Adaptive Damping and Mean Removal for the Generalized Approximate Message Passing Algorithm

Jeremy Vila (Ohio State University), Philip Schniter (Ohio State University), Sundeep Rangan (New York University, Poly), Florent Krzakala (Sorbonne Universités), Lenka Zdeborová (Institut de Physique Théorique)

ICASSP 2015, Brisbane, Australia

Introduction

- **Goal**: Make convergence of generalized approximate message passing (GAMP) [1] robust to the matrix $\boldsymbol{A}$.
- GAMP is a computationally efficient approach to MAP or approximate MMSE inference of $\boldsymbol{x} \in \mathbb{R}^N$ that exploits:
  - known separable signal prior $p_{\mathsf{x}}(\boldsymbol{x}) = \prod_n p_{\mathsf{x}_n}(x_n)$,
  - known separable $p_{\mathsf{y}|\mathsf{z}}(\boldsymbol{y}|\boldsymbol{z}) = \prod_m p_{\mathsf{y}_m|\mathsf{z}_m}(y_m|z_m)$, where $z_m \triangleq \boldsymbol{a}_m^{\mathsf{T}}\boldsymbol{x}$.
  - sufficiently large, dense, and random $\boldsymbol{A} = [\boldsymbol{a}_1, \ldots, \boldsymbol{a}_M]^{\mathsf{T}}$.
- GAMP admits rigorous analysis in the large-system limit for i.i.d zero-mean sub-Gaussian $\boldsymbol{A}$.
- For some $\boldsymbol{A}$, GAMP can diverge.
  - "Swept" GAMP (SwGAMP) [2] improves convergence by estimating $\{\mathsf{x}_n\}_{n=1}^N$ and $\{\mathsf{z}_m\}_{m=1}^M$ sequentially, so it cannot exploit BLAS or fast matrix operations, and thus can be slow.

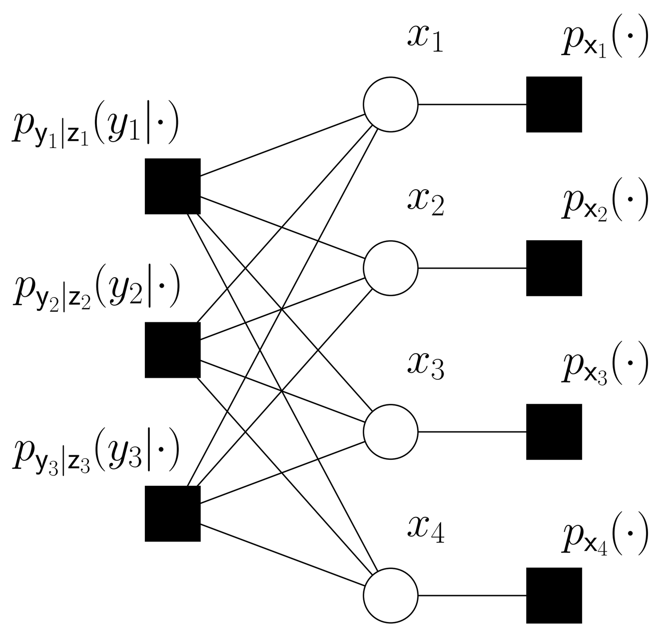


Figure 1 : GAMP factor graph.

Adaptively Damped GAMP

- To make convergence robust to the operator $\boldsymbol{A}$, we damp the updates of GAMP.
  - Not enough damping can result in GAMP divergence, while too much damping can slow GAMP convergence
  - Thus, we damp adaptively by monitoring the cost function.
- When GAMP converges, it minimizes the cost [3],[4]:

$$\text{MAP:} \quad J_{\mathsf{MAP}}(\widehat{\boldsymbol{x}}) \triangleq -\ln p_{\mathsf{y}|\mathsf{z}}(\boldsymbol{y}|\boldsymbol{A}\widehat{\boldsymbol{x}}) - \ln p_{\mathsf{x}}(\widehat{\boldsymbol{x}})$$

$$\text{MMSE:} \quad J_{\text{Bethe}}(b_{\mathsf{x}}, b_{\mathsf{z}}) \triangleq D\big(b_{\mathsf{x}}\|p_{\mathsf{x}}\big) + D\big(b_{\mathsf{z}}\|p_{\mathsf{y}|\mathsf{z}}Z^{-1}\big) + H_N(b_{\mathsf{z}}) \text{ s.t. } \mathrm{E}\{\mathsf{z}|b_{\mathsf{z}}\} = \boldsymbol{A}\,\mathrm{E}\{\mathsf{x}|b_{\mathsf{x}}\}.$$

- To evaluate $J_{\text{Bethe}}(b_{\mathsf{x}}, b_{\mathsf{z}})$ before convergence, we need to evaluate certain fixed-point quantities, and for this we use a Newton-based method.

