Introducción a la Teoría de la Información
Rate Distortion Theory

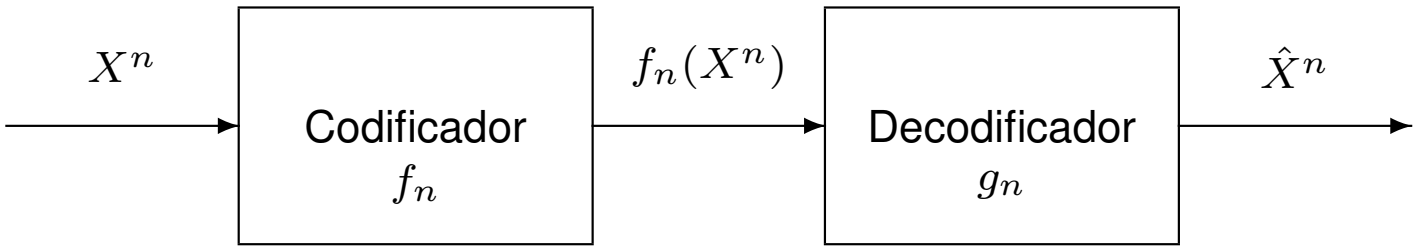
Facultad de Ingeniería, UdelaR

Agenda

Agenda

Introducción

- En la representación de un símbolo $X \in \mathcal{R}$ con un número finito de bits R necesariamente se cometerán errores ("*distorsión*").
- ¿Cuán rápido se puede transmitir sobre un canal *sin ruido* si se permite cometer errores en la representación?
  - Dada una distorsión D mínima "tolerable" para representar en el receptor una señal (la fuente de origen) ¿cuál es la cantidad mínima de información $R(D)$ que se debe transmitir?



$$f_n(X^n) \in \{1, 2, \ldots, 2^{nR}\} \text{ y } g_n(f_n(x^n)) = \hat{X}^n$$

Agenda

Cuantificación de una variable aleatoria real

Definición (Cuantificador Vectorial)

Es un mapeo (función) de un vector de un espacio euclídeo de dimensión k, $\mathcal{R}^k$, en un conjunto finito $\mathcal{C}$ (codebook) conteniendo N vectores de salida $\hat{X}_i \in \mathcal{R}^k$ (codevector)

$$Q : \mathcal{R}^k \rightarrow \mathcal{C} = \{\hat{X}_1, \hat{X}_2, \ldots, \hat{X}_N\}$$

- Asociado a cada uno de los $\hat{X}_i$ existe un región o celda

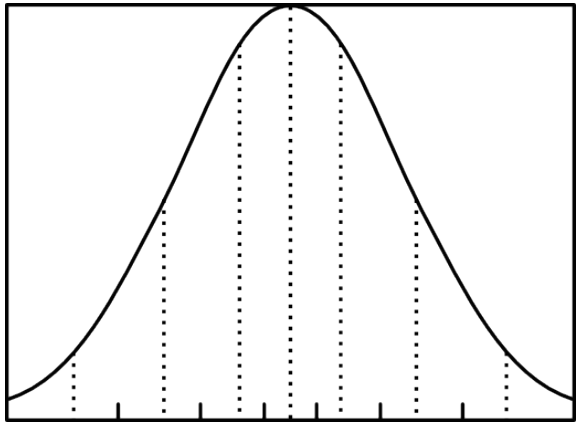
$$\mathcal{R}_i = \{X \in \mathcal{R}^k | Q(X) = \hat{X}_i\}$$

- Es un proceso con pérdidas; hay un error en la representación

$$D = \int_{\mathcal{R}^k} d(X, Q(X)) p_X(X) dX$$

Caso unidimensional

Supongamos $X \sim \mathcal{N}(0, \sigma^2)$ y una representación con R bits ($N = 2^R$).



Dada una $d(\cdot, \cdot)$ se busca minimizar D donde las variables son $\{\hat{X}_1, \hat{X}_2, \ldots, \hat{X}_N\}$ y las regiones (x_j con $j = 1, \ldots, 2^R$).

Ejercicio

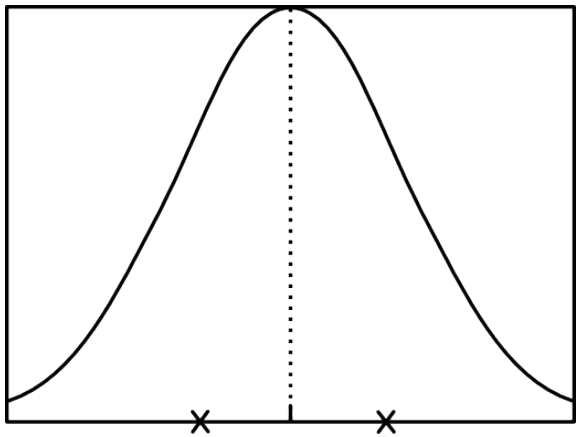
Dada una variable aleatoria $X \sim p_X()$ unidimensional y medida de distorsión $d(x, y) = (x - y)^2$ los codevectors son

$$\hat{X}_i = \frac{\int_{x_i}^{x_{i+1}} x p_X(x) dx}{\int_{x_i}^{x_{i+1}} p_X(x) dx}$$

y los límites de las regiones son $x_i = \frac{1}{2}\left(\hat{X}_i + \hat{X}_{i-1}\right)$.

Caso unidimensional

Cuantificación de una V.A. gaussiana con un bit (R=1) y distancia euclídea.



$$\hat{X}(x) = \begin{cases} \sqrt{\frac{2}{\pi}}\sigma & \text{si } x \geq 0 \\ -\sqrt{\frac{2}{\pi}}\sigma & \text{si } x < 0 \end{cases}$$

Con este valor queda

$$D = 2\int_0^{+\infty} \left(x - \sqrt{\frac{2}{\pi}}\sigma \right)^2 p_X(x)dx = \left(\frac{\pi - 2}{\pi} \right) \sigma^2 = 0{,}3634\sigma^2$$

Cuantificación: condiciones de optimalidad

Existen tres condiciones necesarias que debe cumplir un cuantificador óptimo:

1. *Condición de vecino más cercano*
   Dado un código $\{\hat{X}_i\}$ la distorsión se minimiza mapeando cada elemento X al codevector más cercano. Definiendo las regiones de *Voronoi* o partición de *Dirichlet*.
2. *Condición de centroide*
   Dada una partición la distorsión se minimiza eligiendo los centroides como los elementos del código.
3. *Condición de probabilidad cero en los bordes*

Estas condiciones permiten definir un algoritmo simple para el diseño de un cuantificador vectorial. Este algoritmo es el *algoritmo de Lloyd generalizado*.

Algoritmo de Lloyd generalizado

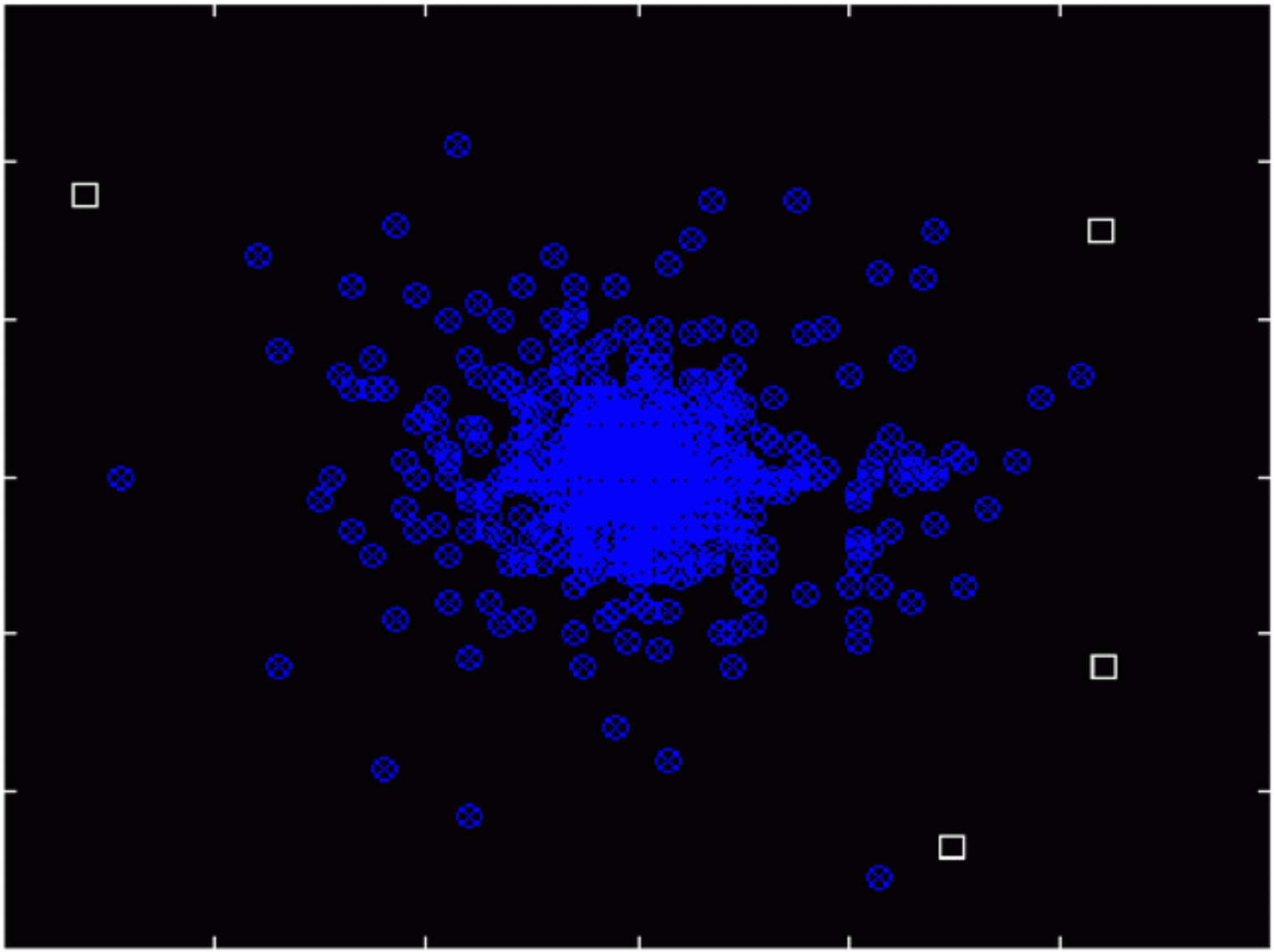
1. Se comienza con un codebook inicial $\mathcal{C}_1$, y $m = 1$
2. Dado el codebook $\mathcal{C}_m$ se obtiene $\mathcal{C}_{m+1}$ mediante:
   1. Encontrar la partición óptima mediante la condición de vecino más cercano.
   2. Utilizar la condición de centroide sobre las nuevas regiones y asignar en $\mathcal{C}_{m+1}$ estos nuevos centroides.
3. Se calcula la distorsión media para $\mathcal{C}_{m+1}$. Si el cambio desde la última iteración es menor que cierto umbral, se detiene el algoritmo. Si no es así se hace $m = m + 1$ y se vuelve al paso 2.

Algoritmo de Lloyd generalizado

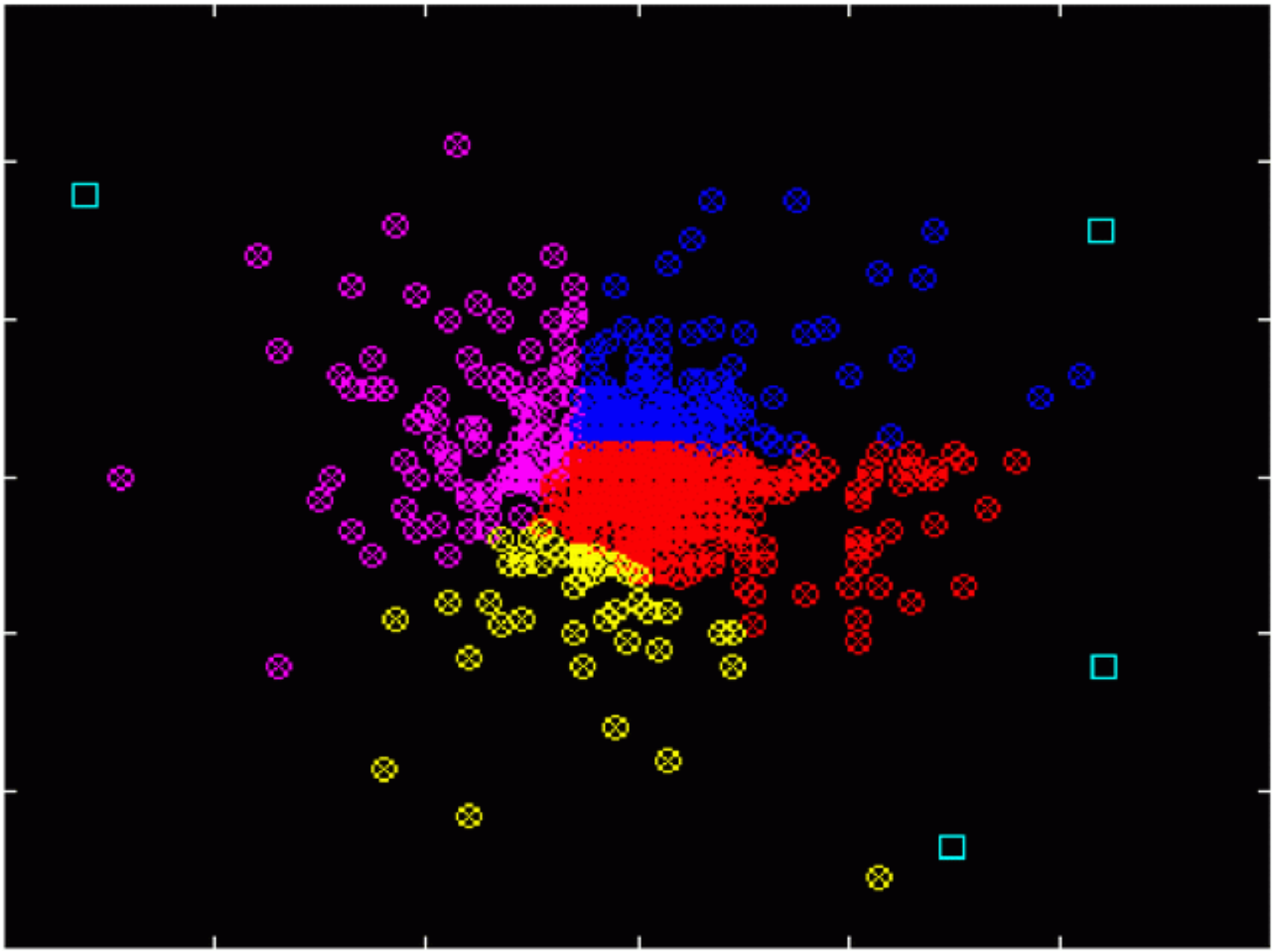
1. Se comienza con un codebook inicial $\mathcal{C}_1$, y $m = 1$
2. Dado el codebook $\mathcal{C}_m$ se obtiene $\mathcal{C}_{m+1}$ mediante:
   1. Encontrar la partición óptima mediante la condición de vecino más cercano.
   2. Utilizar la condición de centroide sobre las nuevas regiones y asignar en $\mathcal{C}_{m+1}$ estos nuevos centroides.
3. Se calcula la distorsión media para $\mathcal{C}_{m+1}$. Si el cambio desde la última iteración es menor que cierto umbral, se detiene el algoritmo. Si no es así se hace $m = m + 1$ y se vuelve al paso 2.

- Verificar la condición de probabilidad cero en los bordes.
- El algoritmo de Lloyd (la iteración) garantiza que la distorsión no crece.
- Converge en un número finito de pasos a un mínimo local.
- Dependencia muy fuerte del codebook inicial que se utilice en el algoritmo.

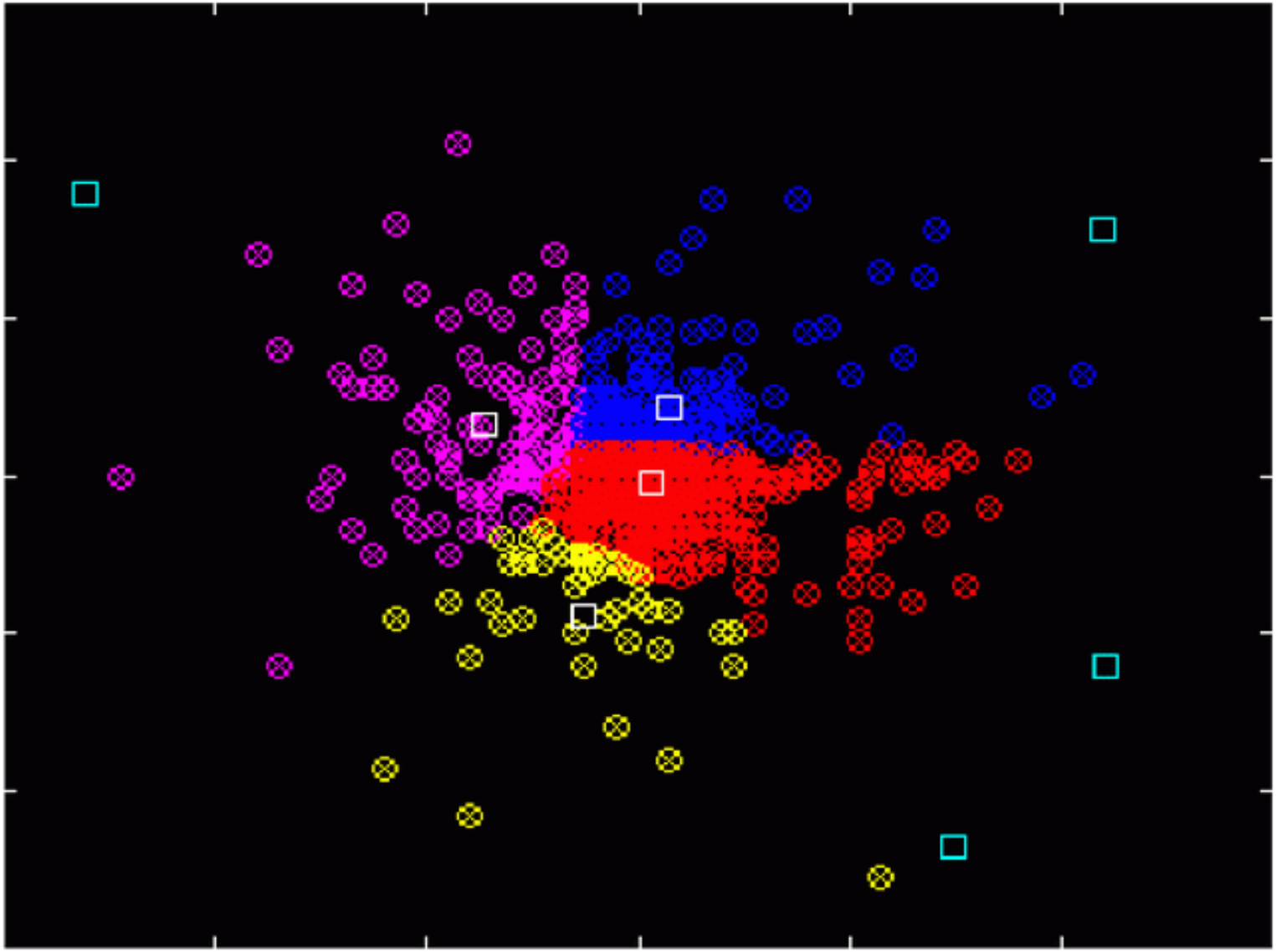
Algoritmo de Lloyd generalizado: ejemplo



