COMP760, SUMMARY OF LECTURE 13.

HAMED HATAMI

1. Mutual information

Let's start from a simple example. Let $B_1, \ldots, B_6$ be independent random bits, i.e. independent Bernoulli random variables with parameter $\frac{1}{2}$. Let $X = (B_1, B_2, B_3, B_4)$, and $Y = (B_2, B_3, B_4, B_5, B_6)$. Then obviously

$$H(X) = 4 \qquad \text{and} \qquad H(Y) = 5.$$

On the other hand

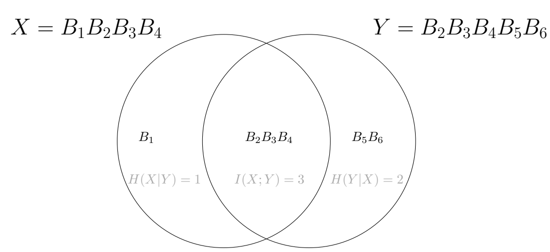
$$H(XY) = 6,$$

as XY is determined by the six random variables $B_1, \ldots, B_6$.

- $H(X|Y) = H(XY) - H(Y) = 1$, so the amount of information left in X after we know Y is 1. We just need to know B_1 and Y to fully recover X.
- $H(Y|X) = H(XY) - H(X) = 2$, so the amount of information left in Y after we know X is 2. We just need to know B_6, B_6 and X to fully recover Y.

Note that by knowing either X or Y we can learn the value of the three independent bits (B_2, B_3, B_4). In other words, we can think of these two bits as the shared information between X and Y. The mutual information $I(X;Y)$ between X and Y is the amount of information that one can learn about X knowing Y, and it turns out this is equal to the amount of the information that one can learn about Y knowing X. This corresponds to the amount of shared information between X and Y. This is demonstrated in Figure 1.

Figure 1. A Venn diagram showing the mutual information between two variables.



Now let us formally define the notion of mutual information.

Definition 1 (Mutual information). *The mutual information between two variables X and Y is defined as*

$$\begin{aligned} I(X;Y) = I(Y;X) &= H(X) - H(X|Y) \\ &= H(Y) - H(Y|X) \\ &= H(X) + H(Y) - H(XY). \end{aligned}$$

By subadditivity of entropy, $I(X;Y) \geq 0$.

Example 2. Let $B_1, \dots, B_5$ be independent random bits, and let $X = (B_1, B_2, B_3)$ and $Y = (B_1 \oplus B_2, B_2 \oplus B_4, B_3 \oplus B_4, B_5)$. Note that the distribution of Y is uniform on $\{0,1\}^4$ as it can be easily seen that its coordinates are mutually independent. Hence obviously

$$H(X) = 3 \qquad \text{and} \qquad H(Y) = 4.$$

On the other hand

$$H(XY) = 5,$$

as XY is determined by the five random variables $B_1, \dots, B_5$.

- $H(X|Y) = H(XY) - H(Y) = 1$, so the amount of information left in X after we know Y is 1. For example we just need to know B_1 and Y to fully recover X.
- $H(Y|X) = H(XY) - H(X) = 2$, so the amount of information left in Y after we know X is 2. For example we just need to know B_4, B_5 and X to fully recover Y.

We have $I(X;Y) = H(X) - H(X|Y) = H(Y) - H(Y|X) = H(X) + H(Y) - H(XY) = 2$. Note that by knowing either X or Y we can learn the value of the two (independent) bits $(B_1 \oplus B_2, B_2 \oplus B_3)$. In other words, we can think of these two bits as the shared information between X and Y. ■

Remark 3. Note that the Venn diagram of Figure 1 has its limitations. For example if X, Y, Z are random bits conditioned on $X \oplus Y \oplus Z = 0$, then they are pairwise independent while for example $I(XY; Z) = 1$. Note that we cannot use a Venn diagram to ilustrate this. ■

We can similarly define the conditional mutual information

Definition 4 (Mutual information). *The mutual information between two variables X and Y conditioned on Z is defined as*

$$\begin{aligned} I(X;Y|Z) &= \mathbb{E}_z I(X;Y|Z=z) \\ &= H(X|Z) - H(XY|Z) = H(Y|Z) - H(Y|XZ) \\ &= H(X|Z) + H(Y|Z) - H(XY|Z) \geq 0. \end{aligned}$$

Recall that X and Y are independent if and only if $H(X) = H(X|Y)$. This leads to the following remark.

