What do MH kernel couplings look like?

Guanyang Wang (Rutgers Stats)

joint work with

John O'Leary Harvard Stats

(Special thanks to Pierre E.Jacob)

International Conference on Monte Carlo Methods and Applications
2021

Outline

Introduction

MH-like coupling: Theory & Algorithms

Our contribution

What is a coupling?

Informal Definition: Given two random variables $X \sim p, Y \sim q$, a coupling of X and Y means a joint distribution such that its first marginal is p and second marginal is q.

What is a coupling?

Informal Definition: Given two random variables $X \sim p, Y \sim q$, a coupling of X and Y means a joint distribution such that its first marginal is p and second marginal is q.

Formal Definition: Let p, q be two probability measures on the same measurable space $(S, \mathcal{S})$. A coupling of (p, q) is a probability measure γ on $(S \times S, \mathcal{S} \times \mathcal{S})$ such that for every $A \in \mathcal{S}$:

$$\gamma(A \times S) = p(A) \qquad \text{and} \qquad \gamma(S \times A) = q(A).$$

What is a coupling?

Informal Definition: Given two random variables $X \sim p, Y \sim q$, a coupling of X and Y means a joint distribution such that its first marginal is p and second marginal is q.
Formal Definition: Let p, q be two probability measures on the same measurable space $(S, \mathcal{S})$. A coupling of (p, q) is a probability measure γ on $(S \times S, \mathcal{S} \times \mathcal{S})$ such that for every $A \in \mathcal{S}$:

$$\gamma(A \times S) = p(A) \qquad \text{and} \qquad \gamma(S \times A) = q(A).$$

Example: If X and Y are both (fair) coin flips, then

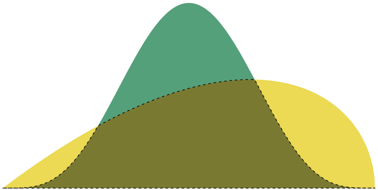
- Identity coupling: Flip a coin once, and set $X = Y =$ the outcome
- Negation coupling: Flip a coin once, and set $X =$ the outcome $= 1 - Y$
- Independent coupling: Flip a coin twice, $X =$ outcome of the first flip, $Y =$ outcome of the second flip.

Coupling inequality and maximal coupling

The Coupling Inequality: Let γ be any coupling of p and q, the coupling inequality says:

$$\mathbb{P}_\gamma(X = Y) \leq 1 - d_{\mathsf{TV}}(p, q)$$

- $1 - d_{\mathsf{TV}}(p, q)$ is the area of the shaded region



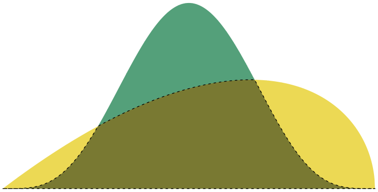
Coupling inequality and maximal coupling

The Coupling Inequality: Let γ be any coupling of p and q, the coupling inequality says:

$$\mathbb{P}_\gamma(X = Y) \leq 1 - d_{\mathsf{TV}}(p, q)$$

- $1 - d_{\mathsf{TV}}(p, q)$ is the area of the shaded region
- A coupling γ_0 is called a maximal coupling if

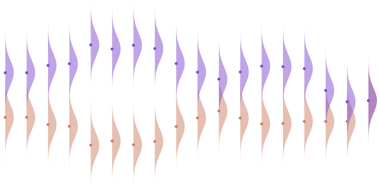
$$\mathbb{P}_{\gamma_0}(X = Y) = 1 - d_{\mathsf{TV}}(p, q)$$



Coupling and Markov chain Monte Carlo

Coupling plays a central role in MCMC theory and methods. People use coupling for

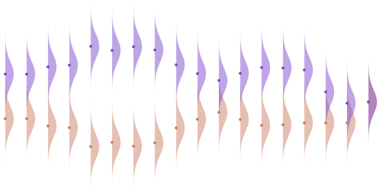
- Analyzing convergence rates
- Perfect sampling
- Unbiased estimation
- Convergence diagnosis
- ...



Coupling and Markov chain Monte Carlo

Coupling plays a central role in MCMC theory and methods. Couping has been used for

- Analyzing convergence rates
- Perfect sampling
- Unbiased estimation
- Convergence diagnosis
- ...



