Information and Coding Theory **Winter 2021**

Lecture 4: January 21, 2021

Lecturer: Madhur Tulsiani

1 Fano's inequality

We first prove an important inequality that lets us understand how well can some "ground truth" random variable X be predicted based on some observed data Y. We state the inequality in the language of Markov chains, which we saw before in the context of data processing inequality. We will denote the Markov chain as $X \to Y \to \widehat{X}$. We can think of X as the choice of an unknown parameter from some finite set $\mathcal{X}$. We think of Y as the "data" generated from this, say a sequence independent samples. Finally, we think of $\widehat{X}$ as a "guess" for X, which depends only on the data. Fano's inequality is concerned with the probability of error in the guess, defined as $p_e = \mathbb{P}\left[\widehat{X} \neq X\right]$. We have the following statement

Lemma 1.1 (Fano's inequaity)**.** *Let $X \to Y \to \widehat{X}$ be a Markov chain, and let $p_e = \mathbb{P}\left[\hat{X} \neq X\right]$. Let $H_2(p_e)$ denote the binary entropy function computed at p_e. Then,*

$$H_2(p_e) + p_e \cdot \log\left(|\mathcal{X}| - 1\right) \;\geq\; H(X|\widehat{X}) \;\geq\; H(X|Y)\,.$$

Proof: We define a binary random variable, which indicates an error i.e

$$E \;:=\; \begin{cases} 1 \text{ if } \widehat{X} \neq X \\ 0 \text{ if } \widehat{X} = X \end{cases}$$

The bound in the ineuality then follows from considering the undertainty that still remains after our prediction, i.e., the entroy $H(X, E|\widehat{X})$.

$$H(X, E|\widehat{X}) \;=\; H(X|\widehat{X}) + H(E|X, \widehat{X}) \;=\; H(X|\widehat{X})\,,$$

since $H(E|X, \widehat{X}) = 0$ (why?) Another way of computing this entropy is

$$\begin{aligned} H(X, E|\widehat{X}) &= H(E|\widehat{X}) + H(X|E, \widehat{X}) \\ &= H(E|\widehat{X}) + p_e \cdot H(X|E = 1, \widehat{X}) + (1 - p_e) \cdot H(X|E = 0, \widehat{X}) \\ &\leq H(E) + p_e \cdot H(X|E = 1, \widehat{X}) \\ &\leq H_2(p_e) + p_e \cdot \log\left(|\mathcal{X}| - 1\right)\,. \end{aligned}$$

Comparing the two expressions then proves the claim. ■

Fano's inequality provides a useful way of lower bounding the error of a predictor, particularly in the case when $|\mathcal{X}| > 2$. As we will see later, in the case when $|\mathcal{X}| = 2$, we will be able to obtain better bounds using the concept of KL-divergence considered later.

2 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. 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 restric our discussion to the uniform distribution. Let us check a couple of examples.

Example 2.1 (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 2.2 (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 2.3. *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 2.4 (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. ■

2.1 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 2.5. *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).*

3 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 3.1. *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 3.2. *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 $\mathrm{Supp}(P) = \{x \mid p(x) > 0\}$. Then, we must have $\mathrm{Supp}(P) \subseteq \mathrm{Supp}(Q)$ if $D(P,Q) < \infty$. We can then assume without loss of generality that $\mathrm{Supp}(Q) = \mathcal{X}$. Using the fact the log is a (strictly) concave function, with Jensen inequality, we have:

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

For the case when $D(P||Q) = 0$, we note that this implies $p(x) = p(x)\ \forall x \in \mathrm{Supp}(P)$, 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 3.3 (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.

3.1 Total variation distance and Pinsker's inequality

We can now relate KL-divergence to some other notions of distance between two probability distributions.

