Deterministic and Monte Carlo Approaches for Joint Iterative Data Detection and Channel Estimation

Ansgar Scherb, Volker Kühn and Karl-Dirk Kammeyer
Department of Communications Engineering
University of Bremen, Otto-Hahn-Allee, D-28359 Bremen, Germany
Email:{scherb, kuehn, kammeyer}@ant.uni-bremen.de

***Abstract*— This paper deals with joint data detection and channel estimation for single input single output systems in presence of inter symbol interference. Therefore, deterministic methods, the Gibbs-sampler and combinations between deterministic and Monte Carlo approaches are compared. The examined methods belong to the class of block by block iterative algorithms alternating between channel estimation and data detection. It will be shown that the deterministic method might get trapped in a local maximum of the likelihood function, whereas the Monte Carlo methods theoretically almost converge to a global maximum. Based on simulation results it will be shown that a performance gain can be achieved at the expense of slower convergence speed or an increased computational effort.**

I. Introduction

The requirements of spectral efficiency for wireless communication systems are still growing. In order to get reliable transmission, the receiver of a wireless communication link requires channel state information. Usually, a pilot data sequence embedded in the data block enables the receiver to estimate the channel. Since this pilot sequence bears no information, the waste of bandwidth caused by pilot symbols should be kept as low as possible. It has been shown that the quality of channel estimates can be improved dramatically by feeding back the decided data as pseudo reference signal for the channel estimation. On the other hand the data detection becomes more reliable using the improved channel estimates. Thus, iterative equalizer structures alternating between data detection and channel estimation promise good performance gains, since the number of required pilot symbols can be decreased and thus more bandwidth can be utilized to transmit information.

A lot of work has been done in this field so far, e.g. Feder and Catipovic have dealt with iterative joint channel and data estimation for finite impulse response (FIR) channels [1] and Kaleh and Valet have established the expectation-maximization-algorithm for this task [2]. Beside block by block processing, sequential methods have become more and more popular. In [3], [4] trellis based implementations for iterative joint maximum likelihood (JML) approximation were presented. However, in the face of bad initialization most deterministic methods suffer from insufficient convergence properties.

Recently, the old idea of Monte Carlo sampling has been revitalized, e.g. in [5]. This class of approaches has its origin in the early fifties by the well known metropolis algorithm [6]. A comprehensive survey of Monte Carlo (MC) methods is given in [7]. One of the most popular block by block algorithms is the Gibbs sampler [8], which was first introduced to unknown FIR channel equalization in [9]. As shown in [7] the Gibbs sampler provides almost global convergence. Also sequential Monte Carlo sampling was applied to wireless communication problems, e.g [10], [11].

The contribution of this paper is to combine deterministic and MC methods and to compare these iterative structures with respect to the quality of the initial guess. We will focus on block-wise signal processing in order to illustrate the general problems occurring in the context of iterative joint maximum likelihood approximation. Throughout this paper, we will assume that an initial channel estimate of low quality is available at the receiver without specifying how to obtain it, e.g. pilot-based or completely blind.

In Section II we introduce the system model. On basis of the joint maximum likelihood criterion, a suboptimal deterministic iterative approach is derived and the EM-algorithm is briefly explained in Section III. In Section IV the Gibbs-sampler is explained and some combinations of deterministic and Markov chain Monte Carlo (MCMC) algorithms are defined. A comparison on basis of numerical results is presented in Section V and the paper is concluded in Section VI.

II. System Model

We consider a block transmission of K M-ary PSK symbols $\mathbf{s} = [s(1), \cdots, s(K)]^T$ over a frequency selective channel of order L. We assume that the interval between two consecutive data blocks is filled by a sufficiently large number of zero symbols, such that no inter symbol interference between two consecutive blocks occurs. Collecting $K+L$ samples in the vector $\mathbf{r}$ at symbol rate, the channel output is given by

$$\mathbf{r} = \mathbf{S}\mathbf{h} + \mathbf{n}, \qquad (1)$$

where $\mathbf{S}$ is the $(L+K \times L+1)$ convolution matrix of $\mathbf{s}$ defined by

$$[\mathbf{S}]_{\nu,\mu} = \begin{cases} s(\nu-\mu+1) & : \quad 0 \le \nu-\mu < K \\ 0 & : \quad \text{else} \end{cases} . \qquad (2)$$

The vector $\mathbf{h} = [h_0, \cdots, h_L]^T$ contains the channel gains $h_l \in \mathcal{C}^1$ including transmit and receive filter and $\mathbf{n} =$

[1] $\mathcal{C}$: set of complex numbers

$[n(1), \cdots, n(L+K)]^T$ is white gaussian noise with power σ^2. Due to the finite symbol alphabet, each transmit vector $\mathbf{s}$ is an element of the set $\mathcal{A} = \{\mathbf{s}_1, \cdots, \mathbf{s}_{M^K}\}$ of all possible symbol sequences.

III. Deterministic Joint Maximum Likelihood Approximation

In view of blind data detection and channel estimation the joint maximum likelihood (JML) criterion may deliver jointly suitable estimates of channel and data [1]. The JML solution is given by

$$\begin{aligned}(\hat{\mathbf{s}}, \hat{\mathbf{h}}) &= \arg \max_{\mathbf{s}\in\mathcal{A}, \mathbf{h}\in\mathcal{C}^{L+1}} p(\mathbf{r}|\mathbf{s}, \mathbf{h}) \\ &= \arg \min_{\mathbf{s}\in\mathcal{A}, \mathbf{h}\in\mathcal{C}^{L+1}} \|\mathbf{r} - \mathbf{S}\mathbf{h}\|^2, \end{aligned} \tag{3}$$

where

$$p(\mathbf{r}|\mathbf{s}, \mathbf{h}) = \frac{1}{(\sigma\pi)^{K+L}} \exp\left(-\|\mathbf{r} - \mathbf{S}\mathbf{h}\|^2/\sigma^2\right) \tag{4}$$

is the probability of receiving $\mathbf{r}$ under the condition, $\mathbf{s}$ was transmitted over the channel $\mathbf{h}$. Please note that (3) has no unique solution due to an unknown complex rotation factor of $\mathbf{h}$ which yields a phase ambiguity. Therefore, throughout this paper we assumed differential encoding. However, accomplishing the JML estimation is not practicable in real applications due to the high computational effort of the exhaustive search over M^K possibilities of $\mathbf{s}$.

