Relations Between Probability Measures

IV125, Adam Ivora

June 5, 2022

1 Definitions

In this TED talk, we will talk about the relation between Kullback-Leibler divergence and the total variation distance between probability distributions. We will focus on **discrete** probability distributions only. Let's start with some definitions and assumed theorems first:

Definition 1.1. S is a sample space, $\mathcal{A} = 2^S$. A probability distribution $P : \mathcal{A} \to [0,1]$ is a probability distribution. P satisfies Kolmogorov axioms of probability.

For $s \in S$, we also use the shorthand notation $P(s) := P(\{s\})$.

Definition 1.2. The Kullback-Leibler divergence between two probability distributions $P(x)$ and $Q(x)$ from discrete probability spaces defined over the same S is

$$\mathrm{D_{KL}}(P||Q) = \sum_{x \in S} P(x) \log \frac{P(x)}{Q(x)}. \tag{1}$$

Definition 1.3. The Manhattan distance (L_1 metric) between two probability distributions $P(x)$ and $Q(x)$ from discrete probability spaces defined over the same S is

$$||P - Q||_1 = \sum_{x \in S} |P(x) - Q(x)|. \tag{2}$$

Definition 1.4. The total variation distance between two probability distributions $P(x)$ and $Q(x)$ from discrete probability spaces defined over the same S is

$$\delta(P, Q) = \max_{A \in 2^S} |P(A) - Q(A)|. \tag{3}$$

Theorem 1.5. Jensen's inequality.

For a convex function f, and reals $p_1, \dots, p_n \geq 0$ such that $\sum_{i=1}^{n} p_i = 1$ it holds that:

$$f\left(\sum_{i=1}^{n} p_i x_i\right) \leq \sum_{i=1}^{n} p_i f(x_i) \tag{4}$$

2 Lower Bound

Theorem 2.1. Pinsker's inequality. For two probability distributions $P(x)$ and $Q(x)$ from discrete probability spaces defined over the same S, it holds that

$$||P - Q||_1 \leq \sqrt{2\,\mathrm{D_{KL}}(P||Q)}.$$

The equivalent inequality is that

$$\mathrm{D_{KL}}(P||Q) \geq \frac{1}{2}||P - Q||_1^2.$$

Proof. Bernoulli distributions case.

Let's denote by P and Q Bernoulli distribution over $S = \{0, 1\}$. Also, denote:

$$p = P(0), 1 - p = P(1)$$
$$q = Q(0), 1 - q = Q(1)$$

We can see that

$$||P-Q||_1 = |p-q| + |1-p-1+q| = 2 \cdot |p-q|$$
$$||P-Q||_1^2 = 4(p-q)^2$$

Let's define $f(p,q) = \mathrm{D_{KL}}(P||Q) - \frac{1}{2}||P-Q||_1^2$. We will analyse the behaviour of the function using basic calculus.

$$\begin{aligned}
f(p,q) &= p\log\frac{p}{q} + (1-p)\log\frac{1-p}{1-q} - 2(p-q)^2 \\
\frac{\partial f}{\partial q} &= p \cdot \frac{q}{p} \cdot \frac{-p}{q^2} + (1-p) \cdot \frac{1-q}{1-p} \cdot (1-p)\frac{-1}{(1-q)^2} \cdot (-1) - 4(p-q) \cdot (-1) \\
&= -\frac{p}{q} + \frac{1-p}{1-q} + 4(p-q) \\
&= \frac{-p+pq+q-pq}{q(1-q)} - 4(q-p) \\
&= (q-p)\left[\frac{1}{q(1-q)} - 4\right]
\end{aligned}$$

We see that with $q \neq \frac{1}{2}$, the sign of the partial derivative depends only on the sign of $q-p$. Therefore, $\frac{\partial f}{\partial q}$ is negative for $q < p$, positive for $q > p$ and 0 for $q = p$. That means that for $q \neq \frac{1}{2}$, $q = p$ is the minimum of $f(p,q)$.

$$\begin{aligned}
f(p,p) &= p\log\frac{p}{p} + (1-p)\log\frac{1-p}{1-p} - 2(p-p)^2 \\
&= 0.
\end{aligned}$$

That means that for $q \neq \frac{1}{2}$, $f(p,q)$ is non-negative and the Pinsker's inequality holds.

