Electrical Engineering 229A Lecture 4 Notes

Daniel Raban

September 7, 2021

1 Convexity of Relative Entropy and the Data Processing Inequality

1.1 Chain rules for entropy, relative entropy, and mutual information

The chain rule for entropy for two random variables says that

$$H(X_1, X_2) = H(X_1) + H(X_2 \mid X_1)$$

For n variables, we have

$$\begin{aligned}
H(X_1^n) &= H(X_1^{n-1}, X_n) \\
&= H(X_1^{n-1}) + H(X_n \mid X_1^{n-1}) \\
&\vdots \\
&= H(X_1) + H(X_2 \mid X_1) + \cdots + H(X_n \mid X_1^{n-1}),
\end{aligned}$$

which we can write as

$$= \sum_{\ell=1}^{n} H(X_\ell \mid X_1^{\ell-1}).$$

Here, the convention is that $X_1^{\ell-1}$ for $\ell = 1$ needs no conditioning.

This also comes from

$$\begin{aligned}
H(X_1^n) &= \mathbb{E}\left[\log \frac{1}{\prod_{\ell=1}^{n} p(X_\ell \mid X_1^{\ell-1})}\right] \\
&= \sum_{\ell=1}^{n} \mathbb{E}\left[\log \frac{1}{p(X_\ell \mid X_1^{\ell-1})}\right] \\
&= \sum_{\ell=1}^{n} H(X_\ell \mid X_1^{\ell-1}).
\end{aligned}$$

Similarly, we can obtain the chain rule for relative entropy from

$$\begin{aligned}
D(p(x_1^n) \mid\mid q(x_1^n)) &= \mathbb{E}_p\left[\log \frac{p(X_1^n)}{q(X_1^n)}\right] \\
&= \mathbb{E}_p\left[\log \frac{\prod_{\ell=1}^n p(X_\ell \mid X_1^{\ell-1})}{\prod_{\ell=1}^n q(X_\ell \mid X_1^{\ell-1})}\right] \\
&= \sum_{\ell=1}^n \mathbb{E}_p\left[\log \frac{p(X_\ell \mid X_1^{\ell-1})}{p(X_\ell \mid X_1^{\ell-1})}\right] \\
&= \sum_{\ell=1}^n D(p(x_\ell \mid x_1^{\ell-1}) \mid\mid q(x_\ell \mid x_1^{\ell-1}) \mid p(x_1^{\ell-1})).
\end{aligned}$$

We can also obtain the chain rule for mutual information:

$$I(X; Y_1, Y_2) = I(X; Y_1) + I(X; Y_2 \mid Y_1).$$

This comes from

$$\begin{aligned}
\mathbb{E}\left[\log \frac{p(X, Y_1)}{p(X)p(Y_1, Y_2)}\right] &= \mathbb{E}\left[\frac{p(X, Y_1)}{p(X)p(Y_1)} \frac{p(X, Y_1, Y_2)p(Y_1)p(Y_2)}{p(Y_1)p(X, Y_1)p(Y_2, Y_1)}\right] \\
&= \mathbb{E}\left[\log \frac{p(X, Y_1)}{p(X)p(Y_1)} \frac{p(X, Y_2 \mid Y_1)}{p(X \mid Y_1)p(Y_2 \mid Y_1)}\right],
\end{aligned}$$

More generally,

$$\begin{aligned}
I(X; Y_1^n) &= I(X; Y_1^{n-1}, Y_n) \\
&= I(X; Y_1^{n-1}) + I(X; Y_n \mid Y_1^{n-1}) \\
&\vdots \\
&= I(X; Y_1) + I(X; Y_2 \mid Y_1) + \cdots + I(X; Y_n \mid Y_1^{n-1}),
\end{aligned}$$

which we can write as

$$= \sum_{\ell=1}^n I(X; Y_\ell \mid Y_1^{\ell-1}).$$

1.2 Convexity of relative entropy and the log-sum inequality

An important property of relative entropy $D(p \mid\mid q)$ is that it is convex in the pair (p, q), where p denotes $(p(x), x \in \mathscr{X})$ and q denotes $(q(x), x \in \mathscr{X})$. That is for all $(p_0, q_0), (p_1, q_1)$ and $\lambda \in [0, 1]$, if we denote $p_\lambda = \lambda p_1 + (1 - \lambda)p_0$ and $q_\lambda = \lambda q_1 + (1 - \lambda)q_0$, then

$$D(p_\lambda \mid\mid q_\lambda) \leq \lambda D(p_1 \mid\mid q_1) + (1 - \lambda)D(p_0 \mid\mid q_0.$$

Remark 1.1. Note that $D(p \,||\, q)$ can take the value $+\infty$.

