FRIDAY 07/30: TA SESSION 2 (THEODOR MISIAKIEWICZ)

"BENIGN OVERFITTING IN LINEAR REGRESSION"

- BARTLETT, LONG, LUGOSI, TSIGLER, 2019.

BACKGROUD:

Parametric family of fcts: $F = \{ f(\cdot, \theta) : \theta \in \mathbb{R}^p \}$

Data: $\{(y_i, x_i)\}_{i \in [n]}$ $\quad x_i \in \mathbb{R}^d$ $\quad y_i \in \mathbb{R}$

$x_i \overset{iid}{\sim} \mathbb{P}$, $\quad y_i = f(x_i, \theta_*) + \varepsilon_i$ ← noise e.g. $\varepsilon_i \overset{iid}{\sim} N(0, \sigma_\varepsilon^2)$

Goal: fit the data with F i.e. estimate $\hat{\theta}$ such that $f(\cdot, \hat{\theta})$ is a 'good prediction' on new data points

e.g. $\mathbb{E}_{y_{new}, x_{new}}\left[(y_{new} - f(x_{new}, \hat{\theta}))^2 \right]$ small. ←

$x_{new} \sim \mathbb{P}$ $\quad y_{new} = f(x_{new}, \theta_*) + \varepsilon_{new}$

Classical approach:

$$\hat{\theta}(\lambda) = \underset{\theta \in \mathbb{R}^p}{\operatorname{argmin}} \left\{ \sum_{i=1}^n (y_i - f(x_i, \theta))^2 + \lambda \|\theta\|_2^2 \right\}$$

Test error: $R(\theta_*, \lambda) = \mathbb{E}_x\left[(f(x, \theta_*) - f(x, \hat{\theta}(\lambda)))^2 \right]$ ←

prediction on new data point: noise $\varepsilon \sim N(0, \sigma_\varepsilon^2)$ $\quad y_{new} = f(x_{new}, \theta_*) + \varepsilon_{new}$

$$\mathbb{E}_{x_{new}, y_{new}}\left[(y_{new} - f(x_{new}, \hat{\theta}))^2 \right] = R(\theta_*, \lambda) + \sigma_\varepsilon^2$$ ←

Bias - Variance decomposition : $y_i = f(x, \theta_*) + \varepsilon_i$

noise $\varepsilon = (\varepsilon_1, \dots, \varepsilon_n)$ $\rightarrow$ $\hat{\theta}(\lambda)$ depends on ε

$$\mathbb{E}_\varepsilon[R(\theta, \lambda)] = \mathbb{E}_{x,\varepsilon}\left[\left(f(x, \theta_*) - f(x, \hat{\theta}(\lambda)) \right)^2 \right]$$

$$= \mathbb{E}_x\left[\left(f(x, \theta_*) - \mathbb{E}_\varepsilon[f(x, \hat{\theta}(\lambda))] \right)^2 \right] + \mathbb{E}_{x,\varepsilon}\left[\left(f(x, \hat{\theta}(\lambda)) - \mathbb{E}_\varepsilon[f(x, \hat{\theta}(\lambda))] \right)^2 \right]$$

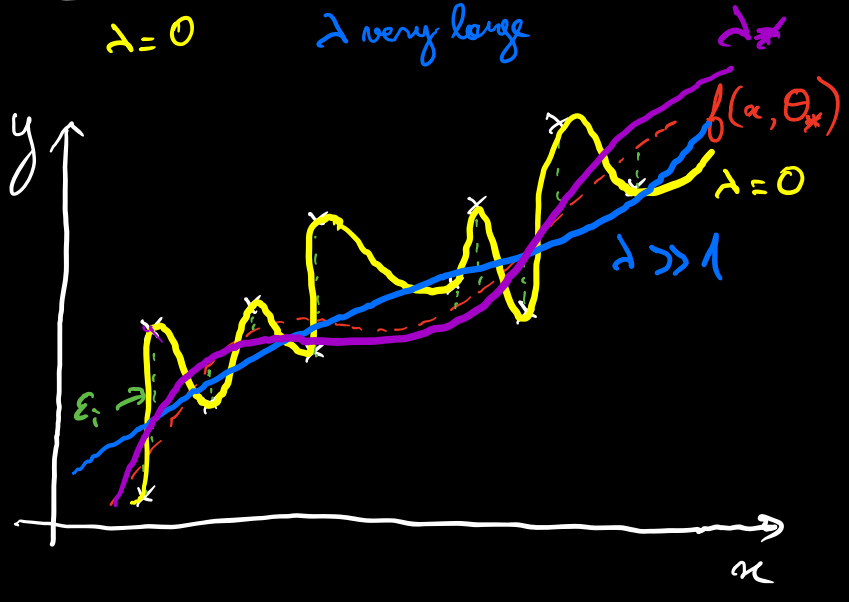
$$= \text{BIAS}(\lambda) + \text{VARIANCE}(\lambda)$$

