Thuật toán

$$\min_{f \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^{n} \ell(y_i, f(x_i)) + \frac{\lambda}{2} \| f \|_{\mathcal{H}}^2$$

Giả sử với mọi i € {1, ..., n}, $k(x_i, x_i) = \| \varphi(x_i) \|^2 \leq R^2$

-Định lý đại diện:

$$\min_{\alpha \in \mathbb{R}^n} \frac{1}{n} \sum_{i=1}^{n} \ell(y_i, (K\alpha)_i) + \frac{\lambda}{2} \alpha^\top K \alpha$$

=>Tối ưu hóa lồi

-Xét trường hợp đặc biệt của hàm mất mát bình phương (Ridge Regression):

$$\min_{\alpha \in \mathbb{R}^n} \frac{1}{2n} \|y - K\alpha\|_2^2 + \frac{\lambda}{2} \alpha^\top K \alpha$$

$$\min_{\alpha \in \mathbb{R}^n} \frac{1}{2n}\|y - K\alpha\|_2^2 + \frac{\lambda}{2}\alpha^\top K\alpha$$

-Tính đạo hàm: $\frac{1}{2n}.(-2K^T).(y - K\alpha) + \frac{\lambda}{2}.(K + K^T).\alpha$

$$\Leftrightarrow \quad (K^2 + n\lambda K)\alpha = Ky$$

$$\Leftrightarrow \quad \alpha = (K + n\lambda I)^{-1}y$$

- K thường có trị riêng rất nhỏ => tối ưu không tốt

+ Hessians = $\frac{1}{n} K \,\mathrm{Diag}(h) K + \lambda K$ $\quad h \in R^n$

- Tính căn bậc 2 của K = $\Phi\Phi^T$

+ Xét $\min_{\beta \in \mathbb{R}^m} \frac{1}{n} \sum_{i=1}^{n} \ell(y_i, (\Phi\beta)_i) + \frac{\lambda}{2} \|\beta\|_2^2$

$$\begin{cases} \Phi \in R^{n \times m} \\ \text{m là hạng của ma trận K} \\ \alpha = \Phi^T \beta \end{cases}$$

- Trong trường hợp Ridge Regression, trị riêng nhỏ nhất > λ => tối ưu tốt

+ Hessian(hàm mục tiêu) = $\frac{1}{n}\Phi^\top \Phi + \lambda I$

Xấp xỉ Nyström

-Việc thực hiện tính trên có thể theo 1 số cách như phân rã Cholesky hoặc SVD

+ Xét K € R^{nxn} $\longrightarrow K \approx K(V,I)K(I,I)^{-1}K(I,V)$

Column sampling: Với K(A,B) là ma trận con của K, có được bằng cách lấy các hàng từ A ⊂ {1, . . . , n}, các cột từ B ⊂ {1, . . . , n} và V = ⊂ {1, . . . , n}.

+ Xét với I = {1, . . . , m} :

=> $\Phi = K\,(V, I)\,K\,(I, I)^{-1/2}$

$\begin{cases} \Phi \in R^{nxm} \\ m = |I| \end{cases}$

$K(I,I)$	$K(I,J)$
$K(J,I)$	$K(J,J)$

Random feature: Một số kernel có dạng đặc biệt dẫn đến xấp xỉ cụ thể sau:

$$k(x, x') = \int_{\mathcal{V}} \varphi(x, v)\varphi(x', v)d\mu(v)$$

dμ : phân phối xác suất trên KG V

φ(x,v) € R

- Có thể xấp xỉ kỳ vọng như sau:

$$\hat{k}(x, x') = \frac{1}{m}\sum_{i=1}^{m} \varphi(x, v_i)\varphi(x', v_i) \longrightarrow \min_{\beta \in \mathbb{R}^m} \frac{1}{n}\sum_{i=1}^{n} \ell(y_i, \hat{\varphi}(x_i)^\top \beta) + \frac{\lambda}{2}\|\beta\|_2^2.$$

-Với φ(x) = $\frac{1}{\sqrt{m}}$φ(x, v_i), i € {1,...,m}

- m phải nhỏ hơn n đáng kể, thường là vừa đủ thực tế

-Việc làm giảm kích thước như trên có thể thực hiện độc lập với dữ liệu đầu vào (Random feature hàm φ(·, v_i)), trái với Column sampling thực hiện phụ thuộc.

ví dụ:

