Exact and Approximate Sampling by Systematic Stochastic Search

Vikash Kumar Mansinghka

Ph.D. Thesis, Chapter 3

Computational Intelligence Seminar April 2010
presented by Dejan Pecevski

Outline

- Motivation
- Sequential rejection sampling
- Adaptive sequential rejection sampling
- Results from experiments
- Discussion

Inference and Multimodality

- Given a probabilistic model P(x,z) and visible variables x we want to find posterior distribution over the causes P(z|x).
- Inference becomes difficult for high-dimensional distributions with widely separated hard-to-find modes.
- Existing approaches: variational methods, convex relaxations and generalized BP often fail to capture the multimodal character of the distribution.
- MCMC estimates like Gibbs sampling also are not efficient in cases of widely separated modes.

Motivation: Systematic and Local Algorithms

Domain	Systematic	Local
Sorting	MergeSort	BubbleSort
SAT solving	DFS, BFS, arc consistency, DPLL	Min constraint
ODEs/PDEs	Shooting (e.g. Euler, RK4)	Relaxation (e.g. Gauss-Seidel)
Linear systems	Algebraic (e.g. LU/QR)	Iterative (e.g. conjugate gradient)

Deterministic Search	depth-first search, backtracking	Hill Climbing
Sampling	?	MCMC methods

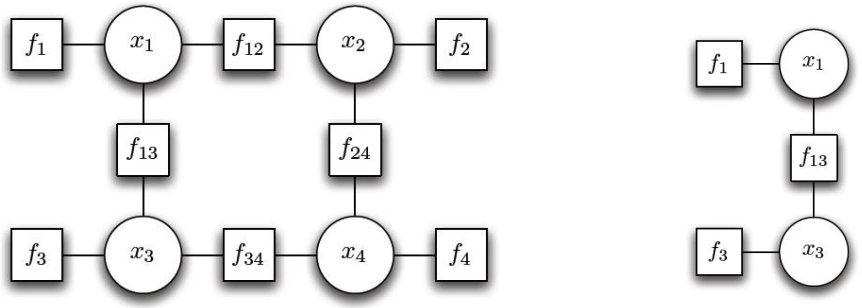
What are stochastic generalizations of systematic deterministic search algorithms?

Key Idea of the Algorithm: Divide and Conquer

- Construct partial samples by adding one variable at a time
- At each step in the sequence, manage correlations between the newly added variable and the variables already in the partial sample

Creating Sequence of Distributions from Graphical Models

- **Restriction of a factor graph** on a subset of variables S is a subgraph that:
  - Includes the subset of variables S
  - Only the factors that have all the variables they depend on in S



- We consider a particular ordering of the variables $x_1, x_2, \cdots, x_n$
- Let $X_{1:k} = \{x_1, \cdots, x_k\}$ are the first k variables in the ordering
- The sequence of nested distributions are the **restrictions** over the subsets $X_{1:k}$

- Good variable ordering puts the more constrained variables first.

Sequential Rejection Sampling

- We assume by induction that we already generated sample $\hat{y}$ from the distribution $\overline{p}'(y) \propto \psi_1(y)$
- We want to generate samples from a discrete distribution $p(x), x \in \mathbf{X}$

$$\overline{p}(x) = \overline{p}(y,z) \propto \psi_1(y)\psi_2(y,z), \quad \mathbf{X} = \mathbf{Y} \times Z$$

- We define the Gibbs transition kernel

$$q_p(z|y) \triangleq \frac{\overline{p}(y,z)}{\sum_{z'} \overline{p}(y,z')} = \frac{\psi_2(y,z)}{\sum_{z'} \psi_2(y,z')}$$

- And use $p' q_p$ as proposal distribution for rejection sampling of p(x)

