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

Phil Schniter

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

June 27, 2012

Outline:

1. Motivation.
   (a) the need for non-linear inference schemes,
   (b) some problems if interest.
2. The focus of this talk:
   (a) compressive sensing (in theory),
   (b) compressive sensing (in practice).
3. Recent approaches to these respective problems:
   (a) approximate message passing (AMP),
   (b) turbo-AMP.
4. Illustrative applications of turbo-AMP:
   (a) compressive imaging,
   (b) compressive tracking,
   (c) communication over sparse channels.

Motivations for nonlinear inference:

- **Linear inference** (e.g., matched filtering, linear equalization, least-squares, Kalman filtering, etc.) has been extremely popular in engineering and statistics due to computational efficiency and a well-developed theory.
  - Indeed, linear inference is **optimal** in problems well-modeled by linear observations and Gaussian signal and noise.
- In many cases, though, linear inference is **not good enough**.
  - The signal or noise may be **non-Gaussian**, or the observation mechanism may be **nonlinear**, in which case linear inference is suboptimal.
  - For example, the observations may be "**compressed**" (i.e., sampled below the Nyquist rate), in which case nonlinear inference becomes essential.
- But is there an **accurate** and **computationally efficient** framework for high-dimensional nonlinear inference?
  - For a wide (and expanding) range of problems, Yes!
  - Based on "belief propagation" or "**message passing**."

A few problems of interest:

Linear additive:

- $\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x} + \boldsymbol{w}$ with $\boldsymbol{A}$ known, $\boldsymbol{x} \sim p_{\boldsymbol{x}}$, $\boldsymbol{w} \sim p_{\boldsymbol{w}}$
- examples: communications, imaging, radar, compressive sensing (CS).

Generalized linear:

- $\boldsymbol{y} \sim p(\boldsymbol{y}|\boldsymbol{z})$ for $\boldsymbol{z} = \boldsymbol{A}\boldsymbol{x}$ with $\boldsymbol{A}$ known, $\boldsymbol{x} \sim p_{\boldsymbol{x}}$
- examples: quantization, phase retrieval, classification.

Generalized bilinear:

- $\boldsymbol{Y} \sim p(\boldsymbol{Y}|\boldsymbol{Z})$ for $\boldsymbol{Z} = \boldsymbol{A}\boldsymbol{X}$ with $\boldsymbol{A} \sim p_{\boldsymbol{A}}$, $\boldsymbol{X} \sim p_{\boldsymbol{X}}$
- examples: dictionary learning, matrix completion, robust PCA.

Parametric nonlinear:

- $\boldsymbol{y} \sim p(\boldsymbol{y}|\boldsymbol{z})$ for $\boldsymbol{z} = \boldsymbol{A}(\boldsymbol{\theta})\boldsymbol{x}$ with $\boldsymbol{A}(\cdot)$ known, $\boldsymbol{\theta} \sim p_{\boldsymbol{\theta}}$, $\boldsymbol{x} \sim p_{\boldsymbol{x}}$
- examples: frequency estimation, calibration, autofocus.

Compressive sensing (in theory):

- 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 well-behaved (e.g., satisfies RIP), the sparsity of $\boldsymbol{x}$ can be exploited for provably accurate reconstruction with computationally efficient algorithms.
  - Caution: usually need to **tune** an algorithmic parameter that balances sparsity with data fidelity. If using "cross-validation," this can be expensive!

- Such $\boldsymbol{A}$ 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 $\boldsymbol{\Phi}$).

Compressive sensing (in practice):

- Usually, real-world applications exhibit **additional structure**. . .
  - in the **support** of large signal coefficients (e.g., block, tree, etc.),
  - among the **values** of large signal coefficients (e.g., correlation, coherence),

  and exploitation of these additional structures may be essential.
- But, exploiting this additional structure **complicates tuning**, since...
  - many more parameters are involved in the model, and
  - mismatch in these parameters can severely bias the signal estimate.
- Also, many real-world applications are **not content with point estimates**. . .
  - since the estimates may be later used for decision-making, control, etc.,
  - in which case confidence intervals are needed, or preferably the *full posterior probability distribution* on the unknowns.

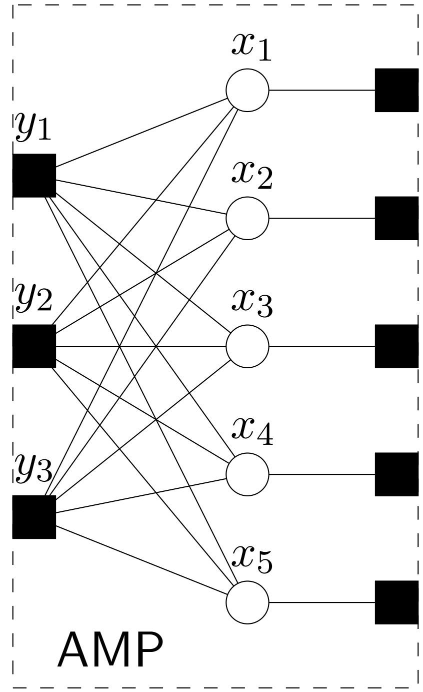
Solving the theoretical CS problem — AMP:

- **Approximate message passing** (AMP) [Donoho/Maleki/Montanari 2009/10] refers to a family of signal reconstruction algorithms that are
  - designed to solve the *theoretical* CS problem,
  - inspired by principled approximations of **belief propagation**.
- AMP highlights:
  - Very computationally efficient: a form of iterative thresholding.
  - Very high performance (with sufficiently large N, M):
    - Can be configured to produce near-MAP or near-MMSE estimates.
  - Admits rigorous asymptotic analysis [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 belief propagation on sparse matrices, where marginal posterior distributions are known to be asymptotically optimal.

