Approximate Message Passing: Can it Work?

Sundeep Rangan (NYU-Poly)

Joint work with Alyson Fletcher (UCSC)

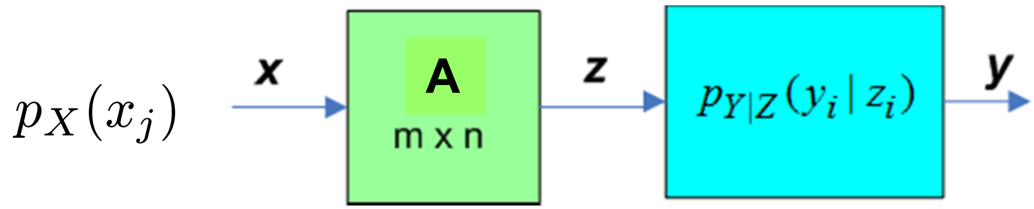
École normale supérieure, Paris, France, 18 Nov 2013

NYU-poly
POLYTECHNIC INSTITUTE OF NEW YORK UNIVERSITY

Outline

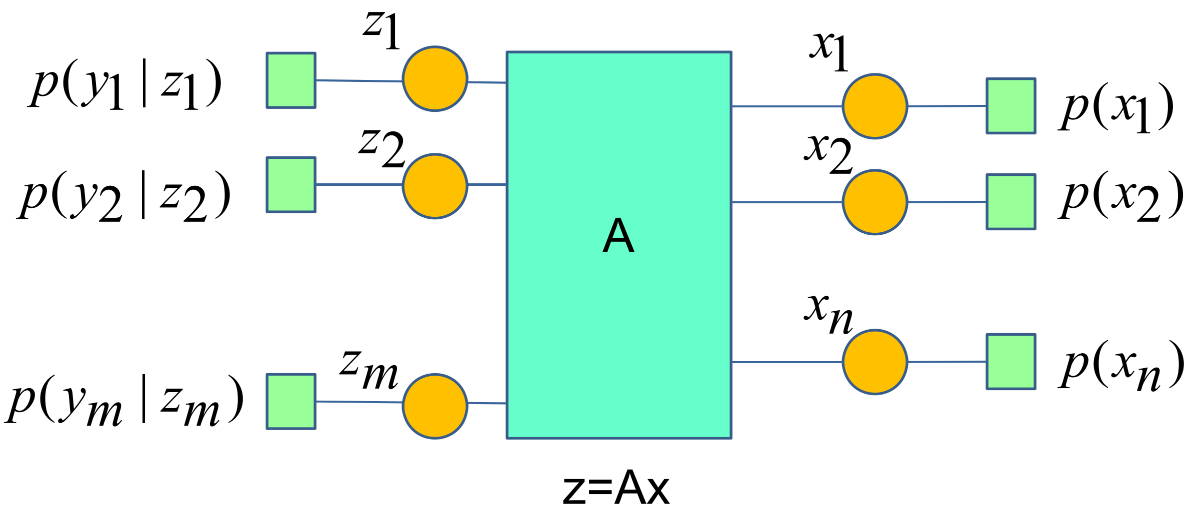
- Generalized approximate messaging (GAMP)
  - Graphical model approach for estimation with linear mixing
  - Challenges with arbitrary matrices
- Max-Sum GAMP: Connections to ADMM
- Sum-Product GAMP: Free energy optimization
- Convergence in AWGN models
- Numerical examples
  - Neural connectivity detection
- Conclusions

Bayesian Estimation with Linear Mixing



- Problem: Estimate **x** and **z** given **y** and **A**
- Many applications
  - Communication channels, linear inverse problems, regularized linear regression or classification
  - Compressed sensing
- Challenge: Generically, optimal estimation is hard
  - Components of vector **x** are coupled in **z**

Factor Graph for Linear Mixing Estimation



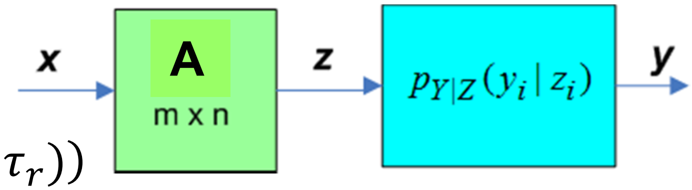
$$p(\mathbf{x} \mid \mathbf{y}) = \frac{1}{Z(\mathbf{y})} \times \prod_{i=1}^{m} p(y_i \mid z_i) \prod_{j=1}^{n} p(x_j)$$

- Posterior p($\mathbf{x}$ | $\mathbf{y}$) factors due to separability assumptions
- Output factors and variables coupled by matrix **A**
- Can apply loopy BP when coupling is sparse.

Generalized Approximate Message Passing

Input node update

$$r = \hat{x} + \tau_s A^* s$$
$$\hat{x} = E(x|r = x + N(0, \tau_r))$$



Output node update

$$p = A\hat{x} - \tau_p s$$
$$\hat{z} = E\big(z \big| y, z = p + N(0, \tau_p)\big)$$
$$s = (z - p)/\tau_p$$

- Traditional loopy BP requires sparse **A**
- GAMP: Use Gaussian and quadratic approximations.
  - Pass mean and variances
- Two variants:
  - Max-sum for MAP estimation
  - Sum-product for estimation of posterior marginals
- Computationally extremely simple
  - Linear transforms + scalar AWGN problems

