Approximate Inference

Probabilistic Graphical Models

Jerónimo Hernández-González

Outline

Approximate Inference

Background

The **variable elimination** and the **clique tree** algorithms can be used to answer queries to a PGM

Their complexity is exponential in the width of the induced graph

What can we do when it is so large?

Approximate Inference

Background

The **variable elimination** and the **clique tree** algorithms can be used to answer queries to a PGM

Their complexity is exponential in the width of the induced graph

What can we do when it is so large?

Approximate inference!

Approximate inference

Alternatives

Sampling:

- Forward sampling
- MCMC
- Gibbs sampling

Optimization:

- Loopy Belief Propagation (previous class)
- Expectation Propagation
- Variational approaches (short intro in the following class)

Approximate inference

Alternatives

Sampling:

- Forward sampling
- MCMC
- Gibbs sampling

Optimization:

- Loopy Belief Propagation (previous class)
- Expectation Propagation
- Variational approaches (short intro in the following class)

Marginal distribution, $P(X_i)$

Given a sample $D = \{\boldsymbol{x}^1, \ldots, \boldsymbol{x}^M\}$ from a probability distribution $P(\boldsymbol{X})$, where $\boldsymbol{X} = (X_1)$ and X_1 is a Bernoulli (0-1) variable.

Find an estimate for $P(X_1 = 1)$

Marginal distribution, $P(X_i)$

Given a sample $D = \{\boldsymbol{x}^1, \dots, \boldsymbol{x}^M\}$ from a probability distribution $P(\boldsymbol{X})$, where $\boldsymbol{X} = (X_1)$ and X_1 is a Bernoulli (0-1) variable.

Find an estimate for $P(X_1 = 1)$

$$P(X_1 = 1) \simeq \frac{\sum_{\boldsymbol{x}^i \in D} \delta[x_1^i = 1]}{M}$$

Marginal distribution, $P(X_i)$

Given a sample $D = \{\boldsymbol{x}^1, \ldots, \boldsymbol{x}^M\}$ from a probability distribution $P(\boldsymbol{X})$, where $\boldsymbol{X} = (X_1, X_2, X_3)$ and all X_i are Bernoulli (0-1) variables.

Find an estimate for $P(X_1 = 1)$

Marginal distribution, $P(X_i)$

Given a sample $D = \{\boldsymbol{x}^1, \ldots, \boldsymbol{x}^M\}$ from a probability distribution $P(\boldsymbol{X})$, where $\boldsymbol{X} = (X_1, X_2, X_3)$ and all X_i are Bernoulli (0-1) variables.

Find an estimate for $P(X_1 = 1)$

$$P(X_1 = 1) \simeq \frac{\sum_{\boldsymbol{x}^i \in D} \delta[x_1^i = 1]}{M}$$

Marginal distribution, $P(X_i)$

Given a sample $D = \{\mathbf{x}^1, \ldots, \mathbf{x}^M\}$ from a probability distribution $P(\mathbf{X})$, where $\mathbf{X} = (X_1, X_2, X_3)$ and all X_i are Bernoulli (0-1) variables.

Find an estimate for $P(X_1 = 1)$

$$P(X_1 = 1) \simeq \frac{\sum_{\mathbf{x}^i \in D} \delta[x_1^i = 1]}{M}$$

The larger the sample size M, the better the estimation

Marginal distribution, $P(\boldsymbol{Y})$

Given a sample $D = \{\boldsymbol{x}^1, \ldots, \boldsymbol{x}^M\}$ from a distribution $P(\boldsymbol{X})$, where $\boldsymbol{X} = (X_1, \ldots, X_d)$ and all X_i are Bernoulli (0-1) variables,

and a subset of variables of interest $\boldsymbol{Y} \subseteq \boldsymbol{X}$

Find an estimate for $P(\boldsymbol{Y} = \boldsymbol{y})$

Marginal distribution, $P(\boldsymbol{Y})$

Given a sample $D = \{\boldsymbol{x}^1, \ldots, \boldsymbol{x}^M\}$ from a distribution $P(\boldsymbol{X})$, where $\boldsymbol{X} = (X_1, \ldots, X_d)$ and all X_i are Bernoulli (0-1) variables,

and a subset of variables of interest $\boldsymbol{Y} \subseteq \boldsymbol{X}$

Find an estimate for $P(\boldsymbol{Y} = \boldsymbol{y})$

$$P(\boldsymbol{Y} = \boldsymbol{y}) \simeq \frac{\sum_{\boldsymbol{x}^i \in D} \delta[\boldsymbol{y}^i = \boldsymbol{y}]}{M}$$

Conditional distribution, $P(\boldsymbol{Y}|\boldsymbol{E}=\boldsymbol{e})$

Given a sample $D = \{\boldsymbol{x}^1, \ldots, \boldsymbol{x}^M\}$ from a distribution $P(\boldsymbol{X})$, where $\boldsymbol{X} = (X_1, \ldots, X_d)$ and all X_i are Bernoulli (0-1) variables,

and two disjoint subsets of variables $\boldsymbol{Y} \subseteq \boldsymbol{X}$, $\boldsymbol{E} \subseteq \boldsymbol{X}$ $(\boldsymbol{Y} \cap \boldsymbol{E} = \emptyset)$.

Find an estimate for $P(\boldsymbol{Y}=\boldsymbol{y}|\boldsymbol{E}=\boldsymbol{e})$

Approximate inference

Solving probabilistic queries by samples

Conditional distribution, $P(\boldsymbol{Y}|\boldsymbol{E} = \boldsymbol{e})$

Given a sample $D = \{\boldsymbol{x}^1, \ldots, \boldsymbol{x}^M\}$ from a distribution $P(\boldsymbol{X})$, where $\boldsymbol{X} = (X_1, \ldots, X_d)$ and all X_i are Bernoulli (0-1) variables,

and two disjoint subsets of variables $\boldsymbol{Y} \subseteq \boldsymbol{X}$, $\boldsymbol{E} \subseteq \boldsymbol{X}$ ($\boldsymbol{Y} \cap \boldsymbol{E} = \emptyset$).

Find an estimate for $P(\boldsymbol{Y} = \boldsymbol{y}|\boldsymbol{E} = \boldsymbol{e})$

$$P(\boldsymbol{Y} = \boldsymbol{y}|\boldsymbol{E} = \boldsymbol{e}) \simeq \frac{\sum_{\boldsymbol{x}^i \in D} \delta[\boldsymbol{y}^i = \boldsymbol{y} \wedge \boldsymbol{e}^i = \boldsymbol{e}]}{M'},$$

where $M' = \sum_{\boldsymbol{x}^i \in D} \delta[\boldsymbol{e}^i = \boldsymbol{e}]$

Good news!

If we have a sample from $P(\boldsymbol{X})$, we can solve queries approximately!

Approximate Inference

Probabilistic Graphical Models

Jerónimo Hernández-González

Good news!

If we have a sample from $P(\boldsymbol{X})$, we can solve queries approximately!

Close the loop

We know how to answer queries given a sample...

but **we don't have a sample** yet!

How can we obtain a sample from our PGM?

Sample from a categorical distribution

- Variable $X \sim Cat(p_1, \ldots, p_{|\Omega_X|})$
- $\Omega_X = \{x_1, \ldots, x_{|\Omega_X|}\}$: set of possible values of X

Unequal roulette

Sample from a categorical distribution

- Variable $X \sim Cat(p_1, \ldots, p_{|\Omega_X|})$
- $\Omega_X = \{x_1, \ldots, x_{|\Omega_X|}\}$: set of possible values of X

Unequal roulette

Sample x' from $p(X)$ as:

1. Obtain sample from a uniform dist.: $v \sim \mathcal{U}[0, 1]$
2. $$x' = x_j : \text{smallest } j \;\text{ s.t. } \left(\sum_{i=1}^{j} p_i\right) > v$$

** It is the same when sampling $p(\boldsymbol{Y}|\boldsymbol{Z} = \boldsymbol{z})$

Exercise

Sampling a categorical distribution

Sample from $p(X)$, where $\Omega_X = \{a, b, c, d\}$ and:

x_i	$p(x_i)$	$\sum_{i=1}^{j} p_i$
a	0.15	
b	0.24	
c	0.47	
d	0.14	

given the following sample from the uniform distribution $[0, 1]$:

$$\{0.61,\ 0.95,\ 0.13,\ 0.88,\ 0.34,\ 0.25,\ 0.23,\ 0.97,\ 0.74,\ 0.12\}$$

Sampling-based Approximate Inference

Forward sampling from a Bayesian network

Bayesian networks (revisited)

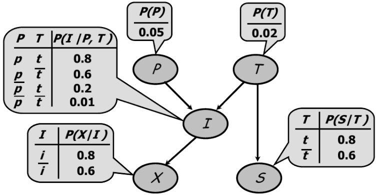
Given a $P_{\mathcal{G}}(\boldsymbol{X})$, each X_i follows a CPD where the conditioning variables are the parents of X_i in $\mathcal{G}$.
A CPD is really a family of distributions: $\sum_{x_i} P(X_i = x_i | \boldsymbol{pa}_i) = 1, \forall \boldsymbol{pa}_i$.

