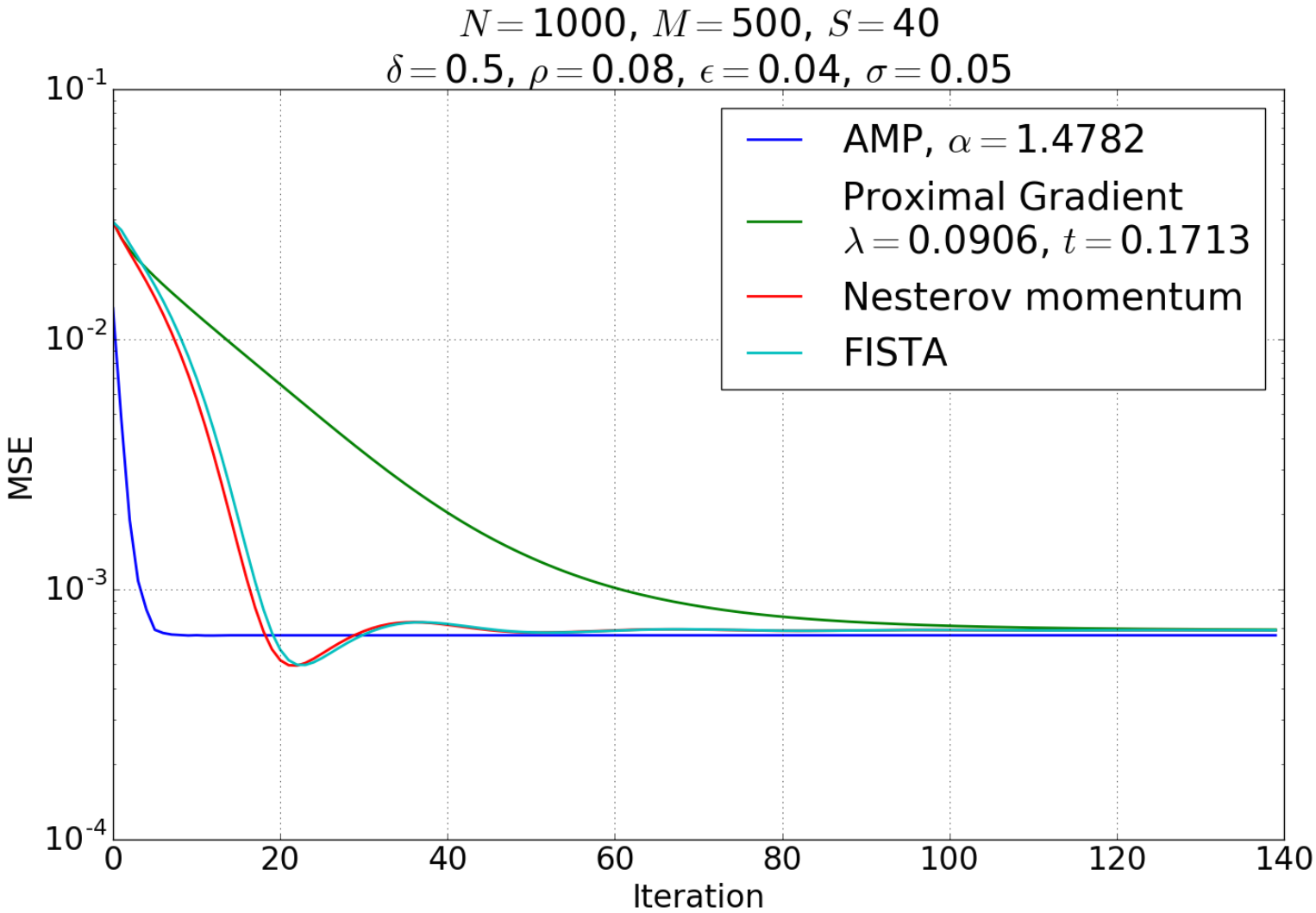


– Here: empirical observation at a single t for specific m, N

– Later: rigorous proof that statistical properties exact in limit of m, N for all t