Definition 3.4. *Let P and Q be two distributions on a finite universe $\mathcal{X}$. Then the* total-variation distance *or* statistical distance *between P and Q is defined as*

$$\delta_{TV}(P,Q) \;=\; \frac{1}{2} \cdot \|P - Q\|_1 \;=\; \frac{1}{2} \cdot \sum_{x \in \mathcal{X}} |p(x) - q(x)| \;.$$

*The quantity $\|P - Q\|_1$ is referred to as the ℓ_1-*distance *between P and Q.*

The total variation distance of P and Q represents the maximum probability with which any test can distinguish between the two distributions *given one random sample*. It may seem that the restriction to one sample severely limits the class of tests, but we can always think of an n-sample test for P and Q as getting one sample from one of the product distributions P^n or Q^n.

Let $f : \mathcal{X} \to \{0,1\}$ be any classifier, which given one sample $x \in \mathcal{X}$, outputs 1 if the guess is that the sample came from P, and 0 if the guess is that it came from Q. The difference in its behavior over the two distributions can be measured by the quantity (which can be thought of as the rate of true positive minus the rate of false positive) $|\mathbb{E}_P[f(x)] - \mathbb{E}_Q[f(x)]|$. The following lemma bounds this in terms of the total variation distance.

Lemma 3.5. *Let P, Q be any distributions on $\mathcal{X}$. Let $f : \mathcal{X} \to [0, B]$. Then*

$$\left|\mathop{\mathbb{E}}_{P}[f(x)] - \mathop{\mathbb{E}}_{Q}[f(x)]\right| \;\le\; \frac{B}{2} \cdot \|P - Q\|_1 \;=\; B \cdot \delta_{TV}(P, Q)\,.$$

Proof:

$$\begin{aligned}
\left|\mathop{\mathbb{E}}_{P}[f(x)] - \mathop{\mathbb{E}}_{Q}[f(x)]\right| &= \left|\sum_{x\in\mathcal{X}} p(x)\cdot f(x) - \sum_{x\in\mathcal{X}} q(x)\cdot f(x)\right| \\
&= \left|\sum_{x\in\mathcal{X}} (p(x) - q(x))\cdot f(x)\right| \\
&= \left|\sum_{x\in\mathcal{X}} (p(x) - q(x))\cdot \left(f(x) - \frac{B}{2}\right) + \frac{B}{2}\cdot\left(\sum_{x\in\mathcal{X}} p(x) - q(x)\right)\right| \\
&\le \sum_{x\in\mathcal{X}} |p(x) - q(x)| \cdot \left|f(x) - \frac{B}{2}\right| \\
&\le \frac{B}{2}\cdot \|P - Q\|_1
\end{aligned}$$

■

Exercise 3.6. *Prove that the above inequality is tight. What is the optimal classifier f?*

In many applications, we want to actually bound the ℓ_1-distance between P and Q but it's easier to analyze the KL-divergence. The following inequality helps relate the two.

Lemma 3.7 (Pinsker's inequality)**.** *Let P and Q be two distributions defined on a universe $\mathcal{X}$. Then*

$$D(P \,\|\, Q) \;\ge\; \frac{1}{2\ln 2}\cdot \|P - Q\|_1^2\,.$$

We will prove the inequality in two steps. Let us first consider a special case when $\mathcal{X} = \{0,1\}$ and P, Q are distributions as below

$$P \;=\; \begin{cases} 1 & \text{w.p. } p \\ 0 & \text{w.p. } 1-p \end{cases} \qquad \text{and} \qquad Q \;=\; \begin{cases} 1 & \text{w.p. } q \\ 0 & \text{w.p. } 1-q \end{cases}$$

In this case, we have

$$D(P\|Q) \;=\; p \cdot \log\left(\frac{p}{q}\right) + (1-p)\cdot \log\left(\frac{1-p}{1-q}\right) \qquad \text{and} \qquad \|P-Q\|_1 \;=\; 2\cdot |p-q| \; .$$

We will first prove Pinsker's inequality for this special case.

