A Kernel Perspective for Regularizing Deep Neural Networks

Alberto Bietti* Grégoire Mialon* Dexiong Chen Julien Mairal

Inria

ICML 2019, Long Beach

Regularization in Deep Learning

Two issues with today's deep learning models:

- Poor performance on **small datasets**
- **Lack of robustness** to adversarial perturbations

Regularization in Deep Learning

Two issues with today's deep learning models:

- Poor performance on **small datasets**
- **Lack of robustness** to adversarial perturbations

Questions:

- Can **regularization** address this?

$$\min_f \frac{1}{n} \sum_{i=1}^{n} \ell(y_i, f(x_i)) + \lambda \Omega(f)$$

- What is a **good choice** of $\Omega(f)$ for deep (convolutional) networks?

Regularization with the RKHS Norm

Kernel methods: $f(x) = \langle f, \Phi(x) \rangle_{\mathcal{H}}$

- $\Phi(x)$ captures useful **properties of the data**
- $\|f\|_{\mathcal{H}}$ controls **model complexity** and **smoothness**:

$$|f(x) - f(y)| \leq \|f\|_{\mathcal{H}} \cdot \|\Phi(x) - \Phi(y)\|_{\mathcal{H}}$$

Regularization with the RKHS Norm

Kernel methods: $f(x) = \langle f, \Phi(x) \rangle_{\mathcal{H}}$

- $\Phi(x)$ captures useful **properties of the data**
- $\|f\|_{\mathcal{H}}$ controls **model complexity** and **smoothness**:

$$|f(x) - f(y)| \leq \|f\|_{\mathcal{H}} \cdot \|\Phi(x) - \Phi(y)\|_{\mathcal{H}}$$

Our work: view generic CNN f_θ as an element of a RKHS $\mathcal{H}$ and regularize using $\|f_\theta\|_{\mathcal{H}}$

Regularization with the RKHS Norm

Kernel methods: $f(x) = \langle f, \Phi(x) \rangle_{\mathcal{H}}$

- $\Phi(x)$ captures useful **properties of the data**
- $\|f\|_{\mathcal{H}}$ controls **model complexity** and **smoothness**:

$$|f(x) - f(y)| \leq \|f\|_{\mathcal{H}} \cdot \|\Phi(x) - \Phi(y)\|_{\mathcal{H}}$$

Our work: view generic CNN f_θ as an element of a RKHS $\mathcal{H}$ and regularize using $\|f_\theta\|_{\mathcal{H}}$

Kernels for deep convolutional architectures (Bietti and Mairal, 2019):

Regularization with the RKHS Norm

Kernel methods: $f(x) = \langle f, \Phi(x) \rangle_{\mathcal{H}}$

- $\Phi(x)$ captures useful **properties of the data**
- $\|f\|_{\mathcal{H}}$ controls **model complexity** and **smoothness**:

$$|f(x) - f(y)| \leq \|f\|_{\mathcal{H}} \cdot \|\Phi(x) - \Phi(y)\|_{\mathcal{H}}$$

Our work: view generic CNN f_θ as an element of a RKHS $\mathcal{H}$ and regularize using $\|f_\theta\|_{\mathcal{H}}$

Kernels for deep convolutional architectures (Bietti and Mairal, 2019):

- $\|\Phi(x) - \Phi(y)\|_{\mathcal{H}} \leq \|x - y\|_2$
- $\|\Phi(x_\tau) - \Phi(x)\|_{\mathcal{H}} \leq C(\tau)$ for a small transformation x_τ of x

Regularization with the RKHS Norm

Kernel methods: $f(x) = \langle f, \Phi(x) \rangle_{\mathcal{H}}$

- $\Phi(x)$ captures useful **properties of the data**
- $\|f\|_{\mathcal{H}}$ controls **model complexity** and **smoothness**:

$$|f(x) - f(y)| \leq \|f\|_{\mathcal{H}} \cdot \|\Phi(x) - \Phi(y)\|_{\mathcal{H}}$$

Our work: view generic CNN f_θ as an element of a RKHS $\mathcal{H}$ and regularize using $\|f_\theta\|_{\mathcal{H}}$

Kernels for deep convolutional architectures (Bietti and Mairal, 2019):

- $\|\Phi(x) - \Phi(y)\|_{\mathcal{H}} \leq \|x - y\|_2$
- $\|\Phi(x_\tau) - \Phi(x)\|_{\mathcal{H}} \leq C(\tau)$ for a small transformation x_τ of x
- CNNs f_θ with ReLUs are (approximately) **in the RKHS** with norm

$$\|f_\theta\|_{\mathcal{H}}^2 \leq \omega(\|W_1\|_2, \ldots, \|W_L\|_2).$$

Approximating the RKHS norm

Our approach: use upper and lower bound approximations of $\|f\|_{\mathcal{H}}$

Approximating the RKHS norm

Our approach: use upper and lower bound approximations of $\|f\|_{\mathcal{H}}$

- **Upper bound**: constraint/penalty on spectral norms

Approximating the RKHS norm

Our approach: use upper and lower bound approximations of $\|f\|_{\mathcal{H}}$

- **Upper bound**: constraint/penalty on spectral norms
- **Lower bounds**: use $\|f\|_{\mathcal{H}} = \sup_{\|u\|_{\mathcal{H}} \leq 1} \langle f, u \rangle_{\mathcal{H}}$
  $\implies$ consider tractable subsets of the RKHS unit ball