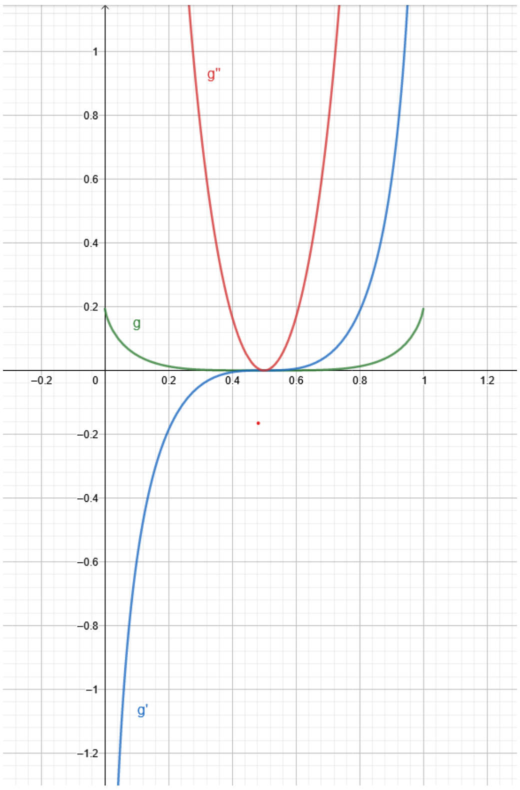
We now analyze

$$\begin{aligned}
g(p) = f\left(p, \frac{1}{2}\right) &= p \cdot \log(2p) + (1-p) \cdot \log(2-2p) - 2 \cdot \left(p - \frac{1}{2}\right)^2 \\
g'(p) &= \log(2p) + p \cdot \frac{1}{2p} \cdot 2 + (-1) \cdot \log(2-2p) + (1-p) \cdot \frac{1}{2-2p} \cdot (-2) - 4 \cdot (p - \frac{1}{2}) \\
&= \log(2p) + 1 - \log(2-2p) - 1 - 4p + 2 \\
&= \log(2p) - \log(2-2p) - 4p + 2. \\
g''(p) &= \frac{1}{2p} \cdot 2 - \frac{1}{2-2p} \cdot (-2) - 4 \\
&= \frac{1}{p} + \frac{1}{1-p} - 4 = \frac{1-p+p-4p+4p^2}{p \cdot (1-p)} \\
&= \frac{4p^2-4p+1}{p-p^2} = \frac{4\left(p-\frac{1}{2}\right)^2}{p \cdot (1-p)}.
\end{aligned}$$

For $p \in (0,1)$, the denominator of the second derivative is always positive and therefore the sign depends only on the nominator. As the nominator is also non-negative for $p \in (0,1)$, the second derivative is always non-negative and the first derivative is therefore a non-decreasing function.

$$\begin{aligned}
g'\left(\frac{1}{2}\right) &= \log 1 - \log 1 - 2 + 2 = 0 \\
g\left(\frac{1}{2}\right) &= \frac{1}{2} \cdot \log 1 + \frac{1}{2} \cdot \log 1 - 0 = 0 \\
g(0) &= 0 \cdot \log 0 + 1 \cdot \log 2 = \infty \\
g(1) &= 1 \cdot \log 2 + 0 \cdot \log 0 = \infty
\end{aligned}$$

The figure sums up our analysis:

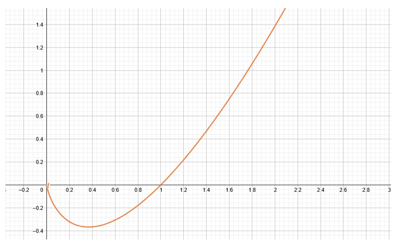


□

Therefore, the Pinsker's inequality holds for two arbitrary Bernoulli distributions. For the general case, we will need the log sum inequality and the information processing inequality:

Lemma 2.2. Log sum inequality Let $p_1, p_2, \ldots, p_n, q_1, q_2, \ldots, q_n \in \mathbb{R}_0^+$ be non-negative real numbers. Let $p = \sum_{i=1}^n p_i$ and $q = \sum_{i=1}^n q_i$. Then

$$\sum_{i=1}^{n} p_i \log \frac{p_i}{q_i} \geq p \log \frac{p}{q}.$$



Proof. Set $f(x) = x \log x$. Notice that f is a convex function. Then,

$$\begin{aligned}
\sum_{i=1}^{n} p_i \log \frac{p_i}{q_i} &= \sum_{i=1}^{n} p_i \frac{q_i}{q_i} \log \frac{p_i}{q_i} \\
&= \sum_{i=1}^{n} q_i f\left(\frac{p_i}{q_i}\right) \\
&= q \sum_{i=1}^{n} \frac{q_i}{q} f\left(\frac{p_i}{q_i}\right) \\
&\geq q \cdot f\left(\sum_{i=1}^{n} \frac{q_i}{q} \cdot \frac{p_i}{q_i}\right) \\
&= q \cdot f\left(\frac{1}{q} \sum_{i=1}^{n} p_i\right) = q \cdot f\left(\frac{p}{q}\right) \\
&= q \cdot \frac{p}{q} \cdot \log \frac{p}{q} \\
&= p \log \frac{p}{q}.
\end{aligned}$$

□

Lemma 2.3. Information processing inequality.

