This full text paper was peer reviewed at the direction of IEEE Communications Society subject matter experts for publication in the IEEE GLOBECOM 2005 proceedings.

A Variational Free Energy Minimization Interpretation of Multiuser Detection in CDMA

Darryl Dexu Lin and Teng Joon Lim
The Edward S. Rogers Sr. Department of Electrical and Computer Engineering, University of Toronto
10 King's College Road, Toronto, Ontario, Canada M5S 3G4
{linde, limtj}@comm.utoronto.ca

Abstract— **We propose a unified approach for deriving and studying multiuser detection algorithms using the concept of variational free energy minimization. Under this generalized framework, we readily arrive at many popular multiuser detection schemes. In addition to its systematic appeal, there are several other advantages of this viewpoint. First of all, by condensing the design of multiuser detectors into the selection of a few key probability distributions, namely $p(\mathbf{b})$, $p(\mathbf{r}|\mathbf{b})$ and $Q(\mathbf{b})$, we provide rigorous justifications for numerous detectors that were proposed on heuristic grounds and recommend new and improved designs. Furthermore, the free energy formulation facilitates convenient joint detection and decoding (utilizing the turbo principle) when error-control codes are incorporated, as well as efficient parameter estimation via the variational expectation maximization (EM) algorithm.**

I. Introduction

Multiuser detection (MUD) in CDMA has been an extensively studied area for a number of years [1]. In addition to the jointly-optimal (JO) and individually-optimal (IO) detectors, there are many sub-optimal multiuser detectors which may be categorized as linear or non-linear. These were derived using radically different approaches: minimizing mean squared error (MSE) at the output of a linear filter, estimating and cancelling interference user by user, and so on. This diversity of viewpoints makes the comparison and analysis of multiuser detectors difficult. Attempts to unify the diverse range of multiuser detectors have been numerous, among which [2] and [3] establish the uncoded linear and optimal detectors as posterior mean estimators of the *Bayes retrochannel* and evaluate their bit error rate (BER) performance in the large system limit, while [4] generalizes iterative multiuser joint decoding in coded CDMA as an approximate sum-product algorithm in a factor-graph, which leads to elegant performance analysis using *density evolution*.

This paper contributes to these pioneering works in that it serves as a natural extension of [2] and [3] into the realm of non-linear (and iterative) detectors and it supplements [4] as a probabilistic-inference interpretation of uncoded iterative MUD by temporarily removing the code constraints (instead of considering the channel and code constraints jointly). The technique we adopt is called *variational inference* [5, pp. 422–436], which, like the sum-product algorithm, is an approximate inference algorithm in probabilistic models. Our goal is to study multiuser detection in coded CDMA, but with focus on the multiuser channel layer by assuming the common MAP decoding in the convolutional code layer. The iterative decoding is naturally incorporated as message passing between the two layers through the relay nodes (Section IV-A).

The implications of this new generalized framework are significant. We will highlight this by examining the algorithms proposed in [6] and [7], which are arguably among the most important literature on turbo MUD. Much work has been devoted to their analysis, yet still little is known to strengthen their theoretical justifications, let alone revealing their intimate connections to each other. In this paper, we will show how both algorithms can be rigorously justified from a single unified principle and systematically improved. All these are accomplished within the framework of the *variational expectation-maximization* (variational EM) algorithm [8], which is a generalized EM algorithm with the exact inference in the E step replaced by variational inference. Central to the variational inference technique is the formulation of the *variational free energy*, which can be used to formulate both basic inference techniques and joint parameter estimation and data detection schemes.

The rest of the paper will be organized as follows: Section II describes the synchronous CDMA channel model and formulates the optimal multiuser detector; Section III introduces the variational inference technique and the unified multiuser detection framework; Section IV demonstrates how to rigorously derive turbo multiuser detectors within the variational EM formulation; Section V provides simulation results that verifies the performance of multiuser detectors incorporating adaptive parameter estimation; Section VI contains the conclusions.

II. Signal Model

Consider a synchronous DS-CDMA wireless link with K users. By sampling the chip matched filter output at chip rate, the received signal in one symbol interval, $\mathbf{r} \in \mathbb{R}^{N\times 1}$, can be written in the well-known vector form:

$$\mathbf{r} = \mathbf{S}\mathbf{A}\mathbf{b} + \mathbf{n}, \tag{1}$$

where $\mathbf{S} = [\mathbf{s}_1, \mathbf{s}_2, \cdots, \mathbf{s}_K]$ is the spreading code matrix consisting of the normalized spreading sequences of the K active users, $\mathbf{A} = \text{diag}(A_1, A_2, \cdots, A_K)$ is the channel matrix representing each user's signal amplitude and $\mathbf{b} = [b_1, b_2, \cdots, b_K]^T$ contains the transmitted BPSK channel bits from each user. $\mathbf{n}$ is a white Gaussian noise vector with distribution $p(\mathbf{n}) = \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I})$.

After bit-level matched filtering at the receiver, we may write the matched filter output, $\mathbf{y} \in \mathbb{R}^{K \times 1}$, as:

$$\mathbf{y} = \mathbf{S}^T \mathbf{r} = \mathbf{RAb} + \mathbf{z}, \tag{2}$$

where $\mathbf{R} = \mathbf{S}^T\mathbf{S}$ is the symmetric normalized signature correlation matrix with unit diagonal elements, and $\mathbf{z}$ is a coloured Gaussian noise vector with distribution $p(\mathbf{z}) = \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{R})$.

