

Approximate Inference

Aayush Bajaj

August 31, 2026

exact inference costs $\exp(\text{treewidth})$, and real networks — image grids, dense diagnostic models, unrolled dbns — routinely have treewidth in the hundreds. approximate inference trades guaranteed answers for tractable ones, along two philosophically different roads: **sampling** replaces the distribution with draws from it (anytime, asymptotically exact, embarrassingly parallel), and **approximate message passing** runs exact-style propagation on graphs where its correctness proof no longer holds. this page walks the sampling ladder in the order each rung fixes the last one's failure — forward, rejection, likelihood weighting, gibbs — then closes with loopy belief propagation.

comp9418 benchmarks all four samplers against variable elimination on the icu-alarm network; the qualitative ranking below is what those experiments show

Contents

1 monte carlo estimation	1
2 sampling the prior	2
2.1 forward sampling	2
2.2 rejection sampling	2
3 likelihood weighting	2
4 gibbs sampling	2
4.1 why (and when) it converges	3
5 loopy belief propagation	3
6 choosing an algorithm	4
7 results	4
8 see also	4

1 monte carlo estimation

every sampler below reduces a query to counting. draw N independent samples $\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(N)}$ from P ; estimate any probability as the fraction of samples where the event holds:

$$\hat{P}(\alpha) = \frac{1}{N} \sum_{s=1}^N \mathbf{1}[\mathbf{x}^{(s)} \models \alpha]. \quad (1)$$

the estimator is unbiased and its standard error shrinks as $O(1/\sqrt{N})$ — *independently of dimension*, which is the entire appeal. the constant matters though: rare

events need $N \gg 1/P(\alpha)$ samples before the count is even nonzero. everything that follows is a battle to spend samples where the posterior mass actually is.

2 sampling the prior

2.1 forward sampling

a bayesian network is a recipe for generating worlds. visit variables in topological order (parents first); for each, draw from its cpt row given the already-sampled parent values. drawing from a row is inverse-cdf sampling on a table: draw $r \sim U[0, 1)$, walk the row accumulating probability, return the first outcome where the running total exceeds r .

this primitive — one uniform draw, one walk along a cpt row — is the only randomness in all four algorithms

forward sampling costs $O(n)$ per sample and hits the prior exactly. it answers *prior* queries beautifully and conditional queries not at all — evidence has nowhere to enter.

2.2 rejection sampling

to estimate $P(\alpha \mid \mathbf{e})$: forward-sample, *discard* every sample inconsistent with $\mathbf{e}$, count α among the survivors. correct, and doomed: the acceptance rate is $P(\mathbf{e})$, so for evidence of any specificity — ten binary observations put $P(\mathbf{e})$ around 2^{-10} — a thousand samples are thrown away per survivor. cost per effective sample: $O(n/P(\mathbf{e}))$.

3 likelihood weighting

stop sampling the evidence variables; *fix* them to their observed values and let each sample pay for the privilege. sample non-evidence variables forward as usual, and give the completed sample a **weight** — the probability the evidence would have taken its observed values along the way:

$$w(\mathbf{x}) = \prod_{E \in \mathbf{E}} P(e \mid \mathbf{u}_E), \quad (2)$$

one cpt lookup per evidence variable, parents $\mathbf{u}_E$ as sampled. estimates become weighted counts:

$$\hat{P}(\alpha \mid \mathbf{e}) = \frac{\sum_s w^{(s)} \mathbf{1}[\mathbf{x}^{(s)} \models \alpha]}{\sum_s w^{(s)}}. \quad (3)$$

this is importance sampling with the network’s own dynamics as proposal: no sample is wasted, and on the icu-alarm benchmarks it improves on rejection sampling by orders of magnitude — the workhorse of the family. its failure mode is *weight degeneracy*: evidence deep downstream barely steers the sampling of its ancestors, so most samples describe worlds the evidence all but rules out and a handful of weights dominate the estimate. many evidence variables, or evidence far from the roots, and the effective sample size collapses.

4 gibbs sampling

the fix for degeneracy is to let the evidence influence *every* variable, which requires abandoning independent samples. **gibbs sampling** builds a markov chain whose

states are complete instantiations and whose stationary distribution is the posterior:

1. initialise all non-evidence variables arbitrarily (evidence stays clamped);
2. repeatedly pick a variable X and resample it from $P(X \mid \text{everything else})$ — which by the markov blanket property needs only the factors mentioning X :

$$P(x \mid \text{rest}) \propto \theta_{x|\mathbf{u}} \prod_{Y \in \text{Children}(X)} \theta_{y|\mathbf{u}_Y}, \quad (4)$$

a handful of lookups per step, however large the network.

