entropy

MDPI

Article

Nested Variational Chain and Its Application in Massive MIMO Detection for High-Order Constellations

Qiwei Wang

School of Telecommunications Engineering, Xidian University, Xi'an 710071, China; qwwang@xidian.edu.cn

Abstract: Multiple input multiple output (MIMO) technology necessitates detection methods with high performance and low complexity; however, the detection problem becomes severe when high-order constellations are employed. Variational approximation-based algorithms prove to deal with this problem efficiently, especially for high-order MIMO systems. Two typical algorithms named Gaussian tree approximation (GTA) and expectation consistency (EC) attempt to approximate the true likelihood function under discrete finite-set constraints with a new distribution by minimizing the Kullback–Leibler (KL) divergence. As the KL divergence is not a true distance measure, 'exclusive' and 'inclusive' KL divergences are utilized by GTA and EC, respctively, demonstrating different performances. In this paper, we further combine the two asymmetric KL divergences in a nested way by proposing a generic algorithm framework named nested variational chain. Acting as an initial application, a MIMO detection algorithm named Gaussian tree approximation expectation consistency (GTA-EC) can thus be presented along with its alternative version for better understanding. With less computational burden compared to its counterparts, GTA-EC is able to provide better detection performance and diversity gain, especially for large-scale high-order MIMO systems.

Keywords: massive multiple input multiple output (MIMO); nested variational chain; Gaussian tree approximation (GTA); expectation consistency (EC)

Citation: Wang, Q. Nested Variational Chain and Its Application in Massive MIMO Detection with High-Order Constellations. *Entropy* **2023**, *25*, 1621. https://doi.org/10.3390/e25121621

Academic Editors: Sébastien Roy, Julian Cheng and Jean-Yves Chouinard

Received: 24 October 2023
Revised: 21 November 2023
Accepted: 29 November 2023
Published: 5 December 2023

1. Introduction

Multiple input multiple output (MIMO) technology has attracted broad attention over the last decade and has been widely applied into practical communication systems. The benefit of MIMO technology lies in the improvement of spectral efficiency and link reliability due to the multiplexing and diversity gain that grows with the number of elements, and a MIMO system is referred to as a *massive* MIMO system when the scale of array elements grows large enough, which brings increasing difficulty to the signal detection due to huge computational burden, hindering the prevailing usage of massive MIMO systems [1,2].

Many research studies have been carried out for signal detection in massive MIMO systems [3–17]. It is well known that the maximum likelihood detection presents the best detection performance with the cost of exponentially growing computational burden [3]. Neglecting the finite-set constraint, the minimum mean square error (MMSE) approach can be applied by solving the least square fit, and a closest lattice point can then be found by treating symbols independently [4]. The MMSE approach normally exhibits a benchmark performance when comparing different detectors, and its performance can be vastly improved by MMSE-SIC when combining with the successive interference cancellation (SIC) technique [5]. However, as MMSE or MMSE-SIC cannot provide satisfied performance, several alternatives have been proposed instead, which can be divided into two major categories, i.e., sub-space searching-based and variational inference-based detectors.

The sub-space searching-based category originates from the idea of reducing the searching space of all possible lattice points with unacceptable complexity. Sphere decoding tries to replicate the maximum likelihood performance by diminishing the searching space,

the dimension of which grows up with the number of antennas as well as the modulation order, making it prohibitive for the large-scale or high-order MIMO systems [6,7]. Another two local searching-based approaches were proposed by the name of likelihood ascending search and reactive tabu search [8–10], and the basic idea behind them is to search through a proximity sub-space around a given initial solution. They present good performance for a large number of antennas with low-order constellations but poor performance for high-order constellations. A layered tabu search algorithm was proposed in [11] by performing detection over layers, requiring a higher order of complexity for high-order constellations, and a Gibbs sampling-based detector was proposed in [12] by performing a serial of one-dimensional searches over iterations. It may provide good performance for low-order constellations with the cost of enormous processing time. Therefore, algorithms in the former category suffer from poor performance or prohibitive computational burden in large-scale high-order MIMO systems.

Proved to be suitable for the detection problem resulted by high-order constellations, the variational inference based category tries to approximate the true likelihood function into a new distribution that is much easier to handle. A Gaussian tree approximation (GTA) algorithm was proposed in [13,14] by transforming the fully connected factor graph into a tree graph, based on which belief propagation based message passing can be proceeded for inference. The GTA algorithm has comparable performance with MMSE-SIC at a similar complexity only to MMSE. The expectation propagation (EP) algorithm was proposed for MIMO detection in [15] by substituting true priors belonging to a discrete finite set with the introduced Gaussian priors being able to be updated over iterations. EP performs the best at a complexity several times that of MMSE, and its alternative named expectation consistency (EC) was then proposed to provide a more general perspective than EP [16]. Two low-complexity EP/EC-based algorithms were proposed for scenarios when the number of transmit antennas is less than that the number of receiver ones [17,18], and a double-EP based iterative detection and decoding was proposed by iteratively exploiting decoders in [19]. EP/EC-based algorithms can also be applied into channel estimation problems in massive MIMO systems [20,21].

In this paper, we would like to expand the variational inference paradigm by proposing a nested variational chain. The basic idea behind it is that 'exclusive' and 'inclusive' KL divergences employed by GTA and EC, respectively, are not exclusive and can be combined in a nested way so as to form an approximation chain, by which both GTA and EC are improved. The major contributions are listed as follows.

- Firstly, the basic idea of the nested variational chain is proposed, and an algorithm is then proposed to establish a general framework. By referring to 'general', it means this framework is able to combine 'exclusive' and 'inclusive' KL divergences, or it degrades to either one as a special case.
- Secondly, providing several examples, we show that existing algorithms, such as MMSE, GTA, and EC, can be regarded as special cases of the variational chain.
- Finally, to provide an initial application of the variational chain into massive MIMO detection, a GTA-embedded Expecatation Consistency (GTA-EC) algorithm is proposed which proves to provide better detection performance, especially for high-order constellations. The complexity of GTA-EC is analyzed as well along with comparisons.

This paper is arranged as follows. Section 2 introduces the system model and MIMO detection problem, based on which the nested variational chain is provided in Section 3 along with a generic algorithm framework. Section 4 derives the GTA-EC algorithm with complexity analyses. Simulation results are demonstrated in Section 5 along with discussions, and the conclusion is presented in Section 6. Throughout this paper, matrices and vectors are denoted by symbols in boldface, and variables are denoted in italics. The notation $\mathbf{A}^\top$ or $\mathbf{a}^\top$ is used to represent the transpose of a vector or matrix, and $\mathbf{I}$ represents a unit matrix.

2. Preliminary

2.1. Signal Model

A multiuser MIMO system is considered, without loss of generality, in which $\widetilde{N}$ transmitters, each equipped with one antenna, communicate with a base station that is equipped with $\widetilde{M}$ antennas. Assume each transmitter transmits any symbol $\widetilde{x}_i \in \mathbb{C}$ that is selected from a Quadrature Amplitude Modulation (QAM) constellation set $\widetilde{\mathcal{A}}$, where $\mathbb{C}$ stands for the complex domain, and the cardinality of the constellation set is $|\widetilde{\mathcal{A}}| = A$. The transmitted symbols can be represented as a vector $\widetilde{\mathbf{x}} \in \widetilde{\mathcal{A}}^{\widetilde{N}\times 1}$ with the average energy of each QAM symbol defined as $\widetilde{E}_s$. After propagating through the wireless channels, the received signal $\widetilde{\mathbf{y}} \in \mathbb{C}^{\widetilde{M}\times 1}$ at the base station can be expressed as

$$\widetilde{\mathbf{y}} = \widetilde{\mathbf{H}}\,\widetilde{\mathbf{x}} + \widetilde{\mathbf{n}}, \tag{1}$$

where $\widetilde{\mathbf{n}} \in \mathbb{C}^{M\times 1}$ stands for additive Gaussian white noises (AWGN), each element having zero mean and σ_n^2 variance. $\widetilde{\mathbf{H}} \triangleq [\widetilde{\mathbf{h}}_1, \ldots, \widetilde{\mathbf{h}}_i, \ldots, \widetilde{\mathbf{h}}_N]$ is defined as a matrix by stacking up channel coefficients with $\widetilde{\mathbf{h}}_i = \mathrm{diag}\{\widetilde{h}_{i,1}, \ldots, \widetilde{h}_{i,m}, \ldots, \widetilde{h}_{i,M}\}$ being Rayleigh flat-fading channel coefficients of the ith symbol. Perfect channel state information (CSI) is assumed such that $\widetilde{\mathbf{H}}$ is known at the base station.

