LETTER
Expectation Propagation Decoding for Sparse Superposition Codes

Hiroki MAYUMI[†], *Nonmember* and Keigo TAKEUCHI[†a)], *Member*

SUMMARY Expectation propagation (EP) decoding is proposed for sparse superposition coding in orthogonal frequency division multiplexing (OFDM) systems. When a randomized discrete Fourier transform (DFT) dictionary matrix is used, the EP decoding has the same complexity as approximate message-passing (AMP) decoding, which is a low-complexity and powerful decoding algorithm for the additive white Gaussian noise (AWGN) channel. Numerical simulations show that the EP decoding achieves comparable performance to AMP decoding for the AWGN channel. For OFDM systems, on the other hand, the EP decoding is much superior to the AMP decoding while the AMP decoding has an error-floor in high signal-to-noise ratio regime.
key words: *sparse superposition codes, orthogonal frequency division multiplexing (OFDM), discrete Fourier transform (DFT) dictionary, approximate message-passing, expectation propagation*

1. Introduction

Sparse superposition (SS) codes [1]–[3] are an error-correcting code achieving the Shannon capacity of the additive white Gaussian noise (AWGN) channel. A codeword of an SS code is generated as the multiplication of a dense dictionary matrix by a sparse information vector. Thus, the codes are called SS codes.

Approximate message-passing (AMP) decoding [4], [5] is a low-complexity and capacity-achieving algorithm for SS codes with zero-mean independent and identically distributed (i.i.d.) dictionary matrices. Numerical simulations in [4] showed that, when a randomized Hadamard dictionary matrix is used instead, AMP achieves good performance comparable to the case of zero-mean i.i.d. Gaussian matrices. Hadamard dictionary matrices allow us to implement low-complexity encoding and decoding of SS codes.

A limitation of AMP is that it fails to converge when the dictionary matrix is ill-conditioned [6]. This convergence issue is practically important since fading in wireless communication systems [7] might convert the dictionary matrix into an ill-conditioned effective dictionary matrix. The purpose of this letter is to propose a novel decoding algorithm that converges in fading channels.

As an important example of fading channels, we consider orthogonal frequency division multiplexing (OFDM). When SS coding is performed across OFDM subcarriers,

the effective dictionary matrix is the product of the original dictionary matrix and a diagonal matrix that consists of channel gains in frequency domain. Fading makes the diagonal matrix ill-conditioned.

We extend expectation propagation (EP) [8] in compressed sensing to the decoding issue of SS codes. EP can be regarded as a Bayes-optimal version of orthogonal AMP [9] or equivalently vector AMP [10], which was originally proposed in [11]. The main advantage of EP is the Bayes-optimality for all unitarily invariant matrices [8], [10], including ill-conditioned effective dictionary matrices. Numerical simulations in this letter show that the proposed EP decoder has good convergence properties for ill-conditioned effective dictionary matrices, where AMP without damping fails to converge.

2. System Model

2.1 OFDM

We consider OFDM transmission of block length N_b [7]. A complex codeword of length $N > N_b$ is sent over $K = N/N_b$ OFDM blocks. For simplicity, we assume that N is divisible by N_b. The frequency-domain received vector $\boldsymbol{y}_k \in \mathbb{C}^{N_b}$ in OFDM block $k \in \{0, \ldots, K-1\}$ is given by

$$\boldsymbol{y}_k = \boldsymbol{\Lambda}_k \boldsymbol{c}_k + \boldsymbol{w}_k, \quad \boldsymbol{w}_k \in \mathcal{CN}(\boldsymbol{0}, N_0 \boldsymbol{I}_{N_b}). \tag{1}$$

In (1), $\boldsymbol{c}_k \in \mathbb{C}^{N_b}$ is part of a codeword $\boldsymbol{c} \in \mathbb{C}^N$ in frequency domain, i.e. $\boldsymbol{c} = (\boldsymbol{c}_0^T, \ldots, \boldsymbol{c}_{K-1}^T)^T$. The vectors $\{\boldsymbol{w}_k\}$ are independent AWGN vectors with variance N_0. The nth diagonal element $\lambda_{k,n}$ of the complex diagonal matrix $\boldsymbol{\Lambda}_k = \text{diag}(\lambda_{k,0}, \ldots, \lambda_{k,N_b-1})$ represents the channel gain of subcarrier $n \in \{0, \ldots, N_b - 1\}$ in OFDM block k. Assuming the use of cyclic prefix longer than the delay spread of the fading channels, we have

$$\lambda_{k,n} = \sum_{p=0}^{N_b-1} h_{k,p} e^{-2\pi j pn/N_b}, \tag{2}$$

where $h_{k,p} \in \mathbb{C}$ denotes the time-domain channel gain of the pth resolvable path in OFDM block k.

