Information and Coding Theory

MA41024/ MA60020/ MA60262

Bibhas Adhikari

Spring 2022-23, IIT Kharagpur

Lecture 7

January 24, 2023

KL divergence

Convexity property of KL divergence

Log-sum inequality Let $a_1, a_2, b_1, b_2 \geq 0$. Then

$$(a_1 + a_2) \log \left(\frac{a_1 + a_2}{b_1 + b_2} \right) \leq a_1 \log \frac{a_1}{b_1} + a_2 \log \frac{a_2}{b_2}$$

KL divergence

Convexity property of KL divergence

Log-sum inequality Let $a_1, a_2, b_1, b_2 \geq 0$. Then

$$(a_1 + a_2) \log \left(\frac{a_1 + a_2}{b_1 + b_2} \right) \leq a_1 \log \frac{a_1}{b_1} + a_2 \log \frac{a_2}{b_2}$$

In general,

$$\left(\sum_i a_i \right) \log \left(\frac{\sum_i a_i}{\sum_i b_i} \right) \leq \sum_i a_i \log \frac{a_i}{b_i}$$

KL divergence

Convexity property of KL divergence

Log-sum inequality Let $a_1, a_2, b_1, b_2 \geq 0$. Then

$$(a_1 + a_2) \log \left(\frac{a_1 + a_2}{b_1 + b_2} \right) \leq a_1 \log \frac{a_1}{b_1} + a_2 \log \frac{a_2}{b_2}$$

In general,

$$\left(\sum_i a_i \right) \log \left(\frac{\sum_i a_i}{\sum_i b_i} \right) \leq \sum_i a_i \log \frac{a_i}{b_i}$$

Proof Recall that $f(x) = x \log x$ is a strictly convex function for all $x > 0$. By Jensen's inequality

$$f \left(\sum_i^n \alpha_i x_i \right) \leq \sum_{i=1}^n \alpha_i f(x_i)$$

where $\sum_{i=1}^n \alpha_i = 1$, $\alpha_i \geq 0$.

KL divergence

Convexity property of KL divergence

Log-sum inequality Let $a_1, a_2, b_1, b_2 \geq 0$. Then

$$(a_1 + a_2) \log \left(\frac{a_1 + a_2}{b_1 + b_2} \right) \leq a_1 \log \frac{a_1}{b_1} + a_2 \log \frac{a_2}{b_2}$$

In general,

$$\left(\sum_i a_i \right) \log \left(\frac{\sum_i a_i}{\sum_i b_i} \right) \leq \sum_i a_i \log \frac{a_i}{b_i}$$

Proof Recall that $f(x) = x \log x$ is a strictly convex function for all $x > 0$. By Jensen's inequality

$$f\left(\sum_i^n \alpha_i x_i \right) \leq \sum_{i=1}^n \alpha_i f(x_i)$$

where $\sum_{i=1}^n \alpha_i = 1$, $\alpha_i \geq 0$.

Kl divergence

Then the proof follows by setting $x_i = \frac{a_i}{b_i}$ and $\alpha_i = \frac{b_i}{\sum_{i=1}^{n} b_i}$.

Kl divergence

Then the proof follows by setting $x_i = \frac{a_i}{b_i}$ and $\alpha_i = \frac{b_i}{\sum_{i=1}^{n} b_i}$.

Question When does the equality hold?

Kl divergence

Then the proof follows by setting $x_i = \frac{a_i}{b_i}$ and $\alpha_i = \frac{b_i}{\sum_{i=1}^n b_i}$.

Question When does the equality hold?

Lemma Let P_1, P_2, Q_1, Q_2 be distributions on a finite set $\mathcal{X}$, and $\alpha \in [0, 1]$. Then

$$\begin{aligned} & D(\alpha P_1 + (1-\alpha)P_2 \,\|\, \alpha Q_1 + (1-\alpha)Q_2) \\ & \leq \alpha D(P_1 \| Q_1) + (1-\alpha) D(P_2 \| Q_2) \end{aligned}$$

Kl divergence

Then the proof follows by setting $x_i = \frac{a_i}{b_i}$ and $\alpha_i = \frac{b_i}{\sum_{i=1}^{n} b_i}$.

Question When does the equality hold?

