A Variational Free Energy Minimization Interpretation of Multiuser Detection in CDMA

Darryl Dexu Lin and Teng Joon Lim

University of Toronto

December 1, 2005

Presented at Globecom 2005, St. Louis, MO, on December 1, 2005.

Outline

- Multi-user signal model, with and without coding;
- Multiuser detection with channel coding – turbo MUD;
- Variational inference – free energy, how to use it for approximate inference, and links to Exp.-Max. (EM) algorithm;
- MUD as variational inference – unified framework for both turbo and uncoded MUD.
- Summary.

Multiuser Received Signal Model

- Synchronous DS-CDMA, K users, received signal in t-th symbol interval is

$$\mathbf{r}_t = \mathbf{S}_t \mathbf{b}_t + \mathbf{n}_t, \quad t = 1, \ldots, T. \tag{1}$$

  - $\mathbf{r}_t = [r_{1,t}, \cdots, r_{N,t}]^T$ is received data vector sampled at chip rate;
  - $\mathbf{S}_t = [\mathbf{s}_{1,t}, \mathbf{s}_{2,t}, \cdots, \mathbf{s}_{K,t}]$ is the matrix of received pulse waveforms;
  - $\mathbf{b}_t = [b_{1,t}, b_{2,t}, \cdots, b_{K,t}]^T$ contains the channel symbols;
  - $\mathbf{n}_t$ is the noise vector.

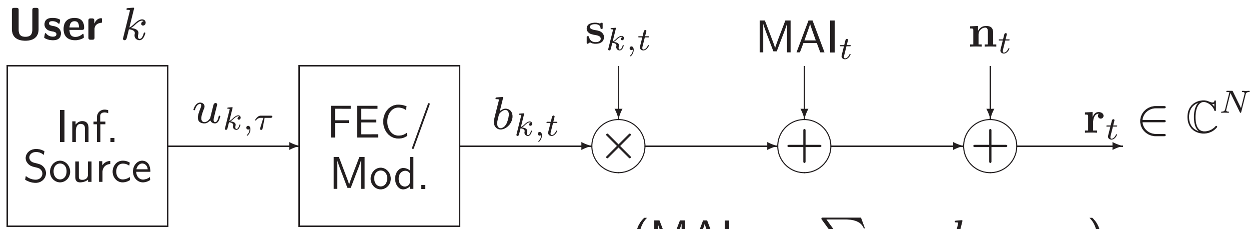
- Equivalently, the matched filter output can be used as sufficient statistics:

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

  where $\mathbf{R}_t = \mathbf{S}_t^T \mathbf{S}_t$.

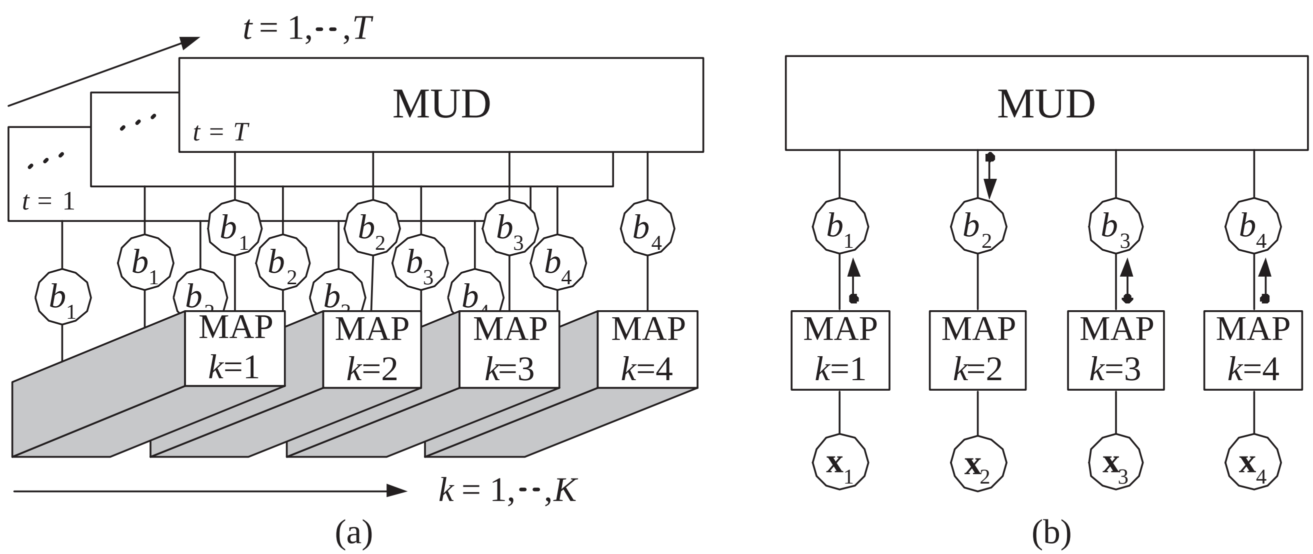
- Note: $\mathbf{b}_t$ can represent coded or uncoded symbols so the model applies to both cases.

