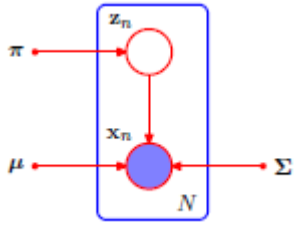
Particle-Based Approximate Inference on Graphical Model

Reference:

Probabilistic Graphical Model Ch. 12 (Koller & Friedman)
CMU, 10-708, Fall 2009 Probabilistic Graphical Models Lectures 18,19 (Eric Xing)
Pattern Recognition & Machine Learning Ch. 11. (Bishop)

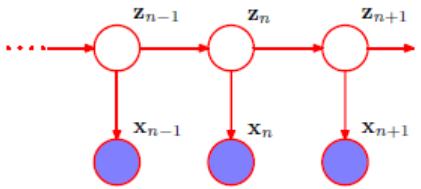
In terms of difficulty, there are 3 types of inference problem.

- Inference which is easily solved with Bayes rule.



- Inference which is tractable using some dynamic programming technique.

  (e.g. Variable Elimination or J-tree algorithm)

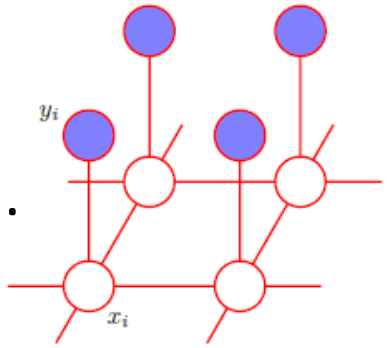


Today's focus

- Inference which is proved intractable

  & should be solved using some Approximate Method.

  (e.g. Approximation with Optimization or Sampling technique.)



Agenda

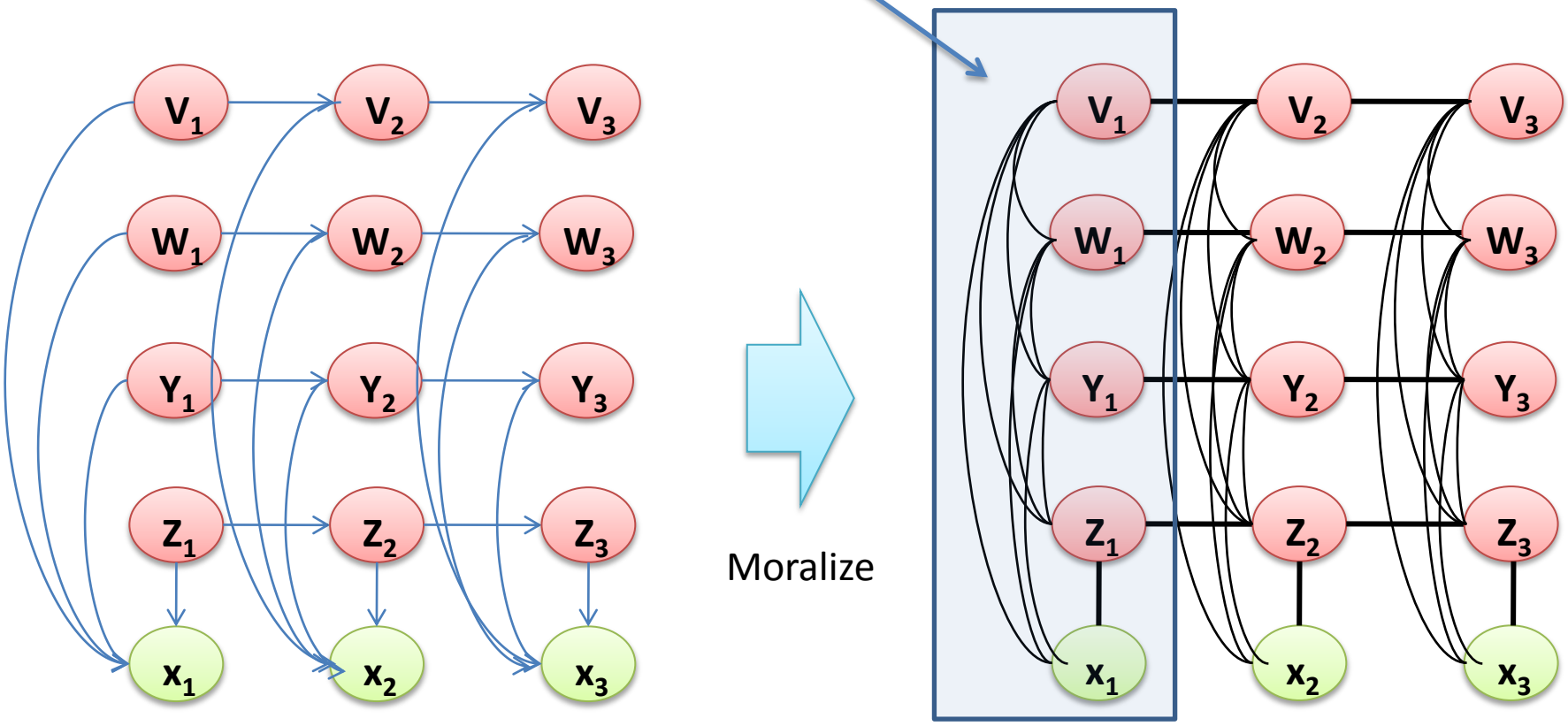
- When to use Particle-Based Approximate Inference ?
- Forward Sampling & Importance Sampling
- Markov Chain Monte Carlo (MCMC)
- Collapsed Particles

Agenda

- When to use Particle-Based Approximate Inference ?
- Forward Sampling & Importance Sampling
- Markov Chain Monte Carlo (MCMC)
- Collapsed Particles

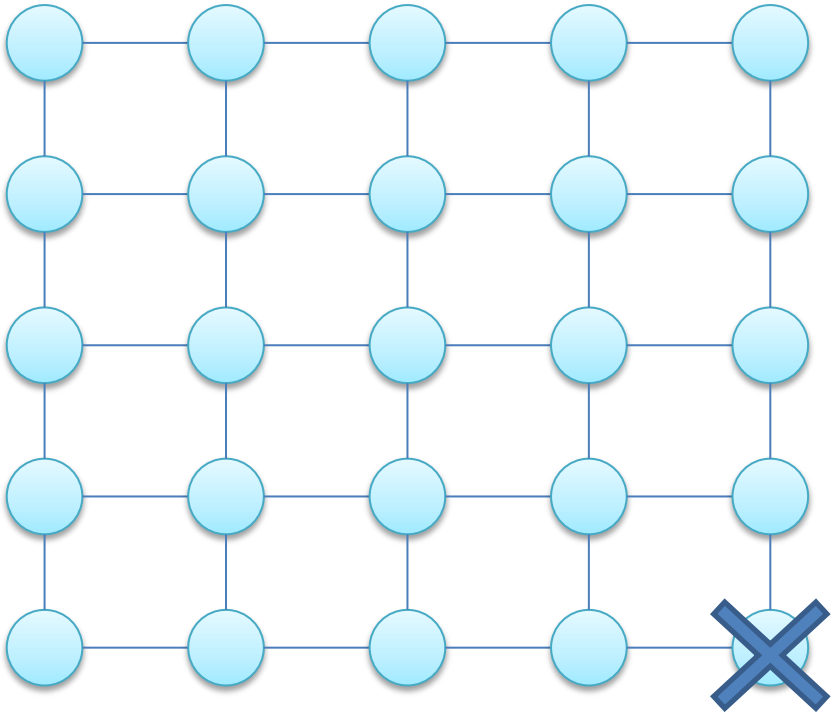
Example : General Factorial HMM

A clique size=5,
intractable most of times.
(No tractable elimination exist…)



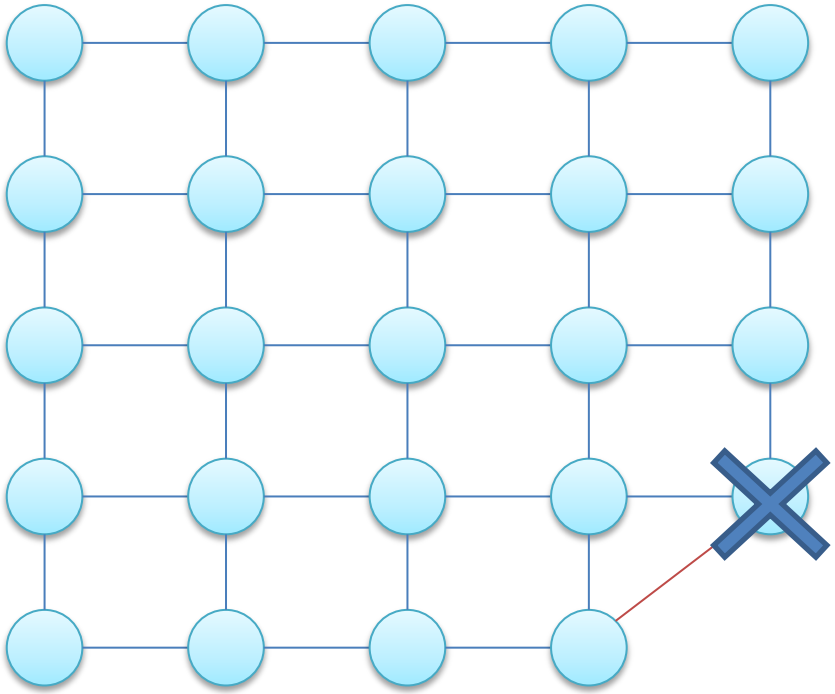
Some Model are Intractable for Exact Inference

Example: A Grid MRF



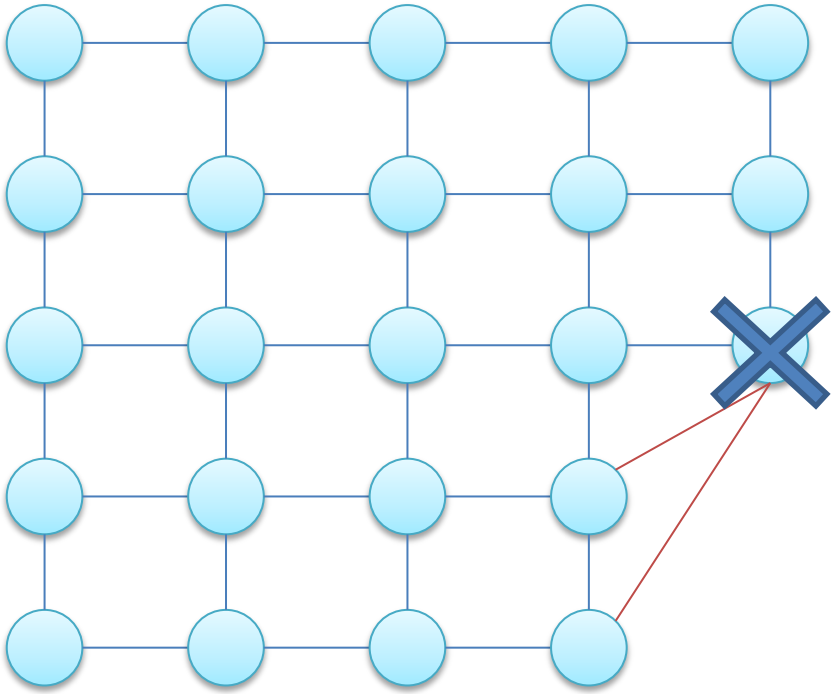
Some Model are Intractable for Exact Inference

Example: A Grid MRF



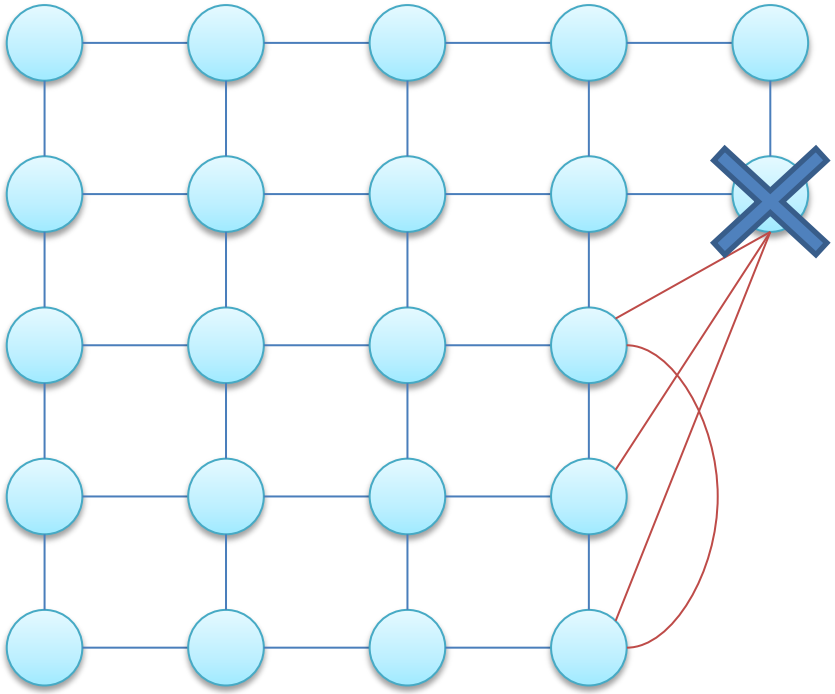
Some Model are Intractable for Exact Inference

Example: A Grid MRF



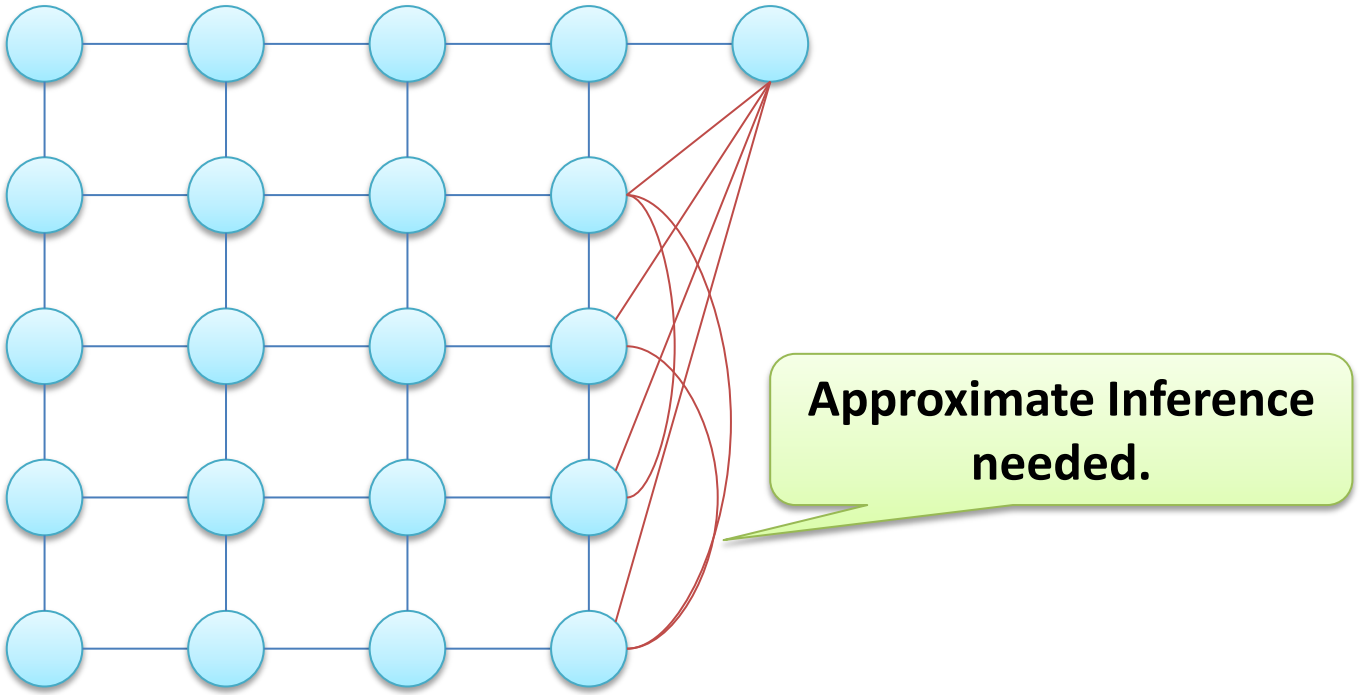
Some Model are Intractable for Exact Inference

Example: A Grid MRF



Some Model are Intractable for Exact Inference

Example: A Grid MRF



Generally, we will have clique of "size N"
for a N*N grid, which is indeed intractable.

General idea of Particle-Based (Monte Carlo) Approximation

Most of Queries we want can be formed as:

Intractable when K→ ∞.

$$E_{P(X)}[f(X)] = \sum_{X_1} \ldots \sum_{X_K} P(X_1 \ldots X_K) * f(X_1 \ldots X_K)$$

which is intractable most of time. Assume we can generate i.i.d. samples $X^{(1)} \ldots X^{(n)}$ from P(X), we can approximate above using:

$$\hat{f} = \frac{1}{N} \sum_{n=1}^{N} f(X^{(n)})$$

It's a unbiased estimator whose variance converges to 0 when N→∞.

$$E[\hat{f}] = \frac{1}{N} E[\sum_{n=1}^{N} f(X^{(n)})] = E[f(X)]$$

$$Var[\hat{f}] = \frac{1}{N^2} Var[\sum_{n=1}^{N} f(X^{(n)})] = \frac{1}{N} Var[f(X)]$$

Var. not Related to dimension of X. Var→0 as N→∞

Which Problem can use Particle-Based (Monte Carlo) Approximation ?

- Type of queries:
  - 1. Likelihood of evidence/assignments on variables
  - 2. Conditional Probability of some variables (given others).
  - ~~3. Most Probable Assignment for some variables (given others).~~

Problem which can be written as following form:

$$E_{P(X)}[f(X)] = \sum_{X_1} \ldots \sum_{X_K} P(X_1 \ldots X_K) * f(X_1 \ldots X_K)$$

Marginal Distribution (Monte Carlo)

To Compute Marginal Distribution on X_k

$$P(X_k = x_k)$$
$$= \sum_{X_{-k}} P(X_k = x_k, X_{-k}) = \sum_{X_k} \sum_{X_{-k}} P(X_k, X_{-k}) * 1\{X_k = x_k\}$$
$$= E_{P(X)}[\,1\{X_k = x_k\}\,]$$

Particle-Based Approximation:

$$\hat{f} = \frac{1}{N} \sum_{n=1}^{N} 1\{X_k^{(n)} = x_k\}$$

(Just count the proportion of samples in which $X_k = x_k$)

Marginal Joint Distribution (Monte Carlo)

To Compute Marginal Distribution on (X_i , X_j)

$$P(X_i = x_i, X_j = x_j)$$

$$= \sum_{X_{-ij}} P(X_i = x_i, X_j = x_j, X_{-ij}) = \sum_{X_{-ij}} \sum_{Xi} \sum_{Xj} P(X_i, X_j, X_{-k}) * 1\{X_i = x_i \,\&\, X_j = x_j\}$$

$$= E_{P(X)}[\, 1\{X_i = x_i \,\&\, X_j = x_j\}\,]$$

Particle-Based Approximation:

$$\hat{f} = \frac{1}{N} \sum_{n=1}^{N} 1\{X_i^{(n)} = x_i, X_j^{(n)} = x_j\}$$

(Just count the proportion of samples in which $X_i=x_i$ & $X_j=x_j$)

So What's the Problem ?

Note what we **can** do is:

"**Evaluate**" the probability/likelihood $P(X_1=x_1,\ldots, X_K=x_K)$.

What we **cannot** do is:

Summation / Integration in high-dim. space: $\sum_X P(X_1,\ldots, X_K)$.

