

# Coupling from the past

Mark Rowland

March 16, 2016



# Coupling from the past

**An **exact** algorithm, based on Markov chains, to sample from any distribution  $\pi$  on a finite state space  $S$ .**

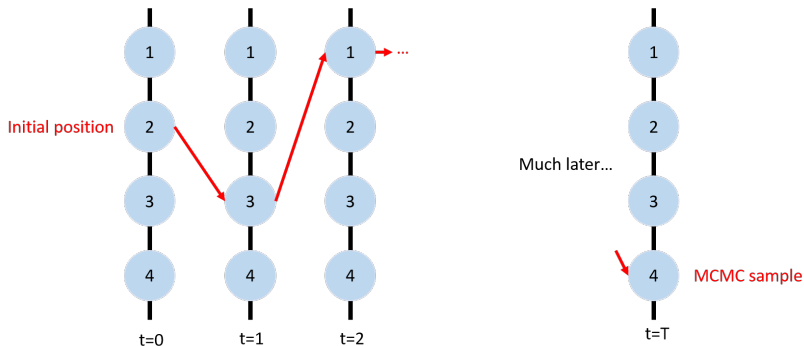
Aims of talk:

- ▶ Describe intuition behind **coupling from the past** as a sampling algorithm
- ▶ Introduce the notion of **coupling** as a proof technique

# Markov chain Monte Carlo

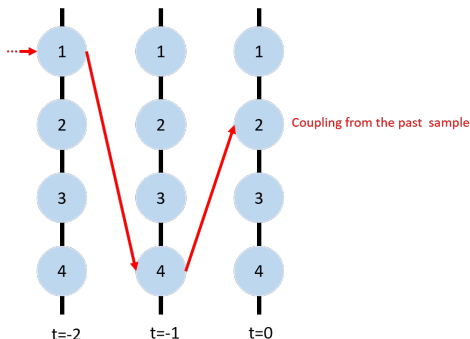
To sample from  $\pi$ , pick Markov chain transition matrix  $P$  with invariant distribution  $\pi$ , initialise the chain somehow at  $t = 0$ , and check the state of the chain at time  $T \gg 0$

$$\mathbb{P}(X_T = j | X_0 = i) \rightarrow \pi_j \quad \text{as } T \rightarrow \infty$$



## Coupling from the past [Propp & Wilson, 1996]

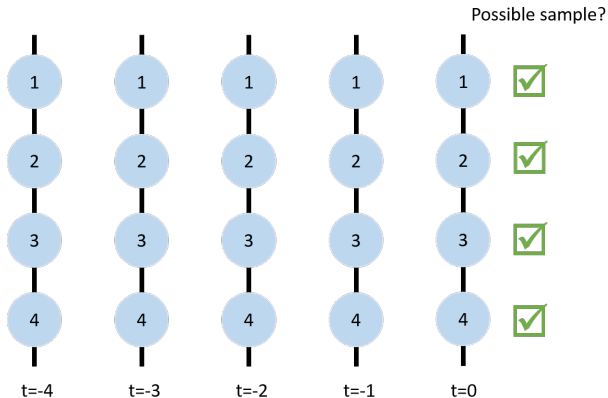
**(Slightly flawed) intuition:** Start the Markov chain at  $t = -\infty$ , then at  $t = 0$ , it would be perfectly mixed, and  $X_0 \sim \pi$ .



**Great insight of Propp & Wilson:** we can **work backwards** from  $t = 0$  to sample from  $X_0$ .

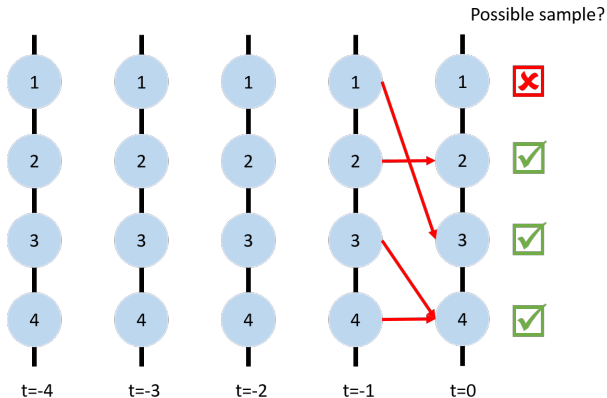
# Coupling from the past

Graphical representation of algorithm:



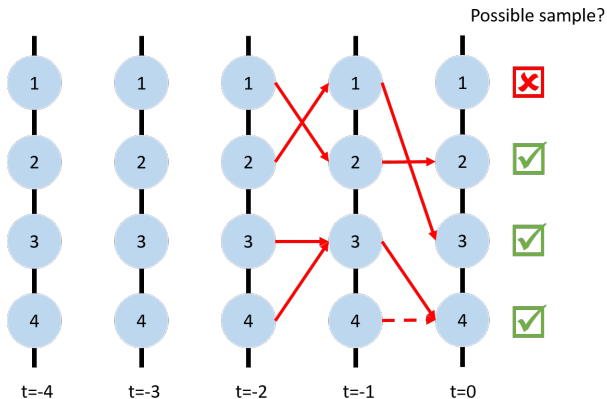
# Coupling from the past

Graphical representation of algorithm:



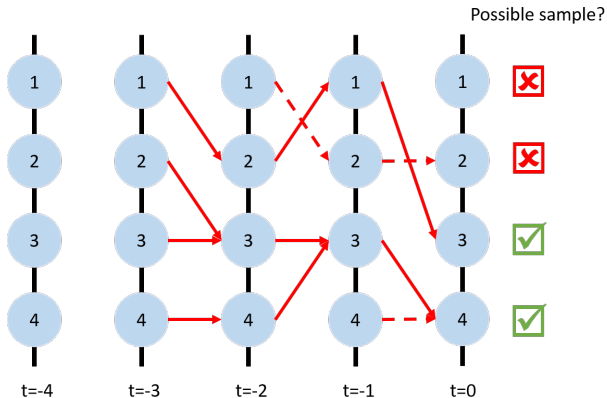
# Coupling from the past

Graphical representation of algorithm:



# Coupling from the past

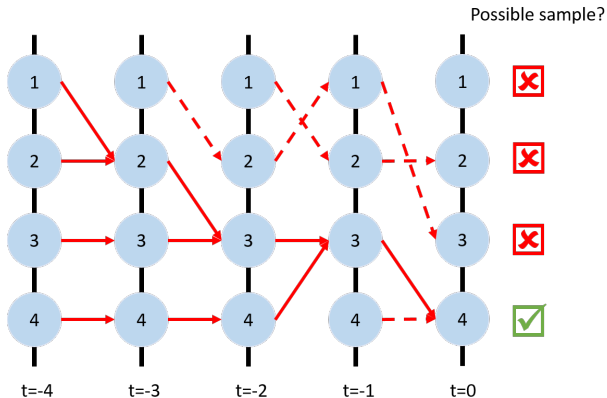
Graphical representation of algorithm:





# Coupling from the past

Graphical representation of algorithm:



# Proof of exactness

We need to prove this procedure really samples from  $\pi$ . This is where *coupling* comes in.

## Theorem

*If the Propp-Wilson algorithm terminates with probability 1, then it samples from  $\pi$ .*

## Nice method of proof [Adapted from Häggström, 2002]:

Let  $T$  be the number of time steps we have to go back for the algorithm to terminate.

Let  $(X_t^{(N)})_{t=-N}^0$  be a Markov chain started in the invariant distribution  $\pi$ , *coupled* to the transitions drawn in the CFTP algorithm.

$$\begin{aligned}\pi_i &= \mathbb{P}(X_0^{(N)} = i) \\ &= \mathbb{P}(X_0^{(N)} = i | T \leq N) \mathbb{P}(T \leq N) + \mathbb{P}(X_0^{(N)} = i | T > N) \mathbb{P}(T > N) \\ &\rightarrow \mathbb{P}(\text{Select state } i)\end{aligned}$$

## Proof of exactness

We need to prove this procedure really samples from  $\pi$ . This is where *coupling* comes in.

### Theorem

*If the Propp-Wilson algorithm terminates with probability 1, then it samples from  $\pi$ .*

**Nice method of proof [Adapted from Häggström, 2002]:**

Let  $T$  be the number of time steps we have to go back for the algorithm to terminate.

Let  $(X_t^{(N)})_{t=-N}^0$  be a Markov chain started in the invariant distribution  $\pi$ , *coupled* to the transitions drawn in the CFTP algorithm.

$$\begin{aligned}\pi_i &= \mathbb{P}(X_0^{(N)} = i) \\ &= \mathbb{P}(X_0^{(N)} = i | T \leq N) \mathbb{P}(T \leq N) + \mathbb{P}(X_0^{(N)} = i | T > N) \mathbb{P}(T > N) \\ &\rightarrow \mathbb{P}(\text{Select state } i)\end{aligned}$$

## Proof of exactness

We need to prove this procedure really samples from  $\pi$ . This is where *coupling* comes in.

### Theorem

*If the Propp-Wilson algorithm terminates with probability 1, then it samples from  $\pi$ .*

### Nice method of proof [Adapted from Häggström, 2002]:

Let  $T$  be the number of time steps we have to go back for the algorithm to terminate.

Let  $(X_t^{(N)})_{t=-N}^0$  be a Markov chain started in the invariant distribution  $\pi$ , *coupled* to the transitions drawn in the CFTP algorithm.

$$\begin{aligned}\pi_i &= \mathbb{P}(X_0^{(N)} = i) \\ &= \mathbb{P}(X_0^{(N)} = i | T \leq N) \mathbb{P}(T \leq N) + \mathbb{P}(X_0^{(N)} = i | T > N) \mathbb{P}(T > N) \\ &\rightarrow \mathbb{P}(\text{Select state } i)\end{aligned}$$

# What I haven't covered

- ▶ Why doesn't this work when composing the transitions forwards in time?
- ▶ What models might this be useful for?
- ▶ Is the algorithm as presented practical?
- ▶ What can be done to reduce memory/computational burden?
- ▶ The ubiquity of coupling in modern probability theory
- ▶ etc.

# Coupling from the past

## References:

*Exact sampling with coupled Markov chains and applications to statistical mechanics* - James Propp & David Wilson

*Finite Markov Chains and Algorithmic Applications* - Olle Häggström

*Markov Chains and Mixing Times* - Levin, Peres & Willmer

Thank you!

## Extra material

### **Exercise:**

Why doesn't this work composing the transitions in the forwards direction?

### **Many extensions:**

Sandwiching trick for coping with large state spaces

Wilson's modification to reduce memory burden



# Coupling of random variables

A **coupling** of random variables  $X$  and  $Y$  is a joint distribution for the random variable  $(X, Y)$ .

Example

$$X, Y \sim N(0, 1)$$

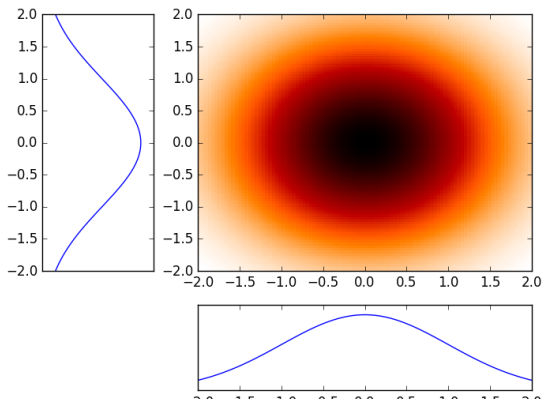
## Coupling of random variables

A **coupling** of random variables  $X$  and  $Y$  is a joint distribution for the random variable  $(X, Y)$ .

### Example

$$X, Y \sim N(0, 1)$$

$$(X, Y) \sim N_2 \left( \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \right)$$



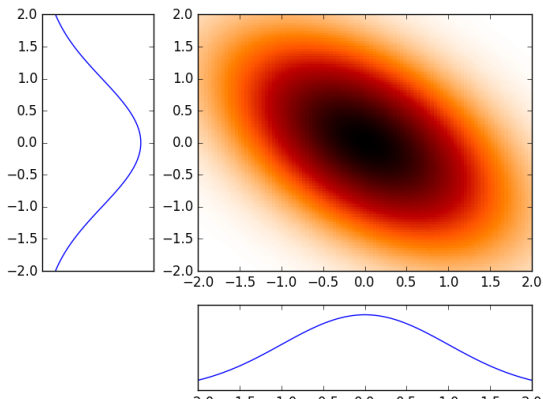
## Coupling of random variables

A **coupling** of random variables  $X$  and  $Y$  is a joint distribution for the random variable  $(X, Y)$ .

### Example

$$X, Y \sim N(0, 1)$$

$$(X, Y) \sim N_2 \left( \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & -0.5 \\ -0.5 & 1 \end{pmatrix} \right)$$



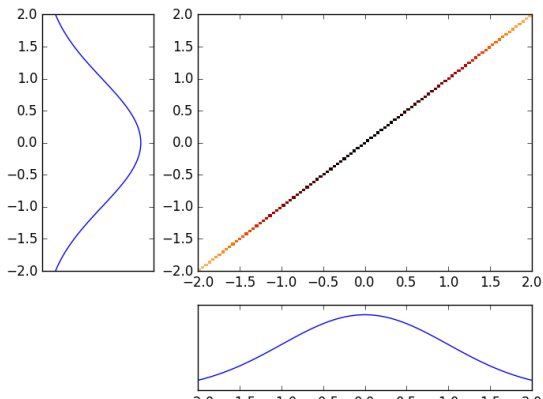
## Coupling of random variables

A **coupling** of random variables  $X$  and  $Y$  is a joint distribution for the random variable  $(X, Y)$ .

### Example

$$X, Y \sim N(0, 1)$$

$$(X, Y) \sim N_2 \left( \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix} \right) \quad (X = Y)$$

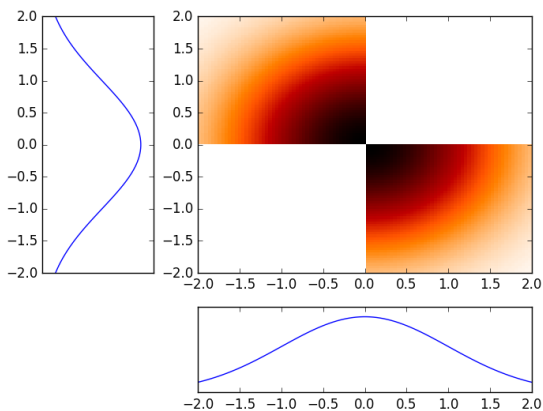


# Coupling of random variables

A **coupling** of random variables  $X$  and  $Y$  is a joint distribution for the random variable  $(X, Y)$ .

## Example

$$X, Y \sim N(0, 1)$$

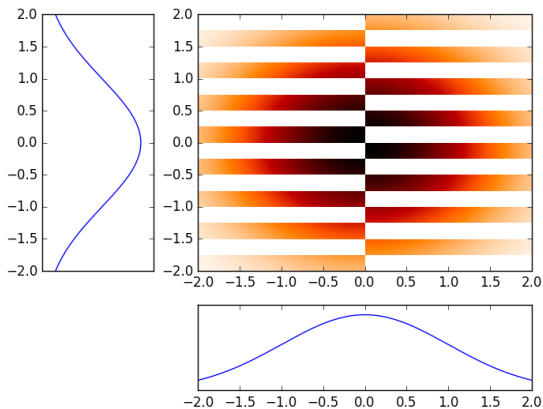


# Coupling of random variables

A **coupling** of random variables  $X$  and  $Y$  is a joint distribution for the random variable  $(X, Y)$ .

## Example

$$X, Y \sim N(0, 1)$$



# Coupling of stochastic processes

$(X_t)_{t \in \mathbb{N}}, (Y_t)_{t \in \mathbb{N}}$  stochastic processes.

Coupling  $\implies$  joint distribution for  $(X_t, Y_t)_{t \in \mathbb{N}}$ .

## Example

$(X_t)_{t \in \mathbb{N}}, (Y_t)_{t \in \mathbb{N}}$  Markov chains on  $\{1, 2, 3, 4\}$  with same transition matrix, but different initial distributions.

# Coupling of stochastic processes

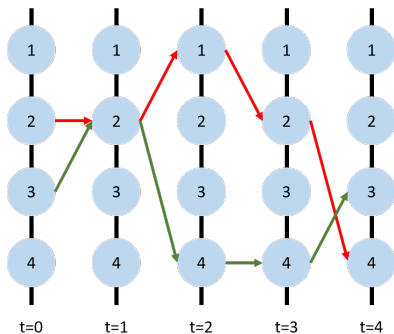
$(X_t)_{t \in \mathbb{N}}$ ,  $(Y_t)_{t \in \mathbb{N}}$  stochastic processes.

Coupling  $\implies$  joint distribution for  $(X_t, Y_t)_{t \in \mathbb{N}}$ .

## Example

$(X_t)_{t \in \mathbb{N}}$ ,  $(Y_t)_{t \in \mathbb{N}}$  Markov chains on  $\{1, 2, 3, 4\}$  with same transition matrix, but different initial distributions.

Example draw from independent coupling:



$$(X_0, X_1, X_2, X_3, X_4) = (2, 2, 1, 2, 4)$$

$$(Y_0, Y_1, Y_2, Y_3, Y_4) = (3, 2, 4, 4, 3)$$



# Coupling of stochastic processes

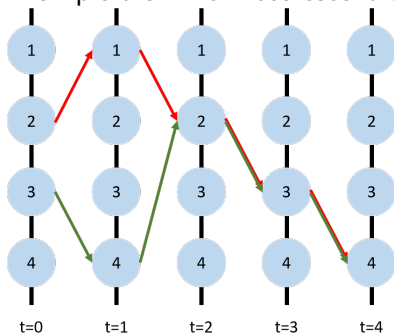
$(X_t)_{t \in \mathbb{N}}$ ,  $(Y_t)_{t \in \mathbb{N}}$  stochastic processes.

Coupling  $\implies$  joint distribution for  $(X_t, Y_t)_{t \in \mathbb{N}}$ .

## Example

$(X_t)_{t \in \mathbb{N}}$ ,  $(Y_t)_{t \in \mathbb{N}}$  Markov chains on  $\{1, 2, 3, 4\}$  with same transition matrix, but different initial distributions.

Example draw from coalescent coupling:



$$(X_0, X_1, X_2, X_3, X_4) = (2, 1, 2, 3, 4)$$

$$(Y_0, Y_1, Y_2, Y_3, Y_4) = (3, 4, 2, 3, 4)$$

# Theory and applications of coupling

- ▶ Used heavily in modern probability theory
- ▶ A way to think about metrics for probability distributions
  - ▶ Total variation distance
  - ▶ Wasserstein distances
- ▶ Underpins *a lot* of Markov chain theory
- ▶ Fundamental in optimal transport theory, useful for:
  - ▶ Meaningfully interpolating between measures
  - ▶ Finding “centres of mass” for collections of probability distributions
  - ▶ Performing PCA-like analysis for collections of probability distributions
- ▶ **Exact sampling algorithms**

# Theory and applications of coupling

- ▶ Used heavily in modern probability theory
- ▶ A way to think about metrics for probability distributions
  - ▶ Total variation distance
  - ▶ Wasserstein distances
- ▶ Underpins *a lot* of Markov chain theory
- ▶ Fundamental in optimal transport theory, useful for:
  - ▶ Meaningfully interpolating between measures
  - ▶ Finding “centres of mass” for collections of probability distributions
  - ▶ Performing PCA-like analysis for collections of probability distributions
- ▶ **Exact sampling algorithms**

# Theory and applications of coupling

- ▶ Used heavily in modern probability theory
- ▶ A way to think about metrics for probability distributions
  - ▶ Total variation distance
  - ▶ Wasserstein distances
- ▶ Underpins *a lot* of Markov chain theory
- ▶ Fundamental in optimal transport theory, useful for:
  - ▶ Meaningfully interpolating between measures
  - ▶ Finding “centres of mass” for collections of probability distributions
  - ▶ Performing PCA-like analysis for collections of probability distributions
- ▶ **Exact sampling algorithms**

# Theory and applications of coupling

- ▶ Used heavily in modern probability theory
- ▶ A way to think about metrics for probability distributions
  - ▶ Total variation distance
  - ▶ Wasserstein distances
- ▶ Underpins *a lot* of Markov chain theory
- ▶ Fundamental in optimal transport theory, useful for:
  - ▶ Meaningfully interpolating between measures
  - ▶ Finding “centres of mass” for collections of probability distributions
  - ▶ Performing PCA-like analysis for collections of probability distributions
- ▶ **Exact sampling algorithms**