History

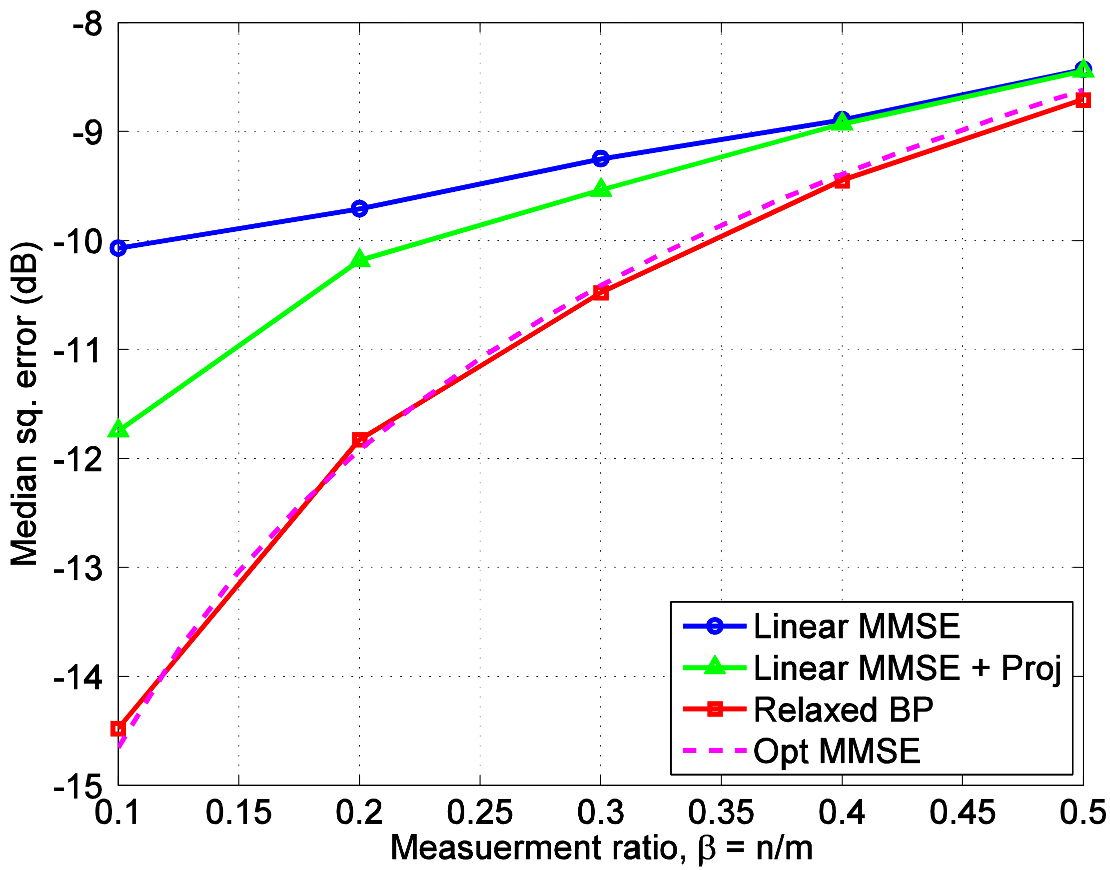
- Gaussian approximations of belief propagation
  - Multiuser CDMA & compressed sensing
  - Boutros & Caire (02), Montanari & Tse (06), Guo & Wang (06), Tanaka & Okada (06), Donoho, Maleki & Montanari (09).
  - Many names: Approximate message passing (AMP), Approx BP, relaxed BP, parallel interference cancellation (PIC), ….
- Closely related to expectation-propagation (Minka 01)
- Extensions :
  - EM: Krzakala, Mezard, Sausset, Sun, Zdeborová (2011,12), Vila, Schniter (2011), Kamilov et al (2012)
  - Turbo-hybrid: Schniter et al (2010+)

Performance of GAMP

- Well-understood for large iid **A**:
  - Scalar state evolution analysis
  - Testable conditions for optimality even when non-convex

- Extensions to new matrices
  - Sparse matrices: BouCai02, MonTse05, GuoW06,07, Ran10
  - Dense iid: DMM09, BayMon10, Ran10, JavMon11
  - Spatially coupled matrices, KrzMSSZ11,12
  - Other matrices: TulCaiVS11(free matrices)

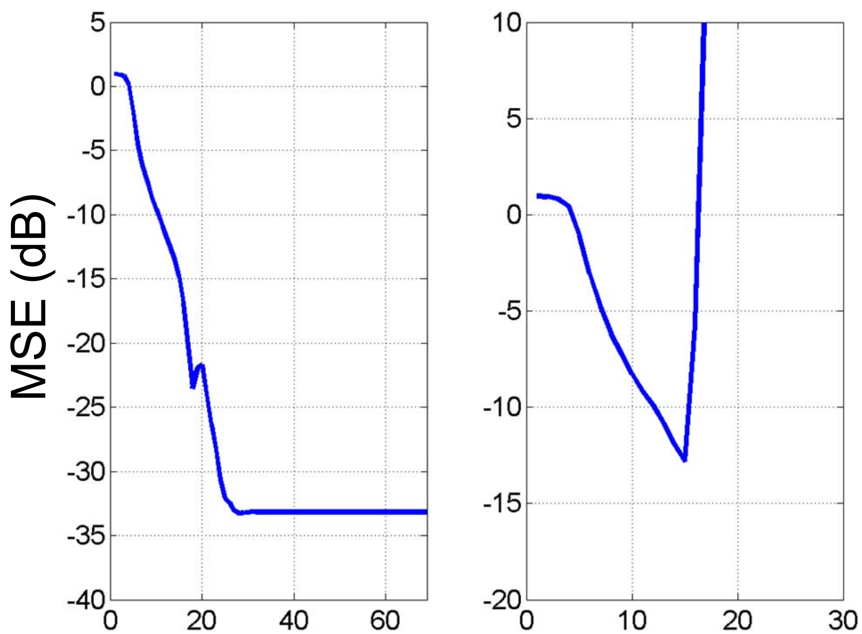
Example Bounded Noise Estimation



- Gaussian input with bounded noise output
- Arises in quantization
- NP-hard problem
- GAMP close to optimal at n=50 and outperforms best known reconstruction methods

Is GAMP only valid for certain iid A?