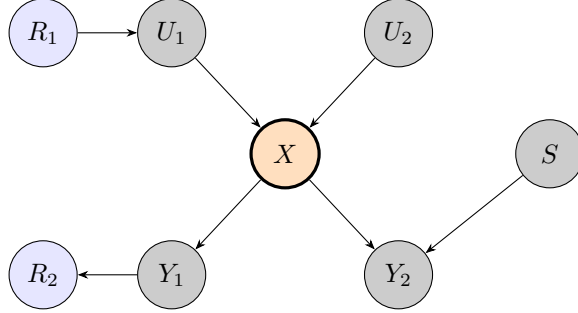


Figure 1: one gibbs step resamples X conditioned only on its markov blanket (shaded): parents, children, and co-parents. everything outside the blanket is irrelevant given it.

4.1 why (and when) it converges

the gibbs chain leaves the posterior invariant by construction; convergence to it additionally needs the chain to be ergodic — irreducible and aperiodic over the instantiations consistent with $\mathbf{e}$. positivity of the distribution suffices; deterministic cpts can strand the chain in disconnected islands. the practical liturgy that follows from mcmc theory:

- **burn-in** — discard the prefix before the chain forgets its initialisation;
- **thinning / autocorrelation** — successive samples are correlated, so N gibbs samples carry fewer than N samples' worth of information;
- **multiple chains** — run several from dispersed starts and compare their estimates; disagreement means none of them has mixed.

empirically (icu-alarm again): gibbs is competitive with likelihood weighting in accuracy and beats it when evidence is deep, but costs far more per sample — every step touches a blanket, and the chain must mix before any of it counts. the general-purpose upgrade is **metropolis–hastings**, which accepts arbitrary proposal moves with probability $\min(1, \text{posterior ratio} \times \text{proposal ratio})$; gibbs is the special case whose proposals are always accepted.

5 loopy belief propagation

the message-passing recursion of the jointree algorithm is defined by local equations that never mention the tree — so run them anyway on a cyclic factor graph: initialise all messages to uniform, update until (hopefully) a fixed point. on trees this is exact; on loopy graphs it is neither guaranteed to converge nor correct when it does, yet

it is often startlingly good — the decoding algorithm of turbo and ldpc codes *is* loopy bp, running within a whisker of the shannon limit. damping the updates and scheduling messages sensibly buys convergence in practice; oscillation is the tell that it won't. the same recursion with gaussian factors in canonical form $(K, \mathbf{h}, g)$ gives gaussian bp, exact on trees and — when it converges — exact in the means even on loops.

the fixed points are extrema of the bethe free energy — the variational view that explains why loopy bp works as well as it does

6 choosing an algorithm

situation	reach for
no evidence, or prior queries	forward sampling
little, likely evidence	likelihood weighting
much or deep evidence, positivity holds	gibbs / metropolis-hastings
huge discrete network, marginals wanted fast	loopy belief propagation
sequential state estimation	particle filtering
treewidth actually small	stop approximating — use the jointree

7 results

- **monte carlo error** — unbiased estimates with standard error $O(1/\sqrt{N})$, dimension-free; rare events need $N \gtrsim 1/P(\alpha)$.
- **forward sampling** — exact prior samples in $O(n)$ via topological-order inverse-cdf draws.
- **rejection sampling** — conditions by discarding; effective cost multiplies by $1/P(\mathbf{e})$.
- **likelihood weighting** — importance sampling with clamped evidence; weight = product of evidence-cpt lookups; degrades as evidence grows deep or numerous.
- **gibbs sampling** — mcmc whose stationary distribution is the posterior; each step needs only the markov blanket; requires ergodicity (positivity suffices), burn-in, and mixing diagnostics.
- **loopy bp** — exact-on-trees message passing iterated on loopy graphs; no general guarantees, excellent in practice, fixed points = bethe free energy extrema.

8 see also

- exact inference — what these methods approximate, and the baseline to benchmark against
- markov chains and hidden markov models — the convergence theory mcmc stands on, and particle filters
- bayesian networks — markov blankets, the object gibbs conditions on
- markov networks — metropolis and simulated annealing as energy-space search