Prox versus Prior?

Who cares—Just learn the damn thing!

Philip Schniter and Jeremy Vila

Supported in part by NSF-I/UCRC grant IIP-0968910, by NSF grant CCF-1018368, and by DARPA/ONR grant N66001-10-1-4090.

BASP Frontiers — Jan '13

Compressive Sensing

- Goal: recover signal $\boldsymbol{x}$ from noisy sub-Nyquist measurements

$$\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x} + \boldsymbol{w} \quad \boldsymbol{x} \in \mathbb{R}^N \quad \boldsymbol{y}, \boldsymbol{w} \in \mathbb{R}^M \quad M < N.$$

  where $\boldsymbol{x}$ is K-sparse with $K < M$, or compressible.

- With sufficient sparsity and appropriate conditions on the mixing matrix $\boldsymbol{A}$ (e.g. RIP, nullspace), accurate recovery of $\boldsymbol{x}$ is possible using polynomial-complexity algorithms.

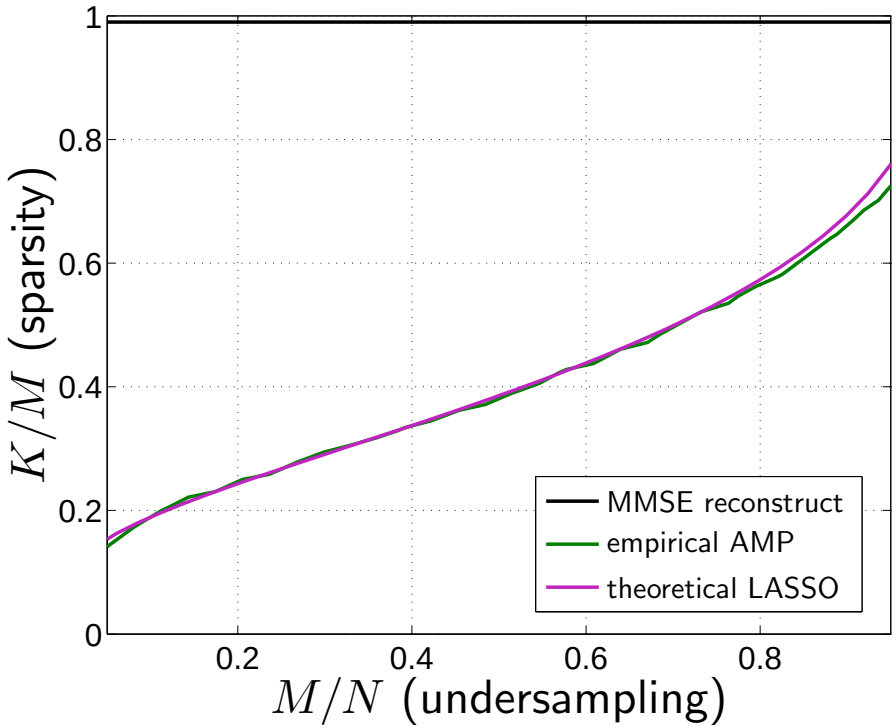
- A common approach (LASSO) is to solve the convex problem

$$\min_{\boldsymbol{x}} \|\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}\|_2^2 + \alpha\|\boldsymbol{x}\|_1$$

  where α can be tuned in accordance with sparsity and SNR.

Phase Transition Curves (PTC)

- The PTC identifies ratios $(\frac{M}{N}, \frac{K}{M})$ for which perfect noiseless recovery of K-sparse $\boldsymbol{x}$ occurs (as $M, N, K \to \infty$ under i.i.d sub-Gaussian $\boldsymbol{A}$).
- Suppose $\{x_n\}$ are drawn i.i.d.
  $$\boxed{p_X(x_n) = \lambda f_X(x_n) + (1-\lambda)\delta(x_n)}$$
  with known $\lambda \triangleq K/N$.
- LASSO's PTC is invariant to $f_X(\cdot)$. Thus, LASSO is robust in the face of unknown $f_X(\cdot)$.
- MMSE-reconstruction's PTC is far better than Lasso's, but requires knowing prior $f_X(\cdot)$.



Wu and Verdú, "Optimal phase transitions in compressed sensing," arXiv Nov. 2011.

Motivations

For practical compressive sensing...

- want minimal MSE
  - distributions are unknown $\Rightarrow$ can't formulate MMSE estimator
  - but there is hope:
    various algs seen to outperform Lasso for specific signal classes
  - really, we want a universal algorithm: good for all signal classes
- want fast runtime
  - especially for large signal-length N (i.e., scalable).
- want to avoid algorithmic tuning parameters,
  - who wants to tweak an algorithm when you can ski in the alps!

Proposed Approach: "EM-GM-AMP"

- Model the signal and noise using flexible distributions:
  - i.i.d Bernoulli Gaussian-mixture (GM) signal

$$p(x_n) = \lambda \sum_{l=1}^{L} \omega_l \mathcal{N}(x_n; \theta_l, \phi_l) + (1-\lambda)\delta(x_n) \quad \forall n$$

  - i.i.d Gaussian noise with variance ψ (but easy to generalize)

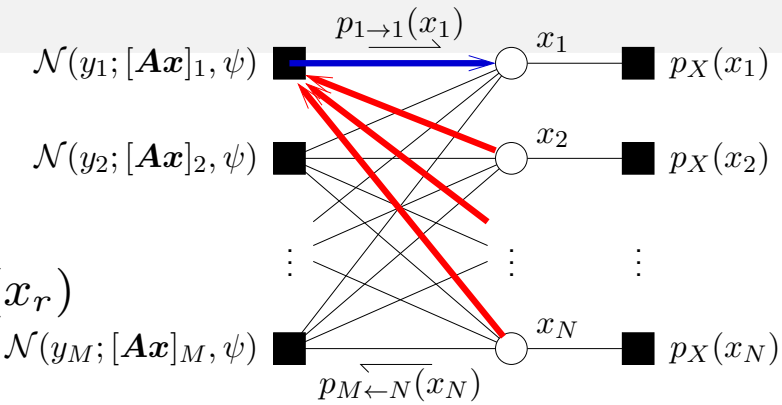
- Learn the prior parameters $\boldsymbol{q} \triangleq \{\lambda, \omega_l, \theta_l, \phi_l, \psi\}_{l=1}^{L}$
  - treat as deterministic and use expectation-maximization (EM)

