Universidad de Buenos Aires
Facultad de Ingenieria
Año 2018 - 2er Cuatrimestre

Aprendizaje Estadístico, Teoría y aplicación

Trabajo Práctico Final

Sbruzzi, José Ignacio - Ingeniería Informática #97452
jose.sbru@gmail.com

Índice

1. Introducción a la primera parte

El teorema 5.2 del Gyorfi indica que la esperanza del riesgo L^2 de un estimador kernel que usa un kernel naive está acotada por una función de la cantidad de puntos de la muestra (n) y la cantidad de dimensiones del domínio (d). Se generan situaciones al azar variando n y d con el objetivo de observar de forma empírica las predicciones del teorema. Para esto se analiza el comportamiento al variar n entre 100 y 1000, utilizando $d \in \{1; 2; 3; 4; 6; 8; 10; 30\}$ y $h_n \in \{0{,}1; 0{,}5\}$.

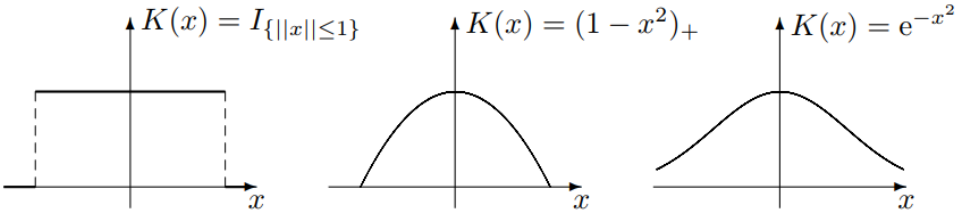
2. Introducción teórica

2.1. Estimador kernel

Un estimador kernel tiene la forma:

$$m_n(x) = \frac{\sum_{i=1}^{n} Y_i K\left(\frac{x - X_i}{h_n}\right)}{\sum_{i=1}^{n} K\left(\frac{x - X_i}{h_n}\right)}$$

y vale 0 cuando el denominador vale 0. Así, un estimador kernel no es otra cosa que un promedio ponderado de cada Y_i según la distancia entre x y X_i, multiplicada por $\frac{1}{h_n}$. A esta distancia se le aplica la función kernel $K(x)$. Generalmente $K(x)$ es grande cuando $|x|$ es grande. A continuación pueden observarse distintos kernels.



2.2. Kernel naive

El kernel naive se define como:

$$K(x) = \mathbb{1}\{|x| \leq 1\}$$

Así, los $x \in \mathbb{R}^d$ para los cuales $K(||\frac{x-X_i}{h_n}||) = 1$ conforman una bola de radio h_n centrada en X_i. Además, los únicos valores que puede tomar $K(y)$ son o bien 1 o bien 0 para cualquier $y \in \mathbb{R}$. Esto tiene como consecuencia una simplificación del análisis teórico.

2.3. Teorema 5.2

Teniendo un estimador kernel que utiliza un kernel naive, asumir que:[1]

$$Var(Y|X = x) \leq \sigma^2, x \in \mathbb{R}^d$$

y

$$|m(x) - m(z)| \leq ||x - z||, x, z \in \mathbb{R}^d$$

y X tiene un soporte compacto S^*. Entonces:[2]

$$\mathbb{E}||m_n - m||^2 \leq \hat{c}\frac{\sigma^2 + sup_{z\in S^*}|m(z)|^2}{n \cdot h_n^d} + C^2 h_n^2$$

donde $\hat{c}$ depende únicamente del diámetro de S^* y de d, entonces, para: [3]

$$h_n = c'\left(\frac{\sigma^2 + sup_{z\in S^*}|m(z)|^2}{C^2}\right)^{1/(d+2)} n^{\left(-\frac{1}{d+2}\right)}$$

[1]For a kernel estimate with a naive kernel assume that
[2]and X has a compact support S^*. Then
[3]where $\hat{c}$ depends only on the diameter of S^* and on d, thus for

tenemos que[4]

$$\mathbb{E}||m_n - m||^2 \leq c'' \left(\sigma^2 + \sup_{z \in S^*} |m(z)|^2 \right)^{2/(d+2)} C^{2d/(d+2)} n^{-2/(d+2)}$$

La demostración de este teorema está incluída en el anexo.

3. Descripción del primer experimento

El objetivo es observar empíricamente las consecuencias teóricas del teorema 5.2 del Gyorfy:

Para esto, se aborda el problema fijando los siguientes parámetros:

- $D = [-1, 1]^d$
- $x_i \in D$
- $|m(z)| \leq 1$ para todo $z \in D$
- El ruido agregado a $m(z)$ para generar los pares x_i, y_i es una normal cuya varianza varía a lo largo de D, pero nunca supera 1, con lo cual $Var(Y|X = x) = 1$, es decir, $\sigma = 1$

Así, queda acotado también C. De esta forma, la última ecuación del teorema puede escribirse como:

$$\mathbb{E}||m_n - m||^2 \leq c''(1+1)^{2/(d+2)} C^{2d/(d+2)} n^{-d/(d+2)}$$

Podemos hacer algo similar con la primera conclusión del teorema:

$$\mathbb{E}||m_n - m||^2 \leq \hat{c} \frac{1+1}{n \cdot h_n^d} + C^2 h_n^2$$

3.1. Estimación de $\mathbb{E}||m_n - m||^2$

A continuación se explican los pasos que usa el programa para estimar este valor para determinados n, d y h_n.

1. generar una función m con $-1 \leq m(x) \leq 1$ para todo $x \in D$. [5]
2. generar una función s con las mismas características.[5]

[4] we have

[5] en un anexo se explica cómo se generaron m y s

3. Generar un conjunto P de n pares (x_i, y_i) tales que $y_i = m(x_i) + S$, donde S tiene una distribución normal centrada en 0 y con una varianza $|s(x_i)| \leq 1$. Los puntos x_i pertenecen a D, es decir, tienen d dimensiones.

4. A partir de este conjunto P de pares, se genera una estimación de la regresión, m_n, usando un naive kernel y el h_n correspondiente.

5. Teniendo $m(x)$ y $m_n(x)$ definidos para todo $x \in D$, se utiliza la librería de python mcint para integrar $(m(x) - m_n(x))^2$ sobre todo D. mcint utiliza técnicas montecarlo para estimar la integral, ya que para d dimensiones la integral es difícil de calcular numericamente (es decir, tarda demasiado). Así se obtiene un $||m - m_n||^2$. Ver el anexo correspondiente para más detalles.

6. Se repite este procedimiento para una cantidad de $m(\cdot)$, $s(\cdot)$ y $m_n(\cdot)$ generadas al azar (en la mayoría de los casos se hicieron 300 experimentos para cada n y d, en otros casos se hicieron 100).

7. Se promedian los $||m - m_n||^2$ para obtener una estimación de la esperanza.

Así, se obtiene la función $encontrarEError(n, d, h_n)$.

3.2. Verificación del teorema

La segunda conclusión del teorema, para la situación tal como se la limitó, es:

$$\mathbb{E}||m_n - m||^2 \leq c''(1+1)^{2/(d+2)} C^{2d/(d+2)} n^{-d/(d+2)}$$

Así, se puede reemplazar $(1 + 1)^{2/(d+2)} C^{2d/(d+2)}$ por una constante c que depende sólo de d, y $d/(d + 2)$ por otra constante k que depende sólo de d. Así, para cada una de las combinaciones de d y h_n probados, al variar n se podría conseguir acotar los errores por una expresión de la forma

$$c(n^{-k})$$

con c y k constantes que se determinan a partir de los datos.

Se busca verificar que al variar n y mantener fijo d y h_n, se cumple que existe una cota de la forma

$$c(n^{-k})$$

tal que:

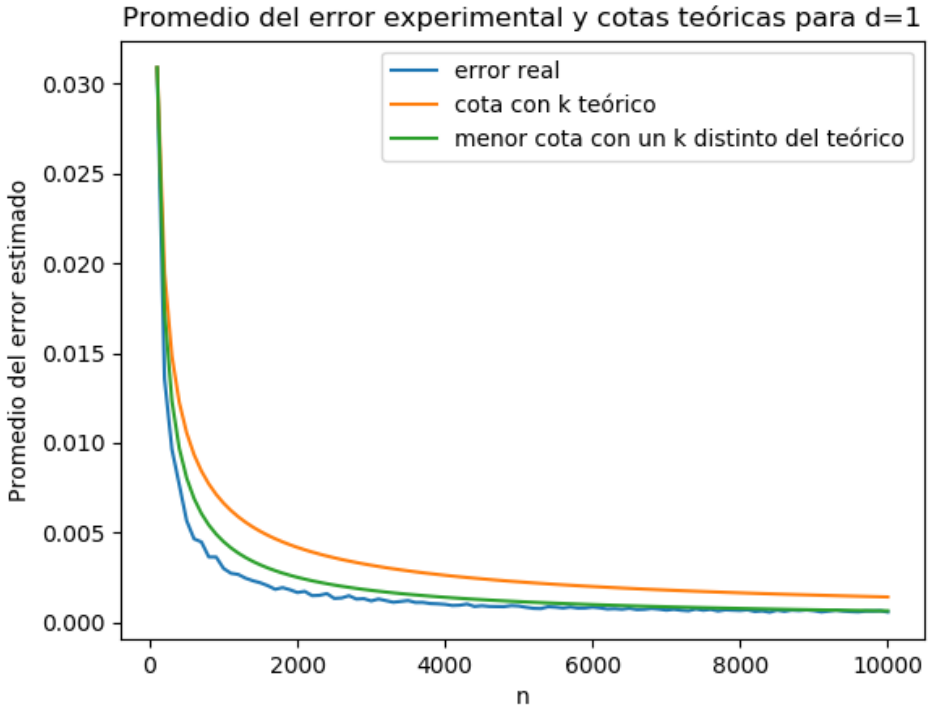
- $c(n^{-k})$ es mayor que todas las estimaciones $encontrarEError(n)$ (d y h_n son fijos)
- Los c y k elegidos deben ser tales que minimicen

$$\sum_{i=1}^{n} \left(c(i^{-k}) - encontrarEError(i)\right)^2$$

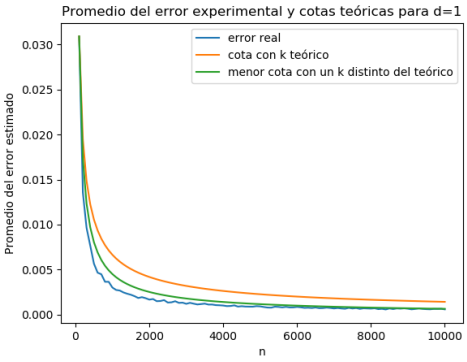
Así, la curva más ajustada a los datos (es decir, con c y k mejores que los que propone el teorema), debería cumplir que k sea mejor al propuesto por el teorema (el teorema indica $k = 2/(d+2)$) para corroborarlo.

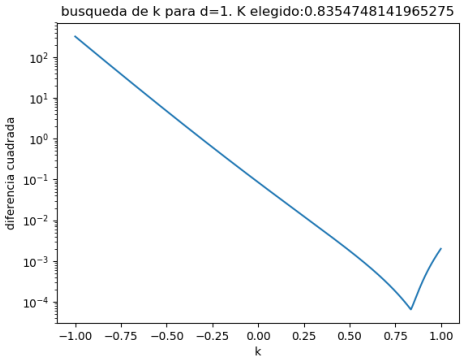
También se analizó la curva que cumple $k = 2/(d+2)$. En este caso simplemente se fijó k y se buscó sólo el c que cumpla las condiciones listadas arriba.

