0.1. Convergencia del modelo

Habiendo entendido la estructura de las funciones que componen este proyecto, se busca demostrar (o dar una idea) por que se tiene una aproximación a f y a f_i's y no una igualdad escrita. Esto se logra, usando principios de álgebra lineal y análisis funcional. Esta discusión sigue los principios planteados por (**bergstrom1985estimation**) en donde se demuestra el caso univariado y una pequeña discusión sobre los *Espacios de Hilbert de Kernel Reproductivo* (RKHS)[1]

Para entender el teorema de convergencia, necesitamos considerar los Espacios de Hilbert. Como introducción a estos, se usa el ejemplo clásico de plantearlo como una generalización del caso euclidiano. En este espacio euclidiano normal $\mathbb{R}^n$ podemos representar cualquier vector como una combinación lineal de un conjunto de bases ortogonales. En espacios abstractos de dimensión infinita, en particular en espacios de funciones, se busca representar *cualquier función*, en este caso f_i's, como una combinación lineal de bases. Los espacios de Hilbert dan idea de que los espacios vectoriales pueden ser suficientemente abstractos para que los vectores no sean simplemente listas ordenadas de números como lo son en $\mathbb{R}^n$. Los vectores pueden ser cualquier objeto en este caso, funciones. Una vez definido el espacio vectorial y sus objetos (que cumplan los correspondientes 8 axiomas presentados en el Apéndice **??**) se les puede denota de *producto interno* y por consecuente una *métrica* la cual induce una topología.

1. Reproducing Kernel Hilbert Space a falta de una mejor traducción.

Formalización matemática y teorema de convergencia

Se dice que $\mathcal{H}$ es un Espacio de Hilbert si $\mathcal{H}$ es un espacio vectorial con producto interno que también es un espacio metrico completo. (**rudin1987real**).

Para hacer la prueba de convergencia, se considera únicamente a las funciones f_i y no a la f general. Se estudia en particular el espacio de Hilbert $\mathcal{H} = \mathbf{L}_2(\mathbf{R}, \mu)$ el *espacio de funciones integrables al cuadrado en* $\mathbf{R} = [a, b]$ con medida de Lesbegue ordinaria μ. Es decir:

$$f \in \mathcal{H} \iff \int_a^b f(x)^2 \, dx < \infty$$

Donde el producto punto es:

$$\langle f, g \rangle = \int_a^b f(x) g(x) \, dx$$

y su norma inducida:

$$||f||_{\mathcal{H}} = \langle f, f \rangle = \int_a^b f(x)^2 \, dx$$

Antes de que se presente el teorema de Bergstrom, se tienen que presentar los tres

supuestos fuertes que hace:

1. Las variables aleatorias $y_1, \ldots, y_n$ son generadas por la ecuación:

$$y_i = h(x_i) + e_i \quad \forall i = 1, \ldots, n$$

y los valores de las covariables están dados por las ecuaciones:

$$\begin{aligned} x_1 &= a + \frac{b-a}{2n} \\ x_{i+1} &= x_i + \frac{b-a}{n} \end{aligned} \tag{1}$$

Con a, b los extremos del intervalo $b > a$ y e_i son ruido aleatorio cumpliendo:

$$\begin{aligned} \mathbb{E}[e_i] &= 0, \quad \forall i = 1, \ldots, n \\ \mathbb{E}[e_i^2] &= \sigma^2, \quad \forall i = 1, \ldots, n \\ \mathbb{E}[e_i e_j] &= 0, \quad \forall i, j = 1, \ldots, n \; i \neq j \end{aligned} \tag{2}$$

2. La función $h(x)$, está definida en el intervalo cerrado $[a, b]$ es acotada y continua en casi todas partes.

3. El conjunto contables de funciones base $\Psi_1(x), \Psi_2(x), \ldots$ es un conjunto *ortonormal* en $\mathcal{H}$. Es decir, estas funciones cumplen:

$$\int_a^b \Psi_j^2(x)\, dx = 1 \quad \forall j \tag{3}$$
$$\int_a^b \Psi_j(x)\Psi_k(x)\, dx = 0 \quad \forall k, j \; k \neq j$$

Estos supuestos son bastante fuertes y hay ciertas ecuaciones que no se cumplen *per se* en el modelo propuesto, sin embargo, vale la pena analizar el resultado pues lleva a cosas aún más interesantes. El primer supuesto es el más problemático. Aunque el modelo generador es idéntico a (**??**) y el ruido aleatorio es un supuesto aceptable (y común) el problema está en (**??**) pues para este trabajo no se asume que el estadista fija las x's sino que se asume una muestra aleatoria de datos. Sin embargo, en la prueba, este supuesto se usa para argumentar que, si $n \to \infty$, los datos cubren de manera homogénea todo el intervalo aproximando una integral. Aunque el propósito es completamente diferente que el de este trabajo en el que se busca suavizar sobre datos dispersos, se decide obviar por ahora el supuesto, en interés de presentar el teorema en su forma más rigurosa.

