Consequences of Mercer's Theorem + NTK

Last time

Theorem (Mercer)

Suppose $\mathcal{X}$ is compact and $K: \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is positive-definite and continuous. Let ν be a finite Borel measure supported on all of $\mathcal{X}$. Then define T_K as

$$(T_K f)(x) = \int_{\mathcal{X}} K(x,y) f(y) \, d\nu(y)$$

Claim. T_K has eigenbasis $e_1, e_2, \ldots$ with eigenvalues $\mu_1, \mu_2, \ldots$

s.t. $K(x,y) = \sum_{m=1}^{\infty} \mu_m e_m(x) e_m(y)$.

Corollary: For any ν,

$$\|f\|_{\mathcal{H}}^2 = \sum_{m=1}^{\infty} c_m^2 / \mu_m, \quad \text{where}$$

$$c_m = \int f(x) e_m(x) \, d\nu(x).$$

Slightly weird b/c χ RHS depends on ν, but LHS does not.

Intuition: $\|f\|_{\mathcal{H}}^2 = \sup_{S=\{x_1,\dots,x_n\}} f_S^T K_S^{-1} f_S$

$f_S = \begin{bmatrix} f(x_1) \\ \vdots \\ f(x_n) \end{bmatrix}$

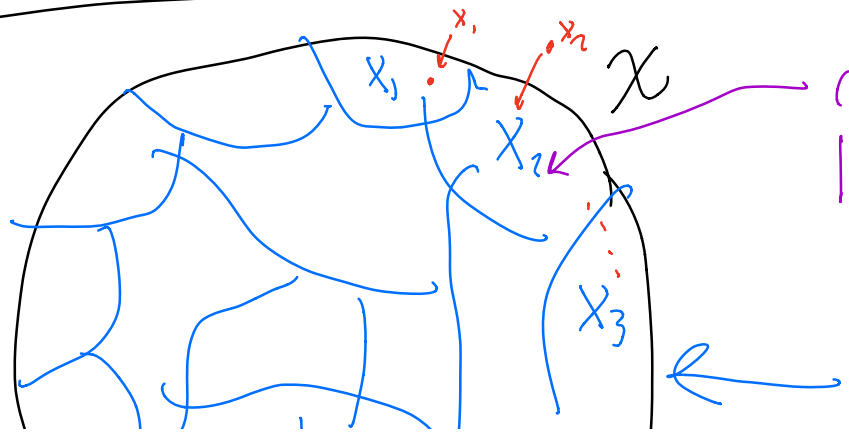
S

Sample $x_1, \dots, x_n \sim \nu$

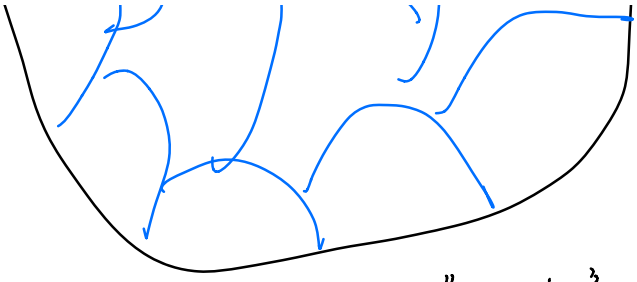
Since $\mathrm{supp}(\nu) = \chi$, "eventually" these samples contain everything.

As $n \to \infty$, $f_S^T K_S^{-1} f_S \longrightarrow \|f\|_{\mathcal{H}}^2$

K_S, f_S clearly do depend on ν — happens independently of ν



On each X_i, $|f(x) - f(y)| \le \delta$ $\forall\ x, y \in X_i$



Think of x_i as "samples", but chosen this way for ease of analysis.

① $S = \{x_1, \dots, x_n\}$

$f_S^T K_S^{-1} f_S \longrightarrow \|f\|_{\mathcal{H}}^2$ as S gets large. (as $S \to \Omega$)

② $f_S^T K_S^{-1} f_S \longrightarrow \sum_{m=1}^{\infty} \frac{1}{\mu_m} \left(\int_{\mathcal{X}} f(x) e_m(x) d\nu(x) \right)^2$

Why ② is true.

