Unbiased Markov chain Monte Carlo methods with couplings

Pierre E. Jacob
Department of Statistics, Harvard University
joint work with John O'Leary, Yves Atchadé, and other fantastic collaborators acknowledged throughout

Research Seminar in Mathematical Statistics
Humboldt-Universität zu Berlin
November 4, 2020

Outline

Outline

Setting

Continuous or discrete space, e.g. $\mathbb{R}^d$, or $\{-1, +1\}^d$, or set of subsets of $\{1, \ldots, n\}$, or set of colourings of graphs, etc.

Target probability distribution π.

Goal: approximate π, i.e

$$\text{approximate} \quad \mathbb{E}_\pi[h(X)] = \int h(x)\pi(x)dx = \pi(h),$$

for a class of "test" functions h.

Unbiased estimator: variable with expectation $\mathbb{E}_\pi[h(X)]$.

Integrals in statistics

Integrals arise in most attempts to quantify uncertainty.

- Probability of some event, $\mathbb{P}(X \in A) = \int \mathbb{1}(x \in A)\pi(dx)$.
- In particular, p-values $\mathbb{P}(T > t^{\text{obs}})$.
- Posterior in Bayesian inference $\mathbb{P}(\text{parameter}|\text{data})$.
- Any latent variable leads to an integral in the likelihood.

Often these computations are not feasible analytically and numerical methods are required.

Markov chain Monte Carlo estimators are biased

Initially, $X_0 \sim \pi_0$, then $X_t | X_{t-1} \sim P(X_{t-1}, \cdot)$ for $t = 1, \ldots, T$.

Estimator:

$$\frac{1}{T-b} \sum_{t=b+1}^{T} h(X_t),$$

where b first iterations are discarded as burn-in.

Markov chain Monte Carlo estimators are biased

Initially, $X_0 \sim \pi_0$, then $X_t | X_{t-1} \sim P(X_{t-1}, \cdot)$ for $t = 1, \ldots, T$.

Estimator:

$$\frac{1}{T-b} \sum_{t=b+1}^{T} h(X_t),$$

where b first iterations are discarded as burn-in.

Biased for $\pi(h)$, for any fixed b, T, if $\pi_0 \neq \pi$.
Running "lots of short chains" and averaging can lead to misleading results.

Markov chain Monte Carlo estimators are biased

Initially, $X_0 \sim \pi_0$, then $X_t | X_{t-1} \sim P(X_{t-1}, \cdot)$ for $t = 1, \ldots, T$.

Estimator:

$$\frac{1}{T-b} \sum_{t=b+1}^{T} h(X_t),$$

where b first iterations are discarded as burn-in.

Biased for $\pi(h)$, for any fixed b, T, if $\pi_0 \neq \pi$.
Running "lots of short chains" and averaging can lead to misleading results.

Consistent as $T \to \infty$. How many iterations are enough?

Example: Metropolis–Hastings kernel P

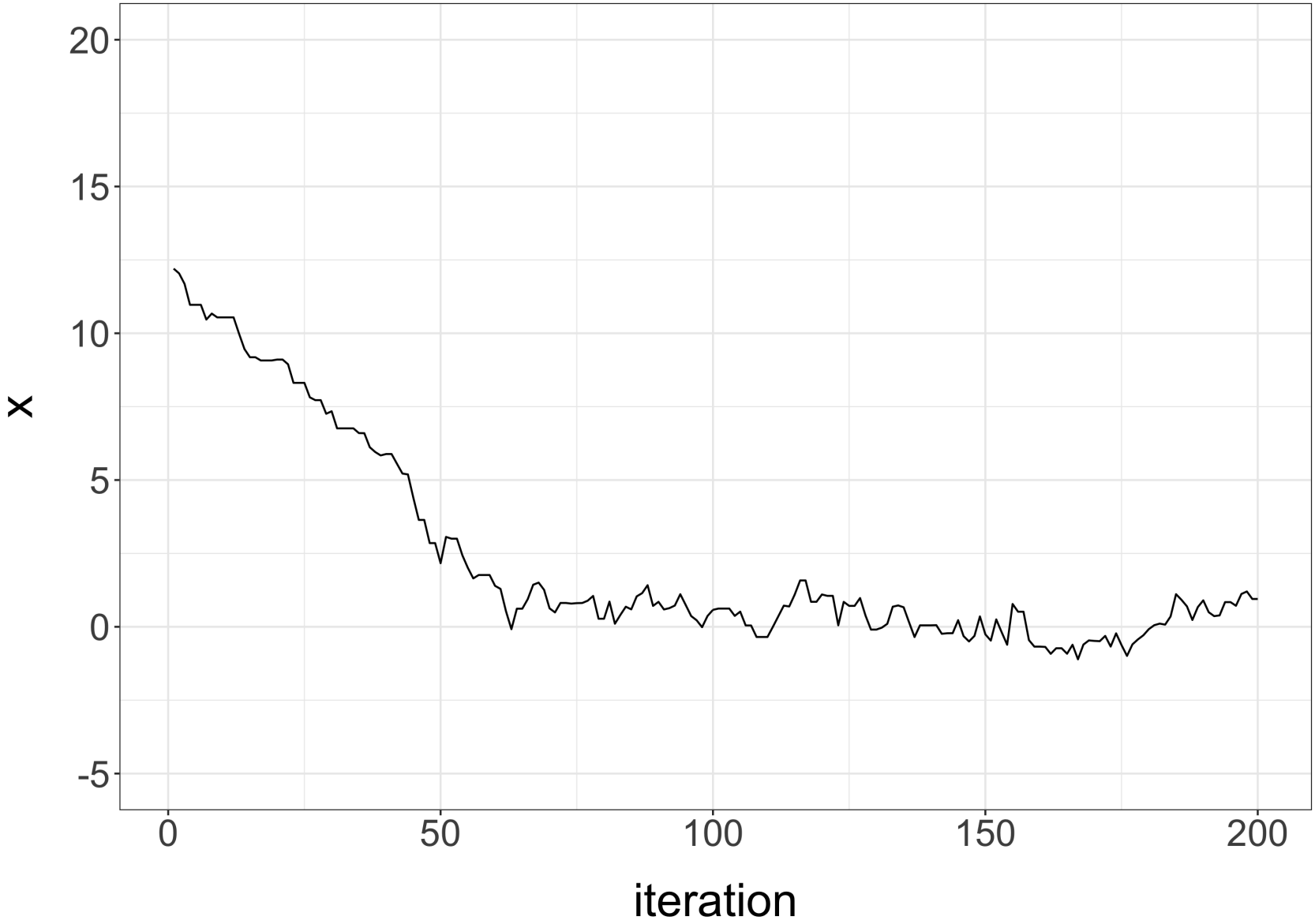
Initialize: $X_0 \sim \pi_0$.

At each iteration $t \geq 1$, with Markov chain at state X_{t-1},

1. propose $X^\star \sim k(X_{t-1}, \cdot)$,
2. sample $U \sim \mathcal{U}(0,1)$,
3. if
$$U \leq \frac{\pi(X^\star)k(X^\star, X_{t-1})}{\pi(X_{t-1})k(X_{t-1}, X^\star)},$$
set $X_t = X^\star$, otherwise set $X_t = X_{t-1}$.

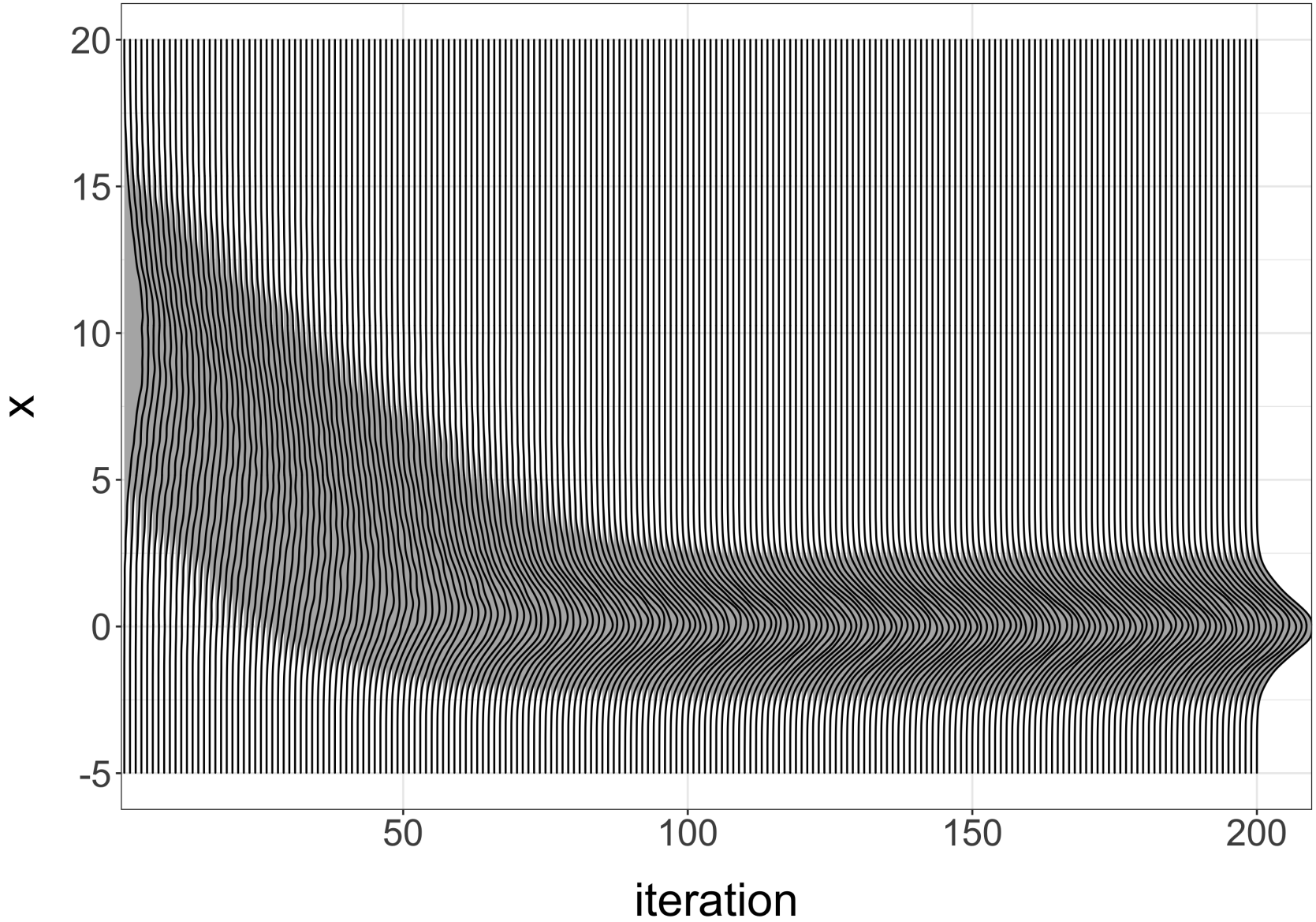
Hastings, *Monte Carlo sampling methods using Markov chains and their applications*, Biometrika, 1970.

Metropolis–Hastings path



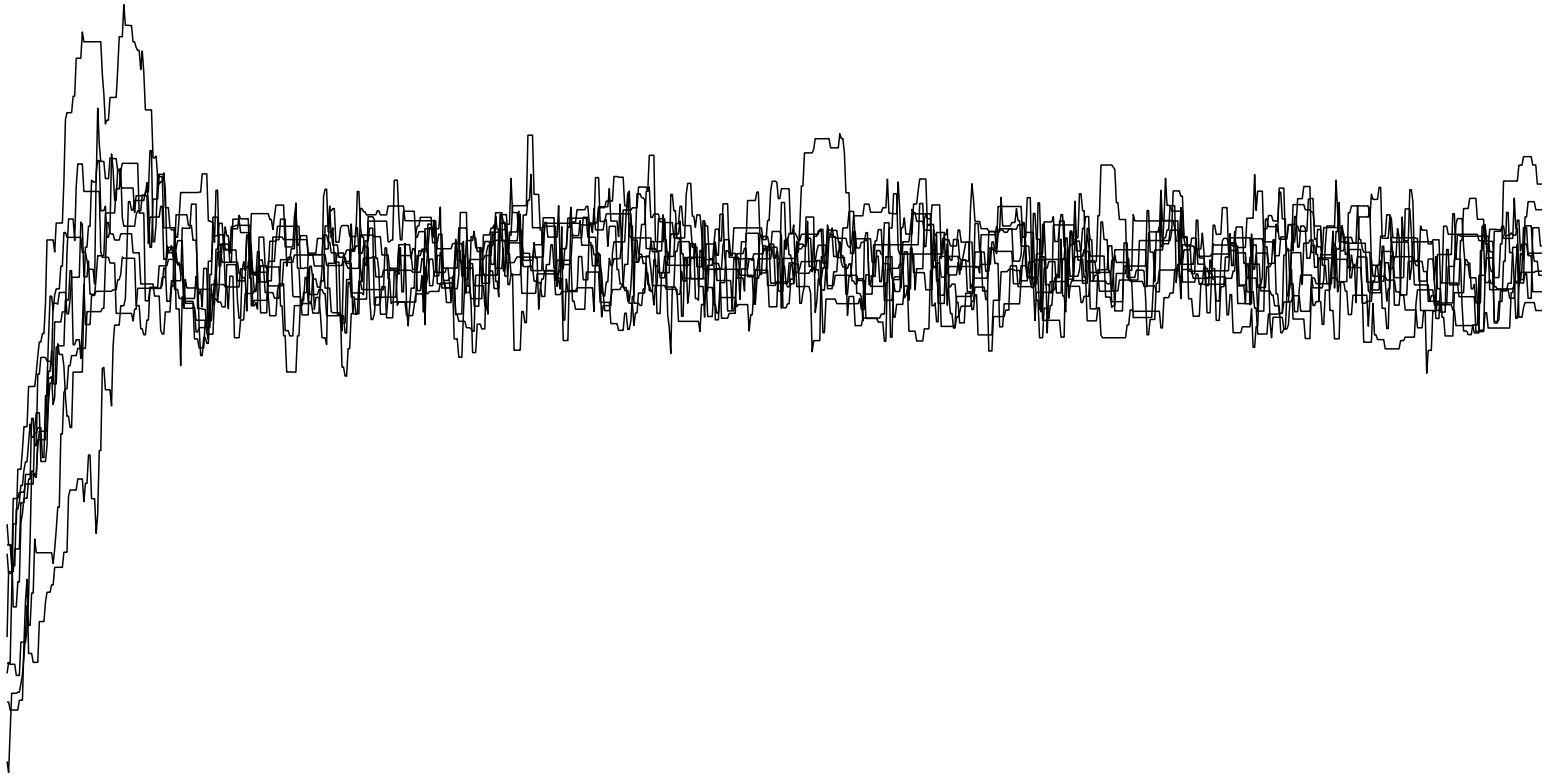
$\pi = \mathcal{N}(0, 1)$, MH with Normal proposal std $= 0.5$, $\pi_0 = \mathcal{N}(10, 3^2)$

Metropolis–Hastings marginal distributions



$\pi_0 = \mathcal{N}(10, 3^2)$, RWMH with Normal proposal std $= 0.5$, $\pi = \mathcal{N}(0, 1)$

Trace plots



. . . such plots can be informative, for real-valued components of the chains.

Trace plots

Consider an MCMC algorithm on subsets $\{k_1, \ldots, k_m\}$ of size m of the set $\{1, \ldots, Q\}$.

Trace plots

Consider an MCMC algorithm on subsets $\{k_1, \ldots, k_m\}$ of size m of the set $\{1, \ldots, Q\}$.

At each iteration, pick one element out of $\{k_1, \ldots, k_m\}$, uniformly at random, and replace it by an element of $\{1, \ldots, Q\} \setminus \{k_1, \ldots, k_m\}$, picked uniformly at random.

Trace plots

Consider an MCMC algorithm on subsets $\{k_1, \ldots, k_m\}$ of size m of the set $\{1, \ldots, Q\}$.

At each iteration, pick one element out of $\{k_1, \ldots, k_m\}$, uniformly at random, and replace it by an element of $\{1, \ldots, Q\} \setminus \{k_1, \ldots, k_m\}$, picked uniformly at random.

Converges to the uniform distribution on m-subsets of $\{1, \ldots, Q\}$... but how long does it take?

Diaconis, Shahshahani, *Time to reach stationarity in the Bernoulli–Laplace diffusion model*, 1987.
Guruswami, *Rapidly Mixing Markov Chains: A Comparison of Techniques*, 2000.

Trace plots

```
{1,2,3} {1,2,4} {1,2,3} {1,3,10} {1,3,4}

{1,3,7} {3,7,8} {3,7,9} {3,5,9} {3,5,7}

{3,5,9} {3,6,9} {5,6,9} {3,5,6} {2,3,6}

{2,6,9} {2,3,6} {3,6,10} {6,8,10} {6,7,8}  . . .
```

. . . has it converged?

Do we care about unbiased estimators?

In classical point estimation, unbiasedness is not crucial.

Larry Wasserman in "All of Statistics" (2003) writes:

> *Unbiasedness used to receive much attention but these days is considered less important.*

On the other hand, Jeff Rosenthal in "Parallel computing and Monte Carlo algorithms" (2000) writes:

> *When running parallel Monte Carlo with many computers, it is more important to start with an unbiased (or low-bias) estimate than with a low-variance estimate.*

Parallel computing with unbiased estimators

Suppose R processors generate, independently, unbiased estimators, one after the other.

Parallel computing with unbiased estimators

Suppose R processors generate, independently, unbiased estimators, one after the other.

There are different ways of aggregating estimators, either discarding on-going calculations at time t or not.

Parallel computing with unbiased estimators

Suppose R processors generate, independently, unbiased estimators, one after the other.

There are different ways of aggregating estimators, either discarding on-going calculations at time t or not.

Different regimes can be considered, e.g.

- as $R \to \infty$ for fixed t,
- as $t \to \infty$ for fixed R,
- as t and R both go to infinity.

Parallel computing with unbiased estimators

Suppose R processors generate, independently, unbiased estimators, one after the other.

There are different ways of aggregating estimators, either discarding on-going calculations at time t or not.

Different regimes can be considered, e.g.

- as $R \to \infty$ for fixed t,
- as $t \to \infty$ for fixed R,
- as t and R both go to infinity.

This has been thoroughly studied, e.g.
Glynn & Heidelberger, *Bias properties of budget constrained simulations*, 1990.
Glynn & Heidelberger, *Analysis of parallel replicated simulations under a completion time constraint*, 1991.

Parallel computing: in details

A machine generates X_k in a time T_k, for $k \geq 1$.

Parallel computing: in details

A machine generates X_k in a time T_k, for $k \geq 1$.

$N(t)$: the number of completed estimators by time t, with $N(t) = 0$ if $t < T_1$.

Parallel computing: in details

A machine generates X_k in a time T_k, for $k \geq 1$.

$N(t)$: the number of completed estimators by time t,
with $N(t) = 0$ if $t < T_1$.

Let $\bar{X}_N = N^{-1} \sum_{k=1}^{N} X_k$ if $N > 0$, and $\bar{X}_N = 0$ if $N = 0$.

Parallel computing: in details

A machine generates X_k in a time T_k, for $k \geq 1$.

$N(t)$: the number of completed estimators by time t,
with $N(t) = 0$ if $t < T_1$.

Let $\bar{X}_N = N^{-1} \sum_{k=1}^{N} X_k$ if $N > 0$, and $\bar{X}_N = 0$ if $N = 0$.

Then

$$\begin{aligned} \mathbb{E}[\bar{X}_{N(t)}] &\neq \mathbb{E}[X], \\ \text{and } \mathbb{E}[\bar{X}_{N(t)+1}] &\neq \mathbb{E}[X], \\ \text{but } \mathbb{E}[\bar{X}_{\max(1,N(t))}] &= \mathbb{E}[X]. \end{aligned}$$

Glynn & Heidelberger, *Bias properties of budget constrained simulations*, 1990 (Theorem 1, Theorem 2, Corollary 7).

Preview

In the proposed approach, each run will involve two coupled chains (X_t) and (Y_t), until some event occurs, which will happen after a random number of iterations denoted by τ.

Preview

In the proposed approach, each run will involve two coupled chains (X_t) and (Y_t), until some event occurs, which will happen after a random number of iterations denoted by τ.

At that point an estimator will be returned, with the guarantee that its expectation is $\mathbb{E}_\pi[h(X)]$ (*unbiasedness*).

Preview

In the proposed approach, each run will involve two coupled chains (X_t) and (Y_t), until some event occurs, which will happen after a random number of iterations denoted by τ.

At that point an estimator will be returned, with the guarantee that its expectation is $\mathbb{E}_\pi[h(X)]$ (*unbiasedness*).

Removing the bias will come at a price in terms of computing cost and variance. The hope is that cost can be reasonable and that variance can be reduced with parallel processors.

Preview

In the proposed approach, each run will involve two coupled chains (X_t) and (Y_t), until some event occurs, which will happen after a random number of iterations denoted by τ.

At that point an estimator will be returned, with the guarantee that its expectation is $\mathbb{E}_\pi[h(X)]$ (*unbiasedness*).

Removing the bias will come at a price in terms of computing cost and variance. The hope is that cost can be reasonable and that variance can be reduced with parallel processors.

The technique provides a concrete way of tackling the question of convergence for MCMC algorithms, provided that adequate coupled chains can be obtained.

Outline

Couplings

Technique to study the convergence of Markov chains.

Construct a joint process (X_t, Y_t) such that $Y_t \sim \pi$ for all $t \geq 0$, (X_t) is the original process of interest started from π_0.

Couplings

Technique to study the convergence of Markov chains.

Construct a joint process (X_t, Y_t) such that $Y_t \sim \pi$ for all $t \geq 0$, (X_t) is the original process of interest started from π_0.

Suppose that there exists τ a random variable such that $X_t = Y_t$ for all $t \geq \tau$, called "meeting time". Then

$$\|\mathcal{L}(X_t) - \mathcal{L}(Y_t)\|_{\mathrm{TV}} \leq \mathbb{P}(\tau \geq t),$$

where $\|\cdot\|_{\mathrm{TV}}$ is the total variation distance. This is called the "coupling inequality".

Couplings

Technique to study the convergence of Markov chains.

Construct a joint process (X_t, Y_t) such that $Y_t \sim \pi$ for all $t \geq 0$, (X_t) is the original process of interest started from π_0.

Suppose that there exists τ a random variable such that $X_t = Y_t$ for all $t \geq \tau$, called "meeting time". Then

$$\|\mathcal{L}(X_t) - \mathcal{L}(Y_t)\|_{\mathrm{TV}} \leq \mathbb{P}(\tau \geq t),$$

where $\|\cdot\|_{\mathrm{TV}}$ is the total variation distance. This is called the "coupling inequality".

Goal is then to make the right-hand side as small as possible, by constructing chains that "meet" as early as possible.

Bru & Yor, *Comments on the life and mathematical legacy of Wolfgang Doeblin*, 2002.

Couplings

Coupling techniques also enable the study of convergence in other metrics than TV.

For example if we can construct coupled processes that "contract" on average,

$$\mathbb{E}[d(X_t, Y_t)] \to 0,$$

then we can bound the 1-Wasserstein distance

$$\|\mathcal{L}(X_t) - \pi\|_{\mathrm{W}} \leq \mathbb{E}[d(X_t, Y_t)].$$

Couplings

Coupling techniques also enable the study of convergence in other metrics than TV.

