in Proc. IEEE ICASSP 2010
Dallas, TX, March 2010, pp. 3886-3889

A PARTIALLY COLLAPSED GIBBS SAMPLER FOR PARAMETERS WITH LOCAL CONSTRAINTS

Georg Kail[a], Jean-Yves Tourneret[b], Franz Hlawatsch[a], and Nicolas Dobigeon[b]

[a]Institute of Communications and Radio-Frequency Engineering, Vienna University of Technology
Gusshausstrasse 25/389, A-1040 Vienna, Austria; e-mail: gkail@nt.tuwien.ac.at

[b]University of Toulouse, IRIT/ENSEEIHT/TéSA
2 rue Charles Camichel, BP 7122, 31071 Toulouse cedex 7, France; e-mail: jean-yves.tourneret@enseeiht.fr

ABSTRACT

We consider Bayesian detection/classification of discrete random parameters that are strongly dependent locally due to some deterministic local constraint. Based on the recently introduced *partially collapsed Gibbs sampler* (PCGS) principle, we develop a Markov chain Monte Carlo method that tolerates and even exploits the challenging probabilistic structure imposed by deterministic local constraints. We study the application of our method to the practically relevant case of nonuniformly spaced binary pulses with a known minimum distance. Simulation results demonstrate significant performance gains of our method compared to a recently proposed PCGS that is not specifically designed for the local constraint.

***Index Terms*—** Markov chain Monte Carlo method, partially collapsed Gibbs sampler, pulse detection, deterministic constraints.

1. INTRODUCTION

We consider Bayesian detection/classification of a sequence of K discrete random parameters $\boldsymbol{\theta} = (\theta_1 \cdots \theta_K)^{\mathrm{T}}$. It is assumed that the parameters θ_k are strongly dependent locally due to some *deterministic local constraint* whereas dependencies between more distant parameters are weak. A simple example is a random sequence of nonuniformly spaced binary pulses with a known minimum distance between any two pulses and weak dependencies otherwise. This example, which will be discussed in the course of this paper, is interesting for many signal processing applications including signal segmentation [1], layer detection [2], and electromyography [3].

A Monte Carlo approximation [4] is often used in detection/classification problems where a direct solution is too complex. The Gibbs sampler (GS) [4] is the simplest and most popular Monte Carlo method. Unfortunately, the GS has been observed to be computationally inefficient if, as in the case studied here, there are strong dependencies among the parameters [5]. The *partially collapsed Gibbs sampler* (PCGS) [6] is a recently proposed generalization of the GS which allows certain modifications that improve the convergence behavior while preserving the target distribution.

In this paper, we develop a PCGS for the detection/classification of a sequence of parameters with a deterministic local constraint. Our method tolerates and even exploits the challenging probabilistic structure imposed by such constraints, which constitutes a significant advantage over more general sampler concepts. We study the application of our method to the practically relevant example mentioned above and show via simulation that it outperforms a recently proposed PCGS that is not specifically tailored to the local constraint.

This paper is organized as follows. In Section 2, we describe the problem and an application example. Section 3 briefly reviews the GS and PCGS principles. The proposed method is presented in Section 4. In Section 5, it is applied to a practically relevant problem. Its performance is assessed experimentally in Section 6.

This work was supported by the FWF under grant S10603-N13 (Statistical Inference) within the National Research Network SISE.

2. PROBLEM STATEMENT

Consider a discrete random parameter vector $\boldsymbol{\theta} = (\theta_1 \cdots \theta_K)^{\mathrm{T}}$ whose elements θ_k are drawn from some M-ary finite alphabet $\mathcal{A}$ (i.e., $|\mathcal{A}| = M$). We assume that $\boldsymbol{\theta}$ is subject to some deterministic local constraint (to be discussed presently). We also allow further random parameters, collected in the vector $\boldsymbol{\phi}$, which are not constrained and may be discrete or continuous. The goal is to detect or classify $\boldsymbol{\theta}$ and to detect, classify, or estimate $\boldsymbol{\phi}$ from an observed data vector $\mathbf{x}$. In the Bayesian setting adopted, this problem is fully described by the posterior distribution $p(\boldsymbol{\theta}, \boldsymbol{\phi}|\mathbf{x})$.

We now define the class of "deterministic local constraints" we will consider. A *deterministic* constraint reduces the set of possible hypotheses for $\boldsymbol{\theta}$ in that $\boldsymbol{\theta} \in \mathcal{C}$ rather than $\boldsymbol{\theta} \in \mathcal{A}^K$, with a "constraint set" $\mathcal{C} \subset \mathcal{A}^K$. The prior of $\boldsymbol{\theta}$ then satisfies $p(\boldsymbol{\theta}) = 0$ for $\boldsymbol{\theta} \notin \mathcal{C}$. The deterministic constraint can be considered *local* if it causes strong dependencies between parameters θ_k that are close to each other but not between distant parameters. Thus, the effect of the local constraint on the prior $p(\boldsymbol{\theta})$ can be described not only in terms of the entire sequence $\boldsymbol{\theta}$ as above but also locally, in terms of neighborhoods. For some fixed positive $d \leq K$ (typically, $d \ll K$) and an arbitrary $k \in \{1, \ldots, K-d+1\}$, let $\mathcal{J}_d(k) \triangleq \{k, \ldots, k+d-1\}$ denote some right-hand neighborhood of k and $\boldsymbol{\theta}_{\mathcal{J}_d(k)} \triangleq (\theta_k \cdots \theta_{k+d-1})^{\mathrm{T}}$ the corresponding subvector of $\boldsymbol{\theta}$. In general, $\boldsymbol{\theta}_{\mathcal{J}_d(k)} \in \mathcal{A}^d$ and thus there would be M^d hypotheses $\boldsymbol{\theta}_{\mathcal{J}_d(k)}$ for any given k. However, due to the deterministic local constraint, the set of hypotheses is reduced to a smaller set $\mathcal{C}_d$, i.e., we know that $\boldsymbol{\theta}_{\mathcal{J}_d(k)} \in \mathcal{C}_d$ and thus

