Couplings of the Random-Walk Metropolis algorithm

John O'Leary*

December 6, 2021

Abstract

Couplings play a central role in contemporary Markov chain Monte Carlo methods and in the analysis of their convergence to stationarity. In most cases, a coupling must induce relatively fast meeting between chains to ensure good performance. In this paper we fix attention on the random walk Metropolis algorithm and examine a range of coupling design choices. We introduce proposal and acceptance step couplings based on geometric, optimal transport, and maximality considerations. We consider the theoretical properties of these choices and examine their implication for the meeting time of the chains. We conclude by extracting a few general principles and hypotheses on the design of effective couplings.

1 Introduction

In commemorating the 50th anniversary of the Metropolis–Hastings (MH) algorithm, Dunson and Johndrow [2020] point to the unbiased estimation method of Jacob et al. [2020] as a leading strategy for the parallelization of Markov chain Monte Carlo (MCMC) algorithms. However, they note a challenge: while it is usually easy to find a transition kernel coupling with properties needed for this approach, that choice is rarely unique, and the wrong selection can result in low estimator efficiency. The design of efficient couplings is, as they write, "an exciting direction that we expect will see growing attention among practitioners." In this study we take up this important question.

From the early days of Markov chain theory [e.g. Doeblin, 1938, Harris, 1955, Pitman, 1976, Aldous, 1983, Rosenthal, 1995], couplings have played a key role in the analysis of convergence to stationarity. In recent years they have also been used to formulate MCMC diagnostics [Johnson, 1996, 1998, Biswas et al., 2019], variance reduction methods, [Neal and Pinto, 2001, Goodman and Lin, 2009, Piponi et al., 2020], and new sampling and estimation strategies [Propp and Wilson, 1996, Fill, 1997, Neal, 1999, Flegal and Herbei, 2012, Glynn and Rhee, 2014, Jacob et al., 2020, Heng and Jacob, 2019]. Couplings that produce smaller meeting times generally yield better results in the form of tighter bounds, more variance reduction, greater computational efficiency, or more precise estimators.

*Department of Statistics, Harvard University, Cambridge, MA, USA. Email: joleary@g.harvard.edu

Thus, the design of efficient couplings has been an important question for almost the entire history of the coupling method. When a coupling is not required to be co-adapted to the chains in question, simple arguments show that a maximal coupling of the chains exists and results in meeting at the fastest rate allowed by the coupling inequality [Griffeath, 1975, Goldstein, 1979]. However when the coupling must be implementable and Markovian, maximal couplings are known only in special cases [Burdzy and Kendall, 2000, Hsu and Sturm, 2013, Böttcher, 2017]. Markovian couplings are easy to work with and are required for many of the applications above, but they are rarely maximal.

In this study we consider transition kernel couplings of the Random Walk Metropolis (RWM) algorithm [Metropolis et al., 1953], which is perhaps the oldest, simplest, and best-understood MCMC method. Transition kernel couplings [Douc et al., 2018, chap. 19] are Markovian by construction. Explicit and implementable couplings of the RWM kernel seem to originate with Johnson [1998]. These methods were taken up in Jacob et al. [2020], which found that apparently minor differences in coupling design can have significant implications for meeting times, especially in relation to the dimension of the state space. In this paper we continue this line of inquiry and take a pragmatic approach to the question of coupling design. We ask: what options are available for coupling the RWM kernel, how do these choices affect meeting times, and what lessons can we learn from this simple case?

We begin by introducing the essential ingredients of an RWM kernel coupling. First, we consider proposal distribution couplings, devoting some attention to maximal couplings of the multivariate normal distribution. Next, we turn to coupling at the accept/reject step. Any coupling of the RWM kernel can be realized as a proposal coupling followed by an acceptance step coupling [O'Leary and Wang, 2021], so this focus on separate proposal and acceptance couplings involves no loss of generality. We conclude with a range of simulation exercises to understand how various coupling design options affect meeting times. We conclude with some stylized facts and advice on the construction of efficient couplings for the RWM algorithm and beyond.

2 Setting and notation

Throughout the following we write $a \wedge b := \min(a, b)$, $a \vee b := \max(a, b)$, and $\mathcal{L}(Z)$ for the law of a random variable Z. We write $\text{Bern}(\alpha)$ for the Bernoulli distribution on $\{0, 1\}$ with $\mathbb{P}(\text{Bern}(\alpha) = 1) = \alpha$, $\text{N}(\mu, \Sigma)$ for the multivariate normal distribution with mean μ and covariance matrix Σ, and $\text{N}(z; \mu, \Sigma)$ for the density of this distribution evaluated at a point z.

Fix a target distribution π on a state space $(\mathbb{R}^d, \mathcal{B})$, where $\mathcal{B}$ is the Borel σ-algebra on $\mathbb{R}^d$. Let $Q : \mathbb{R}^d \times \mathcal{B} \to [0, 1]$ be a proposal kernel. Thus $Q(x, \cdot)$ is a probability measure for all $z \in \mathbb{R}^d$ and $Q(\cdot, A) : \mathbb{R}^d \to [0, 1]$ is measurable for all $A \in \mathcal{B}$. We interpret $Q(x, A)$ as the probability of proposing some point $x' \in A$ when the current state is x. Assume π has density $\pi(\cdot)$ and $Q(x, \cdot)$ has density $q(x, \cdot)$ for $x \in \mathbb{R}^d$, all with respect to Lebesgue measure. The MH acceptance ratio [Hastings, 1970] is then defined as $a(x, x') := 1 \wedge \frac{q(x',x)}{q(x,x')} \frac{\pi(x')}{\pi(x)}$. In this study we focus on the

RWM algorithm with multivariate normal proposal increments, so $Q(x, \cdot) = \mathrm{N}(x, I_d\, \sigma_d^2)$ for all $x \in \mathbb{R}^d$. In this case $q(x, \tilde{x}) = q(\tilde{x}, x)$ for all $x, x' \in \mathbb{R}^d$, and $a(x, \tilde{x}) = 1 \wedge (\pi(\tilde{x})/\pi(x))$.

We construct an MH chain (X_t) as follows. First we initialize the chain with a draw X_0 from an arbitrary distribution π_0 on $(\mathbb{R}^d, \mathcal{B})$. At each iteration t we draw $\tilde{x} \sim Q(x, \cdot)$, where $x = X_t$ is the current state of the chain. We then draw an acceptance indicator $b_x \sim \mathrm{Bern}(a(x, x'))$ and set $X_{t+1} := b_x x' + (1 - b_x)x$. It is often convenient to realize the acceptance indicator draw by taking $U \sim \mathrm{Unif}$ and $b_x := 1(U \leq a(x, x'))$. The chain (X_t) defined above will have a transition kernel P defined by $P(x, A) := \mathbb{P}(X_{t+1} \in A \,|\, X_t = x)$ for $x \in \mathbb{R}^d$ and $A \in \mathcal{B}$.

For any probability measures μ and ν on $(\mathbb{R}^d, \mathcal{B})$, we say that a probability measure γ on $(\mathbb{R}^d \times \mathbb{R}^d, \mathcal{B} \otimes \mathcal{B})$ is a coupling of μ and ν if $\gamma(A \times \mathbb{R}^d) = \mu(A)$ and $\gamma(\mathbb{R}^d \times A) = \nu(A)$ for all $A \in \mathcal{B}$. We write $\Gamma(\mu, \nu)$ for the set of all such couplings of μ and ν. Next suppose that (X_t) and (Y_t) are both Markov chains defined on the same probability space and that both evolve according to the RWM transition kernel P defined above. We say that (X_t, Y_t) follows a transition kernel coupling $\bar{P}$ based on P if there exists a joint kernel $\bar{P} : (\mathbb{R}^d \times \mathbb{R}^d) \times (\mathcal{B} \otimes \mathcal{B}) \to [0, 1]$ with $\bar{P}((x, y), \cdot) \in \Gamma(P(x, \cdot), P(y, \cdot))$ for all $x, y \in \mathbb{R}^d$. We write $\Gamma(P, P)$ for the set of all such kernel couplings. The limitation to couplings of (X_t) and (Y_t) that can be expressed in the form above is not a trivial one, as described further in Kumar and Ramesh [2001].

We write $\tau = \min(t : X_t = Y_t)$ for the first time the chains meet. Couplings $\bar{P}$ with the property that $\mathbb{P}(\tau < \infty) = 1$ are called successful. To obtain successful couplings, we generally need a proposal kernel coupling with $\mathbb{P}(x' = y' \,|\, x, y) > 0$ from at least some state pairs $(X_t, Y_t) = (x, y)$. This will lead us to consider maximal couplings of the proposal distributions, which achieve the highest possible probability $\mathbb{P}(x' = y' \,|\, x, y)$ for each $x, y \in \mathbb{R}^d$. Couplings $\bar{P}$ with the property that $X_t = Y_t$ for all $t \geq \tau$ are called sticky and are also our subject of interest here. Rosenthal [1997] and Dey et al. [2017] point out that stickiness is a non-trivial property, even for Markovian couplings. However, the couplings we consider can always be made sticky by requiring $x' = y' \sim Q(x, \cdot) = Q(y, \cdot)$ and $V = U \sim \mathrm{Unif}$ if $x = y$.

In this paper we consider a range of coupling options for the RWM transition kernel. Our goal is to understand the implications of these options for the distribution of τ and especially for the value of its mean $\mathbb{E}[\tau]$. The average meeting time serves as a convenient summary of the meeting rate and plays a specific role in the efficiency of the estimators described in Jacob et al. [2020]. We will focus on transition kernel couplings that arise by separately coupling the proposals (x', y') and the uniform draws (U, V) underlying the acceptance indicators $(b_x, b_y) = (1(U \leq a(x, x')), 1(V \leq a(y, y')))$. This strategy is fully general except with respect to the acceptance indicator coupling, as noted in O'Leary and Wang [2021].

When it is unlikely to cause confusion, we write $(x, y) = (X_t, Y_t)$ for the current state of a pair of coupled MH chains, $(x', y') = (x + \xi, y + \eta)$ for the proposals, and $(X, Y) = (X_{t+1}, Y_{t+1})$ for the next state pair. We write $r = ||x - y||$ for the Euclidean distance between x and y and $m = (x + y)/2$ for their midpoint. We write $e = (y - x)/r$ for the unit vector pointing from x to y, an important direction in many of the constructions described below. Finally, for any $z \in \mathbb{R}^d$

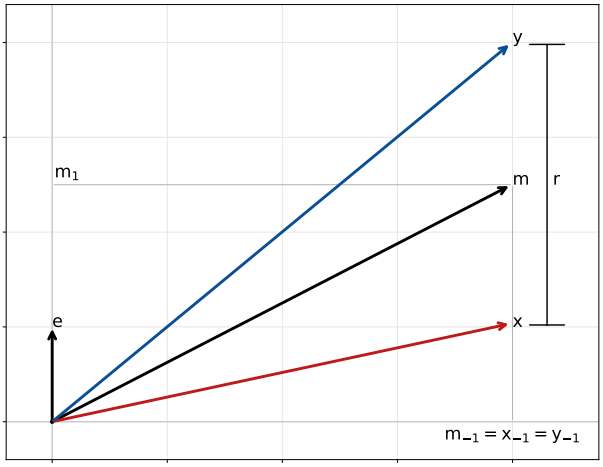


Figure 1: Coupled chain notation and geometry. We denote current points by $x, y \in \mathbb{R}^d$, their separation by $r \geq 0$, and their midpoint by $m \in \mathbb{R}^d$. The unit vector e points in the direction from x to y, $m_1 \in \mathbb{R}$ gives the e component of m, and $m_{-1} = x_{-1} = y_{-1} \in \mathbb{R}^d$ gives the component of these three vectors which is orthogonal to e.

we write $z_1 = e'z \in \mathbb{R}$ for the e component of z and $z_{-1} = (I_d - ee')z \in \mathbb{R}^d$ for the projection of z onto the subspace orthogonal to e. Thus we can express any vector $z \in \mathbb{R}^d$ as $z = ez_1 + z_{-1}$. See Figure 1 for an illustration of these quantities.

