Approximate Message Passing for Statistical Inference and Estimation

Cynthia Rush, Columbia University

Joint work with Ramji Venkataramanan (University of Cambridge)

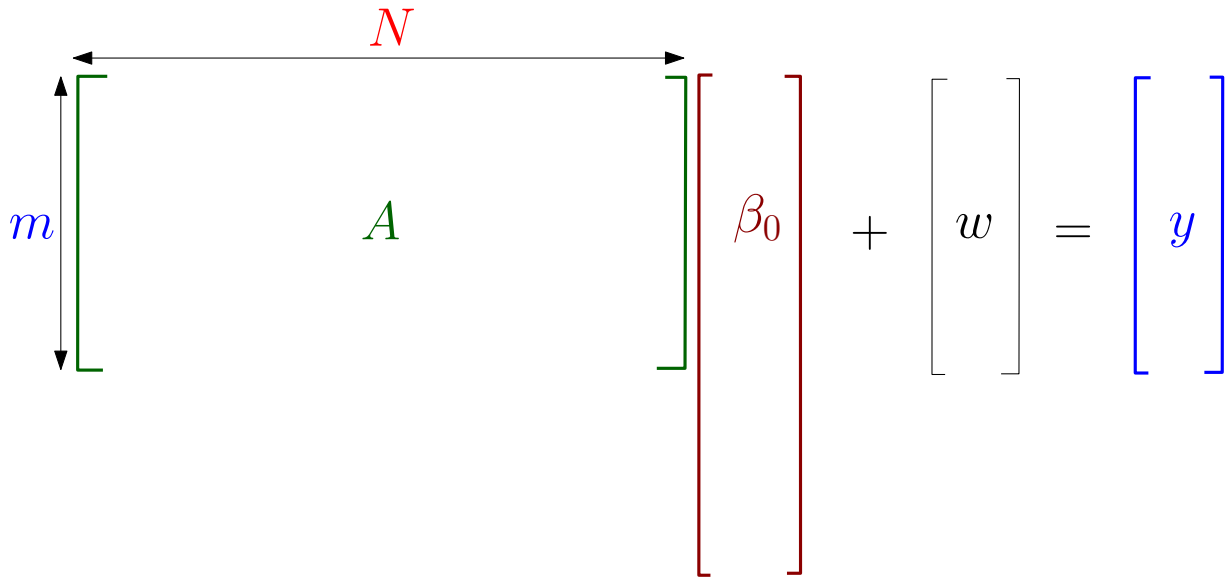
International Centre for Theoretical Sciences
Discussion Meeting on Statistical Physics of Machine Learning
January, 2020

High-dimensional Linear Regression

These days, fields like natural sciences, engineering, and social sciences have lots of data which they use to construct more complex statistical models than ever before.

This requires new methodologies and mathematical techniques for analysis of such models.

Today I overview recent progress on one such prototypical problem in this area: **high-dimensional regression**. Specifically, I focus on finite sample analysis of approximate message passing algorithms.



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

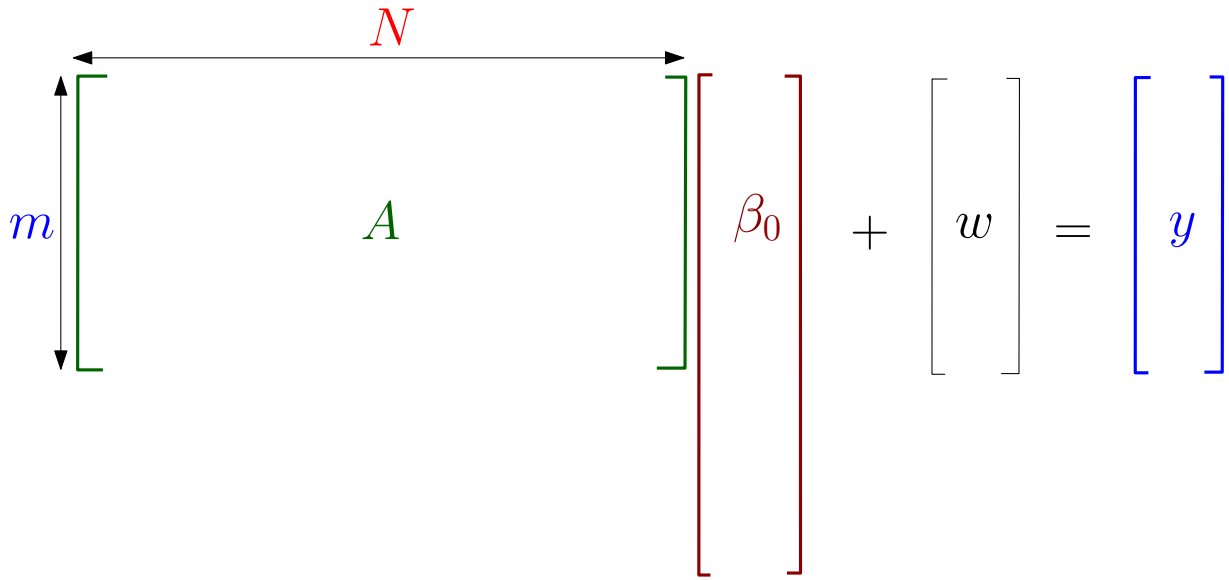
- y: length-m measurement vector
- w: length-m measurement noise
- A: $m \times N$ design matrix with $m < N$, $\frac{m}{N} \to \delta \in \Theta(1)$
- β_0: length-N unknown parameter vector (or 'signal')