What we **want to** do (for approximation) is:

"Draw" samples from $P(X_1,\ldots, X_K)$.

How to make better use of samples ?

How to know we've sampled enough ?

How to draw Samples from P(X) ?

- Forward Sampling
  draw from ancestor to descendant in BN.

- Rejection Sampling
  create samples using Forward Sampling, and reject those inconsistent with evidence.

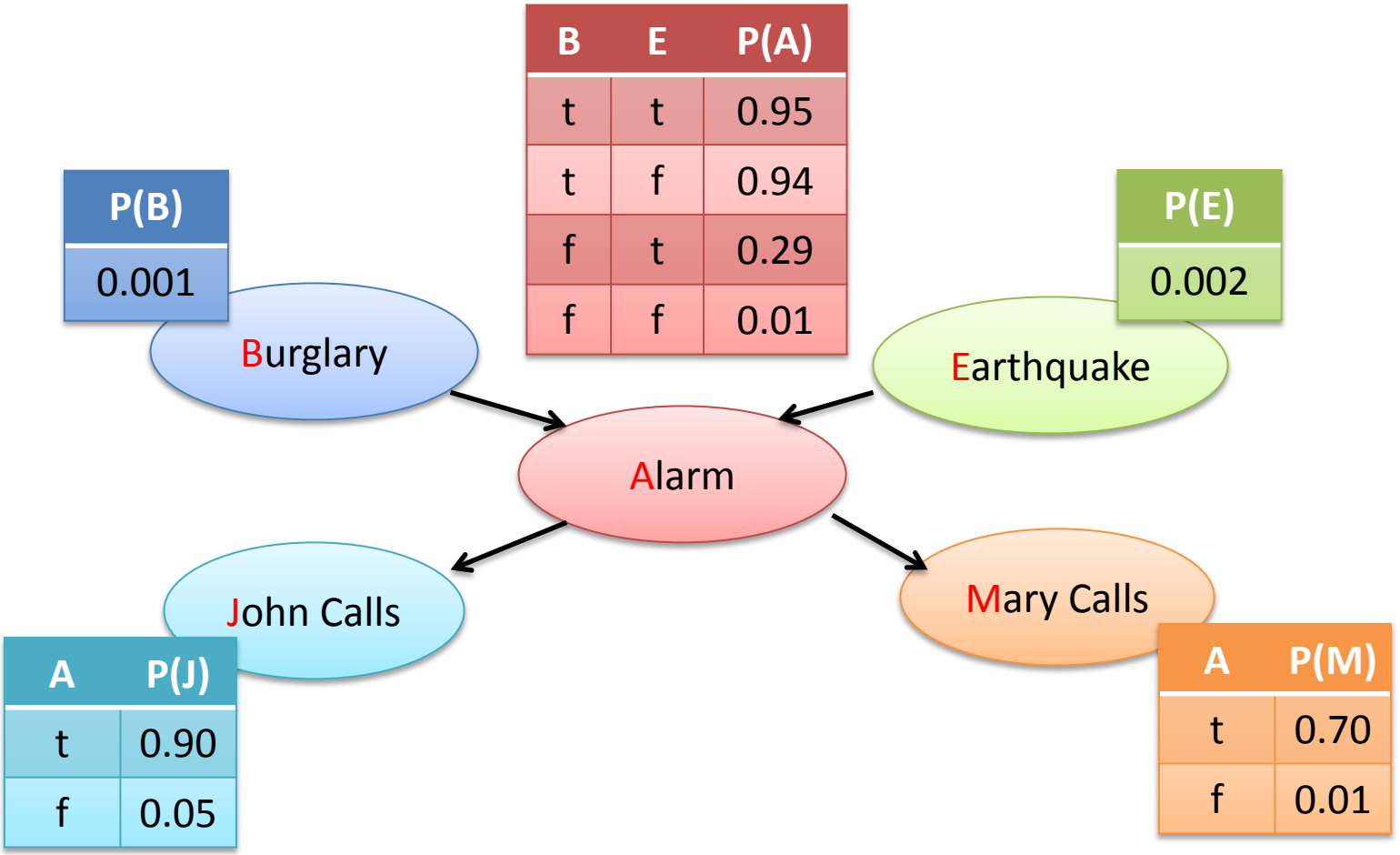
- Importance Sampling
  Sample from proposal dist. Q(X), but give large weight on sample with high likelihood in P(X).

- Markov Chain Monte Carlo
  Define a Transition Dist. $T(x \rightarrow x')$ s.t. samples can get closer and closer to P(X).

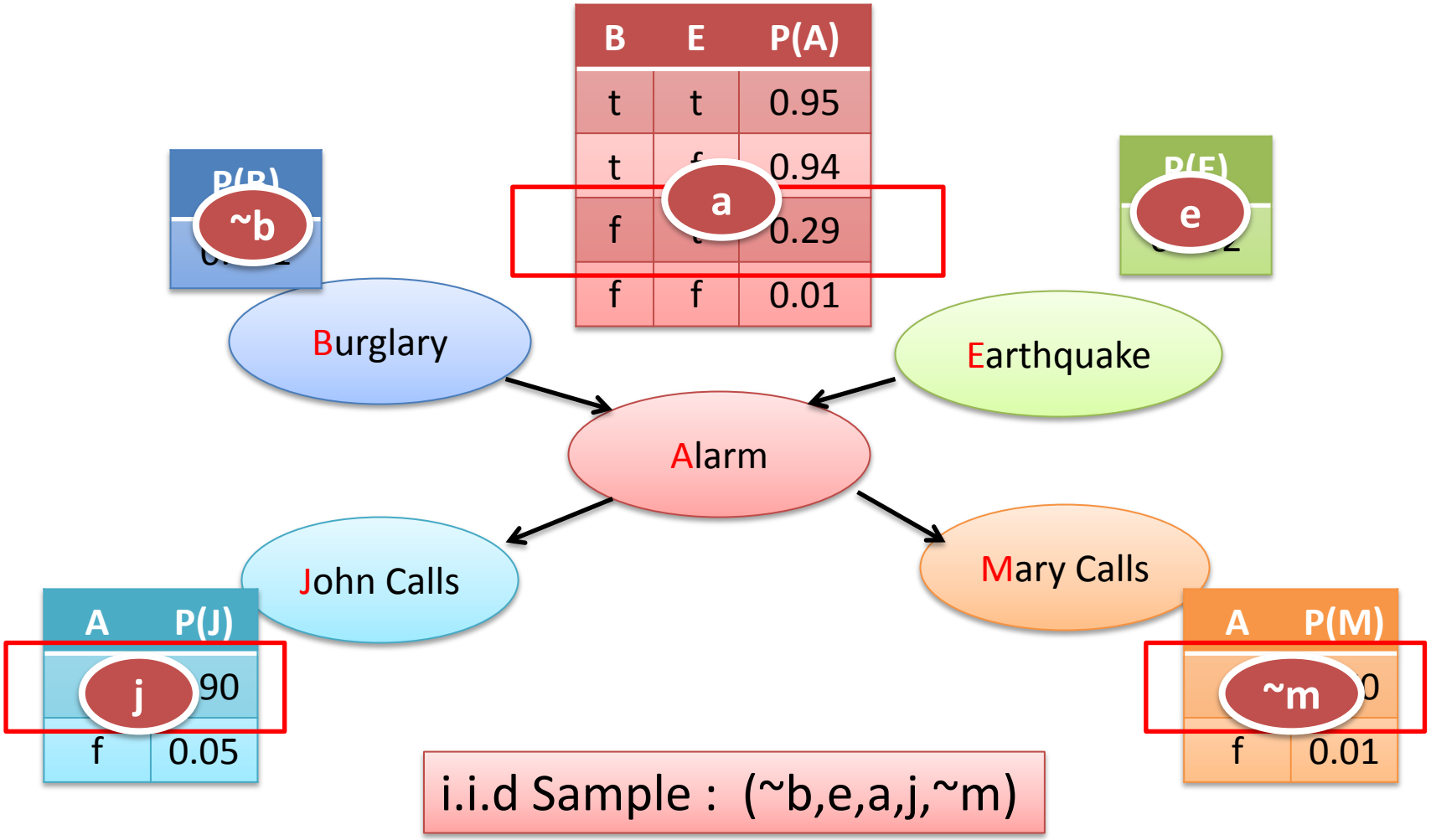
Agenda

- When to use Particle-Based Approximate Inference ?
- Forward Sampling & Importance Sampling
- Markov Chain Monte Carlo (MCMC)
- Collapsed Particles

Forward Sampling



Forward Sampling



Forward Sampling

Samples :
(~b,e,a,j,~m)
…
…
(~b,~e, a,~j, ~m)

Particle-Based Represent
of the joint distribution P(B,E,A,J,M).

$$P(M = m) = \frac{1}{N}\sum_{n=1}^{N} 1\{M^{(n)} = m\}$$

$$P(B = b, M = \sim m) = \frac{1}{N}\sum_{n=1}^{N} 1\{B^{(n)} = b, M^{(n)} = \sim m\}$$

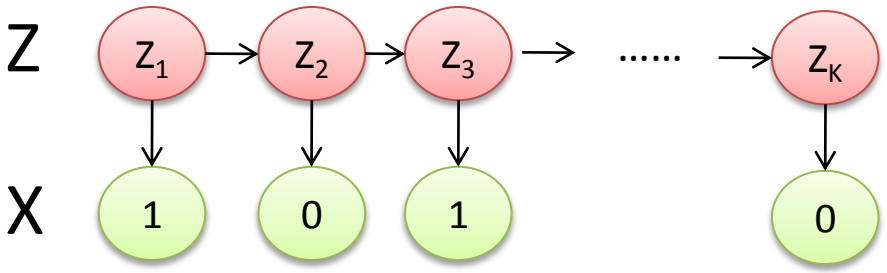
What if we want samples from P(**B, E, A** | J=j , M=~m) ?

1. Collect all samples in which J=j , M=~m .

2. Those samples form the particle-based representation of P(**B, E, A** | J=j, M=~m).

Disadvantage…….

Forward Sampling from P(Z|Data) ?



1. Forward Sampling N times.

2. Collect all samples $(Z^{(n)},X^{(n)})$ in which $X_1=1, X_2=0, X_3=1, \ldots\ldots X_K=0$.

3. Those samples form the particle-based representation of **P(Z|X)**.

How many such samples can we get ??

➔ **N*P(Data) !!** (Less than 1 if N not large enough……)

Solutions…….

Importance Sampling to the Rescue

We need not draw from P(X) to compute $E_{P(X)}[f(X)]$:

$$E_{P(X)}[f(X)] = \sum_X P(X) * f(X)$$

$$= \sum_X Q(X) * (\frac{P(X)}{Q(X)} * f(X)) = E_{Q(X)}[\frac{P(X)}{Q(X)} * f(X)]$$

$$\hat{E}_{P(X)}[f(X)] = \frac{1}{N} \sum_{n=1}^{N} (\frac{P(X^{(n)})}{Q(X^{(n)})}) * f(X^{(n)})$$

That is, we can draw from an arbitrary distribution Q(X), but give larger weights on samples having higher probability under P(X).

Importance Sampling to the Rescue

Sometimes we can only evaluate an unnormalized distribution :

$$\tilde{P}(X) \quad , where \ \frac{\tilde{P}(X)}{Z} = P(X)$$

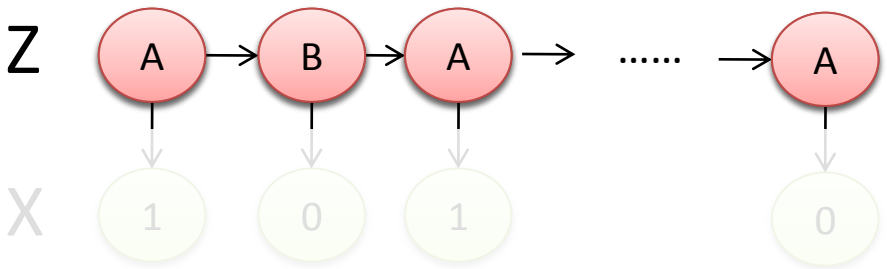
Then we can estimate Z as follows:

$$Z = \sum_X \tilde{P}(X) = \sum_X Q(X) \frac{\tilde{P}(X)}{Q(X)} = E_{Q(X)}[\frac{\tilde{P}(X)}{Q(X)}] \qquad \hat{Z} = \frac{1}{N} \sum_{n=1}^{N} \frac{\tilde{P}(X^{(n)})}{Q(X^{(n)})}$$

Note that we can compute $\hat{Z}$ only if we can evaluate a **normalized distribution** $Q(X)$, that is, we have Z_Q or $Q(X)$ is from a BN.

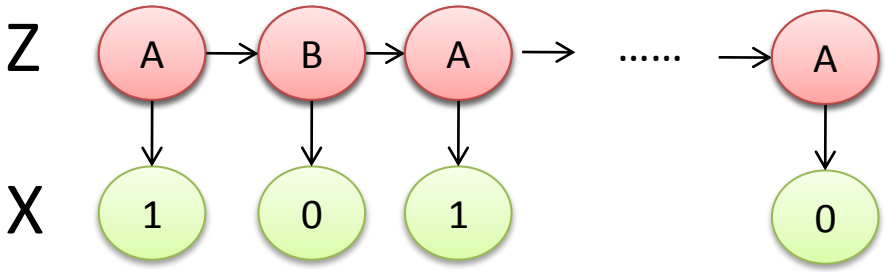
$$E_{P(X)}[f(X)] = \frac{1}{Z} E_{Q(X)}[\frac{\tilde{P}(X)}{Q(X)} * f(X)] \qquad \hat{E}_{P(X)}[f(X)] = \frac{\hat{E}_{\tilde{P}(X)}[f(X)]}{\hat{Z}} = \frac{\sum_{n=1}^{N} \frac{\tilde{P}(X^{(n)})}{Q(X^{(n)})} * f(X^{(n)})}{\sum_{n=1}^{N} \frac{\tilde{P}(X^{(n)})}{Q(X^{(n)})}}$$

Importance Sampling from P(Z|Data) ?



1. Sampling from P(Z), a normalized distribution obtained from BN truncating the part with evidence.

Importance Sampling from P(Z|Data) ?



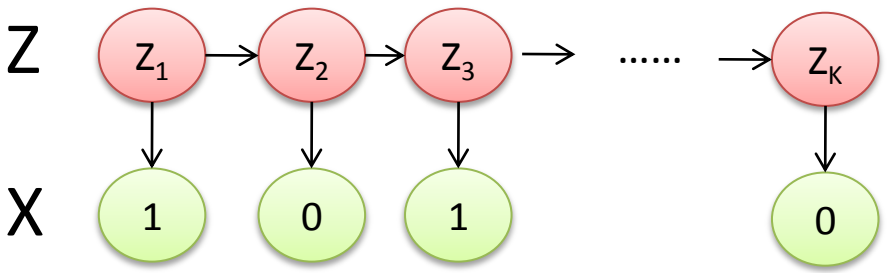
$$w^{(n)} = P(Data \mid Z)$$
$$= P(1 \mid A)P(0 \mid B)P(1 \mid A)....P(0 \mid A)$$

1. Sampling from P(Z), a normalized distribution obtained from BN truncating the part with evidence.

2. Give each sample **(Z(n), X(n))** a weight:

$$w^{(n)} = \frac{\tilde{P}(Z)}{Q(Z)} = \frac{P(Z)P(Data \mid Z)}{P(Z)} = P(Data \mid Z)$$

Importance Sampling from P(Z|Data) ?



(A,B,A,…,A)	w= 0.01
(A,B,A,…,B)	0.3
(B,B,B,…,A)	1.0

N_{eff} = 1.31

P(Data)=N_{eff}/N = 1.31/3

1. Sampling from P(Z), a normalized distribution obtained from BN truncating the part with evidence.

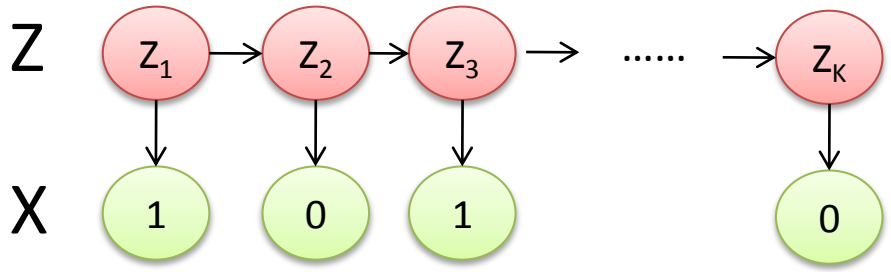
2. Give each sample **(Z(n), X(n))** a weight:

$$w^{(n)} = \frac{\tilde{P}(Z)}{Q(Z)} = \frac{P(Z)P(Data \mid Z)}{P(Z)} = P(Data \mid Z)$$

3. The effective number of samples is $N_{eff} = \sum_{n=1}^{N} w^{(n)}$

$$\left(\hat{P}(\text{Data}) = \frac{1}{N}\sum_{n=1}^{N} w^{(n)} = \frac{1}{N}\sum_{n=1}^{N} P(Data \mid Z^{(n)}) \right)$$

Importance Sampling from P(Z|Data) ?



