Probabilistic Graphical Models
Part 4: Approximate Inference

Manfred Jaeger

Aalborg University
Basic Sampling
Problem Structure

Input: Evidence \( Y = y \), random variable \( X \), value \( x \in \text{Val}(X) \).
Output: Approximation \( q \) for \( p := P(X = x \mid Y = y) \).

Absolute/Relative Error

For \( p, q \in [0, 1] \): \( q \) is approximation for \( p \) with absolute error \( \leq \epsilon \), if

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

\( q \) is approximation for \( p \) with relative error \( \leq \epsilon \), if

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

- Relative error is not symmetric in \( p \) and \( q \) and not invariant under the transition \( p \rightarrow (1 - p) \), \( q \rightarrow (1 - q) \). Use with care!
- When \( q_1, q_2 \) are approximations for \( p_1, p_2 \) with absolute error \( \leq \epsilon \), then no error bounds follow for \( q_1/q_2 \) as an approximation for \( p_1/p_2 \).
- When \( q_1, q_2 \) are approximations for \( p_1, p_2 \) with relative error \( \leq \epsilon \), then \( q_1/q_2 \) approximates \( p_1/p_2 \) with relative error \( \leq (2\epsilon)/(1 + \epsilon) \).
Observation: can use Bayesian network as random generator that produces full instantiations $X = x$ according to distribution $P(X)$.

Example:

- 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.
Given a sample $x_1, \ldots, x_N$ of complete instantiations generated (independently) by the sampling algorithm, approximate $P(Y = y)$ as

$$q^* := \frac{1}{N} \left| \{i \in 1, \ldots, N \mid Y = y \text{ in } x_i \} \right|$$

Similarly, the sample provides an estimate for $P(X = x, Y = y)$. 

Let $\epsilon, \delta > 0$, $p = P(Y = y)$. From Chebyshev’s inequality:

\[ N \geq \frac{1}{\epsilon^2 \delta} \rightarrow P_N(|q^* - p| \leq \epsilon) \geq 1 - \delta. \]

\[ N \geq \frac{1 - p}{p \epsilon^2 \delta} \rightarrow P_N(|1 - q^*/p| \leq \epsilon) \geq 1 - \delta. \]

$P_N$: probability distribution on $Val(X)^N$ induced by independently sampling $N$ instances. Chebyshev’s inequality: for any real-valued random variable $X$:

\[ P(|X - E[X]| \geq \epsilon) \leq \frac{1}{\epsilon^2} \text{Var}[X] \]

- Sharper bounds can be obtained from Hoeffding and Chernoff inequalities

Problems for approximation with relative error bound:

1) $N$ depends on $p$ – but $p$ is not known!

2) $N$ increases as $1/p$. 
Approximation for $P(X = x \mid Y = y)$: \[ \frac{\# \text{ red circles}}{\# \text{ all circles}} \]

Sample with
- not $Y = y$
- $Y = y, X \neq x$
- $Y = y, X = x$
Problem of forward sampling: samples with $Y \neq y$ are useless!

Goal: find algorithm that samples according to $P(X \mid Y = y)$:
MCMC Sampling
Principle: obtain new sample from previous sample by randomly changing the value of only one selected variable.

Procedure Gibbs sampling
\[ x_0 = (x_{0,1}, \ldots, x_{0,n}) := \text{arbitrary instantiation of } X \text{ with } Y = y \]
\[ t := 1 \]
repeat forever
  choose \( X_k \in X \setminus Y \)
  generate randomly \( x_{t,k} \) according to \( P(X_k \mid X \setminus X_k = x_{t-1} \setminus x_{t-1,k}) \)
  set \( x_{t,j} := x_{t-1,j} \) for \( j \neq k \).
  \( t := t + 1 \)
$P(X_k = x_k \mid X \setminus X_k = x_{t-1} \setminus x_{t-1,k})$

$\approx P(X_k = x_k, X \setminus X_k = x_{t-1} \setminus x_{t-1,k})$

$\approx P(X_k = x_k \mid \text{Pa}(X_k) = (x_{t-1} \uparrow \text{Pa}(X_k))) \cdot \prod_{i : X_k \in \text{Pa}(X_i)} P(X_i = x_{t-1,i} \mid \text{Pa}(X_i) \setminus X_k = (x_{t-1} \uparrow \text{Pa}(X_i) \setminus X_k), X_k)$ (*)

where $\approx$ means: equals up to a constant that does not depend on $X_k$.

