Ergodic Behavior of Markov Process

Abstract

セミナーのまとめ。"Ergodic Behavior of Markov Process" Alexei Kulik

1 Recall : total variation distance

Let's consider Markov Chains.Let E be a state space. E is a polish space like $\mathbb{R}$. Let be $\mathcal{F}$ a σ-alg. For example, $\mathcal{B}(\mathbb{R}^d)$. Denote $P(x, dy)$ as a Markov Kernel on $(E, \mathcal{F})$. This means that the probability of the transition form x to dy.[1] These are the propaties of Markov Kernel:

1. $\forall x \in E, P(x, \cdot)$ is a probability measure on $\mathcal{F}$.
2. $\forall A \in \mathcal{F}, P(\cdot, A)$ is a measurable function on E.

By the 2nd property,

- $\forall f$: bounded measurable, $\displaystyle\int_{y\in E} f(y)P(\cdot, dy)$ is measurable

Let $(X_n)_n$ be a Markov Process on E, time homogenious. Assume that for every n,

$$X_n \sim P^n(x, \cdot)$$

which means that if it starts on x at every n, $X_{n+1} \sim P(X_n, \cdot)$ holds. P^n denotes the n-step transition. For example,

$$P^2(x, dy) = \int_{z\in E} P(x, dz)P(z, dy) \tag{1}$$

$$P^n(x, dy) = \int_{x_1,\cdots,x_{n-1}\in E} P(x, dx_1)P(x_1, dx_2)\cdots P(x_{n-1}, dy) \tag{2}$$

and so on.

Our Objective is to study the convergence of (X_n) in tatal variation.

[1] Not y but dy because the state space is countinuous.

1.1 Total Variation

Let $\mathcal{M}(E) = \{\text{signed measures on } E\}$. *i.e.*

$$\mu \in \mathcal{M}(E) \Leftrightarrow \mu = \mu_1 - \mu_2$$

where μ_1, μ_2 are finite positive measures.

Theorem 1.1 (Jordan-Hahn decomposition of a signed-measure)**.**

$$\mu = \mu_+ - \mu_- \quad \text{and} \quad \mu_+ \perp \mu_-$$

The latter says that there exists $A \in \mathcal{F}$ such that $\mu_+(A) = \mu(E), \mu_-(A) = 0$

For the proof, see a measure theory textbook.

Definition 1.2. $||\cdot||_{TV}$ is a total variation for a measure μ, if for any $\mu \in \mathcal{M}(E)$, $\mu = \mu_+ - \mu_-$,

$$||\mu||_{TV} := \mu_+(E) + \mu_-(E)$$

It is clear that if $\mu > 0$, then $||\mu||_{TV} = \mu(E)$

For more detailed theory for tatal variation, you should check a textbook about measure theory, too.

Definition 1.3. For any $\mu, \nu \in \mathcal{M}(E)$, $d_{TV} = ||\mu - \nu||_{TV}$ is a distance and we call d_{TV} "total variation distance".

We can characterize the total varation norm like functional analisys method...?

Proposition 1.4.

$$||\mu||_{TV} = \sup_{\substack{||f||_\infty = 1 \\ f: E \to \mathbb{R} \\ f \text{ is measurable}}} \left| \int f(x)\mu(dx) \right|$$

証明. If $||f||_\infty = 1$,

$$\begin{aligned}
\left| \int f(x)\mu(dx) \right| &\leq ||f||_\infty \left| \int \mu(dx) \right| \\
&\leq 1 \times \left(\left| \int_{A_+} \mu(dx) \right| + \left| \int_{A_-} \mu(dx) \right| \right) \\
&\leq \Big(\mu(A_+) + \mu(A_-) \Big) \\
&= ||\mu||_{TV}
\end{aligned}$$

where $A_+ = \mathrm{supp}(\mu_+), A_- = A_+^c$. [2]

Take $f = 1_{A_+} - 1_{A_-}$, then $||f||_\infty = 1$ and

$$\begin{aligned}\left|\int f(x)\mu(dx)\right| &= \left|\int_{A_+} f(x)\mu(dx) + \int_{A_-} f(x)\mu(dx)\right| \\ &= \left|\int_{A_+} 1_X\mu(dx) - \int_{A_-} 1_X(-\mu_-(dx))\right| \\ &= ||\mu||_{TV}\end{aligned}$$