For example if we can construct coupled processes that "contract" on average,

$$\mathbb{E}[d(X_t, Y_t)] \to 0,$$

then we can bound the 1-Wasserstein distance

$$\|\mathcal{L}(X_t) - \pi\|_{\mathrm{W}} \leq \mathbb{E}[d(X_t, Y_t)].$$

Theoretical construct, in practice we cannot sample from π, so we cannot initiate the (Y_t) process at stationarity.

Couplings

On the theoretical side, coupling techniques have proved very successful, in some cases giving precise rates of convergence.

Jerrum, *Mathematical foundations of the MCMC method*, 1998.
Guruswami, *Rapidly Mixing Markov Chains: A Comparison of Techniques*, 2000.
Roberts & Rosenthal, *General state space Markov chains and MCMC algorithms*, 2004.
Eberle, *Reflection couplings and contraction rates for diffusions*, 2016

Couplings

On the theoretical side, coupling techniques have proved very successful, in some cases giving precise rates of convergence.

Jerrum, *Mathematical foundations of the MCMC method*, 1998.
Guruswami, *Rapidly Mixing Markov Chains: A Comparison of Techniques*, 2000.
Roberts & Rosenthal, *General state space Markov chains and MCMC algorithms*, 2004.
Eberle, *Reflection couplings and contraction rates for diffusions*, 2016

Given X_t, Y_t, can we sample X_{t+1}, Y_{t+1} in a coupled way, such that the chains contract to one another, or even meet?

Couplings

(X, Y) follows a coupling of p and q if $X \sim p$ and $Y \sim q$.

The coupling inequality states that

$$\mathbb{P}(X = Y) \leq 1 - \|p - q\|_{\mathrm{TV}},$$

for any coupling, with $\|p - q\|_{\mathrm{TV}} = \frac{1}{2} \int |p(x) - q(x)| dx$.

Maximal couplings achieve the bound.

Maximal coupling: algorithm

Input: p and q.
Output: pairs (X, Y) from max coupling of p and q.

1. Sample $X \sim p$ and $W \sim \mathcal{U}(0, 1)$.
2. If $W \leq q(X)/p(X)$, set $Y = X$.
3. Otherwise, sample $Y^\star \sim q$ and $W^\star \sim \mathcal{U}(0, 1)$ until $W^\star > p(Y^\star)/q(Y^\star)$, then set $Y = Y^\star$.

e.g. Thorisson, *Coupling, stationarity, and regeneration*, 2000, Chapter 1, Section 4.5.

Examples: couplings of univariate Normals

Consider couplings of $X \sim \mathcal{N}(\mu_x, \sigma^2)$ and $Y \sim \mathcal{N}(\mu_y, \sigma^2)$.

- Independent coupling.

Examples: couplings of univariate Normals

Consider couplings of $X \sim \mathcal{N}(\mu_x, \sigma^2)$ and $Y \sim \mathcal{N}(\mu_y, \sigma^2)$.

- Independent coupling.
- Optimal transport: minimizes $\mathbb{E}[|X - Y|^2]$,
  $X = \mu_x + \varepsilon$, $Y = \mu_y + \varepsilon$, with $\varepsilon \sim \mathcal{N}(0, \sigma^2)$.

Examples: couplings of univariate Normals

Consider couplings of $X \sim \mathcal{N}(\mu_x, \sigma^2)$ and $Y \sim \mathcal{N}(\mu_y, \sigma^2)$.

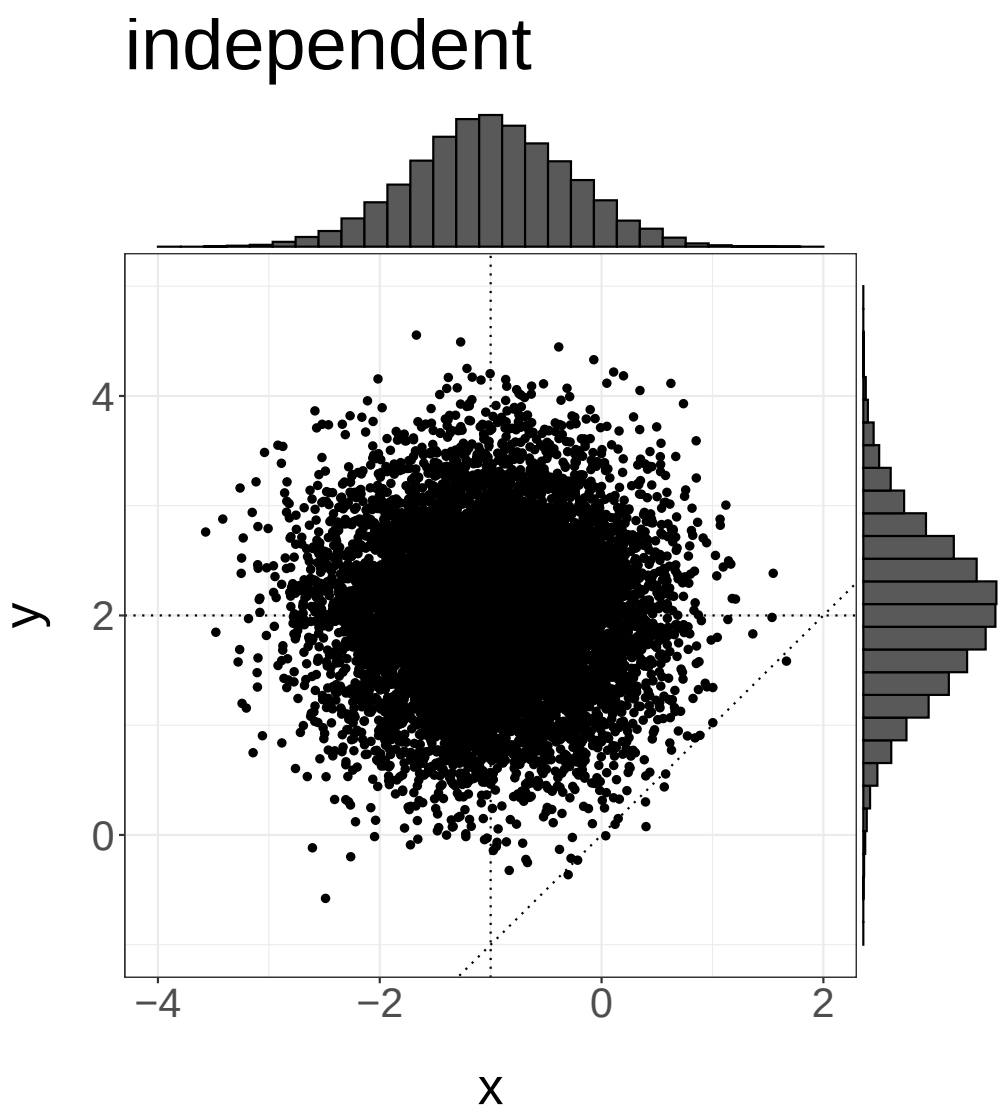
- Independent coupling.
- Optimal transport: minimizes $\mathbb{E}[|X - Y|^2]$,
  $X = \mu_x + \varepsilon$, $Y = \mu_y + \varepsilon$, with $\varepsilon \sim \mathcal{N}(0, \sigma^2)$.
- Reflection: $X = \mu_x + \varepsilon$, $Y = \mu_y - \varepsilon$, with $\varepsilon \sim \mathcal{N}(0, \sigma^2)$.

Examples: couplings of univariate Normals

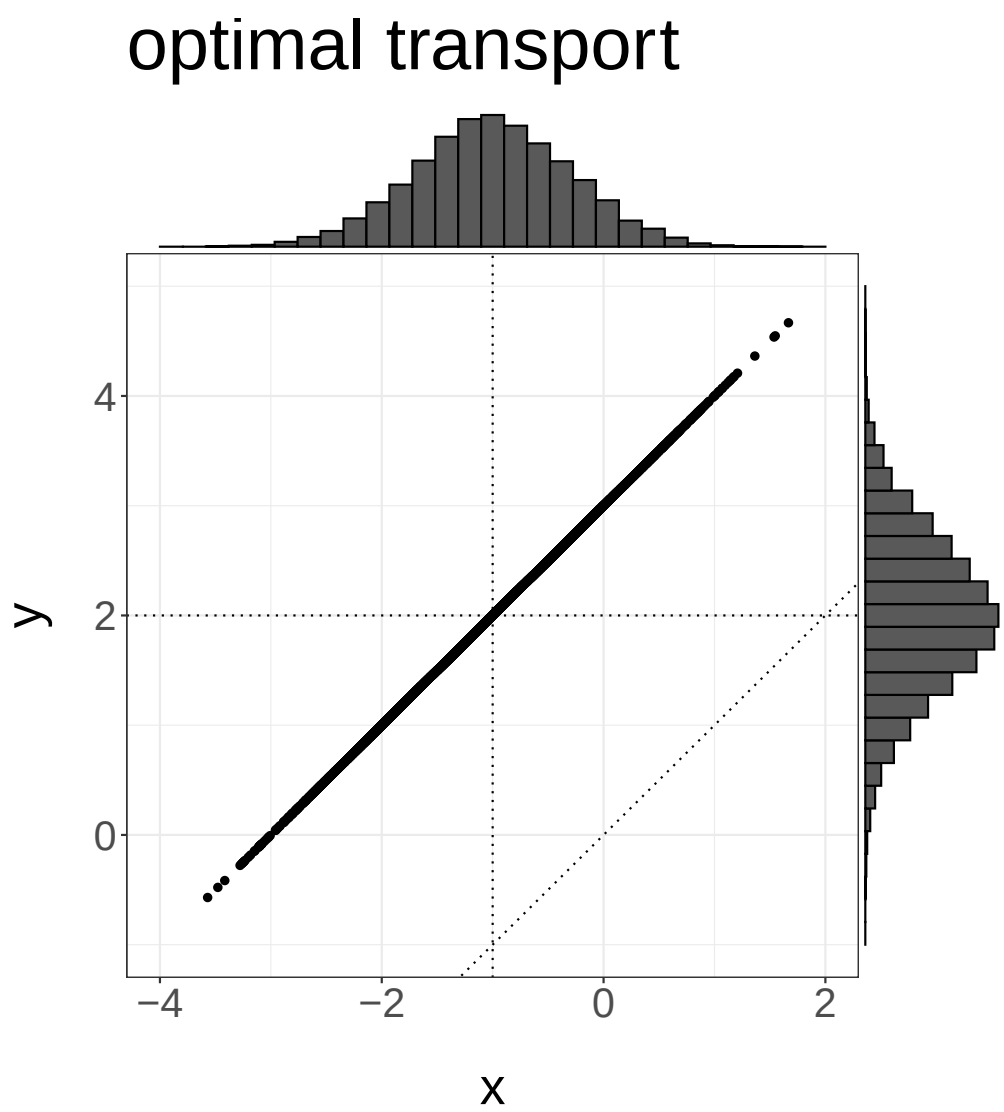
Consider couplings of $X \sim \mathcal{N}(\mu_x, \sigma^2)$ and $Y \sim \mathcal{N}(\mu_y, \sigma^2)$.

- Independent coupling.
- Optimal transport: minimizes $\mathbb{E}[|X - Y|^2]$,
  $X = \mu_x + \varepsilon$, $Y = \mu_y + \varepsilon$, with $\varepsilon \sim \mathcal{N}(0, \sigma^2)$.
- Reflection: $X = \mu_x + \varepsilon$, $Y = \mu_y - \varepsilon$, with $\varepsilon \sim \mathcal{N}(0, \sigma^2)$.
- Maximal: $\mathbb{P}(X = Y)$ is maximized.
  In the event $\{X \neq Y\}$, X and Y could be independent or not, which results in various maximal couplings.

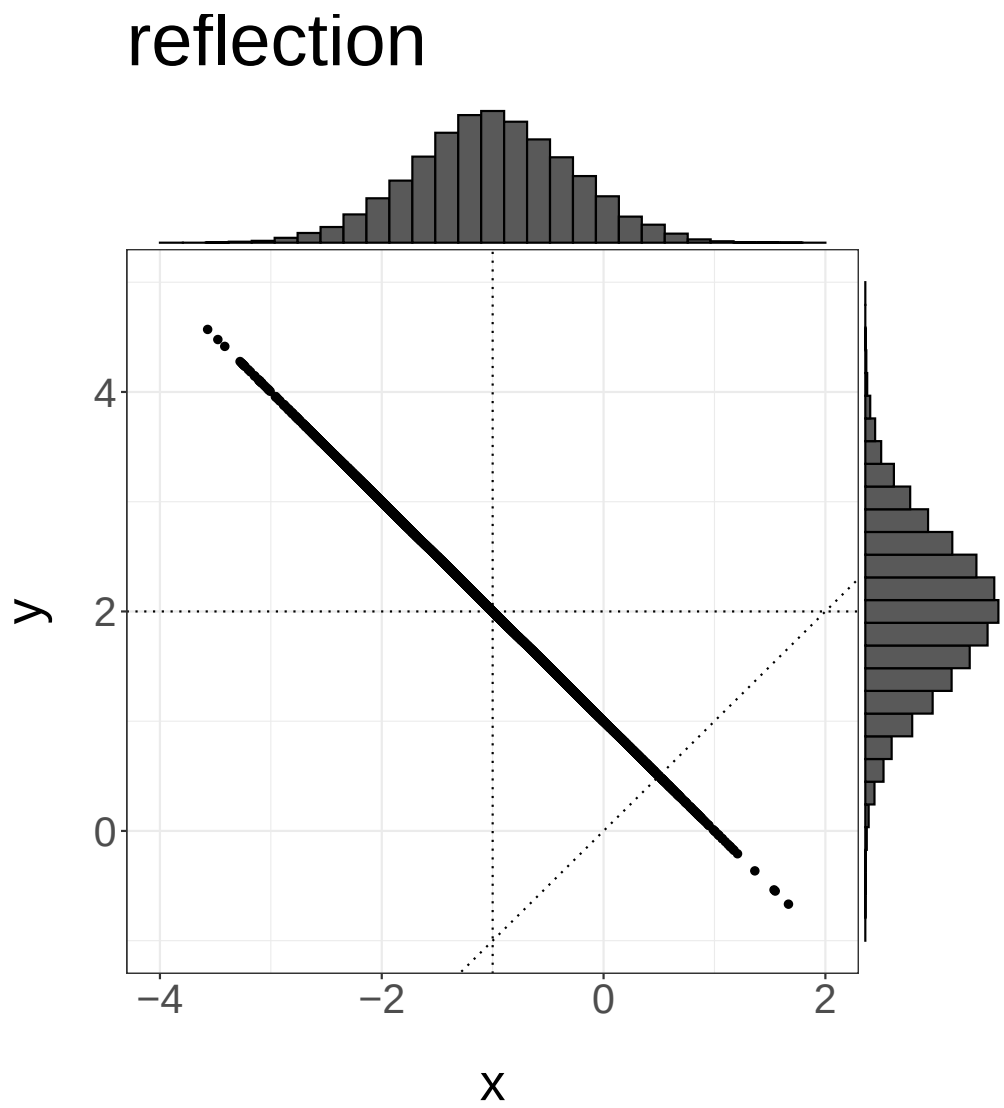
Examples: couplings of univariate Normals



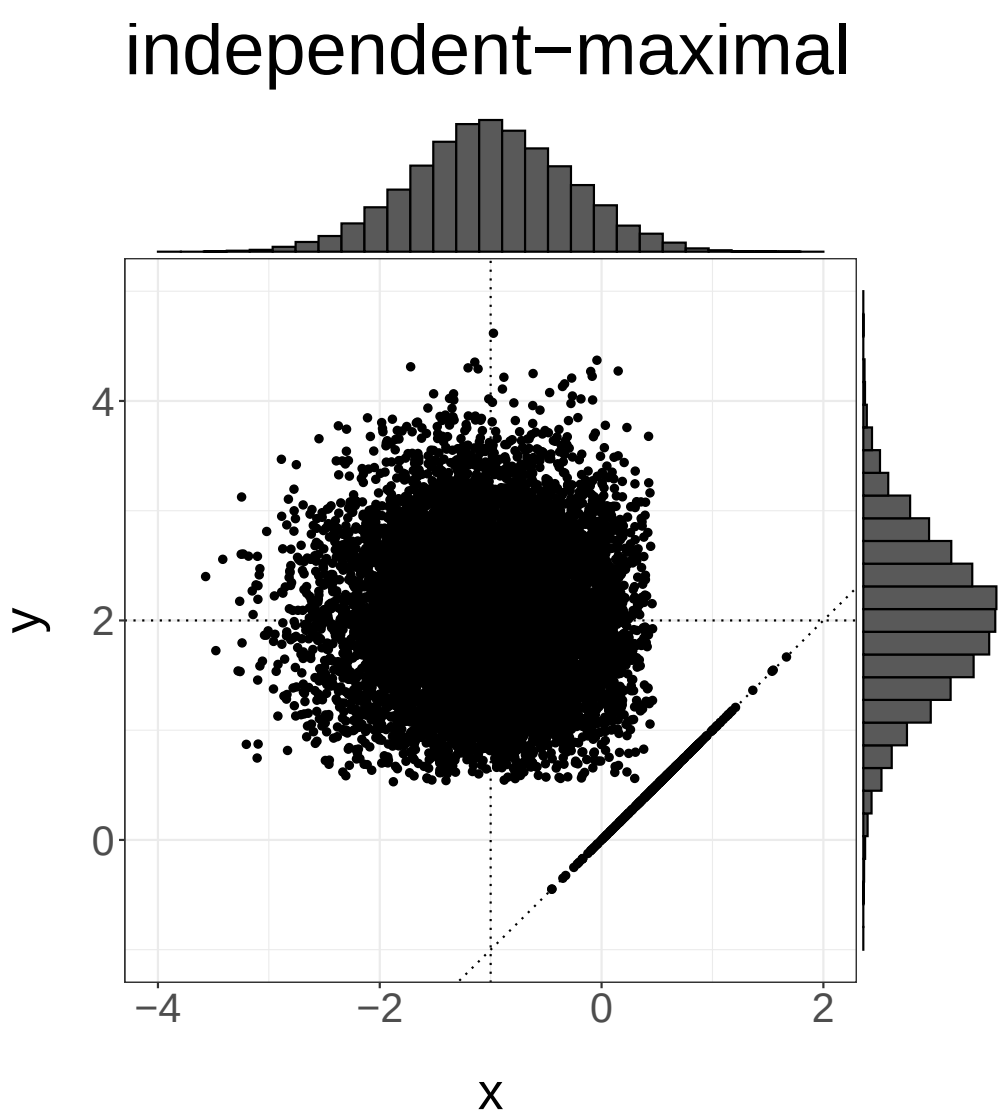
Examples: couplings of univariate Normals



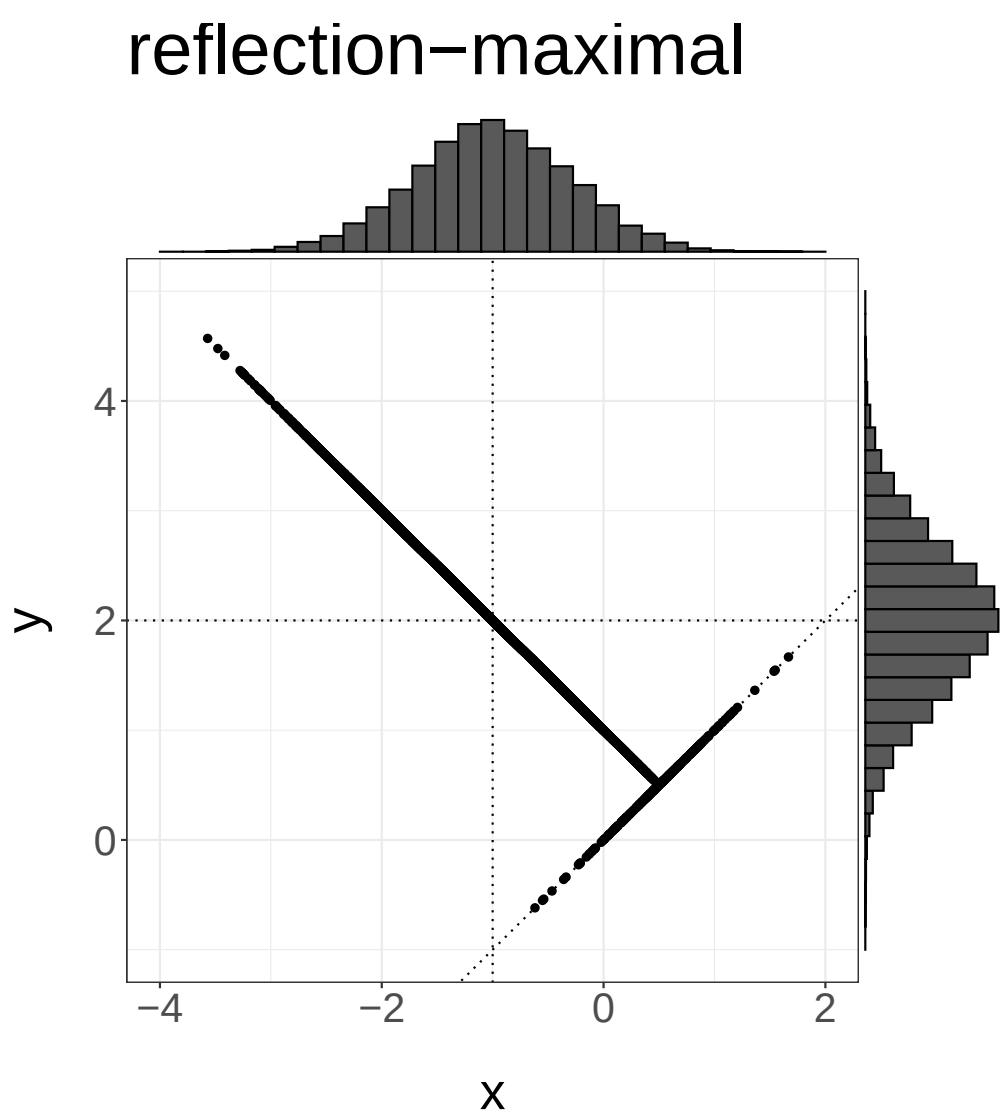
Examples: couplings of univariate Normals



Examples: couplings of univariate Normals



Examples: couplings of univariate Normals



Couplings of MCMC algorithms

Back to the MCMC setting with a Markov kernel P,
we want to sample from a Markov kernel $\bar{P}$,

such that, when (X_t, Y_t) is sampled from $\bar{P}\left((X_{t-1}, Y_{t-1}), \cdot\right)$,

- marginally $X_t | X_{t-1} \sim P(X_{t-1}, \cdot)$, and $Y_t | Y_{t-1} \sim P(Y_{t-1}, \cdot)$,
- there exists some time τ such that $X_t = Y_t$ for all $t \geq \tau$.

Couplings of MCMC algorithms

Many implementable couplings in the literature...

- Propp & Wilson, *Exact sampling with coupled Markov chains and applications to statistical mechanics*, 1996.
- Johnson, *Studying convergence of Markov chain Monte Carlo algorithms using coupled sample paths*, 1996.
- Neal, *Circularly-coupled Markov chain sampling*, 1999.
- Glynn & Rhee, *Exact estimation for Markov chain equilibrium expectations*, 2014.
- Bou-Rabee, Eberle & Zimmer, *Coupling and Convergence for Hamiltonian Monte Carlo*, 2018.