El segundo supuesto no es nada descabellado y se ha usado con anterioridad. Además,

aún permite aproximar un número grande de funciones y es igual de flexible que el modelo anterior. Sin embargo, este supuesto si implica que $h \in \mathcal{H}$. Por lo que esta puede ser representada en su combinación lineal de bases funcionales, es decir:

$$h(x) = \sum_{i=1}^{\infty} w_i \Psi_i(x)$$

diferente a la expansión de bases en de la ecuación (**??**). Esto se deriva, de que ahora se busca encontrar una representación *exacta* de h.[2] El último supuesto, implica la construcción de una base *ortonormal*. El trabajo original, sugiere que se puede lograr una base completa, aplicando un proceso de ortonormalización a las bases canónica polinomial $\{1, x, x^2, \ldots\}$. Por lo tanto, al menos de forma teórica, la base escogida para este trabajo definida en (**??**) también es ortonormalizable independientemente de la elección de J, N y K. Por lo que se cumple el supuesto. Finalmente:

Teorema 0.1. *Sea* $\hat{h}_n^{N^*}(x)$ *el estimador de* h *definido por:*

$$\hat{h}_n^{N^*}(x) = \hat{w}_1 \Psi_1(x) + \ldots + \hat{w}_{N^*} \Psi_{N^*}(x) \tag{4}$$

que depende del número de datos n *y el número de bases funcionales* N^**. Y, con* $\hat{w}_i \quad i = 1, \ldots, N^*$ *los estimadores de mínimos cuadrados, es decir, los valores de* $w_i \quad i = 1, \ldots, N^*$ *que minimizan la expresión:*

2. Antes se buscaba, más que aproximarla, suavizar los datos. Además, se usa el signo de igualdad para no introducir confusión en la exposición.

$$\sum_{i=1}^{n} [y_i - w_1\Psi_1(x_i) - \ldots - w_{N^*}\Psi_{N^*}(x_i)]^2 \tag{5}$$

Para todo $\epsilon > 0$, bajo los supuestos 1 a 3, existe un entero N^* y una función $n_\epsilon(N^*)$ tal que:

$$\mathbb{E}\left[\int_a^b \left(\hat{h}_n^{N^*}(x) - h(x)\right)^2 dx\right] < \epsilon \qquad \forall N \geq N^* \ y \ \forall n \geq n_\epsilon(N^*) \tag{6}$$

Detrás de toda esta verborrea y notación aparatosa, el corazón del teorema está en que, bajo ciertos supuestos, rigurosos más no descabellados, y en caso de existir una función h que genere los datos, está se puede aproximar a un grado de precisión arbitraria.

Aunque no es el objetivo del trabajo, vale la pena hacer una mención a lo sublime que es la demostración, pues utiliza conceptos de análisis, álgebra lineal, optimización y estadística. Los detalles y la utilización de los supuestos, son sutiles, sin embargo es una prueba rigurosa en todo el sentido de la palabra. Además, (**bergstrom1985estimation**) va mucho más allá de únicamente demostrar la existencia. Se demuestran tres teoremas más, para dar estimaciones (bajo un supuesto adicional) de el tamaño de muestra necesario n y el número de bases N^* necesa-

rias para la aproximación arbitraria de h. Sin embargo, este procedimiento, aunque elegante, no es nada practico pues depende de poder generar a merced las x's (con su correspondiente nivel y) aumentando y disminuyendo el tamaño de muestra. Además, se requiere ir generando todas las bases Ψ's de forma que sean ortonormales y sus correspondientes coeficientes de Fourier $w_j = \int_a^b h(x)\Psi_j(x)\,dx \quad \forall j$. El resultado, es más bien teórico en el sentido de que, justifica que estos modelos tienen sentido. Para este trabajo, en especifico, da la *intuición* de que funcionará (obviando un poco el primer supuesto) pues, con una muestra suficientemente grande, las f_i's serán identificables y aunque sean aproximaciónes, estaremos captando los patrones subyacentes y con suerte, podremos hacer predicciones.

Se hace notar que el primer supuesto, da la intuición de *granularidad* de el intervalo $[a, b]$. Bajo la construcción de Bergstrom, tenemos las condiciones exactas para estimar $h(x)$. Sin embargo, en la practica es raro que el estadista tenga el control sobre las covariables y siempre tendremos datos aleatorios. A pesar de esto, al estar trabajando sobre intervalos cerrados, se puede suponer que $X \sim \mathrm{U}(a, b)$ de donde tenemos que si $n \to \infty$ cubrimos todo el intervalo y tenemos algo análogo al supuesto 1 y por ende, el teorema es válido.

Otro teorema de convergencia y RHKS

Este resultado de Bergstrom, es uno de los muchos teoremas que se probaron en la época para justificar la existencia de estos modelos. Otro resultado interesante del mismo año, viene dado por (**stone1985additive**). El, discute varios mode-

los posibles dada una estructura de datos practica. Plantea un GAM[3] tradicional $h^*(\mathbf{x}) = \sum h_i^*(x_i)$ con la restricción (**??**), y prueba que

$$\mathbb{E}[(h(x) - h^*(x))^2]$$

es mínimo, con la igualdad si h es en verdad aditiva. Además, los estimadores de h_j^* que da, son splines (los mismos que usan Hastie y Tibshirani).

Finalmente, se hace notar, que toda esta teoría y resultados, son casos específicos de una teoría más general y mucho más compleja, llamada *Espacios de Hilbert de Kernel Reproductivo* (RKHS) desarrollada en los años 90. Esta va mucho más allá del enfoque de este trabajo, sin embargo, vale la pena mencionarla pues engloba muchos de los modelos usados hoy en día en un marco matemático riguroso y basado en el análisis funcional. Muchas de las ideas presentadas en este trabajo, como lo son la regularización, las expansiones de bases y los espacios de Hilbert, se elevan varios niveles y llevan a resultados todavía más generales. Se recomienda el Capítulo 5.8 de (**hastie2008elements**) y el libro de (**wahba1990splines**) donde se discuten a detalle todas las consideraciones de los RKHS. Además, todos estos resultados son *deterministas* en sus parámetros, en contrapuesta de la filosofía bayesiana. Sin embargo, essto no le quita validez al modelo ni mucho menos a los resultados.

3. T. Hastie y R. Tibshirani publicaron preámbulos a los GAM antes del trabajo citado aquí de 1986. Además, de que la cercanía, Stone en Berkley y ellos en Stanford, ayudó a su colaboración.