Randomized Algorithms

Lecture 16: Total variation distance and coupling

Raul Garcia-Patron

School of Informatics
University of Edinburgh

Markov chain and mixing times

- Sampling from a given probability distribution is a fundamental algorithmic tool.
- We have seen that in some cases one can design a Markov chain that has as stationary distribution our target distribution.
  - After sufficiently many steps we converge to the target distribution regardless of the initial state.
- To achieve our goal, we need to have a guarantee of the convergence to the target distribution, this will be the goal of this and next lecture.
  1. This lecture: notion of distance + coupling as a tool to prove mixing times.
  2. Next lecture: path coupling to prove mixing times.

Total variation distance

Definition (Definition 12.1)

The total variation distance between two distributions D_1 and D_2 on a countable state space S is given by

$$\|D_1 - D_2\| \;=\; \frac{1}{2}\sum_{x \in S} |D_1(x) - D_2(x)|.$$

Properties:

1. Triangle inequality: $\|D_1 - D_3\| \leq \|D_1 - D_2\| + \|D_2 - D_3\|$
2. $\|D_1 - D_2\| = 0$ only if $D_1 = D_2$.
3. $0 \leq \|D_1 - D_2\| \leq 1$

- π being the stationary distribution of a Markov chain M. We want to bound the distance between the distribution of the chain after t steps when starting at state x, i.e., bound $\|p_x^t - \pi\|$.
- We want to show that it becomes ϵ small in number of steps t polynomial on the size of the problem.

Examples

- Two biased coins:
  1. $\{p(0) = p, p(1) = 1 - p\}$,
  2. $\{q(0) = 1 - p, q(1) = p\}$ (where $0 \leq p \leq 1/2$),

  $\|p - q\| = \frac{1}{2}(|p - (1 - p)| + |1 - p - p|) = 1 - 2p.$

- Non-overlapping supports:
  1. For all $W \subseteq A$, $D_1(W) > 0$ and $D_2(W) = 0$,
  2. where for all $W \subseteq \bar{A}$, $D_1(W) = 0$ and $D_2(W) \geq 0$.

$$\begin{aligned} \|D_1 - D_2\| &= \frac{1}{2}\sum_{x \in S} |D_1(x) - D_2(x)| \\ &= \frac{1}{2}\sum_{x \in A} D_1(x) + \frac{1}{2}\sum_{x \in \bar{A}} D_2(x) = 1 \end{aligned}$$

Operational interpretation

Definition (Lemma 12.1)

$$\|D_1 - D_2\| \;=\; \frac{1}{2}\sum_{x\in S}|D_1(x) - D_2(x)|.$$

For any $A \subseteq S$ let $D_i(A) = \sum_{x\in A} D_i(x)$, i.e., the weight of subspace A. Then

$$\|D_1 - D_2\| = \max_{A\subseteq S}|D_1(A) - D_2(A)|. \quad (1)$$

1. For any $B \subseteq S$ we have $\|D_1 - D_2\| \geq |D_1(B) - D_2(B)|$.
   - It can also used to proved non-convergence: if $\exists B$, s.t. $|D_1(B) - D_2(B)| > c$ then also $\|D_1 - D_2\| > c$.
2. If $\|D_1 - D_2\| \leq \epsilon$: D_1 and D_2 can not be distinguish up to error ϵ, i.e., whether you sample from one or the other is indistinguishable on any subset $B \subseteq S$!
   - Probability of guessing distribution 1 or 2 right: $P^{\max}_{\text{guess}} = \frac{1}{2}(1 + \|D_1 - D_2\|)$

Proof of Lemma 12.1

- $\|D_1 - D_2\| = \frac{1}{2}\sum_{x \in S} |D_1(x) - D_2(x)|$.
- $A \subseteq S : \|D_1 - D_2\| = \max_{A \subseteq S} |D_1(A) - D_2(A)|$.

Proof.

1. Let $S^+ \subseteq S$ s.t. $D_1(x) \geq D_2(x)$ and S^- complement
2. $\max_{A \subseteq S} D_1(A) - D_2(A) = D_1(S^+) - D_2(S^+)$
3. $\max_{A \subseteq S} D_2(A) - D_1(A) = D_2(S^-) - D_1(S^-)$
4. $D_1(S^+) + D_1(S^-) = 1 = D_2(S^+) + D_2(S^-)$
   - $D_1(S^+) - D_2(S^+) = D_2(S^-) - D_1(S^-)$
5. $\max_{S \subseteq S} |D_1(A) - D_2(A)| = |D_1(S^+) - D_2(S^+)| = |D_1(S^-) - D_2(S^-)|$
6. $|D_1(S^+) - D_2(S^+)| + |D_1(S^-) - D_2(S^-)| = 2\|D_1 - D_2\|$ (Def. TV)

□

Mixing time

Definition (Definition 12.2)

Let M be a finite, irreducible and aperiodic Markov chain over the state space Ω and let π be its stationary distribution. We define $\Delta_x(t), \Delta(t)$ as

$$\Delta_x(t) \; = \; \|M^t[x, \cdot] - \pi\|, \quad \Delta(t) \; = \; \max_{x \in \Omega} \Delta_x(t).$$

We also define

$$\tau_x(\epsilon) \; = \; \min\{t : \Delta_x(t) \leq \epsilon\}, \quad \tau(\epsilon) \; = \; \max_{x \in \Omega} \tau_x(\epsilon).$$

1. $\tau(\epsilon)$ is called mixing time.
2. A chain is rapidly mixing if $\tau(\epsilon)$ is polynomial in $\log(1/\epsilon)$ and the size of the problem.
3. There are two main techniques for upper-bounding mixing time:
   - Coupling: nice tight bounds when it works.
   - Conductance: worse bounds, works on a larger pool.