Sequential Rejection Sampling 2

- The weight of the generated sample is

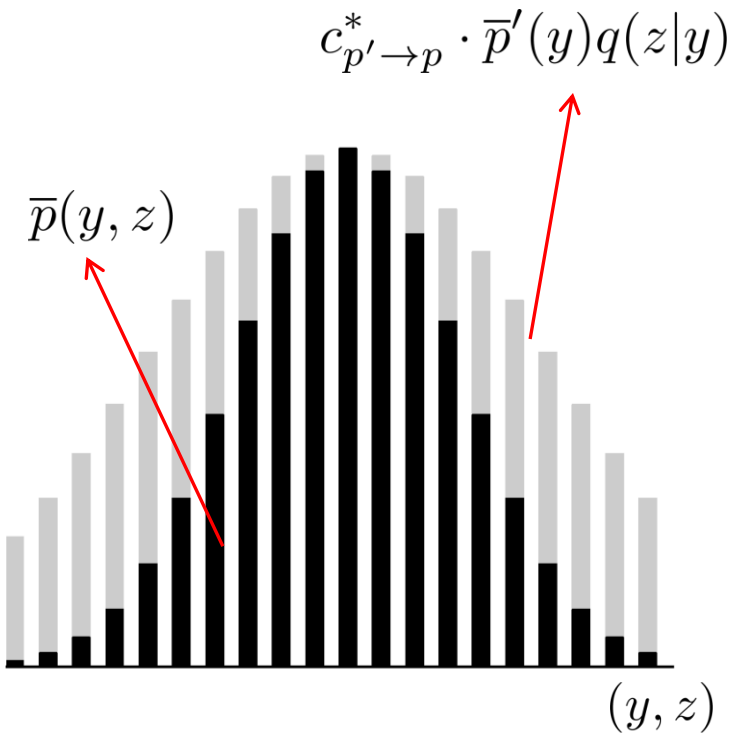
$$w_{p' \to p}(\hat{x}) \triangleq \frac{\overline{p}(\hat{y}, \hat{z})}{\overline{p}'(\hat{y}) q(\hat{z}|\hat{y})} = \frac{\overline{p}(y)}{\overline{p}'(y)} = \sum_{z'} \psi_2(\hat{y}, z')$$

- The sample is accepted with probability

$$\frac{w_{p' \to p}(\hat{y})}{c_{p' \to p}} \qquad c_{p' \to p} > w_{p' \to p}(y) \quad \forall y \in \mathbf{Y}$$

- $c_{p' \to p}$ is calculated as

$$c^*_{p' \to p} \triangleq \max_y w_{p' \to p}(y) = \max_y \sum_{z'} \psi_2(y, z')$$



- If the factor $\psi_2(y, z)$ is a function of only O(log n) dimensions y_i then $c^*_{p' \to p}$ is calculable in polynomial time.

Introducing Adaptation

- Probability of acceptance of the non-adaptive sampler is

$$\alpha_{p' \to p} = \frac{\mathbb{E}_{p'} \left\{ w_{p' \to p}(\hat{y}) \right\}}{c^*_{p' \to p}}$$

- From $w_{p' \to p}(y) \propto \frac{p(y)}{p'(y)}$ follows

$$\alpha_{p' \to p} = \frac{\sum_y p'(y) \frac{p(y)}{p'(y)}}{\max_y \frac{p(y)}{p'(y)}} = \min_y \frac{p'(y)}{p(y)}$$

The goal is to increase $\alpha_{p' \to p}$ and drive it to 1 by adapting the proposal distribution $p'(y)$ to be closer to *p(y)*.

