# Introducción a la Teoría de la Información

## Codificación de fuentes

Facultad de Ingeniería, UdelaR

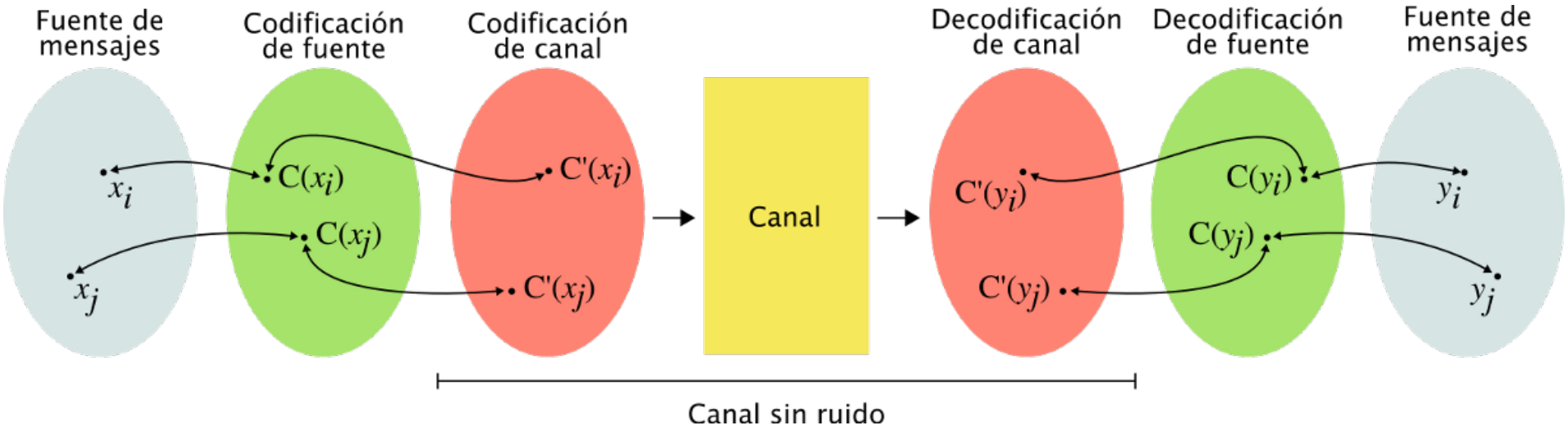
# Agenda

1. Codificación de fuente
   - Definiciones
   - Clasificación de códigos
   - Desigualdad de Kraft
   - Cotas para el largo medio de código

2. Esquemas de codificación
   - Códigos de Huffman
   - Optimalidad competitiva del código de Shannon
   - Códigos de Shannon-Fano-Elias
   - Codificación aritmética

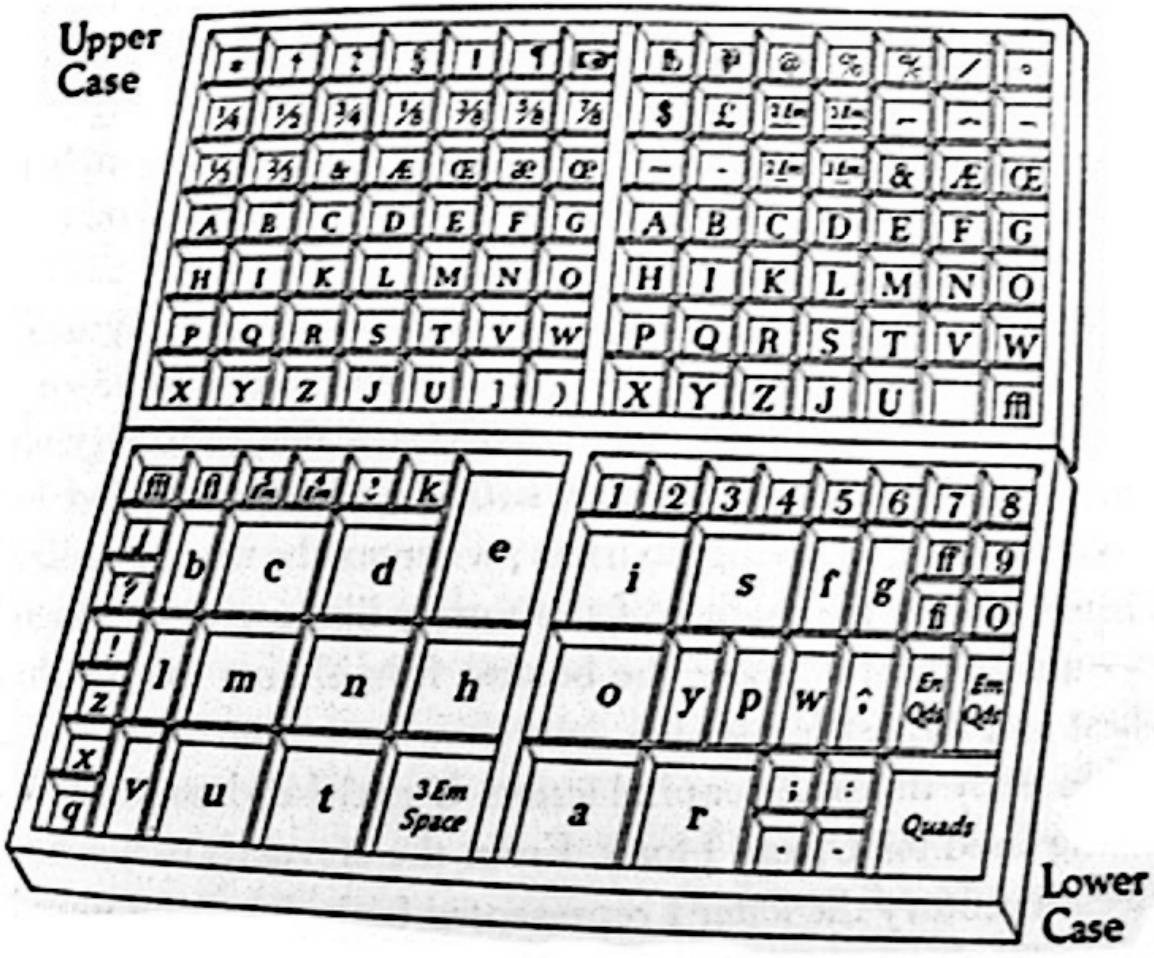
# Codificación de fuente

Fuente generadora de mensajes de un alfabeto fuente $\mathcal{X} = \{x_1, \ldots, x_m\}$ con probabilidades $p_X(x_i)$. A cada uno de los mensajes se le asignará una palabra de código $C(x_i)$.




