Some methods based on couplings of Markov chain Monte Carlo algorithms

Pierre E. Jacob

ECODEP
April 27, 2022

Outline

1 Introduction

2 Couplings

3 Unbiased estimation of target expectations

4 Diagnostics of convergence

5 Asymptotic variance estimation

Outline

1 Introduction

2 Couplings

3 Unbiased estimation of target expectations

4 Diagnostics of convergence

5 Asymptotic variance estimation

Setting & some questions

Target probability distribution π.
MCMC: $X_0 \sim \pi_0$, then $X_t | X_{t-1} \sim P(X_{t-1}, \cdot)$ for $t = 1, 2, \ldots$

Setting & some questions

Target probability distribution π.
MCMC: $X_0 \sim \pi_0$, then $X_t | X_{t-1} \sim P(X_{t-1}, \cdot)$ for $t = 1, 2, \ldots$

Convergence of marginals:

$$|\pi_t - \pi| \to 0.$$

Setting & some questions

Target probability distribution π.
MCMC: $X_0 \sim \pi_0$, then $X_t | X_{t-1} \sim P(X_{t-1}, \cdot)$ for $t = 1, 2, \ldots$

Convergence of marginals:

$$|\pi_t - \pi| \to 0.$$

Central limit theorem:

$$\sqrt{t} \left(t^{-1} \sum_{s=0}^{t-1} h(X_s) - \pi(h) \right) \to \mathcal{N}(0, v(P, h)).$$

Setting & some questions

Target probability distribution π.
MCMC: $X_0 \sim \pi_0$, then $X_t | X_{t-1} \sim P(X_{t-1}, \cdot)$ for $t = 1, 2, \ldots$

Convergence of marginals:

$$|\pi_t - \pi| \to 0.$$

Central limit theorem:

$$\sqrt{t}\left(t^{-1}\sum_{s=0}^{t-1} h(X_s) - \pi(h)\right) \to \mathcal{N}(0, v(P, h)).$$

How to choose t such that the error is small?
How to reduce the error with parallel computers?

Example

Prior $\theta \sim \text{Normal}(0, \sigma^2)$, on the location θ of $\text{Cauchy}(\theta, 1)$ observations $x_1, \ldots, x_n$.

Example

Prior $\theta \sim \text{Normal}(0, \sigma^2)$, on the location θ of $\text{Cauchy}(\theta, 1)$
observations $x_1, \ldots, x_n$.
Posterior:

$$\pi(\theta | x_1, \ldots, x_n) \propto \exp(-\theta^2/2\sigma^2) \prod_{i=1}^{n} \{1 + (\theta - x_i)^2\}^{-1}$$

$$\propto \exp(-\theta^2/2\sigma^2) \prod_{i=1}^{n} \int \exp(\{1 + (\theta - x_i)^2\}\eta_i/2) d\eta_i.$$

Example

Prior $\theta \sim \text{Normal}(0, \sigma^2)$, on the location θ of $\text{Cauchy}(\theta, 1)$
observations $x_1, \ldots, x_n$.
Posterior:

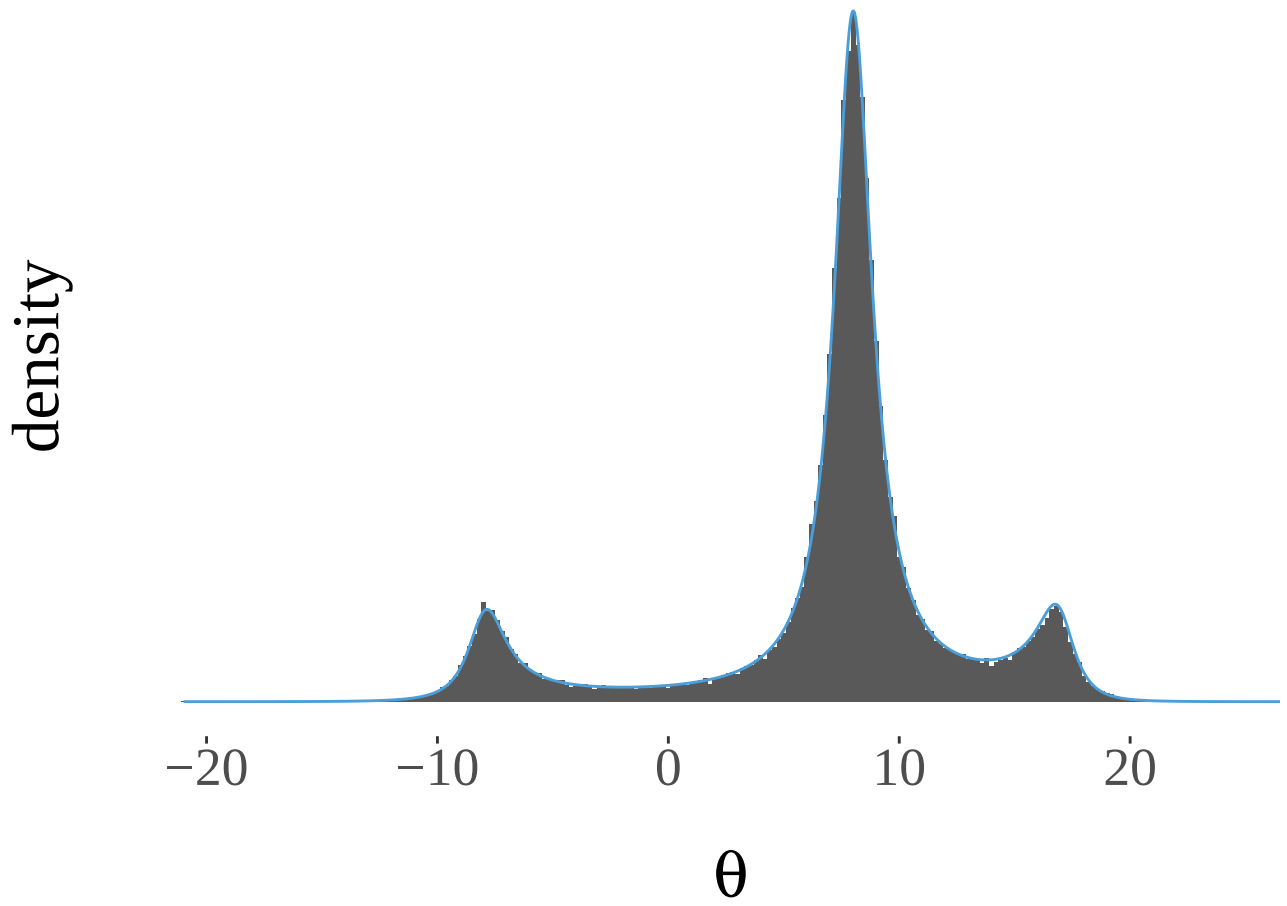
$$\pi(\theta | x_1, \ldots, x_n) \propto \exp(-\theta^2/2\sigma^2) \prod_{i=1}^{n} \{1 + (\theta - x_i)^2\}^{-1}$$
$$\propto \exp(-\theta^2/2\sigma^2) \prod_{i=1}^{n} \int \exp(\{1 + (\theta - x_i)^2\}\eta_i/2) d\eta_i.$$

Gibbs sampler:

$$\eta_i | \theta \sim \text{Exponential}\left(\frac{1 + (\theta - x_i)^2}{2}\right) \quad \forall i = 1, \ldots, n$$
$$\theta' | \eta_1, \ldots, \eta_n \sim \text{Normal}\left(\frac{\sum_{i=1}^n \eta_i x_i}{\sum_{i=1}^n \eta_i + \sigma^{-2}}, \frac{1}{\sum_{i=1}^n \eta_i + \sigma^{-2}}\right).$$

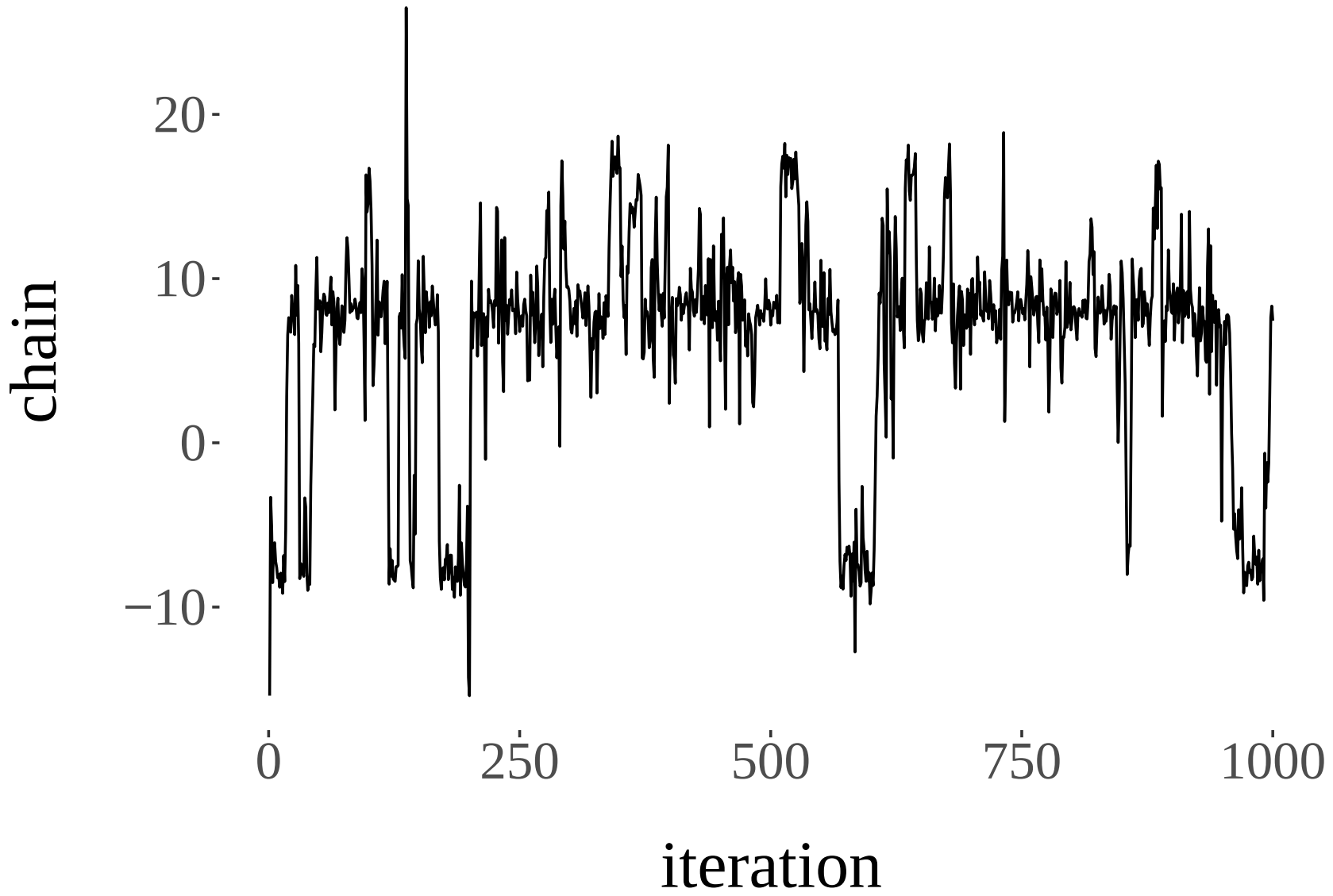
Initial distribution: $\pi_0 = \text{Normal}(0, 1)$.