The channel model above is usually re-expressed in the real domain by taking into consideration real and imaginary parts, respectively. By defining $\mathcal{R}(\cdot)$ and $\mathcal{I}(\cdot)$ as operations to take the real and imaginary part of a variable or matrix, one can define $\mathbf{y} = [\mathcal{R}(\widetilde{\mathbf{y}})^\top\ \mathcal{I}(\widetilde{\mathbf{y}})^\top]^\top$, $\mathbf{x} = [\mathcal{R}(\widetilde{\mathbf{x}})^\top\ \mathcal{I}(\widetilde{\mathbf{x}})^\top]^\top$, $\mathbf{n} = [\mathcal{R}(\widetilde{\mathbf{n}})^\top\ \mathcal{I}(\widetilde{\mathbf{n}})^\top]^\top$, and that

$$\mathbf{H} = \begin{bmatrix} \mathcal{R}(\widetilde{\mathbf{H}}) & -\mathcal{I}(\widetilde{\mathbf{H}}) \\ \mathcal{I}(\widetilde{\mathbf{H}}) & \mathcal{R}(\widetilde{\mathbf{H}}) \end{bmatrix},$$

where $\mathbf{y}, \mathbf{n} \in \mathbb{R}^{M\times 1}$, $\mathbf{x} \in \mathbb{R}^{N\times 1}$, $\mathbf{H} \in \mathbb{R}^{M\times N}$, $M = 2\widetilde{M}$, $N = 2\widetilde{N}$, and $\mathbb{R}$ stands for the real domain. The equivalent model in the real domain is then given as

$$\mathbf{y} = \mathbf{H}\,\mathbf{x} + \mathbf{n}, \tag{2}$$

where the variance of each element of $\mathbf{n}$ equals $\sigma_n^2 = \widetilde{\sigma}_n^2/2$, and $\mathbf{x}$ belongs to the pulse amplitude modulation (PAM) constellation set $\mathcal{A}$ containing real and imaginary parts of the A-QAM alphabets with its cardinality being $|\mathcal{A}| = \sqrt{A}$. The average energy of a PAM symbol is $E_s = \widetilde{E}_s/2$, and the signal-to-noise (SNR) of the MIMO system is then defined as

$$\mathrm{SNR} = 10 \cdot \log_{10}\left(\frac{N\widetilde{E}_s}{\widetilde{\sigma}_n^2}\right) = 10 \cdot \log_{10}\left(\frac{NE_s}{\sigma_n^2}\right)$$

2.2. MIMO Detection

As the received signal in a MIMO system is a superposition of transmitted symbols weighted by channel coefficients, the purpose of MIMO detection is to estimate successfully all transmitted symbols impaired by channel fading and noises. As is well known, the maximum *a posteriori* (MAP) detector could achieve the best detection performance by maximizing the a posteriori probability as follows

$$\widehat{\mathbf{x}} = \arg\max_{\mathbf{x}\in\mathcal{A}} \mathcal{P}(\mathbf{x}|\mathbf{y}, \mathbf{H}),$$

where the *aposteriori* distribution given the received signal $\mathbf{y}$ and CSI $\mathbf{H}$ is expressed as

$$\mathcal{P}(\mathbf{x}|\mathbf{y}, \mathbf{H}) \propto \mathcal{N}\Big(\mathbf{y} : \mathbf{H}\mathbf{x}, \sigma_n^2\mathbf{I}\Big)\mathcal{P}(\mathbf{x}), \tag{3}$$

$\mathcal{N}\big(\mathbf{y} : \mathbf{H}\mathbf{x}, \sigma_n^2\mathbf{I}\big)$ is defined as Gaussian distribution with a mean vector of $\mathbf{H}\mathbf{x}$ and a covariance matrix of $\sigma_n^2\mathbf{I}$, and $\mathcal{P}(\mathbf{x})$ is defined as the a priori probability of symbols. When

$\mathcal{P}(\mathbf{x}) = \prod_{i=1}^{N} \frac{1}{\sqrt{A}} \mathbb{I}_{x_i \in \mathcal{A}}$ is uniformly distributed with $\mathbb{I}_{x_i \in \mathcal{A}}$ being an indication function that takes value one if $x_i \in \mathcal{A}$ and zero otherwise, the MAP detection degrades into the maximum likelihood detection, i.e.,

$$\widehat{\mathbf{x}} = \arg\max_{\mathbf{x}} \mathcal{N}\Big(\mathbf{y} : \mathbf{H}\mathbf{x}, \sigma_n^2 \mathbf{I}\Big) \prod_{i=1}^{N} \mathbb{I}_{x_i \in \mathcal{A}}$$

The complexity of maximum likelihood detection grows up exponentially with the number of symbols N, making it prohibitive for middle- or large-scale MIMO systems with especially high-order constellations.

3. Nested Variational Chain

In order to perform low-complexity MIMO detection, one popular approach is to approximate the true posterior with another distribution that is much simpler to perform inference on, and the KL divergence is commonly used to obtain the desired distribution. Defining $\mathcal{Q}(\mathbf{x})$ as the distribution utilized to approximate the true posterior, the minimization of the 'exclusive' and 'inclusive' KL divergences can be expressed as

$$\begin{aligned} \mathcal{Q}(\mathbf{x}) &= \arg\min_{\mathcal{Q}'(\mathbf{x})} \mathrm{KL}\big(\mathcal{Q}'(\mathbf{x}) || \mathcal{P}(\mathbf{x}|\mathbf{y}, \mathbf{H})\big) \\ &= \arg\min_{\mathcal{Q}'(\mathbf{x})} \int_{\mathbf{x}} \mathcal{Q}'(\mathbf{x}) \log \frac{\mathcal{Q}'(\mathbf{x})}{\mathcal{P}(\mathbf{x}|\mathbf{y}, \mathbf{H})}, \end{aligned} \tag{4}$$

and

$$\begin{aligned} \mathcal{Q}(\mathbf{x}) &= \arg\min_{\mathcal{Q}'(\mathbf{x})} \mathrm{KL}\big(\mathcal{P}(\mathbf{x}|\mathbf{y}, \mathbf{H}) || \mathcal{Q}'(\mathbf{x})\big) \\ &= \arg\min_{\mathcal{Q}'(\mathbf{x})} \int_{\mathbf{x}} \mathcal{P}(\mathbf{x}|\mathbf{y}, \mathbf{H}) \log \frac{\mathcal{P}(\mathbf{x}|\mathbf{y}, \mathbf{H})}{\mathcal{Q}'(\mathbf{x})}. \end{aligned} \tag{5}$$

For instance, the GTA algorithm takes the former way, while the EC algorithm takes the latter one. However, with only one approximation, GTA is unable to update its approximated tree structure, while EC only treats symbols in an independent way rather than exploiting correlation among symbles. In this case, we then would like to demonstrate that the two KL divergences could be combined together, and a nested variational chain is then proposed in what follows.

Suppose there is a desired variational distribution $\mathcal{G}(\mathbf{x})$ that can be obtained with 'exclusive' KL divergengce:

$$\mathcal{G}(\mathbf{x}) = \arg\min_{\mathcal{G}'(\mathbf{x})} \mathrm{KL}\big(\mathcal{Q}(\mathbf{x}) || \mathcal{G}'(\mathbf{x})\big), \tag{6}$$

which is embedded in an optimization for $\mathcal{Q}(\mathbf{x})$ with $\mathcal{Q}(\mathbf{x})$ obtained in the first place as

$$\mathcal{Q}(\mathbf{x}) = \arg\min_{\mathcal{Q}'(\mathbf{x})} \mathrm{KL}\big(\mathcal{P}(\mathbf{x}|\mathbf{y}, \mathbf{H}) || \mathcal{Q}'(\mathbf{x})\big). \tag{7}$$

The processing above actually forms a variational chain with a nested structure given as

$$\mathcal{P}(\mathbf{x}|\mathbf{y}, \mathbf{H}) \Rightarrow \mathcal{Q}(\mathbf{x}) \Rightarrow \mathcal{G}(\mathbf{x}), \tag{8}$$

indicating $\mathcal{Q}(\mathbf{x})$ should be obtained according to the minimization of $\mathrm{KL}(\mathcal{P}(\mathbf{x}|\mathbf{y}, \mathbf{H}) || \mathcal{Q}'(\mathbf{x}))$ with respect to $\mathcal{Q}'(\mathbf{x})$, and that the desired $\mathcal{G}(\mathbf{x})$ could then be obtained by minimizing $\mathrm{KL}(\mathcal{Q}(\mathbf{x}) || \mathcal{G}'(\mathbf{x}))$ with respect to $\mathcal{G}'(\mathbf{x})$. Following the roadmap, we may derive an algorithm for the nested variational chain combining the two asymetric KL divergences.

3.1. A Generic Framework for Nested Variational Chain

To begin with, a general statistical model should first be defined as follows [12],

$$\mathcal{P}(\mathbf{x}) \propto \mathcal{F}(\mathbf{x}) \prod_i t_i(\mathbf{x}), \tag{9}$$

where $\mathcal{F}(\mathbf{x})$ is a function belonging to the exponential family, and $t_i(\mathbf{x})$ for $i = 1, \ldots, I$ are non-negative factors. Normally, it is intractable or prohibitive complex to perform inference over $\mathcal{P}(\mathbf{x})$ such that the variational inference-based approaches provide another distribution $\mathcal{Q}(\mathbf{x})$ that is tractable or easy to handle. The nested variational chain consists of four steps: factor substitution, inner approximation, symbol detection and factor updating.

