# Turbo-AMP: A Graphical-Models Approach to Compressive Inference

Phil Schniter

(Joint work with students **Dr. Subhojit Som** and **Mr. Justin Ziniel**)

(With support from NSF CCF-1018368 and DARPA/ONR N66001-10-1-4090.)

Dec. 16, 2010

## Outline:

1. Problems we consider:
   (a) classical compressive sensing,
   (b) real-world compressive sensing.
2. Recent approaches to these respective problems:
   (a) approximate message passing (AMP) [Donoho/Maleki/Montanari],
   (b) turbo-AMP [Schniter et al.]
3. Illustrative applications of turbo-AMP:
   (a) compressive imaging,
   (b) compressive tracking,
   (c) communication over sparse channels.

## The classical compressive sensing problem:

- Say $N$-length signal of interest $\boldsymbol{u}$ is *sparse* or "*compressible*" in a known orthonormal basis $\boldsymbol{\Psi}$ (e.g., wavelet, Fourier, or identity basis):

$$\boldsymbol{u} = \boldsymbol{\Psi}\boldsymbol{x}, \text{ where } \boldsymbol{x} \text{ has only } K \ll N \text{ large coefficients.}$$

- We observe $M \ll N$ noisy linear measurements $\boldsymbol{y}$:

$$\boldsymbol{y} = \boldsymbol{\Phi}\boldsymbol{u} + \boldsymbol{w} = \boldsymbol{\Phi}\boldsymbol{\Psi}\boldsymbol{x} + \boldsymbol{w} = \boldsymbol{A}\boldsymbol{x} + \boldsymbol{w}$$

  from which we want to recover $\boldsymbol{u}$ (or, equivalently, $\boldsymbol{x}$).

- If $\boldsymbol{A}$ is "*incoherent*," the sparsity of $\boldsymbol{x}$ can be exploited for provably accurate reconstruction with computationally practical algorithms.
- Incoherent $\boldsymbol{A}$ results (with high probability) from $\boldsymbol{\Phi}$ constructed *randomly* (e.g., i.i.d Gaussian) or *semi-randomly* (e.g., from random rows of fixed unitary $\boldsymbol{\Phi}$).
- Note: One usually needs to *calibrate* a "tuning parameter" that trades between sparsity and residual variance. Such "cross-validation" can be very expensive!

## Real-world compressive sensing problems:

- Usually, real-world applications exhibit *additional structure*. . .
  - among the *locations of large signal coefficients* (e.g., block, tree, etc.),
  - among the *values of large signal coefficients* (e.g., correlation),
  - within the *measurement noise* (e.g., non-Gaussian and non-independent),

  and exploitation of this structure may be essential.
- Exploiting this additional structure *complicates calibration*, since...
  - many more parameters are involved in the model, and
  - mismatch in these parameters can severely bias the signal estimate.
- Many real-world applications are *not content with point estimates*. . .
  - since CS-estimates are later used for decisions, classification, control, etc.,
  - in which case CS must provide confidence intervals, or preferably the *full posterior probability distribution* on the unknowns.

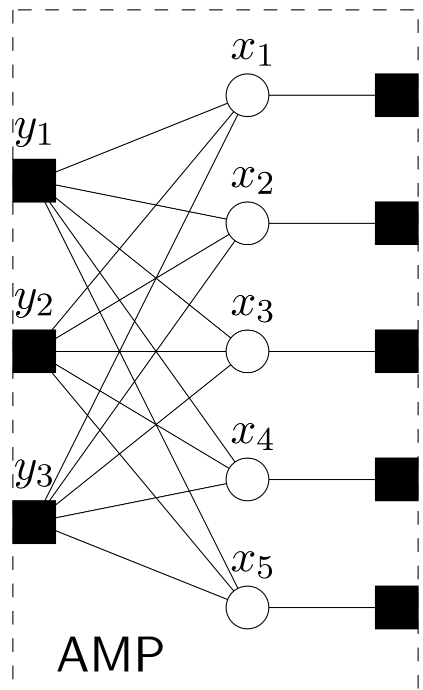
## Solving the classical CS problem — AMP:

- *Approximate message passing* (AMP) [Donoho/Maleki/Montanari 2010] refers to a family of signal reconstruction algorithms that are
  - designed to solve the *classical* CS problem,
  - inspired by (approximate) belief propagation.
- AMP highlights:
  - Very *computationally efficient*: a form of iterative thresholding.
  - Very *high performance* (for i.i.d $\boldsymbol{A}$ and sufficiently large $N, M$):
    - Can configure for LASSO robustness, or ≈MMSE accuracy!
  - Admits *rigorous asymptotic analyses* [Bayati/Montanari 2010, Rangan 2010] (under i.i.d Gaussian $\boldsymbol{A}$ and $N, M \to \infty$ with fixed $N/M$):
    - AMP follows a (deterministic) state-evolution trajectory.
    - Agrees with analysis under the (non-rigorous) replica method.
    - Agrees with sparse-matrix analysis of Guo/Wang's relaxed-BP algorithm that shows asymptotic optimality of marginal posterior estimates.

## Solving real-world compressive inference problems — Turbo-AMP:

- The *Bayesian graphical-model framework* is a flexible and powerful way to handle signal/measurement structure, calibration, and interval estimation.

