sensors

MDPI

Article

Variational Channel Estimation with Tempering: An Artificial Intelligence Algorithm for Wireless Intelligent Networks

Jia Liu [1], Mingchu Li [1], Yuanfang Chen [2,*], Sardar M. N. Islam [3] and Noel Crespi [4]

[1] School of Software Technology and Key Laboratory for Ubiquitous Network and Service Software, Dalian University of Technology, Dalian 116620, China; Jialiudlut@gmail.com (J.L.); mingchul@dlut.edu.cn (M.L.)
[2] School of Cyberspace, Hangzhou Dianzi University, Hangzhou 310018, China
[3] Institute for Sustainable Industries and Liveable Cities, Victoria University, Melbourne 14428, Australia; Sardar.Islam@vu.edu.au
[4] Institut Polytechnique de Paris, Institut Mines-Telecom, Telecom SudParis, 91011 Evry CEDEX, France; noel.crespi@mines-telecom.fr
* Correspondence: chenyuanfang@hdu.edu.cn; Tel.: +86-1373-545-0984

Received: 8 September 2020; Accepted: 18 October 2020; Published: 21 October 2020

Abstract: With the rapid development of wireless sensor networks (WSNs) technology, a growing number of applications and services need to acquire the states of channels or sensors, especially in order to use these states for monitoring, object tracking, motion detection, etc. A critical issue in WSNs is the ability to estimate the source parameters from the readings of a distributed sensor network. Although there are several studies on channel estimation (CE) algorithms, existing algorithms are all flawed with their high complexity, inability to scale, inability to ensure the convergence to a local optimum, low speed of convergence, etc. In this work, we turn to variational inference (VI) with tempering to solve the channel estimation problem due to its ability to reduce complexity, ability to generalize and scale, and guarantee of local optimum. To the best of our knowledge we are the first to use VI with tempering for advanced channel estimation. The parameters that we consider in the channel estimation problem include pilot signal and channel coefficients, assuming there is orthogonal access between different sensors (or users) and the data fusion center (or receiving center). By formulating the channel estimation problem into a probabilistic graphical model, the proposed Channel Estimation Variational Tempering Inference (CEVTI) approach can estimate the channel coefficient and the transmitted signal in a low-complexity manner while guaranteeing convergence. CEVTI can find out the optimal hyper-parameters of channels with fast convergence rate, and can be applied to the case of code division multiple access (CDMA) and uplink massive multi-input-multi-output (MIMO) easily. Simulations show that CEVTI has higher accuracy than state-of-the-art algorithms under different noise variance and signal-to-noise ratio. Furthermore, the results show that the more parameters are considered in each iteration, the faster the convergence rate and the lower the non-degenerate bit error rate with CEVTI. Analysis shows that CEVTI has satisfying computational complexity, and guarantees a better local optimum. Therefore, the main contribution of the paper is the development of a new efficient, simple and reliable algorithm for channel estimation in WSNs.

Keywords: artificial intelligence algorithm; truth inference; channel estimation; message passing

1. Introduction

1.1. Problem Statement

The recent growth of the mobile Internet has brought a tremendous increase in the number and types of data services, creating much more challenging requirements for 5G communications. Meanwhile, the development of the 5G network relies highly on the improvement of WSNs, which are comprised of a large number of cheap devices that can be used to observe and preprocess neighboring physical and environmental changes. To construct an efficient path through the tremendous amount of base stations, millimeter wave technology is one of the critical parts of 5G technology [1–3]. Therefore, for 5G networks, it is essential to know the channel impulse response (CIR) of the millimeter waves, i.e., to estimate the channel pattern between different base stations. Luckily, the technological breakthrough of micro-electro-mechanical systems (MEMS) has promoted the application of large scale WSNs [4–6]. Thanks to the advantageous properties of self-organization and real-time data preprocessing, WSNs are a promising technology that can contribute to future 5G communications [7–9]. However, there are some challenges underlying this realization, since it is still difficult to deploy certain sensor networks to execute specific tasks. One of the challenges is limited energy resources, as it is energy-consuming to deploy many sensors awaiting task assignment and computing duties. It is, therefore, imperative to develop energy efficient algorithms in WSNs.

A central problem in WSNs technology is its use of data collected by a network of spatially distributed sensors to estimate channel parameters. Sensors can cooperate to send their individual raw or preprocessed data to a data fusion center (FC) when requested. The FC then makes use of the collected data to extract the desired information. This process differs from the traditional centralized system, given that in a centralized system, all data are available and can be accessed in real time. Since the centralized estimation problem is no longer an issue, the existing works are all considered to be decentralized estimation.

1.2. Previous Research and Limitations

The issue of channel estimation has attracted broad interest in the past few years, and the existing algorithms can be classified as least square (LS)/minimum mean square error (MMSE)-based, compressive sensing (CS)-based, message passing (MP)-based, deep learning (DL)-based, or VI-based [10–12].

Conventional channel estimation algorithms typically include LS, MMSE, or expectation maximization (EM)-based methods. In 2005, Qian et al. [13] researched into iterative LS-based channel estimation algorithms, and showed that LS works well in mobile networks. In 2010, Simko et al. [14] proposed an approximate linear minimum mean square error (ALMMSE) fast fading CE algorithm. In 2012, Wang et al. [15] proposed a soft-output MMSE channel estimation algorithm under correlated Rician fading MIMO channels. In 2013, Zhang et al. [16] proposed a distributed angle estimation (or channel estimation) algorithm based on space-alternating generalized expectation maximization. Compared with the LS method, MMSE, including linear minimum mean square error (LMMSE, [17]), considers noise and therefore improves the accuracy. However, with matrix inverse computations, LMMSE has higher computation complexity. Therefore, the conventional channel estimation algorithms (LS, MMSE, etc.) are obviously flawed with their problems of high complexity and inability to scale, which can rapidly lead to failure in massive MIMO scenarios. In a massive MIMO, as the number of antennas in each base station (BS) increases, the downlink pilot training overhead and uplink channel state information (CSI) feedback overhead can quickly become overwhelming. Therefore, the pilots for evaluating CSI are inadequate.

A CS-based approach can be used to solve the above issue with conventional channel estimation algorithms. From the simulated study we can see that limited by the number of scatterers in the BS, the user channel matrix tends to be sparse in massive MIMO scenarios. In this case, it is ineffective to use long training symbols to estimate CSI. CS can mitigate this disadvantage and make use of

user channel matrix sparsity. In 2014, Rao et al. [18] proposed a distributed compressive transmitter side channel state information (CSIT) estimation scheme. In their work, they locally observed the compressed measurements while recovering the CSIT at the base station. In 2016, Tseng et al. [19] proposed a new CS-based downlink CSI estimation scheme for frequency division duplexing (FDD) massive MIMO systems (CE for FDD). However, there are still challenges in CS-based channel estimation algorithms, such as the difficulty to exploit joint channel sparsity among multiple users, and to decide how sparsity can affect performance. Although some researchers have tried to analyze this issue, in general, these are specific analyses, and thus cannot be extended.

Owing to its convergence guarantee property and distributed nature, the MP algorithm is also used to implement channel estimation. In 2005, Paskin et al. [20] proposed a robust distributed sensor network inference architecture and exploited MP to compute the global parameter after forming a spanning tree and a junction tree. In 2019, Bellili et al. [21] proposed a generalized approximate message passing (GAMP) algorithm to find the ground truth of the mmWave MIMO channel coefficient matrix. However, MP-based CE algorithms are too specific to generalize in different types of networks.

Some researchers use DL to estimate channel due to its strong computing power. In 2017, Ye et al. [22] proposed a DL-based CE algorithm to train a deep neural network (DNN) under different channel conditions, and then applied it online to recover the transmitted data. In 2018, He et al. [23] proposed a DL-based channel estimation for beam-space mmWave massive MIMO systems. It is interesting that the authors also use a learned denoising-based approximate message passing network (LDAMP), i.e., they combined the use of DL and MP. In 2020, Zhang et al. [24] proposed a deep reinforcement learning algorithm for double coded caching. In summary, DL-based CE algorithms use deep neural networks to learn the channel feature, and obtain the complete channel information from received pilot symbols. Some DL-based CE models treat channel information as a picture and use the CS-based method to solve channel state information. Despite the high accuracy of DL-based CE algorithms, it is impractical to obtain labeled data or train self-adaptive neural networks for channel estimation. Therefore, it is still difficult to develop DL-based CE algorithms that have real-time generality and efficiency properties.

