Teoría información

Notación:
∼: Nota , •: Tema , >: Propiedad , #: Definir,
∗: Ejemplo, ‖: Demostración

• **Matemáticas:**

$$\lim_{p\to 0} p\log(p) = 0,\ \sum_{k=0}^{n} ar^k = a\left(\frac{1-r^{n+1}}{1-r}\right),\ \sum_{n=1}^{\infty} nr^n = \frac{r}{(1-r)^2}$$

$$\sum_{k=k_1}^{k_2} ar^k = a\frac{r^{k_1}-r^{k_2+1}}{1-r},\ \int_a^b x^n f(kx)dx = \frac{1}{k^{n+1}}\int_{ka}^{kb} uf(u)du$$

$$\int_{\mathbb{R}} e^{-(ax^2+bx+c)} = \sqrt{\frac{\pi}{a}}e^{(\frac{b^2}{4a}-c)},\ \delta(p-p') = \frac{1}{2\pi}\int_{\mathbb{R}} e^{i(p-p')\xi}d\xi$$

> **La truca para estas**

$$I = \int_a^b x^n e^{\pm\alpha x}dx,\ \partial_\alpha e^{\pm\alpha x} = (\pm 1)xe^{t\alpha x},\ \partial_\alpha^n e^{\pm\alpha x} = (\pm 1)^n x^n e^{\pm\alpha x}$$

$$I = (\pm 1)^n\partial_\alpha^n\int_a^b e^{\pm\alpha x}dx \Rightarrow I = (\pm 1)^{n+1}\partial_\alpha^n\frac{e^{\pm\alpha x}}{\alpha}\Big|_a^b$$

$$\Rightarrow \int_0^\infty x^n e^{-x/a}dx = n!a^{n+1},\ \int_0^\infty x^{2n+1}e^{-x^2/a^2}dx = \frac{n!}{2}a^{2n+2}$$

$$\int_0^\infty x^{2n}e^{-x^2/a^2}dx = (-1)^n\sqrt{\pi}\partial_\alpha^n(\alpha^{-1/2})$$

Otra: $\int_0^t e^{\alpha\varepsilon}d\varepsilon = \frac{1}{\alpha}(e^{\alpha t}-1)$

Distr. cauchy $f(x;x_0,\gamma) = \dfrac{1}{\pi\gamma\left[1+\left(\frac{x-x_0}{\gamma}\right)^2\right]}$, x_0: ancho, g: ancho a mitad de altura

• **Probabilidades:** A : event A, $\bar{A}$: Not event A

> $P(A)+P(\bar{A}) = 1$

> $P(A\cup B) = P(A)+P(B)-P(A\cap B)$ (A or B)

> $P(A,B) = P(A\cap B) = P(A\mid B)P(B)$ (A and B)

> $P(E_i\mid E) = \frac{P(E_i)P(E|E_i)}{\sum_{j=1}^k P(E|E_j)P(E_j)}$ (Baye's Formula)

> $P(A\mid B) = P(A)$ (Independent Trials)

> $E[X] = \langle x\rangle = \int_{\mathbb{R}} xp(x)dx$ (continuo)
$E[X] = \langle x\rangle = \sum_{i=1}^k x_i p_i$ (discreto)

> $\mathrm{Var}(X) = E\left[(X-E[X])^2\right]$
$\mathrm{Var}(X) = E\left[X^2-2XE[X]+E[X]^2\right]$
$= E\left[X^2\right]-2E[XE[X]]+E[X]^2$
$= E\left[X^2\right]-E[X]^2$

> Probability of one of k mutually exclusive events
$P = P(E_1)+P(E_2)+\ldots+P(E_k)$

> Total Probability, k mutually exclusive events:

$$P(E) = P(E\mid E_1)P(E_1)+\ldots+P(E\mid E_k)P(E_k)$$
$$= \sum_{i=1}^k P(E\mid E_i)P(E_i)$$

> Variable change, prob. density functions:

$$g(\vec{y}) = f(\vec{x})\left|\det\left(\frac{d\vec{x}}{d\vec{y}}\right)\right|$$

I. INTRODUCCIÓN

• **Notación:**

$\mathcal{A}_X$: Alfabeto de la variable aleatoria X

p_i : Probabilidades

l_i : # de preguntas necesarias para llegar al elemento i del alfabeto

H : Entropía

$C(x)$: palabra clave asociada a x.

$\ell(x)$: longitud de $C(x)$, medida en dígitos.

Num. medio de preguntas en pie de igualdad:

$$\left\langle \begin{array}{c}\#\text{ preguntas en pie de igualdad}\\ \text{bien elegidas}\end{array}\right\rangle = \sum_i p_i\ell_i$$

Usando: $p_i = D^{-\ell_i} \Leftrightarrow \ell_i = -\log_D p_i$

$$\text{Se tiene: } \left\langle \begin{array}{c}\#\text{ preguntas}\\ \text{bien elegidas}\end{array}\right\rangle = -\sum_i p_i\log_D p_i$$

Entropía:

$$H = -\sum_i p_i\log_D p_i$$

> **Conversión de unidades:**

$$H_D = -\sum p_i\log_D p_i = -\sum p_i\frac{\log_{D'}(p_i)}{\log_{D'}(D)} = \frac{1}{\log_{D'}(D)}H_{D'}$$

• **Propiedades de H:**

> Anidación: Si $Y = f(X)$, $\mathcal{C}(y)\subset\mathcal{A}_X$ contiene a todo $x : x\xrightarrow{f} y \Rightarrow$

$$H(X) = H(Y)-\sum_y p(y)\sum_{x\in C(y)} p(x\mid y)\log[p(x\mid y)$$

> Cotas y sus implicaciones:

$$\underbrace{0}_{\text{Determinista}} \leq H(X) \leq \underbrace{\log|A_X|}_{\text{Uniforme}}$$

• **Caso bivariado:** Con $\mathbf{X} = (X_1, X_2)$ variable aleatoria bivariada, $\mathbf{X} \in \mathcal{A}_{X_1} \times \mathcal{A}_{X_2}$ con prob. conjunta $p(x_1, x_2)$

Entropía conjunta:

$$H(X_1, X_2) = -\sum_{x_1}\sum_{x_2} p(x_1, x_2) \log_D [p(x_1, x_2)]$$

Entropías marginales(una a una):

$$H(X_1) = -\sum_{x_1} p(x_1) \log_D [p(x_1)]$$

$$H(X_2) = -\sum_{x_2} p(x_2) \log_D [p(x_2)]$$

distr. marginales

$$p(x_1) = \sum_{x_2} p(x_1, x_2), \; p(x_2) = \sum_{x_1} p(x_1, x_2)$$

Entropía condicional:

$$H(X_1 \mid X_2) = -\sum_{x_2} p(x_2) \sum_{x_1} p(x_1 \mid x_2) \log_D [p(x_1 \mid x_2)]$$

$$H(X_2 \mid X_1) = -\sum_{x_1} p(x_1) \sum_{x_2} p(x_2 \mid x_1) \log_D [p(x_2 \mid x_1)]$$

> **Regla de la cadena:**

$$\begin{aligned} H(X,Y) &= H(X) + H(Y \mid X) \\ &= H(Y) + H(X \mid Y) \end{aligned}$$

$$\begin{aligned} H(X,Y) &= -\sum_x \sum_y p(x,y) \log[p(x,y)] \\ &= -\sum_x \sum_y p(x)p(y \mid x) \log[p(x)p(y \mid x)] \\ &= -\sum_x p(x) \sum_y p(y \mid x)\{\log[p(x)] + \log[p(y \mid x)]\} \\ &= -\sum_x p(x) \log[p(x)] \underbrace{\sum_y p(y \mid x)}_{1} \\ &\quad -\sum_x p(x) \sum_y p(y \mid x) \log[p(y \mid x)] \\ &= H(X) + H(Y \mid X) \end{aligned}$$

> **Regla de la cadena multivariable:**

$$H(X_1, X_2, \ldots, X_n) = \sum_{i=1}^{n} H(X_i \mid X_{i-1}, \ldots, X_1)$$

> **Cotas H condicional:**

$$\underbrace{0}_{\text{determinista}} \leq H(X \mid Y) \leq \underbrace{H(X)}_{\text{Independientes}} \leq \underbrace{\log|A_X|}_{\text{Uniforme}}$$

> **Cotas H multivariable:**

$$H(X_1, X_2, \ldots, X_N) \leq \sum_i H(X_i)$$

> $H(X|X) = H(X)$

> **Condición de independencia:**

X, Y son independientes $\Leftrightarrow H(X \mid Y) = H(X)$

X, Y son independientes $\Leftrightarrow H(Y \mid X) = H(Y)$

X, Y son independientes $\Leftrightarrow H(X,Y) = H(X) + H(Y)$

> $H(X, Y \mid Z) = H(X \mid Z) + H(Y \mid X, Z)$

II. CLASE 2:

Función convexa: $f(x)$ es convexa en el intervalo $[a,b] \Leftrightarrow \forall x_1, x_2 \in [a,b]$, $\forall \lambda \in [0,1]$:

$$\underbrace{f[\lambda x_1 + (1-\lambda)x_2]}_{\text{Gráfico de } f \text{ en } x \in (x_1, x_2)} \leq \underbrace{\lambda f(x_1) + (1-\lambda) f(x_2)}_{\text{Cuerda en } x \in (x_1, x_2)}$$

Estrictamente convexa: La función es convexa y únicamente en los extremos se cumple la igualdad.

> Si $f'' > 0 (f'' > 0)$ en un intervalo $I \Rightarrow$ es estrictamente convexa (cóncava) en I.

> **Desigualdad de Jansen:** Si $f : \mathbb{R} \to \mathbb{R}$ es convexa, y X es una variable aleatoria$\Rightarrow$

$$\langle f(X) \rangle \geq f(\langle X \rangle), \quad \forall \text{ distribución } p(x)$$

∼ Si f es estrictamente convexa, $\langle f(X) \rangle = f(\langle X \rangle) \Leftrightarrow X$ es determinista.