Conceptual Block Diagram



$$(\mathsf{MAI}_t = \sum_{j \neq k} b_{k,t} \mathbf{s}_{k,t})$$

Encoder for user k



Multi-user decoder. Note time dependency among symbols of the same user, and user dependency among symbols at the same time.

Main Points

- Many sub-optimal MUD's have been derived using different approaches e.g. linear MMSE, multi-stage SIC, etc.

- Turbo multi-user detection (MUD) has a well-understood, standard error decoding section, but an ad-hoc MUD section:
  - Optimal APP updates for the multi-access interference (MAI) channel is not feasible $\Rightarrow$ simplified methods e.g. soft interference cancellation.

- Main point: variational inference provides coherence and structure to the design of sub-optimal MUD (turbo or not).
  - Secondary point: Variational EM (expectation maximization) makes parameter estimation natural e.g. noise covariance matrix.

Optimal MUD: JO and IO Detectors

- The optimal detector requires maximizing the posterior distribution of $\mathbf{u}$ or $u_{k,t}$ given $\mathbf{r} = [\mathbf{r}_1; \cdots; \mathbf{r}_T]$, subject to $\mathbf{u} \in \{+1, -1\}^{KT}$ – an integer programming problem with exponential complexity.

- Jointly-optimal (JO) detector

$$P(\mathbf{u}|\mathbf{r}) = \frac{p(\mathbf{r}|\mathbf{u})P(\mathbf{u})}{\sum_{\mathbf{u}} p(\mathbf{r}|\mathbf{u})P(\mathbf{u})} \Longrightarrow \hat{\mathbf{u}} = \arg\max_{\mathbf{u}} P(\mathbf{u}|\mathbf{r}). \tag{1}$$

- Individually-optimal (IO) detector

$$P(u_{k,t}|\mathbf{r}) = \frac{p(\mathbf{r}|u_{k,t})P(u_{k,t})}{\sum_{u_{k,t}} p(\mathbf{r}|u_{k,t})P(u_{k,t})} \Longrightarrow \widehat{u_{k,t}} = \arg\max_{u_{k,t}} P(u_{k,t}|\mathbf{r}), \tag{2}$$

where $k = 1, \cdots, K$ and $p(\mathbf{r}|u_{k,t}) = \sum_{\mathbf{u} \setminus u_{k,t}} p(\mathbf{r}|\mathbf{u})P(\mathbf{u})$.

Iterative Decoding for Near-Optimal Performance

- Optimal decoding not practical, even without coding.
- With error control code, turbo approach has been attempted:
  - FEC decoder is standard SISO (forward-backward), derived from factor graph of code;
  - But factor graph for multi-user channel has variable nodes representing $b_{1,t}, \ldots, b_{K,t}$ all connected to one factor node, hence factor graph cannot simplify problem.
  - Various ad-hoc methods (matched-filter based IC, MMSE-based IC) have been used with some success.
- Without coding, sub-optimal detectors such as SIC, PIC, LMMSE, decorrelating decision feedback exist.
- Can coded and uncoded near-optimal MUD's be derived from the same starting point?

