Expectation Propagation-Based Sampling Decoding: Enhancement and Optimization

Zheng Wang, *Member, IEEE*, Shanxiang Lyu, *Member, IEEE*, Yili Xia, *Member, IEEE*, and Qihui Wu, *Senior Member, IEEE*

***Abstract*—In this paper, the paradigm of expectation propagation (EP) algorithm in large-scale MIMO detection is extended by the sampling decoding in an Markov chain Monte Carlo way to boost the approximation of the target posterior distribution. The proposed EP-based sampling decoding scheme not only theoretically addresses the inherent convergence problem of EP, but also is able to achieve the near-optimal decoding performance with the increment of Markov moves. Specifically, the EP-based independent Metropolis-Hastings (MH) is proposed to guarantee the exponential convergence to the target posterior distribution, thus bridging the EP detector and the sampling decoding as a whole. Meanwhile, the output yielded by the EP detector also provides a good initial setup for the sampling decoding, which results in a better convergence performance in the approximation. To further improve the convergence performance and the decoding efficiency, the EP-based Gibbs sampling is given, where the choice of the standard deviation of the discrete Gaussian distribution in the Markov mixing is also studied for a better decoding performance. Moreover, we extend the proposed EP-based Gibbs sampling decoding to the soft-output decoding in MIMO bit-interleaved coded modulation (BICM) systems, which enjoys a flexible decoding trade-off between performance and complexity by the number of Markov moves.**

***Index Terms*—EP algorithm, sampling decoding, MIMO detection, soft-output decoding, lattice decoding, near-ML decoding.**

Manuscript received March 5, 2020; revised September 17, 2020; accepted November 5, 2020. Date of publication November 25, 2020; date of current version December 23, 2020. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Sangarapillai Lambotharan. This work was supported in part by the open research fund of National Mobile Communications Research Laboratory, Southeast University (No. 2019D04), in part by the open research fund of Key Laboratory of Dynamic Cognitive System of Electromagnetic Spectrum Space (Nanjing Univ. Aeronaut. Astronaut.), Ministry of Industry and Information Technology, Nanjing, 211106, China under Grant KF20181913, in part by the National Natural Science Foundation of China under Grants 61801216 and 61771124, in part by the Natural Science Foundation of Jiangsu Province under Grant BK20180420, and in part by the Zhi Shan Young Scholar Program of Southeast University. *(Corresponding author: Zheng Wang.)*

Zheng Wang is with the Key Laboratory of Dynamic Cognitive System of Electromagnetic Spectrum Space, Ministry of Industry and Information Technology, Nanjing University of Aeronautics and Astronautics, Nanjing 211106, China, and also with the National Mobile Communications Research Laboratory, School of Information Science and Engineering, Southeast University, Nanjing 210096, China (e-mail: z.wang@ieee.org).

Shanxiang Lyu is with the College of Cyber Security, Jinan University, Guangzhou 510632, China (e-mail: shanxianglyu@gmail.com).

Yili Xia is with the School of Information Science and Engineering, Southeast University, Nanjing 210096, China (e-mail: yili_xia@seu.edu.cn).

Qihui Wu is with the Key Laboratory of Dynamic Cognitive System of Electromagnetic Spectrum Space (Nanjing Univ. Aeronaut. Astronaut.), Ministry of Industry and Information Technology, Nanjing 211106, China (e-mail: wuqihui2014@sina.com).

Digital Object Identifier 10.1109/TSP.2020.3040047

I. Introduction

NOWADAYS, massive multiple-input multiple-output (MIMO) has been widely applied in 5 G to boost the network capacity on a much greater scale without extra bandwidth [1]–[4]. Besides the large-scale number of antennas at the base station, the total antenna number at the user side also increases dramatically due to the smaller size of antennas and rapid growth of user devices, which imposes a pressing challenge on the uplink signal detection of massive MIMO systems. To this end, a number of works have been made to achieve a better decoding trade-off between performance and complexity [5]–[9]. Nevertheless, the substantial performance gap still does exist especially when the system size goes up.

In [10], the expectation propagation (EP) technique was adopted into the large-scale MIMO detection, which fully takes the advantages of approximate inference over the posterior probability of the transmitted signal. As a generalization of belief propagation (BP) to construct tractable approximations [11], [12], EP shows better efficiency and robustness than BP and Gaussian tree approximation (GTA) [13], [14], and was further generalized by the expectation consistency (EC) technique under a free energy approximation framework [15], [16]. In [17], a low-complexity EP decoding scheme is given to reduce the computational burden of the matrix inversion at each iteration. In [18], EP is considered for MIMO detection under the case of imperfect channel state information (CSI). Furthermore, EP was applied to MIMO generalized frequency division multiplexing (GFDM) system to achieve the near-optimum detection [19]. Besides MIMO detection, EP has also been applied to research fields of low-density parity-check (LDPC) channel decoding [20], [21], Turbo equalization [22], [23] and so on. However, the convergence of EP is not guaranteed in theory, which has always been an inherent problem of EP since it was introduced. Although in a few special cases like the exponential family, the resulting KL-divergence is proven to have a stationary point, the iteration of EP still may not reach it [11], [24]. Even though it seems that EP works well in the scenario of MIMO detection, such a risk does always exist. Suffered from it, the performance gain of EP is always limited no matter how large size of the iteration number L is.

On the other hand, sampling decoding has emerged as a promising detection strategy especially for high-dimensional systems [25]–[29]. Typically, it performs decoding by sampling from a discrete multi-dimensional Gaussian distribution, where

the optimal decoding solution with the smallest Euclidean distance naturally entails the largest probability to be sampled. Most importantly, during such a problem transformation, sampling decoding introduces a new parameter into the decoding framework, i.e., the standard deviation $\sigma > 0$. Because of the unimodal distribution, it is encouraged to set a small σ for a large target sampling probability, thus resulting in an efficient decoding by sampling [30]. Therefore, the problem of sampling decoding chiefly lies on how to successfully sample from the target discrete Gaussian distribution, which is a rather difficult problem in sharp contrast to the case of continuous Gaussian density. Fortunately, the classic Markov chain Monte Carlo (MCMC) method has been demonstrated to approach it with accessible convergence rate [31]–[33]. Nevertheless, sampling decoding still severely suffers from the initial setup as it plays an important role upon the convergence of the approximation [34]–[36]. In this regard, EP detector serves as a good complement to sampling decoding by offering a well chosen starting setup. Most importantly, the posterior distribution sought by EP decoding is essentially the same with the discrete Gaussian distribution in sampling decoding, making it possible to unify them together as a whole.

In this paper, to improve the decoding performance of EP in large-scale MIMO detection, we propose to incorporate the sampling decoding into the EP detector, which leads to the proposed EP-based sampling decoding schemes. From it, several promising merits can be achieved. Firstly, different from the EP decoding, the convergence of the proposed EP-based sampling decoding is guaranteed in theory, which exponentially converges to the target posterior distribution in an Markov chain Monte Carlo (MCMC) way. Meanwhile, the usage of EP detector is also helpful to the following sampling decoding as it provides a better starting setup for the Markov mixing so that extra potential can be further exploited in both convergence and performance. Secondly, based on the convergence, an extra decoding performance gain can be obtained by simply increasing the number of Markov moves T in the sampling decoding. This essentially overcomes the problem of EP decoding in the sense that a near-optimal decoding performance become possible. Thirdly, the proposed EP-based sampling decoding enjoys a flexible decoding trade-off between performance and complexity, as the complexity of Gibbs sampling decoding $O(T \cdot n^2)$ can be freely adjusted by the Markov moves. Therefore, our work unifies EP detector and sampling decoding, and they are actually well complementary of each other for a better decoding trade-off. Moreover, to achieve a near-capacity performance over MIMO channels, the proposed EP-based sampling decoding is further applied to MIMO bit-interleaved coded modulation (BICM) systems for the iterative detection and decoding (IDD) [37]–[40].

The rest of this paper is organized as follows. Section II introduces the background of MIMO detection as well as the soft-output decoding in MIMO-BICM systems, and briefly reviews the basics of EP detector. In Section III, the proposed EP-based independent Metropolis-Hastings (MH) sampling decoding is presented, followed by the demonstration of the convergence as well as the related convergence analysis. In order to further exploit the decoding potential, the EP-based Gibbs sampling decoding is proposed in Section IV. Meanwhile, the reasonable choice of the standard deviation σ in the Markov mixing is also studied. Section V introduces the EP-based Gibbs sampling algorithm to the soft-output decoding in MIMO-BICM systems and in Section VI simulation results based on the massive MIMO detection are presented to illustrate the performance gain and the flexible decoding trade-off. Finally, Section VII concludes the paper.

Notation: Matrices and column vectors are denoted by upper and lowercase boldface letters, and the transpose, inverse, pseudoinverse of a matrix $\mathbf{H}$ by $\mathbf{H}^T, \mathbf{H}^{-1}$, and $\mathbf{H}^\dagger$, respectively. We use $\mathbf{h}_i$ for the ith column of the matrix $\mathbf{H}$, $\widehat{\mathbf{h}}_i$ for the i-th Gram-Schmidt vector of the matrix $\mathbf{H}$, and $h_{i,j}$ for the entry in the ith row and jth column of the matrix $\mathbf{H}$. The multivariate normal distribution of a random vector $\mathbf{y}$ is represented by $\mathcal{N}(\mathbf{y} : \mu, \mathbf{\Sigma})$, where μ denotes the mean vector and $\mathbf{\Sigma}$ represents the variance matrix. In addition, in this paper, the computational complexity is measured by the number of arithmetic operations (additions, multiplications, comparisons, etc.). $\Re(\cdot)$ indicate the real components while $\Im(\cdot)$ standards for the imaginary components. The operator $\|\cdot\|_{TV}$ represents the total variation distance in the measurement between two probability distributions.

II. Preliminaries

In this section, we introduce the background and mathematical tools needed to describe and analyze the following EP-based sampling decoding algorithms.

A. MIMO Detection

Consider the hard detection of an $n_t \times n_r$ ($n_r \geq n_t$) MIMO system. Let $\widetilde{\mathbf{s}} \in \mathbb{C}^n$ denote the transmitted signal, then the corresponding received signal $\widetilde{\mathbf{y}}$ is given by

$$\widetilde{\mathbf{y}} = \widetilde{\mathbf{H}}\widetilde{\mathbf{s}} + \widetilde{\mathbf{w}}. \tag{1}$$

Specifically, the i-th entry of the transmitted signal $\widetilde{\mathbf{s}}$, denoted as $\widetilde{s}_i$, is a modulation symbol taken independently from an M-QAM constellation with Gray mapping and $\widetilde{\mathbf{w}}$ denotes the noise vector with zero mean and variance $\widetilde{\sigma}_w^2$. Meanwhile, it is assumed a flat fading environment, where the channel matrix $\widetilde{\mathbf{H}} \in \mathbb{C}^{n_t \times n_r}$ contains uncorrelated complex Gaussian fading gains with unit variance and remains constant over each frame duration.

In general, the complex model shown in (1) can be simplified as

$$\begin{bmatrix} \Re(\widetilde{\mathbf{y}}) \\ \Im(\widetilde{\mathbf{y}}) \end{bmatrix} = \begin{bmatrix} \Re(\widetilde{\mathbf{H}}) & -\Im(\widetilde{\mathbf{H}}) \\ \Im(\widetilde{\mathbf{H}}) & \Re(\widetilde{\mathbf{H}}) \end{bmatrix} \begin{bmatrix} \Re(\widetilde{\mathbf{s}}) \\ \Im(\widetilde{\mathbf{s}}) \end{bmatrix} + \begin{bmatrix} \Re(\widetilde{\mathbf{w}}) \\ \Im(\widetilde{\mathbf{w}}) \end{bmatrix}, \tag{2}$$

which gives an equivalent $2n_t \times 2n_r$ real-valued MIMO system. Therefore, for the sake of notational simplicity, we consider the $n \times n$ real-valued MIMO system model

$$\mathbf{y} = \mathbf{H}\mathbf{s} + \mathbf{w} \tag{3}$$

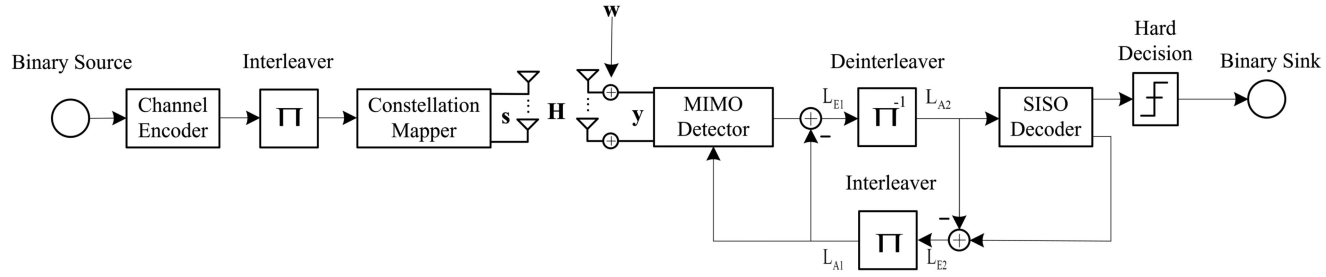


Fig. 1. Illustration of the BICM transmitter and IDD receiver in an MIMO system. The subscript "1" denotes processing blocks that are connected with the inner detection operation while subscript "2" indicates processing blocks connected to the outer decoding operations.

with Gaussian channel matrix $\mathbf{H} \in \mathbb{R}^{n\times n}$, transmitted signal $\mathbf{x} \in \mathcal{X}^n \subseteq \mathbb{Z}^n$ and noise variance $\sigma_w^2 = \widetilde{\sigma}_w^2/2$, where the detection extension of cases $n \times m$ with $n > m$ is straightforward [41]. Typically, given the system model in (3), the optimal maximum likelihood (ML) decoding in MIMO detection computes

$$\widehat{\mathbf{s}}_{\text{ml}} = \arg\min_{\mathbf{s}\in\mathcal{X}^n} \|\mathbf{Hs} - \mathbf{y}\|^2, \tag{4}$$

which essentially amounts to solving the closest vector problem (CVP) in lattice decoding [42]. More precisely, the problem of MIMO detection can be viewed as a special case of lattice decoding with finite state space $\mathbf{s}$ and Gaussian distributed $\mathbf{H}$ [43].

