On the Inductive Bias of Neural Tangent Kernels

Alberto Bietti    Julien Mairal    -    Inria Grenoble

**Inductive Bias and Over-Parameterization**

- Over-parameterized deep networks are very expressive
- Optimization algorithm plays a crucial role for generalization

**Lazy Training**: In certain regimes (over-parameterization, particular initialization), neural networks behave like their linearization near initialization

\[ f(x; \theta) = f(x; \theta_0) + (\theta - \theta_0, \nabla f(x; \theta_0)) \]

**Neural Tangent Kernels (NTK)**: In this regime, generalization properties are controlled by the limiting kernel \( \kappa \) [Jacot et al., 2018]

\( \nabla f(x; \theta_0, \nabla f(x', \theta_0)) \to K(x, x') \)

In particular, with squared loss and infinite width, we get the interpolating solution with minimum RKHS norm.

**Contributions**:
- Derivation of NTK for convolutional networks with a generic linear patch extraction/pooling operators
- Study of smoothness, stability, and approximation properties of functions with finite RKHS norm
- Comparison to other ReLU kernels (e.g., training only last layer with random weights): the NTK has weaker smoothness properties but better approximation.

**Approximation Properties (two layers)**

**Question**: How rich is the RKHS for the NTK \( \kappa_{NTK} \) versus the simpler kernel \( \kappa_1 \) obtained by training just the second layer (random features)?

**Mercer decomposition with spherical harmonics**

**Proposition (Mercer decomposition)**

For any \( x, y \in S^{d-1} \), we have the following decomposition of the NTK \( \kappa_{NTK} \):

\[
\kappa_{NTK}(x, y) = \sum_{K=0}^{\infty} \sum_{j=1}^{N_K} \sum_{k=1}^{N_j} Y_k(x) Y_j(y),
\]

where \( Y_k \) are spherical harmonic polynomials of degree \( k \), and the non-negative eigenvalues \( \mu_k \) satisfy \( \mu_0 > \mu_1 > 0 \), \( \mu_k = 0 \) if \( k < 2j + 1 \) and \( j \geq 1 \), and otherwise \( \mu_k \sim C(p)k^{d-p} \) as \( k \to \infty \).

This gives an explicit characterization of the RKHS norm of a function.

**Approximation results**:
- The RKHS is “larger”: slower decay compared to \( \kappa_1 \), for which \( \mu_k = O(k^{d-p}) \):
- \( f \) with \( \phi \)-bounded derivatives \( \implies f \in \mathcal{H} \) with \( \|f\| \leq \mathcal{O}(\phi) \);
- Weaker requirement compared to \( \kappa_1 \) (need \( p/2 + 1 \) derivatives);
- Better rates for approximating Lipschitz functions on the sphere.

**Relevant References**

**Smoothness and Deformation Stability**

**Two-layer ReLU networks**: The NTK (when training both layers) has weaker smoothness compared to training only the second layer.

**Proposition (Non-Lipschitzness)**

The kernel mapping \( \Phi(x) \) of the two-layer NTK is not Lipschitz:

\[ \sup_{x, y} \| \Phi(x) - \Phi(y) \|_x \to +\infty. \]

It follows that the RKHS \( \mathcal{H} \) contains unit-norm functions with arbitrarily large Lipschitz constant.

**Proposition (Smoothness for ReLU NTK)**

The kernel mapping \( \Phi \) satisfies

\[
\| \Phi(x) - \Phi(y) \| \leq \sqrt{\min(\|x\|, \|y\|)} \|x - y\| + 2 \|x - y\|.
\]

**Deformation stability for deep ReLU CNNs**: Similar assumptions to [Bietti and Mairal, 2019]:
- Continuous signals \( x(u) \) in \( L^2(\mathbb{R}^d) \), \( t : \mathbb{R}^d \to \mathbb{R}^d \), \( C^1 \), deformations \( L \cdot x(u) = x(u - \tau(v)) \);
- Anti-aliasing of the original signal: \( A \cdot x \) instead of \( x \);
- Patch sizes controlled at current resolution: \( \sup_{v \in S_k} \|v\| \leq \beta \sigma_{k-1} \)

**Proposition (Stability of NTK)**

Let \( \Phi_d(x) = \Phi_A(x) \), and assume \( \|\nabla \tau\|_{L^\infty} \leq 1/2 \). We have:

\[
\|\Phi_d(L \cdot x) - \Phi_d(x)\| \leq \left( C_1 \|\nabla \tau\|_{L^\infty}^2 + C_2 \|\tau\|_{L^\infty} + \frac{C_3}{\sigma_0} \|\tau\|_{L^\infty} \right) \|x\|
\]

Worse dependence on \( \|\nabla \tau\|_{L^\infty} \) for small deformations compared to CKN/random feature kernel obtained when training just the last layer!