ENEE 420
FALL 2012
COMMUNICATIONS SYSTEMS

DATA COMPRESSION:

Reminder

With $a > 0$, the logarithm in base a of $x > 0$, denoted $\log_a x$, is the unique scalar such that

$$x = a^{\log_a x}.$$

With $x = 0$ we tdopt he usual convention of defining $\log_a x = -\infty$, and $x \log_a x = 0$ – The latter follows by an easy continuity argument since

$$\lim_{x \downarrow 0} x \log_a x = 0,$$

say by L'Hospital's rule. With $a > 0$ and $b > 0$, it is always the case that

$$\log_b x = (\log_b a) \cdot \log_a x, \quad x > 0$$

Throughout $\log_2 x$ is logarithm in base 2 of $x > 0$, and we use $\log x$ for the natural logarithm which corresponds to $a = e$.

Finite sources

Let $\mathcal{X}$ denote a finite set, hereafter called the *alphabet*, and we refer to an element x of $\mathcal{X}$ as a *symbol*. A probability mass function (pmf) $\boldsymbol{p} = (p(x),\ x \in \mathcal{X})$ on $\mathcal{X}$ is any collection of scalars indexed by $\mathcal{X}$ such that

$$0 < p(x) \leq 1,\ x \in \mathcal{X} \quad \text{with} \quad \sum_{x \in \mathcal{X}} p(x) = 1.$$

A source is simply a pair $(\mathcal{X}, \boldsymbol{p})$ where $\mathcal{X}$ is a finite alphabet and $\boldsymbol{p}$ is a pmf on $\mathcal{X}$. It is sometimes convenient to refer to such a source by the notation $X = (\mathcal{X}, \boldsymbol{p})$ where the $\mathcal{X}$-valued random variable $X : \Omega \to \mathcal{X}$ is defined on some probability space $(\Omega, \mathcal{F}, \mathbb{P})$ such that

(1) $$\mathbb{P}[X = x] = p(x), \quad x \in \mathcal{X}.$$

In short, we can think of $p(x)$ as the likelihood that the source generates symbol x. In principle we could have $p(x) = 0$ for *some* of the values of x in $\mathcal{X}$.

Extensions of a source

Consider a source $(\mathcal{X}, \boldsymbol{p})$ where $\mathcal{X}$ is a finite alphabet and $\boldsymbol{p}$ is a pmf on $\mathcal{X}$. For each $n = 1, 2, \ldots$, its n^{th} extension is the source $(\mathcal{X}^n, \boldsymbol{p}_n)$ where the pmf $\boldsymbol{p}_n$ on $\mathcal{X}^n$ is given by

$$p_n(\boldsymbol{x}^n) = \prod_{i=1}^{n} p(x_i), \quad \boldsymbol{x}^n = (x_1, \ldots, x_n) \in \mathcal{X}^n. \tag{2}$$

It is often useful to view this n^{th} extension in terms of an $\mathcal{X}^n$-valued random variable $\boldsymbol{X}^n : \Omega \to \mathcal{X}^n$ defined on some probability space $(\Omega, \mathcal{F}, \mathbb{P})$ such that

$$\mathbb{P}\left[\boldsymbol{X}^n = \boldsymbol{x}^n\right] = p_n(\boldsymbol{x}^n), \quad \boldsymbol{x}^n \in \mathcal{X}^n.$$

where

$$\boldsymbol{X}^n = (X_1, \ldots, X_n).$$

Under (2) it is easy to check that

$$\mathbb{P}\left[\boldsymbol{X}^n = \boldsymbol{x}^n\right] = \prod_{i=1}^{n} \mathbb{P}\left[X_i = x_i\right], \quad \boldsymbol{x}^n = (x_1, \ldots, x_n) \in \mathcal{X}^n.$$

In other words, the $\mathcal{X}$-valued random variables $X_1, \ldots, X_n$ are independent and identically distributed (i.i.d.) random variables, each distributed according to the pmf $\boldsymbol{p}$.

