A Kernel Perspective for Regularizing Deep Neural Networks

Alberto Bietti* Grégoire Mialon* Dexiong Chen Julien Mairal - Inria

Inria · Microsoft Research - Inria JOINT CENTRE · ENS · Université Grenoble Alpes

Motivation

Issues with Deep Models

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

Can regularization address this?

$$\min_\theta \frac{1}{n}\sum_{i=1}^n \ell(y_i, f_\theta(x_i)) + \lambda\Omega(f_\theta) \quad \text{or} \quad \min_\theta \frac{1}{n}\sum_{i=1}^n \ell(y_i, f_\theta(x_i)) \text{ s.t. } \Omega(f_\theta) \leq C$$

- What is a good $\Omega(f_\theta)$ when f_θ is a (convolutional) neural network?

Our approach:

- View neural network f_θ as element of RKHS for a suitable kernel
- Regularize using (approximations of) the RKHS norm
- We recover existing strategies and obtain new ones

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** (generalization) and **smoothness**:

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

Kernels for deep convolutional networks [Bietti and Mairal, 2019]
For a given CNN architecture, we may define a corresponding multi-layer hierarchical kernel, and RKHS $\mathcal{H}$.

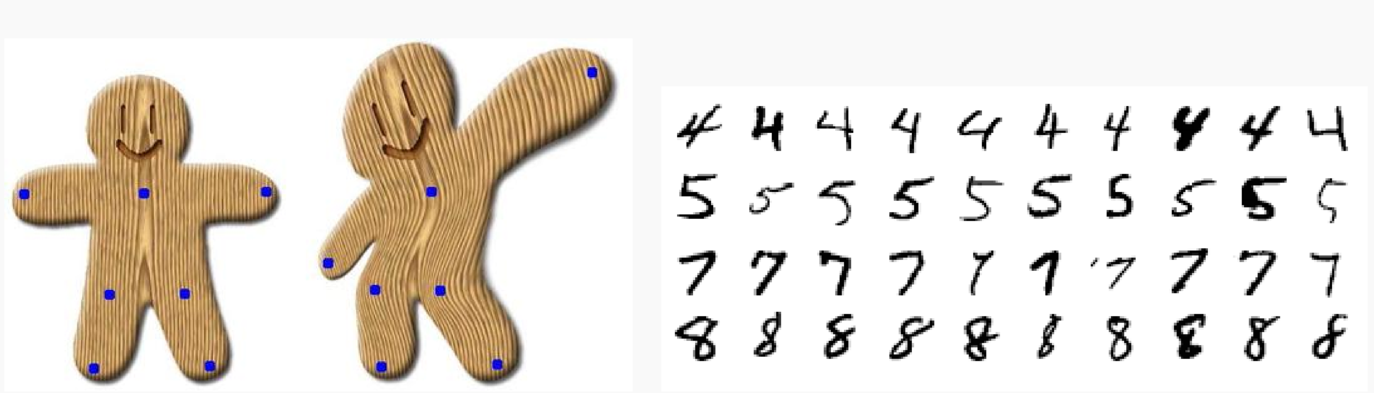
Properties of $\Phi(\cdot)$:

- **Non-expansiveness** (robustness to additive perturbations):

$$\|\Phi(x) - \Phi(y)\|_{\mathcal{H}} \leq \|x - y\|_2.$$

- **Stability to deformations** τ (*e.g.*, translations, rotations etc):

$$\|\Phi(x_\tau) - \Phi(x)\| \leq C(\tau)\|x\|_2$$



Norm of a given CNN f_θ. Consider a ReLU CNN f_θ with parameters $\theta = (W_1, \dots, W_L)$. f_θ is approximately **in the RKHS**, with norm

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

$\|W_k\|_2$ are spectral norms, and ω is increasing.

Regularize generic networks using this norm?

- Unlike traditional kernel methods, $\|f_\theta\|_{\mathcal{H}}$ is **intractable**
- $\implies$ use upper/lower bound **approximations**

RKHS norm approximations: lower and upper bounds

Lower bounds: Use the variational form of Hilbert norms:

$$\|f\|_{\mathcal{H}} = \sup_{\|u\|_{\mathcal{H}} \leq 1} \langle f, u\rangle_{\mathcal{H}}$$

Make it tractable by considering **subsets of the unit ball** $\bar{U} \subset B_{\mathcal{H}}(1)$

Adversarial perturbations: $\bar{U} = \{\Phi(x+\delta) - \Phi(x) : x \in X, \|\delta\|_2 \leq 1\}$

$$\|f\|_{\mathcal{H}} \geq \|f\|_\delta := \sup_{x, \|\delta\|_2 \leq 1} f(x+\delta) - f(x).$$

Similar to adversarial training (PGD), but decoupled from the loss, encourages **global** robustness instead of local only.

