Probability Theory: The Coupling Method

Frank den Hollander

Mathematical Institute, Leiden University,
P.O. Box 9512, 2300 RA Leiden, The Netherlands

email: *denholla@math.leidenuniv.nl*

First draft: June 2010, LaTeX-file prepared by H. Nooitgedagt.
Second draft: September 2012, figures prepared by A. Troiani.
Third draft: December 2012.

ABSTRACT

Coupling is a powerful method in probability theory through which random variables can be compared with each other. Coupling has been applied in a broad variety of contexts, e.g. to prove limit theorems, to derive inequalities, or to obtain approximations.

The present course is intended for master students and PhD students. A basic knowledge of probability theory is required, as well as some familiarity with measure theory. The course first explains what coupling is and what general framework it fits into. After that a number of applications are described. These applications illustrate the power of coupling and at the same time serve as a guided tour through some key areas of modern probability theory. Examples include: random walks, card shuffling, Poisson approximation, Markov chains, correlation inequalities, percolation, interacting particle systems, and diffusions.

PRELUDE 1: A game with random digits.

Draw 100 digits randomly and independently from the set of numbers $\{1, 2, \ldots, 9, 0\}$. Consider two players who each do the following:

1. Randomly choose one of the first 10 digits.
2. Move forward as many digits as the number that is hit (move forward 10 digits when a 0 is hit).
3. Repeat.
4. Stop when the next move goes beyond digit 100.
5. Record the last digit that is hit.

It turns out that the probability that the two players record the *same* last digit is approximately 0.974.

Why is this probability so close to 1? What if N digits are drawn randomly instead of 100 digits? Can you find a formula for the probability that the two players record the same last digit before moving beyond digit N?

PRELUDE 2: A game with two biased coins.

You are given two coins with success probabilities $p, p' \in (0, 1)$ satisfying $p < p'$ (head = success = 1; tail = failure = 0). Clearly, it is less likely for the p-coin to be successful than for the p'-coin. However, if you throw the two coins independently, then it may happen that the p-coin is successful while the p'-coin is not. Can you throw the two coins *together* in such a way that the outcome is always ordered?

The answer is yes! Let $p'' = (p' - p)/(1 - p) \in (0, 1)$. Take a third coin with success probability p''. Throw the p-coin and the p''-coin independently. Let X be the outcome of the p-coin and X'' the outcome of the p''-coin, and put $X' = X \vee X''$. Because $p' = p + (1 - p)p''$, X' has the same distribution as the outcome of the p'-coin (check this statement!). Since $X' \geq X$, you have thus achieved the ordering.

Contents

1 Introduction

In Sections 1.1–1.3 we describe three examples of coupling illustrating both the method and its usefulness. Each of these examples will be worked out in more detail later. The symbol$\mathbb{N}_0$ is used for the set $\mathbb{N} \cup \{0\}$ with $\mathbb{N} = \{1, 2, \ldots\}$. The symbol tv is used for the total variation distance, which is defined at the beginning of Chapter 2. The symbols $\mathbb{P}$ and $\mathbb{E}$ are used to denote probability and expectation.

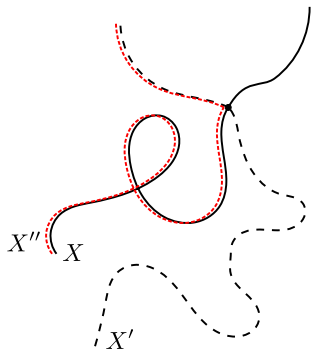
Lindvall [10] explains how coupling was invented in the late 1930's by Wolfgang Doeblin, and provides some historical context. Standard references for coupling are Lindvall [11] and Thorisson [15].

1.1 Markov chains

Let $X = (X_n)_{n\in\mathbb{N}_0}$ be a Markov chain on a countable state space S, with initial distribution $\lambda = (\lambda_i)_{i\in S}$ and transition matrix $P = (P_{ij})_{i,j\in S}$. If X is *irreducible*, *aperiodic* and *positive recurrent*, then it has a unique stationary distribution π solving the equation $\pi = \pi P$, and

$$\lim_{n\to\infty} \lambda P^n = \pi \qquad \text{componentwise on } S. \tag{1.1}$$

This is the standard *Markov Chain Convergence Theorem* (MCCT) (see e.g. Häggström [5], Chapter 5, or Kraaikamp [7], Section 2.2).



A coupling proof of (1.1) goes as follows. Let $X' = (X'_n)_{n\in\mathbb{N}_0}$ be an independent copy of the same Markov chain, but starting from π. Since $\pi P^n = \pi$ for all n, X' is stationary. Run X and X' together, and let

$$T = \inf\{k \in \mathbb{N}_0 \colon\ X_k = X'_k\}$$

be their *first meeting time*. Note that T is a *stopping time*, i.e., for each $n \in \mathbb{N}_0$ the event $\{T = n\}$ is an element of the sigma-algebra generated by $(X_k)_{0\le k\le n}$ and $(X'_k)_{0\le k\le n}$. For $n \in \mathbb{N}_0$, define

$$X''_n = \begin{cases} X_n, & \text{if } n < T, \\ X'_n, & \text{if } n \ge T. \end{cases}$$

By the *strong Markov property* (which says that, for any stopping time T, $(X_k)_{k>T}$ depends on $(X_k)_{k\le T}$ only through X_T), we have that $X'' = (X''_n)_{n\in\mathbb{N}_0}$ is a copy of X. Now write, for

$i \in S$,

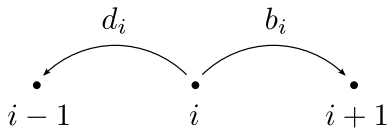
$$\begin{aligned}(\lambda P^n)_i - \pi_i &= \mathbb{P}(X''_n = i) - \mathbb{P}(X'_n = i) \\ &= \mathbb{P}(X''_n = i, T \le n) + \mathbb{P}(X''_n = i, T > n) \\ &\qquad -\mathbb{P}(X'_n = i, T \le n) - \mathbb{P}(X'_n = i, T > n) \\ &= \mathbb{P}(X''_n = i, T > n) - \mathbb{P}(X'_n = i, T > n),\end{aligned}$$

where we use $\mathbb{P}$ as the generic symbol for probability (in later Chapters we will be more careful with the notation). Hence

$$\begin{aligned}\|\lambda P^n - \pi\|_{tv} &= \sum_{i \in S} |(\lambda P^n)_i - \pi_i| \\ &\le \sum_{i \in S} \big[\mathbb{P}(X''_n = i, T > n) + \mathbb{P}(X'_n = i, T > n)\big] = 2\mathbb{P}(T > n).\end{aligned}$$

The left-hand side is the *total variation norm* of $\lambda P^n - \pi$. The conditions in the MCCT guarantee that $\mathbb{P}(T < \infty) = 1$ (as will be explained in Chapter 6). The latter is expressed by saying that *the coupling is successful*. Hence the claim in (1.1) follows by letting $n \to \infty$.

1.2 Birth-Death processes



Let $X = (X_t)_{t \ge 0}$, be the Markov process with state space $\mathbb{N}_0$, birth rates $b = (b_i)_{i \in \mathbb{N}_0}$, death rates $d = (d_i)_{i \in \mathbb{N}_0}$ ($d_0 = 0$), and initial distribution $\lambda = (\lambda_i)_{i \in \mathbb{N}_0}$. Suppose that b and d are such that X is *recurrent* (see Kraaikamp [7], Section 3.6, for conditions on b and d that guarantee recurrence). Let $X' = (X'_t)_{t \ge 0}$ be an independent copy of the same Markovv process, but starting from a different initial distribution $\mu = (\mu_i)_{i \in \mathbb{N}_0}$. Run X and X' together, and let

$$T = \inf\{t \ge 0 \colon\ X_t = X'_t\}$$

be the first time X and X' meet each other.

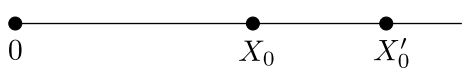
For $t \ge 0$, define

$$X''_t = \begin{cases} X_t, & \text{if } t < T, \\ X'_t, & \text{if } t \ge T. \end{cases}$$

The same argument as in Section 1.1 gives

$$\|\lambda P_t - \mu P_t\|_{tv} \le 2\mathbb{P}(T > t),$$

where P_t is the transition matrix at time t, i.e., $(\lambda P_t)_i = \mathbb{P}(X_t = i)$, $i \in \mathbb{N}_0$. Since transitions can occur between neighboring elements of $\mathbb{N}_0$ only, X and X' *cannot cross without meeting*.



Hence we have

$$T \le \max\{\tau_0, \tau'_0\}$$

with

$$\tau_0 = \{t \geq 0 \colon\ X_t = 0\}, \qquad \tau_0' = \{t \geq 0 \colon\ X_t' = 0\},$$

the first hitting times of 0 for X and X', respectively. By the assumption of recurrence, we have $\mathbb{P}(\tau_0 < \infty) = \mathbb{P}(\tau_0' < \infty) = 1$. This in turn implies that $\mathbb{P}(T < \infty) = 1$, i.e., the coupling is successful, and so we get

$$\lim_{t\to\infty} \|\lambda P_t - \mu P_t\|_{tv} = 0.$$

If X is *positive recurrent* (see Kraaikamp [7], Section 3.6, for conditions on b and d that guarantee positive recurrence), then X has a unique stationary distribution π, solving the equation $\pi P_t = \pi$ for all $t \geq 0$. In that case, by picking $\mu = \pi$ we get

$$\lim_{t\to\infty} \|\lambda P_t - \pi\|_{tv} = 0. \tag{1.2}$$

Remark: The fact that transitions can occur between neighboring elements of $\mathbb{N}_0$ only allows us to deduce, straightaway from the recurrence property, that the coupling is successful. In Section 1.2 this argument was not available, and we had to defer this part of the proof to Chapter 6. In Chapter 6 we will show that the coupling is successful under the stronger assumption of positive recurrence.

1.3 Poisson approximation

Let Y_m, $m = 1, \ldots, n$, be independent $\{0, 1\}$-valued random variables with

$$\mathbb{P}(Y_m = 1) = p_m, \qquad m = 1, \ldots, n,$$

and put $X = \sum_{m=1}^n Y_m$. If all the p_m's are small, then X is *approximately Poisson distributed* with parameter $\sum_{m=1}^n p_m$ (see Rice [13], Section 2.1.5). How good is this approximation?

For $\lambda > 0$, define

$$p_\lambda(i) = e^{-\lambda}\frac{\lambda^i}{i!}, \quad i \in \mathbb{N}_0,$$

which is the Poisson distribution with parameter λ, abbreviated as POISSON(λ). Let X' have distribution p_λ with $\lambda = \sum_{m=1}^n p_m$. Then, for $i \in \mathbb{N}_0$,

$$\begin{aligned}
\mathbb{P}(X = i) - p_\lambda(i) &= \mathbb{P}(X = i) - \mathbb{P}(X' = i) \\
&= \mathbb{P}(X = i, X = X') + \mathbb{P}(X = i, X \neq X') \\
&\qquad -\mathbb{P}(X' = i, X = X') - \mathbb{P}(X' = i, X \neq X') \\
&= \mathbb{P}(X = i, X \neq X') - \mathbb{P}(X' = i, X \neq X')
\end{aligned}$$

and hence

$$\|\mathbb{P}(X \in \cdot) - p_\lambda(\cdot)\|_{tv} \leq 2\mathbb{P}(X \neq X'). \tag{1.3}$$

Thus, in order to get a good approximation it suffices to find a coupling of X and X' that makes them equal with high probability. Choosing them independently will not do. Here is how we proceed.

Let (Y_m, Y_m'), $m = 1, \ldots, n$, be independent $\{0, 1\} \times \mathbb{N}_0$-valued random variables with distribution

$$\mathbb{P}\big((Y_m, Y_m') = (i, i')\big) = \begin{cases} 1 - p_m, & \text{if } i = 0,\ i' = 0, \\ e^{-p_m} - (1 - p_m), & \text{if } i = 1,\ i' = 0, \\ 0, & \text{if } i = 0, i' \in \mathbb{N}, \\ e^{-p_m}\frac{p_m^{i'}}{i'!}, & \text{if } i = 1, i' \in \mathbb{N}, \end{cases} \qquad m = 1, \ldots, n.$$

By summing out over i', respectively, i we see that

$$\mathbb{P}(Y_m = i) = \begin{cases} 1 - p_m, & \text{if } i = 0, \\ p_m, & \text{if } i = 1, \end{cases} \qquad \mathbb{P}(Y'_m = i') = e^{-p_m} \frac{p_m^{i'}}{i'!}, \quad i' \in \mathbb{N}_0,$$

so that the marginal distributions are indeed correct and we have a proper coupling. Now estimate

$$\begin{aligned} \mathbb{P}(X \neq X') &= \mathbb{P}\left(\sum_{m=1}^{n} Y_m \neq \sum_{m=1}^{n} Y'_m \right) \\ &\leq \mathbb{P}\big(\exists\, m \in \{1, \ldots, n\} \colon\, Y_m \neq Y'_m\big) \\ &\leq \sum_{m=1}^{n} \mathbb{P}(Y_m \neq Y'_m) \\ &= \sum_{m=1}^{n} \left[e^{-p_m} - (1 - p_m) + \sum_{i'=2}^{\infty} e^{-p_m} \frac{p_m^{i'}}{i'!} \right] \\ &= \sum_{m=1}^{n} p_m (1 - e^{-p_m}) \\ &\leq \sum_{m=1}^{n} p_m^2. \end{aligned}$$

Hence, for $\lambda = \sum_{m=1}^{n} p_m$, we have proved that

$$\|\mathbb{P}(X \in \cdot) - p_\lambda(\cdot)\|_{tv} \leq 2\lambda M$$

with $M = \max_{m=1,\ldots,n} p_m$. This quantifies the extent to which the approximation is good when M is small. Both λ and M will in general depend on n. Typical applications will have λ of order 1 and M tending to zero as $n \to \infty$.

Remark: The coupling produced above will turn out to be the best possible: it is a *maximal coupling* (see Chapter 2.5). The crux is that $(Y_m, Y'_m) = (0, 0)$ and $(1, 1)$ are given the largest possible probabilities. More details will be given in Chapter 5.

2 Basic theory of coupling

Chapters 2 and 7 provide the theoretical basis for the theory of coupling and consequently are technical in nature. It is here that we arm ourselves with a number of basic facts about coupling that are needed to deal with the applications described in Chapters 3–6 and Chapters 8–10. In Section 2.1 we give the definition of a coupling of two probability measures, in Section 2.2 we state and derive the basic coupling inequality, bounding the total variation distance between two probability measures in terms of their coupling time, in Section 2.3 we look at bounds on the coupling time of two random sequences, in Section 2.4 we introduce the notion of distributional coupling, while in Section 2.5 we prove the existence of a maximal coupling for which the coupling inequality is optimal.

In what follows we use some elementary ingredients from measure theory, for which we refer the reader to standard textbooks.

Definition 2.1 *Given a bounded signed measure $\mathbb{M}$ on a measurable space $(E, \mathcal{E})$ such that $\mathbb{M}(E) = 0$, the total variation norm of $\mathbb{M}$ is defined as*

$$\|\mathbb{M}\|_{tv} = 2 \sup_{A \in \mathcal{E}} \mathbb{M}(A).$$

Remark: The total variation norm of $\mathbb{M}$ is defined as

$$\|\mathbb{M}\|_{tv} = \sup_{\|f\|_\infty \leq 1} \left| \int_E f \, d\mathbb{M} \right|,$$

where the supremum runs over all functions $f \colon E \to \mathbb{R}$ that are bounded and measurable w.r.t. $\mathcal{E}$, and $\|f\|_\infty = \sup_{x \in E} |f(x)|$ is the supremum norm. By the Jordan-Hahn decomposition theorem, there exists a set $D \in \mathcal{E}$ such that $\mathbb{M}^+(\cdot) = \mathbb{M}(\cdot \cap D)$ and $\mathbb{M}^-(\cdot) = -\mathbb{M}(\cdot \cap D^c)$ are both non-negative measures on $(E, \mathcal{E})$. Clearly, $\mathbb{M} = \mathbb{M}^+ - \mathbb{M}^-$ and $\sup_{A \in \mathcal{E}} \mathbb{M}(A) = \mathbb{M}(D) = \mathbb{M}^+(E)$. It therefore follows that $\|\mathbb{M}\|_{tv} = \int_E (1_D - 1_{D^c}) \, d\mathbb{M} = \mathbb{M}^+(E) + \mathbb{M}^-(E)$ (note that the absolute value sign disappears). If $\mathbb{M}(E) = 0$, then $\mathbb{M}^+(E) = \mathbb{M}^-(E)$, in which case $\|\mathbb{M}\|_{tv} = 2\mathbb{M}^+(E) = 2 \sup_{A \in \mathcal{E}} \mathbb{M}(A)$.

2.1 Definition of coupling

A *probability space* is a triple $(E, \mathcal{E}, \mathbb{P})$, with $(E, \mathcal{E})$ a measurable space consisting of a *sample space* E and a σ-algebra $\mathcal{E}$ of subsets of E, and with $\mathbb{P}$ a probability measure on $\mathcal{E}$. Typically, E is a *Polish space* (i.e., complete, separable and metric) and $\mathcal{E}$ consists of its *Borel sets*.

Definition 2.2 *A coupling of two probability measures $\mathbb{P}$ and $\mathbb{P}'$ on the same measurable space $(E, \mathcal{E})$ is any (!) probability measure $\hat{\mathbb{P}}$ on the product measurable space $(E \times E, \mathcal{E} \otimes \mathcal{E})$ (where $\mathcal{E} \otimes \mathcal{E}$ is the smallest sigma-algebra containing $\mathcal{E} \times \mathcal{E}$) whose marginals are $\mathbb{P}$ and $\mathbb{P}'$, i.e.,*

$$\mathbb{P} = \hat{\mathbb{P}} \circ \pi^{-1}, \qquad \mathbb{P}' = \hat{\mathbb{P}} \circ \pi'^{-1},$$

where π is the left-projection and π' is the right-projection, defined by

$$\pi(x, x') = x, \qquad \pi'(x, x') = x', \qquad (x, x') \in E \times E.$$

A similar definition holds for random variables. Given a probability space $(\Omega, \mathcal{F}, \mathbb{Q})$, a *random variable* X is a measurable mapping from $(\Omega, \mathcal{F})$ to $(E, \mathcal{E})$. The image of $\mathbb{Q}$ under X is $\mathbb{P}$, the probability measure of X on $(E, \mathcal{E})$. When we are interested in X only, we may forget about $(\Omega, \mathcal{F}, \mathbb{Q})$ and work with $(E, \mathcal{E}, \mathbb{P})$ only.

Definition 2.3 *A coupling of two random variable X and X' taking values in $(E, \mathcal{E})$ is any (!) pair of random variables $(\hat{X}, \hat{X}')$ taking values in $(E \times E, \mathcal{E} \otimes \mathcal{E})$ whose marginals have the same distribution as X and X', i.e.,*

$$\hat{X} \stackrel{D}{=} X, \qquad \hat{X}' \stackrel{D}{=} X',$$

with $\stackrel{D}{=}$ denoting equality in distribution.

Remark: The law $\hat{\mathbb{P}}$ of $(\hat{X}, \hat{X}')$ is a coupling of the laws $\mathbb{P}, \mathbb{P}'$ of X, X' in the sense of Definition 2.2.

Remark: Couplings are not unique. Two trivial examples are:

$$\begin{array}{lcl} \hat{\mathbb{P}} = \mathbb{P} \times \mathbb{P}' \text{ with } \mathbb{P}, \mathbb{P}' \text{ arbitrary} & \Longleftrightarrow & \hat{X}, \hat{X}' \text{ are independent,} \\ \mathbb{P} = \mathbb{P}' \text{ and } \hat{\mathbb{P}} \text{ lives on the diagonal} & \Longleftrightarrow & \hat{X} = \hat{X}'. \end{array}$$

Non-trivial examples were given in Chapter 1.

In applications the challenge is to find a coupling that makes $\|\mathbb{P} - \mathbb{P}'\|_{tv}$ as small as possible. For this reason *coupling is an art, not a recipe*. We will see plenty of examples as we go along.

2.2 Coupling inequalities

2.2.1 Random variables

The basic coupling inequality for two random variables X, X' with probability distributions $\mathbb{P}, \mathbb{P}'$ reads as follows:

Theorem 2.4 *Given two random variables X, X' with probability distributions $\mathbb{P}, \mathbb{P}'$, any (!) coupling $\hat{\mathbb{P}}$ of $\mathbb{P}, \mathbb{P}'$ satisfies*

$$\|\mathbb{P} - \mathbb{P}'\|_{tv} \leq 2\hat{\mathbb{P}}(\hat{X} \neq \hat{X}').$$

Proof. Pick any $A \in \mathcal{E}$ and write

$$\begin{array}{rcl} \mathbb{P}(X \in A) - \mathbb{P}'(X' \in A) & = & \hat{\mathbb{P}}(\hat{X} \in A) - \hat{\mathbb{P}}(\hat{X}' \in A) \\ & = & \hat{\mathbb{P}}(\hat{X} \in A, \hat{X} = \hat{X}') + \hat{\mathbb{P}}(\hat{X} \in A, \hat{X} \neq \hat{X}') \\ & & \quad -\hat{\mathbb{P}}(\hat{X}' \in A, \hat{X} = \hat{X}') - \hat{\mathbb{P}}(\hat{X}' \in A, \hat{X} \neq \hat{X}') \\ & = & \hat{\mathbb{P}}(\hat{X} \in A, \hat{X} \neq \hat{X}') - \hat{\mathbb{P}}(\hat{X}' \in A, \hat{X} \neq \hat{X}'). \end{array}$$

Hence, by Definition 2.1 (where we write $\mathbb{P}(A) = \mathbb{P}(X \in A)$),

$$\begin{array}{rcl} \|\mathbb{P} - \mathbb{P}'\|_{tv} & = & 2 \sup_{A \in \mathcal{E}} [\mathbb{P}(A) - \mathbb{P}'(A)] \\ & = & 2 \sup_{A \in \mathcal{E}} [\mathbb{P}(X \in A) - \mathbb{P}'(X' \in A)] \\ & \leq & 2 \sup_{A \in \mathcal{E}} \hat{\mathbb{P}}(\hat{X} \in A, \hat{X} \neq \hat{X}') \\ & = & 2\hat{\mathbb{P}}(\hat{X} \neq \hat{X}'), \end{array}$$

where the last equality holds because the supremum is achieved at $A = E \in \mathcal{E}$. ∎

Exercise 2.5 *Let U, V be random variables on $\mathbb{N}_0$ with probability mass functions*

$$f_U(x) = \tfrac{1}{2}\,1_{\{0,1\}}(x), \qquad f_V(x) = \tfrac{1}{3}\,1_{\{0,1,2\}}(x), \qquad x \in \mathbb{N}_0,$$

where 1_S is the indicator function of the set S. (a) Compute the total variation distance. (b) Give two different couplings of U and V. (c) Give a coupling of U and V under which $\{U \geq V\}$ with probability 1.

Exercise 2.6 *Let U, V be random variables on $[0, \infty)$ with probability density functions*

$$f_U(x) = 2e^{-2x}, \qquad f_V(x) = e^{-x}, \qquad x \in [0, \infty).$$

Answer the same questions as in Exercise 2.5.

2.2.2 Sequences of random variables

There is also a version of the coupling inequality for *sequences of random variables*. Let $X = (X_n)_{n\in\mathbb{N}_0}$ and $X' = (X'_n)_{n\in\mathbb{N}_0}$ be two sequences of random variables taking values in $(E^{\mathbb{N}_0}, \mathcal{E}^{\otimes\mathbb{N}_0})$. Let $(\hat{X}, \hat{X}')$ be a coupling of X and X'. Define

$$T = \inf\{n \in \mathbb{N}_0 \colon\ \hat{X}_m = \hat{X}'_m \text{ for all } m \geq n\},$$

which is the *coupling time* of $\hat{X}$ and $\hat{X}'$, i.e., the first time from which the two sequences agree onwards (possibly $T = \infty$).

Theorem 2.7 *For two sequences of random variables $X = (X_n)_{n\in\mathbb{N}_0}$ and $X' = (X'_n)_{n\in\mathbb{N}_0}$ taking values in $(E^{\mathbb{N}_0}, \mathcal{E}^{\otimes\mathbb{N}_0})$, let $(\hat{X}, \hat{X}')$ be a coupling of X and X', and let T be the coupling time. Then*

$$\|\mathbb{P}(X_n \in \cdot) - \mathbb{P}'(X_n \in \cdot)\|_{tv} \leq 2\hat{\mathbb{P}}(T > n).$$

Proof. This follows from Theorem 2.4 because $\{\hat{X}_n \neq \hat{X}'_n\} \subseteq \{T > n\}$. ∎

Remark: In Section 1.3 we already saw an example of sequence coupling, namely, X and X' were two copies of a Birth-Death process starting from different initial distributions. The Markov property implies that T is equal in distribution to the first time $\hat{X}$ and $\hat{X}'$ meet each other.

A stronger form of sequence coupling can be obtained by introducing the *left-shift* θ on $E^{\mathbb{N}_0}$, defined by

$$\theta(x_0, x_1, \ldots) = (x_1, x_1, \ldots),$$

i.e., θ drops the first element of the sequence.

Theorem 2.8 *Let X, X' and T be defined as in Theorem 2.7. Then*

$$\|\mathbb{P}(\theta^n X \in \cdot) - \mathbb{P}'(\theta^n X' \in \cdot)\|_{tv} \leq 2\hat{\mathbb{P}}(T > n).$$

Proof. Because

$$\{\hat{X}_m \neq \hat{X}'_m \text{ for some } m \geq n\} \subseteq \{T > n\},$$

the claim again follows from Theorem 2.4. ∎

Remark: Similar inequalities hold for continuous-time random processes $X = (X_t)_{t\geq 0}$ and $X' = (X'_t)_{t\geq 0}$.

2.2.3 Mappings

Since total variation distance never increases under a mapping, we have the following corollary.