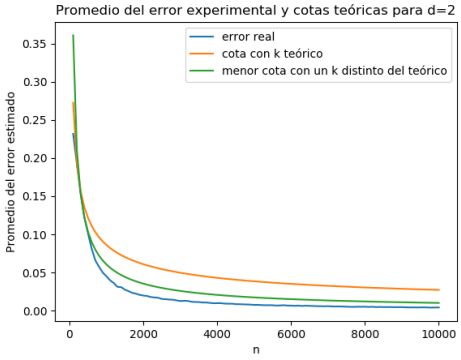
Esta prueba se repitió para $h_n = 0{,}5$ y $h_n = 0{,}1$.

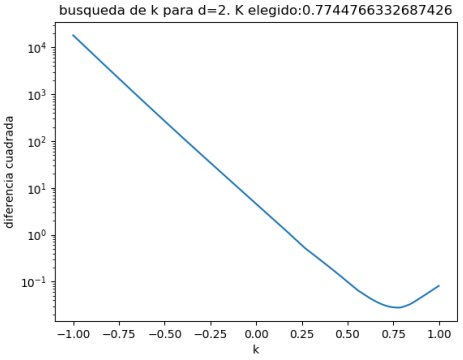


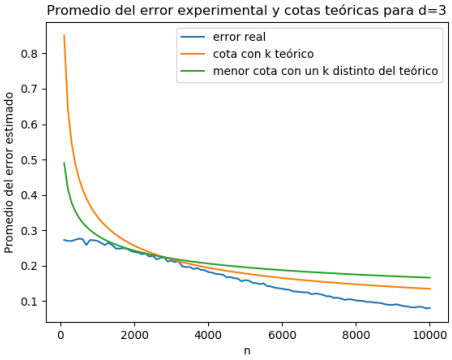
4. Primeros resultados: $h_n = 0{,}1$

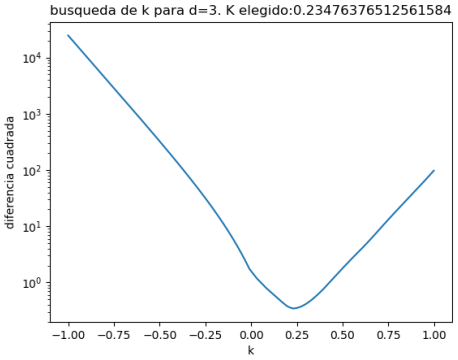


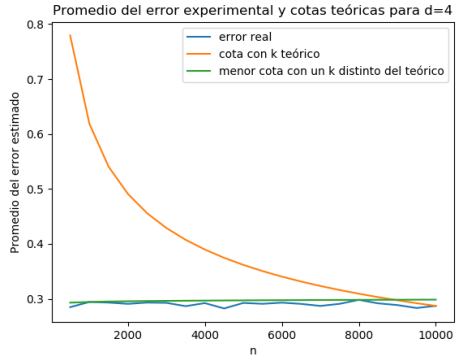


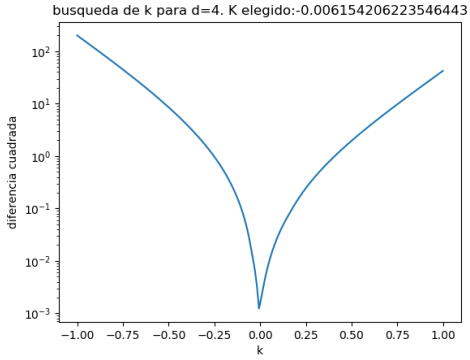


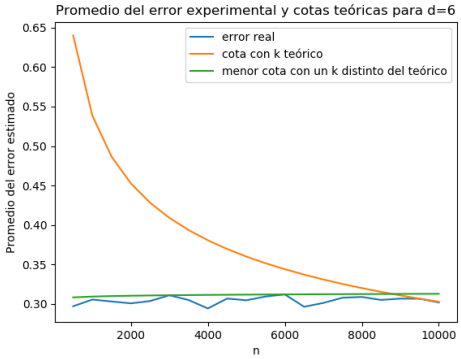


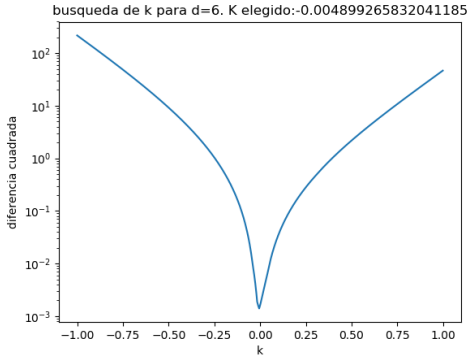


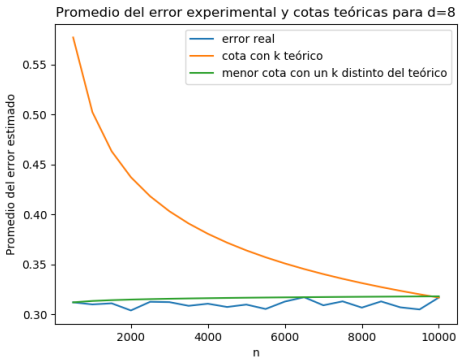


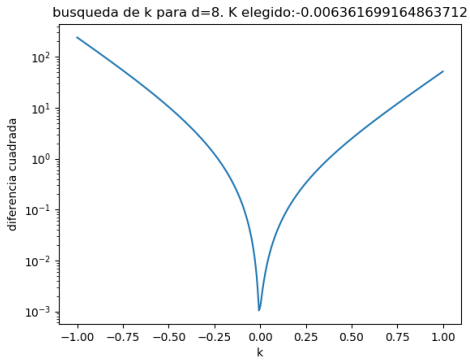

Promedio del error experimental y cotas teóricas para d=4
error real
cota con k teórico
menor cota con un k distinto del teórico
Promedio del error estimado
n

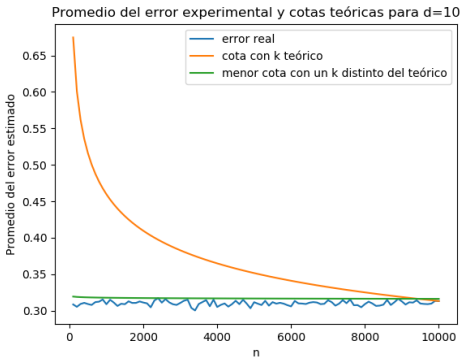

busqueda de k para d=4. K elegido:-0.006154206223546443
diferencia cuadrada
k

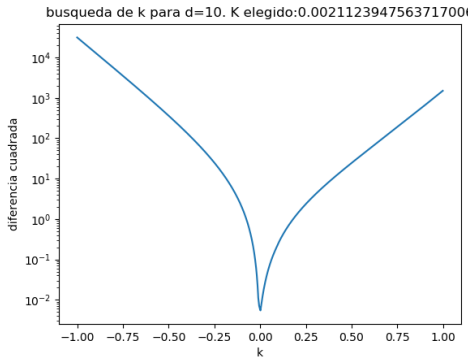

Promedio del error experimental y cotas teóricas para d=6
error real
cota con k teórico
menor cota con un k distinto del teórico
Promedio del error estimado
n

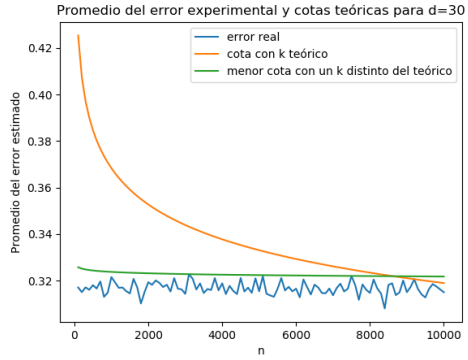

busqueda de k para d=6. K elegido:-0.004899265832041185
diferencia cuadrada
k

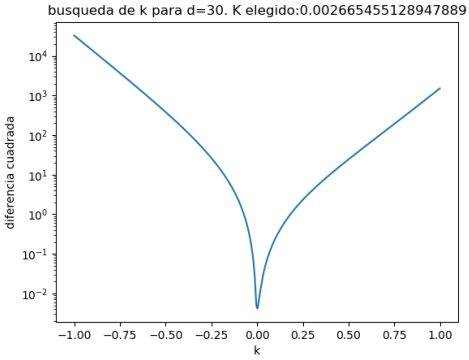

Promedio del error experimental y cotas teóricas para d=8
error real
cota con k teórico
menor cota con un k distinto del teórico
Promedio del error estimado
n

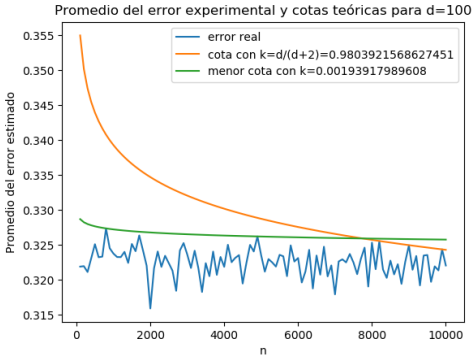

busqueda de k para d=8. K elegido:-0.006361699164863712
diferencia cuadrada
k

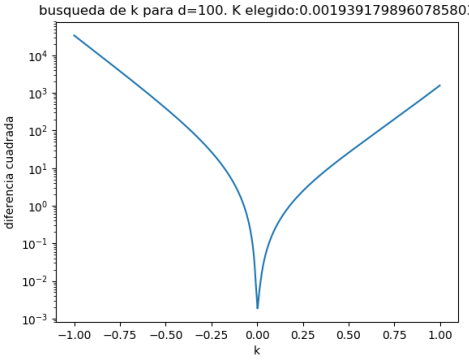

Promedio del error experimental y cotas teóricas para d=10
error real
cota con k teórico
menor cota con un k distinto del teórico
Promedio del error estimado
n

