ECE 520.674
Information Theoretic Methods in Statistics

Sanjeev Khudanpur

January 29 and February 4, 1999

Calligraphic letters such as $\mathcal{X}$ and $\mathcal{Y}$ shall be used to denote discrete finite sets. P, Q, and R shall be used to denote probability mass functions (pmf's) on either $\mathcal{X}$, $\mathcal{Y}$ or $\mathcal{X} \times \mathcal{Y}$ as clarified at the time of usage. We shall also follow the convention that

$$0 \log 0 = 0, \quad 0 \log \frac{0}{0} = 0, \quad t \log \frac{t}{0} = +\infty,$$

for every $t > 0$. $|A|$ denotes the cardinality of a set A.

The unit simplex $\mathbb{P}^k$ in $\mathbb{R}^k$, the k-dimensional space of real numbers, is the set of points for which

- each of the coordinates is nonnegative, and
- the sum of the coordinates is unity.

Observe that a pmf P on $\mathcal{X}$ can be identified with a point on the unit simplex $\mathbb{P}^{|\mathcal{X}|}$ in $\mathbb{R}^{|\mathcal{X}|}$. Recall, also, that restriction of a topology on $\mathbb{R}$ to $\mathbb{P} \subset \mathbb{R}$ means that $E \subseteq \mathbb{P}$ is considered open in $\mathbb{P}$ iff $\exists\ \tilde{E} \subseteq \mathbb{R}$ such that $\tilde{E} \cap \mathbb{P} = E$, and $\tilde{E}$ is open in $\mathbb{R}$.

For all topological notions, such as open and closed sets of pmf's, convergence of a sequence of pmf's, *etc.*, we shall use the usual Euclidean topology on $\mathbb{R}^{|\mathcal{X}|}$ restricted to the unit simplex.

The Kullback-Leibler distance or the information divergence between two pmf's on $\mathcal{X}$, say P and Q, is defined as

$$D(P\|Q) = \sum_{x \in \mathcal{X}} P(x) \log \frac{P(x)}{Q(x)}.$$

We next establish some properties of I-divergence which are useful in proving several important results.

Properties of I-Divergence

The I-divergence $D(\cdot\|\cdot)$ satisfies the following relationships.

1. **Nonnegativity.** The I-divergence between two pmf's satisfies

$$D(P\|Q) \geq 0,$$

   with equality iff $P = Q$.

2. **Lower semicontinuity.** For a sequence of pmf's (P_n, Q_n), $n = 1, 2, \ldots$, which converges to (P, Q),

$$\liminf_{n\to\infty} D(P_n\|Q_n) \geq D(P\|Q).$$

   If $Q(x) > 0$ for each $x \in \mathcal{X}$, then $D(P\|Q)$ is continuous in the pair (P, Q)

3. **Convexity.** For any $\alpha \in [0, 1]$, and pmf's P_1, Q_1, P_2, Q_2,

$$\alpha D(P_1\|Q_1) + (1-\alpha)D(P_2\|Q_2) \geq D(\alpha P_1 + (1-\alpha)P_2\|\alpha Q_1 + (1-\alpha)Q_2).$$

4. **Partition Inequality.** If $\mathcal{A} = \{A_1, \ldots, A_K\}$ is a partition of $\mathcal{X}$, *i.e.*, $\mathcal{X} = \bigcup_{i=1}^K A_i$, and $i \neq j \Rightarrow A_i \bigcap A_j = \phi$, and we define

$$\begin{aligned} P_{\mathcal{A}}(i) &= \sum_{x\in A_i} P(x), \qquad i = 1, \ldots, K, \\ Q_{\mathcal{A}}(i) &= \sum_{x\in A_i} Q(x), \qquad i = 1, \ldots, K, \end{aligned}$$

   then

$$D(P\|Q) \geq D(P_{\mathcal{A}}\|Q_{\mathcal{A}}),$$

   with equality iff $P(x|x \in A_i) = Q(x|x \in A_i)$, $x \in A_i$, for each i.