$$p(\boldsymbol{\theta}) = 0 \quad \text{for all } \boldsymbol{\theta} \text{ such that } \boldsymbol{\theta}_{\mathcal{J}_d(k)} \notin \mathcal{C}_d\,,$$

for all $k \in \{1, \ldots, K-d+1\}$. We note that our results can be generalized to k-dependent sets $\mathcal{A}(k)$ and $\mathcal{C}_d(k)$.

Example. Consider binary parameters $\theta_k \in \{0, 1\}$ (i.e., $M = 2$) with a given minimum distance $d_{\min}$ between successive 1's in the sequence. That is, a $\boldsymbol{\theta}$ for which two elements $\theta_k = 1$ and $\theta_{k'} = 1$ are closer than $d_{\min}$, i.e., $|k - k'| < d_{\min}$, has zero probability. Consider a subvector $\boldsymbol{\theta}_{\mathcal{J}_d(k)} = (\theta_k \cdots \theta_{k+d-1})^{\mathrm{T}}$ for a fixed k. For convenience, we choose $d = d_{\min}$. Without the minimum-distance constraint, there would be 2^d possible hypotheses for $\boldsymbol{\theta}_{\mathcal{J}_d(k)}$. However, due to the constraint, all $\boldsymbol{\theta}$ for which $\boldsymbol{\theta}_{\mathcal{J}_d(k)}$ contains more than one 1 have zero probability. Thus, there are $d+1$ admissible hypotheses $\boldsymbol{\theta}_{\mathcal{J}_d(k)} \in \mathcal{C}_d$, namely one with no 1 and d with one 1.

3. GIBBS SAMPLER AND PCGS

Markov chain Monte Carlo (MCMC) methods [4] are often used when the analytic expression of a detector or an estimator (typically

an MMSE or MAP estimator) is too complex to be calculated directly. MCMC methods generate samples from the distribution of interest by constructing an ergodic Markov chain. Calculations are then based on the samples rather than on the distribution itself. In this section, we briefly review the GS and the PCGS and discuss their suitability for problems with a deterministic constraint.

GS. Let $\boldsymbol{\vartheta}$ be a vector of K parameters, and let $\boldsymbol{\vartheta}_{\sim k}$ denote $\boldsymbol{\vartheta}$ without the kth element ϑ_k (a generalization to the case where ϑ_k is itself a vector is possible). To obtain samples from $p(\boldsymbol{\vartheta})$ (which corresponds to $p(\boldsymbol{\theta}, \boldsymbol{\phi}|\mathbf{x})$ in our problem), the GS iteratively generates samples of each ϑ_k from $p(\vartheta_k|\boldsymbol{\vartheta}_{\sim k})$ in an arbitrary order. This strategy is known to converge to the target distribution $p(\boldsymbol{\vartheta})$. After convergence, K such sampling substeps produce a new sample $\boldsymbol{\vartheta}$ from $p(\boldsymbol{\vartheta})$. A known weakness of the GS is that dependencies between the ϑ_k tend to result in slow convergence [5].

PCGS. The PCGS is an extension of the GS that allows for the following three modifications (see [6] for details).

- *Marginalization.* Rather than sampling only ϑ_k in substep k, some other parameters may be sampled along with ϑ_k instead of being conditioned upon. This can improve the convergence rate significantly, especially for strong dependencies between the parameters. Within one entire PCGS iteration, some parameters are then sampled in more than one substep.
- *Trimming.* If a parameter is sampled in several substeps and is not conditioned upon between these substeps, only the value sampled in the last of these substeps is relevant, since the other values are never used. Such unused parameters can thus be removed from the respective sampling distribution. This reduces the complexity of the sampling steps while not affecting the convergence behavior. Note that the distributions used for sampling are generally no longer conditional distributions associated with the full joint distribution $p(\boldsymbol{\vartheta})$, but rather conditional distributions associated with certain marginal distributions of $p(\boldsymbol{\vartheta})$.
- *Permutation.* It is reasonable to choose the (arbitrary) sampling order such that trimming can be performed to a maximum extent. After trimming, permutations are only allowed if they preserve the justification of the trimming already applied.

These modifications do not change the stationary distribution of the Markov chain. The flexibility of the PCGS regarding the choice of the sampling distributions makes it applicable to many cases in which the posteriors required by the GS cannot be calculated analytically (for potential applications, see [7] and references therein).

Deterministic constraints. The GS is poorly suited to problems with deterministic constraints as considered in Section 2. A constraint that excludes parts of the hypothesis space may even inhibit convergence to $p(\boldsymbol{\vartheta})$ altogether. This is because each of the K sampling substeps constitutes a jump along one of the axes of the K-dimensional hypothesis space. A deterministic constraint may restrict the hypotheses with nonzero probability to disjoint regions such that one region cannot be reached from another by such jumps.

