Compresión de Datos sin Pérdida

Codificación Universal

Álvaro Martín

[1]Instituto de Computación,
Facultad de Ingeniería
almartin@fing.edu.uy

[2]PEDECIBA Informática

Motivación

- Algoritmos de Lempel-Ziv son universales para un conjunto extremadamente amplio de fuentes de información (procesos estacionarios y ergódicos).
- La velocidad de convergencia del largo de código por símbolo a la tasa de entropía es $O\left(\frac{1}{\log n}\right)$.
- Si nos conformamos con una familia más restringida de modelos de fuentes, ¿podemos hacerlo mejor?
- ¿Cuánto mejor?, ¿qué es lo mejor a lo que podemos aspirar?

Universalidad

- Consideramos una clase paramétrica $\mathcal{C} = \{P_\theta\}_{\theta \in \Theta}$ sobre un alfabeto $\mathcal{X}$.
- Informalmente, buscamos una secuencia de códigos, $\{C_n\}_{n>0}$, $C_n : \mathcal{X}^n \to \mathcal{B}^\star$, tal que el largo de código por símbolo de C_n se aproxime al óptimo para P_θ a medida que n crece, cualquiera sea $\theta \in \Theta$.
- Equivalentemente, buscamos una secuencia de distribuciones de probabilidad, $\{Q_n\}_{n>0}$, tal que Q_n aproxima a la distribución sobre $\mathcal{X}^n$ definida por P_θ a medida que n crece, cualquiera sea $\theta \in \Theta$.
- Definiciones formales en breve.
- Con un poco de abuso de nomenclatura y notación, en general nos referimos a $\{C_n\}_{n>0}$ y $\{Q_n\}_{n>0}$ simplemente como "un código" o "una distribución", respectivamente, y omitimos el subíndice n en la notación.

Universalidad

- Universalidad en media:
  - Un código es *universal en media* con respecto a $\mathcal{C}$ si

$$E_{P_\theta}\left[\frac{|C_n(X^n)|}{n}\right] - \frac{H(X^n)}{n} \xrightarrow[n\to\infty]{} 0\,, \quad \forall\theta \in \Theta\,.$$

  - Una distribución es *universal en media* con respecto a $\mathcal{C}$ si

$$\frac{D\left(P_\theta||Q_n\right)}{n} \xrightarrow[n\to\infty]{} 0\,, \quad \forall\theta \in \Theta\,.$$

- Universalidad puntual:
  - Un código es *puntualmente universal* con respecto a $\mathcal{C}$ si

$$\frac{1}{n} \max_{x^n \in \mathcal{X}^n} \left\{|C(x^n)| + \log P_{ML}(x^n)\right\} \xrightarrow[n\to\infty]{} 0\,.$$

  - *Arrepentimiento* de Q_n con respecto a $\mathcal{C}$ para $x^n \in \mathcal{X}^n$

$$\mathrm{REG}_{Q_n}(x^n) \triangleq -\log Q_n(x^n) + \log P_{ML}(x^n)\,.$$

  - Una distribución *puntualmente universal* con respecto a $\mathcal{C}$ si

$$\frac{1}{n} \max_{x^n \in \mathcal{X}^n} \left\{\mathrm{REG}_{Q_n}(x^n)\right\} \xrightarrow[n\to\infty]{} 0\,.$$

¿Cómo buscar códigos universales?

- Estrategias:
  - Códigos en dos partes (por ejemplo ejercicio 6 del práctico 5).
  - Estimación secuencial (plug-in codes).
  - Mezclas (de distribuciones de la clase en consideración)
- Vamos a ilustrar estas estrategias a través de la familia de modelos de Bernoulli, $\mathcal{C}_B = \{P_\theta\}_{\theta \in \Theta}$, donde θ es la probabilidad del símbolo 0 y $\Theta = (0, 1)$.
- Las mismas ideas se generalizan a clases más complejas.
- **Notación:** Denotamos con $n_0(x^n)$ y $n_1(x^n)$ la cantidad de ocurrencias de los símbolos 0 y 1, respectivamente, en la secuencia x^n. A veces omitimos x^n de la notación.

Codificación enumerativa

Codificación en dos partes.

1. Se describe n_0 usando $\lceil \log(n+1) \rceil$ bits.
2. Se describe x^n especificando $\mathrm{I}(x^n)$, el índice de x^n en una enumeración del conjunto

$$\mathcal{T}(x^n) = \{y^n \in \mathcal{X}^n : n_0(y^n) = n_0(x^n)\}.$$

Para esta segunda parte se utilizan $\left\lceil \log \binom{n}{n_0} \right\rceil$ bits.

Codificación enumerativa - Ejemplo

$x^n = 101011$, $n = 6$, $n_0 = 2$, $\binom{n}{n_0} = \frac{6!}{4!2!} = 15$.

