Lecture-14: Total Variation Distance

1 Comparison of Monte Carlo Methods

We have seen that Monte Carlo sampling methods give us the desired stationary distribution starting from a base Markov chain. Assuming one step of the Markov chain takes unit time, we can compare their computational efficiency in terms of number of operations needed per sample, and the number of sampled needed to reach *close enough* to the desired stationary distribution.

1.1 Total variation distance

The closeness of two distributions can be measured by the following distance metric.

Definition 1.1. The **total variation distance** between two probability distributions μ and ν on $\mathcal{X}^N$ is defined by

$$\|\mu - \nu\|_{\mathrm{TV}} = \max\left\{|\mu(A) - \nu(A)| : A \subseteq \mathcal{X}^N\right\}.$$

This definition is probabilistic in the sense that the distance between μ and ν is the maximum difference between the probabilities assigned to a single event by the two distributions.

Proposition 1.2. *Let μ and ν be two probability distributions on the configuration space $\mathcal{X}^N$. Let $B = \{x \in \mathcal{X}^N : \mu(x) \geqslant \nu(x)\}$, then*

$$\|\mu - \nu\|_{TV} = \frac{1}{2}\sum_{x\in\mathcal{X}^N} |\mu(x) - \nu(x)| = \sum_{x\in B}[\mu(x) - \nu(x)].$$

Proof. Let $A \subseteq \mathcal{X}^N$ be any event. Since $\mu(x) - \nu(x) < 0$ for any $x \in A \cap B^c$, we have

$$\mu(A) - \nu(A) \leqslant \mu(A \cap B) - \nu(A \cap B) \leqslant \mu(B) - \nu(B).$$

Similarly, we can show that

$$\nu(A) - \mu(A) \leqslant \nu(B^c) - \mu(B^c) = \mu(B) - \nu(B).$$

It follows that $|\mu(A) - \nu(A)| \leqslant \mu(B) - \nu(B)$ for all events A, and the equality is achieved for $A = B$ and $A = B^c$. Thus, we get that

$$\|\mu - \nu\|_{\mathrm{TV}} = \frac{1}{2}[\mu(B) - \nu(B) + \nu(B^c) - \mu(B^c)] = \frac{1}{2}\sum_{x\in\mathcal{X}^N} |\mu(x) - \nu(x)|.$$

□

Exercise 1.3. Show that $\|\mu - \nu\|_{\mathrm{TV}}$ is a distance.

Proposition 1.4. *Let μ and ν be two probability distributions on the configuration space $\mathcal{X}^N$. For any observable $f : \mathcal{X}^N \to \mathbb{R}$, we have*

$$\|\mu - \nu\|_{TV} = \frac{1}{2}\sup\left\{\sum_{x\in\mathcal{X}^N} f(x)\mu(x) - \sum_{x\in\mathcal{X}^N} f(x)\nu(x) : \max_{x\in\mathcal{X}^N} |f(x)| \leqslant 1\right\}.$$

Proof. If $\max_x |f(x)| \leqslant 1$, then it follows that

$$\frac{1}{2}\left|\sum_x f(x)(\mu(x)-\nu(x)\right| \leqslant \frac{1}{2}\sum_x |\mu(x)-\nu(x)| = \|\mu-\nu\|_{\mathrm{TV}}.$$

For the reverse inequality, we define $f^*(x) = \mathbb{1}_{\{x\in B\}} - \mathbb{1}_{\{x\notin B\}}$ in terms of the set $B = \{x \in \mathcal{X}^N : \mu(x) \geqslant \nu(x)\}$. It is clear that $\max_x |f(x)| = 1$, and we have

$$\frac{1}{2}\left|\sum_x f^*(x)(\mu(x)-\nu(x)\right| = \frac{1}{2}\sum_x |\mu(x)-\nu(x)| = \|\mu-\nu\|_{\mathrm{TV}}.$$

□

1.2 Coupling and total variation distance

Definition 1.5. A **coupling** of two probability distributions μ and ν is a pair of random variables (X,Y) defined on a single probability space such that the marginal distribution of X is μ and the marginal distribution of Y is ν. That is, a coupling (X,Y) satisfies $P\{X=x\} = \mu(x)$ and $P\{Y=y\} = \nu(y)$.

