A PARTIALLY COLLAPSED GIBBS SAMPLER FOR PARAMETERS WITH LOCAL CONSTRAINTS

Georg Kail, Jean-Yves Tourneret, Franz Hlawatsch, and Nicolas Dobigeon

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 \( \theta = (\theta_1, \ldots \theta_K)^T \). It is assumed that the parameters \( \theta_k \) 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.

2. PROBLEM STATEMENT

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

We now define the class of “deterministic local constraints” we will consider. A deterministic constraint reduces the set of possible hypotheses for \( \theta \) in that \( \theta \in C \) rather than \( \theta \in A^K \), with a “constraint set” \( C \subset A^K \). The prior of \( \theta \) then satisfies \( p(\theta) = 0 \) for \( \theta \notin C \). The deterministic constraint can be considered local if it causes strong dependencies between parameters \( \theta_k \) that are close to each other but not between distant parameters. Thus, the effect of the local constraint on the prior \( p(\theta) \) can be described not only in terms of the entire sequence \( \theta \) 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 \( J_d(k) \triangleq \{k, \ldots, k+d-1\} \) denote some right-hand neighborhood of \( k \) and \( \theta_{J_d(k)} \triangleq (\theta_k, \ldots \theta_{k+d-1})^T \) the corresponding subvector of \( \theta \). In general, \( \theta_{J_d(k)} \in A^d \) and thus there would be \( M^d \) hypotheses \( \theta_{J_d(k)} \) for any given \( k \). However, due to the deterministic local constraint, the set of hypotheses is reduced to a smaller set \( C_d \), i.e., we know that \( \theta_{J_d(k)} \in C_d \) and thus

\[
p(\theta) = 0 \quad \text{for all } \theta \text{ such that } \theta_{J_d(k)} \notin C_d,
\]

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

Example. Consider binary parameters \( \theta_k \in \{0, 1\} \) (i.e., \( M = 2 \)) with a given minimum distance \( d_{\text{min}} \) between successive 1’s in the sequence. That is, \( \theta \) for which two elements \( \theta_k = 1 \) and \( \theta_{k+1} = 1 \) are closer than \( d_{\text{min}} \), i.e., \( |k - k'| < d_{\text{min}} \), has zero probability. Consider a subvector \( \theta_{J_d(k)} = (\theta_k, \ldots \theta_{k+d-1})^T \) for a fixed \( k \). For convenience, we choose \( d = d_{\text{min}} \). Without the minimum-distance constraint, there would be \( 2^d \) possible hypotheses for \( \theta_{J_d(k)} \). However, due to the constraint, all \( \theta \) for which \( \theta_{J_d(k)} \) contains more than one 1 have zero probability. Thus, there are \( d+1 \) admissible hypotheses \( \theta_{J_d(k)} \in C_d \), 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 \( \theta \) be a vector of \( K \) parameters, and let \( \theta_{-k} \) denote \( \theta \) without the \( k \)th element \( \theta_k \) (a generalization to the case where \( \theta_k \) is itself a vector is possible). To obtain samples from \( p(\theta) \) (which corresponds to \( p(\theta, \phi|x) \) in our problem), the GS iteratively generates samples of each \( \theta_k \) from \( p(\theta_k | \theta_{-k}) \) in an arbitrary order. This strategy is known to converge to the target distribution \( p(\theta) \). After convergence, \( K \) such sampling substeps produce a new sample \( \theta \) from \( p(\theta) \). A known weakness of the GS is that dependencies between the \( \theta_k \) 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 \( \theta_k \) in substep \( k \), some other parameters may be sampled along with \( \theta_k \) 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 \( p(\theta) \), but rather conditional distributions associated with certain marginal distributions of \( p(\theta) \).

- **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 posterior 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 \( p(\theta) \) 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 \( \theta_k \) 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 \( \theta = (\theta_1, \ldots, \theta_K) \) and an unconstrained parameter vector \( \phi = (\phi_1, \ldots, \phi_L) \). The task is to sample from the posterior \( p(\theta, \phi|x) \). In order to extend the definition of the neighborhoods \( J_{\theta}(k) \) to all \( k \in \{1, \ldots, K\} \), we redefine \( J_{\theta}(k) \) as \( \{k, \ldots, k_{\text{max}}(k)\} \) with \( k_{\text{max}}(k) = \min\{k+d-1, K\} \). 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 \( \theta_k \) from \( p(\theta_k | \theta_{-k}, J_{\theta}(k), \phi, x) \).
- For \( l = 1, \ldots, L \), sample \( \phi_l \) from \( p(\phi_l | \theta, \phi_{-l}, x) \).

