Chapter 3

Convergence and couplings of Markov chains

In this chapter we consider couplings of discrete-time Markov chains. These couplings help in the study of the rate of convergence of the chains to their stationary distribution. Convergence of MCMC algorithms is a vast topic, and we do not try to provide an exhaustive account; instead we focus on how couplings are used in this context. For the account to make sense, we do have to introduce a number of concepts of Markov chain theory.

Section 3.1 gives some general elements of Markov chain theory with an emphasis on couplings and convergence rates. Section 3.2 then provides a short case study, detailing how some specific upper bounds on convergence rates can be obtained for MCMC algorithms.

3.1 Convergence of Markov chains

3.1.1 The convergence question

MCMC methods as recalled in Chapter 1 generate Markov chains starting from $X_0 \sim \pi_0$, where π_0 is the initial distribution. For $t \geq 1$, X_t given $X_0, \ldots, X_{t-1}$ is generated using X_{t-1} only (the "Markov property") from a distribution $K(X_{t-1} \to \cdot)$. Here K is called the transition kernel. We denote by π_t the marginal distribution of X_t at time t. MCMC algorithms are designed such that π, the "target distribution" of interest, is invariant by K, meaning $\pi(dy) = \int K(x \to dy)\pi(dx)$, or in words: if X_{t-1} follows π and X_t given X_{t-1} follows $K(X_{t-1} \to \cdot)$ then X_t follows π. MCMC algorithms are employed to approximate π, or to approximate integrals of "test functions" h with respect to π. There are a number of key practical questions of interest for any MCMC user, and we focus on this one:

> How fast is (X_t) going to π?

To formalize the question we need a notion of distance between distributions, such as the total variation or Wasserstein distance. The question becomes: how should t be chosen as a function of $\varepsilon > 0$ such that $\|\pi - \pi_t\|$ would be less than ε? Furthermore, we might want to know how t also depends on some "features" of the problem. For example if π is a distribution on images, we might want to know how t scales with the size of the image. An equivalent and more common phrasing of the convergence question is as follows: find a function $g(t)$ such that $\|\pi - \pi_t\| \leq g(t)$ for all $t \geq 1$. Again the function g would ideally be explicit in its dependency on some features of π, and in its dependency on π_0. Other distances than TV, such as Wasserstein and Prokhorov, are employed in the MCMC literature.

Among answers to the convergence questions, some are analytical calculations that provide a number of steps $\hat{t}(\varepsilon)$ such that $t \geq \hat{t}(\varepsilon)$ guarantees that $\|\pi - \pi_t\| \leq \varepsilon$ for some $\varepsilon > 0$, e.g. in Jones and Hobert [Section 4 2001], Roberts and Rosenthal [Section 3.5 2004] and references therein. Some are numerical procedures that provide an estimate of $\hat{t}(\varepsilon)$ [e.g. Johnson, 1996, Cowles and Rosenthal, 1998, Biswas et al., 2019]. Numerous algorithmic-specific bounds provide some qualitative results on what might affect the convergence rate, but involve quantities that are hard to estimate and thus cannot directly help users choose their numbers of iterations.

A bit of terminology: whatever the distance $\|\cdot\|$ between distributions, if $\|\pi - \pi_t\| \to 0$ the chain is "ergodic" or "converges to π". If $\|\pi - \pi_t\| \leq C(\pi_0)\rho^t$ for $\rho \in (0, 1)$ and a term $C(\pi_0)$ that is constant in t but might depend on π_0, then the chain is "geometrically ergodic". Note that, since this designation says nothing about the value of ρ, "geometrically ergodic" does not mean "fast" (or "slow"); it is a statement about the form of the dependency of $\|\pi - \pi_t\|$ on t. Furthermore, if $C(\pi_0)$ is independent of π_0 then the chain is termed "uniformly ergodic". On the other hand if $\|\pi - \pi_t\| \leq C(\pi_0)t^{-\kappa}$ for some $\kappa > 0$ then the

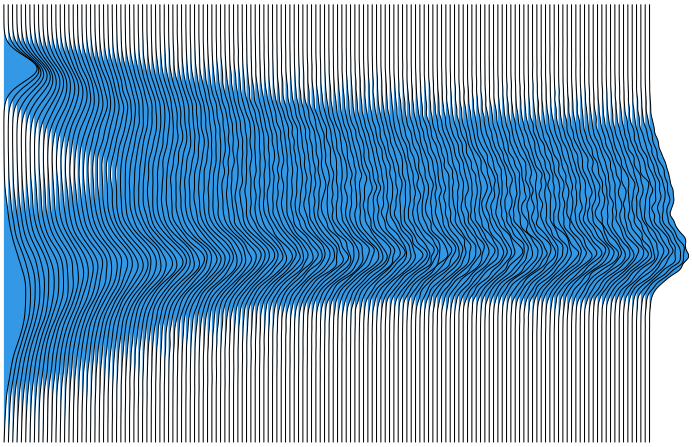

Figure 3.1: Convergence of marginals

chain would be "polynomially ergodic". Here "geometric" and "polynomial" refer to the form of $g(t)$ as a function of t. See Figure 3.1.

3.1.2 Irreducibility and aperiodicity

Let us start by reminding some basic properties of Markov chains. The fact that K is π-invariant is satisfied by design in MCMC settings, and can often be checked by the "detailed balance" condition. In itself π-invariance is not quite enough to guarantee convergence: for instance, if $K(x \to \cdot)$ is a Dirac mass at x, it is π-invariant, for any distribution π, but π_t is not π unless π_0 is.

We will denote by K^t the t-th iterate of the Markov kernel, i.e. $K^t(x \to A) = \mathbb{P}(X_t \in A | X_0 = x)$ for all x, A. Note that the notation is consistent with the linear algebra notation that, in discrete settings, can be used to describe the chain. The chain (X_t) is ϕ-irreducible if there exists a non-zero σ-finite measure ϕ such that, for all $A \in \mathcal{X}$ and $x \in \mathbb{X}$, there exists $t \geq 1$ such that $K^t(x, A) > 0$. It is a case-by-case task to verify that an MCMC algorithm is irreducible, i.e. the generated chain can go from anywhere to anywhere in a certain number of steps.

We will assume throughout that the Markov chain is aperiodic, which means that it is not periodic, which itself means that there is a partition of $\mathbb{X}$ into $\mathbb{X}_1, \ldots, \mathbb{X}_K$ such that if $x \in \mathbb{X}_i$ then $K(x, \mathbb{X}_j) = 1$ for some $j \neq i$, i.e. the chain goes systematically from some subpart of the space to another subpart. For example the Markov chain $X_t = t \mod 2$ is periodic, as it lands in 0 or 1 depending on t being even or odd. With MCMC algorithms it is usually simple to check that the generated chains are aperiodic; for example they are aperiodic whenever they have a non-zero chance of staying at their current state, $K(x \to x) > 0$ for all $x \in \mathbb{X}$.

Irreducibility and aperiodicity are essentially enough for the convergence of X_t to some invariant law. Theorem 4 of Roberts and Rosenthal [2004] (or e.g. Theorem 4.9 in Levin et al. [2009]) for instance states that if a chain is ϕ-irreducible, aperiodic and has stationary distribution π then $\|\pi_t - \pi\|_{\mathrm{TV}} \to 0$ as $t \to \infty$ where π_0 is a Dirac mass on $x \in \mathbb{X}$, for π-almost every x. This motivates the next question: the rate at which this convergence occurs.

3.1.3 Convergence rates and couplings

We define couplings of Markov chains, and what the "coupling inequality" means in the setting of Markov chains. A coupling of two Markov chains $(X_t)_{t\geq 0}$ and $(Y_t)_{t\geq 0}$ will refer to a joint process $(X_t, Y_t)_{t\geq 0}$ on $\mathbb{X} \times \mathbb{X}$, such that, dropping the Y_t's and keeping the X_t's we retrieve exactly the same process as the original $(X_t)_{t\geq 0}$. Likewise if we dropped the X_t's and kept the Y_t's, we retrieve the original process $(Y_t)_{t\geq 0}$. The joint process $(X_t, Y_t)_{t\geq 0}$ itself might be Markovian or not. Marginally (X_t) and (Y_t) might be identical or not, they might start from the same distribution π_0 or not, they might evolve according to the same transition kernel K or not, etc. We will denote the marginal distribution of a random variable X by $\mathcal{L}(X)$.