Divergence

With $a > 0$, the divergence (in base a) between the pmfs $\boldsymbol{p}$ and $\boldsymbol{q}$ on $\mathcal{X}$ is defined by

$$D_a(\boldsymbol{p}\|\boldsymbol{q}) := -\sum_{x \in \mathcal{X}} p(x) \log_a \left(\frac{q(x)}{p(x)}\right).$$

The basic bound

$$D_a(\boldsymbol{p}\|\boldsymbol{q}) \geq 0$$

holds with equality if and only if $\boldsymbol{p} = \boldsymbol{q}$.

Entropy

With $a > 0$, the entropy (in base a) of the pmf $\boldsymbol{p}$ on $\mathcal{X}$ is defined by

$$H_a(\boldsymbol{p}) := -\sum_{x \in \mathcal{X}} p(x) \log_a p(x).$$

This is sometimes denoted $H_a(\mathcal{X}, \boldsymbol{p})$ or $H_a(X)$ where the $\mathcal{X}$-valued random variable $X : \Omega \to \mathcal{X}$ is defined on some probability space $(\Omega, \mathcal{F}, \mathbb{P})$ such that (1) holds.

The basic bounds

$$0 \leq H_a(\boldsymbol{p}) \leq \log_a |\mathcal{X}|$$

hold, and we have

1. The lower bound is achieved if and only if the pmf $\boldsymbol{p}$ is degenerate, i.e.,

$$H_a(\boldsymbol{p}) = 0 \quad \text{if and only if } p(x) = 1 \quad \text{for some } x \in \mathcal{X};$$

2. The upper bound is achieved if and only if the pmf $\boldsymbol{p}$ is the uniform pmf on $\mathcal{X}$, i.e.,

$$H_a(\boldsymbol{p}) = \log_a |\mathcal{X}| \quad \text{if and only if } p(x) = \frac{1}{|\mathcal{X}|}, \quad x \in \mathcal{X}.$$

Compression codes

Let $\mathcal{B}^\star$ denote the collection of all binary words with *finite* length, i.e.,

$$\mathcal{B}^\star = \cup_{n=1}^\infty \{0,1\}^n.$$

A binary *compression* code, hereafter simply a code, for an $\mathcal{X}$-valued source is any mapping

$$C : \mathcal{X} \to \mathcal{B}^\star.$$

For each x in $\mathcal{X}$, $C(x)$ is known as the *codeword* associated with x under C. It is customary to refer to the collection $\{C(x),\ x \in \mathcal{X}\}$ of all codewords as the codebook for C, and to identify it with C.

Some terminology: A code $C : \mathcal{X} \to \mathcal{B}^\star$ is said to be

1. non-singular if $C(x) \neq C(y)$ for any pair of distinct symbols x, y in $\mathcal{X}$;

2. uniquely decipherable if the equality

$$C(x_1) \ldots C(x_n) = C(y_1) \ldots C(y_m)$$

for some $x_1, \ldots, x_n, y_1, \ldots, y_m$ in $\mathcal{X}$ implies

$$n = m \quad \text{and} \quad x_j = y_j,\ j = 1, \ldots, n.$$

3. prefix (or to have the prefix property) if for any symbol x in $\mathcal{X}$, no prefix of $C(x)$ is a codeword for some other symbol in $\mathcal{X}$.

Prefix codes are also known as instantaneous codes. We denote the collection of all prefix codes by $\mathcal{C}_{\rm Pref}$.

Length of codes

Given a code $C : \mathcal{X} \to \mathcal{B}^\star$, let $\ell_C(x)$ denote the length of the binary codeword $C(x)$ associated with the symbol x in $\mathcal{X}$. Given a source $X = (\mathcal{X}, \boldsymbol{p})$, the expected codeword length of the code $C : \mathcal{X} \to \mathcal{B}^\star$ is given by

$$\begin{aligned} L(C; \boldsymbol{p}) &:= \mathbb{E}\left[\ell_C(X)\right] \\ &= \sum_{x \in \mathcal{X}} \ell_C(x) p(x). \end{aligned} \tag{3}$$