Unstructured sparsity with known model parameters

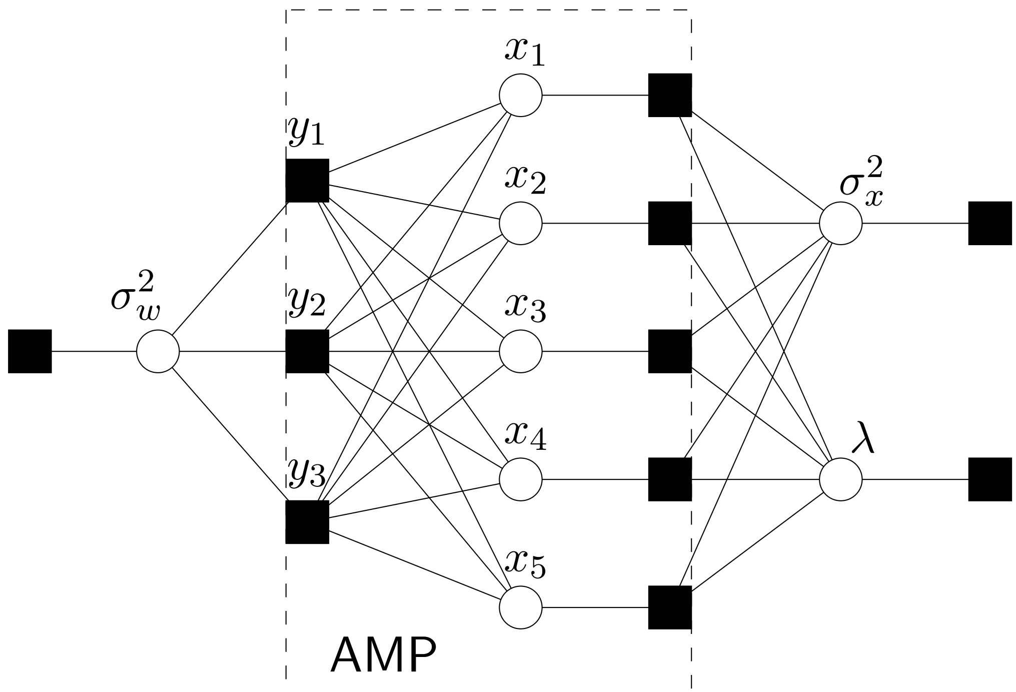



$$p(\boldsymbol{x}|\boldsymbol{y}) \propto \prod_{m=1}^{M} p(y_m|\boldsymbol{x}) \prod_{n=1}^{N} p(x_n)$$

## Solving real-world compressive inference problems — Turbo-AMP:

- The *Bayesian graphical-model framework* is a flexible and powerful way to handle signal/measurement structure, calibration, and interval estimation.

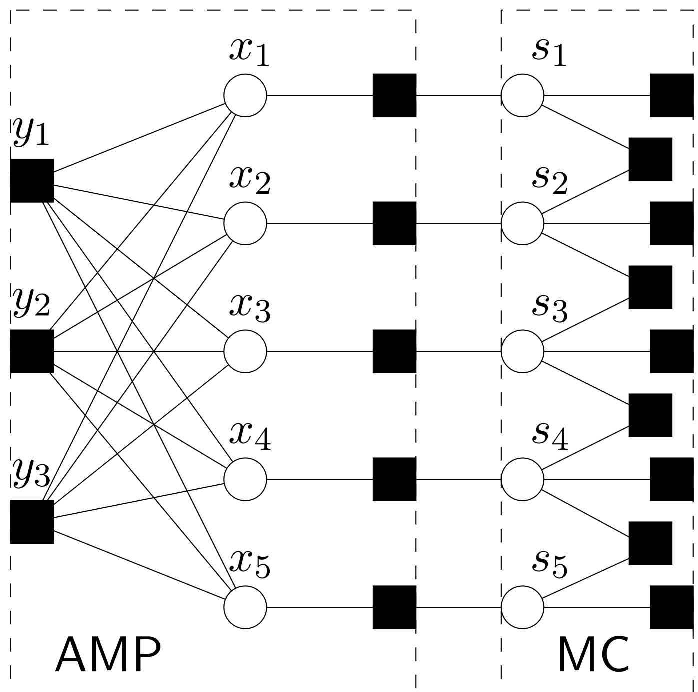
Unstructured sparsity with unknown model parameters



## Solving real-world compressive inference problems — Turbo-AMP:

- The *Bayesian graphical-model framework* is a flexible and powerful way to handle signal/measurement structure, calibration, and interval estimation.

Structured sparsity with known model parameters:

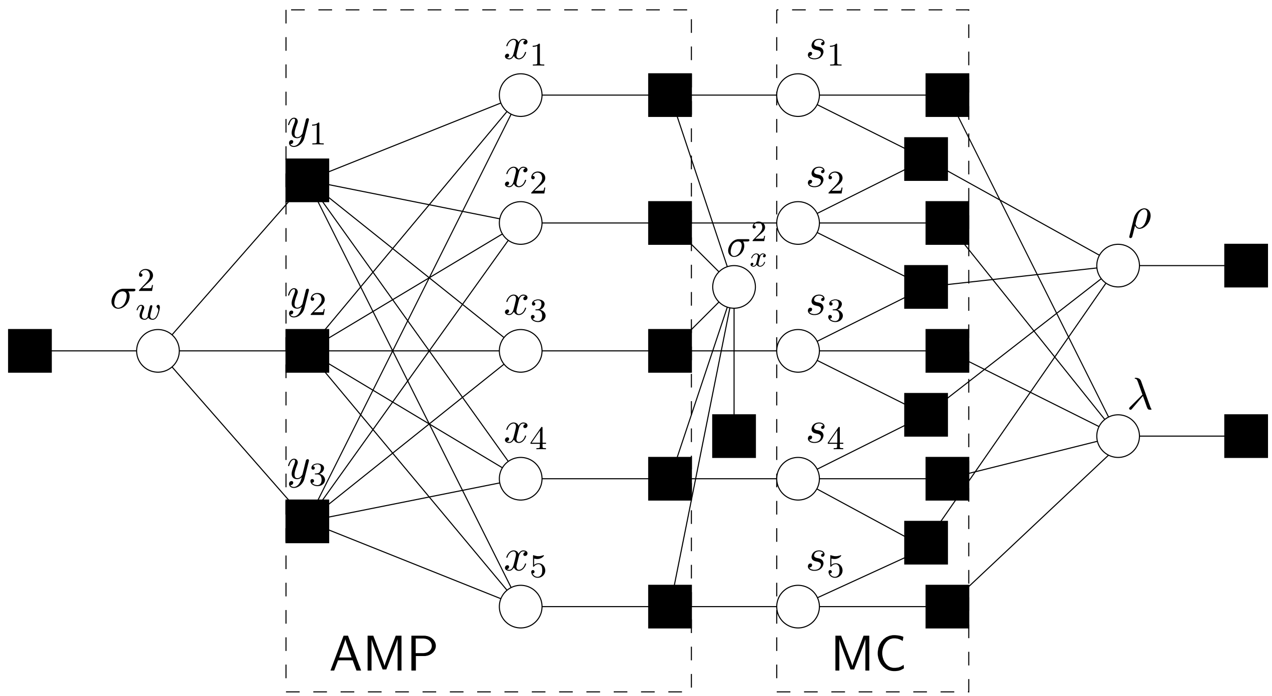



- For these problems, AMP is used as a soft-input soft-output inference block, like a "channel decoder" in a "turbo" receiver. [Schniter CISS 10]