3 Maximal coupling foundations

To obtain finite meeting times we generally need $\mathbb{P}(x' = y' \,|\, x, y) > 0$ from at least some state pairs (x, y) with $x \neq y$. One solution is to draw $(\tilde{x}, \tilde{y})$ from a maximal coupling $\bar{Q} \in \Gamma(Q, Q)$. A coupling of $\tilde{x} \sim Q(x, \cdot)$ and $\tilde{y} \sim Q(y, \cdot)$ is said to be maximal if it achieves the upper bound given by the coupling inequality, $\mathbb{P}(\tilde{x} = \tilde{y} \,|\, x, y) \leq 1 - ||Q(x, \cdot) - Q(y, \cdot)||_{\text{TV}} = 1 - \sup_{A \in \mathcal{B}} |Q(x, A) - Q(y, A)|$. See Thorisson [2000, chap. 1.4] or Levin et al. [2017, chap 4.] for discussion of this bound and its applications. For any probability distributions μ and ν on $(\mathbb{R}^d, \mathcal{B})$, we write $\Gamma^{\max}(\mu, \nu)$ for the set of all maximal couplings of μ and ν. Maximal couplings of the proposal distribution make an appealing starting point, but note that their use is neither necessary nor sufficient to maximize $\mathbb{P}(X = Y \,|\, x, y)$. Gerber and Lee [2020] also observe that the variance of the computational cost to draw from a maximal coupling can blow up when $r = \|y - x\| \to 0$. In such cases one may prefer to use a slightly non-maximal coupling over a maximal one.

The following result, closely related to Douc et al. [2018], Theorem 19.1.6 and Proposition D.2.8, shows that maximality comes with significant constraints on a coupling's behavior:

Lemma 3.1. Let $\bar{Q}((x, y), \cdot)$ be a maximal coupling of $Q(x, \cdot)$ and $Q(y, \cdot)$, distributions with

densities $q(x,\cdot)$ and $q(y,\cdot)$ on $\mathbb{R}^d$. If $(x',y') \sim \bar{Q}((x,y),\cdot)$, then for all $A \in \mathcal{B}$,

$$\begin{aligned}
\mathbb{P}(x' \in A, x' = y' \,|\, x, y) = \mathbb{P}(y' \in A, x' = y' \,|\, x, y) &= \int_A q(x,z) \wedge q(y,z)\, \mathrm{d}z \\
\mathbb{P}(x' \in A, x' \neq y' \,|\, x, y) &= \int_A 0 \vee (q(x,z) - q(y,z))\, \mathrm{d}z \\
\mathbb{P}(y' \in A, x' \neq y' \,|\, x, y) &= \int_A 0 \vee (q(y,z) - q(x,z))\, \mathrm{d}z.
\end{aligned}$$

We can obtain $\mathbb{P}(x' = y' \,|\, x, y)$ by evaluating the first equation at $A = \mathbb{R}^d$. This meeting probability takes a particularly simple form for multivariate normal distributions, as we see in the following extension of Pollard [2005, chap. 3.3]:

Lemma 3.2. If (x', y') follows any maximal coupling of $\mathrm{N}(x, I_d\sigma_d^2)$ and $\mathrm{N}(y, I_d\sigma_d^2)$, then

$$\mathbb{P}(x' = y' \,|\, x, y) = \mathbb{P}\left(\chi_1^2 \geq \tfrac{||y-x||^2}{4\sigma_d^2}\right).$$

Proof. Recall that we write $\mathrm{N}(z; \mu, \Sigma)$ for the density of $\mathrm{N}(\mu, \Sigma)$ and have defined $m = (y+x)/2$, $r = \|y - x\|$, $e = (y-x)/r, z_1 = e'z$ and $z_{-1} = (I_d - ee')z$ for all $z \in \mathbb{R}^d$. In general $\mathrm{N}(z; \mu, I_d\sigma^2) = \mathrm{N}(z_1; \mu_1, \sigma_d^2)\,\mathrm{N}(z_{-1}; \mu_{-1}, I_{d-1}\sigma_d^2)$. We can also decompose x and y into e and $e^\perp$ parts according to $x = m - e\frac{r}{2} = (m_1 - \frac{r}{2})e + m_{-1}$ and $y = m + e\frac{r}{2} = (m_1 + \frac{r}{2})e + m_{-1}$. Combining these expressions with Lemma 3.1 yields the desired conclusion:

$$\begin{aligned}
\mathbb{P}(\tilde{x} = \tilde{y} \,|\, x, y) &= \int \mathrm{N}(z; x, I_d\sigma_d^2) \wedge \mathrm{N}(z; y, I_d\sigma_d^2)\, \mathrm{d}z \\
&= \iint \left(\mathrm{N}(z_1; -\tfrac{r}{2}, \sigma_d^2) \wedge \mathrm{N}(z_1; \tfrac{r}{2}, \sigma_d^2)\right) \mathrm{N}(z_{-1}; m_{-1}, I_{d-1}\sigma_d^2)\, \mathrm{d}z_1\, \mathrm{d}z_{-1} \\
&= \int_{-\infty}^{\infty} \mathrm{N}(z_1; -\tfrac{r}{2}, \sigma_d^2) \wedge \mathrm{N}(z_1; \tfrac{r}{2}, \sigma_d^2)\, \mathrm{d}z_1 = 2\int_0^{\infty} \mathrm{N}(z_1; -\tfrac{r}{2}, \sigma_d^2)\, \mathrm{d}z_1 \\
&= 2\,\mathbb{P}(\mathrm{N}(0,1) \geq \tfrac{r}{2\sigma_d}) = \mathbb{P}(\chi_1^2 \geq \tfrac{r^2}{4\sigma_d^2}).
\end{aligned}$$

□

An important implication of Lemma 3.2 is that as we increase the dimension d, the separation $r = ||y - x||$ needed to hold $\mathbb{P}(x' = y' \,|\, x, y)$ constant must vary in proportion to σ_d^2. Under the typical RWM assumption that $\sigma_d^2 = \ell^2/d$, this means r must shrink at a rate $1/\sqrt{d}$ to maintain a constant probability of meeting proposals. This inverse square-root condition plays a crucial role in determining the dimension scaling behavior of different couplings as we will observe in the simulations of Section 6. Note that the meeting probability derived in Lemma 3.2 also admits the following useful inequalities:

Lemma 3.3. Under any maximal coupling of $x' \sim \mathrm{N}(x, I_d\sigma_d^2)$ and $y' \sim \mathrm{N}(y, I_d\sigma_d^2)$, we have

$$1 - \sqrt{\tfrac{2}{\pi}}\tfrac{||y-x||}{2\sigma_d} \leq \mathbb{P}(x' = y' \,|\, x, y) \leq \tfrac{4\sigma_d^2}{\|y-x\|^2} \quad \text{and} \quad \mathbb{P}(x' = y' \,|\, x, y) \leq \tfrac{1}{\sqrt{1-2s}} \exp\left(-s\tfrac{||y-x||^2}{4\sigma_d^2}\right)$$

for $s \in (0, 1/2)$.

Proof. For the lower bound, let $\phi(z) = \mathrm{N}(z; 0, 1)$ be the standard normal density and let $\Phi(z)$ be the corresponding cumulative distribution function. Since $\Phi(0) = 1/2$ and $\phi(z) \leq 1/\sqrt{2\pi}$ for

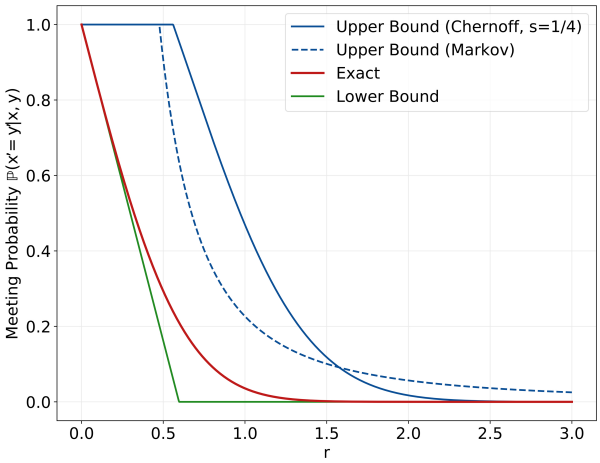


Figure 2: Meeting probability and bounds for maximal couplings of normal distributions on $\mathbb{R}$. The red line ('Exact') gives $\mathbb{P}(x' = y' \,|\, x, y)$ when (x', y') follows a maximal coupling of $\mathrm{N}(x, 1)$ and $\mathrm{N}(y, 1)$, based on the results on Lemma 3.2. The blue and green lines give the bounds derived in Lemma 3.3. Here $r = \|y - x\|$. The exact meeting probabilities involve the complementary CDF of the χ^2_1 distribution, so it can be analytically more convenient to use the given bounds.

all z, then for $a > 0$ we may write $\Phi(a) = \int_{-\infty}^{a} \phi(z)\,\mathrm{d}z \leq \frac{1}{2} + \frac{a}{\sqrt{2\pi}}$. This expression rearranges to $1 - \sqrt{\frac{2}{\pi}}a \leq 2(1 - \Phi(a))$, and plugging in $a = r/(2\sigma_d)$ yields the desired lower bound. The first upper bound follows directly from Markov's inequality:

$$\mathbb{P}(x' = y' \,|\, x, y) = \mathbb{P}(\chi^2_1 \geq \tfrac{r^2}{4\sigma_d^2}) \leq \frac{\mathbb{E}[\chi^2_1]}{r^2/(4\sigma_d^2)} = \frac{4\sigma_d^2}{r^2}.$$

The second upper bound is due to Chernoff's inequality, $\mathbb{P}(\chi^2_1 \geq a) \leq e^{-sa}\,\mathbb{E}[e^{s\chi^2_1}]$ for all $s > 0$. We have $\mathbb{E}[e^{s\chi^2}] = 1/\sqrt{1 - 2s}$ for $s < 1/2$, so plugging in $a = r^2/(4\sigma_d^2)$ yields the desired expression. □

In Figure 2 we plot the value of $\mathbb{P}(\tilde{x} = \tilde{y} \,|\, x, y)$ as derived in Lemma 3.2 along with the upper and lower bounds from Lemma 3.3. We observe that while the lower bound is tight at $r = 0$ and in the limit as $r \to \infty$, the upper bounds only become tight in the large-r limit. When needed, sharper upper and lower bounds can be obtained from more precise Gaussian tail inequalities, see e.g. Abramowitz et al. [1988, chap. 7] and Duembgen [2010].

We close by noting that if we can produce draws from one maximal coupling, we can often transform these into draws from a maximal coupling of a related pair of distributions. Recall that for any measure μ on $(\mathbb{R}^d, \mathcal{B})$ and measurable function $f : \mathbb{R}^d \to \mathbb{R}^d$, the pushforward measure $f_\star\mu$ on $(\mathbb{R}^d, \mathcal{B})$ is defined by $f_\star\mu(A) := \mu(f^{-1}(A))$ for all $A \in \mathcal{B}$. Also, if $x' \sim \mu$ then $f(x') \sim f_\star\mu$. Thus we have the following:

Lemma 3.4. Suppose μ and ν are probability measures on $(\mathbb{R}^d, \mathcal{B})$ and let $f : \mathbb{R}^d \to \mathbb{R}^d$ be a homeomorphism. (x', y') follows a maximal coupling of μ and ν if and only if $(f(x'), f(y'))$ follows a maximal coupling of $f_\star\mu$ and $f_\star\nu$.