Example



Example taken from "Convergence control methods for Markov chain Monte Carlo algorithms", Christian P. Robert, 1995.

Example



MCMC to approximate integrals

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.

MCMC to approximate integrals

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.

Among them, Monte Carlo methods provide state-of-the-art performance in high dimensions.

Outline

1 Introduction

2 Couplings

3 Unbiased estimation of target expectations

4 Diagnostics of convergence

5 Asymptotic variance estimation

Couplings

Technique to study the convergence of Markov chains.

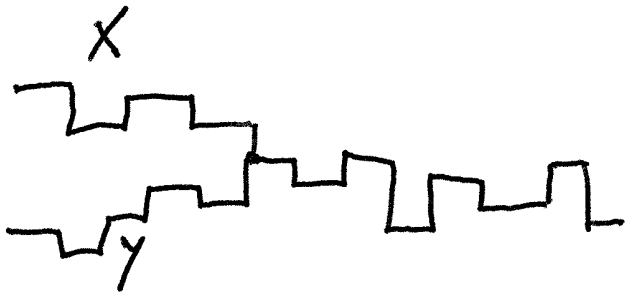
Construct a joint process (X_t, Y_t) such that $Y_t \sim \pi$ for all $t \geq 0$, and marginally both chains evolve according to same kernel P.

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 $t \geq 0$, and marginally both chains evolve according to same kernel P.

Suppose that there exists τ a random variable such that $X_t = Y_t$ for all $t \geq \tau$.

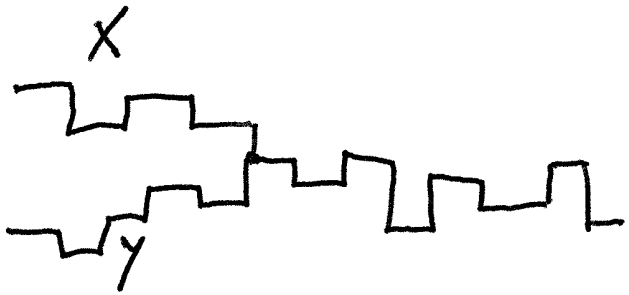


Couplings

Technique to study the convergence of Markov chains.

Construct a joint process (X_t, Y_t) such that $Y_t \sim \pi$ for all $t \geq 0$, and marginally both chains evolve according to same kernel P.

Suppose that there exists τ a random variable such that $X_t = Y_t$ for all $t \geq \tau$.



Then

$$\begin{aligned} \|\pi_t - \pi\|_{\mathrm{TV}} &= \|\mathcal{L}(X_t) - \mathcal{L}(Y_t)\|_{\mathrm{TV}} \\ &\leq \mathbb{P}(X_t \neq Y_t) = \mathbb{P}(\tau > t), \end{aligned}$$

where $\|\cdot\|_{\mathrm{TV}}$ is the total variation distance.

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

Couplings

Coupling techniques have proved very successful, in some cases giving precise rates of convergence.

See for example

Jerrum, *Mathematical foundations of the MCMC method*, 1998.

Couplings

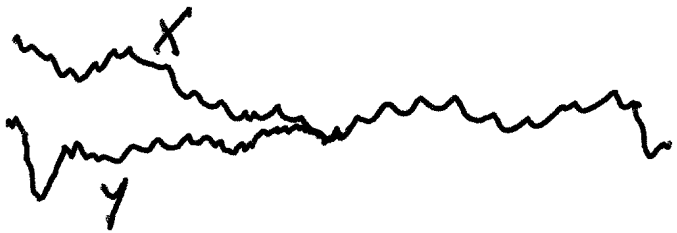
Coupling techniques have proved very successful, in some cases giving precise rates of convergence.

See for example

Jerrum, *Mathematical foundations of the MCMC method*, 1998.

Coupling techniques provide
bounds on other metrics than TV,

$$\|\pi_t - \pi\|_{W_1} = \inf_{X,Y \sim \gamma \in \Gamma(\pi_t, \pi)} \mathbb{E}_\gamma[d(X,Y)]$$
$$\leq \mathbb{E}[d(X_t, Y_t)].$$



Couplings

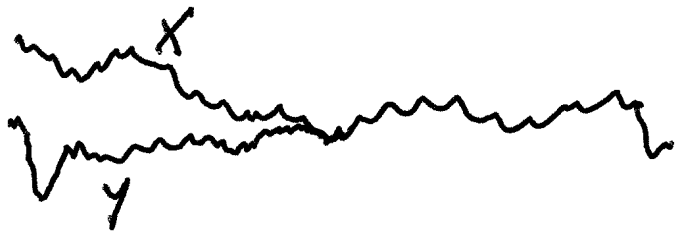
Coupling techniques have proved very successful, in some cases giving precise rates of convergence.

See for example

Jerrum, *Mathematical foundations of the MCMC method*, 1998.

Coupling techniques provide
bounds on other metrics than TV,

$$\|\pi_t - \pi\|_{W_1} = \inf_{X,Y \sim \gamma \in \Gamma(\pi_t, \pi)} \mathbb{E}_\gamma[d(X,Y)]$$
$$\leq \mathbb{E}[d(X_t, Y_t)].$$



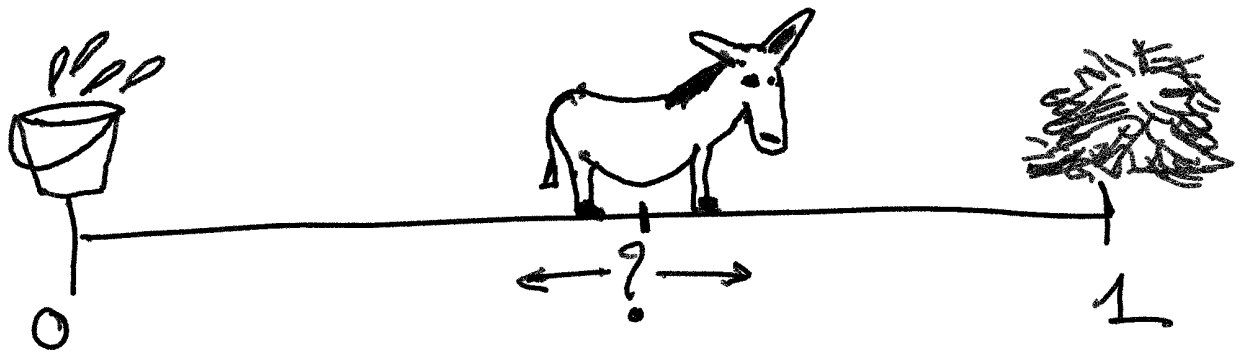
All of this appears theoretical, since we cannot sample $Y_0 \sim \pi$.

Example: donkey walk

Consider the chain (Z_t) on $[0, 1]$ with recursion

$$Z_t = B_{t,1}(1 - B_{t,0})Z_{t-1} + B_{t,0},$$

where $B_{t,1} \sim \text{Beta}(N_1, 1)$ and $B_{t,0} \sim \text{Beta}(1, N_0)$ are independent, and N_0, N_1 are positive integers.



Letac, *Donkey walk and Dirichlet distributions*, 2002.

Jacob, Gong, Edlefsen & Dempster, *A Gibbs sampler for a class of random convex polytopes*, 2021.

Example: donkey walk

A "common random numbers" coupling

$$Z_t = B_{t,1}(1 - B_{t,0})Z_{t-1} + B_{t,0}$$
$$\tilde{Z}_t = B_{t,1}(1 - B_{t,0})\tilde{Z}_{t-1} + B_{t,0},$$

leads to

$$\|\pi_t - \pi\|_{W_1} \leq \left(\frac{N_0}{N_0 + 1} \times \frac{N_1}{N_1 + 1}\right)^t \mathbb{E}\left[\left|Z_0 - \tilde{Z}_0\right|\right].$$

Example: donkey walk

A "common random numbers" coupling

$$Z_t = B_{t,1}(1 - B_{t,0})Z_{t-1} + B_{t,0}$$
$$\tilde{Z}_t = B_{t,1}(1 - B_{t,0})\tilde{Z}_{t-1} + B_{t,0},$$

leads to

$$\|\pi_t - \pi\|_{W_1} \leq \left(\frac{N_0}{N_0 + 1} \times \frac{N_1}{N_1 + 1}\right)^t \mathbb{E}\left[\left|Z_0 - \tilde{Z}_0\right|\right].$$

We can obtain a lower bound converging with the same rate (as pointed out by Guanyang Wang, Rutgers University).

Example: donkey walk

A "common random numbers" coupling

$$Z_t = B_{t,1}(1 - B_{t,0})Z_{t-1} + B_{t,0}$$
$$\tilde{Z}_t = B_{t,1}(1 - B_{t,0})\tilde{Z}_{t-1} + B_{t,0},$$

leads to

$$\|\pi_t - \pi\|_{W_1} \leq \left(\frac{N_0}{N_0 + 1} \times \frac{N_1}{N_1 + 1}\right)^t \mathbb{E}\left[\left|Z_0 - \tilde{Z}_0\right|\right].$$

We can obtain a lower bound converging with the same rate (as pointed out by Guanyang Wang, Rutgers University).

We obtain guidance on the choice of number of iterations t, but there are typically intractable constants in such analyses.

Example: conditional Bernoulli

Heng, Jacob & Ju, *A simple Markov chain for independent Bernoulli variables conditioned on their sum*, on arXiv.

Let $p = (p_1, \ldots, p_N) \in (0,1)^N$ and define $w_n = p_n/(1-p_n)$, the associated odds.

Example: conditional Bernoulli

Heng, Jacob & Ju, *A simple Markov chain for independent Bernoulli variables conditioned on their sum*, on arXiv.

Let $p = (p_1, \ldots, p_N) \in (0,1)^N$ and define $w_n = p_n/(1-p_n)$, the associated odds.

Let $X = (X_1, \ldots, X_N) \in \{0,1\}^N$ such that $X_n \sim \text{Bernoulli}(p_n)$, independently.

Example: conditional Bernoulli

Heng, Jacob & Ju, *A simple Markov chain for independent Bernoulli variables conditioned on their sum*, on arXiv.

Let $p = (p_1, \ldots, p_N) \in (0,1)^N$ and define $w_n = p_n/(1-p_n)$, the associated odds.

Let $X = (X_1, \ldots, X_N) \in \{0,1\}^N$ such that $X_n \sim \text{Bernoulli}(p_n)$, independently.

The conditional distribution of X given $\sum_{n=1}^N X_n = S$ is called "conditional Bernoulli", denoted by $\text{CBernoulli}(p, S)$.

Example: conditional Bernoulli

Heng, Jacob & Ju, *A simple Markov chain for independent Bernoulli variables conditioned on their sum*, on arXiv.

Let $p = (p_1, \ldots, p_N) \in (0,1)^N$ and define $w_n = p_n/(1-p_n)$, the associated odds.

Let $X = (X_1, \ldots, X_N) \in \{0,1\}^N$ such that $X_n \sim \text{Bernoulli}(p_n)$, independently.

The conditional distribution of X given $\sum_{n=1}^N X_n = S$ is called "conditional Bernoulli", denoted by $\text{CBernoulli}(p, S)$.

