DISCRETE-TIME MARKOV CHAIN

Stationary Distribution: Mean Ergodic Theorem

Tanishq Prasad

July 11, 2025

Contents

1 Intuition: What does the theorem say? **1**
1.1 Time and Space Averages 1
1.2 The general notion: Hilbert space version 2

2 The Mean Ergodic Theorem **4**
2.1 Beyond Boundedness: Birkhoff's Ergodic Theorem 8

3 Applications **9**

1 Intuition: What does the theorem say?

We shall digress from discrete time Markov chains for a moment to gain a deeper insight into the statement of the Mean Ergodic Theorem.

1.1 Time and Space Averages

Consider an evolving system; a stochastic process $\{X_i\}_{i\in\mathbb{N}}$, and a (nice enough) function f of it which you are interested in. The values $\{f(X_i)\}_{i\in\mathbb{N}}$ denote the **time evolution** of the process, hence their running average denotes the **time average** of the process.
Alternatively, at any time point, consider the set of values which the function can take; i.e., consider $\{f(X_i(\omega)) : \omega \in \Omega\}$. This represents the **space** of values which

can be taken; an average over this under some particular law represents a **space average**. One can even consider this as an average of a "freeze-frame" across several parallel, identical systems, at the same time point.
An important and interesting question, which might arise naturally at this point, is

Are these two averages related? Can we comment on space averages without having the knowledge of an underlying stationary law by simply considering the time evolution?

Why is this question important? A result linking the two averages could justify the use of long-run data from one process to estimate statistical properties under a stationary distribution. The corresponding results for i.i.d. processes are well known as laws of large numbers, but here the distributional structure is different.
As an example, suppose you are a Madridista, who is constantly being haunted by *La Masia* (FC Barcelona's Academy) forwards. Their latest product, Lamine Yamal, seems to be a nightmare. You are thinking for the future; you want to calculate the average potential of a *La Masia* forward, so that you can appropriately recommend Real Madrid to improve their own academy, or make a Galactico signing once they see a forward on the rise. However, here comes the problem: you do not have information on a large number of *La Masia* forwards playing against Real Madrid. This makes you wonder: under some regularity conditions, can observing Lamine's performances against Madrid, over time, give an idea about the general performance of a *La Masia* forward?

1.2 The general notion: Hilbert space version

We will not spend a lot of time on this digression, and therefore this will be stated without proof. The general statement is the following.

Theorem 1.1 (Ergodic Theorem for Hilbert spaces)**.** *Let $\mathcal{H}$ be a Hilbert space, and $T : \mathcal{H} \to \mathcal{H}$ a linear contraction, i.e., $||Tf|| \leq ||f|| \; \forall f \in \mathcal{H}$. Define the Cesaro averages of $f \in \mathcal{H}$ as*

$$A_n f = \frac{1}{n} \sum_{k=0}^{n-1} T^k f.$$

Then, $\|A_n f - Pf\| \to 0$ as $n \to \infty$, where P is the projection operator onto the space of $T-$invariant functions: $\mathcal{M} = \{g \in \mathcal{H} : Tg = g\}$.

What the theorem essentially means is that for an element f, the average behaviour due to action of T in the long-run is simply the part of f which is "stable" with respect to T: the projection.

This can be used to give a sketch of how the theorem looks like in the context of discrete time Markov Chains; consider one with:

- measurable state-space $(\mathcal{X}, \mathcal{B})$ (the space and the σ algebra),
- transition kernel $P : \mathcal{X} \times \mathcal{B} \to [0, 1]$ (i.e. probability of next state being in a certain measurable set, given the current state),
- stationary measure π on the measurable state-space, i.e

$$\pi P(A) = \int_{\mathcal{X}} P(x, A)\, d\pi(x),\ \forall A \in \mathcal{B}.$$

Note: Think stationary transition ~~kernel~~ matrix, π stationary ~~measure~~ distribution, for countable state space in case you face a difficulty with these terms!