2.2 SS Coding

We consider a complex SS code with length N, L sections, and section size M. The codeword $\boldsymbol{c} = \boldsymbol{D}\boldsymbol{\beta} \in \mathbb{C}^N$

is generated as the multiplication of a complex dictionary matrix $\boldsymbol{D} \in \mathbb{C}^{N \times ML}$ by an information vector $\boldsymbol{\beta} = (\boldsymbol{\beta}^{\mathrm{T}}[0], \ldots, \boldsymbol{\beta}^{\mathrm{T}}[L-1])^{\mathrm{T}}$. We assume the uniform power allocation and write the average power as $P > 0$. Let $\boldsymbol{e}_m$ denote the mth column of the $M \times M$ identity matrix. The information vector $\boldsymbol{\beta}[l] \in \mathbb{C}^M$ in section l is a 1-sparse vector $\sqrt{PM}\boldsymbol{e}_m$, in which the message m is sampled from the index set $\{0, \ldots, M-1\}$ uniformly and randomly. Since we have M different codewords per section, the transmission rate per complex channel use is defined as

$$R = \frac{L}{N} \log_2 M. \tag{3}$$

We use a randomized discrete Fourier transform (DFT) dictionary matrix $\boldsymbol{D}$ obtained by selecting N different rows from the rows of an $ML \times ML$ DFT matrix uniformly and randomly. The norm of each row is normalized to 1. Since $\mathbb{E}[\boldsymbol{\beta}\boldsymbol{\beta}^{\mathrm{H}}] = P\boldsymbol{I}_{ML}$ holds, this normalization implies the average power constraint

$$\frac{1}{N}\mathbb{E}_{\boldsymbol{\beta}}\left[\|\boldsymbol{c}\|^2\right] = \frac{1}{N}\mathrm{Tr}\left(\boldsymbol{D}\mathbb{E}\left[\boldsymbol{\beta}\boldsymbol{\beta}^{\mathrm{H}}\right]\boldsymbol{D}^{\mathrm{H}}\right) = P. \tag{4}$$

The DFT dictionary matrix allows us to implement an efficient SS encoder. When the fast Fourier transform is used, the computational complexity in encoding is $O(ML \log ML)$ since N is smaller than ML in general.

3. Expectation Propagation

Let $\boldsymbol{y} = (\boldsymbol{y}_0^{\mathrm{T}}, \ldots, \boldsymbol{y}_{K-1}^{\mathrm{T}})^{\mathrm{T}}$ and $\boldsymbol{w} = (\boldsymbol{w}_0^{\mathrm{T}}, \ldots, \boldsymbol{w}_{K-1}^{\mathrm{T}})^{\mathrm{T}}$ in (1). To propose EP decoding, we rewrite the SS-coded OFDM system (1) as

$$\boldsymbol{y} = \boldsymbol{A}\boldsymbol{\beta} + \boldsymbol{w}, \quad \boldsymbol{A} = \boldsymbol{\Lambda}\boldsymbol{D}, \tag{5}$$

with $\boldsymbol{\Lambda} = \mathrm{diag}(\boldsymbol{\Lambda}_0, \ldots, \boldsymbol{\Lambda}_{K-1})$. The purpose of the decoder is to estimate the information vector $\boldsymbol{\beta}$ from the knowledge about the received vector $\boldsymbol{y}$ and the effective dictionary matrix $\boldsymbol{A}$.

As derived in Appendix, the proposed EP decoder consists of two modules—called modules A and B. Suppose that the extrinsic mean $\boldsymbol{\beta}_{\mathrm{B}\to\mathrm{A},t} \in \mathbb{C}^{ML}$ and variance $v_{\mathrm{B}\to\mathrm{A},t} > 0$ of the information vector $\boldsymbol{\beta}$ have been passed from module B to module A. The module A uses the linear minimum-mean square error (LMMSE) filter to compute the extrinsic messages $\boldsymbol{\beta}_{\mathrm{A}\to\mathrm{B},t} \in \mathbb{C}^{ML}$ and variance $v_{\mathrm{A}\to\mathrm{B},t} > 0$.