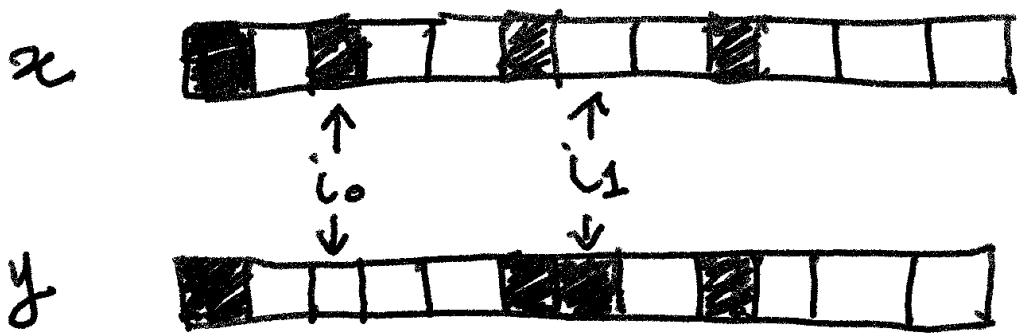
Exact sampling costs $\mathcal{O}(S \cdot N)$, i.e. N^2 if $S \propto N$.

Chen & Liu, *Statistical applications of the Poisson-Binomial and conditional Bernoulli distributions*, Statistica Sinica, 1997.

Example: conditional Bernoulli

A Rosenbluth–Hastings transition goes as follows:

- independently sample $i_0 \in \mathcal{I}_0 = \{n : x_n = 0\}$ and $i_1 \in \mathcal{I}_1 = \{n : x_n = 1\}$ uniformly;
- construct proposed state y with a swap $i_0 \leftrightarrow i_1$;
- accept y as next state with probability $\min\{1, w_{i_0}/w_{i_1}\}$.



Chen, Dempster & Liu, *Weighted finite population sampling to maximize entropy*, Biometrika, 1994.

Relevance of conditional Bernoulli

Identical success probabilities (p_n):

- the chain obtained by successive swaps is known as the Bernoulli-Laplace diffusion model;

Relevance of conditional Bernoulli

Identical success probabilities (p_n):

- the chain obtained by successive swaps is known as the Bernoulli-Laplace diffusion model;
- the chain has been thoroughly studied; if $S = N/2$, mixing occurs in $N/8 \cdot \log N$ iterations (+ cutoff phenomenon).

  Diaconis & Shahshahani, *Time to reach stationarity in the Bernoulli-Laplace diffusion model*, SIAM Journal on Mathematical Analysis, 1987.

Relevance of conditional Bernoulli

Identical success probabilities (p_n):

- the chain obtained by successive swaps is known as the Bernoulli-Laplace diffusion model;
- the chain has been thoroughly studied; if $S = N/2$, mixing occurs in $N/8 \cdot \log N$ iterations (+ cutoff phenomenon).

  Diaconis & Shahshahani, *Time to reach stationarity in the Bernoulli-Laplace diffusion model*, SIAM Journal on Mathematical Analysis, 1987.

Non-identical (p_n): arises in various contexts in statistics, and occurred in our research on agent-based models:

Ju, Heng & Jacob, *Sequential Monte Carlo algorithms for agent-based models of disease transmission*, on arXiv.

Assumptions

- (Condition on the odds). The odds (w_n) are such that there exist $\zeta > 0$, $0 < l < r < \infty$ and $\eta > 0$ such that for all N large enough,

$$\mathbb{P}\left(|\{n \in [N] : w_n \notin (l, r)\}| \leq \zeta N\right) \geq 1 - \exp(-\eta N).$$

- (Condition on S). There exist $0 < \xi \leq 1/2$ and $\eta' > 0$ such that for all N large enough,

$$\mathbb{P}\left(\xi N \leq S\right) \geq 1 - \exp(-\eta' N).$$

Convergence rate of swap chain for conditional Bernoulli

There exist $\kappa > 0$, $\nu > 0$, $N_0 \in \mathbb{N}$ independent of N such that, for any $\epsilon \in (0, 1)$, and for all $N \geq N_0$, with probability at least $1 - \exp(-\nu N)$, we have

$$\|x^{(t)} - \text{CBernoulli}(p, S)\|_{\text{TV}} \leq \epsilon \quad \text{for all} \quad t \geq \kappa N \log(N/\epsilon).$$

Convergence rate of swap chain for conditional Bernoulli

There exist $\kappa > 0$, $\nu > 0$, $N_0 \in \mathbb{N}$ independent of N such that, for any $\epsilon \in (0, 1)$, and for all $N \geq N_0$, with probability at least $1 - \exp(-\nu N)$, we have

$$\|x^{(t)} - \text{CBernoulli}(p, S)\|_{\text{TV}} \leq \epsilon \quad \text{for all} \quad t \geq \kappa N \log(N/\epsilon).$$

- A simple Markov chain provides samples for a cheaper cost than exact sampling: $N \log N$ versus N^2.
- Coupling technique sharp enough to establish a mixing time in $N \log N$.
- But constants appearing in these results are not useful.

Markovian couplings of Markov chains

Moving on,
we will look at practical couplings of MCMC algorithms,
or *coupling MCMC algorithms for practical reasons*.

Markovian couplings of Markov chains

Moving on,
we will look at practical couplings of MCMC algorithms,
or *coupling MCMC algorithms for practical reasons.*

Consider two chains, propagated using a coupled kernel $\bar{P}$.

If $(X', Y') \sim \bar{P}((X, Y), \cdot)$, then

- $X'|(X, Y) \sim P(X, \cdot)$,
- $Y'|(X, Y) \sim P(Y, \cdot)$.

Markovian couplings of Markov chains

We will consider coupled kernels such that

- $\bar{P}(\{X' = Y'\}|X, Y) > 0$ for at least some X, Y,
- $\bar{P}(\{X' = Y'\}|\{X = Y\}) = 1$.

Given an MCMC algorithm, we will try to design a coupled kernel, and aim at obtaining short "meeting times".

Example of coupled kernel

Gibbs sampler:

$$\eta_i | \theta \sim \text{Exponential}\left(\frac{1 + (\theta - x_i)^2}{2}\right) \quad \forall i = 1, \ldots, n$$

$$\theta' | \eta_1, \ldots, \eta_n \sim \text{Normal}\left(\frac{\sum_{i=1}^n \eta_i x_i}{\sum_{i=1}^n \eta_i + \sigma^{-2}}, \frac{1}{\sum_{i=1}^n \eta_i + \sigma^{-2}}\right).$$

Example of coupled kernel

Gibbs sampler:

$$\eta_i|\theta \sim \text{Exponential}\left(\frac{1+(\theta-x_i)^2}{2}\right) \quad \forall i = 1, \ldots, n$$

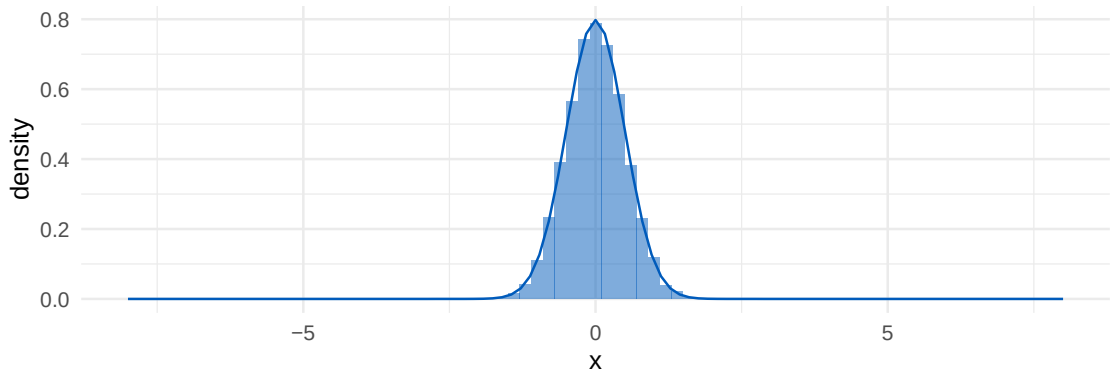
$$\theta'|\eta_1, \ldots, \eta_n \sim \text{Normal}\left(\frac{\sum_{i=1}^n \eta_i x_i}{\sum_{i=1}^n \eta_i + \sigma^{-2}}, \frac{1}{\sum_{i=1}^n \eta_i + \sigma^{-2}}\right).$$

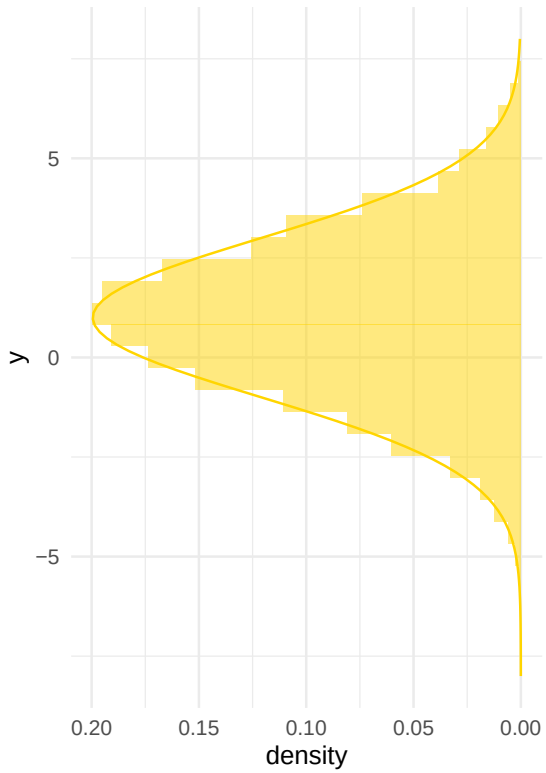
Start from $\theta^{(1)}, \theta^{(2)}$, possibly unequal.

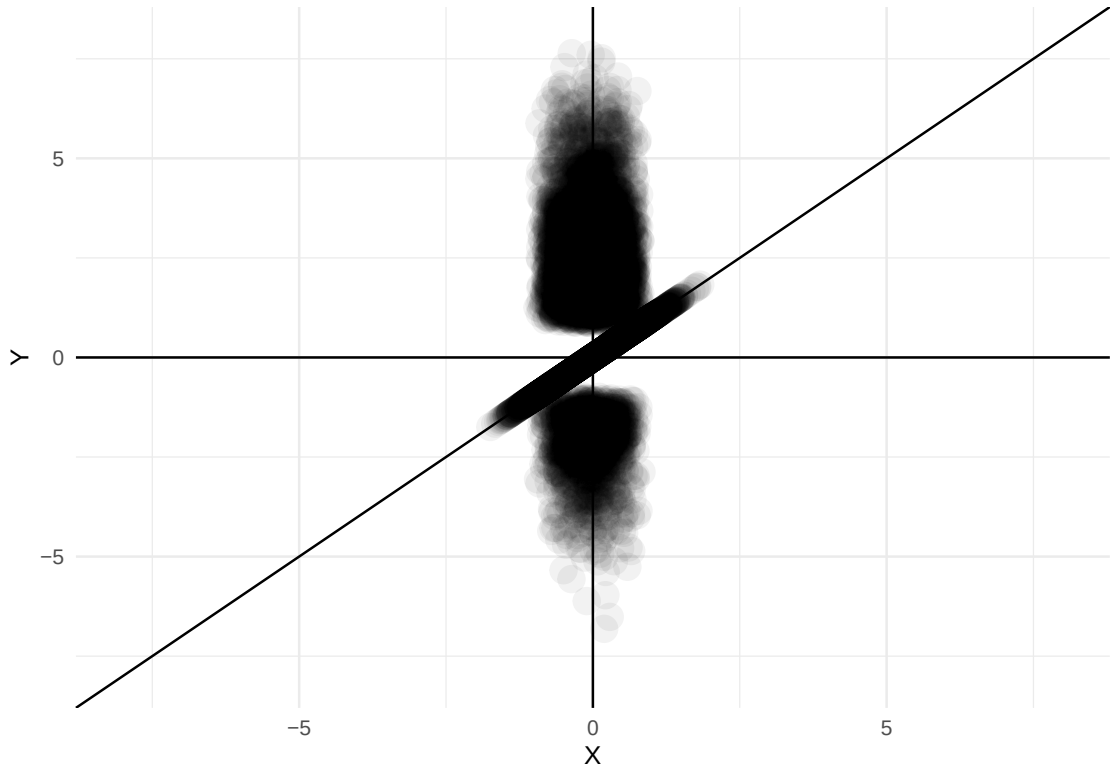
Generate $\eta^{(1)}, \eta^{(2)}$ using common uniforms.