Some researchers try to use statistical methods to solve the CE problem by forming it into a statistical optimization problem. Treating the channel state information as a set of samples generated by a distribution, we can write the joint posterior density to be maximized as the objective function. However, in statistical inference, it is often too difficult to directly optimize the objective function when computing joint posterior density or other statistical variables. Therefore, we opt for an alternative suboptimal approach, which is VI. In VI, we use another distribution that has minimal KL divergence, or variational free energy, or evidence lower bound (ELBO), with the approximated distribution. By approximating the difficult-to-compute distribution, we can guarantee faster convergence and optimum performance (see Ahad et al. [25]). Please note that many studies that use MP, MF and BP are categorized as VI, and use factor graphs as their graphical models. We classify MP-based channel estimation (CE) algorithms as a specific group, because unlike MF and BP, MP is the direct model for autonomous systems that communicate two-way. In 2008, Hu et al. [26] proposed a divergence minimization approach to derive iterative decoding algorithms in coded CDMA systems. They stated that the KL divergence is equivalent to the variational free energy, both of which can be used to measure the distribution distance between true distribution and auxiliary distribution. In 2010, Kirkelund et al. [27] proposed a variational message passing (VMP) architecture that can estimate the channel coefficients and inverse noise variance matrix in a semi-blind way. As with Hu et al. [26], this scheme is also an iterative decoding channel estimation approach, which only differs in proposing the notion of VI. In 2012, Ahmad et al. [25] proposed a simpler suboptimal approach that uses VI and obtains the auxiliary distribution through KL divergence minimizing and joint posterior density maximizing (we call joint distributed CE in the following). In 2019, Bellili et al. [21] proposed a novel Bayes-optimal channel estimation method that makes use of approximate belief propagation Bayesian networks. In 2017, Cheng et al. [28] proposed a different variational Bayesian inference-based

channel estimation algorithm using a Gaussian mixture as a prior distribution. By using this Gaussian mixture in the prior distribution, the proposed algorithm can make the best use of sparsity within a single channel matrix, as well as between different channel matrices. In 2019, Thoota et al. [29] proposed another variational Bayesian channel estimation algorithm to estimate channel coefficients and the hyper-parameter of the transmitted symbol in MIMO cases. They modeled the MIMO system as a directed graphical model and so, despite the advantages offered by VI combined with data subsampling, i.e., providing approximate posterior inference over large data sets, their approach suffers from a poor local optimum (see Mandt et al. [30]). Some researchers use an annealing approach to tackle this problem, but annealing is only effective when choosing the appropriate cooling schedule (see Mandt et al. [30]).

To sidestep the disadvantages of the above methods of CE, we propose a variational channel estimation with tempering, and solve the problem of channel estimation in a MIMO scenario, and test it using simulations. Variational tempering (VT) is derived from Markov Chain Monte Carlo (MCMC) methods. By sampling from different temperatures, a MCMC method can increase the mixing time of the Markov chain [31]. In a similar way, we introduce global temperature. On the contrary, VT algorithms can learn a temperature distribution from the data, and update the weight in every iteration, which is similar to treating the transmitted signal of each sensor user as a distribution as well as its channel coefficiency. Additionally, we apply our CEVTI in coded CDMA case and uplink massive MIMO case and analyze its complexity and optimality guarantee. We find that CEVTI has better accuracy compared with state-of-the-art CE algorithms, satisfying computational cost and guarantees local optimality.

1.3. Objective

To overcome the problems of the algorithms discussed above, it is necessary to develop a new algorithm that is efficient, operational, and relatively simple. We use variational tempering to meet these requirements. In this work, we consider the channel estimation problem of the pilot signal and channel coefficients, assuming there is orthogonal access between the different sensors and the data fusion center. To reduce the complexity of the direct joint posterior density maximization, we propose a simpler VI algorithm that solves the channel estimation problem, adopting the concept of tempering. By formulating the channel estimation problem into a probabilistic graphical model, CEVTI can estimate the channel coefficiency and the transmitted signal in a low-complexity manner while simultaneously guaranteeing convergence at the same time. Simulations show that CEVTI has higher accuracy than state-of-the art algorithms.

The main objective of this paper is to obtain channel coefficiency by making the best use of trained pilot signals. To achieve this objective, we have performed the following tasks:

1. Modeling of the channel estimation problem into a variational message passing algorithm;
2. Evaluation of the performance error bound of our novel CEVTI algorithm;
3. Use of a Monte Carlo simulation to verify the bit error rate (BER), convergence rate, and mutual information of the CEVTI approach; and
4. Established the efficiency and superiority of our new algorithm, CEVTI.

1.4. Contributions

Over the past few years, VI has been explored in artificial intelligence, crowdsourcing and computer vision, etc. It has been proved an efficient way to compute hyper-parameters, and ground truth in various backgrounds that can be modeled as inference problems.

We propose a new channel estimation algorithm for MIMO-OFDM (orthogonal frequency division multiplexing) systems based on the VI approach. The CEVTI algorithm first models the channel estimation problem into a probabilistic graphical model, and then uses a factorized distribution to approximate the marginal distributions of the desired variables. Finally, the CEVTI algorithm solves

this optimization problem in an alternative manner, just like the EM algorithm. Simulations show that CEVTI performs better than the state-of-the-art channel estimation algorithms.

This paper's main contribution is the development of a new algorithmic approach to channel estimation in WSNs, as discussed below. Based on a variational message passing algorithm, a kind of probabilistic graphical model, this approach solves the channel estimation problem. First, we make an assumption about the user's signal generative distribution, and then infer the channel coefficiency by using pilot feedback. Then we apply CEVTI in the MIMO-OFDM scenario and uplink massive MIMO scenario, and analyze its complexity and optimality. Analysis and simulations show that our algorithm can also be used in a higher dimension and results in better performance than the current channel estimation algorithms.

The main contributions of the paper can be summarized as follows:

1. The modeling of the OFDM channel estimation problem into a new variational message passing algorithm;
2. An evaluation of the performance error bound of our innovative variational tempering channel estimation algorithm; and
3. A numerical simulation of the performance of CEVTI to show that in general cases, the proposed CEVTI algorithm performs better than other algorithms.

In the rest of the paper, we first review the background of channel estimation model and variational inference in Section 2, and then define the background assumptions of the inference model in Section 3. Next, we derive our algorithm stepwise in Section 4, and show the application of CEVTI in Section 5. Next, we show how we obtain the results in Section 6, where we also compare the results of our algorithm with the performance of other approaches. Our CEVTI algorithm requires fewer constraints on the prior distributions of users and channels, and it performs better than the existing channel estimation algorithms. Section 7 presents the conclusion of this study.

2. Background

In our CEVTI system, we consider the case of CDMA, in which there are N antennas (which we term users in the rest of this paper) in the base station. At first, each user will send a signal to its antenna within communication range. The signal experiences relay block fading [32]. We assume that the transmitted signal of each antenna is uniformly distributed. Please note that the transmitted signal is first coded and preprocessed. The code rate of each user is R_c. We use C_i to denote the codeword set of user i after interleaving. Each codeword then is multiplexed with pilot symbols, the number of which is L_p.

In this section, we describe the background notations of CEVTI, assuming that in the wireless sensor network, there are M mobile users and N antennas for each BS. We use d_i to denote the transmitted symbol of mobile user i.

Suppose that each antenna has the state value of $\{\pm 1\}$, and that we have defined d_i to denote the transmitted symbol of user i, i.e., $d_i \in \{\pm 1\}$. Therefore, $d_i[l]$ is the lth transmitted signal, and column vector $\mathbf{d}_i = [\mathbf{d}_{i,p}, \mathbf{d}_{i,c}]^T = [d_i[0], ..., d_i[L-1]]$ is the symbol vector transmitted by user i. We use $\mathbf{d}_{i,p}$ to denote the pilot symbol of user i, the length of which is L_p, and $\mathbf{d}_{i,c} \in \mathcal{C}_k$ to denote the codeword of user i, the length of which is L_c. Obviously $L = L_c + L_p$. For ease of reference, we treat both $\mathbf{d}_i$ and $\mathbf{d}_{i,c}$ as the codeword vectors of user i. The transmitted symbol of user i is then modulated by a normalized sequence embedding $s_i[l]$ of length N_c. Through a channel, the transmitted signal then experiences channel gain, including noise, fading, shifting, etc. However, we assume that the channel gain remains constant in one block time and differs in different blocks.

The vector $d_i \in \mathcal{C}_k$ denotes the combination of the pilot symbol vector and the codeword vector; the length of the latter is L_c. Therefore, the dimension of d_i is L (we may also refer to d_i as a codeword to simplify notation). Every symbol is first processed by a signature waveform, and then is further altered by the transmitted channel, including channel gains, block fading etc. We assume the within

each block, consisting of L symbols, the channel influence is constant, while between blocks, channel gains are different.

Next, we use $\mathbf{r}[l] = [r_1[l], ..., r_{N_c}[l]]^T$ to indicate the receiving signal at interval l. So that the received signal after experiencing coding and channel gain is:

$$\mathbf{r}[l] = \mathbf{S}[l]\mathbf{A}\mathbf{d}[l] + \mathbf{w}[l] = \mathbf{S}[l]\mathbf{D}[l]\mathbf{a} + \mathbf{w}[l] \tag{1}$$

where $\mathbf{S}[l]$ is the spreading sequences of all users at signaling interval l, of which $\mathbf{a} \triangleq [a_1, ..., a_N]^T$ is the channel coefficient, and $\mathbf{A} = \text{diag}\{a_1, ..., a_N\}$ is the matrix of diagonalized channel coefficient, of which a_i is the channel coefficient value of user i. Assuming a Rayleigh-fading channel and white Gaussian noise distribution, we have $\mathbf{a} \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma_a})$.

