Learning and Free Energies for Vector Approximate Message Passing

Philip Schniter (Ohio State/Duke)
Alyson Fletcher (UCLA)

ICASSP 2017 – New Orleans
Supported by NSF grants 1527162 and 1254204.

Linear Regression with Unknown Prior/Likelihood

Consider the following linear regression problem:

- Observations:
$$y = Ax + w \quad \text{with} \quad \begin{cases} x : \text{unknown signal} \\ A : \text{known linear operator in } \mathbb{R}^{M\times N} \\ w : \text{white Gaussian noise.} \end{cases}$$
- Prior:
$$p(x;\theta_1) \quad \text{with deterministic unknown parameters } \theta_1.$$
- Likelihood:
$$\ell(x;\theta_2) = \mathcal{N}(y; Ax, \theta_2 I) \quad \text{with deterministic unknown variance } \theta_2.$$

Goal: jointly infer x and estimate $\theta \triangleq [\theta_1, \theta_2]$.

Approach: combine variational inference with ML estimation.

Variational Inference

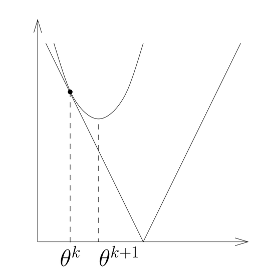
- For now, let's suppose that θ is known.
- We would like to compute the posterior density
$$p(x|y) = \frac{p(x;\theta_1)\ell(x;\theta_2)}{Z(\theta)} \quad \text{for} \quad Z(\theta) \triangleq \int p(x;\theta_1)\ell(x;\theta_2)\,\mathrm{d}x,$$
but the high-dimensional integral in $Z(\theta)$ is difficult to compute.
- We can avoid computing $Z(\theta)$ through variational optimization:
$$\begin{aligned} p(x|y) &= \arg\min_b D\big(b(x)\big\|p(x|y)\big) \text{ where } D(\cdot\|\cdot) \text{ is KL divergence} \\ &= \arg\min_b \underbrace{D\big(b(x)\big\|p(x;\theta_1)\big) + D\big(b(x)\big\|\ell(x;\theta_2)\big) + H\big(b(x)\big)}_{\text{Gibbs free energy}} \\ &= \arg\min_{b_1,b_2}\max_q \underbrace{D\big(b_1(x)\big\|p(x;\theta_1)\big) + D\big(b_2(x)\big\|\ell(x;\theta_2)\big) + H\big(q(x)\big)}_{\triangleq J(b_1,b_2,q;\theta)} \\ &\quad \text{such that } b_1 = b_2 = q, \end{aligned}$$
but the density constraint keeps the problem difficult.
- Expectation consistent approximation (EC) [1] relaxes the density constraint to moment-matching constraints:
$$p(x|y) \approx \arg\min_{b_1,b_2}\max_q J(b_1,b_2,q;\theta) \quad \text{such that } \begin{cases} \mathrm{E}\{x|b_1\} = \mathrm{E}\{x|b_2\} = \mathrm{E}\{x|q\} \\ \mathrm{tr}[\mathrm{Cov}\{x|b_1\}] = \mathrm{tr}[\mathrm{Cov}\{x|b_2\}] = \mathrm{tr}[\mathrm{Cov}\{x|q\}]. \end{cases}$$
- The stationary points of EC are
$$\begin{aligned} b_1(x) &\propto p(x;\theta_1)\mathcal{N}(x;r_1,v_1 I) \\ b_2(x) &\propto \ell(x;\theta_2)\mathcal{N}(x;r_2,v_2 I) \\ q(x) &= \mathcal{N}(x;\widehat{x},\widehat{v}I) \end{aligned} \quad \text{s.t.} \begin{cases} \mathrm{E}\{x|b_1\} = \mathrm{E}\{x|b_2\} = \widehat{x} \\ \mathrm{tr}[\mathrm{Cov}\{x|b_1\}] = \mathrm{tr}[\mathrm{Cov}\{x|b_2\}] = N\widehat{v}. \end{cases}$$