For any function $f : S \to S'$ and probability distributions $X : 2^S \to [0,1]$ and $Y : 2^S \to [0,1]$ defined over S, define

$$\begin{aligned}
X' &: 2^{S'} \to [0,1], \\
Y' &: 2^{S'} \to [0,1].
\end{aligned}$$

For every $i \in S'$, define

$$\begin{aligned}
X'(i) = X(f^{-1}(i)) &= \sum_{w \in f^{-1}(i)} X(w), \\
Y'(i) = Y(f^{-1}(i)) &= \sum_{w \in f^{-1}(i)} Y(w).
\end{aligned}$$

If X' and Y' are probability distributions, then

$$\mathrm{D_{KL}}(X' || Y') \leq \mathrm{D_{KL}}(X || Y).$$

Proof.

$$\begin{aligned}
\mathrm{D_{KL}}(X||Y) &= \sum_{w \in S} X(w) \log \frac{X(w)}{Y(w)} \\
&= \sum_{i \in S'} \sum_{w \in f^{-1}(i)} X(w) \log \frac{X(w)}{Y(w)} \\
&\geq \sum_{i \in S'} X'(i) \log \frac{X'(i)}{Y'(i)} \\
&= \mathrm{D_{KL}}(X'||Y').
\end{aligned}$$

□

Proof. Pinsker's inequality. Given probability distributions $P(x)$ and $Q(x)$ from discrete probability spaces defined over the same S, define $f : S \to \{0,1\}$

$$f(w) = \begin{cases} 1 & P(w) \leq Q(w), \\ 0 & P(w) > Q(w). \end{cases}$$

Define probability distributions $P', Q' : 2^{\{0,1\}} \to [0,1]$ for $i \in \{0,1\}$ as

$$
\begin{aligned}
P'(i) &= P(f^{-1}(i)) = \sum_{w \in f^{-1}(i)} P(w), \\
Q'(i) &= Q(f^{-1}(i)) = \sum_{w \in f^{-1}(i)} Q(w), \\
P'(0) &= \sum_{\{w \in S \ | \ P(w) > Q(w)\}} P(w), \\
Q'(0) &= \sum_{\{w \in S \ | \ P(w) > Q(w)\}} Q(w), \\
P'(1) &= \sum_{\{w \in S \ | \ P(w) \le Q(w)\}} P(w), \\
Q'(1) &= \sum_{\{w \in S \ | \ P(w) \le Q(w)\}} Q(w).
\end{aligned}
$$

From this follows that $P'(0) > Q'(0)$ and $P'(1) \le Q'(1)$.

As P' and Q' are Bernoulli distributions, we know that $\mathrm{D_{KL}}(P'||Q') \ge \frac{1}{2}||P' - Q'||_1^2$ by Pinsker's inequality for Bernoulli distributions.

Also,

$$
\begin{aligned}
||P - Q||_1 &= \sum_{w \in S} |P(w) - Q(w)| \\
&= \sum_{w \in f^{-1}(0)} (P(w) - Q(w)) + \sum_{w \in f^{-1}(1)} (Q(w) - P(w)) \\
&= P'(0) - Q'(0) + Q'(1) - P'(1) \\
&= |P'(0) - Q'(0)| + |P'(1) - Q'(1)| \\
&= ||P' - Q'||_1.
\end{aligned}
$$

Therefore, $\mathrm{D_{KL}}(P'||Q') \ge \frac{1}{2}||P - Q||_1^2$.

By information processing inequality, we know that

$$\mathrm{D_{KL}}(P'||Q') \le \mathrm{D_{KL}}(P||Q).$$

And that is all, folks!

$$\mathrm{D_{KL}}(P||Q) \ge \mathrm{D_{KL}}(P'||Q') \ge \frac{1}{2}||P' - Q'||_1^2 = \frac{1}{2}||P - Q||_1^2.$$

□

3 Upper Bound

There does not exist such a nice lower bound for KL divergence for a simple reason.

3.1 A counterexample

Theorem 3.1. Kullback-Leibler divergence is not upper bounded by the L_1 metric. Formally, for every $\varepsilon > 0$, there exist probability distributions P_ε and Q such that:

$$||P - Q||_1 \le \varepsilon, \text{ but } \mathrm{D_{KL}}(P||Q) = \infty.$$

Proof. Define $P(x)$ and Q as

$$
\begin{aligned}
S &= \{a, b\} \\
Q(a) &= 0, Q(b) = 1 \\
P_\varepsilon(a) &= \frac{\varepsilon}{2}, P_\varepsilon(b) = 1 - \frac{\varepsilon}{2}
\end{aligned}
$$

Then,

$$||P - Q||_1 = \varepsilon,$$
$$\mathrm{D_{KL}}(P||Q) = \frac{\varepsilon}{2} \cdot \log \frac{\frac{\varepsilon}{2}}{0} = \infty.$$

□

3.2 A proof

Theorem 3.2. For two probability distributions $P(x)$ and $Q(x)$ that are defined over the same S, it holds that

$$\mathrm{D_{KL}}(P||Q) \leq \frac{1}{2\alpha_Q}||P - Q||_1^2,$$

where

$$\alpha_Q = \min_{x \in S} Q(x).$$

4 Misc

4.1 Total variation distance vs L_1 norm

Theorem 4.1. Scheffé's lemma.