- Exploit the learned priors in near-MMSE signal reconstruction
  - use the approximate message passing (AMP) algorithm
  - AMP also provides EM with everything it needs to know

Approximate Message Passing (AMP)

- AMP methods infer $\boldsymbol{x}$ from $\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x} + \boldsymbol{w}$ using loopy belief propagation with carefully constructed approximations.
  - The original AMP [Donoho, Maleki, Montanari '09] solves the LASSO problem assuming i.i.d sub-Gaussian matrix $\boldsymbol{A}$.
  - The Bayesian AMP [Donoho, Maleki, Montanari '10] framework tackles MMSE inference under any factorized signal prior, AWGN, and i.i.d $\boldsymbol{A}$.
  - The generalized AMP [Rangan '10] framework tackles MAP or MMSE inference under any factorized signal prior & likelihood, and generic $\boldsymbol{A}$.
- AMP is a form of iterative thresholding, requiring only two applications of $\boldsymbol{A}$ per iteration and ≈ 25 iterations. Very fast!
- Rigorous large-system analyses (under i.i.d sub-Gaussian $\boldsymbol{A}$) have established that (G)AMP follows a state-evolution trajectory with certain optimalities [Bayati, Montanari '10], [Rangan '10].

AMP Heuristics (Sum-Product)



1. Message from y_i node to x_j node:

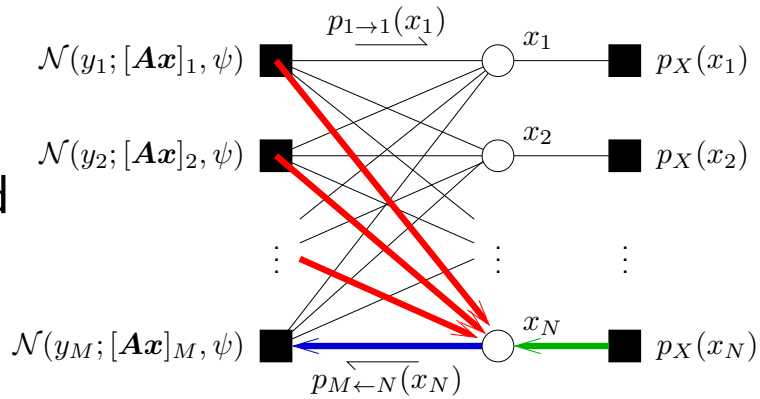
$$p_{i\to j}(x_j) \propto \int_{\{x_r\}_{r\neq j}} \mathcal{N}\big(y_i; \overbrace{\textstyle\sum_r a_{ir}x_r}^{\approx\mathcal{N}\text{ via CLT}}, \psi\big) \textstyle\prod_{r\neq j} p_{i\leftarrow r}(x_r)$$

$$\approx \int_{z_i} \mathcal{N}(y_i; z_i, \psi)\, \mathcal{N}\big(z_i; \hat{z}_i(x_j), \nu_i^z(x_j)\big) \sim \mathcal{N}$$

To compute $\hat{z}_i(x_j), \nu_i^z(x_j)$, the means and variances of $\{p_{i\leftarrow r}\}_{r\neq j}$ suffice, thus Gaussian message passing!

Remaining problem: we have $2MN$ messages to compute (too many!).

2. Exploiting similarity among the messages $\{p_{i\leftarrow j}\}_{i=1}^M$, AMP employs a Taylor-series approximation of their difference whose error vanishes as $M\to\infty$ for dense $\boldsymbol{A}$ (and similar for $\{p_{i\leftarrow j}\}_{i=1}^N$ as $N\to\infty$).
Finally, need to compute only $\mathcal{O}(M+N)$ messages!



The GAMP Algorithm

Require: Matrix $\boldsymbol{A}$, Bayes $\in \{0,1\}$, initializations $\boldsymbol{x}^0$, $\boldsymbol{\nu}_x^0$

$t = 0$, $\boldsymbol{s}^{-1} = \boldsymbol{0}$, $\forall mn : S_{mn} = |A_{mn}|^2$

repeat

$\boldsymbol{\nu}_p^t = \boldsymbol{S}\boldsymbol{\nu}_x^t, \quad \boldsymbol{p}^t = \boldsymbol{A}\boldsymbol{x}^t - \boldsymbol{s}^{t-1}.\boldsymbol{\nu}_p^t$ (gradient step)

if Bayes **then**

$\forall m : \hat{z}_m^t = \mathsf{E}(Z|P; \hat{p}_m^t, \nu_{p_m}^t), \quad \nu_{z_m}^t = \mathsf{var}(Z|P; \hat{p}_m^t, \nu_{p_m}^t),$

else

$\forall m : \hat{z}_m^t = \mathsf{prox}_{\nu_{p_m}^t f_z}(\hat{p}_m^t), \quad \nu_{z_m}^t = \nu_{p_m}^t \mathsf{prox}'_{\nu_{p_m}^t f_z}(\hat{p}_m^t)$

end if

$\boldsymbol{\nu}_s^t = (1 - \boldsymbol{\nu}_z^t./\boldsymbol{\nu}_p^t)./\boldsymbol{\nu}_p^t, \quad \boldsymbol{s}^t = (\boldsymbol{z}^t - \boldsymbol{p}^t)./\boldsymbol{\nu}_p^t$ (dual update)