A = iid, N(0,1) **A** = iid N(0.5,1)



Converges rapidly

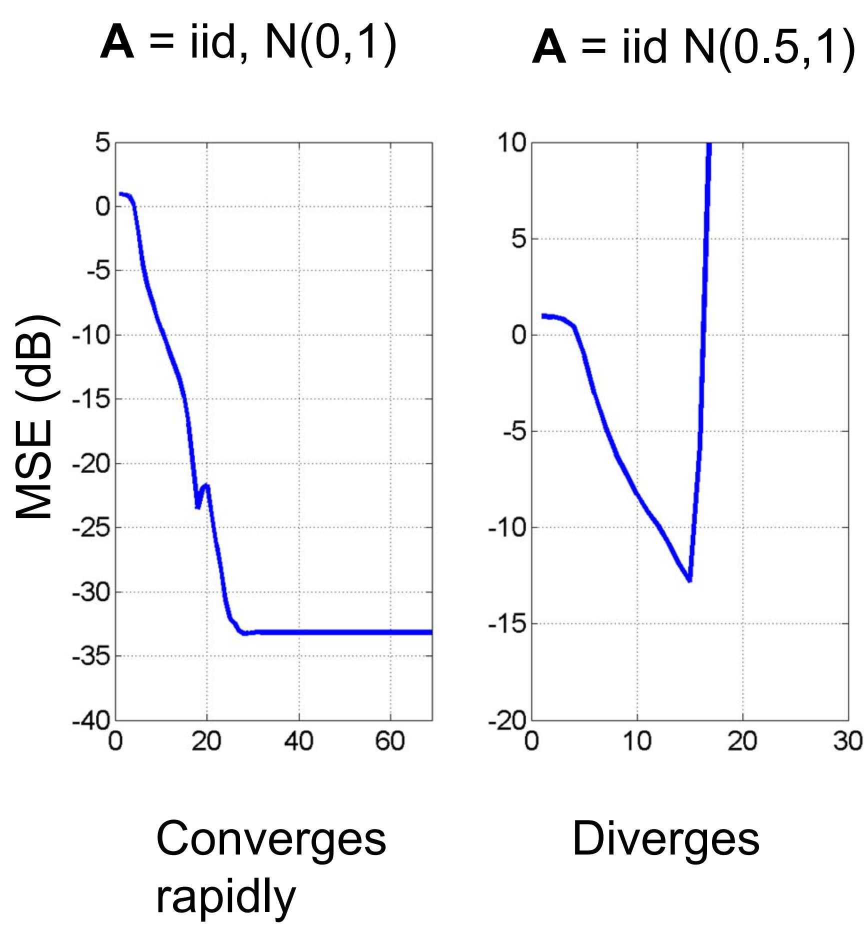
Diverges

- *"Evidently, this promise comes with the caveat that message-passing algorithms are specifically designed to solve sparse-recovery problems for Gaussian matrices…", Felix Hermann, Nuit Blanche blog*

NYU·poly

POLYTECHNIC INSTITUTE OF NEW YORK UNIVERSITY

Goals for this Talk

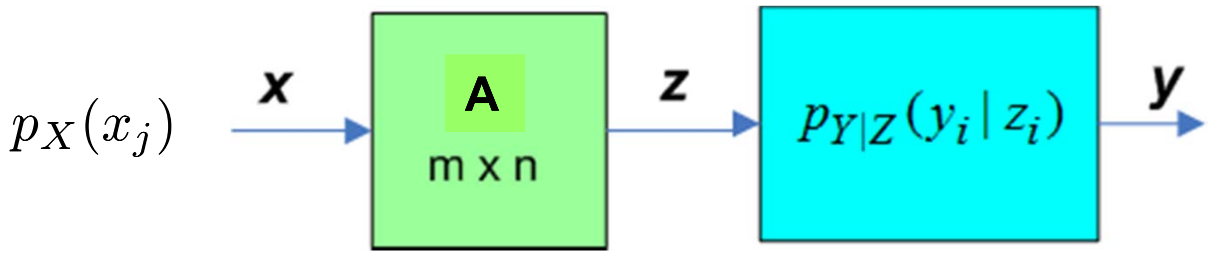


- Characterize the behavior of GAMP for arbitrary matrices
- Optimization formulation
- Relate to classic optimization methods
- Convergence results for AWGN problems
- Insights to fix GAMP

Outline

- Generalized approximate messaging (GAMP)
  - Graphical model approach for estimation with linear mixing
  - Challenges with arbitrary matrices
- Max-Sum GAMP: Connections to ADMM
- Sum-Product GAMP: Free energy optimization
- Convergence in AWGN models
- Numerical examples
  - Neural connectivity detection
- Conclusions

Max-Sum GAMP& MAP Estimation



- Consider constrained optimization:

$$(\hat{\boldsymbol{x}}, \hat{\boldsymbol{z}}) = \arg\min\ f_x(\boldsymbol{x}) + f_z(\boldsymbol{z}) \quad s.t.\ z = Ax$$