For two probability distributions $P(x)$ and $Q(x)$ that are defined over the same S, it holds that

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

Proof. Let's refresh the definition of the total variation distance:

$$\delta(P, Q) = \max_{A \in 2^S} |P(A) - Q(A)|. \tag{5}$$

Denote by $G = \{x \in S \mid P(x) \geq Q(x)\}$. Try to find $A \subset S$ such that $P(A) - Q(A)$ is maximized. Intuitively, it is the case when $A = G$.

Now, try to find $A' \subset S$ such that $Q(A) - P(A)$ is maximized. Intuitively, it is the case when $A' = S \setminus G$.

Therefore, the subset A in $\delta(P, Q) = \max_{A \in 2^S} |P(A) - Q(A)|$ is either $A = G$ or $A' = S \setminus G$. We will show that the maximum is obtained at both A and A':

$$P(G) - Q(G) = (1 - P(S \setminus G)) - (1 - Q(S \setminus G)) = Q(S \setminus G) - P(S \setminus G).$$

So if A maximizes $P(X) - Q(X)$, A' maximizes $Q(X) - P(X)$ and if A' maximizes $P(X) - Q(X)$, then A maximizes $Q(X) - P(X)$.

Now,

$$\begin{aligned}
||P - Q||_1 &= \sum_{x \in S} |P(x) - Q(x)| \\
&= \sum_{x \in G} (P(x) - Q(x)) + \sum_{x \in S \setminus G} (Q(x) - P(x)) \\
&= \delta(P, Q) + \delta(P, Q) \\
&= 2 \cdot \delta(P, Q)
\end{aligned}$$

□

4.2 Inverse Pinsker inequality

Let P and Q be probability distributions on the finite set A. Let $A_+ = \{a : Q(a) > 0\}$ and let $\alpha_Q = \min_{a \in A_+} Q(a)$.

How to prove that if $D(P||Q) < \infty$ then

$$D(P||Q) \leq \frac{d^2(P,Q)}{\alpha_q \cdot \ln 2},$$

where $d(P,Q)$ is the variational distance of distributions P and Q, i.e., $d(P,Q) = \sum_{a \in A} |P(a) - Q(a)|$.

I was given a hint that first should prove that:

$$D(P||Q) \leq \sum_{a \in A_+} \frac{P(a)}{\ln 2} \left(\frac{P(a)}{Q(a)} - 1\right) = \frac{1}{\ln 2} \sum_{a \in A_+} \frac{|P(a) - Q(a)|^2}{Q(a)}.$$

statistics information-theory

Share Cite Follow

asked Mar 1, 2020 at 15:25
tefodos139
1

What did you try, and where are you stuck? Can you prove the hinted inequality? Can you conclude the argument assuming the hint? – stochasticboy321 Mar 5, 2020 at 4:07

Add a comment

swer

Sorted by: Highest score (default)

For the Hint from the right hand side to the middle is easy, hint distract the squared expression.

$$\sum_{a \in A_+} \frac{P(a)}{\ln 2} \left(\frac{P(a)}{Q(a)} - 1\right) = \frac{1}{\ln 2} \sum_{a \in A_+} P(a) exp(ln \frac{P(a)}{Q(a)}) - 1 \geqslant \frac{1}{\ln 2} exp(\sum_{a \in A_+} P(a) ln(\frac{P(a)}{Q(a)})) - 1$$
$$\geqslant exp(D(P||Q)) - 1$$
$$\geqslant 1 + D(P||Q) - 1 = D(P||Q)$$

For the reversed pinker's;

$$D(P||Q) \leqslant \frac{1}{\ln 2} \sum_{a \in A_+} \frac{|P(a) - Q(a)|^2}{Q(a)}$$

$$\leqslant \frac{1}{\ln 2} \sum_{a \in A_+} \frac{|P(a) - Q(a)|^2}{min_{a \in A_+} Q(a)} \leqslant \frac{max_{a \in A_+} |P(a) - Q(a)| . \sum_{a \in A_+} |P(a) - Q(a)|}{\alpha_Q . ln 2} \leqslant \frac{d^2(P,Q)}{\alpha_Q . ln 2}$$

The last inequality you can deduce it after using scheffe's theorem for variational distance.