Mean Removal

- To mitigate problems with non-zero-mean $\boldsymbol{A}$, we augment $\boldsymbol{z} = \boldsymbol{A}\boldsymbol{x}$ via

$$\underbrace{\begin{bmatrix} \boldsymbol{z} \\ z_{M+1} \\ z_{M+2} \end{bmatrix}}_{\triangleq\,\bar{\boldsymbol{z}}} = \underbrace{\begin{bmatrix} \widetilde{\boldsymbol{A}} & b_{12}\boldsymbol{\gamma} & b_{13}\mathbf{1}_M \\ b_{21}\mathbf{1}_N^{\mathsf{H}} & -b_{21}b_{12} & 0 \\ b_{31}\boldsymbol{c}^{\mathsf{H}} & 0 & -b_{31}b_{13} \end{bmatrix}}_{\triangleq\,\bar{\boldsymbol{A}}} \underbrace{\begin{bmatrix} \boldsymbol{x} \\ x_{N+1} \\ x_{N+2} \end{bmatrix}}_{\triangleq\,\bar{\boldsymbol{x}}}$$

- $\boldsymbol{\gamma}$ are row averages of $\boldsymbol{A}$.
- $\boldsymbol{c}$ are shifted column averages of $\boldsymbol{A}$.
- b_{ij} equalize row/column norms.
- $\widetilde{\boldsymbol{A}}$ has zero row and column averages.

with trivial GAMP signal priors and strict likelihoods on the augmented entries, i.e.,

$$p_{\mathsf{y}_m|\mathsf{z}_m}(y_m|z_m) \triangleq \delta(z_m) \quad \text{for} \quad m \in \{M+1, M+2\}, \qquad p_{\mathsf{x}_n}(x_n) \propto 1 \quad \text{for} \quad n \in \{N+1, N+2\}.$$

Selected References

[1] S. Rangan, "Generalized approximate message passing for estimation with random linear mixing," in *Proc. IEEE Int. Symp. Inform. Thy.*, Aug. 2011, pp. 2168–2172, (full version at *arXiv:1010.5141*).

[2] A. Manoel, F. Krzakala, E. W. Tramel, and L. Zdeborová, "Sparse estimation with the swept approximated message-passing algorithm," *arXiv:1406.4311*, Jun. 2014.

[3] S. Rangan, P. Schniter, E. Riegler, A. Fletcher, and V. Cevher, "Fixed points of generalized approximate message passing with arbitrary matrices," in *Proc. IEEE Int. Symp. Inform. Thy.*, Jul. 2013, pp. 664–668, (full version at *arXiv:1301.6295*).

