1 Preliminaries

$\ln z \le z-1$ for $z>0$, eq. iff $z=1$.

$\alpha \neq 1$:
$$\sum_{j=0}^{k-1} j\alpha^{j-1} = \frac{d}{d\alpha}\sum_{j=0}^{k-1}\alpha^j$$
$$= \frac{d}{d\alpha}\frac{\alpha^k-1}{\alpha-1} = \frac{k\alpha^{k-1}(\alpha-1)-\left(\alpha^k-1\right)}{(\alpha-1)^2}$$

Convex set S: $x,y\in S, \lambda\in[0,1] \Rightarrow \lambda x+(1-\lambda)y\in S$.

Convex f $f: S\to\mathbb{R}$, S convex: $\forall x,y\in S, \lambda\in[0,1]$
$$f(\lambda x+(1-\lambda)y) \le \lambda f(x)+(1-\lambda)f(y)$$
strictly convex if $<$ for $x\neq y$, $\lambda\in(0,1)$.

f is **Concave** iff $-f$ convex.

$f\in C^2$: convex iff $f''\ge 0$; $f''>0 \Rightarrow$ strictly convex.

log strictly concave, $z\log z$ strictly convex on $(0,\infty)$.

Jensen (1) f convex $\Rightarrow f(\mathbb{E}[X]) \le \mathbb{E}[f(X)]$ ($\ge$ if concave). If strictly convex, eq. iff X constant a.s.

2 Source coding

Kleene star $\mathcal{A}^*$: all finite sequences over $\mathcal{A}$, incl. null.

(Bin.) source code $c:\mathcal{U}\to\{0,1\}^*$, we assume $|\mathcal{U}|>1$.

Injective if $u\neq v \Rightarrow c(u)\neq c(v)$.

Code product $(c'\times c'')(u',u'') = c'(u')c''(u'')$
$$c^n = c\times\ldots\times c$$
$$c^*(u^n) = c^n(u^n)\forall n$$

Uniquely decodable (UD) if c^* injective.

Prefix-free (PF) if $u\neq v \Rightarrow c(u)$ is not a prefix of $c(v)$ (instantaneously decodable; codewords are leaves).

PF $\Rightarrow$ UD (converse false, e.g. $\{0,01,11\}$). PF $\subset$ UD $\subset$ Inj.

2.1 KraftSums

(2) $\mathbf{KraftSum}(c) = \sum_u 2^{-\text{len } c(u)}$.

(3) If c injective $\Rightarrow \mathrm{KS}(c)\le\log_2(1+|\mathcal{U}|)$.

$\mathrm{KS}(c'\times c'') = \mathrm{KS}(c')\ \mathrm{KS}(c'')$.

$\mathrm{KS}(c^n) = \mathrm{KS}(c)^n$, which together with (2) and (3) means
$$(4)\ c \text{ UD} \Rightarrow \mathrm{KS}(c)\le 1$$
$\sum_u 2^{-\ell(u)}\le 1 \Rightarrow \exists$ PF c with len $c(u)=\ell(u)$.

By the above and (4), c UD $\Rightarrow \exists$ PF c' with same lengths.

2.2 Expected codeword length

(5) $\mathbb{E}[\text{len } c(U)] \ge H(U) - \log_2 \mathrm{KS}(c)$, with equality iff $\exists d>0$ s.t. $p(u) = d\cdot 2^{-\text{len } c(u)}$. It results that $d=\frac{1}{\mathrm{KS}(c)}$.

c inj. $\Rightarrow \mathbb{E}[\text{len } c(U)] \ge H(U) - \log_2\log_2(1+|\mathcal{U}|)$. This follows from (5) and (3).

(6) c UD $\Rightarrow \mathbb{E}[\text{len } c(U)] \ge H(U)$, eq. iff $p(u)=2^{-\text{len } c(u)}$.

(7) $\exists$ PF c: $\mathbb{E}[\text{len } c(U)] \le H(U)+1$ ($\ell(u)=\left\lceil\log_2\frac{1}{p(u)}\right\rceil$).

$\exists$ PF c_n on $\mathcal{U}^n$: $\frac{1}{n}H(U^n)\le\frac{1}{n}\mathbb{E}[\text{len}]\le\frac{1}{n}H(U^n)+\frac{1}{n}$, by (6) and (7) applied to U^n.

c_n inj. $\Rightarrow \frac{1}{n}\mathbb{E}[\text{len}] \ge \frac{1}{n}H(U^n) - \frac{1}{n}\log_2\log_2(1+|\mathcal{U}|^n)$ ($\sim(\log_2 n)/n$).