## Solving real-world compressive inference problems — Turbo-AMP:

- The *Bayesian graphical-model framework* is a flexible and powerful way to handle signal/measurement structure, calibration, and interval estimation.

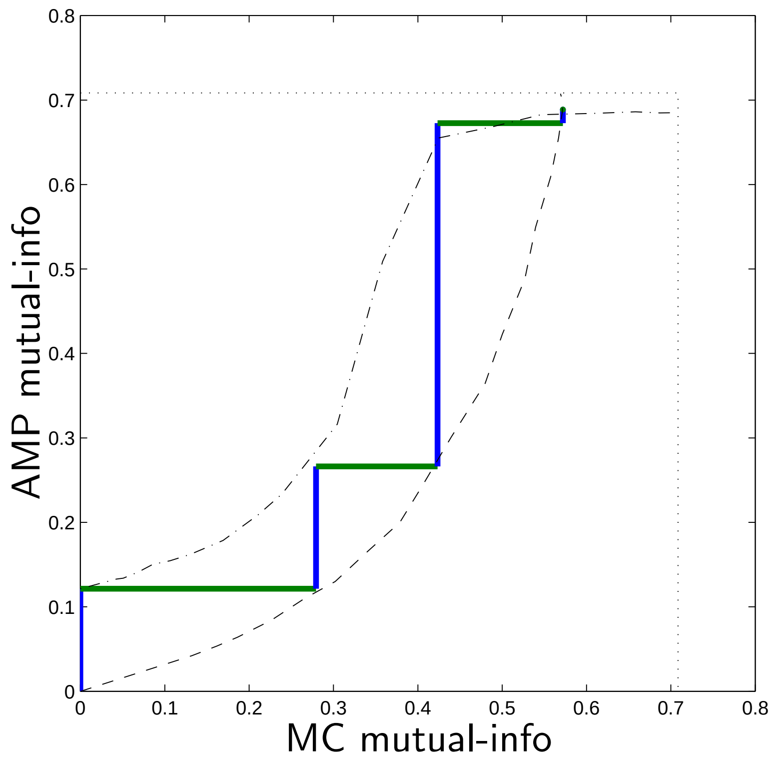
Structured sparsity with unknown model parameters:




- For these problems, AMP is used as a soft-input soft-output inference block, like a "channel decoder" in a "turbo" receiver. [Schniter CISS 10]

## Extrinsic information transfer (EXIT) charts:

EXIT charts, developed to predict the convergence of turbo decoding [ten Brink 01], can be similarly applied to turbo-AMP:




The EXIT chart plots the mutual-information between the estimated and true sparsity indicators $\{s_n\}$.

## We will now detail three applications of the turbo-AMP idea:

1. Compressive imaging

   ...with (persistence across scales) structure in the signal support.

2. Compressive tracking

   ...with (slow variation) structure in the signal's support and coefficients.

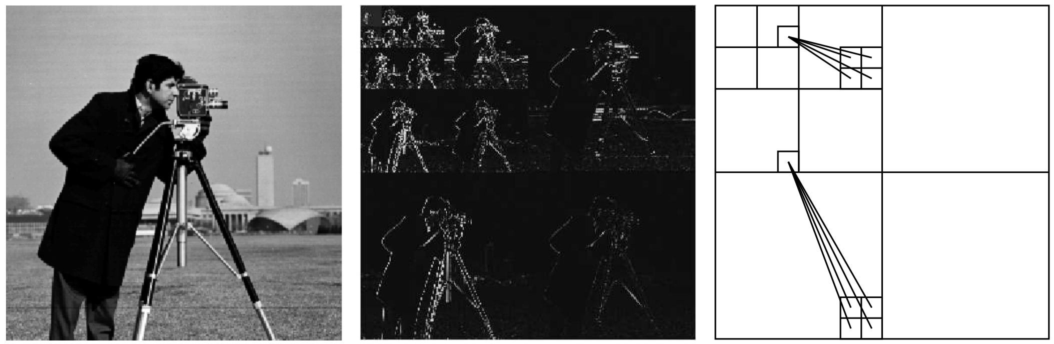
3. Communication over sparse channels

   ...with structured non-AWGN measurements, and

   ...where signal estimates are used in a larger detection problem.

## 1) Compressive imaging:

- Wavelet representations of natural images are not only sparse, but also exhibit *persistence across scales*:



- Can be efficiently modeled using a *Bernoulli-Gaussian hidden Markov-tree*:

$$p(x_n \,|\, s_n) = s_n \mathcal{N}(x_n; 0, \sigma_j^2) + (1 - s_n)\delta(x_n) \quad \text{for } s_n \in \{0,1\}$$

$$p(s_n \,|\, s_m): \text{ state transition mtx } \begin{pmatrix} p_j^{00} & 1{-}p_j^{00} \\ 1{-}p_j^{11} & p_j^{11} \end{pmatrix}, \text{ for } n \in \text{children}(m), \quad j = \text{level}(n)$$

$$\boldsymbol{y} \;=\; \boldsymbol{\Phi u} + \boldsymbol{w} \;=\; \boldsymbol{\Phi \Psi x} + \boldsymbol{w}, \quad \boldsymbol{w} \sim \mathcal{N}(0, \sigma_w^2)$$

- The model parameters $\sigma_w$ and $\{\sigma_j, p_j^{00}, p_j^{11}\}_{j=0}^J$ are treated as *random* with non-informative hyperpriors (Gamma and Beta, respectively). In approximate message passing, we pass only the mean estimates.