$\boldsymbol{\nu}_r^t = 1./(\boldsymbol{S}^T\boldsymbol{\nu}_s^t), \quad \boldsymbol{r}^t = \boldsymbol{x}^t + \boldsymbol{\nu}_r^t\boldsymbol{A}^T\boldsymbol{s}^t$ (gradient step)

if Bayes **then**

$\forall n : \hat{x}_n^{t+1} = \mathsf{E}(X|R; \hat{r}_n^t, \nu_{r_n}^t), \quad \nu_{x_n}^{t+1} = \mathsf{var}(X|R; \hat{r}_n^t, \nu_{r_n}^t),$

else

$\forall n : \hat{x}_n^{t+1} = \mathsf{prox}_{\nu_{r_n}^t f_x}(\hat{r}_n^t), \quad \nu_{x_n}^{t+1} = \nu_{r_n}^t \mathsf{prox}'_{\nu_{r_n}^t f_x}(\hat{r}_n^t)$

end if

$t \leftarrow t+1$

until Terminated

Note connections to primal-dual, ADMM, split-Bregman, proximal FB splitting, DR...

Expectation-Maximization

- We use expectation-maximization (EM) to learn the signal and noise prior parameters $\boldsymbol{q} \triangleq \{\lambda, \boldsymbol{\omega}, \boldsymbol{\theta}, \boldsymbol{\phi}, \psi\}$
  - Alternate between
    - E: maximizing a lower bound on $\ln p(\boldsymbol{y}; \boldsymbol{q})$ for fixed $\boldsymbol{q}$
    - M: maximizing $\boldsymbol{q}$ for fixed lower bounding function.
  - Lower bound via posterior approx $\prod_{n=1}^{N} \hat{p}(x_n|\boldsymbol{y}; \boldsymbol{q}^i) \approx p(\boldsymbol{x}|\boldsymbol{y}; \boldsymbol{q}^i)$.

$$\ln p(\boldsymbol{y}; \boldsymbol{q}) = \underbrace{\sum_{n=1}^{N} \int_{x_n} \hat{p}(x_n|\boldsymbol{y}; \boldsymbol{q}^i) \ln p(x_n; \boldsymbol{q})}_{\text{lower bound}} + \underbrace{D\big(\hat{p}_{\boldsymbol{x}|\boldsymbol{y}} \,\|\, p_{\boldsymbol{x}|\boldsymbol{y}}\big)}_{\geq 0}$$

  - Incremental maximization: $\lambda, \theta_1, \ldots, \theta_L, \phi_1, \ldots, \phi_L, \boldsymbol{\omega}, \psi$
- All quantities needed for the EM updates are provided by AMP!

Parameter Initialization

Initialization matters; EM can get stuck in a local max. Our approach:

- initialize the sparsity λ according to the theoretical LASSO PTC.
- initialize the noise and active-signal variances using known energies $\|\boldsymbol{y}\|_2^2, \|\boldsymbol{A}\|_F^2$ and $\text{SNR}^0 = 20$ dB (or user-supplied value):

$$\psi^0 = \frac{\|\boldsymbol{y}\|_2^2}{(\text{SNR}^0+1)M}, \quad (\sigma^2)^0 = \frac{\|\boldsymbol{y}\|_2^2 - M\psi^0}{\lambda^0 \|\boldsymbol{A}\|_F^2}$$

- fix $L = 3$ and initialize the GM parameters $(\boldsymbol{\omega}, \boldsymbol{\theta}, \boldsymbol{\phi})$ as the best fit to a uniform distribution with variance σ^2.