(A,B,A,...,A) w= 0.01
(A,B,A,...,B) 0.3
(B,B,B,...,A) 1.0

N_{eff} = 1.31

P(Data)=N_{eff}/N = 1.31/3

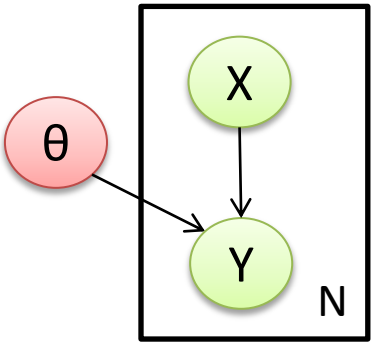
To get estimate of P(Z_1|Data) :

$$\hat{P}(Z_1 = B \mid Data) = \frac{0.01*0 + 0.3*0 + 1.0*1}{1.31} = 0.76$$

$$\hat{P}(Z_1 = A, Z_K = B \mid Data) = \frac{0.01*0 + 0.3*1 + 1.0*0}{1.31} = 0.23$$

Any joint dist. can be estimated. (No “out of clique” problem)

Bayesian Treatment with Importance Sampling



Ex. $P_\theta(Y=1|X) = \text{logistic}(\theta_1 * X + \theta_0)$

Often, Posterior on parameters θ :

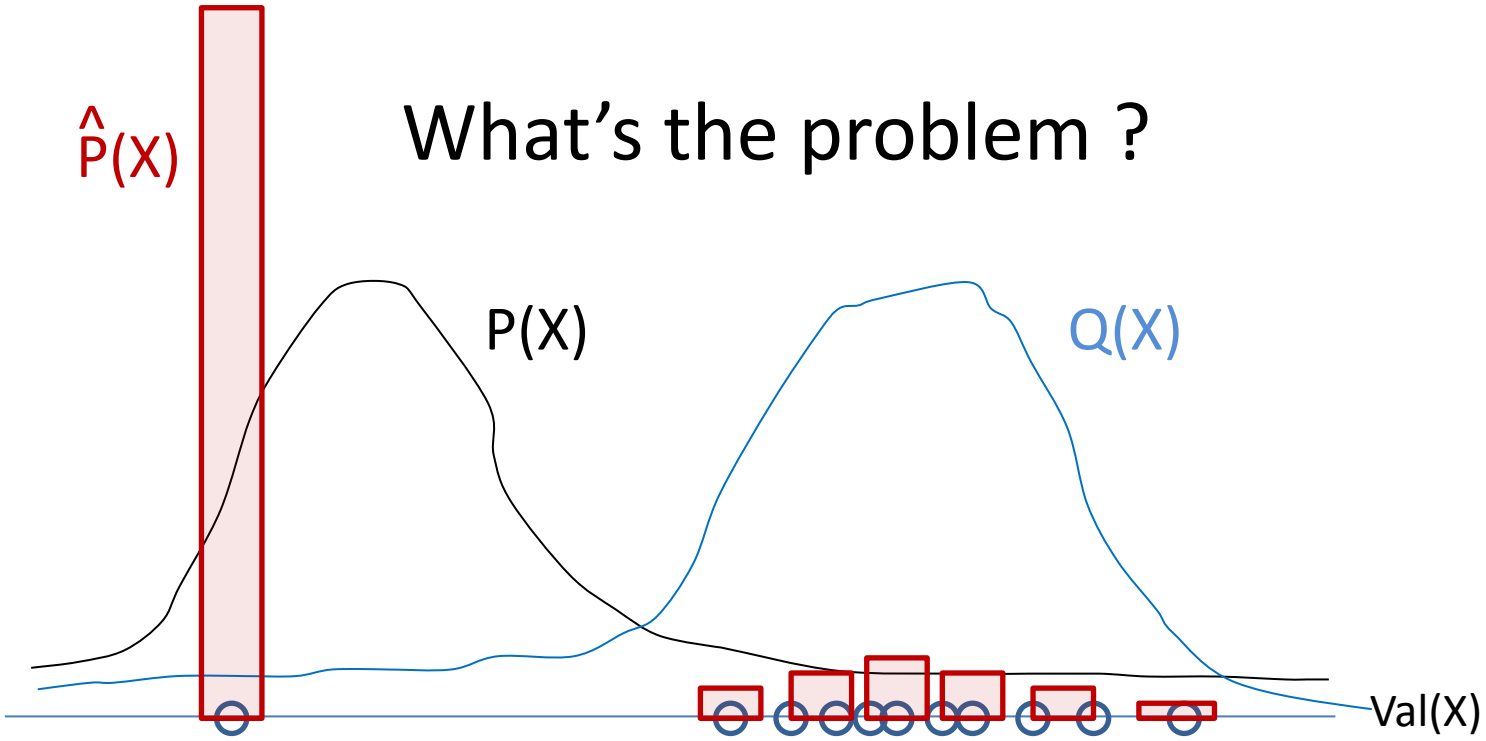
$$P(\theta \mid Data) = \frac{P(Data \mid \theta)P(\theta)}{P(\text{Data})} = \frac{P(Data \mid \theta)P(\theta)}{\int_\theta P(Data \mid \theta)P(\theta)\ d\theta}$$

is **intractable** because many types of $P_\theta(\text{Data}|\theta)$ cannot be integrated analytically.

Approximate with:

$$\hat{P}(\theta = a \mid \text{Data}) = \frac{\sum_{n=1}^{N} P(Data \mid \theta^{(n)} = a)\ 1\{\theta^{(n)} = a\}}{\sum_{n=1}^{N} P(Data \mid \theta^{(n)})} = \frac{P(Data \mid \theta = a)\sum_{n=1}^{N} 1\{\theta^{(n)} = a\}}{\hat{P}(Data)}$$

We need not evaluate "the integration" to estimate P(θ|Data) using Importance Sampling.



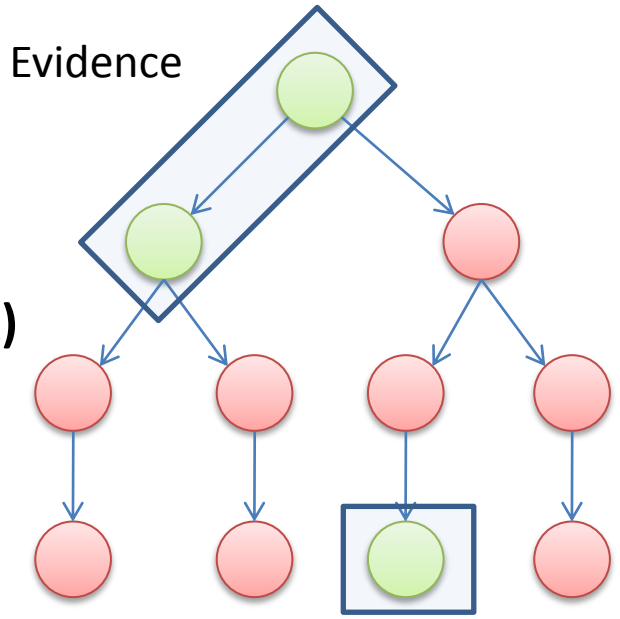
If P(X) and Q(X) not matched properly……

Only small number of samples will fall in the region with high P(X).

➔ Very large N needed to get a good picture of P(X).

How P(Z|X) and Q(Z) Match ?

When evidence is close to root, forward sampling is a good Q(Z), which can generate samples with high likelihood in P(Z|X).

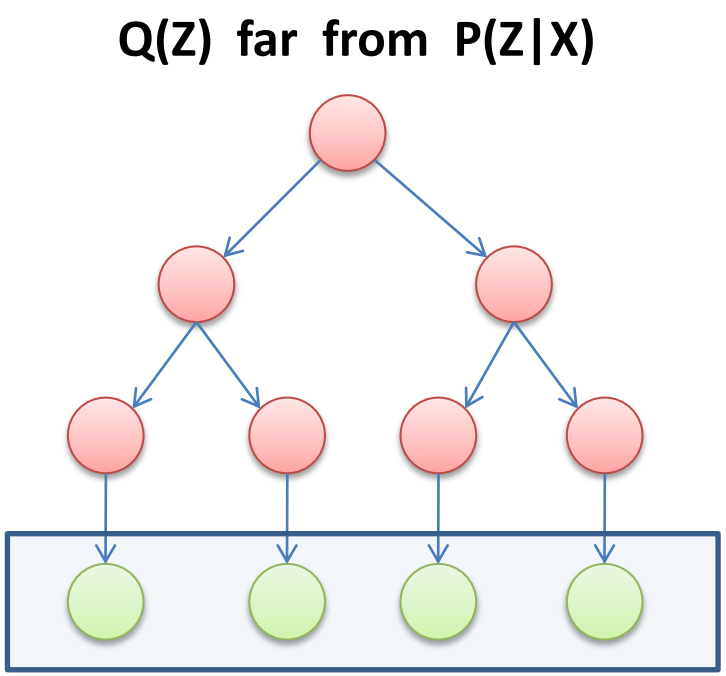


Q(Z) close to P(Z|X)

How P(Z|X) and Q(Z) Match ?

When evidence is on the leaves,
forward sampling is a bad Q(Z),
yields very low likelihood=P(X|Z).

So we need very large sample size
to get a good picture of P(Z|X).

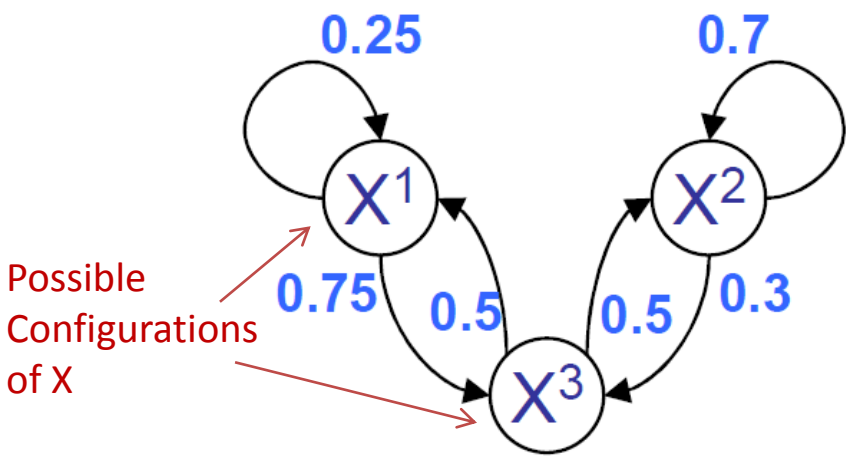


Can we **improve with time** to draw from a distribution
more like the desired P(Z|X) ?
➔MCMC try to draw from a distribution closer and closer to P(Z|X).
(Apply **equally well in BN & MRF**.)

Agenda

- When to use Particle-Based Approximate Inference ?
- Forward Sampling & Importance Sampling
- Markov Chain Monte Carlo (MCMC)
- Collapsed Particles

What is Markov Chain (MC) ?



A set of Random Variables:
$\mathbf{X} = (X_1,....X_K)$

Variables change with Time:
$\mathbf{X^{(t)}} = (X_1^{(t)},....X_K^{(t)})$

which take transition following:
$P(\mathbf{X^{(t+1)}} = x' \mid \mathbf{X^{(t)}} = x) = T(x \rightarrow x')$

There is a stationary distribution $\pi_T(X)$ for Transition T, in which:

$$\pi_T(X=x') = \sum_x \pi_T(X=x) * T(x \rightarrow x')$$

(After transition, still the same distribution over all possible configurations X^1~X^3)

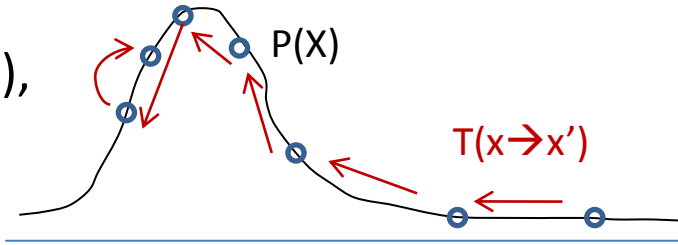
Ex. The MC (Markov Chain) above has only 1 variable X taking on values $\{x^1,x^2,x^3\}$,

There is a $\boldsymbol{\pi_T}$ s.t. $\pi_T * T = \begin{bmatrix}0.2 & 0.5 & 0.3\end{bmatrix}\begin{bmatrix}0.25 & 0 & 0.75\\ 0 & 0.7 & 0.3\\ 0.5 & 0.5 & 0\end{bmatrix} = \begin{bmatrix}0.2 & 0.5 & 0.3\end{bmatrix} = \pi_T$

What is MCMC (Markov Chain Monte Carlo) ?

Importance Sampling is efficient only if **Q(X) matches P(X) well**.
Finding such Q(X) is difficult.

Instead, MCMC tries to find a transition dist. T(x→x'),
s.t. **X tends to transit into states with high P(X)**,
and **finally follows stationary dist. $\pi_T = P(X)$.**

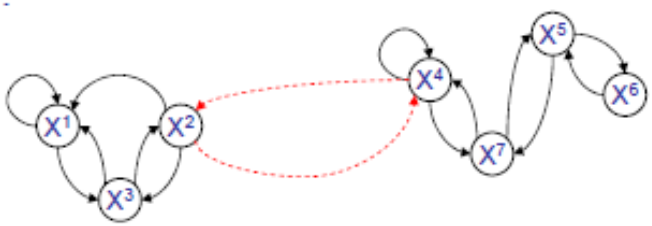


Setting $X^{(0)}$=any initial value, we samples $X^{(1)}, X^{(2)} \ldots\ldots X^{(M)}$ following T(x→x'), and hope that $X^{(M)}$ follows stationary distribution $\boldsymbol{\pi_T = P(X)}$.
If $X^{(M)}$ really does , we got a sample $X^{(M)}$ from P(X).

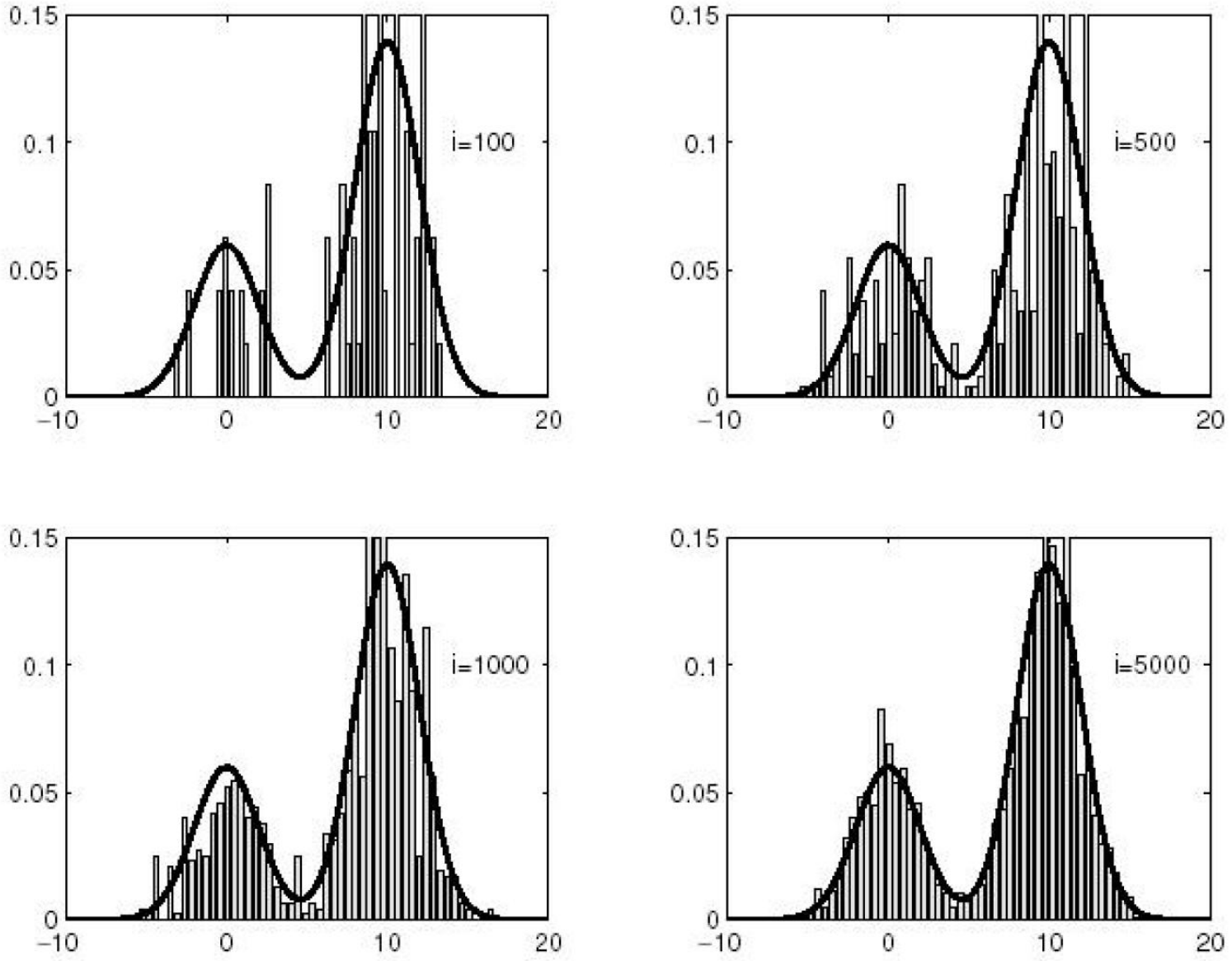
Why will the MC **converge to stationary distribution** ? there is a simple, useful **sufficient** condition:

"Regular " Markov Chain : (for finite state space)
Any state x can reach any other states x' with prob. > 0.
(all entries of Potential/CPD > 0)

➔ **$X^{(M)}$ follows a unique π_T as M large enough.**



Example Result



How to define T(x→x') ? ---- Gibbs Sampling

Gibbs Sampling is the most popular one used in Graphical Model.
In graphical model :

It is easy to draw sample from **"each individual variable given others $P(X_k|\mathbf{X_{-k}})$"**, while drawing from the **joint dist. of $(X_1,X_2,\ldots,X_K)$** is difficult.

So, we define T(X→X') in Gibbs-Sampling as :
Taking transition of X_1 ~ X_K in turn with transition distribution :

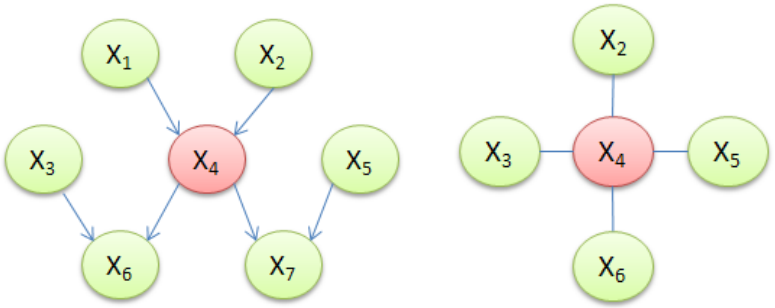
$$T_1(x_1 \rightarrow x_1'),\ T_2(x_2 \rightarrow x_2'),\ \ldots\ldots\ T_K(x_K \rightarrow x_K')$$

Where

$$T_k(x_k \rightarrow x_k') = P(X_k = x_k' | \mathbf{X_{-k}})$$ (Redraw X_k ~ conditional dist. given all others.)

In a Graphical Model,

$P(X_k = x_k' | \mathbf{X_{-k}}) = P(X_k = x_k' |$ Markov Blanket(X_k))

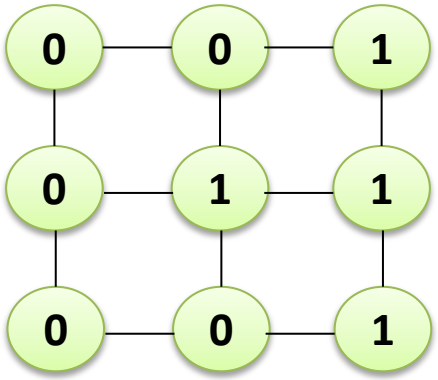


Gibbs Sampling for MRF

Gibbs Sampling :

1. Initialize all variables randomly.
for t = 1~M
    for every variable X
        2. Draw X_t from P(X | $N(X)_{t-1}$).
    end
end

t=1



$$P(X=1 \mid N(X)) = \frac{\prod_{Y \in N(X)} \phi(X=1,Y)}{\prod_{Y \in N(X)} \phi(X=1,Y) + \prod_{Y \in N(X)} \phi(X=0,Y)}$$

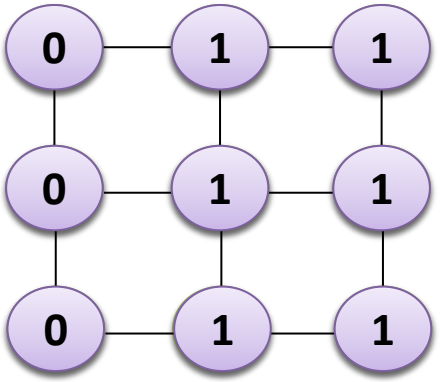
ϕ(X,Y)	0	1
0	5	1
1	1	9

Gibbs Sampling for MRF

Gibbs Sampling :

1. Initialize all variables randomly.
for t = 1~M
    for every variable X
        2. Draw X_t from P(X | $N(X)_{t-1}$).
    end
end

t=2



$$P(X=1|N(X)) = \frac{\prod_{Y\in N(X)} \phi(X=1,Y)}{\prod_{Y\in N(X)} \phi(X=1,Y) + \prod_{Y\in N(X)} \phi(X=0,Y)}$$

For the central node:

$$P(X=1|N(X)) = \frac{1*9*9*1}{1*9*9*1+5*1*1*5} = 0.76$$

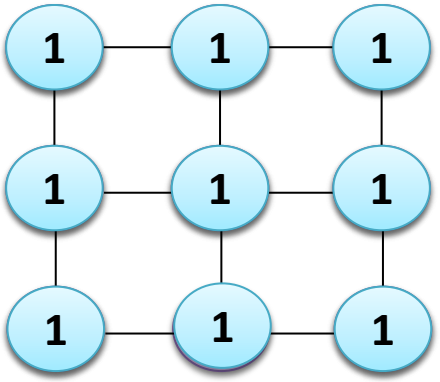
ϕ(X,Y)	0	1
0	5	1
1	1	9

Gibbs Sampling for MRF

Gibbs Sampling :

1. Initialize all variables randomly.
for t = 1~M
    for every variable X
        2. Draw X_t from P(X | $N(X)_{t-1}$).
    end
end

t=3



$$P(X=1|N(X)) = \frac{\prod_{Y\in N(X)} \phi(X=1,Y)}{\prod_{Y\in N(X)} \phi(X=1,Y) + \prod_{Y\in N(X)} \phi(X=0,Y)}$$

ϕ(X,Y)	0	1
0	5	1
1	1	9

For the central node:

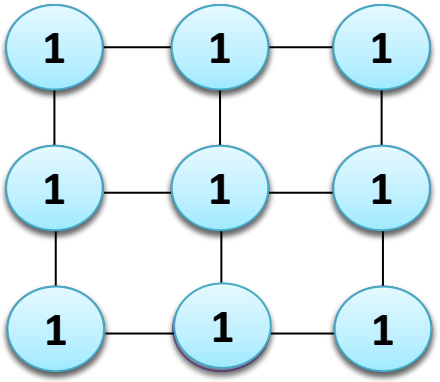
$$P(X=1|N(X)) = \frac{9*9*9*9}{9*9*9*9+1*1*1*1} = 0.99$$

Gibbs Sampling for MRF

Gibbs Sampling :

1. Initialize all variables randomly.
for t = 1~M
    for every variable X
        2. Draw X_t from P(X | $N(X)_{t-1}$).
    end
end

t=3



When M is large enough, $X^{(M)}$ follows stationary dist. :

$$\pi_T(X) = P(X) = \frac{1}{Z}\prod_C \phi(X_C)$$

(Regularity: All entries in the Potential are positive.)