Approximating the RKHS norm

Our approach: use upper and lower bound approximations of $\|f\|_{\mathcal{H}}$

- **Upper bound**: constraint/penalty on spectral norms
- **Lower bounds**: use $\|f\|_{\mathcal{H}} = \sup_{\|u\|_{\mathcal{H}} \leq 1} \langle f, u \rangle_{\mathcal{H}}$
  $\implies$ consider tractable subsets of the RKHS unit ball

$$\|f\|_{\mathcal{H}} \geq \sup_{x, \|\delta\| \leq 1} \langle f, \Phi(x+\delta) - \Phi(x) \rangle_{\mathcal{H}} \quad \text{(adversarial perturbations)}$$

Approximating the RKHS norm

Our approach: use upper and lower bound approximations of $\|f\|_{\mathcal{H}}$

- **Upper bound**: constraint/penalty on spectral norms
- **Lower bounds**: use $\|f\|_{\mathcal{H}} = \sup_{\|u\|_{\mathcal{H}} \leq 1} \langle f, u \rangle_{\mathcal{H}}$
  $\Longrightarrow$ consider tractable subsets of the RKHS unit ball

$$\|f\|_{\mathcal{H}} \geq \sup_{x, \|\delta\| \leq 1} f(x + \delta) - f(x) \quad \text{(adversarial perturbations)}$$

Approximating the RKHS norm

Our approach: use upper and lower bound approximations of $\|f\|_{\mathcal{H}}$

- **Upper bound**: constraint/penalty on spectral norms
- **Lower bounds**: use $\|f\|_{\mathcal{H}} = \sup_{\|u\|_{\mathcal{H}} \leq 1} \langle f, u \rangle_{\mathcal{H}}$
  $\implies$ consider tractable subsets of the RKHS unit ball

$$\|f\|_{\mathcal{H}} \geq \sup_{x, \|\delta\| \leq 1} f(x + \delta) - f(x) \quad \text{(adversarial perturbations)}$$

$$\|f\|_{\mathcal{H}} \geq \sup_{x, C(\tau) \leq 1} f(x_\tau) - f(x) \quad \text{(adversarial deformations)}$$

Approximating the RKHS norm

Our approach: use upper and lower bound approximations of $\|f\|_{\mathcal{H}}$

- **Upper bound**: constraint/penalty on spectral norms
- **Lower bounds**: use $\|f\|_{\mathcal{H}} = \sup_{\|u\|_{\mathcal{H}} \leq 1} \langle f, u \rangle_{\mathcal{H}}$
  $\implies$ consider tractable subsets of the RKHS unit ball

$$\|f\|_{\mathcal{H}} \geq \sup_{x, \|\delta\| \leq 1} f(x + \delta) - f(x) \quad \text{(adversarial perturbations)}$$

$$\|f\|_{\mathcal{H}} \geq \sup_{x, C(\tau) \leq 1} f(x_\tau) - f(x) \quad \text{(adversarial deformations)}$$

$$\|f\|_{\mathcal{H}} \geq \sup_{x} \|\nabla f(x)\|_2 \quad \text{(gradient penalty)}$$

Approximating the RKHS norm

Our approach: use upper and lower bound approximations of $\|f\|_{\mathcal{H}}$

- **Upper bound**: constraint/penalty on spectral norms
- **Lower bounds**: use $\|f\|_{\mathcal{H}} = \sup_{\|u\|_{\mathcal{H}} \leq 1} \langle f, u \rangle_{\mathcal{H}}$
  $\implies$ consider tractable subsets of the RKHS unit ball

$$\|f\|_{\mathcal{H}} \geq \sup_{x, \|\delta\| \leq 1} f(x + \delta) - f(x) \quad \text{(adversarial perturbations)}$$

$$\|f\|_{\mathcal{H}} \geq \sup_{x, C(\tau) \leq 1} f(x_\tau) - f(x) \quad \text{(adversarial deformations)}$$

$$\|f\|_{\mathcal{H}} \geq \sup_{x} \|\nabla f(x)\|_2 \quad \text{(gradient penalty)}$$

- Best performance by **combining upper + lower** bound approaches

More Perspectives and Experiments

Regularization approaches

- **Unified view** on various existing strategies, including links with **robust optimization**

Theoretical insights

- Guarantees on **adversarial generalization** with margin bounds
- Insights on regularization for training generative models

Experiments

- Improved performance on small data scenarios in vision and biological datasets
- Robustness benefits with large adversarial perturbations

Poster #223