Back to Metropolis–Hastings (kernel P)

At each iteration t, Markov chain at state X_{t-1},

1. propose $X^\star \sim k(X_{t-1}, \cdot)$,
2. sample $U \sim \mathcal{U}(0,1)$,
3. set X_t based on $X^\star, X_{t-1}, U$, using MH acceptance ratio.

How to propagate two MH chains from states X_{t-1} and Y_{t-1} such that $\{X_t = Y_t\}$ can happen?

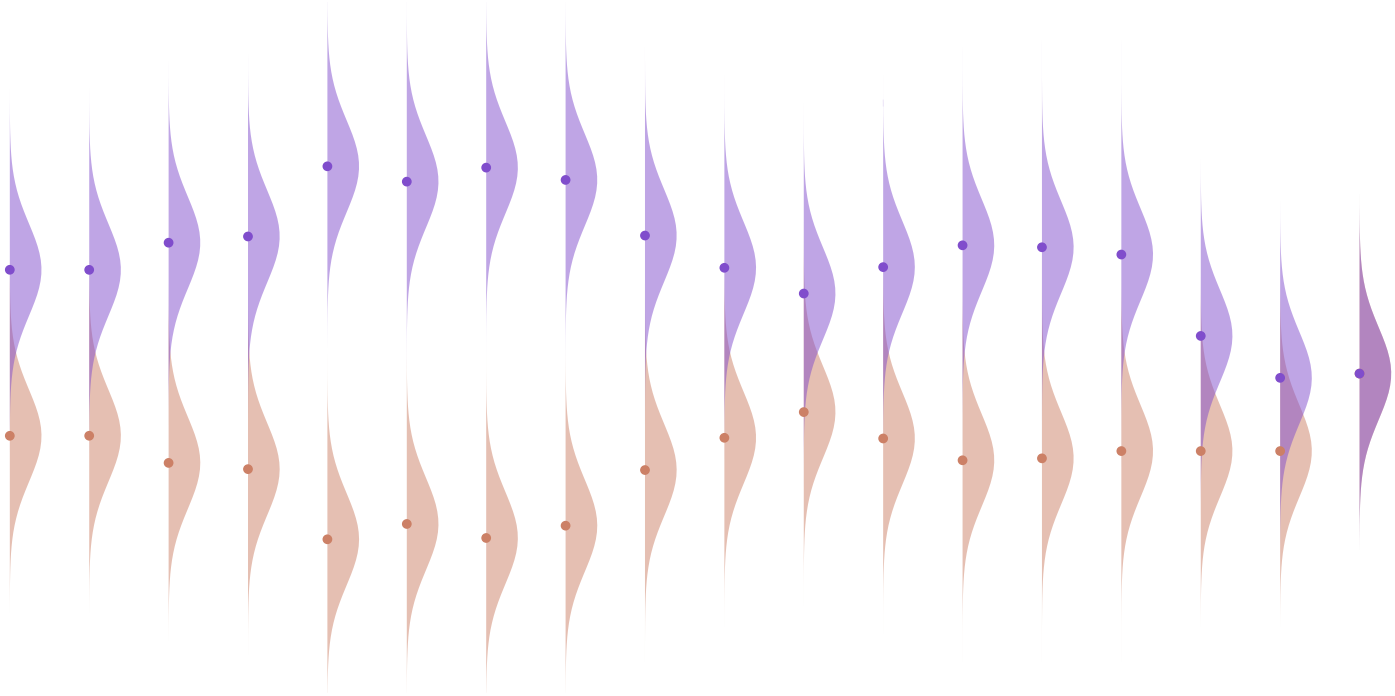
Coupling of Metropolis–Hastings (kernel $\bar{P}$)

At each iteration t, two Markov chains at states X_{t-1}, Y_{t-1},

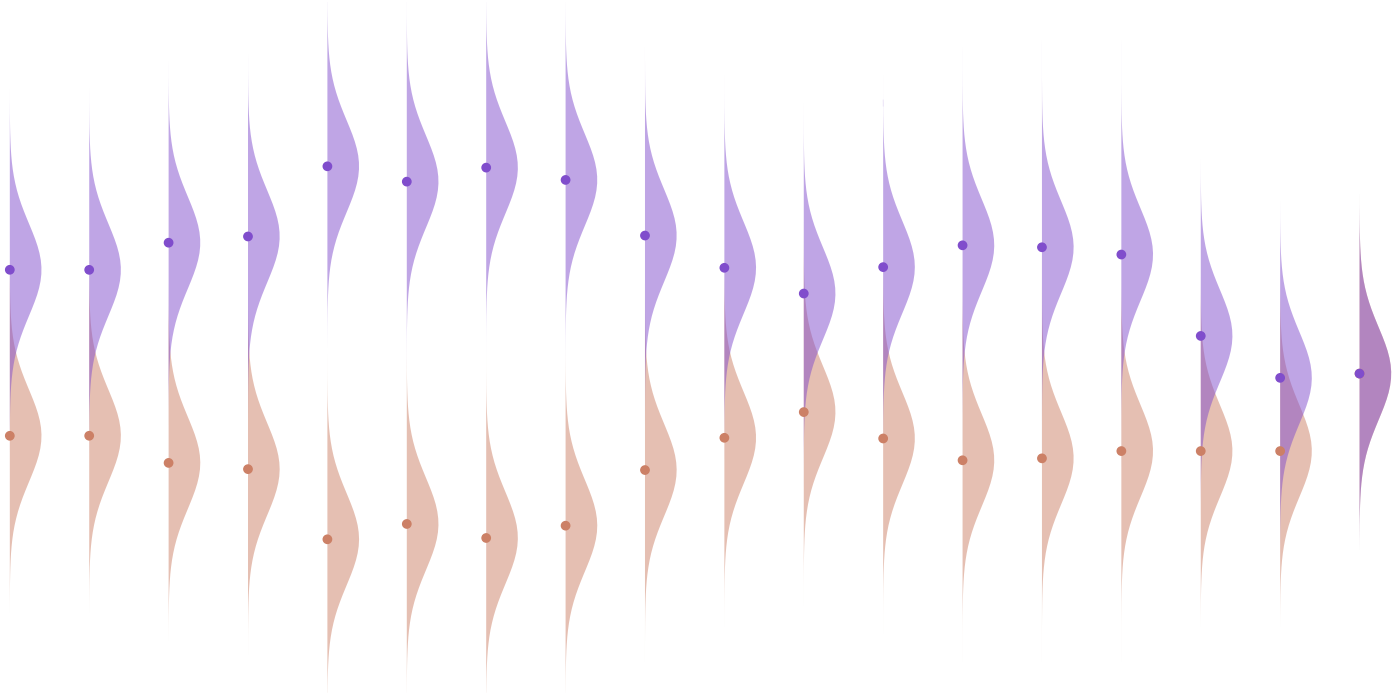
1. propose $(X^\star, Y^\star)$ from coupling of $k(X_{t-1}, \cdot)$, $k(Y_{t-1}, \cdot)$,
2. sample U, V from coupling of $\mathcal{U}(0,1)$ and $\mathcal{U}(0,1)$,
3. set X_t based on $X^\star, X_{t-1}, U$, using MH acceptance ratio,
4. set Y_t based on $Y^\star, Y_{t-1}, V$, using MH acceptance ratio.

O'Leary, Wang & Jacob, *Maximal couplings of the Metropolis–Hastings algorithm*, 2020

Coupling of Metropolis–Hastings

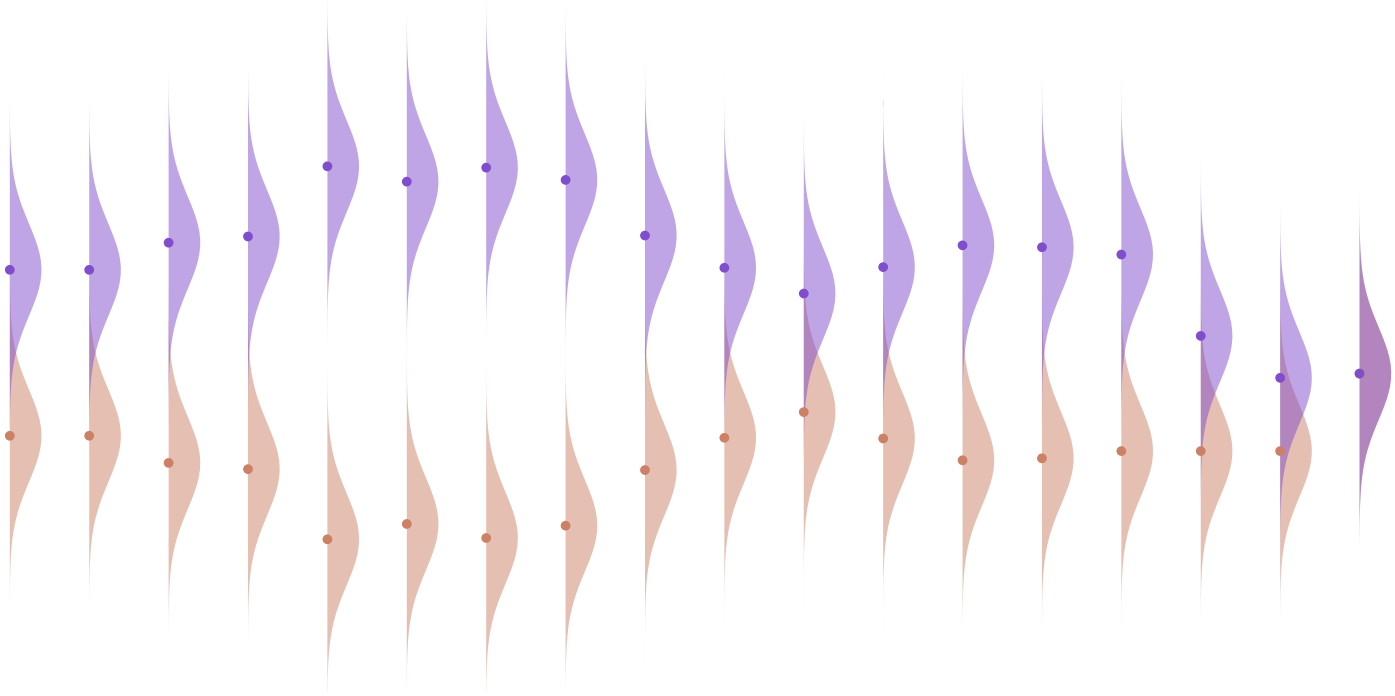


Coupling of Metropolis–Hastings



. . . so we might be able to sample two Markov chains such that they eventually coincide exactly. Great!

Coupling of Metropolis–Hastings



. . . so we might be able to sample two Markov chains such that they eventually coincide exactly. Great!

But how does this help in approximating π?

Outline

Couplings of two lagged chains

Glynn & Rhee, *Exact estimation for MC equilibrium expectations*, 2014.

Couplings of two lagged chains

Glynn & Rhee, *Exact estimation for MC equilibrium expectations*, 2014.

Generate two chains (X_t) and (Y_t) as follows,

- sample X_0 and Y_0 from π_0 (independently, or not),

Couplings of two lagged chains

Glynn & Rhee, *Exact estimation for MC equilibrium expectations*, 2014.

Generate two chains (X_t) and (Y_t) as follows,

- sample X_0 and Y_0 from π_0 (independently, or not),
- sample $X_t|X_{t-1} \sim P(X_{t-1}, \cdot)$, for $t = 1, \ldots, L$, where L is the "lag",

Couplings of two lagged chains

Glynn & Rhee, *Exact estimation for MC equilibrium expectations*, 2014.

Generate two chains (X_t) and (Y_t) as follows,

- sample X_0 and Y_0 from π_0 (independently, or not),
- sample $X_t|X_{t-1} \sim P(X_{t-1}, \cdot)$, for $t = 1, \ldots, L$, where L is the "lag",
- for $t \geq L$, sample $(X_{t+1}, Y_{t+1-L})|(X_t, Y_{t-L}) \sim \bar{P}\left((X_t, Y_{t-L}), \cdot\right)$.

Couplings of two lagged chains

Glynn & Rhee, *Exact estimation for MC equilibrium expectations*, 2014.

Generate two chains (X_t) and (Y_t) as follows,

- sample X_0 and Y_0 from π_0 (independently, or not),
- sample $X_t|X_{t-1} \sim P(X_{t-1}, \cdot)$, for $t = 1, \ldots, L$, where L is the "lag",
- for $t \geq L$, sample $(X_{t+1}, Y_{t+1-L})|(X_t, Y_{t-L}) \sim \bar{P}\left((X_t, Y_{t-L}), \cdot\right)$.

$\bar{P}$ must be such that

- $X_{t+1}|X_t \sim P(X_t, \cdot)$ and $Y_{t+1-L}|Y_{t-L} \sim P(Y_{t-L}, \cdot)$ (thus X_t and Y_t have the same distribution for all $t \geq 0$),

Couplings of two lagged chains

Glynn & Rhee, *Exact estimation for MC equilibrium expectations*, 2014.

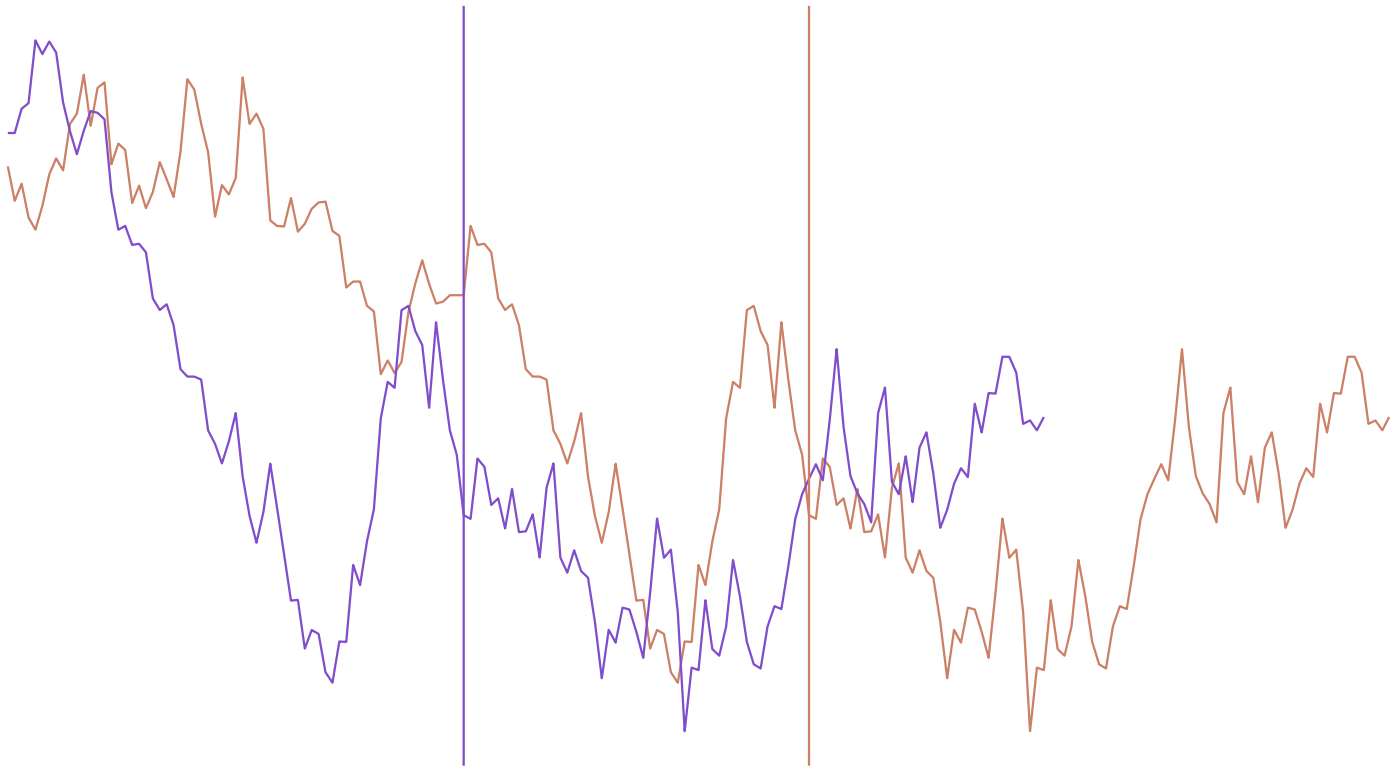
Generate two chains (X_t) and (Y_t) as follows,

- sample X_0 and Y_0 from π_0 (independently, or not),
- sample $X_t|X_{t-1} \sim P(X_{t-1}, \cdot)$, for $t = 1, \ldots, L$, where L is the "lag",
- for $t \geq L$, sample $(X_{t+1}, Y_{t+1-L})|(X_t, Y_{t-L}) \sim \bar{P}\left((X_t, Y_{t-L}), \cdot\right)$.

$\bar{P}$ must be such that

- $X_{t+1}|X_t \sim P(X_t, \cdot)$ and $Y_{t+1-L}|Y_{t-L} \sim P(Y_{t-L}, \cdot)$ (thus X_t and Y_t have the same distribution for all $t \geq 0$),
- there exists a random time τ such that $X_t = Y_{t-L}$ for $t \geq \tau$ (the chains meet).

Coupled Markov chains with a lag



Unbiased estimators with coupled chains

Assume lag $L = 1$. Limit as a telescopic sum, for all $k \geq 0$,

$$\mathbb{E}_\pi[h(X)] = \lim_{t \to \infty} \mathbb{E}[h(X_t)] = \mathbb{E}[h(X_k)] + \sum_{t=k+1}^{\infty} \mathbb{E}[h(X_t) - h(X_{t-1})].$$

Unbiased estimators with coupled chains

Assume lag $L = 1$. Limit as a telescopic sum, for all $k \geq 0$,

$$\mathbb{E}_\pi[h(X)] = \lim_{t \to \infty} \mathbb{E}[h(X_t)] = \mathbb{E}[h(X_k)] + \sum_{t=k+1}^{\infty} \mathbb{E}[h(X_t) - h(X_{t-1})].$$

Since for all $t \geq 0$, X_t and Y_t have the same distribution,

$$= \mathbb{E}[h(X_k)] + \sum_{t=k+1}^{\infty} \mathbb{E}[h(X_t) - h(Y_{t-1})].$$

Unbiased estimators with coupled chains

Assume lag $L = 1$. Limit as a telescopic sum, for all $k \geq 0$,

$$\mathbb{E}_\pi[h(X)] = \lim_{t \to \infty} \mathbb{E}[h(X_t)] = \mathbb{E}[h(X_k)] + \sum_{t=k+1}^{\infty} \mathbb{E}[h(X_t) - h(X_{t-1})].$$

Since for all $t \geq 0$, X_t and Y_t have the same distribution,

$$= \mathbb{E}[h(X_k)] + \sum_{t=k+1}^{\infty} \mathbb{E}[h(X_t) - h(Y_{t-1})].$$

If we can swap expectation and limit,

$$= \mathbb{E}[h(X_k) + \sum_{t=k+1}^{\infty} (h(X_t) - h(Y_{t-1}))],$$

then $h(X_k) + \sum_{t=k+1}^{\infty}(h(X_t) - h(Y_{t-1}))$ is unbiased.

Unbiased estimators with coupled chains

Unbiased estimator is given by

$$H_k(X, Y) = h(X_k) + \sum_{t=k+1}^{\tau-1} (h(X_t) - h(Y_{t-1})),$$

with the convention $\sum_{t=k+1}^{\tau-1}\{\cdot\} = 0$ if $\tau - 1 < k + 1$.

Unbiased estimators with coupled chains

Unbiased estimator is given by

$$H_k(X, Y) = h(X_k) + \sum_{t=k+1}^{\tau-1} (h(X_t) - h(Y_{t-1})),$$

with the convention $\sum_{t=k+1}^{\tau-1}\{\cdot\} = 0$ if $\tau - 1 < k + 1$.

$h(X_k)$ alone is biased; the other terms correct for the bias.

Unbiased estimators with coupled chains

Unbiased estimator is given by

$$H_k(X, Y) = h(X_k) + \sum_{t=k+1}^{\tau-1} (h(X_t) - h(Y_{t-1})),$$

with the convention $\sum_{t=k+1}^{\tau-1}\{\cdot\} = 0$ if $\tau - 1 < k + 1$.

$h(X_k)$ alone is biased; the other terms correct for the bias.

Glynn & Rhee, *Exact estimation for Markov chain equilibrium expectations*, 2014.

Improved unbiased estimators

Efficiency matters, thus we introduce a variation of the previous estimator, defined for integers $k \leq m$ as

$$H_{k:m}(X, Y) = \frac{1}{m - k + 1} \sum_{t=k}^{m} H_t(X, Y).$$

Here k is the "start time" and m the "prospective end time".

Improved unbiased estimators

Efficiency matters, thus we introduce a variation of the previous estimator, defined for integers $k \leq m$ as

$$H_{k:m}(X, Y) = \frac{1}{m - k + 1} \sum_{t=k}^{m} H_t(X, Y).$$

Here k is the "start time" and m the "prospective end time".

It can also be written

$$\frac{1}{m - k + 1} \sum_{t=k}^{m} h(X_t) + \sum_{t=k+1}^{\tau-1} \min\left(1, \frac{t - k}{m - k + 1}\right) (h(X_t) - h(Y_{t-1})).$$

Improved unbiased estimators

Efficiency matters, thus we introduce a variation of the previous estimator, defined for integers $k \leq m$ as

$$H_{k:m}(X, Y) = \frac{1}{m - k + 1} \sum_{t=k}^{m} H_t(X, Y).$$

Here k is the "start time" and m the "prospective end time".

It can also be written

$$\frac{1}{m - k + 1} \sum_{t=k}^{m} h(X_t) + \sum_{t=k+1}^{\tau-1} \min\left(1, \frac{t - k}{m - k + 1}\right) (h(X_t) - h(Y_{t-1})).$$

Standard MCMC average + bias correction terms.

Improved unbiased estimators

Efficiency matters, thus we introduce a variation of the previous estimator, defined for integers $k \leq m$ as

$$H_{k:m}(X, Y) = \frac{1}{m - k + 1} \sum_{t=k}^{m} H_t(X, Y).$$

Here k is the "start time" and m the "prospective end time".

It can also be written

$$\frac{1}{m - k + 1} \sum_{t=k}^{m} h(X_t) + \sum_{t=k+1}^{\tau-1} \min\left(1, \frac{t - k}{m - k + 1}\right) (h(X_t) - h(Y_{t-1})).$$

Standard MCMC average + bias correction terms.

As $k \to \infty$ or $m \to \infty$, bias correction goes to zero.
Works with mild modifications for any choice of lag $L \geq 1$.

Contrast with previous work

Estimator in Glynn & Rhee, 2014,

$$h(X_0) + \sum_{j=1}^{R} \frac{h(X_j) - h(Y_{j-1})}{\mathbb{P}(R \geq j)},$$

where R is a random variable on $\mathbb{N}$. In the context of MCMC, see Agapiou, Roberts & Vollmer, 2018.

Contrast with previous work

Estimator in Glynn & Rhee, 2014,

$$h(X_0) + \sum_{j=1}^{R} \frac{h(X_j) - h(Y_{j-1})}{\mathbb{P}(R \geq j)},$$

where R is a random variable on $\mathbb{N}$. In the context of MCMC, see Agapiou, Roberts & Vollmer, 2018.