In many cases, better coupling design $\approx$ shorter meeting time

Outline

Introduction

MH-like coupling: Theory & Algorithms

Our contribution

MH-like transition kernel

Definition (MH-like transition kernel)

We say a Markov kernel P is MH-like if it is generated by a generated by a proposal kernel Q and an acceptance function $a : \mathcal{X} \times \mathcal{X} \to [0, 1]$, i.e.,

$$x' \sim Q(x, \cdot)$$

and

$$b_x \sim B(x, x') = \mathrm{Bern}(a(x, x'))$$

imply

$$X := b_x x' + (1 - b_x)x \sim P(x, \cdot).$$

- The proposal kernel includes random-walk (RWM), MALA/HMC, informed proposals, and so on.
- The function a can be the MH acceptance function, the Barker's acceptance function, and others (see Andrieu et al. [1]).

Transition kernel coupling, proposal coupling, and acceptance indicator coupling

Definition (Transition kernel/Proposal coupling)

Let P be a MH-like kernel generated by Q and a. We denote by $\Gamma(P, P)$ the set of all the coupling of P. Every $\gamma \in \Gamma(P, P)$ is called a transition kernel coupling. Similarly, every $\mu \in \Gamma(Q, Q)$ is called a proposal coupling.

Definition (Acceptence indicator coupling)

We call $\bar{B} : \mathcal{X}^2 \times \mathcal{X}^2 \to [0, 1]^2$ an acceptance indicator coupling if $\bar{B}((x, y), \cdot)$ is measurable for all $(x, y) \in \mathcal{X}^2$.

- Given (x', y'), $\bar{B}((x, y), (x', y'))$ defines a point in $[0, 1]^2$, which specifies a pair of coupled Bernoulli random variables.

Two types of couplings

- Couplings directly defined in terms of the transition kernel P.
  - Minorization coupling ('splitting and coupling'), one-shot coupling.
  - Very widely used in MCMC theory, see Rosenthal [9], Qin and Hobert [8] for examples.
  - Often unimplementable.

Two types of couplings

- Couplings directly defined in terms of the transition kernel P. (one-step coupling)
  - Minorization coupling ('splitting and coupling'), one-shot coupling.
  - Very widely used in MCMC theory, see Rosenthal [9], Qin and Hobert [8] for examples.
  - Often unimplementable.
- Coupling defined by a proposal coupling followed by an acceptance indicator coupling. (two-step coupling)
  - The Johnson's γ-coupling [5], used also in Jacob [4] et al., Biswas et al. [2] and the subsequent papers.
  - Bou-Rabee et al.'s Reflection coupling [3], used also in Jacob et al.[4].
  - Used in both theory (reflection coupling) and practice (γ-coupling).

Three questions, two bottlenecks, and one solution

Three questions that are difficult to answer:

- What does a kernel coupling generally 'looks like'?
- What is the representability of the implementable couplings?
- How to design a good coupling?

Three questions, two bottlenecks, and one solution

Three questions that are difficult to answer:

- What does a kernel coupling generally 'looks like'?
- What is the representability of the implementable couplings?
- How to design a good coupling?

Two bottlenecks:

- The set of all possible couplings are abstract objects (as a joint measure with specific margins) and there no systematic study on its structure.
- Designing a good coupling is often considered as an artwork.

Three questions, two bottlenecks, and one solution

Three questions:

- What does a kernel coupling generally 'looks like'?
- What is the representability of the implementable couplings?
- How to design a good coupling?

Two bottlenecks:

- The set of all possible couplings are abstract objects (as a joint measure with specific margins) and there no systematic study on its structure.
- Designing a good coupling is often considered as an artwork.

One solution:

- We need some representation theorems!

Some examples of representation theorems

- Structure theorem for finitely generated (f.g.) modules over a PID.
  - Application: Jordan canonical form $\rightarrow$ Tells us what matrices look like (up to a similar transformation).
  - Application: Fundamental theorem for f.g. Abelian groups.
- Riesz's representation theorem.
  - Tells us what the continuous linear functionals look like.
- Cayley's theorem.
  - Tells us what groups look like.
- Nath/Whitney's embedding theorem (geometry), Skorokhod's embedding theorem (probability) ......

Outline

Our contribution: A representation theorem for MH-like kernel couplings

Theorem (O'Leary and G., 2021)

Let P be the MH-like transition kernel on $(\mathcal{X}, \mathcal{F})$ generated by a proposal kernel Q and an acceptance rate function a. Then $\bar{P} \in \Gamma(P, P)$ if and only if $\bar{P}$ is generated by $\bar{Q} \in \Gamma(Q, Q)$ and an acceptance indicator coupling $\bar{B}$ such that if $(b_x, b_y) \sim \bar{B}((x, y), (x', y'))$, then for all $x, y \in \mathcal{X}$:

1. $\mathbb{P}(b_x = 1 \mid x, y, x') = a(x, x')$ *for $Q(x, \cdot)$-almost all x', and*
2. $\mathbb{P}(b_y = 1 \mid x, y, y') = a(y, y')$ *for $Q(y, \cdot)$-almost all y'.*

- Informally: Every MH-like kernel coupling $\Leftrightarrow$ proposal coupling $+$ acceptance indicator coupling.

