COMPSCI 630: Randomized Algorithms	January 31, 2018
Lecture 5	
Lecturer: Rong Ge	*Scribe: Aaron Lowe*

1 Overview

It is useful to be able to measure how similar two random variables are to one another. To this end, we will see two different ways to measure "distances" between the distribution of random variables. If two random variables are close under these measures, we say that they are more similar than if their distances are far apart.

2 Total Variational Distance

The first, and perhaps most natural, distance will be the total variational distance.

Definition 1. *Given two random variables P,Q, letting p,q be the associated density functions, the total variational distance is defined as*

$$\delta(P,Q) = \frac{1}{2}\int |p(x) - q(x)|\, dx.$$

When P,Q are discrete distributions, abusing notation, letting P,Q be probability vectors over the support $\Omega \to [0,1]$, and then we have

$$\delta(P,Q) = \frac{1}{2}\|P - Q\|_1.$$

Remark 1. *Multiplying by $\frac{1}{2}$ is just a normalizing constant, so two disjoint distributions have total variational distance 1. In some texts, this factor is not always included.*

Example 1. If P is random variable describing a coin flip, the associated vector would be $\left(\frac{1}{2},\frac{1}{2}\right)$.

The distance between two random variables can also measure how much we can "distinguish" two random variables. Thinking of functions of the form $f : \Omega \to \{0,1\}$ to be tests we claim the following:

Claim 1. $\delta(P,Q) = \max_{f:\Omega\to\{0,1\}} \mathbb{E}[f(p)] - \mathbb{E}[f(Q)]$.

Proof. For any $i \in \Omega$, to maximize this value, if $P(i) \geq Q(i)$, we should set $f(i)$ to 1. If $P(i) \leq Q(i)$ then we should set $f(i)$ to 0. then we see that

$$\begin{aligned}
\mathbb{E}[f(P)] - \mathbb{E}[f(Q)] &= \mathbb{E}[f(P) - f(Q)] \\
&= \sum_{i\in\Omega} (P(i) - Q(i))\, f(i) \\
&= \sum_{i\in\Omega} (P(i) - Q(i))\, 1_{P(i)\geq Q(i)} \\
&= \frac{1}{2}\|P - Q\|_1 = \delta(P,Q)
\end{aligned}$$

The last step is not obvious, but follows from the fact that since P and Q are both probability distributions,

$$\sum_{i\in\Omega} P(i) - Q(i) = 0$$

so it follows that

$$\sum_{i\in\Omega} (P(i) - Q(i))\, 1_{P(i)\geq Q(i)} = \sum_{i\in\Omega} (Q(i) - P(i))\, 1_{Q(i)\geq P(i)}$$

and we conclude by noting that

$$\begin{aligned}\|P - Q\|_1 &= \sum_{i\in\Omega} |P(i) - Q(i)| \\ &= \sum_{i\in\Omega} (P(i) - Q(i))\, 1_{P(i)\geq Q(i)} + \sum_{i\in\Omega} (Q(i) - P(i))\, 1_{Q(i)\geq P(i)}.\end{aligned}$$

□

We will now see an application of the total variational distance.

Example 2. Suppose we have two different coins P and Q. Letting 0 be the outcome of tails, and 1 be heads, P is fair coin with distribution

$$P = \begin{cases} 0 & \text{with prob. } 1/2 \\ 1 & \text{with prob. } 1/2 \end{cases}$$

and Q is biased with distribution

$$Q = \begin{cases} 0 & \text{with prob. } 1/2 - \varepsilon \\ 1 & \text{with prob. } 1/2 + \varepsilon \end{cases}$$

Then we have that $\delta(P,Q) = \varepsilon$.

Example 3. Now suppose we want to compare the distribution of multiple flips of the coins. Let nP (resp. nQ) be a collection of n random variables $(X_1, X_2, \ldots, X_n)$ with each $X_i \sim P$ with all X_i independent (resp. $(Y_1, Y_2, \ldots, Y_n)$ with $Y_i \sim Q$.) We say that these variables are independent and identically distributed (**i.i.d.**).

Claim 2. $\delta(nP, nQ) \leq n\delta(P,Q)$

Proof. While this is true in general, we will just show this for our example. To prove this we will rely on a useful technique called a coupling argument. This depends on the fact that the measure δ does not depend on the correlation between P and Q. We can exploit this by "designing" P, Q to be highly correlated (while still having the correct marginal distributions.) This will make arguing about the distance of the i.i.d.'s much easier.

We will first let the fair coin to be drawn normally $X = (X_1, \ldots, X_n) \sim nP$. Now, we define

$$Y_i = \begin{cases} X_i & \text{with prob. } 1 - 2\varepsilon \\ 1 & \text{with prob. } 2\varepsilon \end{cases}$$

We make two key observations:

1. Y_i is highly correlated to X_i, and

2. the distribution of Y is exactly nQ.

Now for any test f:

$$\begin{aligned}\mathbb{E}[f(X)] - \mathbb{E}[f(Y)] &= \mathbb{E}[f(Y) - f(X)] \\ &= \mathbb{E}\left[(f(X) - f(Y))\,1_{X \neq Y}\right] \\ &\leq \Pr[X \neq Y] \\ &\leq n\Pr[X_i \neq Y_i] \\ &= n\varepsilon\end{aligned}$$

The last two steps follow from a union bound. For large values of n it can be improved to $1-(1-\varepsilon)^n$ by calculating $\Pr[X \neq Y]$ exactly. □

This bound implies we need at least $n = \frac{1}{2\varepsilon}$ trials to be able to distinguish the fair and biased coin with constant probability.

