UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL
INSTITUTO DE MATEMÁTICA
PROGRAMA DE PÓS-GRADUAÇÃO EM MATEMÁTICA APLICADA

Aspectos Matemáticos do Problema de Aprendizagem em Inteligência Artificial

por

Jean Carlo Pech de Moraes

Dissertação submetida como requisito parcial
para a obtenção do grau de

Mestre em Matemática Aplicada

Prof. Dr. José Afonso Barrionuevo
Orientador

Porto Alegre, maio de 2006.

CIP - CATALOGAÇÃO NA PUBLICAÇÃO

Moraes, Jean Carlo Pech de

Aspectos Matemáticos do Problema de Aprendizagem em Inteligência Artificial / Jean Carlo Pech de Moraes.—Porto Alegre: PPGMAp da UFRGS, 2006.

56 p.: il.

Dissertação (mestrado) —Universidade Federal do Rio Grande do Sul, Programa de Pós-Graduação em Matemática Aplicada, Porto Alegre, 2006.
Orientador: Barrionuevo, José Afonso

Dissertação: Análise Aplicada
Inteligência Artificial, Teoria de Aproximações

Aspectos Matemáticos do Problema de Aprendizagem em Inteligência Artificial

por

Jean Carlo Pech de Moraes

Dissertação submetida ao Programa de Pós-Graduação em Matemática Aplicada do Instituto de Matemática da Universidade Federal do Rio Grande do Sul, como requisito parcial para a obtenção do grau de

Mestre em Matemática Aplicada

Linha de Pesquisa: Análise Aplicada

Orientador: Prof. Dr. José Afonso Barrionuevo

Banca examinadora:

Prof. Dr. Dalcídio M. Claudio
PUC-RS

Profª. Drª. Manuela Longoni de Castro
PPGMAp - UFRGS

Prof. Dr. Eduardo Brietzke
PPGMAT-UFRGS

Dissertação apresentada e aprovada em
25 de maio de 2006.

Profª. Drª. Maria Cristina Varriale
Coordenadora

Ao meu amor,

Gabriela

AGRADECIMENTOS

Aos meus pais pelo incentivo para seguir em frente e pela compreensão nos momentos em que estive ausente ou ocupado demais para dar-lhes atenção.

Ao Conselho Nacional de Desenvolvimento Científico e Tecnológico pelo apoio financeiro, ao Programa de Pós-Graduação em Matemática Aplicada (PPGMAp) por me oferecer a oportunidade de realizar este mestrado. À coordenadora Maria Cristina e ao ex-coordenador Vilmar Trevisan que sempre se mostraram dispostos à ajudar os alunos do PPGMAp.

Ao meu orientador, José Afonso, por direcionar minha pesquisa e auxiliar na execução e nas soluções de dúvidas que surgiram ao longo do trabalho.

Aos professores do PPGMAp, PPGMAT e PPGE, que de alguma maneira contribuiram nesta minha caminhada, em especial aos professores que ministraram as disciplinas que cursei durante estes dois anos.

Ao professor Eduardo Brietzke por sempre estar disposto a ajudar-me e aconselhar-me.

À minha noiva Gabriela que sempre me incentivou, apoiou e ouviu com paciência quando enfrentei os problemas que tive na elaboração deste trabalho.

Conteúdo

LISTA DE SÍMBOLOS

$\lceil \cdot \rceil$	maior inteiro
$\| \cdot \|_K$	norma em H_k
$C(X)$	Espaço das funções contínuas em X
$C^\infty(X)$	Espaço das funções infinitamente diferenciáveis em X
C_K	Supremo (para x e $t \in X$) de $K(x,t)$
$D_{\mu\rho}$	Distorção de μ com respeito a ρ
ε	erro de aproximação
ε_γ	erro regularizado
$\varepsilon_{\gamma,z}$	erro empírico regularizado
$\varepsilon_{I\!H}$	erro em $I\!H$
$\varepsilon_{I\!H,z}$	erro amostral
ε_z	erro empírico
f_γ	função alvo, em relação a ε_γ, em $I\!H$
$f_{I\!H}$	função alvo em $I\!H$
f_z	função alvo empírica
$I\!H_K$	RKHS
$H_s(X)$	O completamento de $C^\infty(X)$ com respeito a norma $\|\|\|_s$
$I\!L^2_\nu(X)$	Espaço das funções de quadrado integrável na medida ν
L_z	função defeito
ρ	medida de probabilidade
ρ_X	medida de probailidade marginal
$\rho_{y/x}$	medida de probailidade condicional
σ^2_ρ	limite inferior do erro
X	Domínio compacto ou variedade num espaço euclidiano
Y	y = Produto cartesiano dos reais k vezes
Z	Produto cartesiano X com Y

RESUMO

O objetivo deste trabalho é apresentar a base teórica para o problema de aprendizagem através de exemplos conforme as ref. [14], [15] e [16]. Aprender através de exemplos pode ser examinado como o problema de regressão da aproximação de uma função multivaluada sobre um conjunto de dados esparsos. Tal problema não é bem posto e a maneira clássica de resolvê-lo é através da teoria de regularização. A teoria de regularização clássica, como será considerada aqui, formula este problema de regressão como o problema variacional de achar a função f que minimiza o funcional

$$Q[f] = \frac{1}{n}\sum_{i=1}^{n}(y_i - f(x_i))^2 + \lambda\|f\|_K^2,$$

onde $\|f\|_K^2$ é a norma em um espaço de Hilbert especial que chamaremos de Núcleo Reprodutivo (Reproducing Kernel Hilbert Spaces), ou somente RKHS, $\mathcal{H}$ definido pela função positiva K, o número de pontos do exemplo n e o parâmetro de regularização λ. Sob condições gerais a solução da equação é dada por

$$f(x) = \sum_{i=1}^{n} c_i K(x, x_i).$$

A teoria apresentada neste trabalho é na verdade a fundamentação para uma teoria mais geral que justifica os funcionais regularizados para a aprendizagem através de um conjunto finito de dados e pode ser usada para estender consideravelmente a estrutura clássica a regularização, combinando efetivamente uma perspectiva de análise funcional com modernos avanços em Teoria de Probabilidade e Estatística.

ABSTRACT

The purpose of this work is to present the theoretical framework for the problem of learning from examples developed in [14], [15] e [16]. Learning from examples can be regarded as the regression problem of approximating a multivariate function from sparse data. The problem is ill-posed and the classical way to solve it is using 0regularization theory. Classical regularization theory, as we will consider here, formulates the regression problem as a variational problem of finding the function f that minimizes the functional

$$Q[f] = \frac{1}{n}\sum_{i=1}^{n}(y_i - f(x_i))^2 + \lambda\|f\|_K^2,$$

where $\|f\|_K^2$ is a norm in a Reproducing Kernel Hilbert Spaces $I\!H$ defined by the positive function K, n is the number os data points or examples and λ is the regularization parameter. Under rather general conditions the solution of equation is

$$f(x) = \sum_{i=1}^{n} c_i K(x, x_i).$$

The theory developed in this work is the foundation for a more general theory that justifies regularization functionals for learning from finite sets and can be used to extend considerably the classical framework of regularization, effectively matching a functional analysis perspective with modern advances in the theory of probability and statistics.

1 INTRODUÇÃO

A compreensão da inteligência é considerada um dos grandes problemas da ciência hoje. Para cientistas renomados como Tomaso Poggio e Steve Smale este será o grande problema deste século [2], assim como a descoberta do código genético foi para a segunda metade do século passado. O problema de aprendizagem representa a passagem entre descobrir como o cérebro funciona até criar máquinas inteligentes que aprendam através de experiências e aumentem sua capacidade de resolução de tarefas. O aprendizado por exemplos se refere a um sistema que é treinado com um conjunto de exemplos ao invés de ser programado. Este tipo de sistema, que podem aprender através de exemplos a executar uma tarefa específica pode ter muitas aplicações, como a gerência de uma conta de ativos no mercado financeiros. Decidir sobre empréstimos a clientes de um determinado banco, ajudar um médico a decidir o melhor tratamento a um dado conjunto de sintomas, etc. No caso de empréstimos à clientes de um determinado banco, consideramos cada empréstimo um ponto em um espaço multidimensional de variáveis caracterizando suas propriedades e a esse ponto é associado um valor binário: "bom"ou "mau"negócio.

O que se supõe é que uma máquina é treinada ao invés de ser programada para desenvolver uma determinada tarefa. Treinar significa minimizar a função que melhor representa a relação entre os dados de entrada e os dados de saída. A principal questão em teoria de aprendizagem é o quanto esta função é bem generalizada, isto é, que confiança podemos ter na previsão dos dados que esta função dá.

A integração de Monte Carlo é um exemplo clássico em Teoria de Aprendizagem. Seja $f : [0,1]^n \to I\!R$. O processo de Monte Carlo para computar a integral $\int_{x\in[0,1]^n} f(x)dx$ consiste em tomar uma amostra aleatória de pontos $x_1, x_2, ..., x_m \in [0,1]^n$ e calcular $I_m(f) = \frac{1}{m}\sum_{i=1}^{m} f(x_i)$. Sob condições amenas sobre f, temos que

$$\lim_{m\to\infty} Prob_{x_1,...,x_m}\{|I_m(f) - \int f| > \epsilon\} = 0.$$

Aprender, neste caso, significa que a medida que m aumenta tem-se que a probabilidade de $I_m(f)$ estar a uma diferença maior que ϵ de $\int f$ tende a zero, ou seja quanto mais exemplos tivermos menor será a probabilidade de erro. Tecnicamente isto é conhecido como aprendizado PAC (Probably Approximately Correct).

O objetivo é então aprender o valor da integral (um número real) através da amostra. Observe que aqui a medida que governa a amostra é conhecida (medida de Lebesgue em $\mathbb{R}^n$), mas a idéia pode ser usada também para uma medida desconhecida. Se ρ_X é a medida de probabilidade em $X \in \mathbb{R}^n$, um domínio ou uma variedade, $I_m(f)$ aproximará com alta probabilidade $\int_{x \in X} f(x) d\rho_X$, para m grande, desde de que $x_1, x_2, ..., x_m$ estejam distribuídos em X de acordo com a medida ρ_X.

A teoria de aprendizagem foi introduzida no final da década de 60. Até 1990 ela era puramente análise teórica do problema de estimar uma função de um conjunto de dados dado. No meio da década de 1990 novos tipos de algoritmos de aprendizagem foram desenvolvidos (chamados support vector machines), baseado na teoria que havia sido desenvolvida. Isto fez da Teoria de Aprendizagem não apenas uma ferramenta para análise teórica, mas também uma ferramenta para criar algoritmos práticos para estimar funções multidimensionais. Este trabalho apresenta uma visão geral dos aspectos teóricos da Teoria de Aprendizagem, que é uma Teoria que está em pleno desenvolvimento, por isso algumas estimativas que serão apresentadas neste trabalho já foram melhoradas, mas apresentá-las foge do objetivo, que é demonstrar as bases da teoria de aprendizagem clásica e apresentar a generalização desta para Teoria de Vapnik.

2 UMA VISÃO GERAL

O objetivo deste capítulo é apenas dar uma idéia geral de como o algorítmo a ser estudado funciona. Para tal, este capítulo descreverá de uma maneira heurística o algorítmo que através de um conjunto de dados $z = (x_i, y_i)$ obtém uma função $f : X \to Y$ (onde X é um subespaço fechado de $\mathbb{R}^n$ e $Y \subset \mathbb{R}$) que generaliza as associações do conjunto.

A partir de um conjunto de dados $(x_i, y_i)_{i=1}^m$ se escolhe um núcleo simétrico positivo definido contínuo $K_x(x') = K(x, x')$ em $X \times X$. Um bom exemplo é o núcleo Gaussiano $K_x(x') = e^{\frac{-\|x-x'\|^2}{2\sigma^2}}$ restrito a $X \times X$.

Após definido o núcleo se define $f : X \to Y$ como $f(x) = \sum_{i=1}^m c_i K_{x_i}$, onde $c = (c_1, c_2, ..., c_m)$ e $(m\gamma I + K)c = y$, onde I é a matriz identidade e K é a raiz da matriz positiva definida com elementos $K_{i,j} = K(x_i, x_j)$, y é o vetor com coordenadas y_i e γ é um parâmetro real positivo. Como K é positivo e $(m\gamma I+K)$ é estritamente positivo então o sistema linear $(m\gamma I + K)c = y$ é bem posto. O número de condicionamento é bom se $m\gamma$ for suficientemente grande.

O algorítmo descrito acima pode ser derivado da teoria de regularização de Tikhonov ref. [16] e [17]. Para encontrar a função que minimiza o erro do problema, se resolve o ERM (Empirical Risk Minimization) que consiste em achar uma função em $\mathbb{H}$ (um espaço de funções) que minimiza

$$\frac{1}{m}\sum_{i=1}^{m}(f(x_i) - y_i)^2,$$

que é mal posto, dependendo das hipóteses sobre $\mathbb{H}$. Seguindo Tikhonov, para resolver o problema, pode-se minimizar o funcional regularizado

$$\frac{1}{m}\sum_{i=1}^{m}(f(x_i) - y_i)^2 + \gamma\|f\|_K^2$$

que é bem posto, onde $\|f\|_K^2$ é a norma em $\mathbb{H}_K$ o Núcleo Reproduzido em Espaços de Hilbert(RKHS) definido por K, o termo $\gamma\|f\|_K^2$ é o termo regulador que garante unicidade e suavidade da solução.

Considere o espaço linear gerado por K_{x_j} e defina o produto interno deste espaço como sendo $\langle K_x, K_{x_j} \rangle = K(x, x_j)$ e estende linearmente a $\sum_{j=1}^{r} a_j K_{x_j}$. O completamento deste espaço na norma associada é o RKHS, que é um espaço de Hilbert com norma $\|f\|_K^2$, como foi dito acima.

Para minimizar o funcional na equação acima, toma-se a derivada do funcional com respeito a f e aplica-se um elemento $\overline{f}$ de $I\!H_K$ e iguala-se a zero. Assim obtém-se:

$$\frac{1}{m} \sum_{i=1}^{m} (y_i - f(x_i)) \overline{f}(x_j) - \gamma \langle f, \overline{f} \rangle = 0.$$

Esta equação é válida para qualquer $\overline{f} \in I\!H_K$. Em particular para $\overline{f} = K_x$ tem-se

$$f(x) = \sum_{i=1}^{m} c_i K_{x_i},$$

onde

$$c_i = \frac{y_i - f(x_i)}{m\gamma}.$$

Substituindo $f(x) = \sum_{i=1}^{m} c_i K_{x_i}$ em $c_i = \frac{y_i - f(x_i)}{m\gamma}$ obtém-se $(m\gamma I + K)c = y$.

No capítulo 7 este algorítmo será demonstrado de maneira formal. Pode-se repetir este mesmo raciocínio para uma função $V(y, f(x))$ ao invés de $(y_i - f(x_i))^2$. Chegar-se-á que $f(x) = \sum_{i=1}^{m} c_i K_{x_i}$, mas o sistema dado por $(m\gamma I + K)c = y$ será diferente, provavelmente não linear, dependendo da forma de V. Esta generalização foi realizada primeiramente por Vapnik e a solução deste problema dá origem a Teoria de Vapnik.

3 DEFINIÇÕES BÁSICAS

Para modelar o estudo de Teoria de Aprendizagem precisa-se definir vários objetos matemáticos que serão os alicerces desta teoria. Um objeto primário do desenvolvimento é a medida de probabilidade ρ que governa a amostra, a qual não se conhece, mas que também não é o objetivo revelar.

Seja X um domínio compacto ou uma variedade compacta em um Espaço Euclidiano, $Y = \mathbb{R}^k$ e seja ρ uma medida de probabilidade em $Z = X \times Y$.

O valor esperado (esperança), denotado $E(\xi)$, e a variância, denotada por $\sigma^2(\xi)$, de uma variável aletória são definidos da seguinte maneira:

$$\begin{aligned} E(\xi) &= \int_Z \xi d\rho \\ \sigma^2(\xi) &= E\left((\xi - E(\xi))^2\right) = E(\xi^2) - (E(\xi))^2 \end{aligned}$$

Como o objetivo principal é encontrar uma função que aprenda determinada tarefa através de um conjunto de treinamento, isto é, encontrar uma f a partir de um conjunto de dados, precisa-se de algo que nos dê uma referência do quão boas são essas funções de aproximações, para isso se define o erro de f como:

$$\varepsilon(f) = \varepsilon_\rho(f) = \int_Z (f(x) - y)^2 \, d\rho, \quad f : X \to Y.$$

Pode-se então escrever problema da seguinte maneira: **Qual é a f que minimiza o erro $\varepsilon(f)$?**

Para analisar esta questão decompor-se-á este erro, $\varepsilon(f)$, como a soma de dois componentes, o erro de aproximação e o erro amostral, nos próximos capítulos se analisará cada um desses dois erros. Para isso considere que para cada $x \in X$, $\rho(y|x)$ é a medida de probabilidade condicional em Y e ρ_X a medida de probabilidade marginal,

$$\rho_X(S) = \rho(\pi^{-1}(S))$$

onde $\pi : X \times Y \to \mathrm{X}$ é a projeção de $X \times Y$ em X .

Para toda função integrável $\varphi : X \times Y \to I\!R$ o Teorema de Fubini, afirma que

$$\int_{X\times Y} \varphi(x,y)d\rho = \int_X \left(\int_Y \varphi(y,x)d\rho(y|x) \right) d\rho_X$$

DEFINIÇÃO: A função regressão f_ρ de ρ é definida por

$$f_\rho(x) = \int_Y y d\rho(y|x) = E[y] \text{ para } y|x.$$

Para cada $x \in X$, f_ρ é o valor esperado de y na fibra $\{x\} \times Y$. Observe que hipóteses de regularidade sobre ρ induzem regularidade sobre f_ρ.

