COMS-E6998: Information Theory in Computer Science **Fall 2017**

Lecture 3: Graph Embeddings, Mutual Information, KL Divergence

Lecturer: Omri Weinstein *Scribes: Geelon So*

Recall that for a *uniformly distributed* random variable X,

$$H(X) = \log \left|\text{supp}(X)\right|.$$

This equation hints the connection between entropy and combinatorics – We already saw this fact is useful for counting applications, and today we'll see a more advanced application for counting embeddings of *general graphs.*

3.1 Application of Shearer's Lemma to Graph Embeddings

Given a collection $X_1, \ldots, X_n$ of n random variables, we can let $X = (X_1, \ldots, X_n)$ package the random variables together into a single vector. Of course, X is just another random variable, so if X is uniformly distributed over its support, $H(X)$ is still the log of the size of its support.

But as a practical matter, $H(X)$ may be difficult to calculate. What we can do instead is leverage the fact that X is built up from X_i's. So, maybe we don't need to examine all n of them at once. Maybe we can just choose a smaller subset $S \subseteq [n]$ and consider only those X_i's for which the indices i are in S.[1]

Shearer's lemma does exactly this. It bounds $H(X)$ with respect to $H(X_S)$ considered over a distribution of subsets S. As a result, Galvin writes, *a global problem—understanding $H(X_1, \ldots, X_n)$—is reduced to a collection of local ones—understanding $H(X_S)$ for each S.*[G2014] Here is Shearer's lemma:

Lemma 3.1.1 (Shearer)**.** *Let $X_1, \ldots, X_n$ be n random variables. Let $\mathcal{F}$ be a family of subsets of $[n]$, and $S \in_R \mathcal{F}$ is a random subset of $[n]$. Suppose that there exists some $\alpha \geq 0$ such that for all $i \in [n]$,*

$$\Pr\left[i \in S\right] \geq \alpha.$$

Then, $H(X) := H(X_1, \ldots, X_n)$ is bounded above by

$$H(X) \leq \frac{1}{\alpha}\mathbb{E}\big[H(X_S)\big].$$

This simply says: form a family of random vectors X_S, each containing 'partial' information relative to X; for each i, the coordinate X_i may or may not be contained in X_S. But as long as the probability

$$\Pr[X_i \text{ is a component of the random variable } X_S]$$

is at least some fixed α, then $H(X)$ may be bounded as above.

Let's quickly refresh on how Shearer's lemma helped us bound the maximal number of embeddings of a triangle into a graph. Then, we'll move to the general case of graph embeddings.

[1] In this way, we construct a 'smaller' vector X_S, the projection of X onto the indices specified in S. Equivalently,

$$X_S = (X_i : i \in S).$$

3.1.1 Embedding Triangles

Let t be the maximal number of embeddings of a triangle into a graph G with ℓ edges. If we want to use Shearer's lemma to bound t, we should convert the combinatorial value t into an entropic one, $H(X)$.

Naturally, this suggests that we want a space whose support is the collection of all triangles in the graph G. Thus, let $X = (X_1, X_2, X_3)$, where each of the X_i's are sampled from the vertex set V of G. Of course, we don't want to uniformly sample from $V \times V \times V$ because not all triples of vertices in G form a triangle.

Instead, we'll give $V \times V \times V$ precisely the probability distribution whose support is those triples that form a triangle in G, and whose distribution is uniform over that support. Now, as there are t triangles, and the number of ways to represent the same triangle as a triple of vertices is $6 = 3!$, we have

$$H(X) = \log 6t.$$

We can now reduce this problem about triangles into a problem about its edges (again, turning a global problem into a local one). When $X = (X_1, X_2, X_3)$ is a triangle, then its edges correspond to the subsets (X_1, X_2), (X_2, X_3), and so on. The probability that X_i is contained in a random tuple (X_j, X_k) is $2/3$.

Shearer's lemma (where S is a random edge) gives

$$\log 6t = H(X) \leq \frac{1}{2/3}\mathbb{E}\left[H(X_S)\right].$$

In particular, X_S is a distribution over all edges contained in a triangle of G. We appeal to the maximality of entropy of a uniform distribution to bound $H(X_S)$.[2] As there are ℓ edges in G, $H(X_S)$ is bounded above by $\log 2\ell$, where the factor of 2 comes from the number of ways the same edge can be represented as a tuple of vertices.

If $H(X_S)$ is bounded above by $\log 2\ell$ for all S, then the the expectation must also be bounded by $\log 2\ell$. Replacing this in the above inequality yields

$$\log 6t \leq \frac{3}{2}\log 2\ell.$$

Raising both sides to the power of 2 gives:

Proposition 3.1.2. *The maximal number t of embeddings of a triangle into a graph with ℓ edges is*

$$t \leq \frac{1}{6}(2\ell)^{3/2}.$$

□

Hopefully, this gives us a better grasp on how Shearer's lemma can lead to combinatorial bounds. Now, let's move on to graph embeddings in general.

3.1.2 Graph Embeddings

Previously, we looked at the maximal number of triangles that can exist in a graph G with ℓ edges. It seems that it's possible to give a bound because a triangle has a lot of structure—it is, after all, the complete graph on three vertices.

[2] Recall, if μ is a distribution over X, the $H_\mu(X) \leq H_{\text{uniform}}(X)$.