## Comparison to other methods:

Average over Microsoft Research class recognition database (591 images):

For $M = 5000$ random measurements of $128 \times 128$ images ($N = 16384$)...

| Algorithm | Authors (year) | Time | NMSE |
|---|---|---|---|
| ModelCS | Baraniuk, Cevher, Duarte, Hegde (2010) | 1205 s | -15.10 dB |
| IRWL1 | Duarte, Wakin, Baraniuk (2008) | 363 s | -14.37 dB |
| MCMC | He & Carin (2009) | 742 s | -20.10 dB |
| Variational Bayes | He, Chen, Carin (2010) | 107 s | -19.04 dB |
| Turbo | Som & Schniter (2010) | 53 s | -20.31 dB |

*Turbo-AMP beats state-of-the-art simultaneously in speed and accuracy!*

## Comparison to other methods:

Original

ModelCS

IRWL1

Variational Bayes

MCMC

Turbo-AMP

## 2) Compressive tracking / Dynamic compressive sensing:

- We observe the sequence of vectors

$$\boldsymbol{y}^{(t)} = \boldsymbol{A}^{(t)}\boldsymbol{x}^{(t)} + \boldsymbol{w}^{(t)}, \quad t = 1:T, \quad w_n^{(t)} \sim \text{i.i.d } \mathcal{N}(0, \sigma_w^2)$$

  with a sparse $\boldsymbol{x}^{(t)}$ whose coefficients and support change slowly with time $t$.

- The slowly varying sparse signal can be modeled as Bernoulli-Gaussian with Gauss-Markov coefficient evolution and Markov-chain support evolution:

$$x_n^{(t)} = s_n^{(t)}\theta_n^{(t)} \text{ for } s_n^{(t)} \in \{0,1\} \text{ and } \theta_n^{(t)} \in \mathbb{R}$$

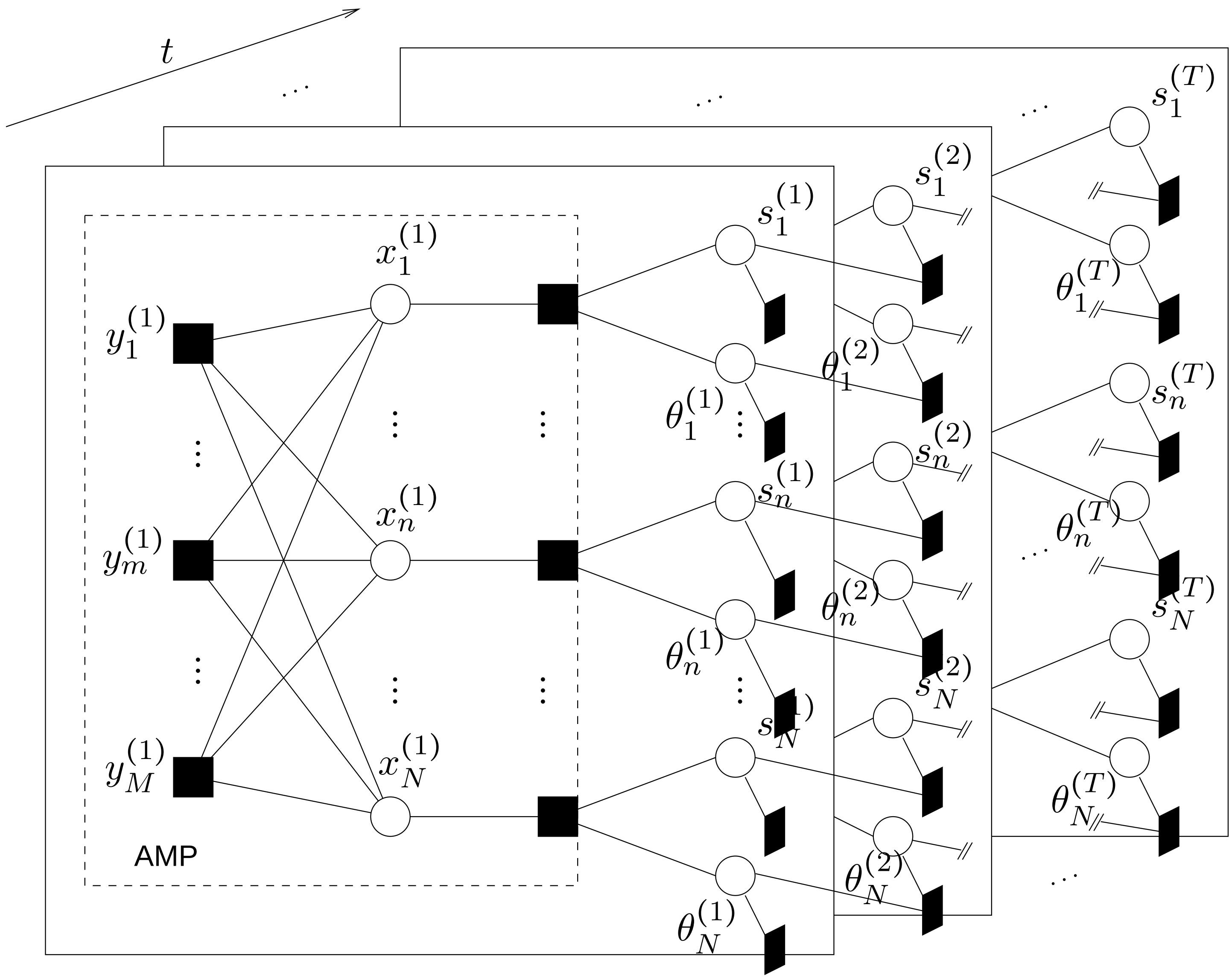
$$\theta_n^{(t)} = (1-\alpha)\theta_n^{(t-1)} + \alpha v_n^{(t)}, \quad v_n^{(t)} \sim \text{i.i.d } \mathcal{N}(0, \sigma_v^2)$$

$$p(s_n^{(t)} \,|\, s_n^{(t-1)}): \text{ state transition matrix } \begin{pmatrix} p^{00} & 1-p^{00} \\ 1-p^{11} & p^{11} \end{pmatrix}$$