A. Iterative Joint Maximum Likelihood Approximation (IJML)

Since the maximization of the probability density function (4) can be simplified keeping either $\mathbf{s}$ or $\mathbf{h}$ fixed, a natural way to approximate JML solution is to estimate iteratively the channel and the data. Let i be an iteration counter. Then the ML channel estimation under the assumption that $\hat{\mathbf{s}}^{(i)}$ was the transmitted data sequence can be performed by maximizing the conditional likelihood

$$\mathrm{L}(\mathbf{h}|\hat{\mathbf{s}}^{(i)}) \propto \exp(-\|\mathbf{r} - \hat{\mathbf{S}}_{(i)}\mathbf{h}\|^2/\sigma^2). \tag{5}$$

and calculating

$$\begin{aligned}\hat{\mathbf{h}}^{(i)} &= \arg \max_{\mathbf{h}\in\mathcal{C}^{L+1}} \mathrm{L}(\mathbf{h}|\hat{\mathbf{s}}^{(i)}) \\ &= \arg \min_{\mathbf{h}\in\mathcal{C}^{L+1}} \|\mathbf{r} - \hat{\mathbf{S}}_{(i)}\mathbf{h}\|^2. \end{aligned} \tag{6}$$

The ML data detector maximizes the conditional likelihood

$$\mathrm{L}(\mathbf{s}|\hat{\mathbf{h}}^{(i)}) \propto \exp(-\|\mathbf{r} - \mathbf{S}\hat{\mathbf{h}}^{(i)}\|^2/\sigma^2). \tag{7}$$

by calculating

$$\begin{aligned}\hat{\mathbf{s}}^{(i+1)} &= \arg \max_{\mathbf{s}\in\mathcal{A}} \mathrm{L}(\mathbf{s}|\hat{\mathbf{h}}^{(i)}) \\ &= \arg \min_{\mathbf{s}\in\mathcal{A}} \|\mathbf{r} - \mathbf{S}\hat{\mathbf{h}}^{(i)}\|^2. \end{aligned} \tag{8}$$

A solution for (8) can be obtained by the well known Viterbi algorithm [12], whereas

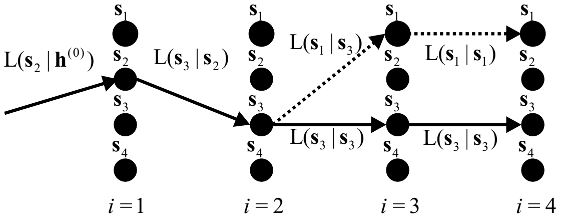
$$\hat{\mathbf{h}}^{(i)} = \left(\hat{\mathbf{S}}_{(i)}^H \hat{\mathbf{S}}_{(i)}\right)^{-1} \hat{\mathbf{S}}_{(i)}^H \mathbf{r} \tag{9}$$

is the maximum likelihood channel estimator with covariance

$$\begin{aligned}\mathbf{C}_{hh}^{(i)} &= E\{(\mathbf{h} - \hat{\mathbf{h}}^{(i)})(\mathbf{h} - \hat{\mathbf{h}}^{(i)})^H|\hat{\mathbf{s}}^{(i)}\} \\ &= \sigma^2 \left(\hat{\mathbf{S}}_{(i)}^H \hat{\mathbf{S}}_{(i)}\right)^{-1}. \end{aligned} \tag{10}$$

B. Discussion of the IJML

The algorithm can not be called totally blind, since in the first step an initial channel estimate is required in order to start the iterative procedure. The suggested procedure is suboptimal in the sense that only a small part of the set $\mathcal{A}$ will be covered. As shown in the example of Fig. 1 the iterative procedure can be interpreted as Markov chain, where the states represent the instantaneous data estimates at iteration step i according to $\mathcal{A} = \{\mathbf{s}_1, \mathbf{s}_2, \mathbf{s}_3, \mathbf{s}_4\}$ and the state transitions are given by



$\mathbf{s}_1$
$\mathbf{s}_2$
$\mathbf{s}_3$
$\mathbf{s}_4$
$\mathrm{L}(\mathbf{s}_2\,|\,\mathbf{h}^{(0)})$
$\mathrm{L}(\mathbf{s}_3\,|\,\mathbf{s}_2)$
$\mathrm{L}(\mathbf{s}_1\,|\,\mathbf{s}_3)$
$\mathrm{L}(\mathbf{s}_1\,|\,\mathbf{s}_1)$
$\mathrm{L}(\mathbf{s}_3\,|\,\mathbf{s}_3)$
$\mathrm{L}(\mathbf{s}_3\,|\,\mathbf{s}_3)$
$i=1$
$i=2$
$i=3$
$i=4$

Fig. 1. Iterative Joint Maximum Likelihood Approximation as Markov Chain

$$\mathrm{L}(\mathbf{s}|\hat{\mathbf{s}}^{(i)}) \propto e^{\left(-\|\mathbf{r} - \mathbf{S}\left(\hat{\mathbf{S}}_{(i)}\hat{\mathbf{S}}_{(i)}^H\right)^{-1}\hat{\mathbf{S}}_{(i)}^H\mathbf{r}\|^2/\sigma^2\right)}. \tag{11}$$

Maximizing $\mathrm{L}(\mathbf{s}|\hat{\mathbf{s}}^{(i)})$ is identical to successively calculating (6) and (8). The algorithm converges, when the likelihood becomes maximum for $\mathrm{L}(\mathbf{s}_k|\hat{\mathbf{s}}^{(i)} = \mathbf{s}_k)$. The obtained maximum does not need to coincide inevitably with the global maximum of (3). The performance of the algorithm depends on the quality of the initial channel estimate. Starting with the initial channel estimate $\hat{\mathbf{h}}^{(0)}$ the solid path in Fig. 1 represents the maximum of (11) at each iteration step. In order to illustrate the problem of trapping into a local minimum assume that the metric $\mathrm{L}(\mathbf{s}_1|\mathbf{s}_1)$ denoted by the dashed line between iteration 3 and 4 is larger than $\mathrm{L}(\mathbf{s}_3|\mathbf{s}_3)$ but also $\mathrm{L}(\mathbf{s}_1|\mathbf{s}_3)$ is lower than $\mathrm{L}(\mathbf{s}_3|\mathbf{s}_3)$ as depicted by the dashed and solid line between iteration 3 and 4. Even if $\mathrm{L}(\mathbf{s}_1|\mathbf{s}_3)$ is very close to $\mathrm{L}(\mathbf{s}_3|\mathbf{s}_3)$ the algorithm does not converge to the global maximum.

A hint on the instantaneous channel estimation quality of the i-th iteration step can be obtained by the covariance matrix given in (10). However, the explained algorithm does not take this quality into account.

