Lecture 6: Kernel methods

Last week: Lazy/linear regime

$\Rightarrow$ optimization regime where optimization is provably tractable

$$\hat{R}_n(\theta_t) \leq \hat{R}_n(\theta_0)\, e^{-c_0 t}$$

Success???

↳ what we actually care about: generalization (test) error $R(\theta_t)$

What is the performance of NNs trained in this regime?
(and in particular, can it explain the success of NNs in practice?)

In the lazy/linear regime, we saw that we can effectively replace the NN by its linearization around θ_0:

$$f_{lin}(x;\theta) = f(x;\theta_0) + \langle \theta - \theta_0, \nabla_\theta f(x;\theta_0) \rangle$$

Denote $a = \theta - \theta_0$

From lecture 2, GD converges to

$$\hat{a} = \operatorname{argmin} \left\{ \|a\|_2 : \langle a, \nabla_\theta f(x_i;\theta_0) \rangle = y_i - f(x_i;\theta_0) \right\}$$

and the model obtained is

$$\hat{f}(x) = f(x;\theta_0) + \langle \hat{a}, \nabla_\theta f(x;\theta_0) \rangle$$

We want to understand: What is the test error of $\hat{f}$?

In fact, this is a special example of a "linear" or "kernel" method also called "kernel machines"

↳ methods that were developped in the 80 - 90's, which were state-of-the-art in ML before being replaced by NNs in the 2010s (disappointing !?!?)

This connection between NNs trained by GD and kernel methods has spurred renewed interest in kernel methods

Are NNs simply glorified kernel methods? (Short answer: no !!!)

Lots of work since 2018 on understanding the fine grained performance of kernel methods. In particular

- interpolating solution
- non-monotonic behavior in learning curve

(both not covered by classical works on kernels)

Goal for today's lecture:

* (Gentle) introduction to kernel methods
* Sharp analysis of kernel ridge regression
* Limitations of kernel methods

$$f_{\text{lin}}(x;\theta) = f(x;\theta_0) + \langle \theta - \theta_0, \nabla_\theta f(x;\theta_0) \rangle$$

offset, not learned (random)

"featurization of the data"

Linear model: $\Phi : \mathbb{R}^d \to \mathbb{R}^p$, $x \mapsto \Phi(x)$ | embedding data in a higher dimensional space

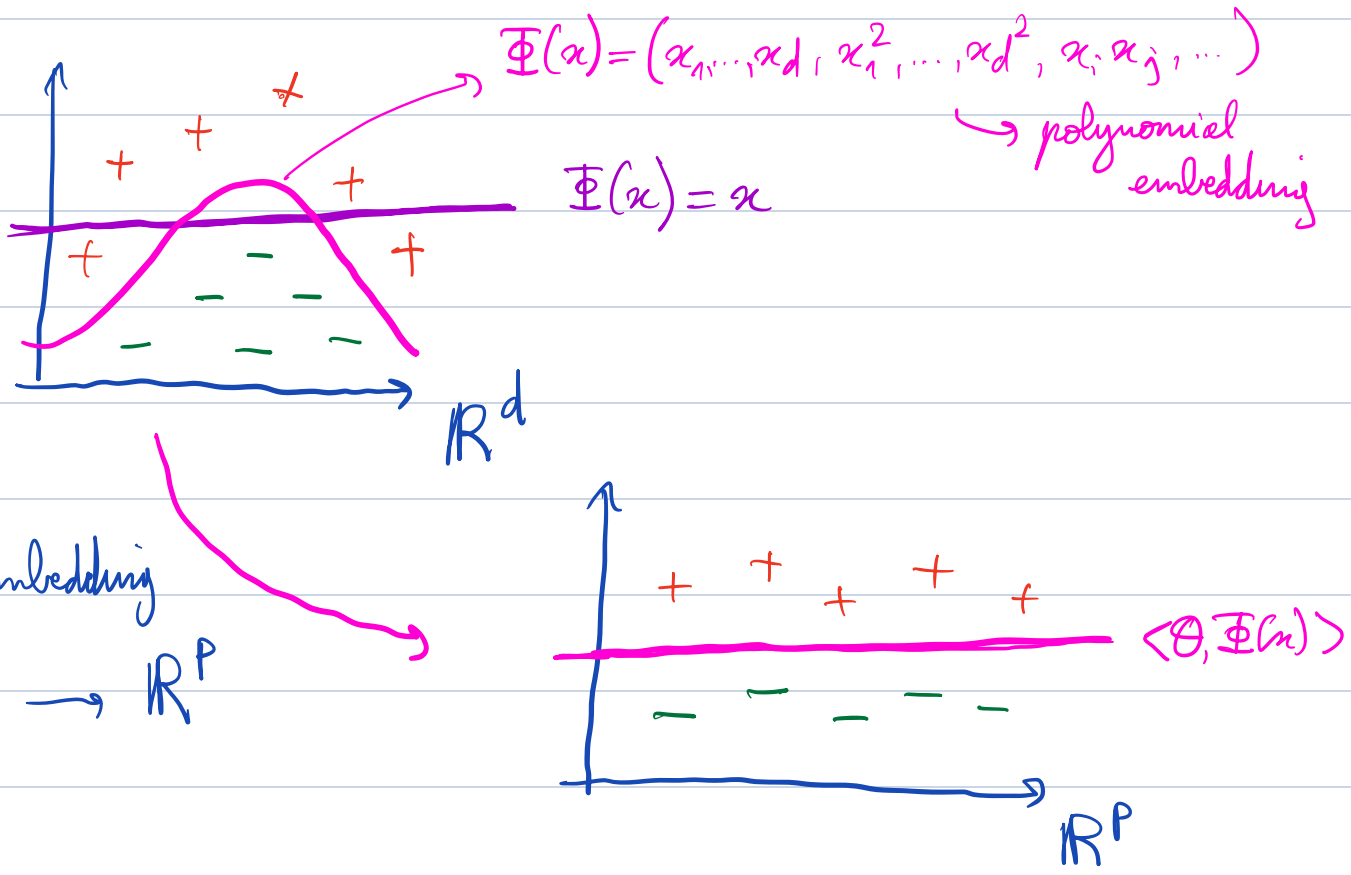
Model: $f(x;\theta) = \langle \theta, \Phi(x) \rangle$ parameter $\theta \in \mathbb{R}^p$

↳ generalization of standard linear model $x \mapsto \langle \theta, x \rangle$

→ here still "linear model": linear in the parameters but non linear in the data x

Why this is interesting?