¿Cómo asignar las palabras de códigos de forma *óptima* y sistemática?

# Código Morse



# Codificación de fuente

## Definición (Código fuente)

Un *código de fuente* $C$ para una variable aleatoria $X$ es una función,

$$C : \mathcal{X} \to \mathcal{D}^* ,$$

donde $\mathcal{D} = \{0, 1, \ldots, D-1\}$ es un conjunto finito, denominado *alfabeto de código ($D$-ario)*, y $\mathcal{D}^*$ es el conjunto de palabras de largo finito sobre $\mathcal{D}$.

- A cada mensaje $x_i$ se asigna la palabra de código $C(x_i)$ de largo $l(x_i) = l_i$.
- El alfabeto de código suele ser binario, que denotamos $\mathcal{B}$.

## Ejemplo

Para $X$ definida sobre $\mathcal{X} = \{x_1, x_2, x_3\}$, $C(x_1) = 0$, $C(x_2) = 10$, $C(x_3) = 11$ es un código de fuente binario.

# Codificación de fuente

## Definición (Largo medio de un código)

El *largo medio* de código, $L(C)$, para una variable aleatoria $X$ con distribución de probabilidad $p(x)$ se define como

$$L(C) = \sum_{x \in \mathcal{X}} p(x) l(x) \,,$$

donde $l(x)$ es el largo de la palabra de código asignada a $x$.

## Ejemplo

- $p_X(\mathcal{X}) = \left\{\frac{1}{2}, \frac{1}{4}, \frac{1}{8}, \frac{1}{8}\right\}$ con $C(X) = \{0, 10, 110, 111\}$, $L(C) = H(X) = 1{,}75$ bits
- $p_X(X) = \left\{\frac{1}{3}, \frac{1}{3}, \frac{1}{3}\right\}$ con $C(X) = \{0, 10, 11\}$, $L(C) = 1{,}66 > H(X) = 1{,}58$ bits

# Clasificación de códigos

## Definición (Código no singular)

Un código es *no singular* si cada elemento de $\mathcal{X}$ se mapea a una palabra de código diferente de $\mathcal{D}^*$

$$x_i \neq x_j \Rightarrow C(x_i) \neq C(x_j)$$

## Definición (Extensión de un código)

La *extensión* $C^*$ de un código $C$ es un mapeo de una secuencia de símbolos de $\mathcal{X}$ en un secuencia de $\mathcal{D}$ definida por

$$C(x_1 x_2 \dots x_n) = C(x_1)C(x_2)\dots C(x_n),$$

donde $C(x_1)C(x_2)\dots C(x_n)$ es la concatenación de las palabras de código $C(x_1), C(x_2), \dots, C(x_n)$.

# Clasificación de códigos

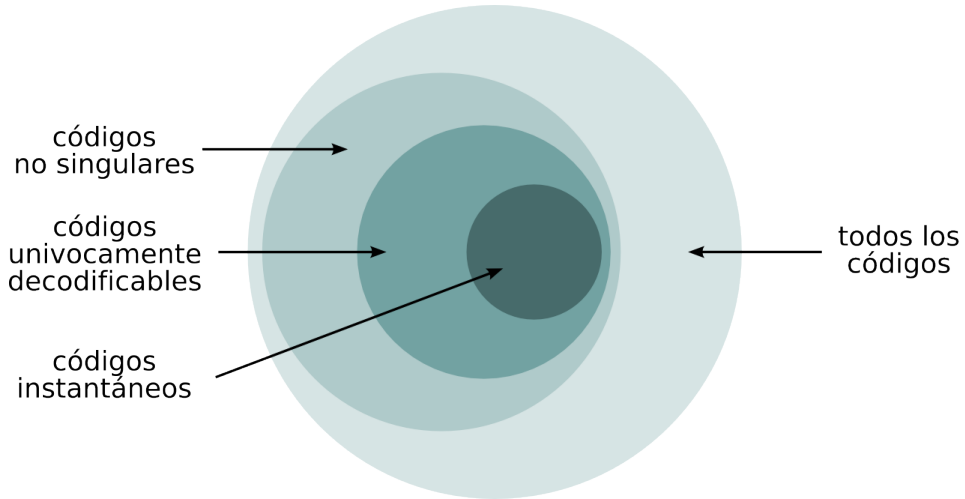
## Definición (Código unívocamente decodificable)

Un código es *unívocamente decodificable* si su extensión es no singular.

O sea, no hay ambigüedades al momento de decodificar una secuencia codificada.

## Definición (Código instantáneo)

Un código es *instantáneo* o *de prefijo* si ninguna palabra de código es prefijo de otra palabra de código.



# Clasificación de códigos

## Ejemplo

| $\mathcal{X}$ | $C_1$ | $C_2$ | $C_3$ | $C_4$ | $C_5$ |
|---|---|---|---|---|---|
| $x_1$ | 0 | 0 | 10 | 0 | 0 |
| $x_2$ | 0 | 010 | 00 | 10 | 10 |
| $x_3$ | 0 | 01 | 11 | 110 | 110 |
| $x_4$ | 0 | 10 | 110 | 1110 | 111 |

- $C_1$ es singular, no sirve para mucho
- $C_2$ es no singular pero no es unívocamente decodificable (UD). La secuencia 010 puede decodificarse como $x_2$, $x_1x_4$ o $x_3x_1$.
- $C_3$ es UD aunque no es instantáneo. Si se recibe $110\ldots$, decodificamos $x_3$ o $x_4$ dependiendo de la paridad de la cantidad de ceros que siguen a 11.
- $C_4$ es instantáneo (es un código de puntuación, el 0 marca el final de cada palabra) *¿es eficiente*?
- $C_5$ es instantáneo.