The "coupling inequality" in the context of Markov chains refers to the following. Let τ be a random variable with support on $\mathbb{N}$ such that $X_t = Y_t$ for all $t \geq \tau$. We allow τ to be $+\infty$ with non-zero probability. Then the inequality reads

$$\|\mathcal{L}(X_t) - \mathcal{L}(Y_t)\|_{\mathrm{TV}} \leq \mathbb{P}(\tau > t). \tag{3.1}$$

It holds for any process $(X_t, Y_t)_{t\geq 0}$ on $\mathbb{X} \times \mathbb{X}$; although if $\tau = +\infty$ almost surely, the right-hand side becomes one, and then the inequality becomes vacuous. The left-hand side concerns a relation between marginal distributions, while the right-hand side concerns the dependency within the pair (X_t, Y_t): the event $\{\tau > t\}$ is equivalent to $\{X_t \neq Y_t\}$. The inequality is obtained from the representation $\|\mathcal{L}(X_t) - \mathcal{L}(Y_t)\|_{\mathrm{TV}} = \inf \mathbb{P}(X_t \neq Y_t)$ over all couplings (X_t, Y_t). The variable τ will often be called the "meeting time". When $\mathbb{P}(\tau < \infty) = 1$ the coupling is sometimes said to be "successful". See Figure 3.2 for an example of pair of trajectories that meets.

How are we going to use (3.1) to obtain convergence rates? Often, by assuming that (Y_t) is a Markov chain "at stationarity", i.e. marginally distributed according to π while (X_t) is our chain of study, started from π_0. Both chains evolve according to the same transition kernel K. Then the left-hand side of (3.1) becomes $\|\pi_t - \pi\|_{\mathrm{TV}}$. The inequality then states that the convergence of (X_t) to stationarity can be upper-bounded by $\mathbb{P}(\tau > t)$, the "tail probabilities" of the meeting time τ. This was originally the setting considered by Wolfgang Doeblin, [see the appendix 'Epilogue' in Lindvall, 2002], credited for the invention of the coupling technique. If we can further obtain the convergence rate of $\mathbb{P}(\tau > t)$ to zero, or an upper bound for it, then we would obtain an upper bound for $\|\pi_t - \pi\|_{\mathrm{TV}}$. For example, if $\mathbb{E}[\varphi(\tau)] < \infty$ with φ non-negative and increasing, then $\varphi(t)\mathbb{P}(\tau > t) \to 0$ and thus $\varphi(t)^{-1}$ gives an upper bound on the rate of

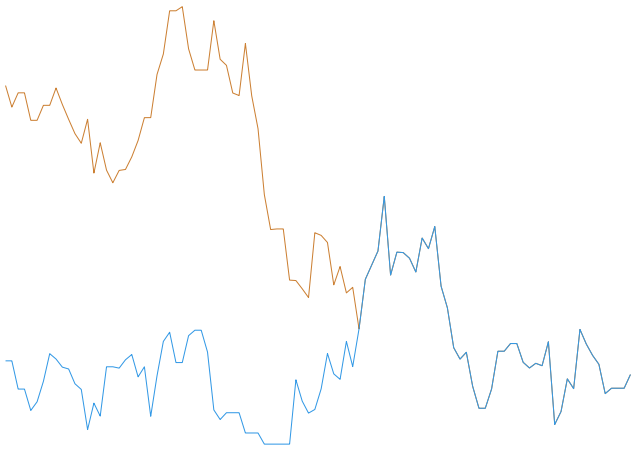

Figure 3.2: Faithful coupling of Markov chains

convergence [Section I.3 in Lindvall, 2002]. Indeed $\varphi(t)\mathbb{P}[\tau > t] \leq \mathbb{E}[\varphi(\tau)\mathbb{1}(\tau > t)] \to 0$ if $\mathbb{E}[\varphi(\tau)] < \infty$, thus $\mathbb{P}(\tau > t) \leq \varphi(t)^{-1}$ as t gets large.

A natural question is whether, for a given pair of processes $(X_t)_{t\geq 0}$ and $(Y_t)_{t\geq 0}$, there exists a coupling such that (3.1) is in fact an equality, for all $t \geq 0$. This would be called a "maximal coupling of the two processes", by analogy with maximal couplings of random variables. This question was first addressed in the works of Griffeath [1975], Goldstein [1979]. Such a maximal coupling can be constructed abstractly in great generality. In certain cases that maximal coupling is unique and Markovian, e.g. Hsu and Sturm [2013] in the case of Euclidean Brownian motions. On the other hand, there are known situations where Markovian couplings can only provide sub-optimal bounds on the convergence rate; see e.g. Anil Kumar and Ramesh [1999], Hayes and Vigoda [2003], Burton and Kovchegov [2006]. In the context of this course, we will not be much concerned with maximal couplings of stochastic processes, nor with obtaining optimal bounds, thus we will be content with non-maximal couplings, especially if they provide practical bounds and can be simulated.

In the previous chapter we considered other distances than TV. In particular, the Wasserstein distance was mentioned in the case of metric spaces, as well as the Prokhorov distance; denote by $d(\cdot, \cdot)$ the ground metric. Such distances are also found in the study of convergence rates of Markov chains. For instance Gibbs [2004], Durmus and Moulines [2015], Douc et al. [Chapter 20, 2018] or Qin and Hobert [2019] for the convergence of Markov chains in Wasserstein, and Diaconis and Freedman [1999] or Mangoubi and Smith [2017] for results formulated in Prokhorov. To obtain a convergence rate in such distances, one does not need to consider coupled chains that exactly meet. It is enough to consider chains that become "close" to one another in some sense. Gibbs [2004] provides a simple argument as follows. Suppose that $\mathbb{E}[d(X_{t+1}, Y_{t+1})|X_t, Y_t] \leq cd(X_t, Y_t)$ for all

$t \geq 0$, almost all X_t, Y_t and some $c \in (0,1)$, in other words the expected distance between two chains $(X_t)_{t\geq 0}$ and $(Y_t)_{t\geq 0}$ decreases on average. Then by induction $\mathbb{E}[d(X_t, Y_t)|X_0, Y_0] \leq c^t d(X_0, Y_0)$ for $t \geq 1$, and the coupling representation of the 1-Wasserstein distance $W_1(\pi_t, \pi)$ yields the bound $W_1(\pi_t, \pi) \leq c^t d(x_0, y_0)$ conditional on $(X_0, Y_0) = (x_0, y_0)$.

We next provide some key arguments that can be used to study the contraction or meeting of pairs of Markov chains. These arguments appear in some form in most convergence analyses of MCMC algorithms.

3.1.4 Minorization

We start with a technique called "minorization". Understanding this technique helps develop some intuition regarding how pairs of chains can meet exactly even on continuous state spaces. Let us assume that there exists an integer $k \in \mathbb{N}$ and some $\epsilon > 0$ such that for all x, A $K^k(x \to A) > \epsilon$. This holds for example when $\mathbb{X}$ is a finite state space or a compact space, and when K is irreducible and aperiodic. It is unlikely to hold when the state space is unbounded. We work under the assumption of bounded spaces momentarily and denote the uniform distribution on $\mathbb{X}$ by $\nu(x) = \mathbb{1}(x \in \mathbb{X})/|\mathbb{X}|$. We can then write the distributions $\{K^k(x \to \cdot), x \in \mathbb{X}\}$ as mixtures of two components: a component that depends on x, and another component that does not. Under the assumption that $K^k(x \to A) > \epsilon$ for all x, A, there exists $\varepsilon > 0$ such that, for all x, A,

$$K^k(x \to A) = \varepsilon \nu(A) + (1 - \varepsilon)\, R(x \to A), \tag{3.2}$$

$$\text{with} \quad R(x \to A) = \frac{K^k(x \to A) - \varepsilon \nu(A)}{1 - \varepsilon}. \tag{3.3}$$

The kernel $R(x \to \cdot)$ can be called the "residual kernel", as it captures the part of $K^k(x \to \cdot)$ that is not in the common part ν. The representation suggests that we can generate coupled Markov chains (X_t, Y_t) as follows. Given (X_t, Y_t), if $X_t = Y_t$, then sample $X_{t+k} \sim K^k(X_t \to \cdot)$ and set $Y_{t+k} = X_{t+k}$. If $X_t \neq Y_t$, sample $U \sim \text{Uniform}(0, 1)$.

- If $U < \varepsilon$, sample $Z \sim \nu$ and set $X_{t+k} = Z$ and $Y_{t+k} = Z$.
- Otherwise ($U \geq \varepsilon$), sample $X_{t+k} \sim R(X_t \to \cdot)$ and $Y_{t+k} \sim R(Y_t \to \cdot)$, independently.

One could, at least theoretically, "fill in" the states $(X_{t+1}, \ldots, X_{t+k-1})$ and $(Y_{t+1}, \ldots, Y_{t+k-1})$ conditional on the states (X_t, Y_t) at time t and (X_{t+k}, Y_{t+k}) at time $t + k$. By construction such chains (X_t) and (Y_t) marginally evolve according to K. Yet the two chains can meet: it occurs when the next states are sampled from ν, which happens with probability ε. Note the similarity between the above construction and what we have previously discussed in the context of maximal couplings. The distribution ν can be said to "minorize" the distributions $K^k(x \to \cdot)$ for all $x \in \mathbb{X}$, because it is the "common part" of all these distributions. Note also that it would actually be quite difficult to sample $(X_t, Y_t)_{t\geq 0}$ as described above in the context of realistic MCMC applications.

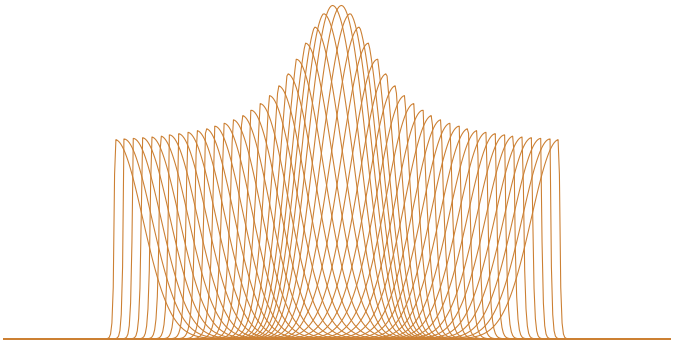

Figure 3.3: Minorization condition

With the above construction the event $\{X_{t+k} = Y_{t+k}\}$ occurs with probability at least ε given (X_t, Y_t) at time t: it occurs whenever two states are drawn from the "common part" ν. Once the chains have met, they remain identical forever. This can be called a "faithful coupling". In this construction, the meeting time is the time for an ε-coin to give a first success, which is Geometric with probability ε. This can lead to a geometric rate of convergence in TV with coefficient $(1-\varepsilon)^{1/k}$ in the original time parametrization; see Theorem 4.9 in Levin et al. [2009] or Theorem 8 in Roberts and Rosenthal [2004].

The minorization condition can be extended and generalized in various ways. In the form given above, with $K^k(x \to A) \geq \varepsilon\nu(A)$ for all x, the assumption seems very strong and precludes most Markov chains on unbounded state spaces. An exception is Metropolis–Hastings with independent proposals, depending on the tails of the proposal and of the target [Tierney, 1994]. A more recent example is the Gibbs sampler of Polson et al. [2013] for logistic regression, as explained in Choi and Hobert [2013]. Even when the reasoning does apply formally, the value of ε might be so small that the rate of convergence might be catastrophically close to one. A common relaxation is to assume that the minorization condition holds only on a subset of the entire space. For instance it is often assumed that there exist $C \subset \mathbb{X}$, $k \in \mathbb{N}$, $\varepsilon > 0$ and a probability measure ν such that $K^k(x \to A) \geq \varepsilon\nu(A)$ for all A whenever x is in the set C. Such sets C are called "small sets" for the Markov chain with kernel K. An adaptation of the above reasoning states that the two chains meet with probability ε when they are simultaneously in the set C. Then another ingredient must be added to justify that the two chains indeed jointly visit the set C, and will do so often enough; see drift conditions in Section 3.1.6.

3.1.5 Contractions and iterated random functions

At the end of Section 3.1.3 we mentioned the argument of Gibbs [2004], stating that if contraction occurs at every step in the sense $\mathbb{E}[d(X_{t+1}, Y_{t+1})|X_t, Y_t] \leq cd(X_t, Y_t)$ for some $c < 1$ and almost all X_t, Y_t, then the distance $W_1(\pi_t, \pi)$ goes

to zero geometrically as c^t. Variants of that reasoning have proved useful in various settings; however it might not be realistic to assume that contraction occurs at every single step in the form written above. We will follow briefly some arguments in Diaconis and Freedman [1999], which surveys various results from the literature, and also provides an opportunity to describe "iterated random functions", which will be useful in some future chapters. Steinsaltz [1999] would be another good reference to follow here.

A Markov chain with kernel K can nearly always be represented as

$$X_t = \psi(X_{t-1}, U_t), \quad U_t \sim \text{Uniform}(0,1), \tag{3.4}$$

where ψ is some deterministic function and (U_t) a sequence of independent uniforms. Steinsaltz [1999] cites Kifer [1986] for that. To see this, first note that with a single uniform $U \sim \text{Uniform}(0,1)$ one can obtain arbitrarily many independent uniforms; at least, conceptually. One way of obtaining two uniform variables from one, for example, goes as follows. Decompose U into its binary expansion, obtaining an infinite sequence (B_k) in $\{0,1\}$ such that $U = \sum_{k\geq 1} 2^{-k} B_k$. The sequence (B_k) is made of independent Bernoulli(1/2). We can then take every odd index $2k-1$ and every even index $2k$, for all k, to construct two infinite sequences of independent Bernoulli(1/2), say $(B_j^{(1)})$ and $(B_j^{(2)})$. Finally we can define $U^{(1)} = \sum_{j\geq 1} 2^{-j} B_j^{(1)}$ and $U^{(2)} = \sum_{j\geq 1} 2^{-j} B_j^{(2)}$, two independent draws from Uniform(0,1), and feed the multitude. Thus, conceptually having access to a single uniform is equivalent to having access to arbitrarily many uniforms.

Furthermore, since the function ψ can be anything, we can convince ourselves that we could "implement" any Markov kernel K with some choice of ψ and arbitrary numbers of uniform inputs. Multiple functions ψ could correspond to the same Markov kernel K. For example if ψ transforms some uniform draws into Normals, it can be done in multiple ways: inverse transform, Box-Muller, ziggurat [Marsaglia and Tsang, 2000], etc. This would correspond to different functions ψ, but the same transition kernel K. Whenever "common random number" strategies are mentioned in this context it refers to the use of a common stream of random inputs (U_t) in some specific ψ, and starting from two multiple states. It is therefore important to keep in mind that such strategies are, in general, sensitive to the particular choice of functions ψ.

An equivalent representation to (3.4) is $X_t = \Psi_t(X_{t-1})$, with Ψ_t being itself a random function defined as $\psi(\cdot, U_t)$, where U_t is uniform. Depending on the situation it might be more convenient to use either notation. The notation $\psi(\cdot, U_t)$ is more precise but can become cumbersome when considering compositions of functions: it is easy to write $\Psi \circ \Psi'$ rather than something like $\psi(\psi(\cdot, u'), u)$. With an iterated random function representation of a Markov chain, we can make the idea of "contraction" precise by assuming that functions $x \mapsto \psi(x,u)$ are Lipschitz, with coefficients depending on u. The Lipschitz coefficient is defined as

$$L(u) = \sup_{x \neq y} \frac{d(\psi(x,u), \psi(y,u))}{d(x,y)},$$

where d is the ground distance. It is related to contraction since $d(\psi(x,u), \psi(y,u))$

can be interpreted as the distance between two chains propagated using ψ with common random input u, from two different starting points x and y. Viewing u as the realization of a Uniform leads to $L(U)$ being a random variable; Diaconis and Freedman [1999] details the measurability of Lipschitz coefficients. Diaconis and Freedman [1999] considers chains that are contracting in the sense that:

- $L(U)$ has polynomial tails, $\mathbb{P}(L(U) > \ell) \leq \alpha \ell^{-\beta}$ for some α, β,
- and $\mathbb{E}[\log L(U)] < 0$. This does not imply $\mathbb{E}[L(U)] < 1$, but the converse is true by Jensen's inequality.

Diaconis and Freedman [1999] assumes further that $\zeta(U) = d(\psi(x_0, U), x_0)$, for an arbitrary point $x_0 \in \mathbb{X}$, has polynomial tails. These assumptions on $L(U)$ and $\zeta(U)$ are shown to have great consequences, including the existence of $A_x < \infty$ and $r \in (0, 1)$ such that $\|\pi_t - \pi\| \leq A_x r^t$ in the Prokhorov distance, and assuming that the chain $(X_t)_{t \geq 0}$ starts at some arbitrary point $x \in \mathbb{X}$.

Without entering into a detailed justification of this geometric ergodicity result, we mention a key idea that can be found both in Diaconis and Freedman [1999] and Steinsaltz [1999], and that involves the "backward process". It plays an important role in "coupling from the past" [Propp and Wilson, 1996], a topic discussed in a later chapter, thus we might as well get used to the idea. First observe that, "unrolling the recursion", we can write

$$X_t = \Psi_t(X_{t-1}) = \Psi_t \circ \ldots \circ \Psi_1(X_0),$$

where (Ψ_t) are independent draws of the form $(\psi(\cdot, U_t))$ for $U_t \sim \text{Uniform}(0, 1)$. Since the functions (Ψ_t) are independent and identically distributed, we can shuffle them arbitrarily without modifying the marginal distribution of X_t. In particular the variable Y_t defined as

$$Y_t = \Psi_1 \circ \ldots \circ \Psi_t(X_0),$$

has the same distribution as X_t. However, the processes (X_t) and (Y_t), termed "forward" and "backward" have markedly different behaviors. Indeed, under the assumptions mentioned above on $L(U)$ and $\zeta(U)$, while the process (X_t) converges in distribution to π, the process (Y_t) converges to a fixed limit in $\mathbb{X}$, that depends on the realization of the infinite sequence (Ψ_t) but not on the starting point X_0. Note that (Y_t) is not the same process as the "time-reversal" of (X_t), which is the process corresponding to viewing the process (X_t) backward in time, i.e. assuming the time index t can be any value in $\mathbb{Z}$, the process $(BX_t)_{t \in \mathbb{Z}}$ defined as $BX_t = X_{t_0 - t}$ for some arbitrary t_0.

