`magic_msu-v08.tex`

Introduction to MPs & Poisson equations

A.Yu.Veretennikov

April 10, 2009

1. General and integration
2. Uniformly ergodic MC's; LLN and CLT.
3. Non-uniformly ergodic MC's. Lyapunov stability. Convergence to stationary distribution in total variation. Harris technique.
4. LLN and CLT, again.
5. Poisson equations. Invariant measures, rate of convergence in total variation. Localization of Dobrushin's conditions for local mixing.
6. Poisson equations with parameters.

1 Introduction: probability spaces and random variables

1. We have to define our objects with which we are going to work. For a probabilist, those objects are, first of all, probability space and random variable. Hence, in the first lecture we have to discuss these notions.

   Eventually, random variable will be a function of outcome, with some restrictions, the latter beaing practically always omitted in the undergraduate level. So we start with *outcomes*. Standard notation for a *probability space* is $(\Omega, \mathcal{F}, P)$. Here Ω is the *space of outcomes*, which formally may be any nonempty set, and whose elements are called *outcomes*. Often Euclidean spaces or their subsets may serve as spaces of outcomes, however, virtually Ω could be any set [1].

2. The next object in the triple $(\Omega, \mathcal{F}, P)$ is $\mathcal{F}$. This is understood as *sigma– algebra* (equivalently, *sigma–field*) of subsets of Ω. Can it consist of all subsets of Ω (in this case $\mathcal{F}$ is denoted usually by 2^{Ω})? Only if Ω itself is finite or countable. This relates to some subtle issues of measure theory and integration which we are not in the position to discuss here. However, even if Ω is finite, there are natural cases where it is unreasonable to include all subsets into $\mathcal{F}$; we will see this a bit later when we turn to conditional expectations.

   To be a sigma–algebra, the family $\mathcal{F}$ must satisfy certain requirements:

   (a) The set Ω and emptyset $\emptyset$ belong to $\mathcal{F}$.

   (b) If $A \in \mathcal{F}$, then also $\Omega \setminus A \in \mathcal{F}$.

   (c) If there is any finite or countable sequence $A_i \in \mathcal{F}$, then also $\bigcup\limits_i A_i \in \mathcal{F}$.

   Elements of $\mathcal{F}$ are usually called *events*. If instead of (c) we require the same property with only any finite number of events, the family is called *algebra*.

3. *Minimal sigma–algebra generated by some family $\mathcal{G} \subset 2^{\Omega}$*. For any such (nonempty) family we can define the minimal sigma–algebra which

[1] Nonempty is a must. We do not touch Set theory axioms here, that is, we admit that the notion of *set* does not require any discussion.

contains all sets from $\mathcal{G}$, this is denoted by $\mathcal{F}^{\mathcal{G}}$. Clearly, there is at least one sigma–algebra which contains $\mathcal{G}$, namely, 2^Ω. On the other hand, the object $\bigcap \mathcal{S}$ which is the intersection of all sigma–algebras containing $\mathcal{G}$, indeed, is a sigma–algebra (An Easy Exercise); clearly, it is minimal. Hence, the notion makes sense.

4. The third element in the triple is P, *probability measure.* This is a *function on elements from $\mathcal{F}$.* That is, we will be talking only of probability of events. To be a probability measure, the function P must also satisfy so,e requirements, usually called *Kolmogorov's axioms.*

   (a) The function P is non-negative, and $P(\Omega) = 1$.

   (b) Additivity: if the events A_i, $1 \le i \le N$, do not intersect (are pairwise exclusive), then $P(\bigcup\limits_i A_i) = \sum_i P(A_i)$.

   (c) Sigma–additivity: if there is a countable family of events A_i which are pairwise exclusive, then $P(\bigcup\limits_i A_i) = \sum_i P(A_i)$.

   Naturally, here (c) implies (b), however, we would like to keep the assumption (b), too.

5. *Completed sigma-algebra.* Once P is introduced, it is possible and in some occasions very useful to extend $\mathcal{F}$ in the following way: denote $\bar{\mathcal{F}} := \{A \subseteq \Omega : \exists B \in \mathcal{F},\ P(B\Delta A) = 0\}$, where $B\Delta A$ is a *symmetric difference* of B and A, that is, $(B \setminus A) \bigcup (A \setminus B)$. It is An Easy Exercise to show that thus defined $\bar{\mathcal{F}}$ is a sigma-algebra which has a name, *completed sigma–algebra* (with respect to the probability P). In the sequel, we always assume that our sigma-algebra $\mathcal{F}$ is completed (and will not use notation $\bar{\mathcal{F}}$ for that).

6. Example. *Discrete probability space.* Here $\Omega = \{\omega_1, \omega_2, \ldots\}$; there are given non-negative numbers p_i, $i \ge 1$, so that $\sup_i p_i = 1$. Now, let $\mathcal{F} = 2^\Omega$. Then, if we have any event A, its probability is defined as $P(A) = \sup_{i:\, \omega_i \in A} p_i$. It is An Easy Exercise, to check the axioms of probability for this case.

7. Example. *Continuous probability space*, with $\Omega = R^1$ or R^d, and a density, $p(\cdot)$. Here for any set A for which "the integral $\int_A p(x)\,dx \equiv \int 1(x \in A)\, p(x)\, dx$ makes sense", probability $P(A)$ is defined as the

latter integral. All *first course textbooks*, of course, use Riemann integration here; accordingly, set A is suitable for any density if it is *Jordan–measurable*, so $\mathcal{F}$ should consist of all Jordan–measurable sets on the line (we do not revise this notion here, however). Unfortunately, this is an algebra, but *not* a sigma-algebra. Hence, this way, – to use Riemann integration, – is potentially dangerous in probability, and may lead to contradictions and counterexamples, because Riemann integral has no sigma-additivity property. It works well just because most of densities and integrals which may arise admit Lebesgue integration, even though this side remains hidden. If the reader is not convinced, he may be offered a little exercise, namely, to construct a counterexample so that the axiom (c) above is violated.

Why do we care about the axiom (c)? Simply because if some event (set) is split into a countable number of mutually exclusive sub-events, the probability of the original event clearly should not depend on the order of those sub-events in the sum; moreover, this will also affect expectations (below).

One could question whether continuous case is necessary at *first course textbooks* at all, given that discrete case is really free from those obstacles. The answer is that as long as you do not deal with, say, Central Limit Theorem or some other limiting procedure where some density shows up, you can well stay with discrete case. However, if you wish to use CLT (χ^2, F, Student, etc.), apparently, it becomes necessary to deal with continuous densities, and you must take care of sigma-additivity.

8. *General probability space* may be neither discrete, nor continuous, but usually some mixture, and may include a singular non-discrete part of the probability measure, too. Singular with respect to what? e.g., with respect to the usual Lebesgue measure on the line. However, we would not like to go into details here.

9. *Random variable.* On any probability space, we usually deal with probabilities and random variables. Probabilities being already defined, it remains to say what is precisely random variable. By definition, this is any function, say, ξ, on Ω, with values from R^1 (more general spaces of values are possible, but this will be, hopefully, discussed later), such that for every $x \in R^1$, the set $\{\omega \in \Omega : \ \xi(\omega) \le x\} \in \mathcal{F}$. In the other words, for every $x \in R^1$ there should be defined probability

$P(\omega \in \Omega : \ \xi(\omega) \le x)$; for brevity, the latter object is usually written as $P(\xi \le x)$; the usual abbreviation r.v. means random variable[2].

10. *Sigma–algebra generated by random variable.* Let ξ be a r.v. on $(\Omega, \mathcal{F}, P)$. Sometimes we will need the object denoted by $\mathcal{F}^{\xi}$, which is called *sigma–algebra generated by the r.v.* ξ. By definition, this is the minimal sigma–algebra containing all sets from the family $\{\omega \in \Omega : \xi(\omega) \le x\}$, $x \in R^1$.

11. *Expectation* of r.v. This is a well-known object denoted by $E\xi$. We are to give a formal and rigorous definition, and here we are: for any $\xi \ge 0$,

$$E\xi := \lim_{N\to\infty} \sum_{k=0}^{N\times 2^N} \frac{k}{2^N}\, P\left(\frac{k}{2^N} \le \xi < \frac{k+1}{2^N}\right).$$

It easily follows (An Easy Exercise) that the limit always exists, – although may take value $+\infty$, – because the sequence under the limit is increasing.

For a general r.v. ξ,

$$E\xi := E\xi_+ - E\xi_-,$$

where $\xi_+ = \xi\, 1(\xi \ge 0)$, $\xi_- = |\xi|\, 1(\xi \le 0)$. In this definition we actually follow [Kolmogorov, Basic notions of probability theory, 1934.][3] This definition is equivalent to the use of Lebesgue integration, however, it looks as if we avoided any mentioning of it. Of course, it is just hidden somewhere, but still the result is that as if Lebesgue integration did not show up.

[2] For advanced reading, it follows from our definition that the function ξ is *measurable* as a function from *measurable space* $(\Omega, \mathcal{F})$ to another measurable space $(R^1, \mathcal{B})$, where the latter couple is the real line equipped with the topology of Borel sets. We will not discuss this here.

[3] Of course, details of the theory of measurable sets and Lebesgue integration are put aside here. However, from this more general point of view, our definition is exactly Lebesgue integral $\int \xi(\omega)P(d\omega)$. Below we give another, a bit unusual elementary interpretation of the same integral; there is some evidence (due to a remark in an S. N. Bernstein's textbook, if the author's memory does not fail) that Kolmogorov was aware of this approach in 30s, but argued that "conventional" Lebesgue integration is far easier. The latter is, of course, correct, however, this "another way" is a fun.

12. *Another standard notation* for the object $E\xi$ is

$$\int \xi(\omega)\, P(d\omega).$$

In the sequel, we may use both as synonyms. The rationale for the latter form is that in the discrete state space, clearly, the two definitions

$$E\xi = \sum_{x_k} x_k\, P(\xi = x_k)$$

and

$$E\xi = \sum_{\omega_i} \xi(\omega_i)\, p_i$$

are equivalent.

13. *Change of variables.* This is a formula often accepted as a definition in undergraduate textbooks; however, the reality is that this is a theorem to be proved (not here), and the sense is that it is, indeed, a change of variables type result. We formulate it firstly in a restricted form, for r.v.'s with a density. Suppose p is a density of the r.v. ξ. Then, for any non-negative function f such that $f(\xi)$ is a random variable,

$$Ef(\xi) = \int f(x)p(x)\, dx.$$

More generally, using notation from the previous paragraph, this may be rewritten as

$$Ef(\xi) = \int f(\xi(\omega))P(d\omega),$$

and in this form it does not depend on whether the density p exists or not. Naturally, if f may change sign, we can split f as $f = f_+ - f_-$, and use the additivity,

$$\int f(x)p(x)\, dx = \int f_+(x)p(x)\, dx - \int f_-(x)p(x)\, dx,$$

or

$$Ef(\xi) = \int f_+(\xi(\omega))P(d\omega) - \int f_-(\xi(\omega))P(d\omega),$$

of course, under an assumption that both integrals in the right hand side here are well-defined.

14. *Advanced reading: another approach to $E\xi$,* or how it is possible to reduce the Lebesgue integral $E\xi$ to Riemann's one. Emphasize that we are not to redefine any Lebesgue integral, but just $E\xi$, which is, however, general enough. We again split ξ into positive and negative parts, i.e., $\xi = \xi_+ - \xi_-$. Thus, it suffices to define what is $E\xi_+$. Hence, assume from the beginning that $\xi \geq 0$. Now, what we are looking for, is often denoted by $\int_{[0,\infty)} x dF_\xi(x)$, where F_ξ is the cumulative distribution function of ξ, i.e., as[4] $F_\xi(x) := P(\xi \leq x)$. This can be treated as Stiltjes (more precisely, Lebesgue–Stiltjes) type integral, however, we prefer to avoid this way. Instead, we consider the c.d.f. $F = F_\xi$. This is, clearly, an increasing function of x. As such, it may have not more than a finite or countable number of jumps, say, at points a_k, $k \geq 1$. Suppose, the jumps are of corresponding sizes $F(a_k) - F(a_k-) = \delta_k$, $k \geq 1$. Hence, if we subtract $F^c(x) := F(x) - \sum_{k:\, a_k \leq x} \delta_k$, the result is an increasing and *continuous* function. Now, define

$$\int_0^\infty x\, dF(x) := \sum_k a_k \delta_k + \int_0^\infty x\, dF^c(x),$$

next, clearly,

$$\int_0^\infty x\, dF^c(x) = \lim_{N\to\infty} \int_{(0,N]} x\, dF^c(x)$$

(remind that $F^c(0) - F^c(0-) = F^c(0) - 0 = 0$), and, finally,

$$\int_{(0,N]} x\, dF^c(x) := N\, F^c(N) - \int_{(0,N]} F^c(x)\, dx,$$

where the latter integral is well defined as Riemenn's one. Of course, we should remember that at any step here the result may turn out to be $+\infty$, but this is possible in any approach.

15. *Radon–Nikodym Theorem.* The last tool that we have to introduce in this chapter is Conditional expectations. For that aim, we need Radon–Nikodym Theorem (accents on both "o"). Suppose ν is a *signed measure* on $(\Omega, \mathcal{F})$, that is, a sigma-additive function on $\mathcal{F}$. It is called *absolutely continuous with respect to* P, if $P(A) = 0$ always implies $\nu(A) = 0$, notation $\nu << P$. The Radon–Nikodym Theorem (not to be

[4] There are two schools, one uses "$\leq$" here, another "$<$". Take care. With our definition, the function F_ξ is continuous from the right.

proved here) claims that if $\nu << P$, then there exists an $\mathcal{F}$–measurable function f called density, – here measurability means that for each x, we have $\{\omega : f(\omega) \le x\} \in \mathcal{F}$, – such that for any $A \in \mathcal{F}$,

$$\nu(A) = \int f(\omega)P(d\omega).$$

This density is unique up to P–a.s. (almost surely). Emphasize that in applications $\mathcal{F}$ may change, as it will in the next paragraph.

16. *Conditional expectation.* Suppose $E\xi < \infty$, and η is another r.v. We call the r.v. ζ *conditional expectation* of ξ given η, with notation $E(\xi \mid \eta)$, or, equivalently, $E(\xi \mid \mathcal{F}^\eta)$, if

    - ζ is measurable with respect to $\mathcal{F}^\eta$, that is, for each x, we have $\{\omega : \zeta(\omega) \le x\} \in \mathcal{F}^\eta$,
    - for any $A \in \mathcal{F}^\eta$,
    $$E\zeta 1(A) = E\xi 1(A).$$

    Since $P(A) = 0$ implies $E1(A)\xi = 0$, the signed measure $\nu(A) := E1(A)\xi$ is absolutely continuous with respect to P on $\mathcal{F}^\eta$ (An Easy Exercise). So, by the R-N Theorem, it has a density ζ, which is $\mathcal{F}^\eta$–measurable, such that for each $A \in \mathcal{F}^\eta$,

    $$\int 1(A)\zeta P(d\omega) = \nu(A) = \int 1(A)\xi P(d\omega).$$

    Hence, this density *is* a conditional expectation of ξ given η, by definition.

    Emphasize that conditional expectation is defined almost surely, i.e., up to a possible change of values on any P–zero event on $\bar{\mathcal{F}}$. For discrete time processes this, of course, is of no harm, but for continuous time precautions have to be made, such as establishing existence of *regular conditional probabilities* (not in this course).

17. *Conditional expectation with respect to some sigma–algebra included in $\mathcal{F}$ but maybe not generated by any r.v.* can be defined, too, absolutely similarly as to the previous paragraph, via the R–N Theorem, however, we will not use it in this course. (Or, if we will use it, we will revise this point later.)

18. *Why consider $\mathcal{F}$ other than 2^Ω in finite state case.* The answer is in the previous paragraph: it may be important to consider sigma-algebras generated by some r.v.'s, for example, because it may be important to compute conditional expectations of some r.v.'s with respect to some others. But those sigma-algebras generated by r.v.'s are usually different from 2^Ω.

19. *Literature:* ((19b) and (19c) for advanced reading)

    (a) Kolmogorov, A. N. Basic notions of probability

    (b) Kolmogorov, A. N., Fomin, S. V. Elements of functional analysis,

    (c) Krylov, N. V.

    (d) Shiryaev, A. N. Probability (any edition). Springer, London et al.,

    (e) Stirzaker

    (f) Wentzell, A. D., A course on stochastic processes.

2 Markov processes, invariant measures, uniform stabilization (ergodic theorem)

2.1 MP

1. *Def. of MP (MC):* $\boxed{P(X_{t+1} \le x \mid X_t) = P(X_{t+1} \le x \mid X_t, X_{t-1}, \ldots, x).}$

2. *First warning.* First of all, there are two, closely connected yet different, notions of Markov process[5] in the literature. One relates, loosely speaking, to *one process* which possesses the so called *Markov property formulated in terms of conditional expectations*, which we discuss below. Another notion relates to a *family of measures* with arbitrary starting points (i.e. "initial measure" could be any delta–measure, δ_x, and later even any distribution, perhaps not concentrated at one point), with a *Markov property formulated in terms of those measures.* Not that those two Markov properties are very different, they just relate to different objects (and, hence, of course, they are different). Some authors

[5] Of course, they are sometimes mixed.

use another term to distinguish, calling the first kind of such processes "Markov stochastic functions", while retaining the original term for the second case.

Example of mixing those two different cases is if you read a sentence like "this Markov process has a stationary version (regime)", which, in particular, means another starting distribution. But a given process with a given initial distribution cannot have another initial distribution, or it will be a new (Markov stochastic) process, with *the same transition matrix* (see below). The sentence could be correctly understood via the second notion of Markov family. The classical monograph on the second case is [Dynkin, E. B. Markov processes].

We will be studying the first case, however, having in mind that, if necessary, it is possible to acquire a wider point of view.

3. *Second warning.* The Markov property is often, – nearly invariably, – formulated in textbooks as "future does not depend on the past given present", or likewise. Although this is a good intuitive description of the precise definition (below) via formulas, it also may be very misleading. By independence we must understand only some probabilistic equalities, but certainly not *functional independence.* In the other words, *trajectories* of a Markov process in the "future" and in the "past" may strongly depend, even given "present", however, this may have nothing to do with Markov property or its lack. Because Markov property is *only and no more than* some probability equations, which may allow however strong functional dependence, so to say, in the algorithm which defines the evolution of the process. On the other hand, functional independence, if it is present, often helps to show Markov property indeed[6].

4. *Transition matrices.* For $m \leq n$, transition from time m to time n is,

$$T(m,n) = (p(m,i;n,j) = P(X_n = j \mid X_m = i))_{i,j=1}^{N}\,.$$

[6] I am sure this is not my invention, but for my undergraduate course on Markov chains I introduced a notion of *simple algorithm*, i.e. of evolution with functional independence of future and past given present. Such algorithms provide Markov property which does not require any additional verification (which, of course, can be done easily, but it is not easy at the UG level: some simple tasks may be difficult to do).

5. *Homogeneous MP's, def.* MP is called homogeneous if the matrices $T(m, m+1)$ does not depend on m. If so, we consider the *transition matrix*
$$T := T(0,1) \quad (\equiv T(m, m+1)).$$

6. *Chapman–Kolmogorov's equations.* We denote,
$$T^{(n)} = (P(X_n = j \mid x = i))_{i,j=1}^{N} \equiv T(0,n).$$
Then, there is a "micro-theorem",
$$T^{(n)} = T^n,$$
or
$$T^{(m+k)} = T^{(m)} T^{(k)}.$$
This is one version of Chapman–Kolmogorov.