Given the prior distribution of $p(\mathbf{b})$ and the conditional distribution $p(\mathbf{r}|\mathbf{b})$, the jointly optimal (JO) detector is obtained by using Bayes rule to compute

$$p(\mathbf{b}|\mathbf{r}) = \frac{p(\mathbf{r}|\mathbf{b})p(\mathbf{b})}{\sum_{\mathbf{b}} p(\mathbf{r}|\mathbf{b})p(\mathbf{b})}. \tag{3}$$

For BPSK modulated information symbols, the decision is made as the binary vector $\mathbf{b} \in \{+1, -1\}^K$ that maximizes $p(\mathbf{b}|\mathbf{r})$. Similarly, the individually optimal (IO) detector is obtained by evaluating the marginal posterior distribution of b_k ($k = 1$ to K):

$$p(b_k|\mathbf{r}) = \frac{p(\mathbf{r}|b_k)p(b_k)}{\sum_{b_k} p(\mathbf{r}|b_k)p(b_k)}, \tag{4}$$

where $p(\mathbf{r}|b_k) = \sum_{\mathbf{b}\setminus b_k} p(\mathbf{r}|\mathbf{b})p(\mathbf{b})$. Due to the discrete nature of the information symbols, both JO and IO detectors entail prohibitive exponential complexity.

The IO detector is the optimal soft-in-soft-out (SISO) multiuser detector in terms of minimizing bit error rate (BER) and the JO is very close to optimal. It is well-known that practical suboptimal SISO multiuser detectors may be derived by taking in the *prior* information $p(b_k)$ and producing a *posterior* estimate $p(b_k|\mathbf{r})$ or $p(b_k|\mathbf{y})$ through some intelligent approximation which does not lead to exponential complexity. Section III will start this discussion by introducing a novel approach such that the "intelligent approximation" can be done systematically.

III. Multiuser Detection as Variational Inference

In [3], the linear multiuser detectors are treated as *posterior mean estimators* of the *Bayes retrochannel* with appropriately postulated distributions $p(\mathbf{b})$ and $p(\mathbf{r}|\mathbf{b})$. For example, if a Gaussian prior is assumed, i.e. $p(\mathbf{b}) = \mathcal{N}(\mathbf{0}, \mathbf{I})$, and the channel is modelled as $p(\mathbf{r}|\mathbf{b}) = \mathcal{N}(\mathbf{SAb}, \sigma^2\mathbf{I})$, the posterior mean estimator, i.e. $E(\mathbf{b}|\mathbf{r})$, is a generalized linear detector given by

$$\hat{\mathbf{b}} = \mathbf{A}^{-1}[\mathbf{S}^T\mathbf{S} + \sigma^2\mathbf{A}^{-2}]^{-1}\mathbf{S}^T\mathbf{r}. \tag{5}$$

By choosing different values for σ, we arrive at different linear detectors. If σ^2 equals the true noise variance, we get the MMSE detector. If $\sigma \to 0$, we approach the decorrelating detector. And if $\sigma \to \infty$, the matched filter output is attained.

However, this model has its limitations in that it excludes some of the useful nonlinear detectors as $p(\mathbf{b})$ can only be modelled as a continuous distribution for the evaluation of the posterior $p(\mathbf{b}|\mathbf{r})$ to be tractable[1], which is obviously unsatisfactory since $\mathbf{b}$ by nature takes on values in a discrete alphabet. In this paper, we wish to extend the coverage of the posterior mean estimator by introducing an additional degree of freedom in approximating the posterior distribution. More specifically, we will not limit ourselves to applying Bayes rule to calculate the posterior, but instead we use the more general and flexible *variational inference* technique.

A. Variational Inference and Free Energy

We shall explain the variational inference method specifically in terms of its application to multiuser detection, while a more general and in-depth treatment, as well as its connection to statistical physics, can be found in [5] and [9].

As stated earlier, the general task of the SISO multiuser detector is to perform inference on $\mathbf{b}$ given the observation $\mathbf{r}$ or $\mathbf{y}$ (we will simply use $\mathbf{r}$ from here on, as it is understood that both are equivalent). Suppose our objective is the JO detector, then the distribution of interest is $p(\mathbf{b}|\mathbf{r})$ (strictly speaking, IO detector minimizes the BER. But since the difference is minimal, we may consider the JO detector for simplicity). Very often, however, the direct evaluation of $p(\mathbf{b}|\mathbf{r})$ is computationally intractable when Bayes rule is applied directly, in particular, when $p(\mathbf{b})$ is a discrete distribution. In such a case, the variational inference technique assumes a tractable approximation to $p(\mathbf{b}|\mathbf{r})$, written as $Q(\mathbf{b})$, where the constant $\mathbf{r}$ is omitted.

A good approximation $Q(\mathbf{b})$ needs to resemble $p(\mathbf{b}|\mathbf{r})$ as closely as possible, and the *Kullback-Leibler divergence* (or *relative entropy*) $D(Q(\mathbf{b})\|p(\mathbf{b}|\mathbf{r}))$ offers an excellent measure of similarity. But since the computation of $p(\mathbf{b}|\mathbf{r})$ is intractable as we have assumed, an equivalent alternative, $p(\mathbf{b}, \mathbf{r})$, is used, where $p(\mathbf{b}, \mathbf{r})$ is called the *complete likelihood function*. $D(Q(\mathbf{b})\|p(\mathbf{b}, \mathbf{r}))$ is the *variational free energy*:

$$\mathcal{F}(Q, p) = \int_{\mathbf{b}} Q(\mathbf{b}) \log \frac{Q(\mathbf{b})}{p(\mathbf{b}, \mathbf{r})} d\mathbf{b}. \tag{6}$$

If no constraints are placed on $Q(\mathbf{b})$, by minimizing $\mathcal{F}(Q, p)$, we reach $Q(\mathbf{b}) = p(\mathbf{b}|\mathbf{r})$ and nothing is gained. Yet if we parameterize $Q(\mathbf{b})$ by assuming that it comes from a restricted family of distributions (for example, a Gaussian), we may very easily find its closed-form expression which approximates $p(\mathbf{b}|\mathbf{r})$ well by minimizing the variational free energy. This method of performing approximate inference is called *variational inference* [9].

One important instance of the variational method is to assume that $Q(\mathbf{b})$ is factorized as $\prod_{k=1}^{K} Q_k(b_k)$ (we shall omit the subscripts in Q_k from here on), and find independent distributions $\{Q(b_k)\}_{k=1}^{K}$ that minimize the free energy. This factorization of a distribution and the independence assumption associated with it is referred to as the *mean-field approximation* in statistical physics. A demonstration of its application will be presented in detail in Section IV-C.