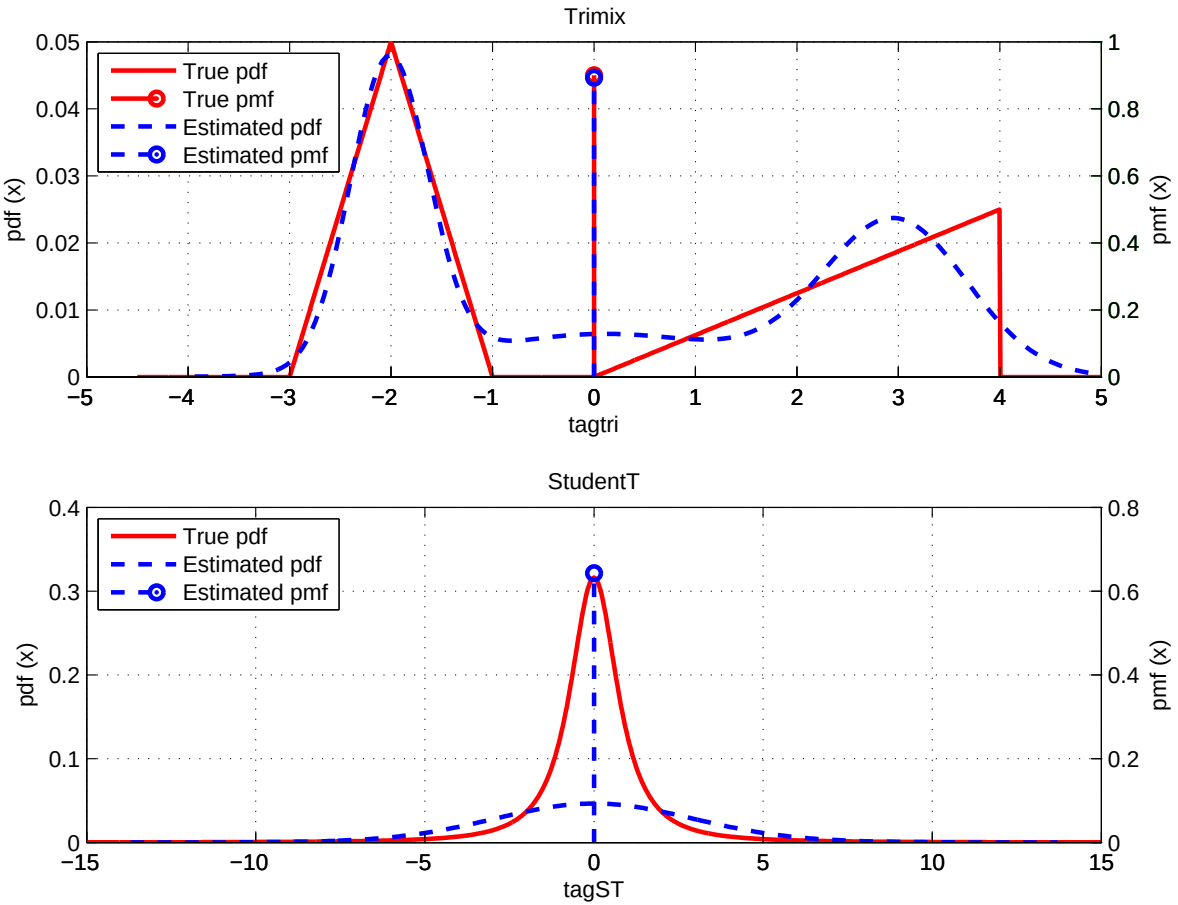
As other options, we provide

- a "heavy tailed" mode that forces zero-mean GM components.
- a model-order-selection procedure to learn L.

Examples of Learned Signal-pdfs

Example comparing EM-GM-AMP-learned priors to the actual priors used to generate the data realization.

The top corresponds to an i.i.d triangle-mixture signal and the bottom an i.i.d Student-t signal.



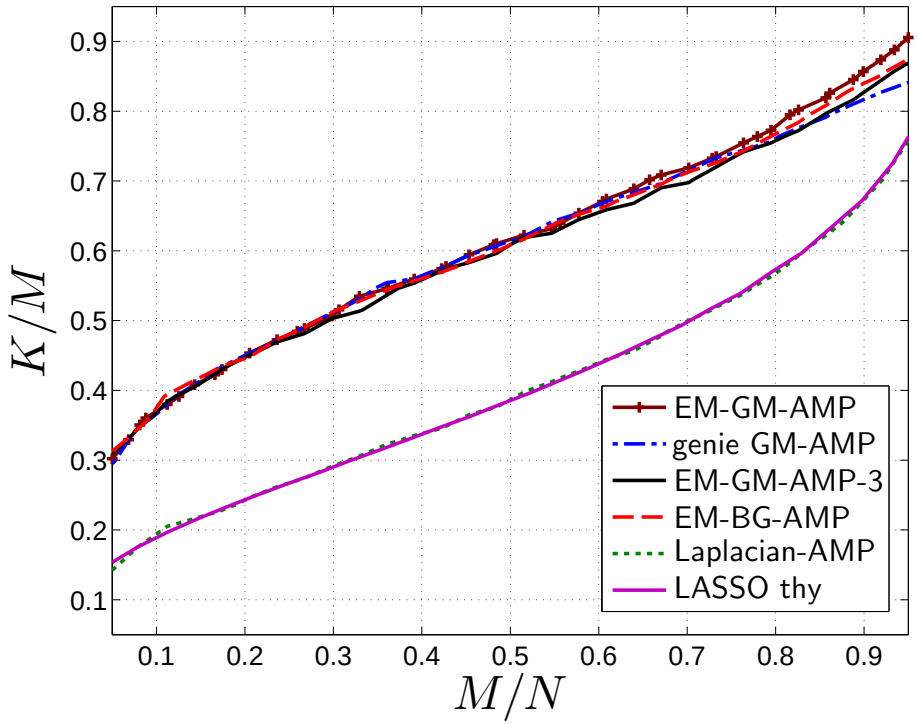
True and EM-GM-AMP-learned signal prior pdfs

Empirical PTCs: Bernoulli-Gaussian signals

- We now evaluate noiseless reconstruction performance via phase-transition curves constructed using $N = 1000$-length i.i.d signals, i.i.d Gaussian $\boldsymbol{A}$, and 100 realizations.

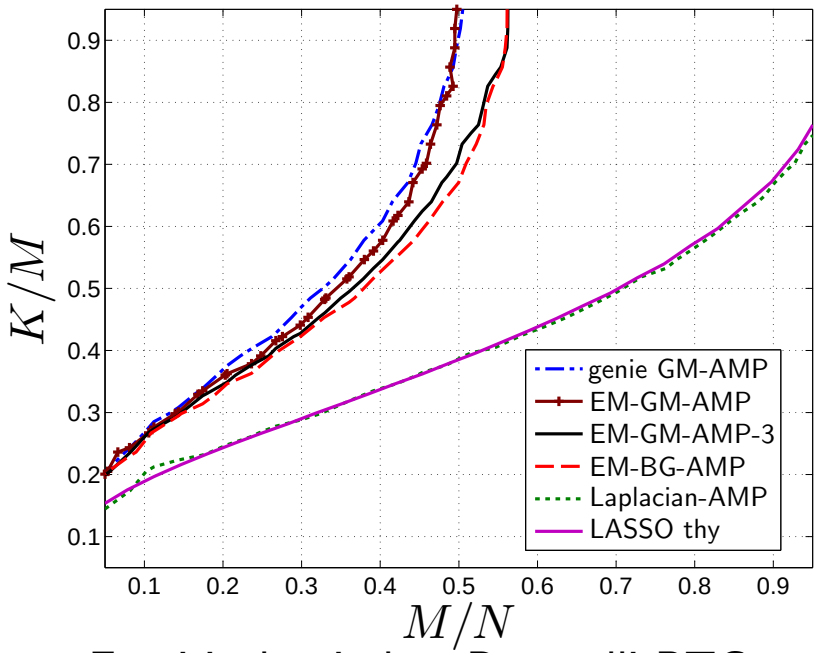
- We see all variants of EM-GM-AMP performing significantly better than LASSO for i.i.d Bernoulli-Gaussian signals.