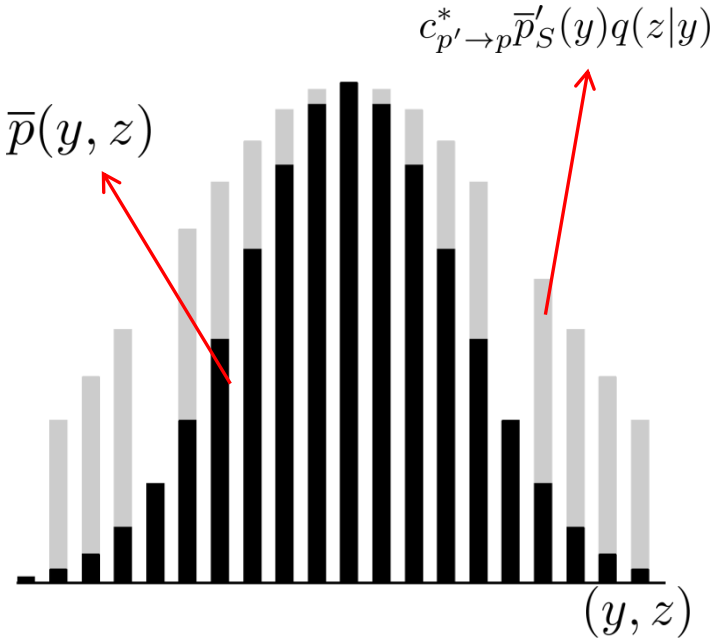
Adaptive Sequential Rejection Sampling

- After proposing a sample $(\hat{y}, \hat{z})$, we augment a set S with $\hat{y}$.

- Instead of p'(y) we sample from the distribution

$$p'_S(y) \propto p'(y) \prod_{y' \in \mathbf{S}} \left(\frac{w_{p' \to p(y)}}{c^*_{p' \to p}} \right)^{\delta_{yy'}}$$

$S \subset \mathbf{Y}$, $\delta_{yy'}$ is the Kronecker delta function.



Then

$$w_{p'_S \to p}(\hat{x}) \triangleq \frac{p(\hat{y}, \hat{z})}{p'_S(\hat{y}) q(\hat{z}|\hat{y}))} = \begin{cases} c^*_{p' \to p} & y \in S, \\ w_{p' \to p}(y) & y \notin S. \end{cases}$$

If S = Y , then $w_{p'_S \to p} = c^*_{p' \to p}$ and every sample is accepted

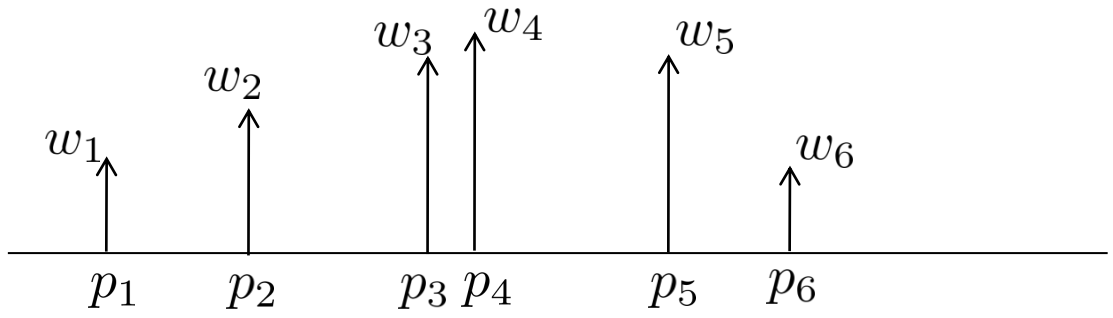
$$\overline{p}'_{S=\mathbf{Y}}(y) \propto \overline{p}'(y) w_{p' \to p}(y) = \overline{p}(y)$$