$$\boldsymbol{\beta}_{\mathrm{A}\to\mathrm{B},t} = \boldsymbol{\beta}_{\mathrm{B}\to\mathrm{A},t} + \gamma_t \boldsymbol{A}^{\mathrm{H}}\boldsymbol{\Xi}_t^{-1}(\boldsymbol{y} - \boldsymbol{A}\boldsymbol{\beta}_{\mathrm{B}\to\mathrm{A},t}), \tag{6}$$

$$v_{\mathrm{A}\to\mathrm{B},t} = \gamma_t - v_{\mathrm{B}\to\mathrm{A},t}, \tag{7}$$

with

$$\boldsymbol{\Xi}_t = N_0\boldsymbol{I}_N + v_{\mathrm{B}\to\mathrm{A},t}\boldsymbol{A}\boldsymbol{A}^{\mathrm{H}}, \tag{8}$$

$$\gamma_t^{-1} = \frac{1}{ML}\mathrm{Tr}\left(\boldsymbol{\Xi}_t^{-1}\boldsymbol{A}\boldsymbol{A}^{\mathrm{H}}\right)^{-1}. \tag{9}$$

In the initial iteration, $\boldsymbol{\beta}_{\mathrm{B}\to\mathrm{A},0} = \boldsymbol{0}$ and $v_{\mathrm{B}\to\mathrm{A},0} = P$ are used.

Remark 1: Module A requires the matrix inversion $\boldsymbol{\Xi}_t^{-1}$ in

(6). While the singular-value decomposition (SVD) of $\boldsymbol{A}$ allows us to circumvent this high-complexity matrix inversion [10], the SVD itself needs high complexity in general. Fortunately, this complexity issue does not occur in OFDM systems because the SVD $\boldsymbol{A} = \boldsymbol{\Lambda}\boldsymbol{D}$ is given explicitly. ☐

Module B uses $\boldsymbol{\beta}_{\mathrm{A}\to\mathrm{B},t} = (\boldsymbol{\beta}_{\mathrm{A}\to\mathrm{B},t}^{\mathrm{T}}[0], \ldots, \boldsymbol{\beta}_{\mathrm{A}\to\mathrm{B},t}^{\mathrm{T}}[L-1])^{\mathrm{T}} \in \mathbb{C}^{ML}$ and $v_{\mathrm{A}\to\mathrm{B},t}$ to compute the posterior mean $\boldsymbol{\beta}_{\mathrm{B},t+1} = (\boldsymbol{\beta}_{\mathrm{B},t+1}^{\mathrm{T}}[0], \ldots, \boldsymbol{\beta}_{\mathrm{B},t+1}^{\mathrm{T}}[L-1])^{\mathrm{T}} \in \mathbb{R}^{ML}$ and variance $v_{\mathrm{B}}^{t+1} > 0$. Consider the virtual AWGN channel for section l

$$\boldsymbol{\beta}_{\mathrm{A}\to\mathrm{B},t}[l] = \boldsymbol{\beta}[l] + \boldsymbol{z}_t[l], \quad \boldsymbol{z}_t[l] \sim \mathcal{CN}(\boldsymbol{0}, v_{\mathrm{A}\to\mathrm{B},t}\boldsymbol{I}_M). \tag{10}$$

The posterior mean and variance are given by

$$\boldsymbol{\beta}_{\mathrm{B},t+1}[l] = \mathbb{E}\left[\boldsymbol{\beta}[l]|\boldsymbol{\beta}_{\mathrm{A}\to\mathrm{B},t}[l]\right], \tag{11}$$

$$v_{\mathrm{B},t+1} = \frac{1}{LM}\sum_{l=0}^{L-1}\mathbb{E}\left[\|\boldsymbol{\beta}[l] - \boldsymbol{\beta}_{\mathrm{B},t+1}[l]\|^2\big|\boldsymbol{\beta}_{\mathrm{A}\to\mathrm{B},t}[l]\right]. \tag{12}$$

Since $\boldsymbol{\beta}[l]$ follows the uniform distribution on the discrete set $\{\sqrt{PM}\boldsymbol{e}_0, \ldots, \sqrt{PM}\boldsymbol{e}_{M-1}\}$, we have the explicit formulas,

$$\beta_{\mathrm{B},t+1,m} = \frac{\sqrt{PM}e^{2\sqrt{PM}\Re[\beta_{\mathrm{A}\to\mathrm{B},t,m}]/v_{\mathrm{A}\to\mathrm{B},t}}}{\sum_{m'=0}^{M-1}e^{2\sqrt{PM}\Re[\beta_{\mathrm{A}\to\mathrm{B},t,m'}]/v_{\mathrm{A}\to\mathrm{B},t}}}, \tag{13}$$

$$v_{\mathrm{B},t+1} = P - \frac{1}{LM}\sum_{l=0}^{L-1}\|\boldsymbol{\beta}_{\mathrm{B},t+1}[l]\|^2. \tag{14}$$

In (13) and (14), $\beta_{\mathrm{A}\to\mathrm{B},t,m}$ and $\beta_{\mathrm{B},t+1,m}$ denote the mth elements of $\boldsymbol{\beta}_{\mathrm{A}\to\mathrm{B},t}$ and $\boldsymbol{\beta}_{\mathrm{B},t+1}$, respectively. The notation $\Re[z]$ means the real part of a complex number $z \in \mathbb{C}$.