   Equivalently, we have a more standard version of it,
$$p_{ij}^{(m+k)} = \sum_{\ell} p_{i\ell}^{(m)} p_{\ell j}^{(k)}.$$
The proof consists of the reference on complete probability formula combined with Markov property.

7. *Stationary measures.* The *probability measure* μ is called stationary for the homogeneous MP with a given transition matrix T, if all *marginal distributions* $\mu_n = (\mu_n(1), \ldots, \mu_n(N)) := (P_\mu(X_n = 1), \ldots, P_\mu(X_n = N))$ do not depend on n.

8. *An Easy Exercise:* Show that equivalently the property of stationarity may be expressed as
$$\boxed{\mu = \mu T.} \qquad \text{(equivalent def.)}$$
(In the r.h.s. here the row vector is multiplied by the matrix from the right, and the result is again a row vector of the same dimension N.)

2.2 Uniform ergodic theorem(s)

1. **Theorem 1** *Let X_n be a (homogeneous) Markov chain with transition matrix T, of size $N \times N$. Then there exists at least one stationary probability measure.*

2. *Proof.* For each i_0, consider the sequence of vectors,

$$\left(\frac{1}{n+1}\sum_{k=0}^{n} p_{i_0 j}^{(k)}\right), \quad n \geq 0.$$

Since this is a bounded sequence, for some sub-sequence $n' \to \infty$ there exists a limit $(\pi_j,\ 1 \leq j \leq N)$, that is,

$$\left(\frac{1}{n'+1}\sum_{k=0}^{n'} p_{i_0 j}^{(k)}\right) \to \pi_j \quad (1 \leq j \leq N), \quad n' \to \infty.$$

Since the state space is finite, the vector $(\pi_j,\ 1 \leq j \leq N)$ is a probability measure, i.e., all $\pi_i \geq 0$, and $\sum_i \pi_i = 1$. Due to Chapman–Kolmogorov,

$$\frac{1}{n'+1}\sum_{k=0}^{n'} p_{i_0 j}^{(k)} = \frac{1}{n'+1}\sum_{k=0}^{n'}\sum_{\ell=1}^{N} p_{i_0 \ell}^{(k-1)} p_{\ell j} + \frac{1}{n'+1} p_{i_0 j}^{(0)}$$

$$= \sum_{\ell=1}^{N} \frac{1}{n'+1}\sum_{k=0}^{n'-1} p_{i_0 \ell}^{(k)} p_{\ell j} + \frac{1}{n'+1}\delta_{i_0 j}$$

$$= \sum_{\ell=1}^{N} p_{\ell j} \frac{1}{n'+1}\sum_{k=0}^{n'} p_{i_0 \ell}^{(k)} + \frac{1}{n'+1}\delta_{i_0 j} - \frac{1}{n'+1}\sum_{\ell=1}^{N} p_{\ell j} p_{i_0 \ell}^{(n')}.$$

(δ_{ij} is a Kronecker symbol). This implies for every j,

$$\lim_{n'\to\infty} \frac{1}{n'+1}\sum_{k=0}^{n'} p_{i_0 j}^{(k)} = \sum_{\ell=1}^{N} p_{\ell j}\pi_\ell$$

Hence,

$$\pi_j = \sum_{\ell=1}^{N} p_{\ell j}\pi_\ell \quad (\forall j) \quad \sim \quad \pi = \pi T,$$

i.e., the distribution (π_j) is stationary. The Theorem 1 is proved[7].

[7] The method is named after N. M. Krylov and N. N. Bogoliubov, suggested in 1930s.

3. **Theorem 2** *Let X_n be a (homogeneous) Markov chain with transition matrix T, of size $N \times N$, where all entries p_{ij} are positive. Then there is a unique stationary probability measure $\pi = (\pi_1, \ldots, \pi_N)$, and, moreover,* $\inf_j \pi_j > 0$, *and*

$$\sup |p_{ij}^{(n)} - \pi_j| \le (1-\kappa)^n, \tag{1}$$

where

$$\kappa = \inf_{i,j} \sum_{\ell} p_{i\ell} \wedge p_{j\ell}. \tag{2}$$

4. *Comment.* From the proof below and from the formulation of the theorem itself it follows easily that (2) suffices for (1), as well as for uniqueness of stationary measure (although, perhaps, not for positiveness of all π_j's).

5. *"Kolmogorov's" Proof.*[8] (A) Denote for any A,

$$m^{(n)}(A) := \min_i P_i(n, A), \quad M^{(n)}(A) := \max_i P_i(n, A).$$

By Chapman–Kolmogorov,

$$m^{(n+1)}(A) = \min_i P_i(n+1, A) = \min_i \sum_j p_{ij} P_j(n, A)$$

$$\ge \min_i \sum_j p_{ij} \min_{j'} P_{j'}(n, A) = m^n(A),$$

which signifies that the sequence $m^n(A)$ does not decrease in n. Similarly, the sequence $M^n(A)$ does not increase in n. Hence, it suffices to show that

$$M^{(n)}(A) - m^{(n)}(A) \le (1-\kappa)^n. \tag{3}$$

(B) Again by Chapman–Kolmogorov,

$$M^n(A) - m^n(A) = \max_i P_i(n, A) - \max_{i'} P_{i'}(n, A)$$

$$= \max_i \sum_j p_{ij} P_j(n-1, A) - \max_{i'} \sum_j p_{i'j} P_j(n-1, A).$$

[8] On the West, apparently the author of this proof is unknown, although W. Doeblin may be sometimes mentioned as its inventor; in fact, another proof belongs to him, which will be presented in the next chapter. In Russian literature the present proof is attributed to Kolmogorov, with the standard reference on some (not every) editions of B. V. Gnedenko's famous textbook on Probability.

Let max here be attained at i_+ while min at i_-. Then,

$$M^n(A) - m^n(A) = \sum_j p_{i_+j} P_j(n-1, A) - \sum_j p_{i_-j} P_j(n-1, A)$$

$$= \sum_j (p_{i_+j} - p_{i_-j}) P_j(n-1, A). \qquad (4)$$

(C) Denote by S^+ the part of the sum in the right hand side of (4) with just $(p_{i_+j} - p_{i_-j}) \geq 0$, and by S^- the part of the sum with $(p_{i_+j} - p_{i_-j}) < 0$. We estimate,

$$S^+ \leq \sum_j (p_{i_+j} - p_{i_-j})_+ M^{(n-1)}(A),$$

and

$$S^- \leq -\sum_j (p_{i_+j} - p_{i_-j})_- m^{(n-1)}(A).$$

Therefore,

$$M^{(n)}(A) - m^{(n)}(A) = S^+ + S^-$$

$$\leq M^{(n-1)}(A) \sum_j (p_{i_+j} - p_{i_-j})_+ + m^{(n-1)}(A) \sum_j (p_{i_+j} - p_{i_-j})_-.$$

(D) It remains to notice that

$$\sum_j (p_{i_+j} - p_{i_-j})_- = -\sum_j (p_{i_+j} - p_{i_-j})_+, \qquad (5)$$

and

$$\sum_j (p_{i_+j} - p_{i_-j})_+ \leq 1 - \kappa. \qquad (6)$$

The first follows from the normalization condition

$$\sum_j p_{i_+j} = \sum_j p_{i_-j} = 1,$$

while the second from

$$\sum_j (p_{i_+j} - p_{i_-j})_+ = \sum_j (p_{i_+j} - \min(p_{i_-j}, p_{i_+j})) \leq 1 - \kappa.$$

So, we find that

$$M^{(n)}(A) - m^{(n)}(A) \le (1-\kappa)\,(M^{(n-1)}(A) - m^{(n-1)}(A)).$$

By induction this implies (3). So, (1) as well as uniqueness of the limits $\pi_j = \lim_{n\to\infty} p_{ij}^{(n)}$ follow. Uniqueness of stationary measure, in turn, follows from (1) (start from "another" stationary distribution μ, then $\mu_j \equiv P_\mu(X_n = j) = \sum_\ell \mu_\ell p_{ij}^{(n)} \to \pi_j,\ n \to \infty$). The Theorem 2 is proved.

3 More on integration

3.1 Integration: Fubini Theorem

We often perform double integration: it could be expectation of a sum, or two sums, or sum of integrals, etc., like $E\sum_k \xi_k$. We often have to change the order of summations, integrations, etc. We formulate (without proof) one popular version of a result of this sort.

Theorem. [Fubini] Let $(\xi_k,\, k \ge 1)$ be a sequence of r.v.'s. Then,

$$E\sum_{k=1}^{\infty} \xi_k = \sum_{k=1}^{\infty} E\xi_k,$$

if the series $\sum_{k=1}^{\infty} E|\xi_k|$ converges.

3.2 Probabilities = expectations

Indicator functions: If $A \in \mathcal{F}$, then we define a *random variable*

$$1(A) = 1(A)(\omega) := \left\{ \begin{array}{ll} 1, & \omega \in A, \\ 0, & \omega \notin A. \end{array} \right.$$

This r.v. is called indicator of A.

An Easy Exercise: show that this is, indeed, a random variable.

Use of indicator functions: we can replace probabilities by expectations and vice versa,

$$\boxed{E1(A) = P(A)} \quad \sim \quad \boxed{P(A) = E1(A)}$$

3.3 Conditional expectation as function of condition

Measurability Lemma. Let ξ, η be two r.v.'s and $E\xi < \infty$. Then for the expression $E(\xi \mid \eta)$ there exists a measurable (Borel) function h such that

$$E(\xi \mid \eta) = h(\eta) \qquad \text{a.s.}$$

Def. Borel function h is a function $h : R^1 \mapsto R^1$ measurable with respect to Borel sigma-algebras $\mathcal{B}$ in the domain and the image space, that is, for any $a \in R^1$, $\{x : h(x) \le a\} \in \mathcal{B}$.

Def. $\mathcal{B}$, the Borel sigma–algebra on R^1, is the σ–algebra generated by all open intervals. But $\mathcal{B}$ is *not* completed by any reasonable measure!

Idea of proof: construct firstly h_N such that $1(k/2^N \le \eta < (k+1)/2^N) \times E(\xi \mid k/2^N \le \eta < (k+1)/2^N) = h_N(\eta)$, then let N go to ∞.

4 Stopping times

4.1 Filtrations, Stopping Times

A big advantage of theory of stochastic processes in comparison to (formally equivalent) measure theory is such a natural tool as *path approach.* E.g., we may track some trajectory (i.e. given ω) until it attains some prescribed state, and study questions like how long on average this takes. Measure theory does allow this, but it looks so unnatural!

Filtration denotes a family of increasing sigma-algebras $(\mathcal{F}_t^X,\, t \ge 0)$, where $\mathcal{F}_t^X = \sigma(X_s,\, s \le t)$.

A r.v. $\tau \in [0, +\infty]$ is called *stopping time* for the MC X iff for every (nonrandom) $t \ge 0$, we have $\{\tau > t\} \in \mathcal{F}_t^X$. In the other words, by time t we can decide if τ "occurred or not".

An Easy Exercise: given any $x \in R$, the r.v. $\tau : \inf(t \ge 0 : X_t \ge x)$ is a stopping time (with a natural convention $\inf(\emptyset) = +\infty$).

4.2 X_τ – stopped MP

In the sequel, we will deal with processes at random times, such as X_τ, where, say, τ is a stopping time. First question is whether this is just a random variable! But where is the problem? Here: we have $X_n : \Omega \mapsto R^1$, and this is a random variable by definition, i.e. $\{\omega : X_n(\omega) \le x\} \in \mathcal{F}$. Now we have some complications: $X_{\tau(\omega)}(\omega)$, hence, this is a composite function. Why then $\{\omega : X_{\tau(\omega)}(\omega) \le x\} \in \mathcal{F}$?

It turns out, however, that in discrete time the problem disappears (unlike in continuous!).

5 Strong MP

5.1 Strong Markov processes

Definition. Let X be a MP. It is called *strong Markov* iff for every $x \in R$,

$$P(X_{\tau+1} \le x \mid \mathcal{F}_\tau) = P(X_{\tau+1} \le x \mid X_\tau).$$

Sigma-algebra $\mathcal{F}^X_\tau$ is defined as follows,

$$\mathcal{F}^X_\tau := \{A \in \mathcal{F} :\, A \bigcap \{\tau > t\} \in \mathcal{F}^X_t, \quad \forall\, t\}.$$

In discrete time, all MP's are *strong* MP's. Crucial is that τ may take not more than countable number of values.

5.2 Strong MP: Theorem

Remind that time is discrete.
Theorem. Let X be a MP, and τ some stopping time. Then for every $x \in R^1$,

$$P(X_{\tau+1} \le x \mid \mathcal{F}^X_\tau) = P(X_{\tau+1} \le x \mid X_\tau),$$

that is, the r.h.s. is a version of the conditional probability from the l.h.s. Equivalently, it suffices to show (a) measurability $P(X_{\tau+1} \le x \mid X_\tau) \in \mathcal{F}^X_\tau$, which is easy (is it?), and (b) that for every $A \in \mathcal{F}^X_\tau$,

$$P(X_{\tau+1} \le x; A) = E1(A)P(X_{\tau+1} \le x \mid X_\tau). \tag{7}$$

The idea is to split $1 = \sum_k 1(\tau = k)$ and use the standard Markov property. But how can we plug in k instead of τ in $P(X_{\tau+1} \le x \mid X_\tau)$? It does not look really explicit.

5.3 Hint

Let us prove that (it will help show (7))

$$P(X_{\tau+1} \le x \mid X_\tau) = \sum_k 1(\tau = k) P(X_{k+1} \le x \mid X_k). \tag{8}$$

The r.h.s. here $\in \sigma(X_\tau)$. (Let $g(X_k) = P(X_{k+1} \le x \mid X_k)$, then r.h.s. equals $\sum 1(\tau = k) g(X_k) = \sum 1(\tau = k) g(X_\tau) = g(X_\tau) \in \sigma(X_\tau)$.) Take any $B \in \sigma(X_\tau)$. Then, $1(B) = h(X_\tau)$ for some measurable h. Hence, – and this shows (8), –

$$\begin{aligned}
E1(B) \times \text{r.h.s.} &\overset{\text{Fubini}}{=} \sum_k Eh(X_\tau) 1(\tau = k) P(X_{k+1} \le x \mid X_k) \\
&= \sum_k Eh(X_k) 1(\tau = k) P(X_{k+1} \le x \mid X_k) \\
= \sum_k E1(B) 1(\tau = k) 1(X_{k+1} \le x) &\overset{\text{Fubini}}{=} E1(B) \times \text{l.h.s.}
\end{aligned}$$

5.4 Proof of the Theorem

Proof. We have, due to (8) and because $1(A)1(\tau = k) \in \mathcal{F}^X_k$, that the r.h.s. from (7) can be presented as

$$\begin{aligned}
& E1(A) P(X_{\tau+1} \le x \mid X_\tau) \\
&= E1(A) \sum_k 1(\tau = k) P(X_{k+1} \le x \mid X_k) \\
&\overset{\text{Fubini}}{=} \sum_k E1(A) 1(\tau = k) P(X_{k+1} \le x \mid \mathcal{F}^X_k) \\
&= \sum_k E1(A) 1(\tau = k) 1(X_{k+1} \le x) \\
&\overset{\text{Fubini}}{=} E \sum_k 1(A) 1(\tau = k) 1(X_{\tau+1} \le x) = P(X_{\tau+1} \le x; A).
\end{aligned}$$

6 Probability spaces: direct products

Let $(\Omega^1, \mathcal{F}^1, P^1)$ and $(\Omega^2, \mathcal{F}^2, P^2)$ be two probability spaces. We can construct a new probability space which is called *direct product* of them,

$$(\Omega, \mathcal{F}, P) := (\Omega^1, \mathcal{F}^1, P^1) \times (\Omega^2, \mathcal{F}^2, P^2).$$

The recipe is as follows. A. $\Omega = \Omega^1 \times \Omega^2$, which simply means that new outcomes are couples, $\omega = (\omega^1, \omega^2)$, with $\omega^i \in \Omega^i$. B. $\mathcal{F} = \mathcal{F}^1 \times \mathcal{F}^2$, which is understood as $\mathcal{F} := \sigma(A^1 \times A^2,\ A^i \in \mathcal{F}^i,\ i = 1, 2)$. Finally, measure P is determined by $P(A^1 \times A^2) := P^1(A^1)P^2(A^2)$. Here we need the **Theorem** [Kolmogorov] about unique extension of a measure from a *semi-ring* to sigma-algebra generated by this semi-ring. We do not discuss it here, see [??].

Let

$$(\Omega, \mathcal{F}, P) = (\Omega^1, \mathcal{F}^1, P^1) \times (\Omega^2, \mathcal{F}^2, P^2).$$

Suppose we have r.v. ξ^i on $(\Omega^i, \mathcal{F}^i, P^i)$. We now can *extend* or re-define both ξ^i on a new probability space $(\Omega, \mathcal{F}, P)$, as follows,

$$\xi^1(\omega^1, \omega^2) := \xi^1(\omega^1), \qquad \xi^2(\omega^1, \omega^2) := \xi^2(\omega^2).$$

Then, (A) distribution of ξ^i on the extended space is the same as on the original one; (B) both r.v.'s are now defined on the same probability space, so we *can* ask whether they are (or are not) independent. The answer is that they are, indeed, independent.

For the finite sate space MP X under the assumptions of Ergodic Theorem $\min p_{ij} > 0$, we can use the following idea. Consider *another* probability space, and on that space another MP $\tilde{X}$ with the *same* transition probabilities, which is *stationary* (which exists, due to the simple Theorem 1 about stationary measures). Let $\tilde{\kappa} := \max_j \min_i p_{ij} =: \min_i p_{ij_0}$.

On *direct product* of the two probability spaces, both processes remain MP. However, now we can define such an object as their *first meeting time* $\tau := \inf(t \geq 0 : \ X_t = \tilde{X}_t)$. This is a *stopping time* (with respect to which filtration?). Let

$$\hat{X}_t := X_t \, 1(t < \tau) + \tilde{X}_t \, 1(t \geq \tau).$$

Lemma 1. $\hat{X}$ is a MP equivalent to X. (An Easy Exercise)

Lemma 2. For each $A \subset \mathcal{S}$,

$$|P(X_n \in A) - P(\tilde{X}_n \in A)| \leq (1 - \tilde{\kappa}^2)^n.$$

Proof.

$$|P(X_n \in A) - P(\tilde{X}_n \in A)| = |P(\hat{X}_n \in A) - P(\tilde{X}_n \in A)|$$

$$= |E1(\hat{X}_n \in A) - E1(\tilde{X}_n \in A)| = |E(1(\hat{X}_n \in A) - 1(\tilde{X}_n \in A))|$$

$$\le E|(1(\hat X_n \in A) - 1(\tilde X_n \in A))|1(n<\tau) \le E1(n<\tau)$$

$$\le (1 - \min_{i,i'} P(\hat X_1 = \tilde X_1 = j_0 \mid \hat X_0 = i,\, \tilde X_0 = i'))^n = (1-\tilde\kappa^2)^n.$$

The bound of Lemma 2 is not optimal. Here is another one.