B. Linear Multiuser Detectors and Free Energy Minimization

In this Section, we will derive the linear multiuser detectors from variational inference, and thus show that changing the postulated distributions $p(\mathbf{b})$, $p(\mathbf{r}|\mathbf{b})$ and $Q(\mathbf{b})$

[1] We loosely define tractability as computations that can be done with polynomial complexity.

is all that is needed to arrive at individual detectors. Although the exercises presented here are somewhat trivial, it lays the foundation for more sophisticated variations in Section IV.

Case 1: Linear Multiuser Detectors

$$\begin{aligned} p(\mathbf{b}) &= \mathcal{N}(\mathbf{0}, \mathbf{I}), \\ p(\mathbf{r}|\mathbf{b}) &= \mathcal{N}(\mathbf{SAb}, \sigma^2\mathbf{I}), \\ Q(\mathbf{b}) &= \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}). \end{aligned} \tag{7}$$

Proof: Evaluating $\mathcal{F}(Q, p)$ as in (6), we have a function of $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$:

$$\begin{aligned} \mathcal{F}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \;=\; & -\tfrac{1}{2}\log|\boldsymbol{\Sigma}| + \tfrac{1}{2\sigma^2}\{\boldsymbol{\mu}^T(\mathbf{A}^T\mathbf{S}^T\mathbf{SA} + \sigma^2\mathbf{I})\boldsymbol{\mu} \\ & + \mathrm{tr}[(\mathbf{A}^T\mathbf{S}^T\mathbf{SA} + \sigma^2\mathbf{I})\boldsymbol{\Sigma}] - 2\mathbf{r}^T\mathbf{SA}\boldsymbol{\mu}\} \end{aligned} \tag{8}$$

The final estimate of $Q(\mathbf{b})$ is given by the minimizers $\hat{\boldsymbol{\mu}}$ and $\hat{\boldsymbol{\Sigma}}$ of $\mathcal{F}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Calculating $\partial\mathcal{F}(\boldsymbol{\mu})/\partial\boldsymbol{\mu}$ and $\partial\mathcal{F}(\boldsymbol{\Sigma})/\partial\boldsymbol{\Sigma}^{-1}$ and equating to zero, we have

$$\begin{aligned} \hat{\boldsymbol{\mu}} &= (\mathbf{A}^T\mathbf{S}^T\mathbf{SA} + \sigma^2\mathbf{I})^{-1}\mathbf{A}^T\mathbf{S}^T\mathbf{r}; \\ \hat{\boldsymbol{\Sigma}} &= \sigma^2(\mathbf{A}^T\mathbf{S}^T\mathbf{SA} + \sigma^2\mathbf{I})^{-1}, \end{aligned} \tag{9}$$

where $\hat{\boldsymbol{\mu}}$ approaches the MMSE, decorrelating and matched filter detector outputs when σ takes on the its true value, 0 and ∞, respectively. Note that given the postulated priors in (7), the exact posterior $p(\mathbf{b}|\mathbf{r})$ is tractable and is in fact Gaussian. Therefore, the solved Q-function, $\mathcal{N}(\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\Sigma}})$, is the exact posterior distribution which could also have been found by applying Bayes rule directly. ■

C. IC-MUD and Coordinate Descent Algorithm

Iterative multiuser detectors, especially their convergence behavior, have been actively researched in the past. In [10], linear SIC and PIC are categorized as the Gauss-Seidel and Jacobi iterations for solving linear equations. SIC is also analyzed in greater depth in [11]. The study is later extended to clipped SIC in [12] through the investigation of the variational inequality (VI) problem. Here we offer an alternative interpretation for the SIC detectors as the coordinate descent algorithm applied to the minimization of variational free energy in (8).

Case 2: Linear/Clipped SIC Detectors
Given (7) and (8), in the i-th iteration, for $k = 1$ to K

$$\hat{\mu}_k^{(i)} = \arg\min_{\mu_k} \mathcal{F}(\mu_1^{(i)}, \cdots, \mu_{k-1}^{(i)}, \mu_k, \mu_{k+1}^{(i-1)}, \cdots, \mu_K^{(i-1)}) \tag{10}$$

s.t. $\mu_{min} \leq \mu_k \leq \mu_{max}$, describes a linear SIC if $\mu_{min} = -\infty$ and $\mu_{max} = \infty$, and a clipped SIC otherwise.

Proof: Formulating $\partial\mathcal{F}(\boldsymbol{\mu})/\partial\mu_k = 0$ based on (8) yields

$$A_k\mathbf{s}_k^T\mathbf{SA}\boldsymbol{\mu} - A_k\mathbf{s}_k^T\mathbf{r} + \sigma^2\mu_k = 0. \tag{11}$$

Rearranging the terms and defining $\boldsymbol{\mu}_{\backslash k} = [\mu_1, \cdots, \mu_{k-1}, 0, \mu_{k+1}, \cdots, \mu_K]^T$, the optimal μ_k is then expressed in the familiar interference cancellation form:

$$\hat{\mu}_k = \frac{1}{A_k^2 + \sigma^2}A_k\mathbf{s}_k^T(\mathbf{r} - \mathbf{SA}\boldsymbol{\mu}_{\backslash k}), \tag{12}$$

if μ_k is unbounded (i.e. $\mu_{min} = -\infty$ and $\mu_{max} = \infty$). Since updating μ_k ($k = 1, \cdots, K$) consecutively subject to $\partial\mathcal{F}(\boldsymbol{\mu})/\partial\mu_k = 0$ is the coordinate descent algorithm for minimizing $\mathcal{F}(\boldsymbol{\mu})$, then similar to the linear detectors, σ taking its true value corresponds to the coordinate descent implementation of the MMSE detector, and $\sigma \to 0$ corresponds to the coordinate descent implementation of the decorrelating detector. On the other hand, if μ_{min} and μ_{max} are finite, we need to solve (11) subject to $\mu_{min} \leq \mu_k \leq \mu_{max}$, which corresponds to clipped SIC. ■