ϕ(X,Y)	0	1
0	5	1
1	1	9

Why Gibbs Sampling has $\pi_T = P(X)$?

To prove P(X) is the stationary distribution, we prove P(X) is invariant under $T_k(x_k \rightarrow x_k')$:

Assume $(X_1,\ldots,X_K)$ currently follows $P(X)= P(X_k|X_{-k})*P(X_{-k})$,

1. After $T_k(x_k \rightarrow x_k')$, **$X_{-K}$ still follows $P(X_{-k})$** because they are unchanged.

2. After $T_k(x_k \rightarrow x_k')=P(X_k=x_k'|X_{-k})$ (new state indep. from current value x_k)

   ➔ **$X_k(t)$ still follows $P(X_k|X_{-k})$.**

So, after $T_1(x_1 \rightarrow x_1')$, ……, $T_1(x_K \rightarrow x_K')$, **$X=(X_1,\ldots,X_K)$ still follows $P(X)$**.

(**Uniqueness** & **Convergence** guaranteed from **Regularity** of MC.)

Gibbs Sampling not Always Work

When drawing from individual variable is not possible:
(We can evaluate P(Y|X) but not P(X|Y).)

Non-linear Dependency :

$$P(Y \mid X) = N(w_0 + w_1 X + w_2 X^2, \sigma^2)$$

$$P(Y \mid X) = \log istic\ (w_0 + w_1 X_1)$$

$$P(Y \mid X) = N(\ \sum_{n=1}^{N} \mathrm{K}(\mathrm{X}, \mathrm{X}^{(n)})\ ,\ \sigma^2\ )\ (\ker nel\ trick)$$

$$P(X \mid Y) = \frac{P(Y \mid X)P(X)}{\int_X P(Y \mid X)P(X)\, dX}$$

(Intractable Integration)

Large State Space : $(In\ Structure\ \ Learning,\ \mathrm{statespace} = G_1, G_2, G_3....)$

$$P(G \mid Data) = \frac{P(Data \mid G)P(G)}{\sum_G P(Data \mid G)P(G)}$$

(Too large state space to do summation)

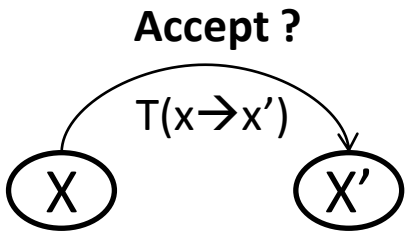
Other MCMC like **Metropolis-Hasting** needed. (see reference.)

Metropolis-Hasting ----MCMC

Metropolis-Hasting (M-H) is a general MCMC method to sample P(X|Y) whenever we can evaluate P(Y|X). (**evaluation of P(X|Y) not needed**)

In M-H, instead of drawing from P(X|Y), we draw from another **Proposal Dist. T(x→x')** based on current sample **x,** and **Accept the Proposal** with probability:

$$P(accept\ \ from\ x\ to\ x') = \begin{cases} 1 & , \ if\ \ P(x')\mathrm{T}(\mathrm{x}' \to \mathrm{x}) > P(x)\mathrm{T}(\mathrm{x} \to \mathrm{x}') \\ \dfrac{P(x')\mathrm{T}(\mathrm{x}' \to \mathrm{x})}{P(x)\mathrm{T}(\mathrm{x} \to \mathrm{x}')} & , \ o.w. \end{cases}$$

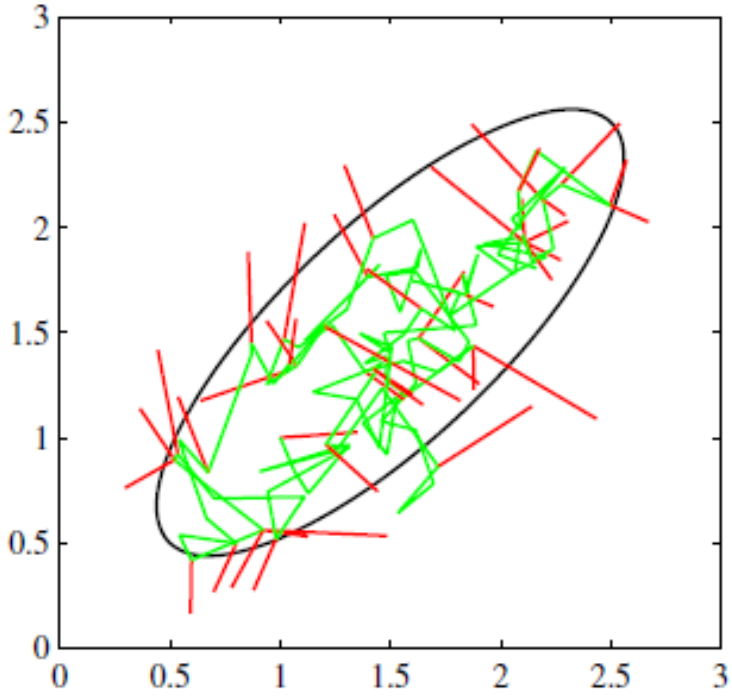


Example : P(X) = N(μ,σ²)

Proposal Dist. T(x→x') = N(x, 0.2^2)

$$P(accept\ \ from\ x\ to\ x') = \begin{cases} 1\ , & if\ \ |x' - \mu| < |x - \mu| \\ \dfrac{N(x';\mu,\sigma^2)}{N(x;\mu,\sigma^2)}\ , & o.w. \end{cases}$$

$$(\ T(x \to x') = T(x' \to x)\ \ this\ \ case.\ )$$



(red: Reject)

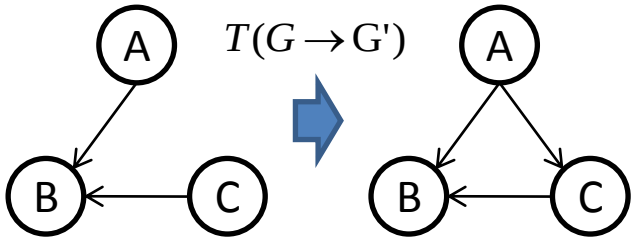
(green: Accept)

Example : Structure Posterior = P(G|Data)

Proposal Distribution:

$T(G \to G')$

$= \mathrm{P}(\text{add/remove a randomly chosen edge of } \mathrm{G} \Rightarrow \mathrm{G}')$



$$P(\textit{accept from } G \textit{ to } G') = \begin{cases} 1 \ , \ \textit{if } P(\mathrm{Data}\,|\,\mathrm{G}') < P(\mathrm{Data}\,|\,\mathrm{G}) \\ \dfrac{P(\mathrm{Data}\,|\,\mathrm{G}')}{P(\mathrm{Data}\,|\,\mathrm{G})} \ , \ \textit{o.w.} \end{cases}$$

$(\ T(G \to G') = T(G' \to G) \ \textit{this case.} \ )$

Why Metropolis-Hasting has $\pi_T = P(X)$?

Detailed-Balance Sufficient Condition:

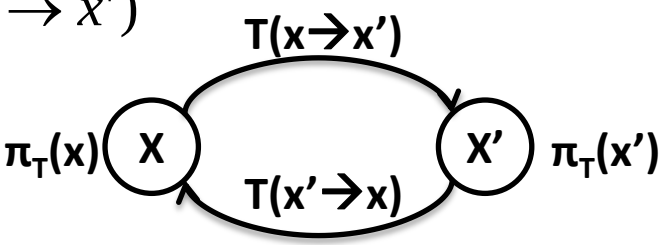
If $\pi_T(x')*T(x' \rightarrow x) = \pi_T(x)*T(x \rightarrow x')$, then $\pi_T(x)$ is **stationary under T.**

Given desired $\pi_T(x)=P(X)$, and a **Proposal dist. T(x→x') ,**
we can let **Detailed Balance** satisfied using **accept prob. A(x→x') :**

$$Assume\ P(x')T(x' \rightarrow x) < P(x)T(x \rightarrow x'),\ then:$$

$$We\ know\ P(x')T(x' \rightarrow x)*1 = P(x)T(x \rightarrow x')*\frac{P(x')T(x' \rightarrow x)}{P(x)T(x \rightarrow x')}$$

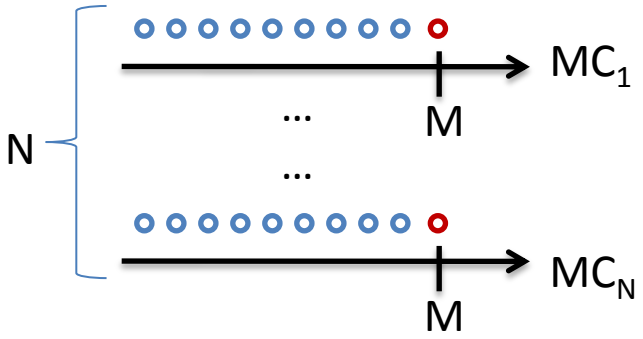
$$define\quad A(x \rightarrow x') = \begin{cases} 1, & P(x')T(x' \rightarrow x) > P(x)T(x \rightarrow x') \\ \dfrac{P(x')T(x' \rightarrow x)}{P(x)T(x \rightarrow x')}, & o.w. \end{cases}$$



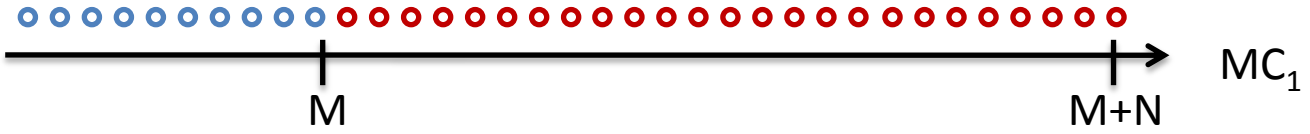
How to Collect Samples ?

Assume we want collecting N samples:

1. Run N times of MCMC and collect their M^{th} samples.



2. Run 1 time of MCMC and collect $(M+1)^{th}$ ~ $(M+N)^{th}$ samples.

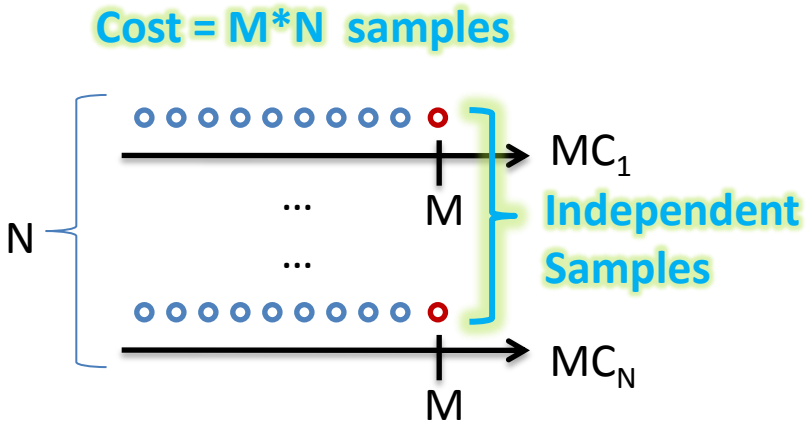


What's the difference ??

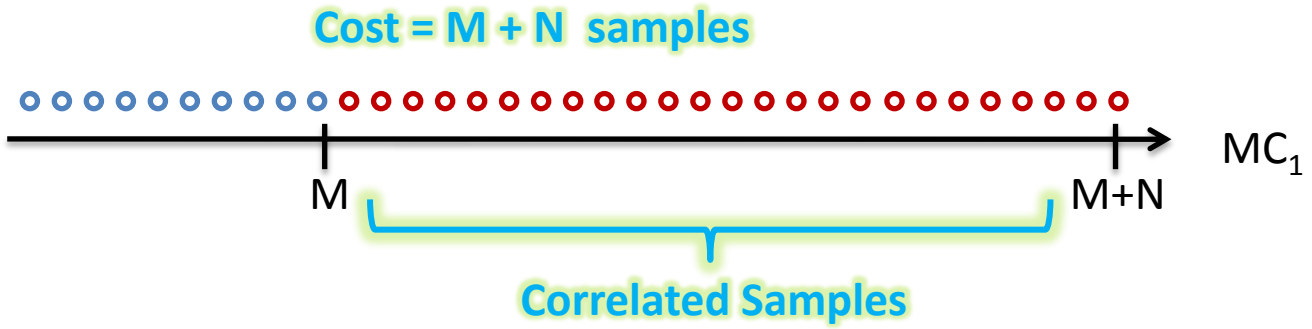
How to Collect Samples ?

Assume we want collecting N samples:

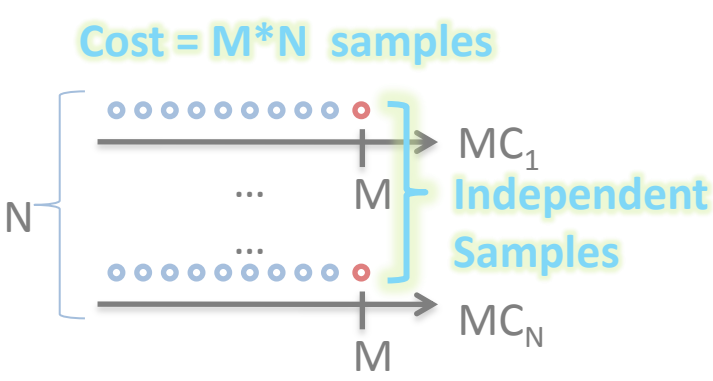
1. Run N times of MCMC and collect their M^{th} samples.

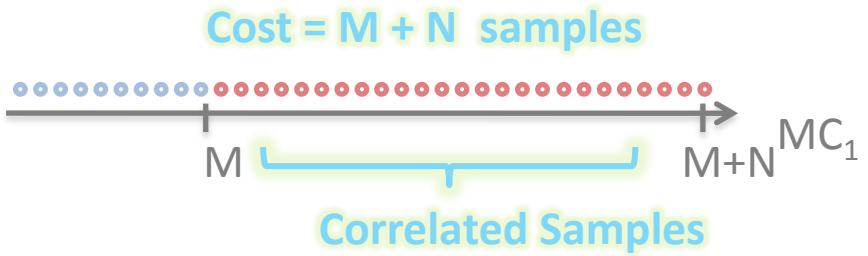


2. Run 1 time of MCMC and collect $(M+1)^{th}$ ~ $(M+N)^{th}$ samples.



Comparison





$$E[\hat{f}] = E[\frac{1}{N}\sum_{n=1}^{N} f(X^{(n)})] = \frac{1}{N}E[\sum_{n=1}^{N} f(X^{(n)})] = \frac{1}{N}\sum_{n=1}^{N} E[f(X^{(n)})] = E[f(X)]$$

No Independent Assumption Used ➔ **Unbiased Estimator in both cases.**

For simple analysis, Take N=2:

$$Var[\hat{f}] = Var[\frac{1}{2}(f(X^{(1)}) + f(X^{(2)})]$$

$$= \frac{1}{4}(Var[f(X^{(1)})] + Var[f(X^{(2)})]) = \frac{Var[f(X)]}{2}$$

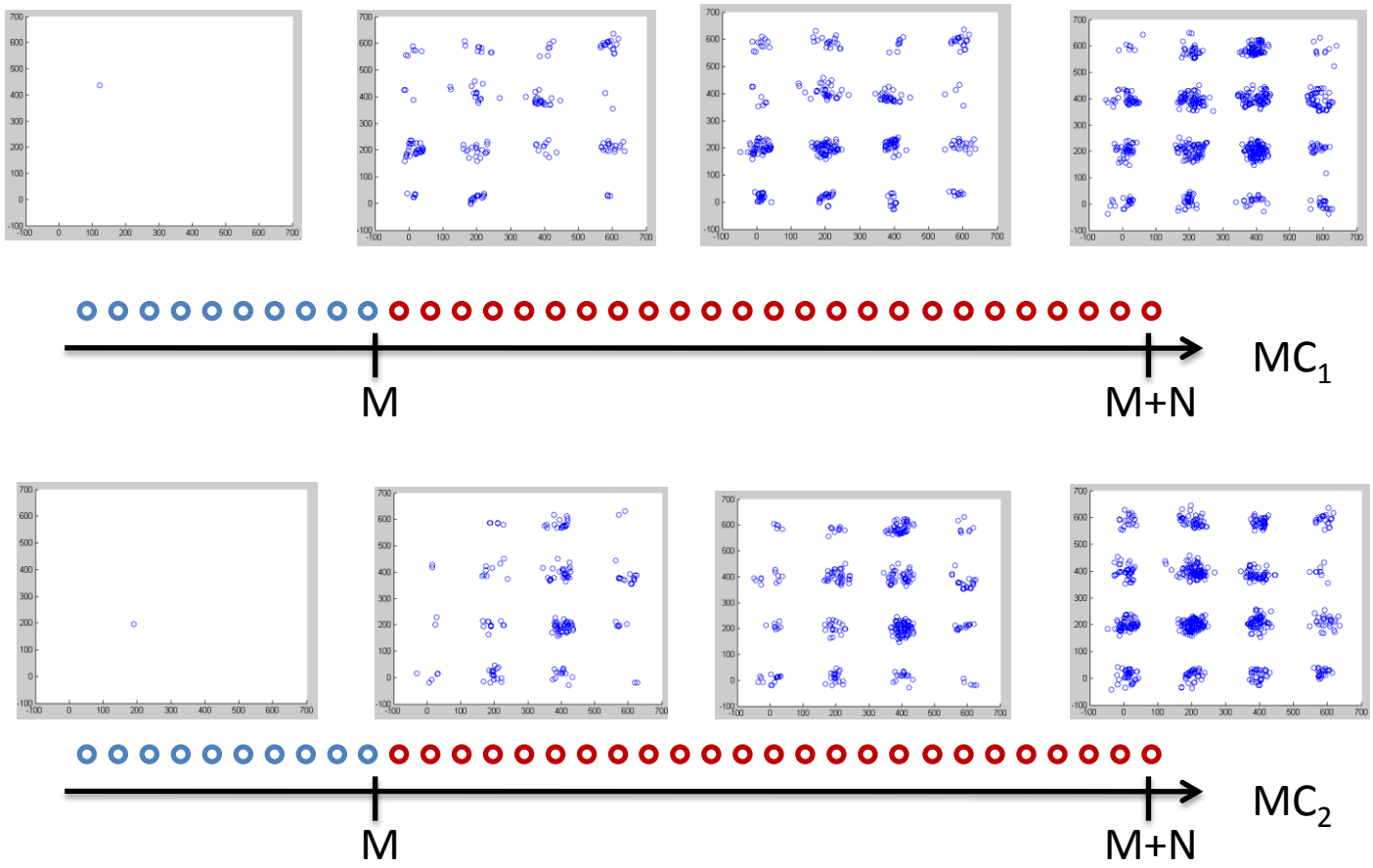
$$Var[\hat{f}] = Var[\frac{1}{2}(f(X^{(1)}) + f(X^{(2)})]$$

$$= \frac{1}{4}(Var[f(X^{(1)})] + Var[f(X^{(2)})] + 2Cov[f(X^{(1)}), f(X^{(2)})])$$

$$= \frac{Var[f(X)]}{2} + \rho_{f(X^{(1)}), f(X^{(2)})} * \frac{Var[f(X)]}{2} > \frac{Var[f(X)]}{2}$$

Practically, **many correlated samples** (right) outperforms **few independent samples** (left).

How to Check Convergence ?



Should be consistent if converge to π_T

$$Check\ \text{Ratio} = \sqrt{\frac{\text{B}}{\text{W}}}\ close\ to\ 1\ enough.\quad (\text{assume K MCs, each with N samples.})$$

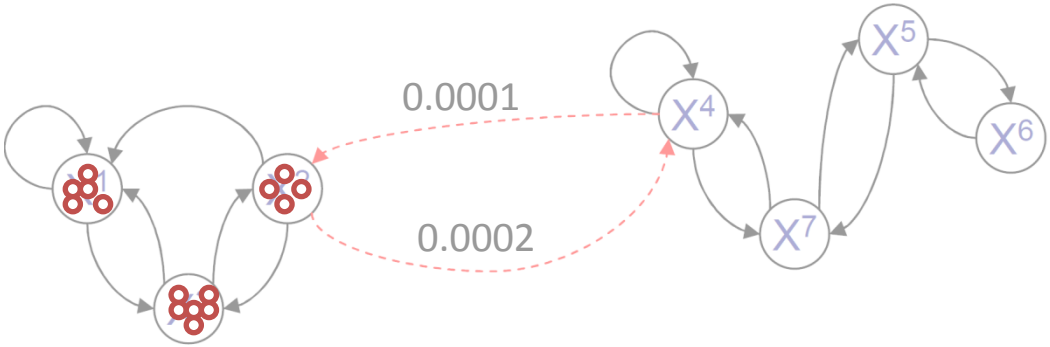
$$\bar{f} = \frac{1}{K}\sum_{k=1}^{K}\bar{f}_k$$

$$B = Var.\,between\ \text{MC} = \frac{\text{N}}{\text{K-1}}\sum_{k=1}^{K}(\bar{f}_k - \bar{f})^2$$

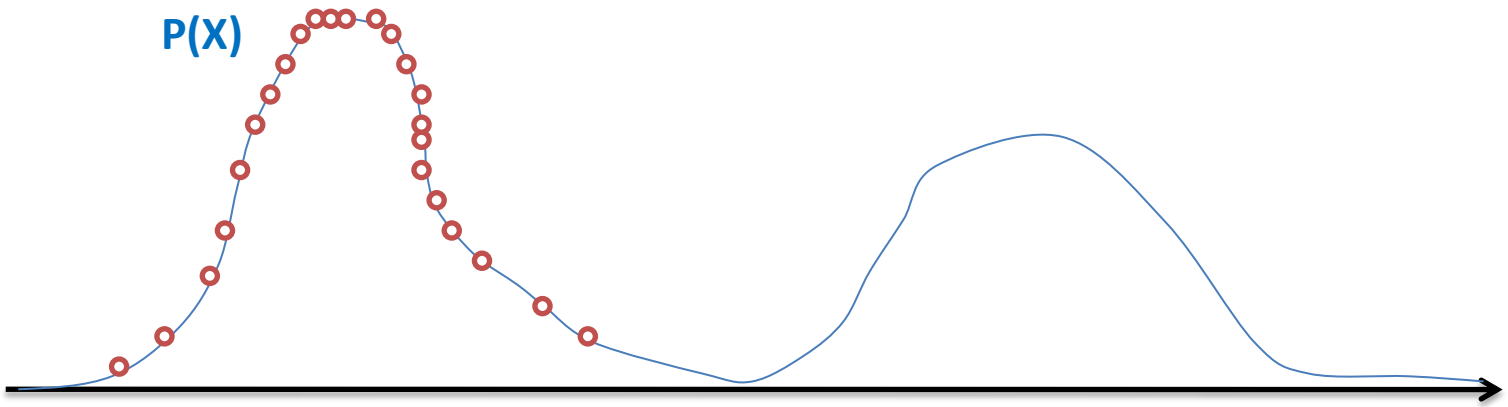
$$\text{W} = Var.\,within\ MC = \frac{1}{K(N-1)}\sum_{k=1}^{K}\sum_{n=1}^{N}(f(X^{(k,n)}) - \bar{f}_k)^2$$

The Critical Problem of MCMC

When **ρ→1 , M→∞ , Var[.] not decreasing with N**
➔ MCMC cannot yield acceptable result in reasonable time.



Taking very large M to converge to π_T .



How to Reduce Correlation (ρ) among Samples ?

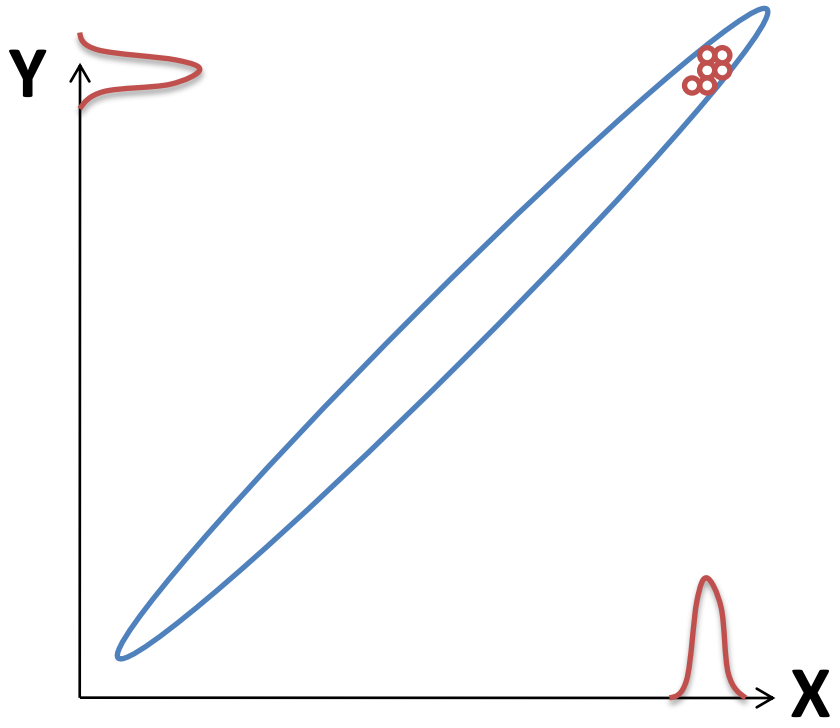
Taking Large Step in Sample Space :

- Block Gibbs Sampling
- Collapsed-Particle Sampling

Problem of Gibbs Sampling

Correlation (ρ) between samples is high,
when correlation among variables X_1~X_K is high.

		X	**0**
Y	ϕ(Y,X)	0	1
0	0	99	1
	1	1	99

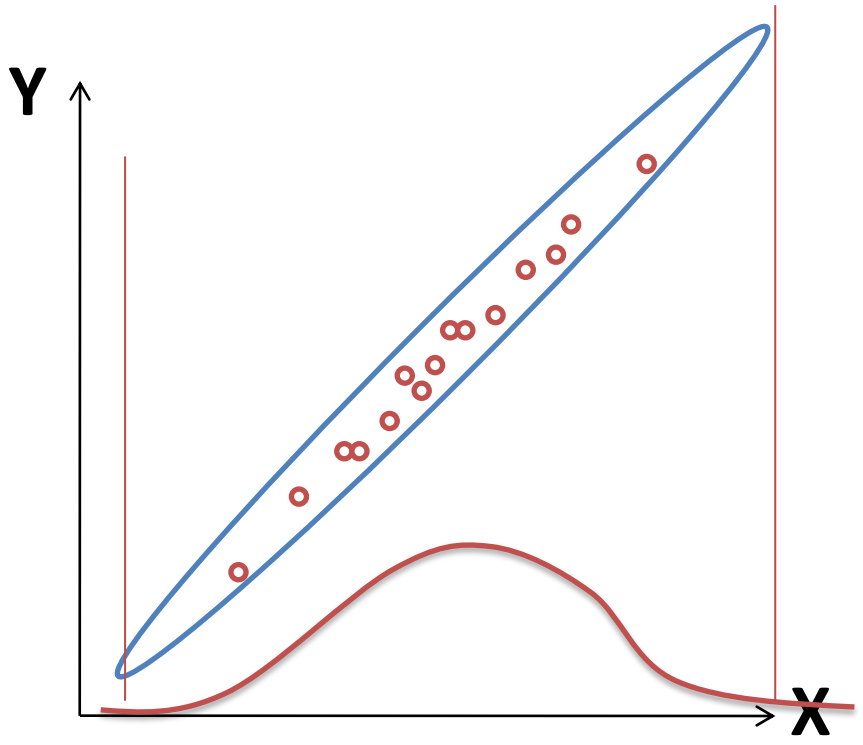


Taking very large M to converge to π_T .

Block Gibbs Sampling

Draw "block" of variables jointly: P(X,Y)=P(X)P(Y|X)

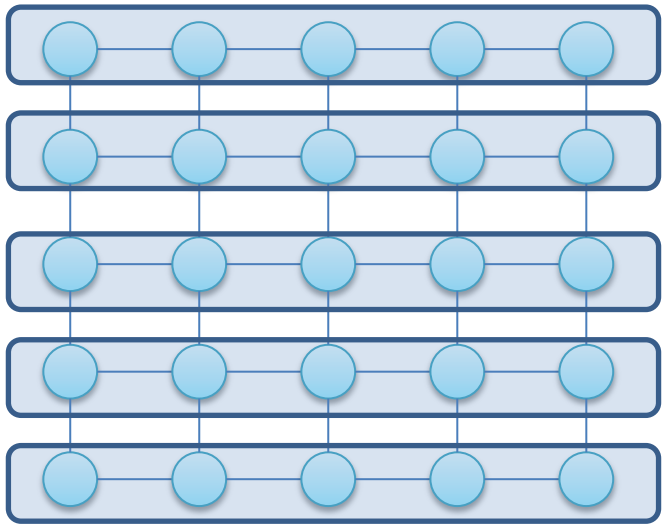
		Marginal	
	X	100	100
Y	ϕ(Y,X)	0	1
	0	99	1
	1	1	99

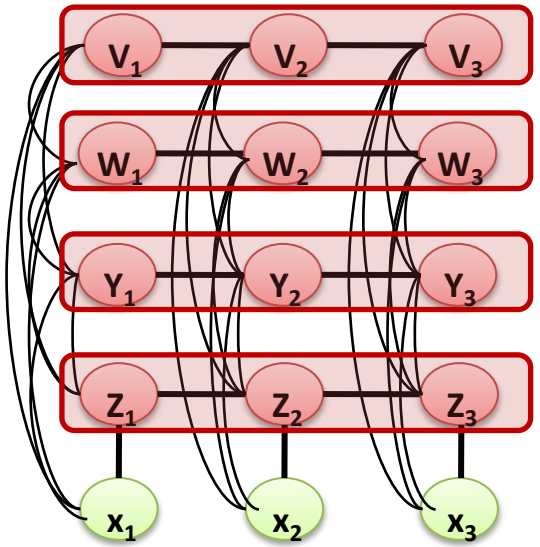


Converge to π_T much quickly.

Block Gibbs Sampling

Divide **X** into several "**tractable blocks**" $\mathbf{X_1}$, $\mathbf{X_2}$, …, $\mathbf{X_B}$.
Each block $\mathbf{X_b}$ can be **drawn jointly** given variables in other blocks.

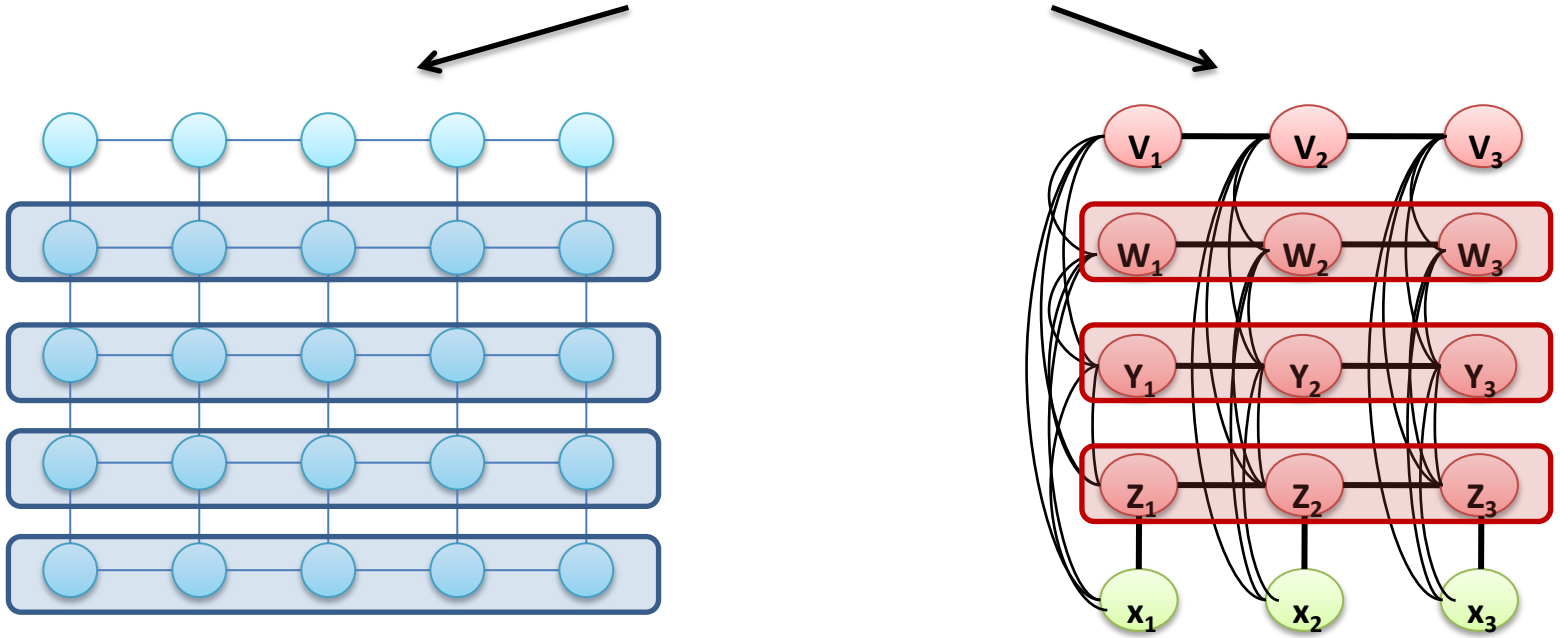




Block Gibbs Sampling

Divide **X** into several "**tractable blocks**" $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_B$.
Each block $\mathbf{X}_b$ can be **drawn jointly** given variables in other blocks.

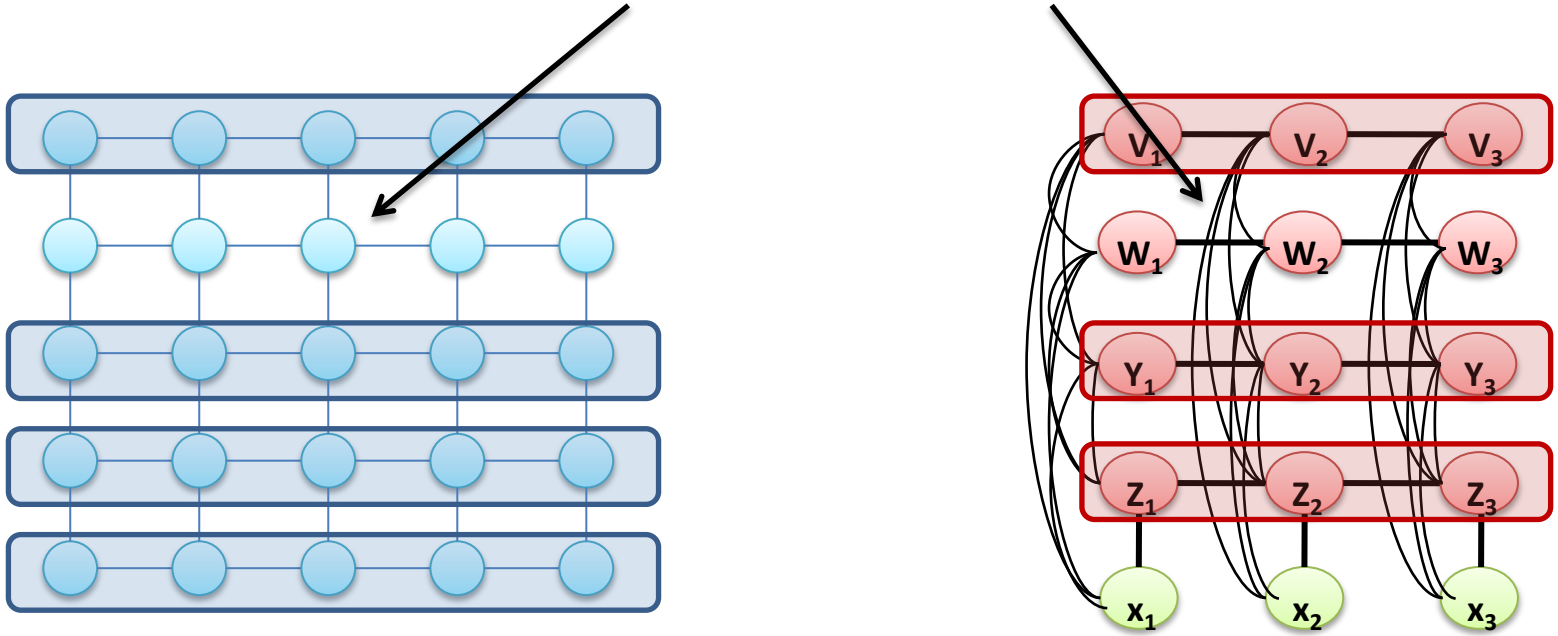
Given samples on other blocks,
Drawing a block jointly from is tractable.



Block Gibbs Sampling

Divide **X** into several "**tractable blocks**" $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_B$.
Each block $\mathbf{X}_b$ can be **drawn jointly** given variables in other blocks.

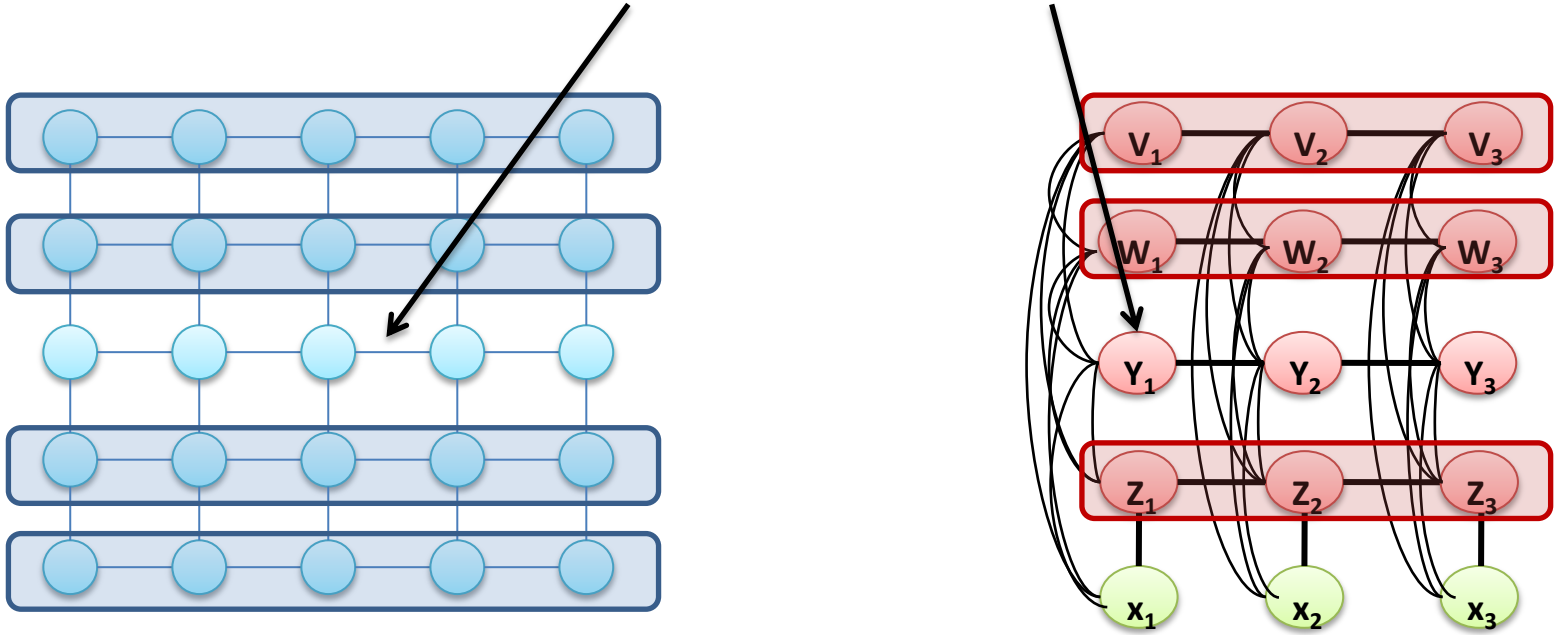
Given samples on other blocks,
Drawing a block jointly from is tractable.



Block Gibbs Sampling

Divide **X** into several "**tractable blocks**" $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_B$.
Each block $\mathbf{X}_b$ can be **drawn jointly** given variables in other blocks.

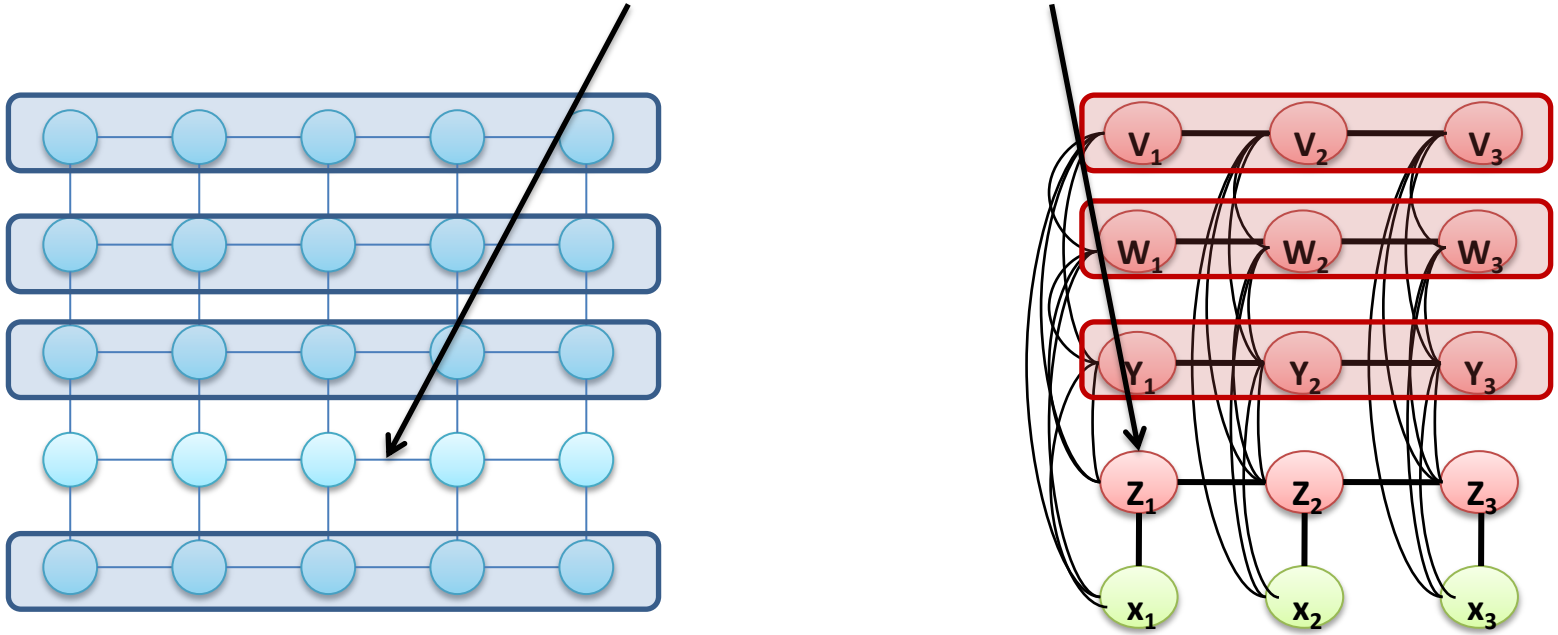
Given samples on other blocks,
Drawing a block jointly from is tractable.



Block Gibbs Sampling

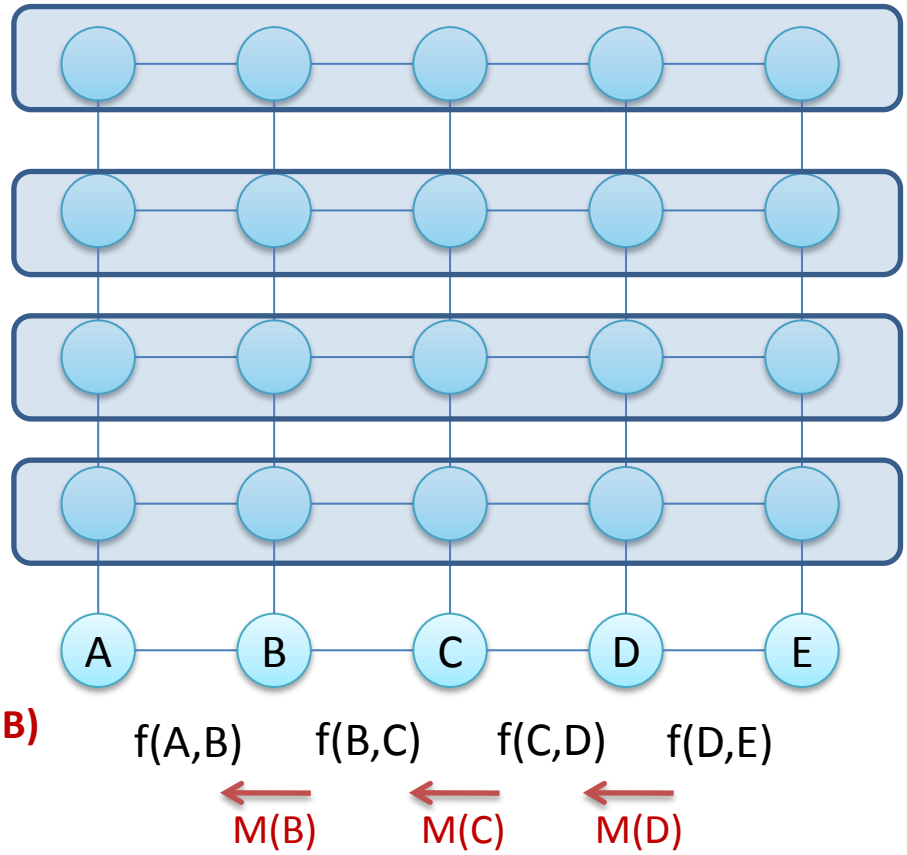
Divide **X** into several "**tractable blocks**" $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_B$.
Each block $\mathbf{X}_b$ can be **drawn jointly** given variables in other blocks.

Given samples on other blocks,
Drawing a block jointly from is tractable.



Block Gibbs Sampling by VE

Drawing from a block $\mathbf{X_b}$ **jointly** may need 1 pass of VE.

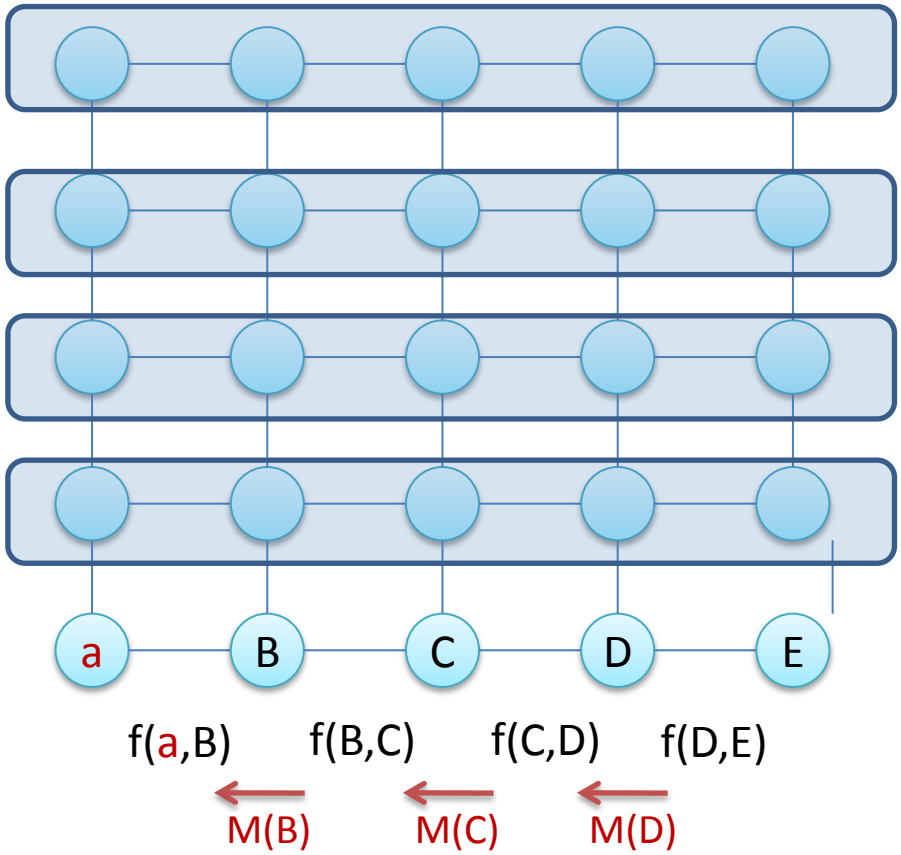


Draw A from:

$M(A)=\sum_B f(A,B)M(B)$

Block Gibbs Sampling by VE

Drawing from a block $\mathbf{X_b}$ **jointly** may need 1 pass of VE.

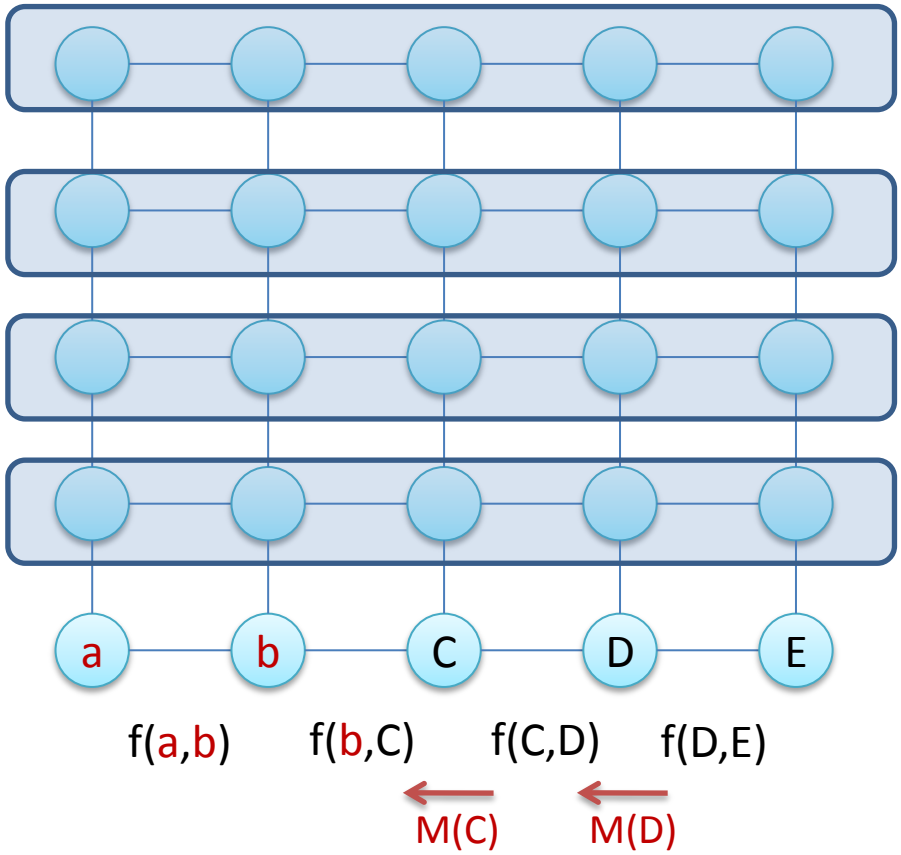


Draw B from:

f(A=a,B)M(B)

Block Gibbs Sampling by VE

Drawing from a block $\mathbf{X_b}$ **jointly** may need 1 pass of VE.

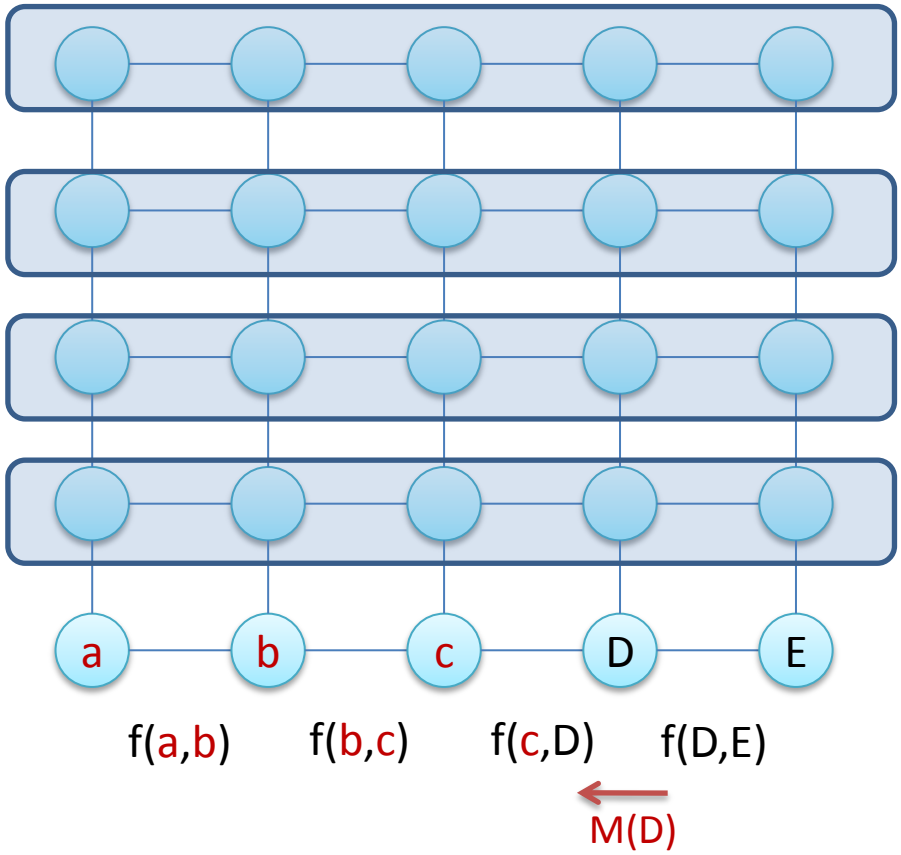


Draw C from:

f(B=b,C)M(C)

Block Gibbs Sampling by VE

Drawing from a block $\mathbf{X_b}$ **jointly** may need 1 pass of VE.

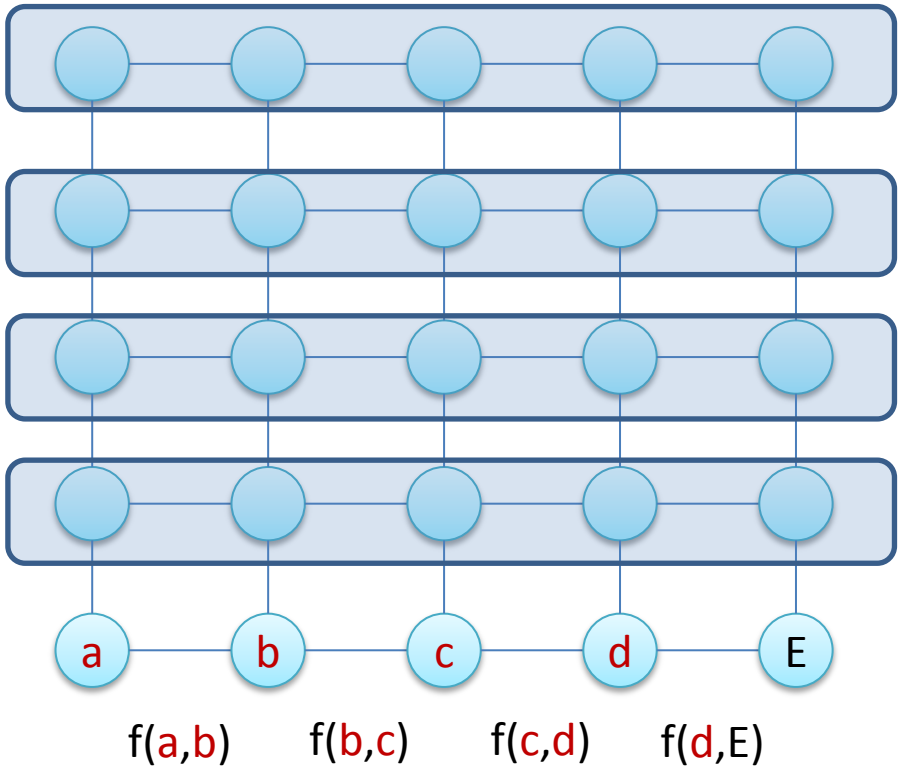


Draw D from:

f(C=c,D)M(D)

Block Gibbs Sampling by VE

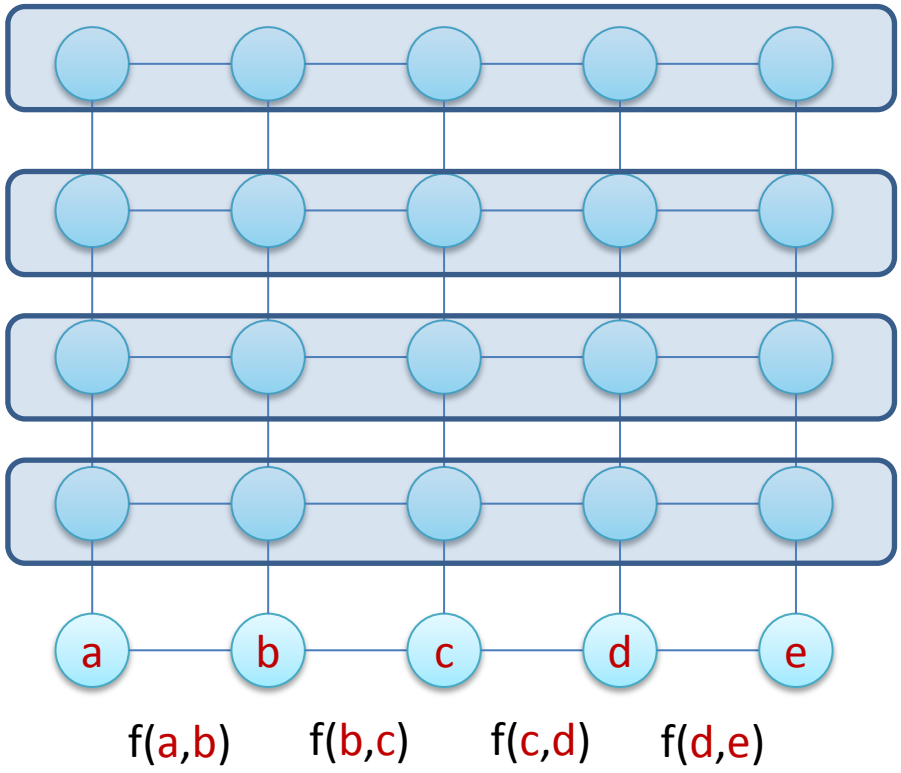
Drawing from a block $\mathbf{X_b}$ **jointly** may need 1 pass of VE.