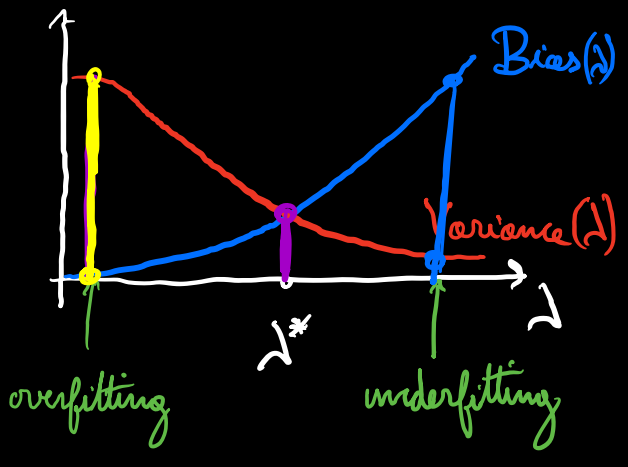
$$= \mathbb{E}_x\left[\left\{ \text{bias}\left(\mathbb{E}_\varepsilon[f(x, \hat{\theta})] \right) \right\}^2 \right] \qquad = \mathbb{E}_x\left[\text{Var}_\varepsilon\left(f(x, \hat{\theta}) \right) \right]$$

CLASSICAL PICTURE:

$$\hat{\theta}(\lambda) = \underset{\theta \in \mathbb{R}^p}{\operatorname{argmin}} \left\{ ER(\theta) + \lambda \|\theta\|_\ell^2 \right\}$$

$$\Leftrightarrow \left\{ \|\theta\| \leq C(\lambda) \right\}$$





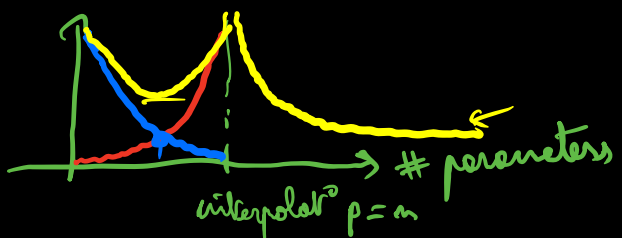
Trade-off between BIAS and VARIANCE

$\rightarrow$ need to carefully choose regularization parameter λ in order to control the complexity of the fitted model $f(x, \hat{\theta}(\lambda))$

$\Rightarrow$ too complex : small bias, large variance (overfitting)

$\Rightarrow$ too simple : large bias, small variance (underfitting)

MODERN APPROACH:



Train until interpolation: $f(x_i, \hat{\theta}) = y_i = f(x_i, \theta_*) + \epsilon_i$

→ IMPLICIT REGULARIZATION: choose interpolator ←

but no explicit control on the bias-variance trade-off

(no control on the complexity of the function)

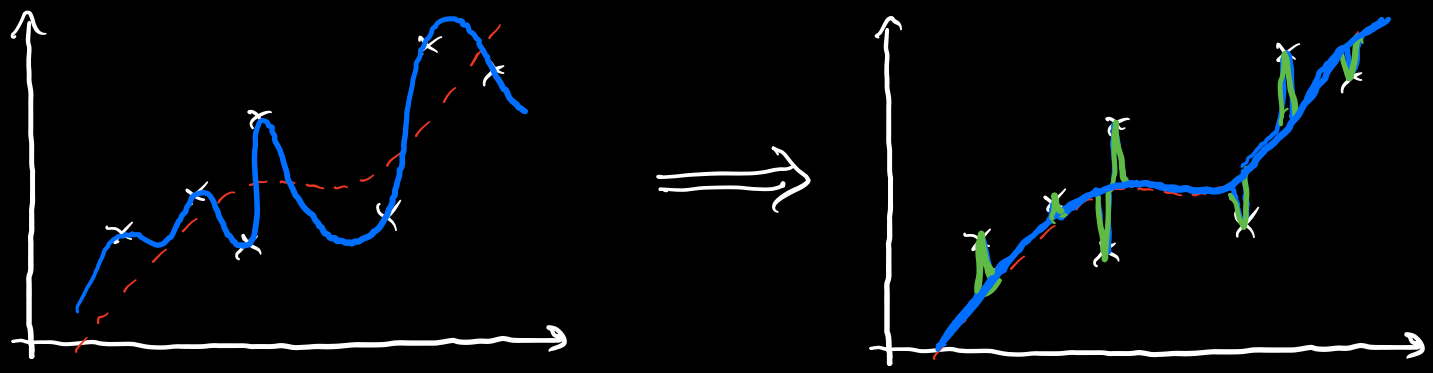
HERE: I will present a scenario where the balance between the bias and the variance is achieved not by carefully tuning an explicit parameter but by a novel phenomena:

SELF-INDUCED REGULARIZATION

≠ implicit regularization: choose an interpolator $\hat{\theta}$

given an interpolator $\hat{\theta}$: self-induced regularization shows that interpolation does not hurt generalization (test error)

Example of **BENIGN OVERFITTING PHENOMENA**



Expected picture

Sometimes "benign"

Benign overfitting:

estimator : $\hat{f}(x, \hat{\theta}) = \hat{f}_0(x, \hat{\theta}_0) + \Delta(x)$

$\hat{f}_0(x, \hat{\theta}_0)$: good for prediction because smooth

$\Delta(x)$: spikes that are useful for interpolation but do not harm prediction because $\mathbb{E}_x[\Delta(x)^2] \ll 1$.

RIDGE REGRESSION IN LINEAR MODELS:

Setting: • **Data:** $\{(y_i, x_i)\}_{i \le n}$ $x_i \in \mathbb{R}^p$, $y_i \in \mathbb{R}$

• $\mathbb{E}[x_i] = 0$ ✓ $\mathbb{E}[x_i x_i^T] = \Sigma$ ✓

$x_i = \Sigma^{\frac{1}{2}} z_i$ where z_i is c-sub-Gaussian

• x_i are iid, $y_i = \langle \theta_*, x_i \rangle + \varepsilon_i$

noise ε_i independent $\mathbb{E}[\varepsilon_i] = 0$, $\mathbb{E}[\varepsilon_i^2] = \sigma_\varepsilon^2$

Fit a linear model: $f(x, \theta) = \langle x, \theta \rangle$ $\theta \in \mathbb{R}^p$

Ridge regression:

$$\hat{\theta}_\lambda(X, y) = \underset{\theta \in \mathbb{R}^p}{\operatorname{argmin}} \left\{ \|y - X\theta\|_2^2 + \lambda \|\theta\|_2^2 \right\}$$

$$X = \begin{matrix} \overset{p}{\longrightarrow} \\ \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix} \end{matrix} \quad \begin{matrix} \uparrow n \\ \downarrow \end{matrix} \qquad y = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}$$

$$= X^T S y \quad \text{where } S = (XX^T + \lambda \mathrm{Id})^{-1}$$

Test error: $R(\hat{\theta}; X, y, \lambda) = \mathbb{E}_x\left[\left(f(x, \theta_*) - f(x, \hat{\theta}_\lambda(X, y)) \right)^2 \right]$

$$= \mathbb{E}_x\left[\langle x, \theta_* - \hat{\theta}_\lambda(X, y) \rangle^2 \right]$$

$$= \|\hat{\theta}_\lambda(X, y) - \theta_*\|_\Sigma^2$$

$$= \mathbb{E}\left[\langle \theta_0 - \hat{\theta}_\lambda, \underbrace{x x^T}_{\Sigma} (\theta_0 - \hat{\theta}) \right]$$

$\left(\|\theta\|_\Sigma^2 := \langle \theta, \Sigma \theta \rangle \right)$

$$R(\hat{\theta}, X, \lambda) = \mathbb{E}_\varepsilon \left[\| \hat{\theta}_\lambda(X, y) - \theta_* \|_\Sigma^2 \right]$$

$$= \mathbb{E}_\varepsilon \left[\| X^T S X \theta_* + X^T S \varepsilon - \theta_* \|_\Sigma^2 \right]$$

$$= B(X, \lambda) + V(X, \lambda)$$

where $B(X, \lambda) = \| (Id - X^T S X) \theta_* \|_\Sigma^2$

$$V(X, \lambda) = \sigma^2 \operatorname{Tr}(S X \Sigma X^T S)$$

SELF-INDUCED REGULARIZATION:

WLOG: assume $\Sigma = \operatorname{diag}(\lambda_1, \lambda_2, \dots, \lambda_p)$