Bayesian View of the Turbo MUD Problem

- Turbo MUD = Generate EXT (extrinsic information) of $\mathbf{b}$ exactly using BCJR in decoder, but update EXT's *approximately* using near-optimal MUD in multi-user section.

- Bayesian inference = use prior distribution on unknown variables $\mathbf{b}$, $P_{in}(\mathbf{b})$, and use appropriate statistical model and observations to update that to $P_{out}(\mathbf{b})$.
  - If $P_{out}(\mathbf{b}) \propto p(\mathbf{y}|\mathbf{b})P_{in}(\mathbf{b})$, then we have *exact inference*;
  - BUT $P_{out}(\mathbf{b})$ may take a different form, easier to compute. Even $P_{in}(\mathbf{b})$ can assume a more convenient structure than its exact expression. Then we have *approximate* inference.

- So the MUD section of a turbo MUD can be designed as an approximate Bayesian inference engine; uncoded MUD can be viewed similarly, but with uniform priors.

- Exact updates too complex because
  - $P(\mathbf{b}|\mathbf{r})$ is a discrete distribution;
  - $P(\mathbf{b}|\mathbf{r})$ does not factorize.

- Variational inference tackles these problems by *postulating* a distribution $Q(\mathbf{b})$ with a convenient form e.g. a Gaussian.

- Then we minimize the Kullback-Leibler (KL) divergence between $Q(\mathbf{b})$ and $P(\mathbf{b}|\mathbf{r})$. This is also known as variational free energy:

$$\mathcal{F}(\lambda_1, \ldots, \lambda_J) = \int_{\mathbf{b}} Q(\mathbf{b}) \log \frac{Q(\mathbf{b})}{p(\mathbf{r}|\mathbf{b})P(\mathbf{b})} d\mathbf{b}$$

  where $\lambda_1, \ldots, \lambda_J$ are the parameters that specify $Q(\mathbf{b})$.

- General procedure for deriving detectors using variational inference:
  - Assume postulated distributions for $p(\mathbf{b})$, $p(\mathbf{r}|\mathbf{b})$ and $Q(\mathbf{b})$;
  - Calculate closed-form expression for $\mathcal{F}(\lambda_1, \ldots, \lambda_J)$;
  - Minimize $\mathcal{F}(\lambda_1, \ldots, \lambda_J)$ (exactly or iteratively).

Case 1: Linear Multiuser Detector

$$\begin{aligned} p(\mathbf{b}) &= \mathcal{N}(\mathbf{0}, \mathbf{I}), \\ p(\mathbf{r}|\mathbf{b}) &= \mathcal{N}(\mathbf{S}\mathbf{b}, \sigma^2\mathbf{I}), \\ Q(\mathbf{b}) &= \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}). \end{aligned} \tag{1}$$

- Now the free energy can be evaluated as

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

- It can be minimized exactly by solving $\partial\mathcal{F}(\boldsymbol{\mu})/\partial\boldsymbol{\mu} = \mathbf{0}$ and $\partial\mathcal{F}(\boldsymbol{\Sigma})/\partial\boldsymbol{\Sigma}^{-1} = \mathbf{0}$.

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

- $\hat{\boldsymbol{\mu}}$ approaches the MMSE, decorrelating and matched filter detector outputs when σ takes on the its true value, 0 and ∞, respectively.

Iterative Minimization: Linear and Clipped SIC

- What if we choose to iteratively minimize $\mathcal{F}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$?

Case 2: Linear/Clipped SIC Detectors
Given the expression for $\mathcal{F}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, in the i-th iteration, for $k = 1$ to K

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

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.