Estimation of the information vector $\boldsymbol{\beta}[l]$ is based on the hard decision of $\boldsymbol{\beta}_{\mathrm{B},t+1}[l]$. To improve the decoding performance, module B feeds the extrinsic messages $\boldsymbol{\beta}_{\mathrm{B}\to\mathrm{A},t+1}$ and $v_{\mathrm{B}\to\mathrm{A},t+1}$ back to module A,

$$\boldsymbol{\beta}_{\mathrm{B}\to\mathrm{A},t+1} = v_{\mathrm{B}\to\mathrm{A},t+1}\left(\frac{\boldsymbol{\beta}_{\mathrm{B},t+1}}{v_{\mathrm{B},t+1}} - \frac{\boldsymbol{\beta}_{\mathrm{A}\to\mathrm{B},t}}{v_{\mathrm{A}\to\mathrm{B},t}}\right), \tag{15}$$

$$\frac{1}{v_{\mathrm{B}\to\mathrm{A},t+1}} = \frac{1}{v_{\mathrm{B},t+1}} - \frac{1}{v_{\mathrm{A}\to\mathrm{B},t}}. \tag{16}$$

The EP decoder may have a bad convergence property for finite-sized systems. To improve the convergence property, we replace the messages $\boldsymbol{\beta}_{\mathrm{B}\to\mathrm{A},t+1}$ and $v_{\mathrm{B}\to\mathrm{A},t+1}$ with the damped messages $\theta\boldsymbol{\beta}_{\mathrm{B}\to\mathrm{A},t+1} + (1-\theta)\boldsymbol{\beta}_{\mathrm{B}\to\mathrm{A},t}$ and $\theta v_{\mathrm{B}\to\mathrm{A},t+1} + (1-\theta)v_{\mathrm{B}\to\mathrm{A},t}$ for damping factor $\theta \in [0, 1]$, respectively.

The proposed EP decoding with T iterations is presented in Algorithm 1. The computational complexity of the EP decoding is dominated by the updates in module A. For a DFT dictionary matrix, they can be computed in $O(ML \log ML)$ time. Thus, the proposed EP decoding has the same complexity per iteration as AMP decoding [4], [5].

Algorithm 1 EP Decoding

Require: Received vector y and the effective dictionary matrix A.
1: Let $\boldsymbol{\beta}_{\mathrm{B}\to\mathrm{A},0} = \mathbf{0}$ and $v_{\mathrm{B}\to\mathrm{A},0} = P$.
2: **for** $t = 0, \ldots, T-1$ **do**
3:     Compute γ_t given in (9).
4:     Compute $\boldsymbol{\beta}_{\mathrm{A}\to\mathrm{B},t}$ and $v_{\mathrm{A}\to\mathrm{B},t}$ given in (6) and (7).
5:     Compute $\boldsymbol{\beta}_{\mathrm{B},t+1}$ and $v_{\mathrm{B},t+1}$ given in (13) and (14).
6:     Compute $\boldsymbol{\beta}_{\mathrm{B}\to\mathrm{A},t+1}$ and $v_{\mathrm{B}\to\mathrm{A},t+1}$ given in (15) and (16).
7:     Update $\boldsymbol{\beta}_{\mathrm{B}\to\mathrm{A},t+1} \leftarrow \theta\boldsymbol{\beta}_{\mathrm{B}\to\mathrm{A},t+1} + (1-\theta)\boldsymbol{\beta}_{\mathrm{B}\to\mathrm{A},t}$.
8:     Update $v_{\mathrm{B}\to\mathrm{A},t+1} \leftarrow \theta v_{\mathrm{B}\to\mathrm{A},t+1} + (1-\theta)v_{\mathrm{B}\to\mathrm{A},t}$.
9: **end for**
10: Output hard decision of $\boldsymbol{\beta}_{\mathrm{B},t+1}$.

---

4. Numerical Simulation

The EP decoding is compared to conventional AMP decoding [4], [5] in terms of section error rate (SER). We simulated damped AMP decoding with two damping factors $\theta_{\mathrm{A}} \in [0,1]$ and $\theta_{\mathrm{B}} \in [0,1]$ shown in Algorithm 2. The average power E_{b} per information bit is defined as $E_{\mathrm{b}} = P/R$, with the rate R given in (3).