Lemma Let P_1, P_2, Q_1, Q_2 be distributions on a finite set $\mathcal{X}$, and $\alpha \in [0, 1]$. Then

$$\begin{aligned} & D(\alpha P_1 + (1-\alpha)P_2 \,\|\, \alpha Q_1 + (1-\alpha)Q_2) \\ & \leq \alpha D(P_1 \| Q_1) + (1-\alpha) D(P_2 \| Q_2) \end{aligned}$$

Meaning When $D(P\|Q)$ is viewed as a function of the inputs P and Q, is jointly convex in both of it's inputs i.e. it is convex in the input (P, Q)

KL divergence

Proof

$$
\begin{aligned}
& D(\alpha P_1 + (1-\alpha)P_2 \,\|\, \alpha Q_1 + (1-\alpha)Q_2) \\
= \quad & \sum_{x \in \mathcal{X}} (\alpha p_1(x) + (1-\alpha)p_2(x)) \log \left(\frac{\alpha p_1(x) + (1-\alpha)p_2(x)}{\alpha q_1(x) + (1-\alpha)q_2(x)} \right)
\end{aligned}
$$

KL divergence

Proof

$$
\begin{aligned}
& D(\alpha P_1 + (1-\alpha)P_2 \,\|\, \alpha Q_1 + (1-\alpha)Q_2) \\
= \; & \sum_{x\in\mathcal{X}} (\alpha p_1(x) + (1-\alpha)p_2(x)) \log\left(\frac{\alpha p_1(x) + (1-\alpha)p_2(x)}{\alpha q_1(x) + (1-\alpha)q_2(x)}\right) \\
\leq \; & \sum_{x\in\mathcal{X}} \alpha p_1(x) \log\left(\frac{\alpha p_1(x)}{\alpha q_1(x)}\right) + (1-\alpha)p_2(x) \log\left(\frac{(1-\alpha)p_2(x)}{(1-\alpha)q_2(x)}\right)
\end{aligned}
$$

KL divergence

Proof

$$
\begin{aligned}
& D(\alpha P_1 + (1-\alpha)P_2 \,\|\, \alpha Q_1 + (1-\alpha)Q_2) \\
= \; & \sum_{x\in\mathcal{X}} (\alpha p_1(x) + (1-\alpha)p_2(x)) \log\left(\frac{\alpha p_1(x) + (1-\alpha)p_2(x)}{\alpha q_1(x) + (1-\alpha)q_2(x)}\right) \\
\leq \; & \sum_{x\in\mathcal{X}} \alpha p_1(x) \log\left(\frac{\alpha p_1(x)}{\alpha q_1(x)}\right) + (1-\alpha)p_2(x) \log\left(\frac{(1-\alpha)p_2(x)}{(1-\alpha)q_2(x)}\right) \\
= \; & \alpha D(P_1\|Q_1) + (1-\alpha)D(P_2\|Q_2)
\end{aligned}
$$

KL divergence

l_1-distance Let P and Q be two distributions on a finite set $\mathcal{X}$. Then the total-variation distance or statistical distance between P and Q is defined as

$$\delta_{\mathsf{TV}}(P,Q) = \frac{1}{2}\|P-Q\|_1 = \frac{1}{2}\sum_x |p(x)-q(x)|$$

KL divergence

l_1-distance Let P and Q be two distributions on a finite set $\mathcal{X}$. Then the total-variation distance or statistical distance between P and Q is defined as

$$\delta_{\text{TV}}(P, Q) = \frac{1}{2}\|P - Q\|_1 = \frac{1}{2}\sum_x |p(x) - q(x)|$$

Question Is it a standard notion of distance?

KL divergence

l_1-distance Let P and Q be two distributions on a finite set $\mathcal{X}$. Then the total-variation distance or statistical distance between P and Q is defined as

$$\delta_{\mathsf{TV}}(P, Q) = \frac{1}{2}\|P - Q\|_1 = \frac{1}{2}\sum_x |p(x) - q(x)|$$

Question Is it a standard notion of distance?

