Lecture 13: Test error of Random Features models

"Generalization error of random features and kernel methods: hypercontractivity and kernel matrix concentration"
[Mei, M., Montanari, 2021]

Here: $n = d^\ell$ "polynomial scaling"

Next lecture: $n = c\,d$ "proportional scaling"

Random Features model:

RKHS: $f(x; a) = \int_\Omega \sigma(x;\theta)\, a(\theta)\, \tau(d\theta)$

where $\|a\|_{L^2}^2 = \int_\Omega a(\theta)^2\, \tau(d\theta) < \infty$

∞^{te} dim space

Random features approx: $\theta_1, \dots, \theta_N \overset{iid}{\sim} \tau$

$$\hat{f}_{RF}(x, a) = \frac{1}{N} \sum_{j=1}^{N} \underbrace{a_j}_{\text{2nd layers trained}} \sigma(x; \underbrace{\theta_j}_{\text{1st layer weights fixed}})$$

[Rahimi - Recht - 2008]

$\dim(RF) = N$ "finite dim. approx." of RKHS

$$H_N(x_1, x_2) = \frac{1}{N} \sum_{j=1}^{N} \sigma(x_1; \theta_j)\, \sigma(x_2; \theta_j)$$

$$\longrightarrow \mathbb{E}_{\theta}\left[H_N(x_1,x_2)\right] = \int_{\Omega} \sigma(x_1,\theta)\,\sigma(x_2,\theta)\,\tau(d\theta)$$

"Modern" interpretat°:

* 2-layers NN : train 2^{nd} layer
* Linearized NNs :

$$f_{NN}(x;\beta) \approx f_{NN}(x;\beta_0) + \langle \beta-\beta_0, \nabla f_{NN}(x;\beta_0)\rangle$$

$$\beta = (a, \theta)$$

RF: linearization wrt 2^{nd} layer weights

$$\hat{f}_{RF}(x;a) = \langle a, \nabla_a f_{NN}(x;\theta_0)\rangle$$

$$= \langle a, \varphi_N(x)\rangle$$

$$\varphi_N(x) = (\sigma(x;\theta_1), \dots, \sigma(x;\theta_N)) \in \mathbb{R}^N$$

Setting: - $(\mathcal{X}, \nu)$ prob space $f_* \in L^2(\mathcal{X})$

$\mathcal{X} \subseteq \mathbb{R}^d$

$\{(y_i, x_i)\}_{i \leq n}$ $y_i = f_*(x_i) + \varepsilon_i$

$x_i \underset{iid}{\sim} \nu$ $\varepsilon_i \underset{iid}{\sim} N(0, \sigma_\varepsilon^2)$

- (Ω, τ) prob space weights

$\{\theta_j\}_{j \leq N}$ $\theta_j \overset{iid}{\sim} \tau$

RF model: $\hat{f}(x; a) = \frac{1}{N} \sum_{j=1}^{N} a_j \sigma(x; \theta_j)$

RF ridge regression (RFRR):

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

Test error:

$$R_{n,N}(f_*) = \mathbb{E}_{x_{new}}\left[\left(f_*(x_{new}) - \hat{f}(x_{new}, \hat{a}(\lambda))\right)^2\right]$$

1) General assumptions

2) Example : $x \sim \mathrm{Unif}(\mathbb{S}^{d-1})$

General assumptions :

$(\mathcal{X}_d, \nu_d) \qquad (\Omega_d, \tau_d) \qquad \mathcal{X} \subseteq \mathbb{R}^d$

Featurization map : $\qquad \sigma \in L^2(\mathcal{X} \times \Omega)$

$$\sigma : \mathcal{X} \times \Omega \longrightarrow \mathbb{R}$$
$$(x ; \theta) \longmapsto \sigma(x ; \theta)$$

- $\mathbb{T} : L^2(\Omega) \longrightarrow L^2(\mathcal{X})$

$$a \longmapsto f(x ; a) = \int_\Omega \sigma(x ; \theta)\, a(\theta)\, \tau(d\theta)$$