Study this estimation problem but ideas extend to other settings

Many Applications

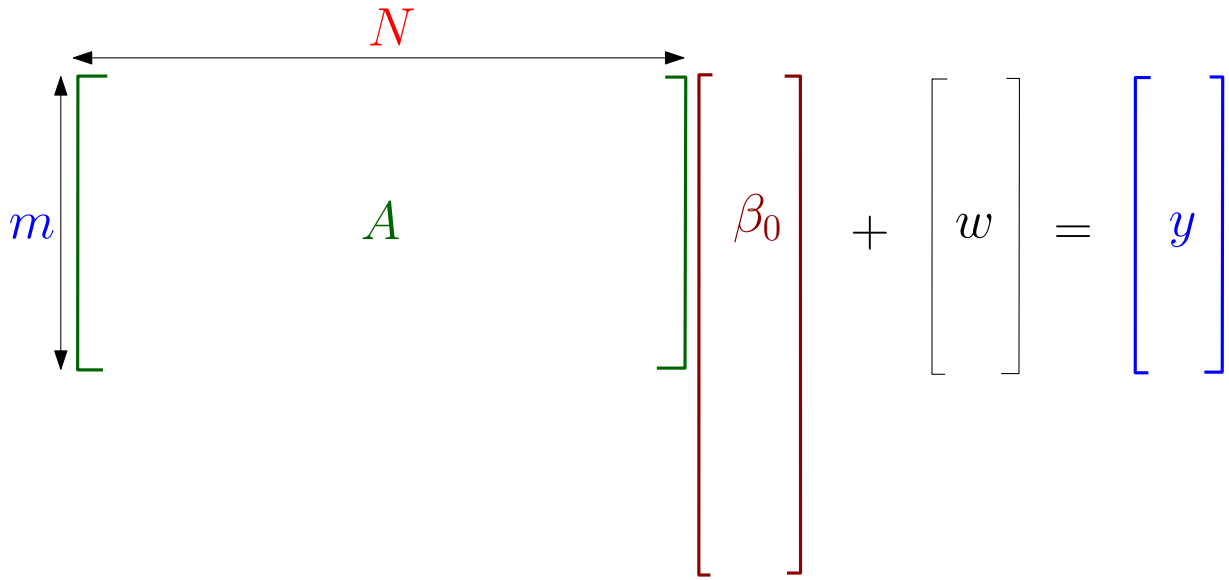
- **Imaging: Medical, Seismic, Compressive Sensing...**
  y = measurements w = sensor noise
  A = basis representation β_0 = sparse image/signal
- **Statistics/Machine Learning**
  y = experimental outcome w = model error
  A = feature data β_0 = prediction coefficients
- **Channel Coding in Communications**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

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



Often assume additional info about β_0, e.g. it has $k < N$ non-zero elements

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



Often assume additional info about β_0, e.g. it has $k < N$ non-zero elements

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

Want to solve:

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

Unfortunately, a very hard combinatorial problem.

Instead, a convex relaxation:

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

Instead, a convex relaxation:

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

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

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

[Donoho'06, Candes-Romberg-Tao'06, Bickel-Ritov-Tsybakov'09,...]

