Upscaling Vector Approximate Message Passing

International Conference on Acoustics, Speech, and Signal Processing

May 4, 2020

Nikolajs Skuratovs, Michael Davies
The University of Edinburgh

This work was supported by the ERC project C-SENSE (ERC-ADG-2015-694888)

The model

Consider the recovery of a random signal **x** from a set of linear measurements

$$\mathbf{y} = \mathbf{A}\mathbf{x} + \mathbf{w}$$

Where

- $\mathbf{x} \in \mathbb{R}^N$
- $\mathbf{y} \in \mathbb{R}^M$
- $\mathbf{w} \sim N(0, v_w \mathbf{I}_M)$
- $\mathbf{A} \in \mathbb{R}^{M \times N}$

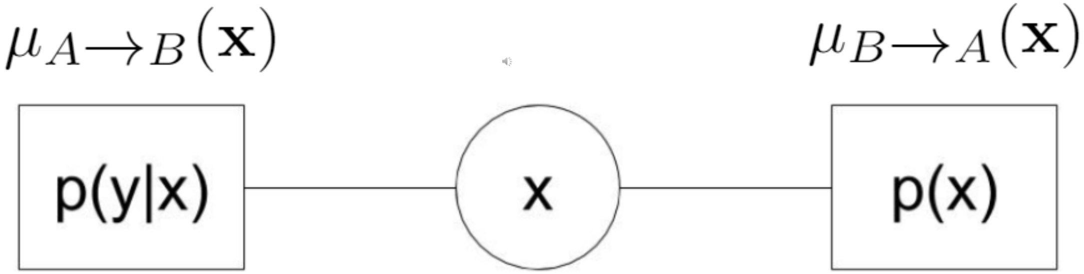
And we consider the compressed sensing scenario $M << N$ with both of a similar order

Inference via Bayes-motivated approach: EP

Assume we can form the posterior for **x** given measurements **y**

$$p(\mathbf{x}|\mathbf{y}) \propto p(\mathbf{y}|\mathbf{A}, \mathbf{x})p(\mathbf{x})$$

which can be represented with a factor graph (FG)



On this FG, we employ EP with isotropic Gaussian approximations:

- p(y|x) is approximated by $\mu_{A\rightarrow B}(\mathbf{x}) = N(\mathbf{x}; \mathbf{x}_{A\rightarrow B}, v_{A\rightarrow B}\mathbf{I}_N)$
- p(x) is approximated by $\mu_{B\rightarrow A}(\mathbf{x}) = N(\mathbf{x}; \mathbf{x}_{B\rightarrow A}, v_{B\rightarrow A}\mathbf{I}_N)$

Expectation Propagation updates

On that simple factor graph, the EP update rules are

$$\mu_{A\to B}(\mathbf{x}) = \frac{\mu_A(\mathbf{x})}{\mu_{B\to A}(\mathbf{x})} = \frac{proj\left[\mu_{B\to A}(\mathbf{x})p(\mathbf{y}|\mathbf{x})\right]}{\mu_{B\to A}(\mathbf{x})}$$

$$\mu_{B\to A}(\mathbf{x}) = \frac{\mu_A(\mathbf{x})}{\mu_{A\to B}(\mathbf{x})} = \frac{proj\left[\mu_{A\to B}(\mathbf{x})p(\mathbf{x})\right]}{\mu_{A\to B}(\mathbf{x})}$$

Where the proj[ ] operator is the KL projection on the family of Gaussian distributions with isotropic covariance matrices

Note that the updates are carried out only in terms of moments: **the mean and the variance**.

$$\mu_{A\to B}(\mathbf{x}) = N(\mathbf{x}; \mathbf{x}_{A\to B}, v_{A\to B}\mathbf{I}_N)$$

$$\mu_{B\to A}(\mathbf{x}) = N(\mathbf{x}; \mathbf{x}_{B\to A}, v_{B\to A}\mathbf{I}_N)$$

