Maximal Couplings of the Metropolis–Hastings Algorithm

John O'Leary[1]
Harvard University

Guanyang Wang[1]
Rutgers University

Pierre E. Jacob
Harvard University

Abstract

Couplings play a central role in the analysis of Markov chain Monte Carlo algorithms and appear increasingly often in the algorithms themselves, e.g. in convergence diagnostics, parallelization, and variance reduction techniques. Existing couplings of the Metropolis–Hastings algorithm handle the proposal and acceptance steps separately and fall short of the upper bound on one-step meeting probabilities given by the coupling inequality. This paper introduces maximal couplings which achieve this bound while retaining the practical advantages of current methods. We consider the properties of these couplings and examine their behavior on a selection of numerical examples.

1 Introduction

Markov chain Monte Carlo (MCMC) methods offer a powerful framework for approximating integrals over a wide range of probability distributions (Brooks et al., 2011). The Metropolis–Hastings (MH) family of algorithms has proved to be especially popular, from its original forms (Metropolis et al., 1953; Hastings, 1970) to modern incarnations such as Hamiltonian Monte Carlo (Duane et al., 1987; Neal, 1993, 2011), the Metropolis-adjusted Langevin algorithm (Roberts and Tweedie, 1996), and particle MCMC (Andrieu et al., 2010). In many settings MH methods present an attractive mix of effectiveness, flexibility, and transparency.

Couplings of MCMC transition kernels have long played a role in the analysis and diagnosis of their con-

vergence (Rosenthal, 1995; Johnson, 1996, 1998; Jerrum, 1998; Rosenthal, 2002; Biswas et al., 2019), as a way of obtaining perfect samples or unbiased estimators (Propp and Wilson, 1996; Neal, 1999; Glynn and Rhee, 2014; Heng and Jacob, 2019; Jacob et al., 2020; Middleton et al., 2019, 2020), and as a variance reduction technique (Neal and Pinto, 2001; Goodman and Lin, 2009; Piponi et al., 2020). Such couplings are usually required to make the chains meet in finite time, with smaller meeting times associated with tighter bounds and greater precision or computational efficiency. In practice it is also essential for couplings to be implementable, in the sense that they require no extra knowledge about the target distribution beyond the requirements of the underlying MCMC algorithm.

We take up the question of coupling continuous state-space MH chains, following Johnson (1996, 1998) and Jacob et al. (2020). In Section 2 we define our setting and review existing methods. In Section 3 we introduce a set of implementable couplings that achieve the largest possible meeting probability at each iteration. These are the first known MH couplings with this property. We compare these algorithms with existing methods and introduce refinements that combine maximality with the benefits of the status quo. In Section 4 we apply our couplings to two numerical examples, gaining further insight into their properties and behavior. Finally, in Section 5 we consider open questions and next steps. Scripts in Python (Van Rossum and Drake, 2009) are available at https://github.com/johnoleary/mh-max-couplings.

2 Metropolis–Hastings Couplings

2.1 Setting and Definitions

We write $x \wedge y = \min(x, y)$, $x \vee y = \max(x, y)$, Unif for the uniform distribution on $[0, 1]$, and Bern(α) for the Bernoulli distribution on $\{0, 1\}$ with $\mathbb{P}(\text{Bern}(\alpha) = 1) = \alpha$.

Let P be a Markov transition kernel with stationary distribution π on $(\mathcal{X}, \mathscr{F}, \lambda)$, a Polish space equipped

---

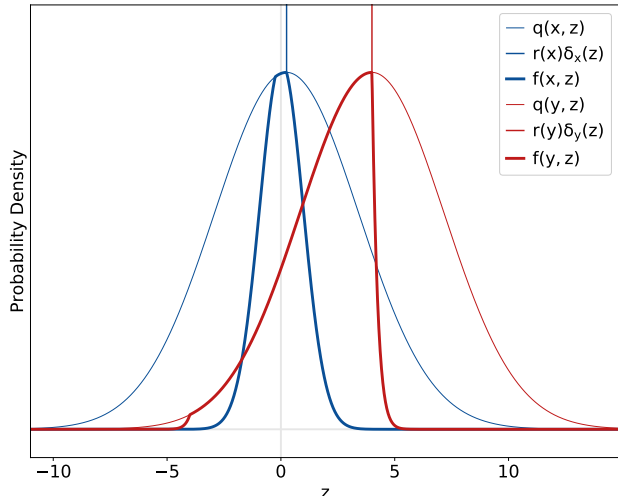[1] The first two authors contributed equally to this work.

Figure 1: Proposal densities q and MH transition densities f, with $\pi = \mathrm{N}(0,1)$, $x = 1/4$, $y = 4$, and $Q(z, \cdot) = \mathrm{N}(z, 10)$. $P(x, \cdot)$ and $P(y, \cdot)$ also contain point masses with weights $r(x) \approx 0.69$ and $r(y) \approx 0.47$, respectively. The algorithms described in this paper construct couplings of such transition kernels, sometimes by way of proposal distribution couplings.

with the standard Borel σ-algebra and some base measure λ. For $x \in \mathcal{X}$ and $A \in \mathcal{F}$, $P(x, A)$ denotes the probability of a transition from x to A. We focus on MH-like kernels P, characterized by the property that we can obtain $X \sim P(x, \cdot)$ by drawing a proposal $x' \sim Q(x, \cdot)$ and an acceptance indicator $B \sim \mathrm{Bern}(a(x, x'))$ and setting $X = Bx' + (1 - B)x$. We assume that for all $x \in \mathcal{X}$, $Q(x, \cdot)$ has density $q(x, \cdot)$ with respect to a base measure on $(\mathcal{X}, \mathcal{F})$. We also assume that the proposal distribution is non-atomic, so that $Q(x, \{y\}) = 0$ for all $x, y \in \mathcal{X}$.

The acceptance rate under MH will be $a(x, x') = 1 \wedge \frac{\pi(x')q(x',x)}{\pi(x)q(x,x')}$, and we allow for alternatives such as Barker's algorithm (Barker, 1965). For $x' \neq x$ we define $f(x, x') := q(x, x')a(x, x')$ and $r(x) := 1 - \int f(x, x')\lambda(\,\mathrm{d}x')$, so that $P(x, \cdot)$ has density $f(x, x')$ except for an atom where $P(x, \{x\}) = r(x)$. See Figure 1 for an illustration of a pair of proposal and transition distributions, and see Table 1 for convenient reference on the main functions and algorithms defined over the course of this paper.
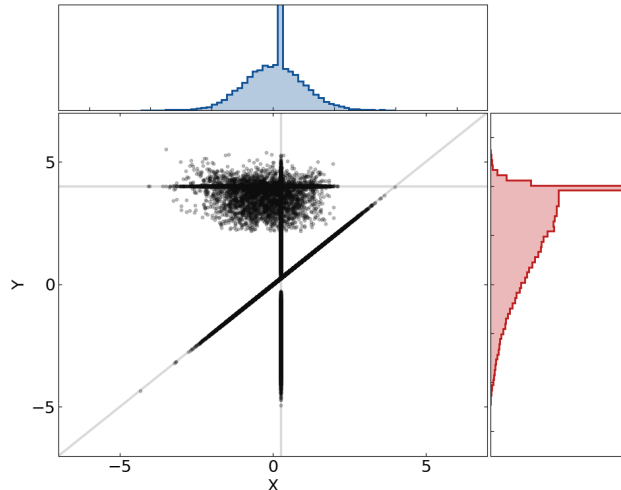

Figure 2: Draws from the coupling $\bar{P}_{\mathrm{SQ}}((x, y), \cdot)$ using $\bar{Q}_{\mathrm{MI}}$ and the same parameters as in Figure 1. The grey lines indicate $X = x$, $Y = y$, and $X = Y$, while the histograms show the marginal distributions of X and Y.

A probability distribution γ on $\mathcal{X} \times \mathcal{X}$ is a coupling of distributions μ and ν on $(\mathcal{X}, \mathcal{F})$ if $\gamma(A \times \mathcal{X}) = \mu(A)$ and $\gamma(\mathcal{X} \times A) = \nu(A)$ for any $A \in \mathcal{F}$. Let $\Gamma(\mu, \nu)$ be the set of all couplings of μ and ν. We say that a joint kernel $\bar{P}$ is a coupling of P with itself and write $\bar{P} \in \Gamma(P, P)$ if $\bar{P}((x, y,), \cdot) \in \Gamma(P(x, \cdot), P(y, \cdot))$ for any $x, y \in \mathcal{X}$. Similar definitions apply to couplings $\bar{Q}$ of proposal distributions and couplings $\bar{B}$ of acceptance indicators.