Dado um espaço $Z = X \times Y$ e uma medida ρ o objetivo é encontrar $f(x) = f_\rho(x) = \int_Y y d\rho(y|x)$. Para fazer isto, primeiro se procura tal função dentro de um espaço de funções $I\!H$, a qual se chama de $f_{I\!H}$, e após aproxima-se está função por uma função de regressão, f_z

Fixando $x \in X$ e considerando a função de Y em $I\!R$ que leva y em $(y - f_\rho(x))$, (o valor esperado é 0), a variância é $\sigma^2(x) = \int_Y (y - f_\rho(x))^2 d\rho(y|x)$.
Fazendo a média sobre X, definimos

$$\sigma_\rho^2 = \int_X \sigma^2(x) d\rho_X = \varepsilon(f_\rho).$$

O número σ_ρ^2 pode ser visto como uma medida de quão precisa é a solução do nosso problema, análogo ao número de condicionamento em álgebra linear.

PROPOSIÇÃO 1: Para toda $f : X \to Y$

$$\varepsilon(f) = \int_X (f(x) - f_\rho(x))^2 + \sigma_\rho^2.$$

PROVA:

$$\begin{aligned}
\varepsilon(f) &= \int_Z (f(x)-y)^2\, d\rho = \int_Z (f(x)-f_\rho(x)+f_\rho(x)-y)^2\, d\rho = \\
&= \int_{X\times Y} (f(x)-f_\rho(x))^2\, d\rho + \int_{X\times Y} (f_\rho(x)-y)^2 d\rho \\
&\quad +2\int_{X\times Y} (f(x)-f_\rho(x))\,(f_\rho(x)-y)\, d\rho = \\
&= \int_X (f(x)-f_\rho(x))^2\, d\rho_X + \int_X \left(\int_Y (f_\rho(x)-y)^2\, d\rho(y|x)\right) d\rho_x \\
&\quad + \int_X \left(\int_Y (f(x)-f_\rho(x))\,(f_\rho(x)-y)\right) d\rho(y|x) d\rho_X = \\
&= \int_X (f(x)-f_\rho(x))^2 + \sigma_\rho^2.
\end{aligned}$$

A $\int_X (f(x)-f_\rho(x))^2 + \sigma_\rho^2$ fornece uma média do erro que cometemos usando f para modelar f_ρ.

Como σ_ρ^2 é independente de f, esta proposição indica que f_ρ tem o menor erro possível entre todas as funções $f : X \to Y$. Assim σ_ρ^2 representa o limite inferior do erro ε e observe que este limite inferior é válido para qualquer f em qualquer espaço de funções. No trabalho se fará estimativa de erros para alguns espaços de funções específicos, mas nenhuma destas estimativas pode ser menor que σ_ρ^2, pois foi mostrado que este é um limite inferior para o erro.

DEFINIÇÃO: Seja $z \in Z^m$, $z = ((x_1,y_1),(x_2,y_2),...,(x_m,y_m))$ uma amostra em Z^m. O erro empírico de f com respeito a z é dado por

$$\varepsilon_z(f) = \frac{1}{m}\sum_{i=1}^{m} \varepsilon(z_i)^2,$$

onde $\varepsilon(z_i) = f(x_i) - y_i$.

DEFINIÇÃO: Para cada função $f : X \to Y$ denotamos por f_Y a função

$$f_Y : X \times Y \to Y$$
$$(x,y) \mapsto f(x) - y$$

Observe que $\varepsilon(f) = \mathbf{E}(f_Y^2)$ e $\varepsilon_z = \mathbf{E}_z(f_Y^2)$ e como a esperança de f_Y é zero, sua variância é σ_ρ^2.

3.1 Convergência em Probabilidade

Dada f e uma amostra z, $\varepsilon(f)$ é desconhecida mas $\varepsilon_z(f)$ pode ser conhecida. O principal Teorema desta seção faz uma estimativa da probabilidade de $\varepsilon(f)$ e $\varepsilon_z(f)$ estarem a uma diferença maior que ϵ. Para facilitar a notação definimos um funcional que mede esta diferença, chama-se este funcional de defeito.

DEFINIÇÃO: Seja $f : X \to Y$. A função defeito de f

$$L_z(f) = L_{\rho,z}(f) = \varepsilon(f) - \varepsilon_z(f)$$

TEOREMA A: Seja $M > 0$ e $f : X \to Y$ tal que $|f(x) - y| \leq M$ q.s. em Z, então para todo $\epsilon > 0$

$$Prob\left\{z \in Z^m : |L_z(f)| < \epsilon\right\} \geq 1 - 2\exp\left(\frac{-m\epsilon^2}{2(\sigma^2 + \frac{1}{2}M^2\epsilon)}\right)$$

onde σ^2 é a variância de f_Y^2.

O Teorema A fala sobre a probabilidade de o erro empírico estar a uma diferença ϵ do erro $\varepsilon(f)$, ou seja, ele limita a $Prob\left\{|L_z(f)| < \epsilon\right\}$ para uma funçao $f : X \to Y$. Além disso pode-se observar que quanto maior for m mais provável será que $|\varepsilon(f) - \varepsilon_z(f)| < \epsilon$, isto é quanto maior for minha amostra maior será a probabilidade da função que minimiza o risco empírico estar próxima da função que minimiza o erro. Em suma o Teorema diz que se $\varepsilon_z(f)$ é pequeno então a probabilidade de que $\varepsilon(f)$ seja pequeno é bastante alta. A demonstração deste teorema é uma conseqüenciencia direta da Desigualdade de Bernstein.

PROPOSIÇÃO 2(Desigualdade de Bernstein): Seja ξ uma variável aleatória no espaço de probabilidade Z com $E(\xi) = \mu$ e $\sigma^2(\xi) = \sigma^2$. Se $|\xi(z) - E(\xi)| \leq M$ para todo $z \in Z$, então para todo $\epsilon > 0$

$$Prob\left\{z \in Z^m : \left|\frac{1}{m}\sum_{i=1}^{m}\xi(z_i) - \mu\right| \geq \epsilon\right\} \leq 2\exp\left(\frac{-m\epsilon^2}{2(\sigma^2 + \frac{1}{3}M\epsilon)}\right)$$

A demonstração desta proposição está no Apêndice A.

PROVA DO TEOREMA A: Adimitindo as hipóteses do Teorema A, seja σ^2 a variância de f_Y^2. Tem-se que

$$Prob\{z \in Z^m : |L_z(f)| < \epsilon\} = 1 - Prob\{z \in Z^m : |L_z(f)| \geq \epsilon\}$$

Usando a desigualdade de Bernstein para a segunda probabilidade obtém-se

$$Prob\{z \in Z^m : |L_z(f)| < \epsilon\} \geq 1 - 2\exp\left(\frac{-m\epsilon^2}{2(\sigma^2 + \frac{1}{2}M^2\epsilon)}\right)$$

3.2 Funções Alvo

O processo de aprendizagem não acontece sem treino, i.e. sem um conjunto de dados. Algumas estruturas precisam estar presentes para que o processo se inicialize. Para desenvolver formalmente a teoria, supõe-se que esta estrutura toma a forma de uma classe de funções. Analisa-se o erro de f supondo que $f : X \to Y$ pertence a um conjunto de funções $I\!H$, pois fazendo certas hipóteses sobre $I\!H$ pode-se estimar o erro. Seja $C(X)$ o espaço de Banach de funções contínuas em X com norma $\|f\| = \sup_{x \in X} |f(x)|$.

O espaço onde o algoritmo irá procurar a melhor aproximação possível para f_ρ é um subconjunto compacto $I\!H$ de $C(X)$ e se chama de espaço da Hipótese.

A principal escolha neste trabalho é um subconjunto compacto de $C(X)$, mas se irá também considerar bolas fechadas nos subespaços de dimensão finita de $C(X)$. É bastante importante esta escolha de $I\!H$, pois desta maneira garantimos a existência de $f_{I\!H}$ e f_z (definidos abaixo).

TEOREMA: Se $I\!H$ é compacto então existe f em $I\!H$ tq f minimiza $\int_Z (f(x) - y)^2 \, d\rho$.

Prova: Primeiramente observe que $\varepsilon : C(X) \to I\!R$ é contínua pois dado $I\!H \subseteq C(X)$ tal que para toda $f \in I\!H$, $|f(x) - y| \leq M$ q.s., então $|\varepsilon(f_1) - \varepsilon(f_2)| \leq 2M \|f_1 - f_2\|_\infty$ e $|\varepsilon_z(f_1) - \varepsilon_z(f_2)| \leq 2M \|f_1 - f_2\|_\infty$, implicando que $\varepsilon, \varepsilon_z : I\!H \to I\!R$ são contínuas em $I\!H$. Como $I\!H$ é compacto temos que o subconjunto dos reais $\varepsilon(I\!H)$ é compacto, logo assume mínimo, ou seja existe uma sequência $f_n \in I\!H$ tq $\varepsilon(f_n)$

tende para o mínimo de $\varepsilon(I\!H)$. Esta sequência $f_n \in I\!H$ tem uma subseqüência que converge para $f \in I\!H$, que tem erro mínimo, pois $I\!H$ é compacto e ε contínua.

DEFINIÇÃO: A função $f_{I\!H}$ é uma função que minimiza o erro $\varepsilon(f)$, satisfazendo

$$\min_{f \in I\!H} \int_Z (f(x) - y)^2 \, d\rho.$$

Como $\varepsilon(f) = \int_X (f - f_\rho)^2 + \sigma_\rho^2$, $f_{I\!H}$ também é um mínimo de

$$\min_{f \in I\!H} \int_X (f - f_\rho)^2 .$$

Aqui $f_{I\!H}$ não é necessariamente única, pode-se garantir a unicidade quando $I\!H$ for convexo.

DEFINIÇÃO: Seja $z \in Z^m$ uma amostra. Definimos a função alvo empírica $f_{I\!H,z} = f_z$ como a função que minimiza o erro empírico $\varepsilon_z(f)$ sobre $f \in I\!H$, i.e. satisfaz

$$\min_{f \in I\!H} \frac{1}{m} \sum_{i=1}^{m} (f(x_i) - y_i)^2 .$$

A existência de f_z segue da compacidade de $I\!H$ e da continuidade de ε_z. Observe que f_z não depende de ρ.

Define-se o erro normalizado de $f \in I\!H$ como

$$\varepsilon_{I\!H}(f) = \varepsilon(f) - \varepsilon(f_{I\!H}) = \int_X (f - f_\rho)^2 \, d\rho - \int_X (f - f_{I\!H})^2 \, d\rho.$$

Denomina-se este erro como normalizado pois $\varepsilon_{I\!H}(f_{I\!H}) = 0$, note também que $\varepsilon_{I\!H}(f) \geq 0$.

Da Proposição 1, tem-se que $\varepsilon(f) = \int_X (f(x) - f_\rho(x))^2 \, d\rho + \sigma_\rho^2$, então $\varepsilon(f_Z) = \varepsilon_{I\!H}(f_Z) + \varepsilon(f_{I\!H}) = \int_X (f_Z - f_\rho)^2 \, d\rho + \sigma_\rho^2$.

Considere a soma $\varepsilon_{I\!H}(f_z) + \varepsilon(f_{I\!H})$. O segundo termo na soma depende da escolha de $I\!H$ mas é independente da amostra. Aqui $\varepsilon(f_{I\!H})$ é o erro da aproximação e $\varepsilon_{I\!H}(f_z)$ é o erro da estimativa (amostral).

Tem-se então dois problemas distintos, estimar o erro amostral e o erro de aproximação, um depende de $\mathbb{H}$ e o outro é dependente de Z. Para um $\mathbb{H}$ fixo o erro amostral cai quando o número de exemplos aumenta (veremos no Teorema C), mas o erro de aproximação diminui quando o tamanho de $\mathbb{H}$ aumenta, então $\varepsilon_{\mathbb{H}}(f_z) = \varepsilon(f_z) - \varepsilon(f_{\mathbb{H}})$ aumenta quando $\mathbb{H}$ aumenta, já que $\varepsilon(f_z)$ não depende de $\mathbb{H}$. Isto é, quando $\mathbb{H}$ aumenta uma das parcelas do erro aumenta (erro amostral) enquanto a outra dimunui (erro de aproximação).

3.3 Estimativas Uniformes do Defeito

O segundo principal resultado, o Teorema B, estende o Teorema A para uma família de funções. Enquanto o Teorema A é uma aplicação imediada da desigualdade de Bernstein, o Teorema B é uma versão da principal estimativa em Teoria de Aprendizagem.

DEFINIÇÃO: Seja S espaço métrico e $s > 0$. O número de cobertura $N(S, s)$ é o número mínimo de bolas de raio s que cobre S. Quando S é compacto (aplicação do Teorema de Heine-Borel), este número é finito.

TEOREMA B: Seja $\mathbb{H}$ um subespaço compacto de $C(X)$. Suponhamos que para todo $f \in \mathbb{H}$, $|f(x) - y| \leq M$ q.s. em Z. Então, para todo $\epsilon > 0$

$$Prob\left\{z \in Z^m : \sup_{f\in\mathbb{H}} |L_z(f)| \leq \epsilon\right\} \geq 1 - N\left(\mathbb{H}, \frac{\epsilon}{8M}\right) 2\exp\left(\frac{-m\epsilon^2}{4(2\sigma^2 + M^2\epsilon/3)}\right),$$

onde

$$\sigma^2 = \sigma^2(\mathbb{H}) = \sup_{f\in\mathbb{H}} \sigma^2(f_Y^2).$$

PROPOSIÇÃO 3: Se $|f_j(x) - y| \leq M$ q.s. para $j = 1, 2$, então para $z \in U^m$

$$|L_z(f_1) - L_z(f_2)| \leq 4M \|f_1 - f_2\|_\infty .$$

PROVA: Primeiro observe que

$$(f_1 - y)^2 - (f_2 - y)^2 = (f_1(x) - f_2(x))\,(f_1(x) - f_2(x) - 2y)$$

Tem-se então

$$\begin{aligned}
|\varepsilon(f_1) - \varepsilon(f_2)| &= \left| \int_Z (f_1 - f_2)(f_1 + f_2 - 2y)\, d\rho \right| \leq \\
&\leq \|f_1 - f_2\|_\infty \int_Z |f_1 - y + f_2 - y|\, d\rho \leq \\
&\leq \|f_1 - f_2\|_\infty \int_Z (|f_1 - y| + |f_2 - y|)\, d\rho \leq \\
&\leq 2M \|f_1 - f_2\|_\infty .
\end{aligned}$$

Também para $Z \in U^m$, tem-se

$$\begin{aligned}
|\varepsilon_z(f_1) - \varepsilon_z(f_2)| &= \frac{1}{m} \left| \sum_{i=1}^{m} (f_1(x_i) - f_2(x_i))(f_1(x_i) + f_2(x_i) - 2y_i) \right| \leq \\
&\leq \|f_1 - f_2\|_\infty \frac{1}{m} \sum_{i=1}^{m} |(f_1(x_i) - y_i) + (f_2(x_i) - y_i)| \leq \\
&\leq \|f_1 - f_2\|_\infty 2M.
\end{aligned}$$

Assim

$$|L_z(f_1) - L_z(f_2)| \leq |\varepsilon(f_1) - \varepsilon_z(f_1) - \varepsilon(f_2) + \varepsilon_z(f_2)| \leq \|f_1 - f_2\|_\infty 4M. \square$$

Seja $I\!H \subseteq C(X)$ tal que para todo $f \in I\!H$, $|f(x) - y| \leq M$ quase sempre, então $|\varepsilon(f_1) - \varepsilon(f_2)| < 2M \|f_1 - f_2\|_\infty$ e $|\varepsilon_z(f_1) - \varepsilon_z(f_2)| \leq 2M \|f_1 - f_2\|_\infty$, implicando que $\varepsilon, \varepsilon_z : I\!H \to I\!R$ são contínuas.