Solving practical compressive inference problems — Turbo-AMP:

- The **Bayesian graphical-model framework** is a flexible and powerful way to incorporate and exploit probabilistic structure.

Simple sparsity with known model parameters

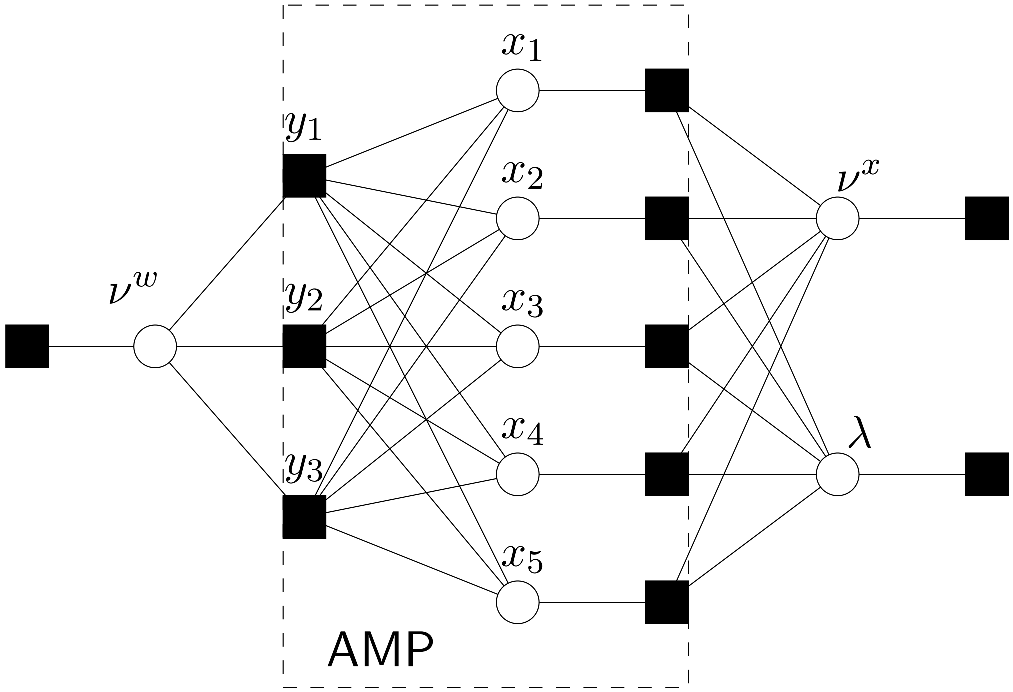


$$p(\boldsymbol{x}|\boldsymbol{y}) \propto \prod_{m=1}^{M} p(y_m|\boldsymbol{x}) \prod_{n=1}^{N} p(x_n)$$

Solving practical compressive inference problems — Turbo-AMP:

- The **Bayesian graphical-model framework** is a flexible and powerful way to incorporate and exploit probabilistic structure.

Simple sparsity with unknown model parameters

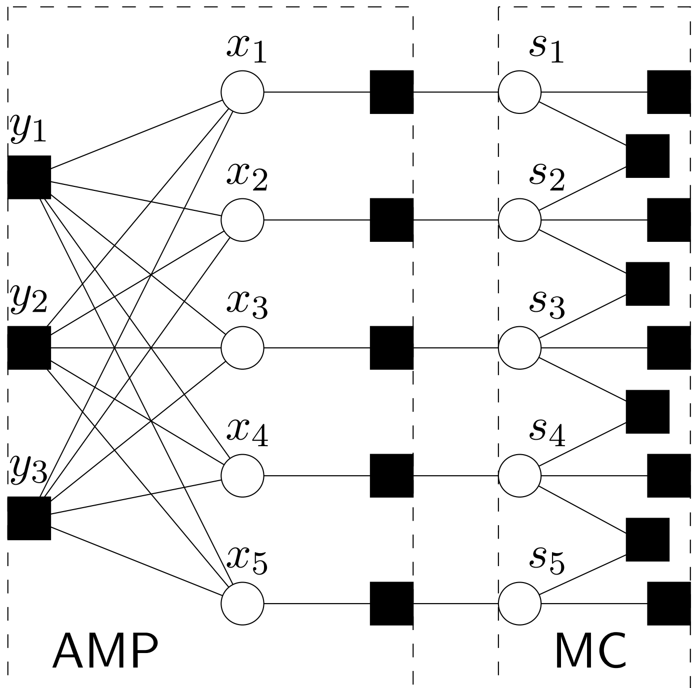


Or treat (ν^w, ν^x, λ) as deterministic unknowns, and do ≈ML estimation via EM.

Solving practical compressive inference problems — Turbo-AMP:

- The **Bayesian graphical-model framework** is a flexible and powerful way to incorporate and exploit probabilistic structure.