To sample $X_{t,k}$:

- evaluate (*) for all $x_k \in \text{Val}(X_k)$
- normalize to obtain $P(X_k \mid X \setminus X_k = x_t \setminus x_{t,k})$
- sample value $x_{t,k}$ according to the resulting distribution
The process of Gibbs sampling can be understood as a random walk in the space of all instantiations with $Y = y$:

Reachable in one step: instantiations that differ from current one by value assignment to at most one variable (assume randomized choice of variable $X_k$).
Goal: want to sample instantiations $\mathbf{x} \in \text{Val}(\mathbf{X})$ from a distribution $\pi(\mathbf{X})$.
(e.g.: $\pi(\mathbf{X}) = P(\mathbf{X} \mid \mathbf{Y} = \mathbf{y})$

Procedure Metropolis Hastings sampling
$\mathbf{x}_0 = (x_{0,1}, \ldots, x_{0,n}) := \text{arbitrary instantiation of } \mathbf{X} \text{ with } \pi(\mathbf{X} = \mathbf{x}_0) > 0$
$t := 1$
repeat forever
    generate $\mathbf{x}_{\text{cand}}$ according to a proposal distribution $Q(\mathbf{X}_t \mid \mathbf{X}_{t-1} = \mathbf{x}_{t-1})$
    compute acceptance probability $p_{\text{acc}} = \min\{1, \frac{\pi(\mathbf{x}_{\text{cand}})Q(\mathbf{X}_t = \mathbf{x}_{t-1} \mid \mathbf{X}_{t-1} = \mathbf{x}_{\text{cand}})}{\pi(\mathbf{x}_{t-1})Q(\mathbf{X}_t = \mathbf{x}_{\text{cand}} \mid \mathbf{X}_{t-1} = \mathbf{x}_{t-1})}\}$
    set $\mathbf{x}_t = \begin{cases} \mathbf{x}_{\text{cand}} & \text{with probability } p_{\text{acc}} \\ \mathbf{x}_{t-1} & \text{with probability } 1 - p_{\text{acc}} \end{cases}$
    $t := t + 1$
Independent sampling from $\pi$

Suppose we can sample from $\pi(X)$. Then M-H sampling with

$$Q(X_t \mid X_{t-1} = x_{t-1}) = \pi(X)$$

produces a series of independent samples from $\pi(X)$.

Gibbs sampling

Let $\pi(X) = P(X \mid Y = y)$, and $Q(X_t \mid X_{t-1} = x_{t-1})$ be defined by resampling a single, randomly selected variable $X_k$. Then

$$\frac{\pi(x_{\text{cand}})}{\pi(x_{t-1})} = \frac{Q(X_t = x_{\text{cand}} \mid X_{t-1} = x_{t-1})}{Q(X_t = x_{t-1} \mid X_{t-1} = x_{\text{cand}})},$$

$p_{\text{acc}} = 1$, and M-H becomes Gibbs sampling.
- Need to design \( Q(X_t \mid X_{t-1} = x_{t-1}) \) so that
  - can effectively sample from \( Q() \)
  - so that acceptance probabilities are not too small

\[
\frac{\pi(x_{cand})}{\pi(x_{t-1})} \approx \frac{Q(X_t = x_{cand} \mid X_{t-1} = x_{t-1})}{Q(X_t = x_{t-1} \mid X_{t-1} = x_{cand})}
\]

- Need to be able to evaluate ratios \( \frac{\pi(x_{cand})}{\pi(x_{t-1})} \)
Two conditions on the Markov Chain $X_0, X_1, \ldots$ defined by M-H sampling:

**Irreducibility**
For all $x, x' \in Val(X)$ with $\pi(x) > 0, \pi(x') > 0$: there exists $k > 0$ with

$$P(X_k = x' \mid X_0 = x) > 0$$

**Aperiodicity**
There exists $x$ with $\pi(x) > 0$, so that for all sufficiently large $k$:

$$P(X_k = x \mid X_0 = x) > 0$$

These are essentially conditions for the proposal distribution!

**Convergence**

If Irreducibility and Aperiodicity hold, then the sampling distribution of $X_t$ converges to $\pi(X)$. 

MCMC Sampling

Estimate probability $\pi(A) \ A \subseteq \text{Val}(X)$: (special case: $\pi(X) = P(X | Y = y), A = \{X = x\}$)

1. Find initial $x$ with $\pi(x) > 0$
2. **Burn in**: Perform $k$ steps of M-H sampling
3. Sample values $x_{k+1}, x_{k+2}, \ldots, x_{k+N}$
4. Estimate:

$$\hat{\pi}(A) = \frac{1}{N} | \{i \mid x_{k+i} \in A\} |$$

Issues:

- How large needs $k$ to be, so that $P(X_k)$ is close to $\pi(X)$?
- The $x_{k+1}, x_{k+2}, \ldots, x_{k+N}$ are not independent. Accuracy of estimate $\hat{\pi}(A)$ very difficult to determine
Markov chain $x_0, x_1, \ldots$ as random walk in $Val(X)$:

- red: states $x$ with $x \in A$
- green: states $x$ with $x \not\in A$
\( P(\mathbf{X}_{k+i}) \) close to \( \pi(\mathbf{X}) \): probability that \( \mathbf{x}_{k+i} \) is in the red region is close to \( \pi(\mathbf{A}) \).

This does not guarantee that \( 1/N | \{ i \mid \mathbf{x}_{k+i} \in \mathbf{A} \} | \) yields a good approximation for \( \pi(\mathbf{A}) \)!
In practice, one tries to counteract these difficulties by running multiple Markov chains: