Consistency of Orlicz Random Fourier Features

Zoltán Szabó – CMAP, École Polytechnique

Joint work with:

- Linda Chamakh @ CMAP & BNP Paribas
- Emmanuel Gobet @ CMAP

EPFL
Lausanne, Switzerland
September 23, 2019

Focus (high level)

- Task: speed up kernel machines on $\mathbb{R}^d$.
- Technique: random Fourier features.
- Interest: high-order derivatives.

Kernel k, RKHS $\mathcal{H}_k \leftarrow$ generalization of $\mathbf{a}^T\mathbf{b}$

Given: $\mathcal{X}$ set. $\mathcal{H}$(ilbert space).

- Kernel:

$$k(a,b) = \langle \varphi(a), \varphi(b) \rangle_{\mathcal{H}}, \quad (\forall a, b \in \mathcal{X}).$$

Kernel k, RKHS $\mathcal{H}_k \leftarrow$ generalization of $\mathbf{a}^T\mathbf{b}$

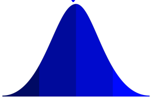
Given: $\mathcal{X}$ set. $\mathcal{H}$(ilbert space).

- Kernel:

$$k(a,b) = \langle \varphi(a), \varphi(b) \rangle_{\mathcal{H}}, \quad (\forall a, b \in \mathcal{X}).$$

- Reproducing kernel of a $\mathcal{H} \subset \mathbb{R}^{\mathcal{X}}$:

$$\underbrace{k(\cdot, a)} \in \mathcal{H}, \qquad \underbrace{f(b) = \langle f, k(\cdot, b) \rangle_{\mathcal{H}}}_{\text{reproducing property}}.$$



Kernel k, RKHS $\mathcal{H}_k \leftarrow$ generalization of $\mathbf{a}^T\mathbf{b}$

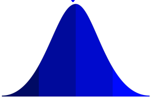
Given: $\mathcal{X}$ set. $\mathcal{H}$(ilbert space).

- Kernel:

$$k(a,b) = \langle \varphi(a), \varphi(b) \rangle_{\mathcal{H}}, \quad (\forall a, b \in \mathcal{X}).$$

- Reproducing kernel of a $\mathcal{H} \subset \mathbb{R}^{\mathcal{X}}$:

$$\underbrace{k(\cdot, a)} \in \mathcal{H}, \qquad \underbrace{f(b) = \langle f, k(\cdot, b) \rangle_{\mathcal{H}}}_{\text{reproducing property}}.$$



$$\xrightarrow{\text{spec.}} k(a,b) = \langle k(\cdot, a), k(\cdot, b) \rangle_{\mathcal{H}}.$$

Kernel k, RKHS $\mathcal{H}_k \leftarrow$ generalization of $\mathbf{a}^T\mathbf{b}$

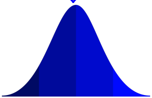
Given: $\mathcal{X}$ set. $\mathcal{H}$(ilbert space).

- Kernel:

$$k(a,b) = \langle \varphi(a), \varphi(b) \rangle_{\mathcal{H}}, \quad (\forall a, b \in \mathcal{X}).$$

- Reproducing kernel of a $\mathcal{H} \subset \mathbb{R}^{\mathcal{X}}$:

$$\underbrace{k(\cdot,a)} \in \mathcal{H}, \qquad \underbrace{f(b) = \langle f, k(\cdot,b) \rangle_{\mathcal{H}}}_{\text{reproducing property}}.$$



$$\xrightarrow{\text{spec.}} k(a,b) = \langle k(\cdot,a), k(\cdot,b) \rangle_{\mathcal{H}}. \quad \boxed{\mathcal{H}_k = \overline{\left\{\sum_{i=1}^n \alpha_i k(\cdot,x_i)\right\}}}.$$

Kernel: continued

- Def-1 (feature space):

$$k(a,b) = \langle \varphi(a), \varphi(b) \rangle_{\mathcal{H}}.$$

- Def-2 (reproducing kernel, constructive):

$$k(\cdot,b) \in \mathcal{H}, \qquad f(b) = \langle f, k(\cdot,b) \rangle_{\mathcal{H}}.$$

Kernel: continued

- Def-1 (feature space):

$$k(a,b) = \langle \varphi(a), \varphi(b) \rangle_{\mathcal{H}}.$$

- Def-2 (reproducing kernel, constructive):

$$k(\cdot,b) \in \mathcal{H}, \qquad f(b) = \langle f, k(\cdot,b) \rangle_{\mathcal{H}}.$$

- Def-3 (Gram matrix): $\mathbf{G} = [k(x_i,x_j)]_{i,j=1}^{n} \in \mathbb{R}^{n\times n} \succeq 0$.

Kernel: continued

- Def-1 (feature space):

$$k(a,b) = \langle \varphi(a), \varphi(b) \rangle_{\mathcal{H}}.$$

- Def-2 (reproducing kernel, constructive):

$$k(\cdot,b) \in \mathcal{H}, \qquad f(b) = \langle f, k(\cdot,b) \rangle_{\mathcal{H}}.$$

- Def-3 (Gram matrix): $\mathbf{G} = [k(x_i,x_j)]_{i,j=1}^n \in \mathbb{R}^{n\times n} \succeq 0$.
- Def-4 (evaluation): $\delta_x(f) = f(x)$ is continuous for all x.

Kernel: continued

- Def-1 (feature space):

$$k(a,b) = \langle \varphi(a), \varphi(b) \rangle_{\mathcal{H}}.$$

- Def-2 (reproducing kernel, constructive):

$$k(\cdot, b) \in \mathcal{H}, \qquad f(b) = \langle f, k(\cdot, b) \rangle_{\mathcal{H}}.$$

- Def-3 (Gram matrix): $\mathbf{G} = [k(x_i, x_j)]_{i,j=1}^{n} \in \mathbb{R}^{n \times n} \succeq 0$.
- Def-4 (evaluation): $\delta_x(f) = f(x)$ is continuous for all x.

- All these definitions are equivalent, $k \overset{1:1}{\leftrightarrow} \mathcal{H}_k$.