Corollary 2.9 *Let ψ be a measurable map from $(E, \mathcal{E})$ to $(E^*, \mathcal{E}^*)$. Let $\mathbb{Q} = \mathbb{P} \circ \psi^{-1}$ and $\mathbb{Q}' = \mathbb{P}' \circ \psi^{-1}$ (i.e., $\mathbb{Q}(B) = \mathbb{P}(\psi^{-1}(B))$ and $\mathbb{Q}'(B) = \mathbb{P}'(\psi^{-1}(B))$ for $B \in \mathcal{E}^*$). Then*

$$\|\mathbb{Q} - \mathbb{Q}'\|_{tv} \leq \|\mathbb{P} - \mathbb{P}'\|_{tv} \leq 2\hat{\mathbb{P}}(\hat{X} \neq \hat{X}').$$

Proof. Simply estimate

$$\begin{aligned}
\|\mathbb{Q} - \mathbb{Q}'\|_{tv} &= 2 \sup_{B \in \mathcal{E}^*} [\mathbb{Q}(B) - \mathbb{Q}'(B)] \\
&= 2 \sup_{B \in \mathcal{E}^*} [\mathbb{P}(\psi(X) \in B) - \mathbb{P}'(\psi(X') \in B)] \\
&\leq 2 \sup_{A \in \mathcal{E}} [\mathbb{P}(X \in A) - \mathbb{P}'(X' \in A)] \qquad (A = \psi^{-1}(B)) \\
&= \|\mathbb{P} - \mathbb{P}'\|_{tv},
\end{aligned}$$

where the inequality comes from the fact that $\mathcal{E}$ may be larger than $\psi^{-1}(\mathcal{E}^*)$. Use Theorem 2.4 to get the bound. ∎

2.3 Rates of convergence

Suppose that we have some control on the moments of the coupling time T, e.g. for some $\phi\colon\ \mathbb{N}_0 \to [0, \infty)$ non-decreasing with $\lim_{n\to\infty} \phi(n) = \infty$ we know that

$$\hat{\mathbb{E}}(\phi(T)) < \infty.$$

Theorem 2.10 *Let X, X' and ϕ be as above. Then*

$$\|\mathbb{P}(\theta^n X \in \cdot) - \mathbb{P}'(\theta^n X' \in \cdot)\|_{tv} = o\big(1/\phi(n)\big) \text{ as } n \to \infty.$$

Proof. Estimate

$$\phi(n)\, \hat{\mathbb{P}}(T > n) \leq \hat{\mathbb{E}}\big(\phi(T) 1_{\{T>n\}}\big).$$

Note that the right-hand side tends to zero as $n \to \infty$ by dominated convergence, because $\hat{\mathbb{E}}(\phi(T)) < \infty$. Use Theorem 2.8 to get the claim. ∎

Typical examples are:

$$\begin{aligned}
&\phi(n) = n^\alpha, \quad \alpha > 0 \text{ (polynomial rate)}, \\
&\phi(n) = e^{\beta n}, \quad \beta > 0 \text{ (exponential rate)}.
\end{aligned}$$

For instance, for finite-state irreducible aperiodic Markov chains, there exists an $M < \infty$ such that $\hat{\mathbb{P}}(T > 2M \mid T > M) \leq \frac{1}{2}$ (see Häggström [5], Chapter 5), which implies that there exists a $\beta > 0$ such that $\hat{\mathbb{E}}(e^{\beta T}) < \infty$. In Section 3 we will see that for random walks we typically have $\hat{\mathbb{E}}(T^\alpha) < \infty$ for all $0 < \alpha < \frac{1}{2}$.

2.4 Distributional coupling

Suppose that a coupling $(\hat{X}, \hat{X}')$ of two random sequences $X = (X_n)_{n\in\mathbb{N}_0}$ and $X' = (X'_n)_{n\in\mathbb{N}_0}$ comes with two random times T and T' such that not only

$$\hat{X} \stackrel{D}{=} X, \qquad \hat{X}' \stackrel{D}{=} X',$$

but also

$$\big(\theta^T \hat{X}, T\big) \stackrel{D}{=} \big(\theta^{T'} \hat{X}', T'\big).$$

Here we compare the two sequences shifted over *different* random times, rather than the same random time.

Theorem 2.11 *Let X, X', T, T' be as above. Then*

$$\|\mathbb{P}(\theta^n X \in \cdot) - \mathbb{P}'(\theta^n X' \in \cdot)\|_{tv} \le 2\hat{\mathbb{P}}(T > n)\Big[= 2\hat{\mathbb{P}}(T' > n)\Big].$$

Proof. Write, for $A \in \mathcal{E}^{\otimes\mathbb{N}_0}$,

$$\begin{aligned}
\hat{\mathbb{P}}(\theta^n \hat{X} \in A, T \le n) &= \sum_{m=0}^{n} \hat{\mathbb{P}}(\theta^{n-m}(\theta^m \hat{X}) \in A, T = m) \\
&= \sum_{m=0}^{n} \hat{\mathbb{P}}(\theta^{n-m}(\theta^m \hat{X}') \in A, T' = m) \\
&= \hat{\mathbb{P}}(\theta^n \hat{X}' \in A, T' \le n).
\end{aligned}$$

It follows that

$$\begin{aligned}
\hat{\mathbb{P}}(\theta^n \hat{X} \in A) - \hat{\mathbb{P}}(\theta^n \hat{X}' \in A) &= \hat{\mathbb{P}}(\theta^n \hat{X} \in A, T > n) - \hat{\mathbb{P}}(\theta^n \hat{X}' \in A, T' > n) \\
&\le \hat{\mathbb{P}}(T > n),
\end{aligned}$$

and hence

$$\begin{aligned}
\|\mathbb{P}(\theta^n X \in \cdot) - \mathbb{P}'(\theta^n X' \in \cdot)\|_{tv} &= 2 \sup_{A\in\mathcal{E}^{\otimes\mathbb{N}_0}} [\mathbb{P}(\theta^n X \in A) - \mathbb{P}'(\theta^n X' \in A)] \\
&= 2 \sup_{A\in\mathcal{E}^{\otimes\mathbb{N}_0}} [\hat{\mathbb{P}}(\theta^n \hat{X} \in A) - \hat{\mathbb{P}}(\theta^n \hat{X}' \in A)] \\
&\le 2\hat{\mathbb{P}}(T > n).
\end{aligned}$$

∎

Remark: A restrictive feature of distributional coupling is that $T \stackrel{D}{=} T'$, i.e., the two random times must have the same distribution. Therefore distributional coupling is more of a theoretical than a practical tool. We will see in Chapter 4 that it plays a role in card shuffling.

Remark: In Section 6.2 we will encounter yet another form of coupling, called *shift-coupling*. This requires the existence of random times T, T' such that

$$\theta^T X = \theta^{T'} X'$$

(which is stronger than $\theta^T X \stackrel{D}{=} \theta^{T'} X'$), but does not require that $T \stackrel{D}{=} T'$. This form of coupling is useful for dealing with time averages. Thorisson [15] contains a critical analysis of how different forms of coupling compare with each other.

2.5 Maximal coupling

Does there exist a "best possible" coupling, one that gives the sharpest estimate on the total variation distance, in the sense that the inequality in Theorem 2.4 becomes an equality? The answer is yes!

Theorem 2.12 *For any two probability measures $\mathbb{P}$ and $\mathbb{P}'$ on a measurable space $(E, \mathcal{E})$ there exists a coupling $\hat{\mathbb{P}}$ such that*

(i) $\|\mathbb{P} - \mathbb{P}'\|_{tv} = 2\hat{\mathbb{P}}(\hat{X} \neq \hat{X}')$.

(ii) $\hat{X}$ *and* $\hat{X}'$ *are independent conditional on* $\{\hat{X} \neq \hat{X}'\}$, *provided the latter event has positive probability.*

Proof. We give an abstract construction of a maximal coupling. Let $\Delta = \{(x, x)\colon\ x \in E\}$ be the diagonal of $E \times E$. Let $\psi\colon\ E \to E \times E$ be the map defined by $\psi(x) = (x, x)$.

Exercise 2.13 *Show that ψ is measurable because E is a Polish.*

Put

$$\lambda = \mathbb{P} + \mathbb{P}', \qquad g = \frac{d\mathbb{P}}{d\lambda}, \quad g' = \frac{d\mathbb{P}'}{d\lambda},$$

and note that g and g' are well defined because $\mathbb{P}$ and $\mathbb{P}'$ are both absolutely continuous w.r.t. λ. Define $\mathbb{Q}$ on $(E, \mathcal{E})$ and $\hat{\mathbb{Q}}$ on $(E \times E, \mathcal{E} \otimes \mathcal{E})$ by

$$\frac{d\mathbb{Q}}{d\lambda} = g \wedge g', \qquad \hat{\mathbb{Q}} = \mathbb{Q} \circ \psi^{-1}.$$

(Both are sub-probability measures.) Then $\hat{\mathbb{Q}}$ puts all its mass on Δ. Call this mass $\gamma = \hat{\mathbb{Q}}(\Delta)$, and put

$$\nu = \mathbb{P} - \mathbb{Q}, \quad \nu' = \mathbb{P}' - \mathbb{Q}, \quad \hat{\mathbb{P}} = \frac{\nu \times \nu'}{1 - \gamma} + \hat{\mathbb{Q}}.$$

Then

$$\hat{\mathbb{P}}(A \times E) = \frac{\nu(A) \times \nu'(E)}{1 - \gamma} + \hat{\mathbb{Q}}(A \times E) = \mathbb{P}(A),$$

because $\nu(A) = \mathbb{P}(A) - \mathbb{Q}(A)$, $\nu'(E) = \mathbb{P}'(E) - \mathbb{Q}(E) = 1 - \gamma$ and $\hat{\mathbb{Q}}(A \times E) = \mathbb{Q}(A)$. Similarly, $\hat{\mathbb{P}}(E \times A) = \mathbb{P}'(A)$, so that the marginals are indeed correct and we have a proper coupling.

To get (i), compute

$$\begin{aligned} \|\mathbb{P} - \mathbb{P}'\|_{tv} &= \int_E |g - g'|\, d\lambda = 2\left[1 - \int_E (g \wedge g')\, d\lambda\right] \\ &= 2\,[1 - \mathbb{Q}(E)] = 2(1 - \gamma) = 2\hat{\mathbb{P}}(\Delta^c) = 2\hat{\mathbb{P}}(\hat{X} \neq \hat{X}'). \end{aligned}$$

Here, the first equality uses the Jordan-Hahn decomposition of signed measures into a difference of non-negative measures, the second equality uses the identity $|g - g'| = g + g' - 2(g \wedge g')$, the third equality uses the definition of $\mathbb{Q}$, the fourth equality uses that $\mathbb{Q}(E) = \hat{\mathbb{Q}}(\Delta) = \gamma$, the fifth equality uses that $\hat{\mathbb{Q}}(\Delta^c) = 0$ and $(\nu \times \nu')(\Delta^c) = \nu(E)\nu'(E) = (1 - \gamma)^2$, while the sixth equality uses the definition of Δ.

Exercise 2.14 *Prove the first equality. Hint: use a splitting as in the remark below Definition 2.1 with $\mathbb{M} = \mathbb{P} - \mathbb{P}'$ and $D = \{x \in E\colon\ g(x) \geq g'(x)\}$.*

To get (ii), note that

$$\hat{\mathbb{P}}(\,\cdot \mid \hat{X} \neq \hat{X}') = \hat{\mathbb{P}}(\,\cdot \mid \Delta^c) = \left(\frac{\nu}{1-\gamma} \times \frac{\nu'}{1-\gamma}\right)(\cdot).$$

∎

Remark: What Theorem 2.12 says is that we can *in principle* find a coupling that gives the correct value for the total variation. Such a coupling is called a *maximal coupling*. However, *in practice* it is often difficult to write out this coupling explicitly (the above is only an abstract construction), and we have to content ourselves with good estimates or approximations. We will encounter examples in Chapter 9.

Exercise 2.15 *Give a maximal coupling of U and V in Exercise* 2.5.

Exercise 2.16 *Give a maximal coupling of U and V in Exercise* 2.6.

Exercise 2.17 *Is the coupling of the two coins in* PRELUDE 2 *a maximal coupling?*

3 Random walks

Random walks on $\mathbb{Z}^d$, $d \geq 1$, are special cases of Markov chains: the transition probability to go from site x to site y only depends on the difference vector $y-x$. Because of this translation invariance, random walks can be analyzed in great detail. A standard reference is Spitzer [14]. One key fact we will use below is that any irreducible random walk whose step distribution has zero mean and finite variance is recurrent in $d = 1, 2$ and transient in $d \geq 3$. In $d = 1$ any random walk whose step distribution has zero mean and finite first moment is recurrent.

In Section 3.1 we look at random walks in dimension 1, in Section 3.2 at random walks in dimension d. In Section 3.3 we use random walks in dimension d to show that bounded harmonic functions on $\mathbb{Z}^d$ are constant. This result has an interesting interpretation in physics: a system in thermal equilibrium has a constant temperature.

3.1 Random walks in dimension 1

3.1.1 Simple random walk

Let $S = (S_n)_{n\in\mathbb{N}_0}$ be a *simple random walk* on $\mathbb{Z}$ starting at 0, i.e., $S_0 = 0$ and $S_n = \sum_{i=1}^n Y_i$, $n \in \mathbb{N}$, where $Y = (Y_i)_{i\in\mathbb{N}}$ are i.i.d. with

$$\mathbb{P}(Y_i = -1) = \mathbb{P}(Y_i = 1) = \tfrac{1}{2}.$$

The following theorem says that, modulo period 2, the distribution of S_n becomes *flat* for large n.

Theorem 3.1 *Let S be a simple random walk. Then, for every $k \in \mathbb{Z}$ even,*

$$\lim_{n\to\infty} \|\mathbb{P}(S_n \in \cdot\,) - \mathbb{P}(S_n + k \in \cdot\,)\|_{tv} = 0.$$

Proof. Let S' denote an independent copy of S starting at $S'_0 = k$. Write $\hat{\mathbb{P}}$ for the joint probability distribution of (S, S'), and let

$$T = \min\{n \in \mathbb{N}_0 \colon\ S_n = S'_n\}.$$

Then

$$\|\mathbb{P}(S_n \in \cdot\,) - \mathbb{P}(S_n + k \in \cdot\,)\|_{tv} = \|\mathbb{P}(S_n \in \cdot\,) - \mathbb{P}(S'_n \in \cdot\,)\|_{tv} \leq 2\hat{\mathbb{P}}(T > n).$$

Now, $\tilde{S} = (\tilde{S}_n)_{n\in\mathbb{N}_0}$ defined by $\tilde{S}_n = S'_n - S_n$ is a random walk on $\mathbb{Z}$ starting at $\tilde{S}_0 = k$ with i.i.d. increments $\tilde{Y} = (\tilde{Y}_i)_{i\in\mathbb{N}}$ given by

$$\tilde{\mathbb{P}}(Y_i = -2) = \tilde{\mathbb{P}}(Y_i = 2) = \tfrac{1}{4}, \qquad \tilde{\mathbb{P}}(Y_i = 0) = \tfrac{1}{2}.$$

This is a simple random walk on $2\mathbb{Z}$ with a "random time delay", namely, it steps only half of the time. Since

$$T = \tilde{\tau}_0 = \{n \in \mathbb{N}_0 \colon\ \tilde{S}_n = 0\}$$

and k is even, it follows from the recurrence of $\tilde{S}$ that $\hat{\mathbb{P}}(T < \infty) = 1$. Let $n \to \infty$ to get the claim. ∎

In analytical terms, Theorem 3.1 says the following. Let $p(\cdot,\cdot)$ denote the transition kernel of the simple random walk, let $p^n(\cdot,\cdot)$, $n \in \mathbb{N}$, denote the n-fold composition of $p(\cdot,\cdot)$, and

let $\delta_k(\cdot)$, $k \in \mathbb{Z}$, denote the vector whose components are 1 at k and 0 elsewhere. Then Theorem 3.1 says that for k even

$$\lim_{n\to\infty} \|\delta_k p^n(\cdot) - \delta_0 p^n(\cdot)\|_{tv} = 0.$$

(This short-hand notation comes from the fact that $\delta_k p^n(\cdot) = \mathbb{P}(S_n \in \cdot \mid S_0 = k)$.) It is possible to prove the latter statement by hand, i.e., by computing $\delta_k p^n(\cdot)$ and $\delta_0 p^n(\cdot)$, evaluating their total variation distance and letting $n \to \infty$. However, this computation turns out to be somewhat involved.

Exercise 3.2 *Do the computation. Hint: Use the formula*

$$\delta_k p^n(l) = (\tfrac{1}{2})^n \binom{n}{\frac{1}{2}(n + |k-l|)}, \qquad k, l \in \mathbb{Z},\ n + |k-l| \text{ even}.$$

The result in Theorem 3.1 cannot be extended to k odd. In fact, because the simple random walk has period 2, the laws of S_n and $S_n + k$ have disjoint support when k is odd, irrespective of n, and so

$$\|\mathbb{P}(S_n \in \cdot\,) - \mathbb{P}(S_n + k \in \cdot\,)\|_{tv} = 2 \qquad \forall\, n \in \mathbb{N}_0,\ k \in \mathbb{Z} \text{ odd}.$$

3.1.2 Beyond simple random walk

Does the same result as in Theorem 3.1 hold for random walks other than the simple random walk? Yes, it does! To formulate the appropriate statement, let S be the random walk on $\mathbb{Z}$ with i.i.d. increments Y satisfying the *aperiodicity condition*

$$\gcd\big\{z' - z\colon\ z, z' \in \mathbb{Z},\ \mathbb{P}(Y_1 = z)\mathbb{P}(Y_1 = z') > 0\big\} = 1. \tag{3.1}$$

Theorem 3.3 *Subject to* (3.1)*,*

$$\lim_{n\to\infty} \|\mathbb{P}(S_n \in \cdot\,) - \mathbb{P}(S_n + k \in \cdot\,)\|_{tv} = 0 \qquad \forall\, k \in \mathbb{Z}.$$

Proof. We try to use the same coupling as in the proof of Theorem 3.1. Namely, we put $\tilde{S}_n = S'_n - S_n$, $n \in \mathbb{N}_0$, we note that $\tilde{S} = (\tilde{S}_n)_{n\in\mathbb{N}_0}$ is a random walk starting at $\tilde{S}_0 = k$ whose i.i.d. increments $\tilde{Y} = (\tilde{Y}_i)_{i\in\mathbb{N}}$ are given by

$$\tilde{\mathbb{P}}(\tilde{Y}_1 = \tilde{z}) = \sum_{\substack{z, z' \in \mathbb{Z} \\ z' - z = \tilde{z}}} \mathbb{P}(Y_1 = z)\mathbb{P}(Y_1 = z'), \qquad \tilde{z} \in \mathbb{Z},$$

we further note that (3.1) written in terms of $\tilde{\mathbb{P}}$ transforms into

$$\gcd\{\tilde{z} \in \mathbb{Z}\colon\ \tilde{\mathbb{P}}(\tilde{Y}_1 = \tilde{z}) > 0\} = 1, \tag{3.2}$$

so that $\tilde{S}$ is an *aperiodic* random walk, and finally we argue that $\tilde{S}$ is recurrent, i.e.,

$$\tilde{\mathbb{P}}(\tilde{\tau}_0 < \infty) = 1,$$

to complete the proof. However, there is a problem: *recurrence may fail*! Indeed, even though $\tilde{S}$ is a symmetric random walk (because $\tilde{\mathbb{P}}(\tilde{Y}_1 = \tilde{z}) = \tilde{\mathbb{P}}(\tilde{Y}_1 = -\tilde{z})$, $\tilde{z} \in \mathbb{Z}$), the distribution of

$\tilde{Y}_1$ may have a *thick tail* resulting in $\tilde{E}(|\tilde{Y}_1|) = \infty$, in which case $\tilde{S}$ is not necessarily recurrent (see Spitzer [14], Section 3).

The lack of recurrence may be circumvented by slightly *adapting* the coupling. Namely, instead of letting the two copies of the random walk S and S' step independently, we let them make *independent small steps*, but *dependent large steps*. Formally, we let Y'' be an independent copy of Y, and we define Y' by putting

$$Y_i' = \begin{cases} Y_i'' & \text{if } |Y_i - Y_i''| \le N, \\ Y_i & \text{if } |Y_i - Y_i''| > N, \end{cases} \tag{3.3}$$

i.e., S' copies the jumps of S'' when they differ from the jumps of S by at most N, otherwise S' copies the jumps of S. The value of $N \in \mathbb{N}$ is arbitrary and will later be taken large enough.

First, we check that S' is a copy of S. This is so because, for every $z \in \mathbb{Z}$,

$$\begin{aligned} \mathbb{P}'(Y_1' = z) &= \hat{\mathbb{P}}(Y_1' = z, |Y_1 - Y_1''| \le N) + \hat{\mathbb{P}}(Y_1' = z, |Y_1 - Y_1''| > N) \\ &= \hat{\mathbb{P}}(Y_1'' = z, |Y_1 - Y_1''| \le N) + \hat{\mathbb{P}}(Y_1 = z, |Y_1 - Y_1''| > N), \end{aligned}$$

and the first term in the right-hand side equals $\hat{\mathbb{P}}(Y_1 = z, |Y_1 - Y_1''| \le N)$ by symmetry (use that Y and Y'' are independent), so that we get $\mathbb{P}'(Y_1' = z) = \mathbb{P}(Y_1 = z)$.

Next, we note from (3.3) that the difference random walk $\tilde{S} = S - S'$ has increments

$$\tilde{Y}_i = Y_i' - Y_i = \begin{cases} Y_i'' - Y_i & \text{if } |Y_i - Y_i''| \le N, \\ 0 & \text{if } |Y_i - Y_i''| > N, \end{cases}$$

i.e., no jumps larger than N can occur. Moreover, by picking N large enough we also have that

$$\tilde{\mathbb{P}}(\tilde{Y}_1 \neq 0) > 0 \quad \text{ and (3.2) holds.}$$

Exercise 3.4 *Prove the last two statements.*

Thus, $\tilde{S}$ is an aperiodic symmetric random walk on $\mathbb{Z}$ with bounded step size. Consequently, $\tilde{S}$ is recurrent and therefore we have $\tilde{\mathbb{P}}(\tilde{\tau}_0 < \infty) = 1$, so that the proof of Theorem 3.3 can be completed in the same way as the proof of Theorem 3.1. ∎

Remark: The coupling in (3.3) is called the *Ornstein coupling*. The idea is that S' manages to stay close to S by copying its large jumps.

Remark: Theorem 3.1 may be sharpened by noting that

$$\hat{\mathbb{P}}(T > n) = O\left(\frac{1}{\sqrt{n}}\right).$$

Indeed, this follows from a classical result for random walks in $d = 1$ with zero mean and finite variance, namely $\mathbb{P}(\tau_z > n) = O(\frac{1}{\sqrt{n}})$ for all $z \neq 0$ with τ_z the first hitting time of z (see Spitzer [14], Section 3). Consequently,

$$\|\mathbb{P}(S_n \in \cdot\,) - \mathbb{P}(S_n + k \in \cdot\,)\|_{tv} = O\left(\frac{1}{\sqrt{n}}\right) \qquad \forall\, k \in \mathbb{Z} \text{ even.}$$

A direct proof of this estimate without coupling turns out to be rather hard, especially for an arbitrary random walk in $d = 1$ with zero mean and finite variance. Even a well-trained analyst typically does not manage to cook up a proof in a day! Exercise 3.2 shows how to proceed for simple random walk.

Exercise 3.5 *Show that, without* (3.1)*, Theorem* 3.3 *holds if and only if k is a multiple of the period.*

3.2 Random walks in dimension d

Question: What about random walks on $\mathbb{Z}^d$, $d \geq 2$? We know that an arbitrary irreducible random walk in $d \geq 3$ is transient, and so the Ornstein coupling does not work to bring the two coupled random walks together with probability 1.

Answer: It still works, provided we do the Ornstein coupling componentwise.

3.2.1 Simple random walk

Here is how the componentwise coupling works. We first consider a simple random walk on $\mathbb{Z}^d$, $d \geq 2$. Pick direction 1, i.e., look at the x_1-coordinate of the random walks S and S', and couple these as follows:

$$\begin{aligned} &Y_i \in \{-e_1, e_1\} \Longrightarrow \text{ draw } Y_i' \in \{-e_1, e_1\} \text{ independently with probability } \tfrac{1}{2} \text{ each,} \\ &Y_i \notin \{-e_1, e_1\} \Longrightarrow \text{ put } Y_i' = Y_i. \end{aligned}$$

The difference random walk $\tilde{S} = S' - S$ has increments $\tilde{Y}$ given by

$$\tilde{\mathbb{P}}(\tilde{Y}_i = -2e_1) = \tilde{\mathbb{P}}(\tilde{Y}_i = 2e_1) = \left(\tfrac{1}{2d}\right)^2, \qquad \tilde{\mathbb{P}}(\tilde{Y}_i = 0) = 1 - 2\left(\tfrac{1}{2d}\right)^2.$$

Start at $\tilde{S}_0 = \tilde{z} \in \mathbb{Z}^d$ with all components $\tilde{z}^1, \ldots, \tilde{z}^d$ even, and use that $\tilde{S}$ is recurrent in direction 1, to get that

$$\tau_1 = \inf\{n \in \mathbb{N}_0 \colon\ \tilde{S}_n^1 = 0\}$$

satisfies $\tilde{\mathbb{P}}(\tau_1 < \infty)$. At time τ_1 change the coupling to direction 2, i.e., do the same but now identify the steps in all directions different from 2 and allow for independent steps only in direction 2. Put

$$\tau_2 = \inf\{n \geq \tau_1 \colon\ \tilde{S}_n^2 = 0\}$$

and note that $\tilde{\mathbb{P}}(\tau_2 - \tau_1 < \infty) = 1$. Continue until all d directions are exhausted. At time

$$\tau_d = \inf\{n \geq \tau_{d-1} \colon\ \tilde{S}_n^d = 0\},$$

for which $\tilde{\mathbb{P}}(\tau_d - \tau_{d-1} < \infty) = 1$, the two walks meet. Since $\tilde{\mathbb{P}}(\tau_d < \infty) = 1$, the coupling is successful and the proof is complete.