$\mathbb{T}$ compact operator

$$\left(\mathbb{T} = \sum_{j=1}^{\infty} \lambda_j \psi_j \phi_j^* \right.$$

$\mathbb{E}_x [\psi_j(x) \psi_k(x)] = \delta_{jk}$

- $\{\psi_j\}_{j \geq 1}$ orthonormal basis of $L^2(\mathcal{X})$
- $\{\phi_j\}_{j \geq 1}$ ———— $L^2(\Omega)$

$$\sigma(x;\theta) = \sum_{j=1}^{\infty} \lambda_j \psi_j(x) \phi_j(\theta)$$

$$\sigma \in L^2 \qquad \|\sigma\|_{L^2}^2 = \sum_{j=1}^{\infty} \lambda_j^2 < \infty$$

- $\mathbb{H} = \mathbb{T}\mathbb{T}^* : L^2(\mathcal{X}) \to L^2(\mathcal{X})$

$$\mathbb{H} f(x) = \int_{\mathcal{X}} \underbrace{H(x,x')} f(x') \, \nu(dx')$$

$$H(x_1,x_2) = \int_{\Omega} \sigma(x_1;\theta) \sigma(x_2;\theta) \, \tau(d\theta)$$

$$\mathbb{H} = \sum_{j=1}^{\infty} \lambda_j^2 \psi_j \psi_j^*$$

- $\mathbb{U} = \mathbb{T}^*\mathbb{T} : L^2(\Omega) \to L^2(\Omega)$

$$\mathbb{U} g(\theta) = \int_{\Omega} U(\theta,\theta') g(\theta') \, \tau(d\theta')$$

$$U(\theta_1,\theta_2) = \int_{\mathcal{X}} \sigma(x;\theta_1) \sigma(x;\theta_2) \, \nu(dx)$$

$$\mathbb{U} = \sum_{j=1}^{\infty} \lambda_j^2 \phi_j \phi_j^*$$

(simplified) <u>assumptions at level (M, m, u)</u>

$M(d)$ associated to $N(d)$

$m(d)$ ——————— $n(d)$

$u \geq \max(M, m)$

① [Hypercontractivity] $\mathcal{X}_{\leq u} = \mathrm{span}\{\psi_j : j \leq u\}$

$\forall k \geq 1, \quad \exists\ C > 0$

$\forall g \in \mathcal{X}_{\leq u}, \quad \|g\|_{L^{2k}(X)} \leq C \cdot \|g\|_{L^2(X)}$

$\|g\|_{L^p} = \mathbb{E}[g^p]^{\frac{1}{p}}$ $\quad p<q : \|g\|_{L^p} \leq \|g\|_{L^q}$

(Jensen)

$\to$ hypercontractivity : reverse inequ

space G "hypercontractive" $\forall g \in G, \|g\|_{L^p} \leq C \|g\|_{L^q}$

$p > q$

functions in G are "delocalized".

Same thing $\Omega_{\leq u} = \mathrm{span}\{\phi_j : j \leq u\}$

② [Concentration of diagonal elements]

$$H_{>m}(x_1, x_2) = \sum_{j>m} \lambda_j^2 \psi_j(x_1)\psi_j(x_2)$$

$$\sup_{i \leq n} \left| H_{>m}(x_i, x_i) - \mathbb{E}_x[H_{>m}(x,x)] \right| = o_{d,\mathbb{P}}(1) \cdot \mathbb{E}_x[H_{>m}(x,x)]$$

$o_{d,\mathbb{P}}$ "small o in probability"

$X_d = o_{d,\mathbb{P}}(Y_d)$ $\qquad \frac{X_d}{Y_d} \to 0$ in $\mathbb{P}$.

Same for $U_{>M}(\theta_i, \theta_i)$

③ [Spectral gap]

- Underparametrized regime : $N \leq n$

$$\frac{1}{\lambda_M^2} \sum_{k=M+1}^{\infty} \lambda_k^2 \ll N \ll \frac{1}{\lambda_{M+1}^2} \sum_{k=M+1}^{\infty} \lambda_k^2$$

gap : $\lambda_M \quad \lambda_{M+1}$

$\to$ M: subspace estimated accurately in under.

