Universidad de Valladolid

Facultad de Ciencias

Trabajo Fin de Grado

Grado en Matemáticas

Introducción a la Regresión no Paramétrica

Autor: Jorge Martín Villafruela

Tutor/es: Eustasio del Barrio Tellado

Índice general

Índice de figuras

Resumen

Este trabajo estudia una clase de estimadores no paramétricos de la regresión: los estimadores locales polinomiales. Esta es una clase de estimadores para que se dispone de una forma eficiente de cálculo. Además, se prueba que estos estimadores son óptimos en tasa, para varias formas de medir el riesgo asociado.

El desarrollo de los resultados de optimalidad entra dentro de la Teoría Minimax en Estadística Matemática. En este trabajo se desarrolla esta teoría junto con las herramientas habituales en este campo relacionados con métricas estadísticas.

Abstract
This paper studies a category of non-parametric regression estimators: local polynomial estimators. These are estimators for which an efficient form of computation is available. Furthermore, it is proved that these estimators are optimal in rate, for various ways of measuring the associated risk.

The development of optimality results falls within the Minimax Theory in Mathematical Statistics. In this paper we develop this theory together with the usual tools in the field related to statistical metrics.

Capítulo 1

Introducción y motivación

En los estudios del Grado ya se ha estudiado el análisis de regresión bajo modelos paramétricos, como la Regresión Lineal. El problema de restringir la función de estimación a una clase "demasiado específica", es cuando la relación entre los atributos y las etiquetas se aleja mucho de cualquier función de la clase. Por ejemplo, supongamos que la relación entre la variable Y y la variable X está dada por

$$Y = \sin X + \epsilon$$

con ϵ una variable aleatoria centrada, independiente de X. Si se dispone de una muestra $(X_1, Y_1), \ldots, (X_n, Y_n)$ de réplicas del modelo anterior, la estimación obtenida a partir de la regresión lineal deja mucho que desear:

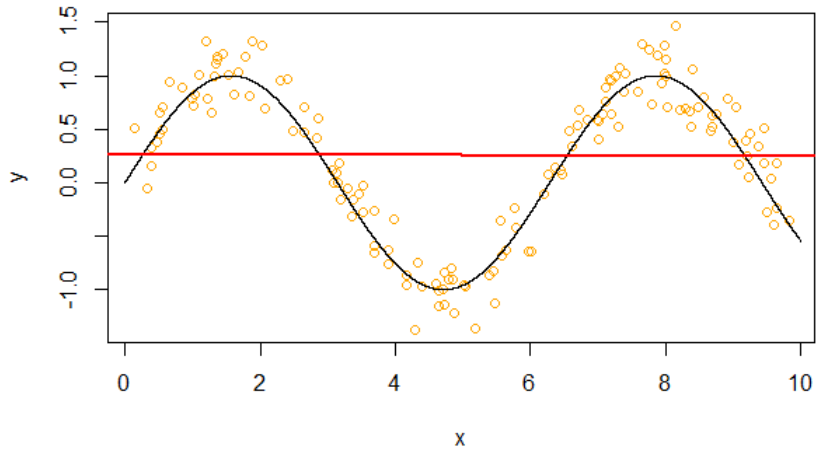


Figura 1.1: Estimación mediante la regresión lineal de la función $f(x) = \sin(x)$. Como se puede observar, el que el modelo se restringa a la funciones lineales hace que la estimación de la función difiera mucho de la regresión $f(x)$ a estimar.

Para evitar este problema, conviene trabajar dentro de una clase más grande, que reduzca el error cometido al restringirse a ella. En concreto, se considerarán clases suficientemente grandes como para no poderse definir a partir de un número finito de parámetros: **las clases no paramétricas**. A su vez, se tendrá que encontrar un estimador que aproxime correctamente las funciones de esa clase, de manera que la función de estimación sea buena aproximación de la regla. Esa idea de "buena aproximación" es dada por el concepto de **riesgo**.

A nivel teórico, cuando se mide el error mediante la pérdida cuadrática, la mejor función de la variable explicativa para predecir la respuesta es la esperanza condicionada (también denominada función de regresión). Para poder estimar esta función de forma directa será necesario disponer de réplicas en cada posible localización del vector de atributos, lo que no ocurre en la práctica. Por ello, se debe recurrir a la información aportada por las observaciones con atributos próximos

ponderando de forma apropiada su influencia. Esa idea llevará a estudiar el estimador de Nadaraya-Watson, relacionando de forma simple con el estimador núcleo de la densidad. Esto se describe a comienzos de la sección 2.2.

El estimador de Nadaraya-Watson puede ser mejorado. Una forma de hacerlo es mediante los estimadores locales locales polinomiales. A finales de la misma Sección 2.2, se describen estos estimadores, así como una forma eficiente de calcularlos numéricamente.

El objetivo principal de este trabajo es estudiar la calidad de los estimadores anteriores. Para poder afrontar este estudio a nivel teórico se procede en dos direcciones.

Por una parte buscamos cotas superiores para el riesgo. Estas cotas se darán de forma uniforme en clases suficientemente regulares de funciones de regresión. Esta uniformidad es importante si tenemos en cuenta que la verdadera función de regresión es desconocida.

Por otro lado, las cotas superiores para el riesgo, aunque sean satisfactorias, no excluyen la probabilidad de que existan mejores estimadores. Para eliminar (en lo esencial) esta posibilidad exploramos en este trabajo la Teoría Minimax. Esta teoría proporciona cotas inferiores para el error de estimación a partir del estudio de ciertas problemas de contraste. Además, en el desarrollo de este teoría son necesarios ciertos resultados relativos a métricas y divergencias habituales en Estadística Matemática. Todo este material se desarrolla en las secciones 2.3, 2.4 y 2.5.

Los resultados principales en este trabajo tienen que ver con la tasa de convergencia del estimador local polinomial de la regresión. Fijándonos en la pérdida cuadrática, se demuestra que, escogiendo el orden adecuado

$$\mathbb{E}_f \left\| \hat{f}_n - f \right\|_2^2 \leq C \frac{1}{n^{\frac{2\beta}{2\beta+1}}},$$

donde β es el orden de suavidad de la función de regresión f. Además, esta tasa es óptima: ningún estimador puede aproximarse más rápido a la función f.

Además del estudio teórico, en este trabajo se han implementado numéricamente varios estimadores y se ha comprobado su funcionamiento práctico. Este material se resume en la sección 3.2.

A lo largo del trabajo ha sido necesario recurrir a algunos resultados técnicos que divergen del objetivo principal. Por esta razón estos resultados se presentan en un Apéndice.

Capítulo 2

Resultados previos

En esta primera parte se obtendrán resultados de carácter general sobre el análisis de regresión (presentando las condiciones del modelo de regresión bajo el que se trabajará) y la Teoría Minimax. En la segunda parte, los resultados en esta parte se aplicarán para el caso específico de trabajar con el estimador local polinomial, también definido en esta parte.

2.1. El aprendizaje supervisado

La regresión no paramétrica es una especificación del problema planteado por el aprendizaje supervisado. Se comenzará dando unas nociones a nivel general sobre el entorno de este problema, para después concretar en el análisis de regresión; definiendo y justificando el modelo que se estudiará.

2.1.1. Nociones generales

El objetivo del aprendizaje automático es encontrar una función que relacione unos valores conocidos (llamados **atributos**) con uno desconocido, denominado **etiqueta**. A dicha función r se le llama **función de estimación** o **regla**.

Dentro del aprendizaje automático, el aprendizaje supervisado es la rama donde se parte de una muestra inicial $(\boldsymbol{X}_i, Y_i)$ de valores para obtener la función de estimación, llamada **conjunto de entrenamiento**.En el aprendizaje supervisado, el objetivo es encontrar una regla que aproxime el valor de la etiqueta para un atributo dado con ayuda del conjunto de entrenamiento.

Dependiendo de cómo sea el dominio de las etiquetas, se puede distinguir entre **clasificadores** y **regresiones**.

En caso en que haya finitos valores para las etiquetas, será una **clasificación**, denominada así puesto que el objetivo del problema es encontrar a qué categoría (asociada a los distintos posibles valores de la etiqueta) pertenece una entidad asociada con una serie de atributos. En caso de que lo atributos y las etiquetas tomen valores reales (es decir, $\boldsymbol{x} \in \mathbb{R}^d$ y $y \in \mathbb{R}$), se habla de **regresión**.

En cualquier caso, para determinar cuán buena es una regla r, se tiene que considerar una función que determine cuán cercano es el valor estimado $(r(\boldsymbol{x}))$ al real (y). La función que tome ese papel se denomina **función de pérdida** $l(r, \boldsymbol{x}, y)$.

Por otra parte, el conjunto de entrenamiento se puede considerar como una muestra aleatoria $(\boldsymbol{X}_i, Y_i)$ de variables i.i.d, siendo entonces el objetivo del aprendizaje supervisado acercase lo máximo posible a la función aleatoria que cumpla que $f(\boldsymbol{X}) = Y$.

Al considerar los atributos y las etiquetas como variables aleatorias, a la hora de medir la calidad de la predicción es conveniente hablar de la esperanza de la función de pérdida, que se le llama **riesgo**:

$$R(r) := \mathbb{E}\left(l(r, \boldsymbol{X}, Y)\right).$$

Aquellas reglas que minimizan el riesgo se las denomina **reglas de Bayes** y se denotarán $r_0 = f$.

No obstante, las distribuciones de $\boldsymbol{X}$ e Y son desconocidas, siéndolo por tanto también la distribución de $R(f)$. Esto requiere de que se tenga que obtener el riesgo de manera alternativa.

Se define al **riesgo empírico** como la media muestral de las funciones de pérdida, para cada elemento del conjunto de entrenamiento:

$$R_n(f) := \frac{1}{n}\sum_{i=1}^{n} l(f, \boldsymbol{X}_i, Y_i) \tag{2.1}$$

Puesto que la distribución de $((\boldsymbol{X}_i, Y_i))_{i=1}^n$ es la misma que la de las variables $(\boldsymbol{X}, Y)$, por la ley de los Grandes Números se cumple que el riesgo empírico converge hacia el riesgo. Es decir, considerando un conjunto de entrenamiento suficiente grande, $R_n(r)$ será suficientemente cercano a $R(r)$. Esto sugiere trabajar con el conjunto de entrenamiento para conseguir una función $\hat{f}_n$ que converja hacia el riesgo.

Puesto que el menor riesgo que se puede obtener es el de la regla de Bayes, se puede considerar el incremento del riesgo obtenido por considerar la regla r en vez de f, el exceso de riesgo. A diferencia del riesgo, el exceso de riesgo indica el aumento de riesgo debido no estar manejando el estimador óptimo, obviando los debido a causas ajenas al estimador elegido, como una mala selección de los atributos de la muestra, o que los atributos y las etiquetas sean variables incorreladas.

Pese a trabajar con el riesgo empírico en vez de con el riesgo, el conjunto de funciones de estimación que se pueden obtener a partir del conjunto de entrenamiento es muy amplio y variado como para poder calcular la regla que minimiza $R_n(r)$. Es por ello que en los problemas de regresión conviene restringir las funciones de estimación a únicamente considerar aquellas que pertenezcan a cierta clase $\mathcal{F}$. En esos casos, el objetivo es obtener aquella/s reglas que minimicen el riesgo empírico dentro de la clase $\mathcal{F}$:

$$r_{\mathcal{F}} = \arg\min_{r\in\mathcal{F}} R_n(r).$$

Obviamente, que $r_{\mathcal{F}}$ sea mejor o peor regla dependerá tanto de la clase $\mathcal{F}$, como del riesgo empírico. Esto se puede puede observar acotando el exceso de riesgo de $r_{\mathcal{F}}$:

$$\begin{aligned} R(r_{\mathcal{F}}) - R(f) &= R(r_{\mathcal{F}}) - R_n(r_{\mathcal{F}}) + R_n(r_{\mathcal{F}}) - R(r) + R(r) - R(f) \\ &\leq R(r_{\mathcal{F}}) - R_n(r_{\mathcal{F}}) + R_n(r) - R(r) + R(r) - R(f) \\ &\leq 2\sup_{r\in\mathcal{F}} |R_n(r) - R(r)| + R(r) - R(f) \qquad \forall\, r \in \mathcal{F} \end{aligned}$$

$$\Rightarrow R(r_{\mathcal{F}}) - R(f) \leq 2\sup_{r\in\mathcal{F}} |R_n(r) - R(r)| + \min_{r\in\mathcal{F}}(R(r) - R(f)). \tag{2.2}$$

Por un lado, se encuentra la diferencia obtenida entre el riesgo y el riesgo empírico dentro de $\mathcal{F}$, que se denomina **error de estimación**. Este error será mayor cuanto más grande sea $\mathcal{F}$, al contener $\mathcal{F}$ más reglas. Por otro lado, se encuentra el exceso de riesgo derivado de trabajar solamente con las reglas dentro de $\mathcal{F}$, que será mayor al ser $\mathcal{F}$ mas restringida. A este se le denomina **error de aproximación**. Una buena elección de $\mathcal{F}$ será aquella que encuentre un equilibrio entre ambos tipos de errores, ya que no siempre se podrá minimizar ambos errores.

2.1.2. Marco de la regresión

En caso de que el aprendizaje supervisado se realice sobre variables reales, entonces se suele denominar **regresión**. En los problemas de regresión, la función de pérdida normalmente utilizada es **pérdida cuadrática**, también denominado **error cuadrático**:

$$l(r, \boldsymbol{x}, y) = (y - r(\boldsymbol{x}))^2.$$

Cuando se considera la pérdida cuadrática, el riesgo recibe el nombre de error cuadrático medio. En ese caso, el riesgo es la esperanza condicionada.

Proposición 2.1.1. Sean $\boldsymbol{X}$ e Y variables aleatorias reales. Para el modelo de regresión de $(\boldsymbol{X}, Y)$, considerando la pérdida cuadrática, la regla de Bayes es la esperanza condicionada $f(\boldsymbol{x}) = \mathbb{E}(Y|\boldsymbol{X} = \boldsymbol{x})$.

Demostración.

Sea $r(\boldsymbol{x})$ una función cualquiera en $\mathbb{R}^d$.

Puesto que $\mathbb{E}(Y|\boldsymbol{X}=\boldsymbol{x})$ es una función en $\mathbb{R}^d$, basta con probar que para toda r se cumple que $R(r) > R(f)$.

Metiendo f en el riesgo de r y descomponiendo, se obtiene:

$$\begin{aligned} R(r) = \mathbb{E}(Y - r(\mathbf{X}))^2 &= \mathbb{E}\left[(Y - f(\mathbf{X})) + (f(\mathbf{X}) - r(\mathbf{X}))\right]^2 \\ &= \mathbb{E}\left[(Y - f(\mathbf{X})\right]^2 + \mathbb{E}\left[(f(\mathbf{X}) - r(\mathbf{X}))\right]^2 + 2\mathbb{E}\left[(Y - f(\mathbf{X}))(f(\mathbf{X}) - r(\mathbf{X}))\right]. \end{aligned}$$

Puesto que la esperanza condicionada tiene la propiedad de que, para cualquier función g de X se cumple que:

$$\mathbb{E}\left[\mathbb{E}(Y|\mathbf{X})g(\mathbf{X})\right] = g(\mathbf{X}) \cdot \mathbb{E}\left[\mathbb{E}(Y|\mathbf{X})\right] = g(\mathbf{X})\mathbb{E}\left[Y\right] \tag{2.3}$$

$$\Rightarrow \mathbb{E}\left[(Y - f(\mathbf{X}))(f(\mathbf{X}) - r(\mathbf{X}))\right] = 0,$$

el último término es nulo, y se obtiene:

$$\begin{aligned} \mathbb{E}(y - r(\mathbf{X}))^2 &= \mathbb{E}\left[(y - f(\mathbf{X})\right]^2 + \mathbb{E}\left[(f(\mathbf{X}) - r(\mathbf{X}))\right]^2 \\ &\geq \mathbb{E}\left[(y - f(\mathbf{X})\right]^2. \end{aligned}$$

Luego, $R(f) \leq R(r)$ para cualquier función r, así que $f(\boldsymbol{x}) = \mathbb{E}(Y|\boldsymbol{X}=\boldsymbol{x})$ es la función con menor error cuadrático posible, y por tanto la regla de Bayes. □

A partir del hecho de que $f(\boldsymbol{x}) = \mathbb{E}(Y|\boldsymbol{X}=\boldsymbol{x})$, se puede considerar la variable aleatoria

$$\epsilon := Y - \mathbb{E}(Y|\boldsymbol{X}=\boldsymbol{x}). \tag{2.4}$$

ϵ será entonces una variable aleatoria con $\mathbb{E}(\epsilon) = 0$ (puesto que $\mathbb{E}(\mathbb{E}(Y|\boldsymbol{X}=\boldsymbol{x})) = \mathbb{E}(Y)$), y varianza finita σ^2. Es habitual además considerar que esta variable sea normal; es decir, $\epsilon \sim \mathcal{N}(0, \sigma^2)$.

Por otra parte, puesto que la regla de Bayes se define puntualmente como $f(\boldsymbol{x}) = \mathbb{E}(Y|\boldsymbol{X}=\boldsymbol{x})$, no hay diferencia entre considerar $\boldsymbol{x}$ fijadas o condicionadas para aproximar la regla. Esto sugiere considerar una función de pérdida de la estimación a nivel puntual.

En este contexto, tiene sentido hablar del riesgo puntual de la estimación:

$$R(r, \boldsymbol{x}) := \mathbb{E}((Y - r(\boldsymbol{x}))^2) = \mathbb{E}((f(\boldsymbol{x}) - r(\boldsymbol{x}) + \epsilon)^2). \tag{2.5}$$

En el caso en el que se esté trabajando con la distancia puntual para un $\boldsymbol{x}$ fijo, si además se considera que los atributos del conjunto de entrenamiento son deterministas (es decir, los atributos $\boldsymbol{X}_i = \boldsymbol{x}_i$ no sean variables aleatorias, sino que pueden ser elegidos), se cumple que

$$Y = f(\boldsymbol{x}) + \epsilon,$$

siendo la variable aleatoria ϵ i.i.d a ϵ_i para todo i. El hecho de los atributos sean deterministas implica que la función de estimación $\hat{f}$ es independiente de ϵ, ya que $\hat{f}$ es construida a partir del conjunto de entrenamiento, cuya aleatoriedad proviene únicamente de las variables $(\epsilon_i)_{i=1}^n$.

Observación 2.1.1. La variable aleatoria ϵ cumple que es incorrelada respecto a la regla de Bayes f:

$$\begin{aligned} Cov(f(\boldsymbol{X}), \epsilon) &= \mathbb{E}(f(\boldsymbol{X})\epsilon) - \mathbb{E}(f(\boldsymbol{X}))\,\mathbb{E}(\epsilon) \underset{\mathbb{E}(\epsilon)=0}{=} \mathbb{E}(f(\boldsymbol{X})\,\epsilon) \\ &= \mathbb{E}\left(\mathbb{E}\left(Y|\boldsymbol{X}\right)\left(Y - E\left(Y|\boldsymbol{X}\right)\right)\right) = \mathbb{E}\left(Y\,\mathbb{E}\left(Y|\boldsymbol{X}\right) - \mathbb{E}\left(Y|\boldsymbol{X}\right)\,\mathbb{E}\left(Y|\boldsymbol{X}\right)\right) \\ &= \mathbb{E}\left(Y\,\mathbb{E}\left(Y|\boldsymbol{X}\right) - f(\boldsymbol{X})\,\mathbb{E}\left(Y|\boldsymbol{X}\right)\right) \underset{(2.3)}{=} \mathbb{E}\left(Y\,\mathbb{E}\left(Y|\boldsymbol{X}\right) - \mathbb{E}\left(\mathbb{E}\left(Y|\boldsymbol{X}\right)Y|\boldsymbol{X}\right)\right) \\ &= \mathbb{E}\left(Y\mathbb{E}\left(Y|\boldsymbol{X}\right) - \mathbb{E}\left(Y|\boldsymbol{X}\right)\right) = 0, \end{aligned}$$

donde las últimas igualdades se dan al cumplirse $\mathbb{E}\left(Y\mathbb{E}\left(Y|\boldsymbol{X}\right)\right) = \mathbb{E}\left(Y|\boldsymbol{X}\right)$.

Considerando lo anterior, el modelo de regresión para la pérdida cuadrática de $(\boldsymbol{X}, Y)$, siendo $\boldsymbol{X}$ e Y dos variables aleatorias reales, es

$$Y = f(\boldsymbol{X}) + \epsilon \tag{2.6}$$

siendo ϵ una variable aleatoria con $\mathbb{E}(\epsilon) = 0$ (puesto que $\mathbb{E}(\mathbb{E}(Y|\boldsymbol{X} = \boldsymbol{x})) = \mathbb{E}(Y)$), con varianza $\sigma^2 < \infty$ e incorrelada con $f(\boldsymbol{X})$ (y por tanto con $\boldsymbol{X}$).

Aunque la incorrelación no implica independencia, esto indica que trabajar considerando que $\boldsymbol{X}$ y ϵ son independientes no sea una condición muy exigente. Además, la independencia de esas variables ayudará al cálculo del riesgo de los estimadores.

Esto sugiere que, para generalizar el modelo con el que se trabaja a considerar unas condiciones menos restrictivas, se opte por tratar con que los atributos del conjunto de entrenamiento sean aleatorios, pero independientes de los errores ϵ_i del modelo.

Esta independencia de $\boldsymbol{x}$ sobre el conjunto de entrenamiento obtiene el siguiente resultado sobre el riesgo:

$$\begin{aligned} R(\hat{f}; \boldsymbol{x}) &= \mathbb{E}_f((f(\boldsymbol{x}) - \hat{f}(\boldsymbol{x}) + \epsilon)^2) = \mathbb{E}_f((f(\boldsymbol{x}) - \hat{f}(\boldsymbol{x}))^2 + 2\epsilon(f(\boldsymbol{x}) - \hat{f}(\boldsymbol{x})) + \epsilon^2) \\ &= \mathbb{E}_f((f(\boldsymbol{x}) - \hat{f}(\boldsymbol{x}))^2) + \mathbb{E}_f(\epsilon^2) = \mathbb{E}_f((f(\boldsymbol{x}) - \hat{f}(\boldsymbol{x}))^2) + \sigma^2 \\ &= sesgo_{\boldsymbol{x}}(\hat{f})^2 + Var(\hat{f}(\boldsymbol{x})) + \sigma^2. \end{aligned}$$

Es decir, $R(\hat{f}; \boldsymbol{x})$ se compone de una parte variable ($sesgo_{\boldsymbol{x}}(\hat{f})^2 + Var(\hat{f}(\boldsymbol{x}))$) y de una parte debida al propio riesgo de f. Esta parte depende de la distribución de ϵ, y es constante e inevitable. Esto sugiere que tratar como riesgo de $\hat{f}$ únicamente la parte variable; o lo que es lo mismo, comparar el estimador no con la etiqueta, sino con la regla de Bayes:

$$R(\hat{f}; \boldsymbol{x}) := \mathbb{E}_f((f(\boldsymbol{x}) - \hat{f}(\boldsymbol{x}))^2) = sesgo_{\boldsymbol{x}}(\hat{f})^2 + Var(\hat{f}(\boldsymbol{x})). \tag{2.7}$$

Además, esta consideración tiene sentido a nivel práctico. A la hora de la obtención de los resultados sobre el riesgo, habría que estar arrastrando la constante σ^2 continuamente, obteniendo una notación más llevadera; todo esto sin que se pierda nada a cambio, ya que no importa el valor del riesgo del estimador, sino su comparación con otros posibles estimadores.

Sin embargo, $R(r; \boldsymbol{x})$ es una medida local. Es por ello que también habrá que encontrar funciones de pérdida a nivel global donde también se pueda considerar esa acepción del riesgo, a la vez de que la regla de Bayes sea la esperanza condicionada.

Esto sugiere considerar la pérdida cuadrática, donde la que ya se ha visto que la regla de Bayes es la esperanza condicionada. Su equivalente del error cuadrático medio es el error cuadrático integrado medio, denotado por MISE (de *Mean Integrated Squared Error*):

$$MISE(r) = \mathbb{E}\left[\int (f - r)^2\right].$$

El MISE está relacionado con el error cuadrático medio. Al ser $(f_n - f)^2$ una función no negativa, se puede intercambiar el orden de la esperanza y de la integral (lema de Tonelli), resultando en:

$$MISE(r) = \int \mathbb{E}\left[(f - r)^2\right] = \int MSE(r).$$

Esto además da lugar a una descomposición sesgo-varianza similar a la obtenida con el error cuadrático medio

$$\begin{aligned} MISE(r) &= \int \left(\mathbb{E}\left(r(\boldsymbol{x})\right) - f(\boldsymbol{x})\right)^2 + Var(r(\boldsymbol{x}))\, d\boldsymbol{x} \\ &= \int sesgo_{\boldsymbol{x}}^2 + \int Var(r(\boldsymbol{x})). \end{aligned}$$

Otra función de pérdida que se puede considerar a nivel global para este propósito es el considerar la mayor entre todas las pérdidas puntuales del estimador. A esta función de pérdida se le denominará **pérdida del supremo**. Al ser una extensión directa de la distancia puntual, hereda directamente

todo lo obtenido para la norma puntual, así que es otra opción a considerar, donde el riesgo (respecto a la reglas de Bayes) será

$$R(r) = \mathbb{E}_f(\sup_{x_0 \in I} |f(x_0) - r(x_0)|^2).$$

Las distintas observaciones y suposiciones comentadas a lo largo de esta sección dan con el siguiente modelo:

Definición 2.1.1 (Modelo de estudio)**.** El modelo de regresión bajo el que se trabajará, cumple que

Condiciones 1 (Condiciones $\boldsymbol{C}$).

C1 Los atributos del conjunto de entrenamiento $(x_i)_{i=1}^n$ son unidimensionales (es decir, toman valores en $\mathbb{R}$) y deterministas; en concreto, se considerará que $x_i = i/n$.

C2 Existen $\epsilon_1, \dots \epsilon_n$ variables aleatorias normales i.i.d centradas con varianza σ^2 tales que

$$Y_i = f(x_i) + \epsilon_i. \tag{2.8}$$

Además, las funciones de pérdida bajo las que se trabajará son:

- La pérdida puntual:

$$|f(x_0) - g(x_0)|^2.$$

- La pérdida cuadrática:

$$\int |f(x) - g(x)|^2 \, dx. \tag{2.9}$$

- La pérdida del supremo:

$$\sup_{x \in I} |f(x) - g(x)|^2. \tag{2.10}$$

El objetivo dentro del modelo será el encontrar una función de estimación $\hat{f}_n$ cuyo riesgo (para esas funciones de pérdida) sea lo suficientemente pequeño dentro de una determinada clase $\mathcal{F}$. En concreto, tendrán mayor interés aquellos que puedan mantener un riesgo bajo trabajando en clases no paramétricas (clases de funciones que no pueden definirse a partir de un número numerable de parámetros).

2.2. Regresión no paramétrica

Durante la sección anterior se ha estado comentado el contexto del aprendizaje supervisado, centrándonos el caso concreto en el se trabaje con regresiones,y definiendo el modelo en el que se va trabajar. Bajo ese modelo, se conoce que la regla de Bayes es $f(x) = \mathbb{E}(Y|X = x)$.

Puesto que ya se sabe cuál es la regla de Bayes, una manera de obtener funciones de estimación será considerar distintas aproximaciones de f. Esto permite poder utilizar métodos que aproximen f a nivel local, haciendo que para estimar $f(x_0)$ para cierto valor x_0, cada etiqueta del conjunto de entrenamiento "influirá" en la estimación dependiendo la cercanía de su atributo relacionado a x_0. Esta "influencia" es modelizada por las funciones núcleo.

En concreto, se estudiarán los estimadores locales polinomiales. Como su nombre indica, estas reglas tratarán de estimar el valor de $f(x)$ por el polinomio local que más se asemeje a f en un determinado entorno. Este estimador será el que se estudie en la segunda parte del documento, en donde se observará como se comporta su riesgo dentro de una clase paramétrica determinada.

2.2.1. Estimador de Rosenblatt

Antes de comenzar con métodos de estimación de regresiones, vamos a comentar los estimadores de Rosenblatt (también llamados estimadores núcleo). Estos son estimadores de la función de densidad; sin embargo, ayudarán a comprender la elección de unos de los estimadores que se van a tratar: el estimador de Nadaraya-Watson.

Sean $X_1, X_2, \ldots, X_n$ i.i.d. con función de densidad p. Denotaremos con $F(x) = \int_{-\infty}^{x} p(t)\,dt$ su respectivas funciones de distribución. Podemos estimar $F(x)$ por $\hat{F}_n(x) = \frac{1}{n}\sum_{i=1}^{n} I(X_i \leq x)$ que, por la ley fuerte de los Grandes Números, converge a $F(x)$ casi seguro.

Asumiendo que la función de densidad p que se quiere estimar es continua (luego, la clase $\mathcal{F}$ en este caso serían las funciones de densidad continuas), se puede aplicar el Teorema del Cálculo Integral, y obtener que

$$p(x) = \lim_{h\to 0} \frac{1}{2h}\int_{x-h}^{x+h} p(z)dz = \lim_{h\to\infty} \frac{F(x+h) - F(x-h)}{2h}.$$

Luego la función de densidad p se puede estimar por el estadístico

$$\begin{aligned} \hat{p}_n^R =& \frac{F_n(x+h) - F_n(x-h)}{2h} \\ \Rightarrow \hat{p}_n^R(x) =& \frac{1}{2hn}\sum_{i=1}^{n} I(x-h \leq X_i \leq x+h) \\ =& \frac{1}{hn}\sum_{i=1}^{n} K_0\left(\frac{x - X_i}{h}\right) \end{aligned}$$

siendo $K_0(u) = \frac{1}{2}I(|u| \leq 1)$. Este estimador es el llamado **estimador de Rosenblatt**.
Cabe destacar el hecho de que $\hat{p}_n^R$ es una función de densidad, ya que, puesto que $\int K_0(x)\,dx = 1$, se tiene que

$$\begin{aligned} \int \hat{p}_n^R(z)dz &= \int \frac{1}{hn}\sum_{i=1}^{n} K_0\left(\frac{X_i - z}{h}\right)dz = \frac{1}{hn}\sum_{i=1}^{n}\int K_0\left(\frac{X_i - z}{h}\right)dz \\ &= \frac{1}{hn}\sum_{i=1}^{n} h = 1 \end{aligned}$$

Esto sugiere que, siempre y cuando K sea una función de densidad, el papel de la función K_0 pueda ser cumplido por cualquier otra función, obteniendo otro estadístico diferente pero que será igualmente una función de densidad, y convergerá a p.

Definición 2.2.1. Se dice que K es una **función núcleo** si $K : \mathbb{R} \to \mathbb{R}$ es una función integrable que satisface que $\int K(u)du = 1$.

Es decir, una función núcleo K es simplemente una función de densidad centrada en 0. Por ejemplo, la función K_0, usada en la definición del estimador de Rosenblatt, es una función núcleo. Algunos ejemplo de núcleos son (denotando por I a la función identidad):

- Núcleo rectangular: $K(u) = \frac{1}{2}I(|u| \leq 1)$
- Núcleo Gaussiano: $K(u) = \frac{1}{\sqrt{2\pi}}\exp\left(-u^2/2\right)$
- Núcleo triangular: $K(u) = (1 - |u|)\,I(|u| \leq 1)$
- Núcleo de Epanechnikov o parabólico: $K(u) = \frac{3}{4}\left(1 - u^2\right)I(|u| \leq 1)$

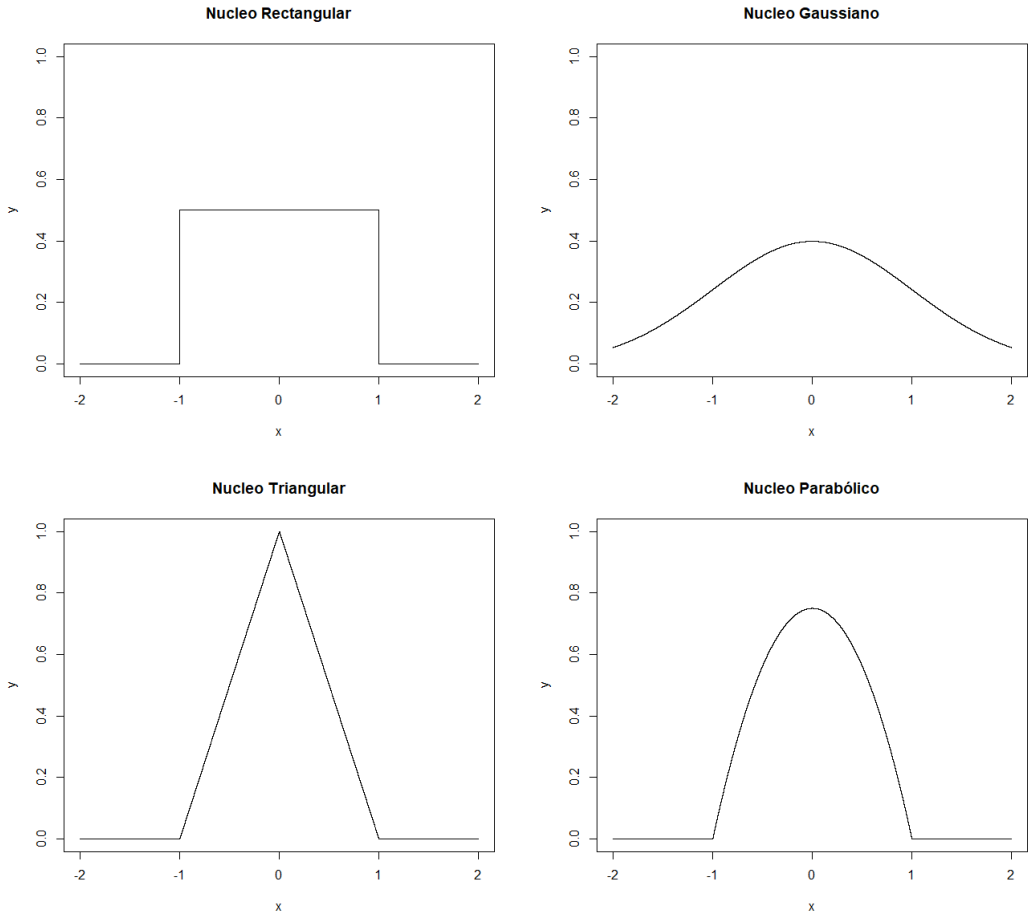


Figura 2.1: Gráficas de las funciones núcleo.

Las funciones núcleo convierten valores dados en funciones de densidad centradas. Una forma entonces de comprender el estimador de Rosenblatt es estimar el valor de la función de densidad en cierto punto x mediante el promedio de las densidades obtenidas de esta forma a partir de los valores de la muestra (centradas en x), haciendo que los elementos de la muestra $(X_i)_{i=1}^n$ suficientemente alejados de x (a grandes rasgos, eso será cuando la distancia entre el ancho de banda y x sea mayor que el ancho de banda h) no se tengan en cuenta. Por ello, también se suele denominar a estos estimadores **estimadores núcleo**.
La elección de un ancho de banda u otro cumple la misión considerar qué valores se tienen en cuenta a la hora de estimar cada valor.

Observación 2.2.1. El estimador de Rosenblatt puede generalizarse al caso multidimensional. Por ejemplo, para $\mathbb{R}^2$, el estimador sería, a partir del conjunto de entrenamiento $((X_i, Y_i))_{i=1}^n$:

$$\hat{p}_n(x,y) = \frac{1}{nh^2}\sum_{i=1}^{n} K\left(\frac{x-X_i}{h}\right) K\left(\frac{Y_i-y}{h}\right).$$

A partir de los estimadores núcleo, se puede construir un estimador de la regla de Bayes. Puesto que la regla de Bayes es $f(x) = \mathbb{E}(Y|X = x)$, se puede expresar a partir de las funciones de densidad $p_{X,Y}(x,y)$ y $p_X(x)$:

$$\mathbb{E}(Y|X=x) = \int y \frac{p_{X,Y}(x,y)}{p_X(x)} dy.$$

Una posible estimación de f será el considerar los estimadores núcleo de dichas funciones de densidad, $\hat{p}_n(x,y)$ y $\hat{p}_n(x)$[1]:

[1]Considerando $\hat{f}_n(x) = 0$ cuando $K\left(\frac{x-X_i}{h_n}\right)$).

$$\hat{f}_n(x) = \frac{\int y\hat{p}_n(x,y)dy}{\hat{p}_n(x)} = \frac{\int \frac{1}{nh^2}\sum_{i=1}^n yK\left(\frac{x-X_i}{h_n}\right)K\left(\frac{Y_i-y}{h_n}\right)dy}{\frac{1}{nh}\sum_{i=1}^n K\left(\frac{x-X_i}{h_n}\right)}$$

Al ser K una función núcleo, K es la función de densidad de cierta variable aleatoria Z centrada. Eso implica que $\frac{1}{h}K\left(\frac{\mu - Z}{h}\right)$ será la función de densidad de la variable aleatoria $h\,Z + \mu$.

