# Maximal Couplings of the Metropolis–Hastings Algorithm


Guanyang Wang (Rutgers Stats)

joint work with

| | |
|---|---|
| John O'Leary | Harvard Stats |
| Pierre E. Jacob | Harvard Stats |


AISTATS 2021

# Outline

Introduction

Status Quo: Johnson's Coupling

Maximal Couplings

Numerical Examples

Concluding Remarks

# 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 $\gamma$ 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 $\gamma$ 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

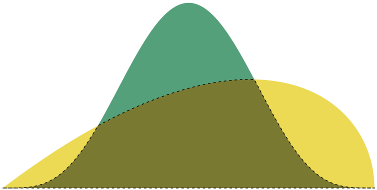
- Identity coupling: Flip a coin once, and set $X = Y =$ the outcome
- Negation coupling: Flip a coin once, and set $X =$ the outcome $= 1 - Y$
- Independent coupling: Flip a coin twice, $X =$ outcome of the first flip, $Y =$ outcome of the second flip.

# Coupling inequality and maximal coupling

The Coupling Inequality: Let $\gamma$ 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



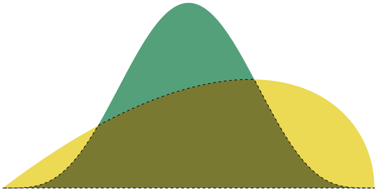
# Coupling inequality and maximal coupling

The Coupling Inequality: Let $\gamma$ 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 $\gamma_0$ is called a maximal coupling if

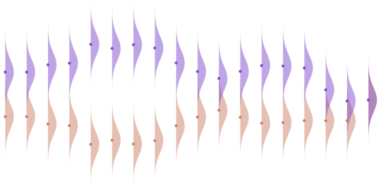
$$\mathbb{P}_{\gamma_0}(X = Y) = 1 - d_{\mathsf{TV}}(p, q)$$



# Coupling and Markov chain Monte Carlo

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

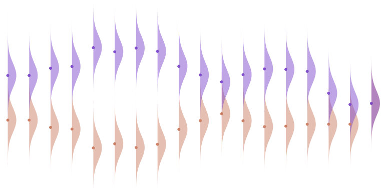
- Analyzing convergence rates
- Perfect sampling
- Unbiased estimation
- Convergence diagnosis
- ...



# 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

Status Quo: Johnson's Coupling

Maximal Couplings

Numerical Examples

Concluding Remarks

## Johnson's Coupling: Max$^2 \neq$ Max

- Theorists can construct any virtual coupling algorithm as long as it is mathematically correct. In practice, it is important for coupling algorithm to be implementable.

## Johnson's Coupling: Max$^2 \neq$ Max

- Theorists can construct any virtual coupling algorithm as long as it is mathematically correct. In practice, it is important for coupling algorithm to be implementable.
- The state of the art implementable coupling of Metropolis-Hastings (MH) transition kernels is introduced in Johnson [3] and used in Jacob, O'Leary, and Atchadé [2].

## Johnson's Coupling: Max$^2 \neq$ Max

- Theorists can construct any virtual coupling algorithm as long as it is mathematically correct. In practice, it is important for coupling algorithm to be implementable.
- The state of the art implementable coupling of Metropolis-Hastings (MH) transition kernels is introduced in Johnson [3] and used in Jacob, O'Leary, and Atchadé [2].
- Idea of Johnson's coupling: a maximal coupling of the proposal distribution followed by a maximal coupling of the acceptance/rejection step.

## Johnson's Coupling: Max$^2 \neq$ Max

- Theorists can construct any virtual coupling algorithm as long as it is mathematically correct. In practice, it is important for coupling algorithm to be implementable.
- The state of the art implementable coupling of Metropolis-Hastings (MH) transition kernels is introduced in Johnson [3] and used in Jacob, O'Leary, and Atchadé [2].
- Idea of Johnson's coupling: a maximal coupling of the proposal distribution followed by a maximal coupling of the acceptance/rejection step.
  - Works for very general MH algorithms, easy to implement
  - Not a maximal coupling: Max proposal coupling $\times$ Max acceptance/rejection coupling $\neq$ Max transition kernel coupling (which may result in slower meeting time)

# Johnson's Coupling: Max$^2 \neq$ Max

- Theorists can construct any virtual coupling algorithm as long as it is mathematically correct. In practice, it is important for coupling algorithm to be implementable.
- The state of the art implementable coupling of Metropolis-Hastings (MH) transition kernels is introduced in Johnson [3] and used in Jacob, O'Leary, and Atchadé [2].
- Idea of Johnson's coupling: a maximal coupling of the proposal distribution followed by a maximal coupling of the acceptance/rejection step.
  - Works for very general MH algorithms, easy to implement
  - Not a maximal coupling: Max proposal coupling $\times$ Max acceptance/rejection coupling $\neq$ Max transition kernel coupling (which may result in slower meeting time)
- Our contribution: We design three classes of implementable maximal couplings of the MH transition kernel.