Implement a maximal coupling to sample $\theta'^{(1)}, \theta'^{(2)}$,
i.e. maximize $\mathbb{P}(\theta'^{(1)} = \theta'^{(2)}|\eta^{(1)}, \eta^{(2)})$.

A maximal coupling of two Normals







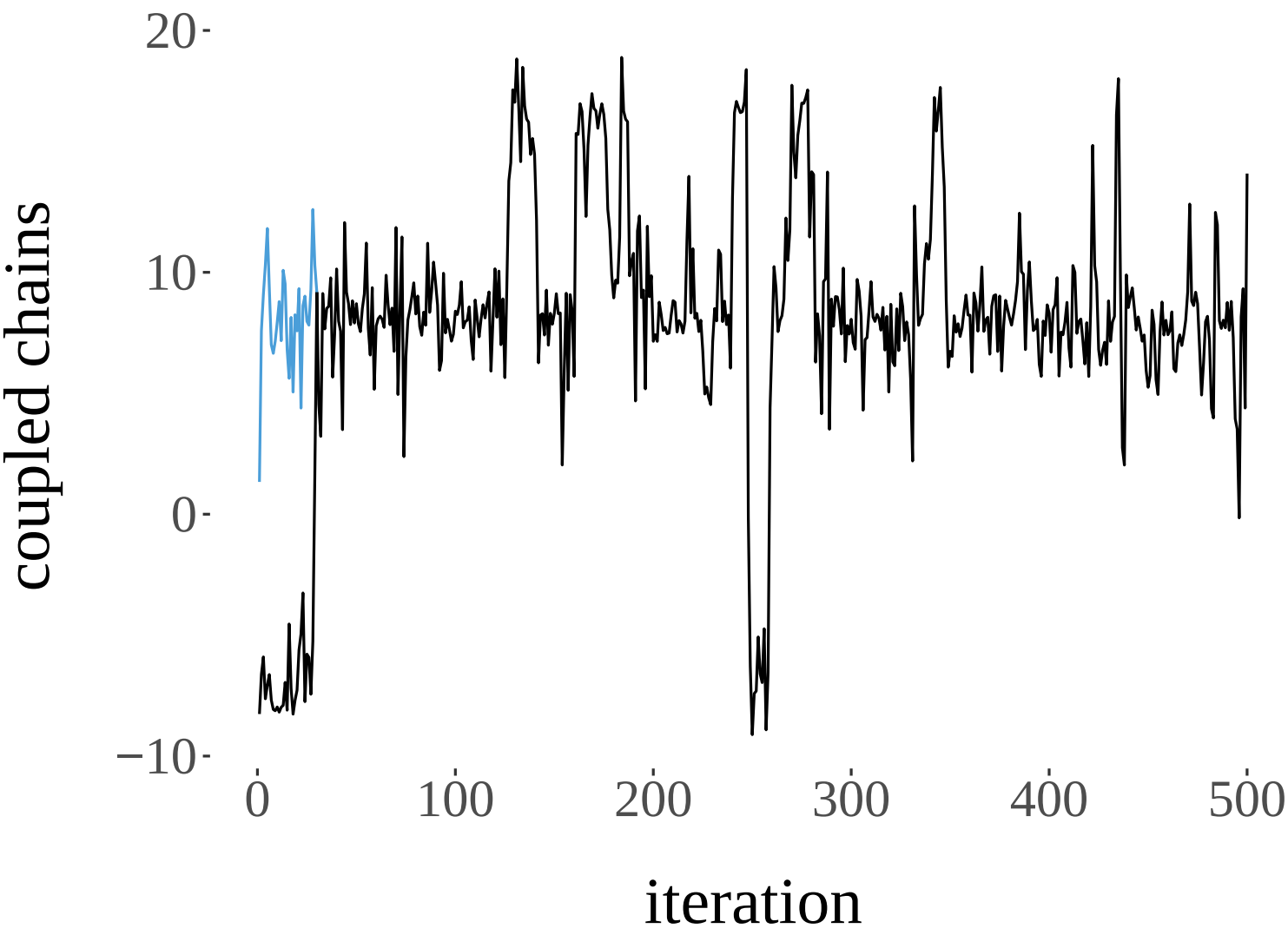
A maximal coupling for two tractable distributions

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

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

Example of coupled trajectories



Practical couplings in realistic MCMC settings

- Niloy Biswas, Anirban Bhattacharya, Pierre E. Jacob & James Johndrow, *Coupling-based convergence assessment of some Gibbs samplers for high-dimensional Bayesian regression with shrinkage priors*, 2022.
- Francisco J. R. Ruiz, Michalis K. Titsias, Taylan Cemgil & Arnaud Doucet, *Unbiased gradient estimation for variational auto-encoders using coupled Markov chains*, 2020.
- Brian L. Trippe, Tin D. Nguyen, Tamara Broderick, *Optimal transport couplings of Gibbs samplers on partitions for unbiased estimation*, 2021.
- Luke J. Kelly, Robin J. Ryder & Grégoire Clarté, *Lagged couplings diagnose Markov chain Monte Carlo phylogenetic inference*, 2022.

Summary

- We can study the convergence of a Markov chain to its limiting distribution using couplings,

- and we might be able to generate pairs of Markov chains, that can exactly meet after a random number of iterations.

Next: new Monte Carlo methods employing such pairs of chains.

Outline

1 Introduction

2 Couplings

3 Unbiased estimation of target expectations

4 Diagnostics of convergence

5 Asymptotic variance estimation

Lagged coupled chains

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 $t = 1, \ldots, L$,
- for $t \geq L+1$, sample $(X_t, Y_{t-L}) | (X_{t-1}, Y_{t-L-1}) \sim \bar{P}\left((X_{t-1}, Y_{t-L-1}), \cdot\right)$.

Lagged coupled chains

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 $t = 1, \ldots, L$,
- for $t \geq L+1$, sample $(X_t, Y_{t-L}) | (X_{t-1}, Y_{t-L-1}) \sim \bar{P}\left((X_{t-1}, Y_{t-L-1}), \cdot\right)$.

Denote by τ the "meeting time" such that $X_t = Y_{t-L}$ for $t \geq \tau$.

Note that $X_t \overset{d}{=} Y_t$ at all times $t \geq 0$.

Unbiased estimators from lagged chains

Here lag $L = 1$ for simplicity. Write limit as a telescopic sum,

$$\begin{aligned}\mathbb{E}_\pi[h(X)] &= \lim_{t\to\infty} \mathbb{E}[h(X_t)] \\ &= \mathbb{E}[h(X_0)] + \sum_{j=1}^{\infty} \mathbb{E}[h(X_j) - h(X_{j-1})].\end{aligned}$$

Unbiased estimators from lagged chains

Here lag $L = 1$ for simplicity. Write limit as a telescopic sum,

$$\mathbb{E}_\pi[h(X)] = \lim_{t \to \infty} \mathbb{E}[h(X_t)]$$
$$= \mathbb{E}[h(X_0)] + \sum_{j=1}^{\infty} \mathbb{E}[h(X_j) - h(X_{j-1})].$$

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

$$= \mathbb{E}[h(X_0)] + \sum_{j=1}^{\infty} \mathbb{E}[h(X_j) - h(Y_{j-1})].$$

Unbiased estimators from lagged chains

Here lag $L = 1$ for simplicity. Write limit as a telescopic sum,

$$\begin{aligned}\mathbb{E}_\pi[h(X)] &= \lim_{t\to\infty} \mathbb{E}[h(X_t)] \\ &= \mathbb{E}[h(X_0)] + \sum_{j=1}^{\infty} \mathbb{E}[h(X_j) - h(X_{j-1})].\end{aligned}$$

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

$$= \mathbb{E}[h(X_0)] + \sum_{j=1}^{\infty} \mathbb{E}[h(X_j) - h(Y_{j-1})].$$

If we cross fingers,

$$= \mathbb{E}\left[h(X_0) + \sum_{j=1}^{\infty} (h(X_j) - h(Y_{j-1}))\right].$$

Unbiased estimators from lagged chains

After some variance reduction tricks and manipulations, an unbiased estimator of $\mathbb{E}_\pi[h(X)]$ is given by

$$\frac{1}{m-k+1}\sum_{t=k}^{m} h(X_t) + \sum_{\ell=k+L}^{\tau-1} \min\left(1, \frac{\lceil (\ell-k)/L \rceil}{m-k+1}\right)(h(X_\ell) - h(Y_{\ell-L})),$$

where user-chosen parameters include L, k and m. Tuning largely an open question.

Unbiased estimators from lagged chains

After some variance reduction tricks and manipulations, an unbiased estimator of $\mathbb{E}_\pi[h(X)]$ is given by

$$\frac{1}{m-k+1}\sum_{t=k}^{m} h(X_t) + \sum_{\ell=k+L}^{\tau-1} \min\left(1, \frac{\lceil (\ell-k)/L \rceil}{m-k+1}\right)(h(X_\ell) - h(Y_{\ell-L})),$$

where user-chosen parameters include L, k and m. Tuning largely an open question.

Benefits of larger lags: comment by Vanetti & Doucet in discussion paper of Jacob, O'Leary & Atchadé, 2020.

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.*

Outline

1 Introduction

2 Couplings

3 Unbiased estimation of target expectations

4 Diagnostics of convergence

5 Asymptotic variance estimation

Upper bounds using couplings

Triangle inequalities with steps of length L between π_t and π,

$$\begin{aligned}\|\pi_t - \pi\|_{\mathrm{TV}} &\leq \sum_{j=1}^{\infty} \|\pi_{t+jL} - \pi_{t+(j-1)L}\|_{\mathrm{TV}} \\ &\leq \sum_{j=1}^{\infty} \mathbb{P}(X_{t+jL} \neq Y_{t+(j-1)L}).\end{aligned}$$

Using coupled lagged chains we estimate $\mathbb{P}(X_{t+jL} \neq Y_{t+(j-1)L})$ by $\mathbb{1}(X_{t+jL} \neq Y_{t+(j-1)L})$, for *all* t, j.