Adaptive Sequential Rejection Sampling 2

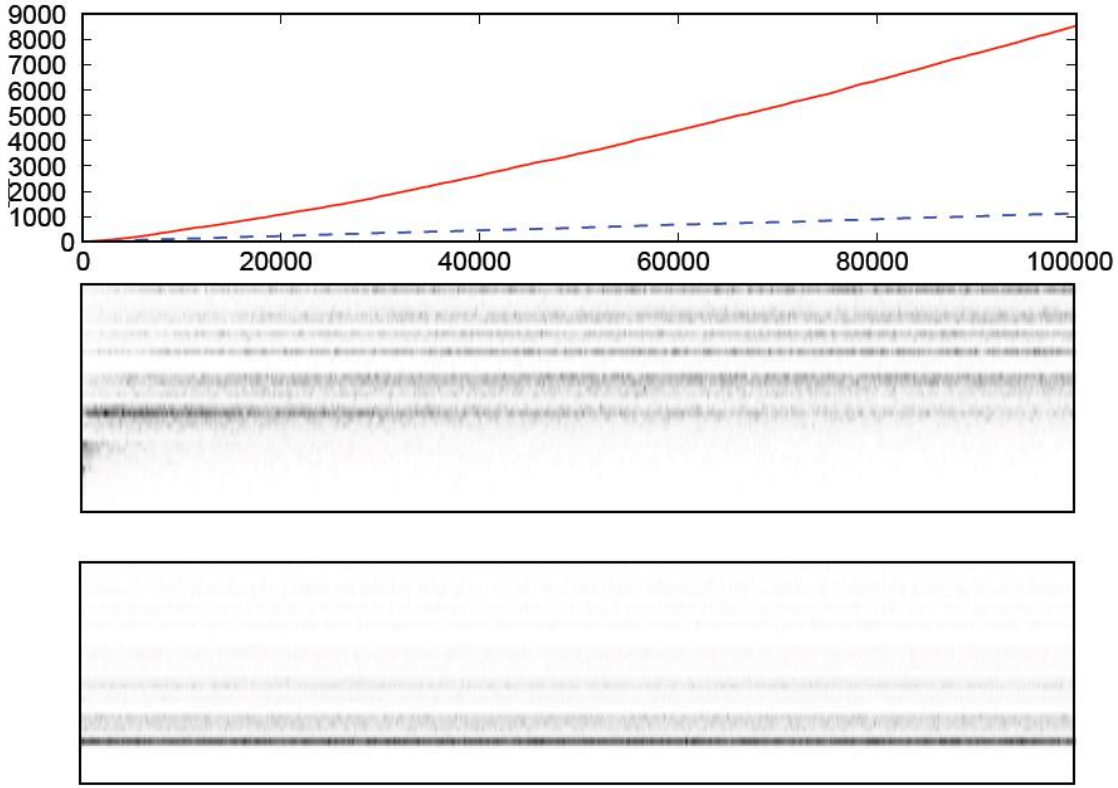
- A hashmap of visited states is stored for every distribution in the sequence.
- The Gibbs transition kernels are modified to reflect the hashmap contents.
- In the deterministic case the algorithm recovers the backtracking meta-heuristics.

Importance Sampling with Resampling Instead of Rejection

- At each stage we have k particles.
- We extend all particles with the value from the z random variable.
- Resample k particles from the discrete distribution over the particles with their rejection weights.



