Approximate inference - BNDG 4.8

Finn V. Jensen and Thomas D. Nielsen

Approximate Inference

Motivation

Because of the (worst-case) intractability of exact inference in Bayesian networks, try to find more efficient approximate inference techniques:
instead of computing exact posterior

$$P(A \mid \mathbf{E} = \mathbf{e})$$

compute approximation

$$\hat{P}(A \mid \mathbf{E} = \mathbf{e})$$

with

$$\hat{P}(A \mid \mathbf{E} = \mathbf{e}) \sim P(A \mid \mathbf{E} = \mathbf{e})$$

Approximate Inference

Absolute/Relative Error

For $p, \hat{p} \in [0,1]$: $\hat{p}$ is approximation for p with *absolute error* $\leq \epsilon$, if

$$| p - \hat{p} | \leq \epsilon, \text{ i.e. } \hat{p} \in [p - \epsilon, p + \epsilon].$$

Approximate Inference

Absolute/Relative Error

For $p, \hat{p} \in [0,1]$: $\hat{p}$ is approximation for p with *absolute error* $\leq \epsilon$, if

$$| p - \hat{p} | \leq \epsilon, \text{ i.e. } \hat{p} \in [p - \epsilon, p + \epsilon].$$

$\hat{p}$ is approximation for p with *relative error* $\leq \epsilon$, if

$$| 1 - \hat{p}/p | \leq \epsilon, \text{ i.e. } \hat{p} \in [p(1-\epsilon), p(1+\epsilon)].$$

Approximate Inference

Absolute/Relative Error

For $p, \hat{p} \in [0,1]$: $\hat{p}$ is approximation for p with *absolute error* $\leq \epsilon$, if

$$\mid p - \hat{p} \mid \leq \epsilon, \text{ i.e. } \hat{p} \in [p - \epsilon, p + \epsilon].$$

$\hat{p}$ is approximation for p with *relative error* $\leq \epsilon$, if

$$\mid 1 - \hat{p}/p \mid \leq \epsilon, \text{ i.e. } \hat{p} \in [p(1-\epsilon), p(1+\epsilon)].$$

This definition is not always fully satisfactory, because it is not symmetric in p and $\hat{p}$ and not invariant under the transition $p \rightarrow (1-p)$, $\hat{p} \rightarrow (1-\hat{p})$. Use with care!

When $\hat{p}_1, \hat{p}_2$ are approximations for p_1, p_2 with absolute error $\leq \epsilon$, then no error bounds follow for $\hat{p}_1/\hat{p}_2$ as an approximation for p_1/p_2.

When $\hat{p}_1, \hat{p}_2$ are approximations for p_1, p_2 with relative error $\leq \epsilon$, then $\hat{p}_1/\hat{p}_2$ approximates p_1/p_2 with relative error $\leq (2\epsilon)/(1+\epsilon)$.

Approximate Inference

Randomized Methods

Most methods for approximate inference are randomized algorithms that compute approximations $\hat{P}$ from random samples of instantiations.

We shall consider:

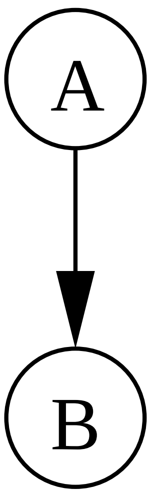
- Forward sampling
- Likelihood weighting
- Gibbs sampling
- Metropolis Hastings algorithm

Approximate Inference

Forward Sampling

Observation: can use Bayesian network as random generator that produces full instantiations $\mathbf{V} = \mathbf{v}$ according to distribution $P(\mathbf{V})$.

Example:



	A: t	A: f
	.2	.8

A	B: t	B: f
t	.7	.3
f	.4	.6

- Generate random numbers r_1, r_2 uniformly from [0,1].
- Set $A = t$ if $r_1 \leq .2$ and $A = f$ else.
- Depending on the value of A and r_2 set B to t or f.

Generation of one random instantiation: linear in size of network.