Draw E from:
f(D=d,E)

Block Gibbs Sampling by VE

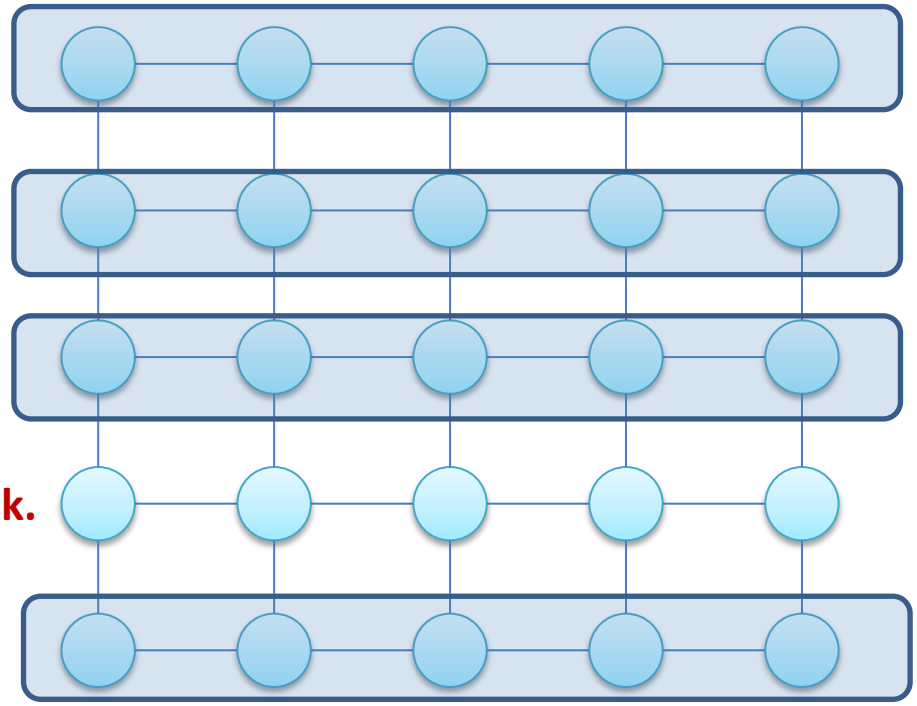
Drawing from a block $\mathbf{X_b}$ **jointly** may need 1 pass of VE.



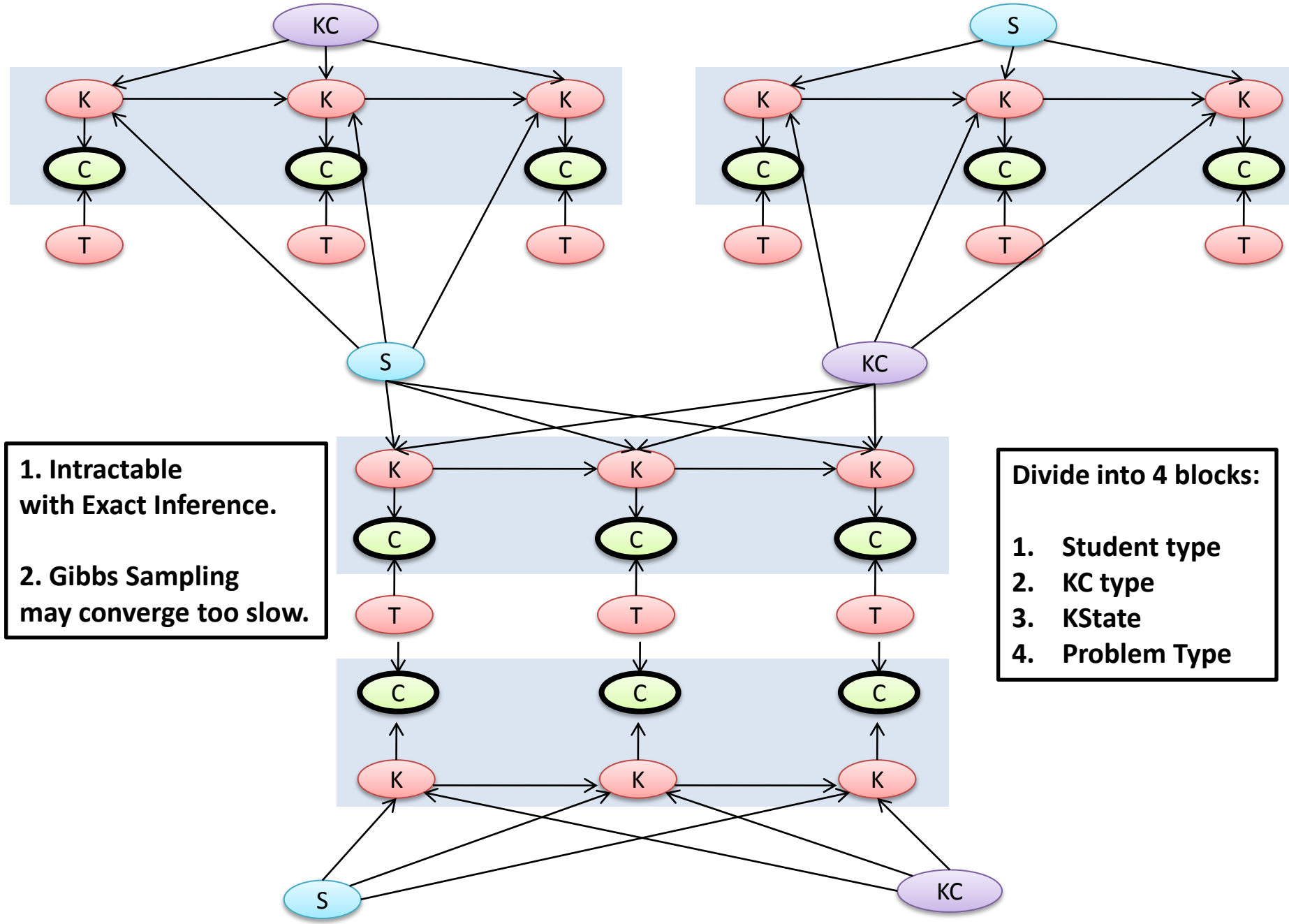
Draw E from:
f(D=d,E)

Block Gibbs Sampling by VE

Drawing from a block $\mathbf{X_b}$ **jointly** may need 1 pass of VE.



Draw Another Block.

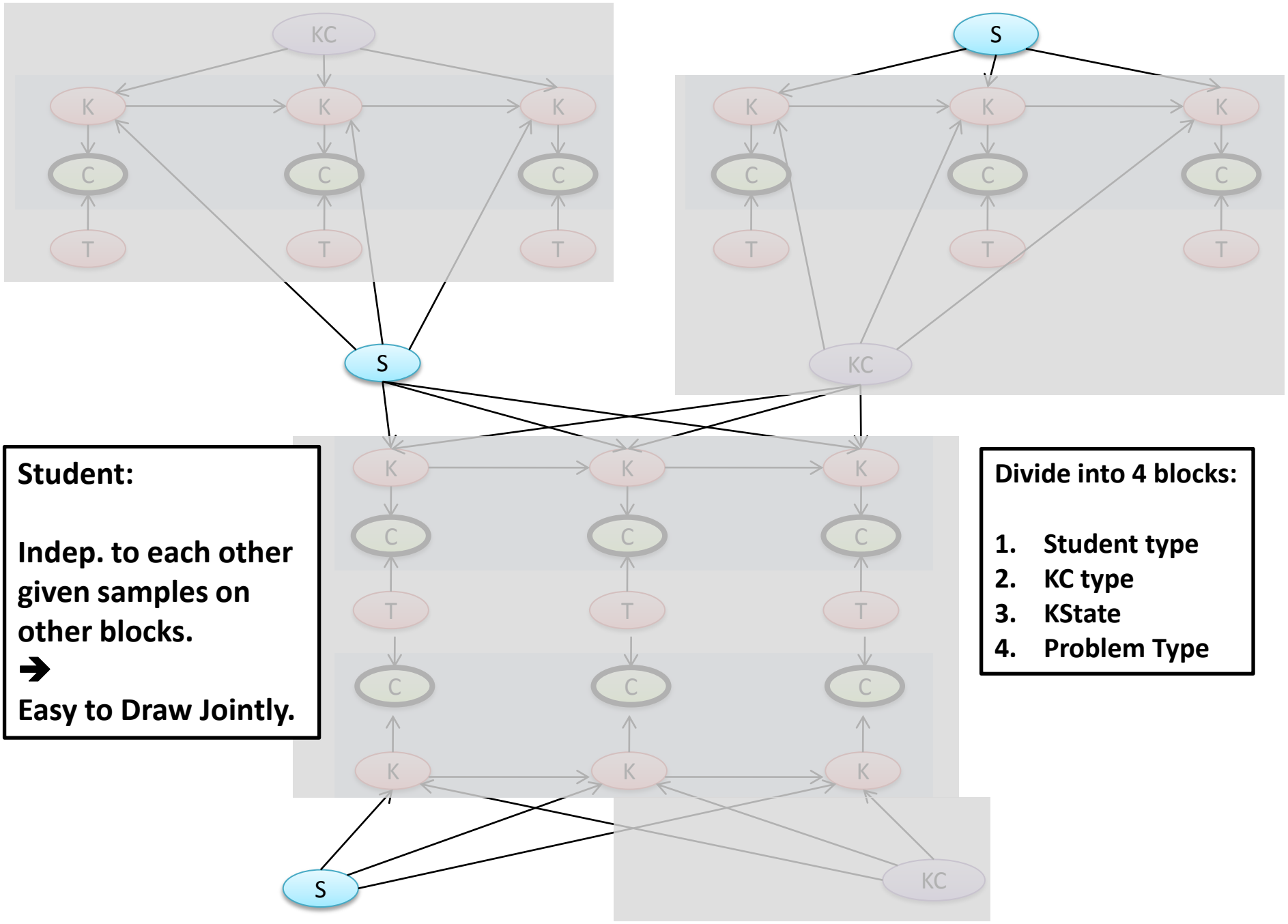


1. Intractable with Exact Inference.

2. Gibbs Sampling may converge too slow.

Divide into 4 blocks:

1. **Student type**
2. **KC type**
3. **KState**
4. **Problem Type**



Student:

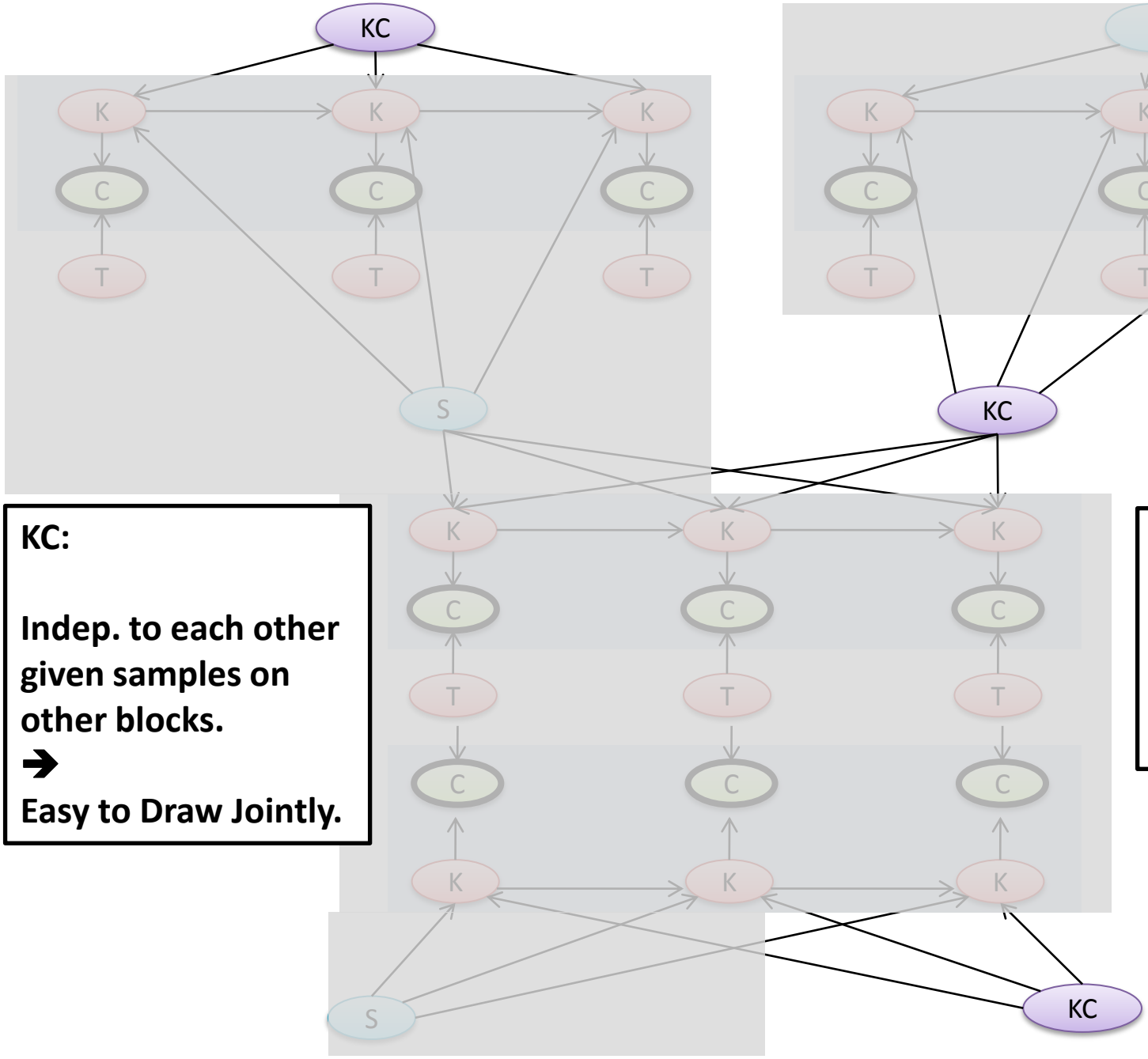
Indep. to each other given samples on other blocks.

➔

Easy to Draw Jointly.

Divide into 4 blocks:

1. **Student type**
2. **KC type**
3. **KState**
4. **Problem Type**



KC:

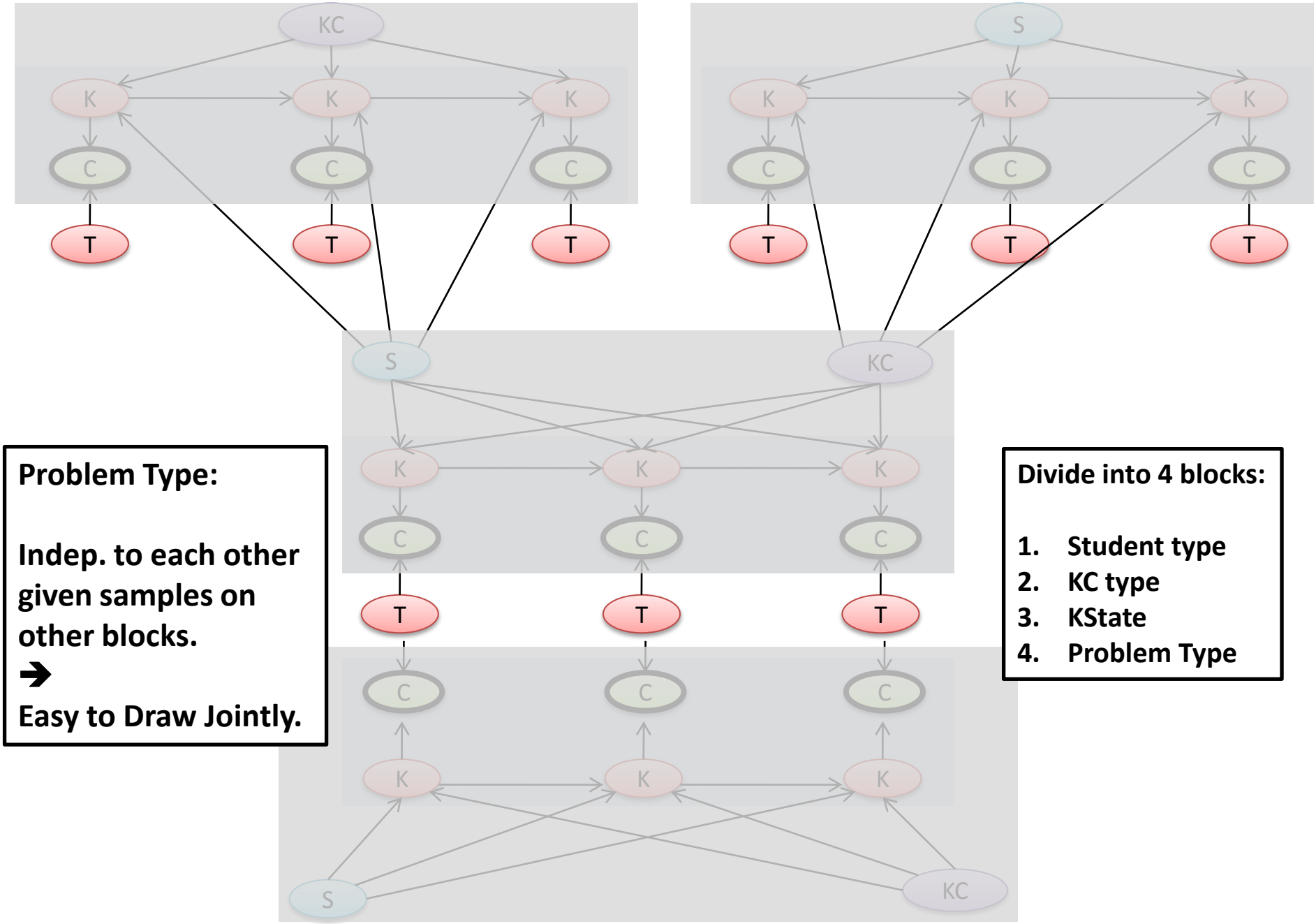
Indep. to each other given samples on other blocks.