E.g. (x_i, y_i) $y_i \in \{\pm 1\}$ predictor sign $(\langle \theta, \Phi(x) \rangle)$



→ can fit way more fcts

→ in fact, typically $p = \infty$ and we can fit any function (if we have enough data) "universal"

Kernel methods

General definition:

* $(F, \langle \cdot, \cdot \rangle_F)$ "feature space" with some inner product $\langle \cdot, \cdot \rangle_F$
  norm $\|f\|_F = \langle f, f \rangle_F^{1/2}$ (examples below)

* $\Phi : \mathcal{X} \to F$ "featurization map" that embeds data x in $\Phi(x) \in F$

* $\mathcal{H} = \left\{ h_\theta : \mathcal{X} \to \mathbb{R} \quad h_\theta(x) = \langle \theta, \Phi(x) \rangle : \theta \in F \right\}$

↳ space of functions which inherit scalar product/norm from F

$$\langle h_\theta, h_{\theta'} \rangle_{\mathcal{H}} = \langle \theta, \theta' \rangle \qquad \|h_\theta\|_{\mathcal{H}} = \|\theta\|_F$$

Examples (1) $F = \mathbb{R}^d$ with standard euclidean scalar product

$$\langle u, v \rangle = u_1 v_1 + \dots + u_d v_d$$

$\Phi(x) = x$

$\Rightarrow$ standard linear model $h_\theta(x) = \langle x, \theta \rangle \qquad \|h_\theta\|_{\mathcal{H}} = \|\theta\|_2$

(2) $F = \mathbb{R}^{k+1}$ with standard euclidean scalar product

$\Phi(x) = (1, x, x^2, \dots, x^k)$

$\Rightarrow$ standard polynomial regression model

$$h_\theta(x) = \theta_1 + \theta_2 x + \dots + \theta_{k+1} x^k$$

(3) (Ω, μ) probability space

$$L^2(\Omega, \mu) = \{ a : \Omega \to \mathbb{R} : \int a(\omega)^2 \, \mu(d\omega) < \infty \}$$

(space of squared integrable fcts on Ω)

$$\langle a, b \rangle_{L^2} = \int a(\omega) \, b(\omega) \, \mu(d\omega)$$

$$\| a \|_{L^2}^2 = \int a(\omega)^2 \, \mu(d\omega)$$

$$\Phi : \mathcal{X} \to L^2(\Omega, \mu) \qquad x \mapsto \{ \omega \mapsto \sigma(x; \omega) \}$$

$$h_a(x) = \langle a, \Phi(x) \rangle_{L^2} = \int a(\omega) \, \sigma(x; \omega) \, \mu(d\omega)$$

$$\| h_a \|_{\mathcal{H}} = \| a \|_{L^2}$$

$\mathcal{H}$ = space of ∞ width 2-layer NNs with $\| a \|_{L^2} < \infty$

Rmk: $(\mathcal{H}, \langle \cdot, \cdot \rangle_{\mathcal{H}})$ forms a "Reproducing kernel Hilbert space" (RKHS)

which is a space uniquely defined by a positive semi-definite (PSD) kernel $K: \mathcal{X} \times \mathcal{X} \to \mathbb{R}$

[i.e.

$$\forall n \in \mathbb{R}, \ \forall x_1, \dots, x_n \in \mathcal{X}, \quad \forall a_1 \dots a_n \in \mathbb{R}$$

$$\sum_{i,j=1}^{n} a_i a_j K(a_i, a_j) \geq 0$$

]

In the above case: $K(x, x') = \langle \Phi(x), \Phi(x') \rangle_{\mathcal{F}}$

(e.g. $K(x, x') = \int \sigma(x; \omega) \sigma(x'; \omega) \mu(d\omega)$)

Those are two equivalent ways of defining $(\mathcal{H}, \langle \cdot, \cdot \rangle_{\mathcal{H}})$

Given K, define $\Phi(x) = \{ z \mapsto K(x, z) \} \in \mathcal{H}$

$K(x, x') = \langle \Phi(x), \Phi(x') \rangle_{\mathcal{H}}$

$\langle h, \Phi(x) \rangle_{\mathcal{H}} = h(x)$ "reproducing property"

(I think defining RKHS through featurizat° map is the most illuminating. ⚠ some technical difficulties that I skipped

⚠ Given $\mathcal{H}$ or K: there are infinite number of possible featurizat° $(\Phi, \mathcal{F})$ which yield some $\mathcal{H}$, but not necessarily equivalent when doing approximat°s etc.

Kernel methods

$$\hat{\theta} = \underset{\theta \in \mathcal{F}}{\operatorname{argmin}} \left\{ \frac{1}{n} \sum_{i=1}^{n} \ell(y_i, \langle \theta, \phi(x_i) \rangle_{\mathcal{F}}) + \lambda \|\theta\|_{\mathcal{F}}^2 \right\}$$

$\Leftrightarrow$

$$\hat{h} = \underset{h \in \mathcal{H}}{\operatorname{argmin}} \left\{ \frac{1}{n} \sum_{i=1}^{n} \ell(y_i, h(x_i)) + \lambda \|h\|_{\mathcal{H}}^2 \right\}$$

Is it tractable? $\Phi(x_i) \rightarrow$ potentially infinite dimensional
optim θ : ∞-te dimensional problem

[doesn't matter if $\mathcal{H}$ is way more expressive than standard linear models if we can't train them efficiently]

"Kernel trick": in fact, reduces to n-dimensional convex problem!
↳ this is the reason kernel methods become so popular