As for factor substitution, the optimization of the KL divergence should first be achieved, i.e., $\mathcal{Q}(\mathbf{x}) = \arg\min_{\mathcal{Q}'(\mathbf{x})} \mathrm{KL}(\mathcal{P}(\mathbf{x})||\mathcal{Q}'(\mathbf{x}))$, for which the EC framework can be employed. The EC algorithm assumes a distribution that belongs to the exponential family:

$$\mathcal{Q}(\mathbf{x}) \propto \mathcal{F}(\mathbf{x}) \prod_i \widetilde{t}_i(\mathbf{x}), \tag{10}$$

where $\widetilde{t}_i(\mathbf{x})$ instead of $t_i(\mathbf{x})$ for $i = 1, \ldots, I$ are modified factors, belonging to the exponential family as well.

It should be noticed that the EC framework replaces each non-negative factor $t_i(\mathbf{x})$ by another $\widetilde{t}_i(\mathbf{x})$ in the exponential family. However, the distribution $\mathcal{F}(\mathbf{x})$ remains constant during this optimization process, which may be further exploited. Based on this idea, another variational distribution could be embedded inside as an inner approximation so as to achieve a final distribution $\mathcal{G}(\mathbf{x})$, and the optimization in (6) could be performed:

$$\begin{aligned} \mathcal{G}(\mathbf{x}) &= \arg\min_{\mathcal{G}'(\mathbf{x})} \mathrm{KL}\big(\mathcal{G}'(\mathbf{x})||\mathcal{Q}(\mathbf{x})\big) \\ &= \arg\min_{\mathcal{G}'(\mathbf{x})} \mathrm{KL}\left(\mathcal{G}'(\mathbf{x})||\mathcal{F}(\mathbf{x}) \prod_i \widetilde{t}_i(\mathbf{x})\right). \end{aligned} \tag{11}$$

The approximation is normally expressed as $\mathcal{G}(\mathbf{x}) = \prod_j G_j(\mathbf{x})$. When $G_j(\mathbf{x})$ for $j = 1, \ldots, J$ are defined as disjoint groups, it is mean-field approximation, and structured approximation can be employed when $G_j(\mathbf{x})$ for $j = 1, \ldots, J$ are overlapped with each other.

Toward symbol detection, a cavity distribution for each factor can then be acquired as

$$\mathcal{G}^{\setminus i}(\mathbf{x}) = \frac{\mathcal{G}(\mathbf{x})}{\widetilde{t}_i(\mathbf{x})} \tag{12}$$

and the final distribution can be represented as

$$\widetilde{\mathcal{G}}_i(\mathbf{x}) \propto \prod_{i=1}^{N} \underbrace{\mathcal{G}^{\setminus i}(\mathbf{x}) t_i(\mathbf{x})}_{p_i(\mathbf{x})}, \tag{13}$$

with $p_i(\mathbf{x}) \triangleq \mathcal{G}^{\setminus i}(\mathbf{x}) t_i(\mathbf{x})$ defined as a new distribution by attaching the true factor.

The moments of $p_i(\mathbf{x})$ are then obtained by exploiting the true distribution $t_i(\mathbf{x})$ as $\mathbb{E}_{p_i(\mathbf{x})}[\phi(\mathbf{x})]$, where $\phi(\mathbf{x})$ stands for the sufficient statistics of the exponential family. A new factor $\widetilde{t}_i^{new}(\mathbf{x})$ is updated as well by satisfying the moment-matching condition $\mathbb{E}_{p_i(\mathbf{x})}[\phi(\mathbf{x})] = \mathbb{E}_{q_i(\mathbf{x})}[\phi(\mathbf{x})]$ with

$$q_i(\mathbf{x}) \propto \mathcal{G}^{\setminus i}(\mathbf{x}) \widetilde{t}_i^{new}(\mathbf{x}), \tag{14}$$

such that the distribution $\mathcal{Q}(\mathbf{x})$ is able to be updated iteratively.

An algorithm is provided in Algorithm 1, which is used to approximate a statistical model $\mathcal{P}(\mathbf{x}) \propto \mathcal{F}(\mathbf{x}) \prod_i t_i(\mathbf{x})$. Associating with the four steps described above, the algorithm

first substites all factors in Step 1 with much easier accessible ones by using the 'inclusive' KL divergence, as seen in the EC algorithm. After that, the algorithm further approximates $\mathcal{Q}(\mathbf{x})$ with a new distribution in Step 2 such that better detection performance is expected. With detection proceeded on the new distribution, moment matching can be achieved in Step 3 so as to update substituted factors in Step 4. Note that any of the steps, such as factor substitution, inner approximation, or factor updating, may be skipped for a certain purpose so as to form a special case. In the next subsection, we would like to demonstrate that the MMSE, GTA and EC algorithms could be deemed as special cases.

Algorithm 1 An algorithm for nested variational chain

Require: A statistical model $\mathcal{P}(\mathbf{x}) \propto \mathcal{F}(\mathbf{x}) \prod_i t_i(\mathbf{x})$.
Ensure:
repeat
(1) Step 1: Factor Substitution.
Substitute each non-negative factor $t_i(\mathbf{x})$ with $\widetilde{t}_i(\mathbf{x})$ for $i = 1, \ldots, I$ such that $\mathcal{Q}(\mathbf{x}) \propto \mathcal{F}(\mathbf{x}) \prod_i \widetilde{t}_i(\mathbf{x})$.
(2) Step 2: Inner Approximation.
Obtain a new distribution to approximate $\mathcal{Q}(\mathbf{x})$ as

$$\mathcal{G}(\mathbf{x}) = \arg\min_{\mathcal{G}'(\mathbf{x})} \mathrm{KL}\big(\mathcal{G}'(\mathbf{x}) || \mathcal{Q}(\mathbf{x})\big)$$

(3) Step 3: Symbol Detection.
for $i \in [1, \ldots, I]$ **do**
Obtain a cavity distribution as $\mathcal{G}^{\setminus i}(\mathbf{x}) = \mathcal{G}(\mathbf{x}) / \widetilde{t}_i(\mathbf{x})$, and then achieve moment matching between $p_i(\mathbf{x}) \propto \mathcal{G}^{\setminus i}(\mathbf{x}) t_i(\mathbf{x})$ and $q_i(\mathbf{x}) \propto \mathcal{G}^{\setminus i}(\mathbf{x}) \widetilde{t}_i^{new}(\mathbf{x})$.
end for
(4) Step 4: Factor Updating.
Substitute $\widetilde{t}_i(\mathbf{x})$ with $\widetilde{t}_i^{new}(\mathbf{x})$ into $\mathcal{Q}(\mathbf{x}) \propto \mathcal{F}(\mathbf{x}) \prod_i \widetilde{t}_i(\mathbf{x})$ and repeat this procedure if necessary.
until Convergence is achieved.
Output: Detection results on the approximated distribution.

3.2. MMSE, GTA and EC MIMO Detectors as Special Cases

In a MIMO system, the distribution $\mathcal{F}(\mathbf{x})$ can be expressed as the likelihood function, i.e., $\mathcal{F}(\mathbf{x}) \propto \mathcal{N}(\mathbf{y} : \mathbf{H}\mathbf{x}, \sigma_n^2 \mathbf{I})$, and each non-negative factor $t_i(\mathbf{x})$ for $i = 1, \ldots, I$ could be regarded as the *apriori* probability with respect to symbols. When a factor $t_i(\mathbf{x})$ corresponds only to one symbol x_i, it reduces to $t_i(x_i) = \mathcal{P}(x_i) = \frac{1}{\sqrt{A}} \mathbb{I}_{x_i \in \mathcal{A}}$. Hence, as there are N symbols in a MIMO system, there would be N factors or priors as well, and the expression for the substituted factor $\widetilde{t}_i(x_i)$ for $i = 1, \ldots, N$ depends on any specific algorithm.