- This is equivalent to a coordinate descent algorithm applied to the minimization of $\mathcal{F}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. And it leads to a proof of guaranteed convergence of linear/clipped SIC with at least linear convergence rate.

- Modify the original formulation by including an unknown parameter θ, which is constant over T realizations of the channel $\mathbf{b}_1, \cdots, \mathbf{b}_T$.
- Let $\lambda_1, \cdots, \lambda_T$ be the parameters of $Q(\mathbf{b}_1), \cdots, Q(\mathbf{b}_T)$.

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

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

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

$$\hat{\theta}^{(i)} = \arg\min_{\hat{\theta}} \sum_{t=1}^{T} \left(\int_{\mathbf{b}_t} Q^{(i)}(\mathbf{b}_t) \log \frac{Q^{(i)}(\mathbf{b}_t)}{p(\mathbf{b}_t, \mathbf{r}_t|\hat{\theta})} d\mathbf{b}_t \right). \tag{2}$$

Relationship to Conventional EM Algorithm

- Conventional E step:

$$U(\theta, \hat{\theta}^{(i)}) = E[\log p(\mathbf{b}|\theta, \mathbf{r})]$$

  where expectation is with respect to the distribution $p(\mathbf{b}|\hat{\theta}^{(i)}, \mathbf{r})$.

- But U may be hard to find. If we approximate $Q(\mathbf{b}) \approx p(\mathbf{b}|\hat{\theta}^{(i)}, \mathbf{r})$ by minimizing variational free energy between the two distributions, we get the E step of the previous page.

- The M step follows from computing U using the postulated distribution, and then maximizing the expression.

- Variational EM gives "hard" parameter estimates and "soft" symbol estimates.

Applying Variational EM to MUD

Case 3: Mean-Field Multiuser Detector

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

- Here ξ_k and γ_k are the prior and posterior probabilities of b_k being 1. This looks more appealing than the Gaussian approximations earlier.

- The approximation made here is the independence of posterior: $Q(\mathbf{b}) = \prod_{k=1}^{K} Q(b_k)$. This is called the *mean field approximation*.

- We assume the noise covariance matrix $\mathbf{\Phi}$ is an unknown parameter, to be estimated together with data iteratively, via variational EM.

Mean Field Multiuser Detector: E Step

- Evaluate the free energy expression considering T realizations.

$$\mathcal{F}(\mathbf{m},\mathbf{\Phi}) = \sum_{t=1}^{T}\left(\sum_{k=1}^{K}\frac{1+m_{k,t}}{2}\log\frac{1+m_{k,t}}{1+\tilde{b}_{k,t}} + \frac{1-m_{k,t}}{2}\log\frac{1-m_{k,t}}{1-\tilde{b}_{k,t}}\right) + \frac{1}{2}\log|\mathbf{\Phi}|$$
$$-\mathbf{r}_t^T\mathbf{\Phi}^{-1}\mathbf{S}\mathbf{m}_t + \frac{1}{2}\mathbf{r}_t^T\mathbf{\Phi}^{-1}\mathbf{r}_t + \frac{1}{2}\mathbf{m}_t^T[\mathbf{S}^T\mathbf{\Phi}^{-1}\mathbf{S} - \mathrm{diag}(\mathbf{S}^T\mathbf{\Phi}^{-1}\mathbf{S})]\mathbf{m}_t \tag{1}$$

  where $\mathbf{m}_t = 2\boldsymbol{\gamma}_t - \mathbf{1}$, $\tilde{\mathbf{b}}_t = 2\boldsymbol{\xi}_t - \mathbf{1}$ and $\mathbf{B} = \mathbf{S}^T\mathbf{\Phi}^{-1}\mathbf{S} - \mathrm{diag}(\mathbf{S}^T\mathbf{\Phi}^{-1}\mathbf{S})$.

