Compressive Phase Retrieval via Bethe Free Energy Minimization

Phil Schniter

The Ohio State University

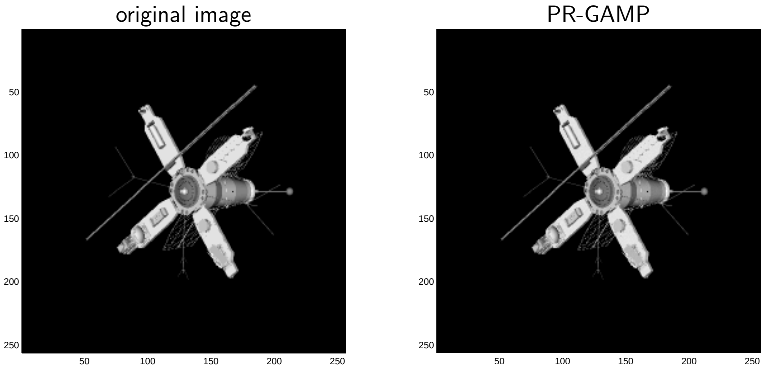
Collaborators: Sundeep Rangan (NYU)

Supported in part by NSF grants CCF-1018368 and CCF-1218754.

AMS Sectional Meeting (East Lansing, MI) — 3.14.15

Compressive Phase Retrieval. . . An Example

65536 image pixels, 32768 measurements, 30dB SNR:



NMSE = -37.5 dB, runtime = 1.8 sec.

Image Recovery

- In image recovery, we want to
  - recover a image $\boldsymbol{x} \in \mathbb{C}^N$
  - from corrupted measurements $\boldsymbol{y} \in \mathbb{C}^M$
  - of hidden linear transform outputs $\boldsymbol{z} = \boldsymbol{A}\boldsymbol{x} \in \mathbb{C}^M$.
- The measurement corruption mechanism might be
  - additive noise: $y_i = z_i + w_i$
  - phase-less: $y_i = |z_i + w_i|$
  - one-bit: $y_i = \operatorname{sgn}(z_i + w_i)$
  - photon-limited (Poisson), etc...
- The image is structured in that $\boldsymbol{\Omega}\boldsymbol{x} \in \mathbb{C}^D$ is . . .
  - sparse (sufficiently few nonzeros)
  - co-sparse (sufficiently many zeros).

  In this talk, we discuss only the case $\boldsymbol{\Omega} = \boldsymbol{I}$ for simplicity.

Statistical Approach to Image Recovery

In the statistical approach to image recovery...

- measurements modeled via likelihood $p(\boldsymbol{y}|\boldsymbol{x}) = \prod_{i=1}^{M} p_{\mathsf{y|z}}(y_i|[\boldsymbol{A}\boldsymbol{x}]_i)$
- image modeled via prior distribution $p(\boldsymbol{x}) = \prod_{j=1}^{N} p_{\mathsf{x}}(x_j)$
- The posterior

$$p(\boldsymbol{x}|\boldsymbol{y}) = \frac{p(\boldsymbol{y}|\boldsymbol{x})p(\boldsymbol{x})}{\int_{\mathbb{C}^N} p(\boldsymbol{y}|\boldsymbol{x}')p(\boldsymbol{x}')\, d\boldsymbol{x}'},$$

  tells *all* we can learn about $\boldsymbol{x}$ from $\boldsymbol{y}$, but is expensive to compute.

- Instead, one usually settles for point estimates like the
  - MAP estimate: $\hat{\boldsymbol{x}}_{\mathsf{MAP}} = \arg\max_{\boldsymbol{x}} p(\boldsymbol{x}|\boldsymbol{y})$
  - MMSE estimate: $\hat{x}_{j,\mathsf{MMSE}} = \mathrm{E}\{x_j|\boldsymbol{y}\} = \int_{\mathbb{C}} x_j\, p(x_j|\boldsymbol{y}) d\boldsymbol{x} \quad \forall j$

  and perhaps marginal uncertainty information like $\mathrm{var}\{x_j|\boldsymbol{y}\}$.