$\mathrm{I}(y^n)$	y^n		$\mathrm{I}(y^n)$	y^n
0	001111		8	101110
1	010111		9	110011
2	011011		10	110101
3	011101		11	110110
4	011110		12	111001
5	100111		13	111010
6	101011		14	111100
7	101101			

Largo parte 1: $\lceil \log 7 \rceil = 3$. Largo parte 2: $\lceil \log 15 \rceil = 4$

$$C(x^n) = \underbrace{010}_{n_0}\,\underbrace{0110}_{\mathrm{I}(x^n)}$$

Estimación secuencial con Estimador de Laplace

- Para cada $i = 1, 2, \ldots, n$, asignamos una distribución de probabilidad condicional para el símbolo en posición i, dada la secuencia de símbolos anteriores, x^{i-1}.
- Estimador de Laplace:

$$Q^{\mathrm{L}}(a|x^{i-1}) = \frac{n_a(x^{i-1}) + 1}{i - 1 + |\mathcal{X}|}, \quad a \in \mathcal{X}.$$

- La probabilidad asignada a la secuencia completa x^n es

$$Q_n^{\mathrm{L}}(x^n) = \prod_{i=1}^{n} Q^{\mathrm{L}}(x_i|x^{i-1}).$$

Estimador de Laplace - alfabeto binario

- Para $\mathcal{X} = \{0, 1\}$ obtenemos

$$Q^{\mathrm{L}}(0|x^{i-1}) = \frac{n_0(x^{i-1}) + 1}{i + 1}, \quad Q^{\mathrm{L}}(1|x^{i-1}) = \frac{n_1(x^{i-1}) + 1}{i + 1}.$$

$$Q_n^{\mathrm{L}}(x^n) = \prod_{i=1}^{n} Q^{\mathrm{L}}(x_i|x^{i-1}) = \frac{n_0! n_1!}{(n+1)!} = \frac{1}{n+1} \binom{n}{n_0}^{-1}.$$

- **Ejemplo:**

$$Q_n^{\mathrm{L}}(0010111) = \frac{1}{2}\frac{2}{3}\frac{1}{4}\frac{3}{5}\frac{2}{6}\frac{3}{7}\frac{4}{8}$$

- Largo de código ideal $\approx$ largo con codificación enumerativa (a menos de redondeos)

$$-\log Q_n^{\mathrm{L}}(x^n) = \log(n+1) + \log \binom{n}{n_0}.$$

Mezcla uniforme

- Si $\omega(\theta)$ es no negativa y $\int_{\theta\in\Theta}\omega(\theta)d\theta = 1$, entonces

$$Q(x^n) = \int_{\theta\in\Theta} P_\theta(x^n)\omega(\theta)d\theta$$

  es una distribución de probabilidad sobre $\mathcal{X}^n$.

- En particular para la familia de Bernoulli, con $\omega(\theta) = 1$ obtenemos un mezcla uniforme

$$\begin{aligned} Q_n^{\mathsf{U}}(x^n) &= \int_0^1 P_\theta(x^n)d\theta \\ &= \int_0^1 \theta^{n_0}(1-\theta)^{n_1}d\theta\,. \end{aligned}$$

Mezcla uniforme

- Para $n_1 = 0$, tenemos $n_0 = n$ y $P_\theta(x^n) = \theta^n$, de modo que

$$Q_n^{\mathsf{U}}(x^n) = \int_0^1 \theta^n d\theta = \left.\frac{\theta^{n+1}}{n+1}\right|_0^1 = \frac{1}{n+1}\,. \tag{1}$$

- Para $n_1 > 0$, integrando por partes obtenemos

$$\int_0^1 \underbrace{\theta^{n_0}}_{g'}\underbrace{(1-\theta)^{n_1}}_{f}\,d\theta = \underbrace{\frac{\theta^{n_0+1}}{n_0+1}}_{g}\underbrace{(1-\theta)^{n_1}}_{f}\Bigg|_0^1 + \int_0^1 \underbrace{\frac{\theta^{n_0+1}}{n_0+1}}_{g}\underbrace{n_1(1-\theta)^{n_1-1}}_{-f'}\,d\theta\,.$$

$$= \frac{n_1}{n_0+1}\int_0^1 \theta^{n_0+1}(1-\theta)^{n_1-1}d\theta\,.$$

Mezcla uniforme

- Para $n_1 > 0$ tenemos

$$\begin{aligned}
\int_0^1 \theta^{n_0}(1-\theta)^{n_1} d\theta &= \frac{n_1}{n_0+1} \int_0^1 \theta^{n_0+1}(1-\theta)^{n_1-1} d\theta \\
&= \frac{n_1}{n_0+1}\frac{n_1-1}{n_0+2} \int_0^1 \theta^{n_0+2}(1-\theta)^{n_1-2} d\theta \\
&\vdots \\
&= \frac{n_1(n_1-1)\dots 1}{(n_0+1)(n_0+2)\dots(n_0+n_1)} \int_0^1 \theta^{n} d\theta \\
&= \binom{n}{n_0}^{-1} \frac{1}{n+1}
\end{aligned}$$