Aplicándolo a la igualdad anterior, se obtiene la siguiente expresión del estimador:

$$\begin{aligned}\hat{f}_n(x) &= \frac{\int \frac{1}{nh^2}\sum_{i=1}^n yK\left(\frac{x-X_i}{h_n}\right)K\left(\frac{Y_i-y}{h_n}\right)dy}{\frac{1}{nh}\sum_{i=1}^n K\left(\frac{x-X_i}{h_n}\right)} = \frac{\frac{1}{nh}\sum_{i=1}^n K\left(\frac{x-X_i}{h_n}\right)\int \frac{1}{h}y\,K\left(\frac{Y_i-y}{h_n}\right)dy}{\frac{1}{nh}\sum_{i=1}^n K\left(\frac{x-X_i}{h_n}\right)} \\ &= \frac{\sum_{i=1}^n K\left(\frac{x-X_i}{h_n}\right)\mathbb{E}(hZ - Y_i))}{\sum_{i=1}^n K\left(\frac{x-X_i}{h_n}\right)} = \frac{\sum_{i=1}^n Y_i K\left(\frac{x-X_i}{h_n}\right)}{\sum_{i=1}^n K\left(\frac{x-X_i}{h_n}\right)}.\end{aligned}$$

En resumen, a partir del estimador de Rosenblatt se obtiene la siguiente función de estimación de f:

$$\hat{f}_n(x) = \frac{\sum_{i=1}^n Y_i K\left(\frac{x-X_i}{h_n}\right)}{\sum_{i=1}^n K\left(\frac{x-X_i}{h_n}\right)}. \tag{2.11}$$

De esta manera, se ha obtenido un primer estimador no paramétrico de la regla de Bayes: el estimador de Nadaraya-Watson.

Observación 2.2.2. De la misma manera que el estimador de Rosenblatt, el estimador de Nadaraya-Watson se puede extender al ámbito multidimensional siguiendo el mismo proceso, aplicando la expresión multidimensional del estimador de Rosenblatt (2.2.1) para obtener $\hat{p}_n(\mathbf{x}, y)$ y $\hat{p}_n(\mathbf{x})$.

2.2.2. Estimadores no paramétricos lineales

A la hora de conseguir un estimador, conviene que se pueda expresar como combinación lineal de las etiquetas, siendo así más fácil de calcular su riesgo.

Definición 2.2.2. Un estimador $\hat{f}_n$ de $f(\mathbf{x})$ es un estimador lineal no paramétrico si existen unos pesos $W_{ni}(x, X_1, \ldots, X_n)$ no negativos (que se denotarán $W_{ni}(\mathbf{x})$) de cara a la simplificación de la notación) que cumplen que $\sum_{i=1}^n W_{ni}(\mathbf{x}) = 1$, tales que $\hat{f}_n$ se puede expresar de la siguiente forma:

$$\hat{f}_n(\mathbf{x}) = \sum_{i=1}^n Y_i W_{ni}(\mathbf{x}). \tag{2.12}$$

Dicho de una manera más coloquial, un estimador no paramétrico lineal es aquel que halla el valor de $\hat{f}_n(x)$ como combinación lineal de las etiquetas del conjunto de entrenamiento.

A partir de los valores que toma el estimador en cada uno de los atributos X_i, se define la **matriz de estimación** W_n:

$$W_n = \begin{pmatrix} W_{n1}(X_1) & W_{n2}(X_1) & \ldots & W_{nn}(X_1) \\ W_{n1}(X_2) & W_{n2}(X_2) & \ldots & W_{nn}(X_2) \\ \vdots & & \ddots & \vdots \\ W_{n1}(X_n) & \ldots & \ldots & W_{nn}(X_n) \end{pmatrix}.$$

Esto permite denotar al estimador en forma matricial:

$$\hat{\boldsymbol{f}}_n = W_n \boldsymbol{Y},$$

donde $\hat{\boldsymbol{f}}_n = (\hat{f}_n(X_1), \ldots, \hat{f}_n(X_n))^T$ y $\boldsymbol{Y} = (Y_1, \ldots, Y_n)^T$.

Ejemplo 2.2.1. Un ejemplo de un estimador lineal es el estimador medio local.

Denotando $I(x) = \{i : |x - X_i| \leq 1/n\}$ y $n(x) = Card(I)$, el estimador medio local es aquel cuyos pesos son

$$W_{ni}(x) = \begin{cases} \dfrac{1}{n(x)} & \text{si } n(x) > 0 \text{ y } i \in I \\ 0 & \text{en otro caso.} \end{cases},$$

siendo $n(x) := Card\{i : |x - X_i| \leq h\}$. Es decir, el estimador consiste en simplemente considerar como estimación de $f(x)$ el valor medio de las etiquetas Y_i cuyas respectivas X_i sean suficientemente cercanas a x. Esa idea de que solo se tengan en cuenta las etiquetas cuyos atributos sean suficiente cercanos al valor a estimar es muy importante, porque puesto que se trata de una estimación de un valor a nivel local, el valor etiquetas demasiado alejadas no es relevante par la aproximación.

En el caso en el que los atributos del conjunto de entrenamiento sean $X_i = i/n$ para todo i y se escogiera $h = 1/n$ (para facilitar los cálculos), su matriz de estimación sería:

$$W_n = \begin{pmatrix} 1/2 & 1/2 & 0 & \dots & 0 \\ 1/3 & 1/3 & 1/3 & \dots & 0 \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ 0 & \dots & 1/3 & 1/3 & 1/3 \\ 0 & \dots & 0 & 1/2 & 1/2 \end{pmatrix}.$$

Entre los estimadores no paramétricos lineales, se encuentra el estimador de Nadarya-Watson, comentado en la sección anterior ((2.11)). Una alternativa de definir ese estimador es:

Definición 2.2.3. Sea K una función núcleo, una muestra de entrenamiento $\{(X_i, Y_i)\}_{i=1}^n$ y $h_n > 0$ un ancho de banda. Se define al estimador de Nadaraya-Watson de la función de regresión como:

$$f_n^{NW}(x) := \arg\min_{\theta \in \Theta} \sum_{i=1}^{n} (Y_i - \theta)^2 K\left(\frac{x - X_i}{h}\right), \tag{2.13}$$

con $f_n^{NW}(x) = 0$ en caso de que $K\left(\frac{x-X_i}{h}\right) = 0$.

Es decir, bajo esa definición, el estimador de Nadaraya-Watson consiste en estimar $f(x)$ por la estimación de mínimos cuadrados de las etiquetas Y_i. Sin embargo, al igual que con el estimador medio local, conviene considerar únicamente las etiquetas cercanas, ya que se quiere hacer una estimación del valor a nivel local. Por ello, se utiliza una función núcleo K como forma de ponderar cuánto debe pesar cada muestra del conjunto de entrenamiento, según su cercanía a x.

Antes de nada, se tiene que demostrar que esta definición del estimador de Nadaraya-Watson da con la misma expresión a la conseguida anteriormente en (2.11). Para ello, basta ver que el mínimo de (2.13) es (2.11).

En caso en el que $K\left(\frac{x-X_i}{h}\right) = 0$, es obvio, por lo que se puede suponer que nos encontramos en el otro caso. Si $K\left(\frac{x-X_i}{h}\right) \neq 0$, entonces K es positivo, lo que implica que ese mínimo es único, y se puede calcular viendo cuándo la derivada respecto a θ se anula:

$$\begin{aligned} f_n^{NW}(x) = \arg\min_{\theta \in \Theta} \sum_{i=1}^{n} (Y_i - \theta)^2 K\left(\frac{x - X_i}{h}\right) \\ \Longrightarrow \sum_{i=1}^{n} 2(Y_i - \theta) K\left(\frac{x - X_i}{h}\right) = 0 \\ 2\left(\sum_{i=1}^{n} Y_i K\left(\frac{x - X_i}{h}\right) - \sum_{i=1}^{n} \theta K\left(\frac{x - X_i}{h}\right)\right) = 0 \\ \sum_{i=1}^{n} Y_i K\left(\frac{x - X_i}{h}\right) = \theta K \sum_{i=1}^{n}\left(\frac{x - X_i}{h}\right) \end{aligned}$$

$$\Longrightarrow f_n^{NW}(x) = \theta = \frac{\sum_{i=1}^n Y_i K\left(\frac{x-X_i}{h}\right)}{\sum_{i=1}^n K\left(\frac{x-X_i}{h}\right)}.$$

Esto da una definición alternativa del estimador, que además indica cómo hallar su valor computacionalmente, puesto que la función K es conocida y el ancho de banda h se considera elegido. Cabe recordar que esta fórmula es la misma que el estimador obtenido a través del estimadora de Rosenblatt. Dicho de otra manera, el estimador NW puede comprenderse a la ver de dos maneras:

- Como el estimador de mínimos cuadrados a nivel local.
- Como el estimador obtenido al estimar la función de densidad de la regla de Bayes por el estimador de Rosenblatt.

Proposición 2.2.1. Sea K una función núcleo, una muestra de entrenamiento $\{(X_i, Y_i)\}_{i=1}^n$ y $h_n > 0$ un ancho de banda. El estimador de Nadaraya-Watson de la función de regresión es:

$$f_n^{NW}(x) = \frac{\sum_{i=1}^n Y_i K\left(\frac{x-X_i}{h}\right)}{\sum_{i=1}^n K\left(\frac{x-X_i}{h}\right)}. \tag{2.14}$$

Observación 2.2.3. El estimador de Nadaraya-Watson es un estimador no paramétrico lineal, ya que, considerando

$$W_{ni}(x) = \frac{K\left(\frac{x-X_i}{h}\right)}{\sum_{j=1}^n K\left(\frac{x-X_j}{h}\right)},$$

se puede escribir el estimador de Nadaraya-Watson tal que

$$f_n^{NW}(x) = \sum_{i=1}^n Y_i W_{ni}(x).$$

Donde los $W_{ni}(x)$ son no negativos al serlo las funciones núcleo, y que $\sum_{i=1}^n W_{ni}(x) = 1$ para todo x.

Visualmente, en el estimador NW (Nadaraya-Watson) se intenta encontrar la constante que minimiza el error cuadrático medio en cierto entonrno de x (la función núcleo cumple el papel de "ponderar" cuán cercanos los atributos X_i tienen que ser a x para considerarse en la estimación).

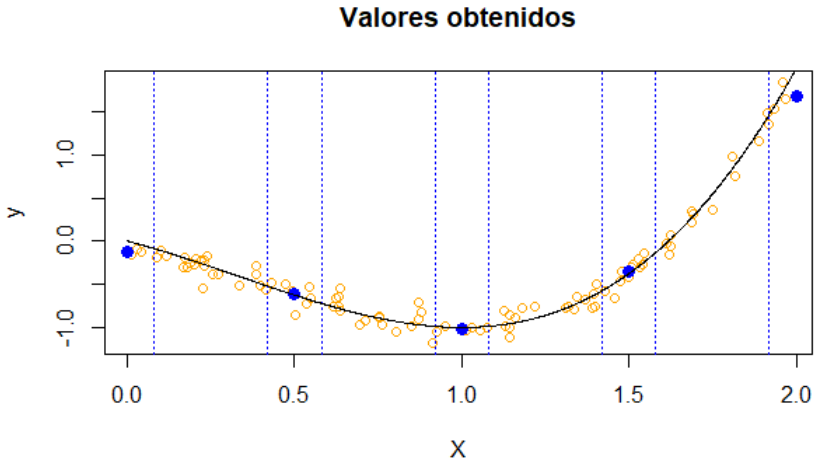


Figura 2.2: Partiendo de un conjunto de entrenamiento de 100 elementos, el estimador de Nadaraya-Watson para x= 0,5,1,1.5 y 2, siendo $f(x) = x^3 - x^2 - x$ la función a estimar, y donde las etiquetas del conjunto de entrenamiento tienen un error normal de varianza $\sigma^2 = 0{,}1$. Las líneas discontinuas indican el ancho de banda escogido (luego, los elementos del conjunto de entrenamiento que se utilizan para estimar cada uno de los valores son los que se encuentren entre medias).

El estimador de Nadaraya-Watson obtiene la constante local que más se asemeja al valor de las etiquetas del conjunto de entrenamiento cercanas. Sin embargo, no tiene en cuenta los posibles

puntos de inflexión o cambios de curvatura que puede tener la regresión cerca de x. Esto sugiere considerar mejorar el estimador para que se pueda tener en cuanta esos cambios a nivel local a la hora de estimar el valor de $f(x)$. Una opción es, en vez de buscar la constante θ que minimiza el error cuadrático, es buscar el polinomio local que lo hace. Esto da lugar a los estimadores locales polinomiales.

2.2.3. Estimadores locales polinomiales

La idea tras los estimadores locales polinomiales es el mejorar los estimadores NW, buscando un estimador que no solo encuentre el valor más cercano a $f(x_0)$, sino el polinomio local de grado l en x_0 que más se asemeje al comportamiento de f en ese entorno.

Un polinomio local tiene la siguiente forma:

$$P_{x_0}(z;\boldsymbol{\theta}) = \theta_0 + \theta_1(z-x_0) + \ldots + \frac{\theta_l}{l!}(z-x_0)^l := \boldsymbol{\theta}^T \boldsymbol{U}\,(z-x_0) \tag{2.15}$$

siendo

$$\boldsymbol{U}(u) := (1, u, u^2/2!, \ldots, u^l/l!)^T.$$

Para poder simplificar los resultados siguientes, se manejará en cambio $\boldsymbol{U}\left(\frac{x-X_i}{h}\right)$ para normalizar su valor, ya que $|x - X_i| \leq h$. Esto hace que entonces el polinomio local se denote

$$P_{x_0}(z;\boldsymbol{\theta}) = \theta_0 + \theta_1(z-x_0) + \ldots + \frac{\theta_l}{l!}(z-x_0)^l := \boldsymbol{\theta'}^T \boldsymbol{U}\left(\frac{z-x_0}{h}\right) \tag{2.16}$$

con $\boldsymbol{\theta'} := (\theta_0, \theta_1 h, \ldots, \theta_l\, h^l)$. Para simplificar la notación, denotaremos directamente $\boldsymbol{\theta}$ a $\boldsymbol{\theta'}$.

Si la función que se quiere estimar es es l-diferenciable (y su derivada $l-1$-ésima es continua), el polinomio local de grado l que mejor aproxima f es:

$$f(z) \approx f(x_0) + f'(x_0)(z-x_0) + \ldots + \frac{f^{(l)}(x_0)}{l!}(z-x_0)^l := \boldsymbol{F}_{x_0} \cdot \boldsymbol{U}\left(\frac{z-x_0}{h}\right), \tag{2.17}$$

además de ser único. Es interesante observar que el vector $\boldsymbol{F}_{x_0}$ no depende del valor de z, solo de x_0 y h.

En el estimador de Nadaraya-Watson, se estima el valor de $f(x)$ por la constante $\hat{\theta}$ que minimiza el error cuadrático medio para con las etiquetas cercanas Y_i, obteniendo una aproximación la función constante $\hat{\theta}$ que más se asemeja a f en un entorno de x.

En vez de eso, lo anterior sugiere estimar el valor de $f(x)$ no por la constante sino por los polinomios locales $P_x(X_i;\boldsymbol{\theta})$ que minimicen el error cuadrático medio, obteniendo un estadístico $\hat{\boldsymbol{\theta}}$ que aproxima el polinomio que más se asemeja a f en un entorno de x: $P_x(X_i;\boldsymbol{F}_x)$.

$$\begin{aligned}\hat{\boldsymbol{\theta}}_n(x) &= \arg\min_{\theta\in\mathbb{R}^{l+1}} \left(\sum_{i=1}^{n} \left[Y_i - P_{X_i}(x;\boldsymbol{\theta})\right]^2 K\left(\frac{x-X_i}{h}\right)\right) \\ &= \arg\min_{\theta\in\mathbb{R}^{l+1}} \left(\sum_{i=1}^{n} \left[Y_i - \boldsymbol{\theta}^T \boldsymbol{U}\left(\frac{x-X_i}{h}\right)\right]^2 K\left(\frac{x-X_i}{h}\right)\right).\end{aligned}$$

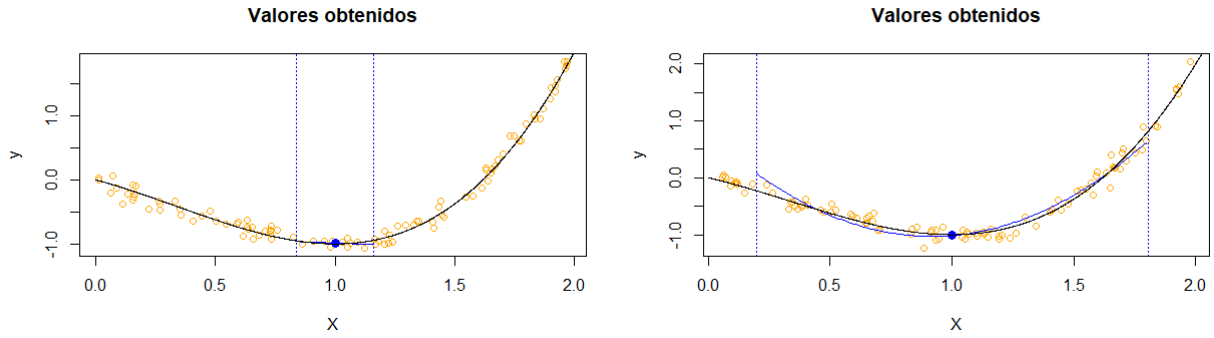
Definición 2.2.4. Sea $((X_i, Y_i))_{i=1}^n$ un conjunto de entrenamiento. Sea K un núcleo, $h > 0$ un ancho de banda. Se considera el estadístico

$$\hat{\boldsymbol{\theta}}_n(x) = \arg\min_{\theta\in\mathbb{R}^{l+1}} \left(\sum_{i=1}^{n} \left[Y_i - \boldsymbol{\theta}^T \boldsymbol{U}\left(\frac{x-X_i}{h}\right)\right]^2 K\left(\frac{x-X_i}{h}\right)\right).$$

El **estimador local polinomial** de orden l de la función de regresión f en x es aquel que estima el valor de $f(x) \approx P_x(x;\boldsymbol{\theta})$ por el valor de $P_x(x;\hat{\boldsymbol{\theta}})$:

$$\hat{f}_n(x) = \boldsymbol{U}^T(0)\hat{\boldsymbol{\theta}}_n(x) = \hat{\theta}_{1n}(x). \tag{2.18}$$

Siendo $\hat{\theta}_{1n}(x)$ la primera componente del vector $\hat{\boldsymbol{\theta}}_n(x)$.



(a) Estimador polinomial de orden $l = 1$ para $x = 1$

(b) Estimador polinomial de orden $l = 2$ para $x = 1$

Figura 2.3: Ejemplo del estimador local polinomial (con $l = 1$ y $l = 2$ respectivamente) para un conjunto de entrenamiento generado de la misma manera que el de la figura 2.2, para estimar el valor de $f(x) = x^3 - x^2 - x$ en $x = 1$. La curva en azul marca el polinomio local $P_x(x, \hat{\theta}_n)$ a partir del cual se obtiene la estimación $\hat{f}_n(1)$, el punto azul.

Al igual que el estimador de Nadaraya-Watson, los estimadores locales polinomiales también se pueden expresar en forma matricial:

Proposición 2.2.2. Considerando

$$\boldsymbol{a}_{nx} = \frac{1}{nh}\sum_{i=1}^{n} Y_i \boldsymbol{U}\left(\frac{x - X_i}{h}\right) K\left(\frac{x - X_i}{h}\right) \tag{2.19}$$

y

$$\mathcal{B}_{nx} = \frac{1}{nh}\sum_{i=1}^{n} \boldsymbol{U}\left(\frac{x - X_i}{h}\right)\boldsymbol{U}^T\left(\frac{x - X_i}{h}\right) K\left(\frac{x - X_i}{h}\right), \tag{2.20}$$

el estimador local polinomial satisface:

$$\hat{\boldsymbol{\theta}}_n(x) = \arg\min_{\theta\in\mathbb{R}^{l+1}} \left(-2\boldsymbol{\theta}^T\boldsymbol{a}_{nx} + \boldsymbol{\theta}^T\mathcal{B}_{nx}\boldsymbol{\theta}\right). \tag{2.21}$$

Demostración.

Denotando

$$\boldsymbol{a} = \sum_{i=1}^{n} Y_i \boldsymbol{U}\left(\frac{x - X_i}{h}\right) K\left(\frac{x - X_i}{h}\right) \tag{2.22}$$

y

$$\mathcal{B} = \sum_{i=1}^{n} \boldsymbol{U}\left(\frac{x - X_i}{h}\right)\boldsymbol{U}^T\left(\frac{x - X_i}{h}\right) K\left(\frac{x - X_i}{h}\right), \tag{2.23}$$

veamos que se tiene que

$$\hat{\boldsymbol{\theta}}_n(x) = \arg\min_{\theta\in\mathbb{R}^{l+1}} \left(-2\boldsymbol{\theta}^T\boldsymbol{a} + \boldsymbol{\theta}^T\mathcal{B}\boldsymbol{\theta}\right). \tag{2.24}$$

Con ello, ya se tendrá lo que se quiere, puesto que

$$\arg\min_{\theta\in\mathbb{R}^{l+1}} \left(-2\boldsymbol{\theta}^T\boldsymbol{a} + \boldsymbol{\theta}^T\mathcal{B}\boldsymbol{\theta}\right) = \arg\min_{\theta\in\mathbb{R}^{l+1}} \left(\frac{1}{nh}(-2\boldsymbol{\theta}^T\boldsymbol{a} + \boldsymbol{\theta}^T\mathcal{B}\boldsymbol{\theta})\right) = \arg\min_{\theta\in\mathbb{R}^{l+1}} \left(-2\boldsymbol{\theta}^T\boldsymbol{a}_{nx} + \boldsymbol{\theta}^T\mathcal{B}_{nx}\boldsymbol{\theta}\right).$$

Hay que demostrar entonces que

$$\hat{\boldsymbol{\theta}}_n(x) = \arg\min_{\theta\in\mathbb{R}^{l+1}} \left(\sum_{i=1}^{n} \left[Y_i - \boldsymbol{\theta}^T \boldsymbol{U}\left(\frac{x-X_i}{h}\right)\right]^2 K\left(\frac{x-X_i}{h}\right)\right) = \arg\min_{\theta\in\mathbb{R}^{l+1}} \left(-2\boldsymbol{\theta}^T\boldsymbol{a} + \boldsymbol{\theta}^T\mathcal{B}\boldsymbol{\theta}\right) :$$

$$\begin{aligned}
&\sum_{i=1}^{n} \left[Y_i - \boldsymbol{\theta}^T \boldsymbol{U}\left(\frac{x-X_i}{h}\right)\right]^2 K\left(\frac{x-X_i}{h}\right) \\
=&\sum_{i=1}^{n} \left[Y_i^2 - 2\boldsymbol{\theta}^T Y_i \boldsymbol{U}\left(\frac{x-X_i}{h}\right) + \left(\boldsymbol{\theta}^T \boldsymbol{U}\left(\frac{x-X_i}{h}\right)\right)^2\right] K\left(\frac{x-X_i}{h}\right) \\
=&\sum_{i=1}^{n} Y_i^2 K\left(\frac{x-X_i}{h}\right) - \sum_{i=1}^{n} 2\boldsymbol{\theta}^T Y_i \boldsymbol{U}\left(\frac{x-X_i}{h}\right) K\left(\frac{x-X_i}{h}\right) + \sum_{i=1}^{n} \left(\boldsymbol{\theta}^T \boldsymbol{U}\left(\frac{x-X_i}{h}\right)\right)^2 K\left(\frac{x-X_i}{h}\right).
\end{aligned}$$

El primer término se puede obviar (puesto que no depende de $\boldsymbol{\theta}$, el término a minimizar). Esto queda con los dos términos restantes, que operando queda:

$$\begin{aligned}
&-2\boldsymbol{\theta}^T \sum_{i=1}^{n} Y_i \boldsymbol{U}\left(\frac{x-X_i}{h}\right) K\left(\frac{x-X_i}{h}\right) + \sum_{i=1}^{n} \left(\boldsymbol{\theta}^T \boldsymbol{U}\left(\frac{x-X_i}{h}\right)\right)\left(\boldsymbol{\theta}^T \boldsymbol{U}\left(\frac{x-X_i}{h}\right)\right)^T K\left(\frac{x-X_i}{h}\right) \\
&= -2\boldsymbol{\theta}^T\boldsymbol{a} + \sum_{i=1}^{n} \left(\boldsymbol{\theta}^T \boldsymbol{U}\left(\frac{x-X_i}{h}\right)\right) \boldsymbol{U}^T\left(\frac{x-X_i}{h}\right)\boldsymbol{\theta} K\left(\frac{x-X_i}{h}\right) \\
&= -2\boldsymbol{\theta}^T\boldsymbol{a} + \sum_{i=1}^{n} \left(\boldsymbol{\theta}^T \boldsymbol{U}\left(\frac{x-X_i}{h}\right) \boldsymbol{U}^T\left(\frac{x-X_i}{h}\right)\boldsymbol{\theta}\right) K\left(\frac{x-X_i}{h}\right) \\
&= -2\boldsymbol{\theta}^T\boldsymbol{a} + \sum_{i=1}^{n} \boldsymbol{\theta}^T \left(\boldsymbol{U}\left(\frac{x-X_i}{h}\right) \boldsymbol{U}^T\left(\frac{x-X_i}{h}\right) K\left(\frac{x-X_i}{h}\right)\right)\boldsymbol{\theta} = \\
&= -2\boldsymbol{\theta}^T\boldsymbol{a} + \boldsymbol{\theta}^T \left(\sum_{i=1}^{n} \boldsymbol{U}\left(\frac{x-X_i}{h}\right) \boldsymbol{U}^T\left(\frac{x-X_i}{h}\right) K\left(\frac{x-X_i}{h}\right)\right)\boldsymbol{\theta} = -2\boldsymbol{\theta}^T\boldsymbol{a} + \boldsymbol{\theta}^T\mathcal{B}\boldsymbol{\theta}.
\end{aligned}$$

Por lo dicho al principio se obtiene que:

$$\begin{aligned}
\hat{\boldsymbol{\theta}}_n(x) &= \arg\min_{\theta\in\mathbb{R}^{l+1}} \left(\sum_{i=1}^{n} \left[Y_i - \boldsymbol{\theta}^T \boldsymbol{U}\left(\frac{x-X_i}{h}\right)\right]^2 K\left(\frac{x-X_i}{h}\right)\right) \\
&= \arg\min_{\theta\in\mathbb{R}^{l+1}} \left(-2\boldsymbol{\theta}^T\boldsymbol{a} + \boldsymbol{\theta}^T\mathcal{B}\boldsymbol{\theta}\right) \\
&= \arg\min_{\theta\in\mathbb{R}^{l+1}} \left(-2\boldsymbol{\theta}^T\boldsymbol{a}_{nx} + \boldsymbol{\theta}^T\mathcal{B}_{nx}\boldsymbol{\theta}\right).
\end{aligned}$$

□

Por el mismo motivo por el que se ha decidido considerar $\boldsymbol{U}\left(\frac{x-X_i}{h}\right)$ en vez de $\boldsymbol{U}(x-X_i)$, para que $\|\boldsymbol{a}_{nx}\| = 1$ y $\|\mathcal{B}_{nx}\| = 1$ y que sea más sencillo operar con ellas, se suele considerar en cambio Como

$$\arg\min_{\theta\in\mathbb{R}^{l+1}} \left(-2\boldsymbol{\theta}^T\boldsymbol{a}_{nx} + \boldsymbol{\theta}^T\mathcal{B}_{nx}\boldsymbol{\theta}\right) = \arg\min_{\theta\in\mathbb{R}^{l+1}} \left(\frac{1}{nh}(-2\boldsymbol{\theta}^T\boldsymbol{a}_{nx} + \boldsymbol{\theta}^T\mathcal{B}_{nx}\boldsymbol{\theta})\right)$$

Observación 2.2.4.

- El estimador de Nadaraya-Watson es el estimador $LP(0)$ (considerando $l = 0$).
- De la misma forma que podemos estimar $f(x)$, el estimador $LP(l)$ nos puede servir para encontrar estimadores de las l primeras derivadas de f, escogiendo sus respectivas coordenadas de $\hat{\boldsymbol{\theta}}_n(x)$.

- Dentro de los estimadores locales polinomiales, los normalmente usados son los $LP(1)$, también denominados **estimadores lineales locales.**

Sin embargo, del mismo modo que sucedía con el estimador Nadaraya-Watson, esta definición no es útil para calcular computacionalmente el valor del estimador. Para ello, se necesitan hipótesis adicionales para que se tenga una expresión del mínimo de $-2\boldsymbol{\theta}^T\boldsymbol{a}_{nx} + \boldsymbol{\theta}^T\mathcal{B}_{nx}\boldsymbol{\theta}$.

Proposición 2.2.3. Si $\mathcal{B}_{nx}$ es definida positiva ($\mathcal{B}_{nx} > 0$), entonces el estimador $LP(l)$ es único, y viene dado por

$$\hat{\boldsymbol{\theta}}_n(x) = \mathcal{B}_{nx}^{-1}\boldsymbol{a}_{nx} \tag{2.25}$$

Demostración.

Podemos calcular mínimos locales de valores de $\hat{\boldsymbol{\theta}}$ viendo cuándo se cumple

$$\frac{d}{d\boldsymbol{\theta}}\left(-2\boldsymbol{\theta}^T\boldsymbol{a}_{nx} + \boldsymbol{\theta}^T\mathcal{B}_{nx}\boldsymbol{\theta}\right) = -2\boldsymbol{a}_{nx} + 2\mathcal{B}_{nx}\boldsymbol{\theta} = \mathbf{0}$$

Como $\mathcal{B}_{nx}$ es definida positiva, es invertible, y la solución es única y un mínimo absoluto. Por tanto,

$$\hat{\boldsymbol{\theta}}_{\boldsymbol{n}}(x) = \arg\min_{\theta\in\mathbb{R}^{l+1}}\left(-2\boldsymbol{\theta}^T\boldsymbol{a}_{nx} + \boldsymbol{\theta}^T\mathcal{B}_{nx}\boldsymbol{\theta}\right) = \mathcal{B}_{nx}^{-1}\boldsymbol{a}_{nx}.$$

□

Proposición 2.2.4. Si la matriz $\mathcal{B}_{nx}$ definida en la proposición anterior es definida positiva, entonces su respectivo estimador local polinomial $\hat{f}_n(x)$ es un estimador lineal.

Demostración.

Para ello, tenemos que ver que $\hat{f}_n(x) = \boldsymbol{U}^T(0)\hat{\boldsymbol{\theta}}_n(x)$ es de la forma

$$f(x) = \sum_{i=1}^{n} Y_i W_{ni}(x)$$

Por la proposición anterior, se deduce inmediatamente que

$$\begin{aligned}\hat{\theta} = \mathcal{B}_{nx}^{-1}\boldsymbol{a}_{nx} &= \mathcal{B}_{nx}^{-1}\left(\frac{1}{nh}\sum_{i=1}^{n} Y_i\boldsymbol{U}\left(\frac{x-X_i}{h}\right)K\left(\frac{x-X_i}{h}\right)\right)\\ &= \frac{1}{nh}\sum_{i=1}^{n} Y_i\mathcal{B}_{nx}^{-1}\boldsymbol{U}\left(\frac{x-X_i}{h}\right)K\left(\frac{x-X_i}{h}\right)\\ &= \sum_{i=1}^{n} Y_i\left(\frac{1}{nh}\mathcal{B}_{nx}^{-1}\boldsymbol{U}\left(\frac{x-X_i}{h}\right)K\left(\frac{x-X_i}{h}\right)\right) = \sum_{i=1}^{n} Y_i W_{ni}(x)\end{aligned}$$

□

Observación 2.2.5.

El resultado anterior implica que no solo se conoce una expresión explícita del estimador local polinomial, sino que además se puede calcular su valor numérico en la práctica:

1. Dar tanto con la matriz $\mathcal{B}_{nx}$ como con el vector $\mathbf{a}_{nx}$. Para ello:
   - Para cada i:
     - Calcular $K(\frac{X_i-x}{h})$.
     - Obtener la matriz $U(\frac{X_i-x}{h})U^T(\frac{X_i-x}{h})$, (al ser simétrica, calcular $\frac{n(n+1)}{2}$ elementos). Esto también implica calcular $U(\frac{X_i-x}{h})$, ya que es su primera columna ($\mathcal{O}(\frac{n^2-n}{2})$).

- Calcular $Y_i \boldsymbol{U}\left(\frac{x-X_i}{h}\right) K\left(\frac{x-X_i}{h}\right)$.$(\mathcal{O}(n))$
- Calcular $\boldsymbol{U}\left(\frac{x-X_i}{h}\right)\boldsymbol{U}^T\left(\frac{x-X_i}{h}\right) K\left(\frac{x-X_i}{h}\right) \to \mathcal{O}(\frac{n^2-n}{2}))$.

Es decir, se obtiene una complejidad de $\mathcal{O}(n^2(n-1))$.

- Sumar los n vectores y las n matrices $\to \mathcal{O}(n)$ y $\mathcal{O}(\frac{n^2-n}{2})$, respectivamente.

2. Resolver el sistema $\mathcal{B}_{nx}\boldsymbol{\theta} = \mathbf{a}_{nx}$. Puesto que B_{nx} es definida positiva, se puede considerar su factorización LU. El coste computacional de resolver un sistema mediante la factorización LU es de $O(\frac{2}{3}n^3)$[2].

En definitiva, **siempre y cuando $\mathcal{B}_{nx}$ sea definida positiva**, podemos calcular el valor del estimador, además de que todas las operaciones a realizar para calcular el valor a estimar son factibles a nivel computacional, siendo ésta la mayor desventaja del estimador.

En el siguiente capítulo (específicamente, en la Sección 3.1), se estudiará el riesgo del estimador local polinomial, pudiéndolo acotar bajo ciertas condiciones.

2.3. Teoría Minimax

Aunque se consiga obtener una acotación sobre el riesgo de cierto estimador un a clase paramétrica, ésta no es interesante si otros posibles estimadores dentro de $\mathcal{F}$ obtienen una mejor estimación de la regla de de Bayes. Es más, puesto que el riesgo depende de la función a estimar, conviene considerar como medida de comparación una medida más estable, que muestre la eficiencia del estimador dentro de toda la clase. De esa manera, se puede considerar comparar el máximo riesgo obtenido a la hora de estimar cualquier función de la clase $\mathcal{F}$ para cada estimador, queriendo idóneamente obtener el menor de todos ellos: el riesgo minimax de la clase $\mathcal{F}$.

La teoría minimax no está restringida a tratar con funciones de regresión, sino que se puede aplicar a cualquier problema de estimación. En general, se puede aplicar a cualquier modelo estadístico $(\mathcal{X}, \mathcal{A}, \{P_\theta : \theta \in \Theta\})$, siempre que se pueda definir una función de pérdida l en el espacio de parámetros Θ.

Para un estimador $\hat{\theta}$ dentro de una clase Θ, se puede considerar cuán buen estimador es a partir del máximo riesgo obtenido para todas las funciones de la clase Θ, que se le denomina **riesgo máximo** de $\hat{\theta}_n$:

$$R_{\text{máx}}(\hat{\theta}_n) := \sup_{\theta\in\Theta} \mathbb{E}_\theta\left[l(\hat{\theta}_n, \theta)\right].$$

El riesgo máximo sirve para medir el rendimiento de un estimador para estimar las funciones de la clase Θ, y así comparar varios estimadores entre ellos. Por ello, tiene interés considerar el riesgo del estimador con menor riesgo máximo: el riesgo minimax.

Definición 2.3.1. Se define el **riesgo minimax** en Θ a:

$$\mathcal{R}_n^* := \inf_{\hat{\theta}_n} \sup_{\theta\in\Theta} \mathbb{E}_\theta\left[l(\hat{\theta}_n, \theta)\right]. \tag{2.26}$$

La **estimación minimax** será aquella cuyo riesgo asociado será el riesgo minimax.

$$\theta^* := \inf_{\hat{\theta}_n} \sup_{\theta\in\Theta} \mathbb{E}_\theta(l(\hat{\theta}_n, \theta)).$$

En un principio se puede considerar entonces que la meta sea encontrar los estimadores en Θ cuya tasa de convergencia sea la óptima:

Definición 2.3.2. Un estimador θ_n^* es **asintóticamente eficiente** en (Θ, d) si cumple que

$$\lim_{n\to\infty} \frac{R_{\text{máx}}(\theta_n^*)}{\mathcal{R}_n^*} = 1. \tag{2.27}$$

Es decir, un estimador será asintóticamente eficiente si la tasa de convergencia del riesgo máximo del estimador es equivalente a la tasa óptima de convergencia.

Sin embargo, el encontrar un estimador asintóticamente eficiente no es posible salvo en modelos estadísticos singulares. Es por ello que resulta más ventajoso extender a también considerar aquellos cuya tasa de convergencia sea "parecida":

Definición 2.3.3. Sea una sucesión $(\psi_n)_{n=1}^{\infty}$ que tiende a 0. Se dice que es el **tasa óptima de convergencia** de los estimadores en (Θ, d) si cumple

$$\limsup_{n\to\infty} \psi_n^{-2}\mathcal{R}_n^* \leq C \quad \text{y} \quad \liminf_{n\to\infty} \psi_n^{-2}\mathcal{R}_n^* \geq c. \tag{2.28}$$

Definición 2.3.4. Un estimador θ_n^* es un **estimador óptimo en tasa** en (Θ, d) si, siendo $(\psi_n)_{n=1}^{\infty}$ una tasa óptima de convergencia y $C < \infty$ una constante, satisface que

$$\sup_{\theta\in\Theta} \mathbb{E}_\theta \left[l(\theta_n^*, \theta)\right] \leq C\psi_n^2. \tag{2.29}$$

Que un estimador sea óptimo en tasa es una condición menos exigente a que sea asintóticamente eficiente. Sin embargo, asegura que la velocidad de decrecimiento del riesgo para conjuntos de entrenamientos grandes es la misma que utilizando un estimador eficiente.