The quantity $\|P - Q\|_1$ is referred to as the l_1-distance between P and Q

KL divergence

Lemma Let P, Q be any distributions on $\mathcal{X}$. Let $f : \mathcal{X} \to [0, B]$. Then

$$|\mathbb{E}_P[f(x)] - \mathbb{E}_Q[f(x)]| \leq \frac{B}{2}\|P - Q\|_1 = B \cdot \delta_{\mathsf{TV}}(P, Q)$$

KL divergence

Lemma Let P, Q be any distributions on $\mathcal{X}$. Let $f : \mathcal{X} \to [0, B]$. Then

$$|\mathbb{E}_P[f(x)] - \mathbb{E}_Q[f(x)]| \leq \frac{B}{2}\|P - Q\|_1 = B \cdot \delta_{\mathsf{TV}}(P, Q)$$

Proof

$$\begin{aligned} & |\mathbb{E}_P[f(x)] - \mathbb{E}_Q[f(x)]| \\ = & \left|\sum_x p(x) \cdot f(x) - \sum_x q(x) \cdot f(x)\right| \end{aligned}$$

KL divergence

Lemma Let P, Q be any distributions on $\mathcal{X}$. Let $f : \mathcal{X} \to [0, B]$. Then

$$|\mathbb{E}_P[f(x)] - \mathbb{E}_Q[f(x)]| \leq \frac{B}{2}\|P - Q\|_1 = B \cdot \delta_{\mathsf{TV}}(P, Q)$$

Proof

$$\begin{aligned} & |\mathbb{E}_P[f(x)] - \mathbb{E}_Q[f(x)]| \\ = & \left| \sum_x p(x) \cdot f(x) - \sum_x q(x) \cdot f(x) \right| \\ = & \left| \sum_x (p(x) - q(x)) \cdot f(x) \right| \end{aligned}$$

KL divergence

Lemma Let P, Q be any distributions on $\mathcal{X}$. Let $f : \mathcal{X} \to [0, B]$. Then

$$|\mathbb{E}_P[f(x)] - \mathbb{E}_Q[f(x)]| \leq \frac{B}{2}\|P - Q\|_1 = B \cdot \delta_{\mathsf{TV}}(P, Q)$$

Proof

$$\begin{aligned}
&|\mathbb{E}_P[f(x)] - \mathbb{E}_Q[f(x)]| \\
= &\left|\sum_x p(x) \cdot f(x) - \sum_x q(x) \cdot f(x)\right| \\
= &\left|\sum_x (p(x) - q(x)) \cdot f(x)\right| \\
= &\left|\sum_x (p(x) - q(x)) \cdot \left(f(x) - \frac{B}{2}\right) + \frac{B}{2} \cdot \left(\sum_x (p(x) - q(x))\right)\right|
\end{aligned}$$

KL divergence

Lemma Let P, Q be any distributions on $\mathcal{X}$. Let $f : \mathcal{X} \to [0, B]$. Then

$$|\mathbb{E}_P[f(x)] - \mathbb{E}_Q[f(x)]| \leq \frac{B}{2}\|P - Q\|_1 = B \cdot \delta_{\mathsf{TV}}(P, Q)$$

Proof

$$\begin{aligned}
& |\mathbb{E}_P[f(x)] - \mathbb{E}_Q[f(x)]| \\
= \ & \left|\sum_x p(x) \cdot f(x) - \sum_x q(x) \cdot f(x)\right| \\
= \ & \left|\sum_x (p(x) - q(x)) \cdot f(x)\right| \\
= \ & \left|\sum_x (p(x) - q(x)) \cdot \left(f(x) - \frac{B}{2}\right) + \frac{B}{2} \cdot \left(\sum_x (p(x) - q(x))\right)\right| \\
\leq \ & \sum_x |p(x) - q(x)| \cdot \left|f(x) - \frac{B}{2}\right| \leq \frac{B}{2}\|P - Q\|_1
\end{aligned}$$

KL divergence

Question What is the interpretation of the above lemma?

KL divergence

Question What is the interpretation of the above lemma?

Let $f : \mathcal{X} \to \{0, 1\}$ be any classifier. For instance, f outputs 1 if the guess is that the sample point came from P and 0 if the guess is that it came from Q.

KL divergence

Question What is the interpretation of the above lemma?

Let $f : \mathcal{X} \to \{0, 1\}$ be any classifier. For instance, f outputs 1 if the guess is that the sample point came from P and 0 if the guess is that it came from Q. Then the rate of true positive minus the rate of false positive can be measured as the difference

$$|\mathbb{E}_P[f(x)] - \mathbb{E}_Q[f(x)]|$$

KL divergence

Question What is the interpretation of the above lemma?

