# On the speed of convergence of Markov chain Monte Carlo methods

# Outline

# Outline

- MCMC

# Outline

- MCMC
- Upper and lower bounds for relaxation times

# Outline

- MCMC
- Upper and lower bounds for relaxation times
- Sophisticated samplers

# Outline

- MCMC
- Upper and lower bounds for relaxation times
- Sophisticated samplers
- Negative results

# Outline

- MCMC
- Upper and lower bounds for relaxation times
- Sophisticated samplers
- Negative results
- Myth busting

# Outline

- MCMC
- Upper and lower bounds for relaxation times
- Sophisticated samplers
- Negative results
- Myth busting
- More negative results

# Outline

- MCMC
- Upper and lower bounds for relaxation times
- Sophisticated samplers
- Negative results
- Myth busting
- More negative results
- Some hope

# Outline

- MCMC
- Upper and lower bounds for relaxation times
- Sophisticated samplers
- Negative results
- Myth busting
- More negative results
- Some hope
- Or proving that there's no hope...

# Markov chain Monte Carlo

"Almost all Markov chains can be modified to converge to a prescribed distribution"

Primer Markov chain with transition probabilities T(Y|X), aperiodic, irreducible and for all X and Y,

$$T(Y|X) \neq 0 \implies T(X|Y) \neq 0$$

Metropolis-Hastings algorithm:

Step 1. (proposal) Propose a $Y \sim T(\cdot|X_n)$

Step 2. (acceptance) Accept Y with probability

$$\min\left\{1, \frac{\pi(Y)T(X_n|Y)}{\pi(X)T(Y|X_n)}\right\}$$

# Mixing of Markov chains

"Mixing correlates with the second largest eigenvalue"

SLEM: Second largest eigenvalue modulus

$$\rho = \min\{\lambda_2, |\lambda_r|\}$$

Relaxation time

$$\tau_{rel} := \frac{1}{1-\lambda_2}$$

$$\tau_i(\varepsilon) := \min\left\{k_0 : \forall k > k_0, d_v(P^k\delta_i, \pi) \leq \varepsilon\right\}$$

Theorems (Aldous, 1982; Diaconis & Stroock,1991)

$$\tau_i(\varepsilon) \leq \frac{1}{1-\rho}\left(\ln\left(\frac{1}{\pi(i)}\right) + \ln\left(\frac{1}{\varepsilon}\right)\right)$$

$$\max_{i\in I} \tau_i(\varepsilon) \geq \frac{\rho}{2(1-\rho)} \ln(1/\varepsilon)$$

# Cheeger inequalities

“A Markov chain mixes slowly if and only if there is a bottleneck in it”

Ergodic flow:

$$F(S) := \sum_{i \in S, j \in i \setminus S} \pi(i) p_{ij}$$

Conductance:

$$\Phi := \min_{S \in I} \left\{ \frac{\sum_{i \in S, j \in i \setminus S} \pi(i) p_{ij}}{\pi(S)} \middle| 0 < \pi(S) \leq \frac{1}{2} \right\}$$

Cheeger inequality:

$$1 - 2\Phi \leq \lambda_2 \leq 1 - \frac{\Phi^2}{2}$$

# MCMC convergence

# MCMC convergence

$D_1$ $D_2$ $D_3$ $D_4$ $D_5$ $D_6$ $D_7$ ... $D_n$

# MCMC convergence

$$\begin{tikzcd}
D_1 \arrow[d] & D_2 \arrow[d] & D_3 \arrow[d] & D_4 \arrow[d] & D_5 \arrow[d] & D_6 \arrow[d] & D_7 \arrow[d] & \dots \arrow[d] & D_n \arrow[d] \\
M_1 & M_2 & M_3 & M_4 & M_5 & M_6 & M_7 & \dots & M_n
\end{tikzcd}$$

# MCMC convergence

$$\begin{tikzcd}
D_1 \arrow[d] & D_2 \arrow[d] & D_3 \arrow[d] & D_4 \arrow[d] & D_5 \arrow[d] & D_6 \arrow[d] & D_7 \arrow[d] & \ldots \arrow[d] & D_n \arrow[d] \\
M_1 & M_2 & M_3 & M_4 & M_5 & M_6 & M_7 & \ldots & M_n
\end{tikzcd}$$

## Fast convergence

$\exists\, k\ \forall\, D_n\ (|D_n| \leq n)$ the series of $1/(n^k(1-\lambda_{n,2}))$ built from the $\lambda_{n,2}$'s of the Markov chains $M_n$ defined by the algorithms converges to 0.

# MCMC convergence

$$\begin{tikzcd}
D_1 \arrow[d] & D_2 \arrow[d] & D_3 \arrow[d] & D_4 \arrow[d] & D_5 \arrow[d] & D_6 \arrow[d] & D_7 \arrow[d] & \dots \arrow[d] & D_n \arrow[d] \\
M_1 & M_2 & M_3 & M_4 & M_5 & M_6 & M_7 & \dots & M_n
\end{tikzcd}$$

## Fast convergence

$\exists k \ \forall D_n$ $(|D_n| \leq n)$ the series of $1/(n^k(1-\lambda_{n,2}))$ built from the $\lambda_{n,2}$'s of the Markov chains $M_n$ defined by the algorithms converges to 0.

## Slow convergence

$\exists D_n$ $(|D_n| \leq n)$ s.t. $\forall k$ the series of $1/(n^k(1-\lambda_{n,2}))$ built from the $\lambda_{n,2}$'s of the Markov chains $M_n$ defined by the algorithms does not converge to 0.

# Fast converging Markov chains

FPRAS:

(Fully Polynomial Randomized Approximation Scheme)

$$\text{given } x \in \Sigma^*, \varepsilon > 0, 0 < \delta < 1:$$
$$\Pr((1-\varepsilon) f(x) \le OUT \le (1+\varepsilon) f(x)) \ge 1-\delta$$

Algorithm polynomial in $|x|$, $1/\epsilon$ and $-\log(\delta)$

#P: Hard counting problem

# Fast converging Markov chains

FPRAS:

(Fully Polynomial Randomized Approximation Scheme)

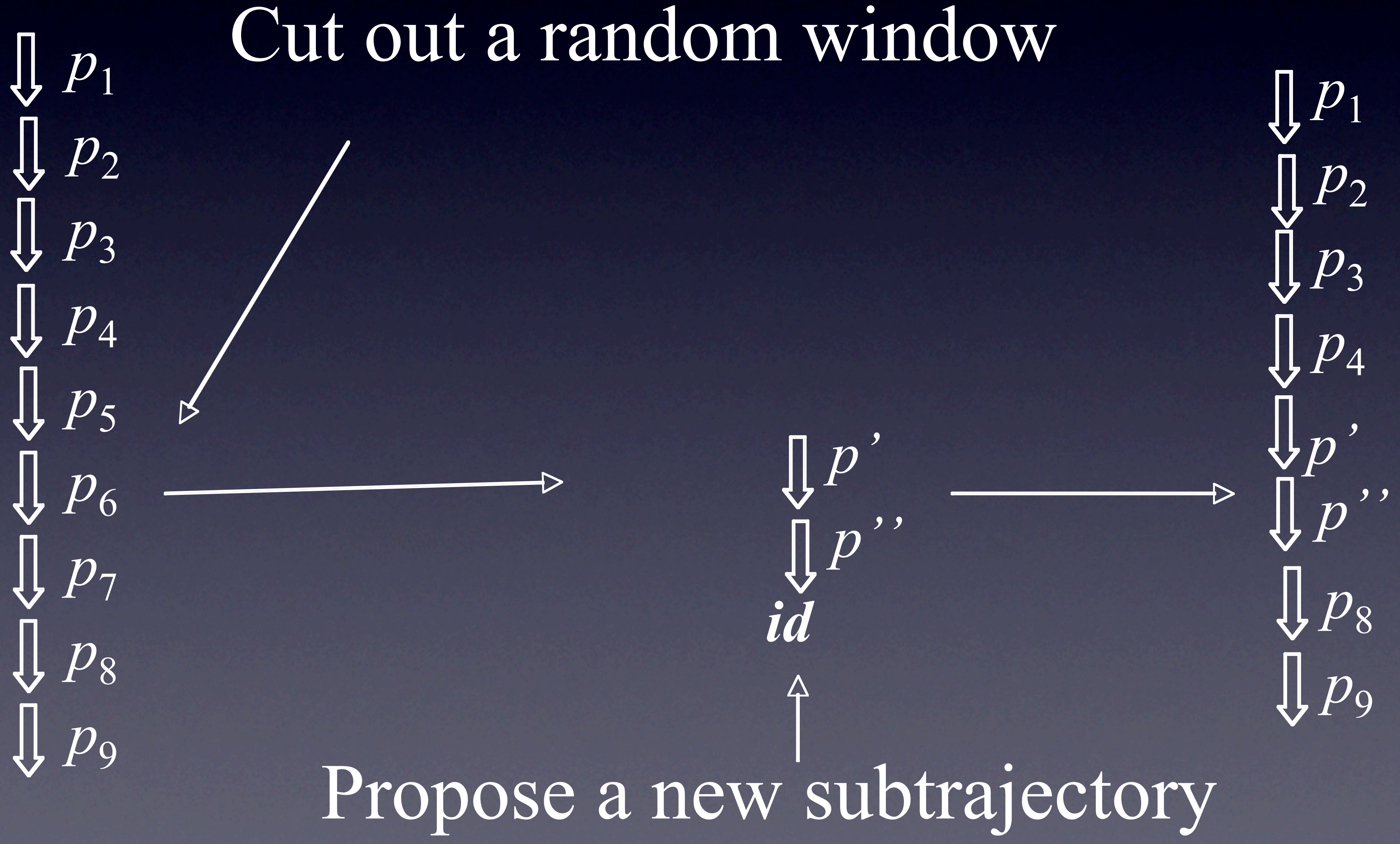
$$\text{given } x \in \Sigma^*, \varepsilon > 0, 0 < \delta < 1:$$
$$\Pr((1-\varepsilon) f(x) \leq OUT \leq (1+\varepsilon) f(x)) \geq 1-\delta$$

Algorithm polynomial in |x|, 1/ϵ and -log(δ)

#P: Hard counting problem

It is known that the intersection of #P and FPRAS is not empty. For example linear extension of partially ordered sets.

# ParIS: Partial Importance Sampling



Cut out a random window

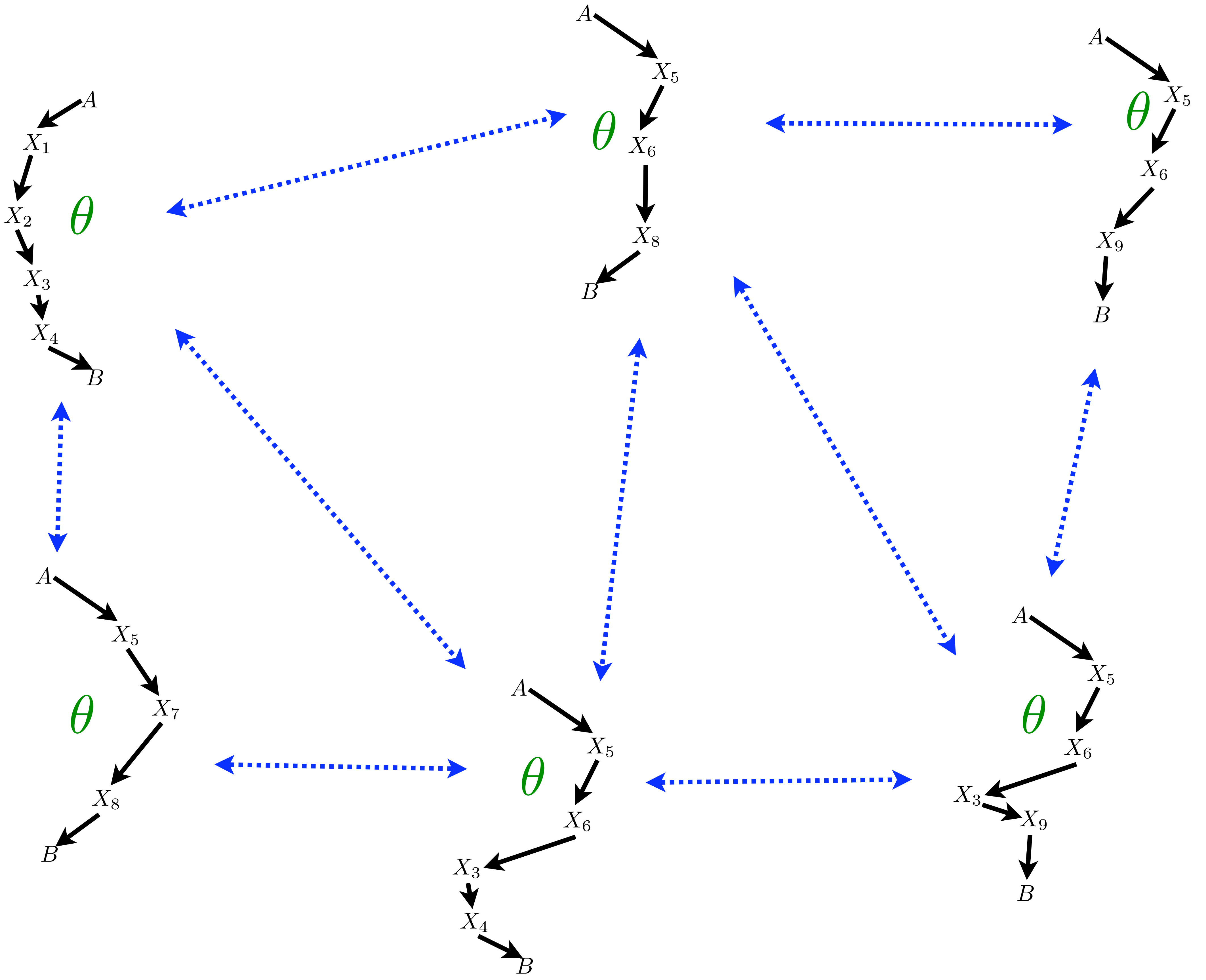
$\Downarrow p_1$
$\Downarrow p_2$
$\Downarrow p_3$
$\Downarrow p_4$
$\Downarrow p_5$
$\Downarrow p_6$
$\Downarrow p_7$
$\Downarrow p_8$
$\Downarrow p_9$

$\Downarrow p'$
$\Downarrow p''$
$\boldsymbol{id}$

$\Downarrow p_1$
$\Downarrow p_2$
$\Downarrow p_3$
$\Downarrow p_4$
$\Downarrow p'$
$\Downarrow p''$
$\Downarrow p_8$
$\Downarrow p_9$

Propose a new subtrajectory

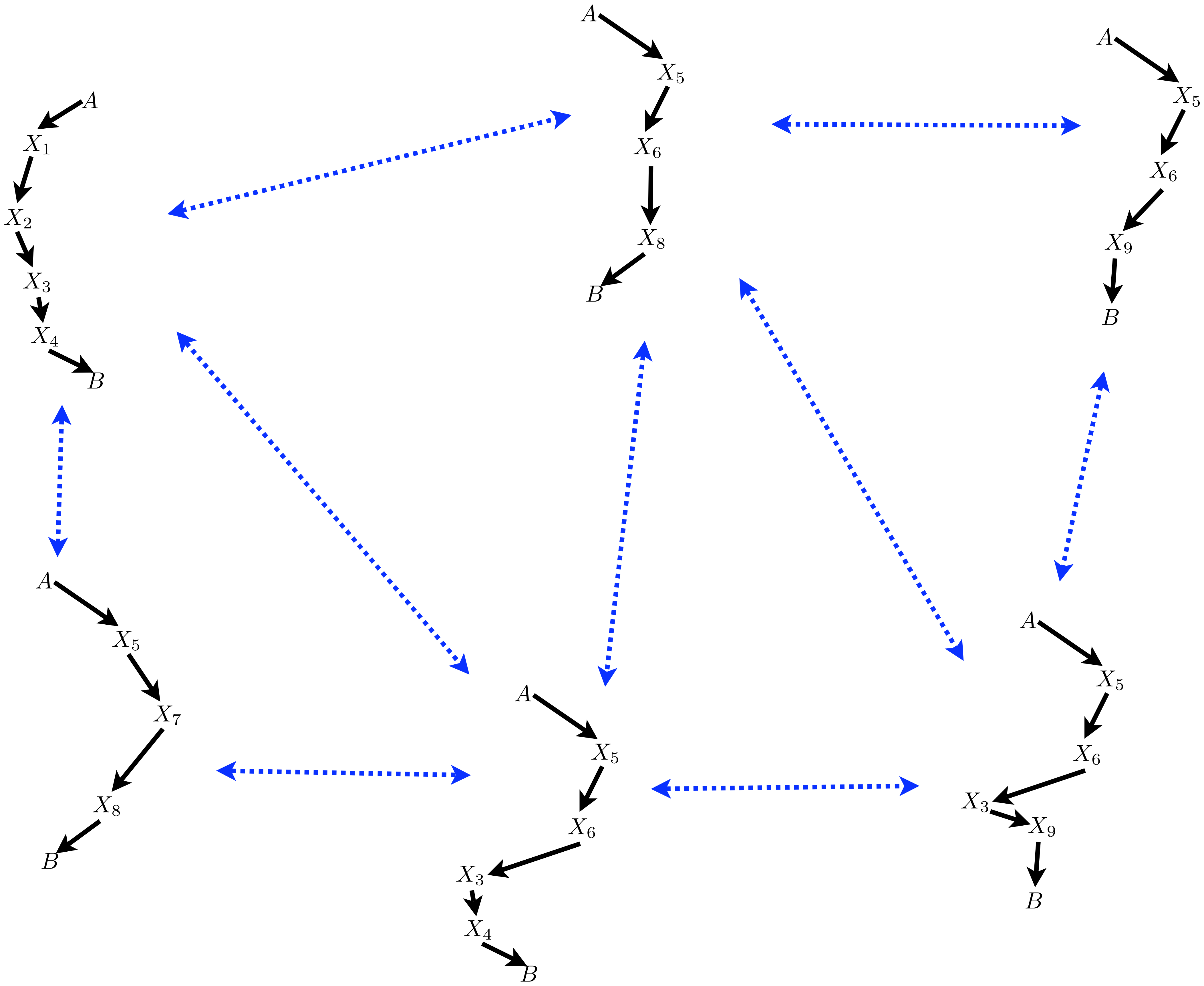
# Random walk on random walks



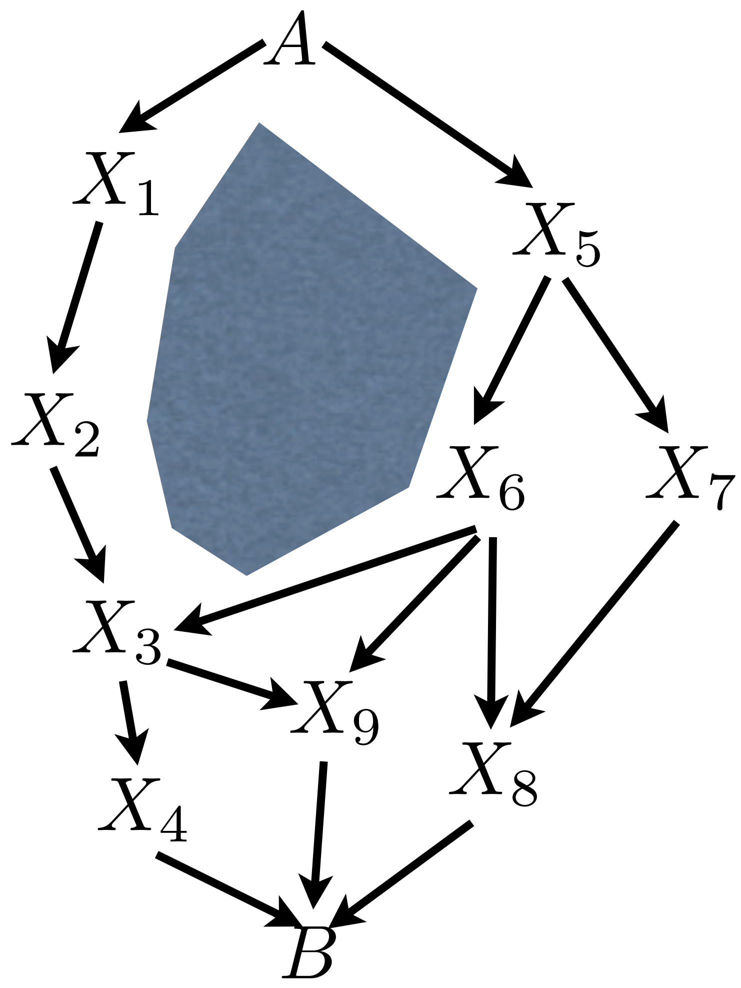
# Examples for RWoRWs

- Genome rearrangement paths
- Ancestral states in the infinite site recombination model
- Sequence alignment
- etc.

# Big islands

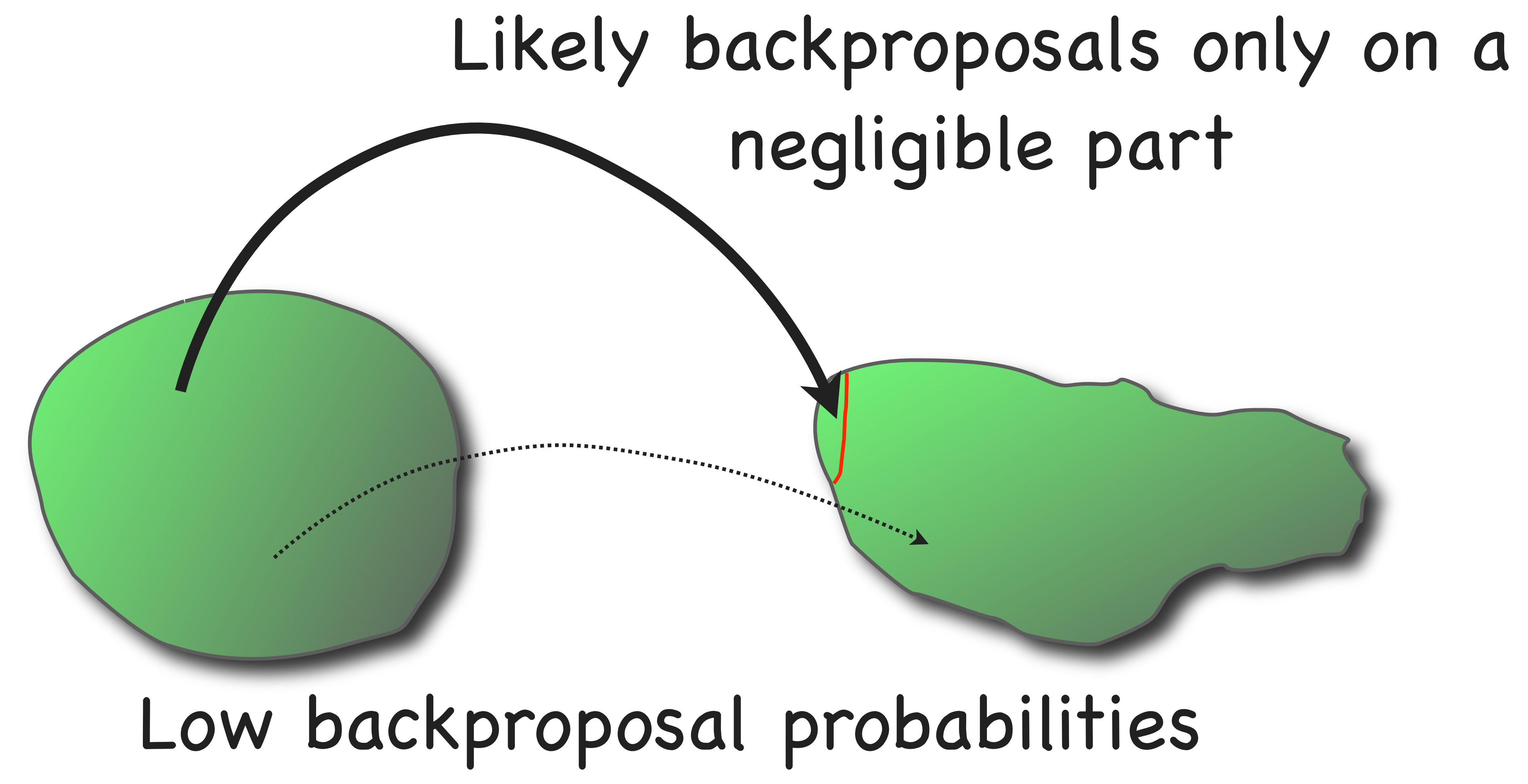

$A$
$X_1$
$X_2$
$X_3$
$X_4$
$B$
$A$
$X_5$
$X_6$
$X_8$
$B$
$A$
$X_5$
$X_6$
$X_9$
$B$
$A$
$X_5$
$X_7$
$X_8$
$B$
$A$
$X_5$
$X_6$
$X_3$
$X_4$
$B$
$A$
$X_5$
$X_6$
$X_3$
$X_9$
$B$

# Big islands



# ParIS mixes slowly on minimal reversal sorting paths

Miklós, Mélykúti, Swenson, manuscript submitted



Likely backproposals only on a negligible part

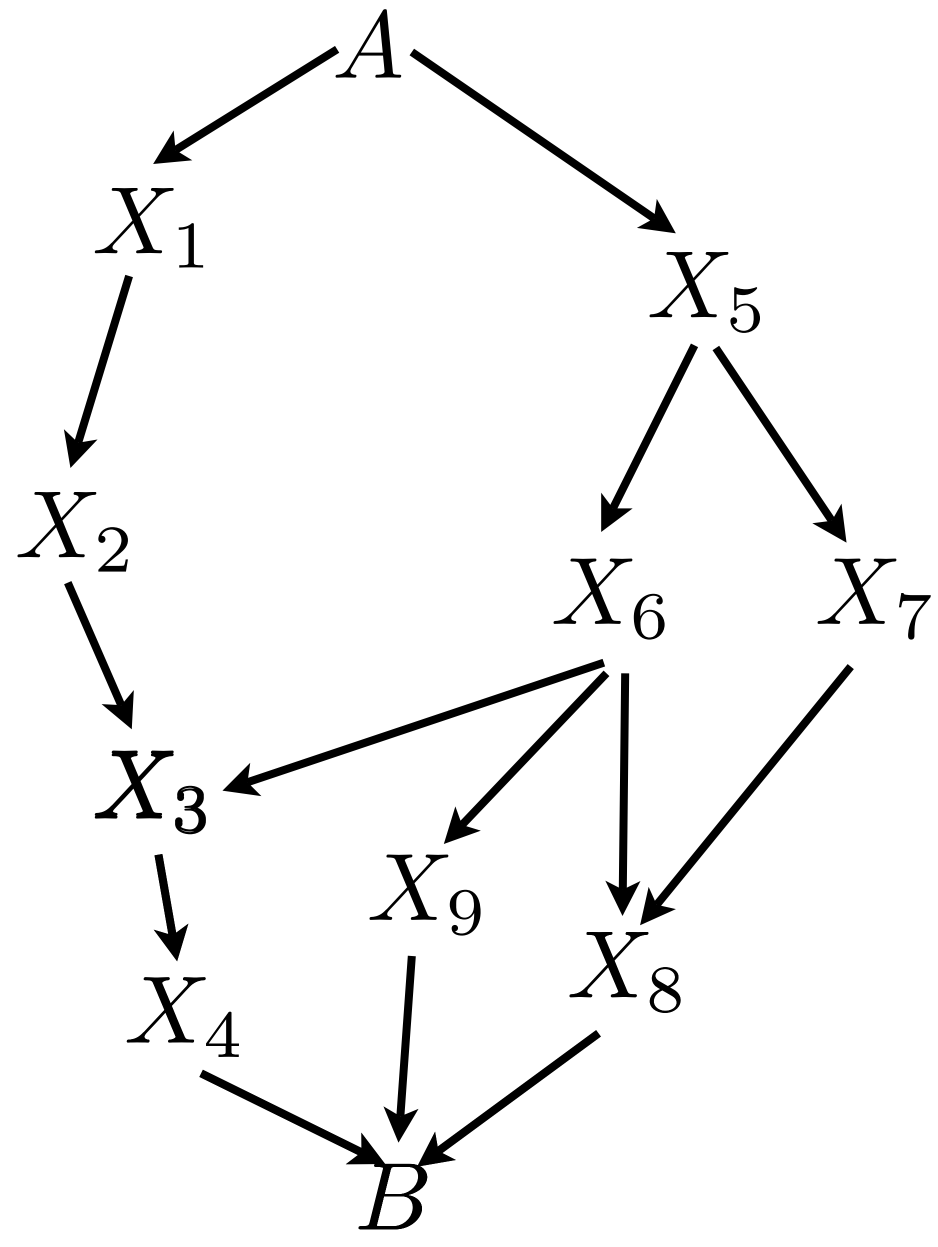
Low backproposal probabilities


Pascal = N
Crack
MOMENT OF IMPACT
NOT TO SCALE
LAWYER FRAME FOR CORRECT MASS DIST.
56" SHOULDER
SANDBAG ON EVERY RUNG
YTH STERS

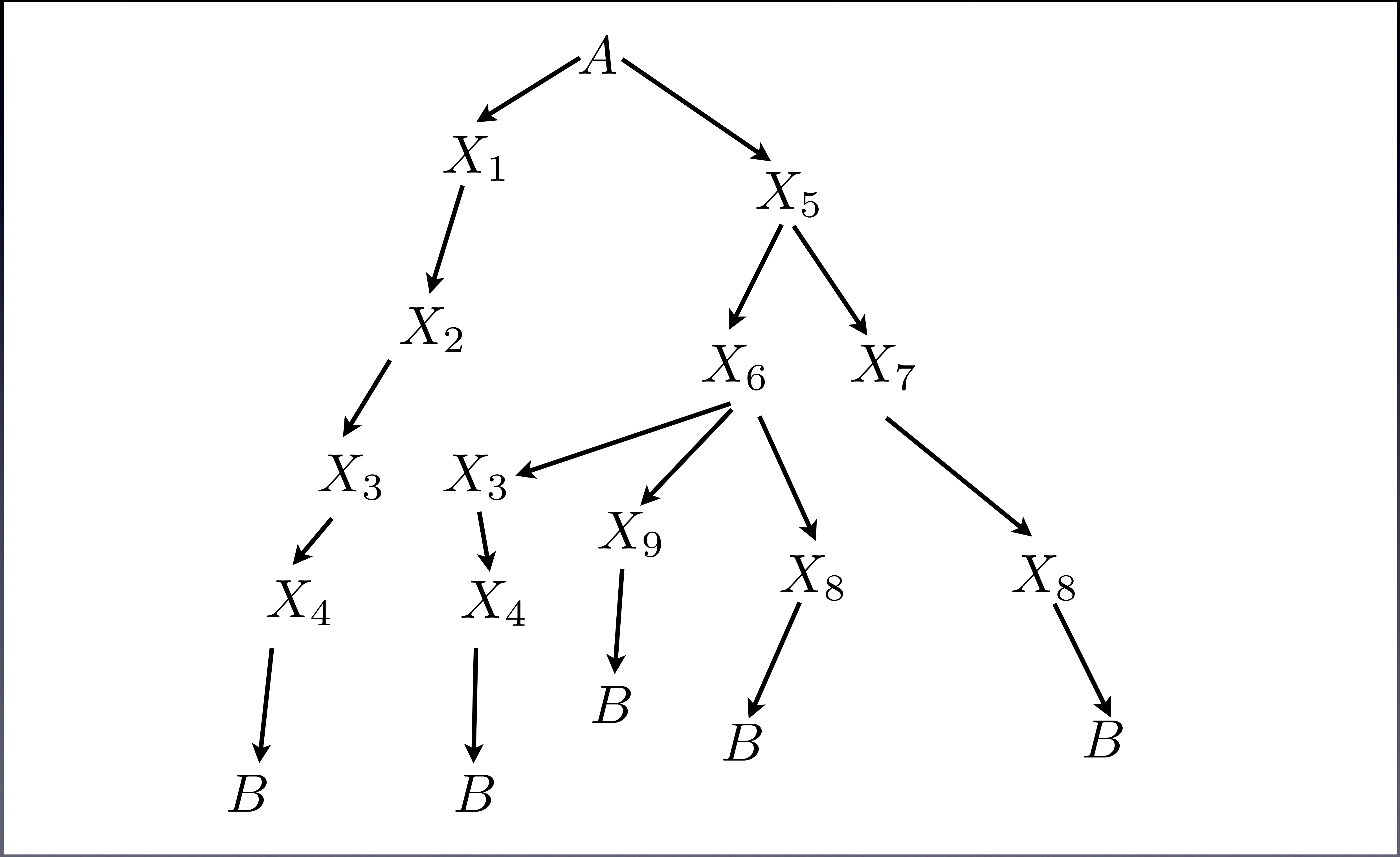
# Myth I.

# It is worth jumping big

# Random walks as decision trees



# Random walks as decision trees



# Myth I.

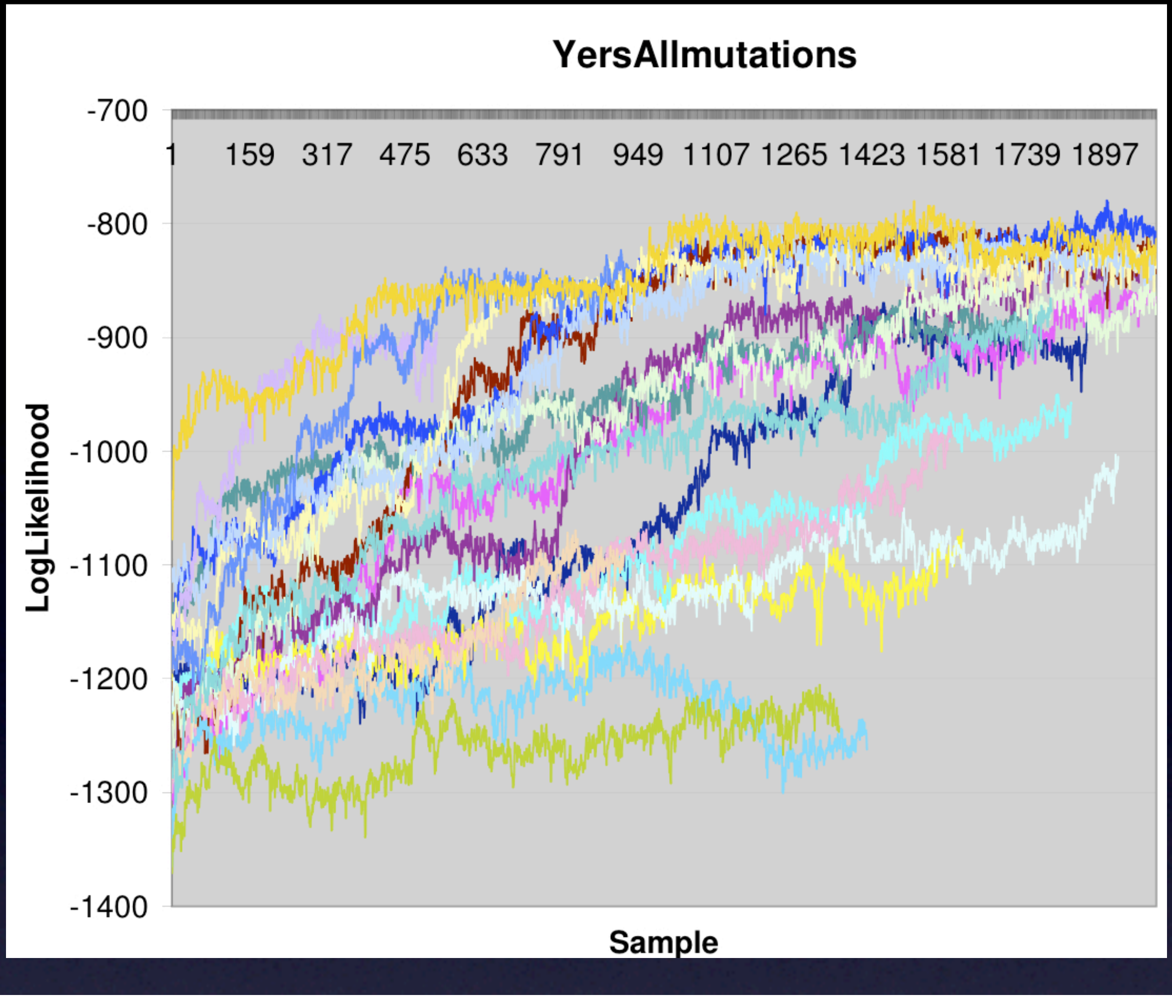
# It is worth jumping big

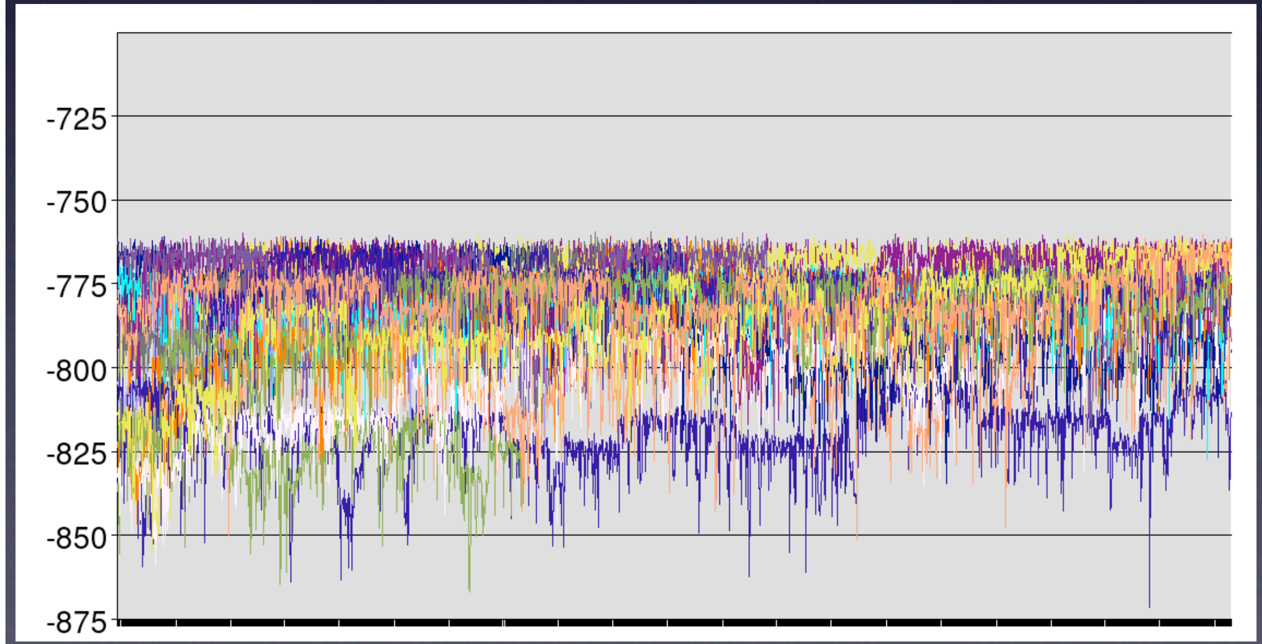
# Myth I.
# It is worth jumping big

We cannot jump big in discrete, high dimensional distributions.

# Myth II.

# Small acceptance rate is caused by bad proposals





Pre-burnin without proposal ratio!

# Myth II.
# Small acceptance rate is caused by bad proposals

# Myth II.
# Small acceptance rate is caused by bad proposals

The ‘curse of small backproposal probabilities’.

# ParIS: indeed speeded up computation?

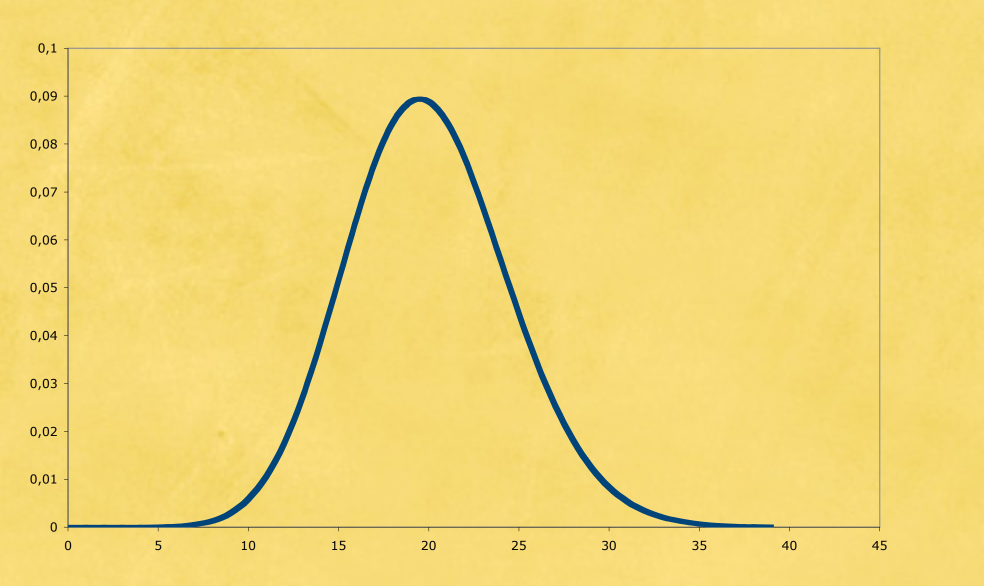
It is enough to calculate proposal probabilities for a 'window' (Green, 1995; Lunter, Miklos, etc, 2005)

$$\pi(X)T(Y \mid X) =$$

$$\pi(X)\sum_i P(Y, w_i \mid X)\min\left(1, \frac{\pi(Y)P(X, w_i \mid Y)}{\pi(X)P(Y, w_i \mid X)}\right) =$$

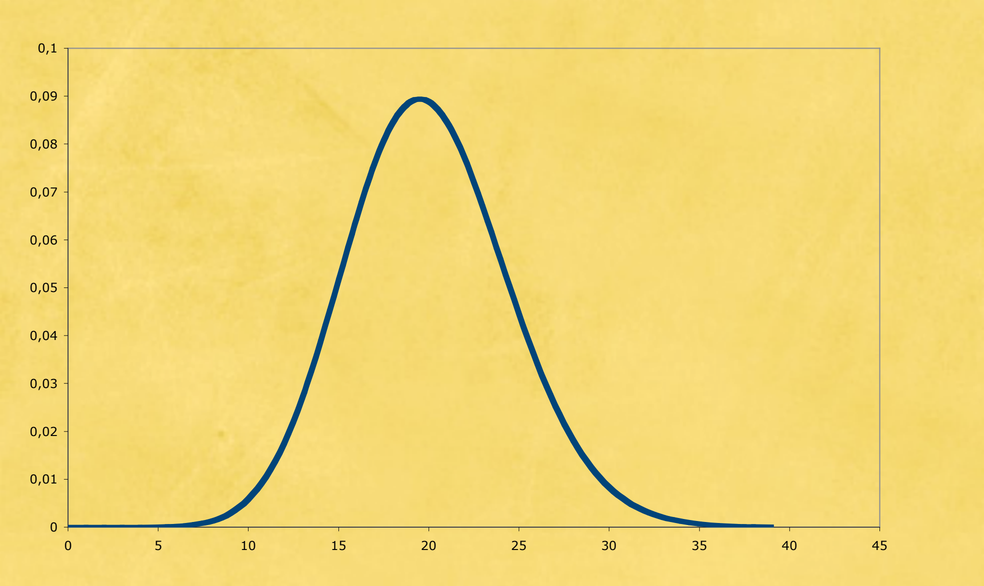
$$\sum_i \min\big(\pi(X)P(Y, w_i \mid X), \pi(Y)P(X, w_i \mid Y)\big)$$

# But does it help?



Proposing alignment window lengths from Poisson distribution

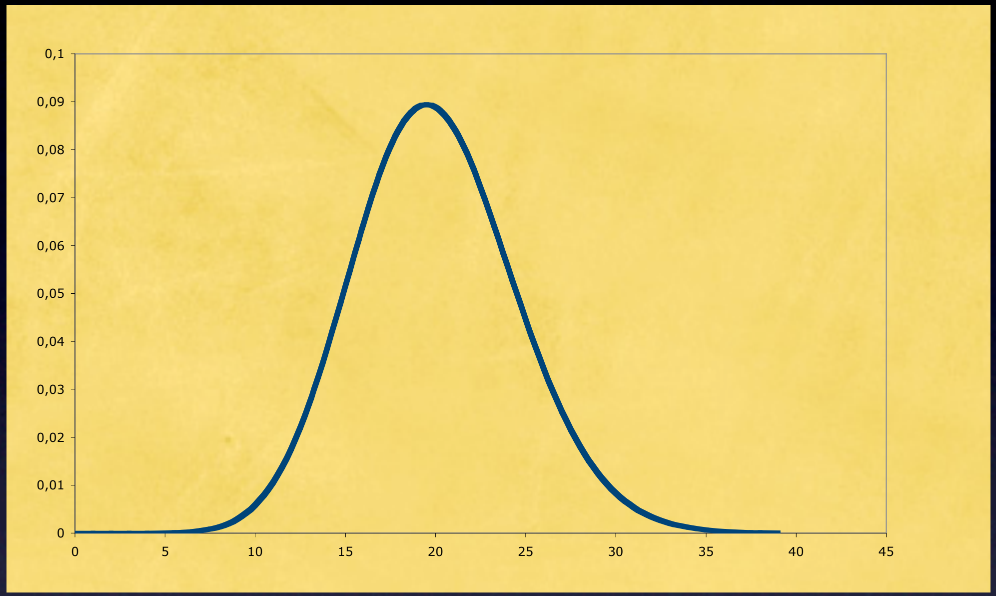
# But does it help?



Proposing alignment window lengths from Poisson distribution

It causes low backproposal probabilities!

# But does it help?

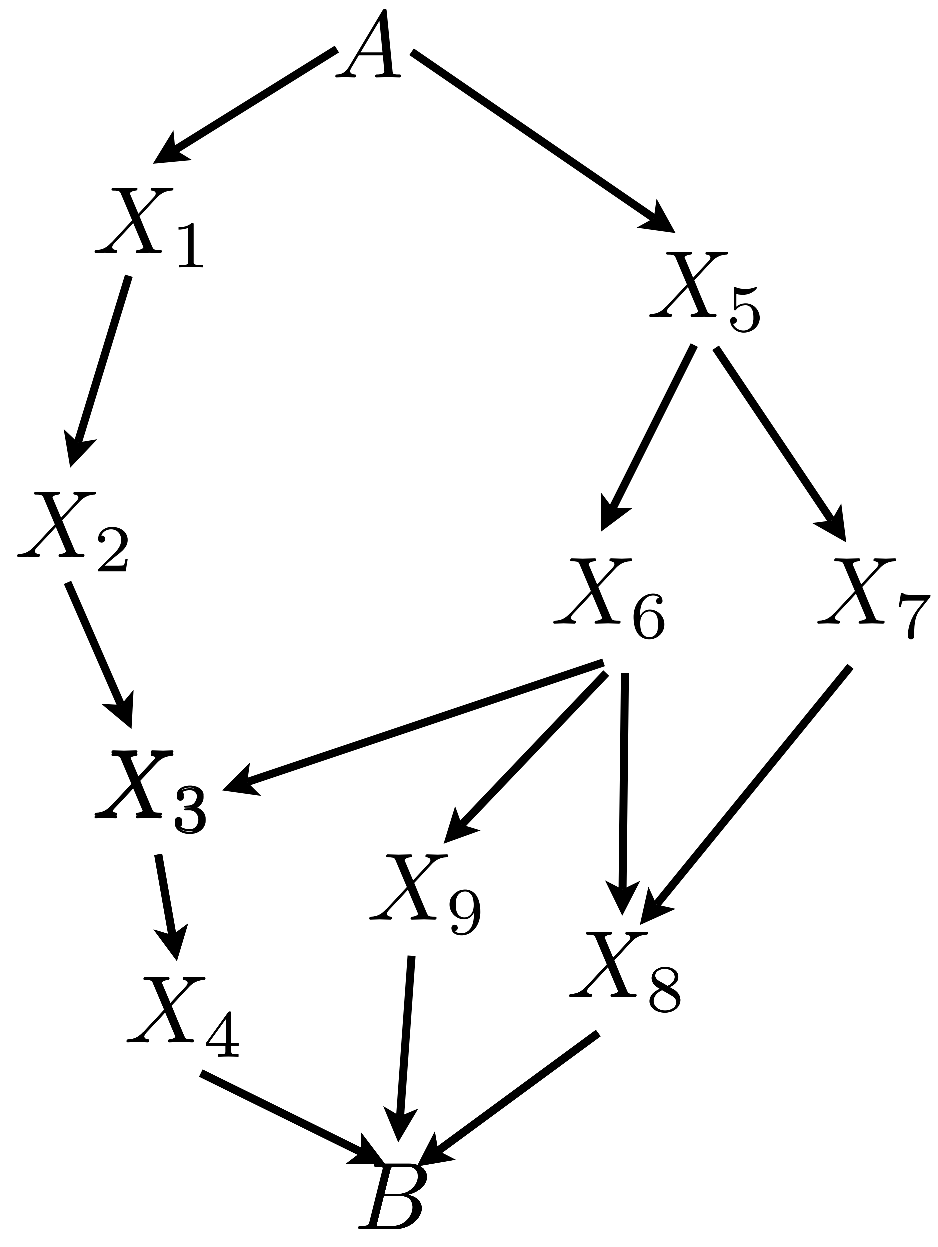


Proposing alignment window lengths from Poisson distribution

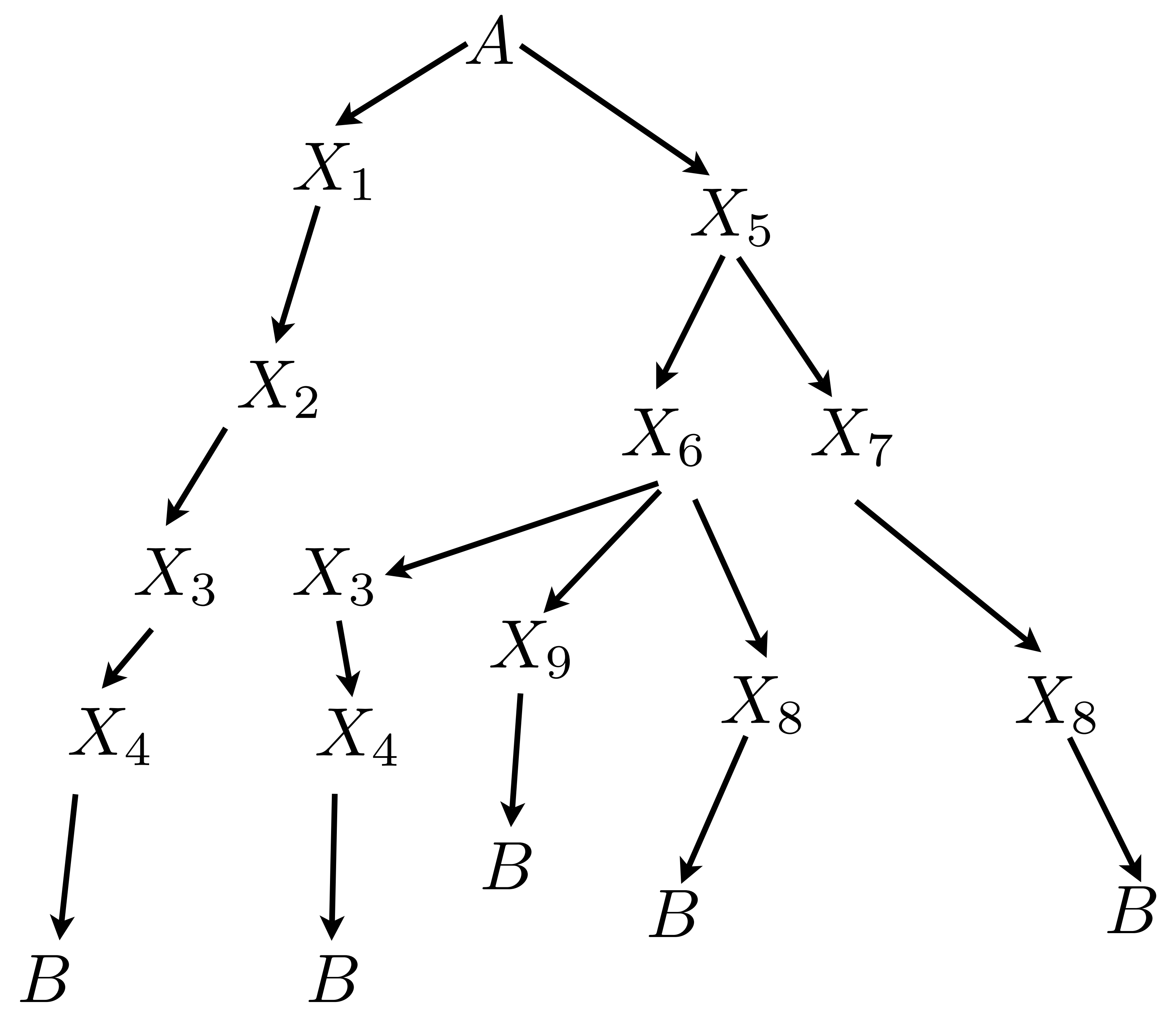
It causes low backproposal probabilities!

A mixture distribution helps

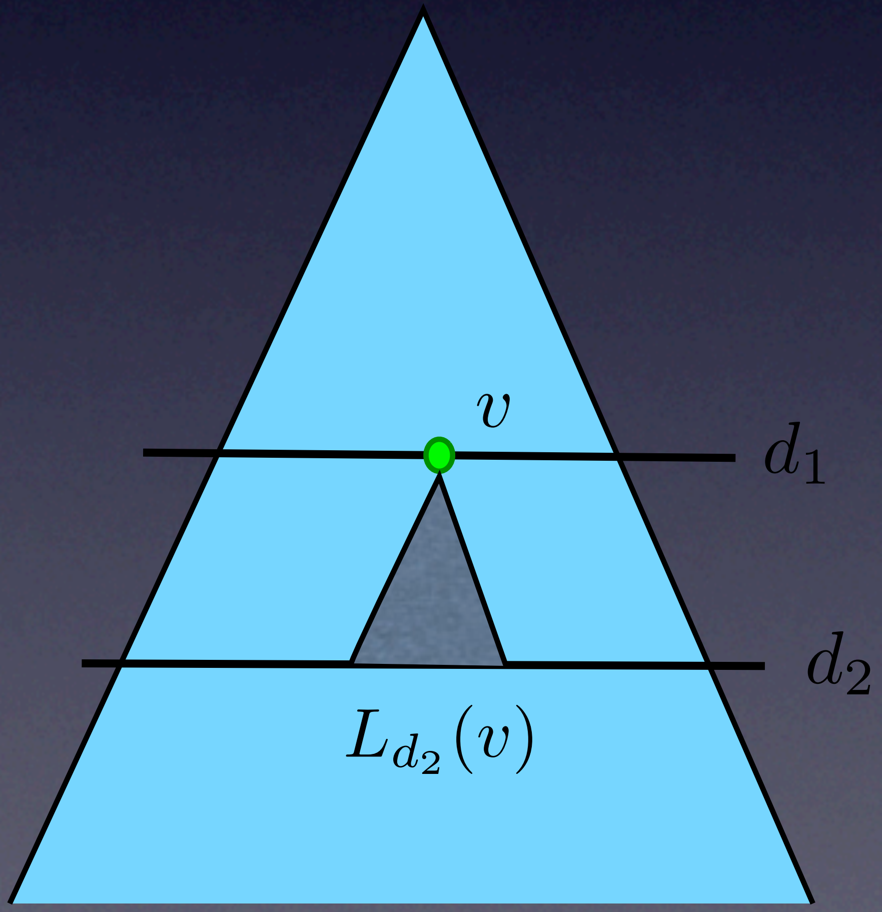
# Random walks as decision trees



# Random walks as decision trees



# Hiddenness of decision trees

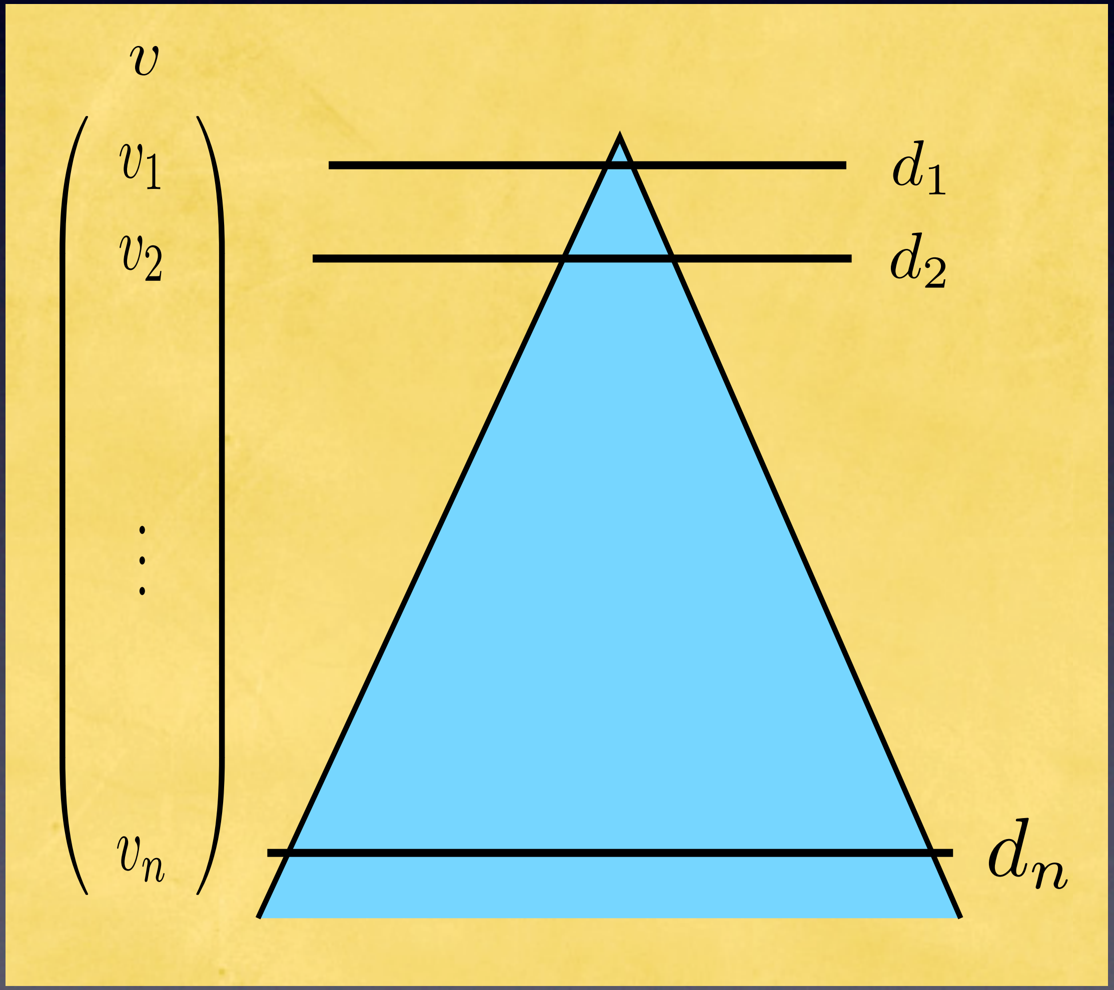


$$|L_{d_2}(v)| \leq k \times \frac{|d_2|}{|d_1|}$$

# MCMCMCMC

Model Changing Metropolis Coupled

Markov Chain Monte Carlo

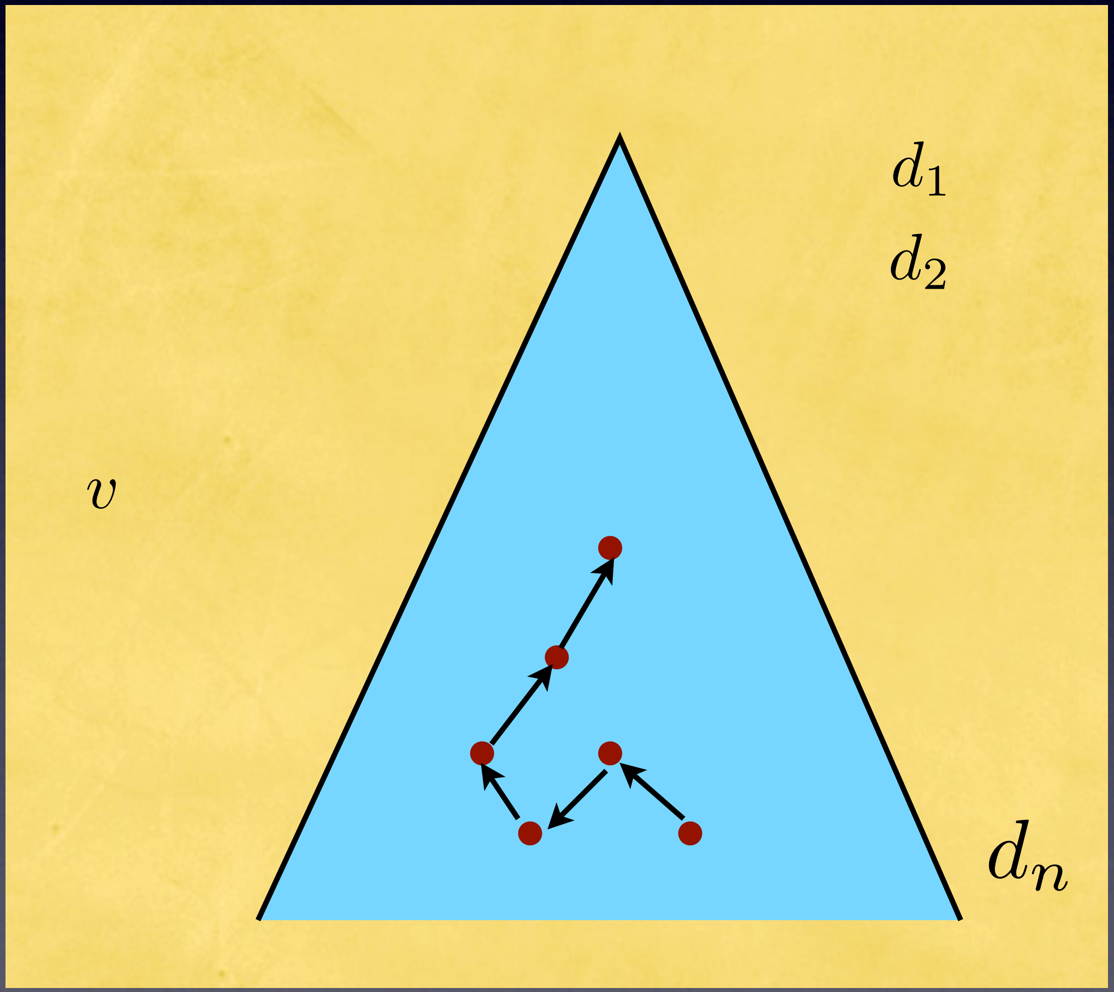


PT-like MSIS

- No need to know the normalizing constant
- Parallelization is possible
- Too many chains?

# MCMCMCMC

Model Changing Metropolis Coupled
Markov Chain Monte Carlo



ST-like MSIS

- Unknown normalizing constant!
- Parallelization is not possible
- Single chain

# Convergence property of MCMCMCMC

For ST-like MSIS

$$\Omega(k) = \tau_{rel} = O(an^2k)$$

For PT-like MSIS

$$\Omega(k) = \tau_{rel} = O(an^{6+\log_2(k)})$$

Where k is the hiddenness rate, ‘a’ is the smallest transition probability, and n is the depth of the tree

# How hidden might the decision tree for reversal sorting paths be?

# How hidden might the decision tree for reversal sorting paths be?

Hiddenness rate might grow exponentially with the length of the signed permutations!

# How hidden might the decision tree for reversal sorting paths be?

Hiddenness rate might grow exponentially with the length of the signed permutations!

Fortunately, we can add more transition kernels, like swapping commuting mutations.

Mixing speed is unknown at the moment...

# Proving that there’s no hope

# Proving that there's no hope

BPP: Bounded-error Probabilistic Polynomial time

It is believed that BPP=P and P$\neq$NP

# Proving that there's no hope

BPP: Bounded-error Probabilistic Polynomial time

It is believed that BPP=P and P $\neq$ NP

When will an FPRAS be a BPP?

# Proving that there's no hope

BPP: Bounded-error Probabilistic Polynomial time

It is believed that BPP=P and P $\neq$ NP

When will an FPRAS be a BPP?

If the exact answer provides solution for a decision problem and it is a polynomial size of the input

# Proving that there's no hope

BPP: Bounded-error Probabilistic Polynomial time

It is believed that BPP=P and P $\neq$ NP

When will an FPRAS be a BPP?

If the exact answer provides solution for a decision problem and it is a polynomial size of the input

If the decision problem is NP-complete, then it is unlikely that an FPRAS algorithm exists, unless BPP $\neq$ P or P=NP

# Further questions - further hope?

# Further questions - further hope?

Is it a problem that we have a hidden part if it has a negligible probability?

# Further questions - further hope?

Is it a problem that we have a hidden part if it has a negligible probability?

The Markov chain gets stuck only if we go there -- which is EXTREMELY unlikely...

# Further questions - further hope?

Is it a problem that we have a hidden part if it has a negligible probability?

The Markov chain gets stuck only if we go there -- which is EXTREMELY unlikely...

What really matters is the variational distance

# Further questions - further hope?

Is it a problem that we have a hidden part if it has a negligible probability?

The Markov chain gets stuck only if we go there -- which is EXTREMELY unlikely...

What really matters is the variational distance

$$d_V(\delta_i^T P^n, \pi^T) \leq c(P, \pi)\rho^n$$

# Further questions - further hope?

Is it a problem that we have a hidden part if it has a negligible probability?

The Markov chain gets stuck only if we go there -- which is EXTREMELY unlikely...

What really matters is the variational distance

$$d_V(\delta_i^T P^n, \pi^T) \leq c(P, \pi)\rho^n$$

$$c(P, \pi) = \min\left\{ \frac{1}{\rho}\left(\frac{p_{ii}}{\pi_i}\right)^{\frac{1}{2}}, \frac{1}{2}\left(\frac{1-\pi_i}{\pi_i}\right)^{\frac{1}{2}} \right\}$$

# Further questions - further hope?

# Further questions - further hope?

Would be we satisfied with a slightly worse algorithmic complexity?

# Further questions - further hope?

Would be we satisfied with a slightly worse algorithmic complexity?

For example, stochastic running time that has a polynomial expectation...

# Further questions - further hope?

Would be we satisfied with a slightly worse algorithmic complexity?

For example, stochastic running time that has a polynomial expectation...

CFTP algorithms (Coupling From the Past) do that!

# Further questions - further hope?

Would be we satisfied with a slightly worse algorithmic complexity?

For example, stochastic running time that has a polynomial expectation...

CFTP algorithms (Coupling From the Past) do that!

How is this complexity class related to FPRAS?

# Conclusions

# Conclusions

I hope I spread non-knowledge significantly today!

# Thank you!