Lecture 3: Coupling and Convergence

John Sylvester Nicolás Rivera Luca Zanetti Thomas Sauerwald

Lent 2019

UNIVERSITY OF CAMBRIDGE

Outline

Total Variation Distance

Coupling

Convergence Theorem - Proof and Application

Mixing Times

How Similar are Two Probability Measures?

Loaded Dice

- I present to you three *loaded* (unfair) dice A, B, C:

x	1	2	3	4	5	6
$\mathbf{P}[A = x]$	1/3	1/12	1/12	1/12	1/12	1/3
$\mathbf{P}[B = x]$	1/4	1/8	1/8	1/8	1/8	1/4
$\mathbf{P}[C = x]$	1/6	1/6	1/8	1/8	1/8	9/24

- Question 1 : Which dice is the least fair ?
- Question 2 : Which dice is the most fair ?

Question 1: Most of you choose A. Why?

Question 2: Dice B and C seem "fairer" than A but which is fairest?

Question 3 : What do we mean by "fair"?

Total Variation Distance

The *Total Variation Distance* between two probability distributions μ and η on a countable state space Ω is given by

$$\|\mu - \eta\|_{tv} = \frac{1}{2}\sum_{\omega\in\Omega} |\mu(\omega) - \eta(\omega)|.$$

- Let $d(\mu,\nu) = \|\mu - \nu\|_{tv}$, then $d(\cdot,\cdot)$ is a metric on the space of measures.

Loaded Dice : let D be a fair dice and observe:

$$\|D - A\|_{tv} = \frac{1}{2}\left(2\left|\frac{1}{6} - \frac{1}{3}\right| + 4\left|\frac{1}{6} - \frac{1}{12}\right|\right) = \frac{1}{3}$$

$$\|D - B\|_{tv} = \frac{1}{2}\left(2\left|\frac{1}{6} - \frac{1}{4}\right| + 4\left|\frac{1}{6} - \frac{1}{8}\right|\right) = \frac{1}{6}$$

$$\|D - C\|_{tv} = \frac{1}{2}\left(3\left|\frac{1}{6} - \frac{1}{8}\right| + \left|\frac{1}{6} - \frac{9}{24}\right|\right) = \frac{1}{6}.$$

Thus

$$\|D - B\|_{tv} = \|D - C\|_{tv} \qquad \text{and} \qquad \|D - C\|_{tv}, \|D - C\|_{tv} < \|D - A\|_{tv}.$$

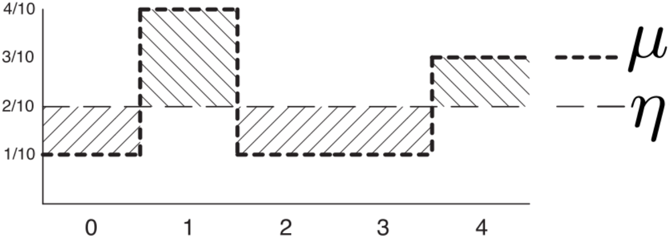
So A is the least “fair” however B and C are equally “fair” (in TV distance).

Lemma

For any probability distributions μ and η on a countable state space Ω

$$\|\mu - \eta\|_{tv} := \frac{1}{2} \sum_{\omega \in \Omega} |\mu(\omega) - \eta(\omega)| = \sup_{A \subset \Omega} |\mu(A) - \eta(A)|.$$

Proof by picture.



□

TV Distances

Let P be a Markov Chain with stationary distribution π.

- Let μ be a prob. vector on $\mathcal{I}$ (might be just one vertex) and $t \geq 0$. Then

$$P_\mu^t := \mathbf{P}_\mu[X_t = \cdot] = \mathbf{P}[\, X_t = \cdot \mid X_0 \sim \mu\,],$$

  is a probability measure on $\mathcal{I}$.

- For any μ,

$$\left\| P_\mu^t - \pi \right\|_{tv} \leq \max_{x \in \mathcal{I}} \left\| P_x^t - \pi \right\|_{tv}.$$

Convergence Theorem (rephrased)

For any finite, irreducible, aperiodic Markov Chain

$$\lim_{t \to \infty} \max_{x \in \mathcal{I}} \left\| P_x^t - \pi \right\|_{tv} = 0.$$

Outline

Total Variation Distance

Coupling

Convergence Theorem - Proof and Application

Mixing Times

Coupling

Ordered Coins

- Two coins: X is fair, Y achieves heads with probability $2/3$. Thus: $X \sim \text{Ber}(1/2)$, $Y \sim \text{Ber}(2/3)$.
- Question : Can you sample (toss) the coins together so that $Y \geq X$?

Answer: Yes!