Let $f : \mathcal{X} \to \{0, 1\}$ be any classifier. For instance, f outputs 1 if the guess is that the sample point came from P and 0 if the guess is that it came from Q. Then the rate of true positive minus the rate of false positive can be measured as the difference

$$|\mathbb{E}_P[f(x)] - \mathbb{E}_Q[f(x)]|$$

Pinsker's inequality Let P and Q be two distributions defined on $\mathcal{X}$. Then

$$D(P\|Q) \geq \frac{1}{2\ln 2}\|P - Q\|_1^2$$

KL divergence

Question What is the interpretation of the above lemma?

Let $f : \mathcal{X} \to \{0, 1\}$ be any classifier. For instance, f outputs 1 if the guess is that the sample point came from P and 0 if the guess is that it came from Q. Then the rate of true positive minus the rate of false positive can be measured as the difference

$$|\mathbb{E}_P[f(x)] - \mathbb{E}_Q[f(x)]|$$

Pinsker's inequality Let P and Q be two distributions defined on $\mathcal{X}$. Then

$$D(P\|Q) \geq \frac{1}{2\ln 2}\|P - Q\|_1^2$$

Special case Let $\mathcal{X} = \{0, 1\}$ and

$$P = \begin{cases} 1, & \text{wp } p \\ 0, & \text{wp } 1 - p \end{cases} \qquad Q = \begin{cases} 1, & \text{wp } q \\ 0, & \text{wp } 1 - q. \end{cases}$$

KL divergence

Then

$$D(P\|Q) = p\log\frac{p}{q} + (1-q)\log\frac{1-p}{1-q} \text{ and } \|P-Q\|_1 = 2\,|p-q|$$

KL divergence

Then

$$D(P\|Q) = p\log\frac{p}{q} + (1-q)\log\frac{1-p}{1-q} \text{ and } \|P-Q\|_1 = 2\,|p-q|$$

Lemma(Pinsker's inequality for $\mathcal{X} = \{0,1\}$) Let P and Q be distributions as above. Then

$$p\log\frac{p}{q} + (1-q)\log\frac{1-p}{1-q} \geq \frac{2}{\ln 2}(p-q)^2$$

KL divergence

Then

$$D(P\|Q) = p\log\frac{p}{q} + (1-q)\log\frac{1-p}{1-q} \text{ and } \|P-Q\|_1 = 2\,|p-q|$$

Lemma(Pinsker's inequality for $\mathcal{X} = \{0,1\}$) Let P and Q be distributions as above. Then

$$p\log\frac{p}{q} + (1-q)\log\frac{1-p}{1-q} \geq \frac{2}{\ln 2}(p-q)^2$$

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

KL divergence

Then

$$D(P\|Q) = p\log\frac{p}{q} + (1-q)\log\frac{1-p}{1-q} \text{ and } \|P-Q\|_1 = 2\,|p-q|$$

Lemma(Pinsker's inequality for $\mathcal{X} = \{0,1\}$) Let P and Q be distributions as above. Then

$$p\log\frac{p}{q} + (1-q)\log\frac{1-p}{1-q} \geq \frac{2}{\ln 2}(p-q)^2$$

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

$$\frac{\partial f}{\partial q} = -\frac{(p-q)}{\ln 2}\left(\frac{1}{q(1-q)} - 4\right)$$

KL divergence

Then

$$D(P\|Q) = p\log\frac{p}{q} + (1-q)\log\frac{1-p}{1-q} \text{ and } \|P-Q\|_1 = 2\,|p-q|$$

Lemma(Pinsker's inequality for $\mathcal{X} = \{0,1\}$) Let P and Q be distributions as above. Then

$$p\log\frac{p}{q} + (1-q)\log\frac{1-p}{1-q} \geq \frac{2}{\ln 2}(p-q)^2$$

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

$$\frac{\partial f}{\partial q} = -\frac{(p-q)}{\ln 2}\left(\frac{1}{q(1-q)} - 4\right)$$

Since $\frac{1}{q(1-q)} - 4 \geq 0$ for all q, $\frac{\partial f}{\partial q} \leq 0$ when $q \leq p$ and $\frac{\partial f}{\partial q} \geq 0$ when $q \geq p$.