Thm ["Representer theorem"] For any loss ℓ (doesn't need to be convex!) solut° $\hat{\theta}$ can be written as

$$\hat{\theta} = \sum_{i=1}^{n} a_i \Phi(x_i)$$

Rmk: To find $\hat{\theta}$, it suffices to restrict $\theta \in \text{span}\{\Phi(x_i) : i \in [n]\}$

↳ n-dim linear subspace

Recall $K(x, x') = \langle \Phi(x), \Phi(x') \rangle_{\mathcal{F}}$

Denote $k_n(x) = (K(x, x_1), \dots, K(x, x_n)) \in \mathbb{R}^n$

$$K_n = \left(K(x_i, x_j)\right)_{i,j=1}^{n}$$

"Empirical kernel matrix"

Then

$$\hat{a} = \underset{a \in \mathbb{R}^n}{\operatorname{argmin}} \left\{ \frac{1}{n} \sum_{i=1}^{n} \ell(y_i, a^T k_n(x_i)) + \lambda a^T K a \right\}$$

→ If ℓ convex then this is a convex optimization problem over n variables

Solution $\hat{\theta} = \sum_{i=1}^{n} \hat{a}_i \Phi(x_i)$

$$\hat{h}(x) = \langle \hat{\theta}, \Phi(x) \rangle = \sum_{i=1}^{n} \hat{a}_i K(x, x_i)$$

Hence: for training and prediction, we never need to compute $\Phi(x)$, just the inner-product $K(x, x') = \langle \Phi(x), \Phi(x') \rangle_{\mathcal{F}}$ which is often easy!

This is known as the "kernel trick".

Proof: Denote $V = \text{span}\{\Phi(x_i) : i \in [n]\}$ linear subspace in F

$F = V \oplus V_\perp$ $\quad V_\perp = \{\theta \in F : \langle \theta, v \rangle_F = 0 \text{ for all } v \in V\}$

$= \text{orthogonal space}$

$\theta = \theta_{/\!/} + \theta_\perp$ $\quad \|\theta\|_F^2 = \|\theta_{/\!/}\|_2^2 + \|\theta_\perp\|_2^2$

$$\min_{\theta_{/\!/},\, \theta_\perp} \frac{1}{n} \sum_{i=1}^{n} \ell\left(y_i, \langle \theta_{/\!/} + \theta_\perp, \phi(x_i) \rangle\right) + \lambda \|\theta_{/\!/}\|_F + \lambda \|\theta_\perp\|_F^2$$

$\theta_\perp$ only appears here so $\theta_\perp = 0$ to minimize

□

Kernel Ridge Regression

A popular example is KRR: $\ell(y, \hat{y}) = (y - \hat{y})^2$

$$\hat{h} = \underset{h \in \mathcal{H}}{\operatorname{argmin}} \left\{ \sum_{i=1}^{n} (y_i - h(x_i))^2 + \lambda \|h\|_{\mathcal{H}}^2 \right\}$$

$$\hat{h} = \sum_{i=1}^{n} \hat{a}_i K(x, x_i) \qquad k_n(x) = \left(K(x, x_i) \right)_{i=1}^{n}$$

$$K_n = \left(K(x_i, x_j) \right)_{i,j=1}^{n}$$

$$\|h\|_{\mathcal{H}}^2 = a^T K_n a \qquad y = (y_1, \dots, y_n)$$

$$\hat{a} = \underset{a \in \mathbb{R}^n}{\operatorname{argmin}} \left\{ \| y - K_n a \|_2^2 + \lambda a^T K_n a \right\}$$

$$= (K_n^2 + \lambda K_n)^{-1} K_n y \overset{\uparrow}{=} (K_n + \lambda I_n)^{-1} y$$

KKT condition

(assume K_n is full rank)

$$\hat{h}(x) = k_n(x)^T (K_n + \lambda I_n)^{-1} y$$

Rmk: ["linear regression on Hilbert space"] $\hat{h}(x) = \langle \hat{\theta}, \Phi(x) \rangle$

$$\hat{\theta} = \Phi(X)^T \left(\Phi(X) \Phi(X)^T + \lambda I_n \right)^{-1} y$$

$$\Phi(X) = \begin{bmatrix} - \phi(x_1) - \\ \vdots \\ - \phi(x_n) - \end{bmatrix} \in \mathbb{R}^{n \times p}$$

Rmk: [Interpolating solution] $\lambda \searrow 0$

$$\hat{h} = \operatorname{argmin} \left\{ \| h \|_{\mathcal{H}} : h(x_i) = y_i \quad i \le n \right\}$$

$$\hat{h}(x) = k_n(x) \, K_n^{-1} \, y$$

Diagonalization of the kernel

The performance of kernel methods depend crucially on the eigendecomposition of the kernel

$x \sim (\mathcal{X}, \nu)$ $\quad L^2(\mu) = \{ f : \mathcal{X} \to \mathbb{R}, \ \|f\|_{L^2}^2 = \int f(x)^2 \nu(dx) < \infty \}$

Can associate to K a linear operator $\mathbb{K} : L^2(\nu) \to L^2(\nu)$

$$(\mathbb{K} f)(x) = \int K(x, x') f(x') \nu(dx')$$

Assume $\mathbb{E}_x[K(x,x)] < \infty$ $\quad$ (K is "trace class")

Then by spectral theorem over bounded linear operator

$$\mathbb{K} = \sum_{j=1}^{\infty} \lambda_j \phi_j \phi_j^*$$

$\lambda_1 \geq \lambda_2 \geq \dots$ are the eigenvalues (non-negative)

$\phi_1, \phi_2, \dots$ are the eigenfunctions $\quad \mathbb{K} \phi_j = \lambda_j \phi_j$

$\langle \phi_j, \phi_k \rangle_{L^2(\nu)} = \delta_{j,k}$

In terms of the kernel function, this implies

$$K(x, x') = \sum_{j=1}^{\infty} \lambda_j \phi_j(x) \phi_j(x')$$

This suggests a "canonical" way of constructing an embedding:

$$\Phi(x) = \left(\lambda_j^{1/2} \phi_j(x) \right)_{j=1}^{\infty}$$

$\Phi(x) \in (\ell_2, <,>_{\ell_2})$ the Hilbert space of squared summable sequences

$$\ell_2 = \left\{ (a_1, a_2, \dots) : \sum_{i=1}^{\infty} a_i^2 < \infty \right\}$$

$$\langle a, b \rangle_{\ell_2} = \sum_{i=1}^{\infty} a_i b_i$$

$$\mathcal{H} = \left\{ h_\theta(x) = \langle \theta, \Phi(x) \rangle = \sum_{j=1}^{\infty} \theta_j \lambda_j^{1/2} \phi_j(x) : \theta \in \ell_2 \right\}$$

