Lecture 11: Learning Under Invariances and Geometric Priors

Recap last week: Benefits of depth in NNs in terms of approximation.

Today: . Look for other sources of regularity; how to incorporate them into learning model.

. Two key principles: (i) symmetries/ invariances

(ii) multiscale separation.

Symmetries and Geometric Stability

→ So far, we have measured the regularity / complexity of a target function $f: X \to \mathbb{R}$ with norms $\gamma(f)$ of the form

$$\gamma(f) = \int |\hat{f}(\xi)|^2 (1+\|\xi\|^2)^s \, d\xi \quad \text{[Sobolev norm]}$$

$$\gamma(f) = \int |\hat{f}(\xi)| \, \|\xi\|^k \, d\xi \quad \text{[Barron norm]}$$

$$\gamma(f) = \inf \left\{ \|g\|^2_{L^2(\Theta,\tau)} \; ; \; f(x) = \int g(\theta) \, \sigma(\langle x, \theta \rangle) \, \tau(d\theta) \right\}$$

[RKHS norm]

→ So far, we have treated X as a "mere" high-dim vector space. — invariant / agnostic to how we parametrize the input.!

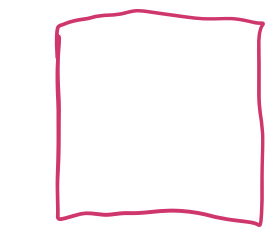
→ In most situations of interest, the input space X has much more structure:

$$X = \{ x : \Omega \to C \}$$

Ω: physical domain

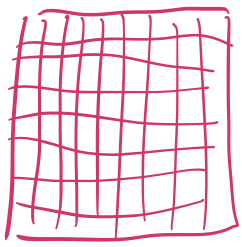
C: feature channels

ex: images



$\Omega = [0,1]^2$

discretised version



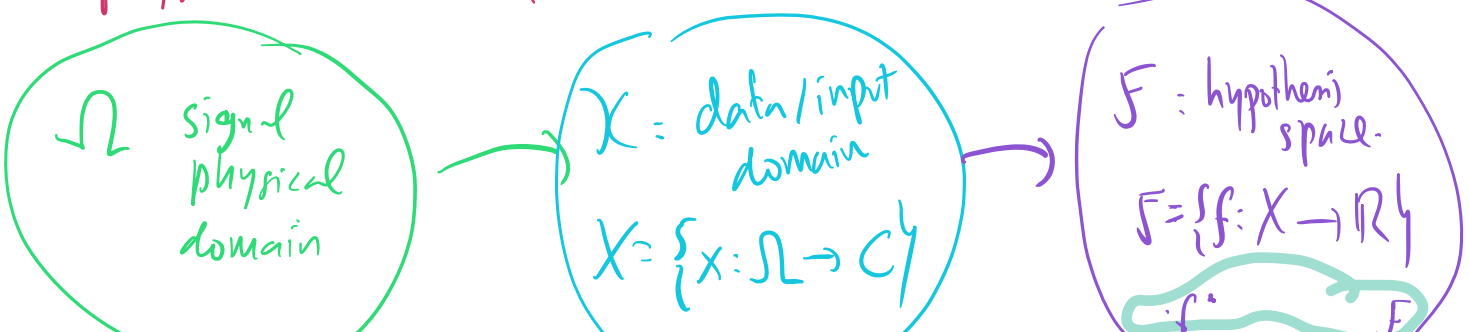
$\Omega = [1, M) \times [1, N]$

C: RGB space of colors.

- speech/music $\Omega = \mathbb{R}$
- particle physics: $\Omega = \{ (p_i, q_i) \quad i = 1 \dots N \}$ ← state of N particles.

$$= (\mathbb{R}^3 \times \mathbb{R}^3)^N$$

→ Q: How to exploit our prior knowledge of the physical domain?



low-dimensional high-dim Very-high-dim.

→ Symmetries are known transformations $T: X \to X$ that do not change "much" the function of interest f:

$T: X \to X$ such that $f(Tx) \approx f(x)$ $\forall x$ $\forall f \in \mathcal{F}$

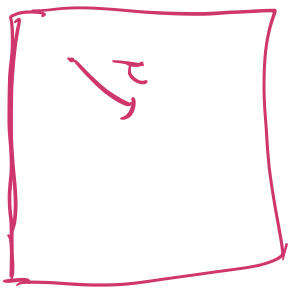
↳ symmetries might be exact or approximate.