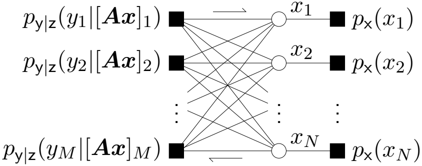
Loopy Belief Propagation: Computing Posterior Marginals

- Factor the posterior, exposing the statistical structure of the problem:

$$p(\boldsymbol{x}|\boldsymbol{y}) = \prod_{\alpha=1}^{N+M} f_\alpha(\boldsymbol{x}_\alpha) \;\propto \prod_{i=1}^{M} p_{\mathsf{y|z}}(y_i|[\boldsymbol{A}\boldsymbol{x}]_i) \prod_{j=1}^{N} p_{\mathsf{x}}(x_j),$$

Visualize using the factor graph:

(White circles are random variables and black boxes are factors.)



- Inference: Pass messages (pdfs) between nodes until they agree. The sum-product algorithm approximates the marginal posteriors $p(x_j|\boldsymbol{y})$ by locally minimizing the Bethe free energy:

$$J(\{q_\alpha\},\{q_\beta\}) = \sum_{\alpha=1}^{N+M} D_{\mathsf{KL}}(q_\alpha \| f_\alpha) + M \sum_{\beta=1}^{N} h(q_\beta)$$

q_α, q_β : cluster marginals s.t. $q_\alpha(x_\beta) = \int q_\alpha(\boldsymbol{x}_\alpha)\, d\boldsymbol{x}_{\alpha\backslash\beta} = q_\beta(x_\beta)\ \forall \alpha, \beta \in \mathfrak{N}_\alpha$

The Blessings of Dimensionality

For general prior/likelihood and $\boldsymbol{A}$, loopy BP is not tractable.

But if $\boldsymbol{A}$ is i.i.d. sub-Gaussian then in the large-system limit . . .

- messages can be approximated as Gaussian pdfs due to CLT,
- differences between messages approximated via Taylor's expansion,[1]
  $\rightarrow$ Approximate Message Passing (AMP) algorithm
- per-iteration behavior characterized by a scalar state-evolution (SE),
- if SE has unique fixed point, the marginal-pdf estimates are exact.[2]

[1] Donoho,Maleki,Montanari–PNAS'09

[2] Bayati,Montanari–IT'11

The Generalized[3] AMP Algorithm

for $t = 1, 2, 3, \ldots$

$1/\sigma_t = \nu_t^x \|\boldsymbol{A}\|_F^2 / M$	stepsize adaptation
$\tilde{\boldsymbol{s}}_{t+1} = G(\boldsymbol{s}_t + \sigma_t \boldsymbol{A}\boldsymbol{x}_n, \sigma_t)$	scalar denoising
$\nu_{t+1}^s = \text{avg}\{\sigma_t G'(\boldsymbol{s}_t + \sigma_t \boldsymbol{A}\boldsymbol{x}_n, \sigma_t)\}$	local sensitivity
$1/\tau_t = \nu_{t+1}^s \|\boldsymbol{A}\|_F^2 / N$	stepsize adaptation
$\tilde{\boldsymbol{x}}_{t+1} = F\big(\boldsymbol{x}_t - \tau_t \boldsymbol{A}^{\mathsf{H}} \tilde{\boldsymbol{s}}_{t+1}, \tau_t\big)$	scalar denoising
$\nu_{t+1}^x = \text{avg}\big\{\tau_t F'\big(\boldsymbol{x}_t - \tau_t \boldsymbol{A}^{\mathsf{H}} \hat{\boldsymbol{s}}_{t+1}, \tau_t\big)\big\}$	local sensitivity
$\begin{bmatrix} \boldsymbol{x}_{t+1} \\ \boldsymbol{s}_{t+1} \end{bmatrix} = \beta_t \begin{bmatrix} \tilde{\boldsymbol{x}}_{t+1} \\ \tilde{\boldsymbol{s}}_{t+1} \end{bmatrix} + (1 - \beta_t) \begin{bmatrix} \boldsymbol{x}_t \\ \boldsymbol{s}_t \end{bmatrix}$	damping, $\beta_t \in (0, 1]$

