# Exploiting Structured Sparsity in Bayesian Experimental Design

Phil Schniter

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

CAMSAP 2011

## Outline:

1. Compressive sensing under **structured** sparsity
2. **Adaptive** compressive sensing via **Bayesian experimental design**
3. **Approximate message passing (AMP)** for structured-sparse recovery
4. How to make AMP (and other algorithms like LASSO) adaptive
5. Empirical performance close to **oracle bounds**.

## Compressive Sensing:

- In compressive sensing, we aim to recover a signal vector $\boldsymbol{u}$ from noisy **underdetermined** linear measurements

$$\boldsymbol{y} = \boldsymbol{\Phi u} + \boldsymbol{w} \in \mathbb{R}^M.$$

- Although the problem is underdetermined, accurate recovery maybe possible if $\boldsymbol{u}$ can be **sparsely** represented in some dictionary $\boldsymbol{\Psi}$, i.e.,

$$\boldsymbol{u} = \boldsymbol{\Psi x} \text{ for } K\text{-sparse } \boldsymbol{x} \in \mathbb{R}^N,$$

where $\boldsymbol{\Psi}$ is "incoherent" with $\boldsymbol{\Phi}$.

- It is common to choose $\boldsymbol{\Phi}$ **randomly** and apply the **LASSO** algorithm to recover an estimate $\hat{\boldsymbol{x}}$, in which case one can guarantee $\|\hat{\boldsymbol{x}} - \boldsymbol{x}\|_2^2 \leq C\|\boldsymbol{w}\|_2^2$, for some constant $C$, with

$$M \geq \mathcal{O}(K \log(N/K)) \text{ measurements.}$$

## Structured Sparsity:

- Often the signal $\boldsymbol{u}$ has a representation $\boldsymbol{x}$ that is not simply sparse but rather **structured sparse**.

  For examples,
  - wavelet coefficients of natural images are **tree-sparse**, and
  - impulse responses of wideband wireless channels are **clustered-sparse**.

- In this case, similar reconstruction guarantees are possible with only

$$M \geq \mathcal{O}(K) \text{ measurements}$$

  using structured-sparse recovery algorithms!

## Adaptive Compressive Sensing:

- In some applications, we can afford $T > 1$ measurement rounds and **adapt** the measurement matrix $\mathbf{\Phi}_t$ for the $t^{\text{th}}$ round based on the knowledge gained from previous rounds.

- In this case, the observation model changes to

$$\underbrace{\begin{bmatrix} \underline{\boldsymbol{y}}_{t-1} \\ \boldsymbol{y}_t \end{bmatrix}}_{\underline{\boldsymbol{y}}_t} = \underbrace{\begin{bmatrix} \underline{\mathbf{\Phi}}_{t-1} \\ \mathbf{\Phi}_t \end{bmatrix}}_{\underline{\mathbf{\Phi}}_t} \boldsymbol{u} + \underbrace{\begin{bmatrix} \underline{\boldsymbol{w}}_{t-1} \\ \boldsymbol{w}_t \end{bmatrix}}_{\underline{\boldsymbol{w}}_t \in \mathbb{R}^{\underline{M}_t}} \begin{matrix} \in \mathbb{R}^{\underline{M}_{t-1}} \\ \in \mathbb{R}^{M_t} \end{matrix},$$

where underbars are used to denote **cumulative** quantities.

*So, how is $\mathbf{\Phi}_t$ designed?*

- In Bayesian experimental design [DeGroot 62], $\mathbf{\Phi}_t$ is chosen to maximize the **expected information gain (EIG)**.

## Bayesian Experimental Design:

- The **information gain** is defined as the **KL divergence** between the **prior** and **posterior** distributions at measurement step $t$:

$$D(\boldsymbol{y}_t) \triangleq \int_{\boldsymbol{u}} q(\boldsymbol{u} \,|\, \boldsymbol{y}_t) \log \frac{q(\boldsymbol{u} \,|\, \boldsymbol{y}_t)}{q(\boldsymbol{u})},$$

where

$$q(\boldsymbol{u}) \triangleq p(\boldsymbol{u} \,|\, \underline{\boldsymbol{y}}_{t-1}) \text{ is the step-}t\textbf{ prior}\text{, and}$$

$$q(\boldsymbol{u} \,|\, \boldsymbol{y}_t) \triangleq p(\boldsymbol{u} \,|\, \underline{\boldsymbol{y}}_{t-1}, \boldsymbol{y}_t) \text{ is the step-}t\textbf{ posterior}.$$

