Suppose we have a graphical model with closed loops.

**Example: Potts Model in 2D**

\[ E(x) = \sum_{i \neq j} \varphi(x_{ij}, x_{ij}) \]

\[ N(ij) = \langle (i-1,j), (i+1,j) \rangle \]

\[ T(i)(x) = \frac{1}{Z} e^{-E(x)} \]

**How to sample from \( T(i)(x) \)?**

We could try SIS — i.e., choose an order \( x_1, \ldots, x_n \) of the variables \( \langle x_{ij} \rangle \).

Select trial distributions \( g(x_t | x_1, \ldots, x_{t-1}) \) and approximate distributions \( T_t(x_t) \)

\( x_t \approx \langle x_t \rangle \)

Such that \( T_0(x_0) = T(i)(x) \).

The problem is that we do not know how to specify \( g( \cdot | \cdot, \ldots) \) and \( T_t(x_t) \). Unless these relate closely to \( T(i)(x) \), the SIS samples will be in the wrong parts of the distribution and have low weight.

So we may need an enormous number of samples in order to get a good estimator.

So SIS will usually not work well.
MCMC is a way to sample from any distribution $\pi(x)$.

It does not sample directly from $\pi(x)$. Instead it proceeds by defining a Markov Chain that converges to samples from $\pi(x)$.

The original MCMC is the Metropolis algorithm (Metropolis, Rosenbluth & Rosenbluth, Teller & Teller 1953). Later generalized to the Metropolis-Hastings algorithm. Strictly speaking, these are not algorithms. They are design principles for algorithms.

For any distribution $\pi(x)$, there are many different Metropolis-Hastings algorithms that we can design to obtain samples from $\pi(x)$. Some will be much more efficient than others.

Not all MCMC are Metropolis-Hastings. First we will specify the most general MCMC. Then we will introduce Metropolis and Metropolis-Hastings.
Markov Chain.

Let \( P(x_{t+1} \mid x_t) = K(x_{t+1} \mid x_t) \) be the transition kernel.

\[ K(\cdot \mid \cdot) \] must obey \( K(x_{t+1} \mid x_t) \geq 0 \), for all \( x_t, x_{t+1} \)

and \( \sum_{x_{t+1}} K(x_{t+1} \mid x_t) = 1 \), for all \( x_t \)

We can obey a set of samples from this chain.

\( x^0 \) — randomly initialized.
\( x^1 \) — sampled from \( K(x^1 \mid x^0) \)
\( x^2 \) — sampled from \( K(x^2 \mid x^1) \)
\( \ldots \)
\( x^n \) — sampled from \( K(x^n \mid x^{n-1}) \)

MCMC is a special MC chosen so that the transition kernel \( K(x \mid y) \) satisfies the fixed point condition

\[ \sum_y K(x \mid y) \pi(y) = \pi(x) \] \hspace{1cm} (1)

or the detailed balance condition

\[ K(x \mid y) \pi(y) = K(y \mid x) \pi(x) \] \hspace{1cm} (2)

Note: detailed balance implies fixed point because

\[ \sum_y K(x \mid y) \pi(y) = \sum_y K(y \mid x) \pi(x) = \pi(x) \sum_y K(y \mid x) = \pi(x) \]

Note: fixed point condition means intuitively that if we sample \( y \) from \( \pi(y) \), next sample \( x \) from \( K(x \mid y) \), then \( x \) is a sample from \( \pi(x) \)

Note: in practice, most transition kernels are chosen to obey Detailed Balance (simpler to check).
We also require an MCMC to be irreducible, so that for any $x, y$ we can find a sequence $x_1, \ldots, x_n$ such that $K(x_1|x_2) K(x_2|x_3) \ldots K(x_n|y) > 0$. (for some $n$)

i.e. there is a set of moves that take us from any point $x$ to any other point $y$.

(Equivalently \[ \sum_{y_1, \ldots, y_n} K(x_1|y_1) K(y_1|y_2) \ldots K(y_{n-1}|y_n) > 0 \])

These conditions ensure that samples from the MCMC will eventually tend to sample from $\pi(x)$.

i.e. $x_0, x_1, \ldots, x_t, x_N, \ldots$ with $x_t$ sampled from $K(x_t|x_t)$

then, for large enough $N$, $x_N$ is a sample $\pi(x)$

(Proof will be given in a later lecture).

This means that we can sample from any distribution $\pi(x)$.

This is too good to be true. We could solve NP-complete problems.

The difficulty is how big $N$ must be in order for $x_N$ to be a sample from $\pi(x)$. 
The Metropolis Algorithm

Suppose $\Pi(x) = e^{-E(x)}$ is the Gibbs distribution, $Z = \sum_x e^{-E(x)}$ unknown, hard to calculate.

For each $x$ define a neighborhood $N(x)$ such that $y \in N(x)$ implies $x \in N(y)$ (for all $x, y$) and the size $|N(x)|$ of the neighborhood is independent of $x$.

Basic Metropolis:

Loop over $t$ state $x^t$ at time $t$.

1. **Propose a move/transition from $x^t$ to $x^{t+1} \in N(x^t)$ with uniform probability — i.e., $p(x^{t+1}) = \frac{1}{|N(x^t)|}$.
2. Accept the move with probability $\min \left\{ 1, \frac{\Pi(x^{t+1})}{\Pi(x^t)} \right\} = \min \left\{ 1, e^{E(x^t) - E(x^{t+1})} \right\}$. 

Note: independent of $Z$.

This has transition kernel:

$K(x^{t+1}|x^t) = 0$, if $x^{t+1} \notin N(x^t)$.

$K(x^{t+1}|x^t) = \frac{1}{|N(x^t)|} \min \left\{ 1, \frac{\Pi(x^{t+1})}{\Pi(x^t)} \right\}$, if $x^{t+1} \in N(x^t)$,

$K(x^t|x^t) = 1 - \sum_{x^{t+1}\notin N(x^t)} K(x^{t+1}|x^t)$.

Metropolis obeys detailed balance because for $x^{t+1}, x^t$:

$K(x^{t+1}|x^t) \Pi(x^t) = \frac{1}{|N(x^t)|} \min \{ \Pi(x^t), \Pi(x^{t+1}) \} = K(x^t|x^{t+1}) \Pi(x^{t+1})$.

(with equality also if $x^t = x^{t+1}$). (Recall $|N(x^{t+1})| = |N(x^t)|$ constant).

Hence Metropolis will generate samples from $\Pi(x)$ (assuming irreducibility).