LEMA 1: Sejam $H = S_1 \cup S_2 \cup \ldots \cup S_l$ e $\epsilon > 0$. Então,

$$Prob\left\{ Z \in Z^M : \sup_{f \in I\!H} |L_Z(f)| \geq \epsilon \right\} \leq \sum_{j=1}^{l} Prob\left\{ \sup_{f \in S_j} |L_Z(f)| \geq \epsilon \right\}.$$

Isto segue da equivalência

$$\begin{aligned}
\sup_{f \in I\!H} |L_z(f)| \geq \epsilon \quad &\Leftrightarrow \quad \exists j \leq l \text{ tal que } \sup_{f \in S_j} |L_Z(f)| \geq \epsilon \\
&\Rightarrow \quad Prob\left\{ z \in Z^m : \sup_{f \in I\!H} |L_z(f)| \geq \epsilon \right\} \leq \sum_{j=1}^{l} Prob\left\{ \sup_{f \in S_j} |L_z(f)| \geq \epsilon \right\},
\end{aligned}$$

pois a probabilidade da união de eventos é limitada pela soma das probabilidades destes eventos. $\square$

PROVA DO TEOREMA B: Seja $l = N(I\!H, \epsilon/(4M))$ e consideramos $f_1, \ldots, f_l$ tais que a união dos discos D_j centrado em f_j e de raio $\epsilon/(4M)$ cobre $I\!H$. Seja U um espaço de medida um em que $|f(x) - y| \leq M$. Pela Proposição 3, para todo $z \in U^m$ e $f \in D_j$

$$|L_z(f) - L_z(f_j)| \leq 4M \|f - f_j\|_\infty \leq 4M \frac{\epsilon}{4M} = \epsilon.$$

Como isto é verdade para todo $z \in U^m$ e $f \in D_j$, temos

$$\sup_{f \in D_j} |L_z(f)| \geq 2\epsilon \Rightarrow |L_z(f_j)| \geq \epsilon.$$

Assim para $j = 1, \ldots, l$

$$\begin{aligned} Prob\left\{ z \in Z^m : \sup_{f \in D_j} |L_z(f)| \geq 2\varepsilon \right\} &\leq Prob\left\{ z \in Z^m : |L_z(f_j)| > \epsilon \right\} \\ &\leq 2 \exp\left(\frac{-m\epsilon^2}{2\left(\sigma^2(f_{jy}^2)) + M^2\epsilon/3\right)} \right). \end{aligned}$$

Usando o Teorema A, o Teorema B segue de

$$\begin{aligned} Prob\left\{ \sup_{f \in I\!H} |L_z(f)| \leq \epsilon \right\} &= 1 - Prob\left\{ \sup_{f \in I\!H} |L_z(f)| \geq \epsilon \right\} \geq \\ &\geq 1 - \sum_{j=1}^{l} 2 \exp\left(\frac{-m\epsilon^2}{2\left(\sigma^2(f_{jy}^2) + M^2\epsilon/3\right)} \right) \geq \\ &\geq 1 - l2 \exp\left(\frac{-m\epsilon^2}{4(2\sigma^2 + M^2\epsilon/3)} \right) = \\ &= 1 - N\left(H, \frac{\epsilon}{8M}\right) 2 \exp\left(\frac{-m\epsilon^2}{4(2\sigma^2 + M^2\epsilon/3)} \right). \square \end{aligned}$$

Observe a semelhança do Teorema B com o Teorema A. No Teorema A estima-se a probabilidade da função defeito de uma f ser menor que ϵ , no Teorema B se faz esta estimativa não para apenas uma função, mas para um espaço de funções, por isso se inclui os números de coberturas.

PROPOSIÇÃO 4: Seja F uma família de funções de um espaço de probabilidade Z em $I\!R$ e d a distância sobre F. Seja $U \subset Z$ um espaço de medida completa tal que

a) $|\xi(Z)| < B$ para todo $\xi \in F$ e todo $Z \in U$ e

b) $|L_Z(\varepsilon_1) - L_Z(\varepsilon_2)| \leq \mathbf{L}d(\varepsilon_1, \varepsilon_2)$ para todo $\varepsilon_1, \varepsilon_2 \in F$ e todo $Z \in U^m$,

onde

$$L_z(\xi) = \int_Z \xi(z)d\rho - \frac{1}{M}\sum_{i=1}^{M} \xi(z_i)$$

Então, para todo $\epsilon > 0$

$$Prob\left\{z \in Z^m : \sup_{\xi \in F} |L_Z(\xi)| \leq \epsilon\right\} \geq 1 - N\left(F, \epsilon/(2L)\right) 2\exp\left(\frac{-M\varepsilon^2}{4\left(2\sigma^2 + B\epsilon/3\right)}\right),$$

onde $\sigma^2 = \sigma^2(F) = \sup_{\xi \in F} \sigma^2(\xi)$.

A prova é similar à prova do Teorema B.

4 ESTIMANDO O ERRO AMOSTRAL

O objetivo nesta seção é saber quão bem f_z aproximará $f_{I\!H}$. Ou em outras palavras, quão pequeno pode-se esperar que o erro amostral $\varepsilon_{I\!H}(f_z)$ seja. O algorítmo, descrito no Capítulo 1, gera uma função baseada na amostra, i.e. uma f_z. Necessita-se, então, saber se esta função é uma boa aproximação para f. Para isso, analisa-se o erro de f_z para $f_{I\!H}$ e de $f_{I\!H}$ para f. O Teorema C, que é uma consequência do Teorema B, faz esta estimativa. Para demonstrá-lo, necessita-se do seguinte Lema:

LEMA 2: Seja $I\!H$ um subconjunto compacto de $C(X)$. Sejam $\epsilon > 0$ e $0 < \delta < 1$ com

$$Prob\left\{z \in Z^m : \sup_{f \in I\!H} |L_z(f)| \leq \epsilon\right\} \geq 1 - \delta.$$

Então,

$$Prob\left\{z \in Z^m : \varepsilon_{I\!H}(f_z) < 2\epsilon\right\} \geq 1 - \delta.$$

PROVA: Por hipótese, tem-se com probabilidade maior que $1 - \delta$ que

$$\varepsilon(f_z) \leq \varepsilon_z(f_z) + \epsilon$$

$$\varepsilon_z(f_{I\!H}) \leq \varepsilon(f_{I\!H}) + \epsilon.$$

Além disso, como f_z minimiza ε_z em $I\!H$, tem-se que

$$\varepsilon_z(f_z) \leq \varepsilon_z(f_{I\!H}).$$

Portanto, com probabilidade pelo menos $1 - \delta$

$$\varepsilon(f_z) \leq \varepsilon_z(f_z) + \epsilon \leq \varepsilon_z(f_{I\!H}) + \epsilon \leq \varepsilon(f_{I\!H}) + 2\epsilon$$

e, assim,

$$\varepsilon_{I\!H}(f_z) \leq 2\epsilon.\square$$

TEOREMA C: Seja $I\!H$ um subconjunto compacto de $C(X)$. Supanhamos que, para todo $f \in I\!H$, $|f(x) - y| \leq M$ q.s.. Seja

$$\sigma^2 = \sigma^2(I\!H) = \sup_{f \in I\!H} \sigma^2\left(f_Y^2\right),$$

onde $\sigma^2\left(f_Y^2\right)$ é a variância de f_Y^2. Então, para todo $\epsilon > 0$

$$Prob\left\{z \in Z^m : \varepsilon_{I\!H}(f_z) \leq \epsilon\right\} \geq 1 - N\left(I\!H, \epsilon/(16M)\right) 2\exp\left(\frac{-m\epsilon^2}{8\left(4\sigma^2 + M^2\epsilon/3\right)}\right).$$

PROVA: Basta trocar ϵ por $\epsilon/2$ no Lema 2 e aplicar o Teorema B.

COROLÁRIO: Seja $I\!H$ um subconjunto compacto de $C(X)$. Assume que, para todo $f \in I\!H$, $|f(x) - y| \leq M$ q.s.. Então para que $\epsilon, \delta > 0$ satisfaça

$$Prob\left\{z \in Z^m : \varepsilon_{I\!H}(f_z) \leq \epsilon\right\} \geq 1 - \delta \tag{4.1}$$

é suficiente que o número de exemplos satisfaça

$$m \geq \frac{8\left(4\sigma^2 + M^2\epsilon/3\right)}{\epsilon^2}\left[\ln\left(2N\left(I\!H, \frac{\epsilon}{16M}\right)\right) + \ln\left(\frac{1}{\delta}\right)\right].$$

PROVA: Seja

$$\delta = N\left(I\!H, \frac{\epsilon}{16M}\right) 2\exp\left(\frac{-m\epsilon^2}{8\left(4\sigma^2 + M^2\epsilon/3\right)}\right).$$

Então pelo Teorema C (4.1) é satisfeita

$$\begin{aligned}
\exp\left(\frac{-m\epsilon^2}{8\left(4\sigma^2 + M^2\epsilon/3\right)}\right) &= \frac{\delta}{N\left(I\!H, \epsilon/(16M)\right)} \\
\left(\frac{-m\epsilon^2}{8\left(4\sigma^2 + M^2\epsilon/3\right)}\right) &= \ln\left[\frac{\delta}{N\left(I\!H, \epsilon/(16M)\right)}\right] \\
m\epsilon^2 &= 8\left(4\sigma^2 + M^2\epsilon/3\right)\ln\left[\frac{N(I\!H, \epsilon/(16M))}{\delta}\right] \\
m &\geq \frac{8\left(4\sigma^2 + M^2\epsilon/3\right)}{\epsilon^2}\ln\left[N(I\!H, \frac{\epsilon}{16M}) + \frac{1}{\delta}\right].\square
\end{aligned}$$

4.1 Estimando Números de Coberturas

Como pode-se perceber, para aplicar os Teoremas B e C, tem-se que saber os valores, ou pelo menos ter uma estimativa, dos números de coberturas. Nesta seção serão feitas estimativas para o número de cobertura de certos conjuntos compactos em espaços de Banach de dimensão finita e espaços de Sobolev.

DEFINIÇÃO: Seja S um espaço métrico. Para $k \geq 1$ define-se

$$e_k(S) = \inf \{\epsilon > 0 : \exists \text{ bolas fechadas } D_1, \ldots, D_k \text{ com raio } \epsilon \text{ cobrindo } S\}.$$

Observe que $e_k(S) \leq \eta \Leftrightarrow N(S, \eta) \leq k$. Note que para todo $R > 0$, $e_k(RS) = Re_k(S)$. Aqui $RS = \{Rx : x \in S\}$.

DEFINIÇÃO: Para $k \geq 1$, define

$$\varphi_k(S) = \sup \{\delta > 0 : \exists x_1, \ldots, x_{k+1} \in S \text{ t.q. para } i \neq j, d(x_i, x_j) > 2\delta\}$$

Por exemplo, S sendo a esfera de raio 1, $\varphi_0(S) = \sqrt{2}$. Observe que para todo $k > 1$, $\varphi_k(S) \leq e_k(S) \leq 2\varphi_k(S)$.

LEMA 3: Sejam E um espaço de Banach de dimensão N e B_1 a bola unitária em E. Para todo $k \geq 1$

$$k^{-1/N} \leq e_k(B_1) \leq 4\,(k+1)^{-1/N}.$$

PROVA: Observe que $\varphi_k(B_1) \leq 1$ para todo $k \in \mathbb{N}$. Seja $\rho < \varphi_k(B_1)$. Então, existem $x_1, \ldots, x_{k+1}$ tais que $d(x_i, x_j) > 2\rho$ para $1 \leq i \neq j \leq k+1$. Seja $D_j = x_j + \rho B_1$, $j = 1, \ldots, k+1$. Claramente $D_i \bigcap D_j = \emptyset$ se $i \neq j$ e além disso, para todo $x \in D_j$, $\|x\| < \|x - x_j\| + \|x_j\| < \rho + 1 < 2$. Portanto $D_j \subseteq B_2$.
Como um espaço vetorial, E é isomórfico ao $\mathbb{R}^N$. Qualquer isomorfismo induz sobre E uma medida V que é invariante sobre a translação e é homogênea de grau N com respeito a homotetias (ie, $V(\lambda B) = \lambda^N V(B)$ para todo conjunto mensurável B).
Usando esta medida, chega-se a

$$\begin{aligned} \sum_{i=1}^{k+1} V(D_i) \leq V(B_2) &\Rightarrow \sum_{i=1}^{k+1} \rho^N V(B_1) \leq 2^N V(B_1) \\ &\Rightarrow (k+1)\rho^N \leq 2^N \Rightarrow \rho \leq 2(k+1)^{-1/N} \end{aligned}$$

Daqui tem-se que $e_k(B_1) < 4(4+1)^{-1/N}$.
Para a outra desigualdade considere $\epsilon > e_k(B_1)$. Então existem bolas fechadas $D_1, \ldots, D_k$ de raio ϵ cobrindo B_1 e consequentemente $V(B_1) \leq k\epsilon^N V(B_1)$. Portanto $k^{-1/N} \leq \epsilon$. $\square$

DEFINIÇÃO: Seja $x \in I\!R$, define-se $\lceil x \rceil$ o maior inteiro menor ou igual a x.

PROPOSIÇÃO 5: Sejam E espaço de Banach de dimensão finita e $B_R = \{x \in E : \|x\| < R\}$ e seja $N = dimE$. Então $\ln N(B_R, \eta) \leq N \ln(4R/\eta)$.

PROVA: Seja $k = \left\lceil \left(\frac{4R}{\eta}\right)^N - 1 \right\rceil$. Então $k+1 > \left(\frac{4R}{\eta}\right)^N$ e

$$4(k+1)^{-1/N} \leq \frac{\eta}{R} \Rightarrow e_k(B_1) \leq \frac{\eta}{R} \Leftrightarrow e_k(B_R) \leq \eta \Leftrightarrow N(B_R, \eta) \leq k.$$

Como $k \leq \left(\frac{4R}{\eta}\right)^N$, então $\ln N(B_R, \eta) \leq N \ln(4R/\eta)$.□

4.1.1 Versão Logarítmica dos Números de Entropia

Para $k \geq 1$ definimos o k-ésimo número de entropia do espaço métrico S como

$$\vartheta_k(S) = \inf\{\varepsilon > 0 : \exists \text{ bolas fechadas } D_1, \ldots, D_{2^k-1} \text{ com raio } \varepsilon \text{ combrindo } S\}$$

DEFINIÇÃO: Seja E e F Espaços de Banach e $T : E \to F$ um operador linear, então define-se $\vartheta_k(T) = \vartheta_k(T(B_1))$.

LEMA 4:
a) $\vartheta_k(T) \leq \eta \Leftrightarrow N(T(B_1), \eta) < 2^k - 1$

b) $\vartheta_k T(B_R) = R\vartheta_k(T)$.

PROVA: a) Usando que $e_k(S) \leq \eta \Leftrightarrow N(S, \eta) \leq k$, temos que

$$e_k(T) \leq \eta \Leftrightarrow e_{2^k-1}(T(B_1)) \leq \eta \Leftrightarrow N(T(B_1), \eta) \leq 2^k - 1.$$

b) Basta ver que

$$e_k(T(B_R)) = e_k(T(RB_1)) = e_k(RT(B_1)) = e_k(RT) = Re_k(T).\square$$

4.1.2 Estimando Número de Entropia para Espaços de Sobolev

Se X é um domínio compacto em $\mathbb{R}^n$ com contorno suave, o espaço $C^\infty(X)$ está bem definido. Para todo $s \in \mathbb{N}$, pode-se definir o produto interno em $C^\infty(X)$ por

$$\langle f, g \rangle_s = \int_X \sum_{|\alpha| \leq s} D^\alpha f D^\alpha g,$$

onde $\alpha \in \mathbb{N}^n$,$|\alpha| = \alpha_1 + \alpha_2 + ... + \alpha_n$ $D^\alpha f$ é a derivada parcial

$$\frac{\partial^{|\alpha|} f}{\partial x_1^{\alpha_1} \partial x_2^{\alpha_2} \dots \partial x_n^{\alpha_n}}$$

e integra-se com respeito à medida de Lebesgue μ e sobre X inerente ao espaço Euclidiano. Denota-se por $\| \ \|_s$ a norma induzida por $\langle \ \rangle_s$. Quando $s = 0$, o produto interno coincide com o produto interno de $L^2_\mu(X)$ tem-se $\| \ \|_0 = \| \ \|_{L^2_\mu}$.

DEFINIÇÃO: O espaço de Sobolev $H^s(X)$ é o completamento de $C^\infty(X)$ com respeito a norma s.

O Teorema da Imersão de Sobolev afirma que, para $s > \eta/2$ a inclusão $J_s : H^s(X) \hookrightarrow C(X)$ é bem definida e limitada. Do Teorema de Rellich segue que esta imersão é na verdade compacta. A definição de $H^s(X)$ pode ser extendida para $s \in \mathbb{R}$, $s \geq 0$ ver Edmunds e Tirebel ref.[4].

Assim, se B_R denota bolas fechadas de raio R em $H(X)$, pode-se tomar $H_{Rs} = H = \overline{J_s(B_R)}$.

Suponha que H é a imagem de B_R em $C(X)$. O principal resultado em número de entropia de espaços de Sobolev afirma que, se $X \subseteq \mathbb{R}^n$ é um domínio compacto com contorno suave e $s > n/2$, então para todo $k \geq 1$

$$\vartheta_k(J_s) \leq C \left(\frac{1}{k}\right)^{s/n}$$

Este é o único resultado téorico não trivial deste trabalho que não será demonstrado, a prova provém do Teorema geral de Edmunds and Triebel [3, página 105], considere $s_1 = s$, $s_2 = 0$, $p_1 = 2$ e $p_2 = \infty$. Observe que C é constante e independente de k (que depende X e s).

PROPOSIÇÃO 6: Seja B_R uma bola fechada de raio R centrada na origem em $H^s(X)$ e $I\!H = \overline{J_s(B_R)}$ sua imagem em $C(X)$. Então, para todo $\epsilon > 0$

$$\ln N(I\!H, \epsilon) \leq \left(\frac{RC}{\epsilon}\right)^{n/s} + 1$$

PROVA: Seja $\eta = R\epsilon$ e $k = \left\lceil \left(\frac{C}{\eta}\right)^{n/s} \right\rceil$. Então $\eta \geq C\left(\frac{1}{k}\right)^{s/n}$. Pela desigualdade $\vartheta_k(J_s) \leq C\left(\frac{1}{k}\right)^{s/n}$, nós temos que $\vartheta_k(J_s) < \eta$ e portanto $N(J_s(B_1, \eta)) \leq 2^k - 1$. Assim

$$\ln\left(N(J_s(B_R), R\eta)\right) = \ln N(J_s(B_1), \eta) < k < \left(\frac{RC}{\epsilon}\right)^{n/s} + 1.\square$$

4.2 Espaço de Hipóteses Convexo

Quando o espaço de Hipóteses $I\!H$ é convexo pode-se garantir a unicidade de $f_{I\!H}$, além disso pode-se garantir que o expoente da estimativa dada pelo Teorema C seja linear em ϵ. Observe que quando $\sigma_\rho^2 = 0$, tem-se que para todo $f \in I\!L_\rho^2(X)$, $\sigma^2(f_Y^2) = 0$. Portanto, $\sigma_{I\!H}^2 = 0$ e o expoente no Teorema C fica $\dfrac{3m\epsilon}{8M^2}$, linear em ϵ. Uma outra maneira de tornar esta dependência linear é supor o espaço de Hipótese $I\!H$ convexo. É verdade que a constante do expoente é maior, mas a dependência em ϵ do expoente deixa de ser quadrática e passa a ser linear. Se $I\!H$ é um subespaço compacto e convexo de $C(X)$ e se para todo $f \in I\!H$, $|f(x) - y| \leq M$. Então, para todo $\epsilon > 0$

$$Prob\left\{z \in Z^m : \epsilon_{I\!H}(f_Z) \leq \epsilon\right\} \geq 1 - N\left(I\!H, \frac{\epsilon}{24M}\right) \exp\left(\frac{-m\epsilon}{288M^2}\right)$$

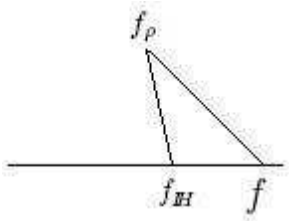
Esse resultado é o Teorema C* que vem a seguir, porém para demonstrá-lo necessita-se dos Lemas e Proposições que serão enunciados e demonstrados a seguir.

