LEARNING IN GAUSSIAN MARKOV RANDOM FIELDS

Thomas J. Riedl, Andrew C. Singer, and Jun Won Choi

University of Illinois at Urbana-Champaign
Email: triedl2@illinois.edu

ABSTRACT

This paper addresses the problem of state estimation in the case where the prior distribution of the states is not perfectly known but instead is parameterized by some unknown parameter. Thus in order to support the state estimator with prior information on the states and improve the quality of the state estimates, it is necessary to learn this unknown parameter first. Here we assume a parameterized Gaussian Markov random field to model the prior distribution of the states and propose an algorithm that is able to learn its parameters from given observations on these states. The effectiveness of this approach is proven experimentally by simulations.

Index Terms— Gaussian Markov random field, EM algorithm, unsupervised learning, sum-product algorithm, factor graph

1. INTRODUCTION

Many problems in Signal Processing can be cast into the framework of state estimation, in which we have state variables $h[i]$ whose values are not directly accessible and variables $y[i]$ whose values are available. Variables of the latter kind are also referred to as observations in this context. Usually there exists a statistical relationship $p(\boldsymbol{y}|\boldsymbol{h})$ between the state variables $h[i]$ and the observations $y[i]$ such that we can infer estimates $\hat{h}[i]$ of the states from the observations. In many cases prior knowledge about the states is also available (usally in form of a probability distribution $p(\boldsymbol{h})$ on the state variables) and we can use that knowledge to refine the state estimate.

In a variety of interesting problems, however, neither the statistical relationship between the state variables and the observations nor the prior distribution are perfectly known and hence are modeled as parameterized distributions $p(\boldsymbol{y}|\boldsymbol{h},\theta)$ and $p(\boldsymbol{h}|\theta)$ with unknown parameters θ. These parameters are then also subject to estimation.

Here we restrict the prior distribution on the hidden state variables to the form of a parametrized Gaussian Markov random field and assume a simple parametrized linear observation model. We shall propose an efficient algorithm to estimate the unknown parameters. Our algorithm can be interpreted as an approximation to the well known expectation maximization (EM) algorithm.

An interesting example of an signal processing problem that fits the framework of state estimation is channel estimation. The widely used wide sense stationary uncorrelated scattering model for the communications channel neglects correlations between different multipath arrivals [1–3], but this seems to oversimplify the real channel in many cases. One example is the underwater acoustic channel, whose impulse response is fairly continuous in delay and hence indeed exhibits a certain correlation structure in delay.

To address this shortcoming [4] introduced a novel channel model that is based on a Gaussian Markov random field (MRF) for the complex channel gains. This graphical model is used to capture the local nature of the statistical dependencies (in time and space) of the channel taps.

In order for the MRF model to fit the actual physical channel well, its parameters must be adapted appropriately. This is the topic of the algorithm proposed in this paper. Once these parameters are known, the MRF model can then either be used for channel estimation [4] or it can be embedded into an iterative (turbo) receiver [5], where it is expected to improve the data estimation performance significantly as the parameterized MRF carries prior knowledge on the channel.

2. PARAMETER ESTIMATION

2.1. Problem Setup

We consider an incomplete data problem where some of the variables are hidden and others are observable. The derivations in this paper assume that the hidden variables are modeled by the generic lattice shaped gaussian Markov Random Field model introduced in [4]. However, with minor modifications the results hold for other Gaussian graphical models as well. Due to the geometric nature of lattices, we prefer to index the state variables by the two dimensional index pair $[i,j]$. The graphical model in [4] is illustrated by the factor graph in Figure 1, and completely characterized by the following

