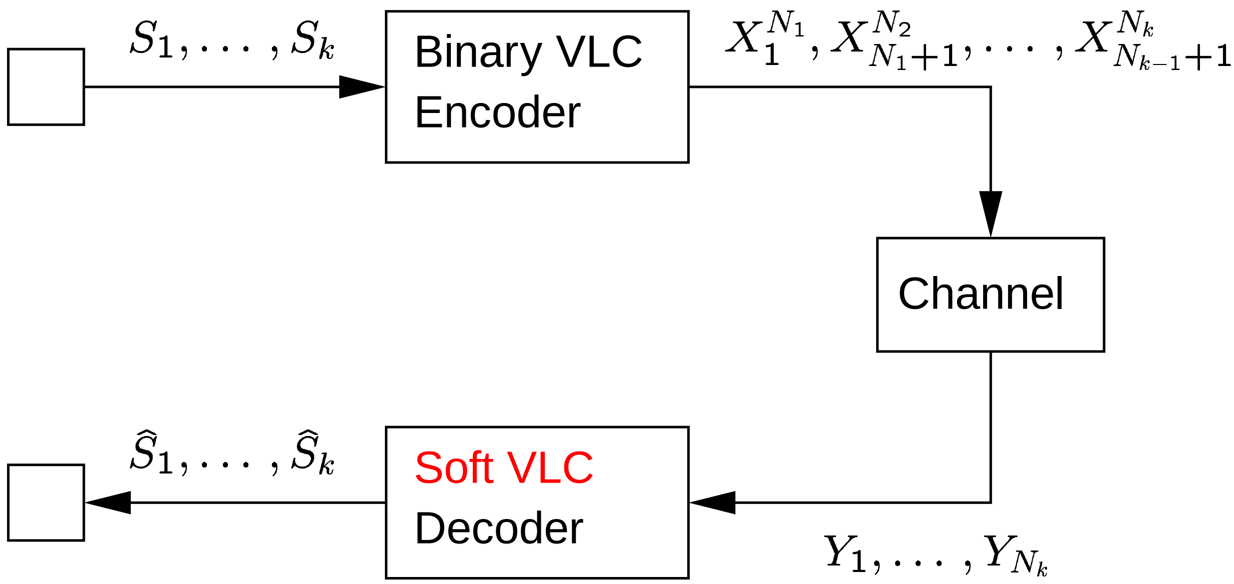
Variable-Length Codes getting Framed

Claudio Weidmann
IRISA-INRIA
Rennes, France

2nd Asian-European Workshop on Information Theory
Breisach, Germany, June 26-29, 2002

Setup



- The decoder knows N_k because the sequence is framed

- The set of all sequences of length $N_k = n$ that are encodings of k symbols forms a nonlinear block code with $d_{min}(k, n)$

- Free distance:

$$d_{free} = \min\{d_{min}(k, n) : k = 1, 2, \ldots ; n \in \mathcal{N}_k\}$$

Variable-length codes

- Length distribution

$$(m_1, m_2, \ldots, m_{\ell_{max}})$$

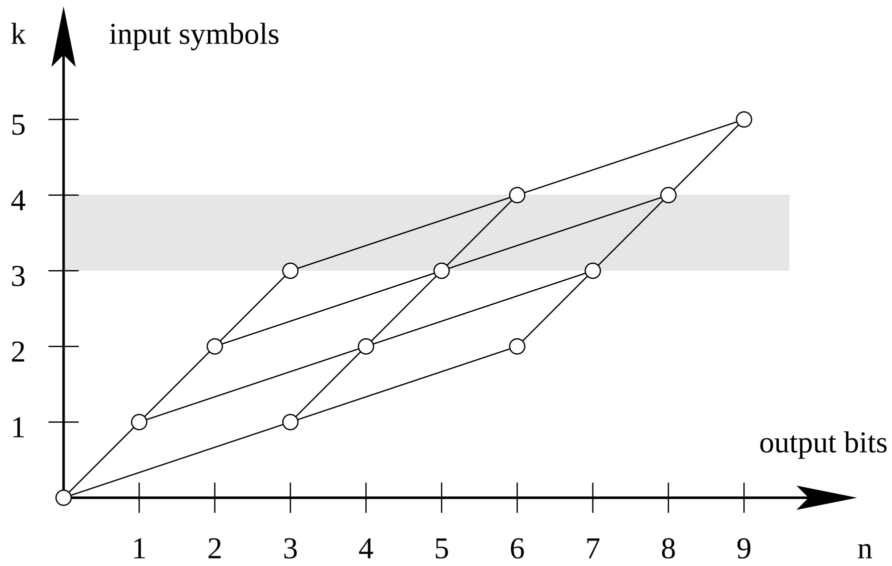
- Uniquely decodable:

$$\sum_{\ell=1}^{\ell_{max}} m_\ell 2^{-\ell} \leq 1$$

- Compact binary VLCs have Kraft sum 1

- Compact prefix-free codes have $d_{free} = 1$

Segmentation trellis



Counting sequences to bound d_{free}

$M(k, n)$ is the number of source sequences of length k that are encoded into n bits.