2.3 Optimal prefix free code

Problem: find integers $\{\ell(u), u\in\mathcal{U}\}$ with $\sum_u 2^{-\ell(u)}\le 1$ minimizing $\sum_u p(u)\ell(u)$. I.e. find $q(u)\in\{1,\frac{1}{2},\frac{1}{4},\ldots\}$ with $\sum_u q(u)\le 1$ minimizing $\sum_u p(u)\log_2\frac{1}{q(u)}$. Relaxing to $q(u)\in[0,1]$ gives optimum $q=p$ with value $H(U)$.

Let c be an optimal PF code with lengths $\ell(u)$. Then:
1. $p(u)>p(v)\Rightarrow\ell(u)\le\ell(v)$.
2. In the binary tree of c, each node is either a codeword or has exactly 2 children.
3. There are ≥ 2 longest codewords (as $|\mathcal{U}|>1$).

$\exists$ optimal PF code with 2 least likely letters are siblings.

Huffman$(\mathcal{U},p)$ $\quad$ $|\mathcal{U}|=k\ge 2$, returns prefix free c

if $k=2$ **then**
  $c(u_1)\leftarrow 0, c(u_2)\leftarrow 1$
  return c
find permutation π s.t. $p(u_{\pi(1)})\ge\ldots\ge p(u_{\pi(k)})$
$a\leftarrow u_{\pi(k-1)}$,v $b\leftarrow u_{\pi(k)}$, create dummy symbol $w\notin\mathcal{U}$
$\mathcal{U}'\leftarrow\mathcal{U}\setminus\{a,b\}\cup\{w\}$
$$p'(x)\leftarrow\begin{cases}p(x) & \text{if } x\neq w\\ p(a)+p(b) & \text{if } x=w\end{cases}$$
$c'\leftarrow$ Huffman$(\mathcal{U}',p')$
$$c(u)\leftarrow\begin{cases}c'(u) & \text{if } u\notin\{a,b\}\\ c'(w)\,\|\,0 & \text{if } u=a\\ c'(w)\,\|\,1 & \text{if } u=b\end{cases}$$
return c

Huffman is optimal. Optimal codes aren't unique (even lengths may differ).

3 Information measures

3.1 Entropy

$H(U) = \sum_u p(u)\log_2\frac{1}{p(u)}$, where $0\log\frac{1}{0}=0$.

$H(U)\ge 0$, eq. iff U deterministic.

$H(U)\le\log_2|\mathcal{U}|$, eq. iff U uniform.

$H(U,V) = \sum_{u,v} p(u,v)\log_2\frac{1}{p(u,v)}$.

Log-sum inequality $a_i,b_i\ge 0$, $A=\sum a_i$, $B=\sum b_i$:
$$\sum_i a_i\log\frac{a_i}{b_i}\ge A\log\frac{A}{B}$$
eq. iff $a_i/A=b_i/B\ \forall i$. ($0\log\frac{0}{B}=0$ if $B\neq 0$, $A\log\frac{A}{0}=+\infty$ if $A\neq 0$, by continuity).

(8) $H(U,V)\le H(U)+H(V)$, eq. iff $U\perp V$.

$H(V|U) = \sum_{u,v}p(u,v)\log_2\frac{1}{p(v|u)} =$
$\sum_u p(u)\sum_v p(v|u)\log_2\frac{1}{p(v|u)} = \sum_u p(u)H(V|U=u)$.

Chain rule (9) $H(U,V) = H(U)+H(V|U)$
$$H(U^n) = \sum_i H(U_i\mid U^{i-1})$$

$H(V|U)\le H(V)$, eq. iff $U\perp V$; follows from (9) and (8).

$I(U;V) = H(U)+H(V)-H(U,V)\ge 0$, with equality iff $U\perp V$ (from (8)); symmetric quantity.

$I(U;V) = H(V)-H(V|U) = H(U)-H(U|V)$.

$I(U;V|W) = H(U|W)-H(U|V,W)$
$$= H(V|W)-H(V|U,W).$$

$I(U;V|W)\ge 0$, eq. iff $U\perp V\mid W$.

From the above follows that $H(U|W)\ge H(U|W,V)$.

From (9) follows $I(U^n;V|W) = \sum_i I(U_i;V|U^{i-1},W)$.

Fact: conditioning can *increase* MI: $X\oplus Y\oplus Z=0$ uniform $\Rightarrow I(X;Y)=0<I(X;Y|Z)=1$.