3. Solution Framework

Here, we introduce the CEVTI algorithm and then form its derivations. To consider a specific application of CEVTI, we apply it to the case of a coded multiuser system. In this multiuser scenario, we have multiple users whose channel coefficients must be calculated.

First, given the output of filters at signaling interval l $\mathbf{r}[l]$, channel coefficient vector $\mathbf{a}$, and transmitted signal $\mathbf{d}$, we obtain the joint probability of transmitted signals, channel coefficients, and output signal matrix, conditioned on the hyper-parameter of a α and the hyper-parameter of d θ:

$$p(\mathbf{d}, \mathbf{a}, \mathbf{r}|\alpha, \theta) \propto \prod_{i \in [N]} p(\mathbf{a}|\alpha) p(\mathbf{d}|\theta) p(\mathbf{r} \mid \mathbf{a}, \mathbf{d}) \tag{2}$$

Please note that we assume channel coefficients, transmitted signals and output signals are independent with each other, for instance, $p(\mathbf{d}|\theta) = \prod_{i \in [N]} p(d|\theta)$, and the joint probability of transmitted signals, channel coefficients, and output signal matrix can be extended using the chain rule in Bayesian network to simplify computation.

According to Bayesian theory, the estimation of d that has minimum error rate is:

$$\begin{aligned} &\hat{d}_i = \arg\max_{d_i} p(d_i|\mathbf{r}, \alpha) \\ &\quad \text{where } p(d_i|\mathbf{r}, \alpha) = \sum_{i \in \mathcal{N}_j \setminus i} \int_a p(\mathbf{a}, \mathbf{d} \mid \mathbf{r}, \alpha) \mathrm{d}\mathbf{a} \end{aligned} \tag{3}$$

of which $\mathcal{N}_j$ is the neighbor of base station antenna j.

We can directly see that computing $p(d_i|\alpha, \theta)$ is trivial since $p(a, d|r, \alpha)$ is unavailable. To solve this problem, we can borrow a concept from belief propagation (BP) and the mean-field method (MF) that uses a more factorable distribution to approximate the conditional probability in the above equation.

3.1. Mean-Field CEVTI

This subsection focuses on the hierarchical Bayesian model. In this model, all users share a global variable that determines the distribution of channel coefficients, while each local hidden variable belongs to each user. We denote $\mathbf{r} = r_{1:N}$ as observed variables, $\mathbf{d} = d_{1:N}$ as local hidden values, and α as global hidden values. Using a variational distribution $q(\mathbf{a}|\alpha)$ to approximate $p(\mathbf{a}|\alpha)$, and a variational distribution $q(\mathbf{d}|\theta)$ to approximate $p(\mathbf{d}|\theta)$, the joint density of the model, according to Equation (2), then becomes:

$$p(\mathbf{a}, \mathbf{r}, \mathbf{d}) = q(\mathbf{a}|\alpha) q(\mathbf{d}|\theta) \prod_{i=1}^{N} p(r_i, d_i|\theta), \tag{4}$$

of which α is the global variable, i.e., the hyper-parameter of this model, and θ is the hidden local variable.

The main computing problem of the Bayesian model is the posterior inference, i.e., the output signal (or received signal). Our objective is to calculate $p(\mathbf{a,d} \mid \mathbf{r})$, the conditional probability of the hidden variable (the transmitted signal) and the global hidden variables (the hyper-parameter of channel coefficients) given the observations (received signal). In many cases, this computation is intractable, and approximation is required.

CEVTI proposes a parameterized family of hidden variables and tries to determine the family of members that are closest to the posterior distribution by KL divergence. That is to say, CEVTI tries to use the variational distribution to replace the target distribution. In VI, this is equal to computing the ELBO in terms of a variational parameter, since maximizing ELBO and minimizing KL divergence between the variational distribution and target distribution, are equivalent to maximizing the posterior probability:

$$\begin{aligned} \log p(\mathbf{r}) &= \mathbb{E}_q[\log(\frac{p(\mathbf{a,\ d,\ r})}{q(\mathbf{a}|\alpha)q(\mathbf{d}|\theta)})] - \mathbb{E}_q[\log(\frac{p(\mathbf{a,d} \mid \mathbf{r})}{q(\mathbf{a}|\alpha)q(\mathbf{d}|\theta)})] \\ &= \mathbb{E}_q[\log p(\mathbf{a,\ d,\ r})] - \mathbb{E}_q[\log q(\mathbf{a,\ d}|v)] - \mathbb{E}_q[\log(\frac{p(\mathbf{a,d} \mid \mathbf{r})}{q(\mathbf{a}|\alpha)q(\mathbf{d}|\theta)})] \\ &= \mathcal{L}(v) + KL(q(\mathbf{a,d})||p(\mathbf{a,d} \mid \mathbf{r})) \end{aligned} \tag{5}$$

of which

$$\mathcal{L}(v) = \mathbb{E}_q[\log p(\mathbf{a,\ d,\ r})] - \mathbb{E}_q[\log q(\mathbf{a,\ d}|v)] \tag{6}$$

Using the fully factored family, we assume that each user and its transmitted signals are independent

$$q(\mathbf{d,a}|v) = q(\mathbf{d}|\lambda)\prod q(a_i|\phi_i) \tag{7}$$

where v is the variational parameter, λ is the global variational variable, ϕ_i is the local variational variable of user i, and therefore $v = \{\lambda, \phi\}$. This kind of simplification method is also called the mean-field method, and offers several computational advantages, especially when deriving the gradients of the objective function. VI optimizes Equation (6) using a gradient or coordinate ascent. To best determine the local optimal, we use tempered variational inference.

Tempered variational inference applies tempering into mean-field variational inference. We first use an additional temperature parameter $T \geq 1$. Given T, the joint probability becomes:

$$p(\mathbf{d,a,r}|T) = \frac{p(\mathbf{d,r} \mid \mathbf{a})^{1/T} p(\mathbf{a}|\alpha)}{c(T)}. \tag{8}$$

where $c(T)$ is the normalizing constant, or the tempered partition function:

$$c(T) = \int p(\mathbf{d,r} \mid \mathbf{a})^{1/T} p(\mathbf{a}|\alpha)\mathbf{drdad d}. \tag{9}$$

The tempered joint means the tempered posterior. Tempered variational inference optimizes the variational distribution $q(\dot{)}$ against a tempered posterior. We start from a higher temperature and end when $T = 1$. The tempered ELBO then becomes:

$$\begin{aligned} \mathcal{L}_A(\lambda, \phi; T) =& \mathbb{E}_q[\log p(\mathbf{a}|\alpha)] - \mathbb{E}_q[\log q(\mathbf{a}|\lambda)] \\ &+ \sum_{i=1}^{N} (\mathbb{E}_q[\log p(d_i, r_i|a)])/T - \mathbb{E}_q[\log p(d_i|\phi_i)] \end{aligned} \tag{10}$$

We introduce a random variable that produces temperature to the joint distribution. We obtain the joint distribution of tempering at this temperature based on the results of random variables and use the following polynomial temperature distribution

$$y \sim Mult(\pi) \tag{11}$$

3.2. Tempered Joint

Using the chain rule of Bayesian theory, the joint distribution can be factorized as $p(\mathbf{r}, \mathbf{d,a}, y) = p(\mathbf{r}, \mathbf{d,a}|y)p(y)$. We use uniform temperature distribution $p(y) = \prod_{m=1}^{M} \pi_m^{y_m}$. Conditioned on y, the tempered joint can be defined as:

$$p(\mathbf{r,d,a}|y) = \frac{p(\alpha)}{c(T_y)} \prod_{i=1}^{N} p(\mathbf{a}, d_i|\alpha)^{1/T_y} \tag{12}$$

Thus, the model becomes:

$$p(\mathbf{r,a,d}, y) = p(\alpha) \prod_{m=1}^{M} (\frac{\pi_m}{c(T_m)} \prod_{i=1}^{N} p(\mathbf{a}, d_i|\alpha)^{1/T_m})^{y_m} \tag{13}$$

