ECE 598 - Fall 2025

Lecture 2: Information Measures

August 28, 2025

Instructor: Shubhanshu Shekhar

We begin this lecture by stating the defintions of the three key information measures for discrete distributions, and discuss how they can be extended to more general distributions. Next, we establish some elementary properties of these information measures, and apply those ideas to study the problem of generating random bits.

1 Information Measures

For simplicity, let us assume that $\mathcal{X}$ and $\mathcal{Y}$ denote discrete alphabets that are either finite or countable (places where finiteness is important will be explicitly mentioned). For random variables X and Y taking values in the alphabets $\mathcal{X}$ and $\mathcal{Y}$ respectively, we use P_X and P_Y to denote their probability distributions; that is, $P_X : 2^{\mathcal{X}} \to [0,1]$ and $P_Y : 2^{\mathcal{Y}} \to [0,1]$ map subsets of $\mathcal{X}$ and $\mathcal{Y}$ to the unit interval. Furthermore, we use $p_X : \mathcal{X} \to [0,1], p_Y : \mathcal{Y} \to [0,1]$ to denote their probability mass functions, with the property

$$P_X(E) = \sum_{x \in E} p_X(x), \quad \text{and} \quad P_Y(F) = \sum_{y \in F} p_Y(y),$$

for $E \subset \mathcal{X}$ and $F \subset \mathcal{Y}$. Similarly, we can define joint distributions P_{XY} (with pmf p_{XY}) and conditional distributions $P_{Y|X}$ (with pmf $p_{Y|X}$). We now introduce the definition of the first important information measure, entropy.

Definition 1.1 (Entropy)**.** Suppose $X \sim P_X$ is an $\mathcal{X}$-valued discrete random variable with pmf p_X. Then, the entropy of X (equivalently entropy of the distribution P_X, or pmf p_X) is defined as

$$H(X) \equiv H(P_X) \equiv H(p_X) = \sum_{x \in \mathcal{X}} p_X(x) \log 1/p_X(x) = \mathbb{E}_{X \sim P_X}\left[\log\left(\frac{1}{p_X(X)}\right)\right].$$

For a pair of random variables (X, Y) with a joint distribution P_{XY}, we can define the joint entropy $H(X,Y) \equiv H(P_{XY})$ and conditional entropy $H(Y|X) \equiv H(P_{Y|X}|P_X)$ as

$$H(X,Y) = \sum_{x,y \in \mathcal{X} \times \mathcal{Y}} p_{XY}(x,y) \log 1/p_{XY}(x,y) = \mathbb{E}_{(X,Y) \sim P_{XY}}\left[\log(1/p_{XY}(X,Y))\right], \quad \text{and}$$

$$H(Y|X) = \sum_{x \in \mathcal{X}} p_X(x) \sum_{y \in \mathcal{Y}} p_{Y|X}(y|x) \log(1/p_{Y|X}(y|x)) = \mathbb{E}_{X \sim P_X}\mathbb{E}_{Y|X \sim P_{Y|X}}\left[\log(1/P_{Y|X}(Y|X))\right].$$

Since each $p_X(x) \in [0,1]$, it follows that $\log(1/p_X(x)) \geq 0$ for all $x \in \mathcal{X}$, which implies that the entropy of P_X is always nonnegative.

Exercise: Suppose $X \sim P_X$ and $f : \mathcal{X} \to \{1, \ldots, |\mathcal{X}|\}$ be a one-to-one map. Then, what is the relation between $H(X)$ and $H(f(X))$.

Example 1.2. *Consider the simplest case of $\mathcal{X} = \{0, 1\}$ and $X \sim \text{Bernoulli}(p)$. Then, the entropy of X, often referred to as the binary entropy and denoted by $h_b(p)$, is equal to*

$$H(X) \equiv h_b(p) = -p \log p - \bar{p} \log \bar{p}, \quad \textit{where} \quad \bar{p} = 1 - p.$$

The variation of $h_b(p)$ with p is shown in the left plot in Figure 1. Notice that qualitatively $h(p)$ looks similar to the variance $p(1-p)$, which is another common measure of uncertainty.