$f(Tx) = f(x)$

$(f(Tx) = f(x)$ yet $\|Tx - x\| = \Theta(\|x\|))$ 1

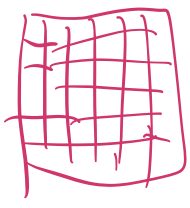
→ Rather than describing symmetries in the data space (high-dim), we rely on the physical space Ω instead. 2

→ Example: let $\Omega = [0,1]^2$ and $\tau \in \Omega$.



$$T_\tau : \Omega \to \Omega$$
$$u \mapsto u - \tau \pmod 1$$

More generally, in a discrete domain $\Omega = [1,n]^2$, we can consider transformations that move pixels around: $T: [1,n]^2 \to [1,n]^2$



if $|\Omega| = N$, T can be identified with a permutation $\sigma \in S_N = \{\sigma : [N] \to [N]$ bijection$\}$

S_N ($\pi : [N] \to [N]$ bijection)

→ A transformation of the domain Ω can be "lifted" to a linear transformation on X:

$$T : \Omega \to \Omega \quad \rightsquigarrow \quad T : X \to X$$

$$u \mapsto Tu \qquad\qquad x \mapsto (Tx) : \Omega \to \mathbb{C}$$

$$u \mapsto (Tx)(u) = x(T^{-1}u)$$

Example: $T_\tau : \Omega \to \Omega$

$$u \mapsto u + \tau$$

↳ $\tilde{T}_\tau : X \to X$

$$x \mapsto (\tilde{T}x)(u) = x(u - \tau)$$

→ observe that $\tilde{T}$ is always linear.

→ The set of exact symmetries of a hypothesis class $\mathcal{F}$

(T such that $f(Tx) = f(x) \quad \forall x, \forall f \in \mathcal{F}$)

forms a group G with composition: if T and T' are exact symmetries then $T \circ T'$ is also an exact symm.

↳ In this case, the action of the group G on X is called a group representation.

Natural Questions (i) How much can be gained by leveraging symmetries?

(ii) How to implement "symmetry-aware" NNs efficiently? ← "Geometric Deep Learning"

Sample Complexity of Learning Under Invariance

→ Setup: $\Omega = \{1, \dots, d\}$ discrete domain with d "pixels"

$X = S^{d-1}$ unit d-dimensional sphere; with $\nu = \mathrm{Unif}(S^{d-1})$

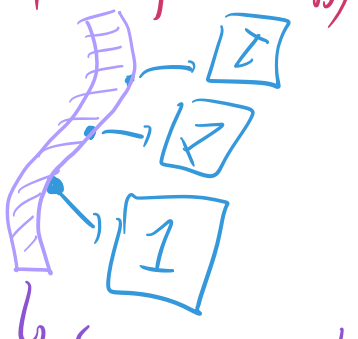
→ Permutations $\sigma : \Omega \to \Omega$ define unitary linear transformations of S^{d-1}: σ can be viewed as a $d \times d$ permutation matrix.

$$\|\sigma \cdot x\| = \|x\|. \qquad x \in S^{d-1}$$

→ Given a subset (or a subgroup, to simplify) G of the permutation group, we define the smoothing operator S_G acting on $L^2(S^{d-1}) = \{f : S^{d-1} \to \mathbb{R}\}$ (with finite energy)

$$S_G f = \frac{1}{|G|} \sum_{\sigma \in G} f(\sigma \cdot x)$$

(averaging f along each orbit).