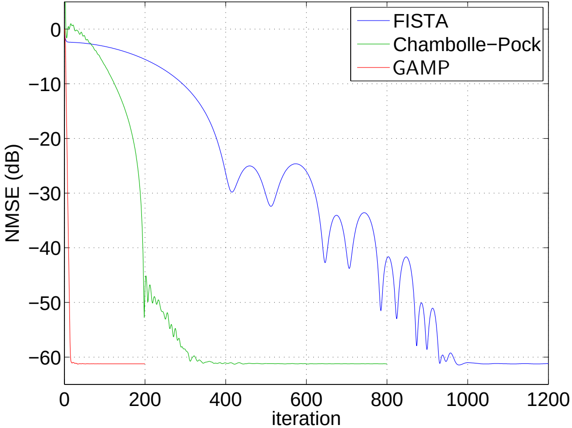
Looks just like a "primal-dual" algorithm, but ...

- prox operators are replaced by MMSE denoisers,
- step-sizes σ_t and τ_t are adapted so that...
- denoiser input is an AWGN-corrupted true $\boldsymbol{x}$ with error variance τ_t.

[3]Rangan—arXiv:1010:5141

How fast is (G)AMP?

Pretty fast, at least for i.i.d. zero-mean Gaussian $\boldsymbol{A}$:



Above: LASSO recovery of a 40-sparse 1000-length Bernoulli-Gaussian signal from 400 AWGN-corrupted measurements.

What about generic matrices $\boldsymbol{A}$?

Here is what we know about sum-product GAMP:

- It may diverge! But...
- Gaussian case: convergence is determined by the peak-to-average ratio of the squared singular-values in $\boldsymbol{A}$. For any $\boldsymbol{A}$, possible to find fixed damping coefficient $\beta_t = \beta$ that guarantees global convergence.[4]
- General case: if it converges, then it converges to a local minimum of the large-system-limit Bethe free energy (LSL-BFE):[5][6]

$$J(b_x, b_z) = D_{\mathsf{KL}}(b_x \| p_{\mathsf{x}}) + D_{\mathsf{KL}}(b_z \| p_{\mathsf{y|z}}) + \bar{h}\big(\operatorname{var}(\boldsymbol{x}|b_x), \operatorname{var}(\boldsymbol{z}|b_z)\big)$$

$$b_x, b_z : \text{separable posteriors pdfs s.t. } \mathrm{E}\{\boldsymbol{A}\boldsymbol{x}|b_x\} = \mathrm{E}\{\boldsymbol{z}|b_z\}$$

LSL-BFE-based damping works empirically, but not provably.

[4] Rangan,Schniter,Fletcher–arXiv:1402.3210

[5] Rangan,Schniter,Riegler,Fletcher,Cevher–arXiv:1301.6295

[6] Krzakala,Manoel,Tramel,Zdeborova–arXiv:1402.1384

ADMM-GAMP: A Provably Convergent Alternative

- Main idea: direct minimization of LSL-BFE:

$$\underset{\text{separable pdfs } b_x, b_z}{\arg\min} \; D_{\mathsf{KL}}(b_x \| p_{\mathsf{x}}) + D_{\mathsf{KL}}(b_z \| p_{\mathsf{y|z}}) + \bar{h}\big(\operatorname{var}(\boldsymbol{x}|b_x), \operatorname{var}(\boldsymbol{z}|b_z)\big) \quad \text{s.t. } \operatorname{E}\{\boldsymbol{A}\boldsymbol{x}|b_x\} = \operatorname{E}\{\boldsymbol{z}|b_z\}$$

- Challenge: $\bar{h}(\operatorname{var}(b))$ is neither convex nor concave in $b \triangleq (b_x, b_z)$.
- Solution: a double loop algorithm:[7]
  - Outer loop: linearize $\bar{h}$ about current guess $\rightsquigarrow$ convex + concave

$$D_{\mathsf{KL}}(b_x \| p_{\mathsf{x}}) + D_{\mathsf{KL}}(b_z \| p_{\mathsf{y|z}}) + \frac{1}{2\boldsymbol{\tau}}^{\mathsf{T}} \operatorname{var}(\boldsymbol{x}|b_x) + \frac{\boldsymbol{\sigma}}{2}^{\mathsf{T}} \operatorname{var}(\boldsymbol{z}|b_z).$$