Vector AMP (VAMP)

- There exist several algorithms (e.g., EC, ADATAP [2], S-AMP [3]) whose fixed points coincide with the EC stationary points, but often they don't converge.
- An exception is Vector AMP [4], which can be derived using a form of approximation message passing on the vector-valued factor graph
$$p(x_1;\theta_1)\ \blacksquare\!-\!\bigcirc\overset{x_1}{-}\underset{\delta(x_1-x_2)}{\blacksquare}\overset{x_2}{-}\bigcirc\!-\!\blacksquare\ \ell(x_2,\theta_2)$$
In particular, VAMP is provably convergent under either
1) strictly log-concave prior $p(x;\theta_1)$ and arbitrary A (after damping),
2) iid prior $p(x;\theta_1)$ and large, right-rotationally invariant A.
- $A = U\,\mathrm{Diag}(s)V^{\mathsf{T}}$ is said to be "right-rotationally invariant" when V is uniformly distributed on the set of unitary matrices.
  - The other SVD quantities, U and S, are deterministic and arbitrary.
  - This model includes mean-perturbed and ill-conditioned A, known to break regular AMP.
- With large, right-rotationally invariant A, VAMP has a rigorous state evolution [4] whose fixed points match the replica prediction of MMSE [5].

Expectation maximization (EM)

- We now return to the problem of estimating θ.
- The maximum-likelihood (ML) estimate is
$$\widehat{\theta} = \arg\max_\theta p(y|\theta) = \arg\min_\theta \{-\ln p(y|\theta)\},$$
which is difficult to compute directly.
- Let's instead consider majorization-minimization: Iteratively minimize a tight upper bound on $-\ln p(y|\theta)$:
$$\widehat{\theta}^{k+1} = \arg\min_\theta \Big\{ -\ln p(y|\theta) + \underbrace{D\big(b^k(x)\big\|p(x|y;\theta)\big)}_{\geq 0} \Big\}$$
$$\text{with } b^k(x) = p(x|y;\widehat{\theta}_k)$$



- The upper bound "$Q(\theta,b^k)$" can be rewritten in the form
$$\begin{aligned} Q(\theta,b^k) &\triangleq -\ln p(y|\theta) + D\big(b^k(x)\big\|p(x|y;\theta)\big) \\ &= -\mathrm{E}\big\{\ln p(y,x;\theta)\,\big|\,b^k\big\} + \text{const}, \end{aligned}$$
which is the usual way of writing the EM algorithm, but it can also be written in terms of the Gibbs free energy
$$\begin{aligned} Q(\theta,b^k) &= D\big(b^k(x)\big\|p(x;\theta_1)\big) + D\big(b^k(x)\big\|\ell(x;\theta_2)\big) + H\big(b^k(x)\big) \\ &= J(b^k,b^k,b^k;\theta) \end{aligned}$$
which yields a variational interpretation of EM [6].

The Proposed EM-VAMP Algorithm

- Recall that VAMP iteratively computes a posterior approximation $b^k(x)$ by minimizing $J(b_1,b_2,q;\theta)$ (under moment constraints) with known θ.
- Likewise, EM iteratively estimates θ by minimizing $J(b^k,b^k,b^k;\theta)$ assuming the posterior approximation $b^k(x) = p(x|y;\theta^k)$ is available.
- We propose to combine EM and VAMP as follows:

Input g_1 and g_2, and initialize $r_1 = 0$ and $v_1 = \infty$.
For $k = 1,2,3,\dots$