B. Soft-Output Decoding

In order to achieve the near-capacity performance over MIMO channels, bit-interleaved coded modulation (BICM) and iterative detection and decoding (IDD) were introduced, where the extrinsic information calculated by a priori probability (APP) detector is taken into account to produce the soft decisions [37]. Typically, as shown in Fig. 1, the extrinsic information L_{E1} is calculated by the MIMO detector based on the received $\mathbf{y}$ and *a priori* information (API) L_{A1} of the transmitted bits which is provided by the SISO decoder. Then L_{E1} is passed through the deinterleaver to become API L_{A2} to the SISO decoder, which computes the new extrinsic information L_{E2} to feed back to the MIMO detector. Clearly, one complete cycle of information exchange between the sections labeled "1" and "2" forms an iteration.

In particular, the extrinsic information in soft-output decoding is always calculated through the *posterior* log-likelihood ratio (LLR) for each information bit associated with the transmitted signal $\mathbf{s}$, i.e.,

$$L(b_i|\mathbf{y}) = \log\frac{P(b_i = 1|\mathbf{y})}{P(b_i = 0|\mathbf{y})} \tag{5}$$

where b_i is the i-th information bit in $\mathbf{s}$, $1 \le i \le nu$. Here, u represents the number of bits per real constellation symbol and $\mathbf{s}$ contains nu information bits in all. Through the exchange of extrinsic information in each iteration, the performance of soft-output decoding improves gradually and the posterior LLR follows

$$L(b_i|\mathbf{y}) = L_{\text{A}}(b_i) + \log\frac{\sum_{\mathbf{s}:b_i=1} P(\mathbf{y}|\mathbf{s})\cdot\exp\sum_{j\in\mathcal{J}_i} L_{\text{A}}(b_j)}{\sum_{\mathbf{s}:b_i=0} P(\mathbf{y}|\mathbf{s})\cdot\exp\sum_{j\in\mathcal{J}_i} L_{\text{A}}(b_j)} \tag{6}$$

where $L_{\text{A}}(b_i)$ denotes the API of each transmitted bit in $\mathbf{s}$

$$L_{\text{A}}(b_i) = \log\frac{P(b_i = 1)}{P(b_i = 0)} \tag{7}$$

and $\mathcal{J}_i$ is the set of indices j with

$$\mathcal{J}_i = \{j|j = 1, \ldots, nu, j \ne i\}. \tag{8}$$

In the absence of API, it is assumed that all the bits in $\mathbf{s}$ have the same probability to be 0 or 1 before $\mathbf{y}$ is observed. Then, the L-value in (5) becomes [37], [44]

$$L(b_i|\mathbf{y}) = \log\frac{\sum_{\mathbf{s}:b_i=1}\exp\left(-\frac{1}{2\sigma_w^2}\|\mathbf{Hs} - \mathbf{y}\|^2\right)}{\sum_{\mathbf{s}:b_i=0}\exp\left(-\frac{1}{2\sigma_w^2}\|\mathbf{Hs} - \mathbf{y}\|^2\right)}. \tag{9}$$

In principle, the straightforward way to calculate the L-value in (9) is MAP algorithm which computes the sums that contain 2^{nu} terms. Unfortunately, the exponentially increased complexity of MAP renders it inapplicable in practice. For this reason, a number of works were dedicated to calculate the L-value in an approximation way [45]–[47].

C. Expectation Propagation Detector

In [10], the expectation propagation (EP) technique from the field of approximate inference was introduced to MIMO detection. Specifically, according to the system model in (3), the posterior probability of the transmitted signal $\mathbf{s}$ given the received signal $\mathbf{y}$ satisfies

$$p(\mathbf{s}|\mathbf{y}) = \frac{p(\mathbf{y}|\mathbf{s})p(\mathbf{s})}{p(\mathbf{y})} \propto \mathcal{N}(\mathbf{y} : \mathbf{Hs}, \sigma_w^2\mathbf{I})\cdot\prod_{i=1}^{n}\mathbb{I}_{s_i\in\mathcal{X}} \tag{10}$$

by *Bayesian theorem*. Here s_i is the component of $\mathbf{s}$, $\mathbb{I}_{s_i\in\mathcal{X}}$ is the indicator function that takes value 1 if $s_i \in \mathcal{X}$ and 0 otherwise. However, it is intractable to directly perform inference over the posterior distribution $p(\mathbf{s}|\mathbf{y})$ in (10). To this end, the expectation propagation was adopted to approximate the intractable posterior distribution $p(\mathbf{s}|\mathbf{y})$ with a tractable distribution $q(\mathbf{s}|\mathbf{y})$, which minimizes the Kullback-Leibler (KL) divergence between them. In fact, this is essentially equivalent

to the moment matching conditions by matching the first and second moments of $p(\mathbf{s}|\mathbf{y})$ and $q(\mathbf{s}|\mathbf{y})$ with each other [10].

In particular, to construct a tractable approximation of $p(\mathbf{s}|\mathbf{y})$, $q(\mathbf{s}|\mathbf{y})$ is initially set as

$$q(\mathbf{s}|\mathbf{y}) = \mathcal{N}(\mathbf{y} : \mathbf{H}\mathbf{s}, \sigma_w^2\mathbf{I}) \cdot e^{-\frac{1}{2}\mathbf{s}^T\Lambda\mathbf{s}+\boldsymbol{\gamma}^T\mathbf{s}} \tag{11}$$

where $\boldsymbol{\gamma} = [\gamma_1, \gamma_2, \ldots, \gamma_n]^T$, and $\boldsymbol{\Lambda} = \text{diag}[\Lambda_1, \ldots, \Lambda_n]$ is a diagonal matrix. In this condition, the mean vector μ and the covariance matrix $\boldsymbol{\Sigma}$ of $q(\mathbf{s}|\mathbf{y})$ follow

$$\begin{aligned} \mu &= \Sigma(\sigma_w^{-2}\mathbf{H}^T\mathbf{y} + \gamma) \\ \Sigma &= (\sigma_w^{-2}\mathbf{H}^T\mathbf{H} + \Lambda)^{-1} \end{aligned} \tag{12}$$

respectively. Then, with the initial setting of $\gamma_i^0 = 0$ and $\Lambda_i^0 = E_s^{-1}$ (E_s standards for the mean symbol energy) for $1 \le i \le n$, the sequential EP detector in [10] alternatively approaches the moment matching between the marginals $p(s_i|\mathbf{y})$ and $q(s_i|\mathbf{y})$ for each i, so that each pair of $(\gamma_i^l, \Lambda_i^l)$ are updated independently along the iteration index $l \ge 0$. To summarize, given the iteration pair $(\gamma_i^l, \Lambda_i^l)$, the $(l+1)$-th iteration of the EP detector works in the following steps:

1) Update the mean μ_i^l and the variance $\sigma_i^{2(l)}$ of the marginal PDF of $q^l(s_i|\mathbf{y})$, where μ_i is the i-th element of μ^l and the variance $\sigma_i^{2(l)}$ equals to the i-th diagonal element of Σ^l.

2) Compute the mean t_i^l and the variance $h_i^{2(l)}$ of the cavity marginal $q_{\backslash i}^l(s_i|\mathbf{y})$ as

$$q_{\backslash i}^l(s_i|\mathbf{y}) = \frac{q^l(s_i|\mathbf{y})}{e^{-\frac{1}{2}\Lambda_i s_i^2 + \gamma_i s_i}}, \tag{13}$$

where

$$\begin{aligned} t_i^l &= h_i^{2(l)}\left(\frac{\mu_i^l}{\sigma_i^{2(l)}} - \gamma_i^l\right) \\ h_i^{2(l)} &= \frac{\sigma_i^{2(l)}}{(1 - \sigma_i^{2(l)}\Lambda_i^l)}. \end{aligned} \tag{14}$$

3) Compute the mean $\mu_{p_i}^l$ and the variance $\sigma_{p_i}^{2(l)}$ of the distribution

$$\widehat{p}^l(s_i|\mathbf{y}) \propto q_{\backslash i}^l(s_i|\mathbf{y}) \cdot \mathbb{I}_{s_i \in \mathcal{X}}. \tag{15}$$

4) Update the pair $(\gamma_i^{l+1}, \Lambda_i^{l+1})$ so that the mean and the variance of the following distribution

$$q_{\backslash i}^l(s_i|\mathbf{y}) \cdot e^{\gamma_i^{l+1}s_i - \frac{1}{2}\Lambda_i^{l+1}s_i^2} \tag{16}$$

match $\mu_{p_i}^l$ and $\sigma_{p_i}^{2(l)}$ respectively by setting

$$\begin{aligned} \Lambda_i^{l+1} &= \frac{1}{\sigma_{p_i}^{2(l)}} - \frac{1}{h_i^{2(l)}} \\ \gamma_i^{l+1} &= \frac{\mu_{p_i}^l}{\sigma_{p_i}^{2(l)}} - \frac{t_i^l}{h_i^{2(l)}}. \end{aligned} \tag{17}$$

Clearly, given Λ^l and γ^l, the mean vector μ^l and the covariance matrix Σ^l in (12) can be computed, which allows parallel updates of all pairs $(\gamma_i^{l+1}, \Lambda_i^{l+1})$. Finally, after a number of iterations (i.e., $l = L$), the components of decoding solution $\widehat{\mathbf{s}}$ in EP is determined by the following hard decision

$$\widehat{s}_i = \arg\min_{s_i \in \mathcal{X}} |s_i - \mu_i^L|^2. \tag{18}$$

Undoubtedly, with the initial setting γ_i^0 and Λ_i^0 for $1 \le i \le n$, the EP detector yields the decoding result in an MMSE sense.

III. EP-Based Sampling Decoding Algorithm

In this section, the concept of sampling decoding is introduced into the EP detector. Instead of outputting the decoding solution by hard decision, random sampling is applied to output the final solution. Besides the decoding gain, such an enhancement version of the EP detector is also convergence guaranteed in an MCMC way. Moreover, we not only show it converges exponentially but also try to specify the lower bound of the convergence rate for the tractable decoding. Most importantly, due to the convergence nature, the performance of the proposed EP-based sampling decoding can be improved by simply increasing the number of Markov moves.

Besides the posterior probability shown in (10), given the received signal $\mathbf{y}$, it is also eligible to establish the probability distribution with respect to the transmitted signal $\mathbf{s}$ as

$$\overline{p}(\mathbf{s}|\mathbf{y}) = \frac{e^{-\frac{1}{2\sigma_w^2}\|\mathbf{H}\mathbf{s}-\mathbf{y}\|^2}}{\sum_{\mathbf{s}\in\mathcal{X}^n} e^{-\frac{1}{2\sigma_w^2}\|\mathbf{H}\mathbf{s}-\mathbf{y}\|^2}}. \tag{19}$$

Clearly, it is straightforward to see that both the posterior distributions $p(\mathbf{s}|\mathbf{y})$ and $\overline{p}(\mathbf{s}|\mathbf{y})$ are essentially the same in the sense of decoding $\mathbf{s}$, i.e.,

$$\begin{aligned} \widehat{\mathbf{s}}_{\text{ml}} &= \arg\min_{\mathbf{s}\in\mathcal{X}^n} \|\mathbf{H}\mathbf{s} - \mathbf{y}\|^2 = \arg\max_{\mathbf{s}\in\mathcal{X}^n} p(\mathbf{s}|\mathbf{y}) \\ &= \arg\max_{\mathbf{s}\in\mathcal{X}^n} \overline{p}(\mathbf{s}|\mathbf{y}), \end{aligned} \tag{20}$$

where the optimal decoding solution naturally entails the largest sampling probability. Intuitively, if sampling from the posterior distribution $\overline{p}(\mathbf{s}|\mathbf{y})$ can be successfully performed, the optimal decoding solution $\widehat{\mathbf{s}}_{\text{ml}}$ will most likely to be encountered by multiple samplings. Therefore, after the approximation of the posterior probability $p(\mathbf{s}|\mathbf{y})$ in (10) by using $q(\mathbf{s}|\mathbf{y})$ with L iterations, sampling decoding can be applied to continue the approximation based on the mean vector μ^L and the covariance matrix Σ^L outputted by EP. As for the way to guarantee the convergence to the target distribution $\overline{p}(\mathbf{s}|\mathbf{y})$, we resort to the classic MCMC methods.

A. EP-Based Independent Metropolis-Hastings Sampling Decoding

We now introduce the proposed EP-based independent Metropolis-Hastings sampling decoding. Specifically, given the mean μ_i^L and the variance $\sigma_i^{2(L)}$ issued by EP, rather than outputting the decoding solution by direct rounding in (18), we

pick up the choice of s_i randomly as follows

$$\widehat{s}_i \sim p(s_i) = \frac{e^{-\frac{1}{2\sigma_i^{2(L)}}|s_i-\mu_i^L|^2}}{\sum_{s_i\in\mathcal{X}} e^{-\frac{1}{2\sigma_i^{2(L)}}|s_i-\mu_i^L|^2}}. \tag{21}$$

Furthermore, by counting the n-times 1-dimensional sampling from s_1 to s_n as one iteration, the sampling probability $p(\widehat{\mathbf{s}})$ can be expressed as

$$p(\widehat{\mathbf{s}}) = \prod_{i=1}^{n} p(\widehat{s}_i) = \frac{e^{-\frac{1}{2}\|\mathbf{D}(\mathbf{s}-\mu^L)\|^2}}{\sum_{\mathbf{s}\in\mathcal{X}^n} e^{-\frac{1}{2}\|\mathbf{D}(\mathbf{s}-\mu^L)\|^2}} \tag{22}$$

with $\mathbf{D} = \text{diag}(\Sigma^L)\cdot\mathbf{I}$. Clearly, with the initial setting of EP, i.e., $\Lambda_i^0 = E_s^{-1}$ and $\gamma_i^0 = 0$ for $1 \le i \le n$, the decoding solution given in (22) corresponds to the randomized variant of MMSE decoding algorithm [48]. As the iteration of EP proceeds, better choices of μ and Σ would be refined gradually, which offers a perspective initial setup for the random sampling in the following.