Proof. First, $(x', y') \in \Gamma(\mu, \nu)$ if and only if $(f(x'), f(y')) \in \Gamma(f_\star\mu, f_\star\nu)$ since f is a bijection. Also

$$\|f_\star\mu - f_\star\nu\|_{\mathsf{TV}} = \sup_{A\in\mathcal{B}} |\mu(f^{-1}(A)) - \nu(f^{-1}(A))| = \sup_{B\in\mathcal{B}'} |\mu(B) - \nu(B)| = \|\mu - \nu\|_{\mathsf{TV}}.$$

Here $\mathcal{B}' = \{f^{-1}(A) : A \in \mathcal{B}\}$, and $\mathcal{B}' = \mathcal{B}$ since f is a homeomorphism. Since f is a bijection, we also have $\mathbb{P}(x' = y') = \mathbb{P}(f(x') = f(y'))$. Thus (x', y') will achieve the coupling inequality bound exactly when $(f(x'), f(y'))$ does. Thus we have shown that the former pair follows a maximal coupling of μ and ν if and only if the latter follows a maximal coupling of $f_\star\mu$ and $f_\star\nu$. □

Lemma 3.4 allows us to efficiently draw from and analyze the maximal independent coupling of distributions like $\mathrm{N}(x, \Sigma)$ and $\mathrm{N}(y, \Sigma)$ in terms of the maximal independent coupling of $\mathrm{N}(0, I_d)$ and $\mathrm{N}(\Sigma^{-1/2}(y - x), I_d)$. It can also be useful in the design of couplings when the proposal kernel arises from a deterministic but well-behaved function of a multivariate normal random variable, as in the case of Hamiltonian Monte Carlo [Duane et al., 1987, Neal, 1993, 2011].

4 Proposal step couplings

In this section we describe a range of proposal kernel couplings $\bar{Q}$ based on the RWM proposal kernel $Q(z, \cdot) = \mathrm{N}(z, I_d\sigma_d^2)$ on $\mathbb{R}^d$. If $(x', y') = (x + \xi, y + \eta) \sim \bar{Q}((x, y), \cdot)$, then marginally $\xi, \eta \sim \mathrm{N}(0, I_d\sigma_d^2)$. These increments can exhibit a complex dependence pattern, and (ξ, η) need not be multivariate normal. The simplest option, however, is the independent coupling $\xi, \eta \overset{iid}{\sim} \mathrm{N}(0, I_d\sigma_d^2)$. One step more complex is the synchronous or 'common random numbers' coupling $\xi = \eta \sim \mathrm{N}(0, I_d\sigma_d^2)$. As noted in Givens and Shortt [1984] and Knott and Smith [1984], the synchronous coupling minimizes the expected squared distance $\mathbb{E}[\|x' - y'\|^2]$ among all joint distributions with $x' \sim \mathrm{N}(x, I_d\sigma_d^2)$ and $y' \sim \mathrm{N}(y, I_d\sigma_d^2)$. We comment further on optimal transport couplings below.

Another slightly more complex option is the simple reflection coupling, in which $\xi \sim \mathrm{N}(0, I_d\sigma_d^2)$ and $\eta = (I_d - 2ee')\xi$. With the notation $z_1 = e'z$ and $z_{-1} = (I_d - ee')z$ for any $z \in \mathbb{R}^d$, we note that the reflection coupling yields $\eta_1 = -\xi_1$ and $\eta_{-1} = \xi_{-1}$. Thus η is the reflection of ξ over the hyperplane $\mathcal{H} = \{z : ||z - x|| = ||z - y||\} = \{z : z_1 = m_1\}$. Taking this geometric logic a step further, we can also consider the full-reflection coupling in which $\xi \sim \mathrm{N}(0, I_d\sigma_d^2)$ and $\eta = -\xi$. This coupling maximizes $\mathbb{E}[\|y' - x'\|^2]$ just as the synchronous coupling minimizes it. The independent, synchronous, reflection, and full-reflection couplings are easy to draw from and straightforward to analyze. They also differ dramatically in the covariance and transport properties that they establish between x' and y', their interactions with various accept/reject procedures, and thus the degree of contraction they produce between coupled chains. However each of these couplings has the property that if $x \neq y$ then $x' \neq y'$ almost

Algorithm 1 Draw from the maximal independent coupling of $Q(x, \cdot)$ and $Q(y, \cdot)$

1. Draw $x' \sim Q(x, \cdot)$ and $W_x \sim \text{Unif}$
2. If $W_x\, q(x, x') \leq q(y, x')$, set $y' = x'$
3. Else:
   (a) Draw $\tilde{y} \sim Q(y, \cdot)$ and $W_y \sim \text{Unif}$
   (b) If $W_y\, q(y, \tilde{y}) > q(x, \tilde{y})$, set $y' = \tilde{y}$
   (c) Else go to 3(a)
4. Return (x', y')

surely. This implies $X \neq Y$, so exclusive reliance on these couplings cannot yield $\mathbb{P}(\tau < \infty) = 1$ unless $X_0 = Y_0$.

4.1 The maximal independent coupling

Suppose $(x', y') \sim \bar{Q}((x, y), \cdot)$ for some $\bar{Q} \in \Gamma^{\max}(Q, Q)$. One consequence of Lemma 3.1 is that all maximal couplings exhibit the same distribution of x' and y' given $x' = y'$. In particular, each of these variables will have conditional density $q^m_{xy}(z) := q(x, z) \wedge q(y, z) / \int q(x, w) \wedge q(y, w)\, \mathrm{d}w$. We refer to the distributions of x' and y' given $x' \neq y'$ as the residuals of $\bar{Q}((x, y), \cdot)$. In light of the above, we differentiate between various maximal couplings according to the behavior of these residuals, i.e. according to the distribution of (x', y') conditional on $x' \neq y'$.

The first and perhaps most famous maximal coupling was introduced by Vaserstein [1969] and termed the γ-coupling by Lindvall [1992]. It is the unique maximal coupling with the property that x' and y' are independent when $x' \neq y'$. Thus we call this the maximal coupling with independent residuals, or simply the maximal independent coupling. When $Q(z, \cdot) = \mathrm{N}(z, I_d \sigma_d^2)$ for $z \in \mathbb{R}^d$, Lemmas 3.1 and 3.2 imply that this coupling approximates the independent coupling of $Q(x, \cdot)$ and $Q(y, \cdot)$ as $r = ||y - x|| \to \infty$.

The references above prove that one can draw from the maximal independent coupling by using the rejection sampling procedure described in Algorithm 1. This method is simple and versatile, although it suffers from a loss of efficiency as a function of dimension. In our setting, each normal density evaluation requires $\mathcal{O}(d)$ computations, and these costs can be a factor in algorithmic performance in high dimensions or when the number of iterations required to obtain a valid y' draw is large. Algorithm 2 offers an alternative, which exploits the symmetries and factorization properties of the multivariate normal distribution. It provides a more efficient way to draw from the maximal coupling of these distributions, and it also lends itself to extensions and variations as we consider below. Lemma 4.1 establishes the validity of this algorithm.

Lemma 4.1. The output of Algorithm 2 is distributed according to the maximal independent coupling of $\mathrm{N}(x, I_d \sigma_d^2)$ and $\mathrm{N}(y, I_d \sigma_d^2)$.

Proof. First we show that the output (x', y') of Algorithm 2 follows a coupling of $\mathrm{N}(x, I_d \sigma_d^2)$ and $\mathrm{N}(y, I_d \sigma_d^2)$. For x' we have $x'_1 \sim \mathrm{N}(e'x, \sigma_d^2)$ and $x'_{-1} \sim \mathrm{N}((I_d - ee')x, (I_d - ee')\sigma_d^2)$ with independence between x'_1 and x'_{-1}. Thus $x' = x'_1 e + x'_{-1} \sim \mathrm{N}(x, I_d \sigma_d^2)$. For y', note that $y'_{-1} \sim \mathrm{N}(y_{-1}, (I_d - ee')\sigma_d^2)$ whether or not $x'_1 = y'_1$. This is trivial when $x'_1 \neq y'_1$. When $x'_1 = y'_1$ we have

Algorithm 2 Draw from the maximal independent coupling of $\mathrm{N}(x, I_d\sigma_d^2)$ and $\mathrm{N}(y, I_d\sigma_d^2)$.

1. Compute $e = (y-x)/\|y-x\|$, $m = (y+x)/2$, $x_1 = x'e$, and $y_1 = y'e$
2. Draw (x_1', y_1') from the maximal independent coupling of $\mathrm{N}(x_1, \sigma_d^2)$ and $\mathrm{N}(y_1, \sigma_d^2)$ using Algorithm 1
3. Independently draw $\tilde{x} \sim \mathrm{N}(x, I_d\sigma_d^2)$ and $\tilde{y} \sim \mathrm{N}(y, I_d\sigma_d^2)$
4. Set $x'_{-1} = (I_d - ee')\tilde{x}$ and $y'_{-1} = (I_d - ee')\tilde{y}$
5. Set $x' = \tilde{x}_1 e + \tilde{x}_{-1}$. If $\tilde{x}_1 = \tilde{y}_1$ set $y' = \tilde{y}_1 e + \tilde{x}_{-1}$, else set $y' = \tilde{y}_1 e + \tilde{y}_{-1}$
6. Return (x', y')

Algorithm 3 Draw from the maximal semi-independent coupling of $\mathrm{N}(x, I_d\sigma_d^2)$ and $\mathrm{N}(y, I_d\sigma_d^2)$.

1. Compute $e = (y-x)/\|y-x\|$, $m = (y+x)/2$, $x_1 = x'e$, and $y_1 = y'e$
2. Draw (x_1', y_1') from the maximal independent coupling of $\mathrm{N}(x_1, \sigma_d^2)$ and $\mathrm{N}(y_1, \sigma_d^2)$ using Algorithm 1
3. Draw $\tilde{z} \sim \mathrm{N}(m, I_d\sigma_d^2)$, set $z'_{-1} = (I_d - ee')\tilde{z}$, $x' = x_1'e + z'_{-1}$, and $y' = y_1'e + z'_{-1}$
4. Return (x', y')

$\mathbb{E}[y'_{-1} \,|\, x, y, x_1' = y_1'] = (I_d - ee')x = (I_d - ee')(m - \frac{r}{2}e) = (I_d - ee')m = (I_d - ee')(m + \frac{r}{2}e) = y_{-1}$. Also, $y_1' \sim \mathrm{N}(y'e, \sigma_d^2)$ and y'_{-1} are independent, so we conclude $y' \sim \mathrm{N}(y, I_d\sigma_d^2)$.

Next, we show that (x', y') follows a maximal coupling. By Lemma 3.2, draws from a maximal coupling of $\mathrm{N}(x, I_d\sigma_d^2)$ and $\mathrm{N}(y, I_d\sigma_d^2)$ must meet with probability $\mathbb{P}(\chi_1^2 \geq \frac{\|y-x\|^2}{4\sigma_d^2})$. By construction we have $x' = y'$ if and only if $x_1' = y_1'$. Applying Lemma 3.2 to the maximal coupling of $\mathrm{N}(x_1, \sigma_d^2)$ and $\mathrm{N}(y_1, \sigma_d^2)$ shows that meeting occurs with probability $\mathbb{P}(\chi_1^2 \geq \frac{(y_1-x_1)^2}{4\sigma_d^2})$. We also have $y_1 - x_1 = e'(y-x) = (y-x)'(y-x)/\|y-x\| = \|y-x\|$, so meeting occurs at the maximal rate.

Finally, we observe that x' and y' are independent conditional on $x' \neq y'$. This holds for x_1 and y_1 since these are drawn from a maximal independent coupling on $\mathbb{R}$, and it holds for x_{-1} and y_{-1} since in the relevant case these are defined using independent random variables. Thus Algorithm 2 produces draws from the maximal independent coupling of $\mathrm{N}(x, I_d\sigma_d^2)$ and $\mathrm{N}(y, I_d\sigma_d^2)$. □

Overall, the meeting time associated with a transition kernel coupling depends on that coupling's probability of producing a meeting at each step together with the dynamics of the chains conditional on not meeting. It is often a good idea to control the variance of $y' - x'$ when $x' \neq y'$, to reduce the tendency of the chains to push apart when meeting does not occur. This motivates what we call the maximal coupling with semi-independent residuals, or the maximal semi-independent coupling, which we define in Algorithm 3. This algorithm differs from the maximal independent coupling in that it has $x'_{-1} = y'_{-1}$ whether or not $x_1' = y_1'$. The validity of this algorithm follows from essentially the same argument as that of Lemma 4.1.

Algorithm 4 Draw from the maximal optimal transport coupling of $\mathrm{N}(x, \sigma^2)$ and $\mathrm{N}(y, \sigma^2)$.