We first compare the EP decoding with the AMP decoding for artificial fading channels. Let $\lambda_n = [\boldsymbol{\Lambda}]_{n,n}$ denote the nth diagonal element of $\boldsymbol{\Lambda}$ in (5). For condition number $\kappa = \lambda_0/\lambda_{N-1} \in [1,\infty)$, we assume $\lambda_n = d\lambda_{n-1}$ for $d = \kappa^{-(N-1)^{-1}}$ and $N^{-1}\sum_{n=0}^{N-1}\lambda_0^2 = 1$. In particular, $\kappa = 1$ implies the AWGN channel $\lambda_n = 1$.

Figure 1 shows the SERs of the EP and AMP decoding for the artificial fading channels. The two algorithms are comparable to each other for the AWGN channel $\kappa = 1$. The AMP decoding has poor performance when the condition number is larger than 2. On the other hand, the EP decoding has comparable SER to the AWGN channel $\kappa = 1$ when κ is below 3. These results imply that the EP decoding is robust against ill-conditioned fading channels.


Algorithm 2 AMP Decoding

Require: Received vector y and the effective dictionary matrix A.
1: Let $\boldsymbol{\beta}_{\mathrm{B},0} = \mathbf{0}$, $v_{\mathrm{B},0} = P$, $\boldsymbol{\beta}_{\mathrm{A},-1} = \mathbf{0}$, $v_{\mathrm{A},-1} = P$, and $z_{-1} = \mathbf{0}$.
2: **for** $t = 0, \ldots, T-1$ **do**
3:     Compute $z_t = y - A\boldsymbol{\beta}_{\mathrm{B},t} + v_{\mathrm{B},t}z_{t-1}/v_{\mathrm{A},t-1}$.
4:     Compute $\boldsymbol{\beta}_{\mathrm{A},t} = N\boldsymbol{\beta}_{\mathrm{B},t}/(ML) + A^{\mathrm{H}}z_t$.
5:     Compute $v_{\mathrm{A},t} = N_0 + v_{\mathrm{B},t}$.
6:     Update $\boldsymbol{\beta}_{\mathrm{A},t} \leftarrow \theta_{\mathrm{A}}\boldsymbol{\beta}_{\mathrm{A},t} + (1-\theta_{\mathrm{A}})\boldsymbol{\beta}_{\mathrm{A},t-1}$.
7:     Update $v_{\mathrm{A},t} \leftarrow \theta_{\mathrm{A}}v_{\mathrm{A},t} + (1-\theta_{\mathrm{A}})v_{\mathrm{A},t-1}$.
8:     Compute $\boldsymbol{\beta}_{\mathrm{B},t+1}$ and $v_{\mathrm{B},t+1}$ given in (13) and (14).
9:     Update $\boldsymbol{\beta}_{\mathrm{B},t+1} \leftarrow \theta_{\mathrm{B}}\boldsymbol{\beta}_{\mathrm{B},t+1} + (1-\theta_{\mathrm{B}})\boldsymbol{\beta}_{\mathrm{B},t}$.
10:     Update $v_{\mathrm{B},t+1} \leftarrow \theta_{\mathrm{B}}v_{\mathrm{B},t+1} + (1-\theta_{\mathrm{B}})v_{\mathrm{B},t}$.
11: **end for**
12: Output hard decision of $\boldsymbol{\beta}_{\mathrm{B},t+1}$.