# Outline

Introduction

Status Quo: Johnson's Coupling

Maximal Couplings

Numerical Examples

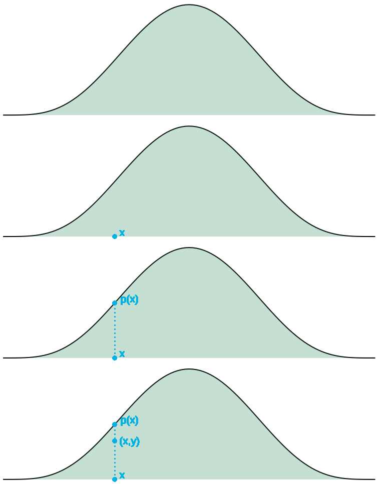
Concluding Remarks

# High level idea: go to a higher dimension

Sample $X \sim p \Leftrightarrow$ Sample $(X, H) \sim \mathrm{Unif}(S)$, $S := \{(x, h) | h \leq p(x)\}$

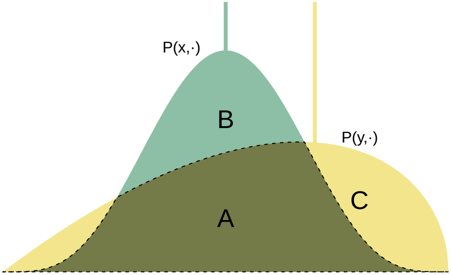
A simple algorithm to sample from $\mathrm{Unif}(S)$:

- Sample $x \sim p$
- Sample $u \sim \mathrm{Unif}[0, 1]$
- Set $h = u \cdot p(x)$
- $(x, h) \sim \mathrm{Unif}(S)$



# Maximal coupling design

Maximally couple $P(x, \cdot)$ and $P(y, \cdot) \Leftrightarrow$ Couple everything in area A

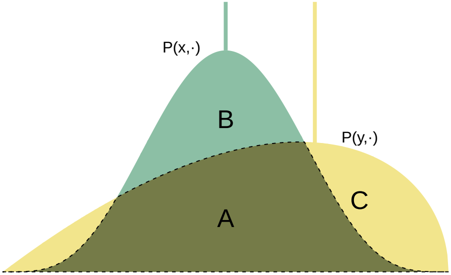


# Maximal coupling design

Maximally couple $P(x, \cdot)$ and $P(y, \cdot)$ $\Leftrightarrow$ Couple everything in area A

A heuristic algorithm:

- Sample $x' \sim P(x, \cdot)$
- Sample $h'$ such that $(x', h') \sim \mathrm{Unif}(A \cup B)$
- If $(x', h')$ in $A$: Set $y' = x'$ and we are done 
- Otherwise: Sample $(y', h'') \sim \mathrm{Unif}(C)$

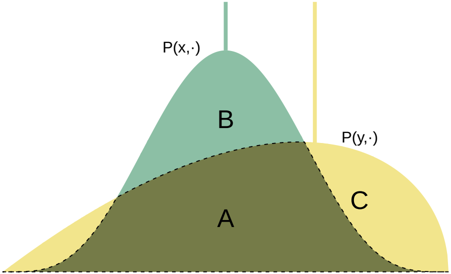


## Maximal coupling design

Maximally couple $P(x, \cdot)$ and $P(y, \cdot) \Leftrightarrow$ Couple everything in area A

A heuristic algorithm:

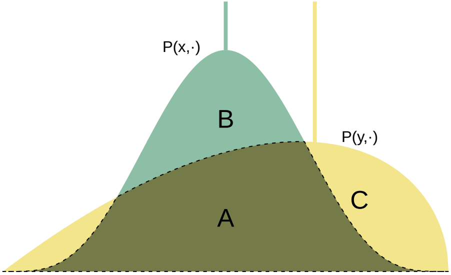
- Sample $x' \sim P(x, \cdot)$
- Sample $h'$ such that $(x', h') \sim \mathrm{Unif}(A \cup B)$
- If $(x', h')$ in $A$: Set $y' = x'$ and we are done ☺
- Otherwise: Sample $(y', h'') \sim \mathrm{Unif}(C)$




The last step still needs to be carefully designed. The easiest way is to repeatedly sample $(y_{\text{new}}, h_{\text{new}}) \sim \mathrm{Unif}(A \cup C)$ until it falls into $C$. Many other choices are available.

# Maximal coupling design

- In our paper, we introduce three classes of maximal couplings: $\bar{P}_{\mathrm{MI}}$, $\bar{P}_{\mathrm{MR}}$, $\bar{P}_{\mathrm{C}}$
- Numerical examples suggest $\bar{P}_{\mathrm{C}}$ works best when proposal has suitable symmetricity.
- Big idea: $\bar{P}_{\mathrm{C}}$ uses reflection coupling for the proposal + conditional acceptance coupling depending on the region.
- In some regions, the acceptance probability is higher than the standard MH rate, while in some other regions is lower.