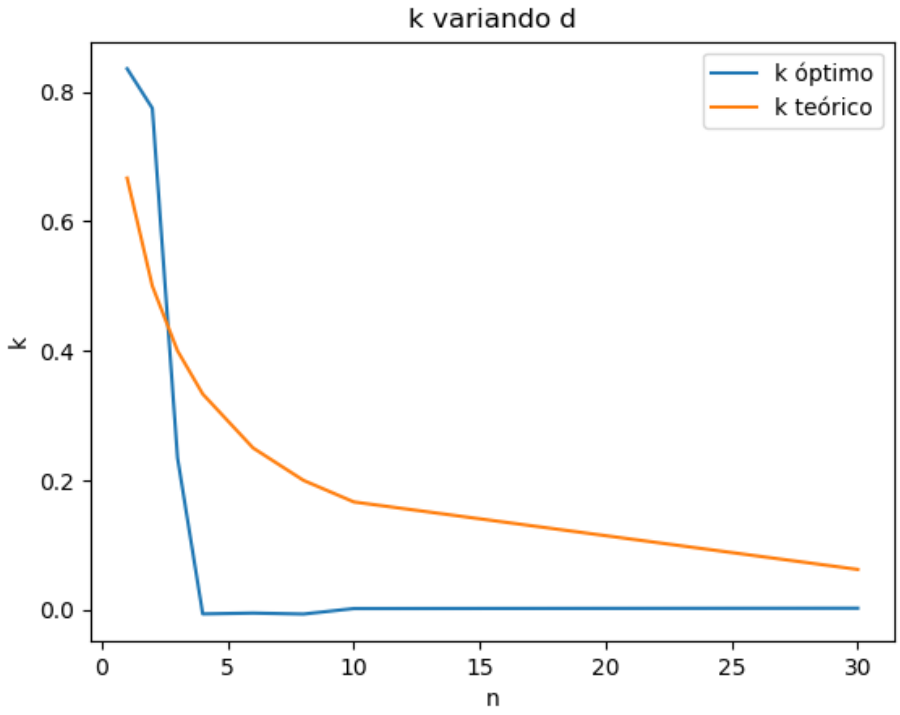

busqueda de k para d=10. K elegido:0.0021123947563717006
diferencia cuadrada
k

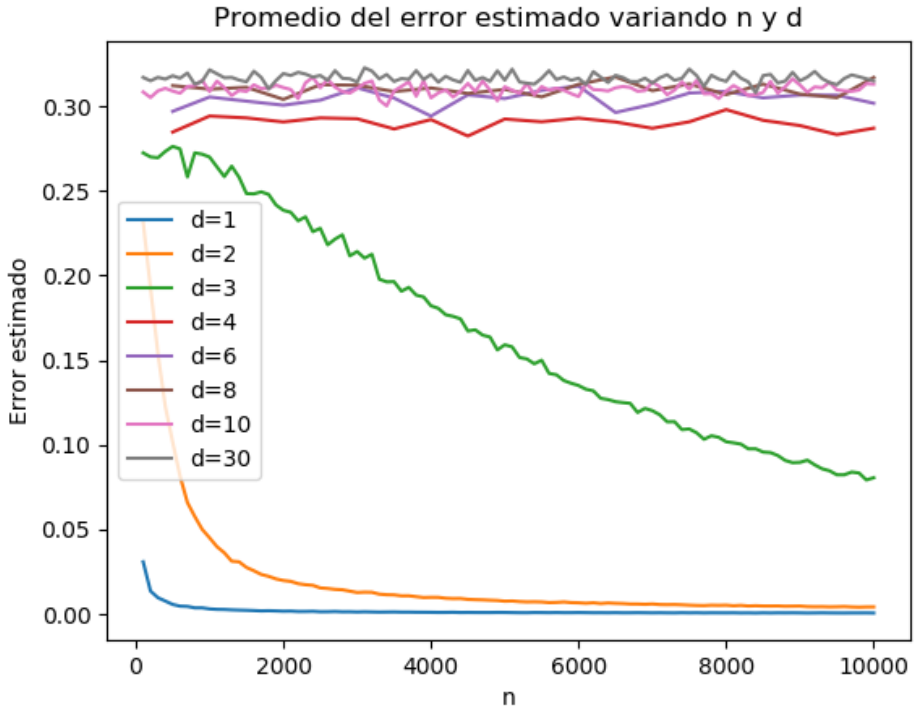

Promedio del error experimental y cotas teóricas para d=30
error real
cota con k teórico
menor cota con un k distinto del teórico
Promedio del error estimado
n


busqueda de k para d=30. K elegido:0.002665455128947889
diferencia cuadrada
k

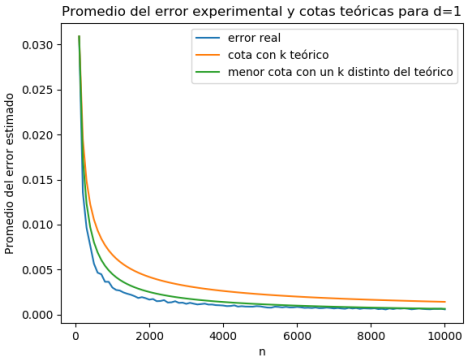

Promedio del error experimental y cotas teóricas para d=100
error real
cota con k=d/(d+2)=0.9803921568627451
menor cota con k=0.00193917989608
Promedio del error estimado
n

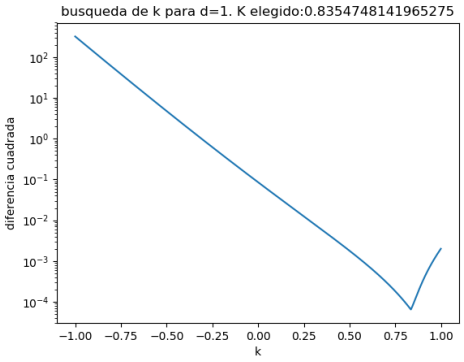

busqueda de k para d=100. K elegido:0.0019391798960785803
diferencia cuadrada
k

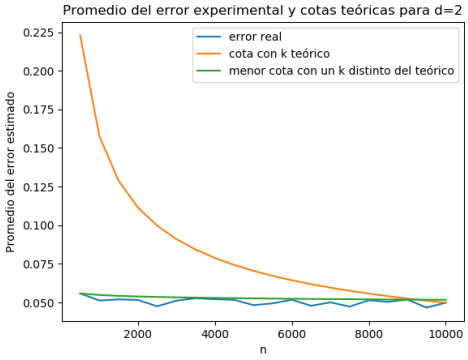

k variando d
k óptimo
k teórico
k
n
0.0
0.2
0.4
0.6
0.8
0
5
10
15
20
25
30

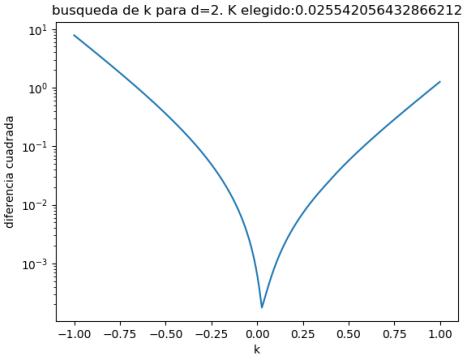

Promedio del error estimado variando n y d
Error estimado
n
d=1
d=2
d=3
d=4
d=6
d=8
d=10
d=30
0.00
0.05
0.10
0.15
0.20
0.25
0.30
0
2000
4000
6000
8000
10000

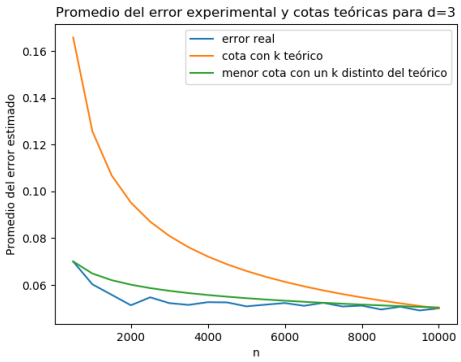
5. Primeros resultados: $h_n = 0{,}5$

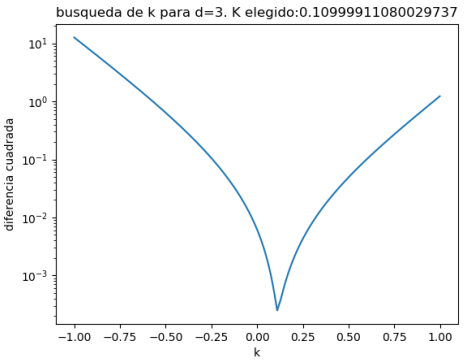


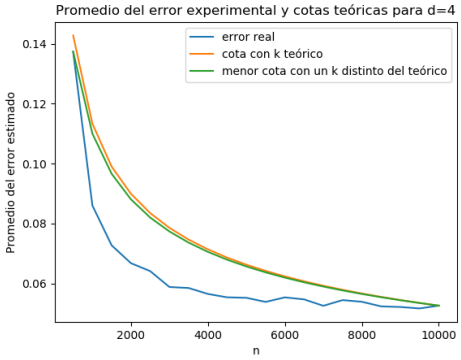


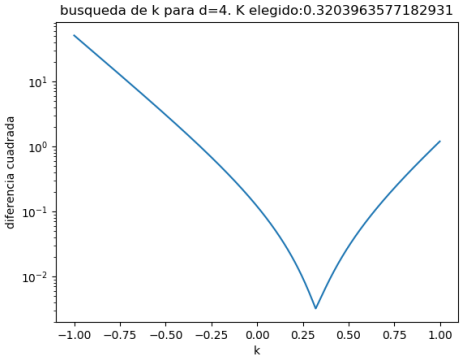


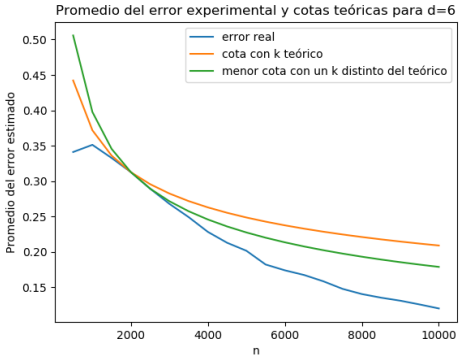


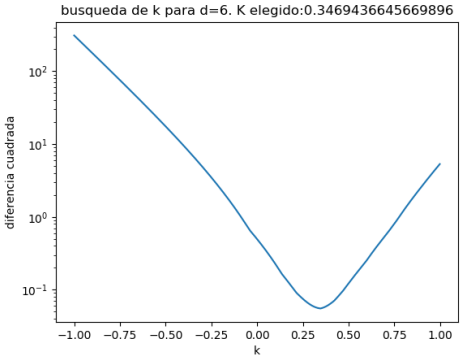


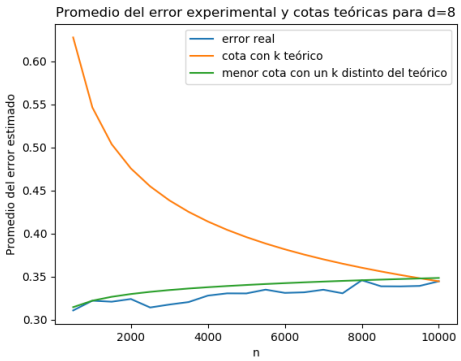


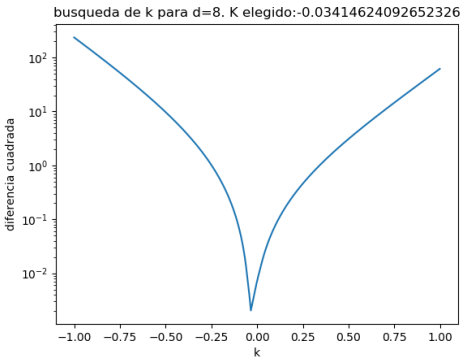

Promedio del error experimental y cotas teóricas para d=4
error real
cota con k teórico
menor cota con un k distinto del teórico
Promedio del error estimado
n

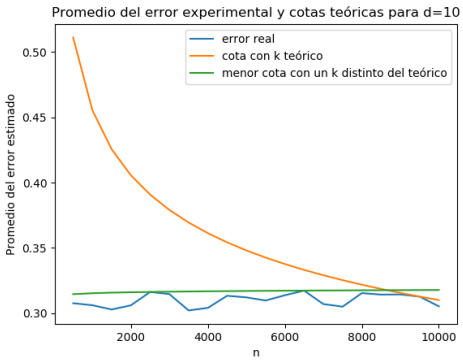

busqueda de k para d=4. K elegido:0.3203963577182931
diferencia cuadrada
k

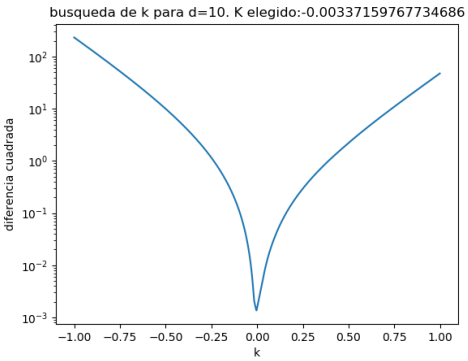

Promedio del error experimental y cotas teóricas para d=6
error real
cota con k teórico
menor cota con un k distinto del teórico
Promedio del error estimado
n

