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 $\mathbf{x}$ from a set of linear measurements

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

Where

- $\mathbf{x} \in \mathbb{R}^N$
- $\mathbf{y} \in \mathbb{R}^M$
- $\mathbf{w} \sim N(0, \sigma_w^2 \mathbf{I}_M)$
- $\mathbf{A} \in \mathbb{R}^{M \times N}$

And we consider the compressed sensing scenario $M \ll N$ with both of a similar order
Inference via Bayes-motivated approach: EP

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

$$p(x|y) \propto p(y|A, x)p(x)$$

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 $\mu_{A\rightarrow B}(x) = N(x; x_{A\rightarrow B}, v_{A\rightarrow B}I_N)$

- $p(x)$ is approximated by $\mu_{B\rightarrow A}(x) = N(x; x_{B\rightarrow A}, v_{B\rightarrow A}I_N)$
Expectation Propagation updates

On that simple factor graph, the EP update rules are

\[
\mu_{A \rightarrow B}(x) = \frac{\mu_A(x)}{\mu_B \rightarrow A(x)} = \frac{\text{proj} [\mu_B \rightarrow A(x)p(y|x)]}{\mu_B \rightarrow A(x)}
\]

\[
\mu_B \rightarrow A(x) = \frac{\mu_A(x)}{\mu_A \rightarrow B(x)} = \frac{\text{proj} [\mu_A \rightarrow B(x)p(x)]}{\mu_A \rightarrow B(x)}
\]

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.

\[
\mu_{A \rightarrow B}(x) = N(x; x_{A \rightarrow B}, v_{A \rightarrow B}I_N)
\]

\[
\mu_{B \rightarrow A}(x) = N(x; x_{B \rightarrow A}, v_{B \rightarrow A}I_N)
\]
EP-based algorithm

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

Initialization: \( x_{0 \rightarrow A}^0 = 0, \tilde{v}_{0 \rightarrow A}^0 = \tilde{v}_x, t = 0 \)

while \( t < T_{max} \) and \( \tilde{v}_{B \rightarrow A}^t \geq \epsilon \) do

\[
\text{Block A}
\]

\[
\mu_A^t = g_A(x_{B \rightarrow A}^t, \tilde{v}_{B \rightarrow A}, \tilde{v}_w)
\]

\[
\gamma_A^t = \frac{1}{N} \nabla (x_{B \rightarrow A}^t) \cdot \left( A^T g_A(x_{B \rightarrow A}^t, \tilde{v}_{B \rightarrow A}^t, \tilde{v}_w) \right)
\]

\[
x_{A \rightarrow B}^t = x_{B \rightarrow A}^t - \gamma_A^t A^T \mu_A^t
\]

\[
\tilde{v}_{A \rightarrow B}^t = f_A(\mu_A^t, x_{B \rightarrow A}^t, \tilde{v}_{B \rightarrow A}, \tilde{v}_w)
\]

\[
\text{computation of } \mu_{A \rightarrow B}(x)
\]

\[
N(x; x_{A \rightarrow B}, v_{A \rightarrow B} I_N)
\]

\[
\text{Block B}
\]

\[
\mu_{B}^{t+1} = g_B(x_{A \rightarrow B}^t, \tilde{v}_{A \rightarrow B}^t)
\]

\[
\gamma_{B}^{t+1} = \frac{1}{N} \nabla x_{A \rightarrow B}^t \cdot g_B(x_{A \rightarrow B}^t, \tilde{v}_{A \rightarrow B}^t)
\]

\[
x_{B \rightarrow A}^{t+1} = \frac{1}{1 - \gamma_B} \left( \mu_{B}^{t+1} - \gamma_{B}^{t+1} x_{A \rightarrow B}^t \right)
\]

\[
\tilde{v}_{B \rightarrow A}^{t+1} = f_B(\mu_{B}^{t+1}, x_{B \rightarrow A}^t, \tilde{v}_{A \rightarrow B})
\]

\[
\text{computation of } \mu_{B \rightarrow A}(x)
\]

\[
N(x; x_{B \rightarrow A}, v_{B \rightarrow A} I_N)
\]

\[
t = t + 1
\]

Output: \( \mu_B^t \)
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

$$A = USV^T$$

the singular vector matrix $V$ is Haar distributed, while $U$ and $S$ can be any.
Implementation of Block B

Block B

\[
\begin{align*}
\mu_{B}^{t+1} &= g_B(x_{A\rightarrow B}^t, \tilde{v}_{A \rightarrow B}^t) \\
\gamma_{B}^{t+1} &= \frac{1}{N} \nabla x_{A \rightarrow B}^t \cdot g_B(x_{A \rightarrow B}^t, \tilde{v}_{A \rightarrow B}^t) \\
x_{B \rightarrow A}^{t+1} &= \frac{1}{1 - \gamma_{B}^{t+1}} \left( \mu_{B}^{t+1} - \gamma_{B}^{t+1} x_{A \rightarrow B}^t \right) \\
\tilde{v}_{B \rightarrow A}^{t+1} &= f_B(\mu_{B}^{t+1}, x_{B \rightarrow A}^t, \tilde{v}_{A \rightarrow B}^t)
\end{align*}
\]

Thus $g_B$ acts as a denoiser with measurements $x_{A \rightarrow B}^t$.