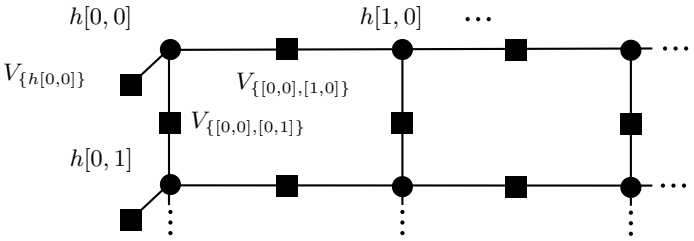

Fig. 1. Factor graph based on MRF model

joint distribution

$$p(\boldsymbol{H}|\boldsymbol{\theta}) = Z(\boldsymbol{\theta})^{-1} \prod_{b\in \mathrm{B}_1} \exp(-\alpha_{[j_b]}|h[i_b,j_b]-\mu[j_b]|^2)$$
$$\cdot \prod_{b\in \mathrm{B}_2} \exp(-\alpha_{[i_b,j_b],[m_b,n_b]}|(h[i_b,j_b]-\mu[j_b])$$
$$-(h[m_b,n_b]-\mu[n_b])|^2)$$

where the vector $\boldsymbol{\theta}$ containes all the model parameters i.e the $\alpha_{[i_b,j_b],[m_b,n_b]}$, the $\alpha_{[j_b]}$ and the $\mu[j]$. The set B_1 comprises all single cliques that correspond to state variables $h[i,j]$ at $i=0$ and the set B_2 contains all the pairwise cliques. So $\boldsymbol{H}$ is jointly Gaussian distributed and $h[i,j]$ has mean $\mu[j]$. The structure of this

lattice shaped graphical model also imposes a certain continuity on the behavior of neighboring state variables, as $p(h[i,j]|h[l,m] : [l,m] \in \mathbf{N}[i,j])$ for $i \neq 0$ then becomes a complex Gaussian distibution with mean

$$\mu[j] + \frac{\sum_{[l,m]\in\mathbf{N}[i,j]} \alpha_{[i,j],[l,m]}(h[l,m]-\mu[m])}{\sum_{[l,m]\in\mathbf{N}[i,j]} \alpha_{[i,j],[l,m]}} \tag{1}$$

and variance

$$\left(\sum_{[l,m]\in\mathbf{N}[i,j]} \alpha_{[i,j],[l,m]}\right)^{-1}. \tag{2}$$

The mean of $h[i,j]$ is shifted by a weighted sum of the differences $h[l,m]-\mu[m]$. So if all the neighbors, for example, assumed values above their means, $h[i,j]$ is likely to assume a value that is above its mean as well. The $\alpha_{[i,j],[l,m]}$'s determine what impact each neighbor has on $h[i,j]$.

The observations $y[i]$ and the hidden variables $\boldsymbol{H}$ are assumed to have the following statistical relationship

$$y[i] = \boldsymbol{h}[i]^{\mathrm{T}}\boldsymbol{x}[i] + w[i], \qquad i = 0,\dots,M-1 \tag{3}$$

where $w[i]$ denotes additive complex white Gaussian noise with zero mean and circular symmetric variance σ_n^2, the $\boldsymbol{x}[i]$ are some complex vectors and $\boldsymbol{h}[i] = \boldsymbol{H}_{(i,:)}$. If we interpret i as a discrete time index, $y[i]$ could be considered as the output of a linear system that has $\boldsymbol{x}[i]$ as its input plus noise. We let $\boldsymbol{\theta}$ now also comprise the noise variance σ_n^2.

Our goal is to estimate the paramters $\boldsymbol{\theta}$ and we do so in a maximum likelihood (ML) fashion.

$$\hat{\boldsymbol{\theta}} = \operatorname*{argmax}_{\boldsymbol{\theta}} p(\boldsymbol{y}|\boldsymbol{\theta}) = \operatorname*{argmax}_{\boldsymbol{\theta}} \int_{\boldsymbol{H}} p(\boldsymbol{y},\boldsymbol{H}|\boldsymbol{\theta}) \tag{4}$$

where it can easily be checked that

$$p(\boldsymbol{y},\boldsymbol{H}|\boldsymbol{\theta}) = p(\boldsymbol{y}|\boldsymbol{H},\boldsymbol{\theta})p(\boldsymbol{H}|\boldsymbol{\theta}) \tag{5}$$

The contribution of this section is the development of an efficient algorithm for the estimation of these parameters.