+ ***kernels tịnh tiến bất biến***: $k(x,y) = q(x-y) = \frac{1}{(2\pi)^d} \int_{\mathbb{R}^d} \hat{q}(\omega) e^{i\omega^\top (x-y)} d\omega$

ω : được lấy mẫu từ phân phối với mật độ $\frac{1}{(2\pi)^d} \frac{q(\omega)}{q(0)}$ (phân phối Gauss cho Gaussian Kernel)

$$\varphi(x,\omega) = \sqrt{q(0)} e^{i\omega^\top x} \in \mathbb{C},$$

->Thay thế feature € C bởi feature € R: $\sqrt{2} \cos(\omega^\top x + b)$ (b € [0,2π])

+ **Mạng nơron với trọng số ngẫu nhiên:** $\varphi(x,v) = \sigma(v^\top x)$ với σ: R → R

$k(x,x') = \frac{\|x\|_2 \|x'\|_2}{2(d+1)\pi} \left[(\pi - \eta) \cos\eta + \sin\eta \right]$ với $\cos\eta = \frac{x^\top x'}{\|x\|_2 \cdot \|x'\|_2}$

=> một mạng nơ-ron với 1 lượng lớn nơ-ron ẩn, với trọng số đầu vào ngẫu nhiên không tối ưu như kernel method.

Thuật toán đối ngẫu

$$f(x) = \langle \varphi(x), \theta \rangle$$ với θ ∈ H

Dùng đối ngẫu Lagrange giải : $\min_{\theta \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^{n} \ell(y_i, \langle \varphi(x_i), \theta \rangle) + \frac{\lambda}{2} \|\theta\|^2$

$$= \min_{\theta \in \mathcal{H},\ u \in \mathbb{R}^n} \frac{1}{n} \sum_{i=1}^{n} \ell(y_i, u_i) + \frac{\lambda}{2} \|\theta\|^2$$ $\forall i \in \{1, \dots, n\}$, $\langle \varphi(x_i), \theta \rangle = u_i$

Xét: $\max_{\alpha \in \mathbb{R}^n} \min_{\theta \in \mathcal{H},\ u \in \mathbb{R}^n} \frac{1}{n} \sum_{i=1}^{n} \ell(y_i, u_i) + \frac{\lambda}{2} \|\theta\|^2 + \lambda \sum_{i=1}^{n} \alpha_i \big(u_i - \langle \varphi(x_i), \theta \rangle\big)$

$$= \max_{\alpha \in \mathbb{R}^n} \Big\{ \frac{1}{n} \sum_{i=1}^{n} \min_{u_i \in \mathbb{R}} \{ \ell(y_i, u_i) + n\lambda \alpha_i u_i \} + \min_{\theta \in \mathcal{H}} \Big\{ \frac{\lambda}{2} \|\theta\|^2 - \lambda \sum_{i=1}^{n} \alpha_i \langle \varphi(x_i), \theta \rangle \Big\} \Big\}$$

$$= \max_{\alpha \in \mathbb{R}^n} \frac{1}{n} \sum_{i=1}^{n} \min_{u_i \in \mathbb{R}} \{ \ell(y_i, u_i) + n\lambda \alpha_i u_i \} - \frac{\lambda}{2} \Big\| \sum_{i=1}^{n} \alpha_i \varphi(x_i) \Big\|^2$$ với $\theta = \sum_{i=1}^{n} \alpha_i \varphi(x_i)$,

$$= \max_{\alpha \in \mathbb{R}^n} \frac{1}{n} \sum_{i=1}^{n} \min_{u_i \in \mathbb{R}} \{ \ell(y_i, u_i) + n\lambda \alpha_i u_i \} - \frac{\lambda}{2} \alpha^\top K \alpha$$

$$\max_{\alpha \in \mathbb{R}^n} \frac{1}{n} \sum_{i=1}^{n} \min_{u_i \in \mathbb{R}} \{\ell(y_i, u_i) + n\lambda\alpha_i u_i\} - \frac{1}{2\lambda}\alpha^\top K \alpha$$

với $\theta = \sum_{i=1}^{n} \alpha_i \varphi(x_i)$; tối ưu

Vì hàm $\alpha_i \mapsto \min_{u_i \in \mathbb{R}} \{\ell(y_i, u_i) + n\lambda\alpha_i u_i\}$ lõm, bài toán trên thành tìm cực đại lõm