Our contribution: A representation theorem for MH-like kernel couplings

Theorem (O'Leary and G., 2021)

Let P be the MH-like transition kernel on $(\mathcal{X}, \mathcal{F})$ generated by a proposal kernel Q and an acceptance rate function a. Then $\bar{P} \in \Gamma(P, P)$ if and only if $\bar{P}$ is generated by $\bar{Q} \in \Gamma(Q, Q)$ and an acceptance indicator coupling $\bar{B}$ such that if $(b_x, b_y) \sim \bar{B}((x, y), (x', y'))$, then for all $x, y \in \mathcal{X}$:

1. $\mathbb{P}(b_x = 1 \mid x, y, x') = a(x, x')$ *for $Q(x, \cdot)$-almost all x', and*
2. $\mathbb{P}(b_y = 1 \mid x, y, y') = a(y, y')$ *for $Q(y, \cdot)$-almost all y'.*

- Informally: Every MH-like kernel coupling $\Leftrightarrow$ proposal coupling + acceptance indicator coupling.
- Additionally: The proof is purely constructive.

Our contribution: A representation theorem for MH-like kernel couplings

Theorem (O'Leary and G., 2021)

Let P *be the MH-like transition kernel on* $(\mathcal{X}, \mathcal{F})$ *generated by a proposal kernel* Q *and an acceptance rate function* a. *Then* $\bar{P} \in \Gamma(P, P)$ *if and only if* $\bar{P}$ *is generated by* $\bar{Q} \in \Gamma(Q, Q)$ *and an acceptance indicator coupling* $\bar{B}$ *such that if* $(b_x, b_y) \sim \bar{B}((x, y), (x', y'))$, *then for all* $x, y \in \mathcal{X}$:

1. $\mathbb{P}(b_x = 1 \mid x, y, x') = a(x, x')$ *for* $Q(x, \cdot)$*-almost all* x', *and*
2. $\mathbb{P}(b_y = 1 \mid x, y, y') = a(y, y')$ *for* $Q(y, \cdot)$*-almost all* y'.

- Informally: Every transition kernel coupling $\Leftrightarrow$ proposal coupling $+$ acceptance indicator coupling.
- Additionally: The proof is purely constructive.
- Even better: We have a similar representation theorem for maximal couplings.

Corollary: Representability of the coupling algorithms

- Our results confirm the two-step couplings used in practice can represent every transition kernel coupling.

Corollary: Representability of the coupling algorithms

- Our results confirm the two-step couplings used in practice can represent every transition kernel coupling.
- Therefore it suffices to only focus on optimizing the parameters in proposal couplings/acceptance indicator couplings.
  - Still a lot to do. O'Leary et al. [7] may be viewed as an attempt.
  - One big question is the scalability of different proposal couplings. Reflection coupling seems to scale very well, but we still need some theory ...
  - Many numerical experiments are reported in O'Leary [6].

Some elements in the proof

- Key difficulty: $\bar{P}$ is complicated. Given $(x, y) \in \mathcal{X} \times \mathcal{X}$, the measure $\bar{P}((x, y), \cdot)$ has singularity on the x-line (rejecting the x-move), on the y-line (rejecting the y-move), and on the single point (x, y) (rejecting both moves).

Some elements in the proof

- Key difficulty: $\bar{P}$ is complicated. Given $(x_0, y_0) \in \mathcal{X} \times \mathcal{X}$, the measure $\bar{P}((x, y), \cdot)$ has singularity on the x-line (rejecting the x-move), on the y-line (rejecting the y-move), and on the single point (x, y) (rejecting both moves).
- Idea: Decompose $\bar{P}$ into 4 parts. Project it onto
  1. The vertical line $(x, \cdot)$,
  2. The horizontal line $(\cdot, y)$,
  3. The single point (x, y),
  4. The remaining part.