- Por lo tanto, combinando con (1), obtenemos en general

$$Q_n^{\mathsf{U}}(x^n) = \frac{1}{n+1}\binom{n}{n_0}^{-1} = Q_n^{\mathsf{L}}(x^n)\,. \tag{2}$$

Todos los caminos conducen a ...

- Siguiendo tres estrategias distintas llegamos a básicamente al mismo resultado.
- El largo de código que obtuvimos es

$$|C(x^n)| = \log n + \log \binom{n}{n_0} + O(1)\,.$$

- El arrepentimiento para estas distribuciones es

$$\mathrm{REG}_{Q_n}(x^n) = \log n + \log \binom{n}{n_0} + \log P_{ML}(x^n) + O(1)\,. \quad (3)$$

- Difícil de interpretar ...

Fórmula de Stirling para factoriales

$$\sqrt{2\pi n}\left(\frac{n}{e}\right)^n e^{1/(12n+1)} \leq n! \leq \sqrt{2\pi n}\left(\frac{n}{e}\right)^n e^{1/12n}\,,\ \ n>0\,.$$

Como $e^{1/12n} \in [1, e^{1/12}]$ y $e^{1/(12n+1)} \in [1, e^{1/13}]$, tenemos

$$n! = \Theta\left(\frac{n^{n+\frac{1}{2}}}{e^n}\right)\,. \tag{4}$$

Entonces, para $n_0 > 0$ y $n_1 > 0$, podemos escribir

$$\begin{aligned}
\binom{n}{n_0} = \frac{n!}{n_0!n_1!} &= \Theta\left(\frac{n^{n+\frac{1}{2}}e^{n_0}e^{n_1}}{e^n n_0^{n_0+\frac{1}{2}} n_1^{n_1+\frac{1}{2}}}\right) = \Theta\left(\frac{n^{n_0+n_1}}{n_0^{n_0}n_1^{n_1}}\left(\frac{n}{n_0 n_1}\right)^{\frac{1}{2}}\right) \\
&= \Theta\left(\left(\frac{n}{n_0}\right)^{n_0}\left(\frac{n}{n_1}\right)^{n_1}\left(\frac{n}{n_0 n_1}\right)^{\frac{1}{2}}\right) \qquad (5) \\
&= \Theta\left(\frac{1}{P_{ML}(x^n)}\left(\frac{n}{n_0 n_1}\right)^{\frac{1}{2}}\right)\,. \qquad (6)
\end{aligned}$$

¿Cómo se compara nuestro largo de código con $-\log P_{ML}(x^n)$?

- Tomando logaritmos en (6) obtenemos

$$\log \binom{n}{n_0} = -\log P_{ML}(x^n) - \frac{1}{2}\log \frac{n_0 n_1}{n} + \Theta(1)\,. \qquad (7)$$

- Sea $\epsilon \in (0,1)$. Si x^n es tal que $\epsilon < \frac{n_0}{n}$ y $\epsilon < \frac{n_1}{n}$, entonces

$$-\frac{1}{2}\log \frac{n}{n_0 n_1} = \frac{1}{2}\log \Big(n \underbrace{\frac{n_0}{n}\frac{n_1}{n}}_{\in(\epsilon^2,1)} \Big) = \frac{1}{2}\log n + \Theta(1)\,, \qquad (8)$$

y por lo tanto en este caso tenemos

$$\log \binom{n}{n_0} = -\log P_{ML}(x^n) - \frac{1}{2}\log n + \Theta(1)\,. \qquad (9)$$