Numerical evaluation of maximum-likelihood estimates is often difficult. As a remedy we will use a powerful optimization method that has been used with great success in many applications: The Expectation Maximization (EM) algorithm [6]. A short review of this algorithm is in order:

1. Make some initial guess $\hat{\boldsymbol{\theta}}^{(0)}$
2. Expectation step: calculate

$$Q(\boldsymbol{\theta},\hat{\boldsymbol{\theta}}^{(k)}) = \langle \log p(\boldsymbol{y},\boldsymbol{H}|\boldsymbol{\theta})\rangle_{p(\boldsymbol{H}|\boldsymbol{y},\hat{\boldsymbol{\theta}}^{(k)})} \tag{6}$$

3. Maximization step: compute

$$\hat{\boldsymbol{\theta}}^{(k+1)} = \operatorname*{argmax}_{\boldsymbol{\theta}} Q(\boldsymbol{\theta},\hat{\boldsymbol{\theta}}^{(k)}) \tag{7}$$

4. Repeat 6-7 until convergence or until the available time is over.

$\langle\cdot\rangle_p$ represents expectation with respect to p. Under rather general conditions this algorithm is proven to yield a nondecreasing sequence $p(\boldsymbol{y}|\hat{\boldsymbol{\theta}}^{(k)})$. However, it is well known that due to the interaction between the $h[i,j]$, the precise calculation of the partition function $Z(\boldsymbol{\theta})$ and the integral in 6 is not computationally feasible [7].

2.2. Solution

One can bypass the requirement of exactly knowing the partition function $Z(\boldsymbol{\theta})$ by approximating the maximization step above by a gradient ascent step. It should be noted, however, that taking all parameters in $\boldsymbol{\theta}$ as independent and distinct parameters would seriously overparameterize our model, since there would be more parameters than available observations. To tackle this problem we assume that all $\alpha_{[i,j],[l,m]}$s that correspond to a vertical pairwise clique are the same and equal α_v and similarly that all $\alpha_{[i,j],[l,m]}$s that correspond to a horizontal pairwise clique are the same and equal α_h. Also we take $\alpha_{[j_b]} = \alpha$.

As mentioned above we substitute the maximization step in the EM algorithm with an gradient ascent step. So lets procede with the calculation of the gradient of $Q(\boldsymbol{\theta},\hat{\boldsymbol{\theta}}^{(k)})$ with respect to $\boldsymbol{\theta}$.

$$\begin{aligned}\frac{\partial}{\partial\sigma_n^2}Q(\boldsymbol{\theta},\hat{\boldsymbol{\theta}}^{(k)}) = &-\frac{M}{\sigma_n^2} - \sum_{i=0}^{M-1}\langle |y[i] \\ &- \sum_{j=0}^{L-1} h[i,j]x[i,j]|^2\rangle_{p(\boldsymbol{H}|\boldsymbol{y},\hat{\boldsymbol{\theta}}^{(k)})}\end{aligned} \tag{8}$$

And the partial derivatives with respect to the MRF parameters θ_j have the following form

$$\begin{aligned}\frac{\partial}{\partial\theta_j}Q(\boldsymbol{\theta},\hat{\boldsymbol{\theta}}^{(k)}) = &\langle\frac{\partial}{\partial\theta_j}\sum_{b\in\mathrm{B}} V_b(\bar{b})\rangle_{p(\boldsymbol{H}|\boldsymbol{\theta})} \\ &-\langle\frac{\partial}{\partial\theta_j}\sum_{b\in\mathrm{B}} V_b(\bar{b})\rangle_{p(\boldsymbol{H}|\boldsymbol{y},\hat{\boldsymbol{\theta}}^{(k)})}\end{aligned} \tag{9}$$