The coupling inequality (Levin et al., 2017, Proposition 4.7) states that the meeting probability $\mathbb{P}(X = Y)$ is less than $1 - \|\mu - \nu\|_{\mathsf{TV}}$ under any coupling $\gamma \in \Gamma(\mu, \nu)$, where $\|\mu - \nu\|_{\mathsf{TV}} = \sup_{A \in \mathcal{F}} |\mu(A) - \nu(A)|$ is the total variation distance. A coupling that achieves this bound is said to be maximal, and we write $\Gamma^{\max}(\mu, \nu) \subset \Gamma(\mu, \nu)$ for the set of maximal couplings of μ and ν.

2.2 Status Quo: the Heuristic Coupling $\bar{P}_{\mathrm{SQ}}$

We begin by describing the state-of-the-art coupling of MH transition kernels, first introduced in Johnson (1998). Recall that draws from a MH-like ker-

Table 1: Notation Reference

Functions		Couplings	
$q(x, y)$	$\frac{\mathrm{d}Q(x, \cdot)}{\mathrm{d}\lambda}(y)$	$\bar{P}_{SQ}$	Algorithm 1
$f(x, y)$	$q(x, y)a(x, y)$	$\bar{Q}_{MI}$	Algorithm 2
$f_{xy}^m(x')$	$f(x, x') \wedge f(y, x')$	$\bar{P}_{MI}$	Algorithm 3
$\tilde{f}_{xy}^r(x')$	$f(x, x') - f_{xy}^m(x')$	$\bar{P}_{MR}$	Algorithm 4
$q_{xy}^m(x')$	$\frac{\mathrm{d}\bar{Q}((x,y),\cdot)|_\Delta}{\mathrm{d}\lambda|_\Delta}(x')$	$\bar{P}_C$	Algorithm 5

nel $P(x, \cdot)$ can be obtained via a proposal $x' \sim Q(x, \cdot)$ which is accepted as a realization for X with probability $a(x, x')$. Thus a simple way to couple P with itself is to draw coupled proposals $(x', y') \sim \bar{Q}((x, y), \cdot)$ and accept or reject these according to coupled indicators $(B_x, B_y) \sim \bar{B}((x, y), (x', y'))$, where $\bar{Q} \in \Gamma(Q, Q)$ and $\bar{B}((x, y), (x', y')) \in \Gamma(\mathrm{Bern}(a(x, x')), \mathrm{Bern}(a(y, y')))$. We refer to this coupling as $\bar{P}_{\mathrm{SQ}}$ and summarize it in Algorithm 1.

For the chains to meet in finite time we need $\mathbb{P}(X = Y \mid x, y) > 0$ from at least some states (x, y), which in turn requires $\mathbb{P}(x' = y' \mid x, y) > 0$. To obtain this we take the joint proposal distribution $\bar{Q}$ to be a maximal coupling of Q with itself. We can draw from such a coupling by sampling $x' \sim Q(x, \cdot)$, using this as a rejection sampling proposal for $Q(y, \cdot)$, and drawing y' in a specified way if this $x' = y'$ proposal is rejected. This approach achieves the coupling inequality upper bound $\mathbb{P}(x' = y' \mid x, y) = 1 - \int q(x, z) \wedge q(y, z) \lambda(\mathrm{d}z)$.

A simple example of this method is the maximal coupling with independent residuals $\bar{Q}_{\mathrm{MI}}$, introduced in Vaserstein (1969) and called the γ-coupling in Lindvall (1992, chap. 1.5). It has the property that x' and y' are independent when $x' \neq y'$. Algorithm 2 describes how to draw from this coupling, and Figure 2 illustrates a set of draws $(X, Y) \sim \bar{P}_{\mathrm{SQ}}((x, y), \cdot)$ based on proposals from $\bar{Q}_{\mathrm{MI}}$.

The maximal coupling with reflection residuals, $\bar{Q}_{\mathrm{MR}}$, is often a better alternative when $\mathcal{X} = \mathbb{R}^d$ and Q is spherically symmetric in the sense that $q(x, y) = g(\|x - y\|)$ for some function g. This coupling was introduced in the context of Hamiltonian and Langevin methods (Bou-Rabee et al., 2020; Eberle et al., 2019) and has its origins in the analysis of continuous-time processes (Lindvall and Rogers, 1986; Eberle, 2011). $\bar{Q}_{\mathrm{MR}}$ is identical to $\bar{Q}_{\mathrm{MI}}$ for $x' = y'$. When $x' \neq y'$ the coupling $\bar{Q}_{\mathrm{MR}}$ sets $y' = T_{xy}(x')$, where $T_{xy}(x') = y + (I - 2ee')(x' - x)$ and $e = (y - x)/\|y - x\|$. We define $T_{yx}(y')$ similarly. Note that the transformation $z \mapsto (I - 2ee')z$ reflects the e component of z while leaving the $e^{\perp}$ component fixed.

A standard choice for the acceptance indicators $(B_x, B_y) \sim \bar{B}((x, y), (x', y'))$ is the unique maximal coupling of $\mathrm{Bern}(a(x, x'))$ and $\mathrm{Bern}(a(y, y'))$, which can be realized by drawing $U \sim \mathrm{Unif}$ and setting $B_x = \mathbf{1}(U \leq a(x, x'))$ and $B_y = \mathbf{1}(U \leq a(y, y'))$.

Algorithm 1 Draw $(X, Y) \sim \bar{P}_{\mathrm{SQ}}((x, y), \cdot)$

1. Draw $(x', y') \sim \bar{Q}((x, y), \cdot)$
2. Draw $(B_x, B_y) \sim \bar{B}((x, y), (x', y'))$
3. Set $X = B_x\, x' + (1 - B_x)x$ and $Y = B_y\, y' + (1 - B_y)y$
4. Return (X, Y)

Algorithm 2 Draw $(x', y') \sim \bar{Q}_{\mathrm{MI}}((x, y), \cdot)$

1. Draw $x' \sim Q(x, \cdot)$ and $U \sim \mathrm{Unif}$
2. If $Uq(x, x') \leq q(y, x')$, set $y' = x'$
3. Else
    (a) Draw $\tilde{y} \sim Q(y, \cdot)$ and $V \sim \mathrm{Unif}$
    (b) If $Vq(y, \tilde{y}) > q(x, \tilde{y})$, set $y' = \tilde{y}$
    (c) Else go to 3(a)
4. Return (x', y')

Among couplings of these distributions, this one yields the maximal probability $\mathbb{P}(B_x = B_y = 1 \mid x', y') = a(x, x') \wedge a(y, y')$. Other objectives, such as the minimization of $\mathbb{E}[\|X - Y\| \mid x', y']$, are also possible.

The coupling $\bar{P}_{\mathrm{SQ}}$ obtained when $\bar{Q}$ is a maximal coupling of proposal distributions and $\bar{B}$ is the maximal coupling of acceptance indicators can yield a relatively high chance of $X = Y$ given the current state pair (x, y), but it typically falls short of the theoretical upper bound. Under $\bar{P}_{\mathrm{SQ}}$, the probability of $X = Y$ is $\int (q(x, z) \wedge q(y, z))(a(x, z) \wedge a(y, z)) \lambda(\mathrm{d}z)$. However Lemma 1 implies that the coupling inequality bound is $\int (q(x, z)a(x, z)) \wedge (q(y, z)a(y, z)) \lambda(\mathrm{d}z)$, which is always an equal or larger quantity. Figure 3 illustrates the gap between the meeting probabilities under $\bar{P}_{\mathrm{SQ}}$ and under any maximal coupling $\bar{P}$. Note that these probabilities will coincide when either $q(x, z)$ or $a(x, z)$ does not depend on x, e.g. for the independence sampler.

Lemma 1. *Let P be an MH-like transition kernel as defined above. Then for $x \neq y$, $\|P(x, \cdot) - P(y, \cdot)\|_{\mathsf{TV}} = 1 - \int f(x, z) \wedge f(y, z) \lambda(\mathrm{d}z)$.*

Proof of Lemma 1. In the spirit of e.g. Lindvall (1992), Chapter I.5, let

$$C_{xy} = \{z : f(x, z) > f(y, z)\} \cup \{x\} \setminus \{y\}$$

and similarly for C_{yx}. We have $\|P(x, \cdot) - P(y, \cdot)\|_{\mathsf{TV}} = \sup_A |P(x, A) - P(y, A)|$, and $|P(x, A) - P(y, A)| = |\int_A f(x, z) - f(y, z)\, \mathrm{d}z + \mathbf{1}(x \in A)r(x) - \mathbf{1}(y \in A)r(y)|$. We must have either $A = C_{xy}$ or $A = C_{yx}$ in the supremum above. Both yield the same value, and so $\|P(x, \cdot) - P(y, \cdot)\|_{\mathsf{TV}} = 1 - \int f(x, z) \wedge f(y, z)\, \mathrm{d}z$. $\qquad \square$

We want to maximize the coupling probability at each step and reduce typical meeting times, so the factors above lead us to ask if there are couplings $\bar{P}$ which are both implementable and maximal. Our contribution answers this question in the affirmative.