Given the sampling probability in (22), the following question is how to guarantee its approximation to $\overline{p}(\mathbf{s}|\mathbf{y})$. In fact, it has been demonstrated in [48] that $p(\widehat{\mathbf{s}})$ with $\Lambda_i^0 = E_s^{-1}$ and $\gamma_i^0 = 0$ approximates $\overline{p}(\mathbf{s}|\mathbf{y})$ within a negligible statistical distance if σ is sufficiently large. However, such a requirement is too stringent, rendering the direct approximation by $p(\widehat{\mathbf{s}})$ inapplicable in most cases of interest. In this condition, we attempt to perform the approximation by establishing a valid Markov chain converging to $\overline{p}(\mathbf{s}|\mathbf{y})$. Once the underlying Markov chain arrives at the stationary distribution, then the exact sampling from $\overline{p}(\mathbf{s}|\mathbf{y})$ can be carried out. In fact, when the convergence rate of the Markov chain is known, the approximation becomes tractable, since the total variation distance between the built distribution and the target distribution can be easily estimated.

Therefore, in order to built the Markov chain for the target distribution $\overline{p}(\mathbf{s}|\mathbf{y})$, the sampling scheme of the Metropolis-Hastings (MH) algorithm is adopted, where the sampling probability $p(\widehat{\mathbf{s}})$ in (22) is applied as the proposal distribution[1] $\underline{q}(\cdot,\cdot)$. To summarize, given the mean vector μ^L and the covariance matrix Σ^L outputted by EP after L iterations, the approximation of $\overline{p}(\mathbf{s}|\mathbf{y})$ induced by the designed Markov chain operates in the following steps:

1) Sample from the independent proposal distribution to obtain the candidate state $\mathbf{g} \in \mathcal{X}^n$ for the Markov move $\mathbf{S}_{t+1}$,

$$\begin{aligned}\underline{q}(\mathbf{s},\mathbf{g}) &= p(\mathbf{g}) \\ &= \frac{e^{-\frac{1}{2}\|\mathbf{D}(\mathbf{g}-\mu^L)\|^2}}{\sum_{\mathbf{g}\in\mathcal{X}^n} e^{-\frac{1}{2}\|\mathbf{D}(\mathbf{g}-\mu^L)\|^2}},\end{aligned} \tag{23}$$

where the sampling of $\mathbf{g}$ is independent of $\mathbf{s}$.

2) Calculate the acceptance ratio $\alpha(\mathbf{s},\mathbf{g})$

$$\alpha(\mathbf{s},\mathbf{g}) = \min\left\{1, \frac{\pi(\mathbf{g})\underline{q}(\mathbf{g},\mathbf{s})}{\overline{p}(\mathbf{s}|\mathbf{y})\underline{q}(\mathbf{s},\mathbf{g})}\right\} = \min\left\{1, \frac{\pi(\mathbf{g})p(\mathbf{s})}{\overline{p}(\mathbf{s}|\mathbf{y})p(\mathbf{g})}\right\}$$

[1]Theoretically, the proposal distribution $\underline{q}(\cdot,\cdot)$ in the MH algorithm can be any fixed distribution from which we can conveniently draw samples.

Algorithm 1: EP-Based Independent MH Sampling Decoding.

Require: $\mathbf{H}, \sigma_w, \mathbf{y}, L, T$;
Ensure: $\widehat{\mathbf{s}}_{\text{output}}$;
1: use EP detector to get μ^L and Σ^L after L iterations
2: let $\widehat{\mathbf{s}}$ in (18) denote the initial state of $\mathbf{S}_0$ and $\widehat{\mathbf{s}}_{\text{output}} = \widehat{\mathbf{s}}$
3: **for** $t = 1, \ldots, T$ **do**
4: sample $\mathbf{g}$ from the proposal distribution $\underline{q}(\mathbf{s},\mathbf{g})$ in (23)
5: calculate the acceptance ratio $\alpha(\mathbf{s},\mathbf{g})$ in (24)
6: generate a sample u from the uniform density $U[0,1]$
7: **if** $u \le \alpha(\mathbf{s},\mathbf{g})$ **then**
8: let $\mathbf{S}_t = \mathbf{g}$
9: **else**
10: $\mathbf{S}_t = \mathbf{s}$
11: **end if**
12: **if** $\|\mathbf{Hg}-\mathbf{y}\| < \|\mathbf{H}\widehat{\mathbf{s}}_{\text{output}}-\mathbf{y}\|$ **then**
13: update $\widehat{\mathbf{s}}_{\text{output}} = \mathbf{g}$
14: **end if**
15: **end for**

$$= \min\left\{1, \frac{e^{-\frac{1}{2\sigma_w^2}\|\mathbf{Hg}-\mathbf{y}\|^2}\cdot e^{-\frac{1}{2}\|\mathbf{D}(\mathbf{s}-\mu^L)\|^2}}{e^{-\frac{1}{2\sigma_w^2}\|\mathbf{Hs}-\mathbf{y}\|^2}\cdot e^{-\frac{1}{2}\|\mathbf{D}(\mathbf{g}-\mu^L)\|^2}}\right\} \tag{24}$$

with

$$\pi(\mathbf{s}) = \overline{p}(\mathbf{s}|\mathbf{y}) = \frac{e^{-\frac{1}{2\sigma_w^2}\|\mathbf{Hs}-\mathbf{y}\|^2}}{\sum_{\mathbf{x}\in\mathcal{X}^n} e^{-\frac{1}{2\sigma_w^2}\|\mathbf{Hs}-\mathbf{y}\|^2}}. \tag{25}$$

3) Make a decision for $\mathbf{S}_{t+1}$ based on $\alpha(\mathbf{s},\mathbf{g})$ to accept $\mathbf{S}_{t+1} = \mathbf{g}$ or not (i.e., $\mathbf{S}_{t+1} = \mathbf{s}$).

In this way, a Markov chain $\{\mathbf{S}_0, \mathbf{S}_1, \ldots\}$ is established with the transition probability $P(\mathbf{s},\mathbf{g})$ as follows:

$$P(\mathbf{s},\mathbf{g}) = \underline{q}(\mathbf{s},\mathbf{g})\cdot\alpha(\mathbf{s},\mathbf{g}) = \min\left\{p(\mathbf{g}), \frac{\pi(\mathbf{g})p(\mathbf{s})}{\overline{p}(\mathbf{s}|\mathbf{y})}\right\}. \tag{26}$$

Note that the generation of the state candidate $\mathbf{g}$ is independent of the previous one $\mathbf{s}$. This is referred to as the independent Metropolis-Hastings algorithm [49], where the connection between two consecutive Markov states only lies in the decision part. Note that the acceptance ratio $\alpha(\mathbf{s},\mathbf{g}) > 0$ can be further written as

$$e^{-\frac{1}{2}[\|\overline{\mathbf{H}}(\mathbf{g}-\mu_{\text{ZF}})\|^2-\|\overline{\mathbf{H}}(\mathbf{s}-\mu_{\text{ZF}})\|^2]-\frac{1}{2}[\|\mathbf{D}(\mathbf{s}-\mu^L)\|^2-\|\mathbf{D}(\mathbf{g}-\mu^L)\|^2]}, \tag{27}$$

where $\overline{\mathbf{H}} = \mathbf{H}/\sigma_w$, and $\mu_{\text{ZF}} = \mathbf{H}^\dagger\mathbf{y}$ denotes the mean vector yielded by zero-forcing (ZF) detector. Accordingly, μ^L can be viewed as the mean vector outputted by an enhanced MMSE detector. Therefore, once $\mathbf{g}$ is a better choice than $\mathbf{s}$ in the sense of Euclidean distance, its acceptance ratio will be improved remarkably due to the first term $\|\overline{\mathbf{H}}(\mathbf{g}-\mu_{\text{ZF}})\|^2 - \|\overline{\mathbf{H}}(\mathbf{s}-\mu_{\text{ZF}})\|^2$ in (27). To summarize, the operation of the proposed EP-based independent MH sampling decoding is presented in detail in Algorithm 1.

B. Convergence Analysis

In principle, it is easy to verify that the proposed Markov chain with transition probability given in (26) is *irreducible* by satisfying

$$P(\mathbf{S}^{t+k}=\mathbf{g}|\mathbf{S}^{t}=\mathbf{s})>0 \tag{28}$$

with Markov move index $t\geq 0$ and positive integer $k>0$, *aperiodic* by fulfilling

$$\gcd\{k: P(\mathbf{S}^{t+k}=\mathbf{g}|\mathbf{S}^{t}=\mathbf{s})>0\}=1 \tag{29}$$

where "gcd" represents the greatest common divisor, and *reversible* due to

$$\overline{p}(\mathbf{s}|\mathbf{y})P(\mathbf{S}^{t+1}=\mathbf{g}|\mathbf{S}^{t}=\mathbf{s})=\pi(\mathbf{g})P(\mathbf{S}^{t+1}=\mathbf{s}|\mathbf{S}^{t}=\mathbf{g}). \tag{30}$$

Therefore, according to the *convergence theorem* of MCMC, we arrive at the following Theorem to show its geometric ergodicity,[2] where the detailed proof can be found in [34, Theorem 4.9].

Theorem 1: Given the target discrete Gaussian distribution $\pi(\mathbf{s})=\overline{p}(\mathbf{s}|\mathbf{y})$, the Markov chain induced by the proposed EP-based independent MH sampling decoding satisfies

$$\|P^{t}(\mathbf{s},\cdot)-\pi\|_{TV}\leq M(1-\delta)^{t} \tag{31}$$

where $0<\delta<1$ and $M>0$ for all states $\mathbf{s}\in\mathcal{X}^{n}$.

Clearly, the convergence of the proposed EP-based sampling decoding to the target distribution $\overline{p}(\mathbf{s}|\mathbf{y})$ is guaranteed theoretically, which implies the optimal decoding solution $\widehat{\mathbf{s}}_{\mathrm{ml}}$ would be returned along with the Markov mixing. This is different from EP as its convergence for MIMO detection is not guaranteed [10], [11]. Most importantly, the distribution induced by the underlying Markov chain converges exponentially fast while the exponential decay coefficient $0<\delta<1$ is the key to determine the upper bound of the convergence rate $(1-\delta)$. Therefore, in what follows, we try to specify the exponential decay coefficient δ, so that the mixing time required by the Markov chain to converge becomes accessible.

Lemma 1: The exponential decay coefficient $0<\delta<1$ in the proposed EP-based independent MH sampling decoding follows

$$\delta\triangleq\min_{\mathbf{g}\in\mathcal{X}^{n}}\{e^{-\frac{1}{2}\|\mathbf{D}(\mathbf{g}-\mu^{L})\|^{2}+\frac{1}{2}\|\overline{\mathbf{H}}(\mathbf{g}-\mu_{\mathrm{ZF}})\|^{2}}\}\cdot\beta \tag{32}$$

with a constant

$$\beta=\frac{\sum_{\mathbf{g}\in\mathcal{X}^{n}}e^{-\frac{1}{2}\|\overline{\mathbf{H}}(\mathbf{g}-\mu_{\mathrm{ZF}})\|^{2}}}{\sum_{\mathbf{g}\in\mathcal{X}^{n}}e^{-\frac{1}{2}\|\mathbf{D}(\mathbf{g}-\mu^{L})\|^{2}}}. \tag{33}$$

Proof: To begin with, it has been demonstrated that for the independent MH algorithm, δ actually denotes the lower bound of the ratio $q(\cdot)/\pi(\cdot)$ [31]. Therefore, according to (22) and (25), we have

$$\frac{q(\mathbf{g})}{\pi(\mathbf{g})}=\frac{e^{-\frac{1}{2}\|\mathbf{D}(\mathbf{g}-\mu^{L})\|^{2}}}{\sum_{\mathbf{g}\in\mathcal{X}^{n}}e^{-\frac{1}{2}\|\mathbf{D}(\mathbf{g}-\mu^{L})\|^{2}}}\cdot\frac{\sum_{\mathbf{g}\in\mathcal{X}^{n}}e^{-\frac{1}{2}\|\mathbf{H}\mathbf{g}-\mathbf{y}\|^{2}}}{e^{-\frac{1}{2}\|\mathbf{H}\mathbf{g}-\mathbf{y}\|^{2}}}$$

$$\geq\min_{\mathbf{g}\in\mathcal{X}^{n}}\{e^{-\frac{1}{2}\|\mathbf{D}(\mathbf{g}-\mu^{L})\|^{2}+\frac{1}{2}\|\overline{\mathbf{H}}(\mathbf{g}-\mu_{\mathrm{ZF}})\|^{2}}\}\cdot\beta$$

$$=\delta, \tag{34}$$

completing the proof. ■

Therefore, given $\mathbf{D}$, $\mathbf{H}$, $\mathbf{y}$ and μ^{L}, δ can be exactly calculated, thus offering an upper bound for the convergence rate of the Markov chain. Another point worth being mentioned is that δ also depends on the size of state space $\mathcal{X}^{n}$, where a larger state space size naturally corresponds to a smaller δ. This is because δ itself only serves as a lower bound for all the possible Markov moves, and the true value in practice could be much larger than it. Consequently, based on δ, the upper bound of the mixing time can be further calculated as [31]

$$t_{\mathrm{mix}}(\epsilon)\leq(-\ln\epsilon)\cdot\left(\frac{1}{\delta}\right),\ \ \epsilon<1, \tag{35}$$

which leads to a tractable Markov mixing. Clearly, the mixing time is proportional to $1/\delta$, and becomes $O(1)$ as $\delta\to 1$. Hence, after a certain burn-in time, the sampling based on the approximation of $\overline{p}(\mathbf{s}|\mathbf{y})$ can be carried out, where the sample with the closest Euclidean distance among all the candidates after T times Markov moves is selected as the final decoding solution.

IV. Enhancement and Optimization

The flexible choice of the proposal distribution $\underline{q}(\cdot,\cdot)$ in MH is beneficial for the establishment of the valid Markov chain. However, its convergence turns out to be slow if the choice of $\underline{q}$ is not well suited. This is mainly due to the acceptance mechanism in MH as the sampling candidate from the proposal distribution could be rejected in a large probability. To this end, in this section, the EP-based Gibbs sampling is proposed to further exploit the decoding potential. Moreover, under the framework of sampling decoding, the standard deviation σ is optimized for a better decoding performance.

A. EP-Based Gibbs Sampling Decoding

As a special case of MH sampling, Gibbs sampling employs univariate conditional sampling to build the Markov chain [51]. Compared to the MH sampling, the Gibbs sampling from MCMC is able to further improve the decoding efficiency and performance in two perspectives. On one hand, the usage of the proposal distribution in the traditional MH sampling is removed, whose setting is flexible but appears difficult to find the optimal one. As the ideal proposal distribution for MH is unknown, the solution offered by Gibbs sampling to perform the 1-dimensional conditional sampling over the target distribution turns out to be more promising for a better convergence performance. On the other hand, Gibbs sampling is easy to implement, where the decision by acceptance ratio in the traditional MH sampling can be avoided (because of the acceptance ratio α in Gibbs sampling is always 1 [34]). This is rather beneficial for the sampling efficiency in most cases of interest.

