

① HYPOTHESIS TESTING

general problem: two probability distributions $P_1(x)$, $P_2(x)$ on finite set X (one null and one alternative)
 we observe n iid samples, x_1, \dots, x_n , drawn from an unknown distribution
 goal is generally to determine which distribution samples came from

② TOTAL VARIATION DISTANCE

prerequisite to successful hypothesis testing is a way to measure distance between distributions

two distributions: $A = A(x_1) \dots A(x_k)$ for $k = |\Omega|$
 $B = B(x_1) \dots B(x_k)$

$$\boxed{\text{TV}(A, B) = \frac{1}{2} \|A - B\|_1 = \frac{1}{2} \sum_{x \in \Omega} |A(x) - B(x)|}$$

note the $\frac{1}{2}$ ensures TVD is bounded between 0 and 1, not 0 and 2

WHY NOT JUST USE KL?

- ① nonsymmetric $= KL$
- ② $1 - TVD$ is lower bound of type I and type II error rates
- ③ many other distances, but this one is most distinguishing

Intuition: TVD is best distinguishing probability supremum over all subsets of Ω of $|A(x) - B(x)|$ [i.e. largest possible difference between the probabilities that two probability distributions can assign to the same event]

more concretely:

null hypothesis $H_0 = A$

alt hypothesis $H_A = B$

have algorithm \mathcal{T} whose job, given samples, outputs after each sample if thinks sample came from distribution A or distribution B

2 common types of error: Type I = false \oplus = incorrectly reject true H_0

Type II = false \ominus = fail to reject false H_0

lets say I reject A when event Z occurs:

$$\text{type I error + type II error} = A(Z) + B(Z^c)$$

$$= A(Z) + (1 - B(Z))$$

$$= 1 + (A(Z) - B(Z))$$

$$\geq 1 + \inf_A [A(Z) - B(Z)]$$

$$= 1 - \sup_A [B(Z) - A(Z)]$$

$$= 1 - \text{TV}(A, B)$$

missing step? $\text{TV}(A, B) = \sup_A (|A - B|)$

$\sup_A (|A - B|) = \sup_A (A - B)$

①

SKIP ③ TVD BEHAVIOR ?

- If $A \neq B$, then $\lim_{n \rightarrow \infty} TVD(A^n, B^n) \rightarrow 1$

- but, $TVD(A^{k+1}, B^{k+1})$ can be equal to $TVD(A^k, B^k)$

PROVE?

④ PINSKER'S INEQUALITY

recall KL divergence $D(P||Q) = \sum p(x) \log \frac{p(x)}{q(x)}$

If P and Q are discrete distributions, then $D(P||Q) \geq \frac{1}{2 \ln(2)} \|P-Q\|_1^2$

so we can use KL to upper bound the TVD: $TVD(P, Q) \leq \frac{1}{2} \sqrt{2 \ln(2) D(P||Q)}$

PROOF: I'll prove a special case of coin flip / Bernoulli trial

$$\text{let } P = \begin{cases} 1 & \text{wp } p \\ 0 & \text{wp } 1-p \end{cases} \quad Q = \begin{cases} 1 & \text{wp } q \\ 0 & \text{wp } 1-q \end{cases}$$

assume $p \geq q$ (but similar steps allow us to prove the other case)

$$D(P||Q) = p \log \frac{p}{q} + (1-p) \log \frac{1-p}{1-q}$$

$$\| (p, 1-p) - (q, 1-q) \|_1 = p - q + p - q = 2(p - q)$$

$$\text{let } f(p, q) = p \log \frac{p}{q} + (1-p) \log \frac{1-p}{1-q} - \frac{1}{2 \ln 2} (2(p - q))^2$$

$$\frac{\partial f}{\partial q} = -\frac{p-q}{\ln 2} \left(\frac{1}{q(1-q)} - 4 \right) \leq 0 \quad \text{i.e. difference between two sides of inequality} \\ \text{f=0 if } q=p \\ \text{since } q \leq 1$$

$$\text{so } f(p, q) \geq 0 \text{ if } p \geq q \text{ so } D(P||Q) \geq \frac{1}{2 \ln 2} \|P-Q\|_1^2 \text{ for this special case}$$

WHY? when increase q : function decreases since $\frac{\partial f}{\partial q} < 0$

we care when $q \leq p$, which means q must decrease. If q decreases, f increases above 0

for general case,

sketch: want to prove for my distributions P and Q , calculate $\|P-Q\|_1$, and simplify to

$\|P_A - Q_A\|_1$, where $P_A \sim \text{Bern}(\sum_x p(x))$ (these two will be equal)

then use chain rule of KL divergence and the fact that KL is non-negative

CHAIN RULE OF KL

$$\begin{aligned} KLLP(X, Y) || Q(X, Y) \\ = KLL(P(X)||Q(X)) \\ + KLLP || Q(X) \end{aligned}$$

(2)

⑤ APPLICATION OF PINSKER: lower bound of coin samples needed to distinguish two coins with slightly different biases

let 1 = heads, 0 tails

we are given one of two coins:

$$P = \begin{cases} 1 & \text{up } \frac{1}{2}-\epsilon \\ 0 & \text{up } \frac{1}{2}+\epsilon \end{cases} \quad Q = \begin{cases} 1 & \text{up } \frac{1}{2} \\ 0 & \text{up } \frac{1}{2} \end{cases} \quad \text{fair}$$

we have our algorithm $A(x_1 \dots x_m) \rightarrow \{0, 1\}$

assume our algorithm is good, i.e.