Kraft Inequality

For any prefix code $C : \mathcal{X} \to \mathcal{B}^\star$, we have

$$\sum_{x \in \mathcal{X}} 2^{-\ell_C(x)} \leq 1.$$

Conversely, for any collection $(\ell(x),\ x \in \mathcal{X})$ of positive integers such that

$$\sum_{x \in \mathcal{X}} 2^{-\ell(x)} \leq 1,$$

there exists a prefix code $C : \mathcal{X} \to \mathcal{B}^\star$ such that

$$\ell_C(x) = \ell(x), \quad x \in \mathcal{X}.$$

Shannon encoding

Set

$$\ell_{\rm SH}(x) = \lceil \log_2 \frac{1}{p(x)} \rceil, \quad x \in \mathcal{X}.$$

Since $2^{\log_2 t} = t$ for all $t > 0$, we find

$$\begin{aligned} \sum_{x \in \mathcal{X}} 2^{-\ell_{\rm SH}(x)} &\leq \sum_{x \in \mathcal{X}} 2^{-\log_2 \frac{1}{p(x)}} \\ &= \sum_{x \in \mathcal{X}} 2^{\log_2 p(x)} \\ &= \sum_{x \in \mathcal{X}} p(x) = 1, \end{aligned} \tag{4}$$

and there exists a prefix code $C_{\rm SH} : \mathcal{X} \to \mathcal{B}^\star$ such that

$$\ell_{C_{\rm SH}}(x) = \ell_{\rm SH}(x), \quad x \in \mathcal{X}. \tag{5}$$

Any code satsifying (5) is known as Shannon encoding.

Note that

$$\begin{aligned} L(C_{\rm SH};\boldsymbol{p}) &= \sum_{x\in\mathcal{X}} p(x)\ell_{\rm SH}(x) \\ &\leq \sum_{x\in\mathcal{X}} p(x)\left(\log_2 \frac{1}{p(x)} + 1\right) \\ &= -\sum_{x\in\mathcal{X}} p(x)\log_2 p(x) + \sum_{x\in\mathcal{X}} p(x) \\ &= H_2(\boldsymbol{p}) + 1. \end{aligned} \tag{6}$$

Shannon encoding comes from within one bit of source entropy!

Average code length and entropy

Consider a prefix code $C : \mathcal{X} \to \mathcal{B}^\star$. Introduce the pmf $\boldsymbol{q}_C$ on $\mathcal{X}$ given by

$$q_C(x) = \frac{2^{-\ell_C(x)}}{\Sigma(C)}, \quad x \in \mathcal{X}$$

where

$$\Sigma(C) = \sum_{x\in\mathcal{X}} 2^{-\ell_C(x)}.$$

We have

$$L(C;\boldsymbol{p}) - H_2(\boldsymbol{p}) = D(\boldsymbol{p}\|\boldsymbol{q}_C) + \log_2\left(\frac{1}{\Sigma(C)}\right) \tag{7}$$

so that

$$L(C;\boldsymbol{p}) \geq H_2(\boldsymbol{p})$$

since $D(\boldsymbol{p}\|\boldsymbol{q}_C) \geq 0$ and $\Sigma(C) \leq 1$ by the Kraft inequality. Equality holds if and only if $D(\boldsymbol{p}\|\boldsymbol{q}_C) = 0$ and $\Sigma(C) = 1$. In other words, equality holds if and only if there exists positive integers $(n(x),\ x \in \mathcal{X})$ such that

$$p(x) = 2^{-n(x)}, \quad x \in \mathcal{X}.$$

A proof of (7)

$$
\begin{aligned}
L(C;\boldsymbol{p}) &= \sum_{x\in\mathcal{X}} \ell_C(x)p(x) \\
&= -\sum_{x\in\mathcal{X}} p(x)\log_2\left(2^{-\ell_C(x)}\right) \\
&= -\sum_{x\in\mathcal{X}} p(x)\log_2\left(\frac{2^{-\ell_C(x)}}{\Sigma(C)}\cdot\Sigma(C)\right) \\
&= -\sum_{x\in\mathcal{X}} p(x)\log_2\left(\frac{q_C(x)}{p(x)}\cdot p(x)\Sigma(C)\right) \\
&= -\sum_{x\in\mathcal{X}} p(x)\left(\log_2\left(\frac{q_C(x)}{p(x)}\right)+\log_2 p(x)+\log_2\Sigma(C)\right) \\
&= -\sum_{x\in\mathcal{X}} p(x)\log_2\left(\frac{q_C(x)}{p(x)}\right) - \sum_{x\in\mathcal{X}} p(x)\log_2 p(x) - \log_2\Sigma(C).
\end{aligned}
$$