- Perhaps not a fair comparison because the true prior is matched to our model.

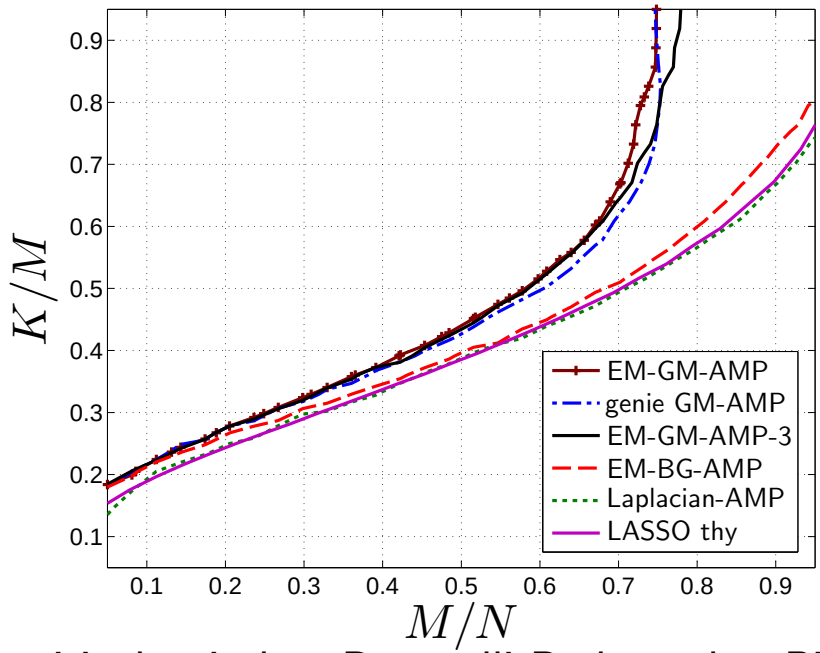


Empirical noiseless Bernoulli-Gaussian PTCs

PTCs for Bernoulli and Bernoulli-Rademacher signals



Empirical noiseless Bernoulli PTCs



Empirical noiseless Bernoulli-Rademacher PTCs

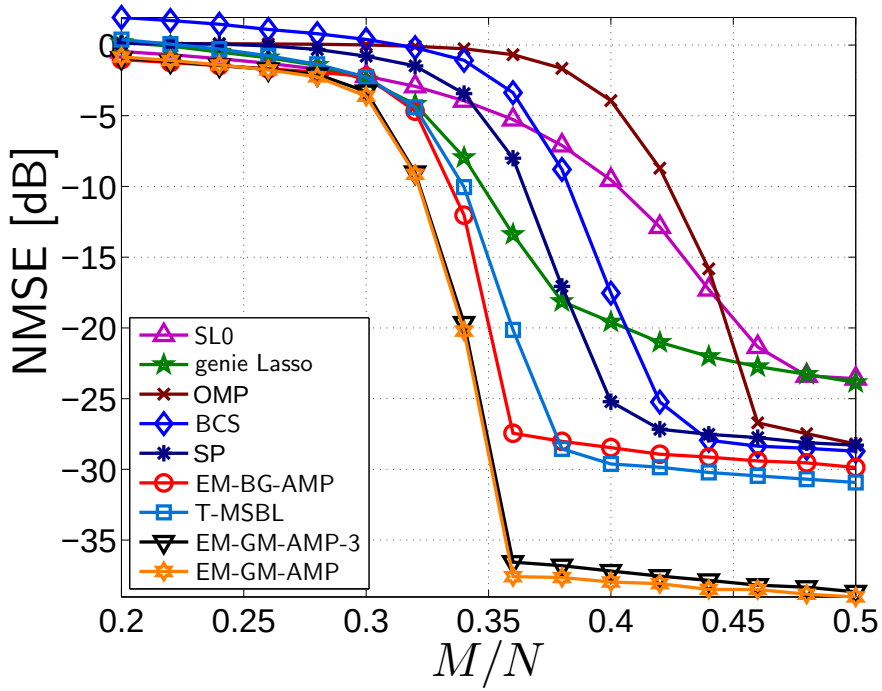
For these signals, we see EM-GM-AMP performing...

- significantly better than LASSO,
- with model-order-selection enabled, as good as genie-aided GM-AMP
- significantly better than EM-BG-AMP with the i.i.d BR signal.