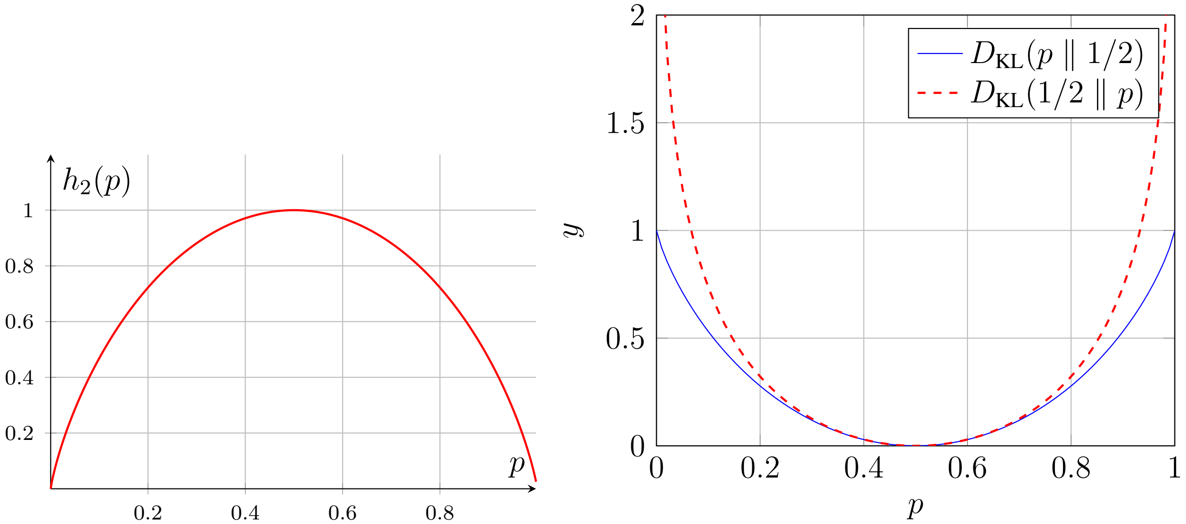

Figure 1: Binary entropy $h_b(p)$ on the left, and $D_{\text{KL}}(p \parallel 1/2)$ and $D_{\text{KL}}(1/2 \parallel p)$ on the right.

We now present the definition of the next term, called relative entropy or KL-divergence, that is a notion of distance between two distributions.

Definition 1.3 (Relative Entropy)**.** Suppose X and X' are two $\mathcal{X}$-valued random variables with distributions P and Q, with pmfs p and q respectively. Then, the relative entropy $D_{\text{KL}} : \mathcal{P}(\mathcal{X}) \times \mathcal{P}(\mathcal{X}) \to \mathbb{R}$ between P and Q is defined as

$$D_{\text{KL}}(P \parallel Q) = \sum_{x \in \mathcal{X}} p(x) \log \left(\frac{p(x)}{q(x)} \right).$$

This definition extends naturally to joint distributions P_{XY} and Q_{XY} both on $\mathcal{X} \times \mathcal{Y}$ as

$$D_{\text{KL}}(P_{XY} \parallel Q_{XY}) = \sum_{x,y} p_{XY}(x,y) \log \left(\frac{p_{XY}(x,y)}{q_{XY}(x,y)} \right) = \mathbb{E}_{P_{XY}} \left[\log \left(\frac{p_{XY}(X,Y)}{q_{XY}(X,Y)} \right) \right],$$

and to the case of conditional distributions

$$\begin{aligned} D_{\text{KL}}(P_{Y|X} \parallel Q_{Y|X} \mid P_X) &= \sum_{x \in \mathcal{X}} p_X(x) \sum_{y \in \mathcal{Y}} p_{Y|X}(y|x) \log \left(\frac{p_{Y|X}(y|x)}{q_{Y|X}(y|x)} \right) \\ &= \mathbb{E}_{P_X} \left[\mathbb{E}_{P_{Y|X}} \left[\log \left(\frac{p_{Y|X}(Y|X)}{q_{Y|X}(Y|X)} \right) \right] \right]. \end{aligned}$$

The conditional relative entropy between $P_{Y|X}$ and $Q_{Y|X}$ averaged according to the marginal P_X can be understood as follows for the case of finite $\mathcal{X}, \mathcal{Y}$: We calculate the relative entropy between $P_{Y|X=x}$ and $Q_{Y|X=x}$ which are the rows of the transition probability matrices $P_{Y|X}$ and $Q_{Y|X}$ corresponding to $X = x$, and then we average these for all x according to the marginal P_X.

