MEXIT: Maximal un-coupling times for stochastic processes[☆☆]

Philip A. Ernst[1,*], Wilfrid S. Kendall [2,**], Gareth O. Roberts [2], Jeffrey S. Rosenthal [3]

Abstract

Classical coupling constructions arrange for copies of the *same* Markov process started at two *different* initial states to become equal as soon as possible. In this paper, we consider an alternative coupling framework in which one seeks to arrange for two *different* Markov (or other stochastic) processes to remain equal for as long as possible, when started in the *same* state. We refer to this "un-coupling" or "maximal agreement" construction as *MEXIT*, standing for "maximal exit". After highlighting the importance of un-coupling arguments in a few key statistical and probabilistic settings, we develop an explicit *MEXIT* construction for stochastic processes in discrete time with countable state-space. This construction is generalized to random processes on general state-space running in continuous time, and then exemplified by discussion of *MEXIT* for Brownian motions with two different constant drifts.

Keywords: adaptive MCMC, copula, coupling, diffusions, Fréchet class, Hahn-Jordan decomposition, Markovian coupling, MCMC, meet measure, MEXIT, one-step minorization, pseudo-marginal MCMC, recognition lemma for maximal coupling, stochastic processes, un-coupling

2010 MSC: Primary: 60J05, 60J25, Secondary: 60J60

[☆]Submitted November 3rd, 2017

[☆☆]This is a theoretical research paper and, as such, no new data were created during this study.

[*]Corresponding author. E-mail: philip.ernst@rice.edu

[**]E-mail: w.s.kendall@warwick.ac.uk

E-mail: gareth.o.roberts@warwick.ac.uk

E-mail: jeff@math.toronto.edu

[1]Department of Statistics, Rice University, Houston, TX 77005, USA

[2]Department of Statistics, University of Warwick, Coventry CV5 6FQ, UK

[3]Department of Statistics, University of Toronto, Toronto, Ontario, Canada M5S 3G3

1. Introduction

Coupling is a device commonly employed in probability theory for learning about distributions of certain random variables by means of judicious construction in ways which depend on other random variables (Lindvall [15] and Thorisson [31]). Such coupling constructions are often used to prove convergence of Markov processes to stationary distributions (Pitman [21]), especially for Markov chain Monte Carlo (MCMC) algorithms (Roberts and Rosenthal [24, and references therein]), by seeking to build two different copies of the *same* Markov process started at two *different* initial states in such a way that they become equal at a fast rate. Fastest possible rates are achieved by the *maximal coupling* constructions which were introduced and studied in Griffeath [11], Pitman [21], and Goldstein [10]. Our results and methods are closely related to the work of Goldstein [10], who deals with rather general discrete-time random processes. Our situation is related to a time-reversal of the situation studied by Goldstein [10]. However our approach seems more direct.

In the current work, we consider what might be viewed as the dual problem where coupling is used to try to construct two *different* Markov (or other stochastic) processes which remain equal for as long as possible, when they are started in the *same* state. That is, we move from consideration of the coupling time to focus on the *un-coupling time* at which the processes diverge, and try to make that as *large* as possible. We refer to this as *MEXIT* (standing for "maximal exit" time). While finalizing our current work, it came to our attention that this construction is the same as the *maximal agreement coupling time* of the August 2016 work of Völlering [32], who additionally derives a lower bound on *MEXIT*. Nonetheless, we believe the current work complements Völlering [32] well. It offers an explicit treatment of discrete-time countable-state-space, generalizes the continuous-time case, and discusses a number of significant applications of *MEXIT*. We note that the work of Völlering [32] does not consider the continuous-time case.

In addition to being a natural mathematical question, *MEXIT* has direct applications to many key statistical and probabilistic settings (see Section 2 below). In particular, couplings which are *Markovian* and *faithful* (Rosenthal [28], i.e. couplings which preserve the marginal update distributions even when conditioning on both processes; alternatively "co-adapted" or "immersion", depending on the extent to which one wishes to emphasize the underlying filtration as in Burdzy and Kendall [5] and Kendall [13]) are the most straightforward to construct, but often are *not* maximal, while more complicated non-Markovian and non-faithful couplings lead to stronger bounds. The same is true in the context of *MEXIT*.

2. Applications

To motivate the natural role of *MEXIT* in the existing literature, we first consider the role of un-coupling arguments in a few statistical and probabilistic settings.

2.1. Bounds on accuracy for statistical tests

Un-coupling has an impact on the theory classical statistical testing. In Farrell [9], amongst other sources, some function of the data (but not the data itself) is assumed to have been observed. A statistical test is then constructed to enable detection of the distribution from which the observed data have been sampled. For example, suppose that data $X_1, X_2, \ldots$ are generated as a draw either from a multivariate probability distribution $\mathbb{P}_1$ or from a multivariate probability distribution $\mathbb{P}_2$. The goal is to determine whether the data was generated from $\mathbb{P}_1$ or from $\mathbb{P}_2$. For some function h of the data, and some acceptance region A, the statistical test decides in favor of $\mathbb{P}_1$ if $h(X_1, \ldots, X_n) \in A$ and otherwise decides in favor of $\mathbb{P}_2$.

Suppose that there exists an un-coupling time T, such that if $X_1, X_2, \ldots$ are generated from $\mathbb{P}_1$, and if $Y_1, Y_2, \ldots$ are generated from $\mathbb{P}_2$ then it is exactly the case that $X_i = Y_i$ for all $1 \leq i \leq T$ (so that $X_i \neq Y_i$ for all $i > T$). We use $\mathbb{P}$ to refer to the joint distribution (in fact, the coupling) of $\mathbb{P}_1$ and $\mathbb{P}_2$.

The following proposition uses the un-coupling probabilities to recover a lower bound on the accuracy of such statistical tests related to Farrell [9, Theorem 1].

Proposition 1. *Under the above assumptions, the sum of the probabilities of Type-I and Type-II errors of our statistical test is at least* $\mathbb{P}\left[T > n\right]$.

Proof. We apply elementary arguments to the sum of the probabilities of Type-I and Type-II errors:

$$
\begin{aligned}
\mathbb{P}_2[h(Y_1,\ldots,Y_n) \in A] + \mathbb{P}_1[h(X_1,\ldots,X_n) \notin A] \quad = & \\
= \quad & \mathbb{P}_2[h(Y_1,\ldots,Y_n) \in A] + 1 - \mathbb{P}_1[h(X_1,\ldots,X_n) \in A] \\
= \quad & 1 - \Big(\mathbb{P}_1[h(Y_1,\ldots,Y_n) \in A] - \mathbb{P}_2[h(X_1,\ldots,X_n) \in A]\Big) \\
\geq \quad & 1 - |\mathbb{P}_1[h(Y_1,\ldots,Y_n] \in A] - \mathbb{P}_2[h(X_1,\ldots,X_n) \in A]| \\
\geq \quad & 1 - \|\mathcal{L}_{\mathbb{P}_1}(X_1,\ldots,X_n) - \mathcal{L}_{\mathbb{P}_2}(Y_1,\ldots,Y_n)\| \quad \text{(definition of total variation distance)} \\
\geq \quad & 1 - \mathbb{P}\left[X_i \neq Y_i \text{ for some } 1 \leq i \leq n\right] \quad \text{(coupling inequality)} \\
= \quad & 1 - (1 - \mathbb{P}\left[X_i = Y_i \text{ for all } \ 1 \leq i \leq n\right]) \\
= \quad & \mathbb{P}\left[X_i = Y_i \text{ for all } 1 \leq i \leq n\right] \quad = \quad \mathbb{P}\left[T > n\right].
\end{aligned}
$$

□

2.2. Two independent coin flips

We now turn to the classical probabilistic paradigm of coin flips. Let X and Y represent two different sequences of i.i.d. coin flips, with probabilities of landing on H (heads) to be q and r respectively, where $0 \leq r \leq q \leq 1/2$. Suppose that we wish to maximise the length of the initial segment for which coin flips agree:

$$
S \quad = \quad \max\{t : X_i = Y_i \text{ for all } 1 \leq i \leq t\}.
$$

For concreteness, we will set $q = 0.5$ and $r = 0.4$ throughout this section; the generalization to other values is immediate.

2.2.1. Markovian Faithful Coupling for Independent Coin Flips

The "greedy" (Markovian and faithful) coupling carries out the best "one-step minorization" coupling possible, separately at each iteration. One-step minorization is essentially maximal coupling for single random variables. In this case, that means that for each flip, $\mathbb{P}\left[X = Y = H\right] = 0.4$, $\mathbb{P}\left[X = Y = T\right] = 0.5$, and $\mathbb{P}\left[X = H,\ Y = T\right] = 0.1$. This preserves the marginal distributions of X and Y, and yields $\mathbb{P}\left[X = Y\right] = 0.9$ at each step. Accordingly, the probability of agreement continuing for at least n steps is given by $\mathbb{P}\left[X_i = Y_i \text{ for } 1 \leq i \leq n\right] = (0.9)^n$.

2.2.2. A Look-ahead Coupling for Independent Coin Flips

Let a "look-ahead" coupling be a coupling which instead uses an n-step minorization couple on the entire sequence of n coin tosses, so that for each sequence s of n different Heads and Tails, it sets $\mathbb{P}[X = Y = s] = \min(\mathbb{P}[X = s], \mathbb{P}[Y = s])$. Consequently, if s has m Heads and $n - m$ Tails, then

$$\mathbb{P}[X = Y = s] \quad = \quad \min\{0.5^n,\, 0.4^m 0.6^{n-m}\}\,.$$

Elementary events for which X and Y disagree are assigned probabilities which preserve the marginal distributions of X and of Y. The simplest way to implement this is to use "independent residuals", but other choices are also possible.

This look-ahead coupling leads to a larger probability that $X = Y$. Indeed, even in the case $n = 2$, the probability of agreement over two coin flips under the greedy coupling is given by

$$\mathbb{P}[X = Y] \quad = \quad (0.9)^2 = 0.81\,.$$

The look-ahead coupling delivers a strictly larger probability of agreement over two coin flips:

$$\begin{aligned}\mathbb{P}[X = Y] \quad &= \quad \min(0.5^2, 0.4^2) + \min(0.5^2, 0.6^2) + 2\min(0.5^2, 0.4 \cdot 0.6)\\ &= \quad 0.4^2 + 0.5^2 + 2 \cdot 0.4 \cdot 0.6 = 0.16 + 0.25 + 0.48 = 0.89\,.\end{aligned}$$

When $n = 2$, the matrix of joint probabilities for X and Y under the look-ahead coupling is calculated to be:

$X\backslash Y$	HH	HT	TH	TT	SUM
HH	0.16	0	0	0.09	0.25
HT	0	0.24	0	0.01	0.25
TH	0	0	0.24	0.01	0.25
TT	0	0	0	0.25	0.25
SUM	0.16	0.24	0.24	0.36	1

Marginalizing this coupling on the initial coin flip ("projecting back" to the initial flip, with $n = 1$), we see that $\mathbb{P}[X_1 = Y_1 = H] = 0.16 + 0.24 = 0.4$, and $\mathbb{P}[X_1 = Y_1 = T] = 0.24 + 0.01 + 0.25 = 0.5$, and $\mathbb{P}[X_1 = H,\ Y_1 = T] = 0.09 + 0.01 = 0.1$. The projection to the initial flip yields the same agreement probability as would have been attained by maximizing the probability of staying together for just one flip ($n = 1$). That is, the $n = 2$ look-ahead coupling construction is *compatible* with the $n = 1$ construction.