To get the same result when $\tilde{z}_1 + \cdots + \tilde{z}_d$ is even (rather than all $\tilde{z}^1, \ldots, \tilde{z}^d$ being even), we argue as follows. There is an even number of directions i for which $\tilde{z}^i$ is odd. Pair these directions in an arbitrary manner, say, $(i_1, j_1), \ldots, (i_l, j_l)$ for some $1 \leq l \leq d$. Do a componentwise coupling in the directions (i_1, j_1), i.e., the jumps of S in direction i_1 are independent of the jumps of S' in direction j_1, while the jumps in all directions other than i_1 and j_1 are copied. Wait until $S' - S$ is even in directions i_1 and j_1, switch to the pair (i_2, j_2), etc., until all components of $S' - S$ are even. After that do the componentwise coupling as before.

Exercise 3.6 *Write out the details of the last argument.*

3.2.2 Beyond simple random walk

The general statement is as follows. Suppose that

$$\big\{z' - z \colon\ z, z' \in \mathbb{Z}^d,\, \mathbb{P}(Y_1 = z)\mathbb{P}(Y_1 = z') > 0\big\} \text{ is not contained in any sublattice of } \mathbb{Z}^d, \tag{3.4}$$

which is the analogue of (3.1).

Theorem 3.7 *Subject to* (3.4)*,*

$$\lim_{n\to\infty} \|\mathbb{P}(S_n \in \cdot\,) - \mathbb{P}(S_n + z \in \cdot\,)\|_{tv} = 0 \qquad \forall\, z \in \mathbb{Z}^d.$$

Proof. Combine the componentwise coupling with the "cut out large steps" in the Ornstein coupling (3.3). ∎

Exercise 3.8 *Write out the details of the proof. Warning: The argument is easy when the random walk can move in only one direction at a time (like simple random walk). For other random walks a projection argument is needed.*

Exercise 3.9 *Show that, without* (3.4)*, Theorem* 3.7 *holds if and only if* z *is an element of the minimal sublattice containing* $\{z' - z\colon\ z, z' \in \mathbb{Z}^d,\ \mathbb{P}(Y_1 = z)\mathbb{P}(Y_1 = z') > 0\}$*.*

3.3 Random walks and the discrete Laplacian

The result in Theorem 3.7 has an interesting corollary. Let Δ denote the *discrete Laplacian* acting on functions $f\colon\ \mathbb{Z}^d \to \mathbb{R}$ as

$$(\Delta f)(x) = \frac{1}{2d} \sum_{\substack{y\in\mathbb{Z}^d \\ \|y-x\|=1}} [f(y) - f(x)], \qquad x \in \mathbb{Z}^d.$$

A function f is called *harmonic* when $\Delta f \equiv 0$, i.e., f is at every site equal to the average of its values at neighboring sites.

Theorem 3.10 *All bounded harmonic functions on* $\mathbb{Z}^d$ *are constant.*

Proof. Let S be a simple random walk starting at 0. Then, by the harmonic property of f, we have

$$\mathbb{E}(f(S_n)) = \mathbb{E}\big(\mathbb{E}(f(S_n) \mid S_{n-1})\big) = \mathbb{E}(f(S_{n-1})),$$

where we use that $\mathbb{E}(f(S_n) \mid S_{n-1} = x) = f(x) + (\Delta f)(x) = f(x)$. Iteration gives $\mathbb{E}(f(S_n)) = f(0)$. Now pick any $x, y \in \mathbb{Z}^d$ such that all components of $x - y$ are even, and estimate

$$\begin{aligned}
|f(x) - f(y)| &= |\mathbb{E}(f(S_n + x)) - \mathbb{E}(f(S_n + y))| \\
&= \Big|\sum_{z\in\mathbb{Z}^d} [f(z + x) - f(z + y)]\mathbb{P}(S_n = z)\Big| \\
&= \Big|\sum_{z\in\mathbb{Z}^d} f(z)\big[\mathbb{P}(S_n = z - x) - \mathbb{P}(S_n = z - y)\big]\Big| \\
&\le M \sum_{z\in\mathbb{Z}^d} |\mathbb{P}(S_n + x = z) - \mathbb{P}(S_n + y = z)| \\
&= M\|\mathbb{P}(S_n + x \in \cdot\,) - \mathbb{P}(S_n + y \in \cdot\,)\|_{tv}
\end{aligned}$$

with $M = \sup_{z\in\mathbb{Z}^d} |f(z)| < \infty$. Let $n \to \infty$ and use Theorem 3.7 to get $f(x) = f(y)$. Extend this equality to $x, y \in \mathbb{Z}^d$ with $\|x - y\|$ even by first doing the coupling in paired directions, as in Section 3.2. Hence we conclude that f is constant on the even and on the odd sublattice of $\mathbb{Z}^d$, say, $f \equiv c_{even}$ and $f \equiv c_{odd}$. But $c_{odd} = \mathbb{E}(f(S_1)) = f(0) = c_{even}$, and so f is constant. ∎

Remark: Theorem 3.10 has an interesting interpretation. Simple random walk can be used to describe the *flow of heat* in a physical system. Space is discretized to $\mathbb{Z}^d$ and time is discretized to $\mathbb{N}_0$. Each site has a temperature that evolves with time according to the Laplace operator. Indeed, if $x \mapsto f(x)$ is the temperature profile at time n, then

$$x \mapsto \frac{1}{2d} \sum_{\substack{y \in \mathbb{Z}^d \\ \|y-x\|=1}} f(y) = f(x) + (\Delta f)(x)$$

is the temperature profile at time $n+1$: heat flows to neighboring sites proportionally to temperature differences. A temperature profile that is *in equilibrium* must therefore be harmonic, i.e., $\Delta f \equiv 0$. Theorem 3.10 shows that on $\mathbb{Z}^d$ the only temperature profile in equilibrium that is bounded is the one where the temperature is constant.

Exercise 3.11 *Give an example of an unbounded harmonic function on* $\mathbb{Z}$.

4 Card shuffling

Card shuffling is a topic that combines coupling, algebra and combinatorics. Diaconis [3] gives key ideas. Levin, Peres and Wilmer [8] provides a broad panorama on mixing poperties of Markov chains, with Chapter 8 devoted to card shuffling. Two examples of random shuffles are described in the MSc-thesis by H. Nooitgedagt [12].

In Section 4.1 we present a general theory of random shuffles. In Section 4.2 we look at a specific random shuffle, called the "top-to-random shuffle", for which we carry out explicit computations.

4.1 Random shuffles

Consider a deck with $N \in \mathbb{N}$ cards, labeled $1, \ldots, N$. An *arrangement* of the deck is an element of the set $\mathcal{P}_N$ of permutations of $(1, \ldots, N)$. We think of the first coordinate in the permutation as the label of the "top card" and the last coordinate as the label of the "bottom card".

Definition 4.1 *A shuffle of the deck is a permutation drawn from $\mathcal{P}_N$ and applied to the deck. A random shuffle is a shuffle drawn according to some probability distribution on $\mathcal{P}_N$.*

Applying *independent* random shuffles to the deck, we get a Markov chain $X = (X_n)_{n\in\mathbb{N}_0}$ on $\mathcal{P}_N$. If each shuffle uses the same probability distribution on $\mathcal{P}_N$, then X is time-homogeneous. In typical cases, X is irreducible and aperiodic, with a unique invariant distribution π that is *uniform* on $\mathcal{P}_N$. (The latter corresponds to a random shuffle that leads to a "random deck" after it is applied many times.) Since $\mathcal{P}_N$ is finite, we know that the distribution of X_n converges to π exponentially fast as $n \to \infty$, i.e.,

$$\|\mathbb{P}(X_n \in \cdot\,) - \pi(\cdot)\|_{tv} \leq e^{-\delta n}$$

for some $\delta = \delta(N) > 0$ and $n \geq n(N, \delta)$.

In what follows we will be interested in establishing a *threshold time*, written t_N, around which the total variation norm drops from being close to 2 to being close to 0, i.e., we want to identify the *time of approach to the invariant distribution* (t_N is also called a "mixing time").

Definition 4.2 *$(t_N)_{N\in\mathbb{N}}$ is called a sequence of threshold times if $\lim_{N\to\infty} t_N = \infty$ and, for all $\epsilon > 0$ small enough,*

$$\begin{aligned}
\lim_{N\to\infty} \inf_{n\leq(1-\epsilon)t_N} \|\mathbb{P}(X_n \in \cdot\,) - \pi(\cdot)\|_{tv} &= 2,\\
\lim_{N\to\infty} \sup_{n\geq(1+\epsilon)t_N} \|\mathbb{P}(X_n \in \cdot\,) - \pi(\cdot)\|_{tv} &= 0.
\end{aligned}$$

It turns out that for card shuffling threshold times typically grow with N in a *polynomial* fashion.

To capture the phenomenon of threshold time, we need the notion of *strong uniform time*.

Definition 4.3 *T is a strong uniform time if the following hold:*

1. *T is a stopping time, i.e., for all $n \in \mathbb{N}_0$ the event $\{T = n\}$ is an element of the σ-algebra $\mathcal{F}_n = \sigma(X_0, X_1, \ldots, X_n)$ containing all events that involve X up to time n.*
2. *$X_T \stackrel{D}{=} \pi$.*

3. X_T and T are independent.

Remark: Think of $T = T_N$ as the *random time* at which the random shuffling of the deck is stopped such that the arrangement of the deck is "completely random" (this is a form of distributional coupling defined in Section 2.4). In typical cases the threshold times $(t_N)_{N\in\mathbb{N}}$ are such that

$$\lim_{N\to\infty} \mathbb{E}(T_N)/t_N = 1, \qquad \lim_{N\to\infty} \mathbb{P}(1-\delta < T_N/t_N < 1+\delta) = 1 \quad \forall\, \delta > 0. \tag{4.1}$$

In Section 4.2 we will construct T_N for a special example of a random shuffle.

Theorem 4.4 *If T is a strong uniform time, then*

$$\|\mathbb{P}(X_n \in \cdot\,) - \pi(\cdot)\|_{tv} \leq 2\mathbb{P}(T > n) \qquad \forall\, n \in \mathbb{N}_0.$$

Proof. By now the intuition behind this inequality should be obvious. For $n \in \mathbb{N}_0$ and $A \subset \mathcal{P}_N$, write

$$\begin{aligned}
\mathbb{P}(X_n \in A, T \leq n) &= \sum_{\sigma\in\mathcal{P}_N} \sum_{i=0}^{n} \mathbb{P}(X_n \in A \mid X_i = \sigma, T = i)\, \mathbb{P}(X_i = \sigma, T = i) \\
&= \sum_{i=0}^{n} \mathbb{P}(T = i) \left[\sum_{\sigma\in\mathcal{P}_N} \mathbb{P}(X_{n-i} \in A \mid X_0 = \sigma)\pi(\sigma) \right] \\
&= \sum_{i=0}^{n} \mathbb{P}(T = i)\ \pi(A) \\
&= \pi(A)\mathbb{P}(T \leq n),
\end{aligned}$$

where the second equality uses that $\mathbb{P}(X_n \in A \mid X_i = \sigma, T = i) = \mathbb{P}(X_{n-i} \in A \mid X_0 = \sigma)$ by the strong Markov property of X, and $\mathbb{P}(X_i = \sigma, T = i) = \mathbb{P}(X_i = \sigma \mid T = i)\mathbb{P}(T = i) = \pi(\sigma)\mathbb{P}(T = i)$ by Definition 4.3, while the third equality holds because π is the invariant distribution. Hence

$$\begin{aligned}
\mathbb{P}(X_n \in A) - \pi(A) &= \mathbb{P}(X_n \in A, T > n) - \pi(A)\mathbb{P}(T > n) \\
&= \big[\mathbb{P}(X_n \in A \mid T > n) - \pi(A)\big]\, \mathbb{P}(T > n),
\end{aligned}$$

from which the claim follows after taking the supremum over A. ∎

Remark: Note that T really is the coupling time to a parallel deck that starts in π, even though this deck is not made explicit.

4.2 Top-to-random shuffle

We will next focus on a particular random shuffle, namely, take the top card and insert it randomly back into the deck, i.e., with probability $1/N$ put it at each of the N possible locations, including the top itself. This is called "top-to-random shuffle".

Theorem 4.5 *For the top-to-random shuffle the sequence* $(t_N)_{N\in\mathbb{N}}$ *with* $t_N = N\log N$ *is a sequence of threshold times.*

Proof. Let $T = \tau_* + 1$, with

$$\tau_* = \text{ the first time that the original bottom card comes on top.}$$

Exercise 4.6 *Show that* T *is a strong uniform time. Hint: The* $+1$ *represents the insertion of the original bottom card at a random position in the deck after it has come on top.*

For the proof it is convenient to view T differently, namely,

$$T \stackrel{D}{=} V \tag{4.2}$$

with V the number of *random draws with replacement* from an urn with N balls until each ball has been drawn *at least once*. To see why this holds, for $i = 0, 1, \ldots, N$ put

$$\begin{aligned} T_i &= \text{ the first time there are } i \text{ cards below the original bottom card,} \\ V_i &= \text{ the number of draws necessary to draw } i \text{ distinct balls.} \end{aligned}$$

Then

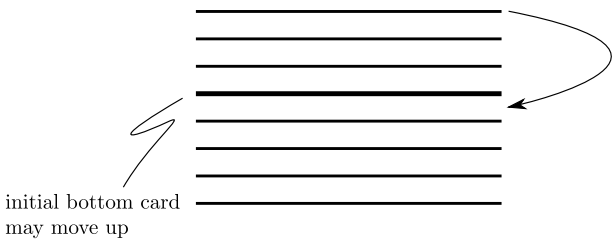
$$T_{i+1} - T_i \stackrel{D}{=} V_{N-i} - V_{N-(i+1)} \stackrel{D}{=} \mathrm{GEO}\left(\frac{i+1}{N}\right),\ i = 0, 1\ldots, N-1, \text{ are independent,} \tag{4.3}$$

where $\mathrm{GEO}(p) = \{p(1-p)^{k-1}\colon\ k \in \mathbb{N}\}$ denotes the geometric distribution with parameter $p \in [0,1]$.

Exercise 4.7 *Prove* (4.3).

Since $T = T_N = \sum_{i=0}^{N-1}(T_{i+1} - T_i)$ and $V = V_N = \sum_{i=0}^{N-1}(V_{N-i} - V_{N-(i+1)})$, (4.3) proves (4.2).

Label the balls $1, \ldots, N$ and let A_i be the event that ball i is not in the first $(1+\epsilon)N\log N$ draws, $i = 1, \ldots, N$ (for ease of notation we will pretend that this number is integer). Then



$$\begin{aligned} \mathbb{P}\big(T > (1+\epsilon)N\log N\big) &= \mathbb{P}\big(V > (1+\epsilon)N\log N\big) \\ &= P\left(\cup_{i=1}^N A_i\right) \le \sum_{i=1}^N \mathbb{P}(A_i) \\ &= N\left(1 - \frac{1}{N}\right)^{(1+\epsilon)N\log N} \\ &= Ne^{-(1+\epsilon)\log N + O(\frac{\log N}{N})} \sim N^{-\epsilon}, \qquad N \to \infty, \end{aligned}$$

which yields the second line of Definition 4.2 via Theorem 4.4.

To get the first line of Definition 4.2, pick $\delta > 0$, pick $j = j(\delta)$ so large that $1/j! < \frac{1}{2}\delta$, and define

$$\begin{aligned} B_N &= \big\{\sigma \in \mathcal{P}_N\colon\ \sigma_{N-j+1} < \sigma_{N-j+2} < \ldots < \sigma_N\big\} \\ &= \text{ set of permutations whose last } j \text{ terms are ordered upwards}, \quad N \geq j. \end{aligned}$$

Then $\pi(B_N) = 1/j!$, and $\{X_n \in B_N\}$ is the event that the order of the original j bottom cards is retained at time n. Since the first time the card with label $N-j+1$ comes to the top is distributed like V_{N-j+1}, we have

$$\mathbb{P}\big(X_{(1-\epsilon)N\log N} \in B_N\big) \geq \mathbb{P}\big(V_{N-j+1} > (1-\epsilon)N\log N\big). \tag{4.4}$$

Indeed, for the upward ordering to be destroyed, the card with label $N-j+1$ must come to the top and must subsequently be inserted below the card with label $N-j+1$. We will show that, for $N \geq N(\delta)$,

$$\mathbb{P}\big(V_{N-j+1} \leq (1-\epsilon)N\log N\big) < \tfrac{1}{2}\delta. \tag{4.5}$$

From this it follows that

$$\begin{aligned} \big\|\mathbb{P}(X_{(1-\epsilon)N\log N)} \in \cdot\,) - \pi(\cdot)\big\|_{tv} &\geq 2\left[\mathbb{P}(X_{(1-\epsilon)N\log N)} \in B_N) - \pi(B_N)\right] \\ &\geq 2\left[1 - \mathbb{P}\big(V_{N-j+1} \leq (1-\epsilon)N\log N\big) - \pi(B_N)\right] \\ &\geq 2\left[1 - \tfrac{1}{2}\delta - \tfrac{1}{2}\delta\right] = 2(1-\delta). \end{aligned}$$

The first inequality uses the definition of total variation, the third inequality uses (4.4) and (4.5). By letting $N \to \infty$ followed by $\delta \downarrow 0$, we get the first line of Definition 4.2.

To prove (4.5), we compute

$$\begin{aligned} \mathbb{E}(V_{N-j+1}) &= \sum_{i=j-1}^{N-1} \mathbb{E}(V_{N-i} - V_{N-i-1}) \\ &= \sum_{i=j-1}^{N-1} \frac{N}{i+1} \sim N\log\frac{N}{j} \sim N\log N \\ \mathrm{Var}(V_{N-j+1}) &= \sum_{i=j-1}^{N-1} \mathrm{Var}(V_{N-i} - V_{N-i-1}) \\ &= \sum_{i=j-1}^{N-1} \left(\frac{N}{i+1}\right)^2 \left(1 - \frac{i+1}{N}\right) \sim c_j N^2,\ c_j = \sum_{k\geq j} k^{-2}. \end{aligned}$$

Here we use that $\mathbb{E}(\mathrm{GEO}(p)) = 1/p$ and $\mathrm{Var}(\mathrm{GEO}(p)) = (1-p)/p^2$. Chebyshev's inequality therefore gives

$$\begin{aligned} \mathbb{P}\big(V_{N-j+1} \leq (1-\epsilon)N\log N\big) &= \mathbb{P}\Big(V_{N-j+1} - \mathbb{E}(V_{N-j+1}) \leq -\epsilon N\log N[1+o(1)]\Big) \\ &\leq \mathbb{P}\Big([V_{N-j+1} - \mathbb{E}(V_{N-j+1})]^2 \geq \epsilon^2 N^2\log^2 N[1+o(1)]\Big) \\ &\leq \frac{\mathrm{Var}(V_{N-j+1})}{\epsilon^2\mathbb{E}(V_{N-j+1})^2}[1+o(1)] \\ &\sim \frac{c_j N^2}{\epsilon^2 N^2\log^2 N} = O\left(\frac{1}{\log^2 N}\right). \end{aligned}$$

This proves (4.5). ∎

Remark: We have shown that $\mathbb{E}(T_N) = 1 + \sum_{i=1}^{N}(N/i) \sim N \log N$ and $\mathrm{Var}(T_N/\mathbb{E}(T_N)) \to 0$ as $N \to \infty$. This in turn implies that $t_N/T_N \to 1$ in probability as $N \to \infty$ and identifies the scaling of the threshold time as $t_N \sim \mathbb{E}(T_N)$, in accordance with the prediction made in (4.1).

5 Poisson approximation

In Section 1.3 we already briefly described coupling in the context of Poisson approximation. We now return to this topic. Let $\mathrm{BINOM}(n,p) = \{\binom{n}{k}p^k(1-p)^{n-k}\colon\ k = 0,\dots,n\}$ be the binomial distribution with parameters $n \in \mathbb{N}$ and $p \in [0,1]$. A classical result from probability theory is that, for every $c \in (0,\infty)$, $\mathrm{BINOM}(n, c/n)$ is close to $\mathrm{POISSON}(c)$ when n is large. In this section we will quantify how close, by developing a general theory for approximations to the Poisson distribution called the *Stein-Chen method*. After suitable modification, the same method also works for approximation to other types of distributions, e.g. the Gaussian distribution, but this will not be pursued.

In Section 5.1 we derive a crude bound for sums of independent $\{0,1\}$-valued random variables. In Section 5.2 we describe the Stein-Chen method, which not only leads to a better bound, but also applies to dependent random variables. In Section 5.3 we look at two specific applications.

5.1 Coupling

Fix $n \in \mathbb{N}$ and $p_1,\dots,p_n \in [0,1)$. Let

$$Y_i \stackrel{D}{=} \mathrm{BER}(p_i), \qquad i = 1,\dots,n, \qquad \text{be independent,}$$

i.e., $\mathbb{P}(Y_i = 1) = p_i$ and $\mathbb{P}(Y_i = 0) = 1 - p_i$, and put $X = \sum_{i=1}^n Y_i$.

Theorem 5.1 *With the above definitions,*

$$\|\mathbb{P}(X \in \cdot\,) - p_\lambda(\cdot)\|_{tv} \le \sum_{i=1}^n \lambda_i^2$$

with $\lambda_i = -\log(1-p_i)$, $\lambda = \sum_{i=1}^n \lambda_i$ *and* $p_\lambda = \mathrm{POISSON}(\lambda)$.

Proof. Let $Y_i' \stackrel{D}{=} \mathrm{POISSON}(\lambda_i)$, $i = 1,\dots,n$, be independent, and put $X' = \sum_{i=1}^n Y_i'$. Then

$$\begin{aligned} Y_i &\stackrel{D}{=} Y_i' \wedge 1, \quad i = 1,\dots,n,\\ X' &\stackrel{D}{=} \mathrm{POISSON}(\lambda), \end{aligned}$$

where the first line uses that $e^{-\lambda_i} = 1 - p_i$ and the second line uses that the independent sum of Poisson random variables with given parameters is again Poisson, with parameter equal to the sum of the constituent parameters. It follows that

$$\begin{aligned} \mathbb{P}(X \neq X') &\le \sum_{i=1}^n \mathbb{P}(Y_i \neq Y_i') = \sum_{i=1}^n \mathbb{P}(Y_i' \ge 2),\\ \mathbb{P}(Y_i' \ge 2) &= \sum_{k=2}^\infty e^{-\lambda_i}\frac{\lambda_i^k}{k!} \le \tfrac{1}{2}\lambda_i^2 \sum_{l=0}^\infty e^{-\lambda_i}\frac{\lambda_i^l}{l!} = \tfrac{1}{2}\lambda_i^2, \end{aligned}$$

where the second inequality uses that $k! \ge 2(k-2)!$ for $k \ge 2$. Since

$$\|\mathbb{P}(X \in \cdot\,) - p_\lambda(\cdot)\|_{tv} = \|\mathbb{P}(X \in \cdot\,) - \mathbb{P}(X' \in \cdot\,)\|_{tv} \le 2\mathbb{P}(X \neq X'),$$

the claim follows. ∎

Remark: The interest in Theorem 5.1 is when n is large, $p_1, \ldots, p_n$ are small and λ is of order 1. (Note that $\sum_{i=1}^n \lambda_i^2 \leq M\lambda$ with $M = \max\{\lambda_1, \ldots, \lambda_n\}$.) A typical example is $p_i \equiv c/n$, in which case $\sum_{i=1}^n \lambda_i^2 = n[-\log(1-c/n)]^2 \sim c^2/n$ as $n \to \infty$.

Remark: In Section 1.3 we derived a bound similar to Theorem 5.1 but with $\lambda_i = p_i$. For $p_i \downarrow 0$ we have $\lambda_i \sim p_i$, and so the difference between the two bounds is minor.

5.2 Stein-Chen method

We next turn our attention to a more sophisticated way of achieving a Poisson approximation, which is called the *Stein-Chen method*. Not only will this lead to better bounds, it will also be possible to deal with random variables that are *dependent*. For details, see Barbour, Holst and Janson [2].

5.2.1 Sums of dependent Bernoulli random variables

Again, we fix $n \in \mathbb{N}$ and $p_1, \ldots, p_n \in [0,1]$, and we let

$$Y_i \stackrel{D}{=} \mathrm{BER}(p_i), \qquad i = 1, \ldots, n.$$

However, we do *not* require the Y_i's to be independent. We abbreviate (note the change of notation compared to Section 5.1)

$$W = \sum_{i=1}^n Y_i, \qquad \lambda = \sum_{i=1}^n p_i, \tag{5.1}$$

and, for $j = 1, \ldots, n$, define random variables U_j and V_j satisfying

$$\begin{aligned} U_j \stackrel{D}{=} W\colon &\quad \mathbb{P}(U_j \in \cdot\,) = \mathbb{P}(W \in \cdot\,), \\ V_j \stackrel{D}{=} W - 1 \mid Y_j = 1\colon &\quad \mathbb{P}(V_j \in \cdot\,) = \mathbb{P}(W - 1 \in \cdot \mid Y_j = 1), \end{aligned} \tag{5.2}$$

where we note that $W - 1 = \sum_{i \neq j} Y_i$ when $Y_j = 1$ (and we put $V_j = 0$ when $\mathbb{P}(Y_j = 1) = 0$). No condition of independence of U_j and V_j is required. Clearly, if $U_j = V_j$, $j = 1 \ldots, n$, with large probability, then we expect the Y_i's to be *weakly dependent*. In that case, if the p_i's are small, then we expect a good Poisson approximation to be possible.

Before we proceed, we state *two core ingredients in the Stein-Chen method*. These will be exploited in Section 5.2.2.