Structured sparsity with known model parameters:

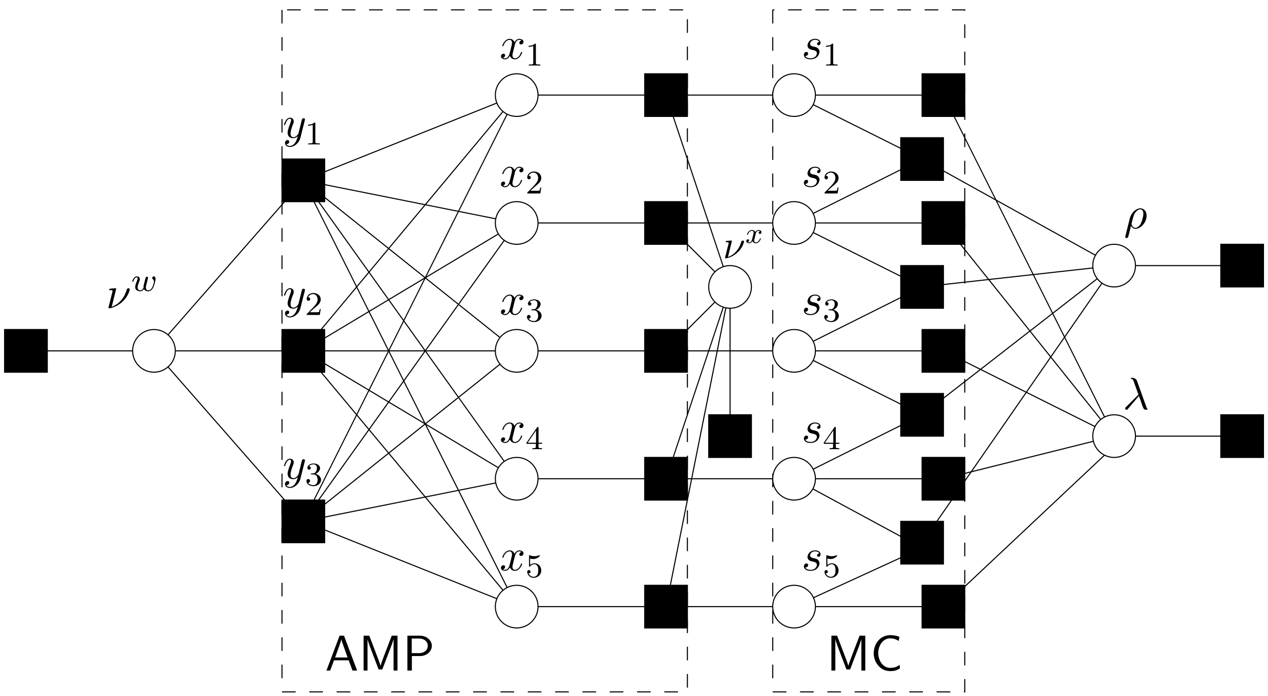


- For these problems, AMP is used as a soft-input soft-output inference block, like a "channel decoder" in a "turbo" receiver. [Schniter CISS 10]

Solving practical compressive inference problems — Turbo-AMP:

- The **Bayesian graphical-model framework** is a flexible and powerful way to incorporate and exploit probabilistic structure.

Structured sparsity with unknown model parameters:

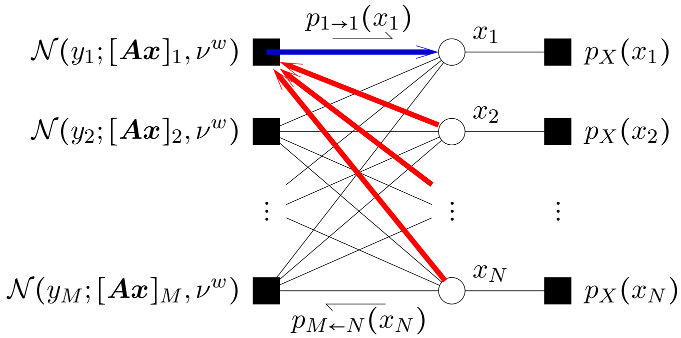


- For these problems, AMP is used as a soft-input soft-output inference block, like a "channel decoder" in a "turbo" receiver. [Schniter CISS 10]

So what approximations lead to AMP?:

Assume sum-product form of AMP. Then...

1. Message from y_i node to x_j node:



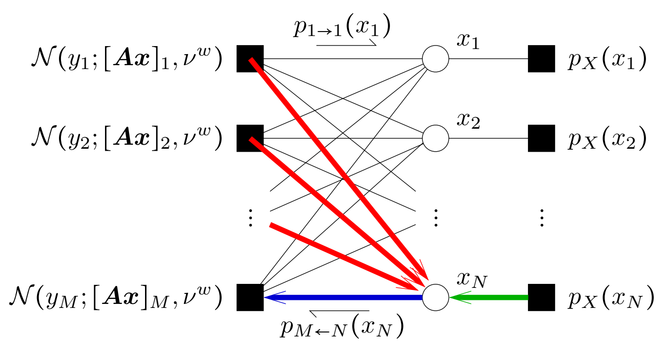
$$p_{i\to j}(x_j) \propto \int_{\{x_r\}_{r\neq j}} \mathcal{N}\big(y_i; \overbrace{\textstyle\sum_r a_{ir}x_r}^{\approx \mathcal{N} \text{ via CLT}}, \nu^w\big) \textstyle\prod_{r\neq j} p_{i\leftarrow r}(x_r)$$