EP-based algorithm

If one derives these update rules, one can get the following algorithm

Initialization: $\mathbf{x}^0_{B\to A} = 0, \tilde{v}^0_{B\to A} = \tilde{v}_x, t = 0$

while $t < T_{max}$ *and* $\tilde{v}^t_{B\to A} \geq \epsilon$ **do**

- **Block A**
  - $\boldsymbol{\mu}^t_A = \mathbf{g}_A(\mathbf{x}^t_{B\to A}, \tilde{v}_{B\to A}, \tilde{v}_w)$
  - $\frac{1}{\gamma^t_A} = \frac{1}{N}\nabla_{(\mathbf{x}^t_{B\to A})} \cdot \left(\mathbf{A}^T \mathbf{g}_A(\mathbf{x}^t_{B\to A}, \tilde{v}^t_{B\to A}, \tilde{v}_w)\right)$
  - $\mathbf{x}^t_{A\to B} = \mathbf{x}^t_{B\to A} - \gamma^t_A \mathbf{A}^T \boldsymbol{\mu}^t_A$
  - $\tilde{v}^t_{A\to B} = f_A(\boldsymbol{\mu}^t_A, \mathbf{x}^t_{B\to A}, \tilde{v}_{B\to A}, \tilde{v}_w)$

  } computation of $\mu_{A\to B}(\mathbf{x})$

  $N(\mathbf{x}; \mathbf{x}_{A\to B}, v_{A\to B}\mathbf{I}_N)$

- **Block B**
  - $\boldsymbol{\mu}^{t+1}_B = \mathbf{g}_B(\mathbf{x}^t_{A\to B}, \tilde{v}^t_{A\to B})$
  - $\gamma^{t+1}_B = \frac{1}{N}\nabla_{\mathbf{x}^t_{A\to B}} \cdot \mathbf{g}_B(\mathbf{x}^t_{A\to B}, \tilde{v}^t_{A\to B})$
  - $\mathbf{x}^{t+1}_{B\to A} = \frac{1}{1-\gamma^t_B}\left(\boldsymbol{\mu}^{t+1}_B - \gamma^{t+1}_B \mathbf{x}^t_{A\to B}\right)$
  - $\tilde{v}^{t+1}_{B\to A} = f_B(\boldsymbol{\mu}^{t+1}_B, \mathbf{x}^t_{B\to A}, \tilde{v}_{A\to B})$

  } computation of $\mu_{B\to A}(\mathbf{x})$

  $N(\mathbf{x}; \mathbf{x}_{B\to A}, v_{B\to A}\mathbf{I}_N)$

- $t = t + 1$

Output: $\boldsymbol{\mu}^t_B$

Other works

An equivalent form of EP, called Vector Approximate Message Passing (VAMP), was first proposed by Rangan *et al* [1]. Shortly after a similar result was presented by Takeuchi [2].

Both of these works studied the dynamics of EP for the considered problem under the assumption that in the SVD of

$$\mathbf{A} = \mathbf{U}\mathbf{S}\mathbf{V}^T$$

the singular vector matrix **V** is **Haar distributed**, while **U** and **S** can be any.

Implementation of Block B

Block B

$$\boldsymbol{\mu}_B^{t+1} = \mathbf{g}_B(\mathbf{x}_{A\to B}^t, \tilde{v}_{A\to B}^t)$$
$$\gamma_B^{t+1} = \frac{1}{N}\nabla_{\mathbf{x}_{A\to B}^t} \cdot \mathbf{g}_B(\mathbf{x}_{A\to B}^t, \tilde{v}_{A\to B}^t)$$
$$\mathbf{x}_{B\to A}^{t+1} = \frac{1}{1-\gamma_B^t}\left(\boldsymbol{\mu}_B^{t+1} - \gamma_B^{t+1}\mathbf{x}_{A\to B}^t\right)$$
$$\tilde{v}_{B\to A}^{t+1} = f_B(\boldsymbol{\mu}_B^{t+1}, \mathbf{x}_{B\to A}^t, \tilde{v}_{A\to B})$$

Thus g_B acts as a denoiser with measurements $\mathbf{x}_{A\to B}^t$

The scalar γ_B^{t+1} is the divergence of the denoiser

The function f_B produces an estimate of the MSE $\frac{1}{N}||\mathbf{x}_{B\to A}^{t+1} - \mathbf{x}||^2$

These components were well studied in [4], [5], [6], [7]