3 Maximal Couplings

We now introduce a collection of implementable couplings $\bar{P} \in \Gamma^{\mathrm{max}}(P, P)$ starting from an arbitrary MH-like transition kernel P. We consider two approaches.
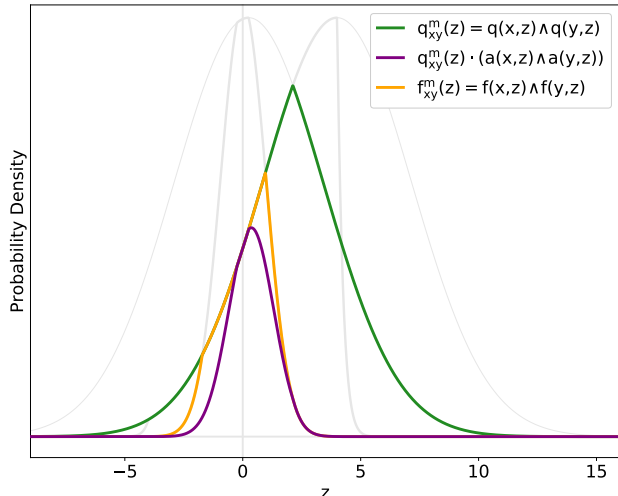
Figure 3: Meeting densities based on the same parameters as in Figure 1. The green line shows the density of $x' = y' = z$ under a maximal coupling $\bar{Q}$. The purple line shows the density of $X = Y = z$ under $\bar{P}_{\mathrm{SQ}}$ assuming the use of a maximal coupling $\bar{B}$. Finally, the orange line shows the density of $X = Y = z$ under a maximal coupling $\bar{P}$. The inequalities suggested here hold in general.
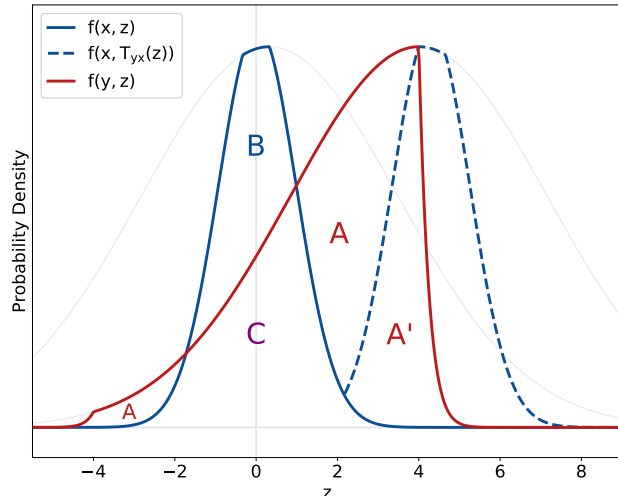

Figure 4: Rejection sampling regions for Algorithms 3 and 4 based on the same parameters as Figure 1. For both options, meeting occurs on samples drawn from region C.

In Sections 3.1 and 3.2 we show how to couple Markov transition kernels without explicitly coupling their underlying proposal or acceptance distributions. We refer to these as full-kernel couplings. In Section 3.3, we modify Algorithm 1 to maximize the probability of accepting proposals $x' = y'$ at the expense of decreasing the acceptance probabilities when $x' \neq y'$.

3.1 $\bar{P}_{\mathrm{MI}}$, a Full-kernel Coupling with Independent Residuals

Our first full-kernel coupling, which we will refer to as $\bar{P}_{\mathrm{MI}}$, is inspired by the procedure described in Algorithm 2 for drawing from the coupling $\bar{Q}_{\mathrm{MI}}$ of proposal distributions. A key difference between coupling proposal distributions and transition kernels is that by assumption the proposal distributions are non-atomic and absolutely continuous with respect to an underlying measure, while MH-like kernels P can have a point mass at the current state. Therefore our rejection sam-

---

Algorithm 3 Draw $(X, Y) \sim \bar{P}_{\mathrm{MI}}((x, y), \cdot)$

---

1. Draw $X \sim P(x, \cdot)$ and $U \sim \mathrm{Unif}$
2. If $X \neq x$ and $U\, f(x, X) \leq f(y, X)$, set $Y = X$
3. Else
   (a) Draw $\tilde{y} \sim P(y, \cdot)$ and $V \sim \mathrm{Unif}$
   (b) If $\tilde{y} = y$, set $Y = \tilde{y}$
   (c) If $\tilde{y} \neq y$ and $V\, f(y, \tilde{y}) > f(x, \tilde{y})$, set $Y = \tilde{y}$
   (d) Else go to 3(a)
4. Return (X, Y)

---

pling procedure must be modified to yield draws X and Y with the correct marginal distributions.

Algorithm 3 contains this modification, which we illustrate in Figure 4. The blue and red curves give the continuous parts of $P(x, \cdot)$ and $P(y, \cdot)$, respectively. We think of the algorithm as sampling uniformly from the region under the graph along with point masses at x and y. If the X draw falls in C we can use the same point for both chains, and otherwise we use rejection sampling to obtain a Y draw from $A \cup A' \cup \{y\}$. Meetings occur exactly on draws from C, and thus it occurs with probability $\int f(x, z) \wedge f(y, z) \lambda(\mathrm{d}z)$. See Appendix A.1 for a detailed proof that $\bar{P}_{\mathrm{MI}}((x, y), \cdot)$ is a maximal coupling of $P(x, \cdot)$ and $P(y, \cdot)$ and an analysis of the computation cost of this algorithm.

3.2 $\bar{P}_{\mathrm{MR}}$, a Full-kernel Coupling with Reflection Residuals

All maximal couplings of a given P have equal meeting probability, but some perform better than others. Although $\mathbb{P}(X = Y \mid x, y)$ is maximal under $\bar{P}_{\mathrm{MI}}((x, y), \cdot)$, X and Y are independent when $X \neq Y$. This creates a strong tendency for $\|Y - X\|$ to grow with the dimension of the state space, as seen in the experiments of Jacob et al. (2020). Even under a maximal coupling, $\mathbb{P}(X = Y \mid x, y)$ can be small until x and y are close. Thus the time for a pair of coupled chains to meet depends on both the meeting probability from each state pair and the degree of contraction between chains when meeting does not occur.

The advantages of using $\bar{Q}_{\mathrm{MR}}$ vs. $\bar{Q}_{\mathrm{MI}}$ in Algorithm 1 appear to be due to this consideration. See Section 4.2 below for numerical justifications.

Algorithm 4 Draw $(X, Y) \sim \bar{P}_{\mathrm{MR}}((x, y), \cdot)$

---

1. Draw $X \sim P(x, \cdot)$ and $U \sim \mathrm{Unif}$
2. If $X \neq x$ and $U f(x, X) \leq f(y, X)$, set $Y = X$
3. Else
   (a) Set $\tilde{y} = T_{xy}(X)$ and draw $V \sim \mathrm{Unif}$
   (b) If $X \neq x$ and $V \tilde{f}^r_{xy}(X) \leq \tilde{f}^r_{yx}(\tilde{y})$, set $Y = \tilde{y}$
   (c) Else
      i. Draw $\tilde{y} \sim P(y, \cdot)$ and $W \sim \mathrm{Unif}$
      ii. If $\tilde{y} = y$, set $Y = \tilde{y}$
      iii. If $\tilde{y} \neq y$ and $W f(y, y') \leq \tilde{f}^t_{yx}(\tilde{y})$, set $Y = \tilde{y}$
      iv. Else go to 3(c)i.
4. Return (X, Y)

It also motivates the following full-kernel maximal coupling with reflection residuals, which we refer to as $\bar{P}_{\mathrm{MR}}$ and describe in detail in Algorithm 4. For this algorithm define $f^m_{xy}(z) := f(x, z) \wedge f(y, z)$, $\tilde{f}^r_{xy}(x') := f(x, x') - f^m_{xy}(x')$, and likewise for $\tilde{f}^r_{yx}(y')$. Finally set $\tilde{f}^t_{yx}(y') := \tilde{f}^r_{yx}(y') - \tilde{f}^r_{yx}(y') \wedge \tilde{f}^r_{xy}(T_{yx}(y'))$, the y residual after reflection evaluated at y'.

Figure 4 provides some intuition into the behavior of $\bar{P}_{\mathrm{MR}}$. The first step of $\bar{P}_{\mathrm{MR}}$ is identical to $\bar{P}_{\mathrm{MI}}$ in that it attempts to draw from the region C, and a meeting occurs if this is successful. Otherwise $\bar{P}_{\mathrm{MR}}$ proposes the reflected point $T_{xy}(X)$ for Y, which succeeds if this point falls in the region A'. If all else fails, we use rejection sampling to obtain a Y draw from $A \cup \{y\}$. See Appendix A.2 for a detailed proof of the validity and maximality, and an analysis of the computation cost of $\bar{P}_{\mathrm{MR}}$.