3.3. Tempered ELBO

Now we have defined the variational objective of the extended model. Next, we implement the mean-field family into a model containing temperature:

$$p(\mathbf{d,a}, y|\phi, \lambda, \gamma) = q(\mathbf{d}|\phi)q(\mathbf{a}|\lambda)q(y|\gamma) \tag{14}$$

in which γ is variational parameter for temperature. The tempered ELBO is:

$$\begin{aligned} \mathcal{L}_T(\lambda, \phi; T) = {} & \mathbb{E}_q[\log p(\mathbf{d})] + \mathbb{E}_q[\log\, p(\mathbf{a}|\alpha)] + \mathbb{E}_q[\log\, p(y)] \\ & - \mathbb{E}_q[\log\, q(\theta)] \\ & - \mathbb{E}_q[1/T_y] \sum_{i=1}^{N} (\mathbb{E}_q[\log p(\mathbf{a}, d_i|\alpha)]) \\ & - \mathbb{E}_q[\log C(T_y)] - \sum_{i=1}^{N} \mathbb{E}_q[\log q(d_i)] - \mathbb{E}_q[\log q(y)] \end{aligned} \tag{15}$$

3.4. Local Variational

To make the calculation easier to process, we can use the local temperature for each user instead of the global temperature. This approach also makes it easier to process future data and to learn the tempering schedule of a single data. This local temperature can be viewed as the local energy potential across the network. We represent t_i as the local temperature of each data, and the joint probability can be expressed as:

$$p(\mathbf{r,d,a}, t) \propto p(\mathbf{a}|\alpha) \prod_{i=1}^{N} [p(r_i, d_i|\mathbf{a})^{1/t_i} p(t_i)] \tag{16}$$

According to the above equation, the temperature can down-weight global hidden variables, making the local distribution more entropic. That is to say, at first, we put little weight on the influence of training data. In this way, we can quickly find a good initial result. In addition, then, we gradually optimize the result by putting more weight on the training data. Thanks to this effect, we will not have to calculate the tempered partition function, because the local tempering likelihood and the tempered partition function will be in the same family as the original prior of user distribution. The disadvantage; however, is that we cannot reach a conjugate model. Local temperature shows the likelihood of the data that comes from CEVTI without tempering. By assigning them higher temperatures, we can better explain outliers. In other words, local temperature enables us to model the data more flexibly, and to learn different tempering schedules for each data point during inference.

4. The CEVTI Algorithm

We now show how we derive our CEVTI algorithm. Our algorithm borrows from the concept of stochastic variational inference (see Hoffman et al. [33]), and combine it further with tempering. As in Hoffman et al., we focus on the conditional conjugate exponential family (CCEF). If in the model, the prior of user distribution and the local conditional probability belong to the same exponential family, and are conjugate, then the model is in the CCEF:

$$p(\mathbf{a}|\alpha) = h(\mathbf{a})\exp\{\alpha^T t(\mathbf{a}) - a_g(\theta)\} \tag{17}$$

$$p(d_i, r_i|\mathbf{a}) = h(r_i, d_i)\exp\{\mathbf{a}^T t(d_i, r_i) - a_l(\mathbf{a})\} \tag{18}$$

where $t(\mathbf{a})$ and $t(d_i, r_i)$ are sufficient statistics of global and local data points, and the sufficient statistics of a are $t(a) = (a, -a_l(a))$ (with slight abuse of notations, since l is the interval indicator of the transmitted signals, the subscripts g and l refer to global and local respectively). Therefore, hyper-parameter α consists of two components, (α_1, α_2), the former is a vector that has the same dimension as a, and the latter is a scalar vector. Notice that the conditional probability of the global variable is in the same exponential family as the prior of natural parameter η_g, of which $a_g(\theta)$ and $a_l(d)$ are the corresponding log normalizers (here a is different from channel coefficients $\mathbf{a}$). We choose CCEF because it allows us to compute expectations analytically.

We treat the following objective function as a function of the global variational parameter

$$\mathcal{L}_T(\lambda, \phi; T) = \mathcal{L}_T(\lambda, \phi_{ij}; T) \tag{19}$$

$$\phi(\lambda, T) = \arg\max_{\phi_{ij}} \mathcal{L}_T(\lambda, \phi_{ij}; T) \tag{20}$$

According to Hoffman et al. [33], the ELBO after combining with tempering is:

$$\begin{aligned}\mathcal{L}_T(\lambda, \phi; T) = \mathbb{E}_q[1/T_y \eta_g(\mathbf{r,d}, \alpha)]^T \nabla_\lambda a_g(\lambda) - \lambda^T \nabla_\lambda a_g(\lambda) \\ + a_g(\lambda) - \mathbb{E}_q[\log C(T_y)] - \mathbb{E}_q[\log q(y)]\end{aligned} \tag{21}$$

of which

$$\eta_g(\mathbf{r,d}, \alpha) = (\alpha_1 + \sum_{i=1}^{N} t(r_i, d_i), \alpha_2 + N) \tag{22}$$

4.1. Updates

According to Hoffman et al. [33], the tempered ELBO's natural gradient with respect to the global variational parameter is:

$$\nabla_\lambda \phi(\lambda, T) = \mathbb{E}_q[\eta_g(\mathbf{r,d}, \alpha)] - \lambda \tag{23}$$

$$\hat{\lambda} = \mathbb{E}_q[\eta_g(\mathbf{r,d}, \alpha)] \tag{24}$$

$$\lambda_{t+1} = \lambda_t + \rho(\hat{\lambda} - \lambda_t) \tag{25}$$

In the first step, we estimate $\hat{\lambda}$, and then update it using λ_t with decreasing learning rate ρ_t. By dividing the expectation of the sufficient statistics, we can reduce the variance of the gradients.

Then we optimize the tempered ELBO, so that the local variational update is:

$$\phi_{ij} = \frac{1}{T}\mathbb{E}_q[\eta_l(t_{i,-j}, a_i, \theta)] \tag{26}$$

of which η_l is local variational parameter.

5. Application of CEVTI

In this section, we apply the CEVTI algorithm to the case of coded Code Division Multiple Access (CDMA) and in the case of uplink massive MIMO systems. As discussed above, CEVTI works as a formal optimization framework based on variational inference with tempering. It has the properties of guaranteed convergence and smooth descent, which lead to a better local optimum.

5.1. Application of CEVTI for CDMA

Assuming that the prior distribution of $\mathbf{a}$ is Gaussian, i.e.,

$$p(\mathbf{a}) \propto \exp\{-\mathbf{a}^{\mathbf{H}}\boldsymbol{\Sigma}_{\mathbf{a}}\mathbf{a}\} \tag{27}$$

we can rewrite the CEVTI objective as:

$$\begin{aligned} \text{Maximize}\mathcal{L}_T(\lambda,\phi;T) &= \mathbb{E}_q[1/T_y\eta_g(\mathbf{r},\mathbf{d},\alpha)]^T\nabla_\lambda a_g(\lambda) \\ &\quad - \lambda^T\nabla_\lambda a_g(\lambda) \\ &\quad + a_g(\lambda) - \mathbb{E}_q[\log C(T_y)] - \mathbb{E}_q[\log q(y)] \\ \text{subject to } &\int_a \mathrm{d}\mathbf{a} = 1 \\ &\sum_i^N d_i = 1 \end{aligned} \tag{28}$$

5.1.1. Channel Coefficient Estimation

According to Equation (26), the update of the channel coefficient vector is:

$$q_{\mathbf{a}}^{t+1}(\mathbf{a}) \propto \frac{1}{T}\exp\sum_{d_i} q_{d_i}^t \eta_g(\mathbf{r},\mathbf{d},\alpha) \tag{29}$$

Similarly, the update of the codeword parameter is:

$$q_{d_i}^{t+1}(d_i) \propto \frac{1}{T}\exp\int \mathrm{d}\mathbf{a}\sum_{d_{-i}} q_{d_{-i}}^t \eta_l(\mathbf{r},\mathbf{d},\alpha) \tag{30}$$

Observe that when $d_i \notin \mathcal{C}_i$, $q_{d_i} = 0$. Furthermore, $\eta_g(\mathbf{r},\mathbf{d},\alpha) \propto \log p(\mathbf{a},\mathbf{d} \mid \mathbf{r})$, which is not intuitive to compute, it can be simplified by computing $p(\mathbf{r}|\mathbf{a},\mathbf{d})$ instead, since by the chain rule, $p(\mathbf{a},\mathbf{d}|$ $textbfr) = p(\mathbf{r}|\mathbf{a},\mathbf{d})p(\mathbf{a})p(\mathbf{d})/p(\mathbf{r})$, and dropping $p(\mathbf{d})$ when updating $p(\mathbf{a})$ does not impact the result. According to Equation (1), we can easily compute the analytical distribution of $p(\mathbf{r}|\mathbf{a},\mathbf{d})$ using Gaussian density equation. Therefore, the above equation can be rewritten as:

$$q_{\mathbf{a}}^{t+1}(\mathbf{a}) \propto \frac{1}{T}\exp\sum_{d_i} q_{d_i}^t \log p(\mathbf{r}|\mathbf{a},\mathbf{d}) \tag{31}$$