Lemma 5.2 *If $Z \stackrel{D}{=} \mathrm{POISSON}(\lambda)$ for some $\lambda \in (0, \infty)$, then for any bounded function $f\colon \mathbb{N} \to \mathbb{R}$,*

$$\mathbb{E}\big(\lambda f(Z+1) - Z f(Z)\big) = 0. \tag{5.3}$$

Proof. In essence, (5.3) is a recursion relation that is *specific* to the Poisson distribution. Indeed, let $p_\lambda(k) = e^{-\lambda}\lambda^k/k!$, $k \in \mathbb{N}_0$, denote the coefficients of $\mathrm{POISSON}(\lambda)$. Then

$$\lambda p_\lambda(k) = (k+1) p_\lambda(k+1), \qquad k \in \mathbb{N}_0, \tag{5.4}$$

and hence

$$\begin{aligned}\mathbb{E}\big(\lambda f(Z+1)\big) &= \sum_{k\in\mathbb{N}_0} \lambda p_\lambda(k) f(k+1) \\ &= \sum_{k\in\mathbb{N}_0} (k+1) p_\lambda(k+1) f(k+1) \\ &= \sum_{l\in\mathbb{N}} p_\lambda(l) l f(l) \\ &= \mathbb{E}(Z f(Z)).\end{aligned}$$

∎

Lemma 5.3 *For $\lambda \in (0,\infty)$ and $A \subset \mathbb{N}_0$, let $g_{\lambda,A}\colon \mathbb{N}_0 \to \mathbb{R}$ be the solution of the recursive equation*

$$\begin{aligned}&\lambda g_{\lambda,A}(k+1) - k g_{\lambda,A}(k) = 1_A(k) - p_\lambda(A), \qquad k \in \mathbb{N}_0, \\ &g_{\lambda,A}(0) = 0.\end{aligned} \tag{5.5}$$

Then, uniformly in A,

$$\|\Delta g_{\lambda,A}\|_\infty = \sup_{k\in\mathbb{N}_0} |g_{\lambda,A}(k+1) - g_{\lambda,A}(k)| \le 1 \wedge \lambda^{-1}.$$

Proof. For $k \in \mathbb{N}_0$, let $U_k = \{0,1,\dots,k\}$. Then the solution of the recursive equation is given by $g_{\lambda,A}(0) = 0$ and

$$g_{\lambda,A}(k+1) = \frac{1}{\lambda p_\lambda(k)} \big[p_\lambda(A \cap U_k) - p_\lambda(A) p_\lambda(U_k)\big], \qquad k \in \mathbb{N}_0, \tag{5.6}$$

as may be checked by induction on k. From this formula we deduce two facts:

$$g_{\lambda,A} = \sum_{j\in A} g_{\lambda,\{j\}}, \tag{5.7}$$

$$g_{\lambda,A} = -g_{\lambda,A^c}, \tag{5.8}$$

with $A^c = \mathbb{N}_0 \setminus A$.

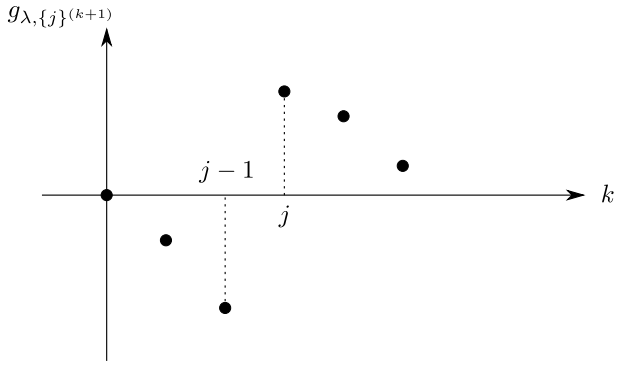
Exercise 5.4 *Check the claims in* (5.6–5.8).

For $A = \{j\}$, the solution reads

$$g_{\lambda,\{j\}}(k+1) = \begin{cases} -\frac{p_\lambda(j)}{\lambda p_\lambda(k)} \sum_{l=0}^{k} p_\lambda(l), & k < j, \\ +\frac{p_\lambda(j)}{\lambda p_\lambda(k)} \sum_{l=k+1}^{\infty} p_\lambda(l), & k \ge j, \end{cases} \tag{5.9}$$

from which we see that

$$k \mapsto g_{\lambda,\{j\}}(k+1) \text{ is } \begin{cases} \text{negative and decreasing for } k < j, \\ \text{positive and decreasing for } k \ge j. \end{cases}$$



Hence $g_{\lambda,\{j\}}(k+1) - g_{\lambda,\{j\}}(k) \leq 0$ for $k \neq j$, while for $k = j$

$$\begin{aligned} g_{\lambda,\{j\}}(j+1) - g_{\lambda,\{j\}}(j) &= \frac{1}{\lambda}\left(\frac{p_\lambda(j)}{p_\lambda(j)}\sum_{l=j+1}^{\infty} p_\lambda(l) + \frac{p_\lambda(j)}{p_\lambda(j-1)}\sum_{l=0}^{j-1} p_\lambda(l)\right) \\ &= \frac{1}{\lambda}\left(\sum_{l=j+1}^{\infty} p_\lambda(l) + \frac{\lambda}{j}\sum_{l=0}^{j-1} p_\lambda(l)\right) \\ &= \frac{1}{\lambda}\left(\sum_{l=j+1}^{\infty} p_\lambda(l) + \sum_{l=1}^{j} p_\lambda(l)\frac{l}{j}\right) \\ &\leq \frac{1}{\lambda}\sum_{l=1}^{\infty} p_\lambda(l) = \frac{1}{\lambda}(1-e^{-\lambda}) \leq 1 \wedge \lambda^{-1}, \end{aligned}$$

where the second and third equality use (5.4). It follows from (5.7) that

$$g_{\lambda,A}(k+1) - g_{\lambda,A}(k) \leq 1 \wedge \lambda^{-1},$$

where we use that the jumps from negative to positive in (5.9) occur at disjoint positions as j runs through A. Combine the latter inequality with (5.8) to get

$$g_{\lambda,A}(k+1) - g_{\lambda,A}(k) \geq -(1 \wedge \lambda^{-1}),$$

so that $\|\Delta g_{\lambda,A}\|_\infty \leq 1 \wedge \lambda^{-1}$. ∎

5.2.2 Bound on total variation distance

We are now ready to state the result we are after. This result will be exploited later on.

Theorem 5.5 *Let $n \in \mathbb{N}$, $p_1, \ldots, p_n \in [0,1)$ and W, U, V as defined above. Then*

$$\|\mathbb{P}(W \in \cdot\,) - p_\lambda(\cdot\,)\|_{tv} \leq 2(1 \wedge \lambda^{-1})\sum_{j=1}^{n} p_j \mathbb{E}(|U_j - V_j|).$$

Proof. Pick any $A \subset \mathbb{N}_0$ and write

$$\begin{aligned}
\mathbb{P}(W \in A) - p_\lambda(A) &= \mathbb{E}\big(1_A(W) - p_\lambda(A)\big) \\
&= \mathbb{E}\big(\lambda g_{\lambda,A}(W+1) - W g_{\lambda,A}(W)\big) \\
&= \sum_{j=1}^{n} \big[p_j \mathbb{E}\big(g_{\lambda,A}(W+1)\big) - \mathbb{E}\big(Y_j g_{\lambda,A}(W)\big)\big] \\
&= \sum_{j=1}^{n} p_j \big[\mathbb{E}\big(g_{\lambda,A}(W+1)\big) - \mathbb{E}\big(g_{\lambda,A}(W) \mid Y_j = 1\big)\big] \\
&= \sum_{j=1}^{n} p_j \mathbb{E}\big(g_{\lambda,A}(U_j+1) - g_{\lambda,A}(V_j+1)\big),
\end{aligned}$$

where the second equality uses (5.5), the third equality uses (5.1), while the fifth equality uses (5.2). Applying (5.5) once more, we get

$$|\mathbb{P}(W \in A) - p_\lambda(A)| \le (1 \wedge \lambda^{-1}) \sum_{j=1}^{n} p_j \mathbb{E}(|U_j - V_j|),$$

and taking the supremum over A we get the claim. ∎

To put Theorem 5.5 to use, we look at a *subclass* of dependent $Y_1, \ldots, Y_n$.

Definition 5.6 *The above random variables $Y_1, \ldots, Y_n$ are said to be negatively related if there exist arrays of random variables*

$$\left.\begin{matrix} Y_{j1}, \ldots, Y_{jn} \\ Y'_{j1}, \ldots, Y'_{jn} \end{matrix}\right\} \qquad j = 1, \ldots, n,$$

such that, for each j with $\mathbb{P}(Y_j = 1) > 0$,

$$\begin{aligned}
&(Y_{j1}, \ldots, Y_{jn}) \stackrel{D}{=} (Y_1, \ldots, Y_n), \\
&(Y'_{j1}, \ldots, Y'_{jn}) \stackrel{D}{=} (Y_1, \ldots, Y_n) \mid Y_j = 1, \\
&Y'_{ji} \le Y_{ji} \qquad \forall\, i \ne j,
\end{aligned}$$

while, for each j with $\mathbb{P}(Y_j = 1) = 0$, $Y'_{ji} = 0$ for $j \ne i$ and $Y'_{jj} = 1$.

What negative relation means is that the condition $Y_j = 1$ has a tendency to force $Y_i = 0$ for $i \ne j$. Thus, negative relation is like negative correlation (although the notion is in fact stronger).

An important consequence of negative relation is that there exists a coupling such that $U_j \ge V_j$ for all j. Indeed, we may pick

$$U_j = \sum_{i=1}^{n} Y_{ji}, \qquad V_j = -1 + \sum_{i=1}^{n} Y'_{ji},$$

in which case (5.2) is satisfied and, moreover,

$$U_j - V_j = \sum_{\substack{i=1,\ldots,n \\ i \ne j}} \underbrace{(Y_{ji} - Y'_{ji})}_{\ge 0} + \underbrace{(1 - Y'_{jj})}_{\ge 0} + \underbrace{Y_{jj}}_{\ge 0} \ge 0.$$

The ordering $U_j \ge V_j$ has the following important consequence.

Theorem 5.7 *If $Y_1, \ldots, Y_n$ are negatively related, then*

$$\|\mathbb{P}(W \in \cdot) - p_\lambda(\cdot)\|_{tv} \leq 2(1 \wedge \lambda^{-1})[\lambda - \mathrm{Var}(W)].$$

Proof. The ordering $U_j \geq V_j$ allows us to compute the sum that appears in the bound in Theorem 5.5:

$$\begin{aligned}
\sum_{j=1}^n p_j \mathbb{E}(|U_j - V_j|) &= \sum_{j=1}^n p_j \mathbb{E}(U_j - V_j) \\
&= \sum_{j=1}^n p_j \mathbb{E}(W) - \sum_{j=1}^n p_j \mathbb{E}(W \mid Y_j = 1) + \sum_{j=1}^n p_j \\
&= \mathbb{E}(W)^2 - \sum_{j=1}^n \mathbb{E}(Y_j W) + \lambda \\
&= \mathbb{E}(W)^2 - \mathbb{E}(W^2) + \lambda \\
&= -\mathrm{Var}(W) + \lambda,
\end{aligned}$$

where the second equality uses (5.2). ∎

Remark: The upper bound in Theorem 5.7 *only contains the unknown quantity* $\mathrm{Var}(W)$. It turns out that in many examples this quantity can be either computed or estimated.

5.3 Two applications

1. Let $Y_1, \ldots, Y_n$ be independent (as assumed previously). Then $\mathrm{Var}(W) = \sum_{i=1}^n \mathrm{Var}(Y_i) = \sum_{i=1}^n p_i(1-p_i) = \lambda - \sum_{i=1}^n p_i^2$, and the bound in Theorem 5.7 reads

$$2(1 \wedge \lambda^{-1}) \sum_{i=1}^n p_i^2,$$

which is better than the bound derived in Section 1.3 when $\lambda \geq 1$.

2. Consider $N \geq 2$ urns and $1 \leq m < N$ balls. Each urn can contain at most one ball. Place the balls "randomly" into the urns, i.e., each of the $\binom{N}{m}$ configurations has equal probability. For $i = 1, \ldots, N$, let

$$Y_i = 1_{\{\text{urn } i \text{ contains a ball}\}}.$$

Pick $n < N$ and let $W = \sum_{i=1}^n Y_i$. Then the probability distribution of W is hypergeometric, i.e.,

$$\mathbb{P}(W = k) = \binom{n}{k}\binom{N-n}{m-k}\binom{N}{m}^{-1}, \qquad k = 0 \vee (m+n-N), \ldots, m \wedge n,$$

where $\binom{n}{k}$ is the number of ways to place k balls in urns $1, \ldots, n$ and $\binom{N-n}{m-k}$ in the number of ways to place $m-k$ balls in urns $n+1, \ldots, N$.

Exercise 5.8 *Check that the right-hand side is a probability distribution. Show that*

$$\begin{aligned}
&\mathbb{E}(W) = n\tfrac{m}{N} = \lambda, \\
&\mathrm{Var}(W) = n\tfrac{m}{N}(1 - \tfrac{m}{N})\tfrac{N-n}{N-1}.
\end{aligned}$$

It is *intuitively* clear that $Y_1, \ldots, Y_n$ are negatively related: if we condition on urn j to contain a ball, then urn i with $i \neq j$ is less likely to contain a ball. More formally, recall Definition 5.6 and, for $j = 1, \ldots, n$, define $Y_{j1}, \ldots, Y_{jn}$ and $Y'_{j1}, \ldots, Y'_{jn}$ as follows:

- Place a ball in urn j.
- Place the remaining $m-1$ balls randomly in the other $N-1$ urns.
- Put $Y'_{ji} = 1_{\{\text{urn } i \text{ contains a ball}\}}$.
- Toss a coin that produces head with probability $\frac{m}{N}$.
- If head comes up, then put $(Y_{j1}, \ldots, Y_{jn}) = (Y'_{j1}, \ldots, Y'_{jn})$.
- If tail comes up, then pick the ball in urn j, place it randomly in one of the $N-m-1$ urns that are empty, and put $Y_{ji} = 1_{\{\text{urn } i \text{ contains a ball}\}}$.

Exercise 5.9 *Check that the above construction produces arrays with the properties required by Definition* 5.6.

We expect that if $m/N, n/N \ll 1$, then W is approximately Poisson distributed. The formal computation goes as follows. Using Theorem 5.7 and Exercise 5.9, we get

$$\begin{aligned} \|\mathbb{P}(W \in \cdot\,) - p_\lambda(\cdot)\|_{tv} &\leq 2(1 \wedge \lambda^{-1})[\lambda - \mathrm{Var}(W)] \\ &= 2(1 \wedge \lambda^{-1})\lambda \left[1 - \left(1 - \frac{m}{N}\right)\frac{N-n}{N-1}\right] \\ &= 2(1 \wedge \lambda^{-1})\lambda \frac{(m+n-1)N - mn}{N(N-1)} \\ &\leq 2\frac{m+n-1}{N-1}. \end{aligned}$$

Indeed, this is small when $m/N, n/N \ll 1$.

6 Markov Chains

In Section 1.1 we already briefly described coupling for Markov chains. We now return to this topic. We recall that $X = (X_n)_{n\in\mathbb{N}_0}$ is a Markov chain on a *countable* state space S, with an initial distribution $\lambda = (\lambda_i)_{i\in S}$ and with a transition matrix $P = (P_{ij})_{i,j\in S}$ that is *irreducible* and *aperiodic*.

There are three cases, which will be treated in Sections 6.1–6.3:

1. positive recurrent,
2. null recurrent,
3. transient.

In case 1 there exists a unique stationary distribution π, solving the equation $\pi = \pi P$ and satisfying $\pi > 0$, such that $\lim_{n\to\infty} \lambda P^n = \pi$ componentwise on S. The latter is the standard *Markov Chain Convergence Theorem*, and we want to investigate *the rate* of convergence. In cases 2 and 3 there is *no* stationary distribution, and $\lim_{n\to\infty} \lambda P^n = 0$ componentwise. We want to investigate the rate of convergence as well, and see what the role is of the initial distribution λ.

In Section 6.4 we take a brief look at "perfect simulation", where coupling of Markov chains is used to simulate random variables *with no error*.

6.1 Case 1: Positive recurrent

For $i \in S$, let

$$T_i = \min\{n \in \mathbb{N}\colon\ X_n = i\},$$
$$m_i = \mathbb{E}_i(T_i) = \mathbb{E}(T_i \mid X_0 = i),$$

which, by positive recurrence, are finite. A basic result of Markov chain theory is that $\pi_i = 1/m_i$, $i \in S$ (see Häggström [5], Chapter 5, and Kraaikamp [7], Section 2.2).

We want to compare two copies of the Markov chain starting from different initial distributions $\lambda = (\lambda_i)_{i\in S}$ and $\mu = (\mu_i)_{i\in S}$, which we denote by $X = (X_n)_{n\in\mathbb{N}_0}$ and $X' = (X'_n)_{n\in\mathbb{N}_0}$, respectively. Let

$$T = \min\{n \in \mathbb{N}_0\colon\ X_n = X'_n\}$$

denote their first meeting time. Then the standard coupling inequality in Theorem 2.7 gives

$$\|\lambda P^n - \mu P^n\|_{tv} \le 2\hat{\mathbb{P}}_{\lambda,\mu}(T > n),$$

where $\hat{\mathbb{P}}_{\lambda,\mu}$ denotes *any* probability measure that couples X and X'. We will choose the *independent coupling* $\hat{\mathbb{P}}_{\lambda,\mu} = \mathbb{P}_\lambda \otimes \mathbb{P}_\mu$, and instead of T focus on

$$T^* = \min\{n \in \mathbb{N}_0\colon\ X_n = X'_n = *\},$$

their first meeting time at $*$ (where $*$ is any chosen state in S). Since $T^* \ge T$, we have

$$\|\lambda P^n - \mu P^n\|_{tv} \le 2\hat{\mathbb{P}}_{\lambda,\mu}(T^* > n). \tag{6.1}$$

The key fact that we will uses is the following.

Theorem 6.1 *Under positive recurrence,*

$$\hat{\mathbb{P}}_{\lambda,\mu}(T^* < \infty) = 1 \qquad \forall\, \lambda, \mu.$$

Proof. The successive visits to $*$ by X and X', given by the $\{0,1\}$-valued random sequences

$$\begin{aligned} Y &= (Y_k)_{k\in\mathbb{N}_0} \quad \text{with } Y_k = 1_{\{X_k=*\}}, \\ Y' &= (Y'_k)_{k\in\mathbb{N}_0} \quad \text{with } Y'_k = 1_{\{X'_k=*\}}, \end{aligned}$$

constitute a *renewal process*: each time $*$ is hit the process of returns to $*$ starts from scratch. Define

$$\hat{Y}_k = Y_k Y'_k, \qquad k \in \mathbb{N}_0.$$

Then also $\hat{Y} = (\hat{Y}_k)_{k\in\mathbb{N}_0}$ is a renewal process. Let

$$I = \{\hat{Y}_k = 1 \text{ for infinitely many } k\}.$$

It suffices to show that $\hat{\mathbb{P}}_{\lambda,\mu}(I) = 1$ for all λ, μ.

If $\lambda = \mu = \pi$, then $\hat{Y}$ is *stationary* and, since $\hat{\mathbb{P}}_{\pi,\pi}(\hat{Y}_0 = 1) = \pi_0^2 > 0$, it follows from the renewal property that $\hat{\mathbb{P}}_{\pi,\pi}(I) = 1$. Since $\pi > 0$, the latter in turn implies that

$$\hat{\mathbb{P}}_{\lambda,\mu}(I) = 1,$$

which yields the claim. ∎

Exercise 6.2 *Check the last statement in the proof.*

Theorem 6.1 combined with (6.1) implies that

$$\lim_{n\to\infty} \|\lambda P^n - \mu P^n\|_{tv} = 0,$$

and by picking $\mu = \pi$ we get the Markov Chain Convergence Theorem.

Remark: If $|S| < \infty$, then the convergence is exponentially fast. Indeed, pick k so large that

$$\min_{i,j\in S} (P^k)_{ij} \stackrel{\text{def}}{=} \rho > 0,$$

which is possible by irreducibility and aperiodicity. Then

$$\hat{\mathbb{P}}_{\lambda,\mu}(X_k \neq X'_k) \leq 1 - \rho \qquad \forall\, \lambda, \mu,$$

and hence, by the Markov property,

$$\hat{\mathbb{P}}_{\lambda,\mu}(T > n) \leq (1-\rho)^{\lfloor n/k \rfloor} \qquad \forall\, \lambda, \mu, n,$$

where $\lfloor \cdot \rfloor$ denotes the lower integer part. Via the standard coupling inequality this shows that

$$\|\lambda P^n - \mu P^n\|_{tv} \leq 2(1-\rho)^{\lfloor n/k \rfloor} = \exp[-cn + o(n)],$$

with $c = \frac{1}{k}\log[1/(1-\rho)] > 0$.

Remark: All rates of decay are possible when $|S| = \infty$: sometimes exponential, sometimes polynomial. With the help of Theorem 2.10 it is possible to estimate the rate when some additional control on the moments of T or T^* is available (recall Section 2.3). This typically requires additional structure. For simple random walk on $\mathbb{Z}$ and $\mathbb{Z}^2$ it is known that $\mathbb{P}(T > n) \asymp 1/\sqrt{n}$, respectively, $\mathbb{P}(T^* > n) \asymp 1/\log n$ (Spitzer [14], Section 3).

6.2 Case 2: Null recurrent

Null recurrent Markov chains do not have a stationary distribution. Consequently,

$$\lim_{n\to\infty} \lambda P^n = 0 \text{ pointwise} \qquad \forall\, \lambda. \tag{6.2}$$

Is it still the case that

$$\lim_{n\to\infty} \|\lambda P^n - \mu P^n\|_{tv} = 0 \qquad \forall\, \lambda, \mu? \tag{6.3}$$

It suffices to show that there exists a coupling $\hat{\mathbb{P}}_{\lambda,\mu}$ such that $\hat{\mathbb{P}}_{\lambda,\mu}(T^* < \infty) = 1$. The proof of Theorem 6.1 for positive recurrent Markov chains does not carry over because there is no stationary distribution. However, it is enough to show that there exists a coupling $\hat{\mathbb{P}}_{\lambda,\mu}$ such that $\hat{\mathbb{P}}_{\lambda,\mu}(T < \infty) = 1$, which seems easier because the two copies of the Markov chain only need to meet *somewhere*, not necessarily at $*$.

Theorem 6.3 *Under null recurrence,*

$$\hat{\mathbb{P}}_{\lambda,\mu}(T < \infty) = 1 \qquad \forall\, \lambda, \mu.$$

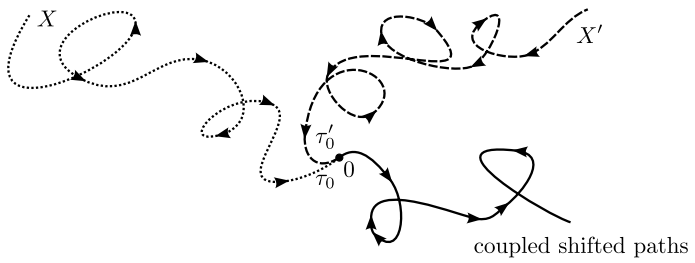
Proof. A proof of this theorem and hence of (6.3) is beyond the scope of the present course. We refer to Lindvall [11], Section III.21, for more details. As a weak substitute we prove the "Cesaro average" version of (6.3):

$$X \text{ recurrent} \implies \lim_{N\to\infty} \left\| \frac{1}{N} \sum_{n=0}^{N-1} \lambda P^n - \frac{1}{N} \sum_{n=0}^{N-1} \mu P^n \right\|_{tv} = 0 \qquad \forall\, \lambda, \mu.$$

The proof uses the notion of *shift-coupling*, i.e., coupling with a *random time shift*. Let X and X' be two independent copies of the Markov chain starting from λ and μ. Write 0 instead of $*$, and let τ_0 and τ_0' denote the first hitting times of 0. Couple X and X' by letting their paths coincide after τ_0, respectively, τ_0':

$$X_{k+\tau_0} = X'_{k+\tau_0'} \qquad \forall\, k \in \mathbb{N}_0.$$

This definition makes sense because $\mathbb{P}(\tau_0 < \infty) = \mathbb{P}(\tau_0' < \infty) = 1$ by recurrence.



Fix any event A. Write

$$\begin{aligned}
&\left|\frac{1}{N}\sum_{n=0}^{N-1}(\lambda P^n)(A) - \frac{1}{N}\sum_{n=0}^{N-1}(\mu P^n)(A)\right| \\
= \;& \frac{1}{N}\left|\sum_{n=0}^{N-1}\hat{\mathbb{P}}_{\lambda,\mu}(X_n \in A) - \sum_{n=0}^{N-1}\hat{\mathbb{P}}_{\lambda,\mu}(X'_n \in A)\right| \\
= \;& \frac{1}{N}\sum_{m,m'\in\mathbb{N}_0}\hat{\mathbb{P}}_{\lambda,\mu}\big((\tau_0,\tau'_0)=(m,m')\big) \\
&\quad\times\left|\sum_{n=0}^{N-1}\hat{\mathbb{P}}_{\lambda,\mu}\big(X_n\in A \mid (\tau_0,\tau'_0)=(m,m')\big) - \sum_{n=0}^{N-1}\hat{\mathbb{P}}_{\lambda\mu}\big(X'_n\in A \mid (\tau_0,\tau'_0)=(m,m')\big)\right| \\
\le \;& \hat{\mathbb{P}}_{\lambda,\mu}(\tau_0\vee\tau'_0\ge M) + \frac{1}{N}\sum_{\substack{m,m'\in\mathbb{N}_0\\ m\vee m'<M}}\hat{\mathbb{P}}_{\lambda,\mu}\big((\tau_0,\tau'_0)=(m,m')\big) \\
&\quad\times\Bigg\{2(m\vee m') + \Bigg|\sum_{k=0}^{(N-m-1)\wedge(N-m'-1)} \\
&\qquad\Big[\hat{\mathbb{P}}_{\lambda,\mu}\big(X_{m+k}\in A \mid (\tau_0,\tau'_0)=(m,m')\big) - \hat{\mathbb{P}}_{\lambda,\mu}\big(X'_{m'+k}\in A \mid (\tau_0,\tau'_0)=(m,m')\big)\Big]\Bigg|\Bigg\} \\
\le \;& \hat{\mathbb{P}}_{\lambda,\mu}(\tau_0\vee\tau'_0\ge M) + \frac{2}{N}\,\mathbb{E}\Big((\tau_0\vee\tau'_0)1_{\{\tau_0\vee\tau'_0<M\}}\Big).
\end{aligned}$$

In the first inequality we take $M \le N$ and note that $m + m' + |m - m'| = 2(m \vee m')$ is the number of summands that are lost by letting the sums start at $n = m$, respectively, $n = m'$, shifting them by m, respectively, m', and afterwards cutting them at $(N-m-1)\wedge(N-m'-1)$. In the second inequality the sum over k is zero by the shift-coupling.