Source coding Theorem (Shannon 1948)

The bounds

(8) $$H_2(\boldsymbol{p}) \le L_{\min}(\boldsymbol{p}) \le H_2(\boldsymbol{p}) + 1$$

hold where

$$L_{\min}(\boldsymbol{p}) := \min\left(L(C;\boldsymbol{p}):\ C\in\mathcal{C}_{\mathrm{Pref}}\right).$$

Moreover,

$$L_{\min}(\boldsymbol{p}) = H_2(\boldsymbol{p})$$

if and only if there exist positive integers $(n(x),\ x\in\mathcal{X})$

$$p(x) = 2^{-n(x)}, \quad x\in\mathcal{X}.$$

Reaching entropy

It is possible to construct examples of sources for which the upper bound in (8) is tight, i.e., for every ϵ in $(0,1)$, there exists a pmf $\boldsymbol{p}_\varepsilon$ on $\mathcal{X}$ such that

$$H_2(\boldsymbol{p}_\varepsilon) + 1 - \varepsilon \le L_{\min}(\boldsymbol{p}_\varepsilon) \le H_2(\boldsymbol{p}_\varepsilon) + 1$$

Let $C^\star_\varepsilon : \mathcal{X} \to \mathcal{B}^\star$ denote the corresponding optimal prefix code, i.e.,

$$L_{\min}(\boldsymbol{p}_\varepsilon) = \mathbb{E}\left[\ell_{C^\star_\varepsilon}(X)\right]$$

Now consider the n^{th} extension $(\mathcal{X}^n, \boldsymbol{p}_{\varepsilon,n})$ of this source. It seems reasonable to encode the symbol $X_1, \ldots, X_n$ according to the optimal prefix code $C^\star_\varepsilon$, resulting in the concatenated binary codeword

$$C^\star_\varepsilon(X_1) \ldots C^\star_\varepsilon(X_n),$$

with length

$$\sum_{k=1}^n \ell_{C^\star_\varepsilon}(X_k).$$

As a result, the expected codeword length is simply

$$\sum_{k=1}^n \mathbb{E}\left[\ell_{C^\star_\varepsilon}(X_k)\right],$$

so that

$$n\left(H_2(\boldsymbol{p}_\varepsilon) + 1 - \varepsilon\right) \leq \sum_{k=1}^n \mathbb{E}\left[\ell_{C^\star_\varepsilon}(X_k)\right] \leq n\left(H_2(\boldsymbol{p}_\varepsilon) + 1\right).$$

As a result, the *expected codeword length per symbol* satisfies

$$H_2(\boldsymbol{p}_\varepsilon) + 1 - \varepsilon \leq \frac{1}{n}\sum_{k=1}^n \mathbb{E}\left[\ell_{C^\star_\varepsilon}(X_k)\right] \leq H_2(\boldsymbol{p}_\varepsilon) + 1.$$

In other words, in this situation a deviation from the entropy bound of close to n bits will occur, a discrepancy that will grow large with n. Equivalently, this reflected in the expected codeword length per symbol being almost one bit away from the netropy of the source. A natural question then arises as to whether this can be improved. That is indeed so is now discussed:

Consider a finite source $X = (\mathcal{X}, \boldsymbol{p})$. Recall that $C_{\rm SH}$ denotes any prefix code for this source which implements Shannon encoding, i.e.,

$$\ell_{C_{SH}}(x) = \left\lceil \log_2 \frac{1}{p(x)} \right\rceil, \quad x \in \mathcal{X}.$$