Under contractive couplings + appropriate truncation variable, these estimators have finite cost and finite variance.

Contrast with previous work

Estimator in Glynn & Rhee, 2014,

$$h(X_0) + \sum_{j=1}^{R} \frac{h(X_j) - h(Y_{j-1})}{\mathbb{P}(R \geq j)},$$

where R is a random variable on $\mathbb{N}$. In the context of MCMC, see Agapiou, Roberts & Vollmer, 2018.

Under contractive couplings + appropriate truncation variable, these estimators have finite cost and finite variance.

But appropriate choice of truncation variable requires information about contraction rate of the coupling.

Contrast with previous work

Estimator in Glynn & Rhee, 2014,

$$h(X_0) + \sum_{j=1}^{R} \frac{h(X_j) - h(Y_{j-1})}{\mathbb{P}(R \geq j)},$$

where R is a random variable on $\mathbb{N}$. In the context of MCMC, see Agapiou, Roberts & Vollmer, 2018.

Under contractive couplings + appropriate truncation variable, these estimators have finite cost and finite variance.

But appropriate choice of truncation variable requires information about contraction rate of the coupling.

Efficiency could be low compared to that of ergodic average.

Validity conditions

1. Marginal chain converges:
   $\mathbb{E}[h(X_t)] \to \mathbb{E}_\pi[h(X)]$,
   and $\exists \eta > 0, D < \infty$ such that $\forall t \geq 0$, $\mathbb{E}[|h(X_t)|^{2+\eta}] \leq D$.

Validity conditions

1. Marginal chain converges:
   $\mathbb{E}[h(X_t)] \to \mathbb{E}_\pi[h(X)]$,
   and $\exists \eta > 0, D < \infty$ such that $\forall t \geq 0$, $\mathbb{E}[|h(X_t)|^{2+\eta}] \leq D$.

2. Meeting time τ has Geometric tails:

$$\exists C < +\infty \quad \exists \delta \in (0,1) \quad \forall t \geq 0 \quad \mathbb{P}(\tau > t) \leq C\delta^t.$$

Validity conditions

1. Marginal chain converges:
   $\mathbb{E}[h(X_t)] \to \mathbb{E}_\pi[h(X)]$,
   and $\exists \eta > 0, D < \infty$ such that $\forall t \geq 0$, $\mathbb{E}[|h(X_t)|^{2+\eta}] \leq D$.

2. Meeting time τ has Geometric tails:

$$\exists C < +\infty \quad \exists \delta \in (0,1) \quad \forall t \geq 0 \quad \mathbb{P}(\tau > t) \leq C\delta^t.$$

3. Chains are faithful: $X_t = Y_{t-L}$ for all $t \geq \tau$.

Validity conditions

1. Marginal chain converges:
   $\mathbb{E}[h(X_t)] \to \mathbb{E}_\pi[h(X)]$,
   and $\exists \eta > 0, D < \infty$ such that $\forall t \geq 0$, $\mathbb{E}[|h(X_t)|^{2+\eta}] \leq D$.

2. Meeting time τ has Geometric tails:

$$\exists C < +\infty \quad \exists \delta \in (0,1) \quad \forall t \geq 0 \quad \mathbb{P}(\tau > t) \leq C\delta^t.$$

3. Chains are faithful: $X_t = Y_{t-L}$ for all $t \geq \tau$.

Proposition 1: under these conditions, $H_{k:m}(X, Y)$ is unbiased, has finite expected cost and finite variance.

Validity conditions

1. Marginal chain converges:
   $\mathbb{E}[h(X_t)] \to \mathbb{E}_\pi[h(X)]$,
   and $\exists \eta > 0, D < \infty$ such that $\forall t \geq 0$, $\mathbb{E}[|h(X_t)|^{2+\eta}] \leq D$.

2. Meeting time τ has Geometric tails:

$$\exists C < +\infty \quad \exists \delta \in (0,1) \quad \forall t \geq 0 \quad \mathbb{P}(\tau > t) \leq C\delta^t.$$

3. Chains are faithful: $X_t = Y_{t-L}$ for all $t \geq \tau$.

Proposition 1: under these conditions, $H_{k:m}(X, Y)$ is unbiased, has finite expected cost and finite variance.

Jacob, O'Leary, Atchadé, *Unbiased MCMC with couplings*, 2020.

Signed measure estimator

Replacing function evaluations by delta masses leads to

$$\hat{\pi}(\cdot) = \frac{1}{m-k+1} \sum_{t=k}^{m} \delta_{X_t}(\cdot) + \sum_{t=k+1}^{\tau-1} \min\left(1, \frac{t-k}{m-k+1}\right) (\delta_{X_t}(\cdot) - \delta_{Y_{t-1}}(\cdot)).$$

Signed measure estimator

Replacing function evaluations by delta masses leads to

$$\hat{\pi}(\cdot) = \frac{1}{m-k+1} \sum_{t=k}^{m} \delta_{X_t}(\cdot) + \sum_{t=k+1}^{\tau-1} \min\left(1, \frac{t-k}{m-k+1}\right) (\delta_{X_t}(\cdot) - \delta_{Y_{t-1}}(\cdot)).$$

This is of the form $\hat{\pi}(\cdot) = \sum_{n=1}^{N} \omega_n \delta_{Z_n}(\cdot)$, where $\sum_{n=1}^{N} \omega_n = 1$ but some ω_n might be negative.

Signed measure estimator

Replacing function evaluations by delta masses leads to

$$\hat{\pi}(\cdot) = \frac{1}{m-k+1} \sum_{t=k}^{m} \delta_{X_t}(\cdot) + \sum_{t=k+1}^{\tau-1} \min\left(1, \frac{t-k}{m-k+1}\right) (\delta_{X_t}(\cdot) - \delta_{Y_{t-1}}(\cdot)).$$

This is of the form $\hat{\pi}(\cdot) = \sum_{n=1}^{N} \omega_n \delta_{Z_n}(\cdot)$, where $\sum_{n=1}^{N} \omega_n = 1$ but some ω_n might be negative.

Denote by $\bar{\pi}$ the average of R independent copies of $\hat{\pi}$.
Proposition 2: uniform convergence of $\bar{\pi}$ to π, as $R \to \infty$.

Efficiency

Writing $H_{k:m}(X, Y) = \mathrm{MCMC}_{k:m} + \mathrm{BC}_{k:m}$,

Efficiency

Writing $H_{k:m}(X, Y) = \mathrm{MCMC}_{k:m} + \mathrm{BC}_{k:m}$,

then, denoting the MSE of $\mathrm{MCMC}_{k:m}$ by $\mathrm{MSE}_{k:m}$,

Efficiency

Writing $H_{k:m}(X, Y) = \mathrm{MCMC}_{k:m} + \mathrm{BC}_{k:m}$,

then, denoting the MSE of $\mathrm{MCMC}_{k:m}$ by $\mathrm{MSE}_{k:m}$,

$$\mathbb{V}[H_{k:m}(X, Y)] \leq \mathrm{MSE}_{k:m} + 2\sqrt{\mathrm{MSE}_{k:m}}\sqrt{\mathbb{E}\left[\mathrm{BC}_{k:m}^2\right]} + \mathbb{E}\left[\mathrm{BC}_{k:m}^2\right].$$

Efficiency

Writing $H_{k:m}(X, Y) = \text{MCMC}_{k:m} + \text{BC}_{k:m}$,

then, denoting the MSE of $\text{MCMC}_{k:m}$ by $\text{MSE}_{k:m}$,

$$\mathbb{V}[H_{k:m}(X, Y)] \leq \text{MSE}_{k:m} + 2\sqrt{\text{MSE}_{k:m}}\sqrt{\mathbb{E}\left[\text{BC}_{k:m}^2\right]} + \mathbb{E}\left[\text{BC}_{k:m}^2\right].$$

Proposition 3: under geometric drift condition and regularity assumptions on h, for some $\delta < 1$, $C < \infty$,

$$\mathbb{E}[\text{BC}_{k:m}^2] \leq \frac{C\delta^k}{(m-k+1)^2},$$

with δ directly related to tails of the meeting time.

Tails of the meeting time

One of the assumptions was that τ has Geometric tails.

Tails of the meeting time

One of the assumptions was that τ has Geometric tails.

Proposition 4: if kernels $P, \bar{P}$ are such that there exists $V : \mathcal{X} \to [1, \infty)$, $\lambda \in (0, 1)$, $b < \infty$, and a small set $\mathcal{C}$ such that

Tails of the meeting time

One of the assumptions was that τ has Geometric tails.

Proposition 4: if kernels $P, \bar{P}$ are such that there exists $V : \mathcal{X} \to [1, \infty)$, $\lambda \in (0, 1)$, $b < \infty$, and a small set $\mathcal{C}$ such that

- for all $x \in \mathcal{X}$, $\int P(x, dx')V(x') \leq \lambda V(x) + b\mathbb{1}(x \in \mathcal{C})$,

Tails of the meeting time

One of the assumptions was that τ has Geometric tails.

Proposition 4: if kernels $P, \bar{P}$ are such that there exists $V : \mathcal{X} \to [1, \infty)$, $\lambda \in (0, 1)$, $b < \infty$, and a small set $\mathcal{C}$ such that

- for all $x \in \mathcal{X}$, $\int P(x, dx')V(x') \leq \lambda V(x) + b\mathbb{1}(x \in \mathcal{C})$,
- $\mathcal{C} = \{x : V(x) \leq \ell\}$ for some ℓ with $\lambda + b/(1 + \ell) < 1$,

Tails of the meeting time

One of the assumptions was that τ has Geometric tails.

Proposition 4: if kernels $P, \bar{P}$ are such that there exists $V : \mathcal{X} \to [1, \infty)$, $\lambda \in (0, 1)$, $b < \infty$, and a small set $\mathcal{C}$ such that

- for all $x \in \mathcal{X}$, $\int P(x, dx')V(x') \leq \lambda V(x) + b\mathbb{1}(x \in \mathcal{C})$,
- $\mathcal{C} = \{x : V(x) \leq \ell\}$ for some ℓ with $\lambda + b/(1 + \ell) < 1$,
- there exists $\varepsilon > 0$ such that meeting occurs with probability at least ε whenever two chains are in $\mathcal{C}$,

Tails of the meeting time

One of the assumptions was that τ has Geometric tails.

Proposition 4: if kernels $P, \bar{P}$ are such that there exists $V : \mathcal{X} \to [1, \infty)$, $\lambda \in (0, 1)$, $b < \infty$, and a small set $\mathcal{C}$ such that

- for all $x \in \mathcal{X}$, $\int P(x, dx')V(x') \leq \lambda V(x) + b\mathbb{1}(x \in \mathcal{C})$,
- $\mathcal{C} = \{x : V(x) \leq \ell\}$ for some ℓ with $\lambda + b/(1 + \ell) < 1$,
- there exists $\varepsilon > 0$ such that meeting occurs with probability at least ε whenever two chains are in $\mathcal{C}$,

then indeed the meeting time τ has Geometric tails.

Polynomial tails

Middleton, Deligiannidis, Doucet, Jacob, *Unbiased MCMC for intractable target distributions*, 2020.

1. Marginal chain converges:
   $\mathbb{E}[h(X_t)] \to \mathbb{E}_\pi[h(X)]$,
   and $h(X_t)$ has $(2+\eta)$-finite moments for all t.

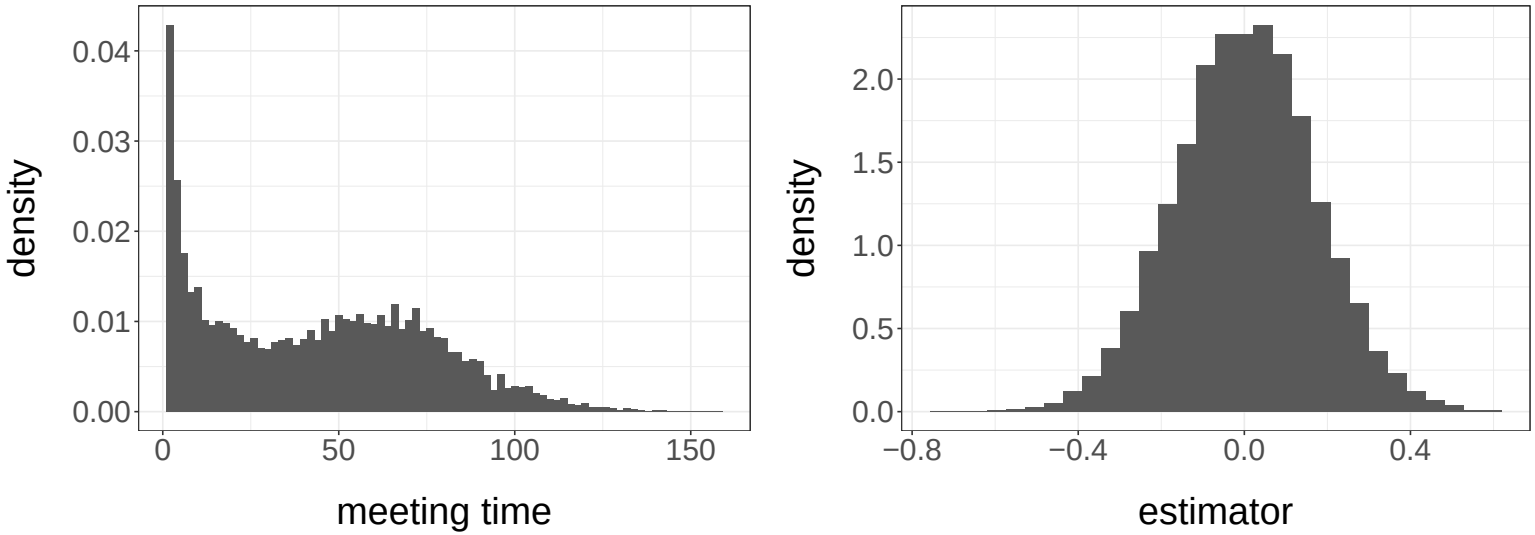
2. Meeting time τ has **polynomial** tails:

$$\exists C < +\infty \quad \exists \delta > 2(2\eta^{-1} + 1) \quad \forall t \geq 0 \quad \mathbb{P}(\tau > t) \leq C t^{-\delta}.$$

3. Chains are faithful: $X_t = Y_{t-L}$ for all $t \geq \tau$.

Condition 2 itself implied by e.g. **polynomial** drift condition.

Unbiased random walk MH on toy example



$\pi = \mathcal{N}(0, 1)$, RWMH with Normal proposal std $= 0.5$, $\pi_0 = \mathcal{N}(10, 3^2)$.

Chains coupled with reflection-maximal couplings of proposals.

Estimators obtained with $k = 200$, $m = 1000$, lag $L = 1$,

average cost ≈ 1048, variance ≈ 0.028.

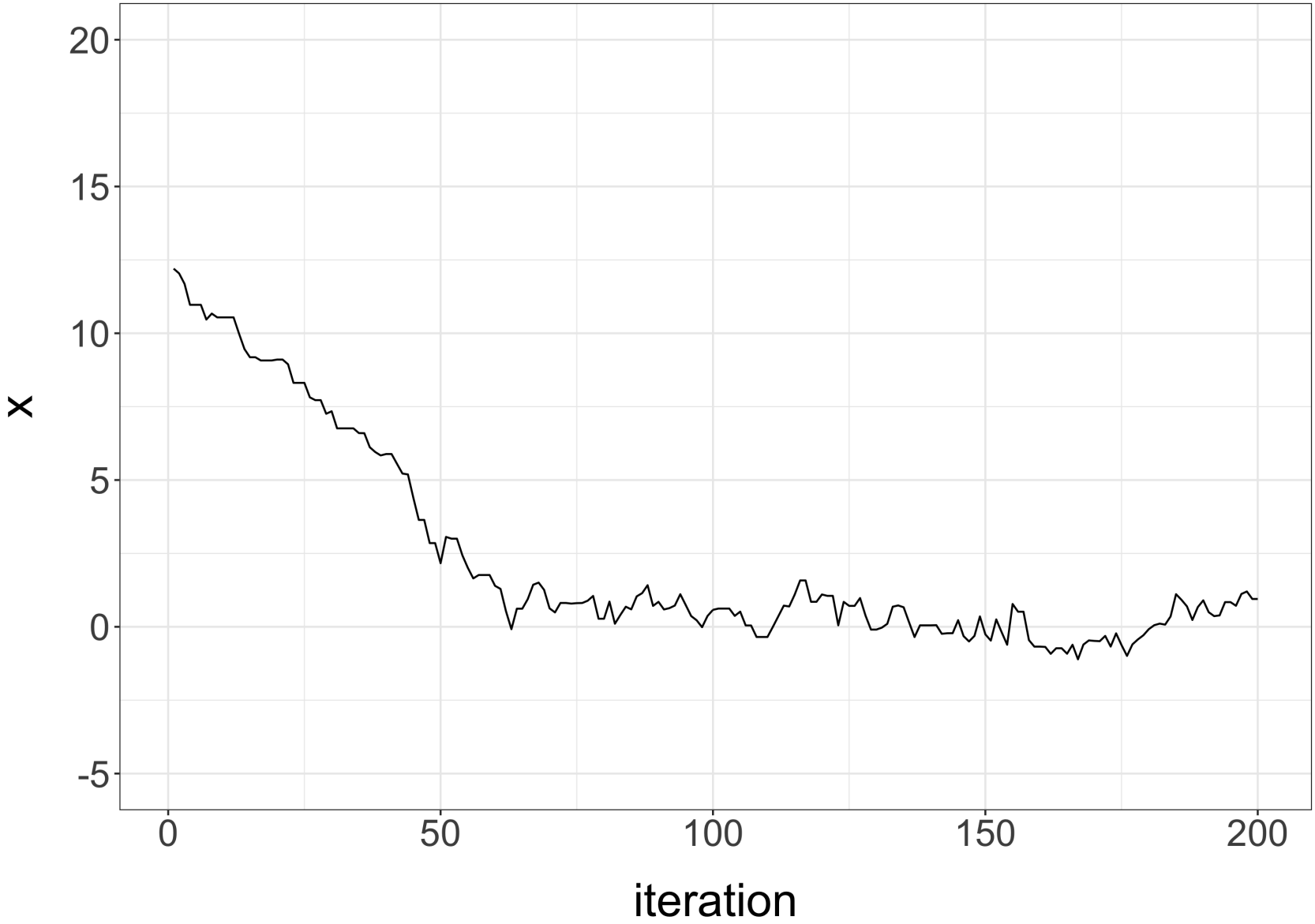
Inefficiency compared to MCMC average: $\times 1.3$ (approximately).

More realistic examples

- Various Gibbs samplers in Jacob, O'Leary & Atchadé, 2020, also parallel tempering.
- Conditional particle filters in Jacob, Lindsten & Schön, 2019, Lee, Singh & Vihola, 2018.
- Hamiltonian Monte Carlo samplers in Heng & Jacob, 2019, see also Piponi, Hoffman & Sountsov, 2020.
- Particle MCMC and exchange algorithm, in Middleton, Deligiannidis, Doucet, Jacob, 2020.
- Unbiased Gradient Estimation for Variational Auto-Encoders, in Ruiz, Titsias, Cemgil & Doucet, 2020.