5. **Data Processing Inequality.** If W is a $|\mathcal{X}|\times|\mathcal{Y}|$ stochastic matrix (whose rows sum to 1), and we define

$$\begin{aligned} P{\circ}W(x,y) &= P(x)W(y|x), \qquad x \in \mathcal{X},\ y \in \mathcal{Y}, \\ Q{\circ}W(x,y) &= Q(x)W(y|x), \qquad x \in \mathcal{X},\ y \in \mathcal{Y}, \\ PW(y) &= \sum_{x\in\mathcal{X}} P{\circ}W(x,y), \qquad y \in \mathcal{Y}, \\ QW(y) &= \sum_{x\in\mathcal{X}} Q{\circ}W(x,y), \qquad y \in \mathcal{Y}, \end{aligned}$$

   then

$$D(P\|Q) \geq D(PW\|QW),$$

   with equality iff the *a posteriori* probability of x given y is the same for each y under both the joint distributions $P{\circ}W$ and $Q{\circ}W$.

6. **Pinsker's Inequality.** The variational distance between pmf's,

$$d(P,Q) = \sum_{x\in\mathcal{X}} |P(x) - Q(x)|,$$

is bounded above by the I-divergence between the pmf's in the sense that

$$D(P\|Q) \geq \frac{1}{2}d^2(P,Q).$$

7. **Parallelogram Identity.** For pmf's P, Q, and R,

$$D(P\|R) + D(Q\|R) = 2D\left(\frac{P+Q}{2}\|R\right) + D\left(P\|\frac{P+Q}{2}\right) + D\left(Q\|\frac{P+Q}{2}\right).$$

An algebraic inequality which, in turn, is very useful in proving the abovementioned relationships is the log-sum inequality.

Log-Sum Inequality: Let $\{a_i\}_{i=1}^n$, and $\{b_i\}_{i=1}^n$ be sequences of nonnegative numbers. Let $a = \sum_{i=1}^n a_i$ and $b = \sum_{i=1}^n b_i$. Then

$$\sum_{i=1}^n a_i \log\frac{a_i}{b_i} \geq a\log\frac{a}{b},$$

with equality iff $\frac{a_i}{b_i} = c$ for every i, where c is some constant.

Proof: First observe that

- it suffices to prove the inequality for $a_i > 0$: Dropping from consideration any index i for which $a_i = 0$ does not change the left side of the inequality, and can only increase the right side by possibly reducing the sum b.

- it suffices to prove the inequality for $b_i > 0$: Otherwise, the left side is $+\infty$ and there is nothing to prove.

- it suffices to prove the inequality for $a = b$: The inequality is invariant to a scaling of the b_i's because

$$\begin{aligned}\sum_{i=1}^n a_i \log\frac{a_i}{\lambda b_i} &= \sum_{i=1}^n a_i\log\frac{a_i}{b_i} + a_i\log\frac{1}{\lambda}, \text{ and}\\ a\log\frac{a}{\sum\lambda b_i} &= a\log\frac{a}{b} + a\log\frac{1}{\lambda}.\end{aligned}$$

Hence it suffices to show that for $\{a_i\}_{i=1}^n$, and $\{b_i\}_{i=1}^n$ such that $a_i, b_i > 0$ for all i, and $a = b$,

$$\sum_{i=1}^n a_i \log \frac{a_i}{b_i} \geq 0,$$

with equality iff $a_i = b_i$ for every $i = 1, \ldots, n$. Next, recall that

$$\log t \leq t - 1, \qquad \text{for } t > 0,$$

with equality iff $t = 1$. Therefore, setting $t_i = \frac{b_i}{a_i}$, we get

$$\begin{aligned} \sum_{i=1}^n a_i \log \frac{a_i}{b_i} &= -\sum_{i=1}^n a_i \log t_i \\ &\geq -\sum_{i=1}^n a_i (t_i - 1) \\ &= -\sum_{i=1}^n (b_i - a_i), \end{aligned}$$

which proves the desired inequality, together with the condition for equality.