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

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!**

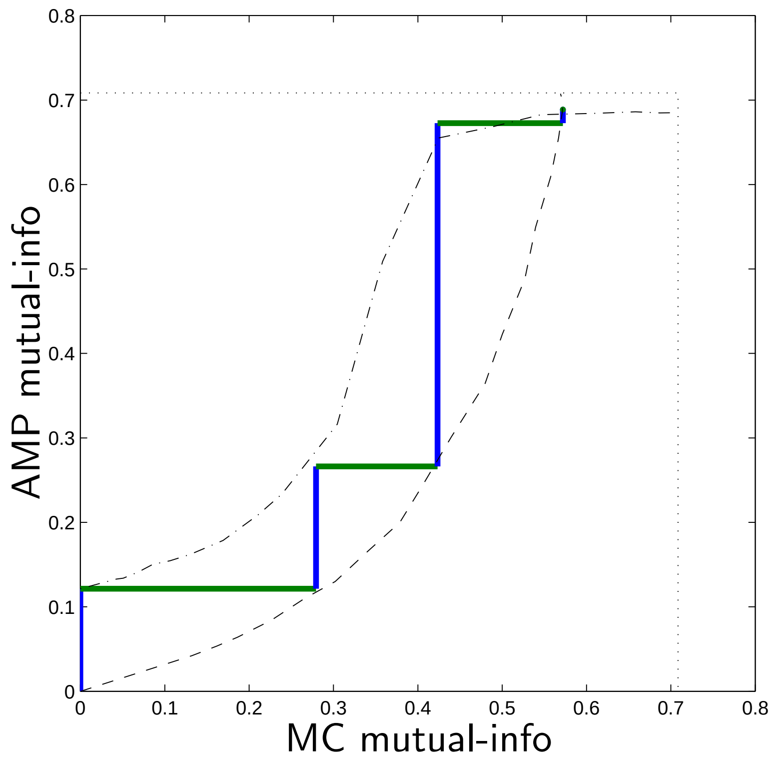
Remaining problem: we have $2MN$ messages to compute (too many!).

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**!



Extrinsic information transfer (EXIT) charts:

EXIT charts, developed to predict the convergence of turbo decoding [ten Brink 01], can help to understand the interaction between turbo-AMP inference blocks:



In this EXIT chart, we are plotting the mutual-information between the true and (AMP or MC)-estimated support pattern $\{s_n\}$.

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

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.

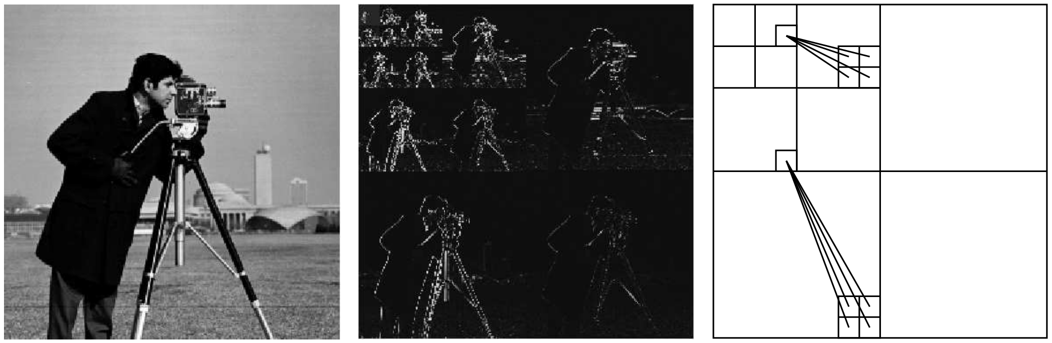
3. **Communication over sparse channels**

   . . . involving a generalized linear model, and

   . . . where AMP is embedded in a larger factor graph.

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**:

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

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

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

- The model parameters ν^w and $\{\nu_j, p_j^{00}, p_j^{11}\}_{j=0}^{J}$ are treated as random with non-informative hyperpriors (Gamma and Beta, respectively). We approximate those messages by passing only the means.

Comparison to other methods:

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

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

Algorithm	Authors (year)	Time	NMSE
IRWL1	Duarte, Wakin, Baraniuk (2008)	363 s	-14.4 dB
ModelCS	Baraniuk, Cevher, Duarte, Hegde (2010)	117 s	-17.4 dB
Variational Bayes	He, Chen, Carin (2010)	107 s	-19.0 dB
MCMC	He & Carin (2009)	742 s	-20.1 dB
Turbo-AMP	Som & Schniter (2010)	**51 s**	**-20.7 dB**

Turbo-AMP beats other approaches simultaneously in speed and accuracy!

Comparison to other methods:

True

ModelCS

IRWL1

Variational Bayes

MCMC

Turbo-AMP

2) Compressive tracking / Dynamic compressive sensing:

- Now say we observe the **vector sequence**

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

with 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:

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

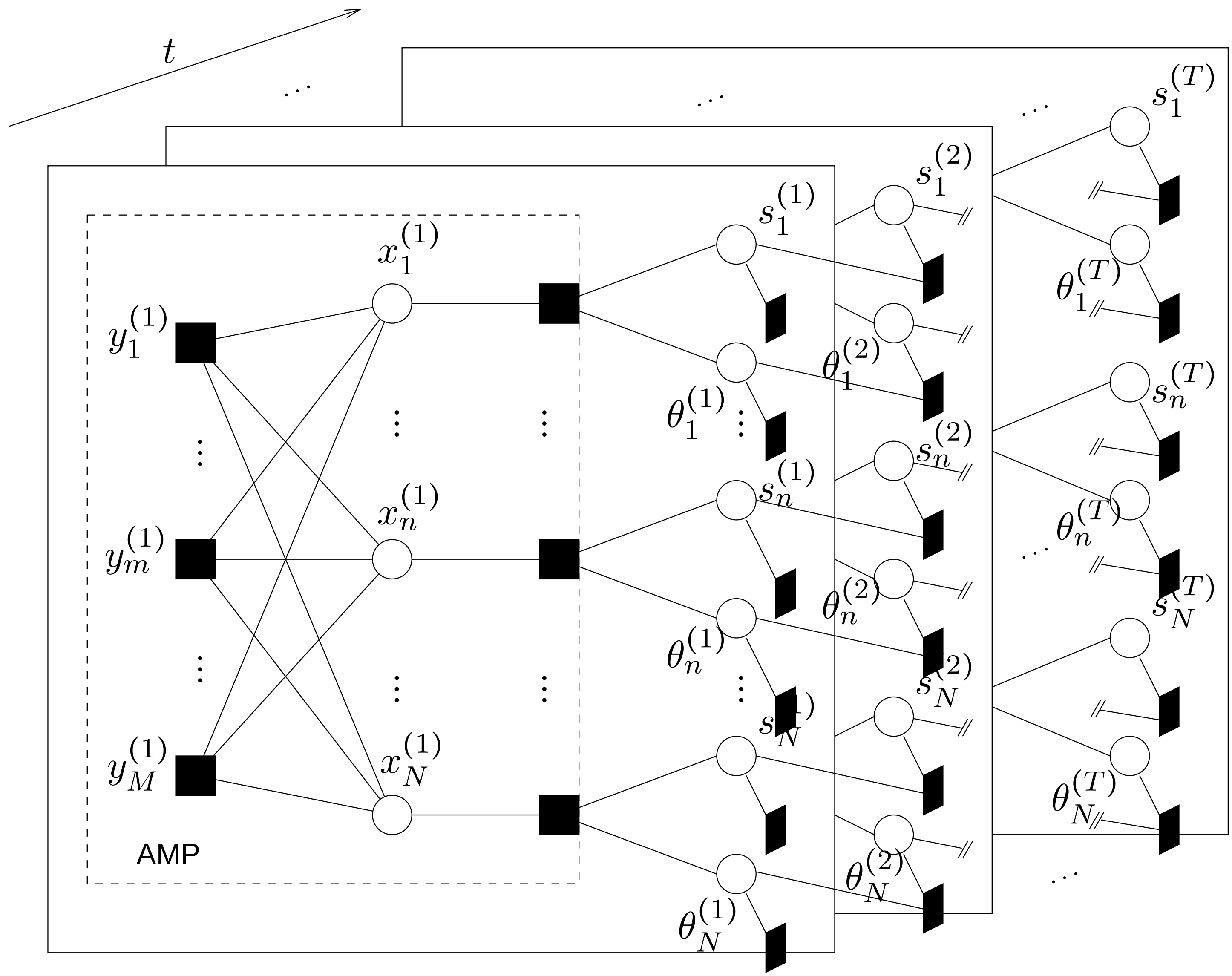
$$\theta_n^{(t)} = (1-\alpha)\theta_n^{(t-1)} + \alpha v_n^{(t)}, \quad v_n^{(t)} \sim \text{i.i.d } \mathcal{N}(0, \nu^v)$$

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

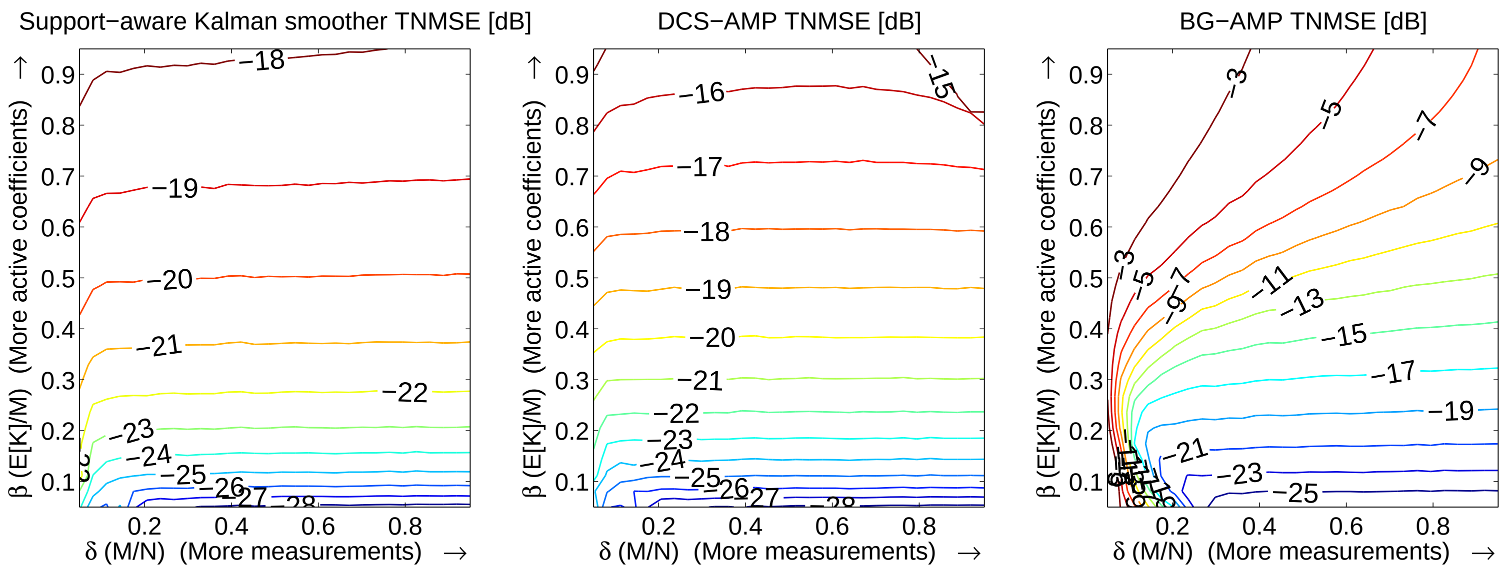
where here the model parameters $\{\nu^w, \nu^v, \alpha, p^{00}, p^{11}\}$ are treated as deterministic unknowns and learned using the EM algorithm.