Remark 5. Note that X and Y are independent if and only if $I(X;Y) = 0$, and similarly X and Y are independent conditioned on Z if and only if $I(X, Y|Z) = 0$. ■

Example 6. Note that conditioning can increase the mutual information. For example if X, Y, Z are random uniform bits conditioned on $X \oplus Y \oplus Z = 0$, then $I(X;Y) = 0$ while $I(X;Y|Z) = 1$ as after knowing Z the value of Y is determined by the value of X. ■

Theorem 7 (Chain Rule). *We have*

$$I(XY; Z) = I(X; Z) + I(Y; Z|X).$$

Proof.

$$I(XY;Z) = H(Z) - H(Z|XY) = H(Z) - H(Z|X) + H(Z|X) - H(Z|XY) = I(X;Z) + I(Y;Z|X).$$

□

The chain rule says that the amount of information that Z shares with XY equals to the amount of information that Z shares with X plus the amount of information that Z shares with Y once one knows X.

Remark 8. The non-negativity of the mutual information is very useful. For example to prove the intuitively obvious fact $I(X;Y) \leq I(X;YZ)$, one notes that $I(X;YZ) = I(X;Y) + I(X;Z|Y) \geq I(X;Y)$. ■

Example 9. Let $X \to Y \to Z$ be a Markov chain. Then since $I(X;Z|Y) = 0$, we have

$$I(X;Z) \leq I(X;YZ) = I(X;Y) + I(X;Z|Y) = I(X;Y),$$

as it is expected. ■

Consider random variables X, Y with joint probability distribution $p(x,y)$. We can write $p(x,y) = p(x)p(y|x)$, where $p(x) = \Pr[X = x]$ and $p(y|x) = \Pr[Y = y|X = x]$.

Theorem 10. *Consider random variables X, Y with joint distribution $p(x,y)$. Suppose $p(x) = \alpha(x)$ and $p(y|x) = \beta(x,y)$. Then $I(X;Y)$ is concave in α and convex in β.*

Proof. **Convexity with respect to α:** Suppose $(X_1,Y_1) \sim (\alpha_1,\beta)$ and $(X_2,Y_2) \sim (\alpha_2,\beta)$, and $(X,Y) \sim (\lambda\alpha_1 + (1-\lambda)\alpha_2, \beta)$. To sample (X,Y) we use a Bernoulli random variable B with parameter λ: If $B = 1$, then we sample (X,Y) using (α_1,β) and otherwise we use (α_2,β). Note that conditioned on $X = x$, Y is sampled according to the function β regardless of the value of B. In other words, conditioned on X, the random variables B and Y are independent: $I(B;Y|X) = 0$. Hence

$$I(BX;Y) = I(X;Y) + I(B;Y|X) = I(X;Y).$$

On the other hand

$$I(BX;Y) = I(B;Y) + I(X;Y|B) \geq I(X;Y|B) = \lambda I(X_1;Y_1) + (1-\lambda)I(X_1;Y_1).$$

This shows

$$I(X;Y) \geq \lambda I(X_1;Y_1) + (1-\lambda)I(X_1;Y_1),$$

as desired.