For example consider the auto-regressive process where $X_t = \rho X_{t-1} + \eta_t$ where $\rho \in (-1, 1)$ and η_t is a sequence of independent standard Normals. This process can be pedantically interpreted as generated by the unadjusted Langevin algorithm targeting a Normal distribution, so there is some MCMC analogy. We can write this process as $X_t = \psi(X_{t-1}, \eta_t)$ with $\psi(x, \eta) = \rho x + \eta$. The Lipschitz coefficient is here independent of η_t and equal to ρ. It might not be a surprise

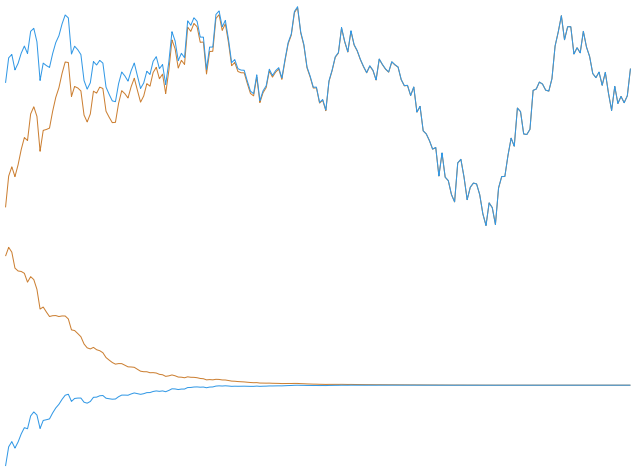

Figure 3.4: Forward processes with common random numbers, and backward processes with common random numbers

that X_t moves around the space as time progresses and converges in distribution to a Normal. On the other hand, unrolling the recursion that defines Y_t, we find

$$Y_t = \rho^t X_0 + \rho^{t-1}\eta_t + \rho^{t-2}\eta_{t-1} + \ldots + \rho^1\eta_2 + \eta_1,$$

and thus $Y_{t+1} = Y_t + \rho^t(\eta_{t+1} + (\rho - 1)X_0)$. From this we see that the increments of Y_t are smaller and smaller in magnitude, in such a way that Y_t converges to a fixed value as $t \to \infty$. Since $Y_t = X_t$ in distribution for all times t, we can imagine that this fixed limiting value of (Y_t) has itself a Normal distribution, "as $(\eta_t)_{t\geq 0}$ varies". Figure 3.4.

To summarize this section, the iterated random function representation of a chain leads to interesting pairs of processes. The forward process (X_t) and the backward process (Y_t) in the above definitions have the same marginal distributions at all times, and thus share the same rate of convergence to their limits. This has proved useful in establishing these rates, as illustrated in Diaconis and Freedman [1999]. Another interesting coupling of chains is obtained by using "common random numbers", say $X_t = \psi(X_{t-1}, U_t)$ and $Y_t = \psi(Y_{t-1}, U_t)$. Under conditions on the Lipschitz coefficients of the functions $x \mapsto \psi(x, u)$ for $u \in [0, 1]$, these "forward" coupled chains might contract towards one another as time progresses.

3.1.6 Drift condition

The preceding discussions involve conditions on Markov chains and their implementations, such as minorization, or contraction, or Lipschitz coefficients of iterated random functions. In many settings, these conditions might hold locally but not globally. They might hold when the chains are simultaneously

in some subset $C \times C$ of the state space $\mathbb{X} \times \mathbb{X}$, but not on the entire space. For example we might have $K(x \to A) \geq \varepsilon\nu(A)$ when $x \in C$ for some set C, or we might observe some contraction between two chains when they are both in some set C. These conditions do not lead to geometric ergodicity on their own; they must now be complemented with additional arguments, that ensure the presence of the chains in those sets. Indeed, properties that only hold on some set C would be of little use if such C was never, or almost never visited by the chains! Here we briefly describe "drift conditions", which are commonly used in the MCMC literature to study the return times to some sets. The combined use of "drift and minorization" is a very standard path towards geometric ergodicity in total variation, on unbounded continuous state spaces, as described succinctly in Rosenthal [2002], and with various examples in Roberts and Rosenthal [2004], Jones and Hobert [2001]. Drift and contraction conditions can be combined to obtain geometric ergodicity in Wasserstein distance as in Durmus and Moulines [2015], Qin and Hobert [2019]. The PhD thesis of Jerison [2016] provides an excellent introduction to drift conditions, which we follow quite closely here.

We say that there is a drift function for a Markov kernel K, a set C, a parameter $\lambda \in (0, 1)$ and a function $V : \mathbb{X} \to [1, \infty)$, if

$$KV(x) = \mathbb{E}[V(X_{t+1})|X_t = x] \leq \lambda V(x) \quad \text{for all } x \notin C. \tag{3.5}$$

In words, it means that whenever the chain goes out of the set C, the univariate process $V(X_t)$ starts decreasing in expectation at a geometric rate. This decrease cannot go on indefinitely because $V(x)$ is above one for all $x \in \mathbb{X}$. Therefore the chain must eventually re-enter the set C. This hopefully provides some intuition regarding why drift conditions helps controlling excursions outside of sets. In passing note how the study shifts to a stochastic process $(V(X_t))_{t\geq 0}$ on $\mathbb{R}$, and not on $\mathbb{X}$. More precisely, Jerison [2016, Lemma 2.6 or Lemma 4.7] states the following result. Denote by $\mathbb{E}_x$ expectations with respect to the Markov chain started at $X_0 = x$, and denote by τ_C the first time that the chain X_t enters C, $\tau_C = \inf\{t \geq 0 : X_t \in C\}$, also called the "hitting time" of the set. Then (3.5) implies that for all $x \notin C$,

$$\mathbb{E}_x \left[\lambda^{-\tau_C}\right] \leq V(x). \tag{3.6}$$

Note that if $x \in C$ then $\tau_C = 0$ and thus the inequality holds too because $V(x) \geq 1$ for all $x \in \mathbb{X}$. The left-hand side, which does not involve V, provides a point-wise lower bound for drift functions with parameter λ. One can check that the function $\tilde{V} : x \mapsto \mathbb{E}_x \left[\lambda^{-\tau_C}\right]$ itself satisfies (3.5). Thus it is the "smallest" drift function for the set C and parameter λ. To prove (3.6), we can follow Jerison [2016, Lemma 4.7] and first consider the following, proved by induction, that for all $t \geq 1$ and $x \notin C$,

$$V(x) \geq \lambda^{-t}\mathbb{E}_x \left[V(X_t)\mathbb{1}(\tau_C > t)\right] + \sum_{s=1}^{t} \lambda^{-s}\mathbb{P}_x(\tau_C = s). \tag{3.7}$$

It follows that $V(x) \geq \lambda^{-t}\mathbb{P}(\tau_C > t) + \sum_{s=1}^{t} \lambda^{-s}\mathbb{P}_x(\tau_C = s)$, since $V \geq 1$. Thus for all $x \notin C$ and all $t \geq 1$, $V(x) \geq \lambda^{-t}\mathbb{P}_x(\tau_C > t) \geq \lambda^{-t}\mathbb{P}_x(\tau_C = \infty)$

and therefore $\mathbb{P}_x(\tau_C = \infty) = 0$. Next, sending t to infinity, we obtain $V(x) \geq \sum_{s=1}^{\infty} \lambda^{-s} \mathbb{P}_x(\tau_C = s) = \mathbb{E}_x[\lambda^{-\tau_C}]$, which completes the proof.

Inequality (3.6) implies that $\mathbb{E}_x[\tau_C] \leq \log V(x) / \log(\lambda^{-1})$ by Jensen's inequality, which provides further insight on the relationship between drift functions and hitting times. In some settings one can numerically approximate $\mathbb{E}_x[\tau_C]$ for various $x \notin C$ and get a sense of how $\log V$ should behave. Also (3.6) implies that $\mathbb{P}_x(\tau_C > t) \leq a\lambda^t$ for all t and some constant $a < \infty$. Indeed $\lambda^{-t}\mathbb{P}_x(\tau_C > t) \leq \mathbb{E}_x[\lambda^{-\tau_C} \mathbb{1}(\tau_C > t)]$, which goes to zero as $t \to \infty$ by dominated convergence. In words, a drift condition such as (3.5) implies that return times to C have at most Geometric tails.

The attentive reader will have noted that we have described drift conditions for single chains (X_t). On the other hand some of the previous discussions required pairs of chains to be simultaneously in some set. Thus, it would seem more important to study the return times of the pair $(X_t, Y_t)_{t \geq 0}$ to a subset $C \times C$ of $\mathbb{X} \times \mathbb{X}$. These are sometimes called "bivariate drift" conditions; they can potentially be derived from univariate drift conditions, as in Roberts and Rosenthal [Proposition 11, 2004].

3.2 Concrete examples

[Work in progress.]

We next give more concrete examples of how rates of convergence are obtained for certain simple or common MCMC algorithms. The point of this section is to see how geometric ergodicity can be established and the role of couplings in such results.