Sabe-se que $f_{I\!H}$ é uma função em $I\!H$, onde a distância em $I\!L_\rho^2(X)$ a f_ρ é mínima. Como dito anteriormente se $I\!H$ é convexo, então f_ρ é única.

LEMA 5: Seja $\mathbb{H}$ um subconjunto convexo de $C(X)$ tal que $f_{\mathbb{H}}$ exista. Então $f_{\mathbb{H}}$ é única em $\mathbb{L}^2_\rho(X)$ é para todo $f \in \mathbb{H}$

$$\int_X (f_{\mathbb{H}} - f)^2 \leq \varepsilon_{\mathbb{H}}(f).$$

PROVA: Seja $S = \overline{f_{\mathbb{H}} f}$ o segmento de reta com extremidades $f_{\mathbb{H}}$ e f. Se $\mathbb{H}$ é convexo então $S \in \mathbb{H}$.



Como $f_{\mathbb{H}}$ minimiza distância em $\mathbb{L}^2_\rho$ a f_ρ sobre $\mathbb{H}$, temos que para todo $g \in S$

$$\|f_{\mathbb{H}} - f_\rho\|^2_\rho \leq \|g - f_\rho\|_\rho .$$

Isto implica que o ângulo $\widehat{f_\rho f_{\mathbb{H}} f}$ é obtuso o que implica

$$\|f_{\mathbb{H}} - f\|^2_\rho \leq \|f - f_\rho\|^2_\rho - \|f_{\mathbb{H}} - f_\rho\|^2_\rho .$$

Pela lei dos cossenos, temos que

$$\|f_\rho - f\|^2_\rho = \|f_\rho - f_{\mathbb{H}}\|^2_\rho + \|f - f_{\mathbb{H}}\|^2_\rho - 2\cos\theta \, \|f_\rho - f_{\mathbb{H}}\| \, \|f - f_{\mathbb{H}}\| .$$

Como $\cos\theta < 0$, tem-se que

$$\|f_\rho - f\|^2_\rho \geq \|f_\rho - f_{\mathbb{H}}\|^2_\rho + \|f - f_{\mathbb{H}}\|^2_\rho .$$

O que implica

$$\|f - f_{\mathbb{H}}\|^2_\rho \leq \|f_\rho - f\|^2_\rho - \|f_\rho - f_{\mathbb{H}}\|^2_\rho .$$

Ou seja

$$\int_X (f_{\mathbb{H}} - f)^2 \leq \varepsilon(f) - \varepsilon(f_{\mathbb{H}}).$$

Observe que o argumento geométrico é válido pois $\mathbb{H}$ um espaço de Hilbert.

Para provar a unicidade de $f_{I\!H}$, suponha por absurdo que $f'_{I\!H}$ e $f''_{I\!H}$ sejam dois mínimos, como $I\!H$ é convexo o segmento de reta $\overline{f'_{I\!H} f''_{I\!H}}$ está em $I\!H$. Pelo mesmo argumento acima tem-se que $\widehat{f_\rho f'_{I\!H} f''_{I\!H}}$ e $\widehat{f_\rho f''_{I\!H} f'_{I\!H}}$ são obtusos, absurdo a não ser que $f'_{I\!H} = f''_{I\!H}$.□

Suponha que além de convexo, $I\!H$ é subconjunto compacto de $C(X)$, então os números de cobertura $N(I\!H, \eta)$ são finitos. Assuma que existe $M > 0$ tal que para todo $f \in I\!H$, $|f(x) - y| < M$ quase sempre.

Para uma amostra $z \in Z^m$, o erro empírico em $I\!H$ de $f \in I\!H$ é $\varepsilon_{I\!H,z}(f) = \varepsilon_Z(f) - \varepsilon_z(f_{I\!H})$. Observe que $\varepsilon_{I\!H,z}(f_z) \leq 0$. Seja $l(f) : Z \to Y$ definido por $f_Y^2 - f_{I\!H,Y}^2$. Assim $El(f) = \varepsilon(f) - \varepsilon(f_{I\!H}) = \varepsilon_{I\!H}(f)$ e, para $z \in Z^M$, $E_z l(f) = \varepsilon_z(f) - \varepsilon_z(f_{I\!H}) = \varepsilon_{I\!H,z}(f)$. Além do mais, vamos supor que para todo $f \in I\!H$, $|l(f)(x,y)| \leq M^2$ q.s.. Com convexidade consegue-se provar o Lema 6. Seja $\sigma^2 = \sigma^2(l(f))$ denotando a variânica de $l(f)$.

LEMA 6: Para toda $f \in I\!H$, $\sigma^2 \leq 4M^2 \varepsilon_{I\!H}(f)$.

PROVA:

$$\sigma^2 = El(f)^2 = E\left[(f_{I\!H} - f)^2 (y - f + y - f_{I\!H})^2\right] \leq 4M^2 E\left[(f_{I\!H} - f)^2\right].$$

Pelo Lema 5 temos que

$$E\left[(f_{I\!H} - f)^2\right] \leq \varepsilon_{I\!H}(f).$$

Logo tem-se que $\sigma^2 \leq 4M^2 \varepsilon_{I\!H}(f)$.□

LEMA 7: Seja $f \in I\!H$. Para todo $\epsilon, \alpha > 0$,$\alpha \leq 1$

$$Prob\left\{z \in Z^m : \frac{\varepsilon_{I\!H}(f) - \varepsilon_{I\!H,z}(f)}{\varepsilon_{I\!H}(f) + \epsilon} \geq \alpha\right\} \leq \exp\left(\frac{-\alpha^2 m \epsilon}{8M^2}\right).$$

PROVA: Seja $\mu = \varepsilon_{I\!H}(f)$. Usando a desigualdade de Bernstein aplicada a $l(f)$, juntamente com o fato de $|l(f)((x,y))| \leq M^2$ q.s. em Z, tem-se

$$Prob\left\{z \in Z^m : \frac{\varepsilon_{I\!H}(f) - \varepsilon_{I\!H,z}(f)}{\mu + \epsilon} \geq \alpha\right\} \leq \exp\left(\frac{-\alpha m (\mu + \epsilon)^2}{2\left(\sigma^2 + M^2 \alpha (\mu + \epsilon)/3\right)}\right).$$

Observe que $2\mu\epsilon + \epsilon^2 \leq (\mu+\epsilon)^2$. Pelo Lema 6 $\frac{\epsilon\sigma^2}{4M^2} \leq \epsilon\mu$, além disso $\frac{\epsilon\alpha\mu}{12} \leq \epsilon\mu$ e $\frac{\epsilon^2\alpha}{12} \leq \epsilon^2$ para $\alpha \leq 1$. Assim,

$$\begin{aligned}(\mu+\epsilon)^2 &\geq 2\mu\epsilon+\epsilon^2 \geq \frac{\epsilon\sigma^2}{4M^2}+\frac{\epsilon\alpha\mu}{12}+\frac{\epsilon^2\alpha}{12}\\ &\Leftrightarrow (\mu+\epsilon)^2 \geq \frac{\epsilon}{4M^2}(\sigma^2+\frac{1}{3}M^2\alpha(\mu+\epsilon)) \Leftrightarrow\\ &\Leftrightarrow \frac{(\mu+\epsilon)^2}{(\sigma^2+\frac{1}{3}M^2\alpha(\mu+\epsilon))} \geq \left(\frac{\epsilon}{8M^2}\right)\end{aligned}$$

Então

$$Prob\left\{z \in Z^m : \frac{\varepsilon_{I\!H}(f)-\varepsilon_{I\!H,Z}(f)}{\varepsilon_{I\!H}(f)+\epsilon} \geq \alpha\right\} \leq e^{\left(\frac{-\alpha^2 m\epsilon}{8M^2}\right)}.\square$$

LEMA 8: Seja $0<\alpha<1$, $\epsilon>0$ e $f \in I\!H$ tal que

$$\frac{\varepsilon_{I\!H}(f)-\varepsilon_{I\!H,Z}(f)}{\varepsilon_{I\!H}(f)+\epsilon} < \alpha.$$

Para toda $g \in I\!H$ tal que $\|f-g\|_\infty \leq \frac{\alpha\epsilon}{4M}$, tem-se

$$\frac{\varepsilon_{I\!H}(g)-\varepsilon_{I\!H,Z}(g)}{\varepsilon_{I\!H}(g)+\epsilon} < 3\alpha.$$

PROVA:

$$\begin{aligned}&\frac{\varepsilon_{I\!H}(g)-\varepsilon_{I\!H,Z}(g)}{\varepsilon_{I\!H}(g)+\epsilon} = \frac{\varepsilon(g)-\varepsilon(f_{I\!H})-\varepsilon_z(g)+\varepsilon_z(f_{I\!H})}{\varepsilon I\!H(g)+\epsilon} = \frac{L_z(g)-L_z(f_{I\!H})}{\varepsilon_{I\!H}(g)+\epsilon} =\\ =\ &\frac{L_z(g)-L_z(f)+L_z(f)-L_z(f_{I\!H})}{\varepsilon_{I\!H}(g)+\epsilon} = \frac{L_z(g)-L_z(f)}{\varepsilon_{I\!H}(g)+\epsilon}+\frac{L_z(f)-L_z(f_{I\!H})}{\varepsilon_{I\!H}(g)+\epsilon}\end{aligned}$$

AFIRMAÇÃO 1: $\frac{L_z(g)-L_z(f)}{\varepsilon_{I\!H}(g)+\epsilon} < \alpha$.
Se o termo é negativo, então não há nada a mostrar, caso contrário, tem-se

$$\frac{L_z(g)-L_z(f)}{\varepsilon_{I\!H}(g)+\varepsilon} \leq \frac{L_z(g)-L_z(f)}{\epsilon} \leq \frac{4M\alpha\varepsilon}{4M\epsilon} = \alpha$$

A última desigualdade segue da Proposição 3 usando que $\|f-g\|_\infty < \frac{\alpha\epsilon}{4M}$.
AFIRMAÇÃO 2: $\frac{L_z(f)-L_z(f_{I\!H})}{\varepsilon_{I\!H}(g)+\epsilon} < 2\alpha$.

$$\begin{aligned}\varepsilon(f)-\varepsilon(g) &= \int (f(x)-g(x))\,(f(x)+g(x)-2y) \leq \|f-g\|_\infty \int ((f-y)+(g-y)) \leq\\ &\leq 2M\,\|f-g\|_\infty\\ &\leq \frac{2M\alpha\epsilon}{4M} < \epsilon.\end{aligned}$$

Como $\alpha < 1$. Isto implica que

$$\varepsilon_{I\!H}(f) - \varepsilon_{I\!H}(g) = \varepsilon(f) - \varepsilon(g) \leq \epsilon \leq \varepsilon_{I\!H}(g) + \epsilon.$$

Ou equivalentemente $\dfrac{\varepsilon_{I\!H}(f) + \epsilon}{\varepsilon_{I\!H}(g) + \epsilon} < 2$. Mas então,

$$\frac{L_z(f) - L_z(f_{I\!H})}{\varepsilon_{I\!H}(g) + \epsilon} = \frac{\varepsilon_{I\!H}(f) - \varepsilon_{I\!H,z}(f)}{\varepsilon_{I\!H}(g) + \epsilon} \leq \alpha \frac{\varepsilon_{I\!H}(f) + \epsilon}{\varepsilon_{I\!H}(g) + \epsilon} < 2\alpha. \square$$

PROPOSIÇÃO 7: Para todo $\epsilon > 0$, $0 < \alpha < 1$

$$Prob\left\{ z \in Z^m : \sup_{f \in I\!H} \frac{\varepsilon_{I\!H}(f) - \varepsilon_{I\!H,z}(f)}{\varepsilon_{I\!H}(f) + \epsilon} \geq 3\alpha \right\} \leq N\left(I\!H, \frac{\alpha\epsilon}{4M}\right) \exp\left(\frac{-\alpha^2 m\epsilon}{8M^2}\right).$$

PROVA: Seja $I\!H = S_1 \cup S_2 \cup \ldots \cup S_l$, S bolas fechadas, $0 < \alpha < 1$, $\epsilon > 0$ $l = N\left(I\!H, \dfrac{\alpha\epsilon}{4M}\right)$, considere $f_1, \ldots, f_l$ tais que as bolas centradas em f_j e com raio $\dfrac{\alpha\epsilon}{4M}$ cobrem $I\!H$. Seja U conjunto com $\rho(U) = 1$ tal que $|f(x) - y| \leq M$

$$\|f(x) - g(x)\|_\infty < \frac{\alpha\epsilon}{4M}.$$

Seja $0 < \alpha < 1$, $\epsilon > 0$ e $f \in I\!H$ tal que

$$\frac{\varepsilon_{I\!H}(f) - \varepsilon_{I\!H,z}(f)}{\varepsilon_{I\!H}(f) + \epsilon} < \alpha$$

Então

$$Prob\left\{ z \in Z^m : \sup_{f \in I\!H} \frac{\varepsilon_{I\!H}(f) - \varepsilon_{I\!H,Z}(f)}{\varepsilon_{I\!H}(f) + \epsilon} \geq 3\alpha \right\} \leq \sum_{i=1}^{l} Prob\left\{ z \in Z^m : \sup_{f \in S_i} \frac{\varepsilon_{I\!H}(f) - \varepsilon_{I\!H,Z}(f)}{\varepsilon_{I\!H}(f) + \epsilon} \geq 3\alpha \right\}$$

Pela Proposição 7, tem-se que

$$Prob\left\{ z \in Z^m : \sup_{f \in I\!H} \frac{\varepsilon_{I\!H}(f) - \varepsilon_{I\!H,Z}(f)}{\varepsilon_{I\!H}(f) + \epsilon} \geq 3\alpha \right\} \leq l \exp\left(\frac{-\alpha^2 m\varepsilon}{8M^2}\right) = N\left(I\!H, \frac{\varepsilon\alpha}{4M}\right) \exp\left(\frac{-\alpha^2 m\varepsilon}{8M^2}\right)$$

$\square$

Finalmente, pode-se mostrar o resultado enunciado no início desta seção, o Teorema C*.

TEOREMA C*: Seja $I\!H$ subconjunto convexo e compacto de $C(X)$. Assume que para todo $f \in I\!H$, $|f(x) - y| \leq M$ q.s.. Então, para todo $\epsilon > 0$

$$Prob\left\{ z \in Z^m : \varepsilon_{I\!H}(f_z) \leq \epsilon \right\} \geq 1 - N\left(I\!H, \frac{\epsilon}{24M}\right) \exp\left(\frac{-m\epsilon}{288M^2}\right).$$

PROVA: Colocando $\alpha = 1/6$ na Proposição 7 tem-se com probabilidade pelo menos $1 - N\left(I\!H, \frac{\epsilon}{24M}\right) \exp\left(\frac{-m\epsilon}{288M^2}\right)$, que $\sup_{f \in I\!H} \frac{\varepsilon_{I\!H}(f) - \varepsilon_{I\!H,Z}(f)}{\varepsilon_{I\!H}(f) + \epsilon} < \frac{1}{2}$ e portanto para toda $f \in I\!H$,

$$\frac{1}{2}\varepsilon_{I\!H}(f) < \varepsilon_{I\!H,Z}(f) + \frac{1}{2} + \epsilon$$

Tomando $f = f_z$ e multiplicando tudo por 2

$$\varepsilon_{I\!H}(f_z) < 2\varepsilon_{I\!H,z}(f_z) + \epsilon$$

Mas $\varepsilon_{I\!H,z}(f_z) < 0$ pela definição de f_z, tem-se que $\varepsilon_{I\!H}(f_z) < \epsilon$ e portanto

$$Prob\left\{z \in Z^m : \varepsilon_{I\!H}(f_z) \leq \epsilon\right\} \geq 1 - N\left(I\!H, \frac{\epsilon}{24M}\right) \exp\left(\frac{-m\epsilon}{288M^2}\right). \square$$

COROLÁRIO 1: Com as hipóteses do Teorema C^*, para todo $\epsilon > 0$,

$$Prob\left\{Z \in Z^M : \int (f_Z - f_{I\!H})^2 \leq \epsilon\right\} \geq 1 - N(I\!H)\frac{\epsilon}{24M} e^{-\frac{m\epsilon}{288M^2}}.$$

A prova é consequência direta do Lema 5 com o Teorema C^*.