Compressed Sensing and its Challenges

Significant work beginning $\sim$2006:

- Many successful applications, e.g. MRI
- Fast algorithms
- Scaling laws for measurements required to recover 'true' sparse vector

Compressed Sensing and its Challenges

Significant work beginning $\sim$2006:

- Many successful applications, e.g. MRI
- Fast algorithms
- Scaling laws for measurements required to recover 'true' sparse vector

Challenges:

- Most analyses only provide bounds; can be conservative
- Methods specific to LASSO and don't generalize
- Lacking results related to theoretically-optimal estimate that minimizes MSE $\mathbb{E}\|\hat{\beta} - \beta_0\|^2$(assuming a known prior on β_0)
  - When can we achieve this and how?
- Limited insight on the distribution of $\hat{\beta}$; needed for inference

Approximate Message Passing (AMP)

Low complexity, scalable algorithm studied extensively for solving high-dimensional linear regression in **compressed sensing**

Benefits of AMP

For certain random matrices,

- Fast convergence
- Asymptotically exact characterization
- Testable conditions for optimality

Though studied extensively for compressed sensing, theory provides insights into many more complex models

(GLMS, logistic regression, phase retrieval, multilayer models, PCA, optimization, ...)

Approximate Message Passing (AMP)

Outline

1. AMP algorithm for the LASSO.
2. General AMP algorithms.
3. State evolution and performance guarantees.

Solving the LASSO:

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

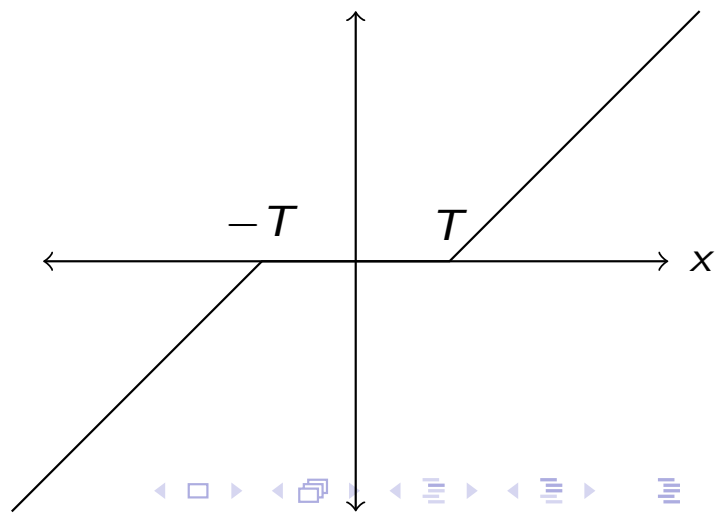
First-order methods: Iteratively generate estimates of β_0 as $\beta^1, \beta^2, \ldots$

1. Proximal Gradient (aka Iterative Soft-Thresholding)

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

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

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



2. Proximal Gradient + Momentum (FISTA/Nesterov)

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

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

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

FISTA is good, but can we use a *message passing* algorithm to address the issues raised earlier...?

- Can we get an asymptotically exact characterization?
- What is the 'optimal' estimate?
- Want faster convergence as N grows large.

Assuming:

- entries of A are iid $\mathcal{N}(0, \frac{1}{m})$
- dimensions of A are large, $\quad \frac{m}{N} \to \delta \quad (\delta \text{ is } \Theta(1))$

AMP 'derived' as approximation of loopy belief propagation (BP) for dense graphs

[Mézard '89, Opper-Winther '96, Kabashima '03, '08, Donoho-Maleki-Montanari '09, Rangan '11, Krzakala et al '12, Schniter '11, ...]

While BP used to derive, need other techniques to analyze

At a high level...

- Message passing operates on messages sent over edges of an undirected graph at time indices t.

At a high level...

- Message passing operates on messages sent over edges of an undirected graph at time indices t.
- Local rules update messages: outgoing messages from a vertex at time $t+1$ are functions of incoming messages to that vertex at time t, except for the message on the same edge.

At a high level...

- Message passing operates on messages sent over edges of an undirected graph at time indices t.
- Local rules update messages: outgoing messages from a vertex at time $t+1$ are functions of incoming messages to that vertex at time t, except for the message on the same edge.
- Describing a general class of dynamical systems; only require locality and 'non-back-tracking'.
- When the graph is a tree, such systems have many interesting properties and have found numerous applications.

(e.g. [Koller-Friedman '09, Richardson-Urbanke '08, Mézard-Montanari '09])

At a high level...

- Message passing operates on messages sent over edges of an undirected graph at time indices t.
- Local rules update messages: outgoing messages from a vertex at time $t+1$ are functions of incoming messages to that vertex at time t, except for the message on the same edge.
- Describing a general class of dynamical systems; only require locality and 'non-back-tracking'.
- When the graph is a tree, such systems have many interesting properties and have found numerous applications.

(e.g. [Koller-Friedman '09, Richardson-Urbanke '08, Mézard-Montanari '09])

AMP can be thought of as the *limit* of message passing algorithms when the underlying graph is completely dense.