- *E Step*: Data Detection – Equating $\partial\mathcal{F}(\mathbf{m},\mathbf{\Phi})/\partial m_{k,t} = 0$.

$$m_{k,t} = \tanh(\boldsymbol{\eta}_k^T\mathbf{r}_t - \boldsymbol{\beta}_k^T\mathbf{m}_t), \tag{2}$$

  where $\boldsymbol{\eta}_k$ and $\boldsymbol{\beta}_k$ are the k-th column vector of $\mathbf{\Phi}^{-1}\mathbf{S}$ and $\mathbf{B}$.

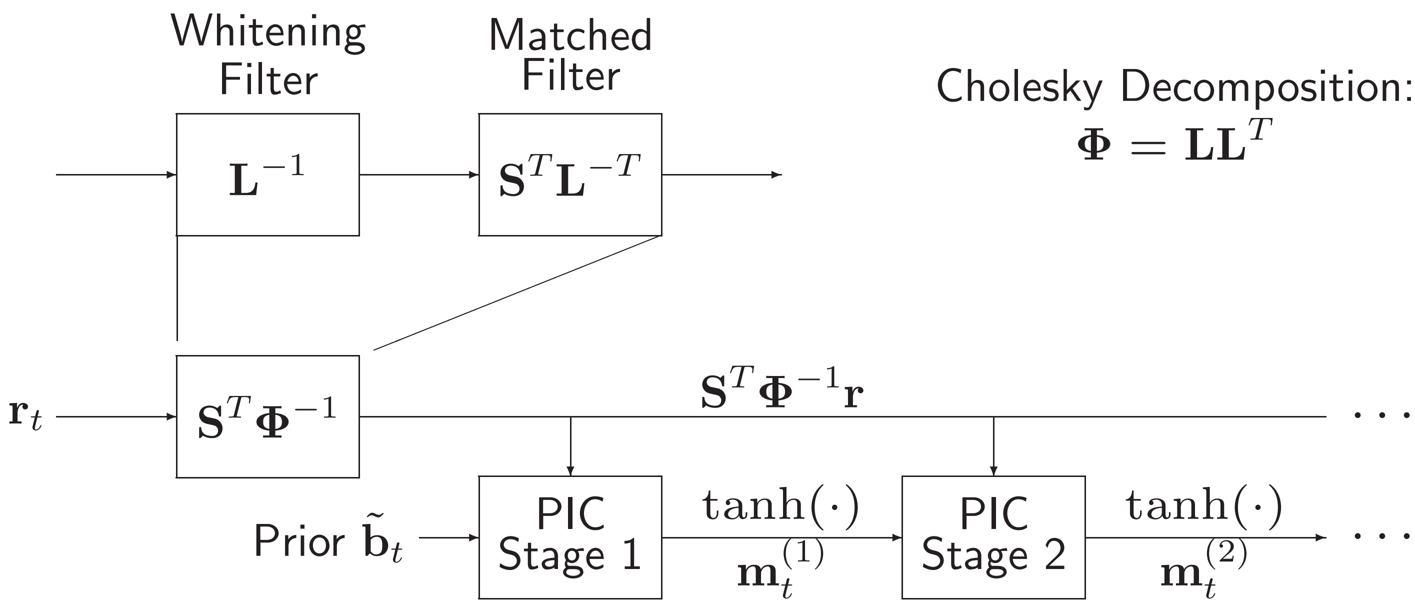
- We can use a coordinate descent approach to minimize $\mathcal{F}(\mathbf{m},\mathbf{\Phi})$ over $\mathbf{m}_t$.

Multi-Stage PIC Interpretation of the E Step

- A multi-stage parallel update would be

$$m_{k,t}^{(p+1)} = \tanh(\boldsymbol{\eta}_k^T \mathbf{r}_t - \boldsymbol{\beta}_k^T \mathbf{m}_t^{(p)})$$

in the p-th iteration, which is a PIC.