To verify that (12) does converge to MMSE or decorrelator solutions, and to gain further insight into the convergence behavior when the optimization constraints are active (Clipped SIC), we invoke the following theorem [13]:

Theorem 1: Consider an optimization problem:

$$min\ f(\mathbf{x}) = g(\mathbf{Ex}) + \mathbf{c}^T\mathbf{x}, \quad s.t.\ \mathbf{x} \in \mathcal{X}, \tag{13}$$

where $\mathcal{X}$ is a box (possibly unbounded) in $\mathbb{R}^n$, f is a proper closed convex function in $\mathbb{R}^n$, g is a proper closed convex function in $\mathbb{R}^m$, $\mathbf{E}$ is an $m \times n$ matrix having no zero column, and $\mathbf{c} \in \mathbb{R}^n$. Also assume

1) *The set of optimal solutions for (13), denoted by $\mathcal{X}^*$ is nonempty;*
2) *dom g is open and g is strictly convex twice continuously differentiable on dom g;*
3) *$\nabla^2 g(\mathbf{Ex}^*)$ is positive definite for all $\mathbf{x}^* \in \mathcal{X}^*$.*

Then if $\{\mathbf{x}^r\}$ is a sequence of iterates generated by coordinate descent method according to the Almost Cyclic Rule or Gauss-Southwell Rule, $\{\mathbf{x}^r\}$ converges at least linearly to an element of $\mathcal{X}^*$.

Since the objective function of optimization, $\mathcal{F}(\boldsymbol{\mu})$, satisfies all conditions in the theorem when the spreading codes are linearly independent, it is clear that this theorem applies to the general linear/clipped SIC setting. Also due to the objective function being quadratic and the constraints being linear, there is a unique optimal solution in $\mathcal{X}^*$. We may thus conclude that linear and clipped SIC are guaranteed to converge to the unique minimum free energy defined by $\mathcal{F}(\boldsymbol{\mu})$ and the constraint ($\mu_{min} \leq \mu_k \leq \mu_{max}$), and the rate of convergence is at least linear. This result is rather significant, and is proved for the first time to our knowledge.

Furthermore, we may relax the conventional cyclic order of iteration for SIC and assert that as long as the coordinates are iterated upon according to either the *Almost Cyclic Rule* or *Gauss-Southwell Rule*, we still can guarantee the at least linear convergence rate. For details regarding the relaxed iteration rules, please refer to [13].

IV. Turbo-MUD and Variational EM Algorithm

With the preliminaries laid out in the preceding sections, we are now ready to investigate more sophisticated implications of this free energy formulation. In Section IV-A, we will first describe how to interface the variational inference multiuser detectors with a MAP convolutional decoder by following the message-passing rule in graphs. The detector and decoder combined in this fashion will be used to justify the formulation of turbo multiuser detection. In Section IV-B, we will introduce an extension of the variational inference algorithm, called variational EM, to facilitate joint parameter estimation in turbo multiuser detection. Then in the two case studies in Sections IV-C and IV-D, we will explore how the variational EM algorithm provides a concrete explanation for existing algorithms and how it leads to additional improvements.

A. Message-Passing in Coded CDMA

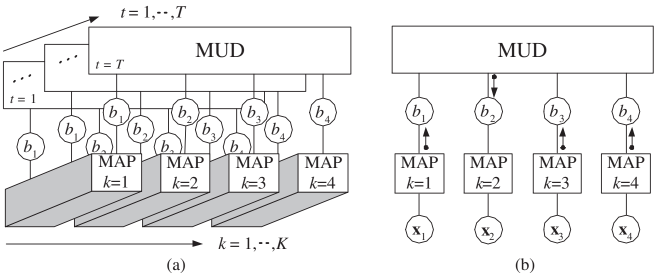

Fig. 1. Message-passing in coded CDMA.

From Fig. 1(a), it is seen that the nodes representing the channel bits $\{b_{t,k}\}_{k=1}^K$ are the relay nodes that separate the graph into two halves, where on one side the decoder runs belief propagation (BCJR) to perform MAP decoding locally and on the other side the multiuser detector runs variational inference. The process by which the MAP decoder retrieves prior information and generates extrinsic information is standard (see [7]) and will be skipped. Here we concentrate on how the multiuser detector based on variational inference accepts and generates information. Although the variational inference is generally not the exact inference, we will assume that $Q(b_k)$ is a good approximation to the exact posterior. Following the message passing rule stated in [4], $Q(b_k)$ calculated by ignoring the prior $p(b_k)$ generates the extrinsic information sent back to the decoder. Fig. 1(b) depicts how a multiuser detector of $K = 4$ users accepts the incoming messages from the decoder for $\{b_k\}_{k\neq 2}$(which are used as the prior $p(\mathbf{b})$), to generate the extrinsic information for b_2 to send back to the MAP decoder. Such a procedure is repeated for all $k = 1, \cdots, K$. After several rounds of message passing between multiuser detectors and MAP decoders, the MAP decoders make final decisions on the information bits $\{\mathbf{x}_k\}_{k=1}^K$.

B. Variational EM

While the variational inference technique offers a powerful means to perform data detection, it assumes perfect knowledge of system parameters, such as the noise variance and channel amplitudes. One way to incorporate the uncertainties of these parameters in the model is through the variational EM algorithm [8].