Sampling parameters jointly, as in the PCGS, corresponds to a jump along the linear span of the axes associated with the respective parameters. Thus, there are more configurations of disjoint regions in the hypothesis space between which the sampler can jump. In the GS, the parameters ϑ_k may be grouped into vectors, too, but these vectors must be disjoint. Therefore, the possible directions of the jumps are still orthogonal to each other. In the restricted hypothesis space, more freedom is needed for the jumps, which is provided by the overlapping subvectors in the PCGS.

4. THE PROPOSED SAMPLER

We now apply the PCGS principle to obtain a fast-converging sampler for problems with a deterministic local constraint. As in Section 2, we consider a constrained discrete parameter vector $\boldsymbol{\theta} = (\theta_1 \cdots \theta_K)^{\mathrm{T}}$ and an unconstrained parameter vector $\boldsymbol{\phi} = (\phi_1 \cdots \phi_L)^{\mathrm{T}}$. The task is to sample from the posterior $p(\boldsymbol{\theta}, \boldsymbol{\phi}|\mathbf{x})$. In order to extend the definition of the neighborhoods $\mathcal{J}_d(k)$ to all $k \in \{1, \ldots, K\}$, we redefine $\mathcal{J}_d(k)$ as $\{k, \ldots, k_{\max}(k)\}$ with $k_{\max}(k) = \min\{k+d-1, K\}$. The choice of the neighborhood width d will be discussed presently. The proposed PCGS algorithm is now stated as follows.

PCGS for problems with a deterministic local constraint:

- For $k = 1, \ldots, K$, sample θ_k from $p\big(\theta_k \big| \boldsymbol{\theta}_{\sim \mathcal{J}_d(k)}, \boldsymbol{\phi}, \mathbf{x}\big)$.
- For $l = 1, \ldots, L$, sample ϕ_l from $p(\phi_l|\boldsymbol{\theta}, \boldsymbol{\phi}_{\sim l}, \mathbf{x})$.

Note that the sampling distribution for θ_k is a conditional distribution associated with $p\big(\theta_k, \boldsymbol{\theta}_{\sim \mathcal{J}_d(k)}, \boldsymbol{\phi}\big|\mathbf{x}\big)$, which is the joint posterior $p(\boldsymbol{\theta}, \boldsymbol{\phi}|\mathbf{x})$ marginalized with respect to all parameters in $\mathcal{J}_d(k)$ except θ_k (i.e., with respect to $\theta_{k+1}, \ldots, \theta_{k_{\max}(k)}$). Accordingly, $\boldsymbol{\theta}_{\mathcal{J}_d(k)}$ is not contained in the condition for θ_k, which is therefore sampled regardless of the previous sample of $\boldsymbol{\theta}_{\mathcal{J}_d(k)}$. This difference from the GS is essential, since it gives our sampler freedom to explore the restricted hypothesis space efficiently.

To see that the proposed sampler is a valid PCGS, we can view it as the trimmed version of a sampler that samples $\boldsymbol{\theta}_{\mathcal{J}_d(k)}$ from $p(\boldsymbol{\theta}_{\mathcal{J}_d(k)}|\boldsymbol{\theta}_{\sim \mathcal{J}_d(k)}, \boldsymbol{\phi}, \mathbf{x})$ in every substep. In that untrimmed sampler, all the sampling distributions are conditionals associated with the full joint posterior. The trimming is justified because, as is easily verified, all elements of $\mathcal{J}_d(k)$ except k itself are also contained in $\mathcal{J}_d(k+1)$. Thus, they will be sampled again after substep k before being conditioned upon, which makes them eligible for trimming. This also explains why we choose a one-sided neighborhood: within $\boldsymbol{\theta}_{\mathcal{J}_d(k)} = (\theta_k \cdots \theta_{k_{\max}(k)})^{\mathrm{T}}$, θ_k is the only parameter that cannot be trimmed because it is conditioned upon in the next substep. Therefore we choose θ_k (and not some other parameter in $\boldsymbol{\theta}_{\mathcal{J}_d(k)}$) as the parameter associated with the neighborhood $\mathcal{J}_d(k)$.

Often, it will not be possible to marginalize $p(\boldsymbol{\theta}, \boldsymbol{\phi}|\mathbf{x})$ with respect to $\theta_{k+1}, \ldots, \theta_{k_{\max}(k)}$ analytically. In this case, we sample $\boldsymbol{\theta}_{\mathcal{J}_d(k)}$ from $p(\boldsymbol{\theta}_{\mathcal{J}_d(k)}|\boldsymbol{\theta}_{\sim \mathcal{J}_d(k)}, \boldsymbol{\phi}, \mathbf{x})$ and then use the θ_k contained in the sample. Sampling from $p(\boldsymbol{\theta}_{\mathcal{J}_d(k)}|\boldsymbol{\theta}_{\sim \mathcal{J}_d(k)}, \boldsymbol{\phi}, \mathbf{x})$ requires that we evaluate these probabilities for all hypotheses. This can be done efficiently by using the expression