Data processing inequality $X-Y-Z$ Markov $(p(z|x,y)=p(z|y))\Rightarrow I(X;Z)\le I(X;Y)$. Pf: $I(X;Y,Z)=I(X;Z)+I(X;Y|Z)=I(X;Y)+\underbrace{I(X;Z|Y)}_{0}$, and $I(X;Y|Z)\ge 0$.

Binary entropy $h_2(\alpha)=\alpha\log_2\frac{1}{\alpha}+(1-\alpha)\log_2\frac{1}{1-\alpha}$; symmetric around $1/2$, $h_2(0)=h_2(1)=0$, $h_2(1/2)=1$, $h_2(0.11)\approx 1/2$, concave.

Let $\alpha \in [0,1]$, then
$H(p_1, ..., p_{k-1}, p_k\alpha, p_k(1-\alpha)) = H(p_1, ..., p_k) + p_k h_2(\alpha)$
Simplex $\mathcal{S}_k = \left\{(x_1, ..., x_k) \in \mathbb{R}^k : \sum_i x_i = 1, x_i \geq 0 \ \forall i\right\}$; convex. With $|\mathcal{U}| = k$, $p_U \in \mathcal{S}_k$.

H is concave on the simplex $\mathcal{S}_k$, i.e. $\forall p, q \in \mathcal{S}_k$, $\lambda \in [0,1]$:
$$\lambda H(p) + (1-\lambda)H(q) \leq H(\lambda p + (1-\lambda)q)$$
Let $g\left(p_U, p_{V|U}\right) = I(U;V)$, $(U,V) \sim p_{UV}$. $\forall \lambda \in [0,1]$:
- fixed $p_{V|U}$, $g\left(\cdot, p_{V|U}\right)$ is concave
- fixed p_U, $g(p_U, \cdot)$ is convex

Entropy rate $\mathcal{H} = \lim_n \frac{1}{n} H(U^n)$ (if exists).
i.i.d. $\mathcal{H} = H(U_1)$.
Markov $p(u_n \mid u^{n-1}) = p(u_n \mid u_{n-1})$; $U_1 - U_2 - ....$
Stationary $(U_{i+1}, ..., U_{i+n}) \sim U^n \ \forall i, n$.
(10) **Cesàro mean** $\alpha_n \to \alpha \Rightarrow \frac{1}{n}\sum_{i \leq n} \alpha_i \to \alpha$.
Stationary $\Rightarrow \mathcal{H} = \lim \frac{1}{n} H(U^n) = \lim H(U_n \mid U^{n-1})$ (both exist; latter decreasing). Can be proven with (10).
KL $D(q\|p) = \sum_u q(u) \log_2 \frac{q(u)}{p(u)}$; not symmetric.
$I(U;V) = D(p_{UV} \parallel p_U p_V)$.
f-div. $f : \mathbb{R}_{\geq 0} \to \mathbb{R} \cup \{+\infty\}$ convex, $f(1) = 0$, strictly convex at 1: $D_f(q\|p) = \sum_u p(u) f\left(\frac{q(u)}{p(u)}\right)$. TV: $f = |z-1|$; χ^2: $f = z^2 - 1$.
KL is the f-div. with $f(z) = z \log_2 z$.
For any f-divergence $D_f(q\|p) \geq 0, = 0$ iff $p = q$.

4 Universal compression

Suppose optimal code for q, but true probability is p:
$\mathbb{E}[\text{len}] = \sum_u p(u) \log_2 \frac{1}{q(u)} = H(p) + D(p\|q)$.
Minimax redundancy $\mathcal{R}^*(\mathcal{P}) = \min_q \max_{p \in \mathcal{P}} D(p\|q)$.
Pointwise regret $\mathcal{R}^\star(\mathcal{P}) = \min_q \max_{p \in \mathcal{P}} \max_u \log_2 \frac{p(u)}{q(u)}$.
Clearly $\mathcal{R}^\star \geq \mathcal{R}^*$.
$\textbf{ShtarkovSum}(\mathcal{P}) = \sum_u \max_{p \in \mathcal{P}} p(u)$
$\mathcal{R}^\star(\mathcal{P}) = \log_2 \text{SS}(\mathcal{P}), \quad q^*(u) = \max_p p(u)/\text{SS}(\mathcal{P})$
Let $\mathcal{U} = \{0,1\}^n$, $\mathcal{P}_n = \left\{p_\theta \mid p_\theta(u^n) = \theta^k(1-\theta)^{n-k}\right\}$ the class of *iid* Bernoulli on $\{0,1\}^n$, where $k = \|u^n\|_1$. Then:
$$\mathcal{R}^\star(\mathcal{P}_n) \leq \log_2(n+1)$$
Designing a code under q^* in this scenario yields:
$$\text{len } c(u^n) - \log_2 \frac{1}{p_\theta}(u^n) \leq ... \leq 1 + \log_2(n+1)$$
It can be shown that $\binom{n}{k} \leq 2^{n h_2(k/n)}$.