> **Desigualdad de la suma de logaritmos:** Dados $a_1, \ldots, a_n \in \mathbb{R}_0^+$, $b_1, \ldots, b_n \in \mathbb{R}^+$

$$\sum_{i=1}^{n} a_i \log\left(\frac{a_i}{b_i}\right) \geq \left(\sum_{i=1}^{n} a_i\right) \log\left(\frac{\sum_{j=1}^{n} a_j}{\sum_{k=1}^{n} b_k}\right)$$

∼ $(\geq)$ se vuelve $(=) \Leftrightarrow \frac{a_i}{b_i} = \text{cte.}$ (no depende de i).

Divergencia de Kullback-Leibler: Dadas $p(x_i)$, $q(x_i)$ distr. de prob. con $x_i \in A_X, \forall i$:

$$D_{\mathrm{KL}}(p\|q) = \sum_i p(x_i) \log\left[\frac{p(x_i)}{q(x_i)}\right]$$

> **Justificando en término de # preguntas sobrantes:**

$$\langle \#^{\text{preguntas}}_{\text{sobrantes}} \rangle = \left\langle \#^{\text{preguntas requeridas}}_{\text{con la estrategia de q}} \right\rangle - \left\langle \#^{\text{preguntas requeridas}}_{\text{con la estrategia de p}} \right\rangle$$

Además, $\langle \#^{\text{preguntas requeridas}}_{\text{con la estrategia de q}} \rangle = \sum_i \underbrace{p(x_i)}$

$$= -\sum_i p(x_i) \log_D [q(x_i)]$$

• **Propiedades D_{KL}:**

> $\underbrace{0}_{p=q} \leq D_{\mathrm{KL}}(p\|q)$

> D_{KL} es convexa

- ∃ (muchos) casos en los que:

  > $D_{\mathrm{KL}}(p\|q) \neq D_{\mathrm{KL}}(q\|p)$ (Asimetría)

  > $D_{\mathrm{KL}}(p\|q) + D_{\mathrm{KL}}(q\|r) < D_{\mathrm{KL}}(r\|p)$ (Desigualdad triangular)

  > D_{KL} diverge ($=\infty$)

Información mutua: Con variables aleatorias X,Y con distr. conjunta $p(x,y)$, la info. mutua $I(X;Y)$ entre ellas se define como

$$\begin{aligned} I(X;Y) &= H(Y) - H(Y \mid X) \\ &= H(X) - H(X \mid Y) \\ &= H(X) + H(Y) - H(X,Y) \\ &= D_{\mathrm{KL}}[p(x,y)\|p(x)p(y)] \\ &= \sum_{x\in\{0,1\}} \sum_{y\in\{0,1\}} p(x,y) \log \left[\frac{p(x,y)}{p(x)p(y)}\right] \end{aligned}$$

∼ Las dos variables se separan por ; Por ejemplo,

$$\begin{aligned} I\left(X_1, X_2; Y_1, Y_2, Y_3\right) &= H\left(X_1, X_2\right) \\ &\quad - H\left(X_1, X_2 \mid Y_1, Y_2, Y_3\right) \end{aligned}$$

Información mutua condicionada:

$$\begin{aligned} I(X;Y \mid Z) &= H(X \mid Z) - H(X \mid Y, Z) \\ &= H(Y \mid Z) - H(Y \mid X, Z) \\ &= H(X \mid Z) + H(Y \mid Z) - H(X, Y \mid Z) \\ &= \sum_z p(z) D_{\mathrm{KL}}[p(x,y \mid z)\|p(x \mid z)p(y \mid z)] \end{aligned}$$

donde X, Y, Z son variables aleatorias.

I **condicionada multivariada:**

$$I\left(X_1, X_2, \ldots, X_n; Y\right) = \sum_{i=1}^{n} I\left(X_i; Y \mid X_{i-1}, X_{i-2}, \ldots, X_1\right)$$

$$\begin{aligned} &I\left(X_1, X_2, \ldots, X_n; Y\right) = \\ &= H\left(X_1, X_2, \ldots, X_n\right) - H\left(X_1, X_2, \ldots, X_n \mid Y\right) \\ &= \sum_{i=1}^{n} H\left(X_i \mid X_{i-1}, \ldots, X_1\right) \\ &- \sum_{i=1}^{n} H\left(X_i \mid X_{i-1}, \ldots, X_1, Y\right) \\ &= \sum_{i=1}^{n} I\left(X_i; Y \mid X_1, X_2, \ldots, X_{i-1}\right) \end{aligned}$$

> $I(X;Y,Z) = I(X;Y) + I(X;Z \mid Y)$

• **Propiedades de I mutua:**

> $I(X)$(y $H(X)$)no depende de X, solo de sus prob.

> $I(X;X) = H(X)$ (viene de $H(X \mid X) = 0$)

> $I(X;Y) = I(Y;X)$ (simetría)

> $Y = f(X), f \overset{\text{funcion}}{\underset{\text{inyectiva}}{}} \Rightarrow I(X;Y) = H(X) = H(Y)$

> Variables aleatorias $X, Y, Z = f(Y)$
$\Rightarrow I(X;Y) \geq I(X;Z)$

> **Cota I mutua:**

$$\begin{aligned} \underbrace{0}_{\text{independiente}} &\leq I(X;Y) \leq \underbrace{\text{mín}[H(X), H(Y)]}_{\text{determinista}} \\ \underbrace{0}_{\text{independiente}} &\leq I(X;Y) \leq \underbrace{\text{mín}\left[\log|\mathcal{A}_X|, \log|\mathcal{A}_Y|\right]}_{\text{uniforme}} \end{aligned}$$

Estadística suficiente: Un mapeo $z = f(y)$ es una estadística suficiente $\Leftrightarrow I(X;Y) = I(X;Z)$.

III. CLASE 3:

Convergencia en probabilidad: Una secuencia de variables aleatorias $(X_1, X_2, \ldots) \to x_0$ en prob. $\Leftrightarrow \forall\varepsilon > 0 : \Pr\left[|X_n - x_0| > \varepsilon\right] \to 0$ cuando $n \to \infty$. Aquí $\Pr\left[|X_n - x_0| > \varepsilon\right] \to 0$ involucra un límite normal, ya que la prob. no es una variable aleatoria.

$$\text{Es decir, } \int_{x_n : |x_n - x_0| > \varepsilon} P(x)\mathrm{d}x \to 0$$

Convergencia en valor cuadratico medio: Una secuencia de variables aleatorias $(X_1, X_2, \ldots) \to x_0$ en valor cuad. medio $\Leftrightarrow \left\langle (X_n - x_0)^2 \right\rangle \to 0$ cuando $n \to \infty$.

$$\text{Es decir, } \int_{-\infty}^{+\infty} P\left(x_n\right)\left(x_n - x_0\right)^2 \, \mathrm{d}x_n \to 0$$

Convergencia con probabilidad 1: Una secuencia de variables aleatorias $(X_1, X_2, \ldots) \to x_0$ con prob. 1, o casi con seguridad $\Leftrightarrow \Pr\left(X_n = x_0\right) \to 1$, cuando $n \to \infty$.

$$\text{Es decir, } \lim_{\delta\to 0} \int_{-\infty}^{+\infty} P\left(x_n\right) \frac{e^{-(x_n - x_0)^2/2S^2}}{\sqrt{2\pi\delta^2}} \to 1$$

la prob. $\Pr\left(X_n = x_0\right)$ es la integral de una δ Dirac, y usamos una aprox. cualquiera de la δ. Lo importante es que $\delta \to 0$ debe tomarse antes que el otro limite.

> **Niveles de exigencia:**

en prob. $<$ en valor cuadr. medio $<$ con prob. 1

iid: independiente(s) e idénticamente distribuida(s).

> **La ley de los grandes números:** Dada una secuencia de variables aleatorias $X_1, X_2, \ldots, X_n$ **iid**s con prob. $p(x)$, donde $x \in \{a_1, \ldots, a_k\}$, la ley establece que:

$$\frac{1}{n}\sum_{i=1}^{n} f(X_i) \to \sum_{k=1}^{k} p(a_k) f(a_k) \text{ en prob.(Forma débil)}$$

Forma débil(en prob.), Forma fuerte(en prob. 1.)

La primera $\sum$ es un promedio de una función f evaluada en una muestra particular, es una variable aleatoria. La segunda $\sum$ es el valor medio de f aplicada a los elementos del alfabeto, no es una variable aleatoria.

> **Teorema de equipartición asintótica:** Si $X_1, X_2, \ldots, X_n$ son variables aleatorias **iid** con prob. $p(x)$, $\Rightarrow$ la variable aleatoria

$$-\frac{1}{n}\log p(X_1, \ldots, X_n) \to H(X) \text{ en probabilidad.}$$

$-\frac{1}{n}\log p(X_1, \ldots, X_n)$ es una variable aleatoria, ya que p(determinista) está evaluada en una variable aleatoria. $H(X)$, en cambio, es determinista.

\# **Conjunto típico:** El conjunto típico $\mathcal{A}_\varepsilon^{(n)}$ con respecto a $p(x)$ es el conjunto de secuencias $(x_1, \ldots, x_n) \in (\mathcal{A}_X)^n$ que cumplen

$$2^{-n[H(X)+\varepsilon]} \leq p(x_1, \ldots, x_n) \leq 2^{-n[H(X)-\varepsilon]}$$

$\sim$ $H(X)$ no depende de la cadena, X es el nombre de la variable aleatoria. La condición para pertenecer al $\mathcal{A}_\varepsilon^{(n)}$ es una condición sobre la prob. de la cadena.

$\sim$ **PUEDE NO SER 2** puede ser otra base, $e, 3$ o lo que convenga para el problema (unidades de info.).