Finally, it is worth noting that the $n = 2$ look-ahead coupling is certainly not faithful. For example, $\mathbb{P}[X_2 = H \,|\, X_1 = Y_1 = H] = 0.4 \neq 0.5$, and $\mathbb{P}[X_2 = H \,|\, X_1 = H,\ Y_1 = T] = 0.9 \neq 0.5$, *etc.*

2.3. A Look-ahead coupling for independent coin flips: the case $n = 3$

The matrix of joint probabilities for X and Y under the look-ahead coupling for $n = 3$ is more complicated, but can be calculated as:

$X\backslash Y$	HHH	HHT	HTH	HTT	THH	THT	TTH	TTT	SUM
HHH	0.064	0	0	0.0078	0	0.0078	0.0078	0.0375	0.125
HHT	0	0.096	0	0.0037	0	0.0037	0.0037	0.0178	0.125
HTH	0	0	0.096	0.0037	0	0.0037	0.0037	0.0178	0.125
HTT	0	0	0	0.125	0	0	0	0	0.125
THH	0	0	0	0.0037	0.096	0.0037	0.0037	0.0178	0.125
THT	0	0	0	0	0	0.125	0	0	0.125
TTH	0	0	0	0	0	0	0.125	0	0.125
TTT	0	0	0	0	0	0	0	0.125	0.125
SUM	0.064	0.096	0.096	0.144	0.096	0.144	0.144	0.216	1

With these probabilities, we compute that

$$\mathbb{P}[X = Y] = 0.064 + 3 \times 0.096 + 4 \times 0.125 = 0.852\,.$$

This is greater than the agreement probability of $0.9^3 = 0.729$ that would have be achieved via the greedy coupling. It is natural to wonder whether or not it is possible always to ensure that such a construction works not just for one fixed time but for all times. We further expound on this point in Sections 3 and 4, where discussion of a much more general context shows that that such constructions always exist.

2.3.1. Optimal Expectation

Until now, this section has focused on maximising $\mathbb{P}\left[X_i = Y_i \text{ for all } 1 \leq i \leq n\right]$, which is to say, maximizing $\mathbb{P}\left[S \geq n\right]$ with S being the time of first disagreement as above. We now consider the related question of maximizing the expected value $\mathbb{E}\left[S\right]$. Using the greedy coupling, clearly

$$\mathbb{E}\left[S\right] \quad = \quad \sum_{j=1}^{\infty} \mathbb{P}\left[S \geq j\right] \quad = \quad \sum_{j=1}^{\infty} 0.9^j \quad = \quad 0.9/(1-0.9) = 9\,.$$

If the different look-ahead couplings are chosen to be compatible, then this shows that $\mathbb{E}\left[S\right]$ is the sum for $r = 1, 2, \ldots$ of the probabilities that the j^{th} look-ahead coupling was successful. The work of Sections 3 and 4 shows that such a choice is always feasible, even for very general random processes indeed.

2.4. Adaptive MCMC

Un-coupling arguments play a natural role in the adaptive MCMC (Markov-chain Monte Carlo) literature, highlighted in particular by the work of Roberts and Rosenthal [25]. Roberts and Rosenthal [25] prove convergence of *adaptive* MCMC by comparing an adaptive process to a process which "*stops* adapting" at some point, and then by showing that the two processes have a high probability of remaining equal long enough such that the second process (and hence also the first process) converge to stationarity. The authors accomplish this by considering a sequence of adaptive Markov kernels $P_{\Gamma_1}, P_{\Gamma_2}, \ldots$ on a state space $\mathcal{X}$, where $\{P_\gamma : \gamma \in \mathcal{Y}\}$ are a collection of Markov kernels each having the same stationary probability distribution π, and the Γ_i are $\mathcal{Y}$-valued random variables which are "adaptive" (i.e., they depend on the previous Markov chain values but not on future values). Under appropriate assumptions, the authors prove that a Markov chain X which evolves *via* the adaptive Markov kernels will still converge to the specified stationary distribution π. The key step in the proof of the central result Roberts and Rosenthal [25, Theorem 5] is an un-coupling approach, highlighted below.

Roberts and Rosenthal [25, Theorem 5] assume that, for any $\varepsilon > 0$, there is a non-negative integer $N = N(\varepsilon)$ such that

$$\|P_\gamma^N(x, \cdot) - \pi(\cdot)\|_{\text{TV}} \quad \leq \quad \varepsilon$$

for all $x \in \mathcal{X}$ and $\gamma \in \mathcal{Y}$ (where $\|\cdot\|_{\mathrm{TV}}$ denotes total variation norm of a signed measure). Furthermore, there is a non-negative integer $n^* = n^*(\varepsilon)$ such that with probability at least $1-\varepsilon/N$,

$$\sup_{x\in\mathcal{X}} \|P_{\Gamma_{n+1}}(x,\cdot) - P_{\Gamma_n(x,\cdot)}\|_{\mathrm{TV}} \quad \leq \quad \varepsilon/N^2$$

for all $n \geq n^*$.

These assumptions are used to prove, for any $K \geq n^* + N$, the existence of a pair of processes X and X' defined for $K-N \leq n \leq K$, such that X evolves *via* the adaptive transition kernels P_{Γ_n}, while X' evolves *via* the fixed kernel $P' = P_{\Gamma_{K-N}}$. With probability at least $1-2\varepsilon$, the two processes remain equal for all times n with $K-N \leq n \leq K$. Hence, their un-coupling probability over this time interval is bounded above by 2ε. Consequently, conditional on X_{K-N} and Γ_{K-N}, the law of X_K lies within 2ε (measured in total variation distance) of the law of X'_K, which in turn lies within ε of the stationary distribution π. Hence, the law of X_K is within 3ε of π. Since this holds for any $\varepsilon > 0$ (for sufficiently large $K = K(\varepsilon)$), it follows that the law of X_K converges to π as $K \to \infty$. Accordingly the adaptive process X is indeed a "valid" Monte Carlo algorithm for approximately sampling from π; namely it converges asymptotically to π. The proof of a more general result (Roberts and Rosenthal [25, Theorem 13]), is quite similar, only requiring one additional ε.

3. *MEXIT* for discrete-time countable state-space

Having motivated the prominence of un-coupling arguments in key statistical and probabilistic settings, we now turn to an explicit construction of *MEXIT*. We begin by considering two discrete-time stochastic processes defined on the same countable discrete state-space, begun at the same initial state s_0. We extend the state-space by keeping track of the past trajectory of each stochastic process (its "genealogy"). The state of one of these stochastic processes at time n will thus be a sequence or genealogy $\mathbf{s} = (s_0, s_1, \ldots, s_n)$ of $n+1$ states, and these stochastic processes are then time-inhomogeneous Markov chains governed at time n by transition probability kernels $p(\mathbf{s}, b)$ and $q(\mathbf{s}, b)$, respectively.. Let $\mathbf{s}\cdot a$ denote the sequence or genealogy $\mathbf{s} = (s_0, s_1, \ldots, s_n, a)$ of $n+2$ states, corresponding to the chain moving to state a at time $n+1$. Note that if the original processes were originally Markov chains then this notation is equivalent to working with path probabilities $p(\mathbf{s}) = p(s_0, s_1)p(s_1, s_2)\ldots p(s_{n-1}, s_n)$, $q(\mathbf{s}) = q(t_0, t_1)q(t_1, t_2)\ldots q(t_{n-1}, t_n)$, with $p(\mathbf{s}\cdot a) = p(\mathbf{s})p(s_n, a)$ *et cetera.*

We define a coupling between the two processes as a random process on the Cartesian product of the (extended) state-space with itself, whose marginal distributions are those of the individual processes.

Definition 2 (Coupling of two discrete-time stochastic processes)**.** A *coupling* of two discrete-time stochastic processes on a countable state space with genealogical probabilities $p(\mathbf{s})$ and $q(\mathbf{t})$ respectively, is a random process (*not* necessarily Markov) with state $(\mathbf{s}, \mathbf{t})$ at time n given by a pair of genealogies $\mathbf{s}$ and $\mathbf{t}$ each of length n, such that if the probability of seeing state $(\mathbf{s}, \mathbf{t})$ at time n is equal to $r(\mathbf{s}, \mathbf{t})$, then

$$\sum_{\mathbf{t}} r(\mathbf{s}, \mathbf{t}) \quad = \quad p(\mathbf{s}) \qquad \text{(row-marginals)}\,, \tag{1}$$

$$\sum_{\mathbf{s}} r(\mathbf{s}, \mathbf{t}) \quad = \quad q(\mathbf{t}) \qquad \text{(column-marginals)}\,. \tag{2}$$

Moreover, probabilities at consecutive times are related by

$$\sum_a \sum_b r(\mathbf{s} \cdot a, \mathbf{t} \cdot b) \quad = \quad r(\mathbf{s}, \mathbf{t}) \qquad \text{(inheritance)}\,. \tag{3}$$

Remark 3. A coupling of two non-genealogical Markov chains can be converted into the above form simply by keeping track of the genealogies.

Remark 4. We assume that both processes begin at the same fixed starting point s_0, so $p((s_0)) = q((s_0)) = 1$, and the processes initially have the same trajectory. *MEXIT* occurs when first the trajectories split apart and disagree: the tree-like nature of genealogical state-space means the genealogical processes will never recombine.

A *MEXIT* coupling is one which achieves the bound prescribed by the Aldous [1] coupling inequality (Lemma 3.6 therein), thus (stochastically) maximising the time at which the chains split apart.

Definition 5 (*MEXIT* coupling)**.** Suppose that the following equation holds for all genealogical states $\mathbf{s}$:

$$r(\mathbf{s}, \mathbf{s}) \quad = \quad p(\mathbf{s}) \wedge q(\mathbf{s})\,. \tag{4}$$

Then the coupling is a *maximal exit coupling* (*MEXIT* coupling).

We now prove that *MEXIT* couplings always exist.

Theorem 6. *Consider two discrete-time stochastic processes taking values in a given countable state-space and started at the same initial state s_0. A* MEXIT *coupling can always be constructed such that the joint probability $r(\cdot,\cdot)$ satisfies the properties* (1)–(4).

Proof. We claim a MEXIT coupling is given by the following recursive definition

$$r(\mathbf{s}\cdot a, \mathbf{t}\cdot b) = \begin{cases} p(\mathbf{s}\cdot a) \wedge q(\mathbf{s}\cdot a) & \text{if } \mathbf{t}=\mathbf{s}, a=b, \\ \pi_1(\mathbf{s}\cdot b)\pi_2(\mathbf{s}\cdot a)\sum_c d^-(\mathbf{s}\cdot c) & \text{if } \mathbf{t}=\mathbf{s}, p(\mathbf{s}) \le q(\mathbf{s}), \\ \pi_1(\mathbf{s}\cdot b)\pi_2(\mathbf{s}\cdot a)\sum_c d^+(\mathbf{s}\cdot c) & \text{if } \mathbf{t}=\mathbf{s}, p(\mathbf{s}) > q(\mathbf{s}), \\ \pi_1(\mathbf{t}\cdot b)\pi_2(\mathbf{s}\cdot a) r(\mathbf{s},\mathbf{t}) & \text{if } \mathbf{t}\neq\mathbf{s}, \end{cases}$$

where

$$\begin{gathered} d^+(\mathbf{s}) = (q(\mathbf{s}) - p(\mathbf{s})) \vee 0, \quad d^-(\mathbf{s}) = (p(\mathbf{s}) - q(\mathbf{s})) \vee 0 \\ \pi_1(\mathbf{t}\cdot b) = \frac{d^+(\mathbf{t}\cdot b)}{\sum_c d^+(\mathbf{t}\cdot c)}, \quad \pi_2(\mathbf{s}\cdot a) = \frac{d^-(\mathbf{s}\cdot a)}{\sum_c d^-(\mathbf{s}\cdot c)}. \end{gathered}$$

We set π_1 (or π_2) to zero if the denominator appearing in the definition is zero. The initial joint probability is given by $r(s_0, s_0) = 1$, which clearly satisfies (1)–(4).