Proposition 3.8 (Pinsker's inequality for $\mathcal{X} = \{0,1\}$)**.** *Let P and Q be distributions as above. Then,*

$$p \cdot \log\left(\frac{p}{q}\right) + (1-p)\cdot \log\left(\frac{1-p}{1-q}\right) \;\geq\; \frac{2}{\ln 2}\cdot (p-q)^2 \, .$$

Proof: Let

$$f(p,q) \;:=\; p \cdot \log\left(\frac{p}{q}\right) + (1-p)\cdot \log\left(\frac{1-p}{1-q}\right) - \frac{2}{\ln 2}\cdot (p-q)^2 \, .$$

We have,

$$\frac{\partial f}{\partial q} = -\frac{(p-q)}{\ln 2}\left(\frac{1}{q(1-q)} - 4\right) \, .$$

Since $\frac{1}{q(1-q)} - 4 \geq 0$ for all q, we have that $\frac{\partial f}{\partial q} \leq 0$ when $q \leq p$ and $\frac{\partial f}{\partial q} \geq 0$ when $q \geq p$. Moreover, $f(p,q) = \infty$ when $q = 0$ and $f(p,q) = 0$ when $q = p$. Thus, the function achieves its minimum value at $q = p$ and is always non-negative, which proves the desired inequality. ■

We can now reduce the general case of Pinsker's inequality, to the case of $\mathcal{X} = \{0,1\}$ considered above.

Proposition 3.9. *Let P and Q be distributions on a finite set $\mathcal{X}$. Then, there exist distributions P', Q' on $\{0,1\}$ such that*

$$\left\|P' - Q'\right\|_1 \;=\; \|P-Q\|_1 \quad \textit{and} \quad D(P\|Q) \;\geq\; D(P'\|Q')$$

Proof: Let $A \subset \mathcal{X}$ be

$$A = \{x \mid p(x) \geq q(x)\} \; .$$

and P' and Q' be

$$P' \;:=\; \begin{cases} 1 & \text{w.p. } \sum_{x \in A} p(x) \\ 0 & \text{w.p. } \sum_{x \notin A} p(x) \end{cases} \qquad \text{and} \qquad Q' \;:=\; \begin{cases} 1 & \text{w.p. } \sum_{x \in A} q(x) \\ 0 & \text{w.p. } \sum_{x \notin A} q(x) \end{cases}$$

Then,

$$\begin{aligned} \|P - Q\|_1 &= \sum_{x \in \mathcal{X}} |p(x) - q(x)| \\ &= \sum_{x \in A} (p(x) - q(x)) + \sum_{x \notin A} (q(x) - p(x)) \\ &= \left| \sum_{x \in A} p(x) - \sum_{x \in A} q(x) \right| + \left| \left(1 - \sum_{x \in A} p(x)\right) - \left(1 - \sum_{x \in A} q(x)\right) \right| \\ &= \left\|P' - Q'\right\|_1 \end{aligned}$$

To calculate the KL-divergence, we define a random variable Z (which is a function of X) as

$$Z \;=\; \begin{cases} 1 & \text{if } x \in A \\ 0 & \text{if } x \notin A \end{cases} .$$

Since Z is a function of X, we can also think of the two distributions P and Q as joint distributions for the random variables (X, Z). Also, note that the marginal distributions of Z are P' and Q'. Applying the chain rule for KL-divergence gives

$$\begin{aligned} D(P\|Q) &= D(P(X,Z) \;\|\; Q(X.Z)) \\ &= D(P(Z) \;\|\; Q(Z)) + D(P(X|Z) \;\|\; Q(X|Z) \\ &\geq D(P(Z) \;\|\; Q(Z)) \\ &= D\left(P'\|Q'\right) \end{aligned}$$

which completes the proof. ■

Finally, we can complete the proof of Pinkser's inequality for the general case, by noting that

$$D(P\|Q) \;\geq\; D(P'\|Q') \;\geq\; \frac{1}{2\ln 2} \cdot \left\|P' - Q'\right\|_1^2 \;=\; \frac{1}{2\ln 2} \cdot \|P - Q\|_1^2 .$$

References

[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. 2

[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. 2, 3