With the above setup in mind, consider the Hilbert space

$$\mathcal{H} = \mathcal{L}^2(\pi) = \{f : \mathcal{X} \to \mathbb{R} : \int_{\mathcal{X}} f^2\, d\pi < \infty\},$$

(or if you have difficulty with the general notion of the integral, you can think of the set $\{f : \mathbb{E}f^2(X) < \infty, X \sim \pi\}$) with usual inner product

$$\langle f, g\rangle = \int_{\mathcal{X}} f(x)g(x)\, d\pi(x).$$

Consider the **Koopman/Markov operator**:

$$\mathcal{P} : \mathcal{L}^2(\pi) \to \mathcal{L}^2(\pi) \text{ such that } \mathcal{P}(f(x)) = \int_{\mathcal{X}} f(y)P(x, dy) = \mathbb{E}[f(X_{n+1})|X_n = x].$$

It can be checked that $\mathcal{P}$ is a contraction, and that the fixed points of $\mathcal{P}$ are the constant functions **under ergodicity**. Thus,

$$\left\| \frac{1}{n} \sum_{i=0}^{n-1} \mathcal{P}^i(f) - \mathbb{E}_\pi[f(X)] \right\| \to 0.$$

Define the *Cesaro average* of the sequence as $A_n = \frac{1}{n}\sum_{i=0}^{n-1} f(X_i)$. Consider its second moment:

$$\begin{aligned}
\mathbb{E}[A_n^2] &= \frac{1}{n^2}\sum_{i,j=0}^{n-1}\mathbb{E}[f(X_i)f(X_j)] \\
&= \frac{1}{n^2}\sum_{i,j=0}^{n-1}\mathbb{E}[f(X_0)f(X_{|i-j|})] \text{ [using stationarity]} \\
&= \frac{1}{n^2}\sum_{i,j=0}^{n-1}\langle f, \mathcal{P}^{|i-j|}f\rangle \text{ [condition on } f(X_0)].
\end{aligned}$$

Note that this converges to μ^2 using the previous theorem, as

$$\begin{aligned}
\left\|\frac{1}{n}\sum_{k=0}^{n-1}\mathcal{P}^k f - \mu\right\|^2 &= \langle \frac{1}{n}\sum_{k=0}^{n-1}\mathcal{P}^k(f-\mu), \frac{1}{n}\sum_{k=0}^{n-1}\mathcal{P}^k(f-\mu)\rangle \\
&= \frac{1}{n^2}\sum_{i,j=0}^{n-1}\langle \mathcal{P}^i(f-\mu), \mathcal{P}^j(f-\mu)\rangle \\
&= \frac{1}{n^2}\sum_{i,j=0}^{n-1}\langle (f-\mu), \mathcal{P}^{|i-j|}(f-\mu)\rangle \text{ [stationarity]} \\
&= A_n^2 - \mu^2.
\end{aligned}$$

Thus, by Chebyshev's inequality, $\mathbf{A_n} \xrightarrow{\mathbb{P}} \mathbb{E}_\pi[\mathbf{f}(\mathbf{X})]$.

2 The Mean Ergodic Theorem

We are now ready to look at the theorem and its proof in the context of discrete time Markov chains. Previously, we used Hilbert space formulation to prove in probability convergence; here we shall prove almost sure convergence. A detailed discussion can be found in [1]

Definition 2.1 (Ergodicity)**.** *An irreducible and positive recurrent DTMC is said to be an* ***ergodic*** *DTMC.*

Theorem 2.1 (Mean Ergodic Theorem for Bounded functions in a DTMC)**.** *Let $\{X_n\}_{n\geq 0}$ be a discrete time, discrete state space Markov chain with irreducible transition matrix $P(x,y)$. Let λ be any initial distribution for the chain. Then for any state i,*

$$\frac{1}{n}\sum_{k=0}^{n-1}\mathbb{I}(X_k = i) \overset{a.e.}{\to} \frac{1}{m_i},$$

where $m_i = \mathbb{E}_i[T_i]$, the expected time of return to state i.
Further, if the chain is positive recurrent (hence ergodic), then for any bounded, measurable function f,

$$\frac{1}{n}\sum_{i=0}^{n-1} f(X_i) \overset{a.e.}{\to} \mathbb{E}_\pi[f(X)],$$

where π is the unique stationary distribution.