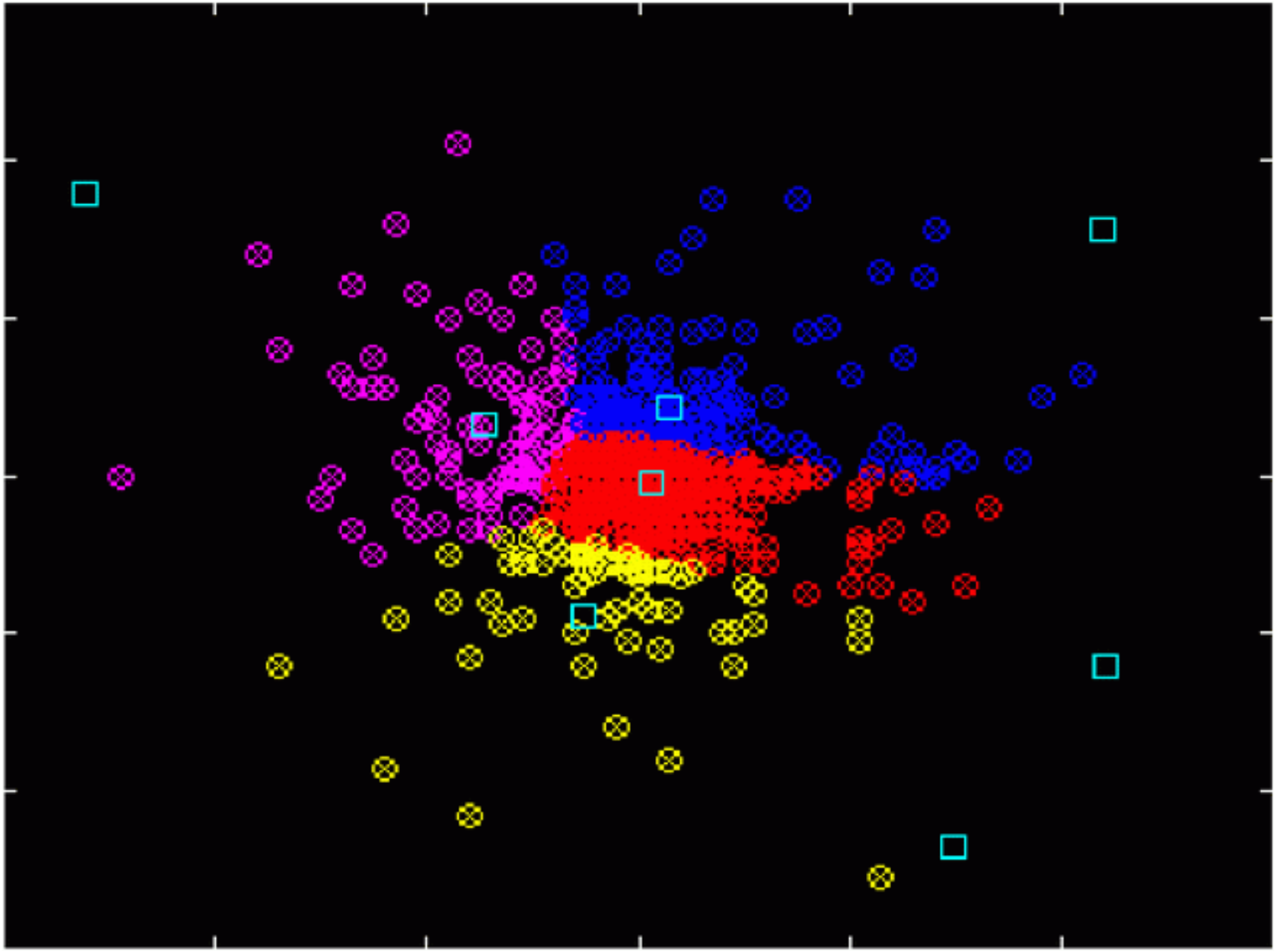
Algoritmo de Lloyd generalizado: ejemplo



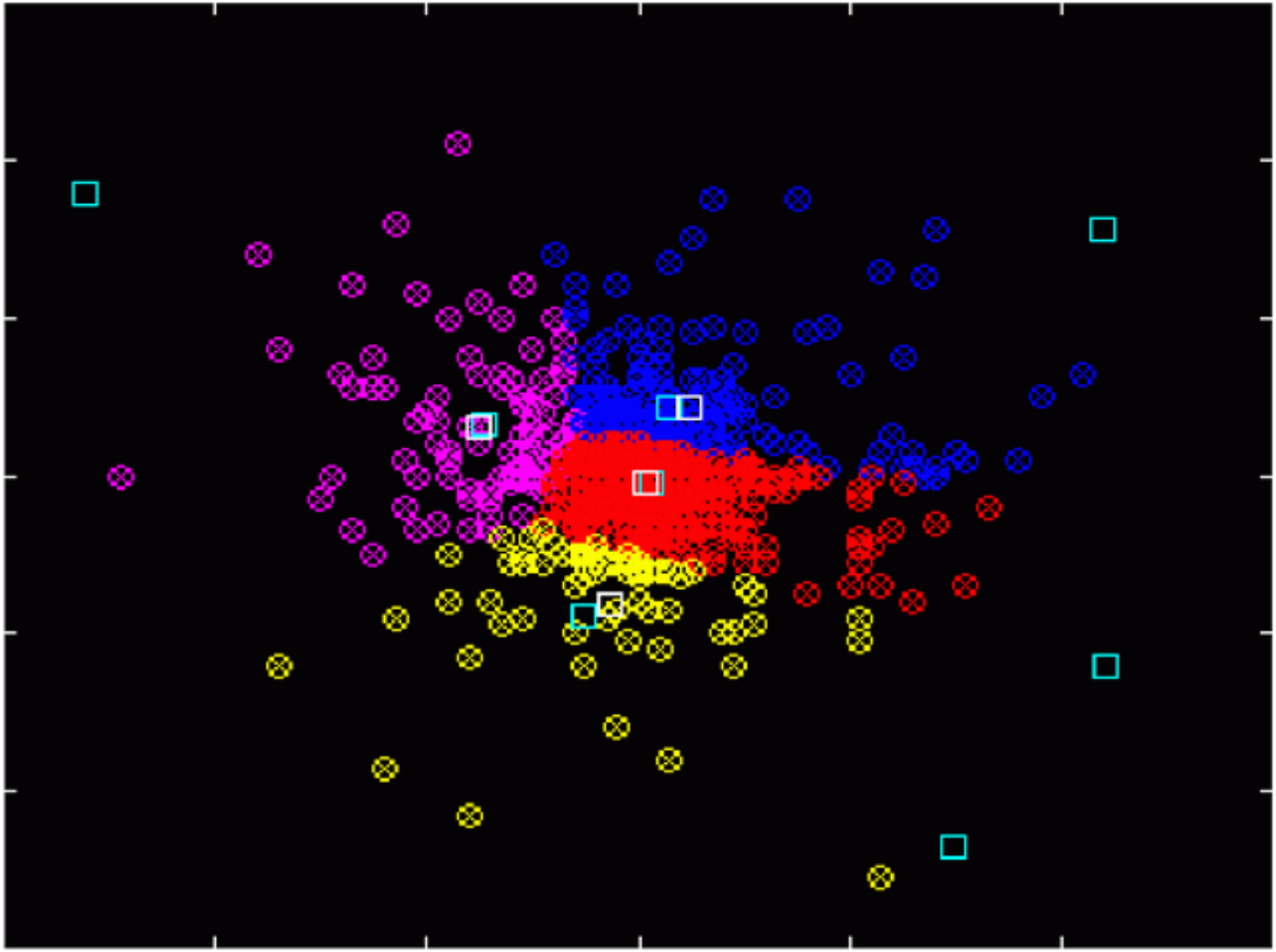
Algoritmo de Lloyd generalizado: ejemplo



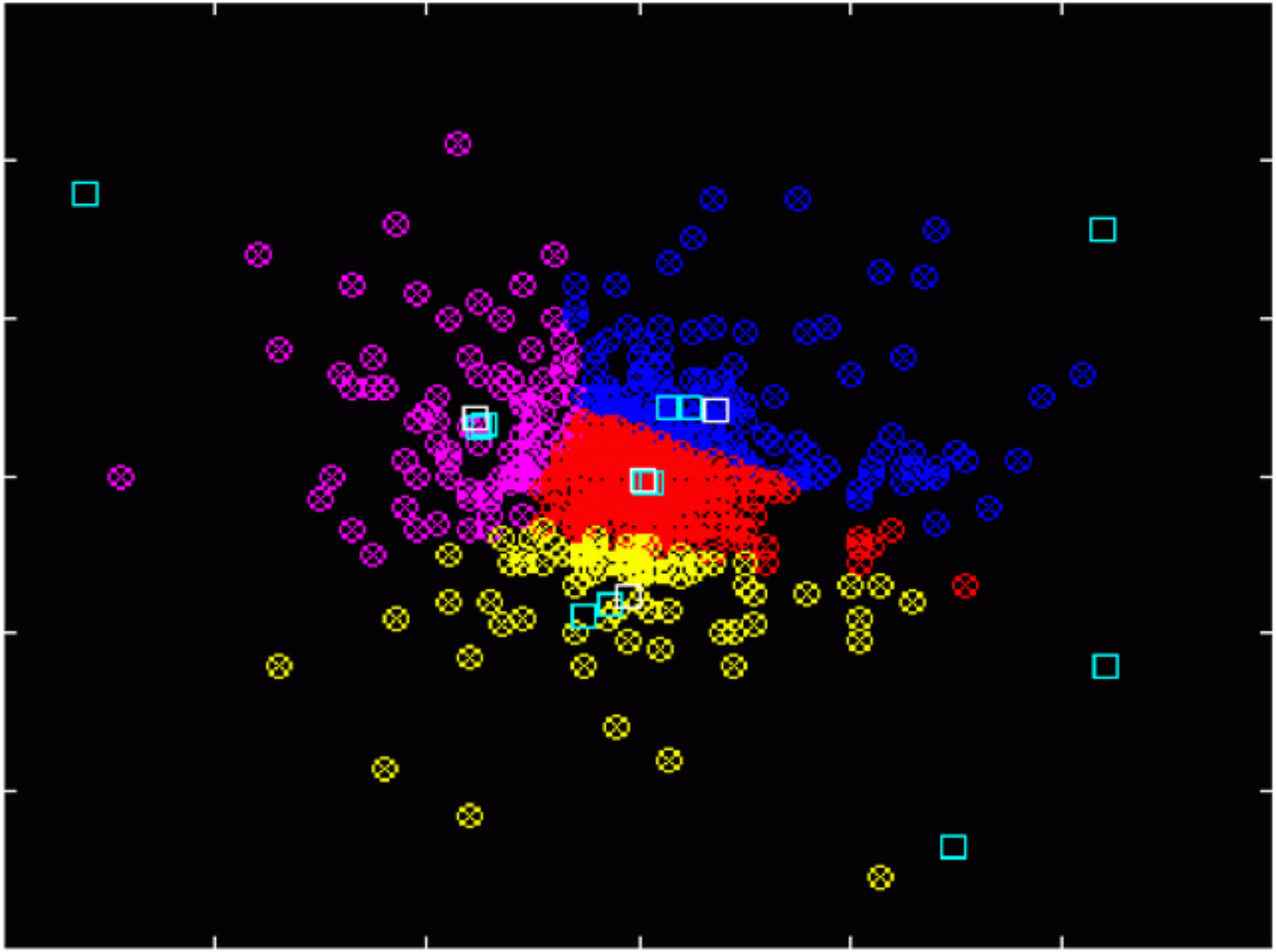
Algoritmo de Lloyd generalizado: ejemplo