Now we verify by induction this construction actually satisfies (1)–(4) at each time n. First, the MEXIT equation (4) holds by construction. Second, if $\mathbf{s}\neq\mathbf{t}$, we immediately have

$$\sum_a \sum_b r(\mathbf{s}\cdot a, \mathbf{t}\cdot b) = r(\mathbf{s},\mathbf{t})$$

since $\sum_c \pi_1(\mathbf{t}\cdot c) = \sum_c \pi_2(\mathbf{s}\cdot c) = 1$. Observe that

$$\sum_c d^-(\mathbf{s}\cdot c) + \sum_c (p(\mathbf{s}\cdot c) \wedge q(\mathbf{s}\cdot c)) = \sum_c p(\mathbf{s}\cdot c) = p(\mathbf{s}),$$

and $d^+(\mathbf{s}\cdot a)d^-(\mathbf{s}\cdot a) = 0$. Hence if $p(\mathbf{s}) \le q(\mathbf{s})$,

$$\begin{aligned} \sum_a \sum_b r(\mathbf{s}\cdot a, \mathbf{s}\cdot b) &= \sum_c (p(\mathbf{s}\cdot c) \wedge q(\mathbf{s}\cdot c)) + \left(\sum_c d^-(\mathbf{s}\cdot c)\right) \sum_b \sum_{a\neq b} \pi_1(\mathbf{s}\cdot b)\pi_2(\mathbf{s}\cdot a) \\ &= \sum_c (p(\mathbf{s}\cdot c) \wedge q(\mathbf{s}\cdot c)) + \left(\sum_c d^-(\mathbf{s}\cdot c)\right) \frac{\sum_a \sum_b d^+(\mathbf{s}\cdot a) d^-(\mathbf{s}\cdot b)}{\left(\sum_c d^+(\mathbf{s}\cdot c)\right)\left(\sum_c d^-(\mathbf{s}\cdot c)\right)} \\ &= \sum_c (p(\mathbf{s}\cdot c) \wedge q(\mathbf{s}\cdot c)) + \sum_c d^-(\mathbf{s}\cdot c) = p(\mathbf{s}). \end{aligned}$$

Similarly, if $p(\mathbf{s}) > q(\mathbf{s})$,

$$\sum_a \sum_b r(\mathbf{s}\cdot a, \mathbf{s}\cdot b) = q(\mathbf{s}).$$

Thus we conclude the inheritance property (3) holds. Intuitively, given $r(\mathbf{s}, \mathbf{t})$ at time n, we can proceed to time $n+1$ by first filling in the diagonals according to (4); then for each big cell $(\mathbf{s}, \mathbf{t})$, the sum of $r(\mathbf{s}\cdot a, \mathbf{t}\cdot b)$ must be equal to $r(\mathbf{s}, \mathbf{t})$ by (3) and we fill in all the remaining cells proportionally by π_1 and π_2.

Now it remains to check the row/column marginal conditions. We shall only check that the row marginal condition holds. If $p(\mathbf{s}) \leq q(\mathbf{s})$, by the induction assumption, we have $r(\mathbf{s}, \mathbf{s}) = p(\mathbf{s})$ and $r(\mathbf{s}, \mathbf{t}) =$ for any $\mathbf{t} \neq \mathbf{s}$. Thus,

$$\begin{aligned}
\sum_{\mathbf{t}} \sum_b r(\mathbf{s}\cdot a, \mathbf{t}\cdot b) &= \sum_b r(\mathbf{s}\cdot a, \mathbf{s}\cdot b) \\
&= (p(\mathbf{s}\cdot a) \wedge q(\mathbf{s}\cdot a)) + \pi_2(\mathbf{s}\cdot a) \sum_c d^-(\mathbf{s}\cdot c) \sum_b \pi_1(\mathbf{s}\cdot b) \\
&= (p(\mathbf{s}\cdot a) \wedge q(\mathbf{s}\cdot a)) + d^-(\mathbf{s}\cdot a) = p(\mathbf{s}\cdot a).
\end{aligned}$$

If $p(\mathbf{s}) > q(\mathbf{s})$, observe that $p(\mathbf{s}) - q(\mathbf{s}) + d^+(\mathbf{s}\cdot c) = d^-(\mathbf{s}\cdot c)$ and thus

$$\begin{aligned}
\sum_{\mathbf{t}} \sum_b r(\mathbf{s}\cdot a, \mathbf{t}\cdot b) &= \sum_{\mathbf{t}\neq\mathbf{s}} \sum_b \pi_1(\mathbf{t}\cdot b)\pi_2(\mathbf{s}\cdot a) r(\mathbf{s}, \mathbf{t}) + \sum_b r(\mathbf{s}\cdot a, \mathbf{s}\cdot b) \\
&= \pi_2(\mathbf{s}\cdot a)(p(\mathbf{s}) - q(\mathbf{s})) + (p(\mathbf{s}\cdot a) \wedge q(\mathbf{s}\cdot a)) + \pi_2(\mathbf{s}\cdot a) \sum_c d^+(\mathbf{s}\cdot c) \\
&= d^-(\mathbf{s}\cdot a) + (p(\mathbf{s}\cdot a) \wedge q(\mathbf{s}\cdot a)) = p(\mathbf{s}\cdot a).
\end{aligned}$$

By symmetry, the column marginal condition holds. □

Remark 7. Note that the above theorem continues to hold if the common initial state s_0 is itself chosen randomly from some initial probability distribution.

Remark 8. MEXIT coupling is not unique in general. We can (over-)parametrize all possible *MEXIT* couplings by replacing the assignations π_1 and π_2 using copulae (Nelsen [18]) to parametrize the dependence between changes in the p-chain and the q-chain.

Recall the coin flip example. The table for $n = 3$ given in Section 2.3 does not satisfy the inheritance principle. Using the construction provided in the proof above, one *MEXIT* coupling is given by

$X\backslash Y$	HHH	HHT	HTH	HTT	THH	THT	TTH	TTT	SUM
HHH	0.064	0	0	0	0	0	0.0105	0.0505	0.125
HHT	0	0.096	0	0	0	0	0.0050	0.0240	0.125
HTH	0	0	0.096	0.019	0	0	0.0017	0.0083	0.125
HTT	0	0	0	0.125	0	0	0	0	0.125
THH	0	0	0	0	0.096	0.019	0.0017	0.0083	0.125
THT	0	0	0	0	0	0.125	0	0	0.125
TTH	0	0	0	0	0	0	0.125	0	0.125
TTT	0	0	0	0	0	0	0	0.125	0.125
SUM	0.064	0.096	0.096	0.144	0.096	0.144	0.144	0.216	1

It is easy to see that *MEXIT* is not unique. Assume all the cells are fixed except the upper-right four cells, which can be seen as a 2×2 table. Then this 2×2 table only need satisfy three constraints: the sum must be 0.9, the sum of the first row must be 0.061, and the sum of the first column must be 0.0155. Hence there is still one degree of freedom.

Having proven the existence of *MEXIT* couplings, we now provide calculations of *MEXIT* rate bounds (Subsection 3.1) and gain further insight into *MEXIT* by considering its connection with the Radon-Nikodym derivative (Subsection 3.2). We finish Section 3 on an applied note with a discussion of *MEXIT* times for MCMC algorithms (Subsection 3.3).

3.1. *MEXIT* rate bound

We now consider *MEXIT* rate bounds.

Proposition 9. *Consider the context of Theorem 6. Suppose we know that there is some $\delta > 0$ such that either:*

(a) for all $\mathbf{s}$ *and* a*,*

$$\frac{p(\mathbf{s}\cdot a)/p(\mathbf{s})}{q(\mathbf{s}\cdot a)/q(\mathbf{s})} \quad \geq \quad 1-\delta$$

or

(b) for all $\mathbf{s}$ *and* a*,*

$$\frac{q(\mathbf{s}\cdot a)/q(\mathbf{s})}{p(\mathbf{s}\cdot a)/p(\mathbf{s})} \quad \geq \quad 1-\delta.$$

Then

$$\mathbb{P}[\text{MEXIT } \textit{at time } n+1 \mid \textit{no } \text{MEXIT } \textit{by time } n] \quad \leq \quad \delta.$$

Proof. Assume (a) (then (b) follows by symmetry). We obtain

$$\mathbb{P}[\text{no } \textit{MEXIT} \text{ by time } n+1 \,|\, \text{no } \textit{MEXIT} \text{ by time } n]$$

$$\begin{aligned}
&= \frac{\sum_{\mathbf{s},a}[p(\mathbf{s}\cdot a)\wedge q(\mathbf{s}\cdot a)]}{\sum_{\mathbf{s}}[p(\mathbf{s})\wedge q(\mathbf{s})]} \\
&\geq \frac{\sum_{\mathbf{s},a}[(1-\delta)q(\mathbf{s}\cdot a)\frac{p(\mathbf{s})}{q(\mathbf{s})}\wedge q(\mathbf{s}\cdot a)]}{\sum_{\mathbf{s}}[p(\mathbf{s})\wedge q(\mathbf{s})]} \\
&= \frac{\sum_{\mathbf{s},a}\frac{q(\mathbf{s}\cdot a)}{q(\mathbf{s})}[(1-\delta)p(\mathbf{s})\wedge q(\mathbf{s})]}{\sum_{\mathbf{s}}[p(\mathbf{s})\wedge q(\mathbf{s})]} \\
&= \frac{\sum_{\mathbf{s}}[(1-\delta)p(\mathbf{s})\wedge q(\mathbf{s})]}{\sum_{\mathbf{s}}[p(\mathbf{s})\wedge q(\mathbf{s})]} \\
&\geq 1-\delta\,.
\end{aligned}$$

□

The above is the discrete state-space version of a bound contained in Völlering [32]. It should be noted that this bound applies equally well to faithful couplings, which typically degenerate in continuous time (see Theorem 28 in the present work for an example of this in the context of suitably regular diffusions.) Two corollaries of Proposition 9 follow immediately:

Corollary 10. *Under the conditions of Proposition 9,* $\mathbb{P}[\textit{no } \text{MEXIT } \textit{by time } n] \geq (1-\delta)^n$.

Corollary 11. *Under the conditions of Proposition 9,* $\mathbb{E}\,[\text{MEXIT } \textit{time}] \geq (1/\delta)$.

3.2. A Radon-Nikodym perspective on *MEXIT*

In this section, we explore a simple and natural connection of *MEXIT* to the value of the Radon-Nikodym derivative of q with respect to p.

In our discussion, it will suffice to consider *MEXIT* when the historical probability of the current path under both p and q are close to being equal, rare big jumps excepting. It follows from our *MEXIT* construction that the probability of *not* "MEXITing" by time n is equal to $\sum_{\mathbf{s}}(p(\mathbf{s}) \wedge q(\mathbf{s}))$, where the sum is over all length-n paths $\mathbf{s}$. Hence, conditional on having followed the path $\mathbf{s}$ up to time n and not "MEXITed," the conditional probability of *not* "MEXITing" at time $n+1$ is equal to

$$\frac{\sum_a (p(\mathbf{s} \cdot a) \wedge q(\mathbf{s} \cdot a))}{p(\mathbf{s}) \wedge q(\mathbf{s})}.$$

Thus, the probability of "MEXITing" at time $n+1$ is

$$1 - \frac{\sum_a (p(\mathbf{s} \cdot a) \wedge q(\mathbf{s} \cdot a))}{p(\mathbf{s}) \wedge q(\mathbf{s})} \quad = \quad \frac{(p(\mathbf{s}) \wedge q(\mathbf{s})) - \sum_a (p(\mathbf{s} \cdot a) \wedge q(\mathbf{s} \cdot a))}{p(\mathbf{s}) \wedge q(\mathbf{s})}.$$

In particular, if $p(\mathbf{s}) > q(\mathbf{s})$ and $p(\mathbf{s} \cdot a) > q(\mathbf{s} \cdot a)$ for all a, then the numerator is zero, so the probability of "MEXITing" is zero. That is, "MEXITing" can only happen when the relative ordering of $(p(\mathbf{s}), q(\mathbf{s}))$ and $(p(\mathbf{s} \cdot a), q(\mathbf{s} \cdot a))$ are different.

We now rephrase the above arguments in the language of Radon-Nikodym derivatives. Let $q(a|\mathbf{s}) = q(\mathbf{s} \cdot a)/q(\mathbf{s})$, and $R(\mathbf{s}) = p(\mathbf{s})/q(\mathbf{s})$. Then the non-*MEXIT* probability is

$$\frac{\sum_a (p(\mathbf{s} \cdot a) \wedge q(\mathbf{s} \cdot a))}{p(\mathbf{s}) \wedge q(\mathbf{s})} \quad = \quad \mathbb{E}_{q(a|\mathbf{s})}\left[\frac{R(\mathbf{s} \cdot a) \wedge 1}{R(\mathbf{s}) \wedge 1}\right] \quad = \quad \mathbb{E}_{p(a|\mathbf{s})}\left[\frac{R(\mathbf{s} \cdot a)^{-1} \wedge 1}{R(\mathbf{s})^{-1} \wedge 1}\right].$$

Note that $\mathbb{E}_{q(a|\mathbf{s})}[R(\mathbf{s} \cdot a)] = R(\mathbf{s})$. Thus, if we have either $R(\mathbf{s}) < 1$ and $R(\mathbf{s} \cdot a) < 1$ for all a, or $R(\mathbf{s}) > 1$ and $R(\mathbf{s} \cdot a) > 1$ for all a, then this non-*MEXIT* probability is one and thus the *MEXIT* probability is zero. That is, *MEXIT* can only occur when the Radon-Nikodym derivative R changes from more than 1 to less than 1 or vice-versa.