Proof. **Case 1: Transient Chain**. In this case, consider $N_i(n) = \sum_{j=0}^{n-1}\mathbb{I}(X_j = i)$, the total visits to i till n steps (including 0). It is an increasing sequence in n with

$$N_i(n) \overset{ptwise}{\to} \sum_{j\geq 0}\mathbb{I}(X_j = i) =: N_i,$$

and,

$$\begin{aligned}
\mathbb{E}_k[N_i(n)] &= \sum_{j=0}^{n-1} p_{ki}^{(j)} \\
&= \sum_{j=0}^{n-1}\sum_{m=1}^{j} f_{ki}^{(m)} p_{ii}^{(j-m)} \\
&= \sum_{r=0}^{n-1}\sum_{m=0}^{n-r-1} f_{ki}^{(m)} p_{ii}^{(r)} \text{ [re-index r=j-m]} \\
&\leq \sum_{r=0}^{n-1} p_{ii}^{(r)} \text{ [as } \sum_m f_{ki}^{(m)} \leq 1].
\end{aligned}$$

Taking limit on both sides yields

$$\mathbb{E}_k[N_i] \leq \sum_{r\geq 0} p_{ii}^{(r)} < \infty \text{ [transient]}.$$

where the LHS is written as $\sum_{j\geq 0}\mathbb{P}(X_j = i)$ using the monotone convergence theorem. Using this, we can now see that:

$$\mathbb{E}[N_i] = \sum_{k\in\mathcal{X}} \mathbb{E}_k[N_i]\lambda(k) < \infty \text{ [finite convex combination]},$$

and hence $\mathbb{P}(N_i < \infty) = 1$. Thus, $\frac{N_i}{n} \overset{a.e.}{\to} 0$, and the first part holds as for transient states $m_i = \infty$.

Case 2: Recurrent Chain. Let $T_j^r \coloneqq \min\left\{i > 0 : \sum_{k=1}^{i} \mathbb{I}(X_k = j) = r\right\}$, i.e., time of the rth visit to state j. Let $w_j^1 \coloneqq T_j^1$ and $w_j^n \coloneqq T_j^{n+1} - T_j^n$ for $n > 1$, i.e., the waiting times between consecutive visits. Note that $w_j^n | X_0 = j$ are *iid* with mean m_j. Thus, by the strong law of large numbers,

$$\frac{T_j^k}{k} = \frac{1}{k}\sum_{n=1}^{k} w_j^n \overset{a.e.}{\to} m_j.$$

Now, suppose $N_j(n) = r$, i.e., at n steps, r many visits have been made; rth visit is before n and the next is after n. Then

$$T_j^r \leq n \leq T_j^{r+1}.$$

Dividing by r,

$$\frac{T_j^r}{r} \leq \frac{n}{N_j(n)} \leq \frac{T_j^{r+1}}{r}.$$

Both sides converge a.e. to m_j; hence using sandwich lemma for real sequences,

$$\frac{N_j(n)}{n} \overset{a.e.}{\to} \frac{1}{m_j}$$

which completes the first part of the theorem.

Assume now that the chain is positive recurrent. Then it has a unique stationary distribution $\pi = \{\pi_i\}_{i\in\mathcal{X}} = \{\frac{1}{m_i}\}_{i\in\mathcal{X}}$. Let f be a bounded, real valued measurable function Also assume without loss of generality $|f| \leq 1$. Let $\mu \coloneqq \mathbb{E}_\pi[f(X)]$. Note

that for any $\mathcal{X}' \subseteq \mathcal{X}$,

$$\begin{aligned}
\left|\frac{1}{n}\sum_{k=0}^{n-1} f(X_k) - \mu\right| &= \left|\sum_{j\in\mathcal{X}}\left(\frac{N_j(n)}{n} - \pi_j\right) f(j)\right| \\
&\leq \sum_{j\in\mathcal{X}'}\left|\left(\frac{N_j(n)}{n} - \pi_j\right)\right| + \sum_{j\notin\mathcal{X}'}\left|\left(\frac{N_j(n)}{n} - \pi_j\right)\right| \\
&\leq \sum_{j\in\mathcal{X}'}\left|\left(\frac{N_j(n)}{n} - \pi_j\right)\right| + \sum_{j\notin\mathcal{X}'}\left(\frac{N_j(n)}{n} + \pi_j\right) \\
&= \sum_{j\in\mathcal{X}'}\left|\left(\frac{N_j(n)}{n} - \pi_j\right)\right| + \sum_{j\notin\mathcal{X}'}\left(\frac{N_j(n)}{n} - \pi_j + 2\pi_j\right) \\
&\leq 2\sum_{j\in\mathcal{X}'}\left|\left(\frac{N_j(n)}{n} - \pi_j\right)\right| + 2\sum_{j\notin\mathcal{X}'}\pi_j.
\end{aligned}$$