3.3 Maximal Coupled Transitions from Maximally Coupled Proposals with $\bar{P}_{\mathrm{C}}$

While the coupling $\bar{P}_{\mathrm{MR}}$ sometimes outperforms $\bar{P}_{\mathrm{MI}}$, it also has a few limitations. First, the reflection proposal strategy requires a high degree of reflection symmetry between the distributions of X and Y conditional on $X \neq Y$ to be successful. For example in the setting of Figure 4, the reflection proposal has more than a 50% chance of being rejected. Without such a symmetry X and Y will tend to be conditionally independent, resulting in poor contraction between chains when meeting does not occur.

Second, any full-kernel coupling is constrained to work directly with the complicated and irregular geometry of P rather than the simple and often tractable form of the proposal distribution Q. In contrast to the range of couplings and optimal transport strategies available when a standard distribution is used for Q, it appears to be more difficult to design high-performance couplings directly in terms of the associated transition kernels P.
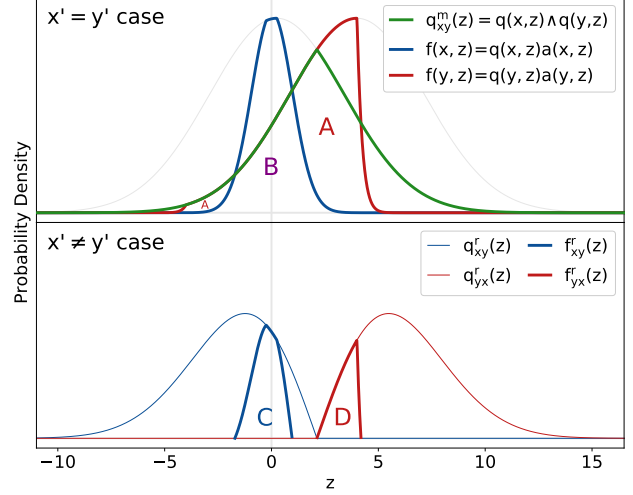

Figure 5: Distributions used in Algorithm 5 based on the same parameters as Figure 1. Upper pane: draws $x' = y'$ follow q^m and are used as proposals for $f(x, \cdot)$ and $f(y, \cdot)$. Lower pane: draws $x' \neq y'$ follow the proposal residuals q^r and are used as rejection sampling proposals for the transition kernel residuals f^r.

This motivates the next coupling, which allows the use of any coupling of Q while still possibly achieving the maximal coupling probability identified in Lemma 1. We refer to this new algorithm as $\bar{P}_{\mathrm{C}}$, since it resembles $\bar{P}_{\mathrm{SQ}}$ up to the conditional use of one or another Bernoulli coupling $\bar{B}$ depending on whether or not a meeting is proposed. Together with the definitions below, Algorithm 5 shows how to draw from $\bar{P}_{\mathrm{C}}$. We can choose a maximal $\bar{Q}$ such as $\bar{Q}_{\mathrm{MI}}$ or $\bar{Q}_{\mathrm{MR}}$ to obtain a maximal $\bar{P}$, although we also obtain a valid $\bar{P} \in \Gamma(P, P)$ for non-maximal choices of $\bar{Q}$.

To define $\bar{P}_{\mathrm{C}}$, let $q^m_{xy}(\cdot)$ be the density of $\bar{Q}((x, y), \cdot)$ on $\Delta = \{(z, z) : z \in \mathcal{X}\}$. We establish the existence and properties of q^m_{xy} in Lemma 2, below. Define the proposal residual $q^r_{xy}(x') := q(x, x') - q^m_{xy}(x')$ and the transition kernel residual $f^r_{xy}(x') := 0 \vee (f(x, x') - q^m_{xy}(x'))$, and similarly for $q^r_{yx}(y')$ and $f^r_{yx}(y')$. We illustrate these functions in Figure 5.

Lemma 2. *If $\bar{Q} \in \Gamma(Q, Q)$ then there exists a density q^m_{xy} for $\bar{Q}((x, y), \cdot)$ on Δ. If $\bar{Q}$ is a maximal coupling, then $q^m_{xy}(z) = q(x, z) \wedge q(y, z)$ for almost all z.*

Algorithm 5 Draw $(X, Y) \sim \bar{P}_{\mathrm{C}}((x, y), \cdot)$

---

1. Draw $(x', y') \sim \bar{Q}((x, y), \cdot)$
2. If $x' = y'$ set $\bar{B} = \bar{B}_1$, else set $\bar{B} = \bar{B}_2$
3. Draw $(B_x, B_y) \sim \bar{B}((x, y), (x', y'))$
4. Set $X = B_x x' + (1 - B_x) x$ and $Y = B_y y' + (1 - B_y) y$
5. Return (X, Y)

Proof of Lemma 2. Let λ denote the base measure on $(\mathcal{X}, \mathscr{F})$ and let λ_Δ be its push-forward to Δ by the map $z \mapsto (z, z)$. Since $\mathcal{X}$ is a Polish space, we have $A_\Delta := \{(z, z) : z \in A\} \in \mathscr{F} \otimes \mathscr{F}$ for any $A \in \mathscr{F}$. Thus $\bar{Q}((x, y), \cdot)$ induces a sub-probability $\bar{Q}_\Delta$ on Δ. Also $\bar{Q}_\Delta \ll \lambda_\Delta$, since if $\lambda_\Delta(A_\Delta) = \lambda(A) = 0$ then $\bar{Q}_\Delta(A_\Delta) \leq \bar{Q}((x, y), A \times \mathcal{X}) = Q(x, A) = 0$. The Radon-Nikodym theorem then guarantees the existence of an integrable function $q_{xy}^m : \mathcal{X} \to [0, \infty)$ such that $\bar{Q}((x, y), A_\Delta) = \int_A q_{xy}^m(z)\, \mathrm{d}z$.

Next, we claim $q_{xy}^m(z) \leq q(x, z) \wedge q(y, z)$ for λ-almost all z. Let $A := \{z : q_{xy}^m(z) > q(x, z) \wedge q(y, z)\} \in \mathscr{F}$, $A_x := \{z : q_{xy}^m(z) > q(x, z)\} \in \mathscr{F}$, and likewise for A_y. Since $A = A_x \cup A_y$, $\lambda(A) > 0$ implies $\lambda(A_x) > 0$ or $\lambda(A_y) > 0$. If $\lambda(A_x) > 0$ then $\bar{Q}((x, y), A_x \times \mathcal{X}) = Q(x, A_x) < \bar{Q}((x, y), (A_x)_\Delta) \leq \bar{Q}((x, y), A_x \times \mathcal{X})$, a contradiction. The case $\lambda(A_y) > 0$ similarly implies a contradiction, so we conclude that $\lambda(A) = 0$.

Finally, if $\bar{Q}$ is maximal, a total variation computation similar to that of Lemma 1 shows that $\bar{Q}((x, y), \Delta) = \int q(x, z) \wedge q(y, z)\, \mathrm{d}z$. Combining this with the above implies $q_{xy}^m(z) = q(x, z) \wedge q(y, z)$ for λ-almost all z. $\square$

Resuming our definition of $\bar{P}_\mathrm{C}$, we set acceptance probabilities $b_{xy}(x') := 1 \wedge (f(x, x')/q_{xy}^m(x'))$ if $q_{xy}^m(x') > 0$ or else $b_{xy}(x') := 1$, $c_{xy}(x') := f_{xy}^r(x')/q_{xy}^r(x')$ if $q_{xy}^r(x') > 0$ or else $c_{xy}(x') := 1$, and likewise for $b_{yx}(y')$ and $c_{yx}(y')$. When $x' = y'$ we require that the acceptance indicator pair (B_x, B_y) follows the maximal coupling $\bar{B}_1$ of $\mathrm{Bern}(b_{xy}(x'))$ and $\mathrm{Bern}(b_{yx}(y'))$. When $x' \neq y'$ we require that (B_x, B_y) follows any coupling $\bar{B}_2 \in \Gamma(\mathrm{Bern}(c_{xy}(x')), \mathrm{Bern}(c_{yx}(y')))$. Simulation results suggest that maximal couplings for $\bar{B}_2$ perform well, but optimal transport couplings that aim to minimize $\|X - Y\|$ are also attractive in this setting.

When $\bar{Q}$ proposes a meeting $x' = y'$, we think of $\bar{P}_\mathrm{C}$ as using these values as rejection sampling proposals for the transition distributions $f(x, z)$ and $f(y, z)$. This results in a higher marginal acceptance rate than we would have under MH. On the other hand, when $x' \neq y'$ we think of this method as falling back to rejection sampling of the residual distributions f from the residual distributions of q, both after the removal of q_{xy}^m. This produces a lower marginal acceptance rate than MH, exactly counterbalancing the above. In Proposition 1 we show that $\bar{P}_\mathrm{C}((x, y), \cdot)$ is a maximal coupling of $P(x, \cdot)$ and $P(y, \cdot)$.