In this scenario, we can define a coding scheme that: sends k, then index among $\binom{n}{k}$: $\frac{1}{n}$ len $c(u^n) \leq$
$$\lceil \log_2(1+n) \rceil + \left\lceil \log_2 \binom{n}{k} \right\rceil \leq h_2\left(\frac{k}{n}\right) + \frac{2 + \log_2(n+1)}{n}$$

4.1 FSM compressors and Lempel-Ziv

Lempel-Ziv$(\mathcal{U}, u = u_1 u_2 \cdots)$ assumes $u_i \in \mathcal{U}$

$i \leftarrow 0, \mathcal{D} \leftarrow \mathcal{U}$
loop
  assign $\lceil \log_2 |\mathcal{D}| \rceil$ bit representation to each $w \in \mathcal{D}$
  find $w \in \mathcal{D}$ which is the longest prefix of $u_{i+1} u_{i+2} \cdots$
  send the binary representation of w
  $\mathcal{D} \leftarrow \mathcal{D} \setminus \{w\} \cup \{w\alpha \mid \alpha \in \mathcal{U}\}$
  $i \leftarrow i + \text{length}(w)$

A FSM compressor has finite states $\mathcal{S}$, starting state s_0. If we are in state $s \in \mathcal{S}$ and read letter $u \in \mathcal{U}$, next-state function $\texttt{next_state} : \mathcal{S} \times \mathcal{U} \to \mathcal{S}$, and output function $\texttt{output} : \mathcal{S} \times \mathcal{U} \to \{0,1\}^*$: $s_{i+1} = \texttt{next_state}(s_i, u_{i+1})$, $y_{i+1} = \texttt{output}(s_i, u_{i+1})$. For $v^n \in \mathcal{U}^*$:
$\texttt{final_state}(s, v^n) \in \mathcal{S}$ state after reading v^n from s;
$\texttt{output}(s, v^n) \in \{0,1\}^*$ concatenation of outputs while reading $v_1, ..., v_n$. WLOG all states reachable from s_0 (via $\texttt{final_state}(s_0, \cdot)$).
Information lossless (IL) $\forall u^n \neq v^m \in \mathcal{U}^*$, at least one:
- $\texttt{output}(s_0, u^n) \neq \texttt{output}(s_0, v^m)$;
- $\texttt{final_state}(s_0, u^n) \neq \texttt{final_state}(s_0, v^m)$.

All $s \in \mathcal{S}$ reachable $\Rightarrow$ (IL $\Leftrightarrow$ the above holds $\forall s \in \mathcal{S}$).
Distinct parsing $u = w_1 ... w_m$, w_i pairwise distinct.
$m^*(u^n)$ = max m of a distinct parsing.
$m^*(u^n) \leq n+1$ (null + single letters)
$m^*(u^n) = o(n)$
$y_1, ..., y_m$ distinct binary: $\sum_i \text{len}(y_i) \geq m \log_2\left(\frac{m}{8}\right)$.
(11) IL FSM with s states: $\ell_s(u^n) \geq m^* \log_2 \frac{m^*}{8s^2}$.
(12) $\ell_{\text{LZ}}(u^n) \leq m^* \log_2(2m^* \ |\mathcal{U}|)$.
(13) $\frac{1}{n}(\ell_{\text{LZ}} - \ell_s) \leq \frac{m^*}{n} \log_2(16 \ |\mathcal{U}| \ s^2) = o(1)$: LZ asymptotically matches IL FSM; follows from (11) and (12).
For a stationary process: $\lim \frac{1}{n}\mathbb{E}[\ell_{\text{LZ}}(U^n)] = \mathcal{H}$ (LZ is universally optimal asymptotically).

Compressibility of u^∞ under $c : \mathcal{U}^* \to \{0,1\}^*$: $\rho_c(u^\infty) = \lim_n \frac{1}{n}$ len $c(u^n)$ (limsup if no limit).
FSM-IL compressibility
$$\rho_{\text{FSM-IL}}(u^\infty) = \inf_{M \text{ IL FSM}} \rho_{c_M}(u^\infty)$$
where $c_M(u^n) = \texttt{output}(s_0, u^n)$.
$\rho_{\text{LZ}} \leq \rho_{\text{FSM-IL}}$; follows from (13).