Approximate Inference

Sampling Algorithm

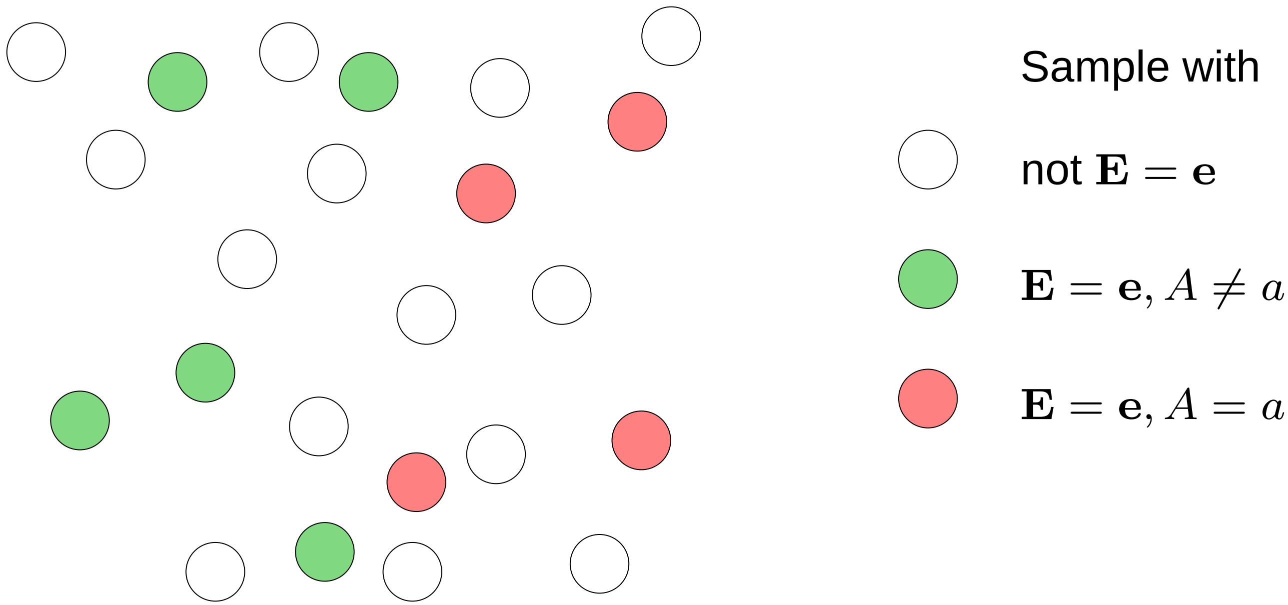
Thus, we have a randomized algorithm S that produces possible outputs from $\mathrm{sp}(\mathbf{V})$ according to the distribution $P(\mathbf{V})$.

Define

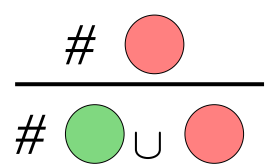
$$\hat{P}(A = a \mid \mathbf{E} = \mathbf{e}) := \frac{|\{i \in 1, \ldots, N \mid \mathbf{E} = \mathbf{e}, A = a \text{ in } S_i\}|}{|\{i \in 1, \ldots, N \mid \mathbf{E} = \mathbf{e} \text{ in } S_i\}|}$$

Approximate Inference

Forward Sampling: Illustration



Approximation for $P(A = a \mid \mathbf{E} = \mathbf{e})$:

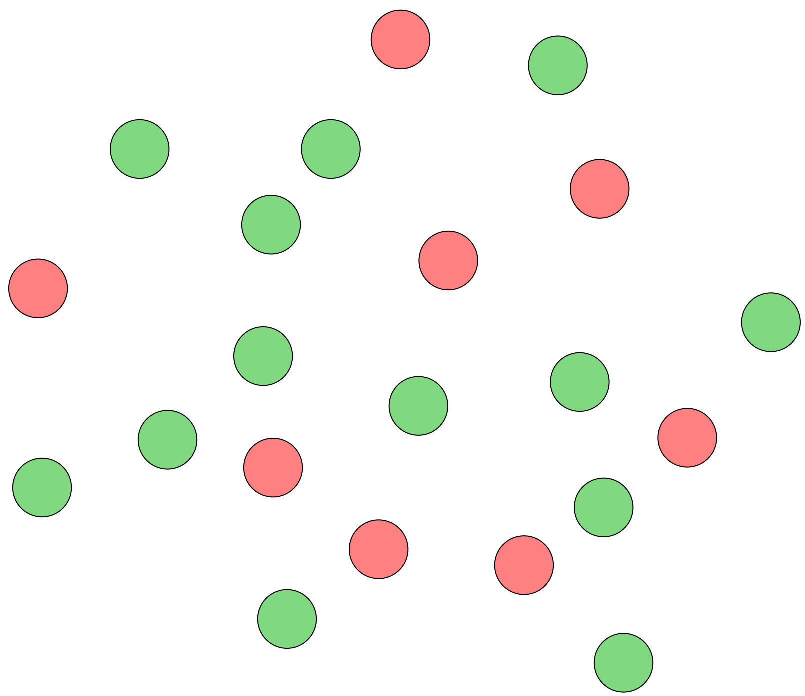


Approximate Inference

Sampling from the conditional distribution

Problem of forward sampling: samples with $\mathbf{E} \neq \mathbf{e}$ are useless!

Idea: find sampling algorithm S_c that produces outputs from $\mathrm{sp}(\mathbf{V})$ according to the distribution $P(\mathbf{V} \mid \mathbf{E} = \mathbf{e})$.

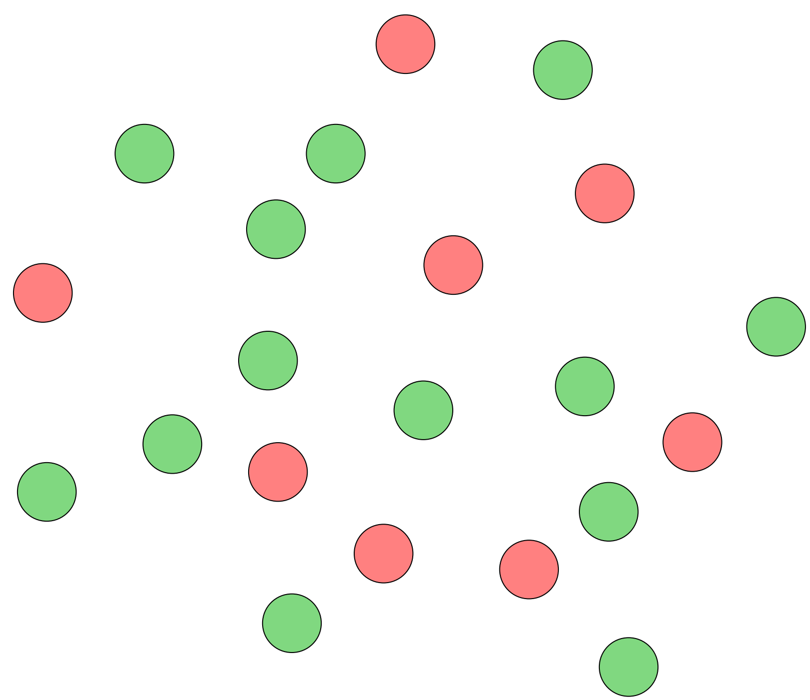


Approximate Inference

Sampling from the conditional distribution

Problem of forward sampling: samples with $\mathbf{E} \neq \mathbf{e}$ are useless!

Idea: find sampling algorithm S_c that produces outputs from $\mathrm{sp}(\mathbf{V})$ according to the distribution $P(\mathbf{V} \mid \mathbf{E} = \mathbf{e})$.