  - Separable functions $f_x(\boldsymbol{x})$ and $f_z(\boldsymbol{z})$
- Equivalent to MAP estimation with :

$$f_x(\boldsymbol{x}) = -\log p(\boldsymbol{x})$$
$$f_z(\boldsymbol{z}) = -\log p(\boldsymbol{y}|\boldsymbol{z})$$

ADMM

- Define Lagrangian:
$$L(x, z, s) = f_x(x) + f_z(z) + s^T(z - Ax)$$

- Alternating direction method of multipliers (ADMM):
$$x^{t+1} = \arg\min f_x(x) - s^{tT}Ax + Q_x(x, x^t, z^t)$$
$$z^{t+1} = \arg\min f_z(z) + s^{tT}z + Q_z(z, x^{t+1}, z^t)$$
$$s^{t+1} = s^t + \alpha(z^{t+1} - Ax^{t+1})$$

- Classic technique in optimization:
  - Convergence with appropriate auxiliary functions
  - Minimizations often have simple closed-form expressions.
  - Reduces to variant of iterative thresholding for compressed sensing

Convergence of ADMM

- "Classic" ADMM uses quadratic penalties

$$Q_x = \frac{\alpha}{2}\|z^t - Ax\|^2, \qquad Q_z = \frac{\alpha}{2}\|z - Ax^t\|^2$$

- When f_x and f_z are convex, ADMM will converge for any α
- But, x-step optimization is not separable
  - Use conjugate gradient steps with variable splitting
  - Method of choice for many compressed sensing solvers
- Can also use inexact methods
  - Bound quadratic with a separable augmenting function.

Max-Sum GAMP as ADMM

- Theorem: Max-sum GAMP is equivalent to inexact ADMM:

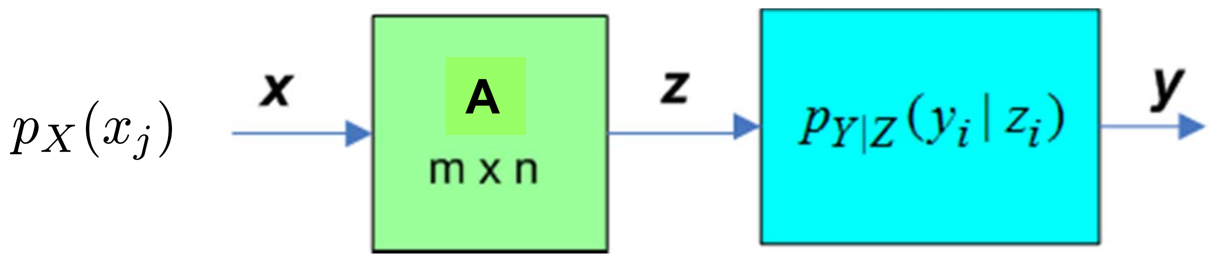
$$x^{t+1} = \arg\min f_x(x) - s^{tT} Ax + \|x - x^t\|^2 / 2\tau_r^t$$
$$z^{t+1} = \arg\min f_z(z) + s^{tT} z + \|z - Ax^{t+1}\|^2 / 2\tau_p^{t+1}$$
$$s^{t+1} = s^t + \ (z^{t+1} - Ax^{t+1}) / 2\tau_p^{t+1}$$

- Implications:
  - Fixed-point of GAMP are local maxima of posterior
  - But, convergence is not guaranteed
  - Adaptive, vector-valued step sizes

Outline

- Generalized approximate messaging (GAMP)
  - Graphical model approach for estimation with linear mixing
  - Challenges with arbitrary matrices
- Max-Sum GAMP: Connections to ADMM
- Sum-Product GAMP: Free energy optimization
- Convergence in AWGN models
- Numerical examples
  - Neural connectivity detection
- Conclusions

Sum-Product GAMP



- Produces estimates of the posterior marginals

$$p(x_j|\boldsymbol{y}) = p(x_j)\exp\left[-(x_j - r_j)^2/(2\tau_r)\right]$$

$$p(z_i|\boldsymbol{y}) = p(y_i|z_i)\exp\left[-(z_i - p_i)^2/(2\tau_p)\right]$$

- Derived based on approximation of loopy BP
- But, no optimization interpretation

Free Energy Optimization in Estimation

- Estimation as an optimization:

$$b_{x,z}(x,z) = \arg\min D(b_{x,z} || p_{x,z})$$

  - Minimize over a tractable class
  - Ex: Mean-field methods => use separable distribution

- Theorem (Yedidia, Freeman, Weiss, 2003):
  Loopy BP minimizes the Bethe free energy.
  - Optimization over marginal distributions
    + consistency constraints

Sum-Product GAMP
Free Energy Minimization

- Consider "energy" function:

$$J\left(b_x, b_z, \tau_p\right) := D(b_x || e^{-f_x}) + D(b_z || e^{-f_z}) + H_{gauss}(b_z, \tau_p)$$

  - Second-order moment matching constraints btw b_x and b_z.

$$E(z|b_z) = AE(x|b_x), \qquad \tau_p = |A|^2 var(x|b_x)$$

  - Similar in form to Bethe free energy

- Theorem: Fixed-points of sum-product GAMP are local minima of $J\left(b_x, b_z, \tau_p\right)$

GAMP Distributions

- Minima of energy function have parametric form:

$$-\log b_z(z_i) = f_{z_i}(z_i) + \frac{1}{2\tau_{p_i}}(z_i - p_i)^2 + c$$

$$-\log b_x(x_j) = f_{x_j}(x_j) + \frac{1}{2\tau_{r_j}}(x_j - r_j)^2 + c$$

- Parameters $p_i, \tau_{p_i}, r_j, \tau_{r_j}$ given by GAMP outputs
- Can be used as approximations of marginal distributions