In particular, in the proposed EP-based Gibbs sampling decoding algorithm, each coordinate of $\mathbf{s}$ is sampled from the

[2]Since the state space of the Markov chain is finite, geometric ergodicity and uniform ergodicity essentially are the same [50].

following 1-dimensional conditional distribution

$$p_{\text{gibbs}}(s_i|\mathbf{s}_{[-i]},\mathbf{y}) = \frac{e^{-\frac{1}{2\sigma_w^2}\|\mathbf{Hs}-\mathbf{y}\|^2}}{\sum_{s_i\in\mathcal{X}} e^{-\frac{1}{2\sigma_w^2}\|\mathbf{Hs}-\mathbf{y}\|^2}} \tag{36}$$

with $\mathbf{s}_{[-i]} \triangleq [s_1,\ldots,s_{i-1},s_{i+1},\ldots,s_n]^T$. During this univariate sampling over s_i, the other $n-1$ variables contained in $\mathbf{s}_{[-i]}$ are keeping fixed but are counted into the sampling as well. By repeating such a procedure with a certain order [51], a valid Markov chain $\{\mathbf{S}^0,\mathbf{S}^1,\ldots\}$ is established. More specifically, it is also straightforward to verify that the underlying Markov chain is irreducible and aperiodic so as to the following Theorem, whose proof is omitted due to its simplicity as well.

Theorem 2: Given the target discrete Gaussian distribution $\pi(\mathbf{s}) = \overline{p}(\mathbf{s}|\mathbf{y})$, the Markov chain induced by the proposed EP-based Gibbs sampling algorithm satisfies

$$\|P^t(\mathbf{s},\cdot) - \pi\|_{TV} \le \overline{M}(1-\overline{\delta})^t \tag{37}$$

where $0<\overline{\delta}<1$ and $\overline{M}>0$ for all states $\mathbf{s}\in\mathcal{X}^n$.

Clearly, the complexity of Gibbs sampling at each Markov move is easily accepted with $O(n^2)$. Due to this attraction, the complexity of each Markov move is often insignificant, whereas the number of Markov moves turns out to be more critical. Compared to the EP-based independent MH sampling decoding, the univariate sampling in EP-based Gibbs sampling decoding fully makes use of the correlation among the components of $\mathbf{s}$, which naturally leads to a better Markov mixing. Besides, different from the former who performs the sampling based on the mean vector μ^L and the variance $\text{diag}(\Sigma^L)$ yielded by EP, the proposed EP-based Gibbs sampling decoding only takes the decoding solution of EP detector (i.e., $\widehat{\mathbf{s}}$ in (18)) as the initial starting point $\mathbf{s}^0$.

In essence, the initial starting point $\mathbf{s}^0$ plays an important role in the Markov mixing. More precisely, for the *small set* $\{\mathbf{s}: V(\mathbf{s}) = \overline{p}(\mathbf{s}|\mathbf{y})^{-c} \le d, c>0\}$ and $d > 2b/(1-\lambda)$, the underlying Markov chain will converge exponentially as [52]

$$\|P^t(\mathbf{s}^0,\cdot) - \pi\|_{TV} \le (1-\delta')^{rt} + \left(\frac{U^r}{\alpha^{1-r}}\right)^t \left(1 + \frac{b}{1-\lambda} + V(\mathbf{s}^0)\right), \tag{38}$$

where $0<r<1$, $0<\lambda<1$, $U = 1+2(d+b)$ and $\alpha = \frac{1+d}{1+2b+\lambda d}$. From (38), starting the Markov chain with $\mathbf{s}^0$ as close to the center of the discrete Gaussian distribution (i.e., the optimal decoding solution $\widehat{\mathbf{s}}_{\text{ml}}$) as possible would be a judicious choice for the efficient convergence, which is accordance with our suggestion of using the decoding solution of EP detector. Since only the output of EP is utilized in the following Gibbs sampling, the proposed sampling decoding is well compatible to the most of EP detectors.

Compared to the EP-based independent MH sampling decoding, the convergence rate analysis of the proposed EP-based Gibbs sampling decoding is hard to carry out. Nevertheless, thanks to the univariate sampling over the target posterior distribution, the Markov chain built by Gibbs sampling still turns out to be a more attractive solution due to the better convergence and efficiency.

B. Choice of σ

Generally, the standard deviation $\sigma>0$ in the discrete Gaussian distribution $\overline{p}(\mathbf{s}|\mathbf{y})$ is set as σ_w by default due to the effect of noises [26], [27], [35], [53]. However, it is interesting to see that the setting of σ actually could be flexible as $\overline{p}(\mathbf{s}|\mathbf{y})$ is a unimodal distribution, which means the optimal decoding solution always entails the largest sampling probability no matter what $\sigma>0$ is, namely,

$$\begin{aligned}\widehat{\mathbf{s}}_{\text{ml}} &= \arg\max_{\mathbf{s}\in\mathcal{X}^n} \frac{e^{-\frac{1}{2\sigma_w^2}\|\mathbf{Hs}-\mathbf{y}\|^2}}{\sum_{\mathbf{s}\in\mathcal{X}^n} e^{-\frac{1}{2\sigma_w^2}\|\mathbf{Hs}-\mathbf{y}\|^2}} \\ &= \arg\max_{\mathbf{s}\in\mathcal{X}^n} \frac{e^{-\frac{1}{2\sigma^2}\|\mathbf{Hs}-\mathbf{y}\|^2}}{\sum_{\mathbf{s}\in\mathcal{X}^n} e^{-\frac{1}{2\sigma^2}\|\mathbf{Hs}-\mathbf{y}\|^2}}.\end{aligned} \tag{39}$$

In other words, this means an extra decoding degree of freedom can be obtained by adjusting σ for a better decoding performance. Intuitively, from (39), a small size σ is preferred as it corresponds to a large sampling probability of $\widehat{\mathbf{s}}_{\text{ml}}$. However, this also intensively increases the Markov mixing time since the chain becomes less dynamic [34]. On the contrary, it has been demonstrated in [54] that when σ is sufficiently large there is no need of MCMC for approximation as the sampling can be successfully carried out directly. Unfortunately, in this condition, the sampling probability of $\widehat{\mathbf{s}}_{\text{ml}}$ would be extremely small due to the near-uniform distribution. In a word, there is a latent trade-off regarding to the choice of σ, which should be carefully investigated. Here, to balance this inherent trade-off for a better decoding performance, a reasonable compromise is to ensure a reliable sampling probability $\overline{p}(\widehat{\mathbf{s}}_{\text{ml}}|\mathbf{y})$ given moderate σ. To this end, in the sequel we give guidelines on how to specify σ.

Typically, with respect to any $\mathbf{s}\in\mathcal{X}^n$ to be sampled, we firstly extract σ from the denominator of $\overline{p}(\mathbf{s}|\mathbf{y})$ by

$$\begin{aligned}\overline{p}(\mathbf{s}|\mathbf{y}) &= \frac{e^{-\frac{1}{2\sigma^2}\|\mathbf{Hs}-\mathbf{y}\|^2}}{\sum_{\mathbf{s}\in\mathcal{X}^n} e^{-\frac{1}{2\sigma^2}\|\mathbf{Hs}-\mathbf{y}\|^2}} \\ &> \frac{e^{-\frac{1}{2\sigma^2}\|\mathbf{Hs}-\mathbf{y}\|^2}}{\sum_{\mathbf{s}\in\mathbb{Z}^n} e^{-\frac{1}{2\sigma^2}\|\mathbf{Hs}-\mathbf{y}\|^2}} \\ &\overset{(a)}{\ge} \frac{e^{-\frac{1}{2\sigma^2}\|\mathbf{Hs}-\mathbf{y}\|^2}}{\sum_{\mathbf{s}\in\mathbb{Z}^n} e^{-\frac{1}{2\sigma^2}\|\mathbf{Hs}\|^2}} \\ &> \frac{e^{-\frac{1}{2\sigma^2}\|\mathbf{Hs}-\mathbf{y}\|^2}}{\int_{\mathbf{s}\in\mathbb{R}^n} e^{-\frac{1}{2\sigma^2}\|\mathbf{Hs}\|^2} d\mathbf{s}} \\ &\overset{(b)}{=} \frac{e^{-\frac{1}{2\sigma^2}\|\mathbf{Hs}-\mathbf{y}\|^2}}{\int_{\mathbf{s}\in\mathbb{R}^n} \rho_{\sqrt{2\pi\sigma^2\mathbf{H}^{-1}(\mathbf{H}^{-1})^T}}(\mathbf{s}) d\mathbf{s}} \\ &\overset{(c)}{=} \frac{e^{-\frac{1}{2\sigma^2}\|\mathbf{Hs}-\mathbf{y}\|^2}}{(\sqrt{2\pi}\sigma)^n \sqrt{\det(\mathbf{H}^{-1}(\mathbf{H}^{-1})^T)}} \\ &= f(\sigma)\cdot\eta\end{aligned} \tag{40}$$

where

$$\eta \triangleq 1/\sqrt{(2\pi)^n \det(\mathbf{H}^{-1}(\mathbf{H}^{-1})^T)} \tag{41}$$

is a positive constant and

$$f(\sigma) \triangleq \frac{e^{-\frac{1}{2\sigma^2}\|\mathbf{Hs}-\mathbf{y}\|^2}}{\sigma^n} \tag{42}$$

is parameterized by σ. Here, (a), (b) and (c) respectively obey the facts from lattice theory ([55, Lemma 2.5]) that

$$\sum_{\mathbf{x}\in\mathbb{Z}^n} e^{-\frac{1}{2\sigma^2}\|\mathbf{Bx}-\mathbf{c}\|^2} \leq \sum_{\mathbf{x}\in\mathbb{Z}^n} e^{-\frac{1}{2\sigma^2}\|\mathbf{Bx}\|^2} \tag{43}$$

for lattice $\mathcal{L} = \mathbf{Bx}$, $\mathbf{B} \in \mathbb{R}^{n\times n}$, $\mathbf{x} \in \mathbb{Z}^n$, and $\rho_{\mathbf{B}^{-1}}(\mathbf{x}) = e^{-\pi\|\mathbf{Bx}\|^2}$ for $\int_{\mathbb{R}^n} \rho_{\sqrt{\Sigma}}(\mathbf{x})d\mathbf{x} = \sqrt{\det\Sigma}$ with $\Sigma = \mathbf{BB}^T$.

From (40), it is natural to see that the sampling probability for any specific $\mathbf{s} \in \mathcal{X}^n$ is actually lower bounded by the function $f(\sigma)$. Furthermore, let the derivative of function $f(\sigma)$ with respect to σ be zero, we have

$$\sigma = \frac{\|\mathbf{Hs}-\mathbf{y}\|}{\sqrt{n}}, \tag{44}$$

which means the sampling probability lower bound of any specific $\mathbf{s} \in \mathcal{X}^n$ can be optimized through it. Therefore, to obtain the target decoding solution $\widehat{\mathbf{s}}_{\text{ml}}$, the corresponding choice of σ_{ml} should vary with $\|\mathbf{H}\widehat{\mathbf{s}}_{\text{ml}} - \mathbf{y}\|$, i.e.,

$$\sigma_{\text{ml}} = \frac{\|\mathbf{H}\widehat{\mathbf{s}}_{\text{ml}} - \mathbf{y}\|}{\sqrt{n}}. \tag{45}$$

Generally speaking, given different configurations of $\mathbf{H}$ and $\mathbf{w}$, such a flexible setting of σ_{ml} is more beneficial to sampling decoding by providing a specific rather than statistical choice[3] σ_w [27]. For a small value of $\|\mathbf{H}\widehat{\mathbf{s}}_{\text{ml}} - \mathbf{y}\|$, σ_{ml} tends to get smaller since $\mathbf{y}$ appears close to $\mathbf{H}\widehat{\mathbf{s}}_{\text{ml}}$ and vice versa, thus adaptively guiding the choice of σ for each specific $\widehat{\mathbf{s}}_{\text{ml}}$. However, in practice, it is impossible to get $\widehat{\mathbf{s}}_{\text{ml}}$ for choosing σ_{ml} in (45), which imposes a significant request on $\mathbf{s}^0$ as an approximation

$$\sigma_{\text{approximate}} = \frac{\|\mathbf{Hs}^0 - \mathbf{y}\|}{\sqrt{n}}. \tag{46}$$

Intuitively, the closer of $\mathbf{s}^0$ to $\widehat{\mathbf{s}}_{\text{ml}}$, the more accurate of the selected σ, which is also in line with the requirement of Markov mixing. Therefore, the initial Markov state $\mathbf{s}^0$ outputted by the EP detector can be applied as a high-quality approximation. Moreover, it is also feasible to follow a dynamic strategy for updating σ, namely, update σ when a smaller size $\|\mathbf{Hs}^j - \mathbf{y}\|$ is obtained:

$$\sigma_{\text{dynamic}} \triangleq \frac{\min_{t\geq j\geq 0}\|\mathbf{Hs}^j - \mathbf{y}\|}{\sqrt{n}}. \tag{47}$$

Hence, it is clear to see that the Euclidean distance $\min_{t\geq j\geq 0}\|\mathbf{Hs}^j - \mathbf{y}\|$ shrinks monotonically along with the Markov moves, which leads to a more accurate approximation of the optimal σ. Note that updating σ dynamically is compatible with the mechanism of MCMC, which is known as adaptive MCMC [56]. Meanwhile, this is also accordance with *simulated annealing* (SA) by gradually cooling down the temperature of the Markov chain, which is widely accepted in various research fields [57].

[3]The common choice σ_w also severely suffers from the stalling problem as shrinks intensively with the increase of SNR.

Algorithm 2: EP-Based Gibbs Sampling Decoding.