SGD (Stochastic Gradient Descent)

$$\min_{\theta \in \mathcal{H}} \mathbb{E}\big[\ell(y, \langle \varphi(x), \theta \rangle)\big] + \frac{\lambda}{2}\|\theta\|^2$$

-Khi tối ưu, SGD dẫn đến đệ quy:

$$\theta_t = \theta_{t-1} - \gamma_t\big[\ell'(y_t, \langle \varphi(x_t), \theta_{t-1}\rangle)\varphi(x_t) + \lambda\theta_{t-1}\big]$$

-Khi khởi tạo $\theta_0 = 0$, θ_t là tổ hợp tuyến tính của $\varphi(x_i)$, i = 1, . . . , t

$$\theta_t = \sum_{i=1}^{t} \alpha_i^{(t)} \varphi(x_i),$$

-Với $\alpha^{(0)} = 0$, đệ quy trên trở thành:

$$\alpha_i^{(t)} = (1 - \gamma_t\lambda)\alpha_i^{(t-1)} \text{ for } i \in \{1, \ldots, t-1\}$$ với $$\alpha_t^{(t)} = -\gamma_t \ell'\Big(y_t, , \sum_{i=1}^{t-1} \alpha_i^{(t-1)} k(x_t, x_i)\Big)$$

+Nếu hàm mất mát G-Lipschitz liên tục với: $F(\theta) = \mathbb{E}\big[\ell(y, \langle\varphi(x), \theta\rangle)\big] + \frac{\lambda}{2}\|\theta\|^2$

$$\mathbb{E}\big[F(\bar{\theta}_t)\big] - \inf_{\theta\in\mathcal{H}} F(\theta) \leqslant \frac{G^2R^2}{\lambda t}$$

$$\mathbb{E}\big[\mathcal{R}(f_{\bar{\theta}_t})\big] \leqslant \frac{G^2R^2}{\lambda n} + \inf_{f\in\mathcal{H}} \Big\{\mathcal{R}(f) + \frac{\lambda}{2}\|f\|_{\mathcal{H}}^2\Big\}$$

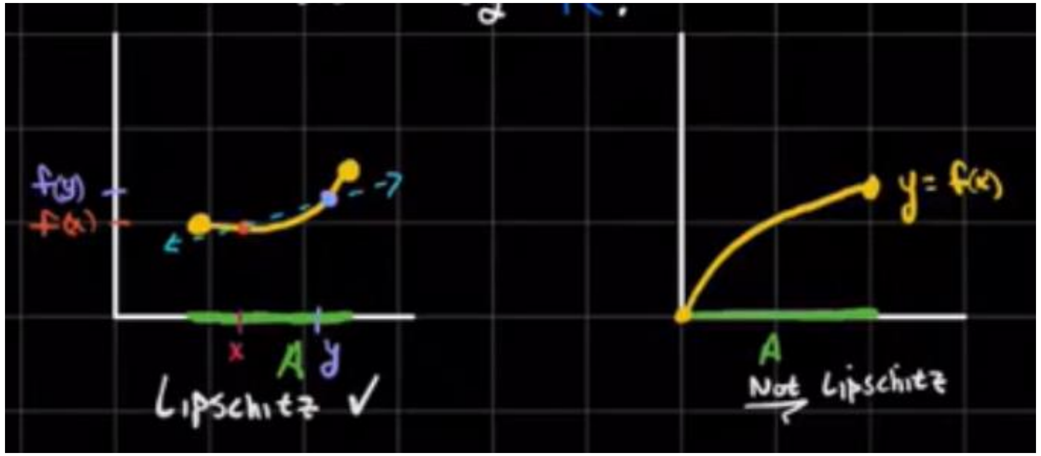
=>Tổng kết 7.4 - "Kernelization" của các thuật toán tuyến tính:
+ học có giám sát, nhiều thuật toán học không giám sát có thể được ''kernelized'' như PCA(Principle component analysis) và CCA(canonical correlation analysis)

Bảo toàn tổng quát– Mất mát Lipschitz liên tục

+Xét hàm mất mát G-Lipschitz, xét cực tiểu $\hat{f}_D^{(c)}$:

$$\min_{f\in\mathcal{H}} \frac{1}{n}\sum_{i=1}^{n}\ell(y_i, f(x_i)) \text{ such that } \|f\|_{\mathcal{H}} \leqslant D$$

$$\min_{f\in\mathcal{H}} \frac{1}{n}\sum_{i=1}^{n}\ell(y_i, f(x_i)) + \frac{\lambda}{2}\|f\|_{\mathcal{H}}^2$$