In this talk

Assumptions:

1. Domain ($\mathcal{X}$): $\mathbb{R}^d$,
2. Kernel (k): continuous, bounded, shift-invariant.

In this talk

Assumptions:

1. Domain ($\mathcal{X}$): $\mathbb{R}^d$,
2. Kernel (k): continuous, bounded, shift-invariant.

Examples:

$$k_p(\mathbf{x}, \mathbf{y}) = (\langle \mathbf{x}, \mathbf{y} \rangle + \gamma)^p$$

In this talk

Assumptions:

1. Domain ($\mathcal{X}$): $\mathbb{R}^d$,
2. Kernel (k): continuous, bounded, shift-invariant.

Examples:

$k_p(\mathbf{x}, \mathbf{y}) = (\langle \mathbf{x}, \mathbf{y} \rangle + \gamma)^p$: NO!

In this talk

Assumptions:

1. Domain ($\mathcal{X}$): $\mathbb{R}^d$,
2. Kernel (k): continuous, bounded, shift-invariant.

Examples:

$$k_p(\mathbf{x},\mathbf{y}) = (\langle \mathbf{x},\mathbf{y}\rangle + \gamma)^p\text{: NO!}, \quad k_G(\mathbf{x},\mathbf{y}) = e^{-\gamma\|\mathbf{x}-\mathbf{y}\|_2^2}\text{: YES!},$$

$$k_e(\mathbf{x},\mathbf{y}) = e^{-\gamma\|\mathbf{x}-\mathbf{y}\|_2}\text{: YES!}, \qquad k_C(\mathbf{x},\mathbf{y}) = \frac{1}{1+\gamma\|\mathbf{x}-\mathbf{y}\|_2^2}\text{: YES!}$$

In this talk

Assumptions:

1. Domain ($\mathcal{X}$): $\mathbb{R}^d$,
2. Kernel (k): continuous, bounded, shift-invariant.

Examples:

$$k_p(\mathbf{x},\mathbf{y}) = (\langle \mathbf{x},\mathbf{y}\rangle + \gamma)^p\text{: NO!},\quad k_G(\mathbf{x},\mathbf{y}) = e^{-\gamma\|\mathbf{x}-\mathbf{y}\|_2^2}\text{: YES!},$$

$$k_e(\mathbf{x},\mathbf{y}) = e^{-\gamma\|\mathbf{x}-\mathbf{y}\|_2}\text{: YES!},\quad k_C(\mathbf{x},\mathbf{y}) = \frac{1}{1+\gamma\|\mathbf{x}-\mathbf{y}\|_2^2}\text{: YES!}$$

Bochner theorem $\Rightarrow$

$$k(\mathbf{x},\mathbf{y}) = \int_{\mathbb{R}^d} \cos\left(\boldsymbol{\omega}^T(\mathbf{x}-\mathbf{y})\right) \mathrm{d}\Lambda(\boldsymbol{\omega}).$$

Cost: function values ← curve fitting

Given sample $\{(\mathbf{x}_n, y_n)\}_{n\in[N]} \subset \mathbb{R}^d \times \mathbb{R}$, kernel k on $\mathbb{R}^d$.

Cost: function values ← curve fitting

Given sample $\{(\mathbf{x}_n, y_n)\}_{n\in[N]} \subset \mathbb{R}^d \times \mathbb{R}$, kernel k on $\mathbb{R}^d$.

1. Kernel ridge regression ($\lambda > 0$):

$$\min_{f\in\mathcal{H}_k} C(f) := \frac{1}{N}\sum_{n\in[N]} [f(\mathbf{x}_n) - y_n]^2 + \lambda \|f\|_{\mathcal{H}_k}^2 .$$

Cost: function values ← curve fitting

Given sample $\{(\mathbf{x}_n, y_n)\}_{n\in[N]} \subset \mathbb{R}^d \times \mathbb{R}$, kernel k on $\mathbb{R}^d$.

1. Kernel ridge regression ($\lambda > 0$):

$$\min_{f\in\mathcal{H}_k} C(f) := \frac{1}{N}\sum_{n\in[N]} [f(\mathbf{x}_n) - y_n]^2 + \lambda \|f\|^2_{\mathcal{H}_k}.$$

2. Classification with hinge loss ($y_n \in \{\pm 1\}$):

$$\min_{f\in\mathcal{H}_k} C(f) := \frac{1}{N}\sum_{n\in[N]} \max(1 - y_n f(\mathbf{x}_n), 0) + \lambda \|f\|^2_{\mathcal{H}_k}.$$

Cost: function values ← curve fitting

Given sample $\{(\mathbf{x}_n, y_n)\}_{n\in[N]} \subset \mathbb{R}^d \times \mathbb{R}$, kernel k on $\mathbb{R}^d$.

1. Kernel ridge regression ($\lambda > 0$):

$$\min_{f\in\mathcal{H}_k} C(f) := \frac{1}{N}\sum_{n\in[N]} [f(\mathbf{x}_n) - y_n]^2 + \lambda \|f\|^2_{\mathcal{H}_k}.$$

2. Classification with hinge loss ($y_n \in \{\pm 1\}$):

$$\min_{f\in\mathcal{H}_k} C(f) := \frac{1}{N}\sum_{n\in[N]} \max(1 - y_n f(\mathbf{x}_n), 0) + \lambda \|f\|^2_{\mathcal{H}_k}.$$

Optimization over function spaces.

Solution

- Representer theorem:

$$f(\cdot) = \sum_{n\in[N]} a_n k(\cdot, \mathbf{x}_n), \quad a_n \in \mathbb{R}.$$

Solution

- Representer theorem:

$$f(\cdot) = \sum_{n\in[N]} a_n k(\cdot, \mathbf{x}_n), \quad a_n \in \mathbb{R}.$$

- Reproducing property:

$$f(\mathbf{x}_i) = \langle f, k(\cdot, \mathbf{x}_i)\rangle_{\mathcal{H}_k} = \sum_{n\in[N]} a_n k(\mathbf{x}_i, \mathbf{x}_n).$$

Solution

- Representer theorem:

$$f(\cdot) = \sum_{n \in [N]} a_n k(\cdot, \mathbf{x}_n), \quad a_n \in \mathbb{R}.$$

- Reproducing property:

$$f(\mathbf{x}_i) = \langle f, k(\cdot, \mathbf{x}_i) \rangle_{\mathcal{H}_k} = \sum_{n \in [N]} a_n k(\mathbf{x}_i, \mathbf{x}_n).$$

- $\Rightarrow$ finite-dimensional optimization problem:

$\min_{f \in \mathcal{H}_k}$ switched to $\min_{\mathbf{a} \in \mathbb{R}^N}$.