- 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/smoothing:



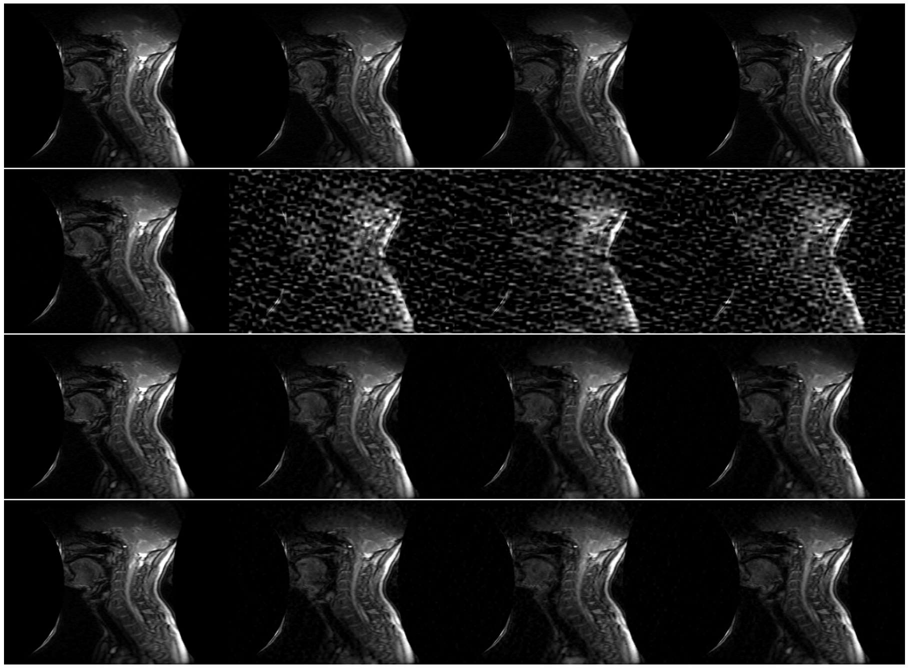
Near-optimal MSE performance:



- With a Bernoulli-Gaussian signal, the **support-aware Kalman smoother** provides an **oracle bound** on MSE.
- The proposed “dynamic” turbo-AMP performs very close to the bound, and much better than standard AMP (which does not exploit temporal structure).

Performance on dynamic-MRI dataset:

Frames 1, 2, 5, and 10 of a dynamic MRI image sequence:



Algorithm	**TNMSE (dB)**	**Runtime**
Basis Pursuit	-17.22	47 min
Modified-CS [Vaswani/Lu 09]	-34.30	7.39 hrs
Turbo-AMP (Filter)	**-34.62**	**8.08 sec**

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 impulse response $\boldsymbol{x}$ with $K < N$ non-zero coefs,

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

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

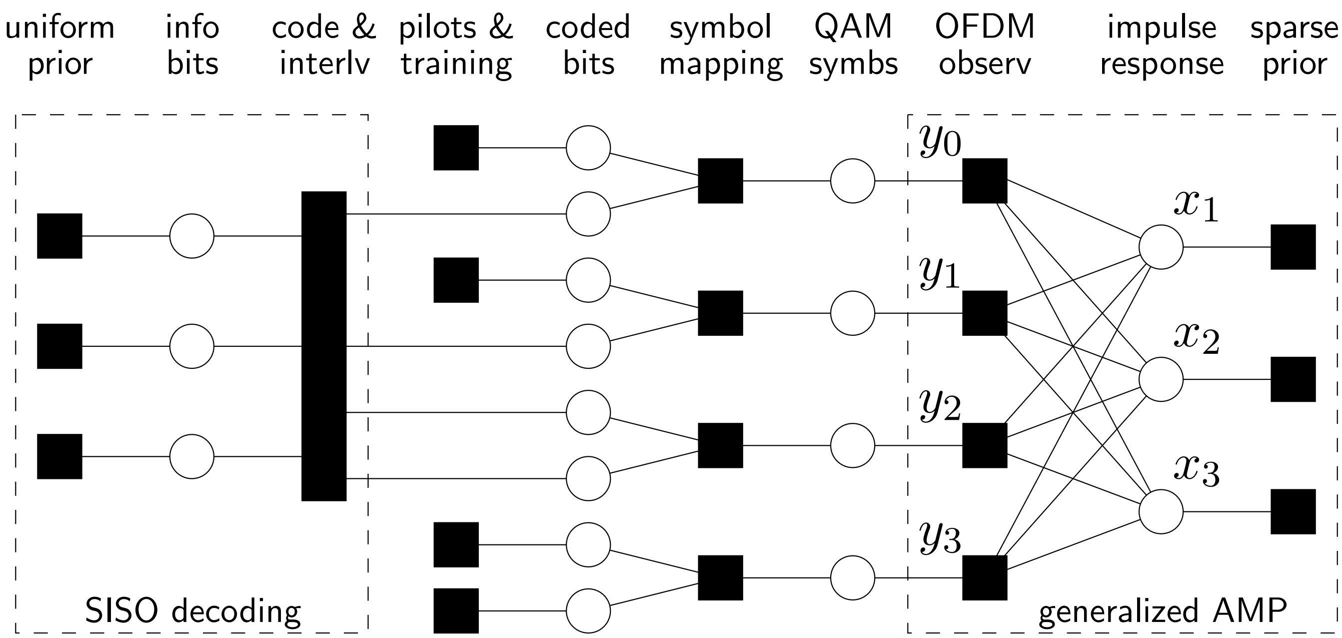
  with known diagonal pilot matrix $\boldsymbol{D}_{\text{p}}$, selection matrix $\boldsymbol{\Phi}_{\text{p}}$, and DFT $\boldsymbol{\Psi}$.
- In "**compressed channel sensing**" (CCS), the channel $\boldsymbol{x}$ is estimated from $\boldsymbol{y}_{\text{p}}$ and the resulting $\hat{\boldsymbol{x}}$ is used to decode 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 prescribe the use of $M = \mathcal{O}(K \operatorname{polylog} N)$ pilots, but communicating near capacity requires using no more than $M = K$ pilots!