Earlier we showed the bounds

$$H_2(\boldsymbol{p}) \leq \mathbb{E}\left[\ell_{C_{SH}}(X)\right] \leq H_2(\boldsymbol{p}) + 1. \tag{9}$$

Now, for a given $n = 2, \ldots$, let $C_{n,\mathrm{SH}}$ denote any prefix code which implements Shannon encoding for the n^{th} extension $(\mathcal{X}^n, \boldsymbol{p}_n)$ of this source. Applying the bounds (9) to this source we get

$$H_2(\boldsymbol{p}_n) \leq \mathbb{E}\left[\ell_{C_{n,\mathrm{SH}}}(X)\right] \leq H_2(\boldsymbol{p}_n) + 1, \tag{10}$$

or equivalently,

$$nH_2(\boldsymbol{p}) \leq \mathbb{E}\left[\ell_{C_{n,\mathrm{SH}}}(X)\right] \leq nH_2(\boldsymbol{p}) + 1. \tag{11}$$

Turning to the expected codeword length per symbol, we conclude that

$$H_2(\boldsymbol{p}) \leq \mathbb{E}\left[\frac{\ell_{C_{n,\mathrm{SH}}}(X)}{n}\right] \leq H_2(\boldsymbol{p}) + \frac{1}{n}. \tag{12}$$

It is now immediate to see that

$$\lim_{n\to\infty} \mathbb{E}\left[\frac{\ell_{C_{n,\mathrm{SH}}}(X)}{n}\right] = H_2(\boldsymbol{p}).$$

Entropy can be reached (in an asymptotic sense).

The more likely the symbol, the shorter its description

Consider a prefix code $C : \mathcal{X} \to \mathcal{B}^\star$. Define a new code $C' : \mathcal{X} \to \mathcal{B}^\star$ as follows: Pick distinct x and y in $\mathcal{X}$, and set

$$C'(z) = \begin{cases} C(z) & \text{if } z \neq x, y \\ C(y) & \text{if } z = x \\ C(x) & \text{if } z = y \end{cases}$$

Obviously,

$$\ell_{C'}(z) = \begin{cases} \ell_C(z) & \text{if } z \neq x, y \\ \ell_C(y) & \text{if } z = x \\ \ell_C(x) & \text{if } z = y \end{cases}$$

so that

$$\begin{aligned} L(C; \boldsymbol{p}) - L(C'; \boldsymbol{p}) &= \sum_{z \in \mathcal{X}} \ell_C(z) p(z) - \sum_{z \in \mathcal{X}} \ell_{C'}(z) p(z) \\ &= (\ell_C(x) p(x) + \ell_C(y) p(y)) - (\ell_C(y) p(x) + \ell_C(x) p(y)) \\ &= (\ell_C(x) - \ell_C(y)) \, p(x) + (\ell_C(y) - \ell_C(x)) \, p(y) \\ &= (\ell_C(x) - \ell_C(y)) \, (p(x) - p(y)) \, . \end{aligned}$$

In short, if $p(y) < p(x)$, then $L(C; \boldsymbol{p}) \leq L(C'; \boldsymbol{p})$ if and only if $\ell_C(x) \leq \ell_C(y)$ – In other words, C is preferable to C' if $p(x) \leq p(y)$. Note that C' is a prefix code if C is a prefix code. This is a simple consequence of the Kraft inequality.

Iterating this step leads to the following conclusion: With the symbols in the alphabet $\mathcal{X}$ relabeled so that

$$p(M) \leq p(M-1) \leq \ldots \leq p(2) \leq p(1),$$

any optimal (prefix) code $C : \mathcal{X} \to \mathcal{B}^\star$ necessarily satisfies

$$\ell_C(1) \leq \ell_C(2) \leq \ldots \leq \ell_C(M-1) \leq \ell_C(M).$$

Reduction step behind Huffman encoding

Consider a code $C : \mathcal{X} \to \mathcal{B}^\star$ with the following property: There exist distinct symbols x and y in $\mathcal{X}$ such that their codewords differ only in their last bit, i.e., for some $\ell = 1, 2, \ldots$, we have

$$C(x) = (b_1, \ldots, b_\ell, 1) \quad \text{and} \quad C(y) = (b_1, \ldots, b_\ell, 0)$$

with $b_1, \ldots, b_\ell$ in $\{0, 1\}$.

