Information theory tools to rank MCMC algorithms on probabilistic graphical models

Firas Hamze, Jean-Noel Rivasseau and Nando de Freitas
Computer Science Department
University of British Columbia
Email: {fhamze,jnriva,nando}@cs.ubc.ca

***Abstract*— We propose efficient MCMC tree samplers for random fields and factor graphs. Our tree sampling approach combines elements of Monte Carlo simulation as well as exact belief propagation. It requires that the graph be partitioned into trees first. The partition can be generated by hand or automatically using a greedy graph algorithm. The tree partitions allow us to perform exact inference on each tree. This enables us to implement efficient Rao-Blackwellised blocked Gibbs samplers, where each tree is sampled by conditioning on the other trees. We use information theory tools to rank MCMC algorithms corresponding to different partitioning schemes.**

I. INTRODUCTION

Undirected probabilistic graphical models play an important role in spatial statistics, artificial intelligence and computer vision (Besag 1986, Besag 1974, Kumar and Hebert 2003, Li 2001, McCallum, Rohanimanesh and Sutton 2003). Existing MCMC algorithms for undirected models tend to be slow and fail to exploit the structural properties of the undirected graphical model (Geman and Geman 1984, Swendsen and Wang 1987). In contrast, variational approximation schemes (Yedidia, Freeman and Weiss 2000, Wainwright, Jaakkola and Willsky 2003) do exploit structural properties, but may often fail to converge.

In this paper, we extend the Rao-Blackwellised MCMC algorithm for undirected probabilistic graphical models that we proposed in (Hamze and de Freitas 2004) to factor graphs. This algorithm exploits the property that graphical models can be split into disjoint trees as shown in Figure 1. We use greedy search algorithms to find these partitions (see (Rivasseau 2005) for details), but in some simple cases such partitions are obvious. For example, a Markov random field (MRF) can be split into two disjoint trees. By carrying out exact inference on each tree, it is possible to sample half of the MRF nodes in a single MCMC step. Our theorem will show that this tree sampling method outperforms simpler MCMC schemes. This theoretical result is mirrored by our numerical examples.

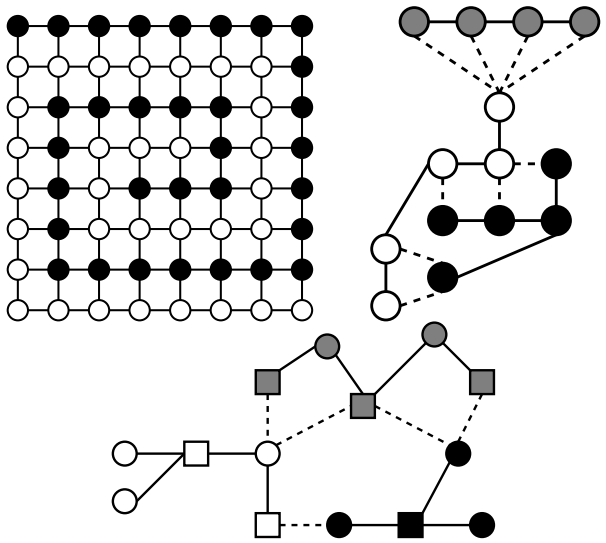

Fig. 1. Tree partitions of two Markov random fields (top) and a factor graph (bottom) generated automatically. One can carry out exact inference on each tree. We adopt a Rao-Blacwellised Gibbs sampler where one samples one-tree-at-a-time efficiently.

II. TREE SAMPLING FOR MRFS

A. MODEL SPECIFICATION

For simplicity of presentation, we focus on the square-lattice MRF but remind the reader that our algorithms apply to other graphs, such as factor graphs and conditional random fields (CRFs). We specify the MRF distribution on a graph $\mathcal{G}(\mathcal{V}, \mathcal{E})$, with edges $\mathcal{E}$ and N nodes $\mathcal{V}$ as shown in Figure 2 left. The clear nodes correspond to the unknown discrete states $\mathbf{x} \in \{1, \ldots, n_x\}$, while the attached black nodes represent discrete observations $\mathbf{y} \in \{1, \ldots, n_x\}$ (they could also be Gaussian). According to this graph, the MRF distribution factorizes into a product of local Markov positive potentials:

$$P(\mathbf{x}_{1:n}, \mathbf{y}_{1:n}) = \frac{1}{Z} \prod_{i \in \mathcal{V}} \phi(\mathbf{x}_i, \mathbf{y}_i) \prod_{(i,j) \in \mathcal{E}} \psi(\mathbf{x}_i, \mathbf{x}_j)$$

where Z is the partition function, $\phi(\cdot)$ denotes the observation potentials and $\psi(\cdot)$ denotes the pair-wise interaction potentials. Our goal is to estimate the marginal posterior distributions (beliefs) $p(\mathbf{x}_i|\mathbf{y}_{1:N})$ and expectations of functions over these distributions.

As shown in Figure 2, *an MRF can be partitioned into two disjoint trees*. The loops in the MRF model cause it to be analytically intractable. However, belief propagation on each of the two trees is a tractable problem. This idea leads naturally to an algorithm that combines analytical and sampling steps. In particular if we have a sample of one of the trees, we can use belief propagation to compute the exact distribution of the other tree by conditioning on this sample. The algorithm

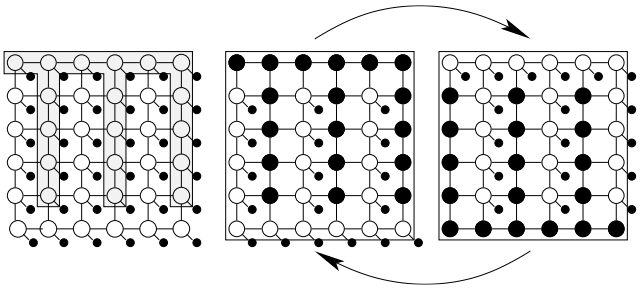

Fig. 2. At left, illustration of a partitioned MRF; nodes in the shaded and white regions are Δ_1, Δ_2 respectively, with the small black circles representing observations. At right, depiction of the two-stage sampler; *sampled* values are *large* black circles. Conditioned on Δ_1, the variables in Δ_2 form a tree. Using this two-stage scheme, Rao-Blackwellised estimators are guaranteed to outperform naive ones.

therefore alternates between computing the trees and sampling them, as shown in Figure 2. Drawing samples in blocks (trees in this case) is well known to have benefits over algorithms that sample one node at-a-time. In Section III we will prove the domination of estimators using this tree sampling algorithm in comparison to other sampling schemes. Before this, we present the algorithm in more detail.

B. TREE SAMPLING ALGORITHM

Consider the 5x5 MRF graph shown in Figure 2 at left. We have partitioned the nodes into two disjoint sets. Denote the set of indices of the shaded nodes as Δ_1 and those of the unshaded nodes as Δ_2, where of course, $\Delta_1 \cup \Delta_2 = I$, the set of all node indices. Let the variables indexed by these nodes be $X_{\Delta_1} \stackrel{\triangle}{=} \{X_j | j \in \Delta_1\}$ and $X_{\Delta_2} \stackrel{\triangle}{=} \{X_j | j \in \Delta_2\}$. If we can sample from the conditional distributions:

$$\begin{array}{l} p(\mathbf{x}_{\Delta_1} | \mathbf{x}_{\Delta_2}, \mathbf{y}) \\ p(\mathbf{x}_{\Delta_2} | \mathbf{x}_{\Delta_1}, \mathbf{y}) \end{array} \tag{1}$$

then we have a two-stage Gibbs sampler called *data augmentation*, which has powerful structural properties that the general Gibbs sampler lacks (Liu 2001, Robert and Casella 1999). In particular, the two Markov chains in data augmentation exhibit the *interleaving property*: $\mathbf{x}_{\Delta_2}^{(t)}$ is independent of $\mathbf{x}_{\Delta_2}^{(t-1)}$ given $\mathbf{x}_{\Delta_1}^{(t)}$; and $(\mathbf{x}_{\Delta_2}^{(t)}, \mathbf{x}_{\Delta_2}^{(t-1)})$ and $(\mathbf{x}_{\Delta_2}^{(t)}, \mathbf{x}_{\Delta_2}^{(t)})$ are identically distributed under stationarity.