+ Đặt $\mathcal{R}(f) = \mathbb{E}\big[\ell(y, f(x))\big]$: Rủi ro kỳ vọng, và f* -- Một trong những cực tiểu

-- Giả sử K(x,x) $\leq R^2$

$$\begin{aligned} \mathcal{R}(f) - \mathcal{R}(f^*) &\leqslant \mathbb{E}[|\ell(y, f(x)) - \ell(y, f^*(x))|] \leqslant G\mathbb{E}[|f(x) - f^*(x)|] \\ &\leqslant G\sqrt{\mathbb{E}[|f(x) - f^*(x)|^2]} = G\|f - f^*\|_{L_2(dp(x))} \end{aligned}$$

$\|f\|_{L_2(dp(x))} \leqslant \|\frac{dp}{dx}\|_\infty^{1/2}\|f\|_{L_2(dx)}$ => thay $G\|f - f^*\|_{L_2(dp(x))}$ bởi $G\|\frac{dp}{dx}\|_\infty^{1/2}\|f - f^*\|_{L_2(dx)}$

a) Phân rã rủi ro

+ Vấn đề hạn chế: Dựa trên phân phối Rademacher, sai số ước tính có giới hạn trên là $\frac{2GDR}{\sqrt{n}}$

$$\mathbb{E}\big[\mathcal{R}(\hat{f}_D^{(c)})\big] - \mathcal{R}(f^*) \leqslant \underbrace{\frac{2GDR}{\sqrt{n}}}_{\text{Sai số ước tính}} + \underbrace{G \inf_{\|f\|_{\mathcal{H}} \leqslant D} \|f - f^*\|_{L_2(dp(x))}}_{\text{Sai số xấp xỉ}}$$

+ Tìm cách cực tiểu hóa theo D bằng đối ngẫu Lagrange:

$$\inf_{D \geqslant 0} \frac{2GRD}{\sqrt{n}} + G \inf_{\|f\|_{\mathcal{H}} \leqslant D} \|f - f^*\|_{L_2(dp)}$$

$$\inf_{D\geqslant 0} \frac{2GRD}{\sqrt{n}} + G \inf_{\|f\|_{\mathcal{H}}\leqslant D} \|f - f^*\|_{L_2(dp)}$$

$$= \inf_{D\geqslant 0} \frac{2GBD}{\sqrt{n}} + G \sup_{\lambda\geqslant 0} \inf_{f\in\mathcal{H}} \|f - f^*\|_{L_2(dp(x))} + \sqrt{\lambda}(\|f\|_{\mathcal{H}} - D)$$

$$\leqslant \sup_{\lambda\geqslant 0} \inf_{D\geqslant 0} GD[\frac{2R}{\sqrt{n}} - \sqrt{\lambda}] + 2G\sqrt{\inf_{f\in\mathcal{H}} \{\|f - f^*\|^2_{L_2(dp(x))} + \lambda\|f\|^2_{\mathcal{H}}\}}$$ Áp dụng $a + b \leqslant 2\sqrt{a^2 + b^2}$

$$= \sup_{\lambda\geqslant 0} G\sqrt{\inf_{f\in\mathcal{H}} \{\|f - f^*\|^2_{L_2(dp(x))} + \lambda\|f\|^2_{\mathcal{H}}\}}$$ với $\sqrt{\lambda} \leqslant \frac{2R}{\sqrt{n}}$

$$\leqslant 2G\sqrt{\inf_{f\in\mathcal{H}} \{\|f - f^*\|^2_{L_2(dp(x))} + \frac{4R^2}{n}\|f\|^2_{\mathcal{H}}\}}$$ với $\lambda^* = \frac{4R^2}{n}$

$$A(\lambda, f^*) = \inf_{f \in \mathcal{H}} \{\|f - f^*\|^2_{L_2(dp(x))} + \lambda \|f\|^2_{\mathcal{H}}\}$$

- A -> 0 khi λ -> 0 trong 1 số trường hợp sau:

- Nếu f* trong H, $A(\lambda, f^*) = \lambda \|f^*\|^2_{\mathcal{H}}$

- Nếu f* không trong H, nhưng có thể tiệm cận khi f* là bao hàm L_2(dp(x))

- Biểu thị Π ¯H (f *) là phép chiếu trực giao trong L_2(dp(x)) của f* trên bao hàm H bằng định lý Py-ta-go $A(\lambda, f^*) = A(\lambda, \Pi_{\bar{\mathcal{H}}}(f^*)) + \|f^* - \Pi_{\bar{\mathcal{H}}}(f^*)\|^2_{L_2(dp(x))}$

Tuy nhiên xảy ra ko thể nén được do lựa chọn không gian hàm không đủ lớn.