Lemma 2'. Let $\bar\kappa := \min_{ij} p_{ij}$. For each $A \subset \mathcal{S}$,

$$|P(X_n \in A) - P(\tilde X_n \in A)| \le (1-\bar\kappa)^n.$$

Proof.

$$|P(X_n \in A) - P(\tilde X_n \in A)| \le P(n<\tau)$$

$$\le (1 - \min_{i} P(\hat X_1 = \tilde X_1 \mid \tilde X_0 = i))^n \le (1-\bar\kappa)^n.$$

Is it really an improvement? I.e., is $\tilde\kappa^2 < \bar\kappa$? (And how it compares to $\kappa := \min_{ii'} \sum_\ell p_{i\ell} \wedge p_{i'\ell}$?) Maybe yes, maybe no, although the *order* has improved, if, say, all p_{ij} are small. Continue our efforts to improve the bounds.

The bound of Lemma 2' is not optimal either. Let us do one more try to improve it. Let $\hat\kappa := \min_{ii'} \sum_k p_{ik}p_{i'k}$.

$$P(X_1 \ne \tilde X_1 \mid X_0 = i, \tilde X_0 = i') = E_{ii'} 1(X_1 \ne \tilde X_1)$$

$$= \sum_k E_{ii'} P(X_1 \ne k \mid \tilde X_1) 1(\tilde X_1 = k)$$

$$= \sum_k E_{ii'} P(X_1 \ne k) 1(\tilde X_1 = k) = \sum_k (1-p_{ik})p_{i'k} =$$

$$= 1 - \sum_k p_{ik}p_{i'k} \le 1 - \hat\kappa.$$

Still, the bound is worse than in the first proof of Ergodic Theorem. How to achieve that bound with κ?

The ultimate improvement may be attained on the following way. Let us reconstruct the process $\hat X$, so that it will have more chances to meet with $\tilde X$ at each step, remaining a MP with the same transition matrix. This turns out to be possible.

Lemma. Let ξ^1, ξ^2 be two r.v.'s, each on its own probability space, with

$$\min_{ii'} \sum_i p_{ij} \wedge p_{i'j} = \kappa \ (> 0). \tag{9}$$

Then there exists a new probability space which is some extension of the direct product of the first two, on which there exists a third r.v. ξ^3, such that distribution of it coincides with the distribution of ξ^1, and, at the same time,

$$P(\xi^3 = \xi^2) = \kappa.$$

$$\text{Let} \quad C := \left\{ \frac{p^1(\xi^2)}{\max(p^1(\xi^2), p^2(\xi^2))} \ge \zeta \right\}, \quad \zeta \sim U[0,1].$$

Define on some independent probability space a r.v. $\eta \sim p^\eta(\cdot) = (p^1 - p^1 \wedge p^2)(\cdot) / \int (p^1 - p^1 \wedge p^2)$, and

$$\xi^3 := \xi^2 1(C) + \eta 1(\bar{C}).$$

Another equivalent representation of ξ^3 with a domain $S_0 = \{(x,y) : \ 0 \le y \le p^1 \wedge p^2(x)\}$, reads,

$$\xi^3 := \xi^2 1((\xi^2, \zeta \ p^2(\xi^2)) \in S_0) + \eta 1((\xi^2, \zeta \ p^2(\xi^2)) \not\in S_0).$$

$$\text{Apparently,} \quad P(\xi^3 = \xi^2) \ge P((\xi^2, \zeta \, p^2(\xi^2)) \in S_0)$$
$$= \int_{p^2 \le p^1} p^2(x) \, dx + \int_{p^2 > p^1} p^2(x) \frac{p^1}{p^2}(x) \, dx = \kappa$$

(in fact, $>$ in the first line is not possible, *An Easy Exercise*).
Next, for any bounded (Borel) function f,

$$\begin{aligned} Ef(\xi^3) = Ef(\xi^2) P(\zeta \le p^1 / \max(p^1, p^2)(\xi^2)) \\ + Ef(\eta) P(\zeta > p^1 / \max(p^1, p^2)(\xi^2)) \\ = Ef(\xi^2) \, p^1 / \max(p^1, p^2)(\xi^2) \\ + Ef(\eta) E(1 - p^1 / \max(p^1, p^2)(\xi^2)) \\ = Ef(\xi^1), \end{aligned}$$

$$\text{because} \qquad Ef(\xi^2)\, p^1/\max(p^1,p^2)(\xi^2)$$

$$= \int_{p^1\ge p^2} f(x)p^2(x)\,dx + \int_{p^1<p^2} f(x)\frac{p^1}{p^2}\,p^2(x)\,dx$$

$$= \int_{p^1\ge p^2} f(x)p^2(x)\,dx + \int_{p^1<p^2} f(x)p^1(x)\,dx,$$

$$\text{and} \qquad Ef(\eta)E(1-p^1/\max(p^1,p^2)(\xi^2))$$

$$= \int f(x)(p^1 - p^1\wedge p^2)(x)(1-\kappa)^{-1}\,dx \times$$

$$\times \int_{p^1<p^2} (1-p^1/p^2)p^2(x)\,dx$$

$$= \int_{p^1\ge p^2} f(x)(p^1-p^2)(x)\,dx.$$

Now, for each step of our MC, using Lemma of three random variables, we *reconstruct* our MC X_n so that it is an equivalent MC, but on each step the probability to coincide with $\tilde X_n$ (stationary version) is at least $\kappa = \min_{ij}\sum_\ell p_{i\ell}\wedge p_{j\ell}$. Hence, probability $P(\tau > n)$ not to meet until time n admits a bound

$$P(\tau > n) \le (1-\kappa)^n.$$

This finishes the proof of Ergodic Theorem with the same exponential bound as in the Kolmogorov proof.

Remark. Although we worked with finite state spaces, there is no change in the above proof for countable ones under the assumption (9).

In Probability an important role belongs to coefficients of *mixing* (weak dependence), in particular, to φ–mixing or *uniformly strong mixing* coefficient [by I. Ibragimov],

$$\varphi(n) := \sup_k \sup_{A\subset\mathcal{S}} |P(X_{k+n}\in A) - P(X_{k+n}\in A \mid \mathcal{F}_k)|.$$

We say that the process (X_n) is φ–mixing iff its φ coefficient satisfies

$$\varphi_n \to 0, \qquad n\to\infty.$$

Ergodic Theorem for finite MC's can be formulated as

$$\boxed{\varphi_n \le (1-\kappa)^n.}$$

However, φ–mixing is applicable to much wider class of processes than MC's, finite or not.

Pétit sets condition; comparison with (9).

In the next chapters, for *non-compact state spaces*, we will use a local version of this condition. Another condition is quite popular in Ergodic theory for MP's with non-compact state spaces, namely, *pétit sets condition.* We have to realise that condition (9) is much more relaxed, and at the same ti;e provides a better rate of convergence. Currently we are doing compact state spaces, and it is appropriate to study this question here.

A "global" analogue of a pétit sets condition was used by Doob, and it assumes that there exist positive value $\epsilon > 0$ and a probability measure ν, such that for any $A \subset \mathcal{S}$ (this is a "globalization", we require that $\mathcal{S}$ itself is pétit)

$$P_i(X_1 \in A) \geq \epsilon\nu(A). \tag{10}$$

Then, the theory provides an exponential convergence bound (with the same notations as above)

$$|P(X_n \in A) - P(\tilde{X}_n \in A)| \leq (1-\epsilon)^n.$$

One trivial aspect of our comparison is, of course, that (10) implies (9) with

$$\boxed{\kappa \geq \epsilon.}$$

More than that, we can easily construct examples where ϵ is *arbitrarily small*, or *even equals zero*, while κ is bounded away from zero or even *arbitrarily close to one.* Let us start with a simple case, with $\mathcal{S} = (1, 2, 3)$ and

$$p_{k,i} = \frac{1}{2} 1(i \neq k).$$

Then, clearly,

$$\epsilon = 0, \qquad \kappa = \frac{1}{2}.$$

Next example, with $\mathcal{S} = (1, 2, 3, \ldots, N)$ and again

$$p_{k,i} = \frac{1}{2} 1(i \neq k).$$

Then, clearly,

$$\epsilon = 0, \qquad \kappa = \frac{N-2}{N-1}.$$

It may be noticed, however, that for the two step transition probability matrix in the first example,

$$\epsilon^{(2)} = \frac{3}{4}.$$

This provides practically the same convergence rate bound as via κ, namely,

$$|P(X_n \in A) - P(\tilde{X}_n \in A)| \le (1 - \epsilon^{(2)})^{[n/2]} \equiv (1 - \frac{3}{4}^{(2)})^{[n/2]} \equiv (\frac{1}{4})^{[n/2]}.$$

For n even, this is equal to 2^{-n}, as with κ. For n odd, however, there is a minor discrepancy which is negligible when $n \to \infty$. Notice also that $\kappa^{(2)} = \frac{3}{4}$, too. In any case, if we use κ, <u>we will always do better than or equal to</u> what we may obtain with ϵ.

$$\boxed{\text{In general, the Q if } 1 - \epsilon^{(n)} \text{ and } 1 - \kappa^{(n)} \text{ may be similar is open.}}$$

My hypothesis is that for finite state spaces they may *be similar, but for general state spaces κ may be always strictly better.*

Some other examples where using κ is much better could be provided by diffusion processes, however they relate to continuous time, so we do not touch them here.

7 LLN

Assumption (A1): Consider a MP satisfying the assumptions of Ergodic Theorem (with $\kappa > 0$).
Theorem 1: [stationary weak LLN] *For a stationary MC under (A1), for any f on the state space $\mathcal{S}$,*

$$\frac{1}{n}\sum_{k=0}^{n-1} f(X_k) \xrightarrow{P} E_{inv} f(X_0),$$

where E_{inv} stands for expectation with respect to the invariant measure, $E_{inv} f(X_0) = \sum_{j \in \mathcal{S}} f(j)\pi_j$.

Such results were one of the goals of A.A.Markov (1856–1922) himself when he introduced his Markov chains, i.e., he wanted to extend limit theorems from the usual IID scheme to some naturally dependent r.v.'s.

Proof. We will use Bienaimé–Chebyshev inequality with variance. Assume $E_{inv}f(X_0) = 0$, otherwise subtract.

$$P_{inv}(|\frac{1}{n}\sum_{k=0}^{n-1} f(X_k)| \ge \epsilon) \le \frac{E_{inv}|\sum_{k=0}^{n-1} f(X_k)|^2}{\epsilon^2 n^2}$$

$$= \frac{\sum_{k=0}^{n-1} E_{inv} f^2(X_k)}{\epsilon^2 n^2} + \frac{2\sum_{k<j}^{n-1} E_{inv} f(X_k) f(X_j)}{\epsilon^2 n^2}. \qquad (11)$$

$$\text{But} \qquad |E_{inv}f(X_k)f(X_j)| = |E_{inv}f(X_k)E(f(X_j) \mid X_k)|$$
$$\le \quad \text{(in fact, =)} \quad |E_{inv}f(X_k)E_{inv}(f(X_j)|$$
$$+|E_{inv}f(X_k)[E(f(X_j) \mid X_k) - E_{inv}(f(X_j)]|,$$

where $|E(f(X_j) \mid X_k) - E_{inv}(f(X_j)]| \le C_f(1-\kappa)^{j-k}$.

Remind that $E_{inv}f(X_k)E_{inv}(f(X_j) = 0$.

Thus, in (58) the first term equals C/n, while the second by modulus does not exceed (all C generic)

$$\frac{C\sum_{k=0}^{n-1}\sum_{j=k+1}^{n-1}(1-\kappa)^{j-k}}{n^2} \le \frac{C\sum_{k=0}^{n-1}\sum_{j-k=1}^{\infty}(1-\kappa)^{j-k}}{n^2} = \frac{C}{n}.$$

This completes the proof of *"Markov Chain LLN"*. Clearly, exponential convergence rate is more than enough, but for finite state MC there is no much choice.

Theorem 2: [non-stationary weak LLN] *Under (A1), for any f on the state space $\mathcal{S}$,*

$$\frac{1}{n}\sum_{k=0}^{n-1} f(X_k) \overset{P}{\to} E_{inv}f(X_0),$$

where E_{inv} stands for expectation with respect to the invariant measure, $E_{inv}f(X_0) = \sum_{j\in\mathcal{S}} f(j)\pi_j$.

Proof. Considering the stationary version $\tilde{X}$ along with the original MC X switched to the stationary after the first meeting (τ), denoted by $\hat{X}$ and equivalent to X, we have,

$$P(|\frac{1}{n}\sum_{k=0}^{n-1} f(\hat{X}_k)| \ge \epsilon) \le P(|\frac{1}{n}\sum_{k=0}^{n-1} f(\tilde{X}_k)| \ge \epsilon/2)$$

$$+P(|\sum_{k=0}^{n-1} f(\hat{X}_k) - \sum_{k=0}^{n-1} f(\tilde{X}_k)| \geq n\epsilon/2)$$

$$\leq P(|\frac{1}{n}\sum_{k=0}^{n-1} f(\tilde{X}_k)| \geq \epsilon/2) + P(C_f\tau \geq n\epsilon/2),$$

where the first term goes to zero due to the Theorem 1, while the second due to the inequality $P(\tau > n) \leq (1-\kappa)^n$.

8 CLT

Theorem 3: [stationary CLT] *Under (A1), for any f on the state space $\mathcal{S}$,*

$$\frac{1}{\sqrt{n}}\sum_{k=0}^{n-1}(f(X_k) - E_{inv}f(X_k)) \overset{P_{inv}}{\Longrightarrow} \sigma Z,$$

where $Z \sim \mathcal{N}(0,1)$ and

$$0 \leq \sigma^2 = E_{inv}(f(X_0) - E_{inv}f(X_0))^2$$

$$+2\sum_{k=1}^{\infty} E_{inv}(f(X_0) - E_{inv}f(X_0))(f(X_k) - E_{inv}f(X_k)).$$

(We consider 0 as a (degenerate) Gaussian r.v.)

Why such "unexpected" (if seen for the first time) σ^2?

$$E_{inv}|\frac{1}{\sqrt{n}}\sum_{k=0}^{n-1}(f(X_k) - E_{inv}f(X_k))|^2$$

$$= \frac{1}{n}E_{inv}\sum_{k=0}^{n-1}\sum_{j=0}^{n-1}(f(X_k) - E_{inv}f(X_k))(f(X_j) - E_{inv}f(X_j))$$

$$= \frac{1}{n}E_{inv}\sum_{k=0}^{n-1}(f(X_k) - E_{inv}f(X_k))^2$$

$$+\frac{2}{n}E_{inv}\sum_{k=0}^{n-1}\sum_{j=k+1}^{n-1}(f(X_k) - E_{inv}f(X_k))(f(X_j) - E_{inv}f(X_j)),$$

and the difference between the latter and σ^2 goes to zero. In our forthcoming notations $\eta_1 = \sum_{k=0}^{m-1} f(X_k)$, with $m \to \infty$, this may be expressed as

$$\mathrm{var}_{inv}(\eta_1)/m \sim \sigma^2.$$

We shall remember this. In the sequel, assume that

$$E_{inv} f(X_0) = 0;$$

otherwise, subtract.

Why at all there is a (weak) limit, and why Gaussian? Intuitively, there is a good reason for both assertions, if we compare the expression $\frac{1}{\sqrt{n}} \sum_{k=0}^{n-1} f(X_k)$ with another one, $\frac{1}{\sqrt{n}} \sum_{k=0}^{\gamma_n} f(X_k)$, with, say, $\gamma_1 := \inf(k > 0 : X_k = X_0)$ and by induction $\gamma_{n+1} := \inf(k > \gamma_n : X_k = X_0)$, which satisfies CLT due to standard IID arguments. However, such way is technically not so easy. We will prove the assertion using another method, historically the first one. Split the (growing as $n \to \infty$) interval $[0, n]$ by larger and smaller partitions, e.g., as follows: take $k := [\frac{n}{[n^{3/4}]}]$ (the total number of long "corridors" of equal length, which (the length) will be chosen in a minute: in any case, it will not exceed $n^{3/4}$ and will be equivalent to that function); $w := [n^{1/5}]$ (the length of short "windows" which separate all consequent corridors); now $m := [\frac{n}{k}] - w = [\frac{n}{k}] - [n^{1/5}]$ (the length of each corridor, except the last one which has the complementary length $n - k[\frac{n}{k}] \leq k$).

Notice that $k \sim n^{1/4}$ as $n \to \infty$ and that the total length of all windows is equivalent to $n^{9/20}$, which satisfies $n^{9/20} << n^{1/2}$; that $m \sim n^{1/4}$, and that the last corridor's length does not exceed k and, hence, asymptotically does not exceed $n^{1/4}$.

The idea is now to take into account only the corridors where we perform summation, while summation over windows is to be dropped, without big consequence for the assertion due to a small total length of windows. Now partial sums over different corridors are "nearly independent", with some clear sense derived from the exponential estimate of the Ergodic Theorem. Because of that, we can nearly repeat the calculus with characteristic functions for the IID case, with the remark that the asymptotical variance σ^2 is already evaluated above. The next pages show an "approximately rigorous" proof.

Denote all partial sums $\sum f(X_s)$ over first k corridors as η_j, $1 \leq j \leq k$. Notice that

$$\frac{1}{\sqrt{n}} (\sum_{s=1}^{n} f(X_s) - \sum_{j=1}^{k} \eta_j) \sim 0, \quad n \to \infty.$$

Hence, it remains to evaluate (for a NON-iid case)

$$E\exp(i\lambda\sum_{j=1}^{k}\eta_j), \quad n\to\infty.$$

We will do this by induction, using on its each step the exponential bound of the Ergodic Theorem.

Notice that

$$|E(\exp(i\lambda\eta_j)\mid\mathcal{F}^X_{(j-1)[n/k]})-E_{inv}\exp(i\lambda\eta_j)|\le C(1-\kappa)^{n^{1/5}}.$$

Hence,

$$E\exp(i\frac{\lambda}{n^{1/2}}\sum_{j=1}^{k}\eta_j)$$

$$=E\exp(i\frac{\lambda}{n^{1/2}}\sum_{j=1}^{k-1}\eta_j)E(\exp(i\frac{\lambda}{n^{1/2}}\eta_k)\mid\mathcal{F}^X_{(k-1)[n/k]})$$

$$=E\exp(i\frac{\lambda}{n^{1/2}}\sum_{j=1}^{k-1}\eta_j)\left(E_{inv}\exp(i\frac{\lambda}{n^{1/2}}\eta_k)+O((1-\kappa)^{n^{1/5}})\right)$$

$$=\ldots=\left(E_{inv}\exp(i\frac{\lambda}{n^{1/2}}\eta_1))\right)^k+O(k(1-\kappa)^{n^{1/5}}).$$

But it can be seen that (remind that $m\sim n/k$)

$$E_{inv}\exp(i\frac{\lambda}{n^{1/2}}\eta_1)=1+i\frac{\lambda}{n^{1/2}}\times 0-\frac{\lambda^2}{n}\,\frac{n}{k}\,\frac{1}{m}\,E_{inv}\eta_1^2+o(1/k)$$

$$=1-\frac{\lambda^2}{n}\,\frac{n}{k}\sigma^2+o(1/k). \qquad (12)$$

From this we get,

$$\left(E_{inv}\exp(i\frac{\lambda}{n^{1/2}}\eta_1))\right)^k\approx\left(1-\frac{\lambda^2\sigma^2}{2k}\right)^k\to\exp(-\lambda^2\sigma^2/2), \qquad (13)$$

as required.