Conditioned on set X_{Δ_2}, the variables X_{Δ_1} form an acyclic graph whose marginals are readily computed using *belief propagation* (Pearl 1987). This enables us to sample efficiently from the joint conditionals in (1) using the *Forward Filtering/Backward Sampling* algorithm (FF/BS) (Carter and Kohn 1994, Wilkinson and Yeung 2001). The details of our extension of FF/BS to factor graphs appear in (Rivasseau 2005). The sampling cycle is graphically shown in Figure 2, which makes it explicit that sampled values act as additional "evidence" to the complementary graph. The algorithm is shown in pseudocode in Figure 3.

In the pseudocode, we are adopting *Rao-Blackwellised* es-

Tree sampling

Initialize

- Set all sums $S_i = 0$
- Partition the set of all nodes I into disjoint, tree-connected sets $\{\Delta_1, \Delta_2\}$
- for $i \in I$, randomly initialize $X_i^{(0)}$

for $t = 1 \ldots T$

- for $i \in \Delta_1$
  - Apply belief propagation to compute the *smoothing* densities $p(\mathbf{x}_i | \mathbf{x}_{\Delta_2}^{(t-1)}, \mathbf{y})$, treating the samples $\Delta_2^{(t-1)}$ as evidence.
  - Compute the expectations for the Rao-Blackwellised estimator $\mathbb{E}[h(X_i) | \mathbf{x}_{\Delta_2}^{(t-1)}, \mathbf{y}] = \sum_{X_i} h(X_i) p(\mathbf{x}_i | \mathbf{x}_{\Delta_2}^{(t-1)}, \mathbf{y})$
  - Set $S_i \leftarrow S_i + \mathbb{E}[h(X_i) | \mathbf{x}_{\Delta_2}^{(t-1)}, \mathbf{y}]$
  - Sample $X_{\Delta_1}^{(t)} \sim p(\mathbf{x}_{\Delta_1} | \mathbf{x}_{\Delta_2}^{(t-1)}, \mathbf{y})$ using *Forward filtering / Backward sampling*.
- for $i \in \Delta_2$
  - Apply belief propagation to compute the *smoothing* densities $p(\mathbf{x}_i | \mathbf{x}_{\Delta_1}^{(t)}, \mathbf{y})$, treating the samples $\Delta_1^{(t)}$ as evidence.
  - Compute the expectations for the Rao-Blackwellised estimator $\mathbb{E}[h(X_i) | \mathbf{x}_{\Delta_1}^{(t)}, \mathbf{y}] = \sum_{X_i} h(X_i) p(\mathbf{x}_i | \mathbf{x}_{\Delta_1}^{(t)}, \mathbf{y})$
  - Set $S_i \leftarrow S_i + \mathbb{E}[h(X_i) | \mathbf{x}_{\Delta_1}^{(t)}, \mathbf{y}]$
  - Sample $X_{\Delta_2}^{(t)} \sim p(\mathbf{x}_{\Delta_1} | \mathbf{x}_{\Delta_1}^{(t)}, \mathbf{y})$ using *Forward filtering / Backward sampling*.

Output Rao-Blackwellised estimates

- $\delta_{rb}(h(X_i)) = \frac{1}{T} S_i$

Fig. 3. Tree sampling.

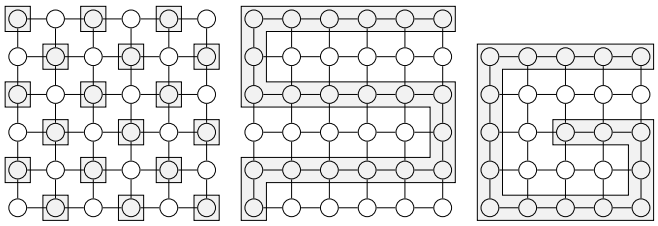

Fig. 4. Alternative partitions of an MRF corresponding to data augmentation methods. Again, the shaded nodes are called X_{Δ_1}. For the leftmost scheme, the elements of Δ_1 are separated by Δ_2, and so the conditional $p(x_{\Delta_1} | x_{\Delta_2}, y)$ is a product of the conditionals of each node. In the rightmost partition there are no unconnected subregions of either partition. The middle case is intermediate.