Concavity with respect to β: Suppose $(X_1,Y_1) \sim (\alpha,\beta_1)$ and $(X_2,Y_2) \sim (\alpha,\beta_2)$, and $(X,Y) \sim (\alpha, \lambda\beta_1 + (1-\lambda)\beta_2)$. To sample (X,Y) we use a Bernoulli random variable B with parameter λ: If $B = 1$, then we sample (X,Y) using (α,β_1) and otherwise we use (α,β_2). Now X and B are independent: $I(X,B) = 0$. Hence

$$I(Y,X) \leq I(BY,X) = I(B,X) + I(Y,X|B) = \lambda I(X_1;Y_1) + (1-\lambda)I(X_1;Y_1).$$

□

1.1. **Some useful inequalities.** The following inequalities concern the case where A and C are independent, and the case where A and C are independent conditioned on B. Note that using

$$I(AB;C) = I(A;C) + I(A;B|C) = I(A;B) + I(A;C|B)$$

we obtain

$$I(A;B) = I(A;B|C) + I(A;C) - I(A;C|B).$$

This shows

$$\begin{aligned} I(A;C) = 0 &\implies I(A;B) \leq I(A;B|C) \\ I(A;C|B) = 0 &\implies I(A;B) \geq I(A;B|C) \\ I(A;C) = I(A;C|B) = 0 &\implies I(A;B) = I(A;B|C). \end{aligned}$$

Remark 11. As we saw earlier if A, B, C are uniform random bits conditioned on $A \oplus B \oplus C = 0$, then $I(A;C) = 0$ and $0 = I(A;B) < I(A;B|C) = 1$. So the first inequality can be strict.

Also A, B, C are random variables that satisfy $B = C$, then $I(A;C|B) = 0$ and also $I(A;B|C) = 0$. So in this case, the second inequality becomes strict if $I(A;B) > 0$.

Further note that the condition $I(A;C) = I(A;C|B) = 0$ is weaker than $I(AB;C) = 0$. Obviously if C is independent from AB, then the chain rule implies that $I(A;C) = I(A;C|B) = 0$. ■

We can also obviously condition all those inequalities on a fourth random variable Z. Let us summarize this as the following theorem which we shall use frequently.

Theorem 12. *Let A, B, C, Z be random variables. Then*

$$I(A;B|Z) = I(A,B|ZC) + I(A;C|Z) - I(A,C|ZB).$$

which shows

$$\begin{aligned} I(A;C|Z) = 0 &\implies I(A;B|Z) \leq I(A;B|CZ) \\ I(A;C|BZ) = 0 &\implies I(A;B|Z) \geq I(A;B|CZ) \\ I(A;C|Z) = I(A;C|BZ) = 0 &\implies I(A;B|Z) = I(A;B|CZ). \end{aligned}$$

1.2. **Information processing inequality.** Suppose that X, Y, Z are random variables, and f is a function. Then

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

Indeed

$$I(f(X);Y|Z) \leq I(Xf(X);Y|Z) = I(X;Y|Z),$$

as $Xf(X)$ has the same underlying distribution as X.

2. Informational Divergence

The *informational divergence* or *Kullback-Liebler divergence* between two probability distributions $p(x)$ and $q(x)$ on the same universe Ω is a measure of distance between them. It is formally defined as

$$\mathbf{D}(p\|q) = \sum_{\substack{x \in \Omega \\ p(x) \neq 0}} p(x) \log \frac{p(x)}{q(x)} = \mathbb{E}_{x \sim p}\left[\log \frac{p(x)}{q(x)}\right].$$

Note that if there is any point x with $q(x) = 0$ and $p(x) > 0$, then $\mathbf{D}(p\|q) = \infty$. So this notion is most useful when $\mathrm{supp}(p) \subseteq \mathrm{supp}(q)$. It can be thought of as a measure of how well p approximates q. Note that a particular case that guarantees $\mathrm{supp}(p) \subseteq \mathrm{supp}(q)$ is when $p(x)$ is the law of a random variable X, and $q(x)$ is the law of the random variable obtained from X by conditioning on an event E.

Let us list some facts about the divergence.