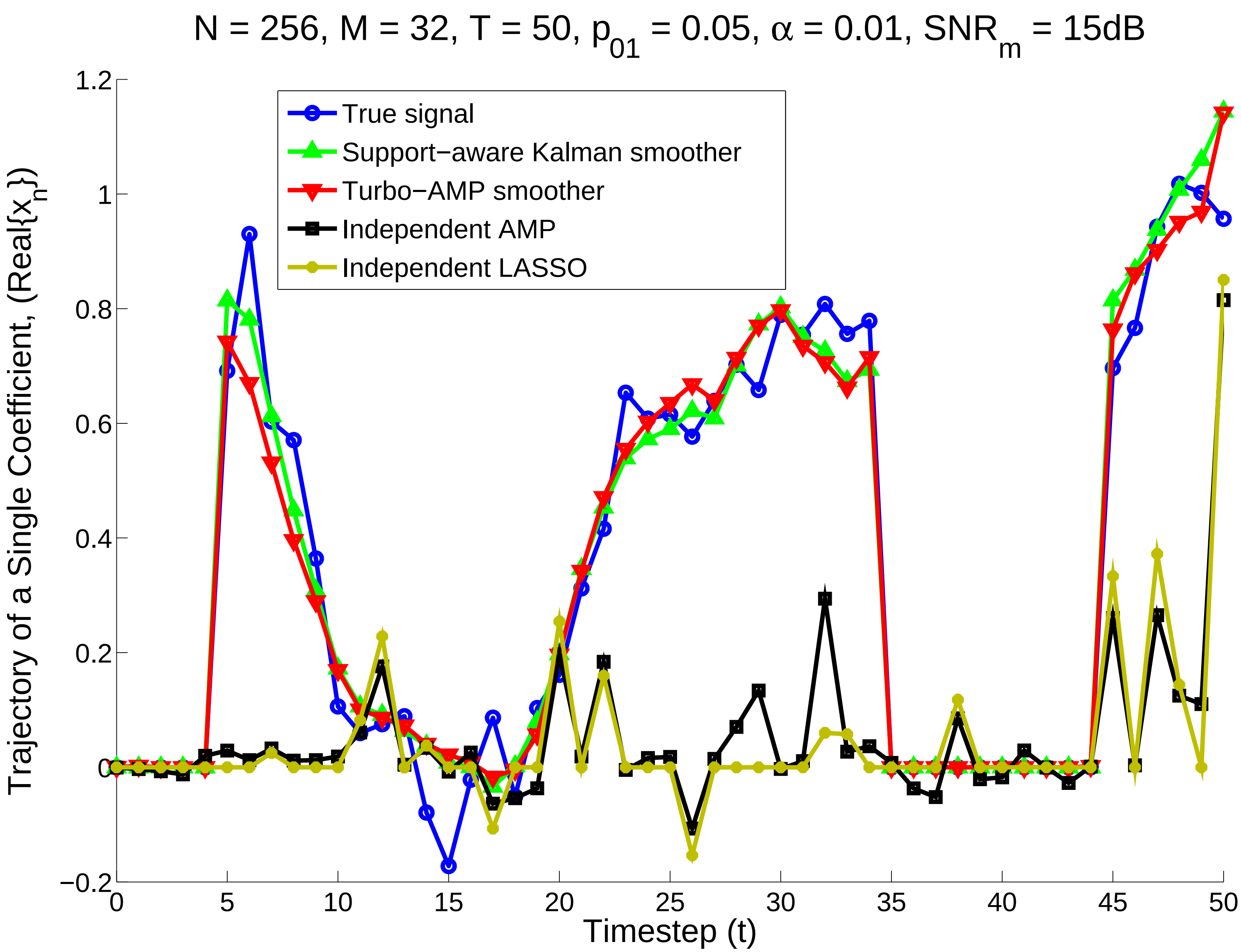
  where, as before, the model parameters $\{\sigma_w^2, \sigma_v^2, \alpha, p^{00}, p^{11}\}$ are treated as random (with non-informative hyperpriors) and learned from the data.

- Note: Our message-passing framework allows a unified treatment of *tracking* (i.e., causal estimation of $\{\boldsymbol{x}^{(t)}\}_{t=1}^T$) and *smoothing*.