Denoting the unknown parameter as θ, we assume θ remains static over T i.i.d. realizations of the channel. In the context of CDMA, this implies that we assume θ, the noise variance σ^2 for example, remains constant when a block of T bits are transmitted by each user (T could be the code word length). The variational EM algorithm extracts point estimates for θ and postulated posterior over the channel bits. Therefore, the new Q-function may be written as:

$$Q(\theta, \mathbf{b}_1, \cdots, \mathbf{b}_T) = \delta(\theta - \hat{\theta}) \prod_{t=1}^{T} Q(\mathbf{b}_t), \tag{14}$$

where $\mathbf{b}_t$ contains the channel bits carried in the t-th realization of the channel. Recall that for i.i.d. data, $p(\mathbf{b}, \mathbf{r}, \theta) = p(\theta)\prod_{t=1}^T p(\mathbf{b}_t, \mathbf{r}_t|\theta)$. Substituting (14) into (6) yields the following free energy:

$$\mathcal{F} = -\log p(\hat{\theta}) + \sum_{t=1}^{T} \left(\int_{\mathbf{b}_t} Q(\mathbf{b}_t) \log \frac{Q(\mathbf{b}_t)}{p(\mathbf{b}_t, \mathbf{r}_t|\hat{\theta})} d\mathbf{b}_t \right). \tag{15}$$

The first term on the right hand side of the equation constitutes the prior knowledge of the parameter. While this is an important factor in general, in this paper we assume no parameter prior knowledge, i.e. $p(\theta) = constant$.

As proven in [14], alternating between minimizing (15) w.r.t. $\{Q(\mathbf{b}_t)\}_{t=1}^T$ in the E step, and w.r.t. $\hat{\theta}$ in the M Step leads to the exact EM algorithm where $\{\mathbf{b}_t\}_{t=1}^T$ are the "complete data" and $\hat{\theta}$ is the unknown parameter of interest. Unfortunately, the exact EM is only possible in special cases because the solution to the E step is $Q(\mathbf{b}_t) = p(\mathbf{b}_t|\mathbf{r}_t, \hat{\theta})$ (s.t. $\int_{\mathbf{b}_t} Q(\mathbf{b}_t) d\mathbf{b}_t = 1$), which is often intractable. But suppose we use a postulated (and simple) distribution $Q(\mathbf{b}_t)$, with parameter λ_t, and then find λ_t that minimizes (15). We then arrive at the *variational EM* algorithm, which consists of the initialization plus the E step and M step in the i-th iteration:

Initialization Choose initial values for $\hat{\theta}^{(0)}$.
E Step Minimize $\mathcal{F}(\lambda_1, \cdots, \lambda_T, \hat{\theta}^{(i-1)})$ in (15) w.r.t. λ_t

$$\lambda_t^{(i)} = \arg\min_{\lambda_t} \int_{\mathbf{b}_t} Q(\mathbf{b}_t) \log \frac{Q(\mathbf{b}_t)}{p(\mathbf{b}_t, \mathbf{r}_t|\hat{\theta}^{(i-1)})} d\mathbf{b}_t, \tag{16}$$

for $t = 1, \cdots, T$.
M Step Minimize $\mathcal{F}(\lambda_1^{(i)}, \cdots, \lambda_T^{(i)}, \hat{\theta})$ in (15) w.r.t. $\hat{\theta}$

$$\hat{\theta}^{(i)} = \arg\min_{\hat{\theta}} \sum_{t=1}^{T} \left(\int_{\mathbf{b}_t} Q^{(i)}(\mathbf{b}_t) \log \frac{Q^{(i)}(\mathbf{b}_t)}{p(\mathbf{b}_t, \mathbf{r}_t|\hat{\theta})} d\mathbf{b}_t \right). \tag{17}$$

C. Mean-Field Multiuser Detection

In [6], a surprisingly simple (linear complexity) multiuser detector was proposed for coded CDMA producing near optimal performance at very high network load. The authors applied a simple interference cancellation scheme and made the following approximation:

$$\begin{aligned} & p(y_k|b_k, \mathbf{b}_{\backslash k} = \hat{\mathbf{b}}_{\backslash k}) \\ = \ & \tfrac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{ -\tfrac{1}{2\sigma^2}(y_k - A_k \mathbf{s}_k^T \mathbf{S}\mathbf{A}\hat{\mathbf{b}}_{\backslash k} - A_k^2 b_k)^2 \right\} \\ = \ & \tfrac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{ -\tfrac{1}{2\sigma^2}[A_k^2 b_k - A_k \mathbf{s}_k^T(\mathbf{r} - \mathbf{S}\mathbf{A}\hat{\mathbf{b}}_{\backslash k})]^2 \right\} \end{aligned} \tag{18}$$

where $\hat{\mathbf{b}}$ is the average bit estimate received from the MAP decoder, and $\hat{\mathbf{b}}_{\backslash k} = [\hat{b}_1, \cdots, \hat{b}_{k-1}, 0, \hat{b}_{k+1}, \cdots, \hat{b}_K]^T$. Here $\sigma^2 = \sigma_n^2 + \sigma_{MU}^2$ is the combination of channel noise and multiple access interference modelled as Gaussian noise, which can be heuristically computed as

$$\sigma^2 = \text{var}[A_k^2 \hat{b}_k - A_k \mathbf{s}_k^T (\mathbf{r} - \mathbf{SA}\hat{\mathbf{b}}_{\backslash k})]. \tag{19}$$

The soft bit decision is then drawn from (18) and fed back to the MAP channel code decoder for the algorithm to iterate. Though free energy appears rather distant from this simple while effective detection scheme, we will show that it is exactly an instance of the variational EM algorithm when σ^2 is an unknown parameter to be estimated. To generalize, we will assume a channel noise covariance matrix $\mathbf{\Phi}$ instead of white noise.

Case 3: Mean-Field Multiuser Detector

$$\begin{aligned} p(\mathbf{b}) &= \prod_{k=1}^K \xi_k^{\frac{1+b_k}{2}} (1-\xi_k)^{\frac{1-b_k}{2}}, \quad b_k \in \{\pm 1\}, \\ p(\mathbf{r}|\mathbf{b}) &= \mathcal{N}(\mathbf{SAb}, \mathbf{\Phi}), \\ Q(\mathbf{b}) &= \prod_{k=1}^K \gamma_k^{\frac{1+b_k}{2}} (1-\gamma_k)^{\frac{1-b_k}{2}}, \quad b_k \in \{\pm 1\}. \end{aligned} \tag{20}$$

Proof: In (20), we omitted the realization index in $\mathbf{b}_t$ and $\mathbf{r}_t$ for simplicity since the E step in (16) is invariant w.r.t. t. Here ξ_k and γ_k are the prior and approximate posterior probability of b_k being 1.