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

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

- r^t is the 'modified residual' after step t
- `soft` denoises the *effective observation* to produce β^{t+1}
- θ_t chosen wrt LASSO penalty λ, changing with t

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

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

- r^t is the 'modified residual' after step t
- `soft` denoises the *effective observation* to produce β^{t+1}
- θ_t chosen wrt LASSO penalty λ, changing with t

Compare to Iterative Soft-Thresholding

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

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

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

With the assumptions:

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

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

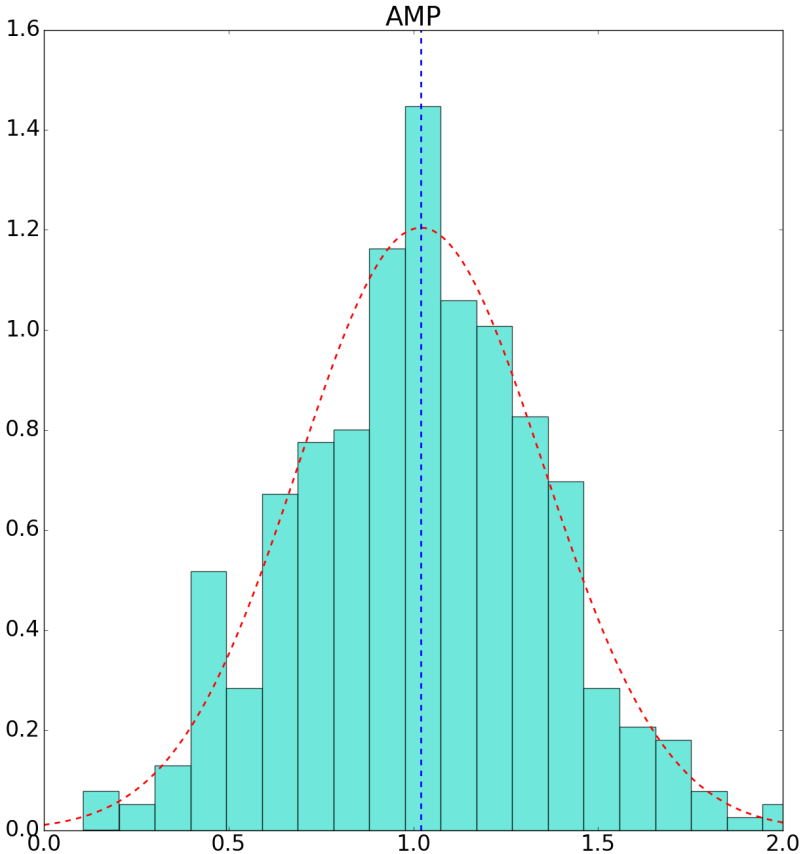
$$A^T r^t + \beta^t \approx \beta_0 + \tau_t Z \quad \text{where } Z \text{ is } \mathcal{N}(0,1)$$

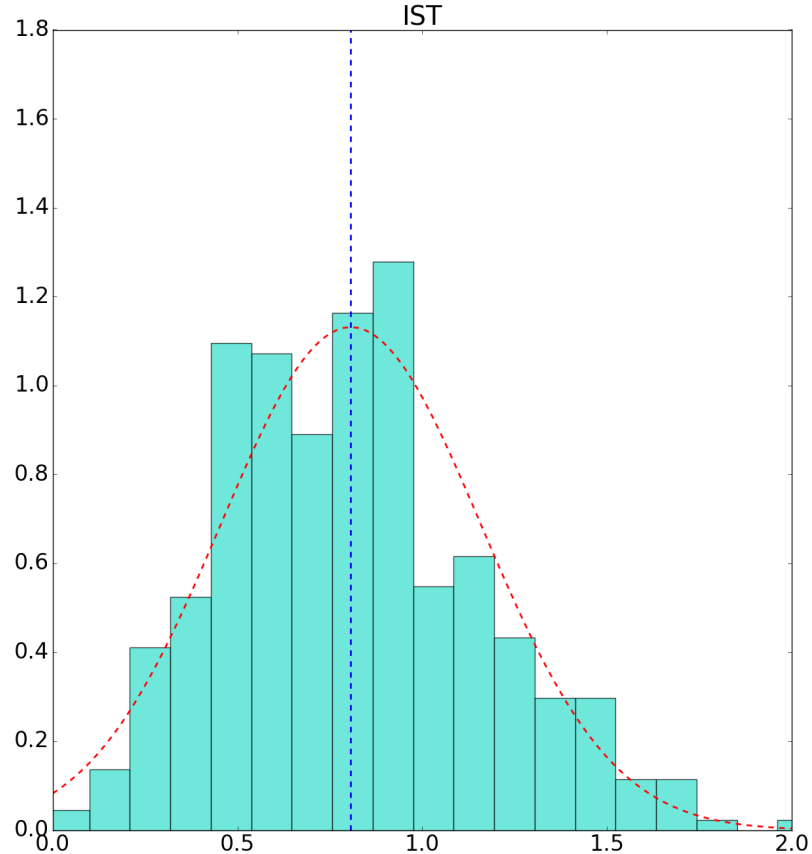
$\Rightarrow$ *Effective observation* $A^T r^t + \beta^t$ is true signal observed in independent Gaussian noise with τ_t predicted by **state evolution**.