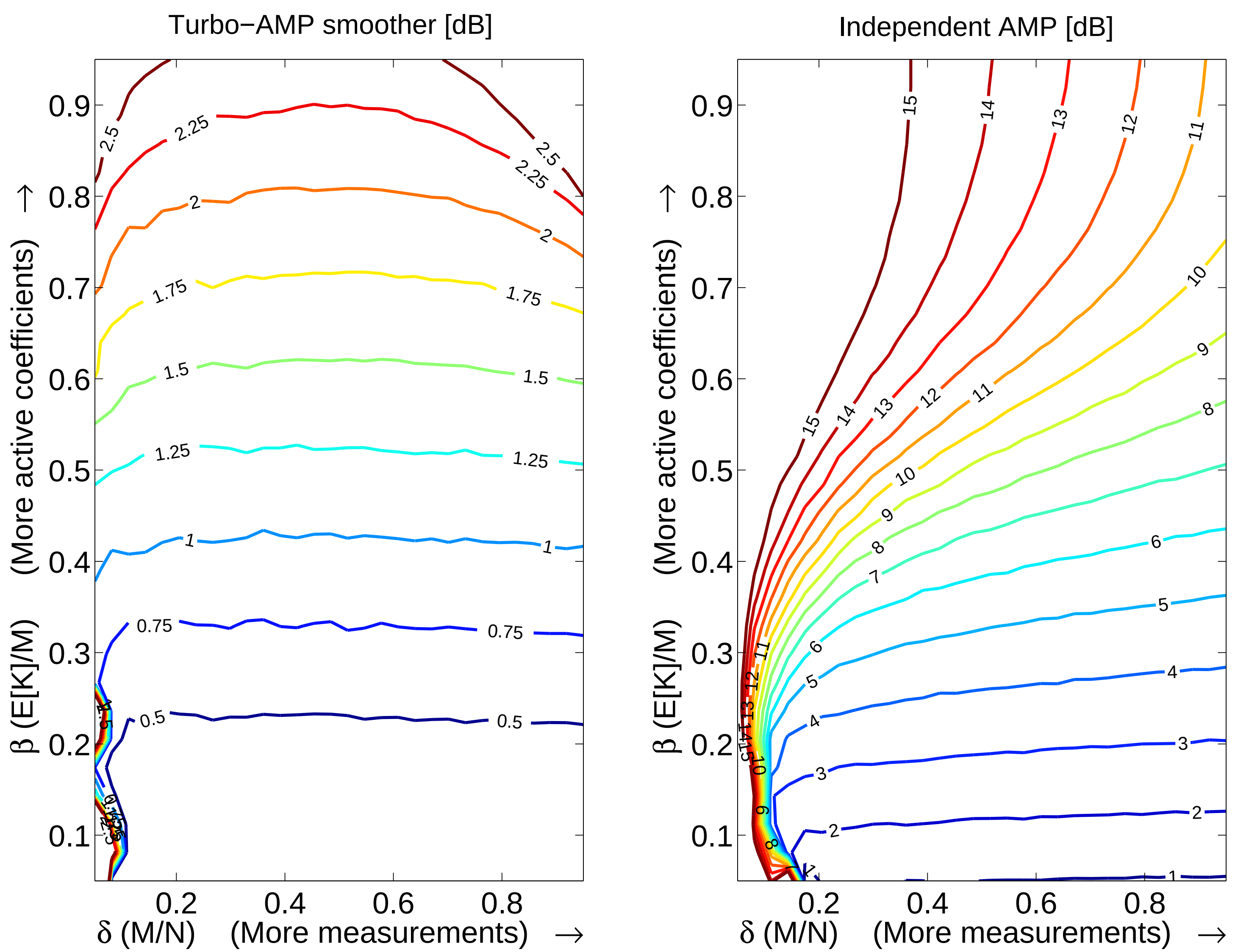
## Factor graph for compressive tracking:



# Example coefficient trajectories:



# NMSE relative to the support-aware Kalman smoother:



## 3) Communication over sparse channels:

- Consider communicating reliably over a channel that is
  - Rayleigh block-fading with block length $B$,
  - frequency-selective with delay spread $N$,
  - sparse with $K$ non-zero impulse-response coefs $\{x_n\}_{n=1}^N$,

  where both coefs and support are unknown to the transmitter & receiver.
- The ergodic capacity is $\boxed{C(\text{SNR}) = \frac{B-K}{B}\log(\text{SNR}) + \mathcal{O}(1) \text{ for high SNR.}}$
- Say, with $B$-subcarrier OFDM, we use $M$ pilot subcarriers, yielding

$$\boldsymbol{y}_\text{p} = \boldsymbol{D}_\text{p}\boldsymbol{\Phi}_\text{p}\boldsymbol{\Psi}\boldsymbol{x} + \boldsymbol{w}_\text{p}$$

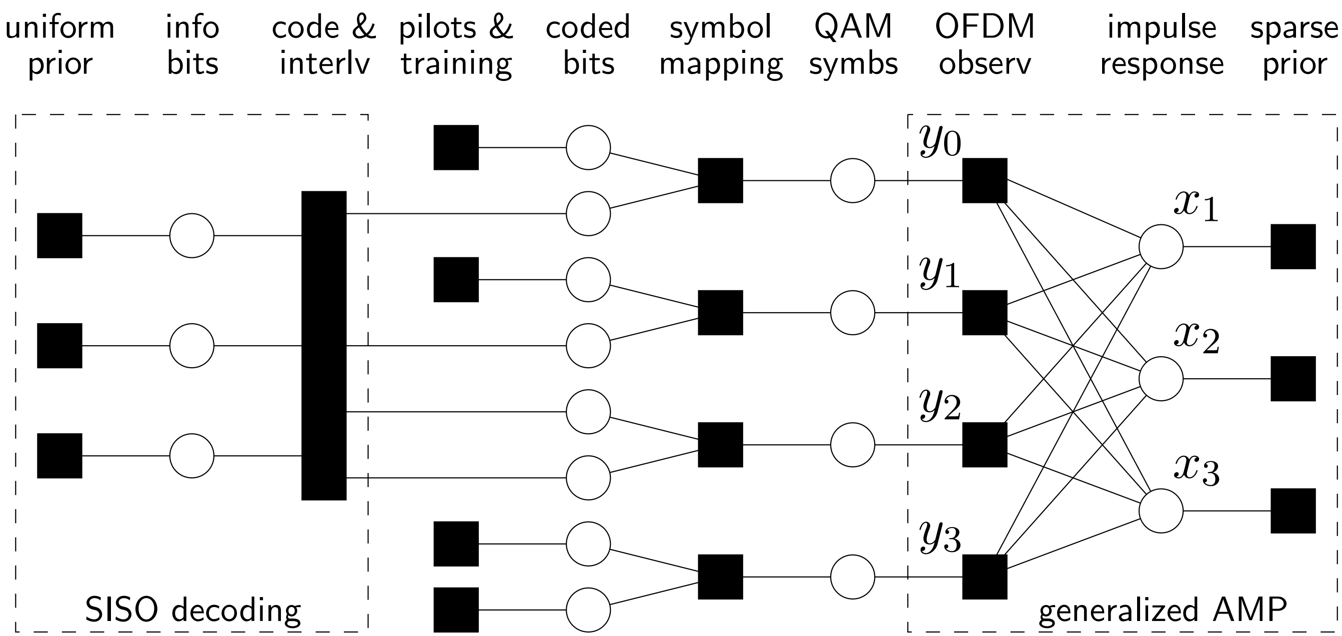
  with known diagonal pilot matrix $\boldsymbol{D}_\text{p}$, selection matrix $\boldsymbol{\Phi}_\text{p}$, and Fourier $\boldsymbol{\Psi}$.
- In "compressed channel sensing" (CCS), we estimate $\boldsymbol{x}$ from $\boldsymbol{y}_\text{p}$ and then assume $\boldsymbol{x} = \hat{\boldsymbol{x}}$ when decoding the $(B-M)$ data subcarriers