➔

Easy to Draw Jointly.

Divide into 4 blocks:

1. **Student type**
2. **KC type**
3. **KState**
4. **Problem Type**



Problem Type:

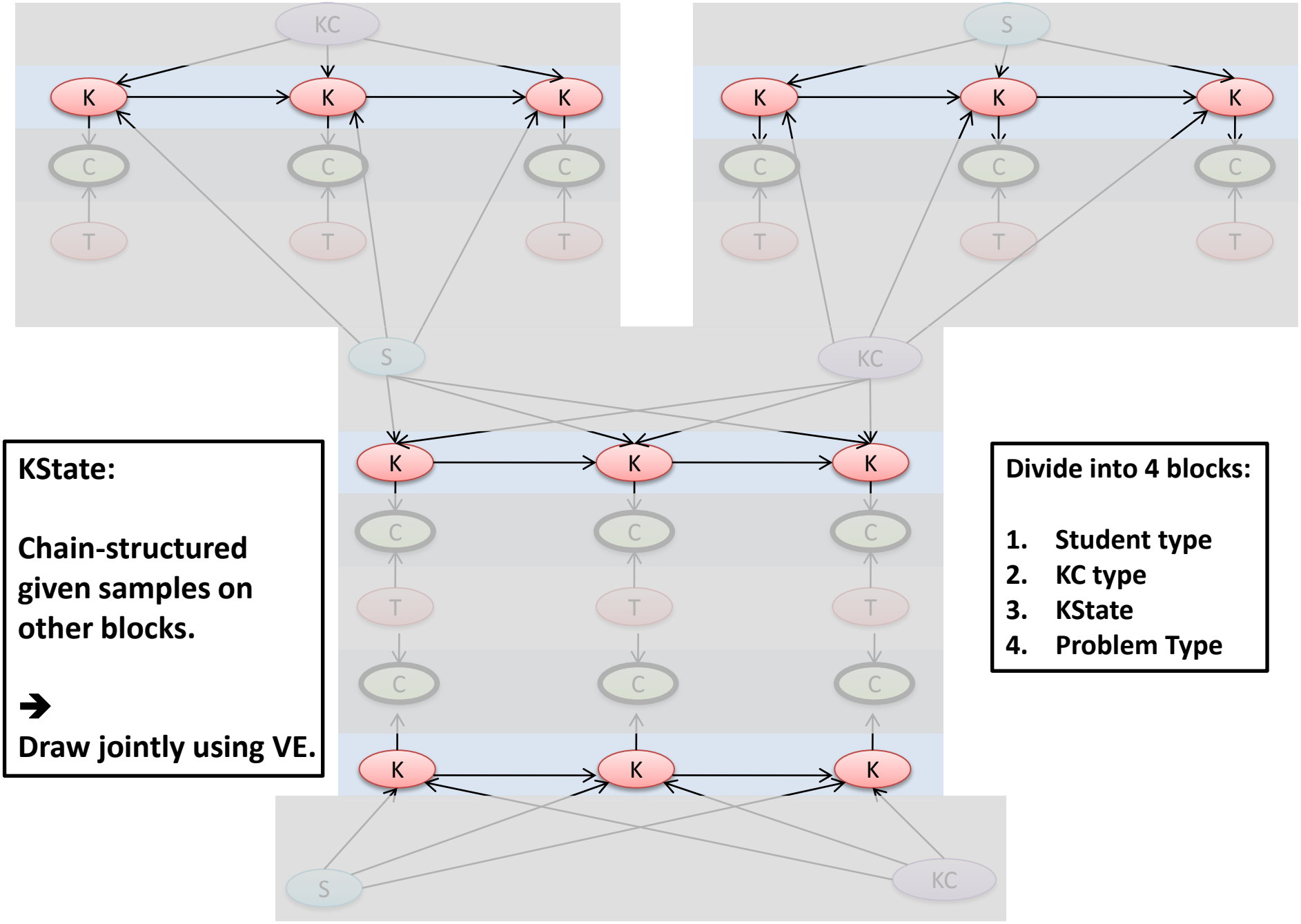
Indep. to each other given samples on other blocks.

➔

Easy to Draw Jointly.

Divide into 4 blocks:

1. **Student type**
2. **KC type**
3. **KState**
4. **Problem Type**



KState:

Chain-structured given samples on other blocks.

➔

Draw jointly using VE.

Divide into 4 blocks:

1. **Student type**
2. **KC type**
3. **KState**
4. **Problem Type**

Agenda

- When to use Approximate Inference ?
- Forward Sampling & Importance Sampling
- Markov Chain Monte Carlo (MCMC)
- Collapsed Particles

Collapsed Particle

Exact: $E_{P(X)}[f(X)] = \sum_X P(X) * f(X)$

Particle-Based: $\hat{f} = \frac{1}{N} \sum_{n=1}^{N} f(X^{(n)})$

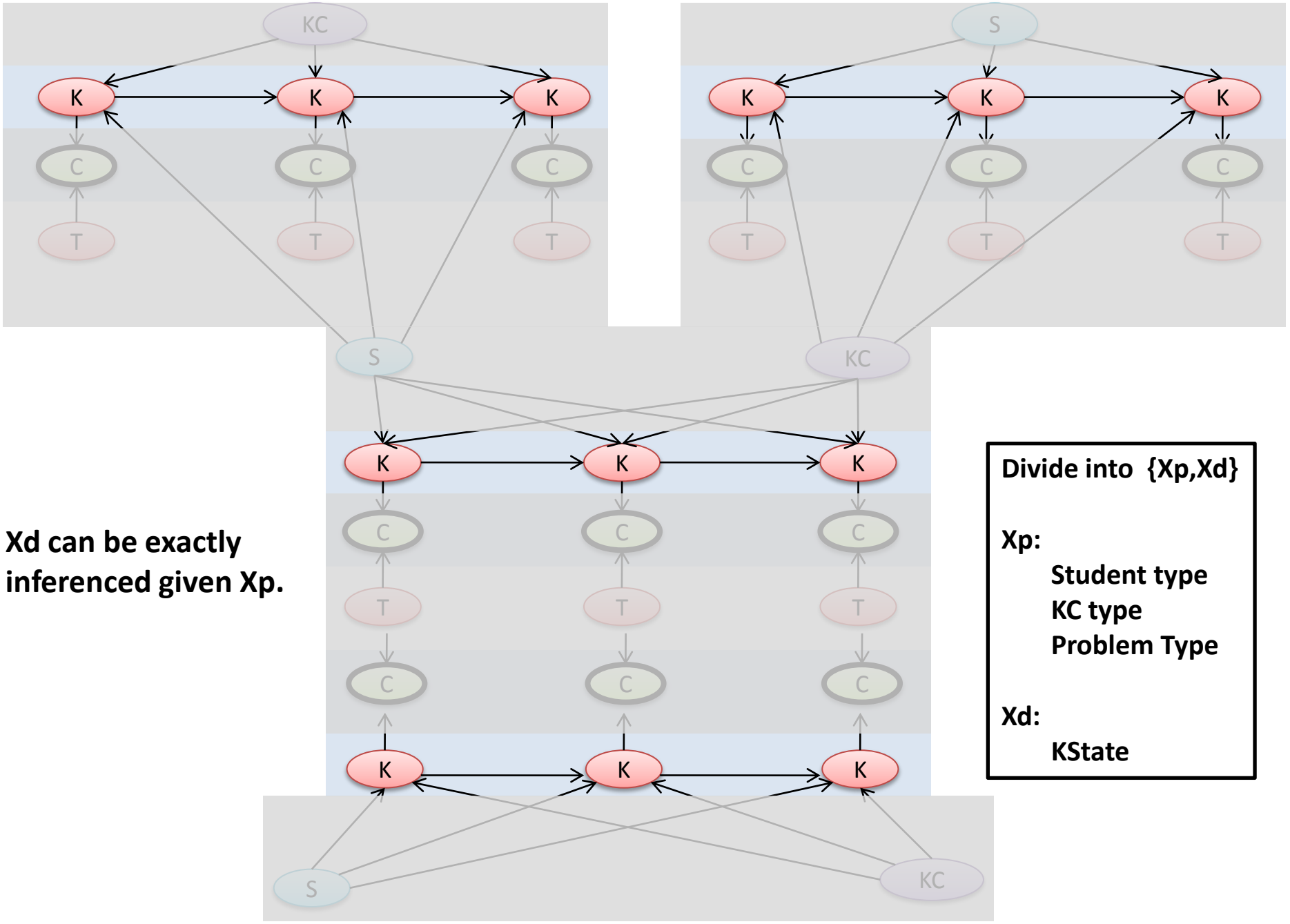
Collapsed-Particle:

Divide X into 2 parts {X_p, X_d} , where X_d can do inference given X_p

$$E_{P(X)}[f(X)] = \sum_X P(X) * f(X) = \sum_{Xp} P(X_p) \sum_{Xd} P(X_d \mid X_p) * f(X)$$

$$\hat{E}_{P(X)}[f(X)] = \frac{1}{N} \sum_{n=1}^{N} \left(\sum_{Xd} P(X_d \mid X_p^{(n)}) f(X_d, X_p^{(n)}) \right)$$

(If **X_p** contains few variables, Var. can be much reduced !!)



Xd can be exactly inferenced given Xp.

Divide into {Xp,Xd}

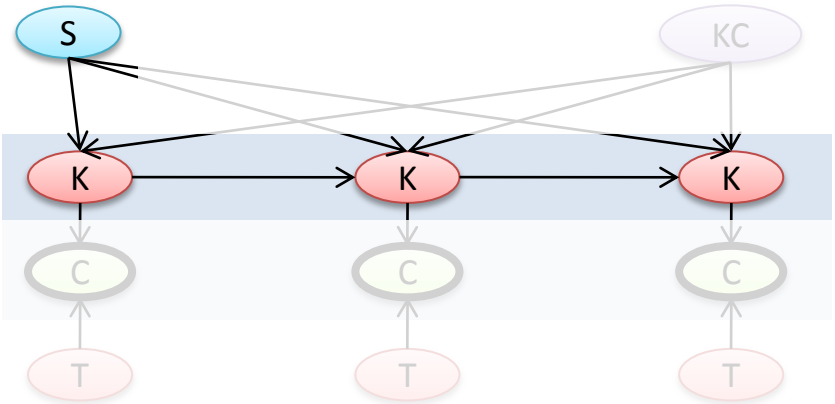
Xp:

- **Student type**
- **KC type**
- **Problem Type**

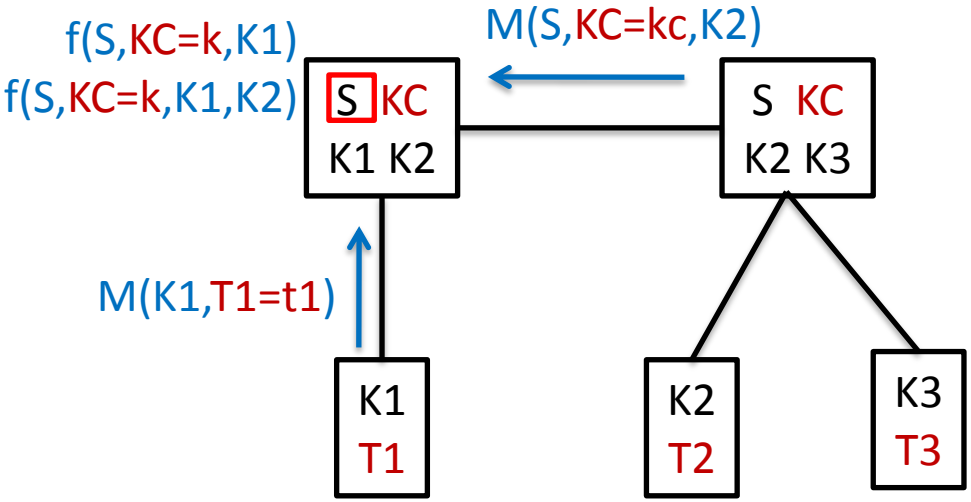
Xd:

- **KState**

Collapsed Particle with VE



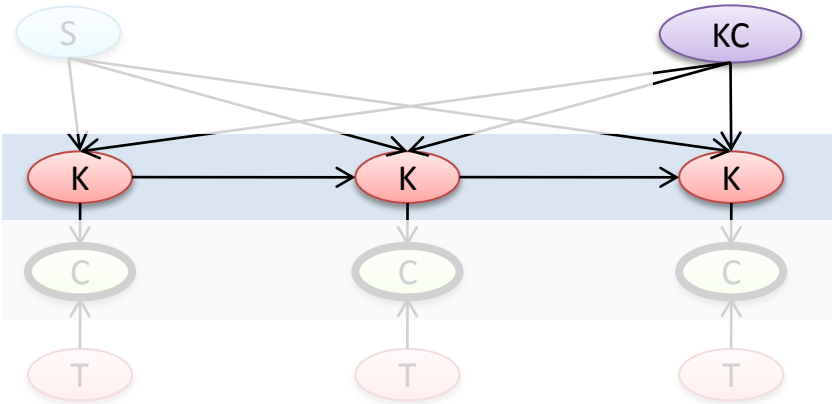
To draw $\mathbf{X_k}$,
Given all other variables in **Xp**
sum out all other variables in **Xd**



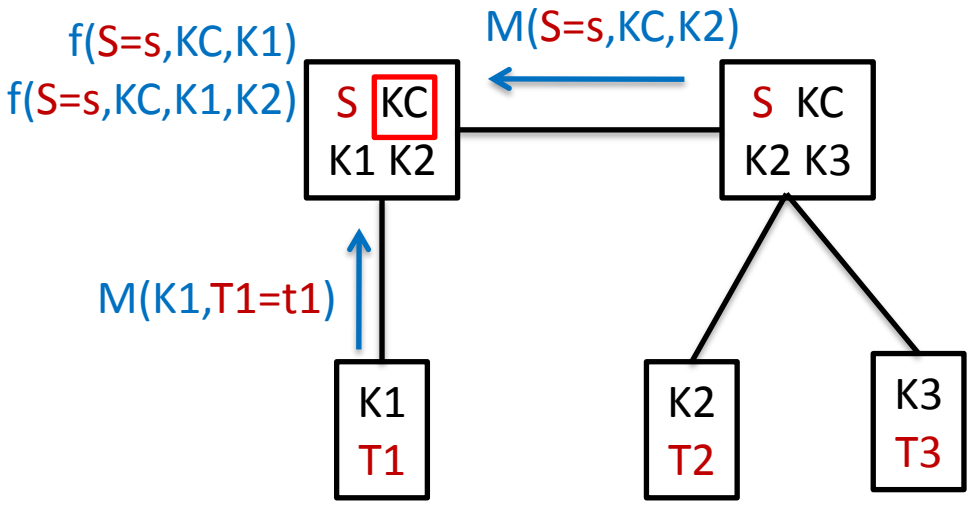
Draw S (given KC=k & T=t) from:

$$M(S) = \sum_{K1,K2} F(S,KC=k,K1,K2)\, M(K1,T1=t1)\, M(S, KC=k,K2)$$

Collapsed Particle with VE



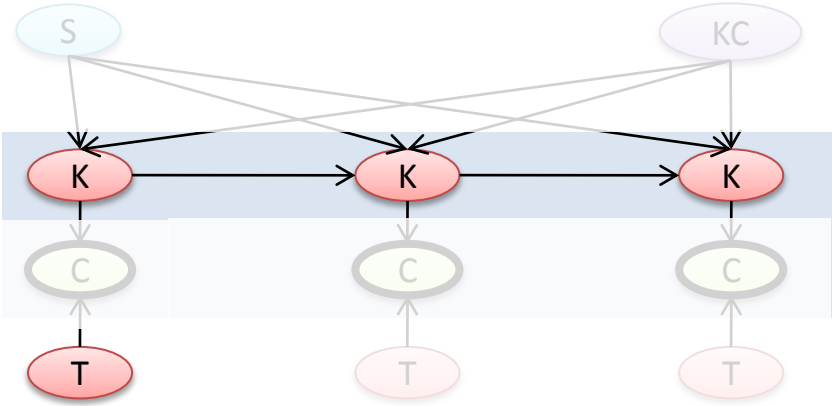
To draw $\mathbf{X_k}$,
Given all other variables in **Xp**
sum out all other variables in **Xd**



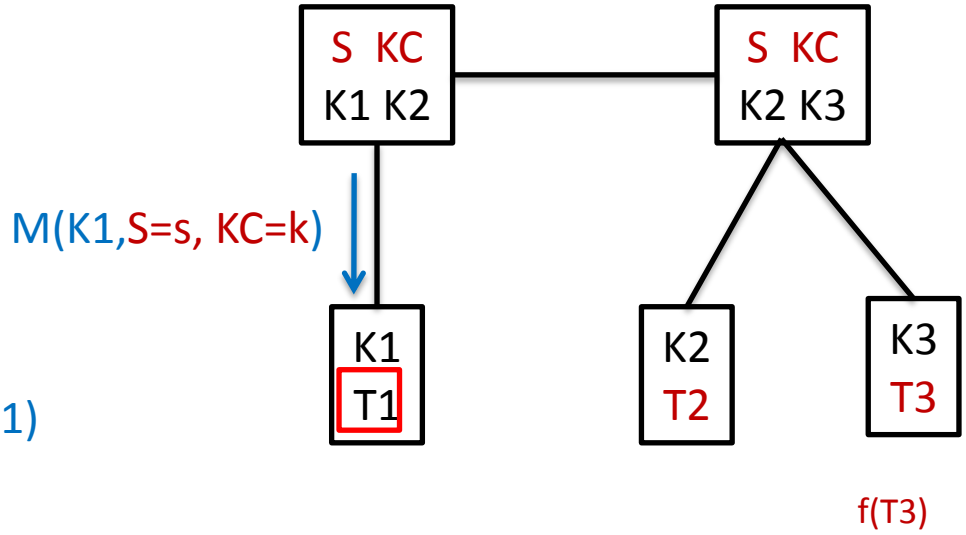
Draw KC (given S=s & T=t) from:

$$M(KC) = \sum_{K1,K2} F(S=s,KC,K1,K2)\ M(K1,T1=t1)\ M(S=s,KC,K2)$$

Collapsed Particle with VE



To draw $\mathbf{X_k}$,
Given all other variables in **Xp**
sum out all other variables in **Xd**



Draw T1 (given S=s & KC=k) from:

$$M(T1) = \sum_{K1} M(K1,S=s,KC=k)\ F(K1,T1)$$

Collect Samples

Xp	**Xd**
(S, KC, T1, T2, T3)	(K1, K2, K3)
(Intel, Quick, Hard, Easy, Hard)	({1/3,1/3,1/3} , {1/4,1/4,1/2}, {1/2,1/2,0})
(Intel, Slow, Easy, Easy, Hard)	({1/2,1/2,1/4} , {1/5,4/5,0}, {1/4,1/4,1/2})
…..	…..
(Dull, Slow, Easy, Easy, Hard)	({1/3,1/3,1/3} , {1/4,1/4,1/2}, {1/2,1/2,0})
Average	Average

$$\hat{E}_{P(X)}[f(X)] = \frac{1}{N}\sum_{n=1}^{N} \left(\sum_{Xd} P(X_d \mid X_p^{(n)}) f(X_d, X_p^{(n)})\right)$$