# Desigualdad de Kraft

## Teorema (Desigualdad de Kraft)

*Para todo código instantáneo sobre un alfabeto de tamaño $D$ y largos de palabra $l(x_1), l(x_2), \ldots, l(x_m)$ se cumple*

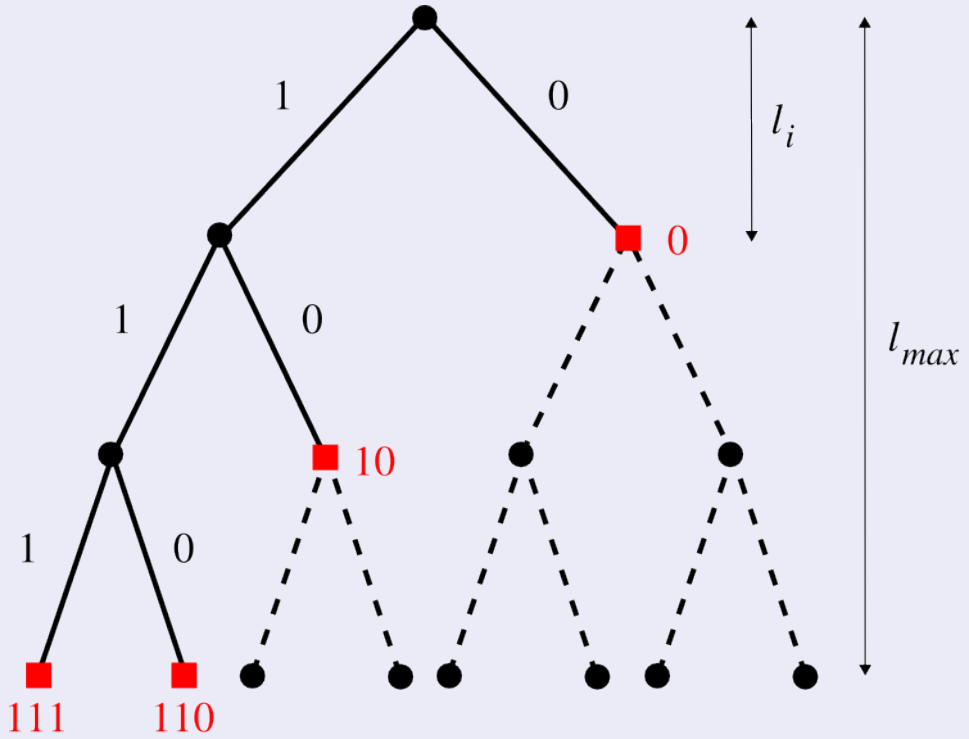
$$\sum_{x \in \mathcal{X}} D^{-l(x)} \leq 1 \,.$$

*Recíprocamente, dado un conjunto de largos de palabra de código que satisfacen esta desigualdad, existe un código instantáneo con esos largos.*

- Las longitudes de las palabras no pueden ser todas "cortas", si hay una muy corta debe haber otras más largas.

# Desigualdad de Kraft

## Demostración




$$\sum_i D^{l_{\text{máx}}-l_i} \leq D^{l_{\text{máx}}}$$

# Desigualdad de Kraft extendida

## Teorema (Desigualdad de Kraft extendida)

*Los largos de palabra de un código $D$-ario instantáneo definido para un alfabeto numerable satisfacen*

$$\sum_{i=1}^{+\infty} D^{-l_i} \leq 1.$$

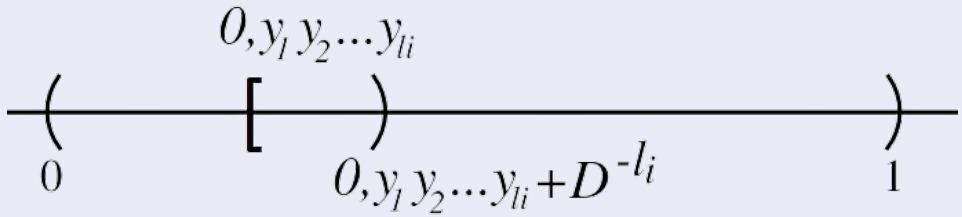
*Recíprocamente, dado un conjunto numerable de largos de palabra de código que satisfacen esta desigualdad, existe un código instantáneo con esos largos.*

# Desigualdad de Kraft extendida

## Demostración

Sea $y_1 y_2 \dots y_{l_i}$ la $i$-ésima palabra, y sea $0.y_1 y_2 \dots y_{l_i} = \sum_{j=1}^{l_i} y_j D^{-j}$ el número real que representa en base $D$.

La representación D-aria de todos los reales en $\left[0.y_1 y_2 \dots y_{l_i}, 0.y_1 y_2 \dots y_{l_i} + D^{-l_i}\right)$ empieza con $0.y_1 y_2 \dots y_{l_i}$.




Todos estos intervalos son disjuntos por la condición de prefijo.

Por lo tanto, la suma de los anchos de estos intervalos debe ser menor o igual que la del intervalo $[0, 1]$ que los cubre,

$$\sum_{i=1}^{+\infty} D^{-l_i} \leq 1 \,.$$

# Desigualdad de Kraft extendida (recíproco)

- Ordenamos los largos de menor a mayor, $l_1 \leq l_2 \leq \ldots$.
- Para $m \geq 1$, definimos la palabra de código $c_m$ como los $l_m$ dígitos a la derecha de la coma en la representación $D$-aria del número

$$f_m = \sum_{i=1}^{m-1} D^{-l_i} .$$

- Sean $m$ y $m'$ arbitrarios, con $m < m'$. Tenemos

$$f_{m'} \geq \sum_{i=1}^{m} D^{-l_i} = f_m + D^{-l_m} .$$

- La representación $D$-aria de $f_{m'}$ difiere de la de $f_m$ en alguno de los primeros $l_m$ dígitos a la derecha de la coma.
- Por lo tanto $c_m$ y $c_{m'}$ no pueden ser una prefija de la otra.

# Desigualdad de Kraft para códigos UD I

La clase de los codigos UD es más grande que la de los instantáneos, sin embargo no presentan ninguna ventaja respecto a la longitud de las palabras de código.

## Teorema (McMillan)

*Los largos de palabra de código $l_i$ de un código UD cumplen la desigualdad de Kraft,*

$$\sum D^{-l_i} \leq 1 \,.$$

# Desigualdad de Kraft para códigos UD II

## Demostración

Asumimos primero que el código es finito y, para un natural $k$, escribimos

$$\begin{aligned}
\left(\sum_{x\in\mathcal{X}} D^{-l(x)}\right)^k &= \sum_{x_1\in\mathcal{X}}\sum_{x_2\in\mathcal{X}}\cdots\sum_{x_k\in\mathcal{X}} D^{-l(x_1)}D^{-l(x_2)}\ldots D^{-l(x_k)} \\
&= \sum_{x_1,x_2,\ldots,x_k\in\mathcal{X}^k} D^{-l(x_1)}D^{-l(x_2)}\ldots D^{-l(x_k)} \\
&= \sum_{x^k\in\mathcal{X}^k} D^{-l(x^k)}, \quad \text{donde } l(x^k)\triangleq\sum_{i=1}^{k} l(x_i) \\
&= \sum_{m=1}^{kl_{\text{máx}}} a(m)D^{-m},
\end{aligned}$$

donde $a(m)$ es la cantidad de secuencias $x^k$ con $l(x^k)=m$.