We next compare the EP decoding with the AMP decoding in OFDM systems. We assume independent Rayleigh fading $h_{k,p} \sim \mathcal{CN}(0, \sigma_p^2)$ in (2) with the exponential power decay $\sigma_p^2 = Ce^{-0.1p}$, in which C is the normalization constant to impose $\sum_{p=0}^{N_{\mathrm{b}}-1}\sigma_p^2 = 1$. The average condition number of this fading channel is very large, e.g. approximately
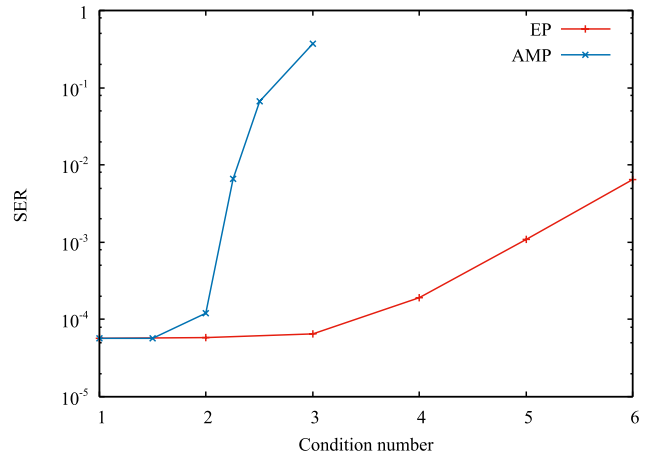
Fig. 1 SER versus the condition number κ for the artificial fading channels. $L = 256$, $M = 16$, $N = 512$, $E_{\mathrm{b}}/N_0 = 4$ dB, 20 iterations, $\theta = 0.9$ in the EP decoding, $\theta_{\mathrm{A}} = 1$ and $\theta_{\mathrm{B}} = 0.9$ in the AMP decoding.
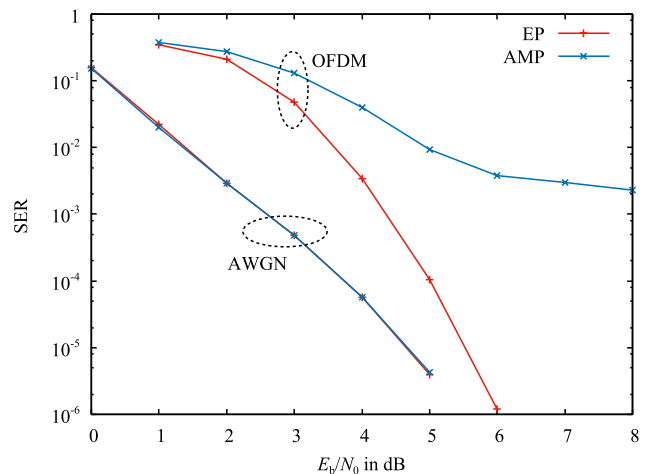
Fig. 2 SER versus E_{b}/N_0 for $L = 256$, $M = 16$, and $N = 512$. For OFDM systems with block length $N_{\mathrm{b}} = 64$, 50 iterations, $\theta = 0.8$ in the EP decoding, and $\theta_{\mathrm{A}} = \theta_{\mathrm{B}} = 0.5$ in the AMP decoding were used. For the AWGN channel, 20 iterations, $\theta = 0.9$, $\theta_{\mathrm{A}} = 1$, and $\theta_{\mathrm{B}} = 0.9$ were used.

100 for $N = 512$ and $N_{\mathrm{b}} = 64$.

Figure 2 shows the SERs of the EP and AMP decoding for both OFDM and AWGN channels. For the AWGN channel, the EP decoding is comparable to the AMP decoding for all signal-to-noise ratios (SNRs). For OFDM systems, the EP decoding is much superior to the AMP decoding while the AMP decoding has an error-floor in the high SNR regime. The poor performance of the AMP decoding is due to ill-conditioned effective dictionary matrices. The EP decoding has a good convergence property even for such an ill-conditioned case.

18H01441), Japan.

References

[1] A. Joseph and A.R. Barron, "Least squares superposition codes of moderate dictionary size are reliable at rates up to capacity," IEEE Trans. Inf. Theory, vol.58, no.5, pp.2541–2557, May 2012.

[2] A. Joseph and A.R. Barron, "Fast sparse superposition codes have near exponential error probability for $R < C$," IEEE Trans. Inf. Theory, vol.60, no.2, pp.919–942, Feb. 2014.

[3] Y. Takeishi, M. Kawakita, and J. Takeuchi, "Least squares superposition codes with Bernoulli dictionary are still reliable at rates up to capacity," IEEE Trans. Inf. Theory, vol.60, no.5, pp.2737–2750, May 2014.

[4] C. Rush, A. Greig, and R. Venkataramanan, "Capacity-achieving sparse superposition codes via approximate message passing decoding," IEEE Trans. Inf. Theory, vol.63, no.3, pp.1476–1500, March 2017.

[5] J. Barbier and F. Krzakala, "Approximate message-passing decoder and capacity achieving sparse superposition codes," IEEE Trans. Inf. Theory, vol.63, no.8, pp.4894–4927, Aug. 2017.

[6] S. Rangan, P. Schniter, A. Fletcher, and S. Sarkar, "On the convergence of approximate message passing with arbitrary matrices," IEEE Trans. Inf. Theory, vol.65, no.9, pp.5339–5351, Sept. 2019.

[7] D.N.C. Tse and P. Viswanath, Fundamentals of Wireless Communication, Cambridge University Press, Cambridge, UK, 2005.

[8] K. Takeuchi, "Rigorous dynamics of expectation-propagation-based signal recovery from unitarily invariant measurements," IEEE Trans. Inf. Theory, vol.66, no.1, pp.368–386, Jan. 2020.

[9] J. Ma and L. Ping, "Orthogonal AMP," IEEE Access, vol.5, pp.2020–2033, Jan. 2017.