Input: $\mathbf{H}, \sigma_{\text{dynamic}}, \mathbf{y}, L, T$;
Output: $\widehat{\mathbf{s}}_{\text{output}}$;
use EP detector to get $\widehat{\mathbf{s}}$ in (18) and let $\widehat{\mathbf{s}}_{\text{output}} = \widehat{\mathbf{s}}$
let $\widehat{\mathbf{s}}$ denote the initial state of $\mathbf{S}_0$ and let $\sigma = \frac{\|\mathbf{H}\widehat{\mathbf{s}}-\mathbf{y}\|}{\sqrt{n}}$
for t =1, ..., T **do**
  for i =n, ..., 1 **do**
    sample s_i^t from $p_{\text{gibbs}}(s_i|\mathbf{s}_{[-i]}, \mathbf{y})$ shown in (36)
  end for
  if $\|\mathbf{Hs}^t - \mathbf{y}\| < \|\mathbf{Hs}_{\text{output}} - \mathbf{y}\|$ **then**
    update $\widehat{\mathbf{s}}_{\text{output}} = \mathbf{s}^t$ and $\sigma = \frac{\|\mathbf{H}\widehat{\mathbf{s}}_{\text{output}}-\mathbf{y}\|}{\sqrt{n}}$
  end if
end for

Here, because the state space of $\mathcal{X}$ is formed by the finite and discontinuous integers rather than the integer space $\mathbb{Z}$, necessary adjustment has to be made in practice to ensure the sampling performance. Typically, because the bias brought by the usage of $\mathbf{s}^j$, the sampling has the risk to be uniform especially for the limited state space $\mathcal{X}$. Therefore, to avoid the related performance degradation, an upper bound is helpful to the choice of σ_{dynamic} in practice. To summarize, the operation of the proposed EP-based Gibbs sampling decoding with dynamic choice of σ is illustrated in Algorithm 2. Additionally, we point that the application of Gibbs sampling also introduces some extra degrees of freedom in exploited the decoding potential (i.e., the scan order of the Gibbs sampling, parallel sampling and so on), which will be the future work in our research.

V. Extension to Soft-Output Decoding in MIMO-BICM Systems

As shown in (9), different from MIMO detection, the optimal MAP decoding in soft-output detection takes all the possible candidates $\mathbf{s} \in \mathcal{X}^n$ in the state space into account for the LLR calculation, which is inapplicable in practice. Therefore, for the sake of complexity consideration, in this section we introduce the proposed EP-based Gibbs sampling decoding to the MIMO-BICM systems for the iterative detection and decoding. Typically, we show that by collecting the samples from the discrete Gaussian distribution

$$\overline{p}(\mathbf{s}|\mathbf{y}) = \frac{e^{-\frac{1}{2\sigma^2}\|\mathbf{Hs}-\mathbf{y}\|^2}}{\sum_{\mathbf{s}\in\mathcal{X}^n} e^{-\frac{1}{2\sigma^2}\|\mathbf{Hs}-\mathbf{y}\|^2}} \tag{48}$$

with $\sigma = \|\mathbf{H}\widehat{\mathbf{s}}_{\text{ml}} - \mathbf{y}\|/\sqrt{n}$, the proposed EP-based Gibbs sampling decoding achieves a flexible trade-off between performance and complexity, and near-MAP decoding performance can be achieved based on a restricted decoding set

$$\mathcal{C} = \{\mathbf{s} \in \mathcal{X}^n : \|\mathbf{Hs} - \mathbf{y}\| \leq \sqrt{2\pi}\|\mathbf{H}\widehat{\mathbf{s}}_{\text{ml}} - \mathbf{y}\|\}. \tag{49}$$

We point out that the decoding output of EP detector $\mathbf{s}^0$ is still applied here to serve as the initial choice of σ by approximating $d(\Lambda, \mathbf{y})$. Similarly, σ can also be updated in a dynamic way like (47). Note that decoding based on the derived decoding set $\mathcal{C}$ is in the same spirit as the list sphere decoding proposed in [37], but here we give an explicit size of the sphere radius to achieve the near-MAP performance.

To concisely state the analysis, we extend the discrete Gaussian distribution in (48) into a more general expression as

$$D_{\mathbf{H},\sigma,\mathbf{y}}(\mathbf{s}) = \frac{e^{-\frac{1}{2\sigma^2}\|\mathbf{Hs}-\mathbf{y}\|^2}}{\sum_{\mathbf{s}\in\mathbb{Z}^n} e^{-\frac{1}{2\sigma^2}\|\mathbf{Hs}-\mathbf{y}\|^2}}, \tag{50}$$

which is known as *lattice Gaussian distribution* in literature due to $\mathbf{s} \in \mathbb{Z}^n$ [54], [58]. Therefore, we will take $\mathbb{Z}^n$ as the state space of $\mathbf{s}$ in the following while each specific $\mathbf{Hs}$ in referred to as lattice point belonging to the lattice $\Lambda = \mathbf{Hs} \subseteq \mathbb{R}^n$. On the other hand, because $\|\mathbf{H}\widehat{\mathbf{s}}_{\text{ml}} - \mathbf{y}\|$ indicates the Euclidean distance between the query point $\mathbf{y}$ and lattice Λ, we use

$$d(\Lambda, \mathbf{y}) \triangleq \|\mathbf{H}\widehat{\mathbf{s}}_{\text{ml}} - \mathbf{y}\| \tag{51}$$

to represent it for short.

A. Sphere Radius for Near-MAP Decoding Performance

First of all, according to (43), the summation term $\sum_{\mathbf{s}\in\mathbb{Z}^n} e^{-\frac{1}{2\sigma_w^2}\|\mathbf{Hs}-\mathbf{y}\|^2}$ in (9) for the calculation of LLR is upper bounded by

$$\sum_{\mathbf{s}\in\mathbb{Z}^n} e^{-\frac{1}{2\sigma_w^2}\|\mathbf{Hs}-\mathbf{y}\|^2} \le \sum_{\mathbf{s}\in\mathbb{Z}^n} e^{-\frac{1}{2\sigma_w^2}\|\mathbf{Hs}\|^2}. \tag{52}$$

Next, because of the following inequality [59]

$$\sum_{\mathbf{x}\in\mathbb{Z}^n} e^{-\pi t\|\mathbf{Bx}\|^2} \le t^{-\frac{n}{2}} \sum_{\mathbf{x}\in\mathbb{Z}^n} e^{-\pi\|\mathbf{Bx}\|^2} \tag{53}$$

for $0 < t < 1$, it follows that

$$\sum_{\mathbf{s}\in\mathbb{Z}^n} e^{-\frac{1}{2\sigma_w^2}\|\mathbf{Hs}-\mathbf{y}\|^2} \le (2\pi\sigma_w^2)^{n/2} \cdot \sum_{\mathbf{s}\in\mathbb{Z}^n} e^{-\pi\|\mathbf{Hs}\|^2}, \tag{54}$$

for $\sigma_w \ge 1/\sqrt{2\pi}$.[4] Meanwhile, due to the following relationship [60]

$$\sum_{\mathbf{x}\in\mathbb{Z}^n} e^{-\frac{1}{2\sigma^2}\|\mathbf{Bx}-\mathbf{c}\|^2} \ge e^{-\frac{d^2(\mathcal{L},\mathbf{c})}{2\sigma^2}} \cdot \sum_{\mathbf{x}\in\mathbb{Z}^n} e^{-\frac{1}{2\sigma^2}\|\mathbf{Bx}\|^2}, \tag{55}$$

we have

$$\sum_{\mathbf{s}\in\mathbb{Z}^n} e^{-\frac{1}{2\sigma_w^2}\|\mathbf{Hs}-\mathbf{y}\|^2} \le (2\pi\sigma_w^2)^{\frac{n}{2}} \cdot e^{\pi d^2(\Lambda,\mathbf{y})} \cdot \sum_{\mathbf{s}\in\mathbb{Z}^n} e^{-\pi\|\mathbf{Hs}-\mathbf{y}\|^2} \tag{56}$$

for $\sigma_w \ge 1/\sqrt{2\pi}$.

On the other hand, according to (55), it is easy to verify that

$$\begin{aligned}
\sum_{\mathbf{s}\in\mathbb{Z}^n} e^{-\frac{1}{2\sigma_w^2}\|\mathbf{Hs}-\mathbf{y}\|^2} &\ge e^{-\frac{d^2(\Lambda,\mathbf{y})}{2\sigma_w^2}} \cdot \sum_{\mathbf{s}\in\mathbb{Z}^n} e^{-\frac{1}{2\sigma_w^2}\|\mathbf{Hs}\|^2} \\
&= e^{-\frac{d^2(\Lambda,\mathbf{y})}{2\sigma_w^2}} \cdot \sum_{\mathbf{s}\in\mathbb{Z}^n} e^{-\frac{\pi}{2\pi\sigma_w^2}\|\mathbf{Hs}\|^2} \\
&\ge e^{-\frac{d^2(\Lambda,\mathbf{y})}{2\sigma_w^2}} \cdot \sum_{\mathbf{s}\in\mathbb{Z}^n} e^{-\pi\|\mathbf{Hs}\|^2} \\
&\ge e^{-\frac{d^2(\Lambda,\mathbf{y})}{2\sigma_w^2}} \cdot \sum_{\mathbf{s}\in\mathbb{Z}^n} e^{-\pi\|\mathbf{Hs}-\mathbf{y}\|^2}
\end{aligned} \tag{57}$$

for $\sigma_w \ge 1/\sqrt{2\pi}$. Therefore, by combining (56) and (57), it is clear to see that for $\sigma_w \ge 1/\sqrt{2\pi}$, the sums $\sum_{\mathbf{s}\in\mathbb{Z}^n} e^{-\frac{1}{2\sigma_w^2}\|\mathbf{Hs}-\mathbf{y}\|^2}$ can be characterized by the term $\sum_{\mathbf{s}\in\mathbb{Z}^n} e^{-\pi\|\mathbf{Hs}-\mathbf{y}\|^2}$ and a function made up by σ_w and $d(\Lambda, \mathbf{y})$:

$$\sum_{\mathbf{s}\in\mathbb{Z}^n} e^{-\frac{1}{2\sigma_w^2}\|\mathbf{Hs}-\mathbf{y}\|^2} = g(\sigma_w, d(\Lambda,\mathbf{y})) \cdot \sum_{\mathbf{s}\in\mathbb{Z}^n} e^{-\pi\|\mathbf{Hs}-\mathbf{y}\|^2}, \tag{58}$$

where function $g(\cdot)$ is bounded by

$$(2\pi\sigma_w^2)^{n/2} \cdot e^{\pi d^2(\Lambda,\mathbf{y})} \ge g(\sigma_w, d(\Lambda,\mathbf{y})) \ge e^{-\frac{d^2(\Lambda,\mathbf{y})}{2\sigma_w^2}}. \tag{59}$$

Therefore, in what follows, we will show that the term $\sum_{\mathbf{s}\in\mathbb{Z}^n} e^{-\pi\|\mathbf{Hs}-\mathbf{y}\|^2}$ in (58) is mainly determined by the lattice points within the sphere radius $\sqrt{2\pi}d(\Lambda, \mathbf{y})$ centered at $\mathbf{y}$.

In particular, with $\sigma_w \ge 1/\sqrt{2\pi}$, it follows that

$$\begin{aligned}
\sum_{\mathbf{s}\in\mathbb{Z}^n} e^{-\frac{1}{2\sigma_w^2}\|\mathbf{Hs}-\mathbf{y}\|^2} &= \sum_{\mathbf{s}\in\mathbb{Z}^n} e^{-\pi\cdot\frac{1}{2\pi\sigma_w^2}\|\mathbf{Hs}-\mathbf{y}\|^2} \\
&= \sum_{\mathbf{s}\in\mathbb{Z}^n} e^{\pi(1-\frac{1}{2\pi\sigma_w^2})\|\mathbf{Hs}-\mathbf{y}\|^2} e^{-\pi\|\mathbf{Hs}-\mathbf{y}\|^2} \\
&> \sum_{\mathbf{s}\in\mathbb{Z}^n, \|\mathbf{Hs}-\mathbf{y}\|\ge\sqrt{2\pi}d(\Lambda,\mathbf{y})} e^{\pi(1-\frac{1}{2\pi\sigma_w^2})\|\mathbf{Hs}-\mathbf{y}\|^2} e^{-\pi\|\mathbf{Hs}-\mathbf{y}\|^2} \\
&> e^{(1-\frac{1}{2\pi\sigma_w^2})2\pi^2\, d^2(\Lambda,\mathbf{y})} \cdot \sum_{\mathbf{s}\in\mathbb{Z}^n, \|\mathbf{Hs}-\mathbf{y}\|\ge\sqrt{2\pi}d(\Lambda,\mathbf{y})} e^{-\pi\|\mathbf{Hs}-\mathbf{y}\|^2}.
\end{aligned} \tag{60}$$

Hence, according to (56) and (60), we have

$$\sum_{\mathbf{s}\in\mathbb{Z}^n} e^{-\pi\|\mathbf{Hs}-\mathbf{y}\|^2} > \frac{e^{(1-\frac{1}{2\pi\sigma_w^2})2\pi^2\, d^2(\Lambda,\mathbf{y})}}{(2\pi\sigma_w^2)^{\frac{n}{2}} \cdot e^{\pi d^2(\Lambda,\mathbf{y})}} \sum_{\mathbf{s}\in\mathbb{Z}^n, \|\mathbf{Hs}-\mathbf{y}\|\ge\sqrt{2\pi}d(\Lambda,\mathbf{y})} e^{-\pi\|\mathbf{Hs}-\mathbf{y}\|^2}, \tag{61}$$

so as to

$$\sum_{\mathbf{s}\in\mathbb{Z}^n, \|\mathbf{Hs}-\mathbf{y}\|<\sqrt{2\pi}d(\Lambda,\mathbf{y})} e^{-\pi\|\mathbf{Hs}-\mathbf{y}\|^2} > [k(\sigma_w) - 1] \cdot \sum_{\mathbf{s}\in\mathbb{Z}^n, \|\mathbf{Hs}-\mathbf{y}\|\ge\sqrt{2\pi}d(\Lambda,\mathbf{y})} e^{-\pi\|\mathbf{Hs}-\mathbf{y}\|^2}, \tag{62}$$

where

$$k(\sigma_w) = (2\pi\sigma_w^2)^{-\frac{n}{2}} \cdot e^{(2\pi^2 - \frac{\pi}{\sigma_w^2} - \pi)d^2(\Lambda,\mathbf{y})}. \tag{63}$$

Therefore, if the function $k(\sigma_w)$ is sufficiently large, the value of $\sum_{\mathbf{s}\in\mathbb{Z}^n} e^{-\pi\|\mathbf{Hs}-\mathbf{y}\|^2}$ is dominantly decided by the lattice points within $\|\mathbf{Hs} - \mathbf{y}\| < \sqrt{2\pi}d(\Lambda, \mathbf{y})$.