C. EM Algorithm

In this section we give a brief explanation of the EM-algorithm which is also an iterative procedure and is a standard tool for the missing data problem in statistical mathematics. It was firstly applied to the joint data and channel estimation of FIR channels by Kaleh and Valet [2]. One of the most useful features of the EM-algorithm is that the desired likelihood is increased at each iteration step. Thus, it can be guaranteed that at least a local maximum is reached.

The target of the EM-algorithm is to maximize the likelihood $p(\mathbf{r}|\mathbf{h})$ with respect to $\mathbf{h}$ by

$$\hat{\mathbf{h}} = \arg\max_{\mathbf{h}} p(\mathbf{r}|\mathbf{h}) \tag{12}$$

Since it is very difficult to solve this expression, the hidden variable $\mathbf{s}$ is introduced by

$$p(\mathbf{r}|\mathbf{h}) = \sum_{\mathbf{s}} p(\mathbf{r},\mathbf{s}|\mathbf{h}). \tag{13}$$

The EM-approach consists of an expectation step with respect to the hidden variable $\mathbf{s}$ and a maximization step with respect of $\mathbf{h}$, which are given by

$$\text{E-step:} \qquad \mathrm{U}(\mathbf{h}|\hat{\mathbf{h}}^{(i)}) = \mathrm{E}\{\log p(\mathbf{r},\mathbf{s}|\mathbf{h})|\hat{\mathbf{h}}^{(i)}\} \tag{14}$$

and

$$\text{M-step:} \qquad \hat{\mathbf{h}}^{(i+1)} = \arg\max_{\mathbf{h}} \mathrm{U}(\mathbf{h}|\hat{\mathbf{h}}^{(i)}). \tag{15}$$

Since without the a-priori information $p(\mathbf{s})$ the density $p(\mathbf{r},\mathbf{s}|\mathbf{h})$ is proportional to $p(\mathbf{r}|\mathbf{h},\mathbf{s})$, the E-step can be expressed as

$$\begin{aligned} \mathrm{U}(\mathbf{h}|\hat{\mathbf{h}}^{(i)}) &= \mathrm{E}\{\log p(\mathbf{h}|\mathbf{s},\mathbf{r})|\hat{\mathbf{h}}^{(i)}\} \\ &= \mathrm{E}\{\|\mathbf{r}-\mathbf{S}\hat{\mathbf{h}}^{(i)}\|^2/\sigma^2\} \\ &= \|\mathbf{r}\|^2 - 2\Re(\mathbf{r}^H\hat{\mathbf{S}}\hat{\mathbf{h}}^{(i)}) \\ &\quad +(\hat{\mathbf{h}}^{(i)})^H\mathbf{C}_{ss}^{(i)}\hat{\mathbf{h}}^{(i)}, \end{aligned} \tag{16}$$

where $\hat{\mathbf{S}_{(i)}} = \mathrm{E}\{\mathbf{S}|\hat{\mathbf{h}}^{(i)}\}$ as well as $\mathbf{C}_{ss}^{(i)} = \mathrm{E}\{\mathbf{S}^H\mathbf{S}|\hat{\mathbf{h}}^{(i)}\}$ can be obtained by an forward backward algorithm similarly to the BCJR-algorithm. Note that the estimate $\hat{\mathbf{S}}_{(i)}$ consists of soft values. For a detailed derivation, see [2]. The solution of the M-step (15) is given by

$$\hat{\mathbf{h}}^{(i+1)} = \left(\mathbf{C}_{ss}^{(i)}\right)^{-1}\hat{\mathbf{S}}_{(i)}\mathbf{r}. \tag{17}$$

In a sense the estimation of $\mathbf{s}$ is a byproduct of the EM-algorithm. Therefore, this approach seems at the first view a little bit astonishing, since usually the receiver is more interested to obtain an estimate of the data $\mathbf{s}$ than to know the channel impulse response $\mathbf{h}$. Thus, it could be more natural to treat $\mathbf{h}$ as the hidden variable. However, experience has shown that this version outperforms the complementary EM-approach. In contrast to IJML this approach take into account the variance of the current data estimates in terms of $\mathbf{C}_{ss}^{(i)}$. Similarly as illustrated in section III-B for IJML, the EM-algorithm may trap into a local maximum.

IV. Markov Chain Monte Carlo (MCMC)

A way to achieve global convergence may be obtained by the so called Markov chain Monte Carlo approaches. In the following we will describe the Gibbs-sampler [8] as one of the most popular MCMC schemes, which is the basis for the MC related methods examined in this paper. As well as before the Gibbs-sampler is an iterative procedure alternating between data detection and channel estimation. But in contrast to deterministic algorithms as IJML neither the channel estimation nor the data detection is deterministic in the sense that the Gibbs-sampler does not deliver an unique output for a certain input. The task of the Gibbs-sampler is to generate a sequence of random numbers according to an appropriate pdf. On the basis of this random numbers the desired estimator output can be approximated. The technical realization of random number generation is out of the range of this paper (e.g. pseudo random numbers by m-sequences, etc. [7]). We will focus on the derivation of the pdf's which corresponds to the required random numbers. In order to distinguish between deterministic estimates and random samples, throughout the paper all random samples are labelled by a bar and all estimates are labelled by a hat.

In our case the objective is to approximate the conditional expectation

$$\hat{\mathbf{s}} = E\{\mathbf{s}|\mathbf{r}\} = \sum_{\mathbf{s}\in\mathcal{A}} \mathbf{s}\,p(\mathbf{s}|\mathbf{r}). \tag{18}$$

Please note, that in contrast to the output of the IJML $\hat{\mathbf{s}} \notin \mathcal{A}$ has soft values. Due to the large number of members of the set $\mathcal{A}$ the analytical calculation of (18) is not tractable. The key idea is to draw I random samples $\bar{\mathbf{s}}^{(i)}$ from $p(\mathbf{s}|\mathbf{r})$ in order to approximate the expectation by

$$\hat{\mathbf{s}} \approx \frac{1}{I}\sum_{i=1}^{I}\bar{\mathbf{s}}^{(i)}. \tag{19}$$

For sufficiently smooth probability density functions (pdf) a quite good approximation of $p(\mathbf{s}|\mathbf{r})$ can be obtained by a small number of random samples. Since it is very difficult to obtain $p(\mathbf{s}|\mathbf{r})$ analytically, it would be desirable to perform the sampling procedure without having to calculate the exact density. To this end, $p(\mathbf{s}|\mathbf{r})$ can be expressed as a function dependent on the conditional pdf $p(\mathbf{h}|\mathbf{r})$ by making use of the marginalization

$$p(\mathbf{s}|\mathbf{r}) = \int p(\mathbf{s}|\mathbf{h},\mathbf{r})p(\mathbf{h}|\mathbf{r})d\mathbf{h}. \tag{20}$$