- Since $\boldsymbol{y}_t$ is not yet known, we consider **expected** information gain:

$$\begin{aligned}\mathsf{EIG}_t \triangleq \mathrm{E}\{D(\boldsymbol{y}_t) \,|\, \underline{\boldsymbol{y}}_{t-1}\} &= \int_{\boldsymbol{y}_t} \underbrace{p(\boldsymbol{y}_t \,|\, \underline{\boldsymbol{y}}_{t-1})}_{\triangleq\, q(\boldsymbol{y}_t)} \int_{\boldsymbol{u}} q(\boldsymbol{u} \,|\, \boldsymbol{y}_t) \log \frac{q(\boldsymbol{u} \,|\, \boldsymbol{y}_t)}{q(\boldsymbol{u})} \\ &= \int_{\boldsymbol{y}_t} \int_{\boldsymbol{u}} q(\boldsymbol{u}, \boldsymbol{y}_t) \log \frac{q(\boldsymbol{u}, \boldsymbol{y}_t)}{q(\boldsymbol{u}) q(\boldsymbol{y}_t)} = \mathrm{I}(\mathbf{u}; \mathbf{y}_t),\end{aligned}$$

i.e., the **mutual information** between $\mathbf{u} \sim q(\boldsymbol{u})$ and $\mathbf{y}_t \sim q(\boldsymbol{y}_t)$.

## Gaussian Experimental Design:

- Evaluating the expected information gain is often **difficult**.
- However, when all distributions are **Gaussian**, it becomes easy.

  For example, if

$$\begin{aligned} \text{noise:} \quad & \boldsymbol{w} \sim \mathcal{N}(\mathbf{0}, v_w \boldsymbol{I}) \\ \text{step-}t \text{ signal prior:} \quad & \boldsymbol{u}|\underline{\boldsymbol{y}}_{t-1} \sim \mathcal{N}(\boldsymbol{\mu}_u, \boldsymbol{\Sigma}_u) \end{aligned}$$

  then it is straightforward to show that

$$\mathsf{EIG}_t = \tfrac{1}{2}\log\left|\tfrac{1}{v_w}\boldsymbol{\Phi}_t \boldsymbol{\Sigma}_u \boldsymbol{\Phi}_t^{\mathsf{T}} + \boldsymbol{I}\right|.$$

- Of course, in compressive sensing, the signal priors are **non-Gaussian** and thus the above could only be used after approximations are made.

## Gaussian design of $\mathbf{\Phi}_t$:

*What is the EIG-maximizing $\mathbf{\Phi}_t$ subject to the energy constraint $\|\mathbf{\Phi}_t\|_F^2 \leq \mathcal{E}$?*

- Previous works [Seeger 08, Ji/Xu/Carin 08] studied the case of one **scalar** measurement per step (i.e., $M_t = 1$).

  In this case, $\mathbf{\Phi}_t$ is a row vector and so $\text{EIG}_t = \frac{1}{2}\log\left|\frac{1}{v_w}\mathbf{\Phi}_t\mathbf{\Sigma}_u\mathbf{\Phi}_t^\mathsf{T} + \boldsymbol{I}\right|$ is maximized by the **dominant eigenvector** of $\mathbf{\Sigma}_u$.

- In practice, though, we may want $M_t \gg 1$ measurements per step.

  For this case, we show that the EIG is maximized by **waterfilling**:

  **Lemma 1** *Say that $(\lambda_m, \boldsymbol{v}_m)_{m=1}^{M_t}$ are the $M_t$ dominant (eigenvalue, eigenvector) pairs of $\mathbf{\Sigma}_u$. Then for $\{E_m\}_{m-1}^{M_t}$ and "water level" $L$ satisfying*

  $$E_m = \max\left\{L - v_w/\lambda_m, 0\right\} \quad \forall m \in \{1,\dots,M_t\}$$

  $$\textstyle\sum_{m=1}^{M_t} E_m = \mathcal{E},$$

  *the $m^{th}$ row of the EIG-maximizing $\mathbf{\Phi}_t$ equals $\sqrt{E_m}\boldsymbol{v}_m$.*