Proposition 1. *For any $\bar{Q} \in \Gamma(Q, Q)$, the output (X, Y) of Algorithm 5 will follow a coupling $\bar{P} \in \Gamma(P, P)$. If $\bar{Q}$ is maximal then $\bar{P}$ will be as well.*

Proof. At any point $x' \neq x$, X will have density

$$q_{xy}^m(x')b_{xy}(x') + q_{xy}^r(x')c_{xy}(x')$$
$$= \left(q_{xy}^m(x') \wedge f(x, x')\right) + \left(q_{xy}^r(x') \wedge f_{xy}^r(x')\right)$$
$$= \left(q_{xy}^m(x') \wedge f(x, x')\right) + f_{xy}^r(x') = f(x, x').$$

The second equality holds because $f_{xy}^r(x') \leq q_{xy}^r(x')$. This in turn holds because $f(x, x') \leq q(x, x')$ and $q_{xy}^m(x') \leq q(x, x')$ by Lemma 2, so

$$f_{xy}^r(x') = \left(f(x, x') \vee q_{xy}^m(x')\right) - q_{xy}^m(x')$$
$$\leq q(x, x') - q_{xy}^m(x') = q_{xy}^r(x').$$

Integrating the density of X over all $x' \neq x$ yields $\mathbb{P}(X = x) = 1 - \int f(x, x')\, \mathrm{d}x' = r(x)$, so we conclude $X \sim P(x, \cdot)$. A similar argument shows that $Y \sim P(y, \cdot)$. Thus (X, Y) follows the desired coupling.

If $\bar{Q}$ is maximal then by Lemma 2 the probability density at the proposal $(z, z) \sim \bar{Q}((x, y), \cdot)$ will be $q_{xy}^m(z) = q(x, z) \wedge q(y, z)$. By the definition of $\bar{B}_1$, the probability of accepting a proposal $x' = y' = z$ for both X and Y will be

$$\left(1 \wedge \frac{f(x, z)}{q_{xy}^m(z)}\right) \wedge \left(1 \wedge \frac{f(y, z)}{q_{yx}^m(z)}\right) = \frac{f(x, z) \wedge f(y, z)}{q(x, z) \wedge q(y, z)}.$$

Combining this with the proposal density implies that the overall coupled transition kernel density at $X = Y = z$ will be $f(x, z) \wedge f(y, z)$. Thus $\mathbb{P}(X = Y) = \int f(x, z) \wedge f(y, z)\lambda(\mathrm{d}z)$, and by Lemma 1 we conclude that $\bar{P}$ is maximal. $\square$

We observe that $\bar{P}_\mathrm{C}$ matches the flexibility and computational efficiency of $\bar{P}_\mathrm{SQ}$ while offering a higher meeting probability at each iteration. The ability to chose an arbitrary $\bar{Q} \in \Gamma(Q, Q)$ is also a significant advantage of $\bar{P}_\mathrm{C}$ over the full-kernel couplings. The extra effort required to construct, validate, and draw from $\bar{P}_\mathrm{MR}$ relative to $\bar{P}_\mathrm{MI}$ shows how challenging such refinements can be. Finally, we note that $\bar{P}_\mathrm{C}$ can be more computationally efficient than the full-kernel couplings $\bar{P}_\mathrm{MI}$ and $\bar{P}_\mathrm{MR}$, in that it avoids the 'while' loops of Step 3 of Algorithms 3 and 4.

4 Numerical Examples

4.1 Biased Random Walk MH

For our first example we consider a toy model that emphasizes the differences between $\bar{P}_\mathrm{SQ}$ and the maximal couplings $\bar{P}_\mathrm{MI}$, $\bar{P}_\mathrm{MR}$, and $\bar{P}_\mathrm{C}$. We assume an Exponential target distribution $\pi = \mathrm{Expo}(1)$ and draw proposals from $Q(z, \cdot) = \mathrm{N}(z + \kappa, \sigma^2)$ with $\kappa > 0$. We then accept or reject these proposals at the usual MH rate, so $a(z, z') = 1 \wedge \mathbb{1}(z' \geq 0) \exp\left((z - z')(2\kappa/\sigma^2 + 1)\right)$. Our assumption on κ implies $a(x, z) \wedge a(y, z) = a(x, z)$

Table 2: Section 4.1 Example Results

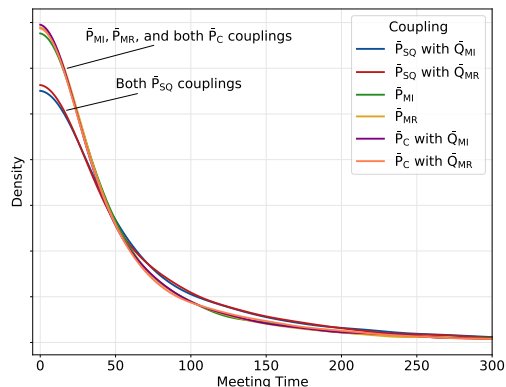| Coupling	Avg. Meeting Time	S.E.
$\bar{P}_{SQ}$ with $\bar{Q}_{MI}$	74.0	0.94
$\bar{P}_{SQ}$ with $\bar{Q}_{MR}$	75.6	0.99
$\bar{P}_{MI}$	60.5	0.84
$\bar{P}_{MR}$	60.9	0.87
$\bar{P}_C$ with $\bar{Q}_{MI}$	61.3	0.87
$\bar{P}_C$ with $\bar{Q}_{MR}$	62.2	0.89

Figure 6: Distribution of meeting times for the example described in Section 4.1. The four maximal couplings yield almost identical distributions of shorter meeting times times while the two $\bar{P}_{SQ}$ methods yield almost identical distribution of longer ones. In this example the requirement that $\bar{P}_{SQ}$ accept meeting and non-meeting proposals at exactly the MH rate puts it at a significant disadvantage vs. the maximal couplings.

if $x \leq y$ and $a(x,z) = 1$ if $z \leq x$. Thus Q will tend to propose increasing values while a favors decreasing ones. This tension between the proposal and target distributions is characteristic of MH kernels that do not mix rapidly.

We construct the transition kernel couplings $\bar{P}_{SQ}$ with $\bar{Q}_{MI}$, $\bar{P}_{MI}$, $\bar{P}_{MR}$, and $\bar{P}_C$ with $\bar{Q}_{MI}$ as described in Sections 2 and 3. We also define a simple generalization of the maximal coupling with reflection residuals, such that either $x' = y'$ or $y' - (y + \kappa) = (x + \kappa) - x'$. We use the resulting $\bar{Q}_{MR}$ as an alternative proposal kernel coupling for $\bar{P}_{SQ}$ and $\bar{P}_C$. We set $\kappa = \sigma^2 = 3$, draw initial values independently from the target π, and run 10,000 replications for each coupling option. For each replication we record the meeting time $\tau = \min(t \geq 0 : X_t = Y_t)$. As described in Section 1, such meeting times are of theoretical and practical importance, and they also make a good measure of coupling performance. We summarize the average behavior of these meeting times in Table 2 and present their full distribution in Figure 6.

We find that both non-maximal couplings deliver average meeting times around 75 iterations, while the four maximal couplings deliver meeting times around 61 iterations. We recall that for a given state pair (x,y), the maximal couplings produce one value of $\mathbb{P}(X = Y \mid x,y)$ and two $\bar{P}_{SQ}$ couplings produce another. The observed clustering of algorithms is consistent with the idea that in this example meeting times are driven by these one-step meeting probabilities rather than by behavior when meeting does not occur, which varies significantly by algorithm.

For a better understanding of these differences, we contrast the behavior of $\bar{P}_{SQ}$ and $\bar{P}_C$. Although we use the same underlying proposal coupling $(x', y') \sim \bar{Q}((x,y), \cdot)$ in each case, the two MH transition kernel couplings differ in that $\bar{P}_{SQ}$ accept its proposals at exactly the MH rate while $\bar{P}_C$ uses a higher acceptance probability when $x' = y'$ and a lower one when $x' \neq y'$. In this example, most proposals have a relatively low MH acceptance probability to begin with. Thus $\bar{P}_C$ meets more quickly by concentrating the little acceptance probability available on the draws

where they could result in a meeting $X = Y$, while $\bar{P}_{SQ}$ is forced to accept its proposals at the same relatively low rate whether or not a meeting is proposed.

The above leads us to expect that maximal couplings might provide the greatest advantage over the status quo when low acceptance probabilities are typical, either due to the presence of a challenging target, an imperfect proposal distribution, or both. This example emphasizes simplicity over realism, but we would expect its principles to hold more broadly, especially in cases where mixing is relatively slow.