Universalidad de Enumerativa/Laplace/Mezcla uniforme

- Para probabilidades empíricas "lejos" de cero, a partir de (3) y (9) obtenemos

$$\mathrm{REG}_{Q_n}(x^n) = \frac{1}{2}\log n + \Theta(1)\,.$$

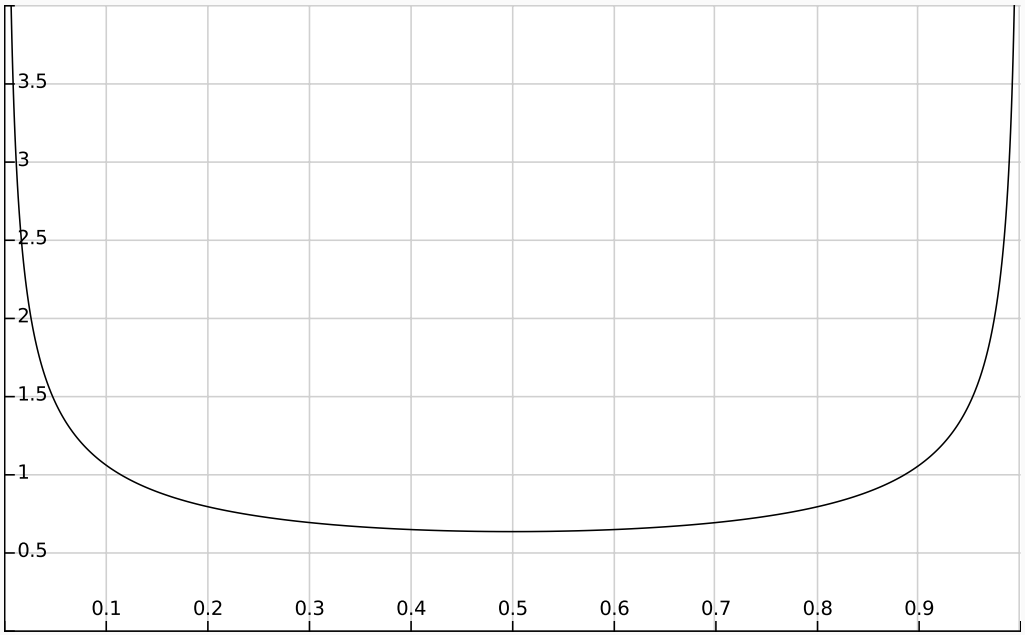
- Cerca de las puntas $\mathrm{REG}_{Q_n}(x^n)$ se acerca a $\log n$ (ver (3) para $n_0 = 0$ o (8) para $n_0 = 1$).
- En cualquier caso, el código obtenido es puntualmente universal para la familia de modelos de Bernoulli, aunque el exceso de largo de código por encima de $-\log P_{ML}(x^n)$ se comporta diferente en cada caso.
- ¿Se puede hacer mejor?

Asignación de Krichevsky-Trofimov (KT)

- Mezcla (Jeffreys prior). Para Bernoulli:

$$Q_n^{\mathrm{KT}}(x^n) = \int_0^1 P_\theta(x^n)\omega(\theta)d\theta\,, \quad \omega(\theta) = \left(\pi\sqrt{\theta(1-\theta)}\right)^{-1}.$$

- $\omega(\theta)$ asigna más peso a "las puntas"



Asignación de Krichevsky-Trofimov (KT)

- Mezcla (Jeffreys prior). Para Bernoulli:

$$Q_n^{\mathrm{KT}}(x^n) = \int_0^1 P_\theta(x^n)\omega(\theta)d\theta\,, \quad \omega(\theta) = \left(\pi\sqrt{\theta(1-\theta)}\right)^{-1}\,.$$

- En función de contadores n_a, $a \in \mathcal{X}$, para modelos i.i.d.:

$$Q_n^{\mathrm{KT}}(x^n) = \frac{\prod_{a\in\mathcal{X}}\prod_{i=0}^{n_a-1}\left(i+\frac{1}{2}\right)}{\prod_{i=0}^{n-1}\left(i+\frac{|\mathcal{X}|}{2}\right)}\,. \tag{10}$$

- Regla de asignación secuencial de probabilidad condicional

$$Q^{\mathrm{KT}}(a|x^{i-1}) = \frac{n_a(x^{i-1})+\frac{1}{2}}{i-1+\frac{|\mathcal{X}|}{2}}\,.$$

Aproximación de arrepentimiento de KT vía Stirling

Para $m > 0$,

$$\begin{aligned}
\prod_{i=0}^{m-1}\left(i+\frac{1}{2}\right) &= \frac{1}{2^m}\prod_{i=0}^{m-1}(2i+1) = \frac{1\times 3\times 5\times \ldots \times(2m-1)}{2^m} \\
&= \frac{(2m)!}{2^m}\frac{1}{2\times 4\times 6\times \ldots \times 2m} \\
&= \frac{(2m)!}{2^m}\frac{1}{2^m\times m!} \\
&= 2^{-2m}\frac{(2m)!}{m!} \\
&= \Theta\left(2^{-2m}\frac{(2m)^{2m+\frac{1}{2}}}{e^{2m}}\frac{e^m}{m^{m+\frac{1}{2}}}\right) \\
&= \Theta\left(\frac{m^m}{e^m}\right). \qquad (11)
\end{aligned}$$