- We have $\mathbf{D}(p\|p) = 0$.
- Unlike mutual information, $\mathbf{D}(p\|q)$ is *not* symmetric.
- Suppose that p and q are respectively uniform distributions on sets $P \subseteq Q$. Then

$$\mathbf{D}(p\|q) = \log \frac{|Q|}{|P|}.$$

- More generally if p is obtained from q by conditioning on the event that x belongs to a set $E \subseteq \mathrm{supp}(q)$, then

$$\mathbf{D}(p\|q) = \log\frac{1}{q(E)} = \log\frac{1}{\Pr_{x\sim q}[x \in E]}.$$

- Always $\mathbf{D}(p\|q) \ge 0$. Indeed by convexity of $-\log(x)$, we have

$$\mathbf{D}(p\|q) = -\mathbb{E}_{x\sim p}\left[\log\frac{q(x)}{p(x)}\right] \ge -\log\mathbb{E}_{x\sim p}\left[\frac{q(x)}{p(x)}\right] = -\log 1 = 0.$$

2.1. **Divergence and Entropy.** Let X be a random variable with the law $p(x)$ supported on a set χ. Intuitively the entropy of X is related to how much $p(x)$ diverges from the uniform distribution ν on χ. The more $p(x)$ diverges the lesser its entropy is, and vice versa. Indeed it is straightforward to verify that

$$H(X) = \log|\chi| - \mathbf{D}(p\|\nu).$$

2.2. **Divergence and Mutual information.** Mutual information $I(X;Y)$ can also be expressed as a divergence, of the product $p(x)\times p(y)$ of the marginal distributions of the two random variables X and Y, from $p(x,y)$ the random variables' joint distribution:

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

Indeed

$$\begin{aligned}\mathbf{D}(p(x,y)\|p(x)p(y)) &= \sum_{x,y} p(x,y)\log\frac{p(x,y)}{p(x)p(y)} \\ &= \sum_{x,y} p(x,y)\left(\log\frac{1}{p(x)} + \log\frac{1}{p(y)} - \log\frac{1}{p(x,y)}\right) \\ &= H(X) + H(Y) - H(XY) = I(X;Y).\end{aligned}$$

Note further that

$$\begin{aligned}I(X;Y) &= \mathbf{D}(p(x,y)\|p(x)p(y)) = \sum_{x,y} p(x,y)\log\frac{p(x,y)}{p(x)p(y)} \\ &= \sum_y p(y)\sum_x p(x|y)\log\frac{p(x|y)}{p(x)} = \sum_y p(y)\mathbf{D}(p(x|y)\|p(x)) \\ &= \mathbb{E}_y\mathbf{D}(p(x|y)\ \|\ p(x)) = \mathbb{E}_{y\sim Y}\mathbf{D}(X|_{Y=y}\ \|\ Y)\end{aligned}$$

If X and Y are random variables on the same probability space with distributions $p(x)$ and $q(x)$, we might also write $\mathbf{D}(X\|Y)$ to denote $\mathbf{D}(p\|q)$. We summarize as the following theorem.

Theorem 13. *Let A, B be random variables in the same probability space. Then*

$$I(A;B) = \mathbb{E}_{a\sim A}\mathbf{D}\left(B|_{A=a}\ \|\ B\right),$$

and more generally if C is also a random variable in the same probability space:

$$I(A;B|C) = \mathbb{E}_{\substack{a\sim A\\ c\sim C}}\mathbf{D}\left(B|_{A=a,C=c}\ \|\ B|_{C=c}\right).$$

3. Things to add

Pinsker's inequality, Divergenec as a measure of surprise with emperical experiments, Normal distribution as highest entropy with fixed expected value and variance, super additivity of divergence,

Theorem 14. *Let $X = (X_1, \ldots, X_n)$ be independent random variables, and let $E = E(X)$ be an event with $\Pr[E] \geq 2^{-\epsilon n}$. Then for most coordinates $D(p_{x_i|E} \| p_{x_i})$ is small.*

References

School of Computer Science, McGill University, Montréal, Canada
E-mail address: hatami@cs.mcgill.ca