1. Draw $x' \sim \mathrm{N}(x, \sigma^2)$ and $W_x \sim \mathrm{Unif}$
2. If $W_x \sim N(x'; x, \sigma^2) \leq \mathrm{N}(x'; y, \sigma^2)$, set $y' = x'$
3. Else set $y' = t_{xy}(x')$ using the transport map t_{xy} as defined in Lemma 4.2
4. Return (x', y')

4.2 Optimal transport couplings

As noted above, the joint distribution of $(\tilde{x}, \tilde{y})$ given $\tilde{x} \neq \tilde{y}$ plays an important role in determining the distribution of meeting times. This is especially important since Lemma 3.2 shows that the probability $\mathbb{P}(x' = y' \,|\, x, y)$ of meeting proposals must be small until the chains are relatively close. Thus it is natural to consider not just ways to limit the variance of $y' - x'$ when $x' \neq y'$, but methods for making this quantity as small as possible.

Given a metric δ on $\mathbb{R}^d$, we say that $\bar{Q}((x, y), \cdot) \in \Gamma(Q(x, \cdot), Q(y, \cdot))$ is an optimal transport coupling if $\bar{Q}((x, y), \cdot)$ minimizes $\mathbb{E}_{(x', y') \sim \tilde{Q}}[\delta(x', y;)]$ among all couplings $\tilde{Q} \in \Gamma(Q(x, \cdot), Q(y, \cdot))$. In this study we set $\delta(x', y') = \|y' - x'\|^2$. Below, we show how to construct an optimal transport coupling between the residuals of a maximal coupling of $\mathrm{N}(x, I_d \sigma_d^2)$ and $\mathrm{N}(y, I_d \sigma_d^2)$. Optimal transport couplings are not usually available in closed form, but the symmetries of the multivariate normal distribution present an opportunity. We begin with the following result in one dimension:

Lemma 4.2. Suppose $\bar{Q}((x, y), \cdot) \in \Gamma^{\max}(\mathrm{N}(x, \sigma^2), \mathrm{N}(y, \sigma^2))$, and define the residual distributions $\mu(A) := \mathbb{P}(x' \in A \,|\, x' \neq y', x, y)$ and $\nu(A) = \mathbb{P}(y' \in A \,|\, x' \neq y', x, y)$ where $(x', y') \sim \bar{Q}((x, y), \cdot)$ and $A \in \mathcal{B}$. Let Φ_x and Φ_y be the cumulative distribution functions of μ and ν on $\mathbb{R}$, and define the transport map $t_{xy}(x') := \Phi_y^{-1}(\Phi_x(x'))$. If $x' \sim \mu$, then $(x', t_{xy}(x'))$ is an optimal transport coupling of μ and ν. Also Φ_x and Φ_y have the functional forms given in the proof below.

Proof. The main result is due to the cumulative distribution function characterization of optimal transport maps for non-atomic distributions on $\mathbb{R}$, see e.g. Rachev and Rüschendorf [1998, chap. 3.1]. If $x' \sim \mu$ and $y' \sim \nu$ then by Lemma 3.1, x' and y' the following CDFs:

$$\Phi_x(x') = \begin{cases} \frac{F_x(x' \wedge m) - F_y(x' \wedge m)}{F_x(m) - F_y(m)} & \text{if } x < y \\ 1 - \frac{F_x(x' \vee m) - F_y(x' \vee m)}{F_x(m) - F_y(m)} & \text{if } x \geq y \end{cases} \qquad \Phi_y(y') = \begin{cases} \frac{F_y(y' \wedge m) - F_x(y' \wedge m)}{F_y(m) - F_x(m)} & \text{if } y < x \\ 1 - \frac{F_y(y' \vee m) - F_x(y' \vee m)}{F_y(m) - F_x(m)} & \text{if } y \geq x. \end{cases}$$

Here $m = (x + y)/2$ and $F_z(\cdot)$ is the CDF of $\mathrm{N}(z, \sigma^2)$ for $z \in \mathbb{R}$. □

We say that $\bar{Q}$ is a maximal coupling with optimal transport residuals, or a maximal optimal transport coupling, if $\bar{Q}((x, y), \cdot) \in \Gamma^{\max}(Q(x, \cdot), Q(y, \cdot))$ and if the residuals of $\bar{Q}((x, y), \cdot)$ follow an optimal transport coupling. The result above suggests an algorithm for drawing from the maximal optimal transport coupling of one-dimensional normal distributions. See Algorithm 4 for the details of this method and Lemma 4.3 for a proof of its validity.

Algorithm 5 Draw from the maximal optimal transport coupling of $\mathrm{N}(x, I_d\sigma_d^2)$ and $\mathrm{N}(y, I_d\sigma_d^2)$

1. Compute $e = (y-x)/\|y-x\|$, $m = (y+x)/2$, $x_1 = x'e$, and $y_1 = y'e$
2. Draw (x_1', y_1') from the maximal optimal transport coupling of $\mathrm{N}(x_1, \sigma_d^2)$ and $\mathrm{N}(y_1, \sigma_d^2)$ using Algorithm 4
3. Draw $\tilde{z} \sim \mathrm{N}(m, I_d\sigma_d^2)$, set $z_{-1}' = (I_d - ee')\tilde{z}$, $x' = x_1'e + z_{-1}'$, and $y' = y_1'e + z_{-1}'$
4. Return (x', y')

Lemma 4.3. The output of Algorithm 4 follows a maximal optimal transport coupling of $\mathrm{N}(x, \sigma^2)$ and $\mathrm{N}(y, \sigma^2)$.

Proof. $x' \sim \mathrm{N}(x, \sigma^2)$ by construction. By Lemma 3.1 and the validity of Algorithm 1, we have $\mathbb{P}(y' \in A, y' = x') = \int_A \mathrm{N}(y'; y, \sigma^2) \wedge \mathrm{N}(y'; x, \sigma^2)\,\mathrm{d}y'$ for $A \in \mathcal{B}$. Lemmas 3.1 and 4.2 also imply $\mathbb{P}(y' \in A, y' \neq x') = \int_A (\mathrm{N}(y'; y, \sigma^2) - \mathrm{N}(y'; x, \sigma^2)) \vee 0\,\mathrm{d}y'$ for $A \in \mathcal{B}$. Together these imply $y' \sim \mathrm{N}(y, \sigma^2)$, so (x', y') follows some coupling of $\mathrm{N}(x, \sigma^2)$ and $\mathrm{N}(y, \sigma^2)$. Finally, we note that Algorithm 4 has exactly the same probability of $x' = y'$ as Algorithm 1 does when $Q(z, \cdot) = \mathrm{N}(z, \sigma^2)$. We know that the coupling implemented in Algorithm 1 is maximal, so we conclude that the present one is as well. □

Finally, we combine the result of Lemma 4.3 with the logic of Algorithm 3 to obtain an algorithm for drawing from the maximal coupling with optimal transport residuals on $\mathbb{R}^d$. See Algorithm 5 for a statement of this method and Lemma 4.4 for a proof of its validity.

Lemma 4.4. The output of Algorithm 5 follows a maximal optimal transport coupling of $\mathrm{N}(x, I_d\sigma_d^2)$ and $\mathrm{N}(y, I_d\sigma_d^2)$.

Proof. The proof that (x', y') follows a maximal coupling of $\mathrm{N}(x, I_d\sigma_d^2)$ and $\mathrm{N}(y, I_d\sigma_d^2)$ is almost identical to the argument of Lemma 4.1, except we now use the same draw for $y_{-1}' = x_{-1}'$ rather than independent draws $x_{-1}', y_{-1}' \sim \mathrm{N}(m_{-1}, (I_d - ee')\sigma_d^2)$. To see that (x', y') follows an optimal transport coupling conditional on $x' \neq y'$, we apply Theorem 2.1 of Knott and Smith [1984]. That result says that if we can write $y' = T_{xy}(x')$ such that y' has the correct distribution and $\partial T_{xy}(x')/\partial x'$ is symmetric and positive definite, then $(x', T_{xy}(x'))$ is an optimal transport coupling for $\delta(x', y') = \|y' - x'\|^2$. In this case we have $y' = T_{xy}(x') = t_{xy}(x_1')e + x_{-1}'$. Symmetry follows immediately and positive definiteness follows since t_{xy} is monotonically increasing. □

4.3 The maximal reflection coupling

Another coupling in the spirit of the previous section is the maximal coupling with reflection residuals, also called the maximal reflection coupling. It is the maximal analogue to the reflection coupling defined near the beginning of Section 4, and it has previously been considered in Eberle and Majka [2019], Bou-Rabee et al. [2020], and Jacob et al. [2020]. We say that $(x', y') \sim \bar{Q}((x, y), \cdot)$ is a maximal reflection coupling of $\mathrm{N}(x, I_d\sigma_d^2)$ and $\mathrm{N}(y, I_d\sigma_d^2)$ if it is a maximal coupling and if $x' \neq y'$ implies $\eta = (I_d - 2ee')\xi$, where we define $\xi = x' - x$ and $\eta = y' - y$. When $x' = y'$, the maximal reflection coupling yields the same distribution of (x', y') as any

Algorithm 6 Draw from the maximal reflection coupling of $\mathrm{N}(x, \sigma^2)$ and $\mathrm{N}(y, \sigma^2)$

1. Draw $x' \sim \mathrm{N}(x, \sigma^2)$ and $W_x \sim \mathrm{Unif}$
2. If $W_x \sim N(x'; x, \sigma^2) \leq \mathrm{N}(x'; y, \sigma^2)$, set $y' = x'$
3. Else set $\xi = x' - x$, $\eta = -\xi$, and $y' = y + \eta$
4. Return (x', y')

other maximal coupling, and when $x' \neq y'$ it reflects the increments of each chain over the hyperplane equidistant between x and y.

This coupling is related to the reflection coupling of diffusions described in Lindvall and Rogers [1986], Eberle [2011], Hsu and Sturm [2013], and other studies. These continuous-time reflection couplings are sometimes maximal couplings of processes, in the strong sense that they produce the fastest meeting times allowed by the coupling inequality. We will see that using the maximal reflection coupling for RWM proposals also delivers good meeting time performance. In our setting this seems to arise from a felicitous interaction between reflection couplings and the Metropolis accept/reject step. Understanding the analogy between the continuous- and discrete-time settings remains an interesting open question, especially for reflection couplings.

As with the maximal independent and optimal transport couplings, we describe an efficient method for drawing from the maximal reflection coupling. We begin with Algorithm 6, which yields draws from the maximal reflection coupling on $\mathbb{R}$. The validity of this algorithm is established in Bou-Rabee et al. [2020, sec. 2] and Jacob et al. [2020, sec. 4]. Algorithm 7 produces draws from the general form of this coupling on $\mathbb{R}^d$, and we establish the validity of this algorithm in Lemma 4.5. For the algorithm and its validity proof, recall that we have defined $z_1 := e'z$ and $z_{-1} = (I_d - ee')z$ for any $z \in \mathbb{R}^d$.

Lemma 4.5. The output (x', y') of Algorithm 7 is distributed according to a maximal reflection coupling of $\mathrm{N}(x, I_d\sigma_d^2)$ and $\mathrm{N}(y, I_d\sigma_d^2)$.