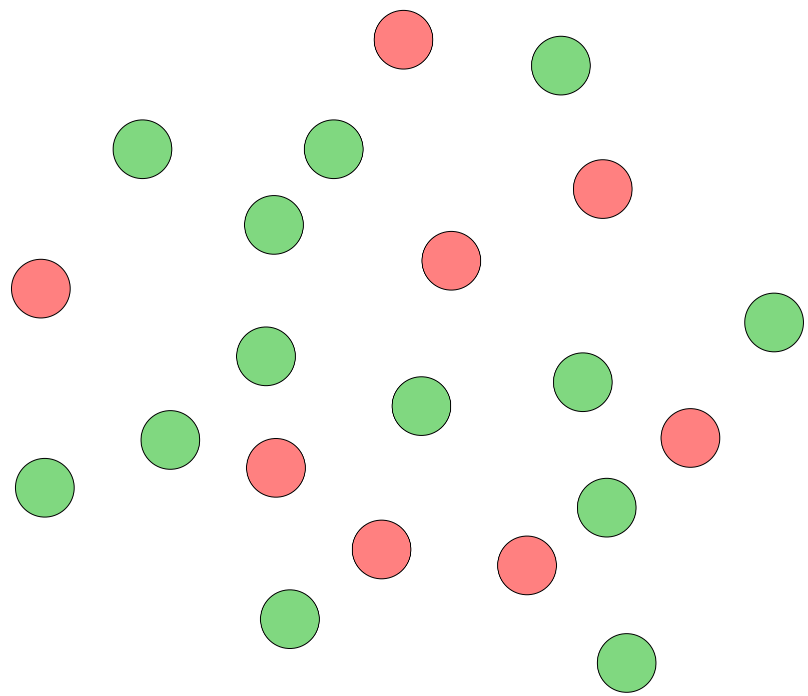
A tempting approach: Fix the variables in $\mathbf{E}$ to $\mathbf{e}$ and sample from the nonevidence variables only!

Approximate Inference

Sampling from the conditional distribution

Problem of forward sampling: samples with $\mathbf{E} \neq \mathbf{e}$ are useless!

Idea: find sampling algorithm S_c that produces outputs from $\mathrm{sp}(\mathbf{V})$ according to the distribution $P(\mathbf{V} \mid \mathbf{E} = \mathbf{e})$.



A tempting approach: Fix the variables in $\mathbf{E}$ to $\mathbf{e}$ and sample from the nonevidence variables only! Problem: Only evidence from the ancestors are taken into account!

Approximate Inference

Likelihood weighting

We would like to sample from ($\mathrm{pa}(X)''$ are the parents in $\mathbf{E}$)

$$P(\mathcal{U},\mathbf{e}) = \prod_{X\in\mathcal{U}\setminus\mathbf{E}} P(X \mid \mathrm{pa}(X)', \mathrm{pa}(X)'' = \mathbf{e}) \times \prod_{X\in\mathbf{E}} P(X = e \mid \mathrm{pa}(X)', \mathrm{pa}(X)'' = \mathbf{e}),$$

but by applying forward sampling with fixed $\mathbf{E}$ we actually sample from:

$$\text{Sampling distribution} = \prod_{X\in\mathcal{U}\setminus\mathbf{E}} P(X \mid \mathrm{pa}(X)', \mathrm{pa}(X)'' = \mathbf{e}).$$

Likelihood weighting

We would like to sample from ($\mathrm{pa}(X)''$ are the parents in $\mathbf{E}$)

$$P(\mathcal{U}, \mathbf{e}) = \prod_{X \in \mathcal{U} \setminus \mathbf{E}} P(X \mid \mathrm{pa}(X)', \mathrm{pa}(X)'' = \mathbf{e}) \times \prod_{X \in \mathbf{E}} P(X = e \mid \mathrm{pa}(X)', \mathrm{pa}(X)'' = \mathbf{e}),$$

but by applying forward sampling with fixed $\mathbf{E}$ we actually sample from:

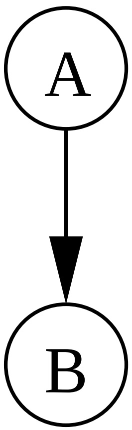
$$\text{Sampling distribution} = \prod_{X \in \mathcal{U} \setminus \mathbf{E}} P(X \mid \mathrm{pa}(X)', \mathrm{pa}(X)'' = \mathbf{e}).$$