To sample X_i, we pick a single distribution from the family. Which?

Forward sampling

Sample one variable X_i at a time, following an ancestral ordering.

1. Thus, we know that, when it is the turn of X_i –with CPD $p(X_i|\boldsymbol{PA}_i)$– we have already sampled the parents, $\boldsymbol{pa}_i$
2. The distribution we need to sample is $X_i \sim p(X_i|\boldsymbol{PA}_i = \boldsymbol{pa}_i)$



Forward sampling

Idea: Obtain the values $\boldsymbol{pa}_i$ for parent variables **before** sampling the conditional distribution $p(X_i|\boldsymbol{pa}_i)$.

Following an ancestral ordering we know that, everytime that we pick a X_i, we already visited all $\boldsymbol{PA}_i$.

I.e., we have a $\boldsymbol{pa}_i \rightarrow$ we know which distribution to pick from the CPD (family) of X_i.

So, a sample $\boldsymbol{x} = (x_1, ..., x_d)$ is obtained in d steps:

sampling one-at-a-time the d distributions, $P(X_i|\boldsymbol{PA}_i)$

Sampling-based Approximate Inference

Forward sampling from a Bayesian network

procedure ForwardSample($\mathcal{B}$, $(X_1, \ldots, X_d)$) ▷ $\mathcal{B}$ is a BN
▷ $(X_1, \ldots, X_d)$ follows an ancestral ord. w.r.t. $\mathcal{B}$

  for $i \in \{1, \ldots, d\}$ **do**

    $\boldsymbol{pa}_i \leftarrow \boldsymbol{x}[\boldsymbol{PA}_i]$

    $x' \sim P(X_i | \boldsymbol{pa}_i)$ ▷ Sample the CPD of X_i

    $\boldsymbol{x}[X_i] \leftarrow x'$ ▷ x_i (or $\boldsymbol{x}[X_i]$) is the sampled value

  end for

  return $\boldsymbol{x} = (x_1, \ldots, x_d)$

end procedure

Sampling-based Approximate Inference

Forward sampling from a Bayesian network

procedure ForwardSample($\mathcal{B}$, $(X_1, \ldots, X_d)$) $\triangleright$ $\mathcal{B}$ is a BN
$\triangleright$ $(X_1, \ldots, X_d)$ follows an ancestral ord. w.r.t. $\mathcal{B}$
  for $i \in \{1, \ldots, d\}$ **do**
    $\boldsymbol{pa}_i \leftarrow \boldsymbol{x}[\boldsymbol{PA}_i]$
    $x' \sim P(X_i | \boldsymbol{pa}_i)$ $\triangleright$ Sample the CPD of X_i
    $\boldsymbol{x}[X_i] \leftarrow x'$ $\triangleright$ x_i (or $\boldsymbol{x}[X_i]$) is the sampled value
  end for
  return $\boldsymbol{x} = (x_1, \ldots, x_d)$
end procedure

Close the loop for BNs

We know how to answer queries given a sample...

and we know how to obtain a sample from our BN!

Marginal Probability Query

Given a Bayesian network, $\mathcal{B}$, for $\boldsymbol{X} = (X_1, \ldots, X_n)$, and a subset of variables $\boldsymbol{Y} \subseteq \boldsymbol{X}$.

Find an estimate for $P(\boldsymbol{Y} = \boldsymbol{y})$:

Procedure:

1. Generate samples from the BN, $\mathcal{B}$
2. Compute the fraction of samples where $\boldsymbol{Y} = \boldsymbol{y}$

Conditional Probability Query

Given a Bayesian network, $\mathcal{B}$, for $\boldsymbol{X} = (X_1, \ldots, X_n)$, s.t. $\boldsymbol{X} = (\boldsymbol{Y}, \boldsymbol{H}, \boldsymbol{E})$.