El objetivo será entonces obtener cuál es la tasa óptima de convergencia $(\psi_n)_{n=1}^{\infty}$ en $(\mathcal{F}, d)$ (para cada una de las posibles d). Es decir, encontrar la sucesión $(\psi_n)_{n=1}^{\infty}$ que cumpla las desigualdades (2.28).

La acotación superior se puede obtener a través de la obtenida para un estimador $\hat{\theta}_n$ en Θ

$$\sup_{\theta\in\Theta} \mathbb{E}_\theta \left[l(\hat{\theta}_n, \theta)\right] \leq C.$$

La verdadera dificultad se encuentra en encontrar la cota inferior:

$$\forall \hat{\theta}_n : \quad \sup_{\theta\in\Theta} \mathbb{E}_\theta \left[l(\hat{\theta}_n, \theta)\right] \geq c\psi_n^2. \tag{2.30}$$

Además, encontrar las acotaciones de esta forma implicará que el estimador $\hat{\theta}$ utilizado para la cota superior será óptimo en tasa, que será el objetivo a cumplir en la segunda parte con el estimador local polinomial.

En el problema que se considera, se asumirá que el espacio de parámetros Θ es un espacio (pseudo)métrico (Θ, d) [2], donde la función de pérdida será d^2 (puesto que se necesitará de la desigualdad triangular).

Observación 2.3.1. En el caso específico de nuestro modelo, Θ es la clase no paramétrica de funciones $\mathcal{F}$. Por tanto, la familia de medidas probabilísticas $\{P_\theta : \theta \in \mathcal{F}\}$ será la formada por las medidas probabilísticas de las leyes de las etiquetas del conjunto de entrenamiento respecto a la función a estimar f. Es decir, $P_f = \mathcal{L}((Y_1, \ldots, Y_n)|f)$.

En cuanto a las (psuedo)distancias consideradas, serán las asociadas a cada una de las funciones de pérdida que se consideran en el modelo. Es decir:

- La distancia L^2 $(d(f,g) = \|f - g\|_2)$
- La distancia del supremo $(d(f,g) = \|f - g\|_\infty)$
- La distancia puntual para x_0 fijo: $d(f,g) = |f(x_0) - g(x_0)|$.

A partir de ahora entonces se trabajará con que (Θ, d) es un espacio métrico, donde se considera como función de pérdida d^2. Esto implica que el riesgo de $\hat{\theta}_n$ en (Θ, d) será $R(\hat{\theta}_n) = \mathbb{E}_\theta(d^2(\hat{\theta}_n, \theta))$, y el riesgo minimax

$$\mathcal{R}_n^* := \inf_{\hat{\theta}_n} \sup_{\theta\in\Theta} \mathbb{E}_f \left[d^2(\hat{\theta}_n, \theta)\right], \tag{2.31}$$

denotando con $\hat{\theta}_n$ a los posibles estimadores partiendo de un conjunto de entrenamiento de n elementos.

[2] Las propiedades que se necesitan de d es que sea una función no negativa y que cumpla la desigualdad triangular, por lo que sea una pseudométrica es suficiente. Sin embargo, a veces se denominará a d distancia por generalidad.

Observación 2.3.2. La desigualdad presentada en (2.30) para (Θ, d) medible es equivalente a

$$\forall \hat{\theta}_n : \quad \sup_{\theta\in\Theta} \mathbb{E}_\theta \left[\psi_n^{-2} d^2(\hat{\theta}_n, \theta)\right] = \sup_{\theta\in\Theta} \mathbb{E}_\theta \left[\left(\psi_n^{-1} d(\hat{\theta}_n, \theta)\right)^2\right] \geq c, \tag{2.32}$$

que será la desigualdad que se intentará conseguir.

Para obtener una acotación inferior de $\mathcal{R}_n^*$, se puede utilizar la desigualdad de Markov:

$$P(|X| \geq A) \leq \frac{\mathbb{E}(X^2)}{A}.$$

Aplicándolo al contexto del riesgo minimax, implica que se cumple que

$$P_\theta(d(\hat{\theta}_n, \theta) \geq s_n) \leq \mathbb{E}_\theta(\psi^{-2} d^2(\hat{\theta}_n, \theta)),$$

siendo $s_n = A\psi_n$. Dicho de otra manera:

$$\mathcal{R}_n^* \geq \inf_{\hat{\theta_n}} \sup_{\theta\in\Theta} P_\theta(d(\hat{\theta}_n, \theta) \geq s_n). \tag{2.33}$$

Es más, puesto que el máximo de un conjunto es mayor que el de cualquier subconjunto,

$$\mathcal{R}_n^* \geq \inf_{\hat{\theta_n}} \max_{0\leq M} P_{\theta_j}(d(\hat{\theta}_n, \theta_j) \geq s_n) \tag{2.34}$$

para cada $\{\theta_0, \ldots, \theta_M\} \subset \Theta$.

Para manejar mejor la expresión presentada en (2.34), se presenta a continuación la probabilidad de error minimax en un problema de contraste.

Supongamos que se pretende decidir entre $M+1$ hipótesis $\{\theta_0, \ldots, \theta_M\}$. Se puede basar la decisión en el estimador $\hat{\theta}_n$, mediante la regla

$$\Psi^* = \arg \min_{0\leq k\leq M} d(\hat{\theta}_n, \theta_k). \tag{2.35}$$

es decir, escoger la hipótesis más cercana a $\hat{\theta}_n$.

Ψ^* es una **función test** (en este contexto, esto significa que es una función medible de $\mathcal{X}$ en $\{0, 1, \ldots, M\}$).

La **probabilidad de error minimax** asociada a estas reglas de mínima distancia es

$$p_{e,M} := \inf_{\hat{\theta}_n} \max_{0\leq j\leq M} P_{\theta_j}(\Psi^* \neq j). \tag{2.36}$$

Supongamos ahora que las hipótesis $\{\theta_0, \ldots, \theta_M\}$ se eligen tales que $d(\theta_j, \theta_k) \geq 2s$ si $j \neq k$.

Entonces, $\Psi^* \neq j$ implica que existe $k \neq j$ tal que $d(\hat{\theta_n}, \theta_k) < d(\hat{\theta_n}, \theta_j)$. Esto a su vez implica que $d(\hat{\theta_n}, \theta_k) \geq s$ (en caso contrario, se cumpliría que $d(\theta_j, \theta_k) \leq d(\theta_j, \hat{\theta_n}) + d(\hat{\theta_n}, \theta_j) \leq 2s$, en contra de que las hipótesis se hayan eligiendo cumpliendo que $d(\theta_j, \theta_k) \geq 2s$), y por tanto se obtiene el siguiente resultado:

Teorema 2.3.1. Sean $\{\theta_0, \ldots, \theta_M\} \subset M$ tales que $d(\theta_j, \theta_k) \geq 2s$ si $j \neq k$. Entonces, se cumple que

$$\mathcal{R}_n^* \geq \inf_{\hat{\theta_n}} \max_{0\leq M} P_{\theta_j}(d(\hat{\theta}_n, \theta_j) \geq s_n) \geq p_{e,M}. \tag{2.37}$$

Esto justifica que para acotar el riesgo minimax en problemas de estimación, se recurra a cotas inferiores para la probabilidad del error minimax. Estas cotas se estudiarán en la Sección 2.5.

2.4. Métricas probabilísticas

Recordemos que en la Teoría Minimax, se trabaja bajo el espacio de parámetros (Θ, d), construido a partir del modelo estadístico $(\mathcal{X}, \mathcal{A}, \{P_\theta : \theta \in \Theta\})$. En (Θ, d), se quiere encontrar la tasa de convergencia óptima, para lo que hace falta buscar una sucesión $(\psi)_{n=1}^{\infty}$ que cumpla la desigualad (2.30).

Como se vio al final de la Sección 2.3 hablando sobre la Teoría Minimax, una manera de poder encontrar acotaciones inferiores del riesgo minimax es obtenerlas para la probabilidad de error minimax $p_{e,M}$. Para acotar $p_{e,M}$, se tendrá que trabajar en el espacio muestral $(\mathcal{X}, \mathcal{A})$, usando las probabilidades P_θ. Para ello, se necesitará definir varias métricas en ese espacio probabilístico.

Existen varias posibles métricas en un espacio probabilístico. En concreto, se estudiarán la distancia de Hellinger y la variación total. Aún no siendo una distancia, también se hablará de la divergencia de Kullback-Leibler, puesto que se suele utilizar para este mismo propósito, sobre todo en la Teoría de la Información.

Para presentar dichas distancias, se trabajará con P y Q dos probabilidades en el espacio medible $(\mathcal{X},\mathcal{A})$, y ν una medida σ-finita en $(\mathcal{X},\mathcal{A})$), tal que $P, Q \ll \nu$. Se denotará $p = dP/d\nu$ y $q = dQ/d\nu$ para simplificar (siempre existe una probabilidad μ que cumpla con dicha propiedad, ya que la probabilidad $\nu = \frac{P+Q}{2}$ lo cumple).

Distancia de Hellinger

Definición 2.4.1. La distancia de Hellinger entre P y Q se define como:

$$H(P,Q) := \left(\int (\sqrt{p} - \sqrt{q})^2 \, d\nu \right)^{1/2}. \tag{2.38}$$

Proposición 2.4.1. $H(P,Q)$ no depende de la elección de ν.

Demostración.

Supongamos que $P, Q \gg \nu'$, siendo ν' una medida distinta a ν. Se quiere ver que $H(P,Q) = \left(\int (\sqrt{p} - \sqrt{q})^2 \, d\nu \right)^{1/2}$.

Sea ν'' tal que $\nu \ll \nu''$ y $\nu' \ll \nu''$ (por ejemplo $\nu'' = \nu + \nu'$). Tal y como se ha definido, ν'' cumple que $P \ll \nu''$ y $Q \ll \nu''$.

Utilizado la regla de la cadena de la derivada de Radon-Nykodym (Proposición A.2.2), se obtiene que:

$$\begin{aligned}
H^2_{\nu''}(P,Q) &= \int \left(\sqrt{\frac{dP}{d\nu''}} - \sqrt{\frac{dQ}{d\nu''}} \right)^2 d\nu'' \\
&= \int \left(\sqrt{\frac{dP}{d\nu}\frac{d\nu}{d\nu''}} - \sqrt{\frac{dQ}{d\nu}\frac{d\nu}{d\nu''}} \right)^2 d\nu'' \\
&= \int \left(\sqrt{\frac{dP}{d\nu}} - \sqrt{\frac{dQ}{d\nu}} \right)^2 \frac{d\nu}{d\nu''} \, d\nu'' \\
&= \int \left(\sqrt{\frac{dP}{d\nu}} - \sqrt{\frac{dQ}{d\nu}} \right)^2 d\nu = H^2_\nu(P,Q).
\end{aligned}$$

Puesto que el papel que ha cumplido ν también lo puede hacer ν' (ya que solo se ha utilizado que $\nu \ll \nu''$), también cumple que $H^2_{\nu''}(P,Q) = H^2_{\nu'}(P,Q)$. Uniendo ambas desigualdades, se tiene que $H^2_{\nu'}(P,Q) = H^2_\nu(P,Q)$, la que implica que $H_{\nu'}(P,Q) = H_\nu(P,Q)$ quedando entonces demostrado que la distancia de Hellinger no depende de la medida ν escogida.

□

Proposición 2.4.2. La distancia de Hellinger tiene las siguientes propiedades:

1. $H(P,Q)$ es una distancia. Para demostrarlo, basta con observar que la distancia de Hellinger es la distancia asociada la norma L^2 para las funciones $\sqrt{p}$ y $\sqrt{q}$.

$$H(P,Q)\left(\int(\sqrt{p}-\sqrt{q})^2 d\nu\right)^{\frac{1}{2}} = \|\sqrt{p}-\sqrt{q}\|_2 = d(\sqrt{p},\sqrt{q}).$$

2. $H^2(P,Q) = 2\left(1-\int\sqrt{pq}\,d\nu\right) = 2\left(1-\sqrt{dP\,dQ}\right)$
3. $0 \leq H^2(P,Q) \leq 2$, que se deduce del hecho de que sea una distancia y que se cumpla la propiedad anterior.

Distancia de la variación total

Definición 2.4.2. La distancia de la variación total (denominada simplemente variación total) entre P y Q se define como

$$V(P,Q) := \sup_{A\in\mathcal{A}}|P(A)-Q(A)| = \sup_{A\in\mathcal{A}}\left|\int_A (p-q)d\nu\right|. \tag{2.39}$$

Cabe destacar la similitud entre esta distancia **probabilística** y la asociada la norma del supremo $\|\|\|_\infty$.

Observación 2.4.1. Del mismo modo que se demostró que la distancia de Hellinger no depende de ν, se puede demostrar lo mismo para la variación total. Esto se puede ver más claramente para esta distancia, donde la primera expresión con la se define no depende de ν.

Una expresión alternativa de la distancia de variación total es dada por el Teorema de Scheffé.

Lema 2.4.3 (Teorema de Scheffé)**.**

$$V(P,Q) = \frac{1}{2}\int|p-q|d\nu = 1-\int \text{mín}(dP,dQ)$$

Demostración.

La segunda desigualdad se cumple debido a que p y q son funciones de densidad:

$$\begin{aligned} 2 = 1+1 = \int p\,d\nu + \int q\,d\nu &= \int|p-q|\,d\nu + \int \text{mín}(p,q)d\nu + \int \text{mín}(p,q)d\nu \\ \Rightarrow 2 &= \int|p-q|\,d\nu + 2\int \text{mín}(p,q)d\nu \\ \Rightarrow \frac{1}{2}\int|p-q|\,d\nu &= 1-\int \text{mín}(p,q)d\nu \end{aligned}$$

Para la primera igualdad, se verá que se cumple tanto que $V(P,Q) \geq \frac{1}{2}\int|p-q|\,d\nu$ como que $V(P,Q) \leq \frac{1}{2}\int|p-q|\,d\nu$.

Sea $A_0 = \{x\in\mathcal{X} : q(x)\geq p(x)\}$. Se obtiene que

$$\int|p-q|\,d\nu = \int_{A_0} q-p\,d\nu + \int_{A_0^c} p-q\,d\nu \geq 2\int_{A_0}(q-p)\,d\nu.$$

Con ello, se obtiene la primera desigualdad:

$$V(P,Q) \geq Q(A_0)-P(A_0) = \frac{1}{2}\int|p-q|\,d\nu.$$

Falta entonces ver que también se cumple la desigualdad contraria para obtener la primera igualdad.

Para todo $A\in\mathcal{A}$, se puede considerar la división $A = (A\cap A_0)\cup(A\cap A_0^c)$. Trabajando sobre ello :

$$\left|\int_A (q-p)d\nu\right| = \left|\int_{A\cap A_0}(q-p)d\nu + \int_{A\cap A_0^c}(q-p)\,d\nu\right|$$
$$\leq \text{máx}\left\{\int_{A_0}(q-p)\,d\nu, \int_{A_0^c}(p-q)\,d\nu\right\} = \frac{1}{2}\int |p-q|\,d\nu = Q(A_0) = Q(A_0) + P(A_0).$$

Por tanto, se obtiene que

$$V(P,Q) = Q(A_0) + P(A_0)$$

de lo que se deduce lo que queríamos demostrar.

□

Proposición 2.4.4. La distancia de la variación total tiene las siguientes propiedades:

1. $V(P,Q)$ es una distancia, puesto que, por el Teorema de Scheffé, es la mitad de la distancia asociada a la norma $L1$ entre p y q ($V(P,Q) = 1/2\,\|p-q\|_1$).
2. $0 \leq V(P,Q) \leq 1$.Se deduce de que $V(P,Q) = 1 - \int \text{mín}(dP, dQ)$.

Divergencia de Kullback-Liebler

Definición 2.4.3. La **divergencia de Kullback** entre P y Q se define por

$$K(P,Q) := \begin{cases} \int \log\dfrac{dP}{dQ}dP & si P \ll Q \\ +\infty & \text{en otro caso} \end{cases}$$

Tal y como la divergencia de Kullback ha sido definida, nada asegura que para el caso $P \ll Q$ la integral de Lebesgue no esté bien definida, puesto que si

$$\int \left(\log\frac{dP}{dQ}\right)_+ dP = \infty \qquad y \qquad \int \left(\log\frac{dP}{dQ}\right)_- dP = \infty$$

la integral de Lebesgue resultará en una indeterminación. Para ver que no es el caso, se comprueba que $\int \log\dfrac{dP}{dQ}dP = \infty$ positiva.

Proposición 2.4.5. Si $P \ll Q$, $K(P,Q) \geq 0$, y por tanto, está bien definida.

Demostración.

Si $P \ll Q$, se tiene que $\{q > 0\} \supseteq \{p > 0\}$. Esto implica $\{q \cdot p > 0\} = \{p > 0\}$. Esto implica que se trabaja con que p y q son del mismo signo:

$$K(P,Q) = \int \log\frac{dP}{dQ}dP = \int_{p>0} p\log\frac{p}{q}d\nu = \int_{pq>0} p\log\frac{p}{q}d\nu = \int_{s>0} q\cdot s\log s\,d\nu = \int_{s>0} q\cdot h(s)\,d\nu.$$

Denotando $h(s) = s\log s$ y $g(s) = s-1$, $h(s) \leq g(s)$ para $s > 0$. Esto es debido a que $h(0) \to 0 > -1 = g(0)$, y h tiene una mayor derivada que g: $h'(s) = 1 + \log(s) > 1 = g'(s)$. Por tanto,

$$\int_{s>0} q\cdot h(s)d\nu \leq \int_{s>0} q\cdot g(s)d\nu \leq \int_{s>0} q(s-1)d\nu \geq 0.$$

□

Observación 2.4.2. $K(Q,P)$ **no** es una distancia. Es más, $K(Q,P) \neq K(P,Q)$, ya que el papel que cumplen P y Q son distintos.

Sin embargo, si que cumple ciertas propiedades asociadas a las distancias que necesarias para nosotros, como la desigualdad triangular o el hecho de que sea no negativa.

Todas estas distancias se pueden aplicar a probabilidades producto. La siguiente pregunta sería pues cómo se relaciona la distancia entre probabilidades producto con las distancias de las probabilidades que lo componen. Para cada una de las distancias hay una relación diferente. Como para nuestro objetivo de estudio solo se va a necesitar la asociada a la divergencia de Kullback, será la única que se mencionará.

Proposición 2.4.6. Si P y Q son medidas producto ($P = \bigotimes_{i=1}^{n} P_i$, y $Q = \bigotimes_{i=1}^{n} Q_i$), se tiene que

$$K(P,Q) = \sum_{i=1}^{n} K(P_i, Q_i). \tag{2.40}$$

Demostración.

Basta con ver el caso $P \ll Q$ (el otro caso es obvio), donde las propiedades del logaritmo consiguen la igualdad:

$$\begin{aligned} K(P,Q) &= \int \log \frac{dP}{dQ} dP = \int \log dP\, dP - \int \log dQ\, dP \\ &= \int \log \left(\prod_{i=1}^{n} dP_i \right) dP - \int \log \left(\prod_{i=1}^{n} dQ_i \right) dP \\ &= \int \sum_{i=1}^{n} \log dP_i\, dP - \int \sum_{i=1}^{n} \log dQ_i\, dP \\ &= \sum_{i=1}^{n} \int (\log dP_i - \log dQ_i)\ dP = \sum_{i=1}^{n} K(P_i, Q_i). \end{aligned}$$

□

2.4.1. Relaciones entre las métricas

El objetivo de presentar estas métricas probabilísticas es para posteriormente conseguir acotaciones a probabilidades según su distancia. Para construir dichas acotaciones de manera eficaz, primero se tendrá que ver cómo están relacionadas entre sí. Hay varias desigualdades que relacionan entre sí las tres distancias estudiadas.

Lema 2.4.7 (Desigualdad de Le Cam)**.**

$$\int \text{mín}(dP, dQ) \geq \frac{1}{2} \left(\int \sqrt{dP dQ} \right)^2 = \frac{1}{2} \left(1 - \frac{H^2(P,Q)}{2} \right)^2$$

$$\Rightarrow \frac{1}{2} H^2(P,Q) \leq V(P,Q) \leq H(P,Q) \sqrt{1 - \frac{H^2(P,Q)}{4}} \tag{2.41}$$

Demostración.

Como $\int p = 1$ y $\int q = 1$, se obtiene que $\int \text{máx}(p,q) + \int \text{mín}(p,q) = 2$. Junto a la desigualdad de Jensen (para $f(x) = \sqrt{x}$):

$$\begin{aligned} \left(\int \sqrt{pq} \right)^2 &= \left(\int \sqrt{\text{mín}(p,q)\, \text{máx}(p,q)} \right)^2 \leq \int \text{mín}(p,q)\, \text{máx}(p,q) \\ &= \int \text{mín}(p,q) \left(2 - \int \text{mín}(p,q) \right) \end{aligned}$$

Con esto se obtiene la desigualdad primera. La igualdad de los dos últimos términos se da por las propiedades de la distancia de Hellinger (ver en la Sección 2.4).

Traduciendo este resultado a las distancias $V(\cdot,\cdot)$ y $H^2(\cdot,\cdot)$:

$$\left(1-\frac{H^2(P,Q)}{2}\right)^2 = \left(\int\sqrt{pq}\right)^2 \leq \int \text{mín}(p,q)\left(2-\int \text{mín}(p,q)\right)$$
$$= (1-V(P,Q))(1+V(P,Q)) = 1-V^2(P,Q)$$
$$\Rightarrow V(P,Q) \leq H(P,Q)\sqrt{1-\frac{H^2(P,Q)}{4}}$$

Además, como $\int \text{mín}(p,q) \geq \int\sqrt{pq}$, también se tiene que:

$$V(P,Q) = 1-\int \text{mín}(p,q) \geq 1-\int\sqrt{pq} = H^2(P,Q)/2$$

Si se juntan ambas desigualdades, obtenemos:

$$H^2(P,Q)/2 \leq V(P,Q) \leq H(P,Q)\sqrt{1-\frac{H^2(P,Q)}{4}}$$

como queríamos probar. □

Observación 2.4.3. Es interesante observar que la desigualdad de Le Cam muestra que

$$H^2(P,Q)/2 \leq V(P,Q) \leq 2\sqrt{1-\frac{H^2(P,Q)}{4}}$$
$$\Rightarrow f_1(H^2(P,Q)) \leq V(P,Q) \leq f_2(H^2(P,Q)),$$

poniendo de manifiesto una relación de equivalencia entre la distancia de variación total y $H(\cdot,\cdot)$.

Lema 2.4.8.

$$H^2(P,Q) \leq K(P,Q). \tag{2.42}$$

Demostración.

Es suficiente con probar el caso en el que $K(P,Q) < \infty$ (ya que $H^2(P,Q) < 2$), por lo que podemos asumir que $P \ll Q$.
Como $-\log(x+1) \geq -x$ si $x > -1$, obtenemos:

$$K(P,Q) = \int_{pq>0} p\left(\log\frac{p}{q}\right) = 2\int_{pq>0} p\log\left(\sqrt{\frac{p}{q}}\right)$$
$$= -2\int_{pq>0} p\log\left(\left[\sqrt{\frac{p}{q}}-1\right]+1\right)$$
$$\leq -2\int_{pq>0} p\log\left(\left[\sqrt{\frac{p}{q}}-1\right]+1\right)$$
$$= -2\int_{pq>0} p\left[\sqrt{\frac{p}{q}}-1\right]$$
$$= -2\int\sqrt{pq}-1 = H^2(P,Q).$$

□

Juntando los lemas anteriores, se obtiene el siguiente resultado:

Teorema 2.4.9.

$$V(P,Q) \leq H(P,Q) \leq \sqrt{K(P,Q)}. \tag{2.43}$$

Sin embargo, ésta no es la desigualdad más ajustada que se puede obtener entre $V(P,Q)$ y $\sqrt{K(P,Q)}$. Para esa desigualdad, hace falta un lema previo:

Lema 2.4.10. Si $P \ll Q$, entonces

$$\int \left(\log \frac{dP}{dQ}\right)_- dP \leq V(P,Q), \tag{2.44}$$

siendo $(a)_- = \max\{0, -a\}$.

Demostración.

Mediante el mismo razonamiento que el utilizado en 2.4.5, se cumple que:

$$\int \left(\log \frac{dP}{dQ}\right)_- d\nu = \int_{pq>0} p \left(\log \frac{p}{q}\right)_- d\nu$$

Sean $A_0 = \{x \in \mathcal{X} : q(x) \geq p(x)\}$, y $A_1 = \{q(x) \geq p(x) > 0\} = A_0 \cap \{p > 0\}$. Puesto que $s \log s \leq s - 1$, se cumple que

$$\begin{aligned}\int_{pq>0} p \left(\log \frac{p}{q}\right)_- d\nu &= \int_{p>0} p \log \frac{p}{q} d\nu = \int_{A_1} p \log \frac{p}{q} d\nu \\ &\leq \int_{A_1} (q-p) d\nu = Q(A_1) - P(A_1) \leq V(P,Q).\end{aligned}$$

□

Proposición 2.4.11 (Desigualdades de Pinsker.)**.**

1. $$V(P,Q) \leq \sqrt{K(P,Q)/2} \tag{2.45}$$

2. Si $P \ll Q$:
$$\int \left|\log \frac{dP}{dQ}\right| dP \equiv \int_{pq>0} p \left|\log \frac{p}{q}\right| d\nu \leq K(P,Q) + \sqrt{2K(P,Q)} \tag{2.46}$$

$$\Rightarrow \int \left(\log \frac{dP}{dQ}\right)_+ dP \leq K(P,Q) + \sqrt{K(P,Q)/2} \tag{2.47}$$

Demostración.

1. Consideramos la función

$$\psi(x) = \begin{cases} x \log x - x + 1 & \text{si } x > 0 \\ 0 & \text{si } x = 0 \end{cases}.$$

Observemos que $\psi(0) = 1$, $\psi(1) = 0$, $\psi'(1) = 0$ y $\psi''(x) = 1/x \geq 0$. Además, $\psi(x) \geq 0$ $\forall x \geq 0$.

Para poder obtener la desigualdad deseada, se tiene que demostrar que

$$\left(\frac{4}{3} + \frac{2}{3}x\right) \psi(x) \geq (x-1)^2. \tag{2.48}$$

Para el caso $x = 0$ es obvio, así que solo hace falta ver cuando $x > 0$.

Sea la función

$$g(x) := (x-1)^2 - \left(\frac{4}{3} + \frac{2}{3}x\right) \psi(x).$$

Dicha función cumple

$$g(1) = 0, \quad g'(1) \quad g''(x) = -\frac{4\psi(x)}{3x} \leq 0,$$

Así que, considerando ϵ tal que $|\epsilon - 1| < |x - 1|$, tenemos que considerando la extensión de Taylor de g:

$$g(x) = g(1) + g'(1)(x-1) + \frac{g''(\epsilon)}{2}(x-1)^2 = 0 + 0 - \frac{4\psi(x)}{3x} \leq 0$$

Luego se cumple la cota (2.48).

De (2.48), se obtiene que para $P \ll Q$:

$$\begin{aligned} V(P,Q) = \frac{1}{2}\int |p-q| &= \frac{1}{2}\int_{q>0} |\frac{p}{q} - 1| q \\ &\leq \frac{1}{2}\int_{q>0} q\sqrt{\left(\frac{4}{3} + \frac{2}{3}\frac{p}{q}\right)\psi\left(\frac{p}{q}\right)} && (2.48) \\ &\leq \frac{1}{2}\sqrt{\int\left(\frac{4q}{3} + \frac{2p}{3}\right)}\sqrt{\int_{q>0} q\left(\frac{p}{q}\right)} && (Cauchy - Schwarz) \\ &= \sqrt{\frac{1}{2}\int_{pq>0} p\log\frac{p}{q}} = \sqrt{K(P,Q)/2} \end{aligned}$$

2. A partir de descomponer la divergencia de Kullback en parte positiva y negativa, $K(P,Q) = \int_{pq>0} p\left(\log(p/q)\right)_+ d\nu - \int_{pq>0} p\left(\log(p/q)\right)_- d\nu$, se tiene que, aplicando tanto la primera desigualdad como el lema anterior:

$$\begin{aligned} \int_{pq>0} p\left|\log(p/q)\right| d\nu &= \int_{pq>0} p\left(\log(p/q)\right)_+ d\nu + \int_{pq>0} p\left(\log(p/q)\right)_- d\nu \\ &= K(P,Q) + 2\int_{pq>0} p\left(\log(p/q)\right)_- d\nu \\ \leq K(P,Q) + 2V(P,Q) \leq K(P,Q) + 2\sqrt{K(P,Q)/2} &= K(P,Q) + \sqrt{2K(P,Q)}. \end{aligned}$$

Trabajando sobre la parte positiva con la desigualdad obtenida:

$$\begin{aligned} \int_{pq>0} p\left(\log(p/q)\right)_+ d\nu &= \int_{pq>0} p\left(\log(p/q)\right)_+ d\nu - \int_{pq>0} p\left(\log(p/q)\right)_- d + \int_{pq>0} p\left(\log(p/q)\right)_- d\nu \\ &= K(P,Q) + \int_{pq>0} p\left(\log(p/q)\right)_- d\nu \leq K(P,Q) + V(P,Q) \leq K(P,Q) + \sqrt{K(P,Q)/2}. \end{aligned}$$

□

Otra relación que se puede encontrar entre la divergencia de Kullback y $V(\cdot,\cdot)$ es la siguiente:

Lema 2.4.12.

$$V(P,Q) \leq -\frac{1}{2}\exp(-K(P,Q)). \tag{2.49}$$

Demostración.

Como en casos anteriores, basta con demostrar el caso $K(P,Q) < \infty$, puesto que el otro caso es obvio.

Del mismo modo que a la hora de probar la primera desigualdad de Pinkster, la manera de relacionar la variación total con la divergencia de Kullback es a partir de la función logaritmo $f(x) = log(x)$. Lo más intuitivo sería utilizar la desigualdad de Jensen, pero $f(x) = log(x)$ no es

una función convexa así que no se puede aplicar. Sin embargo, $g(x) = -f(x) = -log(x)$ sí que lo es. Por tanto, aplicando la desigualdad de Jensen a g y se obtiene que que:

$$\begin{aligned} -\log\left(\int x d\nu\right) &\leq \int -\log(x)\, d\nu \\ \Rightarrow \log\left(\int x d\nu\right) &\geq \int \log(x)\, d\nu. \end{aligned}$$

Como $P \ll Q$, y considerando la desigualdad anterior, se obtiene que (recordemos que $p = dP/d\nu$):

$$\begin{aligned} \left(\int \sqrt{pq}\right)^2 &= \exp\left(2\log\int_{pq>0}\sqrt{pq}\right) = \exp\left(2\log\int_{pq>0} p\sqrt{\frac{q}{p}}\right) \\ &\geq \exp\left(2\int_{pq>0} p\log\sqrt{\frac{q}{p}}\right) = \exp(-K(P,Q)) \end{aligned}$$

Finalmente, usando la desigualdad de Le Cam (2.41):

$$\begin{aligned} \int \text{mín}(dP, dQ) &\geq \frac{1}{2}\left(\int\sqrt{dPdQ}\right)^2 \geq \frac{1}{2}\exp(-K(P,Q)) \\ \Rightarrow V(P,Q) &= 1 - \int \text{mín}(dP, dQ) \geq 1 - \frac{1}{2}\exp(-K(P,Q)) \end{aligned}$$

□

Observación 2.4.4. A la hora de hacer comparaciones entre la divergencia de Kullback y la variación total, se han obtenido varias posibles desigualdades. Juntando ambas desigualdades obtenidas, se puede considerar la siguiente relación:

$$V(P,Q) \leq \text{mín}\left\{1 - \frac{1}{2}\exp(-K(P,Q)), \sqrt{K(P,Q)/2}\right\}. \tag{2.50}$$

2.5. Acotaciones a través de la probabilidad del error minimax

El objetivo de esta sección es probar un inferior general para el riesgo minimax. Concretamente (ver Corolario 2.5.8), se demostrará que si se eligen $\{\theta_0, \ldots, \theta_M\} \subset \Theta$ tales que

1. $d(\theta_j, \theta_k) \geq 2s$, con $j \neq k$.
2. $P_{\theta_j} \ll P_{\theta_0}$ para $j = 1, \ldots, M$, y

$$\frac{1}{M}\sum_{j=1}^{M} K(P_{\theta_j}, P_{\theta_0}) \leq \alpha \log M \qquad \text{con } 0 < 1/8.$$

Es decir, el promedio de las divergencias de Kullback de las hipótesis está "controlado".

Entonces, si $A > 0$, y $\psi = s/A$, se obtiene que

$$\inf_{\hat{\theta}_n}\sup_{\theta\in\Theta} \mathbb{E}_\theta(\psi^{-2}d^2(\hat{\theta}_n, \theta)) \geq c(\alpha)\cdot A^2,$$

donde $c(\alpha)$ es una constante que únicamente dependerá de α ((2.64)).

De acuerdo con lo estudiado en la Sección 2.3, una forma de buscar acotaciones inferiores del riesgo es buscándolas para $p_{e,M}$. Durante el resto del capítulo se intentarán conseguir cotas de la probabilidad de error minimax, terminando con el Teorema Principal de las Cotas inferiores del

Riesgo, enunciado en el Teorema 2.5.7.

Primero se considerará el caso más simple, cuando $M = 1$. Este caso corresponde con el problema de contraste entre dos hipótesis, θ_0 y θ_1. Por simplificación, se denotarán a las respectivas probabilidades de cada hipótesis $P_i = P_{\theta_i}$.

Por el Teorema de Descomposición de Jordan-Hahn, (Teorema A.2.4 de la Sección A.2), P (al ser una medida positiva) se puede descomponer $P_0 = P_0^a + P_0^s$, donde P_0^a es la componente absolutamente continua de P_0 respecto a P_1, y P_0^s la parte singular respecto a P_1. Esto sugiere considerar $\frac{dP_0^a}{dP_1}$. Puesto que P_0^a puede expresarse en la probabilidad P_1 ($P_0^a \ll P_1$), se puede sacar una acotación del valor de $p_{e,1}$ a partir del comportamiento de $\frac{dP_0^a}{dP_1}$.

Proposición 2.5.1. Sean θ_0, θ_1 hipótesis en Θ, y P_0, P_1 sus respectivas probabilidades en $(\mathcal{X}, \mathcal{A})$. Denotando $p_{e,1}$ a la probabilidad del error minimax, se tiene que:

$$p_{e,1} \geq \sup_{\tau>0} \left\{ \frac{\tau}{1+\tau} P_1 \left(\frac{dP_0^a}{dP_1} \geq \tau \right) \right\}.$$

Demostración.

Se denotará

$$p := P_1(\psi = 1) \qquad y \qquad \alpha_1 := P_1 \left(\frac{dP_0^a}{dP^1} \geq \tau \right)$$

.

Con esta notación, nuestro objetivo es ver que

$$p_{e,1} \geq \sup_{\tau>0} \left\{ \frac{\tau}{1+\tau}(1-\alpha_1) \right\}.$$

Sea $\tau > 0$ fijo. Para cualquier test $\Psi : \mathcal{X} \to \{0, 1\}$, tenemos:

$$\begin{aligned} P_0(\Psi \neq 0) = P_0(\Psi = 1) &\geq P_0^a(\Psi = 1) \\ &= \int I(\Psi = 1) \frac{dP_0^a}{dP_1} dP_1 \\ &\geq \tau \int I\left(\{\Psi = 1\} \cap \left\{ \frac{dP_0^a}{dP_1} \geq \tau \right\} \right) dP_1 \\ &\geq \tau \left(P_1(\Psi = 1) - P_1 \left(\frac{dP_0^a}{dP_1} < \tau \right) \right) = \tau(p - \alpha_1). \end{aligned}$$

De esta manera se obtiene una acotación superior para $P_0(\Psi \neq 0)$. Puesto que $P_1(\Psi \neq 1) = 1 - P_1(\Psi = 1) = 0$, también se tiene una actoación para $P_1(\Psi \neq 1)$, así juntando ambas se consigue una acotación para la probabilidad de error minimax:

$$p_{e,1} = \inf_{\Psi} \max_{j=0,1} P_j(\Psi \neq j) \geq \min_{0 \leq p \leq 1} \max \{\tau(p - \alpha_1), 1 - p\} = \frac{\tau(1-\alpha_1)}{1+\tau}$$

para todo $\tau > 0$.

Como la desigualdad se cumple para todo τ:

$$p_{e,1} \geq \sup_{\tau>0} \left\{ \frac{\tau}{1+\tau} P_1 \left(\frac{dP_0^a}{dP_1} \geq \tau \right) \right\}.$$

□

Esta proposición implica que para obtener una cota inferior a la probabilidad de error minimax, es suficiente con encontrar valores constantes para $\tau > 0$ y $0 < \alpha < 1$ tal que se cumpla que

$$P_1\left(\frac{dP_0^a}{dP_1} \geq \tau\right) \geq 1 - \alpha. \tag{2.51}$$

Juntando este resultado con el Teorema 2.3.1, se obtiene el siguiente corolario

Corolario 2.5.2. Sean $\theta_0, \theta_1 \in \Theta$ tales que satisfacen $d(\theta_0, \theta_1) \geq 2s > 0$. Entonces, se cumple que

$$R_n^* \geq \inf_{\hat{\theta}_n} \sup_{\theta \in \{0,1\}} P_\theta(d(\hat{\theta}, \theta) \geq s) \geq \sup_{\tau > 0} \left\{ \frac{\tau}{1+\tau} P_1\left(\frac{dP_0^a}{dP_1} \geq \tau\right) \right\}.$$

Observación 2.5.1.