Example 1.4. *Consider* $P = \mathrm{Bernoulli}(p)$ *and* $Q = \mathrm{Bernoulli}(0.5)$. *Then, we have*

$$D_{KL}(P \parallel Q) = d_{KL}(p \parallel 0.5) = p \log(2p) + \bar{p} \log(2\bar{p}).$$

The variation of the binary relative entropy with p *is shown in the right plot on Figure 1. The figure indicates that* $d_{KL}(p \parallel 0.5)$ *appears like a quadratic function, and in fact we can show that* $d_{KL}(p \parallel 0.5) = \mathcal{O}(|p - 0.5|^2)$.

The final information measure we introduce is the mutual information, which can be defined in terms of the either the relative entropy or the entropy.

Definition 1.5 (Mutual Information)**.** The mutual information $I : \mathcal{P}(\mathcal{X} \times \mathcal{Y}) \to \mathbb{R}$ between two random variables X and Y with joint distribution P_{XY} is defined as

$$I(X;Y) \equiv I(P_{XY}) = \sum_{x,y \in \mathcal{X} \times \mathcal{Y}} p_{XY}(x,y) \log \left(\frac{p_{XY}(x,y)}{p_X(x) p_Y(y)} \right).$$

It immediately follows that the above definition is equivalent to

$$I(X;Y) = I(P_{XY}) = D_{\mathrm{KL}}(P_{XY} \parallel P_X \times P_Y),$$

where $P_X \times P_Y$ denotes the product of the marginals.

We will see later on that mutual information can also be written in terms of entropy as follows:

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

The last two definitions provide some intuitive understanding of information as the reduction of uncertainty: if $H(Y)$ is the total uncertainty about Y and $H(Y|X)$ is the uncertainty remaining about Y given the knowledge of X, then $I(X;Y)$ is the reduction in uncertainty about Y that is achieved by knowing X.

1.1 Extension to general alphabets

While we have only introduced the information measures for the case of discrete distributions, we now show that for the case of relative entropy (and hence mutual information), this restriction to finite alphabets is essentially without loss of generality. In particular, suppose $\mathcal{X}$ denotes some general alphabet, and $\mathcal{B}$ is a sigma-field or sigma-algebra of subsets of $\mathcal{X}$. Let $\mathcal{E}$ denote the collection of all finite partitions on $\mathcal{X}$ constructed using sets in $\mathcal{B}$. Let $\pi = (E_1, \ldots, E_m)$ denote one such partition of $\mathcal{X}$, and for any two probability measures P, Q on $(\mathcal{X}, \mathcal{B})$, denote the restriction of P, Q to π with P_π and Q_π. That is, P_π is a discrete distribution over an alphabet of size m, with pmf $p_\pi \equiv (P(E_1), \ldots, P(E_m))$. With these notations, we can now state the following result.

Definition 1.6. Let $(\mathcal{X}, \mathcal{B})$ denote a general measurable space, and let $P : \mathcal{B} \to [0,1]$ and $Q : \mathcal{B} \to [0,1]$ denote two probability measures defined on this measurable space. Then, the relative entropy between P and Q is defined as

$$D_{\text{KL}}(P \parallel Q) := \sup_{\pi \in \mathcal{E}} D_{\text{KL}}(P_\pi \parallel Q_\pi) = \sup_{\pi \in \mathcal{E}} \sum_{E \in \pi} P(E) \log \left(\frac{P(E)}{Q(E)} \right).$$

Since mutual information between X and Y is also defined as the relative entropy between the joint distribution P_{XY} and the product of its marginals, $P_X \times P_Y$, the above definition also immediately gives us an analogous variational definition of mutual information.