Due to the Bayesian law

$$p(\mathbf{s}|\mathbf{h},\mathbf{r}) = \frac{p(\mathbf{r}|\mathbf{s},\mathbf{h})p(\mathbf{s})}{p(\mathbf{r}|\mathbf{h})} \tag{21}$$

the relation

$$p(\mathbf{s}|\mathbf{h},\mathbf{r}) \propto p(\mathbf{r}|\mathbf{s},\mathbf{h}) \tag{22}$$

holds, if $\mathbf{h}$ is given and no a-priori information $p(\mathbf{s})$ exists. As shown in (4) the density $p(\mathbf{r}|\mathbf{s},\mathbf{h})$ can be easily determined. Thus, assuming that several random samples $\bar{\mathbf{h}}^{(i)}$ for $i = 1,\cdots,I$ according $p(\mathbf{h}|\mathbf{r})$ are available $p(\mathbf{s}|\mathbf{r})$ can be approximated in a similar way as illustrated in (18) and (19).

In order to obtain the desired random variables $\bar{\mathbf{h}}$ the density $p(\mathbf{h}|\mathbf{r})$ is required. Similarly as for $p(\mathbf{s}|\mathbf{r})$, this pdf can be expressed as a function depending on the pdf $p(\mathbf{s}|\mathbf{r})$ by

$$p(\mathbf{h}|\mathbf{r}) = \sum_{\mathbf{s}\in\mathcal{A}} p(\mathbf{h}|\mathbf{s},\mathbf{r})p(\mathbf{s}|\mathbf{r}). \tag{23}$$

Again, due to

$$p(\mathbf{h}|\mathbf{s},\mathbf{r}) = \frac{p(\mathbf{r}|\mathbf{s},\mathbf{h})p(\mathbf{h})}{p(\mathbf{r}|\mathbf{s})} \tag{24}$$

the relation

$$p(\mathbf{h}|\mathbf{s},\mathbf{r}) \propto p(\mathbf{r}|\mathbf{s},\mathbf{h}) \tag{25}$$

holds, if $\mathbf{s}$ is given and no a-priori information $p(\mathbf{s})$ exists. Therefore, $p(\mathbf{h}|\mathbf{r})$ can be approximated by a sufficient large number of random sample $\bar{\mathbf{s}}^{(i)}$ for $i = 1, \cdots, I$ according to $p(\mathbf{s}|\mathbf{r})$.

Obviously, $p(\mathbf{s}|\mathbf{r})$ and $p(\mathbf{h}|\mathbf{r})$ can be approximated by a set of random variables $\bar{\mathbf{h}}^{(i)}$ and $\bar{\mathbf{s}}^{(i)}$ and on the other hand these densities are required to generate the desired random variable. Unfortunately, neither the receiver knows the desired densities nor has the desired random variables. Therefore, the Gibbs-sampler approximates the desired pdf's $p(\mathbf{s}|\mathbf{r})$ and $p(\mathbf{h}|\mathbf{r})$ by $p(\mathbf{r}|\bar{\mathbf{h}}^{(i)},\mathbf{s})$ and $p(\mathbf{r}|\bar{\mathbf{s}}^{(i)},\mathbf{s})$. It generates iteratively random samples starting by an initial guess $\hat{\mathbf{h}}^{(0)}$. Therefore, the both steps

$$\bar{\mathbf{s}}^{(i+1)} \sim p(\mathbf{r}|\bar{\mathbf{h}}^{(i)},\mathbf{s}) \tag{26}$$

and

$$\bar{\mathbf{h}}^{(i)} \sim p(\mathbf{r}|\bar{\mathbf{s}}^{(i)},\mathbf{s}) \tag{27}$$

are repeated alternating, where a detailed description of (26) and (27) is given in section IV-D and IV-C, respectively.

A. Convergence properties

The transition probability from n-th to the m-th member of the set $\mathcal{A}$ is given by

$$a_{n,m} = p(\mathbf{s}_n|\mathbf{s}_m,\mathbf{r}) = \int p(\mathbf{s}_n|\mathbf{h},\mathbf{s})p(\mathbf{h}|\mathbf{s}_m,\mathbf{s})d\mathbf{h}. \tag{28}$$

Due to the fact, that (for given $\mathbf{r}$) any $\bar{\mathbf{s}}^{(i+1)}$ only depends on $\bar{\mathbf{s}}^{(i)}$ from the previous iteration step, the sampling procedure can be modelled as stationary Markov chain, where the term "stationary" means that the transition probabilities from state n to state m does not change during the iterations (Fig. 2). Defining the $(M^L \times M^L)$ transition matrix $[\mathbf{A}]_{m,n} = a_{m,n}$, the joint transition probabilities of i subsequent iteration steps are given by $\mathbf{A}^i$. Let

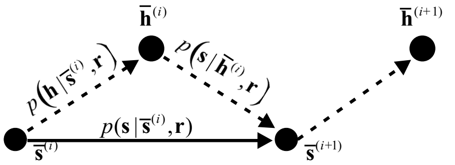

$\bar{\mathbf{h}}^{(i)}$
$\bar{\mathbf{h}}^{(i+1)}$
$p(\mathbf{h}|\bar{\mathbf{s}}^{(i)},\mathbf{r})$
$p(\mathbf{s}/\bar{\mathbf{h}}^{(i)},\mathbf{r})$
$p(\mathbf{s}|\bar{\mathbf{s}}^{(i)},\mathbf{r})$
$\bar{\mathbf{s}}^{(i)}$
$\bar{\mathbf{s}}^{(i+1)}$