→ f G- invariant function $\left(f \text{ st } f(\sigma x) = f(x) \quad \forall x \ \forall \sigma \in G\right)$ $\{\sigma \cdot x ; \sigma \in G\}$

$$S_G f = f$$

→ **f is G- stable if $f = S_G \tilde{f}$ for some $\tilde{f} \in H$**

→ Recall from the spherical harmonic recitation: an orthonormal basis of $L^2(S^{d-1})$

$$\{Y_{kj}\}_{k=1\dots\infty} \leftrightarrow \text{frequency magnitude.}$$
$$j=1 \dots N_{d,k} \leftrightarrow \text{frequency "direction".}$$

$$N_{d,k} \sim (2k+d-2)\binom{k+d-3}{d-2}$$

$\{Y_{kj}\}_j$ are harmonic polynomials of degree k.

$$f = \sum_{k=1}^{\infty} \sum_{j=1}^{N_{d,k}} \langle f, Y_{j,k} \rangle \cdot Y_{j,k}$$

→ Dot-product kernels $K(x,x') = \varphi(\langle x, x' \rangle)$ ← non-linear function. define an RKHS $H_\varphi \subset L^2(S^{d-1})$, given by functions

$$H_\varphi = \left\{ f = \sum_{j,k} \alpha_{j,k} Y_{j,k} \ ; \ \sum_{k,j} \mu_k^{-1} \alpha_{j,k}^2 < +\infty \right\}$$

$$\mu_k = C_d \cdot \int_{-1}^{1} \varphi(t) \cdot \left(P_{k,d}(t)\sqrt{(1-t^2)}^{(d-3)}\right) dt$$

-1

↳ Legendre / Gegenbauer polynomial.

$P_{k,d} : [-1, 1] \to \mathbb{R}$

$\varphi : [-1, 1] \to \mathbb{R}$.

Fact: [Mei, Misiakiewicz, Montanari '21] The smoothing operator leaves each harmonic space $V_k = \text{span}\{Y_{k,j}\}_j$ invariant: $S_G V_k = \bar{V}_k \subseteq V_k$.

$\bar{V}_k$: harmonic polynomials of degree k that are G-invariant.

$\bar{N}_{d,k}$: dimension of $\bar{V}_k$.

$p(x) = x_1 x_2$

↓

$\frac{1}{|G|} \sum_{\sigma} x_{\sigma(1)} x_{\sigma(2)}$

Fact: [MMM'21] : We have

$$\bar{N}_{d,k} = N_{d,k} \left[\frac{1}{|G|} \sum_{\sigma \in G} \mathbb{E}_{x \sim \nu} \left[P_{d,k}(\langle x, \sigma \cdot x \rangle) \right] \right]$$

$1 + \sum_{\sigma \neq \mathrm{Id}} \cdots$

$\gamma_d(k)$: : fraction of degree-k harmonic polynomials that are G-invariant.

(Moral: $\gamma_d(k) \sim \frac{1}{|G|}$)

Q: If our goal is to learn f^*, and we have the promise that f^* is G-invariant, how can

we adapt the kernel?

→ G- invariant kernel associated with K_φ:

$$K_\varphi^G(x,x') = \frac{1}{|G|} \sum_{\sigma \in G} \varphi(\langle x, \sigma \cdot x' \rangle)$$

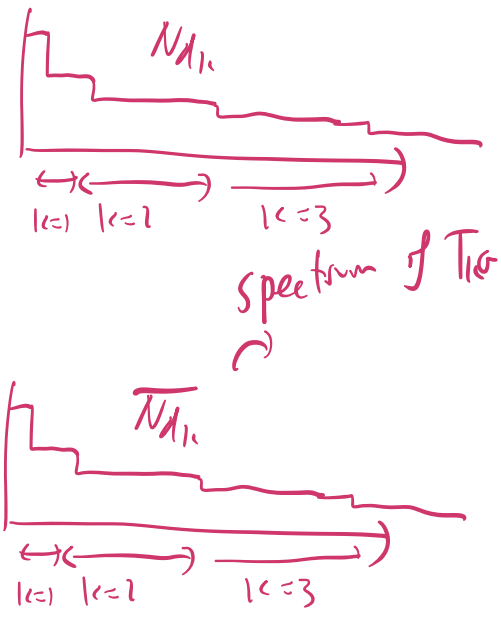
Remark: This is not a user-friendly formula!

→ Associated Integral operators:

$$T_k f(x) = \int K(x,x') f(x') \nu(dx')$$

$$T_{kG} f(x) = \int K_G(x,x') f(x') \nu(dx').$$

↳ T_k and T_{kG} share the same eigenfunctions / eigenvalues; only difference is the multiplicity ($N_{d,k} \rightsquigarrow \overline{N}_{d,k}$).