Solution: Instead of letting each sample count as 1, use

$$w(\mathbf{x}, \mathbf{e}) = \prod_{X \in \mathbf{E}} P(X = e \mid \mathrm{pa}(X)', \mathrm{pa}(X)'' = \mathbf{e}).$$

Approximate Inference

Likelihood weighting: example



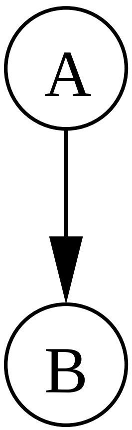
	A	
	t	f
	.2	.8

	B	
A	t	f
t	.7	.3
f	.4	.6

- Assume evidence $B = t$.
- Generate a random number r uniformly from [0,1].
- Set $A = t$ if $r \leq .2$ and $A = f$ else.
- If $A = t$ then let the sample count as $w(t, t) = 0.7$; otherwise $w(f, t) = 0.4$.

Approximate Inference

Likelihood weighting: example



	A	
	t	f
	.2	.8

	B	
A	t	f
t	.7	.3
f	.4	.6

- Assume evidence $B = t$.
- Generate a random number r uniformly from [0,1].
- Set $A = t$ if $r \leq .2$ and $A = f$ else.
- If $A = t$ then let the sample count as $w(t,t) = 0.7$; otherwise $w(f,t) = 0.4$.

With N samples $(a_1, \ldots, a_N)$ we get

$$\hat{P}(A = t \mid B = t) = \frac{\sum_{i=1}^{N} w(a_i = t, e)}{\sum_{i=1}^{N} (w(a_i = t, e) + w(a_i = f, e))}.$$

Approximate Inference

Gibbs Sampling

For notational convenience assume from now on that for some l: $\mathbf{E} = V_{l+1}, V_{l+2}, \ldots, V_n$.
Write $\mathbf{W}$ for $V_1, \ldots, V_l$.

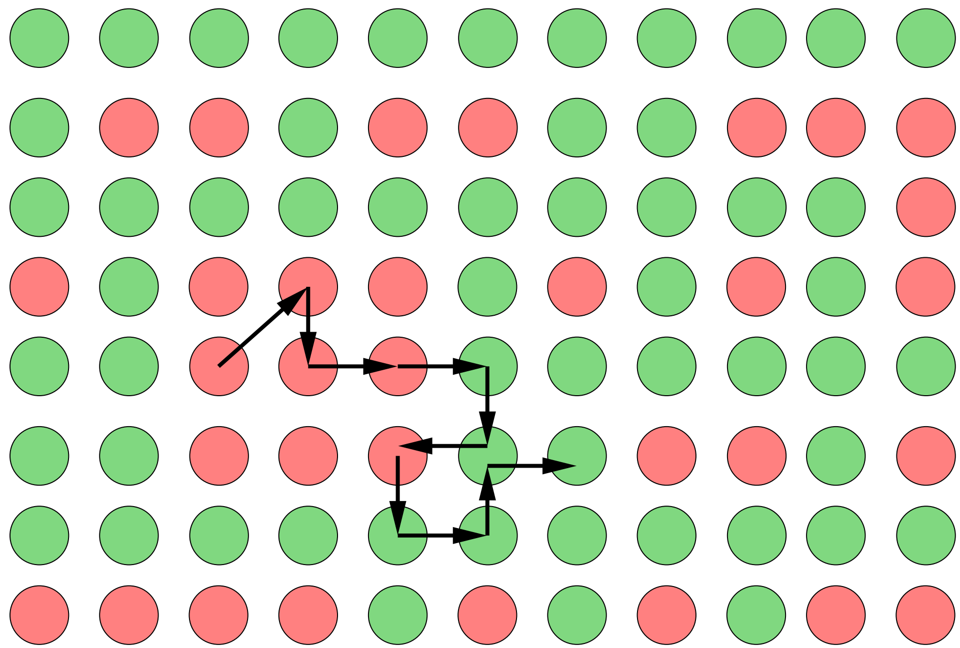
Principle: obtain new sample from previous sample by randomly changing the value of only one selected variable.

```
Procedure Gibbs sampling
v_0 = (v_{0,1}, ..., v_{0,l}) := arbitrary instantiation of W
i := 1
repeat forever
      choose V_k ∈ W        # deterministic or randomized
      generate randomly v_{i,k} according to distribution
            P(V_k | V_1 = v_{i-1,1}, ..., V_{k-1} = v_{i-1,k-1},
                  V_{k+1} = v_{i-1,k+1}, ..., V_l = v_{i-1,l}, E = e)
      set v_i = (v_{i-1,1}, ... v_{i-1,k-1}, v_{i,k}, v_{i-1,k+1}, ..., v_{i-1,l})
      i := i + 1
```