Proof. Essentially the same argument as in Lemmas 4.1 and 4.4 establishes that (x', y') follows a maximal coupling. For the reflection condition we recall that $y = m + r/2e$ and $x = m - r/2e$, which implies $y = y_1 e + m_{-1}$ and $x = x_1 e + m_{-1}$. Thus $y' - y = (y_1' - y_1)e + \zeta_{-1}$ and $x' - x = (x_1' - x_1)e + \zeta_{-1}$. We also have $(I_d - 2ee')e = -e$ and $(I_d - 2ee')(I_d - ee') = (I_d - ee')$. By the definition of Algorithm 6, $y_1' - y_1 = -(x_1' - x_1)$ when $x_1' \neq y_1'$. Thus $x' \neq y'$ implies

$$(I_d - 2ee')(x' - x) = -(I_d - 2ee')(y_1' - y_1)e + (I_d - 2ee')\zeta_{-1} = (y_1 - y_1)e + \zeta_{-1} = y' - y.$$

Algorithm 7 Draw from the maximal reflection coupling of $\mathrm{N}(x, I_d\sigma_d^2)$ and $\mathrm{N}(y, I_d\sigma_d^2)$

1. Compute $e = (y - x)/\|y - x\|$, $m = (y + x)/2$, $x_1 = x'e$, $y_1 = y'e$, and $m_{-1} = (I_d - ee')m$
2. Draw (x_1', y_1') from the maximal reflection coupling of $\mathrm{N}(x_1, \sigma_d^2)$ and $\mathrm{N}(y_1, \sigma_d^2)$, by the method of Algorithm 6
3. Draw $\zeta \sim \mathrm{N}(0, I_d\sigma_d^2)$ and set $x' = x_1'e + m_{-1} + \zeta_{-1}$, and $y' = y_1'e + m_{-1} + \zeta_{-1}$
4. Return (x', y')

We conclude that (x', y') satisfies the reflection condition when $x' \neq y'$, and so the output of Algorithm 7 follows a maximal reflection coupling of $\mathrm{N}(x, I_d\sigma_d^2)$ and $\mathrm{N}(y, I_d\sigma_d^2)$. □

4.4 Hybrid couplings

It is also possible to choose among the coupling strategies described above – or any valid coupling of the proposal distributions – depending on the current state pair (x, y). As implied by Lemma 3.2, maximal couplings of Gaussian distributions have very little chance of producing $x' = y'$ unless $r = \|y - x\|$ is relatively small. Thus we can deploy one coupling method such as a maximal coupling when r is below some threshold and a different coupling when r is above it. This is reminiscent of the two-coupling strategies used in Smith [2014], Pillai and Smith [2017], and Bou-Rabee et al. [2020]. This approach can be deployed to produce faster meeting between chains. It can also simplify some theoretical arguments since, for example, the simple reflection coupling is simpler to analyze than its maximal coupling counterpart.

5 Acceptance step couplings

Recall that we write P for the MH transition kernel generated by the proposal kernel Q and acceptance rate function a. We construct our MH kernel coupling $\bar{P} \in \Gamma(P, P)$ as follows. First, we draw (X_0, Y_0) such that $X_0, Y_0 \sim \pi_0$ for some arbitrary initial distribution π_0. We begin each iteration t by drawing proposals $(x', y') \sim \bar{Q}((x, y), \cdot)$, where $(x, y) = (X_t, Y_t)$. Then we draw acceptance indicators (b_x, b_y) from some joint distribution $\bar{B}((x, y), (x', y'))$ on $\{0, 1\}^2$. Finally we set $(X_{t+1}, Y_{t+1}) = (X, Y)$ where $X = b_x x' + (1 - b_x)x$ and $Y = b_y y' + (1 - b_y)y$. For any $x, y \in \mathbb{R}^d$ and $A \in \mathcal{B} \otimes \mathcal{B}$, we write $\bar{P}((x, y), A) := \mathbb{P}((X, Y) \in A \,|\, x, y)$ for the resulting joint transition distribution. We want $\bar{P}((x, y), \cdot) \in \Gamma(P(x, \cdot), P(y, \cdot))$, and which implies a few constraints on the acceptance indicator coupling $\bar{B}((x, y), (x', y'))$.

Given any mapping $\bar{B}$ from current state and proposal pairs $(x, y), (x', y')$ to probabilities on $\{0, 1\}^2$, we can define joint acceptance rate functions $a_x((x, y), (x', y')) := \mathbb{P}(b_x = 1 \,|\, x, y, x', y')$ and $a_y((x, y), (x', y')) := \mathbb{P}(b_y = 1 \,|\, x, y, x', y')$, where $(b_x, b_y) \sim \bar{B}((x, y), (x', y'))$. These definitions make $\bar{B}((x, y), (x', y'))$ a coupling of $\mathrm{Bern}(a_x((x, y), (x', y')))$ and $\mathrm{Bern}(a_y((x, y), (x', y')))$. We want the transition pair (X, Y) defined above to imply $X \sim P(x, \cdot)$ and $Y \sim P(y, \cdot)$ conditional on (x, y). In O'Leary and Wang [2021], this is shown to hold if $(x', y') \sim \bar{Q}((x, y), \cdot)$ for some $\bar{Q} \in \Gamma(Q, Q)$ and if

$$\mathbb{P}(b_x = 1 \,|\, x, y, x') = \mathbb{E}[a_x((x, y), (x', y')) \,|\, x, y, x'] = a(x, x') \quad \text{for } Q(x, \cdot)\text{-almost all } x'$$
$$\mathbb{P}(b_y = 1 \,|\, x, y, y') = \mathbb{E}[a_y((x, y), (x', y')) \,|\, x, y, y'] = a(y, y') \quad \text{for } Q(y, \cdot)\text{-almost all } y'.$$

These conditions are intuitive, but they allow for relatively complicated forms of a_x, a_y, and $\bar{B}$. For example, this flexibility is used in O'Leary et al. [2020] to formulate an acceptance indicator coupling $\bar{B}$ which yields a maximal transition kernel coupling $\bar{P} \in \Gamma(P, P)$ any time it is used with a maximal proposal coupling $\bar{Q} \in \Gamma(Q, Q)$. For now, we focus on acceptance

indicator couplings with $a_x((x,y),(x',y')) = a(x,x')$ and $a_y((x,y),(x',y')) = a(y,y')$. The resulting acceptance couplings take a simple form, as described in the following result.

Lemma 5.1. Suppose $(b_x, b_y) \sim \bar{B}((x,y),(x',y')) \in \Gamma(\mathrm{Bern}(a(x,x')), \mathrm{Bern}(a(y,y')))$ for state pairs $(x,y),(x',y') \in \mathbb{R}^d \times \mathbb{R}^d$. Then for some $\rho_{xy} \in [0 \vee (a(x,x') + a(y,y') - 1), a(x,x') \wedge a(y,y')]$,

$$\begin{aligned} \mathbb{P}(b_x = 1, b_y = 1) &= \rho_{xy} & \mathbb{P}(b_x = 1, b_y = 0) &= a(x,x') - \rho_{xy} \\ \mathbb{P}(b_x = 0, b_y = 1) &= a(y,y') - \rho_{xy} & \mathbb{P}(b_x = 0, b_y = 0) &= 1 - a(x,x') - a(y,y') + \rho_{xy}. \end{aligned}$$

Proof. Set $\rho_{xy} := \mathbb{P}(b_x = 1, b_y = 1)$. The values for $\mathbb{P}(b_x = 1, b_y = 0)$ and $\mathbb{P}(b_x = 0, b_y = 1)$ follow from the margin conditions, and then the value of $\mathbb{P}(b_x = 0, b_y = 0)$ follows from the requirement that $\sum_{i,j \in \{0,1\}} \mathbb{P}(b_x = i, b_y = j) = 1$. The constraints on ρ_{xy} follow from the requirement that all of these joint probabilities must fall in $[0,1]$. □

Note for example that independent draws $b_x \sim \mathrm{Bern}(a(x,x'))$ and $b_y \sim \mathrm{Bern}(a(y,y'))$ imply $\rho_{xy} = a(x,x')a(y,y')$. This satisfies the given bounds, since $\rho_{xy} = a(x,x')a(y,y') \le a(x,x') \wedge a(y,y')$ and $\rho_{xy} = a(x,x')a(y,y') \ge a(x,x') + a(y,y') - 1$ from the fact that $(1 - a(x,x'))(1 - a(y,y')) \ge 0$.

The two chains can only meet if $\tilde{x} = \tilde{y}$ is proposed and both proposals are accepted. This suggests that we should maximize the probability of $(b_x, b_y) = (1,1)$ by choosing $\rho_{xy} = a(x,x') \wedge a(y,y')$. By Lemma 5.1, this also maximizes the probability of $(b_x, b_y) = (0,0)$ and of $b_x = b_y$. Thus, this ρ_{xy} corresponds to using the maximal coupling of $\mathrm{Bern}(a(x,x'))$ and $\mathrm{Bern}(a(y,y'))$, which is unique in this case. Simulation results suggest that some couplings tend to produce contraction between chains when both proposals are accepted, no change in the separation between chains when both are rejected, and an increase in separation when one chain is accepted and the other is rejected. This further argues for drawing (b_x, b_y) from its maximal coupling conditional on (x,y) and (x',y').

Write $A \triangle B = (A \setminus B) \cup (B \setminus A)$ for the symmetric difference of $A, B \in \mathcal{B}$. As described in Section 1, it is convenient to describe acceptance indicators and their couplings in terms of uniform random variables. In particular we have the following:

Lemma 5.2. Fix $a_x, a_y \in [0,1]$. $\tilde{B} \in \Gamma(\mathrm{Bern}(a_x), \mathrm{Bern}(a_y))$ if and only if there exists a coupling $\bar{U} \in \Gamma(\mathrm{Unif}, \mathrm{Unif})$ such that $(b_x, b_y) \sim \tilde{B}$ for $b_x = 1(U \le a_x)$ and $b_y = 1(V \le a_y)$. In particular, $\mathbb{P}(b_x = b_y \,|\, x, y, x', y')$ is maximized when $U = V$ and minimized when $V = 1 - U$.

Proof. Suppose $\bar{U} \in \Gamma(\mathrm{Unif}, \mathrm{Unif})$ and (b_x, b_y) are defined as in the statement above. $b_x \sim \mathrm{Bern}(a_x)$ since $\mathbb{P}(b_x = 1) = \mathbb{P}(U \le a_x) = a_x$, and similarly for b_y. Thus the law of (b_x, b_y) is a coupling of $\mathrm{Bern}(a_x)$ and $\mathrm{Bern}(a_y)$. For the converse, by Lemma 5.1 any coupling $\tilde{B}$ will be characterized by $\rho = \mathbb{P}(b_x = b_y = 1)$. Thus we must find a coupling $\bar{U} \in \Gamma(\mathrm{Unif}, \mathrm{Unif})$ such that if $(U,V) \sim \bar{U}$ then $\mathbb{P}(U \le a_x, V \le a_y) = \rho$. One such coupling is the distribution on $[0,1]^2$

with density

$$f(u,v) = \begin{cases} \frac{\rho}{a_x a_y} & \text{if } u \le a_x, v \le a_y \\ \frac{a_y - \rho}{(1-a_x)a_y} & \text{if } u > a_x, v \le a_y \end{cases} \qquad \begin{cases} \frac{a_x - \rho}{a_x(1-a_y)} & \text{if } u \le a_x, v > a_y \\ \frac{1-a_x-a_y+\rho}{(1-a_x)(1-a_y)} & \text{if } u > a_x, v > a_y. \end{cases}$$

Note that when $U = V \sim \text{Unif}$, we obtain $\mathbb{P}(b_x = b_y = 1) = a(x,x') \wedge a(y,y')$, the maximal value of ρ, and when $1 - V = U \sim \text{Unif}$ we achieve $\mathbb{P}(b_x = b_y = 1) = 0 \vee (a(x,x') + a(y,y') - 1)$, the minimal value of ρ. By Lemma 5.1, the probability of $b_x = b_y$ is maximized when ρ is maximized and minimized when ρ is minimized. □

While the acceptance indicator couplings described above are appealing in their simplicity, we may wonder if we can do better by adapting our choice of ρ_{xy} depending on the current state pair (x, y) or the proposals (x', y'). In particular, we consider an 'optimal transport' approach to selecting ρ_{xy}, in which we aim to minimize the expected distance between state $X = b_x x' + (1 - b_x)x$ and $Y = b_y y' + (1 - b_y)y$ after the joint accept/reject step. For each $x, y, x', y' \in \mathbb{R}^d$, we solve

$$\min_{\rho} \{ \mathbb{E}[\delta(X,Y) \,|\, x, y, x', y'] \;:\; \rho \in [(a(x,x') + a(y,y') - 1) \vee 0, a(x,x') \wedge a(y,y')] \}.$$

As above, we set $\delta(X,Y) = \|Y - X\|^2$. Note that this is a linear program with linear constraints in ρ, so for typical proposal couplings $\bar{Q}$ the above will almost surely have solution either $\rho = 0 \vee (a(x,x') + a(y,y') - 1)$ or $\rho = a(x,x') \wedge a(y,y')$. Qualitatively, the lower-bound solution will be optimal when $\delta(x,y')$ and $\delta(x',y)$ are small relative to $\delta(x,y)$ and $\delta(x',y')$. Below, we see that this is uncommon for the proposal distributions we consider.