## Leveraging Gaussian design for Adaptive CS:

- In CS, the step-$t$ prior (i.e., step-$(t-1)$ posterior) $p(\boldsymbol{u}|\underline{\boldsymbol{y}}_{t-1})$ is non-Gaussian, and so a **Gaussian posterior approximation** must be made.
- Previous works have tackled this using a **Gaussian prior approximation**:
  - Say $p(\boldsymbol{x}\,|\,\underline{\boldsymbol{y}}_{t-2}) \approx \prod_{n=1}^{N} \mathcal{N}(x_n; 0, \alpha_n^{-1})$ with "precision" $\alpha_n$.
  - Then $p(\boldsymbol{x}\,|\,\underline{\boldsymbol{y}}_{t-1}) \approx \mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}_x, \boldsymbol{\Sigma}_x)$ with

$$\begin{aligned}\boldsymbol{\Sigma}_x &\triangleq \left(\tfrac{1}{v_w}\underline{\boldsymbol{A}}_{t-1}^{\mathsf{T}}\underline{\boldsymbol{A}}_{t-1} + \mathcal{D}(\boldsymbol{\alpha})\right)^{-1}\\ \boldsymbol{\mu}_x &\triangleq \tfrac{1}{v_w}\boldsymbol{\Sigma}_x\underline{\boldsymbol{A}}_{t-1}^{\mathsf{T}}\underline{\boldsymbol{y}}_{t-1}\\ \underline{\boldsymbol{A}}_{t-1} &\triangleq \underline{\boldsymbol{\Phi}}_{t-1}\boldsymbol{\Psi}\end{aligned}$$

    and so $p(\boldsymbol{u}\,|\,\underline{\boldsymbol{y}}_{t-1}) \approx \mathcal{N}(\boldsymbol{u}; \boldsymbol{\mu}_u, \boldsymbol{\Sigma}_u)$ with $\boldsymbol{\mu}_u = \boldsymbol{\Psi}\boldsymbol{\mu}_x$ and $\boldsymbol{\Sigma}_u = \boldsymbol{\Psi}\boldsymbol{\Sigma}_x\boldsymbol{\Psi}^{\mathsf{T}}$.
  - To estimate $\boldsymbol{\alpha}$, [Ji/Xu/Carin 08] used Tipping's RVM ("Bayesian CS").
- Other works used different Gaussian posterior approximations:
  - [Seeger 08] assumed Laplacian $\boldsymbol{x}$ and expectation propagation, and
  - [Seeger/Nickisch 11] used variational methods.

## Approximate Message Passing:

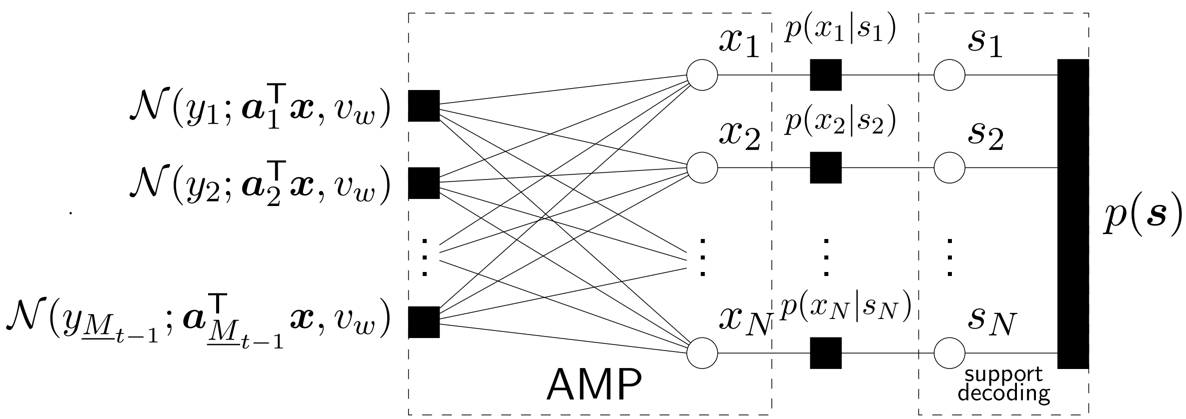
- Efficient sparse reconstruction algorithms have been constructed using loopy belief propagation with carefully constructed message approximations:
  - The **LASSO AMP** [Donoho/Maleki/Montanari 09] assumes i.i.d Laplacian signal, Gaussian noise, and i.i.d constructed $\boldsymbol{A}$.
  - The **Bayesian AMP** [Donoho/Maleki/Montanari 10] accepts generic signal priors, Gaussian noise, and i.i.d constructed $\boldsymbol{A}$.
  - The **generalized AMP** [Rangan 10] accepts generic signal and noise priors and arbitrary $\boldsymbol{A}$. *(We need this one!)*
- These AMP algorithms are **very fast iterative thresholding** algorithms. Their complexity is dominated by one application of $\boldsymbol{A}$ and $\boldsymbol{A}^\mathsf{T}$ per iteration, and $\lesssim 50$ iterations (for any $M$ and $N$) . . . many fewer than FISTA.