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 indices where $\beta_0 = +1$ at $t = 10$





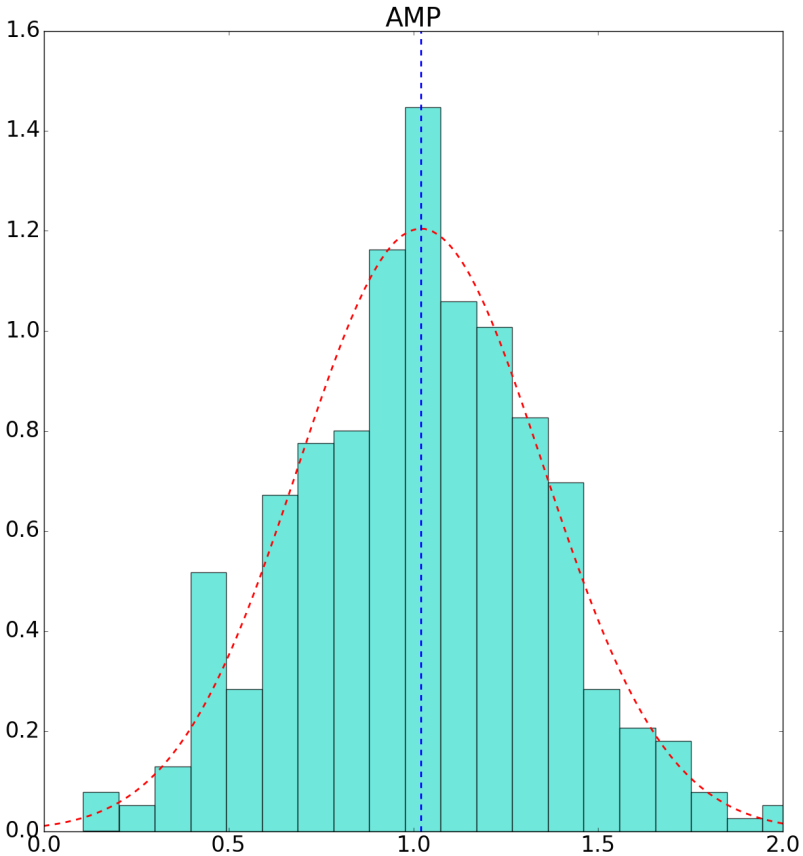
with $r^t = y - A^T \beta^t + r^{t-1} \frac{\|\beta^t\|_0}{m}$