Find an estimate for $P(\boldsymbol{Y} = \boldsymbol{y} | \boldsymbol{E} = \boldsymbol{e})$:

Procedure:

1. Generate M samples from the BN, $\mathcal{B}$
2. Keep only the M' samples where $\boldsymbol{E} = \boldsymbol{e}$, $\quad (M' < M)$
3. Compute the fraction (over M') of samples where $\boldsymbol{Y} = \boldsymbol{y}$

Exercise

Rejecting Samples

Consider the regular process of rejecting samples which do not match $\boldsymbol{e}$ for the posterior distribution $P(\boldsymbol{Y}|\boldsymbol{e})$.
If we want to obtain M' samples, which is the expected number of samples M that would need to be drawn from $P(\boldsymbol{X})$?

a) $M' \cdot P(\boldsymbol{e})$
b) $M' \cdot P(\boldsymbol{Y}|\boldsymbol{e})$
c) $M' \cdot (1 - P(\boldsymbol{e}))$
d) $M' \cdot (1 - P(\boldsymbol{Y}|\boldsymbol{e}))$
e) $M'/P(\boldsymbol{e})$
f) $M'/(1 - P(\boldsymbol{e}))$

Sampling-based Approximate Inference

Forward sampling for answering CPQs

Conditional Probability Query

Given a Bayesian network, $\mathcal{B}$, for $\boldsymbol{X} = (X_1, \ldots, X_n)$, s.t. $\boldsymbol{X} = (\boldsymbol{Y}, \boldsymbol{H}, \boldsymbol{E})$.

Find an estimate for $P(\boldsymbol{Y} = \boldsymbol{y} | \boldsymbol{E} = \boldsymbol{e})$:

Procedure:

1. Generate M samples from the BN, $\mathcal{B}$
2. Keep only the (M') samples where $\boldsymbol{E} = \boldsymbol{e}$
3. Compute the fraction of samples where $\boldsymbol{Y} = \boldsymbol{y}$

Computational issue!

The number of samples (M) required for a *good approximation* grows exponentially with $|\boldsymbol{E}|$ (no. observed variables)!!

Sampling-based Approximate Inference

Likelihood weighted *sampling for answering CPQs*

Conditional Probability Query

Given a Bayesian network, $\mathcal{B}$, for $\boldsymbol{X} = (X_1, \ldots, X_n)$, with $\boldsymbol{X} = (\boldsymbol{Y}, \boldsymbol{H}, \boldsymbol{E})$,

find an estimate for $P(\boldsymbol{Y} = \boldsymbol{y} | \boldsymbol{E} = \boldsymbol{e})$

Procedure:

1. Generate M weighted samples, $WS = \left\{(\boldsymbol{x}^i, w^i)\right\}_{i=1}^{M}$, that concur with $\boldsymbol{E} = \boldsymbol{e}$ from the BN, $\mathcal{B}$
2. Compute the weighted ratio of samples satisfying $\boldsymbol{Y} = \boldsymbol{y}$

$$P(\boldsymbol{Y} = \boldsymbol{y} | \boldsymbol{E} = \boldsymbol{e}) \simeq \frac{\sum_{(\boldsymbol{x}^i, w^i) \in WS} w^i \cdot \delta[\mathbf{y}^i = \boldsymbol{y}]}{\sum_{(\boldsymbol{x}^i, w^i) \in WS} w^i}$$

Sampling-based Approximate Inference

Likelihood weighted sampling for answering CPQs

procedure LW-Sample($\mathcal{B}$, $\boldsymbol{E} = \boldsymbol{e}$, $(X_1, \ldots, X_d)$) $\triangleright$ $\mathcal{B}$ is a BN
$\triangleright$ $(X_1, \ldots, X_d)$ follows an ancestral ord. w.r.t. $\mathcal{B}$