Here, we resort to statistics with respect to $d(\Lambda, \mathbf{y})$ to illustrate the relationship behind $k(\sigma_w)$. Typically, because $\mathbf{w}$ in (5) entails the additive white Gaussian noise (AWGN) with zero mean and variance σ_w^2, the expectation of $\|\mathbf{Hs} - \mathbf{y}\|^2$ satisfies

$$E[\|\mathbf{Hs} - \mathbf{y}\|^2] = n\sigma_w^2 \tag{64}$$

[4]This requirement of σ_w is easy to fulfill in practice for the soft-output decoding of MIMO systems especially for high-dimensional systems.

by the law of large numbers. Hence, we have

$$E[k(\sigma_w)] = \left(\frac{e^{(2\pi^2-\pi)\sigma_w^2-\pi}}{\sqrt{2\pi}\sigma_w} \right)^n. \quad (65)$$

Clearly, given $\sigma_w \geq 1/\sqrt{2\pi}$, the average value of $k(\sigma_w)$ increases exponentially with the system dimension n (e.g., with $\sigma_w^2 = 0.3$, $E[k] = 4.577^n$), making the lattice points within sphere radius $\|\mathbf{Hs}-\mathbf{y}\| < \sqrt{2\pi}d(\Lambda,\mathbf{y})$ dominant in the sums $\sum_{\mathbf{s}\in\mathbb{Z}^n} e^{-\pi\|\mathbf{Hs}-\mathbf{y}\|^2}$. In fact, we have to admit that the bound in (60) is rather loose while the true value of $E[k]$ could be much larger than that of (65).

From the relationship shown in (62), the sums in (58) can be well approximated by

$$\sum_{\mathbf{s}\in\mathbb{Z}^n} e^{-\frac{1}{2\sigma_w^2}\|\mathbf{Hs}-\mathbf{y}\|^2} \approx g(\sigma_w, d(\Lambda,\mathbf{y})) \sum_{\mathbf{s}\in\mathbb{Z}^n, \|\mathbf{Hs}-\mathbf{y}\|<\sqrt{2\pi}d(\Lambda,\mathbf{y})} e^{-\pi\|\mathbf{Hs}-\mathbf{y}\|^2}, \quad (66)$$

which indicates that a near-MAP decoding situation is enabled by those lattice points within the sphere radius $\|\mathbf{Hs}-\mathbf{y}\| < \sqrt{2\pi}d(\Lambda,\mathbf{y})$ centered at $\mathbf{y}$.

B. Flexible Trade-Off in Soft-Output Decoding

We now show that the proposed EP-based Gibbs sampling decoding with $\sigma = d(\Lambda,\mathbf{y})/\sqrt{n}$ enjoys a flexible decoding trade-off by efficiently collecting lattice points within the sphere radius $\|\mathbf{Hs}-\mathbf{y}\| < \sqrt{2\pi}d(\Lambda,\mathbf{y})$ centered at $\mathbf{y}$.

Theorem 3: With the choice $\sigma = d(\Lambda,\mathbf{y})/\sqrt{n}$, the collected lattice points $\mathbf{Hs}$ by the proposed EP-based Gibbs sampling decoding satisfy

$$P_{\mathbf{s}\sim D_{\mathbf{H},\sigma,\mathbf{y}}}[\|\mathbf{Hs}-\mathbf{y}\|<\sqrt{2\pi}rd(\Lambda,\mathbf{y})] \geq 1-2^{-\Omega(n)} \quad (67)$$

for $r \geq 1/\sqrt{2\pi}$.

Proof: To start with, we invoke the following Lemma from [61].

Lemma 2 ([61]): For any lattice $\mathcal{L} = \mathbf{Bx} \subset \mathbb{R}^n$, $\sigma > 0$, $\mathbf{c} \in \mathbb{R}^n$ and $r \geq 1/\sqrt{2\pi}$, it follows that

$$P_{\mathbf{x}\sim D_{\mathbf{B},\sigma,\mathbf{c}}}[\|\mathbf{Bx}-\mathbf{c}\| \geq r\sqrt{2\pi n}\sigma] < \frac{\rho_\sigma(\mathcal{L})}{\rho_\sigma(\mathcal{L}-\mathbf{c})}(\sqrt{2\pi e r^2}e^{-\pi r^2})^n, \quad (68)$$

where $\mathbf{x} \in \mathbb{Z}^n$ is sampled from the lattice Gaussian distribution $D_{\mathbf{B},\sigma,\mathbf{c}}$, and $\rho_\sigma(\mathcal{L}-\mathbf{c}) = \sum_{\mathbf{x}\in\mathbb{Z}^n} e^{-\frac{1}{2\sigma^2}\|\mathbf{Bx}-\mathbf{c}\|^2}$.

According to Lemma 2, given $\sigma = d(\Lambda,\mathbf{y})/\sqrt{n}$, we have

$$\begin{aligned} P_{\mathbf{s}\sim D_{\mathbf{H},\sigma,\mathbf{y}}}[\|\mathbf{Hs}-\mathbf{y}\| \geq \sqrt{2\pi}rd(\Lambda,\mathbf{y})] &< (\sqrt{2\pi}r \cdot e^{1-\pi r^2})^n \\ &= 2^{-\Omega(n)}, \end{aligned} \quad (69)$$

and the inequality holds due to (55), completing the proof. ■

Based on Theorem 3, it is straightforward to observe that with $r = 1$, the probability of lattice points sampled from $D_{\mathbf{H},\sigma,\mathbf{y}}$ locating within the sphere radius $\sqrt{2\pi}d(\Lambda,\mathbf{y})$ centered at $\mathbf{y}$ is lower bounded by

$$P_{\mathbf{s}\sim D_{\mathbf{H},\sigma,\mathbf{y}}}[\|\mathbf{Hs}-\mathbf{y}\|<\sqrt{2\pi}d(\Lambda,\mathbf{y})] \geq 1-0.2945^n, \quad (70)$$

which is almost 1 especially in high-dimensional systems. On the other hand, since the lattice points closer to the center point $\mathbf{y}$ naturally have larger probabilities to be sampled, the proposed EP-based Gibbs sampling decoding is able to efficiently collect those high-quality lattice points from $D_{\mathbf{H},\sigma,\mathbf{y}}$. Consequently, in the soft-output decoding, the proposed EP-based Gibbs sampling decoding enjoys a flexible trade-off between performance and complexity by simply adjusting the number of Markov moves, and the near-MAP decoding performance can be achieved with the increase of the sample size.

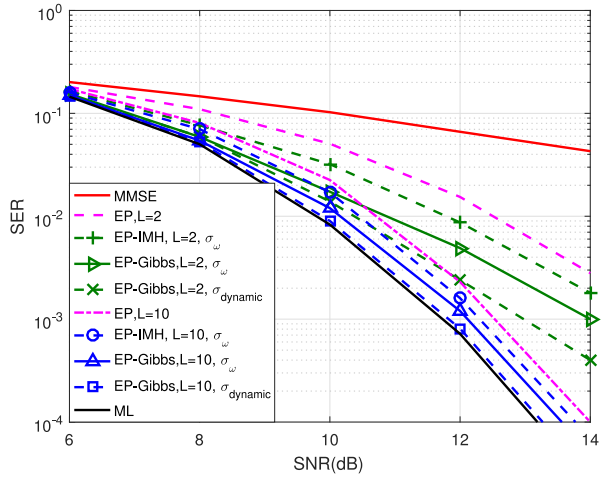


Fig. 2. Symbol error rate versus average SNR for the uncoded 12×12 MIMO system using 4-QAM.

VI. Simulation

In this section, the performance of the proposed EP-based sampling decoding schemes in MIMO systems are studied by simulations in full details.

In Fig. 2, the performance comparison about the proposed EP-based sampling decoding is illustrated in a 12×12 uncoded MIMO system with 4-QAM. The decoding performance is evaluated in terms of the symbol error rates (SERs), and the iteration numbers of the basic EP detector are set as $L = 2$ and $L = 10$ respectively. Then, based on the EP detector, the proposed independent MH sampling and Gibbs sampling schemes are shown, where the numbers of Markov moves are set as $T = 50$. Clearly, the proposed EP-based independent MH sampling achieves a better decoding performance than the original EP detector for both cases of $L = 2$ and $L = 10$. As expected, with the standard deviation σ_w, the EP-based Gibbs sampling outperforms the EP-based independent MH sampling, implying a better convergence performance. Note that the EP-based Gibbs sampling with $L = 10$ obtains a better decoding performance than that with $L = 2$, which indicates the importance of the choice of the starting point. Furthermore, an extra decoding gain can be obtained by the application of σ_{dynamic}, which is able to achieve the near-optimal decoding performance. This is also accordance with the afore-mentioned analysis about the choice of σ. Therefore, considering the superiority of σ_{dynamic}, we apply it to the EP-based Gibbs sampling by default in the rest of performance comparisons.

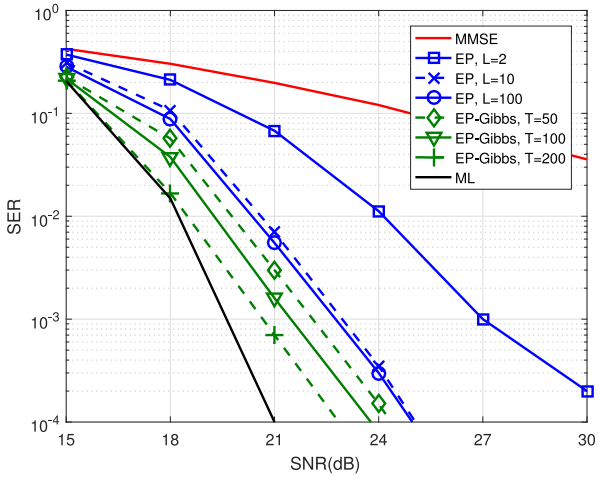


Fig. 3. Symbol error rate versus average SNR for the uncoded 12×12 MIMO system using 16-QAM.

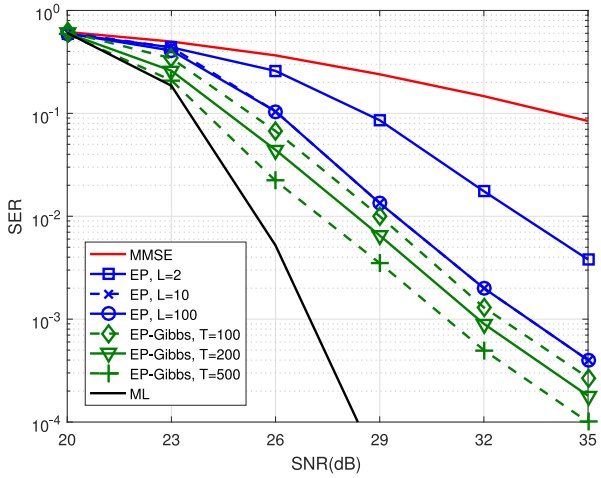


Fig. 4. Symbol error rate versus average SNR for the uncoded 12×12 MIMO system using 64-QAM.

In Fig. 3, the symbol error rates (SERs) of the EP decoding with different iteration numbers (i.e., $L = 2, 10, 100$) and the EP-based Gibbs sampling decoding with different numbers of Markov moves (i.e., $T = 50, 100, 200$) are illustrated in a 12×12 uncoded MIMO system with 16-QAM. Here, we point out that all the EP-based Gibbs sampling decoding schemes applied in the following are based on the decoding output of EP detector with $L = 10$. This is straightforward to understand as the decoding performance of EP is limited by convergence and there is no obvious difference between the cases of $L = 10$ and $L = 100$. For a better comparison, the performance of an MMSE detector is also presented. Clearly, with $L = 0$, the EP detector yields the equivalent MMSE decoding performance while a better decoding performance is achieved with the increment of L. Meanwhile, the optimal ML decoding, which is impractical due to the exponentially increased decoding complexity, serves as a baseline for the performance comparison.

Undoubtedly, the performance of the EP decoding with $L = 100$ almost remains the same with that of $L = 10$. In other words, the decoding performance of EP detector can not be always improved by simply increasing L. On the other hand, as for the EP-based Gibbs sampling decoding, an extra decoding performance gain can be achieved with the increment of Markov moves T. Clearly, with an increase in Markov moves, the near-optimal performance can be achieved by the proposed EP-based Gibbs sampling decoding. We emphasize that this is quite different from EP detector since the decoding performance gain of EP vanishes rapidly along with the number of iterations L, thus resulting in a performance limit after a number of iterations. In addition, the usage of EP also offers a good initial setting for the underlying Markov mixing, thus leads to a better decoding performance due to the faster convergence to the target distribution.

In Fig. 4, to investigate the impact of modulation orders, the performance comparison of the proposed EP-Gibbs sampling decoding is illustrated in a 12×12 uncoded MIMO system with 64-QAM. As can be seen, the decoding performance of the EP detector can not be further improved by simply increasing L. Different from it, the performance of EP-Gibbs sampling decoding still improves accordingly with the increment of T. However, due to the higher modulation order, more Markov moves are needed to guarantee the performance gain. This is expected because the whole state space of $\mathbf{s}$ is greatly expanded, resulting in a slower convergence rate. Nevertheless, the convergence of the Markov chain is still ensured and the near-ML decoding performance can be got if a large enough T is allowed. Note that the complexity of the proposed EP-based Gibbs sampling decoding turns out to be mild (the sampling complexity by Gibbs sampling is $O(T \cdot n^2)$) so as to a flexible decoding trade-off between performance and complexity. Another observation should be mentioned is that all the EP-based sampling decoding schemes fail to achieve the full receive diversity, which makes the EP-based decoding less attractive in the area of high SNRs.