Proofs of the seven properties.

1. The nonnegativity of $D(P\|Q)$ follows trivially from the log-sum inequality, as does the condition for equality.

2. The lower semicontinuity follows from the three cases listed below.

   (a) For each $x \in \mathcal{X}$, $Q(x) > 0$: Since, $Q_n(x) \to Q(x)$, $Q_n(x) > 0$ for all $n > n_0$ for some $n_0(x)$. Therefore

   $$P_n(x) \log \frac{P_n(x)}{Q_n(x)} \to P(x) \log \frac{P(x)}{Q(x)} < +\infty.$$

   and since $D(P_n\|Q_n)$ is a sum of a finite number of such terms, each of which goes to a finite limit,

   $$D(P_n\|Q_n) \to D(P\|Q).$$

   (b) There is an $x' \in \mathcal{X}$ such that $Q(x') = 0$ and $P(x') > 0$: In this case, $D(P\|Q) = +\infty$ and

   $$\begin{aligned} &D(P_n\|Q_n) \\ &= P_n(x') \log \frac{P_n(x')}{Q_n(x')} + \sum_{x \neq x'} P_n(x) \log \frac{P_n(x)}{Q_n(x)} \end{aligned}$$

$$
= \begin{cases} P_n(x') \log \frac{P_n(x')}{Q_n(x')} + 0 & \text{if } P_n(x') = 1, \\ P_n(x') \log \frac{P_n(x')}{Q_n(x')} + \infty & \text{if } P_n(x') < 1 \\ & \& \; Q_n(x') = 1, \\ P_n(x') \log \frac{P_n(x')}{Q_n(x')} + \sum_{x \neq x'} P_n(x) \left[\log \frac{\frac{P_n(x)}{1-P_n(x')}}{\frac{Q_n(x)}{1-Q_n(x')}} + \log \frac{1-P_n(x')}{1-Q_n(x')} \right] & \text{otherwise.} \end{cases}
$$

Since $Q_n(x') \to 0$, it is easy to verify that the first term above goes to $+\infty$, and the second term, in every case, does not go to $-\infty$ and therefore

$$
D(P_n \| Q_n) \to +\infty.
$$

(c) For every $x' \in \mathcal{X}$ such that $Q(x') = 0$, we also have that $P(x') = 0$: In this case

$$
\begin{aligned}
& D(P_n \| Q_n) \\
= & \sum_{x:\ Q(x)>0} P_n(x) \log \frac{P_n(x)}{Q_n(x)} + \sum_{x':\ Q(x')=0} P_n(x') \log \frac{P_n(x')}{Q_n(x')} \\
= & \sum_{x:\ Q(x)>0} P_n(x) \log \frac{P_n(x)}{Q_n(x)} + \sum_{x':\ Q(x')=0} P_n(x') \log P_n(x') - P_n(x') \log Q_n(x') \\
\geq & \sum_{x:\ Q(x)>0} P_n(x) \log \frac{P_n(x)}{Q_n(x)} + \sum_{x':\ Q(x')=0} P_n(x') \log P_n(x')
\end{aligned}
$$

It is again easy to verify that the first term above converges to $D(P\|Q)$, and the second term to zero. Thus

$$
\liminf_{n \to \infty} D(P_n \| Q_n) \geq D(P \| Q).
$$

3. The convexity follows from the log-sum inequality as

$$
\begin{aligned}
& \alpha P_1(x) \log \frac{P_1(x)}{Q_1(x)} + (1-\alpha) P_2(x) \log \frac{P_2(x)}{Q_2(x)} \\
= & \ \alpha P_1(x) \log \frac{\alpha P_1(x)}{\alpha Q_1(x)} + (1-\alpha) P_2(x) \log \frac{(1-\alpha) P_2(x)}{(1-\alpha) Q_2(x)} \\
\geq & \ [\alpha P_1(x) + (1-\alpha) P_2(x)] \log \frac{[\alpha P_1(x) + (1-\alpha) P_2(x)]}{[\alpha Q_1(x) + (1-\alpha) Q_2(x)]}.
\end{aligned}
$$