timators (Casella and Robert 1996, Gelfand and Smith 1990):

$$\delta_{rb}(h(X_i)) = \frac{1}{T} \sum_{t=1}^{T} \mathbb{E}[h(X_i) | \mathbf{x}_{\Delta_1}^{(t)}, \mathbf{y}]$$

where T denotes the number of samples and , in this case, $i \in \Delta_2$. The alternative Monte Carlo histogram estimator is:

$$\delta_{mc}(h(X_i)) = \frac{1}{T} \sum_{t=1}^{T} h(X_i)$$

Both estimators converge to $\mathbb{E}(h(X_i)$, (Liu, Wong and Kong

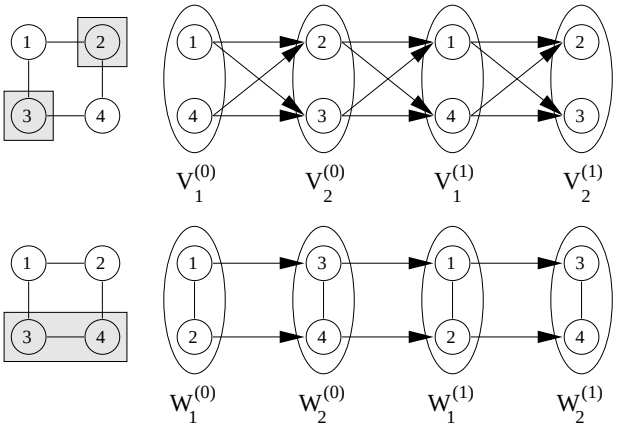

Fig. 5. Illustration of chessboard (top) and fully-connected (bottom) partitioning of variables for a 2x2 MRF. In the top and bottom schemes, the shaded nodes delineate variable sets V_2 and W_2 respectively. The chains beside each show the spatiotemporal dependencies of variables in the resulting Markov chains for each scheme. Theory and experiment reveal that the fully-connected scheme yields a superior RB estimator.

1994) have proved that sampling from data augmentation is a *sufficient condition* for the Rao-Blackwellised estimator to dominate(have lower variance.) To obtain estimates of the node beliefs, we can simply choose h to be the set indicator function, yielding:

$$\widehat{p}(\mathbf{x}_i|\mathbf{y}) = \frac{1}{T}\sum_{t=1}^{T} p(\mathbf{x}_i|\mathbf{x}_{\Delta_k}^{(t)}, \mathbf{y})$$

where $k = 1$ if $i \in \Delta_2$ or 2 if $i \in \Delta_1$. Rao-Blackwellised estimators are more expensive per sample as they require exact computation of the expectations. In practice, the large gain in variance reduction typically offsets this extra cost. Our experiments will show this in our setting.

As shown in Figure 4, we can partition MRFs into trees in several different ways. (We address the theoretical advantages of these and other partitioning schemes in Section III.) Thus it is possible at each iteration of the MCMC sampler to draw a uniform random variable and choose a particular tree partition. In particular, if we consider two different tree sampling algorithms with Markov transition kernels K_1 and K_2 and stationary distribution $p(\mathbf{x}|\mathbf{y})$, then the mixture kernel $K_{mix} = \alpha K_1 + (1-\alpha)K_2$, with $0 \leq \alpha \leq 1$, also converges to $p(\mathbf{x}|\mathbf{y})$ (Tierney 1994).

III. INFORMATION THEORETIC ANALYSIS

The question of what partitions should be adopted must be addressed. As shown in Figure 4 one could partition the graph using trees or, as proposed in (Greenwood, McKeague and Wefelmeyer 1996), using a checker board. In this section, we will show that our tree partitioning scheme results in lower dependency between Markov chain samples.

The question of comparing different univariate sampling schemes has been addressed in (Liu et al. 1994) using the variance decomposition lemma and functional analysis of Markov chains. Our problem is more complex in that the samples are multivariate and conform to a specific graph structure.