(1) Minimum Mean Square Error

The MMSE approach could be obtained by assuming that each non-negative factor $t_i(x_i)$ for $i = 1, \ldots, N$ can be replaced by a Gaussian distributed factor $\widetilde{t}_i(x_i) = \mathcal{N}(x_i : 0, E_s)$ of zero-mean and a variance of E_s, and the modified distribution with factor substitution for MMSE is then given as

$$\mathcal{Q}_{MMSE}(\mathbf{x}) \propto \mathcal{N}\Big(\mathbf{y} : \mathbf{H}\mathbf{x}, \sigma_n^2 \mathbf{I}\Big) \prod_{i=1}^{N} \mathcal{N}(x_i : 0, E_s), \tag{15}$$

whose second-order and first-order moments are derived as

$$\begin{cases} \boldsymbol{\Sigma}_{MMSE} = \left(\mathbf{H}^\top \mathbf{H} + \dfrac{\sigma_n^2}{E_s} \mathbf{I}\right)^{-1} \\ \boldsymbol{\mu}_{MMSE} = \boldsymbol{\Sigma}_{MMSE} \mathbf{H}^\top \mathbf{y} \end{cases}$$

Not mentioned though before, there is actually a simple inner approximation for MMSE to approximate the distribution $\mathcal{Q}_{MMSE}(\mathbf{x})$. With a fully factorized distribution $\mathcal{G}_{MMSE}(\mathbf{x}) = \prod_i^N G_{MMSE}(x_i)$, each factorized one can be obtained as

$$\begin{aligned} G_{MMSE}(x_i) &\propto \exp\Big(\langle \ln \mathcal{Q}_{MMSE}(\mathbf{x})\rangle_{\sim G_{MMSE}(x_i)}\Big) \\ &\propto \mathcal{N}(x_i : \mu_{i,MMSE}, \Sigma_{i,MMSE}), \end{aligned} \tag{16}$$

which is known as the mean-field approximation. The expression $\langle\cdot\rangle_{\sim G_{MMSE}(x_i)}$ refers to expectation with respect to all factors $G_{MMSE}(x_j)$ for $j = 1, \ldots, N$ except for $G_{MMSE}(x_i)$. This process is equivalent to marginalization of $\mathcal{Q}_{\text{MMSE}}(\mathbf{x})$ with $\mu_{i,MMSE}$ and $\Sigma_{i,MMSE}$ being the i^{th} element of $\boldsymbol{\mu}_{\text{MMSE}}$ and of the diagonal of $\boldsymbol{\Sigma}_{MMSE}$.

The MMSE approach skips factor updating, but instead it may output directly the hard detection results. The final distribution of MMSE is expressed as

$$\widetilde{\mathcal{G}}_{MMSE}(\mathbf{x}) \propto \prod_{i=1}^{N} \underbrace{G_{MMSE}(x_i) t_i(x_i)}_{p_{MMSE,i}(x_i)}, \tag{17}$$

where $p_{MMSE,i}(x_i) \triangleq G_{MMSE}(x_i) t_i(x_i)$ is defined as a new distribution by attaching true priors, based on which symbol detection can be proceeded for each symbol independently.

(2) Expectation Consistency

The EC algorithm defines a substitution factor for each symbol as well. It replaces the prior $t_i(x_i) = \frac{1}{\sqrt{A}}\mathbb{I}_{x_i \in \mathcal{A}}$ to $\widetilde{t}_i(x_i) \propto e^{\gamma_i x_i - \frac{1}{2}\Lambda_i x_i^2}$ so the posterior can be expressed as

$$\mathcal{Q}_{EC}(\mathbf{x}) \propto \mathcal{N}\Big(\mathbf{y} : \mathbf{H}\mathbf{x}, \sigma_n^2 \mathbf{I}\Big) \prod_{i=1}^{N} e^{\gamma_i x_i - \frac{1}{2}\Lambda_i x_i^2} \tag{18}$$

Note that $\widetilde{t}_i(x_i) \propto e^{\gamma_i x_i - \frac{1}{2}\Lambda_i x_i^2}$ is Gaussian distributed. In this regard, it can be noticed that EC relates essentially to MMSE with the difference that it is able to update priors. The second-order and first-order moments of $\mathcal{Q}_{\text{EC}}(\mathbf{x})$ are derived as

$$\begin{cases} \boldsymbol{\Sigma}_{EC} = \Big(\sigma_n^{-2}\mathbf{H}^\top \mathbf{H} + \boldsymbol{\Lambda}\Big)^{-1} \\ \boldsymbol{\mu}_{EC} = \boldsymbol{\Sigma}_{EC}\Big(\mathbf{H}^\top \mathbf{y} + \boldsymbol{\gamma}\Big) \end{cases}$$

where $\boldsymbol{\Lambda}$ is a diagonal matrix containing Λ_i, and $\boldsymbol{\gamma}$ is a vector containing γ_i for $i = 1, \ldots, N$.

The EC algorithm employs mean-field approximation for inner approximation as well, by which the fully factorized distribution is defined as $\mathcal{G}_{EC}(\mathbf{x}) = \prod_i^N G_{EC}(x_i)$, and each factorized distribution $G_{EC}(x_i)$ is Gaussian distributed such that:

$$\begin{aligned} G_{EC}(x_i) &\propto \exp\Big(\langle \ln \mathcal{Q}_{EC}(\mathbf{x})\rangle_{\sim G_{EC}(x_i)}\Big) \\ &\propto \mathcal{N}(x_i : \mu_{i,EC}, \Sigma_{i,EC}), \end{aligned} \tag{19}$$

with $\mu_{i,EC}$ and $\Sigma_{i,EC}$ being the i^{th} element of $\boldsymbol{\mu}_{EC}$ and of the diagonal of $\boldsymbol{\Sigma}_{EC}$, respectively. By doing so, factor updating is then operated with a cavity distribution:

$$G_{EC}^{\setminus i}(x_i) = \frac{G_{EC}(x_i)}{\widetilde{t}_i(x_i)}, \tag{20}$$

and the final distribution for EC is represented as

$$\widetilde{\mathcal{G}}_{EC}(\mathbf{x}) \propto \prod_{i=1}^{N} \underbrace{G_{EC}^{\setminus i}(x_i) t_i(x_i)}_{p_{EC,i}(x_i)}, \tag{21}$$

where $p_{EC,i}(x_i) \triangleq G_{EC}^{\setminus i}(x_i)t_i(x_i)$ is defined as a new distribution. Symbol detection can then be performed to achieve the moment-matching condition so the pairs (γ_i, Λ_i) for $i = 1, \ldots, N$ are updated in parallel.

(3) Gaussian Tree Approximation

The GTA algorithm was proposed based on the modified distribution of MMSE, and its distribution with substituted factors can be represented as

$$\begin{aligned} \mathcal{Q}_{GTA}(\mathbf{x}) = \mathcal{Q}_{MMSE}(\mathbf{x}) &\propto \mathcal{F}(\mathbf{x}) \prod_{i=1}^{N} \widetilde{t}_i(x_i) \\ &\propto \mathcal{N}\left(\mathbf{y} : \mathbf{H}\mathbf{x}, \sigma_n^2 \mathbf{I}\right) \prod_{i=1}^{N} \mathcal{N}(x_i : 0, E_s). \end{aligned} \tag{22}$$

As for inner approximation, the GTA algorithm chooses to optimally approximate the distribution with a tree graph, which can be constructed based on $\mathcal{Q}_{GTA}(\mathbf{x})$ as

$$\begin{aligned} \mathcal{G}_{GTA}(\mathbf{x}) &= \arg\min_{\mathcal{G}'(\mathbf{x})} \mathrm{KL}\big(\mathcal{G}'(\mathbf{x}) || \mathcal{Q}_{GTA}(\mathbf{x})\big) \\ &= \prod_i G_{GTA}\left(x_i | x_{pa(i)}\right), \end{aligned} \tag{23}$$

where $G_{GTA}(x_i|x_{pa(i)})$ stands for the conditional probability of x_i given its parent $x_{pa(i)}$, and $G_{GTA}(x_i|x_{pa(i)}) = G_{GTA}(x_i)$ in case that x_i is the root of the tree.

This leads to a result that GTA skips factor updating as well, similar to MMSE, and the performance of the GTA algorithm is subject to the fixed initial distribution $\mathcal{Q}_{GTA}(\mathbf{x})$ that is not able to be updated. In this case, by directly attaching the true priors $t_i(x_i) = \frac{1}{\sqrt{A}} \mathbb{I}_{x_i \in \mathcal{A}}$ for $i = 1, \ldots, N$, the final distribution of GTA is then represented as

$$\begin{aligned} \widetilde{\mathcal{G}}_{GTA}(\mathbf{x}) &\propto \prod_{i=1}^{N} G_{GTA}\left(x_i | x_{pa(i)}\right) \prod_{i=1}^{N} t_i(x_i) \\ &= \prod_{i=1}^{N} \underbrace{G_{GTA}\left(x_i | x_{pa(i)}\right) t_i(x_i)}_{p_{GTA,i}(x_i)}. \end{aligned} \tag{24}$$

where $p_{GTA,i}(x_i) \triangleq G_{GTA}(x_i)t_i(x_i)$. Proceeding on such a loop-free tree graph, message passing can then be utilized to perform efficient detection during all but one iteration.

4. Applications into MIMO High-Order Detection

Introducing the nested variational chain for MIMO detection, it can be seen that all existing approaches employ factor substitution. As for inner approximation, MMSE and EC actually perform mean-field approximation with fully factorized distribution, while GTA performs the maximum spanning tree approximation. Finally, only EC performs factor updating, while MMSE and GTA choose to perform direct detection.

This analysis puts forward the question of whether any improvement can be achieved when one enables GTA to update its substituted factors or whether any better inner approximation can be derived for EC rather than being fully factorized. Both thoughts lead us to an idea that it is worth trying to update the GTA factors iteratively since the approximated Gaussian tree is capable of capturing correlation among symbols rather than keeping independence among them. Following this idea, an initial application of the nested variational chain can be performed. By utilizing EC as an outer approximation, an algorithm named GTA-embedded EC (GTA-EC) is proposed in the following.