Therefore

$$
\alpha D(P_1 \| Q_1) + (1-\alpha) D(P_2 \| Q_2) \geq D(\alpha P_1 + (1-\alpha) P_2 \| \alpha Q_1 + (1-\alpha) Q_2).
$$

4. For a partition $\mathcal{A} = \{A_1, \ldots, A_K\}$,

$$
D(P \| Q) = \sum_{x \in \mathcal{X}} P(x) \log \frac{P(x)}{Q(x)} \quad = \quad \sum_{i=1}^{K} \sum_{x \in A_i} P(x) \log \frac{P(x)}{Q(x)}
$$

$$
\begin{aligned}
&\geq \sum_{i=1}^{K} \left[\sum_{x \in A_i} P(x) \right] \log \frac{\sum_{x \in A_i} P(x)}{\sum_{x \in A_i} Q(x)} \\
&= \sum_{i=1}^{K} P_{\mathcal{A}}(i) \log \frac{P_{\mathcal{A}}(i)}{Q_{\mathcal{A}}(i)} \\
&= D(P_{\mathcal{A}} \| Q_{\mathcal{A}}).
\end{aligned}
$$

Observe that the log-sum inequality has been used separately for every i. Therefore, for equality to hold, the condition that must be satisfied for every i is that

$$
\begin{aligned}
\frac{P(x)}{Q(x)} &= c_i, \qquad x \in A_i, \text{ and thus} \\
\sum_{x \in A_i} P(x) &= c_i \sum_{x \in A_i} Q(x), \text{ or} \\
\frac{P(x \in A_i)}{Q(x \in A_i)} &= c_i,
\end{aligned}
$$

which gives the desired necessary and sufficient condition for equality.

5. Note that, for the pmf's on $\mathcal{X} \times \mathcal{Y}$ represented by $P \circ W$ and $Q \circ W$,

$$
\begin{aligned}
D(P \circ W \| Q \circ W) &= \sum_{x \in \mathcal{X}} \left[\sum_{y \in \mathcal{Y}} P(x) W(y|x) \log \frac{P(x) W(y|x)}{Q(x) W(y|x)} \right] \\
&= \sum_{x \in \mathcal{X}} \left[\sum_{y \in \mathcal{Y}} P(x) W(y|x) \right] \log \frac{P(x)}{Q(x)} \\
&= D(P \| Q).
\end{aligned}
$$

Next, the collection of sets $\mathcal{A} = \{A_y, y \in \mathcal{Y}\}$, where $A_y = \bigcup_{x' \in \mathcal{X}} \{(x', y)\}$ is a partition of $\mathcal{X} \times \mathcal{Y}$, and the pmf's corresponding to the partition $\mathcal{A}$ are

$$
\begin{aligned}
(P \circ W)_{\mathcal{A}}(y) &= \sum_{(x', y') \in A_y} P \circ W \ (x', y') = PW(y), \text{ and} \\
(Q \circ W)_{\mathcal{A}}(y) &= \sum_{(x', y') \in A_y} Q \circ W \ (x', y') = QW(y).
\end{aligned}
$$

From the partition inequality, it follows that

$$
D(P \circ W \| Q \circ W) \geq D(PW \| QW).
$$

Furthermore, the condition for equality in the partition inequality is that for every y, $P \circ W$ and $Q \circ W$ must satisfy

$$
P \circ W \ (x, y | (x, y) \in A_y) = Q \circ W \ (x, y | (x, y) \in A_y),
$$

which is easily seen to be the same as following condition on the posterior probability of x given y:

$$
P \circ W \ (x|y) = Q \circ W \ (x|y), \qquad x \in \mathcal{X}, \ y \in \mathcal{Y}.
$$