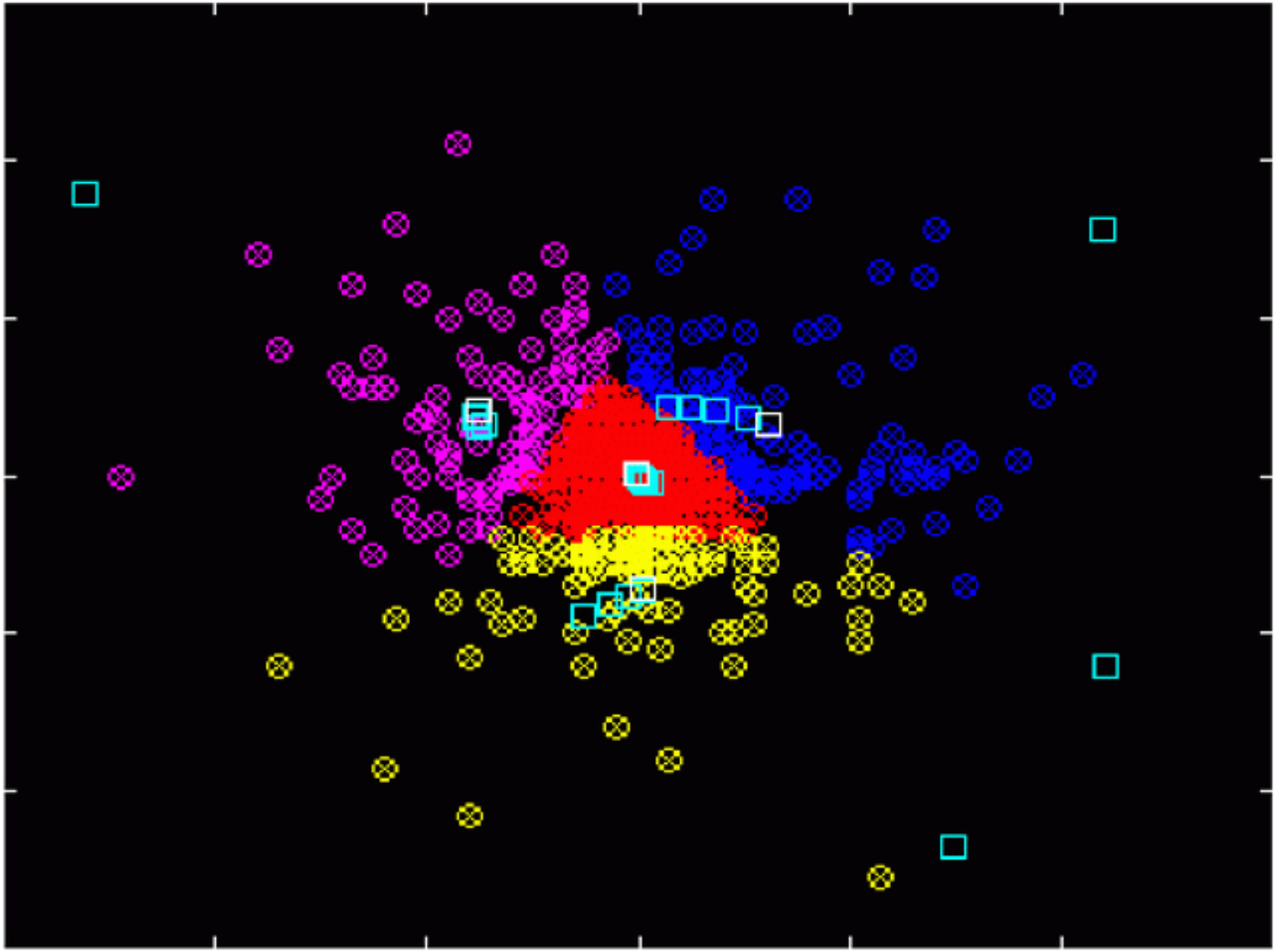
Algoritmo de Lloyd generalizado: ejemplo



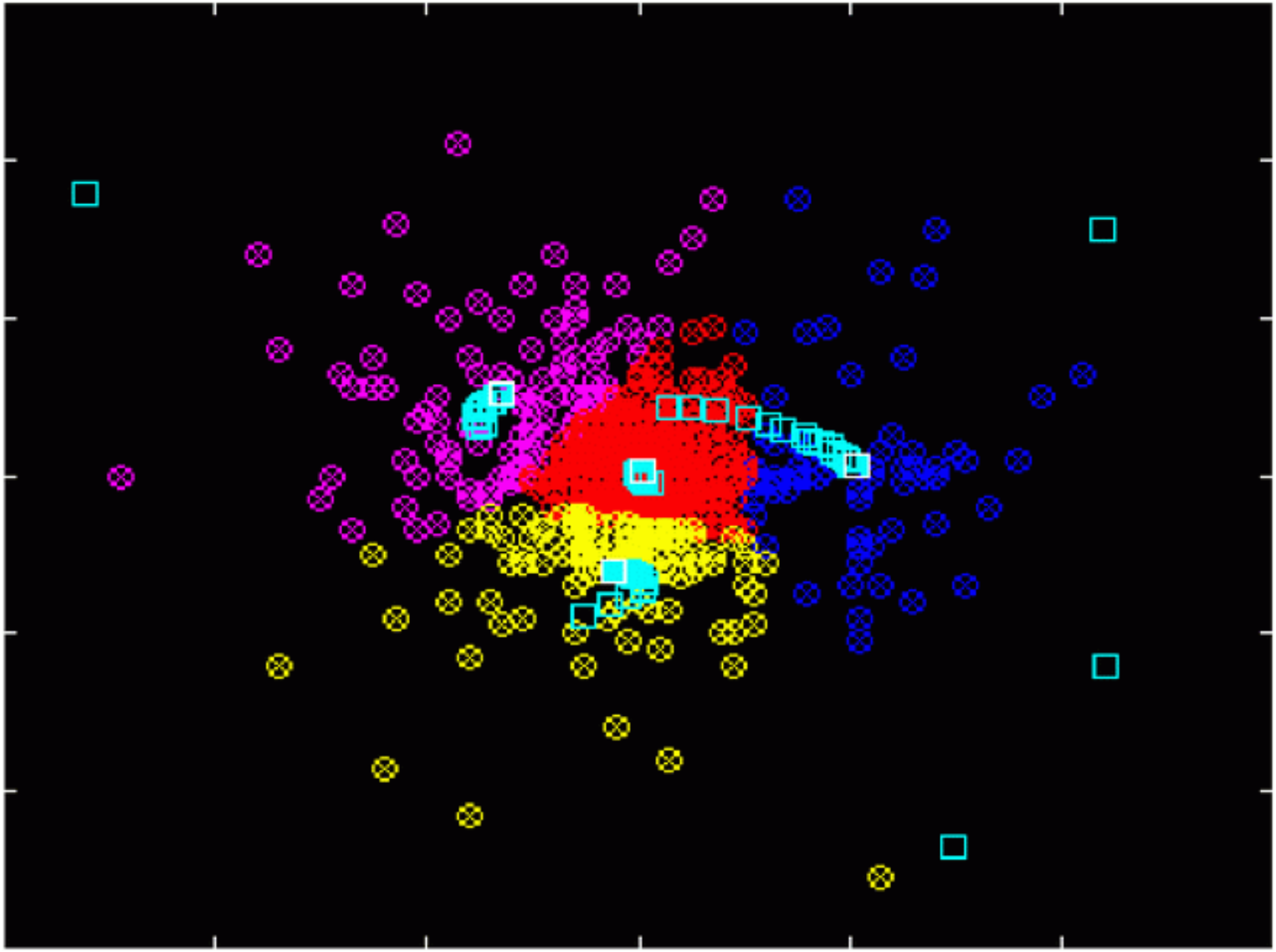
Algoritmo de Lloyd generalizado: ejemplo



Algoritmo de Lloyd generalizado: ejemplo



Algoritmo de Lloyd generalizado: ejemplo



Cuantificación vectorial: un ejemplo

Cuantificación vectorial de una imágen, calculando el codificador óptimo.

"Vectores" de dimensión 4 (bloques de 2×2), codebook de tamaño 32. Distorsión $D = 95{,}1$ y $\mathrm{PSNR} = 28,35\ \mathrm{dB}$.