Vấn đề hiệu chuẩn:

$$\mathbb{E}[\mathcal{R}(\hat{f}_\lambda^{(r)})] - \mathcal{R}(f^*) \leqslant \frac{32G^2R^2}{\lambda n} + \inf_{f\in\mathcal{H}} \{G\|f - f^*\|_{L_2(dp(x))} + \frac{\lambda}{2}\|f\|_{\mathcal{H}}^2\}.$$

- Cực tiểu hóa giới hạn với λ, $\lambda^* = \frac{8RG}{\sqrt{n}}$

$$G \inf_{f\in\mathcal{H}} \{\|f - f^*\|_{L_2(dp(x))} + \frac{8R}{\sqrt{n}}\|f\|_{\mathcal{H}}\} \leqslant 2G\sqrt{\inf_{f\in\mathcal{H}} \{\|f - f^*\|_{L_2(dp)}^2 + \frac{64R^2}{n}\|f\|_{\mathcal{H}}^2\}}$$

- biểu thức bị hạn chế ở 1 số ràng buộc nhưng là 1 cách tối ưu đc sử dụng phổ biến trong thực tế. Tham số Regularization tỉ lệ với R^2/n.

b) Sai số xấp xỉ cho "kernels bất biến liên tục" trên R^d

-Với 1 phân phối dp(x), mục tiêu là tính:

$$A(\lambda, f^*) = \inf_{f \in \mathcal{H}} \|f - f^*\|^2_{L_2(dp(x))} + \lambda \|f\|^2_{\mathcal{H}}$$

-Giả sử $\|f - f^*\|^2_{L_2(dp(x))} \leqslant C \|f - f^*\|^2_{L_2(dx)}$

C = ‖dp/dx‖∞

dp/dx là mật độ của dp(x)

$\|f^*\|_{L_2(dx)}$ hữu hạn

-Đưa ra giới hạn trên $\widetilde{A}(\lambda, f^*) = \inf_{f \in \mathcal{H}} \|f - f^*\|^2_{L_2(dx)} + \lambda \|f\|^2_{\mathcal{H}}$

-Nếu f* € H (best case) => $\widetilde{A}(\lambda, f^*) = \lambda \|f^*\|^2_{\mathcal{H}}$

Xấp xỉ tường minh:

- Đối với kernels bất biến liên tục, $\|f\|_{\mathcal{H}}^2 = \frac{1}{(2\pi)^d}\int_{\mathbb{R}^d} \frac{|\hat{f}(\omega)|^2}{\hat{q}(\omega)} d\omega$

=> $$\widetilde{A}(\lambda, f^*) = \inf_{\hat{f}\in L_2(d\omega)} \frac{1}{(2\pi)^d}\int_{\mathbb{R}^d} \Big[|\hat{f}(\omega) - \hat{f}^*(\omega)|^2 + \lambda\frac{|\hat{f}(\omega)|^2}{\hat{q}(\omega)}\Big] d\omega$$

- Việc tối ưu được thực hiện độc lập với ω, trên là hàm bậc 2

=> Tính đạo hàm theo $\hat{f}(\omega)$

$$0 = 2(\hat{f}(\omega) - \hat{f}^*(\omega)) + 2\lambda\frac{\hat{f}(\omega)}{\hat{q}(\omega)}$$

$$\Leftrightarrow \qquad \hat{f}_\lambda(\omega) = \frac{\hat{f}^*(\omega)}{1+\lambda\hat{q}(\omega)^{-1}}$$

- Đối với hàm mục tiêu, ta có:

$$\widetilde{A}(\lambda, f^*) \;=\; \frac{1}{(2\pi)^d}\int_{\mathbb{R}^d}\left[|\hat{f}^*(\omega)|^2\left(1-\frac{1}{1+\lambda\hat{q}(\omega)^{-1}}\right)\right]d\omega = \frac{1}{(2\pi)^d}\int_{\mathbb{R}^d}\left[|\hat{f}^*(\omega)|^2\frac{\lambda}{\hat{q}(\omega)+\lambda}\right]d\omega.$$