5 ERRO DE APROXIMAÇÃO

Dado um espaço de hipóteses $\mathbb{H}$, pode-se decompor o erro da função alvo empírica f_Z como a soma do erro da aproximação e o erro de $f_{\mathbb{H}}$, i.e.,

$$\varepsilon(f_Z) = \varepsilon_{\mathbb{H}}(f_Z) + \varepsilon(f_{\mathbb{H}}).$$

No capítulo anterior foi analisado o erro amostral, neste capítulo o objetivo será estudar $\varepsilon(f_{\mathbb{H}})$, o erro de aproximação. Pode-se notar que $\varepsilon(f_{\mathbb{H}})$ depende somente de $\mathbb{H}$ e de ρ. Pela Proposição 1, tem-se que

$$\varepsilon(f_{\mathbb{H}}) = \int_X (f_{\mathbb{H}} - f_\rho)^2 + \sigma_\rho^2,$$

onde

$$\sigma_\rho^2 = \int_X \int_Y (f_\rho(x) - y)^2 .$$

Observe que σ_ρ^2 não depende da escolha do espaço de hipótese $\mathbb{H}$. O fato de f_ρ não ser conhecido de antemão e de não se fazer hipóteses sobre ela além de ser limitada irá restringir o que se pode dizer sobre o erro de aproximação. Pode-se facilmente verificar que se f_ρ está em $\mathbb{H}$ então $\varepsilon(f_{\mathbb{H}}) = \sigma_\rho^2$ pois $\int_X (f_{\mathbb{H}} - f_\rho)^2 = 0$, já que $f_\rho = f_{\mathbb{H}}$.

DEFINIÇÃO: a) Um operador linear limitado $L : H \to H$, H espaço de Hilbert separável, é auto-adjunto se, para todo $f,\ g \in H$, $\langle Lf, g\rangle = \langle f, Lg\rangle$.

b) $L : H \to H$, é dito positivo (estritamente positivo) se é auto-adjunto e se para todo $f \in H$, $f \neq 0$, $\langle Lf, f\rangle \geq 0$ ($\langle Lf, f\rangle > 0$). L é compacto se L(B) é relativamente compacto para todo B limitado.

TEOREMA ESPECTRAL PARA OPERADORES COMPACTOS: Seja L um operador linear limitado compacto autoadjunto em um espaço de Hilbert de dimensão infinita H, então existe em H um sistema ortonornal completo $\{\phi_1, \phi_2, \ldots\}$

consistindo de autovetores de L. Se λ_k é o autovalor correspondente de ϕ_k, então o conjunto $\{\lambda_k\}$ é ou finito ou $\lambda_k \to 0$ quando $k \to \infty$. Além disso, $\max_{k\geq 1} |\lambda_k| = \|L\|$ os autovalores são reais se L é auto-adjunto, e se L é positivo então $\lambda_k \geq 0$ para todo $k \geq 1$, e se L é estritamente positivo então $\lambda_k > 0$ para todo $k \geq 1$.

A demonstração deste teorema pode ser encontrada em [3].

DEFINIÇÃO: Se L é um operador estritamente positivo, então para todo $s > 0$, definimos $L^s(\sum a_k\phi_k) = \sum \lambda_k^s a_k\phi_k$. Se $s < 0$, definimos $L^s(\sum a_k\phi_k) = \sum \lambda_k^s a_k\phi_k$ no subespaço $T_s = \{\sum a_k\phi_k : \sum (|a_k\lambda_k^s|^2)$ é convergente$\}$.
Para $s < 0$, a expressão $\|L^s a\|$ é, por definição, infinito se $a \notin T_s$.

TEOREMA 3: Seja H um espaço de Hilbert, $A : H \to T_{-1}$ operador compacto, auto-adjunto, estritamente positivo. Sejam $p,\ r \in I\!R$ tais que $p > r > 0$.
(1) Seja $\gamma > 0$. Então, para todo $a \in H$

$$\min_{b\in H} \left(\|b-a\|^2 + \gamma \left\|A^{-p}b\right\|^2 \right) \leq \gamma^r \left\|A^{-pr}a\right\|^2 .$$

(2) Seja $R > 0$. Então para todo $a \in H$

$$\min_{b\ t.q. \|A^{-p}b\|\leq R} \|b-a\| \leq \left(\frac{1}{R}\right)^{r/(p-r)} \left\|A^{-r}a\right\|^{r/(p-r)} .$$

Nos dois casos o mínimo $\widehat{b}$ existe e é único. Além disso, em (1), $\widehat{b} = (I + \gamma A^{-2p})^{-1} a$.

PROVA: Sem perda de generalidade, pode-se trocar A por A^p. Assim reduz-se o problema em (1) e (2) para $p = 1$. Define-se

$$\varphi(b) = \|b-a\|^2 + \gamma \left\|A^{-1}b\right\|^2 . \tag{5.1}$$

Primeiramente observe que A não é necessariamente invertível em H, mas é invertível em T_{-1} que é um subconjunto de H. Caso b não esteja em T_{-1} então b não pode ser o mínimo de φ, pois $\|A^{-1}b\| = \infty$. Se um ponto $\widehat{b}$ minimiza φ, então o operador derivada $D\varphi$ deve ser zero neste ponto. Derivando, substituindo $\widehat{b}$ e igualando a zero, verifica-se que $\widehat{b}$ satisfaz $(I + \gamma A^{-2})\,\widehat{b} = a$, implicando que

$$\left(I + \gamma A^{-2}\right)^{-1} a = \widehat{b}. \tag{5.2}$$

O operador γA^{-2} é positivo, o que implica que γA^{-2} é um operador injetor em T_{-2}, logo é invertível em T_{-2} e portanto $I + \gamma A^{-2}$ é invertível em T_{-2}. Logo garante-se que $\widehat{b}$ existe.

Se $\lambda_1 \geq \lambda_2 \geq \ldots > 0$ são os autovalores de A, temos substituindo (5.2) em (5.1) que

$$\begin{aligned}
\varphi(\widehat{b}) &= \left\|\left((I+\gamma A^{-2})^{-1} - I\right) a\right\|^2 + \gamma \left\|A^{-1}\left(I+\gamma A^{-2}\right)^{-1} a\right\|^2 = \\
&= \left\{\sum_{k=1}^{\infty}\left(\frac{1}{1+\gamma\lambda_k^{-2}} - 1\right)^2 + \gamma\sum_{k=1}^{\infty}\left(\frac{1}{\lambda(1+\gamma\lambda_k^{-2})}\right)^2\right\} a_k^2 = \\
&= \sum_{k=1}^{\infty}\left(\frac{\gamma^2\lambda_k^{-4} + \gamma\lambda_k^{-2}}{\lambda_k^{-2}(\lambda_k^2+\gamma)^2}\right) a_k^2 = \gamma\sum_{k=1}^{\infty}\left(\frac{1}{\lambda_k^2+\gamma}\right) a_k^2 = \\
&= \gamma\sum_{k=1}^{\infty}\left(\frac{\lambda_k^{2r}}{\lambda_k^2+\gamma}\right)\lambda_k^{-2r} a_k^2 \leq \gamma \sup_{t\in\mathbb{R}^+}\left(\frac{t^r}{t+\gamma}\right)\left\|A^{-r}a\right\|^2.
\end{aligned}$$

Considere $\psi(t) = \frac{t^r}{t+\gamma}$ e observe que $0 < r \leq 1$, senão $\sup_{t\in\mathbb{R}^+}(\frac{t^r}{t+\gamma})$ será infinito, então tem-se que

$$\psi'(t) = \frac{rt^{r-1}(t+\gamma) - t^r}{(t+\gamma)^2}$$

$$\begin{aligned}
\psi'(t) = 0 &\Leftrightarrow rt^{r-1}(t+\gamma) = t^r \Leftrightarrow r(t+\gamma) = t \Leftrightarrow \\
&\Leftrightarrow rt - t = -r\gamma \Leftrightarrow (r-1)t = r\gamma \Leftrightarrow \\
&\Leftrightarrow t = -r\frac{\gamma}{r-1} \Leftrightarrow t = \frac{r\gamma}{1-r} := \widehat{t}.
\end{aligned}$$

$$\begin{aligned}
\psi(\widehat{t}) &= \frac{r^r\gamma^r}{(1-r)^r}\frac{1}{\frac{r\gamma}{1-r}+\gamma} = \frac{r^r\gamma^r}{(1-r)^r}\frac{(1-r)}{r\gamma+\gamma(1-r)} = \\
&= \frac{r^r\gamma^{r-1}}{(1-r)^{r-1}}\frac{1}{r+(1-r)} = \frac{r^r\gamma^{r-1}}{(1-r)^{r-1}+1} \leq \gamma^{r-1}.
\end{aligned}$$

Assim conclui-se que

$$\varphi(\widehat{b}) = \min_{b\in H}\left(\|b-a\|^2 + \gamma\left\|A^{-1}b\right\|^2\right) \leq \gamma^r\left\|A^{-r}a\right\|^2.$$

Para demonstrar a parte (2), primeiro observe que se $\|A^{-1}a\| \leq R$, então

$$\min_{b\ t.q.\|A^{-1}b\|\leq R}\|b-a\| = 0.$$

Tomando $a = b$ a desigualdade é satisfeita trivialmente, então supõe-se que $\|A^{-1}a\| > R$. Primeiramente observa-se que se $\widehat{b}$ é ponto de mínimo de $\|a-b\|$ no subconjunto de H dado por $\|A^{-1}b\| \leq R$ então $\widehat{b}$ está na fronteira deste subconjunto, ou seja, $\left\|A^{-1}\widehat{b}\right\| = K$. Pois caso $\widehat{b}$ não esteja na fronteira (ou seja $\|A^{-1}b\| < R$) então pode-se tomar um $b' = b - \epsilon(b-a)$ tal que $\|A^{-1}b'\| < R$ $\|a-b'\| = (1-\epsilon)\|a-b\| < \|a-b\|$.

Observe que existe $\gamma \geq 0$ (multiplicador de Lagrange) tal que $\widehat{b}$ é um zero do lagrangeano

$$D\left(\|b-a\|^2\right) + \gamma D\left(\left\|A^{-1}b\right\|^2\right).$$

Este lagrangeano é a mesma coisa que $D\varphi$ na parte (1), que diz de $\varphi(\widehat{b}) \leq \gamma^r \|A^{-r}a\|^2$. Assim tem-se que

$$\gamma R^2 \leq \gamma^r \left\|A^{-r}a\right\|^2. \tag{5.3}$$

A parte da esquerda na inequação vem do fato que $\widehat{b}$ está na fronteira e a direita da parte (1). Com $\gamma > 0$, tem-se que

$$\left\|\widehat{b}-a\right\|^2 \leq \gamma^r \left\|A^{-r}a\right\|^2. \tag{5.4}$$

De (5.3) tem-se que

$$\gamma \leq \left(\frac{1}{R}\right)^{2/(1-r)} \left\|A^{-r}a\right\|^{2/(1-r)}. \tag{5.5}$$

Juntando (5.4) com (5.5) tem-se que

$$\left\|\widehat{b}-a\right\|^2 \leq \left(\frac{1}{R}\right)^{2r/(1-r)} \left\|A^{-r}a\right\|^{2r/(1-r)} \left\|A^{-r}a\right\|^2.$$

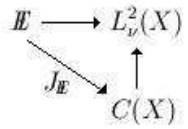
O que implica

$$\begin{aligned} \left\|\widehat{b}-a\right\| &\leq \left(\frac{1}{R}\right)^{r/(1-r)} \left\|A^{-r}a\right\|^{r/(1-r)} \left\|A^{-r}a\right\| = \\ &= \left(\frac{1}{R}\right)^{r/(1-r)} \left\|A^{-r}a\right\|^{1/(1-r)}. \square \end{aligned}$$

DEFINIÇÃO: Seja ν uma medida em X e $\Lambda : \mathbb{L}^2_\nu(X) \to \mathbb{L}^2_\nu(X)$ um operador compacto estritamente positivo. Fixando $p > 0$ seja $\mathbb{E} = \{g \in \mathbb{L}^2_\nu(X) : \|A^{-p}g\|_\nu < \infty\}$. $\mathbb{E}$ é um espaço de Hilbert com produto interno

$$\langle g, h\rangle_{\mathbb{E}} = \left\langle A^{-p}g, A^{-p}h\right\rangle_\nu$$

Assim, $A^{-s} : \mathbb{L}^2_\nu(X) \to \mathbb{E}$ é um isomorfismo de espaços de Hilbert. A inclusão $\mathbb{E} \hookrightarrow \mathbb{L}^2_\nu(X)$ fatora



com $J_{\mathbb{E}}$ compacto.

Portanto o espaço de hipóteses $H = H_{\mathbb{E},R}$ é $\overline{J_{\mathbb{E}}(B_R)}$ onde B_R é a bola de raio R em $\mathbb{E}$. Observe que a função alvo f_H é o $\widehat{b}$ do Teorema 3, para $H = \mathbb{L}^2_\nu(X)$.

DEFINIÇÃO: A distorção de ν com respeito a ρ,$D_{\nu\rho}$, é a norma $\|J\|$ onde J é a identidade entre $\mathbb{L}^2_\nu(X)$ e $\mathbb{L}^2_\rho(X)$. $\mathbb{L}^2_\nu(X) \overset{J}{\longrightarrow} \mathbb{L}^2_\rho(X)$. $D_{\nu\rho}$ mede quanto ρ distorce o ambiente de medida μ. É razoável supor que a distorção é finita.

TEOREMA 4: Em um conjunto de um espaços de Hilbert, para $0 < r < p$ o erro de aproximação satisfaz

$$\varepsilon(f_{\mathbb{H}}) = \|f_{\mathbb{H}} - f_\rho\|^2_\rho + \sigma^2_\rho \leq D^2_{\nu\rho} \left(\frac{1}{R}\right)^{2r/(p-r)} \left\|A^{-r} f_\rho\right\|^{2p/(p-r)} + \sigma^2_\rho. \tag{5.6}$$

PROVA:

$$\|f_\rho - f_{\mathbb{H}}\| = \min_{g \in B_R} \|f_\rho - g\|_\rho \leq D_{\nu\rho} \min_{g \in B_R} \|f_\rho - g\|_\mu \leq D_{\nu\rho} \left(\frac{1}{R}\right)^{r/(p-r)} \left\|A^{-r} f_\rho\right\|^{p/(p-r)}$$

onde a última desigualdade provém do Teorema 3 com $H = \mathbb{L}^2_\nu$ e $a = f_\rho$.

5.1 Erro de Aproximação em Espaços de Sobolev e Núcleos Reprodutivos em Espaços de Hilbert (RKHS)

Nesta seção irá se analizar o erro de aproximação para espaços de Sobolev. Seja $X \subseteq \mathbb{R}^n$ um domínio compacto com contorno suave.

TEOREMA 5: Sejam $p > n/2$ e r tais que $0 < r < p$. Seja $R > 0$ e B_R a bola de raio R em $H^p(X)$ e $\mathbb{H} = \overline{J_p(B_R)}$. Então o erro de aproximação satisfaz

$$\varepsilon(f_{\mathbb{H}}) \leq D_{\mu\rho}^2 C \left(\frac{1}{R}\right)^{2r/(p-r)} \left(\|f_\rho\|_r\right)^{2p/(p-r)} + \sigma_p^2, \tag{5.7}$$

onde C é uma constante que depende apenas de p, r e X.

PROVA: Sejam $\Delta : H^2(X) \to \mathbb{L}_\mu^2$ o Laplaciano e $A = (-\Delta + I)^{-1/2}$. Para $s > 0$, $A^s : \mathbb{L}_\mu^2(X) \to H^s(X)$ é um operador compacto linear com inversa limitada. Então existem $C_0, C_1 > 0$, tais que para todo $g \in H^s(X)$

$$C_0 \|g\|_s \leq \left\|A^{-s} g\right\|_\mu \leq C_1 \|g\|_s .$$

Seja $\mathbb{E}$ o espaço definido neste conjunto com $A = A$ e $p = s$. Então a bola $B_{RC_0}(\mathbb{E})$ de raio RC_0 em $\mathbb{E}$ é incluída na bola $B_R(H^p(X))$ em $H^p(X)$ e consequentemente

$$\varepsilon(f_H) = \min_{g \in B_R(H^p(X))} \|f_\rho - g\|_\rho^2 + \sigma_\rho^2 < \min_{g \in B_{RC_0}(\mathbb{E})} \|f_\rho - g\|_\rho^2 + \sigma_\rho^2.$$

Agora, aplicamos o Teorema 4 para obter

$$\min_{g \in B_{RC_0}} \|f_\rho - g\|_\rho^2 + \sigma_\rho^2 \leq D_{\nu\rho}^2 \left(\frac{1}{RC_0}\right)^{2r/(p-r)} \left\|A^{-r} f_A\right\|_\mu^{2p/(p-r)} + \sigma_\rho^2. \tag{5.8}$$

Finalmente, aplicando a desigualdade do Teorema 5, com $s = r$, chega-se

$$\left\|A^{-r} f_\rho\right\|_\mu \leq C_1 \|f_\rho\|_r .$$

Tomando $C_0^{-2r/(p-r)} C_1^{2p/(p-r)} = C$, tem-se que

$$\varepsilon(f_H) \leq D_{\mu\rho}^2 C \left(\frac{1}{R}\right)^{2r/(p-r)} \left(\|f_\rho\|_r\right)^{2p/(p-r)} + \sigma_\rho^2 . \square$$

Também é possível usar o Teorema 4 para achar os limites para o erro de aproximação de espaços associados a um núcleo.

DEFINIÇÃO: Um núcleo de Mercer K é uma função $K : X \times X \to \mathbb{R}$ que é contínua, simétrica e positiva definida, i.e. para todo conjunto finito $x_1, x_2, ..., x_k \subset X$ a matriz quadrada de ordem k, $K[\mathbf{X}]$ é positiva definida, onde o termo $a_{i,j}$ da matriz $K[\mathbf{X}]$ é $K[x_i, x_j]$.