Agenda

Medida de Distorsión

Definición (Medida de distorsión)

Una función de distorsión o medida de distorsión es un mapeo

$$d : \mathcal{X} \times \hat{\mathcal{X}} \to \mathcal{R}^+$$

Es una medida del costo de representar $x \in \mathcal{X}$ por $\hat{x} \in \hat{\mathcal{X}}$.

Definición (Distorsión entre secuencias)

La distorsión entre dos secuencias x^n y $\hat{x}^n$ es

$$d(x^n, \hat{x}^n) = \frac{1}{n} \sum_{i=1}^{n} d(x_i, \hat{x}_i)$$

Medida de Distorsión

Definición (Medida de distorsión acotada)

Una función de distorsión o medida de distorsión es acotada si su máximo valor es finito

$$d_{\text{máx}} = \max_{x \in \mathcal{X}, \hat{x} \in \hat{\mathcal{X}}} d(x, \hat{x}) < \infty$$

Example

- Distorsión de Hamming (Probabilidad de error)

$$d(x, \hat{x}) = \begin{cases} 0 & \text{si } x = \hat{x} \\ 1 & \text{si } x \neq \hat{x} \end{cases}$$

- Distorsión del Error Cuadrático Medio (MSE)

$$d(x, \hat{x}) = (x - \hat{x})^2$$

Código $(2^{nR}, n)$ con distorsión

Definición (Código $(2^{nR}, n)$ con distorsión)

Un código $(2^{nR}, n)$ con distorsión consiste de una función de codificación f_n

$$f_n : \mathcal{X} \to \{1, 2, \ldots, 2^{nR}\}$$

y una función de decodificación g_n

$$g_n : \{1, 2, \ldots, 2^{nR}\} \to \hat{\mathcal{X}}$$

Definición (Codebook)

$\{g_n(1), g_n(2), \ldots, g_n(2^{nR})\} = \left\{\hat{X}^n(1), \hat{X}^n(2), , \ldots, \hat{X}^n(2^{nR})\right\}$ es el codebook.

Definición (Distorsión de un Código)

La distorsión asociada a este codigo se define como

$$\begin{aligned} D &= E_{p(x^n)}\left[d(X^n, g_n(f_n(X^n)))\right] \\ &= \sum_{x^n} p(x^n) d(x^n, g_n(f_n(x^n))) \end{aligned}$$

Par tasa-distorsión (R, D) alcanzable

Definición (Par tasa-distorsión (R, D) alcanzable)

Un par tasa-distorsión (R, D) se dice alcanzable si existe una secuencia de códigos $(2^{nR}, n)$, (f_n, g_n), con

$$\lim_{n \uparrow \infty} E_{p(x^n)} \left[d(X^n, g_n(f_n(X^n))) \right] \leq D$$

Definición (Región tasa-distorsión)

La región tasa-distorsión para una fuente es la clausura del conjunto de pares tasa-distorsión (R, D) alcanzables.

$R(D)$ y $D(R)$

Definición (Función tasa-distorsión $R(D)$)

La función tasa-distorsión $R(D)$ es el ínfimo de las tasas R tal que el par (R, D) está en la región tasa-distorsión de una fuente para una distorsión dada D.

- Esta es la definición *operacional* de tasa-distorsión.
- Otro (R', D) (alcanzable) será $R'(D) \geq R(D)$, lo que implica que tendrá una distorsión no mayor que D, pero a costa de una descripción de la fuente con más bits.

Definición (Función distorsión-tasa $D(R)$)

La función distorsión-tasa $D(R)$ es el ínfimo de las distorsiones tal que el par (R, D) está en la región tasa-distorsión de una fuente para una distorsión dada R.

Función tasa-distorsión *informacional* $R^I(D)$

Definición (Función tasa-distorsión *informacional* $R^I(D)$)

La función tasa-distorsión *informacional* $R^I(D)$ para una fuente X con medida de distorsión $d(x, \hat{x})$ es

$$R^I(D) = \min_{p(\hat{x}|x):\sum_{(x,\hat{x})} p(x)p(\hat{x}|x)d(x,\hat{x}) \leq D} I(X;\hat{X})$$

Teorema fundamental

Theorem (Teorema tasa-distorsión)

La función tasa-distorsión $R(D)$ para una fuente X i.i.d. con distribución $p(x)$ y función distorsión acotada $d(x, \hat{x}) \leq d_{\text{máx}} < \infty$ es igual a la función tasa-distorsión informacional $R^I(D)$

$$R(D) = R^I(D)$$

- Primero veremos el cálculo de $R(D)$ para casos sencillos
- Luego el recíproco ($R \geq R(D)$ para cualquier código con distorsión D)
- Finalmente las ideas de la demostración del directo (alcanzabilidad: existen códigos con tasa $R^I(D)$ con distorsión D)

Agenda

Cálculo de $R(D)$ para una fuente binaria I

Theorem

La función tasa-distorsión para una fuente $X \sim Bernoulli(p)$ con distancia de Hamming (proporción de error) menor o igual que D es

$$R(D) = R^I(D) = \begin{cases} H(p) - H(D) & \text{si } 0 \leq D \leq \min\{p, 1-p\} \\ 0 & \text{si } D > \min\{p, 1-p\} \end{cases}$$

$$R^I(D) = \min_{p(\hat{x}|x):\sum_{(x,\hat{x})} p(x)p(\hat{x}|x)d(x,\hat{x}) \leq D} I(X;\hat{X})$$

Vale $P(X \neq \hat{X}) \leq D$ y supondremos $P(X = 1) = p < \frac{1}{2}$.
Notaremos $\oplus_2$ la suma módulo 2. El evento $X \oplus_2 \hat{X} = 1$ es equivalente a $X \neq \hat{X}$.
Acotemos $I(X;\hat{X})$

$$\begin{aligned} I(X;\hat{X}) &= H(X) - H(X|\hat{X}) && (1) \\ &= H_2(p) - H(X \oplus_2 \hat{X}|\hat{X}) && (2) \\ &\geq H_2(p) - H(X \oplus_2 \hat{X}) && (3) \\ &\geq H_2(p) - H_2(D) && (4) \end{aligned}$$

Cálculo de $R(D)$ para una fuente binaria II

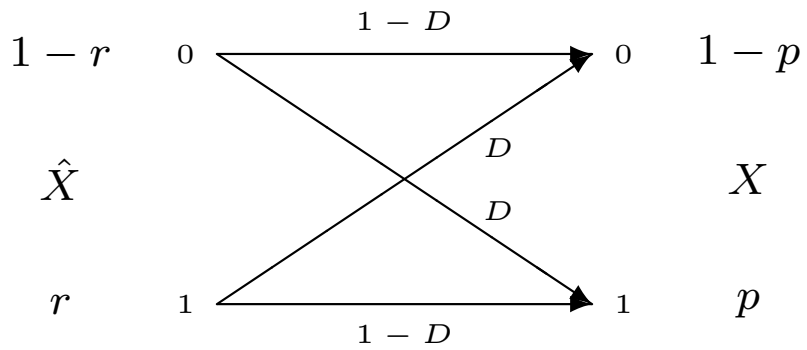
(4) pues $H(X \oplus_2 \hat{X}) = H_2(P(X \neq \hat{X})) \leq H_2(D)$ para $D \leq \frac{1}{2}$
Entonces

$$R(D) \geq H(p) - H(D).$$

Demostraremos que la cota se alcanza encontrando una distribución conjunta $p(x, \hat{x})$ que verifica la igualdad

$$I(X; \hat{X}) = R(D)$$

con $0 \leq D \leq \text{mín}\{p, 1-p\}$
La encontraremos planteando el siguiente BSC.



Cálculo de $R(D)$ para una fuente binaria III

Sabemos que $P(X=1) = P(\hat{X}=0)P(\hat{X} \neq X) + P(\hat{X}=1)P(\hat{X}=X)$

$$\Rightarrow r = \frac{p-D}{1-2D}$$

Con esto $P(\hat{X} \neq X) = (1-r)D + rD = D$ si $D \leq p$

$$\Rightarrow p(x|\hat{x}) = \begin{cases} D & \text{si } X \neq \hat{X} \\ 1-D & \text{si } X = \hat{X} \end{cases}$$

y

$$I(X;\hat{X}) = H(X) - H(X|\hat{X}) = H_2(p) - H_2(D)$$

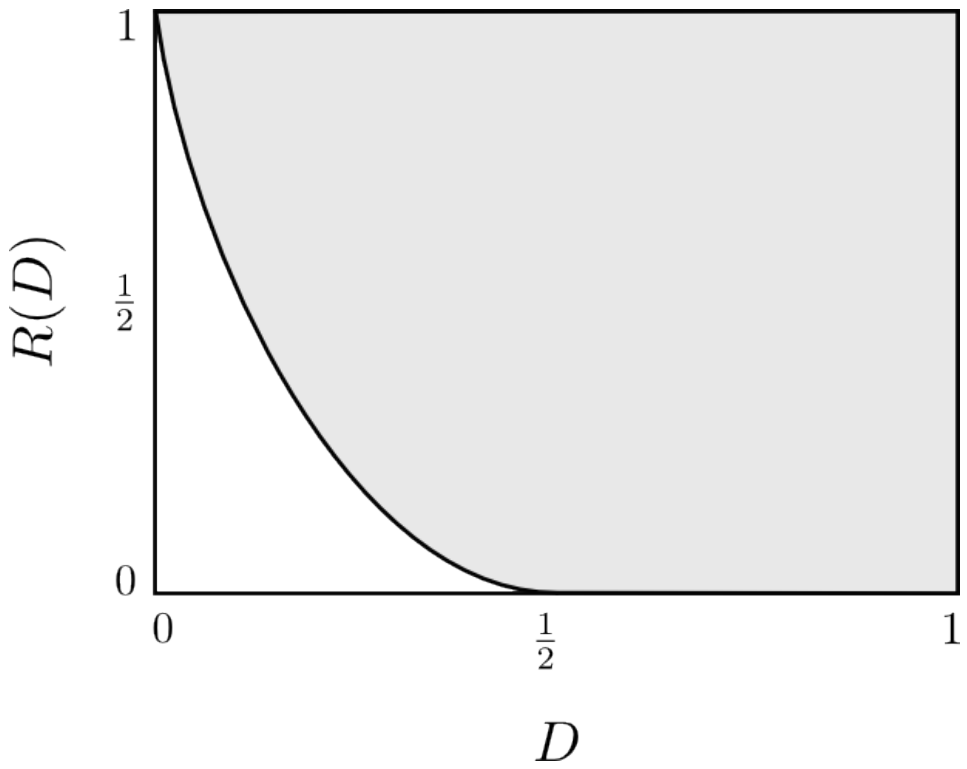
Si $D \geq p$, tomamos $r = 0$ ($P(\hat{X}=0) = 1$), entonces $I(X;\hat{X}) = 0$ y la distorsión máxima que se tendrá es $p \leq D$.