AMP vs the Rest

$$y = A\beta_0 + w, \quad w \text{ iid } \sim \mathcal{N}(0, \sigma^2), \quad MSE = \frac{1}{N}\|\beta^t - \beta_0\|^2$$



General AMP Framework

In the talk so far:

- Sparse signal (unknown signal prior distribution)
- Goal to minimize LASSO cost

Generalization:

- Known signal prior distribution (sparsity-inducing or not)
- Goal to minimize mean squared error (MSE)

General AMP Framework

In the talk so far:

- Sparse signal (unknown signal prior distribution)
- Goal to minimize LASSO cost

Generalization:

- Known signal prior distribution (sparsity-inducing or not)
- Goal to minimize mean squared error (MSE)

Let $y = A\beta_0 + w, \qquad \beta_0 \text{ iid} \sim p_\beta, \qquad w \text{ iid} \sim \mathcal{N}(0, \sigma^2)$

$$r^t = y - A\beta^t + \frac{r^{t-1}}{m} \sum_{i=1}^{N} \eta'_{t-1}(A^T r^{t-1} + \beta^{t-1})_i$$

$$\beta^{t+1} = \eta_t(A^T r^t + \beta^t)$$

Function η_t chosen to denoise *effective observation* producing β^{t+1}

Choosing $\eta_t(\cdot)$

Let $\quad y = A\beta_0 + w, \quad \beta_0 \text{ iid} \sim p_\beta, \quad w \text{ iid} \sim \mathcal{N}(0, \sigma^2)$

$$r^t = y - A\beta^t + \frac{r^{t-1}}{m} \sum_{i=1}^{N} \eta'_{t-1}(A^T r^{t-1} + \beta^{t-1})_i$$

$$\beta^{t+1} = \eta_t(A^T r^t + \beta^t)$$

KEY: For large m, N, at each time step t

$$A^T r^t + \beta^t \approx \beta_0 + \tau_t Z \quad \text{where } Z \text{ is } \mathcal{N}(0,1)$$

- p_β known: Bayes-optimal η_t choice minimizes $\mathbb{E}\|\beta_0 - \beta^{t+1}\|^2$. Equals

$$\eta_t(s) = \mathbb{E}[\beta_0 \mid \beta_0 + \tau_t Z = s]$$

- p_β unknown: partial knowledge about β_0 can guide η_t choice.

General AMP Framework

To summarize:

LASSO:

- Sparse signal (unknown signal prior distribution)
- Goal to minimize LASSO cost
- Use denoiser $\eta(\cdot)$ as soft-threshold

Generalization:

- Known signal prior distribution (sparsity-inducing or not)
- Goal to minimize mean squared error (MSE)
- Use denoiser $\eta_t(s) = \mathbb{E}[\beta_0 \,|\, \beta_0 + \mathcal{N}(0, \tau_t) = s]$.

In both cases, $A^T r^t + \beta^t \approx \beta_0 + \mathcal{N}(0, \tau_t)$.

Choice of denoiser determines the type of problem AMP solves.

The Modified Residual [Donoho-Maleki-Montanari '09]

Assume $A_{ij} \sim \mathcal{N}(0, 1/m)$ and $w_i \sim \mathcal{N}(0, \sigma^2)$.

Suppose instead,

$$r^t = y - A\beta^t$$

The Modified Residual [Donoho-Maleki-Montanari '09]

Assume $A_{ij} \sim \mathcal{N}(0, 1/m)$ and $w_i \sim \mathcal{N}(0, \sigma^2)$.

Suppose instead,

$$r^t = y - A\beta^t$$

Then effective observation:

$$\beta^t + A^T r^t \;=\; \beta_0 \;+\; A^T w \;+\; (\mathrm{I} - A^T A)(\beta_0 - \beta^t)$$

The Modified Residual [Donoho-Maleki-Montanari '09]

Assume $A_{ij} \sim \mathcal{N}(0, 1/m)$ and $w_i \sim \mathcal{N}(0, \sigma^2)$.