- Let $Z \sim \text{Ber}(1/3)$.
- Sample X and Z independently and let $Y = \max\{X, Z\}$.
- Because

$$\begin{aligned} \mathbf{P}[\,Y = 1\,] &= \mathbf{P}[\,X = 1\,] + \mathbf{P}[\,X = 0, Z = 1\,] \\ &= 1/2 + (1/2) \cdot (1/3) = 2/3, \end{aligned}$$

we have $Y \sim \text{Ber}(q)$. However, $Y = \max\{X, Z\} \geq X$.

Let μ and η be two probability measures on a finite Ω. A probability measure ν on $\Omega \times \Omega$ is a *Coupling* of (μ, η) if for every $x \in \Omega$,

$$\sum_{y \in \Omega} \nu(x, y) = \mu(x) \qquad \text{and} \qquad \sum_{y \in \Omega} \nu(y, x) = \eta(x).$$

Coupling Lemma

Let μ and η be two probability measures on a finite Ω and ν be a coupling of (μ, η) . If (X, Y) is distributed according to ν then

$$\|\mu - \eta\|_{tv} \leq \mathbf{P}[\, X \neq Y \,] .$$

Furthermore, there always is a coupling which achieves equality.

Proof: [Proof of $\leq$ only] For any $\omega \in \Omega$ we have

$$\mu(\omega) = \mathbf{P}[\, X = \omega \,] = \mathbf{P}[\, X = \omega, Y = X \,] + \mathbf{P}[\, X = \omega, Y \neq X \,]$$
$$= \mathbf{P}[\, Y = \omega, Y = X \,] + \mathbf{P}[\, X = \omega, Y \neq X \,] \leq \eta(\omega) + \mathbf{P}[\, X = \omega, Y \neq X \,] .$$

The same calculation with μ and η reversed gives

$$\eta(\omega) \leq \mu(\omega) + \mathbf{P}[\, Y = \omega, Y \neq X \,] .$$

Let $\Omega^+ = \{\omega : \mu(\omega) \geq \eta(\omega)\}$ and $\Omega^- = \{\omega : \mu(\omega) < \eta(\omega)\}$ then observe

$$\begin{aligned} 2 \cdot \|\mu - \eta\|_{tv} &= \sum_{\omega \in \Omega} |\mu(\omega) - \eta(\omega)| \\ &= \sum_{\omega \in \Omega^+} (\mu(\omega) - \eta(\omega)) + \sum_{\omega \in \Omega^-} (\eta(\omega) - \mu(\omega)) \\ &\leq \sum_{\omega \in \Omega^+} \mathbf{P}[\, X = \omega, Y \neq X \,] + \sum_{\omega \in \Omega^-} \mathbf{P}[\, Y = \omega, Y \neq X \,] \leq 2\mathbf{P}[\, Y \neq X \,] . \end{aligned}$$

$\square$

Coupling of Markov Chains

A Coupling of a Markov chain P on $\mathcal{I}$ is a M.C. $Z_t = (X_t, Y_t)$ on $\mathcal{I} \times \mathcal{I}$ s.t.:

$$\mathbf{P}\left[\, X_{t+1} = x' \mid Z_t = (x, y) \,\right] = \mathbf{P}\left[\, X_{t+1} = x' \mid X_t = x \,\right]$$
$$\mathbf{P}\left[\, Y_{t+1} = y' \mid Z_t = (x, y) \,\right] = \mathbf{P}\left[\, Y_{t+1} = y' \mid Y_t = y \,\right].$$

Tom & Jerry Chain

- A house $\mathcal{H} = \{0, 1, 2\}$ has a ground floor, first floor and a garden.
- Cat (T_i) and a Mouse (J_i) each move to one of the two adjacent areas with probability $1/2$ each time step. (P is SRW on 3-Cycle)

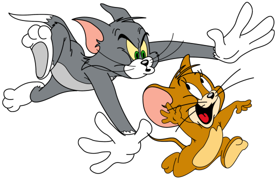
Question : Can the Mouse avoid the Cat indefinitely? (Provided $T_0 \neq J_0$)

Answer: YES! - Jerry's Coupling

Markov chain $Z_t = (T_t, J_t)$ on $\mathcal{H} \times \mathcal{H}$:

- Run the Cat T_t as normal
- Mouse J_t moves according to the rule:

$$J_{t+1} = \begin{cases} J_t + 1 \mod 3 & \texttt{if} \quad T_{t+1} = T_t + 1 \mod 3 \\ J_t - 1 \mod 3 & \texttt{if} \quad T_{t+1} = T_t - 1 \mod 3 \end{cases}$$