  - Inner loop: Minimize linearized LSL-BFE using ADMM under constraints $E(\boldsymbol{x}|b_x) = \boldsymbol{v}$, $\operatorname{E}(\boldsymbol{z}|b_z) = \boldsymbol{A}\boldsymbol{v}$ using penalty vectors $\frac{1}{2\boldsymbol{\tau}}$ and $\frac{\boldsymbol{\sigma}}{2}$, respectively.
  - Result is basically GAMP plus one additional LS step for $\boldsymbol{v}$.
- Global linear convergence proven for strongly concave $\log p_{\mathsf{x}}$ & $\log p_{\mathsf{y|z}}$.

[7] Rangan, Fletcher, Schniter, Kamilov–arXiv:1501.01797

Tuning the Hyperparameters

- The prior p_{x} often has tunable parameters (e.g., sparsity). How to choose them?
  - The input to GAMP's denoiser is an AWGN corrupted version of the truth with known error variance. Thus,
    1. learn prior via EM[8] (deconvolution of blurred pdf), or
    2. apply Stein's Unbiased Risk Estimator.[9]
  - Can "learn prior" by tuning a high-order Gaussian-mixture model p_{x}.
- The likelihood $p_{\mathsf{y|z}}$ also has tunable parameters (e.g., noise variance). How to choose them?
  - Use the LSL-BFE as a negative-log-likelihood upper-bound. The AWGN case admits simple closed-form tuning.[10] For the non-AWGN case, we proposed a Newton-based algorithm.[11]

[8] Vila,Schniter–SAHD'11 & TSP'13

[9] Mousavi,Maleki,Baraniuk–arXiv:1311.0035 / Guo,Davies–arXiv:1409.0440

[10] Krzakala,Mezard,Sausset,Sun,Zdeborova–JSM'12

[11] Schniter,Rangan–arXiv:1405.5618

Application to Phase Retrieval

Need a likelihood function $p_{\mathsf{y|z}}(y_i|z_i)$ relating the noisy intensity measurements y_i to the noiseless transform outputs $z_i = [\boldsymbol{Ax}]_i$.

1. Pre-intensity additive noise: $y_i = |z_i + w_i|$.

   If $w_i \sim \mathcal{CN}(0, \nu^w)$, then likelihood is Rician:

   $$p_{\mathsf{y|z}}(y_m|z_m;\nu^w) = \frac{2y_m}{\nu^w}\exp\Big(-\frac{y_m^2+|z_m|^2}{\nu^w}\Big)I_0\Big(\frac{2y_m|z_m|}{\nu^w}\Big)1_{y_m\geq 0},$$

   where $I_0(\cdot)$ is the 0^{th}-order modified Bessel function of the first kind. LSL-BFE-based tuning of ν^w is detailed in paper.[12]

2. Post-intensity additive noise: $y_i = q(|z_i|) + w_i$ for some $q(\cdot)$.

   Can handle this for generic $q(\cdot)$ and p_{w}. See details in paper.[12]

3. Non-additive noise: e.g., Poisson model.

   Can handle this as well since we allow generic $p_{\mathsf{y|z}}(y_m|z_m)$.

[12] Schniter, Rangan–arXiv:1405.5618

Synthetic Experiments

For these numerical results we generated random. . .

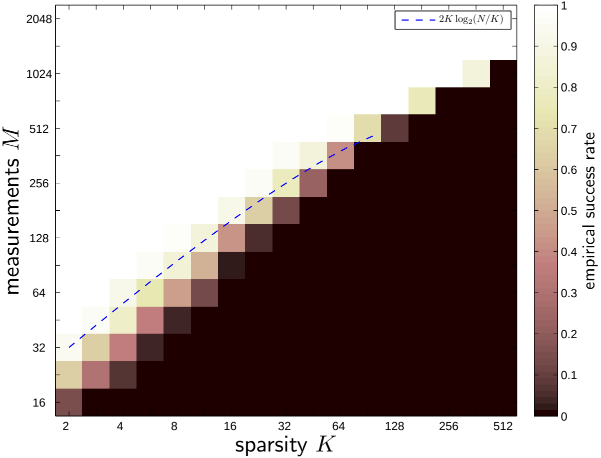
- signals $\boldsymbol{x}_0$ as K-sparse, $N = 512$-length, Bernoulli-circular-Gaussian,
- measurement matrices $\boldsymbol{A}$ as i.i.d circular Gaussian,
- pre-intensity additive noise $\boldsymbol{w}$ as circular white Gaussian,