3.2.1. An example: *MEXIT* for simple random walks

To further elucidate the connection of *MEXIT* with the Radon-Nikodym derivative, we consider a concrete example: two simple random walks that both start at 0. Let "p" be simple random walk with up probability $\eta < 1/2$ and down probability $1 - \eta$. Similarly, let "q" be a simple random walk with up probability $1 - \eta$ and down probability η. The Radon-Nikodym derivative at time n can be computed as

$$R(\mathbf{s}) = \frac{p(\mathbf{s})}{q(\mathbf{s})} = \left(\frac{\eta}{1-\eta}\right)^{x_n + y_n - n},$$

where x_n and y_n denote the number of upward moves of chain "p" and "q" respectively. Hence $R(\mathbf{s}) = 0$ if and only if $x_n + y_n = n$. Before *MEXIT*, the two chains are coupled such that $x_n = y_n$, which further implies that *MEXIT* only occurs at 0, i.e. $x_n = y_n = n/2$. Indeed, the "pre-*MEXIT*" process (i.e., the joint process, conditional on *MEXIT* not having yet occurred) evolves with the following dynamics (for simplicity, we use P to denote the transition probability of either chain conditional on that *MEXIT* has not occurred.)

- For $k > 0$, $P(k, k+1) = \eta$, and $P(k, k-1) = 1 - \eta$.
- For $k < 0$, $P(k, k+1) = 1 - \eta$, and $P(k, k-1) = \eta$.
- $P(0, 1) = P(0, -1) = \eta$ with *MEXIT* probability $1 - 2\eta$ when we are at 0.

For $n = 2$, the joint distribution of the two chains is given by

$q \backslash p$	$++$	$+-$	$-+$	$--$	Sum
$++$	η^2	0	0	$1-2\eta$	$(1-\eta)^2$
$+-$	0	$\eta(1-\eta)$	0	0	$\eta(1-\eta)$
$-+$	0	0	$\eta(1-\eta)$	0	$\eta(1-\eta)$
$--$	0	0	0	η^2	η^2
Sum	η^2	$\eta(1-\eta)$	$\eta(1-\eta)$	$(1-\eta)^2$	1

Note that the chain P is defective at 0, but otherwise has a drift towards the *MEXIT* point 0. Consider the joint process, with death when MEXIT occurs. Let Q_t denote the number of times this process hits 0 up to and including time t. Then

$$\mathbb{P}[\textit{MEXIT} \text{ by time } t \mid Q_{t-1}] \quad = \quad 1 - (2\eta)^{Q_{t-1}}. \tag{5}$$

Hence,

$$P[\text{no } \textit{MEXIT} \text{ by time } t \mid Q_{t-1}] = (2\eta)^{Q_{t-1}}.$$

In particular, since $\eta < 1/2$, and the joint process is recurrent conditional on not yet "MEXITing", eventual MEXIT is certain.

3.3. An application: noisy MCMC

The purpose of this section is to provide an application of *MEXIT* for discrete-time countable state-spaces. We do so by comparing the *MEXIT* time τ of the *penalty method* MCMC algorithm with the usual Metropolis-Hastings algorithm.

In the usual Metropolis-Hastings algorithm, starting at a state X, we propose a new state Y, and then accept it with probability $1 \wedge A(X,Y)$, where $A(X,Y)$ is an appropriate acceptance probability formula in order to make the resulting Markov chain reversible with respect to the *target* density π. In *noisy MCMC* (specifically, the *penalty method* MCMC, see Ceperley and Dewing [6]; Nicholls *et al.* [19]; Medina-Aguayo *et al.* [17]; Alquier *et al.* [2]) which is similar to but different from the *pseudo-marginal MCMC* method of Andrieu and Roberts [3]), we accept with probability $\widehat{\alpha}(X,Y) := 1 \wedge (\hat{A}(X,Y))$, where $\hat{A}(X,Y)$ is an estimator of $A(X,Y)$ obtained from some auxiliary random experiment.

Noisy Metropolis-Hastings is popular in situations where the target density π is either not available or its pointwise evaluations are very computationally expensive. However replacing A by $\hat{A}$ interferes with detailed balance and therefore usually the invariant distribution of noisy Metropolis-Hastings (if it even exists) is biased (ie different from π). Quantifying the bias is therefore an important theoretical question. It is not our intention to give a full analysis of this here, as this is well-studied for example Medina-Aguayo *et al.* [17]. However a crucial component in the argument used in that paper is the construction of a coupling between a standard and a noisy Metropolis-Hastings chain in such a way that, with high probability, MEXIT occurs at a time after both chains have more or less converged to equilibrium. Here therefore we shall just focus on lower bounds for the MEXIT time.

For this example we shall assume that $W = \exp(N)$ where $N \sim \text{Normal}(-\sigma^2/2,\, \sigma^2)$ for some fixed $\sigma > 0$ (so that $\mathbb{E}[W] = \mathbb{E}[\exp(N)] = 1$), i.e. that $\widehat{\alpha}(X,Y) := 1 \wedge (A(X,Y)\exp(N))$. We now show that the *penalty method MCMC* produces a Metropolis-Hastings algorithm with sub-optimal acceptance probability.

Proposition 12. *The penalty method MCMC produces a Metropolis-Hastings algorithm with (sub-optimal) acceptance probability* $\widetilde{\alpha}(X,Y,\sigma) := \mathbb{E}[\widetilde{\alpha}(X,Y) \,|\, X,Y]$ *given by*

$$\widetilde{\alpha}(X,Y,\sigma) \quad = \quad \Phi\left[\frac{\log A(X,Y)}{\sigma} - \frac{\sigma}{2}\right] + A(X,Y)\,\Phi\left[-\frac{\sigma}{2} - \frac{\log A(X,Y)}{\sigma}\right].$$

Proof. We invoke Proposition 2.4 of Roberts *et al.* [23], which states that if $B \sim \text{Normal}(\mu, \sigma^2)$, then

$$\mathbb{E}\left[1 \wedge e^B\right] \quad = \quad \Phi\left(\frac{\mu}{\sigma}\right) + \exp(\mu + \sigma^2/2)\,\Phi\left[-\sigma - \frac{\mu}{\sigma}\right].$$

Note

$$\widetilde{\alpha}(X,Y,\sigma) \quad = \quad \mathbb{E}\left[\widehat{\alpha}(X,Y)\right] \quad = \quad \mathbb{E}\left[1 \wedge (A(X,Y)e^N)\right] \quad = \quad \mathbb{E}\left[1 \wedge e^{N(-\sigma^2/2+\log A(X,Y),\ \sigma^2)}\right].$$

After straightforward algebra, the right-hand side of the last equality simplifies to

$$\Phi\left[\frac{\log A(X,Y)}{\sigma} - \frac{\sigma}{2}\right] + A(X,Y)\,\Phi\left[-\frac{\sigma}{2} - \frac{\log A(X,Y)}{\sigma}\right].$$

$\square$

Proposition 13. $A(X,Y)\,\phi\left[-\frac{\sigma}{2} - \frac{\log A(X,Y)}{\sigma}\right] \quad = \quad \phi\left[\frac{\log A(X,Y)}{\sigma} - \frac{\sigma}{2}\right]$.

Proof. We calculate

$$\begin{aligned}
& A(X,Y)\,\phi\left[-\frac{\sigma}{2} - \frac{\log A(X,Y)}{\sigma}\right] \\
= \quad & \frac{1}{\sqrt{2\pi}} \exp\left(\log A(X,Y) - \frac{1}{2}\left(-\frac{\sigma}{2} - \left(\frac{\log A(X,Y)}{\sigma}\right)^2\right)\right) \\
= \quad & \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{\log A(X,Y)}{\sigma} - \frac{\sigma}{2}\right)^2\right) \\
= \quad & \phi\left(\frac{\log A(X,Y)}{\sigma} - \frac{\sigma}{2}\right).
\end{aligned}$$

$\square$

Proposition 14. *For any $a, s > 0$, we have that*

$$\frac{1}{a}\,\phi\left(\frac{\log a}{s} - \frac{s}{2}\right) \quad \leq \quad \frac{1}{\sqrt{2\pi}}. \tag{6}$$

Proof. This follows from noting

$$\begin{aligned}
& \frac{1}{a}\,\phi\left(\frac{\log a}{s} - \frac{s}{2}\right) \\
= \quad & \frac{1}{\sqrt{2\pi}} \exp\left(-\log a - \frac{1}{2}\left(\frac{\log a}{s} - \frac{s}{2}\right)^2\right) \\
= \quad & \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{\log a}{s} + \frac{s}{2}\right)^2\right) \leq \frac{1}{\sqrt{2\pi}}.
\end{aligned}$$

$\square$

Let $r(X)$ and $\widetilde{r}(X)$ be the probabilities of rejecting the proposal when starting at X for the original Metropolis-Hastings algorithm and the *penalty method* MCMC, respectively. We now proceed with Proposition 15.

Proposition 15. *For all X, Y in the state space, and $\sigma \geq 0$, the following seven statements hold*

(1) $\widetilde{\alpha}(X,Y) \leq \alpha(X,Y)$.

(2) $\widetilde{r}(X) \geq r(X)$.

(3) $\lim_{\sigma \searrow 0} \widetilde{\alpha}(X,Y,\sigma) = \alpha(X,Y)$.

(4) $\frac{d}{d\sigma}\widetilde{\alpha}(X,Y,\sigma) = -\phi\left[\frac{\log A(X,Y)}{\sigma} - \frac{\sigma}{2}\right]$.

(5) $0 \geq \frac{d}{d\sigma}\widetilde{\alpha}(X,Y,\sigma) \geq -1/\sqrt{2\pi}$.

(6) $\widetilde{\alpha}(X,Y,\sigma) \geq \alpha(X,Y) - \sigma/\sqrt{2\pi}$.

(7) $\frac{\widetilde{\alpha}(X,Y,\sigma)}{\alpha(X,Y)} \geq 1 - \sigma/\sqrt{2\pi}$.

Proof. For statement (1), apply Jensen's inequality. Note that

$$\begin{aligned}
\mathbb{E}\left[\widetilde{\alpha}(X,Y) \,|\, X,Y\right] = \mathbb{E}\left[1 \wedge (A(X,Y)e^N) \,|\, X,Y\right] \quad &\leq 1 \wedge \mathbb{E}\left[(A(X,Y)e^N)\right] \\
= 1 \wedge (A(X,Y)\,\mathbb{E}\left[e^N\right]) \qquad\qquad &= 1 \wedge A(X,Y) \qquad = \alpha(X,Y)\,.
\end{aligned}$$

Statement (2) follows immediately from statement (1) by taking the complements of the expectations of the $\alpha(X,Y)$ and $\widetilde{\alpha}(X,Y)$ with respect to Y.

For statement (3), note that if $A(X,Y) > 1$ then $\lim_{\sigma \searrow 0} \widetilde{\alpha}(X,Y,\sigma) = \Phi[+\infty] + A(X,Y)\,\Phi[-\infty] = 1$, while if $A(X,Y) < 1$ then $\lim_{\sigma \searrow 0} \widetilde{\alpha}(X,Y,\sigma) = \Phi[-\infty] + A(X,Y)\,\Phi[+\infty] = 0 + A(X,Y)\,1 = A(X,Y)$. Further, if $A(X,Y) = 1$ then $\lim_{\sigma \searrow 0} \widetilde{\alpha}(X,Y,\sigma) = \Phi[0] + A(X,Y)\,\Phi[0] = (1/2) + (1)(1/2) = 1$. Thus, in all cases, $\lim_{\sigma \searrow 0} \widetilde{\alpha}(X,Y,\sigma) = 1 \wedge A(X,Y) = \alpha(X,Y)$.