Any two distributions μ and ν have an independent coupling. However, when the two distributions are not identical, it will not be possible for the random variables to always have the same value. Total variation distance between μ and ν determine how close can a coupling get to having X and Y identical.

Definition 1.6. For two distributions μ and ν on the configuration space $\mathcal{X}^N$, the coupling (X,Y) is **optimal** if

$$\|\mu-\nu\|_{\mathrm{TV}} = P\{X \neq Y\}.$$

Proposition 1.7. *Let μ and ν be two distributions on the configuration space $\mathcal{X}^N$, then*

$$\|\mu-\nu\|_{TV} = \inf\{P\{X\neq Y\} : (X,Y) \text{ a coupling of distributions } \mu,\nu\}.$$

Proof. For any coupling (X,Y) of the distributions μ,ν and any event $A \subseteq \mathcal{X}^N$, we have

$$\mu(A)-\nu(A) = P\{X\in A\} - P\{Y\in A\} \leqslant P\{X\in A, Y\notin A\} \leqslant P\{X\neq Y\}.$$

Therefore, it follows that $\|\mu-\nu\|_{\mathrm{TV}} \leqslant P\{X\neq Y\}$ for all couplings (X,Y) of distributions μ,ν.

Next we find a coupling (X,Y) for which $\|\mu-\nu\|_{\mathrm{TV}} = P\{X\neq Y\}$. In terms of the set $B = \{x\in\mathcal{X}^N : \mu(x)\geqslant\nu(x)\}$, we can write

$$p \triangleq \sum_{x\in\mathcal{X}} \mu(x)\wedge\nu(x) = \mu(B^c)+\nu(B) = 1-(\mu(B)-\nu(B)) = 1-\|\mu-\nu\|_{\mathrm{TV}}.$$

By the definition of p, the function $\frac{\mu\wedge\nu}{p} : \mathcal{X}^N \to [0,1]$ is a probability distribution on $\mathcal{X}^N$. Let us call this distribution as $\gamma_3(x) \triangleq \frac{\mu(x)\wedge\nu(x)}{p}$. We also define the following two function from the configuration space $\mathcal{X}^N$ to $[0,1]$ as

$$\gamma_1(x) \triangleq \frac{\mu(x)-\nu(x)}{\|\mu-\nu\|_{\mathrm{TV}}}\mathbb{1}_{\{x\in B\}}, \qquad \gamma_2(x) \triangleq \frac{\nu(x)-\mu(x)}{\|\mu-\nu\|_{\mathrm{TV}}}\mathbb{1}_{\{x\notin B\}}.$$

From the definition of the set B, we can easily verify that $\gamma_1(\mathcal{X}^N) = \gamma_1(B) = \gamma_2(B^c) = \gamma_2(\mathcal{X}^N) = 1$. We define a binary random variable $\xi\in\{0,1\}$ such that $\mathbb{E}\xi = p$, and the conditional distribution of (X,Y) such that

$$P((X,Y)=(x,y) \mid \xi) = \gamma_3(x)\mathbb{1}_{\{x=y\}}\xi + (1-\xi)\gamma_1(x)\gamma_2(y).$$

Since $\gamma_1,\gamma_2,\gamma_3$ are distributions, it follows that $P\{(X,Y)=(x,y)\} = p\gamma_3(x)\mathbb{1}_{\{x=y\}} + (1-p)\gamma_1(x)\gamma_2(y)\mathbb{1}_{\{x\neq y\}}$ is a joint distribution function. From the definition of the set B, we observe that

$$P\{X=x\} = p\gamma_3(x)+(1-p)\gamma_1(x) = \mu(x)\wedge\nu(x)+(\mu(x)-\nu(x))\mathbb{1}_{\{x\in B\}} = \mu(x)$$
$$P\{Y=y\} = p\gamma_3(y)+(1-p)\gamma_2(y) = \mu(y)\wedge\nu(y)+(\nu(y)-\mu(y))\mathbb{1}_{\{y\notin B\}} = \nu(y).$$

That is, (X,Y) is a coupling of the distributions μ,ν and $P\{X\neq Y\} = 1-p = \|\mu-\nu\|_{\mathrm{TV}}$. □