1. En caso de que $P_0 \ll P_1$ (y que, por tanto, se tenga que $P_0^a = P_0$), $\frac{dP_0}{dP_1}(\mathbf{X})$ es la **razón de verosimilitud** entre las hipótesis θ_0 y θ_1, y $\left(\frac{dP_0^a}{dP_1} \geq \tau\right)$ es la región de rechazo del test más potente de θ_1 frente a θ_0 (por la notación que se está utilizando, θ_1 juega el papel de la hipótesis nula, aunque no sea la habitual).

   El Lema de Neyman Pearson dicta que el test más potente para contrastar las dos hipótesis es $P_1\left(\frac{dP_0^a}{dP_1} \geq \tau\right)$. Esto implica que, cuando $P_0 \ll P_1$, la ecuación (2.51) -y por tanto el Teorema 2.5.2- viene a acotar inferiormente la probabilidad del error minimax por la potencia del test más potente para cierto α.

2. La condición entre P_0^a y P_1 que nos da la desigualdad (2.51) significa que ambas probabilidades son "lo suficientemente cercanas".

   En el caso extremo en el que $P_0 = P_1$, se tiene que

   $$P_0 = P_1 \Rightarrow P_0^a = P_1^a = P_1 \Rightarrow \frac{P_0^a}{dP_1} = \frac{dP_1}{dP_1} = 1.$$

   Substituyendo en (2.51), se obtiene

   $$P_1(1 \geq \tau) \geq 1 - \alpha$$

   así que se puede considerar $\tau = 1$ y $\alpha = 0$ ($P_1(1 \geq 1) \geq 1$). Substituyendo dichos valores en la desigualdad de la proposición anterior, se obtiene

   $$\begin{aligned} p_{e,1} &\geq \sup_{\tau > 0} \left\{ \frac{\tau}{1+\tau} P_1\left(\frac{dP_0^a}{dP_1} \geq \tau\right) \right\} \\ &\underset{\tau=1}{\geq} \frac{1}{1+1} P_1\left(\frac{dP_1}{dP_1} \geq 1\right) = \frac{1}{2}. \end{aligned}$$

   Esto implica que en caso de que $P_0 = P_1$, la mejor acotación que se puede obtener de esta manera sería $1/2$, que no resulta muy útil.

El Teorema 2.5.2 da una primera acotación a la probabilidad de error minimax (y, por tanto, al riesgo minimax). Sin embargo, esta acotación supone calcular el supremo de varias probabilidades, algo que no siempre será posible. Esto supone tener que buscar ciertas condiciones extra sobre las probabilidades de las hipótesis P_0 y P_1 para poder acotar $p_{e,1}$.

Puesto que en la función test Ψ^* se considera la distancia entre P_0 y P_1, puede buscarse una cota según la distancia en probabilidad entre las hipótesis. En concreto, se podrá obtener una acotación a partir del valor de $V(P_0, P_1)$ (que se podrá extender al resto de distancias estudiadas a partir de las desigualdades obtenidas en el Apartado 2.4.1).

Teorema 2.5.3. Sean P_0 y P_1 dos probabilidades en $(\mathcal{X}, \mathcal{A})$.

1. Si $V(P_1, P_0) \leq \alpha < 1$, entonces

   $$p_{e,1} \geq \frac{1-\alpha}{2} \tag{2.52}$$

2. Si $H^2(P_1, P_0) \leq \alpha < 2$, entonces:

$$p_{e,1} \geq \frac{1}{2}(1 - \sqrt{\alpha(1 - \alpha/4)}) \tag{2.53}$$

3. Si $K(P_1, P_0) \leq \alpha < \infty$, entonces

$$p_{e,1} \geq \text{máx}\left(\frac{1}{4}\exp(-\alpha), \frac{1 - \sqrt{\alpha/2}}{2}\right) \tag{2.54}$$

Demostración.

1. Primero, recordemos que $p_{e,1}$ cumple que:

$$\begin{aligned} p_{e,1} &= \inf_{\Psi} \max_{j=0,1} P_j(\Psi \neq j) \geq \frac{1}{2} \inf_{\Psi}(P_0(\Psi \neq 0) + P_1(\Psi \neq 1)) \\ &= \frac{1}{2}(P_0(\Psi^* \neq 0) + P_1(\Psi^* \neq 1)) \end{aligned}$$

siendo Ψ^* el test de máxima verosimilitud:

$$\Psi^* = \begin{cases} 0 & \text{si } p_0 \geq p_1 \\ 1 & \text{si } p_0 < p_1 \end{cases}$$

siendo p_0, p_1 las respectivas densidades de las probabilidades P_0 y P_1.
Esto implica que $\text{máx}(P_0, P_1) = P_0$ si $\Psi^* = 0$ y $\text{máx}(P_0, P_1) = P_1$ si $\Psi^* = 1$. Dicho de otra manera,

$$\begin{aligned} &\int \text{máx}(dP_0, dP_1) = P_0(\Psi^* = 0) + P_1(\Psi^* = 1) \\ \Rightarrow &\int \text{mín}(dP_0, dP_1) = 2 - \int \text{máx}(dP_0, dP_1) = 2 - (P_0(\Psi^* = 0) + P_1(\Psi^* = 1)) \\ &= 1 - (P_0(\Psi^* = 0) + 1 - P_1(\Psi^* = 1)) = P_0(\Psi^* \neq 0) + P_1(\Psi^* \neq 1). \end{aligned}$$

Aplicando la igualdad del Teorema de Scheffé (2.4.3) a lo anterior, se obtiene que:

$$\frac{1}{2}(P_0(\Psi^* \neq 0) + P_1(\Psi^* \neq 1)) = \frac{1}{2}\int \text{mín}(dP_0, dP_1) = (1 - V(P_0, P_1))/2.$$

Por tanto, obtenemos:

$$\begin{aligned} p_{e,1} &\geq \frac{1 - V(P_0, P_1)}{2} \\ \Longrightarrow p_{e,1} &\geq \frac{1 - V(P_0, P_1)}{2} \geq \frac{1 - \alpha}{2} \end{aligned}$$

.

2. Aplicando la desigualdad de LeCam (2.41) :

$$V(P, Q) \leq H(P, Q)\sqrt{1 - H^2(P, Q)/4},$$

por lo que

$$p_{e,1} \geq \frac{1 - V(P_0, P_1)}{2} \geq \frac{1 - H(P, Q)\sqrt{1 - H^2(P, Q)/4}}{2} \geq \frac{1 - \alpha\sqrt{1 - \alpha/4}}{2}$$

3. Si en vez de aplicar la desigualdad de LeCam aplicamos lo dicho en la Observación 2.4.4, tenemos que

$$p_{e,1} \geq \frac{1 - V(P_0, P_1)}{2} \geq \frac{1 - \text{mín}\left\{1 - \frac{1}{2}\exp(-K(P,Q)), \sqrt{K(P,Q)/2}\right\}}{2}$$
$$\geq \text{máx}\left(\frac{1}{4}\exp(-\alpha), \frac{1-\sqrt{\alpha/2}}{2}\right)$$

□

Con esto, hemos obtenido una probabilidad del error minimax inferior al error medio para cualquiera de las tres distancias.

Ya hemos visto varias acotaciones que podemos conseguir para cuando se manejan dos hipótesis. Sin embargo, habrá casos donde la acotación obtenida para la probabilidad de riesgo con dos hipótesis no es suficiente para obtener una cota suficientemente buena del estimador.

Es por ello que se necesita extender las proposiciones obtenidas para dos hipótesis (en concreto, las Proposiciones 2.5.1 y 2.5.2, y el Teorema 2.5.3) al caso general. Para ello, se seguirán los resultados dados en [3, sec. 2.6] .

El primer resultado que habría que generalizar es la acotación a la probabilidad del riesgo, dada en el Teorema 2.5.1:

$$p_{e,1} \geq \sup_{\tau>0}\left\{\frac{\tau}{1+\tau} P_1\left(\frac{dP_0^a}{dP_1} \geq \tau\right)\right\}.$$

El mayor problema para extender este resultado al caso general es el hecho de que cada valor de $A_j := \left\{\frac{dP_{0,j}^a}{dP_j}\right\}$ es distinto para cada j, por lo que se tendrá que buscar una manera de "normalizar" esos valores para poder obtener una desigualdad similar a la de la Proposición 2.5.1.

Proposición 2.5.4. Sean $\theta_0, \theta_1, \ldots, \theta_M$ y conjunto de hipótesis en Θ, y $P_0, P_1, \ldots, P_M$ sus respectivas probabilidades en $(\mathcal{X}, \mathcal{A})$. Entonces,

$$p_{e,M} \geq \sup_{\tau>0}\left\{\frac{\tau M}{1+\tau M}\left[\frac{1}{M}\sum_{j=1}^{M} P_j\left(\frac{dP_{0,j}^a}{dP_j} \geq \tau\right)\right]\right\} \tag{2.55}$$

siendo $P_{0,j}$ la componente absolutamente continua de la probabilidad P_j en al medida P_0.

Demostración.

Sea Ψ una función test que toma valores en $\{0, 1, \ldots, M\}$.

Notemos que:

$$\bigcup_{j=1}^{M}\{\Psi = j\} = \{\Psi \neq 0\},$$

y que, al ser Ψ una función test:

$$\{\Psi = j\} \cap \{\Psi \neq k\} = \emptyset \text{ para todo } k \neq j.$$

Denotemos a los eventos aleatorio $A_j := \left\{\frac{dP_{0,j}^a}{dP_j} \geq \tau\right\}$, y

$$p_0 := \frac{1}{M}\sum_{j=1}^{M} P_j(\Psi = j) \quad y \quad \alpha := \frac{1}{M}\sum_{j=1}^{M} P_j\left(\frac{dP_{0,j}^a}{dP_j} > \tau\right).$$

Se obtiene que:

$$\begin{aligned}
P_0(\Psi \neq 0) &= \sum_{j=1}^{M} P_0(\Psi = j) \geq \sum_{j=1}^{M} P_{0,j}^a(\Psi = j) \\
&\geq \sum_{j=1}^{M} \tau P_j(\{\Psi = j\} \cap A_j) \\
&\geq \tau M \left(\frac{1}{M} \sum_{j=1}^{M} P_j(\Psi = j) \right) - \tau \sum_{j=1}^{M} P_j(A_j^C) \\
&= \tau M(p_0 - \alpha).
\end{aligned}$$

Entonces, para todo τ se tiene:

$$\begin{aligned}
\max_{0 \leq j \leq M} P_j(\Psi \neq j) &= \max \left\{ P_0(\Psi \neq 0), \max_{\Psi \neq j} P_j(\Psi \neq j) \right\} \\
&\geq \max \left\{ \tau M(p_0 - \alpha), \frac{1}{M} \sum_{j=1}^{M} P_j(\Psi \neq j) \right\} \\
&\geq \max \{ \tau M(p_0 - \alpha), 1 - p_0 \} \\
&\geq \min_{0 \leq p \leq 1} \max \{ \tau M(p - \alpha), 1 - p \}
\end{aligned}$$

.

Como $\tau M(p - \alpha)$ es creciente respecto al valor de τ, a la vez que $1 - p$ es decreciente, el mínimo se dará en el valor que coincidan (siempre y cuando ese valor se encuentre en $[0, 1]$):

$$\begin{aligned}
\tau M(p - \alpha) &= 1 - p \\
p + \tau M p - \tau M \alpha &= 1 \\
p(1 + \tau M) &= 1 + \tau M \alpha \\
p &= \frac{1 + \tau M \alpha}{1 + \tau M}
\end{aligned}$$

luego el mínimo se dará cuando $p = \dfrac{1 + \tau M \alpha}{1 + \tau M}$, lo que implica que

$$\begin{aligned}
\max_{0 \leq j \leq M} P_j(\Psi \neq j) &\geq \min_{0 \leq p \leq 1} \max \{ \tau M(p - \alpha), 1 - p \} \\
&= 1 - \frac{1 + \tau M \alpha}{1 + \tau M} = \frac{\tau M(1 - \alpha)}{1 + \tau M}.
\end{aligned}$$

Como esto se cumple para todo τ, también se dará para el supremo:

$$p_{e,M} \geq \sup_{\tau > 0} \left(\frac{\tau M}{1 + \tau M}(1 - \alpha) \right) \geq \sup_{\tau > 0} \left(\frac{\tau M}{1 + \tau M} \left[\frac{1}{M} \sum_{j=1}^{M} P_j \left(A_j \right) \right] \right).$$

□

Si se conoce una acotación de $\frac{1}{M} \sum_{j=1}^{M} P_j \left(\frac{dP_{0,j}^a}{dP_j} \geq \tau \right)$, entonces se puede dar una acotación a la probabilidad de error minimax:

Proposición 2.5.5. Suponemos que Θ contiene los elementos $\theta_0, \theta_1, \ldots, \theta_M$, que cumplen:

1. $d(\theta_j, \theta_k) \geq 2s > 0$, $\forall\, 0 \leq j < k \leq M$
2. Existe $\tau > 0$ y $0 < \alpha < 1$ tales que:

$$\frac{1}{M} \sum_{j=1}^{M} P_j \left(\frac{dP_{0,j}^a}{dP_j} \geq \tau \right) \geq 1 - \alpha \tag{2.56}$$

siendo $P_{0,j}^a$ la componente absolutamente continua de la medida $P_0 = P_{\theta_0}$ respecto a $P_j = P_{\theta_j}$.

Entonces, se cumple que

$$\inf_{\hat{\theta}_n} \sup_{\theta\in\Theta} P_\theta(d(\hat{\theta}_n,\theta) \geq s) \geq p_{e,M} \geq \frac{\tau M}{1+\tau M}(1-\alpha).$$

Demostración.

Al cumplirse que las hipótesis $d(\theta_j,\theta_k) \geq 2s > 0$, $\forall\, 0 \leq j < k \leq M$, se puede acotar el riesgo minimax por la probabilidad del riesgo (Teorema 2.3.1) y obtener que

$$\inf_{\hat{\theta}_n} \sup_{\theta\in\Theta} P_\theta(d(\hat{\theta}_n,\theta) \geq s_n) \geq \inf_{\hat{\theta}_n} \max_{0\leq j\leq M} P_j(\Psi^* \neq j) = p_{e,M}.$$

En la proposición anterior, se obtuvo que

$$p_{e,M} \geq \sup_{\tau>0}\left\{\frac{\tau M}{1+\tau M}\left[\frac{1}{M}\sum_{j=1}^{M} P_j\left(\frac{dP^a_{0,j}}{dP_j} \geq \tau\right)\right]\right\}.$$

Por hipótesis, se asume existen $\tau^* > 0$ y $0 < \alpha < 1$ tales que:

$$\frac{1}{M}\sum_{j=1}^{M} P_j\left(\frac{dP^a_{0,j}}{dP_j} \geq \tau^*\right) \geq 1-\alpha,$$

lo que implica que

$$\begin{aligned}\inf_{\hat{\theta}_n} \sup_{\theta\in\Theta} P_\theta(d(\hat{\theta}_n,\theta) \geq s_n) \geq p_{e,M} &\geq \sup_{\tau>0}\left\{\frac{\tau M}{1+\tau M}\left[\frac{1}{M}\sum_{j=1}^{M} P_j\left(\frac{dP^a_{0,j}}{dP_j} \geq \tau\right)\right]\right\}\\ &\geq \frac{\tau M}{1+\tau M}\left[\frac{1}{M}\sum_{j=1}^{M} P_j\left(\frac{dP^a_{0,j}}{dP_j} \geq \tau^*\right)\right]\\ &\geq \frac{\tau M}{1+\tau M}(1-\alpha),\end{aligned}$$

que da con la desigualdad deseada.

□

Esto implica que para tener un resultado similar al Teorema 2.5.3, basta con encontrar ciertas condiciones sobre las distancias entre las distintas probabilidades P_j para que se cumpla la desigualdad (2.56). Para varias hipótesis, solo se podrá considerar la divergencia de Kullback.

Proposición 2.5.6. Sean $P_0, P_1, \ldots, P_M$ medidas probabilísticas en $(\mathcal{X},\mathcal{A})$ tales que:

$$\frac{1}{M}\sum_{j=1}^{M} K(P_j,P_0) \leq \alpha_* \tag{2.57}$$

con $0 < \alpha_* < \infty$. Entonces

$$p_{e,M} \geq \sup_{\tau>0}\left[\frac{\tau M}{1+\tau M}\left(1+\frac{\alpha_*+\sqrt{\alpha_*/2}}{\log\tau}\right)\right]. \tag{2.58}$$

Demostración.

Por la proposición anterior, es suficiente con ver que para todo $0 < \tau < 1$ se cumple que :

$$\frac{1}{M}\sum_{j=1}^{M} P_j\left(\frac{dP^a_{0,j}}{dP_j} \geq \tau\right) \geq 1 - \frac{\alpha_*+\sqrt{\alpha_*/2}}{\log\tau}.$$

Es decir, que $\alpha = -\dfrac{\alpha_*+\sqrt{\alpha_*/2}}{\log\tau}$.

Como $K(P_j, P_0)$ es finito para todo j por (2.57), se tiene que $P_j \ll P_0$. Esto implica que $\frac{dP_j}{dP_0} = \frac{dP_j}{dP^a_{0,j}}$ casi siempre en la medida P_j. Por tanto,

$$\begin{aligned} P_j\left(\frac{dP_j}{dP^a_{0,j}} \geq \tau\right) &= P_j\left(\frac{dP_0}{dP_j} \leq \frac{1}{\tau}\right) = 1 - P_j\left(\log\frac{dP_j}{dP_0} > \log\frac{1}{\tau}\right) \\ &\geq 1 - \frac{1}{\log(1/\tau)}\int\left(\log\frac{dP_j}{dP_0}\right)_+ dP_j \qquad \text{Desigualdad de Markov} \\ &\geq 1 - \frac{1}{\log(1/\tau)}\left[K(P_j,P_0) + \sqrt{K(P_j,P_0)/2}\right] \qquad 2^{\underline{a}} \text{ desigualdad de Pinsker(2.47)} \end{aligned}$$

Por otra parte, como $f(x) = -\sqrt{x}$ es una función convexa, podemos aplicar la desigualdad de Jensen, obtener

$$\frac{1}{M}\sum_{j=1}^{M} -\sqrt{K(P_j,P_0)} = \frac{1}{M}\sum_{j=1}^{M} f(K(P_j,P_0)) \geq f\left(\frac{1}{M}\sum_{j=1}^{M} K(P_j,P_0)\right) \leq -\sqrt{\alpha_*} \tag{2.59}$$

Esto implica que

$$\begin{aligned} \frac{1}{M}\sum_{j=1}^{M} P_j\left(\frac{dP^a_{0,j}}{dP_j} \geq \tau\right) &= \frac{1}{M}\sum_{j=1}^{M}\left(1 - \frac{1}{\log(1/\tau)}\left[K(P_j,P_0) + \sqrt{K(P_j,P_0)/2}\right]\right) \\ &= 1 - \left[\frac{1}{M}\frac{1}{\log(1/\tau)}\sum_{j=1}^{M}\left(K(P_j,P_0) + \sqrt{K(P_j,P_0)/2}\right)\right] \\ &= 1 - \frac{1}{\log(1/\tau)}\left[\frac{1}{M}\sum_{j=1}^{M} K(P_j,P_0) + \frac{1}{M}\sum_{j=1}^{M}\sqrt{K(P_j,P_0)/2}\right] \\ &\geq 1 - \frac{1}{\log(1/\tau)}\left[\alpha_* + \sqrt{\alpha_*/2}\right], \end{aligned}$$

que es a lo que queríamos llegar.

□

Teorema 2.5.7 (Teorema Principal de las Cotas Inferiores del Riesgo)**.** Sea $M \geq 2$, y sean $\theta_0, \theta_1, \ldots, \theta_M$ elementos de Θ tales que cumplen:

1. $d(\theta_j, \theta_k) \geq 2s > 0$, $\forall\, 0 \leq j < k \leq M$;
2. $P_j \ll P_0$, $\forall\, j = 0, \ldots, M$, y

$$\frac{1}{M}\sum_{j=1}^{M} K(P_j, P_0) \leq \alpha \log M \tag{2.60}$$

con $0 < \alpha < 1/8$ y $P_j = P_{\theta_j}$ para $j = 0, 1, \ldots, M$.

Entonces, se tiene que:

$$\inf_{\hat\theta}\sup_{\theta\subset\Theta}\left(P_\theta(d(\hat\theta,\theta) \geq s)\right) \geq \frac{\sqrt{M}}{1+\sqrt{M}}\left(1 - 2\alpha - \sqrt{\frac{2\alpha}{\log M}}\right) > 0. \tag{2.61}$$

Demostración.

Basta con aplicar el teorema anterior, considerando $\alpha_* = \alpha \log M$ con $\alpha < 1/8$:

$$\begin{aligned}
p_{e,M} &\geq \sup_{\tau>0}\left[\frac{\tau M}{1+\tau M}\left(1+\frac{\alpha_* + \sqrt{\alpha_*/2}}{\log \tau}\right)\right] \\
&\geq \frac{\sqrt{M}}{1+\sqrt{M}}\left(1+\frac{\alpha_* + \sqrt{\alpha_*/2}}{\log(1/\sqrt{M})}\right) && (\text{ Considerando } \tau = 1/\sqrt{M}\ ) \\
&= \frac{\sqrt{M}}{1+\sqrt{M}}\left(1+\frac{\alpha_* + \sqrt{\alpha_*/2}}{-\frac{1}{2}\log(M)}\right) && (\text{ } \log(1/\sqrt{M}) = -1/2 \log M\ ) \\
&= \frac{\sqrt{M}}{1+\sqrt{M}}\left(1-\frac{\alpha_*}{1/2\log M} - \sqrt{\frac{\alpha_*}{2/4\log^2 M}}\right) \\
&= \frac{\sqrt{M}}{1+\sqrt{M}}\left(1-2\frac{\alpha_*}{\log M} - \sqrt{\frac{2\alpha_* \log M}{\log^2 M}}\right) \\
&= \frac{\sqrt{M}}{1+\sqrt{M}}\left(1-2\frac{\alpha_*}{\log M} - \sqrt{\frac{2\alpha_*}{\log M}}\right)
\end{aligned}$$

Esta cota obtenida es no nula, puesto que $M \geq 2 \Rightarrow \log M \geq \log 2$, lo que implica que

$$\begin{aligned}
&\frac{\sqrt{M}}{1+\sqrt{M}}\left(1-2\frac{\alpha_*}{\log M} - \sqrt{\frac{2\alpha}{\log M}}\right) \\
&\geq \frac{\sqrt{M}}{1+\sqrt{M}}\left(1-2\alpha - \sqrt{\frac{2\alpha}{\log 2}}\right) > 0
\end{aligned}$$

□

Observación 2.5.2. El resultado anterior se cumple para cualquier $M \geq 2$. Se podría considerar extendiéndolo al caso $M = 1$, pero puesto que $log(1) = 0$, la tercera hipótesis implicaría que $K(P_0, P_1) = 0$. En ese caso, se puede considerar cualquier valor de α en el Teorema 2.5.3. Como para $0 < \alpha < 1/8$

$$\text{máx}\left(\frac{1}{4}\exp(-\alpha), \frac{1-\sqrt{\alpha/2}}{2}\right) = \frac{1-\sqrt{\alpha/2}}{2}$$

(puesto que $1/4 \exp(-\alpha) \leq 1/4$ para $\alpha > 0$), y $\frac{1-\sqrt{\alpha/2}}{2}$ es una función decreciente, se cumple que bajo las mismas hipótesis que el teorema anterior que

$$p_{e,1} \leq \frac{1-\sqrt{0/2}}{2} = \frac{1}{2}.$$

Lo dicho en la observación anterior implica que el teorema anterior se puede extender para $M = 1$:

Corolario 2.5.8. Sea $M \geq 1$, y $\theta_0, \theta_1, \ldots, \theta_M \in \Theta$ tales que:

1. $d(\theta_j, \theta_k) \geq 2s > 0$, $\forall\, 0 \leq j < k \leq M$;
2. $P_j \ll P_0$, $\forall\, j = 0, \ldots, P_0$, y
$$\frac{1}{M}\sum_{j=1}^{M} K(P_j, P_0) \leq \alpha \log M \tag{2.62}$$
con $0 < \alpha < 1/8$ y $P_j = P_{\theta_j}$ para $j = 0, 1, \ldots, M$.

Entonces, se tiene que:

$$\inf_{\hat{\theta}} \sup_{\theta\in\Theta} \mathbb{E}_\theta\left[\psi^{-2} d^2(\hat{\theta}, \theta)\right] \geq \inf_{\hat{\theta}} \sup_{\theta\subset\Theta}\left(P_\theta(d(\hat{\theta}, \theta) \geq s)\right) \geq c(\alpha) > 0, \tag{2.63}$$

siendo

$$c(\alpha) := \begin{cases} 1/2 & \text{si M=1} \\ \dfrac{\sqrt{M}}{1+\sqrt{M}}\left(1-2\alpha-\sqrt{\dfrac{2\alpha}{\log 2}}\right) & \text{si } M \geq 2 \end{cases} \tag{2.64}$$

En casos en los que las cotas obtenidas utilizando dos hipótesis no sean satisfactorias, se tendrá que hacer uso de las acotaciones obtenidas en esta sección. Sin embargo, cabe recordar que las proposiciones presentadas en esta sección requieren de una condición adicional sobre las probabilidades de las hipótesis seleccionas. Esto implicará que habrá que hacer una selección meticulosa de las hipótesis para cuando se necesite considerar varias, teniendo que escogerlas tales que $\frac{1}{M}\sum_{j=1}^{M} K(P_j, P_0) \leq \alpha \log M$ además de que $d(\theta_j, \theta_k) \geq 2s$.

2.6. Refinamientos: Lema de Fano

Las acotaciones conseguidas a partir de la probabilidad de error minimax ya ofrecen una acotación inferior al riesgo minimax que nos permite obtener la tasa óptima de convergencia del conjunto. El Lema de Fano no sólo permitirá ajustar más la acotación obtenida, sino que además la cota obtenida será sobre la probabilidad de error media, que permitirá extender dicha acotación inferior obtenida a un modelo más general en el capítulo final del documento (Sección 3.4).

Definición 2.6.1. Sea $P_0, \ldots, P_M$ probabilidades sobre el espacio medible $(\mathcal{X}, \mathcal{A})$. Para un test $\Psi : \mathcal{X} \to \{1, \ldots, \theta_M\}$ se define la **probabilidad del error media** $\overline{p}_{e,M}$ a la media muestra de la probabilidad de fallo del test:

$$\overline{p}_{e,M}(\Psi) := \frac{1}{M+1}\sum_{j=0}^{M} P_j(\Psi \neq j)$$

y la mínima probabilidad de error media $\overline{p}_{e,M}$ es

$$\overline{p}_{e,M} := \inf_{\Psi} \overline{p}_{e,M}(\Psi)$$

Del mismo modo, se denotará **probabilidad media** a la siguiente probabilidad en $(\mathcal{X}, \mathcal{A})$:

$$\overline{P} := \frac{1}{M+1}\sum_{j=0}^{M} P_j$$

Se puede escribir la probabilidad de error media de cierto test en términos de la esperanza respecto a la probabilidad media

Proposición 2.6.1. Sea Ψ una función test y $P_0, \ldots, P_M$ portabilidades en $(\mathcal{X}, \mathcal{A})$.

Denotando

$$b_j = I(A_j) \qquad y \qquad p_j := \frac{1}{M+1}\frac{dP_j}{d\overline{P}},$$

se obtiene que

$$\overline{p}_{e,M}(\Psi) = \mathbb{E}_{\overline{P}}\left[\sum_{j=0}^{M} b_j p_j\right] = \mathbb{E}_{\overline{P}}\left[\sum_{j\neq j_0} p_j\right],$$

para cierto $j_0 \in \{0, \ldots, M\}$.

Demostración.

Se denotará $A_j = \{\Psi \neq j\}$.

A partir de la definición de $\overline{p}_{e,M}$, y de la regla de la cadena de la derivada de Radon-Nikodym (Teorema A.2.2):

$$\begin{aligned}\overline{p}_{e,M}(\Psi) &= \frac{1}{M+1}\sum_{j=0}^{M} P_j(\Psi \neq j) = \frac{1}{M+1}\sum_{j=0}^{M} \mathbb{E}_j\left[I(A_j)\right] \\ &= \frac{1}{M+1}\sum_{j=0}^{M} \mathbb{E}_{\overline{P}}\left[I(A_j)\frac{dP_j}{d\overline{P}}\right] = \frac{1}{M+1}\mathbb{E}_{\overline{P}}\left[\sum_{j=0}^{M} I(A_j)\frac{dP_j}{d\overline{P}}\right] \\ &= \mathbb{E}_{\overline{P}}\left[\sum_{j=0}^{M} b_j\, p_j\right],\end{aligned}$$

Por una parte, al ser Ψ una función inyectiva, se cumple que $b_j \in \{0,1\}$ para todo j y que $\sum_{j=0}^{M} b_j = M$. Es decir, existe $j_0 \in \{0,\dots,M\}$ aleatorio tal que $b_{j_0} = 0$, y $b_j = 1$ para el resto de valores.

Por otra parte,

$$\sum_{j=0}^{M} \frac{dP_j}{d\overline{P}} = \sum_{j=0}^{M} \frac{dP_j}{d\overline{P}} = \sum_{j=0}^{M} \frac{P_j}{\overline{P}} = M+1,$$

por lo que $\sum_{j=0}^{M} p_j = 1$.

Uniendo ambas cosas, se obtiene que

$$\sum_{j=0}^{M} b_j\, p_j = \sum_{j\neq j_0} p_j.$$

Dando con la igualdad final:

$$\overline{p}_{e,M}(\Psi) = \mathbb{E}_{\overline{P}}\left[\sum_{j\neq j_0} p_j\right], \tag{2.65}$$

□

Esto sugiere intentar relacionar la esperanza de (2.65) con la divergencia de Kullback entre las probabilidades P_j y la probabilidad media $\overline{P}$.

Para obtener esa relación, se define la función

$$g(x) := x\log M + \mathcal{H}(x) \quad \text{siendo} \quad \mathcal{H}(x) := \begin{cases} -x\log(x) - (1-x)\log(1-x) & \text{si } 0 < x \leq 1 \\ 0 & \text{si } x = 0 \end{cases}. \tag{2.66}$$

Lema 2.6.2. Para todo $j_0 \in \{0,1,\dots,M\}$, y $p_0, p_1, \dots, p_M$ positivos tales que $\sum_{j=0}^{M} p_j = 1$. Entonces, se cumple que

$$g\left(\sum_{j\neq j_0} p_j\right) \geq -\sum_{j=0}^{M} p_j \log p_j,$$

siendo g la función definida en (2.66).

Demostración.

Es suficiente con probar el caso en el que $\sum_{j\neq j_0} p_j \neq 0$, ya que el otro caso es obvio ($g(0) = 0$).

Separando el sumando de j_0 del sumatorio y operando, se obtiene que:

$$\begin{aligned}
\sum_{j=0}^{M} p_j \log p_j &= p_{j_0} \log p_{j_0} + \sum_{j\neq j_0} p_j \log p_j \\
&= p_{j_0} \log p_{j_0} + \sum_{j\neq j_0} p_j \left[\log\left(\sum_{i\neq j_0} p_i\right) - \log\left(\sum_{i\neq j_0} p_i\right) + \log\left(p_j\right) \right] \\
&= p_{j_0} \log p_{j_0} + \sum_{j\neq j_0} p_j \left[\log\left(\sum_{i\neq j_0} p_i\right) + \log\left(\frac{p_j}{\sum_{i\neq j_0} p_i}\right) \right] \\
&= p_{j_0} \log p_{j_0} + \left(\sum_{j\neq j_0} p_j\right) \left(\log \sum_{j\neq j_0} p_j\right) + \sum_{i\neq j_0} p_j \log\left(\frac{p_j}{\sum_{i\neq j_0} p_i}\right) \\
&= -\mathcal{H}\left(\sum_{j\neq j_0} p_j\right) + \left(\sum_{j\neq j_0} p_j\right) \left(\sum_{j\neq j_0} \frac{p_j}{\sum_{i\neq j_0} p_i} \log \frac{p_j}{\sum_{i\neq j_0} p_i}\right) \\
&:= -\mathcal{H}\left(\sum_{j\neq j_0} p_j\right) + \left(\sum_{j\neq j_0} p_j\right) \left(\sum_{j\neq j_0} q_j \log q_j\right),
\end{aligned}$$

donde en la última igualdad se ha denotado $q_j = \dfrac{p_j}{\sum_{i\neq j_0} p_i}$.

Obsérvese que $\sum_{j\neq j_0} q_j = 1$. Supongamos que $q_j > 0$ para todo j (en caso contrario, se ignoran los términos nulos para el siguiente paso). Puesto que $-\log x$ es una función convexa para $x > 0$, se puede aplicar la desigualdad de Jensen obteniendo que

$$\sum_{j\neq j_0} q_j \log q_j = -\sum_{j\neq j_0} q_j \log(1/q_j) \geq -\log \sum_{j\neq j_0} q_j/q_j = -\log M.$$

Juntando a la igualdad anterior, se obtiene que

$$\begin{aligned}
\sum_{j=0}^{M} p_j \log p_j &\geq -\mathcal{H}\left(\sum_{j\neq j_0} p_j\right) + \left(\sum_{j\neq j_0} p_j\right)(-\log M) = -g\left(\sum_{j\neq j_0} p_j\right) \\
&\Longrightarrow g\left(\sum_{j\neq j_0} p_j\right) \geq -\sum_{j=0}^{M} p_j \log p_j.
\end{aligned}$$

□

Las proposiciones anteriores permiten obtener este resultado previo al Lema de Fano:

Lema 2.6.3. Sean $P_0, P_1, \ldots, P_M$ probabilidades en $(\mathcal{X}, \mathcal{A})$, con $M \geq 1$. Siendo

$$g(x) := x \log M + \mathcal{H}(x) \qquad y \qquad \mathcal{H}(x) := \begin{cases} -x\log(x) - (1-x)\log(1-x) & \text{si } 0 < x \leq 1 \\ 0 & \text{si } x = 0 \end{cases},$$

con g definida en $[0, 1]$, se cumple que

$$g(\overline{p}_{e,M}) \geq \log(M+1) - \frac{1}{M+1}\sum_{j=0}^{M} K(P_j, \overline{P}) \qquad y \qquad \overline{p}_{e,M} \leq \frac{M}{M+1}. \tag{2.67}$$

Demostración.

Por la Proposición 2.6.1,

$$\overline{p}_{e,M}(\Psi) = \mathbb{E}_{\overline{P}}\left[\sum_{j\neq j_0} p_j\right],$$

siendo

$$p_j = \frac{1}{M+1}\frac{dP_j}{d\overline{P}}.$$

Además, en esa proposición también se demostró que $\sum_{j=0}^{M} p_j = 1$, lo que implica que se puede aplicar el lema anterior a los valores $p_0, p_1, \ldots, p_M$. Eso, junto con aplicar la desigualdad de Jensen a $g(x)$ (que es cóncava en $[0,1]$), para todo test Ψ se obtiene

$$\begin{aligned} g\left(\overline{p}_{e,M}(\Psi)\right) &= g\left(\mathbb{E}_{\overline{P}}\left[\sum_{j=0}^{M} b_j\, p_j\right]\right) \geq \mathbb{E}_{\overline{P}}\left[g\left(\sum_{j\neq j_0} p_j\right)\right] \\ &\geq \mathbb{E}_{\overline{P}}\left[-\sum_{j=0}^{M} p_j \log p_j\right] = \mathbb{E}_{\overline{P}}\left[-\sum_{j=0}^{M} \frac{1}{M+1}\frac{dP_j}{d\overline{P}} \log\left(\frac{1}{M+1}\frac{dP_j}{d\overline{P}}\right)\right] \\ &= -\frac{1}{M+1}\mathbb{E}_{\overline{P}}\left[\sum_{j=0}^{M} \frac{dP_j}{d\overline{P}} \log \frac{dP_j}{d\overline{P}}\,(-\log(M+1))\right] \\ &= \log(M+1) - \frac{1}{M+1}K(P_j, \overline{P}). \end{aligned}$$

Al ser $\overline{p}_{e,M}$ un mínimo, existe una sucesión de tests $(\Psi_k)_{k=1}^{\infty}$ tales que $\overline{p}_{e,M}(\Psi) \longrightarrow \overline{p}_{e,M}$ cuando $k \to \infty$. La continuidad de g implica que:

$$g\left(\overline{p}_{e,M}\right) = \log(M+1) - \frac{1}{M+1}\sum_{j=0}^{M} K(P_j, \overline{P}).$$

Ahora solo falta probar que $\overline{p}_{e,M} \leq \dfrac{M}{M+1}$.

Considerando el test $\Psi^* \equiv 1$:

$$\overline{p}_{e,M} = \inf_{\Psi} \overline{p}_{e,M}(\Psi) \leq \overline{p}_{e,M}(\Psi^*) = \frac{1}{M+1}\sum_{j=0}^{M} P_j(j \neq 1) = \frac{M}{M+1}.$$

□

Este lema obtiene una acotación de $\overline{p}_{e,M}$, y de manera intrínseca se ha acotado $p_{e,M}$:

$$\begin{aligned} p_{e,M} = \inf_{\Psi} \max_{0\leq j\leq M} P_j(\Psi \neq j) &\leq \inf_{\Psi} \overline{p}_{e,M}(\Psi) = \overline{p}_{e,M} \\ &\geq g^{-1}\left(\log(M+1) - \frac{1}{M+1}K(P_j, \overline{P})\right), \end{aligned}$$

siendo g^{-1} la función inversa de g en $[0, M/(M+1)]$, que existe puesto que g es una función continua creciente en dicho intervalo. Además, cumple que $g(0) = 0$ y $g(\frac{M}{M+1}) = \log(M+1)$.