Coupling as upper-bound of TV distance

Definition (Definition 12.2)

A coupling of two probability distributions μ and ν is a pair of random variables (X, Y) defined on a single probability space, i.e., a joint probability distribution q on $\Omega \times \Omega$ such that

$$\sum_{y \in \Omega} q(x, y) = \mu(x) \textit{and} \sum_{x \in \Omega} q(x, y) = \nu(x) \tag{2}$$

Definition (Lemma 12.3)

Given distributions $\mu(x)$ and $\nu(x)$ on state space Ω. All couplings (X, Y) satisfy the condition

$$\inf \Pr(X \neq Y) \geq \|\mu - \nu\|. \tag{3}$$

- This will allow us to upper-bound distances between two Markov chains at step t, and also with respect to the stationary distribution, which leads to upper-bounds on mixing times.

Coupling example I

Coupling: $$\sum_{y\in\Omega} q(x,y) = \mu(x) \text{ and } \sum_{x\in\Omega} q(x,y) = \nu(x)$$

Bound on TV: $$\inf \Pr(X \neq Y) \geq \|\mu - \nu\|$$

Consider two fair coins: $\mu(x) = \nu(x) = c(x) = 1/2$.

It is trivial to see that $\|\mu - \nu\| = \|c - c\| = 0$ as both are equal.
A trivial coupling consist of two independent coins

- Because $q(x,y) = c(x)c(y) \Rightarrow \sum_{y\in\Omega} q(x,y) = c(x) = 1/2$.
- Remark that $\Pr(X \neq Y) = 1/2 > 0 = \|\mu - \nu\|$

Coupling example II

Coupling: $\sum_{y\in\Omega} q(x,y) = \mu(x)$ and $\sum_{x\in\Omega} q(x,y) = \nu(x)$

Bound on TV: $\inf \Pr(X \neq Y) \geq \|\mu - \nu\|$

Consider two fair coins: $\mu(x) = \nu(x) = c(x) = 1/2$.

It is trivial to see that $\|\mu - \nu\| = \|c - c\| = 0$ as both are equal.
Consider now a perfectly correlated coin:

- Because $q(0,0) = q(1,1) = 1/2 \Rightarrow \sum_{y\in\Omega} q(x,y) = c(x) = 1/2$.
- Remark that $\Pr(X \neq Y) = 0 = \|\mu - \nu\|$
- Can we saturated the lower-bound building correlated joint distribution!

Coupling example III

Consider the distributions

1. $\{\mu(0) = 3/4, \mu(1) = 1/4\}$
2. $\{\nu(0) = 1/4, \nu(1) = 3/4\}$

- Remark that $||\mu - \nu|| = 1/2$.
- The following algorithm generates a coupling.

Algorithm COUPLING COINS()

1. Generate a random bit with $p(b_1 = 0) = 1/2$.
2. **if** $b_1 = 0$ **then**
   Generate perfect random bit b_2 and fix $X = Y = b_2$.
3. **else**
   Fix $X = 0$ and $Y = 1$

- Always $\exists$ a coupling saturating $|D_1 - D_2||$

Proof Lemma 12.3 (upper-bound)

- $A \subseteq S : \|\mu - \nu\| = \max_{A \subseteq S} |\mu(A) - \nu(A)|$.
- Bound on TV: $\inf \Pr(X \neq Y) \geq \|\mu - \nu\|$

Proof.

1. $\mu(A) - \nu(A) = \Pr\{X \in A\} - \Pr\{Y \in A\}$
2.

$$\begin{aligned}
\mu(A) - \nu(A) &= (\Pr\{X \in A, y \in A\} + \Pr\{X \in A, Y \notin A\}) - \\
&\quad (\Pr\{X \in A, Y \in A\} + \Pr\{X \notin A, Y \in A\}) \\
&= \Pr\{X \in A, Y \notin A\} - \Pr\{X \notin A, Y \in A\} \\
&\leq \Pr\{X \notin A, Y \in A\} \\
&\leq \Pr\{X \neq Y\}
\end{aligned}$$

□

It has to hold for all coupling (X, Y), including the one that provides the minimum of $\Pr(X \neq Y)$.

