Lecture 1. Coupling and Markov chain, basics

- Self-intro, email
- X: prob space, typically discrete & finite

1.1. Coupling

Motivating example:

Two players A, B each tosses a coin. A's coin has head prob = 0.75, B's coin has head prob = 0.5, each tosses 100 times.

Prove: $\mathbb{P}(A \text{ gets} \geq K \text{ heads}) \geq \mathbb{P}(B \text{ gets} \geq K \text{ heads})$

[Formal: $\mathbb{P}(\text{Binom}(100, 0.75) \geq k) \geq \mathbb{P}(\text{Binom}(100, 0.5) \geq k)$.]

At each time, let $X_A \in \{0,1\}$ be the outcome of A,

$X_B \in \{0,1\}$ — — — of B

We can couple X_A and X_B s.t. $X_A \geq X_B$

Therefore the inequality becomes trivial.

Observation: LHS only depends on A's distribution

RHS — — — B's distribution

Therefore we can "design" their correlation in whatever way.

Def: A coupling of two prob distributions $\mu, \nu \in \mathcal{P}(\mathcal{X})$ is a joint distribution on $\mathcal{X} \times \mathcal{X}$.

π (or $\pi \rightsquigarrow$)

if $(X, Y) \sim \pi$ then

$$\mathbb{P}(X = x) = \mu(x) \qquad \mathbb{P}(Y = y) = \nu(y) \qquad \forall\, x, y$$

P. Joint distribution that maintains marginal

Given μ, ν, a coupling can be described as a matrix $q \in \mathbb{R}^{|\mathcal{X}| \times |\mathcal{Y}|}$

with $q(x, y) = \mathbb{P}(X = x, Y = y)$ satisfying

$$\sum_y q(x, y) = \mu(x), \qquad \sum_x q(x, y) = \nu(y)$$

Example:

① Couple Ber(0.5) and Ber(0.75)

② Couple Ber(0.5) and Ber(0.5)

③ Couple Ber(p) and Ber(q) (2-d bond percolation)

1.2.

Coupling and total variation distance

Def: Given μ, ν on $\mathcal{X}$, their TV distance is

$$\|\mu - \nu\|_{TV} := \sup_A |\mu(A) - \nu(A)|$$

Lemma: $\|\mu - \nu\|_{TV} = \sup_A \mu(A) - \nu(A)$

$$= \sup_B \nu(B) - \mu(B)$$

$$= \sum_{\{x : \mu(x) > \nu(x)\}} \mu(x) - \nu(x)$$

$$= \frac{1}{2} \sum_{x \in \mathcal{X}} |\mu(x) - \nu(x)|$$

[TV distance is L^1-distance]

Coupling lemma: Fix μ, ν on $\mathcal{X}$, let $\Gamma(\mu, \nu)$ be the set of all the couplings between μ and ν. Then

$$\|\mu - \nu\|_{TV} = \inf_{(X,Y) \sim \Gamma(\mu,\nu)} \mathbb{P}(X \neq Y)$$

$$= 1 - \sup_{(X,Y) \sim \Gamma(\mu,\nu)} \mathbb{P}(X = Y)$$

D. Any coupling satisfying $\|\mu - \nu\|_{TV} = 1 - \mathbb{P}(X = Y)$ is called "maximal coupling".

Pf:

$$\|\mu - \nu\|_{TV} = \frac{1}{2} \sum_x |\mu(x) - \nu(x)|$$

$$= \frac{1}{2} \sum_{\{x : \mu(x) \geq \nu(x)\}} \mu(x) - \nu(x) + \frac{1}{2} \sum_{\{x : \nu(x) > \mu(x)\}} \nu(x) - \mu(x)$$

$$= \frac{1}{2} \sum_x \max\{\mu(x), \nu(x)\} - \min\{\mu(x), \nu(x)\}$$

$$= 1 - \sum_x \min\{\mu(x), \nu(x)\}$$

For any coupling $q \in \Gamma(\mu, \nu)$

$$P(X = Y = x) \leq \min\{\mu(x), \nu(x)\}$$

$$\Rightarrow P(X = Y) \leq \sum_x \min\{\mu(x), \nu(x)\}. \quad \square$$

Lemma: Maximal coupling always exists

Def $\tilde{\mu}(x) \propto \mu(x) - \min\{\mu(x), \nu(x)\}$

$\tilde{\nu}(x) \propto \nu(x) - \min\{\mu(x), \nu(x)\}$

Consider the following way of generating (X, Y)

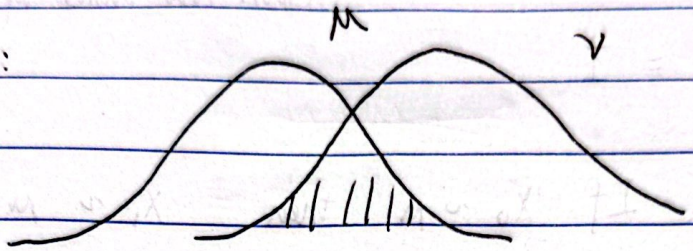
① Flip coin with head prob $\sum_x \min\{\mu(x), \nu(x)\}$

② If heads, generate $X \propto \min\{\mu, \nu\}$ set $Y = X$

If tails, generate $X \sim \tilde{\mu}$, $Y \sim \tilde{\nu}$ independently

Therefore $P(X=Y) \geq \sum_x \min\{\mu(x), \nu(x)\}$ □

Geometric:



TV distance = 1 − shaded area

1.3. Markov chain

A Markov chain on $\mathcal{X}$ is a stochastic process $(X_0, X_1, \cdots)$ such that for any (x, y) and any $x_0, \cdots x_{t-1}$, and any t such that $P(X_0 = x_0, X_1 = x_1, \cdots X_{t-1} = x_{t-1}, X_t = x) \neq 0$ we have

$$P(X_{t+1} = y \mid X_0 = x_0, \cdots X_{t-1} = x_{t-1}, X_t = x)$$
$$= P(X_{t+1} = y \mid X_t = x) \quad \text{[Markov property]}.$$

▷ Future depends only on present, but not past

If $P(X_{t+1} = y \mid X_t = x)$ does not depend on t, then Markov chain is homogeneous.