[10] S. Rangan, P. Schniter, and A.K. Fletcher, "Vector approximate message passing," IEEE Trans. Inf. Theory, vol.65, no.10, pp.6664–6684, Oct. 2019.

[11] M. Opper and O. Winther, "Expectation consistent approximate inference," J. Mach. Learn. Res., vol.6, pp.2177–2204, Dec. 2005.

Appendix: Derivation of EP Decoding

We follow [8] to derive the EP decoding. The purpose of EP is to compute the marginal posterior distribution $p(\boldsymbol{\beta}[l]|\boldsymbol{y}, \boldsymbol{A})$ approximately. The marginal posterior distribution is approximated via a tractable distribution

$$q_{\mathrm{A}}(\boldsymbol{\beta}) \propto p(\boldsymbol{y}|\boldsymbol{A}, \boldsymbol{\beta}) \prod_{l=0}^{L-1} q_{\mathrm{B} \to \mathrm{A}}(\boldsymbol{\beta}[l]), \qquad (\mathrm{A}\cdot 1)$$

with

$$q_{\mathrm{B} \to \mathrm{A}}(\boldsymbol{\beta}[l]) \propto \exp\left(-\frac{\|\boldsymbol{\beta}[l] - \boldsymbol{\beta}_{\mathrm{B} \to \mathrm{A}}[l]\|^2}{v_{\mathrm{B} \to \mathrm{A}}}\right), \qquad (\mathrm{A}\cdot 2)$$

where $f(\boldsymbol{x}) \propto g(\boldsymbol{x})$ means that there is an $\boldsymbol{x}$-independent constant $C > 0$ such that $f(\boldsymbol{x}) = Cg(\boldsymbol{x})$ holds. Let $\boldsymbol{\beta}_{\mathrm{B} \to \mathrm{A}} = (\boldsymbol{\beta}_{\mathrm{B} \to \mathrm{A}}[0]^{\mathrm{T}}, \ldots, \boldsymbol{\beta}_{\mathrm{B} \to \mathrm{A}}[L-1]^{\mathrm{T}})^{\mathrm{T}}$ and $\boldsymbol{\beta}_{\mathrm{A}} = (\boldsymbol{\beta}_{\mathrm{A}}[0]^{\mathrm{T}}, \ldots, \boldsymbol{\beta}_{\mathrm{A}}[L-1]^{\mathrm{T}})^{\mathrm{T}}$. Following [8], we can evaluate the marginalization of (A·1) over $\{\boldsymbol{\beta}[l'] : l' \neq l\}$ as

$$q_{\mathrm{A}}(\boldsymbol{\beta}[l]) \propto \exp\left(-\frac{\|\boldsymbol{\beta}[l] - \boldsymbol{\beta}_{\mathrm{A}}[l]\|^2}{v_{\mathrm{A}}}\right), \qquad (\mathrm{A}\cdot 3)$$

with

$$\boldsymbol{\beta}_{\mathrm{A}} = \boldsymbol{\beta}_{\mathrm{B} \to \mathrm{A}} + v_{\mathrm{B} \to \mathrm{A}} \boldsymbol{A}^{\mathrm{H}} \boldsymbol{\Xi}^{-1}(\boldsymbol{y} - \boldsymbol{A}\boldsymbol{\beta}_{\mathrm{B} \to \mathrm{A}}), \qquad (\mathrm{A}\cdot 4)$$

$$v_{\mathrm{A}} = v_{\mathrm{B} \to \mathrm{A}} - \gamma^{-1}(v_{\mathrm{B} \to \mathrm{A}}) v_{\mathrm{B} \to \mathrm{A}}^2, \qquad (\mathrm{A}\cdot 5)$$

$$\boldsymbol{\Xi} = N_0 \boldsymbol{I}_N + v_{\mathrm{B} \to \mathrm{A}} \boldsymbol{A}\boldsymbol{A}^{\mathrm{H}}, \qquad (\mathrm{A}\cdot 6)$$

$$\gamma^{-1}(v) = \frac{1}{LM} \mathrm{Tr}\left(\boldsymbol{\Xi}^{-1} \boldsymbol{A}\boldsymbol{A}^{\mathrm{H}}\right). \qquad (\mathrm{A}\cdot 7)$$

The main difference between conventional [8] and proposed EP is in the update rule of $q_{\mathrm{B} \to \mathrm{A}}$. We define the extrinsic distribution of $\boldsymbol{\beta}[l]$ as

$$q_{\mathrm{A} \to \mathrm{B}}(\boldsymbol{\beta}[l]) \propto \frac{q_{\mathrm{A}}(\boldsymbol{\beta}[l])}{q_{\mathrm{B} \to \mathrm{A}}(\boldsymbol{\beta}[l])}. \qquad (\mathrm{A}\cdot 8)$$