3.2.1 Metropolis–Hastings on Bayesian logistic regression

In the first chapter, we have described a Bayesian logistic regression on the titanic data set. The target distribution is the posterior distribution in a Bayesian analysis. That distribution is not quite Normal, but in the case of logistic regression with the titanic data set, it is not very far: it can be shown that the posterior distribution is "strongly log-concave". Indeed, we can differentiate the logarithm of the posterior density, and obtain

$$
\begin{aligned}
-\log \pi(\beta) &= \frac{\sigma^{-2}}{2}\beta^T\beta + y^T x\beta + \sum_{i=1}^{n} \log\left(1 + \exp\left(-x_i^T\beta\right)\right), \\
-\nabla_\beta \log \pi(\beta) &= \sigma^{-2}\beta + x^T y + \sum_{i=1}^{n} \frac{x_i}{1 + \exp(x_i^T\beta)}, \\
-\nabla^2_\beta \log \pi(\beta) &= \sigma^{-2} I + \sum_{i=1}^{n} \frac{\exp(x_i^T\beta)}{(1 + \exp(x_i^T\beta))^2} x_i x_i^T.
\end{aligned}
$$

Here σ^2 is the prior variance on each coefficient, and x the $n \times p$ matrix with rows (x_i), and y the vector of binary outcomes. From this we can see that

the target distribution is "strongly log-concave"... which implies unimodal, and with sub-Gaussian tails. More precisely, the spectral norm of $-\nabla^2_\beta \log \pi(\beta)$ is lower bounded by $\sigma^{-2} > 0$ and upper bounded by the largest eigenvalue of $\sigma^{-2} I + 4^{-1} \sum_{i=1}^n x_i x_i^T$. Therefore [give some reference], we can explicitly find the constants m and M such that $m \leq M$ and for all $x, x' \in \mathbb{X} = \mathbb{R}^p$,

$$U(x) - U(x') - \nabla U(x')^T (x - x') \geq \frac{m}{2} \|x - x'\|_2^2,$$
$$\|\nabla U(x) - \nabla U(x')\|_2 \leq M \|x - x'\|_2.$$

where the "potential" U is $-\log \pi$. The potential is said to be m-strongly convex, with an M-Lipschitz gradient.

These are favorable conditions for MCMC methods. Indeed log-concavity has often been an assumption that formalizes what a "easy" target distribution might look like, see Frieze et al. [1994], Brooks [1998] or more recently Dwivedi et al. [2019]. Note that there are also a variety of sampling techniques outside of MCMC that have been designed specifically for log-concave distributions [e.g. Wild and Gilks, 1993, Görür and Teh, 2011, Devroye, 2012].

For random walk Metropolis–Hastings with e.g. Normal increments, assumptions on the tails of the target can be shown to imply geometric ergodicity, as in Mengersen and Tweedie [1996]. Jarner and Hansen [2000] provide a result applicable to the present setting in their Section 4, using drift and minorization conditions. The drift function is chosen to be $V : \beta \mapsto c\pi(\beta)^{-1/2}$ with some constant c chosen such that $V(\beta) \geq 1$ for all β. Remarkably the article provides results in the reverse direction: if a random walk Metropolis–Hastings algorithm is geometrically ergodic, then the tails of the target distribution must decay at least exponentially. In these articles, couplings are not explicitly mentioned, but they are underlying the use of minorization condition to provide geometric ergodicity results in total variation; see e.g. Rosenthal [2002] or Roberts and Rosenthal [2004].

The log-concavity assumption has been used extensively in studies on the rate of convergence of MCMC algorithms with gradient-based proposals. The main examples are Langevin and Hamiltonian Monte Carlo algorithms. For example Dalalyan [Theorem 2, 2017] provides a geometric ergodicity result in total variation for the unadjusted Langevin algorithm. The rate of convergence is very explicit in its dependency on m, M and the dimension p. Such constants m and M can be improved in practice by the use of preconditioners. A further interest in this type of result is in the comparison against classical results in convex optimization, and thus this allows to compare the difficulty of sampling versus optimizing for log-concave distributions.

For Hamiltonian Monte Carlo Mangoubi and Smith [2017] provides results in TV and in Wasserstein under assumption of strongly log-concavity of the target. The approach of Mangoubi and Smith [2017] relies strongly on "common random numbers" couplings, also called "synchronous coupling". It is a remarkable property of Hamiltonian dynamics on strongly concave targets that a pair of trajectories, propagated using common initial velocities, tends to contract. See Figure 3.5 for an illustration, on the posterior distribution associated with the

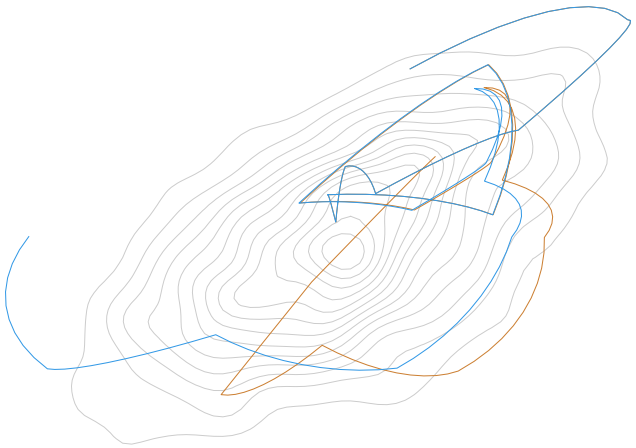

Figure 3.5: Coupled Hamiltonian Monte Carlo trajectories, driven by common random numbers

Titanic data set. The article of Bou-Rabee et al. [2018] considers alternative couplings using reflection, with the aim of improving the contraction properties and relaxing the strong log-concavity assumption. The contraction of HMC can be worked out explicitly with a Gaussian target distribution, as illustrated in Heng and Jacob [2019].

3.2.2 Ising model and image restoration

We can consider another example in a high-dimensional continuous state space, related to the Ising model. The example is taken from Gibbs [2004] and provides an example of geometric ergodicity in Wasserstein distance obtained via the contraction of chains propagated forward using common random numbers.

The model is designed for restoring a grey-scale image $y = \{y[i]\}_{i=1}^N$ with N pixels. It is assumed that each pixel $y[i]$ is a value in $[0, 1]$, representing a level of "grey", e.g. 0 is black and 1 is white. Each $y[i]$ is assumed to be a noisy version of an underlying level of grey $x[i] \in [0, 1]$. The model on $\{x[i]\}_{i=1}^N$ assumes that nearby pixels in the image are more likely to be of a similar colour. All in all, the target distribution is defined on $[0, 1]^N$ as

$$\pi(x) \propto \exp\left(-\frac{1}{2}\left\{\sigma^{-2}\sum_{i=1}^{N}(x[i]-y[i])^2 + \gamma^2 \sum_{i\sim j}(x[i]-x[j])^2\right\}\right),$$

with σ^2 parametrizing the scale of the observation noise, and γ^2 parametrizing the similarity in colour between nearby pixels. The neighborhood structure, denoted via $i \sim j$, can e.g. give to a pixel i four neighbours, corresponding to

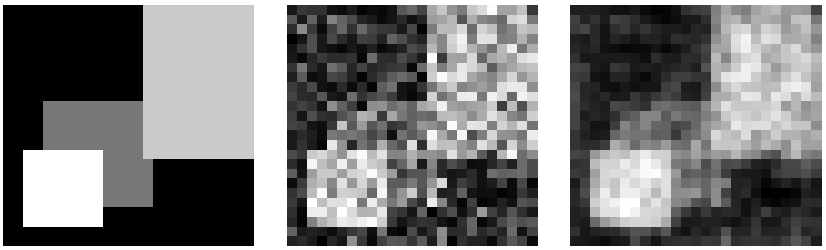

Figure 3.6: Image, noisy image, restored image

the pixels above, below, on the left and on the right to i; the boundaries must be handled also, somehow.

Gibbs [2004] considers a Gibbs sampler for this target, noting that all the full conditionals are truncated Normal distributions:

$$\begin{aligned}
&\pi(x[i]|x[\backslash i]) \propto \mathcal{N}_{[0,1]}(x[i], \mu(x[\backslash i]), \sigma^2(x[\backslash i])), \\
&\text{with} \quad \mu(x[\backslash i]) = (\sigma^{-2} + n_i\gamma^2)^{-1} \left(\sigma^{-2} y[i] + \gamma^2 \sum_{j \sim i} x[j] \right), \\
&\qquad \sigma^2(x[\backslash i]) = (\sigma^{-2} + n_i\gamma^2)^{-1},
\end{aligned}$$

where n_i is the number of neighbours of i, and $\sum_{j\sim i} x[j]$ the sum of the values $x[j]$ of neighbours j of i. The Normal distribution is truncated on the interval $[0, 1]$. This truncated distribution can be sampled from via inverse transform sampling. See Figure 3.6.