With the source $X = (\mathcal{X}, \boldsymbol{p})$, we associate a new source $X' = (\mathcal{X}', \boldsymbol{p}')$ as follows: The new alphabet $\mathcal{X}'$ is obtained by combining the two symbols x and y, i.e.,

$$\mathcal{X}' := (\mathcal{X} - \{x, y\}) \cup \{\star\}$$

where $\star$ denotes the new symbol obtained by combining x and y. Next, the pmf $\boldsymbol{p}'$ on $\mathcal{X}'$ is naturally derived from $\boldsymbol{p}$, namely

$$p'(z) = \begin{cases} p(z) & \text{if } z \neq x, y \\ \\ p(x) + p(y) & \text{if } z = \star. \end{cases}$$

With C we now associate a new code $C' : \mathcal{X}' \to \mathcal{B}^\star$ for this new source $X' = (\mathcal{X}', \boldsymbol{p}')$ given by

$$C'(z) = \begin{cases} C(z) & \text{if } z \neq x, y \\ \\ (b_1, \ldots, b_\ell) & \text{if } z = \star. \end{cases}$$

Therefore,

$$\ell_{C'}(z) = \begin{cases} \ell_C(z) & \text{if } z \neq x, y \\ \\ \ell & \text{if } z = \star. \end{cases}$$

With these definitions,

$$\begin{aligned} L(C', \boldsymbol{p}') &= \sum_{z \in \mathcal{X}'} \ell_{C'}(z) p'(z) \\ &= \sum_{z \in \mathcal{X} - \{x,y\}} \ell_{C'}(z) p'(z) + \ell_{C'}(\star) p'(\star) \\ &= \sum_{z \in \mathcal{X} - \{x,y\}} \ell_C(z) p(z) + \ell \left(p(x) + p(y) \right) \\ &= \sum_{z \in \mathcal{X} - \{x,y\}} \ell_C(z) p(z) + \ell p(x) + \ell p(y) \end{aligned}$$

$$= \sum_{z \in \mathcal{X} - \{x,y\}} \ell_C(z) p(z) + (\ell_C(x) - 1)\, p(x) + (\ell_C(y) - 1)\, p(y)$$
$$= \sum_{z \in \mathcal{X}} \ell_C(z) p(z) - (p(x) + p(y))\, .$$

In short,

(13) $$L(C', \boldsymbol{p}') = L(C, \boldsymbol{p}) - (p(x) + p(y))\, .$$

As a consequence, if the optimal prefix code for the new source $X' = (\mathcal{X}', \boldsymbol{p}')$ were known, then the optimal prefix code for the original source $X = (\mathcal{X}, \boldsymbol{p})$ would be easily available.

Properties of optimal prefix codes

For notational convenience, assume that the symbols in the alphabet $\mathcal{X}$ are relabeled so that

$$p(M) \leq p(M-1) \leq \ldots \leq p(2) \leq p(1)$$

with $|\mathcal{X}| = M$.

1. If a (prefix) code $C : \mathcal{X} \to \mathcal{B}^\star$ is optimal, then necessarily

$$\ell_C(1) \leq \ell_C(2) \leq \ldots \leq \ell_C(M-1) \leq \ell_C(M)$$

2. If the prefix code $C : \mathcal{X} \to \mathcal{B}^\star$ is optimal, then necessarily

$$\ell_C(M-1) = \ell_C(M)$$

3. The optimal prefix code $C : \mathcal{X} \to \mathcal{B}^\star$ can always be selected so that $C(M-1)$ and $C(M)$ differ only in the last bit, i.e., if $C(M-1) = (a_1, \ldots, a_\ell)$ and $C(M) = (b_1, \ldots, b_\ell)$ where $\ell = \ell_C(M-1) = \ell_C(M)$, then

$$a_k = b_k, \quad k = 1, \ldots, \ell - 1.$$