Sum Product GAMP as ADMM

- Define Lagrangian:

$$L = J\left(b_x, b_z, \tau_p\right) + s^T\left(E(z|b_z) - AE(x|b_x)\right)$$

  - Additional constraint $\tau_p = S\tau_x,\ S = |A|^2$

- GAMP iterations look like inexact ADMM and IST:

$$b_z^t = argmin\ L\left(b_x^t, b_z, \tau_p^t\right) + (1/2\tau_p^t)\ \|E(z) - Ax^t\|^2$$

$$b_x^{t+1} = argmin\ L\left(b_x, b_z^t, \tau_p^t\right) + (1/2\tau_r^t)\ \|E(x) - x^t\|^2$$

$$+(\tau_s^t)^* S\tau_x$$

$$\tau_p^t = S\tau_x^t$$

$$s^t = s^{t-1} + \frac{1}{\tau_p^t}\left(E(z|b_z^t) - AE(x|b_x^t)\right)$$

Outline

- Generalized approximate messaging (GAMP)
  - Graphical model approach for estimation with linear mixing
  - Challenges with arbitrary matrices
- Max-Sum GAMP: Connections to ADMM
- Sum-Product GAMP: Free energy optimization
- Convergence in AWGN models
- Numerical examples
  - Neural connectivity detection
- Conclusions

Linear Gaussian Models

- Study convergence with simple Gaussian models:

$$x_j \sim N(0, \tau_{0j}), \qquad y_i = z_i + N(0, \tau_{wi})$$

  - GAMP is not best algorithm: Exact solution is available

- But, convergence on Gaussian models may provide insight:
  - Johnson, Malioutov, Willsky, NIPS 2006

- Note: When AWGN-GAMP converges:
  - Means will be correct, but not variances in general
  - Weiss, Freeman, 2001

Variance Convergence

- AWGN vector-valued variance updates:

$$\tau_p^t = S\tau_x^t, \qquad \tau_s^t = \frac{1}{\tau_p^t + \tau_w},$$

$$\tau_r^t = \frac{1}{S^*\tau_s^t}, \qquad \tau_x^{t+1} = \frac{\tau_r^t \tau_0}{\tau_r^t + \tau_0}$$

  - $S = |A|^2 =$ componentwise magnitude squared

- Theorem: For any τ_w and τ_0,
  the AWGN variance updates converge to unique fixed points

- Subsequent results will consider algorithm with fixed variance vectors.

Proof of the Variance Convergence

- Define vector valued functions:

$$g_s\colon \tau_x^t \mapsto \tau_s^t, \qquad g_x\colon \tau_s^t \mapsto \tau_x^{t+1}, \qquad g = g_x \circ g_s$$

- Verify g satisfies:
  - Monotonically increasing
  - $g(\alpha\tau_s) \leq \alpha g(\tau_s)$ for $\alpha \geq 1$.

- Convergence now follows from R. D. Yates, "A framework for uplink power control in cellular radio systems", 1995
  - Used for convergence of power control loops

Convergence of the Means

Uniform Variance Update

- Consider constant case:
  - Constant variances: $\tau_{0j} = \tau_0,\ \tau_{wi} = \tau_w.$
  - Uniform variance updates in GAMP

- Theorem: The means of the AWGN GAMP will converge for all τ_0 and τ_w if and only if

$$\sigma_{max}^2(A) < \frac{2(m+n)}{mn} ||A||_F^2$$

  - $\sigma_{max}(A)$: maximum singular value
  - $||A||_F^2$ = Frobenius norm = sum of singular values

NYU·poly
POLYTECHNIC INSTITUTE OF NEW YORK UNIVERSITY

Some Matrices Work…

$$\sigma_{max}^2(A) < \frac{2(m+n)}{mn}\|A\|_F^2$$

- Convergence depends on bounded spread of singular values.

- Examples of convergent matrices:
  - Random iid: Converges due to Marcenko-Pastur
  - Subsampled unitary: $\sigma_{max}^2(A)=1$, $\|A\|_F^2 = \min(m,n)$
  - Total variation operator: $(Ax)_i = x_i - x_{i-1}$
  - Walk summable matrices:
    - Generalizes result by Maliutov, Johnson and Willsky (2006)

But, Many Matrices Diverge

$$\sigma_{max}^2(A) < \frac{2(m+n)}{mn} \|A\|_F^2$$

- Examples of matrices that do not converge:
  - Low rank: If A has r equal singular values and other are zero:
  $$2r(m+n) > mn \Rightarrow r > \min(m,n)/2$$
  - $A \in R^{m \times m}$ is a linear filter: $Ax = h * x$ for some filter h
  $$\sup_\theta |H(e^{i\theta})| < \frac{1}{2} \frac{1}{2\pi} \int |H(e^{i\theta})|^2 d\theta$$
  - Some matrices with large non-zero means:
  $$A = A_0 + \mu 1^T$$

Proof of Convergence

- With constant variances system is linear:

$$\begin{bmatrix} s^t \\ x^{t+1} \end{bmatrix} = G \begin{bmatrix} s^{t-1} \\ x^t \end{bmatrix} + b$$