For statement (4), we use Proposition 13 to compute

$$\begin{aligned}
& \frac{d}{d\sigma}\widetilde{\alpha}(X,Y,\sigma) \\
= \;& \frac{d}{d\sigma}\left(\Phi\left[\frac{\log A(X,Y)}{\sigma} - \frac{\sigma}{2}\right] + A(X,Y)\,\Phi\left[-\frac{\sigma}{2} - \frac{\log A(X,Y)}{\sigma}\right]\right) \\
= \;& \phi\left[\frac{\log A(X,Y)}{\sigma} - \frac{\sigma}{2}\right]\,\left(-\frac{\log A(X,Y)}{\sigma^2} - \frac{1}{2}\right) + A(X,Y)\;\phi\left[-\frac{\sigma}{2} - \frac{\log A(X,Y)}{\sigma}\right] \\
= \;& -\frac{1}{2} + \frac{\log A(X,Y)}{\sigma^2} \;\;=\; -\phi\left[\frac{\log A(X,Y)}{\sigma} - \frac{\sigma}{2}\right].
\end{aligned}$$

Since $0 \leq \phi(\cdot) \leq \frac{1}{\sqrt{2\pi}}$, statement (5) follows immediately. Statement (6) then follows by integrating from 0 to σ. For statement (7), note that if $A(X,Y) \geq 1$ then $\alpha(X,Y) = 1$ and the result then follows from statement (6). If instead $A(X,Y) < 1$, then $\alpha(X,Y) = A(X,Y)$, and we may invoke Proposition 14 to obtain

$$\begin{aligned}
\frac{\widetilde{\alpha}(X,Y,\sigma)}{\alpha(X,Y)} &= 1 - \frac{\alpha(X,Y) - \widetilde{\alpha}(X,Y,\sigma)}{\alpha(X,Y)} \\
&= 1 - \int_{u=0}^{\sigma} \frac{1}{\alpha(X,Y)} \frac{d}{du} \widetilde{\alpha}(X,Y,u) du \\
&= 1 - \int_{u=0}^{\sigma} \frac{1}{A(X,Y)} \phi\left[\frac{\log A(X,Y)}{\sigma} - \frac{\sigma}{2}\right] du \\
&\geq 1 - \int_{u=0}^{\sigma} \frac{1}{\sqrt{2\pi}}\, du \quad = \quad 1 - \frac{\sigma}{\sqrt{2\pi}}\,.
\end{aligned}$$

This concludes the proof. □

Let P be the law of a Metropolis-Hastings algorithm, and $\widetilde{P}$ the law of a corresponding noisy MCMC. We now prove Proposition 16 below, whose Corollary 17 uses *MEXIT* to control the discrepancy between the Metropolis-Hastings algorithm and the noisy MCMC algorithm.

Proposition 16.

$$\frac{d\widetilde{P}^{t+1}(\mathbf{s}\cdot a)}{dP^{t+1}(\mathbf{s}\cdot a)} \quad \geq \quad \frac{d\widetilde{P}^{t}(\mathbf{s})}{dP^{t}(\mathbf{s})} \left(1 - \frac{\sigma}{\sqrt{2\pi}}\right).$$

Proof. Note first that $\frac{d\widetilde{P}^t(\mathbf{s})}{dP^t(\mathbf{s})} = \gamma_1\gamma_2\ldots\gamma_n$ where each γ_i equals either $\frac{\widetilde{\alpha}(X_{i-1},X_i)}{\alpha(X_{i-1},X_i)}$ if the move from X_{i-1} to X_i is accepted and otherwise $\frac{\widetilde{r}(X)}{r(X)}$ if the move is rejected. Statement (2) of Proposition 15 tells us that, if we reject,

$$\frac{d\widetilde{P}^{t+1}(\mathbf{s}\cdot a)}{dP^{t+1}(\mathbf{s}\cdot a)} \quad \geq \quad \frac{d\widetilde{P}^{t}(\mathbf{s})}{dP^{t}(\mathbf{s})} \quad \geq \quad \frac{d\widetilde{P}^{t}(\mathbf{s})}{dP^{t}(\mathbf{s})} \left(1 - \frac{\sigma}{\sqrt{2\pi}}\right).$$

However, if we accept, then by statement (7) in Proposition 15, $\frac{d\widetilde{P}^{t+1}(\mathbf{s}\cdot a)}{dP^{t+1}(\mathbf{s}\cdot a)} \geq \frac{d\widetilde{P}^{t}(\mathbf{s})}{dP^{t}(\mathbf{s})}(1 - \frac{\sigma}{\sqrt{2\pi}})$, as claimed. □

The following Corollary to Proposition 16 now follows immediately.

Corollary 17. $\frac{d\widetilde{P}^{t}(\mathbf{s})}{dP^{t}(\mathbf{s})} \geq \left(1 - \frac{\sigma}{\sqrt{2\pi}}\right)^t$.

Applying Proposition 16 to Proposition 9 in Subsection 3.1, with $\delta = \frac{\sigma}{\sqrt{2\pi}}$, the following Corollary follows immediately.

Corollary 18. *The* MEXIT *time* τ *of the above penalty method MCMC algorithm, compared to the regular Metropolis-Hastings algorithm, satisfies the following two inequalities:*

$$\mathbb{P}[\tau > n] \quad \geq \quad \left(1 - \frac{\sigma}{\sqrt{2\pi}}\right)^n$$

and

$$\mathbb{E}\left[\tau\right] \quad \geq \quad \sqrt{2\pi}/\sigma.$$

Of course, unless σ is small, MEXIT will likely occur substantially before Markov chain mixing, reflecting the fact that successful couplings usually have to bring chains together and not just stop them from separating. Therefore these results are usually not useful for explicitly estimating the bias of noisy Metropolis-Hastings. However they are particularly useful for demonstrating robustness results for both noisy and pseudo-marginal chains as in Medina-Aguayo *et al.* [17] and Andrieu and Roberts [3].

4. *MEXIT* for general random processes

The methods and results of Section 3 generalize to the case when the two processes are general time-inhomogeneous random processes in discrete time with countable state-space: such processes, with state augmented to include genealogy, become Markov chains. In fact the methods and results extend to still more general processes: in this section we deal with the case of random processes for which the state-space is a general Polish space (a σ-algebra arising from a complete separable metric space).

4.1. Case of one time-step

To establish notation, we first review the simplest case of just one time-step. We require the state-space to be Polish (we note that in principle one might be able to generalize a little beyond this, but the prospective rewards of such a generalization seem to be not very substantial). In the case of Polish space, the diagonal set $\Delta = \{(x, x) : x \in E\} \subset E \times E$ belongs to the product σ-algebra $\mathcal{E} * \mathcal{E}$ (counterexamples for some more general spaces are provided in Stoyanov [29, Subsection 1.6]; in principle one could seek to exploit the fact that Δ is in general analytic with respect to $\mathcal{E} * \mathcal{E}$, but some kind of assumption about the state-space would still be required to take care of further complications).

Consider two E-valued random variables X_1^+ and X_1^-, measurable with respect to $\mathcal{E}$ on E, with distributions $\mathcal{L}\left(X_1^+\right) = \mu_1^+$ and $\mathcal{L}\left(X_1^-\right) = \mu_1^-$ on $(E, \mathcal{E})$. We recall that the *meet measure* $\hat{\mu}_1 = \mu_1^+ \wedge \mu_1^-$ of the probability measures μ^+ and μ^- in the lattice of non-negative measures on $(E, \mathcal{E}_1)$ can be described explicitly using the Hahn-Jordan decomposition (Halmos [12, §28]) as

$$\mu_1^+ - \mu_1^- \quad = \quad \nu_1^+ - \nu_1^- \tag{7}$$

for unique non-negative measures ν_1^+ and ν_1^- of disjoint support. The condition of disjoint support implies that

$$\hat{\mu}_1 \quad = \quad \mu_1^+ - \nu_1^+ \quad = \quad \mu_1^- - \nu_1^- \tag{8}$$

is the maximal non-negative measure $\widetilde{\mu}$ such that

$$\widetilde{\mu}(D) \quad \leq \quad \min\{\mu_1^+(D), \mu_1^-(D)\} \qquad \text{for all } D \in \mathcal{E}\,.$$

Lemma 19. *Consider two random variables X_1^+ and X_1^- taking values in the same measurable space $(E, \mathcal{E})$ which is required to be Polish. The simplest* MEXIT *problem is solved by maximal coupling of the two marginal probability measures $\mu_1^+ = \mathcal{L}\left(X_1^+\right)$ and $\mu_1^- = \mathcal{L}\left(X_1^-\right)$ using a joint probability measure m_1 on the product measure space $(E \times E, \mathcal{E} * \mathcal{E})$ such that*

1. *m_1 has marginal distributions μ_1^+ and μ_1^- on the two coordinates,*
2. *$m_1 \geq \imath_{\Delta *}\hat{\mu}_1$, where the non-negative measure $\hat{\mu}_1 = \mu_1^+ \wedge \mu_1^-$ is the meet measure for μ_1^+ and μ_1^-, and $\imath_{\Delta *}$ is the push-forward map corresponding to the $(\mathcal{E} : \mathcal{E} * \mathcal{E})$-measurable "diagonal injection" $\imath_\Delta : E \to E \times E$ given by $\imath_\Delta(x) = (x, x)$.*

Proof. One possible explicit construction for m_1 is

$$m_1 \quad = \quad \imath_{\Delta *}\hat{\mu}_1 + \frac{1}{\nu_1^+(E)}\nu_1^+ \otimes \nu_1^-\,, \tag{9}$$

where $\nu_1^\pm$ are defined by the Hahn-Jordan decomposition in (7) and $\nu_1^+ \otimes \nu_1^-$ is the product measure on $(E \times E, \mathcal{E} * \mathcal{E})$. It follows directly from (7) that $\nu_1^+(E) = \nu_1^-(E)$. Maximality of the coupling (which is to say, maximality of $m_1(\Delta) = \hat{\mu}_1(E)$ compared to all other probability measures with these marginals) follows from maximality of the meet measure $\hat{\mu}$. This completes the proof. $\square$

Given this construction, we can realize X_1^+ and X_1^- as the coordinate maps for $E \times E$: the probability statements

$$\mathbb{P}\left[X_1^+ \in D\;;\; X_1^+ = X_1^-\right] \quad = \quad \hat{\mu}_1(D) \qquad \text{for all } D \in \mathcal{E} \tag{10}$$

hold for any maximal coupling of X_1^+ and X_1^-.

It is convenient at this point to note a quick way to recognize when a given coupling is maximal.

Lemma 20 (Recognition Lemma for Maximal Coupling)**.** *Suppose the measurable space $(E, \mathcal{E})$ is Polish. Given a coupling probability measure m^* for $(E, \mathcal{E})$-valued random variables X_1^+ and X_1^- (with distributions $\mathcal{L}\left(X_1^+\right) = \mu_1^+$ and $\mathcal{L}\left(X_1^-\right) = \mu_1^-$), this coupling is maximal if the two non-negative measures*

$$\nu_1^{\pm,*} \;:\; D \quad \mapsto \quad m^*[X_1^{\pm} \in D\;;\; X_1^+ \neq X_1^-] \tag{11}$$

(defined for $D \in \mathcal{E}$) are supported by two disjoint $\mathcal{E}$-measurable sets. Moreover in this case the meet measure for the two probability distributions $\mathcal{L}\left(X_1^+\right)$ and $\mathcal{L}\left(X_1^-\right)$ is given by

$$\hat{\mu}_1(D) \quad = \quad m^*[X_1^+ \in D\;;\; X_1^+ = X_1^-] \qquad \textit{for all } D \in \mathcal{E}\,. \tag{12}$$

Proof. This follows immediately from the uniqueness of the non-negative measures $\nu_1^{\pm}$ of disjoint support appearing in the Hahn-Jordan decomposition, since a sample-wise cancellation of events shows that

$$\mu_1^+ - \mu_1^- \quad = \quad \mathcal{L}\left(X_1^+\right) - \mathcal{L}\left(X_1^-\right) \quad = \quad \nu_1^{+,*} - \nu_1^{-,*}\,.$$

□

4.2. Case of n time-steps

The next step is to consider the extent to which Theorem 6 generalizes to the case of discrete-time random processes taking values in general Polish state-spaces. We first note that the generalization beyond Polish spaces cannot always hold. Based on the work of Rigo and Thorisson [22], and dating back to Doob [7, p.624], Halmos [12, p.210], and Billingsley [4, Chapter 33], consider the following counterexample.