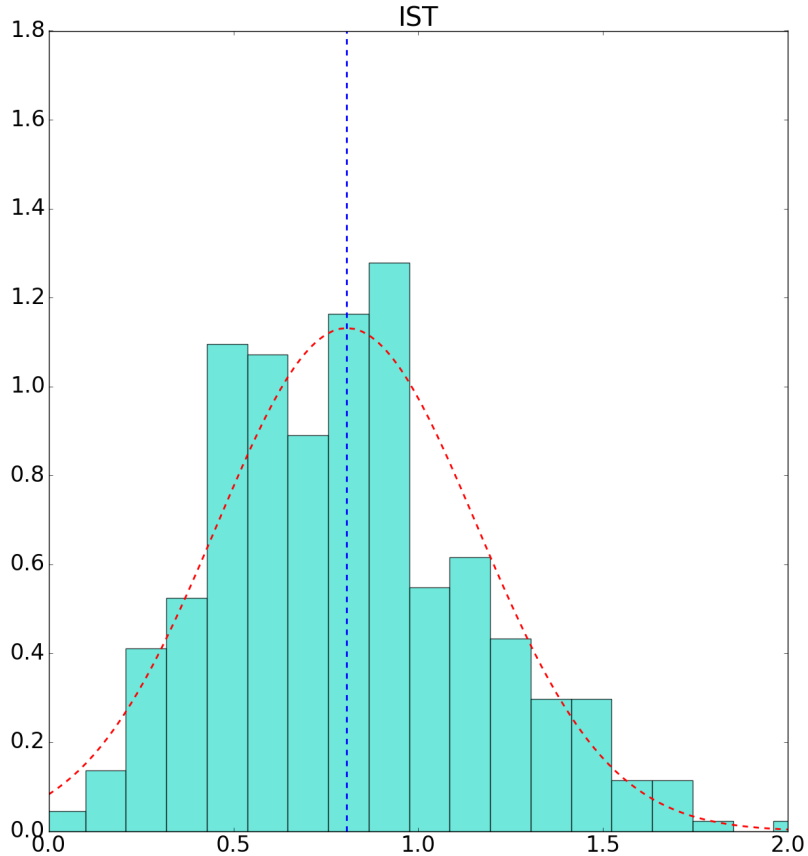
with $r^t = y - A^T \beta^t$

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 indices where $\beta_0 = +1$ 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

Theorem ((Rush '20) LASSO: exact asymptotics)

Assume:

- *Entries of A are iid* $\mathcal{N}(0, \frac{1}{m})$
- *Dimensions of A are large,* $\frac{m}{N} \to \delta$ (δ *is constant*)
- *Signal* $\beta_0 \overset{i.i.d.}{\sim} p_\beta$ *sub-Gaussian and noise* $w \overset{i.i.d.}{\sim} N(0, \sigma_w^2)$

Let α_* and τ_*^2 be the unique solution to the pair of equations

$$\lambda = \alpha_* \Big\{ 1 - \frac{1}{\delta} P\big(|\beta + \tau_* Z| > \alpha_*\big) \Big\},$$

$$\tau_*^2 = \sigma_w + \frac{1}{\delta} \mathbb{E}\Big\{ \big[soft(\beta + \tau_* Z; \alpha_*) - \beta \big]^2 \Big\}.$$

where $\beta \sim p_\beta$ independent of $Z \sim N(0, 1)$.

Theorem ((Rush '20) LASSO: exact asymptotics)

Assume:

- *Entries of A are iid $\mathcal{N}(0, \frac{1}{m})$*
- *Dimensions of A are large, $\frac{m}{N} \to \delta$ (δ is constant)*
- *Signal $\beta_0 \overset{i.i.d.}{\sim} p_\beta$ sub-Gaussian and noise $w \overset{i.i.d.}{\sim} N(0, \sigma_w^2)$*

Let α_* and τ_*^2 be the unique solution to the pair of equations

$$\lambda = \alpha_* \Big\{ 1 - \frac{1}{\delta} P\big(|\beta + \tau_* Z| > \alpha_*\big) \Big\},$$

$$\tau_*^2 = \sigma_w + \frac{1}{\delta} \mathbb{E}\Big\{ \big[soft(\beta + \tau_* Z; \alpha_*) - \beta \big]^2 \Big\}.$$

where $\beta \sim p_\beta$ independent of $Z \sim N(0,1)$. Then, for $\epsilon \in (0,1)$,

$$P\Big(\Big| \frac{1}{N} ||\hat{\beta} - \beta_0||_2^2 - \mathbb{E}\Big\{ \big[soft(\beta + \tau_* Z; \alpha_*) - \beta \big]^2 \Big\} \Big| \geq \epsilon + \delta_t \Big) \leq K K_t e^{-\kappa \kappa_t N \epsilon^2},$$

where κ_t, K_t depend on t but not n, ϵ and $\delta_t < K_1 e^{\kappa_1 t}$.

Theorem ((Rush '20) LASSO: exact asymptotics)

Assume:

- *Entries of A are iid* $\mathcal{N}(0, \frac{1}{m})$
- *Dimensions of A are large,* $\frac{m}{N} \to \delta$ (δ *is constant*)
- *Signal* $\beta_0 \overset{i.i.d.}{\sim} p_\beta$ *sub-Gaussian and noise* $w \overset{i.i.d.}{\sim} N(0, \sigma_w^2)$

Let α_* and τ_*^2 be the unique solution to the pair of equations

$$\lambda = \alpha_* \Big\{ 1 - \frac{1}{\delta} P\big(|\beta + \tau_* Z| > \alpha_*\big) \Big\},$$

$$\tau_*^2 = \sigma_w + \frac{1}{\delta} \mathbb{E}\Big\{ \big[soft(\beta + \tau_* Z; \alpha_*) - \beta \big]^2 \Big\}.$$

where $\beta \sim p_\beta$ independent of $Z \sim N(0,1)$. Then, for $\epsilon \in (0,1)$,

$$P\Big(\Big| \frac{1}{N} ||\hat{\beta} - \beta_0||_2^2 - \mathbb{E}\Big\{ \big[soft(\beta + \tau_* Z; \alpha_*) - \beta \big]^2 \Big\} \Big| \geq \epsilon + \delta_t \Big) \leq K K_t e^{-\kappa \kappa_t N \epsilon^2},$$

where κ_t, K_t depend on t but not n, ϵ and $\delta_t < K_1 e^{\kappa_1 t}$.