4.1. The GTA-EC Algorithm

Given $t_i(x_i) \propto \mathbb{I}_{x_i \in \mathcal{A}}$ for $i = 1, \ldots, N$, the algorithm starts from the likelihood function with discrete priors as in (3), i.e.,

$$\mathcal{Q}_{GTA-EC}(\mathbf{x}) \propto \mathcal{N}\left(\mathbf{y} : \mathbf{H}\mathbf{x}, \sigma_n^2\mathbf{I}\right) \prod_{i=1}^{N} \mathbb{I}_{x_i \in \mathcal{A}}, \tag{25}$$

which could be divided into two parts, i.e.,

$$f_q(\mathbf{x}) = \mathcal{N}\left(\mathbf{y} : \mathbf{H}\mathbf{x}, \sigma_n^2\mathbf{I}\right), \tag{26}$$

$$f_r(\mathbf{x}) = \prod_{i=1}^{N} \mathbb{I}_{x_i \in \mathcal{A}} \tag{27}$$

It is then possible to define a new distribution $q(\mathbf{x})$ as

$$q(\mathbf{x}) \propto f_q(\mathbf{x}) \exp\left(\gamma_q^\top \mathbf{x} - \frac{\mathbf{x}^\top \mathbf{\Lambda}_q \mathbf{x}}{2}\right), \tag{28}$$

of which the moments can be expressed as

$$\begin{cases} \mathbf{\Sigma}_q = \left(\mathbf{H}^\top \mathbf{H} + \mathbf{\Lambda}_q\right)^{-1} \\ \boldsymbol{\mu}_q = \mathbf{\Sigma}_q\left(\sigma_n^{-2}\mathbf{H}^\top \mathbf{y} + \gamma_q\right) \end{cases}$$

Note that the pair $(\gamma_q, \mathbf{\Sigma}_q)$ acts as priors of all symbols to be updated, and that the definition of $q(\mathbf{x})$ actually serves as factor substitution.

To achieve moment consistency, another distribution $s(\mathbf{x})$ is then defined as

$$s(\mathbf{x}) \propto \exp\left(\gamma_s^\top \mathbf{x} - \frac{\mathbf{x}^\top \mathbf{\Lambda}_s \mathbf{x}}{2}\right), \tag{29}$$

where moment matching between $s(\mathbf{x})$ and $q(\mathbf{x})$ should be achieved so as to obtain γ_s and $\mathbf{\Lambda}_s$. The EC algorithm assumes another distribution:

$$r(\mathbf{x}) \propto \exp\left(\gamma_r^\top \mathbf{x} - \frac{\mathbf{x}^\top \mathbf{\Sigma}_r \mathbf{x}}{2}\right) f_r(\mathbf{x}), \tag{30}$$

with moments derived as

$$\begin{cases} \gamma_r = \gamma_s - \gamma_q \\ \mathbf{\Lambda}_r = \mathbf{\Lambda}_s - \mathbf{\Lambda}_q \end{cases}$$

It can be observed that $\exp\left(\gamma_r^\top \mathbf{x} - \frac{\mathbf{x}^\top \mathbf{\Sigma}_r \mathbf{x}}{2}\right)$ partly in $r(\mathbf{x})$ actually serves as a cavity distribution of symbols by subtracting their substituted priors $(\gamma_q, \mathbf{\Sigma}_q)$.

The next step involves inner approximation. Since fully factorization for $r(\mathbf{x})$ neglects correlation among symbols, we instead propose utilizing the Gaussian approximation tree to perform detection according to the moments $\boldsymbol{\mu}_r$, $\mathbf{\Sigma}_r$, $\boldsymbol{\mu}_q$, and $\mathbf{\Sigma}_q$. This is because the Gaussian approximation tree may capture correlation among symbols rather than treating them independently. In this case, we define a new Gaussian tree-based distribution $g(\mathbf{x})$ rather than $r(\mathbf{x})$ as

$$g(\mathbf{x}) \propto f_r(\mathbf{x}) \prod_{i=1}^{N} G^{\setminus i}\left(x_i | x_{pa(i)}\right) = \prod_{i=1}^{N} \underbrace{G^{\setminus i}\left(x_i | x_{pa(i)}\right) \mathbb{I}_{x_i \in \mathcal{A}}}_{p'_{i|pa(i)}(x_i)} \tag{31}$$

where $p'_{i|pa(i)} \triangleq G^{\setminus i}\left(x_i | x_{pa(i)}\right) \mathbb{I}_{x_i \in \mathcal{A}}$ is a new distribution by attaching true priors, and the conditional distribution can be represented as

$$G^{\backslash i}\left(x_i|x_{pa(i)}\right) \propto exp\left\{-\frac{1}{2}\frac{\left[(x_i-\mu_i^r)-\frac{\Sigma_{i,pa(i)}}{\Sigma_{pa(i),pa(i)}}\left(x_{pa(i)}-\mu_{pa(i)}\right)\right]^2}{\Sigma_{i,i}^r-\frac{\Sigma_{i,pa(i)}}{\Sigma_{pa(i),pa(i)}}}\right\} \tag{32}$$

where μ_i^r and $\Sigma_{i,i}^r$ for $i=1,\ldots,N$ are taken from $\boldsymbol{\mu}_r$ and the diagonal of $\boldsymbol{\Sigma}_r$, respectively, while μ_i and $\Sigma_{i,i}$ for $i=1,\ldots,N$ are taken from $\boldsymbol{\mu}_q$ and the diagonal of $\boldsymbol{\Sigma}_q$.

Based on $g(\mathbf{x})$, message passing on the Gaussian tree can then be proceeded:

$$\mathcal{M}_{i\to pa(i)}\left(x_{pa(i)}\right)=\sum_{\sim x_i}\underbrace{G^{\backslash i}\left(x_i|x_{pa(i)}\right)t_i(x_i)}_{p_{i|pa(i)}(x_i)}\prod_{j|pa(j)=i}\mathcal{M}_{j\to i}(x_i) \tag{33}$$

and

$$\begin{aligned}\mathcal{M}_{pa(i)\to i}(x_i)=&\sum_{\sim x_{pa(i)}}\underbrace{G^{\backslash i}\left(x_i|x_{pa(i)}\right)t_i(x_i)}_{p_{i|pa(i)}(x_i)}\\&\times\prod_{j|j\neq i,pa(j)=pa(i)}\mathcal{M}_{j\to p(i)}\left(x_{pa(i)}\right)\mathcal{M}_{pa(pa(i))\to pa(i)}\left(x_{pa(i)}\right)\end{aligned} \tag{34}$$

To achieve consistency, the distribution $s(\mathbf{x})$ is finally utilized once again to achieve moment matching between $g(\mathbf{x})$ and $s(\mathbf{x})$ so as to obtain $\boldsymbol{\gamma}_s$ and $\boldsymbol{\Sigma}_s$, and the a priori moments can be updated:

$$\begin{cases}\boldsymbol{\Lambda}_q^{new}=\beta(\boldsymbol{\Lambda}_s-\boldsymbol{\Lambda}_r)+(1-\beta)\boldsymbol{\Lambda}_q\\ \boldsymbol{\gamma}_q^{new}=\beta(\boldsymbol{\gamma}_s-\boldsymbol{\gamma}_r)\ +(1-\beta)\boldsymbol{\gamma}_q\end{cases} \tag{35}$$

The GTA-EC algorithm is concluded and depicted in detail in Algorithm 2. In step 1, the GTA-EC algorithm initiliazes the distribution $q(\mathbf{x})$, which behaves as an outer approximation by substituting true factors. In step 2, the inner approximation is applied to $q(\mathbf{x})$ by using its moments, such that a maximum spanning tree is constructed. With the derived tree structure, the algorithm repeats step 3 and step 4 over iterations such that factors can be updated by performing symbol detection and moment matching, and hard outputs can then be obtained according to the final distribution.

Algorithm 2 The GTA-EC Algorithm

Require: $\mathbf{y}$, $\mathbf{H}$, E_s and σ_n^2. Initialize $\mathcal{M}_{i\to pa(i)}(x_{pa(i)})=1/\sqrt{A}$, $\mathcal{M}_{pa(i)\to i}(x_i)=1/\sqrt{A}$, $\gamma_i=0$ and $\Lambda_i=E_s^{-1}$ for $i=1,\ldots,N$.
Ensure:
(1) Step 1: Factor Substitution.
Initial $q(\mathbf{x})\propto f_q(\mathbf{x})\exp\left(\boldsymbol{\gamma}_q^\top\mathbf{x}-\frac{1}{2}\mathbf{x}^\top\boldsymbol{\Lambda}_q\mathbf{x}\right)$.
(2) Step 2: Inner Approximation.
The maximum Gaussian spanning tree is constructed according to the initial covariance matrix such that the tree structure and relationship among symbols can be obtained.
repeat
(3) Step 3: Symbol Detection.
Obtain $r(\mathbf{x})$ by achieving consistency between $q(\mathbf{x})$ and $s(\mathbf{x})$, and obtain $g(\mathbf{x})$ according to the established tree structure and derived moments. Perform message passing in updating $\mathcal{M}_{i\to pa(i)}(x_{pa(i)})$ and $\mathcal{M}_{pa(i)\to i}(x_i)$ according to (33) and (34), and obtain the *aposteriori* statistics by achieving consistency between $g(\mathbf{x})$ and $s(\mathbf{x})$.
(4) Step 4: Factor Updating.
Update $\boldsymbol{\gamma}_q^{new}$ and $\boldsymbol{\Lambda}_q^{new}$ such that $q^{new}(\mathbf{x})$ can be updated.
until A maximum number of iterations has been achieved.
Output: Hard outputs according to the first-order moments of the latest $q^{new}(\mathbf{x})$.