$\widehat{\theta}_1 \leftarrow \arg\max_{\theta_1} \mathrm{E}\{\ln p(x;\theta_1)\,\vert\, r_1, v_1, \widehat{\theta}_1\}$	EM update
$\widehat{v}_1 \leftarrow N^{-1}\,\mathrm{tr}[\mathrm{Cov}\{x\vert r_1, v_1, \widehat{\theta}_1\}]$	posterior variance
$\widehat{x}_1 \leftarrow \mathrm{E}\{x\vert r_1, v_1, \widehat{\theta}_1\}$	denoising
$1/v_2 \leftarrow 1/\widehat{v}_1 - 1/v_1$	variance update
$r_2 \leftarrow (\widehat{x}_1/\widehat{v}_1 - r_1/v_1)v_2$	Onsager correction
$\widehat{\theta}_2 \leftarrow \arg\max_{\theta_2} \mathrm{E}\{\ln \ell(x;\theta_2)\,\vert\, r_2, v_2, \widehat{\theta}_2\}$	EM update
$\widehat{v}_2 \leftarrow N^{-1}\,\mathrm{tr}[\mathrm{Cov}\{x\vert r_2, v_2, \widehat{\theta}_2\}]$	posterior variance
$\widehat{x}_2 \leftarrow \mathrm{E}\{x\vert r_2, v_2, \widehat{\theta}_2\}$	LMMSE estimation
$1/v_1 \leftarrow 1/\widehat{v}_2 - 1/v_2$	variance update
$r_1 \leftarrow (\widehat{x}_2/\widehat{v}_2 - r_2/v_2)v_1$	Onsager correction

where
$$\mathrm{E}\{f(x)\,|\,r_1,v_1,\theta_1\} \triangleq \int f(x)\frac{p(x;\theta_1)\mathcal{N}(x;r_1,v_1 I)}{\int p(x';\theta_1)\mathcal{N}(x';r_1,v_1 I)\,\mathrm{d}x'}\,\mathrm{d}x$$
$$\mathrm{E}\{f(x)\,|\,r_2,v_2,\theta_2\} \triangleq \int f(x)\frac{\ell(x;\theta_2)\mathcal{N}(x;r_2,v_2 I)}{\int \ell(x';\theta_2)\mathcal{N}(x';r_2,v_2 I)\,\mathrm{d}x'}\,\mathrm{d}x$$
and similar for the covariances.

