Information and Coding Theory **Winter 2025**

Lecture 4: January 16, 2025

Lecturer: Madhur Tulsiani

1 Graph Entropy

We now consider an application of mutual information, using the concept of Graph Entropy defined by Körner [Kör73], and later used by Newman and Wigderson [NW95] for certain circuit (formula) lower bound problems (see also the book by Jukna [Juk12]). This also provides an example of the scenario we discussed in the previous lecture, when the mutual information $I(X;Y$ is being optimized over our choice of random variables X, Y, rather than being computed for given random variables.

Given a graph $G = (\mathcal{V}, \mathcal{E})$, we define the graph entropy $H(G)$ as

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

Note that while the concept is called "entropy", we are defining it as a mutual information. The name entropy comes from the original definition related to the best (asymptotic) transmission rate for a random variable distributed over the vertices of the graph, when we are required to use different symbols for vertices connected by edges (but not necessarily otherwise). It can be proved that this asymptotic limit comes out to be equal to the mutual information above, and we will use this version of the definition. Also, while the graph entropy can be defined with respect to any distribution P on the vertex set $\mathcal{V}$, we will restrict our discussion to the uniform distribution. Let us check a couple of examples.

Example 1.1 (Complete graph)**.** *Let K_n denote the complete graph on n vertices. Then $H(K_n) = \log n$. This follows from the fact that any independent set is of size at most one, and thus, we must have $Y = X$. This gives*

$$I(X;Y) \;=\; H(X) - H(X|Y) \;=\; \log n - 0 \;=\; \log n\,.$$

Also note that $\log n$ is the maximum possible value for a graph with $|\mathcal{V}| = n$.

Example 1.2 (Bipartite graph)**.** *Let G be a bipartite graph, with n_1 vertices on one side and n_2 vertices on the other. Then, for any vertex v, all the vertices on the side of v form an indepdent set*

containing v. If X is a uniformly random vertex, and Y equals all the vertices on the side of X, then

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

Since $H(G)$ is the minimum of $I(X;Y)$ over all (X,Y), we get that $H(G)\leq 1$.

Exercise 1.3. *Let $\alpha(G)$ denote the size of the maximum independent set in a graph G. Prove that $H(G) \;\geq\; \log\left(\frac{n}{\alpha(G)}\right)$.*

An important property of graph entropy that we need, is that it is *sub-additive* under union of edges.

Proposition 1.4 (Sub-additivity of graph entropy)**.** *Let $G_1=(\mathcal{V},\mathcal{E}_1)$ and $G_2=(\mathcal{V},\mathcal{E}_2)$ be two graphs, and let $G=(\mathcal{V},\mathcal{E}_1\cup\mathcal{E}_2)$, which we denote by $G=G_1\cup G_2$. Then,*

$$H(G) \;=\; H(G_1\cup G_2) \;\leq\; H(G_1)+H(G_2)\,.$$

Proof: Let (X,Y_1) and (X,Y_2) be pairs of random variables achieving $H(G_1)$ and $H(G_2)$ (note that in both cases X is a uniform vertex from $\mathcal{V}$). We can define (why?) a joint distribution on the tuple (X,Y_1,Y_2) such that Y_1 and Y_2 are independent conditioned on any value of X. Take this to be the joint distribution of the tuple (X,Y_1,Y_2) and let $Y=Y_1\cap Y_2$. Note that if Y_1,Y_2 are independent sets containing X in G_1 and G_2 respectively, then $Y_1\cap Y_2$ is an indepdent set in G, containing X. This gives,

$$\begin{aligned}
H(G_1\cup G_2) \;&\leq\; I(X;Y)\\
&\leq\; I(X;(Y_1,Y_2)) && \text{(data processing inequality)}\\
&=\; H(Y_1,Y_2)-H(Y_1,Y_2\mid X)\\
&=\; H(Y_1,Y_2)-H(Y_1\mid X)-H(Y_2\mid X) && \text{(conditional independence)}\\
&\leq\; H(Y_1)+H(Y_2)-H(Y_1\mid X)-H(Y_2\mid X) && \text{(sub-additivity of entropy)}\\
&=\; H(G_1)+H(G_2)\,,
\end{aligned}$$

which proves the claim. ∎

1.1 Covering the complete graph with bipartite graphs

The properties of graph entropy considered so far can be used to provide a very simple answer to the following combnatorial question: what is the minimum number of bipartite graphs $G_1,\ldots,G_r$ such that their edges cover all the edges of the complete graph i.e.,

$$K_n \;=\; G_1\cup\cdots\cup G_r\,.$$

Note that just counting edges does not give a very strong bound since K_n has $n(n-1)/2$ edges, while even a single bipartite graph can have $n^2/4$ edges. On the other hand, graph entropy will yield a (tight!) boound of $\log n$. This also proves a special case of the formula-size lower bounds considered by Newman and Wigderson [NW95], when considering $\bigvee \bigwedge \bigvee$ formulas (three alternating layers of OR, AND, and OR gates, with AND gates having fan-in 2) for the threshold function checking $\sum_{i=1}^{n} x_i \geq 2$. Take a look at the paper for more details.

Back to the case of graphs, when $K_n = G_1 \cup \cdots \cup G_r$, we have

$$\log n = H(K_n) \leq H(G_1) + \cdots + H(G_r) \leq r\,,$$

where we used the bounds on the graph entropy of complete and bipartite graphs, as computed earlier.

Exercise 1.5. *Prove that the above bound is tight. In particular, when n is a power of 2, find a covering of K_n with $\log n$ bipartite graphs (Hint: Think of each vertex as a $(\log n)$-bit string).*

2 Kullback Leibler divergence

The Kullback-Leibler divergence (KL-divergence), also known as relative entropy, is a measure of how different two distributions are. Note that here we will talk in terms of distributions instead of random variables, since this is how KL-divergence is most commonly expressed. It is of course easy to think of a random variable corresponding to a given distribution and vice-versa. We will use capital letters like $P(X)$ to denote a distribution for the random variable X and lowercase letters like $p(x)$ to denote the probability for a specific element x.

Let P and Q be two distributions on a universe $\mathcal{X}$, then the KL-divergence between P and Q is defined as:

$$D(P||Q) := \sum_{x \in U} p(x) \log \left(\frac{p(x)}{q(x)} \right)$$

Let us consider a simple example.

Example 2.1. *Suppose $\mathcal{X} = \{a, b, c\}$, and $p(a) = \frac{1}{3}$, $p(b) = \frac{1}{3}$, $p(c) = \frac{1}{3}$ and $q(a) = \frac{1}{2}$, $q(b) = \frac{1}{2}$, $q(c) = 0$. Then*

$$D(P||Q) = \frac{2}{3} \log \frac{2}{3} + \infty = \infty\,.$$

$$D(Q||P) = \log \frac{3}{2} + 0 = \log \frac{3}{2}\,.$$

The above example illustrates two important facts: $D(P||Q)$ and $D(Q||P)$ are not necessarily equal, and $D(P||Q)$ may be infinite. Even though the KL-divergence is not symmetric,

it is often used as a measure of "dissimilarity" between two distribution. Towards this, we first prove that it is non-negative and is 0 if and only if $P = Q$.

Lemma 2.2. *Let P and Q be distributions on a finite universe $\mathcal{X}$. Then $D(P||Q) \geq 0$ with equality if and only if $P = Q$.*

Proof: Let $\text{Supp}(P) = \{x \mid p(x) > 0\}$. Then, we must have $\text{Supp}(P) \subseteq \text{Supp}(Q)$ if $D(P, Q) < \infty$. We can then assume without loss of generality that $\text{Supp}(Q) = \mathcal{X}$. Using the fact the log is a (strictly) concave function, with Jensen inequality, we have:

$$\begin{aligned}
D(P||Q) = \sum_{x \in \mathcal{X}} p(x) \log \frac{p(x)}{q(x)} &= \sum_{x \in \text{Supp}(P)} p(x) \log \frac{p(x)}{q(x)} \\
&= - \sum_{x \in \text{Supp}(P)} p(x) \log \frac{q(x)}{p(x)} \\
&\geq - \log \left(\sum_{x \in \text{Supp}(P)} p(x) \cdot \frac{q(x)}{p(x)} \right) \\
&= - \log \left(\sum_{x \in \text{Supp}(P)} q(x) \right) \\
&\geq - \log 1 = 0 \,.
\end{aligned}$$

For the case when $D(P||Q) = 0$, we note that this implies $p(x) = p(x)\ \forall x \in \text{Supp}(P)$, which in turn gives that $p(x) = q(x)\ \forall x \in \mathcal{X}$. ■

Like entropy and mutual information, we can also derive a chain rule for KL-divergence. Let $P(X, Y)$ and $Q(X, Y)$ be two distributions for a pair of variables X and Y. We then have the following expression for $D(P(X, Y)||Q(X, Y))$.

Proposition 2.3 (Chain rule for KL-divergence)**.** *Let $P(X, Y)$ and $Q(X, Y)$ be two distributions for a pair of variables X and Y. Then,*

$$\begin{aligned}
D(P(X, Y) \parallel Q(X, Y)) &= D(P(X) \parallel Q(X)) + \mathop{\mathbb{E}}_{x \sim P} [D(P(Y|X = x) \parallel Q(Y|X = x))] \\
&= D(P(X) \parallel Q(X)) + D(P(Y|X) \parallel Q(Y|X))
\end{aligned}$$

Here $P(X)$ and $Q(X)$ denote the marginal distributions for the first variable, and $P(Y|X = x)$ denotes the conditional distribution of Y.

Proof: The proof follows from (by now) familiar manipulations of the terms inside the

log function.

$$\begin{aligned} D(P(X,Y) \parallel Q(X,Y)) &= \sum_{x,y} p(x,y) \log \frac{p(x,y)}{q(x,y)} \\ &= \sum_{x,y} p(x)p(y|x) \log \left(\frac{p(x)}{q(x)} \cdot \frac{p(y|x)}{q(y|x)} \right) \\ &= \sum_{x} p(x) \log \frac{p(x)}{q(x)} \sum_{y} p(y|x) + \sum_{x} p(x) \sum_{y} p(y|x) \log \frac{p(y|x)}{q(y|x)} \\ &= D(P(X) \parallel Q(X)) + \sum_{x} p(x) \cdot D(P(Y|X=x) \parallel Q(Y|X=x)) \\ &= D(P(X) \parallel Q(X)) + D(P(Y|X) \parallel Q(Y|X)) \end{aligned}$$

■

Note that if $P(X,Y) = P_1(X)P_2(Y)$ and $Q(X,Y) = Q_1(X)Q_2(Y)$, then $D(P||Q) = D(P_1||Q_1) + D(P_2||Q_2)$.

We note that KL-divergence also has an interesting interpretation in terms of source coding. Writing

$$D(P||Q) = \sum p(x) \log \frac{p(x)}{q(x)} = \sum p(x) \log \frac{1}{q(x)} - \sum p(x) \log \frac{1}{p(x)},$$

we can view this as the number of extra bits we use (on average) if we designed a code according to the distribution P, but used it to communicate outcomes of a random variable X distributed according to Q. The first term in the RHS, which corresponds to the average number of bits used by the "wrong" encoding, is also referred to as cross entropy.

2.1 Convexity of KL-divergence

Before we consider applications, let us prove an important property of KL-divergence. We prove below that $D(P \parallel Q)$, when viewed as a function of the inputs P and Q, is jointly convext in both it's inputs i.e., it is convex in the input (P,Q) when viewed as a tuple.

Proposition 2.4. *Let P_1, P_2, Q_1, Q_2 be distributions on a finite universe $\mathcal{X}$, and let $\alpha \in [0,1]$. Then,*

$$D(\alpha \cdot P_1 + (1-\alpha) \cdot P_2 \parallel \alpha \cdot Q_1 + (1-\alpha) \cdot Q_2) \leq \alpha \cdot D(P_1 \parallel Q_1) + (1-\alpha) \cdot D(P_2 \parallel Q_2).$$

Proof: For this proof, we will use an inequality called the log-sum inequality, the proof of which is left is an exercise. The inequality states that for $a_1, a_2, b_1, b_2 \geq 0$

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

Using the above inequality, we can bound the LHS as

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

■

Exercise 2.5 (Log-sum inequality)**.** *Prove that for* $a_1, a_2, b_1, b_2 \geq 0$

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

References

[Juk12] Stasys Jukna, *Boolean function complexity - advances and frontiers*, Algorithms and combinatorics, vol. 27, Springer, 2012. 1

[Kör73] János Körner, *Coding of an information source having ambiguous alphabet and the entropy of graphs*, 6th Prague conference on information theory, 1973, pp. 411–425. 1

[NW95] Ilan Newman and Avi Wigderson, *Lower bounds on formula size of Boolean functions using hypergraph entropy*, SIAM Journal on Discrete Mathematics **8** (1995), no. 4, 536–542. 1, 3