Aproximación de arrepentimiento de KT vía Stirling

- Para $|\mathcal{X}|$ par, el denominador en (10) es

$$\prod_{i=0}^{n-1}\left(i+\frac{|\mathcal{X}|}{2}\right)=\frac{\left(n-1+\frac{|\mathcal{X}|}{2}\right)!}{\left(\frac{|\mathcal{X}|}{2}-1\right)!}=\Theta\left(n!n^{\frac{|\mathcal{X}|}{2}-1}\right)=\Theta\left(e^{-n}n^{n+\frac{|\mathcal{X}|-1}{2}}\right)$$

- Para $|\mathcal{X}|$ impar, usando (11), el denominador en (10) es

$$\prod_{i=0}^{n-1}\left(i+\frac{|\mathcal{X}|}{2}\right)=\prod_{i=0}^{n-1}\left(i+\frac{|\mathcal{X}|-1}{2}+\frac{1}{2}\right)=\prod_{j=\frac{|\mathcal{X}|-1}{2}}^{n+\frac{|\mathcal{X}|-1}{2}-1}\left(j+\frac{1}{2}\right)$$

$$=\Theta\left(\prod_{j=0}^{n+\frac{|\mathcal{X}|-1}{2}-1}\left(j+\frac{1}{2}\right)\right)=\Theta\left(\frac{\left(n+\frac{|\mathcal{X}|-1}{2}\right)^{n+\frac{|\mathcal{X}|-1}{2}}}{e^{n+\frac{|\mathcal{X}|-1}{2}}}\right)$$

$$=\Theta\left(e^{-n}n^{n+\frac{|\mathcal{X}|-1}{2}}\right) \qquad (12)$$

Aproximación de arrepentimiento de KT vía Stirling

- Usando (11), el numerador en (10) es

$$\prod_{a\in\mathcal{X}}\prod_{i=0}^{n_a-1}\left(i+\frac{1}{2}\right)=\Theta\left(\prod_{a\in\mathcal{X},n_a>0}\frac{n_a^{n_a}}{e^{n_a}}\right)=\Theta\left(e^{-n}\prod_{a\in\mathcal{X},n_a>0}n_a^{n_a}\right)$$

- De las expresiones para numerador y denominador obtenemos,

$$Q_n^{\mathrm{KT}}(x^n)=\Theta\left(\frac{\prod_{a\in\mathcal{X},n_a>0}n_a^{n_a}}{n^{n+\frac{|\mathcal{X}|-1}{2}}}\right), \tag{13}$$

de modo que

$$\begin{aligned}-\log Q_n^{\mathrm{KT}}(x^n) &= -\sum_{a\in\mathcal{X},n_a>0} n_a\log n_a + n\log n + \frac{|\mathcal{X}|-1}{2}\log n + O(1)\\ &= -\log P_{ML}(x^n) + \frac{|\mathcal{X}|-1}{2}\log n + \Theta(1)\,.\end{aligned}$$

- Tenemos $\mathrm{REG}_{Q_n^{\mathrm{KT}}}(x^n)=\frac{1}{2}\log n+\Theta(1)$ para toda $x^n\in\mathcal{X}^n$.

Arrepentimiento mín − máx

- Resumiendo:
  - Con codificación enumerativa, asignación de Laplace, o mezcla uniforme de modelos de Bernoulli, obtenemos un arrepentimiento para la clase $\mathcal{C}_B$ de modelos de Bernoulli no mayor a $\log n + \Theta(1)$
  - Para secuencias con probabilidades empíricas alejadas de cero es aún menor, $\frac{1}{2}\log n + \Theta(1)$.
  - Con la asignación de Krichevsky-Trofimov $\mathrm{REG}_{Q_n^{\mathrm{KT}}}(x^n) = \frac{1}{2}\log n + \Theta(1)$ para toda $x^n \in \mathcal{X}^n$.
- ¿Podemos seguir mejorando? ¿Cuánto?
- ¿Qué significa "mejor"?
- Definimos el *arrepentimiento* mín − máx *de una clase* $\mathcal{C}$

$$\mathrm{MMR}_n(\mathcal{C}) = \inf_{Q_n \text{ sobre } \mathcal{X}^n} \left\{ \max_{x^n \in \mathcal{X}^n} \mathrm{REG}_{Q_n}(x^n) \right\}. \qquad (14)$$

Máxima verosimilitud normalizada (NML)

- Distribución de probabilidad NML para una clase $\mathcal{C}$:

$$Q_n^{\mathrm{NML}}(x^n) = \frac{P_{ML}(x^n)}{\sum_{y^n \in \mathcal{X}^n} P_{ML}(y^n)}\,, \quad x^n \in \mathcal{X}^n\,. \tag{15}$$

- $\mathrm{REG}_{Q_n^{\mathrm{NML}}}(x^n)$ es independiente de x^n,

$$\begin{aligned} \mathrm{REG}_{Q_n^{\mathrm{NML}}}(x^n) &= -\log Q_n^{\mathrm{NML}}(x^n) + \log P_{ML}(x^n) && (16)\\ &= \log \sum_{y^n \in \mathcal{X}^n} P_{ML}(y^n)\,. && (17) \end{aligned}$$

- El hecho de que ninguna secuencia sea "favorecida" sobre otras hace que Q_n^{NML} sea óptima en sentido mín $-$ máx.