# Desigualdad de Kraft para códigos UD III

## Demostración

Como $C$ es UD, su extensión es no singular, lo cual implica que $a(m) \leq D^m$. Por lo tanto,

$$\left( \sum_{x \in \mathcal{X}} D^{-l(x)} \right)^k = \sum_{m=1}^{k l_{\text{máx}}} a(m) D^{-m} \leq \sum_{m=1}^{k l_{\text{máx}}} D^m D^{-m} = k l_{\text{máx}} \, .$$

Entonces

$$\sum_{x \in \mathcal{X}} D^{-l(x)} \leq (k l_{\text{máx}})^{\frac{1}{k}} \xrightarrow[k \to \infty]{} 1 \, .$$

Finalmente, si el código no es finito, cualquier subconjunto finito de él es UD. Por lo tanto,

$$\sum_{x \in \mathcal{X}} D^{-l(x)} = \sum_{i=1}^{\infty} D^{-l_i} = \lim_{N \to \infty} \sum_{i=1}^{N} D^{-l_i} \leq 1 \, .$$

□

# Desigualdad de Kraft, largo medio de código y entropía

## Ejemplo

Con los códigos analizados, y la fuente $p_X(\mathcal{X}) = \left\{\frac{1}{2}, \frac{1}{4}, \frac{1}{8}, \frac{1}{8}\right\}$ de $H(X) = 1{,}75$bits

| $\mathcal{X}$ | $C_1$ | $C_2$ | $C_3$ | $C_4$ | $C_5$ |
|---|---|---|---|---|---|
| $x_1$ | 0 | 0 | 10 | 0 | 0 |
| $x_2$ | 0 | 010 | 00 | 10 | 10 |
| $x_3$ | 0 | 01 | 11 | 110 | 110 |
| $x_4$ | 0 | 10 | 110 | 1110 | 111 |
| K | 2 | 1.125 | 0.875 | 0.9375 | 1 |
| L | 1 | 1.75 | 2.25 | 1.875 | 1.75 |
| H/L | 1.75 | 1 | 0.778 | 0.933 | 1 |

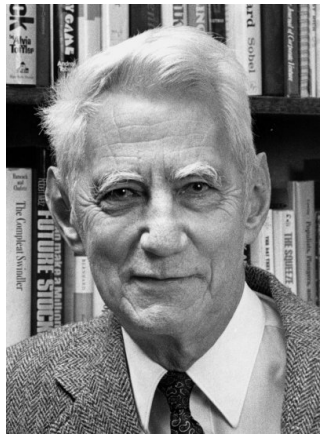
# Primer Teorema de Shannon

## Teorema (Recíproco del Teorema de Codificación de Fuente)

*El largo medio de un código $C$ instantáneo, $D$-ario para una variable aleatoria $X$ es mayor o igual a la entropía de $X$*

$$L(C) \geq H_D(X)$$

*y la igualdad se cumple sii $p_i = D^{-l_i}$.*



- Es una cota para la longitud media para la descripción de una fuente, "no se puede comprimir más allá de la entropía".
- Para que se pueda alcanzar la igualdad:
  - la distribución debe ser $D$-ádica,
  - los largos deben satisfacer $l_i = -\log_D p_i$.

# Teorema de Codificación de Fuente

## Demostración

Sea $K = \sum_i D^{-l_i}$ y $q_i = D^{-l_i}/K$ una distribución de probabilidad sobre $\mathcal{X}$.

$$\begin{aligned} D(p||q) &= \sum_i p_i \log_D \frac{p_i}{q_i} = \sum_i p_i \log_D p_i - \sum_i p_i \log_D q_i \\ &= -H_D(X) - \sum_i p_i \log_D \frac{D^{-l_i}}{K} \\ &= -H_D(X) - \sum_i p_i \log_D D^{-l_i} + \sum_i p_i \log_D K \\ &= -H_D(X) + \sum_i p_i l_i + \log_D K \\ &= -H_D(X) + L(C) + \log_D K\,. \end{aligned}$$

Por lo tanto se cumple

$$L(C) - H_D(X) = D(p||q) - \log_D K \geq 0\,,$$

con igualdad sii $D(p||q) = 0$ y $K = 1$. □

# Cotas para el largo de código medio óptimo

El largo de código "ideal" para $x_i$, $l_i^* \triangleq -\log_D p_i$, puede no ser entero. Sin embargo, tomando

$$l_i = \lceil -\log_D p_i \rceil \geq l_i^* \,,$$

vemos que se cumple la desigualdad de Kraft

$$\sum_i D^{-\lceil \log_D p_i \rceil} \leq \sum_i D^{\log_D p_i} = \sum_i p_i = 1 \,.$$

- Se llama *código de Shannon* a un código con estos largos.
- Como $l_i < -\log_D p_i + 1$, el código de Shannon satisface

$$L < H_D(X) + 1 \,.$$

- Por lo tanto, el largo de código medio óptimo, $L^*$, cumple

$$H_D(X) \leq L^* < H_D(X) + 1 \,.$$

# Cotas para el largo de código medio óptimo

Si consideremos una secuencia $x^n = (x_1, x_2, \ldots, x_n)$ como un símbolo individual de la fuente extendida $\mathcal{X}^n$, el largo de código medio óptimo para esta fuente extendida satisface

$$H(X_1, X_2, \ldots, X_n) \leq E\left[l(X_1, X_2, \ldots, X_n)\right] < H(X_1, X_2, \ldots, X_n) + 1\,.$$

Si los símbolos $X_1, X_2, \ldots, X_n$ son i.i.d., $H(X_1, X_2, \ldots, X_n) = nH(X)$, de modo que el largo de código medio *por símbolo*,

$$L_n \triangleq \frac{1}{n} E\left[l(X_1, X_2, \ldots, X_n)\right]\,,$$

satisface

$$H(X) \leq L_n < H(X) + \frac{1}{n}\,.$$

# Cotas para el largo de código medio óptimo

Si la secuencia proviene de un proceso estocástico estacionario,

$$H(X_1, X_2, \ldots, X_n) \leq E\left[l(X_1, X_2, \ldots, X_n)\right] < H(X_1, X_2, \ldots, X_n) + 1$$

$$\frac{H(X_1, X_2, \ldots, X_n)}{n} \leq L_n < \frac{H(X_1, X_2, \ldots, X_n)}{n} + \frac{1}{n}$$

recordando que $\frac{H(X_1, X_2, \ldots, X_n)}{n} \xrightarrow[n\to\infty]{} H(\mathcal{X})$ es la tasa de entropía del proceso $\{X_i\}_{i\geq 1}$,

## Teorema

*El largo medio mínimo por símbolo para un proceso estocástico cumple*

$$\frac{H(X_1, X_2, \ldots, X_n)}{n} \leq L_n^* < \frac{H(X_1, X_2, \ldots, X_n)}{n} + \frac{1}{n}$$

*y si el proceso es estacionario con tasa de entropía $H(\mathcal{X})$,*

$$L_n^* \longrightarrow H(\mathcal{X})\,.$$

# La distribución incorrecta

¿Qué pasa si utilizamos una distribución $q(x) \neq p(x)$?