  - $G = \begin{bmatrix} I & 0 \\ D(\tau_x)A^* & D(\tau_x\tau_r^{-1}) \end{bmatrix} \begin{bmatrix} D(\tau_p\tau_s) & -D(\tau_s)A \\ 0 & I \end{bmatrix}$
  - $D(\tau) = diag(\tau)$
- System is stable if and only if $\lambda_{max}(G) < 1$
- Eigenvalue condition related to singular values of

$$F = D\left(\tau_s^{1/2}\right) A D\left(\tau_x^{1/2}\right)$$

Non-Uniform Variance Updates

- Definition: Given a matrix $A \in R^{m \times n}$, vectors u and v are row-column normalizers for A if:
$$\tilde{A} = \mathrm{diag}\left(u^{1/2}\right) A \mathrm{diag}\left(v^{1/2}\right)$$
has equal row magnitudes and column magnitudes
  - $\tilde{A}$ is unique up to a constant

- Theorem: For non-uniform variance update GAMP, the means converge for all τ_0 and τ_w if and only if

$$\sigma_{max}^2\left(\tilde{A}\right) < \frac{2(m+n)}{mn} \left\|\tilde{A}\right\|_F^2$$

Damping

- Damped updates: $\theta_s, \theta_x \leq 1$

$$s^t = (1-\theta_s)s^{t-1} + \theta_s g_{out}(p^t, \tau_p^t)$$
$$x^{t+1} = (1-\theta_x)x^t + \theta_x g_{in}(r^t, \tau_r^t)$$

- Theorem: AWGN GAMP will converge for all τ_0 and τ_w if

$$\theta_s \theta_x \sigma_{max}^2(\tilde{A}) < \frac{2(m+n)}{mn} \|\tilde{A}\|_F^2$$

  - Sufficiently large damping guarantees convergence
  - But, slower rate
  - How to perform damping adaptively?

SVD Variable Splitting

- Take SVD $A = USV^*$.
- Write $z = Uw, w = SV^*x$ so that

$$\begin{bmatrix} z \\ 0 \end{bmatrix} = \begin{bmatrix} 0 & U \\ SV^* & -I \end{bmatrix} \begin{bmatrix} x \\ w \end{bmatrix} = A_{new} \begin{bmatrix} x \\ w \end{bmatrix}$$

- New matrix A_{new} can be row-column normalized to have small range in singular values.

- Attractive solution for small to mid-size problems
  - Cost of SVD is one time
- But, not feasible for large problems.
  - Maybe detect dominant singular vectors?

Beyond AWGN Problems

- With constant variances, nonlinear updates of the form
$$(s^t, x^{t+1}) = G(s^{t-1}, x^t)$$
  - Derivative of
$$\mathrm{G}' = \begin{bmatrix} I & 0 \\ G'_{in} D(\tau_r) A^* & G'_{in} \end{bmatrix} \begin{bmatrix} G'_{out} & -G'_{out} D(\tau_p^{-1}) A \\ 0 & I \end{bmatrix}$$

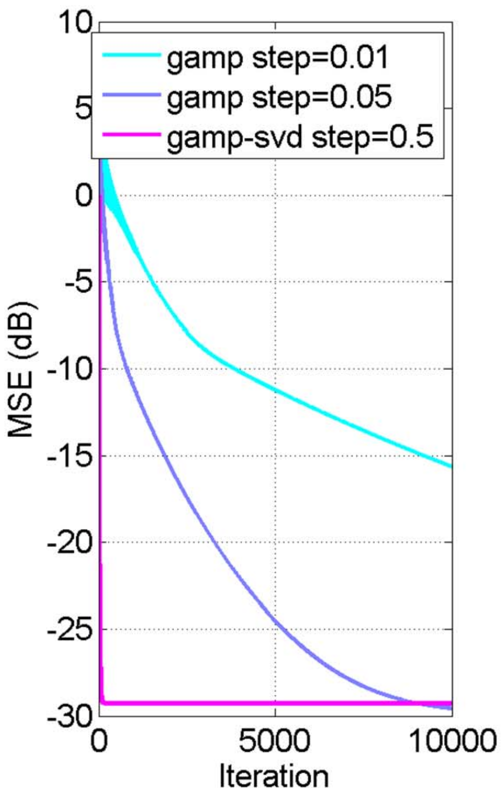
- Similar proof as AWGN case can be used since g_{in} and g_{out} are always contractions.
  - Will provide conditions for global stability of GAMP in general.

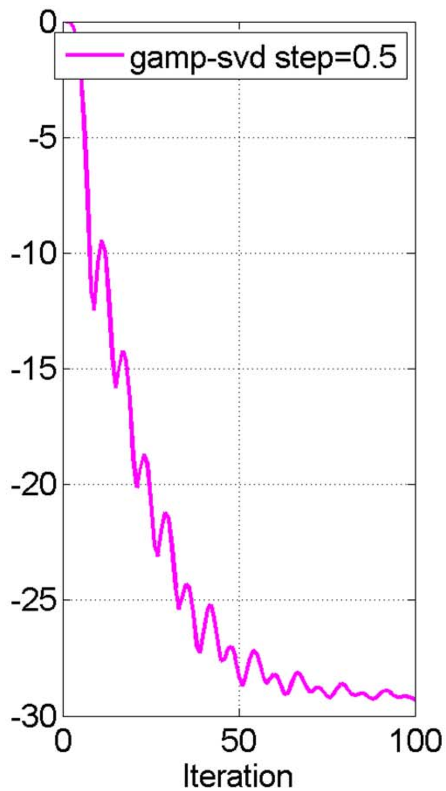
- Key challenge is that variances are not constant.