Definition 1.7. Suppose $(\mathcal{X}, \mathcal{B}_X)$ and $(\mathcal{Y}, \mathcal{B}_Y)$ denote two measurable spaces, and let $(X, Y) \sim P_{XY}$ be a joint distribution on the product space $(\mathcal{X} \times \mathcal{Y}, \mathcal{B}_X \times \mathcal{B}_Y)$. Then, the mutual information between (X, Y) is equal to

$$I(X;Y) \equiv I(P_{XY}) = \sup_{\pi \in \mathcal{E}_{XY}} \sum_{E \in \pi} P_{XY}(E) \log \left(\frac{P_{XY}(E)}{(P_X \otimes P_Y)(E)} \right).$$

where $\mathcal{E}_{XY}$ is the collection of all finite measurable partition of the product space $\mathcal{X} \times \mathcal{Y}$. Since the product sigma-algebra is generated by the "rectangular sets", in fact, we can show that we can restrict our attention to such rectangular partitions:

$$I(X;Y) = \sup_{\pi_X \in \mathcal{E}_X, \pi_Y \in \mathcal{E}_Y} \sum_{E_x \in \pi_X} \sum_{E_y \in \pi_Y} P_{XY}(E_X \times E_Y) \log \left(\frac{P_{XY}(E_X \times E_Y)}{P_X(E_X) P_Y(E_Y)} \right).$$

Thus, the key takeaway from these definitions is that our definitions of D_{KL} and I for discrete distributions is essentially without loss of generality. Everything we prove for discrete distributions can be extended, with some careful limiting arguments, to the case of most general distributions.

The case of entropy is a bit more nuanced. Suppose we define the following for an $\mathcal{X}$ valued random variable X:

$$H'(X) := \sup_{\pi \in \mathcal{E}} H(X_\pi) = \sup_{\pi \in \mathcal{E}} \sum_{E \in \pi} -P_X(E) \log \left(P_X(E) \right).$$

It can be verified that the above definition of H' is finite only when the random variable X is discrete (with possibly countable support). [for example, try this out on Uniform$[0,1]$.] A simple alternative might be to define the so-called *differential entropy*

$$h(X) = \int -\log(p_X) p_X d\mu,$$

where p_X denotes the density of P_X with respect to some reference measure μ. However, this approach has some drawbacks: (i) $h(X)$ can be negative, (ii) the value of $h(X)$ is dependent on the reference measure μ, (iii) $h(X)$ loses some operational meanings (as the fundamental limit of compression) that the usual discrete entropy $H(X)$ possesses.

Remark 1.8. An alternative approach to defining entropy for general distributions is through the notion of *information dimension* as introduced by Rényi (1959). Informally, the idea is to define the information dimension

$$d_{\text{info}} = \lim_{\delta \downarrow 0} \frac{H(X_\delta)}{\log(1/\delta)},$$

where X_δ is the discretization of the random variable X over a partition of $\mathcal{X}$ whose "granularity" is δ (i.e., we require the space $\mathcal{X}$ to be a metric space, and then require the diameter of sets in the partition to be no larger than 2δ). Then, it turns out that an appropriate notion of entropy is

$$h(X) = \lim_{\delta \to 0} \left[H(X_\delta) - d_{\text{info}} \log(1/\delta) \right],$$

assuming this limit exists.
Exercise: What is d_{info} if X is a discrete distribution? What if X is $\mathrm{Uniform}([0,1])$?

2 Properties of Information Measures

In this section, we state and prove certain elementary proprties of the information measures. Like most inequalities, these results can essentially be traced back to the concavity of the map $x \mapsto x \log x$ or concavity of $x \mapsto \log x$.

We begin with the fundamental result, often referred to as Gibbs' inequality.

Theorem 2.1 (Gibbs Inequality). *Suppose P and Q are two distributions over a common finite alphabet $\mathcal{X}$. Then, we have*

$$D_{KL}(P \parallel Q) \geq 0,$$

with equality if and only if $P = Q$.