$$p\big(\boldsymbol{\theta}_{\mathcal{J}_d(k)}\big|\boldsymbol{\theta}_{\sim \mathcal{J}_d(k)}, \boldsymbol{\phi}, \mathbf{x}\big) \;\propto\; p(\boldsymbol{\theta}, \boldsymbol{\phi}, \mathbf{x}) = p(\mathbf{x}|\boldsymbol{\theta}, \boldsymbol{\phi})\, p(\boldsymbol{\theta}, \boldsymbol{\phi})\,, \tag{1}$$

where $\boldsymbol{\theta}$ is the composition of $\boldsymbol{\theta}_{\mathcal{J}_d(k)}$ and $\boldsymbol{\theta}_{\sim \mathcal{J}_d(k)}$. Since the right-hand side contains the prior, we can now exploit the deterministic constraint, i.e., the fact that the number of hypotheses $\boldsymbol{\theta}_{\mathcal{J}_d(k)} \in \mathcal{C}_d$ is smaller than M^d: we only need to evaluate (1) for these fewer hypotheses and normalize the result. This convenient way of sampling $\boldsymbol{\theta}_{\mathcal{J}_d(k)}$ is enabled by the discrete nature of $\boldsymbol{\theta}$, and it is the only aspect of our sampler that presupposes a discrete $\boldsymbol{\theta}$. Otherwise, our sampler is valid also for continuous or partly continuous $\boldsymbol{\theta}$, although sampling from $p(\theta_k|\boldsymbol{\theta}_{\sim \mathcal{J}_d(k)}, \boldsymbol{\phi}, \mathbf{x})$ or from $p(\boldsymbol{\theta}_{\mathcal{J}_d(k)}|\boldsymbol{\theta}_{\sim \mathcal{J}_d(k)}, \boldsymbol{\phi}, \mathbf{x})$ may then require additional steps (e.g., a rejection method), depending on the specific distributions. In Section 5, we will consider an example where $\boldsymbol{\theta}$ is partly continuous.

The choice of the neighborhood width d corresponds to a tradeoff: a larger d grants more freedom for jumping in the hypothesis space but also increases the computational complexity.

5. APPLICATION EXAMPLE

We illustrate the proposed sampler by discussing its use for the example of Section 2. This example is relevant to several signal processing applications (e.g., [2]). For reasons that will become clear later, the binary sequence is now denoted by b_k, rather than by θ_k.

Signal model and priors. We consider a binary sequence $b_k \in \{0,1\}$, $k = 1,\dots,K$ that is multiplied by random weights a_k, convolved with a random pulse f_k, and corrupted by additive noise n_k (see [2] for more details). Thus, the observed sequence is given by

$$x_k \;=\; (a_k b_k) * f_k + n_k\,, \qquad k = 1,\dots,K\,. \tag{2}$$

The distance between successive values $b_k = 1$ is at least d. Let $\mathbf{b}$, $\mathbf{a}$, $\mathbf{f}$, and $\mathbf{n}$ denote the length-K vectors corresponding to b_k, a_k, f_k, and n_k, respectively. Using the $K\times K$ diagonal matrix $\mathbf{B} \triangleq \mathrm{diag}(\mathbf{b})$ and the $K \times K$ Toeplitz matrix $\mathbf{F} \triangleq \mathrm{toep}(\mathbf{f})$, we can write (2) as $\mathbf{x} = \mathbf{FBa} + \mathbf{n}$.

The prior of $\mathbf{b}$ is chosen as $p(\mathbf{b}) \propto \mathcal{B}(\mathbf{b};\pi_1)\,\mathrm{I}_{\mathcal{C}}(\mathbf{b})$, where $\mathcal{B}(\mathbf{b};\pi_1) \triangleq \pi_1^{\sum_{k=1}^K b_k}(1-\pi_1)^{K-\sum_{k=1}^K b_k}$ is the product of independent Bernoulli distributions with "1-probability" π_1 (a fixed hyperparameter) and $\mathrm{I}_{\mathcal{C}}(\mathbf{b})$ is an indicator function enforcing the minimum-distance constraint $\mathbf{b} \in \mathcal{C}$. The pulse $\mathbf{f}$ is represented by a basis expansion $\mathbf{f} = \mathbf{H}\boldsymbol{\alpha}$, where $\boldsymbol{\alpha}$ is a length-N random coefficient vector and $\mathbf{H}$ is a known $K \times N$ matrix (chosen as in [2]). The weights $\mathbf{a}$, pulse coefficients $\boldsymbol{\alpha}$, and noise $\mathbf{n}$ are modeled as mutually independent with complex Gaussian priors $p(\mathbf{a}) = \mathcal{CN}(\mathbf{a};\mathbf{0},\sigma_a^2\mathbf{I})$, $p(\boldsymbol{\alpha}) = \mathcal{CN}(\boldsymbol{\alpha};\mathbf{0},\sigma_\alpha^2\mathbf{I})$, and $p(\mathbf{n}|\sigma_n^2) = \mathcal{CN}(\mathbf{n};\mathbf{0},\sigma_n^2\mathbf{I})$. Here, the variances σ_a^2 and σ_α^2 are fixed whereas the noise variance σ_n^2 is treated as a random hyperparameter; it will be estimated jointly with the model parameters using a hierarchical Bayesian model. For the prior of σ_n^2, we use an inverse gamma distribution $\mathcal{IG}(\sigma_n^2;\xi,\eta)$ as suggested in [1], where ξ and η are fixed hyperparameters. Note that the distributions of $\mathbf{a}$, $\boldsymbol{\alpha}$, and σ_n^2 are conjugate priors for the likelihood $p(\mathbf{x}|\mathbf{b},\mathbf{a},\boldsymbol{\alpha},\sigma_n^2) = \mathcal{CN}(\mathbf{x};\mathbf{FBa},\sigma_n^2\mathbf{I})$, with $\mathbf{F} = \mathrm{toep}(\mathbf{H}\boldsymbol{\alpha})$.