Note that the sampling distribution for \( \theta_k \) is a conditional distribution associated with \( p(\theta_k | \theta_{-k}, J_{\theta}(k), \phi, x) \), which is the joint posterior \( p(\theta, \phi|x) \) marginalized with respect to all parameters in \( J_{\theta}(k) \) except \( \theta_k \) (i.e., with respect to \( \theta_{k+1}, \ldots, \theta_{\text{max}}(k) \)). Accordingly, \( \theta_{J_{\theta}(k)} \) is not contained in the condition for \( \theta_k \), which is therefore sampled regardless of the previous sample of \( \theta_{J_{\theta}(k)} \). 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 \( \theta_{J_{\theta}(k)} \) from \( p(\theta_{J_{\theta}(k)} | \theta_{-J_{\theta}(k)}, \phi, x) \) 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 \( J_{\theta}(k) \) except itself are also contained in \( J_{\theta}(k+1) \). 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 \( \theta_{J_{\theta}(k)} = (\theta_1, \ldots, \theta_{\text{max}}(k)) \), \( \theta_k \) is the only parameter that cannot be trimmed because it is conditioned upon in the next substep. Therefore we choose \( \theta_k \) (and not some other parameter in \( \theta_{J_{\theta}(k)} \)) as the parameter associated with the neighborhood \( J_{\theta}(k) \).

Often, it will not be possible to marginalize \( p(\theta, \phi|x) \) with respect to \( \theta_{k+1}, \ldots, \theta_{\text{max}}(k) \) analytically. In this case, we sample \( \theta_{J_{\theta}(k)} \) from \( p(\theta_{J_{\theta}(k)} | \theta_{-J_{\theta}(k)}, \phi, x) \) and then use the \( \theta_k \) contained in the sample. Sampling from \( p(\theta_{J_{\theta}(k)} | \theta_{-J_{\theta}(k)}, \phi, x) \) requires that we evaluate these probability distributions for all hypotheses. This can be done efficiently by using the expression

\[
p(\theta_{J_{\theta}(k)} | \theta_{-J_{\theta}(k)}, \phi, x) \propto p(\theta, \phi, x) = p(x | \theta, \phi) p(\theta, \phi),
\]

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

### 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 \( \theta_k \).
Signal model and priors. We consider a binary sequence $b_k \in \{0, 1\}, k = 1, \ldots, K$ 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

$$x_k = (a_k b_k) f_k + n_k, \quad k = 1, \ldots, K. \quad (2)$$

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

The prior of $b$ is chosen as $p(b) \propto B(b; \pi_1)I_c(b)$, where $B(b; \pi_1)$ is the product of independent Bernoulli distributions with “1-probability” $\pi_1$ (a fixed hyperparameter) and $I_c(b)$ is an indicator function enforcing the minimum-distance constraint $b \in \mathcal{C}$. The pulse $f$ is represented by a basis expansion $f = Ha$, where $a$ is a length-$N$ random coefficient vector and $H$ is a known $K \times N$ matrix (chosen as in [2]). The weights $a$, pulse coefficients $\alpha$, and noise $n$ are modeled as mutually independent with complex Gaussian priors $p(a) = \mathcal{CN}(a; 0, \sigma^2\beta_I)$, $p(\alpha) = \mathcal{CN}(\alpha; 0, \sigma^2\alpha_I)$, and $p(n|\sigma^2_n) = \mathcal{CN}(n; 0, \sigma^2_n)$. Here, the variances $\sigma^2\alpha$ and $\sigma^2\beta$ are fixed whereas the noise variance $\sigma^2_n$ is treated as a random hyperparameter; it will be estimated jointly with the continuous parameters besides the discrete parameter $b_k$. The sampler is a PCGS, which is stated as follows.

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 $k = 1, \ldots, K$, sample $b_k$ from $p(b_k|b_{-k}, \alpha, \sigma^2_n, x)$.
- Sample $a$ from $p(a|b, \alpha, \sigma^2_n, x)$.
- Sample $\alpha$ from $p(\alpha|b, a, \sigma^2_n, x)$.
- Sample $\sigma^2_n$ from $p(\sigma^2_n|b, a, \alpha, x)$.