Similarly, the update of the codeword parameter is:

$$q_{d_i}^{t+1}(d_i) \propto \frac{1}{T}\exp\int \mathrm{d}\mathbf{a}\sum_{d_{-i}} q_{d_{-i}}^t \log p(\mathbf{r}|\mathbf{a},\mathbf{d}) \tag{32}$$

Assuming the noise vector $w[l]$ of Equation (1) is Σ_w, and $\mathbf{a}$ is also Gaussian distribution, we need to update the mean and covariance of $\mathbf{a}$ separately, i.e.,

$$q_{\mathbf{a}}^{t+1}(\mathbf{a}) \propto \frac{1}{T}\exp\sum_{d_i} q_{d_i}^t \log p(\mathbf{r}|\mathbf{a},\mathbf{d}) \tag{33}$$

in which

$$\begin{aligned} \log p(\mathbf{r}|\mathbf{a},\mathbf{d},\Sigma_w) &= \text{const} * |\Sigma_w|^{-1} \\ *\exp(r[l] - S[l]D[l]\mathbf{a})\Sigma_w^{-1}(r[l] - S[l]D[l]\mathbf{a})^H \end{aligned} \tag{34}$$

Defining

$$\begin{aligned} d_i^t[l] &\triangleq \mathbb{E}_{q_{\mathbf{d}}^t}\{d_i[l]\} = \mathbb{E}_{q_{d_i}^t}\{d_i[l]\} \\ &= \sum_{d_i \in \mathcal{C}_i, d_i[l]=1} q_{d_i}^t(d_i) - \sum_{d_i \in \mathcal{C}_i, d_i[l]=-1} q_{d_i}^t(d_i) \end{aligned} \tag{35}$$

$$\mathbb{E}_{q_d^t}\{d_i[l] * d_j[l]\} \triangleq \begin{cases} d_i[l] * d_j[l], i \neq j, \\ 1, i = j \end{cases} \tag{36}$$

and

$$(\Omega_w^t)^{-1} \triangleq \mathbb{E}_{q_{\Sigma_w^{-1}}^t}\{\Sigma_w^{-1}\} \tag{37}$$

so that:

$$q_{\mathbf{a}}^{t+1}(\mathbf{a}) \propto \frac{1}{T}\exp[-(\mathbf{a} - \mathbf{a}^{t+1})^H {\Sigma_{\mathbf{a}}^{t+1}}^{-1}(\mathbf{a} - \mathbf{a}^{t+1})] \tag{38}$$

where **a** is:

$$\begin{aligned} \mathbf{a}_{\text{mean}}^{t+1} = \Sigma_{\mathbf{a}}^{t+1}(&\sum_{l=0}^{L_p-1} D_p[l]^H S[l]^H (\Omega_w^t)^{-1} r[l] \\ &+ \sum_{l=L_p}^{L-1} \tilde{D}^t[l]^H S[l]^H (\Omega_w^t)^{-1} r[l]) \end{aligned} \tag{39}$$

and the covariance of **a** is:

$$\begin{aligned} \Sigma_{\mathbf{a}}^{t+1} = \frac{1}{T^2}(&\Sigma_{\mathbf{a}}^{-1} + \sum_{l=0}^{L_p-1} D_p[l]^H S[l]^H (\Omega_w^t)^{-1} S[l] D_p[l] \\ &+ \sum_{l=L_p}^{L-1} \tilde{D}^t[l]^H S[l]^H (\Omega_w^t)^{-1} S[l] \tilde{D}[l] \\ &+ \sum_{l=L_p}^{L-1} \mathbf{E}^t[l]^H \text{Diag}\{S[l]^H (\Omega_w^t)^{-1} S[l]\} \mathbf{E}^t[l])^{-1} \end{aligned} \tag{40}$$

in which $\mathbf{E}^t = \text{diag}\{1 - (\tilde{d}_1[l]^t)^2, \cdots, 1 - (\tilde{d}_N[l]^t)^2\}$.

5.1.2. Noise Covariance Estimation

Here we estimate the noise covariance matrix of our CDMA channel estimation model. Notice that in the original derivation of CEVTI, we do not include Σ_w^{-1} as the targeted parameter to be estimated, because CEVTI can be extended to multiple dimensions flexibly, without complex modification.

Similarly, we reach the update equation of noise covariance as follows:

$$q_{\Sigma_w^{-1}}^{t+1}(\Sigma_w^{-1}) \propto \frac{1}{T} p(\Sigma_w^{-1}) \exp \int_{\mathbf{a}} d\mathbf{a} \sum_{d_i} q_{d_i}^t \log p(\mathbf{r}|\mathbf{a},\mathbf{d}) \tag{41}$$

Defining

$$\begin{aligned} B_t &\triangleq \exp \int_{\mathbf{a}} \mathrm{d}\mathbf{a} \sum_{d_i} q_{d_i}^t \log p(\mathbf{r}|\mathbf{a},\mathbf{d}) \\ &= \sum_{l=0}^{L_p-1} ((r[l] - S[l]D_p[l]\mathbf{a}^t)(r[l] - S[l]D_p[l]\mathbf{a}^t)^H \\ &\quad + S[l]D_p[l]\Sigma_{\mathbf{a}}^t {D_p}^t[l]^H S[l]^H) \\ &\quad + \sum_{l=L_p}^{L-1} ((r[l] - S[l]\tilde{D}^t[l]\mathbf{a}^t)(r[l] - S[l]\tilde{D}^t[l]\mathbf{a}^t)^H \\ &\quad + S[l]\mathbf{E}^t[l]\mathbf{A}^t(\mathbf{A}^t)^H(\mathbf{E}^t[l])^H(S[l]^H) \\ &\quad + S[l]\mathbf{E}^t[l]\mathrm{Diag}\{\Sigma_{\mathbf{a}}^t\}(\mathbf{E}^t[l])^H(S[l]^H) \\ &\quad + S[l]\tilde{D}^t[l]\Sigma_{\mathbf{a}}^t \tilde{D}^t[l]^H S[l]^H) \end{aligned} \tag{42}$$

$$q_{\Sigma_w^{-1}}^{t+1}(\Sigma_w^{-1}) \propto \exp \int_{\mathbf{a}} \mathrm{d}\mathbf{a} \sum_{d_i} q_{d_i}^t |\Sigma_w^{-1}|^L \exp[-tr\{\Sigma_w^{-1} B^t\}] \tag{43}$$

Therefore, Σ_w^{-1} is a Wishart distribution, whose expectation is the multiplication of dimension and the inverse covariance matrix, which is

$$(\Omega_w^t)^{-1} = \frac{L+N+1}{B^t} \tag{44}$$

5.1.3. Codeword Distribution Estimation

We estimate the codeword distribution matrix of our CDMA channel estimation model in this subsection.

Similar to the noise covariance estimation, we obtain the updated equation of the codeword distribution as follows:

$$q_{{d_i}^{t+1}}({d_i}^{t+1}) \propto \frac{1}{T} p(d_i) \exp \int_{\mathbf{a}} \mathrm{d}\mathbf{a} \sum_{d_i} q_{d_i}^t \log p(\mathbf{r}|\mathbf{a},\mathbf{d}) \tag{45}$$

Assuming that the distribution of the codeword is a uniform distribution, and defining:

$$\begin{aligned} C_i^{t+1}[l] &\triangleq \exp \int_{\Sigma_w^{-1}} \mathrm{d}\Sigma_w^{-1} \int_{\mathbf{a}} \mathrm{d}\mathbf{a} \sum_{d_i} q_{d_i}^t \log p(\mathbf{r}|\mathbf{a},\mathbf{d}) \\ &\propto \exp\{-tr(\Omega_w^t)^{-1} \sum_{l=0}^{L_p-1} (r[l] - S[l]\tilde{D}_i[l]\mathbf{a}^t)(r[l] - S[l]\tilde{D}_i[l]\mathbf{a}^t)^H \\ &\quad + S[l]\tilde{D}_i^t[l]\Sigma_a^t \tilde{D}_i^t[l]^H S[l]^H\} \end{aligned} \tag{46}$$

where $\tilde{D}_i^t[l] = \mathrm{diag}\{\dots \tilde{d}_{i-1}^t[l] d_i^t[l] \tilde{d}_{i+1}^t[l] \dots\}$. We can get:

$$q_{{d_i}^{t+1}}({d_i}^{t+1}) \propto \exp\{\sum_{l=0}^{L_p-1} d_i[l] C_i^{t+1}[l]\} \tag{47}$$

5.2. Application of CEVTI in Massive MIMO

In this section, we apply the CEVTI algorithm to the case of uplink massive MIMO systems with low resolution ADCs. Similar to the above case in CDMA, in uplink massive MIMO systems with low resolution ADCs, base stations will quantize the received signals before decoding them, i.e.

$$\mathbf{Y} = \mathcal{Q}(\mathbf{r}) \tag{48}$$

of which $\mathbf{r}$ is the received signal before quantization (or modulation), $\mathbf{Y}$ is the received signal after quantization, and $\mathcal{Q}$ is the quantization operator. Since the antennas of massive MIMO systems use M-Quadrature Amplitude Modulation (M-QAM) to modulate the received signals, we need to transform the received signals with imaginary parts and real parts into real number. Therefore, the joint probability of transmitted signals, received signals before modulation, received signals after modulation, and channel coefficients are

$$p(\mathbf{r,d,a,Y},t) \propto p(\mathbf{a}|\alpha)\prod_{i=1}^{N}[p(Y|r_i)p(r_i,d_i|\mathbf{a})^{1/t_i}p(t_i)] \tag{49}$$

Assume the noise is Gaussian distributed, the conditional distribution of the signals and channel coefficients are given as

$$p(\mathbf{r}\,|\,\mathbf{d,a}) \propto \exp(-\frac{1}{\sigma_w^2}\sum_{l=0}^{L-1}||r_i[l]-\mathbf{a}d_i[l]||^2) \tag{50}$$

$$p(\mathbf{a}|\alpha) \propto \exp(-\sum_{j=0}^{M}\frac{1}{\alpha_j}||a_j||^2) \tag{51}$$

$$p(\mathbf{Y}\,|\,\mathbf{d}) = \mathbb{1}(d_i \in [d_i^{lo}, d_i^{hi}]) \tag{52}$$

of which $[d_i^{lo}, d_i^{hi}]$ are soft bounds of d_i.