Suppose instead,

$$r^t = y - A\beta^t$$

Then effective observation:

$$\beta^t + A^T r^t = \beta_0 + \underbrace{A^T w}_{\approx \mathcal{N}(0,\sigma^2)} + (\mathrm{I} - A^T A)(\beta_0 - \beta^t)$$

The Modified Residual [Donoho-Maleki-Montanari '09]

Assume $A_{ij} \sim \mathcal{N}(0, 1/m)$ and $w_i \sim \mathcal{N}(0, \sigma^2)$.

Suppose instead,

$$r^t = y - A\beta^t$$

Then effective observation:

$$\beta^t + A^T r^t = \beta_0 + \underbrace{A^T w}_{\approx \mathcal{N}(0,\sigma^2)} + \underbrace{(\mathrm{I} - A^T A)}_{\approx \mathcal{N}(0,1/m)}(\beta_0 - \beta^t)$$

The Modified Residual [Donoho-Maleki-Montanari '09]

Assume $A_{ij} \sim \mathcal{N}(0, 1/m)$ and $w_i \sim \mathcal{N}(0, \sigma^2)$.

Suppose instead,

$$r^t = y - A\beta^t$$

Then effective observation:

$$\beta^t + A^T r^t = \beta_0 + \underbrace{A^T w}_{\approx \mathcal{N}(0,\sigma^2)} + \underbrace{(\mathrm{I} - A^T A)}_{\approx \mathcal{N}(0,1/m)}(\beta_0 - \beta^t)$$

$$\approx \beta_0 + \sqrt{\sigma^2 + \frac{\mathbb{E}\|\beta_0 - \beta^t\|^2}{m}} Z$$

The Modified Residual [Donoho-Maleki-Montanari '09]

Assume $A_{ij} \sim \mathcal{N}(0, 1/m)$ and $w_i \sim \mathcal{N}(0, \sigma^2)$.

Suppose instead,

$$r^t = y - A\beta^t \; + \; \frac{r^{t-1}}{m} \sum_{i=1}^{N} \eta_t' \left([A^T r^{t-1} + \beta^{t-1}]_i \right)$$

Then effective observation:

$$\beta^t + A^T r^t \; = \; \beta_0 \; + \; \underbrace{A^T w}_{\approx \mathcal{N}(0,\sigma^2)} \; + \; \underbrace{(\mathrm{I} - A^T A)}_{\approx \mathcal{N}(0,1/m)} (\beta_0 - \beta^t)$$

$$\approx \beta_0 + \sqrt{\sigma^2 + \frac{\mathbb{E}\|\beta_0 - \beta^t\|^2}{m}} Z$$

State Evolution

Define τ_t^2 as noise variance in the effective observation after step t.

$$\beta^t + A^T r^t \approx \beta_0 + \tau_t Z, \quad Z \sim \mathcal{N}(0, \mathbb{I}).$$

If $\tau_1, \tau_2, \ldots$ is decreasing, getting a more 'pure' view of β_0 as algorithm iterates.

State Evolution

Define τ_t^2 as noise variance in the effective observation after step t.

$$\beta^t + A^T r^t \approx \beta_0 + \tau_t Z, \quad Z \sim \mathcal{N}(0, \mathbb{I}).$$

If $\tau_1, \tau_2, \ldots$ is decreasing, getting a more 'pure' view of β_0 as algorithm iterates.

SE Equations

Set $\tau_0^2 = \sigma^2 + \frac{\mathbb{E}\|\beta\|^2}{m}$,

$$\tau_t^2 = \sigma^2 + \frac{\mathbb{E}\|\beta - \beta^t\|^2}{m} = \sigma^2 + \frac{\mathbb{E}\|\beta - \eta_t(\beta + \tau_{t-1}Z)\|^2}{m}$$

$Z \sim \mathcal{N}(0,1)$ independent of $\beta \sim p_\beta$.