Our comparison will focus the checker-board (CB) and two-tree sampling (TS) schemes shown in Figure 4. We define the index sets of the fully-connected and checker-board schemes to be (W_1, W_2), (V_1, V_2) respectively (again, $W_1 \cup W_2 = V_1 \cup V_2 = I$). Figure 5 shows a very simple 2x2 MRF sampled using the TS and CB schemes. Adjacent to each is a corresponding "unrolled" sampling sequence showing the *spatiotemporal* relationships of the sampled variables in the usual manner of graphical models. In the ovals are the variables corresponding to the sampling blocks; the superscripts denote the *time* indices of the samples. Arrows indicate the sampling direction and reveal the "parent/child" temporal dependencies between variables. The undirected links in the ovals of the TS case reflect the spatial dependence of the variables in the blocks. In this example, our samples are multivariate. Hence, we need to compare functions of *all* the variables in a block, say $h(x_1, x_4)$ in TS against $h(x_1, x_2)$ in CB sampling. Here, there is the additional difficulty that the variables are shuffled to different times in the sampling schemes (*e.g.* x_4 in TS does not match x_2 in CB directly).

Instead of using correlations as a measure of dependency between samples (the standard approach in statistics), we propose the use of information theory measures to assess this dependency. The following theorem demonstrates how we can use information theory measures of dependency to compare MCMC sampling schemes.

Theorem 1: Under the stationary distribution, the *mutual information* between samples generated from CB is *larger* than that between samples from TS:

$$I_{CB}(X^{(t+1)}; X^{(t)}) \geq I_{TS}(X^{(t+1)}; X^{(t)})$$

Proof: See appendix.

That is, by integrating out variables in long chains (trees), we reduce the dependency between the Markov chain samples.

IV. NUMERICAL RESULTS

A. IMAGE RECONSTRUCTION

Our first experiment was the classic "reconstruction" of states from noisy observations. We used a 50×50 pixel "patch" image (consisting of shaded, rectangular regions) with an isotropic 11-state prior model. Noise was added by randomly flipping states. Each sampler was run for 1000 iterations on each of 50 separate trials. An important aspect to assess is the *sensitivity* of the estimator, that is, is our good estimate a matter of luck or is it robust over trials? The plot in Figure 6 shows the median reconstruction error as a function of *computation time* showing that the gain is considerable. In fact in this case, the checker-board (CB) sampler is hardly distinguishable from plain Gibbs (PG), again a predictable consequence of the theoretical considerations. For larger graphs, far from expecting any kind of breakdown of these results, we *predict that the difference will become even sharper.* The error bars show that the low reconstruction error of our sampler is highly robust across trials compared to that of PG and CB. We also

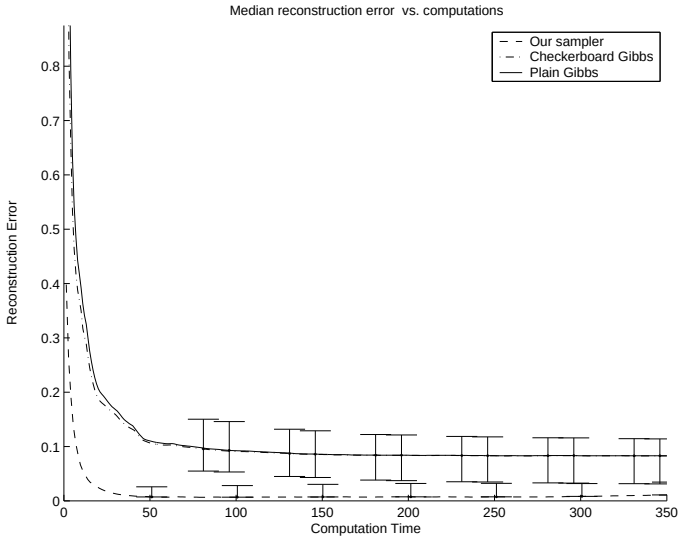

Fig. 6. Reconstruction error against computation time for a 50×50 pixel 11-state image corrupted with noise. The 3 plots show the median reconstruction error history of PG, CB and our sampler over 50 runs. The bars represent the standard deviation of the error across these runs. Clearly, aside from achieving a lower error, our sampler *is more robust*, that is, consistently achieves these results. The gain of CB over PG in this case is negligible, again predictable from the theory of Rao-Blackwellisation

ran loopy on these images, which took about the same number of iterations (around 30 passes through the MRF) to achieve the same error reduction as our method, suggesting that our method might be computationally comparable to loopy but guaranteed to converge. It is very important to realize that the gain is *not* merely due to "blocking"; the CB sampler *is also* a 2-block Rao-Blackwellised scheme, but *does not take advantage of RB as well*.