So we have even proved that the sup is a max. □

Proposition 1.5. Let μ, ν be probability measures. Then

$$||\mu - \nu||_{TV} = 2\max_{A\in\mathcal{F}}(\mu(A) - \nu(A))$$

証明. Firstly,

$$\mu - \nu = (\mu - \nu)_+ - (\mu - \nu)_-$$

so,

$$||\mu - \nu||_{TV} = (\mu - \nu)_+(E) + (\mu - \nu)_-(E)$$

and now for all $A \in \mathcal{F}$,

$$\begin{aligned}\mu(A) - \nu(A) &= (\mu - \nu)_+(A) - (\mu - \nu)_-(A) \\ &\le (\mu - \nu)_+(E) + (\mu - \nu)_-(E) \\ &= ||\mu - \nu||_{TV}\end{aligned}$$

Take A as $A_+ = \mathrm{supp}((\mu - \nu)_+)$. Then

$$\mu(A_+) - \nu(A_+) = (\mu - \nu)_+(A_+) - (\mu - \nu)_-(A_+) = (\mu - \nu)_+(E) + 0$$

,and put $A_- = A_+^c$, we can do similary

$$\mu(A_-) - \nu(A_-) = -(\mu - \nu)_-(E)$$

,finally sum up these two formulas

$$\begin{aligned}\mu(A_+) - \nu(A_+) + \mu(A_+) - \nu(A_+) &= (\mu - \nu)_+(E) + 1 - \mu(A_-) - (1 - \nu(A_-)) \\ &= (\mu - \nu)_+(E) - (\mu - \nu)_-(E) \\ &= ||\mu - \nu||_{TV}\end{aligned}$$

□

[2] Let (X, T) be a topology space. Consider $\mathcal{B}(X)$. For any measure μ on X, $\mathrm{supp}(\mu)$ $:= \{x \in X \mid x \in N_x \in T \Rightarrow \mu(N_x) > 0\}$ where N_x is a open neiborhood of x.

Let's see another property of total variation. Next one is related to the Radon-Nikodym theorem.

Proposition 1.6. Let λ be a measure on E such that $\mu \ll \lambda$ and $\mu \ll \lambda \quad (\mu, \nu \in \mathcal{P}(E))$[3] [4]. And put

$$A_+ := \left\{ x;\ \frac{d\mu}{d\lambda} \geq \frac{d\nu}{d\lambda} \right\}, \quad A_- := \left\{ x;\ \frac{d\mu}{d\lambda} < \frac{d\nu}{d\lambda} \right\}$$

Then

$$||\mu - \nu||_{TV} = \int \left| \frac{d\mu}{d\lambda} - \frac{d\nu}{d\lambda} \right| d\lambda = \mathbf{E}_\lambda \left| \frac{d\mu}{d\lambda} - \frac{d\nu}{d\lambda} \right|$$

and the function in the expectatio is L^1-distance.[5]

証明. the presenter didn't prove this in the seminar. Sadly, I need to show by myself... □

1.2 Coupling Lemma

Next Theorem is a very important and useful tool to deal with a Markov Chain. The distance of two measures, like between a n-step transition kernel and an IPM, is able to calculate by the following theorem in the sense of total variation. Its name is "Coupling lemma". It will be used when proving the propositions in the following sections, which are about the speed of convergence of a Markov Chain.

Lemma 1.7 (Coupling Lemma)**.** Let $\mu, \nu \in \mathcal{P}(E)$. Then

$$||\mu - \nu||_{TV} = 2 \min_{\kappa \in \mathcal{C}(\mu,\nu)} \kappa(\{(x, y)\ ;\ x \neq y\})$$

where κ is a probability measure on $E \times E$ with marginals μ and ν.

Remark 1.8. According to Wasserotein,

$$W_\gamma(\mu, \nu) = \left(\inf_{\kappa \in \mathcal{C}(\mu,\nu)} \int ||x - y||^r \kappa(dx, dy) \right)^{\frac{1}{\gamma}}$$

...What is W??????

[3] $\mathcal{P}(E)$ denotes the set of probability measures on E.
[4] For example, let $\lambda := \frac{1}{2}(\mu + \nu)$. Then $\frac{d\mu}{d\lambda}, \frac{d\nu}{d\lambda}$ exist.
[5] L^1-distance? Clarify the meaning of the statement.