Theorem ((Bayati-Montanari '15) LASSO: exact asymptotics)

Assume:

- *Entries of A are iid* $\mathcal{N}(0, \frac{1}{m})$
- *Dimensions of A are large,* $\frac{m}{N} \to \delta$ (δ *is constant*)
- *Signal* $\beta_0 \overset{i.i.d.}{\sim} p_\beta$ *and noise* $w \overset{i.i.d.}{\sim} N(0, \sigma_w^2)$

Let α_* *and* τ_*^2 *be the unique solution to the pair of equations*

$$\lambda = \alpha_* \Big\{ 1 - \frac{1}{\delta} P\big(|\beta + \tau_* Z| > \alpha_*\big) \Big\},$$

$$\tau_*^2 = \sigma_w + \frac{1}{\delta} \mathbb{E}\Big\{ \big[soft(\beta + \tau_* Z; \alpha_*) - \beta \big]^2 \Big\}.$$

where $\beta \sim p_\beta$ *independent of* $Z \sim N(0,1)$. *Then,*

$$\lim_{m,N \to \infty} \frac{1}{N} ||\hat{\beta} - \beta_0||_2^2 \overset{a.s.}{=} \mathbb{E}\Big\{ \big[soft(\beta + \tau_* Z; \alpha_*) - \beta \big]^2 \Big\}.$$

Theorem ((Bayati-Montanari '15) LASSO: exact asymptotics)

Assume:

- *Entries of A are iid* $\mathcal{N}(0, \frac{1}{m})$
- *Dimensions of A are large,* $\frac{m}{N} \to \delta$ (δ *is constant*)
- *Signal* $\beta_0 \overset{i.i.d.}{\sim} p_\beta$ *and noise* $w \overset{i.i.d.}{\sim} N(0, \sigma_w^2)$

Let α_* *and* τ_*^2 *be the unique solution to the pair of equations*

$$\lambda = \alpha_* \Big\{ 1 - \frac{1}{\delta} P\big(|\beta + \tau_* Z| > \alpha_*\big) \Big\},$$

$$\tau_*^2 = \sigma_w + \frac{1}{\delta} \mathbb{E}\Big\{ \big[soft(\beta + \tau_* Z; \alpha_*) - \beta \big]^2 \Big\}.$$

where $\beta \sim p_\beta$ *independent of* $Z \sim N(0,1)$. *Then,*

$$\lim_{m,N \to \infty} \frac{1}{N} ||\hat{\beta} - \beta_0||_2^2 \overset{a.s.}{=} \mathbb{E}\Big\{ \big[soft(\beta + \tau_* Z; \alpha_*) - \beta \big]^2 \Big\}.$$

Proof requires demonstrating that the AMP algorithm just introduced converges rapidly to $\hat{\beta}$ and can be analyzed *exactly*.

This result actually applies in more generality than presented:

- 'Pseudo-Lipschitz' loss functions [Bayati-Montanari '12]
- Matrices with finite second moments and lite tails

  [Korada-Montanari '11, Bayati-Lelarge-Montanari '15, Oymak-Tropp '15]
- Beyond i.i.d. matrices: some empirics and heuristics

  [Donoho-Tanner '09, Tulino-Caire-Verdu-Shamai '13, Javanmard-Montanari '14b]
- Non-asymptotic analysis [Rush '20]

General AMP Framework

In the talk so far:

- LASSO motivated by a sparse signal (and unknown signal prior distribution)
- Goal to minimize LASSO cost
- Note: theorem doesn't require sparsity of the assumed prior. This suggests....

First generalization:

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

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

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

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

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

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

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

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

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

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.

Historically...

- With Bayesian denoisers, the fixed point version of the AMP iteration dates back to the work of Thouless, Anderson, Palmer (TAP) on mean field spin glasses.
- Iterative solutions of the TAP equations [Bolthausen '14].
- General (non-Bayesian) formulation developed and analyzed in [Donoho-Maleki-Montanari '09, Bayati-Montanari '11] motivated by applications to compressed sensing.

To summarize:

LASSO for compressed sensing:

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

First generalization:

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

In both cases, as $N \to \infty$ with t fixed, AMP admits an *asymptotically exact characterization* via **state evolution**, such that $A^T r^t + \beta^t \approx \beta_0 + \mathcal{N}(0, \tau_t)$.