## Turbo-AMP for Structured Sparsity:

- AMP has been extended to generic **structured-sparse** reconstruction using an approach inspired by **turbo** equalization and decoding.
- For this, the prior pdf is chosen as $p(\boldsymbol{x}) = p(\boldsymbol{s}) \prod_{n=1}^{N} p(x_n \,|\, s_n)$ with a generic support prior $p(\boldsymbol{s})$ and Bernoulli-Gaussian amplitudes:

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

In this case, the factor graph becomes




and we pass **extrinsic likelihoods** on $\{s_n\}$ back and forth between the two soft-input/soft-output "decoders" [Schniter 10].

## Turbo-AMP for Adaptive CS:

- To leverage Gaussian experiment design, we propose a variation on the **Gaussian prior approximation** used in [Ji/Xu/Carin 08]:

$$p(\boldsymbol{x}\,|\,\underline{\boldsymbol{y}}_{t-2}) \approx \prod_{n=1}^{N} \mathcal{N}(x_n; 0, \alpha_n^{-1})$$

- Instead of using the RVM to ML-estimate $\{\alpha_n\}$, we we use **AMP's marginal posteriors**

$$p(x_n\,|\,\underline{\boldsymbol{y}}_{t-1}) \approx \mathcal{N}(x_n; \hat{x}_n, \nu_n) \quad \text{and} \quad \Pr\{s_n\!=\!1\,|\,\underline{\boldsymbol{y}}_{t-1}\} \approx \lambda_n.$$

In particular, we propose several **surrogates** for the inverse precisions $\alpha_n^{-1}$:

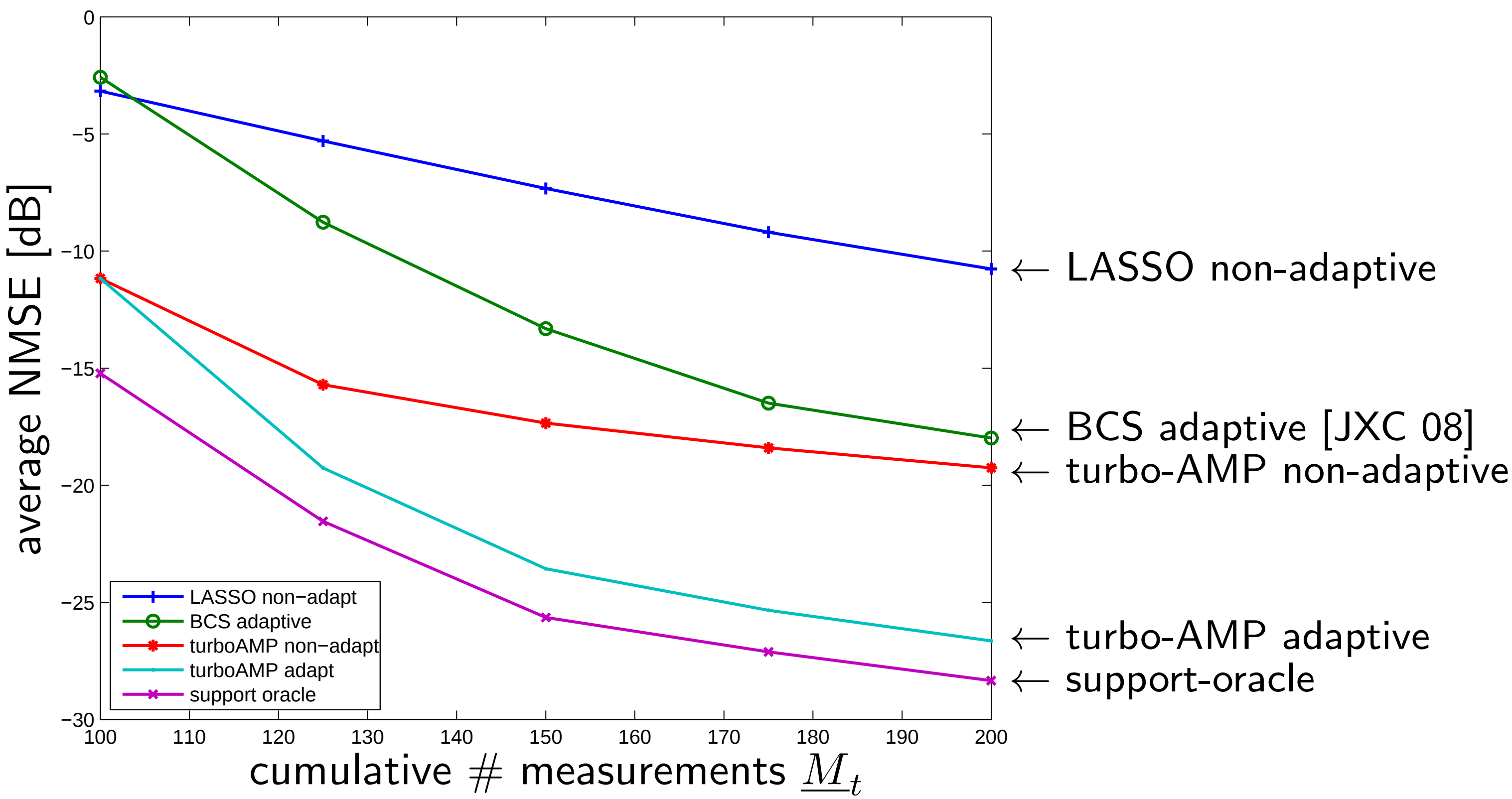
1. "Variance": $\hat{\alpha}_n^{-1} = \nu_n$.
2. "Mean": $\hat{\alpha}_n^{-1} = |\hat{x}_n|^2$ ... only point estimates ($\rightsquigarrow$ **adaptive Lasso**!)
3. "Energy": $\hat{\alpha}_n^{-1} = |\hat{x}_n|^2 + \nu_n$
4. "Support": $\hat{\alpha}_n^{-1} = \lambda_n v_x$ ,

## Empirical Study:

We now present empirical evidence showing that the proposed **adaptive turbo-AMP** performs very close to **oracle bounds**.