# Outline

Introduction

Status Quo: Johnson's Coupling

Maximal Couplings

Numerical Examples

Concluding Remarks

- Question: In practice, is maximal coupling a big improvement?

- Question: In practice, is maximal coupling a big improvement?

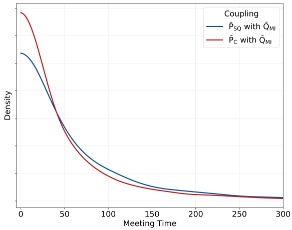
- Answer: Sometimes, but not always!

- Question: In practice, is maximal coupling a big improvement?

- Answer: Sometimes, but not always!
  - In low dimensional cases, the maximal couplings will outperform non-maximal couplings,
  - In high dimensional cases, the advantage of maximal coupling is limited. Sometimes the non-maximal coupling perform better than than maximal couplings.

## Example (Biased Random Walk MH)

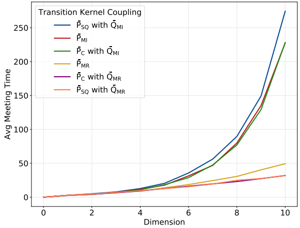
- Take $\pi = \mathrm{Expo}(1)$ and $Q(z, \cdot) = \mathrm{N}(z + \kappa, \sigma^2)$ with $\kappa > 0$.
- $Q$ tends to propose increasing values while $a$ favors decreasing ones.
- We set $\kappa = \sigma^2 = 3$, draw $X_0, Y_0 \overset{iid}{\sim} \pi$, run 10,000 replications for each coupling, and record the meeting times.

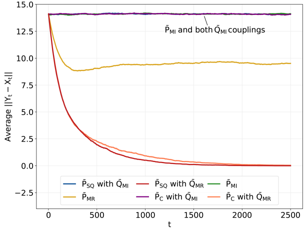



| Coupling | Avg. Meeting Time | S.E. |
|---|---|---|
| Non-maximal | 74.0 | 0.94 |
| Maximal | 61.3 | 0.87 |

# Example (Dimension Scaling with a Normal Target)

- Take $\pi = \mathrm{N}(0, I_d)$ and $Q(z, \cdot) = \mathrm{N}(z, I_d \sigma_d^2)$, $\sigma_d^2 = \frac{2.38^2}{d}$
- Upper plot: we draw $X_0, Y_0 \overset{iid}{\sim} \pi$, run 1,000 replications for each coupling, and record the meeting times for $d \in \{1, 2, \cdots, 10\}$.
- Lower plot: we draw $X_0, Y_0 \overset{iid}{\sim} \pi$, run 1,000 replications for each coupling, and record the distance $\|X_t - Y_t\|$ for $d = 100$.
- In low dimensions they all perform similarly, in high dimensions $\bar{P}_{\mathrm{C}}$ performs best while $\bar{P}_{\mathrm{MI}}$ performs worst.



# Outline

Introduction

Status Quo: Johnson's Coupling

Maximal Couplings

Numerical Examples

Concluding Remarks

# Take-home Message

- We introduce three classes of implementable maximal couplings of MH transition kernels.

# Take-home Message

- We introduce three classes of implementable maximal couplings of MH transition kernels.
- Their cost is roughly the same as the non-maximal couplings.

## Take-home Message

- We introduce three classes of implementable maximal couplings of MH transition kernels.
- Their cost is roughly the same as the non-maximal couplings.
- They have advantages: easy to implement; can naturally generalize to other methods such as Metropolis-adjusted Langevin algorithm, pseudo-marginal Monte Carlo, MH-within-Gibbs; outperforms previous methods at least in low dimensional settings.
- They have limitations: The improvement in high-dimensional settings are very limited, sometimes not as good as previous methods.

# Take-home Message

- We introduce three classes of implementable maximal couplings of MH transition kernels.
- Their cost is roughly the same as the non-maximal couplings.
- They have advantages: easy to implement; can naturally generalize to other methods such as Metropolis-adjusted Langevin algorithm, pseudo-marginal Monte Carlo, MH-within-Gibbs; outperforms previous methods at least in low dimensional settings.
- They have limitations: The improvement in high-dimensional settings are very limited, sometimes not as good as previous methods.
- What we have learned:
  - Maximizing contraction rates (the whole process) may be more important than maximizing the single step meeting probability (every single step).

## Take-home Message

- Formal descriptions of algorithms, more theory and discussions are in our paper:

  "Maximal couplings of the Metropolis–Hastings algorithm." by O'Leary, Wang, and Jacob (2020).

- Code available on Github:

  https://github.com/johnoleary/mh-max-couplings

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

  Couplings and Monte Carlo

Thanks!

Pierre E Jacob.
Lecture notes: Couplings and monte carlo.

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