¡Si la distorsión tolerable es *grande* es innecesario mandar algo!

En resumen

$$R(D) = R^I(D) = \begin{cases} H(p) - H(D) & \text{si } 0 \leq D \leq \text{mín}\{p, 1-p\} \\ 0 & \text{si } D > \text{mín}\{p, 1-p\} \end{cases}$$

Cálculo de $R(D)$ para una fuente binaria IV



Cálculo de $R(D)$ para una fuente gaussiana I

El Teorema de tasa-distorsión también se puede enunciar y probar con fuentes continuas con buen comportamiento y medidas de distorsión no acotadas.

Theorem

La función tasa-distorsión para una fuente $X \sim f_X = \mathcal{N}(0, \sigma^2)$ con medida de distorsión de error cuadrático medio menor o igual que D es

$$R(D) = R^I(D) = \begin{cases} \frac{1}{2} \log \frac{\sigma^2}{D} & \text{si } 0 \leq D \leq \sigma^2 \\ 0 & \text{si } D > \sigma^2 \end{cases}$$

$$R(D) = R^I(D) = \min_{f_X(\hat{x}|x): E_{f_X}\left[(X-\hat{X})^2\right] \leq D} I(X; \hat{X})$$

Cálculo de $R(D)$ para una fuente gaussiana II

Haremos el mismo procedimiento que en el caso anterior.

$$
\begin{aligned}
I(X;\hat{X}) &= h(X) - h(X|\hat{X}) && (1)\\
&= \frac{1}{2}\log 2\pi e\sigma^2 - h(X-\hat{X}|\hat{X}) && (2)\\
&\geq \frac{1}{2}\log 2\pi e\sigma^2 - h(X-\hat{X}) && (3)\\
&\geq \frac{1}{2}\log 2\pi e\sigma^2 - h(\mathcal{N}(0, E\left[(X-\hat{X})^2\right])) && (4)\\
&= \frac{1}{2}\log 2\pi e\sigma^2 - \frac{1}{2}\log 2\pi e E\left[(X-\hat{X})^2\right] && (5)\\
&\geq \frac{1}{2}\log 2\pi e\sigma^2 - \frac{1}{2}\log 2\pi e D && (6)\\
&= \frac{1}{2}\log \frac{\sigma^2}{D} && (7)
\end{aligned}
$$

$$R(D) = I(X;\hat{X}) \geq \frac{1}{2}\log\frac{\sigma^2}{D}$$

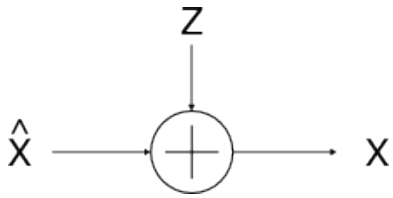
Construiremos $f_X(x|\hat{x})$ para alcanzar la igualdad de la cota.

Cálculo de $R(D)$ para una fuente gaussiana III

Si $D \leq \sigma^2$, elegimos

$$X = \hat{X} + Z, \quad \hat{X} \sim \mathcal{N}(0, \sigma^2 - D), \quad Z \sim \mathcal{N}(0, D)$$

con $\hat{X}$ y Z independientes.



$$\begin{aligned} I(X;\hat{X}) &= h(X) - h(X|\hat{X}) = h(X) - h(\hat{X} + Z|\hat{X}) \\ &= h(X) - h(Z) = \frac{1}{2}\log 2\pi e\sigma^2 - \frac{1}{2}\log 2\pi eD \\ &= \frac{1}{2}\log\frac{\sigma^2}{D} \end{aligned}$$

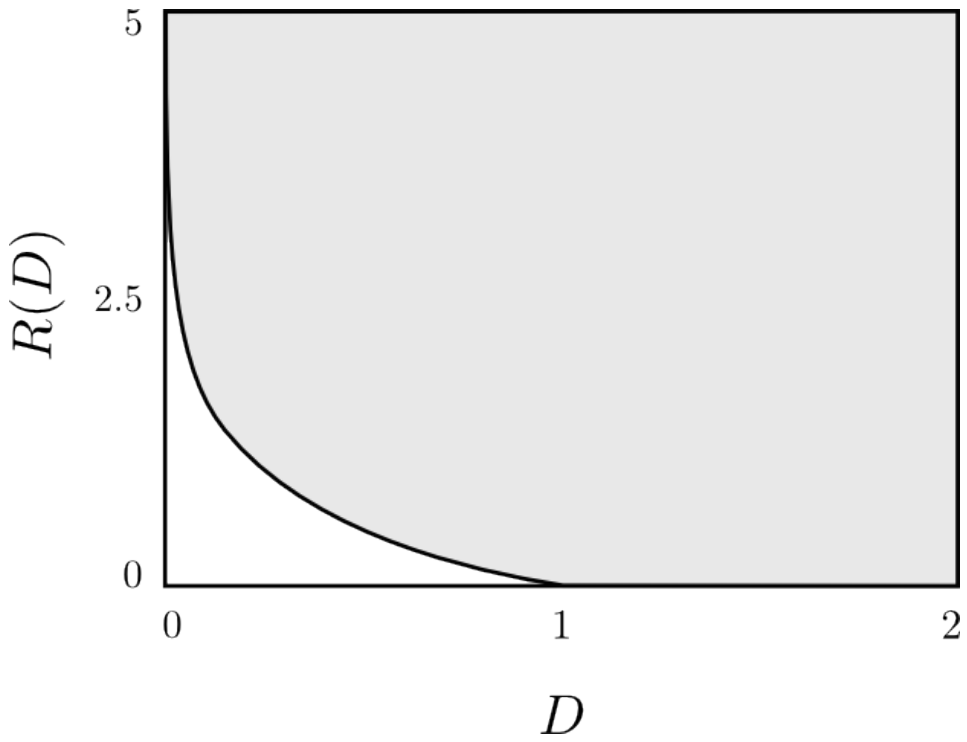
Cálculo de $R(D)$ para una fuente gaussiana IV

Finalmente, si $D \leq \sigma^2$

$$E\left[(X-\hat{X})^2\right] = E\left[Z^2\right] = D$$

Si $D > \sigma^2$ se toma $P(\hat{X}=0)=1$ y $R(D)=0$. La distorsión queda $\sigma^2 < D$.

$$R(D) = R^I(D) = \begin{cases} \frac{1}{2}\log\frac{\sigma^2}{D} & \text{si } 0 \leq D \leq \sigma^2 \\ 0 & \text{si } D > \sigma^2 \end{cases}$$



Cálculo de $R(D)$ para una fuente gaussiana V

Podemos reescribir

$$D(R) = \sigma^2 2^{-2R}$$

- Con $R = 1$ tenemos la cota para el cuantificador de un bit.

$$D(R = 1) = \frac{1}{4}\sigma^2$$

- Cada bit que agregamos en la descripción reduce la distorsión $\frac{1}{4}\sigma^2$
- La cuantificación calculada en la diapositiva 7 da

$$\tilde{D} = 0{,}3634\sigma^2 > 0{,}25\sigma^2 = D(1)$$

Como es posible qué se pueda hacer mejor que lo que se hizo en el ejemplo?

Es mejor (menor distorsión) describir varias variables independientes en bloque que describir cada una de ellas por separado.

Descripción simultánea de variables gaussianas I

Sea

$$X^m = (X_1, \ldots, X_m)$$

un vector de m variables aleatorias gaussianas independientes $X_i \sim \mathcal{N}(0, \sigma_i^2)$ y como medida de distorsión el error cuadrático medio

$$d(X^m, \hat{X}^m) = \sum_{i=1}^{m} (X_i - \hat{X}_i)^2.$$

¿Cómo distribuir R bits en la representación del vector X^m de forma de minimizar la distorsión?

Descripción simultánea de variables gaussianas II

$$
\begin{aligned}
I(X^m;\hat{X}^m) &= h(X^m) - h(X^m|\hat{X}^m) && (1)\\
&= \sum_{i=1}^{m} h(X_i) - \sum_{i=1}^{m} h(X_i|X^{i-1}\hat{X}^m) && (2)\\
&\geq \sum_{i=1}^{m} h(X_i) - \sum_{i=1}^{m} h(X_i|\hat{X}_i) && (3)\\
&= \sum_{i=1}^{m} I(X_i;\hat{X}_i) && (4)\\
&\geq \sum_{i=1}^{m} R(D_i) && (5)\\
&= \sum_{i=1}^{m} \text{máx}\left\{\frac{1}{2}\log\frac{\sigma_i^2}{D_i}, 0\right\} && (6)
\end{aligned}
$$

Descripción simultánea de variables gaussianas III

Tomando $f(x^m|\hat{x}^m) = \prod_{i=1}^m f(x_i|\hat{x}_i)$ y eligiendo $\hat{X}_i \sim \mathcal{N}(0, \sigma_i^2 - D_i)$ como en el ejemplo anterior se alcanza la igualdad de la cota

$$R(D) = \min_{\sum D_i = D} \sum_{i=1}^m \max\left\{\frac{1}{2}\log\frac{\sigma_i^2}{D_i}, 0\right\} = \min_{\sum D_i = D} \sum_{i=1}^m \left(\frac{1}{2}\log\frac{\sigma_i^2}{D_i}\right)^+$$

Ahora se debe elegir la distorsión a asignar a cada una de las variables D_i según su varianza. Construimos el funcional $J(D)$ usando el multiplicador de Lagrange λ

$$J(D) = \sum_{i=1}^m \frac{1}{2}\log\frac{\sigma_i^2}{D_i} + \lambda \sum_{i=1}^m D_i$$

Descripción simultánea de variables gaussianas IV

$$\frac{\partial J}{\partial D_i} = -\frac{1}{2}\frac{1}{D}_i + \lambda = 0 \qquad \Rightarrow \qquad D_i = \frac{1}{2\lambda} = \lambda'$$

El óptimo se alcanza repartiendo la misma cantidad de bits para cada variable.