Some elements in the proof

- Key difficulty: $\bar{P}$ is complicated. Given $(x, y) \in \mathcal{X} \times \mathcal{X}$, the measure $\bar{P}((x, y), \cdot)$ has singularity on the x-line (rejecting the x-move), on the y-line (rejecting the y-move), and on the single point (x, y) (rejecting both moves).
- Idea: Decompose $\bar{P}$ into 4 parts. Project it onto
  1. The vertical line $(x, \cdot)$,
  2. The horizontal line $(\cdot, y)$,
  3. The single point (x, y),
  4. The remaining part.
- Construct the proposal $\bar{Q}$ carefully to assign its mass to the four parts while maintaining the constraints on the acceptance probability.

Some elements in the proof

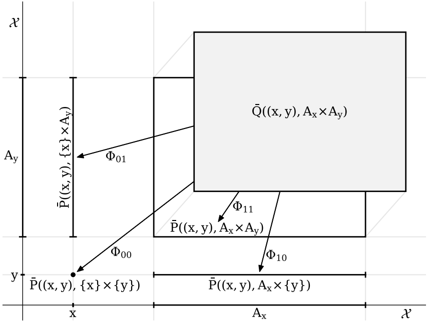


Figure: Schematic diagram of a coupled acceptance mechanism Φ relating a proposal coupling $\bar{Q}$ and a transition kernel coupling $\bar{P}$. Here $A_x \times A_y$ is a measurable rectangle in $\mathcal{X} \times \mathcal{X}$. $\bar{Q}((x, y), A_x \times A_y)$ gives the probability of a proposal $(x', y') \in A_x \times A_y$. The coupled acceptance mechanism $\Phi = (\Phi_{11}, \Phi_{10}, \Phi_{01}, \Phi_{00})$ distributes this probability into contributions to the probability $\bar{P}((x, y), \cdot)$ of a transition from (x, y) to the sets $A_x \times A_y$, $A_x \times \{y\}$, $\{x\} \times A_y$, and $\{x\} \times \{y\}$.

Take-home Message

- We prove a representation theorem for the transition kernel couplings of MH-like algorithms.
- Every transition kernel coupling $\Leftrightarrow$ a proposal coupling + an acceptance indicator coupling.
- The framework is overall quite general. It works for both continuous and discrete spaces, 'lazy' implementations where the proposal distribution is not absolutely continuous with respect to the base measure, and methods like Barker's algorithm.
- It allows us to analyze the relatively complicated set of all MH-like transition kernel couplings in simpler and more tractable ingredients.
- More examples including the detailed characterization of every RWM/MALA coupling, MH on a discrete distribution, and minorization coupling are included in our paper https://arxiv.org/abs/2102.00366 (new version to be uploaded).

Take-home Message

- Formal descriptions of theorems, detailed proofs, examples and discussions are in our paper:

  "Transition kernel couplings of the Metropolis-Hastings algorithm."
  by O'Leary, Wang.

- To learn more about coupling and Monte Carlo, please check out Pierre's excellent lecture notes:

  Couplings and Monte Carlo

- Questions/comments are welcome! Let's couple our ideas together

Thanks!

References I

Christophe Andrieu, Anthony Lee, and Sam Livingstone.
A general perspective on the Metropolis-Hastings kernel.
arXiv preprint arXiv:2012.14881, 2020.

Niloy Biswas, Pierre E Jacob, and Paul Vanetti.
Estimating convergence of markov chains with l-lag couplings.
Advances in neural information processing systems, 32, 2019.

Nawaf Bou-Rabee, Andreas Eberle, and Raphael Zimmer.
Coupling and convergence for hamiltonian monte carlo.
The Annals of Applied Probability, 30(3):1209–1250, 2020.

Pierre E Jacob, John O'Leary, and Yves F Atchadé.
Unbiased Markov chain Monte Carlo methods with couplings.
Journal of the Royal Statistical Society: Series B (Statistical Methodology), 82(3):543–600, 2020.

Valen E Johnson.
A coupling-regeneration scheme for diagnosing convergence in markov chain monte carlo algorithms.
Journal of the American Statistical Association, 93(441):238–248, 1998.

John O'Leary.
Couplings of the Random-Walk Metropolis algorithm.
arXiv preprint arXiv:2102.01790, 2021.

John O'Leary, Guanyang Wang, and Pierre E Jacob.
Maximal couplings of the Metropolis-Hastings algorithm.
arXiv preprint arXiv:2010.08573, 2020.

References II

Qian Qin and James P Hobert.
Wasserstein-based methods for convergence complexity analysis of MCMC with applications.
The Annals of Applied Probability, 2021.

Jeffrey S Rosenthal.
Minorization conditions and convergence rates for Markov chain Monte Carlo.
Journal of the American Statistical Association, 90(430):558–566, 1995.