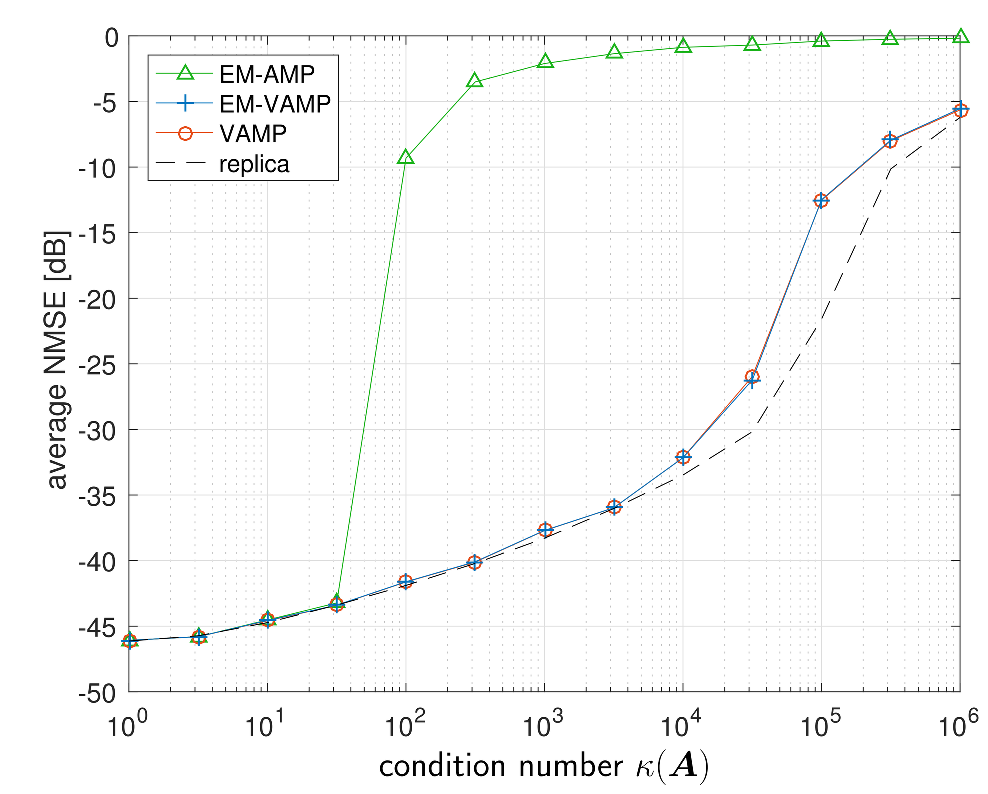
- If the SVD $A = U\,\mathrm{Diag}(s)V^{\mathsf{T}}$ is precomputed, then
$$\widehat{x}_2 \leftarrow V\big(v_2\,\mathrm{Diag}(s)^2 + \theta_2 I\big)^{-1}\big(v_2\,\mathrm{Diag}(s)U^{\mathsf{H}}y + \theta_2 V^{\mathsf{H}} r_2\big)$$
$$\widehat{v}_2 \leftarrow \frac{1}{N}\sum_{n=0}^{N-1}\frac{1}{s_n^2/\theta_2 + 1/v_2}, \qquad \widehat{\theta}_2 \leftarrow \frac{1}{N}\Big[\|y - Ar_2\|^2 + \sum_n \frac{s_n^2}{s_n^2/\theta_2 + 1/v_2}\Big],$$
so EM-VAMP requires only two matrix-vector mults per iteration.
- Other algorithmic variants result when θ_1 and/or θ_2 are updated more or less often than once per VAMP iteration.

Theorem: Fixed Points of EM-VAMP

At any fixed point of EM-VAMP we have
$$\widehat{v}_1 = \widehat{v}_2 = \frac{v_1 v_2}{v_1 + v_2} \triangleq \widehat{v}$$
$$\widehat{x}_1 = \widehat{x}_2 = \Big(\frac{r_1}{v_1} + \frac{r_2}{v_2}\Big)\widehat{v} \triangleq \widehat{x}.$$
Also, EM-VAMP's fixed-points are stationary points of the EM-EC optimization
$$\min_\theta \min_{b_1,b_2}\max_q J(b_1,b_2,q;\theta)$$
$$\text{such that } \begin{cases} \mathrm{E}\{x|b_1\} = \mathrm{E}\{x|b_2\} = \mathrm{E}\{x|q\} \\ \mathrm{tr}[\mathrm{Cov}\{x|b_1\}] = \mathrm{tr}[\mathrm{Cov}\{x|b_2\}] = \mathrm{tr}[\mathrm{Cov}\{x|q\}]. \end{cases}$$