To get a better bound, we will need another way to measure distances between distributions.

3 KL Divergence

Another measure of similarity between two probability distributions is the Kullback-Leibler divergence, or more commonly the KL-divergence.

Definition 2. *The KL divergence from P to Q is*

$$KL[P \parallel Q] = \sum_{i \in \Omega} P(i) \log \frac{P(i)}{Q(i)}.$$

Remark 2. *We note that if for some $i, P(i) = 0, Q(i) \neq 0$, then we think of the term in the sum as the limit, so it will contribute 0. On the other hand if $P(i) \neq 0$ and $Q(i) = 0$, then the KL divergence is defined to be ∞.*

The KL divergence from P to Q "intuitively" measures the amount of information needed to sample from distribution P given distribution Q.

Note that the KL divergence is not a metric. First, it is not usually symmetric; from Remark 2, we see in particular it is not symmetric if the supports are not equal. Second, it does not satisfy the triangle inequality. However it does have some useful properties.

Claim 3. $KL[P \parallel Q] \geq 0$

Proof.

$$\begin{aligned}\mathrm{KL}[P \parallel Q] &= \sum_{i \in \Omega} P(i) \left(-\log \frac{Q(i)}{P(i)}\right) \\ &\geq -\log\left(\sum_{i \in \Omega} P(i) \frac{Q(i)}{P(i)}\right) && \text{By Jensen's inequality, since } -\log \text{ is convex} \\ &= -\log 1 = 0.\end{aligned}$$

□

Claim 4. *Chain Rule of KL*

$$KL\left[(P,P') \parallel (Q,Q')\right] = KL[P \parallel Q] + KL\left[P'|P \parallel Q'|Q\right]$$

where

$$KL\left[P|P' \parallel Q|Q'\right] = \mathbb{E}_{x\sim P}\left[KL\left[P'|P=x \parallel Q'|Q=x\right]\right].$$

Corollary 5. *If P and P' are independent (and likewise Q and Q'), then*

$$KL\left[(P,P') \parallel (Q,Q')\right] = KL[P \parallel Q] + KL\left[P' \parallel Q'\right].$$

Example 3. Returning to the problem of distinguishing a fair and biased coins from Example 2 recall we have fair coin P with distribution

$$P = \begin{cases} 0 & \text{with prob. } 1/2 \\ 1 & \text{with prob. } 1/2 \end{cases}$$

and Q is biased with distribution

$$Q = \begin{cases} 0 & \text{with prob. } 1/2-\varepsilon \\ 1 & \text{with prob. } 1/2+\varepsilon \end{cases}$$

Then we see

$$\begin{aligned} \mathrm{KL}[Q \parallel P] &= (1/2+\varepsilon)\log\left(\frac{1/2+\varepsilon}{1/2}\right) + (1/2-\varepsilon)\log\left(\frac{1/2-\varepsilon}{1/2}\right) \\ &= (1/2+\varepsilon)\log(1+2\varepsilon) + (1/2-\varepsilon)\log(1-2\varepsilon) \\ &\approx 2\varepsilon^2 \qquad \text{using } 2^{nd} \text{ order Taylor expansion.} \end{aligned}$$

By Corollary 5, we have that $\mathrm{KL}[nQ \parallel nP] = n\mathrm{KL}[Q \parallel P] \approx 2n\varepsilon^2$. Finally we will relate this back to total variation using the following

Theorem 6. *Pinsker's Inequality*

$$KL[P \parallel Q] \geq \frac{1}{2}\|P-Q\|_1^2 = 2\delta(P,Q)^2.$$

This implies for the total variational distance $\delta(nP,nQ)$ to be at least than 1/2 then it follows that $\mathrm{KL}[nP \parallel nQ] \geq 1/2$. But combining with our analysis of the KL divergence gives us that $2n\varepsilon^2 \geq 1/2$. This gives a bound on the number of trials needed: $n \geq \frac{1}{4\varepsilon^2}$. This bound is tighter than that given from total variational distance alone.

Finally we will show that $O(1/\varepsilon^2)$ trials do suffice to distinguish the two coins. The expected number of heads of P is $\frac{n}{2}$, while the expectation for Q is $\frac{n}{2}(1+2\varepsilon)$. Further using the laws of variance, we have that the variance of the number of heads of P is $n/4$, and that of Q is also at most $n/4$. Since the difference in means is εn while the standard deviations is $\sqrt{n}/2$ the two distributions can be distinguished with high probability. To choose the right n, suppose that two standard deviations above the mean of nP and two below nQ do not overlap (which will give the probability of this happening being less than 1/2, which can be formalized using Chebychev's inequality.)

Then we have

$$\begin{aligned} \varepsilon n &\geq 4\frac{\sqrt{n}}{2} \\ \sqrt{n} &\geq \frac{2}{\varepsilon} \\ n &\geq \frac{4}{\varepsilon^2}. \end{aligned}$$

So by taking $O(1/\varepsilon^2)$ trials we should be able to distinguish P and Q with constant probability.

4 Summary

In this lecture we saw two different methods of measuring the distance between probability distributions: total variational distance, and KL divergence. We also saw how these measures can be applied to answer questions about distributions, such as how many trials one needs to be able to distinguish between two different random variables.