State evolution is a scalar recursion that allows us to predict the performance of AMP at every iteration!

Assumptions for Performance Guarantees

We make the following assumptions:

- **Measurement matrix**: i.i.d. $\sim \mathcal{N}(0, 1/m)$.
- **Signal**: i.i.d. $\sim p_\beta$, sub-Gaussian.
- **Measurement noise**: i.i.d. $\sim p_W$, sub-Gaussian, $\mathbb{E}[w_i^2] = \sigma^2$.
- **De-noising Functions** η_t: Lipschitz continuous with weak derivative η_t' which is differentiable except possibly at a finite number of points, with bounded derivative everywhere it exists.

Performance Guarantees

Theorem (Rush, Venkataramanan '16)

Under the assumptions of the previous slide, with constants K_t, κ_t, for $\Delta \in (0,1)$ and $t \geq 0$,

$$P\left(\left|\frac{1}{m}\|\beta^{t+1} - \beta_0\|^2 - (\tau_{t+1}^2 - \sigma^2)\right| \geq \Delta\right) \leq K_t e^{-\kappa_t N \Delta^2}.$$

Constants in the Bound:

- Constants $K_t = K_1(K_2)^t(t!)^{10}$ and $\kappa_t = \kappa_1 \kappa_2^{-t}(t!)^{-18}$ where $K_1, K_2, \kappa_1, \kappa_2 > 0$ are universal constants.
- Indicates how large t can get for deviation prob. $\to 0$: $t = o\left(\frac{\log N}{\log \log N}\right)$

Performance Guarantees

Theorem (Rush, Venkataramanan '16)

Under the assumptions of the previous slide, with constants K_t, κ_t, for $\Delta \in (0,1)$ and $t \geq 0$,

$$P\left(\left|\frac{1}{m}\|\beta^{t+1} - \beta_0\|^2 - (\tau_{t+1}^2 - \sigma^2)\right| \geq \Delta\right) \leq K_t e^{-\kappa_t N \Delta^2}.$$

- Result holds for more general class of loss functions (beyond MSE).
- Refines an asymptotic result proved by Bayati, Montanari [Trans. IT '11]
- The finite-sample result above implies the asymptotic result (via Borel-Cantelli), i.e. with $\delta = m/N$

$$\lim_{N\to\infty} \frac{1}{N}\|\beta^{t+1} - \beta_0\|^2 \overset{a.s.}{=} \delta(\tau_{t+1}^2 - \sigma^2).$$

Back to LASSO

It can be shown [Bayati, Montanari '12],

$$\lim_{t\to\infty} \lim_{N\to\infty} \frac{1}{N}\|\beta^t - \hat{\beta}\|^2 \overset{a.s.}{=} 0,$$

for $\hat{\beta}$, the LASSO minimizer, and β^t, the AMP estimate at time t.

(AMP threshold has one-to-one map with LASSO parameter λ. Assumes i.i.d. Gaussian A.)

Moreover, AMP performance guarantees with the above imply an asymptotic result for the LASSO minimizer:

$$\lim_{N\to\infty} \frac{1}{N}\|\beta_0 - \hat{\beta}\|^2 \overset{a.s.}{=} \delta(\tau_*^2 - \sigma^2),$$

where $\tau_*^2 = \lim_{t\to\infty} \tau_t^2$ with τ_t given by state evolution.

Proof Idea of Performance Guarantees

Show $\beta^t + A^T r^t \sim \beta_0 + \tau_t Z$, with τ_t given by state evolution.

Proof Idea of Performance Guarantees

Show $\beta^t + A^T r^t \sim \beta_0 + \tau_t Z$, with τ_t given by state evolution.

Steps:

1. Characterize the conditional distribution of the effective observation and residual as sum of i.i.d. Gaussians plus deviation term.