- Khi $\lambda \to 0$, với mỗi ω, $\hat{f}_\lambda(\omega) \to \hat{f}(\omega)$

-Theo định lí hội tụ, $\widetilde{A}(\lambda, f^*) \to 0$ khi $\lambda \to 0$

Không gian Sobolev

- Giả sử: $\frac{1}{(2\pi)^d}\int_{\mathbb{R}^d}(1+\|\omega\|_2^2)^t|\hat{f}^*(\omega)|^2 d\omega$ hữu hạn khi t>0

- Đối với f* có đạo hàm riêng của tích phân bình phương lên đến bậc t, ta có thể xét:

$$\widetilde{A}(\lambda, f^*) \leqslant \frac{1}{(2\pi)^d}\int_{\mathbb{R}^d}(1+\|\omega\|_2^2)^t|\hat{f}^*(\omega)|^2 d\omega \times \sup_{\omega\in\mathbb{R}^d}\left\{\frac{\lambda}{\hat{q}(\omega)+\lambda}\frac{1}{(1+\|\omega\|_2^2)^t}\right\}$$

- Nếu giả sử $\hat{q}(\omega) \propto (1+\|\omega\|_2^2)^{-s}$ (Matern Kernels) với s>d/2 để có RKHS,

- t≥s, f* € H => $\widetilde{A}(\lambda, f^*) = \lambda\|f^*\|_{\mathcal{H}}^2$

- t<s, f* không € H => $\sup_{\omega\in\mathbb{R}^d}\left\{\frac{\lambda}{\hat{q}(\omega)+\lambda}\frac{1}{(1+\|\omega\|_2^2)^t}\right\} \leqslant \sup_{\omega\in\mathbb{R}^d}\left\{\frac{\lambda}{\hat{q}(\omega)^{t/s}\lambda^{1-t/s}}\frac{1}{(1+\|\omega\|_2^2)^t}\right\} = O(\lambda^{t/s}).$

+Áp dụng $a+b \geqslant a^{t/s}b^{1-t/s}$)

Giới hạn xấp xỉ

- Xấp xỉ f* -> ε với RKHS norm nhỏ nhất có thể:

$$A(\lambda, f^*) = \inf_{f \in \mathcal{H}} \left\{ \|f - f^*\|^2_{L_2(dp(x))} + \lambda \|f\|^2_{\mathcal{H}} \right\}$$

Có dạng $c\lambda^{\alpha}$ với α € (0,1) :

$$\inf_{f \in \mathcal{H}} \|f\|^2_{\mathcal{H}} \text{ such that } \|f - f^*\|_{L_2(dp(x))} \leqslant \varepsilon$$

$$= \inf_{f \in \mathcal{H}} \sup_{\mu \geqslant 0} \|f\|^2_{\mathcal{H}} + \mu(\|f - f^*\|^2_{L_2(dp(x))} - \varepsilon^2)$$ Dùng đối ngẫu Lagrange

$$= \sup_{\mu \geqslant 0} \mu A(\mu^{-1}, f^*) - \mu\varepsilon^2 \leqslant \sup_{\mu \geqslant 0} \mu c \mu^{-\alpha} - \mu\varepsilon^2$$

$$\sup_{\mu \geqslant 0} \mu A(\mu^{-1}, f^*) - \mu\varepsilon^2 \leqslant \sup_{\mu \geqslant 0} \mu c \mu^{-\alpha} - \mu\varepsilon^2$$

- μ tối ưu thỏa mãn $(1 - \alpha)c\mu^{-\alpha} = \varepsilon^2$

➔ $\varepsilon^{2(1 - 1/\alpha)} = \varepsilon^{-2(1 - \alpha)/\alpha}$

-Với $\alpha = t/s$ => RKHS norm tỉ lệ với $\varepsilon^{-(1 - \alpha)/\alpha}$ để có được sai số nhỏ hơn $\|f - f^*\|_{L_2(dx)}$

- Xét t =1 và s>d/2 (Sobolev kernel), có norm bậc:

$$\varepsilon^{-(1/\alpha - 1)} = \varepsilon^{-(s-1)} \geqslant \varepsilon^{-d/2+1}$$

->kích thước tăng nhanh theo cấp số mũ

➔ Đây là 1 cách để phát sinh số chiều khác