In the E step, we keep $\mathbf{\Phi}$ constant. Hence, with some mathematical manipulation, we have

$$\begin{aligned} \mathcal{F}(\gamma, \mathbf{\Phi}) &= \sum_{k=1}^K \gamma_k \log \frac{\gamma_k}{\xi_k} + (1-\gamma_k) \log \frac{1-\gamma_k}{1-\xi_k} \\ &\quad + 2[\gamma^T \mathbf{B} \gamma - (\mathbf{1}^T \mathbf{B} + \mathbf{r}^T \mathbf{\Phi}^{-1} \mathbf{SA}) \gamma], \end{aligned} \tag{21}$$

where $\mathbf{B} = \mathbf{A}^T \mathbf{S}^T \mathbf{\Phi}^{-1} \mathbf{SA} - \text{diag}(\mathbf{A}^T \mathbf{S}^T \mathbf{\Phi}^{-1} \mathbf{SA})$. (21) is obtained by utilizing the property that $E(\mathbf{b}^T \mathbf{C} \mathbf{b}) = (2\gamma - \mathbf{1})^T [\mathbf{C} - \text{diag}(\mathbf{C})](2\gamma - \mathbf{1}) + \mathbf{1}^T \text{diag}(\mathbf{C}) \mathbf{1}$ for $\mathbf{b} \in \{\pm 1\}^K$. $\text{diag}(\mathbf{C})$ denotes a diagonal matrix with the diagonal elements of square matrix $\mathbf{C}$ on its diagonal.

Rearranging $\partial \mathcal{F}(\gamma, \mathbf{\Phi}) / \partial \gamma_k = 0$ gives

$$\log \frac{\gamma_k}{1-\gamma_k} = \log \frac{\xi_k}{1-\xi_k} + 2[\boldsymbol{\eta}_k^T \mathbf{r} - \boldsymbol{\beta}_k^T (2\gamma - \mathbf{1})], \tag{22}$$

where $\boldsymbol{\eta}_k$ and $\boldsymbol{\beta}_k$ are the k-th column vector of $\mathbf{\Phi}^{-1}\mathbf{SA}$ and $\mathbf{A}^T \mathbf{S}^T \mathbf{\Phi}^{-1} \mathbf{SA}$, respectively. Applying the message passing rule discussed in Section IV-A, we extract the extrinsic information by setting $\xi_k = 1/2$. Let $m_k = 2\gamma_k - 1$ represent the posterior average of the estimate for b_k, then from (22) m_k can be evaluated as

$$m_k = \tanh(\boldsymbol{\eta}_k^T \mathbf{r} - \boldsymbol{\beta}_k^T \mathbf{m}). \tag{23}$$

This equation describes a coordinate descent algorithm to minimize the free energy in (21) with $\xi_k = 1/2$, whose convergence cannot be guaranteed, however, since (21) is not a strictly convex function. In [6], instead of solving a system of K nonlinear equations, an approximation is essentially made by letting $m_k = \tanh(\boldsymbol{\eta}_k^T \mathbf{r} - \boldsymbol{\beta}_k^T \hat{\mathbf{b}})$ (which follows from (18)). Thus we have completed the E step and rigorously justified the detection step in [6].

In the M step, we consider all T realizations of the channel to estimate $\mathbf{\Phi}$. Therefore,

$$\begin{aligned} \mathcal{F}(\gamma, \mathbf{\Phi}) &= \sum_{t=1}^T \tfrac{1}{2} \log |\mathbf{\Phi}| - \mathbf{r}^{T(t)} \mathbf{\Phi}^{-1} \mathbf{SAm}_t + \tfrac{1}{2} \mathbf{r}^{T(t)} \mathbf{\Phi}^{-1} \mathbf{r}_t \\ &\quad + \tfrac{1}{2} \mathbf{m}^{T(t)} [\mathbf{A}^T \mathbf{S}^T \mathbf{\Phi}^{-1} \mathbf{SA} - \text{diag}(\mathbf{A}^T \mathbf{S}^T \mathbf{\Phi}^{-1} \mathbf{SA})] \mathbf{m}_t \end{aligned} \tag{24}$$

Solving $\partial \mathcal{F}(\gamma, \mathbf{\Phi}) / \partial \mathbf{\Phi}^{-1} = \mathbf{0}$ as in (17) yields (we omit the technical details of this operation for limited space):

$$\begin{aligned} \mathbf{\Phi} &= \tfrac{1}{T} \sum_{t=1}^T \{(\mathbf{r}_t - \mathbf{SAm}_t)(\mathbf{r}_t - \mathbf{SAm}_t)^T \\ &\quad + \sum_{k=1}^K (1 - m_k^{2(t)}) A_k^2 \mathbf{s}_k \mathbf{s}_k^T\}. \end{aligned} \tag{25}$$

It is clear that $m_k \to \{\pm 1\}$ as the algorithm converges. Hence, the last term in (25) eventually vanishes. By omitting the vanishing term, this is exactly the equation to estimate $\sigma^2 = \sigma_n^2 + \sigma_{MU}^2$ in [6] in the case of white noise. Together with (23), we have provided a theoretical explanation for [6], and extended it to the case of arbitrary noise. ■

In preparing this paper, we discovered that a similar mean-field multiuser detector has been independently proposed in [15], but our paper formulates the coded version and establishes the connection to [6] through variational EM.

D. SISO MMSE Multiuser Detection

A rather different turbo multiuser detection scheme was proposed in [7] which involves a two stage process: First, the soft bit estimate from the MAP decoder is remodulated and subtracted from the matched filter output; Second, a linear MMSE filter is used to further suppress the residual interference. We will now demonstrate that with variational EM formulation, the two-stage process can be derived from a single optimization procedure.