→ Let's assume that f^* is G- invariant. Then the approximation error is not affected by smoothing:

$$\inf \|f - f^*\|^2 + \lambda \|f\|^2_{\mathcal{H}}$$

$f \in \mathcal{H}_K$ $\qquad$ $\| \cdot \|_{\mathcal{H}_K}$

$$\inf_{f \in \mathcal{H}_{KG}} \|f - f^*\|^2 + \lambda \|f\|^2_{\mathcal{H}_{KG}}$$

→ Generalisation error of kernel ridge regression can be studied using the degrees of freedom:

$$V_K = \overline{V}_K \oplus (V_K \ominus \overline{V}_K)$$

N_{dK} $\qquad$ ∞ $N_{dK/G}$

$$\mathcal{N}_K(\lambda) = \mathrm{Tr}\left(\Sigma_K \left(\Sigma_K + \lambda I \right)^{-1} \right) = \sum_m \frac{\lambda_m}{\lambda_m + \lambda}$$

$\{\lambda_m\}_m$ are the eigenvalues of T_K.

$\mathcal{N}_K(\lambda)$ measures "how big" is the RKHS.

→ Punchline: Invariant kernel has smaller degrees of freedom

$$\mathcal{N}_{KG}(\lambda) \leq \left(\sup_x \gamma_d(x) \right) \cdot \mathcal{N}_K(\lambda)$$

↳ fraction of invariant harmonic polynomials.

→ Theorem [Bietti, B, Venturi '21]

Assume (i) $f^* = T_K^r g$ for $\|g\|_{L^2} \leq C$ (source cond)

(ii) $\mathcal{N}_K(\lambda) \leq C_K \lambda^{-1/\alpha}$ $\alpha > 1$ (capacity cond)

(iii) f^* is G-invariant. Then

$$\mathbb{E}\,\|\hat{f}_{\lambda,G} - f^\circ\| \simeq \left(\frac{1}{|G|\cdot n}\right)^{\frac{2\alpha r}{2\alpha r+1}}$$ using K_G,

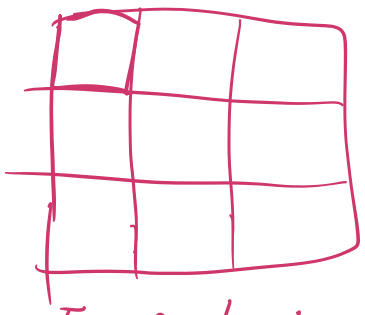
instead of $$\mathbb{E}\,\|\hat{f}_\lambda - f^\circ\| \simeq \left(\frac{1}{n}\right)^{\frac{2\alpha r}{2\alpha r+1}}$$

Remarks: ⊕ Confirm that learning with K_G is more sample efficient, $n \rightsquigarrow n_{\text{eff}} = n\cdot|G|$

$|G|$ can be either

small ($|G| = d$ for 1d-shifts)

or potentially exponential in dimension (local translations).



⊖ Invariance cannot change the rate. In particular, if f° is Lipschitz and invariant,

$$\frac{2\alpha r}{2\alpha r+1} \sim \frac{1}{d}$$

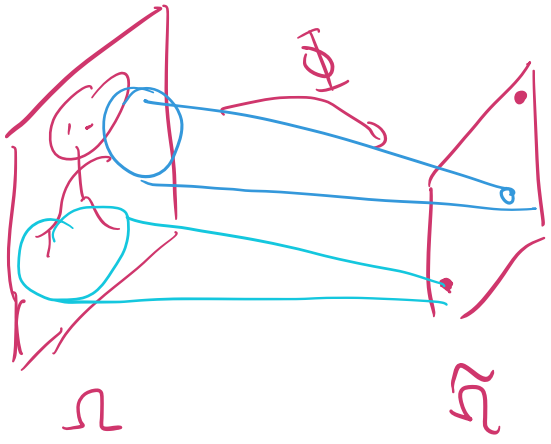
$$\left(\frac{1}{|G|n}\right)^{1/d}$$, so learning is still cursed by dimension, even when $|G|$ is exp(d).

→ Conclusion: To break the curse of dim, we need another prior on top of symmetry/invariance.

→ What is missing? Scale separation / Multiscale prior.

↕

"compositionality"



$\tilde{f}$

.

$f \approx \tilde{f} \circ \phi$