TEOREMA 6: Sejam K um Núcleo de Mercer, ν uma medida em X, $R > 0$ e $I\!H = \overline{I_K(B_R)}$. O erro de aproximação satisfaz, para $0 < r < 1$

$$\varepsilon(f_{I\!H}) \leq D_{\nu\rho}^2 \left(\frac{1}{R}\right)^{2r/(1-r)} \left\| L_K^{-r/2} f_\rho \right\|_\nu^{2/(1-r)} + \sigma_\rho^2.$$

PROVA: Tomamos $A = L_K^{1/2}$ e $s = l$ no Teorema 4. Para todo $f \in L_\nu^2(X)$, $\|f\|_K = \|A^{-1} f\|_\nu$. Aplicando então o Teorema 4, tem-se que

$$\varepsilon(f_{I\!H}) \leq D_{\nu\rho}^2 \left(\frac{1}{R}\right)^{2r/(1-r)} \left\| L_K^{-r/2} f_\rho \right\|_\nu^{2/(1-r)} + \sigma_\rho^2.$$

6 O PROBLEMA DE *BIAS* VARIANÇA

O objetivo desta seção é achar o valor de R que minimiza o limite para o erro $\varepsilon(f_z)$ com confiança $1-\delta$. Cada $R > 0$ determina um espaço de hipóteses, em particular dentre a família de espaços de hipótese parametrizados por R. Foi visto nos capítulos anteriores que $\varepsilon(f_z)$ pode ser visto como a soma de dois outros erros, onde um desses erros aumenta quando R cresce e o outro diminui, então o problema consiste em achar um R ótimo para esta soma de erros.

DEFINIÇÃO: Considere um espaço de Hilbert. Dada uma amostra de tamanho m e confiança $1-\delta$ com $0 < \delta < 1$, para cada $R > 0$ o espaço de hipóteses $\mathbb{H} = \mathbb{H}_{E,R}$ é bem determinado, sendo assim podemos tomar $f_{\mathbb{H}}$ e, para $z \in Z^m$, f_Z. O problema de Bias Variança no conjunto generalizado consiste em achar R tal que o "limite natural" para o erro $\varepsilon(f_Z)$ seja o menor possível.

TEOREMA 7: Para todo $m \in \mathbb{N}$ e $\delta \in \mathbb{R}$, $0 < \delta < 1$, e para todo r com $0 < r < p$, existe uma solução única R^* do problema de Bias Variança.

PROVA: Temos que $\varepsilon(f_Z) = \varepsilon_{\mathbb{H}}(f_Z) + \varepsilon(f_{\mathbb{H}})$, além disso temos que o Teorema 4 limita o erro de aproximação $\varepsilon(f_{\mathbb{H}})$. Para $0 < r < p$, por $\alpha(R)$, onde $\alpha(R)$ é dada por.

$$\alpha(R) = D_{\nu\rho}^2 \left(\frac{1}{R}\right)^{2r/(p-r)} \left\|A^{-r} f_\rho\right\|_\nu^{2p/(p-r)} + \sigma_\rho^2.$$

Para achar um limitante de $\varepsilon(f_Z)$, tem-se que achar um limitante para $\varepsilon_{\mathbb{H}}(f_Z)$. Para conseguir este limitante, observe que

$$\begin{aligned} |f(x) - y| &\leq |f(x)| + |y| \leq |f(x)| + |y - f_\rho(x)| + |f_\rho(x)| \leq \\ &\leq \|J_{\mathbb{E}}\| R + M_\rho + \|f_\rho\|_\infty := M(R). \end{aligned}$$

Pelo Teorema C^*, o erro amostral ε com confiança $1-\delta$ satisfaz

$$\begin{aligned} & N(I\!H, \epsilon/(24M)) \exp\left(\frac{-m\epsilon}{288M^2}\right) \geq \delta \Rightarrow \\ \Rightarrow \quad & \frac{m\epsilon}{288M^2} - \ln\left(\frac{1}{\delta}\right) - \ln\left(N(I\!H, \epsilon/(24M))\right) \leq 0 \Rightarrow \\ \Rightarrow \quad & \frac{m\epsilon}{288M^2} - \ln\left(\frac{1}{\delta}\right) - \left(\frac{24M^2 C_{I\!E}}{\|J_{I\!E}\|\epsilon}\right)^{1/l_{I\!E}} \leq 0 \end{aligned}$$

Usando que $R\|J_{I\!E}\| \leq M$ e escrevendo $v = \epsilon/(M^2)$, então a equação fica

$$c_0 v - c_1 - c_2 v^{-d} \leq 0,$$

onde

$$\begin{aligned} c_0 &= \frac{m}{288} \\ c_1 &= \ln\left(\frac{1}{\delta}\right) \\ c_2 &= \frac{24 C_{I\!E}}{\|J_{I\!E}\|} \\ d &= \frac{1}{l_{I\!E}}. \end{aligned}$$

Agora observe que a equação $c_0 v - c_1 - c_2 v^{-d} = 0$ tem apenas uma solução positiva para v, lembre que $d > 0$ ($v = \epsilon/(M^2)$, então v é positivo). Seja $v^* = (m, \delta)$ esta solução. Então, $\epsilon(R) = M^2 v^*(m, \delta)$ é o melhor limitante que se pode conseguir através de C^* para o erro amostral.

Tem-se então que $\varepsilon(f_Z) \leq \alpha(R) + \varepsilon(R)$. Minimizando $\alpha(R) + \varepsilon(R)$, tem-se que R é minimo de $\alpha(R) + \varepsilon(R)$ se $-\alpha'(R) = +\varepsilon'(R)$.

Derivando, tem-se que

$$\begin{aligned} \alpha'(R) &= C_A \left(-\frac{2r}{p-r}\right) R^{-(p+r)/(p-r)} \\ \alpha''(R) &= C_A \left(\frac{2r(p+r)}{(p-r)^2},\right) R^{-2r/(p-r)} \end{aligned}$$

Onde

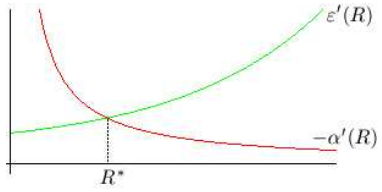
$$\begin{aligned} C_A &= D_{\nu\rho}^2 \left\|A^{-r} f_\rho\right\|^{2p/(p-r)} \\ \varepsilon'(R) &= 2M v^*(m, \delta). \end{aligned}$$

Como $C_A \geq 0$, tem-se que $-\alpha'(R)$ é uma função positiva monótona decrescente em $(0, \infty)$. Por outro lado, com $v^*(m, \delta) > 0$, temos que $\varepsilon'(R)$ é uma função estritamente crescente em $(0, \infty)$.

Como

$$\begin{aligned} \lim_{R\to\infty} \varepsilon'(R) &= \infty \\ \lim_{R\to\infty} -\alpha'(R) &= 0 \\ \lim_{R\to 0} -\alpha'(R) &= \infty \end{aligned}$$

e $\varepsilon'(0)$ é finito, tem-se por continuidade das funções $\alpha(R)$ e $\varepsilon(R)$ que existe um único R^* tal que $\varepsilon'(R^*) = -\alpha'(R^*)$.



Com $J_{\mathbb{E}} : \mathbb{E} \to C(X)$ pode-se estimar os números de entropia para $J_{\mathbb{E}}$ na forma $C_K(J_{\mathbb{E}}) \leq C_{\mathbb{E}}(1/K)^{l_{\mathbb{E}}}$, para algumas constantes positivas $C_{\mathbb{E}}$, $l_{\mathbb{E}}$.

7 OPERADORES DEFINIDOS POR UM NÚCLEO

Lembre-se que X é um domínio compacto ou uma variedade no espaço Euclidiano com $dim(X) = n$, mas para a formalização do algoritmo será suficiente que X seja um espaço métrico completo.

Seja ν uma medida de Borel qualquer em X e $\mathbb{L}^2_\nu(X)$ o espaço de Hilbert das funções de quadrado e integrável em X. A medida de Lebesgue μ e a medida marginal ρ_X são casos particulares desta medida.

Seja $K : X \times X \to \mathbb{R}$ uma função contínua. Então, o operador linear $L_K : \mathbb{L}^2_\nu(X) \to C(X)$, dado pela transformação integral $(L_K f)(X) = \int K(x,t)f(t)d\mu(t)$ está bem definido.

Compondo com a inclusão $C(X) \hookrightarrow \mathbb{L}^2_\nu(X)$, chega-se num operador linear $L_K : \mathbb{L}^2_\nu(X) \to \mathbb{L}^2_\nu(X)$, que também será denotado por L_K.

A função K é o núcleo de L_K e muitas das propriedades de L_K seguem das propriedades de K.

DEFINIÇÃO: Seja $C_K = \sup_{x,t\in X} \|K(x,t)\|$ e $K_x : X \to \mathbb{R}$ dada por $K_x(t) = K(x,t)$, para $x \in X$.

PROPOSIÇÃO 8: Se K é contínuo, então L_K está bem definido e é compacto. Além disso, $\|L_K\| \leq \sqrt{\nu(X)}C_K$, onde $\nu(X)$ denota a medida de X.

PROVA: L_K está bem definido: se $L_K f$ é continua para toda $f \in \mathbb{L}^2_\nu(X)$, seja $f \in \mathbb{L}^2_\nu(X)$ e $x_1, x_2 \in X$. Então

$$\begin{aligned} |(L_K f)(x_1) - (L_K f)(x_2)| &= \left| \int (K(x_1,t) - K(x_2,t))\, f(t) \right| \\ &\leq \|K_{x_1} - K_{x_2}\|\, \|f\| \quad \text{por Cauchy-Schwarz} \\ &\leq \sqrt{\nu(X)} \max_{t\in X} |K(x_1,t) - K(x_2,t)| \|f\|. \end{aligned}$$

Como K é contínua e X é compacto, então K é uniformemente contínua, o que implica que $L_K f$ é contínua.

Mas observe que $|(L_K f)(x)| \leq \sqrt{\nu(X)} \sup_{t \in X} |K(x,t)| \|f\|$ e, portanto, $|(L_K f)(x)| \leq \sqrt{\nu(X)} C_K$.

Por último, para mostrar que L_K é compacto, seja (f_n) uma seqüência limitada em $I\!L^2_\nu(X)$. Como $\|L_K f\|_\infty \leq C_K \|f\|$, temos que $(L_K f_n)$ é uniformemente limitada.

$$|L_K f_n(x_1) - L_K f_n(x_2)| \leq \sqrt{\nu(X)} \max_{t \in X} |K(x_1,t) - K(x_2,t)|$$

então, tem-se que a seqüência $(L_K f_n)$ é equicontínua. Pelo Teorema de Arzela (§1.1.4 de [24]) $(L_K f_n)$ contém uma subseqüência uniformemente convergente.□

PROPOSIÇÃO 9:

a) Se K é simétrico, então $L_K : I\!L^2_\nu(X) \to I\!L^2_\nu(X)$ é auto-adjunto;

b) Se K é positiva definida, então L_K é positiva.

PROVA:

a) Tem-se que

$$\begin{aligned}
\langle L_K f, g \rangle &= \left\langle \int K(x,t) \int K(x,t) f(t), g(x) \right\rangle = \\
&= \int \left(\int K(x,t) L_K f(t) \right) g(x) d\nu = \\
&= \int f(x) \int K(x,t) L_K g(t) d\mu = \\
&= \langle f, L_K g \rangle ,
\end{aligned}$$

isto é, L_K é auto-adjunto.

b) Além disso,

$$\begin{aligned}
\int \int K(x,t) f(x) f(t) &= \lim_{k \to \infty} \frac{\nu(X)}{k^2} \sum_{i,\gamma=1}^{k} K(x_i, x_\gamma) f(x_i) f(x_\gamma) \\
&= \lim_{k \to \infty} \frac{\nu(X)}{k^2} f_{\mathbf{X}}^T K[\mathbf{X}] f_{\mathbf{X}},
\end{aligned}$$

onde para todo $k \geq 1$, $x_1, \ldots, x_k \in X$ é um conjunto de pontos convenientemente escolhidos; $f_{\mathbf{x}} = (f(x_1), \ldots, f(x_k))^{\mathbf{T}}$ e $K[\mathbf{X}]$ é a matriz quadrada de ordem k cujo

elemento $K_(i,j)$ é $K(x_i, x_j)$. Como a matriz é positiva definida então L_K também é positiva definida.□

Como $L_K : I\!L^2_\nu(X) \to I\!L^2_\nu(X)$ é um operador auto-adjunto, compacto e positivo definido, podemos aplicar o Teorema Espectral (Teorema 2 do capítulo 2).

Sejam λ_k, $k \geq 1$, os autovalores de L_K e ϕ_k as correspondentes autofunções. Se $\lambda_k \neq 0$ então ϕ_k é contínua em X, basta ver que $L_K(\phi_k) = \lambda_k \phi_k$ e, portanto, $\phi_k = \frac{1}{\lambda_k} L_K \phi_k$.
De agora em diante, sem perda de generalidade, iremos supor que $\lambda_k > \lambda_{k+1}$ para todo k.

7.1 Teorema de Mercer

Um núcleo de Mercer K é uma função $K : X \times X \to I\!R$ que é contínua simétrica e positiva definida.

Sejam $f \in I\!L^2_\nu(X)$ e $\{\phi_1, \phi_2, \ldots\}$ uma base de Hilbert de $I\!L^2_\nu(X)$. Então, f pode ser escrito de maneira única como uma combinação linear dos elementos da base, i.e., $f = \sum_{k=1}^{\infty} a_k \phi_k$, onde as somas parciais $\sum_{k=1}^{\infty} a_k \phi_k$ convergem a f em $I\!L^2_\nu(X)$.

TEOREMA 8: Sejam X um domínio compacto ou uma variedade, ν uma medida em X e $K : X \times X \to I\!R$ um núcleo de Mercer. Sejam λ_k o k-ésimo autovalor de L_K e $\{\phi_k\}_{k\geq 1}$ os correspondentes autovetores. Então para todo $x, t \in X$, $K(x,t) = \sum_{k=1}^{\infty} \lambda_k \phi_k(x) \phi_k(t)$ onde as somas parciais $\sum_{k=1}^{\infty} \lambda_k \phi_k(x) \phi_k(t)$ convergem absolutamente e uniformemente em X.

A demonstração deste teorema pode ser encontrada em [6].

COROLÁRIO 3: $\sum \lambda_k$ é convergente e $\sum_{k=1}^{\infty} \lambda_k = \int_K K(x,x) \leq \nu(X) C_K$. Portanto, para todo $k \geq 1$, $\lambda_k \leq \left(\frac{\nu(X) C_K}{k}\right)$.

PROVA: Tomando $x = t$ no Teorema 1, tem-se $K(x,x) = \sum_{k=1}^{\infty} \lambda_k \phi_k^2(x)$ e integrando ambos os lados da desigualdade,

$$\sum_{k=1}^{\infty} \int_X \phi_k^2(x) = \int_X K(x,x) \leq \nu(X) C_K.$$

Como $\{\phi_1, \phi_2, \ldots\}$ é uma base de Hilbert, $\int \phi_k^2 = 1$ para todo $k \geq 1$ e segue que

$$\sum_{k=1}^{\infty} \lambda_k = \int_X K(x,x) \leq \nu(X) C_K.$$

Como $\lambda_i \geq \lambda_j$ para $i < j$, vale que

$$\sum_{i=1}^{k} \lambda_i \leq \sum_{i=1}^{\infty} \lambda_i \leq \nu(X) C_K.$$

Mas para todo $i < k$, $\lambda_i \geq \lambda_k$

$$k\lambda_k \leq \sum_{i=1}^{k} \lambda_i \Rightarrow k\lambda_k \leq \mu(X) C_K \Rightarrow \lambda_k \leq \left(\frac{\nu(X) C_K}{k} \right).$$

7.2 Núcleo Reprodutivo em Espaço de Hilbert - RKHS

Um RKHS é um espaço de Hilbert de funções definido sobre um domínio limitado $X \subset I\!R^K$ com propriedade que para cada $x \in X$ o funcional $F_x[f] = f(x)$ é um funcional linear e limitado em $I\!H$. O RKHS é um subespaço de $I\!L_\nu^2$ onde $F_x[f] = f(x)$ é contínuo. Pelo Teorema de Riesz tem-se que existe K_x tal que $f(x) = \langle f, K_x \rangle = \int f(y) K_x(y) d\nu$.

Todo RKHS $I\!H$ corresponde a uma única função positiva definida $K(x,y)$ de duas variáveis em X, chamado de Núcleo Reprodutivo de $I\!H$, que tem a propriedade $f(x) = \langle f(y), K(y,x) \rangle_{I\!H}$ $\forall f \in I\!H$, onde $\langle \cdot, \cdot \rangle_{I\!H}$ denota o produto interno em $I\!H$.

DEFINIÇÃO: Se $f = \sum(a_k \phi_k)$ e $g = \sum(b_k \phi_k)$, λ_k autovalores de L_K, então define-se $\langle f, g \rangle_K = \sum_{k=1}^{\infty} \frac{a_k b_k}{\lambda_k}$

TEOREMA 9: Dado X, K- núcleo de Mercer, existe um único espaço de Hilbert $I\!H_K$ de funções em X satisfazendo as seguintes condições

i) Para todo $x \in X$, $K_x \in I\!H_K$;

ii) O ger$\{K_x : x \in X\}$, i.e. o conjunto das combinações lineares de K_x é denso em $I\!H_K$;

iii) Para toda $f \in I\!H_K$, $f(x) = \langle K_x, f \rangle_K$.