B. QMR DATABASE

Our second experiment was conducted on the QMR (Quick Medical Reference) factor graph. Here, the interaction potentials, ϕ, represent the priors over binary random variables (corresponding to diseases), and ψ represent the potentials associated with positive findings. Note that in factor graphs, the spatial priors are no longer defined only over pairs of variables as in the simple MRF, but over graph cliques C_i:

$$\psi_i(\mathbf{x}_{C_i}) = 1 - \left((1 - q_{i0}) \prod_{j \in C_i} (1 - q_{ij})^{x_j} \right), \qquad (2)$$

where q_{i0} and the q_{ij} are the parameters of the model. The parameter q_{i0} represents the *leak* probability for the finding i. This is the probability that the finding is caused by other diseases than the ones included in the model. q_{ij} is the probability that disease j, if present, will cause a positive finding i.

The model in our experiment consisted of 40 diseases and 14 potentials (findings). We kept the model relatively small so that we could carry out exact inference for evaluation purposes. A Bernoulli prior was chosen for every disease, with a parameter of 0.01 for the presence of the disease. We used 15 runs in each experiment. For each run, we chose the graph parameters uniformly at random between 0 and 1, and generated the QMR graph randomly. We obtained the true marginals (posteriors) of the variables via the junction tree algorithm. Figure 7 shows that the tree sampling algorithm outperforms naive methods such as Gibbs Sampling and loopy belief propagation (LBP).

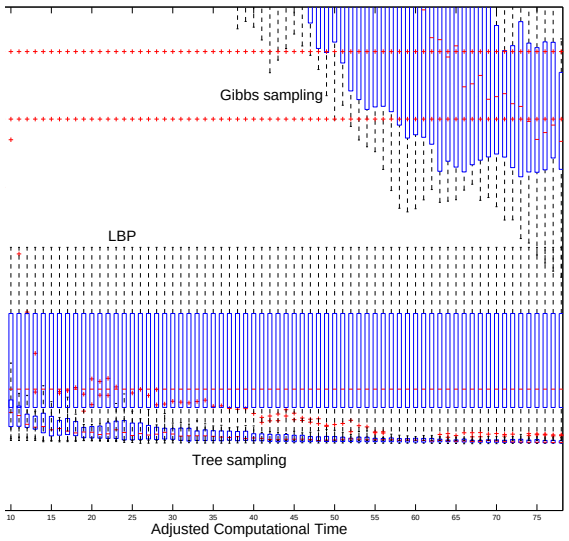

Fig. 7. Performance comparison on a medium-leak QMR network. All algorithms obtain the correct marginals (even if LBP and Gibbs have much higher errors than tree sampling). Tree sampler is clearly the fastest mixing method.

V. CONCLUSION

We presented efficient MCMC algorithms for pairwise undirected graphical models and factor graphs. We showed that information theory measures can be applied to the analysis of MCMC algorithms. In particular, we showed that the *mutual information* between samples from CB is higher than the one for TS. Our experimental results confirmed these assertions. More experiments, details of implementation of the forward filtering backward sampling algorithms for undirected Markov models and factor graphs and, finally, details of implementation of the greedy algorithms for finding automatic tree partitions appear in (Rivasseau 2005).

APPENDIX: PROOF OF THEOREM 1

We show this for the small 2x2 MRF shown in Figure 5; the extension to larger MRFs is inductive. The proof depends on the decomposition of the transition kernel in both cases; it will turn out that we can do a term-by-term analysis of the resulting decompositions and show that some of the TS terms are CB terms with additional integration.