We now develop some PCGSs for our application. The goal is to obtain samples from the posterior distribution $p(\mathbf{b},\mathbf{a},\boldsymbol{\alpha},\sigma_n^2|\mathbf{x}) \propto p(\mathbf{x}|\mathbf{b},\mathbf{a},\boldsymbol{\alpha},\sigma_n^2)\,p(\mathbf{b})\,p(\mathbf{a})\,p(\boldsymbol{\alpha})\,p(\sigma_n^2)$. From these samples, the parameters can be detected and estimated. To link our problem to the notation used in Sections 2 and 4, we first set $\boldsymbol{\theta} \triangleq \mathbf{b}$ and $\boldsymbol{\phi} \triangleq (\mathbf{a},\boldsymbol{\alpha},\sigma_n^2)$. A different correspondence will be considered later.

Reference sampler (RS). As a reference method for performance comparison, we first consider the following sampler that does *not* take into account the minimum-distance constraint.

RS:
- For $k = 1,\dots,K$, sample b_k from $p\big(b_k\big|\mathbf{b}_{\sim k},\boldsymbol{\alpha},\sigma_n^2,\mathbf{x}\big)$.
- Sample $\mathbf{a}$ from $p(\mathbf{a}|\mathbf{b},\boldsymbol{\alpha},\sigma_n^2,\mathbf{x})$.
- Sample $\boldsymbol{\alpha}$ from $p(\boldsymbol{\alpha}|\mathbf{b},\mathbf{a},\sigma_n^2,\mathbf{x})$.
- Sample σ_n^2 from $p(\sigma_n^2|\mathbf{b},\mathbf{a},\boldsymbol{\alpha},\mathbf{x})$.

Note that $\mathbf{a}$ and $\boldsymbol{\alpha}$ need not be sampled elementwise because their posteriors are jointly Gaussian. This sampling algorithm (up to minor modifications) was proposed in [5] for a Bernoulli-Gaussian sequence $(b_k a_k)$. Note, however, that our signal model is different because the Bernoulli-Gaussian prior distribution of $(b_k a_k)$ is modified by the minimum-distance constraint. The sampler is a PCGS, not a classical GS, because the sampling distribution for b_k is not a conditional distribution associated with the full joint posterior.

Proposed sampler I (PS-I). We next propose a PCGS that takes into account the minimum-distance constraint. This sampler is obtained by specializing the PCGS presented in Section 4 to the application example considered.

PS-I:
- For $k = 1,\dots,K$, sample b_k from $p\big(b_k\big|\mathbf{b}_{\sim\mathcal{J}_d(k)},\boldsymbol{\alpha},\sigma_n^2,\mathbf{x}\big)$.
- Sample $\mathbf{a}$ from $p(\mathbf{a}|\mathbf{b},\boldsymbol{\alpha},\sigma_n^2,\mathbf{x})$.
- Sample $\boldsymbol{\alpha}$ from $p(\boldsymbol{\alpha}|\mathbf{b},\mathbf{a},\sigma_n^2,\mathbf{x})$.
- Sample σ_n^2 from $p(\sigma_n^2|\mathbf{b},\mathbf{a},\boldsymbol{\alpha},\mathbf{x})$.

As mentioned in Section 2, the width d of the neighborhood $\mathcal{J}_d(k)$ equals the minimum distance $d_{\min}$. The sampling distributions used in this sampler are as follows:

$$p\big(\mathbf{b}_{\mathcal{J}_d(k)}\big|\mathbf{b}_{\sim\mathcal{J}_d(k)},\boldsymbol{\alpha},\sigma_n^2,\mathbf{x}\big) \propto |\boldsymbol{\Sigma}_{\mathbf{a}}|\exp(\boldsymbol{\mu}_{\mathbf{a}}^{\mathrm{H}}\boldsymbol{\Sigma}_{\mathbf{a}}^{-1}\boldsymbol{\mu}_{\mathbf{a}})\,p(\mathbf{b}) \tag{3}$$

$$p(\mathbf{a}|\mathbf{b},\boldsymbol{\alpha},\sigma_n^2,\mathbf{x}) = \mathcal{CN}(\mathbf{a};\boldsymbol{\mu}_{\mathbf{a}},\boldsymbol{\Sigma}_{\mathbf{a}})$$

$$p(\boldsymbol{\alpha}|\mathbf{b},\mathbf{a},\sigma_n^2,\mathbf{x}) = \mathcal{CN}(\boldsymbol{\alpha};\boldsymbol{\mu}_{\boldsymbol{\alpha}},\boldsymbol{\Sigma}_{\boldsymbol{\alpha}}) \tag{4}$$

$$p(\sigma_n^2|\mathbf{b},\mathbf{a},\boldsymbol{\alpha},\mathbf{x}) = \mathcal{IG}\big(\sigma_n^2;\xi+K,\eta+\|\mathbf{x}-\mathbf{FBa}\|^2\big)\,, \tag{5}$$