Esto es válido siempre que $\lambda' < \sigma_i^2 \ \forall i$. Si la distorisón aumenta y $\lambda' > \sigma_j^2$ la solución encontrada no es válida. El problema de minimización cambia y hay que usar las condiciones de Kuhn-Tucker, quedando

$$\frac{\partial J}{\partial D_i} = -\frac{1}{2}\frac{1}{D}_i + \lambda$$

donde λ se calcula para que

$$\frac{\partial J}{\partial D_i} \begin{cases} = 0 & \text{si } D_i < \sigma_i^2 \\ \leq 0 & \text{si } D_i \geq \sigma_i^2 \end{cases}$$

Descripción simultánea de variables gaussianas V

Theorem

Sea $X^m = (X_1, \ldots, X_m)$ *con* $X_i \sim \mathcal{N}(0, \sigma_i^2) \quad i = 1, \ldots, m$ *variables aleatorias gaussianas independientes con* $d(X^m, \hat{X}^m) = \sum_{i=1}^m (X_i - \hat{X}_i)^2$*. La función tasa-distorsión queda*

$$R(D) = \sum_{i=1}^{m} \left(\frac{1}{2} \log \frac{\sigma_i^2}{D_i} \right)$$

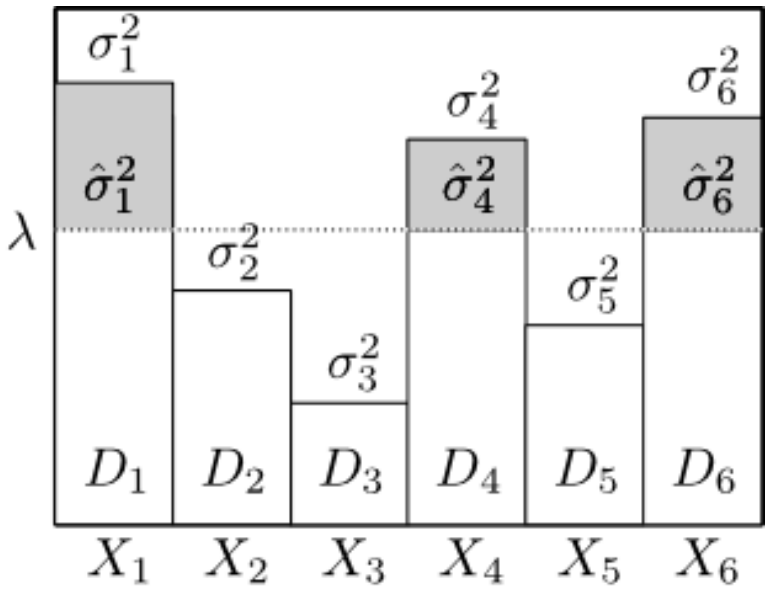
donde

$$D_i = \text{mín}\{\lambda, \sigma_i^2\}$$

con λ *para* $\sum_{i=1}^m D_i = D$

Esto implica un procedimiento similar al *water-filling* inverso...

Descripción simultánea de variables gaussianas VI



En este ejemplo la distorsión queda: $D = \lambda + \sigma_2^2 + \sigma_3^2 + \lambda + \sigma_5^2 + \lambda$

Elegida la constante λ solamente se describirán las variables cuya varianza es mayor que λ. Las variables con varianza menor que λ no son codificadas (no se les asigna ninguno de los R bits).

Agenda

Recíproco del teorema de tasa-distorsión

Teorema (Recíproco del teorema de tasa-distorsión)

No se puede obtener distorsión menor que D si describimos X con tasa menor a $R(D)$ donde

$$R(D) = \min_{p(\hat{x}|x):\sum_{(x,\hat{x})} p(x)p(\hat{x}|x)d(x,\hat{x})\leq D} I(X;\hat{X})$$

- Se prueba viendo que todo código $(2^{nR}, n)$ con distorsión menor que D implica $R \geq R(D)$.
- El punto clave de la demostración se basa en el hecho de que $R(D)$ es una función *no creciente* y *convexa* de D, lo que permite aplicar Jensen. (Demostrarlo, o leerlo del libro p.316)

Prueba, parte 1 de 3

Considerar un código con distorsión $(2^{nR}, n)$ cualquiera, (f_n, g_n) las funciones de codificación/decodificación y asumir $E\left[d(X^n, \hat{X}^n) \leq D\right]$

$$nR \quad \geq \quad H\left(f_n(X^n)\right) = H\left(\hat{X}^n\right)$$

Porque . . .

hay a lo sumo 2^{nR} valores de $\hat{X}^n$ y la entropía está acotada por el log de la cardinalidad del conjunto.

Prueba, parte 1 de 3

Considerar un código con distorsión $(2^{nR}, n)$ cualquiera, (f_n, g_n) las funciones de codificación/decodificación y asumir $E\left[d(X^n, \hat{X}^n) \leq D\right]$

$$\begin{aligned} nR &\geq H\left(f_n(X^n)\right) = H\left(\hat{X}^n\right) \\ &= H\left(\hat{X}^n\right) - H\left(\hat{X}^n | X^n\right) \end{aligned}$$

Porque . . .

$\hat{X}^n$ es *función* de X^n por lo que $H\left(\hat{X}^n | X^n\right) = 0$.

Prueba, parte 1 de 3

Considerar un código con distorsión $(2^{nR}, n)$ cualquiera, (f_n, g_n) las funciones de codificación/decodificación y asumir $E\left[d(X^n, \hat{X}^n) \leq D\right]$

$$\begin{aligned} nR &\geq H\left(f_n(X^n)\right) = H\left(\hat{X}^n\right) \\ &= H\left(\hat{X}^n\right) - H\left(\hat{X}^n | X^n\right) \\ &= I\left(X^n; \hat{X}^n\right) = H\left(X^n\right) - H\left(X^n | \hat{X}^n\right) \end{aligned}$$

Porque . . .

la definición de información mutua.

Prueba, parte 1 de 3

Considerar un código con distorsión $(2^{nR}, n)$ cualquiera, (f_n, g_n) las funciones de codificación/decodificación y asumir $E\left[d(X^n, \hat{X}^n) \leq D\right]$

$$
\begin{aligned}
nR &\geq H\left(f_n(X^n)\right) = H\left(\hat{X}^n\right) \\
&= H\left(\hat{X}^n\right) - H\left(\hat{X}^n | X^n\right) \\
&= I\left(X^n; \hat{X}^n\right) = H\left(X^n\right) - H\left(X^n | \hat{X}^n\right) \\
&= \sum_{i=1}^{n} H\left(X_i\right) - H\left(X^n | \hat{X}^n\right)
\end{aligned}
$$

Porque . . .

X_i son independientes.

Prueba, parte 1 de 3

Considerar un código con distorsión $(2^{nR}, n)$ cualquiera, (f_n, g_n) las funciones de codificación/decodificación y asumir $E\left[d(X^n, \hat{X}^n) \leq D\right]$

$$
\begin{aligned}
nR &\geq H\left(f_n(X^n)\right) = H\left(\hat{X}^n\right) \\
&= H\left(\hat{X}^n\right) - H\left(\hat{X}^n|X^n\right) \\
&= I\left(X^n; \hat{X}^n\right) = H\left(X^n\right) - H\left(X^n|\hat{X}^n\right) \\
&= \sum_{i=1}^{n} H\left(X_i\right) - H\left(X^n|\hat{X}^n\right) \\
&= \sum_{i=1}^{n} H\left(X_i\right) - \sum_{i=1}^{n} H\left(X_i|\hat{X}^n, X^{i-1}\right)
\end{aligned}
$$

Porque . . .

se aplica regla de la cadena.

Prueba, parte 1 de 3

Considerar un código con distorsión $(2^{nR}, n)$ cualquiera, (f_n, g_n) las funciones de codificación/decodificación y asumir $E\left[d(X^n, \hat{X}^n) \leq D\right]$

$$
\begin{aligned}
nR &\geq H\left(f_n(X^n)\right) = H\left(\hat{X}^n\right) \\
&= H\left(\hat{X}^n\right) - H\left(\hat{X}^n|X^n\right) \\
&= I\left(X^n; \hat{X}^n\right) = H\left(X^n\right) - H\left(X^n|\hat{X}^n\right) \\
&= \sum_{i=1}^{n} H\left(X_i\right) - H\left(X^n|\hat{X}^n\right) \\
&= \sum_{i=1}^{n} H\left(X_i\right) - \sum_{i=1}^{n} H\left(X_i|\hat{X}^n, X^{i-1}\right) \\
&\geq \sum_{i=1}^{n} H\left(X_i\right) - \sum_{i=1}^{n} H\left(X_i|\hat{X}_i\right) = \sum_{i=1}^{n} H\left(X_i\right) - H\left(X_i|\hat{X}_i\right)
\end{aligned}
$$

Porque . . .

el condicionamiento reduce entropía.