To show (12), and, hence, (13) rigorously, we have to evaluate the third moment of η_1 with respect to the invariant measure, namely

$$E_{inv}\eta_1^3\equiv E\sum_{0\le k_1,k_2,k_3\le m-1}\xi_{k_1}\xi_{k_2}\xi_{k_3}, \qquad (14)$$

where we set $\xi_j := f(X_j) - E_{inv}f(X_j)$. We shall see that $E_{inv}\eta_1^3$ is *of the order* m, at most. Let us split the sum above into three terms,

$$E \sum_{0\le k_1,k_2,k_3\le m-1} \xi_{k_1}\xi_{k_2}\xi_{k_3} \equiv E \sum_{0\le k_1\le m-1} \xi_{k_1}^3 + E \sum_{0\le k_1,k_2\le m-1; k_1\ne k_2} \xi_{k_1}^2\xi_{k_2}$$

$$+E \sum_{0\le k_1,k_2,k_3\le m-1;\, k_1\ne k_2\ne k_3\ne k_1} \xi_{k_1}\xi_{k_2}\xi_{k_3} =: \Sigma^1 + \Sigma^2 + \Sigma^3.$$

For the first term, we have straight away,

$$|\Sigma^1| \le Cm. \tag{15}$$

Consider the second term,

$$\Sigma^2 = E \sum_{0\le k_1,k_2\le m-1; k_1<k_2} \xi_{k_1}^2\xi_{k_2} + E \sum_{0\le k_1,k_2\le m-1; k_1>k_2} \xi_{k_1}^2\xi_{k_2} =: \Sigma^{2a} + \Sigma^{2b}.$$

We estimate,

$$\Sigma^{2a} = E \sum_{0\le k_1,k_2\le m-1; k_1<k_2} \xi_{k_1}^2\xi_{k_2}$$

$$= \sum_{0\le k_1,k_2\le m-1; k_1<k_2} \sum_{j_1}\sum_{j_2} f^2(j_1)f(j_2)\pi_{j_1}p_{j_1,j_2}^{(k_2-k_1)}$$

$$= \sum_{0\le k_1\le m-1} \sum_{j_1}\sum_{j_2} f^2(j_1)f(j_2)\pi_{j_1} \sum_{k_2:k_2>k_1} (p_{j_1,j_2}^{(k_2-k_1)} - \pi_{j_2}).$$

Denote

$$A(m, j_1, j_2) := \sum_{k_2:k_2>k_1} (p_{j_1,j_2}^{(k_2-k_1)} - \pi_{j_2}).$$

Due to the Ergodic Theorem, all values $A(\cdot)$ are uniformly bounded. Hence,

$$\Sigma^{2a} = \sum_{0\le k_1\le m-1} \sum_{j_1}\sum_{j_2} f^2(j_1)f(j_2)\pi_{j_1}A(m, j_1, j_2)$$

is of the order m, at most.

Next, we have,

$$\Sigma^{2b} = E \sum_{0\le k_1,k_2\le m-1; k_1>k_2} \xi_{k_1}^2\xi_{k_2}$$

$$= \sum_{0\le k_1,k_2\le m-1; k_1>k_2} \sum_{j_1}\sum_{j_2} f^2(j_1)f(j_2)\pi_{j_1}p_{j_1,j_2}^{(k_1-k_2)}$$

$$= \sum_{0\le k_2\le m-1} \sum_{j_1}\sum_{j_2} f^2(j_1)f(j_2)\pi_{j_1} \sum_{k_1:k_1>k_2} (p_{j_1,j_2}^{(k_1-k_2)} - \pi_{j_2}).$$

$$B(m, j_1, j_2) := \sum_{k_1 : k_1 > k_2} (p_{j_1,j_2}^{(k_1 - k_2)} - \pi_{j_2}).$$

Due to the Ergodic Theorem, all values $B(\cdot)$ are uniformly bounded. Hence,

$$\Sigma^{2b} = \sum_{0 \le k_2 \le m-1} \sum_{j_1} \sum_{j_2} f^2(j_1) f(j_2) \pi_{j_1} \sum_{k_1 : k_1 > k_2} B(m, j_1, j_2)$$

is of the order m, at most. Hence, so is the whole term,

$$|\Sigma^2| \le Cm. \tag{16}$$

Finally, consider the <u>third term</u>,

$$\Sigma^3 \equiv 6\, E \sum_{0 \le k_1, k_2, k_3 \le m-1;\, k_1 < k_2 < k_3} \xi_{k_1} \xi_{k_2} \xi_{k_3}$$

$$\equiv 6 \sum_{k_2=1}^{m-2} \sum_{k_1 : k_1 < k_2} \sum_{k_3 : k_3 > k_2} \sum_{j_1} \sum_{j_2} \sum_{j_3} f(j_1) f(j_2) f(j_3) \pi_{j_1} p_{j_1,j_2}^{(k_2 - k_1)} p_{j_2,j_3}^{(k_3 - k_2)}$$

$$\equiv 6 \sum_{k_2=1}^{m-2} \sum_{k_1 : k_1 < k_2} \sum_{k_3 : k_3 > k_2} \sum_{j_1} \sum_{j_2} \sum_{j_3} f(j_1) f(j_2) f(j_3) \pi_{j_1} p_{j_1,j_2}^{(k_2 - k_1)} (p_{j_2,j_3}^{(k_3 - k_2)} - \pi_{j_3})$$

$$\equiv 6 \sum_{k_2=1}^{m-2} \sum_{k_1 : k_1 < k_2} \sum_{j_1} \sum_{j_2} \sum_{j_3} f(j_1) f(j_2) \pi_{j_1} p_{j_1,j_2}^{(k_2 - k_1)} \sum_{k_3 : k_3 > k_2} f(j_3) (p_{j_2,j_3}^{(k_3 - k_2)} - \pi_{j_3}).$$

Denote

$$A(m, j_2) := \sum_{j_3} \sum_{k_3 : k_3 > k_2} f(j_3) (p_{j_2,j_3}^{(k_3 - k_2)} - \pi_{j_3}).$$

Due to the Ergodic Theorem estimate

$$|p_{j_2,j_3}^{(k_3 - k_2)} - \pi_{j_3}| \le (1 - \kappa)^{(k_3 - k_2)},$$

and since f is bounded, the values $A(m, j_2)$ are all uniformly bounded by $C_f \times N \times \kappa^{-1}$. Rewrite Σ^3,

$$\Sigma^3 \equiv 6 \sum_{k_2=1}^{m-2} \sum_{k_1 : k_1 < k_2} \sum_{j_1} \sum_{j_2} f(j_1) f(j_2) \pi_{j_1} p_{j_1,j_2}^{(k_2 - k_1)} A(m, j_2)$$

$$\equiv 6 \sum_{k_2=1}^{m-2} \sum_{j_2} \sum_{j_1} \sum_{k_1 : k_1 < k_2} f(j_1) f(j_2) \pi_{j_1} (p_{j_1,j_2}^{(k_2 - k_1)} - \pi_{j_2}) A(m, j_2),$$

the latter because

$$\sum_{k_2=1}^{m-2}\sum_{j_2}\sum_{j_1}\sum_{k_1:k_1<k_2} f(j_1)f(j_2)\pi_{j_1}\pi_{j_2}A(m,j_2)$$

$$=\sum_{k_2=1}^{m-2}\sum_{j_2}\sum_{k_1:k_1<k_2} f(j_2)\pi_{j_2}A(m,j_2)\sum_{j_1} f(j_1)\pi_{j_1}=0.$$

Hence, we continue,

$$\Sigma^3 \equiv 6\sum_{k_2=1}^{m-2}\sum_{j_2}\sum_{j_1} f(j_1)f(j_2)\pi_{j_1}A(m,j_2)\sum_{k_1:k_1<k_2}(p_{j_1,j_2}^{(k_2-k_1)}-\pi_{j_2}).$$

Denote

$$B(m,j_1):=\sum_{k_1:k_1<k_2}(p_{j_1,j_2}^{(k_2-k_1)}-\pi_{j_2}).$$

Due to the Ergodic Theorem exponential bound, the values $B(n,j_1)$ are all uniformly bounded. Hence, it remains,

$$\Sigma^3 \equiv 6\sum_{k_2=1}^{m-2}\sum_{j_2}\sum_{j_1} f(j_1)f(j_2)\pi_{j_1}A(m,j_2)B(m,j_1),$$

which clearly is of order m. So,

$$|\Sigma^3|\le Cm. \tag{17}$$

Combining all three bounds for $\Sigma^{1,2,3}$, we have,

$$|E_{inv}\eta_1^3|\le Cm. \tag{18}$$

This shows that in (12) we have an even better estimate,

$$n^{-3/2}|E_{inv}\eta_1^3|\le C\frac{m}{n^{3/2}}\sim C\frac{n^{1/4}}{n^{3/2}}=C\frac{1}{n^{5/4}}\sim C\frac{1}{k^5}=o(\frac{1}{k}).$$

The last concern could be a possibly non-complete last corridor, with maybe a smaller number of terms in η_{k+1}, say, $m'\le m$. In this case we just notice that this relates to only one very last multiple $E_\pi\exp(i(\lambda/\sqrt{n})\sum_{j=1}^{m'}f(X_j))$

which tends to 1 as $n \to \infty$. Indeed, the third moment of the sum above admits a similar bound,

$$|E_\pi(\sum_{j=1}^{m'} f(X_j))^3| \le Cm' \le Cm,$$

and for the variance, $\sigma^2_{m'} := E_\pi(\sum_{j=1}^{m'} f(X_j))^2$,

$$|E_\pi(\sum_{j=1}^{m'} f(X_j))^2| \le Cm' \le Cm.$$

So, we have with some $\theta \in [0,1]$,

$$E_\pi \exp(i(\lambda/\sqrt{n})\sum_{j=1}^{m'} f(X_j)) = 1 - \frac{\lambda^2\sigma^2_{m'}}{n} + O(\frac{\lambda^3\theta^3 m'}{n^{3/2}}) \to 1, \;\; n \to \infty.$$

Thus, this last corridor cannot spoil the overall estimate, and, hence, the conclusion (13) is proved.

Theorem 3a: [non-stationary CLT] *Under (A1), for any f on the state space $\mathcal{S}$,*

$$\frac{1}{\sqrt{n}}\sum_{k=0}^{n-1}(f(X_k) - E_{inv}f(X_k)) \overset{P}{\Longrightarrow} \sigma Z,$$

where $Z \sim \mathcal{N}(0,1)$ and

$$0 \le \sigma^2 = E_{inv}(f(X_0) - E_{inv}f(X_0))^2$$
$$+2\sum_{k=1}^{\infty} E_{inv}(f(X_0) - E_{inv}f(X_0))(f(X_k) - E_{inv}f(X_k)).$$

Notice that here the measure P relates to an arbitrary initial distribution μ_0. The proof can be reduced to the Theorem 3 by the same trick as we used for the LLN. We have, for $\hat{X}$ which is equivalent to X,

$$\frac{1}{\sqrt{n}}\sum_{k=0}^{n-1}(f(X_k) - E_\pi f(\hat{X}_k)) \equiv \frac{1}{\sqrt{n}}\sum_{k=0}^{n-1}(f(\hat{X}_k) - E_\pi f(X_k))$$

$$= \frac{1}{\sqrt{n}}\sum_{k=0}^{n-1}(f(\tilde{X}_k) - E_\pi f(\tilde{X}_k)) + \frac{1}{\sqrt{n}}\sum_{k=0}^{n-1}(f(\hat{X}_k) - f(\tilde{X}_k)).$$

Here the first term converges weakly to σZ,

$$\frac{1}{\sqrt{n}}\sum_{k=0}^{n-1}(f(\tilde{X}_k) - E_\pi f(\tilde{X}_k)) \overset{P}{\Longrightarrow} \sigma Z,$$

while the second term equals

$$\frac{1}{\sqrt{n}}\sum_{k=0}^{n-1}(f(\hat{X}_k) - f(\tilde{X}_k)) \equiv \frac{1}{\sqrt{n}}\sum_{k=0}^{n-1}(f(\hat{X}_k) - f(\tilde{X}_k))1(k < \tau).$$

Therefore,

$$E_{\mu_0}|\frac{1}{\sqrt{n}}\sum_{k=0}^{n-1}(f(\hat{X}_k) - f(\tilde{X}_k))| \le C_f\frac{E_{\mu_0}\tau}{\sqrt{n}} \to 0, \qquad n \to \infty.$$

So, the second term converges *in probability to zero.* This implies the claim of the Theorem 3a.

9 LD's

This is another side of "convergence rate" problem: instead of different normalisation coefficients, it is also reasonable to ask, what is the rate of convergence for probability $P(|\frac{1}{n}\sum_{k=0}^{n-1} f(X_k)| \ge \epsilon)$, as $n \to \infty$. In "good cases" (including the case of finite state space under Ergodic Theorem hypothesis) this rate turns out to be exponential. To start with, let us recollect how this may be achieved in the IID case. Hence, for a minute assume that X_k's are (bounded) IID with $Ef(X_1) = 0$ (otherwise subtract). We are going to apply an exponential version of the Bienaimé–Chebyshev inequality. For any $\beta > 0$,

$$P(|\frac{1}{n}\sum_{k=0}^{n-1} f(X_k)| \ge \epsilon) = P(|\beta\sum_{k=0}^{n-1} f(X_k)| \ge \beta n\epsilon)$$

$$\le \exp(-\beta n\epsilon)\left(E\exp(\beta\sum_{k=0}^{n-1} f(X_k)) + E\exp(-\beta\sum_{k=0}^{n-1} f(X_k))\right)$$

$$= \exp(-\beta n\epsilon)\left((E\prod_{k=0}^{n-1}\exp(\beta f(X_k)) + (E\prod_{k=0}^{n-1}\exp(-\beta f(X_k))\right)$$

$$= \exp(-\beta n\epsilon)\left((E\exp(\beta f(X_1))^n + (E\exp(-\beta f(X_1))^n\right)$$

$$= \exp(-n(\beta\epsilon - \ln E\exp(\beta f(X_1)))$$

$$+ \exp(-n(\beta\epsilon - \ln E\exp(-\beta f(X_1))).$$

For $\beta > 0$ small the term $\beta\epsilon$ is linear in β, while $\ln E\exp(\pm\beta f(X_1))$ only quadratic (because $Ef(X_1) = 0$). Hence, for $\beta > 0$ small the difference $\beta\epsilon - \ln E\exp(\pm\beta f(X_1))$ is positive, so with any such β we get an exponential bound.

In fact, more can be said about the exponential bounds above. It is possible to show that the best (i.e. the largest possible) $L(\epsilon)$ in the exponential bound

$$\lim_{n\to\infty}\frac{1}{n}\ln P(|\frac{1}{n}\sum_{k=0}^{n-1} f(X_k)| \geq \epsilon) \leq -L(\epsilon),$$

is given by

$$L(\epsilon) = \sup_{\beta\in R}(\beta\epsilon - \varphi(\beta)) \wedge \sup_{\beta\in R}(\beta\epsilon - \varphi(-\beta)) \equiv \sup_{\beta\in R}(\beta\epsilon - \max(\varphi(\beta), \varphi(-\beta))),$$

with

$$\ln E\exp(\beta f(X_1)) =: \varphi(\beta).$$

Now, how this may be extended to MP's? The following way may be tried.
1° Show $\forall\ \beta \in R$,

$$\exists \quad H(\beta) := \lim_{n\to\infty}\frac{1}{n}\ln\ E\exp(\beta\sum_{k=0}^{n-1} f(X_k)) < \infty.$$

2° Show

$$\exists \quad H'(0) = 0.$$

If this is doable, we may repeat the main hint for the IID case (about $\epsilon\beta - H(\beta) > 0$ and $\epsilon\beta - H(-\beta) > 0$) and find $\beta > 0$ which would provide an exponential bound.

Consider the operator (matrix) depending on β, defined by its action on any function h on the state space:

$$T^\beta h(j) := E_j h(X_1)\exp(\beta f(j)) \equiv \exp(\beta f(j))\sum_{k\in\mathcal{S}} p_{jk}h(k).$$

In the other words, T^β is simply the transition matrix T multiplied by the factor $\exp(\beta f(x)) > 0$. Clearly, this matrix function is analytic with respect to the variable β. We will use the following famous theorem about positive matrices.

Frobenius Theorem. Under the Ergodic Theorem assumptions, the matrix T^β has an isolated eigenvalue λ_0 (called spectral radius) which is real and positive, all other eigenvalues by modulus being strictly less that λ_0, and the eigenvector, say, e_0, which corresponds to this eigenvalue, is positive. The function $\lambda_0(\beta)$ is analytic.

Corollary. The limit in 1° does exist and equals

$$H(\beta) = \ln \lambda_0(\beta).$$

Proof. (The lower bound follows similarly.)

$$\begin{aligned}
&\overline{\lim}_{n\to\infty} \ln\, E_x \exp(\beta \sum_{k=0}^{n-1} f(X_k)) \\
\le\, &\overline{\lim}_{n\to\infty} \ln\, C E_x e_0(X_n) \exp(\beta \sum_{k=0}^{n-1} f(X_k)) \\
=\, &\lim \frac{1}{n} \ln C + \lim \frac{1}{n} \ln (T^\beta)^n e_0(x) \\
=\, &\lim \frac{1}{n} \ln (\lambda_0(\beta))^n e_0(x) = \ln \lambda_0(\beta).
\end{aligned}$$

Next, consider $H_n(\beta, x) = n^{-1} \ln\, E_x \exp(\beta \sum_{k=0}^{n-1} f(X_k))$, and

$$H_n'(\beta, x) = n^{-1} \frac{E_x(\sum_{k=0}^{n-1} f(X_k))\, \exp(\beta \sum_{k=0}^{n-1} f(X_k))}{E_x \exp(\beta \sum_{k=0}^{n-1} f(X_k))}$$

Here $H' = \partial H/\partial\beta$. At $\beta = 0$ we get, due to the LLN,

$$H_n'(0, x) = \frac{E_x \sum_{k=0}^{n-1} f(X_k)}{n} \to 0, \quad n \to \infty.$$

All functions $H_n(\cdot, x)$ and $H(\cdot, x)$ are *convex* in β, in which case *from convergence of functions convergence (locally uniform) of their derivatives follows,* see, e.g., [T. Rockafellar, Convex Analysis, Theorem 25.7]. So $H'(0) = 0$. This shows 2°. So, 1° and 2° imply the existence of $L(\epsilon) > 0$ such that

$$\lim_{n\to\infty} \frac{1}{n} \ln P(|\frac{1}{n} \sum_{k=0}^{n-1} f(X_k)| \ge \epsilon) \le -L(\epsilon).$$

It is possible to show that the best $L(\epsilon)$ here is given by

$$L(\epsilon) = \sup_{\beta\in R}(\beta\epsilon - H(\beta)) \wedge \sup_{\beta\in R}(\beta\epsilon - H(-\beta)),$$

similarly to the IID case.