4.2 Dimension Scaling with a Normal Target

For our second example we consider MH on $\mathbb{R}^d$ with a target distribution $\pi = \mathrm{N}(0, I_d)$ and proposals $Q(z, \cdot) = \mathrm{N}(z, I_d \sigma_d^2)$. Following e.g. Roberts et al. (1997) and Christensen et al. (2005) we set $\sigma_d^2 = \ell^2/d$ with $\ell = 2.38$. This example allows us to consider the role of the dimension and examine differences between the couplings when meeting probabilities are just one important aspect of their behavior.

As above, our main object of interest in this example is the number of iterations required for a pair of coupled chains to meet. We initialize chains on independent draws from π, consider dimensions $d = 1, \ldots, 10$, and run 1,000 replications for each algorithm. We use a maximal coupling of acceptance indicators for $\bar{P}_{SQ}$ and $\bar{P}_C$, which appears to yield the best results among the simple acceptance indicator couplings.
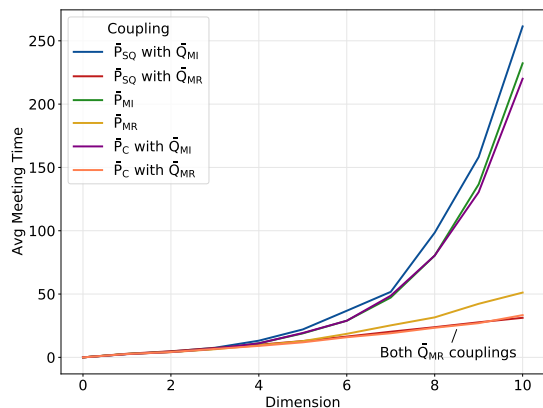
We present the results of this experiment in Figure 7.

Figure 7: Average meeting times for a range of transition kernel couplings, as described in Section 4.2. Coupling strategies that involve reflections of the proposals appear to outperform the others; among the others, maximal couplings seem to perform better than non-maximal ones.
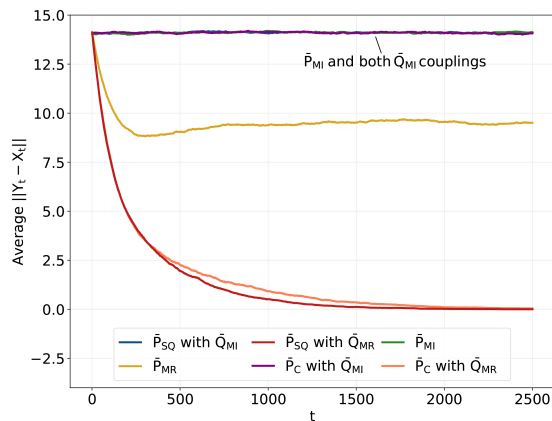

Figure 8: Distance between coupled MH chains by iteration, as described in Section 4.2. Couplings in which $X \neq Y$ implies X, Y independent or almost independent display little contraction, while those based on the maximal reflection coupling of proposal distributions $\bar{Q}_{\mathrm{MR}}$ display strong contraction.

There we observe that $\bar{P}_{\mathrm{SQ}}$ using $\bar{Q}_{\mathrm{MI}}$, $\bar{P}_{\mathrm{C}}$ using $\bar{Q}_{\mathrm{MI}}$, and $\bar{P}_{\mathrm{MI}}$ seem to yield meeting times that increase exponentially in dimension, although the maximal couplings outperform the non-maximal coupling. This blow-up is expected since these options involve independent or weakly dependent behavior when $X \neq Y$. The coupling $\bar{P}_{\mathrm{MR}}$ delivers somewhat better behavior, with $\bar{P}_{\mathrm{SQ}}$ and $\bar{P}_{\mathrm{C}}$ with $\bar{Q}_{\mathrm{MR}}$ delivering the best performance. In higher dimensions it appears that the $\bar{Q}_{\mathrm{MR}}$ version of $\bar{P}_{\mathrm{SQ}}$ may outperform its $\bar{P}_{\mathrm{C}}$ counterpart, an interesting and perhaps counterintuitive result.

To understand these differences in meeting times, we must consider what happens under each coupling $\bar{P}$ when a meeting does not occur. For $Q(z, \cdot) = \mathrm{N}(z, I_d \sigma_d^2)$, any $\bar{Q} \in \Gamma^{\max}(Q, Q)$ will yield $\mathbb{P}(x' = y' | x, y) = \mathbb{P}(\chi_1^2 \geq \|y - x\|^2 / (4\sigma_d^2))$, see e.g. Pollard (2005, chap. 3.3). Since a meeting cannot occur unless one is proposed, this expression is an upper bound on $\mathbb{P}(X = Y \mid x, y)$. Thus we should expect the probability of $X = Y$ to fall off rapidly in $\|y - x\|$, so that couplings that do not promote contraction between chains in the absence of meeting will yield rapidly increasing meeting times.

Figure 8 is consistent with these observations. As above, we run 1,000 replications per algorithm, initializing each chain with an independent draw from the target distribution. We set $d = 100$ to isolate the effects of contraction due to meeting from other contractive behavior and then track the average distance between chains as a function of iteration.

We find that $\bar{P}_{\mathrm{MI}}$, $\bar{P}_{\mathrm{SQ}}$ with $\bar{Q}_{\mathrm{MI}}$, and $\bar{P}_{\mathrm{C}}$ with $\bar{Q}_{\mathrm{MI}}$ produce little or no contraction within the distance

achieved by independent draws from the target. They then remain far enough apart that the probability of meeting is negligible. The coupling $\bar{P}_{\mathrm{MR}}$ displays contraction up to a point. Finally we observe that both the maximal and non-maximal couplings based on $\bar{Q}_{\mathrm{MR}}$ contract rapidly to within a radius where meeting can occur. $\bar{P}_{\mathrm{SQ}}$ appears to contract more rapidly than $\bar{P}_{\mathrm{C}}$, which is consistent with its slightly stronger performance in high dimensions.

5   Discussion

Couplings play a central role in the analysis of MCMC convergence and increasingly appear in new methods and estimators. Until now, no general-purpose algorithm has been available to sample from a maximal coupling of MH transition kernels. We fill this gap by introducing three such algorithms, which are implementable under the standard assumptions that one can draw proposals from a distribution Q and compute the density q and acceptance rate a at any points of the state space. The proposed couplings are simple to apply and can be used with a variety of MH strategies including the Metropolis-adjusted Langevin algorithm, pseudo-marginal methods. Our method is also straightforward to generalize to the case where the proposal kernel is the combination of several kernels, such as the MH-within-Gibbs algorithm.

The experiments in Section 4.1 show that the gains from using these methods can be large, especially when there is a tension between the proposal density q and the acceptance rate a. On the other hand, the example of Section 4.2 shows that maximality is sometimes less

important than other properties of a coupling, such as the contraction behavior when a meeting does not occur. The two examples considered here are simple ones, and experiments with a wider range of MH algorithms and target distributions would clarify the strengths and weaknesses of the proposed couplings.

This work raises several questions. First, it is not known if all couplings of MH kernels might be represented in the form of Algorithm 5 for some appropriate choice of proposal and acceptance couplings; see Nüsken and Pavliotis (2019) for the treatment of a similar question in the setting of continuous-time Markov chains. It would also be interesting to consider the use of sub-maximal proposal distribution couplings in Algorithm 5, as suggested in the comment of Gerber and Lee (2020) on Jacob et al. (2020).

Both meeting probabilities and contraction rates influence meeting times, and one might wonder about deriving maximally contractive couplings in analogy to the present work on meeting times. Reflection couplings seem particularly effective and are known to be optimal in special cases (Lindvall and Rogers, 1986). In other cases, synchronous or 'common random number' couplings yield strong contraction (e.g. Diaconis and Freedman, 1999). In most other scenarios the user must construct a coupling tailored to the problem at hand. The methods proposed here represent a step forward in coupling design, but many important questions remain.

From a more theoretical point of view, while the MH kernel P is known to be the projection of the proposal Q onto the set of π-reversible kernels in a certain metric (Billera and Diaconis, 2001), it is not known how $\bar{Q}$ relates to the set of maximal couplings of π-reversible kernels. In particular, it would be interesting to know if the strategy proposed in Section 3.3 corresponds to a projection of $\bar{Q}$ onto that set.

Finally, the coupling strategies mentioned above are all Markovian. In some cases non-Markovian couplings are known to deliver more satisfactory performance than Markovian ones (e.g. Smith, 2014). The design of practical non-Markovian couplings for MCMC is a topic deserving further attention.