Rmk: If $\{\phi_j\}_{j=1}^{\infty}$ is a basis of $L^2(\nu)$, then for all $f \in L^2(\nu)$, $f = \sum_{j=1}^{\infty} \beta_j \phi_j = \sum_{j=1}^{\infty} \frac{\beta_j}{\lambda_j^{1/2}} \lambda_j^{1/2} \phi_j(x)$

$\rightarrow \langle \theta_*, \Phi(x) \rangle$

$$\|f\|_{L^2}^2 = \sum_{j=1}^{\infty} \beta_j^2 \qquad \|f\|_{\mathcal{H}}^2 = \sum_{j=1}^{m} \frac{\beta_j^2}{\lambda_j} = \|\theta_*\|_{\ell_2}^2$$

Test error : Bias - Variance decomposition

$(x_i, y_i)_{i=1}^n \qquad y_i = h_*(x_i) + \varepsilon_i \qquad \varepsilon_i \sim N(0, \sigma_\varepsilon^2)$

(excess test error) $\quad R(\hat{h}) = \mathbb{E}_x\left[(h_*(x) - \hat{h}(x))^2 \right]$

$$= \mathbb{E}_x\left[(h_*(x) - k_n(x)^T (K_n + \lambda)^{-1} y)^2 \right]$$

$$\mathbb{E}_\varepsilon\left[R(\hat{h})\right] = B_n(\lambda) + V_n(\lambda)$$

$$V_n(\lambda) = \sigma_\varepsilon^2 \operatorname{Tr}\left((K_n + \lambda)^{-2} \mathbb{E}_x\left[k_n(x) k_n(x)^T \right] \right)$$

$$B_n(\lambda) = \langle h_{*,n}, (K_n + \lambda)^{-1} M (K_n + \lambda)^{-1} h_{*,n} \rangle - 2 \langle v, (K_n + \lambda)^{-1} h_{*,n} \rangle + \| h_* \|_{L^2}^2$$

where $\quad M_{ij} = \mathbb{E}_x\left[K(x_i, x) K(x, x_j) \right]$

$$v_i = \mathbb{E}_x\left[K(x_i, x) h_*(x) \right]$$

$$h_{*,n} = (h_*(x_1), \dots, h_*(x_n))$$

Do exercise at home !!

Sharp characterization of Bias / Variance

Let us use the "canonical featurization" from eigendecomposition

$$\Phi(x) = \left(\lambda_j^{1/2}\, \phi_j(x)\right)_{j=1}^{\infty} \qquad y_i = \langle \theta_*, \Phi(x_i)\rangle + \varepsilon_i$$

$$z_i := \Phi(x_i) \qquad Z = \begin{bmatrix} - z_1 - \\ - z_n - \end{bmatrix} \in \mathbb{R}^{n \times \infty}$$

z_i iid $\qquad \mathbb{E}[z z^T] = \Sigma = \begin{pmatrix} \lambda_1 & & \\ & \lambda_2 & \\ & & \ddots \end{pmatrix}$

$$B_n(\lambda) = \lambda^2 \langle \theta_*, (Z^T Z + \lambda)^{-1} \Sigma (Z^T Z + \lambda)^{-1} \theta_* \rangle$$

$$V_n(\lambda) = \sigma_\varepsilon^2 \operatorname{Tr}\left(\Sigma Z^T Z (Z^T Z + \lambda)^{-2}\right)$$

Using random matrix theory, we have very sharp characterization of these quantities

$\rightarrow$ well approximated by "deterministic equivalents"

Define "effective regularization" λ_* as unique solution of fixed point equation:

$$n - \frac{d}{\lambda_*} = \mathrm{Tr}\left(\Sigma(\Sigma+\lambda_*)^{-1}\right)$$

Define:

$$\overline{V}_n(\lambda_*) = \sigma_\varepsilon^2 \cdot \frac{\mathrm{Tr}\left(\Sigma^2(\Sigma+\lambda_*)^{-2}\right)}{n - \mathrm{Tr}\left(\Sigma^2(\Sigma+\lambda_*)^{-2}\right)}$$

$$\overline{B}_n(\lambda_*) = \frac{\lambda_*^2 \langle \theta_*, \Sigma(\Sigma+\lambda_*)^{-2}\theta_* \rangle}{1 - \frac{1}{n}\mathrm{Tr}\left(\Sigma^2(\Sigma+\lambda_*)^{-2}\right)}$$

$$\overline{R}_n(\lambda_*) = \overline{B}_n(\lambda_*) + \overline{V}_n(\lambda_*)$$

↳ deterministic quantity that only depend on n, d, Σ and θ_* (coefficients of the target fct)

Can show: $B_n(\lambda) = (1 + o_n(1))\,\overline{B}_n(\lambda_*)$

$$V_n(\lambda) = (1 + o_n(1))\,\overline{V}_n(\lambda_*$$

Assumption [Hanson-Wright] Assume that for every P.S.D matrix $A \in \mathbb{R}^{\infty \times \infty}$ s.t $\mathrm{Tr}(\Sigma A) < \infty$

$$\mathbb{P}\left(|z^T A z - \mathrm{Tr}(A\Sigma)| \geq t \cdot \|\Sigma^{\frac{1}{2}} A \Sigma^{\frac{1}{2}}\|_F\right) \leq C e^{-ct}$$

E.g. $\Sigma^{-\frac{1}{2}} z$ is iid sub Gaussian or has log-concave density

↳ can relax this assumption

Thm [Cheng, Montanari, '24, M., Saeed, '24] (sketch)