We say $P \in \mathbb{R}_+^{|\mathcal{X}| \times |\mathcal{X}|}$ is the transition matrix of M.C. if $\mathbb{P}(X_{t+1} = y \mid X_t = x) = P(x, y)$

(x-th row, y-th col)

▷ Transition matrix : ~~non-zero~~ non-negative entry

each row sums up to 1.

▷ $P\vec{1} = \vec{1}$.

Exercise : If $X_0 \sim \mu$ then $X_1 \sim \mu P$

Exercise : $\mathbb{P}^{\mu}(X_{s+t} = y \mid X_s = x) = P^t(x, y)$

(Transition matrix after t steps is just P^t).

Def : μ is called stationary distribution of P if $\mu P = P$.

Def : P is called irreducible if for any (x, y) there is t s.t $P^t(x, y) > 0$.

P is called aperiodic if $\forall x$, $\gcd T(x)$

$:= \gcd \{n, P^n(x, x) > 0\} = 1$

P is reversible with respect to π

if $\pi(x) P(x, y) = \pi(y) P(y, x)$.

Prop: P reversible w.r.t. $\pi \Rightarrow \pi P = \pi$.

Example : Random walk on $\mathbb{Z}$ ($* \ \mathcal{X} = \mathbb{Z}$).

$P(x, x+1) = P(x, x-1) = \frac{1}{2}$.

Example : Random walk on graphs.

A graph $G = (V, E)$ $\quad E \subseteq \{\{x, y\} \in V \times V, \ x \neq y\}$

vertex, edge

When $\{x, y\} \in E$, we say y is a neighbor of x,

$\deg(x) =$ # of neighbors of x.

Fix G, define $P(x, y) = \begin{cases} \frac{1}{\deg(x)} & y \sim x \\ 0 & \text{otherwise} \end{cases}$

▷ $G =$ 

▷ check : P reversible w.r.t. $\pi(x) \propto \deg(x)$.

▷ application : search, pagerank, spectral graph theory.

Example : Metropolis-Hastings algorithm.

Target : sample from π

Input : Initialization $x_0 \in \mathcal{X}$, proposal distribution $q(\cdot, \cdot)$

For t in $1 : T$:

draw $y \sim q(x, \cdot)$

set $X_t = \begin{cases} y & \text{w.p. } \frac{\pi(y) q(y, x)}{\pi(x) q(x, y)} \wedge 1 \ * \\ x_{t-1} & \text{w.p. } 1 - * \end{cases}$

Check: reversible w.r.t. π.

application: physics, statistics / ML, theoretical CS.

Example: shuffling

$\mathcal{X} = S_n$

$$P(\sigma, \tilde{\sigma}) = \begin{cases} \frac{1}{\binom{n}{2}} & \text{if } \tilde{\sigma} = \tau \circ \sigma \text{ for some transposition } \tau \\ 0 & \text{otherwise} \end{cases}$$

check: reversible w.r.t. Unif(S_n).

Example: Kac's walk on sphere ($* \ \mathcal{X} = S^{p-1}$)

$x_0 = (x_{0,1}, \cdots x_{0,p}) \in \mathbb{R}^p$ with $\sum_{i=1}^{p} x_{0,i}^2 = 1$

Each time pick two entries (i, j)

generate $\theta \sim \text{Unif}[0, 2\pi)$

change $\begin{pmatrix} x_i \\ x_j \end{pmatrix}$ to $\begin{pmatrix} \cos(\theta) & -\sin(\theta) \\ \sin(\theta) & \cos(\theta) \end{pmatrix} \begin{pmatrix} x_i \\ x_j \end{pmatrix}$

- Spectral gap (AoP, 2001)
  Spectral gap (Acta, 2003)
  Mixing time (AAP, 2017).
- Application: physics, quantum.

1.4 Stationary distribution and convergence.

Thm: Let P be the transition kernel of a M.C. Suppose it is irreducible and aperiodic, then there exists a unique stationary distribution μ.

Pf: Existence: Skipped, see Levin & Peres 1.5.4 or Perron-Frobenius thm.

Uniqueness: we show the following lemma

Lem: If $Ph = h$ then h must be const.

Pf: If not, assume h maximized at x_0 and there is y_0 s.t. $h(y_0) < h(x_0)$ and r s.t. $P^r(x,y) > 0$

Then

$$(P^r h)(x_0) = \sum_y h(y) P^r(x_0, y) \leq h(x_0)$$

contradiction

convince yourself.

This implies

$$\ker(P - I) = \{\lambda \vec{1} \mid \lambda \in \mathbb{R}\}$$

If π_1, π_2 stationary distribution, then

$$\pi_1 (P - I) = \pi_2 (P - I) = 0$$

$$(P^T - I)\, \pi_1^T = (P^T - I)\, \pi_2^T = 0$$

$$\Rightarrow \dim \ker(P^T - I) \geq \dim \{\text{span}\{\pi_1^T, \pi_2^T\}\} \geq 2$$

But $\dim(\ker(P^T - I)) = \dim(\ker(P - I))$. $\square$