Além disso, $I\!H_K$ consiste em funções contínuas e a inclusão $I_K : I\!H_K \to C(X)$ é limitada com $\|I_K\| \leq C_K^{1/2}$.

PROVA: Seja $H_0 = \text{ger}\{K_x : x \in X\}$. Seja $\langle \cdot, \cdot \rangle$ produto interno entre $f = \sum_{i=1}^{s} \alpha_i K_{x_i}$ e $g = \sum_{j=1}^{r} \beta_j K_{t_j}$ em H_0 dado por

$$\langle f, g \rangle = \sum_{\substack{1 \leq i \leq s \\ 1 \leq j \leq r}} \alpha_i \beta j K(x_i, t_j)$$

Seja $I\!H_K$ o completamento de H_0 com a norma associada ao produto interno, pode-se provar que $I\!H_K$ satisfaz as três condições do Teorema, precisa-se então provar que $I\!H_K$ é único.

Seja H um outro espaço de Hilbert de funções em X satisfazendo as condições i), ii), iii). Tem-se que mostrar que $H = I\!H_K$ e $\langle\ ,\ \rangle_H = \langle\ ,\ \rangle_{I\!H_K}$.
Primeiramente observe que $H_0 \subset H$. Também para qualquer $x, t \in X$, $\langle K_x, K_t \rangle_H = K(x,t) = \langle K_x, K_t \rangle_{I\!H_K}$. Por linearidade, para toda $f, g \in H_0$, $\langle f, g \rangle_H = \langle f, g \rangle_{I\!H_K}$.
Como H e $I\!H_K$ são completamentos de H_0, então $H = I\!H_K$ e $\langle\ ,\ \rangle_H = \langle\ ,\ \rangle_{I\!H_K}$.

Por último, considere $f \in I\!H_K$ e $x \in X$. Então

$$|f(x)| = |\langle K_x, f \rangle| \leq \|f\| \|K_x\| = \|f\| \sqrt{K(x,x)} \Rightarrow \|f\|_\infty \leq \sqrt{C_K} \|f\|_{I\!H_K} \Rightarrow \|I_K\| \leq \sqrt{C_K}$$

Logo, a convergência em $\|\ \|_{I\!H_K}$ implica em convergência em $\|\ \|_\infty$ e assim temos que f é contínua pois é o limite de elementos contínuos de H_0.□

TEOREMA 10: A função

$$\begin{aligned} \Phi : X &\to l^2 \\ x &\longmapsto \left(\sqrt{\lambda_k} \phi_k(x) \right),\ k \in I\!N \end{aligned}$$

está bem definida, é contínua e satisfaz $K(x,t) = \langle \Phi(x), \Phi(t) \rangle$.

PROVA: Para verificar que $\Phi(x) \in l^2$, basta ver que para todo $x \in X$, $\sum \lambda_k \phi_k^2(x)$ converge (pelo Teorema 8), e que além disso, para todo $x, t \in X$ tem-se $K(x,t) = \sum_{k=1} \lambda_k \phi_k(x) \phi_k(t) = \langle \Phi(x), \Phi(t) \rangle$.
Para todo $x, t \in X$, $\|\Phi(x) - \Phi(t)\| = \langle \Phi(x), \Phi(x) \rangle + \langle \Phi(t), \Phi(t) \rangle - 2 \langle \Phi(x), \Phi(t) \rangle = K(x,x) + K(t,t) - 2K(x,t)$. Quando $x \to t$, temos que $\|\Phi(x) - \Phi(t)\| \to 0$ por K é contínua, logo $\Phi(x)$ também é continua.□

Denomina-se em teoria de aprendizagem o espaço em questão de espaço característico e Φ de função característica.

COROLÁRIO 4: Para todo $x, t \in X$, $|K(x,t)| \leq K(x,x)^{1/2} K(t,t)^{1/2}$

PROVA: Pelo Teorema 10, tem-se que

$$\begin{aligned} K(x,t) &= \langle \Phi(x), \Phi(t) \rangle \Rightarrow |K(x,t)| = \\ &= |\langle \Phi(x), \Phi(t) \rangle| \leq \|\Phi(x)\| \|\Phi(t)\| = K(x,x)^{1/2} K(t,t)^{1/2}. \square \end{aligned}$$

O Teorema 2 do Capítulo 2 garante que $\lambda_k \geq 0$ para todo $k \geq 1$, a partir de agora supõe-se, sem perda de generalidade, que $\lambda_k > 0$ para todo $k \geq 1$.
Observe que não há perda de generalidade, pois se existe algum $\lambda_k = 0$, então trocamos $I\!L^2_\nu(X)$ pelo espaço gerado pelos autovetores correspondentes aos autovalores diferentes de zero. Caso este espaço seja de dimensão finita n, então trocamos l^2 por $I\!R^n$.
Seja

$$H_K = \left\{ f \in I\!L^2_\nu(X) : f = \sum_{k=1}^{\infty} a_k \phi_k \text{ com } \left(\frac{a_k}{\sqrt{\lambda_k}} \right) \in l^2 \right\}.$$

Temos que H_K é espaço de Hilbert com produto interno $\langle f, g \rangle_K = \sum_{k=1}^{\infty} \frac{a_k b_k}{\lambda_k}$, onde $f = \sum a_k \phi_k$ e $g = \sum b_k \phi_k$. Note que o operador

$$\begin{aligned} L_K^{1/2} : I\!L^2_\nu(X) &\to H_K \\ \sum a_k \phi_k &\longmapsto \sum a_k \sqrt{\lambda_k} \phi_k \end{aligned}$$

define um isomorfismo entre espaços de Hilbert. Este operador é considerado como a raiz quadrada de L_K, pois $L_K = L_K^{1/2} \circ L_K^{1/2}$.

PROPOSIÇÃO 10: Os elementos de H_K são funções contínuas em X. Além disso, para $f \in H_K$, se $f = \sum a_k \phi_k$, então esta série converge absoluta e uniformemente para f.

PROVA: Sega $g \in H_K$, $g = \sum g_k \phi_k$ e $x \in X$, então temos

$$|g(x)| = \left| \sum_{k=1}^{\infty} g_k \phi_k(x) \right| = \left| \sum_{k=1}^{\infty} \frac{g_k}{\sqrt{\lambda_k}} \sqrt{\lambda^k} \phi_k(x) \right| \leq \|g\|_K \, \|\Phi(x)\| = \|g\|_K K(x,x)^{1/2}$$

Assim, $\|g\|_\infty \leq \sqrt{C_K} \|g\|_K$. Portanto a convergência em $\| \ \|_K$ implica em convergência em $\| \ \|_\infty$. Aplicando em $g_N = f - \sum_{k=1}^{N} a_k \phi_k$, temos que $\sum a_k \phi_k$ converge uniformemente para f. f é contínua, pois pelo Corolário 2 ϕ_k é contínua.
A convergência absoluta segue do fato que

$$\sum |g_k \phi_k| \leq \|g\|_K \|\Phi(x)\|. \square$$

Observe que para $x \in X$, a função $\varphi_x : X \to I\!R$ definida por $\varphi_x(t) = \langle \Phi(x), \Phi(t) \rangle$, pelo Teorema 10, $\varphi_x(t)$ pertence a $I\!H_K$.$\square$

PROPOSIÇÃO 11: Para toda $f \in I\!H_K$ e todo $x \in X$, $f(x) = \langle f, K_x \rangle_K$.

PROVA: Para $f \in I\!H_K$, $f = \sum w_k \phi_k$, temos que

$$\begin{aligned} \langle f, K_x \rangle_K &= \sum_{k=1}^{\infty} w_k \langle \phi_k, K_x \rangle_K = \sum_{k=1}^{\infty} \frac{w_k}{\lambda_k} \langle \phi_k, K_x \rangle = \\ &= \sum_{k=1}^{\infty} \frac{w_k}{\lambda_k} \int \phi_k(t) K(x,t) = \sum_{k=1}^{\infty} \frac{w_k}{\lambda_k} \left(L_K \phi_k \right)(x) \\ &= \sum_{k=1}^{\infty} \frac{w_k}{\lambda_k} \lambda_k \phi_k(x) = f(x). \square \end{aligned}$$

TEOREMA 11: Os espaços de Hilbert $I\!H_K$ e H_K são o mesmo espaço de funções em X com o mesmo produto interno.

PROVA: Para qualquer $x \in X$, a função K_x coincide, pelo Teorema 10, com a função φ_x definida acima, e assim $\varphi_x \in H_K$. Além disso, a Preposição 11 mostra

que para toda $f \in H_K$ e para todo $x \in X$, $f(x) = \langle f, K_x \rangle_K$.
Para mostrar que é gerado de $\{K_x : x \in X\}$ é denso em H_K, assuma que para $f \in H_K$, $\langle f, K_t \rangle_K = 0$ para todo $t \in X$. Então, como $\langle f, K_t \rangle_K = f(t)$, temos que $f = 0$ em X, o que implica, que $\{K_x : x \in X\}$ é denso em $I\!H_K$.□

7.3 Números de Cobertura em RKHS

Nesta seção se estimará o número de cobertura $I\!N\left(\overline{I_K(B_R)}, \eta\right)$ para $R, \eta > 0$. Para poder demonstrar o resultado desta seção, será preciso demonstrar dois lemas.

LEMA 9: Sejam $0 < r < s$ e $a \in I\!L^2_\mu(X)$. Suponhamos que existe $C > 0$ tal que para todo $R > 0$

$$\min_{b \ t.q. \|b\|_s \leq R} \|b - a\| \leq C \left(\frac{1}{R}\right)^{r/(s-r)}.$$

Então, para todo $\delta > 0$, $\|a\|_{r-\delta} \leq c_\delta C^{(s-r)/s}$.

PROVA: A demonstração pode ser encontrada em [14].

LEMA 10: Seja K um núcleo de Mercer C^∞, então a imagem de L_K está em $H^\tau(X)$ para todo $\tau \geq 0$. Considere L_K um funcional linear de $I\!L^2_\nu$ a $H^\tau(X)$ limitado.

PROVA: Para $f \in I\!L^2_\nu(X)$

$$\begin{aligned} \|L_K f\|^2_\tau &= \int_{x \in X} \sum_{|\alpha| < \tau} \left(D^\alpha (L_K f)(x)\right)^2 = \int_{x \in X} \sum_{|\alpha| \leq \tau} \left(\int_{t \in X} D^\alpha_x K_t(x) f(t)\right)^2 \\ &\leq \int_{x \in X} \sum_{|\alpha| < \tau} \int_{t \in X} \left(D^\alpha_x K_t(x)\right)^2 \int_{t \in X} f(t)^2 \leq \|f\|^2_0 \mu(x) \sum_{|\alpha| < \tau} \sup_{x,t \in X} \left(D^\alpha_x K_t(x)\right)^2 .\square \end{aligned}$$

Com esses dois lemas pode-se provar o Teorema D, que estima os números de cobertura em RKHS.

TEOREMA D: Seja $K : X \times X \to I\!R$ um núcleo de Mercer C^∞, e $I\!H_K$ seu correspondente RKHS. Então a inclusão $I_K : I\!H_K \hookrightarrow C(X)$ é compacta e seu número de entropia satisfaz $e_k(I_K) \leq c'_h K^{-h/(2n)}$, para todo $h > n$, onde c'_h é independente

de k. Consequentemente, para $h > n$, $\eta > 0$ e $R > 0$

$$\ln I\!N\left(\overline{I_K(B_R)}, \eta\right) \leq \left(\frac{Rc_n}{\eta}\right)^{2n/h},$$

onde c_n é uma constante pouco maior que c'_h.

PROVA: Seja $f \in I\!H_K$ e $R > 0$. Pelo Teorema 2, Capítulo 2, com $A = L_K$, $s = 1$, $r = 1/2$ e $a = f$ tem-se que

$$\min_{g\ t.q.\left\|L_K^{-1}g\right\| \leq R} \|g - f\| \leq \frac{1}{R}\left\|L_K^{-1/2} f\right\|^2 = \frac{1}{R}\|f\|_K^2$$

Sejam $\tau > 0$ e $c_\tau = \|L_K\|$ para $L_K : I\!L_\mu^2(X) \to H^\tau(X)$. Pelo Lema 9 temos

$$\min_{g\ t.q. \|g\|_\tau \leq R/\tau} \|g - f\| \leq \frac{1}{R}\|f\|_K^2$$

ou, substituindo R/c_τ por R,

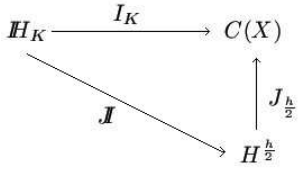
$$\min_{g\ t.q \|g\|_\tau \leq R} \|g - f\| \leq \frac{c_\tau}{R}\|f\|_K^2.$$

Como esta desigualdade vale para todo $R > 0$, podemos aplicar o Lema 2, tomando $s = \tau = 3h/2$, $r = 3h/4$, $\delta = h/4$ e $c = c_\tau \|f\|_K^2$, de onde obtemos

$$\|f\|_{\frac{h}{2}} \leq c' \|f\|_K,$$

com $c' = c_\delta \sqrt{c_{\frac{3h}{2}}}$.

Isto prova a existência de uma imersão limitada $I\!H_K \hookrightarrow H^{\frac{h}{2}}$. Como $h > n$, então aplicando o Teorema da Imersão de Sobolev e o Teorema de Reillich temos a imersão compacta $H^{\frac{h}{2}} \hookrightarrow C(X)$. Assim, temos a seguinte fatoração



que mostra que I_K é compacto.

Além disso pela desigualdade [3.1] tem-se que $\varphi_K(J_{\frac{h}{2}}) \leq c\left(\frac{1}{K}\right)^{h/(2n)}$ para uma constante c independente de K. Portanto

$$\varphi_K(I_K) = \varphi_K\left(J_{\frac{h}{2}} I\!I\right) \leq \varphi_K\left(J_{\frac{h}{2}}\right)\|I\!I\| < c'c\left(\frac{1}{K}\right)^{h/(2n)}$$

que prova a primeira afirmação tomando $c'_h = c'c$. Para provar a segunda, basta usar que $N\left(\overline{I_K(B_R)}, \eta\right) \leq 2^k - 1$ se e somente se $\varphi_K(I_K) \leq \eta/R$ e então

$$\ln N\left(\overline{I_K(B_R)}, \eta\right) \leq \left(\frac{Rc_n}{\eta}\right)^{2n/h}.$$

8 ALGORÍTMO

Considere X, $I\!L^2_\nu(X)$, K, $\| \ \|_K$ e $I\!H_K$ como definidos anteriormente. Aqui se irá redirecionar os estudos, ao invés de considerar um espaço de hipóteses compacto como no Capítulo 1, considera-se-á que $H = I\!H_K$, ou seja $I\!H$ é um espaço linear e assim considera-se o erro regularizado ε_γ definido por

$$\varepsilon_\gamma(f) = \int_Z (f(x) - y)^2 + \gamma \|f\|_K^2$$

para uma $\gamma > 0$ fixo.

DEFINIÇÃO: O erro empírico regularizado $\varepsilon_{\gamma,Z}$ de f é definido por

$$\varepsilon_{\gamma,Z}(f) = \frac{1}{m} \sum_{i=1}^{m} (y_i - f(x_i))^2 + \gamma \|f\|_K^2$$

A função alvo f_γ é a função que minimiza $f_{\gamma,Z}$ em $I\!H$. O objetivo agora é mostrar que esta função existe e é única.

PROPOSIÇÃO 12: Para todo $\gamma > 0$ a função $f_\gamma = (Id + \gamma L_K^{-1})^{-1} f_\rho$ é o único mínimo de ε_γ em $I\!H$.

PROVA: Aplicando o Teorema 3 do Capítulo 2 com $H = I\!L^2_\nu(X)$, $s = 1$, $A = L_K^{1/2}$ e $a = f_\rho$, como para toda $f \in I\!H_K$, $\|f\|_K = \|L_K^{-1/2} f\|_\nu$ a expressão $\|b-a\|^2 + \gamma \|A^{-s} b\|^2$ é $\varepsilon_\gamma(b)$. Assim, f_γ é o $\widehat{b}$ do Teorema 3, isto é

$$f_\gamma = \left(Id + \gamma L_K^{-1}\right)^{-1} f_\rho$$

e também pelo Teorema 9 temos que f_γ é único. □

Pode-se agora finalmente enunciar a proposição que fornece um algorítmo que aproxima as funções alvo, trabalhando num espaço de dimensão infinita $I\!H_K$.

PROPOSIÇÃO 13: Seja $z \in Z^m$ e $\gamma \in I\!R$, $\gamma > 0$. A função alvo empírico, isto é, a função que minimiza

$$\frac{1}{m} \sum_{i=1}^{m} (y_i - f(x_i))^2 + \gamma \|f\|_K^2$$

em $\mathbb{H}_K$ pode ser expressado como

$$f_Z = \sum_{i=1}^{m} a_i K(x, x_i)$$

onde $a = (a_1, \ldots, a_m)$ é a única solução do sistema linear bem posto em $\mathbb{R}^m$

$$(\gamma m Id + K[x])\, a = y$$

onde $K[x]$ é uma matriz $m \times m$ cuja entrada (i, j) é $K(x_i, x_j)$, $x = (x_1, \ldots, x_m) \in X^m$ e $y = (y_1, \ldots, y_m) \in Y^m$ tal que $Z = ((x_1, y_1), \ldots, (x_m, y_m))$.