Next we follow Lemma 3.2 in Gibbs [2004]. We consider two chains, (X_t^{high}) and (X_t^{low}) started respectively from $x_0^{high} = 1$ in all entries and $x_0^{low} = 0$ in all entries. The chains are propagated using common random numbers. At some step in the algorithm, consider the full conditional for an entry i: the two chains will share the same variance $\sigma^2(x^{high}[\backslash i]) = \sigma^2(x^{low}[\backslash i] = (\sigma^{-2} + n_i\gamma^2)^{-1}$ but the means will be respectively

$$\mu(x^{high}[\backslash i]) = (\sigma^{-2} + n_i\gamma^2)^{-1} \left(\sigma^{-2} y[i] + \gamma^2 \sum_{j \sim i} x^{high}[j] \right),$$

$$\mu(x^{low}[\backslash i]) = (\sigma^{-2} + n_i\gamma^2)^{-1} \left(\sigma^{-2} y[i] + \gamma^2 \sum_{j \sim i} x^{low}[j] \right),$$

and thus $\sum_{j\sim i} x^{high}[j] > \sum_{j\sim i} x^{low}[j]$ implies $\mu(x^{high}[\backslash i]) > \mu(x^{low}[\backslash i])$. This implies a similar ordering for the tail probabilities of the two truncated Normal distributions, and consequently, using the same uniform in inverse transform sampling, the next draws $x^{high}[i], x^{low}[i]$ would be similarly ordered. Note that

this sampling strategy corresponds to an optimal transport coupling of the full conditional distributions.

Furthermore, the expected distance between the two chains can be shown to decrease on average, as follows. Introduce $d(x^1, x^2) = \sum_{i=1}^N n_i |x^1[i] - x^2[i]|$, the pixelwise distance weighted by the number of neighbours. By updating one arbitrary pixel i and leaving the others unchanged, using inverse transform sampling with common random numbers for both chains, calculations lead to

$$\mathbb{E}[d(X_t^{high}, X_t^{low}) | X_{t-1}^{high}, X_{t-1}^{low}] \leq \frac{N - 1 + \frac{n_{\max}\gamma^2}{\sigma^{-2} + n_{\max}\gamma^2}}{N} d(X_t^{high}, X_t^{low}),$$

where $n_{\max}$ is the maximum number of neighbours of a pixel. The contraction coefficient is strictly less than one; denoting it by c we have

$$\mathbb{E}[d(X_t^{high}, X_t^{low}) | X_0^{high} = 1, X_0^{low} = 0] \leq c^t \sum_{i=1}^N n_i.$$

We cannot directly conclude that the rate of convergence of $W_1(\pi_t, \pi)$ to zero is geometric, because the above reasoning implies to specific chain, one started at $X_0^{high} = 1$ and one started at $X_0^{low} = 0$, neither of which is started at stationarity. We can however write for an arbitrary x_0,

$$\begin{aligned} W_1(K^t(x_0 \to \cdot), \pi) &= W_1(K^t(x_0 \to \cdot), \int K^t(x \to \cdot)\pi(dx)) \\ &\leq \int W_1(K^t(x_0 \to \cdot), \int K^t(x \to \cdot))\pi(dx), \end{aligned}$$

from the triangular inequality. It is not clear to me why $W_1(K^t(x_0 \to \cdot), \int K^t(x \to \cdot))$ would be less than $W_1(K^t(x^{high} \to \cdot), \int K^t(x^{low} \to \cdot))$ as argued in Gibbs [2004], but I might be missing something. In any case by triangular inequality,

$$\begin{aligned} W_1(K^t(x_0 \to \cdot), \int K^t(x \to \cdot)) \leq W_1(K^t(x_0 \to \cdot), \int K^t(x^{high} \to \cdot)) \\ + W_1(K^t(x \to \cdot), \int K^t(x^{high} \to \cdot)), \end{aligned}$$

and by the coupling representation of the Wasserstein distance, and the monotonocity of the coupling constructed above, it is less than twice $c^t \sum_{i=1}^N n_i$. This gives a geometric rate of convergence in Wasserstein for this Gibbs sampler.

3.2.3 Sampling k-subsets

We next consider target distribution that arise in combinatorics and related problems. We follow Guruswami [2000], Section 6.

Let us first consider the problem of sampling uniformly in the set of all subsets of size $k \geq 2$ of a set $[n] = \{0, \ldots, n-1\}$ with $n \geq 2$. Denote the complementary of a set A in $[n]$ by A^c, so $A^c = [n] \setminus A$. Consider the following Markov chain $(X_t)_{t \geq 0}$. Initially we start from $X_0 = \{0, \ldots, k-1\}$. Then, given X_{t-1}, we define X_t as follows:

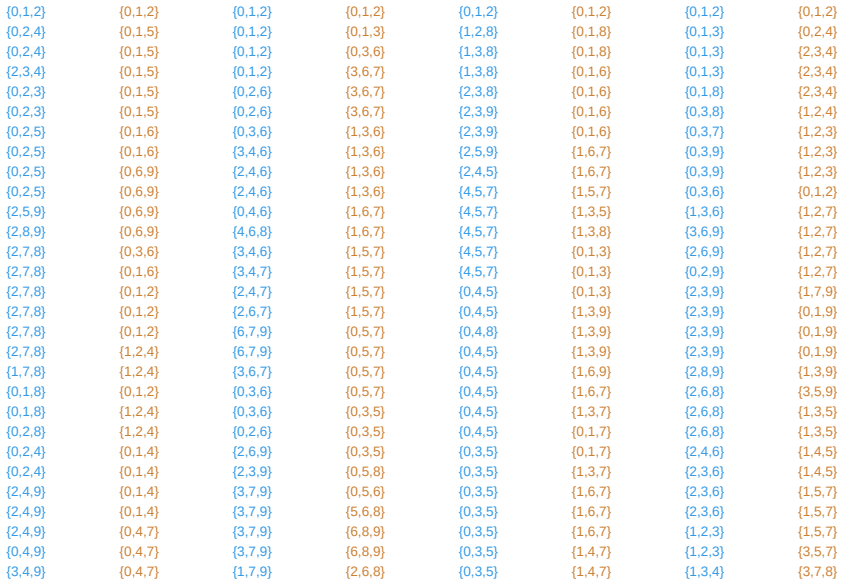

Figure 3.7: Eight independent chains, targeting the uniform distribution on subsets of $\{0, \ldots, 9\}$ of size 3.

- with probability $1/2$, set $X_t = X_{t-1}$,
- with probability $1/2$, draw i uniformly in X_{t-1} and j uniformly in X_{t-1}^c, and define $X_t = X_{t-1} \setminus \{i\} \cup \{j\}$.

This defines a Markov kernel K. Given two distinct subsets of size k, it is reasonably straightforward to imagine a succession of transitions that the chain can undergo from one state to another. This implies that the chain is irreducible. Furthermore, $\mathbb{P}(X_t = X_{t-1}) = 1/2$ and thus the chain is aperiodic. This implies that the chain converges to a stationary distribution. By verifying the detailed balance condition one can check that the stationary distribution is the uniform distribution on subsets of size k, with $\pi(x) = \binom{n}{k}^{-1}$ for all subset x of size k. Figure 3.7 shows a traceplot of eight independent chains, started at $\{0, 1, 2\}$.

Following Theorem 6.2 in Guruswami [2000], consider coupled chains $(X_t, Y_t)_{t \geq 0}$, which, given (X_{t-1}, Y_{t-1}) at time $t-1$, evolves as follows.

- If $X_{t-1} = Y_{t-1}$, sample $X_t \sim K(X_{t-1} \to \cdot)$ and set $Y_t = X_t$.
- Else, with probability $1/2$, set $X_t = X_{t-1}$ and $Y_t = Y_{t-1}$,
- otherwise, set $S = X_{t-1} \setminus Y_{t-1}$ and $T = Y_{t-1} \setminus X_{t-1}$, pick a bijection $g : S \to T$, draw i uniformly in X_{t-1} and j uniformly in X_{t-1}^c, define $X_t = X_{t-1} \setminus \{i\} \cup \{j\}$, and

  - if $i \in X_{t-1} \cap Y_{t-1}$, set $i' = i$, else $i' = g(i)$,
  - if $j \notin Y_{t-1}$, set $j' = j$, else $j' = g^{-1}(j)$.
- and set $Y_t = Y_{t-1} \cup \{j'\} \setminus \{i'\}$.

One way of thinking about this coupled transition is in connection with maximal couplings. A proposal here involves a pair of indices (i, j), uniformly distributed in $X_{t-1} \times X_{t-1}^c$; for the second chain it involves a pair (i', j') in $Y_{t-1} \times Y_{t-1}^c$. We can try maximize the probability that $i = i'$ and the probability that $j = j'$, under the marginal uniformity constraints. It can be seen that the maximal probability of $\{i = i'\}$ is precisely $|X_{t-1} \cap Y_{t-1}|/k$, and is achieved by the above procedure. Likewise the maximal probability of $\{j = j'\}$ is also achieved in the above coupling.