    $w \leftarrow 1$
    for $i \in \{1, \ldots, d\}$ **do**
        $\boldsymbol{pa}_i \leftarrow \boldsymbol{x}[\boldsymbol{PA}_i]$
        if $X_i \notin \boldsymbol{E}$ **then**
            $x' \sim P(X_i|\boldsymbol{pa}_i)$ $\triangleright$ Sample the CPD of X_i
            $\boldsymbol{x}[X_i] \leftarrow x'$ $\triangleright$ x_i (or $\boldsymbol{x}[X_i]$) is the sampled value
        else
            $\boldsymbol{x}[X_i] \leftarrow \boldsymbol{e}[X_i]$ $\triangleright$ x_i is the observed value e_i
            $w \leftarrow w \cdot P(x_i|\boldsymbol{pa}_i)$
            $\triangleright$ the weight w considers the plausibility of e_i in $\boldsymbol{x}$
        end if
    end for
    return $\boldsymbol{x} = (x_1, \ldots, x_d), w$
end procedure

Approximate Inference

Probabilistic Graphical Models

Jerónimo Hernández-González

Conditional Probability Query

Given a Markov network, $\mathcal{H}$, for $\boldsymbol{X} = (X_1, \ldots, X_n)$, with $\boldsymbol{X} = (\boldsymbol{Y}, \boldsymbol{H}, \boldsymbol{E})$,

find an estimate for $P(\boldsymbol{Y} = \boldsymbol{y} | \boldsymbol{E} = \boldsymbol{e})$

Procedure:

1. Generate samples from the MN, $\mathcal{H}$ How?!
2. Remove the samples where $\boldsymbol{E} \neq \boldsymbol{e}$
3. Compute the fraction of samples where $\boldsymbol{Y} = \boldsymbol{y}$

Gibbs Sampling

procedure Gibbs-Sample($P^{(0)}(\boldsymbol{X})$, Φ, M) $\triangleright$ $\mathcal{H}$: PGM; $P^{(0)}(\boldsymbol{X})$: initial state distr.

  $\boldsymbol{x}^0 \sim P^{(0)}(\boldsymbol{X})$

  for $i \in \{1, \ldots, M\}$ **do**

    $\boldsymbol{x}^i \leftarrow \boldsymbol{x}^{i-1}$

    for $X_j \in \boldsymbol{X}$ **do**

      $x' \sim P_{\mathcal{H}}(X_j | \boldsymbol{x}^i_{-j})$

      $x^i_j \leftarrow x'$

    end for

  end for

  return $\{\boldsymbol{x}^0, \ldots, \boldsymbol{x}^M\}$

end procedure

Gibbs Sampling

$P(X_j = x_j | \boldsymbol{x}_{-j})$ for a Markov Network, $\mathcal{H}$

Given that Φ is the set of factors of $\mathcal{H}$, one can sample x_j from:

$$
\begin{aligned}
P(X_j = x_j | \boldsymbol{x}_{-j}) &= \frac{\frac{1}{Z}\widetilde{P}(x_j, \boldsymbol{x}_{-j})}{\sum_{x_j'} \frac{1}{Z}\widetilde{P}(x_{j'}, \boldsymbol{x}_{-j})} = \frac{\frac{1}{Z}\widetilde{P}(x_j, \boldsymbol{x}_{-j})}{\frac{1}{Z}\sum_{x_j'} \widetilde{P}(x_{j'}, \boldsymbol{x}_{-j})} = \\
&= \frac{\prod_{\phi \in \Phi} \phi[x_j, \boldsymbol{x}_{-j}]}{\sum_{x_j'} \prod_{\phi \in \Phi} \phi[x_j', \boldsymbol{x}_{-j}]} = \\
&= \frac{\prod_{\phi \in \Phi : X_j \in Scope(\phi)} \phi(x_j, \boldsymbol{x}_{-j})}{\sum_{x_j'} \prod_{\phi \in \Phi : X_j \in Scope(\phi)} \phi(x_j', \boldsymbol{x}_{-j})} \\
&\propto \prod_{\phi \in \Phi : X_j \in Scope(\phi)} \phi[x_j, \boldsymbol{x}_{-j}]
\end{aligned}
$$

**This is the Markov blanket!

Conditional Probability Query (Revisited)

Given a Markov network, $\mathcal{H}$, for $\boldsymbol{X} = (X_1, \ldots, X_n)$, with $\boldsymbol{X} = (\boldsymbol{Y}, \boldsymbol{H}, \boldsymbol{E})$,

find an estimate for $P(\boldsymbol{Y} = \boldsymbol{y} | \boldsymbol{E} = \boldsymbol{e})$

Procedure (Gibbs sampling):