But really, all we used was that every vertex of a triangle is contained in two-third of that triangle's edges. And so, it seems that we should be able to generalize the solution to all graphs, and not just triangles. That is, if we let T be any graph, what is the greatest number of 'copies' of T in a graph G with ℓ edges?

Call this number $N(T, \ell)$ for short. Certainly, $N(T, \ell)$ must depend on the number of edges in T itself and something about how those edges are connected in T. And in fact, it turns out that we can get a bound on $N(T, \ell)$ dependent only on the number of edges and the so-called *fractional independent set number* of T, a result proved by Friedgut and Kahn, extending Alon's work.[FK1996, A1981]

We'll need a few definitions before proving their result. First, to capture the way a graph is connected, we introduce the following matrix:

Definition 3.1.3. *The* vertex-edge incidence matrix *of a graph H is a 0-1 matrix M whose rows are indexed by $V(H)$ and columns by $E(H)$, where the ij-entry indicates whether vertex i belongs to edge j.*

The vertex-edge incidence matrix lets us define the fractional independent set number of a graph. But to make its introduction less *ad hoc*, we'll briefly provide some background. However, if you're comfortable with manipulating symbols as formal objects, feel free to skip directly to Definition 3.1.5.

Definition 3.1.4. *An* independent set *of a graph H is a collection of vertices $Y \subset V(H)$ such that no two vertices in Y are joined by an edge of H. The* independence number *$\alpha(H)$ of a graph H is the size of a maximal independent set.*

Every subset $Y \subset V(H)$ may be represented by an indicator function $\psi : V(H) \to \{0, 1\}$, where $\psi(i)$ indicates whether the ith vertex is contained in Y. The condition that ψ is an independent set is the requirement: for every edge $e = (u, v) \in E(H)$,

$$\psi(u) + \psi(v) \leq 1.$$

This is written more compactly by thinking of ψ as a vector, where the ith coordinate $\psi_i = \psi(i)$ indicates whether i is contained in the independent set. Then the condition that ψ is an independent set is now just

$$M^T \psi \leq \mathbf{1}.$$

Furthermore, because the cardinality of an independent set is $\mathbf{1}^T \psi$, the independence number α is

$$\alpha = \max_{\psi} \mathbf{1}^T \psi.$$

Written this way, we easily see that determining the independence number of a graph is an integer program—a linear optimization problem with integral constraints ($M^T \psi \leq \mathbf{1}$) and an integral objective function ($\mathbf{1}^T \psi$).

We define the *fractional independent set number* as the *relaxation* of this integer program to a linear program—the same optimization problem where the ψ_i's are allowed to take on real values.[3]

Definition 3.1.5. *A* fractional independent set *is a map $\psi : V(H) \to [0, 1]$ such that for all edges $e = (u, v) \in E(H)$,*

$$\psi(u) + \psi(v) \leq 1.$$

The fractional independent set number *$\alpha^*(H)$ is the maximal value of $\sum_{v \in V(H)} \psi(v)$ over all ψ.*

Determining α^* is just a linear program (LP) that maximizes $\mathbf{1}^T \psi$ subject to the constraints $M^T \psi \leq \mathbf{1}$ and and $\psi \geq 0$, where M is the same vertex-edge incidence matrix as before.

[3] There are other ways to interpret the fractional independent set number. See [G2014] for more details.

Every optimization problem has a dual problem. Here, the dual is formally the LP that minimizes $\mathbf{1}^T\phi$ subject to the constraints $M\phi \geq \mathbf{1}$ and $\phi \geq 0$. Denote the minimum value of $\mathbf{1}^T\phi$ by γ^*, which we call the *fractional covering number* of the graph H. It'll do us some good to understand γ^* geometrically.

Recall that M is the vertex-edge incidence matrix, so ϕ is a vector indexed by the edges of H. We can think of each ϕ_i as a weight on the ith edge of H. The condition that $M\phi \geq \mathbf{1}$ is precisely the condition: for every vertex $v \in V(H)$, the sum of weights over all edges containing v is at least 1.

This dual LP is actually the linear relaxation of another integer program: can we find a minimal collection of edges such that every vertex is contained in an edge? That is, what is the minimal collection of edges that cover the vertices? Thus,

Definition 3.1.6. *A* fractional edge covering *is a map $\phi : E(H) \to [0,1]$ such that for all vertices $v \in V(H)$,*

$$\sum_{e\in E(H):v\in e} \phi(e) \geq 1,$$

where $e \in E(H) : v \in e$ indicates a summation over all edges e containing the vertex v. The fractional covering number γ^* *is the minimum value of $\sum_{e\in E(H)} \phi(e)$ over all fractional edge coverings ϕ.*

It turns out that $\gamma^* = \alpha^*$, a fact that follows from a direct application of the following theorem (proof given in reference):

Theorem 3.1.7 ([SU2011] Theorem A.3.1)**.** *A linear program and its dual have the same value.*

We are now in a position to state and prove a bound for $N(T, \ell)$.

Theorem 3.1.8 (Friedgut, Kahn, Alon 1996)**.** *Let T be a graph. Then $N(T,\ell) = \theta(\ell^{\alpha^*(T)})$. In particular, if $m := |E(T)|$ is the number of edges of T,*

$$\left(\frac{\ell}{m}\right)^{\alpha^*(T)} \lesssim N(T,\ell) \leq (2\ell)^{\alpha^*(T)}.$$

Proof. We'll prove the lower bound by construction (i.e. produce a graph G with at most ℓ edges such that there are $(\ell/m)^{\alpha^*(T)}$ embeddings of T into G).