- Overp. regime: $n \leq N$

$$\frac{1}{\lambda_m^2} \sum_{k=m+1}^{\infty} \lambda_k^2 \ll n \ll \frac{1}{\lambda_{m+1}^2} \sum_{k=m+1}^{\infty} \lambda_k^2$$

$f_* \in L^2(\mathcal{X}) \qquad f_* = \sum_{k=1}^{\infty} \langle f_*, \psi_k \rangle_{L^2(\mathcal{X})} \psi_k$

$$P_{>\ell} f_* = \sum_{k=\ell+1}^{\infty} \langle f_* ; \psi_k \rangle_{L^2(\mathcal{X})} \psi_k$$

<u>Thm</u>: [M, M, M, 21] Assumpt° at level (M, m, u)

- <u>Overp. regime</u>: $N \gtrsim d^\delta n$ for $\delta > 0$

$\lambda \in [0, \lambda_o]$, any fixed $\eta > 0$:

$$R_{n,N}(f_*) = \| P_{>m} f_* \|_{L^2}^2 + o_{d,\mathbb{P}}(1) \cdot \| f_* \|_{L^{2+\eta}}^2$$

- <u>Under regime</u>: $n \gtrsim d^\delta N \qquad \delta > 0$

$\lambda \in [0, \lambda_u]$, any fixed $\eta > 0$

$$\left[R_{n,N}(f_*) = \| P_{>M} f_* \|_{L^2}^2 + o_{d,\mathbb{P}}(1) \dots \right.$$

$$\hat{f}_{RF} \simeq \begin{cases} P_{\leq n} f_* & N \geq d^\delta n \\ P_{\leq M} f_* & n \geq d^\delta N \end{cases}$$

<u>Rmk:</u> (1) n, N symmetric role

<u>Approx. error:</u> N finite, $n = \infty$

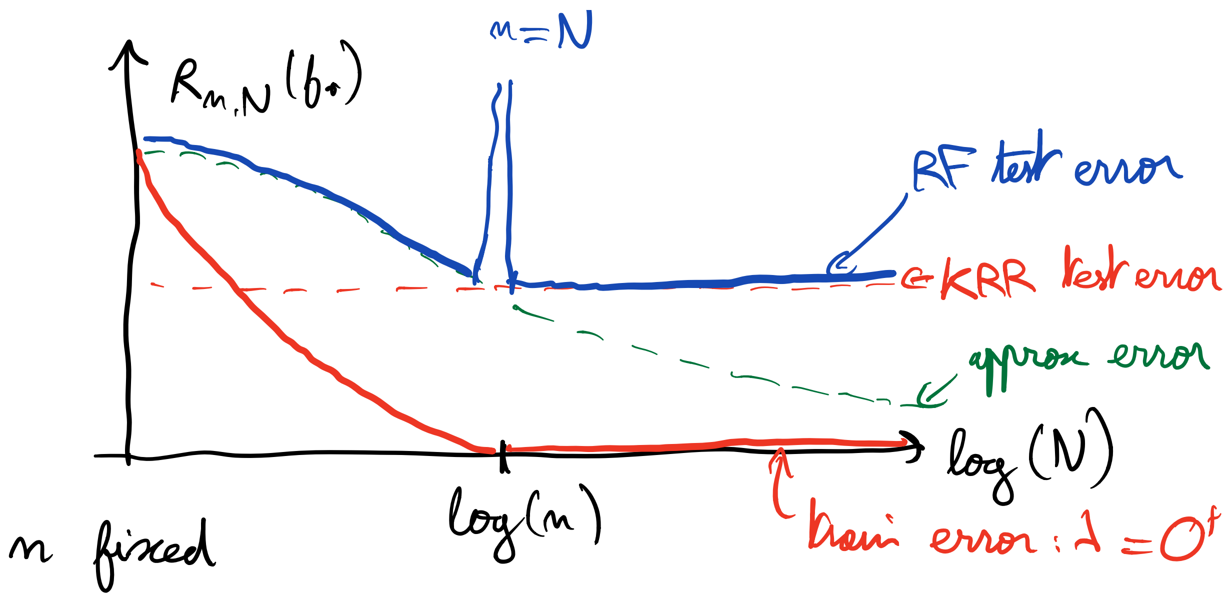
$$R_N^{app}(f_*) = \inf_a \| f_* - \sum_{j=1}^N a_j \sigma(\cdot\,; \theta_j) \|_{L^2}^2$$