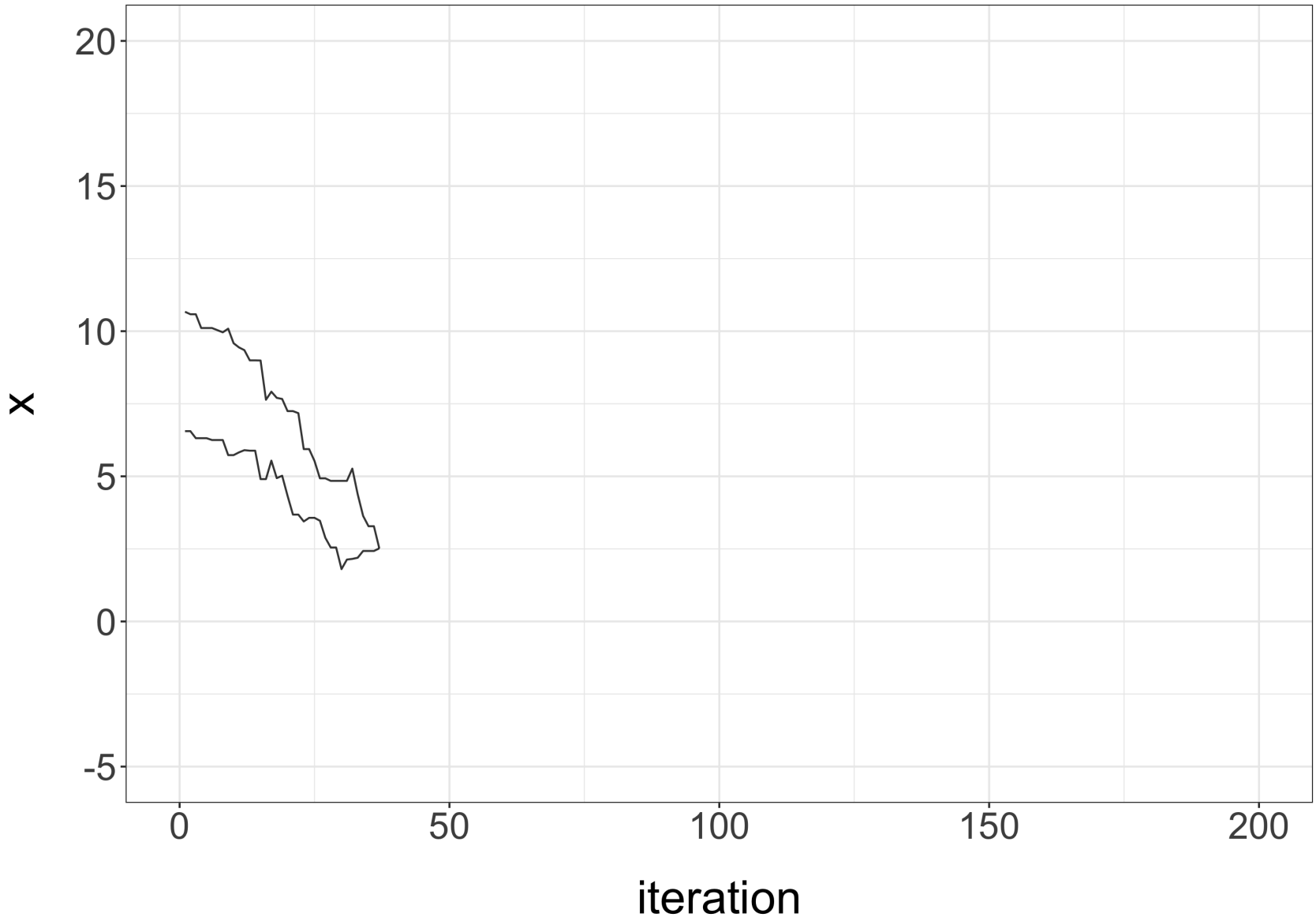
Outline

RWMH on Normal target: trace plot



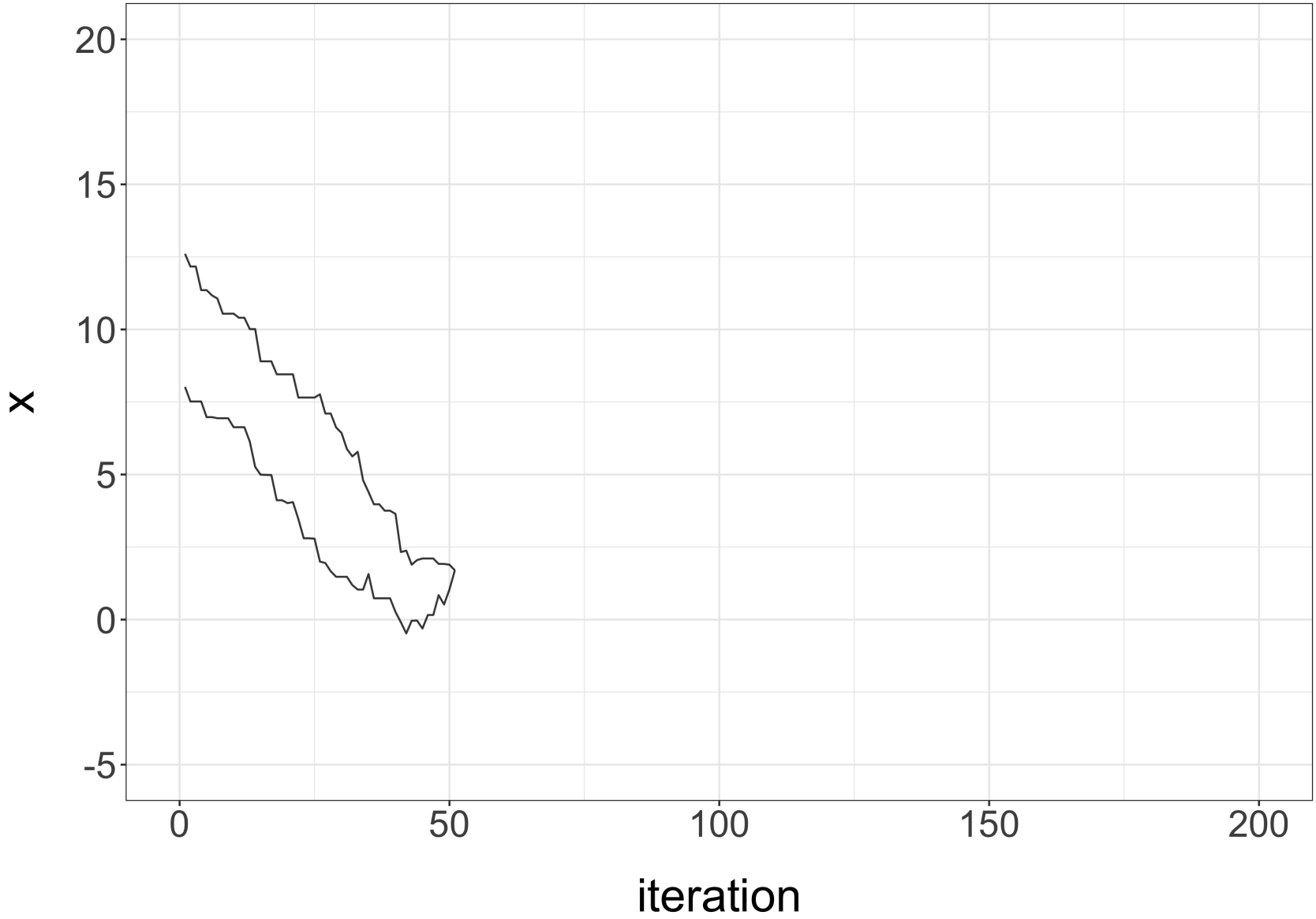
$\pi = \mathcal{N}(0, 1)$, MH with Normal proposal std $= 0.5$, $\pi_0 = \mathcal{N}(10, 3^2)$

RWMH on Normal target: coupled paths



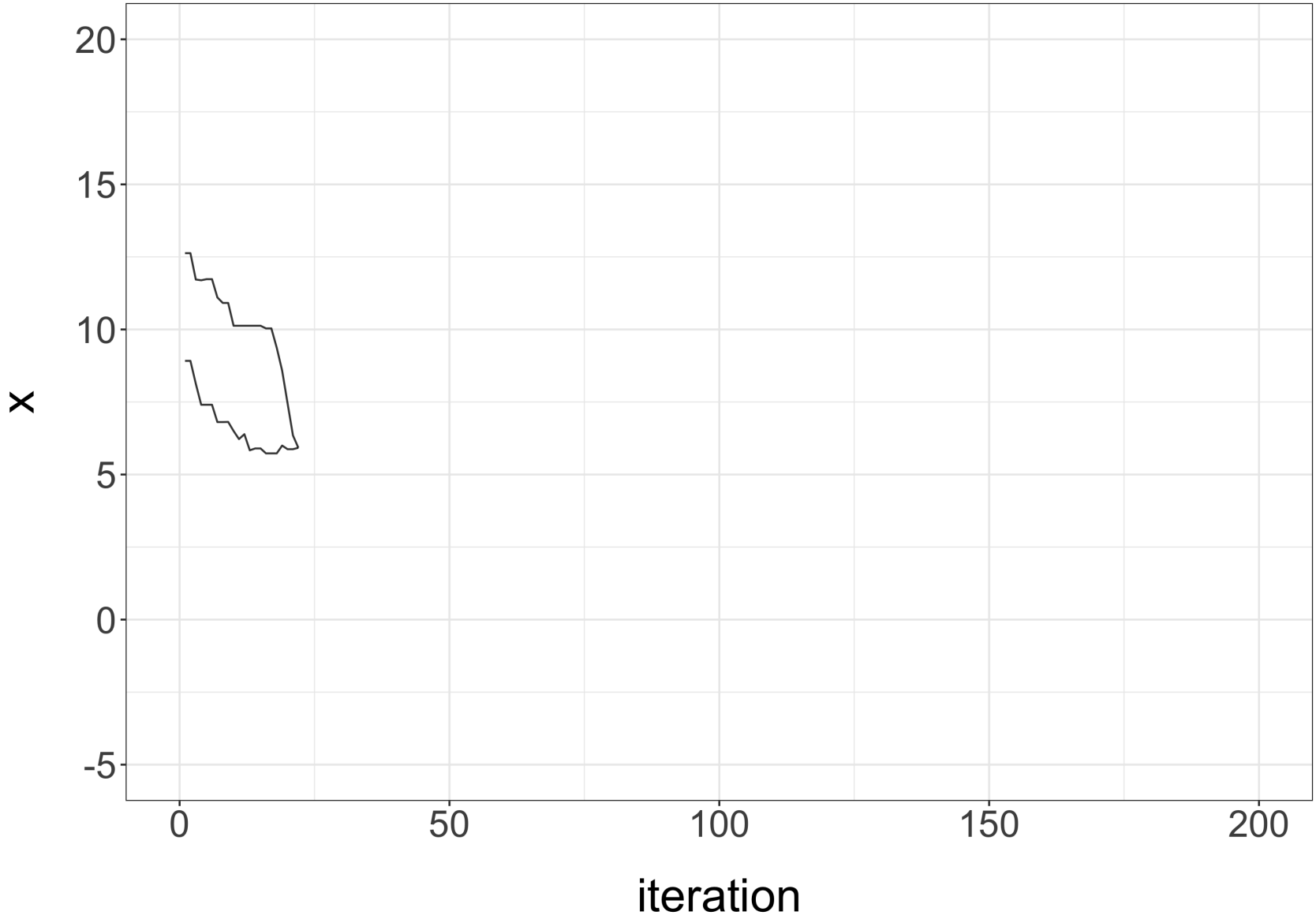
$\pi = \mathcal{N}(0, 1)$, MH with Normal proposal std $= 0.5$, $\pi_0 = \mathcal{N}(10, 3^2)$

RWMH on Normal target: coupled paths



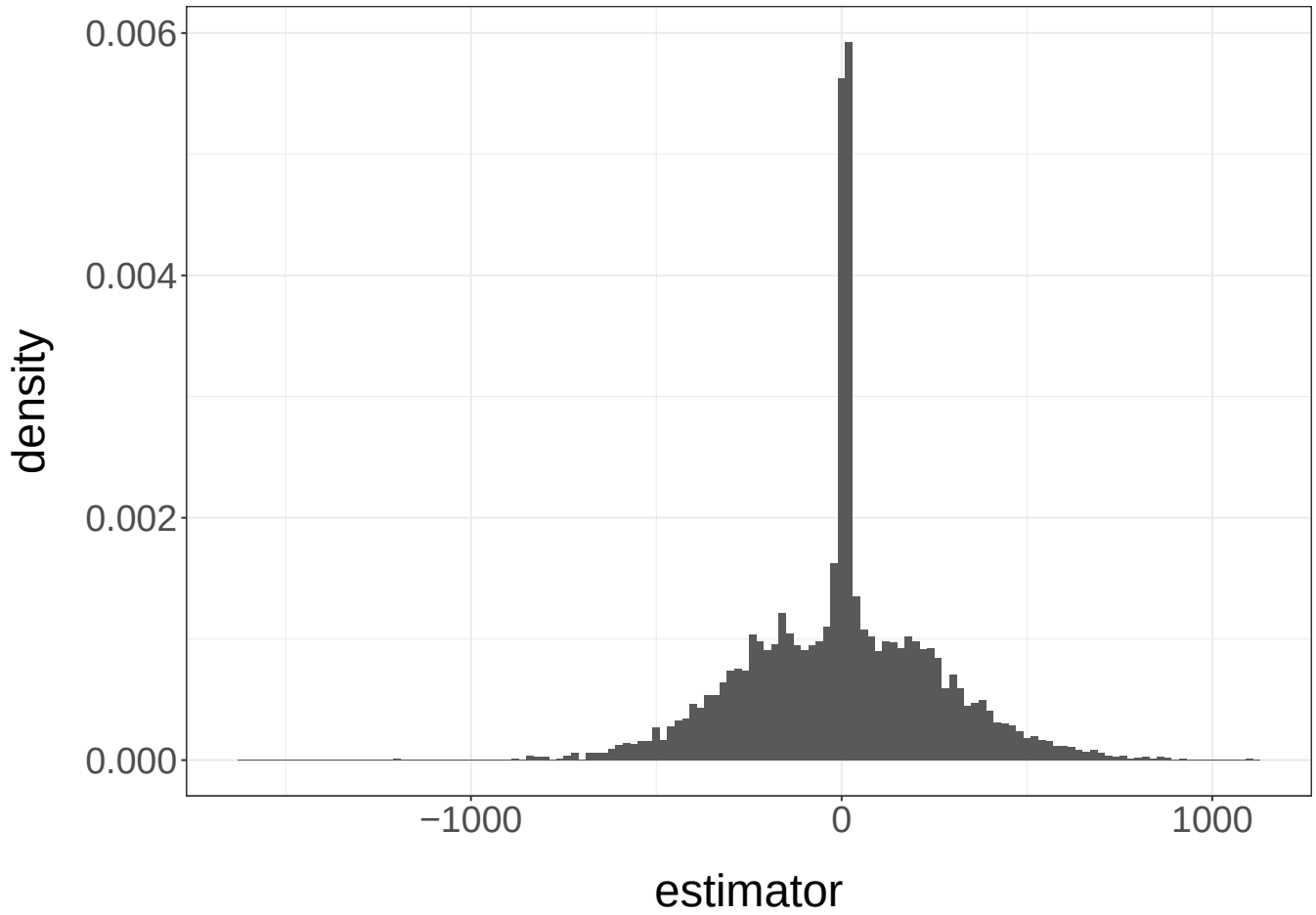
$\pi = \mathcal{N}(0, 1)$, MH with Normal proposal std $= 0.5$, $\pi_0 = \mathcal{N}(10, 3^2)$

RWMH on Normal target: coupled paths



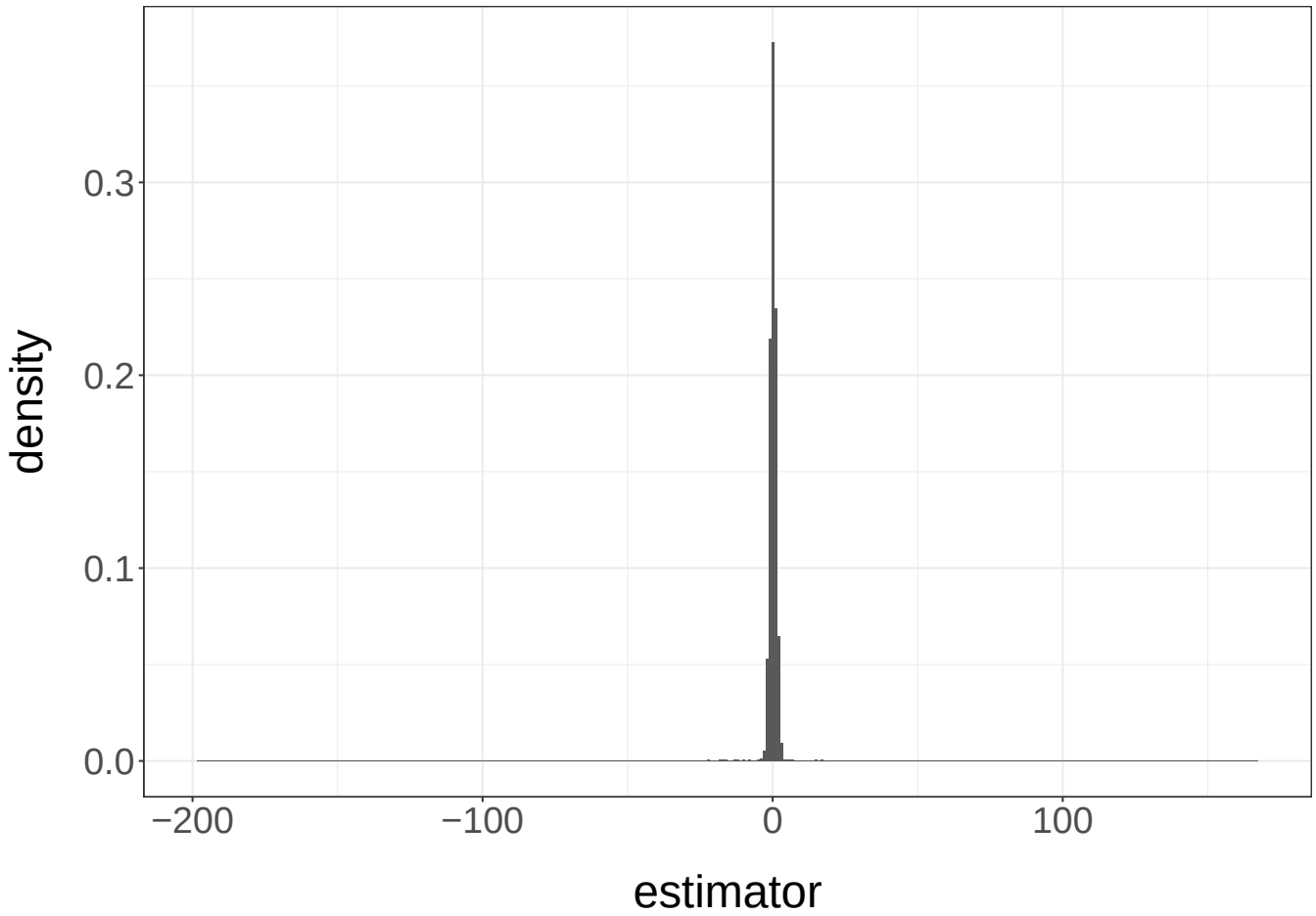
$\pi = \mathcal{N}(0, 1)$, MH with Normal proposal std $= 0.5$, $\pi_0 = \mathcal{N}(10, 3^2)$

RWMH on Normal target: estimators of $\mathbb{E}_\pi[X]$



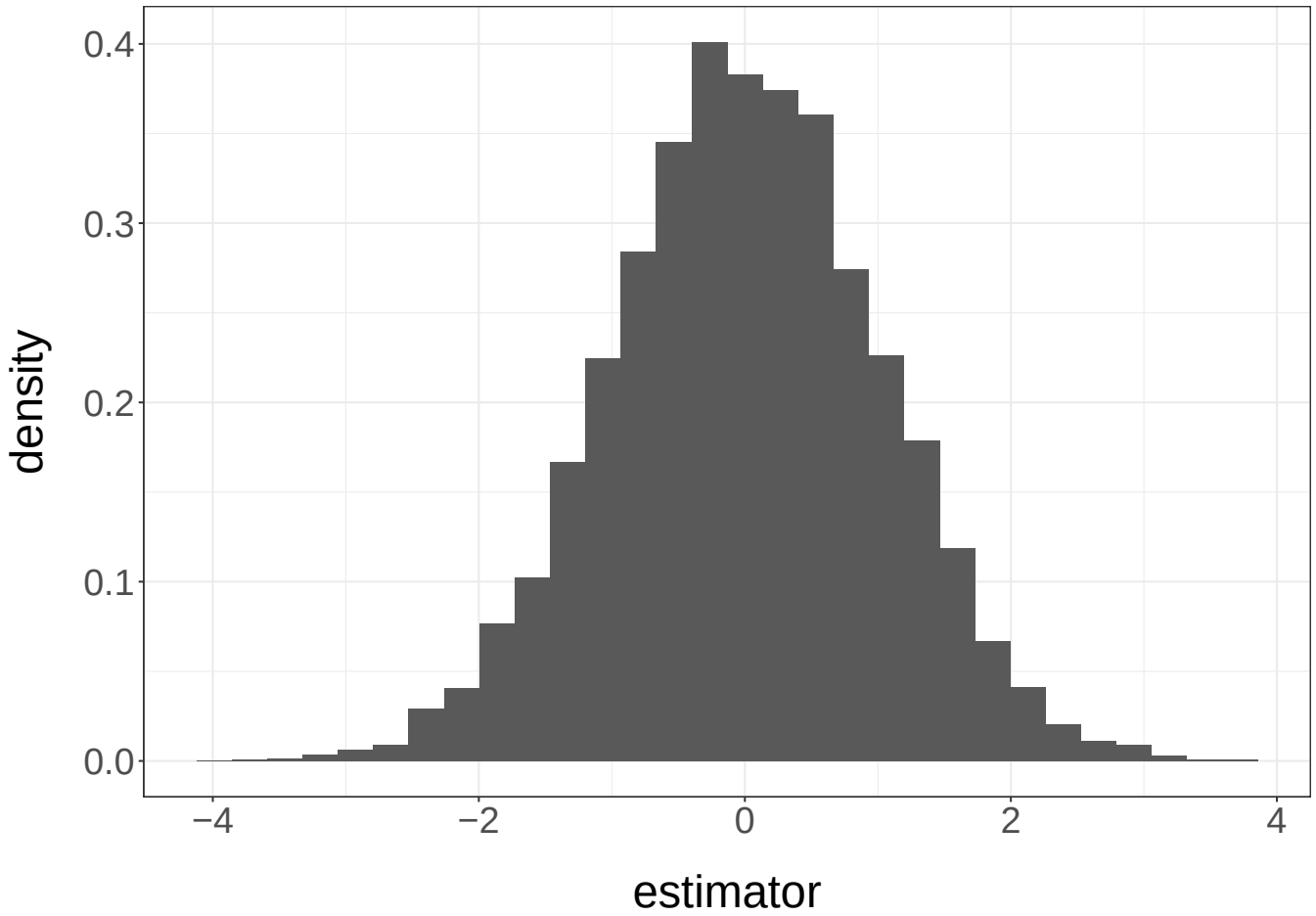
$k = 0$, $\mathbb{E}[2\tau] \approx 96$, $\mathbb{V}[H_0(X, Y)] \approx 65,000$.
$\pi = \mathcal{N}(0, 1)$, RWMH with Normal proposal std $= 0.5$, $\pi_0 = \mathcal{N}(10, 3^2)$

RWMH on Normal target: estimators of $\mathbb{E}_\pi[X]$



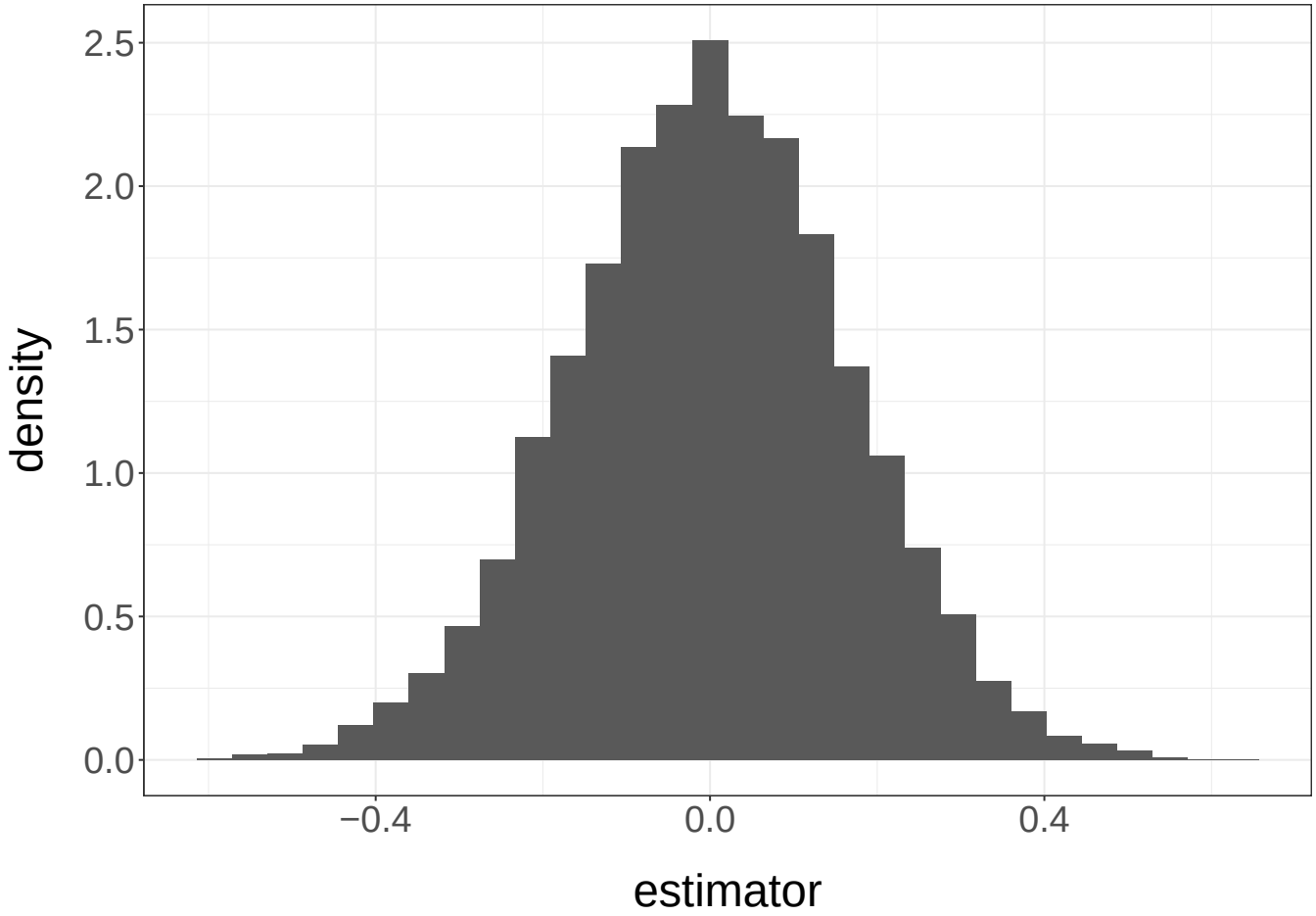
$k = 100$, $\mathbb{E}[\max(k + \tau, 2\tau)] \approx 148$, $\mathbb{V}[H_k(X, Y)] \approx 100$.
$\pi = \mathcal{N}(0, 1)$, RWMH with Normal proposal std $= 0.5$, $\pi_0 = \mathcal{N}(10, 3^2)$

RWMH on Normal target: estimators of $\mathbb{E}_\pi[X]$



$k = 200$, $\mathbb{E}[\max(k+\tau, 2\tau)] \approx 248$, $\mathbb{V}[H_k(X,Y)] \approx 1$.
$\pi = \mathcal{N}(0,1)$, RWMH with Normal proposal std $= 0.5$, $\pi_0 = \mathcal{N}(10, 3^2)$

RWMH on Normal target: estimators of $\mathbb{E}_\pi[X]$



$k = 200$, $m = 1000$, $\mathbb{E}[\max(m + \tau, 2\tau)] \approx 1048$,
$\mathbb{V}[H_k(X, Y)] \approx 0.028$.
$\pi = \mathcal{N}(0, 1)$, RWMH with Normal proposal std $= 0.5$, $\pi_0 = \mathcal{N}(10, 3^2)$

Experiments in multivariate settings

Target is d-dimensional $\mathcal{N}(0_d, V)$.