Consider the interval $\Omega = [0, 1]$ equipped with Lebesgue measure. There exists a set $M \subset \Omega$ with outer measure 1 and inner measure 0, e.g. a Vitali set with outer measure 1. Let $\mathcal{B}$ be the Borel σ-algebra on Ω and consider the σ-algebra $\sigma(\mathcal{B}, M)$. It can be shown that any set $A \in \sigma(\mathcal{B}, M)$

can be written as

$$A = (M \cap B_1) \cup (M^c \cap B_2), \quad B_1, B_2 \in \mathcal{B}.$$

The representation is not unique. However, using the identity $\mathrm{Leb}^*(M) = \mathrm{Leb}^*(M \cap B_1) + \mathrm{Leb}^*(M \cap B_1^c)$ (since B_1 is Lebesgue measurable), we can show $\mathrm{Leb}^*(M \cap B_1) = \mathrm{Leb}(B_1)$ where Leb^* is the Lebesgue outer measure. Similarly, $\mathrm{Leb}^*(M^c \cap B_2) = \mathrm{Leb}(B_2)$. Hence if there is another representation $A = (M \cap B_3) \cup (M^c \cap B_4)$ where B_3 and B_4 are Borel, we must have $\mathrm{Leb}(B_1) = \mathrm{Leb}(B_3)$ and $\mathrm{Leb}(B_2) = \mathrm{Leb}(B_4)$. Now we can define the probability measures $m^\pm$ on $\sigma(\mathcal{B}, M)$ by

$$m^+(A) = \mathrm{Leb}(B_1), \quad m^-(A) = \mathrm{Leb}(B_2).$$

It is straightforward to verify that they are probability measures. Note that for any Borel set B, we have $m^+(B) = m^-(B) = \mathrm{Leb}(B)$. Set $\mathcal{E}_1 = \mathcal{B}$ and $\mathcal{E}_2 = \sigma(\mathcal{B}, M)$. Consider two random sequences (X_1^+, X_2^+) and (X_1^-, X_2^-). Let $X_2^\pm(\omega) = \omega$ be random variables defined on $(\Omega, \mathcal{E}_2, m^\pm)$. Let $X_1^\pm$ be defined on $(\Omega, \mathcal{E}_1)$ and set $X_1^\pm = X_2^\pm$ (this is allowed because the function $X(\omega) = \omega$ is Borel measurable). Since for any $B \in \mathcal{B}$,

$$\mathbb{P}\left[X_1^+ \in B\right] = \mathbb{P}\left[X_2^+ \in B\right] = m^+(B) = \mathrm{Leb}(B),$$

$X_1^\pm$ have the same law (the Lebesgue measure) and thus any realization of MEXIT would have to have $\mathbb{P}\left[X_1^+ = X_1^-\right] = 1$, which further implies $\mathbb{P}\left[X_2^+ = X_2^-\right] = 1$. On the other hand, since $m^+(M) = 1$ and $m^-(M) = 0$, we have $||m^+ - m^-||_{\mathrm{TV}} = 1$ w.r.t $\mathcal{E}_2$. So for any coupling of $X_2^\pm$, denoted by $(\Omega^2, \overline{\mathcal{E}^2}, \boldsymbol{\mu})$, where $\overline{\mathcal{E}^2}$ denotes the completion of $\mathcal{E}_2 \times \mathcal{E}_2$ w.r.t. $\boldsymbol{\mu}$, we must have $\boldsymbol{\mu}(\{(\omega, \omega) : \omega \in \Omega\}) = 0$. This gives a contradiction.

However the existence of MEXIT follows easily in the case of Polish spaces, as also noted by Völlering [32]. Here follows a proof by induction.

Theorem 21. *Consider two discrete-time random processes X^+ and X^-, begun at the same fixed initial point, taking values in a measurable state-space $(E, \mathcal{E})$ which is Polish, and run up to a finite time n. Maximal* MEXIT *couplings exist.*

Proof. The case $n = 1$ follows directly from the general state-space arguments of Lemma 19. The countable product of Polish spaces is again Polish, so an inductive argument completes the proof if we can establish the following.

Suppose $X^{\pm}$ are two random variables taking values in a measurable space $(E, \mathcal{E}_2)$ which is Polish, with laws $\mu_2^{\pm}$. Suppose $\mathcal{E}_1 \subseteq \mathcal{E}_2$ is a sub-σ-algebra such that $(E, \mathcal{E}_1)$ is also Polish, and let $\mu_1^{\pm}$ be the laws of $X^{\pm}$ viewed as random variables taking values in the Polish space $(E, \mathcal{E}_1)$. Suppose m_1 is a maximal coupling with marginals $\mu_1^{\pm}$ on $(E \times E, \mathcal{E}_1 * \mathcal{E}_1)$. The claim is that there then exists a maximal coupling m_2 with marginals $\mu_2^{\pm}$ on $(E \times E, \mathcal{E}_2 * \mathcal{E}_2)$ which equals m_1 when restricted to $\mathcal{E}_1 * \mathcal{E}_1$.

To see this, first note from Lemma 19 that $m_1|_\Delta = \imath_{\Delta *}\hat{\mu}_1$, where $\hat{\mu}_1$ is the sub-probability measure given by $\hat{\mu}_1 = \mu_1^+ \wedge \mu_1^-$. Moreover, if $\hat{\mu}_2$ is the sub-probability measure given by $\hat{\mu}_2 = \mu_2^+ \wedge \mu_2^-$, then we can use the infimum characterization following (8) to show that $\hat{\mu}_2$ satisfies $\hat{\mu}_2(A) \leq \hat{\mu}_1(A)$ for all $A \in \mathcal{E}_1$. Write $(1-\pi_1)\,\mathrm{d}\hat{\mu}_1 = \mathrm{d}(\hat{\mu}_2|_{\mathcal{E}_1})$ to define the $\mathcal{E}_1$-measurable random variable π_1 (with $0 \leq \pi_1 \leq 1$) as *the conditional probability of MEXIT immediately after time* 1. Because $(1-\pi_1)\,\mathrm{d}\hat{\mu}_1$ and $\mathrm{d}\hat{\mu}_2$ agree on $\mathcal{E}_1$, and because we are working with Polish spaces, we can construct a regular conditional probability kernel $\hat{k}_{12}(x, B)$ (a probability measure on $\mathcal{E}_2$ for each fixed x, and $\mathcal{E}_1$-measurable in x) such that

$$\mathrm{d}\hat{\mu}_2 \quad = \quad (1-\pi_1)\hat{k}_{12} * \mathrm{d}\hat{\mu}_1\,. \tag{13}$$

Similarly we can construct regular conditional probability kernels $k_{12}^{\pm}(x, B)$ such that

$$\mathrm{d}\mu_2^{\pm} \quad = \quad k_{12}^{\pm} * \mathrm{d}\mu_1^{\pm}\,. \tag{14}$$

Now $(1-\pi_1)\imath_{\Delta *}(\hat{k}_{12} * \mathrm{d}\hat{\mu}_1) = \imath_{\Delta *}\,\mathrm{d}\hat{\mu}_2$ defines a sub-probability measure on $(E \times E, \mathcal{E}_2 * \mathcal{E}_2)$ with marginals equal to each other and given by $\hat{\mu}_2$ (as a consequence of (13)). The proof of the claim will be completed if we can establish the existence of a sub-probability measure Γ_2 on $(E \times E, \mathcal{E}_2 * \mathcal{E}_2)$ with marginals defined by $\mu_2^{\pm} - \hat{\mu}_2$, and agreeing on $\mathcal{E}_1 * \mathcal{E}_1$ with the measure defined by $\mathrm{d}m_1 - (1-\pi_1)\imath_{\Delta *}\,\mathrm{d}\hat{\mu}_1$. Consider

$$\mathrm{d}\Gamma_2 \quad = \quad (k_{12}^+ \otimes k_{12}^-) * (\,\mathrm{d}m_1 - (1-\pi_1)\imath_{\Delta *}\,\mathrm{d}\hat{\mu}_1)\,,$$

where $(k_{12}^+ \otimes k_{12}^-)((x^+, x^-), B^+ \times B^-) = k_{12}^+(x^+, B^+) \times k_{12}^-(x^-, B^-)$ and we use the theory of product measure to extend to a kernel of product measures $k_{12}^+(x^+, \cdot) \otimes k_{12}^-(x^-, \cdot)$. Exactly because $(k_{12}^+ \otimes k_{12}^-)$ is itself a regular conditional probability kernel, it follows that Γ_2 agrees on $\mathcal{E}_1 * \mathcal{E}_1$ with the measure defined by $\mathrm{d}m_1 - (1-\pi_1)\,\mathrm{d}\hat{\mu}_1$. On the other hand, because Γ_2 is built from appropriate product regular conditional probabilities, Γ_2 has marginals defined by $k_{12}^{\pm}\,\mathrm{d}\mu_1^{\pm} - (1-\pi_1)\,\mathrm{d}\hat{\mu} = \mathrm{d}\mu_2^{\pm} - \mathrm{d}\hat{\mu}_2$ as required.

In summary, the required maximal coupling at the level of $\mathcal{E}_2 * \mathcal{E}_2$ is defined by

$$\imath_{\Delta_*}\,\mathrm{d}\hat{\mu}_2 + \mathrm{d}\Gamma_2 \quad = \quad (1-\pi_1)\imath_{\Delta_*}(\hat{k}_{12} * \mathrm{d}\hat{\mu}_1) + (k_{12}^{+} \otimes k_{12}^{-}) * (\,\mathrm{d}m_1 - (1-\pi_1)\imath_{\Delta_*}\,\mathrm{d}\hat{\mu}_1)\,. \tag{15}$$

□

Remark 22. As in the $n=1$ case of Lemma 19, we can generate a whole class of maximal couplings by using measurable selections from Fréchet classes to replace the product regular conditional probability kernel $(k_{12}^{+} \otimes k_{12}^{-}) * (\,\mathrm{d}m_1 - (1-\pi_1)\imath_{\Delta_*}\,\mathrm{d}\hat{\mu}_1)$. Equally, as in the $n=1$ case of Lemma 19, this clearly does not exhaust all the possibilities.

4.3. Unbounded and/or continuous time

MEXIT for all times (with no upper bound on time) follows easily so long as the Kolmogorov Extension Theorem (Doob [8, §V.6]) can be applied. This is certainly the case if the state-space is Polish; we state this formally as a corollary to Theorem 21 of the previous section. (For an example of what can go wrong in a more general measure-theoretic context for the Kolmogorov Extension Theorem, see Stoyanov [29, §2.3].)

Corollary 23. *Consider two discrete-time random processes X^+ and X^-, begun at the same fixed initial point, taking values in a measurable state-space $(E, \mathcal{E})$ which is Polish.* MEXIT *couplings exist through all time.*

Under the requirement of Polish state-space, it is also straightforward to establish a continuous-time version of the *MEXIT* result for càdlàg processes. The result requires this preliminary elementary properties about joint laws with given marginals.