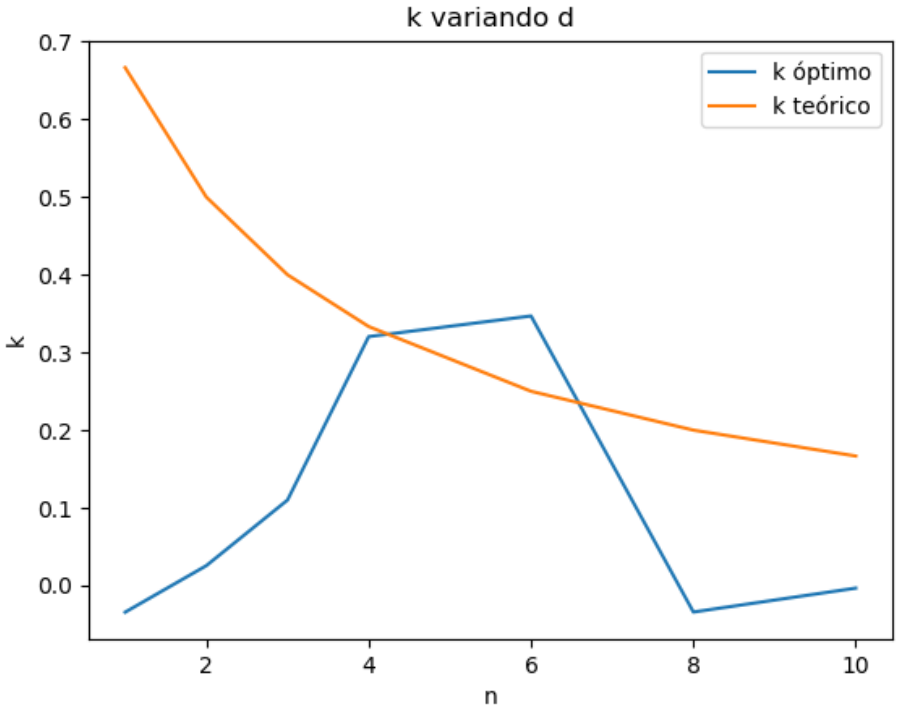

busqueda de k para d=6. K elegido:0.3469436645669896
diferencia cuadrada
k

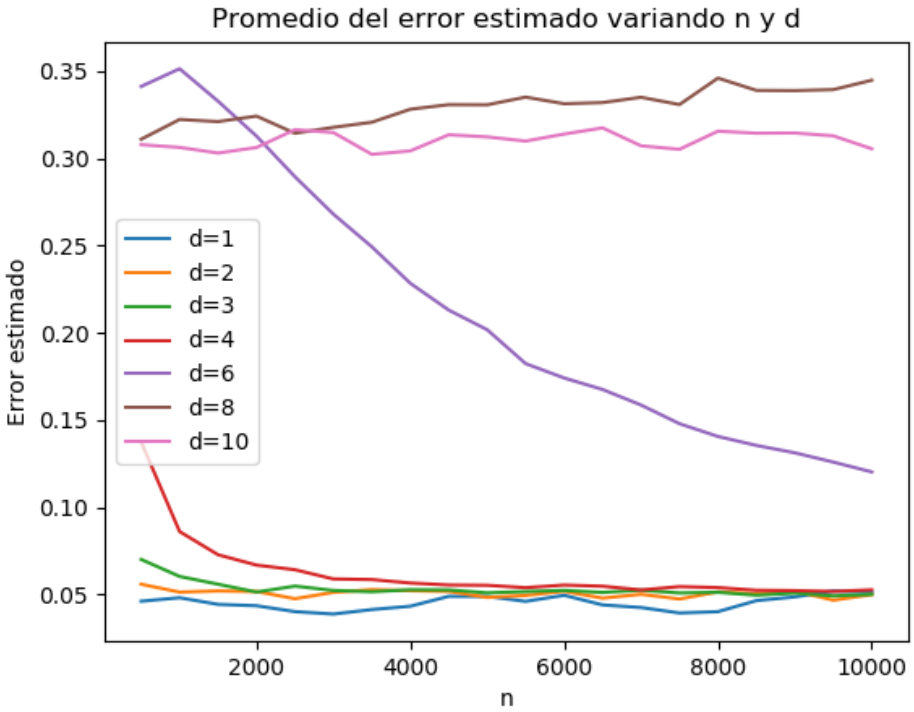

Promedio del error experimental y cotas teóricas para d=8
error real
cota con k teórico
menor cota con un k distinto del teórico
Promedio del error estimado
n


busqueda de k para d=8. K elegido:-0.034146240926523 26
diferencia cuadrada
k


Promedio del error experimental y cotas teóricas para d=10
error real
cota con k teórico
menor cota con un k distinto del teórico
Promedio del error estimado
n


busqueda de k para d=10. K elegido:-0.00337159767734686
diferencia cuadrada
k


k variando d
k óptimo
k teórico
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
k
2
4
6
8
10
n



6. Conclusiones para h_n constante

Para $h_n = 0{,}1$ se obtuvieron resultados buenos (que verifican) para $d = 1$, $d = 2$ y $d = 3$, pero para d superiores no se logró la verificación.

Esto es razonable cuando se tienen en cuenta las reglas usadas en la práctica çomún"de machine learning: al aumentar la cantidad de dimensiones y no subsanar esto con más datos se tiene underfitting.

En este caso en particular también es importante la maldición de la dimensionalidad: a medida que crece d, la cantidad de puntos a una distancia fija h_n cae (en realidad, el mismísimo significado de la distancia es el que se pierde: todos los puntos tienden a estar a una distancia muy similar unos de otros).

Aunque las reglas prácticas del machine learningz la maldición de la dimensionalidad explican los resultados, estos contradicen al teorema que se busca corroborar empiricamente (es decir, la interpretación que se había hecho del mismo).

Con el objetivo de observar el comportamiento del algoritmo en dimensiones altas, se utilizó $h_n = 0{,}5$ para realizar una nueva prueba.

Los resultados de esta prueba son muy importantes: para dimensiones

"medianas"($d = 4$, $d = 6$), con $h_n = 0{,}5$ se logra lo que no se puede con $h_n = 0{,}1$: se tiene una mejora gradual y se corrobora el teorema para $d = 6$.

Esto nos lleva a que h_n distintos funcionan bien para d distintos, con lo cual h_n debe estar relacionado a d. La necesidad de un h_n mayor para mayor d es consecuencia de la maldición de la dimensionalidad: a medida que crece d, la distancia promedio entre los puntos es mayor y un h_n que funciona para $d = 1$ no es significativo para d mayores.

El motivo por el cual no se pudieron observar empíricamente las consecuencias del teorema es que fue ignorada la condición sobre h_n que requiere la segunda conclusión del teorema:

$$h_n = c' \left(\frac{\sigma^2 + sup_{z \in S^*} |m(z)|^2}{C^2} \right)^{1/(d+2)} n^{\left(-\frac{1}{d+2}\right)}$$

Además no se tuvo en cuenta el teorema 5.1, que establece condiciones sobre h_n para que un estimador kernel tenga convergencia universal débil al variar d. Así, según este último teorema, un h_n constante sólo funcionará bien para n y d constante.

7. Introducción a la segunda parte

En la segunda parte se analiza el teorema 5.1 del Gyrofi, que establece condiciones para las cuales cualquier estimador kernel cuyo kernel cumple ciertas condiciones (llamadas *boxed kernel*), converge de forma universalmente débil. Siendo que el kernel naive cumple las condiciones de un *boxed kernel*, y que se utiliza un h_n que varía según lo determina el teorema, se espera que, tal estimador converja para cualquier d. Nuevamente se generan situaciones al azar variando n y d y se analizar n entre 100 y 1000, utilizando $d \in \{1, 2, 3, 4, 6, 8, 10\}$ y $h_n \in \{0{,}8548 \cdot (n^{(} - 1/4{,}054 \cdot d)); 10^{-1/d}\}$. Se realiza el mismo análisis que en el teorema 5.2, para observar si con las nuevas restricciones sobre h_n, éste también se cumple.

8. Descripción del segundo experimento

Es imposible fijar

$$h_n = c' \left(\frac{\sigma^2 + sup_{z \in S^*} |m(z)|^2}{C^2} \right)^{1/(d+2)} n^{\left(-\frac{1}{d+2}\right)}$$

ya que para eso sería necesario conocer c'.

Entonces se intenta corroborar el teorema 5.1, que establece:

> **Teorema 5.1:** Asumiendo que se tienen bolas $S_{0,r}$ de radio r y bolas $S_{0,R}$ de radio R centradas en el origen $(0 < r \leq R)$, y una constante $b > 0$ tal que [6]
>
> $$\mathbb{1}\{x \in S_{0,R}\} \geq K(x) \geq b\mathbb{1}\{x \in S_{0,r}\}$$
>
> (boxed kernel), y considérese el estimador kernel m_n si $h_n \to 0$ y $nh_n^d \to \infty$. Entonces el estimador kernel es debilmente y universalmente consistente.[7]

El kernel naive es un boxed kernel, por lo tanto, si se cumplen las condiciones sobre h_n que establece este teorema, se obtiene la consistencia debil.

Así, se llevaron adelante dos pruebas: una con h_n dependiendo de n y d, y otra en la cual depende sólo de d. Así, para la primera, se usó

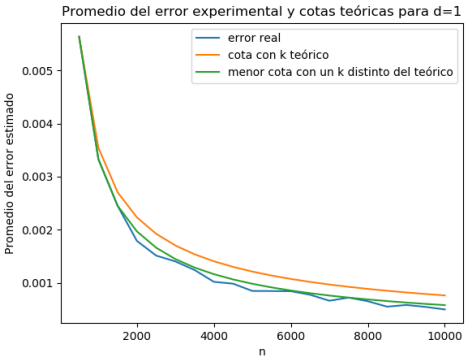
$$0{,}8548(n^{(-1/(4{,}054 \cdot d))})$$

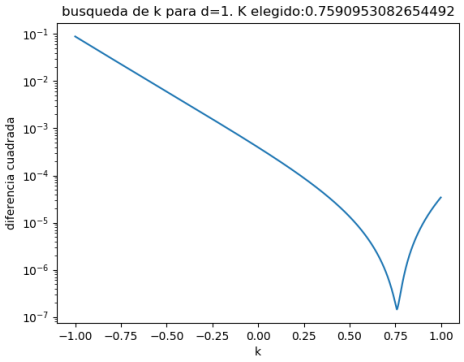
, lo cual cumple las condiciones y además cumple que h_n es aproximadamente 0,1 cuando $d = 1$ y $n = 8000$, y aproximadamente 0,5 cuando $d = 4$ y $n = 8000$. Para la segunda corrida de pruebas se utilizó $h_n = 10^{-1/d}$, elegido con el mismo criterio.

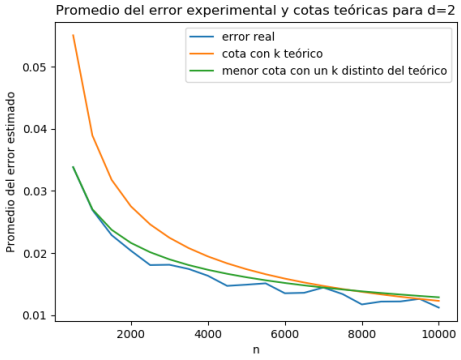
[6] Assume that there are balls $S_{0,r}$ of radius r and balls $S_{0,R}$ of radius R centered at the origin $(0 < r \leq R)$, and constant $b > 0$ such that

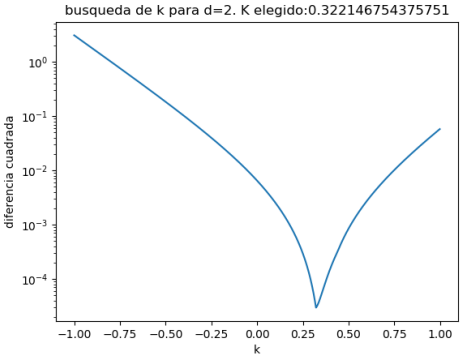
[7] and consider the kernel estimate m_n if $h_n \to 0$ and $nh_n^d \to \infty$, then the kernel estimate is weakly universally consistent.