Mean Field Multiuser Detector: M Step

- *M Step*: Parameter Estimation – Solve $\partial\mathcal{F}(\mathbf{m},\mathbf{\Phi})/\partial\mathbf{\Phi}^{-1}=\mathbf{0}$.

$$\mathbf{\Phi}=\frac{1}{T}\sum_{t=1}^{T}\{(\mathbf{r}_t-\mathbf{S}\mathbf{m}_t)(\mathbf{r}_t-\mathbf{S}\mathbf{m}_t)^T+\sum_{k=1}^{K}(1-m_{k,t}^2)A_k^2\mathbf{s}_k\mathbf{s}_k^T\}. \quad (1)$$

- The last term can be omitted because $m_k \rightarrow \{\pm 1\}$ as the algorithm converges.

- This mean field multiuser detector is exactly the same as the one proposed in *[Alexander, Grant and Reed, Euro. Trans. Telecommun., Oct. 1998]*, which was derived heuristically.

- Another important turbo multiuser detector. *[Wang and Poor, Trans. Comms., July 1999]*

- It can also be derived from variational free energy minimization by postulating the following distributions:

Case 4: SISO MMSE Multiuser Detector

$$\begin{aligned} p(\mathbf{b}) &= \mathcal{N}(\tilde{\mathbf{b}}, \mathbf{W}), \\ p(\mathbf{r}|\mathbf{b}) &= \mathcal{N}(\mathbf{S}\mathbf{b}, \mathbf{\Phi}), \\ Q(\mathbf{b}) &= \mathcal{N}(\boldsymbol{\mu}, \mathbf{\Sigma}). \end{aligned} \tag{1}$$

- $\mathbf{W} = \text{diag}([1 - \tilde{b}_1^2, \cdots, 1 - \tilde{b}_K^2]^T)$ and $\tilde{\mathbf{b}} = [\tilde{b}_1, \cdots, \tilde{b}_K]^T$ are the soft bit estimates from the MAP decoder.

SISO MMSE Multiuser Detector (Cont.)

$$\mathcal{F}(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\Phi}) = \sum_{t=1}^{T} \frac{1}{2}[(\mathbf{r}_t - \mathbf{S}\boldsymbol{\mu}_t)^T \boldsymbol{\Phi}^{-1}(\mathbf{r}_t - \mathbf{S}\boldsymbol{\mu}_t) + \log|\boldsymbol{\Phi}| + \mathrm{tr}(\mathbf{S}^T\boldsymbol{\Phi}^{-1}\mathbf{S}\boldsymbol{\Sigma}_t)]. \quad (1)$$

- *E Step*: Solve $\partial\mathcal{F}(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\Phi})/\partial\boldsymbol{\mu}_t = \mathbf{0}$.