Show:

$$(\beta^t + A^T r^t - \beta_0)|_{\{\text{past}, \beta_0, w\}} \stackrel{d}{=} \tau_t Z_t + \Delta_t,$$

$$(r^t - w)|_{\{\text{past}, \beta_0, w\}} \stackrel{d}{=} \sqrt{\tau_t^2 - \sigma^2}\, \tilde{Z}_t + \tilde{\Delta}_t,$$

Proof Idea of Performance Guarantees

Show $\beta^t + A^T r^t \sim \beta_0 + \tau_t Z$, with τ_t given by state evolution.

Steps:

1. Characterize the conditional distribution of the effective observation and residual as sum of i.i.d. Gaussians plus deviation term.

Show:

$$(\beta^t + A^T r^t - \beta_0)|_{\{\text{past}, \beta_0, w\}} \stackrel{d}{=} \tau_t Z_t + \Delta_t,$$
$$(r^t - w)|_{\{\text{past}, \beta_0, w\}} \stackrel{d}{=} \sqrt{\tau_t^2 - \sigma^2}\, \tilde{Z}_t + \tilde{\Delta}_t,$$

2. Concentration results show the deviation terms are small with high probability.

AMP Extensions/Generalizations

- **Non-Gaussian noise distributions (GAMP)** [Rangan '11]
- **Different measurement matrices:**
  - **Sub-Gaussian** [Bayati, Lelarge, Montanari '15]
  - **Right orthogonally-invariant (VAMP)** [Schniter, Rangan, Fletcher '16, '17]
  - **Spatially-coupled (for improved MSE performance)** [Donoho, Javanmard, Montanari '13]
- **Signals with dependent entries and non-separable denoisers** [Ma, Rush, Baron '17], [Berthier, Montanari, Nguyen '17]

AMP Extensions/Generalizations

Different measurement models:

- **Bilinear Models** [Parker, Schniter, Cevhar '14]
  - **Multiple Measurement Vectors** [Ziniel, Schniter '13]
  - **Matrix Factorization** [Kabashima, Krzakala, Mézard, Sakata, Zdeborová '16]
  - **Blind Deconvolution**
- **Low-rank Matrix Estimation** [Rangan, Fletcher '12], [Lesieur, Krzakala, Zdeborová '15]
  - **Principle Component Analysis** [Deshpandre, Montanari '14], [Montanari, Richard '16]
  - **Stochastic Block Model** [Deshpandre, Abbe, Montanari '16]
  - **Replica Method** [Barbier, Dia, Macris, Krzakala, Lesieur, Zdeborová '15]

AMP Summary

$$y = A\beta_0 + w$$

$$r^t = y - A\beta^t + \frac{r^{t-1}}{m} \sum_{i=1}^{N} \eta'_{t-1}(A^T r^{t-1} + \beta^{t-1})_i$$

$$\beta^{t+1} = \eta_t(A^T r^t + \beta^t)$$

AMP: First-order iterative algorithm

- Theory assumes Gaussian A and iid/exchangeable p_β
- Sharp theoretical guarantees determined by simple scalar iteration. E.g.,

$$\frac{1}{N}\|\beta_0 - \beta^{t+1}\|^2 \approx \delta(\tau_{t+1}^2 - \sigma^2)$$

- AMP can be run even without knowing p_β
  (our result shows that τ_t^2 concentrates on $\|r^t\|^2/m$)
- Knowing p_β can help choose a good denoiser η_t

Open Questions

- Theoretical results for general A matrices
  (iid uniform Bernoulli, partial DFT, ...)
- Connections between AMP and classical optimization techniques

AMP

$$r^t = y - A\beta^t + r^{t-1}\frac{\|\beta^t\|_0}{m}$$
$$\beta^{t+1} = \eta(A^T r^t + \beta^t;\, \alpha\tau_t)$$

Nesterov/FISTA

$$\tilde{\beta}^t = \beta^t + \frac{t-1}{t+2}(\beta^t - \beta^{t-1})$$
$$r^t = y - A\tilde{\beta}^t$$
$$\beta^{t+1} = \eta(\tilde{\beta}^t + sA^T r^t;\, s\lambda)$$