$$\underset{x \in P^m}{\Pr}(A(x)=0) \geq \frac{9}{10} \quad \underset{x \in Q^m}{\Pr}(A(x)=1) \geq \frac{9}{10}$$

goal: derive a lower bound for m (without knowing A)

rewrite assumption conditions in terms of expectations:

$$\underset{x \in P^m}{\mathbb{E}}[A(x)] \leq \frac{1}{10} \quad \underset{x \in Q^m}{\mathbb{E}}[A(x)] \geq \frac{9}{10}$$

since $\frac{9}{10}(0) + \frac{1}{10}(1) = \frac{1}{10}$

$$\text{so } \underset{x \in P^m}{\mathbb{E}}[A(x)] - \underset{x \in Q^m}{\mathbb{E}}[A(x)] \geq \frac{8}{10}$$

apply this lemma: \tilde{P} and \tilde{Q} are distributions on some universal probability world \mathcal{U}
 let $f: \mathcal{U} \rightarrow [0, B]$ B represents a discrete upper bound

$$\left| \underset{\tilde{P}}{\mathbb{E}} f(x) - \underset{\tilde{Q}}{\mathbb{E}} f(x) \right| \leq \frac{B}{2} \|\tilde{P} - \tilde{Q}\|_1$$

Proof: rewrite using definition of expected value / LOTUS

$$\begin{aligned} \left| \sum_x \tilde{P}(x)f(x) - \sum_x \tilde{Q}(x)f(x) \right| &= \left| \sum_x [\tilde{P}(x) - \tilde{Q}(x)]f(x) \right| \\ &= \left| \sum_x (\tilde{P}(x) - \tilde{Q}(x))(f(x) - \frac{B}{2}) + \frac{B}{2}(\sum_x \tilde{P}(x) - \tilde{Q}(x)) \right| \quad \text{add and subtract same thing} \\ &\leq \sum_x |\tilde{P}(x) - \tilde{Q}(x)| |f(x) - \frac{B}{2}| \\ &\leq \frac{B}{2} \|\tilde{P} - \tilde{Q}\|_1. \end{aligned}$$

now if $f = A$, $\tilde{P} = P^m$, $\tilde{Q} = Q^m$

$$\|P^m - Q^m\|_1 \geq 2 \left| \mathbb{E}_{x \in P^m} A(x) - \mathbb{E}_{x \in Q^m} A(x) \right| = \frac{8}{5}$$

recall from lecture 3 that

$$\begin{aligned} m D(P||Q) &= D(P^m || Q^m) \\ &\geq \frac{1}{2 \ln 2} \left(\frac{8}{5} \right)^2 \text{ by Pinsker's inequality} \end{aligned}$$

$$m \geq \frac{1}{2 \ln 2} \frac{1}{D(P||Q)} \left(\frac{8}{5} \right)^2$$

now last thing is to bound $D(P||Q)$:

$$\begin{aligned} D(P||Q) &= \left(\frac{1}{2} - \epsilon \right) \log \left(\frac{\frac{1}{2} - \epsilon}{\frac{1}{2}} \right) + \left(\frac{1}{2} + \epsilon \right) \log \left(\frac{\frac{1}{2} + \epsilon}{\frac{1}{2}} \right) \\ &= \frac{1}{2} \log((1-2\epsilon)(1+2\epsilon)) + \epsilon \log \left(\frac{1+2\epsilon}{1-2\epsilon} \right) \\ &\leq \frac{\epsilon}{\ln 2} \ln \left(1 + \frac{4\epsilon}{1-2\epsilon} \right) \\ &\leq \frac{4\epsilon^2}{\ln 2} \frac{1}{1-2\epsilon} \quad \text{since } \ln(1+x) \leq e^x \end{aligned}$$

if we assume $\epsilon \leq \frac{1}{4}$ i.e. is small

$$D(P||Q) \leq \frac{8\epsilon^2}{\ln 2}$$

$$\text{so } m \geq \frac{1}{2 \ln 2 D(P||Q)} \left(\frac{8}{5} \right)^2 \geq \frac{4}{256}$$

can show this is upto constants using Chernoff bound

⑥ LOWER BOUND ON TVD

[Vadhan's PhD thesis]

"Direct Product" lemma: X and Y are distributions such that $\text{TVD}(X, Y) = \delta$
then $\forall k \in \mathbb{N}$, $|1 - 2e^{-k\delta^2/2} \leq \text{TVD}(X^k, Y^k)|$

Proof: recall that TVD is the best distinguishing probability so there exists some set S such that

$$\text{P}(X \in S) - \text{P}(Y \in S) = \delta$$

$$\text{Let } p = \text{P}(Y \in S) \text{ so } \text{P}(X \in S) = p + \delta$$

then in k samples of X , expected # of those that lie in S are $(p + \delta)k$ and similarly for Y is pk

we now apply Chernoff bounds:

$$\text{P}(\text{at least } (p + \frac{\delta}{2})k \text{ components of } Y^k \text{ lies in } S) \geq e^{-k\delta^2/2}$$

$$\text{P}(\text{at most } (p + \frac{\delta}{2})k \text{ components of } X^k \text{ lie in } S) \geq e^{-k\delta^2/2}$$

If S' is the set of k tuples that contain more than $(p + \frac{\delta}{2})k$ components that lie in S , then:

$$\begin{aligned} \text{TVD}(X^k, Y^k) &\geq \text{P}(X^k \in S') - \text{P}(Y^k \in S') \\ &\geq 1 - 2e^{-k\delta^2/2} \end{aligned}$$

so we can lower bound TVD and upper bound TVD for Bernoulli coin flip
direct product lemma Pinsker's inequality