6 Simulations

We now consider a set of simulations on the relationship between coupling design and meeting times, with a focus on the role of dimension. High-dimensional target distributions are a common challenge in applications of MCMC, and previous studies such as Jacob et al. [2020] suggest that apparently similar couplings can produce unexpected and sometimes dramatic differences in meeting behavior with increasing dimension d. We expect that theory will someday provide definitive guidance on the design of couplings that scale well with dimension. For now, simulations like the following suggest the use of some couplings over others and offer a range of hypotheses for further analysis.

Our simulations target the standard multivariate normal distribution $\pi_d = \text{N}(0, I_d)$ with a range of dimensions d. We focus on the RWM algorithm with proposal kernel $Q(x, \cdot) = \text{N}(x, I_d \sigma_d^2)$ with $\sigma_d^2 = \ell^2/d$ and $\ell = 2.38$. This form of proposal variance yields an acceptance rate converging to the familiar 0.234 value as $d \to \infty$, and diffusion limit arguments suggest that this choice produces rapid or even optimal mixing behavior at stationarity [Gelman et al., 1996, Roberts et al., 1997] and in the transient phase [Christensen et al., 2005, Jourdain et al.,

Table 1: Average meeting times (1000 replications each, $d = 10$)

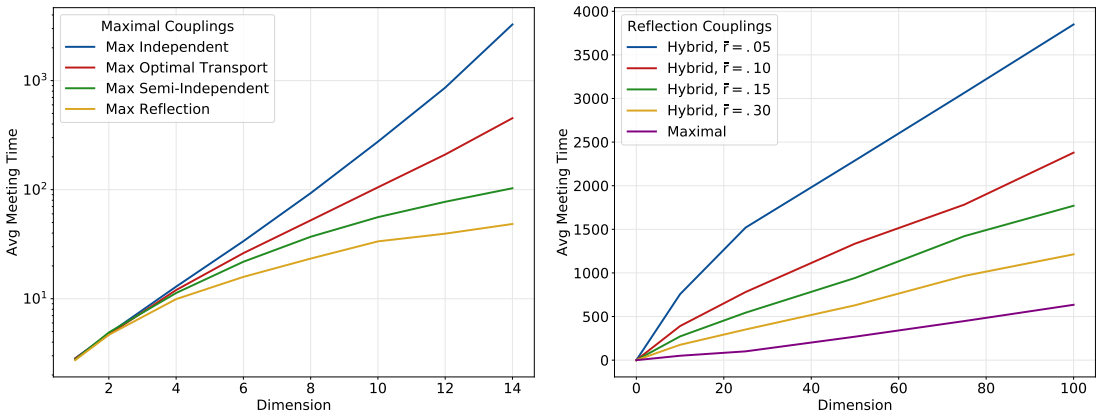
Acceptance Coupling	$V = U$		Independent		$V = 1 - U$	
Proposal Coupling	Avg τ	S.E.	Avg τ	S.E.	Avg τ	S.E.
Maximal Reflection	30	0.8	51	1.4	68	2.0
Maximal Semi-Independent	54	1.5	85	2.4	105	3.3
Maximal Optimal Transport	104	3.0	155	4.6	183	5.7
Maximal Independent	279	8.5	302	9.4	354	11.2

2014]. Meeting times typically depend on both the coupling and the marginal behavior of each chain. Our multivariate normal setting is a particularly simple one, but it allows us to isolate the effect of each coupling decision without concern about the mixing behavior of the marginal kernel.

We begin by considering a full grid of proposal and acceptance coupling combinations in $d = 10$. We initialize each chain using an independent draw from the target distribution. We then iterate until meeting occurs and record the observed meeting time over 1000 replications with each pair of coupling options. The averages and standard deviations of these meeting times τ appear in Table 1. We will consider each of these options in more detail below, but for now it is important to note a few facts. First, even in low dimension like $d = 10$, we already observe more than an order of magnitude difference in the meeting times associated with the best and worst coupling combinations. In the simulations below we generally evaluate high-performance couplings over a much wider range of dimensions than low-performance couplings to avoid unnecessary computational expense.

Second, among the acceptance couplings considered above, the $U = V$ coupling consistently outperforms the $U, V \overset{iid}{\sim} \text{Unif}$ coupling, which consistently outperforms the $V = 1 - U$ coupling. These relationships hold for each choice of proposal coupling. A similar relationship exists among the proposal couplings, with the maximal reflection coupling delivering the best meeting times and the maximal independent coupling delivering the worst ones, again for any choice of acceptance indicator coupling. These robust and monotone relations also hold in higher and lower dimensions and with other initialization and proposal variance options. Although the meeting times shown here arise from a complex interplay of proposal and acceptance behavior, these simulations suggest that some options can be regarded as generally better or worse than others.

In this exercise and in many of the simulations below, we initialize chains using independent draws from the target distribution. When $\pi_d = \text{N}(0, I_d)$, the initialization method does not seem to affect the relative performance of the couplings considered below. This may not hold for all targets, e.g. mixture distributions with well-separated modes. The development of couplings and initialization strategies to address especially challenging targets stands as an important topic for future research.



(a) Maximal proposal distribution couplings (b) Maximal and hybrid reflection couplings

Figure 3: Scale behavior of the average meeting time of coupled RWM chains as a function of the dimension of the target distribution $\pi = \mathrm{N}(0, I_d)$, under proposal distribution coupling options defined in Section 4. The chains are initialized at independent draws from the target and a $V = U$ coupling is used at the Metropolis step. Left: Scaling under four maximal couplings of the proposal distribution. Right: Scaling under the maximal reflection coupling and simple/maximal reflection coupling hybrids under a range of cutoff parameters $\bar{r}$.

6.1 Proposal couplings

We now consider the relationship between proposal couplings and meeting times in more detail. As above, we initialize each chain with an independent draw from the target distribution. Here we use the $U = V \sim$ Unif coupling at the Metropolis step, which maximizes the probability of making the same accept/reject decision for both chains. As illustrated in Table 1, this acceptance coupling seems to produce the fastest meeting times for a range of proposal couplings. For each proposal coupling and dimension, we run 1000 pairs of chains until meeting occurs. The average meeting times from this test appear in Figure 3.

In Figure 3a, we show the average meeting times for the maximal couplings with independent, optimal transport, semi-independent, and reflection residuals, as defined in Section 4. Figure 3b presents the corresponding results for the maximal-reflection coupling and for hybrid couplings that deploy the maximal reflection coupling when $r = ||y - x|| < \bar{r}/\sqrt{d}$ but use the simple reflection coupling when the chains are further apart. We consider hybrid couplings with a range of values of the cutoff parameter $\bar{r}$.

These results suggest that meeting times grow exponentially in dimension under the maximal coupling with independent residuals, close to linearly in dimension under the maximal reflection coupling, and somewhere in between for the other two maximal couplings. The hybrid couplings and maximal reflection coupling show a similar order of dependence on dimension, and the hybrid couplings display an inverse relationship between average meting time and $\bar{r}$. This reflects an increasing number of missed opportunities to meet under the hybrid couplings, since smaller values of $\bar{r}$ result in more situations when $r \geq \bar{r}/\sqrt{d}$ even though r is small enough to produce a reasonable probability of meeting under a maximal coupling.

Any maximal coupling of proposal distributions produces meeting proposals with the same

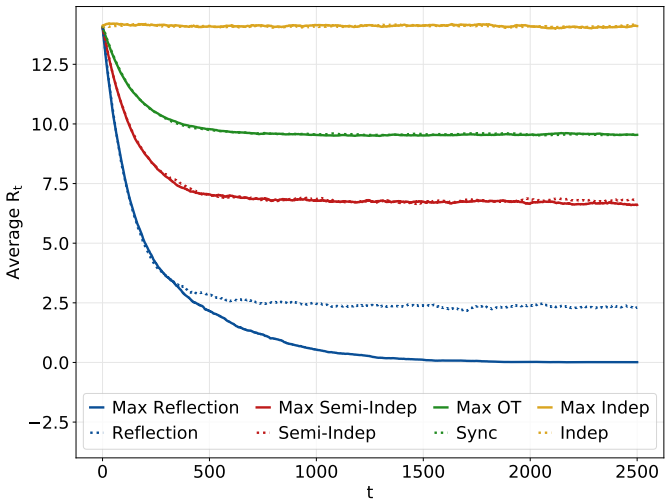


Figure 4: Average distance between chains over 1000 replications as a function of iteration t. Here $d = 100$, and as in Figure 3 these chains are initialized at independent draws from the target. At the meeting time τ, chains switch to a sticky coupling in which $x' = y'$ and $U = V$. Maximal couplings and their non-maximal equivalents produce similar dynamics until R_t is small enough for there to be a reasonable chance of meeting. This effect is visible for the reflection couplings but not the other options, which stop contracting before the chains are close enough for these effects to make a difference.

probability, as a function of $R_t = \|Y_t - X_t\|$. Thus the variation in average τ reflects differences in coupled chain dynamics conditional on not meeting. The degree of contraction between chains seems to play a particularly important role. As noted above, just after Lemma 3.2, the chains must be within a distance $r_d = \mathcal{O}(1/\sqrt{d})$ to maintain a fixed probability of proposed meetings as d increases. Thus the combination of a proposal and acceptance coupling must generate contraction $\mathbb{E}[R_{t+1} - R_t \,|\, X_t, Y_t] < 0$ to within a range r_d to avoid a fall-off in meeting probability as a function of dimension. The results above suggest that some proposal couplings do this better than others.

To visualize this behavior, we run 1000 pairs of coupled chains under a range of maximal and non-maximal couplings, as described in Section 4. We fix $d = 100$, initialize chains independently from the target, and use the $U = V \sim$ Unif coupling at the accept reject/step. We run all pairs of chains for 2500 iterations and use the sticky coupling described in Section 2 to maintain $X_t = Y_t$ for $t \geq \tau$. Finally, we compute $R_t = \|Y_t - X_t\|$ and plot the average distance over replications as a function of the iteration t. See Figure 4 for these results.

The dynamics produced in this exercise provide a compelling explanation for the meeting time behavior observed in Figure 3. In the absence of meeting, each coupling seems to produce contraction down to a certain degree of separation between chains. For the maximal independent coupling that appears to be almost exactly $\mathbb{E}[\|Y_0 - X_0\|]$, the distance obtained by independent draws from the target distribution. The maximal optimal transport coupling and maximal semi-independent coupling produce contraction to within a smaller radius. The explosive increase in meeting times under these couplings suggests that these critical distances do not keep pace with the $\mathcal{O}(1/\sqrt{d})$ rate noted above. By the same token, the maximal reflection coupling appears

to produce sufficient contraction to eventually meet with high probability. Among the four maximal couplings considered here and their four non-maximal counterparts, only the maximal reflection coupling produces a high enough meeting probability to eventually diverge from its non-maximal counterpart.

We can also visualize these differences in drift directly, by creating pairs (X_t, Y_t) with a specific $R_t = r$, running a single step of the coupled MH kernel, and recording the resulting distance R_{t+1}. We show the output of such a test in Figure 5. Again we set $d = 100$ and use the $U = V \sim \text{Unif}$ coupling at the acceptance step. We consider a range of r values and initialize (x, y) to have $e = (1, 0, \dots)$, $m_1 = 1$, and $||m|| = \sqrt{d}$. In this case we run 10,000 replications for each coupling and r value. Consistent with the results above, we find that the different proposal couplings display a range of contraction behavior as a function of the distance between the chains, although all are contractive when the chains are far apart and repulsive when the chains are close together. Except for the reflection coupling, the R_t value where each contraction line crosses the x-axis corresponds to the long-run average value of R_t, as one would expect for a chain in close to a stable equilibrium around this point.

We conclude by noting that the meeting time, separation, and drift behavior illustrated in the plots above agrees with our expectations in some cases more than others. For instance, it is not surprising that using a maximal coupling with independent residuals in the proposal step produces poor contraction behavior as a function of dimension. Although the proposal variance σ_d^2 shrinks in d, this coupling produces almost independent values of X and Y conditional on $x' \neq y'$, whose separation can be expected to increase linearly in the number of independent dimensions. Thus the flat line in Figure 4 agrees with intuition.