Fig. 2. Gibbs sampler

$$\mathbf{A} = \mathbf{V}\mathbf{\Lambda}\mathbf{V}^{-1} \tag{29}$$

be the eigen decomposition of $\mathbf{A}$, where $\mathbf{\Lambda} = \mathrm{diag}(\lambda_1, \cdots, \lambda_{M^L})$ are the eigenvalues of $\mathbf{A}$ with $|\lambda_1| > \cdots > |\lambda_{m^L}|$ and $\mathbf{V} = [\mathbf{v}_1, \cdots, \mathbf{v}_{M^L}]$ the corresponding eigenvectors. The maximum eigenvalue is always $\lambda_1 = 1$ [7]. Then the eigen decomposition of $\mathbf{A}^i$ is given by

$$\mathbf{A}^i = \mathbf{V}\mathbf{\Lambda}^i\mathbf{V}^{-1}, \tag{30}$$

where $\mathbf{\Lambda}^i = \mathrm{diag}(\lambda_1^i, \cdots, \lambda_{M^L}^i)$. Since except of λ_1 all eigenvalues are smaller than one, for a large i the values $\lambda_2^i, \cdots, \lambda_{m^L}^i$ tend to zero. Therefore, for a sufficiently large number of iteration steps i the current state of the Markov chain becomes independent from the initial value. In this case the probability of meeting a certain state is given by the density $p(\mathbf{s}|\mathbf{r})$, which is equivalent to the eigenvector $\mathbf{v}_1$ corresponding to the maximal eigenvalue λ_1. In other words for sufficiently large i we will obtain random samples of $\bar{\mathbf{s}}^{(i)}$ and $\bar{\mathbf{h}}^{(i)}$ according to the conditional densities (20) and (23). As it can be seen from (30), the convergence speed of the Gibbs-sampler depends on the ratio of the maximum to the subsequent eigenvalue λ_1/λ_2 and is geometric.

Since performing (26) and (27) requires a current estimate of the noise power $\hat{\sigma}^2_{(i)}$, in the next section a noise power estimator is presented.

B. Noise power estimation

Given $\bar{\mathbf{s}}^{(i)}$ and $\bar{\mathbf{h}}^{(i)}$ the maximum likelihood estimator for $\hat{\sigma}^2$ is given by

$$\begin{aligned}\hat{\sigma}^2_{(i)} &= \arg\max_{\sigma^2} p(\mathbf{r}|\sigma^2,\bar{\mathbf{s}}^{(i)},\bar{\mathbf{h}}^{(i)}) \\ &= \arg\min_{\sigma^2} \Big[(K+L)\log(\pi\sigma^2) \\ &\qquad +1/\sigma^2\|\mathbf{r}-\bar{\mathbf{S}}_{(i)}\bar{\mathbf{h}}^{(i)}\|^2\Big].\end{aligned} \tag{31}$$

Replacing $\bar{\mathbf{h}}^{(i)}$ in (31) by the r.h.s of (9), the solution can be calculated by differentiating the resulting expression with respect to σ^2:

$$\hat{\sigma}^2_{(i)} = \frac{1}{K+L}\mathbf{r}^H\mathbf{Q}^{(i)}\mathbf{r} \tag{32}$$

with

$$\mathbf{Q}^{(i)} = \left(\mathbf{I} - \bar{\mathbf{S}}_{(i)}\left(\bar{\mathbf{S}}^H_{(i)}\bar{\mathbf{s}}_{(i)}\right)^{-1}\bar{\mathbf{S}}^H_{(i)}\right). \tag{33}$$

Obviously, $\mathbf{Q}^{(i)}$ represents an orthogonal projection matrix with respect to the hypothetic signal space $\mathcal{S}_{(i)}$ spanned by the column vectors of $\bar{\mathbf{S}}_{(i)}$. Hence, $\hat{\sigma}^2_{(i)}$ corresponds to the squared distance between the observation $\mathbf{r}$ and the point inside $\mathcal{S}_{(i)}$, which is nearest to $\mathbf{r}$. However, the part of the noise inside the signal space is neglected by the proposed estimator. Therefore, even if $\bar{\mathbf{s}}^{(i)}$ corresponds to the true data the noise power will be underestimated, resulting in a biased estimator. In order to compensate this bias, the result can be multiplied with a correction factor $\phi = (K+L)/K$, which represents the ratio between the dimensions of the observation space of $\mathbf{r}$ and the noise space complementary to the signal space $\mathcal{N}_{(i)} \perp \mathcal{S}_{(i)}$.

Due to the fact that in most cases $\hat{\sigma}^2_{(i)}$ corresponding to a good guess $\bar{\mathbf{s}}^{(i)}$ is lower than for a bad guess, the current noise power estimate can be interpreted as a measure for the quality of the current guess of $\bar{\mathbf{s}}^{(i)}$. As the variance of the conditional pdf's corresponding to data and channel are closely related to $\hat{\sigma}^2_{(i)}$, the estimated noise power plays an important role in Monte Carlo sampling, whereas the ML estimators of data and channel used in IJML do not need to have this knowledge. Therefore, this value can be understood as a coefficient, which weights the strength of the randomness in MCMC.

C. Random sampling of the channel impulse response

The conditional pdf $p(\mathbf{h}|\bar{\mathbf{s}}^{(i)}, \mathbf{r}) \propto p(\mathbf{r}|\bar{\mathbf{s}}^{(i)}, \mathbf{h})$ is gaussian and completely determined by its mean $\hat{\mathbf{h}}$ as given in (9) and its covariance matrix $\mathbf{C}_{hh}^{(i)}$ as given in (10). Thus, a sample of the channel impulse response according to (27) can be obtained by

$$\bar{\mathbf{h}}^{(i)} = \hat{\mathbf{h}}^{(i)} + \left(\mathbf{C}_{hh}^{(i)}\right)^{1/2} \boldsymbol{\eta}, \tag{34}$$

where $\boldsymbol{\eta}$ is the output of a white gaussian noise generator with $E\{\boldsymbol{\eta}\boldsymbol{\eta}^H\} = \mathbf{I}$. Since $\mathbf{C}_{hh}^{(i)}$ is always positive semidefinite, its root can be obtained by the Cholesky factorization. Please note that in order to calculate $\mathbf{C}_{hh}^{(i)}$ an estimate of the noise power is needed.

D. Random sampling of the data sequence

We present now forward backward sampling method, which was originally derived from [13].