Esto sugiere obtener acotaciones inferiores de la probabilidad de error (y, por tanto, del riesgo minimax) a partir de la desigualdad anterior. Solo hace falta considerar unas hipótesis iniciales para las que las probabilidades $P_0, \ldots, P_M$ cumplan que

$$\log(M+1) - \frac{1}{M+1}\sum_{j=0}^{M} K(P_j, \overline{P}) > 0$$

para que se pueda considerar la inversa de g.

Una opción es escoger las probabilidades de manera que cumplan que

$$\frac{1}{M+1}\sum_{j=0}^{M} K(P_j, P_0) \leq \alpha \log M. \tag{2.68}$$

Basta con observar que

$$\frac{1}{M+1}\sum_{j=0}^{M} K(P_j, P_0) = \frac{1}{M+1}\sum_{j=0}^{M} K(P_j, \overline{P}) + K(\overline{P}, P_0),$$

y, por el Lema 2.6.3, se cumple entonces que

$$g(\overline{p}_{e,M}) \geq \log(M+1) - \sum_{j=0}^{M} K(P_j, P_0) \tag{2.69}$$

$$\Rightarrow \overline{p}_{e,M} \geq g^{-1}\left(\log(M+1) - \alpha \log M\right). \tag{2.70}$$

Aunque se haya conseguido una acotación de la probabilidad del riesgo, ésta no es muy útil, debido a tener que manejar $g^{-1}(x)$ para obtener el valor. El Lema de Fano pule ese resultado para obtener una cota mucho más manejable:

Lema 2.6.4 (Lema de Fano)**.** Sean $P_0, \ldots, P_M$ en $(\mathcal{X}, \mathcal{A})$, con $M \geq 2$, tales que cumplen que

$$\frac{1}{M+1}\sum_{j=0}^{M} K(P_j, P_0) \leq \alpha \log M$$

para $0 < \alpha < 1$.

Entonces, se obtiene la siguiente desigualdad:

$$p_{e,M} \geq \overline{p}_{e,M} \geq \frac{\log(M+1) - \log 2}{\log M} - \alpha.$$

Demostración.

Bajo esas condiciones, se cumple que:

$$g(\overline{p}_{e,M}) \geq \log(M+1) - \frac{1}{M+1}\sum_{j=0}^{M} K(P_j, P_0).$$

Puesto que $g(x) = x \log M + \mathcal{H}(x)$, y $\mathcal{H}(x) \leq 2$ para todo $x \leq 2$.

Por una parte, por el Lema 2.6.3 se tiene que

$$g(\overline{p}_{e,M}) \underset{(2.67)}{\geq} \log(M+1) - \frac{1}{M+1}\sum_{j=0}^{M} K(P_j, P_0).$$

Por la otra, por la definición de g:

$$g(\overline{p}_{e,M}) = \overline{p}_{e,M} \log(M) + \mathcal{H}(\overline{p}_{e,M})$$

Juntando ambas se obtiene:

$$\overline{p}_{e,M} \log(M) + \mathcal{H}(\overline{p}_{e,M}) \geq \log(M+1) - \frac{1}{M+1}\sum_{j=0}^{M} K(P_j, P_0).$$

Despejando $\overline{p}_{e,M}$ de la desigualdad:

$$\begin{aligned}\overline{p}_{e,M} &\geq \frac{\log(M+1) - \mathcal{H}(\overline{p}_{e,M})}{\log M} - \frac{1}{M+1}\frac{\sum_{j=0}^{M} K(P_j, P_0)}{\log M} \\ &\geq \frac{\log(M+1) - \log 2}{\log M} - \alpha.\end{aligned}$$

□

Obsérvese que no se tiene en cuenta el caso $M = 1$. Esto es debido a que, para $M = 1$, se obtiene:

$$p_{e,1} \geq \overline{p}_{e,1} \geq \mathcal{H}^{-1}(\log(2) - \alpha/2).$$

Esta acotación no resulta interesante, puesto que en 2.54, ya se obtuvo una mejor cota bajo las mismas condiciones:

$$p_{e,1} \geq \text{máx}\left(\frac{1}{4}e^{-\alpha}, \frac{1-\sqrt{\alpha/2}}{2}\right).$$

Sin embargo, para el caso en el que se consideran varias hipótesis ($M \geq 2$), la acotación obtenida es mejor que la obtenida en el Corolario 2.5.8, puesto que

$$\frac{\log(M+1) - \log 2}{\log M} - \alpha \geq \frac{\sqrt{M}}{1+\sqrt{M}}\left(1 - 2\alpha - \sqrt{\frac{2\alpha}{\log 2}}\right), \tag{2.71}$$

cuando ambas acotaciones se han conseguido bajo las mismas hipótesis.

Es más, en el siguiente capítulo, donde se estudie la optimalidad de los estimadores lineales locales, se utilizará el Lema de Fano, pero no para encontrar la tasa de convergencia óptima (el Corolario 2.5.8 servirá para ese propósito), sino para poder extender los resultados a un modelo de regresión más general. Esto será posible debido a que el Lema de Fano obtiene la acotación a partir de la esperanza respecto a la probabilidad media (Sección 3.4).

Capítulo 3

Estudio de los estimadores locales polinomiales

En esta parte se estudiará cuán buen estimador es el estimador local polinomial, definido en la Definición 2.2.4, siendo el objetivo final demostrar que es un estimador óptimo en tasa.

Por una parte, se querrá obtener, bajo una clase suficientemente regular, cotas superiores de su riesgo, viendo cómo la acotación conseguida es mayor cuanto mayor es la suavidad considerada (el grado l del estimador). Esto quedará más reflejado en la sección 3.2, donde se se comentarán varios ejemplos de estimación comparando el estimador Nadaraya-Watson ($LP(0)$) con el estimador local lineal ($LP(1)$).

Por otra parte, para demostrar que el estimador $LP(l)$ sea óptimo en tasa, se tendrá que acotar inferiormente el riesgo minimax, y que la tasa de acotación conseguida sea la misma que con la que sea ha conseguido acotar el riesgo del estimador $LP(l)$. Para obtener esta acotación, se usará el Teorema Principal de las Cotas Inferiores (Teorema 2.5.7), siendo la tarea entonces de encontrar la funciones parámetro que cumplan con las condiciones del teorema.

Estos resultados serán obtenidos bajo el modelo comentado en al final de la Sección 2.1.2, es decir:

- Los atributos del conjunto de entrenamiento $(x_i)_{i=1}^n$ son unidimensionales y deterministas; en concreto, se considerará que $x_i = i/n$.
- Existen $\epsilon_1, \dots \epsilon_n$ variables aleatorias normales $\mathcal{N}(0, \sigma^2)$ i.i.d tales que

$$Y_i = f(x_i) + \epsilon_i. \tag{3.1}$$

- Los resultados se conseguirán para las tres funciones de pérdidas consideradas: para la pérdida puntual (sirviendo los resultados para esta como apoyo para el resto), la pérdida cuadrática, y la pérdida del supremo.

3.1. Riesgo de los estimadores locales polinomiales

Para acotar el riesgo del estimador $LP(l)$, se necesitará poder delimitar el valor de los términos del polinomio local: esto sugiere considerar una función l-veces derivable y con derivada $l+1$ continua (para que su polinomio local sea su desarrollo de Taylor de orden l) y que tanto la función como de sus derivadas estén acotadas. La clase de Hölder $\Sigma(L, \beta)$, con $l = \lfloor \beta \rfloor$ cumple con esos requisitos, haciendo de ella una excelente opción sobre clase no paramétrica sobre la que trabajar.

Definición 3.1.1. Sea $L \in \mathbb{R}$ y un intervalo $I \subset \mathbb{R}$. Sea $\beta > 0$ y $l = \lfloor \beta \rfloor$. La clase de Hölder con parámetros L, $\beta > 0$, denotada por $\Sigma(L, \beta)$, viene dada por el conjunto de funciones que son l veces diferenciables, y su l-ésima derivada cumple que, para todo $x, y \in I$,

$$|f^{(l)}(x) - f^{(l)}(y)| \leq L\|x - y\|^{\beta - l}.$$

Hay que recordar que el estimador local polinomial no siempre es fácil de manejar. En la Proposición 2.2.4 se indica que, mientras que la matriz $\mathcal{B}_{nx}$

$$\mathcal{B}_{nx} = \frac{1}{nh}\sum_{i=1}^{n} \boldsymbol{U}\left(\frac{x-x_i}{h}\right)\boldsymbol{U}^T\left(\frac{x-x_i}{h}\right)K\left(\frac{x-x_i}{h}\right)$$

sea definida positiva, el estimador $LP(l)$ será un estimador lineal. Ese será el caso que en que se quiere trabajar, puesto que es cuando se sabe calcular el valor del estimador. Cuando esto ocurra, siendo $\hat{f}_n$ el estimador $LP(l)$, el estimador se puede escribir como

$$f(x) = \sum_{i=1}^{n} Y_i W_{ni}(x),$$

siendo

$$W_{ni}(x) = \frac{1}{nh}\mathcal{B}_{nx}^{-1}\boldsymbol{U}\left(\frac{x-x_i}{h}\right)K\left(\frac{x-x_i}{h}\right). \tag{3.2}$$

Para obtener una acotación de $LP(l)$ para esa clase, se tendrá que poder acotar sus pesos $W_{ni}(x)$. Es por ello que será necesario considerar trabajar bajo ciertas condiciones sobre los $W_{ni}(x)$ para garantizar poder trabajar con ellos.

Condiciones 2 (Suposiciones **LP**)**.** Sean W_{ni} descritos en (3.2). Las hipótesis LP sobre la matriz $\mathcal{B}_{nx}$, el conjunto de entrenamiento $(X_i)_{i=1}^n$, y el núcleo K son

LP1 Existe un número real $\lambda_0 > 0$, y $n_0 \in \mathbb{N}$ tal que

$$\lambda_{\text{mín}}(\mathcal{B}_{nx}^{-1}) \geq \lambda_0 \tag{3.3}$$

para todo $n > n_0$. Esto implica que $\mathcal{B}_{nx}$ sea definida positiva, y que por tanto se pueda considerar que $LP(l)$ es un estimador lineal por la Proposición 2.2.4.

Además, como es una matriz simétrica, también se obtiene que

$$||\mathcal{B}_{nx}v|| \leq ||v||/\lambda_0 \tag{3.4}$$

Para todo $n \geq n_0$, $x \in [0,1]$ y todo $v \in \mathbb{R}^{l+1}$.

LP2 Existe un número real $a_0 > 0$ tal que, para cualquier intervalo A de $[0,1]$, y todo $n \in \mathbb{N}$

$$\frac{1}{n}\sum_{i=1}^{n} I(x_i \in A) \leq a_0 \,\text{máx}\,(m_{Leb}(A), 1/n). \tag{3.5}$$

Esto significa que los atributos del conjunto de entrenamiento sean suficiente variados dentro del intervalo donde se trabaja.

Como en el modelo los atributos del conjunto de entrenamiento son deterministas de la forma $x_i = i/n$, cumplen con esta suposición para $a_0 = 2$, ya que:

- Si $m_{Leb}(A) < 1/n$, por como son los atributos, como mucho habría un solo X_i perteneciente a A, lo que implica que

$$\frac{1}{n}\sum_{i=1}^{n} I(x_i \in A) \leq \frac{1}{n} \leq \frac{2}{n}.$$

- De la misma manera, si $\frac{j}{n} \leq m_{Leb}(A) < \frac{j}{n}$ para cierto j, como mucho j atributos pertenecerán a A:

$$\frac{1}{n}\sum_{i=1}^{n} I(x_i \in A) \leq \frac{j}{n} \leq \frac{2j}{n} \leq 2\, m_{Leb}(A).$$

Es decir, la condición del modelo C1 implica que se cumpla la condición LP2.

Un caso en el que no se cumpla esta suposición es si se trabajase con que todos atributos del conjunto de entrenamiento $x_i = 1/n$. En este caso, sería imposible obtener una constante a_0 que satisfaga el intervalo $A = [0, 1/n]$, ya que debería de depender del tamaño del conjunto de entrenamiento n.

LP3 El núcleo K elegido tiene soporte compacto en $[-1, 1]$, y existe una constante $K_{\text{máx}} < \infty$ tal que

$$|K(u)| \leq K_{\text{máx}} \tag{3.6}$$

Para todo $u \in \mathbb{R}$.

Observación 3.1.1. La hipótesis LP3 restringe a tener que utilizar únicamente núcleos con soporte compacto, así que hay núcleos, como el núcleo gausiano, que quedan fuera al considerar esta hipótesis. Sin embargo, es una hipótesis bastante útil, puesto que permite reducir las muestras del conjunto de entrenamiento que se tienen que considerar a sólo aquellas que cumplan que $|x - X_i| \leq h$.

Las condiciones LP implican las siguientes acotaciones para los pesos W_{ni}:

Teorema 3.1.1. Sea $((x_i, Y_i))_{i=1}^n$ el conjunto de entrenamiento, y $x \in \mathbb{R}$. Supongamos que se cumplen las hipótesis LP. Entonces, las funciones peso W_{ni} cumplen que

1. $\sup_{i,x} |W_{ni}(x)| \leq \dfrac{C_1}{nh}$, siendo C_1 la constante dada por $C_1 = \dfrac{2K_{max}}{\lambda_0}$.
2. $\sum_{i=1}^n |W_{ni}(x)| \leq C_2$ siendo C_2 la constante dada por $C_2 = \dfrac{4K_{max}a_0}{\lambda_0}$.
3. $W_{ni}(x) = 0$ si $|x - x_i| > h$.

Demostración.

1. Debido a que $||\mathcal{B}_{nx}v|| \leq ||v||/\lambda_0$ (por LP1), usando LP3 y $\|\boldsymbol{U}(0)\| = 1$:

$$\begin{aligned}
|W_{ni}| &\leq \frac{1}{nh}\left\|\mathcal{B}_{nx}^{-1}\boldsymbol{U}\left(\frac{x-x_i}{h}\right)K\left(\frac{x-x_i}{h}\right)\right\| \\
&\underset{LP1}{\leq} \frac{1}{nh\lambda_0}\left\|\boldsymbol{U}\left(\frac{x-x_i}{h}\right)K\left(\frac{x-x_i}{h}\right)\right\| \\
&\leq \frac{K_{\text{máx}}}{nh\lambda_0}\left\|\boldsymbol{U}\left(\frac{x-x_i}{h}\right)I\left(\left|\frac{x-x_i}{h}\right| \leq 1\right)\right\| \\
&\underset{LP3}{\leq} \frac{K_{\text{máx}}}{n\,h\,\lambda_0}\sqrt{1+1+\frac{1}{(2!)^2}+\ldots+\frac{1}{(l!)^2}}| \leq \frac{2K_{\text{máx}}}{n\,h\,\lambda_0}.
\end{aligned}$$

2. Mediante el mismo razonamiento, usando ahora LP2:

$$\begin{aligned}
\sum_{n=1}^n |W_{ni}| &\leq \frac{K_{\text{máx}}}{nh\lambda_0}\sum_{n=1}^n\left\|\boldsymbol{U}\left(\frac{x-x_i}{h}\right)I\left(\left|\frac{x-x_i}{h}\right| \leq 1\right)\right\| \\
&\leq \frac{2K_{\text{máx}}}{nh\lambda_0}\sum_{n=1}^n I(x-h \leq x_i \leq x+h) \\
&\underset{LP2}{\leq} \frac{2K_{\text{máx}}a_0}{nh\lambda_0}\,\text{máx}\left(2, \frac{1}{nh}\right) \leq \frac{4K_{\text{máx}}a_0}{\lambda_0}.
\end{aligned}$$

3. Como $|x - x_i| > h$, entonces $1 \geq \dfrac{x-x_i}{h}$. Como por LP3 el soporte de la función núcleo está contenido en $[-1, 1]$, en esos casos $K\left(\dfrac{x-x_i}{h}\right) = 0$, así que $W_{ni}(x) = 0$.

□

Primero se trabajará para obtener el riesgo para la pérdida puntual y cuadrática, para luego extender los resultados a la pérdida del supremo.

3.1.1. Riesgo para la pérdida cuadrática y distancia puntual.

Sea x_0 un punto fijo. Como se comentó en la Sección 2.1.2, el riesgo puntual del estimador $\hat{f}_n$ se puede descomponer en la parte del sesgo y la parte de la varianza:

$$R(\hat{f}_n, x_0) = \left(\mathbb{E}_f(\hat{f}_n(x_0)) - f(x_0)\right)^2 + \mathbb{E}_f((\hat{f}_n(x_0))^2) - \left(\mathbb{E}_f(\hat{f}_n(x_0))\right)^2 = sesgo^2_{x_0} + \sigma^2(x_0).$$

Esto sugiere que a la hora de delimitar el riesgo de los estimadores LP se opte por intentar acotar tanto el sesgo como la varianza por separado.

Antes de hallar una cota del riesgo para la clase de Hölder, se va a estudiar qué sucede en la clase de los polinomios de grado l.

Lema 3.1.2. Sea $x \in \mathbb{R}$ tal que $\mathcal{B}_{nx} > 0$, y Q un polinomio de grado l. Entonces, si W_{ni} son los pesos del estimador $LP(l)$ de Q, se cumple que:

$$\sum_{i=1}^{n} Q(X_i) W_{ni}(x) = Q(x)$$

para cualquier muestra $(X_1, \ldots, X_n)$. En particular,

$$\sum_{i=1}^{n} W_{ni}(x) = 1 \qquad \text{y} \qquad \sum_{i=1}^{n} (x - X_i)^k \, W_{ni}(x) = 0 \qquad \text{para } k = 1, \ldots, l. \tag{3.7}$$

Demostración.

Como Q es un polinomio de grado menor o igual a l, se cumple la siguiente igualdad

$$\begin{aligned} Q(X_i) &= Q(x) + Q'(x)(x - X_i) + \ldots + \frac{Q^{(l)(x)}}{l!}(x - X_i)^l \\ &= \mathbf{q}^T(x) \cdot \boldsymbol{U}\left(\frac{x - X_i}{h}\right) \end{aligned}$$

siendo $\mathbf{q} = (Q(x), Q'(x)h, \ldots, Q^{(l)}h^l)^T$.

El estimador $LP(l)$ del polinomio Q será, por tanto,

$$\begin{aligned} \hat{\theta}_n(x) &= \arg \min_{\boldsymbol{\theta} \in \mathbb{R}^{l+1}} \sum_{i=1}^{n} \left(Q(X_i) - \boldsymbol{\theta}^T \boldsymbol{U}\left(\frac{x - X_i}{h}\right)^2\right) K\left(\frac{x - X_i}{h}\right) \\ &= \arg \min_{\boldsymbol{\theta} \in \mathbb{R}^{l+1}} \sum_{i=1}^{n} \left(\mathbf{q}(x) - \boldsymbol{\theta}\right)^T \mathcal{B}_{nx} \left(\mathbf{q}(x) - \theta\right). \end{aligned}$$

Como $\mathcal{B}_{nx} > 0$, el $\boldsymbol{\theta}$ mínimo será aquel que cumpla que $\boldsymbol{\theta}_n(x) = \mathbf{q}(x)$, por lo que $\hat{f}_n(x) = Q(x)$.

Por tanto, como $Q_k(y) := (y - x)^k$ es un polinomio de grado l, tenemos que

$$\sum_{i=1}^{n} (x - X_i)^k W_{ni}(x) = \sum_{i=1}^{n} Q_k(X_i) W_{ni}(x) = Q_k(x) = (x - x)^k = 0 \tag{3.8}$$

□

Es decir, en el caso en el que la función a estimar es un polinomio de grado l, la regresión no paramétrica local polinomial obtendría el valor exacto.

El hecho de que se pueda acotar los pesos del estimador bajo las condiciones LP por el Teorema 3.1.1, hará posible se pueda obtener una cota superior del riesgo en $\Sigma(L,\beta)$. basándose en el lema anterior.

Proposición 3.1.3. Sea $f \in \Sigma(\beta, L)$ en $[0,1]$, y consideramos $\hat{f}_n$ el estimador $LP(l)$ de f, con $l = \lfloor \beta \rfloor$ tal que:

1. Se cumplen LP1, LP2 y LP3.
2. Los atributos del modelo $x_1, \ldots, x_n$ son deterministas (condición C1).

Entonces, para todo $x_0 \in [0,1]$, $n \geq n_0$ y $h \geq \dfrac{1}{2n}$, se obtiene

$$|b(x_0)| \leq q_1 h^{\beta} \qquad \sigma^2(x_0) \leq \frac{q_2}{nh} \tag{3.9}$$

Es decir, el error cuadrático medio de $\hat{f}_n$ está acotado por:

$$R(\hat{f}_n, x_0) = MSE \leq q_1^2 h^{2\beta} + \frac{q_2}{nh}. \tag{3.10}$$

Demostración.

Como $f \in \Sigma(\beta, L)$, es l veces derivable se tiene que

$$f(x) = f(x_0) + f'(x_0)(x - x_0) + \ldots + \frac{f^{(l)}(x_0 + \tau(x - x_0))}{l!}(x - x_0)^l$$

con $0 \leq \tau \leq 1$. Luego, podemos tratar a f como polinomio de grado l. Esto implica que podemos aplicar la proposición anterior, obteniendo:

$$\begin{aligned}
b(x_0) &= E_f[\hat{f}_n(x_0)] - f(x_0) = \sum_{i=1}^{n} E_f[Y_i] - f(x_0) \\
&= \sum_{i=1}^{n} f(X_i) W_{ni}(x_0) - f(x_0) \\
&= \sum_{i=1}^{n} f(X_i) W_{ni}(x_0) - f(x_0) \sum_{i=1}^{n} W_{ni} \\
&= \sum_{i=1}^{n} \left(f(X_i) - f(x_0)\right) W_{ni}(x_0) \\
&= \sum_{i=1}^{n} \frac{f^{(l)}(x_0 + \tau_i(x_0 - x_i)) - f(x_0)}{l!} (x_0 - x_i)^l W_{ni}(x_0) \qquad \left(\sum_{i=1}^{n} (x - x_i)^k W_{ni}(x) = 0\right) \\
&= \sum_{i=1}^{n} \frac{f^{(l)}(x_0 + \tau_i(x_0 - x_i)) - f^{(l)}(x_0)}{l!} |x_0 - x_i|^l W_{ni}(x_0) \\
&\leq \sum_{i=1}^{n} \frac{L\left[(x_0 + \tau_i(x_0 - x_i)) - x_0\right]^{\beta - l}}{l!} |x_0 - x_i|^l W_{ni}(x_0) \\
&\leq= \sum_{i=1}^{n} \frac{L|x_0 - x_i|^{\beta}}{l!} W_{ni}(x_0)
\end{aligned}$$

con $0 \leq \tau_i \leq 1$, donde en la desigualdad final se ha utilizado que $f \in \Sigma(\beta, L)$.

Ahora, al cumplirse las condiciones LP, obtenemos la siguiente acotación de $b(x_0)$:

$$
\begin{aligned}
|b(x_0)| &\leq \sum_{i=1}^{n} \frac{L|x_0 - x_i|^\beta}{l!} W_{ni}(x_0) \\
&= L \sum_{i=1}^{n} \frac{|x_0 - x_i|^\beta}{l!} W_{ni}(x_0) I(|x_0 - x_i| \leq h) \\
&\underset{LP2}{\leq} \sum_{i=1}^{n} \frac{h^\beta}{l!} W_{ni}(x_0) \leq \frac{LC_2}{l!} h^\beta = q_1 h^\beta,
\end{aligned}
$$

donde se han aplicado LP2 y el lema 3.1.1.

Una vez acotado el sesgo, falta acotar $\sigma(x_0)$.

Utilizando nuevamente el Lema 3.1.1:

$$
\begin{aligned}
\sigma^2(x_0) = E\left[\left(\sum_{i=1}^{n} \epsilon_i W_{ni}(x_0)\right)^2\right] &= \sum_{i=1}^{n} \left(W_{ni}(x_0)\right)^2 E(\epsilon_i^2) \\
&\leq \sigma^2_{\text{máx}} \sup_{i,x} |W_{ni}(x)| \sum_{i=1}^{n} |W_{ni}(x_0)| \\
&\underset{3.1.1}{\leq} \frac{\sigma^2_{\text{máx}} C_1 C_2}{nh} = \frac{q_2}{nh}
\end{aligned}
$$

Por tanto, al haber conseguido acotar tanto el sesgo como la varianza, se ha acotado el error cuadrático medio:

$$
MSE \leq q_1^2 h^{2\beta} + \frac{q_2}{nh}. \tag{3.11}
$$

□

Esa última desigualdad nos da una **cota superior del riesgo** obtenido a través estimadores locales polinomiales par ala pérdida puntual. Esta acotación depende del ancho de banda h_n, lo implica que se puede optimizar escogiendo un ancho de banda h_n determinado.

Considerando $\delta(h) = q_1^2 h^{2\beta} + \frac{q_2}{nh}$, en la proposición anterior se ha hallado que $MSE \leq \delta(h_n)$, para todo valor de h. Esa cota superior depende del ancho de banda escogido, por lo que se puede obtener un menor riesgo escogiendo el ancho de banda que optimiza $\delta(h_n)$.

Calculando el mínimo de δ de la manera habitual (derivando y despejando), se obtiene que el ancho de banda que minimiza el riesgo de los estimadores locales polinomiales es:

$$
\hat{h}_n = \left(\frac{q_2}{2\beta q_1^2}\right)^{\frac{1}{2\beta+1}} n^{-\frac{1}{2\beta+1}}. \tag{3.12}
$$

En estas gráficas se puede observar como el escoger un ancho de banda óptimo o no a la hora de calcular el estimador influye en su su rendimiento:

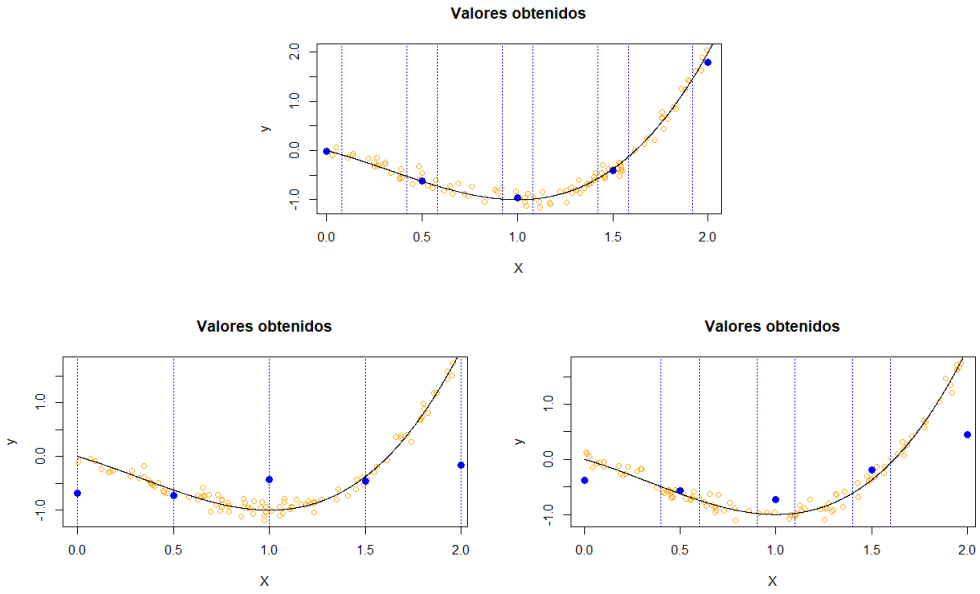


Figura 3.1: Comparación de las estimaciones obtenidas escogiendo distintos anchos de banda, a partir de una muestra generada de la misma forma que la de la Figura 2.2.
Arriba: $h = 0{,}8$ (el más cercano al ancho de banda óptimo).
Izquierda: $h = 0{,}6$
Derecha: $h = 1$

Escogiendo el ancho de banda óptimo, se obtiene por tanto el siguiente corolario, que da una acotación superior al riesgo tanto para la pérdida puntual como para la pérdida cuadrática.

Corolario 3.1.4. Sea $f \in \Sigma(\beta, L)$ en $[0,1]$, y consideramos $\hat{f}_n$ el estimador $LP(l)$ de f, con $l = \lfloor \beta \rfloor$ tal que:

1. Se cumplen las condiciones C del modelo.
2. Se dan las condiciones LP.

Si el ancho de banda escogido es de la forma $h = h_n = \alpha n^{-\frac{1}{2\beta+1}}$, con α siendo una constante positiva, existe $C \geq 0$ tal que

$$\limsup_{n\to\infty} \sup_{f\in\Sigma(\beta,L)} \sup_{x_0\in[0,1]} \mathbb{E}_f \left[\psi_n^{-2} |\hat{f}_n(x_0) - f(x_0)|^2\right] \leq C < \infty, \tag{3.13}$$

siendo $\psi_n = n^{-\frac{\beta}{2\beta+1}}$ la tasa de convergencia.

Esto además implica que se cumpla:

$$\limsup_{n\to\infty} \sup_{f\in\Sigma(\beta,\mathcal{L})} \mathbb{E}_f \left[\psi_n^{-2} \|\hat{f}_n - f\|_2^2\right] \leq C < \infty \tag{3.14}$$

Demostración.

La acotación en (3.13) se obtiene a partir de la proposición anterior, a partir de la desigualdad (3.11) para el ancho de banda óptimo. Partiendo de ella, se tiene que, para n suficientemente grande:

$$\sup_{f\in\Sigma(\beta,L)} \sup_{x_0\in[0,1]} \mathbb{E}_f \left[\psi_n^{-2} |\hat{f}_n(x_0) - f(x_0)|^2\right] \leq C' = C + \epsilon < \infty,$$

luego

$$\mathbb{E}_f \left[\psi_n^{-2} |\hat{f}_n(x) - f(x)|^2\right] \leq C'$$
$$\mathbb{E}_f \left[|\hat{f}_n(x) - f(x)|^2\right] \leq \frac{C'}{\psi_n^{-2}}$$

para todo $f \in \Sigma(\beta, L)$ y $x \in [0,1]$.

$$\Longrightarrow R(\hat{f}) = \mathbb{E}_f \left\| \hat{f} - f \right\|_2^2 \leq \sup_{\boldsymbol{x} \in [0,1]} \mathbb{E}_f \left(\hat{f}(\boldsymbol{x}) - f(\boldsymbol{x}) \right) \tag{3.15}$$

Además, se cumple que, para todo x_0

$$\begin{aligned} \mathbb{E}_f \left\| \hat{f} - f \right\|_2^2 &\leq \mathbb{E}_f \left(\int \hat{f}(\boldsymbol{x}) - f(\boldsymbol{x})\, d\boldsymbol{x} \right) \leq \mathbb{E}_f \left(\int \sup_{\boldsymbol{x} \in [0,1]} \hat{f}(x) - f(\boldsymbol{x})\, d\boldsymbol{x} \right) \\ &= \mathbb{E}_f \left(\sup_{\boldsymbol{x} \in [0,1]} \hat{f}(\boldsymbol{x}) - f(\boldsymbol{x}) \right) \leq \mathbb{E}_f \left(\hat{f}(\boldsymbol{x}_0) - f(\boldsymbol{x}_0) \right). \end{aligned}$$

Por la desigualdad (3.15), se tiene que

$$\mathbb{E}_f \left\| \hat{f} - f \right\|_2^2 \leq \sup_{\boldsymbol{x} \in [0,1]} \mathbb{E}_f \left(\hat{f}(\boldsymbol{x}) - f(\boldsymbol{x}) \right). \tag{3.16}$$

Esto implica que se puede extrapolar la cota del riesgo puntual al riesgo cuadrático:

$$\limsup_{n \to \infty} \sup_{f \in \Sigma(\beta, \mathcal{L})} \mathbb{E}_f \left[\psi_n^{-2} \| \hat{f}_n - f \|_2^2 \right] \leq C' < \infty.$$

□

Recordemos que para llegar a esta acotación hay que comprobar que se cumplan las hipótesis LP. A la hora de aplicarlo a nuestro modelo de regresión, se tiene que ver que cuáles de esas condiciones esas condiciones están cubiertas por las propias condiciones del modelo, y cuáles en cambio, se necesitan considerar como condiciones adicionales.

El mayor problema se encuentra al tratar con la condición LP1 (ya que la condición C1 implica que se cumple LP2, y la condición LP3 simplemente restringe la función núcleo con la que trabajar). La suposición LP1 depende de la matriz $\mathcal{B}_{nx}$ ((2.23)). Dicha matriz depende tanto de n como de x, haciendo que en un principio se tenga que comprobar que se cumpla LP1 para cada valor de x (o n) distinto.

Por la convergencia de las sumas parciales a la integral, podemos considerar que para n suficientemente grande, $\mathcal{B}_{nx}$ se acerca a la matriz

$$\mathcal{B} = \int \boldsymbol{U}(u)\boldsymbol{U}^T(u) K(u)\, du, \tag{3.17}$$

que es independientes de n y x, por lo que es más cómoda de utilizar.

Lema 3.1.5. Sea un núcleo $K : \mathbb{R} \to [0, \infty)$ con medida de Lebesgue positiva (puesto que K ees una función de desidad, implica que $m_{Leb}(\{u : K(u) > 0\}) > 0$). Entonces, la matriz $\mathcal{B}$ es definida positiva.

Demostración.

Es suficiente con probar que, para todo $v \in \mathbb{R}^{l+1}$ no nulo se cumpla que $v^T \mathcal{B} v > 0$, pero eso se cumple al ser $\mathcal{B}$ definida positiva,

$$\int \left(v^T \boldsymbol{U}(u) \right)^2 K(u)\, du \geq 0.$$

Si existiese v no nulo tal que $v^T \mathcal{B} v = 0$, por el Teorema de la anulación de la integral, tenemos que $v^T \boldsymbol{U}(u) = 0$ para casi todo $u \in \{u : K(u) > 0\}$. Sin embargo, puesto que sabemos que $f(u) = v^t \boldsymbol{U}(u)$ es un polinomio de grado $\leq l$, $v^T \boldsymbol{U}(u) = 0$ sólo un número finito de veces, lo que lleva a un absurdo con lo dicho anteriormente. □

Utilizando este lema, podemos dar unas condiciones habituales para las cuales la suposición LP1 se da, sin que dependa de $\mathcal{B}_{nx}$ (y, por tanto, de los valores de n y x):

Lema 3.1.6. Supongamos que existe un $K_{\text{mín}} > 0$ y $\Delta > 0$ tal que

$$K(u) \geq K_{\text{mín}}\, I(|u| \leq \Delta) \quad \forall u \in \mathbb{R},$$

y que los atributos del conjunto de entrenamiento sean $x_i = i/n$ para $i = 1, \ldots, n$. Sea $(h_n)_{n=1}^{\infty} = \mathbf{h}$ una sucesión que satisface:

$$h_n \to 0 \quad y \quad nh_n \to \infty, \tag{3.18}$$

para $n \to \infty$.

Entonces, la suposición LP1 se cumple.

Demostración.

Veamos que

$$\inf_{\|\mathbf{v}\|=1} \mathbf{v}^T \mathcal{B}_{nx} \mathbf{v} \geq \lambda_{\text{mín}}$$

para n suficientemente grande, lo que implicará que se cumple la hipótesis LP1.

Por hipótesis, se sabe que (denotando $z_i = \frac{x - x_i}{h}$):

$$\begin{aligned}\mathbf{v}^T \mathcal{B}_{nx} \mathbf{v} &= \mathbf{v}^T \left[\sum_{i=1}^{n} \boldsymbol{U}\left(\frac{x - x_i}{h}\right) \boldsymbol{U}^T\left(\frac{x - x_i}{h}\right) K\left(\frac{x - x_i}{h}\right) \right] \mathbf{v} \\ &= \frac{1}{nh} \sum_{i=1}^{n} (\mathbf{v}^T \boldsymbol{U}(z_i))^2 K(z_i) \geq \frac{K_{\text{mín}}}{nh} \sum_{i=1}^{n} (\mathbf{v}^T \boldsymbol{U}(z_i))^2 I(|z_i| \leq \Delta).\end{aligned}$$

Observemos que, puesto que $x_i = \frac{i}{n}$:

$$z_1 = \frac{1}{nh} - \frac{x}{h} \leq \frac{1}{nh}, \quad z_i - z_{i-1} = \tfrac{1}{nh}, \quad z_n = \frac{1 - x}{h} \geq 0.$$

- Si $x < 1 - \Delta h$, entonces $z_n > \frac{\Delta h}{h} = \Delta$, lo que implica que $z_1, \ldots, z_n$ es una red de puntos que recubre $[0, \Delta]$ con una distancia entre ellos de $1/nh$. Como $nh \to \infty$, la suma de Riemann converge a la integral:

$$\begin{aligned}\frac{1}{nh} \sum_{i=1}^{n} (\mathbf{v}^T \boldsymbol{U}(z_i))^2 I(|z_i| \leq \Delta) &\geq \frac{1}{nh} \sum_{i=1}^{n} (\mathbf{v}^T \boldsymbol{U}(z_i))^2 I(0 \leq z_i \leq \Delta) \\ &\xrightarrow{n \to \infty} \int_0^{\Delta} (\mathbf{v}^T \boldsymbol{U}(z))^2 dz\end{aligned}$$

- Si $x \geq 1 - \Delta h$, $z_1 < -\Delta$ para n suficientemente grande. Por el mismo razonamiento que en el caso anterior, pero considerando valores negativos:

$$\begin{aligned}\frac{1}{nh} \sum_{i=1}^{n} (\mathbf{v}^T \boldsymbol{U}(z_i))^2 I(|z_i| \leq \Delta) &\geq \frac{1}{nh} \sum_{i=1}^{n} (\mathbf{v}^T \boldsymbol{U}(z_i))^2 I(-\Delta \leq z_i \leq 0) \\ &\xrightarrow{n \to \infty} \int_{-\Delta}^{0} (\mathbf{v}^T \boldsymbol{U}(z))^2 dz.\end{aligned}$$

Ambas convergencias son uniformes en $\{\|\mathbf{v}\| = 1\}$, puesto que $\boldsymbol{U}(u) = (1, u, u^2, \ldots, u^l/l!)^T$, así que $(\mathbf{v}^T \boldsymbol{U}(z))^2$, es un polinomio de grado l.