Proof. Recall that the function $f(x) = x \log x$ is convex, and thus we have

$$D_{\text{KL}}(P \parallel Q) = \mathbb{E}_Q \left[f\left(\frac{P(X)}{Q(X)} \right) \right] \overset{(i)}{\geq} f\left(\mathbb{E}_Q \left[\frac{P(X)}{Q(X)} \right] \right) = f(1) = 0,$$

where (i) follows due to Jensen's inequality and the convexity of f. For this inequality to hold with equality, we require $P(X)/Q(X) = 1$ for all $x \in \mathcal{X}$, which happens if and only if $P = Q$. □

2.1 Nonnegativity of Information Measures

This simple inequality leads to several important consequences, some of which we list below:

- *Nonnegativity of mutual information.* Since $I(X;Y) = D_{\text{KL}}(P_{YX} \parallel P_X \times P_Y)$, it immediately follows that the mutual information between two random variables is also always non-negative. Furthermore, $I(X;Y) = 0$ if and only if $P_{XY} = P_X \times P_Y$ or $X \perp Y$. Thus, mutual information is a notion of correlation between X and Y.

- *Conditioning reduces entropy.* Since $I(X;Y) = H(X) - H(X|Y) \geq 0$, we have $H(X|Y) \leq H(X)$. Similary, we have $H(Y|X) \leq H(Y)$. In both cases, we have equality only if $X \perp Y$.

2.2 Chain Rules

The next important result is about chain rules.

Theorem 2.2 (Chain Rules)**.** *Let $X_1, X_2, \ldots, X_n$ denote a collection of random variblables taking values in $\mathcal{X}_1, \mathcal{X}_2, \ldots, \mathcal{X}_n$ respectively. Consider two possible joint distributions of X^n, denoted by P_{X^n} and Q_{X^n}. Then, we have the following:*

$$D_{KL}(P_{X^n} \parallel Q_{X^n}) = \sum_{i=1}^{n} D_{KL}\left(P_{X_i|X^{i-1}} \parallel Q_{X_i|X^{i-1}} \mid P_{X^{i-1}}\right).$$

Similarly, for the case of mutual information and entropy, we have

$$I(X^n; Y) = \sum_{i=1}^{n} I(X_i; Y \mid X^{i-1}), \quad \textit{and} \quad H(X^n) = \sum_{i=1}^{n} H(X_i|X^{i-1}) \leq \sum_{i=1}^{n} H(X_i).$$

These statements can be proved by simply factoring out the joint pmfs involved in the definitions into appropriate products of the conditional pmfs. We write the explicit steps for the case of entropy with $n = 2$, but the same idea works for the other two measures as well and general n:

$$\begin{aligned} H(X_1, X_2) &= \sum_{x_1, x_2} -p_{X_1,X_2}(x_1, x_2) \log p_{X_1,X_2}(x_1, x_2) \\ &= \sum_{x_1, x_2} -p_{X_1,X_2}(x_1, x_2) \left(\log p_{X_1}(x_1) + \log p_{X_2|X_1}(x_2|x_1)\right) \\ &= -\sum_{x_1} p_{X_1}(x_1) \log p_{X_1}(x_1) - \sum_{x_1} p_{X_1}(x_1) \sum_{x_2} p_{X_2|X_1}(x_2|x_1) \log p_{X_2|X_1}(x_2|x_1) \\ &= H(X_1) + \sum_{x_1} p_{X_1}(x_1) H(X_2|X_1 = x_1) = H(X_1) + H(X_2|X_1). \end{aligned}$$

This extends to the general $n \geq 2$ case directly by induction. Similar arguments work for relative entropy and mutual information (start with $n = 2$, and use induction).

2.3 Data Processing Inequalities

To state the next result, we need to introduce the notion of a channel. In the simplest case of discrete and finite alphabets $\mathcal{X}$ and $\mathcal{Y}$, a channel $K_{Y|X}$ is simply a transition probability matrix. For every input symbol $x \in \mathcal{X}$, the channel outputs $Y \sim K_{Y|X=x}$.

Theorem 2.3 (Data Processing Inequality)**.** *Let $K_{Y|X}$ denote a channel over alphabets $\mathcal{X}$ (input alphabet) and $\mathcal{Y}$ (ouptut alphabet). Let P_X, Q_X denote two input distributions, and let $P_Y = P_X K_{Y|X}$ and $Q_Y = Q_X K_{Y|X}$ denote the output distribution after passing P_X and Q_X through the same channel $K_{Y|X}$. Then, the data processing inequality says that*

$$D_{KL}(P_Y \parallel Q_Y) \leq D_{KL}(P_X \parallel Q_X).$$

In other words, the "distance" between two distributions cannot be increased by passing them through a common channel (or stochastic kernel).