$$\boldsymbol{y}_\text{d} = \boldsymbol{D}_\text{d}\boldsymbol{\Phi}_\text{d}\boldsymbol{\Psi}\boldsymbol{x} + \boldsymbol{w}_\text{d}$$

  RIP analyses suggest the need for $M = \mathcal{O}(K \operatorname{polylog} N)$ pilots, but communicating near capacity requires using no more than $M = K$ pilots!

## Rethinking communication over sparse channels:

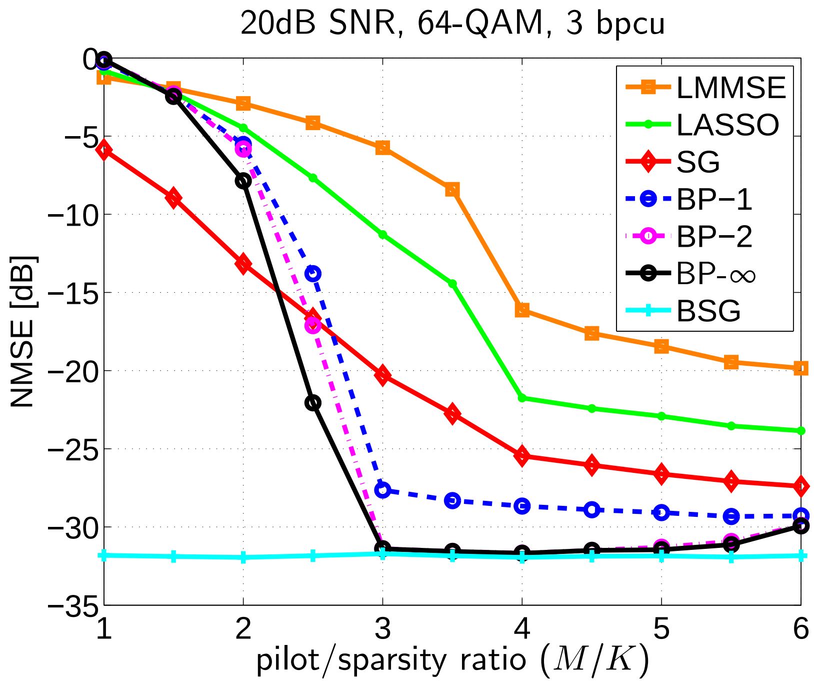
- The fundamental problem with the conventional CCS approach is the *separation* between channel-estimation and decoding.
- To communicate at rates near capacity, we need *joint* estimation/decoding, which is feasible using approximate belief propagation:

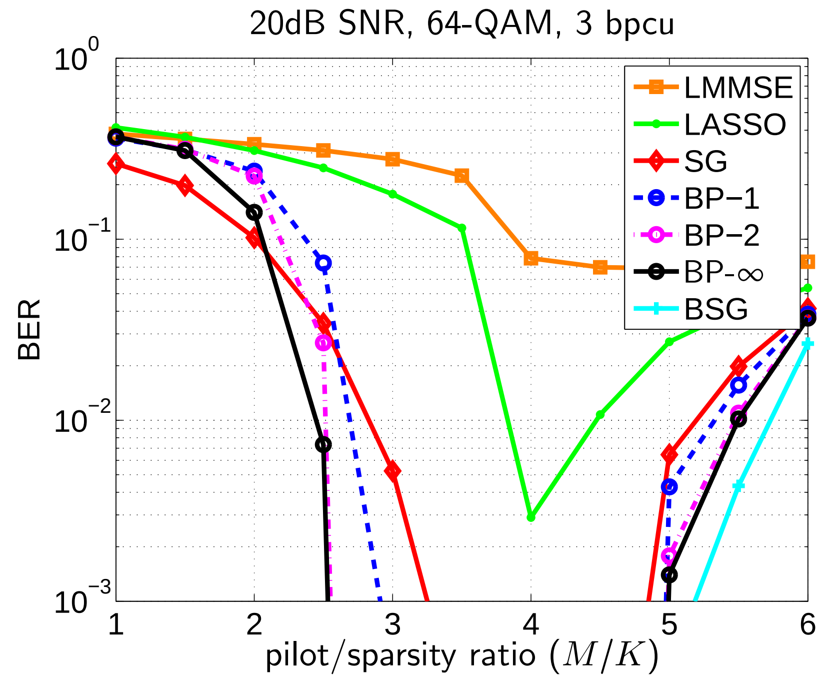



- Note: The sparse channel posterior is used in a larger detection problem!
- Note: We can now place pilots at the bit-level, rather than the symbol level.

## NMSE & BER versus pilot/sparsity ratio ($M/K$):

- Assume $B = 1021$ subcarriers with $\mathrm{E}\{K\} = 64$-sparse channels of length $N = 256$.