10 Generator and Poisson equation

Remind that for any MC, its *generator* is an operator L,

$$Lh(x) = E_x h(X_1) - h(x).$$

Dynkin's formula, to be proved by induction

$$(D1) \qquad \boxed{Eh(X_n) = h(x) + \sum_{k=0}^{n-1} ELh(X_k).}$$

The base is clear, and if the formula is true for some n, then

$$\begin{aligned} &Eh(X_{n+1}) - h(x) - \sum_{k=0}^{n} ELh(X_k) \; (\pm Eh(X_n)) \\ &= E\left(E(h(X_{n+1}) - ELh(X_n) - Eh(X_n)) \mid X_n\right) = 0. \end{aligned}$$

In the ergodic case, as we have seen the invariant measure is the only stationary probability for the MC. The problem which interests us is whether those stationary probabilities are smooth with respect to some parameter, provided that transition probabilities depend on this parameter in a regular way. The answer is positive: invariant probabilities have so many derivatives as the transition ones. Briefly, this follows from the perturbation theory of operators (matrices) with a spectral gap (i.e. the property that the spectral radius is bounded away from the rest of the spectrum by modulus), because the invariant measure is the eigenvector corresponding to the spectral radius. See, e.g., [Kato] or other monographs on Functional Analysis.

Poisson equation is the equation of the sort,

$$Lu(x) = -f(x), \qquad x \in \mathcal{S},$$

for some given function f. (Here for finite state space, any "function" is just a vector on $\mathcal{S}$.)

Notice here that a PE may make sense not for an arbitrary function f, if we wish the equation to be satisfied for all values of x; or else, we may be asked to solve the equation for all values of x but one or several given values, say, x_0, where we should impose some "boundary condition" on the unknown function u. We will consider both options.

Let $A \subset \mathcal{S}$ which is strictly less than $\mathcal{S}$.

$$(PE1) \qquad Lu(x) = -f(x), \qquad x \in \mathcal{S} \setminus A, \quad u|_A = g.$$

Let $\tau := \inf(n \geq 0 : \, X_n \in A)$, and

$$v(x) := E_x \left(\sum_{k=0}^{\tau-1} f(X_k) + g(X_\tau) \right).$$

Theorem 1. v is a unique solution of the (PE1) above.

Proof-1. Clearly, if $x \in A$, then $\tau = 0$, and $v(x) = g(x)$, as required. To verify the equation, we will need a lemma.

Lemma. $E_x \tau < \infty$. Moreover, there exists $\alpha > 0$ such that $E_x \exp(\alpha \tau) < \infty$.

On each step, probability to get to A is positive, and there is a minimal probability for that over $\mathcal{S} \setminus A$, say, p_A. Due to the Markov property, probability not to get to A during n steps is at most $(1 - p_A)^n$. This implies both assertions, the second one with any $\alpha < -\ln(1 - p_A)$. (And if, by some chance, $p_A = 1$, under our ergodic assumptions this would mean that $A = \mathcal{S}$.)

From this Lemma it follows that the expression for $v(\cdot)$ is finite uniformly on $\mathcal{S}$. Now we can continue the proof of the Theorem.

Proof-2. Let $x \notin A$. Then $\tau > 0$. We have, due to the Markov property,

$$\begin{aligned} v(x) &= f(x) + \sum_y E_x 1(X_1 = y) E_y \left(\sum_{k=0}^{\tau-1} f(X_k) + g(X_\tau) \right) \\ &= f(x) + \sum_y p_{xy} v(y) = f(x) + E_x v(X_1). \end{aligned}$$

From this, it follows clearly the statement,

$$Lv(x) = E_x v(X_1) - v(x) = -f(x).$$

Proof-3. Why solution is unique? Consider two solutions, then their difference, say, $w(\cdot) = v^1(\cdot) - v^2(\cdot)$ satisfies the (PE1) with $f \equiv 0$ and $g \equiv 0$. So what? The matter is that any solution of the *Laplace equation* (i.e. with $f \equiv 0$) satisfies

Maximum principle*: its maximum value (and minimum, too) is attained at the boundary, i.e. at A.*

The latter is simply because, due to the equation at any $x \notin A$, the value $v(x)$ equals the average of the rest. Hence, if at some particular $x \notin A$ the maximal value of v is attained, no other value may be less (otherwise that average must be less, too). The same about mininum.

This suffices for uniqueness, because $\max w|_A = \min w|_A = 0$.

$$(PE2) \qquad\qquad Lu(x) = -f(x), \qquad \forall x \in \mathcal{S}.$$

Assume f "centered": $\sum f(j)\pi_j = 0$. Let

$$(PE2 - sol) \qquad\qquad \boxed{v(x) := \sum_{k=0}^{\infty} E_x f(X_k).}$$

Theorem 2. v is a unique solution of the (PE2) above which is centered itself. For any constant c, $v+c$ is also a solution of (PE2). (But not centered, of course.)

1. Firstly, why the series converges? Because $E_{inv} f(X_k) = 0$ and $|E_x f(X_k) - E_{inv} f(X_k)| \le q^k$, by the Ergodic Theorem.

2. Why it satisfy the equation? Due to the same calculus as for the (PE1). Namely, "due to the Markov property,

$$v(x) = f(x) + \sum_y E_x 1(X_1 = y) E_y \sum_{k=0}^{\infty} f(X_k)$$

$$= f(x) + \sum_y p_{xy} v(y) = f(x) + E_x v(X_1).$$

From this, it follows clearly the statement of the Theorem,

$$Lv(x) = E_x v(X_1) - v(x) = -f(x).$$

3. Why is v centered? We check,

$$\sum_x \pi_x v(x) = \sum_x \pi_x \sum_{k=0}^{\infty} E_x f(X_k)$$

$$= \sum_x \pi_x \sum_{k=0}^{\infty} \sum_y p_{xy}^{(k)} f(y) = \sum_{k=0}^{\infty} \sum_y \sum_x \pi_x p_{xy}^{(k)} f(y)$$

$$= \sum_{k=0}^{\infty} \sum_y \pi_y f(y) = 0.$$

Of course, in the last line we have used centering condition on f.

4. How do we verify uniqueness (with centering condition) by Maximum Principle?

Suppose there are two such solutions. Their difference then is also centered and satisfies the Laplace equation $Lw = 0$ everywhere. Suppose w has some positive values, then at some point it attains maximum. Then, all values of w must coincide with that maximum. But in this case centering condition is impossible. Hence, $\max w \le 0$. By similar reasons, also $\min w \ge 0$. Uniqueness is proved.

Now we can explain why for the equation without boundary conditions (PE2) the centering condition is necessary. Suppose it is not satisfied, i.e.

$$a := E_{inv} f(X_1) \neq 0.$$

But due to the Ergodic Theorem,

$$Ef(X_k) \to a, \qquad k \to \infty.$$

Hence, the series in the definition of $v(x) = \sum_{k=0}^{\infty} E_x f(X_k)$ clearly diverges.

All theories above – i.e., LLN, CLT, LD's and Poisson equations of both types, – admit the following generalization. Instead of $\kappa > 0$, we may assume that *there exists $k_0 \ge 1$ such that the transition matrix $T^{k_0} \equiv T^{(k_0)}$ satisfies the same assumption,*

$$\kappa^{(k_0)} := \inf_{x,x'} \sum_y p^{(k_0)}_{x,y} \wedge p^{(k_0)}_{x',y} > 0.$$

E.g., the assertion of the Ergodic Theorem then remains valid, with replacement of the convergence rate from $(1-\kappa)^n$ by $(1-\kappa)^{[n/k_0]}$. Frobenius Theorem also remains valid. In general state spaces, a natural analogue of the condition above is known as Dobrushin's one.

Parameters

We will be now interested in the following problem: suppose all transition probabilities depend on some parameter, $p_{xy}(\theta)$. The range of questions arise: let $p_{xy}(\theta)$ possess some regularity with respect to θ (continuity, Hölder continuity, Lipschitz, C^1, C^N, C^∞, analytic dependence). Will this property or some its version hold true for invariant probabilities $\pi_j(\theta)$, or for expressions like $E_{inv} g(X_0)$. Some of those limiting properties can be derived straight from the Ergodic Theorem; some others follow from advanced perturbation methods [Kato] combined with Frobenius Theorem. We will study some simple method(s) which later in this course will be extended to certain non-finite and non-compact state spaces.

Theorem 1. *Suppose all transition probabilities $p_{xy}(\cdot) \in C$ in θ, and $\inf_\theta \kappa(\theta) > 0$. Then all $\pi_y(\theta)$ are continuous in θ.*

Indeed, from Chapman–Kolmogorov's equations

$$p_{xy}^{(k)}(\theta) = \sum_z p_{xz}^{(k-1)}(\theta) p_{zy}(\theta),$$

$p_{xy}(\cdot) \in C \Longrightarrow p_{xy}^{(k)}(\cdot) \in C$, $\forall k$. Continuous functions $p_{xy}^{(k)}(\cdot)$ converge *uniformly*, so the limit is continuous in θ: $\forall \epsilon > 0$,

$$\begin{aligned} |\pi_y(\theta) - \pi_y(\theta')| &\le |\pi_y(\theta) - p_{xy}^{(k)}(\theta)| \\ &+ |p_{xy}^{(k)}(\theta) - p_{xy}^{(k)}(\theta')| + |p_{xy}^{(k)}(\theta') - \pi_y(\theta')|, \end{aligned} \tag{19}$$

where 1st + 3rd terms $\le \epsilon/2$ for k large enough, and 2nd $\le \epsilon/2$ with this fixed k as $\theta - \theta'$ is small, due to $p_{xy}^{(k)}(\cdot) \in C$.

Clearly, we have also a uniform version of the Theorem 1, i.e., if all (there is only a finite number of them!) $p_{xy}(\cdot)$ are *uniformly* continuous, then the limiting probabilities are also uniformly continuous. The proof follows from the same calculus as above, which suits both continuity at just a single point as well as a uniform one.

How about Lipschitz or Hölder continuity? There is a "limited" result in this direction. It shows that to make $\pi \in C^1$ and further, we apparently will need some new methods.

Theorem 2. *Suppose all transition probabilities $p_{xy}(\cdot) \in H^\alpha$ with some $0 < \alpha \le 1$, in θ. Then all $\pi_y(\theta) \in H^{\alpha'}$ in θ, for every $\alpha' < \alpha$.*

In the other words, this *method* does not provide *the same* continuity for the limiting probabilities as for transition ones.

Proof. Assuming for some $0 < \alpha \le 1$,

$$\sup_{xy} |p_{xy}(\theta) - p_{xy}(\theta')| \le L|\theta - \theta'|^\alpha,$$

let $\epsilon = |\theta - \theta'|^\alpha$. We estimate, from (19) and from the exponential bound of the Ergodic Theorem, with $q = 1 - \kappa$, and taking $k = [\ln(\epsilon/4)/\ln q)] + 1$,

$$|\pi_y(\theta) - p_{xy}^{(k)}(\theta)| + |p_{xy}^{(k)}(\theta') - \pi_y(\theta')| \le 2q^k \le \frac{\epsilon}{2},$$

and from Chapman–Kolmogorov (L is new),

$$|p^{(k)}_{xy}(\theta) - p^{(k)}_{xy}(\theta')| \le k\,L\,|\theta-\theta'|^{\alpha} \le C\epsilon\frac{\ln(\epsilon/4)}{\ln q}. \qquad (20)$$

Since $\epsilon^{\frac{\ln(\epsilon/4)}{\ln q}} << \epsilon^{1-\delta}$ for every $\delta > 0$, when $\epsilon \to 0$, then we obtain, $|\pi_y(\theta) - \pi_y(\theta')| \le C|\theta-\theta'|^{\alpha'}$, with any $\alpha' < \alpha$.

Let us show (20) by induction, i.e.,

$$|p^{(k)}_{xy}(\theta) - p^{(k)}_{xy}(\theta')| \le k\,L\,|\theta-\theta'|^{\alpha}. \qquad (21)$$

For $k = 1$ this is an assumption. Assume the inequality (21) holds true for some k. We get from Chapman–Kolmogorov,

$$|p^{(k+1)}_{xy}(\theta) - p^{(k+1)}_{xy}(\theta')|$$

$$\le \sum_z |p^{(k)}_{xz}(\theta) - p^{(k)}_{zy}(\theta')| p_{zy}(\theta) + \sum_z p^{(k)}_{xz}(\theta')|p_{zy}(\theta) - p_{zy}(\theta')|$$

$$\le NkL|\theta-\theta'|^{\alpha} + L|\theta-\theta'|^{\alpha} \le (k+1)\,NL\,|\theta-\theta'|^{\alpha},$$

as required.

11 Heat equations

Poisson equations studied above were "elliptic" ones. Consider another type of equation, of "parabolic" or "heat" type,

$$\text{(HE1)} \qquad v(n+1,x) - v(n,x) - Lv(n,x) = +f(n,x), \quad \forall\, x \in \mathcal{S},$$

that is, without boundary conditions. In this case, we have to impose some *initial data*, say,

$$\text{(HE2)} \qquad v(0,x) = v_0(x), \quad \forall\, x \in \mathcal{S}.$$

Now we may speak of Cauchy problem for the equation (HE). *By iterations, solution of this problem, of course, exists and is unique.* As we shall see now, there is another Dynkin's formula for solving such equations, which is a useful *representation.*

Theorem 3. *A unique solution of the equation (HE) is given by the formula,*

$$\text{(HE-sol)} \qquad v(n,x) = E_x\left(\sum_{k=1}^{n} f(k-1, X_{n-k}) + v_0(X_n)\right).$$

Proof. Initial data is check easily: at $n = 0$, we get from the (HE-sol), $v(0,x) = 0 + v_0(x) = v_0(x)$, as required.

To verify the equation, for $n \geq 0$ we compute,

$$\begin{aligned} Lv(n,x) = E_x\left(\sum_{k=1}^{n} f(k-1, X_{n+1-k})\right) + E_x v_0(X_{n+1}) \\ -E_x\left(\sum_{k=1}^{n} f(k-1, X_{n-k})\right) - E_x v_0(X_n), \end{aligned}$$

and

$$\begin{aligned} v(n+1,x) - v(n,x) &= E_x\left(\sum_{k=1}^{n+1} f(k-1, X_{n+1-k}) + v_0(X_{n+1})\right) \\ &\quad - E_x\left(\sum_{k=1}^{n} f(k-1, X_{n-k}) + v_0(X_n)\right) \\ &= Lv(n,x) + f(n,x). \end{aligned}$$

Uniqueness follows simply by induction: clearly, if there are two solutions, v^1 and v^2, then their difference w also satisfies a similar equation *with zero initial data and zero right hand side* f, that is,

$$w(n+1) - w(n,x) - Lw(n,x) = 0, \quad w(0,x) = 0.$$

By induction (nothing to solve here!), we get

$$w(n,x) = 0, \quad \forall\, n \geq 0.$$

Another proof of the representation (HE-sol) can be also performed by induction.

We can formulate the hint from this proof as a form of Maximum Principle for heat equations:

Maximum Principle. If function v satisfies (HE) with $f \equiv 0$, then for any $n \geq 0$,

$$\boxed{\sup_x v(n,x) \leq \sup_x v_0(x).}$$

Similarly, of course,

$$\inf_x v(n,x) \geq \inf_x v_0(x).$$

Both inequalities follow straight away from the representation (HE-sol).

Derivative $\partial_\theta p_{xy}^{(k)}(\theta)$

For the rest of this section assume $\exists\ \partial_\theta p_{xy}(\theta),\ \forall\ x,y$. From Chapman–Kolmogorov,

$$p_{xy}^{(k+1)}(\theta) = \sum_z p_{xz}(\theta) p_{zy}^{(k)}(\theta),$$

hence, differentiating wrt θ, we find

$$\partial_\theta p_{xy}^{(k+1)}(\theta) = \sum_z \partial_\theta \left(p_{xz}(\theta) p_{zy}^{(k)}(\theta)\right)$$
$$= \sum_z p_{zy}^{(k)}(\theta) \partial_\theta \left(p_{xz}(\theta)\right) + \sum_z p_{xz}(\theta) \partial_\theta p_{zy}^{(k)}(\theta).$$

Denote

$$q_{zy}(\theta) := \partial_\theta p_{zy}(\theta), \quad q_{zy}^{(k)}(\theta) := \partial_\theta p_{zy}^{(k)}(\theta).$$

Subtracting $\partial_\theta p_{xy}^{(k)}(\theta)$ and dropping θ, we get,

$$q_{xy}^{(k+1)} - q_{xy}^{(k)} = \sum_z p_{zy}^{(k)} q_{xz} + \sum_z p_{xz} q_{zy}^{(k)} - q_{xy}^{(k)}. \tag{22}$$

Notice that

$$\sum_z p_{xz} q_{zy}^{(k)} = E_x q_{X_1 y}^{(k)}, \quad \Longrightarrow \quad \sum_z p_{xz} q_{zy}^{(k)} - q_{xy}^{(k)} = L q_{\cdot y}^{(k)}(x).$$

So, if we denote $f^1(k,x,y) := \sum_z p_{zy}^{(k)} q_{xz}$ and $v(k,x) := q_{xy}^{(k)}$, then the equation (22) with y fixed may be rewritten as a *heat equation*,

$$v(k+1,x) - v(k,x) = Lv(k,x) + f^1(k,x), \tag{23}$$

initial data being just $v_0(x) = \partial_\theta p_{xy}^{(0)} = \partial_\theta 1(x=y) = 0$.

Of course, we could have also started with $k=1$, where the initial data would have been $q_{xy} = \partial_\theta p_{xy}$. This would not change our further analysis, because if we started from $k = 0$ as admitted in (23), the next value would have been exactly

$$q^{(1)}_{xy} = q^{(0)}_{xy} + \sum_z p^{(0)}_{zy} q_{xz} + \sum_z p_{xz} q^{(0)}_{zy} - q^{(0)}_{xy} = q_{xy},$$

as required; so we could have continued by induction with the same results.

Solution of the equation (23) (reminder:)

$$v(k+1, x) - v(k, x) = Lv(k, x) + f^1(k, x), \quad v_0(x) = 0,$$

is given by Dynkin's formula–2:

$$\boxed{\boxed{q^{(n)}_{xy} = v(n, x) = E_x \textstyle\sum_{k=1}^n f^1(k-1, X_{n-k}).}}$$

where $f^1(k, x) = \sum_z p^{(k)}_{zy} q_{xz}$.

Btw, equivalently,

$$\boxed{q^{(n)}_{xy} = v(n, x) = E_x \textstyle\sum_{k=1}^n f^1(n-k, X_{k-1}).}$$

Which expression is better? (We shall see soon that the first one.)

So what? Why do we care of the representation at all? Of course, we might have expected something of the sort, simply because we can differentiate Chapman–Kolmogorov's equations in the right hand side...

We will now use the new representation to make a guess, about how could be looking a formula for the *limiting* object, $\partial_\theta \pi_y$: namely, let us try

$$\boxed{q^{(\infty)}_y := \sum_{k=1}^{\infty} E_{inv} f^1(k-1, X_0).}$$

Why X_0? Just under the invariant measure there is no difference. It turns out that the other representation formula is not easily interpreted with ∞ in place of n. The first was a better one.

(Remind that there was y in f^1 which we dropped for a while.) Now our plan is as follows. 1. Show that the series converges. 2. Show it *is* a derivative of $\pi_y(\theta)$ wrt θ. 3. In fact, simultaneously with part 2, we will see that the derivative of $p_{xy}^{(k)}$ tends to $q_y^{(\infty)}$ geometrically fast. Start with the first part of the plan.