Initialization from either the target, or from $\mathcal{N}(1_d, I_d)$.

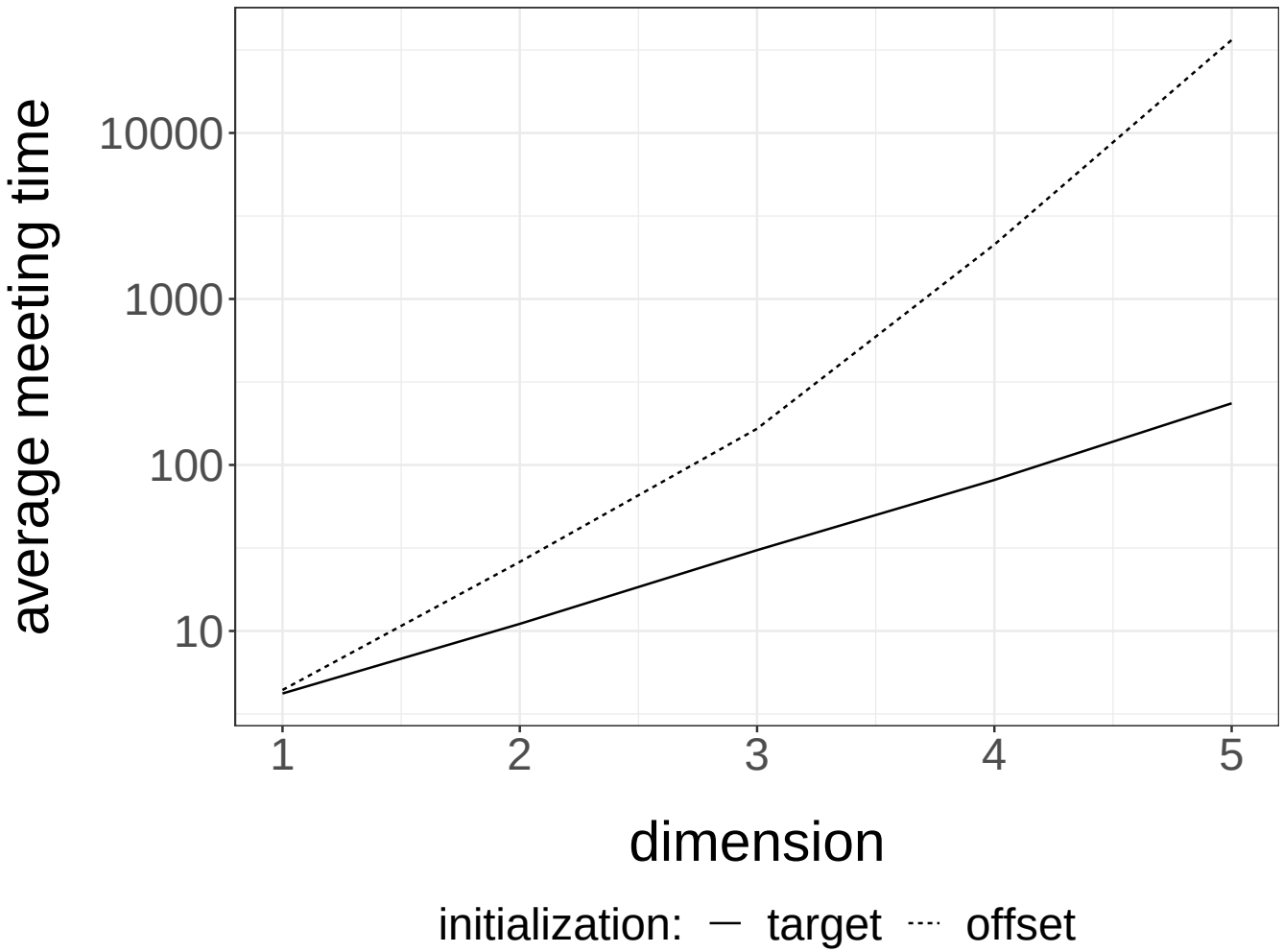
MH with Normal random walk proposal, with variance V/d.

Later, Gibbs sampler updating individual components, with

- V sampled from inverse Wishart, with identity scale and d degrees of freedom, leading to dense matrix,
- or V constructed as $V_{ij} = 0.5^{-|i-j|}$.

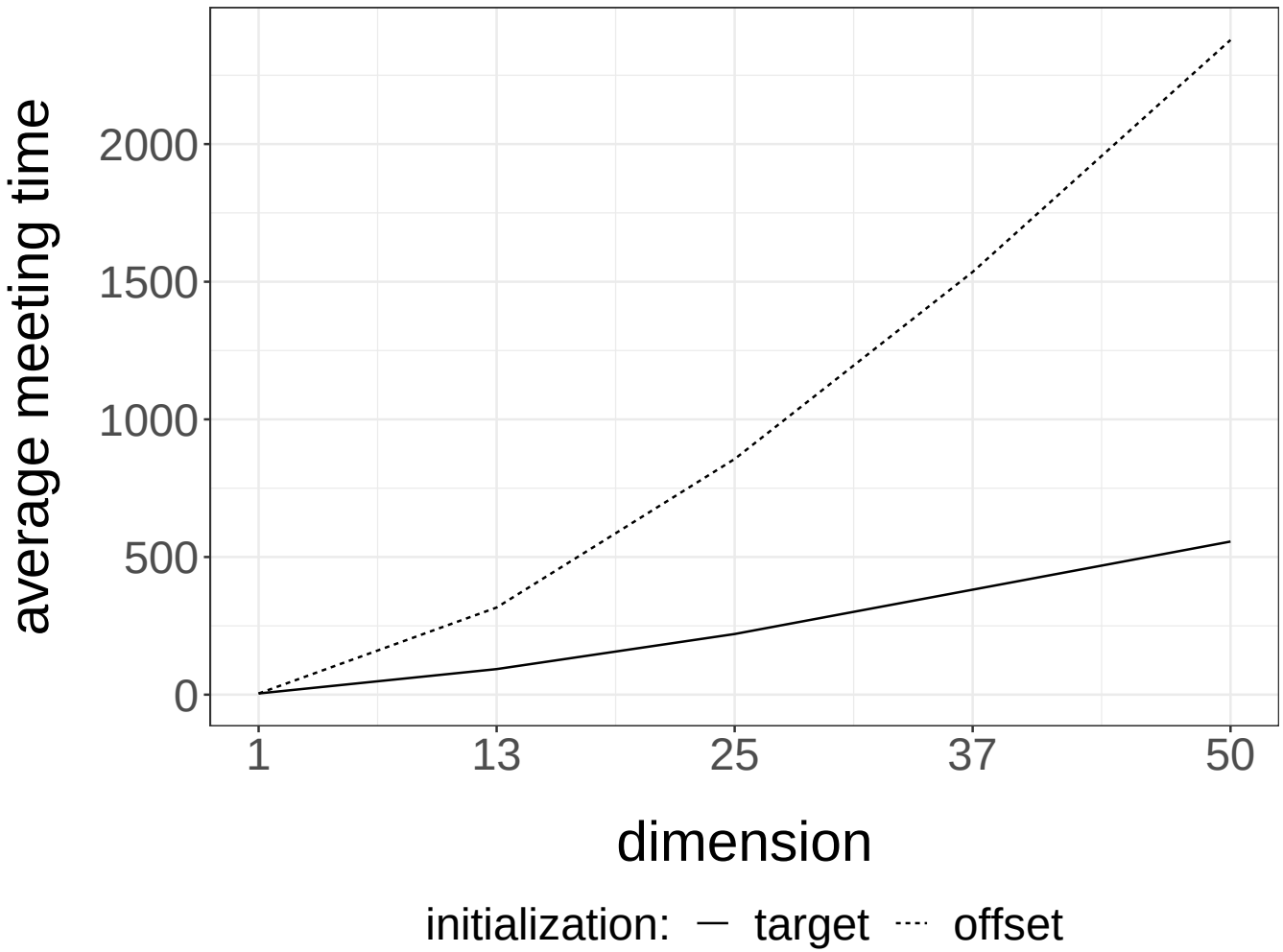
Scaling with dimension: Metropolis–Hastings

With independent-maximal couplings:



Scaling with dimension: Metropolis–Hastings

With reflection-maximal couplings:



Coupling of Gibbs

Gibbs sampler: update component i of the chain,
leaving $\pi(dx^i|x^1, \ldots, x^{i-1}, x^{i+1}, \ldots, x^d)$ invariant.

For instance, we can propose $X^\star \sim k(X_t^i, \cdot)$ to replace X_t^i,
and accept or not with a Metropolis–Hastings step.

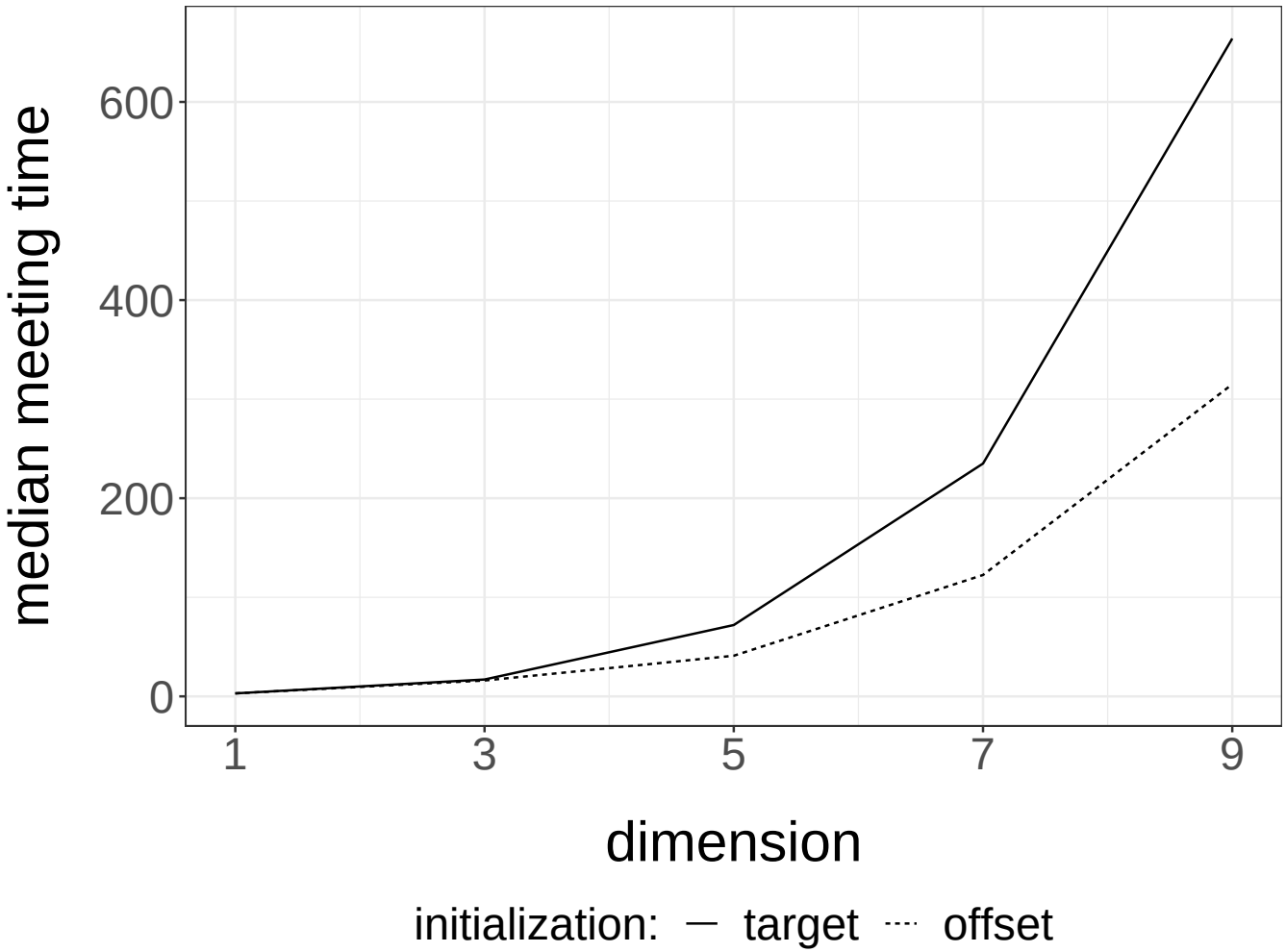
These proposals can be coupled across two chains,
during each component's update.

The chains meet when all components have met.

Likewise we can couple "ensemble" of chains, as in parallel tempering, and meeting occurs when both ensembles of chains coincide.

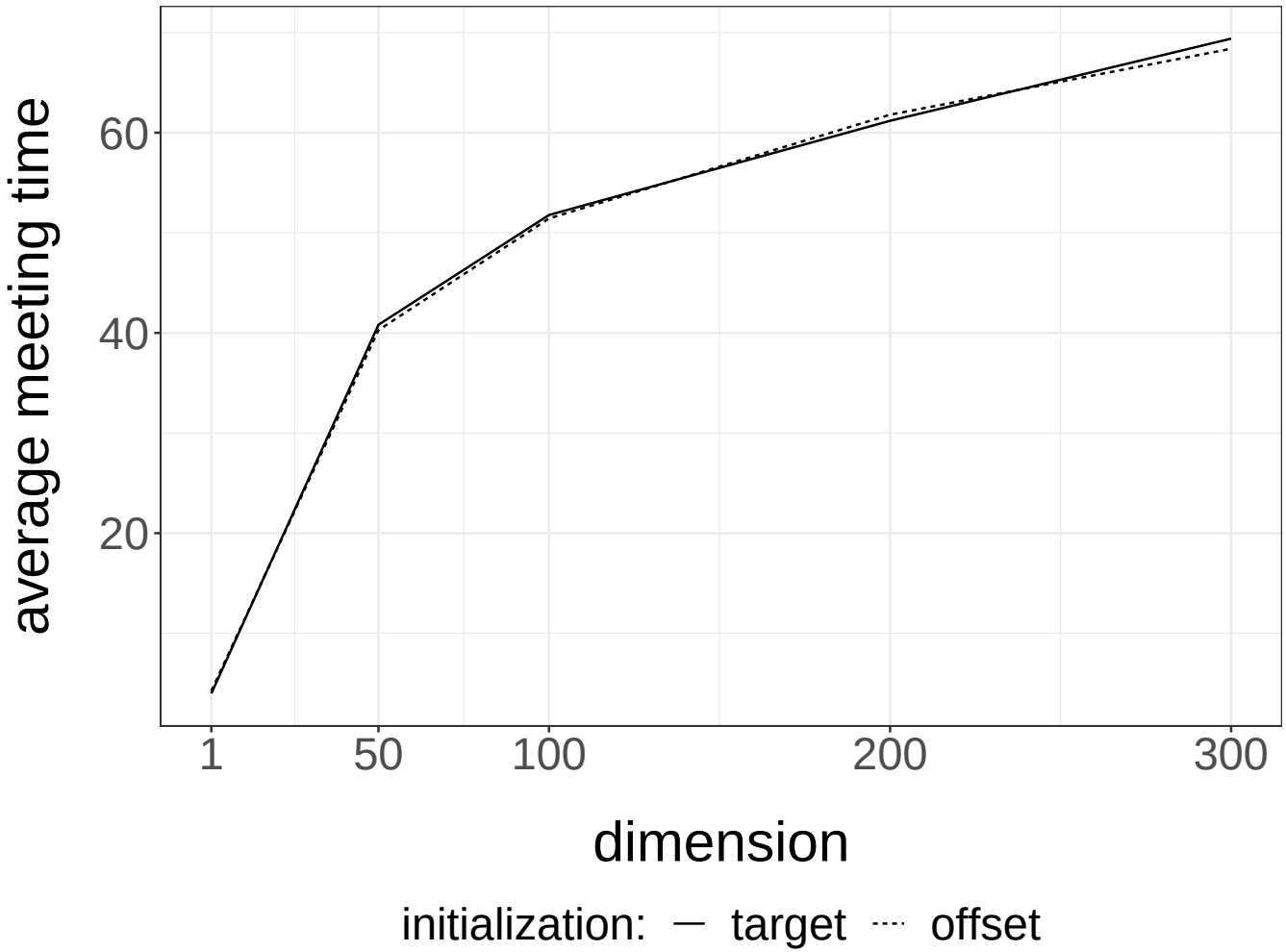
Scaling with dimension: Gibbs

When the target has high correlations (V inverse Wishart):



Scaling with dimension: Gibbs

When the target has low correlations ($V_{ij} = 0.5^{-|i-j|}$):



Hamiltonian Monte Carlo

Introduce potential energy $U(q) = -\log \pi(q)$,
and total energy $E(q,p) = U(q) + \frac{1}{2}|p|^2$.

Hamiltonian dynamics for $(q(s), p(s))$, where $s \geq 0$:

$$\frac{d}{ds} q(s) = \nabla_p E(q(s), p(s)),$$
$$\frac{d}{ds} p(s) = -\nabla_q E(q(s), p(s)).$$

Solving Hamiltonian dynamics exactly is not feasible,
so discretization + Metropolis–Hastings correction.

Common random numbers can make two HMC chains contract,
under assumptions on the target such as strong log-concavity.

Coupling of Hamiltonian Monte Carlo

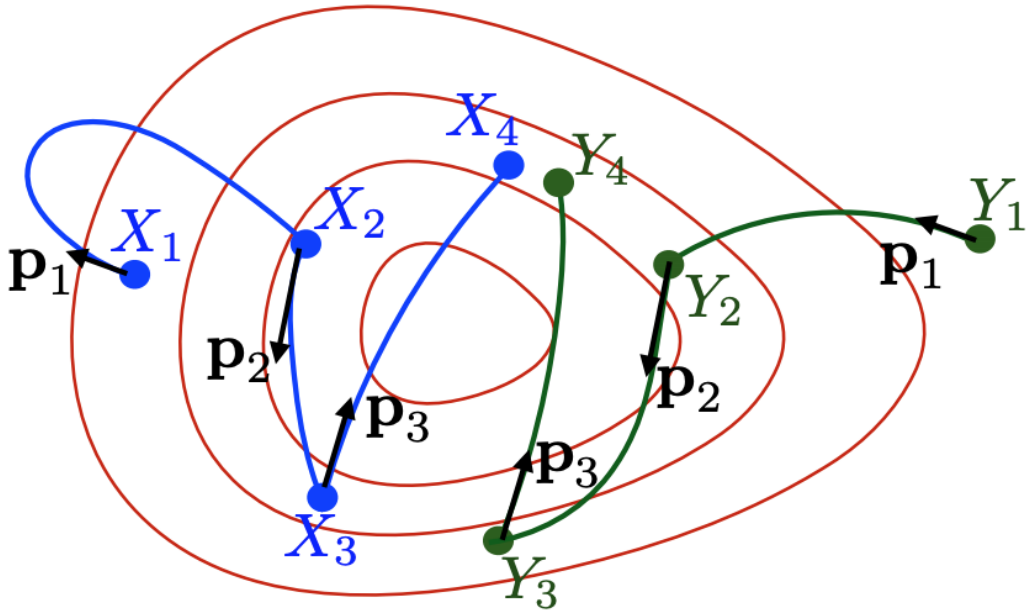
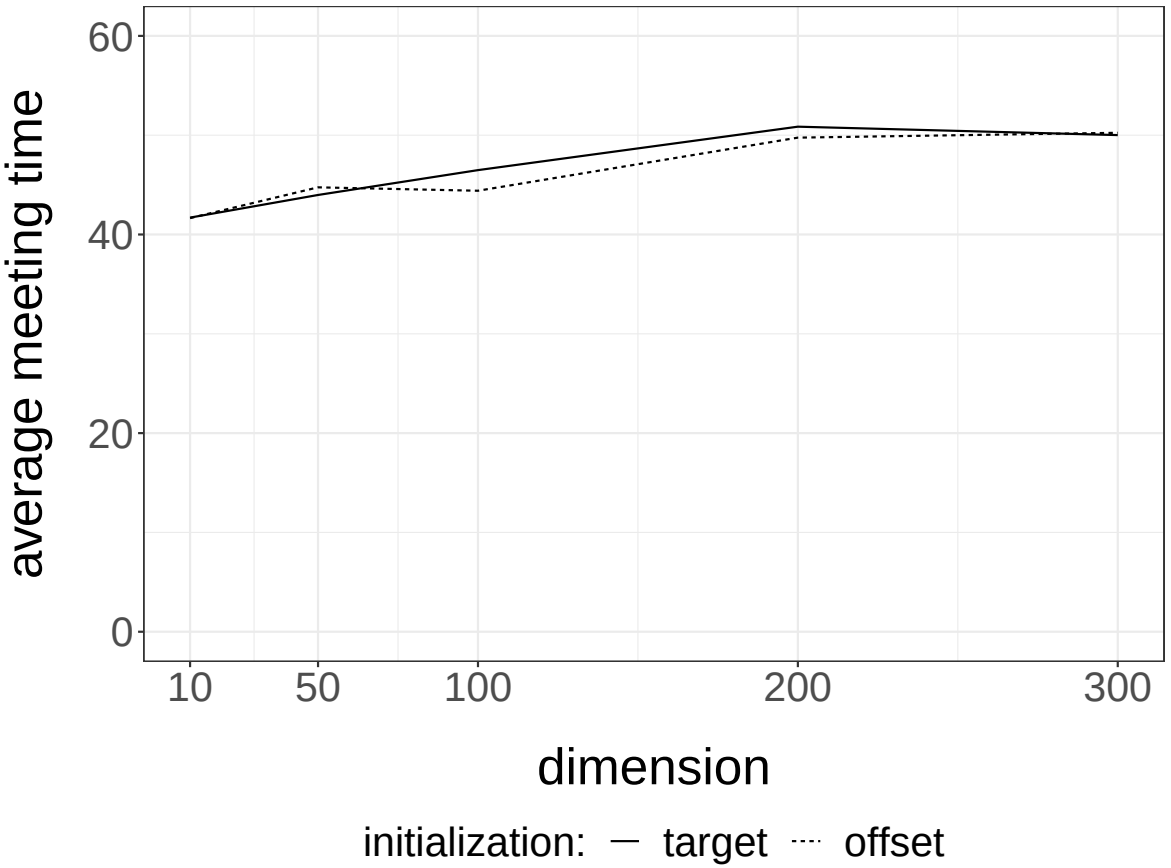


Figure 2 of Mangoubi & Smith, *Rapid mixing of HMC strongly log-concave distributions*, 2017.