1. Generate samples from the MN, $\mathcal{H}$ How?!
   - Use Gibbs Sampling as indicated,
     $$\hat{x}_j^{(t)} \sim P_{\mathcal{H}}(X_j | \boldsymbol{x}_{-j}^{(t)}), \quad \forall j \in \{1, \ldots, d\}$$
     to generate samples from $P_{\mathcal{H}}$
2. Remove the samples where $\boldsymbol{E} \neq \boldsymbol{e}$
3. Compute the fraction of samples where $\boldsymbol{Y} = \boldsymbol{y}$

Exercise

Gibbs Sampling on Bayesian networks

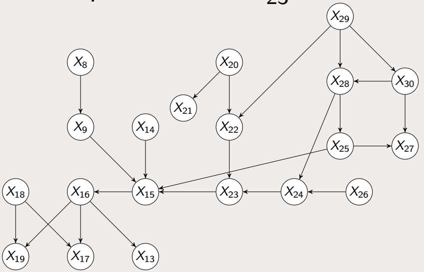
We run Gibbs sampling on the chain BN $X \to Y \to Z$.
If the current sample is (x_0, y_0, z_0) and we now sample Y, which is the probability that the next sample is (x_0, y_1, z_0)?

a) $P(x_0, y_1, z_0)$
b) $P(x_0, z_0 | y_1)$
c) $P(y_1 | x_0, z_0)$
d) $P(y_1 | x_0)$

Exercise

Gibbs Sampling on Bayesian networks

If we are sampling the variable X_{23} of this BN as a substep of Gibbs sampling, which is the closed form for the distribution that we use to sample the value x'_{23}?



a) $P(x'_{23}|x_{22}, x_{24})$

b) $P(x'_{23}|x_{22}, x_{24}) \cdot P(x_{15}|x'_{23}, x_{14}, x_9, x_{25})$

c) $P(x'_{23}|\boldsymbol{x}_{-23})$ where $\boldsymbol{x}_{-23}$ is the tuple of values for all the rest of variables but X_{23}

d) None, as these are all either incorrect or not in closed form

e) $\dfrac{P(x'_{23}|x_{22},x_{24})P(x_{15}|x'_{23},x_{14},x_9,x_{25})}{\sum_{x''_{22},x''_{24},x''_{15},x''_{14},x''_9,x''_{25}} P(x'_{23}|x''_{22},x''_{24})P(x''_{15}|x'_{23},x''_{14},x''_9,x''_{25})}$

f) $\dfrac{P(x'_{23}|x_{22},x_{24})P(x_{15}|x'_{23},x_{14},x_9,x_{25})}{\sum_{x''_{23}} P(x''_{23}|x_{22},x_{24})P(x_{15}|x''_{23},x_{14},x_9,x_{25})}$

Summary

General insights

Given a distribution P,

- If P can be sampled, we can answer queries with the empirical distribution
- If not, we can try to build a Markov chain that converges to P, and then run it to get samples (MCMC)
- If P is multivariate, Gibbs sampling (a type of MCMC) sequentially samples each variable from its conditional distr.

Practical ideas

- For sampling a Bayesian network: Forward sampling (Easy!!)
- For sampling a Markov network: Gibbs sampling

(if all factors are positive, the Gibbs chain is guaranteed to be regular)

Approximate Inference

Probabilistic Graphical Models

Jerónimo Hernández-González

Hoeffding bound

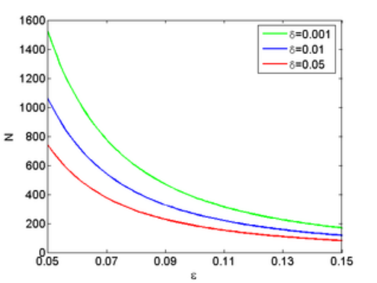
- Actual unknown probabilistic (Bernoulli) distribution $P(\boldsymbol{y})$
- N **independent** trials $D \sim P(\boldsymbol{y})$
- Bound:

$$P_D(\hat{P}_D(\boldsymbol{y}) \notin [P(\boldsymbol{y}) - \epsilon, P(\boldsymbol{y}) + \epsilon]) \leq 2e^{-2N\epsilon^2}$$