Under above assumption, we have with high probs,

$$|B_n(\lambda) - \overline{B}_n(\lambda_*)| \leq \frac{C}{\sqrt{n}} \overline{B}_n(\lambda_*)$$

$$|V_n(\lambda) - \overline{V}_n(\lambda_*)| \leq \frac{C}{n} \overline{V}_n(\lambda_*)$$

Rmks

1) **Relative error:** $R_n(\lambda) = \overline{R}_n(\lambda)(1 + o_n(1))$

↳ very sharp prediction and extremely good agreement with numerical simulations

2) Dimension-free: applies to $p = \infty$

Rmk: [Gaussian sequence model] Equivalence with following model:

Observe $y_j^s = \lambda_j^{\frac{1}{2}} \theta_{*,j} + \frac{\omega}{\sqrt{n}} g_j \quad j = 1, 2, \dots$

$Y^s = \Sigma^{\frac{1}{2}} \theta_* + \frac{\omega}{\sqrt{n}} g \qquad g_j \sim N(0, I_p)$

Goal: recover θ_* from y^s

$$\hat{\theta}^s = \underset{\theta \in \mathbb{R}^\infty}{\operatorname{argmin}} \left\{ \| Y^s - \Sigma^{\frac{1}{2}} \theta \|_2^2 + \lambda_* \| \theta \|_2^2 \right\}$$

ω fixed pt of: $\omega^2 = \sigma_\varepsilon^2 + \mathbb{E}_g \left[\| \hat{\theta}^s - \theta_* \|_\Sigma^2 \right]$

$$R_n^s = \mathbb{E}_g \left[\| \hat{\theta}^s - \theta_* \|_\Sigma^2 \right] = \overline{B}_n(\lambda_*) + \overline{V}_n(\lambda_*)$$

↑ same as KRR

	Original problem	Sequence model
design	X random	$\Sigma^{\frac{1}{2}}$ deterministic
reg.	λ	$\lambda_*(\lambda)$
noise	σ_ε^2	ω^2

Rmk: Even when $\lambda = 0^+$ (interpolating solution)

$\lambda_*(0^+) > 0$: self induced regularization

Model has more regularization than what we naively expect from $\lambda = 0$

Rmk: Test error $\approx$ det equiv : doesn't depend on particular data distribution, only on eigenvalues

$\longrightarrow$ "Gaussian universality" or just "universality"

Inner-Product kernel

Last week : 2 layer neural network

e.g. $f_{RF}(x;a) = \langle a, \Phi_{RF}(x) \rangle$

$$\Phi_{RF}(x) = \frac{1}{\sqrt{M}} \begin{pmatrix} \sigma(\langle x, \omega_1^0 \rangle) \\ \vdots \\ \sigma(\langle x, \omega_M^0 \rangle) \end{pmatrix}$$

Linearizing 2nd layer weights

Associated kernel: $\omega_0 \sim N(0, Id_d)$

$$K_M(x_1, x_2) = \frac{1}{M} \sum_{j \in [M]} \sigma(\langle \omega_j^0, x_1 \rangle) \sigma(\langle \omega_j^0, x_2 \rangle)$$

$$\xrightarrow[M \to \infty]{} \mathbb{E}_{\omega^0 \sim N(0, Id_d)} \left[\sigma(\langle \omega, x_1 \rangle) \sigma(\langle \omega, x_2 \rangle) \right]$$

$\|x_1\|_2 = \|x_2\|_2 = 1 \qquad r = \langle x_1, x_2 \rangle$

$$= \mathbb{E}_{G_1, G_2 \sim N(0,1)} \left[\sigma(G_1) \sigma(r G_1 + \sqrt{1-r^2}\, G_2) \right]$$

$$= h(\langle x_1, x_2 \rangle)$$

More generally limiting kernel as $M \to \infty$ in fully connected NNs with Gaussian initialization are inner product kernels

What is generalization error of inner-product kernel?

$$(x_i, y_i)_{i=1}^n \qquad x_i \sim \mathrm{Unif}\left(\mathbb{S}^{d-1}\right)$$

$$y_i = f_*(x_i) + \varepsilon_i \qquad f_* \in L^2$$

We saw that the test error depends crucially on the eigendecomposition of the kernel.

For data uniformly distributed on the sphere, there is an explicit eigendecomposition in terms of spherical harmonics

$$L^2(\mathbb{S}^{d-1}) = \bigoplus_{k=0}^{\infty} V_k$$

$V_k \to$ space of d° k spherical harmonics
i.e space of polynomials of degree k perpendicular to all polynomials of degree $\leq k-1$.

$$B_{d,k} := \dim(V_k) = \frac{d+2k-2}{k}\binom{d+k-3}{k-1} = \Theta(d^k)$$

$\rightarrow$ basis $\{Y_{ks} : 1 \le s \le B_{d,k}\}$

then

$$h(\langle x_1, x_2\rangle) = \sum_{k=0}^{\infty} \mu_k \sum_{s=1}^{B_{d,k}} Y_{ks}(x_1)\, Y_{ks}(x_2)$$

each eigenvalue μ_k has degeneracy $B_{d,k}$

$\rightarrow$ eigenspace associated to μ_k is V_k

$$\mathrm{Tr}(K) = \mathbb{E}_x[K(x,x)] = h(1) = \sum_{k=0}^{\infty} \mu_k B_{d,k}$$

meaning that $\mu_k \simeq \frac{1}{B_{d,k}} \simeq d^{-k}$

We are interested in the test error in the polynomial high-dimensional regime where $n, d \to \infty$

$$\frac{n}{d^k} \to \gamma \qquad k, \gamma > 0$$

Consider empirical kernel matrix $K_n = \left(K(x_i, x_j)\right)_{i,j=1}^{n}$

$$K_n = \sum_{k=0}^{\infty} \lambda_k Y_k Y_k^T$$

where

$$Y_k = \begin{bmatrix} Y_{k1}(x_1) & \cdots & Y_{k B_{d,k}}(x_1) \\ Y_{k1}(x_n) & \cdots & Y_{k B_{d,k}}(x_n) \end{bmatrix} \in \mathbb{R}^{n \times B_{d,k}}$$