- **Propiedades conjunto típico $\mathcal{A}_\varepsilon^{(n)}$:**

> Si una cadena $\in \mathcal{A}_\varepsilon^{(n)}$, el logaritmo de su prob. está muy cercana a $[-nH(X)]$. Esto es,

$$(x_1, \ldots, x_n) \in \mathcal{A}_\varepsilon^{(n)} \Leftrightarrow$$
$$H(X) - \varepsilon \leq -\frac{1}{n}\log_2 p(x_1, \ldots, x_n) \leq H(X) + \varepsilon$$

> La prob. de muestrear una cadena del conjunto típico se hace tan grande como querramos, aumentando la longitud de las cadenas. Esto es,

$$\forall \delta > 0, \exists n_0 \in \mathbb{N} / \forall n > n_0, \text{Prob}\left[(X_1, \ldots, X_n) \in \mathcal{A}_\varepsilon^{(n)}\right] > 1-\delta$$

> Cota a cardinalidad del conjunto típico:

$$\underbrace{(1-\varepsilon)2^{n[H(X)-\varepsilon]}}_{\forall \varepsilon > 0, \exists n_0 \in \mathbb{N}: \forall n > n_0 (n \text{ big enough})} \leq \left|\mathcal{A}_\varepsilon^{(n)}\right| \leq \underbrace{2^{n[H(X)+\varepsilon]}}_{siempre}$$

> **Teorema de codificación de una fuente:** Una secuencia de n variables aleatorias **iid**, cada una con entropía H, puede representarse por $n(H+\varepsilon)$ dígitos, con ε tan chico como queramos, y con prob. de pérdida de info. también tan chica como queramos, si n es lo suficientemente grande. Si se representan con menos de nH dígitos $\Rightarrow$ seguro que hay pérdida de info. (no infinitesimal).

IV. CLASE 4:

\# **Código fuente:** Dada una variable aleatoria X con alfabeto $\mathcal{A}_X$, y un alfabeto $\mathcal{A}_Y = \{0, 1, \ldots, D-1\}$, un código fuente para X es un mapeo $C : \mathcal{A}_X \to {}^{\text{cadenas finitas}}_{\text{de elementos de}}\ \mathcal{A}_Y$, que también suelen llamarse "cadenas D -arias finitas".

\# **Código fuente no singular:** Un código fuente es no singular si es invertible(como las matrices). Es decir, si símbolos distintos del alfabeto son mapeados en secuencias D-arias distintas.

$*$ **Ejemplo no singular y singular:** Los código C_0, C_1 mapean el alfabeto $\{a, b, c, d\}$ en tiras compuestas de {0,1} , tal que

No singular:

x	$C_0(x)$	$\ell(x)$
a	1000	4
b	0100	4
c	0010	4
d	0001	4

Singular:

x	$C_1(x)$	$\ell(x)$
a	10	2
b	0	1
c	10	2
d	010	3

\# **Código fuente extendido:** C^+ es un código fuente extendido para la variable aleatoria X si $C^+ : \mathcal{A}_X^+ \to \{0, 1, \ldots, D-1\}^+$, donde el supra-índice $+$ representa concatenación sin puntuación.

$*$ **Ejemplo, código fuente extendido:**

$$C_0(abacdb) = \underbrace{1000}_{C_0(a)}\underbrace{0100}_{C_0(b)}\underbrace{1000}_{C_0(a)}\underbrace{0010}_{C_0(c)}\underbrace{0001}_{C_0(d)}\underbrace{0100}_{C_0(b)}$$

\# **Código es unívocamente decodificable:** Código con extension no singular, es decir, si el código extendido es invertible.

\# **Prefijo/ autopuntuado/ indice/ instantáneo:** Un código fuente donde ninguna palabra clave es prefijo de otra palabra clave.

> **Prop.:** indices $\subset$ unívocamente decodificables $\subset$ no singulares $\subset$ códigos

\# **Longitud media:** La longitud media de un código fuente para la variable aleatoria X con alfabeto $\mathcal{A}_X = \{x_1, \ldots, x_k\}$ con probabilidades $p(x_1), \ldots p(x_k)$ es:

$$L = \sum_{x \in A_X} p(x)\ell(x)$$

> **Desigualdad de Kraft:** Con alfabetos $\mathcal{A}_X = \{x_1, \ldots, x_M\}$, $\mathcal{A}_Y = \{0, 1, \ldots, D-1\}$ y un conjunto $\{\ell_1, \ldots, \ell_M\} \subset \mathbb{Z}^+$,
$\exists$ código indice $C : \mathcal{A}_X \to {}^{\text{cadenas finitas}}_{\text{de elementos de}}\ \mathcal{A}_Y$, de longitudes $\{\ell_1, \ldots, \ell_M\} \Leftrightarrow$

$$\sum_{i=1}^{M} D^{-\ell_i} \leq 1$$

> **Cota a la longitud media:** La longitud media L de un código D-ario para una variable aleatoria X cumple:

$$H_D(X) \leq L$$

∼ $L = H_D(X) \Leftrightarrow \forall x, -\log_D[p(x)] \in \mathbb{Z}$ (prob. D-arias) Es decir, si todas las prob. son potencias de $\frac{1}{D}$.

> **Código Shannon-Elias-Fano:** Con longitudes

$$\ell(x) = [-\log_D p(x)]_+$$

donde, $[\cdot]_+$ significa redondeo entero *hacia arriba*. Es decir, si las prob. no son D-arias $\Rightarrow \exists x \in \mathcal{A}_X$ para el cual $-\log_D[p(x)] \notin \mathbb{Z}$. !!!! COMPLETAR con el algoritmo

> Siempre se puede codificar una variable aleatoria (cualquier prob.) con un código indice tal que

$$H_D(X) \leq L < H_D(X) + 1$$

> **Código Huffman:**

- Ordenar de mayor a menor(3,2,1) símbolos por su prob.
- Formar un arbolito $\mathbb{A}$ con nodo donde anoto la prob. de la suma de los menos probables y con hijos los símbolos menos probables, este nodo va en un nivel superior.
- Repetir los dos pasos anteriores.
- Completado $\mathbb{A}$ se usa para codificar. Para indicar el símbolo, se baja por $\mathbb{A}$, las ramas, ordenadas de izquierda a derecha (por decir un orden) indican un dígito D-ario que codifica el símbolo. Cada elección de rama al bajar es un dígito distinto, juntando los dígitos al bajar por el $\mathbb{A}$ se tiene el código fuente del símbolo que esta al final de la rama.

V. CLASE 5:

> **Convertir $\zeta \in [0,1] \subset \mathbb{R}$ a base D-aria(ζ_D):**

- Multiplicar ζ por D. La parte entera del resultado es el primer dígito de la expansión en binario.
- Con ζ_D hasta orden r, tomar la parte decimal del número obtenido en la multiplicación anterior (descartando la parte entera) y multiplicar por D. La parte entera del resultado es el dígito $(r+1)$-ésimo de ζ_D.

> **Transformación de la acumulada:** Dadas X, Y variables aleatorias definidas en intervalos finitos, con Y uniforme. Se puede convertir X en Y mediante:

$$y = F_X(x) = \int_{-\infty}^{x} f_X(\xi)d\xi, \quad f_X : \text{Dens. de prob. de } X$$

Codificación aritmética:

La entrada: Tomada como la representación en base K de $\zeta \in [0,1] \subset \mathbb{R}$ la fuente genera una secuencia $z_1 z_2 z_3 \ldots$, con $z_i \in \{0, 1, 2, \ldots, K-1\}$.

$$\zeta = \sum_{n=1}^{+\infty} z_n K^{-n}, \ f_Z(z) \text{ No uniforme}$$

La salida: Tomada como la representación en base D de un $F_z(\zeta)$, la fuente genera una secuencia $y_1 y_2 y_3 \ldots$, con $y_i \in \{0, 1, 2, \ldots, D-1\}$.

$$F_z(\zeta) = \sum_{n=1}^{+\infty} y_n D^{-n}, \ f_Y(y) \text{ Uniforme}$$

∼ Si ζ_D en un paso le da al palo, vale el valor del palo. Si $\zeta_{D=2}$ por ej. $\zeta_{D=2} < \frac{1}{2}$ poner 0, $\zeta_{D=2} \geq \frac{1}{2}$ poner 1.

VI. CLASE 6:

Canal discreto: $(\mathcal{A}_X, \mathcal{A}_Y, p(y \mid x))$ con el alfabeto de entrada $\mathcal{A}_X$, un alfabeto de salida $\mathcal{A}_Y$ y un conjunto de prob. de transición $p(y \mid x)$.

Canal sin memoria:

$$p(y_1, \ldots, y_n \mid x_1, \ldots, x_n) = \prod_{i=1}^{n} p(y_i \mid x_i)$$

Capacidad de un canal(sin memoria):

$$C = \max_{p(x)} I(X;Y)$$
$$= \max_{p(x)} \sum_{x,y} p(x)p(y \mid x) \log \left[\frac{p(y \mid x)}{\sum_{x'} p(x') p(y \mid x')} \right]$$

Optimización de un canal:

$$\boxed{\text{Fuente}} \xrightarrow{W} \boxed{\text{Compresor}} \xrightarrow{Z} \boxed{\text{Codificador}} \xrightarrow{X} \boxed{\text{Canal}} \xrightarrow{Y}$$

A_W puede no ser equiprobable, A_Z es equiprobable, A_X es tal que optimiza el canal.

Canal sin pérdida:
$H(X|Y) = 0 \equiv I(X;Y) = H(X)$
(Puede agregar ruido, pero no hay perdida de info.)
∼ $H(X|Y) = 0 \Leftrightarrow X$ es función de Y.
∼ $C = \max_{p(x)} I(X;Y) = \max_{p(x)} H(X) = \log|A_X|$. Lo cual ocurre para una p uniforme(Donde $H(X)$ es máxima).

Canal es determinista:
$H(Y|X) = 0 \equiv I(X;Y) = H(Y)$
(No agrega ruido, aunque puede comprimir dos entradas en una salida (perdiendo info.))
∼ $H(Y|X) = 0 \Leftrightarrow Y$ es función de X.
∼ $C = \max_{p(x)} I(X;Y) = \max_{p(x)} H(Y) = \log|A_Y|$. Lo cual ocurre para una p uniforme(Donde $H(Y)$ es máxima).