Cost: function values & derivatives ← curve & slope fitting

1. Hermite learning with gradient data:

$$\min_{f\in\mathcal{H}_k} C(f) := \frac{1}{N}\sum_{n\in[N]}\left(\left[f(\mathbf{x}_n)-y_n\right]^2+\left\|f'(\mathbf{x}_n)-\mathbf{y}'_n\right\|_2^2\right)+\lambda\left\|f\right\|_{\mathcal{H}_k}^2.$$

Cost: function values & derivatives ← curve & slope fitting

1. Hermite learning with gradient data:

$$\min_{f\in\mathcal{H}_k} C(f) := \frac{1}{N}\sum_{n\in[N]}\left([f(\mathbf{x}_n) - y_n]^2 + \left\|f'(\mathbf{x}_n) - \mathbf{y}'_n\right\|_2^2\right) + \lambda \left\|f\right\|_{\mathcal{H}_k}^2 .$$

2. Nonlinear variable selection:

$$\min_{f\in\mathcal{H}_k} C(f) := \frac{1}{N}\sum_{n\in[N]}[f(\mathbf{x}_n) - y_n]^2 + \sum_{j\in[d]}\left\|\partial_j f\right\|,$$

$$\left\|g\right\| = \sqrt{\frac{1}{N}\sum_{n\in[N]}|g(x_n)|^2}.$$

Cost: function values & derivatives ← curve & slope fitting

1. Hermite learning with gradient data:

$$\min_{f\in\mathcal{H}_k} C(f) := \frac{1}{N}\sum_{n\in[N]}\left([f(\mathbf{x}_n) - y_n]^2 + \left\|f'(\mathbf{x}_n) - \mathbf{y}'_n\right\|_2^2\right) + \lambda \|f\|_{\mathcal{H}_k}^2 .$$

2. Nonlinear variable selection:

$$\min_{f\in\mathcal{H}_k} C(f) := \frac{1}{N}\sum_{n\in[N]}[f(\mathbf{x}_n) - y_n]^2 + \sum_{j\in[d]}\|\partial_j f\|,$$

$$\|g\| = \sqrt{\frac{1}{N}\sum_{n\in[N]}|g(x_n)|^2}.$$

3. Exponential family:

$$p_{\boldsymbol{\theta}}(\mathbf{x}) \propto e^{\left\langle \boldsymbol{\theta}, \overbrace{\mathbf{T}(\mathbf{x})}^{\text{sufficient statistics}} \right\rangle}$$

1. Hermite learning with gradient data:

$$\min_{f\in\mathcal{H}_k} C(f) := \frac{1}{N}\sum_{n\in[N]}\left([f(\mathbf{x}_n)-y_n]^2 + \left\|f'(\mathbf{x}_n)-\mathbf{y}_n'\right\|_2^2\right) + \lambda\left\|f\right\|_{\mathcal{H}_k}^2.$$

2. Nonlinear variable selection:

$$\min_{f\in\mathcal{H}_k} C(f) := \frac{1}{N}\sum_{n\in[N]}[f(\mathbf{x}_n)-y_n]^2 + \sum_{j\in[d]}\left\|\partial_j f\right\|,$$

$$\|g\| = \sqrt{\frac{1}{N}\sum_{n\in[N]}|g(x_n)|^2}.$$

3. Infinite-dimensional exponential family (score matching):

$$p_{\boldsymbol{\theta}}(\mathbf{x}) \propto e^{\left\langle \boldsymbol{\theta}, \overbrace{\mathbf{T}(\mathbf{x})}^{\text{sufficient statistics}} \right\rangle} \Rightarrow p_f(\mathbf{x}) \propto e^{\langle f, k(\cdot,\mathbf{x})\rangle}$$

Cost: function values & derivatives ← curve & slope fitting

1. Hermite learning with gradient data:

$$\min_{f\in\mathcal{H}_k} C(f) := \frac{1}{N}\sum_{n\in[N]}\left([f(\mathbf{x}_n)-y_n]^2 + \left\|f'(\mathbf{x}_n)-\mathbf{y}'_n\right\|_2^2\right) + \lambda\|f\|_{\mathcal{H}_k}^2 .$$

2. Nonlinear variable selection:

$$\min_{f\in\mathcal{H}_k} C(f) := \frac{1}{N}\sum_{n\in[N]}[f(\mathbf{x}_n)-y_n]^2 + \sum_{j\in[d]}\|\partial_j f\|,$$

$$\|g\| = \sqrt{\frac{1}{N}\sum_{n\in[N]}|g(x_n)|^2}.$$

3. Infinite-dimensional exponential family (score matching):

$$p_{\boldsymbol{\theta}}(\mathbf{x}) \propto e^{\langle \boldsymbol{\theta}, \overbrace{\mathbf{T}(\mathbf{x})}^{\text{sufficient statistics}} \rangle} \Rightarrow p_f(\mathbf{x}) \propto e^{\langle f, k(\cdot,\mathbf{x})\rangle} = e^{f(\mathbf{x})} \quad (f\in\mathcal{H}_k).$$

Cost: function values & derivatives – continued

A bit more generally:

$$\min_{f \in \mathcal{H}_k} C\left(\left\{\partial^{\mathbf{p}} f(\mathbf{x}_n)\right\}_{\substack{n \in [N] \\ \mathbf{p} \in D_n}}, \|f\|_{\mathcal{H}_k}^2\right) \quad \partial^{\mathbf{p}} f(\mathbf{x}_n) := \frac{\partial^{p_1 + \ldots + p_d} f(\mathbf{x}_n)}{\partial_{x_1}^{p_1} \cdots \partial_{x_d}^{p_d}}.$$

Cost: function values & derivatives – continued

A bit more generally:

$$\min_{f \in \mathcal{H}_k} C\left(\{\partial^{\mathbf{p}} f(\mathbf{x}_n)\}_{\substack{n\in[N]\\ \mathbf{p}\in D_n}}, \|f\|^2_{\mathcal{H}_k}\right) \quad \partial^{\mathbf{p}} f(\mathbf{x}_n) := \frac{\partial^{p_1+\ldots+p_d} f(\mathbf{x}_n)}{\partial^{p_1}_{x_1} \cdots \partial^{p_d}_{x_d}}.$$

Examples: semi-supervised learning with gradient information [Zhou, 2008], nonlinear variable selection [Rosasco et al., 2010, Rosasco et al., 2013], learning of piecewise-smooth functions [Lauer et al., 2012], multi-task gradient learning [Ying et al., 2012], structure optimization in parameter-varying ARX processes [Duijkers et al., 2014], density estimation with infinite-dimensional exponential families [Sriperumbudur et al., 2017], Bayesian inference (adaptive samplers) [Strathmann et al., 2015].