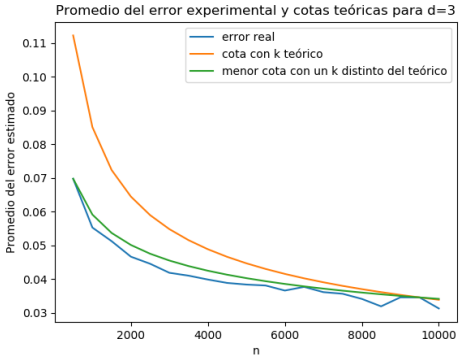
9. Resultados para h_n variando con d y n

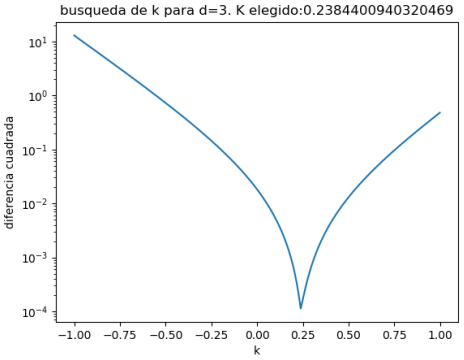


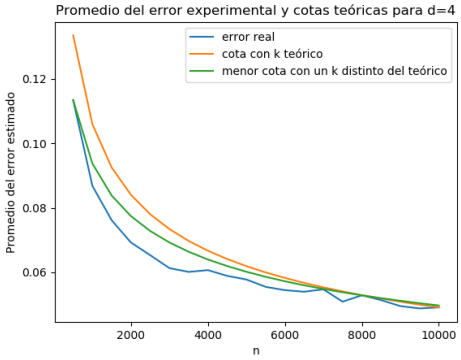


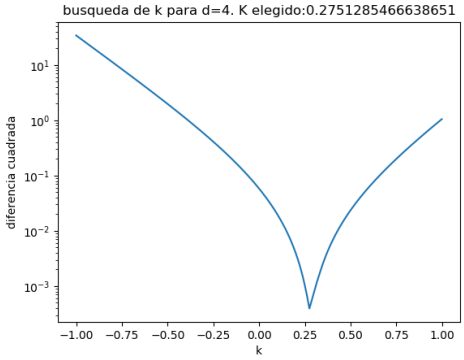


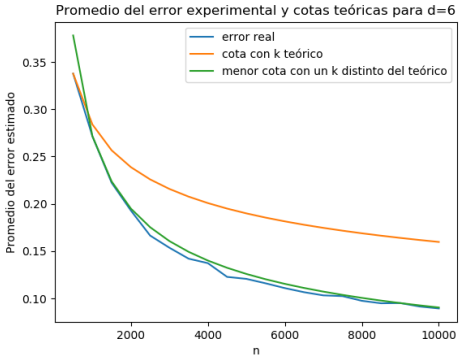


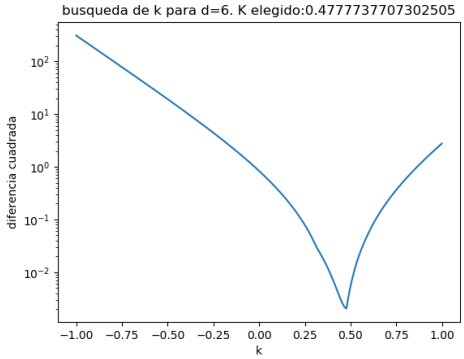


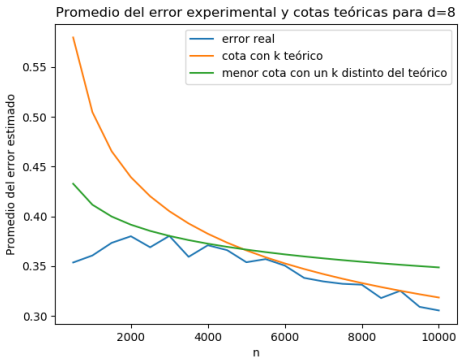


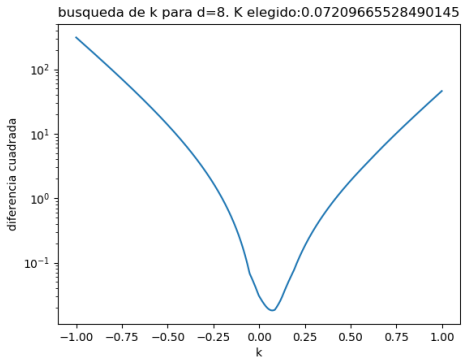

Promedio del error experimental y cotas teóricas para d=4
error real
cota con k teórico
menor cota con un k distinto del teórico
Promedio del error estimado
n

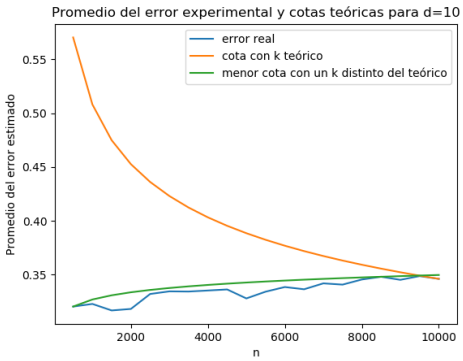

busqueda de k para d=4. K elegido:0.2751285466638651
diferencia cuadrada
k

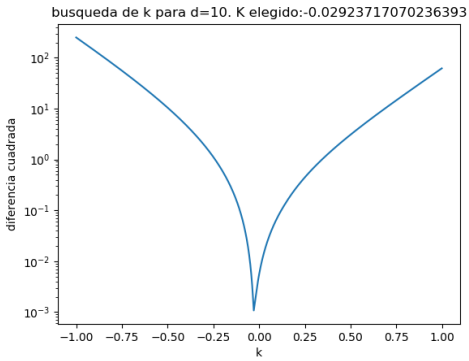

Promedio del error experimental y cotas teóricas para d=6
error real
cota con k teórico
menor cota con un k distinto del teórico
Promedio del error estimado
n

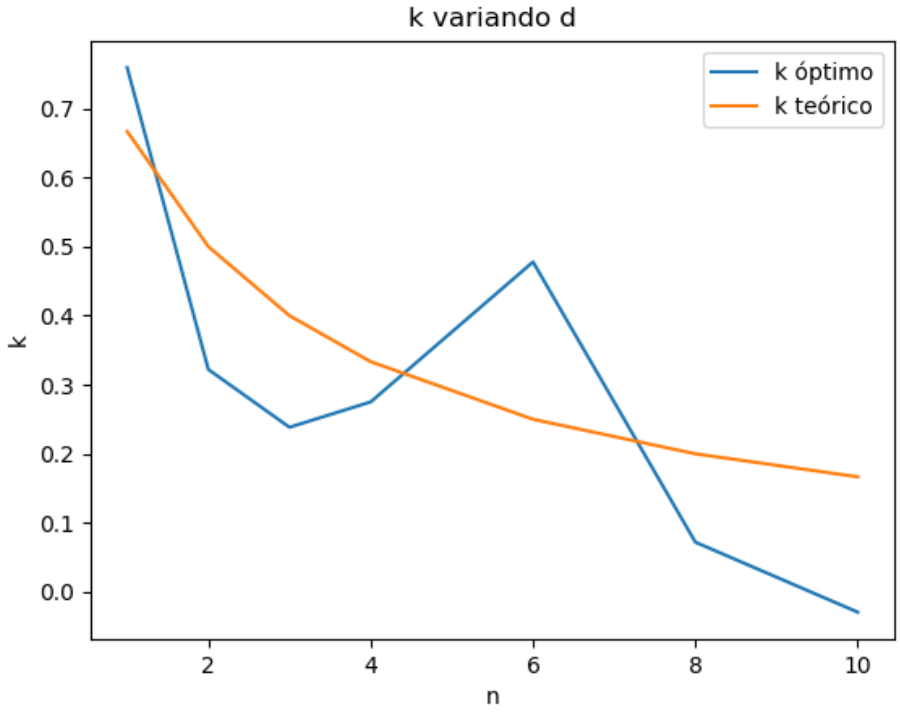

busqueda de k para d=6. K elegido:0.4777737707302505
diferencia cuadrada
k

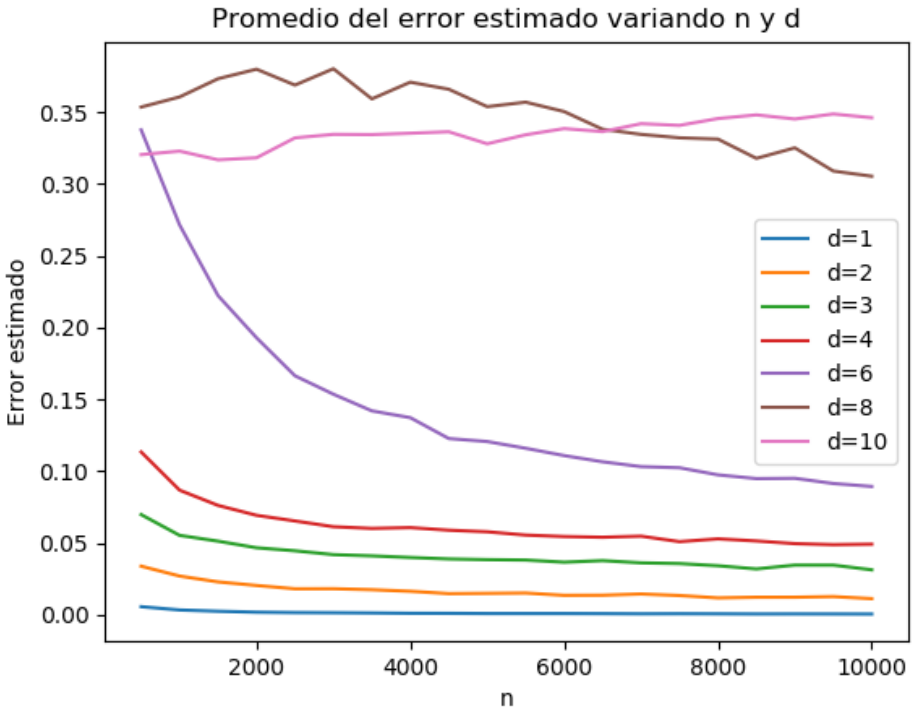

Promedio del error experimental y cotas teóricas para d=8
error real
cota con k teórico
menor cota con un k distinto del teórico
Promedio del error estimado
n


busqueda de k para d=8. K elegido:0.07209665528490145
diferencia cuadrada
k

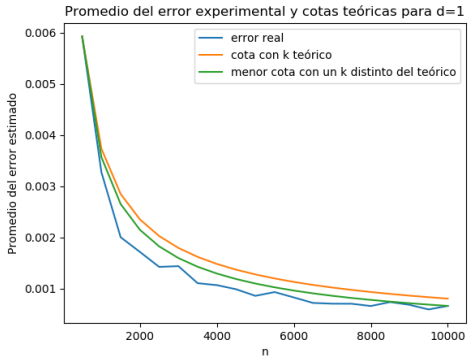

Promedio del error experimental y cotas teóricas para d=10
error real
cota con k teórico
menor cota con un k distinto del teórico
Promedio del error estimado
n