Note on the other hand, that the above coupling is not a maximal coupling of $K(X_{t-1} \to \cdot)$ and $K(Y_{t-1} \to \cdot)$. At least I don't think it is, because if X_{t-1} is one move away from Y_{t-1}, a higher probability of $\{X_t = Y_t\}$ would be obtained by allowing the possibility of updating one of the two states while keeping the other fixed.

Denoting by $X_t \oplus Y_t$ the set made of elements that are in one of the two sets but not in both (i.e. the "symmetric difference" also denoted $X_t \triangle Y_t$), we can show that $q_t = |X_t \oplus Y_t|$ decreases geometrically in expectation, at every step. We do this by considering the multiple possibilities that can occur: the selected i, guaranteed to be in X_{t-1}^c, might be either in Y_{t-1} or in Y_{t-1}^c, the selected j might be either in Y_{t-1} or Y_{t-1}^c, etc. Each case leads to a different change in $q_t = |X_t \oplus Y_t|$.

It's convenient to consider the different cases with numerical examples, e.g. $X = \{1, 2, 3\}$, $Y = \{1, 2, 4\}$, $[n] = \{1, 2, 3, 4, 5, 6\}$, $S = \{3\}$, $T = \{4\}$.

- If $i \in Y$ and $j \in Y$, then $i' = i$ and $j' = g^{-1}(j)$. This leads to i being removed from both X and Y, and j is added to X which decreases q by one, and j' is added to Y which decreases q by one: q is decreased by two. This occurs with probability $|Y \cap X|/k$ times $|X^c \cap Y|/(n-k)$, which is $(k - q/2)/k \times q/2/(n-k)$.

- If $i \notin Y$ and $j \notin Y$, then $i' = g(i)$ and $j' = j$. This leads to different elements being removed, and the same element being added, so q decreases by two. This occurs with probability $|X \cap Y^c|/k \times |X^c \cap Y^c|/(n-k)$ which is $(q/2)/k \times (n - k - q/2)/(n-k)$.

- If $i \notin Y$ and $j \in Y$, then $i' = g(i)$ and $j' = g^{-1}(j)$, and if $j \neq g(i)$, then elements that are not in common are removed in favor of elements that will be common to X and Y, thus reducing the discrepancy q by four. This occurs with probability $(q/2)/k$ times $(q/2 - 1)/(n-k)$.

- If $i \in Y$ and $j \in Y^c$, then $i' = i$ and $j' = j$, then the same element i is removed and the same element j is added, so q is unchanged. Likewise, if $i \notin Y$ and $j \in Y$, then $i' = g(i)$ and $j' = g^{-1}(j)$, and if $j = g(i)$, then

elements (i, j) are swapped in different senses for X and Y and there are no changes in q.

In expectation, we can compute $\mathbb{E}[q_t|q_{t-1}] \leq \left(1 - \frac{1}{k}\right) q_{t-1}$. Now $\mathbb{P}(X_t \neq Y_t) = \mathbb{P}(q_t > 0) \leq \mathbb{E}[q_t] \leq (1 - k^{-1})^t q_0$ and $q_0 \leq 2k$. We get an explicit rate of convergence, with some explicit dependence on k.

3.2.4 Colourings of graphs

Essentially the strategy is described in Jerrum [1998], Section 5.3 for a Gibbs sampler targeting the uniform distribution on the set of colourings of low-degree graphs. With n vertices in the set V, with edge set E, consider a graph G with maximal degree Δ. A (valid) colouring X is an allocation of colors, $X(v)$ in $[q] = \{1, \ldots, q\}$ for each vertex $v \in V$, in such a way that for all $(v, v') \in E$, $X(v) \neq X(v')$. That is, vertices connected by an edge are coloured differently, and there are q colours in total. The problem is to sample uniformly over the set of valid colourings. Here we will assume $q \geq 2\Delta + 1$ for simplicity.

A chain (X_t) might operate as follows, starting from a legal colouring X_0 (obtained... how?).

- Sample a vertex v uniformly in V.
- Sample a colour Z uniformly in legal colours for v, i.e. the set $[q] \setminus X_t(N(v))$ where $N(v)$ is the set of neighbours of v and $X_t(N(v))$ is the set of colours found among these neighbours.
- Set $X_{t+1}(v) = Z$, and do not change the other colours, i.e. $X_{t+1}(v') = X_t(v')$ for all $v' \neq v$.

This corresponds to a Gibbs sampler, since the above procedure samples from the full conditional distribution of $X(v)$. Figure 3.8 shows a "traceplot".

Next consider the following coupling, given X_t, Y_t at step $t \geq 0$.

- Sample v uniformly in V.
- Sample Z, Z' from a maximal coupling of the uniform distributions, respectively on $[q] \setminus X_t(N(v))$ and $[q] \setminus Y_t(N(v))$.
- Set $X_{t+1}(v) = Z$, $Y_{t+1}(v) = Z$, and $X_{t+1}(v') = X_t(v')$, $Y_{t+1}(v') = Y_t(v')$ for all $v' \neq v$.

Therefore, at each step the same vertex is considered for both chains. Denote by D_t the set of vertices v such that $X_t(v) \neq Y_t(v)$ and A_t the set of vertices such that $X_t(v) = Y_t(v)$. Jerrum [1998], Section 5.3, shows that

$$\mathbb{E}[|D_{t+1}||D_t] \leq (1 - a)D_t.$$

for some explicit $a > 0$. This implies that $\mathbb{E}[|D_t|] \leq (1 - a)^t n$, since $|D_0| \leq n$. Therefore $\mathbb{P}(X_t \neq Y_t) = \mathbb{P}(D_t \neq 0) \leq \mathbb{E}|D_t| \leq (1 - a)^t n$, and geometric convergence in total variation is then estalished with a rate $1 - a$. The proof

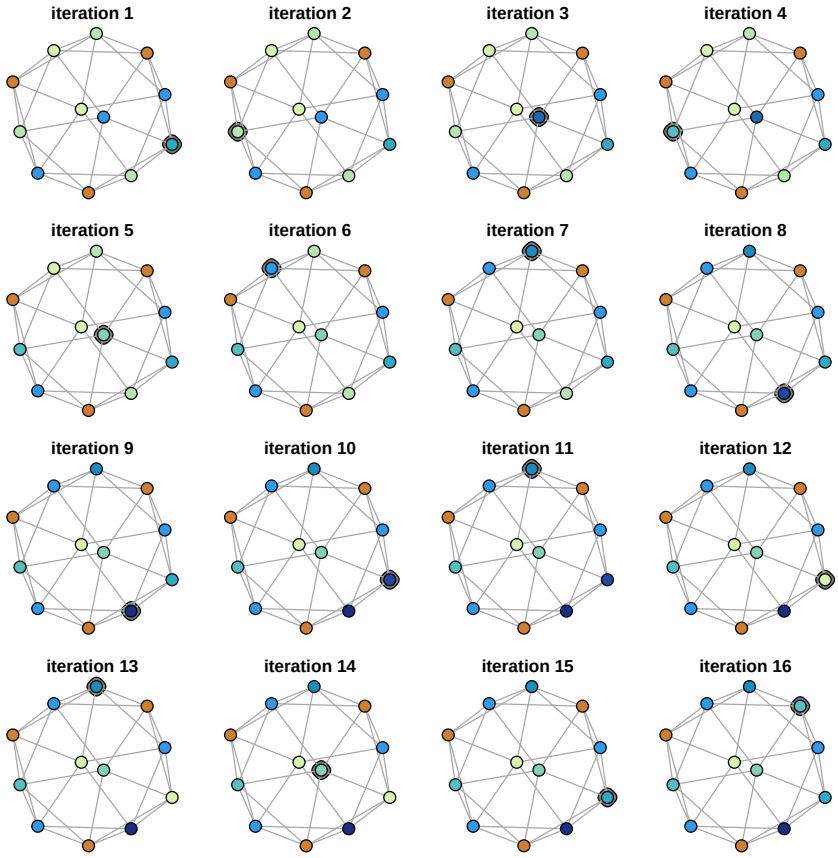

Figure 3.8: Chain targeting the uniform distribution on colourings of a isocahedral graph, with 12 colours.

of the above inequality considers the probabilities $\mathbb{P}(|D_{t+1}| = |D_t| + 1|D_t)$, $\mathbb{P}(|D_{t+1}| = |D_t||D_t)$ and $\mathbb{P}(|D_{t+1}| = |D_t| - 1|D_t)$, corresponding to the three possibilities. Summing up over the possibilities yields the result with the explicit constant $a = (q - 2\Delta)/(n(q - \Delta))$.

3.3 Summary and references

Markov chains generated by (reasonable) MCMC algorithms will converge to their limiting distributions, but it is not obvious to determine how many iterations are enough for chains to be close to stationarity. Closeness can be measured in total variation distance or with any other distance or divergence between probability distributions, such as Wasserstein or Prokhorov. Using coupled Markov chains, one can obtain upper bounds on rates of convergence. These type of bounds are mostly of theoretical interest, but can also provide some practical information regarding how algorithms behave or "scale" according to certain aspects of the targets, e.g. the number of pixels in an image, or the number of covariates in logistic regression.