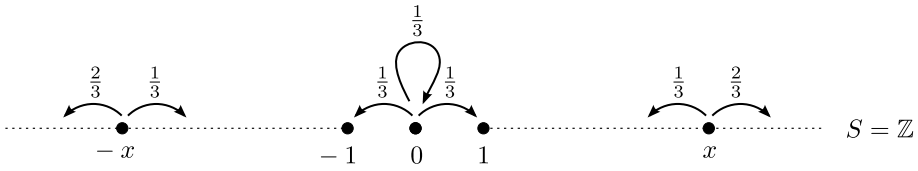
Since the bound is uniform in A, we get the claim by taking the supremum over A and letting $N \to \infty$ followed by $M \to \infty$. ∎

6.3 Case 3: Transient

There is no general result for transient Markov chains: (6.2) always holds, but (6.3) *may hold or may fail*. For the special case of random walks on $\mathbb{Z}^d$, $d \ge 1$, we saw with the help of the Ornstein coupling that (6.3) holds. We also mentioned that for arbitrary random walk

$$\hat{\mathbb{P}}_{\lambda,\mu}(T > n) = O(1/\sqrt{n}),$$

the rate of the componentwise coupling. Here is an example of a Markov chain for which (6.3) fails:



At site x the random walk has:

$$\begin{array}{ll} \text{zero drift with pausing} & \text{for } x = 0, \\ \text{positive drift} & \text{for } x > 0, \\ \text{negative drift} & \text{for } x < 0. \end{array}$$

This Markov chain is irreducible and aperiodic, with $\lim_{x\to\infty} P_x(\tau_0 = \infty) = \lim_{x\to-\infty} P_x(\tau_0 = \infty) = 1$. As a result, we have

$$\lim_{x\to\infty} \liminf_{n\to\infty} \|\delta_x P^n - \delta_{-x} P^n\|_{tv} = 2.$$

6.4 Perfect simulation

The results in Section 6.1 are important for *simulation*. Suppose that we are given a finite set S, and a probability distribution ρ on S from which we want to *draw random samples*. Then we can proceed as follows. Construct an irreducible and aperiodic Markov chain on S whose stationary distribution is ρ. The Markov Chain Convergence Theorem tells us that if we start this Markov chain at any site $i^* \in S$, then after a long time its distribution will be close to ρ. Thus, any *late observation* of the Markov chain provides us with a *good approximation* of a random draw from ρ.

The above approach needs two ingredients:

1. A way to find a transition matrix P on S whose stationary distribution π is equal to the given probability distribution ρ.

2. A rate of convergence estimate that provides an upper bound on the total variation distance $n \mapsto \|\delta_{i^*} P^n - \pi\|_{tv}$ for a given $i^* \in S$, so that any desired accuracy of the approximation can be achieved by running the Markov chain long enough.

Both these ingredients give rise to a *theory of simulation*, for which an extensive literature exists (see e.g. Levin, Peres and Williams [8]).

The drawback is that the simulation is *at best approximate*: no matter how long we run the Markov chain, its distribution is never perfectly equal to ρ (at least in typical situations). Häggström [5], Chapters 10–12, contain an outline of a different approach, through which it is possible to achieve a *perfect simulation*, i.e., to obtain a random sample whose distribution is equal to ρ with *no error* (!) In this approach, independent copies of the Markov chain are started from each site of S "far back in the past", and the simulation is stopped at time zero when all the copies "have collided prior to time zero". The observation of the Markov chain at time zero provides the perfect sample.

The details of the construction are somewhat delicate and we refer the reader to the relevant literature. Concrete examples are discussed in [5].

7 Probabilistic inequalities

In Chapters 1 and 3–6 we have seen coupling at work in a number of different situations. We now return to the basic theory that was started in Chapter 2. Like the latter, the present chapter is somewhat technical.

We will show that the existence of an *ordered* coupling between random variables or random processes is *equivalent* to the respective probability measures being ordered themselves. In Sections 7.1 we look at fully ordered state spaces, in Section 7.2 at partially ordered state spaces. In Section 7.3 we state and derive the Fortuin-Kasteleyn-Ginibre inequality, in Section 7.4 the Holley inequality. Both are inequalities for expectations of functions on partially ordered state spaces.

7.1 Fully ordered state spaces

Let $\mathbb{P}, \mathbb{P}'$ be two probability measures on $\mathbb{R}$ such that

$$\mathbb{P}([x,\infty)) \leq \mathbb{P}'([x,\infty)) \qquad \forall\, x \in \mathbb{R},$$

We say that $\mathbb{P}'$ *stochastically dominates* $\mathbb{P}$, and write $\mathbb{P} \preceq \mathbb{P}'$. In terms of the respective cumulative distribution functions F, F', defined by $F(x) = \mathbb{P}((-\infty, x])$ and $F'(x) = \mathbb{P}'((-\infty, x])$, $x \in \mathbb{R}$, this property is the same as

$$F'(x) \leq F(x) \qquad \forall\, x \in \mathbb{R},$$

i.e., $F' \leq F$ pointwise.

Theorem 7.1 *Let X, X' be $\mathbb{R}$-valued random variables with probability measures $\mathbb{P}, \mathbb{P}'$. If $\mathbb{P} \preceq \mathbb{P}'$, then there exists a coupling $(\hat{X}, \hat{X}')$ of X and X' with probability measure $\hat{\mathbb{P}}$ such that*

$$\hat{\mathbb{P}}(\hat{X} \leq \hat{X}') = 1.$$

Proof. The proof provides an explicit coupling of X and X'. Let F^*, F'^* denote the generalized inverse of F, F' defined by

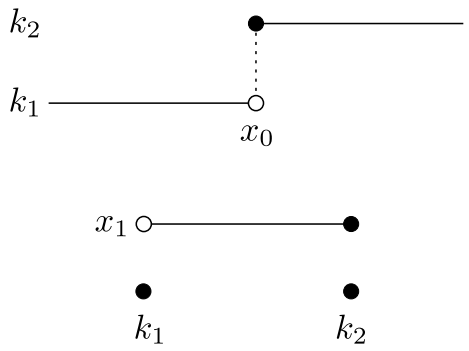
$$\begin{aligned} F^*(u) &= \inf\{x \in \mathbb{R} \colon F(x) \geq u\}, \\ F'^*(u) &= \inf\{x \in \mathbb{R} \colon F'(x) \geq u\}, \end{aligned} \qquad u \in (0,1).$$

Let $U = \mathrm{UNIF}(0,1)$, and put

$$\hat{X} = F^*(U), \qquad \hat{X}' = F'^*(U).$$

Then $\hat{X} \stackrel{D}{=} X$, $\hat{X}' \stackrel{D}{=} X'$, and $\hat{X} \leq \hat{X}'$ because $F' \leq F$ implies $F^* \leq F'^*$. This construction, via a common U, provides the desired coupling. ∎

If F has a point mass $(k_2 - k_1)\delta_{x_0}$ for some $k_2 > k_1$ and $x_0 \in \mathbb{R}$, then this pointmass gives rise to a flat piece in F^* over the interval $(k_1, k_2]$ at height x_1 that solves $F(x_1) = k_2$.



Exercise 7.2 (Examples 2.5–2.6 repeated) *Let U, V be the random variables in Exercises* 2.5–2.6. *Give a coupling of U and V such that $\{U \leq V\}$ with probability 1.*

Theorem 7.3 *If $\mathbb{P} \preceq \mathbb{P}'$, then*

$$\int_{\mathbb{R}} f d\mathbb{P} \leq \int_{\mathbb{R}} f d\mathbb{P}'$$

for all $f\colon \mathbb{R} \to \mathbb{R}$ that are measurable, bounded and non-decreasing.

Proof. Use the coupling in Theorem 7.1 to obtain

$$\int_{\mathbb{R}} f d\mathbb{P} = \mathbb{E}(f(X)) = \hat{\mathbb{E}}(f(\hat{X})) \leq \hat{\mathbb{E}}(f(\hat{X}')) = E'(f(X')) = \int_{\mathbb{R}} f d\mathbb{P}'.$$

■

Actually, the converses of Theorems 7.1 and 7.3 are also true, as is easily seen by picking sets $[x, \infty)$ and functions $x \mapsto 1_{[x,\infty)}$ for $x \in \mathbb{R}$. Therefore the following equivalence holds:

Theorem 7.4 *The three statements*
1. $\mathbb{P} \preceq \mathbb{P}'$,
2. $\exists \hat{\mathbb{P}}\colon \hat{\mathbb{P}}(\hat{X} \leq \hat{X}') = 1$,
3. $\int_{\mathbb{R}} f d\mathbb{P} \leq \int_{\mathbb{R}} f d\mathbb{P}'$ *for all f measurable, bounded and non-decreasing,*

are equivalent.

Exercise 7.5 *Prove the converse of Theorems* 7.1 *and* 7.3.

7.2 Partially ordered state spaces

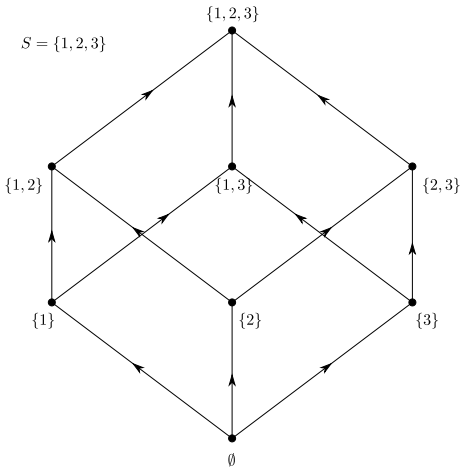
What we did in Section 7.1 can be extended to partially ordered state spaces.

7.2.1 Ordering for probability measures

We will show that the equivalence in Theorem 7.4 continues to hold for more general state spaces, provided it is possible to put a partial ordering on them. In what follows, E is Polish and $\mathcal{E}$ is the σ-algebra of Borel subsets of E.

Definition 7.6 *A relation $\preceq$ on a space E is called a partial ordering if*
1. $x \preceq x$,
2. $x \preceq y$, $y \preceq z \implies x \preceq z$,
3. $x \preceq y$, $y \preceq x \implies x = y$,

where x, y, z are generic elements of E.



Definition 7.7 *Given two probability measures $\mathbb{P}, \mathbb{P}'$ on E, we say that $\mathbb{P}'$ stochastically dominates $\mathbb{P}$, and write $\mathbb{P} \preceq \mathbb{P}'$, if*

$$\mathbb{P}(A) \leq \mathbb{P}'(A) \text{ for all } A \in \mathcal{E} \text{ non-decreasing},$$

where A non-decreasing means

$$x \in A \Longrightarrow A \supset \{y \in E\colon\ x \preceq y\},$$

or equivalently if

$$\int_E f d\mathbb{P} \leq \int_E f d\mathbb{P}' \text{ for all } f\colon\ E \to \mathbb{R} \text{ measureable, bounded and non-decreasing},$$

where f non-decreasing means

$$x \preceq y \Longrightarrow f(x) \leq f(y).$$

The following result is known as *Strassen's theorem*.

Theorem 7.8 *If $\mathbb{P} \preceq \mathbb{P}'$, then there exists a coupling $\hat{\mathbb{P}}$ of $(\mathbb{P}, \mathbb{P}')$ such that*

$$\hat{\mathbb{P}}\big(\{(x, x') \in E^2\colon\ x \preceq x'\}\big) = 1.$$

Proof. Intuitively the result is plausible: if $\mathbb{P}'$ stochastically dominates $\mathbb{P}$, then $\mathbb{P}'$ can be obtained from $\mathbb{P}$ by "moving mass upwards in the partial ordering". However, the technicalities are far from trivial. We refer to Lindvall [11], Section IV.1, for the full proof. ∎

The analogue of Theorem 7.4 reads:

Theorem 7.9 *The three statements*
1. $\mathbb{P} \preceq \mathbb{P}'$,
2. $\exists\, \hat{\mathbb{P}}\colon\ \hat{\mathbb{P}}(\hat{X} \preceq \hat{X}') = 1$,
3. $\int_E f d\mathbb{P} \leq \int_E f d\mathbb{P}'$ *for all f measurable, bounded and non-decreasing,*

are equivalent.

Examples:

- $E = \{0,1\}^{\mathbb{Z}}$, $x = (x_i)_{i\in\mathbb{Z}} \in E$, $x \preceq y$ if and only if $x_i \le y_i$ for all $i \in \mathbb{Z}$. For $p \in [0,1]$, let $\mathbb{P}_p$ denote the probability measure on E under which $X = (X_i)_{i\in\mathbb{Z}}$ has i.i.d. BER(p) components. Then $\mathbb{P}_p \preceq \mathbb{P}_{p'}$ if and only if $p \le p'$.

- It is possible to build in dependency. For instance, let $Y = (Y_i)_{i\in\mathbb{Z}}$ be defined by

$$Y_i = 1_{\{X_{i-1}=X_i=1\}},$$

  and let $\widetilde{\mathbb{P}}_p$ be the law of Y induced by the law $\mathbb{P}_p$ of X. Then the components of Y are not independent, but again $\widetilde{\mathbb{P}}_p \preceq \widetilde{\mathbb{P}}_{p'}$ if and only if $p \le p'$.

Exercise 7.10 *Prove the last two claims.*

More examples will be encountered in Chapter 9.

Exercise 7.11 *Does $\preceq$ in Definition 7.7 define a partial ordering on the space of probability measures?*

7.2.2 Ordering for Markov chains

The notions of partial ordering and stochastic domination are important also for *Markov chains*. Let E be a polish space equipped with a partial ordering $\preceq$. A *transition kernel* K on $E \times E$ is a mapping from $E \times \mathcal{E}$ to $[0,1]$ such that:

1. $K(x,\cdot)$ is a probability measure on E for every $x \in E$;
2. $K(\cdot, A)$ is a measurable mapping from E to $[0,1]$ for every $A \in \mathcal{E}$.

The meaning of $K(x,A)$ is the probability for the Markov chain to jump from x into A. An example is

$$E = \mathbb{R}^d, \quad K(x,A) = \frac{1}{|B_1(x)|}\,|B_1(x) \cap A|,$$

which corresponds to a "Lévy flight" on $\mathbb{R}^d$, i.e., a random walk that makes i.i.d. jumps drawn randomly from the unit ball $B_1(0)$ around the origin. The special case where E is a countable set leads to transition kernels taking the form $K(i,A) = \sum_{j\in A} P_{ij}$, $i \in E$, for some transition matrix $P = (P_{ij})_{i,j\in E}$.

Definition 7.12 *Given two transition kernels K and K' on $E \times E$, we say that K' stochastically dominates K if*

$$K(x,\cdot) \preceq K'(x',\cdot) \text{ for all } x \preceq x'.$$

If $K = K'$ and the latter condition holds, then we say that K is monotone.

Remark: Not all transition kernels are monotone, which is why we cannot write $K \preceq K'$ for the property in Definition 7.12, i.e., there is no partial ordering on the set of transition kernels.

Lemma 7.13 *If $\lambda \preceq \mu$ and K' stochastically dominates K, then*

$$\lambda K^n \preceq \mu K'^n \text{ for all } n \in \mathbb{N}_0.$$

Proof. The proof is by induction on n. The ordering holds for $n = 0$. Suppose that the ordering holds for n. Let f be an arbitrary bounded and non-decreasing function on E^{n+2}. Then

$$\begin{aligned}&\int_{E^{n+2}} f(x_0,\dots,x_n,x_{n+1})(\lambda K^{n+1})(dx_0,\dots,dx_n,dx_{n+1})\\ &\quad=\int_{E^{n+1}}(\lambda K^n)(dx_0,\dots,dx_n)\int_E f(x_0,\dots,x_n,x_{n+1})K(x_n,dx_{n+1}),\end{aligned}\tag{7.1}$$

where $(\lambda K^n)(dx_0,\dots,dx_n)$ is an abbreviation for $\lambda(dx_0)K(x_0,dx_1)\times\cdots\times K(x_{n-1},dx_n)$. The last integral is a function of $x_0,\dots,x_n$. Since f is non-decreasing and K' stochastically dominates K, this integral is bounded from above by

$$\int_E f(x_0,\dots,x_n,x_{n+1})K'(x_n,dx_{n+1}),\tag{7.2}$$

where we use Definitions 7.7 and 7.12.

Exercise 7.14 *Check the above computation.*

Since the ordering holds for n and (7.2) is a non-decreasing function of $(x_0,\dots,x_n)$, the right-hand side of (7.1) is bounded from above by

$$\int_{E^{n+1}}(\mu K'^n)(dx_0,\dots,dx_n)\int_E f(x_0,\dots,x_n,x_{n+1})K'(x_n,dx_{n+1}),$$

which equals

$$\int_{E^{n+2}} f(x_0,\dots,x_n,x_{n+1})(\mu K'^{n+1})(dx_0,\dots,dx_n,dx_{n+1}).$$

This proves the claim by Definition 7.6. ∎

By using the *Kolmogorov extension theorem*, the result in Lemma 7.13 can be extended to $n=\infty$, i.e., the ordering also holds for infinite sequences. This has the following consequence.

Theorem 7.15 *If $\lambda\preceq\mu$ and K' stochastically dominates K, then there exist E-valued random processes*

$$Z=(Z_n)_{n\in\mathbb{N}_0},\quad Z'=(Z'_n)_{n\in\mathbb{N}_0},$$

such that

$$\begin{aligned}(Z_0,\dots,Z_n)&\stackrel{D}{=}\lambda K^n,\\ (Z'_0,\dots,Z'_n)&\stackrel{D}{=}\mu K'^n,\end{aligned}\qquad\forall\, n\in\mathbb{N}_0,$$

and $Z_0\preceq Z'_0$, $Z_1\preceq Z'_1$, ... a.s. w.r.t. the joint law of (Z,Z').

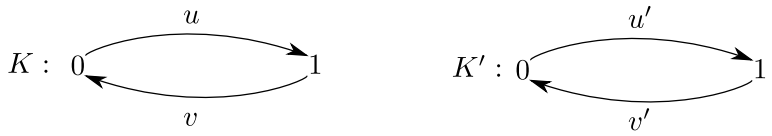
Remark: The last ordering is denoted by $Z\preceq_\infty Z'$. All components are ordered w.r.t. $\preceq$.

Examples:

1. $E=\mathbb{R}$, $\preceq$ becomes $\leq$. The result says that if $\lambda\stackrel{D}{\leq}\mu$ and $K(x,\cdot)\stackrel{D}{\leq}K'(x,\cdot)$ for all $x\leq x'$, then the two Markov chains on $\mathbb{R}$ can be coupled so that they are *ordered for all times*.
2. $E=\{0,1\}^{\mathbb{Z}}$. Think of an infinite sequence of lamps, labelled by $\mathbb{Z}$, that can be either "off" or "on". The initial distributions are $\lambda=\mathbb{P}_p$ and $\mu=\mathbb{P}_{p'}$ with $p<p'$. The transition kernels K and K' are such that the lamps *change their state independently* at rates

$$\begin{aligned}K:&\quad 0\stackrel{u}{\longrightarrow}1,\quad 1\stackrel{v}{\longrightarrow}0,\\ K':&\quad 0\stackrel{u'}{\longrightarrow}1,\quad 1\stackrel{v'}{\longrightarrow}0,\end{aligned}$$

with $u'>u$ and $v'<v$, i.e., K' flips more rapidly on and less rapidly off compared to K.



Exercise 7.16 *Give an example where the flip rate of a lamp depends on the states of the two neighboring lamps.*

7.3 The FKG inequality

Let S be a finite set and let $\mathcal{P}(S)$ be the set of all subsets of S (called the power set of S). Then $\mathcal{P}(S)$ is partially ordered by inclusion. A probability measure μ on $\mathcal{P}(S)$ is called *log-convex* if

$$\mu(a \cup b)\mu(a \cap b) \geq \mu(a)\mu(b) \qquad \forall\, a, b \in \mathcal{P}(S). \tag{7.3}$$

A function f on $\mathcal{P}(S)$ is called *non-decreasing* if

$$f(b) \geq f(a) \qquad \forall\, a, b \in \mathcal{P}(S) \text{ with } a \subset b. \tag{7.4}$$

Abbreviate $\mu[f] = \sum_{a \in \mathcal{P}(S)} f(a)\mu(a)$ for the expectation of f under μ.

Theorem 7.17 (Fortuin-Kastelyn-Ginibre inequality) *If μ is log-convex and f, g are non-decreasing, then*

$$\mu[fg] \geq \mu[f]\mu[g].$$

Proof. The following proof is taken from den Hollander and Keane [6] and proceeds via induction on $|S|$. The claim is trivially true when $|S| = 1$. Suppose that the claim holds for all S with $|S| \leq n$. Let $|S| = n + 1$, pick an element $s \in S$, put $S' = S \setminus \{s\}$ and, for $a \in \mathcal{P}(S')$, let

$$\begin{aligned}
\mu'(a) &= \mu(a) + \mu(a + \{s\}), \\
f'(a) &= \frac{1}{\mu'(a)} \left[f(a)\mu(a) + f(a \cup \{s\})\mu(a \cup \{s\}) \right], \\
g'(a) &= \frac{1}{\mu'(a)} \left[g(a)\mu(a) + g(a \cup \{s\})\mu(a \cup \{s\}) \right],
\end{aligned}$$

i.e., μ' is the marginal of μ on S', and f' and g' are the conditional expectations with respect to μ given the value on S'. To proceed with the proof we need the following lemma.

Lemma 7.18 *Let s_1, s_2, s_3, s_4 and t_1, t_2, t_3, t_4 be non-negative reals such that*

$$s_1 s_2 \geq t_1 t_2, \qquad s_3 s_4 \geq t_3 t_4, \qquad s_2 s_3 \geq t_1 t_4 \vee t_2 t_3.$$

Then $(s_1 + s_3)(s_2 + s_4) \geq (t_1 + t_3)(t_2 + t_4)$.

Exercise 7.19 *Check this lemma.*

The proof continues in three steps:

Step 1: *μ' is log-convex on $\mathcal{P}(S')$:*

Use (7.3) and Lemma 7.18 with $a, b \in \mathcal{P}(S')$ and

$$\begin{array}{ll} s_1 = \mu(a \cup b) & t_1 = \mu(a) \\ s_2 = \mu(a \cap b) & t_2 = \mu(b) \\ s_3 = \mu([a \cup b] \cup \{s\}) & t_3 = \mu(a \cup \{s\}) \\ s_4 = \mu([a \cap b] \cup \{s\}) & t_4 = \mu(b \cup \{s\}) \end{array}$$

to obtain $\mu'(a \cup b)\mu'(a \cap b) \geq \mu'(a)\mu'(b)$.

Exercise 7.20 *Check the latter inequality.*

Step 2: *f', g' are non-decreasing on $\mathcal{P}(S')$:*

For $a, b \in \mathcal{P}(S')$ with $a \subset b$, write

$$\begin{aligned} f'(b) - f'(a) &= \frac{1}{\mu'(a)\mu'(b)} \Big\{ [\mu(a) + \mu(a \cup \{s\})][f(b)\mu(b) + f(b \cup \{s\})\mu(b \cup \{s\})] \\ &\qquad - [\mu(b) + \mu(b \cup \{s\})][f(a)\mu(a) + f(a \cup \{s\})\mu(a \cup \{s\})] \Big\} \\ &= \frac{1}{\mu'(a)\mu'(b)} [\mu(a) + \mu(a \cup \{s\})] \\ &\qquad \times \Big\{ \underbrace{[f(b) - f(a)]}_{\geq 0} \mu(b) + \underbrace{[f(b \cup \{s\}) - f(a \cup \{s\})]}_{\geq 0} \mu(b \cup \{s\}) \Big\} \\ &\qquad\quad + \underbrace{[f(a \cup \{s\}) - f(a)]}_{\geq 0} \underbrace{[\mu(a)\mu(b \cup \{s\}) - \mu(a \cup \{s\})\mu(b)]}_{\geq 0} \geq 0. \end{aligned}$$

The right-hand side is a sum of products of non-negative terms (use (7.3–7.4)), and so $f'(b) \geq f'(a)$.

Step 3: *$\mu[fg] \geq \mu'[f'g']$:*

Write

$$\mu[fg] = \sum_{a \in \mathcal{P}(S)} (fg)(a)\mu(a) = \sum_{a \in \mathcal{P}(S')} (fg)'(a)\mu'(a),$$

and use that

$$\begin{aligned} &\mu'(a)^2 \left[(fg)'(a) - f'(a)g'(a) \right] \\ &\quad = [\mu(a) + \mu(a \cup \{s\})] \left[(fg)(a)\mu(a) + (fg)(a \cup \{s\})\mu(a \cup \{s\}) \right] \\ &\qquad - [f(a)\mu(a) + f(a \cup \{s\})\mu(a \cup \{s\})] \left[g(a)\mu(a) + g(a \cup \{s\})\mu(a \cup \{s\}) \right] \\ &\quad = \mu(a)\mu(a \cup \{s\}) \underbrace{[f(a \cup \{s\}) - f(a)]}_{\geq 0} \underbrace{[g(a \cup \{s\}) - g(a)]}_{\geq 0} \geq 0. \end{aligned}$$

Hence

$$\mu[fg] \geq \sum_{a \in \mathcal{P}(S)} f'(a)g'(a)\mu'(a).$$

By the induction assumption in combination with Steps 1 and 2, we have

$$\mu'[f'g'] \geq \mu'[f']\mu'[g'].$$

But $\mu'[f'] = \mu[f]$ and $\mu'[g'] = \mu[g]$, and so with Step 3 we are done when $\mu > 0$ on $\mathcal{P}(S)$. ∎

Exercise 7.21 *Explain how to remove the restriction that $\mu > 0$ on $\mathcal{P}(S)$.*

Remark: By taking a "projective limit" $|S| \to \infty$, it is trivial to extend Theorem 7.17 to countable sets S. The inequality in (7.3) must then be assumed for arbitrary *cylinder sets*. It is even possible to extend to uncountable sets S.