Noisy Recovery: Bernoulli-Rademacher (± 1) signals

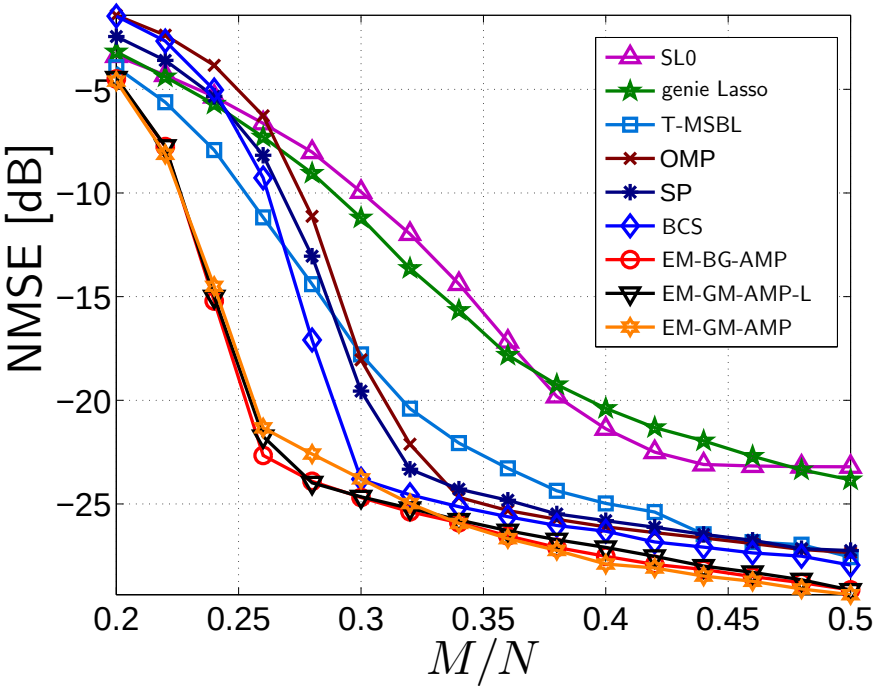
- We now compare the normalized MSE of EM-GM-AMP to several state-of-the-art algorithms (SL0, T-MSBL, BCS, Lasso via SPGL1) for the task of noisy i.i.d signal recovery under i.i.d Gaussian $\boldsymbol{A}$.
- For this, we fixed $N = 1000$, $K = 100$, SNR $= 25$dB and varied M.

- For these i.i.d BR signals, we see EM-GM-AMP outperforming the other algorithms for all undersampling ratios M/N.
- Notice that the EM-BG-AMP algorithm cannot accurately model the Bernoulli-Rademacher prior.

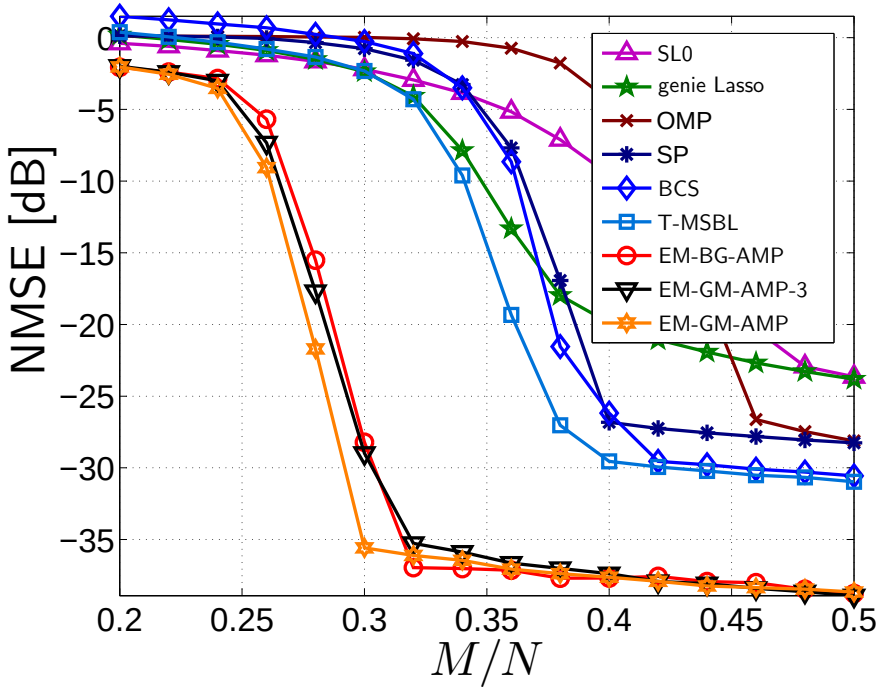


Noisy Bernoulli-Rademacher recovery NMSE.

Noisy Recovery: Bernoulli-Gaussian and Bernoulli signals



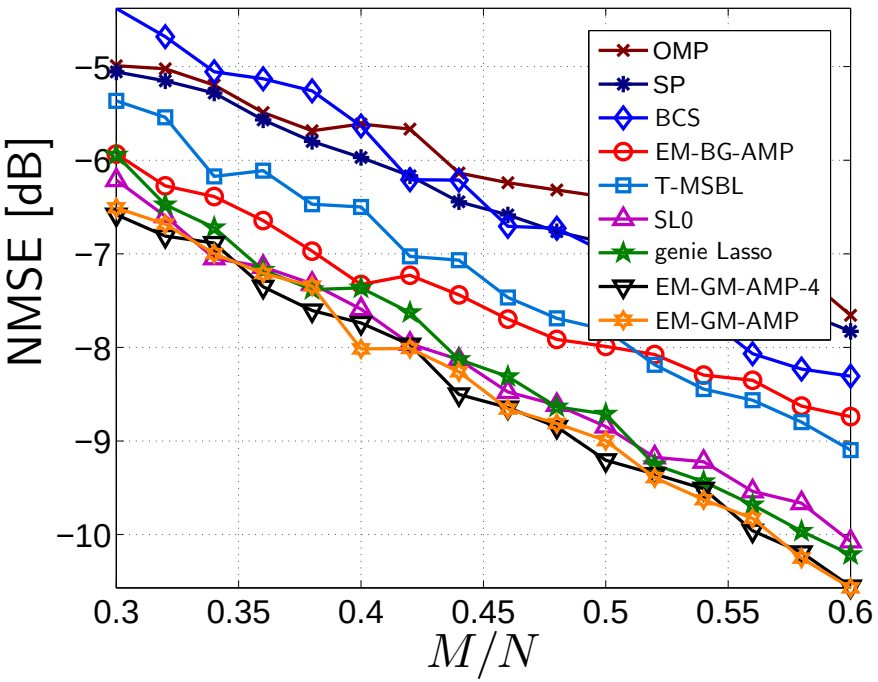
Noisy Bernoulli-Gaussian recovery NMSE.