Solution: representer theorem & derivative-reproducing property

- Previously:

$$f(\cdot) = \sum_{n \in [N]} a_n k(\cdot, \mathbf{x}_n), \quad a_n \in \mathbb{R}.$$

Solution: representer theorem & derivative-reproducing property

- Previously:

$$f(\cdot) = \sum_{n \in [N]} a_n k(\cdot, \mathbf{x}_n), \quad a_n \in \mathbb{R}.$$

- Now:

$$f(\cdot) = \sum_{\substack{n \in [N] \\ \mathbf{p} \in D_n}} a_{n,\mathbf{p}} \partial^{\mathbf{p},\mathbf{0}} k(\cdot, \mathbf{x}_n) \quad a_{n,\mathbf{p}} \in \mathbb{R}.$$

Solution: representer theorem & derivative-reproducing property

- Previously:

$$f(\cdot) = \sum_{n \in [N]} a_n k(\cdot, \mathbf{x}_n), \quad a_n \in \mathbb{R}.$$

- Now:

$$f(\cdot) = \sum_{\substack{n \in [N] \\ \mathbf{p} \in D_n}} a_{n,\mathbf{p}} \partial^{\mathbf{p},\mathbf{0}} k(\cdot, \mathbf{x}_n) \quad a_{n,\mathbf{p}} \in \mathbb{R}.$$

Finite-dimensional optimization problem $\left[\partial^{\mathbf{p},\mathbf{q}} k(\mathbf{x},\mathbf{y}) := \frac{\partial^{\sum_{i=1}^{d}(p_i+q_i)} k(\mathbf{x},\mathbf{y})}{\partial_{x_1}^{p_1} \cdots \partial_{x_d}^{p_d} \partial_{y_1}^{q_1} \cdots \partial_{y_d}^{q_d}}\right]$:

$$\min_{\mathbf{a}} C\left(\left\{\sum_{\substack{m \in [N] \\ \mathbf{q} \in D_m}} a_{m,\mathbf{q}} \partial^{\mathbf{p},\mathbf{q}} k(\mathbf{x}_n, \mathbf{x}_m)\right\}_{\substack{n \in [N] \\ \mathbf{p} \in D_n}}, \sum_{\substack{n,m \in [N] \\ \mathbf{p} \in D_n \\ \mathbf{q} \in D_m}} a_{n,\mathbf{p}} a_{m,\mathbf{q}} \partial^{\mathbf{p},\mathbf{q}} k(\mathbf{x}_n, \mathbf{x}_m)\right).$$

Random Fourier feature (RFF) trick

$\min_{\mathbf{a}}$: can still be computionally heavy.

RFF [Rahimi and Recht, 2007]:

$\min_{\mathbf{a}}$: can still be computionally heavy.

RFF [Rahimi and Recht, 2007]:

- Recall:

$$k(\mathbf{x}, \mathbf{y}) = \int_{\mathbb{R}^d} \cos\left(\boldsymbol{\omega}^T(\mathbf{x} - \mathbf{y})\right) \mathrm{d}\Lambda(\boldsymbol{\omega}), \quad f(\mathbf{x}) = \langle f, k(\cdot, \mathbf{x})\rangle_{\mathcal{H}_k}.$$

Random Fourier feature (RFF) trick

$\min_{\mathbf{a}}$: can still be computionally heavy.

RFF [Rahimi and Recht, 2007]:

- Recall:

$$k(\mathbf{x}, \mathbf{y}) = \int_{\mathbb{R}^d} \cos\left(\boldsymbol{\omega}^T(\mathbf{x} - \mathbf{y})\right) \mathrm{d}\Lambda(\boldsymbol{\omega}), \quad f(\mathbf{x}) = \langle f, k(\cdot, \mathbf{x})\rangle_{\mathcal{H}_k}.$$

- Explicit low-dimensional feature approximation (Λ_M):

$$k(\mathbf{x}, \mathbf{x}') \approx \left\langle \varphi(\mathbf{x}), \varphi(\mathbf{x}') \right\rangle_{\mathbb{R}^{2M}}, \qquad \hat{f}_{\mathbf{w}}(\mathbf{x}) = \langle \mathbf{w}, \varphi(\mathbf{x})\rangle_{\mathbb{R}^{2M}}.$$

Random Fourier feature (RFF) trick

$\min_{\mathbf{a}}$: can still be computionally heavy.

RFF [Rahimi and Recht, 2007]:

- Recall:

$$k(\mathbf{x},\mathbf{y}) = \int_{\mathbb{R}^d} \cos\left(\boldsymbol{\omega}^T(\mathbf{x}-\mathbf{y})\right) \mathrm{d}\Lambda(\boldsymbol{\omega}), \quad f(\mathbf{x}) = \langle f, k(\cdot,\mathbf{x})\rangle_{\mathcal{H}_k}.$$

- Explicit low-dimensional feature approximation (Λ_M):

$$k(\mathbf{x},\mathbf{x}') \approx \left\langle \varphi(\mathbf{x}), \varphi(\mathbf{x}')\right\rangle_{\mathbb{R}^{2M}}, \qquad \hat{f}_{\mathbf{w}}(\mathbf{x}) = \langle \mathbf{w}, \varphi(\mathbf{x})\rangle_{\mathbb{R}^{2M}}.$$

- Estimate $\mathbf{w}$ by leveraging fast linear primal solvers.