$\nu(X_1)$ small but $\nu(X_2)$ is large

$f_S = \begin{bmatrix} f(x_1) \\ \vdots \\ f(x_n) \end{bmatrix}$ $K_S = \begin{bmatrix} k(x_1, x_1) & \cdots & \vdots \\ \vdots & & \vdots \\ k(x_n, x_1) & \cdots & k(x_n, x_n) \end{bmatrix}$

$$f_\nu = \begin{bmatrix} \sqrt{\nu(X_1)}\, f(X_1) \\ \vdots \\ \sqrt{\nu(X_n)}\, f(X_n) \end{bmatrix} \qquad (K_\nu)_{ij} = \sqrt{\nu(X_i)\nu(X_j)}\, k(X_i, X_j)$$

$$f_\nu^T f_\nu = \sum_i \nu(X_i) f(X_i)^2$$

expand this $\approx \int_{\mathcal{X}} f(x)^2 \, d\nu(x)$

using Mercer

$$\boxed{f_\nu^T K_\nu^{-1} f_\nu = f_S^T K_S^{-1} f_S}$$

$$f_\nu = \mathrm{diag}(\nu(X_i))^{1/2} f_S$$

$$K_\nu = \mathrm{diag}(\nu(X_i))^{1/2} K_S \, \mathrm{diag}(\nu(X_i))^{1/2}$$

$$f_\nu^T K_\nu^{-1} f_\nu = f_\nu^T \left(\sum_{m=1}^{\infty} \frac{1}{\mu_m} \begin{bmatrix} \nu(X_1)^{1/2} e_m(X_1) \\ \vdots \\ \nu(X_n)^{1/2} e_m(X_n) \end{bmatrix} \begin{bmatrix} \nu(X_1)^{1/2} e_m(X_1) \\ \vdots \\ \nu(X_n)^{1/2} e_m(X_n) \end{bmatrix}^T \right)^{-1} f_\nu$$

$D = \mathrm{diag}(v(X_1), \ldots, v(X_n))$

$D^{1/2} K D^{1/2} = \sum_m \mu_m e_m e_m^T$

μ_m, e_m: $K(x,y) = \sum_{m=1}^{\infty} \mu_m e_m(x) e_m(y)$

And e_m is an eigenfunction of T_K w/ eigenvalue μ_m

Note. $\begin{bmatrix} v(X_1)^{1/2} e_m(X_1) \\ \vdots \\ v(X_n)^{1/2} e_m(X_n) \end{bmatrix} = h_m$

$\langle h_m, h_\ell \rangle \approx 0$ if $m \neq \ell$

≈ 1 if $m = \ell$

$\langle h_m, h_\ell \rangle = \sum_{i=1}^{n} v(X_i) e_m(X_i) e_\ell(X_i)$

$$\approx \int e_m(x) e_l(x) \, d\nu(x)$$

$$= \mathbb{I}[m = l]$$

$$f_\nu^T K_\nu^{-1} f_\nu = f_\nu^T \left(\sum_{m=1}^{\infty} \frac{1}{\mu_m} h_m h_m^T \right) f_\nu$$

$$= \sum_{m=1}^{\infty} \frac{1}{\mu_m} \langle f_\nu, h_m \rangle^2$$

$$= \sum_{m=1}^{\infty} \frac{1}{\mu_m} \left| \sum_{i=1}^{n} \nu(x_i) f(x_i) e_m(x_i) \right|^2$$

$$\approx \sum_{m=1}^{\infty} \frac{1}{\mu_m} \left| \int_{\mathcal{X}} f(x) e_m(x) \, \nu(x) \right|^2 = \sum_{m=1}^{\infty} c_m^2 / \mu_m$$

$$\sup_x K(x,x) < \infty \quad (\le M)$$

$$\|f\|_{\mathcal{H}} < \infty$$

$$\Rightarrow \|f\|_\infty < \infty$$

$$\|f\|_{\mathcal{H}}^2 = \sup_S f_S^T K_S^{-1} f_S$$

$$\ge \sup_x f(x) K(x,x)^{-1} f(x)$$

$$\ge \frac{1}{M} \sup_x f(x)^2$$

$$\Rightarrow |f(x)| \le \|f\|_{\mathcal{H}} \cdot \sqrt{M}$$

$\nu = p^*$; eigenvalues $\mu_m \Rightarrow$ eigenvalues λ_m of S

If $K(a,a)$ is t times differentiable, this is enough to bound μ_m.