Acknowledgments

References

Andrieu, C., Doucet, A., and Holenstein, R. (2010). Particle Markov chain Monte Carlo methods. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 72(3):269–342. 1

Barker, A. A. (1965). Monte Carlo calculations of the radial distribution functions for a proton–electron plasma. *Australian Journal of Physics*, 18(2):119–134. 2

Billera, L. J. and Diaconis, P. (2001). A geometric interpretation of the Metropolis–Hastings algorithm. *Statistical Science*, pages 335–339. 9

Biswas, N., Jacob, P. E., and Vanetti, P. (2019). Estimating convergence of Markov chains with L-lag couplings. In *Advances in Neural Information Processing Systems*, pages 7391–7401. 1

Bou-Rabee, N., Eberle, A., Zimmer, R., et al. (2020). Coupling and convergence for Hamiltonian Monte Carlo. *Annals of Applied Probability*, 30(3):1209–1250. 3

Brooks, S., Gelman, A., Jones, G., and Meng, X.-L. (2011). *Handbook of Markov chain Monte Carlo*. CRC press. 1

Christensen, O. F., Roberts, G. O., and Rosenthal, J. S. (2005). Scaling limits for the transient phase of local Metropolis–Hastings algorithms. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2):253–268. 7

Diaconis, P. and Freedman, D. (1999). Iterated random functions. *SIAM Review*, 41(1):45–76. 9

Duane, S., Kennedy, A. D., Pendleton, B. J., and Roweth, D. (1987). Hybrid Monte Carlo. *Physics Letters B*, 195(2):216–222. 1

Eberle, A. (2011). Reflection coupling and Wasserstein contractivity without convexity. *Comptes Rendus Mathematique*, 349(19-20):1101–1104. 3

Eberle, A., Guillin, A., and Zimmer, R. (2019). Couplings and quantitative contraction rates for Langevin dynamics. *The Annals of Probability*, 47(4):1982–2010. 3

Gerber, M. and Lee, A. (2020). Discussion on the paper by Jacob, O'Leary, and Atchadé. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 82(3):584–585. 9, 11

Glynn, P. W. and Rhee, C. H. (2014). Exact estimation for Markov chain equilibrium expectations. *Journal of Applied Probability*, 51(A):377–389. 1

Goodman, J. B. and Lin, K. K. (2009). Coupling control variates for Markov chain Monte Carlo. *Journal of Computational Physics*, 228(19):7127–7136. 1

Hastings, W. K. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 57(1):97–109. 1

Heng, J. and Jacob, P. E. (2019). Unbiased Hamiltonian Monte Carlo with couplings. *Biometrika*, 106(2):287–302. 1

Jacob, P. E., O'Leary, J., and Atchadé, Y. F. (2020). Unbiased Markov chain Monte Carlo methods with couplings. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 82(3):543–600. 1, 4, 9, 11

Jerrum, M. (1998). Mathematical foundations of the Markov chain Monte Carlo method. In *Probabilistic Methods for Algorithmic Discrete Mathematics*, pages 116–165. Springer. 1

Johnson, V. E. (1996). Studying convergence of Markov chain Monte Carlo algorithms using coupled sample paths. *Journal of the American Statistical Association*, 91(433):154–166. 1

Johnson, V. E. (1998). A coupling-regeneration scheme for diagnosing convergence in Markov chain Monte Carlo algorithms. *Journal of the American Statistical Association*, 93(441):238–248. 1, 2

Levin, D. A., Peres, Y., and Wilmer, E. L. (2017). *Markov Chains and Mixing Times*, volume 107. American Mathematical Soc. 2

Lindvall, T. (1992). *Lectures on the Coupling Method*. John Wiley and Sons, Inc., New York. 3

Lindvall, T. and Rogers, L. C. G. (1986). Coupling of Multidimensional Diffusions by Reflection. *The Annals of Probability*, 14(3):860–872. 3, 9

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., and Teller, E. (1953). Equation of state calculations by fast computing machines. *The journal of Chemical Physics*, 21(6):1087–1092. 1

Middleton, L., Deligiannidis, G., Doucet, A., and Jacob, P. E. (2019). Unbiased smoothing using particle independent Metropolis–Hastings. volume 89 of *Proceedings of Machine Learning Research*, pages 2378–2387. PMLR. 1

Middleton, L., Deligiannidis, G., Doucet, A., and Jacob, P. E. (2020). Unbiased Markov chain Monte Carlo for intractable target distributions. *Electronic Journal of Statistics*, 14(2):2842–2891. 1

Neal, R. and Pinto, R. (2001). Improving Markov chain Monte Carlo estimators by coupling to an approximating chain. Technical report, Department of Statistics, University of Toronto. 1

Neal, R. M. (1993). Bayesian learning via stochastic dynamics. In *Advances in Neural Information Processing Systems*, pages 475–482. 1

Neal, R. M. (1999). Circularly-coupled Markov chain sampling. Technical report, Department of Statistics, University of Toronto. 1

Neal, R. M. (2011). MCMC using Hamiltonian dynamics. *Handbook of Markov Chain Monte Carlo*, 2(11):2. 1

Nüsken, N. and Pavliotis, G. A. (2019). Constructing sampling schemes via coupling: Markov semigroups and optimal transport. *SIAM/ASA Journal on Uncertainty Quantification*, 7(1):324–382. 9

Piponi, D., Hoffman, M., and Sountsov, P. (2020). Hamiltonian Monte Carlo swindles. volume 108 of *Proceedings of Machine Learning Research*, pages 3774–3783, Online. PMLR. 1

Pollard, D. (2005). Asymptopia. *Manuscript, Yale University, Department of Statistics*. 8

Propp, J. G. and Wilson, D. B. (1996). Exact sampling with coupled Markov chains and applications to statistical mechanics. *Random Structures and Algorithms*, 9(12):223–252. 1

Roberts, G. O., Gelman, A., and Gilks, W. R. (1997). Weak convergence and optimal scaling of random walk Metropolis algorithms. *Annals of Applied Probability*, 7(1):110–120. 7

Roberts, G. O. and Tweedie, R. L. (1996). Exponential convergence of Langevin distributions and their discrete approximations. *Bernoulli*, 2(4):341–363. 1

Rosenthal, J. S. (1995). Minorization conditions and convergence rates for Markov chain Monte Carlo. *Journal of the American Statistical Association*, 90(430):558–566. 1

Rosenthal, J. S. (2002). Quantitative convergence rates of Markov chains: A simple account. *Electronic Communications in Probability*, 7:123–128. 1

Smith, A. (2014). A Gibbs sampler on the n-simplex. *The Annals of Applied Probability*, 24(1):114–130. 9

Van Rossum, G. and Drake, F. L. (2009). *Python 3 Reference Manual*. CreateSpace, Scotts Valley, CA. 1

Vaserstein, L. N. (1969). Markov processes over denumerable products of spaces, describing large systems of automata. *Problemy Peredachi Informatsii*, 5(3):64–72. 3

A   Appendix

In Appendix A.1, we prove that the coupling $\bar{P}_{\mathrm{MI}}((x,y), \cdot)$ described in Algorithm 3 is a maximal coupling of $P(x, \cdot)$ and $P(y, \cdot)$ and analyze its computational cost. Similarly, in Appendix A.2 we prove that the $\bar{P}_{\mathrm{MR}}$ described in Algorithm 4 is valid and maximal and analyze its cost.

A.1   Validity, maximality, and computation cost of $\bar{P}_{\mathrm{MI}}$

We prove that Algorithm 3 defines a coupling of the correct marginal distributions and that it attains the maximum one-step meeting probability.

Proposition 2. *The draws (X, Y) produced by Algorithm 3 follow a maximal coupling of $P(x, \cdot)$ and $P(y, \cdot)$, with the property that X and Y are conditionally independent given $X \neq Y$. Moreover, the coupling probability is maximized among all possible couplings, so that $\bar{P}_{\mathrm{MI}}((x,y), \{X = Y\}) = 1 - \|P(x, \cdot) - P(y, \cdot)\|_{\mathsf{TV}}$.*

Proof. It suffices to show that Y has marginal distribution $P(y, \cdot)$ and the coupling probability equals $1 - \|P(x, \cdot) - P(y, \cdot)\|_{\mathsf{TV}}$. We define the residual measure of y as $\tilde{P}_{yx}(\cdot) \propto r(y)\delta_y(\cdot) + \tilde{f}_{yx}^r(\cdot)$, where δ_y is the point mass at state y and $\tilde{f}_{yx}^r(y') := f(y, y') - f(y, y') \wedge f(x, y')$ is the 'unnormalized' y-residual density evaluated at y'. These definitions are consistent with the ones given in Section 3.2. From this point of view, Step 3 can be seen as a standard rejection sampler with proposal measure $P(y, \cdot)$ and target measure $\tilde{P}_{yx}(\cdot)$. For any $y' \neq y$, let $f_{yx}(y')$ denote the transition density of the Y-chain from y to y'. Then $f_{yx}(y')$ can be written as

$$f_{yx}(y') = f_{yx}(y', \text{Step 2}) + f_{yx}(y', \text{Step 3}).$$

The first term works out to $f_{yx}(y', \text{Step 2}) = f(x, y') \wedge f(y, y')$, while the second term can be computed as

$$f_{yx}(y', \text{Step 3}) = \mathbb{P}(\text{Step 3}) f_{yx}(y' \mid \text{Step 3})$$
$$= \left(1 - \int f(x, z) \wedge f(y, z)dz\right) \frac{1}{c(x, y)} \tilde{f}_{yx}^r(y').$$

Here $c(x, y) = 1 - \int f(y, z) \wedge f(x, z)dz$ is the normalizing constant of $\tilde{P}_{yx}(\cdot)$, which also equals $r(y) + \int \tilde{f}_{yx}^r(z)dz$.