The fundamental idea behind this statement can be summarized as:

> The output likelihood ratio at any y is a convex combination of the input likelihood ratios. Apply Jensen's.

Proof. To simplify the notation, we will write K instead of $K_{Y|X}$. Then, observe that $P_Y(y) = (P_X K)(y) = \sum_{x \in \mathcal{X}} P_X(x) K_{Y|X}(y|x)$ and $Q_Y(y) = (Q_X K)(y) = \sum_{x \in \mathcal{X}} Q_X(x) K_{Y|X}(y|x)$. Then, we have

$$D_{\text{KL}}(P_Y \parallel Q_Y) = \sum_{y \in \mathcal{Y}} P_Y(y) \log \left(\frac{P_Y(y)}{Q_Y(y)} \right) = \sum_{y \in \mathcal{Y}} Q_Y(y) \frac{P_X K(y)}{Q_X K(y)} \log \left(\frac{P_X K(y)}{Q_X K(y)} \right).$$

Let us introduce the notation $\ell_x = P_X(x)/Q_X(x)$ and $w_x(y) = Q_X(x)K(y|x)/Q_Y(y)$, and observe that

$$\sum_{x \in \mathcal{X}} w_x(y) = \frac{\sum_{x \in \mathcal{X}} Q_X(x) K(y|x)}{Q_Y(y)} = 1, \quad \text{and} \quad \sum_{x \in \mathcal{X}} w_x(y) \ell_x = \frac{P_X K(y)}{Q_X K(y)}.$$

Hence by the convexity of $f(u) = u \log u$, we have

$$\begin{aligned} D_{\text{KL}}(P_Y \parallel Q_Y) &= \sum_y Q_Y(y) f\left(\frac{P_X K(y)}{Q_X K(y)} \right) = \sum_y Q_Y(y) f\left(\sum_{x \in \mathcal{X}} w_x(y) \ell_x \right) \\ &\overset{(i)}{\leq} \sum_y Q_Y(y) \sum_{x \in \mathcal{X}} w_x(y) f\left(\ell_x\right) = \sum_{y,x} Q_X(x) K(y|x) \frac{P_X(x)}{Q_X(x)} \log \left(\frac{P_X(x)}{Q_X(x)} \right) \\ &= \sum_{x \in \mathcal{X}} P_X(x) \log \left(\frac{P_X(x)}{Q_X(x)} \right) \sum_y K(y|x) = D_{\text{KL}}(P_X \parallel Q_X), \end{aligned}$$

where (i) follows from an application of Jensen's inequality. □

This general result can be used to derive analogous DPI for mutual information and entropy. In particular, suppose $U \longrightarrow V \longrightarrow W$ form a Markov chain (that is, $U \perp W \mid V$). Then Theorem 2.3 can be used to show the following.

Corollary 2.4. *If $U \longrightarrow V \longrightarrow W$, then we have $I(U;W) \leq I(U;V)$, which implies $H(U|W) \leq H(U|V)$.*

Proof. Due to the Markov property, we have

$$P_{UW} = P_{UV} K_{W|V}, \quad \text{and} \quad P_U P_W = (P_U P_V) K_{W|V}.$$

Hence, by applying the DPI for relative entropy, we have

$$\begin{aligned} I(U;W) &= D_{\mathrm{KL}}(P_{UW} \parallel P_U P_W) = D_{\mathrm{KL}}(P_{UV} K_{W|V} \parallel (P_U P_V) K_{W|V}) \\ &\leq D_{\mathrm{KL}}(P_{UV} \parallel P_U P_V) = I(U;V). \end{aligned}$$

Replacing $I(U;W) = H(U) - H(U|W)$ and $I(U;V) = H(U) - H(U|V)$, we get the require result for entropy. □

3 Application: Extracting Purely Random Bits

Consider the following setting: Let $X^n = (X_1, \ldots, X_n)$ denote n i.i.d. draws from a $\mathrm{Bernoulli}(p)$ distribution, with unknown $p \in (0,1)$. Our goal is to extract L purely random bits from X^n; that is, we want to construct a procedure K such that $K(X^n) = (L, Z^L)$, where $L \in \{0, \ldots, n\}$ and $z^l \in \{0,1\}^l$ for each $l \in \{0, \ldots, n\}$, satisfying the property

$$\mathbb{P}\left(L = l, Z^L = z^l\right) = \mathbb{P}(L = l) 2^{-l}, \quad \text{for all} \quad z^l \in \{0,1\}^l, \ l \in \{0, \ldots, n\}. \tag{1}$$