with

$$\boldsymbol{\mu}_{\mathbf{a}} = \frac{1}{\sigma_n^2}\boldsymbol{\Sigma}_{\mathbf{a}}\mathbf{B}^{\mathrm{H}}\mathbf{F}^{\mathrm{H}}\mathbf{x}\,, \qquad \boldsymbol{\Sigma}_{\mathbf{a}} = \left(\frac{1}{\sigma_n^2}\mathbf{B}^{\mathrm{H}}\mathbf{F}^{\mathrm{H}}\mathbf{FB} + \frac{1}{\sigma_a^2}\mathbf{I}\right)^{-1}$$

$$\boldsymbol{\mu}_{\boldsymbol{\alpha}} = \frac{1}{\sigma_n^2}\boldsymbol{\Sigma}_{\boldsymbol{\alpha}}\mathbf{H}^{\mathrm{H}}\mathrm{toep}(\mathbf{Ba})^{\mathrm{H}}\mathbf{x}$$

$$\boldsymbol{\Sigma}_{\boldsymbol{\alpha}} = \left(\frac{1}{\sigma_n^2}\mathbf{H}^{\mathrm{H}}\mathrm{toep}(\mathbf{Ba})^{\mathrm{H}}\mathrm{toep}(\mathbf{Ba})\mathbf{H} + \frac{1}{\sigma_\alpha^2}\mathbf{I}\right)^{-1}.$$

As explained in Section 4, b_k is sampled by evaluating (3) for all hypotheses $\mathbf{b}_{\mathcal{J}_d(k)} \in \mathcal{C}_d$. There are $d+1$ such hypotheses, namely one with no 1 and d with one 1. After sampling $\mathbf{b}_{\mathcal{J}_d(k)}$ from the distribution in (3), we use the b_k contained in the sample.

Proposed sampler II (PS-II). An alternative PCGS for our problem can be obtained by setting $\boldsymbol{\theta} \triangleq (\mathbf{b},\mathbf{a})$ and $\boldsymbol{\phi} \triangleq (\boldsymbol{\alpha},\sigma_n^2)$. Thus, $\boldsymbol{\theta}$ contains the continuous parameters a_k besides the discrete parameters b_k. Sampling $\theta_k = (b_k,a_k)$ can be done by first sampling b_k, then a_k conditioned on b_k. The resulting sampler, which also complies with the PCGS of Section 4, is stated as follows.

PS-II:
- For $k = 1,\dots,K$,
  - sample b_k from $p\big(b_k\big|\mathbf{b}_{\sim\mathcal{J}_d(k)},\mathbf{a}_{\sim\mathcal{J}_d(k)},\boldsymbol{\alpha},\sigma_n^2,\mathbf{x}\big)$;
  - sample a_k from $p\big(a_k\big|b_k,\mathbf{b}_{\sim\mathcal{J}_d(k)},\mathbf{a}_{\sim\mathcal{J}_d(k)},\boldsymbol{\alpha},\sigma_n^2,\mathbf{x}\big)$.
- Sample $\boldsymbol{\alpha}$ from $p(\boldsymbol{\alpha}|\mathbf{b},\mathbf{a},\sigma_n^2,\mathbf{x})$.
- Sample σ_n^2 from $p(\sigma_n^2|\mathbf{b},\mathbf{a},\boldsymbol{\alpha},\mathbf{x})$.

This is the sampler we previously proposed in [2], up to slight modifications. The sampling distributions now are

$$p\big(\mathbf{b}_{\mathcal{J}_d(k)}\big|\mathbf{b}_{\sim\mathcal{J}_d(k)},\mathbf{a}_{\sim\mathcal{J}_d(k)},\boldsymbol{\alpha},\sigma_n^2,\mathbf{x}\big) \;\propto\; \sigma^2 e^{|\mu|^2/\sigma^2}p(\mathbf{b})$$

$$p\big(a_k\big|b_k,\mathbf{b}_{\sim\mathcal{J}_d(k)},\mathbf{a}_{\sim\mathcal{J}_d(k)},\boldsymbol{\alpha},\sigma_n^2,\mathbf{x}\big) \;=\; \mathcal{CN}(a_k;\mu,\sigma)\,,$$

with

$$\mu \;=\; \frac{\sigma^2\mathbf{b}_{\mathcal{J}_d(k)}^{\mathrm{H}}\mathbf{F}_{\mathcal{J}_d(k)}^{\mathrm{H}}\big(\mathbf{x}-\mathbf{F}_{\sim\mathcal{J}_d(k)}\mathbf{B}^{\sim\mathcal{J}_d(k)}\mathbf{a}_{\sim\mathcal{J}_d(k)}\big)}{\sigma_n^2}$$

$$\sigma^2 = \left(\frac{\mathbf{b}_{\mathcal{J}_d(k)}^{\mathrm{H}}\mathbf{F}_{\mathcal{J}_d(k)}^{\mathrm{H}}\mathbf{F}_{\mathcal{J}_d(k)}\mathbf{b}_{\mathcal{J}_d(k)}}{\sigma_n^2} + \frac{1}{\sigma_a^2}\right)^{-1}.$$

Here, $\mathbf{F}_{\mathcal{J}_d(k)}$ consists of the columns of $\mathbf{F}$ indexed by $\mathcal{J}_d(k)$, $\mathbf{F}_{\sim\mathcal{J}_d(k)}$ denotes $\mathbf{F}$ without these columns, and $\mathbf{B}^{\sim\mathcal{J}_d(k)} \triangleq \mathrm{diag}(\mathbf{b}_{\sim\mathcal{J}_d(k)})$. The sampling distributions for $\boldsymbol{\alpha}$ and σ_n^2 equal those in (4) and (5).