Coupling of Markov chains

1. We have two chains X_t and Y_t that independently behave like the original one, governed by the transition rule P.
2. We couple the chains X and Y, on a joint chain $Z = (X, Y)$.
3. We design a Markov process M acting on Z (both X and Y), such that locally on each chain it still behaves as P, but globally the process is correlated.
   - $\Pr(X_{t+1} = x' | Z_t = (x, y)) = P(x, x')$,
   - $\Pr(Y_{t+1} = y' | Z_t = (x, y)) = P(x, x')$.
4. We are interested in chains that:
   - Bring the two copies of the chain to the same state
   - Once in same state they will make exactly the same move and remain equal

Coupling Lemma

Lemma (Lemma 12.2|Coupling lemma)

Let $Z_t = (X_t, Y_t)$ be a coupling for a Markov chain M on a state space S. Suppose there is a T such that, for every $x, y \in S$,

$$Pr(X_T \neq Y_T | X_0 = x, Y_0 = y) \leq \epsilon$$

Then

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

Shuffling cards

- We want to couple two chains: think of having two decks of card arranged in different configurations X_0 and Y_0.
- Each configuration is a given arrangement of the cards of a deck.
- Coupling:
  1. Choose a position j uniformly at random from deck 1 and then generate X_{t+1} from X_t by moving the j-th card to the top. Let's call that card C.
  2. Search for card C on the second deck and move it to the top to obtain Y_{t+1} from Y_t.
- The movements of card when looking at the deck independently is a standard reshuffling of card with probability $1/n$.
- Once a card is moved to the top it follows the same trajectory on both decks.
- Mixing when all cards have been moved to the top.
- We have mapped the problem to coupon collector: $\tau(\epsilon) = n \log(n/\epsilon)$.

Lazy random Walk on Hypercube

- Lazy random walk on the hypercube $\bar{x} = (x_1, x_2, ..., x_n)$.
  1. Select a coordinate uniformly at random from 1 to n.
  2. Set the value to 0 or 1 with equal probability $1/2$.
- Remark that with probability $1/2$ you remain in the same configuration $\Rightarrow$ aperiodicity.
- Coupling between X_t and Y_t via implementing the same move on both chains.
- Once the i-th coordinate chosen both chain will agree on i in future moves.
- The problem is again mapped to a coupon collector: $\tau(\epsilon) = n \log(n/\epsilon)$.

Path coupling: the big picture (Book section 12.6)

Neighborhood: states $y \in \Omega$ reachable from x in a single step.

- Distance $d(X, Y)$: the amount of steps to reach y from x. Neighbors if $d(X, Y) = 1$. Many times $d(X, Y) \leq |V|$.
  - Distance at step t of MC: $d_t = d(X_t, Y_t)$
- $\Pr(X_T \neq Y_T | X_0 = x, Y_0 = y) \leq \max_{x,y} \Pr(d_T \geq 1) \leq \max_{x,y} E[d_T]$
- Our goal is to bound $\max_{x,y} \mathbf{E}[d_T] \leq \epsilon$.
- After some work...
- $\mathbf{E}[d_{t+1}] =\leq \beta \mathbf{E}[d_t]$, with $\beta < 1$ (Contraction of expect. distance)
- Iterate $\mathbf{E}[d_T] =\leq \beta^T d_0 \leq \beta^T |V|$

Therefore the chain is guaranteed to have mixed for all times, such that $\beta^T |V| \leq \epsilon$, leading to

$$\tau(\epsilon) = \frac{1}{\log(1/\beta)} \left(\log|V| + \log(1/\epsilon)\right).$$

Many times we can write $\beta = e^{-\alpha/|V|}$, leading to $\tau(\epsilon) = \frac{|V|}{\alpha}\left(\log|V| + \log(1/\epsilon)\right)$, where α can itself depend on parameters of the problem.

Coupling Lemma

Lemma (Lemma 12.2|Coupling lemma)

Let $Z_t = (X_t, Y_t)$ be a coupling for a Markov chain M on a state space S. Suppose there is a T such that, for every $x, y \in S$,

$$Pr(X_T \neq Y_T | X_0 = x, Y_0 = y) \leq \epsilon$$

Then

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

Proof of Lemma 12.2: Coupling lemma

- $\Pr(X_T \neq Y_T | X_0 = x, Y_0 = y) \leq \epsilon \quad \Rightarrow \quad \tau(\epsilon) \leq T$
- Choose Y_0 according the uniform distribution and X_0 takes an arbitrary value. For T, ϵ such that lemma is satisfied and for any $A \subseteq S$:

$$\begin{aligned}
\Pr(X_T \in A) \geq\ & \Pr((X_T = Y_T) \cap (Y_T \in A)) \\
& = 1 - \Pr((X_T \neq Y_T) \cup Y_T \notin A) \\
& \geq (1 - \Pr(Y_T \notin A)) - \Pr(X_T \neq Y_T) \\
& \geq \Pr(Y_T \in A) - \epsilon \\
& = \pi(A) - \epsilon
\end{aligned}$$

- The same argument for $S - A$ shows: $\Pr(X_T \notin A) \geq \pi(S - A) - \epsilon$ or equivalently $\Pr(X_T \in A) \leq \pi(A) + \epsilon$.
- It follows:

$$||p_x^T - \pi|| = \max_{x,A} |p_x^T(A) - \pi(A)| \leq \epsilon.$$