First of all, notice that there exists a limit

$$E_{inv} f^1(k-1, X_0, y) \to 0, \quad k \to \infty.$$

Why? Indeed, $f^1(k,x) = \sum_z p_{zy}^{(k)} q_{xz}$ (y being still fixed), so,

$$f^1(k,r) = \sum_z p_{zy}^{(k)} \partial_\theta p_{rz}.$$

Here if we let $k \to \infty$, then the r.h.s. will tend to

$$\sum_z \pi_y \partial_\theta p_{rz} = \pi_y \sum_z \partial_\theta p_{rz}$$
$$= \pi_y \, \partial_\theta \sum_z p_{rz} = \pi_y \, (\partial_\theta \, 1) = 0.$$

Moreover, under our Ergodic Theorem assumptions, this convergence is geometric, since $|\partial_\theta p_{xy}| \le C$, and

$$|p_{zy}^{(k)} - \pi_y| \le q^k, \quad \Longrightarrow \quad |f^1(k,r)| \le C q^k.$$

Hence, the series in the r.h.s. of

$$q_y^{(\infty)} := \sum_{k=1}^{\infty} E_{inv} f^1(k-1, X_0, y)$$

converges geometrically fast. This completes the first part of our plan.

Now, let us do the *third part* (before the second one): show that

$$q_{xy}^{(n)} \to q_y^{(\infty)}, \quad n \to \infty,$$

hopefully also with some geometric rate, or, at least, uniformly. This is where we will need in full our representation for $q_{xy}^{(n)}$.

Remind that

$$q_{xy}^{(n)} = E_x \sum_{k=1}^{n} f^1(k-1, X_{n-k}) = \sum_{k=1}^{n} E_x f^1(k-1, X_{n-k}),$$

and

$$q_y^{(\infty)} := \sum_{k=1}^{\infty} E_{inv} f^1(k-1, X_0, y).$$

We have just explained that the general term of the series for $q^{(\infty)}$ goes to zero geometrically fast, say, $\le C\, q^n$, with some $q < 1$. In particular, we may use the bound,

$$\sum_{k=n/2}^{\infty} |E_{inv} f^1(k-1, X_0, y)| \le C\, q^{n/2}.$$

The same holds true for the series for $q_{xy}^{(n)}$,

$$\sum_{k=n/2}^{n} |E_x f^1(k-1, X_{n-k}, y)| \le C\, q^{n/2}.$$

Hence, it remains to consider the difference,

$$\sum_{k=1}^{n/2} E_x f^1(k-1, X_{n-k}) - \sum_{k=1}^{n/2} E_{inv} f^1(k-1, X_0).$$

But f^1 being bounded, again by the Ergodic Theorem,

$$\sum_{k=1}^{n/2} C q^{(n-k)} \le C\, q^{n/2}.$$

Altogether, we have, with some new C and q,

$$\boxed{|q_{xy}^{(n)} - q_y^{(\infty)}| \le C q^n.} \tag{24}$$

Btw, notice that since $q_{xy}^{(k)}(\cdot)$ is continuous (see its representation!), clearly, $q_y^{(\infty)}(\cdot)$ *is also continuous.*

Otherwise, the same conclusion may be derived from the uniform wrt θ convergence of the series for $q^{(\infty)}$ and continuity of each term in that series, due to the Theorem 1.

Now, the part 3 of our plan being fulfilled, let us show the second part, i.e. that $q_y^{(\infty)} = \partial_\theta \pi_y$. We have,

$$p_{xy}^{(k)}(\theta') - p_{xy}^{(k)}(\theta) = \int_\theta^{\theta'} q_{xy}^{(k)}(t)\, dt.$$

In this identity we can pass to the limit as $k \to \infty$, to obtain,

$$\pi_y(\theta') - \pi_y(\theta) = \int_\theta^{\theta'} q_y^{(\infty)}(t)\, dt.$$

The function $q_y^{(\infty)}$ being continuous in t, this clearly (e.g., by the First Theorem of the Calculus) means that $\pi_y(\cdot) \in C^1$ and that $q_y^{(\infty)}$ is its derivative wrt θ. We have proved the following result:

Theorem 4. *Let assumptions of the Ergodic Theorem be satisfied with* $\inf_\theta \kappa(\theta) > 0$*, and let all transition probabilities* $p_{xy}(\theta)$ *belong to the class* C^1*. Then the (uniquely determined) invariant probabilities* $\pi_y(\theta) \in C^1$*, too.*

Similarly *by induction*, with the use of the same calculus, the following result can be proved. We shall derive the equations of the heat type for higher derivatives. Observe, however, that with each iteration the rate of convergence becomes weaker (but still exponential).

Theorem 5. *Let assumptions of the Ergodic Theorem be satisfied with* $\inf_\theta \kappa(\theta) > 0$*, and let all transition probabilities* $p_{xy}(\theta)$ *belong to the class* C^N*,* $1 \le N \le \infty$ *(i.e. including* $+\infty$*). Then the (uniquely determined) invariant probabilities* $\pi_y(\theta) \in C^N$*, too.*

Under the Ergodic Theorem assumptions, *analytic dependence* of invariant probabilities on θ also follows from analytic dependence of transition probabilities. For example, – beside perturbation methods [Kato], – it follows from the uniform convergence (24), or, even easier, from the uniform convergence $p_{xy}^{(k)} \to \pi_y$. (*An Easy Exercise.*)

Naturally, a similar statements that any C^N property for transition probabilities is inherited by any expression like $\sum_z g(z)\pi_z(\theta) = E_{inv} g(X_0)$ can be made. This does not require any further efforts. If this property was valid for any function g, this would have been equivalent to the theorems above for invariant probabilities. Notice, however, that in the continuous time case, for Markov diffusions, the latter statement is wrong: expectations require milder assumptions. What about solutions of Poisson equations? The case (PE1) (with boundary conditions) is easier, and the answer is positive: solutions also inherit any C^N property from transition probabilities. In the case (PE2) (without boundary conditions under the centering assumption) this is also true, because of uniform convergence with some geometric rate for the derivatives in question. E.g., for the first derivative, this can be easily justified by using (24) and the representation formula (PE2-sol).

12 Non-compact framework

We will study processes in non-compact spaces such as R^d, like non-linear auto-regressions,

$$(AR) \qquad X_{n+1} = X_n + f(X_n) + W_{n+1}, \qquad (W_n) \quad \text{i.i.d.}$$

This will be our main example. Often W_n are standard Gaussian, although many other distributions are covered by this approach, too. The word "auto-regression" suggests that there is a flavour of stationarity in this process, hence, it should be somehow ergodic. Clearly, for that we need some "stability conditions" on the function f. E.g., it *may* be

$$|x + f(x)| \le q|x|, \qquad \forall\, x, \qquad 0 < q < 1.$$

About the "noise" (W_n), in addition to i.i.d., we will eventually assume also *non-degeneracy* in the sense that there exists a density $p(\cdot)$ w.r.t. the Lebesgue measure.

For our main example above (AR) there is no further technical assumptions required to define a MP. However, for some future purposes, we mention that if any more general case is considered, we will always assume that our process has a *transition kernel*

$$Q(x, dx') = P_x(X_1 \in dx'),$$

which kernel is a probability measure for each x, and for any Borel A, $Q(\cdot, A)$ is a Borel function in x, or, otherwise, that for any Borel bounded f, the function $\int f(x')Q(x, dx')$ is again Borel in x.

Remind that in discrete time we can always consider our process X_n as *strong* Markov.

Strategy

Our main goals are again: *LLN, CLT, invariant measures, Poisson and heat equations, dependence on parameters.* The strategy will be to have a "good domain", usually a bounded neighbourhood of the origin, where a "good local mixing" occurs, and quick return to such good area once the process leaves it for some excursion.

Hence, technically we will study two practically separate parts: *recurrence*, and how *local mixing* with recurrence imply convergence. In compare to the finite state space case, even local mixing would now require some more attention, because, informally speaking, when we arrange to glue two copies of our MP, they should be now both in a "good domain", so we should study recurrence *for couples* of processes, not just one MP.

12.1 Recurrence

Let $B_R = \{x \in R^d : \ |x| \le R\}$. Under a suitable choice of R, this will be a convenient "good domain" for gluing. Denote

$$\tau_R := \inf(n \ge 0 : \ X_n \in B_R).$$

From the general theory of MP's it is known that ergodic properties depend on whether our MP is *"recurrent"* ($P_x(\tau_R < \infty) = 1$ for each x), *"positive recurrent"* ($E_x \tau_R < \infty$ for each x), or further bounds like $E_x \tau_R^k < \infty$, or $\exists \lambda > 0, \ E_x \exp(\lambda \tau_R) < \infty$ for each x could be established. Positive recurrent processes "usually" possess a probability invariant measure, while recurrent but not positive recurrent – like Wiener process – may have an invariant measure, but it may be infinite (like the Lebesgue one).

Assume that $\exists q \in (0,1)$ such that $\forall\, |x| \ge 1/(1-q^2)$,

$$|x + f(x)| \le q|x|, \tag{25}$$

$$s^2 := E|W_1|^2 < \infty. \tag{26}$$

and for simplicity of presentation also

$$EW_1 = 0. \tag{27}$$

(Why "for simplicity": because otherwise we could subtract EW_1 and include into f, only slightly changing (25).)

Theorem 1. If $R \ge (1-q^2)^{-1}$, then for any x,

$$E_x \tau_R \le |x|^2. \tag{28}$$

The proof is a typical application of simplest martingale ideas. Let $x \notin B_R$ (otherwise $\tau_R = 0$). Given $X_n \notin B_R$, consider $E_{X_n}|X_{n+1}|^2$: due to (25) and (27),

$$E_{X_n}|X_n + f(X_n) + W_{n+1}|^2 - |X_n|^2 \le -(1-q^2)|X_n|^2 + s^2.$$

From this it follows,

$$\begin{aligned} 1(\tau_R > n)\ &\Big(E_{X_n}|X_{n+1}|^2 - |X_n|^2\Big) \\ &\le 1(\tau_R > n)\,(-(1-q^2)|X_n|^2 + s^2), \end{aligned}$$

if R is chosen so that $-(1-q^2)R^2 + s^2 \le -1$. Then,

$$1(\tau_R > n)\ \Big(E_{X_n}|X_{n+1}|^2 - |X_n|^2\Big) \le -1(\tau_R > n).$$

$$1(\tau_R > n)\ \Big(E_{X_n}|X_{n+1}|^2 - |X_n|^2\Big) \le -1(\tau_R > n),$$

so,

$$\sum_{n=0}^{N-1} 1(\tau_R > n)\ \Big(E_{X_n}|X_{n+1}|^2 - |X_n|^2\Big) \le -\sum_{n=0}^{N-1} 1(\tau_R > n).$$

Since $1(\tau_R > n) \le 1(\tau_R > n-1)$, this implies,

$$\begin{aligned} \sum_{n=0}^{N-1} 1(\tau_R > n)\, E_{X_n}|X_{n+1}|^2 - \sum_{n=0}^{N-1} 1(\tau_R > n-1)\, |X_n|^2 \\ \le -\sum_{n=0}^{N-1} 1(\tau_R > n). \end{aligned}$$

By the way, why we are going to use

$$1(\tau_R > n) \le 1(\tau_R > n-1), \tag{29}$$

and not, say, $1(\tau_R > n+1) \le 1(\tau_R > n)$, which is also correct? Because we will take expectations, and in the term $1(\tau_R > n)\, E_{X_n}|X_{n+1}|^2$ the indicator $1(\tau_R > n)$ is measurable with respect to $\mathcal{F}^X_n$. So, by Markov's property,

$$\begin{aligned} &E_x 1(\tau_R > n)\, E_{X_n}|X_{n+1}|^2 \\ &= E_x 1(\tau_R > n)\, E_x(|X_{n+1}|^2 \mid \mathcal{F}^X_n) \\ &= E_x 1(\tau_R > n)|X_{n+1}|^2). \end{aligned}$$

Such trick is only possible with (29).

Thus, taking expectations, we get,

$$\begin{aligned} \sum_{n=0}^{N-1} E_x 1(\tau_R > n)|X_{n+1}|^2 - \sum_{n=0}^{N-1} E_x 1(\tau_R > n-1)\, |X_n|^2 \\ \le -\sum_{n=0}^{N-1} E_x 1(\tau_R > n), \end{aligned}$$

or,

$$\sum_{n=0}^{N-1} E_x 1(\tau_R - 1 \geq n)$$
$$\leq \sum_{n=0}^{N-1} E_x 1(\tau_R > n-1)\,|X_n|^2 - \sum_{n=0}^{N-1} E_x 1(\tau_R > n)|X_{n+1}|^2.$$

Notice that in the r.h.s. here all terms are canceled out, except the first one and the last one, i.e.,

$$E_x \sum_{n=0}^{N-1} 1(\tau_R - 1 \geq n) \leq E_x 1(\tau_R \geq 0)\,|x|^2$$
$$-E_x 1(\tau_R > N)|X_{N+1}|^2 \leq |x|^2.$$

But

$$\sum_{n=0}^{N-1} 1(\tau_R - 1 \geq n) = \sum_{n=0}^{(N-1)\wedge(\tau_R-1)} 1 = \tau_R \wedge N.$$

So, $\boxed{E_x(\tau_R \wedge N) \leq |x|^2.}$

Our definition of expectation was via monotone convergence (if not, this should have been a lemma), i.e., in particular,

$$E_x\tau_R = \lim_{N\uparrow\infty} E_x(\tau_R \wedge N).$$

Hence, because $|x|^2$ does not depend on N, we get,

$$E_x\tau_R \leq |x|^2,$$

as required. This was done under the assumption $|x| > R$. If $|x| \leq R$, then clearly $\tau_R = 0$. So, in all cases,

$$\boxed{E_x\tau_R \leq |x|^2.}$$

Under $E|W_1|^2 < \infty$, we have no opportunity to work with higher moments. However, there was at least one place where we could relax our assumptions, namely, when we were to choose R large enough. Here is one possible generalization. Assume

$$\limsup_{|x|\to\infty} (|x + f(x)| - |x|) < 0, \quad \& \quad f \text{ locally bounded.} \tag{30}$$

Theorem 2. Under (30), (26)/(2nd moment for W), and (27)/(zero expectation of W), for any R large enough there exists $C > 0$ such that for any x,

$$\boxed{E_x \tau_R \le C|x|^2.}$$

Notice that now the assumption (27) *is* essential.

Under (30), if there exists some exponential moment of W, another exponential moment is finite for hitting time τ_R like in the Theorem 3 (under the more restrictive (25)) below [Gulinsky, AYV, 1993, the chapter about mixing.]

For r large enough, under

$$\limsup_{|x|\to\infty} \left(\frac{|x + f(x)|}{|x|} - 1\right) |x|^2 =: -r < 0, \quad \& \quad f \ \text{loc bdd}, \tag{31}$$

some polynomial moment can be estimated [AYV, 2000]

$$E_x \tau_R^k \le C(1 + |x|^m),$$

with m depending on k and the moment assumptions on W; possibly $m = 2k$. For sub-exponential bounds see [Klokov, AYV, 2003, et al.].

Thus, under (30), (26), and (27), we have $E_x\tau_R \le C|x|^2$. Let us show that this suffices for existence of at least one invariant measure (we are not talking of uniqueness so far). We consider the "process on B_R", that is, $X_n^{B_R} := X_{\tau^n}$, $n \ge 0$, where

$$\tau^1 := \tau_R, \quad \text{and} \quad \tau^{n+1} := \inf(k > \tau^n;\, X_k \in B_R).$$

The sequence of distributions $P(X_n^{B_R} \in \cdot)$ is *compact* in the sense of Prokhorov (or weak convergence), i.e., they are all supported by some compact domain (by definition, up to any $\epsilon > 0$, but in our case with $\epsilon = 0$). Then, the sequence $(n+1)^{-1}\sum_{t=1}^n P(X_k^{B_R} \in \cdot) =: \nu^n$ is also compact. So (Lemma) there exists a subsequence $n' \to \infty$ such that $\nu^{n'}$ has a weak limit ν. Then ν is stationary for X^{B_R}.

The fact that ν is stationary can be proved similarly to the finite state space case. The Lemma about existence of a convergent subsequence follows by the diagonalisation procedure from the fact that the space of continuous functions on R^1 (R^d) is separable, that is, for any $\epsilon > 0$ there exists a not more than countable ϵ–net (which is a family $f_1, f_2, \ldots \in C$ such that for any

$f \in C$, there exists f_m such that $\rho(f, f_m) < \epsilon$). We then take a sequence of $\epsilon_k \to 0$, and all functions from the ϵ_k–nets, say, $\{f_{k,m}\}$. On each of them we may have a convergence

$$\int f_{k,m} d\nu^{n'}, \quad n' \to \infty,$$

over some subsequence $\{n'\}$. Convergence on any $f \in C$ then follows from the convergence on any $f_{k,m}$.

Notice that this consideration is valid on the whole R^1 (R^d), too, however, the measures may converge to zero or to some sub-probability measure (as the measure may in some sense "escape" to infinity). But for a compact family of measures this is impossible.

In this way, we have shown that there exists a measure on B_R, say, ν, which is stationary for our MP ($X_n^{B_R}$, $n \geq 0$). *Harris' principle* then suggests how construct an invariant measure for the original process (X_n, $n \geq 0$): for any $A \in \mathcal{B}^1$ ($A \in \mathcal{B}^d$), let

$$\mu(A) := c \int_{B_R} \left(E_x \sum_{t=0}^{T-1} 1(X_t \in A) \right) \nu(dx), \tag{32}$$

where $T := \inf(t > 0 : \ X_t \in B_R)$, and $c \int_{B_R} E_x T \, \nu(dx) = 1$.

Proof. We are under conditions of the Theorems 1 or 2.

First of all, let us see why μ is well-defined. We have,

$$\begin{aligned} E_x T = P_x(X_1 \in B_R) + E_x 1(X_1 \notin B_R)(1 + E_{X_1} \tau_R) \\ \leq 1 + E_x 1(X_1 \notin B_R)(1 + C|X_1|^2) \\ \leq 2 + C E_x |X_1|^2 = 2 + C E_x |x + f(x) + W_1|^2 \\ \leq 2 + 2C(\sup_{|x| \leq R} |x + f(x)|^2 + s^2) < \infty, \end{aligned}$$

for $|x| \leq R$. Hence,

$$\mu(A) := c \int_{B_R} \left(E_x \sum_{t=0}^{T-1} 1(X_t \in A) \right) \nu(dx) \leq c \sup_{|x| \leq R} E_x T < \infty.$$

Now let us see why the formula (32) defines a stationary measure. We have,

$$\mu(dy) = c \int_{B_R} \left(E_x \sum_{t=0}^{T-1} 1(X_t \in dy) \right) \nu(dx).$$

So,

$$\int_{R^d} P_y(X_1 \in A)\mu(dy)$$

$$= c\int_{R^d} P_y(X_1 \in A) \int_{B_R} (E_x \sum_{t=0}^{T-1} 1(X_t \in dy))\, \nu(dx)$$

$$+c\int_{B_R} (E_x \sum_{t=0}^{T-1} 1(X_t \in dy)\, \int_{R^d} P_y(X_1 \in A))\, \nu(dx).$$

But clearly,

$$c\int_{B_R} (E_x \sum_{t=0}^{T-1} 1(X_t \in dy)\, \int_{R^d} P_y(X_1 \in A))\, \nu(dx)$$

$$= c\int_{B_R} (E_x \sum_{t=1}^{T} 1(X_t \in A)\, \nu(dx)$$

$$= \mu(A) + c\int_{B_R} (E_x 1(X_T \in A) - E_x 1(x \in A))\, \nu(dx)$$

$$= \mu(A) + c\int_{B_R} (E_x 1(X_1^{B_R} \in A)\, \nu(dx) - c\nu(A) = \mu(A),$$

the latter since ν is invariant on B_R.