A finite state machine representation of the communication link is given by

$$r(k) = s(k)h_0 + \mathbf{z}_k^T \tilde{\mathbf{h}} + n(k), \tag{35}$$

where $N = M^L$ is the number of states according to all possible inputs $\mathbf{z}_k = [s(k-L), \cdots, s(k-1)]^T \in \mathcal{Z}$ and $\tilde{\mathbf{h}} = [h_1, \cdots, h_L]$ is the reduced channel vector. Please note that given $\mathbf{r}$ and $\mathbf{h}$ any state $\mathbf{z}_k$ only depends on the neighboring states $\mathbf{z}_{k-1}$ and $\mathbf{z}_{k+1}$. Therefore, the conditional probability distribution (20) can be factorized by the chain rule as

$$\begin{aligned} p(\mathbf{s}|\mathbf{h}, \mathbf{r}) &= p(\mathbf{z}_{K+L}|\mathbf{h}, \mathbf{r}) p(\mathbf{z}_{K+L-1}|\mathbf{z}_{K+L}, \mathbf{h}, \mathbf{r}) \\ &\quad \cdots p(\mathbf{z}_0|\mathbf{z}_1, \mathbf{h}, \mathbf{r}) \\ &= p(\mathbf{z}_{K+L}|\mathbf{h}, \mathbf{r}) \prod_{k=0}^{K+L-1} p(\mathbf{z}_k|\mathbf{z}_{k+1}, \mathbf{h}, \mathbf{r}). \end{aligned} \tag{36}$$

In order to calculate the last chain link $p(\mathbf{z}_{K+L}|\mathbf{h}, \mathbf{r})$, we define

$$\alpha_k(\zeta) = p(\mathbf{z}_k = \zeta, \mathbf{r}_{\leq k}|\mathbf{h}) \tag{37}$$

and

$$\gamma_k(\zeta, \zeta') = p(\mathbf{z}_k = \zeta, r(k)|\mathbf{z}_{k-1} = \zeta', \mathbf{h}), \tag{38}$$

where $\mathbf{r}_{\leq k} = [r(1), \cdots, r(k)]$ is the reduced observation vector and $\zeta, \zeta' \in \mathcal{Z}$. Starting from $\alpha_0(\zeta)$, the subsequent values of $\alpha_k(\zeta)$ for $k = 1, \cdots, K+L$ can be calculated by applying the forward update rule

$$\alpha_k(\zeta) = \sum_{\zeta' \in \mathcal{Z}} \gamma_k(\zeta, \zeta') \alpha_{k-1}(\zeta'). \tag{39}$$

After updating (39) $k = K+L$ times we obtain $\alpha_{K+L}(\zeta) = p(\mathbf{z}_{K+L} = \zeta, \mathbf{r}|\mathbf{h})$. The chain link $p(\mathbf{z}_{K+L}|\mathbf{h}, \mathbf{r})$ can be obtained by normalizing

$$p(\mathbf{z}_{K+L} = \zeta|\mathbf{h}, \mathbf{r}) = \frac{\alpha_{K+L}(\zeta)}{\sum\limits_{\zeta'} \alpha_{K+L}(\zeta')}. \tag{40}$$

Thus, the state variable $\mathbf{z}_{K+L}$ can be drawn from the distribution $\bar{\mathbf{z}}_{K+L} \sim p(\mathbf{z}_{K+L}|\mathbf{h}, \mathbf{r})$. Taking into account the current guess the next chain link of (36) can be obtained by the backward updating rule

$$p(\mathbf{z}_k = \zeta|\bar{\mathbf{z}}_{k+1}, \mathbf{h}, \mathbf{r}) = \frac{\gamma_{k+1}(\bar{\mathbf{z}}_{k+1}, \zeta,)\alpha_k(\zeta)}{\sum\limits_{\zeta} \gamma_{k+1}(\bar{\mathbf{z}}_{k+1}, \zeta)\alpha_k(\zeta)}. \tag{41}$$

The subsequent state variable is drawn from

$$\mathbf{z}_k \sim p(\mathbf{z}_k|\bar{\mathbf{z}}_{k+1}, \mathbf{h}, \mathbf{r}). \tag{42}$$

Repeating these two steps up to $k = 0$ we will obtain $\bar{\mathbf{s}}^{(i)} = f(\bar{\mathbf{z}}_0, \cdots, \bar{\mathbf{z}}_{K+L})$ as a random sample according to $p(\mathbf{s}|\bar{\mathbf{h}}^{(i)}, \mathbf{r})$.

In order to avoid floating point overflow, the calculations of the values α and γ are usually realized in the logarithmic scale. Recall that the forward loop of the presented method is identical to the forward loop of the BCJR-algorithm, whereas the backward loop is similar to the backward loop of the Viterbi algorithm. The computational effort of the presented sampler compared to the BCJR-algorithm is approximately cut in half.

E. Averaging

The random character of the instantaneous data estimates $\bar{\mathbf{s}}^{(i)}$ may result in bad estimates. Therefore, the estimates should be averaged over several iterations. After running the iterations I samples of the channel $\bar{\mathbf{h}}^{(i)}$ and data $\bar{\mathbf{s}}^{(i)}$ are available at the receiver. Due to the forgetfulness of Markov chains later samples are more reliable than the former. We can substantially distinguish between two averaging methods:

The straight forward (SF) approach is

$$\hat{\mathbf{s}} = \frac{1}{I} \sum_{i=1}^{I} \bar{\mathbf{s}}^{(i)} \tag{43}$$

and an averaging scheme which is often referred as Rao-Blackwellization (RB) is given by

$$\hat{\mathbf{s}} = \frac{1}{I} \sum_{i=1}^{I} E\{\mathbf{s}|\bar{\mathbf{h}}^{(i)}, \mathbf{r}\}. \tag{44}$$

It can be shown (e.g. [7]) that Rao-Blackwellization has always the lower variance. The calculation of $E\{\mathbf{s}|\bar{\mathbf{h}}^{(i)}, \mathbf{r}\}$ can be performed by the well known BCJR algorithm [14], which is also a forward backward method and can be combined with the presented sampler. However, the computational complexity is at least twice as high as in the first averaging scheme.

F. Combined deterministic and MC structures

All presented iterative block by block equalizers can be described as in Tab. I.

Calling the channel update according to (6) as deterministic method (det.) and according to (27) as Monte Carlo (MC) method, and similarly calling the data update according to (8) as deterministic method and according to (26) as Monte Carlo methods, we have examined several combinations of these parts as shown in Tab. II.

TABLE I

ITERATIVE EQUALIZER

```
init channel by inital guess $\hat{\mathbf{h}}_{(0)}$
for i= 1:I
    • update the data estimates $\mathbf{s}_{(i)}$
    • (if necessary) update the noise power estimates (section IV-B)
    • update the channel estimates $\mathbf{h}_{(i)}$
end
average over $\mathbf{s}_{(i)}$ (if necessary)
```

TABLE II

COMBINED DETERMINISTIC AND MC APPROACHES

termed as	channel	data	averaging	comp. complexity
IJML	det.	det.	none	low
EM	max.	exp.	none	high
MCMCv1	MC	det.	SF	middle
MCMCv2	det.	MC	RB	middle - high
Gibbs	MC	MC	RB	middle - high

V. NUMERICAL RESULTS

Fig. 3(a) and 3(b) compares the BER vs. SNR of all schemes presented in Tab. II after 20 iterations for D-BSK modulated signals, where the normalized squared distance between true channel and initial channel estimate

$$\|\mathbf{h}-\bar{\mathbf{h}}^{(0)}\|^2/\|\mathbf{h}\|^2 \tag{45}$$

was 0 dB. The complex channel gain were independently complex gaussian distributed and the power of the overall impulse response was normalized to $\mathbf{h}^H\mathbf{h}=1$.