Properties of Block A

Block A

$$\left|\begin{array}{l} \boldsymbol{\mu}_A^t = \mathbf{g}_A(\mathbf{x}_{B\to A}^t, \tilde{v}_{B\to A}, \tilde{v}_w) \\ \frac{1}{\gamma_A^t} = \frac{1}{N}\nabla_{(\mathbf{x}_{B\to A}^t)} \cdot \left(\mathbf{A}^T\mathbf{g}_A(\mathbf{x}_{B\to A}^t, \tilde{v}_{B\to A}^t, \tilde{v}_w)\right) \\ \mathbf{x}_{A\to B}^t = \mathbf{x}_{B\to A}^t - \gamma_A^t \mathbf{A}^T \boldsymbol{\mu}_A^t \\ \tilde{v}_{A\to B}^t = f_A(\boldsymbol{\mu}_A^t, \mathbf{x}_{B\to A}^t, \tilde{v}_{B\to A}, \tilde{v}_w) \end{array}\right.$$

It was shown that that the mean $\mathbf{x}_{B\to A}^t$ of the approximated density $\mu_{B\to A}(\mathbf{x})$ is equal to

$$\mathbf{x}_{B\to A}^t = \mathbf{x} + \mathbf{q}_t$$

The function g_A is the LMMSE estimator $\mathbf{g}_A(\mathbf{x}_{B\to A}^t) = \mathbf{W}_t^{-1}(\mathbf{y} - \mathbf{A}\mathbf{x}_{B\to A}^t)$

$$\mathbf{W}_t = \tilde{v}_w \mathbf{I}_M + \tilde{v}_{B\to A}^t \mathbf{A}\mathbf{A}^T$$

- Directly compute the inverse – very slow
- Use SVD – requires storing large matrices; intractable amount of memory

The scalar $\frac{1}{\gamma_A^t}$ is the divergence of $\mathbf{A}^T\mathbf{g}_A$

- The same problems as with g_A

The Block A is intractable when the dimensions of the system are large as in many imaging problems. Alternatives?

Conjugate Gradient (CG) approximation

Use **a few iterations** of CG to approximate the LMMSE

$$\mathbf{g}_A(\mathbf{x}^t_{B\to A}) = \mathbf{W}_t^{-1}(\mathbf{y} - \mathbf{A}\mathbf{x}^t_{B\to A}) = \mathbf{z}_t$$

What about the divergence of the resulting $\mathbf{A}^T\mathbf{g}_A$ and the MSE $\tilde{v}^t_{A\to B}$?

Takeuchi and Wen shown [3] that under Haar **V** this divergence can be estimated for *i* iterations of CG if one has access to *2i + 2* moments of the singular spectrum of **S**

What if we don't have the access to those moments?

The divergence of CG

In [3] it was shown that shown that as $N \to \infty$ and with Haar **V**, the CG function becomes a linear mapping

$$\mathbf{g}_A^{i[t]} = \mathbf{U}\mathbf{H}_t^{i[t]}\mathbf{U}^T$$

of the vector $\mathbf{z}_t$ and the diagonal matrix $\mathbf{H}_t^{i[t]}$ is a function of **S**, v_w and $v_{B\to A}^t$ only.

The from the definition of $\gamma_A^{t,i[t]}$ we can show that

$$\frac{1}{\gamma_A^t} = Tr\left\{\mathbf{H}_t^{i[t]}\mathbf{S}\mathbf{S}^T\right\} = \frac{\frac{1}{N}\mathbf{q}_t^T\mathbf{A}^T\mathbf{g}_A^{i[t]}(\mathbf{z}_t)}{\tilde{v}_{B\to A}^t}$$

which is independent of a particular realization of **w** and $\mathbf{q}_t$ but is only a function of its statistics

Estimating the divergence of CG

Since the divergence is independent of a particular realization of **w** and $\mathbf{q}_t$ but is only a function of its statistics, synthesize

$$\dot{\mathbf{z}}_t = \dot{\mathbf{w}} - \mathbf{A}\dot{\mathbf{q}}_t$$

with

$$\dot{\mathbf{w}} \sim \mathbf{N}(\mathbf{0}, \mathbf{v_w I_M})$$
$$\dot{\mathbf{q}}_\mathbf{t} \sim \mathbf{N}(\mathbf{0}, \mathbf{v^t_{B\to A} I_N})$$