Case 1: $l < \kappa < l+1$

$$K_n = \underbrace{\sum_{k=0}^{l} \lambda_k Y_k Y_k^T}_{K_{n,\le l}} + \underbrace{\sum_{k=l+1}^{\infty} \lambda_k Y_k Y_k^T}_{K_{n,>l}}$$

* $\boxed{k \le l}$: "low degree part"

$$\lambda_{\min}(Y_k Y_k^T) \simeq n \qquad \lambda_k n \gtrsim d^{\kappa - l} \xrightarrow[d \to \infty]{} \infty$$

hence eigenvalues of this part $\longrightarrow \infty$

furthermore rank $1 + B_{d,1} + B_{d,2} + \ldots + B_{d,l} = O(d^l)$

$$\ll n$$

Hence $K_{n,\leq \ell}$ low rank spiked matrix

$\longrightarrow$ Thanks to that, KRR will fit perfectly $P_{\leq \ell} f_*$ the projection of f_* on degree-ℓ polynomials

(i.e, best degree-ℓ polynomial approx of f_*)

* $\boxed{k \geq \ell+1}$ $\quad Y_k \in \mathbb{R}^{n \times d^k}$ $\quad \frac{Y_k Y_k^T}{B_{d,k}} \simeq I_n$

$$K_{n,>\ell} \approx \underbrace{\left(\sum_{k=\ell+1}^{\infty} \xi_k B_{d,k} \right)}_{\mu_*} I_n$$

"self-induced regularization"

$\longrightarrow$ Because of that, KRR will not fit at all the higher degree part of f_*

Hence: $\quad n \asymp d^k \qquad \ell < k < \ell+1$

$$\hat{f}_{KRR}(x) = P_{\leq \ell} f_*(x) + \underbrace{\Delta(x)}_{L^2 \text{ vanishing}}$$

$\longrightarrow$ can interpolate data if $\lambda = 0^+$

$$R(\hat{\beta}_{KRR}) = \| P_{>\ell} \beta_* \|_{L^2}^2 + o_d(1)$$

Case 2: $n \simeq d^\ell$

$$K_n = K_{n, \leq \ell-1} + \xi_\ell Y_\ell Y_\ell^\top + K_{n, >\ell}$$

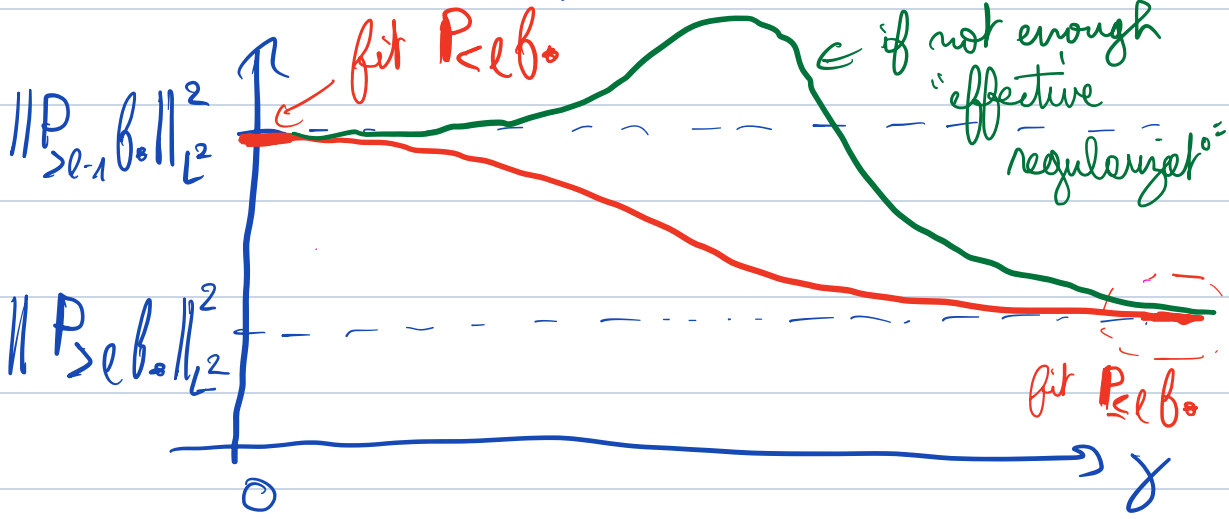
$K_{n, \leq \ell-1}$: low rank spikes → fit perfectly $P_{\leq \ell-1} \beta_*$

$\xi_\ell Y_\ell Y_\ell^\top$: $\xi_\ell \sim B_{d,\ell}$, $\frac{Y_\ell Y_\ell^\top}{B_{d,\ell}} \sim \mathcal{O}(1)$, behave as a Wishart matrix