6. Let $\mathcal{A} = \{A_1, A_2\}$, where $A_1 = \{x : \ P(x) \geq Q(x)\}$ and $A_2 = \{x : \ P(x) < Q(x)\}$. First, note that

$$\begin{aligned} d(P,Q) &= \sum_{x\in\mathcal{X}} |P(x) - Q(x)| \\ &= \sum_{x\in A_1} (P(x) - Q(x)) - \sum_{x\in A_2} (P(x) - Q(x)) \\ &= (P_{\mathcal{A}}(1) - Q_{\mathcal{A}}(1)) + (Q_{\mathcal{A}}(2) - P_{\mathcal{A}}(2)) \\ &= d(P_{\mathcal{A}}, Q_{\mathcal{A}}), \end{aligned}$$

and the partition inequality assures us that

$$D(P||Q) \geq D(P_{\mathcal{A}}||Q_{\mathcal{A}}).$$

Therefore, it suffices to prove that

$$D(P_{\mathcal{A}}||Q_{\mathcal{A}}) \geq \frac{1}{2} d^2(P_{\mathcal{A}}, Q_{\mathcal{A}}),$$

i.e., it suffices to prove Pinsker's inequality for the case $|\mathcal{X}| = 2$. Now, for $\mathcal{X} = \{0,1\}$, let $P = (p, 1-p)$ and $Q = (q, 1-q)$, and consider

$$g(q) = p \log \frac{p}{q} + (1-p)\log\frac{1-p}{1-q} - 4c(p-q)^2$$

as a function of q for fixed values of c, p. Note that $g(p) = 0$, and as long as $q \neq 0$ or 1,

$$g_c'(q) = -\frac{p}{q} + \frac{1-p}{1-q} + 8c(p-q) = (q-p)\left[\frac{1}{(1-q)q} - 8c\right].$$

Since $q(1-q) \leq \frac{1}{4}$, the choice of $c \leq \frac{1}{2}$ guarantees that $g(q)$ achieves a minimum at $q = p$. Therefore, if $c \leq \frac{1}{2}$, then

$$\begin{aligned} g_c(q) &= D(P||Q) - c\left(|p-q| + |(1-p) - (1-q)|\right)^2 \\ &= D(P||Q) - cd^2(P,Q) \\ &\geq 0 \end{aligned}$$

Setting $c = \frac{1}{2}$ yields Pinsker's inequality.

7. The parallelogram identity follows from direct algebraic manipulation of the right side as

$$\begin{aligned} &2D\left(\frac{P+Q}{2}||R\right) + D\left(P||\frac{P+Q}{2}\right) + D\left(Q||\frac{P+Q}{2}\right) \\ &= \sum_{x\in\mathcal{X}} 2\frac{P(x)+Q(x)}{2}\log\frac{\frac{P(x)+Q(x)}{2}}{R(x)} + P(x)\log\frac{P(x)}{\frac{P(x)+Q(x)}{2}} + Q(x)\log\frac{Q(x)}{\frac{P(x)+Q(x)}{2}} \end{aligned}$$