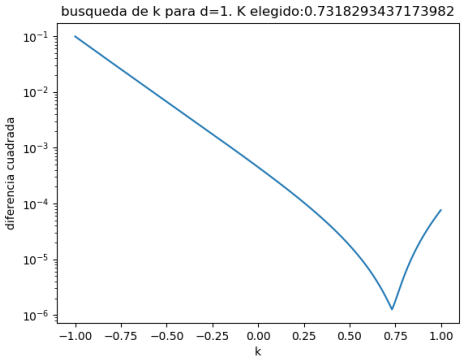

busqueda de k para d=10. K elegido:-0.02923717070236393
diferencia cuadrada
k

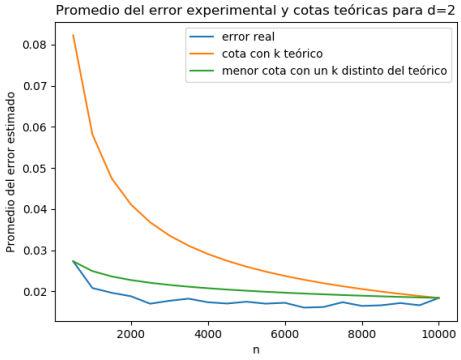

k variando d
k óptimo
k teórico
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
k
2
4
6
8
10
n

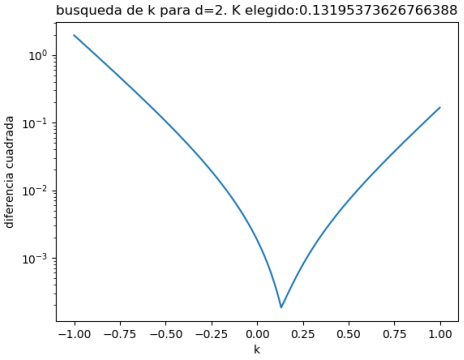

Promedio del error estimado variando n y d
Error estimado
0.00
0.05
0.10
0.15
0.20
0.25
0.30
0.35
2000
4000
6000
8000
10000
n
d=1
d=2
d=3
d=4
d=6
d=8
d=10

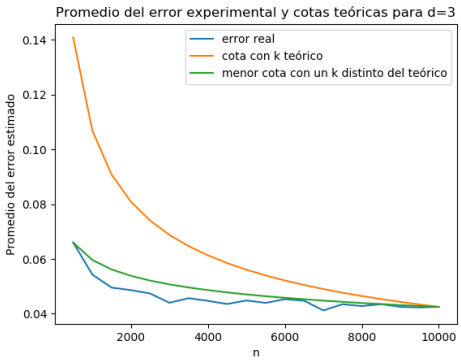
10. Resultados para h_n variando sólo con d

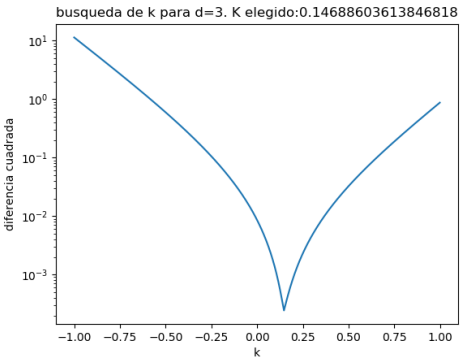


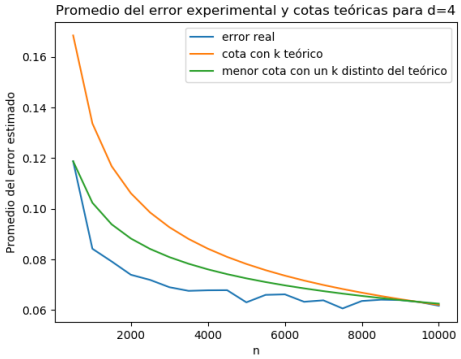


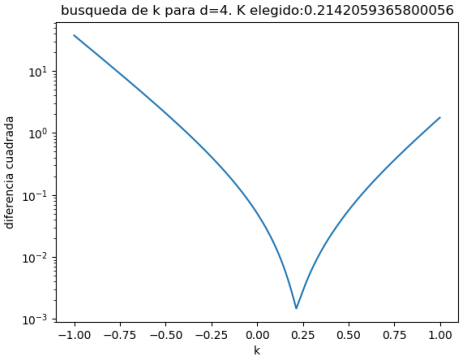


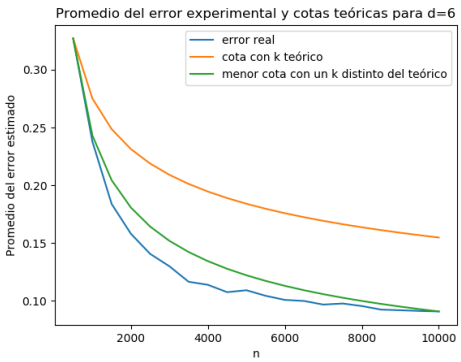


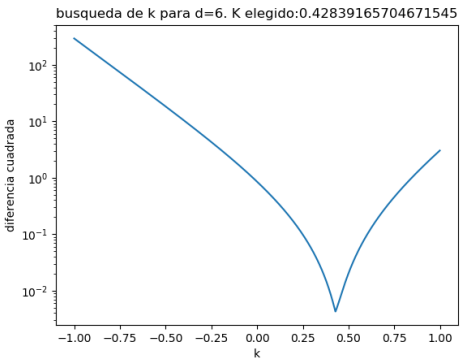


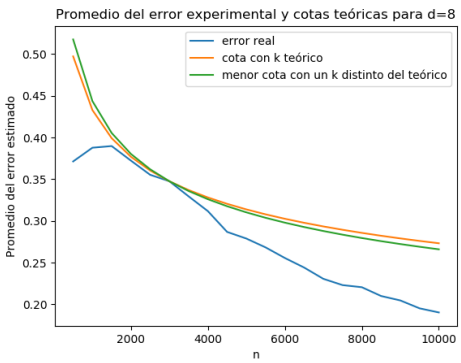


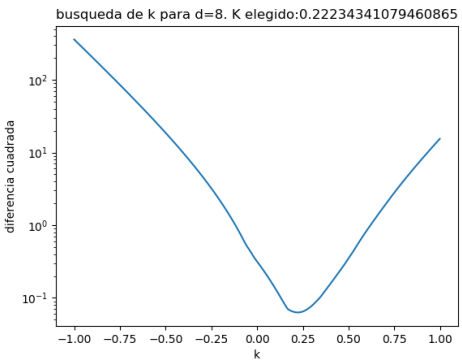

Promedio del error experimental y cotas teóricas para d=4
error real
cota con k teórico
menor cota con un k distinto del teórico
Promedio del error estimado
n

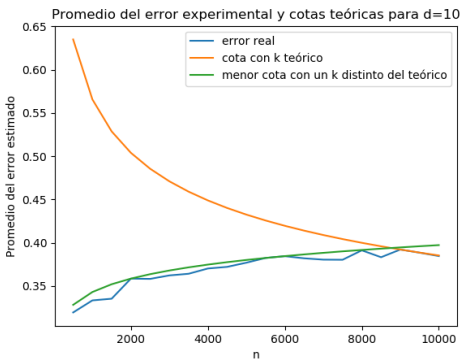

busqueda de k para d=4. K elegido:0.2142059365800056
diferencia cuadrada
k

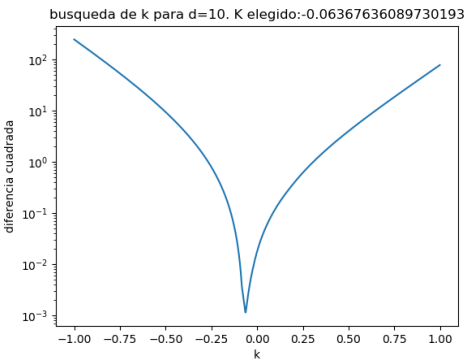

Promedio del error experimental y cotas teóricas para d=6
error real
cota con k teórico
menor cota con un k distinto del teórico
Promedio del error estimado
n

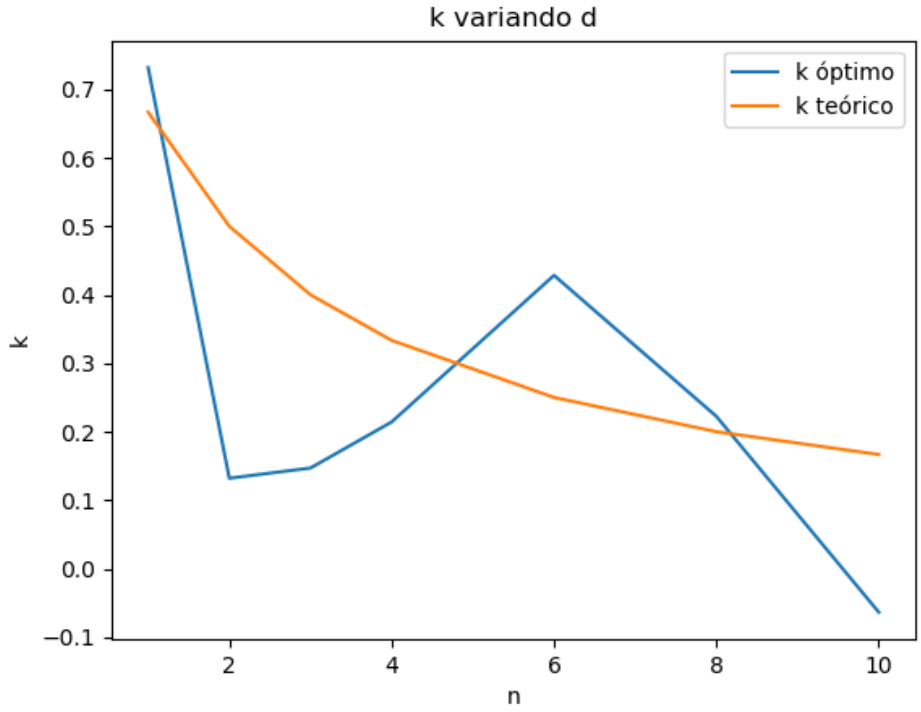

busqueda de k para d=6. K elegido:0.42839165704671545
diferencia cuadrada
k

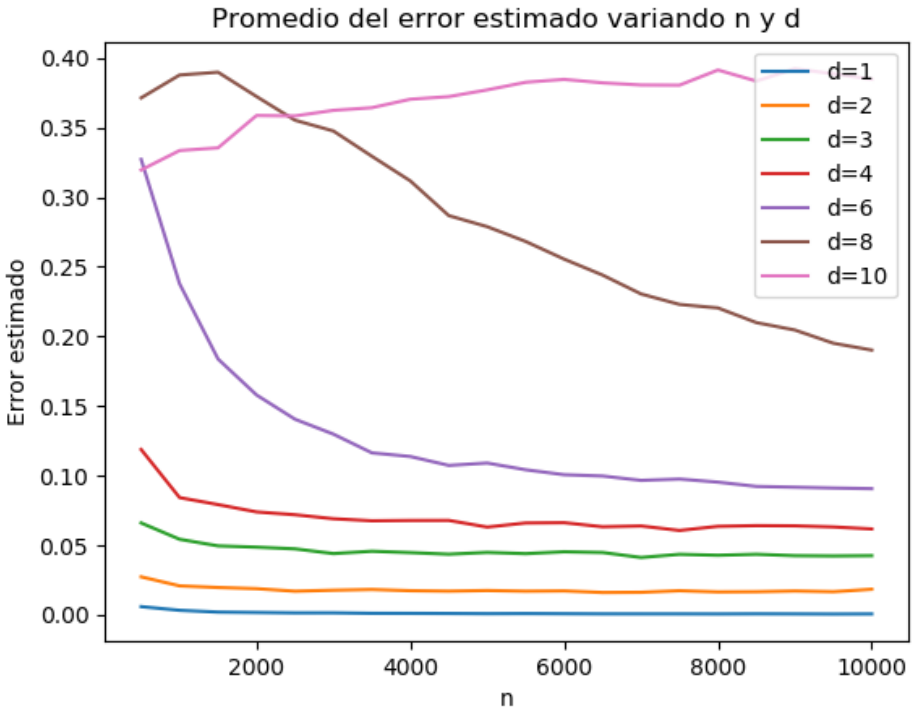

Promedio del error experimental y cotas teóricas para d=8
error real
cota con k teórico
menor cota con un k distinto del teórico
Promedio del error estimado
n


busqueda de k para d=8. K elegido:0.22234341079460865
diferencia cuadrada
k


Promedio del error experimental y cotas teóricas para d=10
error real
cota con k teórico
menor cota con un k distinto del teórico
Promedio del error estimado
n


busqueda de k para d=10. K elegido:-0.06367636089730193
diferencia cuadrada
k


k variando d
k óptimo
k teórico
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
−0.1
k
2
4
6
8
10
n



11. Conclusiones para h_n variable

En caso de h_n variable, se obtuvieron mejores resultados para todas las curvas: hubo aprendizaje para d entre 1 y 6, lo cual era de esperar debido al criterio utilizado para elegir las constantes, discutido previamente; y también hubo aprendizaje para $d = 8$, lo cual no se había observado para $h_n = 0{,}1$ ni $h_n = 0{,}5$. Esto es un indício fuerte para opinar que es posible corroborar el teorema 5.1.