4.2. Complexity Analysis

The calculation of GTA-EC resides mainly on three parts. The first one involves the factor substitution step, which necessitates the calculation of second-order and first-order moments in (29), the same as MMSE in (16) or EC in (19). As is well known, its complexity in one iteration can be given as $\mathcal{O}(NM^2)$. The second part involves construction of the tree graph for inner approximation, which needs only to be initialized at the very beginning of iterations. The construction is based on Prim's algorithm, whose complexity is $\mathcal{O}(M^2)$. The last part involves the calculation of message passing and factor updating. For each iteration, the major complexity lies in calculating messages in (33) and (34), each requiring the maximum likelihood detection on the conditional distribution with the cardinality of PAM constellation being $|\mathcal{A}| = \sqrt{A}$. Since there are $M-1$ conditional distributions in the tree graph, the complexity can be represented as $\mathcal{O}\left(M|\mathcal{A}|^2\right) = \mathcal{O}(MA)$. Therefore, by defining N_{iter} as the number of iterations to proceed, the total complexity can be expressed as $\mathcal{O}\left((N_{iter}+1)NM^2 + M^2 + N_{iter}MA\right) \approx \mathcal{O}\left((N_{iter}+1)NM^2\right)$ due to the reason that $NM^2 \gg MA$ is normally satisfied in a massive MIMO system. This indicates that the complexity of GTA-EC is about N_{iter} times more than that of MMSE or GTA, namely $\mathcal{O}\left(NM^2\right)$. As a comparison, the complexity of EC can be expressed as $\mathcal{O}((N_{iter}+1)NM^2 + M + N_{iter}M\sqrt{A}) \approx \mathcal{O}\left((N_{iter}+1)NM^2\right)$, suggesting that the complexity of GTA-EC is approximately in the same order. The less iterations one algorithm needs to perform, the less complexity it requires. In the next section, when comparing the performance of GTA-EC with EC, the number of iterations should be utilized for complexity comparison. A summary of complexity comparison is demonstrated in Table 1, in which it can be found that the total complexity is dominated by the complexity of factor substitution as well as the number of iterations.

Table 1. Comparisons of complexity.

Algorithm	Factor Substitution	Inner Approximation	Detection and Factor Updating	Total Complexity
MMSE	$\mathcal{O}\left(NM^2\right)$	$\mathcal{O}(M)$	$\mathcal{O}\left(M\sqrt{A}\right)$	$\mathcal{O}\left(NM^2\right)$
GTA	$\mathcal{O}\left(NM^2\right)$	$\mathcal{O}\left(M^2\right)$	$\mathcal{O}(MA)$	$\mathcal{O}\left(NM^2\right)$
EC	$\mathcal{O}\left((N_{iter}+1)NM^2\right)$	$\mathcal{O}(M)$	$\mathcal{O}\left(N_{iter}M\sqrt{A}\right)$	$\mathcal{O}\left((N_{iter}+1)NM^2\right)$
GTA-EC	$\mathcal{O}\left((N_{iter}+1)NM^2\right)$	$\mathcal{O}\left(M^2\right)$	$\mathcal{O}(N_{iter}MA)$	$\mathcal{O}\left((N_{iter}+1)NM^2\right)$

5. Numerical Results

5.1. Simulation Parameters

In this section, the detection performance of a MIMO system is evaluated in terms of bit error rate (BER). Uncorrelated scattering flat-fading channel model is assumed with channel coefficients being modeled as complex Gaussian distributed variables that are independently generated for all antennas. During the simulation, 20,000 realizations of the channel matrix are employed with each used to send one message. As a comparison, several existing algorithm are evaluated as well such as the MMSE, GTA, and EC algorithms. And we mainly take into consideration the 'worst-case' scenarios of load $\alpha = N/M = 1$ when $N = M = 16$ and $N = M = 64$ with high-order constellations 16-QAM, 64-QAM, and 256-QAM considered. The factor β is set as 0.2 for all algorithms, and the iteration number of EC and GTA-EC is set as 2, 4, and 6 since convergence can be achieved within *six* iterations.

5.2. Performance Evaluation

Figures 1–3 demonstrate the BER comparison of the GTA-EC algorithm with existing algorithms. The number of antennas deployed at both the transmitter and receiver in the system is set as $N = M = 16$ with the constellations being 16-QAM, 64-QAM and 256-QAM, respectively. It can be found that GTA-EC outperforms EC with the same

number of iterations, and that GTA-EC with *four* iterations outperforms EC with *six* iterations, indicating that GTA-EC may achieve better performances than EC does with lower complexity. While in Figures 2 and 3, GTA-EC with *two* iterations almost exibits better performance than EC with *six* iterations, revealing better performance gain when high-order constellations are employed. One can further obseve that the BER slopes of GTA-EC decrease faster than that of EC, demonstrating that superior divergence gain can also be obtained by GTA-EC in a high SNR regime.

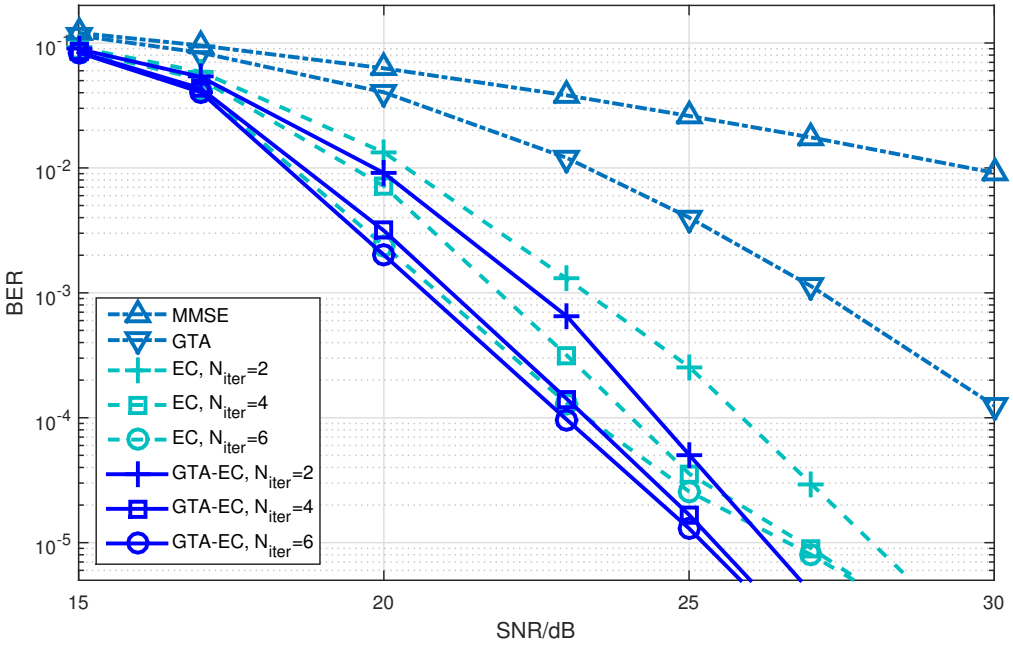

Figure 1. BER comparison of GTA-EC with existing algorithms when $N = M = 16$ with 16-QAM.

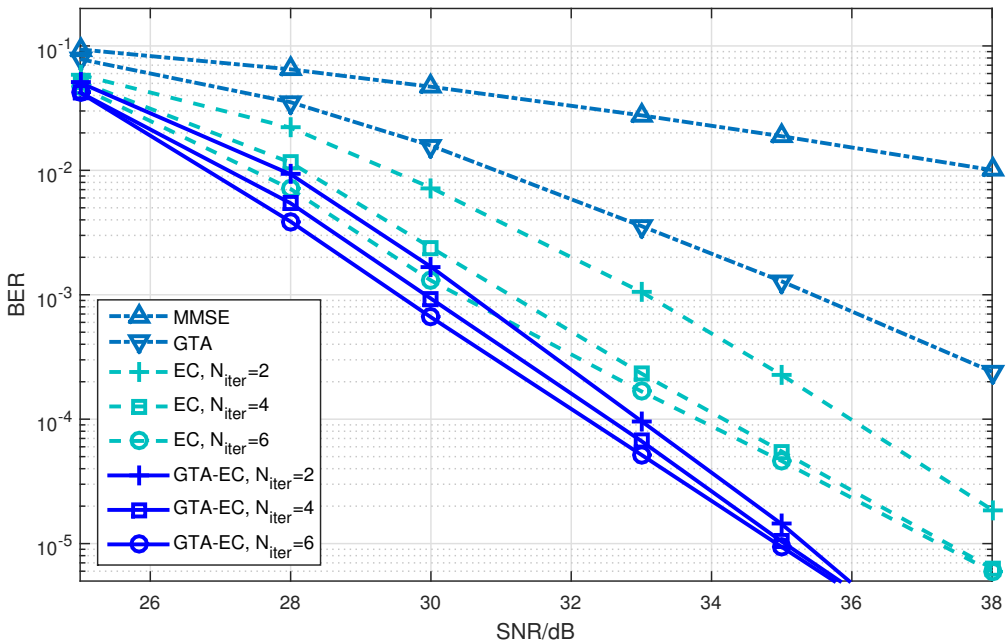

Figure 2. BER comparison of GTA-EC with existing algorithms when $N = M = 16$ with 64-QAM.