Since the problem is better conditioned for a large blocklength, the BER performance illustrated in 3(a) is generally better than in 3(a). However, it can be observed that the difference between deterministic and Monte Carlo approaches becomes smaller in the case of increasing blocklength.

Fig. 3(a) shows that in the high SNR region all MC schemes significantly outperform the IJML, whereby the best results were delivered by the Gibbs-Sampler. At low SNR the BER of IJML is slightly better. The results concerning the low SNR region are astonishing since the Gibbs sampler theoretically always converge to maximum likelihood. However, concerning the BER the joint maximum likelihood might be not the best criterion in the presence of a-prioiri channel state information, which is inherently included by the initial channel guess.

In Fig. 4 the normalized mean squared error (NMSE) defined as in (45) between the true and the estimated channel after 20 iterations is shown. It can be seen that all methods does not significantly differ.

Figure 5(a) shows the convergence behavior of the presented methods at 16 dB SNR. After 4 iterations IJML converges without any further improvements, whereas even after 20 iterations the performances of all MC schemes slightly improve. MCMCv1 and the Gibbs sampler have the slowest convergence speed but intersect MCMCv2 after approximately 8 iterations, whereas MCMCv2 is as fast as IJML and outperforms IJML after 4 iterations. The EM-algorithm converges faster than the Gibbs-sampler, but will be outperformed after approximately 15 iterations.

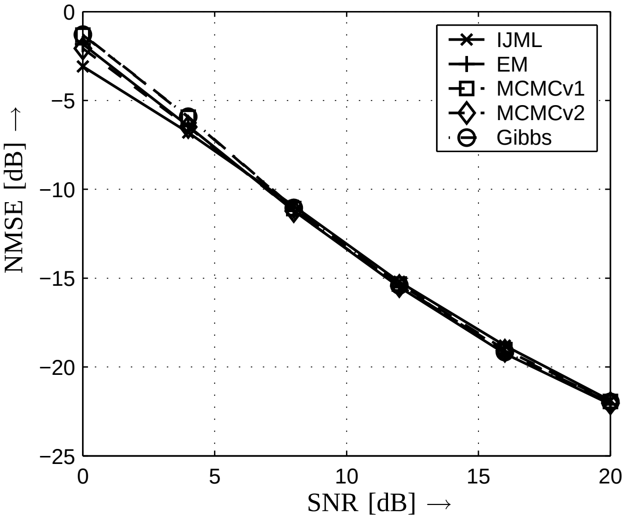


Fig. 4. NMSE vs. SNR with blocklength 10

In Figure 5(b) the average of estimated noise versus iteration steps is plotted for the same configuration as in Fig. 5(a). It can be seen that the current estimate of noise power is an indicator for the BER performance. Please note, that assuming the signal power $|s(k)|^2 = 1$ the noise power estimation is also suitable to estimate the SNR by $SNR = 1/\hat{\sigma}^2$. Thus, the noise power is slightly underestimated.

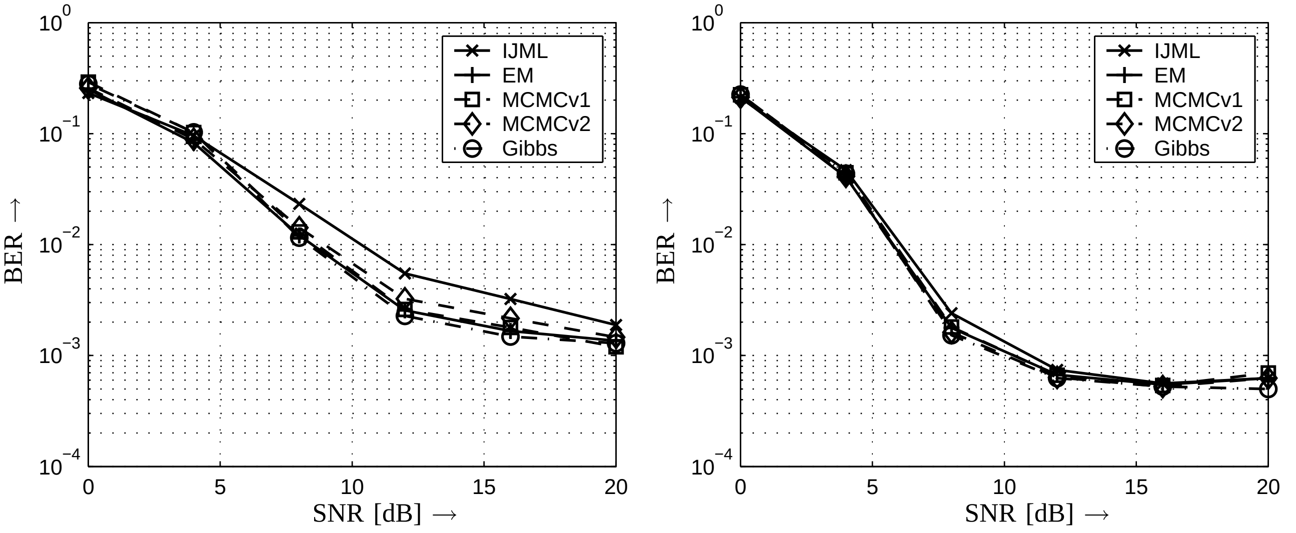
Finally, in Fig. V the BER versus the NMSE of the initial channel estimates is plotted at 8 dB SNR. Although, the MC related methods promise global convergence, their performance depends strongly on the quality of the initialization. Probably, in the bad initialized cases the number of considered iterations are not sufficient. This problem becomes worse with increasing blocklength. Therefore, the blocklength should be kept very small. A possibility for large bloncklength may be obtained by applying sequentiell joint data and channel estimation methods.

VI. CONCLUSION

In comparison to the deterministic channel estimator and data detection schemes the MCMC procedures seem to lead to a degradation of the overall performance due to the artificial deterioration of the estimates. On the other hand IJML neither makes use of the estimation variance of $\hat{\mathbf{h}}$ nor utilizes the pdf of $\mathbf{s}$. In all presented MC-methods the strength of randomness is weighted by an estimate of the noise power, which can be interpreted as a measure of the quality of the current data guess. As it was shown by the numerical results all presented MC-schemes outperform IJML after few iterations. By combining MC and deterministic methods, we can smartly trade off between computational complexity, convergence speed und overall performance. A deterministic alternative to IJML is the EM-algorithm, which suffer from high computational effort.