Compared to PS-I, the alternative sampler PS-II is "less collapsed" since its sampling distributions are conditioned on more parameters. Therefore, the gain in convergence rate is slightly smaller than for PS-I. Nevertheless, PS-II is significantly less complex than PS-I because it does not require inversion of $\boldsymbol{\Sigma}_{\mathbf{a}}$.

Reducing complexity. We can speed up both PS-I and PS-II as follows. First, after generating a 1 when sampling b_k, the rest of $\mathbf{b}_{\mathcal{J}_d(k)}$ can be set to 0 and the next substeps until $k_{\max}(k)$ can be skipped. This is due to the minimum-distance constraint represented by the factor $p(\mathbf{b})$ in (3). It is also valid for PS-II, because $b_l = 0$ means that a_l is irrelevant. Second, if $b_k = 0$ before b_k is sampled

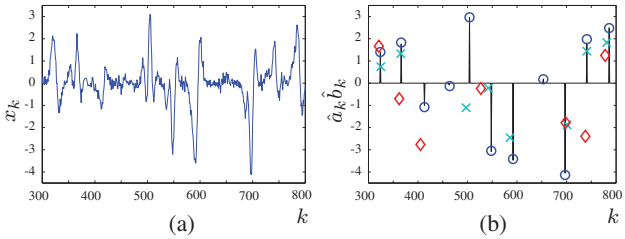

Fig. 1. Results of detection/estimation of $\mathrm{diag}(\mathbf{b})\mathbf{a}$: (a) signal $\mathbf{x}$, (b) estimates. In (b), circles represent PS-I and PS-II results after 10 iterations (they coincide), crosses represent RS-A results, and diamonds represent RS-B results, both after 200 iterations. The vertical lines indicate the true $\mathrm{diag}(\mathbf{b})\mathbf{a}$. Real parts are shown.

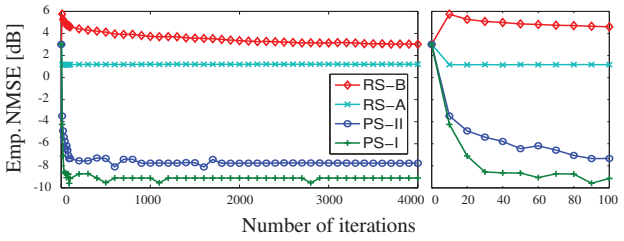

Fig. 2. Empirical NMSE of the estimate $\mathrm{diag}(\hat{\mathbf{b}})\hat{\mathbf{a}}$ versus the number of iterations. Left: all 4000 iterations, right: the first 100 iterations.

and the sampling produces again $b_k = 0$, only one hypothesis is new; thus, for the other hypotheses, the probabilities calculated in the previous substep can be reused.

The most complex part of RS and PS-I is the inversion of $\mathbf{\Sigma_a}$. This problem can be mitigated by calculating (3) recursively, as shown in [5]. An expression equivalent to (3) is

$$p\big(\mathbf{b}_{\mathcal{J}_d(k)}\big|\mathbf{b}_{\sim\mathcal{J}_d(k)}, \boldsymbol{\alpha}, \sigma_n^2, \mathbf{x}\big) \propto \frac{\exp(-\mathbf{x}^{\mathrm{H}}\mathbf{\Sigma}_{\mathbf{b}}^{-1}\mathbf{x})\, p(\mathbf{b})}{|\mathbf{\Sigma}_{\mathbf{b}}|} \;,$$

with $\mathbf{\Sigma_b} = \sigma_a^2\mathbf{F}\mathbf{B}\mathbf{F}^{\mathrm{H}} + \sigma_n^2\mathbf{I}$. In [5], an efficient update of $\mathbf{\Sigma_b}$ and $\exp(-\mathbf{x}^{\mathrm{H}}\mathbf{\Sigma}_{\mathbf{b}}^{-1}\mathbf{x})\, p(\mathbf{b})/|\mathbf{\Sigma_b}|$ is described (with slight modifications due to the different signal model) for the special case that only one b_k is changed. This recursive computation can be used in PS-I to calculate the probabilities of the hypotheses $\mathbf{b}_{\mathcal{J}_d(k)}$. First, the probability of the all-zero hypothesis is calculated as an update of the probability of the previous sample of $\mathbf{b}_{\mathcal{J}_d(k)}$. Since the latter may contain at most one 1, it differs from the all-zero hypothesis in at most one position. Then, the probabilities of all other hypotheses are calculated as updates of the probability of the all-zero hypothesis. Each of them contains one 1, and thus differs from the all-zero hypothesis in one position. In spite of this improvement of RS and PS-I, PS-II is still significantly less complex because the inversion of $\mathbf{\Sigma_a}$ or $\mathbf{\Sigma_b}$ is avoided altogether.