Let the respective joint/conditional probabilities under CB and TS be p_{CB} and p_{TS}. For the CB sampler, the one-step

transition kernel of the joint *chain is:*

$$K_{CB}(x^{(0)}, x^{(1)}) = p_{CB}(x_1^{(1)}|x_2^{(0)}, x_3^{(0)}) p_{CB}(x_2^{(1)}|x_1^{(1)}, x_4^{(1)}) \times p_{CB}(x_3^{(1)}|x_1^{(1)}, x_4^{(1)}) p_{CB}(x_4^{(1)}|x_2^{(0)}, x_3^{(0)})$$

while for the TS sampler, the kernel is:

$$K_{TS}(x^{(0)}, x^{(1)}) = p_{TS}(x_1^{(1)}|x_3^{(0)}, x_4^{(0)}) p_{TS}(x_2^{(1)}|x_1^{(1)}, x_4^{(0)}) \times p_{TS}(x_3^{(1)}|x_1^{(1)}, x_2^{(1)}) p_{TS}(x_4^{(1)}|x_2^{(1)}, x_3^{(1)})$$

Using the reversibility of the augmentation chains, the assumption of stationarity, and the spatial properties of the graph, we see that all *of the above conditional distributions are conditionals from π, the stationary distribution. For example $p_{CB}(x_2^{(1)}|x_1^{(1)}, x_4^{(1)}) = \pi(x_2|x_1, x_4)$. Also note that $p_{CB}(x_2^{(1)}|x_1^{(1)}, x_4^{(1)}) = p_{TS}(x_2^{(1)}|x_1^{(1)}, x_4^{(0)}) = \pi(x_2|x_1, x_4)$ and $p_{CB}(x_4^{(1)}|x_2^{(0)}, x_3^{(0)}) = p_{TS}(x_4^{(1)}|x_2^{(1)}, x_3^{(1)}) = \pi(x_4|x_2, x_3)$. By applying the spatial Markov property in "reverse,"*

$$\pi(x_1^{(1)}|x_2^{(0)}, x_3^{(0)}) = \pi(x_1^{(1)}|x_2^{(0)}, x_3^{(0)}, x_4^{(0)})$$

and

$$\pi(x_3^{(1)}|x_1^{(1)}, x_4^{(1)}) = \pi(x_3^{(1)}|x_1^{(1)}, x_2^{(1)}, x_4^{(1)})$$

If we define the following functions (for conciseness domain variables are omitted.)

$$\begin{aligned} j_1 &= \pi(x_1^{(1)}|x_2^{(0)}, x_3^{(0)}, x_4^{(0)}) \\ j_2 &= \pi(x_2^{(1)}|x_1^{(1)}, x_4^{(1)}) \\ j_3 &= \pi(x_3^{(1)}|x_1^{(1)}, x_2^{(1)}, x_4^{(1)}) \\ j_4 &= \pi(x_4^{(1)}|x_2^{(0)}, x_3^{(0)}) \end{aligned}$$

and:

$$\begin{aligned} k_1 &= \int j_1 \pi(x_2^{(0)}|x_3^{(0)}, x_4^{(0)}) dx_2^{(0)} \\ k_2 &= \pi(x_2^{(1)}|x_1^{(1)}, x_4^{(0)}) \\ k_3 &= \int j_3 \pi(x_4^{(1)}|x_1^{(1)}, x_2^{(1)}) dx_4^{(0)} \\ k_4 &= \pi(x_4^{(1)}|x_2^{(1)}, x_3^{(1)}) \end{aligned}$$

then we can write the transition kernel of CB as:

$$K_{CB}(x^{(0)}, x^{(1)}) = j_1 j_2 j_3 j_4 \tag{3}$$

and that of TS as:

$$K_{TS}(x^{(0)}, x^{(1)}) = k_1 k_2 k_3 k_4 \tag{4}$$

Now let the conditional entropies be

$$\begin{aligned} H_{TS} &\triangleq H(X_{W_1}^{(1)}, X_{W_2}^{(1)}|X_{W_1}^{(0)}, X_{W_2}^{(0)}) = H_{TS}(X^{(1)}|X^{(0)}) \\ H_{CB} &\triangleq H(X_{V_1}^{(1)}, X_{V_2}^{(1)}|X_{V_1}^{(0)}, X_{V_2}^{(0)}) = H_{CB}(X^{(1)}|X^{(0)}) \end{aligned}$$