Upper bounds using couplings

Triangle inequalities with steps of length L between π_t and π,

$$\begin{aligned}\|\pi_t - \pi\|_{\mathrm{TV}} &\leq \sum_{j=1}^{\infty} \|\pi_{t+jL} - \pi_{t+(j-1)L}\|_{\mathrm{TV}} \\ &\leq \sum_{j=1}^{\infty} \mathbb{P}(X_{t+jL} \neq Y_{t+(j-1)L}).\end{aligned}$$

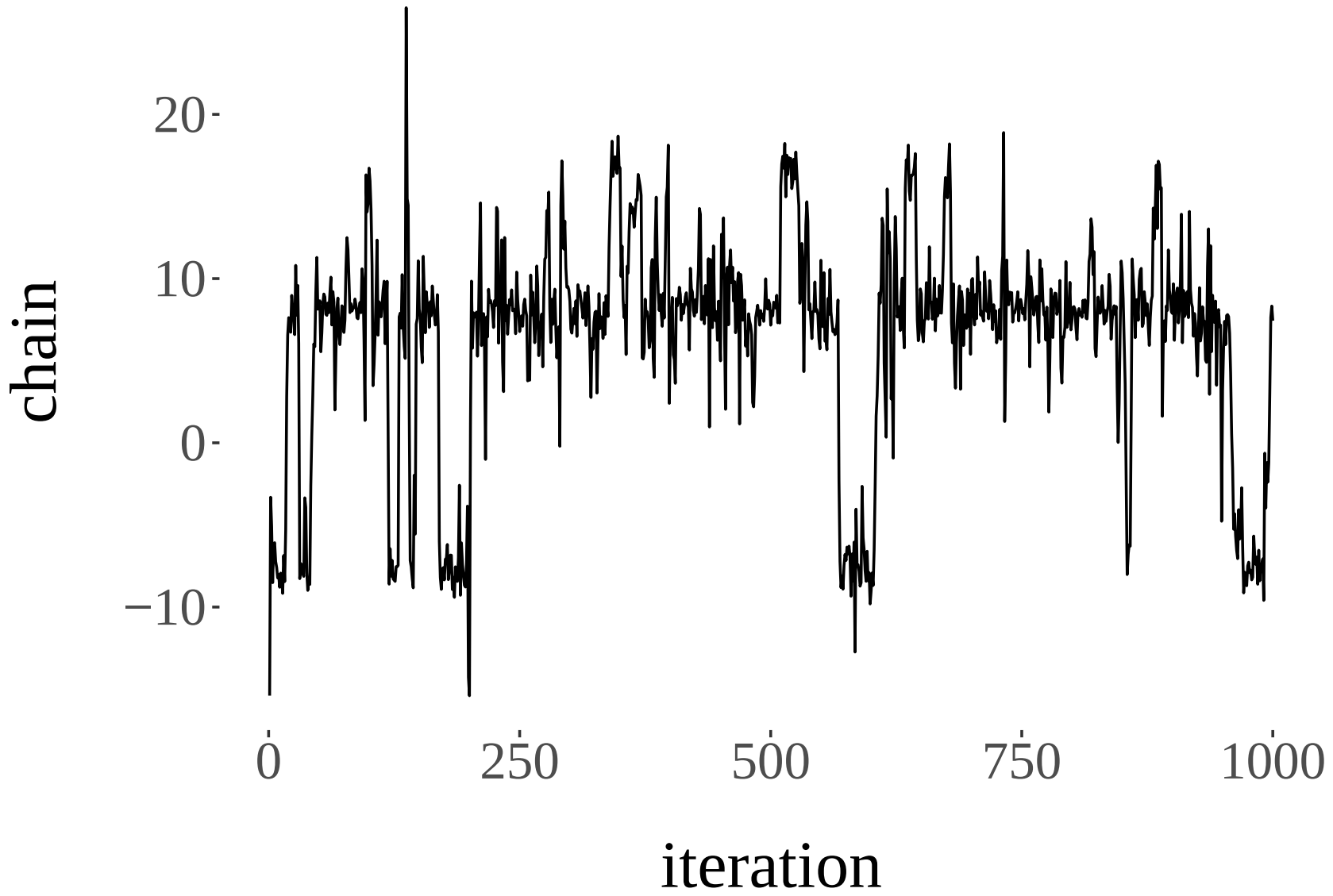
Using coupled lagged chains we estimate $\mathbb{P}(X_{t+jL} \neq Y_{t+(j-1)L})$ by $\mathbb{1}(X_{t+jL} \neq Y_{t+(j-1)L})$, for *all* t, j.

Then, upon an exchange of expectation and limit,

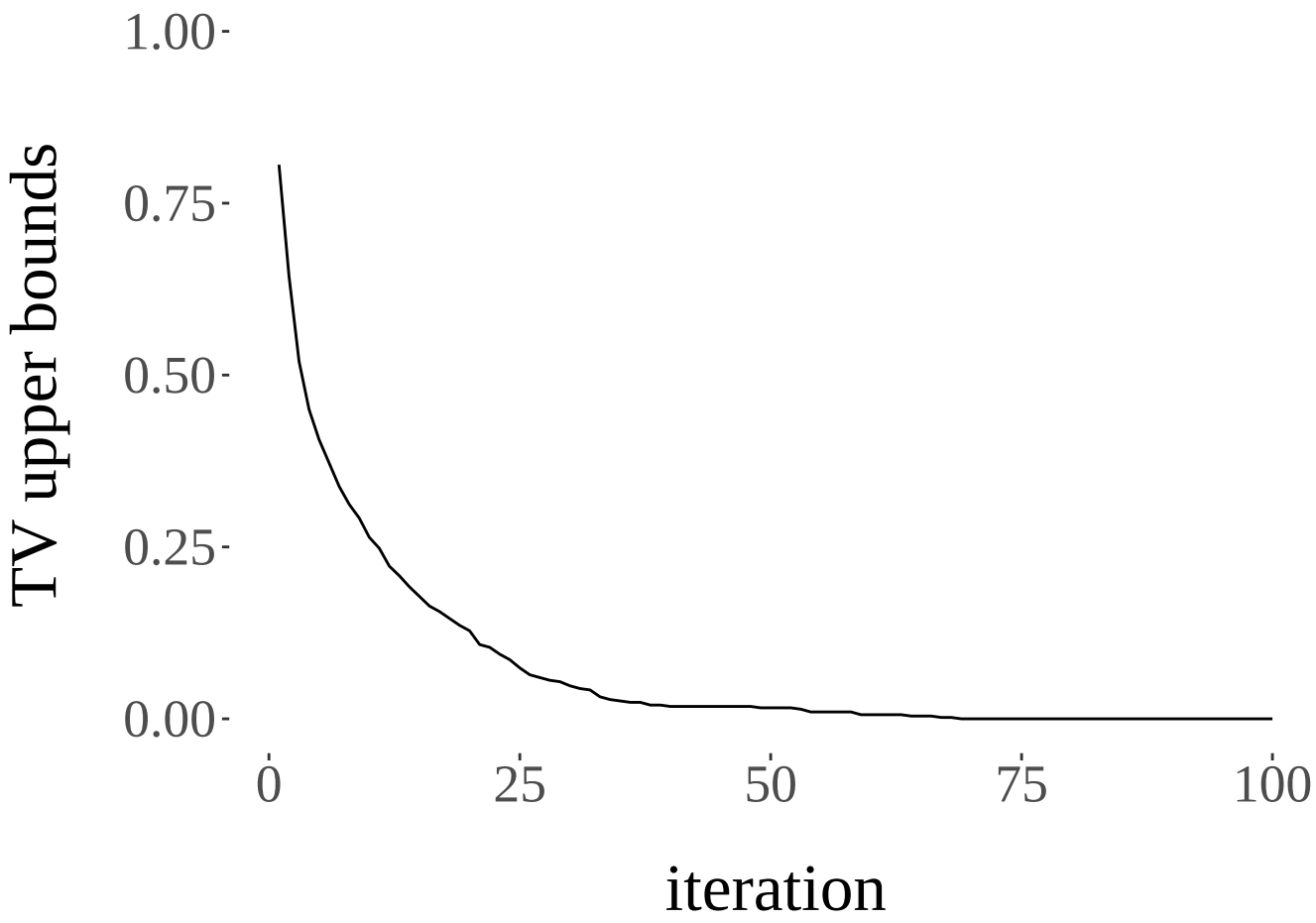
$$\|\pi_t - \pi\|_{\mathrm{TV}} \leq \mathbb{E}[\max(0, \lceil (\tau - L - t)/L \rceil)].$$

Then we estimate the expectation by an empirical average over independent replicates.