RFF trick: a few applications

Differential privacy preserving [Chaudhuri et al., 2011], fast function-to-function regression [Oliva et al., 2015], learning message operators in expectation propagation [Jitkrittum et al., 2015], causal discovery [Lopez-Paz et al., 2015, Strobl et al., 2019], independence testing [Zhang et al., 2017], prediction and filtering in dynamical systems [Downey et al., 2017], convolution neural networks [Cui et al., 2017], bandit optimization [Li et al., 2018], estimation of Gaussian mixture models [Keriven et al., 2018].

RFF trick: a few applications

Differential privacy preserving [Chaudhuri et al., 2011], fast function-to-function regression [Oliva et al., 2015], learning message operators in expectation propagation [Jitkrittum et al., 2015], causal discovery [Lopez-Paz et al., 2015, Strobl et al., 2019], independence testing [Zhang et al., 2017], prediction and filtering in dynamical systems [Downey et al., 2017], convolution neural networks [Cui et al., 2017], bandit optimization [Li et al., 2018], estimation of Gaussian mixture models [Keriven et al., 2018].

10-year test-of-time award (NIPS-2017).

Goodness of RFFs – related & optimal work

- Kernel values [Rahimi and Recht, 2007, Sutherland and Schneider, 2015]

$$\left\|k - \widehat{k}\right\|_{L^\infty(S_M)} = \mathcal{O}_p\left(|S_M|\sqrt{\frac{\log M}{M}}\right)$$

Goodness of RFFs – related & optimal work

- Kernel values [Rahimi and Recht, 2007, Sutherland and Schneider, 2015], [Sriperumbudur and Szabó, 2015]:

$$\left\|k - \hat{k}\right\|_{L^\infty(S_M)} = \mathcal{O}_p\left(|S_M|\sqrt{\frac{\log M}{M}}\right), \text{ afterwards}$$

$$\left\|k - \hat{k}\right\|_{L^\infty(S_M)} = \mathcal{O}_{a.s.}\left(\sqrt{\frac{\log |S_M|}{M}}\right).$$

Goodness of RFFs – related & optimal work

- Kernel ridge regression [Rudi and Rosasco, 2017, Li et al., 2019]:
  - $\mathcal{O}\left(\frac{1}{\sqrt{N}}\right)$ generalization with $M = o(N) = \mathcal{O}\left(\sqrt{N}\log N\right)$ / less RFFs.

Goodness of RFFs – related & optimal work

- Kernel PCA [Sriperumbudur and Sterge, 2018, Ullah et al., 2018], classification with 0-1 loss [Gilbert et al., 2018]: $M = o(N)$ RFFs, spectrum decay.

Goodness of RFFs – related & optimal work

- Kernel derivatives [Szabó and Sriperumbudur, 2019]: same bound as for kernel values (unbounded empirical processes, Bernstein condition).

Bernstein condition on Λ

$d = 1$:

- Gaussian kernel, $f_\Lambda(\omega) \propto e^{-\omega^2}$, analytical moments

Bernstein condition on Λ

$d = 1$:

- Gaussian kernel, $f_\Lambda(\omega) \propto e^{-\omega^2}$, analytical moments $\Rightarrow$ $n := p + q \leq 2$.

$d = 1$:

- Gaussian kernel, $f_\Lambda(\omega) \propto e^{-\omega^2}$, analytical moments $\Rightarrow$ $n := p + q \leq 2$.
- more generally: $f_\Lambda(\omega) \propto e^{-\omega^{2\ell}}$, $\Rightarrow$ $n \leq 2\ell$.

Bernstein condition on Λ

$d = 1$:

- Gaussian kernel, $f_\Lambda(\omega) \propto e^{-\omega^2}$, analytical moments $\Rightarrow$ $n := p + q \leq 2$.
- more generally: $f_\Lambda(\omega) \propto e^{-\omega^{2\ell}}$, $\Rightarrow$ $n \leq 2\ell$.

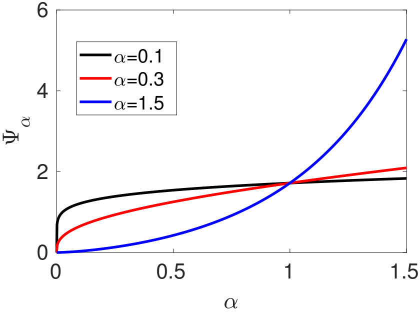
Our question

- Avoid the Bernstein condition.
- With (essentially) $f_\Lambda(\omega) \propto e^{-|\omega|^\alpha}$: guarantees for $\alpha \leq n$.

α-exponential Orlicz norm $(\alpha > 0)$

With $f_\Lambda(\omega) \propto e^{-|\omega|^\alpha}$ in mind,

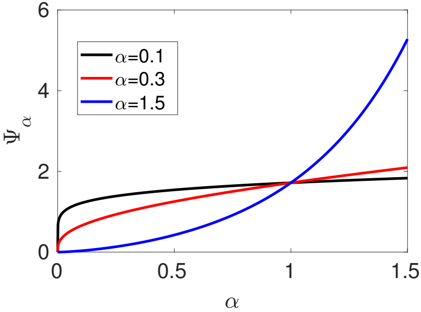
- Let $\Psi_\alpha : x \in \mathbb{R}^{\geq 0} \mapsto e^{x^\alpha} - 1 \in \mathbb{R}^{\geq 0}$.



α-exponential Orlicz norm $(\alpha > 0)$

With $f_\Lambda(\omega) \propto e^{-|\omega|^\alpha}$ in mind,

- Let $\Psi_\alpha : x \in \mathbb{R}^{\geq 0} \mapsto e^{x^\alpha} - 1 \in \mathbb{R}^{\geq 0}$.

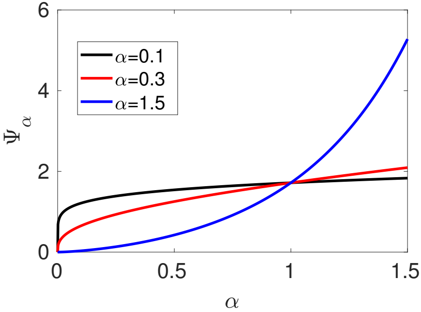


- $L_{\Psi_\alpha} := \left\{\Lambda : \|\Lambda\|_{\Psi_\alpha} := \inf\left\{c > 0 : \mathbb{E}_{\boldsymbol{\omega}\sim\Lambda}\Psi_\alpha\left(\frac{\|\boldsymbol{\omega}\|_2}{c}\right) \leq 1\right\} < +\infty\right\}$.