KL divergence

Moreover, $f(p,q) = \infty$ when $q = 0$ and $f(p,q) = 0$ when $q = p$.

KL divergence

Moreover, $f(p,q) = \infty$ when $q = 0$ and $f(p,q) = 0$ when $q = p$. Thus, the function achieves its minimum value at $q = p$ and is always non-negative.

KL divergence

Moreover, $f(p,q) = \infty$ when $q = 0$ and $f(p,q) = 0$ when $q = p$. Thus, the function achieves its minimum value at $q = p$ and is always non-negative.

Lemma Let P and Q be distributions on a finite set $\mathcal{X}$. Then there exist distributions P', Q' on $\{0,1\}$ such that

$$\|P' - Q'\|_1 = \|P - Q\|_1, \text{ and } D(P\|Q) \geq D(P'\|Q')$$

KL divergence

Moreover, $f(p, q) = \infty$ when $q = 0$ and $f(p, q) = 0$ when $q = p$. Thus, the function achieves its minimum value at $q = p$ and is always non-negative.

Lemma Let P and Q be distributions on a finite set $\mathcal{X}$. Then there exist distributions P', Q' on $\{0, 1\}$ such that

$$\|P' - Q'\|_1 = \|P - Q\|_1, \text{ and } D(P\|Q) \geq D(P'\|Q')$$

Proof Let $A \subset \mathcal{X}$ be $A = \{x : p(x) \geq q(x)\}$ and P', Q' be

$$P' = \begin{cases} 1, & \text{wp } \sum_{x \in A} p(x) \\ 0, & \text{wp } \sum_{x \notin A} p(x) \end{cases} \qquad Q' = \begin{cases} 1, & \text{wp } \sum_{x \in A} q(x) \\ 0, & \text{wp } \sum_{x \notin A} q(x). \end{cases}$$

KL divergence

Then

$$
\begin{aligned}
\|P-Q\|_1 &= \sum_{x\in\mathcal{X}} |p(x)-q(x)| \\
&= \sum_{x\in A}(p(x)-q(x)) + \sum_{x\notin A}(q(x)-p(x)) \\
&= \left|\sum_{x\in A} p(x) - \sum_{x\in A} q(x)\right| + \left|\left(1-\sum_{x\in A} p(x)\right) - \left(1-\sum_{x\in A} q(x)\right)\right| \\
&= \|P'-Q'\|_1
\end{aligned}
$$

KL divergence

To calculate the KL-divergence, define a random variable Z as

$$Z = \begin{cases} 1 \text{ if } x \in A \\ 0 \text{ if } x \notin A \end{cases}$$

KL divergence

To calculate the KL-divergence, define a random variable Z as

$$Z = \begin{cases} 1 \text{ if } x \in A \\ 0 \text{ if } x \notin A \end{cases}$$

Since Z is a function of X, the distributions P and Q can be thought of as joint distributions for the random variables (X, Z).

KL divergence

To calculate the KL-divergence, define a random variable Z as

$$Z = \begin{cases} 1 \text{ if } x \in A \\ 0 \text{ if } x \notin A \end{cases}$$

Since Z is a function of X, the distributions P and Q can be thought of as joint distributions for the random variables (X, Z). Besides, the marginal distributions of Z are P' and Q'.

KL divergence

To calculate the KL-divergence, define a random variable Z as

$$Z = \begin{cases} 1 \text{ if } x \in A \\ 0 \text{ if } x \notin A \end{cases}$$

Since Z is a function of X, the distributions P and Q can be thought of as joint distributions for the random variables (X, Z). Besides, the marginal distributions of Z are P' and Q'. Then

$$D(P \| Q) = D(P(X, Z) \| Q(X, Z))$$

KL divergence

To calculate the KL-divergence, define a random variable Z as

$$Z = \begin{cases} 1 \text{ if } x \in A \\ 0 \text{ if } x \notin A \end{cases}$$

Since Z is a function of X, the distributions P and Q can be thought of as joint distributions for the random variables (X, Z). Besides, the marginal distributions of Z are P' and Q'. Then

$$\begin{aligned} D(P\|Q) &= D(P(X,Z)\|Q(X,Z)) \\ &= D(P(Z)\|Q(Z)) + D(P(X|Z)\|Q(X|Z)) \end{aligned}$$