Canal biyectivo: Es sin perdida y determinista.

Canal inutil: $I(X;Y) = 0$

Canal simétrico: La matriz $p(y \mid x)$ es tal que:

Todas sus $\overset{\textbf{FILAS}}{\textbf{COLUMNAS}}$son permutaciones unas de otras.

∼ En un canal simétrico $I(X;Y) = H(Y) - H(Y \mid X)$ Con, $H(Y \mid X) = -\sum_x p(x) \sum_y p(y \mid x) \log[p(y \mid x)]$

Capacidad de un canal simétrico:

$$C = I(X;Y)\big|_{p(x)\text{uniforme}}$$

- **Propiedades de la capacidad:**

> $C \geq 0$ (ya que $I(X;Y) \geq 0$)

> $C \leq \log|\mathcal{A}_X|$ (ya que $I(X;Y) \leq \log|\mathcal{A}_X|$)

> $C \leq \log|\mathcal{A}_Y|$ (ya que $I(X;Y) \leq \log|\mathcal{A}_Y|$)

> La $p(x)$:$\max_{p(x)} I(X;Y)$ puede no ser única.

> Si $p(x)$: $I(X;Y) = \log|\mathcal{A}_X|$ o $\log|\mathcal{A}_Y| \Rightarrow C = \log|\mathcal{A}_Y|$ o $\log|\mathcal{A}_Y|$ con $p(x)$.

> **Canal binario simétrico:**

$p(y \mid x)$	$y=0$	$y=1$
$x=0$	$1-f$	f
$x=1$	f	$1-f$

$\Rightarrow C = 1+f\log f+(1-f)\log(1-f)$

> **Canal máquina de escribir ruidosa:**

$p(y \mid x)$	$y=0$	$y=1$	$y=2$	$\ldots$	$y=N-1$
$x=0$	$\frac{1}{2}$	$\frac{1}{2}$	0	$\ldots$	0
$x=1$	0	$\frac{1}{2}$	$\frac{1}{2}$	$\ldots$	0
$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$	$\vdots$
$x=N-1$	$\frac{1}{2}$	0	0	$\ldots$	$\frac{1}{2}$

En cada fila hay $N-2$ ceros y dos $1/2$. Lo mismo para cada columna $\therefore$ el canal es simétrico $\Rightarrow$ para calcular capacidad, basta con calcular la I mutua para $p(x) = \frac{1}{M}$,

$$I(X;Y) = H(Y) - H(Y \mid X) = \log(N) - \log(2) = \log\left(\frac{N}{2}\right)$$

el término $\log(N)$ sale de que las salidas son equiprobables, y el $\log(\frac{1}{2})$ es $H(Y \mid X)$

VII. CLASE 7:

\# **Conjunto típico:** El conjunto $\tilde{\mathcal{A}}^{(n)}_{(X;Y)\varepsilon}$ de secuencias $(x^n;y^n) = (x_1,\ldots,x_n;y_1,\ldots,y_n)$ conjuntamente típicas es

$$\tilde{\mathcal{A}}^{(n)}_{(X;Y)\varepsilon} = \{(x^n;y^n) \in \mathcal{A}_X \times \mathcal{A}_Y \; / x^n \in \mathcal{A}^{(n)}_{X\varepsilon} \wedge y^n \in \mathcal{A}^{(n)}_{Y\varepsilon} \wedge (x^n;y^n) \in \mathcal{A}^{(n)}_{(X;Y)\varepsilon}\}$$

Esta definición incluye tres condiciones:

- $x^n \in \mathcal{A}^{(n)}_{X\varepsilon}$, $\Rightarrow x^n$ es típica en $\mathcal{A}^n_X \Leftrightarrow$

$$\left|-\frac{1}{n}\log\left[p\left(x_1,\ldots,x_n\right)\right] - H(X)\right| \leq \varepsilon$$

- $y^n \in \mathcal{A}^{(n)}_{Y\varepsilon}$, $\Rightarrow y^n$ es típica en $\mathcal{A}^n_Y \Leftrightarrow$

$$\left|-\frac{1}{n}\log\left[p\left(y_1,\ldots,y_n\right)\right] - H(Y)\right| \leq \varepsilon$$

- $(x^n;y^n) \in \mathcal{A}^{(n)}_{(X;Y)\varepsilon}$, $\Rightarrow (x^n;y^n)$ es típica en $(\mathcal{A}_X \times \mathcal{A}_Y)^n \Leftrightarrow$

$$\left|-\frac{1}{n}\log\left[p\left(x_1,\ldots,x_n,y_1,\ldots,y_n\right)\right] - H(X,Y)\right| \leq \varepsilon$$

$p\left(x^n,y^n\right) = p\left(x_1,\ldots,x_n,y_1,\ldots,y_n\right)$: prob. conjunta.
$p\left(x^n\right) = p\left(x_1,\ldots,x_n\right)$: prob. marginal de las x
$p\left(y^n\right) = p\left(y_1,\ldots,y_n\right)$: prob. marginal de las y

$$p\left(x^n\right) = \int \mathrm{d}y_1 \cdots \int \mathrm{d}y_n p\left(x^n,y^n\right)$$

$$p\left(y^n\right) = \int \mathrm{d}x_1 \cdots \int \mathrm{d}x_n p\left(x^n,y^n\right)$$

\# **Canal sin memoria:**

$$p\left(x^n;y^n\right) = \prod_{i=1}^{n} p\left(x_i,y_i\right)$$

$\sim$ **Condición canal sin memoria:** Trabajaremos siempre bajo la suposición de que el canal no tiene memoria

> **Propiedades de una cadena conjuntamente típica:**

> **Probabilidad de muestrear una cadena conjuntamente típica:**

$$\text{Prob}\left[(x^n;y^n) \in \tilde{\mathcal{A}}^{(n)}_{(X;Y)\varepsilon}\right] \to 1, \text{ cuando } n \to +\infty$$

> **Número de cadenas conjuntamente típicas:**

$$\underbrace{(1-\varepsilon)2^{n[H(X,Y)-\varepsilon]}}_{\text{si } n\to\infty} < \left|\tilde{\mathcal{A}}^{(n)}_{(X,Y)\varepsilon}\right| \leq 2^{n[H(X,Y)+\varepsilon]}$$

> **Probabilidad de muestrear por casualidad secuencias conjuntamente típicas:** Si x^n e y^n son secuencias de longitud n muestreadas de una distr. de prob. $q\left(x^n,y^n\right) = p\left(x^n\right)p\left(y^n\right)$, $\Rightarrow$

$$\underbrace{(1-\varepsilon)2^{-n[I(X;Y)+3\varepsilon]}}_{n \text{ suficientemente grande}} < \Pr\left[(x^n;y^n) \in \tilde{\mathcal{A}}^{(n)}_{(X,Y)}\right] \leq 2^{-n[I(X;Y)-3\epsilon]}$$

$\sim$ Esto $\Rightarrow$ si, en vez de muestrear tiras $(x^n;y^n)$ usando prob. $p\left(x^n,y^n\right)$, las muestreamos con $p\left(x^n\right)p\left(y^n\right)$, es poco probable que $(x^n;y^n) \in \tilde{\mathcal{A}}^{(n)}_{(X,Y)}$ $(\approx 2^{-nI(X;Y)})$.

\# **Código:** Un código $(\boldsymbol{M},\boldsymbol{n})$ para un canal discreto $[\mathcal{A}_X,\mathcal{A}_Y,p(y \mid x)]$ es (tres cosas):

- Un conjunto de índices $\{1,2,\ldots,M\}$,
- Función codificadora $f : \{1,\ldots,M\} \to \mathcal{A}^n_X$, que a cada mensaje $w \in \{1,\ldots,M\}$ le asigna una cadena de caracteres $f(w) \in \mathcal{A}^n_X$. El conjunto de palabras clave $\{f(1),f(2),\ldots,f(M)\}$ se llama libro de códigos,
- Función decodificadora $g : \mathcal{A}^n_Y \to \{1,\ldots,M\}$.

\# **Prob. error condicional** λ_i**:** Asociada al mensaje i es la prob. de no decodificar i cuando se transmite i. Es decir,

$$\lambda_i = \text{Prob}\left[g\left(y^n\right) \neq i / x^n = f(i)\right] = \sum_{y^n / g(y^n)\neq i} \underbrace{\text{Prob}\left[y^n \mid f(i)\right]}_{\substack{\text{Probabilidad de recibir } y^n \\ \text{cuando se transmite } x^n = f(i)}}$$

Máxima probabilidad de error:

$$\text{del código } (M,n) \text{ es } \lambda_{\max} = \max_i \lambda_i$$

Probabilidad de error media del código:

$$\bar{\lambda} = \frac{1}{M}\sum_{i=1}^{M} \lambda_i$$

Tasa: La tasa R de un código (M,n) es

$$R = \frac{\log(M)}{n}$$

Tasa realizable: Existe una sucesión de códigos $\left(2^{nR}, n\right)$ cuya máxima prob. de error λ_{mix} tiende a cero cuando $n \to \infty$.

> **Teorema de codificación de un canal:** En un canal con capacidad C, todas las tasas $R < C$ son realizables. Recíprocamente, toda secuencia de códigos $\left(2^{nR}, n\right)$ para la cual $\lambda_{\max} \to 0$ cuando $n \to \infty$ cumple que $R < C$.

VIII. CLASE 8:

En esta sección se asume:

- El canal es sin memoria
- La entrada es una secuencia ∞ de símbolos que $\in$ un alfabeto, con muestreo equiprobable e independiente.

> **Aritmética modular:**
Cuando $D = 2$:

- El resultado de $\bar{a} \oplus_2 \bar{1}$ da $\bar{0}$ si $\bar{a} = \bar{1}$ y da $\bar{1}$ si $\bar{a} = \bar{0}$ (es decir, al sumar un uno, se invierte el dígito),
- $\forall \bar{a} \odot_2 \overline{par} = \bar{0}$
- $\forall a,\ \overline{\pm 1} \odot_2 \bar{a} = \bar{a}$.
- Restar dos números es lo mismo que sumarlos.