where

$$\begin{aligned}2\frac{\partial}{\partial\mu[j]^\star}\sum_{b\in\mathrm{B}} V_b(\bar{b}) = &2\Big(\alpha_v\Big(\sum_i h[i,j-1] - M\mu[j-1]\Big) \\ &+\alpha_v\Big(\sum_i h[i,j+1] - M\mu[j+1]\Big) \\ &-2\alpha_v\Big(\sum_i h[i,j] - M\mu[j]\Big) \\ &-\alpha\Big(h[1,j]-\mu[j]\Big)\Big)\end{aligned} \tag{10}$$

$$\frac{\partial}{\partial\alpha}\sum_{b\in\mathrm{B}} V_b(\bar{b}) = \sum_j |h[1,j]-\mu[j]|^2 \tag{11}$$

$$\begin{aligned}\frac{\partial}{\partial\alpha_v}\sum_{b\in\mathrm{B}} V_b(\bar{b}) = &\sum_{i,j}|(h[i,j]-\mu[j]) \\ &-(h[i,j+1]-\mu[j+1])|^2\end{aligned} \tag{12}$$

$$\frac{\partial}{\partial\alpha_h}\sum_{b\in\mathrm{B}} V_b(\bar{b}) = \sum_{i,j}|(h[i,j]-(h[i+1,j])|^2 \tag{13}$$

and $\frac{\partial}{\partial\mu[j]^\star}$ denotes the Wirtinger derivative with respect to $\mu[j]^\star$. Note that the partial derivatives of $\sum_{b\in\mathrm{B}} V_b(\bar{b})$ with respect to the real and imaginary component of $\mu[j]$ coincide with the real and imaginary component of $2\frac{\partial}{\partial\mu[j]^\star}\sum_{b\in\mathrm{B}} V_b(\bar{b})$, respectively. Clearly in order to actually calculate the gradient at $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}^{(k)}$ we first need to know the moments $\langle h[i,j]\rangle$, $\mathrm{Cov}(h[i,j])$, $\mathrm{Cov}([h[i,j],h[i,j+1]])$ and $\mathrm{Cov}([h[i,j],h[i+1,j]])$ with respect to $p(\boldsymbol{H}|\hat{\boldsymbol{\theta}}^{(k)})$ and also with respect to $p(\boldsymbol{H}|\boldsymbol{y},\hat{\boldsymbol{\theta}}^{(k)})$. This can be achieved by use of the

sum-product algorithm [8]. Note that the random variables in $\boldsymbol{H}$ are jointly Gaussian and hence the messages that the sum-product algorithm passes along the edges of a factor graph are Gaussian as well. Gaussian messages are parameterized by a mean vector and a covariance matrix and so the required moments are readily computed by the operation of the sum-product algorithm. The factor graph the sum-product algorithm operates on for calculation of the moments with respect to $p(\boldsymbol{H}|\boldsymbol{y}, \hat{\boldsymbol{\theta}}^{(k)})$ is shown in Figure 2. The convenient