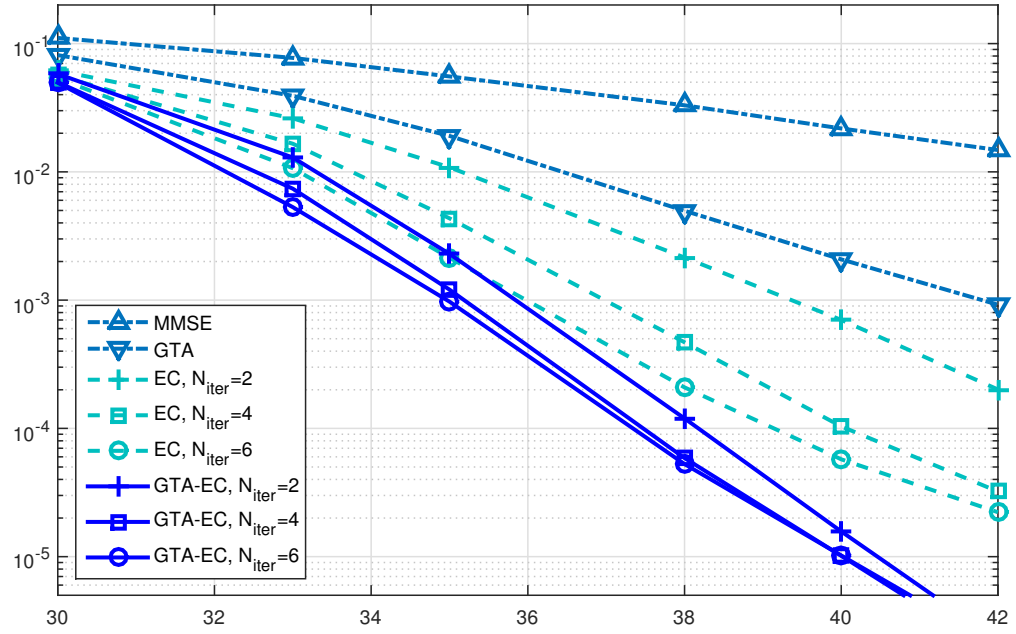

Figure 3. BER comparison of GTA-EC with existing algorithms when $N = M = 16$ with 256-QAM.

Figures 4–6 demonstrate a BER comparison of the GTA-EC algorithm with existing algorithms. The number of antennas deployed at both the transmitter and receiver is given as $N = M = 64$ with the constellations being 16-QAM, 64-QAM and 256-QAM, respectively. In these figures, it can be found that GTA-EC outperforms EC with the same number of iterations, while GTA-EC with $four$ iterations may have similar performance to that of EC with six iterations. This indicates that GTA-EC exhibits better performance than EC does at the same order of complexity or that GTA-EC presents similar performance to that of EC with lower complexity. And in Figures 4 and 5, one can further observe that the BER slope of GTA-EC decreases faster than that of EC, leading to better performance in a high SNR regime.

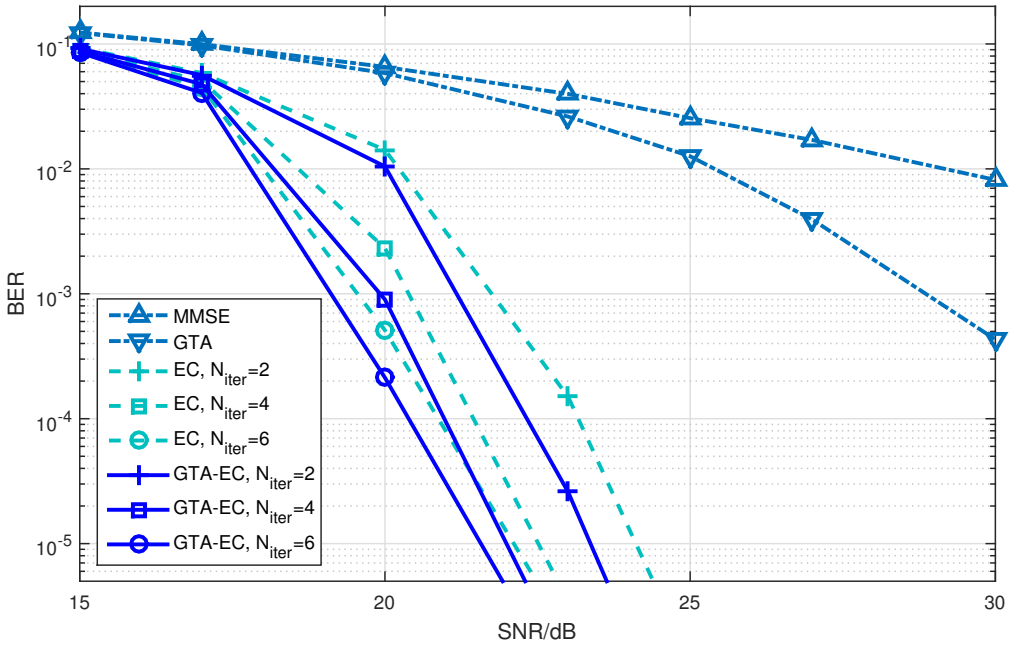

Figure 4. BER comparison of GTA-EC with existing algorithms when $N = M = 64$ with 16-QAM.

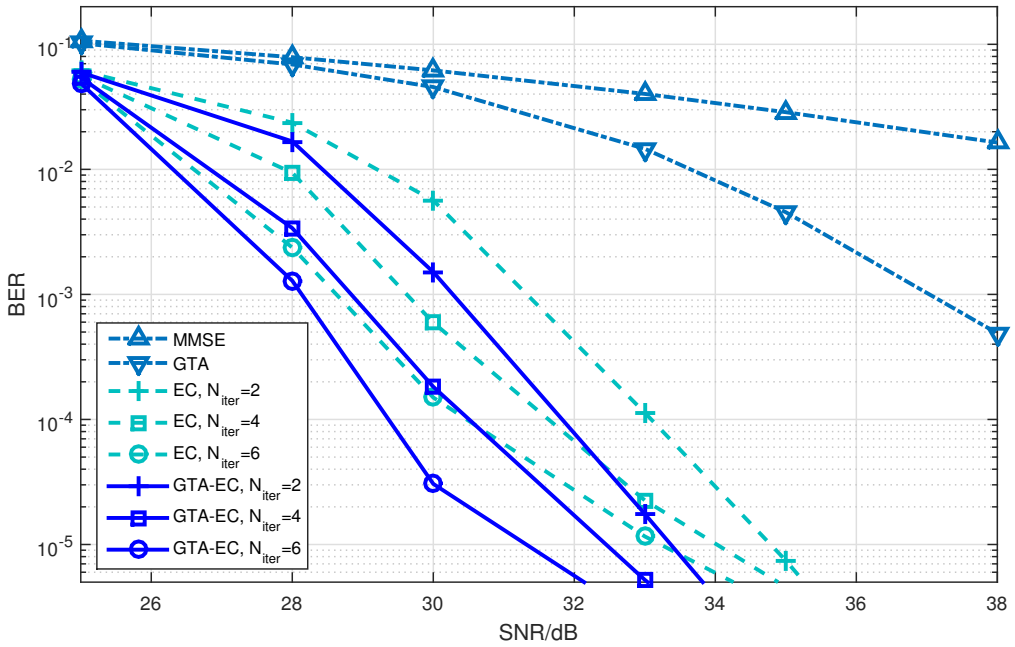

Figure 5. BER comparison of GTA-EC with existing algorithms when $N = M = 64$ with 64-QAM.