- Probability δ of obtaining an estimator with error bound ϵ:

$$2e^{-2N\epsilon^2} \leq \delta$$

$$N \geq \frac{\log(2/\delta)}{2\epsilon^2}$$



Exercise

Forward Sampling

An alternative strategy for obtaining an estimate of the conditional probability $P(\boldsymbol{y}|\boldsymbol{e})$ is by using forward sampling to estimate $P(\boldsymbol{y}, \boldsymbol{e})$ and $P(\boldsymbol{e})$ separately, and then computing the ratio. We can use the Hoeffding Bound to obtain a bound on both the numerator and the denominator. Assuming that M is large, when does the resulting bound provide meaningful guarantees? Recall that we need $M \leq \frac{\log(2/\delta)}{2\epsilon^2}$ to get an additive error bound ϵ that holds with probability $(1-\delta)$ for our estimate.

a) It always provides meaningful guarantees.

b) It never provides a meaningful guarantee.

c) It provides a meaningful guarantee, but only when δ is small relative to $P(\boldsymbol{y}, \boldsymbol{e})$ and $P(\boldsymbol{e})$.

d) It provides a meaningful guarantee, but only when ϵ is small relative to $P(\boldsymbol{y}, \boldsymbol{e})$ and $P(\boldsymbol{e})$.

Approximate Inference

Probabilistic Graphical Models

Jerónimo Hernández-González

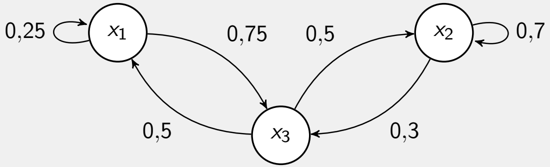
Markov Chain

Definition

A Markov Chain defines a probabilistic transition model over states x,

$$\mathcal{T}(x \to x')$$

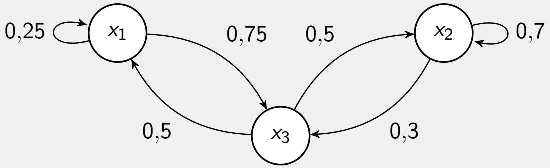
such that $\sum_{x'} \mathcal{T}(x \to x') = 1$



In our case, the states x are the possible values of a random variable X: $x \in \Omega_X$

Markov Chain

Stationary distribution



We look for a probability distribution $P(x) = \pi(x)$ that is stable under the transition of the MC.

What does it mean *stable*? That for all x':

$$\pi(x') = \sum_x \pi(x) \mathcal{T}(x \to x')$$

Exercise

Stationary Distributions

On this Markov chain, by definition, which properties must satisfy a stationary distribution π?

a) $\pi(x_3) = 0{,}4\pi(x_1) + 0{,}75\pi(x_2)$

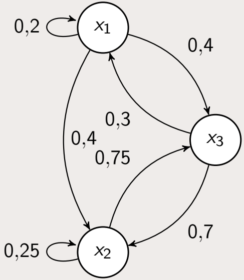
b) $\pi(x_1) = 0{,}2\pi(x_1) + 0{,}4\pi(x_2) + 0{,}4\pi(x_3)$

c) $\pi(x_1) = 0{,}2\pi(x_1) + 0{,}3\pi(x_3)$

d) $\pi(x_1) = \pi(x_2) = \pi(x_3)$

e) $\pi(x_1) + \pi(x_2) + \pi(x_3) = 1$

f) $\pi(x_2) = 0{,}25\pi(x_2) + 0{,}75\pi(x_3)$



Markov Chain

Regularity

A MC is regular if, for every two states (x, x'), there exists a number S such that the probability of $x \to x'$ in S steps is positive

Sufficient (but not necessary) conditions for regularity:

1. Every two states are connected (positive probability path)
2. Every state has a self-transition loop

Property

A regular MC **converges** to a unique stationary distribution regardless of the start state

Markov Chain

Regularity

A MC is regular if, for every two states (x, x'), there exists a number S such that the probability of $x \to x'$ in S steps is positive

Sufficient (but not necessary) conditions for regularity:

1. Every two states are connected (positive probability path)
2. Every state has a self-transition loop

Property

A regular MC **converges** to a unique stationary distribution regardless of the start state

What does it mean **convergence** in this context?