Coupling two copies $X_1, X_2, \ldots$ (blue) and $Y_1, Y_2, \ldots$ (green) of HMC by choosing same momentum p_i at every step.

Scaling with dimension: Hamiltonian Monte Carlo



Mangoubi & Smith, 2017, Bou-Rabee, Eberle & Zimmer, 2018.

Heng & Jacob, 2019.

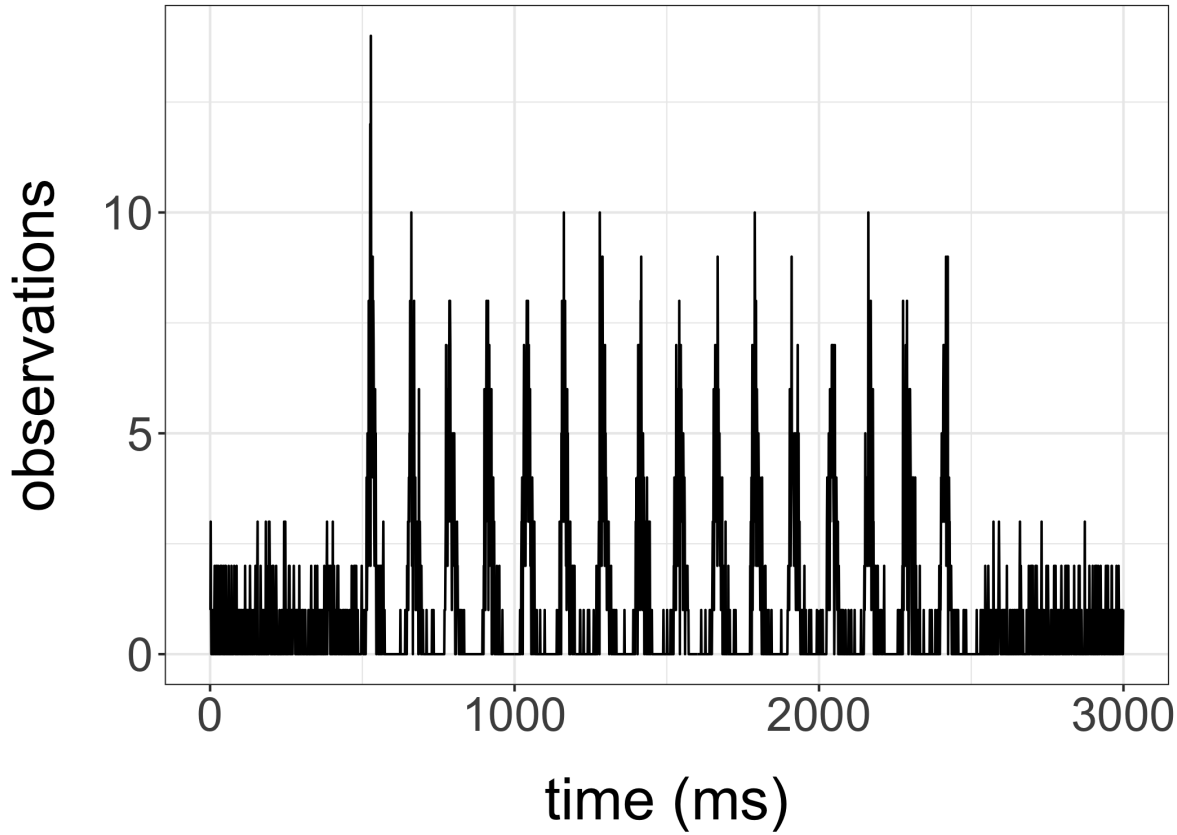
Whisker experiment

Activation of neurons of rats, as their whiskers are being moved with a periodic stimulus.

The activation of a neuron is recorded as a binary variable for each time and each experiment. These activation variables are then aggregated by summing over the M experiments at each time step.

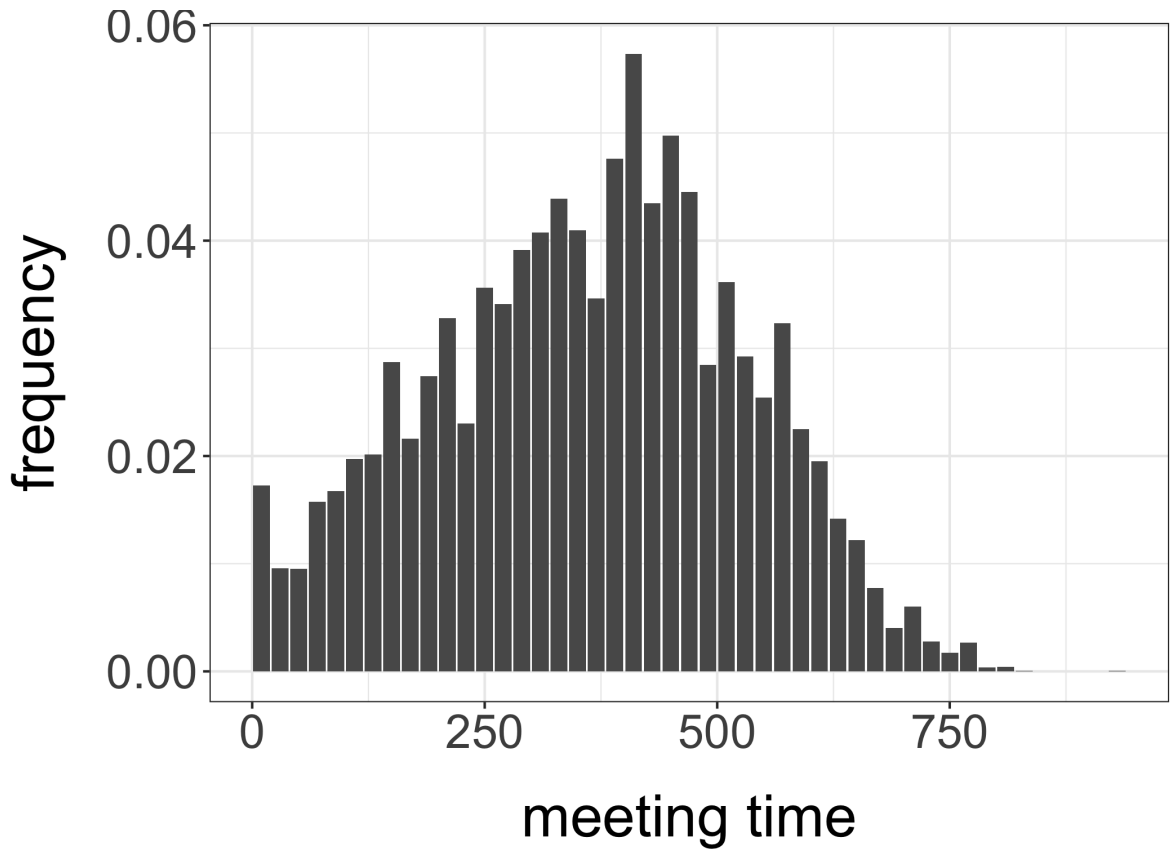
Data kindly shared by Prof. Demba Ba (Harvard SEAS).

Whisker experiment: data



$$\text{Model: } X_0 \sim \mathcal{N}(0,1), \quad X_t|X_{t-1} \sim \mathcal{N}(aX_{t-1}, \sigma_X^2),$$
$$Y_t|X_t \sim \text{Binomial}(M, (1+\exp(-X_t))^{-1}).$$

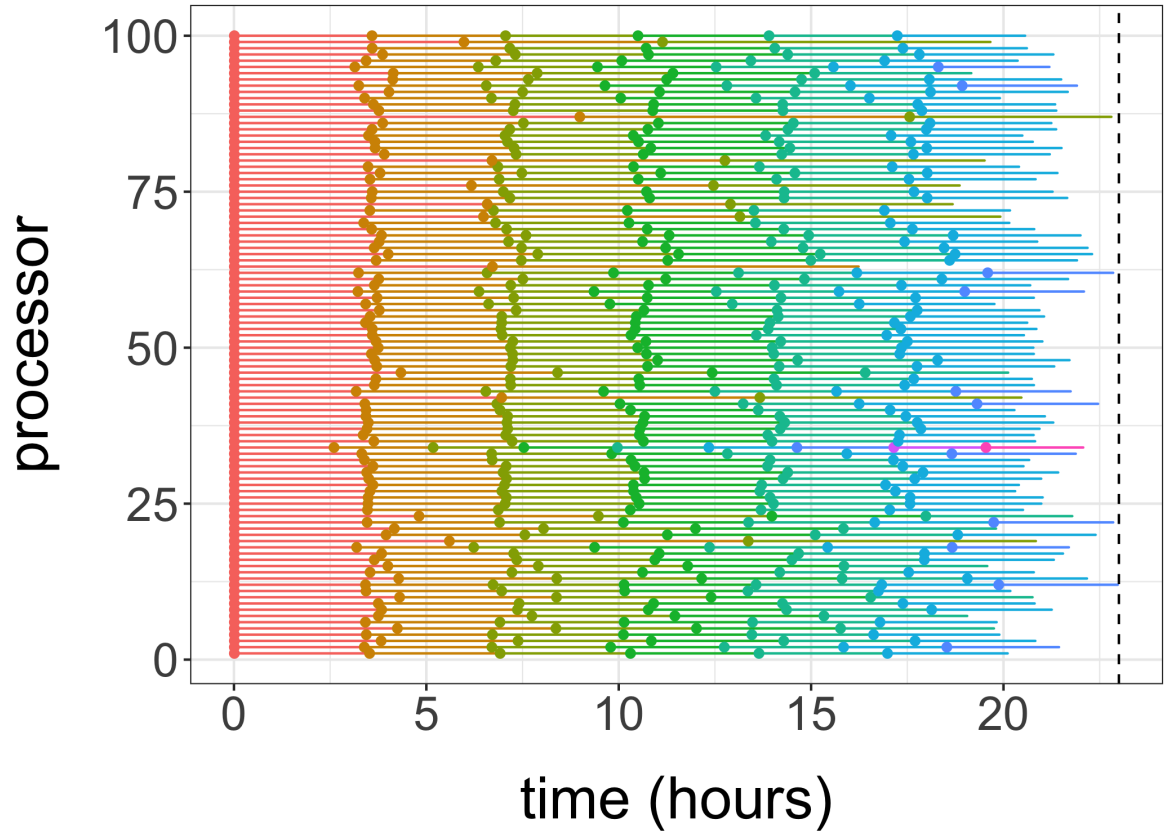
Whisker experiment: meeting times



Middleton, Deligiannidis, Doucet, Jacob, *Unbiased MCMC for intractable target distributions*, 2020.

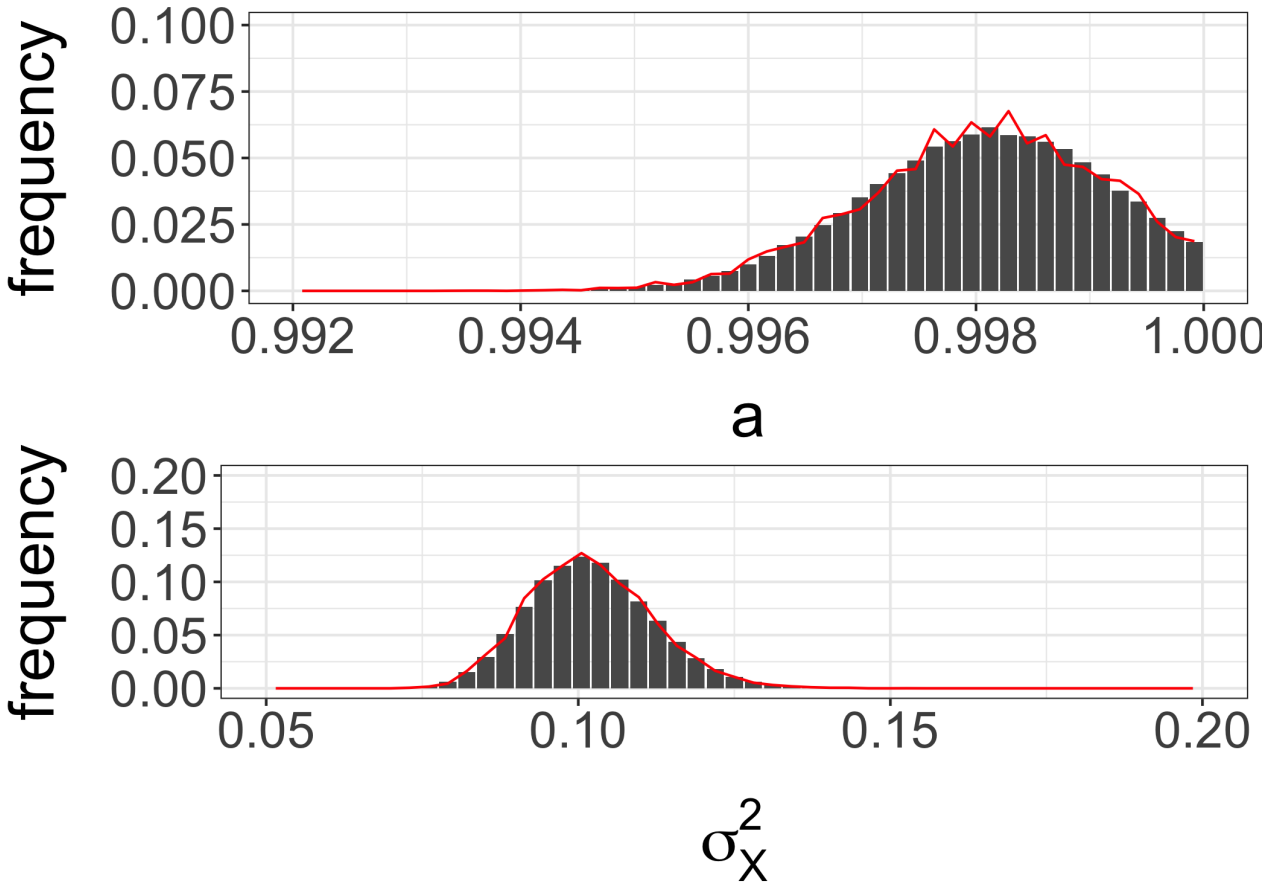
Heng, Bishop, Deligiannidis & Doucet, *Controlled SMC*, 2020.

Whisker experiment: in parallel



Computation chronology, on 100 processors over 23 hours, with $k = 1,000$, $m = 10k$.

Whisker experiment: results



Approximations of marginal posteriors, $k = 1,000$, $m = 10k$.
Red line corresponds to PMMH run (250,000 iterations).

Variable selection

Integers p, n, with possibly $n < p$.

$X_1, \ldots, X_p \in \mathbb{R}^n$: covariates.
$Y \in \mathbb{R}^n$: response variable.

$\gamma \in \{0, 1\}^p$ represents which covariates to select.

$|\gamma| = \sum_{i=1}^{p} \gamma_i$: number of selected variables.

Variable selection

X_γ is the $n \times |\gamma|$ matrix of covariates selected by γ.

Model:

$$Y = X_\gamma \beta_\gamma + w, \quad \text{where} \quad w \sim \mathcal{N}(0, \sigma^2 I_n).$$

Conjugate prior leads to

$$\pi(Y|X,\gamma) \propto \frac{(1+g)^{-|\gamma|/2}}{(1+g(1-R^2_\gamma))^{n/2}}, \text{ where } R^2_\gamma = \frac{Y' X_\gamma (X'_\gamma X_\gamma)^{-1} X'_\gamma Y}{Y'Y}.$$

The prior on γ is set as $\pi(\gamma) \propto p^{-\kappa|\gamma|} \mathbb{1}(|\gamma| \leq s_0)$.

Yang, Wainwright & Jordan, *On the computational complexity of high-dimensional Bayesian variable selection*, 2016.

Variable selection

MCMC algorithm randomly alternates between

- flipping a component, from 1 to 0 or from 0 to 1, accept or not with MH step;
- randomly swapping a 0 and an 1, accept or not with MH step.

Maximal couplings of these proposals can be implemented.

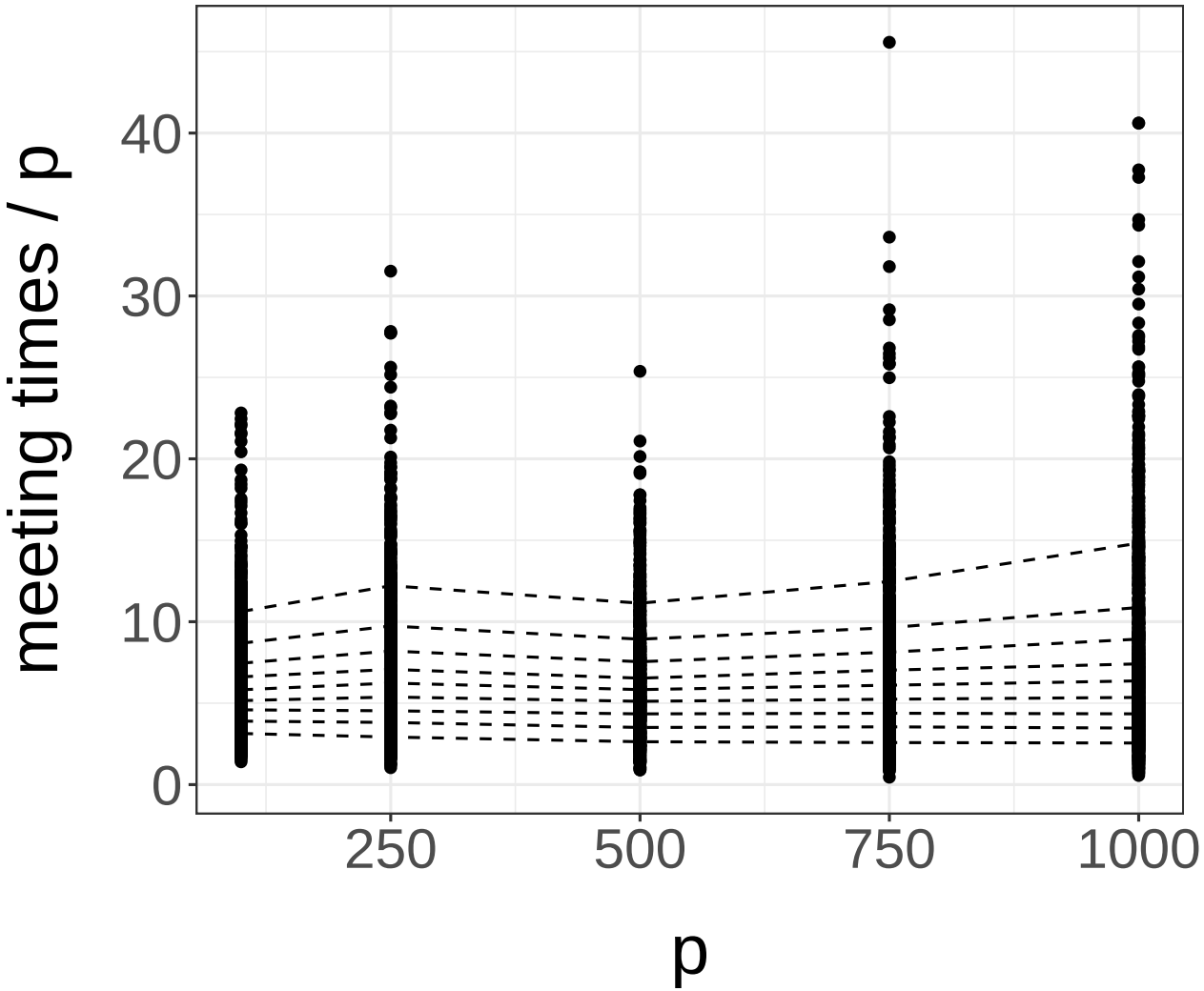
Variable selection

Generate $n = 500$ rows, with covariates X
generated as independent standard Normals.

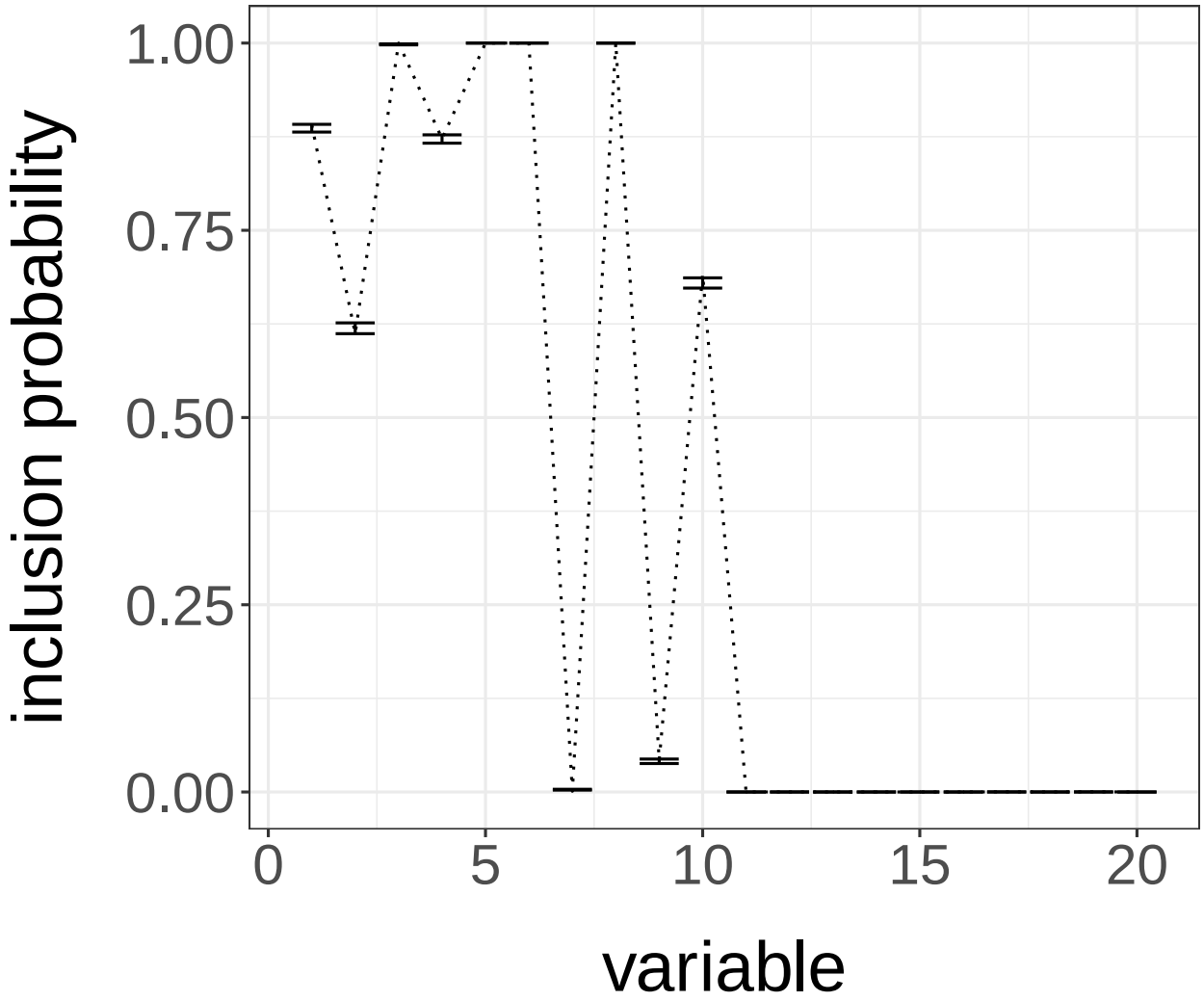
Outcome variable Y generated from the model, with a
coefficient $\beta^\star$ such that first ten entries are non-zero, and $\sigma = 1$.