## Teorema

*Sea $q(x)$ una distribución de probabilidad sobre $\mathcal{X}$ usada para elegir un código $C$ con largos $l(x) = \lceil -\log q(x) \rceil$.*
*El largo medio de $C$ aplicado a $X \sim p(x)$ satisface*

$$H(X) + D(p||q) \leq E_p\left[l(X)\right] < H(X) + D(p||q) + 1\,.$$

$D(p||q)$ representa el incremento en el largo de código medio por haber asumido $q$ en lugar de $p$.

# La distribución incorrecta

## Demostración

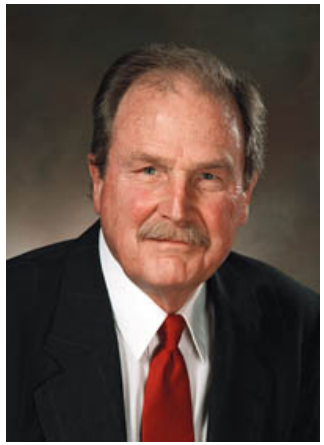
$$
\begin{aligned}
E_p\left[l(X)\right] &= \sum_x p(x)\left\lceil \log\frac{1}{q(x)}\right\rceil \\
&< \sum_x p(x)\left(\log\frac{1}{q(x)}+1\right) \\
&= \sum_x p(x)\log\left(\frac{p(x)}{q(x)}\frac{1}{p(x)}\right)+1 \\
&= \sum_x p(x)\log\frac{p(x)}{q(x)}+\sum_x p(x)\log\frac{1}{p(x)}+1 \\
&= D(p||q)+H(X)+1
\end{aligned}
$$

□

# Códigos de Huffman

David A. Huffman propuso en 1952 un algoritmo para construir un código instantáneo óptimo para una distribución de probabilidad arbitraria sobre un alfabeto finito.

Es un procedimiento recursivo que en cada paso agrupa los $D$ símbolos menos probables para formar un nuevo símbolo.



## Ejemplo

Para una fuente con $\mathcal{X} = \{x_1, x_2, x_3, x_4, x_5\}$, con probabilidades $p = \{0{,}25, 0{,}25, 0{,}2, 0{,}15, 0{,}15\}$ hallar el código de Huffman con $D = 2$ y $D = 3$.

# Códigos de Huffman

- En cada iteración se reduce en $D-1$ el número de símbolos del alfabeto.
- Por lo tanto, para que $|\mathcal{X}|-1$ símbolos sean eliminados sucesivamente hasta que el alfabeto colapse en un único símbolo (la raíz del árbol), $|\mathcal{X}|-1$ debe ser múltiplo de $D-1$.
- Si $D>2$ puede ocurrir que no haya suficientes símbolos para combinar en cada iteración.
- En este caso, antes de comenzar el algoritmo se agregan símbolos "falsos" con probabilidad cero; tantos como sean necesarios para que $|\mathcal{X}|-1$ sea múltiplo de $D-1$.
- Estos símbolos falsos recibirán palabras de código que luego descartaremos.

# Comentarios sobre los códigos de Huffman

- Códigos de Huffman y número de preguntas.
  Para hallar el número óptimo de preguntas (con respuesta sí/no) para determinar un objeto (conociendo las probabilidades de los objetos) ¿cuál es la secuencia de preguntas más eficientes? El número medio de preguntas $E[Q]$ siguiendo el esquema de Huffman cumple
  $$H(X) \leq E[Q] < H(X) + 1\,.$$

- Huffman con pesos.
  El algoritmo de Huffman minimiza $\sum p_i l_i$ para números no negativos $\{p_i\}$, independientemente de $\sum p_i$. Se puede darle "pesos" $\omega_i$ a los largos $l_i$ y aplicar el algoritmo para minimizar $\sum \omega_i l_i$.
- Códigos de Huffman y códigos de Shannon.
  El código de Shannon asigna longitudes de palabras de largo $\lceil -\log p_i \rceil$, lo cual no tiene por qué ser óptimo. Si $p(X) = \{0{,}9, 0{,}1\}$, $l_S(X) = \{4, 1\}$ mientras que $l_H(X) = \{1, 1\}$. ¿Siempre es más corto?
  Sea $p(X) = \{1/3, 1/3, 1/4, 1/12\}$. Las posibles longitudes del código de Huffman son $(2, 2, 2, 2)$ y $(1, 2, 3, 3)$. Las longitudes del código de Shannon son $(2, 2, 2, 4)$.
  A pesar de que uno puede ser más corto que el otro para algún símbolo en particular, *en media* el código de Huffman **nunca** es más largo.

# Optimalidad de los códigos de Huffman

Demostraremos la optimalidad de los códigos de Huffman binarios ($D = 2$).
Consideramos un alfabeto $\mathcal{X} = \{x_1, x_2, \ldots, x_m\}$ de símbolos de una fuente y asumimos, sin pérdida de generalidad, que las probabilidades están ordenadas $p_1 \geq p_2 \geq \cdots \geq p_m$.