Lemma 24. *Suppose that $\{X_i^+\}$ and $\{X_i^-\}$ are two collections of random variables on the probability space $(\Omega, \mathcal{F}, \mathbb{P})$ taking values on a metric space (E, d). Suppose that $\{\mathcal{L}\left(X_i^+\right)\}$ and $\{\mathcal{L}\left(X_i^-\right)\}$ are both tight. Then $\{\mathcal{L}\left(X_i^+, X_i^-\right)\}$ is tight on $(E \times E, \tilde{d})$ where $\tilde{d}$ denotes the Euclidean product measure $d \times d$.*

Proof. For any $\epsilon > 0$, we can find compact sets S^+, S^- such that $\mathbb{P}(X_i^+ \in S^+) > 1 - \epsilon/2$ and $\mathbb{P}(X_i^- \in S^-) > 1 - \epsilon/2$ for all i. But $S^+ \times S^-$ is $\tilde{d}$–compact and clearly $\mathbb{P}((X_i^+, X_i^-) \in S^+ \times S^-) > 1 - \epsilon$, so that $\{\mathcal{L}\left(X_i^+, X_i^-\right)\}$ is tight on $(E \times E, \tilde{d})$. □

Theorem 25. *Consider two continuous-time random processes X^+ and X^-, begun at the same fixed initial point, with càdlàg paths in a given complete separable metric state-space E (with Borel σ-algebra).* MEXIT *couplings exist through all time.*

Proof. We work first up to a fixed time T.

The space of càdlàg paths in a complete separable metric state-space over a fixed time interval $[0,T]$ can be considered as a Polish space (Maisonneuve [16, Théorème 1]), using a slight modification of the Skorokhod metric, namely the following *Maisonneuve distance*: if $\tau(t) : [0,T] \to [0,T]$ is a non-decreasing function determining a change of time, and if $|\tau| = \sup_t |\tau(t) - t| + \sup_{s\neq t} \log\left(\frac{\tau(t)-\tau(s)}{t-s}\right)$, then the Maisonneuve distance is given by

$$\operatorname{dist}_M(\omega, \widetilde{\omega}) \quad = \quad \inf_\tau \{|\tau| + \operatorname{dist}_E((\omega \circ \tau) - \widetilde{\omega})\}\,, \tag{16}$$

where ω and $\widetilde{\omega}$ are two càdlàg paths $[0,T] \to E$.

Consider a sequence of discretizations σ_n $(n = 1, 2, \ldots)$ of time-space $[0,T]$ whose meshes tend to zero, each discretization being a refinement of its predecessor. Note that by "discretization" we mean an ordered sequence $\sigma = (t_1, t_2, ...)$ where $0 < t_1 < t_2 < \ldots$. Let $X^{\pm,n}(t) = X^{\pm}(\inf\{s \in \sigma_n : s \geq t\})$ define discretized approximations of $X^{\pm}$ with respect to the discretization σ_n. Invoking Theorem 21, we require $X^{+,n}$, $X^{-,n}$ to be maximally coupled as discrete-time random processes sampled only at the discretization σ_n: since they are constant off σ_n, this extends to a maximal coupling of $X^{+,n}$, $X^{-,n}$ viewed as piecewise-constant processes defined over all continuous time.

For a given càdlàg path ω, the discretization of ω by σ_n converges to ω in Maisonneuve distance. This follows by observing that, for each fixed $\varepsilon > 0$, the time interval $[0,T]$ can be covered by pointed open intervals $t \in (t_-, t_+)$ such that $|\omega(s) - \omega(t-)| < \varepsilon/2$ if $s \in (s_-, t)$ and $|\omega(s) - \omega(t)| < \varepsilon/2$ if $s \in (t, s_+)$. By compactness we can select a finite sub-cover. For sufficiently fine discretizations σ we can then ensure the Maisonneuve distance between ω and the resulting discretization is smaller than ε. Consequently, both sequences $\{\mathcal{L}\left(X^{+,n}\right) : n = 1, 2, \ldots\}$, $\{\mathcal{L}\left(X^{-,n}\right) : n = 1, 2, \ldots\}$ are tight, and therefore by Lemma 24 we know that the sequence of joint distributions $\{\mathcal{L}\left(X^{+,n}, X^{-,n}\right) : n = 1, 2, \ldots\}$ is also tight.

Therefore (selecting a weakly convergent subsequence if necessary) we may suppose the joint distribution $(X^{+,n}, Y^{+,n})$ converges weakly to a limit which we denote by $(\widetilde{X}^+, \widetilde{X}^-)$. Note that the marginal laws of $(\widetilde{X}^+, \widetilde{X}^-)$ are the laws of X^+, X^-; however the joint distribution (which is to say, the specific coupling described by $(\widetilde{X}^+, \widetilde{X}^-)$) may well be affected by the exact choice of weakly convergent subsequence.

Suppose time t belongs to one of the discretizations in the sub-sequence, and thus eventually to all (since each discretization is a refinement of its predecessor). Consider the subspace of the Cartesian square of Skorokhod space given by $A_t = [\mathit{MEXIT} \geq t]$. This is a closed subset of the Cartesian square of Skorokhod space. Hence, by the Portmanteau Theorem of weak convergence (Billingsley [4, Theorem 2.1]),

$$\limsup_{n\to\infty} \mathbb{P}\left[(X^{+,n}, X^{-,n}) \in A_t\right] \quad \leq \quad \mathbb{P}\left[(\widetilde{X}^{+}, \widetilde{X}^{-}) \in A_t\right] .$$

If n is large enough so that $t \in \sigma_n$, then $\mathbb{P}\left[(X^{+,n}, X^{-,n}) \in A_t\right]$ is equal to the probability that *MEXIT* happens at t or later. By maximality of *MEXIT* for the discretized processes at time $t \in \sigma_n$, it follows for such n that $\mathbb{P}\left[(\widetilde{X}^{+}, \widetilde{X}^{-}) \in A_t\right] = \mathbb{P}\left[(X^{+,n}, Y^{+,n}) \in A_t\right]$ is itself maximal. The càdlàg property then implies maximality of the limiting coupling for all times $t \leq T$. The result on *MEXIT* for all time follows by application of the Kolmogorov Extension theorem as in Corollary 23. This concludes the proof. ∎

Remark 26. Sverchkov and Smirnov [30] prove a similar result for maximal couplings by means of general martingale theory.

Remark 27. Note that Théorème 1 of Maisonneuve [16] can be viewed as justifying the notion of the space of càdlàg paths: this space is the completion of the space of step functions under the Maisonneuve distance dist_M. Thus in some sense Theorem 25 is a maximally practical result concerning *MEXIT*!

5. *MEXIT* for diffusions

The results of Section 4 apply directly to diffusions, which therefore exhibit *MEXIT*. This section discusses the solution of a *MEXIT* problem for Brownian motions, which can be viewed as the limiting case for random walk *MEXIT* problems.

It is straightforward to show that *MEXIT* will generally have to involve constructions not adapted to the shared filtration of the two diffusion in question. By "faithful" *MEXIT* we mean a *MEXIT* construction which generates a coupling between the diffusions which is Markovian with respect to the joint and individual filtrations (see Rosenthal [28] and Kendall [13] for further background). We consider the case of elliptic diffusions X^+ and X^- with continuous coefficients.

Theorem 28. *Suppose X^+ and X^- are coupled elliptic diffusions, thus with continuous semi-martingale characteristics given by their drift vectors and volatility (infinitesimal quadratic variation) matrices, begun at the same point, with this initial point lying in the open set where either or both of the drift and volatility characteristics disagree. Faithful* MEXIT *must happen immediately.*

Proof. Let T be the *MEXIT* time, which by faithfulness will be a stopping time with respect to the common filtration. If X^+ and X^- are semimartingales agreeing up to the random time T, then the localization theorems of stochastic calculus tell us that the integrated drifts and quadratic variations of X^+ and X^- must also agree up to time T. It follows that X^+ and X^- agree as diffusions up to time T. Were the faithful *MEXIT* stopping time to have positive chance of being positive then the diffusions would have to agree on the range of the common diffusion up to faithful *MEXIT*; this would contradict our assertion that the initial point lies in the open set where either or both of the drift and volatility characteristics disagree. □

By way of contrast, *MEXIT* can be described explicitly in the case of two real Brownian motions X^+ and X^- with constant but differing drifts. Because of re-scaling arguments in time and space, there is no loss of generality in supposing that both X^+ and X^- begin at 0, with X^+ having drift $+1$ and X^- having drift -1.

Theorem 29. *If $X^\pm$ is Brownian motion begun at* 0 *with drift ± 1, then* MEXIT *between X^+ and X^- exists and is almost surely positive.*

Proof. The existence of *MEXIT* directly follows from Theorem 27. The almost surely positiveness will be shown in Subsection 5.2 below, through a limiting version of the random walk argument in Subsection 3.2.1. Alternatively one can argue succinctly and directly using the excursion-theoretic arguments of Williams' [1974] celebrated path-decomposition of Brownian motion with constant drift (an exposition in book form is given in Rogers and Williams [27]).

Calculation shows that the *bounded* positive excursions of X^+ (respectively $-X^-$) from 0 are those of the positive excursions of a Brownian motion of *negative* drift -1, while the *bounded* negative excursions of X^+ (respectively $-X^-$) from 0 are those of the negative excursions of a Brownian motion of *positive* drift $+1$. (The unbounded excursion of X^+ follows the law of the distance from its starting point of Brownian motion in hyperbolic 3-space, while the unbounded excursion of X^- has the distribution of the mirror image of the unbounded excursion of X^+.)

Viewing $X^\pm$ as generated by Poisson point processes of excursions indexed by local time, it follows that we may couple X^+ and X^- to share the same bounded excursions, with unbounded excursions being the reflection of each other in 0. Moreover the processes have disjoint support once they become different. So the Recognition Lemma for Maximal Coupling (Lemma 20) applies, and hence this is a *MEXIT* coupling. ☐

5.1. Explicit calculations for Brownian *MEXIT*

Let X^+ and X^- begin at 0, with X^+ having drift $+\theta$ and X^- having drift $-\theta$ with $\theta > 0$. The purpose of this section is to offer explicit calculations of *MEXIT* and *MEXIT* means.

Calculation 1. The meet of the distributions of X_t^+ and X_t^- is the meet of $N(\theta t, t)$ and $N(-\theta t, t)$, and the probability of *MEXIT* happening after time t is given by the total mass of this meet sub-probability distribution. Therefore:

$$\begin{aligned}
\mathbb{P}\left[MEXIT \geq t\right] &= \mathbb{P}\left[N(0,t) < -\theta t\right] + \mathbb{P}\left[N(0,t) > \theta t\right] \\
&= 2\,\mathbb{P}\left[N(0,t) > \theta t\right] \\
&= \frac{2}{\sqrt{2\pi}} \int_{\theta\sqrt{t}}^{\infty} e^{-u^2/2} du.
\end{aligned}$$

Thus,

$$\mathbb{E}\left[MEXIT\right] = \frac{2}{\sqrt{2\pi}} \int_0^{\infty} \int_{\theta\sqrt{t}}^{\infty} e^{-u^2/2} du dt = \theta^{-2}.$$

Remark 30. Excursion theoretic arguments can be used to confirm this is mean time to *MEXIT* for the specific construction given in Theorem 29.

Calculation 2. We now consider the expected amount of time T during which processes agree before *MEXIT* happens.

$$\begin{aligned}
\mathbb{E}\left[T\right] &= \int_0^{\infty} \mathbb{E}_W \left[\min\{e^{\theta W_t - \theta^2 t/2}, e^{-\theta W_t - \theta^2 t/2}\}\right] dt \\
&= 2\int_0^{\infty} \mathbb{E}_W \left[e^{-\theta W_t - \theta^2 t/2}; W_t > 0\right] dt \\
&= 2\int_0^{\infty} \int_0^{\infty} \frac{1}{\sqrt{2\pi t}} \exp\left(-\frac{(w+\theta t)^2}{2t}\right) dw dt \\
&= \theta^{-2}.
\end{aligned}$$

5.2. An explicit construction for *MEXIT* for Brownian motions with drift

In this section, we continue the scenario of Calculation 2 above. We see that *MEXIT* should have the complementary cumulative distribution function

$$\mathbb{P}\left[MEXIT \geq t\right] = 2\Phi(-\theta\sqrt{t}), \tag{17}$$

where $\Phi(y) = \int_{-\infty}^{y} (2\pi)^{-1/2} e^{-u^2/2} du$. A natural question to ask is as follows: how can one explicitly construct and understand this *MEXIT* time in a way that relates to the random walk constructions of Subsection 3.2.1? In this section we first deduce a candidate coupling and EXIT time, and then we proceed to show by direct calculation that our construction indeed gives the correct *MEXIT* time distribution above.