Execute CG on the synthesized data. We expect

$$\frac{1}{\dot{\gamma}_A^{t,i[t]}} = \frac{\frac{1}{N}\dot{\mathbf{q}}_t^T \mathbf{A}^T \mathbf{g}_A^{i[t]}(\dot{\mathbf{z}}_t)}{\tilde{v}^t_{B\to A}}$$

to be close to the result with the real data. Use $\dot{\gamma}_A^{t,i[t]}$ as an estimate of $\gamma_A^{t,i[t]}$

Efficient estimator of MSE $v_{B\to A}^t$

We still need to compute the MSE $\tilde{v}_{A\to B}^t = \frac{1}{N}||\mathbf{x}_{A\to B}^t - \mathbf{x}||^2$

Using the definition of $\mathbf{x}_{A\to B}^t$ and of $\dot{\gamma}_A^{t,i[t]}$, one can show that is it equal to

$$\tilde{v}_{A\to B}^t = (N)^{-1}(\dot{\gamma}_A^{t,i[t]})^2(\boldsymbol{\mu}_A^{t,i[t]})^T\mathbf{A}\mathbf{A}^T\boldsymbol{\mu}_A^{t,i[t]} - v_{B\to A}^t$$

All the components are available

State Evolution of CG-VAMP

It can be shown that the exact solution to

$$\mathbf{g}_A(\mathbf{x}^t_{B\to A}) = \mathbf{W}_t^{-1}(\mathbf{y} - \mathbf{A}\mathbf{x}^t_{B\to A})$$

gives the optimal performance of VAMP w.r.t. the choice of $\mathbf{g}_A(\mathbf{x}^t_{B\to A})$

When we use CG, we sacrifice both **convergence rate** and **the quality of the fixed point** of VAMP

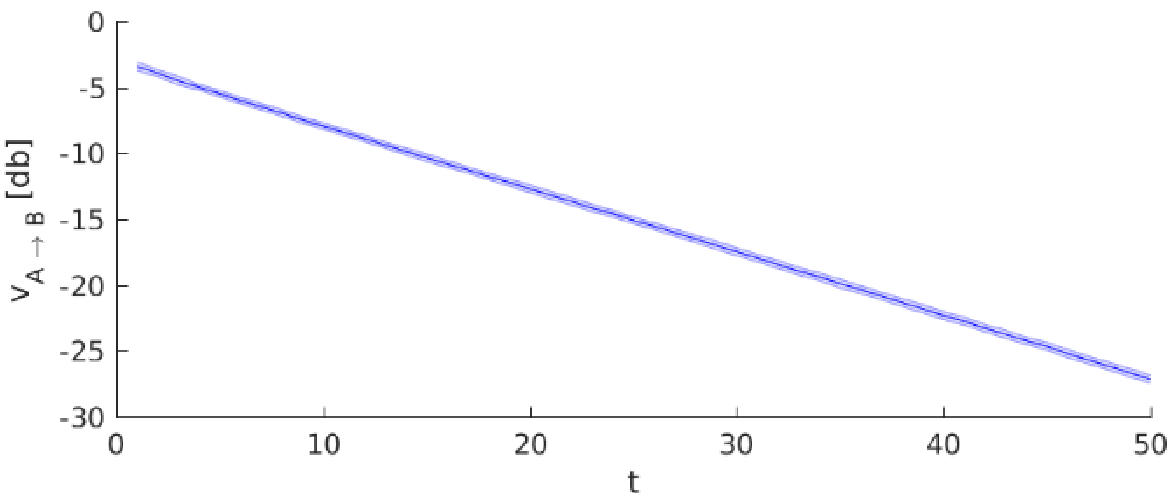
In order to preserve the efficiency and the quality, we adaptively choose the number of CG iterations and iterate while