Prueba, parte 2 de 3

$$
\begin{aligned}
nR &\geq \sum_{i=1}^{n} H\left(X_i\right) - H\left(X_i|\hat{X}_i\right) \\
&= \sum_{i}^{n} I\left(X_i;\hat{X}_i\right)
\end{aligned}
$$

Porque . . .

definición de información mutua

Prueba, parte 2 de 3

$$
\begin{aligned}
nR &\geq \sum_{i=1}^{n} H\left(X_i\right) - H\left(X_i|\hat{X}_i\right) \\
&= \sum_{i}^{n} I\left(X_i;\hat{X}_i\right) \\
&\geq \sum_{i}^{n} R\left(Ed(X_i,\hat{X}_i)\right)
\end{aligned}
$$

Porque . . .

definición de $R(D)$

Prueba, parte 2 de 3

$$
\begin{aligned}
nR &\geq \sum_{i=1}^{n} H\left(X_i\right) - H\left(X_i|\hat{X}_i\right) \\
&= \sum_{i}^{n} I\left(X_i;\hat{X}_i\right) \\
&\geq \sum_{i}^{n} R\left(Ed(X_i,\hat{X}_i)\right) \\
&= n\sum_{i}^{n} \frac{1}{n} R\left(Ed(X_i,\hat{X}_i)\right)
\end{aligned}
$$

Porque . . .

"sin palabras"

Prueba, parte 2 de 3

$$
\begin{aligned}
nR &\geq \sum_{i=1}^{n} H\left(X_i\right) - H\left(X_i|\hat{X}_i\right) \\
&= \sum_{i}^{n} I\left(X_i;\hat{X}_i\right) \\
&\geq \sum_{i}^{n} R\left(Ed(X_i,\hat{X}_i)\right) \\
&= n\sum_{i}^{n} \frac{1}{n} R\left(Ed(X_i,\hat{X}_i)\right) \\
&\geq nR\left(\sum_{i}^{n} \frac{1}{n} Ed(X_i,\hat{X}_i)\right)
\end{aligned}
$$

Porque . . .

convexidad de $R(D)$ y Jensen

Prueba, parte 3 de 3

$$\begin{aligned} nR &\geq nR\left(\sum_{i}^{n}\frac{1}{n}Ed(X_i,\hat{X}_i)\right) \\ &= nR\left(Ed(X^n,\hat{X}^n)\right) \end{aligned}$$

Porque . . .

definición de distorsión para bloques de largo n.

Prueba, parte 3 de 3

$$\begin{aligned} nR &\geq nR\left(\sum_i \frac{1}{n} Ed(X_i, \hat{X}_i)\right) \\ &= nR\left(Ed(X^n, \hat{X}^n)\right) \\ &= nR(D). \end{aligned}$$

Porque . . .

R es no creciente en D y $E\left[d(X^n, \hat{X}^n) \leq D\right]$.

□

Separación de la codificación de fuente y de canal con distorsión

Argumentos similares se pueden aplicar cuando se envía una fuente codificada a través de un canal con ruido y se puede demostrar un teorema de separación fuente-canal en el mismo espíritu del que ya se demostró.

Distorsión ϵ-típica

Definición (Distorsión ϵ-típica)

Sea $p(x, \hat{x})$ la distribución de probabilidad conjunta de $\mathcal{X} \times \hat{\mathcal{X}}$ y $d(x, \hat{x})$ una medida de distorsión sobre $\mathcal{X} \times \hat{\mathcal{X}}$. Para todo $\epsilon > 0$ un par de secuencias $(x^n, \hat{x}^n)$ son *ϵ-tipicas con distorsión* si:

$$\begin{aligned}
\left| -\tfrac{1}{n} \log p(x^n) - H(X) \right| &< \epsilon \\
\left| -\tfrac{1}{n} \log p(\hat{x}^n) - H(\hat{X}) \right| &< \epsilon \\
\left| -\tfrac{1}{n} \log p(x^n, \hat{x}^n) - H(X^n, \hat{X}^n) \right| &< \epsilon \\
\left| d(x^n, \hat{x}^n) - E\left[d(X, \hat{X}) \right] \right| &< \epsilon
\end{aligned}$$

- Se define conjunto de distorsión típica respecto a $d(\cdot)$, $A_{d,\epsilon}^{(n)} \subset A_{\epsilon}^{(n)}$.
- Se prueban las mismas propiedades que para las otras AEP trivialmente a partir de la ley de grandes números.
- En particular, cuando x^n y $\hat{x}^n$ se eligen i.i.d., $\Pr\left\{ A_{d,\epsilon}^{(n)} \right\} \overset{n\uparrow\infty}{\to} 1$.

Agenda

Directo del teorema de tasa-distorsión

Teorema (Teorema tasa-distorsión)

La función tasa-distorsión $R(D)$ para una fuente X i.i.d. con distribución $p(x)$ y función distorsión acotada $d(x, \hat{x}) \leq d_{\text{máx}} < \infty$ es igual a la función tasa-distorsión informacional $R^I(D)$

$$R(D) = R^I(D)$$

Para cualquier D y $R > R(D)$ se puede obtener una secuencia de códigos con distorsión asintóticamente igual a D.

Lineamientos de la prueba I

Tiene una técnica de demostración similar a la usada en Capacidad de Canal.

Configuración

- Generación del *codebook*: generar aleatoriamente $\mathcal{C}$ de tamaño 2^{nR} secuencias (*codewords*) $\hat{X}^n \sim \prod p(\hat{x}_i)$, indizadas $\omega = \{1, 2, \ldots, 2^{nR}\}$
- Codificación: se codifica X^n con ω el índice del primer *codeword* tal que $(X^n, \hat{X}^n(\omega)) \in A_{d,\epsilon}^{(n)}$. Si no hay tal *codeword*, se elige arbitrariamente $\omega = 1$.
- Decodificación: se decodifica con $\hat{X}^n(w)$
- Cálculo de la distorsión: se calcula sobre los *codebooks* $\mathcal{C}$ aleatorios como

$$\overline{D} = E_{X^n,\mathcal{C}}\left[d\left(X^n, \hat{X}^n\right)\right]$$

Lineamientos de la prueba II

Cálculo y cota de la distorsión

Se acota la distorsión total

$$E\left[d\left(X^n, \hat{X}^n(X^n)\right)\right]$$

separando el cálculo en dos sumas:

1. Para las x^n que tienen un codeword ϵ-típico a ellas, $d\left(x^n, \hat{x}^n(w)\right) < D + \epsilon$ por definición. Éstas aportan a lo sumo $D + \epsilon$ a la distorsión total.
2. Para las x^n que no lo tienen, se toma su probabilidad total P_e y se acota la contribución total a la distorsión como $P_e d_{\text{máx}}$, donde $d_{\text{máx}} = \text{máx}\, d\left(X^n, \hat{X}^n\right)$ existe por hipótesis.

$$E\left[d\left(X^n, \hat{X}^n(X^n)\right)\right] \leq \underbrace{D + \epsilon}_{1} + \underbrace{P_e\, d_{\text{máx}}}_{2} \leq D + \delta$$

Dado ϵ puede encontrarse un δ que verifique lo anterior si se prueba que P_e es suficientemente pequeño. El resto de la prueba trata de acotar P_e.

Lineamientos de la prueba III

Cota de la P_e

- Para esta cota se usa la definición de la distorsión ϵ-típica.
- También es necesario que

$$R > I(X;\hat{X}) + 3\epsilon$$

para que

$$e^{-2^{n(R-I(X;\hat{X})-3\epsilon)}}$$

tienda a cero exponencialemente rápido con n, para lo cual se elige $p(\hat{x}|x)$ que alcanza el mínimo de la función tasa-distorsión, entonces $R > R(D)$ implica $R > I((X;\hat{X})$.

- De esta forma puede hacer P_e suficientemente pequeña con una adecuada selección de ϵ y n.

Mostramos que dado $\delta > 0$ existen ϵ y n que para un R aleatorio sobre bloques de largo n la distorsión media queda acotada por $D + \delta$. Entonces, debe existir por lo menos un código con ese R y largo de bloque n haciendo (R, D) alcanzables.

□ .

Codificación de canal para un canal gaussiano

La similitudes entre el teorema de codificación de canal y el de tasa-distorsión no quedan solo en las técnicas de demostración.

Canal gaussiano

Considerar $Y_i = X_i + Z_i$ con $Z_i \sim \mathcal{N}(0, N)$ i.i.d., y una restricción de potencia P por símbolo transmitido; con n transmisiones.

- Las secuencias transmitidas están en una esfera de radio $\sqrt{nP}$
- Se buscan 2^{nR} secuencias en esta esfera tal que las esferas de radio $\sqrt{nN}$ centradas en ellas sean disjuntas.
- El máximo de esferas que se pueden colocar es el cociente entre sus n-ésimas potencias de los radios

$$M \leq \frac{\left(\sqrt{n(P+N)}\right)^n}{\left(\sqrt{n(N)}\right)^n} \left(\frac{P+N}{N}\right)^{\frac{n}{2}} = 2^{nC}$$

Esto es empaquetamiento de esferas (*sphere packing*).

Tasa-distorsión para un canal gaussiano

Canal gaussiano

Considerar $X \sim \mathcal{N}(0, \sigma^2)$. Un código $(2^{nR}, n)$ con distorsión D para X es un conjunto de 2^{nR} secuencias en $\mathcal{R}^n$. Se quiere que:

- las secuencias de tamaño n a menos de $\sqrt{n\sigma^2}$ de alguna palabra de código (la mayoría de las secuencias de la fuente de tamaño n)...
- ...están a distancia menor de $\sqrt{nD}$ de alguna palabra de código.
- El mínimo de esferas que se pueden colocar es

$$2^{nR(D)} = \left(\frac{\sigma^2}{D}\right)^{\frac{n}{2}}$$

Esto es cubrimiento de esferas (*sphere covering*).

Concluyendo

- Se prueba que la mínima tasa es asintóticamente alcanzable, esto es, existen una colección de esferas de radio $\sqrt{nD}$ que *cubren* el espacio de la fuente, excepto por un conjunto de probabilidad arbitrariamente pequeña.
- Un buen código de canal se puede transformar en un buen código con distorsión.
- En codificación de canal se busca el código más grande que tenga la mayor distancia mínima entre las palabras de código. (Sphere packing)
- Mientras, en codificación con distorsión (tasa-distorsión) se busca el código más chico que cubre el espacio de entrada (Sphere covering)