Case 4: SISO MMSE Multiuser Detector

$$\begin{aligned} p(\mathbf{b}) &= \mathcal{N}(\tilde{\mathbf{b}}, \mathbf{W}), \\ p(\mathbf{r}|\mathbf{b}) &= \mathcal{N}(\mathbf{SAb}, \mathbf{\Phi}), \\ Q(\mathbf{b}) &= \mathcal{N}(\boldsymbol{\mu}, \mathbf{\Sigma}). \end{aligned} \tag{26}$$

Proof: In (26), $\mathbf{W} = \text{diag}([1 - \tilde{b}_1^2, \cdots, 1 - \tilde{b}_K^2]^T)$ and $\tilde{\mathbf{b}} = [\tilde{b}_1, \cdots, \tilde{b}_K]^T$ are the soft bit estimates from the MAP decoder. We assume perfect knowledge of $\mathbf{\Phi}$ in the E step and it can be shown that

$$\begin{aligned} \mathcal{F}(\boldsymbol{\mu}, \mathbf{\Sigma}, \mathbf{\Phi}) &= \tfrac{1}{2} [\boldsymbol{\mu}^T (\mathbf{A}^T \mathbf{S}^T \mathbf{\Phi}^{-1} \mathbf{SA} + \mathbf{W}^{-1}) \boldsymbol{\mu} \\ &\quad - 2(\mathbf{r}^T \mathbf{\Phi}^{-1} \mathbf{SA} + \tilde{\mathbf{b}} \mathbf{W}^{-1}) \boldsymbol{\mu}]. \end{aligned} \tag{27}$$

Solving $\partial \mathcal{F}(\boldsymbol{\mu}, \mathbf{\Sigma}, \mathbf{\Phi}) / \partial \boldsymbol{\mu} = \mathbf{0}$, we arrive at

$$\boldsymbol{\mu} - \tilde{\mathbf{b}} = (\mathbf{A}^T \mathbf{S}^T \mathbf{\Phi}^{-1} \mathbf{SA} + \mathbf{W}^{-1})^{-1} \mathbf{A}^T \mathbf{S}^T \mathbf{\Phi}^{-1} (\mathbf{r} - \mathbf{SA}\tilde{\mathbf{b}}). \tag{28}$$

The extrinsic information is extracted following the rule described in Section IV-A, by ignoring the prior information for b_k. That is, for the k-th bit, we define $\tilde{\mathbf{b}}_k = [\tilde{b}_1, \cdots, \tilde{b}_{k-1}, 0, \tilde{b}_{k+1}, \cdots, \tilde{b}_K]^T$ and $\mathbf{W}_k$ similarly. And thus

$$\mu_k = \mathbf{e}_k^T (\mathbf{A}^T \mathbf{S}^T \mathbf{\Phi}^{-1} \mathbf{SA} + \mathbf{W}_k^{-1})^{-1} \mathbf{A}^T \mathbf{S}^T \mathbf{\Phi}^{-1} (\mathbf{r} - \mathbf{SA}\tilde{\mathbf{b}}_k), \tag{29}$$

where $\mathbf{e}_k$ denotes a K-vector of all zeros, except for the k-th element being 1. Assuming white noise, i.e. $\mathbf{\Phi} = \sigma^2 \mathbf{I}$, we get, after some matrix manipulation

$$\mu_k = A_k \mathbf{e}_k^T[\mathbf{A}^T\mathbf{W}_k\mathbf{A}+\sigma^2(\mathbf{S}^T\mathbf{S})^{-1}]^{-1}[(\mathbf{S}^T\mathbf{S})^{-1}\mathbf{S}\mathbf{r}-\mathbf{A}\tilde{\mathbf{b}}_k], \tag{30}$$

which is equivalent to z_k in [7]. From μ_k, the extrinsic information for b_k can be generated.

It is also straightforward to incorporate the iterative noise variance estimation by continuing with the M step similar to Case 3. Considering the entire data block of size T,

$$\begin{aligned}\mathcal{F}(\boldsymbol{\mu},\boldsymbol{\Sigma},\boldsymbol{\Phi}) = & \textstyle\sum_{t=1}^{T}\frac{1}{2}[(\mathbf{r}_t-\mathbf{S}\mathbf{A}\boldsymbol{\mu}_t)^T\boldsymbol{\Phi}^{-1}(\mathbf{r}_t-\mathbf{S}\mathbf{A}\boldsymbol{\mu}_t) \\ & +\log|\boldsymbol{\Phi}|+\mathrm{tr}(\mathbf{A}^T\mathbf{S}^T\boldsymbol{\Phi}^{-1}\mathbf{S}\mathbf{A}\boldsymbol{\Sigma}_t)].\end{aligned} \tag{31}$$

Solving $\partial\mathcal{F}(\boldsymbol{\mu},\boldsymbol{\Sigma},\boldsymbol{\Phi})/\partial\boldsymbol{\Phi}^{-1}=\mathbf{0}$ yields

$$\boldsymbol{\Phi} = \frac{1}{T}\sum_{t=1}^{T}(\mathbf{r}_t-\mathbf{S}\mathbf{A}\boldsymbol{\mu}_t)(\mathbf{r}_t-\mathbf{S}\mathbf{A}\boldsymbol{\mu}_t)^T+\mathbf{S}\mathbf{A}\boldsymbol{\Sigma}_t\mathbf{A}^T\mathbf{S}^T, \tag{32}$$

where $\boldsymbol{\Sigma}$ is found in the E step to be $\boldsymbol{\Sigma}=(\mathbf{A}^T\mathbf{S}^T\boldsymbol{\Phi}^{-1}\mathbf{S}\mathbf{A}+\mathbf{W}^{-1})^{-1}$. When the noise is white, we then have an update for σ^2 of the form

$$\begin{aligned}\sigma^2 = & \textstyle\frac{1}{KT}\sum_{t=1}^{T}[(\mathbf{y}_t-\mathbf{R}\mathbf{A}\boldsymbol{\mu}_t)^T\mathbf{R}^{-1}(\mathbf{y}_t-\mathbf{R}\mathbf{A}\boldsymbol{\mu}_t) \\ & +\mathrm{tr}(\mathbf{A}^T\mathbf{R}\mathbf{A}\boldsymbol{\Sigma}_t)].\end{aligned} \tag{33}$$

■

V. Simulations

In this section we briefly study the performance of the iterative noise variance estimation algorithm as an improvement to [7]. In the simulation, we perform the M step before the E step by letting $\boldsymbol{\mu}_t^{(0)}=\mathbf{0}$ and $\boldsymbol{\Sigma}_t^{(0)}=\mathbf{0}$.