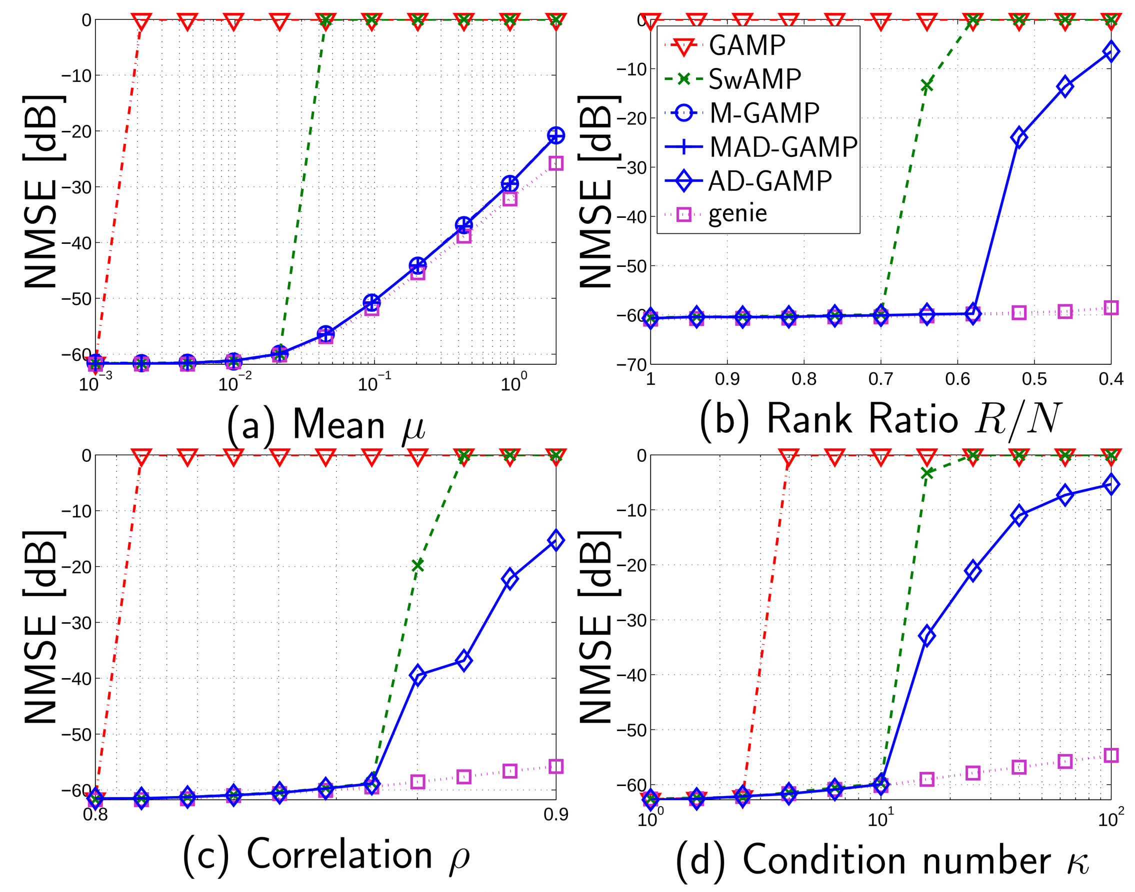
[4] F. Krzakala, A. Manoel, E. W. Tramel, and L. Zdeborová, "Variational free energies for compressed sensing," in *Proc. IEEE Int. Symp. Inform. Thy.*, Jul. 2014, pp. 1499–1503, (see also *arXiv:1402.1384*).

Numerical Results

- **Goal**: Recover $N = 1000$ Bernoulli-Gaussian $\boldsymbol{x}$ under various $p_{\mathsf{y}|\mathsf{z}}(\boldsymbol{y}|\boldsymbol{A}\boldsymbol{x})$ and "difficult" $\boldsymbol{A}$:

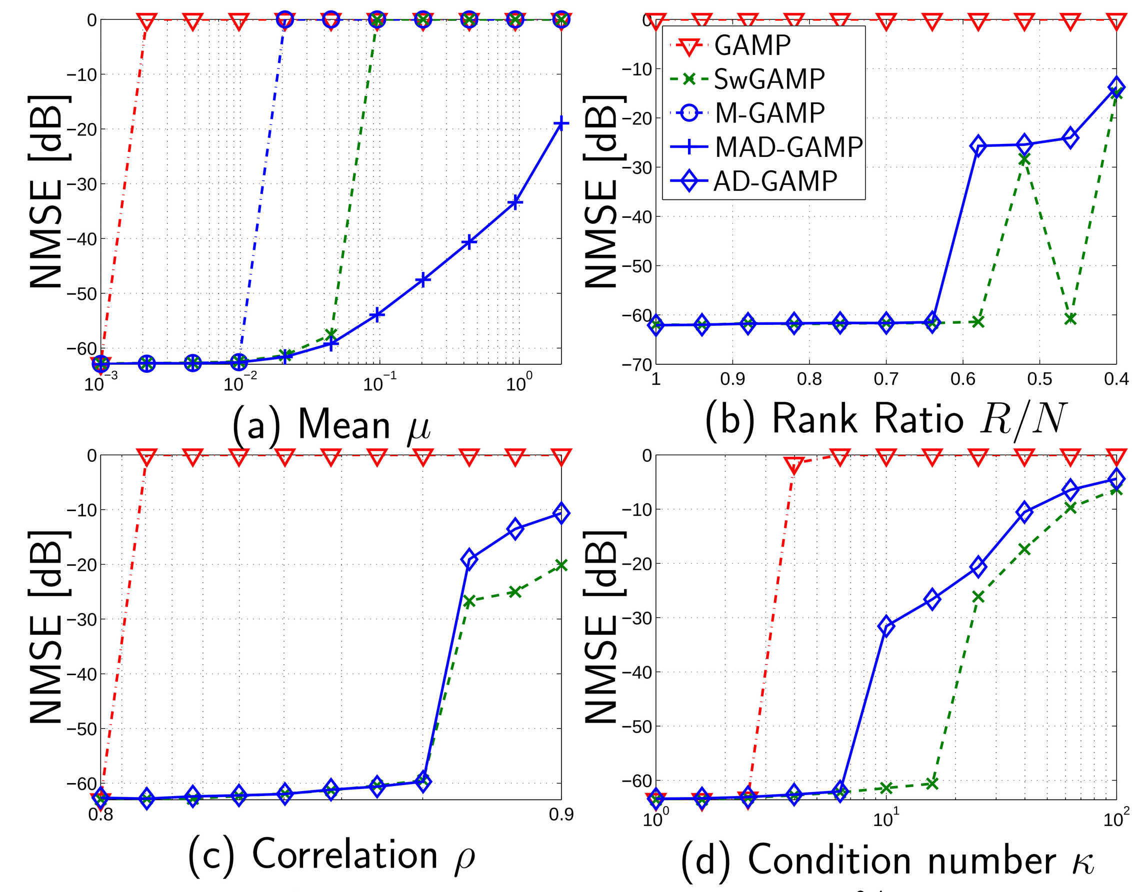
(a) **Non-zero mean**: i.i.d $\mathsf{a}_{mn} \sim \mathcal{N}(\mu, \frac{1}{N})$ for a specified $\mu \neq 0$.
(b) **Low-rank product**: $\mathbf{A} = \frac{1}{N}\mathbf{U}\mathbf{V}$ with $\mathbf{U} \in \mathbb{R}^{M\times R}$, $\mathbf{V} \in \mathbb{R}^{R\times N}$, and i.i.d $\mathsf{u}_{mr}, \mathsf{v}_{rn} \sim \mathcal{N}(0,1)$, for a specified R.
(c) **Column-correlated**: Rows of $\mathbf{A}$ are ind. Gauss-Markov processes with corr. $\rho = \mathrm{E}\{\mathsf{a}_{mn}\mathsf{a}_{m,n+1}^{\mathsf{H}}\}/\,\mathrm{E}\{|\mathsf{a}_{mn}|^2\}$.
(d) **Ill-conditioned**: $\mathbf{A} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^{\mathsf{H}}$ where $\mathbf{U}$ and $\mathbf{V}^{\mathsf{H}}$ are the left and right singular vector matrices of an i.i.d $\mathcal{N}(0,1)$ matrix and $[\mathbf{\Sigma}]_{i,i}/[\mathbf{\Sigma}]_{i+1,i+1} = (\kappa)^{1/\min\{M,N\}}$, with a specified condition number $\kappa > 1$.

- Report average NMSE $\triangleq \|\boldsymbol{x} - \widehat{\boldsymbol{x}}\|_2^2/\|\boldsymbol{x}\|$, runtime, and iterations over 100 realizations.