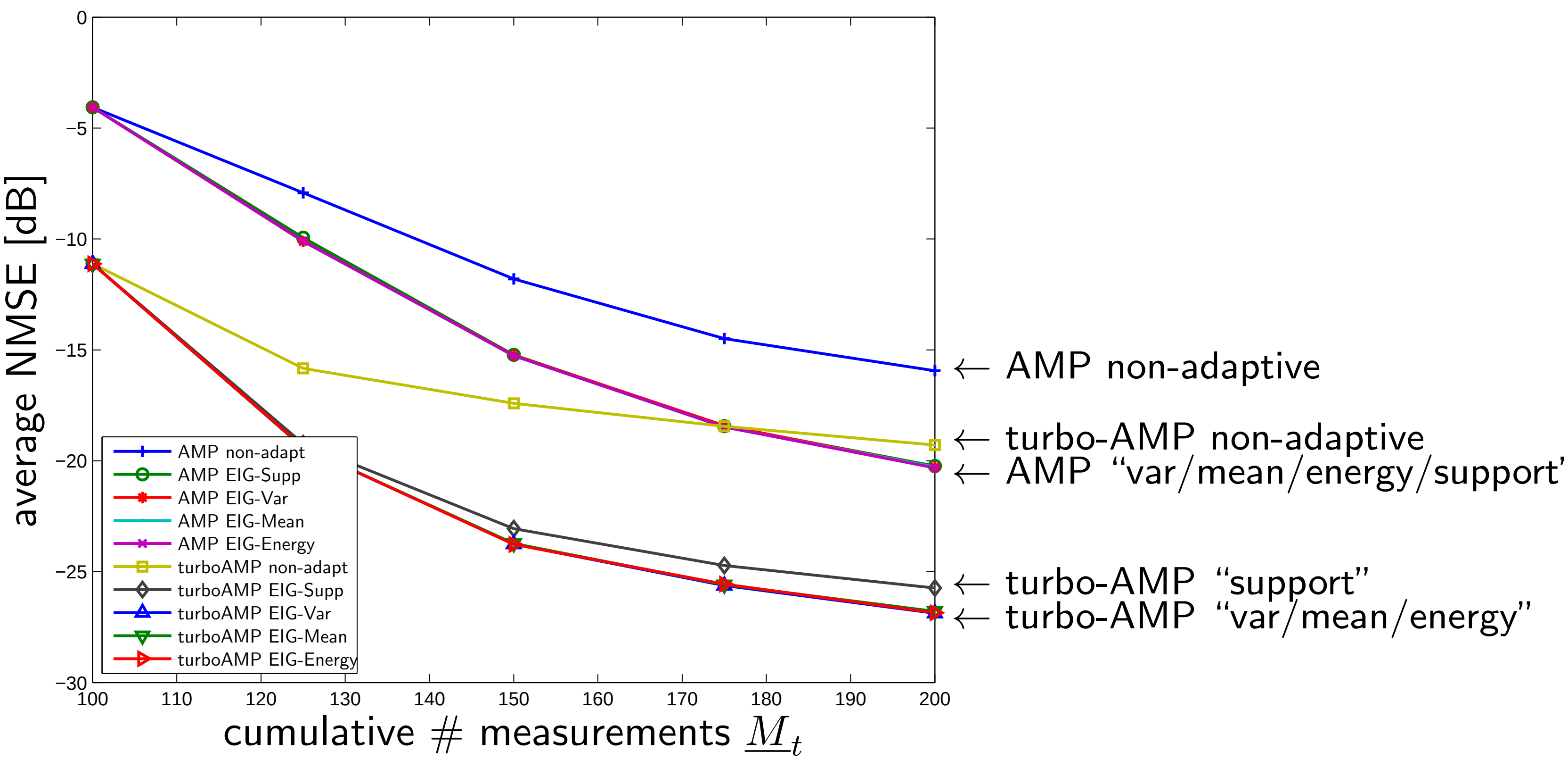
- Clustered-sparse Bernoulli-Gaussian signal:
  - length $N = 500$,
  - sparsity $K = 50$,
  - average cluster-size $= 11$.
- Canonical sparsifying dictionary $\boldsymbol{\Psi} = \boldsymbol{I}$ (i.e., $\boldsymbol{u} = \boldsymbol{x}$).
- AWGN yielding average SNR $= 15$dB.
- $T = 5$ measurement steps, with $M_0 = 100$ i.i.d-$\mathcal{N}$, then subsequently $M_t = 50$.
- We report NMSE $\|\hat{\boldsymbol{x}} - \boldsymbol{x}\|_2^2 / \|\boldsymbol{x}\|_2^2$ averaged over $500$ realizations.
- We compare to the **support oracle**, for which signal is Gaussian, and so both EIG-maximizing $\boldsymbol{\Phi}_t$ and MSE-minimizing $\hat{\boldsymbol{x}}$ can be computed in closed form.