Fig. 5 shows the SER of the proposed EP-Gibbs sampling decoding in a 24×24 uncoded MIMO system with 16-QAM. This corresponds to a lattice decoding scenario with the restricted state space in dimension $n = 48$. Similarly, with the increase of iteration number L, the performance of EP decoding improves gradually but is limited at $L = 10$. As can be seen, no extra performance gain can be obtained even if $L = 100$. On the other hand, a better decoding performance can be achieved when the proposed EP-Gibbs sampling decoding is applied, where cases of $T = 100$, $T = 200$ and $T = 500$ are given. Therefore, a flexible trade-off with respect to the EP-Gibbs sampling decoding is established by adjusting the parameter T. Moreover, the related performance comparison regarding to a 32×32 uncoded MIMO system with 16-QAM is presented in Fig. 6. Compared to cases of 12×12 and 24×24, it has a much higher system dimension so as to a much larger state space of $\mathbf{s}$, which means more Markov moves are required to ensure a certain performance gain. Note that the EP and EP-based Gibbs sampling decoding fail to achieve the full receive diversity gain

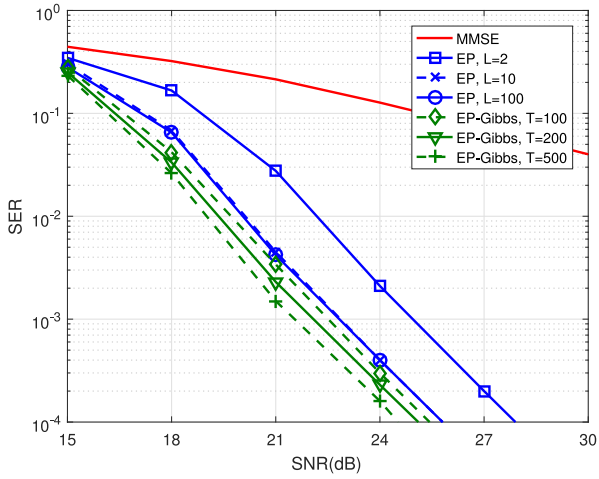


Fig. 5. Symbol error rate versus average SNR for the uncoded 24×24 MIMO system using 16-QAM.

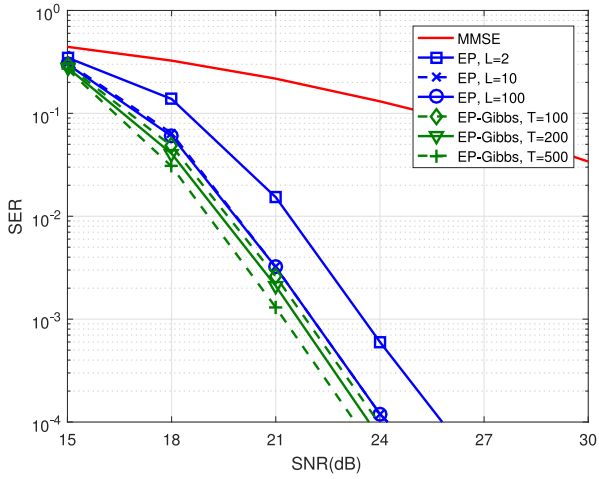


Fig. 6. Symbol error rate versus average SNR for the uncoded 32×32 MIMO system using 16-QAM.

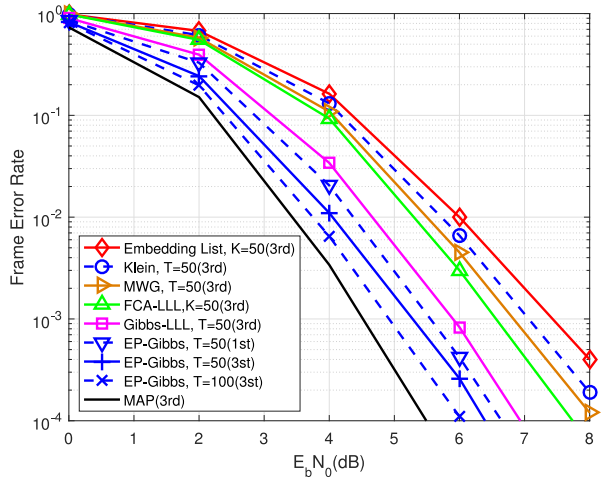


Fig. 7. Frame error rate versus average SNR per bit in the coded 8×8 MIMO BICM-IDD system using 4-QAM.

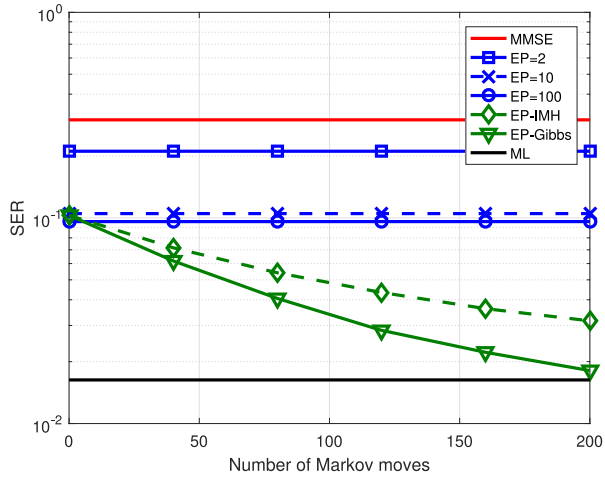


Fig. 8. Symbol error rate versus number of Markov moves for the uncoded 12×12 MIMO system using 16-QAM with SNR $= 18$ dB.

in all the cases. Because of this, the superiority of EP-based algorithms vanishes gradually with the increment of SNR. This means considerable decoding potential can be further exploited, which is one of our research work in future.

Fig. 7 shows the frame error rate (FER) for a coded 8×8 MIMO bit-interleaved coded modulation iterative detection and decoding system with 4-QAM, using a rate-1/2, irregular (256,128,3) low-density parity-check (LDPC) code of codeword length 256 (i.e., 128 information bits). Each codeword spans one channel realization and a random bit interleaver is used. The parity check matrix is randomly constructed, but cycles of length 4 are eliminated. The maximum number of decoding iterations for LDPC is set at 50. Clearly, after three iterations between MIMO detector and soft-output decoder in IDD, the proposed EP-based Gibbs sampling decoding with $T = 50$ Markov moves performs better than FCA, embedding decoding [62], Metropolis-within-Gibbs sampling [63] and Klein's sampling [64]. Note that the EP-based Gibbs sampling decoding also outperforms the conventional Gibbs sampling with starting point outputted by SIC-LLL decoder due to a better starting point for the Markov mixing. For a better comparison, the performance of the proposed sampling algorithm after one iteration is also given. On the other hand, the proposed EP-based Gibbs sampling decoding is able to achieve the near-MAP decoding performance with the increment of Markov moves. Therefore, by adjusting the number of Markov moves T, the whole system enjoys a flexible trade-off between performance and complexity.

In Fig. 8 illustrates the SER decoding performance of EP-based sampling algorithms with fixed SNR $= 18$ dB in a 12×12 uncoded MIMO system with 16-QAM. For a fair comparison, MMSE decoding and ML decoding are applied to serve as the baselines. Clearly, with the increase of iterations from $L = 2$ to $L = 10$, the decoding performance of EP decoding gradually improves. However, the performance limitation does exist as

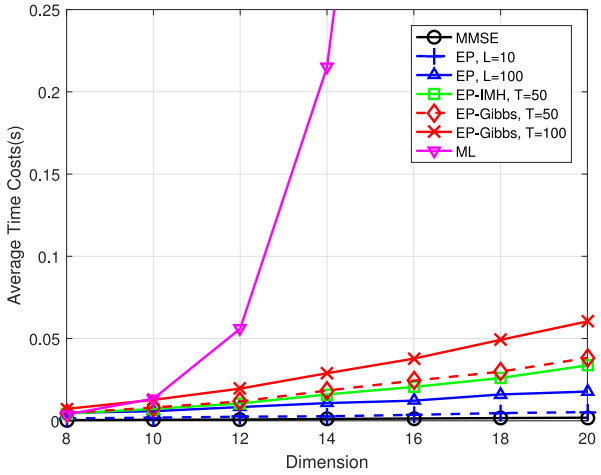


Fig. 9. Complexity comparison in average time cost for the uncoded MIMO system using 16-QAM at SNR $= 18$ dB.

few gain can be obtained even by $L = 100$. Differen from EP decoding, the EP-based sampling decoding schemes (based on EP with $L = 10$) are able to achieve a better performance with the number of Markov moves. More specifically, near-optimal decoding performance can be obtained by EP-based Gibbs sampling with $T = 200$ while the decoding trade-off between performance and complexity is flexible by adjusting T. Note that EP-based Gibbs sampling has a better decoding performance than EP-based IMH sampling due to a better convergence performance. This is easy to understand as the former fully takes advantages of the correlation among components of $\mathbf{s}$.

As a complement to illustrate the computational cost, Fig. 9 is given to show the complexity comparison in terms of the average elapsed running times. In particular, the uncoded MIMO system takes 16-QAM at SNR $= 18$ dB, and the simulation is conducted by MATLAB R2019a on a single computer, with an Intel Core i7 processor at 2.3 GHz, a RAM of 8 GB. Clearly, the average elapsed running time of EP and EP-based decoding scheme increase mildly with the increase of system dimension from 8×8 to 20×20 MIMO systems. On the contrary, the optimal ML decoding from [65] takes an exponentially increasing average elapsed running time, which is unaffordable in high-dimensional cases. As expected, under the same number of Markov moves $T = 50$, the complexity of the proposed EP-based Gibbs sampling decoding is comparable to that of the EP-based IMH sampling decoding. This is accordance to the derived complexity $O(Tn^2)$, making them easy to be implemented especially in high-dimensional MIMO systems.

VII. Conclusion

In this paper, the framework of the EP-based sampling decoding scheme has been proposed to achieve a better decoding trade-off between performance and complexity in large-scale MIMO detection. By performing the random sampling over the mean vector yielded by EP, extra decoding gain can be obtained, and the convergence of the approximation of the target posterior distribution can be ensured as well. In order to further improve the convergence performance and decoding efficiency, we have proposed the EP-based Gibbs sampling decoding algorithm. Meanwhile, the choice of the standard deviation σ of the target discrete Gaussian distribution has also been investigated, thus resulting in a better trade-off between the target sampling probability and the Markov mixing. Moreover, we have integrated the proposed algorithm with the soft-output decoding in MIMO-BICM systems, where a flexible decoding trade-off can be achieved by adjusting the number of Markov moves. Finally, simulation results based on the massive MIMO detection have been presented to confirm the advanced decoding trade-off in both MIMO detection and soft-output decoding.

Acknowledgment

The authors would like to thank Dr. Cong Ling (Imperial College London, U.K.) for his helpful discussions and insightful suggestions.

References

[1] T. L. Marzetta, "Noncooperative cellular wireless with unlimited numbers of base station antennas," *IEEE Trans. Wireless Commun.*, vol. 9, no. 11, pp. 3590–3600, Nov. 2010.
[2] E. G. Larsson, O. Edfors, F. Tufvesson, and T. L. Marzetta, "Massive MIMO for next generation wireless systems," *IEEE Commun. Mag.*, vol. 52, no. 2, pp. 186–195, Feb. 2014.
[3] F. Rusek *et al.*, "Scaling up MIMO: Opportunities and challenges with very large arrays," *IEEE Signal Process. Mag.*, vol. 30, no. 1, pp. 40–60, Jan. 2013.
[4] H. Q. Ngo, E. G. Larsson, and T. L. Marzetta, "Energy and spectral efficiency of very large multiuser MIMO systems," *IEEE Trans. Commun.*, vol. 61, no. 4, pp. 1436–1449, Apr. 2013.
[5] N. Srinidhi, T. Datta, A. Chockalingam, and B. S. Rajan, "Layered Tabu search algorithm for large-MIMO detection and a lower bound on ML performance," *IEEE Trans. Commun.*, vol. 59, no. 11, pp. 2955–2963, Nov. 2011.
[6] L. Dai, X. Gao, X. Su, S. Han, C. I, and Z. Wang, "Low-complexity soft-output signal detection based on Gauss-Seidel method for uplink multiuser large-scale MIMO systems," *IEEE Trans. Veh. Technol.*, vol. 64, no. 10, pp. 4839–4845, Oct. 2015.
[7] A. Lu, X. Gao, Y. R. Zheng, and C. Xiao, "Low complexity polynomial expansion detector with deterministic equivalents of the moments of channel Gram matrix for massive MIMO uplink," *IEEE Trans. Commun.*, vol. 64, no. 2, pp. 586–600, Feb. 2016.
[8] S. Wu, L. Kuang, Z. Ni, J. Lu, D. Huang, and Q. Guo, "Low-complexity iterative detection for large-scale multiuser MIMO-OFDM systems using approximate message passing," *IEEE J. Sel. Topics Signal Process.*, vol. 8, no. 5, pp. 902–915, Oct. 2014.
[9] P. Som, T. Datta, N. Srinidhi, A. Chockalingam, and B. S. Rajan, "Low-complexity detection in large-dimension MIMO-ISI channels using graphical models," *IEEE J. Sel. Topics Signal Process.*, vol. 5, no. 8, pp. 1497–1511, Dec. 2011.
[10] J. Cspedes, P. M. Olmos, M. Snchez-Fernndez, and F. Perez-Cruz, "Expectation propagation detection for high-order high-dimensional MIMO systems," *IEEE Trans. Commun.*, vol. 62, no. 8, pp. 2840–2849, Aug. 2014.
[11] T. Minka, "Expectation propagation for approximate Bayesian inference," in *Proc. 17th Conf. Uncertainty Artif. Intell.*, 2001, pp. 362–369.
[12] T. Minka and Y. Qi, "Tree-structured approximations by expectation propagation," in *Proc. Neural Inf. Process. Syst.*, Dec. 2004, pp. 362–369.
[13] J. Goldberger and A. Leshem, "MIMO detection for high-order QAM based on a Gaussian tree approximation," *IEEE Trans. Inf. Theory*, vol. 57, no. 8, pp. 4973–4982, Aug. 2011.
[14] J. Goldberger, "Improved mimo detection based on successive tree approximations," in *Proc. IEEE Int. Symp. Inf. Theory*, Jul. 2013, pp. 2004–2008.

[15] J. Cspedes, P. M. Olmos, M. Snchez-Fernndez, and F. Perez-Cruz, "Probabilistic MIMO symbol detection with expectation consistency approximate inference," *IEEE Trans. Veh. Technol.*, vol. 67, no. 4, pp. 3481–3494, Apr. 2018.

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

[17] X. Tan, Y. Ueng, Z. Zhang, X. You, and C. Zhang, "A low-complexity massive MIMO detection based on approximate expectation propagation," *IEEE Trans. Veh. Technol.*, vol. 68, no. 8, pp. 7260–7272, Aug. 2019.