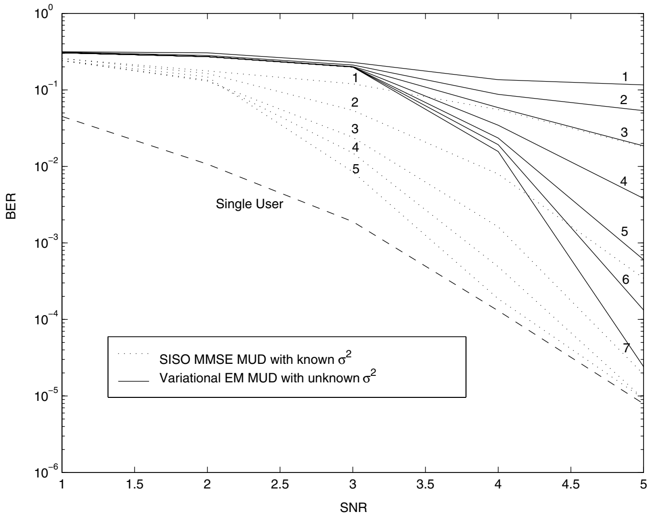

Fig. 2. BER vs. SNR performance for SISO MMSE detector in turbo MUD.

Fig. 2 compares the performance of the SISO MMSE multiuser detector with perfect channel knowledge compared to unknown noise variance. The settings are the same as in [7]: A four-user system is assumed with equal cross-correlation $\rho_{ij}=0.7$. All users have equal power and employ the same rate $1/2$ convolutional code with generators 10011 and 11101. The variational EM algorithm is utilized to iteratively update the estimate for σ^2 in each round of iteration. It is seen that at high SNR, there is no performance degradation as the variational EM detector converges to the single user bound after seven iterations similar to the SISO MMSE detector with perfect noise variance information.

VI. Conclusions

The concept of free energy is a far-reaching one [8][16]. One contribution of this paper is adopting the notion of free energy minimization in establishing and analyzing multiuser detectors. Furthermore, we extended the initiative to variational-EM-based multiuser detectors, in which channel parameters may be efficiently and blindly estimated in conjunction with turbo multiuser detection. This theoretical study finds an interesting fundamental link between two celebrated turbo multiuser detectors in [6] and [7]. We believe such an elegant generalization is not coincidental, but just one example of many potential applications of variational inference in communications.

References

[1] S. Verdù, *Multiuser Detection*. Cambridge, U.K.: Cambridge University Press, 1998.

[2] T. Tanaka, "A statistical-mechanics approach to large-system analysis of CDMA multiuser detectors," *IEEE Trans. Info. Theory*, vol. 48, no. 11, pp. 2888–2910, Nov. 2002.

[3] D. Guo and S. Verdú, "Randomly spread CDMA: asymptotics via statistical physics," *IEEE Trans. Info. Theory*, vol. 51, no. 6, pp. 1983–2010, June 2005.

[4] J. Boutros and G. Caire, "Iterative multiuser joint decoding: Unified framework and asymptotic analysis," *IEEE Trans. Info. Theory*, vol. 48, no. 7, pp. 1772–1793, July 2002.

[5] D. MacKay, *Information Theory, Inference and Learning Algorithms*. Cambridge U.K.: Cambridge University Press, 2003.

[6] P. D. Alexander, A. J. Grant, and M. C. Reed, "Iterative detection in code-division multiple-access with error control coding," *Euro. Trans. Telecommun.*, vol. 9, no. 5, pp. 419–426, Oct. 1998.

[7] X. Wang and H. V. Poor, "Iterative (turbo) soft interference cancellation and decoding for coded CDMA," *IEEE Trans. Comms.*, vol. 47, no. 7, pp. 1046–1061, July 1999.

[8] B. J. Frey and N. Jojic, "Advances in algorithms for inference and learning in complex probability models for vision," *to appear on IEEE Trans. Patt. Anal. Mach. Int.*

[9] B. J. Frey, *Graphical Models for Machine Learning and Digital Communication*. Cambridge, Massachusetts: MIT Press, 1998.

[10] H. Elders-Boll, H. D. Schotten, and A. Busboom, "Efficient implementation of linear multiuser detectors for asynchronous CDMA systems by linear interference cancellation," *Euro. Trans. Telecommun.*, vol. 9, no. 5, pp. 427–438, Oct. 1998.

[11] L. K. Rasmussen, T. J. Lim, and A. L. Johansson, "A matrix-algebraic approach to successive interference cancellation in CDMA," *IEEE Trans. Comms.*, vol. 48, no. 1, pp. 145–151, Jan. 2000.

[12] P. H. Tan, L. K. Rasmussen, and T. J. Lim, "Constrained maximum-likelihood detection in CDMA," *IEEE Trans. Comms.*, vol. 49, no. 1, pp. 142–153, Jan. 2001.

[13] Z. Q. Luo and P. Tseng, "On the convergence of the coordinate descent method for convex differentiable minimization," *J. Optim. Theory Appl.*, vol. 72, no. 1, pp. 7–35, Jan. 1992.

[14] R. M. Neal and G. E. Hinton, "A view of the EM algorithm that justifies incremental, sparse, and other variants," in *Learning in Graphical Models*, M. I. Jordan, Ed. Kluwer Academic Publishers, 1998.

[15] T. Fabricius and O. Nøklit, "Approximations to joint-ML and ML symbol-channel estimators in MUD CDMA," *Proc. Globecom'02*, vol. 1, pp. 389–393, 2002.

[16] J. Yedidia, W. T. Freeman, and Y. Weiss, "Generalized belief propagation," in *Advances in Neural Information Processing Systems 13*. MIT Press, 2001, pp. 689–695.