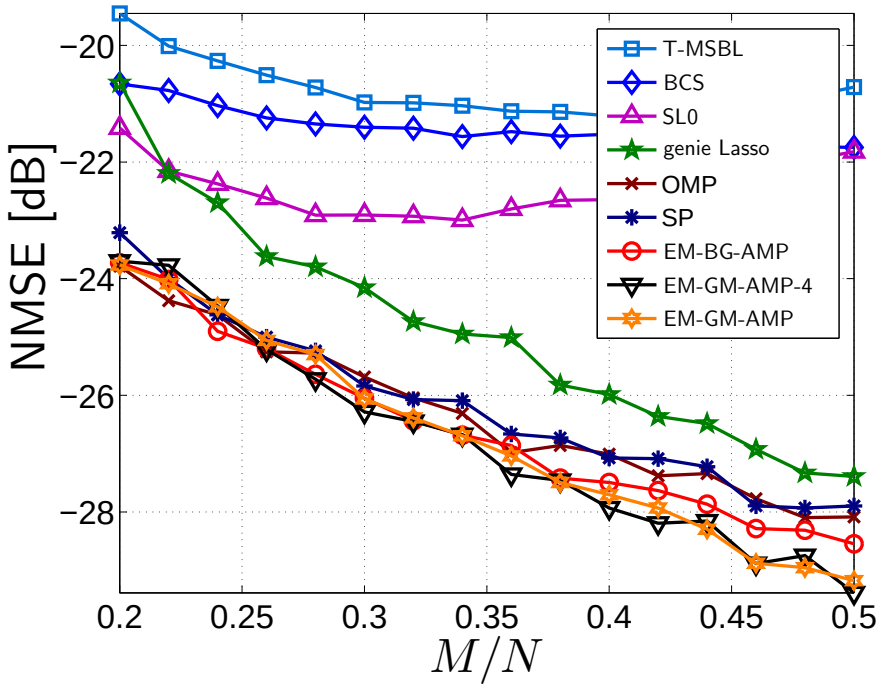
Noisy Bernoulli recovery NMSE.

- For i.i.d Bernoulli-Gaussian and i.i.d Bernoulli signals, EM-GM-AMP again dominates the other algorithms.
- We attribute the excellent performance of EM-GM-AMP to its ability to learn and exploit the true signal prior.

Noisy Recovery of Heavy-Tailed signals



Noisy Student-t recovery NMSE.

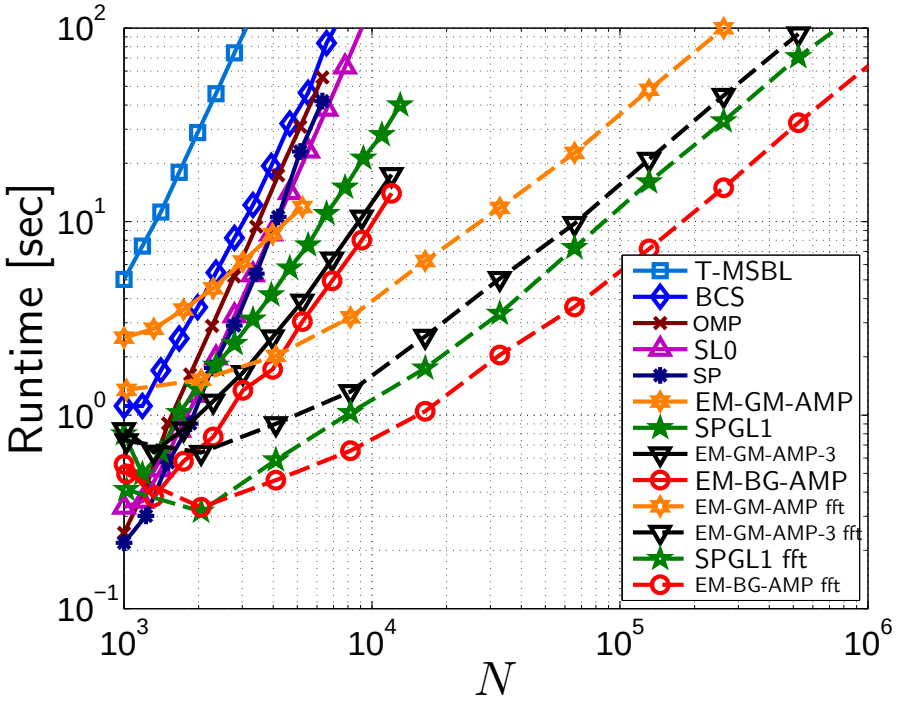


Noisy log-normal recovery NMSE.

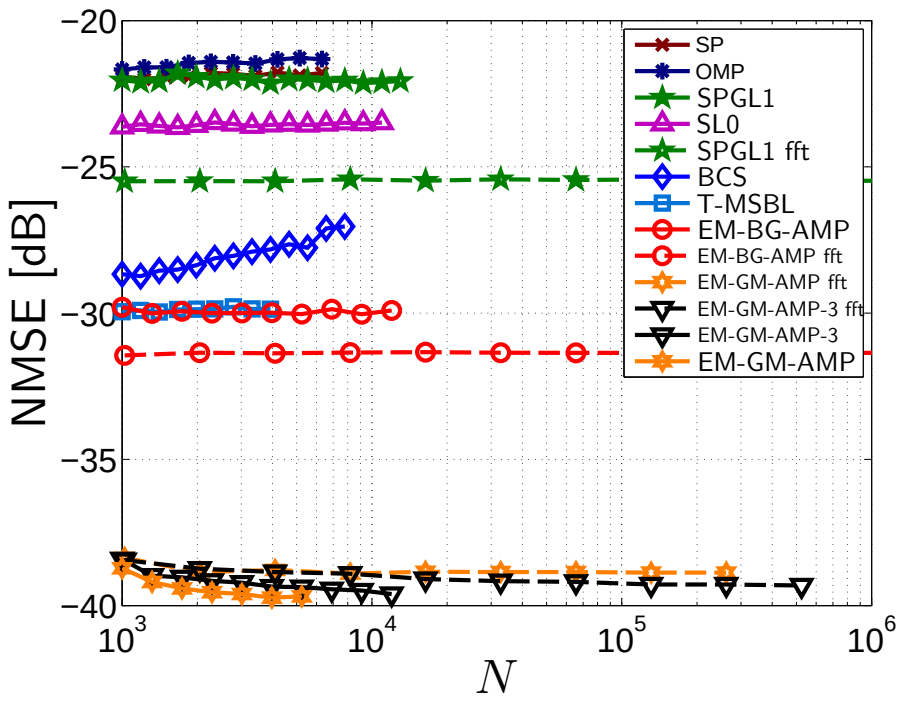
- In its "heavy tailed" mode, EM-GM-AMP again uniformly outperforms all other algorithms.
- Rankings among other algorithms differ across signal types. (Compare OMP and SL0 performances.)