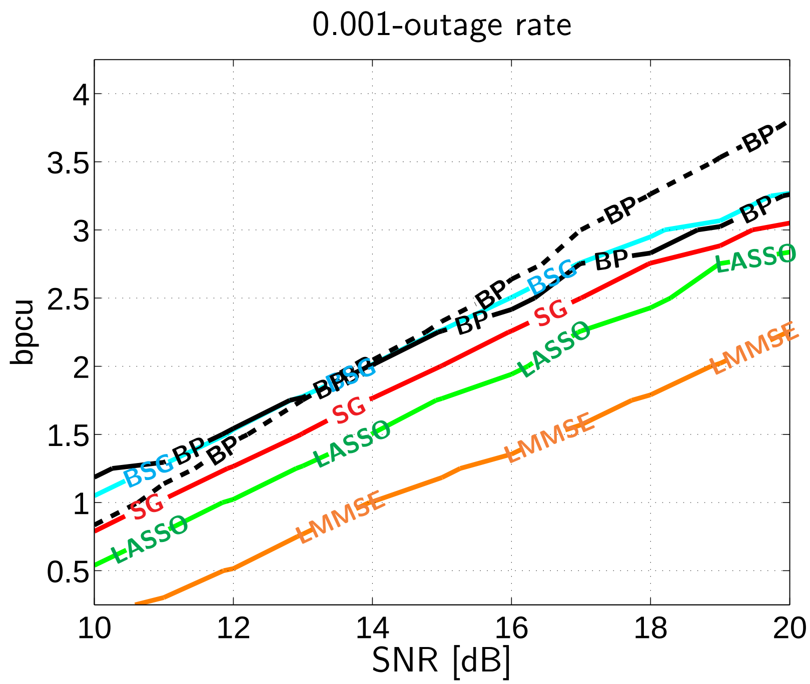


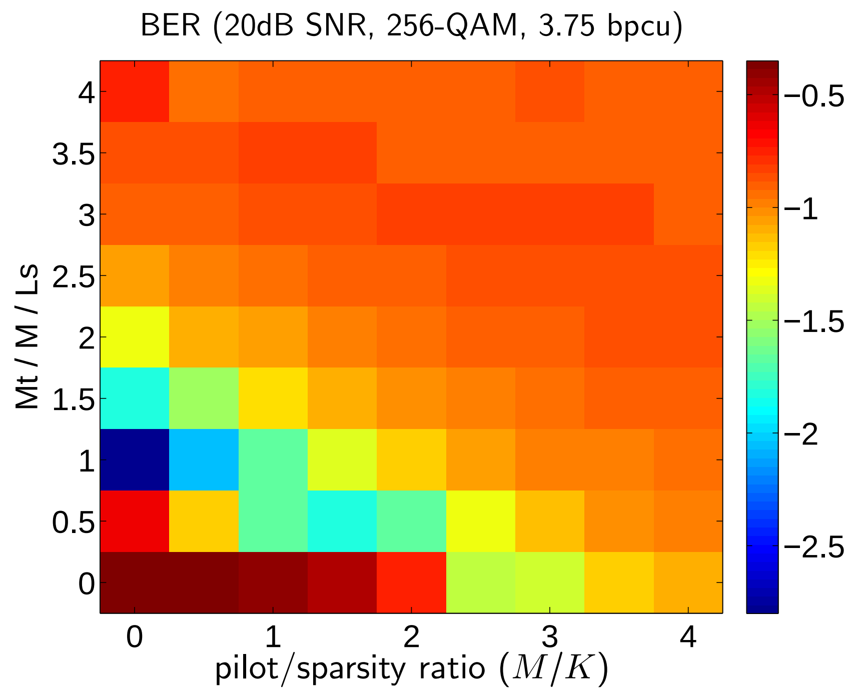




| implementable schemes | reference schemes |
|---|---|
| LMMSE = LMMSE-based CCS<br>LASSO = LASSO-based CCS<br>BP-n = BP after n turbo iterations | SG = support-aware genie<br>BSG = bit- and support-aware genie |

- For the plots above, we used $M$ uniformly spaced pilot subcarriers.
- Since spectral efficiency is fixed, more pilots necessitates a weaker code!

# Outage rate and the importance of bit-level pilots:




- Solid-line rates used 64-QAM and $M = 4K = N$ pilot subcarriers.
- Dashed-line rate used 256-QAM with $K \log_2(256)$ pilot bits as MSBs.

*BP-based scheme achieves the channel capacity's prelog factor!*

## Conclusions:

- The AMP algorithm of Donoho/Maleki/Montanari offers a state-of-the-art solution to the *classical* compressed sensing problem.
- Using the graphical-models framework, more complicated compressive inference tasks, with
  - structured signals (e.g., Markov structure in imaging & tracking),
  - structured non-Gaussian observations (e.g., code structure in comms),
  - self-calibration (e.g., noise variance, sparsity, Markov parameters), and
  - posterior reporting,

  can be solved using a "turbo" approach with AMP as a sub-block.
- Future work includes
  - applying turbo-AMP approach to challenging new problems, and
  - rigorously analyzing turbo-AMP convergence/performance.

*Thanks!*

## Bibliography:


- D. L. Donoho, A. Maleki, and A. Montanari, "Message passing algorithms for compressed sensing," *Proc. National Academy of Sciences*, Nov. 2009.
- D. L. Donoho, A. Maleki, and A. Montanari, "Message passing algorithms for compressed sensing: I. Motivation and construction," *Information Theory Workshop*, Jan. 2010
- M. Bayati and A. Montanari, "The dynamics of message passing on dense graphs, with applications to compressed sensing," arXiv:1001.3448, Jan. 2010.
- D. Guo and C.-C. Wang, "Random sparse linear systems observed via arbitrary channels: A decoupling principle," *ISIT*, June 2007.
- S. Rangan, "Generalized approximate message passing for estimation with random linear mixing," arXiv:1010.5141, Oct 2010.
- P. Schniter, "Turbo reconstruction of structured sparse signals," *CISS* (Princeton), Mar. 2010.
- S. Som, L. C. Potter, and P. Schniter, "Compressive imaging using approximate message passing and a Markov-tree prior," *Asilomar Conf.*, Nov. 2010.
- J. Ziniel, L. C. Potter, and P. Schniter, "Tracking and smoothing of time-varying sparse signals via approximate belief propagation," *Asilomar Conf.*, Nov. 2010.
- P. Schniter, "Joint estimation and decoding for sparse channels via relaxed belief propagation," *Asilomar Conf.*, Nov. 2010.