Then, according to Equation (14), we can obtain the variational joint containing temperature:

$$p(\mathbf{d,Y,a,y}|\phi,\zeta,\lambda,\gamma) = q(\mathbf{d}|\phi)q(\mathbf{Y}|\zeta)q(\mathbf{a}|\lambda)q(y|\gamma) \tag{53}$$

Therefore, the tempered ELBO is:

$$\begin{aligned}\mathcal{L}_T(\lambda,\phi;T) = {} & \mathbb{E}_q[\log p(\mathbf{d})] + \mathbb{E}_q[\log p(\mathbf{Y})] + \mathbb{E}_q[\log\, p(\mathbf{a}|\alpha)] + \mathbb{E}_q[\log\, p(y)] \\ & - \mathbb{E}_q[\log\, q(\theta)] \\ & - \mathbb{E}_q[1/T_y]\sum_{i=1}^{N}(\mathbb{E}_q[\log p(\mathbf{a},d_i|\alpha)]) \\ & - \mathbb{E}_q[\log C(T_y)] - \sum_{i=1}^{N}\mathbb{E}_q[\log q(d_i)] - \sum_{i=1}^{N}\mathbb{E}_q[\log q(Y_i)] - \mathbb{E}_q[\log q(y)]\end{aligned} \tag{54}$$

According to Equation (26), the update of the channel coefficient vector is:

$$\mathbf{a}_{\text{mean}}^{t+1} = \sigma_w^{t+1}(\sum_{l=0}^{L-1}D_i[l]^H(\sigma_w^t)^{-1}\mathbf{Y}[l]) \tag{55}$$

and the covariance of $\mathbf{a}$ is:

$$\sigma_{\mathbf{a}}^{t+1} = \frac{\sigma_w^2}{\sum_{l=0}^{L-1}D_i^t[l] + \frac{\sigma_w^2}{\alpha_j}} \tag{56}$$

And similarly, we reach the solution of $d_i[l]$ and $r_i[l]$:

$$q(d_i[l]=m) = \frac{\exp(-\frac{1}{\sigma_w^2}f(m))}{\sum_{m'}\exp(-\frac{1}{\sigma_w^2}f(m'))} \tag{57}$$

of which m is the value in M-QAM space, and

$$f(m) = \mathbb{E}_q[\|a_j\|^2]|m|^2 + 2\mathcal{R}[(\sum_{j'=1,j'\neq j}^{M}a_{j'}^H a_j d_i^{j'}[l]^* - r[l]^H a_j)m] \tag{58}$$

$$r_i^{mean}[l] = \mathbb{E}_q[\mathbf{a}d_i[l]] \tag{59}$$

$$r_i[l] \sim \mathcal{N}(\mathbb{E}_q[\mathbf{a}d_i[l]], \frac{\sigma_w}{\sqrt{2}}) \tag{60}$$

6. Simulations and Results

In this section, we will numerically assess the performance of the five most popular channel estimation algorithms for the simulations in which we test the CEVTI algorithm. After comparing their performances, we analyze the CEVTI algorithm's complexity and optimality guarantee.

6.1. Simulations

We perform Monte Carlo simulations in a typical communication system, and assume that the signal generating parameter is uniformly distributed as well as the channel gain. Some of these parameters are listed in Table 1.

Table 1. Parameter notations and values.

Parameter	Meaning	Value
L_p	pilot number	32
N	number of users	8
M	number of antennas	4
r_{ij}	receiver j's signal toward user i	
$[\alpha_1, \beta_1]$	signal prior	[0.5,1]
$[\alpha_2, \beta_2]$	user prior	[0.4,0.6]
ϵ	convergence tolerance	10^{-6}

We use the BER, convergence rate, and mutual information to show the accuracy of our CEVTI. In our pilot scheme, all transmitters transmit at the same time frequency and experience the same Rayleigh-fading channel. Furthermore, we assume there is no correlation between the transmitting antennas and the receiving antennas.

We compare CEVTI with five channel estimation algorithms: LS, LLMSE, joint distributed CE, CE for FDD, LDAMP and VMP. All the variational-based algorithms use the same initialization, i.e., a pilot-based channel estimation and a MIMO channel model.

Figure 1 shows the impact of noise variance on the estimation of the channel coefficient. It is intuitive that the BER increases with the noise variance at an acceleration rate. Please note that the sensor noise variance also has a huge impact on the estimation accuracy, and is more difficult to correct using CE algorithms, unless very large data samples are used. However, CEVTI displays higher accuracy compared with others, which means that CEVTI guarantees better local optimum.

Figure 2 shows the BER of the signal-to-noise ratio (SNR, i.e., E_b/N_0). We iterate 10 times to simplify the comparison. The results show that CEVTI has a better transmitter performance than the five other existing CE algorithms. The BER of CEVTI is significantly lower than LLMSE-based algorithms. In addition, the convergence rate of CEVTI is non-degenerate compared to other iterative CE algorithms. We can see that LDAMP also works well due to its computational power and use of large training data. The CEVTI algorithm is also expected to produce a local minimum BER, as it can make the best use of local data samples, even when the sample size is small.

Figure 3 shows the convergence behavior of CEVTI compared with LS, LLMSE, joint distributed CE, CE for FDD, LDAMP and VMP under different SNR. The results show that CEVTI converges faster and smoother than other CE algorithms. The reason is, as explained before, VT has guaranteed convergence and smooth descent. We can see from Figure 3, joint distributed CE and VMP also behave well due to the application of VI. For instance, at $SNR = 15$ dB, CEVTI converges in around 15 iterations, and joint distributed CE and VMP converge in 17 and 16 iterations, respectively. The simulated executing time of CEVTI is also better, showing that CEVTI has satisfying computational complexity.

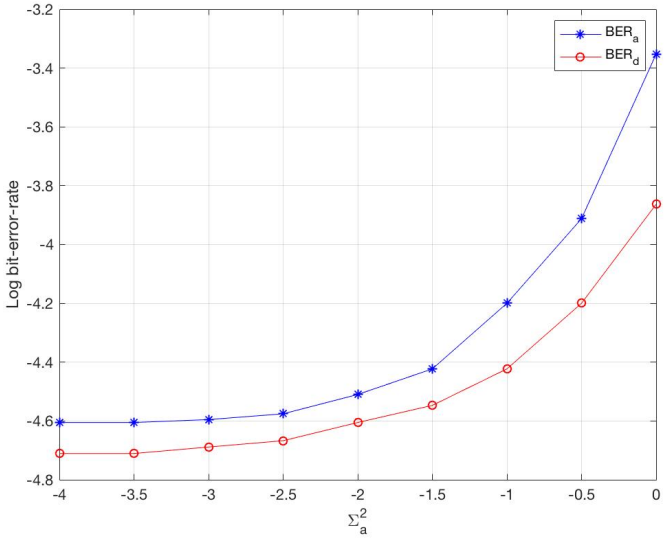

Figure 1. Noise Variance Impact.

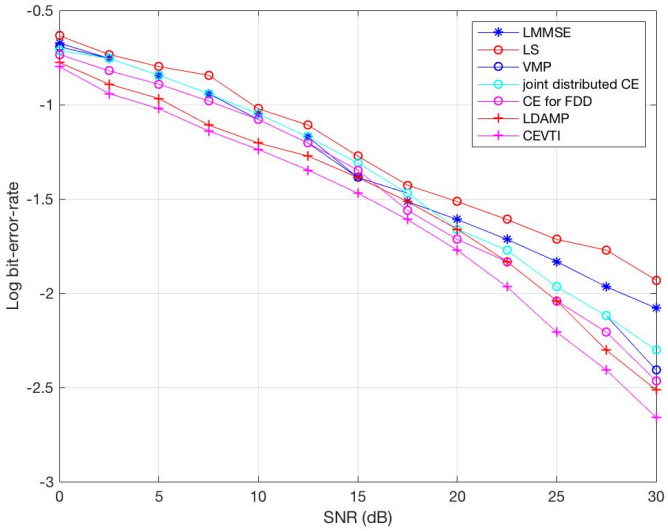

Figure 2. BER comparison.