Por el lema anterior, esto implica que, si $(h_n)_{n=1}^{\infty}$ cumple las convergencias de 3.18,

$$\inf_{\|\mathbf{v}\|=1} \mathbf{v}^T \mathcal{B}_{nx} \mathbf{v} \geq \frac{K_{\text{mín}}}{2} \text{mín} \left\{ \inf_{\|\mathbf{v}\|=1} \int_0^{\Delta} (\mathbf{v}^T \boldsymbol{U}(z))^2 dz, \inf_{\|\mathbf{v}\|=1} \int_{-\Delta}^{0} (\mathbf{v}^T \boldsymbol{U}(z))^2 dz \right\}$$

para n suficientemente grande. Aplicando el lema anterior, tenemos que $\mathcal{B}$ será una matriz definida positiva, debido a que, considerando $K_1(u) = I(0 \leq u \leq \Delta)$ y $K_2(u) = I(-\Delta \leq u \leq 0)$, ambas lo son. Es decir, la hipótesis LP1 se cumple.

□

El ancho de banda óptimo obtenido en 3.12 es $h_n = \left(\frac{q_2}{2\beta q_1^2}\right)^{\frac{1}{2\beta+1}} n^{-\frac{1}{2\beta+1}}$, que cumple con las hipótesis del lema anterior. Esto implica que escogiendo el ancho de banda óptimo se dan las hipótesis LP (y que, por tanto, se cumple el resultado 3.1.4) para nuestro modelo, considerando un ancho de banda y un núcleo determinados:

Corolario 3.1.7. Sea una función $f \in \Sigma(\beta, L)$ en $[0,1]$, con $\beta > 0$, y $L > 0$. Sea $\hat{f}_n$ el estimador $LP(l)$ de f, siendo $l = \lfloor \beta \rfloor$. Supongamos además que se cumple:

1. Se trabaja bajo las condiciones C del modelo; es decir, los atributos son $x_i = \dfrac{i}{n}$ y las variables aleatorias ϵ_i[1] son independientes entre sí, y satisfacen, para todo $i = 1, \ldots, n$:

$$\mathbb{E}(\epsilon_i) = 0 \qquad\qquad E(\epsilon_i^2) \leq \sigma_{max}^2 < \infty$$

2. Existen $K_{min} > 0$, $\Delta > 0$ y $K_{max} < \infty$ constantes tales que:

$$K_{min} I(|u| \leq \Delta) \leq K(u) \leq K_{max} I(|u| \leq 1) \qquad\qquad \forall u \in \mathbb{R}$$

3. El ancho de banda escogido es de la forma $h_n = \alpha n^{-\frac{1}{2\beta+1}}$, siendo $\alpha > 0$.

Entonces, el estimador $\hat{f}_n$ satisface el Corolario 3.1.4 . En concreto, el riesgo de $\hat{f}_n$ está acotado para la perdida cuadrática y la puntual:

$$\limsup_{n\to\infty} \sup_{f \in \Sigma(\beta, \mathcal{L})} \mathbb{E}_f \left[\left(n^{-\frac{\beta}{2\beta+1}} \right)^{-2} |\hat{f}_n(x_0) - f(x_0)|^2 \right] \leq C < \infty \qquad \forall x_0 \in [0,1].$$

$$\limsup_{n\to\infty} \sup_{f \in \Sigma(\beta, \mathcal{L})} \mathbb{E}_f \left[\left(n^{-\frac{\beta}{2\beta+1}} \right)^{-2} \|\hat{f}_n - f\|_2^2 \right] \leq C < \infty.$$

Demostración.

Por el lema anterior anterior, bajo las hipótesis de este corolario se cumplen las hipótesis del Corolario 3.1.4, obteniendo el resultado deseado.

□

3.1.2. Riesgo para la pérdida del supremo

Hasta ahora, se ha considerado la pérdida cuadrática para el riesgo. En el modelo, también se considera como posible función de pérdida la pérdida del supremo. Es decir, aquella cuyo riesgo es:

$$R(\hat{f}_n) = \mathbb{E}_f \left(\left\| f - \hat{f}_n \right\|_{\infty}^2 \right). \tag{3.19}$$

Para poder obtener una acotación para esta distancia a partir de la acotación ya obtenida, es importante considerar el siguiente resultado [3, Cor. 1.3]:

[1] Para este resultado no es necesario que los errores ϵ_i sean normales, ni que tengan la misma varianza.

Lema 3.1.8. Sean $\mu_1, \ldots, \mu_M$ variables aleatorias tales que $\max_{1\leq j\leq M} \mathbb{E}[\exp(\alpha_0 \mu_j^2)] \leq C_0 M$ para ciertas constantes $M, a_0 > 0$ y $C_0 < \infty$. Entonces, se cumple que

$$\mathbb{E}\left[\max_{1\leq j\leq M} \mu_j^2\right] \leq \frac{1}{\alpha_0} \log(C_0 M). \tag{3.20}$$

En caso de que $\mu_1, \ldots, \mu_M$ variables sean variables gausianas, la desigualdad es

$$\mathbb{E}\left[\max_{1\leq j\leq M} \|\mu_j\|^2\right] \leq 4d\sigma^2_{\max} \log(\sqrt{2}Md).$$

Teorema 3.1.9. Sea $f \in \Sigma(\beta, L)$ en $[0,1]$, y $\hat{f}_n$ su correspondiente estimador LP de orden $l = \lfloor \beta \rfloor$, escogiendo de ancho de banda

$$h_n = \alpha \left(\frac{\log n}{n}\right)^{\frac{1}{2\beta+1}}$$

para cierto α. Además, suponemos que:

(a) El modelo de regresión cumple las condiciones C.

(b) Se cumplen las suposiciones LP.

(c) El núcleo K es una función lipschziana.

Entonces, existe una constante $C < \infty$ tal que

$$\limsup_{n\to\infty} \sup_{f\in\Sigma(\beta,L)} \mathbb{E}_f\left[\psi_n^{-2} \|\hat{f}_n - f\|^2_\infty\right] \leq C, \tag{3.21}$$

siendo

$$\psi_n = \left(\frac{\log n}{n}\right)^{\frac{\beta}{2\beta+1}}. \tag{3.22}$$

Demostración.

Por brevedad, se denotarán $\mathbb{E} = \mathbb{E}_f$, y $b(x_0) = \mathbb{E}_f[\hat{f}_n(x_0)] - f(x_0)$.
Usando las acotaciones anteriores de la Proposición 3.1.3 (las usadas para estudiar el riesgo asociado a la pérdida cuadrática), y aplicando la desigualdad triangular

$$\begin{aligned}
\mathbb{E}\|\hat{f}_n - f\|^2_\infty &\leq \mathbb{E}\left[\|\hat{f}_n - \mathbb{E}(\hat{f}_n)\|_\infty + \|\mathbb{E}(\hat{f}_n) - f\|_\infty\right]^2 \\
&\leq \mathbb{E}\left[2\|\hat{f}_n - \mathbb{E}(\hat{f}_n)\|_\infty\right] \mathbb{E}\left[2\|f - \mathbb{E}(\hat{f}_n)\|_\infty\right]^2 \\
&\leq 2\,\mathbb{E}\|\hat{f}_n - \mathbb{E}(\hat{f}_n)\|^2_\infty + 2\left(\sup_{x\in[0,1]} |b(x)|\right)^2 \\
&\leq 2\,\mathbb{E}\|\hat{f}_n - \mathbb{E}(\hat{f}_n)\|^2_\infty + 2q_1^2 h_n^{2\beta}.
\end{aligned}$$

De la misma manera que en la Proposición 3.1.3, esto acota la parte del sesgo. Solo falta entonces delimitar el término asociado a la varianza.

Considerando que el estimador $\hat{f}_n$ es un estimador lineal $\left(\hat{f}_n(x) = \sum_{i=1}^n Y_i W_{ni}\right)$:

$$\begin{aligned}
\mathbb{E}\|\hat{f}_n - \mathbb{E}\hat{f}_n\|^2_\infty &= \mathbb{E}\left[\sup_{x\in[0,1]} \left|\hat{f}_n(x) - \mathbb{E}\left[\hat{f}_n(x)\right]\right|^2\right] \\
&= \mathbb{E}\left[\sup_{x\in[0,1]} \left|\sum_{i=1}^n \epsilon_i W_{ni}(x)\right|^2\right].
\end{aligned}$$

Para agilizar la notación, se denotarán $S_i(x)$ y A a:

$$S_i(x) := \boldsymbol{U}\left(\frac{x-x_i}{h}\right)K\left(\frac{x-x_i}{h}\right) \qquad \text{y} \qquad A := \sup_{x\in[0,1]}\left|\sum_{i=1}^{n}\epsilon_i W_{ni}(x)\right|.$$

Al darse LP1, se cumple que:

$$\left|\sum_{i=1}^{n}\epsilon_i W_{ni}(x)\right| = \left\|\sum_{i=1}^{n}\epsilon_i \mathcal{B}_{nx}S_i(x)\right\|_2 \underset{LP1}{\leq} \frac{1}{\lambda_0 nh}\left\|\sum_{i=1}^{n}\epsilon_i S_i(x)\right\|_2.$$

Fijamos $M = n^2$, y consideramos los puntos fijos $x_j = j/M$. Se pueden utilizar los valores x_j para "pivotar" y acotar considerando el superior entre los x con menos de 1/M de distancia a entre ellos por lo dicho anteriormente:

$$\begin{aligned} A &\leq \frac{1}{\lambda_0 nh}\sup_{x\in[0,1]}\left\|\sum_{i=1}^{n}\epsilon_i S_i(x)\right\|_2 \\ &\leq \frac{1}{\lambda_0 nh}\left(\max_{1\leq j\leq M}\left\|\sum_{i=1}^{n}\epsilon_i S_i(x_j)\right\|_2 + \sup_{x,x':|x-x'|<1/M}\left\|\sum_{i=1}^{n}\epsilon_i(S_i(x)-S_i(x'))\right\|_2\right). \end{aligned}$$

Por otra parte, al ser K y U (al ser polinomial) lipschzianas, y el soporte de K estar contenido en $[-1,1]$ (por LP3), existe $\hat{L}$ tal que

$$\|\boldsymbol{U}(u)K(u) - \boldsymbol{U}(u')K(u')\| = \|\boldsymbol{S}(u) - \boldsymbol{S}(u')\| \leq \hat{L}|u-u'|, \qquad \forall u,u' \in \mathbb{R}$$

por lo que

$$\sup_{x,x':|x-x'|<1/M}\left\|\sum_{i=1}^{n}\epsilon_i(S_i(x)-S_i(x'))\right\|_2 = \frac{\hat{L}}{Mh}\sum_{i=1}^{n}|\epsilon_i|.$$

Entonces, podemos acotar el cuadrado de A por

$$\begin{aligned} A^2 &\leq \left(\frac{1}{\lambda_0 nh}\right)^2\left(\max_{1\leq j\leq M}\left\|\sum_{i=1}^{n}\epsilon_i S_i(x_j)\right\| + \frac{\hat{L}}{Mh}\sum_{i=1}^{n}|\epsilon_i|\right)^2 \\ &\leq \frac{2}{\lambda_0^2 nh}\left[\max_{1\leq j\leq M}\left\|\frac{1}{\sqrt{nh}}\sum_{i=1}^{n}\epsilon_i S_i(x_j)\right\|^2\right] + \frac{2\hat{L}^2}{\lambda_0^2 n^2h^4M^2}\left(\sum_{i=1}^{n}|\epsilon_i|\right)^2 \\ &\leq \frac{2}{\lambda_0^2 nh}\left[\max_{1\leq j\leq M}\|\eta_j\|^2\right] + \frac{2\hat{L}^2}{\lambda_0^2 n^2h^4M^2}\left(\sum_{i=1}^{n}|\epsilon_i|\right)^2. \end{aligned}$$

Con

$$\eta_j := \frac{1}{\sqrt{nh}}\sum_{i=1}^{n}\epsilon_i S_i(x_j).$$

Ahora, por la linealidad de la esperanza:

$$\mathbb{E}\|\hat{f}_n - \mathbb{E}\hat{f}_n\|_\infty^2 = \mathbb{E}(A^2) \leq \frac{2}{\lambda_0^2 nh}\mathbb{E}\left[\max_{1\leq j\leq M}\|\mu_i\|^2\right] + \frac{2\hat{L}^2}{\lambda_0^2 n^2h^4M^2}\mathbb{E}\left[\left(\sum_{i=1}^{n}|\epsilon_i|\right)^2\right] \tag{3.23}$$

Por una parte, al ser ϵ_i variables gausianas i.i.d con varianza $\sigma_\epsilon^2 < \infty$, y que se había fijado $M = n^2$:

$$\frac{1}{n^2h^4M^2}\mathbb{E}\left[\left(\sum_{i=1}^{n}|\epsilon_i|\right)^2\right] \leq \frac{\mathbb{E}(\epsilon_1^2)}{M^2h^4} = \frac{\sigma_\epsilon^2}{(nh)^4} = o\left(\frac{1}{nh}\right). \tag{3.24}$$

Por otro lado, μ_j son variables centradas (al serlo ϵ_j). Mediante el mismo razonamiento que el usado en la demostración del Lema 3.1.1:

$$\begin{aligned}\mathbb{E}[\|\mu_j\|^2] &= \sum_{i=1}^{n}\sigma_\epsilon^2\left\|\boldsymbol{U}\left(\frac{x-x_i}{h}\right)\right\|^2 K^2\left(\frac{x-x_{ij}}{h}\right)\\ &\leq \frac{4K_{\text{máx}}^2\sigma_\epsilon^2}{nh}\sum_{i=1}^{n}I(|x-x_{ij}|\leq h)\\ &\underset{LP2}{\leq} 4K_{\text{máx}}^2\sigma_\epsilon^2 a_0\,\text{máx}\left(2,\frac{1}{nh}\right) = q_3.\end{aligned}$$

Como ϵ_j son variables aleatorias normales, por el Lema 3.1.8 del inicio de la sección:

$$\mathbb{E}\left[\text{máx}\,\|\mu_j\|^2\right] = O(\log M) = O(\log n^2) \equiv O(\log n) \quad \text{cuando } n\to\infty \tag{3.25}$$

Con esto, ya tenemos la cota para el primer término:

$$\mathbb{E}\|\hat{f}_n - \mathbb{E}\hat{f}_n\|_\infty^2 \leq \frac{q_3\log n}{nh}.$$

Juntándolo con la cota lograda anteriormente para la parte asociada a la varianza, obtenemos finalmente una acotación del riesgo:

$$\Rightarrow\mathbb{E}\|\hat{f}_n - f\|_\infty^2 \leq \frac{q_3\log n}{nh} + 2\,q_1^2h^{2\beta}.$$

Derivando y despejando como en (3.12), obtenemos que si se considera

$$h_n = \left(\frac{q_3}{2\beta q_1^2}\frac{\log n}{n}\right)^{\frac{1}{2\beta+1}} = \alpha\cdot\left(\frac{\log n}{n}\right)^{\frac{1}{2\beta+1}}, \tag{3.26}$$

se minimiza el riesgo respecto al ancho de banda.

De manera similar al Corolario 3.1.4, esto implica que se tiene que existe una constante $C>0$ tal que

$$\limsup_{n\to\infty}\sup_{f\in\Sigma(\beta,L)}\mathbb{E}_f\left[\left(\frac{\log n}{n}\right)^{\frac{-\beta}{2\beta+1}}\|\hat{f}_n - f\|_\infty^2\right] \leq C, \tag{3.27}$$

que es lo que queríamos probar.

□

Observación 3.1.2. Este teorema implica que la tasa de convergencia ψ_n que obtenemos de los estimadores polinomiales no paramétricos para la pérdida del supremo es mayor que la obtenida para la pérdida cuadrática, ya que tiene un factor $\log n$ que ralentiza la convergencia. Sin embargo, a nivel práctico a veces convendrá utilizar la pérdida del supremo, al ser más fácil de calcular.

Tanto este resultado como el análogo para la pérdida cuadrática ofrecen una acotación superior al riesgo del estimador LP(l) siempre y cuando se escoja el ancho de banda óptimo en cada caso. en concreto, se ha obtenidos que para todo $f \in \Sigma(L, \beta)$, el estimador $LP(l)$ cumple que

$$R(\hat{f}) = \mathbb{E}_f \left\| \hat{f}_n - f \right\|_2^2 \leq \frac{C}{n^{\frac{2\beta}{2\beta+1}}}$$

y

$$R(\hat{f}) = \mathbb{E}_f \left\| \hat{f}_n - f \right\|_\infty^2 \leq \frac{C}{\left(\frac{n}{\log n}\right)^{\frac{2\beta}{2\beta+1}}}.$$

Esta acotación del riesgo obtenida es mejor para mayores grados del polinomio de estimación (lo que tiene sentido, al añadir mayor suavidad y complejidad). Mientras que la acotación (para la pérdida cuadrática) del estimador Nadaraya-Watson ($LP(0)$) es una constante, para $LP(1)$ es $\frac{C}{n^{2/3}}$ (luego el riesgo decrece con un mayor conjunto de entrenamiento). En el caso idílico de utilizar un polinomio local de grado infinito, se lograría incluso a obtener una tasa de convergencia inversamente proporcional al tamaño del conjunto de entrenamiento utilizado ($\frac{C}{n}$).

Esto se ilustrará en la siguiente sección en el estudio de aplicación de los estimadores $LP(l)$ sobre unos ejemplos, que ayudarán a comprender varios detalles y características de estos estimadores que se han estado mencionando a lo largo de este a sección.

3.2. Comparación de los estimadores locales polinomiales de distintos grados

En este apartado, se pondrá de manifiesto la diferencia entre utilizar distintos estimadores locales, y cómo su eficacia variará según la función que se quiera estimar. En concreto, se analizará la diferencia entre utilizar el estimador Nadaraya-Watson (que es LP(0)), a usar un estimador LP(1), a veces llamado estimador lineal local .

Para ello, se comparará la estimación conseguida con ambos estimadores, a partir de un conjunto de entrenamiento generado de manera uniforme de 200 elementos en el intervalo $I = [0, 10]$, a las que a las etiquetas se les ha imbuido un error ϵ de varianza $0,2$.

Como se comentó a la hora de acotar el riesgo, la elección de uno u otro ancho de banda afectará al riesgo del estimador. En concreto, como aparece, en la ecuaciones (3.12) y (3.26), el ancho de banda óptimo sería, en cada caso,

$$h_n = \left(\frac{q_2}{2\beta q_1^2}\right)^{\frac{1}{2\beta+1}} n^{-\frac{1}{2\beta+1}} \qquad y \qquad h_n = \left(\frac{q_3}{2\beta q_1^2}\frac{\log n}{n}\right)^{\frac{1}{2\beta+1}}.$$

Sin embargo, las constantes q_1, q_2 y q_3 dependen de constantes desconocidas fuera de un entorno teórico, como la varianza de los errores ϵ_i, o la constante L de la clase de Hölder $\Sigma(L, \beta)$ a la que pertenece la función a estimar. Esto sugiere que, en los supuestos prácticos como este, se estime el ancho de banda utilizado. Un método para obtener una buena estimación es mediante el método de la **validación cruzada** (ver Sección A.1 del Apéndice).

Las funciones que se considerarán poseen ciertas características que será interesante ver cómo el estimador se comporta frente a ellas, como el poseer un polinomio local de grado infinito o el no ser derivable en ciertos puntos. Las funciones a estimar van a ser:

$f(x) = x^2 - 1$(función cuyo desarrollo de Taylor es un polinomio local de segundo grado).

Este primer ejemplo ya pone de manifiesto la mejoría en la estimación conseguida al considerar estimadores locales polinomiales de mayor grado, siendo los valores optindos por el estimador local polinomial (PE) mucho más regulares a simple vista.

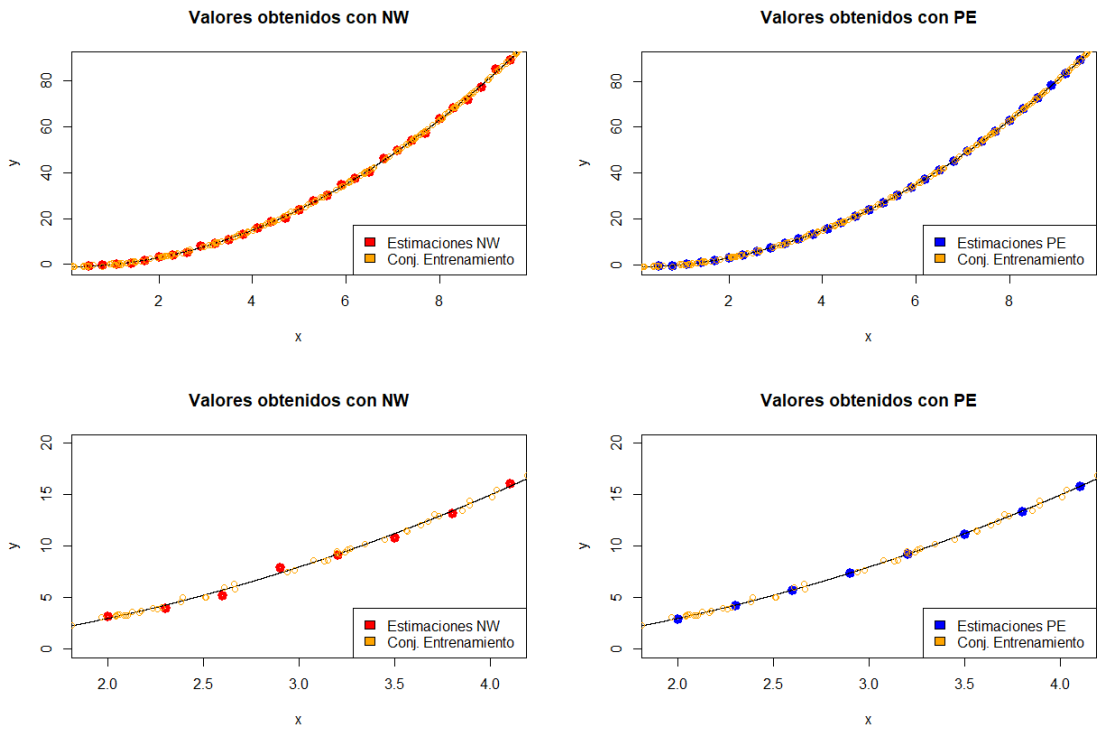


Figura 3.2: Estimaciones de la función $f(x) = x^2 - 1$.

La función a estimar es un polinomio de segundo grado, lo que implica que el polinomio local de f de grado 2 sería su desarrollo de Taylor:

$$f(x) = x_0^2 - 1 + 2x_0(x - x_0) + (x - x_0)^2.$$

Para el estimador lineal local, hay que considerar el polinomio local de grado 1. Es decir, el único término que no se considerará a la hora de estimar es $(x - x_0)^2$, que será pequeño. Sin embargo, para el estimador Nadaraya-Watson, tampoco se considera $2x_0(x - x_0)$. Esto provoca que el error cometido con el estimador de Nadaraya Watson (denotado NW) sea mayor, acentuándose un poco para mayores valores de x.

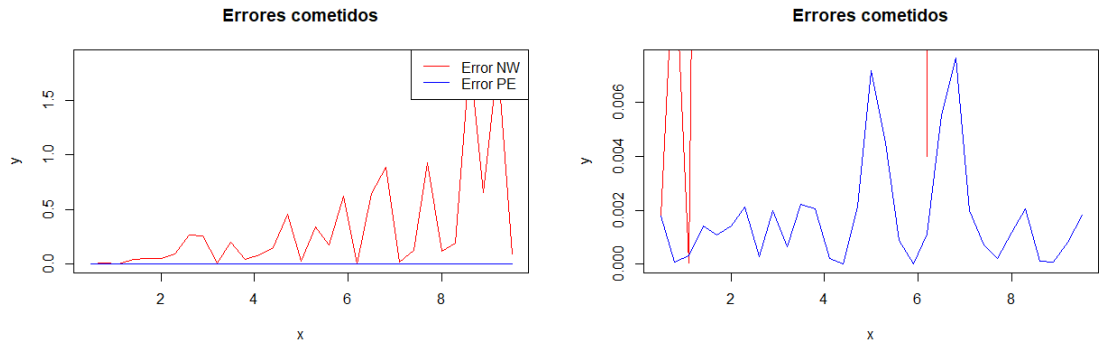


Figura 3.3: Comparación de los errores obtenidos con cada estimador para $f(x) = x^2 - 1$.

Además, en la gráfica anterior se muestra también que hay cierta variabilidad del error a lo largo del intervalo. Esta variabilidad es debida a la mayor o menor cantidad de atributos del conjunto de entrenamiento cercanos al valor a estimar, poniendo en manifiesto el por qué para acotar el riesgo de estos estimadores se es necesaria la hipótesis LP2.

$f(x) = \sin(x^2)$

En este caso, se va a estimar una función cuyo desarrollo de Taylor tiene infinitos términos:

$$f(x) = \sin(x_0^2) + 2x_0\cos(x_0^2)(x - x_0) + \frac{2(\cos(x_0) - x_0^2\sin(x_0^2))}{2}(x - x_0)^2 + \dots$$
$$\Longrightarrow f(x) = \sin(x_0^2) + 2x_0\cos(x_0^2)(x - x_0) + \frac{2(\cos(x_0) - x_0^2\sin(x_0^2))}{2}(x - x_0)^2 + O(x^3),$$

por lo que para ambos estimadores, al igual que le sucedía al estimador NW en el ejemplo anterior,los polinomios locales que utilizarán los estimadores no contemplarán todos los términos del desarrollo de Taylor de la función.

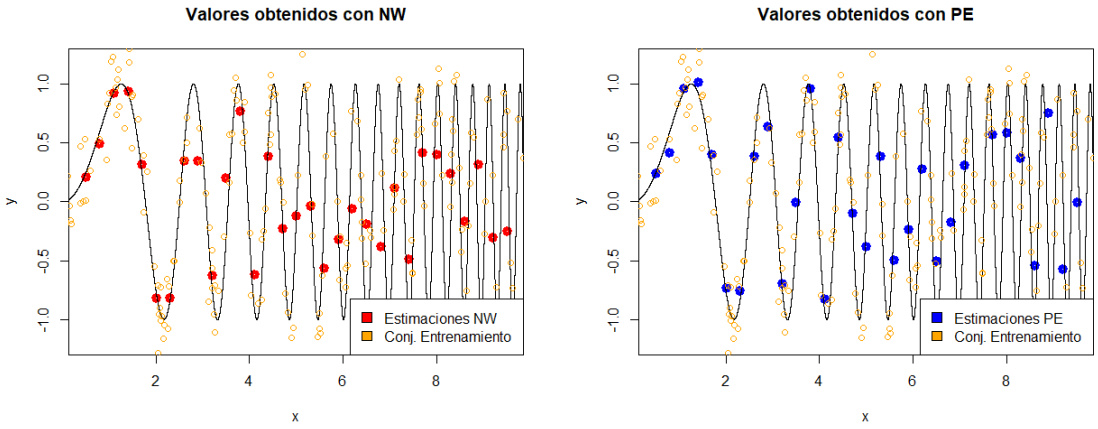


Figura 3.4: Estimaciones de la función $f(x) = \sin(x^2)$

Como se puede observar, en $f(x) = \sin(x^2)$, cuanto mayor es el valor de x, más "estrecha" será la onda, lo que implica que en cada periodo de la onda las derivadas serán cada vez mayores, siendo entonces mayor el error cometido por los estimadores cuanto mayor sea el valor del que estimar la etiqueta.

De la misma manera, en esos mismos valores es donde también se nota la diferencia entre utilizar un estimador u otro, ya que el error cometido por el PE (*Polinomial Estimator*) es menor al considerar la derivada local.

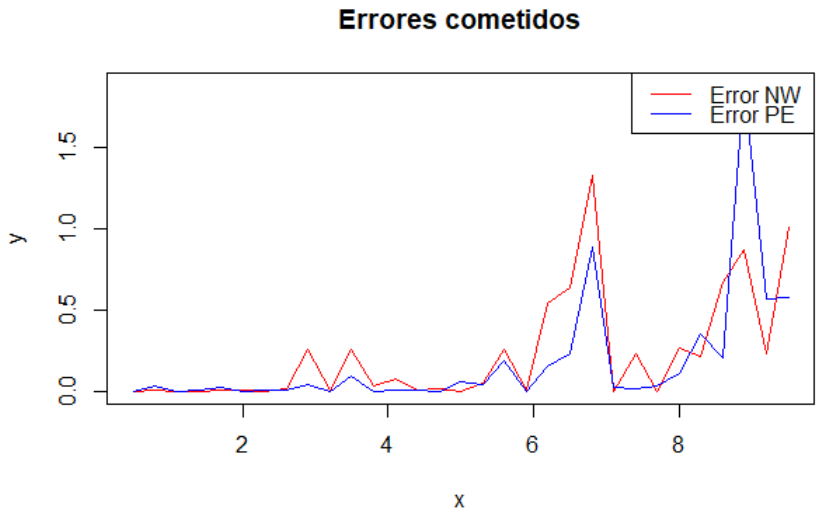


Figura 3.5: Errores cometidos al estimar la función $f(x) = \sin(x^2)$

$f(x) = \log(x)$ (función con asíntota vertical).

En el análisis de datos y la estadística en general, es bastante común encontrar relaciones logarítmicas. Por ello, otra función interesante para analizar las estimaciones es $f(x) = \log(x)$. Con los resultados conseguidos para ambos estimadores, nuevamente queda reflejada la dificultad que tiene el estimador NW al estimar con modelos cuya distribución no sea muy uniforme.

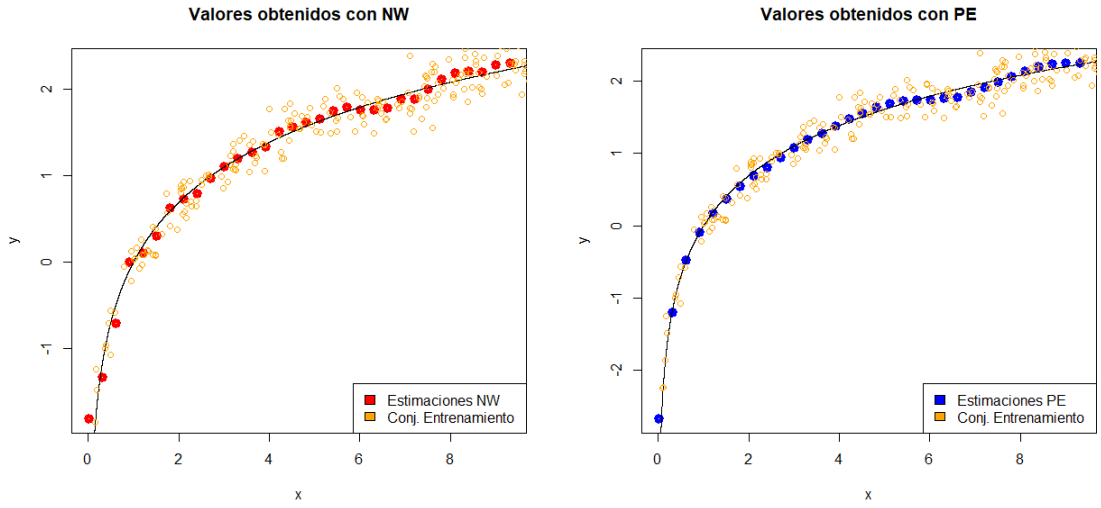


Figura 3.6: Estimaciones de la función $f(x) = \log x$

Por otra parte, es muy interesante observar la gráfica de los errores obtenidos por ambos estimadores, puesto que ambos se disparan cerca del 0 (donde se encuentra la asíntota) [2]:

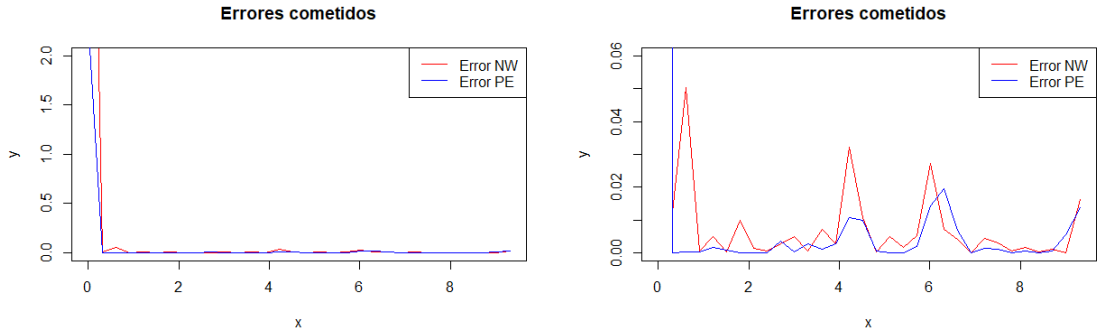


Figura 3.7: Estimaciones de la función $f(x) = \log x$ para ambos estimadores. A la izquierda, visualizando el aumento del error para el valor más cercano a 0 ($x = 0{,}01$); a la izquierda obviando ese valor.

Se puede observar que en ambos casos el error se dispara para valores cercanos a $x = 0$ (de varios órdenes de magnitud en ambos casos), pero es más leve para el caso de $LP(1)$. Este crecimiento ya que la función f es de Hölder en $[\epsilon, 10]$, pero no en $[0, 10]$. Cuanto el valor de x sea más cercano a 0, la constante L será mayor, aumentando la cota del error.

$f(x) = |x - 5|$ (función no derivable en ciertos puntos).

Por último, se va a presentar cómo a los estimadores les afecta los puntos singulares. Para ello, se va a considerar la función $f(x) = |x - 5|$, que no es derivable en $x = 5$.

[2] Por eso mismo motivo, para realizar estas estimaciones el primer valor estimado no ha sido $x = 0$, sino $x = 0{,}01$

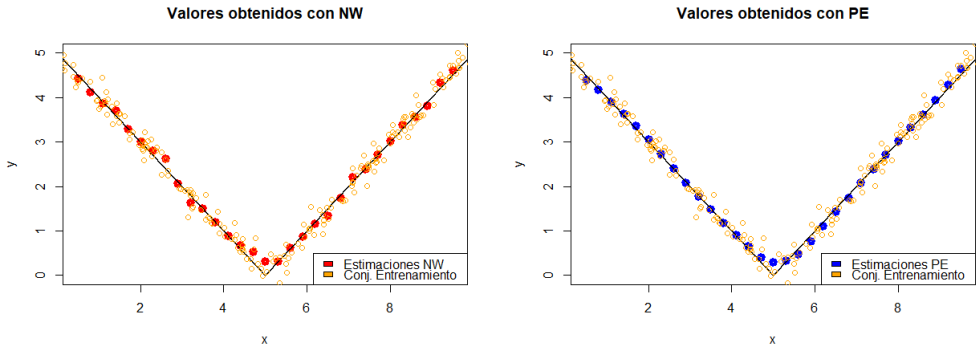


Figura 3.8: Estimaciones de la función $f(x) = |x - 5|$

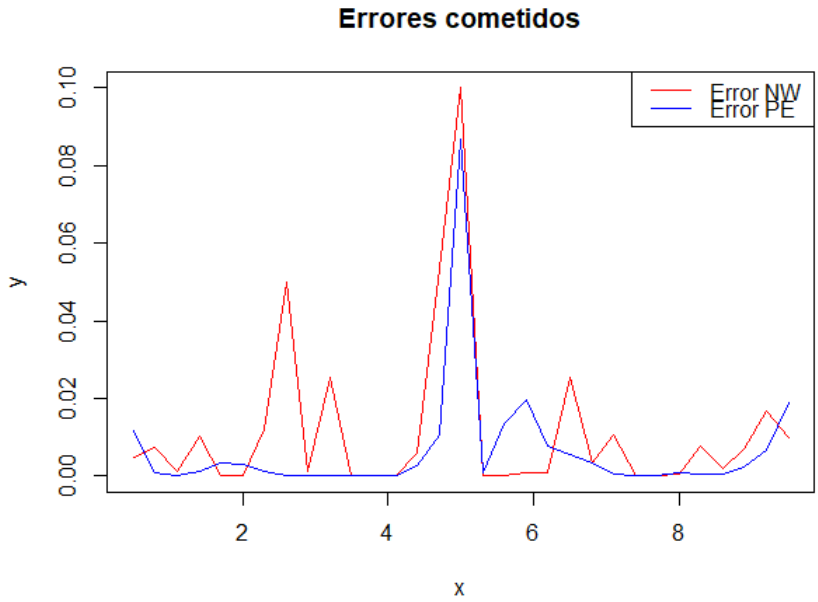


Figura 3.9: Errores de la función $f(x) = |x - 5|$

Para valores cercanos a $x = 5$ (donde f no es derivable, pero sí lineal) se puede apreciar que la estimación polinomial se curva, puesto intenta estimarla por una función derivable. Esto produce que en puntos cercanos a 5 el error del estimador polinomial sea bastante mayor de lo habitual. Sin considerar el valor puntual de $x = 5$, en el resto de su dominio f sí es de Hölder de grado 1, y por ello ahí el estimador lineal local funciona bien, obteniendo errores similares a los obtenidos para $f(x) = x^2 - 1$.