Putting all the terms together, we have $f_{yx}(y') = f(x, y') \wedge f(y, y') + \tilde{f}_{yx}^r(y') = f(y, y')$ as desired, which justifies the validity of Algorithm 3.

We can also observe that the coupling probability equals the probability that Algorithm 3 stops at Step 2. Therefore, the coupling probability satisfies

$$\bar{P}_{\mathrm{MI}}((x,y), \{X = Y\}) = \int f(x, y') \wedge f(y, y')dy' = 1 - \|P(x, \cdot) - P(y, \cdot)\|_{\mathsf{TV}}.$$

We conclude that Algorithm 3 maximizes the coupling probability in one step. $\square$

We analyze the computation cost of Algorithm 3 as follows. To draw one sample from Algorithm 3, one needs to run Step 1 once with probability 1, Step 2 once with probability $1 - r(x)$, Step 3 for N times where N is a random variable which equals 0 with probability $1 - \|P(x, \cdot) - P(y, \cdot)\|_{\mathsf{TV}}$, and otherwise a Geometric random variable with success probability $\|P(x, \cdot) - P(y, \cdot)\|_{\mathsf{TV}}$. Meanwhile, Step 1 contains one draw from P, Step 2 contains two evaluations, Step 3 contains one draw from P and 0 or 2 evaluations, with probability $r(y)$ and $1 - r(y)$ respectively.

Therefore, the expected number of draws from P is 2 and the expected number of evaluations is $4 - 2r(x) - 2r(y)$. Taking account of the fact that each draw from P contains one draw from q and one evaluation of the acceptance ratio, then Algorithm 3 contains 2 draws from q and $6 - 2r(x) - 2r(y)$ evaluations in expectation. The variance of the computing cost depends the total variation distance between $P(x, \cdot)$ and $P(y, \cdot)$, and goes to infinity as $\|P(x, \cdot) - P(y, \cdot)\|_{\mathsf{TV}}$ goes to zero. This can motivate the consideration of sub-maximal coupling strategies, as described in the comment of Gerber and Lee (2020) on Jacob et al. (2020).

A.2   Validity, maximality, and computation cost of $\bar{P}_{\mathrm{MR}}$

We prove that Algorithm 4 defines a valid coupling and attains the maximal coupling probability.

Proposition 3. *The draws $(X, Y) \sim \bar{P}_{\mathrm{MR}}((x, y), \cdot)$ produced by Algorithm 4 follow a maximal coupling of $P(x, \cdot)$ and $P(y, \cdot)$.*

Proof. As in the proof of Proposition 2, it suffices to show that Y is distributed according to $P(y, \cdot)$ and that the meeting probability equals $1 - \|P(x, \cdot) - P(y, \cdot)\|_{\mathsf{TV}}$. Let the functions f_{xy}^m, $\tilde{f}_{xy}^r$, $\tilde{f}_{yx}^r$, and $\tilde{f}_{yx}^t$ have the definitions given in Section 3.2. For any $y' \neq y$, the Markov transition density $f(y, y')$ can be written as the sum of three terms:

$$f(y, y') = f_{xy}^m(y') + \tilde{f}_{yx}^r(y') \wedge \tilde{f}_{xy}^r(T_{yx}(y')) + \tilde{f}_{yx}^t(y'). \tag{1}$$

For any $y' \neq y$, let $f_{yx}(y')$ denote the transition density of the Y-chain from y to y' according to Algorithm 4. Then $f_{yx}(y')$ can also be written as the sum of three terms:

$$f_{yx}(y') = f_{yx}(y', \text{Step 2}) + f_{yx}(y', \text{Step 3}(b)) + f_{yx}(y', \text{Step 3}(c)). \tag{2}$$

We confirm that each term in Formula (2) matches the corresponding term in (1). For the first term, we have

$$f_{yx}(y', \text{Step 2}) = f(x, y') \, \mathbb{P}(U f(x, y') \leq f(y, y')) = f(y, y') \wedge f(x, y').$$

For the second term, let $x' = T_{yx}(y')$ be the preimage of y' through T_{xy}. It is not difficult to verify that $T_{xy}^{-1} = T_{yx}$ and that the Jacobian of both T_{xy} and T_{yx} equals 1. Thus the density of moving from y to y' through Step 3(b) will be

$$f_{yx}(y', \text{Step 3}(b)) = \tilde{f}_{xy}^r(x') \, \mathbb{P}(U \tilde{f}_{xy}^r(x') \leq \tilde{f}_{yx}^r(y')) |J(T_{xy}^{-1})| = \tilde{f}_{yx}^r(y') \wedge \tilde{f}_{xy}^r(T_{yx}(y')).$$

This matches the second term in (1).

Step 3(c) is again a rejection sampler with proposal $P(y, \cdot)$ and target $\tilde{P}_{yx}^t(\cdot) \propto r(y)\delta_y(\cdot) + \tilde{f}_{yx}^t(\cdot)$. Therefore,

$$f_{yx}(y', \text{Step 3}(c)) = \mathbb{P}(\text{Step 3}(c)) \cdot f_{yx}(y' | \text{Step 3}(c)) = \mathbb{P}(\text{Step 3}(c)) \frac{1}{\tilde{c}(x, y)} \tilde{f}_{yx}^t(y'),$$

where $\tilde{c}(x, y) = 1/(r(y) + \int \tilde{f}_{yx}^t(z) \, dz)$ is the normalizing constant of $\tilde{P}_{yx}^t(\cdot)$.

Meanwhile, we have

$$\mathbb{P}(\text{Step 3}(c)) = 1 - \int f_{xy}^m(y') dy' - \int \tilde{f}_{yx}^r(y') \wedge \tilde{f}_{xy}^r(T_{yx}(y')) \, dy'$$

$$= r(y) + \int f(y, y') \, dy' - \int f_{xy}^m(y') \, dy' - \int \tilde{f}_{yx}^r(y') \wedge \tilde{f}_{xy}^r(T_{yx}(y')) \, dy'$$

$$= r(y) + \int \tilde{f}_{yx}^t(z) \, dz.$$

Here the final equality uses (1). This yields $f_{yx}(y', \text{Step 3}(c)) = \tilde{f}_{yx}^t(y')$, which concludes the proof.

We also observe that the meeting probability equals the probability that Algorithm 4 stops at Step 2. Thus the coupling probability $\bar{P}_{\mathrm{MR}}((x, y), \{X = Y\}) = \int f(x, y') \wedge f(y, y') dy' = 1 - \|P(x, \cdot) - P(y, \cdot)\|_{\mathsf{TV}}$ attains the upper bound given by the coupling inequality. $\square$

The computation cost of Algorithm 4 can be analyzed in a similar way as the cost of Algorithm 3. We define the one-step coupling probability $p_c := \int f_{xy}^m(y') \, dy' = 1 - \|P(x, \cdot) - P(y, \cdot)\|_{\mathsf{TV}}$ and the one-step rejection probability $p_r := \tilde{f}_{yx}^r(y') \wedge \tilde{f}_{xy}^r(T_{yx}(y')) \, dy'$. To draw one sample from Algorithm 4, one needs to run Step 1 once with probability 1, run Step 2 once with probability $1 - r(x)$, Step 3(a) once with probability $1 - p_c$, and Step 3(b) once with probability $1 - r(x) - p_c$. The number of runs of Step 3(c) will be zero with probability $1 - p_c - p_r$. Otherwise it will follow a Geometric random variable with success probability $1 - p_c - p_r$.

Meanwhile, Step 1 contains one draw from P, Step 2 contains two evaluations, Step 3(a) contains one evaluation, Step 3(b) contains two evaluations, Step 3(c) contains one draw from P and 0 or 2 evaluations, with probability

$r(y)$ and $1 - r(y)$ respectively. Therefore, the expected number of draws from P is 2, and the expected number of evaluations is $7 - 4r(x) - 2r(y) - 3p_c$. If one takes into account the fact that each draw from P itself contains one draw from q and one evaluation of the acceptance ratio, then Algorithm 3 contains 2 draws from q and $7 - 4r(x) - 2r(y) - 3p_c$ evaluations, in expectation. This is greater than the expected cost of Algorithm 3 by $3 - 3p_c - 2r(x)$. This quantity is between 0 and 3 as $p_c + r(x) \leq 1$. The variance of the computation cost also depends on the total variation distance between $P(x, \cdot)$ and $P(y, \cdot)$, and goes to infinity as $\|P(x, \cdot) - P(y, \cdot)\|_{\mathsf{TV}}$ goes to zero, as noted above.