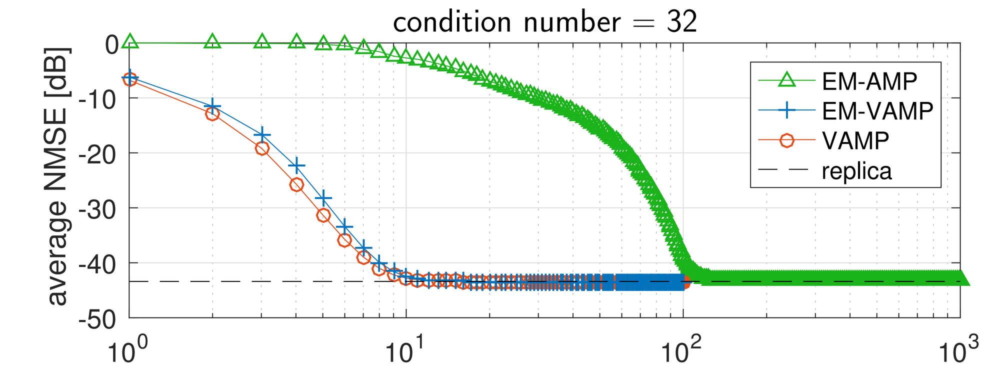
Numerical Experiments

Goal recover $N = 1024$-length i.i.d. Bernoulli-Gaussian x
$$p(x_n;\theta_1) = (1-\theta_{11})\delta(x_n) + \theta_{11}\mathcal{N}(x_n;\theta_{12},\theta_{13}) \quad \text{with} \quad \theta_1 = [0.1, 0, 1]$$
from $M = 512$ measurements
$$y = Ax + \mathcal{N}(0, \theta_2 I) \quad \text{with} \quad \theta_2 \text{ giving SNR=40 dB.}$$
Here, $A = U\,\mathrm{Diag}(s)V^{\mathsf{T}}$ with random orthogonal U, V and $s_n/s_{n-1} = \phi\ \forall n$, where ϕ determines the condition number $\kappa(A)$.

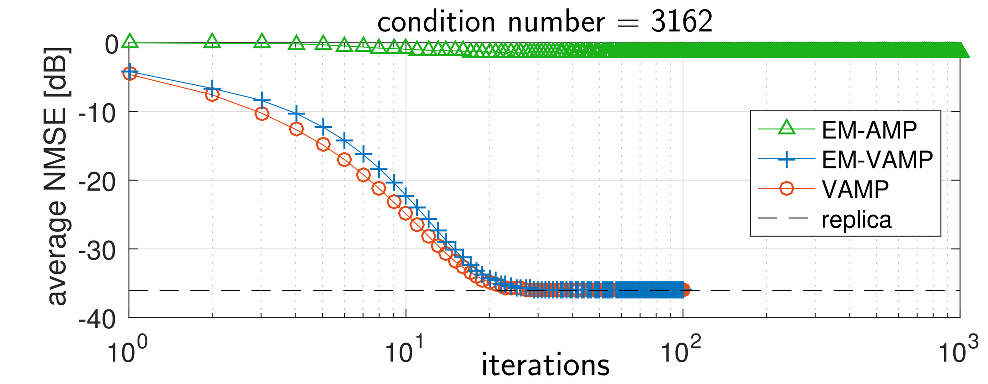


EM-VAMP matches known-θ-VAMP which matches the replica prediction of MMSE for a wide range of $\kappa(A)$.

EM-AMP [7] only works at small $\kappa(A)$.



EM-VAMP and known-θ-VAMP take ~ 10 iterations to converge, whereas EM-AMP takes ~ 100.



EM-VAMP and known-θ-VAMP take only ~ 25 iterations to converge, while EM-AMP fails.

References

[1] M. Opper and O. Winther, "Expectation consistent approximate inference," *J. Mach. Learning Res.*, 2005. .

[2] M. Opper and O. Winther, "Adaptive and self-averaging Thouless-Anderson-Palmer mean-field theory for probabilistic modeling," *Phy. Rev. E*, 2001. .

[3] B. Çakmak, O. Winther, and B.H. Fleury, "S-AMP: Approximate Message Passing for General Matrix Ensembles," *ISIT* 2014.

[4] S. Rangan, P. Schniter, and A. K. Fletcher, "Vector Approximate Message Passing," *arXiv:1610.03082*.

[5] A. M. Tulino, G. Caire, S. Verdú, and S. Shamai (Shitz), "Support recovery with sparsely sampled free random matrices," *IEEE Trans. Info. Thy.*, 2013.

[6] R. Neal and G. Hinton, "A view of the EM algorithm that justifies incremental, sparse, and other variants," in *Learning in Graphical Models*, 1998.

[7] J. P. Vila and P. Schniter, "Expectation-maximization Gaussian-mixture approximate message passing," *IEEE TSP*, 2013.