Rethinking communication over sparse channels:

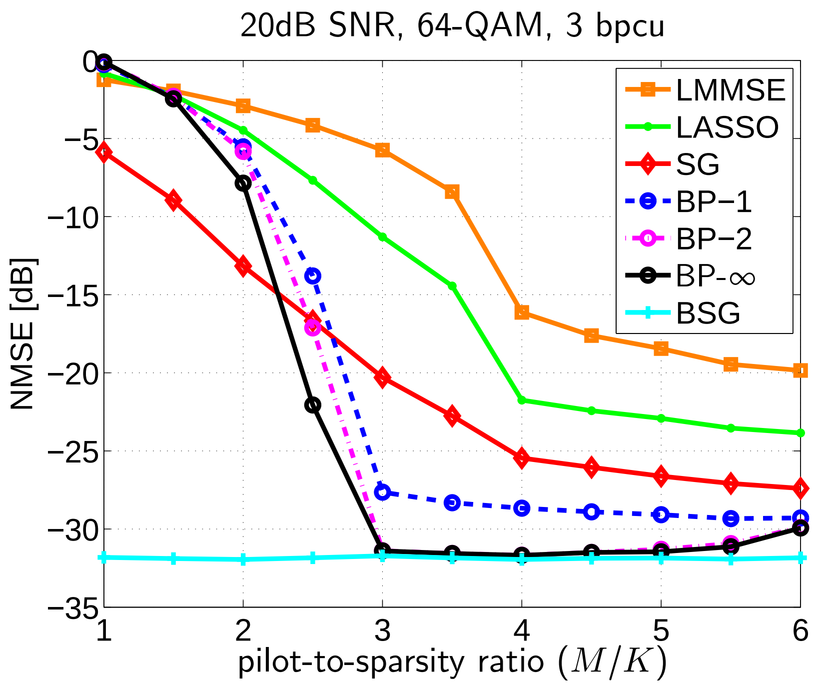
- The fundamental problem with the conventional CCS approach is the separation between channel-estimation and data decoding.
- To communicate at rates near capacity, we need joint estimation/decoding, which we can do using **turbo-AMP**:

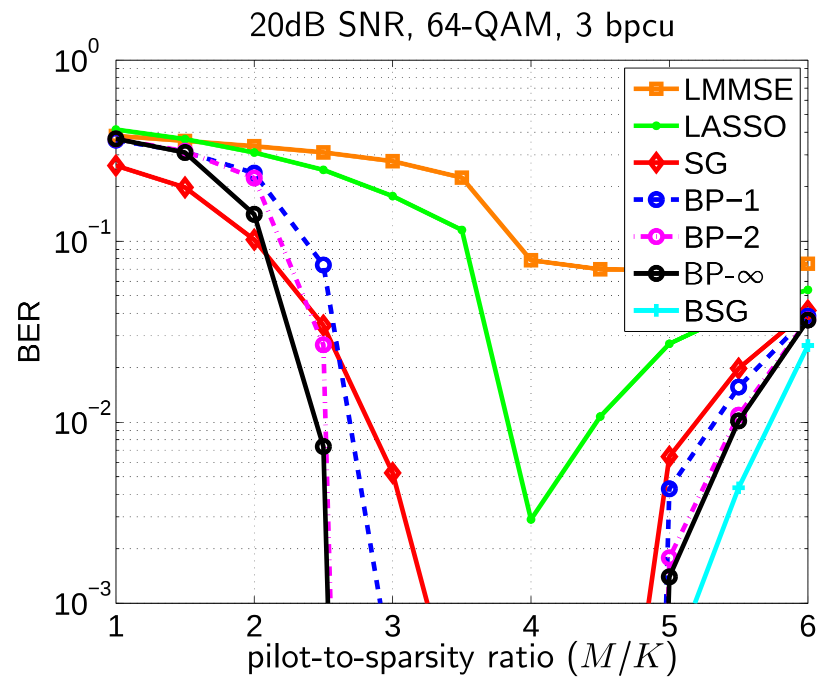


- Note: Here we need to use the generalized AMP from [Rangan 10]
- Note: We can now place pilots at the bit-level, rather than the symbol level.

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

- Assume $B = 1024$ subcarriers with $K = 64$-sparse channels of length $N = 256$.