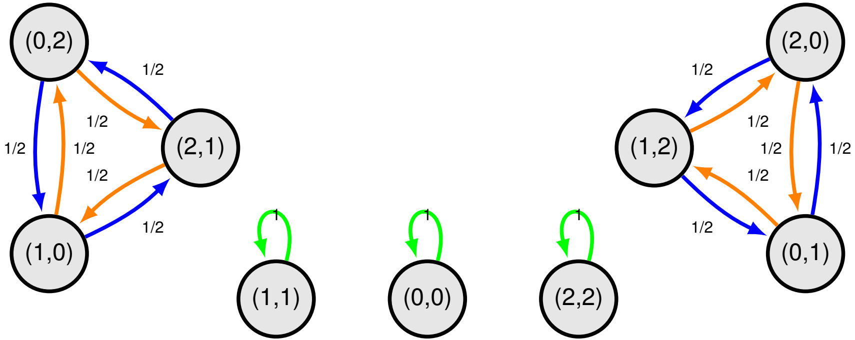


Mouse never shares a state with Cat, but is it a coupling?

Jerry's Coupling

Is it a coupling? Clearly the Cat's marginal distribution is correct, what about mouse? Let $\mathbf{P}_{(x,y)}[\cdot] := \mathbf{P}[\,\cdot \mid Z_t = (x,y)\,]$, then for $x, y, z \in \mathcal{H}$

$$\mathbf{P}_{(x,y)}[J_{t+1} = x] = \mathbf{P}[\,J_{t+1} = x,\, T_t = y\,] + \mathbf{P}[\,J_{t+1} = x,\, T_t = z\,] = 0 + 1/2 = 1/2$$
$$\mathbf{P}_{(x,y)}[J_{t+1} = y] = \mathbf{P}[\,J_{t+1} = y,\, T_t = y\,] + \mathbf{P}[\,J_{t+1} = y,\, T_t = z\,] = 0 + 0 = 0$$
$$\mathbf{P}_{(x,y)}[J_{t+1} = z] = \mathbf{P}[\,J_{t+1} = z,\, T_t = y\,] + \mathbf{P}[\,J_{t+1} = z,\, T_t = z\,] = 1/2 + 0 = 1/2.$$



Outline

Total Variation Distance

Coupling

Convergence Theorem - Proof and Application

Mixing Times

Convergence Theorem

Convergence Theorem (rephrased)

For any finite, irreducible, aperiodic M.C., $\lim_{t\to\infty} \max_{x\in\mathcal{I}} \left\| P_x^t - \pi \right\|_{tv} = 0.$

Proof: Let $x, y \in \mathcal{I}$ and X_t, Y_t be copies of P with $X_0 = x$ and $Y_0 \sim \pi$. Couple (X_t, Y_t) by running X_t, Y_t independently until the first time (τ) they meet, then they move together. More formally let $\widetilde{P}$ be the chain on $\mathcal{I} \times \mathcal{I}$ where

$$\widetilde{P}_{(x_1,y_1),(x_2,y_2)} = \begin{cases} P_{x_1,x_2} \cdot P_{y_1,y_2} & \text{if } x_1 \neq y_1 \\ P_{x_1,x_2} & \text{if } x_1 = y_1 \text{ and } x_2 = y_2 \\ 0 & \text{otherwise} \end{cases}.$$

This is the *Doblin Coupling*

As P is finite, aperiodic and irreducible there exists some time $T < \infty$ such that, for every $w, z \in \mathcal{I}$, $P_{w,z}^T > 0$. (you will prove this in problem class).

Let $C := \min_{w,z\in\mathcal{I}} P_{w,z}^T > 0$ so that $P_{x_1,z}^T \cdot P_{x_2,z}^T \geq C^2$ for all triples (x_1, x_2, z).

Thus after T steps X_T and Y_T meet with probability at least C^2.
Since X_t, Y_t are independent until they meet

$$\mathbf{P}[\,\tau \geq k \cdot T\,] \leq (1 - C^2)^k, \qquad k \in \mathbb{Z}_+.$$

Coupling Lemma: $\max_{x\in\mathcal{I}} \left\| P_x^t - \pi \right\|_{tv} \leq \mathbf{P}[\,X_t \neq Y_t\,] = \mathbf{P}[\,\tau > t\,] \xrightarrow[t\to\infty]{} 0.$ $\square$

Applications of Markov Chain Convergence

Markov Chain Monte Carlo (MCMC): Sampling, Counting, Integration, ...

Example : Markov Chain for Sampling a Matching of G.