Recall the Cauchy-Normal example

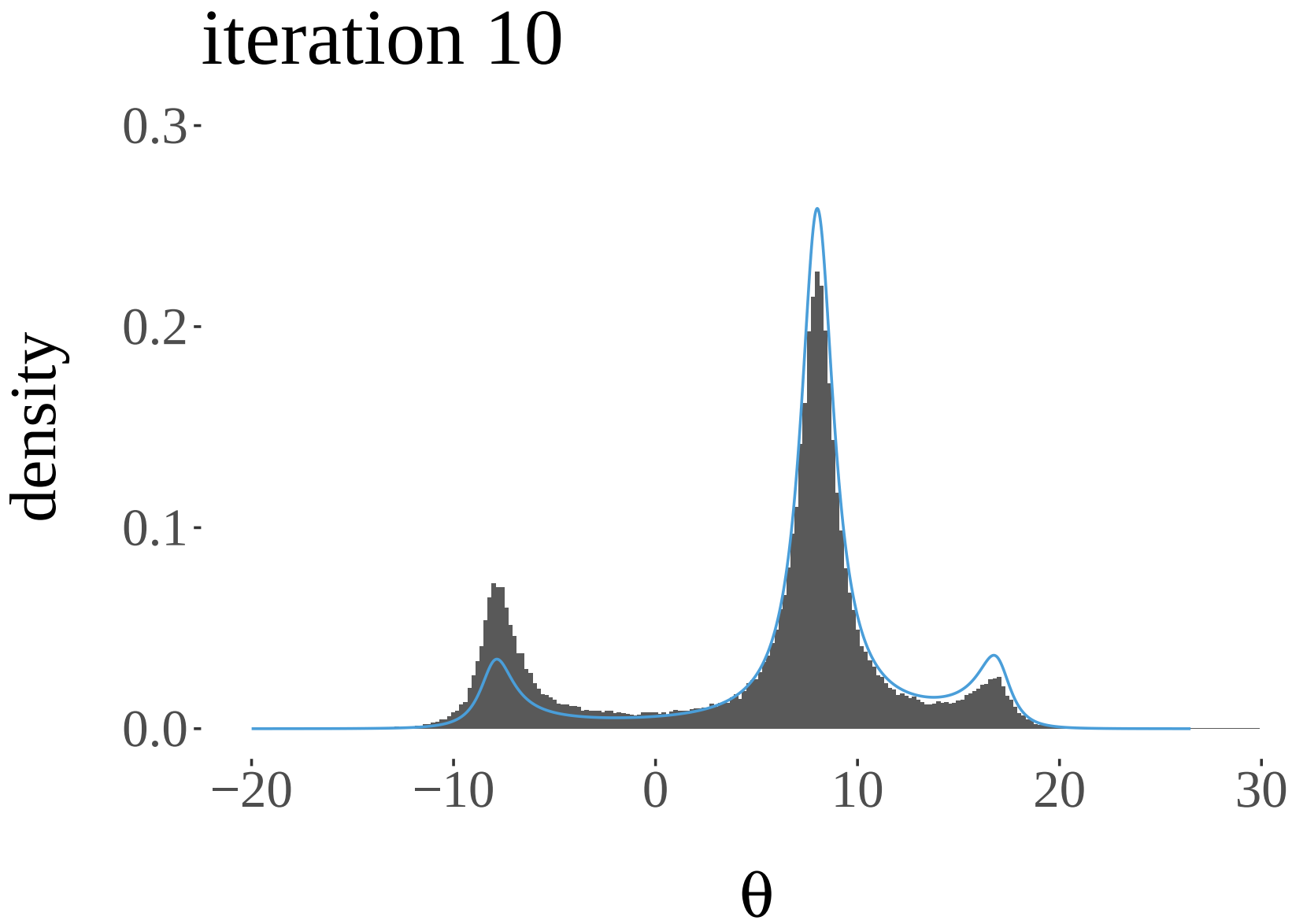


TV upper bounds in the Cauchy-Normal example

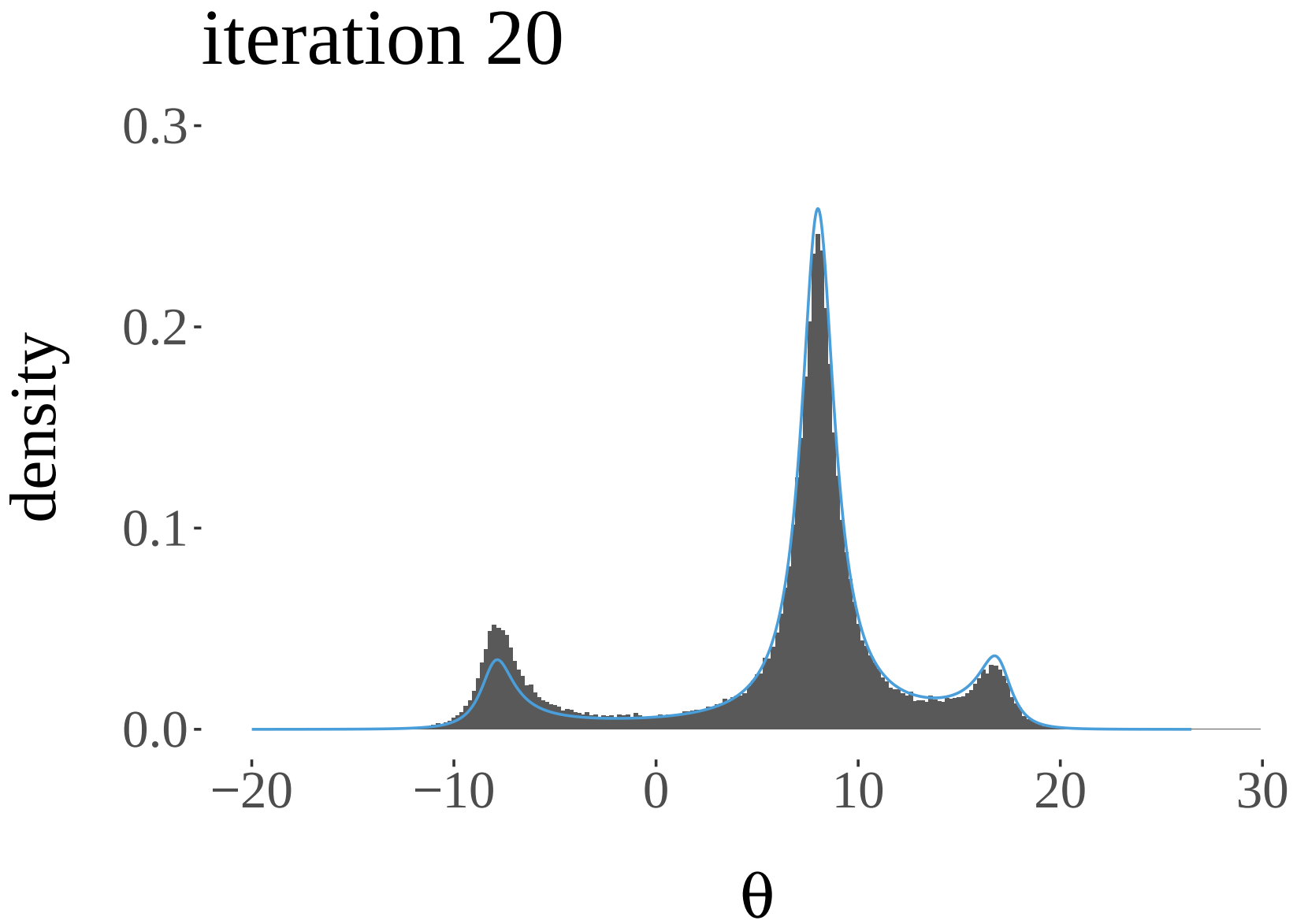


From 500 independent meeting times, with lag $L = 50$.

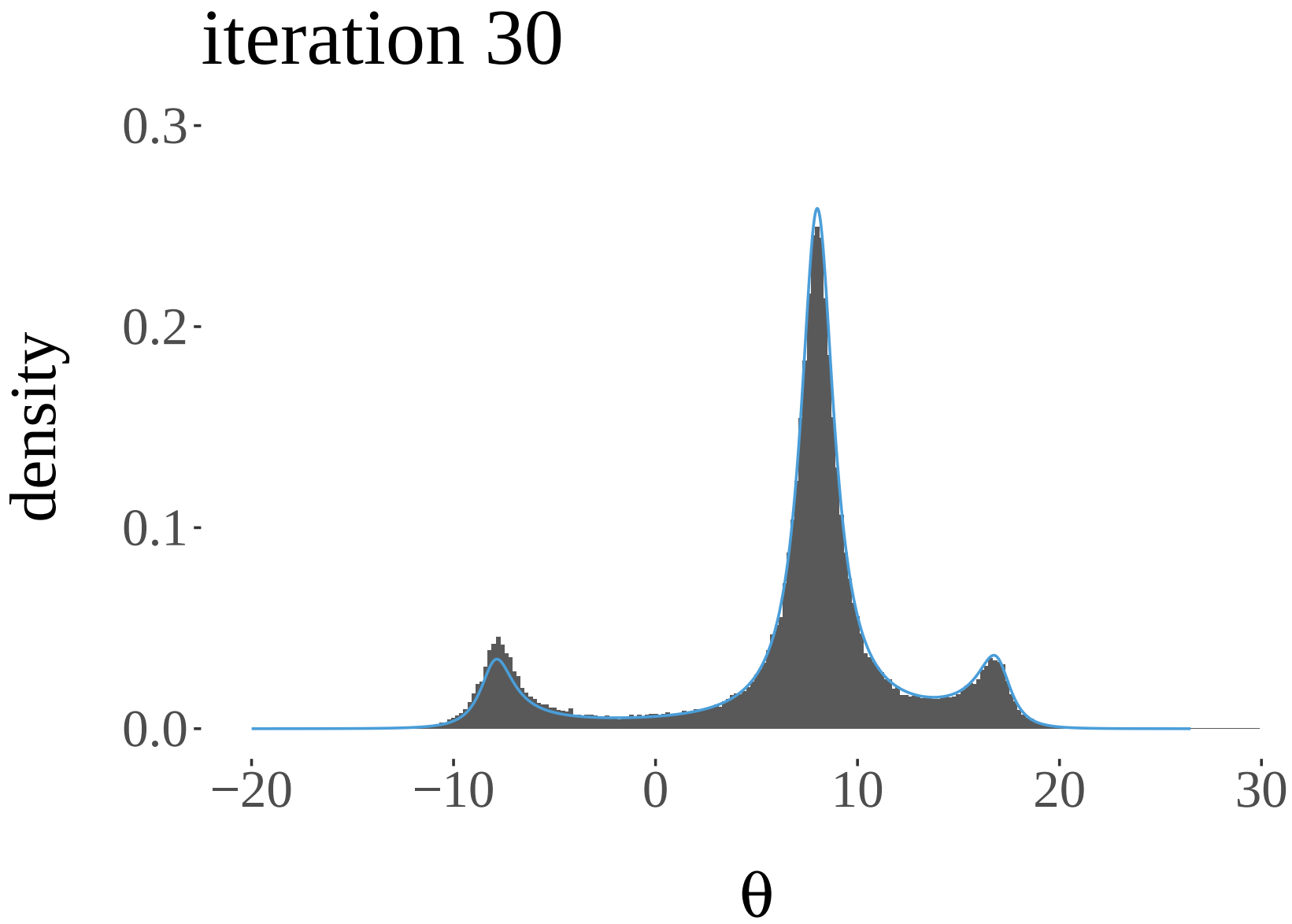
Time marginals in the Cauchy-Normal example



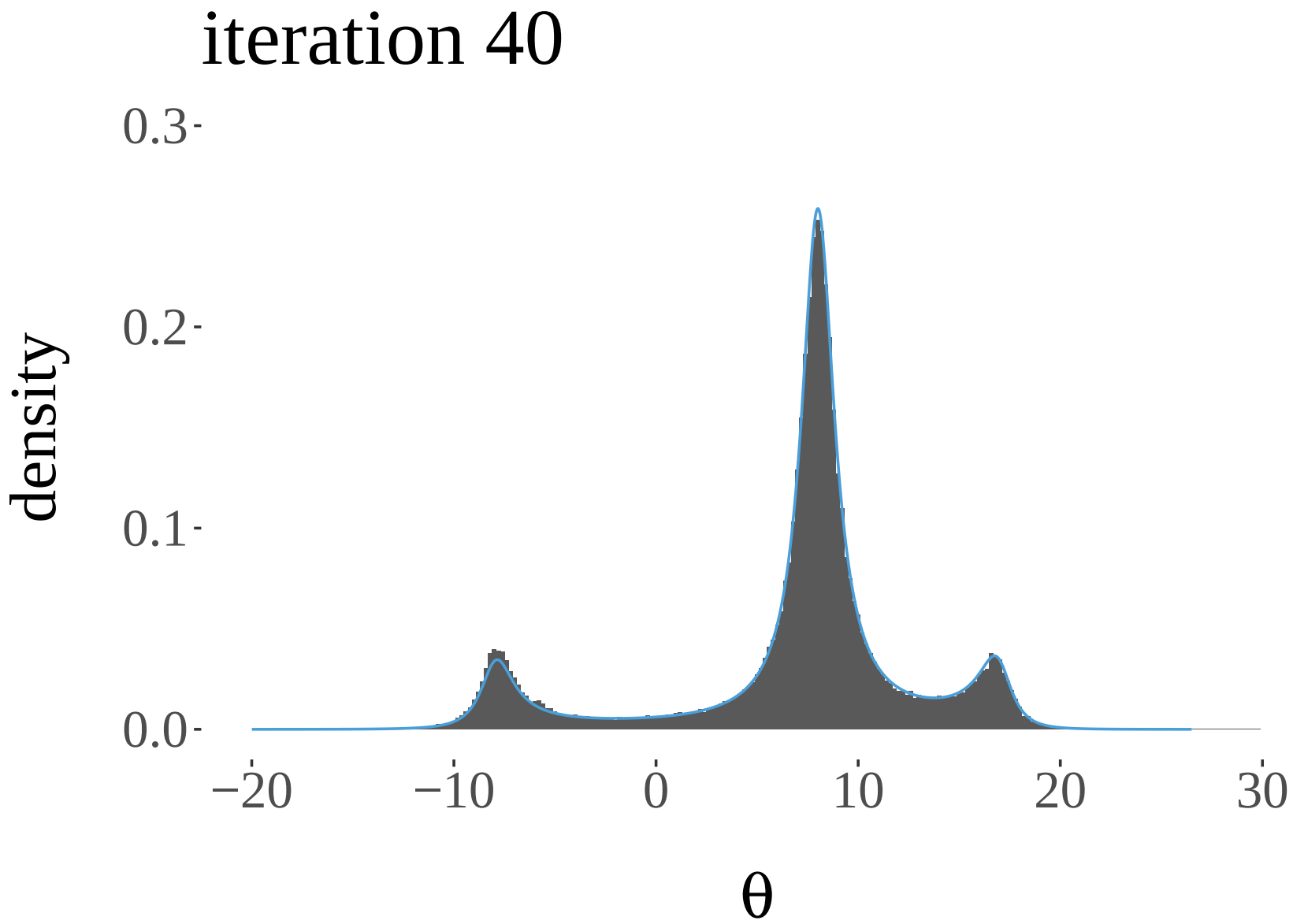
Time marginals in the Cauchy-Normal example