In other words, conditioned on $L = l$, the random variable Z^l is uniformly distributed over $\{0,1\}^l$. This problem is often stated in the following way: *Suppose we have a coin with unknown probability of heads p. How many fair coin tosses can we simulate using n tosses of this possibly biased coin?*

The information theoretic limit. We first establish a fundamental, algorithm-agnostic, limit on the expected number of fair coin tosses that can be extracted from the biased coin. Let K denote any extraction procedure satisfying (1). Then, we have the following chain:

$$\begin{aligned} nh_b(p) &= nH(X_1) = H(X^n) = H(X^n) + H(K(X^n)|X^n) \\ &= H(X^n, K(X^n)) = H(K(X^n)) + H(X^n|K(X^n)) && \text{(chain rule)} \\ &\geq H(K(X^n)) = H(L, Z^L) && (\text{since } H(X^n|K(X^n)) \geq 0) \\ &= H(L) + H(Z^L|L) && \text{(chain rule)} \\ &\geq H(Z^L|L) = \sum_{l=0}^{n} \mathbb{P}(L = l) H(Z^l|L = l) && (\text{since } H(L) \geq 0) \\ &= \sum_{l=0}^{n} \mathbb{P}(L = l) l = \mathbb{E}[L]. && (Z^L|L = l \text{ is uniform}) \end{aligned}$$

Thus, we have proved that any feasible extractor must satisfy

$$\frac{\text{expected \# of fair coin tosses}}{\text{\# number of biased coin tosses}} = \frac{\mathbb{E}[L]}{n} \leq h(p) = -p \log p - \bar{p} \log \bar{p}. \tag{2}$$

Von Neumann's (suboptimal) extractor. We now describe a very popular suboptimal approach, attributed to John von Neumann, for extracting fair coin tosses from a biased coin. Assume $n = 2m$, and proceed as follows:

- Pair off the n coin tosses into m pairs $(X_1, X_2), \dots, (X_{2i-1}, X_{2i}), \dots, (X_{n-1}, X_n)$.
- For each $i \in \{1, \dots, m\}$, define B_i as follows:

$$B_i = \begin{cases} 1, & \text{if } X_{2i-1} \neq X_{2i}, \\ 0, & \text{if } X_{2i-1} = X_{2i}. \end{cases}$$

  Set $L = \sum_{i=1}^{n} B_i$.

- Let $I_1, \dots, I_L$ denote the indices of the pairs for which $B_i = 1$. Then, define $Z^L = (Z_1, \dots, Z_L)$, with $Z_j = 1$ if $(X_{2I_j-1}, X_{2I_j}) = (0, 1)$ and $Z_j = 0$ if $(X_{2I_j-1}, X_{2I_j}) = (1, 0)$.

Proposition 3.1. *We can show that the* (L, Z^L) *constructed using von Neumann's extractor satisfies* (1)*, and furthermore, we have*

$$\frac{\mathbb{E}[L]}{n} = 2p(1-p) \leq h(p).$$

Remark 3.2. To see that the above procedure is suboptimal, for $p = 0.5$ (i.e., if the given coin is already fair), we have $h(p) = 1$ bit, while $2p(1-p) = 0.5$. Hence, von Neumann's procedure discards half of the (already) fair coin tosses unnecessarily in this case.

Peres (1992) proposed a simple iterative strategy that feeds back the discarded bits (i.e., with $B_i = 0$) and an "XoR stream" into a recursive mechanism, to get an extraction procedure that achieves the optimal rate of (2) in the limit of $n \to \infty$ and infinite recursion steps (Proposition 3 of Peres (1992)).

References

Y. Peres. Iterating von Neumann's procedure for extracting random bits. *The Annals of Statistics*, pages 590–597, 1992.

A. Rényi. On the dimension and entropy of probability distributions. *Acta Mathematica Academiae Scientiarum Hungarica*, 10:193–215, 1959.