El aprendizaje para $d = 8$ es muy notorio para h_n dependiente sólo de d, y menos notorio para h_n dependiente de n, lo cual podría resultar contradictorio al teorema.

El hecho de que para $d = 10$ no se dé el aprendizaje podría deberse a dos motivos:

1. Las constantes elegidas para las funciones $h_n(n, d)$ y $h_n(d)$ son incorrectas.

2. Las formas generales elegidas para $h_n(n, d)$ y $h_n(d)$ son incorrectas.

Otra posibilidad es que el teorema sea imposible de corroborar en altas dimensiones debido a la maldición de la dimensionalidad: que el aprendizaje sea demasiado lento como para observarlo.

12. Reconstrucción de la prueba del teorema 5.2

12.1. Esquema de la prueba

1. Descomponer $\mathbb{E}[(m_n(x) - m(x))^2|X_1, ..., X_n]$ en dos términos
2. Acotar el primer término aplicando esas definiciones
3. Acotar el segundo término utilizando la propiedad de Lipschitz
4. Utilizar la descomposición anterior para descomponer $\mathbb{E}||m_n - m||^2$
5. Acotar el primer término de la segunda descomposición
6. Acotar el segundo término de la segunda descomposición
7. Calcular la integral $\int_{S^*} \frac{1}{n\mu(S_{x,h_n})}\mu(dx)$
8. Aplicar las cotas y el resultado de la integral en la segunda descomposición

12.2. Anexo: Prueba del teorema 5.2

12.2.1. Descomponer $\mathbb{E}[(m_n(x) - m(x))^2|X_1, ..., X_n]$ en dos términos

$$\begin{aligned}\mathbb{E}[(m_n(x) - m(x))^2|X_1, ..., X_n] = \\ &= \mathbb{E}[(m_n(x) - \hat{m_n}(x))^2|X_1, ..., X_n] + (\hat{m_n}(x) - m(x))^2 \\ &= A + B\end{aligned}$$

$$\hat{m}_n(x) = \frac{\sum_{i=1}^n m(X_i)\mathbb{1}\{X_i \in S_{x,h_n}\}}{n\mu_n(S_{x,h_n})}$$

12.2.2. Acotar el primer término aplicando esas definiciones

Sean:

$$B_n(x) = \{n\mu_n(S_x, h_n) > 0\}$$

$$m_n(x) = \frac{\sum_{i=1}^n Y_i\mathbb{1}\{X_i \in S_{x,h_n}\}}{n\mu_n(S_{x,h_n})}$$

$$m(x) = \mathbb{E}[Y|X = x]$$

$$\mu_n(A) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}\{X_i \in A\}$$

Entonces, aplicando la segunda definición:

$$\begin{aligned}
A &= \mathbb{E}[(m_n(x) - \hat{m}_n(x))^2 | X_1, ..., X_n] \\
&= \mathbb{E}\left[\left(\frac{\sum_{i=1}^n (Y_i - m(X_i)) \mathbb{1}\{X_i \in S_{x,h_n}\}}{n\mu_n(S_{x,h_n})} \right)^2 | X_1, ..., X_n \right] \\
&= \mathbb{E}\left[\left(\frac{\sum_{i=1}^n (Y_i - \mathbb{E}[Y|X = X_i]) \mathbb{1}\{X_i \in S_{x,h_n}\}}{n\mu_n(S_{x,h_n})} \right)^2 | X_1, ..., X_n \right] \\
&= \mathbb{E}\left[\frac{\sum_{i=1}^n (Y_i - \mathbb{E}[Y|X = X_i])^2 \mathbb{1}\{X_i \in S_{x,h_n}\}}{(n\mu_n(S_{x,h_n}))^2} | X_1, ..., X_n \right] \\
&= \frac{\sum_{i=1}^n \mathbb{E}[(Y_i - \mathbb{E}[Y|X = X_i])^2] \mathbb{1}\{X_i \in S_{x,h_n}\}}{(n\mu_n(S_{x,h_n}))^2} \\
&= \frac{\sum_{i=1}^n Var(Y_i|X_i) \mathbb{1}\{X_i \in S_{x,h_n}\}}{(n\mu_n(S_{x,h_n}))^2} \\
&= Var(Y|X) \frac{\sum_{i=1}^n \mathbb{1}\{X_i \in S_{x,h_n}\}}{(n\mu_n(S_{x,h_n}))^2} \\
&= Var(Y|X) \frac{n\mu_n(S_{x,h_n})}{(n\mu_n(S_{x,h_n}))^2} \\
&= Var(Y|X) \frac{1}{n\mu_n(S_{x,h_n})} \mathbb{1}\{B_n(x)\} \\
&\leq \sigma^2 \frac{1}{n\mu_n(S_{x,h_n})} \mathbb{1}\{B_n(x)\}
\end{aligned}$$

Aclaración: el $\mathbb{1}\{B_n(x)\}$ final se agrega para que la última expresión pueda valer 0 en el caso de que para ningún i se cumpla $X_i \in S_{x,h_n}$, de esta forma se preserva la igualdad en ese caso.

Nota: en la demostración se asume que un único X_i puede pertenecer a un S_{x,h_n}, de esta forma se puede operar con el cuadrado tal como se hace en el desarrollo arriba (página 78 gyorfi).

12.2.3. Acotar el segundo término utilizando la propiedad de Lipschitz

La constante de Lipschitz C es la menor $c \in \mathbb{R}$ que cumple

$$|f(x_1) - f(x_2)| \leq c|x_1 - x_2|$$

Así, $|m(X_i) - m(x)| \leq Ch_n$ porque $|X_i - x| \leq h_n$.

Con lo cual, operando con el cuadrado tal como en la sección anterior, tenemos (página 78):

$$
\begin{aligned}
B &= (\hat{m}_n(x) - m(x))^2 \\
&= \left(\frac{\sum_{i=1}^n (m(X_i) - m(x)) \mathbb{1}\{X_i \in S_{x,h_n}\}}{n\mu_n(S_{x,h_n})} \right)^2 \mathbb{1}\{B_n(x)\} + m(x)^2 \mathbb{1}\{\text{ no } B_n(x)\} \\
&= \frac{\sum_{i=1}^n (m(X_i) - m(x))^2 \mathbb{1}\{X_i \in S_{x,h_n}\}}{n\mu_n(S_{x,h_n})} \mathbb{1}\{B_n(x)\} + m(x)^2 \mathbb{1}\{\text{ no } B_n(x)\} \\
&\leq \frac{\sum_{i=1}^n (Ch_n)^2 \mathbb{1}\{X_i \in S_{x,h_n}\}}{n\mu_n(S_{x,h_n})} \mathbb{1}\{B_n(x)\} + m(x)^2 \mathbb{1}\{\text{ no } B_n(x)\} \\
&= (Ch_n)^2 \frac{\sum_{i=1}^n \mathbb{1}\{X_i \in S_{x,h_n}\}}{n\mu_n(S_{x,h_n})} \mathbb{1}\{B_n(x)\} + m(x)^2 \mathbb{1}\{\text{ no } B_n(x)\} \\
&= (Ch_n)^2 \mathbb{1}\{B_n(x)\} + m(x)^2 \mathbb{1}\{\text{ no } B_n(x)\} \\
&\leq (Ch_n)^2 + m(x)^2 \mathbb{1}\{\text{ no } B_n(x)\}
\end{aligned}
$$

Nota: la primera descomposición en una suma se debe a que se prserva el valor de la igualdad cuando se cumple $B_n(x)$ y cuando no se cumple.

12.2.4. Utilizar la descomposición anterior para descomponer $\mathbb{E}||m_n - m||^2$

$$
\begin{aligned}
&\mathbb{E}||m_n - m||^2 \\
&\qquad = \mathbb{E}\left\{ \int (m_n(x) - m(x))^2 \mu(dx) \right\} \\
&\qquad = \mathbb{E}\left\{ \int_{S^*} (A + B) \mu(dx) \right\} \\
&\qquad = \mathbb{E}\left\{ \int_{S^*} A\mu(dx) \right\} + \mathbb{E}\left\{ \int_{S^*} B\mu(dx) \right\} \\
&\qquad = \mathbb{E}\left\{ \int_{S^*} \sigma^2 \frac{1}{n\mu_n(S_{x,h_n})} \mathbb{1}\{B_n(x)\} \mu(dx) \right\} + \\
&\qquad + \mathbb{E}\left\{ \int_{S^*} (Ch_n)^2 + m(x)^2 \mathbb{1}\{\text{ no } B_n(x)\} \mu(dx) \right\} \\
&\qquad = A' + B'
\end{aligned}
$$

12.2.5. Acotar el primer término de la segunda descomposición

Por la definición de $\mu_n(A)$ y la de $B_n(x)$

$$\begin{aligned}
A' &= \\
&= \mathbb{E}\left\{ \int_{S^*} \sigma^2 \frac{1}{n\mu_n(S_{x,h_n})} \mathbb{1}\{B_n(x)\}\mu(dx) \right\} \\
&= \sigma^2 \mathbb{E}\left\{ \int_{S^*} \frac{\mathbb{1}\{B_n(x)\}}{n(\frac{1}{n}\sum_{i=1}^n \mathbb{1}\{X_i \in S_{x,h_n}\})} \mu(dx) \right\} \\
&= \sigma^2 \mathbb{E}\left\{ \int_{S^*} \frac{\mathbb{1}\{\sum_{i=1}^n \mathbb{1}\{X_i \in S_{x,h_n}\} > 0\}}{\sum_{i=1}^n \mathbb{1}\{X_i \in S_{x,h_n}\}} \mu(dx) \right\}
\end{aligned}$$

$\sum_{i=1}^n \mathbb{1}\{X_i \in S_{x,h_n}\}$ es una variable aleatoria con una distribución binomial de parámetros $n = n$ y $p = \mathbb{P}(X \in S_{x,h_n}) = \mu(S_{x,h_n})$ por la definición de $\mu(A)$.