## Lema

*Todo código instantáneo óptimo para* $[p_1, p_2, \ldots, p_m]$ *satisface*

1. *si* $p_j > p_k$, *entonces* $l_j \leq l_k$,
2. *para toda palabra de código de largo máximo existe otra del mismo largo que sólo difiere en el último símbolo.*

*Existe un código óptimo que adicionalmente satisface*

3. *las palabras de código asociadas a* $x_m$ *y* $x_{m-1}$ *difieren entre sí sólo en el último símbolo.*

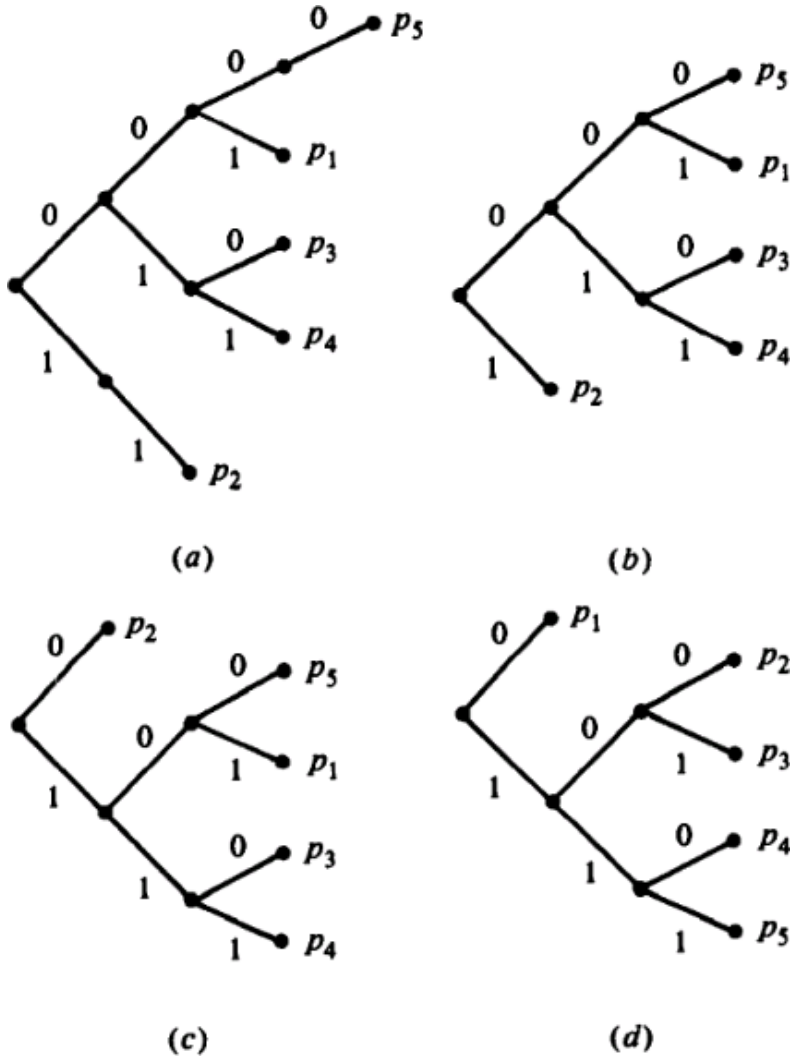
# Optimalidad de los códigos de Huffman

## Demostración

1. Si $p_j > p_k$ pero $l_j > l_k$, intercambiando las palabras $j$ y $k$ obtenemos un código con largo medio menor.
2. Sea $w$ una palabra de código de largo máximo y sea $w'$ la cadena que resulta de invertir el último símbolo de $w$. Si $w'$ no es una palabra de código, entonces podemos eliminar el último símbolo de $w$, obteniendo un nuevo código que es de prefijo y tiene un largo medio menor.
3. Sea $C$ un código óptimo y sean $w$, $w'$ palabras de código de largo máximo que difieren entre sí sólo en el último símbolo. Como $p_m \leq p_{m-1} \leq p_i$, $1 \leq i < m-1$, intercambiando estas palabras de código con las asignadas a los símbolos $x_m$ y $x_{m-1}$ el largo medio no aumenta.

# Optimalidad de los códigos de Huffman

Si $p_1 \geq p_2 \geq \cdots \geq p_m$ existe un código óptimo con largos tal que $l_1 \leq l_2 \leq \cdots \leq l_{m-1} = l_m$ y $C(x_{m-1})$ y $C(x_m)$ difieren en el último bit.



# Definición inductiva de código de Huffman

Un código de Huffman, $C_H^{(m)}$, para una variable aleatoria $X$ definida sobre un alfabeto de tamaño $m$, $\mathcal{X} = \{x_1, x_2, \ldots, x_m\}$, con probabilidades $p_1 \geq p_2 \geq \cdots \geq p_m$, se define inductivamente como

1. Para $m = 2$, definimos $C_H^{(m)} = \{0, 1\}$.
2. Para $m > 2$, sean

$$\begin{aligned} \mathcal{Y} &= \{y_1, y_2, \ldots, y_{m-2}, y_{m-1}\}, \\ p_Y &= (p_1, p_2, \ldots, p_{m-2}, p_{m-1} + p_m), \end{aligned}$$

y sea $C_H^{(m-1)}$ un código de Huffman para $Y \sim p_Y$. Definimos

$$C_H^{(m)}(x_i) = \begin{cases} C_H^{(m-1)}(y_i), & 1 \leq i < m-1, \\ C_H^{(m-1)}(y_{m-1})0, & i = m-1, \\ C_H^{(m-1)}(y_{m-1})1, & i = m. \end{cases}$$

# Demostración por inducción de la optimalidad del código de Huffman

Para $m = 2$ es obvio; para $m > 2$:

- Sean $\mathcal{Y}$, $p_Y$ y $C_H^{(m-1)}$ como en la definición.
- Sea $C^{(m)}$ un código óptimo para $X$ tal que

$$C^{(m)}(x_{m-1}) = w0\,, \quad C^{(m)}(x_m) = w1\,.$$

- Sea $C^{(m-1)}$ el código para $Y$ (no necesariamente óptimo) definido como

$$C^{(m-1)}(y_i) = \begin{cases} C^{(m)}(x_i)\,, & 1 \le i < m-1\,, \\ w\,, & i = m-1\,. \end{cases}$$

- Los largos esperados de estos códigos satisfacen

$$\begin{aligned} L_H^{(m)} &= L_H^{(m-1)} + p_{m-1} + p_m\,, \\ L^{(m-1)} &= L^{(m)} - (p_{m-1} + p_m)\,. \end{aligned}$$

  Sumando y reordenando obtenemos

$$L_H^{(m)} = L^{(m)} + \left(L_H^{(m-1)} - L^{(m-1)}\right) \le L^{(m)}\,.$$

# Optimalidad competitiva del código de Shannon

Probamos que los códigos de Huffman son óptimos en media y vimos un ejemplo donde un código de Shannon asigna una palabra de código de menor longitud que Huffman para un símbolo particular (pero no en media).