Each of the other three couplings has the property that $x'_{-1} = y'_{-1}$ when $x' \neq y'$, which limits the potential for variance from these components as a function of dimension. However, the relative performance of the maximal semi-independent, maximal optimal transport, and maximal reflection couplings is almost the opposite of what one might expect. The optimal transport coupling seems to produce the least contraction in spite of producing the smallest values of $\|x' - y'\|$ conditional on $x' \neq y'$. At the same time, the reflection coupling appears to produce the most contraction despite maximizing the variance of $y'_1 - x'_1$. These differences seem likely to stem from the interaction of these couplings with the acceptance step.

6.2 Acceptance couplings

Next we consider couplings of the accept/reject step. As noted in Section 5, we focus on acceptance indicator couplings that accept both chains at exactly the MH rate for any pair of proposal states. In light of Lemma 5.2, it is convenient to define acceptance indicators $b_x = 1(U \leq a(x, x'))$ and $b_y = 1(V \leq a(y, y'))$ in terms of underlying uniform random variables. We can realize three basic couplings by drawing $U \sim \text{Unif}$ and then either drawing $V \sim \text{Unif}$ independently, setting $V = U$, or setting $V = 1 - U$. The $V = U$ coupling maximizes the probability of $b_x = b_y$ while the $V = 1 - U$ coupling minimizes it. We also consider the

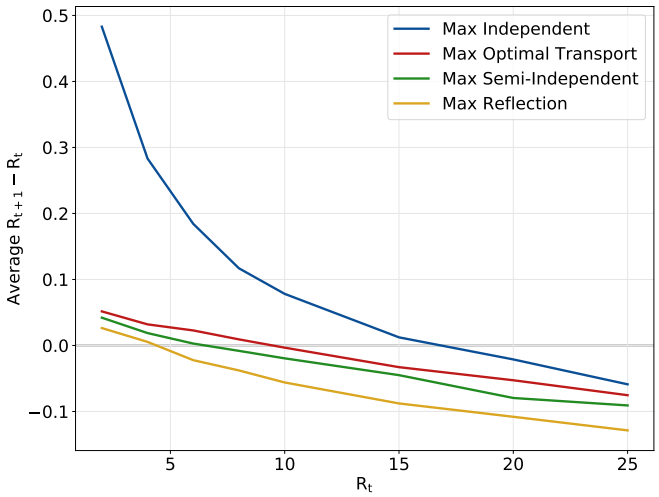


Figure 5: Average contraction as a function of the current distance between chains for four maximal couplings. For each point $R_t = r$ we construct a state pair such that $\|Y_t - X_t\| = r$. We then run one RWM iteration with the specified proposal coupling and the $U = V$ acceptance coupling, and record the resulting $R_{t+1} = \|Y_{t+1} - X_{t+1}\|$. We compute averages over 10,000 replications for each coupling and r to obtain the curves shown here. The depicted drift behavior appears consistent with the meeting times and time series dynamics of Figures 3 and 4.

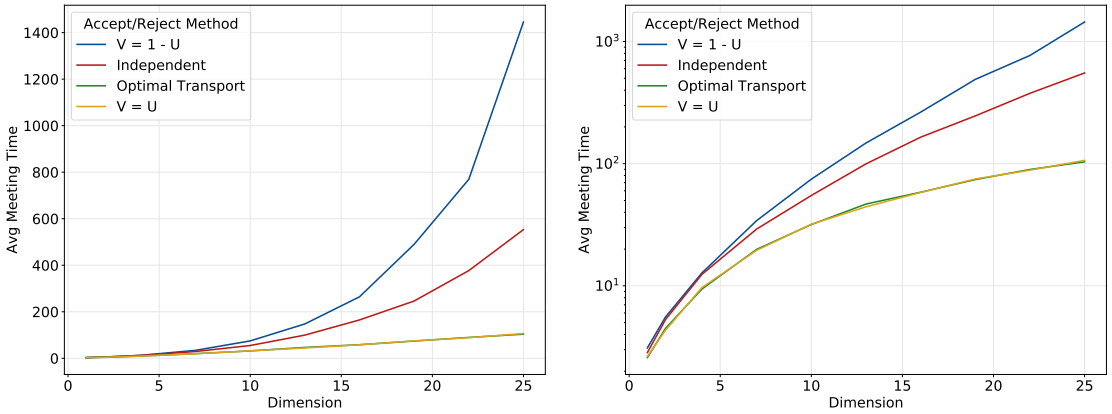
'optimal transport' acceptance indicator coupling described in Section 5. We recall that this option almost surely coincides with either the $V = U$ or $V = 1 - U$ coupling at each iteration, depending on which of these minimizes the expected distance $\mathbb{E}[\|Y - X\|^2 \,|\, x, y, x', y']$.

We present simulation results on these four acceptance step couplings in Figure 6. In each case we use the maximal reflection coupling of proposal distributions and initialize using independent draws from the target. In Figure 6a, we see that the $V = U$ coupling produces meeting times that scale approximately linearly in dimension, while these increase more rapidly under the independent and $V = 1 - U$ couplings. The log-linear plot in Figure 6b suggests that these latter meeting times may still be less than exponential in dimension.

The optimal transport and $U = V$ couplings produce nearly identical results when applied to the maximal reflection coupling of proposal distributions. A closer look at this scenario reveals that the optimal transport coupling coincides with the $U = V$ coupling in all 1000 replications when $d > 1$ and in approximately 96.6% of replications in the $d = 1$ case. We observe qualitatively identical behavior when the proposals are maximally coupled with optimal transport and semi-independent residuals.

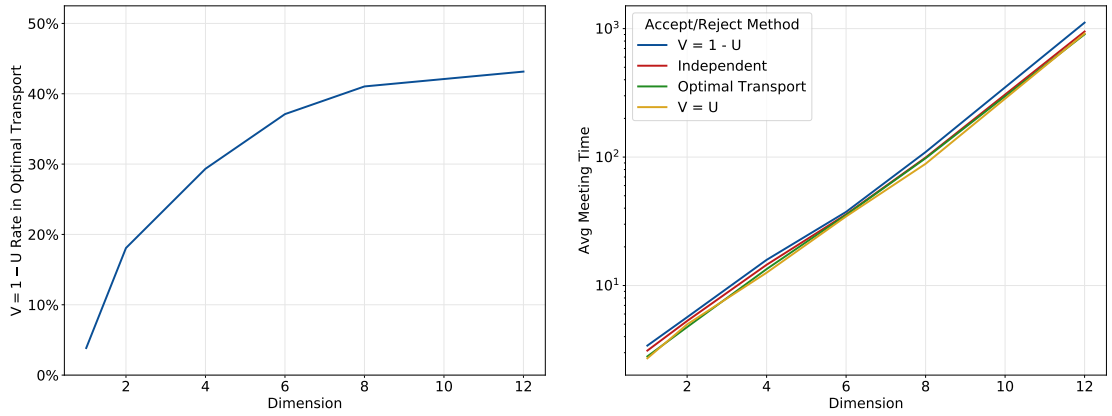
Figure 7 shows that the acceptance step optimal transport coupling displays more complex behavior when the proposal distributions follow a maximal coupling with independent residuals. Here $V = 1 - U$ is optimal in a fraction of iterations going to 50% as d increases. Nevertheless, the resulting meeting times are almost indistinguishable from the meeting times delivered by the $V = U$ and $V = 1 - U$ couplings. This suggests that under the maximal coupling with independent residuals, the rapid growth in meeting times is due to the proposal coupling more than any particular choice of acceptance indicator coupling.

As in the case of proposal couplings, we can also understand the meeting times associated



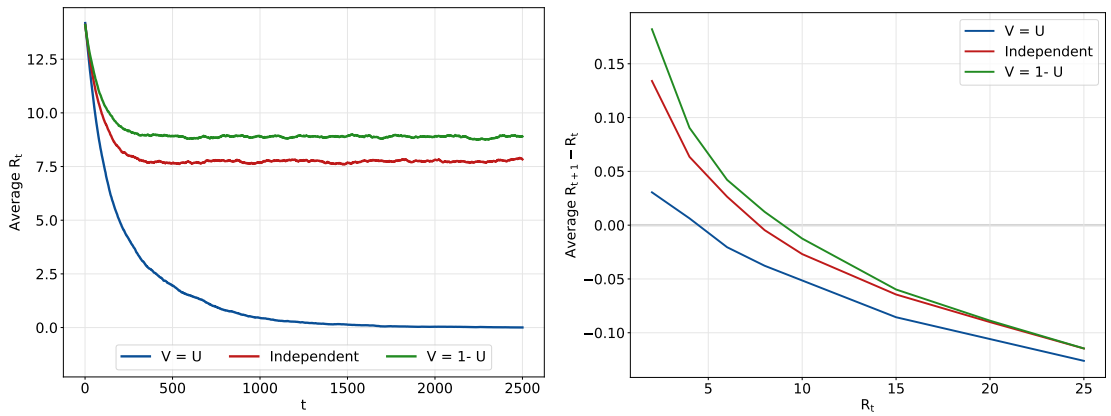
(a) Linear scale (b) Log scale

Figure 6: Scale behavior of the average meeting time of coupled MH chains as a function of the dimension of the target distribution $\pi = \mathrm{N}(0, I_d)$, under various acceptance step couplings defined in Section 5. Here the chains are initialized at independent draws from the target, the proposal distributions are related by a maximal reflection coupling, and averages are taken over 1000 replications for each dimension and coupling. Left: the $V = 1 - U$ and independent V, U methods produce meeting times which grow rapidly in dimension, in contrast to the $V = U$ and optimal transport couplings, which produce approximately linear behavior. Right: A log scale plot suggests that the $V = 1 - U$ and independent methods grow at less than an exponential rate, which would appear as a straight line on this plot.



(a) Optimal transport coupling choice (b) Meeting times (log scale)

Figure 7: Behavior of the optimal transport acceptance step coupling, as a function of the dimension of the target distribution $\pi = \mathrm{N}(0, I_d)$. Here the proposal distributions are related by a maximal coupling with *independent* residuals, and as usual the chains are initialized by independent draws from the target distribution and replicated 1000 times per coupling and dimension. Left: the optimal transport coupling coincides with the $V = 1 - U$ coupling at a rate approaching 50% as the dimension d of the target increases. Right: all four acceptance step couplings deliver similar, exponentially growing meeting times under the maximal independent coupling of proposal distributions.



(a) Average separation by iteration t (b) Contraction vs. distance between chains

Figure 8: Time series and drift function properties of the distance between chains under the three simple acceptance step couplings. Left: we repeat the experiment shown in Figure 4 for the acceptance step couplings. Only the $V = U$ option allows the chains to get close enough to produce a significant probability of meeting. Right: we repeat the experiment shown in Figure 5 for these couplings. The $V = U$ coupling displays more contraction and a smaller x-intercept than the other options. The behavior of this intercept as a function of d has a significant effect on meeting times.

with different acceptance indicator couplings in terms of the contraction between chains. In Figure 8a, we present the average distance between chains under the three simple acceptance indicator couplings. We run all of these using the maximal reflection coupling of proposal distributions. As in the proposal coupling case we set $d = 100$, we initializing each chain with an independent draw from the target, and we use a sticky coupling to ensure $X_t = Y_t$ for $t \geq \tau$. The $U = V$ coupling is able to produce sufficient contraction for meeting to take place, while this seems out of reach for the independent and $V = 1 - U$ couplings.

We also consider the effect of different acceptance couplings on the drift $R_{t+1} - R_t$ as a function of the current distance between chains R_t. We use the maximal reflection coupling at the proposal step, and we run 10,000 replications for each value of R_t and coupling option. As in the case of proposal couplings, the time series behavior of R_t shown in Figure 8a is consistent with relationship between R_t and $\mathbb{E}[R_{t+1} - R_t \,|\, X_t, Y_t]$ observed in Figure 8b.

