Information-theoretic inequalities and Correlated sampling of a one-bit message

Contents

Information-processing and log-sum inequalities

1 Lemma (log-sum inequality). For any pair of sequences $p_1, \ldots, p_n$ and $q_1, \ldots, q_n$ of positive real numbers, we have

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

where $p = \sum_i p_i$ and $q = \sum_i q_i$.

Proof. The inequality is equivalent to

$$\sum_{i=1}^{n} p_i \log \frac{qp_i}{pq_i} \geq 0.$$

But since $\log \frac{1}{x} \geq 1 - x$ for all positive x, and $\frac{\lambda q_i}{p_i}$ is positive, given that p_i, q_i and p/q are positive, then:

$$\sum_{i=1}^{n} p_i \log \frac{p_i}{\lambda q_i} \geq \sum_{i=1}^{n} p_i \left(1 - \frac{pq_i}{qp_i}\right) = (p - \frac{p}{q}q) = 0. \qquad \blacksquare$$

2 *Information Processing Inequality.* For any f,

$$D_{\mathrm{KL}}(f(X), f(Y)) \leq D_{\mathrm{KL}}(X, Y).$$

Proof. By using the log-sum inequality, we derive:

$$\begin{aligned} D_{\mathrm{KL}}(X,Y) &= \sum_{w\in\mathcal{X}} P_X(w)\log\frac{P_X(w)}{P_Y(w)} \\ &= \sum_{i\in f(\mathcal{X})}\sum_{w\in f^{-1}(i)} P_X(w)\log\frac{P_X(w)}{P_Y(w)} \\ &\geq \sum_{i\in f(\mathcal{X})} P_{f(X)}(i)\log\frac{P_{f(X)}(i)}{P_{f(Y)}(i)} \\ &= D_{\mathrm{KL}}(f(X),f(Y)) \end{aligned}$$

■

3 Corollary. For any f, $I(X:Y)\geq I(f(X):Y)$.

Pinsker's inequality

4 *Δ vs D_{KL} — Pinsker's inequality.*

$$\|X-Y\|_1 \leq \sqrt{2D_{\mathrm{KL}}(X\,\|\,Y)} \quad \text{i.e.,} \quad \frac{1}{2}\|X-Y\|_1^2 \leq D_{\mathrm{KL}}(X\,\|\,Y)$$

Proof. Let us first prove it when X,Y are distributions over one bit. Let $p=\mathsf{Pr}[X=0]$, $q=\mathsf{Pr}[Y=0]$. Define

$$g(q)=D_{\mathrm{KL}}(X\,\|\,Y)-\frac{1}{2}\|X-Y\|_1^2 = p\log\frac{p}{q}+(1-p)\log\frac{1-p}{1-q}-2(p-q)^2.$$

Then g' is

$$\begin{aligned} g'(q) &= -\frac{p}{q}+\frac{1-p}{1-q}+4(p-q)=\frac{(1-p)q-p(1-q)}{q(1-q)}-4(q-p) \\ &= (q-p)\left[\frac{1}{q(1-q)}-4\right] \end{aligned}$$

The second factor is always non-negative. Hence $g'(q)$ is negative for $q<p$, positive for $q>p$, and 0 for $q=p$. Hence $q=p$ is a minimum for g, but $g(p)=0$, so g is non-negative.

If X,Y are not one bit, then define $f(w)=1$ if $P_X(w)\leq P_Y(w)$ and $f(w)=0$ otherwise. Then by what we just proved,

$$D_{\mathrm{KL}}(f(X)\,\|\,f(Y)\geq\frac{1}{2}\|f(X)-f(Y)\|_1^2.$$

But also:

$$\begin{aligned}
\|X - Y\|_1 &= \sum_w |P_X(w) - P_Y(w)| \\
&= \sum_{w \in f^{-1}(0)} (P_X(w) - P_Y(w)) + \sum_{w \in f^{-1}(1)} (P_Y(w) - P_X(w)) \\
&= \mathsf{Pr}[f(X) = 0] - \mathsf{Pr}[f(Y) = 0] + \mathsf{Pr}[f(Y) = 1] - \mathsf{Pr}[f(X) = 1] \\
&= \|f(X) - f(Y)\|_1
\end{aligned}$$

The result for any (not-necessarily 1-bit) distribution now follows from the information-processing inequality $D_{\mathrm{KL}}(f(X) \,\|\, f(Y)) \leq D_{\mathrm{KL}}(X \,\|\, Y)$. ∎

Correlated sampling of a one-bit message

5 *Correlated sampling of a one-bit message.* Suppose Alice has input X and Bob Y. Alice wants to send a 1-bit message $M = M(X, R_a)$ to Bob, and this message reveals little information about X to Bob, i.e. $I = I(M : X|Y)$ is close to zero. Let us show how to do this with zero communication, and error probability $\sqrt{\frac{1}{2}I}$ (which is also close to zero, if not quite as close as I).

6 *How Alice samples M.* We can think of M as being sampled in the following way. To each possible input $X = x$ corresponds a value $p_x = \mathsf{Pr}[M = 0|X = x]$. Alice will pick a uniformly-random real-number $v \in [0, 1]$, and set $M = 0$ if $v \leq p_x$ and set $M = 1$ if $v > p_x$.

7. Bob doesn't know p_X because he doesn't know X, but to the extent that X and Y are correlated, Bob will have some estimate of what X is, and hence some estimate for p_x. His best guess for p_x is the value

$$q_y = \mathsf{Pr}[M = 0|Y = y] = \mathop{\mathsf{E}}_{X|Y=y}[q_x] \,.$$

How close are q_y and q_x? It turns out that because $I(X : M|Y)$ is small, we can expect them to be pretty close.

8. Indeed, the distributions of M when one knows x versus M when one knows only y are close, in terms of KL-divergence, because:

$$I = I(X : M|Y) = \mathop{\mathsf{E}}_{Y}\left[\mathop{\mathsf{E}}_{X}\left[D_{\mathrm{KL}}(M|_{x,y} \,\|\, M|_y)\right]\right]$$

(and we think of I as being small). Because $M = M(x)$, it follows that $M|_{x,y} = M|_x$. Let us define

$$I_{x,y} = D_{\mathrm{KL}}(M|_x \,\|\, M|_y)$$

9. Now from the Pinsker inequality, it follows that

$$\sqrt{2I_{x,y}} \geq \|M|_x - M|_y\|_1 = 2|p_x - q_y|.$$

And so $|p_x - q_y| \leq \sqrt{\frac{1}{2}I_{x,y}}$, which is small on average.

10 *The correlated sampling protocol.* So here is a strategy for jointly sampling the bit M without communication: Alice and Bob use shared randomness to sample v, Alice chooses M as before, and Bob assumes that $M = 0$ if $v \leq q_y$, and that $M = 1$ otherwise.

The only case when he is wrong is when v happens to be greater than p_x but smaller than q_y (if $p_x \leq q_y$, or the other way around if $p_x > q_y$). So he will be wrong, on inputs x, y, with probability exactly $|p_x - q_y|$.

Over the input distributions X and Y, the probability that Bob is wrong about M is

$$\mathop{\mathsf{E}}_{X,Y}[|p_x - q_y|] \leq \mathop{\mathsf{E}}_{X,Y}\left[\sqrt{2I_{x,y}}\right] \leq \sqrt{2\mathop{\mathsf{E}}_{X,Y}[I_{x,y}]} = \sqrt{2I},$$

where the last inequality follows from the concavity of the square-root.