## Teorema

*Sea $l(x)$ la longitud de palabra asociada al código de Shannon y $l'(x)$ la longitud de palabra asociada por cualquier otro código UD, entonces*

$$\Pr\left\{l(X) \geq l'(X) + c\right\} \leq 2^{1-c}\,.$$

## Ejemplo

La probabilidad que $l'(X)$ sea 5 bits más corta que $l(X)$ es menor a $\frac{1}{16} = 0{,}0625$

# Optimalidad competitiva del código de Shannon

## Demostración

$$
\begin{aligned}
\Pr\left\{l(X) \geq l'(X) + c\right\} &= \Pr\left\{\lceil -\log p(X) \rceil \geq l'(X) + c\right\} \\
&\leq \Pr\left\{-\log p(X) \geq l'(X) + c - 1\right\} \\
&= \Pr\left\{p(X) \leq 2^{-l'(X)-c+1}\right\} \\
&= \sum_{x:p(x)\leq 2^{-l'(x)-c+1}} p(x) \\
&\leq \sum_{x:p(x)\leq 2^{-l'(x)-c+1}} 2^{-l'(x)-c+1} \\
&\leq \sum_{x} 2^{-l'(x)} 2^{1-c} \leq 2^{1-c}
\end{aligned}
$$

# Optimalidad competitiva del código de Shannon

## Teorema

*Para una distribución de probabilidad diádica $p(x)$, sea $l(x) = -\log p(x)$ la longitud de palabra asociada al código de Shannon y $l'(x)$ la longitud de palabra asociada por cualquier otro código UD. Entonces,*

$$\Pr\left\{l(X) < l'(X)\right\} \geq \Pr\left\{l(X) > l'(X)\right\} .$$

*La igualdad se da si y sólo si $l(X) = l'(X)$ para todo $X$.*

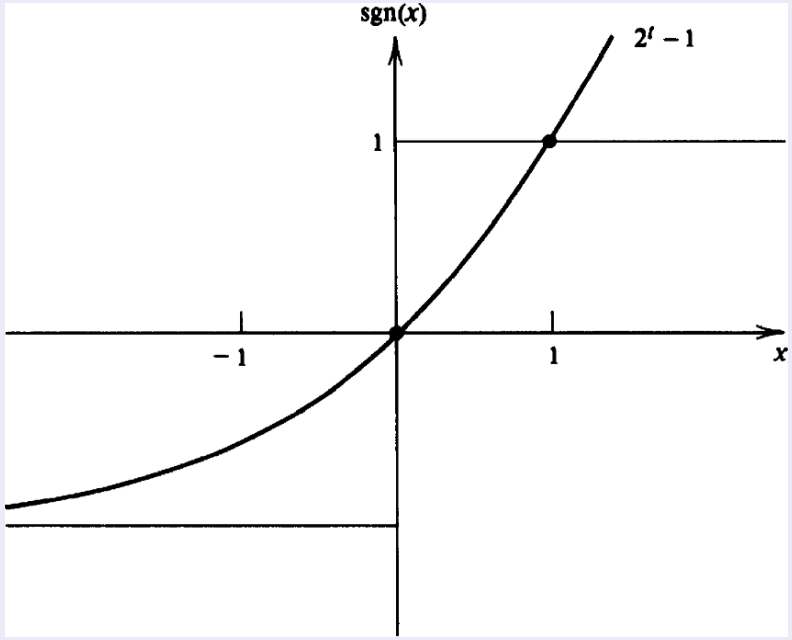
# Optimalidad competitiva del código de Shannon

## Demostración

Usaremos la función

$$\text{sgn}(t) = \begin{cases} 1 & \text{si } t > 0 \\ 0 & \text{si } t = 0 \\ -1 & \text{si } t < 0 \end{cases}$$

$$\text{sgn}(t) \leq 2^t - 1 \, \forall t \in \mathbb{N}$$




(cont.)

# Optimalidad competitiva del código de Shannon

## Demostración

$$
\begin{aligned}
\Pr\left\{l'(X) < l(X)\right\} \quad - \quad & \Pr\left\{l'(X) > l(X)\right\} = \\
= \quad & \sum_{x:l'(x)<l(x)} p(x) - \sum_{x:l'(x)>l(x)} p(x) \\
= \quad & \sum_{x} p(x)\mathsf{sgn}(l(x) - l'(x)) \\
\leq \quad & \sum_{x} p(x)(2^{l(x)-l'(x)} - 1) \\
= \quad & \sum_{x} 2^{-l(x)}(2^{l(x)-l'(x)} - 1) \\
= \quad & \sum_{x} 2^{-l'(x)} - \sum_{x} 2^{-l(x)} \\
= \quad & \sum_{x} 2^{-l'(x)} - 1 \leq 1 - 1 = 0
\end{aligned}
$$

# Optimalidad competitiva del código de Shannon

## Corolario

*Para distribuciones de probabilidad no diádicas*

$$E\left[sgn(l(X) - l'(X) - 1)\right] \leq 0$$

*donde* $l(x) = \lceil -\log p(x) \rceil$ *y* $l'(x)$ *es el largo de cualquier otro código UD.*

# Optimalidad competitiva del código de Shannon

## Demostración

$$
\begin{aligned}
E\left[\mathsf{sgn}(l(X) - l'(X) - 1)\right] &= \sum_x p(x)\mathsf{sgn}(l(x) - l'(x) - 1) \\
&\leq \sum_x p(x)(2^{l(x)-l'(x)-1} - 1) \\
&= \sum_x p(x)(2^{\lceil -\log p(x)\rceil - l'(x)-1} - 1) \\
&\leq \sum_x p(x)(2^{-\log p(x) - l'(x)} - 1) \\
&= \sum_x p(x)\left(\frac{2^{-l'(x)}}{p(x)} - 1\right) \\
&= \sum_x 2^{-l'(x)} - 1 \leq 1 - 1 = 0
\end{aligned}
$$