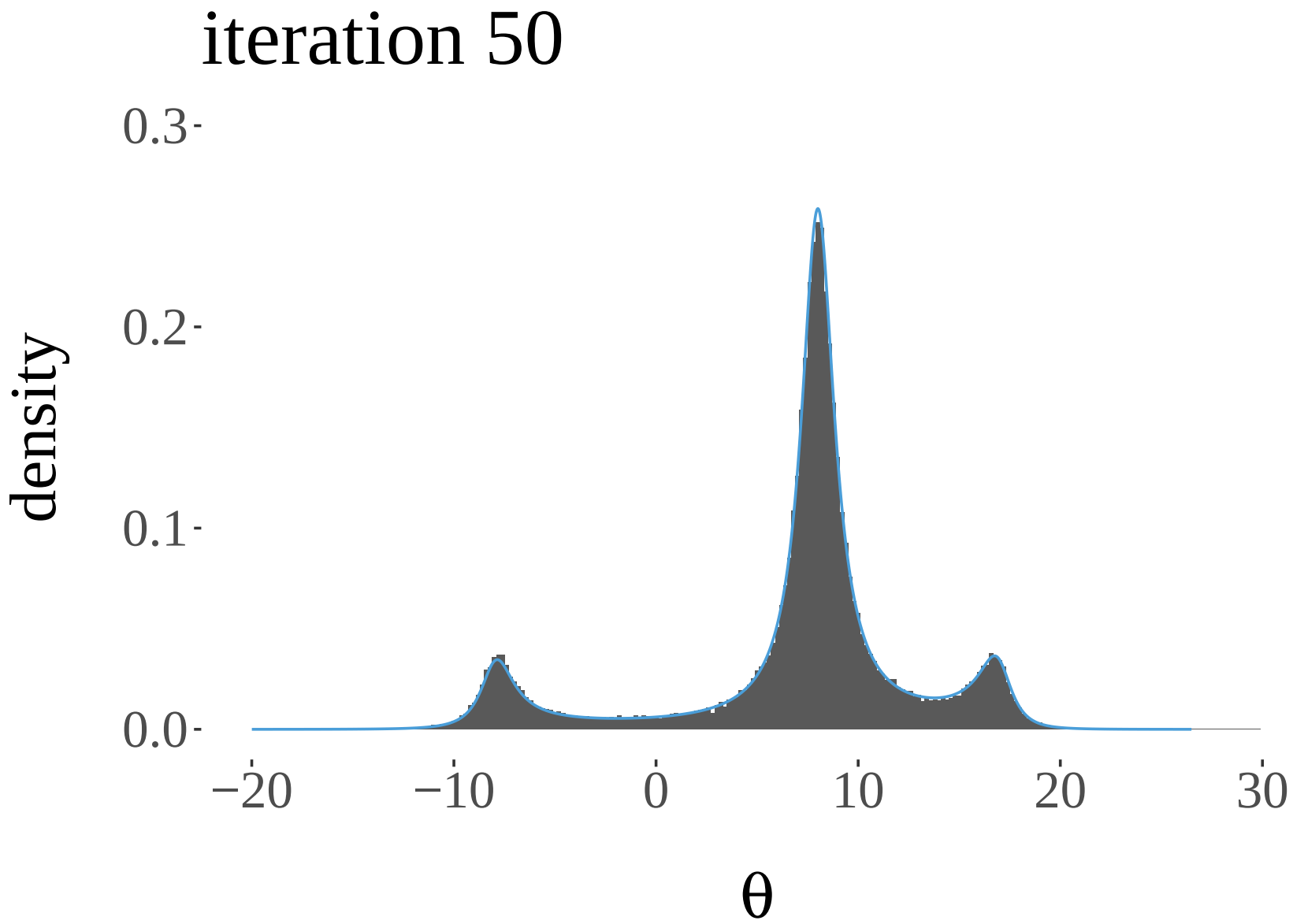
Time marginals in the Cauchy-Normal example



Time marginals in the Cauchy-Normal example



Time marginals in the Cauchy-Normal example



Example: large-scale Bayesian regression

Biswas, Bhattacharya, Jacob & Johndrow, *Coupling-based convergence assessment of some Gibbs samplers for high-dimensional Bayesian regression with shrinkage priors*, 2022.

Linear regression setting, n rows, p columns with $p \gg n$.

$$
\begin{aligned}
& Y \sim \mathcal{N}(X\beta, \sigma^2 I_n), \\
& \sigma^2 \sim \text{InverseGamma}(a_0/2, b_0/2), \\
& \xi^{-1/2} \sim \text{Cauchy}^+, \\
\text{for } j = 1, \ldots, p \quad & \beta_j \sim \mathcal{N}(0, \sigma^2/\xi\eta_j), \quad \eta_j^{-1/2} \sim t(\nu)^+.
\end{aligned}
$$

Global precision ξ, local precision η_j for $j = 1, \ldots, p$.

Example: large-scale Bayesian regression

Gibbs sampler:

- For $j = 1, \ldots, p$, η_j given β, ξ, σ^2 can be sampled from, exactly or by slice sampling.
- Given η, we can sample β, ξ, σ^2:
  - ξ given η using RH step,
  - σ^2 given η, ξ from InverseGamma,
  - β given η, ξ, σ^2 from p-dimensional Normal.

Algorithm has $n^2 p$ cost per iteration.

Example: large-scale Bayesian regression

Gibbs sampler:

- For $j = 1, \ldots, p$, η_j given β, ξ, σ^2 can be sampled from, exactly or by slice sampling.
- Given η, we can sample β, ξ, σ^2:
  - ξ given η using RH step,
  - σ^2 given η, ξ from InverseGamma,
  - β given η, ξ, σ^2 from p-dimensional Normal.

Algorithm has $n^2 p$ cost per iteration.

Coupling strategy involves maximal couplings and common random numbers, combined in certain ways depending on distance between chains.

Example: large-scale Bayesian regression

Gibbs sampler:

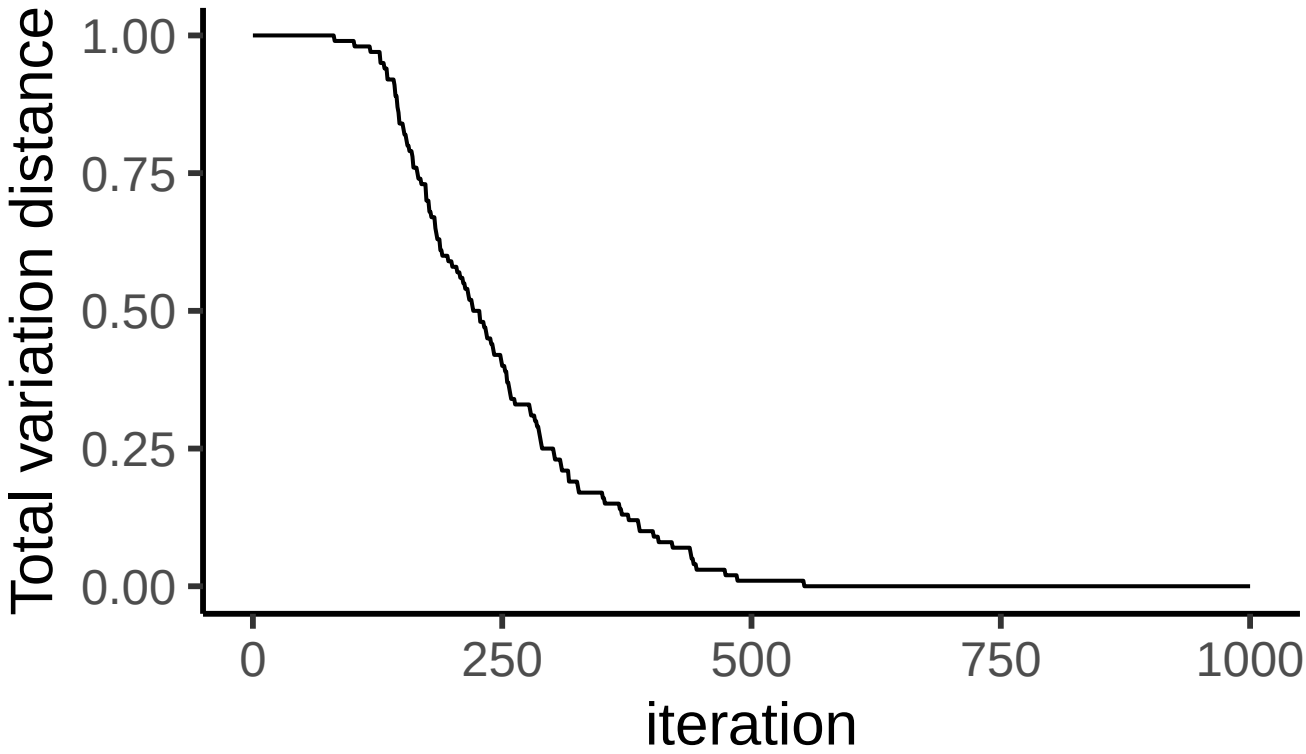
- For $j = 1, \ldots, p$, η_j given β, ξ, σ^2 can be sampled from, exactly or by slice sampling.
- Given η, we can sample β, ξ, σ^2:
  - ξ given η using RH step,
  - σ^2 given η, ξ from InverseGamma,
  - β given η, ξ, σ^2 from p-dimensional Normal.

Algorithm has $n^2 p$ cost per iteration.

Coupling strategy involves maximal couplings and common random numbers, combined in certain ways depending on distance between chains.

Genome-wide association study with $n = 2,266$ and $p = 98,385$.
Outcome: average number of days for silk emergence in maize.
Covariates: single nucleotide polymorphisms of maize.

Example: large-scale Bayesian regression



From 100 independent meeting times, with lag $L = 750$.

Outline

1 Introduction

2 Couplings

3 Unbiased estimation of target expectations

4 Diagnostics of convergence

5 Asymptotic variance estimation

Central limit theorem

Markov kernel P, test function h, might satisfy

$$\sqrt{t}\left(t^{-1}\sum_{s=0}^{t-1} h(X_s) - \pi(h)\right) \to \mathcal{N}(0, v(P,h)),$$

where $v(P,h)$ is called the asymptotic variance.