Let $\varepsilon > 0$. Since the choice of $\mathcal{X}'$ is in arbitrary, choose a finite $\mathcal{X}'$ such that

$$\sum_{j\notin\mathcal{X}'}\pi_j \leq \frac{\varepsilon}{4}.$$

Also, from the first part, we know that

$$\mathbb{P}\left(\omega : \forall\eta > 0, \exists\, N(\eta) \in \mathbb{N},\ \text{s.t.}\ \left|\sum_{j\in\mathcal{X}'}\left(\frac{N_j(n)(\omega)}{n} - \pi_j\right)\right| < \eta\ \forall n > N(\eta)\right) = 1.$$

Hence, $\forall\eta > 0$, $\exists\, N(\eta)$ such that

$$\mathbb{P}\left(\omega : \left|\sum_{j\in\mathcal{X}'}\left(\frac{N_j(n)(\omega)}{n} - \pi_j\right)\right| < \eta\right) = 1\ \ \forall n > N(\eta).$$

Thus,

$$\mathbb{P}\left(\omega : \left|\frac{1}{n}\sum_{k=0}^{n-1} f(X_k(\omega) - \mu\right| < \varepsilon\right) \geq \mathbb{P}\left(\omega : \left|\sum_{j\in\mathcal{X}'}\left(\frac{N_j(n)(\omega)}{n} - \pi_j\right)\right| < \frac{\varepsilon}{4}\right).$$

Thus, $\forall n > N\left(\frac{\varepsilon}{4}\right)$,

$$\mathbb{P}\left(\omega : \left|\frac{1}{n}\sum_{k=0}^{n-1} f(X_k(\omega) - \mu\right| < \varepsilon\right) = 1.$$

Thus, we have

$$\frac{\mathbf{1}}{\mathbf{n}}\sum_{\mathbf{k=0}}^{\mathbf{n-1}}\mathbf{f}(\mathbf{X_k}) \overset{\mathbf{a.e.}}{\to} \mathbb{E}_\pi[\mathbf{f}(\mathbf{X})].$$

□

2.1 Beyond Boundedness: Birkhoff's Ergodic Theorem

We had previously proved in probability convergence (using strong Hilbert space results) but we did not assume boundedness. In the previous section, we proved almost sure convergence but assumed boundedness. **Birkhoff's ergodic theorem** is a stronger result, which extends it to the weaker class of integrable functions instead of bounded functions. However, it also requires a stronger assumption that the chain starts in its stationary distribution. The actual statement is measure theoretic and beyond the scope of this discussion. One can also, however, try extending to integrable functions using the previous result; we will look at an argument sketch (which unfortunately could not be completed, so you are free to try).

Sketch of Argument. *For an integrable function f and $M \in \mathbb{N}$, define the M-truncation as $f_M(x) \coloneqq f(x)\mathbb{I}(|f(x)| \leq M)$ and the error term as $h_M(x) \coloneqq |f - f_M|(x)$. Note that,*

$$\begin{aligned}\left|\frac{1}{n}\sum_{k=0}^{n-1} f(X_k) - \mathbb{E}_\pi[f(X)]\right| &\leq \left|\frac{1}{n}\sum_{k=0}^{n-1} f_M(X_k) - \mathbb{E}_\pi[f_M(x)]\right| \\ &+ |\mathbb{E}_\pi[f_M(x)] - \mathbb{E}_\pi[f(X)]| \\ &+ \frac{1}{n}\sum_{k=0}^{n-1} |f - f_M|(X_k).\end{aligned}$$

The first term converges a.e. to 0 as $n \to \infty$ using the mean ergodic theorem for bounded functions. The second converges to 0 as $M \to \infty$ using monotone convergence theorem. It remains to be shown that the third term also "vanishes" in some sense. The idea is to use truncation again to show that

$$\frac{1}{n}\sum_{k=0}^{n-1} h_M(X_k) \overset{a.e.}{\to} \mathbb{E}_\pi[h_M(X)],$$

and then an application of the dominated convergence theorem would show that taking