$\lambda_1 \geq \lambda_2 \geq \dots$

Consider top k eigenspace (corresponding top $\lambda_1 \dots \lambda_k$)

write $x^T = [\, x_0^T \,,\, x_+^T \,]$

x_0^T: k first coordinates; x_+^T: $p-k$ last coordinates

$$X = \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix} = \begin{matrix} \uparrow \\ n \\ \downarrow \end{matrix} \begin{bmatrix} \overleftarrow{\quad k \quad} & \overleftarrow{\quad p-k \quad} \\ X_0 & X_+ \end{bmatrix}$$

$$\Sigma = \begin{matrix} \uparrow \\ k \\ \downarrow \end{matrix} \left(\begin{array}{c|c} \Sigma_0 & \\ \hline & \Sigma_+ \end{array} \right)$$

$\leftarrow p-k \rightarrow$

$\theta = (\theta_0, \theta_+)$

$\theta_0 = (\theta_1, \dots, \theta_k)$ $\quad$ $\theta_+ = (\theta_{k+1}, \dots, \theta_p)$

$$XX^T = X_0 X_0^T + X_+ X_+^T$$

$$\left(\frac{\lambda_{max}(X_+X_+^T)}{\lambda_{min}(X_+X_+^T)} \leq L\right)$$

$$[X_0\ X_+]\begin{bmatrix} X_0^T \\ X_+^T \end{bmatrix} \text{ imagine } = \gamma M \quad \text{with} \quad \frac{1}{c}\mathrm{Id} \preccurlyeq M \preccurlyeq c\,\mathrm{Id}$$

Recall $\hat{\theta}_{\lambda=0}(X,y) = X^T S\, y \simeq X^T (X_0 X_0^T + \gamma \mathrm{Id})^{-1} y$ (+λId)

k 1st coordinates: $\hat{\theta}_0 = X_0^T (X_0 X_0^T + (\lambda+\gamma)\mathrm{Id})^{-1} y$

$\longrightarrow$ the same as

$$\hat{\theta}_0 = \underset{\theta_0 \in \mathbb{R}^k}{\operatorname{argmin}} \left\{ \|y - X_0\theta_0\|_2^2 + \gamma \|\theta_0\|_2^2 \right\}$$

($\langle x_0, \theta_0 \rangle$, $(\lambda+\gamma)$)

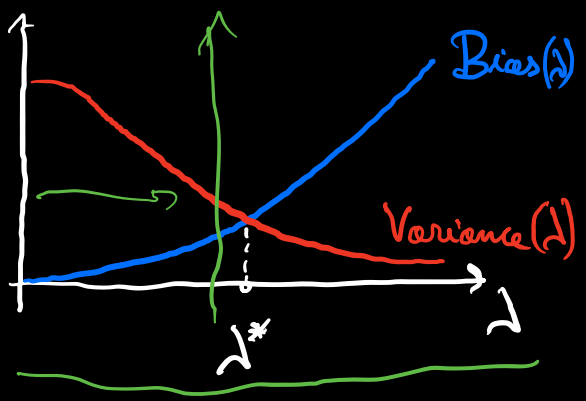
self-induced regularization

Heuristic: X_0 "signal part" of the features

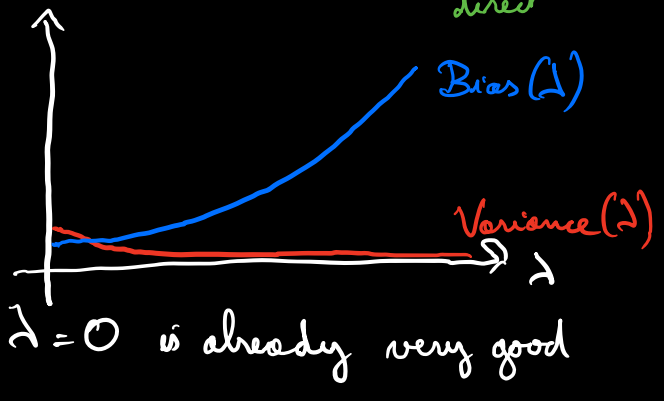
X_+ "noise part" of the features

$\longrightarrow$ noise part act as an effective ridge regularization

Effective regularization: $\lambda \to \lambda_{eff} = \lambda + \gamma$, $\gamma = \sum_{i>k} \lambda_i$ low variance direct°



$\longrightarrow$



$\lambda = 0$ is already very good

Main result:

Recall: $S = (\lambda Id + XX^T)^{-1} = (\lambda Id + X_+ X_+^T + X_0 X_0^T)^{-1}$

$S_+ = (\lambda Id + X_+ X_+^T)^{-1}$ $\qquad$ $S = (S_+^{-1} + X_0 X_0^T)^{-1}$

γM

Thm [Bartlett et al., '19] $\qquad$ Assume there exists k:

Assume $\dfrac{\lambda_{max}(S_+)}{\lambda_{min}(S_+)} \leq L$ with high probability (whp)

Then whp:

$$Bias(\lambda) \lesssim L^4 \left[\|\theta_0^*\|^2_{\Sigma_0^{-1}} \left(\frac{\lambda + \sum_{i>k} \lambda_i}{n} \right)^2 + \|\theta_+^*\|^2_{\Sigma_+} \right]$$

$$Variance(\lambda) \lesssim \sigma_\varepsilon^2 L^2 \left[\frac{k}{n} + \frac{n \sum_{i>k} \lambda_i^2}{(\lambda + \sum_{i>k} \lambda_i)^2} \right]$$

Rmks: * Matching lower bounds up to multiplicative constants.

* In particular, interpolators $\lambda = 0$ are near optimal.

* Does not fit θ_+^* at all:

prediction $f(x, \hat{\theta}) = \underbrace{\langle \hat{\theta}_0, x_0 \rangle}_{\hat{f}_0(x)} + \underbrace{\langle \hat{\theta}_+, x_+ \rangle}_{\Delta(x)}$

prediction component $\Rightarrow$ fit well θ_0^*

interpolation component

* Small variance: need to choose k so that

(1) $k \ll n$

(2) $\dfrac{\left(\sum_{i>k} \lambda_i\right)^2}{\sum_{i>k} \lambda_i^2} \gg n$

E.g. $d_{k+1} := \#$ eigenvalues of Σ in $\left[\dfrac{\lambda_{k+1}}{\epsilon}, \lambda_{k+1}\right]$

and assume eigenvalues $< \dfrac{\lambda_{k+1}}{2}$ are negligible.

$\Rightarrow$ (2) requires $\dfrac{\left(d_{k+1} \cdot \lambda_{k+1}\right)^2}{d_{k+1} \cdot \lambda_{k+1}^2} \gg n \Rightarrow \boxed{d_{k+1} \gg n}$

and we have

$$\text{Variance} \lesssim \frac{k}{n} + \frac{n}{d_{k+1}}$$

- $\frac{k}{n}$: Variance of fitting θ_0^*; parametric rate of fitting θ_0^*
- $\frac{n}{d_{k+1}}$: variance of fitting the rest; effect of overfitting $\Rightarrow$ negligible

TO GET BENIGN OVERFITTING HERE: Σ $p \gg n$

1) Overparametrization: number of directions in parameter space that are unimportant for prediction $\gg$ sample size

2) Smallest eigenvalues of Σ must decay slowly.

$$d_{k+1} \gg n$$

Self-induced regularization: (if x satisfies some small ball properties)

Lemma: $\exists\, c > 0$ s.t. with proba at least $1 - 2\exp(-\frac{n}{c})$

$$\frac{\lambda_{max}(X_t X_t^T)}{\lambda_{min}(X_t X_t^T)} \leq c \left(\frac{\sum_{i \geq k} \lambda_i + n \lambda_{k+1}}{\sum_{i \geq k} \lambda_i - c\, n \lambda_{k+1}} \right) \leq L$$

Properties of Σ are crucial in this setting.

$n \leq \frac{1}{2c} \frac{\sum_{i \geq k} \lambda_i}{\lambda_{k+1}}$

$n \ll \frac{d_{k+1} \lambda_{k+1}}{\lambda_{k+1}}$

Andrea's lecture: Kernel Ridge Regression

$$K = \left(h(\langle x_i, x_j \rangle) \right)_{i,j \in [n]}$$

fit a fct: $\hat{f}_\lambda(x; \hat{a}) = \sum_{i=1}^{n} \hat{a}_i(\lambda)\, h(\langle x_i, x \rangle)$

where $\hat{a} = (K + \lambda \mathrm{Id})^{-1} y$

Here: self induced regularization comes from non polynomial part of the kernel fct:

$$h(x) = \sum_{k=0}^{\infty} \frac{h^{(k)}(0)}{k!} x^k$$

$$K = \sum_{k=0}^{\infty} \frac{h^{(k)}(0)}{k!} Q_k \quad \text{where} \quad Q_k = \left(\langle x_i, x_j \rangle^k \right)_{i,j \in (n)}$$

When if $d^\ell \ll n \ll d^{\ell+1}$,

$$\left\| \sum_{k=\ell+1}^{\infty} \frac{h^{(k)}(0)}{k!} Q_k - \gamma \, \mathrm{Id} \right\|_{op} = o_d(1).$$

$K \simeq K_{\leq \ell} + \underbrace{\gamma \, \mathrm{Id}}_{\text{self-induced regularization}}$

$\Rightarrow$ leads also to benign overfitting.