[18] K. Ghavami and M. Naraghi-Pour, "MIMO detection with imperfect channel state information using expectation propagation," *IEEE Trans. Veh. Technol.*, vol. 66, no. 9, pp. 8129–8138, Sep. 2017.

[19] D. Zhang, L. L. Mendes, M. Matth, I. S. Gaspar, N. Michailow, and G. P. Fettweis, "Expectation propagation for near-optimum detection of MIMO-GFDM signals," *IEEE Trans. Wireless Commun.*, vol. 15, no. 2, pp. 1045–1062, Feb. 2016.

[20] P. M. Olmos, J. J. Murillo-Fuentes, and F. Prez-Cruz, "Tree-structure expectation propagation for LDPC decoding over the BEC," *IEEE Trans. Inf. Theory*, vol. 59, no. 6, pp. 3354–3377, Jun. 2013.

[21] L. Salamanca, P. M. Olmos, J. J. Murillo-Fuentes, and F. Perez-Cruz, "Tree expectation propagation for ML decoding of LDPC codes over the BEC," *IEEE Trans. Commun.*, vol. 61, no. 2, pp. 465–473, Feb. 2013.

[22] I. Santos, J. J. Murillo-Fuentes, E. Arias-de-Reyna, and P. M. Olmos, "Turbo EP-based equalization: A filter-type implementation," *IEEE Trans. Commun.*, vol. 66, no. 9, pp. 4259–4270, Sep. 2018.

[23] S. Sahin, A. M. Cipriano, C. Poulliat, and M. Boucheret, "Iterative equalization with decision feedback based on expectation propagation," *IEEE Trans. Commun.*, vol. 66, no. 10, pp. 4473–4487, Oct. 2018.

[24] A. Gelman, A. Vehtari, P. Jylanki, C. Robert, N. Chopin, and J. P. Cunningham, "Expectation propagation as a way of life," [Online]. Available: https://arxiv.org/abs/1412.4869

[25] S. Liu, C. Ling, and D. Stehlé, "Decoding by sampling: A randomized lattice algorithm for bounded distance decoding," *IEEE Trans. Inform. Theory*, vol. 57, no. 9, pp. 5933–5945, Sep. 2011.

[26] T. Datta, N. Kumar, A. Chockalingam, and B. Rajan, "A novel Monte Carlo sampling based receiver for large-scale uplink multiuser MIMO systems," *IEEE Trans. Veh. Technol.,* vol. 62, no. 7, pp. 3019–3038, Sep. 2013.

[27] B. Hassibi, M. Hansen, A. Dimakis, H. Alshamary, and W. Xu, "Optimized Markov Chain Monte Carlo for signal detection in MIMO systems: An analysis of the stationary distribution and mixing time," *IEEE Trans. Signal Process.*, vol. 62, no. 17, pp. 4436–4450, Sep. 2014.

[28] J. Choi, "An MCMC-MIMO detector as a stochastic linear system solver using successive overrelexation," *IEEE Trans. Wireless Commun.*, vol. 15, no. 2, pp. 1445–1455, Feb. 2016.

[29] Z. Wang, S. Liu, and C. Ling, "Decoding by sampling—Part II: Derandomization and soft-output decoding," *IEEE Trans. Commun.*, vol. 61, no. 11, pp. 4630–4639, Nov. 2013.

[30] P. Klein, "Finding the closest lattice vector when it is unusually close," in Proc. *ACM-SIAM Symp. Discrete Algorithms*, 2000, pp. 937–941.

[31] Z. Wang and C. Ling, "On the geometric ergodicity of Metropolis-Hastings algorithms for lattice Gaussian sampling," *IEEE Trans. Inf. Theory*, vol. 64, no. 2, pp. 738–751, Feb. 2018.

[32] Z. Wang and C. Ling, "Lattice Gaussian sampling by Markov chain Monte Carlo: Bounded distance decoding and trapdoor sampling," *IEEE Trans. Inf. Theory*, vol. 65, no. 6, pp. 3630–3645, Jun. 2019.

[33] Z. Wang, Y. Huang, and S. Lyu, "Lattice-reduction-aided gibbs algorithm for lattice Gaussian sampling: Convergence enhancement and decoding optimization," *IEEE Trans. Signal Process.*, vol. 67, no. 16, pp. 4342–4356, Aug. 2019.

[34] D. A. Levin, Y. Peres, and E. L. Wilmer, *Markov Chains and Mixing Time*. American Mathematical Society, 2008.

[35] P. Aggarwal and X. Wang, "Multilevel sequential Monte Carlo algorithms for MIMO demodulation," *IEEE Trans. Wireless Commun.*, vol. 6, no. 2, pp. 750–758, Feb. 2007.

[36] R. Chen, J. Liu, and X. Wang, "Convergence analysis and comparisons of Markov chain Monte Carlo algorithms in digital communications," *IEEE Trans. Signal Process.*, vol. 50, no. 2, pp. 255–270, Feb. 2002.

[37] B. M. Hochwald and S. ten Brink, "Achieving near-capacity on a multiple-antenna channel," *IEEE Trans. Commun.*, vol. 51, no. 3, pp. 389–399, Mar. 2003.

[38] H. Vikalo, B. Hassibi, and T. Kailath, "Iterative decoding for MIMO channels via modified sphere decoding," *IEEE Trans. Wireless Commun.*, vol. 3, no. 6, pp. 2299–2311, Nov. 2004.

[39] P. Silvola, K. Hooli, and M. Juntti, "Suboptimal soft-output map detector with lattice reduction," *IEEE Signal Process. Lett.*, vol. 13, no. 6, pp. 321–324, Jun. 2006.

[40] W. Zhang and X. Ma, "Low-complexity soft-output decoding with lattice-reduction-aided detectors," *IEEE Trans. Commun.*, vol. 58, no. 9, pp. 2621–2629, Sep. 2010.

[41] B. Hassibi and H. Vikalo, "On the sphere-decoding algorithm I. Expected complexity," *IEEE Trans. Signal Process.*, vol. 53, no. 8, pp. 2806–2818, Aug. 2005.

[42] E. Agrell, T. Eriksson, A. Vardy, and K. Zeger, "Closest point search in lattices," *IEEE Trans. Inform. Theory*, vol. 48, no. 8, pp. 2201–2214, Aug. 2002.

[43] O. Regev, "Lattices in computer science," Lecture notes, 2004, [Online]. Available: https://www.cs.tau.ac.il/odedr/

[44] E. Larsson and J. Jalden, "Fixed-complexity soft MIMO detection via partial marginalization," *IEEE Trans. Signal Process.*, vol. 56, no. 8, pp. 3397–3407, Aug. 2008.

[45] R. Wang and G. B. Giannakis, "Approaching MIMO channel capacity with soft detection based on hard sphere decoding," *IEEE Trans. Commun.*, vol. 54, no. 4, pp. 587–590, Apr. 2006.

[46] C. Studer and H. Bolcskei, "Soft-input soft-output single tree-search sphere decoding," *IEEE Trans. Inform. Theory*, vol. 56, no. 10, pp. 4827–4842, Oct. 2010.

[47] J. Lee, B. Shim, and I. Kang, "Soft-input soft-output list sphere detection with a probabilistic radius tightening," *IEEE Trans. Wireless Commun.*, vol. 11, no. 8, pp. 2848–2857, Aug. 2012.

[48] C. Peikert, "An efficient and parallel Gaussian sampler for lattices," in *Proc. CRYPTO*, 2010, pp. 80–97.

[49] S. P. Meyn and R. L. Tweedie, *Markov Chains and Stochastic Stability*. Cambridge, U.K.: Cambridge Univ. Press, 2009.

[50] G. O. Roberts, "General state space Markov chains and MCMC algorithms," *Probability Surveys*, vol. 1, pp. 20–71, 2004.

[51] J. S. Liu, *Monte Carlo Strategies in Scientific Computing*. Berlin, Germany: Springer-Verlag, 2001.

[52] J. S. Rosenthal, "Minorization conditions and convergence rates for Markov chain Monte Carlo," *J. Amer. Statist. Assoc.*, vol. 90, pp. 558–566, 1995.

[53] H. Zhu, B. Farhang-Boroujeny, and R.-R. Chen, "On performance of sphere decoding and Markov chain Monte Carlo detection methods," *IEEE Signal Process. Lett.*, vol. 12, no. 10, pp. 669–672, Oct. 2005.

[54] D. Micciancio and O. Regev, "Worst-case to average-case reductions based on Gaussian measures," in *Proc. Ann. Symp. Found. Comput. Sci.*, Rome, Italy, Oct. 2004, pp. 372–381.

[55] D. Aggarwal, D. Dadush, O. Regev, and N. Stephens-Davidowitz, "Solving the shortest vector problem in 2^n time via discrete Gaussian sampling," in *Proc. 47th Annu. ACM Symp. Theory Comput.*, 2015, pp. 733–742.

[56] K. Latuszyński, G. O. Roberts, and J. S. Rosenthal, "Adaptive Gibbs samplers and related MCMC methods," *Ann. Appl. Probability*, vol. 23, no. 1, pp. 66–98, 2013.

[57] C. J. Geyer and E. A. Thompson, "Annealing Markov chain Monte Carlo with applications to ancestral inference," *J. Amer. Statist. Assoc.*, vol. 90, pp. 909–920, 1995.

[58] C. Gentry, C. Peikert, and V. Vaikuntanathan, "Trapdoors for hard lattices and new cryptographic constructions," in *Proc. 40th Annu. ACM Symp. Theory Comput.*, Victoria, Canada, 2008, pp. 197–206.

[59] W. Banaszczyk, "New bounds in some transference theorems in the geometry of numbers," *Math. Ann.*, vol. 296, pp. 625–635, 1993.

[60] D. Aharonov and O. Regev, "Lattice problems in NP ∩ coNP." *J. ACM*, vol. 52, no. 5, pp. 749–765, 2005.

[61] N. Stephens-Davidowitz, "Discrete Gaussian sampling reduces to CVP and SVP," in *Proc. 27th Annu. ACM-SIAM Symp. Discrete Algorithms*, 2016, pp. 1748–1764.

[62] L. Luzzi, D. Stehlé, and C. Ling, "Decoding by embedding: correct decoding radius and DMT optimality," *IEEE Trans. Inform. Theory*, vol. 59, no. 5, pp. 2960–2973, May 2013.

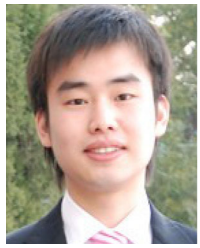
[63] Z. Wang and C. Ling, "Symmetric Metropolis-within-Gibbs algorithm for lattice Gaussian sampling," in *Proc. IEEE Inf. Theory Workshop*, 2016, pp. 394–398.

[64] Z. Wang, C. Ling, and G. Hanrot, "Markov chain Monte Carlo algorithms for lattice Gaussian sampling," in *Proc. IEEE Int. Symp. Inf. Theory*, Honolulu, USA, Jun. 2014, pp. 1489–1493.

[65] M. O. Damen, H. E. Gamal, and G. Caire, "On maximum-likelihood detection and the search for the closest lattice point," *IEEE Trans. Inform. Theory*, vol. 49, no. 10, pp. 2389–2401, Oct. 2003.

Zheng Wang (Member, IEEE) received the B.S. degree in electronic and information engineering from the Nanjing University of Aeronautics and Astronautics, Nanjing, China, in 2009, and the M.S. degree in communications from the Department of Electrical and Electronic Engineering, University of Manchester, Manchester, U.K., in 2010. He received the Ph.D. degree in communication engineering from Imperial College London, U.K., in 2015. From 2015 to 2016, he was a Research Associate with Imperial College London, U.K. From 2016 to 2017, he was a Senior Engineer with Radio Access Network R&D division, Huawei Technologies Company. From 2017, he is an Associate Professor with the College of Electronic and Information Engineering, Nanjing University of Aeronautics and Astronautics (NUAA), Nanjing, China. His current research interests include MIMO systems, machine learning and data analytics over wireless networks, and lattice theory for wireless communications.

Shanxiang Lyu (Member, IEEE) received the B.Eng. and M.Eng. degrees in electronic and information engineering from the South China University of Technology, Guangzhou, China, in 2011 and 2014, respectively, and the Ph.D. degree from the Electrical and Electronic Engineering Department, Imperial College London, in 2018. He is currently a Lecturer with the College of Cyber Security, Jinan University. His research interests include lattice theory, algebraic number theory, and their applications.



Yili Xia (Member, IEEE) received the B.Eng. degree in information engineering from Southeast University, Nanjing, China, in 2006, and the M.Sc. (Distinction) degree in communications & signal processing from the Department of Electrical and Electronic Engineering, Imperial College London, London, U.K., in 2007. He received the Ph.D. degree in adaptive signal processing from Imperial College London in 2011.

Since 2013, he has been an Associate Professor in Signal Processing with the School of Information and Engineering, Southeast University, Nanjing, China, where he is currently the Deputy Head with the Department of Information and Signal Processing Engineering. His research interests include complex and hyper-complex statistical analysis, detection and estimation, linear and nonlinear adaptive filters, as well as their applications on communications and power systems. He was the recipient of the Best Student Paper Award at the International Symposium on Neural Networks (ISNN) in 2010 (coauthor), and the Education Innovation Award at the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP) in 2019. He is currently an Associate Editor for the IEEE TRANSACTIONS ON SIGNAL PROCESSING.

Qihui Wu (Senior Member, IEEE) received the B.S. degree in communications engineering, the M.S. and Ph.D. degrees in communications and information systems from the Institute of Communications Engineering, Nanjing, China, in 1994, 1997, and 2000, respectively. From 2003 to 2005, he was a Postdoctoral Research Associate with Southeast University, Nanjing, China. From 2005 to 2007, he was an Associate Professor with the College of Communications Engineering, PLA University of Science and Technology, Nanjing, China, where he was a Full Professor from 2008 to 2016. Since May 2016, he has been a Full Professor with the College of Electronic and Information Engineering, Nanjing University of Aeronautics and Astronautics, Nanjing, China. From March 2011 to September 2011, he was an Advanced Visiting Scholar with the Stevens Institute of Technology, Hoboken, USA. His current research interests span the areas of wireless communications and statistical signal processing, with emphasis on system design of software defined radio, cognitive radio, and smart radio.