- **Decodificación MAP(Max. a posteriori)**

$$\xrightarrow{i} \boxed{\begin{matrix}\text{Codificador}\\ f\end{matrix}} \xrightarrow{X} \boxed{\text{Canal}} \xrightarrow{Y} \boxed{\begin{matrix}\text{Decodificador}\\ g\end{matrix}} \xrightarrow{i}$$

Buscamos una función $g : \{0,1\}^n \to \{1,2,\ldots,M\}$ que maximice la prob. de acierto. Esta prob. es un promedio sobre las palabras claves x^i transmitidas por el canal, de la prob. de ir a caer a una cadena y que es correctamente decodificado como el mensaje i, es decir,

$$\text{Prob(acierto)} = \sum_{\boldsymbol{x}^i} \text{Prob}\left(\boldsymbol{x}^i\right) \text{Prob}\left[g(\boldsymbol{y}) = i \mid \boldsymbol{x}^i\right]$$

A su vez, Prob $\left[g(\boldsymbol{y}) = i \mid \boldsymbol{x}^i\right]$ es la prob. de que, cuando se transmite $\boldsymbol{x}^i$, el canal autputee un dado vector y que pertenece a la preimagen de i por la función g, es decir,

$$\text{Prob}\left[g(\boldsymbol{y}) = i \mid \boldsymbol{x}^i\right] = \sum_{\boldsymbol{y}:g(\boldsymbol{y})=i} \text{Prob}\left[\boldsymbol{y} \mid \boldsymbol{x}^i\right]$$

donde la suma barre sobre todos los y que, cuando se les aplica la función g, van a parar al verdadero índice i de la cadena transmitida. A su vez, la prob. condicionada de ir a caer a un dado y cuando se transmite un dado x^i vale

$$\text{Prob}\left[\boldsymbol{y} \mid \boldsymbol{x}^i\right] = (1-f)^{k_{\text{bien}}} f^{k_{\text{mal}}}$$

> **Probabilidad de acierto:**

Por digito: $1 - f = p(y=0|x=0) = p(y=1|x=1)$

Por codificación:

$$\text{Prob(acierto)} = \sum_{\boldsymbol{x}^i} \text{Prob}\left(\boldsymbol{x}^i\right) \sum_{y:g(\boldsymbol{y})=i} (1-f)^{n-d\left(\boldsymbol{x}^i,y\right)} f^{d\left(\boldsymbol{x}^i,\boldsymbol{y}\right)}$$

Mensajes equiprobables $\Rightarrow$ Prob $\left(x^i\right)$ uniforme. $\therefore$

$$\text{Prob(acierto)} = \frac{1}{M}\sum_i \sum_{y:g(y)=i} (1-f)^{n-d\left(\boldsymbol{x}^i,y\right)} f^{d\left(\boldsymbol{x}^i,\boldsymbol{y}\right)}$$

La g que maximiza la prob. de acierto es aquella que hace que $d\left(\boldsymbol{x}^i, \boldsymbol{y}\right)$ sea mínimo.

Distancia Hamming: Entre dos vectores z^1 y z^2 La Distancia de Hamming: $d\left(z^1, z^2\right) = \#$ dígitos donde z_i^1 y z_i^2 difieran

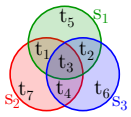
$$d(z^1, z^2) = \sum_{n=1}^{N} \left|z_n^1 - z_n^2\right|$$

> **Cota a distancia hamming:** Sean $x^1, \ldots, x^M$ palabras clave de longitud n que, $e \in \mathbb{Z}^+$, cumplen:

$d\left(\boldsymbol{x}^i, \boldsymbol{x}^j\right) \geq 2e+1,\ i \neq j,\ \therefore$ errores corregibles hasta e

$d\left(\boldsymbol{x}^i, \boldsymbol{x}^j\right) \geq 2e,\ i \neq j,\ \therefore$ $\begin{matrix}\text{errores corregibles hasta } e-1\\ \text{detectables hasta } e\end{matrix}$

> **Corregir errores:**



$(s1, s2, s3) = 011 \Rightarrow$ error en $t4$

$\sim$ Recíprocamente si conozco e la desigualdad me indica como son las distancias Hamming.

> **Cota de Hamming(Cond. nec.):** Si un código de M palabras clave de longitud n corrige errores hasta orden $e \Rightarrow$

$$M \leq \frac{2^n}{\sum_{i=0}^{e}\binom{n}{i}}$$ M:#palabras clave, e:orden errores; n:#dígitos(long. de palabra clave(dim. espacio))

Código de Hamming: Detecta y corrige hasta errores simples.

$$\text{Hamming}(a,b), \begin{matrix}a\text{: \# dígitos}\\ b\text{: \# bits de mensaje}\end{matrix} \qquad a-b\text{: \# bits de paridad}$$

Síndrome o corrector: Con un mensaje transmitido t, un error e y matriz de paridad A, el síndrome s:

$$s = Ay = A(t+e) = Ae, \quad \text{size}(A) = (a-b)\times a$$

Palabras clave:

$$\mathbf{t} : A\mathbf{t} = \mathbf{0},\ \text{Hamming}(a,b) \Rightarrow 2^b \text{palabras clave}$$

Caso Hamming(7,4):

$$\boldsymbol{t}^i = \begin{pmatrix} t_1^i \\ t_2^i \\ t_3^i \\ t_4^i \\ t_5^i \\ t_6^i \\ t_7^i \end{pmatrix}, \quad \begin{array}{l} t_1^i = x_1^i, \quad t_2^i = x_2^i \\ t_3^i = x_3^i, \quad t_4^i = x_4^i \\ t_5^i = x_1^i \oplus x_2^i \oplus x_3^i, \\ t_6^i = x_2^i \oplus x_3^i \oplus x_4^i \\ t_7^i = x_1^i \oplus x_3^i \oplus x_4^i \end{array} \quad \begin{array}{c} \vec{t^i} : i\text{-ésimo mensaje} \\ \text{transmitido,} \\ a \times 1 (1^o s\ b \text{ dígitos} \\ \text{son el mensaje)} \\ \oplus : \text{suma xor} (\oplus_2) \end{array}$$

> **Tabla de decodificación:** Para Hamming(a,b), $r = a-b$.

Síndrome	Error
$\underbrace{0\ldots0}_{b}$	$\underbrace{0\ldots0}_{a}$
A^T	I_a
síndromes restantes	respectivos errores

*Ej. :

Síndrome	Error
000	00000
101	10000
010	01000
011	00100
111	00010
100	00001
110	01001
001	01100

$\sim$ Arrancas con los **s** y buscas sus respectivos **e**

> **Corregir errores:** La matriz de chequeo de paridad A corrige errores de orden hasta $e \Leftrightarrow$ todo conjunto de $2e$ columnas de A es linealmente independiente.

> **Dígitos de chequeo de paridad(cond. suf.):** El número ℓ de dígitos de chequeo de paridad debe cumplir

$$2^\ell > \sum_{i=0}^{2e-1} \binom{n-1}{i}$$

> **Construir matriz chequeo de paridad:** Conociendo las dimensiones. Creo tantas columnas L.I. como bits de paridad. Luego creo columnas restantes como suma de las columnas anteriores

IX. CLASE 9:

Entropía diferencial conjunta: La entropía diferencial conjunta asociada a una densidad conjunta $\rho(x_1,\ldots,x_n)$:

$$h(X_1,\ldots,X_n) = -\int \rho(x_1,\ldots,x_n)\log\left[\rho(x_1,\ldots,x_n)\right] \mathrm{d}x_1\ldots\mathrm{d}x_n$$

$$h(\boldsymbol{X}) = -\int \rho(\boldsymbol{x})\log[\rho(\boldsymbol{x})]\mathrm{d}\boldsymbol{x},\ \boldsymbol{x} = (x_1,\ldots,x_n)$$

> **Change variable:** $y = g(x)$

$$\begin{aligned} f_Y(y) &= \frac{dF_Y(y)}{dy} = f_X\left(g^{-1}(y)\right)\left|\frac{dg^{-1}(y)}{dy}\right| = f_X(x)\left|\frac{dx}{dy}\right| \\ h(Y) &= -\int f_Y(y)\log f_Y(y)dy \\ &= -\int f_Y(y)\log f_X\left(g^{-1}(y)\right)dy - \int f_Y(y)\log\left|\frac{dx}{dy}\right|dy \\ &= h(X) - \mathrm{E}\left[\log\left|\frac{dx}{dy}\right|\right] \end{aligned}$$

Entropía diferencial condicionada: La entropía diferencial condicionada de un conjunto de variables $\boldsymbol{X} = (X_1,\ldots,X_n)$ condicionadas a los valores que toma otro grupo de variables $\boldsymbol{Y} = (Y_1,\ldots,Y_m)$ vale

$$h(\boldsymbol{X}\mid\boldsymbol{Y}) = -\int \rho(\boldsymbol{x},\boldsymbol{y})\log[\rho(\boldsymbol{x}\mid\boldsymbol{y})]\mathrm{d}\boldsymbol{x}\mathrm{d}\boldsymbol{y}$$

$$\begin{aligned} > \quad h(\boldsymbol{X},\boldsymbol{Y}) &= h(\boldsymbol{X}) + h(\boldsymbol{Y}\mid\boldsymbol{X}) \\ &= h(\boldsymbol{Y}) + h(\boldsymbol{X}\mid\boldsymbol{Y}) \end{aligned}$$

Kullback-Leibler: Dadas dos densidades de prob. conjuntas $\rho_1(\boldsymbol{x})$ y $\rho_2(\boldsymbol{x})$ definidas sobre una variable aleatoria multivariada $\boldsymbol{X} = (X_1,\ldots,X_n)$, la divergencia de Kullback-Leibler entre ellas es

$$D_{\mathrm{KL}}(\rho_1\|\rho_2) = \int \rho_1(\boldsymbol{x})\log\left[\frac{\rho_1(\boldsymbol{x})}{\rho_2(\boldsymbol{x})}\right]\mathrm{d}\boldsymbol{x}$$