KL divergence

To calculate the KL-divergence, define a random variable Z as

$$Z = \begin{cases} 1 \text{ if } x \in A \\ 0 \text{ if } x \notin A \end{cases}$$

Since Z is a function of X, the distributions P and Q can be thought of as joint distributions for the random variables (X, Z). Besides, the marginal distributions of Z are P' and Q'. Then

$$\begin{aligned} D(P\|Q) &= D(P(X,Z)\|Q(X,Z)) \\ &= D(P(Z)\|Q(Z)) + D(P(X|Z)\|Q(X|Z)) \\ &\geq D(P(Z)\|Q(Z)) \end{aligned}$$

KL divergence

To calculate the KL-divergence, define a random variable Z as

$$Z = \begin{cases} 1 \text{ if } x \in A \\ 0 \text{ if } x \notin A \end{cases}$$

Since Z is a function of X, the distributions P and Q can be thought of as joint distributions for the random variables (X, Z). Besides, the marginal distributions of Z are P' and Q'. Then

$$\begin{aligned} D(P\|Q) &= D(P(X,Z)\|Q(X,Z)) \\ &= D(P(Z)\|Q(Z)) + D(P(X|Z)\|Q(X|Z)) \\ &\geq D(P(Z)\|Q(Z)) \\ &= D(P'\|Q') \end{aligned}$$

Kl divergence

An application of Pinsker's inequality How do you distinguish two coins of slightly different biases?

Kl divergence

An application of Pinsker's inequality How do you distinguish two coins of slightly different biases?

Suppose we are one of the two coins is given to us with the probability distributions

$$P = \begin{cases} 1 \text{ wp } \frac{1}{2} \\ 0 \text{ wp } \frac{1}{2} \end{cases} \quad \text{and } Q = \begin{cases} 1 \text{ wp } \frac{1}{2} + \epsilon \\ 0 \text{ wp } \frac{1}{2} - \epsilon \end{cases}$$

Kl divergence

An application of Pinsker's inequality How do you distinguish two coins of slightly different biases?

Suppose we are one of the two coins is given to us with the probability distributions

$$P = \begin{cases} 1 \text{ wp } \frac{1}{2} \\ 0 \text{ wp } \frac{1}{2} \end{cases} \quad \text{and } Q = \begin{cases} 1 \text{ wp } \frac{1}{2} + \epsilon \\ 0 \text{ wp } \frac{1}{2} - \epsilon \end{cases}$$

$$D(P\|Q) = \frac{1}{2} \log \frac{1/2}{1/2 + \epsilon} + \frac{1}{2} \log \frac{1/2}{1/2 - \epsilon}$$

Kl divergence

An application of Pinsker's inequality How do you distinguish two coins of slightly different biases?

Suppose we are one of the two coins is given to us with the probability distributions

$$P = \begin{cases} 1 \text{ wp } \frac{1}{2} \\ 0 \text{ wp } \frac{1}{2} \end{cases} \quad \text{and } Q = \begin{cases} 1 \text{ wp } \frac{1}{2} + \epsilon \\ 0 \text{ wp } \frac{1}{2} - \epsilon \end{cases}$$

$$\begin{aligned} D(P\|Q) &= \frac{1}{2}\log\frac{1/2}{1/2+\epsilon} + \frac{1}{2}\log\frac{1/2}{1/2-\epsilon} \\ &= \frac{1}{2}\log\frac{1}{1-4\epsilon^2} \end{aligned}$$

Kl divergence

An application of Pinsker's inequality How do you distinguish two coins of slightly different biases?

Suppose we are one of the two coins is given to us with the probability distributions

$$P = \begin{cases} 1 \text{ wp } \frac{1}{2} \\ 0 \text{ wp } \frac{1}{2} \end{cases} \quad \text{and } Q = \begin{cases} 1 \text{ wp } \frac{1}{2} + \epsilon \\ 0 \text{ wp } \frac{1}{2} - \epsilon \end{cases}$$

$$\begin{aligned} D(P\|Q) &= \frac{1}{2}\log\frac{1/2}{1/2+\epsilon} + \frac{1}{2}\log\frac{1/2}{1/2-\epsilon} \\ &= \frac{1}{2}\log\frac{1}{1-4\epsilon^2} \\ &= \frac{1}{2\ln 2}\ln\left(1+\frac{4\epsilon^2}{1-4\epsilon^2}\right) \end{aligned}$$

Kl divergence

An application of Pinsker's inequality How do you distinguish two coins of slightly different biases?