Using the expressions for $K_{CB}(x^{(0)}, x^{(1)})$ $K_{TS}(x^{(0)}, x^{(1)})$ and the fact that conditioning reduces entropy, it follows that

$$H_{TS} \geq H_{CB}$$

Under stationarity, the marginal *entropy $H(X^{(t)})$ of* both *schemes is the same. Hence by the decomposition of mutual information in terms of marginal and conditional entropies, we have our final result:*

$$I_{CB}(X^{(1)}; X^{(0)}) \geq I_{TS}(X^{(1)}; X^{(0)})$$

REFERENCES

Besag, J. E. (1974). Spatial interaction and the statistical analysis of lattice systems, *Journal of the Royal Statistical Society B* **36**: 192–236.

Besag, J. E. (1986). On the statistical analysis of dirty pictures, *Journal of the Royal Statistical Society B* **48**(3): 259–302.

Carter, C. K. and Kohn, R. (1994). On Gibbs sampling for state space models, *Biometrika* **81**(3): 541–553.

Casella, G. and Robert, C. P. (1996). Rao-Blackwellisation of sampling schemes, *Biometrika* **83**(1): 81–94.

Gelfand, A. E. and Smith, A. F. M. (1990). Sampling-based approaches to calculating marginal densities, *Journal of the American Statistical Association* **85**(410): 398–409.

Geman, S. and Geman, D. (1984). Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images, *IEEE Transactions on Pattern Analysis and Machine Intelligence* **6**(6): 721–741.

Greenwood, P., McKeague, I. and Wefelmeyer, W. (1996). Outperforming the Gibbs sampler empirical estimator for nearest neighbor random fields, *Annals of Statistics* **24**: 1433–1456.

Hamze, F. and de Freitas, N. (2004). From fields to trees, *Uncertainty in Artificial Intelligence*.

Kumar, S. and Hebert, M. (2003). Discriminative fields for modeling spatial dependencies in natural images, *in proc. advances in Neural Information Processing Systems (NIPS)*.

Li, S. Z. (2001). *Markov random field modeling in image analysis*, Springer-Verlag.

Liu, J. S. (ed.) (2001). *Monte Carlo Strategies in Scientific Computing*, Springer-Verlag.

Liu, J., Wong, W. H. and Kong, A. (1994). Covariance structure of the Gibbs sampler with applications to the comparisons of estimators and augmentation schemes, *Biometrika* **81**(1): 27–40.

McCallum, A., Rohanimanesh, K. and Sutton, C. (2003). Dynamic conditional random fields for jointly labeling multiple sequences, *NIPS Workshop on Syntax, Semantics and Statistics*.

Pearl, J. (1987). Evidential reasoning using stochastic simulation, *Artificial Intelligence* **32**: 245–257.

Rivasseau, J. (2005). *Growing Trees for Markov Chain Monte Carlo Inference*, PhD thesis, Department of Computer Science, University of British Columbia.

Robert, C. P. and Casella, G. (1999). *Monte Carlo Statistical Methods*, Springer-Verlag, New York.

Swendsen, R. H. and Wang, J. S. (1987). Nonuniversal critical dynamics in Monte Carlo simulations, *Physical Review Letters* **58**(2): 86–88.

Tierney, L. (1994). Markov chains for exploring posterior distributions, *The Annals of Statistics* **22**(4): 1701–1762.

Wainwright, M., Jaakkola, T. and Willsky, A. (2003). Tree-reweighted belief propagation and approximate ML estimation by pseudo-moment matching, *AI-STATS*, Florida, USA.

Wilkinson, D. J. and Yeung, S. K. H. (2001). Conditional simulation from highly structured gaussian systems, with application to blocking-MCMC for the Bayesian analysis of very large linear models, *Statistics and Computing* **11**: To appear.

Yedidia, J. S., Freeman, W. T. and Weiss, Y. (2000). Generalized belief propagation, *in* S. Solla, T. Leen and K.-R. Müller (eds), *Advances in Neural Information Processing Systems 12*, MIT Press, pp. 689–695.