Información mutua: Dadas las variables aleatorias continuas multivariadas $\boldsymbol{X}$ e $\boldsymbol{Y}$ con densidades conjuntas $\rho(\boldsymbol{x},\boldsymbol{y})$, la info. mutua $I(\boldsymbol{X};\boldsymbol{Y})$ vale

$$\begin{aligned} I(\boldsymbol{X};\boldsymbol{Y}) &= D_{\mathrm{KL}}[\rho(\boldsymbol{x},\boldsymbol{y})\|\rho(\boldsymbol{x})\rho(\boldsymbol{y})] \\ &= h(\boldsymbol{x}) - h(\boldsymbol{x}\mid\boldsymbol{y}) \\ &= h(\boldsymbol{y}) - h(\boldsymbol{y}\mid\boldsymbol{x}) \\ &= h(\boldsymbol{x}) + h(\boldsymbol{y}) - h(\boldsymbol{x},\boldsymbol{y}) \end{aligned}$$

- La I no diverge
- I y D_{KL} pueden seguir interpretándose como el # de preguntas ahorradas (I) o el # de preguntas extras (D_{KL}).
- $0 \le I$ y $0 \le D_{KL}$ no están acotadas superiormente.
- No hay problema en calcular la I entre una variable continua y otra discreta.
- La D_{KL} y I, permanecen inalteradas ante cambios de variables inyectivos del tipo $X' = f(X), Y' = g(Y)$.

$$\bullet p(\mu) = \begin{cases} \frac{1}{\mu_0}, \mu \in \left[\mu_0 - \frac{\mu_0}{2}, \mu_0 + \frac{\mu_0}{2}\right] \\ 0, \text{c.o.c} \end{cases}$$

$$h(\mu) = \log(\mu_0)$$

$$\bullet p(\mu) = \frac{\mathrm{e}^{-\mu/\mu_0}}{\mu_0}$$

$$h(\mu) = 1 + \log(\mu_0)$$

$$\bullet p(\mu) = \frac{2}{\pi\mu_0}\mathrm{e}^{-\mu^2/\pi\mu_0^2}$$

$$h(\mu) = \frac{1}{2} + \log(\pi/2) + \log(\mu_0)$$

$$\bullet p(\mu) = \frac{k^k}{\Gamma(k)\mu_0^k}\mu^{k-1}\mathrm{e}^{-k\mu/\mu_0}$$

$$h(\mu) = k + \log(\mu_0/k) + \log[\Gamma(k)] + (1-k)\psi_0(k)$$

X. CLASE 10:

- **Técnica Maxentropy**
  - Listar todas las restricciones que queremos tener en cuenta.
  - Construir el funcional $\mathcal{F}[P]$ del problema.
  - Maximizar el funcional.

$\mathcal{F}[P(x)]$ es un funcional que depende de una función P, que a su vez depende de la variable continua x. La dependencia es a través de una ecuación integral

$$\mathcal{F}[P(x)] = -\int_{-\infty}^{+\infty} P(x)\log[P(x)]\mathrm{d}x + \sum_{j=1}^{n}\lambda_j\left[\int_{-\infty}^{+\infty} P(x)f(x)\mathrm{d}x - \alpha_j\right]$$

el primer término es $h(x)$, y la $\sum$ contiene todas las restricciones, cada una con su multip. de Lagrange λ_j y su valor fijado α_j. Una de las restricciones (digamos la i-ésima) es de normalización($f_i(x) = 1$, $\alpha_i = 1$).

$$\begin{cases} 0 = \frac{\partial\mathcal{F}[P(\mu)]}{\partial P(x^*)} = -1 - \log[P(x^*)] + \sum_{j=1}^{n}\lambda_j f_j(x^\star) \\ 0 = \int_{-\infty}^{+\infty} P(x)f_1(x)\mathrm{d}x - \alpha_1 \\ \vdots \\ 0 = \int_{-\infty}^{+\infty} P(x)f_n(x)\mathrm{d}x - \alpha_n \end{cases}$$

∼ Vemos que la derivada funcional puede calcularse derivando a lo bestia, considerando que $P(x_1), P(x_2)$, etc., son las variables independientes de la función $\mathcal{F}[P(x)]$ e interpretando las integrales como sumas.

∼ Cuando todas las restricciones fijan el valor medio(μ) de determinadas funciones de la variable aleatoria (lo que equivale a decir que las restricciones son lineales en $P(\mu)$), Maxentropy arroja siempre exponencial en las funciones cuyo μ está fijado

∼ Se puede aplicar este mismo proceso en el discreto

∼ La distr. uniforme maximiza la $h(x)$ sin restricciones (más allá de la normalización).

∼ En variables $\in \mathbb{R}_0$, la distr. exponencial es la que maximiza la $h(x)$ cuando se fija el μ.

∼ La distr. Gaussiana maximiza $h(x)$ cuando se fija la varianza(es irrelevante si se fija o no la media). Esta propiedad (junto al Teorema Central del Lim., cuando se puede suponer que su distr. proviene de la suma de diversos factores descorrelacionados), favorece a la distr. Gaussiana en procesos de los cuales solo conocemos el tamaño típico de las fluctuaciones.

• **Canal gaussiano:** Sean Y_i, X_i, Z_i variables aleatorias. Y_i la salida, X_i la entrada, Z_i un ruido gaussiano y muestreado indepen. del tiempo y de los X_j.

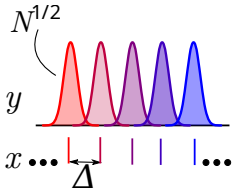
$$Y_i = X_i + Z_i,\ \rho(z) = \frac{e^{-z^2/2N}}{\sqrt{2\pi N}}, \rho(z) : \text{Dens. de prob. de } z$$

La magnitud N de la varianza está fijada por el canal. Por cada valor X_i que tengamos a la entrada, vamos a tener una Gaussiana de media X_i y varianza N:

$$\rho(y_i \mid x_i) = \frac{e^{-(y-x_i)^2/2N}}{\sqrt{2\pi N}}$$

Por esto, la capacidad del canal puede crecer con solo aumentar la separación entre símbolos de entrada.

∗ Imaginemos que trabajamos con un alfabeto $\mathcal{A}_X = \{\ldots, -2\Delta, -\Delta, 0, \Delta, 2\Delta, \ldots\}$.



La salida podemos obtener cualquier valor de y. Sin embargo, la superposición entre Gaussianas puede achicarse tanto como queramos, separando las entradas.

Cuando $\Delta \gg \sqrt{N}$ podemos decodificar la entrada con mucha precisión. Si elegimos los X_i de entrada con distr. uniforme, la info. transmitida $\to \infty$ cuando $\Delta \to \infty$. la solución con capacidad ($C \to \infty$), presupone que podemos usar 4 señales de entrada arbitrariamente grandes, no es posbile. Se trata de maximizar la capacidad restringiendo la potencia (es decir, la varianza) de las X a la entrada.

> **Estrategia 1:** en vez de ∞ entradas discretas, tan solo un número finito, por ejemplo 2. Para separar máximamente las entradas y aún cumplir con la restricción sobre la varianza, proponemos que X pueda valer solo $\pm\sqrt{P}$, es decir.

$$\rho(x) = \frac{1}{2}\delta(x - \sqrt{P}) + \frac{1}{2}\delta(x + \sqrt{P})$$

Esta es una posible estrategia. De hecho, es la mejor estrategia que podemos tomar si nuestro objetivo es minimizar el error de decodificación. Sin embargo, se paga el costo de reducir drásticamente la entropía $H(X)$: del continuo de valores posibles que inicialmente teníamos para X, elegimos usar solo dos. Agregar más entradas aumenta $H(X)$, pero también $H(X \mid Y)$.

> **Estrategia 2:** Conviene maximizar $I(X;Y)$, que es lo que hacemos a continuación. Es decir, buscamos una distr. de entrada $\rho(x)$-sin pérdida de generalidad, podemos suponer que tiene media nula-que maximice la info. transmitida por el canal, sujeta a la restricción

$$\langle X^2 \rangle = \int_{-\infty}^{+\infty} \rho(x) x^2 \, \mathrm{d}x \leq P$$

para una constante conocida P, que representa la potencia de entrada que estamos dispuestos a costear. La I entre la entrada y la salida vale

$$\begin{aligned} I(X;Y) &= h(Y) - h(Y \mid X) = h(Y) - h(X+Z \mid X) \\ &= h(Y) - h(Z \mid X) = h(Y) - h(Z) \\ &= h(Y) - \frac{1}{2}\ln(2\pi e N). \end{aligned}$$

$\sim$ $h(X+Z \mid X) = h(Z \mid X)$, por ser X fijo.

$\sim$ $h(Z \mid X) = h(Z)$ porque X, Z son independientes.

$\sim$ **Entropía diferencial de una gaussiana:** $h(Z) = \ln(2\pi e N)/2$

Como $h(Z)$ está fijada por el canal, maximizar la I sujeta a potencia de entrada P a maximizar la $h(Y)$ sujeta a potencia P.

$$\begin{aligned} \langle Y^2 \rangle = \langle (X+Z)^2 \rangle \quad &= \langle X^2 \rangle + 2\langle XZ \rangle + \langle Z^2 \rangle \\ = \langle X^2 \rangle + \langle Z^2 \rangle \quad &= \langle X^2 \rangle + N \end{aligned}$$

asi, la condición $\langle X^2 \rangle \leq P \Rightarrow \langle Y^2 \rangle \leq P+N$. Hay que encontrar una distr. de entrada $\rho(x)$ que maximice la $h(Y)$ sujeta a la restricción de la ecuación anterior. Usando Maxentropy, se tiene que la distr. $\rho(y)$ es:

$$\rho(y) = \frac{e^{-y^2/2(P+N)}}{\sqrt{2\pi(P+N)}}$$

Por la forma de $\rho(y)$:

$$\rho(x) = \frac{e^{-x^2/2P}}{\sqrt{2\pi P}}, \quad \rho(y \mid x) = \frac{e^{-(y-x)^2/2N}}{\sqrt{2\pi N}}$$

$\rho(x)$ muestrea valores cerca de $0 \Rightarrow$ habrá confusiones a la salida. Pero si la varianza de entrada está fija, al elegir una Gaussiana es más lo que ganamos en $h(X)$ que lo que perdemos en $h(X \mid Y)$.