This is a consequence of the **log-sum inequality**:

Lemma 1.1 (log-sum inequality)**.** *Suppose $a_i, b_i > 0$ for $i \in \mathscr{X}$.*

$$\sum_{i \in \mathscr{X}} a_i \log \frac{a_i}{b_i} \geq a \log \frac{a}{b},$$

where $a = \sum_{i \in \mathscr{X}} a_i$ and $b = \sum_{i \in \mathscr{X}} b_i$.

Proof. This comes from the convexity of $u \log u$ for $u \geq 0$. The left hand side is

$$\begin{aligned}
\sum_{i \in \mathscr{X}} a_i \log \frac{a_i}{b_i} &= \sum_{i \in \mathscr{X}} \frac{b_i}{b} \left(\frac{a_i}{b_i} \log \frac{a_i}{b_i} \right) \\
&\geq b \sum_i \frac{b_i}{b} \frac{a_i}{b_i} \log \left(\sum_i \frac{b_i}{b} \frac{a_i}{b_i} \right) \\
&= a \log \frac{a}{b}.
\end{aligned}$$

□

Corollary 1.1. *$D(p \,||\, q)$ is convex in the pair (p, q).*

Proof.

$$\begin{aligned}
\lambda D(p_1 \,||\, q_1) + (1-\lambda) D(p_0 \,||\, q_0) &= \sum_x \lambda p_1(x) \log \frac{p_1(x)}{q_1(x)} + (1-\lambda) p_0(x) \log \frac{p_0(x)}{q_0(x)} \\
&= \sum_x \lambda p_1(x) \log \frac{\lambda p_1(x)}{\lambda q_1(x)} + (1-\lambda) p_0(x) \log \frac{(1-\lambda) p_0(x)}{(1-\lambda) q_0(x)}
\end{aligned}$$

Using the log-sum inequality,

$$\begin{aligned}
&\geq \sum_x (\lambda p_1(x) + (1-\lambda) p_0(x)) \log \frac{\lambda p_1(x) + (1-\lambda) p_0(x)}{\lambda p_1(x) + (1-\lambda) p_0(x)} \\
&= D(p_\lambda \,||\, q_\lambda).
\end{aligned}$$

□

Remark 1.2. The inequality is still true if any of the terms $= +\infty$.

A good book on convex functions is the book by Rockefeller.

1.3 The data processing inequality

The data processing inequality says that if you are looking at the mutual information between X and Y and then you process Y in a way that does not use X, the mutual information can only decrease. How do we make this notion precise?

Definition 1.1. Given 3 random variables X, Y, Z, we write $Y - X - Z$ to indicate that Y and Z are conditionally independent given X. We may say that they form a **Markov chain** in this order. In probability notation, we may use the notation $Y \amalg_X Z$.

Recall that conditional independence says that $p(y, z \mid x) = p(y \mid x)p(z \mid x)$. Since

$$p(y, z \mid x) = p(y \mid x, z)p(z \mid x),$$

the assumed conditional independence gives

$$p(y \mid x, z) = p(y \mid x).$$

This argument can be run backwards, hence the "Markov" terminology.

Remark 1.3. Running the argument in the other direction gives $p(z \mid x, y) = p(z \mid x)$ if $Y - X - Z$.

Theorem 1.1 (Data processing inequality)**.** *Suppose $Y - X - Z$ form a Markov chain. Then*

$$I(Y; Z) \leq I(Y; X).$$

Proof. Use the chain rule in two different orders:

$$I(Y; X, Z) = I(Y; X) + I(Y; Z \mid X),$$

$$I(Y; X, Z) = I(Y; Z) + I(Y; X \mid Z).$$

Because $Y \amalg_X Z$, $I(Y; Z \mid X) = 0$. In fact, each $I(Y; Z \mid X = x)$ equals 0. So

$$I(Y; X) \geq I(Y; Z),$$

as desired. □

Remark 1.4. The condition for equality is $I(Y; X \mid Z) = 0$, i.e. $Y \amalg_Z X$. This has interesting implications in statistics. Say we try to find an estimate for a random variable Θ (in a Bayesian framework) based on observations X. We might ask for some function $T(X)$ such that $\Theta - X - T(X)$. When is it true that $I(\Theta; T(X)) = I(\Theta; X)$? This happens precisely when $\Theta - T(X) - X$.

A typical example (not in a discrete context) is when Θ is the mean of the marginal, where each marginal is normal with variance 1. So conditioned on $\Theta = \theta$, each $X_i \sim N(0, 1)$ for $1 \leq i \leq n$. If $T(X) = \frac{1}{n}\sum_i X_i$, then $\Theta - T(X) - X$. By the data processing inequality, we should study $T(X)$ instead of X in a statistical context because it contains at least as much information as X in terms of estimating Θ.