$M \to \infty$ would lead to the 0 limit. Define $h_{M,N}(x) \coloneqq h_M(x)\mathbb{I}(x \leq N)$. Then,

$$\begin{aligned} \liminf_n \frac{1}{n}\sum_{k=0}^{n-1} h_M(X_k) &\geq \lim_n \frac{1}{n}\sum_{k=0}^{n-1} h_{M,N}(X_k) \\ &\overset{a.e.}{=} \mathbb{E}_\pi[h_{M,N}(X)] \text{ [M.E.T. for bdd fns].} \end{aligned}$$

and hence taking $N \to \infty$,

$$\liminf_n \frac{1}{n}\sum_{k=0}^{n-1} h_M(X_k) \geq \mathbb{E}_\pi[h_M(X)] \text{ a.e. [M.C.T.].}$$

One can also show that

$$\limsup_n \frac{1}{n}\sum_{k=0}^{n-1} h_M(X_k) \leq \mathbb{E}_\pi[h_M(X)] \text{ a.e.}$$

using similar simple arguments.

See [2] for a detailed discussion on Birkhoff's ergodic theorem.

Note (aperiodicity in ergodicity): In the definition of ergodicity we used above, we just needed irreducibility and positive recurrence to go through the proofs. However, talking about the mean under the stationary distribution would make sense in that setting if the chain ever attained the stationary distribution. This would certainly be the case if the chain started with the stationary distribution. But what if the chain started with an arbitrary distribution? The convergence would still hold, but talking of the mean under the stationary distribution would not be very natural. Thus, more often than not, aperiodicity is included in the definition of ergodicity. This is because aperiodicity would guarantee that starting from an arbitrary distribution, the limiting distribution would be the unique stationary distribution, and thus talking about the mean under that would be somewhat more sensible.
Aperiodicity, however, becomes more important when using Birkhoff's ergodic theorem to prove the convergence for integrable functions, since measure theoretic ergodicity is sufficiently guaranteed by irreducibility, positive recurrence and aperiodicity.

3 Applications

Below are some direct consequences of the mean ergodic theorem.

1. **Proportion of time in a set** Let $A \subset \mathcal{X}$ be a measurable set. Then, the long run proportion of time spent in the set would be strongly consistent for the stationary probability of being in that set, i.e.,

$$\frac{1}{n}\sum_{k=0}^{n-1} \mathbb{I}(X_k \in A) \overset{a.e.}{\to} \mathbb{P}_\pi(A).$$

   This also means that the empirical distribution converges pointwise.

2. **Return time estimation** Let $f_i(x) = \mathbb{I}(x = i)$ for some state i. Then,

$$\frac{1}{n}\sum_{k=0}^{n-1} f_i(X_k) \overset{a.e.}{\to} \frac{1}{\mathbb{E}_i[T_i]},$$

   and therefore, the inverse can be used as an estimate for the return time.

3. **Estimating the transition matrix** For a finite state space, consider the new process $Y_k = (X_k, X_{k+1}) \subset \mathcal{X}^2$ (for $k \geq 0$). Check that this is Markov, ergodic and stationary! Define $f_{i,j}(x, y) = \mathbb{I}(x = i, y = j)$. Let $N_{i,j}(n) = \sum_{k=0}^{n-1} f_{i,j}(Y_k)$. Then,

$$\frac{1}{n} N_{i,j}(n) \overset{a.e.}{\to} \pi(x)P(x, y).$$

   We already have an estimate for $\pi(x)$ in the following sense.

$$\frac{1}{n}\sum_{k=0}^{n-1} \mathbb{I}(X_k = x) \overset{a.e.}{\to} \pi(x).$$

   The ratio of these two estimates is a strongly consistent estimate for the transition matrix.

References

[1] J.R. Norris. *Markov Chains*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 1998. ISBN: 9780521633963. URL: https://books.google.co.in/books?id=qM65VRmOJZAC.

[2] E.M. Stein and R. Shakarchi. *Real Analysis: Measure Theory, Integration, and Hilbert Spaces*. Vol. 1. Princeton Lectures in Analysis. Princeton, NJ: Princeton University Press, 2005. ISBN: 9780691113869. URL: https://books.google.co.in/books?id=2Sg3Vug65AsC&redir_esc=y.