Results: Ising Models

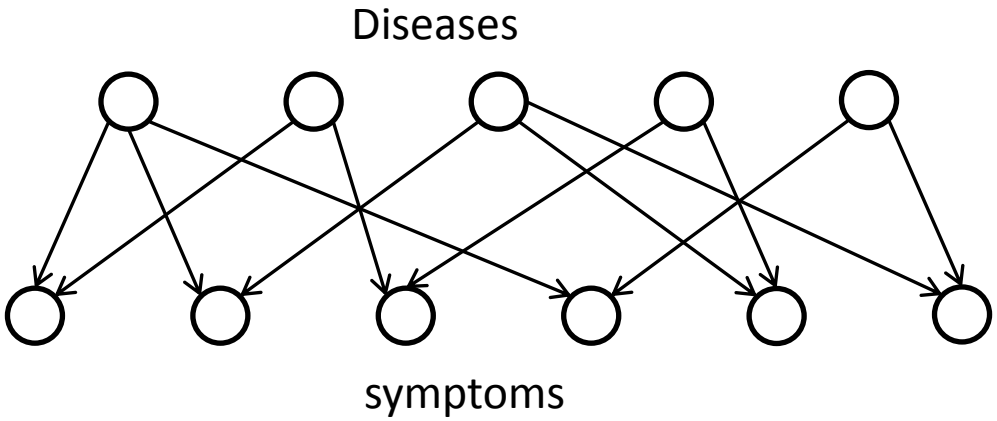


$$P(\mathbf{x}) \propto \exp(-\sum_{(i \neq j)} J_{ij} x_i x_j)$$

$$x \in \{-1, 1\}^n$$

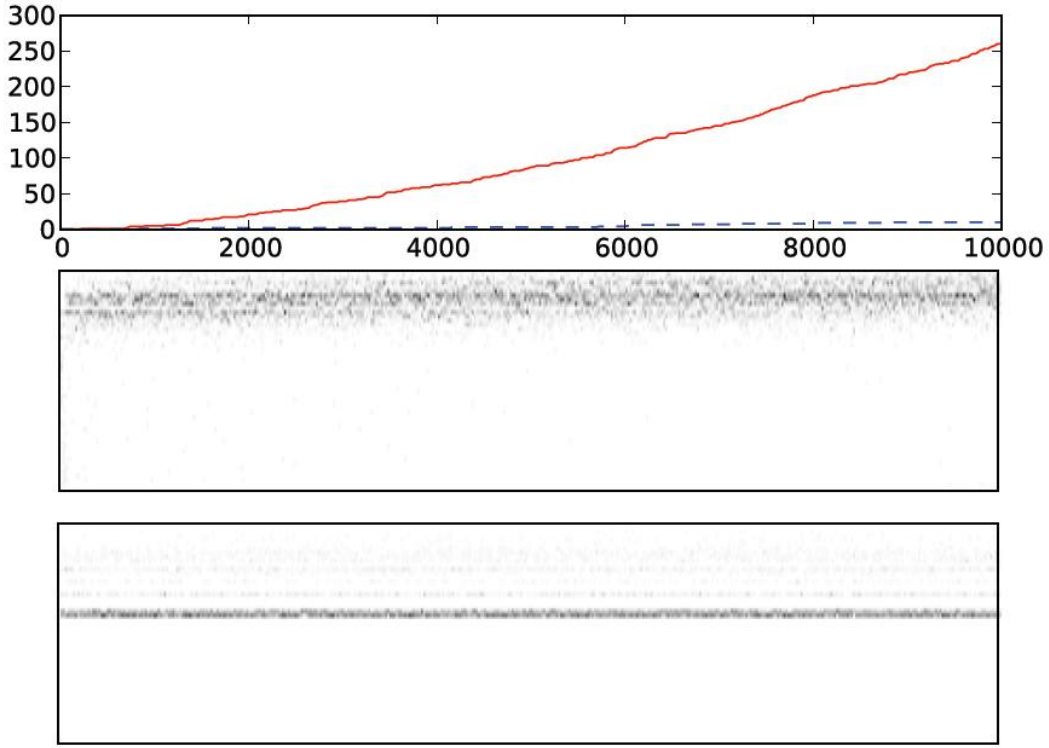
- 36-dimensional Ising model with uniform [-2, 2] distributed coupling parameters.
- Cumulative samples after 100000 iterations shown.

Medical Diagnosis Network



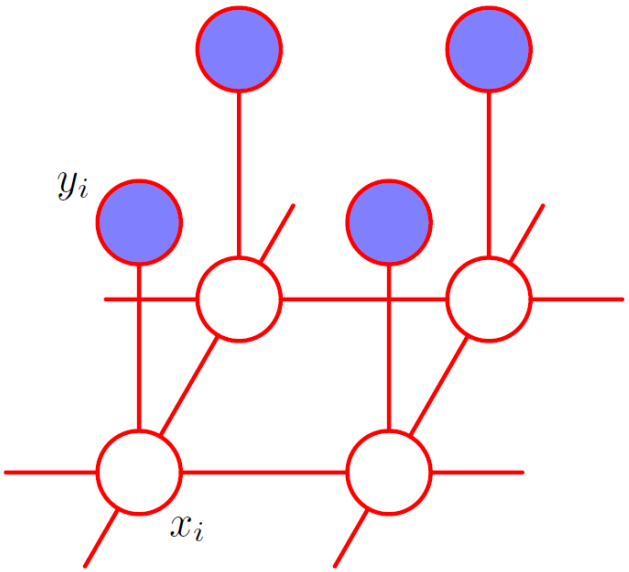
- 20 diseases and 30 symptoms, randomly generated links
- Belief propagation methods do not work for this network type
- Large factors over the diseases