Both of these tests support the impression that the choice of the acceptance coupling has a significant effect on the contraction properties of the resulting chains. At a high level, it appears that the right combination of proposal and acceptance strategies can lead to powerful contraction between chains down to a point where meeting is reasonably probable under a maximal coupling. The combination of a maximal reflection proposal coupling and a maximal acceptance coupling has this property while most other combinations do not, leading to a rapid growth in meeting times as a function of dimension.

7 Discussion

In the sections above we have identified a range of options for use in the design of RWM transition kernel couplings. Our analysis and simulations suggest a few principles for the choice of these elements, which we summarize as follows.

First, the coupling inequality imposes a significant constraint on the ability of any $\bar{Q} \in \Gamma(Q, Q)$ to propose meetings. This suggests using a maximal or nearly maximal coupling to obtain meetings at the highest rate possible. A hybrid approach may also be practical in some cases. When a meeting is not proposed, it seems advantageous to minimize the degrees of freedom in the displacement $y' - x'$ between proposals. These degrees of freedom accumulate in higher dimensions and eventually create a barrier to contraction between chains. This may explain the poor performance of the maximal independent coupling relative to the maximal semi-independent coupling.

Since the probability of a meeting is typically small until the chains are close together, it is important to construct a transition kernel coupling that yields strong and persistent contraction between chains. Surprisingly, the reflection couplings seem to do the best job of this among the proposal options considered above. These couplings do not have good contraction properties on their own, but they seem to set up a favorable interaction with the Metropolis step, especially with the $U = V$ coupling. The precise nature of this interaction is an important open question. For now, it appears safe to recommend the reflection coupling for inducing contraction between chains.

The success of the reflection coupling raises two additional questions. First, we may consider the extent to which this behavior depends on the log-concavity of the target distribution. It seems reasonable to think that this coupling may not work as well with irregular targets. With log-concave targets, like $\pi_d = \mathrm{N}(0, I_d)$, we can also ask how close the MH transition kernels based on a maximal coupling with reflection residuals at the proposal step comes to a maximal coupling with optimal transport residuals of the transition kernels themselves. This question seems amenable to either theoretical and numerical methods.

On the acceptance indicator side, the $U = V$ coupling has a strong a priori appeal. This coupling gives the highest chance of turning a proposed meeting into an actual meeting. It also minimizes the probability of accepting one proposal and rejecting the other, which often leads to a jump in the distance between chains. While the $U = V$ coupling dominates the other options in this study, we recall that we have focused our attention on the subset of acceptance indicator couplings in which the conditional acceptance rates $a_x((x, y), (x', y'))$ and $a_y((x, y), (x', y'))$ agree with the MH rates $a(x, x')$ and $a(y, y')$. The analysis of more general acceptance couplings deserves further attention.

We emphasized the simple case of a multivariate normal target distribution in the simulations above. It would be interesting to know the extent to which our conclusions generalize to more challenging examples such as targets with heavy tails, multi-modality, difficult geometries, and examples in large discrete state spaces. One might also extend the coupling strategies described

above to other common MH algorithms such HMC [Duane et al., 1987, Neal, 1993, 2011], the Metropolis-adjusted Langevin algorithm [Roberts and Tweedie, 1996], and particle MCMC [Andrieu et al., 2010]. We expect that couplings for these extensions would involve some of the same principles as above, but with more moving parts and fewer symmetries to exploit.

Perhaps the most important open questions in the area of coupling design concern the development of theoretical tools to relate proposal and acceptance options to meeting times. Such tools would enable a systematic understanding of the interaction between proposal and acceptance steps. This would also support work on how to pair these to produce as much possible contraction as possible between chains. One approach to this might exploit the drift and minorization approach of Rosenthal [1995, 2002], especially the pseudo-small set concept of Roberts and Rosenthal [2001]. The analyses and simulations above mark a step forward in our understanding of the options for coupling MH transition kernels. They suggest that some options might be better than others and hint at why.

Acknowledgements

The author thanks Pierre E. Jacob, Yves Atchadé, and Niloy Biswas for their helpful comments. He also gratefully acknowledge support by the National Science Foundation through grant DMS-1844695.

References

M. Abramowitz, I. A. Stegun, and R. H. Romer. *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables*. American Association of Physics Teachers, 1988. ISBN 0002-9505. 6

D. Aldous. Random walks on finite groups and rapidly mixing Markov chains. In *Séminaire de Probabilités XVII 1981/82*, pages 243–297. Springer, 1983. 1

C. Andrieu, A. Doucet, and R. Holenstein. Particle Markov chain Monte Carlo methods. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 72(3):269–342, 2010. 24

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

B. Böttcher. Markovian maximal coupling of Markov processes. *arXiv preprint arXiv:1710.09654*, 2017. 2

N. Bou-Rabee, A. Eberle, and R. Zimmer. Coupling and convergence for Hamiltonian Monte Carlo. *Annals of Applied Probability*, 30(3):1209–1250, 2020. 11, 12, 13

K. Burdzy and W. S. Kendall. Efficient Markovian couplings: examples and counterexamples. *Annals of Applied Probability*, pages 362–409, 2000. 2

O. F. Christensen, G. O. Roberts, and J. S. Rosenthal. 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, 2005. 15

D. Dey, P. Dutta, and S. Biswas. A note on faithful coupling of Markov chains. *arXiv preprint arXiv:1710.10026*, 2017. 3

W. Doeblin. Exposé de la théorie des chaînes simples constantes de Markov à un nombre fini d'états. *Mathématique de l'Union Interbalkanique*, 2(77-105):78–80, 1938. 1

R. Douc, E. Moulines, P. Priouret, and P. Soulier. *Markov Chains*. Springer, 2018. 2, 4

S. Duane, A. D. Kennedy, B. J. Pendleton, and D. Roweth. Hybrid Monte Carlo. *Physics Letters B*, 195(2):216–222, 1987. 7, 24

L. Duembgen. Bounding standard gaussian tail probabilities. *arXiv preprint arXiv:1012.2063*, 2010. 6

D. B. Dunson and J. Johndrow. The Hastings algorithm at fifty. *Biometrika*, 107(1):1–23, 2020. 1

A. Eberle. Reflection coupling and Wasserstein contractivity without convexity. *Comptes Rendus Mathematique*, 349(19-20):1101–1104, 2011. 12

A. Eberle and M. B. Majka. Quantitative contraction rates for Markov chains on general state spaces. *Electronic Journal of Probability*, 24, 2019. 11

J. A. Fill. An interruptible algorithm for perfect sampling via Markov chains. In *Proceedings of the Twenty-Ninth Annual ACM Symposium on Theory of Computing*, pages 688–695, 1997. 1

J. M. Flegal and R. Herbei. Exact sampling for intractable probability distributions via a Bernoulli factory. *Electronic Journal of Statistics*, 6:10–37, 2012. 1

A. Gelman, G. O. Roberts, and W. R. Gilks. Efficient metropolis jumping rules. *Bayesian Statistics*, 5(599-608):42, 1996. 15

M. Gerber and A. Lee. Discussion on the paper by Jacob, O'Leary, and Atchadé. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 82(3):584–585, 2020. 4

C. R. Givens and R. M. Shortt. A class of Wasserstein metrics for probability distributions. *The Michigan Mathematical Journal*, 31(2):231–240, 1984. 7

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

S. Goldstein. Maximal coupling. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete*, 46(2):193–204, 1979. 2

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

D. Griffeath. A maximal coupling for Markov chains. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete*, 31(2):95–106, 1975. 2

T. E. Harris. On chains of infinite order. *Pacific Journal of Mathematics*, 5(Suppl. 1):707–724, 1955. 1

W. K. Hastings. Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 57(1):97–109, 1970. 2

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

E. P. Hsu and K.-T. Sturm. Maximal coupling of Euclidean Brownian motions. *Communications in Mathematics and Statistics*, 1(1):93–104, 2013. 2, 12

P. E. Jacob, J. O'Leary, and Y. 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. 1, 2, 3, 11, 12, 15

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

V. 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. 1, 2

B. Jourdain, T. Lelièvre, and B. Miasojedow. Optimal scaling for the transient phase of Metropolis Hastings algorithms: The longtime behavior. *Bernoulli*, 2014. doi: 10.3150/13-BEJ546. 15

M. Knott and C. S. Smith. On the optimal mapping of distributions. *Journal of Optimization Theory and Applications*, 43(1):39–49, 1984. 7, 11

V. S. A. Kumar and H. Ramesh. Coupling vs. conductance for the Jerrum-Sinclair chain. *Random Structures and Algorithms*, 2001. doi: 10.1002/1098-2418(200101)18:1<1::AID-RSA1>3.0.CO;2-7. 3

D. A. Levin, Y. Peres, and E. L. Wilmer. *Markov Chains and Mixing Times*, volume 107. American Mathematical Soc., 2017. ISBN 1470429624. 4

T. Lindvall. *Lectures on the Coupling Method*. Dover Books on Mathematics, 1992. ISBN 0-486-42145-7. 8

T. Lindvall and L. C. G. Rogers. Coupling of multidimensional diffusions by reflection. *The Annals of Probability*, 14(3):860–872, 1986. 12

N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller. Equation of state calculations by fast computing machines. *The Journal of Chemical Physics*, 21(6): 1087–1092, 1953. 2

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

R. M. Neal. Bayesian learning via stochastic dynamics. In *Advances in Neural Information Processing Systems*, pages 475–482, 1993. 7, 24

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

R. M. Neal. MCMC using Hamiltonian dynamics. *Handbook of Markov Chain Monte Carlo*, 2 (11):2, 2011. 7, 24

J. O'Leary and G. Wang. Transition kernel couplings of the Metropolis-Hastings algorithm. *arXiv preprint arXiv:2102.00366*, 2021. 2, 3, 13

J. O'Leary, G. Wang, and P. E. Jacob. Maximal couplings of the Metropolis–Hastings algorithm. *arXiv preprint arXiv:2010.08573*, 2020. 13

N. S. Pillai and A. Smith. Kac's walk on n-sphere mixes in $n \log n$ steps. *The Annals of Applied Probability*, 27(1):631–650, 2017. 13

D. Piponi, M. Hoffman, and P. Sountsov. Hamiltonian Monte Carlo swindles. *Proceedings of Machine Learning Research*, 108:3774–3783, 26–28 Aug 2020. 1

J. Pitman. On coupling of Markov chains. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete*, 35(4):315–322, 1976. 1

D. Pollard. *Asymptopia*. Yale University, Department of Statistics, 2005. 5

J. G. Propp and D. B. Wilson. Exact sampling with coupled Markov chains and applications to statistical mechanics. *Random Structures and Algorithms*, 9(1-2):223–252, 1996. 1

S. T. Rachev and L. Rüschendorf. *Mass Transportation Problems*, volume 1. Springer Science & Business Media, 1998. 10

G. O. Roberts and J. S. Rosenthal. Small and pseudo-small sets for Markov chains. *Stochastic Models*, 17(2):121–145, 2001. 24

G. O. Roberts and R. L. Tweedie. Exponential convergence of Langevin distributions and their discrete approximations. *Bernoulli*, 2(4):341–363, 1996. 24

G. O. Roberts, A. Gelman, and W. R. Gilks. Weak convergence and optimal scaling of random walk Metropolis algorithms. *The Annals of Applied Probability*, 7(1):110–120, 1997. 15

J. S. Rosenthal. Minorization conditions and convergence rates for Markov chain Monte Carlo. *Journal of the American Statistical Association*, 90(430):558–566, 1995. 1, 24

J. S. Rosenthal. Faithful couplings of Markov chains: now equals forever. *Advances in Applied Mathematics*, 18(3):372–381, 1997. 3

J. S. Rosenthal. Quantitative convergence rates of Markov chains: A simple account. *Electronic Communications in Probability*, 7:123–128, 2002. 24

A. Smith. A Gibbs sampler on the n-simplex. *The Annals of Applied Probability*, 24(1):114–130, 2014. 13

H. Thorisson. *Coupling, Stationarity, and Regeneration*, volume 14 of *Probability and Its Applications*. Springer New York, 2000. 4

L. N. Vaserstein. Markov processes over denumerable products of spaces, describing large systems of automata. *Problemy Peredachi Informatsii*, 5(3):64–72, 1969. 8