El lema 4.1 del libro indica que, siendo $B(n,p)$ una variable aleatoria de distribución binomial con parámetros n y p, vale que

$$\mathbb{E}[\frac{1}{B(n,p)}\mathbb{1}\{B(n,p) > 0\}] \leq \frac{2}{(n+1)p} \leq \frac{2}{np}$$

Aplicando este lema:

$$\begin{aligned}
A' &= \\
&= \sigma^2 \int_{S^*} \mathbb{E}\left\{ \frac{\mathbb{1}\{\sum_{i=1}^n \mathbb{1}\{X_i \in S_{x,h_n}\} > 0\}}{\sum_{i=1}^n \mathbb{1}\{X_i \in S_{x,h_n}\}} \right\} \mu(dx) \\
&\leq \sigma^2 \int_{S^*} \frac{2}{n\mu(S_{x,h_n})} \mu(dx) \\
&= 2\sigma^2 \int_{S^*} \frac{1}{n\mu(S_{x,h_n})} \mu(dx)
\end{aligned}$$

12.2.6. Acotar el segundo término de la segunda descomposición

$$
\begin{aligned}
B' &= \\
&= \mathbb{E}\Bigg\{ \int_{S^*} \Big((Ch_n)^2 + m(x)^2 \mathbb{1}\{ \text{ no } B_n(x)\} \Big) \mu(dx) \Bigg\} \\
&= \mathbb{E}\Bigg\{ \int_{S^*} \Big(m(x)^2 \mathbb{1}\{ \text{ no } B_n(x)\} \Big) \mu(dx) + \int_{S^*} (Ch_n)^2 \mu(dx) \Bigg\} \\
&= \mathbb{E}\Bigg\{ \int_{S^*} \Big(m(x)^2 \mathbb{1}\{ \text{ no } B_n(x)\} \Big) \mu(dx) + (Ch_n)^2 \int_{S^*} \mu(dx) \Bigg\} \\
&= \mathbb{E}\Bigg\{ \int_{S^*} \Big(m(x)^2 \mathbb{1}\{ \text{ no } B_n(x)\} \Big) \mu(dx) \Bigg\} + \mathbb{E}\Bigg\{ (Ch_n)^2 \int_{S^*} \mu(dx) \Bigg\} \\
&= \int_{S^*} \mathbb{E}\Bigg\{ m(x)^2 \mathbb{1}\{ \text{ no } B_n(x)\} \Bigg\} \mu(dx) + \mathbb{E}\Bigg\{ (Ch_n)^2 \cdot 1 \Bigg\} \\
&= \int_{S^*} m(x)^2 \mathbb{E}\Bigg\{ \mathbb{1}\{ \text{ no } B_n(x)\} \Bigg\} \mu(dx) + (Ch_n)^2
\end{aligned}
$$

A continuación se analiza $\mathbb{E}\Bigg\{ \mathbb{1}\{ \text{ no } B_n(x)\} \Bigg\}$:

$$
\begin{aligned}
\mathbb{E}\Bigg\{ \mathbb{1}\{ \text{ no } B_n(x)\} \Bigg\} &= \\
&= \mathbb{E}\Bigg\{ \mathbb{1}\{\mu_n(S_{x,h_n}) = 0\} \\
&= \mathbb{P}\Bigg\{ \mu_n(S_{x,h_n}) = 0 \Bigg\} \cdot 1 + \mathbb{P}\Bigg\{ \mu_n(S_{x,h_n}) > 0 \Bigg\} \cdot 0 \\
&= \mathbb{P}\Bigg\{ \sum_{i=1}^{n} \mathbb{1}\{X_i \in S_{x,h_n}\} = 0 \Bigg\} \\
&= \mathbb{P}\Bigg\{ Binomial(n, \mu(S_{x,h_n})) = 0 \Bigg\} \\
&= (1 - \mu(S_{x,h_n}))^n
\end{aligned}
$$

Con lo cual

$$\begin{aligned} B' &= \\ &= \int_{S^*} m(x)^2 \mathbb{E}\Big\{ \mathbb{1}\{ \text{ no } B_n(x)\} \Big\} \mu(dx) + (Ch_n)^2 \\ &= \int_{S^*} m(x)^2 (1 - \mu(S_{x,h_n}))^n \mu(dx) + (Ch_n)^2 \end{aligned}$$

A continuación se acota la integral:

$$\begin{aligned} B' &= \\ &\leq (Ch_n)^2 + \sup_{z \in S^*} m(z)^2 \int_{S^*} e^{-n\mu(S_{x,h_n})} \mu(dx) \\ &\leq (Ch_n)^2 + \sup_{z \in S^*} m(z)^2 \max_u u e^{-u} \int_{S^*} \frac{1}{n\mu(S_{x,h_n})} \mu(dx) \end{aligned}$$

12.2.7. Calcular la integral $\int_{S^*} \frac{1}{n\mu(S_{x,h_n})} \mu(dx)$

En el libro se indica que tal integral se calcula de la siguiente forma, admitiendo $S \subset S^*$ (página 76):

Se eligen $z_1, ..., z_{M_n}$ tales que la unión de $S_{z_1,h_n/2}, ..., S_{z_{M_n},h_n/2}$ cubre S, y

$$M_n \leq \frac{\tilde{c}}{h_n^d}$$

Siendo $\tilde{c}$ una constante.

Inicialmente se acota la integral sobre S con la suma de las integrales

sobre los $S_{z_i,h_n/2}$

$$\begin{aligned}\int_S \frac{1}{n\mu(S_{x,h_n})}\mu(dx) \leq \\ &\leq \sum_{j=1}^{M_n} \int_S \frac{\mathbb{1}\{x \in S_{z_j,h_n/2}\}}{n\mu(S_{x,h_n})}\mu(dx) \\ &= \sum_{j=1}^{M_n} \int_{S_{z_j,h_n/2}} \frac{1}{n\mu(S_{x,h_n})}\mu(dx) \\ &\leq \sum_{j=1}^{M_n} \int_{S_{z_j,h_n/2}} \frac{1}{n\mu(S_{z_j,h_n/2})}\mu(dx) \\ &= \sum_{j=1}^{M_n} \frac{1}{n\mu(S_{z_j,h_n/2})} \int_{S_{z_j,h_n/2}} \mu(dx)\end{aligned}$$

Por la definición de μ:

$$\begin{aligned}&= \sum_{j=1}^{M_n} \frac{1}{n\mu(S_{z_j,h_n/2})}\mu(S_{z_j,h_n/2}) \\ &= \sum_{j=1}^{M_n} \frac{1}{n} \\ &= \frac{M_n}{n} \\ &\leq \frac{\tilde{c}}{nh_n^d}\end{aligned}$$

12.2.8. Aplicar las cotas y el resultado de la integral en la segunda descomposición

$$\begin{aligned}A' + B' \leq \\ &\leq 2\sigma^2 \int_{S^*} \frac{1}{n\mu(S_{x,h_n})}\mu(dx) + \\ &+ (Ch_n)^2 + \\ &+ \sup_{z \in S^*} m(z)^2 \max_u u e^{-u} \int_{S^*} \frac{1}{n\mu(S_{x,h_n})}\mu(dx) \\ &\leq 2\sigma^2 \frac{\tilde{c}}{nh_n^d} + (Ch_n)^2 + \sup_{z \in S^*} m(z)^2 \max_u u e^{-u} \frac{\tilde{c}}{nh_n^d}\end{aligned}$$

Debido a que ue^{-u} es máximo cuando $u = 1$:

$$\begin{aligned} A' + B' \leq & \\ & \leq 2\sigma^2 \frac{\tilde{c}}{nh_n^d} + (Ch_n)^2 + \sup_{z \in S^*} m(z)^2 \frac{\tilde{c}}{nh_n^d} \\ & = \frac{\tilde{c}}{nh_n^d}(2\sigma^2 + \sup_{z \in S^*} m(z)^2) + (Ch_n)^2 \\ & = \tilde{c}\frac{2\sigma^2 + sup_{z \in S^*} m(z)^2}{nh_n^d} + (Ch_n)^2 \\ & \leq \hat{c}\frac{\sigma^2 + sup_{z \in S^*} m(z)^2}{nh_n^d} + (Ch_n)^2 \end{aligned}$$

Donde $\hat{c} = 2\tilde{c}$

12.2.9. Análisis de la segunda proposición del teorema

Al reemplazar el h_n propuesto en la expresión final alcanzada, se obtiene la segunda proposición del teorema, en la que $c'' = c' + \hat{c}(c')^{-d} = c' + 2\tilde{c}(c')^{-d}$.

13. Generación de las funciones m y s

- Se genera un número al azar entre 5 y 8, que luego se multiplica por d. Llámese esta cantidad p
- Se generan p puntos P_i al azar, pertenecientes a $D = [-1, 1]^d$.
- se generan p valores al azar $Q_i \in [-1, 1]$.
- Sea q el promedio del Q_i más alto y el más bajo.
- Se define:

$$g(x) = \frac{\sum_{i=1}^{p} Q_i K\left(\frac{x - P_i}{0,2}\right)}{\sum_{i=1}^{p} K\left(\frac{x - P_i}{0,2}\right)}$$

  Con K un kernel gaussiano que utiliza una gaussiana con $\sigma = 1$.
- La función f generada es:

$$f(x) = g(x) - q$$

14. Cálculo de la integral $\int_D \big(m(x) - m_n(x)\big)^2 dx$

Para calcular una integral se tienen tres alternativas:

- Encontrar la primitiva.
- Aplicar métodos numéricos
- Encontrar una estimación

14.1. Dificultades de encontrar la primitiva de $g(x) = \big(m(x) - m_n(x)\big)^2$

La función $g(x)$ puede escribirse como:

$$\left(\frac{\sum_{i=1}^{p} Q_i K_G\left(\frac{x-P_i}{0,2}\right)}{\sum_{i=1}^{p} K_G\left(\frac{x-P_i}{0,2}\right)} - q - \frac{\sum_{i=1}^{n} Y_i K_N\left(\frac{x-X_i}{h_n}\right)}{\sum_{i=1}^{n} K_N\left(\frac{x-X_i}{h_n}\right)} \right)^2$$

donde K_G es el kernel gaussiano y K_N es el kernel naive. Las dificultades de obtener la primitiva son:

- La función primitiva de e^{-x^2} se conoce como *función error* $(erf(x))$, y no se conoce su fórmula.
- Sería necesario obtener todos los sub conjuntos de D para los cuales se .ªctivan"de forma distinta los kernels. La parametrización de los límites de la integral sería muy compleja.
- Obtener la primitiva y realizar los cálculos a partir de la misma no implica que los mismos sean precisos. Así, es posible que la primitiva -si se pudiera encontrar- tenga una expresión muy compleja, de forma que la precisión del cálculo sea menor que la que se podría obtener por medio de un método numérico o una estimación.

14.2. Dificultades de calcular la integral por medio de métodos numéricos

Se analizaron diversas alternativas, poniendo mayor atención en los paquetes de cálculo numérico de integrales disponibles para python. Entre estos, se puso mayor atención en los provistos por scipy.

- La mayoría de los métodos provistos por scipy son para funciones de una dimensión

- Muchos métodos de integración requieren la derivada, la cual no puede ser calculada, ya que el kernel naive no puede derivarse.
- Los métodos numericos encontrados asumen que la función es contínua, lo cual no se da este caso, debido al kernel naive.
- Los métodos numéricos que ofrece scipy son para 1, 2 ó 3 dimensiones, no más.
- Muchos métodos numéricos requieren un muestreo o una grilla, con lo cual el tiempo consumido por el algoritmo es exponencial a la cantidad de dimensiones.

14.3. Razones para calcular la integral por medio de métodos montecarlo

- Se tiene una aproximación de la integral que o bien puede ser arbitrariamente precisa o bien puede tomar una cantidad arbitraria de tiempo. Aquí se decidió tomar una cantidad arbitraria de muestras: 1000.
- Es util para realizar un prototipo o análisis exploratorio de la situación, ya que requiere muy poco desarrollo.
- Indica la precisión aproximada alcanzada.

14.4. Método montecarlo utilizado

El método utilizado es muy sencillo y consiste en generar una cantidad fija de puntos aleatorios sobre D y calcular $g(x)$ sobre estos puntos. El promedio de los $g(x)$ calculados es el valor aproximado de la integral y la estimación de la varianza de los mismos conforma la estimación del error.