Adversarial deformations: $\bar{U} = \{\Phi(x_\tau) - \Phi(x) : x \in X, C(\tau) \leq 1\}$

$$\|f\|_{\mathcal{H}} \geq \|f\|_\tau := \sup_{x, C(\tau) \leq 1} f(x_\tau) - f(x).$$

Gradient penalties: $\bar{U} = \{\frac{\Phi(x) - \Phi(y)}{\|x-y\|_2} : x, y \in X\}$

$$\|f\|_{\mathcal{H}} \geq \|\nabla f\| := \sup_x \|\nabla f(x)\|_2 \quad (= \|f\|_{\text{Lip}})$$

Recently used for regularizing GANs. Also related to double backpropagation (gradient on the loss instead of predictions).

Link with robust optimization: Another lower bound

$$\frac{1}{n}\sum_{i=1}^n \sup_{\|\delta\|_2 \leq \epsilon} \ell(y_i, f(x_i + \delta)) \leq \frac{1}{n}\sum_{i=1}^n \ell(y_i, f(x_i)) + \epsilon\|f\|_{\mathcal{H}}$$

But: $\|f\|_{\mathcal{H}}$ may be poorly controlled in favor of data fit.

Upper bounds:
Control spectral norms $\|W_k\|_2$ using **penalties** or **constraints**.

Combined approaches: lower bound + spectral norm constraint

Theoretical Guarantees and Insights

Guarantees on *adversarial generalization*: upper bound on test error in the presence of adversarial perturbations:

$$\text{err}_{\mathcal{D}}(f, \epsilon) := P_{(x,y)\sim\mathcal{D}}(\exists \|\delta\|_2 \leq \epsilon : \; yf(x+\delta) < 0).$$

Theorem (Robust margin bound)

With prob. $1 - \delta$ *we have, for all* $\gamma > 0$ *and* $f \in \mathcal{H}$,

$$\text{err}_{\mathcal{D}}(f, \epsilon) \leq L_n^{\gamma + 2\epsilon\|f\|_{\mathcal{H}}}(f) + \tilde{O}\left(\frac{\|f\|_{\mathcal{H}} R}{\gamma\sqrt{n}}\right),$$

with $L_n^\gamma(f) := \frac{1}{n}\sum_{i=1}^n \mathbf{1}\{y_i f(x_i) < \gamma\}$, $R = \max_i \|\Phi(x_i)\|_{\mathcal{H}}$

Insights on regularizing GANs: MMD objective is

$$\min_\phi \sup_{\|f\|_{\mathcal{H}} \leq 1} \mathbb{E}_{x\sim D_x}[f(x)] - \mathbb{E}_{z\sim D_z}[f(G_\phi(z))].$$

Suggests using CNN discriminators f with spectral norm constraints. Similar form to Wasserstein GAN, but better sample complexity!

Experiments on Small Datasets

Vision datasets with few training examples:

CIFAR10 (with/without data augm)

Method	1k VGG-11	1k ResNet-18
No weight decay	50.70 / 43.75	45.23 / 37.12
Weight decay	51.32 / 43.95	44.85 / 37.09
SN penalty (PI)	54.64 / 45.06	47.01 / 39.63
SN projection	54.14 / **46.70**	47.12 / 37.28
VAT	50.88 / 43.36	47.47 / 42.82
PGD-ℓ_2	51.25 / 44.40	45.80 / 41.87
grad-ℓ_2	**55.19** / 43.88	**49.30** / **44.65**
$\|f\|_\delta^2$ penalty	51.41 / 45.07	48.73 / 43.72
$\|\nabla f\|^2$ penalty	54.80 / 46.37	**48.99** / **44.97**
PGD-ℓ_2 + SN proj	54.19 / **46.66**	47.47 / 41.25
grad-ℓ_2 + SN proj	**55.32** / **46.88**	48.73 / 42.78
$\|f\|_\delta^2$ + SN proj	54.02 / **46.72**	48.12 / 43.56
$\|\nabla f\|^2$ + SN proj	**55.24** / **46.80**	**49.06** / **44.92**

infinite MNIST (* = data augmentation)