Remark: The condition of log-convexity of μ is not necessary on fully ordered spaces. Indeed, pick $S = \mathbb{R}$, let f, g be any two non-decreasing functions on $\mathbb{R}$, and write

$$\begin{aligned}\mu[fg] - \mu[f]\mu[g] &= \int_{\mathbb{R}} \mu(dx)\, f(x)g(x) - \int_{\mathbb{R}} \mu(dx)\, f(x) \int_{\mathbb{R}} \mu(dy)\, g(y) \\ &= \tfrac{1}{2} \int_{\mathbb{R}} \mu(dx) \int_{\mathbb{R}} \mu(dy)\ \underbrace{[f(x) - f(y)][g(x) - g(y)]}_{\geq 0} \geq 0.\end{aligned}$$

The two factors in the integrand are either both ≥ 0 or both ≤ 0, and hence $\mu[fg] \geq \mu[f]\mu[g]$.

Remark: The intuition behind log-convexity is the following. First, note that the inequality in (7.3) holds for all $a, b \in \mathcal{P}(S)$ if and only if

$$\frac{\mu(a \cup \{s\})}{\mu(a)} \geq \frac{\mu(\{s\})}{\mu(\emptyset)} \qquad \forall\, a \in \mathcal{P}(S),\ s \in S\backslash a. \tag{7.5}$$

Next, let $X \in \mathcal{P}(S)$ be the random variable with distribution $\mathbb{P}(X = a) = \mu(a)$, $a \in \mathcal{P}(S)$. Define

$$p(a, \{s\}) = \mathbb{P}\big(s \in X \mid X \cap S\backslash\{s\} = a\big), \qquad \forall\, a \in \mathcal{P}(S),\ s \in S\backslash a, \tag{7.6}$$

and note that

$$p(a, \{s\}) = \left(1 + \left(\frac{\mu(a \cup \{s\})}{\mu(a)}\right)^{-1}\right)^{-1}.$$

Therefore (7.5) is the same as

$$p(a, \{s\}) \geq p(\emptyset, \{s\}) \qquad a \in \mathcal{P}(S),\ s \in S\backslash a.$$

In view of (7.6), the latter says: "larger X are more likely to contain an extra point than smaller X", a property referred to as "attractiveness".

Example: [Percolation model]
Take S to be a finite set in $\mathbb{Z}^d$, $\mathcal{P}(S) = \{0, 1\}^S$,

$$\mu(a) = p^{|a|}(1 - p)^{|S\backslash a|} \text{ with } p \in (0, 1), \tag{7.7}$$

$A, B \subset S$ and $f(\cdot) = 1_{\{\cdot \supset A\}}$, $g(\cdot) = 1_{\{\cdot \supset B\}}$. Then

$$\mu(\text{all } 1's \text{ on } A \cup B) \geq \mu(\text{all } 1's \text{ on } A)\mu(\text{all } 1's \text{ on } B). \tag{7.8}$$

Exercise 7.22 *Prove* (7.8) *by checking that μ is log-convex.*

Under the probability distribution in (7.7), each site in S carries a particle (= 1) with probability p or a vacancy (= 0) with probability $1 - p$, independently for different sites. This is referred to as percolation and will be treated in more detail in Chapter 8.

Example: [Ising model]
Take S to be a finite torus in $\mathbb{Z}^d$ (with periodic boundary conditions), $\mathcal{P}(S) = \{0,1\}^S$,

$$\begin{aligned}&\mu(a) = \tfrac{1}{Z_\beta} \exp\left[\beta|\{x,y \in a\colon \|x-y\| = 1\}|\right] \text{ with } \beta \in (0,\infty),\\ &\text{where } Z_\beta \text{ is the normalizing constant,}\end{aligned} \tag{7.9}$$

$A, B \subset S$ and $f(\cdot) = 1_{\{\cdot \supset A\}}$, $g(\cdot) = 1_{\{\cdot \supset B\}}$.

Exercise 7.23 *Prove* (7.8) *by checking that μ is log-convex.*

The probability distribution in (7.9) gives a probabilistic reward e^β to every pair of 1's in S that are located at nearest-neighbor sites. It is used in statistical physics to describe a system consisting of particles that have *a tendency to stick to each other when they are close to each other*, due to the so-called van der Waals force (1 = particle, 0 = vacancy). The parameter β plays the role of the "inverse temperature": the lower the temperature, the larger β, and hence the larger the tendency to stick to each other. This is referred to as ferromagnetism and will be treated in more detail in Chapter 9.

7.4 The Holley inequality

A variant of the FKG-inequality is the following. Given two probability measures μ_1, μ_2 on $\mathcal{P}(S)$, we say that μ_1 is *log-convex with respect to* μ_2 if

$$\mu_1(a \cup b)\mu_2(a \cap b) \geq \mu_1(a)\mu_2(b) \qquad \forall\, a, b \in \mathcal{P}(S). \tag{7.10}$$

Note that a probability measure μ on $\mathcal{P}(S)$ is log-convex with respect to itself if and only it is log-convex in the sense of (7.3).

Theorem 7.24 *If μ_1 is log-convex with respect to μ_2 and f is non-decreasing, then*

$$\mu_1[f] \geq \mu_2[f].$$

Proof. See den Hollander and Keane [6] for a proof similar to that of Theorem 7.17. Here we give a proof that uses Lemma 7.13. Again we assume that $\mu_1, \mu_2 > 0$ on $\mathcal{P}(S)$, a restriction that is easily removed afterwards.

We construct a coupling of two continuous-time Markov chains

$$\eta = (\eta_t)_{t \geq 0}, \qquad \zeta = (\zeta_t)_{t \geq 0},$$

on $\mathcal{P}(S)$, with S finite, such that:
(1) η has stationary distribution μ_2,
(2) ζ has stationary distribution μ_1,
(3) the coupling prevents the pair (η, ζ) to exit the set $\{(a,b) \in \mathcal{P}(S)^2\colon\ a \subset b\}$.

The rates of the coupled Markov chain are chosen as follows. For $s \in S$, let η^s denote the element of $\mathcal{P}(S) = \{0,1\}^S$ obtained from η by flipping the variable at s (either $0 \to 1$ or $1 \to 0$). Allow only the following transitions:

$$(\eta,\zeta) \to (\eta^s,\zeta) \quad \text{at rate } \begin{cases} 1 & \text{if } (\eta(s),\zeta(s)) = (0,1),\\ \frac{\mu_2(\eta^s)}{\mu_2(\eta)} - \frac{\mu_1(\zeta^s)}{\mu_1(\zeta)} & \text{if } (\eta(s),\zeta(s)) = (1,1),\end{cases}$$

$$(\eta,\zeta) \to (\eta,\zeta^s) \quad \text{at rate } \tfrac{\mu_1(\zeta^s)}{\mu_1(\zeta)} \text{ if } (\eta(s),\zeta(s)) = (0,1),$$

$$(\eta,\zeta) \to (\eta^s,\zeta^s) \quad \text{at rate } \begin{cases} 1 & \text{if } (\eta(s),\zeta(s)) = (0,0),\\ \frac{\mu_1(\zeta^s)}{\mu_1(\zeta)} & \text{if } (\eta(s),\zeta(s)) = (1,1).\end{cases}$$

Exercise 7.25 *Check property* (3) *by showing that the allowed transitions preserve the ordering of the Markov chains, i.e., if* $\eta \subseteq \zeta$*, then the same is true after every allowed transition. Consequently,*

$$\eta_0 \subseteq \zeta_0 \quad \Longrightarrow \quad \eta_t \subseteq \zeta_t \; \forall \, t > 0. \tag{7.11}$$

Check properties (1) *and* (2)*. Condition* (7.10) *is needed to ensure that*

$$\frac{\mu_2(\eta^s)}{\mu_2(\eta)} \geq \frac{\mu_1(\zeta^s)}{\mu_1(\zeta)} \text{ when } \eta \subseteq \zeta \text{ with } (\eta(s), \zeta(s)) = (1,1).$$

From (7.11) we get $\mathbb{E}_{\eta_0}(f(\eta_t)) \leq \mathbb{E}_{\zeta_0}(f(\zeta_t))$ for all $t \geq 0$ when $\eta_0 \subset \zeta_0$, and the Holley inequality follows because $\mathbb{E}_{\eta_0}(f(\eta_t)) \to \mu_2[f]$ and $\mathbb{E}_{\zeta_0}(f(\zeta_t)) \to \mu_1[f]$ as $t \to \infty$. Pick $\eta_0 = \emptyset$ and $\zeta_0 = S$ to make sure that $\eta_0 \subseteq \zeta_0$. ∎

Remark: The coupling used in the above proof is a *maximal coupling* in the sense of Section 2.5.

Remark: By viewing the above rates locally, we can extend the Holley inequality to countable sets S via a "projective limit" argument. The inequality in (7.10) must then be assumed for arbitrary cylinder sets. It is even possible to extend to uncountable sets S.

What is important about Theorem 7.24 is that it provides an *explicit criterion* on μ_1, μ_2 such that $\mu_2 \preceq \mu_1$, as is evident from Theorem 7.9. Note that "log-convex with respect to" is not a partial ordering: as noted above, μ is log-convex with respect to itself if and only if it is log-convex. In particular, the reverse of Theorem 7.24 is false.

Exercise 7.26 *Return to the example of the Ising model at the end of Section* 7.3*. Pick* $\beta_1 > \beta_2$*, and let* $\mu_i = \mu_{\beta_i}$*,* $i = 1, 2$*, with* μ_β *the probability measure in* (7.9)*. Show that* μ_{β_1} *is log-convex with respect to* μ_{β_2}*.*

Remark: FKG follows from Holley by choosing

$$\mu_1 = \frac{\mu g}{\mu[g]}, \qquad \mu_2 = \mu. \tag{7.12}$$

Exercise 7.27 *Check this claim.*

8 Percolation

In Sections 8.1 we look at ordinary percolation on $\mathbb{Z}^d$, in Section 8.2 at invasion percolation on $\mathbb{Z}^d$. In Section 8.3 we take a closer look at invasion percolation on regular trees, where explicit computations can be carried through.

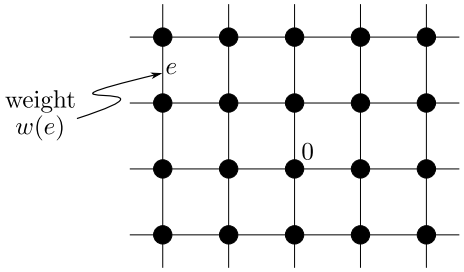
A standard reference for percolation theory is Grimmett [4].

8.1 Ordinary percolation

Consider the d-dimensional integer lattice $\mathbb{Z}^d$, $d \geq 2$. Draw edges between neighboring sites. Associate with each edge e a random variable $w(e)$, drawn *independently* from UNIF$(0,1)$. This gives

$$w = (w(e))_{e \in (\mathbb{Z}^d)^*},$$

where $(\mathbb{Z}^d)^*$ is the set of edges.



Pick $p \in [0,1]$, and partition $\mathbb{Z}^d$ into *p-clusters* by connecting all sites that are connected by edges whose weight is $\leq p$, i.e.,

$$x \overset{p}{\longleftrightarrow} y$$

if and only if there is a path π connecting x and y such that $w(e) \leq p$ for all $e \in \pi$. (A path is a collection of neighboring sites connected by edges.) Let $C_p(0)$ denote the p-cluster containing the *origin*, and define

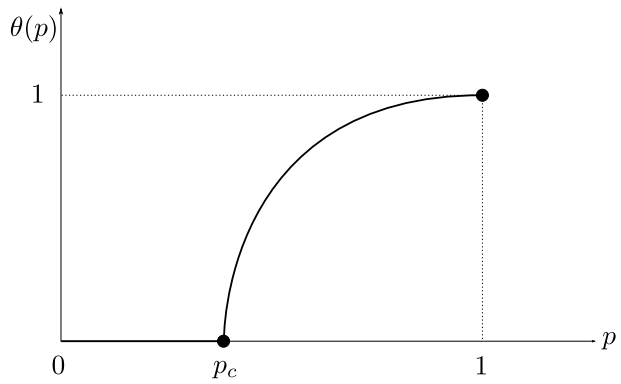
$$\theta(p) = \mathbb{P}(|C_p(0)| = \infty)$$

with $\mathbb{P}$ denoting the law of w. Clearly,

$$C_0(0) = \{0\}, \quad C_1(0) = \mathbb{Z}^d, \quad p \mapsto C_p(0) \text{ is non-decreasing},$$

so that

$$\theta(0) = 0, \quad \theta(1) = 1, \quad p \mapsto \theta(p) \text{ is non-decreasing}.$$



Define

$$p_c = \sup\{p \in [0,1]\colon\ \theta(p) = 0\}.$$

It is known that $p_c \in (0,1)$ (for $d \geq 2$), and that $p \mapsto \theta(p)$ is continuous for all $p \neq p_c$. Continuity is expected to hold also at $p = p_c$, but this has only been proved for $d = 2$ and $d \geq 19$. It is further known that $p_c = \frac{1}{2}$ for $d = 2$, while no explicit expression for p_c is known for $d \geq 3$. There are good numerical approximations available for p_c, as well as expansions in powers of $\frac{1}{2d}$ for d large.

At $p = p_c$ a *phase transition* occurs:

$$\begin{array}{ll} p < p_c\colon & \text{all clusters are } \textit{finite}, \\ p > p_c\colon & \text{there are } \textit{infinite} \text{ clusters.} \end{array}$$

It is known that in the supercritical phase there is a *unique* infinite cluster.

Exercise 8.1 *Why is the uniqueness not obvious?*

Remark: Note that the $C_p(0)$'s for different p's are *coupled* because we use the same w for all of them. Indeed, we have

$$C_p(0) \subseteq C_{p'}(0) \text{ when } p < p'.$$

With $\preceq$ the parial ordering on $\{0,1\}^{\mathbb{Z}^d}$ obtained by inclusion, the random fields $X = (X_z)_{z\in\mathbb{Z}^d}$ and $X' = (X'_z)_{z\in\mathbb{Z}^d}$ defined by

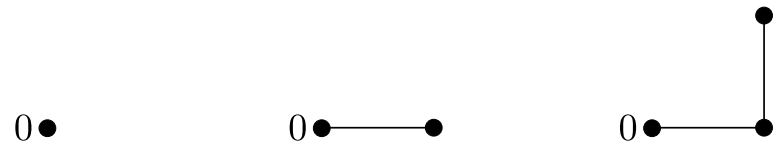
$$X_z = 1_{\{z\in C_p(0)\}}, \qquad X'_z = 1_{\{z\in C_{p'}(0)\}},$$

satisfy $X \preceq X'$ when $p < p'$.

8.2 Invasion percolation

Again consider $\mathbb{Z}^d$ and $(\mathbb{Z}^d)^*$ with the random field of weights w. Grow a cluster from 0 as follows:

1. Invade the origin: $I(0) = \{0\}$.
2. Look at all the edges touching $I(0)$, choose the edge with the smallest weight, and invade the vertex at the other end: $I(1) = \{0, x\}$, with $x = \text{argmin}_{y:\ \|y\|=1} W(\{0,y\})$.
3. Repeat 2 with $I(1)$ replacing $I(0)$, etc.



In this way we obtain a sequence of growing sets $I = (I(n))_{n\in\mathbb{N}_0}$ with $I(n) \subset \mathbb{Z}^d$ and $|I(n)| \leq n+1$. (The reason for the inequality is that the vertex at the other end may have been invaded before. The set of invaded edges at time n has cardinality n.) The *invasion percolation cluster* is defined as

$$C_{\mathrm{IPC}} = \lim_{n\to\infty} I(n).$$

This is an *infinite* subset of $\mathbb{Z}^d$, which is random because w is random. Note that the sequence I is uniquely determined by w (because no two edges have the same weight).

Remark: Invasion percolation may serve as a model for the spread of a virus through a computer network: the virus is "greedy" and invades the network along the weakest links.

The first question we may ask is whether $C_{\mathrm{IPC}} = \mathbb{Z}^d$. The answer is no:

$$C_{\mathrm{IPC}} \subsetneq \mathbb{Z}^d \quad a.s.$$

In fact, C_{IPC} turns out to be a *thin set*, in the sense that

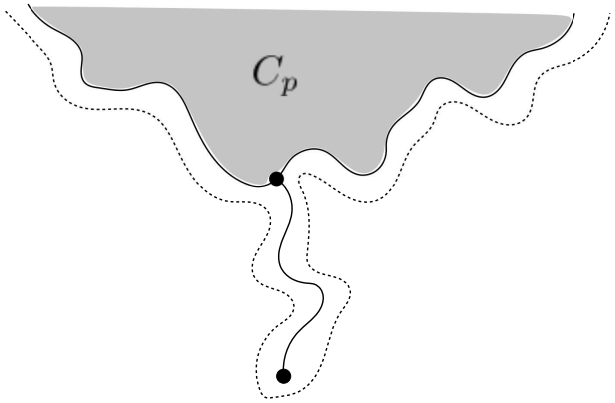
$$\lim_{N\to\infty} \frac{1}{|B_N|} |B_N \cap C_{\mathrm{IPC}}| = 0 \quad a.s. \quad \text{with } B_N = [-N, N]^d \cap \mathbb{Z}^d. \tag{8.1}$$

A key result for invasion percolation is the following. Let W_n denote the weight of the edge that is traversed in the n-th step of the growth of C_{IPC}, i.e., in going from $I(n-1)$ to $I(n)$.

Theorem 8.2 $\limsup_{n\to\infty} W_n = p_c \quad a.s.$

Proof. Pick $p > p_c$. Then the union of all the p-clusters contains a unique infinite component (recall Section 8.1), which we denote by C_p. Note that the asymptotic density of C_p is $\theta(p) > 0$ and that C_p does not necessarily contain the origin. All edges incident to C_p have weight $> p$. Let τ_p denote the first time a vertex in C_p is invaded:

$$\tau_p = \inf\{n \in \mathbb{N}_0 \colon\ I(n) \cap C_p \neq \emptyset\}.$$



We first show that this time is finite a.s.

Lemma 8.3 $\mathbb{P}(\tau_p < \infty) = 1 \quad$ *for all* $p > p_c$.

Proof. Each time I "breaks out" of the box with center 0 it is currently contained in, it sees a "never-before-explored" region containing a half-space. There is an independent probability $\theta(p) > 0$ that it hits C_p at such a break out time. Therefore it will eventually hit C_p with probability 1. (This observation tacitly assumes that $p_c(\mathbb{Z}^d) = p_c(\text{halfspace})$.) ∎

Exercise 8.4 *Work out the details of the proof.*

We proceed with the proof of Theorem 8.2. The edge invaded at time τ_p, being incident to C_p, has weight $> p$. Since the invasion took place along this edge, all edges incident to $I(\tau_p - 1)$ (which includes this edge) have weight $> p$ too. Thus, all edges incident to $I(\tau_p) \cup C_p$ have weight $> p$. However, all edges connecting the vertices of C_p have weight $\leq p$, and so after time τ_p the invasion will be "stuck inside C_p forever". Not only does this show that $C_{\text{IPC}} = I(\tau_p) \cup C_p \subsetneq \mathbb{Z}^d$, it also shows that $W_n \leq p$ for all n large enough a.s. Since $p > p_c$ is arbitrary, it follows that

$$\limsup_{n\to\infty} W_n \leq p_c \quad a.s.$$

In fact, it is trivial to see that equality must hold. Indeed, suppose that $W_n \leq \tilde{p}$ for all n large enough for some $\tilde{p} < p_c$. Then

$$C_{\text{IPC}} \subseteq C_{\tilde{p}} \cup I(\tau(\tilde{p}))$$

with

$$\tau(\tilde{p}) = \inf\big\{m \in \mathbb{N}_0 \colon\ W_n \leq \tilde{p}\ \forall\, n \geq m\big\}.$$

But $|C_{\tilde{p}}| < \infty$ and $|I(\tau(\tilde{p}))| < \infty$ a.s., and this contradicts $|C_{\text{IPC}}| = \infty$. Note that

$$\limsup_{N\to\infty} \frac{1}{|B_N|}\,|B_N \cap C_{\text{IPC}}| \leq \theta(p) \quad a.s. \qquad \forall\, p > p_c,$$

which proves (8.1) because $\lim_{p\downarrow p_c} \theta(p) = 0$. ∎

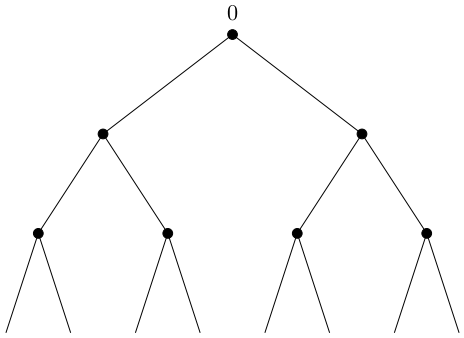
Theorem 8.2 shows that invasion percolation is an example of a stochastic dynamics that exhibits *self-organized criticality*: C_{IPC} is in some sense close to C_{p_c} for ordinary percolation. Informally this can be expressed by writing

$$C_{\text{IPC}} = C_{p_c+} = \lim_{p\downarrow p_c} C_p.$$

Very little is known about the probability distribution of C_{IPC}.

8.3 Invasion percolation on regular trees

If we replace $\mathbb{Z}^d$ by T_σ, the *rooted* tree with branching number $\sigma \in \mathbb{N}\backslash\{1\}$, then a lot can be said about C_{IPC} in detail. What follows is taken from Angel, Goodman, den Hollander and Slade [1].



We again assign independent UNIF$(0,1)$ weights $w = (w(e))_{e \in (T_\sigma)^*}$ to the edges $(T_\sigma)^*$ of the tree, and use this to define ordinary percolation and invasion percolation. We will compare C_{IPC} with the *incipient infinite cluster*, written C_{IIC} and defined informally as

$$C_{\mathrm{IIC}} = C_{p_c} \ \mid \ \{|C_{p_c}| = \infty\},$$

i.e., take the critical cluster of ordinary percolation and condition it to be infinite. A more formal construction is

$$\mathbb{P}(C_{\mathrm{IIC}} \in \cdot\,) = \lim_{n \to \infty} \mathbb{P}(C_{p_c} \in \cdot \mid 0 \leftrightarrow H_n)$$

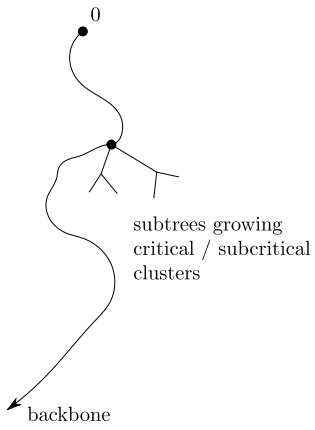
with $H_n \subset T_\sigma$ the set of vertices at height n below the origin. The existence of the limit is far from trivial.

Theorem 8.5 *There exists a coupling of C_{IPC} and C_{IIC} such that $C_{\mathrm{IPC}} \subseteq C_{\mathrm{IIC}}$ a.s.*

Proof. We begin by noting that both C_{IPC} and C_{IIC} consist of a random *back bone* with random *finite branches* hanging off.

IPC: Suppose that with positive probability there is a vertex in C_{IPC} from which there are two disjoint paths to infinity. Conditioned on this event, let M_1 and M_2 denote the *maximal weight* along these paths. It is not possible that $M_1 > M_2$, since this would cause the entire second path to be invaded before the piece of the first path above its maximum weight is invaded. For the same reason $M_1 < M_2$ is not possible either. But $M_1 = M_2$ has probability zero, and so there is a single path to infinity.

IIC: The backbone guarantees the connection to infinity. The cluster is a *critical branching process* with offspring distribution BIN$(\sigma, 1/\sigma)$ conditioned on each generation having at least one child.



We next give *structural representations* of C_{IPC} and C_{IIC}:

Lemma 8.6

C_{IIC}: *The branches hanging off the backbone are critical percolation clusters.*

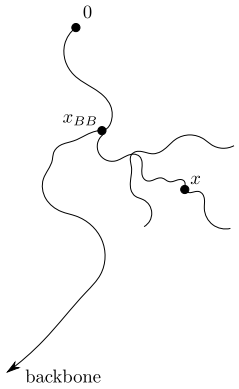
C_{IPC}: *The branches hanging off the backbone at height k are supercritical percolation clusters with parameter $W_k > p_c$ conditioned to be finite, where*

$$W_k = \text{ maximal weight on the backbone above height } k.$$

Proof. We give the proof for C_{IPC} only. By symmetry, all possible backbones are equally likely. Condition on the backbone, abbreviated BB. Conditional on $W = (W_k)_{k\in\mathbb{N}_0}$, the following is true for every vertex $x \in T_\sigma$:

$$x \in C_{\mathrm{IPC}} \iff \text{ every edge on the path between } x_{BB} \text{ and } x \text{ has weight} < W_k,$$

where $x_{BB} = x_{BB}(x)$ is the unique vertex where the path upwards from x to 0 hits BB, and $k = k(x)$ is the height of x_{BB}.



Therefore, the event $\{BB = bb, W = w\}$ is the same as the event that for all $k \in \mathbb{N}_0$ there is no percolation below level W_k (i.e., for p-percolation with $p < W_k$) in each of the branches off BB at height k, and the forward maximal weights along bb are equal to $w = (w_k)_{k\in\mathbb{N}_0}$. ∎

On the tree, there is a nice *duality relation* between subcritical and supercritical percolation.