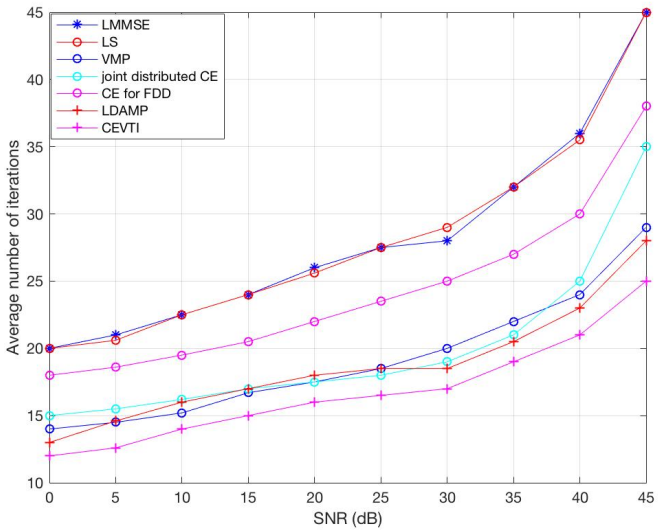

Figure 3. Average number of iterations until convergence.

Figure 4 shows the mutual information comparison with LS, LLMSE, joint distributed CE, CE for FDD, LDAMP and VMP under different SNR. The mutual information is between users and BS antennas, which is a good indicator of the impact of channel estimation accuracy on the throughput of the average bandwidth. The results show that CEVTI has non-degenerate throughput compared with others. For instance, at $SNR = 20$ dB, CEVTI's mutual information is 41, same as VMP and joint CE, slightly higher than others, showing that CEVTI has theoretical satisfying network throughput.

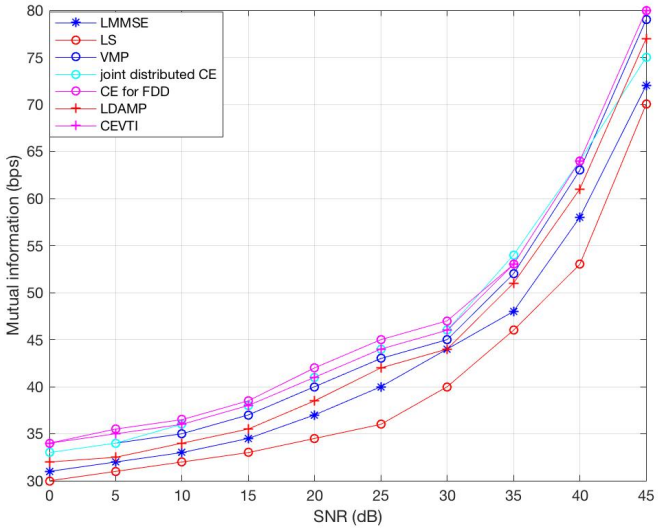

Figure 4. Mutual information comparison.

6.2. Complexity Analysis

Here, we analyze the computational complexity of CEVTI. Since the main computing burden of CEVTI is the matrix inverse operation, we can only consider Equations (38)–(40). Since a $N \times N$ matrix takes N^3 calculations to calculate its inverse, we can see that Equation (38) takes $\mathcal{O}(N^3)$ calculations, and Equations (39) and (40) takes $\mathcal{O}(L_p^3 + (L - L_p)^3)$ calculations. Therefore, the total complexity of CEVTI is $\mathcal{O}(N^3 + L_p^3 + (L - L_p)^3)$.

6.3. Optimality Guarantee

We now prove how our CEVTI has the local optimum guarantee.

Theorem 1 (Optimality Guarantee)**.** *Our CEVTI has optimality guarantee.*

Proof of Theorem 1. Recall that the optimum of a constrained optimization problem, also called stationary point, or sometimes fixed point, has a set of gradients that is orthogonal to the constraint set [34]. Therefore, we need to prove that the result of CEVTI is a stationary point.

We first redefine our factor graph as a clique tree, with a set of vertex v $= \{$M users, N antennas$\}$, and a set of edges $e = \{\text{user}i - \text{antenna}j\}$. Please note that any graph that has a maximum clique size of 1 can also be treated with a clique tree. In this case, our factor graph is a clique tree with clique size 1. According to Koller et al. [35], the configuration **q** of a stationary point of a clique tree only exists when its energy potential has the following form:

$$f_{e_{i-j}} \propto \sum_{v_i - e_{i-j}} f_{v_i}\Big(\prod_{k \in Nb_i - j} f_{e_{k-i}}\Big) \tag{61}$$

$$b_i \propto f_{v_i}\Big(\prod_{j \in Nb_i - j} f_{e_{j-i}}\Big) \tag{62}$$

in which f is the energy of the vertex and the edges, and b is the belief of the vertex. Recall that Equation (33) can be rewritten as:

$$\begin{aligned} ln(q_{\mathbf{a}}^{t+1}(\mathbf{a})) &\propto \sum_{d_i} q_{d_i}^{t} \log p(\mathbf{r}|\mathbf{a},\mathbf{d}) \\ &\propto \sum_{d_i} q_{d_i}^{t} \log p(\mathbf{a},\mathbf{d},\mathbf{r}) \\ &\propto \sum_{d_i} q_{d_i}^{t} \log p(\mathbf{a}) \log p(\mathbf{d}) \log p(\mathbf{r}), \end{aligned} \tag{63}$$

which is exactly the form of Equation (61) if we treat **a**, **d**, **r** as the vertex, and the link between them as the edges. Therefore, the solution of our CEVTI is the stationary point of the variational optimization. □

7. Conclusions and Future Directions

To overcome the problems with the existing algorithms in channel state inference in WSNs, such as high complexity, poor generalization, impracticability and so on, we develop a new channel estimation method with variational tempering for MIMO-OFDM scenario. A ground truth information extraction algorithm based on variational tempering is proposed and implemented. Our CEVTI provides insights that account for multiple factors in inferring truth and can generalize to higher dimensions and finds better local optimum. As a variation of SVI, CEVTI can find out the optimal hyper-parameters of channels with fast convergence rate, and can be applied to the case of CDMA and uplink massive MIMO easily. As can be seen in Section 5, CEVTI can iteratively minimize the objective function with multiple dimensions. We demonstrate the performance of CEVTI through numerical simulation. The BER, convergence rate, and mutual information comparisons with the five existing CE algorithms show that CEVTI outperforms others under different noise variance and signal-to-noise ratio. Furthermore, the results show that the more parameters that are considered in each iteration, the faster the convergence rate and the lower the non-degenerate bit error rate with CEVTI. Analysis shows that CEVTI has satisfying computational complexity, and guarantees better local optimum. Therefore, this paper has contributed to the quest for developing efficient algorithms in artificial advanced sensor networks. Possible future research directions include investigation of how the graph structure can impact the performance of CEVTI, and constructing inference algorithms that can suit more complex situations.

Author Contributions: Conceptualization, J.L. and Y.C.; methodology, J.L. and Y.C.; software, J.L.; validation, M.L and N.C., and S.M.N.I.; writing–original draft preparation, J.L.; writing–review and editing, S.M.N.I.; supervision, Y.C. and N.C. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the National Natural Science Foundation of China (Grant No. 61802097), the Project of Qianjiang Talent (Grant No. QJD1802020), and the Key Research & Development Plan of Zhejiang Province (Grant No. 2019C01012).