Results: Medical Diagnosis Network



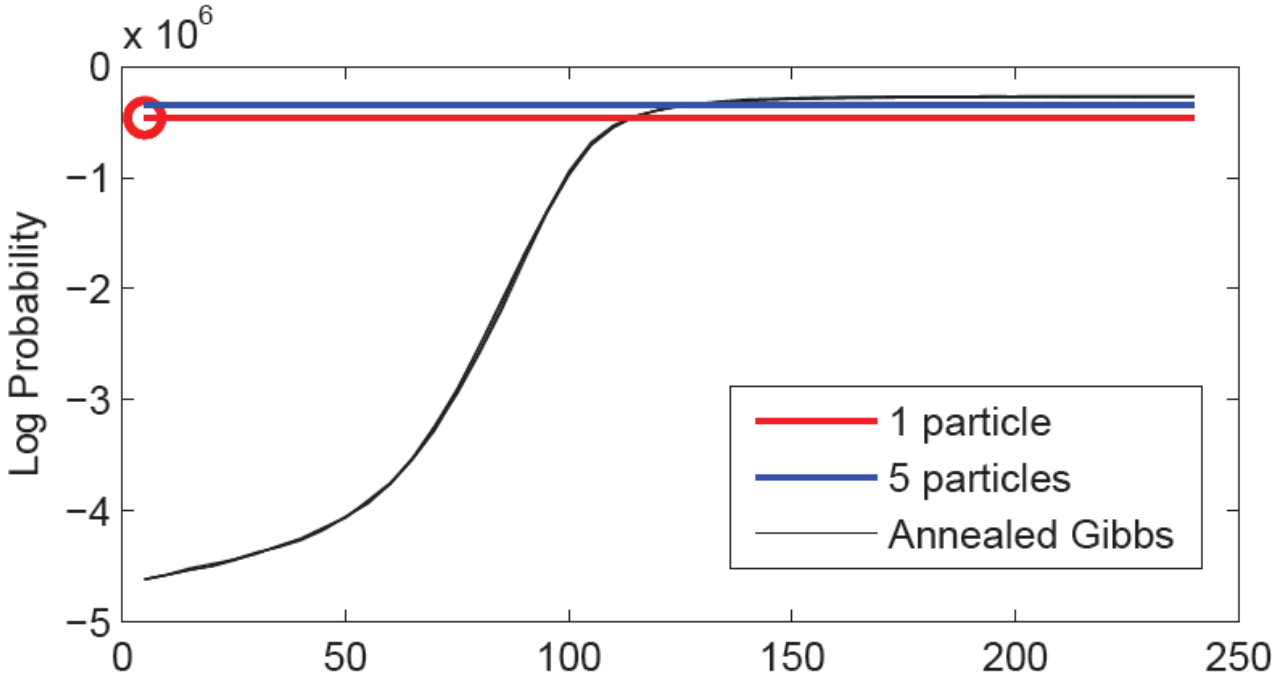
- Adaptation leads to reduced number of rejections, because the partial diagnosis is improved.

Disparity Estimation in Stereo Vision



- 61 344 nodes, each with 30 states
- We want to find configuration of disparities compatible with the pixel intensity difference
- Importance based variant of the algorithm used => approximate samples

Results: Disparity Estimation in Stereo Vision



- The performance is comparable with Annealed Gibbs sampling with 130 performed iterations.

Extensions of the Approach and Its Implications

- Hybrid systematic/local algorithms for sampling
  - Combining the algorithm with Markov chain methods
- Divide and conquer with more aggressive combination of partial samples
  - E.g. merge sort for sorting
- Deterministic search: richer and more developed field
  => Many other algorithms and ideas from deterministic search could potentially be extended to stochastic simulation.
- Cognitive modeling – the execution of a recursive sampler seems to better match the dynamics of probabilistic reasoning than stochastic fixed-point iteration

Summary

- Ideas from deterministic systematic search were generalized in the stochastic setting to produce the **adaptive sequential rejection sampling** method.
- The method performs well in case of multimodality.
- It finds high-probability regions by managing correlations one variable at a time.
- In the extreme case of deterministic constraints, it recovers depth first search with backtracking.