6. SIMULATION RESULTS

For the minimum-distance constrained problem considered in Section 5, we compare the performance of the proposed methods PS-I and PS-II with that of the reference method RS. We consider two versions of RS. The first, denoted RS-A, is based on the true signal model, in which $p(\mathbf{b}) \propto \mathcal{B}(\mathbf{b}; \pi_1)\, \mathrm{I}_{\mathcal{C}}(\mathbf{b})$. Since RS behaves like a classical GS with respect to dependencies within $\mathbf{b}$, it is not well suited to this restricted prior. We thus consider also a second RS version, denoted RS-B, which is based on a signal model that is more compatible with the algorithm, namely, a Bernoulli-Gaussian model (at the cost of a model mismatch). We note that RS was originally designed in [5] for Bernoulli-Gaussian sequences. We thus interpret realizations from $p(\mathbf{b})$ as Bernoulli sequences from $\tilde{p}(\mathbf{b}) = \mathcal{B}(\mathbf{b}; \tilde{\pi}_1)$, with $\tilde{\pi}_1^{-1} = d + \pi_1^{-1}$. (This way, the average distance between 1's in sequences drawn from $p(\mathbf{b})$ and $\tilde{p}(\mathbf{b})$ is identical.) Consistently replacing $p(\mathbf{b})$ by $\tilde{p}(\mathbf{b})$ and π_1 by $\tilde{\pi}_1$ in RS-A, we obtain RS-B.

We generated 170 realizations of $\mathbf{x}$ from parameters randomly drawn according to the priors given in Section 5, using $K = 1024$, $N = 5$, $d = 40$, $\pi_1 = 0.15$, $\sigma_a^2 = 10$, and $\sigma_\alpha^2 = 2.4$ (these are known hyperparameters fixed *a priori*), as well as $\xi = 11$ and $\eta = 0.5$ (to provide a noninformative prior for σ_n^2). For each realization of $\mathbf{x}$, we generated a Markov chain according to each of the four sampler algorithms. Detection and estimation were performed as in [2]. The result of one such simulation run (corresponding to one realization of $\mathbf{x}$) is shown in Fig. 1. The detected/estimated sequences $\mathrm{diag}(\hat{\mathbf{b}})\hat{\mathbf{a}}$ of both PS-I and PS-II coincide with the true values after only 10 iterations. The results of RS-A and RS-B after 200 iterations are significantly worse.

To assess the convergence rates, Fig. 2 shows the empirical normalized mean-square error (NMSE) of detection/estimation of $\mathbf{b}$ and $\mathbf{a}$ versus the number of iterations. The empirical NMSE is defined as the average (over 170 realizations) of $\|\mathrm{diag}(\hat{\mathbf{b}})\hat{\mathbf{a}} - \mathrm{diag}(\mathbf{b})\mathbf{a}\|^2$ normalized by the average of $\|\mathrm{diag}(\mathbf{b})\mathbf{a}\|^2$. The number of iterations equals the total length of the Markov chain. Out of each chain, the last 25% of the iterations were used for detection/estimation. It can be seen that RS-A and RS-B do not yield satisfactory results within 4000 iterations, since the NMSE is still above 0 dB. Because RS-A and RS-B treat the constrained sequence $\mathbf{b}$ as a classical GS would, they can be expected to converge after a much larger number of iterations. By contrast, PS-I and PS-II successfully exploit their knowledge of the constraint, since they reach an NMSE of below -7 dB within the first 100 iterations.

7. CONCLUSION

We studied Bayesian detection/classification of discrete random parameters that are subject to a deterministic local constraint. Such a constraint implies strong dependencies for which the classical Gibbs sampler exhibits slow convergence. We demonstrated that this problem can be overcome by a new Monte Carlo method based on the recently introduced *partially collapsed Gibbs sampler* principle. The proposed method was successfully applied to a detection-estimation problem involving nonuniformly spaced binary pulses with a known minimum distance. Our results are potentially useful for many signal processing applications including signal segmentation, layer detection, and electromyography.

8. REFERENCES

[1] N. Dobigeon, J.-Y. Tourneret, and M. Davy, "Joint segmentation of piecewise constant autoregressive processes by using a hierarchical model and a Bayesian sampling approach," *IEEE Trans. Signal Processing*, vol. 55, pp. 1251–1263, Apr. 2007.

[2] G. Kail, C. Novak, B. Hofer, and F. Hlawatsch, "A blind Monte Carlo detection-estimation method for optical coherence tomography," in *Proc. IEEE ICASSP-2009*, Taipei, Taiwan, pp. 493–496, April 2009.

[3] D. Ge, E. Le Carpentier, and D. Farina, "Unsupervised Bayesian decomposition of multi-unit EMG recordings using Tabu search," *IEEE Trans. Biomed. Eng.*, vol. 56, pp. 1–9, Dec. 2009.

[4] C. P. Robert and G. Casella, *Monte Carlo Statistical Methods*. New York: Springer, 2004.

[5] D. Ge, J. Idier, and E. Le Carpentier, "A new algorithm for blind Bernoulli-Gaussian deconvolution," in *Proc. EUSIPCO-08*, Lausanne, Switzerland, Aug. 2008.

[6] D. A. van Dyk and T. Park, "Partially collapsed Gibbs samplers: Theory and methods," *J. Am. Statist. Assoc.*, vol. 103, pp. 790–796, June 2008.

[7] T. Park and D. A. van Dyk, "Partially collapsed Gibbs samplers: Illustrations and applications," *J. Comput. Graph. Statist.*, vol. 18, pp. 283–305, June 2009.