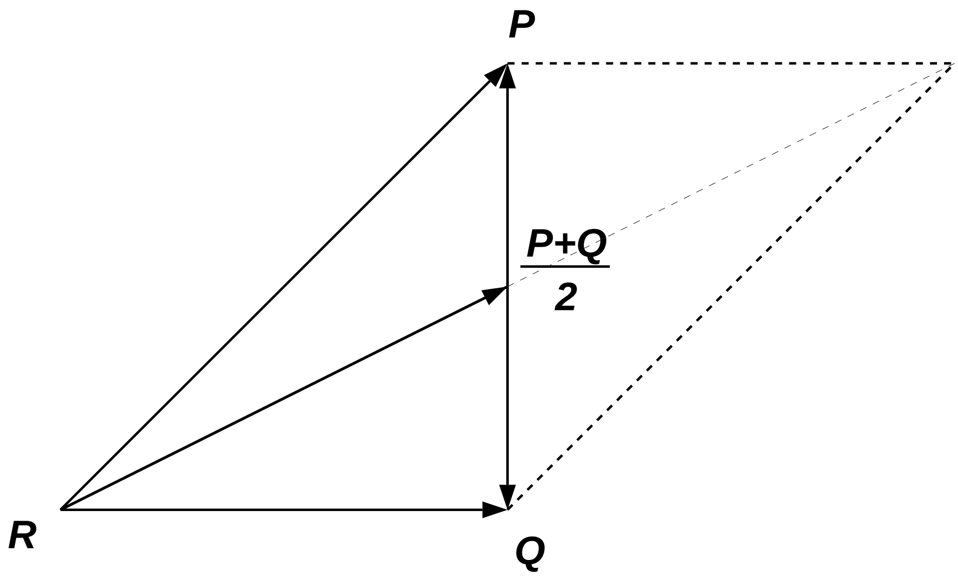

Figure 1: Parallelogram Identity for I-divergences.

$$
\begin{aligned}
&= \sum_{x \in \mathcal{X}} P(x) \log \frac{\frac{P(x)+Q(x)}{2}}{R(x)} \frac{P(x)}{\frac{P(x)+Q(x)}{2}} + Q(x) \log \frac{\frac{P(x)+Q(x)}{2}}{R(x)} \frac{Q(x)}{\frac{P(x)+Q(x)}{2}} \\
&= \sum_{x \in \mathcal{X}} P(x) \log \frac{P(x)}{R(x)} + Q(x) \log \frac{Q(x)}{R(x)} \\
&= D(P\|R) + D(Q\|R).
\end{aligned}
$$

To see a geometric analogue of the parallelogram identity, let $\|P - Q\|$ represent the usual Euclidean (L_2) distance between two pmf's P and Q, and note from the cosine rule for the sides of a triangle that

$$
\begin{aligned}
&\|P - R\|^2 \\
&\quad = \left\|P - \frac{P+Q}{2}\right\|^2 + \left\|\frac{P+Q}{2} - R\right\|^2 - 2\left\|P - \frac{P+Q}{2}\right\| \cdot \left\|\frac{P+Q}{2} - R\right\| \cdot \cos\theta_1, \\
&\|Q - R\|^2 \\
&\quad = \left\|Q - \frac{P+Q}{2}\right\|^2 + \left\|\frac{P+Q}{2} - R\right\|^2 - 2\left\|Q - \frac{P+Q}{2}\right\| \cdot \left\|\frac{P+Q}{2} - R\right\| \cdot \cos\theta_2,
\end{aligned}
$$

where the angles θ_1 and θ_2 are complimentary, as seen in Figure 6. Also, since $\left\|Q - \frac{P+Q}{2}\right\| = \left\|P - \frac{P+Q}{2}\right\|$,

$$
\|P - R\|^2 + \|Q - R\|^2 = 2\left\|\frac{P+Q}{2} - R\right\|^2 + \left\|P - \frac{P+Q}{2}\right\|^2 + \left\|Q - \frac{P+Q}{2}\right\|^2,
$$

which is strikingly similar to the corresponding relationship between the I-divergences. Thus the I-divergence between two pmf's may be thought of as a squared Euclidean distance between them. Note, however, that the analogy must be used only as a guide for understanding, and $D(P\|Q)$ should not be confused with a metric. In particular, the I-divergence is not symmetric and does not satisfy the triangle inequality, *i.e.*, in general

$$\begin{aligned} D(P\|Q) &\neq D(Q\|P) \\ D(P\|Q) + D(Q\|R) &\not> D(P\|R). \end{aligned}$$

Even a symmetric definition based on I-divergences, such as

$$\Delta(P, Q) = \frac{1}{2}\left(D(P\|Q) + D(Q\|P)\right)$$

fails to satisfy the triangle inequality, and thus it is difficult to construct a *bona fide* metric using I-divergences.