$K_{n, >\ell} \simeq \mu_* I_n$: does not fit $P_{>\ell} \beta_*$ at all

using RMT we can compute the test error

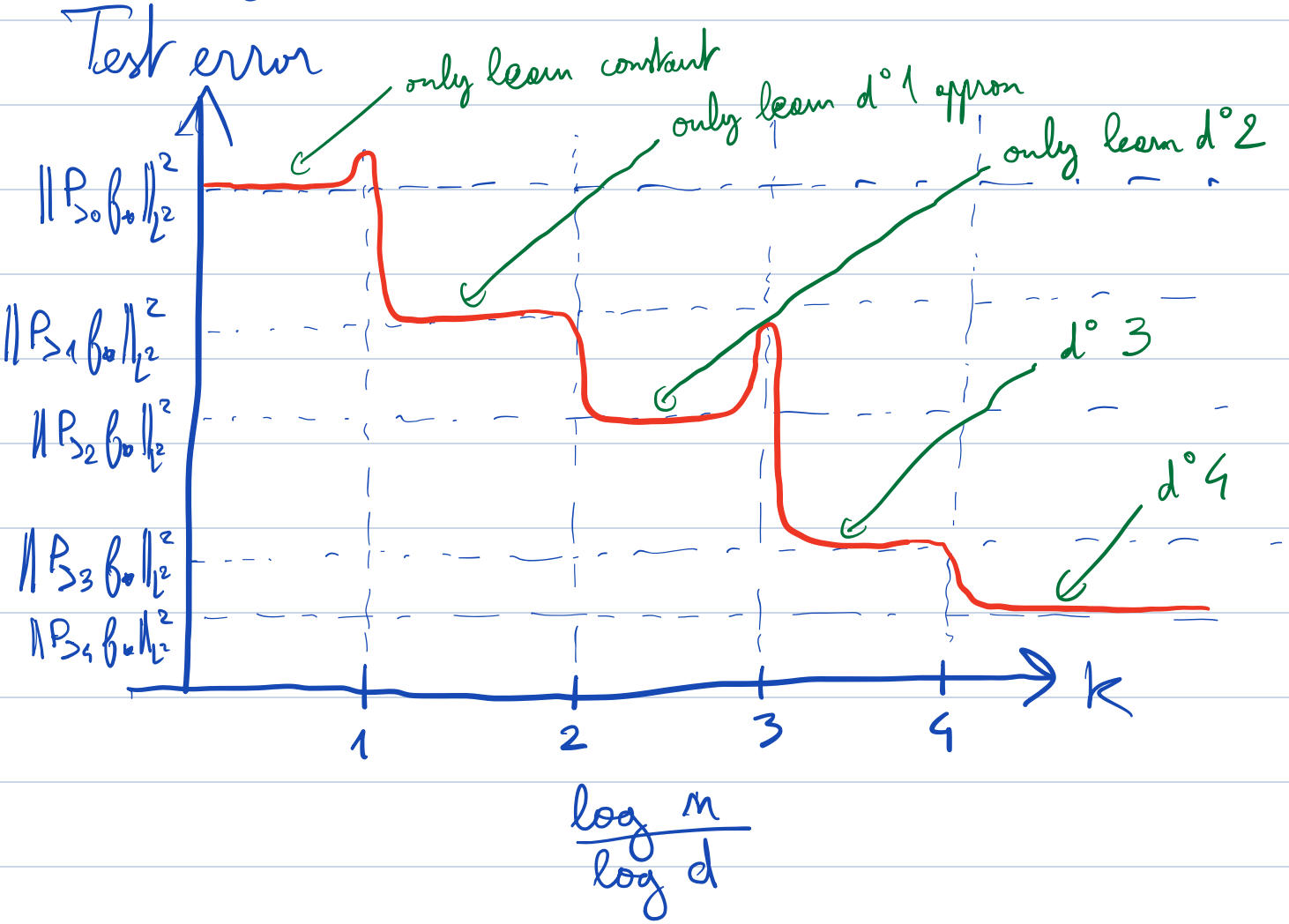
$\frac{n}{d^\ell} \to \gamma$



$\frac{n}{d^\ell} \to \gamma$ $\quad \gamma \searrow 0$ doesn't fit $P_\ell f_*$ at all

$\gamma \to \infty$ fit $P_\ell f_*$ perfectly

In between, if not enough "effective regularization" there might be a peak at $n = B_{d,\ell}$



[Ghorbani et al '21] [Misiakiewicz, '22]

Conclusion: let's say $d = 1000$ (typical in applications)

When to get $d^\circ 1$ approx $n \simeq 1000$

$d^\circ 2$ approx $n \simeq 1\,000\,000$

$d^\circ 3$ approx $n \simeq 1\,000\,000\,000$

Quite bad !!! Curse of dimensionality !!!

⚠ Often data is not isotropic but is concentrated on a smaller subspace of dimension d_{eff}

→ in that case $n \simeq d_{eff}^{k}$ for $d^\circ k$ approx

E.g. CIFAR-10 $d_{eff} \simeq 30$

$n \simeq 60\,000$

→ can only get $d^\circ 3$ approx

Power and limitations of kernel methods

What KRR does high level:

$$K = \sum_{j=1}^{\infty} \lambda_j \phi_j \phi_j^* \qquad \lambda_1 \geq \lambda_2 \geq \dots$$

Target fct: $f_* = \sum_{j=1}^{\infty} \beta_j \phi_j$ and n data pts

$$\Rightarrow \text{KRR} \quad \hat{f} \approx \sum_{j=1}^{n} \beta_j \phi_j$$

i.e fit the projection of f_* on top n eigenspaces

E.g. To fit d^k polynomial approximation: subspace of dim $\Theta(d^k)$

↳ need $n \gtrsim d^k$

→ KRR with inner product kernel

This is minimax optimal: to fit all d^k polynomial we need

$$n = \Omega(d^k)$$

↳ no method can do better!

Classical results on rates of kernels

(1) $\mathcal{G} = \{ f_* : \mathbb{R}^d \to \mathbb{R} \quad L\text{-lipschitz} \}$

$$\sup_{f_* \in \mathcal{G}} R_{test}(f_*, \hat{f}) \asymp n^{-1/d}$$

↳ worst-case, to get error $\leq \varepsilon$ need $n \geq \left(\frac{1}{\varepsilon}\right)^d$

CURSE OF DIMENSIONALITY

⟶ no methods can do better (minimax optimal)

(2) $\mathcal{G}_S = \{ f_* \text{ with } S \text{ first derivatives bounded} \}$

$$\sup_{f_* \in \mathcal{G}_S} R_{test}(f_*, \hat{f}) \asymp n^{-S/S+d}$$

To get ε error $n \asymp \left(\frac{1}{\varepsilon}\right)^{1+\frac{d}{S}}$ if $S \asymp d$

$\downarrow$

no curse of dim

but needs to be extremely smooth

→ again no method can do better
↳ KRR minimax optimal

Hence: KRR is adaptive to smoothness of the fct
↳ smoother fct will be easier to fit

E.g.: • $S \simeq d$ smooth $\quad n \simeq \left(\frac{1}{\varepsilon}\right)^c$

• d° l polynomial $\quad n \simeq d^l$

KRR adaptive to other properties of target fct?

E.g. $n \geq d^l$ to fit any degree-l polynomial no matter the polynomial !!