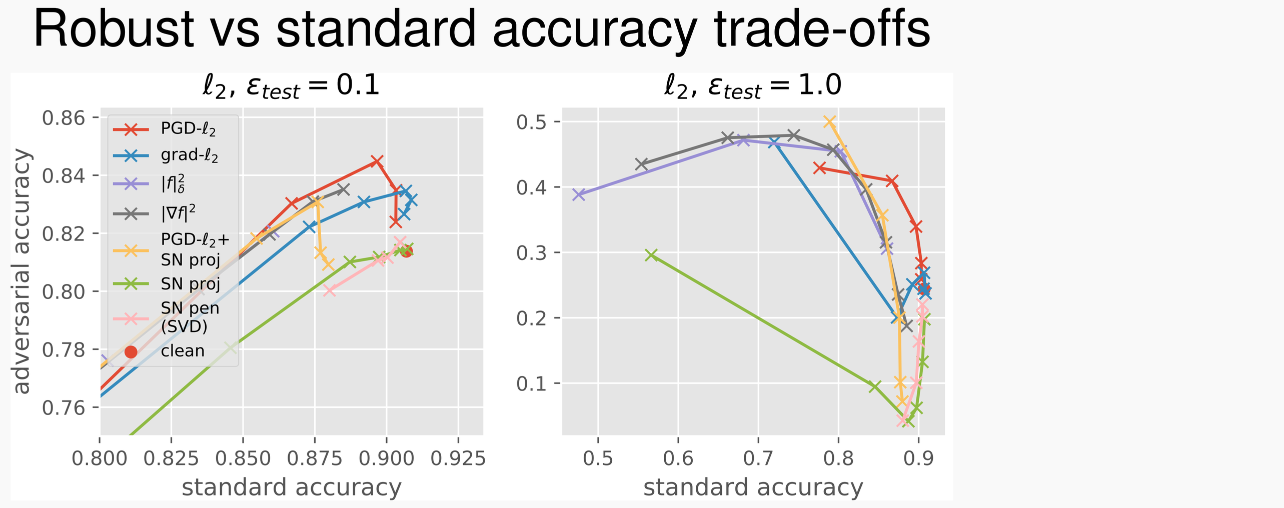
Method	300 VGG	1k VGG
Weight decay	89.32	94.08
SN projection	90.69	95.01
grad-ℓ_2	93.63	96.67
$\|f\|_\delta^2$ penalty	94.17	96.99
$\|\nabla f\|^2$ penalty	94.08	96.82
Weight decay $(*)$	92.41	95.64
grad-ℓ_2 $(*)$	95.05	97.48
$\|D_\tau f\|^2$ penalty	94.18	96.98
$\|f\|_\tau^2$ penalty	94.42	97.13
$\|f\|_\tau^2 + \|\nabla f\|^2$	94.75	97.40
$\|f\|_\tau^2 + \|f\|_\delta^2$	95.23	**97.66**
$\|f\|_\tau^2 + \|f\|_\delta^2$ $(*)$	**95.53**	**97.56**
$\|f\|_\tau^2 + \|f\|_\delta^2$ + SN proj	95.20	**97.60**
$\|f\|_\tau^2 + \|f\|_\delta^2$ + SN proj $(*)$	**95.40**	**97.77**

Protein homology detection: 102 datasets with 100 protein sequences each. Hyperparameters tuned on half the datasets, we report average auROC50 on the other half (DA = data augmentation)

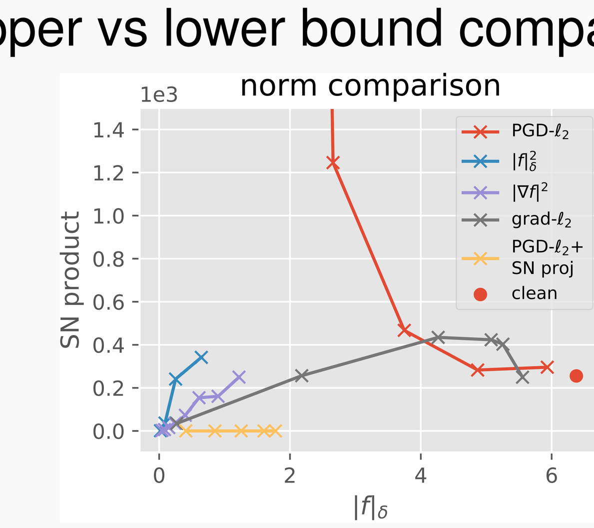
Method	No DA	DA
No weight decay	0.421	0.541
Weight decay	0.432	0.544
SN proj	0.583	0.615
PGD-ℓ_2	0.488	0.554
grad-ℓ_2	0.551	0.570
$\|f\|_\delta^2$	0.577	0.611
$\|\nabla f\|^2$	0.566	0.598
PGD-ℓ_2 + SN proj	**0.615**	**0.622**
grad-ℓ_2 + SN proj	0.581	**0.634**
$\|f\|_\delta^2$ + SN proj	**0.631**	**0.639**
$\|\nabla f\|^2$ + SN proj	0.576	**0.617**

Robustness Experiments

Robust vs standard accuracy trade-offs



Upper vs lower bound comparison



Relevant References

A. Bietti and J. Mairal (2019).
Group Invariance, Stability to Deformations, and Complexity of Deep Convolutional Representations.