REFERENCES

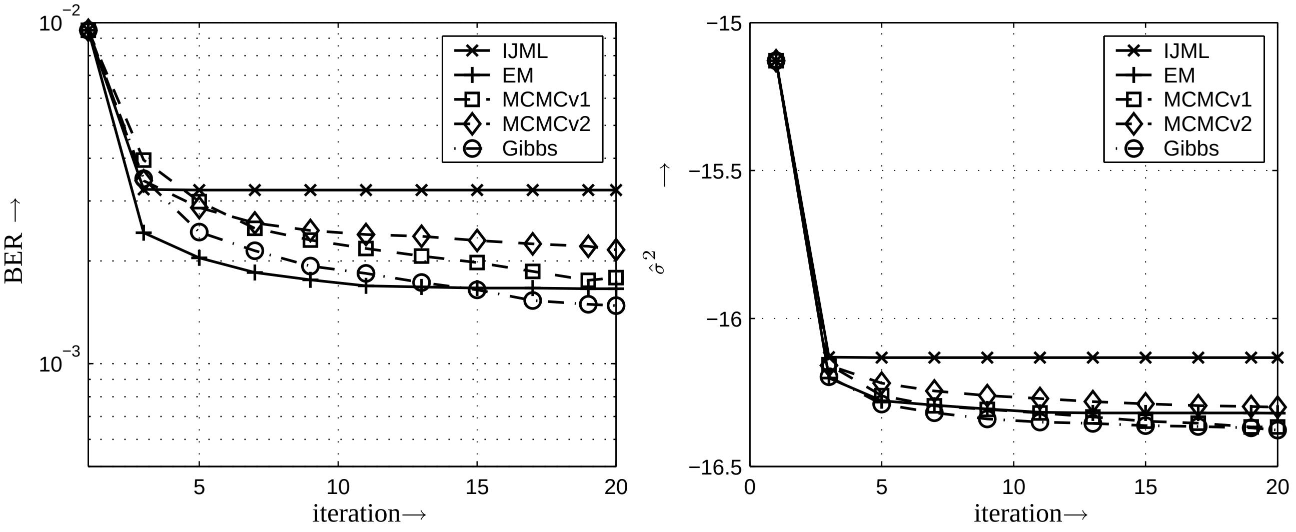
[1] M. Feder and J.A. Catipovic. Algorithms for Joint Channel Estimation and Data Recovery - Application to Equalization in Underwater Com-



(a) blocklength 10

(b) blocklength 40

Fig. 3. BER vs. SNR



(a) BER vs. iteration

(b) $\hat{\sigma}^2$ vs. iteration

Fig. 5. Convergence behavior

munications. *IEEE Journal of Oceanic Engineering*, 16:42–55, January 1991.

[2] G. K. Kaleh and R. Vallet. Joint parameter estimation and symbol detection for linear or nonlinear unknown channels. *IEEE Trans. on Communications*, 42:2406 – 2413, 1994.

[3] X.M. Cheng and X.M. Hoeher. Blind Equalization with Iterative Joint Channel and Data Estimation for Wireless DPSK Systems. *Proc. IEEE GLOBECOM*, pages 274–279, November 2001.

[4] N. Seshadri. Joint Data and Channel Estimation Using Blind Trellis Search Techniques. *IEEE Trans. on Communications*, 42:1000–1011, February 1994.

[5] P.M. Djuric, J.H. Kotecha, J. Zang, Y. Huang, T. Ghirmai, M.F. Bugallo, and J. Miguez. Particle Filtering: A review of the theory and how it can be used for solving problems in wireless communications. *IEEE Signal Processing Magazine*, pages 19 – 38, September 2003.

[6] N. Metropolis, A.W. Rosenbluth, M.N. Rosenbluth, A.H. Teller, and E. Teller. Equations of state calculations by fast computing machines. *Journal of Chemival Physics*, 21-6:1087 – 1091, 1953.

[7] J.S. Liu. *Monte Carlo Strategies in Scientific Computing*. Springer, 2003.

[8] S. Geman and D. Geman. Stochastic relaxation, Gibbs distribution and the Bayesian restoration of images . *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 6:721 – 741, 1984.

[9] R. Chen and T.-H. Li. On Blind restauration of lineary degraded discrete signals by Gibbs sampler. *IEEE Trans. on Signal Processing*, 43:2410 – 2413, September 1995.

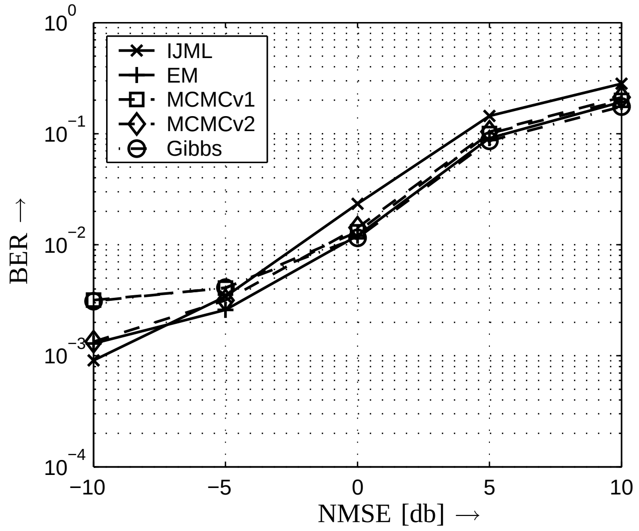


Fig. 6. Influence of the initial channel estimation

[10] J.S. Liu and R. Chen. Blind deconvolution via sequential imputations . *J. Amer. Statist. Assoc.*, 90:567–576, 1995.

[11] J. Miguez and P.M. Djuric. Blind equalization by sequential importance sampling . In *IEEE ISCAS*, pages 845–848., Phoenix, AZ, 2002.

[12] G.D. Forney. Maximum Likelihood Sequence Estimation of Digital Sequence in the Presence of Intersymbol Interference. *IEEE Trans. on Information Theory*, 18:363–378, May 1972.

[13] C.K. Carter and R. Kohn. On Gibbs sampling for state space models. *Biometrika*, 81:541 − 553, 1994.

[14] L. R. Bahl, J. Cocke, F. Jelinek, and J. Raviv. Optimal decoding of linear codes for minimizing symbol error rate. *IEEE Trans. on Information Theory*, 22:284–287, March 1974.