Approximate Inference

Illustration

The process of Gibbs sampling can be understood as a *random walk* in the space of all instantiations with $\mathbf{E} = \mathbf{e}$:



Reachable in one step: instantiations that differ from current one by value assignment to at most one variable (assume randomized choice of variable V_k).

Approximate Inference

Implementation of Sampling Step

The sampling step

$$\texttt{generate randomly } v_{i,k} \text{ according to distribution}$$
$$P(V_k \mid V_1 = v_{i-1,1}, \ldots, V_{k-1} = v_{i-1,k-1}, V_{k+1} = v_{i-1,k+1}, \ldots, V_l = v_{i-1,l}, \mathbf{E} = \mathbf{e})$$

requires sampling from a conditional distribution. In this special case (all but one variables are instantiated) this is easy: just need to compute for each $v \in \mathrm{sp}(V_k)$ the probability

$$P(V_1 = v_{i-1,1}, \ldots, V_{k-1} = v_{i-1,k-1}, V_k = v, V_{k+1} = v_{i-1,k+1}, \ldots, V_l = v_{i-1,l}, \mathbf{E} = \mathbf{e})$$

(linear in network size), and choose $v_{i,k}$ according to these probabilities (normalized).

This can be further simplified by computing the distribution on $\mathrm{sp}(V_k)$ only in the *Markov blanket* of V_k, i.e. the subnetwork consisting of V_k, its parents, its children, and the parents of its children.

Approximate Inference

Convergence of Gibbs Sampling

Under certain conditions: the distribution of samples converges to the posterior distribution $P(\mathbf{W} \mid \mathbf{E} = \mathbf{e})$:

$$\lim_{i \to \infty} P(\mathbf{v}_i = \mathbf{v}) = P(\mathbf{W} = \mathbf{v} \mid \mathbf{E} = \mathbf{e}) \quad (\mathbf{v} \in \mathrm{sp}(\mathbf{W})).$$

Sufficient conditions are:

- in the `repeat` loop of the Gibbs sampler, variable V_k is randomly selected (with non-zero selection probability for all $V_k \in \mathbf{W}$), and
- the Bayesian network has no zero-entries in its cpt's

Approximate Inference

Approximate Inference using Gibbs Sampling

1. Start Gibbs sampling with some starting configuration $\mathbf{v}_0$.
2. Run the sampler for N steps ("Burn in")
3. Run the sampler for M additional steps; use the relative frequency of state $\mathbf{v}$ in these M samples as an estimate for $P(\mathbf{W} = \mathbf{v} \mid \mathbf{E} = \mathbf{e})$.