α-exponential Orlicz norm $(\alpha > 0)$

With $f_\Lambda(\omega) \propto e^{-|\omega|^\alpha}$ in mind,

- Let $\Psi_\alpha : x \in \mathbb{R}^{\geq 0} \mapsto e^{x^\alpha} - 1 \in \mathbb{R}^{\geq 0}$.



- $L_{\Psi_\alpha} := \left\{\Lambda : \|\Lambda\|_{\Psi_\alpha} := \inf\left\{c > 0 : \mathbb{E}_{\boldsymbol{\omega}\sim\Lambda}\Psi_\alpha\left(\frac{\|\boldsymbol{\omega}\|_2}{c}\right) \leq 1\right\} < +\infty\right\}$.
- $\Lambda \in L_{\Psi_2}$: sub-Gaussian, $\Lambda \in L_{\Psi_1}$: sub-exponential.

- Intuition:
  - $f_\Lambda(\omega) \propto e^{-|\omega|^\alpha} \Rightarrow \Lambda \in L_{\Psi_\alpha}$ (polynomial decorations: ✓).

α-exponential Orlicz norm: factory

- Intuition:
  - $f_\Lambda(\omega) \propto e^{-|\omega|^\alpha} \Rightarrow \Lambda \in L_{\Psi_\alpha}$ (polynomial decorations: ✓).
- Tensor product kernels:

If

- $k_i \leftrightarrow \Lambda_i \in L_{\Psi_{\alpha_i}}$

α-exponential Orlicz norm: factory

- Intuition:
  - $f_\Lambda(\omega) \propto e^{-|\omega|^\alpha} \Rightarrow \Lambda \in L_{\Psi_\alpha}$ (polynomial decorations: ✓).
- Tensor product kernels:

If

- $k_i \leftrightarrow \Lambda_i \in L_{\Psi_{\alpha_i}}$ and
- $k(\mathbf{x}, \mathbf{y}) = \prod_{i \in [d]} k_i(x_i, y_i)$, i.e. $\Lambda = \otimes_{i \in [d]} \Lambda_i$,

α-exponential Orlicz norm: factory

- Intuition:
  - $f_\Lambda(\omega) \propto e^{-|\omega|^\alpha} \Rightarrow \Lambda \in L_{\Psi_\alpha}$ (polynomial decorations: ✓).
- Tensor product kernels:

If

- $k_i \leftrightarrow \Lambda_i \in L_{\Psi_{\alpha_i}}$ and
- $k(\mathbf{x}, \mathbf{y}) = \prod_{i \in [d]} k_i(x_i, y_i)$, i.e. $\Lambda = \otimes_{i \in [d]} \Lambda_i$,

then $\Lambda \in L_{\Psi_\alpha}$ with $\alpha = \min_{i \in [d]} \alpha_i$.

Kernel examples with α-exp. Orlicz spectrum: $d=1$

Spectrum	$f_\Lambda(\omega)$	α
Gaussian	$\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{\omega^2}{2\sigma^2}}$	2
Laplace	$\frac{\sigma}{2}e^{-\sigma\|\omega\|}$	1
generalized Gaussian	$\frac{\alpha}{2\beta\Gamma\left(\frac{1}{\alpha}\right)}e^{-\frac{\|\omega\|}{\beta}^{\alpha}}$	α
variance Gamma	$\frac{\sigma^{2b}\|\omega\|^{b-\frac{1}{2}}K_{b-\frac{1}{2}}(\sigma\|\omega\|)}{\sqrt{\pi}\Gamma(b)(2\sigma)^{b-\frac{1}{2}}}$	1
Weibull (S)	$\frac{s}{2\lambda}\left(\frac{\|\omega\|}{\lambda}\right)^{s-1}e^{-\left(\frac{\|\omega\|}{\lambda}\right)^s}$	s
exponentiated exponential (S)	$\frac{\alpha}{2\lambda}\left(1-e^{-\frac{\|\omega\|}{\lambda}}\right)^{\alpha-1}e^{-\frac{\|\omega\|}{\lambda}}$	1

$I_a(z)=\sum_{n\in\mathbb{N}}\frac{1}{n!\Gamma(n+a+1)}\left(\frac{z}{2}\right)^{2n+a}$, $K_a(z)=\frac{\pi}{2}\frac{I_{-a}(z)-I_a(z)}{\sin(a\pi)}$ for $z\in\mathbb{R}$ and non-integer a; when a is an integer the limit is taken.

Kernel examples with α-exponential Orlicz spectrum - 2

Spectrum	$f_\Lambda(\omega)$	α
exponentiated Weibull (S)	$\frac{\alpha s}{2\lambda}\left(\frac{\lvert\omega\rvert}{\lambda}\right)^{s-1}\left[1-e^{-\left(\frac{\lvert\omega\rvert}{\lambda}\right)^s}\right]^{\alpha-1}\times$ $\times e^{-\left(\frac{\lvert\omega\rvert}{\lambda}\right)^s}$	s
Nakagami (S)	$\frac{m^m}{\Gamma(m)\Omega^m}\lvert\omega\rvert^{2m-1}e^{-\frac{m\omega^2}{\Omega}}$	2
chi-squared (S)	$\frac{1}{2^{\frac{s}{2}+1}\Gamma\left(\frac{s}{2}\right)}\lvert\omega\rvert^{\frac{s}{2}-1}e^{-\frac{\lvert\omega\rvert}{2}}$	1
Erlang (S)	$\frac{\lambda^s\lvert\omega\rvert^{s-1}e^{-\lambda\lvert\omega\rvert}}{2(s-1)!}$	1
Gamma (S)	$\frac{1}{2\Gamma(s)\theta^s}\lvert\omega\rvert^{s-1}e^{-\frac{\lvert\omega\rvert}{\theta}}$	1
generalized Gamma (S)	$\frac{p/a^D}{2\Gamma\left(\frac{D}{p}\right)}\lvert\omega\rvert^{D-1}e^{-\left(\frac{\lvert\omega\rvert}{a}\right)^p}$	p