Outline

- Generalized approximate messaging (GAMP)
  - Graphical model approach for estimation with linear mixing
  - Challenges with arbitrary matrices
- Max-Sum GAMP: Connections to ADMM
- Sum-Product GAMP: Free energy optimization
- Convergence in AWGN models
- Numerical examples
  - Neural connectivity detection
- Conclusions

Ex 1. AWGN with Mean Shift





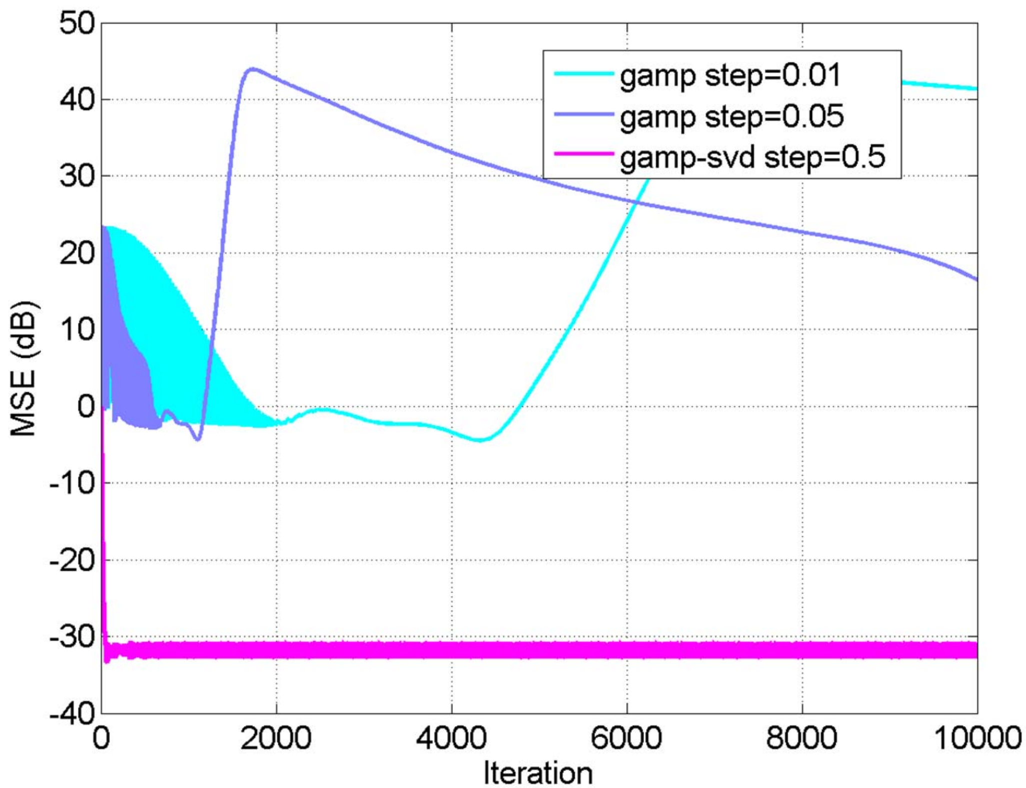
- $A \in R^{200\times100}$
- $A_{ij} \sim N(0,0.1) + 10$
- AWGN, SNR=30 dB

- Damping can get convergence
- But very slow.

- SVD method converges in ~100 iterations

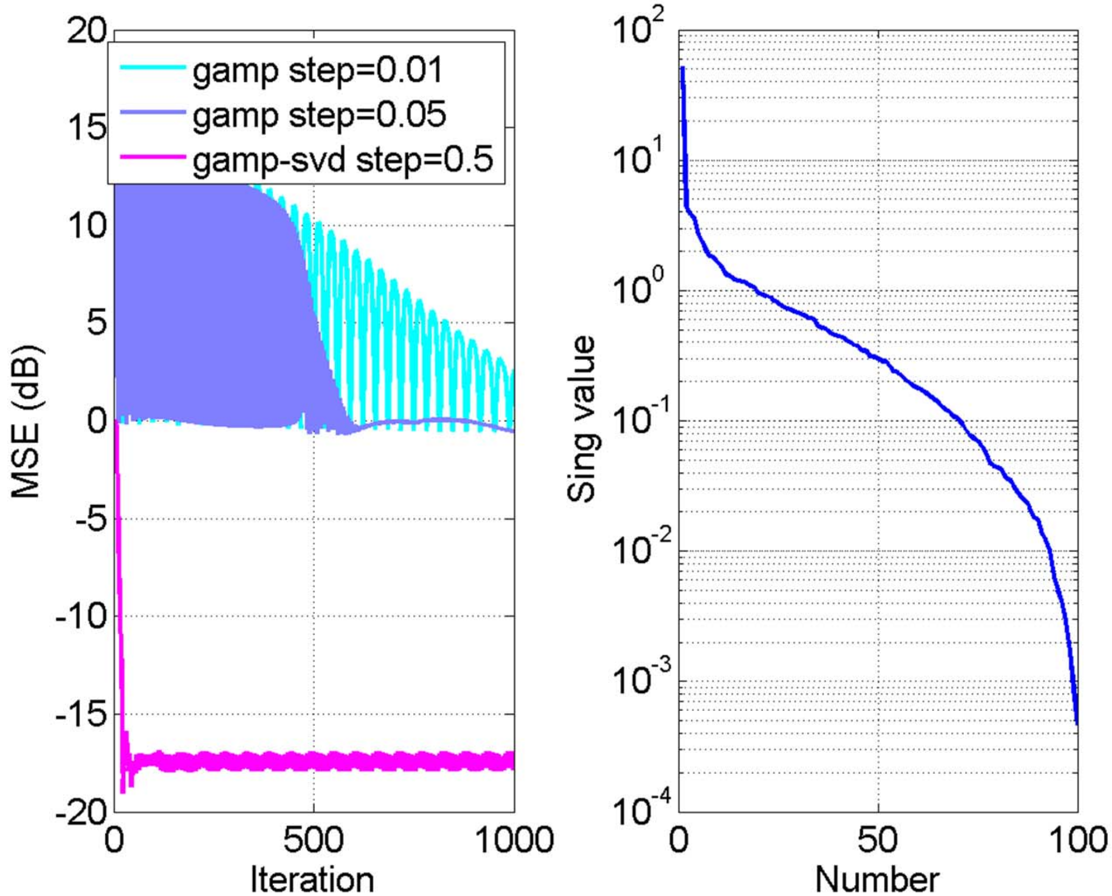
NYU·poly POLYTECHNIC INSTITUTE OF NEW YORK UNIVERSITY