証明. Let ξ, η be random variables such that $\xi \sim \mu$ and $\eta \sim \nu$. By the previous propositions,

$$\begin{aligned} ||\mu - \nu||_{TV} &= 2 \max_{A \in \mathcal{F}} (\mu(A) - \nu(A)) \\ &= 2 \max_{A \in \mathcal{F}} \mathbf{E}(1_{\xi \in A} - 1_{\eta \in A}) \\ &\leq 2 \max_{A \in \mathcal{F}} \mathbf{P}(\xi \in A, \eta \notin A) \\ &\leq 2\mathbf{P}(\xi \neq \eta) \\ &= 2\kappa(\{x \neq y\}) \end{aligned}$$

Next, we construct explicitly the joint law κ which turns inequality above into an equality. Put $\lambda := \frac{1}{2}(\mu + \nu)$, then there exists

$$f = \frac{d\mu}{d\lambda}, \quad g = \frac{d\nu}{d\lambda}$$

and put

$$h := \min(f, g), \quad p := \int h \, d\lambda$$

p is finite because

$$\int h \, d\lambda \leq \int f \, d\lambda = \int \frac{d\mu}{d\lambda} \, d\lambda = 1$$

If $p = 1$, it implies $\mu = \nu$ and we can put $\kappa(A_1 \times A_2) = \mu(A_1 \cap A_2)$. In this case, ξ, η actually equal to one another, with the same law $\mu = \nu$. Otherwise, we decompose

$$\mu = p\theta + (1 - p)\sigma_1, \quad \nu = p\theta + (1 - p)\sigma_2,$$

where θ is a probability measure on E such that $\theta \ll \lambda$ and

$$d\theta = \frac{1}{p} h \, d\lambda.$$

If $p = 0$, $\theta = \lambda$. In fact, σ_1 is defined to be

$$\sigma_1 := \frac{\mu - p\theta}{1 - p}.$$

σ_2 is also defined like this. We should note that we decomposed μ and ν just by a probability measure θ and σ_1, σ_2 are used for only shortening. Define κ a probability measure on $E \times E$ as

$$\kappa(A_1 \times A_2) = p\theta(A_1 \cap A_2) + (1 - p)\sigma_1(A_1)\sigma_2(A_2)$$

We need to check that the marginals of κ is really μ and ν:

$$\begin{aligned}\kappa(A \times E) &= p\theta(A \cap E) + (1-p)\sigma_1(A)\sigma_2(E) \\ &= p\theta(A) + (1-p)\sigma_1(A) \\ &= p\theta(A) + (1-p)\frac{\mu(A) - p\theta(A)}{1-p} \\ &= \mu(A).\end{aligned}$$

Similary, $\kappa(E \times A) = \nu(A)$. On the other hand,

$$\kappa(\{(x,y)\ ;\ x = y\}) \geq p\theta(E) = p,$$

and so

$$\kappa(\{(x,y)\ ;\ x \neq y\}) \leq 1 - p = 1 - \int h\, d\lambda = ||\mu - \nu||_{TV}$$

because[6]

$$\begin{aligned}1 - \int h\, d\lambda &= \int \left(1 - \min\left(\frac{d\mu}{d\lambda}, \frac{d\nu}{d\lambda}\right)\right) d\lambda \\ &= \int_{A_+} \left(1 - \frac{d\mu}{d\lambda}\right) d\lambda + \int_{A_+} \left(1 - \frac{d\mu}{d\lambda}\right) d\lambda \\ &= \cdots \\ &= ||\mu - \nu||_{TV}\end{aligned}$$

This completes the proof □

2 Uniform Ergodicity: The Dobrushin theorem

"uniform" is in the sense that the corresponding bound for the total variation distance between the transition probabilities is uniform w.r.t. initial positions of the chain.

Theorem 2.1 (The Dobrushin Theorem)**.** Assume that $\exists m$ such that :

$$\sup_{x_1, x_2} ||P^m(x_1, \cdot) - P^m(x_2, \cdot)||_{TV} < 2.$$

Then $\exists C, \rho > 0$ such that

$$\sup_{x_1, x_2} ||P^m(x_1, \cdot) - P^m(x_2, \cdot)||_{TV} \leq Ce^{-\rho n}, \quad n \geq 1.$$

[6]Is it really true? You should check.