Optimalidad de NML en sentido mín − máx

Theorem (Shtarkov, 87)
Sea $\mathcal{C}$ una clase paremétrica de modelos. Tenemos

$$MMR_n(\mathcal{C}) = \log \sum_{y^n \in \mathcal{X}^n} P_{ML}(y^n) = \max_{x^n \in \mathcal{X}^n} REG_{Q_n^{NML}}(x^n)\,. \qquad (18)$$

Demostración.

$$\begin{aligned}
\mathrm{MMR}_n(\mathcal{C}) &= \inf_{Q_n \text{ sobre } \mathcal{X}^n} \left\{ \max_{x^n \in \mathcal{X}^n} \left\{ -\log Q_n(x^n) + \log P_{ML}(x^n) \right\} \right\} \\
&= \inf_{Q_n \text{ sobre } \mathcal{X}^n} \left\{ \max_{x^n \in \mathcal{X}^n} \log \frac{P_{ML}(x^n)}{Q_n(x^n)} \right\} \\
&= \inf_{Q_n \text{ sobre } \mathcal{X}^n} \left\{ \log \max_{x^n \in \mathcal{X}^n} \frac{P_{ML}(x^n)}{Q_n(x^n)} \right\}\,.
\end{aligned}$$

Ahora,

$$\log \max_{x^n \in \mathcal{X}^n} \frac{P_{ML}(x^n)}{Q_n(x^n)} \geq \log \sum_{x^n \in \mathcal{X}^n} Q_n(x^n) \frac{P_{ML}(x^n)}{Q_n(x^n)} = \max_{x^n \in \mathcal{X}^n} \mathrm{REG}_{Q_n^{\mathrm{NML}}}(x^n)\,,$$

y la igualdad se alcanza para $Q_n = Q_n^{\mathrm{NML}}$. □

Aproximación de $\mathrm{MMR}_n(\mathcal{C})$ para la clase $\mathcal{C}_B$

$$\begin{aligned}\sum_{y^n\in\mathcal{X}^n} P_{ML}(y^n) &= \sum_{n_0=0}^{n}\binom{n}{n_0}\left(\frac{n_0}{n}\right)^{n_0}\left(\frac{n-n_0}{n}\right)^{n-n_0}\\ &= 2+\sum_{n_0=1}^{n-1}\binom{n}{n_0}\left(\frac{n_0}{n}\right)^{n_0}\left(\frac{n-n_0}{n}\right)^{n-n_0}.\end{aligned}$$

Acotando $\binom{n}{n_0}$ según (5),

$$\begin{aligned}\sum_{y^n\in\mathcal{X}^n} P_{ML}(y^n) &= \Theta\left(n^{\frac{1}{2}}\sum_{n_0=1}^{n-1} n_0^{-\frac{1}{2}}(n-n_0)^{-\frac{1}{2}}\right)\\ &= \Theta\left(n^{\frac{1}{2}}\sum_{n_0=1}^{n-1}\left(\frac{n_0}{n}\right)^{-\frac{1}{2}}\left(1-\frac{n_0}{n}\right)^{-\frac{1}{2}}\frac{1}{n}\right)\\ &= \Theta\left(n^{\frac{1}{2}}\int_0^1\theta^{-\frac{1}{2}}(1-\theta)^{-\frac{1}{2}}\,d\theta\right)=\Theta\left(n^{\frac{1}{2}}\right)\quad(19)\end{aligned}$$

Optimalidad asintótica de KT en sentido mín − máx para $\mathcal{C}_B$

- De (18) y (19) obtenemos

$$\text{MMR}_n(\mathcal{C}_B) = \log \sum_{y^n \in \mathcal{X}^n} P_{ML}(y^n) = \frac{1}{2} \log n + \Theta(1)\,. \qquad (20)$$

- Por otra parte, para la asignación KT tenemos

$$\max_{x^n \in \mathcal{X}^n} \left\{ \text{REG}_{Q_n^{\text{KT}}}(x^n) \right\} = \frac{1}{2} \log n + \Theta(1)\,.$$

- En este sentido, decimos que Q_n^{KT} es asintóticamente óptima en el sentido mín − máx para $\mathcal{C}_B$, ya que el máximo arrepentimiento con Q_n^{KT} difiere en $\Theta(1)$ de $\text{MMR}_n(\mathcal{C}_B)$.