Problems:

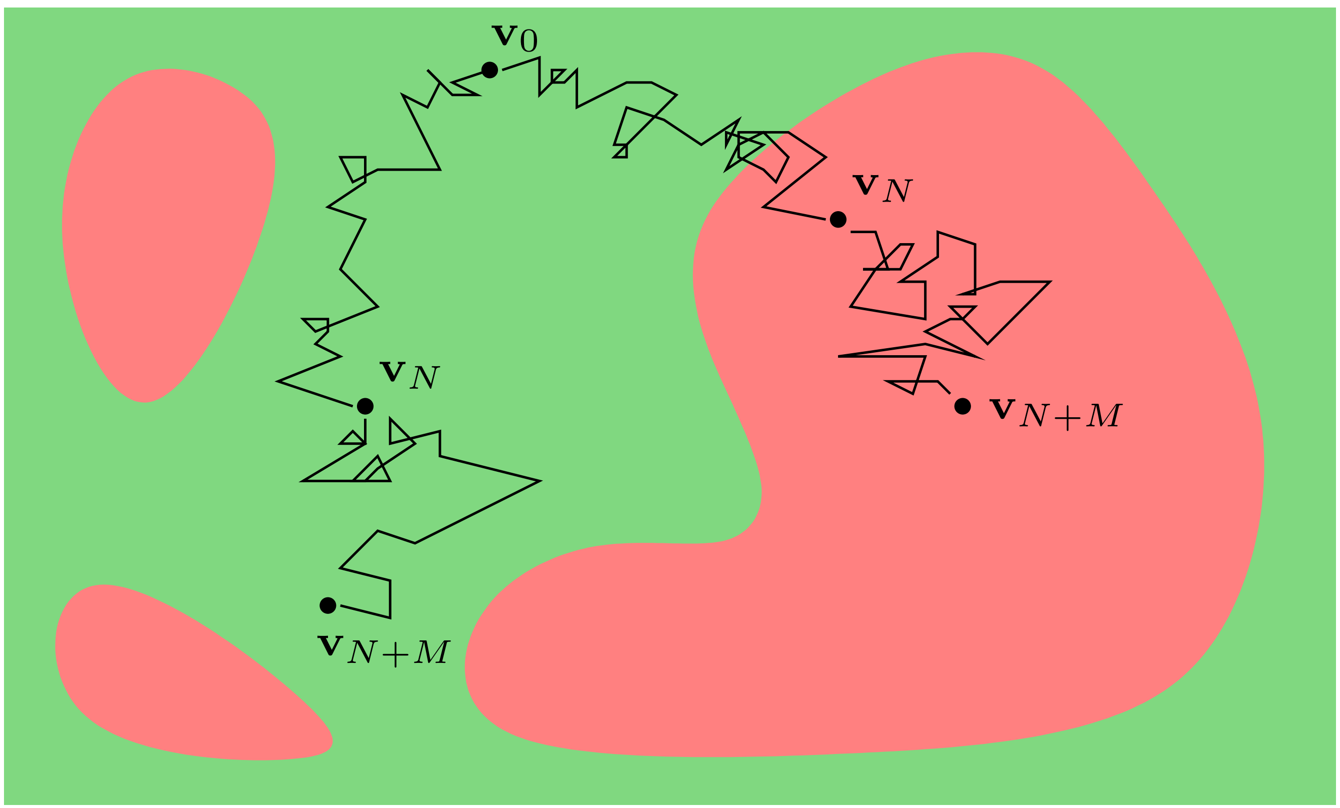
- How large must N be chosen? Difficult to say how long it takes for Gibbs sampler to converge!
- Even when sampling is from the stationary distribution, samples are not independent. Result: error cannot be bounded as function of M using Chebyshev's inequality (or related methods).

Approximate Inference

Effect of dependence

$P(\mathbf{v}_N = \mathbf{v})$ close to $P(\mathbf{W} = \mathbf{v} \mid \mathbf{E} = \mathbf{e})$: probability that $\mathbf{v}_N$ is in the red region is close to $P(A = a \mid \mathbf{E} = \mathbf{e})$.

This does not guarantee that the fraction of samples in $\mathbf{v}_N, \mathbf{v}_{N+1}, \ldots, \mathbf{v}_{N+M}$ that are in the red region yields a good approximation to $P(A = a \mid \mathbf{E} = \mathbf{e})$!



Approximate Inference

Multiple starting points

In practice, one tries to counteract these difficulties by restarting the Gibbs sampling several times (often with different starting points):