A priori bounds

What we need next is *a priori* bounds like

$$\sup_n E_x|X_n|^2 \le h(x) < \infty. \qquad (33)$$

Intuitively, there is quite a good hope to have such inequality, because, so to say, "if $|X_n|$ is small, its square along with expectation of this square remains not large, and while if it is large, its expectation should decrease with time" (see the proof of the $E_x\tau_R < \infty$). However, it is a bit unclear how to realise this hope. Then, there are two ways. Firstly, we may build our further analysis on a weaker bound

$$\sup_n E_x|X_n|^2 1(\tau_R \ge n) < \infty. \qquad (34)$$

Secondly, one more opportunity is to assume more about moments of W_n and to work with exponential moments. Why? Simply because for exponentials it is usually easier to get stronger *a priori* bounds

$$\sup_n E_x \exp(\lambda|X_n|) < \infty. \qquad (35)$$

Here we will prefer the second way, although the first one is also OK. To start with, we, of course, have to assume that

$$E\exp(\lambda|W_1|) < \infty$$

with some $\lambda > 0$. To *simplify* the presentation, we will work on R^1, although it may be repeated on R^d, too (where the calculus would be a bit more involved).

12.2 Stability

Exponential *a priori* bounds

So, we assume

$$E\exp(\lambda|W_1|) < \infty, \qquad \exists\, \lambda > 0,$$

and also the assumption of the Theorem 2 above,

$$\limsup_{|x|\to\infty} (|x + f(x)| - |x|) < 0, \quad \& \quad f \text{ locally bounded},$$

and

$$EW_1 = 0.$$

Under such nice conditions, we may have much more than $E\tau_R < \infty$, which itself hardly suffices for any good quantitative convergence rate, say, in the Ergodic Theorem. (By the way, in fact, so far we are *not* yet under any Ergodic Theorem conditions at all.)

12.3 Exponential hitting bounds

Assume *for simplicity of presentation*,

$$|x + f(x)| \le q|x|, \qquad \forall\, x, \qquad 0 < q < 1, \tag{36}$$

$$\psi(\lambda) := E\exp(\lambda|W_1|) < \infty, \tag{37}$$

and for simplicity of presentation also

$$EW_1 = 0. \tag{38}$$

(Here "for simplicity" means that we can do the same under the Theorem 2 stability assumptions instead of (51).)

Theorem 3. If R is large enough, then $\exists\, C$, α, $\lambda > 0$ such that for any x,

$$E_x \exp(\alpha\tau_R) \le C \exp(\lambda(|x| - R)_+). \tag{39}$$

Proof of Theorem 3

We start with some provisional estimates now. Let $x \notin B_R$ (otherwise $\tau_R = 0$). Given $X_n \notin B_R$, consider $E_{X_n} \exp(\lambda|X_{n+1}|)$: due to (51) and (38),

$$\frac{E_{X_n} \exp(\lambda|X_n + f(X_n) + W_{n+1}|)}{\exp(\lambda|X_n|)} \le \psi(\lambda) \exp(-\lambda|X_n|(1-q)).$$

This does not exceed some $\kappa < 1$ if R is large enough. So,

$$\begin{gathered} 1(\tau_R > n)\ (E_{X_n} \exp(\lambda|X_n + f(X_n) + W_{n+1}|) - \exp(\lambda|X_n|)) \\ \le -1(\tau_R > n)\,(1-\kappa)\exp(\lambda|X_n|), \\ \text{or,} \\ 1(\tau_R > n)\ (E_{X_n} \exp(\lambda|X_{n+1}|) - \exp(\lambda|X_n|)) \\ \le -1(\tau_R > n)(1-\kappa)\exp(\lambda|X_n|). \end{gathered}$$

$$\begin{gathered} 1(\tau_R > n)\ (E_{X_n} \exp(\lambda|X_{n+1}|) - \exp(\lambda|X_n|)) \\ \le -1(\tau_R > n)(1-\kappa)\exp(\lambda|X_n|), \end{gathered}$$

so,

$$\begin{aligned} &\sum_{n=0}^{N-1} 1(\tau_R > n)\ (E_{X_n} \exp(\lambda|X_{n+1}|) - \exp(\lambda|X_n|)) \\ &\qquad\qquad \le -(1-\kappa)\sum_{n=0}^{N-1} 1(\tau_R > n)\exp(\lambda|X_n|). \end{aligned} \tag{40}$$

Since $1(\tau_R > n-1) \ge 1(\tau_R > n)$, this implies, as earlier,

$$\begin{aligned} &\sum_{n=0}^{N-1} 1(\tau_R > n) E_{X_n} \exp(\lambda|X_{n+1}|) - \sum_{n=0}^{N-1} 1(\tau_R > n-1)\exp(\lambda|X_n|) \\ &\qquad\qquad \le -(1-\kappa)\sum_{n=0}^{N-1} 1(\tau_R > n)\exp(\lambda|X_n|). \end{aligned}$$

Canceling equal terms and taking expectations, we get,

$$\begin{aligned}(1-\kappa)E_x\sum_{n=0}^{N-1}1(\tau_R>n)\exp(\lambda|X_n|)\\ \le E_x1(\tau_R\ge 0)\exp(\lambda|x|)-E_x1(\tau_R>N)E_{X_N}\exp(\lambda|X_{N+1}|).\end{aligned}$$

In the other words,

$$(1-\kappa)E_x\sum_{n=0}^{(\tau_R-1)\wedge(N-1)}\exp(\lambda|X_n|)\le\exp(\lambda|x|).$$

As earlier, from the monotone convergence it follows,

$$(1-\kappa)E_x\sum_{n=0}^{\tau_R-1}\exp(\lambda|X_n|)\le\exp(\lambda|x|).$$

This inequality and the Harris principle, – see (32), – imply the first *a priori* bound, for any invariant measure μ,

$$\int\exp(\lambda|x|)\,\mu(dx)<\infty. \tag{41}$$

Rather close to the latter is another observation:

$$\frac{E_{X_n}\exp(\lambda|X_{n+1}|)}{\exp(\lambda|X_n|)}\le\kappa<1, \qquad |x|\ge R,$$

and

$$\sup_{|x|\le R}E_x\exp(\lambda|X_1|)\le C<\infty,$$

imply that for every X_n,

$$E_{X_n}\exp(\lambda|X_{n+1}|)\le C+\kappa\exp(\lambda|X_n|).$$

Let $z_n:=E_x\exp(\lambda|X_n|)$ $(<\infty)$. Then, by induction,

$$\sup_n E_x\exp(\lambda|X_n|)=\sup_n z_n\le\frac{C}{1-\kappa}. \tag{42}$$

Now, similarly to (40), but with a *new Lyapunov function* $f(n,x)=\exp(\alpha n+\lambda|x|)$, we get,

$$\begin{aligned}&1(\tau_R>n)\;(E_{X_n}\exp(\lambda|X_{n+1}|)-\exp(\lambda|X_n|))\\ &\quad\le -1(\tau_R>n)(1-\kappa\exp(\alpha))\exp(\lambda|X_n|),\end{aligned}$$

so with a new constant $\kappa' = \kappa \exp(\alpha) < 1$ (if α is small enough), we repeat the earlier calculus to get,

$$\sum_{n=0}^{N-1} 1(\tau_R > n) E_{X_n} f(n+1, X_{n+1}) - \sum_{n=0}^{N-1} 1(\tau_R > n-1) f(n, X_n)$$
$$\le -(1-\kappa') \sum_{n=0}^{N-1} 1(\tau_R > n) f(n, X_n).$$

Cancelling equal terms and taking expectations, we get,

$$(1-\kappa') E_x \sum_{n=0}^{N-1} 1(\tau_R > n) \exp(\alpha n + \lambda |X_n|)$$
$$\le E_x 1(\tau_R \ge 0) \exp(\lambda |x|).$$

In the other words,

$$(1-\kappa') E_x \sum_{n=0}^{(\tau_R-1)\wedge(N-1)} \exp(\alpha n + \lambda |X_n|) \le \exp(\lambda |x|).$$

As earlier, from the monotone convergence it follows,

$$(1-\kappa') E_x \sum_{n=0}^{\tau_R-1} \exp(\alpha n + \lambda |X_n|) \le \exp(\lambda |x|).$$

This implies,

$$\exp((\lambda |x| - R)_+)/(1-\kappa')$$
$$\ge E_x \sum_{n=0}^{\tau_R-1} \exp(\alpha n) = \frac{E_x \exp(\alpha \tau_R)}{\exp(\alpha) - 1}.$$

So, the desired bound (53) is proved.

Open questions

1. For generalizations which would rather require more subtle tool (34), one strangely unsolved question is whether the stronger *a priori* bounds like (33) can be achieved, *without local mixing conditions* (usually required in the available texts). The matter is that they are achieved only *after* establishing some good rate of convergence towards a unique invariant measure, which, of course, do require some local mixing.

2. There is a gap between some minimal critical r value under which some rate of convergence is known and another critical r, below which there is no invariant measure, generally speaking. This is the most intriguing, in particular, because of a great interest about heavy tails of distributions in various applications.

Uniqueness?

How about uniqueness of invariant measure, and convergence rate towards it?

13 Non-compact framework

Local mixing condition

We are studying strong Markov chains in R^d, like AR's

$$(AR) \qquad X_{n+1} = X_n + f(X_n) + W_{n+1}, \qquad (W_n) \quad \text{i.i.d.},$$

which admit the bounds from the Theorem 3. Now we add the "local mixing condition" on the *density* p_W *of the r.v.* W:

$$\boxed{\inf_{|x|\le R} p_W(x) > 0, \qquad \forall\, R > 0.} \tag{43}$$

Theorem 4. Under the assumptions of Theorem 3 and (56),

$$\|\mu_t - \mu_{inv}\|_{TV} \to 0, \qquad t \to \infty, \tag{44}$$

where μ_{inv} is the unique invariant measure of the MC X, and $\mu_t = \mathcal{L}(X_t)$ (= marginal distribution of X).

Total variation metric

The total variation distance between two measures μ, ν on $(R^d, \mathcal{B}^d)$ is defined as

$$\|\mu - \nu\|_{TV} := \sup_{A\in\mathcal{B}^d} (\mu - \nu)(A) + \sup_{B\in\mathcal{B}^d} (\nu - \mu)(B).$$

Clearly, each term here in the right hand side is non-negative, because $\emptyset \in \mathcal{B}^d$. For probabilistic measures

$$\|\mu - \nu\|_{TV} = 2 \sup_{A\in\mathcal{B}^d} (\mu - \nu)(A).$$

If there are densities w.r.t. (e.g.) Lebesgue's measure, p_μ and p_ν, then also

$$\|\mu - \nu\|_{TV} = \int |p_\mu(x) - p_\nu(x)|\, dx \quad (\overset{\text{btw}}{=} 2 - 2\kappa).$$

Reminder: Theorem 3

We will use the conclusions of the Theorem 3:

Theorem 3. If R is large enough, then $\exists\, C$, α, $\lambda > 0$ such that for any x,

$$E_x \exp(\alpha \tau_R) \le C \exp(\lambda(|x| - R)_+). \tag{45}$$

Also,

$$\int \exp(\lambda |x|)\, \mu(dx) < \infty, \tag{46}$$

and

$$\sup_n E_x \exp(\lambda |X_n|) \le C < \infty. \tag{47}$$

Remind that Theorem 3 does not use condition (56).

To extend our analysis performed for finite state spaces, we will need two technical lemmas. As we have seen, under the assumptions of any Theorem 1–3 from the previous section, there exists at least one invariant measure. Now we are going to consider two independent copies of our MC, (X_n) and $(\tilde{X}_n)$, he second being in the invariant regime, and we would like to arrange their gluing. We use notations,

$$\tau_R = \inf(n \ge 0 : \ X_n \in B_R), \quad \tilde{\tau}_R = \inf(n \ge 0 : \ \tilde{X}_n \in B_R),$$

and

$$\gamma_R = \inf(n \ge 0 : \ X_n \in B_R, \ \underline{\text{and}}\ \tilde{X}_n \in B_R).$$

Bound for γ_R

Lemma. For R large enough,

$$E_{x,x'} \exp(\alpha \gamma_R) \le C \exp(\lambda(|x| - R)_+ + \lambda(|x'| - R)_+). \tag{48}$$

Proof. We apply the same calculus as for one copy of the process in the Theorem 3, with minor changes.

Notice that under our assumptions, for any X_n,

$$E_{X_n} \exp(\lambda |X_{n+1}|) \le C \exp(\lambda |X_n|),$$

with some $C > 0$. So, either (if $X_n \notin B_R, \tilde{X}_n \notin B_R$)

$$\begin{aligned} & E_{X_n,\tilde{X}_n} \exp(\lambda(|X_{n+1}| + |\tilde{X}_{n+1}|)) \\ & \le C \exp(-\lambda |X_n| (1-q)) \exp(\lambda |X_n|) \\ & \times C \exp(-\lambda |\tilde{X}_n| (1-q)) \exp(\lambda |\tilde{X}_n|), \end{aligned}$$

Bound for γ_R
or

$$\begin{aligned} & E_{X_n,\tilde{X}_n} \exp(\lambda(|X_{n+1}| + |\tilde{X}_{n+1}|)) \\ & \le C \exp(-\lambda |X_n| (1-q)) \exp(\lambda |X_n|) \times C \exp(\lambda |\tilde{X}_n|), \end{aligned}$$

if only $X_n \notin B_R$, or, at last and similarly,

$$\begin{aligned} & E_{X_n,\tilde{X}_n} \exp(\lambda(|X_{n+1}| + |\tilde{X}_{n+1}|)) \\ & \le C \exp(-\lambda |\tilde{X}_n| (1-q)) \exp(\lambda |\tilde{X}_n|) \times C \exp(\lambda |X_n|), \end{aligned}$$

if only $\tilde{X}_n \notin B_R$. In all cases, for R large enough,

$$E_{X_n,\tilde{X}_n} \exp(\lambda(|X_{n+1}| + |\tilde{X}_{n+1}|)) \le q \, \exp(\lambda(|X_n| + |\tilde{X}_n|)), \quad q < 1.$$

So the estimate can be completed as in the Theorem 3.

Lemma of three r.v.'s

Lemma. Let ξ^1, ξ^2 be two r.v.'s, each on its own probability space, with

$$\int q_{\xi^1}(x) \wedge q_{\xi^2}(x) \, dx = \kappa > 0.$$

Then there exists a new probability space which is some extension of the direct product of the first two, on which there exists a third r.v. ξ^3, such that distribution of it coincides with the distribution of ξ^1, and, at the same time,

$$P(\xi^3 = \xi^2) = \kappa.$$

Remark. The densities may be considered wrt to some measure other than Lebesgue's.

Application of Lemma of three r.v.'s

For our MC, we assume *local Dobrushin's condition,*

$$\min_{x,\tilde{x}\in B_R} \int q(x,x') \wedge q(\tilde{x},x')\,dx' = \kappa > 0.$$

Then, each time when the two process are both in B_R, we can reconstruct one of them (non-stationary) using the Lemma of three random variables, so that a new version of the process, say, $(\hat{X}_n)$, has the same transition density, and coincides with the stationary one with probability at least κ. Each time when both processes are in B_R, they have a chance to meet at least κ. Convergence rate would have been exponential ($\le (1-\kappa)^n$), were the processes always in B_R. However, in general they may have excursions outside, and we wait until they are both back. Under Theorem 3, the overall estimate should remain exponential.

14 Convergence

Convergence in total variation

We consider the couple of independent processes, X and *an* invariant version $\tilde{X}$. Consider the sequence of stopping times γ^n:

$$\gamma^1 = \gamma, \qquad \gamma^{n+1} := \gamma(\gamma^n) \quad \text{(informally speaking)}.$$

At each γ^n, the "reconstructed" process X has a chance ($\ge \kappa$) to meet and to be glued with $\tilde{X}$. By the (strong) Markov property,

$$P(L > \gamma^n) \le (1-\kappa)^{n-1},$$

where by L we denote the first moment of meeting. Let $\epsilon > 0$, and choose n so that $(1-\kappa)^{n-1} \le \epsilon/2$.

Since each $P(\gamma^n < \infty) = 1$, this implies that

$$\sup_A (P(X_t \in A) - P(\tilde{X}_t \in A)) \le P(L > t) \tag{49}$$
$$\le P(L > \gamma^n) + P(L > t, L \le \gamma^n) \le \epsilon/2 + P(t < \gamma^n).$$

But for given n and random, but finite γ^n, clearly, $P(t < \gamma^n) \to 0$, $t \to \infty$. Hence, there exists t such that $P(t < \gamma^n) \le \epsilon/2$, and, therefore,

$$\sup_A (P(X_t \in A) - P(\tilde{X}_t \in A)) \le \epsilon.$$

Hence, we obtain (44),

$$\|\mu_t - \mu_{inv}\|_{TV} \to 0, \qquad t \to \infty.$$

Theorem 5. Under the assumptions of Theorem 3 and (56),

$$\|\mu_t - \mu_{inv}\|_{TV} \le C(x)\exp(-ct), \tag{50}$$

where $x = X_0$.

Proof. Let us use the hint (49) more accurately. By *Rogers–Hölder's* inequality, with $1/a + 1/b = 1$,

$$\begin{aligned} P(L > t) &= \sum_{n=0}^{\infty} E1(L > t)1(\gamma^n \le t < \gamma^{n+1}) \\ &\le \sum_{n\ge 0} P(L > \gamma^n)^{1/a} P(\gamma^{n+1} > t)^{1/b} \\ \le \sum_{n\ge 0} &(1-\kappa)^{n/a} P(\gamma^{n+1} > t)^{1/b}/(1-\kappa)^{1/a}. \end{aligned}$$

Convergence rate

By Bienaimé–Chebyshev's inequality and by induction,

$$P(\gamma^{n+1} > t) \le e^{-\alpha t} E e^{\alpha\gamma^{n+1}}$$

$$= e^{-\alpha t}\, E\, e^{\alpha(\gamma^1 + \sum_{k=1}^{n}(\gamma^{k+1} - \gamma^k))}$$

$$\le e^{-\alpha t} C_R^n\, C \exp(\lambda((|x| - R)_+ + (|x'| - R)_+)).$$

Hence,

$$P(L > t) \le C \exp(\lambda((|x| - R)_+ + (|x'| - R)_+))/1 - \kappa)^{1/a}$$

$$\times \exp(-\alpha b^{-1} t) \sum_{n\ge 0} \exp(-n(a^{-1}\ln(1-\kappa)^{-1} - b^{-1}\ln C_R)).$$

Convergence rate in total variation

By choosing $a > 1, b > 1$, so that

$$a^{-1}\ln q^{-1} - b^{-1}\ln(C_R) > 0,$$

which is possible because

$$\lim_{b\to\infty} b^{-1}\ln(C_R) = 0$$

and

$$\lim_{a\to 1} a^{-1}\ln(1-\kappa)^{-1} = \ln(1-\kappa)^{-1} > 0,$$

we get here in the right hand side a convergent series in n which does not depend on t, and, hence, the required bound (57). The Theorem 5 is proved.