Aproximación de MMR_n para clases más generales

Theorem
Sea $\mathcal{C}_F = \{P_\theta\}_{\theta\in\Theta}$ la clase de modelos definida sobre un alfabeto finito $\mathcal{X}$ por una FSM F con conjunto de estados S, donde las distribuciones de probabilidad condicionales en los estados se definen mediante el parámetro $\theta \in \Theta$, de dimensión $|S|(|\mathcal{X}| - 1)$. Se cumple

$$MMR_n(\mathcal{C}_F) = \frac{|S|(|\mathcal{X}| - 1)}{2} \log n + \Theta(1)\,. \tag{21}$$

Cota de Rissanen (Hipótesis)

Theorem (Rissanen, 84)
Sea $\mathcal{C} = \{P_\theta\}_{\theta \in \Theta}$ una clase de modelos sobre un alfabeto finito $\mathcal{X}$, donde Θ es un subconjunto acotado de $\mathbb{R}^k$. Para $\theta \in \Theta$, definimos $E_n(\theta)$ como la bola de radio $\frac{\log n}{\sqrt{n}}$ y centro θ,

$$E_n(\theta) = B_{\frac{\log n}{\sqrt{n}}}(\theta)\,.$$

Definimos $\mathcal{X}_n(\theta)$ como el conjunto de secuencias $x^n \in \mathcal{X}^n$ cuyo estimador de máxima verosimilitud, $\hat{\theta}(x^n)$, pertenece a $E_n(\theta)$,

$$\mathcal{X}_n(\theta) = \{x^n \in \mathcal{X}^n : \hat{\theta}(x^n) \in E_n(\theta)\}\,.$$

Supongamos que $P_\theta\Big(\mathcal{X}_n(\theta)\Big) \geq 1 - \delta_n$ para todo $\theta \in \Theta$, donde δ_n tiende a 0 con n.

Cota de Rissanen (Tesis)

Tesis:
Para toda distribución $\{Q_n\}_n > 0$ y todo $\epsilon > 0$ se cumple

$$D(P_\theta || Q_n) \geq (1-\epsilon)\frac{k}{2}\log n\,, \qquad \forall \theta \in \Theta \setminus \Psi_n\,, \qquad (22)$$

donde el volumen de Ψ_n tiende a 0 con n.

Interpretación:
En otras palabras, para todo código unívocamente decodificable, $\{C_n\}_{n>0}$, y todo $\epsilon > 0$ se cumple

$$E_{P_\theta}\left[\frac{|C_n(X^n)|}{n}\right] \geq H_\theta(X^n) + (1-\epsilon)\frac{k}{2}\log n\,, \qquad \forall \theta \in \Theta \setminus \Psi_n\,, \quad (23)$$

donde el volumen de Ψ_n tiende a 0 con n, y $H_\theta(X^n)$ denota la entropía de $X^n \sim P_\theta$.

Cota de Rissanen (Demostración)

Sea

$$\Psi'_n = \left\{ \theta \in \Theta : \sum_{x^n \in \mathcal{X}_n(\theta)} P_\theta(x^n) log \frac{P_\theta(x^n)}{Q_n(x^n)} < \left(1 - \frac{\epsilon}{2}\right) \frac{k}{2} \log n \right\}.$$

Para $\theta \in \Theta \setminus \Psi'_n$ tenemos

$$D(P_\theta || Q_n) \geq \left(1 - \frac{\epsilon}{2}\right) \frac{k}{2} \log n + \sum_{x^n \in \mathcal{X}^n \setminus \mathcal{X}_n(\theta)} P_\theta(x^n) log \frac{P_\theta(x^n)}{Q_n(x^n)},$$

y usando que $\ln 1/z \geq 1 - z$ obtenemos

$$\begin{aligned} D(P_\theta || Q_n) &\geq \left(1 - \frac{\epsilon}{2}\right) \frac{k}{2} \log n + (\log e) \sum_{\substack{x^n \in \mathcal{X}^n \setminus \mathcal{X}_n(\theta) \\ P_\theta(x^n) > 0}} \left(P_\theta(x^n) - Q_n(x^n)\right) \\ &\geq \left(1 - \frac{\epsilon}{2}\right) \frac{k}{2} \log n - (\log e) \\ &\geq (1 - \epsilon) \frac{k}{2} \log n, \quad n > M. \end{aligned}$$

Cota de Rissanen (Demostración)

- Definimos

$$\Psi_n = \begin{cases} \Psi'_n\,, & \text{para } n > M\,, \\ \Theta\,, & \text{para } n \leq M\,. \end{cases}$$