Lemma 8.7 *A supercritical percolation cluster with parameter $p > p_c$ conditioned to stay finite has the same law as a subcritical percolation cluster with dual parameter $\hat{p} < p_c$ given by*

$$\hat{p} = p\zeta(p)^{\sigma-1}$$

with $\zeta(p) \in (0,1)$ the probability that the cluster along a particular branch from 0 is finite.

Proof. For $v \in T_\sigma$, let $C_p(v)$ denote the forward cluster of v for p-percolation. Let U be any finite subtree of T_σ with, say, m edges, and hence with $(\sigma-1)m+\sigma$ boundary edges. Then

$$\begin{aligned}\mathbb{P}\big(U \subset C_p(v) \mid |C_p(v)| < \infty\big) &= \frac{\mathbb{P}(U \subset C_p(v), |C_p(v)| < \infty)}{\mathbb{P}(|C_p(v)| < \infty)} \\ &= \frac{p^m \zeta(p)^{(\sigma-1)m+\sigma}}{\zeta(p)^\sigma},\end{aligned}$$

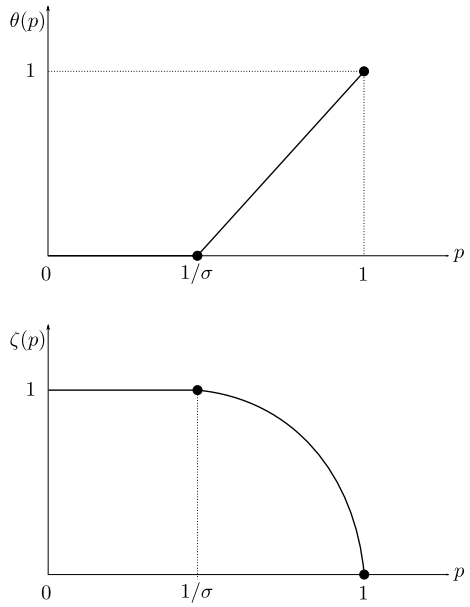
the numerator being the probability of the event that the edges of U are open and there is no percolation from any of the sites in U. The right-hand side equals

$$\hat{p}^m = \mathbb{P}(U \subset C_p(v)),$$

which proves the duality. To see that $p > p_c$ implies $\hat{p} < p_c$, note that

$$\theta(p) = 1 - \zeta(p)^\sigma, \qquad \zeta(p) = 1 - p\theta(p).$$

Exercise 8.8 *Check the latter display.*



We can now complete the proof of $C_{\mathrm{IPC}} \subseteq C_{\mathrm{IIC}}$: since C_{IPC} has subcritical clusters hanging off its backbone, these branches are all *stochastically smaller* than the critical clusters hanging off the backbone of C_{IIC}. The subcritical clusters can be coupled to the critical clusters so that they are contained in them. ∎

9 Interacting particle systems

In Section 9.1 we define what an interacting particle system is. In Sections 9.2–9.3 we focus on shift-invariant spin-flip systems, which constitute a particularly tractable class of interacting particle systems, and look at their convergence to equilibrium. In Section 9.4 we give three examples in this class: Stochastic Ising Model, Contact Process, Voter Model. In Section 9.5 we take a closer look at the Contact Process.

The standard reference for interacting particle systems is Liggett [9].

9.1 Definitions

An *Interacting Particle System* (IPS) is a Markov process $\xi = (\xi_t)_{t\geq 0}$ on the state space $\Omega = \{0,1\}^{\mathbb{Z}^d}$ (or $\Omega = \{-1,1\}^{\mathbb{Z}^d}$), $d \geq 1$, where

$$\xi_t = \{\xi_t(x)\colon\ x \in \mathbb{Z}^d\}$$

denotes the configuration at time t, with $\xi_t(x) = 1$ or 0 meaning that there is a "particle" or "hole" at site x at time t, respectively. Alternative interpretations are

$$\begin{aligned} 1 &= \text{infected/spin-up/democrat} \\ 0 &= \text{healthy/spin-down/republican.} \end{aligned}$$

The configuration changes with time and this models how a virus spreads through a population, how magnetic atoms in iron flip up and down as a result of noise due to temperature, or how the popularity of two political parties evolves in an election campaign.

The evolution is modeled by specifying a set of *local transition rates*

$$c(x,\eta), \qquad x \in \mathbb{Z}^d,\ \eta \in \Omega, \tag{9.1}$$

playing the role of the *rate at which the state at site x changes in the configuration η*, i.e.,

$$\eta \to \eta^x$$

with η^x the configuration obtained from η by changing the state at site x (either $0 \to 1$ or $1 \to 0$). Since there are only two possible states at each site, such systems are called *spin-flip systems*.

Remark: It is possible to allow *more* than two states, e.g. $\{-1,0,1\}$ or $\mathbb{N}_0$. It is also possible to allow *more* than one site to change state at a time, e.g. swapping of states $01 \to 10$ or $10 \to 01$. In what follows we focus entirely on spin-flip systems.

If $c(x,\eta)$ depends on η only via $\eta(x)$, the value of the spin at x, then ξ consists of *independent* spin-flips. In general, however, the rate to flip the spin at x may depend on the spins in the neighborhood of x (possibly even on all spins). This dependence models an "interaction" between the spins at different sites. In order for ξ to be well-defined, some *restrictions* must be placed on the family in (9.1), e.g. $c(x,\eta)$ must depend only "weakly" on the states at "far away" sites (formally, $\eta \mapsto c(x,\eta)$ is continuous in the product topology), and must not be "too large" (formally, bounded away from infinity in some appropriate sense).

9.2 Shift-invariant attractive spin-flip systems

Typically it is assumed that

$$c(x,\eta) = c(x+y,\tau_y\eta) \qquad \forall\, y \in \mathbb{Z}^d \tag{9.2}$$

with τ_y the shift of space over y, i.e., $(\tau_y\eta)(x) = \eta(x-y)$, $x \in \mathbb{Z}^d$. Property (9.2) says that the flip rate at x only depends on the configuration η *as seen relative to* x, which is natural when the interaction between spins is "homogeneous in space". Another useful and frequently used assumption is that the interaction *favors spins that are alike*, i.e.,

$$\eta \preceq \eta' \to \begin{cases} c(x,\eta) \le c(x,\eta') & \text{if } \eta(x) = \eta'(x) = 0, \\ c(x,\eta) \ge c(x,\eta') & \text{if } \eta(x) = \eta'(x) = 1. \end{cases} \tag{9.3}$$

Property (9.3) says that the spin at x flips up faster in η' than in η when η' is everywhere larger than η, but flips down slower. In other words, the dynamics preserves the order $\preceq$. Spin-flip systems with this property are called *attractive*.

Exercise 9.1 *Give the proof of the above statement with the help of maximal coupling.*

We next give three examples of systems satisfying properties (9.2) and (9.3).

1. *(ferromagnetic) Stochastic Ising model* (SIM):
   This model is defined on $\Omega = \{-1,1\}^{\mathbb{Z}^d}$ with rates

   $$c(x,\eta) = \exp[-\beta\eta(x)\sum_{y\sim x}\eta(y)], \qquad \beta \ge 0,$$

   which means that spins prefer to align with the majority of the neighboring spins.

2. *Contact process* (CP):
   This model is defined on $\Omega = \{0,1\}^{\mathbb{Z}^d}$ with rates

   $$c(x,\eta) = \begin{cases} \lambda \sum\limits_{y\sim x}\eta(y), & \text{if } \eta(x) = 0, \\ 1, & \text{if } \eta(x) = 1, \end{cases} \qquad \lambda > 0,$$

   which means that infected sites become healthy at rate 1 and healthy sites become infected at rate λ times the number of infected neighbors.

3. *Voter model* (VM):
   This model is defined on $\Omega = \{0,1\}^{\mathbb{Z}^d}$ with rates

   $$c(x,\eta) = \frac{1}{2d}\sum_{y\sim x} 1_{\{\eta(y)\neq\eta(x)\}},$$

   which means that sites choose a random neighbor at rate 1 and adopt the opinion of that neighbor.

Exercise 9.2 *Check that these three examples indeed satisfy properties* (9.2) *and* (9.3).

In the sequel we will discuss each model in some detail, with *coupling techniques* playing a central role. We will see that properties (9.2) and (9.3) allow for a number of interesting conclusions about the *equilibrium* behavior of these systems, as well as the *convergence* to equilibrium.

9.3 Convergence to equilibrium

Write [0] and [1] to denote the configurations $\eta \equiv 0$ and $\eta \equiv 1$, respectively. These are the *smallest*, respectively, the *largest* configurations in the partial order, and hence

$$[0] \preceq \eta \preceq [1], \qquad \forall\, \eta \in \Omega.$$

Since the dynamics preserves the partial order, we can obtain information about what happens when the system starts from any $\eta \in \Omega$ by comparing with what happens when it starts from [0] or [1].

Interacting particles can be described by *semigroups of transition kernels* $P = (P_t)_{t\geq 0}$. Formally, P_t is an operator acting on $C_b(\Omega)$, the space of bounded continuous functions on Ω, as

$$(P_t f)(\eta) = \mathbb{E}_\eta[f(\xi_t)], \qquad \eta \in \Omega,\ f \in C_b(\Omega).$$

If this definition holds on a dense subset of $C_b(\Omega)$, then it uniquely determines P_t.

Exercise 9.3 *Check that P_0 is the identity and that $P_{s+t} = P_t \circ P_s$ for all $s, t \geq 0$ (where $\circ$ denotes composition). For the latter, use the Markov property of ξ at time s.*

Alternatively, the semigroup can be viewed as acting on the space of probability measures μ on Ω via the duality relation

$$\int_\Omega f\, d(\mu P_t) = \int_\Omega (P_t f)\, d\mu, \qquad f \in C_b(\Omega).$$

See Liggett [9] for more details.

Lemma 9.4 *Let $P = (P_t)_{t\geq 0}$ denote the semigroup of transition kernels associated with ξ. Write $\delta_\eta P_t$ to denote the law of ξ_t conditional on $\xi_0 = \eta$ (which is a probability distribution on Ω). Then*

$$\begin{aligned} t &\mapsto \delta_{[0]} P_t \quad \textit{is stochastically increasing,} \\ t &\mapsto \delta_{[1]} P_t \quad \textit{is stochastically decreasing.} \end{aligned}$$

Proof. For $t, h \geq 0$,

$$\begin{aligned} \delta_{[0]} P_{t+h} &= (\delta_{[0]} P_h) P_t \succeq \delta_{[0]} P_t, \\ \delta_{[1]} P_{t+h} &= (\delta_{[1]} P_h) P_t \preceq \delta_{[1]} P_t, \end{aligned}$$

where we use that $\delta_{[0]} P_h \succeq \delta_{[0]}$ and $\delta_{[1]} P_h \preceq \delta_{[1]}$ for any $h \geq 0$, and also use Strassen's theorem (Theorem 7.8) to take advantage of the coupling representation that goes with the partial order. ∎

Corollary 9.5 *Both*

$$\begin{aligned} \underline{\nu} &= \lim_{t\to\infty} \delta_{[0]} P_t \ \textit{(“lower stationary law”),} \\ \overline{\nu} &= \lim_{t\to\infty} \delta_{[1]} P_t \ \textit{(“upper stationary law”),} \end{aligned}$$

exist as probability distributions on Ω and are equilibria for the dynamics. Any other equilibrium π satisfies $\underline{\nu} \preceq \pi \preceq \overline{\nu}$.

Proof. This is an immediate consequence of Lemma 9.4 and the sandwich $\delta_{[0]}P_t \preceq \delta_\eta P_t \preceq \delta_{[1]}P_t$ for $\eta \in \Omega$ and $t \geq 0$. ∎

The class of *all* equilibria for the dynamics is a *convex* set in the space of signed bounded measures on Ω. An element of this set is called *extremal* if it is not a proper linear combination of any two distinct elements in the set, i.e., not of the form $p\nu_1 + (1-p)\nu_2$ for some $p \in (0,1)$ and $\nu_1 \neq \nu_2$.

Lemma 9.6 *Both $\underline{\nu}$ and $\overline{\nu}$ are extremal.*

Proof. We give the proof for $\overline{\nu}$, the proof for $\underline{\nu}$ being analogous. Suppose that $\overline{\nu} = p\nu_1 + (1-p)\nu_2$. Since ν_1 and ν_2 are equilibria, we have by Corollary 9.5 that

$$\int_\Omega f d\nu_1 \leq \int_\Omega f d\overline{\nu}, \qquad \int_\Omega f d\nu_2 \leq \int_\Omega f d\overline{\nu},$$

for any f increasing. But since

$$\int_\Omega f d\overline{\nu} = p\int_\Omega f d\nu_1 + (1-p)\int_\Omega f d\nu_2$$

and $p \in (0,1)$, it follows that both inequalities must be equalities. Since the integrals of increasing functions determine the measure w.r.t. which is integrated, it follows that $\nu_1 = \overline{\nu} = \nu_2$. ∎

Exercise 9.7 *Prove that integrals of increasing functions determine the measure.*

Corollary 9.8 *The following three properties are equivalent (for shift-invariant spin-flip systems):*

1. *ξ is ergodic (i.e., $\delta_\eta P_t$ converges to the same limit distribution as $t \to \infty$ for all η),*
2. *there is a unique stationary distribution,*
3. *$\underline{\nu} = \overline{\nu}$.*

Proof. Obvious because of the sandwiching of all the configurations between [0] and [1]. ∎

9.4 Three examples

9.4.1 Example 1: Stochastic Ising Model

For $\beta = 0$, $c(x,\eta) = 1$ for all x and η, in which case the dynamics consists of independent spin-flips, up and down at rate 1. In that case $\overline{\nu} = \underline{\nu} = (\frac{1}{2}\delta_{-1} + \frac{1}{2}\delta_{+1})^{\otimes \mathbb{Z}^d}$.

For $\beta > 0$ the dynamics has a tendency to *align* spins. For small β this tendency is weak, for large β it is strong. It turns out that in $d \geq 2$ there is a *critical* value $\beta_d \in (0,\infty)$ such that

$$\begin{aligned} \beta \leq \beta_d\colon &\quad \underline{\nu} = \overline{\nu}, \\ \beta > \beta_d\colon &\quad \underline{\nu} \neq \overline{\nu}. \end{aligned}$$

The proof uses the so-called "Peierls argument", which we will encounter in Section 9.5. In the first case ("high temperature"), there is a unique ergodic equilibrium, which depends on

β and is denoted by ν_β. In the second case ("low temperature"), there are two extremal equilibria, both of which depend on β and are denoted by

$$\begin{array}{rcl} \nu_\beta^+ & = & \text{"plus state" with } \int_\Omega \eta(0)\nu_\beta^+(d\eta) > 0, \\ \nu_\beta^- & = & \text{"minus-state" with } \int_\Omega \eta(0)\nu_\beta^+(d\eta) < 0, \end{array}$$

which are called the magnetized states. Note that ν_β^+ and ν_β^- are images of each other under the swapping of $+1$'s and -1's. It can be shown that in $d = 2$ *all* equilibria are a convex combination of ν_β^+ and ν_β^-, while in $d \geq 3$ *also other* equilibria are possible (e.g. not shift-invariant) when β is large enough. It turns out that $\beta_1 = 0$, i.e., in $d = 1$ the SIM is ergodic for all $\beta > 0$.

9.4.2 Example 2: Contact Process

Note that [0] is a trap for the dynamics (if all sites are healthy, then no infection will ever occur), and so

$$\underline{\nu} = \delta_{[0]}.$$

For small λ infection is transmitted slowly, for large λ rapidly. It turns out that in $d \geq 1$ there is a *critical* value $\lambda_d \in (0, \infty)$ such that

$$\begin{array}{lll} \lambda \leq \lambda_d\colon & \overline{\nu} = \delta_{[0]} & (\text{"extinction", no epidemic}), \\ \lambda > \lambda_d\colon & \overline{\nu} \neq \delta_{[0]} & (\text{"survival", epidemic}). \end{array}$$

Lemma 9.9 *(i)* $d\lambda_d \leq \lambda_1$,
(ii) $2d\lambda_d \geq 1$,
(iii) $\lambda_1 < \infty$.

The proof of this lemma will be given in Section 9.5. It uses a number of comparison arguments based on coupling. Note that (i–iii) combine to yield that $0 < \lambda_d < \infty$ for all $d \geq 1$, so that the phase transition occurs at a non-trivial value of the infection rate parameter.

Remark: Sharp estimates are available for λ_1, but these require heavy machinery. Also, a series expansion of λ_d in powers of $1/2d$ is known up to several orders, but again the proof is very technical.

9.4.3 Example 3: Voter Model

Note that [0] and [1] are both traps for the dynamics (if all sites have the same opinion, then no change of opinion occurs), and so

$$\underline{\nu} = \delta_{[0]}, \qquad \overline{\nu} = \delta_{[1]}.$$

It turns out that in $d = 1, 2$ these are the only extremal equilibria, while in $d \geq 3$ there is a *1-parameter family of extremal equilibria*

$$(\nu_\rho)_{\rho \in [0,1]}$$

with ρ the density of 1's, i.e., $\nu_\rho(\eta(0) = 1) = \rho$. This is remarkable because the VM has no parameter to play with. For $\rho = 0$ and $\rho = 1$ these equilibria coincide with $\delta_{[0]}$ and $\delta_{[1]}$, respectively.

Remark: The dichotomy $d = 1, 2$ versus $d \geq 3$ is directly related to simple random walk being recurrent in $d = 1, 2$ and transient in $d \geq 3$.

9.5 A closer look at the Contact Process

We will next prove (i-iii) in Lemma 9.9. This will take up some space, organized into Sections 9.5.1–9.5.4. In the proof we need a property of the CP called *self-duality*. We will not explain in detail what this is, but only say that it means the following:

> CP *locally* dies out (in the sense of weak convergence) starting from $\delta_{[1]}$ if and only if CP *fully* dies out when starting from a configuration with finitely many infections, e.g., $\{0\}$.

For details we refer to Liggett [9].

9.5.1 Uniqueness of the critical value

Pick $\lambda_1 < \lambda_2$. Let $c_{\lambda_1}(x,\eta)$ and $c_{\lambda_2}(x,\eta)$ denote the local transition rates of the CP with parameters λ_1 and λ_2, respectively. Then it is easily checked that

$$\eta \preceq \eta' \to \begin{cases} c_{\lambda_1}(x,\eta) \le c_{\lambda_2}(x,\eta') & \text{if } \eta(x) = \eta'(x) = 0, \\ c_{\lambda_1}(x,\eta) \ge c_{\lambda_2}(x,\eta') & \text{if } \eta(x) = \eta'(x) = 1, \end{cases} \qquad \forall\, x \in \mathbb{Z}^d,\ \eta \in \Omega.$$

(For the CP the last inequality is in fact an equality.) Consequently,

$$\delta_{[1]} P_t^{\lambda_1} \preceq \delta_{[1]} P_t^{\lambda_2} \qquad \forall t \ge 0,$$

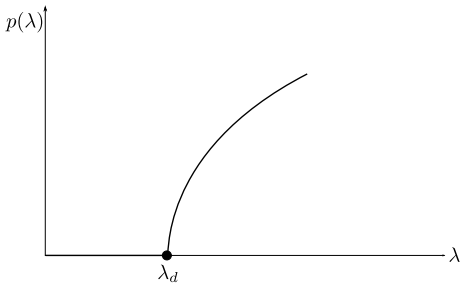
by the maximal coupling, with $P^\lambda = (P_t^\lambda)_{t\ge 0}$ denoting the semigroup of the CP with parameter λ. Letting $t \to \infty$, we get

$$\nu_{\lambda_1} \preceq \nu_{\lambda_2}$$

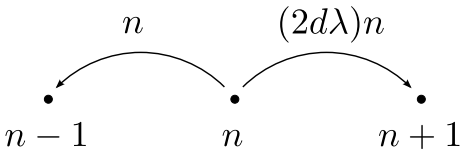
with ν_λ the upper invariant measure of the CP with parameter λ. With $\rho(\lambda) = \nu_\lambda(\eta(0) = 1)$ denoting the density of 1's in equilibrium, it follows that $\rho(\lambda_1) \le \rho(\lambda_2)$. Hence

$$\lambda_d = \inf\{\lambda \ge 0\colon\ \rho(\lambda) > 0\} = \sup\{\lambda \ge 0\colon\ \rho(\lambda) = 0\}$$

defines a unique critical value, separating the phase of (local) extinction of the infection from the phase of (local) survival of the infection. The curve $\lambda \mapsto \rho(\lambda)$ is continuous on $\mathbb{R}$. The continuity at $\lambda = \lambda_d$ is hard to prove.



9.5.2 Lower bound on the critical value



Pick A_0 finite and consider the CP in dimension d with parameter λ starting from the set A_0 as the set of infected sites. Let $A = (A_t)_{t\geq 0}$ with A_t the set of infected sites at time t. Then

$$\begin{aligned} &|A_t| \text{ decreases by 1 at rate } |A_t|, \\ &|A_t| \text{ increases by 1 at rate } \leq 2d\lambda|A_t|, \end{aligned}$$

where the latter holds because each site in A_t has at most $2d$ non-infected neighbors. Now consider the two random process $X = (X_t)_{t\geq 0}$ with $X_t = |A_t|$ and $Y = (Y_t)_{t\geq 0}$ given by the birth-death process on $\mathbb{N}_0$ that moves at rate n from n to $n-1$ (death) and at rate $(2d\lambda)n$ from n to $n+1$ (birth), both starting from $n_0 = |A_0|$. Then X and Y can be *coupled* such that

$$\hat{\mathbb{P}}(X_t \leq Y_t \ \forall\, t \geq 0) = 1,$$

where $\hat{\mathbb{P}}$ denotes the coupling measure. Note that $n = 0$ is a trap for both X and Y. If $2d\lambda < 1$, then this trap is hit with probability 1 by Y, i.e., $\lim_{t\to\infty} Y_t = 0$ a.s., and hence also by X, i.e., $\lim_{t\to\infty} X_t = 0$ a.s. Therefore $\nu_\lambda = \delta_{[0]}$ when $2d\lambda < 1$. Consequently, $2d\lambda_d \geq 1$. ∎

9.5.3 Upper bound on the critical value

The idea is to *couple* two CP's that live in dimensions 1 and d. Again, let $A = (A_t)_{t\geq 0}$ with A_t the set of infected sites at time t of the CP in dimension d with parameter λ, this time starting from $A_0 = \{0\}$. Let $B = (B_t)_{t\geq 0}$ be the same as A, but for the CP in dimension 1 with parameter λd, starting from $B_0 = \{0\}$.

Define the projection $\pi_d\colon\ \mathbb{Z}^d \to \mathbb{Z}$ as

$$\pi_d(x_1, \ldots, x_d) = x_1 + \cdots + x_d.$$

We will construct a coupling $\hat{\mathbb{P}}$ of A and B such that

$$\hat{\mathbb{P}}(B_t \subseteq \pi_d(A_t) \ \forall\, t \geq 0) = 1.$$

From this we get

$$\mathbb{P}\big(A_t \neq \emptyset \mid A_0 = \{0\}\big) = \mathbb{P}\big(\pi_d(A_t) \neq \emptyset \mid A_0 = \{0\}\big) \geq \mathbb{P}\big(B_t \neq \emptyset \mid B_0 = \{0\}\big),$$

which implies that if A dies out then also B dies out. In other words, if $\lambda \leq \lambda_d$, then $\lambda d \leq \lambda_1$, which implies that $d\lambda_d \leq \lambda_1$ as claimed.

The construction of the coupling is as follows. Fix $t \geq 0$. Suppose that $A_t = A$ and $B_t = B$ with $B \subset \pi_d(A)$. For each $y \in B$ there is *at least one* $x \in A$ with $y = \pi_d(x)$. Pick *one* such x for every y (e.g. choose the closest up or the closest down). Now couple:

- If x becomes healthy, then y becomes healthy too.
- If x infects any of the d sites $x - e_i$ with $i = 1, \ldots, d$, then y infects $y - 1$.
- If x infects any of the d sites $x + e_i$ with $i = 1, \ldots, d$, then y infects $y + 1$.
- Anything that occurs at other x''s such that $\pi_d(x') = y$, has no effect on y.

(This is a mapping that defines how B_t evolves given how A_t evolves.)

Exercise 9.10 *Check that this coupling has the right marginals and preserves the inclusion $B_t \subseteq \pi_d(A_t)$.*

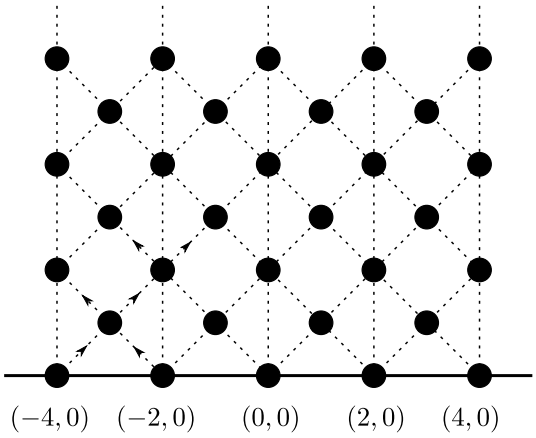
Since $A_0 = B_0 = \{0\}$ and $\{0\} \subset \pi_d(\{0\})$, the proof is complete.

9.5.4 Finite critical value in dimension 1

The proof proceeds via comparison with *directed site percolation* on $\mathbb{Z}^2$. We first make a digression into this part of percolation theory.

Each site is *open* with probability p and *closed* with probability $1-p$, independently of all other sites, with $p \in [0,1]$. The associated probability law on configuration space is denoted by $\mathbb{P}_p$. We say that y is connected to x, written as $x \rightsquigarrow y$, if there is a path from x to y such that

1. all sites in the path are open (including x and y);
2. the path traverses bonds in the *upward* direction.