The scalar $\gamma_{B}^{t+1}$ is the divergence of the denoiser.

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

These components were well studied in [4], [5], [6], [7]
Properties of Block A

\[
\begin{align*}
\mu^t_A &= g_A(x^t_{B \rightarrow A}, \tilde{v}_{B \rightarrow A}, \tilde{v}_w) \\
\frac{1}{\gamma^t_A} &= \frac{1}{N} \nabla (x^t_{B \rightarrow A}) \cdot \left( A^T g_A(x^t_{B \rightarrow A}, \tilde{v}_{B \rightarrow A}, \tilde{v}_w) \right) \\
X^t_{A \rightarrow B} &= x^t_{B \rightarrow A} - \frac{1}{\gamma^t_A} A^T \mu^t_A \\
\tilde{v}^t_{A \rightarrow B} &= \tilde{f}_A(\mu^t_A, x^t_{B \rightarrow A}, \tilde{v}_{B \rightarrow A}, \tilde{v}_w)
\end{align*}
\]

It was shown that the mean \(x^t_{B \rightarrow A}\) of the approximated density \(\mu_{B \rightarrow A}(x)\) is equal to

\[
x^t_{B \rightarrow A} = x + q_t
\]

The function \(g_A\) is the LMMSE estimator

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

The scalar \(\frac{1}{\gamma^t_A}\) is the divergence of \(A^T g_A\)

- 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

$$g_A(x_{B \rightarrow A}^t) = W_t^{-1}(y - Ax_{B \rightarrow A}^t) = z_t$$

What about the divergence of the resulting $A^T g_A$ and the MSE $\tilde{v}_{A \rightarrow B}^t$?

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 as $N \to \infty$ and with Haar $V$, the CG function becomes a linear mapping

$$g^i_{A[t]} = U H^i_{t} U^T$$

of the vector $z_t$ and the diagonal matrix $H^i_{t}$ is a function of $S$, $v_w$ and $v_B^t \to A$ only.

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

$$\frac{1}{\gamma^t_A} = Tr \left\{ H^i_{t} S S^T \right\} = \frac{1}{N} q^T_A A^T g^i_{A[t]}(z_t) \tilde{v}_B^t \to A$$

which is independent of a particular realization of $w$ and $q_t$ 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 $q_t$ but is only a function of its statistics, synthesize

$$\hat{z}_t = \hat{w} - A\hat{q}_t$$

with

$$\hat{w} \sim N(0, v_wI_M)$$
$$\hat{q}_t \sim N(0, v_{t \rightarrow A}^t I_N)$$

Execute CG on the synthesized data. We expect

$$\frac{1}{\gamma_{t,i[t]}^A} = \frac{1}{N} \hat{q}_t^T A^T g_{A}^i[t](\hat{z}_t)$$

$$\tilde{v}_{B \rightarrow A}^t$$

to be close to the result with the real data. Use $\gamma_{t,i[t]}^A$ as an estimate of $\gamma_{t,i[t]}^A$.
Efficient estimator of MSE \( \hat{\nu}_{B \rightarrow A}^t \)

We still need to compute the MSE \( \hat{\nu}_{A \rightarrow B}^t = \frac{1}{N} \| \mathbf{x}_{A \rightarrow B}^t - \mathbf{x} \|^2 \)

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

\[
\hat{\nu}_{A \rightarrow B}^t = (N)^{-1} (\dot{\gamma}_A^{t,i[t]})^2 (\mu_A^{t,i[t]})^T A A^T \mu_A^{t,i[t]} - \nu_{B \rightarrow A}^t
\]

All the components are available
State Evolution of CG-VAMP

It can be shown that the exact solution to

$$g_A(x^t_{B\rightarrow A}) = W_t^{-1}(y - Ax^t_{B\rightarrow A})$$

gives the optimal performance of VAMP w.r.t. the choice of $g_A(x^t_{B\rightarrow A})$

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

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

for some constant $c<1$ that is larger than for the exact $g_A(x^t_{B\rightarrow A})$.
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
- $\text{SNR} = 40\text{dB}$
- 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