and we monitored the phase-corrected normalized MSE

$$\text{NMSE} \triangleq \min_{\theta} \frac{\|\hat{\boldsymbol{x}} - e^{\mathrm{i}\theta}\boldsymbol{x}_0\|_2^2}{\|\boldsymbol{x}_0\|_2^2}.$$

Empirical Success Rate

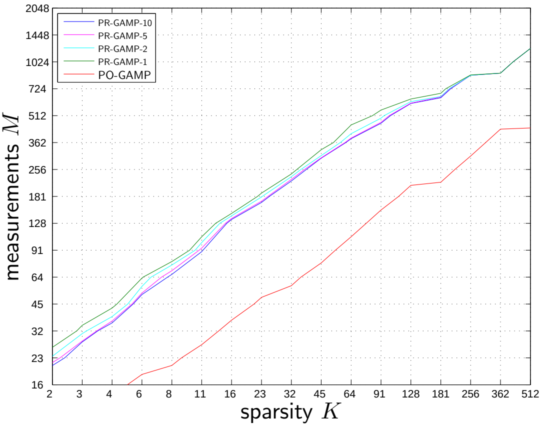
Empirical rate of success $\triangleq \{\text{NMSE} < 10^{-6}\}$, averaged over 100 realizations at SNR $= 100$ dB:



- Note "non-compressive" phase retrieval means $M \gtrsim 4N = 2048$.
- Dashed curve shows $M = 2K\log_2(N/K)$ for reference.

Phase-retrieval GAMP vs. Phase-oracle GAMP

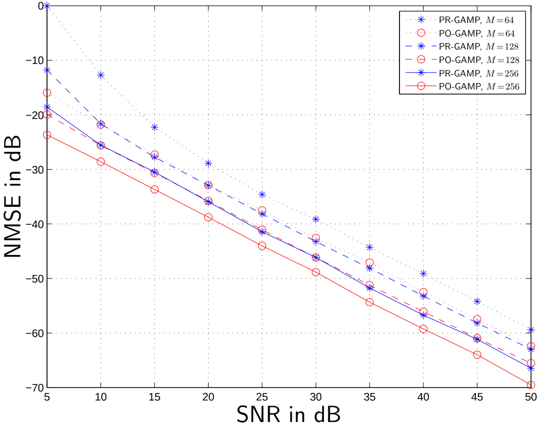
50%-success contours averaged over 100 realizations at SNR $= 100$ dB:



- Phase-retrieval GAMP requires $\approx 4\times$ the number of measurements as phase-oracle GAMP. (Very interesting!)
- Randomly restarting PR-GAMP doesn't help much (for this family of $\boldsymbol{A}$).

Robustness to Noise

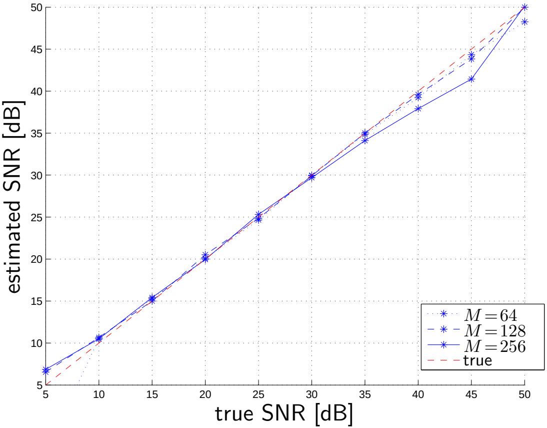
The median NMSE for sparsity $K = 4$ over 200 realizations:



- PR-GAMP loses ≈ 3 dB to PO-GAMP at medium-to-high SNR.
- $(K, M) = (4, 64)$ is near the boundary of the phase transition.