We write the mean and variance of $\boldsymbol{\beta}[l]$ with respect to $q_{\mathrm{A} \to \mathrm{B}}(\boldsymbol{\beta}[l])p(\boldsymbol{\beta}[l])$ as

$$\boldsymbol{\beta}_{\mathrm{B}}[l] = \frac{\sum_{\boldsymbol{\beta}[l]} \boldsymbol{\beta}[l] q_{\mathrm{A} \to \mathrm{B}}(\boldsymbol{\beta}[l]) p(\boldsymbol{\beta}[l])}{\sum_{\boldsymbol{\beta}[l]} q_{\mathrm{A} \to \mathrm{B}}(\boldsymbol{\beta}[l]) p(\boldsymbol{\beta}[l])}, \qquad (\mathrm{A}\cdot 9)$$

$$v_{\mathrm{B}}[l] = \frac{1}{M} \frac{\sum_{\boldsymbol{\beta}[l]} \|\boldsymbol{\beta}[l]\|^2 q_{\mathrm{A} \to \mathrm{B}}(\boldsymbol{\beta}[l]) p(\boldsymbol{\beta}[l])}{\sum_{\boldsymbol{\beta}[l]} q_{\mathrm{A} \to \mathrm{B}}(\boldsymbol{\beta}[l]) p(\boldsymbol{\beta}[l])}$$
$$- \frac{1}{M} \|\boldsymbol{\beta}_{\mathrm{B}}[l]\|^2. \qquad (\mathrm{A}\cdot 10)$$

Let $q_{\mathrm{B} \to \mathrm{A}}^{\mathrm{new}}(\boldsymbol{\beta}[l])$ denote an updated message of $q_{\mathrm{B} \to \mathrm{A}}(\boldsymbol{\beta}[l])$ given by

$$q_{\mathrm{B} \to \mathrm{A}}^{\mathrm{new}}(\boldsymbol{\beta}[l]) \propto \exp\left(-\frac{\|\boldsymbol{\beta}[l] - \boldsymbol{\beta}_{\mathrm{B} \to \mathrm{A}}^{\mathrm{new}}[l]\|^2}{v_{\mathrm{B} \to \mathrm{A}}^{\mathrm{new}}}\right). \qquad (\mathrm{A}\cdot 11)$$

Then, the message $q_{\mathrm{B} \to \mathrm{A}}(\boldsymbol{\beta}[l])$ is updated so as to satisfy the moment matching conditions

$$\boldsymbol{\beta}_{\mathrm{B}}[l] = \frac{\sum_{\boldsymbol{\beta}[l]} \boldsymbol{\beta}[l] q_{\mathrm{A} \to \mathrm{B}}(\boldsymbol{\beta}[l]) q_{\mathrm{B} \to \mathrm{A}}^{\mathrm{new}}(\boldsymbol{\beta}[l])}{\sum_{\boldsymbol{\beta}[l]} q_{\mathrm{A} \to \mathrm{B}}(\boldsymbol{\beta}[l]) q_{\mathrm{B} \to \mathrm{A}}^{\mathrm{new}}(\boldsymbol{\beta}[l])}, \qquad (\mathrm{A}\cdot 12)$$

$$M \sum_{l=0}^{L-1} v_{\mathrm{B}}[l] = \sum_{l=0}^{L-1} \frac{\sum_{\boldsymbol{\beta}[l]} \|\boldsymbol{\beta}[l]\|^2 q_{\mathrm{A} \to \mathrm{B}}(\boldsymbol{\beta}[l]) q_{\mathrm{B} \to \mathrm{A}}^{\mathrm{new}}(\boldsymbol{\beta}[l])}{\sum_{\boldsymbol{\beta}[l]} q_{\mathrm{A} \to \mathrm{B}}(\boldsymbol{\beta}[l]) q_{\mathrm{B} \to \mathrm{A}}^{\mathrm{new}}(\boldsymbol{\beta}[l])}$$
$$- \sum_{l=0}^{L-1} \|\boldsymbol{\beta}_{\mathrm{B}}[l]\|^2. \qquad (\mathrm{A}\cdot 13)$$

The remaining derivation is similar to in [8], so that we omit it. The extrinsic distribution $q_{\mathrm{A} \to \mathrm{B}}$ in conventional EP [8] was defined element-wisely since i.i.d. signals were assumed. On the other hand, (A·8) has been defined for each section because $\boldsymbol{\beta}[l]$ has dependent elements. Here is the main difference in the derivations of conventional and proposed EP algorithms.