Note that $\alpha$ and $\sigma$ 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 ($b_{-k}$). Note, however, that our signal model is different because the Bernoulli-Gaussian prior distribution of $(b_{-k})$ 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 $k = 1, \ldots, K$, sample $b_k$ from $p(b_k|b_{-k}, \alpha, \sigma^2_n, x)$.
- Sample $a$ from $p(a|b, \alpha, \sigma^2_n, x)$.
- Sample $\alpha$ from $p(\alpha|b, a, \sigma^2_n, x)$.
- Sample $\sigma^2_n$ from $p(\sigma^2_n|b, a, \alpha, x)$.

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

$$p(b_j(b_l|b_{-j}(k), \alpha, \sigma^2_n, x) \propto |\Sigma_n| \exp(\mu_n^\text{T} \Sigma_n^{-1} \mu_n) \cdot p(b) \quad (3)$$

$$p(a|b, \alpha, \sigma^2_n, x) = \mathcal{CN}(a; \mu_n, \Sigma_n) \quad (4)$$

$$p(\alpha|b, a, \sigma^2_n, x) = \mathcal{IG}(\sigma^2_n; \xi + K, \eta + |x - FBa|^2) \quad (5)$$

As explained in Section 4, $b_k$ is sampled by evaluating (3) for all hypotheses $b_j(k) \in \mathcal{C}_d$. There are $d + 1$ such hypotheses, namely, one with no 1 and $d$ with one 1. After sampling $b_{-j}(k)$, $d_{\text{min}}$ is chosen as

$$d_{\text{min}} = \left\{ \frac{1}{\sigma^2_n} \Sigma_n B^\text{T} F^\text{H} x, \quad \Sigma_n = \left( \frac{1}{\sigma^2_n} B^\text{T} F^\text{H} F B + \frac{1}{\sigma^2_n} I \right)^{-1} \right\}.$$
Fig. 1. Results of detection/estimation of diag(\(b\))a: (a) signal \(x\), (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 diag(\(b\))a. Real parts are shown.

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 \(\Sigma_a^*\). This problem can be mitigated by calculating (3) recursively, as shown in [5]. An expression equivalent to (3) is

\[
p(b_{\tilde{J}_d(k)} | b_{\tilde{J}_d(k)}, \alpha, \sigma^2_a, x) \propto \exp(-x^H \Sigma_a^{-1} x) p(b) |\Sigma_b|^{-1},
\]

with \(\Sigma_b = \sigma^2_b FFB^H + \sigma^2_d I\). In [5], an efficient update of \(\Sigma_b\) and \(\exp(-x^H \Sigma_a^{-1} x) p(b)/|\Sigma_b|\) is derived (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 \(b_{\tilde{J}_d(k)}\). First, the probability of the all-zero hypothesis is calculated as an update of the probability of the previous sample of \(b_{\tilde{J}_d(k)}\). 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 \(\Sigma_a^*\) or \(\Sigma_b\) 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 \(p(b) \propto B(b; \pi_1) I_c(b)\). Since RS behaves like a classical GS with respect to dependencies within \(b\), 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 \(p(b)\) as Bernoulli sequences from \(p(b) = B(b; \pi_1)\), with \(\pi_1^+ = d + \pi_1^-\). (This way, the average distance between 1’s in sequences drawn from \(p(b)\) and \(p(b)\) is identical.) Consistently replacing \(p(b)\) by \(p(b)\) and \(\pi_1\) by \(\pi_1\) in RS-A, we obtain RS-B.

We generated 170 realizations of \(x\) from parameters randomly drawn according to the priors given in Section 5, using \(K = 1024\), \(N = 5\), \(d = 40\), \(\pi_1 = 0.15\), \(\sigma^2_a = 10\), and \(\sigma^2_d = 2.4\) (these are known hyperparameters fixed a priori), as well as \(\xi = 11\) and \(\eta = 0.5\) (to provide a noninformative prior for \(\sigma^2_d\)). For each realization of \(x\), 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 \(x\)) is shown in Fig. 1. The detected/estimated sequences diag(\(b\))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 \(b\) and \(a\) versus the number of iterations. The empirical NMSE is defined as the average (over 170 realizations) of \(||diag(\(b\))a - diag(\(b\))a||^2\) normalized by the average of \(||diag(\(b\))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 \(b\) as a classical GS, 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