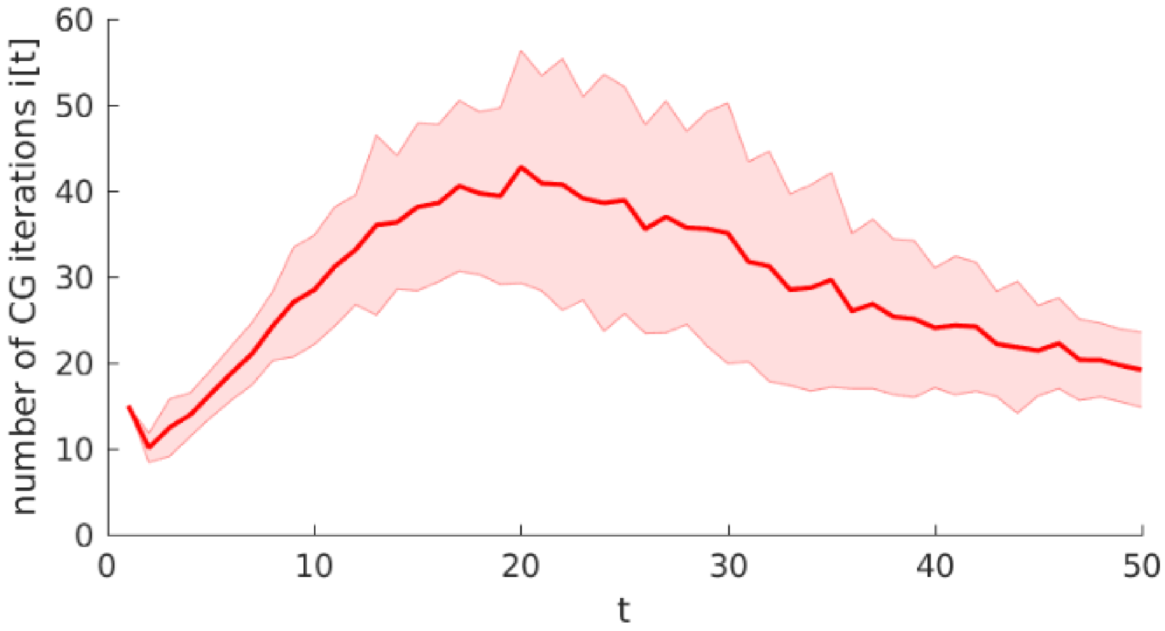
$$\tilde{v}^t_{A\to B}(i) \leq c\tilde{v}^{t-1}_{A\to B}$$

for some constant *c<1* that is larger than for the exact $\mathbf{g}_A(\mathbf{x}^t_{B\to A})$

Simulation results of adaptive CG for VAMP

- x is Bernoulli-Gaussian signal
- $N = 2^{14}$, $M = 2^{13}$
- geometric singular values
- condition number 10 000
- SNR = 40dB
- constant *c=0.9* for the variance reduction





Conclusions

- This work has presented efficient on-the-fly estimation of the variance and divergence terms for CG-VAMP using the concept of a synthetic statistical system
- This implementation does not rely on any prior information about the singular values of **A**
- We have presented an adaptive implementation of CG-VAMP in order to ensure a good convergence rate
- Simulations (not shown) based on Fast ill-conditioned Johnson-Lindenstrauss operators result in both fast and accurate reconstruction

References

[1] Sundeep Rangan, Philip Schniter, and Alyson K. Fletcher. “Vector Approximate Message Passing”

[2] Keigo Takeuchi. “Rigorous Dynamics of Expectation-Propagation-Based Signal Recovery from Unitarily Invariant Measurements”

[3] Keigo Takeuchi and Chao-Kai Wen. “Rigorous dynamics of expectation-propagation signal detection via the conjugate gradient method

[4] K. Dabov et al. “Image Denoising by Sparse 3-DTransform-Domain Collaborative Filtering

[5] Alyson K. Fletcher et al. “Plug-in Estimation in High-Dimensional Linear Inverse Problems: A Rigorous Analysis

[6] Philip Schniter, Sundeep Rangan, and Alyson K.Fletcher. “Denoising based Vector Approximate Message Passing

[7] Chunli Guo and Mike E. Davies. “Near optimal com-pressed sensing without priors: Parametric SURE Ap-proximate Message Passing