## NMSE versus cumulative measurements $\underline{M}_t$:



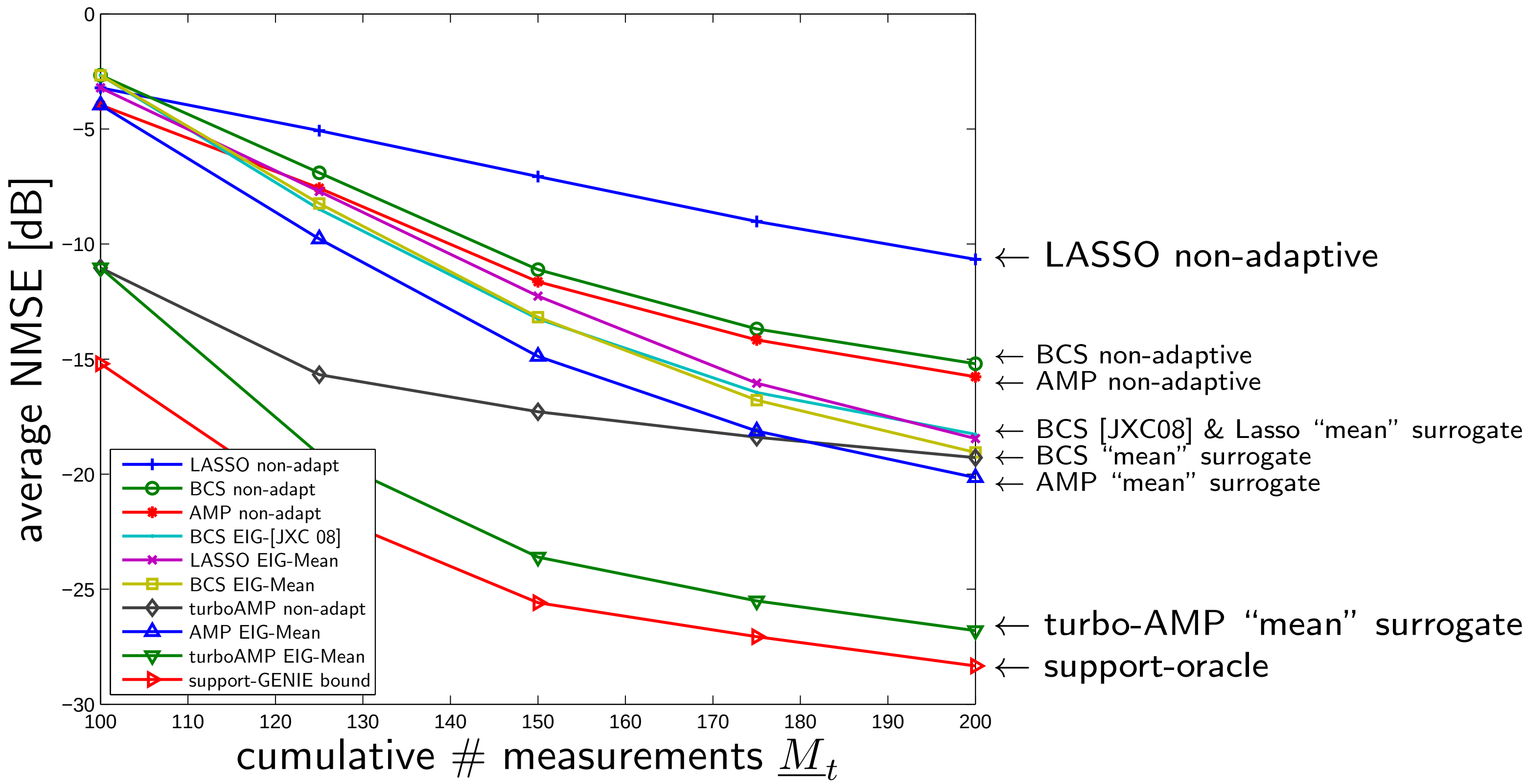

- Performances gain from structured sparsity, adaptivity, and the combination.
- **Adaptive turbo-AMP** performs **1.5 dB from the support-oracle bound**!

# Effect of surrogate choice in Gaussian prior approximation:




Relatively **insensitive** to the Gaussian-prior-approximation used in $\mathbf{\Phi}_t$ design.

# Using the “mean” surrogate to create new algorithms:




- Adaptation using our “mean” surrogate yields an **adaptive LASSO**.
- Adaptation using our “mean” surrogate **improves** BCS over [JXC 08].

## Summary and ongoing work:

- Main focus:

  Merging **Bayesian experim. design** with **structured-sparse** recovery.

- Contributions:
  - **Waterfilling** solves Gaussian experimental design for $M_t > 1$ meas/step.
  - Novel adaptation heuristics leading to **adaptive LASSO**, etc.
  - An **adaptive turbo-AMP** empirically performing near **oracle bounds**.

- Ongoing work:
  - Optimal design of **initial** $\boldsymbol{\Phi}_0$.
  - Theoretical analysis using AMP's **state evolution**.
  - Extension to **pre**-measurement noise model $\boldsymbol{y} = \boldsymbol{\Phi}(\boldsymbol{\Psi}\boldsymbol{x} + \boldsymbol{v}) + \boldsymbol{w}$.
  - Adaptation under **constrained** $\boldsymbol{\Phi}$ (e.g., Toeplitz).
  - Development/analysis of **simplified** schemes (no eigendecomposition).

*Thanks!*