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

WSNs	Wireless Sensor Networks
VI	Variational Inference
CEVTI	Channel Estimation Variational Tempering Inference
CE	Channel Estimation
CS	Compressive Sensing
CSI	Channel State Information
MP	Message Passing
ELBO	Evidence Lower Bound
MCMC	Monte Carlo Markov Chain

References

1. Alibakhshikenari, M.; Virdee, B.S.; Khalily, M.; Shukla, P.; See, C.H.; Abd-Alhameed, R.; Falcone, F.; Limiti, E. Beam-scanning leaky-wave antenna based on CRLH-metamaterial for millimetre-wave applications. *IET Microwaves Antennas Propag.* **2019**, *13*, 1129–1133. [CrossRef]
2. Alibakhshikenari, M.; Virdee, B.S.; Limiti, E. Compact Single-Layer Traveling-Wave Antenna Design Using Metamaterial Transmission Lines. *Radio Sci.* **2017**, *52*, 1510–1521. [CrossRef]
3. Mallat, N.K.; Ishtiaq, M.; Rehman, A.U.; Iqbal, A. Millimeter-Wave in the Face of 5G Communication Potential Applications. *IETE J. Res.* **2020**, *3*, 1–9. [CrossRef]
4. Mohammadi, M.; Kashani, F.H.; Ghalibafan, J. Backfire-to-endfire scanning capability of a balanced metamaterial structure based on slotted ferrite-filled waveguide. *Waves Random Complex Media* **2019**, 1–15. [CrossRef]
5. Awan, W.; Hussain, N.; Naqvi, S.; Iqbal, A.; Striker, R.; Mitra, D.; Braaten, B. A Miniaturized Wideband and Multi-band On-Demand Reconfigurable Antenna for Compact and Portable Devices. *AEU Int. J. Electron. Commun.* **2020**, *122*, 153266. [CrossRef]
6. Alibakhshikenari, M.; Virdee, B.S.; See, C.H.; Abdalhameed, R.; Falcone Lanas, F.J.; Limiti, E. High-gain metasurface in polyimide on-chip antenna based on CRLH-TL for sub-terahertz integrated circuits. *Sci. Rep.* **2020**, *10*, 4298. [CrossRef] [PubMed]
7. Watanabe, K.; Kojima, S.; Akao, T.; Katsuno, M.; Ahn, C.J. Modified Pilot Selection for Systematic Polar Coded MIMO-OFDM Channel Estimation. In Proceedings of the 2018 International Symposium on Intelligent Signal Processing and Communication Systems (ISPACS 2018), Okinawa, Japan, 27–30 November 2018; pp. 237–241.
8. Khan, I.; Singh, D. Efficient compressive sensing based sparse channel estimation for 5G massive MIMO systems. *AEU Int. J. Electron. Commun.* **2018**, *89*, 181–190. [CrossRef]
9. Motade, S.; Kulkarni, A. Channel Estimation and Data Detection Using Machine Learning for MIMO 5G Communication Systems in Fading Channel. *Technologies* **2018**, *6*, 72. [CrossRef]
10. Ma, X.; Yang, F.; Liu, S.; Song, J.; Han, Z. Sparse Channel Estimation for MIMO-OFDM Systems in High-Mobility Situations. *IEEE Trans. Veh. Technol.* **2018**, *67*, 6113–6124. [CrossRef]
11. Dai, J.; Liu, A.; Lau, V.K.N. FDD Massive MIMO Channel Estimation With Arbitrary 2D-Array Geometry. *IEEE Trans. Signal Process.* **2018**, *66*, 2584–2599. [CrossRef]
12. Nayebi, E.; Rao, B.D. Semi-blind Channel Estimation for Multiuser Massive MIMO Systems. *IEEE Trans. Signal Process.* **2018**, *66*, 540–553. [CrossRef]
13. Qiao, Y.T.; Yu, S.Y.; Su, P.C.; Zhang, L.J. Research on an iterative algorithm of LS channel estimation in MIMO OFDM systems. *IEEE Trans. Broadcast.* **2005**, *51*, 149–153. [CrossRef]
14. Simko, M.; Mehlfuhrer, C.; Wrulich, M.; Rupp, M. Doubly dispersive channel estimation with scalable complexity. In Proceedings of the 2010 International Itg Workshop on Smart Antennas (WSA 2010), Bremen, Germany, 23–24 February 2010; pp. 251–256
15. Wang, J.; Chen, H.; Li, S. Soft-output MMSE V-BLAST receiver with MMSE channel estimation under correlated Rician fading MIMO channels. *Wirel. Commun. Mob. Comput.* **2012**, *12*, 1363–1370. [CrossRef]
16. Daasch, R.R.; Shirley, C.G.; Chen, H.; Gao, F.; Ansari, N. Distributed Angle Estimation for Localization in Wireless Sensor Networks. *IEEE Trans. Wirel. Commun.* **2013**, *12*, 527–537.
17. Wang, X.; Poor, H.V. Iterative (turbo) soft interference cancellation and decoding for coded CDMA. *IEEE Trans. Commun.* **1999**, *47*, 1046–1061. [CrossRef]
18. Rao, X.; Lau, V.K.N. Distributed Compressive CSIT Estimation and Feedback for FDD Multi-User Massive MIMO Systems. *IEEE Trans. Signal Process.* **2014**, *62*, 3261–3271.
19. Tseng, C.C.; Wu, J.Y.; Lee, T.S. Enhanced Compressive Downlink CSI Recovery for FDD Massive MIMO Systems Using Weighted Block -Minimization. *IEEE Trans. Commun.* **2016**, *64*, 1055–1067. [CrossRef]
20. Paskin, M.; Guestrin, C.; McFadden, J. A robust architecture for distributed inference in sensor networks. In Proceedings of the 4th International Symposium on Information Processing in Sensor Networks (IPSN 2005), Los Angeles, CA, USA, 24–27 April 2005; pp. 55–62.
21. Bellili, F.; Sohrabi, F.; Yu, W. Generalized Approximate Message Passing for Massive MIMO mmWave Channel Estimation with Laplacian Prior. *IEEE Trans. Commun.* **2019**, *67*, 3205–3219. [CrossRef]

22. Ye, H.; Li, G.Y.; Juang, B.H.F. Power of Deep Learning for Channel Estimation and Signal Detection in OFDM Systems. *IEEE Wirel. Commun. Lett.* **2017**, *7*, 114–117. [CrossRef]
23. He, H.; Wen, C.-K.; Jin, S.; Li, G.Y. Deep learning-based channel estimation for beamspace mmwave massive mimo systems. *IEEE Wirel. Commun. Lett.* **2018**, *7*, 852–855. [CrossRef]
24. Zhang, Z.; Chen, H.; Hua, M.; Li, C.; Huang, Y.; Yang, L. Double Coded Caching in Ultra Dense Networks: Caching and Multicast Scheduling via Deep Reinforcement Learning. *IEEE Trans. Commun.* **2020**, *68*, 1071–1086. [CrossRef]
25. Ahmad, A.; Serpedin, E.; Nounou, H.; Nounou, M. Joint distributed parameter and channel estimation in wireless sensor networks via variational inference. In Proceedings of the 2012 Conference Record of the Forty Sixth Asilomar Conference on Signals, Systems and Computers (ASILOMAR), Pacific Grove, CA, USA, 4–7 November 2012; pp. 830–834.
26. Hu, B.; Land, I.; Rasmussen, L.; Piton, R.; Fleury, B. A Divergence Minimization Approach to Joint Multiuser Decoding for Coded CDMA. *IEEE J. Sel. Areas Commun.* **2008**, *26*, 432–445.
27. Kirkelund, G.E.; Manchon, C.N.I.; Christensen, L.P.B.; Riegler, E.; Fleury, B.H. Variational Message-Passing for Joint Channel Estimation and Decoding in MIMO-OFDM. In Proceedings of the 2010 IEEE Global Communications Conference (GLOBECOM 2010), Miami, FL, USA, 6–10 December 2010; pp. 1–6.
28. Cheng, X.; Sun, J.; Li, S. Channel Estimation for FDD Multi-User Massive MIMO: A Variational Bayesian Inference-Based Approach. *IEEE Trans. Wirel. Commun.* **2017**, *16*, 7590–7602. [CrossRef]
29. Thoota, S.S.; Murthy, C.R.; Annavajjala, R. Quantized Variational Bayesian Joint Channel Estimation and Data Detection for Uplink Massive MIMO Systems with Low resolution ADCS. In Proceedings of the 2019 IEEE 29th International Workshop on Machine Learning for Signal Processing (MLSP 2019), Pittsburgh, PA, USA, 13–16 October 2019; pp. 1–6.
30. Mandt, S.; Mcinerney, J.; Abrol, F.; Ranganath, R.; Blei, D. Variational Tempering. In Proceedings of the 19th International Conference on Artificial Intelligence and Statistics (AISTATS 2016), Cadiz, Spain, 9–11 May 2016; pp. 704–712.
31. Buchholz, A. Quasi-Monte Carlo Variational Inference. *arXiv* **2004**, arXiv:1807.01604v1
32. Wu, N.; Yuan, W.; Guo, Q.; Kuang, J. A Hybrid BP-EP-VMP Approach to Joint Channel Estimation and Decoding for FTN Signaling over Frequency Selective Fading Channels. *IEEE Access* **2017**, *5*, 6849–6858. [CrossRef]
33. Hoffman, M.D.; Blei, D.M.; Wang, C.; Paisley, J. Stochastic Variational Inference. *J. Mach. Learn. Res.* **2013**, *14*, 1303–1347.
34. Subrahmanyam, P.V. Brouwer's Fixed-Point Theorem. In *Elementary Fixed Point Theorems;* Springer: Berlin, Germany, 2018.
35. Koller, D.; Friedman, N.; Bach, F. *Probabilistic graphical models: Principles and techniques*; MIT Press: Cambridge, MA, USA, 2009.