Convergence for beta–mixing

Reminder

Remind that we are studying strong Markov chains in R^d,

$$(AR) \qquad X_{n+1} = X_n + f(X_n) + W_{n+1}, \qquad (W_n) \quad \text{i.i.d.},$$

under the assumptions

$$|x+f(x)| \le q|x|, \qquad \forall\, x, \qquad 0<q<1, \tag{51}$$

$$\psi(\lambda) := E\exp(\lambda|W_1|) < \infty, \quad \& \quad EW_1 = 0. \tag{52}$$

Reminder: Theorem 3. Under (51) and (52), if R is large enough, then $\exists\, C$, α, $\lambda > 0$ such that for any x,

$$E_x\exp(\alpha\tau_R) \le C\exp(\lambda(|x|-R)_+). \tag{53}$$

Also, *a priori* bounds hold true for $0<\epsilon\le\epsilon_0$,

$$\int \exp(\epsilon|x|)\,\mu(dx) < \infty, \tag{54}$$

$$\sup_n E_x\exp(\epsilon|X_n|) \le C\exp(\epsilon|x|) < \infty. \tag{55}$$

Next step was done when we added the "local mixing condition" on the *density* p_W *of the r.v.* W:

$$\boxed{\inf_{|x|\le R} p_W(x) > 0, \qquad \forall\, R>0.} \tag{56}$$

Reminder: Theorem 5. Under the assumptions of Theorem 3 and (56),

$$\|\mu_t^x - \mu_{inv}\|_{TV} \le C\exp(\epsilon|x|)\exp(-ct), \tag{57}$$

where $x = X_0$, μ_{inv} is the unique invariant measure of the MC X, and $\mu_t = \mathcal{L}(X_t)$ (= marginal distribution of X).

15 LLN

LLN for stationary MC: non-compact case

Assumption (A1'): Consider a MP satisfying the assumptions of Theorems 3 and 5 above.

Theorem 1': [stationary weak LLN] *For a stationary MC under (A1'), for any f on the state space R^d with $\int f^2\, d\mu_{inv} < \infty$,*

$$\frac{1}{n}\sum_{k=0}^{n-1} f(X_k) \overset{P}{\to} E_{inv} f(X_0),$$

where E_{inv} stands for expectation with respect to the invariant measure, $E_{inv} f(X_0) = \int f\, d\mu_{inv} < \infty$.

Proof. We will use Bienaimé–Chebyshev inequality with variance. Assume $E_{inv} f(X_0) = 0$, otherwise subtract.

$$\begin{aligned} P_{inv}(|\frac{1}{n}\sum_{k=0}^{n-1} f(X_k)| \ge \epsilon) &\le \frac{E_{inv}|\sum_{k=0}^{n-1} f(X_k)|^2}{\epsilon^2 n^2} \\ = \frac{\sum_{k=0}^{n-1} E_{inv} f^2(X_k)}{\epsilon^2 n^2} &+ \frac{2\sum_{k<j}^{n-1} E_{inv} f(X_k) f(X_j)}{\epsilon^2 n^2}. \end{aligned} \tag{58}$$

$$\begin{aligned} \text{But} \qquad |E_{inv} f(X_k) f(X_j)| &= |E_{inv} f(X_k) E(f(X_j) \mid X_k)| \\ \le \quad \text{(in fact, =)} \quad & |E_{inv} f(X_k) E_{inv}(f(X_j))| \\ &+ |E_{inv} f(X_k)[E(f(X_j) \mid X_k) - E_{inv}(f(X_j)]| \\ &= |E_{inv} f(X_k)[E(f(X_j) \mid X_k) - E_{inv}(f(X_j)]|. \end{aligned}$$

Remind that $E_{inv} f(X_k) E_{inv} f(X_j) = 0$.

$$\text{Since} \qquad \|\mu_t^x - \mu_{inv}\|_{TV} \le C\exp(\epsilon|x|)\exp(-ct),$$

and due to (54) and (55), we have,

$$\begin{aligned} & |[E(f(X_j) \mid X_k) - E_{inv}(f(X_j)]| \\ &= |\int f(x') p^{(j-k)}(X_k, x')\, dx' - \int f(x') p_\infty(x')\, dx'| \\ &= |\int f(x') \times 1 \times (p^{(j-k)}(X_k, x') - p_\infty(x'))\, dx'| \\ &\le |\int f^2(x') |p^{(j-k)}(X_k, x') - p_\infty(x')|\, dx'|^{1/2} \\ &\quad \times (\int 1^2 |p^{(j-k)}(X_k, x') - p_\infty(x')|\, dx')^{1/2} \\ &\le C\exp(\epsilon|X_k|/2)\exp(-c(j-k)/2). \end{aligned}$$

Now we can estimate $|E_{inv} f(X_k)[E(f(X_j) \mid X_k) - E_{inv}(f(X_j)]|$:

$$
\begin{aligned}
&|E_{inv} f(X_k)[E(f(X_j) \mid X_k) - E_{inv}(f(X_j)]| \\
&\le E_{inv} f(X_k)\, C\, \exp(\epsilon|X_k|/2)\exp(-c(j-k)/2) \\
&= C\exp(-c(j-k)/2) E_{inv} f(X_k)\, \exp(\epsilon|X_k|/2) \\
&\le C\exp(-c(j-k)/2)(E_{inv} f^2(X_k))^{1/2}\,(E_{inv}\exp(\epsilon|X_k|))^{1/2} \\
&\le C\exp(-c(j-k)/2),
\end{aligned}
$$

due to the condition $\int f^2\, d\mu_{inv} < \infty$ for the first multiple and a priori bounds from the Theorem 5 for the second multiple. So, the second term in (58) admits the bound,

$$
\frac{2\sum_{k<j}^{n-1} E_{inv} f(X_k) f(X_j)}{\epsilon^2 n^2} \le C\epsilon^{-2} n^{-2}.
$$

And the first term does not exceed Cn^{-1}. Thus, the Theorem 1' is proved.

LLN for non-stationary MC

Theorem 2': [non-stationary weak LLN] *Under (A1'), for any* $f \in L_2(\mu_{inv})$,

$$
\frac{1}{n}\sum_{k=0}^{n-1} f(X_k) \overset{P}{\to} E_{inv} f(X_0).
$$

Of course, condition $f \in L_2(\mu_{inv})$ can be easily relaxed to $f \in L_{1+\delta}(\mu_{inv})$ with any $\delta > 0$ (then we should use Hölder's inequality instead of CBS), and even, – although, less easily, – to $f \in L_1(\mu_{inv})$ (however, the latter is not the aim of this course).

Proof. For simplicity, consider f bounded. Let $\tilde{X}$ be the stationary MC, X the original one, and $\hat{X}$ denote X *switched* to the stationary after the first meeting (τ). We have,

$$
\begin{aligned}
P(|\frac{1}{n}\sum_{k=0}^{n-1} f(\hat{X}_k)| \ge \epsilon) &\le P(|\frac{1}{n}\sum_{k=0}^{n-1} f(\tilde{X}_k)| \ge \epsilon/2) \\
&+ P(|\sum_{k=0}^{n-1} f(\hat{X}_k) - \sum_{k=0}^{n-1} f(\tilde{X}_k)| \ge n\epsilon/2) \\
\le P(|\frac{1}{n}\sum_{k=0}^{n-1} f(\tilde{X}_k)| \ge \epsilon/2) &+ P(C_f \tau \ge n\epsilon/2) \to 0,
\end{aligned}
$$

as $n \to \infty$, because the second term here satisfies $P_x(\tau > n) \le C \exp(\epsilon|x|)$ from the *proof* of the Theorem 5.

16 CLT

Non-compact CLT

Theorem 3': [stationary CLT in non-compact case] *Under (A1'), for any $f \in L_{2+\delta}(\mu_{inv})$ with any $\delta > 0$,*

$$\frac{1}{\sqrt{n}} \sum_{k=0}^{n-1} (f(X_k) - E_{inv} f(X_k)) \overset{P_{inv}}{\Longrightarrow} \sigma Z,$$

where $Z \sim \mathcal{N}(0,1)$ and

$$0 \le \sigma^2 = E_{inv}(f(X_0) - E_{inv} f(X_0))^2$$
$$+2 \sum_{k=1}^{\infty} E_{inv}(f(X_0) - E_{inv} f(X_0))(f(X_k) - E_{inv} f(X_k)).$$

(We consider 0 as a (degenerate) Gaussian r.v.)

Comment on σ^2:

$$E_{inv} | \frac{1}{\sqrt{n}} \sum_{k=0}^{n-1} (f(X_k) - E_{inv} f(X_k)) |^2$$
$$= \frac{1}{n} E_{inv} \sum_{k=0}^{n-1} \sum_{j=0}^{n-1} (f(X_k) - E_{inv} f(X_k))(f(X_j) - E_{inv} f(X_j))$$
$$= \frac{1}{n} E_{inv} \sum_{k=0}^{n-1} (f(X_k) - E_{inv} f(X_k))^2$$
$$+ \frac{2}{n} E_{inv} \sum_{k=0}^{n-1} \sum_{j=k+1}^{n-1} (f(X_k) - E_{inv} f(X_k))(f(X_j) - E_{inv} f(X_j)),$$

and the difference between the latter and σ^2 goes to zero.

In our forthcoming notations $\eta_1 = \sum_{k=0}^{c-1} f(X_k)$, with $c \to \infty$, this may be expressed as

$$\mathrm{var}_{inv}(\eta_1)/c \sim \sigma^2.$$

We shall remember this.

We will prove the assertion using S. Bernstein's "windows and corridors", as earlier. *For simplicity,* assume f bounded. [(Ibragimov's conjecture relates to ϕ-mixing and $f \in L_2$.)]

"Corridors and windows" by S. Bernstein

Split the (growing as $n \to \infty$) interval $[0, n]$ by larger and smaller partitions, e.g., as follows: take $k := [\frac{n}{[n^{3/4}]}]$ (the total number of long "corridors" of equal length, which (the length) will be chosen in a minute: in any case, it will not exceed $n^{3/4}$ and will be equivalent to that function); $w := [n^{1/5}]$ (the length of short "windows" which separate all consequent corridors); now $c := [\frac{n}{k}] - w = [\frac{n}{k}] - [n^{1/5}]$ (the length of each corridor, except the last one which has the complementary length $n - k[\frac{n}{k}] \le k$).

Notice that $k \sim n^{1/4}$ as $n \to \infty$ and that the total length of all windows is equivalent to $n^{9/20}$, which satisfies $n^{9/20} << n^{1/2}$; that $c \sim n^{1/4}$, and that the last corridor's length does not exceed k and, hence, asymptotically does not exceed $n^{1/4}$.

Denote all partial sums $\sum f(X_s)$ over first k corridors as η_j, $1 \le j \le k$. Notice that

$$\frac{1}{\sqrt{n}}(\sum_{s=1}^{n} f(X_s) - \sum_{j=1}^{k} \eta_j) \sim 0, \quad n \to \infty.$$

Hence, it remains to evaluate (for a NON-iid case)

$$E \exp(i\lambda \sum_{j=1}^{k} \eta_j), \quad n \to \infty.$$

We will do this by induction, using on its each step the exponential bound of the Theorem 5.

Notice that (more accurately, one can write r.h.s.$\wedge 1$)

$$\begin{aligned}|E(\exp(i\lambda\eta_j) \mid \mathcal{F}^X_{(j-1)[n/k]}) - E_{inv} \exp(i\lambda\eta_j)| \\ \le C \exp(\epsilon|X_{(j-1)[n/k]}|) \exp(-cn^{1/5}).\end{aligned}$$

Naturally,

$$\begin{aligned}E_{inv}|E_{inv}(\exp(i\lambda\eta_j) \mid \mathcal{F}^X_{(j-1)[n/k]}) - E_{inv} \exp(i\lambda\eta_j)| \\ \le C E_{inv} \exp(\epsilon|X_{(j-1)[n/k]}|) \exp(-cn^{1/5}) \\ \le C' \exp(-cn^{1/5}),\end{aligned}$$

the latter due to the a priori bound of Theorem 3.

Hence, with an *additive* error that does not exceed $C'\exp(-cn^{1/5})$, one can replace the last r.v. η_n in the expression

$$E_{inv}\exp(i\lambda(\eta_1+\ldots+\eta_k))$$

by some *independent and identically distributed r.v.*, say, $\tilde{\eta}_k$.

This error may be regarded as additive, because

$$|\exp(i\lambda(\eta_1+\ldots+\eta_{k-1}))| = 1.$$

Repeating this procedure, we can replace all r.v.'s η_k in this expression by independent ones with the same distribution (w.r.t. the invariant measure), say, $\tilde{\eta}_k$'s. All errors are additive. So, in the end we get by induction,

$$\begin{aligned} &|E_{inv}\exp(i\lambda(\eta_1+\ldots+\eta_k)) - (E_{inv}\exp(i\lambda(\tilde{\eta}_1))^k| \\ &\qquad \le C'\,k\,\exp(-cn^{1/5}). \end{aligned} \tag{59}$$

(On a few next pages there is some illustration of this kind of calculus. I am not sure if it really helps, i.e., probably it should be better dropped. I leave it, just in case.)

E.g., for two last r.v.'s,

$$\begin{aligned} &\Big|E_{inv}(\exp(i\lambda(\eta_{j-1}+\eta_j)) \mid \mathcal{F}^X_{(j-2)[n/k]}) \\ &\quad -(E_{inv}\exp(i\lambda\eta_{j-1}))(E_{inv}\exp(i\lambda\eta_j)) \\ &\mp(E_{inv}\exp(i\lambda\eta_j))(E_{inv}(\exp(i\lambda\eta_{j-1}) \mid \mathcal{F}^X_{(j-2)[n/k]}))\Big| \\ &\le \Big|E_{inv}(\exp(i\lambda(\eta_{j-1}+\eta_j)) \mid \mathcal{F}^X_{(j-2)[n/k]}) \\ &\quad -(E_{inv}\exp(i\lambda\eta_j))(E(\exp(i\lambda\eta_{j-1}) \mid \mathcal{F}^X_{(j-2)[n/k]}))\Big| \\ &+\Big|(E_{inv}\exp(i\lambda\eta_j))(E(\exp(i\lambda\eta_{j-1}) \mid \mathcal{F}^X_{(j-2)[n/k]})) \\ &\quad -(E_{inv}\exp(i\lambda\eta_{j-1}))(E_{inv}\exp(i\lambda\eta_j))\Big|. \end{aligned}$$

Consider the second term here,

$$\begin{aligned}\Big|(E_{inv}\exp(i\lambda\eta_j))(E_{inv}(\exp(i\lambda\eta_{j-1}) \mid \mathcal{F}^X_{(j-2)[n/k]}))&\\ -(E_{inv}\exp(i\lambda\eta_{j-1}))(E_{inv}\exp(i\lambda\eta_j))\Big|\,.&\\ \le C\exp(\epsilon|X_{(j-2)[n/k]}|)\exp(-cn^{1/5}).&\end{aligned}$$

Consider the first term (rewriting it in an equivalent form),

$$\begin{aligned}\Big|E_{inv}(\exp(i\lambda(\eta_{j-1})\,E_{inv}(\exp(i\lambda(\eta_j)) \mid \mathcal{F}^X_{(j-1)[n/k]}) \mid \mathcal{F}^X_{(j-2)[n/k]})&\\ -(E_{inv}\exp(i\lambda\eta_j))(E_{inv}(\exp(i\lambda\eta_{j-1}) \mid \mathcal{F}^X_{(j-2)[n/k]}))\Big|&\\ = \Big|E_{inv}(\exp(i\lambda(\eta_{j-1})\,(E_{inv}(\exp(i\lambda(\eta_j)) \mid \mathcal{F}^X_{(j-1)[n/k]})&\\ -E_{inv}\exp(i\lambda\eta_j)) \mid \mathcal{F}^X_{(j-2)[n/k]}))\Big|\,.&\end{aligned}$$

Here we estimate,

$$\begin{aligned}\Big|E_{inv}(\exp(i\lambda(\eta_{j-1})\,(E_{inv}(\exp(i\lambda(\eta_j)) \mid \mathcal{F}^X_{(j-1)[n/k]})&\\ -E_{inv}\exp(i\lambda\eta_j)) \mid \mathcal{F}^X_{(j-2)[n/k]}))\Big|&\\ \le E_{inv}\,(|\exp(i\lambda(\eta_{j-1}))|&\\ \times|E_{inv}(\exp(i\lambda(\eta_j)) \mid \mathcal{F}^X_{(j-1)[n/k]}) - E_{inv}\exp(i\lambda\eta_j) \mid \mathcal{F}^X_{(j-2)[n/k]}|\Big)&\\ \le E_{inv}C\exp(\epsilon|X_{(j-1)[n/k]}|)\exp(-cn^{1/5}) \le C'\exp(-cn^{1/5}).&\end{aligned}$$

Return to evaluating the characteristic function of the normed sum, $(\eta_1 + \ldots + \eta_k)/\sqrt{n}$. As in the finite state space, using (59), we may now conclude that

$$\begin{aligned}&E\exp(i\frac{\lambda}{n^{1/2}}\sum_{j=1}^{k}\eta_j)\\ &= \left(E_{inv}\exp(i\frac{\lambda}{n^{1/2}}\eta_1))\right)^k + O(k\exp(-cn^{1/5})).\end{aligned}$$

But it can be seen that (remind that $c \sim n/k$)

$$\begin{aligned}E_{inv}\exp(i\frac{\lambda}{n^{1/2}}\eta_1) &\approx 1 + i\frac{\lambda}{n^{1/2}}\times 0 - \frac{\lambda^2}{n}\,\frac{n}{k}\,\frac{1}{c}\,E_{inv}\eta_1^2 + o(1/k)\\ &\approx 1 - \frac{\lambda^2}{n}\,\frac{n}{k}\sigma^2 + o(1/k).\end{aligned}$$

From this we get, as n (and k) tend to infinity, that

$$\left(E_{inv}\exp(i\frac{\lambda}{n^{1/2}}\eta_1))\right)^k \approx \left(1-\frac{\lambda^2\sigma^2}{2k}\right)^k \to \exp(-\lambda^2\sigma^2/2),$$

as required. The "non-compact" CLT is proved.

17 Parameters in non-compact case

17.1 Coupling by using only integration

See av04.pdf, the section 2.4 for a presentation of coupling method via "normal" integration and just a minimum of probability. The calculus will not be repeated in this file.

17.2 Invariant measures with parameters

See av04.pdf, the section 4.1 - Assumptions. In our previous file(s), we, indeed, explained why all the assumptions in that section are valid, except that (A3_i) and (A4_i) must be *assumed*.

The main result here is the Theorem 7 from av04.pdf about $p_\infty(x,\cdot) \in C^i$ (here $\cdot$ stands for parameters).

For the proof, we firstly check (27)–(31). The main tool is Chapman–Kolmogorov's equation.

Secondly etc., we follow the steps of the proof in av04.pdf. They resemble our analysis in the finite state space case.

Further reading – for Markov diffusions – is three papers by [E.Pardoux and A.V. in AP, 2001, 2003, 2003].