Ex 2. Bernoulli-Gaussian



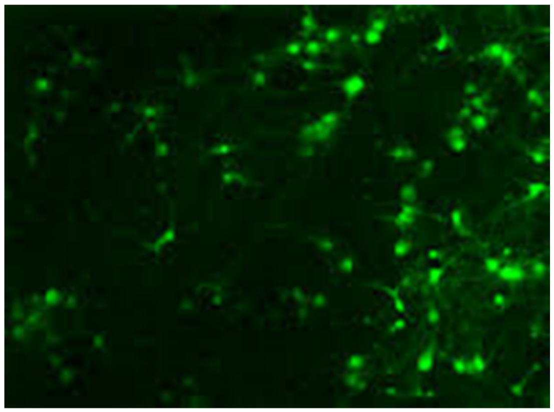
- $A \in R^{100\times 200}$
- $A_{ij} \sim N(0,0.1) + 10$
- x_j: sparsity $= 0.1$
- Damping does not converge
- But, SVD method converges in ~100 iterations

Ex 3: Large Range in Singular Values



- Matrix w/ exponentially distributed singular values
- Bernoulli-Gaussian prior
- Damping ineffective
- But, SVD method works

Neural Dynamical System



Ca imaging from David F. Meany lab, U Penn

- Infer connectivity from statistical correlations in spike patterns
- Neural dynamical system

$$x^{t+1} = \alpha x^t + \boldsymbol{W}\xi^t$$
$$\xi^t \sim Poisson(\phi(x^t))$$

- Measure ξ^t from Ca-image
- Infer connectivity $\boldsymbol{W}$

GLM model

- Neural dynamical system can be rewritten:

$$x^t = Wu^t + v^t, \qquad v^{t+1} = \alpha v^t + \xi^t$$

- Generalized Linear Model

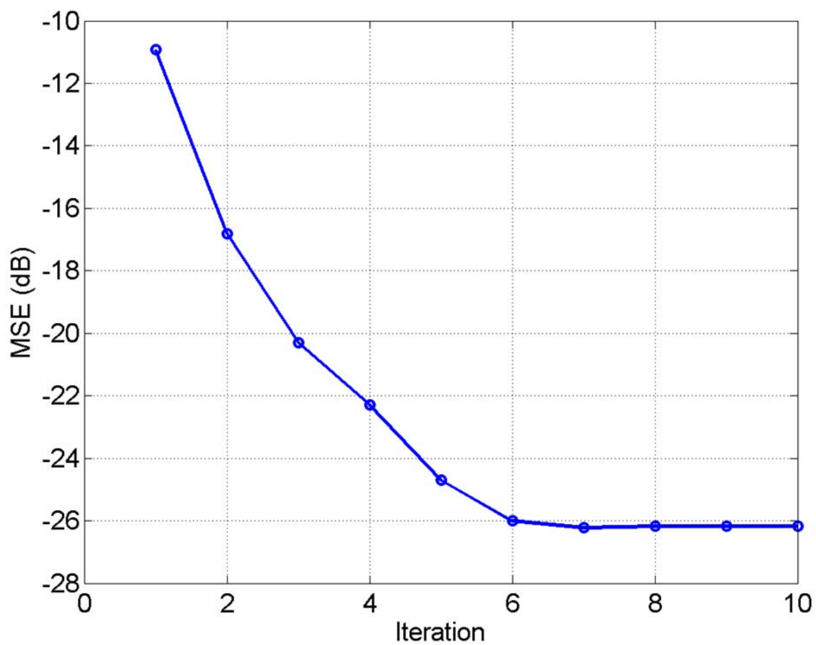
$$\xi^t \sim \text{Poisson}(\,\phi(\,Wu^t))$$

- Apply GAMP with matrix

$$A = [u^0\ u^1\ \ \cdots u^{T-1}]^*$$

  - Matrix is not i.i.d
  - Columns correlated by filtering
  - Components are non-zero mean

Fast Convergence



- SVD method converges rapidly
  - 6 to 10 iterations
- SVD can be approximately computed via Fourier transform

Outline

- Generalized approximate messaging (GAMP)
  - Graphical model approach for estimation with linear mixing
  - Challenges with arbitrary matrices
- Max-Sum GAMP: Connections to ADMM
- Sum-Product GAMP: Free energy optimization
- Convergence in AWGN models
- Numerical examples
  - Neural connectivity detection
- Conclusions

Conclusions

- AMP is a powerful algorithm for certain random matrices
- Reliable extension to arbitrary matrices remains main outstanding obstacle to widespread use
  - Conventional optimization methods likely to remain dominant
- This talk:
  - Optimization interpretation of GAMP
  - Applies to max-sum and sum-product with arbitrary matrices
  - Characterizes fixed points
  - Convergence understood for linear AWGN models
- Still many questions…