□

# Codigos de Shannon-Fano-Elias

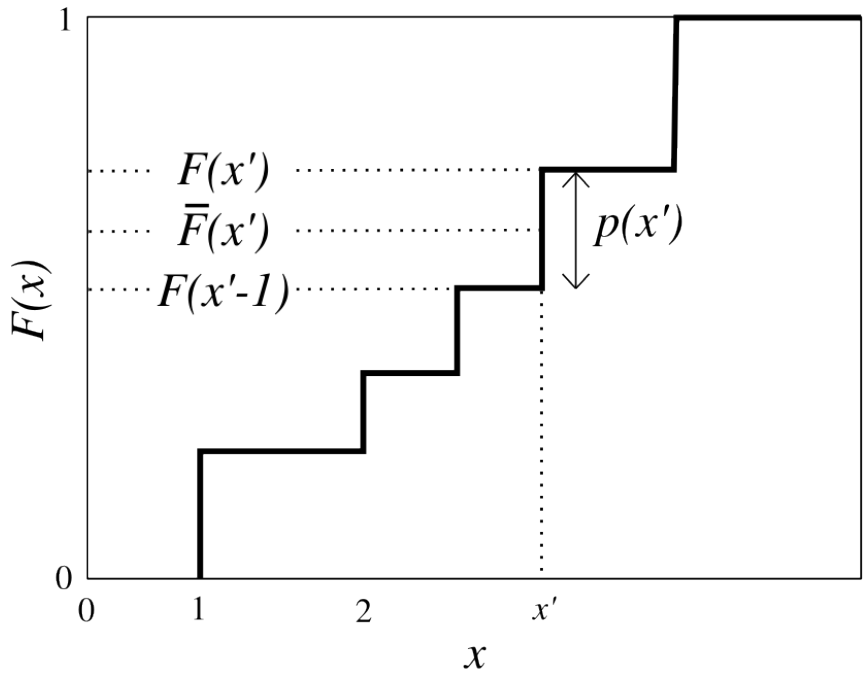
Es un procedimiento constructivo que utiliza la función de distribución acumulativa para asignar palabras de código.
Supongamos que $\mathcal{X} = \{1, 2, \ldots, m\}$ y $p(x) > 0 \ \forall x \in \mathcal{X}$. La función de distribución acumulativa es

$$F(x) = \sum_{a \leq x} p(a) \, .$$

Definimos

$$\bar{F}(x) = \sum_{a < x} p(a) + \frac{1}{2} p(x) \, .$$

- $\bar{F}(a) \neq \bar{F}(b)$ si $a \neq b$
- $\bar{F}(x)$ identifica a $x$



# Codigos de Shannon-Fano-Elias

- Sea $\lfloor \bar{F}(x) \rfloor_{l(x)}$ el resultado de truncar la representación binaria de $\bar{F}(x)$ a $l(x)$ dígitos, donde $l(x) = \lceil -\log p(x) \rceil + 1$. Entonces

$$\bar{F}(x) - \lfloor \bar{F}(x) \rfloor_{l(x)} \leq 2^{-l(x)} \leq \frac{p(x)}{2} = \bar{F}(x) - F(x-1)\,,$$

  donde la última desigualdad surge de que $l(x) \geq -\log p(x) + 1$ y extendemos la definición de $F$ a $F(0) = 0$.
- Adicionalmente, se cumple $\lfloor \bar{F}(x) \rfloor_{l(x)} \leq \bar{F}(x) < F(x)$.
- Entonces $\lfloor \bar{F}(x) \rfloor_{l(x)}$ pertenece al intervalo $\big[F(x-1), F(x)\big)$, que identifica a $x$.
- El código resultante es instantáneo porque para $y > x$ tenemos

$$\begin{aligned} \lfloor \bar{F}(y) \rfloor_{l(y)} - \lfloor \bar{F}(x) \rfloor_{l(x)} &\geq F(x) - \bar{F}(x) \\ &= \frac{p(x)}{2} \\ &\geq 2^{-l(x)}\,, \end{aligned}$$

  por lo cual las representaciones de $\lfloor \bar{F}(y) \rfloor_{l(y)}$ y $\lfloor \bar{F}(x) \rfloor_{l(x)}$ difieren en alguno de los primeros $l(x)$ dígitos.

# Códigos de Shannon-Fano-Elias

Recordando que $l(x) = \lceil -\log p(x) \rceil + 1$, el largo de código medio satisface

$$L = \sum_x p(x) l(x) = \sum_x p(x) \left( \lceil -\log p(x) \rceil + 1 \right) < H(X) + 2$$

## Ejemplo

| $x_i$ | $p(x)$ | $F(x)$ | $\bar{F}(x)$ | $\bar{F}(x)$ en binario | $l(x)$ | palabra |
|---|---|---|---|---|---|---|
| $x_1$ | 0.25 | 0.25 | 0.125 | 0.00100 | 2+1 | 001 |
| $x_2$ | 0.5 | 0.75 | 0.5 | 0.10000 | 1+1 | 10 |
| $x_3$ | 0.125 | 0.875 | 0.8125 | 0.11010 | 3+1 | 1101 |
| $x_4$ | 0.125 | 1 | 0.9375 | 0.11110 | 3+1 | 1111 |

- $L = 2{,}75$ bits, y $H(X) = 1{,}75$ bits. (Huffman alcanza la entropía.)
- Algunas palabras de código se pueden acortar, mejorando el largo medio; pero si se quita el último bit de *todas* las palabras, se pierde la condición de prefijo.

# Códigos de Shannon-Fano-Elias

## Ejemplo

| $x_i$ | $p(x)$ | $F(x)$ | $\bar{F}(x)$ | $\bar{F}(x)$ en binario | $l(x)$ | palabra |
|---|---|---|---|---|---|---|
| $x_1$ | 0.25 | 0.25 | 0.125 | 0.001 | 3 | 001 |
| $x_2$ | 0.25 | 0.25 | 0.375 | 0.011 | 3 | 011 |
| $x_3$ | 0.2 | 0.7 | 0.6 | 0.1**0011**... | 4 | 1001 |
| $x_4$ | 0.15 | 0.85 | 0.775 | 0.110**0011**... | 4 | 1100 |
| $x_5$ | 0.15 | 1 | 0.925 | 0.111**0110**... | 4 | 1110 |

- La entropía es $H(X) = 2{,}28$ bits. El largo medio es $L(C) = 3{,}5$ bits, $1{,}2$ bits más que el de Huffman.

# Codificación aritmética

- El largo medio de código de Shannon-Fano-Elias puede ser mayor que el de Huffman, como en los ejemplos anteriores.
- Sin embargo, a diferencia de Huffman, la construcción de este código admite su aplicación a extensiones a bloques grandes de la fuente de forma eficiente.
- A esta extensión del código de Shannon-Fano-Elias, atendiendo a las limitaciones impuestas por la precisión aritmética finita con que se implementa, se le denomina codificación aritmética.
- Para muchos modelos de fuente, el cálculo de la palabra de código se puede realizar secuencialmente recurriendo a la siguiente relación

$$\begin{aligned}\sum_{y^n<x^n} P(y^n) &= \sum_{y^{n-1}<x^{n-1},a\in\mathcal{X}} P(y^{n-1}a) + \sum_{a<x_n} P(x^{n-1}a) \\ &= \sum_{y^{n-1}<x^{n-1}} P(y^{n-1}) + P(x^{n-1}) \sum_{a<x_n} P(a|x^{n-1})\end{aligned}$$