When the chain is at stationarity (i.e. $X_t \sim \pi$ for all t) we have

$$v(P,h) = \mathbb{V}^\star(h(X_0)) + 2\sum_{t=1}^{\infty} \mathbb{C}\mathrm{ov}^\star(h(X_0), h(X_t)).$$

Difficult to approximate $v(P,h)$ *a priori*, because MCMC chains are not stationary and the sum has infinitely many terms.

The Poisson equation

Write $Ph(x) = \int P(x, dx')h(x') = \mathbb{E}[h(X_1)|X_0 = x]$.

The Poisson equation

Write $Ph(x) = \int P(x, dx')h(x') = \mathbb{E}[h(X_1)|X_0 = x]$.

A function $\tilde{h}$ in $L^1(\pi)$ is said to be a solution of the Poisson equation associated with h and P, if

$$\tilde{h} - P\tilde{h} = h - \pi(h).$$

For brevity we say that $\tilde{h}$ is fishy.

The Poisson equation

Write $Ph(x) = \int P(x, dx')h(x') = \mathbb{E}[h(X_1)|X_0 = x]$.

A function $\tilde{h}$ in $L^1(\pi)$ is said to be a solution of the Poisson equation associated with h and P, if

$$\tilde{h} - P\tilde{h} = h - \pi(h).$$

For brevity we say that $\tilde{h}$ is fishy.

If $\sum_{t \geq 0} \|P^t\{h - \pi(h)\}\|_{L^1(\pi)} < \infty$ then fishy functions exist.

Marie Duflo, *Opérateurs potentiels des chaînes et des processus de Markov irréductibles*, 1970.

Central limit theorem

Aiming for a CLT for Markov chain ergodic averages, write

Central limit theorem

Aiming for a CLT for Markov chain ergodic averages, write

$$\sum_{s=0}^{t-1}\{h(X_s) - \pi(h)\} = \sum_{s=1}^{t}\left\{\tilde{h}(X_s) - P\tilde{h}(X_{s-1})\right\} + \tilde{h}(X_0) - \tilde{h}(X_t).$$

Central limit theorem

Aiming for a CLT for Markov chain ergodic averages, write

$$\sum_{s=0}^{t-1}\{h(X_s) - \pi(h)\} = \sum_{s=1}^{t}\left\{\tilde{h}(X_s) - P\tilde{h}(X_{s-1})\right\} + \tilde{h}(X_0) - \tilde{h}(X_t).$$

Then apply the central limit theorem for martingale difference sequences, leading to the asymptotic variance

$$v(P, h) = \mathbb{E}^\star[\{\tilde{h}(X_1) - P\tilde{h}(X_0)\}^2].$$

Chapter 21 in
Douc, Moulines, Priouret & Soulier, *Markov chains*, 2018.

Asymptotic variance

The more familiar form of the asymptotic variance is

$$\lim_{t\to\infty} \mathbb{V}\left(t^{-1/2}\sum_{s=0}^{t-1} h(X_s)\right) = \mathbb{V}^\star(h(X)) + 2\sum_{s=1}^{\infty} \mathbb{C}\mathrm{ov}^\star\left(h(X_0), h(X_s)\right).$$

Asymptotic variance

The more familiar form of the asymptotic variance is

$$\lim_{t\to\infty} \mathbb{V}\left(t^{-1/2}\sum_{s=0}^{t-1} h(X_s)\right) = \mathbb{V}^\star(h(X)) + 2\sum_{s=1}^{\infty} \mathbb{C}\mathrm{ov}^\star\left(h(X_0), h(X_s)\right).$$

This expression is equivalent to $\mathbb{E}^\star[\{\tilde{h}(X_1) - P\tilde{h}(X_0)\}^2]$.

Asymptotic variance

The more familiar form of the asymptotic variance is

$$\lim_{t\to\infty} \mathbb{V}\left(t^{-1/2}\sum_{s=0}^{t-1} h(X_s)\right) = \mathbb{V}^\star(h(X)) + 2\sum_{s=1}^{\infty} \mathbb{C}\mathrm{ov}^\star\left(h(X_0), h(X_s)\right).$$

This expression is equivalent to $\mathbb{E}^\star[\{\tilde{h}(X_1) - P\tilde{h}(X_0)\}^2]$.

Use $\tilde{h} = \sum_{t=0}^{\infty} P^t\{h - \pi(h)\}$, and $\tilde{h} = h - \pi(h) + P\tilde{h}$,

Asymptotic variance

The more familiar form of the asymptotic variance is

$$\lim_{t\to\infty} \mathbb{V}\left(t^{-1/2}\sum_{s=0}^{t-1} h(X_s)\right) = \mathbb{V}^\star(h(X)) + 2\sum_{s=1}^{\infty} \mathbb{C}\mathrm{ov}^\star\left(h(X_0), h(X_s)\right).$$

This expression is equivalent to $\mathbb{E}^\star[\{\tilde{h}(X_1) - P\tilde{h}(X_0)\}^2]$.

Use $\tilde{h} = \sum_{t=0}^{\infty} P^t\{h - \pi(h)\}$, and $\tilde{h} = h - \pi(h) + P\tilde{h}$,

$$\mathbb{E}^\star[\{\tilde{h}(X_1) - P\tilde{h}(X_0)\}^2] = \pi(\{h - \pi(h)\}^2) + 2\pi(\{h - \pi(h)\} \cdot P\tilde{h}).$$

Asymptotic bias

Consider the bias of the average $t^{-1} \sum_{s=0}^{t-1} h(X_s)$.

Asymptotic bias

Consider the bias of the average $t^{-1} \sum_{s=0}^{t-1} h(X_s)$.

Its expectation is $t^{-1} \sum_{s=0}^{t-1} P^s h(x_0)$, given $X_0 = x_0 \in \mathbb{X}$.

Asymptotic bias

Consider the bias of the average $t^{-1} \sum_{s=0}^{t-1} h(X_s)$.

Its expectation is $t^{-1} \sum_{s=0}^{t-1} P^s h(x_0)$, given $X_0 = x_0 \in \mathbb{X}$.

Therefore

$$\lim_{t \to \infty} t \times \left\{ \mathbb{E}_{x_0}[t^{-1} \sum_{s=0}^{t-1} h(X_s)] - \pi(h) \right\} = \tilde{h}(x_0),$$

where $\tilde{h}$ is the fishy function as before.

Kontoyiannis & Dellaportas, *Notes on using control variates for estimation with reversible MCMC samplers*, 2009.

Unbiased estimation of fishy functions

Choose an arbitrary $y \in \mathbb{X}$. The function

$$x \mapsto \tilde{h}(x) = \sum_{t=0}^{\infty} \left\{ P^t h(x) - P^t h(y) \right\},$$

is fishy. It *wants* to be estimated with coupled Markov chains.

Unbiased estimation of fishy functions

Choose an arbitrary $y \in \mathbb{X}$. The function

$$x \mapsto \tilde{h}(x) = \sum_{t=0}^{\infty} \left\{ P^t h(x) - P^t h(y) \right\},$$

is fishy. It *wants* to be estimated with coupled Markov chains.

If we set $X_0 = x$, $Y_0 = y$, and generate X_t, Y_t such that

$$\begin{cases} X_t | X_{t-1} \sim P(X_{t-1}, \cdot) \\ Y_t | Y_{t-1} \sim P(Y_{t-1}, \cdot) \end{cases} \quad \text{and} \quad \forall t \geq \tau_{x,y} \quad X_t = Y_t,$$

Unbiased estimation of fishy functions

Choose an arbitrary $y \in \mathbb{X}$. The function

$$x \mapsto \tilde{h}(x) = \sum_{t=0}^{\infty} \left\{ P^t h(x) - P^t h(y) \right\},$$

is fishy. It *wants* to be estimated with coupled Markov chains.

If we set $X_0 = x$, $Y_0 = y$, and generate X_t, Y_t such that

$$\begin{cases} X_t | X_{t-1} \sim P(X_{t-1}, \cdot) \\ Y_t | Y_{t-1} \sim P(Y_{t-1}, \cdot) \end{cases} \quad \text{and} \quad \forall t \geq \tau_{x,y} \quad X_t = Y_t,$$

then

$$\tilde{H}(x) = \sum_{t=0}^{\tau_{x,y}-1} \left\{ h(X_t) - h(Y_t) \right\},$$

has expectation equal to $\tilde{h}(x)$.

Unbiased estimation of the asymptotic variance

We can write

$$v(P, h) = 2\pi(\{h - \pi(h)\}\tilde{h}) - v(\pi, h),$$

where $v(\pi, h) = \pi(h^2) - \pi(h)^2$

Unbiased estimation of the asymptotic variance

We can write

$$v(P, h) = 2\pi(\{h - \pi(h)\}\tilde{h}) - v(\pi, h),$$

where $v(\pi, h) = \pi(h^2) - \pi(h)^2$

We can obtain unbiased approximations $\hat{\pi}$ of π, and we can estimate $\tilde{h}$ unbiasedly, point-wise.

Unbiased estimation of the asymptotic variance

We can write

$$v(P, h) = 2\pi(\{h - \pi(h)\}\tilde{h}) - v(\pi, h),$$

where $v(\pi, h) = \pi(h^2) - \pi(h)^2$

We can obtain unbiased approximations $\hat{\pi}$ of π, and we can estimate $\tilde{h}$ unbiasedly, point-wise.

Estimating $v(P, h)$ is an exercise in "nested Monte Carlo".

Unbiased estimation of the asymptotic variance

1. Obtain $\hat{\pi}^{(1)}$ and $\hat{\pi}^{(2)}$, two independent approximations of π.
2. Write $\hat{\pi}^{(1)}(\cdot) = \sum_{n=1}^{N} \omega_n \delta_{Z_n}$. For $r = 1, \ldots, R$,
   - sample $\ell^{(r)} \sim (\xi_1, \ldots, \xi_N)$,
   - generate $\tilde{H}^{(r)}$ with expectation $\tilde{h}(Z_{\ell^{(r)}})$.
3. Estimate

$$\pi(\{h - \pi(h)\}\tilde{h}) \quad \text{with} \quad R^{-1} \sum_{r=1}^{R} w_{\ell^{(r)}} (h(Z_{\ell^{(r)}}) - \hat{\pi}^{(2)}(h)) \tilde{H}^{(r)} / \xi_{\ell^{(r)}};$$

$$v(\pi, h) \quad \text{with} \quad \frac{1}{2}\{\hat{\pi}^{(1)}(h^2) + \hat{\pi}^{(2)}(h^2)\} - \hat{\pi}^{(1)}(h) \times \hat{\pi}^{(2)}(h).$$

Douc, Jacob, Lee & Vats,
Estimation of fishy functions with couplings, on-going work.

Discussion

- Some basic questions about MCMC are still largely open, such as: "how long should we run the chain?"

- Couplings are powerful for theoretical analysis

  but they can also be implemented, leading to new methods.

Discussion

- Some basic questions about MCMC are still largely open, such as: "how long should we run the chain?"

- Couplings are powerful for theoretical analysis
  but they can also be implemented, leading to new methods.

Thank you all for listening!

Collaborators mentioned in these slides: Yves Atchadé, Anirban Bhattacharya, Niloy Biswas, Arthur Dempster, Randal Douc, Arnaud Doucet, Paul Edlefsen, Ruobin Gong, Jeremy Heng, James Johndrow, Nianqiao Ju, Anthony Lee, John O'Leary, Paul Vanetti, Dootika Vats, Guanyang Wang.