Let ψ be an optimal fractional independent set of T. By definition,

$$\sum_{v\in V(T)} \psi(v) = \alpha^*(T),$$

and for every edge (u,v), we have $\psi(u) + \psi(v) \leq 1$. Let $c_v := (\ell/m)^{\psi(v)}$. For each v, create a cluster C_v with c_v vertices. If there is an edge between u and v, then connect each vertex in C_u with each vertex in C_v with an edge. That is, form the complete bipartite graph between C_u and C_v. Performing this process for each edge of T produces a graph G.

As a result, if (u,v) is an edge of T, then there are c_uc_v edges between C_u and C_v, we get a bound

$$c_uc_v = \left(\frac{\ell}{m}\right)^{\psi(u)}\left(\frac{\ell}{m}\right)^{\psi(v)} = \left(\frac{\ell}{m}\right)^{\psi(u)+\psi(v)} \leq \frac{\ell}{m},$$

recalling that $\psi(u) + \psi(v) \leq 1$. Since there are m edges in T, we can bound the number of edges in G

$$|E(G)| = \sum_{(u,v)\in E(T)} c_uc_v \leq \sum_{(u,v)\in E(T)} \frac{\ell}{m} \leq m \cdot \frac{\ell}{m} = \ell.$$

Embedding T into G is a simple matter of choosing a vertex from each C_v for each $v \in V(T)$, so[4]

$$\begin{aligned}\#\{\text{embeddings of } T \text{ into } G\} &= \prod_{v \in V(T)} |C_v| \\ &= \prod_{v \in V(T)} \left(\frac{\ell}{m}\right)^{\psi(v)} = \left(\frac{\ell}{m}\right)^{\sum_{v \in V(T)} \psi(v)} = \left(\frac{\ell}{m}\right)^{\alpha^*(T)}.\end{aligned}$$

This proves the lower bound.

The upper bound is where we'll use Shearer's lemma. The proof will pretty much follow the proof for the triangle case. Since we want to bound $N(T, \ell)$, let G be a maximal graph (that is, there are $N(T, \ell)$ embeddings of T into G). Let K be the collection of embeddings in G.

We'll convert the combinatorial value of $N(T, \ell) = |K|$ into an entropic one by letting $\sigma \in_R K$ be a random embedding chosen uniformly from K, so that

$$H(\sigma) = \log N(T, \ell).$$

Once we have an embedding σ, we look at random edges S of σ. We can calculate the probability that a random edge contains a vertex v of σ. If we can bound this probability $\Pr(v \in S)$ from below for all $v \in V(\sigma)$, then we can apply Shearer's lemma:

$$\log N(T, \ell) = H(\sigma) \le \frac{1}{\min_{v \in V(\sigma)} \Pr(v \in S)} \cdot \mathbb{E}\left[H(\sigma_S)\right]. \tag{3.1}$$

Note that while S is a random edge of a fixed σ, σ_S is a random edge of σ, which itself is a random embedding in G. Thus, σ_S is a random edge of G. But recall that entropy is maximized when the distribution is uniform. Since there are ℓ edges in G (thus 2ℓ possible pairs (u, v) that represent edges), the entropy of σ_S is bounded above by

$$H(\sigma_S) \le \log 2\ell.$$

It follows that the expectation must also be bounded by $\log 2\ell$. By replacing $\mathbb{E}\left[H(\sigma_S)\right]$ and clearing the logs in Equation 3.1, we get

$$N(T, \ell) \le (2\ell)^{1/\min_{v \in V(\sigma)} \Pr(v \in S)}. \tag{3.2}$$

Producing the best bound on $N(T, \ell)$ with this technique is now just a matter of choosing a distribution for S that maximizes

$$c := \min_{v \in V(\sigma)} \Pr(v \in S).$$

Equivalently, what is the best way to assign probabilities to edges such that

$$\sum_{e \in E(\sigma) : v \in e} \Pr(S = e) \ge c$$

for all vertices $v \in V(\sigma)$? But this should be a familiar expression—it's almost exactly the constraint equation to determine the fractional edge covering number $\gamma^*(T)$. In fact, the only difference is sum must be at least c instead of at least 1. So, let's just divide by c

$$\sum_{e \in E(\sigma) : v \in e} \frac{1}{c} \Pr(S = e) \ge 1.$$

[4] Technically, $(\ell/m)^{\alpha^*(T)}$ is an upper-bound on the number of embeddings of T into this particular G, since $|C_v|$ is actually $\lfloor (\ell/m)^{\psi(v)} \rfloor$ and not $(\ell/m)^{\psi(v)}$. So this only really shows that $N(T, \ell) = \Omega\left(\ell^{\alpha^*(T)}\right)$, and not $N(T, \ell) \ge (\ell/m)^{\alpha^*(T)}$.