Other interesting classes of fcts: e.g. $f(x) = g(U^T x)$
("Multi-index fcts")
$U \in \mathbb{R}^{d \times p}$

i.e fcts that only depend on a low dimensional project° of the data

minimax rate is $n^{-\Theta\left(\frac{1}{p}\right)}$

→ KRR will not reach this minimax rate

(e.g. inner-product kernel doesn't care at all that only depend on low dim project°)

Kernel methods will not be adaptive to low-dim structure

At the same time for NNs: $f_{NN}(x;\theta) = \sum a_j \sigma(\langle \omega_j, x \rangle)$

Idea: can align ω_j with support U, then the model effectively dimension p and we can achieve minimax rate

$$\hat{f}_{ERM} = \underset{\theta}{\operatorname{argmin}} \frac{1}{n} \sum_{i=1}^{n} (y_i - f_{NN}(x_i;\theta))^2 + \lambda \|f\|_{F_1}^2$$

then $R(f_*, \hat{f}_{ERM}) \simeq n^{-\Theta(\frac{1}{p})}$

Of course, we would need to show this rate for a NN trained by SGD!

$\Rightarrow$ See next two lectures!!!

Summary

* Lazy regime:
  - → show trackability of optimization ✓
  - → NNs in this regime collapse on kernel methods

* Kernel methods are (relatively) well understood learning methods that have limited adaptivity
  - ↳ heuristically, from n samples, can only fit fcts in a n-dimensional subspace $\text{span}\{\Phi(x_i)\}_{i=1}^n$ (indep of labels)
  - ↳ "fixed features methods"

$$f(x) = \langle \theta, \Phi_{(w_0)}(x) \rangle$$

↳ NT model at initialization

"Fixed embedding"

"Fixed representation"

⇒ suffer curse of dimensionality on interesting fct classes that can be approximated efficiently by NNs

* When NNs are trained beyond lazy regime

W^t moves away from W^0

$$f_t(x) = \langle \theta, \Phi_{W^t}(x) \rangle$$

↳ embedding changes

→ NNs can learn a good representat° of the data

beyond lazy regime : "feature learning" or "rich" regimes

E.g. $a^T \sigma(W_t x)$ W_t can align with the support of target fct $g(U_*^T x)$

↳ then fitting a becomes much easier

(effectively p-dimensional problem instead of d-dimensional problem)

In the next 2 lectures, we will show such an example of optimization regime with feature learning.

Dimension lower bound for Kernel methods

Below, I give a very simple lower bound on learning with kernel methods, which I think encapsulates the limitation of learning with kernels

→ we can only fit a subspace of dimension n independent of the target function

→ no adaptivity

Prop [Hsu, '22, Abbe, Boix-Adsera, Misiakiewicz '22]

Let $\mathcal{R}$ a Hilbert space with inner product $\langle \cdot, \cdot \rangle_{\mathcal{R}}$

Let $\{f_1, \dots, f_M\} \subset \mathcal{R}$ with $\|f_i\|_{\mathcal{R}} = 1$

Let $T \subset \mathcal{R}$ linear subspace of $\mathcal{R}$ of dimension n

Let $\varepsilon := \frac{1}{M} \sum_{i \in [M]} \inf_{g \in T} \|g - f_i\|_{\mathcal{R}}^2$) average approximation error of f_i by T

Then

$$n \geq \frac{M}{\|G\|_{op}} (1 - \varepsilon)$$

where $G = (\langle f_i, f_j \rangle_{\mathcal{R}})_{i,j=1}^{n}$

Rmk: What does it mean?

Consider $\mathcal{H} = L^2(\mathcal{X})$

When fitting with a kernel method with n data points

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

Take $T = \operatorname{span}\{ x \mapsto K(x, x_i) : i \in [n] \}$

$\hookrightarrow \dim(T) = n$

Test error when fitting target fct $f_i \geq \inf_{g \in T} \|g - f_i\|_{L^2}^2$

Hence if test error $\leq \varepsilon$ for M orthogonal fcts f_i we must have

$$n \geqslant M(1-\varepsilon)$$

Proof: Let $\phi_1, \dots, \phi_n$ be an orthonormal basis of T and let Π_T be the orthogonal projection onto T in $(\mathbb{R}, \langle \cdot, \cdot \rangle_R)$.

In particular $\Pi_T f = \sum_{j=1}^{n} \phi_j \langle \phi_j, f \rangle_R$

$$\varepsilon = \frac{1}{M} \sum_{i=1}^{M} \inf_{g \in T} \|g - f_i\|_R^2 \qquad \text{(definition)}$$

$$= \frac{1}{M} \sum_{i=1}^{M} 1 - \|\Pi_T f_i\|_R^2 \qquad (\|f_i\|_R^2 = 1)$$

$$= 1 - \frac{1}{M} \sum_{i=1}^{M} \sum_{j=1}^{n} \langle \phi_j, f_i \rangle_R^2$$

Use that

$$\sum_{i=1}^{M} \langle g, f_i \rangle_R^2 = \langle g, \sum_{i=1}^{M} \langle g, f_i \rangle f_i \rangle_R$$

$$\leq \|g\|_R \left(\sum_{i,j=1}^{M} \langle g, f_i \rangle_R \langle g, f_j \rangle_R \langle f_i, f_j \rangle_R \right)^{1/2}$$

$$= \|g\|_R \left(b^T G b \right)^{1/2} \leq \|g\|_R \, \|G\|_{op}^{1/2} \, \|b\|_2$$

where $b = (\langle g, f_i \rangle)_{i=1}^{M}$

Note that $\|e\|_2^2$ is the left-hand side of the above inequality. Hence

$$\sum_{i=1}^{M} \langle g, f_i \rangle_R^2 \leq \|g\|_R^2 \, \|G\|_{op}$$

Hence $$1 - \varepsilon = \frac{1}{M} \sum_{j=1}^{n} \sum_{i=1}^{M} \langle \phi_j, f_i \rangle^2$$

$$\leq \frac{n}{M} \|G\|_{op}$$

$$\Rightarrow \quad n \geq \frac{M}{\|G\|_{op}} (1 - \varepsilon)$$

□