Con estas densidades I es máxima $\therefore C = I$:

$$\begin{aligned} C = h(Y) - h(Y \mid X) &= \frac{1}{2}\ln[2\pi e(P+N)] - \frac{1}{2}\ln[2\pi e N] \\ &= \frac{1}{2}\ln\left(1 + \frac{P}{N}\right), \quad \left(\frac{P}{N} \text{ es un cociente } \frac{\text{señal}}{\text{ruido}}\right) \end{aligned}$$

XI. CLASE 11:

Estimadores: Dada una variable aleatoria X de alfabeto $\mathcal{A}_X$ con distr. de prob. $P(x \mid \vartheta)$ que depende de un parámetro $\vartheta \in \Theta$ y un conjunto de muestras independientes $X_1, \ldots, X_M$ un estimador de ϑ es una función $\hat{\vartheta} : \mathcal{A}_X^M \to \Theta$.

Estimador consistente: Un estimador $\hat{\vartheta}$ es consistente si $n \to \infty, \hat{\vartheta}(X_1, \ldots, X_n) \to \vartheta$ (limite variables aleatorias).

Sesgo: El sesgo $B(\vartheta)$ de un estimador es el error promedio que produce, donde el promedio se calcula para ϑ fijo. Es decir,

$$B(\vartheta) = \langle \hat{\vartheta} \rangle - \vartheta = \int \mathrm{d}^n x p(x_1, \ldots, x_n \mid \vartheta)\, \hat{\vartheta}(x_1, \ldots, x_n) - \vartheta$$

Un estimador es no sesgado si $B(\vartheta) = 0, \forall \vartheta$

$\sim$ Como vemos la notación $bias(\hat{\theta}) \neq B(\vartheta)$. Pero se refieren a lo mismo

Error cuadrático medio(MSE) estimador:

$$\mathbb{E}\left[(\hat{\theta} - \theta)^2\right]$$

> **Error MSE:**

$$\begin{aligned} \mathbb{E}\left[(\hat{\theta} - \theta)^2\right] &= \mathbb{E}\left[\left(\hat{\theta} - \mathbb{E}\hat{\theta} + \mathbb{E}\hat{\theta} - \theta\right)^2\right] \\ &= \mathbb{E}\left[\left(\hat{\theta} - \mathbb{E}\hat{\theta}\right)^2\right] + \left[\mathbb{E}(\hat{\theta}) - \theta\right]^2 \\ &+ 2\left[\mathbb{E}(\hat{\theta}) - \theta\right]\mathbb{E}\left[\hat{\theta} - \mathbb{E}\hat{\theta}\right] \\ &= \mathrm{var}_\theta(\hat{\theta}) + \mathrm{bias}^2(\hat{\theta}), \quad \mathrm{bias}(\hat{\theta}) = \mathbb{E}(\hat{\theta}) - \theta \end{aligned}$$

Información de Fisher: Denotada $J(\vartheta)$ asociada a una distr. de prob. $p(x \mid \vartheta)$ es la varianza del puntaje. Es decir,

$$\begin{aligned} J(\vartheta) &= \langle (V - \langle V \rangle)^2 \rangle = \langle V^2 \rangle \\ &= \int_{-\infty}^{+\infty} p(x \mid \vartheta) \left[\frac{\partial}{\partial \vartheta} \ln p(x \mid \vartheta)\right]^2 \mathrm{d}x \end{aligned}$$

$$J(\vartheta) = \left\langle \left(\frac{\partial}{\partial \vartheta} \ln P(x_1, \ldots, x_n \mid \vartheta)\right)^2 \right\rangle$$

y los corchetes $\langle\rangle$ denotan promedio pesado con la distr. $P(x_1, \ldots, x_n \mid \vartheta)$.

> **Forma alternativa:**

$$\begin{aligned} J(\vartheta) &= -\left\langle \frac{\partial^2}{\partial \vartheta^2} \ln[p(x \mid \vartheta)] \right\rangle \\ &= -\int_{-\infty}^{+\infty} p(x \mid \vartheta) \frac{\partial^2}{\partial \vartheta^2} \ln p(x \mid \vartheta) \mathrm{d}x \end{aligned}$$

Error cuadrático medio:

$$E^2(\vartheta) = E[\hat{\vartheta} - \vartheta] = \int_{-\infty}^{\infty} p(x_1, \ldots, x_n \mid \vartheta) \left[\hat{\vartheta}(x_1, \ldots, x_n) - \vartheta\right]^2 d^n x$$

Cota de Crámer-Rao:

$$E^2(\vartheta)J(\vartheta) \geq 1$$
$$\left\{\int_{-\infty}^{+\infty} p(x \mid \vec{\vartheta})[\hat{\vartheta}(x) - \vartheta]^2 \, \mathrm{d}x\right\} \cdot$$
$$\left\{\int_{-\infty}^{+\infty} p(x \mid \vartheta)\left[\frac{\partial}{\partial\vartheta}\ln p(x \mid \vartheta)\right]^2 \, \mathrm{d}x\right\} \geq 1$$

Eficiente: Un estimador $\hat{\theta}$ es eficiente si su MSE es pequeño, (según la Cota de Crámer Rao):

$$E^2(\vartheta)J(\vartheta) \geq 1,$$

Puntaje: El puntaje de una variable aleatoria X muestreada con prob. $p(x \mid \vartheta)$ es una nueva variable aleatoria V que se obtiene de transformar X con la función

$$V = \frac{\partial}{\partial\vartheta}\ln[p(X \mid \vartheta)]$$

$\sim$ $p(x \mid \vartheta)$ debe ser derivable (y $\therefore$, continua) en ϑ, salvo un número finito de puntos, pero, puede no ser continua en x. X puede ser una variable discreta, $p(x \mid \vartheta)$ puede solo estar definida en un número finito de $x's$, una suma de δ.

> **Valor medio del puntaje:** $\langle V\rangle = 0$

> **No negatividad:** $J(\vartheta) \geq 0$

> **Aditividad:** Si $(X_1, \ldots, X_n)$ son n muestras indepen. de $p(x \mid \vartheta) \Rightarrow J_n(\vartheta)$:

$$J_n(\vartheta) = nJ_1(\vartheta), \begin{matrix} J_n \text{usa } p\left(x_1, \ldots, x_n \mid \vartheta\right) \\ J_1 \text{usa } p\left(x_1, \ldots, x_n \mid \vartheta\right)\end{matrix}$$

> **Cambio de variable:**

$$I^\theta(\theta) = I^\eta(\eta)\left(J_\theta^\eta\right)^2$$
$$I^\theta(\theta) = I^\eta(\eta)\left(\frac{\partial\eta}{\partial\theta}\right)^2 \text{(1D)}$$

∗ **Binomial:** ϑ es la prob. de éxito:

$$P(k \mid \vartheta) = \frac{N!}{k!(N-k)!}\vartheta^k(1-\vartheta)^{N-k}, N \in \mathbb{N}^{\#\text{ of trials}}_{(parameter)}$$
$$, k \in \{0, 1, \ldots, n\}, ^{\#\text{ of successes}}_{(variable)}$$

$$\mathrm{E}(X) = \sum_{k=0}^{n} k\binom{n}{k}p^k q^{n-k} = \sum_{k=1}^{n} k\binom{n}{k}p^k q^{n-k}$$
$$= \sum_{k=1}^{n} n\binom{n-1}{k-1}p^k q^{n-k} \quad \left(k\binom{n}{k} = n\binom{n-1}{k-1}\right)$$
$$= np\sum_{k=1}^{n}\binom{n-1}{k-1}p^{k-1}q^{(n-1)-(k-1)}$$
$$\text{(sacar } np \text{ y } (n-1)-(k-1) = n-k)$$
$$= np\sum_{j=0}^{m}\binom{m}{j}p^j q^{m-j} \quad \left(\text{putting } {}^{m=n-1}_{j=k-1}\right)$$
$$= np\,(\text{ Binomial Theorem and } p+q=1)$$

El puntaje vale

$$V(K) = \frac{\partial}{\partial\theta}\ln[P(K \mid v)]$$
$$= \frac{\partial}{\partial\theta}\{\ln(N!) - \ln(K!) - \ln[(N-K)!]$$
$$+ k\ln(\vartheta) + (N-K)\ln(1-\vartheta)\}$$
$$= \frac{K}{\vartheta} - \frac{N-K}{1-\vartheta} = \frac{K - N\vartheta}{\partial(1-\vartheta)}$$

Dado que $\langle K\rangle = N\vartheta$, en este ejemplo, V es proporcional a $K - \langle K\rangle$. Por esto el valor medio $= 0$.

$$\text{Informacion fisher: } J(\vartheta) = \frac{N}{\vartheta(1-\vartheta)}$$

∗ **Exponencial:**

$$^{\text{parametrizada con un}}_{\text{tiempo de vida medio } \vartheta_1} : p\left(x \mid \partial_1\right) = \frac{e^{-x/\theta_1}}{\vartheta_1}$$

El puntaje vale

$$V = \frac{\partial}{\partial\vartheta_1}\ln\left[p\left(X \mid \vartheta_1\right)\right] = \frac{\partial}{\partial\vartheta_1}\left[-\frac{X}{\partial_1} - \ln\left(\partial_1\right)\right]$$
$$= \frac{X}{\vartheta_1^2} - \frac{1}{\vartheta_1} = \frac{X - \vartheta_1}{\vartheta_1^2}$$

Nuevamente obtuvimos $V \propto X - \langle X\rangle$. Por esto el valor medio se anula.

$$\text{Información Fisher: } J\left(\vartheta_1\right) = \frac{1}{\vartheta_1^2}$$