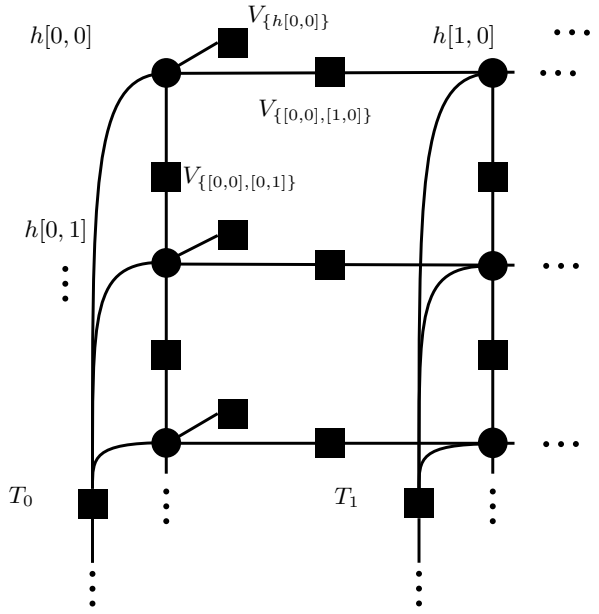

Fig. 2. Factor graph for MAP channel estimation

factorization of $p(\boldsymbol{H}|\boldsymbol{y}, \hat{\boldsymbol{\theta}}^{(k)})$ that the MRF framework provides reduces the complexity of the marginalization tremendously. The factor graph corresponding to $p(\boldsymbol{H}|\hat{\boldsymbol{\theta}}^{(k)})$ coincides with the one in Figure 2 when the functions T_i are eliminated.

The remainder of this section is dedicated to the implementation of the sum-product algorithm for the calculation of the moments. The factor graph corresponding to $p(\boldsymbol{H}|\hat{\boldsymbol{\theta}}^{(k)})$ is contained in the factor graph corresponding to $p(\boldsymbol{H}|\boldsymbol{y}, \hat{\boldsymbol{\theta}}^{(k)})$ and hence the message passing on $p(\boldsymbol{H}|\hat{\boldsymbol{\theta}}^{(k)})$ is just a special case of the one on $p(\boldsymbol{H}|\boldsymbol{y}, \hat{\boldsymbol{\theta}}^{(k)})$. For that reason it suffices to derive the message passing rules for the factor graph associated with $p(\boldsymbol{H}|\boldsymbol{y}, \hat{\boldsymbol{\theta}}^{(k)})$.

The message passing rules required for calculation of the moments $\langle h[i,j] \rangle$ and $\mathrm{Cov}(h[i,j])$ are equivalent to the ones presented in [4].

We see that the messages that are sent along the edges of our factor graph are of three different kinds. Messages of the first kind come from a variable node, messages of the second kind come from one of the potential functions and messages of the third kind come from one of the functions T_i. These three different types of messages are illustrated in Figure 3. The type of the message is superscriped in each case. The derivation of the update rules can be found in [4] and we summarize the results here.

update rule for messages of the first kind

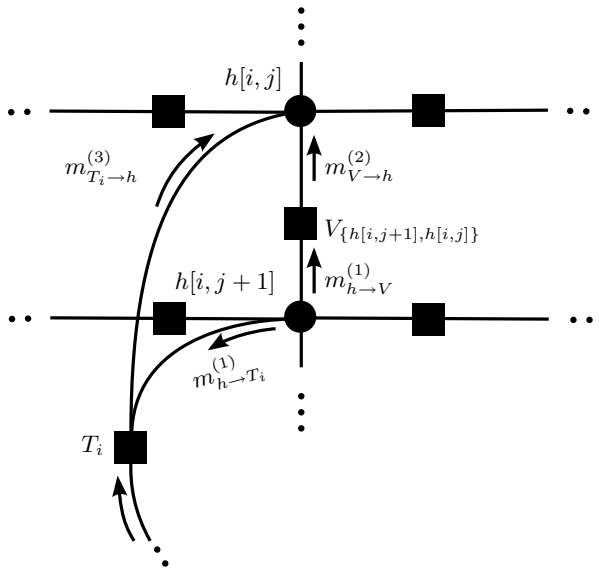

Fig. 3. A closer look at the factor graph from Figure 2

$$\begin{aligned} \mu_{m_{h\to f}} &= \frac{\sum_{g\in n(h)\setminus\{f\}} \sigma_g^{-2}\mu_g}{\sum_{g\in n(h)\setminus\{f\}} \sigma_g^{-2}} \\ \sigma^{-2}_{m_{h\to f}} &= \sum_{g\in n(h)\setminus\{f\}} \sigma_g^{-2} \end{aligned} \tag{14}$$

where μ_g and σ_g^2 are the mean and the variance of the message $m_{g\to h}(h)$, respectively.