Because our goal is to maximize c, we should minimize

$$\sum_{e \in E(\sigma)} \frac{1}{c} \Pr(S = e) \quad = \quad \frac{1}{c}.$$

So, our objective is precisely the objective function that determines $\gamma^*(T)$. Thus, the best we can do is

$$\gamma^*(T) = \sum_{e \in E(\sigma)} \frac{1}{c} \Pr(S = e).$$

Rearranging, we get

$$c = \frac{\sum_{e \in E(\sigma)} \Pr(S = e)}{\gamma^*(T)} = \frac{1}{\gamma^*(T)}.$$

Replacing $\min_{v \in V(\sigma)} \Pr(v \in S)$ with $1/\gamma^*(T)$ in Equation 3.2, we get

$$N(T, \ell) \leq \left(\frac{\ell}{m}\right)^{\gamma^*(T)}.$$

Recall that $\alpha^*(T) = \gamma^*(T)$. So, we have proved our theorem. □

Now, we'll shift to a different topic: Mutual Information and KL Divergence. Reasoning about the *correlation* between random variables is one of the obstacles in many of the applications and lower bounds we shall see in the future of this course (mostly, when we start talking about *interactive* communication). In this lecture we will introduce and formalize this notion (which measures correlation in an "entropic" way), its useful properties and some applications.

3.2 Mutual Information

Given random variables with joint distribution $(X, Y) \sim \mu$, the *mutual information* between X and Y is

$$I(X; Y) := H(X) - H(X|Y).$$

Intuitively, the mutual information function measures the entropy reduction in X when we condition on a random value of Y. That difference is therefore their shared information. The following Venn diagram helps us visualize this.

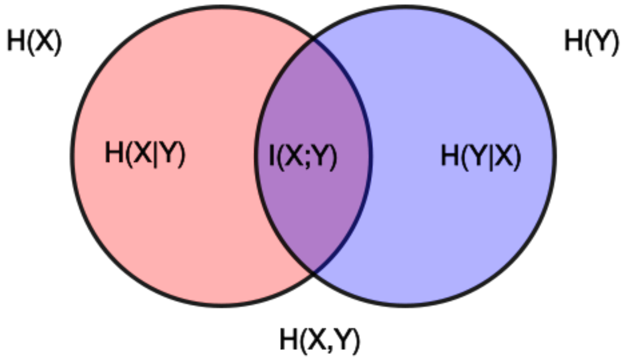

Figure 3.1: Relationship between different quantities of information. *Image courtesy of Wikipedia.*

Mutual information has the basic properties:

(1) I is symmetric in X and Y. That is, $I(X, Y) = I(Y, X)$.

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

(3) $I(X; Y) = 0$ if and only if X and Y are independent.

(4) $I(X; Y) \geq 0$.

Properties 1–3 are easily seen from Figure 3.1,[5] and Property 4 follows from the fact that conditioning does not increase entropy. That is, $H(X) \geq H(X|Y)$.

Example 3.2.1. *Consider X, Y from Example 2.3.2. That is, $X \in_R \{0, 1, 2, 3\}$ and $Y = X \mod 2$. Then $I(X; Y) = H(X) - H(X|Y) = 2 - 1 = 1$.*

3.2.1 Conditioning and Mutual Information

Now, let's define *conditional mutual information.* Intuitively, this answers, *given Z, how many bits of information does Y give on X?*

[5] We could also have shown these equalities from

$$H(X, Y) = H(X) + H(Y|X) = H(Y) + H(X|Y).$$

Definition 3.2.2 (Conditional Mutual Information)**.**

$$I(X;Y|Z) := \mathbb{E}_z[I(X;Y|Z=z)] = H(X|Z) - H(X|YZ).$$

Since the expectation is taken over the nonnegative random variable $I(X;Y|Z=z)$, it's true that

$$I(X;Y|Z) \geq 0.$$

However, unlike entropy, conditioning does not necessarily decrease mutual information. Let's take a look at an example where $I(X;Y|Z) \leq I(X;Y)$ is false.

Example 3.2.3. *Let Y, Z be uniform independent in $\{0,1\}$, and let $X = Y \oplus Z$, where $\oplus$ indicates addition mod 2. Then, $I(X;Y) = 0$, but $I(X;Y|Z) = 1$.*

Naturally, after introducing conditional mutual information, we can prove the corresponding chain rule:

Proposition 3.2.4 (Chain Rule for Mutual Information)**.** $I(X_1 \ldots X_n; Y) = \sum_i I(X_i; Y|X_{<i})$.

Proof. First, let's show that $I(X_1, X_2; Y) = I(X_1;Y) + I(X_2;Y|X_1)$. Generalizing this to n random variables $Y_1, \ldots, Y_n$ is just a matter of letting $X_1 = Y_1$ and $X_2 = (Y_2, \ldots, Y_n)$, then inducting.

So, following our nose, we expand out the definition of mutual information,

$$I(X_1, X_2; Y) = H(X_1, X_2) - H(X_1, X_2|Y).$$

Applying the chain rule for entropy, we have

$$\begin{aligned} I(X_1, X_2; Y) &= H(X_1) + H(X_2|X_1) - [H(X_1|Y) - H(X_2|Y, X_1)] \\ &= [H(X_1) - H(X_1|Y)] + [H(X_2|X_1) - H(X_2|Y, X_1)] = I(X_1;Y) + I(X_2;Y|X_1). \end{aligned}$$

This completes our proof. □

Here's an example where the chain rule helps us calculate mutual information:

Example 3.2.5. *Let $J \in_R \{0, \ldots, k-1\}$ be chosen uniformly at random. Randomly choose $X, Y \in_R \{0, 1, \ldots, 2^k - 1\}$, encoded in binary, so that*

$$X = (X_0, \ldots, X_{k-1}) \quad \textit{and} \quad Y = (Y_0, \ldots, Y_{k-1}),$$

such that $X_{<j} = Y_{<j}$, $X_j \neq Y_j$, and $X_{>j}$ and $Y_{>j}$ are arbitrary. In other words, j is the first coordinate that X and Y differ.