∗ **Exponencial:**

$$\text{Con una tasa media } \vartheta_2 : p\left(x \mid \vartheta_2\right) = \vartheta_2 \mathrm{e}^{-\theta_2 x}$$

El puntaje vale

$$V = \frac{\partial}{\partial\vartheta_2}\ln\left[p\left(X \mid \vartheta_2\right)\right] = \frac{\partial}{\partial\vartheta_2}\left[-\vartheta_2 X + \ln\left(\vartheta_2\right)\right]$$
$$= -X + \frac{1}{\vartheta_2}$$

Una vez más, obtuvimos $V \propto X - \langle X\rangle$, con una constante de proporcionalidad negativa.

$$\text{Informacion fisher: } J\left(\vartheta_2\right) = \frac{1}{\vartheta_2^2}$$

La información de Fisher como expresión local de la divergencia de Kullback-Leibler:

$$D_{\mathrm{KL}}[p(x \mid \vartheta + \mathrm{d}\vartheta)\|p(x \mid \vartheta)] \approx \frac{(\mathrm{d}\vartheta)^2}{2}J(\vartheta)$$
$$D_{\mathrm{KL}}[p(x \mid \vartheta\|p(x \mid \vartheta + \mathrm{d}\vartheta))] \approx \frac{(\mathrm{d}\vartheta)^2}{2} \cdot J(\vartheta)$$

• **Repaso de tensores métricos:** Un tensor métrico M debe ser simétrico y definido positivo (todos sus autovalores positivos), define un producto escalar

$$\left\langle \boldsymbol{v}^\alpha, \boldsymbol{v}^b\right\rangle = \left(\boldsymbol{v}^a\right)^t M\boldsymbol{v}^b,$$

donde t representa "traspuesto". Las dist. y angulos

$$d\left(\boldsymbol{v}^{\alpha},\boldsymbol{v}^{b}\right)=\left|\boldsymbol{v}^{a}-\boldsymbol{v}^{6}\right|,\ |\boldsymbol{v}|=\sqrt{\langle\boldsymbol{v},\boldsymbol{v}\rangle}$$
$$\alpha\left(\boldsymbol{v}^{\alpha},\boldsymbol{v}^{b}\right)=\arccos\left(\frac{\left\langle\boldsymbol{v}^{a},\boldsymbol{v}^{b}\right\rangle}{|\boldsymbol{v}^{a}|\,|\mathbf{v}^{b}|}\right)$$

Información de Fisher, parámetro multidimensional: Si ϑ proviene de un espacio de parámetros Θ de dimensión n, la información de Fisher $J(\boldsymbol{\vartheta})$ es un tensor de rango 2 de dimensión $n\times n$ cuyas componentes valen

$$\begin{aligned}J(\boldsymbol{\vartheta})_{k\ell}&=\left\langle\frac{\partial\ln[p(x\mid\boldsymbol{\vartheta})]}{\partial\boldsymbol{\vartheta}_{k}}\frac{\partial\ln[p(x\mid\boldsymbol{\vartheta})]}{\partial\vartheta_{k}}\right\rangle\\&=-\left\langle\frac{\partial^{2}\ln[p(x\mid\boldsymbol{\vartheta})]}{\partial\boldsymbol{\vartheta}_{k}\partial\boldsymbol{\vartheta}_{k}}\right\rangle\end{aligned}$$

donde los valores medios se calculan pesados con la distr. $p(x\mid\vartheta)$. Para el continuo:

$$\begin{aligned}J_{k\ell}&=\int p(x\mid\vartheta)\left[\frac{\partial\ln[p(x\mid\vartheta)]}{\partial v_{k}}\right]\left[\frac{\partial\ln[p(x\mid\vartheta)]}{\partial v_{\ell}}\right]\mathrm{d}x\\&=-\int p(x\mid\boldsymbol{\vartheta})\left[\frac{\partial^{2}\ln[p(x\mid\boldsymbol{v})]}{\partial\vartheta_{k}\partial\vartheta_{\ell}}\right]\mathrm{d}x\end{aligned}$$

y con variables discretas, las integrales se reemplazan por sumas.

> **Propiedades del tensor $J(\vartheta)$:**

- Es simétrico, por definición.
- Es definido no negativo (autovalores no negativos)

Cota de Crámer-Rao(Tensorial): se transforma en una ecuación matricial

$$E^{2}(\boldsymbol{\vartheta})J(\boldsymbol{\vartheta})\geq\mathbb{I}$$

donde el error cuadrático medio $E(\boldsymbol{\theta})$ es una matriz de $n\times n$ de componentes

$$E_{k\ell}^{2}=\left\langle\left(\hat{\vartheta}_{k}(X)-\vartheta_{k}\right)\left(\hat{\nu}_{\ell}(X)-\vartheta_{\ell}\right)\right\rangle,$$

y la desigualdad $EJ\geq\mathbb{I}$ significa que todos los autovalores de la matriz producto $E.J$ son ≥ 1.

La información de Fisher como expresión local de la divergencia de Kullback-Leibler(Tensorial):

$$D_{\mathrm{KL}}[p(x\mid\boldsymbol{\vartheta}+\mathrm{d}\boldsymbol{\vartheta})\|p(x\mid\boldsymbol{\theta})]\approx\frac{1}{2}\,\mathrm{d}\boldsymbol{\vartheta}^{t}J(\boldsymbol{\vartheta})\mathrm{d}\boldsymbol{\vartheta}$$

> **Transformación:** Si el parámetro ϑ es transformado en un nuevo parámetro $\varphi=F(\vartheta)$, $\Rightarrow$ la $J(\boldsymbol{v})$ transforma como

$$J(\boldsymbol{v})=C^{t}J(\varphi)C,\ C_{k\ell}=\frac{\partial F_{k}}{\partial\vartheta_{\ell}}\ \begin{array}{l}\text{es la matriz jacobiana}\\\text{de la transformación}\end{array}$$

> **Teorema de procesamiento de datos:** Si transformamos la variable aleatoria X en $Y=f(X)\Rightarrow$

$$J_{X}(\boldsymbol{\vartheta})\geq J_{Y}(\boldsymbol{\vartheta})$$

- **Tensor métrico:** $J_{ik}(\vartheta)$ Permite calcular dist. en el espacio de parámetros. Calculemos la dist. entre los parámetros $\boldsymbol{\theta}^{a}=(\mu^{a},\sigma^{a})^{t}$ y $\boldsymbol{\theta}^{b}=\left(\mu^{b},\sigma^{b}\right)^{t}$

$$\begin{aligned}\mathrm{Dist}\left(\boldsymbol{\vartheta}^{a},\boldsymbol{\vartheta}^{b}\right)&=\int_{\mathrm{Camino}}|\mathrm{d}\ell|\\&=\int_{\mathrm{Camino}}\sqrt{\mathrm{d}\boldsymbol{\vartheta}(t)^{T}J[\boldsymbol{\vartheta}(t)]\mathrm{d}\boldsymbol{\vartheta}(t)}\\&=\int_{0}^{1}\sqrt{\dot{\boldsymbol{\vartheta}}^{t}J(\boldsymbol{\vartheta})\dot{\boldsymbol{\vartheta}}}\mathrm{d}t\end{aligned}$$

* **Tensor métrico:** usando la métrica de Fisher de la distr. Gaussiana. Conectamos los puntos a través de un camino lineal

$$\boldsymbol{v}(t)=\begin{pmatrix}\mu(t)\\\sigma(t)\end{pmatrix}=\begin{pmatrix}t\mu^{b}+(1-t)\mu^{a}\\t\sigma^{b}+(1-t)\sigma^{a}\end{pmatrix}$$

con $t\in[0,1]$. La dist. se calcula integrando diferenciales de dist. $d\hat{\ell}$ a lo largo del camino, teniendo en cuenta que

$$\mathrm{d}\ell=\mathrm{d}\boldsymbol{\vartheta}(t)=\dot{\boldsymbol{\vartheta}}\mathrm{d}\boldsymbol{t}=\begin{pmatrix}\mu^{b}-\mu^{a}\\\sigma^{b}-\sigma^{a}\end{pmatrix}\mathrm{d}t,$$

donde $\dot{\vartheta}=\mathrm{d}\vartheta(t)/\mathrm{d}t$. La distancia es una cantidad siempre positiva.

$$\begin{aligned}\mathrm{Dist}\left(\boldsymbol{\vartheta}^{a},\boldsymbol{\vartheta}^{b}\right)&=\frac{\sqrt{\left(\mu^{b}-\mu^{a}\right)^{2}+2\left(\sigma^{b}-\sigma^{2}\right)^{2}}}{\sigma^{b}-\sigma^{a}}\int_{0}^{\sigma^{b}-\sigma^{a}}\frac{\mathrm{d}z}{\sigma^{a}+z}\\&=\frac{\sqrt{\left(\mu^{b}-\mu^{a}\right)^{2}+2\left(\sigma^{b}-\sigma^{2}\right)^{2}}}{\sigma^{b}-\sigma^{a}}\ln\left(\frac{\sigma^{b}}{\sigma^{a}}\right)\end{aligned}$$

~ Si trasladamos μ^{a} y μ^{b} en una cantidad fija, la *Dist.* no varía. Pero si desplazamos las σ, la *Dist.* varia. De hecho, si alguna $\sigma\to 0\Rightarrow$ *Dist.* $\to\infty$. Si alguna $\sigma\to 0$, la distr. correspondiente tiende a una δ Dirac.

* **Tensor:** de Fisher de 2×2 para la distr. Gaussiana Dada la distr.

$$p(x\mid\mu,\sigma)=\frac{e^{-(z-\mu)^{2}/2\sigma^{2}}}{\sqrt{2\pi\sigma^{2}}}$$

$$J(\mu,\sigma)=\begin{pmatrix}J_{\mu\mu}&J_{\mu\sigma}\\J_{\mu\sigma}&J_{\sigma\sigma}\end{pmatrix}\ =\frac{1}{\sigma^{2}}\begin{pmatrix}1&0\\0&2\end{pmatrix}.$$