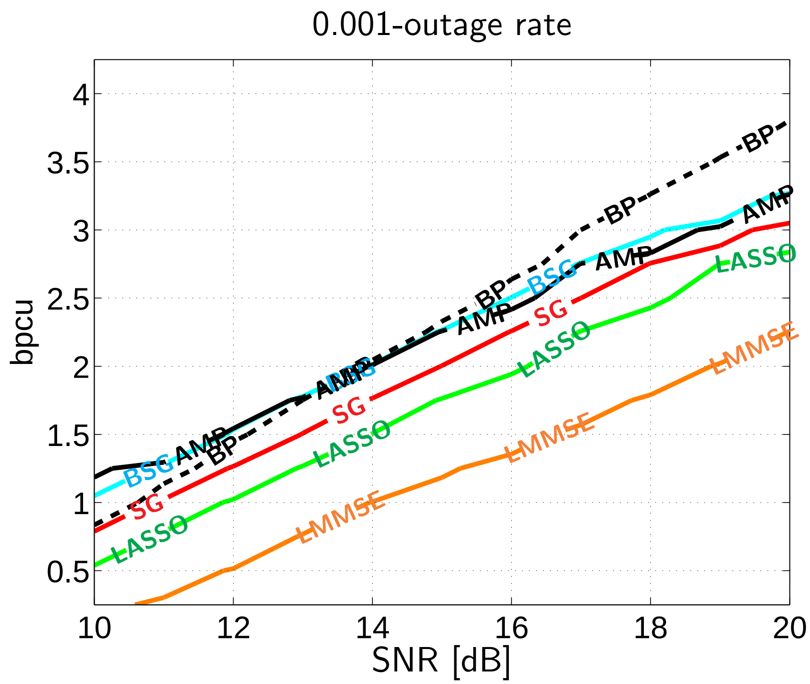


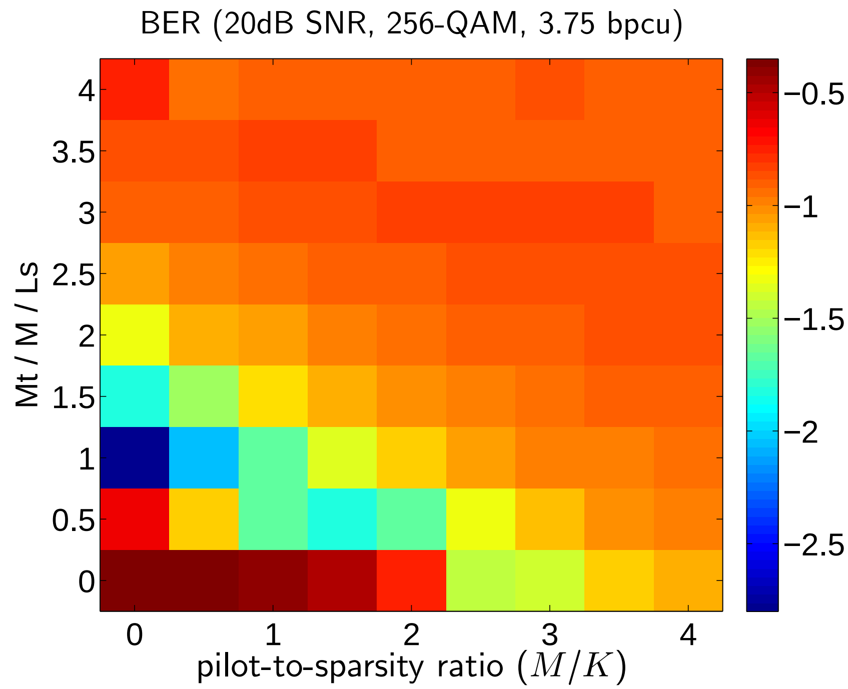
implementable schemes	reference schemes
LMMSE = LMMSE-based CCS LASSO = LASSO-based CCS BP-n = BP after n turbo iterations	SG = support-aware genie BSG = bit- and support-aware genie

- 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.

turbo-AMP achieves the channel capacity's prelog factor!

Conclusions:

- The **AMP** algorithm of Donoho/Maleki/Montanari offers a state-of-the-art solution to the *theoretical* compressed sensing problem.
- Using a graphical-models framework, we can handle more complicated compressive inference tasks, with
  - structured signals (e.g., Markov structure in imaging & tracking),
  - structured generalized-linear measurements (e.g., code structure in comms),
  - self-tuning (e.g., noise variance, sparsity, Markov parameters), and
  - soft outputs,

  using the **turbo-AMP** approach, which leverages AMP as a sub-block.
- Ongoing work includes
  - **applying** turbo-AMP approach to challenging new problems, and
  - **analyzing** turbo-AMP convergence/performance.
- Matlab code is available at

  `http://ece.osu.edu/~schniter/EMturboGAMP`

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," *Trans. IT*, Feb. 2011.
- 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 and P. Schniter, "Compressive imaging using approximate message passing and a Markov-tree prior," *Trans. SP* 2012. (See also Asilomar 2010)
- J. Ziniel and P. Schniter, "Dynamic Compressive Sensing of Time-Varying Signals via Approximate Message Passing," *arXiv:1205.4080*, 2012. (See also Asilomar 2010)
- W. Lu and N. Vaswani, "Modified compressive sensing for real-time dynamic MR imaging," *ICIP*, Nov. 2009.
- A. P. Kannu and P. Schniter, "On Communication over Unknown Sparse Frequency-Selective Block-Fading Channels," *Trans. IT*, Oct. 2011.
- P. Schniter, "'A Message-Passing Receiver for BICM-OFDM over Unknown Clustered-Sparse Channels," *JSTSP*, 2011. (See also Asilomar 2010)