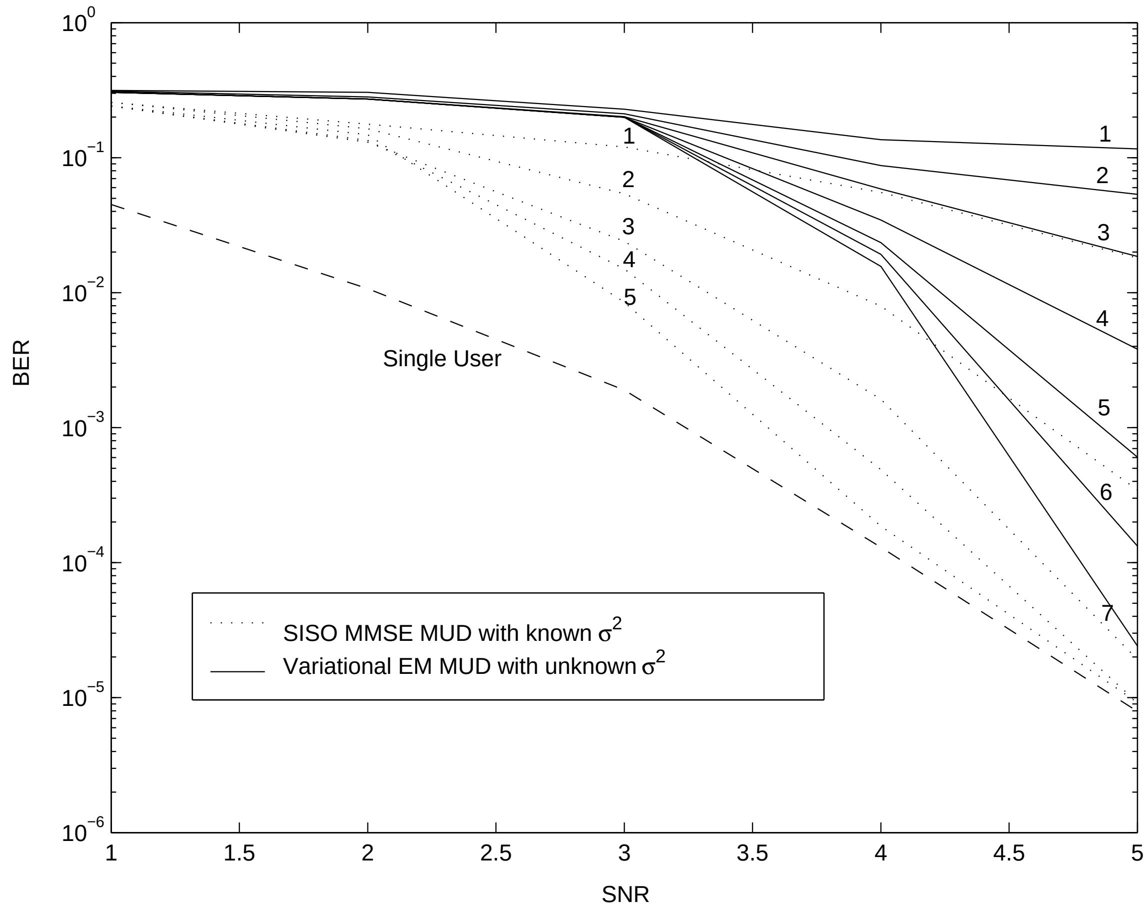
$$\mu_{k,t} = \mathbf{e}_k^T(\mathbf{S}^T\boldsymbol{\Phi}^{-1}\mathbf{S} + \mathbf{W}_{k,t}^{-1})^{-1}\mathbf{S}^T\boldsymbol{\Phi}^{-1}(\mathbf{r}_t - \mathbf{S}\tilde{\mathbf{b}}_{k,t}), \quad (2)$$

  - $\mathbf{e}_k = [0, \cdots, 0, 1, 0, \cdots, 0]^T$, all zero vector with kth element being 1;
  - $\tilde{\mathbf{b}}_{k,t} = [\tilde{b}_{1,t}, \cdots, \tilde{b}_{k-1,t}, 0, \tilde{b}_{k+1,t}, \cdots, \tilde{b}_{K,t}]^T$;
  - $\mathbf{W}_{k,t} = \mathrm{diag}([1-\tilde{b}_{1,t}^2, \cdots, 1-\tilde{b}_{k-1,t}^2, 1, 1-\tilde{b}_{k+1,t}^2, \cdots, 1-\tilde{b}_{K,t}^2]^T)$.

- *M Step*: Solve $\partial\mathcal{F}(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\Phi})/\partial\boldsymbol{\Phi}^{-1} = \mathbf{0}$.

$$\boldsymbol{\Phi} = \frac{1}{T}\sum_{t=1}^{T}(\mathbf{r}_t - \mathbf{S}\boldsymbol{\mu}_t)(\mathbf{r}_t - \mathbf{S}\boldsymbol{\mu}_t)^T + \mathbf{S}\boldsymbol{\Sigma}_t\mathbf{S}^T \quad (3)$$

- The original SISO MMSE detector assumes known noise variance. Let's see how well variational EM works without knowing σ^2.



Summary

- Optimal MUD and its difficulty.
- Concept of variational inference.
- Deriving linear MUD and linear/clipped SIC using variational inference.
- Deriving turbo multiuser detectors using variational inference.

Thank You!