Estos ejemplos muestran cómo, a pesar de que su tiempo de computación es mayor al estimador NW, el rendimiento de los estimadores locales polinomiales es bastante superior, haciendo que merezca la pena el coste computacional que suponen. La elección del grado l del estimador a considerar dependerá de cuanto coste computacional se está dispuesto a aceptar, ya que el estimador $LP(l)$ requiere de la inversión de la matriz $\mathcal{B}_{nx}$, una matriz de dimensión $l + 1$.

En esta sección, se ha estado comparando cómo el rendimiento del estimador polinomial mejora cuando aumenta el grado l elegido; implicando que el trabajar sobre una clase de Hölder $\Sigma(L, \beta)$ de mayor orden aumentará la efectividad de la estimación (a costa del rendimiento computacional).

Tras haber dejado clara la eficiencia de los estimadores polinomiales loclaes, el siguiente paso entonces será saber si estos estimadores de f son los mejores posibles trabajando en las clases de Hölder, para lo que se tendrá que aplicar la Teoría Minimax presentada en la Sección 2.3, y las acotaciones obtenidas en la Sección 2.5.

3.3. Optimalidad de los estimadores $LP(l)$.

El último punto a tratar sobre los estimadores locales polinomiales es ver que son óptimos en tasa en la clase de Hölder en el modelo estudiado.

Como se comentó en 2.3, puesto que en la Sección 3.1 ya se ha demostrado que los estimadores $LP(l)$ cumplen que

$$\sup_{f\in\Sigma(L,\beta)} \mathbb{E}_f\left[d^2(\hat{f}_n,f)\right] \leq C\psi_n^2, \tag{3.28}$$

para cierta sucesión $(\psi_n)_{n=1}^{\infty}$ ($\psi_n = n^{-\frac{\beta}{2\beta+1}}$ para la pérdida puntual/cuadrática, y $\psi_n = \left(\frac{\log n}{n}\right)^{\frac{\beta}{2\beta+1}}$ para la pérdida del supremo), para demostrar que son estimadores óptimos en tasa en $\Sigma(L,\beta)$ (y, a la vez, que $(\psi_n)_{n=1}^{\infty}$ es la tasa de convergencia óptima), es suficiente con demostrar que se cumple la siguiente desigualdad:

$$\liminf_{n\to\infty} \psi_n^{-2}\mathbb{E}_f\left[d^2(\hat{f}_n,f)\right] \geq c. \tag{3.29}$$

Para dar con ella, se utilizará el Teorema Principal de las Cotas Inferiores[3] (2.5.7), pasando entonces a ser el objetivo de la sección el encontrar las funciones hipótesis dentro de $\Sigma(L,\beta)$ que cumplan las condiciones del problema, y así obtener nuestro objetivo.

3.3.1. Estimadores óptimos puntuales en la clase de Hölder

Primero, se considerará la distancia entre la hipótesis y el estimador en un **punto fijo** x_0. Es decir, se trabajará como pseudodistancia dentro del modelo estadístico la asociada a la pérdida puntual, para luego extenderlo a la pérdida cuadrática y a la del supremo.

El objetivo será entonces obtener una cota del riesgo minimax en (Θ,d), siendo $\Theta = \Sigma(\beta,L)$, y considerando la distancia $d(f,g) = |f(x_0) - g(x_0)|$ **para el punto fijo** $x_0 \in [0,1]$.

Por lo comentado anteriormente, para ver que $LP(l)$ es óptimo en tasa, basta con demostrar que el riesgo minimax $\mathcal{R}_n^*$ de $(\Sigma(L\beta),d)$ está acotado inferiormente por la misma tasa de convergencia; es decir, existe $c>0$ tal que:

$$\liminf_{n\to\infty} \psi_n^{-2}\mathcal{R}_n^* \geq c, \tag{3.30}$$

ya que mediante esa desigualdad se demostrará que la tasa de convergencia del estimador polinomial, $(\psi_n)_{n=0}^{\infty} es$, la tasa de convergencia óptima.

El Teorema 2.3.1 indica que si existen $f_{0n}, f_{1n} \in \Sigma(L,\beta)$ tales que $d(f_{0n},f_{1n}) \geq 2s$ que cumplen

$$\inf_{T_n} \max_{f\in\{f_{0n},f_{1n}\}} P_f(|T_n(x_0) - f(x_0)| \geq s) \geq c' > 0, \tag{3.31}$$

siendo $s = A\psi_n$ (para cierta constante $A>0$), entonces por el Teorema 2.3.1 se cumplirá que

$$\inf_{T_n} \sup_{f\in\Sigma(L,\beta)} \mathbb{E}_f\left[d^2(T_n,f)\right] \geq c'\psi_n^2,$$

dando la cota inferior requerida, obteniendo así que en $(\Sigma(L,\beta),d)$ el estimador $LP(l)$ es óptimo en tasa.

Para ello, se tendrá que encontrar ciertas funciones hipótesis f_{0n} y f_{1n} tales que $d(f_{0n},f_{1n}) \geq 2s$ con las que se pueda obtener una acotación suficientemente buena. Para ello, se considerarán como hipótesis

$$f_{0n}(x) \equiv 0 \qquad y \qquad f_{1n}(x) = Lh_n^{\beta}K\left(\frac{x-x_0}{h_n}\right) \tag{3.32}$$

[3] Se podría utilizar de igual maner el Lema de Fano 2.6.4 para este cometido.

siendo $h_n = c_0 n^{-\frac{1}{2\beta+1}}$, y K siendo una función que satisface[4]:

$$K \in \Sigma(\beta, 1/2) \cap C^\infty(\mathbb{R}) \qquad y \qquad K(u) > 0 \Leftrightarrow u \in (-1/2, 1/2). \tag{3.33}$$

Para la demostración, se utilizará el Teorema 2.5.3, a partir de la divergencia de Kullback. Esto implica que P_0 y P_1 tienen que cumplir que $K(P_0, P_1) < \infty$, por lo que se necesitará un lema anterior:

Lema 3.3.1. Sean

$$f_{0n}(x) \equiv 0 \qquad y \qquad f_{1n}(x) = L h_n^\beta K\left(\frac{x - x_0}{h_n}\right).$$

Entonces, considerando $\alpha = L^2 K_{\text{máx}}^2 n h^{2\beta+1}$, las probabilidades $P_i = \mathcal{L}(Y_1, \ldots, Y_n | f_{in})$ cumplen que $K(P_0, P_1) < \alpha < \infty$ bajo las condiciones C.

Demostración.

Como P_j es la función de distribución de $Y_1, \ldots, Y_n$ para $f = f_{jn}$, implica que su función de densidad cumple que, denotando p_ϵ a la función de densidad de la variable aleatoria del error ϵ:

$$p_j(u_1, \ldots, u_n) = \prod_{i=1}^n p_\epsilon(u_i - f_{jn}(X_i)) \tag{3.34}$$

para $j = 0, 1$.

Cuando $h_n = c_0^\beta n^{-\frac{1}{2\beta+1}}$, $nh_n \to \infty$; por lo que existe un n_0 (que depende de los valores que conforta h: $c_0, L, \beta, K_{\text{máx}}, v_0$ (v_0 es la constante que acota)), tal que para todo $n > n_0$, se cumple que $nh_n \geq 1$ y $L h_n^\beta K_{\text{máx}} \leq v_0$.

Puesto que se considera que los errores ϵ_i siguen una distribución normal $\mathcal{N}(0, \sigma^2)$ (por la condición C2 del Modelo), se cumple que

$$\int p_\epsilon(u) \log \frac{p_\epsilon(u)}{p_\epsilon(u+v)}\, du = -\int p_\epsilon(u) \log p_\epsilon(u+v)\, du + \int p_\epsilon(u) \log p_\epsilon(u)\, du$$

$$= \log \frac{\sigma}{\sigma} + \frac{\sigma^2 + (0-v)^2}{2\sigma^2} - \frac{1}{2} = v^2/2. \tag{3.35}$$

Con esa desigualdad, la divergencia de Kullback entre las medidas producto P_0 y P_1 (por lo que utilizando la Proposición 2.4.6) cumple:

$$\begin{aligned}
K(P_0, P_1) &= \int \log \frac{dP_0}{dP_1} dP_0 \underset{(2.40)}{=} \int \ldots \int \log \prod_{i=1}^n \frac{p_\epsilon(u_i)}{p_\epsilon(u_i - f_{1n}(x_i))} \prod_{i=1}^n [p_\epsilon(u_i) du_i] \\
&= \sum_{i=1}^n \int \log \prod_{i=1}^n \frac{p_\epsilon(y)}{p_\epsilon(y - f_{1n}(x_i))} p_\epsilon(y) dy \\
&\underset{(3.35)}{=} \frac{1}{2} \sum_{i=1}^n f_{1n}^2(x_1) = \frac{1}{2} L^2 h_n^{2\beta} \sum_{i=1}^n K^2\left(\frac{x_i - x_0}{h_n}\right) \\
&\underset{(3.33)}{\leq} \frac{1}{2} L^2 h_n^{2\beta} K_{\text{máx}}^2 \sum_{i=1}^n I\left(\left|\frac{x_i - x_0}{h_n}\right| \leq \frac{1}{2}\right) \\
&\underset{LP2}{\leq} \frac{1}{2} L^2 h_n^{2\beta} K_{\text{máx}}^2 a_0 \operatorname{máx}(nh_n, 1) \\
&= \frac{1}{2} a_0 L^2 K_{\text{máx}}^2 n h_n^{2\beta+1}.
\end{aligned}$$

[4] En A.3.1 está demostrada la existencia de una función que cumple dichas características

Puesto que la condición C2 implica la condición LP2 para $a_0 = 2$, la acotación finalmente obtenida es

$$K(P_0, P_1) \leq L^2 K_{\text{máx}}^2 n h_n^{2\beta+1} = \alpha.$$

□

Teorema 3.3.2. Sean β y $L > 0$. Bajo las condiciones C, para todo $x_0 \in [0,1]$, $t > 0$, se cumple:

$$\liminf_{n\to\infty} \inf_{T_n} \sup_{f\in\Sigma(\beta,L)} P_f \left(n^{\frac{\beta}{2\beta+1}} |T_n(x_0) - f(x_0)| \geq t^{\frac{\beta}{2\beta+1}} \right) \geq V_0(ct), \tag{3.36}$$

siendo T_n los estimadores escogidos y $V_0(x) := \text{máx}\left(1/4\, e^{-x}, \frac{1-\sqrt{x/2}}{2}\right)$. Además, $c > 0$ solo depende de β, L y a_0. También

$$\liminf_{n\to\infty} \inf_{T_n} \sup_{f\in\Sigma(\beta,L)} \mathbb{E}_f \left[n^{\frac{2\beta}{2\beta+1}} |T_n(x_0) - f(x_0)|^2 \right] \geq c_1 > 0, \tag{3.37}$$

donde c_1 también solo depende de β, L y a_0.

Demostración.

Si se encuentran dos hipótesis f_{0n} y f_{1n} tales que $f_{jn} \in \Sigma(\beta, L)$, que cumplan que $d(f_{0n}, f_{1n}) \geq 2s$, por el Teorema 2.3.1, significará que

$$\begin{aligned} &\liminf_{n\to\infty} \inf_{T_n} \sup_{f\in\Sigma(\beta,L)} P_f \left(n^{\frac{\beta}{2\beta+1}} |T_n(x_0) - f(x_0)| \geq t^{\frac{\beta}{2\beta+1}} \right) \\ \geq &\liminf_{n\to\infty} \inf_{T_n} \sup_{f\in\{f_{0n},f_{1n}\}} P_f \left(n^{\frac{\beta}{2\beta+1}} |T_n(x_0) - f(x_0)| \geq t^{\frac{\beta}{2\beta+1}} \right). \end{aligned}$$

El objetivo será encontrar una acotación para la última expresión.

Veamos que las funciones hipótesis definidas en (3.32) cumplen esas condiciones:

$$f_{0n}(x) \equiv 0 \qquad y \qquad f_{1n}(x) = L h_n^\beta K\left(\frac{x - x_0}{h_n}\right),$$

se cumplen .

1. $f_{jn} \in \Sigma(\beta, L)$

   Primero, es obvio que $f_{0n} \equiv 0 \in \Sigma(\beta, L)$.

   Por otra parte, la l-ésima derivada de f_{1n} (siendo $l = \lfloor \beta \rfloor$) es:

$$f_{1n}^{(l)}(x) = L h^{\beta-l} K^{(l)}\left(\frac{x - x_0}{h_n}\right).$$

   Veamos que para cualquier l, $f_{1n}^{(l)}$ cumple la condición de Hölder, lo que implica que $f_{1n} \in \Sigma(\beta, L)$:

$$\begin{aligned} |f_{1n}^{(l)}(x) - f_{1n}^{(l)}(x')| &= L h_n^{\beta-l} \left| K^{(l)}\left(\frac{x - x_0}{h_n}\right) - K^{(l)}\left(\frac{x' - x_0}{h_n}\right) \right| \\ &\leq L h_n^{\beta-l} \frac{1}{2} \left| \left(\frac{x - x_0}{h_n}\right) - \left(\frac{x' - x_0}{h_n}\right) \right|^{\beta-l} \\ &= \frac{L}{2} |x - x'|^{\beta-l}, \end{aligned}$$

   donde se ha utilizado que $K^{(l)}$ cumple la condición de Hölder (eligiendo K como en lo dicho en la Proposición A.3.1). Por tanto, $f_{0n}, f_{1n} \in \Sigma(\beta, L)$ en $[0,1]$.

2. $d(f_{0n}, f_{1n}) \geq 2s$

$$\begin{aligned} d(f_{1n}, f_{0n}) &= |f_{1n}(x_0) - f_{0n}(x_0)| = |f_{1n}(x_0)| = \\ &= Lh_n^{\beta} K(0) = Lc_0^{\beta} K(0)\, n^{-\frac{\beta}{2\beta+1}}. \end{aligned}$$

Por tanto, si consideramos $s = A\psi_n := \frac{1}{2} Lc_0^{\beta} K(0)\, n^{-\frac{\beta}{2\beta+1}}$, lo tenemos.

Puesto que $K(P_0, P_1) < \alpha < \infty$ por el lema anterior, se puede considerar la cota obtenida a partir de la divergencia de Kullback en el Teorema 2.5.3, denominado V_0 a dicha función:

$$p_{e,1} \geq \text{máx}(\frac{1}{4} e^{-\alpha}, \frac{1 - \sqrt{\alpha/2}}{2}) = V_0(\alpha).$$

Esto sugiere intentar demostrar la siguiente desigualdad para obtener el resultado deseado:

$$\liminf_{n\to\infty} \inf_{T_n} \sup_{f_{0n}, f_{1n}} P_f \left(n^{\frac{\beta}{2\beta+1}} |T_n(x_0) - f(x_0)| \geq t^{\frac{\beta}{2\beta+1}} \right) \geq V_0(ct).$$

Por la cota que se obtiene de utilizar la divergencia de Kullback, para todo $n > n_0$, y cualquier estimador T_n:

$$\begin{aligned} \sup_{f \in \Sigma(\beta, L)} P_f(|T_n(x_0) - f(x_0)| \geq s_n) &\geq \max_{P_j} P_f(|T_n(x_0) - f(x_0)| \geq s_n) \\ &\geq \text{máx}(\frac{1}{4} e^{-\alpha}, \frac{1 - \sqrt{\alpha/2}}{2}) = V_0(\alpha). \end{aligned}$$

Sea $t > 0$. Recordemos que $s = \frac{1}{2} Lc_0^{\beta} K(0) \psi_n = A\psi_n$.

Considerando $A = t^{\frac{\beta}{2\beta+1}}$, (o lo que es lo mismo, $t = A^{\frac{2\beta+1}{\beta}} > 0$) se cumple que, al ser $\alpha = L^2 K_{\text{máx}}^2 n h^{2\beta+1}$,

$$t = \left(\frac{1}{2} Lc_0^{\beta} K(0) \right)^{\frac{2\beta+1}{\beta}} = \left(\frac{1}{2} LK(0) \right)^{\frac{2\beta+1}{\beta}} c_0^{2\beta+1} = \left(\frac{1}{2} LK(0) \right)^{\frac{2\beta+1}{\beta}} \frac{\alpha}{L^2 K_{\text{máx}}^2} := \frac{\alpha}{c},$$

donde

$$c := \frac{L^2 K_{\text{máx}}^2}{\left(\frac{1}{2} LK(0) \right)^{\frac{2\beta+1}{\beta}}}$$

solo depende de los valores de β y L (la función K es conocida).

Con esto en mente:

$$\begin{aligned} &\liminf_{n\to\infty} \inf_{T_n} \sup_{f \in \Sigma(\beta, L)} P_f \left(n^{\frac{\beta}{2\beta+1}} |T_n(x_0) - f(x_0)| \geq t^{\frac{\beta}{2\beta+1}} \right) \\ =& \liminf_{n\to\infty} \inf_{T_n} \sup_{f \in \Sigma(\beta, L)} P_f \left(n^{\frac{\beta}{2\beta+1}} |T_n(x_0) - f(x_0)| \geq A \right) \\ =& \liminf_{n\to\infty} \inf_{T_n} \sup_{f \in \Sigma(\beta, L)} P_f \left(|T_n(x_0) - f(x_0)| \geq s_n \right) \geq V_0(\alpha) = V_0(c\,t) \end{aligned}$$

para todo $t > 0$.

Por último, fijando el valor de t a $t_1 > 0$ se tiene que $V(c\,t_1) = c_1 < \infty$. Mediante la desigualdad (2.37), se obtiene la última desigualdad deseada:

$$\begin{aligned}
&\liminf_{n\to\infty} \inf_{T_n} \sup_{f\in\Sigma(\beta,L)} \mathbb{E}_f \left[n^{\frac{2\beta}{2\beta+1}} \left(T_n(x_0) - f(x_0)\right)^2 \right] \\
=&\liminf_{n\to\infty} \inf_{T_n} \sup_{f\in\Sigma(\beta,L)} \mathbb{E}_f \left[\left(n^{\frac{\beta}{2\beta+1}} \left|T_n(x_0) - f(x_0)\right| \right)^2 \right] \\
\geq&\liminf_{n\to\infty} \inf_{T_n} \sup_{f\in\Sigma(\beta,L)} P_f \left(n^{\frac{\beta}{2\beta+1}} \left|T_n(x_0) - f(x_0)\right| \geq t_1^{\frac{\beta}{2\beta+1}} \right) \geq c_1.
\end{aligned}$$

□

Es decir, en el caso en el que se considere el modelo estadístico $(\Sigma(\beta, L), d)$, siendo d la distancia en un punto fijo x_0, la tasa óptima de convergencia es $\psi_n = n^{-\frac{\beta}{2\beta+1}}$, y el estimador local polinomial de grado $l = \lfloor\beta\rfloor$, es óptimo en tasa en $(\Sigma(\beta, L), d)$ bajo el modelo de estudio (Definición 2.1.1).

En el caso de que no se considere una distancia puntual (es decir, considerar por ejemplo $d(\cdot,\cdot) = \|\|\|_2$), intentar obtener cotas inferiores a partir de solo dos hipótesis no da buenos resultados:

Ejemplo 3.3.1. Vamos qué sucede si intenta aplicar esta acotación a la pérdida cuadrática; es decir, considerar

$$d(f,g) = \|f-g\|_2 = \left(\int_0^1 (f(x) - g(x))^2\, dx \right)^{1/2}.$$

.

Si se intenta lo mismo que para la distancia puntual, considerando nuevamente las hipótesis (3.32):

$$f_{0n}(x) \equiv 0 \qquad y \qquad f_{1n}(x) = L h_n^\beta K\left(\frac{x - x_0}{h_n} \right),$$

se sigue obteniendo que $h_n = O\left(\exp\left(\frac{-1}{2\beta+1} \right) \right)$, puesto que hasta ese punto el papel de la norma escogida es irrelevante. Sin embargo, en este caso la distancia entre las hipótesis no se comporta de la misma manera, no obteniendo la acotación deseada:

$$\begin{aligned}
d(f_{0n}, f_{1n}) = \|f_{0n} - f_{1n}\|_2 &= \left(\int_0^1 (f_{1n})^2\, dx \right)^{1/2} \\
&= L h_n^\beta \left(\int_0^1 K^2 \left(\frac{x - 1/2}{h_n} \right) dx \right)^{1/2} \\
&= L h_n^{\beta+\frac{1}{2}} \left(\int_0^1 K^2(u)\, du \right)^{1/2}.
\end{aligned}$$

Por tanto, $d(f_{0n}, f_{1n}) \asymp h_n^{\beta+\frac{1}{2}} = O(n^{-1/2})$, bastante menor que orden de convergencia de la cota superior que habíamos obtenido para el riesgo L^2 en $\Sigma(\beta, L)$, $n^{-\frac{\beta}{2\beta+1}}$.

Es decir, para la distancia L^2, la cota inferior obtenida con solo dos hipótesis no es suficientemente buena para obtener información relevante. Por ello, en este caso será necesario trabajar con más hipótesis.

3.3.2. Estimadores óptimos en la clase de Hölder para la pérdida cuadrática y del supremo.

En esta sección se tratará de extrapolar el resultado obtenido en la sección anterior a la pérdida cuadrática. Si, al igual que en el caso puntual, la tasa de convergencia obtenida de esta manera

coincide con el orden de convergencia de la cota superior del riesgo de los estimadores locales polinomiales, se habrá obtenido entonces que éstos son estimadores óptimos en tasa en $(\Sigma(\beta, L), L^2)$.

Es decir, el objetivo de esta sección será que, considerando un modelo de regresión bajo las hipótesis B, tomando como distancia

$$d(f,g) = \|f-g\|_2 = \left(\int_0^1 (f(x)-g(x))^2\,dx\right)^{1/2}, \tag{3.38}$$

obtener que los estimadores locales polinomiales son óptimos en $\Sigma(\beta, L)$ de manera similar a lo que se ha obtenido para distancia puntual.

Sin embargo, como se demostró en el ejemplo 3.3.1, hace falta considerar más de dos hipótesis para obtener una acotación relevante en este caso.

Siguiendo el razonamiento seguido en la sección anterior, el objetivo será encontrar $M+1$ hipótesis $(f_{jn})_{j=1}^M$ tales que $d(f_{jn}, f_{kn}) \geq s$, siendo $s = s_n = A\psi_n = n^{\frac{\beta}{2\beta+1}}$ con $A > 0$, para que se pueda aplicar la Proposición 2.3.1 y hallar una cota inferior al riesgo minimax c' a partir de la probabilidad de error, obteniendo así que ψ_n es la tasa de convergencia óptima y, por tanto, que los estimadores locales polinomiales son óptimos en tasa dentro de la clases de Hölder.

Es decir, para obtener que $LP(l)$ es un estimador óptimo en tasa en $(\Sigma(L,\beta), d)$ en nuestro modelo, se tiene que conseguir encontrar una constante c' tal que cumpla

$$\inf_{T_n} \max_{f\in\{f_{0n},\ldots,f_{Mn}\}} P_f(\|T_n - f\| \geq A\psi_n) \geq c' > 0$$

donde $\inf_{T_n}$ es el inferior en todos los estimadores posibles.

Sin embargo, por ahora se desconoce qué funciones f_{jn} hay que considerar, o siquiera cuál es el valor M de hipótesis que se necesitan.

Considerando como punto de partida las funciones hipótesis del caso puntual, se quieren definir varias funciones en $\Sigma(\beta, L)$ que nos permitan aplicar el Teorema Principal de las Cotas Inferiores del Riesgo (Teorema 2.5.7) para la distancia L^2.

Para ello, se consideran (recordemos que $s = A\psi_n = \frac{1}{2}Lc_0^\beta K(0)\, n^{-\frac{\beta}{2\beta+1}}$):

$$m := \lceil c_0 n^{\frac{1}{2\beta+1}} \rceil, \qquad h_n := \frac{1}{m}, \tag{3.39}$$

$$x_k := \frac{k-1/2}{m} \qquad \varphi_k(x) = Lh_n^\beta K\left(\frac{x-x_k}{h_n}\right) \tag{3.40}$$

siendo $k = 1,\ldots,m$, $x \in [0,1]$, y $K : \mathbb{R} \to [0,\infty)$ siendo una función que cumple que

$$K \in \Sigma(\beta, 1/2) \cap C^\infty(\mathbb{R}) \text{y} \qquad K(u) > 0 \Leftrightarrow u \in (-1/2, 1/2).$$

Las funciones φ_k ya fueron definidas al conseguir la acotación para el caso puntual, cuando además se demostró que $\varphi_k \in \Sigma(\beta, L/2)$ para todo k. Además, denotando

$$\Delta_1 := [0, 1/m) \qquad \Delta_k := ((k-1)/m, k/m)] \text{ si } 1 < k \leq m, \tag{3.41}$$

las funciones ϕ_k serán nulas fuera de Δ_k.

Ahora, se necesitan construir las hipótesis f_{0k}, de manera que $d(f_{0k}, f_{0j}) \leq s$. Sea el conjunto de todas m-uplas binarias:

$$\Omega := \{\boldsymbol{\omega} = (\omega_1,\ldots,\omega_m),\ \omega_i \in \{0,1\}\} = \{0,1\}^m$$

Trabajamos con hipótesis tales que pertenezcan a $\mathcal{E} := \{f_{\boldsymbol{\omega}}(x) = \sum_{k=1}^{m} \omega_k \varphi_k(x), \boldsymbol{\omega} \in \Omega\}$. Se obtiene que, para $\boldsymbol{\omega}, \boldsymbol{\omega}' \in \Omega$

$$d(f_\omega, f_{\omega'}) = \left[\int_0^1 (f_{\boldsymbol{\omega}(x)} - f_{\boldsymbol{\omega}'}(x))^2 \, dx\right]^{1/2} \tag{3.42}$$

$$\underset{(3.41)}{=} \left[\sum_{k=1}^{m} (\omega_k - \omega_k')^2 \int_{\Delta_k} \varphi_k^2(x) \, dx\right]^{1/2} \tag{3.43}$$

$$= Lh_n^{\beta+1/2} \|K\|_2 \left[\sum_{k=1}^{m} (\omega_k - \omega_k')^2\right]^{1/2} \tag{3.44}$$

$$= Lh_n^{\beta+1/2} \|K\|_2 (\delta(\boldsymbol{\omega} - \boldsymbol{\omega}'))^{1/2} \tag{3.45}$$

denotando $\delta(\boldsymbol{\omega}, \boldsymbol{\omega}')$ a la distancia de Hamming, que se define como $\delta(\boldsymbol{\omega}, \boldsymbol{\omega}') = \sum_{k=1}^{m} I(\omega_k \neq \omega_k')$ (Puesto que $\omega_k \in \{0, 1\}$, $(\omega_k - \omega_k')^2 = I(\omega_k \neq \omega_k')$).

Si escogemos $\boldsymbol{\omega}, \boldsymbol{\omega}'$ tales que $\sqrt{\delta(\boldsymbol{\omega}, \boldsymbol{\omega}')} \asymp h_n^{-1/2} \Rightarrow \delta(\boldsymbol{\omega}, \boldsymbol{\omega}') \asymp m$, obtenemos que

$$\begin{aligned} d(f_{\boldsymbol{\omega}}, f_{\boldsymbol{\omega}'}) &= Lh_n^{\beta+1/2} \|K\|_2 (\delta(\omega_k - \omega_k'))^{1/2} = Lh_n^{\beta+1} \|K\|_2 = L\lceil c_0 n^{-\frac{\beta}{2\beta+1}} \rceil \|K\|_2 \\ &\asymp Lc_0 n^{-\frac{\beta}{2\beta+1}} \|K\|_2 = An^{-\frac{\beta}{2\beta+1}} = A\psi_n = s \end{aligned}$$

Ya tenemos una manera de construir f_{jn} para nuestro objetivo. Basta con obtener f_{jn} con $j = 0, 1, \ldots, M$ de manera que $d(f_{jn}, f_{kn}) \geq 2s_n$. Eso es equivalente a que las $\boldsymbol{\omega}$ asociadas a cada f_{jn} tengan una diferencia entre ellas de m. Sin embargo, ¿se pueden escoger en Ω suficientes tuplas $\boldsymbol{\omega}$ como para que M sea suficientemente elevado?

Un cota inferior de la cardinalidad de ese conjunto es dada por la **cota de Varshamov-Gilbert**, resultado de la teoría de la Información contenido en el Apéndice (Sección A.4).

Con ese lema, se obtiene que si $m \geq 8$, entonces existirán $\{\boldsymbol{\omega}^{(0)}, \ldots, \boldsymbol{\omega}^{(M)}\} \in \Omega$, siendo M cualquier entero tal que $M \geq 2^{m/8}$, de que

$$\delta(\boldsymbol{\omega}^{(k)}, \boldsymbol{\omega}^{(j)}) \geq \frac{m}{8} \qquad \forall\, 0 \leq j < k \leq M.$$

Es decir, en el espacio Ω se pueden empaquetar al menos $2^{m/8}$ bolas de radio $m/8$.

Observación 3.3.1. Para la construcción y definición de que estas hipótesis no ha necesitado de que el conjunto de entrenamiento sea determinista, lo que implica que más adelante para hacer una acotación más generalizada a través del Lema de Fano se podrán considerar estas mismas funciones hipótesis.

Aplicado a nuestro caso, esto implica que, escogiendo $f_{jn}(x) = f_{\boldsymbol{\omega}^{(j)}}(x)$ para cada $j = 0, \ldots, M = m/8$, y eligiendo los $\boldsymbol{\omega}^{(j)}$ en Ω como lo hecho en este teorema, se puede aplicar el Teorema Principal de las Cotas Inferiores (Teorema 2.5.7), obteniendo una acotación inferior de la probabilidad del riesgo minimax.

Proposición 3.3.3. Escogiendo $f_{jn}(x) = f_{\boldsymbol{\omega}^{(j)}}(x)$ para cada $j = 0, \ldots, M$, eligiendo los $\boldsymbol{\omega}^{(j)}$ en Ω como lo hecho en en el teorema anterior, para n suficientemente grande, se cumplen las hipótesis del Teorema 2.5.7, es decir

1. $f_{jn} \in \Sigma(\beta, L)$ para todo $j = 0, \ldots, M$.
2. $d(f_{jn}, f_{kn}) \geq 2s > 0$, para $j \neq k$.
3. $\frac{1}{M} \sum_{j=1}^{M} K(P_j, P_0) \leq \alpha \log M$

Demostración.

Primero, recordemos que la construcción que obtuvimos de los f_{jn} es

$$f_{jn} = f_{\boldsymbol{\omega}^{(j)}}(x) = \sum_{k=1}^{M} \omega_k^{(j)} \varphi_k(x),$$

con $\boldsymbol{\omega}^{(j)} = (\omega_1^{(j)}, \ldots, \omega_m^{(j)}) \in \Omega$, y $\varphi_k \in \Sigma(k, L/2)$ disjuntas.

1. $f_{jn} \in \Sigma(\beta, L)$ para todo $j = 0, \ldots, M$.

   Como $\varphi_k \in \Sigma(\beta, L/2)$, las componentes de los $\boldsymbol{\omega}$ están acotadas ($|\omega_i| \leq 1$), y φ_k tienen soporte disjunto, $f_{\boldsymbol{\omega}}(x) = \sum_{k=1}^{M} \omega_k \varphi_k(x)$ pertenece a $\Sigma(\beta, L)$ para todo $\boldsymbol{\omega} \in \Omega$, incluidos $(f_{jn})_{j=1}^{m}$.

2. $d(f_{jn}, f_{kn}) \geq 2s > 0$, para $j \neq k$.

   Por la cota de Varshamov-Gilbert (Teorema A.4.1), y que $d(f_\omega, f_{\omega'}) = Lh_n^{\beta+1/2}\|K\|_2(\delta(\omega_k - \omega_k'))^{1/2}$(3.42), tenemos que

$$\begin{aligned}
\|f_{jn} - f_{kn}\|_2 &= \|f_{\omega_m^{(j)}} - f_{\omega_m^{(k)}}\|_2 \\
&= Lh^{\beta+\frac{1}{2}}\|K\|_2\sqrt{\delta(\omega_m^{(j)}, \omega_m^{(k)})} \\
&\geq Lh^{\beta+\frac{1}{2}}\|K\|_2\sqrt{\frac{1}{16}} \\
&= \frac{L}{4}\|K\|_2 h_n^\beta = \frac{L}{4}\|K\|_2 m^{-\beta}
\end{aligned}$$

   cuando $m \geq 8$.

   Como se indicó al inicio de la construcción de las hipótesis en (3.39), $m = \lceil c_0 n^{\frac{1}{2\beta+1}} \rceil$. Esto implica que, escogiendo $n \geq n_0 = (7/c_0)^{2\beta+1}$, se obtiene que, si $m \geq 8$:

$$\begin{aligned}
m^\beta \leq (c_0 n^{\frac{1}{2\beta+1}} + 1)^\beta &\leq \left(c_0 n^{\frac{1}{2\beta+1}} + \left(\frac{n}{n_0}\right)^{\frac{1}{2\beta+1}}\right)^\beta \\
&= \left(c_0 n^{\frac{1}{2\beta+1}} + n^{\frac{1}{2\beta+1}}\left(\frac{c_0}{7}\right)^{\frac{2\beta+1}{2\beta+1}}\right)^\beta = \left(c_0 n^{\frac{1}{2\beta+1}}\left(1 + \frac{1}{7}\right)^\beta\right) \\
&= c_0^\beta n^{\frac{\beta}{2\beta+1}}\left(1 + \frac{1}{7}\right)^\beta \leq c_0^\beta n^{\frac{\beta}{2\beta+1}} 2^\beta
\end{aligned}$$

   Por tanto, considerando $A = \frac{L}{8}\|K\|_2(2c_o)^{-\beta}$, se obtiene:

$$\begin{aligned}
\|f_{jn} - f_{kn}\|_2 \geq \frac{L}{4}\|K\|_2 m^{-\beta} &\geq 2\frac{L}{8}\|K\|_2\left(\left(2c_0^{-\beta}\right) n^{-\frac{\beta}{2\beta+1}}\right) \\
&= 2An^{-\frac{\beta}{2\beta+1}} = 2A\psi_n = 2s
\end{aligned}$$

3. $\frac{1}{M}\sum_{j=1}^{M} K(P_j, P_0) \leq \alpha \log M$.

   Recordemos que P_j son las funciones de distribución de $Y_1, \ldots, Y_n$ para las hipótesis $f = f_{jn}$. Es por ello que la densidad conjunta se puede descomponer de la siguiente manera:

$$p(u_1, \ldots, u_n) = \prod_{i=1}^{n} p_\epsilon(u_i - f_{jn}(X_i)))$$

   Entonces, existe un entero n_0 tal que para todo $n > n_0$ se obtiene que nh_n y $Lh_n^\beta K_{\text{máx}} \leq v_0$ (por lo que depende de $c_0, L, \beta K_{\text{máx}}, v_0$).

   Mediante el mismo razonamiento que para demostrar que $K(P_0, P_1) < \alpha < \infty$ en 3.3.1, para $n > n_0$ se obtiene:

$$\begin{aligned} K(P_j, P_0) &= \int \log \frac{dP_0}{dP_j} dP_0 \\ &= \int \dots \int \log \prod_{i=1}^{n} \frac{p_\epsilon(u_i)}{p_\epsilon(u_i - f_{jn}(X_i))} \prod_{i=1}^{n} [p_\epsilon(u_i) du_i] \\ &= \sum_{i=1}^{n} \int \frac{p_\epsilon(y)}{p_\epsilon(y - f_{jn}(X_i))} p_\epsilon(y) dy \underset{(3.35)}{\leq} \frac{1}{2} \sum_{i=1}^{n} f_{jn}^2(X_i) \end{aligned}$$

Como esto se cumple para todo $j \in \{1, ..., m\}$, se puede extender a $M + 1$ hipótesis:

$$\begin{aligned} K(P_j, P_0) &\leq \frac{1}{2} \sum_{i=1}^{n} f_{jn}^2(X_i) \leq \frac{1}{2} \sum_{k=1}^{m} \sum_{i: X_i \in \Delta_K} \varphi_k^2(X_i) \\ &\leq \frac{1}{2} L^2 K_{\text{máx}}^2 h_n^{2\beta} \sum_{k=1}^{m} car(\{i : X_i \in \Delta_k\}) \\ &= \frac{1}{2} L^2 K_{\text{máx}}^2 n h_n^{2\beta} \leq \frac{1}{2} L^2 K_{\text{máx}}^2 c_0^{-(2\beta+1)} m. \end{aligned}$$

La cota de Varshamov-Gilbert (Teorema A.4.1) dicta que $M \geq 2^{m/8}$. Despejando m, se tiene que $m \leq 8 \log M / \log 2$.

Escogiendo $c_0 = \left(\frac{4L^2 K_{\text{máx}}^2}{\alpha \log 2} \right)^{\frac{1}{2\beta+1}}$:

$$\begin{aligned} K(P_j, P_0) &\leq \frac{1}{2} L^2 K_{\text{máx}}^2 c_0^{-(2\beta+1)} m \\ &= \frac{1}{2} L^2 K_{\text{máx}}^2 \left(\frac{4L^2 K_{\text{máx}}^2}{\alpha \log 2} \right)^{-1} m \\ &= \frac{\alpha \log 2}{8} m < \alpha \log M \end{aligned}$$

Esto implica que

$$\frac{1}{M} \sum_{j=1}^{M} K(P_j, P_0) \leq \frac{1}{M} \sum_{j=1}^{M} \alpha \log M = \alpha \log M.$$

Así que también se cumple esta hipótesis.