Suppose we are one of the two coins is given to us with the probability distributions

$$P = \begin{cases} 1 \text{ wp } \frac{1}{2} \\ 0 \text{ wp } \frac{1}{2} \end{cases} \quad \text{and } Q = \begin{cases} 1 \text{ wp } \frac{1}{2} + \epsilon \\ 0 \text{ wp } \frac{1}{2} - \epsilon \end{cases}$$

$$\begin{aligned} D(P\|Q) &= \frac{1}{2}\log\frac{1/2}{1/2+\epsilon} + \frac{1}{2}\log\frac{1/2}{1/2-\epsilon} \\ &= \frac{1}{2}\log\frac{1}{1-4\epsilon^2} \\ &= \frac{1}{2\ln 2}\ln\left(1+\frac{4\epsilon^2}{1-4\epsilon^2}\right) \\ &\leq \frac{1}{2\ln 2}\frac{4\epsilon^2}{1-4\epsilon^2} \leq \frac{8\epsilon^2}{2\ln 2}, \text{ using } 1+z \leq e^z, \epsilon \leq \frac{1}{4} \end{aligned}$$

KL divergence

Consider the output of n independent coin tosses.

KL divergence

Consider the output of n independent coin tosses. Then

$$nD(P\|Q) = D(P^n\|Q^n).$$

KL divergence

Consider the output of n independent coin tosses. Then

$$nD(P\|Q) = D(P^n\|Q^n).$$

Suppose there is an algorithm $T(x_1, \ldots, x_n) \to \{0, 1\}$ which outputs 0 if the coin is with distribution P, and 1 if the coin is with distribution Q such that T identifies both coins with probability at least 0.9 i.e.

$$\mathbb{P}_{x \in P^n}[T(x) = 0] \geq \frac{9}{10} \text{ and } \mathbb{P}_{x \in Q^n}[T(x) = 1] \geq \frac{9}{10}$$

KL divergence

Consider the output of n independent coin tosses. Then

$$nD(P\|Q) = D(P^n\|Q^n).$$

Suppose there is an algorithm $T(x_1, \ldots, x_n) \to \{0, 1\}$ which outputs 0 if the coin is with distribution P, and 1 if the coin is with distribution Q such that T identifies both coins with probability at least 0.9 i.e.

$$\mathbb{P}_{x \in P^n}[T(x) = 0] \geq \frac{9}{10} \text{ and } \mathbb{P}_{x \in Q^n}[T(x) = 1] \geq \frac{9}{10}$$

Question Find a lower bound of n without knowing anything about T.

KL divergence

Note that

$$\mathbb{E}_{x \in P^n}[T(x)] \leq \frac{1}{10} \text{ and } \mathbb{E}_{x \in Q^n}[T(x)] \geq \frac{9}{10}$$

KL divergence

Note that

$$\mathbb{E}_{x\in P^n}[T(x)] \leq \frac{1}{10} \text{ and } \mathbb{E}_{x\in Q^n}[T(x)] \geq \frac{9}{10}$$

which gives

$$\mathbb{E}_{x\in Q^n}[T(x)] - \mathbb{E}_{x\in P^n}[T(x)] \geq \frac{8}{10} \Rightarrow \|P^n - Q^n\|_1 \geq \frac{8}{5}$$

KL divergence

Note that

$$\mathbb{E}_{x \in P^n}[T(x)] \leq \frac{1}{10} \text{ and } \mathbb{E}_{x \in Q^n}[T(x)] \geq \frac{9}{10}$$

which gives

$$\mathbb{E}_{x \in Q^n}[T(x)] - \mathbb{E}_{x \in P^n}[T(x)] \geq \frac{8}{10} \Rightarrow \|P^n - Q^n\|_1 \geq \frac{8}{5}$$

Then

$$nD(P \parallel Q) \geq \frac{1}{2\ln 2}\left(\frac{8}{5}\right)^2 \Rightarrow n \geq \frac{1}{2\ln 2 \cdot D(P \parallel Q)}\left(\frac{8}{5}\right)^2 \geq \frac{1}{8\epsilon^2}\left(\frac{8}{5}\right)^2$$

KL divergence

Recall

$$I(X;Y) = \sum_{x,y} p(x,y) \log \frac{p(x,y)}{p(x)p(y)} = D(p(x,y) \| p(x)p(y))$$

[1] Körner, János (1973). "Coding of an information source having ambiguous alphabet and the entropy of graphs". 6th Prague conference on information theory: 411–425.