Let $\mathbb{H} = \{x = (x^1, x^2) \in \mathbb{Z}^2 \colon\ x^2 \geq 0\}$. The random set

$$C_0 = \{x \in \mathbb{H} \colon\ 0 \rightsquigarrow x\}$$

is called the *cluster of the origin* ($C_0 = \emptyset$ if 0 is closed). The percolation function is

$$\theta(p) = \mathbb{P}_p(|C_0| = \infty),$$

and the *critical percolation threshold* is

$$p_c = \inf\{p \in [0,1] \colon\ \theta(p) > 0\} = \sup\{p \in [0,1] \colon\ \theta(p) = 0\}.$$

The uniqueness of p_c follows from the *monotonicity* of $p \mapsto \theta(p)$ proved in Section 8.1.

Lemma 9.11 $p_c \leq \frac{80}{81}$.

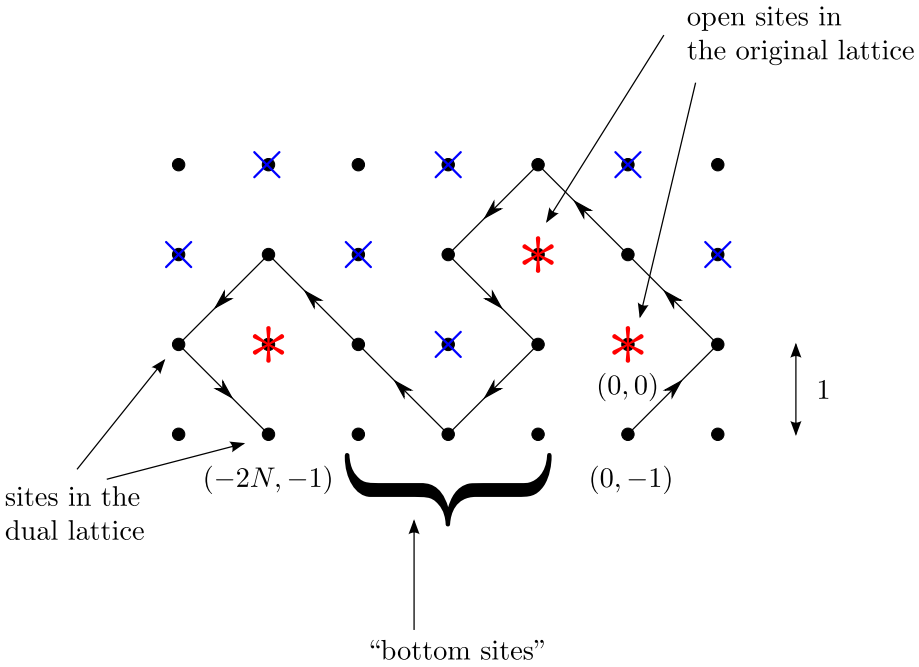
Proof. Pick $N \in \mathbb{N}$ large. Let

$$\begin{aligned} C_N &= \cup_{i=0}^{N} \{x \in \mathbb{H} \colon\ (-2i, 0) \rightsquigarrow x\} \\ &= \text{all sites connected to the lower left boundary of } \mathbb{H} \\ &\quad \text{(including the origin).} \end{aligned}$$

We want to lay a *contour* around C_N. To do so, we consider the oriented lattice that is obtained by shifting all sites and bonds downward by 1. We call this the *dual lattice*, because

the two lattices together make up $\mathbb{Z}^2$ (with upward orientation). Now define

Γ_N = the exterior boundary of the set of all faces in the dual lattice containing a site of C_N or one of the boundary sites$(-2i+1,-1)$, with $i = 1, \ldots, N$.



Think of Γ_N as a path from $(0,-1)$ to $(-2N,-1)$ in the dual lattice, enclosing C_N and being allowed to cross bonds in *both* directions. We call Γ_N the contour of C_N (this contour may be infinite). We need the following observations:

(i) There are at most $4\,3^{n-2}$ contours of length n.
(ii) Any contour of length n has at least $n/4$ closed sites adjacent to it on the outside.

Exercise 9.12 *Prove the observations above.*

We can now complete the proof as follows. Since the shortest possible contour has length $2N$, it follows from (i) and (ii) that

$$\mathbb{P}_p(|C_N| < \infty) = \mathbb{P}_p(|\Gamma_N| < \infty) \leq \sum_{n=2N}^{\infty} 4\,3^{n-2}(1-p)^{n/4}.$$

If $p > 80/81$, then $3(1-p)^{1/4} < 1$ and the sum is < 1 for N sufficiently large, i.e., $\mathbb{P}_p(|C_N| = \infty) > 0$ for $N \geq N_0(p)$. Using the translation invariance, we have

$$\mathbb{P}_p(|C_N| = \infty) \leq (N+1)\mathbb{P}_p(|C_0| = \infty).$$

Hence, if $p > 80/81$, then $\mathbb{P}_p(|C_0| = \infty) > 0$, which implies that $p_c \leq 80/81$. ∎

The contour argument above is referred to as a *Peierls argument*. A similar argument works for many other models as well (such as SIM).

Lemma 9.11 is the key to proving that $\lambda_1 < \infty$, as we next show. The proof uses a *coupling argument* showing that the one-dimensional CP observed at times $0, \delta, 2\delta, \ldots$ with $\delta = \frac{1}{\lambda}\log(\lambda+1)$ *dominates* oriented percolation with $p = p(\lambda)$ given by

$$p(\lambda) = \left(\frac{\lambda}{\lambda+1}\right)^2 \left(\frac{1}{\lambda+1}\right)^{\frac{2}{\lambda}}.$$

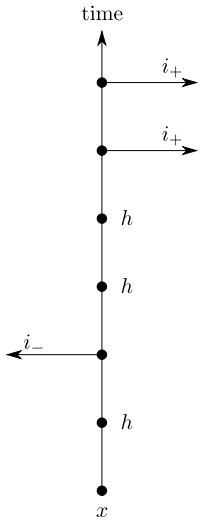
Since $\lim_{\lambda\to\infty} p(\lambda) = 1$ and $p_c \leq \frac{80}{81} < 1$, the infection (locally) survives for λ large enough.

Lemma 9.13 *The one-dimensional CP survives if*

$$p(\lambda) > \frac{80}{81}.$$

Proof. Again consider the half-lattice $\mathbb{H}$ that was used for directed percolation. Pick $\delta > 0$ and shrink the vertical direction by a factor δ. Add dotted vertical lines that represent the time axes associated with the sites of $\mathbb{Z}$. In this graph we are going to construct CP and orientated percolation *together*. This construction comes in three steps.

Step 1: With each time axis we associate three Poisson point processes:



1. Points labeled h (= healthy) at rate 1.
2. Points with right arrows labeled i_+ (= right infection) at rate λ.
3. Points with left arrows labeled i_- (= left infection) at rate λ.

All Poisson point processes at all time axes are independent. Given their realization, we define

$$\begin{aligned} A_t \;=\; & \text{the set of } x \in \mathbb{Z} \text{ such that } (x,t) \text{ can be reached from } (0,0) \\ & \text{by a path that only goes upwards along stretches } \textit{without} \\ & h\text{'s and sidewards along arrows } \textit{with } i_+ \text{ or } i_-. \end{aligned}$$

Exercise 9.14 *Show that $A = (A_t)_{t\geq 0}$ is the CP with parameter λ starting from $A_0 = \{0\}$.*

Step 2: We say that site $(x, n\delta)$ is *open* if
(i) between time $(n-1)\delta$ and $(n+1)\delta$ there is *no* h;

(ii) between time $n\delta$ and $(n+1)\delta$ there are *both* an i_+ and an i_-.
Define

$$B_{n\delta} = \text{the set of } x \in \mathbb{Z} \text{ such that } 0 \rightsquigarrow (x, n\delta).$$

Exercise 9.15 *Show that* $B_{n\delta} = \{x \in \mathbb{Z}\colon\ (x, n\delta) \in C_0\}$, *where* C_0 *is the cluster at the origin in orientated percolation with* $p = e^{-2\delta}(1 - e^{-\delta\lambda})^2$.

Step 3: The key part of the coupling is

$$A_{n\delta} \supset B_{n\delta} \qquad \forall\, n \in \mathbb{N}_0.$$

Exercise 9.16 *Prove this inclusion.*

Let $\mathbb{P}$ denote the joint law of the three Poisson point processes at all the sites. By combining Steps 1–3 and noting that

$$\begin{aligned} \mathbb{P}\big(A_{n\delta} \neq \emptyset\ \forall\, n \in \mathbb{N}_0 \mid A_0 = \{0\}\big) &\geq \mathbb{P}\big(B_{n\delta} \neq \emptyset\ \forall\, n \in \mathbb{N}_0 \mid B_0 = \{0\}\big) \\ &= \mathbb{P}_p(|C_0| = \infty), \end{aligned}$$

we obtain, with the help of Fact 9.11, that the one-dimensional CP with parameter λ survives if

$$\sup_{\delta > 0} e^{-2\delta}(1 - e^{-\delta\lambda})^2 > \frac{80}{81},$$

where in the left-hand side we *optimize* over δ, which is allowed because the previous estimates hold for all $\delta > 0$. The supremum is attained at

$$\delta = \frac{1}{\lambda}\log(\lambda + 1),$$

which yields the claim in Fact 9.13. ∎

Since $\lim_{\lambda\to\infty} p(\lambda) = 1$, it follows from Lemma 9.13 that $\lambda_1 < \infty$.

Remark: The bound in Lemma 9.13 yields $\lambda_1 \leq 1318$. This is a large number because the estimates that were made are crude. The true value is $\lambda_1 \approx 1.6494$, based on simulations and approximation techniques.

10 Diffusions

In Section 10.1 we couple diffusions in dimension 1, in Section 10.2 diffusions in dimension d.

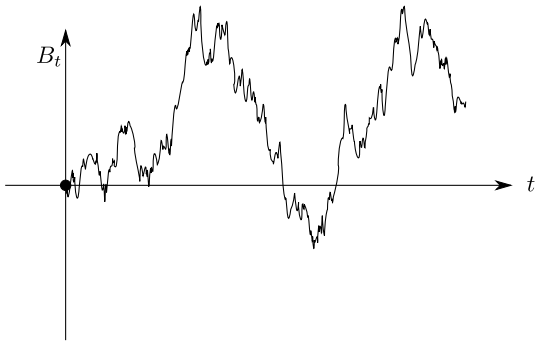
10.1 Diffusions in dimension 1

10.1.1 General properties

Let $S = (S_n)_{n\in\mathbb{N}_0}$ be *simple random walk* on $\mathbb{Z}$, i.e., $S_0 = 0$ and $S_n = X_1 + \ldots + X_n$, $n \in \mathbb{N}$, with $X = (X_n)_{n\in\mathbb{N}}$ i.i.d. with $\mathbb{P}(X_1 = -1) = \mathbb{P}(X_1 = 1) = \frac{1}{2}$. The limit of S under *diffusive scaling* is a *Brownian motion*:

$$\left(\frac{1}{\sqrt{n}} S_{\lceil nt\rceil}\right) \quad \stackrel{n\to\infty}{\Longrightarrow} \quad (B_t)_{t\geq 0}$$

with $\lceil\cdot\rceil$ the upper integer part. Here, $\Longrightarrow$ denotes convergence in *path space* endowed with a metric that is "a kind of flexible supremum norm", called the *Skorohod norm*.



Brownian motion $B = (B_t)_{t\geq 0}$ is a Markov process taking walues in $\mathbb{R}$ and having *continuous paths*. The law of B is called the *Wiener measure*, a probability measure on the set of continuous paths such that increments over *disjoint* time intervals are *independent* and *normally* distributed. To define B properly requires a formal construction that is part of *stochastic analysis*, a subarea of probability theory that uses functional analytic machinery to study continuous-time random processes taking values in $\mathbb{R}$. B is an example of a diffusion.

Definition 10.1 *A diffusion $X = (X_t)_{t\geq 0}$ is a Markov process on $\mathbb{R}$ with continuous paths having the strong Markov property.*

We write $\mathbb{P}_x$ to denote the law of X given $X_0 = x \in \mathbb{R}$. The sample space Ω is the space of continuous functions with values in $\mathbb{R}$, written $C_{\mathbb{R}}[0,\infty)$, endowed with the Borel σ-algebra $\mathcal{C}_{\mathbb{R}}[0,\infty)$ of subsets of $C_{\mathbb{R}}[0,\infty)$ with the *Skorohod topology*.

Remark: The time interval need not be $[0,\infty)$. It can also be $(-\infty,\infty)$, $[0,1]$, etc., depending on what X describes. It is also possible that X takes values in $\mathbb{R}^d$, $d \geq 1$, etc.

An example of a diffusion is X solving the *stochastic differential equation*

$$dX_t = b(X_t)\,dt + \sigma(X_t)\,dB_t, \tag{10.1}$$

where $b(X_t)$ denotes the local drift function and $\sigma(X_t)$ the dispersion function. The integral form of (10.1) reads

$$X_t = X_0 + \int_0^t b(X_s)\,ds + \int_o^t \sigma(X_s)\,dB_s,$$

where the last integral is a so-called "Itô-integral". Equation (10.1) is short-hand for the statement:

> The increments of X over the infinitesimal time interval $[t, t+dt)$ is a sum of two parts, $b(X_t)dt$ and $\sigma(X_t)dB_t$, with dB_t the increment of B over the same time interval.

Again, a formal definition of (10.1) requires functional analytic machinery. The functions $b\colon \mathbb{R} \to \mathbb{R}$ and $\sigma\colon \mathbb{R} \to \mathbb{R}$ need to satisfy *mild regularity properties*, e.g. locally Lipschitz continuous and modest growth at infinity. The solution of (10.1) is called an Itô-diffusion. The special case with $b \equiv 0$, $\sigma \equiv 1$ is Brownian motion itself. The interpretation of X is:

> X is a Brownian motion whose increments are blown up by a factor $\sigma(\cdot)$ and shifted by a factor $b(\cdot)$, both of which depend on the value of the process itself.

Definition 10.2 *A diffusion is called regular if*

$$\mathbb{P}_x(\tau_y < \infty) > 0 \qquad \forall\, x, y \in \mathbb{R}$$

with $\tau_y = \inf\{t \in [0,\infty)\colon\ X_t = y\}$ *the hitting time of* y.

Regularity is analogous to irreducibility for Markov processes taking values in countable state spaces. Every regular diffusion has the property

$$\mathbb{P}_x(\tau_b < \tau_a) = \frac{s(x) - s(a)}{s(b) - s(a)} \qquad \forall\, a, b \in \mathbb{R},\ a < x < b,$$

for some $s\colon \mathbb{R} \to \mathbb{R}$ continuous and strictly increasing. This s is called the *scale function* for X. A diffusion is "in natural scale" when s is the identity. An example of such a diffusion is Brownian motion B. More generally, $Y = (Y_t)_{t\geq 0}$ with $Y_t = s(X_t)$ is in natural scale, and is an Itô-diffusion with $b \equiv 0$.

Exercise 10.3 *Check the last claim.*

Definition 10.4 *A diffusion is called recurrent if*

$$\mathbb{P}_x(\tau_y < \infty) = 1 \qquad \forall\, x, y \in \mathbb{R}.$$

10.1.2 Coupling on the half-line

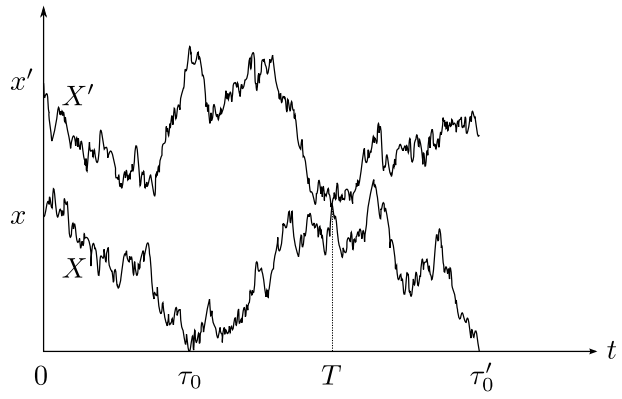
For recurrent diffusions on the *half-line* we have a successful coupling starting from any two starting points. Indeed, let

$$T = \inf\{t \in [0,\infty)\colon\ X_t = X'_t\}$$

be the coupling time of $X = (X_t)_{t\geq 0}$ and $X' = (X'_t)_{t\geq 0}$. Because X and X' are continuous ("skip-free"), we have

$$T \leq \tau_0 \vee \tau'_0,$$

and so recurrence implies that $\hat{\mathbb{P}}_{xx'}(T < \infty) = 1$ for all $x, x' \in \mathbb{R}$, with $\hat{\mathbb{P}}_{xx'} = \mathbb{P}_x \otimes \mathbb{P}_{x'}$ the independent coupling.



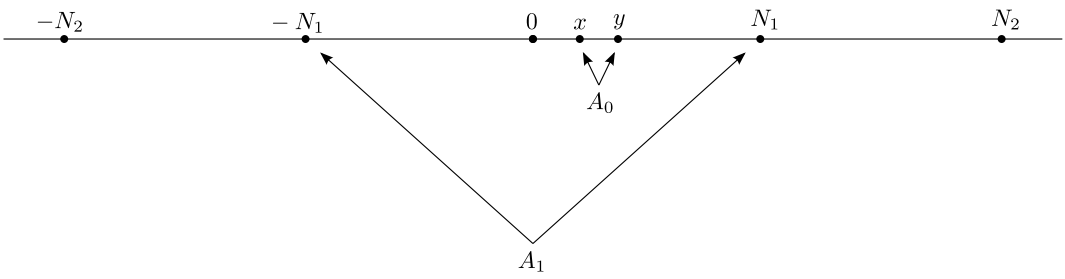
Consequently, the coupling inequality

$$\|\mathbb{P}_x(X_t \in \cdot) - \mathbb{P}_y(X_t \in \cdot)\|_{tv} \leq 2\hat{\mathbb{P}}_{xy}(T > t)$$

gives

$$\lim_{t\to\infty} \|\mathbb{P}_x(X_t \in \cdot) - \mathbb{P}_y(X_t \in \cdot)\|_{tv} = 0 \qquad \forall\, x, y \in \mathbb{R}.$$

10.1.3 Coupling on the full-line



For recurrent diffusions on the *full-line* a similar result holds. The existence of a successful coupling is proved as follows. Without loss of generality we assume that X is in natural scale. Fix $x < y$ and pick $0 < N_1 < N_2 < \cdots$ such that

$$|\mathbb{P}_z(\tau_{A_k} = N_k) - \tfrac{1}{2}| \leq \tfrac{1}{4} \qquad z \in A_{k-1},\, k \in \mathbb{N},$$

with $A_k = \{-N_k, N_k\}$ and $A_0 = \{x, y\}$. Then, by the skip-freeness, we have

$$\hat{\mathbb{P}}_{xy}\left(X_{\tau_{A_k}} \leq X'_{\tau'_{A_k}} \text{ for } 1 \leq k \leq l\right) \leq \left[1 - \left(\tfrac{1}{4}\right)^2\right]^l, \quad l \in \mathbb{N},$$

which in turn implies that

$$\hat{\mathbb{P}}_{xy}(T < \infty) = \hat{\mathbb{P}}_{xy}(X_s = X'_s \text{ for some } s \geq 0) = 1.$$

Indeed, if $X_{\tau_{A_k}} > X'_{\tau'_{A_k}}$ for some $k \in \mathbb{N}$, then there exists an $s \leq \tau_{A_k} \wedge \tau'_{A_k}$ such that $X_s = X'_s$. Via the coupling inequality, we again get loss of memory.

Theorem 10.5 *Regular diffusions X have the strong Feller property, i.e., for any bounded $f\colon \mathbb{R} \to \mathbb{R}$ and any $t > 0$, the function $P_t f$ defined by*

$$(P_t f)(x) = \mathbb{E}_x[f(X_t)], \qquad x \in \mathbb{R},$$

is continuous.

Proof. Fix $t > 0$. Let X and X' be independent copies of the diffusion starting from x and x', respectively. Then

$$\begin{aligned} \big|(P_t f)(x) - (P_t f)(x')\big| &= \Big|\hat{\mathbb{E}}_{xx'}[f(X_t)] - \hat{\mathbb{E}}_{xx'}[f(X'_t)\Big| \\ &\leq 2\hat{\mathbb{P}}_{xx'}(T > t)\, \|f\|_\infty. \end{aligned}$$

The claim follows from the fact that

$$\lim_{x' \to x} \hat{\mathbb{P}}_{xx'}(T > t) = 0 \qquad \forall\, t > 0,$$

which is intuitively obvious. ∎

Exercise 10.6 *Prove the latter statement by using an argument of the type given for the successful coupling on the full-line, but now with shrinking rather than growing intervals.*

The Feller property is important because it says that the space of bounded continuous functions is *preserved* by the semigroup $P = (P_t)_{t \geq 0}$. Since this set is dense in the space of continuous functions, the Feller property allows us to *control* very large sets of functionals of diffusions.

Theorem 10.7 *Let $P = (P_t)_{t \geq 0}$ be the semigroup of a regular diffusion. Then*

$$\lambda \leq \mu \quad \Longrightarrow \quad \lambda P_t \leq \mu P_t \quad \forall\, t \geq 0.$$

Proof. This is immediate from the skip-freeness, by which $\lambda \leq \mu$ allows $X_0 \leq X'_0$, and hence $X_t \leq X'_t$ for all $t \geq 0$, when X_0, X'_0 start from λ, μ. ∎

10.2 Diffusions in dimension d

Let $S = (S_n)_{n \in \mathbb{N}}$ be simple random walk on $\mathbb{Z}^d$, $d \geq 1$: $S_0 = 0$, $S_n = X_1 + \cdots + X_n$, $n \in \mathbb{N}$, with $X = (X_n)_{n \in \mathbb{N}}$ i.i.d. with $\mathbb{P}(X_1 = -e_i) = \mathbb{P}(X_1 = e_i) = \frac{1}{2d}$, $i = 1 \ldots, d$, where $e_1, \ldots, e_d$ are the unit vectors in $\mathbb{Z}^d$.

The limit of S under *diffusive scaling* is Brownian motion on $\mathbb{R}^d$:

$$\left(\frac{1}{\sqrt{n}} S_{\lceil nt \rceil}\right) \quad \stackrel{n \to \infty}{\Longrightarrow} \quad (B_t)_{t \geq 0},$$

where the right-hand side is Markov process with values in $\mathbb{R}^d$ and with continuous paths. In fact,

$$B_t = (B_t^1, \ldots, B_t^n)$$

such that the d components form *independent* Brownian motions on $\mathbb{R}$ (moving at $1/d$ times the rate of one-dimensional Brownian motion). The main definitions of what a diffusion is on $\mathbb{R}^d$ carry over from $d = 1$. Regularity becomes

$$\mathbb{P}_x(\tau_{B_\epsilon(y)} < \infty) > 0 \qquad \forall\, x, y \in \mathbb{R}^d,\, \epsilon > 0,$$

and recurrence becomes

$$\mathbb{P}_x(\tau_{B_\epsilon(y)} < \infty) = 1 \qquad \forall\, x, y \in \mathbb{R}^d,\ \epsilon > 0,$$

i.e., points are replaced by small balls around points in all statements about hitting times.

Itô-diffusions are defined by

$$dX_t = b(X_t)\, dt + \sigma(X_t)\, dB_t, \tag{10.2}$$

where $b\colon\ \mathbb{R}^d \to \mathbb{R}^d$ and $\sigma\colon\ \mathbb{R}^d\mathbb{R}^d \times \mathbb{R}^d$ are the *vector* local drift function and the *matrix* local dispersion function, both subject to regularity properties.

Diffusions in $\mathbb{R}^d$, $d \geq 2$, are *more difficult* to analyze than in $\mathbb{R}$. A lot is known for special classes of diffusions (e.g. with certain symmetry properties). Stochastic analysis has developed a vast arsenal of ideas, results and techniques. The stochastic differential equation in (10.2) is very important because it has a wide range of application, e.g. in transport, finance, filtering, coding, statistics, genetics, etc.

References

[1] O. Angel, J. Goodman, F. den Hollander and G. Slade, Invasion percolation on regular trees, Annals of Probability 36 (2008) 420–466.

[2] A.D. Barbour, L. Holst and S. Janson, *Poisson Approximation*, Oxford Studies in Probability 2, Clarendon Press, Oxford, 1992.

[3] P. Diaconis, The cutoff phenomenon in finite Markov chains, Proc. Natl. Acad. Sci. USA 93 (1996) 1659–1664.

[4] G.R. Grimmett, *Percolation*, Springer, Berlin, 1989.

[5] O. Häggström, *Finite Markov Chains and Algorithmic Applications*, London Mathematical Society Student Texts 52, Cambridge University Press, Cambridge, 2002.

[6] F. den Hollander and M.S. Keane, Inequalities of FKG type, Physica 138A (1986) 167–182.

[7] C. Kraaikamp, *Markov Chains: an introduction*, lecture notes TU Delft, 2010.

[8] D.A. Levin, Y. Peres and E.L. Wilmer, *Markov Chains and Mixing Times*, American Mathematical Society, Providence RI, 2009.

[9] T.M. Liggett, *Interacting Particle Systems*, Grundlehren der mathematische Wissenschaften 276, Springer, New York, 1985.

[10] T. Lindvall, W. Doeblin 1915–1940, Annals of Probability 19 (1991) 929–934.

[11] T. Lindvall, *Lectures on the Coupling Method*, John Wiley & Sons, New York, 1992. Reprint: Dover paperback edition, 2002.

[12] H. Nooitgedagt, Two convergence limits of Markov chains: Cut-off and Metastability, MSc thesis, Mathematical Institute, Leiden University, 31 August 2010.

[13] J.A. Rice, *Mathematical Statistics and Data Analysis* (3rd edition), Duxbury Advanced Series, Thomson Brooks/Cole, Belmont, California, 2007.

[14] F. Spitzer, *Principles of Random Walk*, Springer, New York, 1976.

[15] H. Thorisson, *Coupling, Stationarity and Regeneration*, Springer, New York, 2000.