Other parameters: $s_0 = 100$, $g = p^3$, $\kappa = 0.1$.

Variable selection



Variable selection



Outline

Assessing convergence of MCMC

Total variation distance between X_t and π:

$$\|\mathcal{L}(X_t) - \pi\|_{\mathrm{TV}} = \frac{1}{2} \sup_{h:|h|\leq 1} |\mathbb{E}[h(X_t)] - \mathbb{E}_\pi[h(X)]|$$

Assessing convergence of MCMC

Total variation distance between X_t and π:

$$\|\mathcal{L}(X_t) - \pi\|_{\text{TV}} = \frac{1}{2} \sup_{h:|h|\leq 1} |\mathbb{E}[h(X_t)] - \mathbb{E}_\pi[h(X)]|$$

For any test function h with $|h| \leq 1$,

$$\frac{1}{2}|\mathbb{E}[\sum_{s=t+1}^{\tau-1} h(X_s) - h(Y_{s-1})]| \leq \frac{1}{2}\mathbb{E}[\sum_{s=t+1}^{\tau-1} |h(X_s) - h(Y_{s-1})|]$$
$$= \mathbb{E}[\max(0, \tau - t - 1)],$$

using triangle inequalities and $|h(x) - h(y)| \leq 2$.

Assessing convergence of MCMC

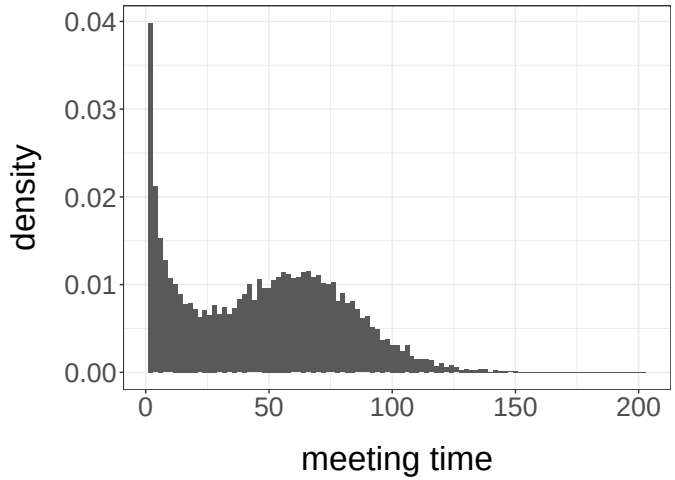
Total variation distance between X_t and π:

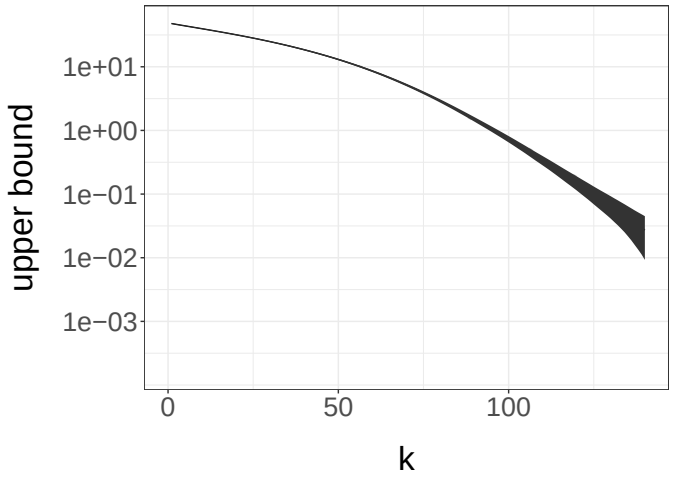
$$\|\mathcal{L}(X_t) - \pi\|_{\mathrm{TV}} = \frac{1}{2} \sup_{h:|h|\leq 1} |\mathbb{E}[h(X_t)] - \mathbb{E}_\pi[h(X)]|$$

For any test function h with $|h| \leq 1$,

$$\frac{1}{2}|\mathbb{E}[\sum_{s=t+1}^{\tau-1} h(X_s) - h(Y_{s-1})]| \leq \frac{1}{2}\mathbb{E}[\sum_{s=t+1}^{\tau-1} |h(X_s) - h(Y_{s-1})|]$$
$$= \mathbb{E}[\max(0, \tau - t - 1)],$$

using triangle inequalities and $|h(x) - h(y)| \leq 2$.





Assessing convergence of MCMC

With L-lag couplings, $\tau^{(L)} = \inf\{t \geq L : X_t = Y_{t-L}\}$,

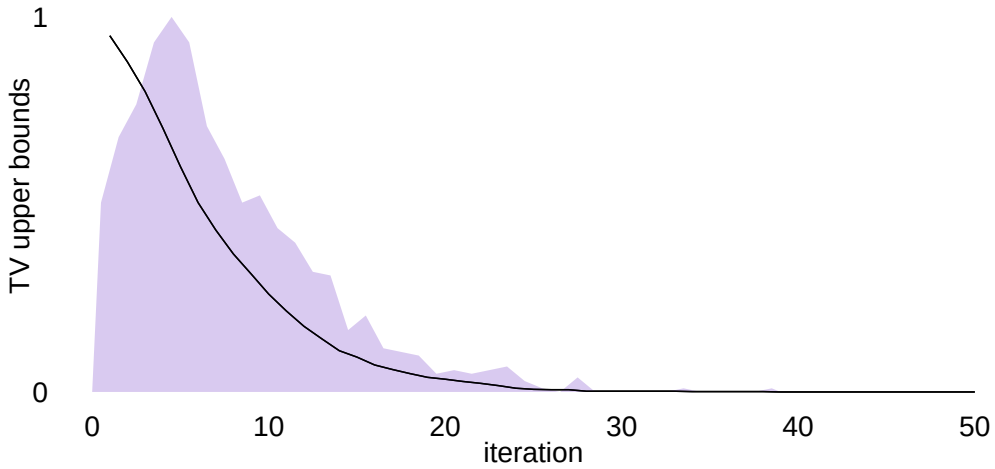
$$\|\mathcal{L}(X_t) - \pi\|_{\mathrm{TV}} \leq \mathbb{E}\Big[\max(0, \lceil(\tau^{(L)} - t - L)/L\rceil)\Big].$$

Assessing convergence of MCMC

With L-lag couplings, $\tau^{(L)} = \inf\{t \geq L : X_t = Y_{t-L}\}$,

$$\|\mathcal{L}(X_t) - \pi\|_{\text{TV}} \leq \mathbb{E}\Big[\max(0, \lceil(\tau^{(L)} - t - L)/L\rceil)\Big].$$

For 3-subsets out of 10 elements,



Biswas, Jacob & Vanetti, *Estimating Convergence of Markov chains with L-Lag Couplings*, 2019.
See also comment by Vanetti & Doucet in discussion of
Jacob, O'Leary, Atchadé, *Unbiased MCMC with couplings*, 2020.

Outline

Bimodal target

Target is mixture of univariate Normal distributions:

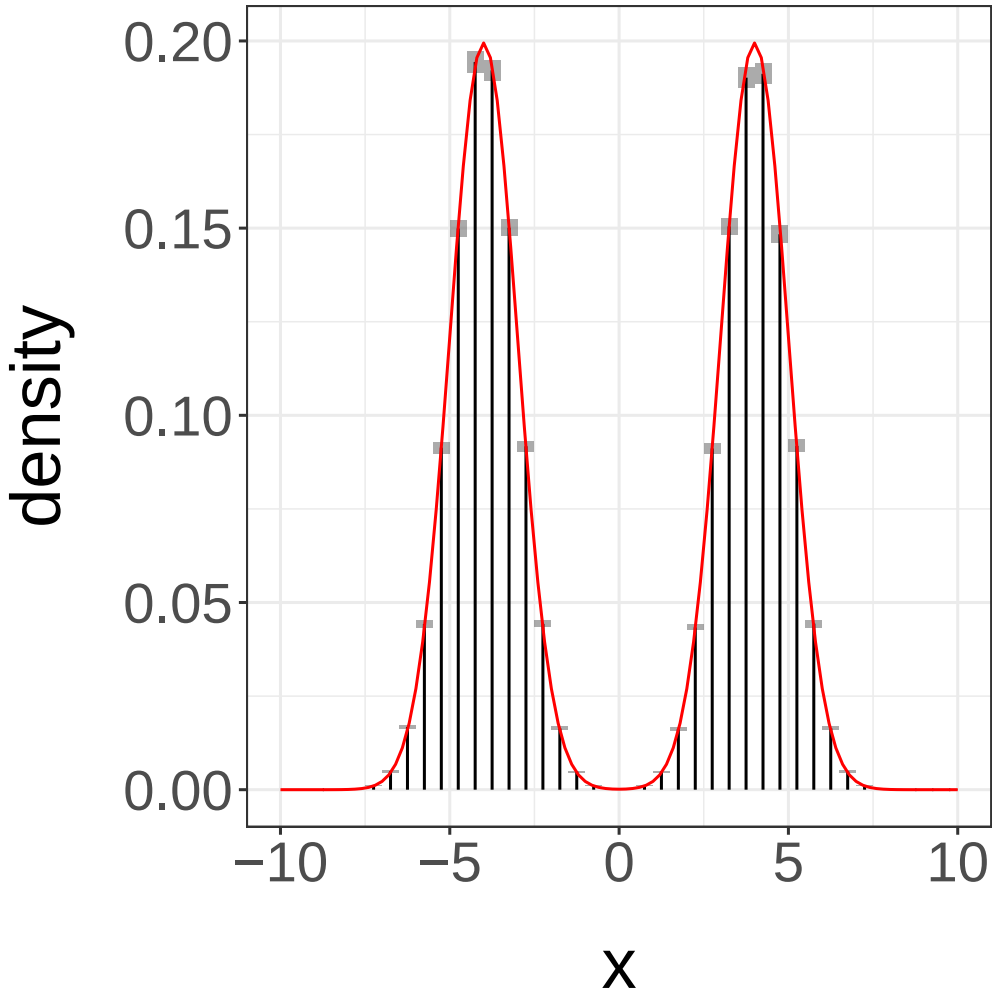
$$\pi = 0.5 \cdot \mathcal{N}(-4, 1) + 0.5 \cdot \mathcal{N}(+4, 1).$$

MCMC: random walk Metropolis–Hastings,
with proposal standard deviation σ that will vary.

Initial distribution π_0 will vary.

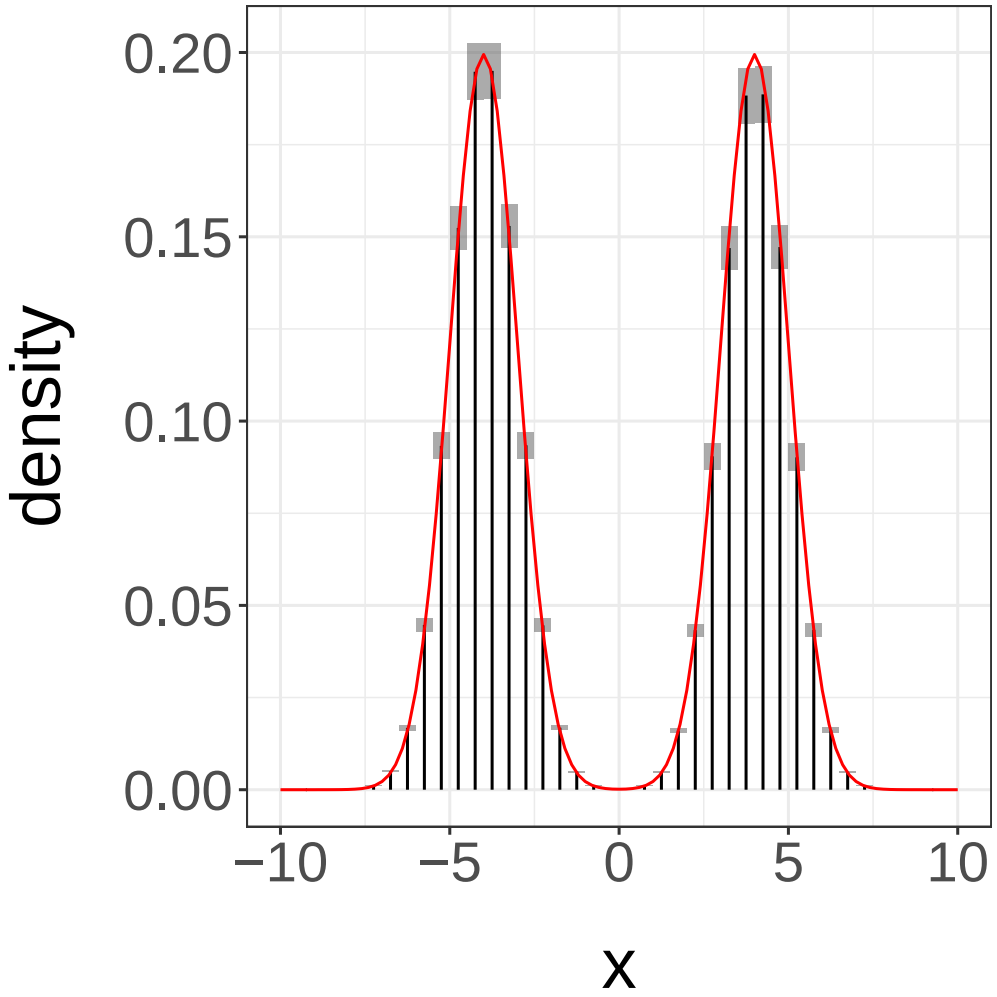
Bimodal target

With $\sigma = 3$, $\pi_0 = \mathcal{N}(10, 10^2) \ldots$



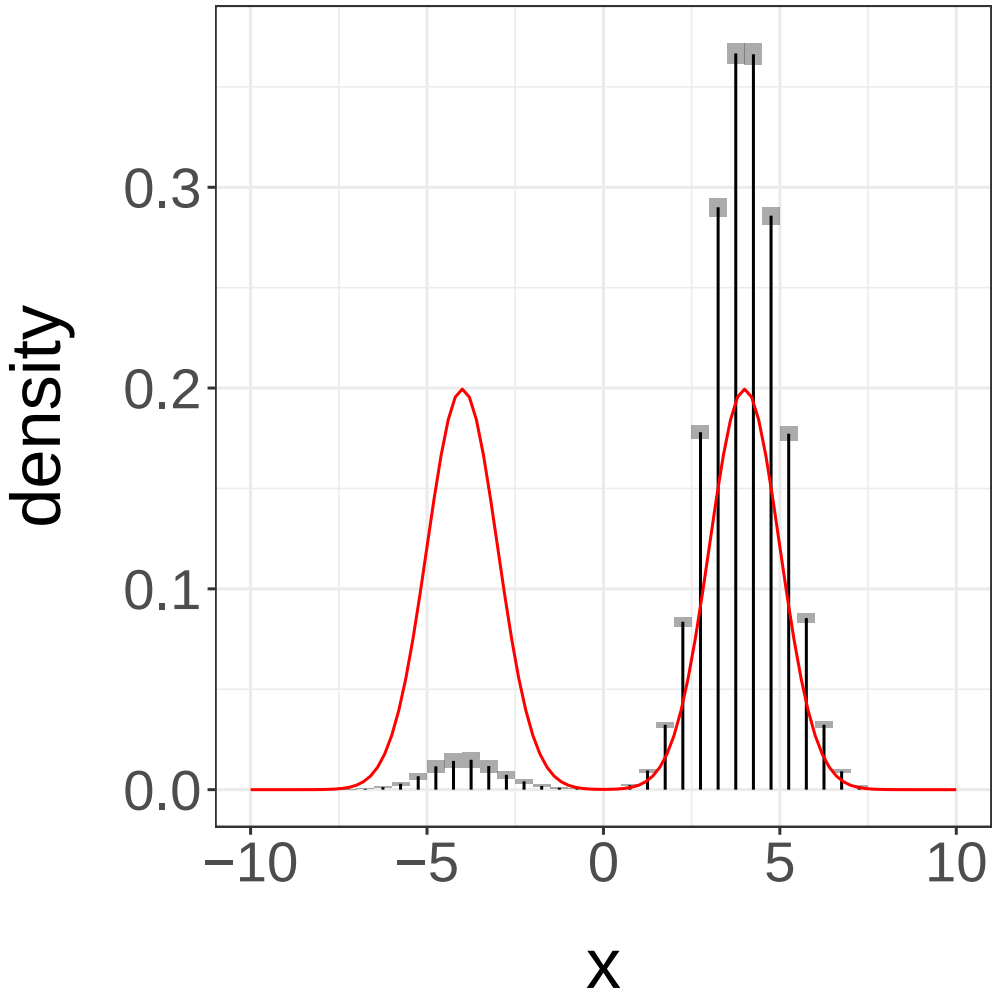
Bimodal target

With $\sigma = 1$, $\pi_0 = \mathcal{N}(10, 10^2) \ldots$



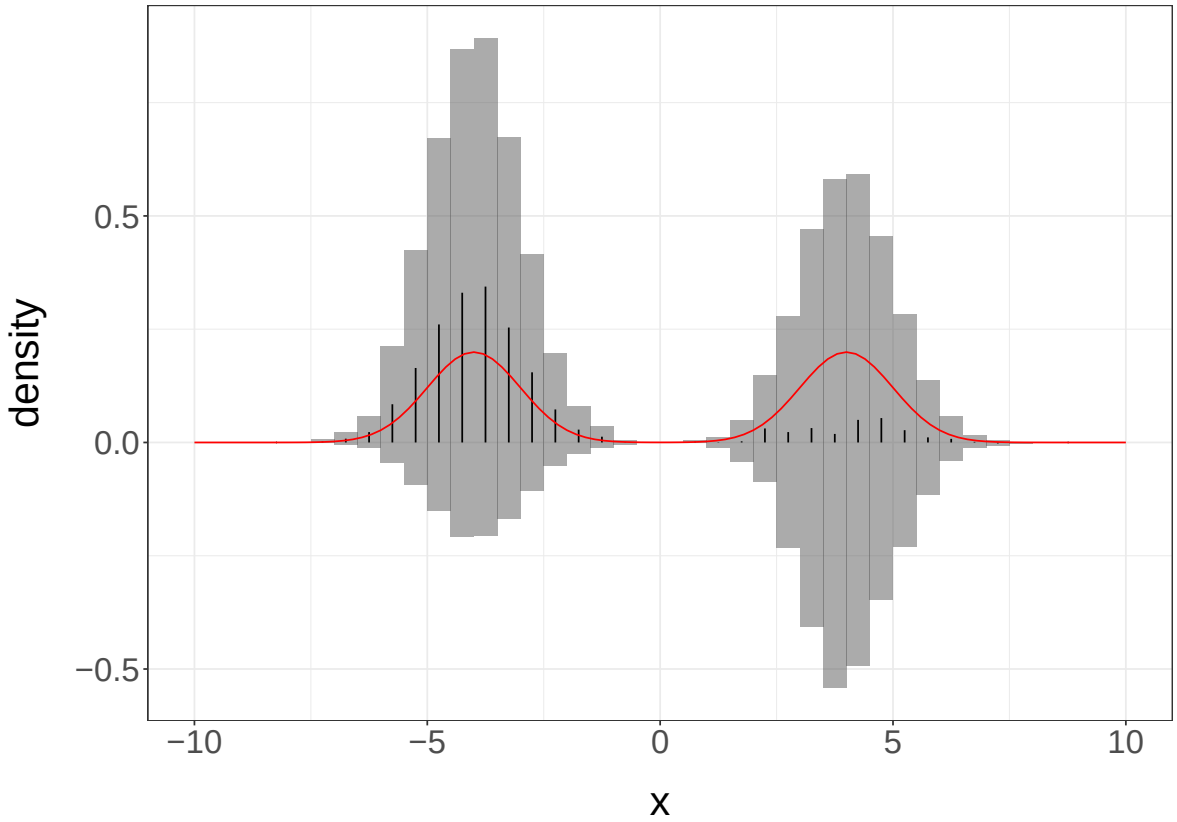
Bimodal target

With $\sigma = 1$, $\pi_0 = \mathcal{N}(10, 1^2) \ldots 1,000$ repeats.



Bimodal target

With $\sigma = 1$, $\pi_0 = \mathcal{N}(10, 1^2) \ldots 10,000$ repeats.



Limitations of Markovian couplings

There are known cases where Markovian couplings cannot yield meeting times that are commensurate with mixing time.

For example, "swapping" algorithm to sample permutations of $\{1, \ldots, n\}$. Markovian couplings result in meeting times of order $n^2 \log n$ while mixing happens in $n \log n$.

Burdzy & Kendall, *Efficient Markovian couplings: examples and counterexamples*, 2000.
Banerjee & Kendall, *Rigidity for Markovian maximal couplings of elliptic diffusions*, 2017.

Discussion

- The proposed technique delivers unbiased estimators and diagnostics of convergence in a variety of settings.
  Couplings are algorithmic-specific and not target-specific; alternative to regeneration methods,

  e.g. Brockwell & Kadane, 2005.

Discussion

- The proposed technique delivers unbiased estimators and diagnostics of convergence in a variety of settings.
  Couplings are algorithmic-specific and not target-specific; alternative to regeneration methods,

  e.g. Brockwell & Kadane, 2005.

- Choice of parameters k, m (and lag L) impacts efficiency.
  How should we choose them?

Discussion

- The proposed technique delivers unbiased estimators and diagnostics of convergence in a variety of settings.
  Couplings are algorithmic-specific and not target-specific; alternative to regeneration methods,

  e.g. Brockwell & Kadane, 2005.

- Choice of parameters k, m (and lag L) impacts efficiency.
  How should we choose them?

- If underlying MCMC algorithm is not well-suited to the target, meeting times will be long, no matter what.

Discussion

- The proposed technique delivers unbiased estimators and diagnostics of convergence in a variety of settings.
  Couplings are algorithmic-specific and not target-specific; alternative to regeneration methods,

  e.g. Brockwell & Kadane, 2005.

- Choice of parameters k, m (and lag L) impacts efficiency. How should we choose them?

- If underlying MCMC algorithm is not well-suited to the target, meeting times will be long, no matter what.

- Perfect samplers, designed to sample i.i.d. from π, would provide same benefits and more.

Thanks for listening!

Main reference for this talk:

Jacob, O'Leary, Atchadé, *Unbiased MCMC with couplings*, 2020.

See the germinal work of ...

McLeish, *A general method for debiasing a Monte Carlo estimator*, 2010.

Glynn & Rhee, *Exact estimation for Markov chain equilibrium*, 2014.

Funding provided by NSF grants DMS-1712872, DMS-1844695.

Code on Github: github.com/pierrejacob/
Blog: statisfaction.wordpress.com/
Material on unbiased MCMC collected by Darren Wilkinson:
github.com/darrenjw/unbiased-mcmc