Intuitively, $I(X;Y) = O(k/2)$ because we should learn 1/2 of the bits of X by seeing Y, since j is expected to reveal the first $k/2$ most significant bits (MSB) of X. As an exercise, pause and try proving this.

Proof. From the chain rule, we have two ways to decompose $I(X, J; Y)$:

$$I(X, J; Y) = I(X;Y) + I(J;Y|X) = I(J;Y) + I(X;Y|J).$$

It follows that

$$I(X;Y) = I(J;Y) + I(X;Y|J) - I(J;Y|X).$$

The first term $I(J;Y)$ is zero, since knowing something about Y won't tell us anything about J. The next term $I(X;Y|J)$ is more interesting: if we know $J = j$, then the jth bit is the first bit of X and Y to differ.

So, the mutual information conditioned on $J = j$ is $j + 1$ (since $Y_0, \ldots, Y_j$ determines $X_0, \ldots, X_j$). Thus, the conditional mutual information is

$$I(X;Y|J) = \mathbb{E}_J\left[I(X;Y|J=j)\right] = \mathbb{E}_J\left[j+1|J=j\right] = \frac{1}{k}\sum_{j=0}^{k-1} j+1 = \frac{k+1}{2}.$$

Finally, the last term $I(J;Y|X)$ gives how much information on J we gain from Y, if we know X. Since X and Y completely determine J (taken to be the first coordinate they differ), it follows that $I(J;Y|X) = H(J)$. Since J was uniform over a set with k elements, $H(J) = \lg k$. Thus,

$$I(X;Y) = \frac{k+1}{2} - \lg k = \frac{k}{2} + \lg\frac{k}{\sqrt{2}},$$

proving $I(X;Y) = O(k/2)$. □

Remark 3.2.6. *Notice at the end of the previous proof, if we let $k = 2$, then $I(X;Y) = 1/2$. Let's try computing this directly. By definition,*

$$I(X,Y) = H(X) - H(X|Y) = 2 - \mathbb{E}_Y[H(X)|Y=y].$$

To calculate $\mathbb{E}_Y[H(X)|Y=y]$, we note that there are four cases in how X_i and Y_i are related. In particular, the associated probabilities for $(X_0 \overset{?}{=} Y_0, X_1 \overset{?}{=} Y_1)$ are:

$$\Pr(=,=) = 0, \quad \Pr(=,\neq) = \frac{1}{2}, \quad \Pr(\neq,=) = \frac{1}{4}, \quad \Pr(\neq,\neq) = \frac{1}{4}.$$

Conditioned on $Y = y$, each of these cases completely determine X. And as these probabilities are independent of the value of y,

$$\mathbb{E}_Y[H(X)|Y=y] = \frac{1}{2}\lg 2 + \frac{1}{4}\lg 4 + \frac{1}{4}\lg 4 = \frac{3}{2}.$$

So, $I(X,Y)$ is indeed $1/2$. On the other hand, $I(X;Y|J)$ is explicitly calculated by

$$I(X;Y|J) = \mathbb{E}_J[I(X;Y|J=j)] = \frac{1}{2}I(X;Y|J=0) + \frac{1}{2}I(X;Y|J=1) = \frac{1}{2}\cdot 1 + \frac{1}{2}\cdot 2 = \frac{3}{2}.$$

This is another case where $I(X;Y) \leq I(X;Y|J)$.

This leads us to wonder when can we say $I(X;Y) \geq I(X;Y|Z)$ or $I(X;Y) \leq I(X;Y|Z)$? Certainly, both cases are possible. We've already seen two examples where $I(X;Y) \leq I(X;Y|Z)$. And are two examples of the other case:

Example 3.2.7. *We always have the following:*

$$I(X;Y) \geq I(X;Y|Y) = 0.$$

We may interpret $I(X;Y|Y)$ as the information shared by X and Y after explaining away Y *(expanding $I(X;Y|Y)$ gives the same interpretation and inequality).*

Example 3.2.8. *Let $X = \{0,1\}^3$ be three bits, $X = (X_0, X_1, X_2)$. Let Y be the first two bits $Y = (X_0, X_1)$ and let Z be the latter two bits $Z = (X_1, X_2)$. We easily see that $I(X;Y) = 2$, while $I(X;Y|Z) = 1$. Like before, $I(X;Y|Z)$ is the number of bits shared by X and Y after explaining away the bits in Z. So,*

$$I(X;Y) \geq I(X;Y|Z).$$

In general, there are two important cases:

1. If X and Z are independent (i.e. $X \coprod Z$), then $I(X;Z) = 0$. So $I(X;Y) \leq I(X;Y|Z)$ because

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

   Intuitively, even though X and Z may be independent, Z could contain information about X that may be retrieved with Y.

2. If X and Z are independent conditioned on Y (i.e. $X \coprod Z|Y$), then $I(X;Y|Z) \leq I(X;Y)$ since

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

   Intuitively, if Y explains away any correlation between X and Z, then symmetrically, Z might explain mutual information between X and Y.