Kernel examples with α-exponential Orlicz spectrum - 3

Spectrum	$f_\Lambda(\omega)$	α
Rayleigh (S)	$\frac{\lvert\omega\rvert}{2\sigma^2}e^{-\frac{\omega^2}{2\sigma^2}}$	2
Maxwell-Boltzmann (S)	$\frac{1}{\sqrt{2\pi}}\frac{\omega^2 e^{-\frac{\omega^2}{2a^2}}}{a^3}$	2
chi (S)	$\frac{1}{2^{\frac{s}{2}}\Gamma\left(\frac{s}{2}\right)}\lvert\omega\rvert^{s-1}e^{-\frac{\omega^2}{2}}$	2
exponential-logarithmic (S)	$-\frac{1}{2\log(p)}\frac{\beta(1-p)e^{-\beta\lvert\omega\rvert}}{1-(1-p)e^{-\beta\lvert\omega\rvert}}$	1
Weibull-logarithmic (S)	$-\frac{1}{2\log(p)}\frac{\alpha\beta(1-p)\lvert\omega\rvert^{\alpha-1}e^{-\beta\lvert\omega\rvert^\alpha}}{1-(1-p)e^{-\beta\lvert\omega\rvert^\alpha}}$	α
Gamma/Gompertz (S)	$\frac{bse^{b\lvert\omega\rvert}\beta^s}{2\left(\beta-1+e^{b\lvert\omega\rvert}\right)^{s+1}}$	bs

Kernel examples with α-exponential Orlicz spectrum - 4

Spectrum	$f_\Lambda(\omega)$	α
hyperbolic secant	$\frac{1}{2}\operatorname{sech}\left(\frac{\pi}{2}\omega\right)$	1
logistic	$\frac{e^{-\frac{\omega}{s}}}{s\left[1+e^{-\frac{\omega}{s}}\right]^2} = \frac{1}{4s}\operatorname{sech}^2\left(\frac{\omega}{2s}\right)$	1
normal-inverse Gaussian	$\frac{\alpha\delta K_1\left(\alpha\sqrt{\delta^2+\omega^2}\right)}{\pi\sqrt{\delta^2+\omega^2}}e^{\delta\alpha}$	1
hyperbolic	$\frac{1}{2\delta K_1(\delta\alpha)}e^{-\alpha\sqrt{\delta^2+\omega^2}}$	1
generalized hyperbolic	$\frac{(\alpha/\delta)^\lambda}{\sqrt{2\pi}K_\lambda(\delta\gamma)}\frac{K_{\lambda-\frac{1}{2}}\left(\alpha\sqrt{\delta^2+\omega^2}\right)}{\left(\frac{\sqrt{\delta^2+\omega^2}}{\alpha}\right)^{\frac{1}{2}-\lambda}}$	1

$$\operatorname{sech}(x) = \frac{1}{\cosh(x)} = \frac{2}{e^x+e^{-x}}.$$

Spectrum $\mapsto$ kernel examples $(b > \frac{1}{2},\ s > 0)$

Kernel name	$k(x,y)$	Spectrum
Gaussian	$e^{-\frac{\sigma^2(x-y)^2}{2}}$	Gaussian
Cauchy / inverse quadric	$\frac{\sigma^2}{\sigma^2+(x-y)^2}$	Laplace
inverse multiquadric	$\left[\frac{\sigma^2}{\sigma^2+(x-y)^2}\right]^b$	variance Gamma
–	$\operatorname{sech}(x-y)$	hyperbolic secant
–	$\frac{\pi s(x-y)}{\sinh(\pi s(x-y))}$	logistic

Spectrum $\mapsto$ kernel examples $(b > \frac{1}{2},\, s > 0)$

Kernel name	$k(x,y)$	Spectrum
Gaussian	$e^{-\frac{\sigma^2(x-y)^2}{2}}$	Gaussian
Cauchy / inverse quadric	$\frac{\sigma^2}{\sigma^2+(x-y)^2}$	Laplace
inverse multiquadric	$\left[\frac{\sigma^2}{\sigma^2+(x-y)^2}\right]^b$	variance Gamma
–	$\operatorname{sech}(x-y)$	hyperbolic secant
–	$\frac{\pi s(x-y)}{\sinh(\pi s(x-y))}$	logistic

+Analytical kernel values: generalized Gaussian, Weibull (S), chi-squared (S), Erlang (S), Gamma (S), Rayleigh (S), chi (S), Weibull-logarithmic (S), Gamma/Gompertz (S), normal-inverse Gaussian, hyperbolic, generalized hyperbolic.

Our result (finite-sample guarantee $\Rightarrow$ asymptotics)

Assume:

- k: continuous, bounded, shift-invariant kernel on $\mathbb{R}^d$.
- $\Lambda \in L_{\Psi_\alpha}$ $(\alpha > 0)$.
- Let $\mathbf{p}, \mathbf{q} \in \mathbb{N}^d$, $[\mathbf{p}; \mathbf{q}] \neq \mathbf{0}$, $n := \sum_{i \in [d]} (p_i + q_i)$, $\alpha \leq n$.

Our result (finite-sample guarantee $\Rightarrow$ asymptotics)

Assume:

- k: continuous, bounded, shift-invariant kernel on $\mathbb{R}^d$.
- $\Lambda \in L_{\Psi_\alpha}$ $(\alpha > 0)$.
- Let $\mathbf{p}, \mathbf{q} \in \mathbb{N}^d$, $[\mathbf{p}; \mathbf{q}] \neq \mathbf{0}$, $n := \sum_{i \in [d]} (p_i + q_i)$, $\alpha \leq n$.

Then

$$\left\| \partial^{\mathbf{p},\mathbf{q}} k - \widehat{\partial^{\mathbf{p},\mathbf{q}} k} \right\|_{L^\infty(S_M)} = \mathcal{O}_{a.s.}\left(|S_M| \frac{\log^r(M)}{\sqrt{M}} \right), \quad r = \frac{n}{\alpha}.$$

Summary

- Focus: RFF-based acceleration & high-order derivatives.
- Result:
  - spectrum: α-exponential Orlicz assumption.
  - $n \geq \alpha$-order derivative: ✓
- Preprint: in the oven.

Summary

- Focus: RFF-based acceleration & high-order derivatives.
- Result:
  - spectrum: α-exponential Orlicz assumption.
  - $n \geq \alpha$-order derivative: ✓
- Preprint: in the oven.