$$= \| P_{>M} f_* \|_{L^2}^2 + o_{d,\mathbb{P}}(1).$$

<u>Statistical error:</u> n finite, $N = \infty$ (KRR)

$$R_n^{KRR}(f_*) = \| P_{>n} f_* \|_{L^2}^2 + o_{d,\mathbb{P}}(1)$$

$$R_{n,N}(f_*) \text{ "=" } \max\Big(\underset{N \ll n}{R_N^{app}(f_*)} \,,\, \underset{N \gg n}{R_n^{KRR}(f_*)} \Big)$$



② <u>Optimal overparametrization:</u> $N \geq d^\delta n$

③ <u>Optimality interpolators:</u> $\lambda \in [0, \lambda_*]$

$$\lambda = 0_+$$

$$y_i = f_*(x_i) + \varepsilon_i$$

Example: $(\mathcal{X}, \nu) = (\Omega, \tau) = (\mathbb{S}^{d-1}(\sqrt{d}), \text{Unif})$

$= \{x \in \mathbb{R}^d, \|x\|_2 = \sqrt{d}\}$

$\sigma(x;\theta) = \sigma(\langle x, \theta\rangle/\sqrt{d})$

$\sigma : \mathbb{R} \to \mathbb{R} \quad \searrow \quad O(1)$

$\mathbb{E}_{x,\theta}[\sigma(\langle x,\theta\rangle/\sqrt{d})] = \mathbb{E}_{x_1}[\sigma(x_1)]$

$\Downarrow d$

$N(0,1)$

$\leq C$

Thm [MMM 21] σ satisfying "genericity" conditions

$d^{s+\delta} \leq n \leq d^{s+1-\delta} \qquad d^{S+\delta} \leq N \leq d^{S+1-\delta}$

- Over regime: $N \geq d^{\delta} n$

$$R_{n,N}(f_*) = \|\overline{P}_{>s} f_*\|_{L^2}^2 + o_{d,\mathbb{P}}(1)$$

$\overline{P}_{>s}$: project° orthogonal to polynomials $d^\circ \leq s$

$\hat{f}_{RF} \rightarrow$ fit exactly polynomials $d^\circ \leq s$

- Under regime: $n \geq d^\delta N$

$$R_{n,N}(f_*) = \| \overline{P}_{>S} f_* \|_{L^2}^2 + o_{d,\mathbb{P}}(1)$$

$\square$

Proof: checking the conditions

Funct° space: $L^2(S^{d-1}(\sqrt{d}), \mathrm{Unif})$

$$L^2(S^{d-1}) = \bigoplus_{\ell=0}^{\infty} V_{d,\ell}$$

$\downarrow$ linear subspace of $d^\circ \ell$ polynomials

- $\dim(V_{d,\ell}) = B_\ell = \frac{2\ell+d-2}{d-2}\binom{\ell+d-3}{\ell} = \Theta(d^\ell)$

- orthonormal basis: $\{Y_{\ell j}\}_{j\in[B_\ell]}$ spherical harmonics

$$\mathbb{E}[Y_{\ell j} Y_{k j'}] = \delta_{\ell k}\delta_{jj'}$$

Integral operator: $\mathbb{T}$ commute with $SO(d)$

* $\sigma(\langle x, \Theta\rangle/\sqrt{d}) = \sum_{k=0}^{\infty} \xi_k \sum_{j\in[B_k]} Y_{kj}(x) Y_{kj}(\Theta)$

eigenvalues (λ_k) $\quad \xi_k \quad$ degeneracy B_k

* $\xi_k^2 B_k \leq \mathbb{E}[\sigma^2] < \infty$

$$\xi_k^2 = O(d^{-k})$$

$$\to \xi_k^2 = \Theta(d^{-k})$$

$$\mu_k(\sigma) = \mathbb{E}_G[\sigma(G) He_k(G)] \neq 0$$

Check the assumptions:

$d^s \leq n \leq d^{s+1}$ $\quad\quad$ $d^S \leq N \leq d^{S+1}$

$m = \sum_{\ell\leq s} B_\ell$ $\quad\quad$ $M = \sum_{\ell\leq S} B_\ell$

$u \geq \max(m, M)$

$\xi_k \propto d^{-\frac{k}{2}}$

① Spaces of low-d° pol. on the sphere are hypercontractive

[Beckner, 92] f degree ℓ polynomial

$$\|f\|^2_{L^q(\mathbb{S}^{d-1})} \leq (q-1)^\ell \|f\|^2_{L^2(\mathbb{S}^{d-1})}$$

② $$H_{>m}(x_i, x_i) = \sum_{k=s+1}^{\infty} \xi_k^2 \underbrace{\sum_{j \in [B_k]} Y_{kj}(x_i)^2}_{= B_k}$$

$>m$

$$= \sum_{k=s+1}^{\infty} \xi_k^2 B_k$$

$$U_{>M}(\theta_i, \theta_i) = \sum_{k=S+1}^{\infty} \xi_k^2 B_k$$

③ $$\sum_{k=\ell+1}^{\infty} \lambda_k^2 = O(1)$$

$$= \omega(1)$$ 6 not polynomial

Underparam. regime:

(1) $$\frac{1}{\lambda_M^2}\sum_{k>M}\lambda_k^2 \asymp \frac{1}{\xi_S^2} \asymp d^S$$

(2) $$\frac{1}{\lambda_{M+1}^2}\sum_{k>M}\lambda_k^2 \asymp \frac{1}{\xi_{S+1}^2} \asymp d^{S+1}$$

$$(1) \ll N \ll (2)$$

$$d^S \ll N \ll d^{S+1}$$

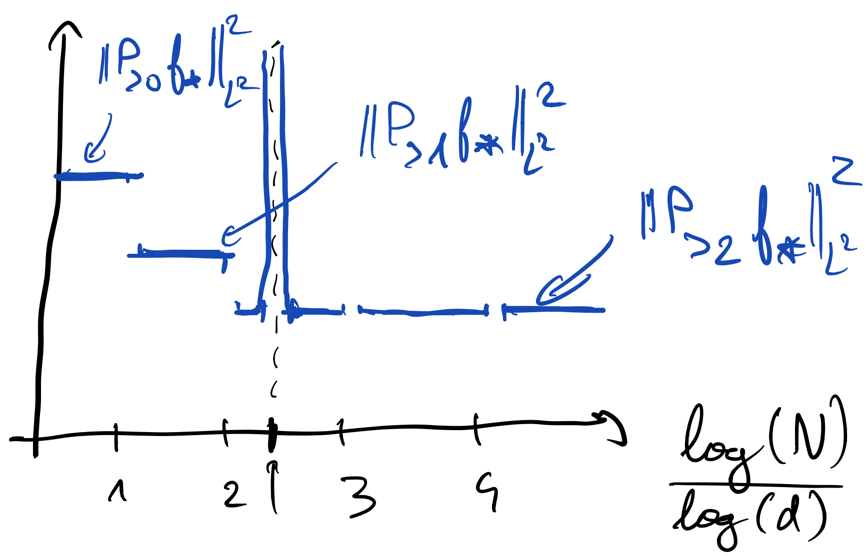
Overp. regime

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

$\square$

Figure:

$$n = d^{2.4}$$



$$\frac{\log(n)}{\log(d)} = \frac{\log(N)}{\log(d)}$$

$N = c_1 d \qquad n = c_2 d$

$N = n$ Double descent phenomenon