□

Esto nos lleva al siguiente resultado, directo de aplicar todo lo anterior:

Teorema 3.3.4. Sea $\beta > 0$ y $L > 0$. Bajo las condiciones C, se tiene que

$$\liminf_{n \to \infty} \inf_{T_n} \sup_{f \in \Sigma(\beta, L)} \mathbb{E}_f \left[n^{\frac{2\beta}{2\beta+1}} \|T_n - f\|_2^2 \right] \geq c$$

siendo $\inf_{T_n}$ el ínfimo entre todos los posibles estimadores, y c siendo una constante que depende solo de β y L.

Demostración.

Por la Proposición 3.3.3, escogiendo las funciones $\{f_{0n}, \dots, f_{Mn}\}$ como se indica en la demostración de la cota de Varshamov-Gilbert (Teorema A.4.1) cumplen las condiciones para poder aplicar el

Teorema Principal de las Cotas Inferiores del Riesgo (Teorema 2.5.7), obteniendo entonces que se cumple que

$$\inf_{T_n} \max_{f\in\{f_{0n},\dots,f_{Mn}\}} P_f(\|T_n - f\| \geq A\psi_n) \geq c' > 0. \tag{3.46}$$

De la misma manera, la Proposición 3.3.3 también afirma que $\{f_{0n}, \dots, f_{Mn}\}$ cumplen con las hipótesis del Teorema 2.3.1, pudiendo pasar la desigualdad al riesgo minimax, obteniendo

$$\liminf_{n\to\infty} \inf_{T_n} \sup_{f\in\Sigma(\beta,L)} \mathbb{E}_f \left[n^{\frac{2\beta}{2\beta+1}} \|T_n - f\|_2^2 \right] \geq c.$$

□

Observación 3.3.2. Se podría considerar una demostración alternativa, donde en vez de acotar por el Teorema Principal de las Cotas Inferiores del Riesgo (2.5.7), se acota a través del Lema de Fano. Puesto que más adelante se dará utilidad al Lema de Fano, este resultado se ha dado mediante el Teorema de las Cotas Inferiores.

Corolario 3.3.5. Sea un modelo de regresión no paramétrico que cumple las condiciones C.

Entonces, para todo valor de $\beta > 0$ y $L > 0$, se cumple que $\psi_n = n^{\frac{-\beta}{2\beta+1}}$ es la tasa de convergencia óptima en $(\Sigma(\beta, L), \|\cdot\|_2)$.

En concreto, se obtiene que para $l = \lfloor\beta\rfloor$, el estimador local polinomial $LP(l)$, con función núcleo K tal que existen $K_{min} > 0$, $\Delta > 0$ y $K_{max} < \infty$ constantes tales que:

$$K_{min} I(|u| \leq \Delta) \leq K(u) \leq K_{max} I(|u| \leq 1) \qquad \forall u \in \mathbb{R}$$

y escogiendo como ancho de banda h_n es de la forma $h_n = \alpha n^{-\frac{1}{2\beta+1}}$, es óptimo en tasa en $(\Sigma(\beta, L), \|\cdot\|_2)$.

Demostración.

Bajo las condiciones enunciadas en el corolario se puede aplicar el Teorema 3.1.7, que obtiene que el estimador $LP(l)$ con ancho de banda h_n cumple que

$$\liminf_{n\to\infty} \sup_{f\in\Sigma(\beta,L)} \mathbb{E}_f \left[n^{\frac{2\beta}{2\beta+1}} \|T_n - f\|_2^2 \right] \leq C,$$

lo que implica que

$$\liminf_{n\to\infty} \inf_{T_n} \sup_{f\in\Sigma(\beta,L)} \mathbb{E}_f \left[n^{\frac{2\beta}{2\beta+1}} \|T_n - f\|_2^2 \right] \leq C.$$

Juntado esto al resultado obtenido en el teorema anterior, se tiene que ψ_n es la tasa de convergencia óptima en $(\Sigma(\beta, L), \|\cdot\|_2)$.

De la misma manera, esto implica que el estimador $LP(l)$ con ancho de banda h_n tiene como tasa de convergencia $(\psi_n)_{i=1}^n$ y que sea un estimador óptimo en tasa en $(\Sigma(\beta, L), \|\cdot\|_2)$.

□

Con este último corolario ya ha encontrado finalmente unas condiciones donde no sólo se conoce la tasa de convergencia óptima dentro de las clases de Hölder, sino que que se conoce cómo calcular un estimador óptimo en tasa. Sin embargo, las condiciones del modelo que se ha estado estudiando son demasiados estrictas. Concretamente, la condición C1 delimita a considerar solo conjuntos de entrenamiento deterministas (es decir, poder elegir las muestras del conjunto de entrenamiento). Eso será el siguiente paso a seguir.

Mediante el mismo razonamiento, estos resultados se pueden obtener de igual manera para la norma del supremo[5]:

[5]La demostración es la misma que para la norma L^2, salvo utilizando los teoremas para la pérdida del supremo de la Sección 3.1.2 en vez de los de la Sección 3.1.1 para la pérdida cuadrática. Estos resultados pueden verse desarrollados en [3] p(108-110).

Teorema 3.3.6. Sea $\beta > 0$ y $L > 0$. Bajo las condiciones C y las condiciones LP, se obtiene que

$$\liminf_{n\to\infty} \inf_{T_n} \sup_{f\in\Sigma(\beta,L)} \mathbb{E}_f \|T_n - f\|_\infty^2 \geq c,$$

siendo $\inf_{T_n}$ el estimador óptimo en $\Sigma(\beta, L)$, siendo $c > 0$ una constante que únicamente depende de β y L.

Corolario 3.3.7. Se considera un modelo de regresión que cumple con las condiciones C Entonces la tasa óptima de convergencia para $(\Sigma(\beta, L), \|\cdot\|_\infty)$ es $\psi_n = \left(\frac{\log n}{n}\right)^{\frac{\beta}{2\beta+1}}$.

En más, el estimador local polinomial $LP(l)$, con $l = \lfloor\beta\rfloor$; escogiendo como ancho de banda $h_n = \alpha\left(\frac{\log n}{n}\right)^{\frac{1}{2\beta+1}}$ para $\alpha > 0$, y bajo las hipótesis del Teorema 3.1.9 sobre el núcleo K elegido, es un estimador óptimo en $(\Sigma(\beta, L), \|\cdot\|_\infty)$.

Es decir, trabajando bajo el un modelo de regresión que cumpla las condiciones C, los estimadores $LP(l)$ son óptimos. Este resultado no se puede extender a un modelo más general, ya que la cota superior obtenida para el riesgo de $LP(l)$ necesita de atributos deterministas. Sin embargo, a partir del Lema de Fano, se puede extender la cota inferior del riesgo minimax a un modelo más general.

3.4. Mejora del modelo a partir del Lema de Fano.

En la Sección 2.6, se acotó inferiormente la probabilidad de error media inferior $\overline{p}_{e,M}$. Esta acotación refina la acotación del Teorema Principal (Teorema 2.5.7). Sin embargo, en este caso nos interesa trabajar con $\overline{p}_{e,M}$ directamente, puesto que a partir de ella se puede extender dicha acotación a considerar un conjunto de entrenamiento aleatorio.

En concreto, se va a intentar extender la cota inferior obtenida con el modelo definido en 2.1.1 (bajo las condiciones C) a un modelo de regresión que cumpla las siguientes condiciones:

Condiciones 3 (Modelo generalizado)**.** Se dirá que un modelo de regresión cumple con las condiciones del Modelo Generalizado si

> $[C^*1]$Los atributos $(X_i)_{i=1}^n$ del conjunto de entrenamiento son **variables aleatorias arbitrarias en** $[0, 1]$ tal que el vector aleatorio $(X_1, \ldots, X_n)$ es independiente del vector aleatorio de los errores $(\epsilon_1, \ldots, \epsilon_n)$. . Existen $\epsilon_1, \ldots \epsilon_n$ variables aleatorias normales i.i.d con $\mathbb{E}(\epsilon_i) = 0$ tales que
>
> $$Y_i = f(x_i) + \epsilon_i. \tag{3.47}$$

La ventaja que supone trabajar con la probabilidad de error media es que trabaja con el promedio en vez de con el supremo como la probabilidad de error minimax.

Teorema 3.4.1. Sea $\beta > 0$ y $L > 0$, y sea $\psi_n = n^{-\frac{\beta}{2\beta+1}}$. Se supone que el modelo de regresión bajo el que se trabaja cumple con las siguientes condiciones:

Entonces, se cumple que

$$\liminf_{n\to\infty} \inf_{T_n} \sup_{f\in\Sigma(\beta,L)} \mathbb{E}_f \left[n^{\frac{2\beta}{2\beta+1}} \|T_n - f\|_2^2\right] \geq c.$$

Donde $\inf_{T_n}$ es el ínfimo entre todos los posibles estimadores en $\Sigma(\beta, L)$ y $c > 0$ es una constante que solo depende de β y L.

Demostración.

Se denotará con $\mathbb{E}_{\mathbf{X}} = \mathbb{E}_{(X_1,\ldots,X_n)}$ la esperanza de la distribución conjunta de $(X_1, \ldots, X_n) = \mathbf{X}$.

Sean $f_{0n}, \ldots, f_{Mn}$ de la misma manera que en la Proposición 3.3.3. Hay que recordar que en esa misma proposición se demostró que se cumple que $\|f_{jn} - f_{kn}\| > 2s$ para $j \neq k$, siendo $s = A\psi_n$.

Para cualquier estimador T_n, se cumple que:

Puesto que $P_f = \mathcal{L}((Y_1, \dots, Y_n)|f)$, se cumple que

$$\begin{aligned}
&\sup_{f\in\Sigma(\beta,l)} \mathbb{E}_f[\psi_n^{-2} \|T_n - f\|^2] \\
&\underset{(2.34)}{\geq} A^2 \max_{f\in\{f_{0n},\dots,f_{Mn}\}} P_f\left(\|T_n - f\| \geq A\psi_n\right) \\
&\geq A^2 \frac{1}{M+1} \sum_{j=0}^{M} P_j\left(\|T_n - f\| \geq A\psi_n\right) \\
&= A^2 \frac{1}{M+1} \sum_{j=0}^{M} \mathbb{E}_{\mathbf{X}}\left[P_j(\|T_n - f\| \geq s)\right].
\end{aligned}$$

Por C*1, $\boldsymbol{X}$ es independiente de los errores $(\epsilon_1, \dots, \epsilon_n)$, por lo que $\mathbb{P}(\|T_n - f\| \geq s) = P(\|T_n - f\| \geq s|\boldsymbol{X})$ para toda probabilidad P.

$$\begin{aligned}
&A^2 \frac{1}{M+1} \sum_{j=0}^{M} \mathbb{E}_{\mathbf{X}}\left[P_j(\|T_n - f\| \geq s)\right] \\
&A^2 \frac{1}{M+1} \sum_{j=0}^{M} \mathbb{E}_{\mathbf{X}}\left[P_j(\|T_n - f\| \geq s\,|\mathbf{X})\right] \\
&= A^2\, \mathbb{E}_{\mathbf{X}}\left[\frac{1}{M+1} \sum_{j=0}^{M} P_j(\|T_n - f\| \geq s\,|\mathbf{X})\right] \\
&\underset{(2.37)}{\geq} A^2\, \mathbb{E}_{\mathbf{X}}\left[\inf_{\Psi} \frac{1}{M+1} \sum_{j=0}^{M} P_j(\Psi \neq j\,|\mathbf{X})\right].
\end{aligned}$$

Fijamos el conjunto de entrenamiento (y por tanto $X_1, \dots, X_n$). De la Proposición 3.3.3, también se saca que las probabilidades $P_0, \dots, P_M$ cumplen que

$$\frac{1}{M+1} \sum_{j=0}^{M} K(P_j, P_0) \leq \alpha \log(M) \qquad \text{para } 0 < \alpha < 1/8 < 1,$$

lo que implica que se puede aplicar el Lema de Fano, obteniendo que

$$p_{e,M} \geq \overline{p}_{e,M} = \inf_{\Psi} \frac{1}{M+1} \sum_{j=0}^{M} P_j(\Psi \neq j|\mathbf{X}) \geq \frac{\log(M+1) - \log 2}{\log M} - \alpha.$$

Uniéndolo todo, se consigue la desigualdad desseada.

$$\begin{aligned}
\sup_{f\in\Sigma(\beta,l)} \mathbb{E}_f[\psi_n^{-2} \|T_n - f\|^2] &\geq A^2\, \mathbb{E}_{\mathbf{X}}\left[\inf_{\Psi} \frac{1}{M+1} \sum_{j=0}^{M} P_j(\Psi \neq j\,|\mathbf{X})\right] \\
&\geq A^2\, \mathbb{E}_{\mathbf{X}}\left[\frac{\log(M+1) - \log 2}{\log M} - \alpha\right] := c.
\end{aligned}$$

□

De esta manera, se ha podido obtener una acotación del riesgo minimax en el Modelo generalizado.

El papel que juega en este resultado el Lema de Fano es muy relevante, porque, al ser la acotación sobre la probabilidad de error media (un promedio) en vez de sobre la probabilidad de error minimax (un supremo), permite intercambiar el orden con $\mathbb{E}_{\boldsymbol{X}}[\cdot]$, algo que no sería posible con la probabilidad de error minimax, puesto que $\max(\mathbb{E}_{\boldsymbol{X}}[\cdot]) \not\geq \mathbb{E}_{\boldsymbol{X}}[\max(\cdot)]$.

Capítulo 4

Conclusiones

En el aprendizaje supervisado, se quiere encontrar una función estimación de la regla de Bayes. Para encontrar ese estimador, se suele trabajar dentro de una determinada clase, donde el error cometido dentro de ella sea pequeño. Lo deseado es que se pueda utilizar el estimador sobre una "clase grande". Por ello tiene gran de interés los estimadores no paramétricos, porque su riesgo se puede acotar dentro de una clase no paramétrica determinada.

El hecho de que se conozca cuál es la regla de Bayes en los problemas de regresión, hace que la obtención de funciones de estimación se pueda enfocar como un problema de aproximación de funciones. Esto es posible gracias a la funciones núcleo, que permiten ayudarse de observaciones suficientemente parecidas a la que con que se trabaja para hallar su valor. El estimador local polinomial es un estimador obtenido a partir de este enfoque, donde se aproxima la regla de Bayes por el polinomio local más parecido en un determinado ancho de banda.

Es interesante el estudio del estimador local polinomial, puesto que dentro de la clase no paramétrica de funciones $\Sigma(L, \beta)$ bajo el modelo de regresión sobre el que se ha trabajado se consigue una acotación del riesgo. Esta acotación es muy interesante, puesto que mejor cuanto mayor es el grado del estimador escogido, justificando su mayor coste computacional con una mejor aproximación. Este resultado es extremadamente útil sobre todo para la pérdida cuadrática, ya que en ese caso hipótesis consideradas a priori para trabajar correctamente con el estimador no acaban siendo necesarias.

La Teoría Minimax proporciona un contexto y unos fundamentos para poder comparar el estimador local polinomial con otros posibles estimadores. Sin embargo, es baste costoso obtener resultados directamente sobre ella. Sin embargo, cotas inferiores para el error pueden ser obtenidas, a partir del estudio de ciertos problemas de contraste, obteniendo como resultados más importantes el Teorema Principal de las Cotas Inferiores y el Lema de Fano, que permiten (de manera independiente) demostrar no solo que el riesgo del estimador local polinomial es bajo, sino que es un estimador óptimo en tasa trabajando en $\Sigma(L, \beta)$, lo que implica que su riesgo decrece según aumenta tamaño del conjunto de entrenamiento tan rápido como el más óptimo posible.

Esto, sumado a que se puede aumentar su eficiencia subiendo su grado (a coste de mayor tiempo de computación), y que se puede calcular su valor numéricamente (con ayuda de la Validación Cruzada), hace de los estimadores lineales una gran opción para trabajar con problemas de regresión bajo el modelo considerado.

Apéndice A

Nociones complementarias

A.1. Validación cruzada aplicada a optimizar el ancho de banda

En la Proposición 3.1.3, se obtuvo que el valor del ancho de banda h en el que se minimiza el riesgo de los estimadores locales polinomiales es

$$\hat{h}_n = \left(\frac{q_2}{2\beta q_1^2}\right)^{\frac{1}{2\beta+1}} n^{-\frac{1}{2\beta+1}},$$

siendo

$$q_1 = \frac{L\, 4K_{max}\, a_0}{\lambda_0 \cdot l!} \qquad q_2 = \sigma^2_{\text{máx}} C_1 C_2 = \sigma^2_{\text{máx}} \frac{2K_{max}}{\lambda_0} \frac{4K_{max}\, a_0}{\lambda_0} = \frac{8K^2_{max}\, \sigma^2_{\text{máx}}\, a_0}{\lambda_0^2}.$$

Puesto que λ_0 es una cota de los autovalores de la matriz $\mathcal{B}_{nx}$, así como la constante a_0 depende del conjunto de entrenamiento (se define al considerar la hipótesis LP2), haría falta un estudio exhaustivo del caso específico para considerar unos valores más o menos óptimos de q_1 y q_2.

En vez de ello, se puede tratar de minimizar el riesgo (que se denotará $R(h)$ para simplificar notación), como función de h: $R(h) = \mathbb{E}\left(\frac{1}{n}\sum_{i=1}^{n}(\hat{f}_n(X_i) - f(X_i))^2\right)$. Un método comúnmente utilizado es la validación cruzada.

El proceso de la validación cruzada consiste en, para cierto valor h_0, estimar el valor $R(h_0)$ utilizando el conjunto de entrenamiento a la vez tanto como de conjunto de entrenamiento y fabricar el estimador, como comprobante de cuán buen estimador se ha construido.

La validación cruzada se compone de los siguientes pasos:

1° Se separa la muestra en dos partes, una denominada **test** $T = X_{t_1}, .., X_{t_j}$ y la otra **de entrenamiento** $E = X_{e_1}, .., X_{e_k}$. Para que sea insesgado, la parte de entrenamiento y los test tienen que ser disjuntos.

2° Se calcula la estimación de los valores test considerando solo los valores de entrenamiento E. Denominaremos a esa función de estimación $\hat{f}^E$.

3° Se calcula el error cuadrático medio cometido, comparando para cada $i = 1...j$ la estimación $\hat{f}^E(X_{t_i})$ con el valor test Y_{t_i}, obteniendo así el riesgo de f^E:

$$\frac{1}{k}\sum_{i=1}^{j}(\hat{f}_n^E(X_{t_i}) - Y_{t_i}))^2$$

4° Se repite el proceso para diferentes conjuntos de entrenamiento y valores test [1]. Se estima el riesgo para el ancho de banda elegido por la media muestral de todos los errores cuadráticos obtenidos para dicho valor de h_0.

De esta manera, se puede repetir este proceso para los valores de h que deseemos, y escoger el menor como estimación del ancho de banda mínimo.

Ejemplo A.1.1. Partiendo del contexto de la Sección 3.2, se quiere obtener una valor del ancho de banda para realizar correctamente la estimaciones de la función $f(x) = \sin(x^2)$.

Para ello, se realiza el algoritmo de la validación cruzada ciertos valores de h. En concreto, para 200 valores de h entre 0 y 0.5; escogiedo como tamño del conjunto test 3.

El promedio (para cada división test-entrenamiento considerada) del riesgo empírico obtenido para cada uno de los estimadores forma las siguientes gráficas:

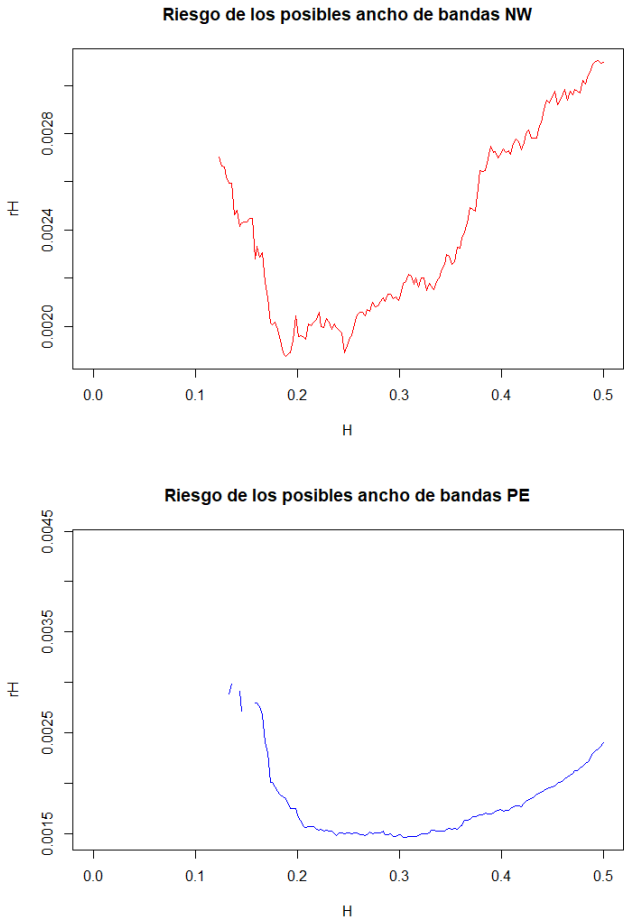


Figura A.1: Riesgos de $f(x) = \sin(x^2)$ para distintos anchos de banda

Se escogerá entonces como ancho de banda óptimo aquél donde el riesgo optenido sea menor, siendo distinto para cada estimador (0.1884422 para Nadaraya-Watson y 0.3040201 para el estimador local polinomial).

[1] Normalmente, los conjuntos de prueba y entrenamiento se construyen dividiendo la muestra en partes de j elementos, y cada una de ellas serán los conjuntos test. El caso más extremo es el **leave-one-out**, en el que se considera como conjuntos test cada uno de los elementos de la muestra de manera individual (es decir, el caso en el que se escoge $j = 1$). Dicho caso es el que más carga computacional posee, pero a la vez es el más preciso.

A.2. Derivada de Radon-Nykodym y descomposición de Jordan-Hahn

Tanto el concepto de la derivada de Radon-Nikodym como la descomposición de Jordan-Hahn son resultados de la Teoría de la Medida que se utilizan en las demostraciones del documento, específicamente en el manejo de la probabilidad del error minimax.

Definición A.2.1. Sean μ, ν dos medidas positivas sobre el mismo espacio medible. Se dice que la medida μ es **absolutamente continua** respecto de ν cuando, para cada conjunto medible A se tiene que $\nu(A) = 0$ implica que $\mu(A) = 0$. Se escribe $\mu \ll \nu$.

Un concepto en antagónico a éste es el de la relación de singularidad:

Definición A.2.2. Sean μ, ν dos medidas positivas sobre el mismo espacio medible. Se dice que la medida μ es **singular** respecto de ν cuando existen dos conjuntos disjuntos A_ν y A_μ tales que $\mu(A_\mu^c) = 0$ y $\nu(A_\nu^c) = 0$.

La derivada de Radon-Nikodym nace de la idea de extender el concepto de derivada a los espacios medibles σ-finitos, cumpliendo un papel similar a la derivada ordinaria.

Teorema A.2.1 (Derivada de Radon-Nikodym)**.** Sean μ, ν dos medidas sobre el mismo espacio medible, tal que μ sea finita y ν positiva y σ-finita, y tales que $\mu \ll \nu$. Entonces, existe una función real f tal que, para todo conjunto $A \in \Sigma$.

$$\mu(A) = \int_A f d\nu. \tag{A.1}$$

Además la función f es única casi seguro (todas las funciones que lo cumplen solo difieren de f en un conjunto de ν-medida nula).

(Billingsley [1], p423.)

A f se le denomina **derivada de Radon-Nikodym** de ν respecto de μ, y se suele denotar por

$$f = \frac{d\mu}{d\nu}. \tag{A.2}$$

Una de las implicaciones de la definición es que, para cualquier función g medible, se obtiene que $\int g\, d\nu$ es integrable si y solo sí $\int g \frac{d\mu}{d\nu} d\mu$ lo es. Además, en caso positivo las integrales con iguales:

$$\int g\, d\mu = \int g \frac{d\mu}{d\nu}\, d\nu. \tag{A.3}$$

A partir de esta propiedad se pueden deducir varias propiedades de la derivada de Radon-Nikodym, como por ejemplo la contraparte de la regla de la cadena de la derivada habitual:

Proposición A.2.2 (Regla de la cadena)**.** Sean ν, μ y ρ tres medidas σ- medibles, tales que $\mu \ll \nu$ y $\nu \ll \rho$. Entonces, $\mu \ll \rho$ y

$$\frac{d\mu}{d\rho} = \frac{d\mu}{d\nu} \frac{d\nu}{d\rho}.$$

Demostración.

Basta con aplicar la propiedad anterior para $g = \frac{d\mu}{d\nu}$. Esto es posible ya que por la definición de la derivada de Radon-Nikodym (Teorema A.2.1) , g es medible:

$$\int_A \frac{d\mu}{d\rho}\, d\rho = \mu(A) = \int_A \frac{d\mu}{d\nu}\, d\nu = \int_A \frac{d\mu}{d\nu} \frac{d\nu}{d\rho}\, d\rho.$$

□

Por otro lado, la descomposición de Jordan-Hahn intenta abordar la relación entre μ y ν cuando μ no sea absolutamente continua respecto de ν. Para ello, primero es necesario presentar el Teorema de Hahn:

Teorema A.2.3 (Teorema de Hahn)**.** Sea un espacio medible (Ω, Σ), y μ una medida positiva en él. Entonces existe $P \in \Sigma$ tal que para todo $A \in \Sigma$ tal que $A \subset P$ se cumple que $\mu(A) \geq 0$, y para todo $A \subset N = \Omega \setminus P$ se cumple que $\mu(A) \leq 0$.

Al par (P, N) se le denomina **descomposición de Hahn**

Demostración.

Bilingsley[1], p.420-421 □

Teorema A.2.4 (Descomposición de Jordan-Hahn)**.** Sea un espacio medible (Ω, Σ), y μ una medida positiva en él. Entonces existe una única descomposición $\nu = \nu^+ + \nu^-$ tal que para toda descomposición de Hahn (P,N) ν cumple que $\nu^+(A) = 0$ si $A \subset N$ y $\nu^-(A) = 0$ si $A \subset P$.

Demostración.

Fischer, Tom (2012) - Existence, uniqueness, and minimality of the Jordan measure decomposition.

arXiv:1206.5449 [math.ST] □

Considerando la definición de la continuidad absoluta, para una medida ν positiva y σ-finita cualquiera se puede considerar P el subconjunto de Ω donde $\nu(A) = 0$. Por la descomposición de Jordan-Hahn, en P también se da que $\mu^-(A) = 0$, luego $\mu^- \ll \nu$. Mientras, por la misma razón, μ^+ es singular respecto a ν (considerando $A_\nu = N$ y $A_{\mu^+} = P$).

En resumen, dadas una medida ν positiva y σ-finita y μ una medida finita, μ se puede descomponer en dos medidas μ^s y μ^a, donde μ^s es singular respecto a ν y μ^a absolutamente continua respecto a ν.

Puesto que las probabilidades son medidas positivas y finitas, tanto la descomposición de Jordan-Hahn como la derivada de Radon-Nikodym se pueden extrapolar a probabilidades, como sucede en este documento.

A.3. Construcción de la función K

Proposición A.3.1. Existe una función K que cumple:

$$K \in \Sigma(\beta, 1/2) \cap C^{\infty}(\mathbb{R}) \qquad y \qquad K(u) > 0 \Leftrightarrow u \in (-1/2, 1/2).$$

Demostración.

Veamos que la función

$$K(u) = a \exp\left(-\frac{1}{1-4u^2}\right) I(|u| < 1/2)$$

cumple ambas propiedades de (3.33).
Primero, podemos tratar con

$$K_0(u) = \exp\left(-\frac{1}{1-u^2}\right) I(|u| \leq 1),$$

ya que $K(u) = aK_0(2u)$.

- $K_0 \in \mathcal{C}^{\infty}((-1,1))$, puesto que es combinación de funciones que lo son (en $(-1,1)$). Además, las derivadas de K_0 son de la forma $K_0^{(l)}(x) = \dfrac{P(x)}{Q(x-1)} K_0(x)$, lo que implica que se cumple que

  $$\lim_{|x|\to 1} K_0^{(l)}(x) = \lim_{|x|\to 1} \frac{P(x)K_0(x)}{Q(x-1)} = 0.$$

  Como por construcción $K_0(u) = 0$ cuando $|x| \geq 1$, implica que $K_0 \in \mathcal{C}^{\infty}(\mathbb{R})$.

- $K_0^{(i)}$ son funciones continuas en el intervalo compacto $[-1,1]$, luego son acotadas en él. Como fuera de ese intervalo son nulas, son acotadas en todo $\mathbb{R}$. Al ser acotadas, $K_0^{(l)}$ son lipschitzianas con constante de Lipschitz $L = 1$ para todo l, lo que implica que $K^{(l)}$ son lipschitzianas con constante de Lipschitz $L = 1/2$ para todo l, por lo que $K \in \Sigma(\beta, 1/2)$ para cualquier $\beta > 0$.

- $K_0(u) > 0$ para todo $u \in (-1,1)$, ya que e^x es positiva para todo $x \in \mathbb{R}$. Además, fuera de dicho intervalo es nula, así que también se cumple la implicación recíproca.

□

A.4. Cota de Varshamov-Gilbert

La cota de Varshamov-Gilbert es un resultado de la teoría de la Información que acota el **orden de empaquetamiento** de un conjunto de tuplas (palabras). En nuestro caso, las "palabras" son las tuplas $\boldsymbol{\omega}^{(j)}$ que definen las hipótesis f_{0j} que se tendrán que considerar para obtener una buena acotación de la probabilidad del error minimax.

Esta cota es un resultado muy interesante, además de tedioso y complejo. Es por ello que solo se mostrará el esqueleto de la demostración.

Teorema A.4.1. Sea $m \geq 8$. Entonces existen $\boldsymbol{\omega}^{(0)}, \ldots, \boldsymbol{\omega}^{(M)} \in \Omega$, con $\boldsymbol{\omega}^{(0)} = (0, \ldots, 0)$, tales que

$$\delta(\boldsymbol{\omega}^{(k)}, \boldsymbol{\omega}^{(j)}) \geq \frac{m}{8} \qquad\qquad \forall\, 0 \leq j < k \leq M$$

Además, se cumple que M, el número de esos elementos, se pueden elegir de manera que

$$M \geq 2^{\frac{m}{8}}$$

Demostración.

Obviamente, $\|\Omega\| = 2^m$.

Para construir las $M+1$ m-uplas necesarias comenzamos escogiendo $\boldsymbol{\omega}^{(0)} = (0, \ldots, 0)$, y consideramos $D = \lfloor m/8 \rfloor$. Excluimos todos los $\boldsymbol{\omega} \in \Omega$ que se encuentren en el entorno de radio D de $\boldsymbol{\omega}^{(0)}$ (es decir, aquellos puntos tales que $\delta(\boldsymbol{\omega}^{(0)}, \boldsymbol{\omega}) \leq D$).

Se define

$$\Omega_1 := \left\{ \boldsymbol{\omega} \in \Omega : \delta(\boldsymbol{\omega}, \boldsymbol{\omega}^{(0)}) > D \right\}$$

y consideramos un elemento cualquiera de Ω_1 al que denotaremos $\boldsymbol{\omega}^{(1)}$.

De esta manera, podemos definir, de manera recursiva el conjunto Ω_2 (volviendo a excluir todos los $\boldsymbol{\omega}$ pertenecientes al entorno de $\boldsymbol{\omega}^{(1)}$) y elegimos un $\boldsymbol{\omega}^{(2)}$ entre los sobrantes, quedándonos con los subconjuntos

$$\Omega_j := \left\{ \boldsymbol{\omega} \in \Omega_{j-1} : \delta(\boldsymbol{\omega}, \boldsymbol{\omega}^{(j-1)}) > D \right\}$$

y eligiendo $\boldsymbol{\omega}^{(j-1)} \in \Omega_j$ cualquiera para cada iteración.

Obsérvese que se cumple que $\Omega_j \subset \Omega_{j-1}$, y que para todo $j \neq k$ se da que $\delta(\boldsymbol{\omega}^{(j)}, \boldsymbol{\omega}^{(k)}) \geq D + 1 = \lfloor m/8 \rfloor + 1 \geq m/8$.

Sea M el menor entero en para el que se cumpla que $\Omega_{M+1} = \emptyset$. Se definen los subconjuntos A_j de Ω:

$$A_j := \{ \boldsymbol{\omega} \in \Omega_j : \delta(\boldsymbol{\omega}, \boldsymbol{\omega}^{(j)}) \leq D \} = \Omega_j \setminus \Omega_{j+1}$$

es decir, los elementos descartados en la interacción j para obtener el subconjunto Ω_{j+1}, y $n_j = \|A_j\| = Car(A_j)$.

Para tener una distancia de Hamming D se debe tener al menos D componentes distintas. Puesto que en Ω hay $\sum_{i=0}^{D} \binom{m}{i}$ elementos con D o menos elementos con componentes distintas a un elemento dado $\boldsymbol{\omega} \in \Omega$, se tiene que:

$$Car(A_j) = n_j \leq \sum_{i=0}^{D} \binom{m}{i} \tag{A.4}$$

para todo $j = 1, .., M$.

Esto, sumado al hecho de que $A_0, A_1, \ldots, A_M$ conforman una partición disjunta de Ω, obtenemos que:

$$2^m = \|\Omega\| = n_0 + n_1 + \ldots + n_m \leq (M+1)\sum_{i=0}^{D}\binom{m}{i}$$

Denotando

$$p^* := \mathbb{P}(Bin(m,1/2) \leq \lfloor m/8 \rfloor)) = \sum_{i=0}^{D} 2^{-m}\binom{m}{i},$$

se obtiene que, despejando de la desigualdad anterior,

$$M+1 \geq \frac{1}{p^*}.$$

Para $X_1, .., X_n$ i.i.d y a_i, b_i, t tales que $a_i \leq X_i \leq b_i$ y $t > 0$, se tiene la siguiente desigualdad, denominada **desigualdad de Hoeffding**[3, A.4]:

$$\mathbb{P}\left(\sum_{i=1}^{n}(X_i - \mathbb{E}(X_i)) \geq t\right) \leq \exp\left(-\frac{2t^2}{\sum_{i=1}^{n}(b_i - a_i)^2}\right).$$

Como ya sabemos, $Bin(m,1/2) = \sum_{i=1}^{m} Z_i$, siendo Z_i Bernuilli i.i.d de parámetro $1/2$. Considerando $a_i = 0$ y $b_i = 1$ para todo i, se puede aplicar la desigualdad la Hoeffding a las variables Z_i:

$$\begin{aligned}
&\mathbb{P}\left(\sum_{i=1}^{m}(Z_i - \mathbb{E}(Z_i)) \geq t\right) \leq \exp\left(-\frac{2t^2}{\sum_{i=1}^{m}(1)^2}\right) \\
\Rightarrow &\mathbb{P}\left(\sum_{i=1}^{m}(Z_i - 1/2) \geq t\right) \leq \exp\left(-\frac{2t^2}{m}\right).
\end{aligned}$$

Aplicándolo a p^*:

$$\begin{aligned}
\mathbb{P}(Bin(m,1/2) \leq \lfloor m/8 \rfloor)) &= \mathbb{P}\left(\sum_{i=1}^{m} Z_i \leq \lfloor m/8 \rfloor\right) \\
&\leq \mathbb{P}\left(\sum_{i=1}^{m} Z_i \leq m/8\right) \\
&= \mathbb{P}\left(\sum_{i=1}^{m} Z_i \geq 7m/8\right) && (\ Z_i \text{ son simétricas respecto a } 1/2) \\
&= \mathbb{P}\left(\sum_{i=1}^{m}(Z_i - 1/2) \geq \frac{7m}{8} - \frac{m}{2}\right) \\
&= \mathbb{P}\left(\sum_{i=1}^{m}(Z_i - 1/2) \geq \frac{3m}{8}\right) \\
&\leq \exp\left(-\frac{2\left(\frac{3m}{8}\right)^2}{m}\right) = \exp\left(-\frac{9m}{32}\right)
\end{aligned}$$

Luego, obtenemos que

$$\begin{aligned}
&p^* \leq \exp(-9m/32) < 2^{-m/4} \\
&\Rightarrow M+1 \geq 2^{m/4} \geq 2^{m/8} + 1 \\
&\Rightarrow M \geq 2^{m/8}
\end{aligned}$$

para todo $m > 8$, obteniendo lo que queríamos demostrar.

□

Bibliografía

[1] Patrick Billingsley. *Probability and Measure*. en. 3.ª ed. Wiley Series in Probability & Mathematical Statistics: Probability & Mathematical Statistics. Nashville, TN: John Wiley & Sons, mayo de 1995.

[2] J. M. Sanz-Serna. *Diez Lecciones de Calculo Numerico*. Universidad de Valladolid, Secretariado de Publicaciones e Intercambio Editorial, 2010.

[3] Alexandre B Tsybakov. *Introduction to Nonparametric Estimation*. en. Springer Series in Statistics. New York, NY: Springer, feb. de 2010.

[4] Larry Wasserman. *All of Nonparametric Statistics*. 1.ª ed. Springer Texts in Statistics. New York, NY: Springer, mayo de 2007.

[5] Larry Wasserman. *All of statistics*. en. 1.ª ed. Springer Texts in Statistics. New York, NY: Springer, sep. de 2004.