update rule for messages of the second kind

$$\begin{aligned} \mu_{m_{V_b\to h}} &= \mu_u + \mu[j_h] - \mu[j_u] \\ \sigma^{-2}_{m_{V_b\to h}} &= \frac{\sigma_u^{-2}\alpha_{[i_h,j_h],[i_u,j_u]}}{\sigma_u^{-2} + \alpha_{[i_h,j_h],[i_u,j_u]}} \end{aligned} \tag{15}$$

where μ_u and σ_u^2 are the mean and the variance of the message $m_{u\to V_b}(u)$, $u \in n(V_b)\setminus\{h\}$, respectively, and $[i_h, j_h]$ and $[i_u, j_u]$ are the coordinates of the nodes h and u, respectively.

update rule for messages of the third kind

$$\begin{aligned} \sigma^{-2}_{m_{T_i\to h[i,j]}} &= |x[i-j]|^2\sigma_n^{-2}(1+z)^{-1} \\ z &= \sigma_n^{-2}\sum_{l=0,l\neq j}^{L-1} |x[i-l]|^2\sigma^2_{m_{h[i,l]\to T_i}} \\ \mu_{m_{T_i\to h[i,j]}} &= x[i-j]^{-1}(y[i] - \sum_{l=0,l\neq j}^{L-1} x[i-l]\mu_{m_{h[i,l]\to T_i}}) \end{aligned} \tag{16}$$

In order to obtain the moments $\mathrm{Cov}([h[i,j], h[i,j+1]])$ and $\mathrm{Cov}([h[i,j], h[i+1,j]])$ as well, we need to modify our current factor graph setup slightly. We cluster the corresponding pairs of nodes $h[i,j]$. Figure 4 shows what effect the clustering of $(h[i,j], h[i,j+1])$ has on the factor graph in Figure 3. The update rule for messages of the third kind as listed above must then be modified as follows.

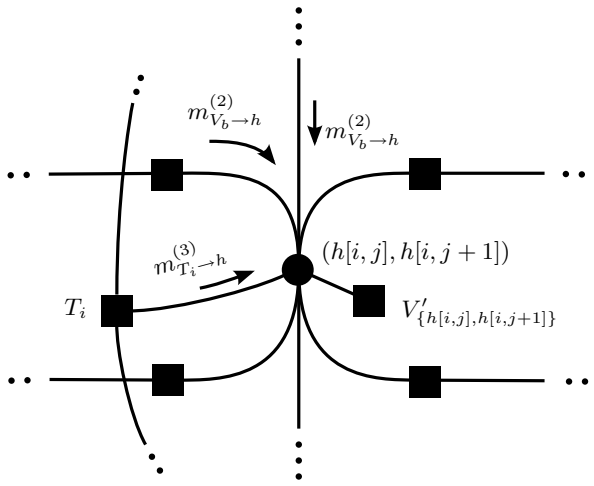

Fig. 4. factor graph

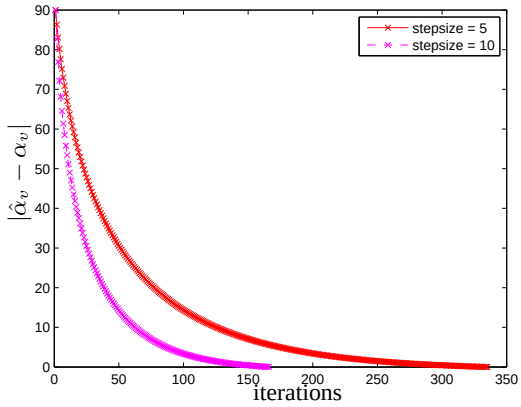

Fig. 5. The absolute estimation error of α_v over iterations

update rule for messages of the third kind

$$\boldsymbol{\Sigma}^{-1}_{m_{T_i\to(h[i,j],h[i,j+1])}} = \begin{bmatrix} x[i-j] \\ x[i-j-1] \end{bmatrix}^{\star} \begin{bmatrix} x[i-j] \\ x[i-j-1] \end{bmatrix}^{\mathrm{T}} \cdot\sigma_n^{-2}(1+z)^{-1} \tag{17}$$