- Con esta definición, se cumple (22).
- Resta probar que $Vol(\Psi_n) \to 0$.
  - Definimos N_n como la máxima cantidad de bolas $E_n(\theta)$ que pueden ubicarse con centros θ en Ψ_n de forma que estas bolas sean disjuntas (recordar que Θ es acotado).
  - Sea $C_n \subseteq \Psi_n$ un conjunto de N_n centros de bolas $E_n(\theta)$ disjuntas.
  - Duplicando el radio de las bolas, cubrimos todo Ψ_n,

$$\Psi_n \subseteq \bigcup_{\theta \in C_n} B_{\frac{2\log n}{\sqrt{n}}}(\theta)\,,$$

de modo que

$$Vol(\Psi_n) = O\left(N_n \left(\frac{2\log n}{\sqrt{n}}\right)^k\right) = O\left(N_n n^{-\frac{k}{2}} \log^k n\right)\,. \quad (24)$$

Cota de Rissanen (Demostración)

- Vamos a probar que existe $c \in (0,1)$ tal que, para n suficientemente grande,

$$Q_n\Big(\mathcal{X}_n(\theta)\Big) \geq n^{-c\frac{k}{2}}\,, \quad \forall \theta \in \Psi_n\,. \tag{25}$$

- Como los conjuntos $\mathcal{X}_n(\theta)$, $\theta \in C_n$, son disjuntos, (25) implica que

$$N_n n^{-c\frac{k}{2}} \leq \sum_{\theta \in C_n} Q_n\Big(\mathcal{X}_n(\theta)\Big) \leq 1\,. \tag{26}$$

- Por lo tanto tenemos $N_n \leq n^{c\frac{k}{2}}$, y reemplazando en (24) obtenemos

$$Vol(\Psi_n) = O\left(n^{(c-1)\frac{k}{2}} \log^k n\right) = o(1)\,. \tag{27}$$

Cota de Rissanen (Prueba de (25))

Sea $\theta \in \Psi_n$, $n > M$. Por la definición de Ψ_n tenemos

$$\begin{aligned}
\left(1-\frac{\epsilon}{2}\right)\frac{k}{2}\log n &> P_\theta\left(\mathcal{X}_n(\theta)\right)\sum_{x^n\in\mathcal{X}_n(\theta)}\frac{P_\theta(x^n)}{P_\theta\left(\mathcal{X}_n(\theta)\right)}\left(-log\frac{Q_n(x^n)}{P_\theta(x^n)}\right)\\
&\geq -P_\theta\left(\mathcal{X}_n(\theta)\right)\log\sum_{\substack{x^n\in\mathcal{X}_n(\theta)\\ P_\theta(x^n)>0}}\frac{P_\theta(x^n)}{P_\theta\left(\mathcal{X}_n(\theta)\right)}\frac{Q_n(x^n)}{P_\theta(x^n)}\\
&= -P_\theta\left(\mathcal{X}_n(\theta)\right)\log\sum_{\substack{x^n\in\mathcal{X}_n(\theta)\\ P_\theta(x^n)>0}}\frac{Q_n(x^n)}{P_\theta\left(\mathcal{X}_n(\theta)\right)}\\
&\geq -P_\theta\left(\mathcal{X}_n(\theta)\right)\log\sum_{x^n\in\mathcal{X}_n(\theta)}\frac{Q_n(x^n)}{P_\theta\left(\mathcal{X}_n(\theta)\right)}\\
&= P_\theta\left(\mathcal{X}_n(\theta)\right)\log\frac{P_\theta\left(\mathcal{X}_n(\theta)\right)}{Q_n\left(\mathcal{X}_n(\theta)\right)}. \qquad (28)
\end{aligned}$$

Cota de Rissanen (Prueba de (25))

Despejando de (28) obtenemos

$$\begin{aligned}
\log Q_n\left(\mathcal{X}_n(\theta)\right) &> \log P_\theta\left(\mathcal{X}_n(\theta)\right) - \left(1-\frac{\epsilon}{2}\right)\frac{k}{2}\frac{\log n}{P_\theta\left(\mathcal{X}_n(\theta)\right)} \\
&= \left(\frac{\log P_\theta\left(\mathcal{X}_n(\theta)\right)}{\frac{k}{2}\log n} - \frac{1-\frac{\epsilon}{2}}{\log P_\theta\left(\mathcal{X}_n(\theta)\right)}\right)\frac{k}{2}\log n \\
&\geq \left(\frac{\log(1-\delta_n)}{\frac{k}{2}\log n} - \frac{1-\frac{\epsilon}{2}}{\log(1-\delta_n)}\right)\frac{k}{2}\log n \\
&> -c\frac{k}{2}\log n\,, \quad c > 1-\frac{\epsilon}{2}\,, n > M'\,. \qquad (29)
\end{aligned}$$

Exponenciando en (29) obtenemos (25).