Under the assumption on the effective observation:

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

If $\tau_1, \tau_2, \ldots$ is decreasing, getting a more 'pure' view of β_0 as algorithm iterates. The **state evolution** computes τ_t^2.

Under the assumption on the effective observation:

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

If $\tau_1, \tau_2, \ldots$ is decreasing, getting a more 'pure' view of β_0 as algorithm iterates. The **state evolution** computes τ_t^2.

State evolution equations

Set $\tau_0^2 = \sigma^2 + \frac{1}{m}\mathbb{E}\|\beta\|^2$,

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

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

State evolution is a scalar recursion that allows for prediction of the performance of AMP at any iteration. Now we make this rigorous.

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' that's differentiable except possibly at a finite num. of points, with bounded derivative everywhere it exists.

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' that's differentiable except possibly at a finite num. of points, with bounded derivative everywhere it exists.

Pseudo-Lipschitz (PL) Loss Functions

A function $\phi : \mathbb{R}^m \to \mathbb{R}$ is PL if there exists a constant $L > 0$ such that for all $x, y \in \mathbb{R}^m$,

$$|\phi(x) - \phi(y)| \leq L(1 + \|x\| + \|y\|)\|x - y\|.$$

E.g. $\phi(x) = \|x\|_2^2$ (squared-error loss), or $\phi(x) = \|x\|_1$ (ℓ_1 loss).

Performance Guarantees

Theorem (Rush, Venkataramanan '18)

Under the assumptions of the previous slide, for any PL function $\phi : \mathbb{R}^2 \to \mathbb{R}$, $\Delta \in (0,1)$, *and* $t \geq 0$,

$$P\Big(\Big|\frac{1}{N}\sum_{i=1}^{N} \phi(\beta_i^{t+1}, \beta_{0_i}) - \delta \mathbb{E}[\phi(\eta_t(\beta + \tau_t Z), \beta)]\Big| \geq \Delta\Big) \leq K_t e^{-\kappa_t N \Delta^2},$$

for $Z \sim \mathcal{N}(0,1)$, $\beta \sim p_\beta$ *independent with constants* K_t, κ_t.

For PL loss functions, can essentially consider AMP estimate β^{t+1} as having i.i.d. entries with each entry $\sim \eta_t(\beta + \tau_t Z)$.

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 $t \geq 0$,

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

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* $t \geq 0$,

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

- Refines the asymptotic result proved by [Bayati-Montanari '11]
- The finite-sample result above implies the asymptotic result (via Borel-Cantelli), i.e. with $\delta = m/N$

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

Performance Guarantees

Theorem (Rush, Venkataramanan '18)

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 $K_t = K_1 (K_2)^t (t!)^{K_3}$ and $\kappa_t = \kappa_1 \kappa_2^{-t} (t!)^{-\kappa_3}$ where $K_1, K_2, K_3, \kappa_1, \kappa_2, \kappa_3 > 0$ are universal constants.
- Indicates how large t can get for deviation prob. $\to 0$:

$$t = o\left(\frac{\log N}{\log \log N}\right)$$

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 conditional dist. of effective observation and residual as sum of i.i.d. Gaussians plus deviation.

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 conditional dist. of effective observation and residual as sum of i.i.d. Gaussians plus deviation.

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. Inductively show that norms of the deviation terms concentrate to zero.

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, Rush-Hsieh-Venkataramanan '18]
- **Signals with dependent entries and non-separable denoisers** [Ma-Rush-Baron '17, Berthier-Montanari-Nguyen '17]

AMP Extensions/Generalizations

Statistics Applications:

- **False discoveries in LASSO and SLOPE estimation**

  [Su-Bogdan-Candes etal '17, Bu-Klusowski-Rush-Su '20]

- **Power of knock-off variable selection** [Weinstein-Barber-Candes etal '17]

- **Exact asymptotic performance guarantees for penalized regression tasks**
  - **LASSO** [Bayati-Montanari '12]
  - **M-estimation** [Donoho-Montanari '16]
  - **SLOPE** [Bu-Klusowski-Rush-Su '19]

- **Bias and variance of the MLE for high-dimensional logistic regression** [Sur-Candes '18]

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 i.i.d. (sub)Gaussian A
- 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 prior 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 more general A matrices
  (i.i.d. uniform Bernoulli, partial DFT, correlated columns, . . .)
- AMP can diverge outside of i.i.d. Gaussian: want to know when and why
- Develop AMP methods for inference and learning in deep networks
- 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

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