The arguments sketched above are simple, and certainly do not do justice to the highly sophisticated works that have been done to analyze many MCMC algorithms. These simple arguments hopefully help understanding at least where the "geometric" in geometric ergodicity comes from. In part, it comes from geometric tails of return times to some nice, central sets of the state space $\mathbb{X}$, on which particular properties of the chains might hold; this can be established using drift conditions. Further, on such nice sets a pair of chains might meet with some minimum probability, leading to the time to first meeting being analogous to a Geometric variable.

Couplings introduced in various theoretical studies might be implementable. For example, couplings based on common random numbers or on maximal couplings can be implemented, possibly using some of the techniques described in the previous chapter. In the next chapter we will see that such couplings can then be employed to derive practical rates of convergence, as well as techniques to remove the effect of the starting distribution from MCMC estimators; we will refer to those as "debiasing" techniques.

Some review articles on the topic of convergence of MCMC include Jerrum [1998], Jones and Hobert [2001], Nummelin [2002], Roberts and Rosenthal [2004]. Books on MCMC typically also contain some material on Markov chains, e.g. Robert and Casella [1999]. Conversely, various books of Markov chains mention connections to MCMC. Meyn and Tweedie [1993] is a very important reference on Markov chains. Douc et al. [2018] provide a more modern treatment, with Chapter 19 specifically on couplings of Markov chains. Levin et al. [2009] and Eberle [2015] also provide instructive accounts, the former with a detailed treatment of Ising models, and the latter with more emphasis on continuous state spaces and log-concave targets. Countless articles are dedicated to the topic of geometric ergodicity results for specific MCMC algorithms.

Bibliography

V. Anil Kumar and H. Ramesh. Markovian coupling vs. conductance for the Jerrum–Sinclair chain. In *40th Annual Symposium on Foundations of Computer Science*, pages 241–251. IEEE, 1999. 5

N. Biswas, P. E. Jacob, and P. Vanetti. Estimating convergence of Markov chains with L-lag couplings. In *Advances in Neural Information Processing Systems 32*, pages 7391–7401. Curran Associates, Inc., 2019. URL http://papers.nips.cc/paper/8958-estimating-convergence-of-markov-chains-with-l-lag-couplings.pdf. 2

N. Bou-Rabee, A. Eberle, and R. Zimmer. Coupling and convergence for Hamiltonian Monte Carlo. *arXiv preprint arXiv:1805.00452*, 2018. 14

S. P. Brooks. Mcmc convergence diagnosis via multivariate bounds on log-concave densities. *The Annals of Statistics*, 26(1):398–433, 1998. 13

R. Burton and Y. Kovchegov. Mixing times via super-fast coupling. *arXiv preprint math/0609568*, 2006. 5

H. M. Choi and J. P. Hobert. The Polya-Gamma Gibbs sampler for Bayesian logistic regression is uniformly ergodic. *Electronic Journal of Statistics*, 7: 2054–2064, 2013. 7

M. K. Cowles and J. S. Rosenthal. A simulation approach to convergence rates for Markov chain Monte Carlo algorithms. *Statistics and Computing*, 8(2): 115–124, 1998. 2

A. S. Dalalyan. Theoretical guarantees for approximate sampling from smooth and log-concave densities. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 79(3):651–676, 2017. 13

L. Devroye. A note on generating random variables with log-concave densities. *Statistics & Probability Letters*, 82(5):1035–1039, 2012. 13

P. Diaconis and D. Freedman. Iterated random functions. *SIAM review*, 41(1): 45–76, 1999. 5, 8, 9, 10

R. Douc, E. Moulines, P. Priouret, and P. Soulier. *Markov chains*. Springer, 2018. 5, 21

A. Durmus and É. Moulines. Quantitative bounds of convergence for geometrically ergodic Markov chain in the Wasserstein distance with application to the Metropolis Adjusted Langevin Algorithm. *Statistics and Computing*, 25 (1):5–19, 2015. 5, 11

R. Dwivedi, Y. Chen, M. J. Wainwright, and B. Yu. Log-concave sampling: Metropolis-hastings algorithms are fast. *Journal of Machine Learning Research*, 20(183):1–42, 2019. 13

A. Eberle. Markov processes, 2015. 21

A. Frieze, R. Kannan, and N. Polson. Sampling from log-concave distributions. *The Annals of Applied Probability*, pages 812–837, 1994. 13

A. L. Gibbs. Convergence in the Wasserstein metric for Markov chain Monte Carlo algorithms with applications to image restoration. *Stochastic Models*, 20(4):473–492, 2004. doi: 10.1081/STM-200033117. URL https://doi.org/10.1081/STM-200033117. 5, 7, 14, 15, 16

S. Goldstein. Maximal coupling. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, 46(2):193–204, 1979. 5

D. Görür and Y. W. Teh. Concave-convex adaptive rejection sampling. *Journal of Computational and Graphical Statistics*, 20(3):670–691, 2011. 13

D. Griffeath. A maximal coupling for Markov chains. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, 31(2):95–106, 1975. 5

V. Guruswami. Rapidly mixing Markov chains: A comparison of techniques. *Available: cs. washington. edu/homes/venkat/pubs/papers. html*, 2000. 16, 17

T. P. Hayes and E. Vigoda. A non-Markovian coupling for randomly sampling colorings. In *44th Annual IEEE Symposium on Foundations of Computer Science, 2003. Proceedings.*, pages 618–627. IEEE, 2003. 5

J. Heng and P. E. Jacob. Unbiased Hamiltonian Monte Carlo with couplings. *Biometrika*, 106(2):287–302, 2019. 14

E. P. Hsu and K.-T. Sturm. Maximal coupling of Euclidean Brownian motions. *Communications in Mathematics and Statistics*, 1(1):93–104, 2013. 5

S. F. Jarner and E. Hansen. Geometric ergodicity of Metropolis algorithms. *Stochastic processes and their applications*, 85(2):341–361, 2000. 13

D. Jerison. *The Drift and Minorization Method for Reversible Markov Chains*. PhD thesis, Stanford University, 2016. 11

M. Jerrum. Mathematical foundations of the Markov chain Monte Carlo method. In *Probabilistic methods for algorithmic discrete mathematics*, pages 116–165. Springer, 1998. 19, 21

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. 2

G. L. Jones and J. P. Hobert. Honest exploration of intractable probability distributions via Markov chain Monte Carlo. *Statistical Science*, pages 312–334, 2001. 2, 11, 21

Y. Kifer. *Ergodic theory of random transformations*, volume 10. Birkhauser, 1986. 8

D. A. Levin, Y. Peres, and E. L. Wilmer. Markov chains and mixing times. *American Mathematical Soc., Providence*, 2009. 4, 7, 21

T. Lindvall. *Lectures on the coupling method*. Dover Books on Mathematics, 2002. 4, 5

O. Mangoubi and A. Smith. Rapid mixing of Hamiltonian Monte Carlo on strongly log-concave distributions. *arXiv preprint arXiv:1708.07114*, 2017. 5, 13

G. Marsaglia and W. W. Tsang. The ziggurat method for generating random variables. *Journal of statistical software*, 5(8):1–7, 2000. 8

K. L. Mengersen and R. L. Tweedie. Rates of convergence of the Hastings and Metropolis algorithms. *The annals of Statistics*, 24(1):101–121, 1996. 13

S. P. Meyn and R. L. Tweedie. *Markov chains and stochastic stability*. Springer Science & Business Media, 1993. 21

E. Nummelin. MC's for MCMC'ists. *International Statistical Review*, 70(2): 215–240, 2002. 21

N. G. Polson, J. G. Scott, and J. Windle. Bayesian inference for logistic models using pólya–gamma latent variables. *Journal of the American statistical Association*, 108(504):1339–1349, 2013. 7

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

Q. Qin and J. P. Hobert. Geometric convergence bounds for Markov chains in Wasserstein distance based on generalized drift and contraction conditions. *arXiv preprint arXiv:1902.02964*, 2019. 5, 11

C. Robert and G. Casella. *Monte Carlo statistical methods*. Springer Science & Business Media, 1999. 21

G. O. Roberts and J. S. Rosenthal. General state space markov chains and MCMC algorithms. *Probability surveys*, 1:20–71, 2004. 2, 4, 7, 11, 12, 13, 21

J. Rosenthal. Quantitative convergence rates of Markov chains: A simple account. *Electronic Communications in Probability*, 7:123–128, 2002. 11, 13

D. Steinsaltz. Locally contractive iterated function systems. *Annals of Probability*, pages 1952–1979, 1999. 8, 9

L. Tierney. Markov chains for exploring posterior distributions. *The Annals of Statistics*, pages 1701–1728, 1994. 7

P. Wild and W. Gilks. Algorithm as 287: Adaptive rejection sampling from log-concave density functions. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 42(4):701–709, 1993. 13