Moreover, $\exists!\,\mu$ an IPM for X, and

$$\sup_{x_1} ||P^m(x_1,\cdot) - \mu||_{TV} \leq Ce^{-\rho n}, \quad n \geq 1.$$

In other words, for all x, P_x^m converges in d_{TV} and the convergence speed is uniform independent of initial state x.

証明. Using coupling argument.

Assume $m = 1$. take x_1, x_2 and let Q be the optimal coupling kernel for $P(x_1,\cdot)$ and $P(x_2,\cdot)$. "optimal" means that Q is equivalent to κ in the coupling lemma in the last section. Then[7], this is a transition kernel on $E \times E$, and if

$$Y_n := (Y_n^1, Y_n^2) \sim \kappa^n((x_1,x_2),\cdot)$$

then

$$Y_n^1 \sim P^n(x_1,\cdot), \quad Y_n^2 \sim P^n(x_2,\cdot).$$

So, by the coupling lemma and the definition of Y and Q, we get

$$||P^n(x_1,\cdot) - P^n(x_2,\cdot)||_{TV} = 2\mathbf{P}(Y_n^1 \neq Y_n^2)$$

Denote $\mathcal{D} = \{(x,y)\ ;\ x = y\}$ and define[8]

$$q(z) := Q(z,\mathcal{D}) \quad (z \in E \times E)$$

so

$$q((x_1,x_2)) = 1 - \frac{1}{2}||P^n(x_1,\cdot) - P^n(x_2,\cdot)||_{TV} = \begin{cases} 1 & (\text{if} \quad x_1 = x_2) \\ \geq 1-\alpha & (\text{otherwise}) \end{cases}$$

where α is a constant probability($0 < \alpha < 1$). Namely, Y_{n+1} stays on $\mathcal{D}$ if $Y_n \in \mathcal{D}$, or goes to $\mathcal{D}$ with probability$\geq (1-\alpha)$ if $Y_n \notin \mathcal{D}$. So, $\mathbf{P}(Y_n^1 \neq Y_n^2) \leq \alpha^n$. Now, we have a "geometric law to reach $\mathcal{D}$". Thus, we can get the inequality in the statement; we put ρ as $-\log(1-\frac{1}{2}\alpha)$

$$\mathbf{P}(Yn \notin \mathcal{D}) \leq \left(1 - \frac{1}{2}\alpha\right)^n \leq Ce^{-\rho n}.$$

Next, we should also to prove that $P^n(x,\cdot) \longrightarrow \mu$. Let $x \in E$. Then, $(P^n(x,\cdot))_n$ is Caucy seq. in $(\mathcal{P}(E), d_{TV})$. So for $m < n$

$$\begin{aligned} ||P^n(x,\cdot) - P^m(x,\cdot)||_{TV} &\leq \left|\left|\int (P^m(x,\cdot) - P^m(x,\cdot))P^{n-m}(x,dy)\right|\right|_{TV} \\ &\leq \int ||(P^m(x,\cdot) - P^m(x,\cdot))||_{TV} P^{n-m}(x,dy) \\ &\leq \int Ce^{-\rho n} P^{n-m}(x,dy) \\ &= Ce^{-\rho n}. \end{aligned}$$

[7] In the textbook, the perocess Y is expressed as Z a "greedy coupling for X".
[8] i.e. $q(Y_n) = \mathbf{P}(Y_{n+1} \in \mathcal{D} \mid Y_n)$

And $(\mathcal{P}(E), d_{TV})$ is a complete space. Therefore there exists μ_x such that

$$P^n(x, \cdot) \longrightarrow \mu_x.$$

But, our inequality impies that

$$||P^n(x_1, \cdot) - P^m(x_2, \cdot)||_{TV} \longrightarrow 0$$

so $\mu_{x_1} = \mu_{x_2}$. We can say that μ_x does not depend on x. This is the μ. Finally, let's prove μ is an IPM.

$$P^{n+1}(x, A) = \int P(y, A) P^n(x, dy),$$

by taking $n \longrightarrow \infty$,

$$\mu(x, A) = \int P(y, A) \mu(x, dy).$$

What if $m \neq 1$? We obtain

$$||P^{nm}(x_1, \cdot) - P^{nm}(x_2, \cdot)|| \leq Ce^{-\rho n}$$

□