We note from the discrete constructions of Section 3 (in particular Subsection 3.2) that *MEXIT* is only possible when the Radon-Nikodym derivative between the "p" and "q" processes moves from being below 1 to above 1 or moves from being above 1 to below 1. Let $\mathbb{P}^+$, $\mathbb{P}^-$ denote the probability laws of X^+, X^- respectively. We have that

$$\frac{d\mathbb{P}^+}{d\mathbb{P}^-}(W_{[0,T]}) = \exp\{2\theta W_T\},$$

which is continuous in time with probability 1 under both $\mathbb{P}^+$ and $\mathbb{P}^-$. By analogy to the discrete case, the region in which *MEXIT* could possibly occur corresponds to the interface $\frac{d\mathbb{P}^+}{d\mathbb{P}^-}(W_{[0,T]}) = 1$ (that is, where $W_T = 0$).

Now we shall focus on the random walk example at the end of Subsection 3.2. We note that the *MEXIT* distribution given in (5) can be constructed as the first time the occupation time of 0 exceeds a geometric random variable with "success" probability $1 - 2\eta$. We aim to give a similar interpretation for the Brownian motion case. To do this, we shall use a sequence of random walks converging to the appropriate Brownian motions. To this end, let

$$\eta_n = \frac{1}{2}\left(1 - \frac{\theta}{n}\right),$$

and set X^{n+} and X^{n-} to be the respective simple random walks with up probability $1 - \eta_n$ and η_n and sped up by factor n^2. We assume (unless otherwise stated) that all processes begin at 0 so that we have that

$$X^{n+}(t) = \sum_{i=1}^{\lfloor n^2 t \rfloor} Z_i^{n+},$$

where $\{X_i^{n+}\}$ denote dichotomous random variable taking the value $+1$ with probability $1-\eta_n$ and -1 with probability η_n. We define X^{n-} analogously.

Given this setup, we have the classical weak convergence results that the law of X^{n+} converges weakly to that of X^+, and similarly X^{n-} converges weakly to X^-. Moreover the joint pre-*MEXIT* process described in Subsection 3.2 will have drift $-sgn(X_t)\theta$. The following holds for the *MEXIT* probability in (5)

$$\mathbb{P}\left[MEXIT > t\right] = \left(1 - \frac{\theta}{n}\right)^{n\ell_t^n} \longrightarrow e^{-\theta \ell_t^n},$$

where ℓ_t^n is the Local Time at 0 of the pre-*MEXIT* process for the nth approximation random walk.

In the (formal) limit as $n \to \infty$, this recovers the construction in Theorem 29 of Brownian motion *MEXIT* time, as follows. Let X be the diffusion with drift $-\operatorname{sgn}(X)\theta$ and unit diffusion coefficient started at 0 and let ℓ_t denote its local time at level 0 and time t. Then set E to be an exponential random variable with mean θ^{-1}. Then the pre-*MEXIT* dynamics are described by X until $\ell_t > E$ at which time *MEXIT* occurs. $E > 0$ w.p. 1 and hence *MEXIT* is positive a.s. since the local time is a continuous process.

We shall now verify that this construction does indeed achieve the valid *MEXIT* probability given in (17). By integrating out E we are required to show that

$$\mathbb{E}\left[e^{-\theta \ell_t}\right] = 2\Phi(-\theta\sqrt{t}) \ .$$

We proceed to do so. Firstly, we note that by symmetry, we may set ℓ_t to be the local time at level 0 of Brownian motion with drift $-\theta$ reflected at 0. Note that by an extension of Lévy's Theorem (see Peskir [20]) that the law of ℓ_t is the same as that of the maximum of Brownian motion with drift θ, i.e. that of X^+. Now this law is well-known as the Bachelier-Lévy formula (see for example Lerche [14]):

$$\mathbb{P}\left[\ell_t < a\right] = \Phi\left(\frac{a}{\sqrt{t}} - \theta\sqrt{t}\right) - e^{2a\theta}\Phi\left(\frac{-a}{\sqrt{t}} - \theta\sqrt{t}\right),$$

with density

$$f_{\ell_t}(a) = \frac{1}{\sqrt{t}}\left(\phi\left(\frac{a}{\sqrt{t}} - \theta\sqrt{t}\right) + e^{2a\theta}\phi\left(\frac{-a}{\sqrt{t}} - \theta\sqrt{t}\right) - 2\sqrt{t}\theta e^{2a\theta}\Phi\left(\frac{-a}{\sqrt{t}} - \theta\sqrt{t}\right)\right),$$

where ϕ is the standard normal density function $\phi(y) = (2\pi)^{-1/2}e^{-y^2/2}$. By direct manipulation of the exponential quadratic in the second of the three terms above, it can readily be shown to equal the first term. Thus

$$f_{\ell_t}(a) = \frac{1}{\sqrt{t}}\left(2\phi\left(\frac{a}{\sqrt{t}} - \theta\sqrt{t}\right) - 2\sqrt{t}\theta e^{2a\theta}\Phi\left(\frac{-a}{\sqrt{t}} - \theta\sqrt{t}\right)\right).$$

We now directly calculate the Laplace transform of this distribution to obtain (17).

$$\begin{aligned}\mathbb{E}\left[e^{-\theta\ell_t}\right] &= \frac{2}{\sqrt{t}}\int_0^\infty e^{-\theta a}\left(\phi\left(\frac{a}{\sqrt{t}} - \theta\sqrt{t}\right) - \sqrt{t}\theta e^{2a\theta}\Phi\left(\frac{-a}{\sqrt{t}} - \theta\sqrt{t}\right)\right) da \\ &=: \frac{2}{\sqrt{t}}(T_1 - T_2)\ .\end{aligned}$$

Using integration by parts, we easily work with T_2 to obtain

$$T_2 = T_1 - \sqrt{t}\Phi(-\theta\sqrt{t}),$$

which implies the assertion in (17), as required.

6. Conclusion

In this paper, we have studied an alternative coupling framework in which one seeks to arrange for two different Markov processes to remain equal for as long as possible, when started in the same state. We call this "un-coupling" or "maximal agreement" construction MEXIT, standing for "maximal exit" time. *MEXIT* sharply differs from the more traditional maximal coupling constructions studied in Griffeath [11], Pitman [21], and Goldstein [10] in which one seeks to build two different copies of the same Markov process started at two different initial states in such a way that they become equal as soon as possible.

This work begins with practical motivation for *MEXIT* by highlighting the importance of un-coupling/maximal agreement arguments in a few key statistical and probabilistic settings. With this motivation established, we develop an explicit *MEXIT* construction for Markov chains in discrete time with countable state-space. We then generalize the construction of *MEXIT* to random processes on Polish state-space in continuous time. We conclude with the solution of a *MEXIT* problem for Brownian motions.

As noted in Remark 8, the approach that we have followed in the construction of *MEXIT* introduces the role of copula theory in parametrising varieties of maximal couplings for random processes. Our future work will aim to establish a definitive role for *MEXIT* (as well as for probabilistic coupling theory in general) in copula theory.

Acknowledgments

We are grateful to the two anonymous referees whose suggestions helped improve the quality of this manuscript. The first author thanks Professor Larry Shepp for talking with him in 2011-2012 about the ideas that would eventually become MEXIT. The second author thanks Mateusz Majka and Aleks Mijatović for useful conversations, and acknowledges support by EPSRC grant EP/K013939 and also by the Alan Turing Institute under the EPSRC grant EP/N510129/1. The third author was supported by EPSRC grants EP/K034154/1, EP/K014463/1, EP/D002060/1. The fourth author received research support from NSERC of Canada.

7. References

[1] Aldous, D. (1983) Random walks on finite groups and rapidly mixing Markov chains. *Séminaire de probabilités de Strasbourg*, **17**, 243–297.

[2] Alquier, P. and Friel, N. and Everitt, R. and Boland, A. (2016) Noisy Monte Carlo: Convergence of Markov chains with approximate transition kernels. *Statistics and Computing*, **26**, 29–47.

[3] Andrieu, C. and Roberts, G. O. (2009) The pseudo-marginal approach for efficient Monte Carlo computations. *The Annals of Statistics*, **37**, 697–725.

[4] Billingsley, P. (1968) *Convergence of Probability Measures.* New York: John Wiley & Sons Inc.

[5] Burdzy, K. and Kendall, W. S. (2000) Efficient Markovian couplings: examples and counterexamples. *The Annals of Applied Probability*, **10**, 362–409.

[6] Ceperley, D. M. and Dewing, M. (1999) The penalty method for random walks with uncertain energies. *The Journal of Chemical Physics*, **110**, 9812–9820.

[7] Doob, J. L. (1953) *Stochastic processes.* New York: Wiley.

[8] Doob, J. L. (1994) *Measure Theory: Graduate Texts in Mathematics.* New York: Springer.

[9] Farrell, R. H. (1964) Asymptotic behavior of expected sample size in certain one sided tests. *The Annals of Mathematical Statistics*, **35**, 36–72.

[10] Goldstein, S. (1978) Maximal coupling. *Probability Theory and Related Fields*, **46**, 193–204.

[11] Griffeath, D. (1975) A maximal coupling for Markov chains. *Probability Theory and Related Fields*, **31**, 95–106.

[12] Halmos, P. R. (1978) *Measure Theory.* New York: Springer.

[13] Kendall, W. S. (2015) Coupling, local times, immersions. *Bernoulli*, **21**, 1014–1046.

[14] Lerche, H. R. (2013) *Boundary Crossing of Brownian motion: Its relation to the Law of the Iterated Logarithm and to Sequential Analysis: Volume 40.* New York: Springer Science & Business Media.

[15] Lindvall, T. (1992) *Lectures on the Coupling Method*. Wiley Series in Probability and Mathematical Statistics: Probability and Mathematical Statistics. New York: John Wiley & Sons Inc.

[16] Maisonneuve, B. (1972) Topologies du type de Skorohod. *Séminaire de probabilités de Strasbourg*, **6**, 113–117.

[17] Medina-Aguayo, F. and Lee, A. and Roberts, G. O. (2015) Stability of noisy Metropolis–Hastings. *Statistics and Computing*, **26**, 1187–1211.

[18] Nelsen, R. B. (2006) *An Introduction to Copulas*. New York: Springer Series in Statistics.

[19] Nicholls, G. K. and Fox, C. and Watt, A. M. (2012) Coupled MCMC with a randomized acceptance probability. *arXiv preprint arXiv:1205.6857*.

[20] Peskir, G. (2006). On reflecting Brownian motion with drift. In *Proc. Symp. Stoch. Syst.*, pp.1-5.

[21] Pitman, J. W. (1976) On coupling of Markov chains. *Probability Theory and Related Fields*, **35**, 315–322.

[22] Rigo, P. and Thorisson, H. (2016) Transfer theorems and right-continuous processes. *Theory of Stochastic Processes*, *to appear.*

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

[24] Roberts, G. O. and Rosenthal, J. S. (2004) General state space Markov chains and MCMC algorithms. *Probability Surveys*, **1**, 20–71.

[25] Roberts, G. O. and Rosenthal, J. S. (2007) Coupling and ergodicity of adaptive MCMC. *Journal of Applied Probability*, **44**, 458–475.

[26] Rogers, L. C. G. (1989) A guided tour through excursions. *Bulletin of the London Mathematical Society*, **44**, 458–475.

[27] Rogers, L. C. G and Williams, D. (1987) *Diffusions, Markov processes, and Martingales: Volume II.* New York: John Wiley & Sons.

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

[29] Stoyanov, J. (1997) *Counterexamples in Probability*. New York: John Wiley & Sons.

[30] Sverchkov, M. Y. and Smirnov, S. N. (1990) Maximal coupling for processes in D[0,∞]. *Dokl. Akad. Nauk SSSR*, **311**, 1059–1061.

[31] Thorisson, H. (2000) *Coupling, Stationarity, and Regeneration*. New York: Springer-Verlag.

[32] Völlering, F. (2016) On maximal agreement couplings. *Arxiv preprint 1608.01511.*

[33] Williams, D. (1974) Path decomposition and continuity of local time for one-dimensional diffusions, I. *Proceedings of the London Mathematical Society*, **s3-28**, 738–768.