Accuracy of Noise-Variance Learning

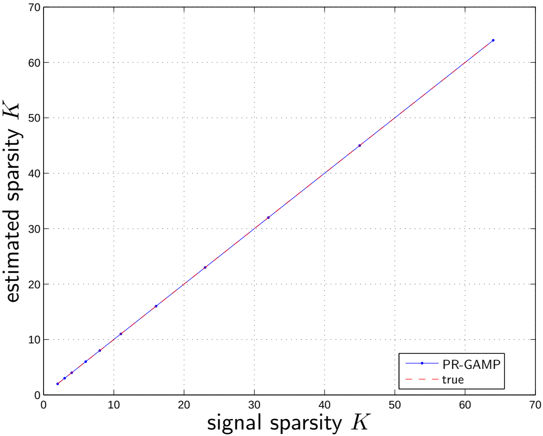
The average estimated noise variance for sparsity $K = 4$ at several M over 10 realizations:



- The LSL-BFE-based likelihood-tuning method is accurate across a wide SNR range.

Accuracy of Sparsity-Rate Learning

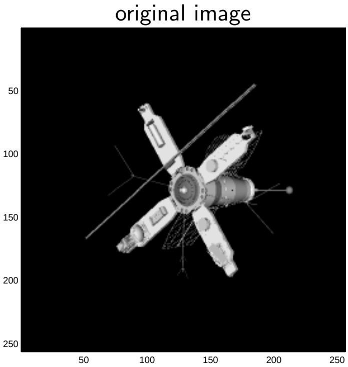
The average estimated sparsity for $M = 512$ over 10 realizations:

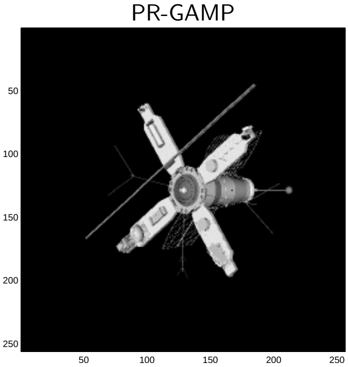


- The EM-based prior-tuning method is accurate across a wide sparsity range.

Compressive Image Recovery

65536 image pixels, 32768 measurements, 30dB SNR:





NMSE = -37.5 dB, runtime = 1.8 sec.

Compressive Image Recovery: Details

- Measurements operators used blurring and masking:

$$A = \begin{bmatrix} B_1 & \\ & B_2 \end{bmatrix} \begin{bmatrix} F & \\ & F \end{bmatrix} \begin{bmatrix} D_1 \\ D_2 \end{bmatrix}$$

  - B_i: banded blur operators, 10 i.i.d-Gaussian entries per column
  - F: 2D FFT
  - D_i: masks with binary $\{0, 1\}$ diagonal entries

- Over 100 random measurement & noise realizations at SNR$=30$dB:
  - NMSE < -36 dB in 99 trials,
  - median runtime $= 3.3$ sec.

PR-GAMP: Ongoing Work

PR-GAMP is a work-in-progress. Things we are working on include:

- Derivation of the state evolution.
- Incorporation of analysis-form priors (i.e., $\boldsymbol{\Omega} \neq \boldsymbol{I}$).[13]
- Incorporation of non-additive (e.g., Poisson) corruption models.[14]
- MAP formulation of PR-GAMP.

[13] Borgerding,Schniter—arXiv:1312.3968

[14] Fletcher,Rangan,Varshney,Bhargava—NIPS'11

Conclusions

- (Compressive) phase retrieval is a longstanding problem that is experiencing a rebirth through compressive sensing and convex relaxation.
- We proposed a new approach to CPR based on generalized approximate message passing (GAMP), which minimizes the large-system limit Bethe free energy.
- Our approach can automatically learn the noise variance and signal sparsity.
- Empirical results show an excellent phase transition ($4\times$measurements of phase-oracle), excellent noise robustness (~ 3 dB worse than phase-oracle), and very fast runtimes.
- As a practical demonstration, we accurately recovered a 64k-pixel image from 32k noisy measurements in only 1.8 seconds.

All of these methods are integrated into GAMPmatlab:
`http://sourceforge.net/projects/gampmatlab/`

Thanks!