Theorem (Kühn, 1987).

Suppose $\mathcal{X}$ is compact d-dimensional smooth manifold and K has continuous t^{th} derivative. Then $\mu_m = O(m^{-t/d-1})$.

Consequences

$\nu = p^*$ $\Rightarrow$ $T_K \to \frac{1}{n}K$ on $x_1, \dots, x_n \sim p^*$ as $n \to \infty$

$\to$ (S)

eigenvalues: power law decay $m^{-\alpha}$ w/ $\alpha = \frac{t}{d} + 1$.

$\Rightarrow$ error: $n^{\frac{1}{\alpha}-1}$ if σ large $\Rightarrow$ $n^{-\frac{t}{t+d}}$

$n^{1-\alpha}$ if r small $\Rightarrow$ $n^{-t/d}$

(Classical) Most people about learning rates for these smoothness classes.

Another application: Random features models

Kernels $\longleftrightarrow$ random features

$$\varphi(x) = \begin{bmatrix} \varphi_1(x) \\ \vdots \\ \varphi_m(x) \end{bmatrix} / \sqrt{m} \qquad \varphi_j \sim P$$

$$\langle \varphi(x), \varphi(y) \rangle = \frac{1}{m} \sum_{j=1}^{m} \varphi_j(x) \varphi_j(y)$$

$$\to \mathbb{E}_{\varphi \sim P}[\varphi(x)\varphi(y)]$$

$$\exists K(x,y) = \mathbb{E}_{\varphi\sim p}[\varphi(x)\varphi(y)]$$

$$K(x,y) = \sum_\ell \mu_\ell e_\ell(x) e_\ell(y)$$

$$\varphi = \sqrt{Z}\, e_\ell \quad \text{with probability} \quad \frac{\mu_\ell}{Z}$$

$$Z = \mu_1 + \mu_2 + \cdots .$$

$$\left(\varphi \sim p \Rightarrow K = \mathbb{E}_\varphi[\varphi(x)\varphi(y)] \right.$$

$$\left. K \Rightarrow (\mu_\ell, e_\ell) \Rightarrow P(\varphi = e_\ell) \propto \mu_\ell \right)$$

Theorem (Bochner). $K(x,y) = K(x-y)$

(

$$k(x,y) = \int e^{-i\langle w, x-y\rangle} \, d\mu(w)$$

$$\mu \propto \hat{k}(w) = \int e^{i\langle w, \Delta\rangle} k(\Delta) \, d\Delta$$

Mercer: $k(x,y) = \sum_m \mu_m e_m(x) \overline{e_m(y)}$

$$e_m(x) = e^{i\langle w, x\rangle}$$

$$\sum_m \mu_m e_m(x) \overline{e_m(y)} = \int \mu_m e^{i\langle w, x-y\rangle}$$

Example 1. $k(x,y) = e^{-\|x-y\|_2^2/2\sigma^2}$

$$k(\Delta) = e^{-\|\Delta\|_2^2/2\sigma^2}$$

$$\hat{k}(w) \propto e^{-\sigma^2\|w\|_2^2/2}$$

$$k(x,y) \propto \int e^{i\langle w, x-y\rangle} e^{-\sigma^2\|w\|_2^2/2} \, dw$$

⇒ random features:

Feature j: sample $\omega_j \sim N(0, \frac{1}{\sigma^2} I)$

$$\phi_j(x) = e^{i\langle \omega_j, x \rangle}$$

← has complex numbers ...

$$\mathbb{E}_\omega[\phi(x)\overline{\phi(y)}] = k(x,y)$$

$$\phi_j(x) = \cos(\langle \omega_j, x \rangle + b), \quad b \sim \text{Uniform}[0, 2\pi]$$

Example 2.

$$k(x,y) = e^{-\|x-y\|_1/\lambda}$$

$$k(\Delta) = e^{-\|\Delta\|_1/\lambda}$$

$$\hat{k}(\omega) \propto \prod_{j=1}^{d} \frac{1}{1+\lambda^2\omega_j^2}$$

⇒ Sample ω from Cauchy distribution

$$\cos(\langle \omega, x \rangle + b)$$

Rahimi & Recht, '07