- Singleton bound:

$$d_{min} \leq n + 1 - \log_2 M.$$

- Plotkin bound:

$$d_{min} \le \frac{M}{M-1} \cdot \frac{n}{2}.$$

Proof: Sum of all distances:

$$S = \sum_{x \in \mathcal{C}} \sum_{y \in \mathcal{C}} d(x,y) \ge M(M-1)d_{min}. \quad (1)$$

Express S as sum over bit positions (M_i is the number of codewords with a 1 in position i):

$$S = \sum_{i=1}^{n} 2M_i(M - M_i) \le n\frac{M^2}{2}. \quad (2)$$

$\square$

- Tightened Plotkin bound (optimize over $j = 0, 1, \ldots, n-1$):

$$d_{min} \le \frac{M}{M - 2^j} \cdot \frac{n-j}{2}.$$

Proof: Divide $\mathcal{C}$ into 2^j classes of codewords which are identical in the first j positions. Let M_j be the size of the respective class. Then $d_{min} \le \frac{M_j}{M_j - 1} \cdot \frac{n-j}{2}$ and the bound follows by observing that $\max\{M_j\} \ge M/2^j$. $\square$

Examples

1. A compact code with length distribution (1,1,1,2) has $d_{free} \leq 2$, since $M(3, 6) = 13$ ($6 = 1 + 1 + 4 = 1 + 2 + 3 = 2 + 2 + 2$).

2. A compact code with length distribution (0,1,1,3,6,12,8) has $d_{free} \leq 3$, due to e.g. $m_4 = 3$. If we further consider $A(n, d)$ for e.g. $n = 5$, we see that $d_{free} \leq 2$ (Plotkin bound too weak in this case).

3. The reversible (a.k.a. fix-free) VLC

$$\mathcal{C}_{12} = \{00, 11, 010, 101, 0110\}$$

   (Takishima, Wada, Murakami 1995) with Kraft sum 0.8125 has $d_{free} \leq 2$, due to $m_2 = 2$.
   Proposition: $d_{free} = 2$.
   Sketch of proof: $d_{min}(1, \ell) \geq 2$ and "prefix, resp. suffix distances" ≥ 1.

   All fix-free VLCs have $d_{free} \geq 2$ if the distance between equal-length codewords is ≥ 2.

4. The binary comma code contains the words $0^{\ell-1}1$ for $\ell = 1, \ldots, \ell_{max}$ (Kraft sum $1 - 2^{-\ell_{max}}$) and has $d_{free} = 2$. The last bit in a frame may be deleted.

5. Compact two-length codes have $d_{free} = 1$.

$$m_1 2^{-\ell_1} + m_2 2^{-\ell_2} = 1$$

Lemma 1 *Let $\mathcal{C}_n$ be a binary $(n, 2^{n-1}, 2)$ code. Then $\mathcal{C}_n$ consists of either all the even weight words or all the odd weight words of length n.*

Proof: Suppose $\mathcal{C}_n$ contains both even and odd weight words. Then the "coset" $\overline{\mathcal{C}}_n = \mathcal{C}_n + 0^{n-1}1$ is also a $(n, 2^{n-1}, 2)$ code and must be equal to $GF(2)^n \setminus \mathcal{C}_n$. But the minimum distance between even and odd weight words in $\mathcal{C}_n$ is at least three, therefore there exists a pair of words (x, y) with $d(x, y) = 1$ that is neither in $\mathcal{C}_n$ nor in $\overline{\mathcal{C}}_n$, which is a contradiction. $\square$

Proposition 1 *Any uniquely decodable VLC with multiplicity* $m_\ell = 2^{\ell-1}$ *for some* $\ell > \ell_{min}$ *has* $d_{free} = 1$*.*

Proof: The Singleton bound implies $d_{free} \leq 2$. Let $\mathcal{C}_\ell$ be the subset of codewords of length ℓ. To have $d_{free} = 2$, Lemma 1 requires $\mathcal{C}_\ell = \mathcal{E}_\ell$ (the even weight words of length ℓ) or $\mathcal{C}_\ell = \mathcal{O}_\ell$ (odd weight). Then for any $x \in \mathcal{C}_{\ell_{min}}$ we can always find a $y \in \{0, 1\}^{\ell-\ell_{min}}$ such that $xy \in \mathcal{C}_\ell$. But this implies that also $yx \in \mathcal{C}_\ell$ and therefore the sequence xyx is not uniquely decodable. $\square$

Asymptotia: Coding Extension Sources

- Binary memoryless source with $\Pr\{S=1\} = p$

- Ideal codeword length for k-th extension (k_1 ones):

$$n = -k_1 \log p - (k - k_1) \log(1 - p)$$

- Decoder knows type $Q = (\frac{k_1}{k}, 1 - \frac{k_1}{k})$ of source sequence if $p \leq \frac{1}{3}$ or $p \geq \frac{2}{3}$

- Size of type class:

$$\log |T_Q^k| = \log \binom{k}{k_1} \approx kH(Q) - \frac{1}{2} \log \frac{2\pi(k - k_1)k_1}{k}$$

- For $k_1 = \lfloor pk \rfloor \gg 1$:

$$n - \log |T_Q^k| \approx kD(Q||P) + \frac{1}{2} \log(2\pi kp(1-p))$$

- Singleton bound:

$$d_{min}(k) \lesssim 1 + \frac{1}{2} \log(2\pi p(1 - p)) + \frac{1}{2} \log k$$

Conclusions

- d_{free} of most practical VLCs is limited by the multiplicities of short codewords.

- Challenge: find a compact VLC with $d_{free} = 2$ or prove that none exists.

- Criteria such as "speed" of resynchronization might be more important in applications.