Runtime versus signal-length N

- We fix $M/N = 0.5$, $K/N = 0.1$, SNR $= 25$dB, and average 50 trials.



Noisy Bernoulli-Rademacher recovery time.

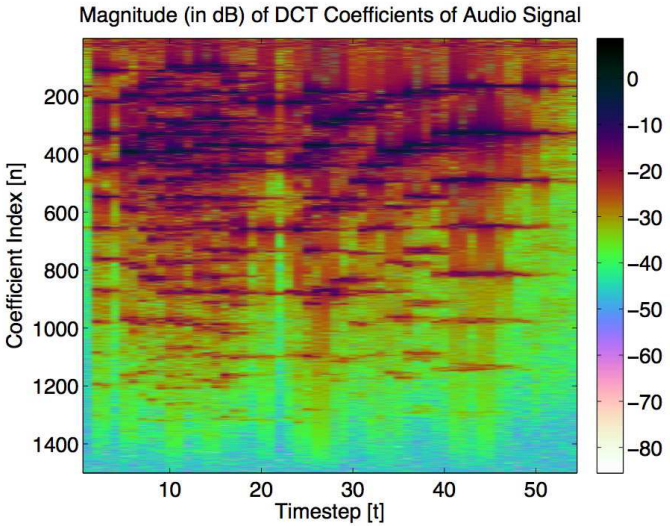


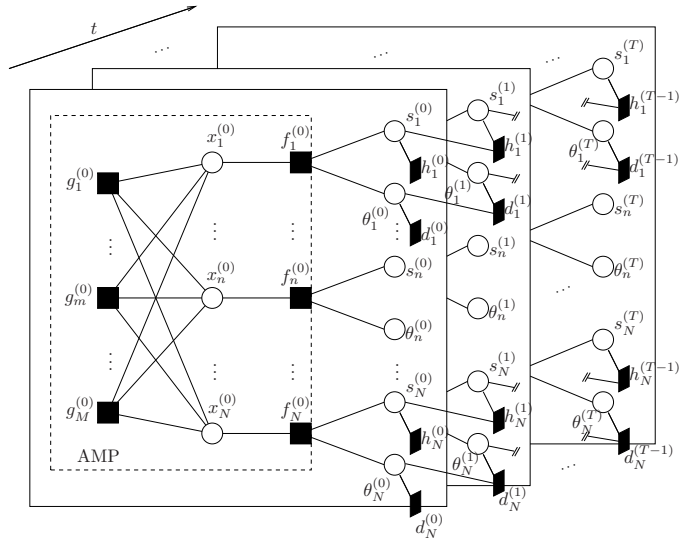
Noisy Bernoulli-Rademacher recovery NMSE.

- For all $N > 1000$, EM-GM-AMP has the fastest runtime!
- EM-GM-AMP can also leverage fast operators for $\boldsymbol{A}$ (e.g., FFT).

Extension to structured sparsity (Justin Ziniel)

- Recovery of an audio signal sparsified via DCT $\boldsymbol{\Psi}$ and compressively sampled via i.i.d Gaussian $\boldsymbol{\Phi}$ (so that $\boldsymbol{A} = \boldsymbol{\Phi}\boldsymbol{\Psi}$).
- Exploit persistence of support across time via Markov chains and turbo AMP.





algorithm	$M/N = 1/5$		$M/N = 1/3$		$M/N = 1/2$	
EM-GM-AMP-3	-9.04 dB	8.77 s	-12.72 dB	10.26 s	-17.17 dB	11.92 s
turbo EM-GM-AMP-3	-12.34 dB	9.37 s	-16.07 dB	11.05 s	-20.94 dB	12.96 s

Conclusions

- We proposed a sparse reconstruction alg that uses EM and AMP to learn and exploit the GM-signal prior and AWGN variance.
- Advantages of EM-GM-AMP: for signal length $N \gtrsim 1000$,...
  - State-of-the-art NMSE performance with all tested i.i.d priors.
  - State-of-the-art complexity
  - Minimal tuning: choose between "sparse" or "heavy-tailed" modes.
- Ongoing related work:
  - Extensions beyond AWGN (e.g., phase retrieval, binary classification).
  - Universal learning/exploitation of structured sparsity.
  - Extensions to matrix completion, dict learning, robust PCA, NNMF.
- EM-AMP Theory:
  - Asymptotic consistency under a matched prior with certain identifiability conditions:

    Kamilov, Rangan, Fletcher, Unser, "Approximate message passing with consistent parameter estimation and applications to sparse learning," *NIPS* 2012.

EM-GM-AMP is integrated into GAMPmatlab:
`http://sourceforge.net/projects/gampmatlab/`

Interface and examples at
`http://ece.osu.edu/~vilaj/EMGMAMP/EMGMAMP.html`

Thanks!