Stress Test
RISK Management and Financial Steering

Acks: This research benefited from the support of the Chair Stress Test, RISK Management and Financial Steering, led by the French Ecole polytechnique and its Foundation and sponsored by BNP Paribas.

Proof idea

Decomposition into 3 terms:

1. Unbounded part: Talagrand & Hoffman-Jorgensen inequalities.
2. Bounded part: Klein-Rio inequality & Dudley entropy integral bound.
3. Truncation: bound on the incomplete Gamma function.

Chaudhuri, K., Monteleoni, C., and Sarwate, A. D. (2011).
Differentially private empirical risk minimization.
Journal of Machine Learning Research, 12:1069–1109.

Cui, Y., Zhou, F., Wang, J., Liu, X., Lin, Y., and Belongie, S. (2017).
Kernel pooling for convolutional neural networks.
In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2921–2930.

Downey, C., Hefny, A., Boots, B., Li, B., and Gordon, G. (2017).
Predictive state recurrent neural networks.
In *Advances in Neural Information Processing Systems (NIPS)*, pages 6053–6064.

Duijkers, R., Tóth, R., Piga, D., and Laurain, V. (2014).
Shrinking complexity of scheduling dependencies in LS-SVM based LPV system identification.

In *IEEE Conference on Decision and Control*, pages 2561–2566.

Gilbert, A., Tewari, A., and Sung, Y. (2018).
But how does it work in theory? Linear SVM with random features.
In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 3379–3388.

Jitkrittum, W., Gretton, A., Heess, N., Eslami, A., Lakshminarayanan, B., Sejdinovic, D., and Szabó, Z. (2015).
Kernel-based just-in-time learning for passing expectation propagation messages.
In *Conference on Uncertainty in Artificial Intelligence (UAI)*, pages 405–414.

Keriven, N., Bourrier, A., Gribonval, R., and Pérez, P. (2018).
Sketching for large-scale learning of mixture models.
Information and Inference: A Journal of the IMA, 7:447–508.

Lauer, F., Le, V. L., and Bloch, G. (2012).

Learning smooth models of nonsmooth functions via convex optimization.
In *International Workshop on Machine Learning for Signal Processing (IEEE-MLSP)*.

Li, L., Jamieson, K., DeSalvo, G., Rostamizadeh, A., and Talwalkar, A. (2018).
Hyperband: A novel bandit-based approach to hyperparameter optimization.
Journal of Machine Learning Research, 18(185):1–52.

Li, Z., Ton, J.-F., Oglic, D., and Sejdinovic, D. (2019).
A unified analysis of random Fourier features.
In *International Conference on Machine Learning (ICML; PMLR)*.

Lopez-Paz, D., Muandet, K., Schölkopf, B., and Tolstikhin, I. (2015).
Towards a learning theory of cause-effect inference.

International Conference on Machine Learning (ICML), 37:1452–1461.

Oliva, J., Neiswanger, W., Póczos, B., Xing, E., and Schneider, J. (2015).
Fast function to function regression.
In *International Conference on Artificial Intelligence and Statistics (AISTATS; PMLR)*, pages 717–725.

Rahimi, A. and Recht, B. (2007).
Random features for large-scale kernel machines.
In *Advances in Neural Information Processing Systems (NIPS)*, pages 1177–1184.

Rosasco, L., Santoro, M., Mosci, S., Verri, A., and Villa, S. (2010).
A regularization approach to nonlinear variable selection.
In *International Conference on Artificial Intelligence and Statistics (AISTATS; PMLR)*, volume 9, pages 653–660.

Rosasco, L., Villa, S., Mosci, S., Santoro, M., and Verri, A. (2013).
Nonparametric sparsity and regularization.
Journal of Machine Learning Research, 14:1665–1714.

Rudi, A. and Rosasco, L. (2017).
Generalization properties of learning with random features.
In *Advances in Neural Information Processing Systems (NIPS)*, pages 3218–3228.

Sriperumbudur, B. and Sterge, N. (2018).
Approximate kernel PCA using random features: Computational vs. statistical trade-off.
Technical report, Pennsylvania State University. (https://arxiv.org/abs/1706.06296).

Sriperumbudur, B. K., Fukumizu, K., Gretton, A., Hyvärinen, A., and Kumar, R. (2017).
Density estimation in infinite dimensional exponential families.
Journal of Machine Learning Research, 18(57):1–59.

Sriperumbudur, B. K. and Szabó, Z. (2015).
Optimal rates for random Fourier features.
In *Advances in Neural Information Processing Systems (NIPS)*, pages 1144–1152.

Strathmann, H., Sejdinovic, D., Livingstone, S., Szabó, Z., and Gretton, A. (2015).
Gradient-free Hamiltonian Monte Carlo with efficient kernel exponential families.
In *Advances in Neural Information Processing Systems (NIPS)*, pages 955–963.

Strobl, E. V., Zhang, K., and Visweswaran, S. (2019).
Approximate kernel-based conditional independence tests for fast non-parametric causal discovery.
Journal of Causal Inference, 7(1).
(https://doi.org/10.1515/jci-2018-0017).

Sutherland, D. J. and Schneider, J. (2015).
On the error of random Fourier features.

In *Conference on Uncertainty in Artificial Intelligence (UAI)*, pages 862–871.

Szabó, Z. and Sriperumbudur, B. K. (2019).
On kernel derivative approximation with random Fourier features.
In *International Conference on Artificial Intelligence and Statistics (AISTATS; PMLR)*, volume 89, pages 827–836.

Ullah, E., Mianjy, P., Marinov, T. V., and Arora, R. (2018).
Streaming kernel PCA with $\tilde{O}(\sqrt{n})$ random features.
In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 7311–7321.

Ying, Y., Wu, Q., and Campbell, C. (2012).
Learning the coordinate gradients.
Advances in Computational Mathematics, 37:355–378.

Zhang, Q., Filippi, S., Gretton, A., and Sejdinovic, D. (2017).
Large-scale kernel methods for independence testing.
Statistics and Computing, pages 1–18.

Zhou, D.-X. (2008).
Derivative reproducing properties for kernel methods in learning theory.
Journal of Computational and Applied Mathematics, 220:456–463.