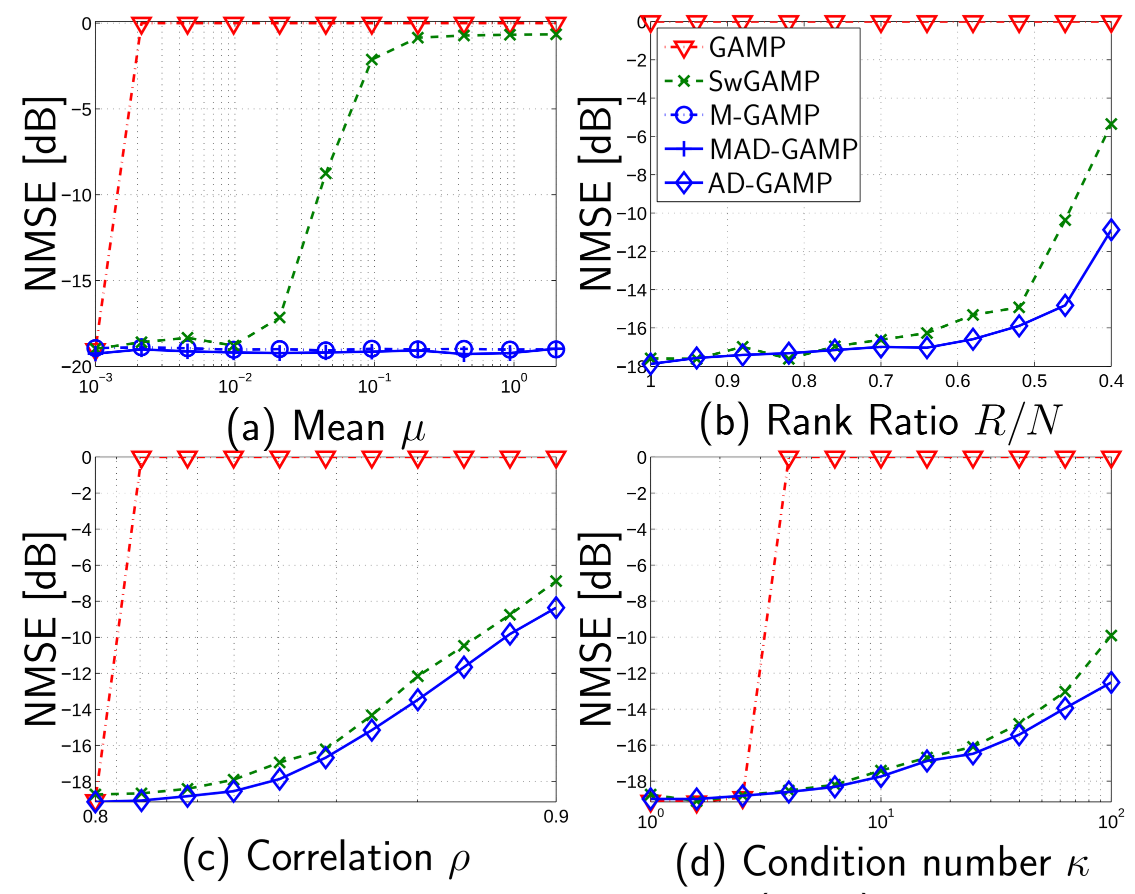
(a) Mean μ (b) Rank Ratio R/N (c) Correlation ρ (d) Condition number κ

Figure 2 : SNR $= 60$ dB AWGN.



(a) Mean μ (b) Rank Ratio R/N (c) Correlation ρ (d) Condition number κ

Figure 3 : SNR $= 60$ dB AWGN with 10% corruption



(a) Mean μ (b) Rank Ratio R/N (c) Correlation ρ (d) Condition number κ

Figure 4 : $\boldsymbol{y} = \operatorname{sgn} \boldsymbol{A}\boldsymbol{x}$ (1-bit)

		$\mu = 0.021$		$R/N = 0.64$		$\rho = 0.8$		$\log_{10}\kappa = 1$	
		MAD-GAMP	SwAMP	AD-GAMP	SwAMP	AD-GAMP	SwAMP	AD-GAMP	SwAMP
seconds	AWGN	**1.06**	1.90	**0.88**	2.74	**1.36**	3.84	**0.81**	1.49
	1-bit	**53.34**	83.21	**49.22**	137.46	**42.32**	149.40	**50.25**	117.62
	Robust	**3.47**	8.81	**2.66**	11.13	**3.33**	15.70	**2.38**	12.22
# iters	AWGN	42.9	**39.2**	130.0	**109.5**	221.9	**153.2**	121.4	**58.8**
	1-bit	947.8	**97.4**	942.7	**160.8**	866.2	**175.8**	927.3	**136.3**
	Robust	187.3	**42.2**	208.7	**56.1**	269.1	**79.2**	187.7	**61.7**

Table 1 : Average runtime (in seconds) and # iterations for various problem types and matrix types.

- The original GAMP algorithm diverges for mildly non-iid $\boldsymbol{A}$.
- The proposed Mean-removed adaptive damping GAMP (MAD)-GAMP is much more robust to non-iid $\boldsymbol{A}$.
- SwGAMP is also robust to some types of non-iid $\boldsymbol{A}$, but less so for non-zero-mean matrices.
- SwGAMP takes fewer iterations to converge, but (M)AD-GAMP has faster average runtime.

Acknowledgments

This work has been supported in part by NSF grants IIP-0968910, CCF-1018368, and CCF-1218754, an allocation of computing time from the Ohio Supercomputer Center, and by European Union's 7th Framework Programme (FP/2007-2013)/ERC Grant Agreement 307087-SPARCS.