Sampling-based Approximate Inference

MCMC for answering CPQs

procedure MCMC-Sample($P^{(0)}(\boldsymbol{X})$, $\mathcal{T}$, M)
$\triangleright$ $\mathcal{T}$: transition model; $P^{(0)}(\boldsymbol{X})$: initial state distr.
  $\boldsymbol{x}^0 \sim P^{(0)}(\boldsymbol{X})$
  for $i \in \{1, \ldots, M\}$ **do**
    $\boldsymbol{x}^i \sim \mathcal{T}(\boldsymbol{x}^{i-1} \rightarrow \boldsymbol{X})$
  end for
  return $\{\boldsymbol{x}^0, \ldots, \boldsymbol{x}^M\}$
end procedure

Sampling-based Approximate Inference

MCMC for answering CPQs

procedure MCMC-Sample($P^{(0)}(\boldsymbol{X})$, $\mathcal{T}$, M)
$\quad\triangleright$ $\mathcal{T}$: transition model; $P^{(0)}(\boldsymbol{X})$: initial state distr.
$\quad\boldsymbol{x}^0 \sim P^{(0)}(\boldsymbol{X})$
$\quad$**for** $i \in \{1, \ldots, M\}$ **do**
$\quad\quad\boldsymbol{x}^i \sim \mathcal{T}(\boldsymbol{x}^{i-1} \to \boldsymbol{X})$
$\quad$**end for**
$\quad$**return** $\{\boldsymbol{x}^0, \ldots, \boldsymbol{x}^M\}$
end procedure

Convergence

Convergence means that the probability of observing a specific $\boldsymbol{x}$ stabilizes after a number of samples M

Markov Chain

A tool for answering queries

Conditional Probability Query

Given an unknown (and unsampleable) distribution P for $\boldsymbol{X} = (X_1, \ldots, X_n)$.

Find an estimate for $P(\boldsymbol{X}_1 = \boldsymbol{x}_1)$

The **MCMC** approach:

1. Try to build a Markov Chain which has P as stationary distribution
2. Sample the MC
3. Remove the first b samples (*burn-in*)
4. Compute the fraction of remaining samples where $\boldsymbol{X}_1 = \boldsymbol{x}_1$

Conditional Probability Query (Revisited)

Given a Markov network, $\mathcal{H}$, for $\boldsymbol{X} = (X_1, \ldots, X_n)$, with $\boldsymbol{X} = (\boldsymbol{Y}, \boldsymbol{H}, \boldsymbol{E})$,

find an estimate for $P(\boldsymbol{Y} = \boldsymbol{y} | \boldsymbol{E} = \boldsymbol{e})$

Procedure (MCMC):

1. Generate samples from the MN, $\mathcal{H}$ How?!
   - Build a Markov Chain which has $P_{\mathcal{H}}$ as stationary distribution,

     $$P_{\mathcal{H}} \approx \pi_{MC}$$

   - Use the MC to generate samples from $P_{\mathcal{H}}$
2. Remove the samples where $\boldsymbol{E} \neq \boldsymbol{e}$
3. Compute the fraction of samples where $\boldsymbol{Y} = \boldsymbol{y}$

Conditional Probability Query (Revisited)

Given a Markov network, $\mathcal{H}$, for $\boldsymbol{X} = (X_1, \ldots, X_n)$, with $\boldsymbol{X} = (\boldsymbol{Y}, \boldsymbol{H}, \boldsymbol{E})$,

find an estimate for $P(\boldsymbol{Y} = \boldsymbol{y} | \boldsymbol{E} = \boldsymbol{e})$

Procedure (MCMC):

1. Generate samples from the MN, $\mathcal{H}$ How?!
   - Build a Markov Chain which has $P_{\mathcal{H}}$ as stationary distribution,

     ** Note that $P_{\mathcal{H}}$ factorizes!!

     $$P_{\mathcal{H}} \approx \pi_{MC}$$
   - Use the MC to generate samples from $P_{\mathcal{H}}$
2. Remove the samples where $\boldsymbol{E} \neq \boldsymbol{e}$
3. Compute the fraction of samples where $\boldsymbol{Y} = \boldsymbol{y}$

Gibbs chain

Definition

Let $P(X)$ be a distribution that factorizes as

$$P(\boldsymbol{X}) = \frac{1}{Z} \prod_{\phi \in \boldsymbol{\Phi}} \phi$$

Build a Markov Chain such that:

- MC state space: complete assignments $\boldsymbol{x}$ to $\boldsymbol{X}$
- Per variable transition model:

$$\mathcal{T}(\boldsymbol{x}^t \rightarrow \boldsymbol{X}) \sim P(X_i | \boldsymbol{x}_{-i})$$