Corollary 3.2.9. *If $X_1, \ldots, X_n$ are* independent *random variables, then*

$$I(X_1, \ldots, X_n; Y) \geq \sum_i I(X_i; Y).$$

Proof. We begin with

$$I(X_1, \ldots, X_n; Y) = \sum_i I(X_i; Y|X_{<i}),$$

where each of the $X_{<i}$ are independent to X_i. This is case 1 from above, where $I(X_i; Y|X_{<i}) \geq I(X_i; Y)$. Substituting back in, we get

$$I(X_1, \ldots, X_n; Y) \geq \sum I(X_i; Y).$$

□

Exercise 3.2.10. *Can a Shearer inequality for Mutual Information exist?*

Note that unlike entropy, mutual information is a measure on *random variables*, not on *distributions*. However, as it was defined in terms of entropy, we should be able to reason about the joint distribution of correlated random variables directly.

Let's return to one of the earlier equalities:

$$I(X;Y) = H(X) + H(Y) - H(X,Y).$$

And let's think back to one of our interpretations of entropy a distribution μ: the smallest (or optimal) expected number of bits needed to code an event with distribution μ. Perhaps this helps us further elucidate the meaning of $I(X;Y)$, which is the difference between $H(X) + H(Y)$ and $H(X,Y)$.

The first expression $H(X) + H(Y)$ gives the expected number of bits we would use to code events drawn from (X,Y) using the optimal coding scheme for X and Y separately. However, when X and Y are not independent random variables (i.e. they carry information about each other), we'd expect that using an optimal coding scheme designed for their joint distribution to perform better, using on average $H(X,Y)$ bits per event. It makes sense that $I(X;Y)$, their difference, gives a measure of how correlated X and Y are.

It's illuminating to expand $I(X;Y)$ back into their expectations:

$$I(X;Y) = \sum_{x \sim X} \mu(x) \lg \frac{1}{\mu(x)} + \sum_{y \sim Y} \mu(y) \lg \frac{1}{\mu(y)} - \mathbb{E}_{X,Y}\left[\lg \frac{1}{\mu(x,y)}\right].$$

We can combine the first two expectations by introducing the factors $\mu(y|x)$ and $\mu(x|y)$, respectively:

$$I(X;Y) = \sum_{(x,y)\sim(X,Y)} \left[\mu(y|x)\cdot\mu(x)\lg\frac{1}{\mu(x)} + \mu(x|y)\cdot\mu(y)\lg\frac{1}{\mu(y)}\right] - \mathbb{E}_{X,Y}\left[\lg\frac{1}{\mu(x,y)}\right]$$

Of course, since $\mu(x,y) = \mu(y|x)\cdot\mu(x) = \mu(x|y)\cdot\mu(y)$, we really have

$$I(X;Y) = \sum_{(x,y)\sim(X,Y)} \mu(x,y)\left[\lg\frac{1}{\mu(x)} + \lg\frac{1}{\mu(y)}\right] - \mathbb{E}_{X,Y}\left[\lg\frac{1}{\mu(x,y)}\right].$$

Recognizing the summation (the non-grayed out terms) as an expectation over the joint distribution X,Y, we may combine expectations:

$$I(X;Y) = \mathbb{E}_{X,Y}\left[\lg\frac{1}{\mu(x)\mu(y)} - \lg\frac{1}{\mu(x,y)}\right] = \mathbb{E}_{X,Y}\left[\lg\frac{\mu(x,y)}{\mu(x)\mu(y)}\right]. \tag{3.3}$$

In fact, now we see that Equation 3.3 formalizes the notion we stated earlier: mutual information measures the expected improvement of bits by using a coding scheme optimized for the joint distribution $\mu(x,y)$ over using a coding scheme optimized for X and Y individually, $\mu(x)\mu(y)$.

This discussion helps motivate our next concept: the Kullback-Leibler (KL) divergence, which helps us generalize a way to measure the relative entropy of a distribution ν with respect to a 'true' distribution μ, in the same way that we compared $\mu(x)\mu(y)$ with the actual joint distribution $\mu(x,y)$.

3.3 Kullback-Leibler Divergence

KL divergence, relative entropy, or information divergence, acts like a distance measure between probability distributions on the same universe.

Definition 3.3.1 (KL Divergence)**.** *For two distributions μ,ν on the same universe U, the* KL divergence *between μ and ν is*

$$\mathsf{D}\left(\frac{\mu}{\nu}\right) := \sum_{x\in U}\mu(x)\lg\frac{\mu(x)}{\nu(x)} = \mathbb{E}_{\mu}\left[\lg\frac{\mu(x)}{\nu(x)}\right].$$

Notation 3.3.2. *It is more common in literature for KL to be denoted by $\mathsf{D}(\mu||\nu)$, although as a result, equations easily run off the page. As a another note, following the usual convention in this course, we will interchangeably write $\mathsf{D}(X||Y)$ to represent the KL divergence between the underlying distributions of X and Y (this will be more convenient soon when we start talking abut conditional random variables).*

An immediate result of our discussion at the end of the previous section is the following lemma:

Lemma 3.3.3. $I(X;Y) = \mathsf{D}\left(\frac{\mu(x,y)}{\mu(x)\mu(y)}\right)$.

Proof. See Equation 3.3. □

Although $I(X;Y)$ is symmetric, KL divergence is *not* (notice in the previous lemma, writing $I(Y;X)$ is tantamount to switching x and y in the KL term, and not flipping the arguments. So even though we'd like to think of KL divergence as a distance measure between distributions, it is not a *metric*. However, it is a *positive-definite function.* Let's take a closer look at its properties.