[Bartlett et al] result: applies also to $\rho = \infty$

→ x: could be the vector of an orthonormal basis of $\mathcal{H}$.

→ however not subgaussian: requires very different techniques

Nonetheless:

Kernel: eigenvalues

$$\exists_k \simeq d^{-k}$$ with degeneracy $O_d(d^k)$

↳ associated to polynomials of degree k.

$\Rightarrow$ to fit all degree ℓ polynomials: $\frac{k}{n} + \frac{n}{d_{k+1}}$

"k" $= \sum_{s=0}^{\ell} d^s \simeq d^\ell$

"d_{k+1}" $\simeq d^{\ell+1}$

$$\text{Variance} \simeq \frac{d^\ell}{n} + \frac{n}{d^{\ell+1}}$$

$$+ \quad \text{Bias} \simeq \| P_{>\ell} f_* \|_{L^2}^2$$

$$K(x_i, x_j) = h(\langle x_i, x_j \rangle)$$

$$K_n = \left(K(x_i, x_j)\right)_{i,j \in [n]}$$

$$h(u) = \sum_{k=0}^{\infty} \frac{h^{(k)}(0)}{k!} u^k$$

$$K_n = \sum_{k=0}^{\infty} \frac{h^{(k)}(0)}{k!} Q_k \qquad Q_k = \left(\langle x_i, x_j \rangle^k\right)_{i,j \in [n]}$$

$$n \ll d^{k+1}$$

$$\left\| \sum_{k=k+1}^{\infty} \frac{h^{(k)}(0)}{k!} Q_k - \gamma \, \mathrm{Id} \right\|_{op} \ll 1$$

$$\hat{\theta} = (K + \lambda \mathrm{Id})^{-1} y$$

$$\simeq (K_{\leq k} + (\lambda + \gamma)\mathrm{Id})^{-1} y$$

$\longrightarrow$ able to fit $P_{\leq k} f_*$ and nothing else

$$f_*(a) = \langle a, \theta_* \rangle \longrightarrow p = \infty$$

$$h(\langle x, y \rangle) = \sum_{i=1}^{\infty} \lambda_i \, \Phi_i(x) \, \Phi_i(y)$$

$\{\Phi_i\}$ orthonormal basis $L^2(\mathbb{P})$

$$f_*(x) = \sum_{i=1}^{\infty} \sqrt{\lambda_i} \, \Phi_i(x) \, \theta_i^*$$

$$= \langle a, \theta_i^* \rangle$$

$$\underline{a = \left(\sqrt{\lambda_i} \, \Phi_i(x) \right)_{i \geq n}}$$

$$Q_k = \left(\langle x_i, x_j \rangle^k \right)_{i,j \in [n]} \qquad x_i \sim \mathrm{Unif}(S^{d-1}(1))$$

$Q_k = \mathrm{Id} + \Delta \longrightarrow$ off diagonal elements of Q_k

$\langle x_i, x_j \rangle \leq \frac{1}{\sqrt{d}}$ whp.

$$\max_{i \neq j} \langle x_i, x_j \rangle \leq \frac{1}{d^{\frac{1}{2}-\varepsilon}}$$

$$\|\Delta\|_F^2 = \sum_{i \neq j} \langle x_i, x_j \rangle^{2k}$$

$$\leq n^2 \, d^{-k + \varepsilon 2k} \to 0$$

$$n \leq d^{\frac{k}{2}} \leftarrow \text{Frobenius}$$

$$n \leq d^{k - \delta}$$

$$n \leq d^{\quad} \leftarrow \text{complicated}$$

$$d^\ell \ll n \ll d^{\ell+1}$$

$\longrightarrow$ fit exactly $P_{\leq \ell} f_*$

$\longrightarrow$ even $\lambda = 0$

$h \longrightarrow \lambda_i > 0$ for any eigenfct of polynomials $\leq \ell$

$\lambda = 0 \longrightarrow h$ not a polynomial $d^\circ \leq \ell$