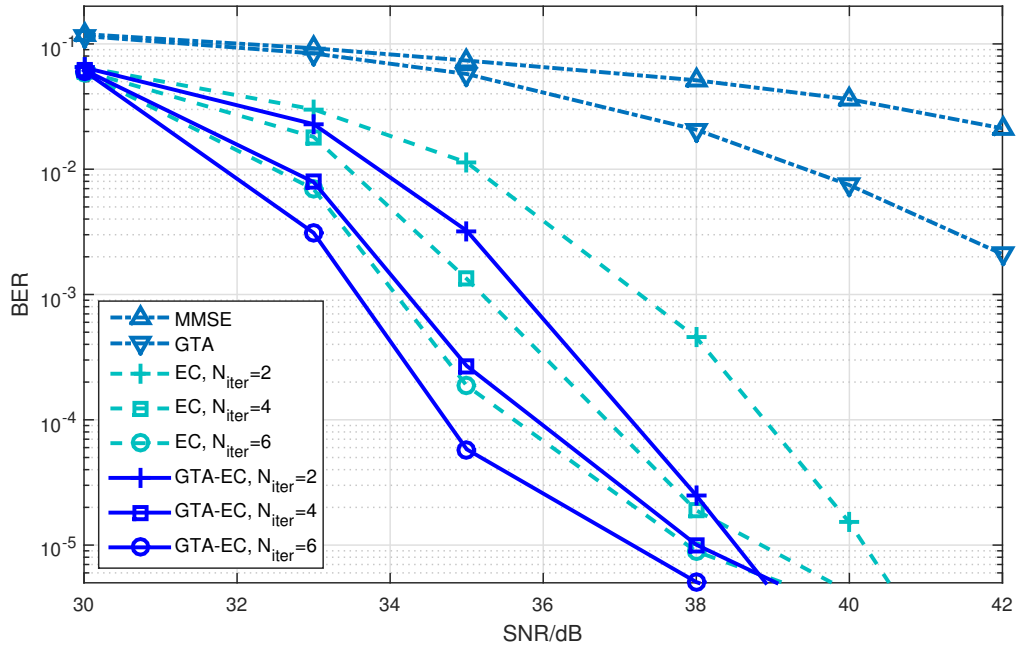

Figure 6. BER comparison of GTA-EC with existing algorithms when $N = M = 64$ with 256-QAM.

By observing and analyzing the figures in different scenarios, we may come to conclusions about the performance comparison of GTA-EC with existing algorithms.

- On one hand, both EC and GTA-EC significantly outperform existing algorithms such as MMSE and GTA. In most scenarios, GTA-EC can obviously outperform EC with either 16-QAM, 64-QAM, or 256-QAM employed. The performance gain of GTA-EC becomes larger when high-order constellation is employed. For example, both the 64-QAM and 256-QAM cases exhibit larger gain than the 16-QAM case when employing 16 or 64 antennas. This indicates that GTA-EC has superior performance gain and is especially suitable for high-order constellations. We believe that the performance gain comes from exploiting additonal relations (correlation) among symbols rather than treating them independently.

- On the other hand, as for the complexity issue, GTA-EC with $four$ iterations may outperform or have comparable performance to EC with six iterations, suggesting that GTA-EC requires less complexity than EC by recalling that their computational burdens are dominated by the number of iterations needed. As a result, $four$ iterations are recommended for GTA-EC according to the simulation results, and hence the complexity of GTA-EC is approximately $four$ times more than MMSE, indicating that it is a practical method for massive MIMO systems.

6. Conclusions

A nested variational chain is proposed along with an algorithm provided, which combines two asymmetic KL divergences. Introduced into MIMO systems, it can be found that several existing algorithms such as MMSE, GTA, and EC can be regarded as special cases. As initial applications for MIMO detection, an algorithm named GTA-EC is proposed with complexity analysis, and numerical results prove that it may achieve better detection performance with less complexity compared to existing algorithms. As for further research topics, it is suggested that one can find better inner approximation that may capture much more correlation among symbols by applying this framework to other detection fields, such as space code multiple access (SCMA), orthogonal time frequency space (OTFS), or low-density parity check (LDPC) decoding systems.

Funding: This research was funded by the National Natural Science Foundation of China, grant number 61801352 and 62371363.

Data Availability Statement: Data are contained within the article.

Conflicts of Interest: The author declares no conflict of interest.

References

1. Yang, S.; Hanzo, L. Fifty Years of MIMO Detection: The Road to Large-Scale MIMOs. *IEEE Commun. Surv. Tutor.* **2015**, *17*, 1941–1988. [CrossRef]
2. Albreem, M.A.; Juntti, M.; Shahabuddin, S. Massive MIMO Detection Techniques: A Survey. *IEEE Commun. Surv. Tutorials* **2019**, *21*, 3109–3132. [CrossRef]
3. Chen, J.-C. A low complexity data detection algorithm for uplink multiuser massive MIMO systems. *IEEE J. Selecive Areas Commun.* **2017**, *35*, 1701–1714. [CrossRef]
4. Caire, G.; Muller, R.; Tanaka, T. Iterative multiuser joint decoding: Optimal power allocation and low-complexity implementation. *IEEE Trans. Inf. Theory* **2004**, *50*, 1950–1973. [CrossRef]
5. Liu, T.; Liu, Y.-L. Modified fast recursive algorithm for efficient MMSE-SIC detection of the V-BLAST system. *IEEE Trans. Wirel. Commun.* **2008**, *7*, 3713–3717.
6. Albreem, M.A.M.; Salleh, M.F.M. Radius selection for lattice sphere decoder-based block data transmission systems. *Wirel. Netw.* **2016**, *22*, 655–662. [CrossRef]
7. Cui, T.; Han, S.; Tellambura, C. Probability-distribution-based node pruning for sphere decoding. *IEEE Trans. Veh. Technol.* **2013**, *62*, 1586–1596. [CrossRef]
8. Chockalingam, A. Low-complexity algorithms for large-MIMO detection. In Proceedings of the 2010 4th International Symposium on Communications, Control and Signal Processing (ISCCSP), Limassol, Cyprus, 3–5 March 2010; pp. 1–6.
9. Pereira, A.A., Jr.; Sampaio-Neto, R. A random-list based LAS algorithm for near-optimal detection in large-scale uplink multiuser MIMO systems. In Proceedings of the 19th International ITG Workshop on Smart Antennas, Ilmenau, Germany, 3–5 March 2015; pp. 1–5.
10. Datta, T.; Srinidhi, N.; Chockalingam, A.; Rajan, B.S. Random restart reactive Tabu search algorithm for detection in large-MIMO systems. *IEEE Commun. Lett.* **2010**, *14*, 1107–1109. [CrossRef]
11. Srinidhi, N.; Datta, T.; Chockalingam, A.; Rajan, B.S. Layered Tabu search algorithm for large-MIMO detection and a lower bound on ML performance. *IEEE Trans. Commun.* **2010**, *59*, 2955–2963. [CrossRef]
12. Bai, L.; Li, T.; Liu, J.; Yu, Q.; Choi, J. Large-scale MIMO detection using MCMC approach with blockwise sampling. *IEEE Trans. Commun.* **2016**, *64*, 3697–3707. [CrossRef]
13. Goldberger, J.; Leshem, A. MIMO detection for high-order QAM based on a Gaussian tree approximation. *IEEE Trans. Inf. Theory* **2011**, *57*, 4973–4982. [CrossRef]
14. Goldberger, J. Improved MIMO detection based on successive tree approximations. In Proceedings of the 2013 IEEE International Symposium on Information Theory, Istanbul, Turkey, 7–12 July 2013; pp. 2004–2008.

15. Céspedes, J.; Olmos, P.M.; Sánchez-Fernández, M.; Perez-Cruz, F. Expectation propagation detection for high-order high-dimensional MIMO systems. *IEEE Trans. Commun.* **2014**, *62*, 2840–2849. [CrossRef]
16. Céspedes, J.; Olmos , P.M.; Sánchez-Fernández, M.; Perez-Cruz, F. Probabilistic MIMO Symbol Detection With Expectation Consistency Approximate Inference. *IEEE Trans. Veh. Technol.* **2018**, *67*, 3481–3494. [CrossRef]
17. Tan, X.; Ueng, Y.; Zhang, Z.; You, X.; Zhang, C. A Low-Complexity Massive MIMO Detection Based on Approximate Expectation Propagation. *IEEE Trans. Veh. Technol.* **2019**, *68*, 7260–7272. [CrossRef]
18. Ge, Y.; Tan, X.; Ji, Z.; Zhang, Z.; You, X.; Zhang, C. Improving Approximate Expectation Propagation Massive MIMO Detector With Deep Learning. *IEEE Wirel. Commun. Lett.* **2021**, *10*, 2145–2149. [CrossRef]
19. Fuentes, J.J.; Santos, I.; Aradillas, J.C.; Sánchez-Fernández, M. A Low-Complexity Double EP-based Detector for Iterative Detection and Decoding in MIMO. *IEEE Trans. Commun.* **2021**, *69*, 1538–1547. [CrossRef]
20. Wataru, T.; Keigo, T. Pilot Decontamination in Spatially Correlated Massive MIMO Uplink via Expectation Propagation. *IEICE Trans. Fundam. Electron. Commun. Comput. Sci.* **2021**, *E104/A*, 723–733.
21. Rashid, M.; Naraghi-Pour, M. Clustered Sparse Channel Estimation for Massive MIMO Systems by Expectation Maximization-Propagation (EM-EP). *IEEE Trans. Veh. Technol.* **2023**, *72*, 9145–9159. [CrossRef]