KL divergence

Recall

$$I(X;Y) = \sum_{x,y} p(x,y) \log \frac{p(x,y)}{p(x)p(y)} = D(p(x,y) \| p(x)p(y))$$

Graph entropy[1] Let $G = (V, E)$. A subset S of the vertices V of an undirected graph $G = (V, E)$ is independent if no edge in the graph has both endpoints in S.

[1] Körner, János (1973). "Coding of an information source having ambiguous alphabet and the entropy of graphs". 6th Prague conference on information theory: 411–425.

KL divergence

Recall

$$I(X;Y) = \sum_{x,y} p(x,y) \log \frac{p(x,y)}{p(x)p(y)} = D(p(x,y) \| p(x)p(y))$$

Graph entropy[1] Let $G = (V, E)$. A subset S of the vertices V of an undirected graph $G = (V, E)$ is independent if no edge in the graph has both endpoints in S.

Define the graph entropy $H(G)$ as

$$\begin{aligned} & \min_{X,Y} I(X;Y) \\ \text{s.t.} \quad & X \text{ is uniformly distributed over } V \\ & Y \text{ is an independent set in } G \text{ containing } X \end{aligned}$$

[1] Körner, János (1973). "Coding of an information source having ambiguous alphabet and the entropy of graphs". 6th Prague conference on information theory: 411–425.

KL divergence

Recall

$$I(X;Y) = \sum_{x,y} p(x,y) \log \frac{p(x,y)}{p(x)p(y)} = D(p(x,y) \| p(x)p(y))$$

Graph entropy[1] Let $G = (V, E)$. A subset S of the vertices V of an undirected graph $G = (V, E)$ is independent if no edge in the graph has both endpoints in S.

Define the graph entropy $H(G)$ as

$$\begin{aligned} & \min_{X,Y} I(X;Y) \\ \text{s.t.} \quad & X \text{ is uniformly distributed over } V \\ & Y \text{ is an independent set in } G \text{ containing } X \end{aligned}$$

Question It is defined in terms of mutual information, why is it called entropy?

[1] Körner, János (1973). "Coding of an information source having ambiguous alphabet and the entropy of graphs". 6th Prague conference on information theory: 411–425.

KL divergence

Let $\mathcal{I}$ denote the independent vertex sets in G. Then we wish to find the joint distribution of (X, Y) on $V \times \mathcal{I}$ with the lowest mutual information such that (i) the marginal distribution of X is uniform and (ii) in samples from the distribution, the Y contains X almost surely. The mutual information of X and Y is then called the entropy of G.

KL divergence

Let $\mathcal{I}$ denote the independent vertex sets in G. Then we wish to find the joint distribution of (X, Y) on $V \times \mathcal{I}$ with the lowest mutual information such that (i) the marginal distribution of X is uniform and (ii) in samples from the distribution, the Y contains X almost surely. The mutual information of X and Y is then called the entropy of G.

Examples

$\rightarrow$ Let K_n denote the complete graph on n vertices. Then

$$H(G) = \log n$$

KL divergence

Let $\mathcal{I}$ denote the independent vertex sets in G. Then we wish to find the joint distribution of (X, Y) on $V \times \mathcal{I}$ with the lowest mutual information such that (i) the marginal distribution of X is uniform and (ii) in samples from the distribution, the Y contains X almost surely. The mutual information of X and Y is then called the entropy of G.

Examples

→ Let K_n denote the complete graph on n vertices. Then

$$H(G) = \log n$$

→ Let G be a bipartite graph, with n_1 vertices on one class and n_2 vertices on the other. Then, for any vertex v, all the vertices in the class which contains v form an independent set containing v. If X is a uniformly random vertex and Y is the set of all vertices which contains X then

$$I(X;Y) \leq H(Y) = \frac{n_1}{n_1+n_2}\log\left(\frac{n_1+n_2}{n_1}\right) + \frac{n_2}{n_1+n_2}\log\left(\frac{n_1+n_2}{n_2}\right) \leq 1$$