3.3.1 Important and useful properties of KL

Example 3.3.4 (KL is not symmetric)**.** *Let μ be the distribution of a fair coin while ν the distribution of a completely biased coin (e.g. $\nu(\mathrm{H}) = 1$ while $\nu(\mathrm{T}) = 0$). Then,*

$$\mathsf{D}\left(\frac{\mu}{\nu}\right) = \mu(\mathrm{H}) \lg \frac{\mu(\mathrm{H})}{\nu(\mathrm{H})} + \mu(\mathrm{T}) \lg \frac{\mu(\mathrm{T})}{\nu(\mathrm{T})} = \frac{1}{2} \cdot \lg \frac{1/2}{1} + \frac{1}{2} \lg \frac{1/2}{0} = \infty,$$

and

$$\mathsf{D}\left(\frac{\nu}{\mu}\right) = \nu(\mathrm{H}) \lg \frac{\nu(\mathrm{H})}{\mu(\mathrm{H})} = \lg 2 = 1.$$

In fact, this shows that KL divergence may not be finite. In particular,

Proposition 3.3.5. $\mathsf{D}\left(\frac{\mu}{\nu}\right) = \infty \iff \mathrm{Supp}(\mu) \not\subseteq \mathrm{Supp}(\nu)$.

Proof. The KL divergence is infinite iff there is a point $x \in U$ such that $\mu(x) \neq 0$ and $\nu(x) = 0$. That is, $\mathrm{Supp}(\mu) \not\subseteq \mathrm{Supp}(\nu)$. □

This proposition lets us easily come up with examples to show KL is not symmetric, just by choosing distributions μ and ν where the support of one is strictly contained in the support of the other. Before we prove the KL divergence is positive-semidefinite, let's take a look at a few more examples.

Example 3.3.6. *We can generalize Example 3.3.4, considering two Bernoulli distributions with corresponding probabilities p and q. Then,*

$$\mathsf{D}\left(\frac{p}{q}\right) = p \lg \frac{p}{q} + (1-p) \lg \frac{1-p}{1-q}.$$

For the case where $p = \frac{1}{2} + \epsilon$ and $q = \frac{1}{2}$, we get

$$\mathsf{D}\left(\frac{\frac{1}{2}+\epsilon}{\frac{1}{2}}\right) = \left((2\epsilon)^2 - \frac{(2\epsilon)^2}{2}\right) + \left(\frac{(2\epsilon)^4}{3} - \frac{(2\epsilon)^4}{4}\right) + \left(\frac{(2\epsilon)^6}{5} - \frac{(2\epsilon)^6}{6}\right) + \cdots = O(\epsilon^2).$$

And in fact, recall the quadratic decay of $H\left(\frac{1}{2}+\epsilon\right)$, for it is not coincidence that both are order of ϵ^2.

Example 3.3.7. *Let $X \sim \mu$ be a random variable and $\mathrm{Supp}(X) = [n]$. Then, we can expand out*

$$\mathsf{D}\left(\frac{X}{\mathrm{uniform}(X)}\right) = \mathbb{E}_\mu\left[\lg \mu(x) - \lg \frac{1}{n}\right] = -H(X) + \lg n,$$

giving us the identity

$$H(X) = \lg(n) - \mathsf{D}\left(\frac{X}{\mathrm{uniform}(X)}\right).$$

The term $\lg(n)$ is the maximum entropy, while $\mathsf{D}(X||\mathrm{uniform}(X))$ is the amount of bits we save coding according to X rather than the uniform distribution. Intuitively, the difference should give $H(X)$, as above.

Now, we'll work to show that the KL divergence is positive semi-definite:

Proposition 3.3.8. $\mathsf{D}\left(\frac{\mu}{\nu}\right) \geq 0$*, and* $\mathsf{D}\left(\frac{\mu}{\nu}\right) = 0$ *if and only if* $\mu = \nu$*.*

Lemma 3.3.9 (Gibb's inequality)**.** $\mathsf{D}\left(\frac{\mu}{\nu}\right) \geq 0.$

Proof. Since KL divergence is the expectation of a log, we'll apply a common trick—flip the argument of the logarithm, so now we have an expectation of a convex function. We can apply Jensen's inequality to move the expectation past the logarithm. Then, after some basic algebra, we get Gibb's inequality:

$$\begin{aligned}
\mathsf{D}\left(\frac{\mu}{\nu}\right) &= \mathbb{E}_{\mu}\left[\lg \frac{\mu(x)}{\nu(x)}\right] \\
&= \mathbb{E}_{\mu}\left[-\lg \frac{\nu(x)}{\mu(x)}\right] \geq -\lg\left(\mathbb{E}_{\mu}\left[\frac{\nu(x)}{\mu(x)}\right]\right) \\
&= -\lg\left(\sum_{x} \mu(x)\frac{\nu(x)}{\mu(x)}\right) = -\lg(1) = 0.
\end{aligned}$$

□

To prove the convexity of KL, we'll use the log-sum inequality, stated here without proof:

Lemma 3.3.10 (Log-Sum Inequality, [D2014] Remark 5.3)**.** *Let $a_1, \ldots, a_n, b_1, \ldots, b_n$ be nonnegative reals. Then*

$$\sum_{i=1}^{n} a_i \log \frac{a_i}{b_i} \geq \left(\sum_{i=1}^{n} a_i\right) \log\left(\frac{\sum_{i=1}^{n} a_i}{\sum_{i=1}^{n} b_i}\right).$$

Lemma 3.3.11 (Convexity of KL)**.** *Let $\mu = \sum_i \alpha_i \mu_i$, $\nu = \sum_i \alpha_i \nu_i$. Then*

$$\sum_{i} \alpha_i \mathsf{D}\left(\frac{\mu_i}{\nu_i}\right) \geq \mathsf{D}\left(\frac{\mu}{\nu}\right).$$

Proof. Expand out the left-hand side:

$$\sum_{i=1}^{n} \sum_{x \in X} \alpha_i \mu_i(x) \lg \frac{\mu_i(x)}{\nu_i(x)} = \sum_{x \in X} \sum_{i=1}^{n} \alpha_i \mu_i(x) \lg \frac{\alpha_i \mu_i(x)}{\alpha_i \nu_i(x)}.$$

We can now apply the log-sum inequality to the inner sum, whence

$$\sum_{i=1}^{n} \alpha_i \mathsf{D}\left(\frac{\mu_i}{\nu_i}\right) \geq \sum_{x \in X}\left(\sum_{i=1}^{n} \alpha_i \mu_i(x)\right) \lg \frac{\sum_i \alpha_i \mu_i(x)}{\sum_i \alpha_i \nu_i(x)} = \sum_{x \in X} \mu(x) \lg \frac{\mu(x)}{\nu(x)} = \mathsf{D}\left(\frac{\mu}{\nu}\right).$$

□

It should be no surprise that KL also satisfies the chain rule:

Proposition 3.3.12 (Chain Rule for KL Divergence)**.**

$$\mathsf{D}\left(\frac{X_1, \ldots, X_n}{Y_1, \ldots, Y_n}\right) = \sum_{i} \mathbb{E}_{x_{<i}}\left[\mathsf{D}\left(\frac{X_i | x_{<i}}{Y_i | x_{<i}}\right)\right].$$

Proof. Prove as exercise using the properties of the log function. □

An immediate and useful corollary from convexity and chain rule is:

Lemma 3.3.13. *If* $Y_1, \ldots, Y_n$ *are* independent *random variables, then*

$$\mathsf{D}\left(\frac{X_1, \ldots, X_n}{Y_1, \ldots, Y_n}\right) \geq \sum_i \mathsf{D}\left(\frac{X_i}{Y_i}\right).$$

Proof. The proof is immediate by the chain rule and convexity of KL, observing that we can drop the conditioning on $x_{<i}$ in the *denominator*. First, expand out the left-hand side:

$$\mathsf{D}\left(\frac{X_1, \ldots, X_n}{Y_1, \ldots, Y_n}\right) = \sum_{i=1}^{n} \mathbb{E}_{X_{<i}}\left[\mathsf{D}\left(\frac{X_i|x_{<i}}{Y_i|x_{<i}}\right)\right].$$

We can drop the conditioning on x_i in the denominator because the Y_i's are independent. And so the above is equal to

$$\sum_{i=1}^{n} \mathbb{E}_{X_{<i}}\left[\mathsf{D}\left(\frac{X_i|x_{<i}}{Y_i}\right)\right].$$

For each $x_{<i}$, we get a distribution for $\Pr(X_i|x_{<i})$. By definition, the marginal distribution for X_i is the sum

$$\sum_{x_{<i}} \Pr(X_{<i} = x_{<i}) \Pr(X_i|x_{<i}).$$

But these $\Pr(X_{<i} = x_{<i})$ are exactly parts of the sum in the expectation:

$$\sum_{i=1}^{n} \mathbb{E}_{X_{<i}}\left[\mathsf{D}\left(\frac{X_i|x_{<i}}{Y_i}\right)\right] = \sum_{i=1}^{n} \sum_{x_{<i}} \Pr(X_{<i} = x_{<i}) \Pr(X_i|x_{<i}) \mathsf{D}\left(\frac{X_i|x_{<i}}{Y_i}\right).$$

So, we may apply convexity, and replacing the right-hand side of the inequality with the marginal distribution:

$$\sum_{i=1}^{n} \mathbb{E}_{X_{<i}}\left[\mathsf{D}\left(\frac{X_i|x_{<i}}{Y_i}\right)\right] \geq \sum_{i=1}^{n} \mathsf{D}\left(\frac{X_i}{Y_i}\right),$$

as desired. □

References

[A1981] Alon, Noga. "On the number of subgraphs of prescribed type of graphs with a given number of edges." Israel Journal of Mathematics 38 (1981): 116-130.

[D2014] Dannan, Fozi M., Patrizio Neff, and Christian Thiel. "On the sum of squared logarithms inequality and related inequalities." arXiv preprint arXiv:1411.1290 (2014).

[FK1996] Friedgut, Ehud, and Jeff Kahn. "On the number of copies of one hypergraph in another." Israel Journal of Mathematics 105.1 (1998): 251-256.

[G2014] Galvin, David. "Three tutorial lectures on entropy and counting." arXiv preprint arXiv:1406.7872 (2014).

[SU2011] Scheinerman, Edward R., and Daniel H. Ullman. Fractional graph theory: a rational approach to the theory of graphs. Courier Corporation, 2011.