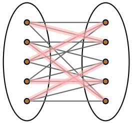
Pick some initial matching M (may have no edges)

1. With probability $1/2$ stay at M
2. Otherwise pick $uv \in E$ and let

$$M' = \begin{cases} M - \{uv\} & \text{if } uv \in M \\ M \cup \{uv\} & \text{if } uv \text{ can be added to } M \\ M \cup \{uv\} - \{e'\} & \text{if either } u \text{ or } v \text{ is matched by } e' \in M \\ M & \text{otherwise} \end{cases}$$

3. Let $M = M'$ and repeat steps $1 - 3$.

- Markov Chain on Matchings of G.
- Satisfies the Convergence theorem.
- Has uniform stationary distribution.
- Thus run it "long enough" then halt to return a uniform matching on G.



$$A = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 \\ 1 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$

Outline

Total Variation Distance

Coupling

Convergence Theorem - Proof and Application

Mixing Times

Mixing Time of a Markov Chain

Convergence Theorem: "Nice" Markov chains converge to stationarity.

Question How fast do they converge?

The *Mixing time* $\tau(\epsilon)$ of a Markov chain P with stationary distribution π is

$$\tau(\epsilon) = \min\left\{t\colon \max_x \left\|P_x^t - \pi\right\|_{TV} \leq \epsilon\right\}.$$

- This is how long we need to wait until we are "ε close" to stationarity .
- We often take $\varepsilon = 1/4$, indeed let $t_{mix} := \tau(1/4)$.
- For any fixed $0 < \epsilon < \delta < 1/2$ we have

$$\tau(\epsilon) \leq \left\lceil \frac{\ln \epsilon}{\ln 2\delta} \right\rceil \tau(\delta).$$

Thus for any $\epsilon < 1/4$

$$\tau(\epsilon) \leq \left\lceil \log_2 \epsilon^{-1} \right\rceil t_{mix}.$$

Bounding Mixing Times Using a Coupling

Coupling Lemma for Mixing

Let $Z_t = (X_t, Y_t)$ be a coupling for a Markov chain P on $\mathcal{I}$. Suppose that there exists a T such that, for every $x, y \in \mathcal{I}$,

$$\mathbf{P}[\, X_T \neq Y_T \mid X_0 = x, Y_0 = y \,] \leq \varepsilon.$$

Then $\tau(\epsilon) \leq T$.

Proof: Let X_i, Y_i be coupled (e.g. Doblin coupling) copies of P starting from $x \in \mathcal{I}$ and π respectively. Now for any $A \subseteq \mathcal{I}$ and T, ϵ as above we have

$$\begin{aligned}
\mathbf{P}[\, X_T \in A \,] &\geq \mathbf{P}[\, Y_T \in A, Y_T = X_T \,] \\
(\textit{Complementry events}) &= 1 - \mathbf{P}[\, \{X_T \neq Y_T\} \cup \{Y_T \notin A\} \,] \\
(\textit{Union bound}) &\geq 1 - \mathbf{P}[\, Y_T \notin A \,] - \mathbf{P}[\, X_T \neq Y_T \,] \\
(\textit{Hypothesis}) &\geq \mathbf{P}[\, Y_T \in A \,] - \epsilon \\
(\textit{Y stationary}) &= \pi(A) - \epsilon.
\end{aligned}$$

The same steps show $\mathbf{P}[\, X_T \notin A \,] \geq \pi(A) - \epsilon$ thus $\mathbf{P}[\, X_T \in A \,] \leq \pi(A) + \epsilon$. We now observe that $\tau(\epsilon)$ is at most T since

$$\max_{x \in \mathcal{I}} \left\| P_x^T - \pi \right\|_{tv} := \max_{x \in \mathcal{I}, A \subseteq \mathcal{I}} |P_x^t(A) - \pi(A)| \leq \epsilon. \qquad \square$$

Card Shuffling

A *Permutation* σ of $[n] = \{1, \dots, n\}$ is a bijection $\sigma : [n] \to [n]$.

Let Σ_n be the set of all $n!$ permutations of $[n]$.

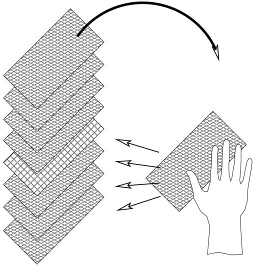
Sampling from uniform.

Given an ordered set $[n]$ we wish to sample a permutation of $[n]$ uniformly.

Top-to-Random (T-to-R) Shuffling

Given a deck of n cards take the top card and place it at random position in the deck.

- Markov chain on Σ_n with π uniform.



Prisoner Problem

Prisoner Problem

- There are 100 numbered prisoners in room A.
- Room B has a cupboard with 100 numbered drawers.
- The warden places each prisoners number an empty draw at random.
- The prisoners must go into room B alone and open at most 50 draws.
- Once finished they close the draws and return to their own cell.
- If all find their numbers then all survive, otherwise they all die.
- Prisoners fix a strategy pre-game, no communication after room B.

Question : What is the best strategy?

Question : What is $\mathbf{P}[\text{Success}]$ for this strategy?