PROVA: Seja

$$Q(f) = \frac{1}{m} \sum_{i=1}^{m} (y_i - f(x_i))^2 + \gamma \|f\|_K^2$$

e pode-se escrever, para qualquer $f \in \mathbb{H}_K$, $f = \sum_{k=1}^{\infty} c_k \phi_k$. Lembrando que $\|f\|_K^2 = \sum_{k=1}^{\infty} \frac{c_k^2}{\lambda_k}$, para todo $k \geq 1$

$$\frac{\partial Q}{\partial c_k} = \frac{1}{m} \sum_{i=1}^{m} -2(y_i - f(x_i))\phi_k(x_i) + 2\gamma \frac{c_k}{\lambda_k}$$

Se f é o mínimo de Q então, para cada k, temos que ter $\frac{\partial Q}{\partial c_k} = 0$ ou, resolvendo para c_k,

$$c_k = \lambda_k \sum_{i=1}^{m} a_i \phi_k(x_i)$$

onde $a_i = \frac{y_i - f(x_i)}{\gamma m}$. Assim

$$\begin{aligned} f(x) &= \sum_{k=0}^{\infty} c_k \phi_k(x) = \sum_{k=0}^{\infty} \lambda_k \sum_{i=1}^{m} a_i \phi_k(x_i) \phi_k(x) \\ &= \sum_{i=1}^{m} a_i \sum_{k=1}^{m} \lambda_k \phi_k(x_i) \phi_k(x) = \sum_{i=1}^{m} a_i K(x_i, x) \end{aligned}$$

Substituindo $f(x) = \sum_{i=1}^{m} a_i K(x_i, x)$ na definição de a_i obtemos

$$a_i = \frac{y_i - \sum_{i=1}^{m} a_i K(x_i, x)}{\gamma m}$$

O que implica em

$$\begin{aligned} & \gamma m a_i = y_i - \sum_{i=1}^{m} a_i K(x_i, x) \\ \Rightarrow \quad & y_i = \gamma m a_i + \sum_{i=1}^{m} a_i K(x_i, x) \\ \Rightarrow \quad & (\gamma m Id + K[x])\, a = y \end{aligned}$$

O sistema é bem posto pois $K[\mathbf{X}]$ é positivo e a $\gamma m Id$ é estritamente positiva.□

Chega-se então ao objetivo que era mostrar que a solução do problema, conhecido como regularização clássica,

$$\min_{f \in H} Q[f] = \frac{1}{n} \sum_{i=1}^{n} (y_i - f(x_i))^2 + \lambda \|f\|_K^2$$

é dado por

$$f(x) = \sum_{i=1}^{n} c_i K(x, x_i).$$

9 CONCLUSÃO

Através do algoritmo descrito no Capítulo 7 podemos obter uma função onde dado um conjunto de dados, esta função é uma aproximação entre os dados de entrada e de saída, além disso com a teoria desenvolvida nos capítulos anteriores podemos verificar quão provável é que esta função acerte a relação entre os dados de entrada e de saída. O trabalho desenvolvido aqui é apenas uma releitura dos aspectos básicos do problema de aprendizagem em inteligência artificial e se propõe ser um texto para quem quiser começar a estudar este assunto. Este trabalho é baseado nas referências [11], [14], [15], com a diferença que este se propõe a trazer o assunto de uma maneira mais elucidada, com algumas demonstrações que por serem consideradas não tão importantes ou básicas não constam naqueles textos. Quem quiser seguir a partir deste ponto tem dois caminhos, encontrar um problema prático para implementar o algoritmo e estimar o quão provável seja que a função aproxime o conjunto de treino ou continuar a estudar outras variações de problemas teóricos em Teoria de Aprendizagem. A partir de agora, veremos brevemente uma nova técnica que surgiu recentemente e está se tornando muito popular por causa da sua boa performance e pelo fato de estar teoricamente bem embasada. Esta técnica se chama support vector machines (SVMs) proposta por Vladimir Vapnik (1995).

A relação entre a regularização clássica e a SVMs é que elas fornecem o mesmo tipo de solução, ($f(x) = \sum_{i=1}^{n} c_i K(x, x_i)$, mas a função é treinada de uma maneira diferente e por isso fornece diferentes valores para o peso c_i após o treino. Na verdade, em SVM muitos coeficientes são usualmente zero e os coeficientes x_i correspondentes aos coeficientes diferentes de zero são chamados vetores suportes, eles capturam todas as informações relevantes do conjunto de treino.

A teoria de Vapnik justifica a teoria descrita neste trabalho e além pode estendê-la consideralvelmente. Ao invés de considerar funcionais da forma

$$H[f] = \frac{1}{n} \sum_{i=1}^{n} (y_i - f(x_i))^2 + \lambda \|f\|_K^2,$$

considera-se funcionais da forma

$$H[f] = \frac{1}{n}\sum_{i=1}^{n} V(y_i, f(x_i)) + \lambda \|f\|_K^2,$$

onde $V(\cdot,\cdot)$ é a função perda. O problema clássico de regularização corresponde a minimização de H quando

$$V(y_i, f(x_i)) = (y_i - f(x_i))^2,$$

enquanto que o SVM corresponde a minimização do operador H quando

$$V(y_i, f(x_i)) = |y_i - f(x_i)|_\epsilon,$$

onde $|\cdot|_\epsilon$ é a norma epsilon de Vapnik definida por: $|x|_\epsilon = 0$ se $|x| < \epsilon$ e $|x|_\epsilon = |x| - \epsilon$ se $|x| > \epsilon$.

Esta discussão é apenas para mostrar que este trabalho é apenas uma parte muito pequena do estudo de aprendizagem. Exitem, ainda, outras formulações conhecidas para o problema de aprendizagem, onde em cada uma há uma $V(y_i, f(x_i))$ distinta, porém a regularização clássica e SVM são as mais importantes atualmente, uma pelo seu contexto histórico e a outra de fato pela sua utilidade. Entretanto a área é ainda muito recente e espera-se descobrir outras $V(y_i, f(x_i))$ para resolver alguns problemas específicos em teoria de aprendizagem.

Apêndice A DESIGUALDADE DE BERNSTEIN

Para demonstrar a Desigualdade de Bernstein, necessita-se da Desigualdade de Chebyshev Modificada e da Desigualdade de Hoeffding.

LEMA (Desigualdade de Chebyshev Modificada): Seja X_i uma variável aleatória no espaço de probabilidade Z com $E(X_i) = \mu$ e $\sigma^2(X_i) = \sigma^2$, então para todo $\varepsilon > 0$

$$Prob\left\{ z \in Z^m : \left| \frac{1}{m} \sum_{i=1}^{M} \xi(z_i) - \mu \right| \geq \epsilon \right\} \leq \frac{\sigma^2}{m\epsilon^2}.$$

PROVA: Observe que

$$E(\xi) = \int_Z \xi d\rho \Rightarrow E\left(\sum_{i=1}^{M} \xi(z_i) \right) = \int_Z \sum_{i=1}^{M} \xi(z_i) d\rho = \sum_{i=1}^{M} \int_Z \xi(z_i) d\rho = \sum_{i=1}^{M} E(\xi_i) = \mu.$$

Da desigualdade de clássica de Chebyshev, tem-se

$$Prob\left\{ Z \in Z^M : \left| \sum \xi_i - \sum E(\xi_i) \right| \geq \epsilon \right\} \leq \frac{Var\{\xi_i\}}{\epsilon^2}.$$

Escrevendo $\sigma^2 = \frac{1}{M} \sum Var X_i$

$$Prob\left\{ \left| \frac{1}{m} \sum \xi(z_i) - \mu \right| \geq \epsilon \right\} \leq \frac{\sum Var\{\xi_i\}}{M^2\epsilon^2} = \frac{\sigma^2}{m\epsilon^2}.\square$$

LEMA (Desigualdade de Hoeffding): Seja X uma variável aleatória com $EX = 0$, $a \leq X \leq b$. Então para $s > 0$,

$$E[e^{sX}] \leq e^{s^2(b-a)^2/8}.$$

PROVA: Como e^x é uma função convexa tem-se que

$$e^{sx} \leq \frac{x-a}{b-a} e^{sb} + \frac{b-x}{b-a} e^{sa} \qquad \text{para} \qquad a \leq x \leq b.$$

Usando que $EX = 0$ e que $p := \frac{a}{b-a}$, obtém-se que

$$\begin{aligned} Ee^{sX} &\leq \frac{b}{b-a} e^{sa} - \frac{a}{b-a} e^{sb} = \\ &= (1 - p + pe^{s(b-a)}) e^{-ps(b-a)} := e^{\phi(u)}, \end{aligned}$$

onde $u = s(b-a)$, e $\phi(u) = -pu + ln(1-p+pe^u)$. Assim tem-se que $\phi'(u) = -p + \frac{p}{p+(1-p)e^{-u}}$, portanto $\phi(0) = \phi'(0) = 0$. Além disso,

$$\phi''(u) = \frac{p(1-p)e^{-u}}{(p+(1-p)e^{-u})^2} \leq \frac{1}{4}.$$

Portanto, pelo Teorema de Taylor, para algum $y \in [0,u]$,

$$\phi(u) = \phi(0) + u\phi'(0) + \frac{u^2}{2}\phi''(y) \leq \frac{u^2}{8} = \frac{s^2(b-a)^2}{2}$$

Portanto $E[e^{sX}] \leq e^{s^2(b-a)^2/8}$. □

PROPOSIÇÃO (Desigualdade de Bernstein): Seja X_i uma variável aleatória no espaço de probabilidade Z com $E(\xi_i) = \mu$ e $\sigma^2(\xi_i) = \sigma^2$. Se $|\xi(Z) - E(\xi)| \leq M$ para todo $Z \in Z$, então para todo $\epsilon > 0$

$$Prob\left\{ z \in Z^m : \left| \frac{1}{m}\sum_{i=1}^{m} \xi(z_i) - \mu \right| \geq \epsilon \right\} \leq 2\exp\left(\frac{-m\epsilon^2}{2(\sigma^2 + \frac{1}{3}M\epsilon)} \right).$$

PROVA: Suponhamos sem perda de generalizações que $E\xi_i = 0$ para todo i e seja $S_n = \Sigma_{i=1}^n \xi_i$.

$$P\left\{ S_n - \sum S_n \geq \epsilon \right\} \leq e^{-s\epsilon E[exp(s\sum_{i=1}^n \xi_i - E\xi_i)]}. \qquad \text{(Teorema de Chernoff)} \quad \text{(A.1)}$$

Então,

$$P\left\{ S_n - \sum S_n \geq \epsilon \right\} \leq \exp(-s\epsilon) \prod E(\exp[s(\xi_i - E\xi_i)]). \quad \text{(por independência)} \quad \text{(A.2)}$$

Seja X v.a. com $EX = 0$, $-M \leq \xi_i \leq M$ para todo $s > 0$

$$E\left[\exp(s\xi_i)\right] \leq \exp\left(\frac{s^2M^2}{2}\right). \qquad \text{(A.3)}$$

Seja $F_i = \sum_{r=2}^{\infty} \frac{s^{r-2}E\left[\xi_i^r\right]}{r!\sigma_i^2}$. Como $\exp(sx_i) = 1 + sx_i + \sum_{r=2}^{\infty} \frac{s^r x_i^r}{r!}$, então

$$\begin{aligned} E[\exp(sx)] &= 1 + sE[x_i] + s^2\sigma_i^2 F_i = 1 + s^2\sigma_i^2 F_i, \text{pois } E[x_i] = 0 \\ &\leq \exp\left(s^2\sigma_i^2 F_i\right). \end{aligned}$$

Como $|\xi_i| < M$ para todo $r > 2$, então $E\left|\xi_i^r\right| \leq M^{r-2}\sigma_i^2$.

Assim,

$$F_i \leq \sum_{r=2}^{\infty} \frac{s^{r-2}M^{r-2}\sigma^2}{r!\sigma^2} = \frac{1}{(sM)^2}\sum_{r=2}^{\infty} \frac{(sM)^r}{r!} = \frac{\exp(SM) - 1 - sM}{(sM)^2},$$

implicando

$$E[\exp(s\xi_i)] \leq \exp\left(s^2\sigma_i^2 \frac{\exp(sM) - 1 - sM}{(sM)^2}\right).$$

Usando que

$$P\left\{|S_n| > \epsilon\right\} \leq \exp(-s\epsilon) \prod E\left[s^{s\xi_i}\right] \leq \exp\left(\frac{n\sigma^2(\exp(sM) - 1 - sM)}{M^2} - s\epsilon\right),$$

e escolhendo o s que minimiza, obtemos

$$s = \log\left(1 + \frac{M\epsilon}{n\sigma^2}\right)^{1/n}.$$

Substituindo, obtém-se

$$\begin{aligned} P\left\{|S_n| > \epsilon\right\} &\leq \exp\left\{\frac{n\sigma^2}{M^2}\left(1 + \epsilon M - 1 - \log\left(\frac{1 + M\epsilon}{n\sigma^2}\right)\right) - \epsilon\log\left(\frac{1 + \epsilon M}{n\sigma^2}\right)\right\} \\ &\leq \exp\left\{\frac{-n\sigma^2}{M^2}\left(\left(1 + \frac{M\epsilon}{n\sigma^2}\right)\log\left(1 + \frac{M\epsilon}{n\sigma^2}\right) - \frac{M\epsilon}{n\sigma^2}\right)\right\} \\ &\leq \exp\left\{\frac{-n\sigma^2}{M^2} h\left(\frac{M\epsilon}{n\sigma^2}\right)\right\}, \end{aligned}$$

onde $h(u) = (1 + u)\ln(1 + u) - u$, $u > 0$. Note que

$$\begin{aligned} h'(u) &= \frac{1 + u}{1 + u} + \ln(1 + u) - 1 = \ln(1 + u), \\ f(u) &= \frac{u^2}{2 + 2u/3}, \\ \text{e} f'(u) &= \frac{2u(2 + 2u/3) - 2u^2/3}{(2 + 2u/3)^2} = \frac{4u + 4u^2/3 - 2u^2/3}{(2 + 2u/3)^2} = \frac{2u(2 + u/3)}{(2 + 2u/3)^2} \end{aligned}$$

Observe também que $h(0) - f(0) = 0$ e que $\frac{d[h(u) - f(u)]}{du} \geq 0$ logo $h(u) \geq \frac{u^2}{2+2u/3}$, $u \geq 0$, e portanto

$$h\left(\frac{M\epsilon}{n\sigma^2}\right) \geq \frac{\frac{(M\epsilon)^2}{n^2\sigma^4}}{2 + \frac{2}{3}\left(\frac{M\epsilon}{n\sigma^2}\right)}.$$

Então

$$\begin{aligned} P\left\{\frac{1}{n}\sum \xi_i > \epsilon\right\} &\leq \exp\left[\frac{-n\sigma^2}{M^2}\left(\frac{M^2\epsilon^2}{n^2\sigma^2.(2\sigma^2 + 2M\epsilon/3)}\right)\right]. \\ &\leq \exp\left[\frac{-M\epsilon^2}{2\sigma^2 + 2M\epsilon/3}\right] \end{aligned}$$

Então,

$$P\left\{\frac{1}{n}\left|\sum \xi_i >\right| \epsilon\right\} \leq 2\exp\left[\frac{-M\epsilon^2}{2\sigma^2 + 2M\epsilon/3}\right],$$

pois basta separar $\sum_{\xi_i}$ em $\sum \xi_i^+ - \sum \xi_i^-$ e obtém para cada um a majoração $\exp\left[\frac{-M\epsilon^2}{2\sigma^2+2M\epsilon/3}\right]$.

REFERÊNCIAS BIBLIOGRÁFICAS

[1] N. Aronszajn, Theory of reproducing kernels, Transactions of the Amer. Math. Soc.68, 1950, 337-404.

[2] A. Björck, Numerical methods for least squares problems, Applied Mathematics Review, Vol. 50, No. 2, 1997.

[3] L. Debnath e p. Mikusinski, Introduction to Hilbert Space with applications, 2º edição, Academic Prees, 1999.

[4] D.E. Edmunds and H. Triebel, Function spaces, entropy numbers, differential operators, Cambridge University Press, 1996.

[5] T. Evgeniou, M. Pontil, and T. Poggio, Regularization Networks and Support Vector Machines, Advances in Computational Mathematics 13, 2000, 1-50.

[6] H. Hochstadt, Integral equations, John Wiley e Sons, 1973.

[7] D. Haussler, Decision theoretic generalizations of the PAC model for neural net and other learning applications, Information and Computation 100, 1992, 78-150.

[8] A.N. Kolmogorov and V.M. Tikhomirov, Entropy and capacity of sets in function spaces, Uspecki 14, 1959, 3-86.

[9] A.N. Kolmogorov and S.V. Fomin, Introductory real analysis, Dover Publications Inc., 1975.

[10] G.G. Lorentz, M. Golitschek, and Y. Makovoz, Constructive approximation, advanced problems, Springer-Verlag, 1996.

[11] G. Lugosi, Concentration of measure inequalities, Lecture Notes of machine Learning, 2002.

[12] T. Poggio and C.R.Shelton, Machine learning, machine vision, and the brain, AI Magazine 20, 1999, 33-55.

[13] F. Riesz e B. Nagy, Functional Analysis, Dover Books, 1990.

[14] S. Smale and D.-X. Zhou, Estimating the approximation error in learning theory, Analysis and Applications 1, 2003, 1-25.

[15] S.Smale e F. Cucker, On the mathematical foundations of learning, Bulletin of the american mathematical sociey 39, number 1, (2001), pages 1-49.

[16] S.Smale e T. Poggio, The mathematics of learning:dealing with data, Notices of the AMS 50, number 5, (2003), pages 537-544.

[17] V. Vapnik, Statistical learning theory, John Wiley e Sons, 1998.