$$z = \sigma_n^{-2} \sum_{l=0,l\neq j,j+1}^{L-1} |x[i-l]|^2 \sigma^2_{m_{h[i,l]\to T_i}}$$

$$\boldsymbol{\mu}_{m_{T_i\to(h[i,j],h[i,j+1])}} = \begin{bmatrix} 1 \\ 0 \end{bmatrix} x[i-j]^{-1} \cdot(y[i] - \sum_{l=0,l\neq j,j+1}^{L-1} x[i-l]\mu_{m_{h[i,l]\to T_i}}) \tag{18}$$

2.3. Analysis

As mentioned above it is well known that the EM algorithm yields a nondecreasing sequence of likelihoods $p(\boldsymbol{y}|\hat{\boldsymbol{\theta}}^{(k)})$. This property remains true even if the maximization step is replaced by a gradient ascent step as proposed in this paper. A proof of this result can be found in the appendix. Note, however, that the above algorithm only approximates this gradient. As our factor graph does contain cycles, the sum-product algorithm only approximately calculates the moments required for setting up the gradient [9].

3. SIMULATION RESULTS

We chose to evaluate the performance of the proposed estimator on synthetic data as this enables us to compare the obtained parameter estimate against the actual value of the parameter. So we draw realisations of the state variables from the probability distribution $p(\boldsymbol{H}|\theta)$, observed some noisy observations $\boldsymbol{y}$ and finally employed the proposed algorithm on these observations to obtain an estimate $\hat{\boldsymbol{\theta}}$ of $\boldsymbol{\theta}$. For our simulation we parameterized the MRF model with the following values. The parameters α, α_v, α_h and $\boldsymbol{\mu}$ were set to 100, 100, 1000 and $[0\,0.6\,0.4\,1\,0.2]$, respectively. Figure 5 shows how the absolute estimation error of one of the parameters approaches zero over the iterations. The convergence speed of the gradient ascent algorithm depends on the step size and so does the algorithm presented here.

4. REFERENCES

[1] S. Song, A. C. Singer, and K.-M. Sung, "Soft input channel estimation for turbo equalization," *IEEE Transactions on Signal Processing*, vol. 52, no. 10, pp. 2885–2894, October 2004.

[2] K. J. Kim and R. A. Iltis, "Joint detection and channel estimation algorithms for qs-cdma signals over time-varying channels," *IEEE Transactions on Communications*, vol. 50, no. 5, pp. 845–855, May 2002.

[3] W. Ling and L. Ting, "Kalman filter channel estimation based on comb-type pilot in time-varying channel," *IEEE Transactions on Communications*, vol. 50, no. 5, pp. 845–855, May 2002.

[4] T. J. Riedl, J. W. Choi, and A. C. Singer, "Channel estimation by inference on gaussian markov random fields," *Proc. of the Asilomar Conference on Signals, Systems and Computers*, 2009, accepted for publication, to appear.

[5] A. P. Worthen and W. E. Stark, "Unified design of iterative receivers using factor graphs," *IEEE Transactions on Information Theory*, vol. 47, no. 2, pp. 843–849, February 2001.

[6] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the em algorithm," *Journal of the Royal Statistical Society*, vol. 39, no. 1, pp. 1–38, 1977.

[7] D. Chandler, "Introduction to modern statistical mechanics," *Oxford University Press*, 1987.

[8] F. R. Kschischang, B. J. Frey, and H.-A. Loeliger, "Factor graphs and the sum-product algorithm," *IEEE Transactions on Information Theory*, vol. 47, no. 2, pp. 498–519, February 2001.

[9] Y. Weiss and W. T. Freeman, "Correctness of belief propagation in gaussian graphical models of arbitrary topology," *Neural Computation*, vol. 13, no. 10, pp. 2173–2200, 2001.