Optimal rates for random Fourier feature kernel approximations

Zoltán Szabó*
Gatsby Unit, UCL

Joint work with Bharath K. Sriperumbudur*
Department of Statistics, PSU
(*equal contribution)

AMPLab, UC Berkeley
November 20, 2015

Outline

- Kernels and kernel derivatives.
- Random Fourier features (RFFs).
- Guarantees on RFF approximation: uniform, L^r.

Kernel, RKHS

- $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ kernel on $\mathcal{X}$, if
  - $\exists \varphi : \mathcal{X} \to H$(ilbert space) feature map,
  - $k(a, b) = \langle \varphi(a), \varphi(b) \rangle_H$ $(\forall a, b \in \mathcal{X})$.

Kernel, RKHS

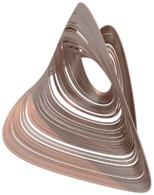
- $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ kernel on $\mathcal{X}$, if
  - $\exists \varphi : \mathcal{X} \to H$(ilbert space) feature map,
  - $k(a, b) = \langle \varphi(a), \varphi(b) \rangle_H$ $(\forall a, b \in \mathcal{X})$.
- Kernel examples: $\mathcal{X} = \mathbb{R}^d$ $(p > 0,\ \theta > 0)$
  - $k(a, b) = (\langle a, b \rangle + \theta)^p$: polynomial,
  - $k(a, b) = e^{-\|a-b\|_2^2/(2\theta^2)}$: Gaussian,
  - $k(a, b) = e^{-\theta\|a-b\|_2}$: Laplacian.

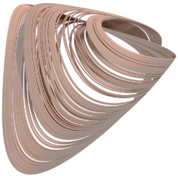
Kernel, RKHS

- $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ kernel on $\mathcal{X}$, if
  - $\exists \varphi : \mathcal{X} \to H$(ilbert space) feature map,
  - $k(a, b) = \langle \varphi(a), \varphi(b) \rangle_H$ $(\forall a, b \in \mathcal{X})$.
- Kernel examples: $\mathcal{X} = \mathbb{R}^d$ $(p > 0,\ \theta > 0)$
  - $k(a, b) = (\langle a, b \rangle + \theta)^p$: polynomial,
  - $k(a, b) = e^{-\|a-b\|_2^2/(2\theta^2)}$: Gaussian,
  - $k(a, b) = e^{-\theta\|a-b\|_2}$: Laplacian.
- In the $H = H(k)$ RKHS $(\exists!)$: $\varphi(u) = k(\cdot, u)$.

Kernel: example domains ($\mathcal{X}$)

- Euclidean space: $\mathcal{X} = \mathbb{R}^d$.
- Graphs, texts, time series, dynamical systems, distributions.





Kernel: application example – ridge regression

- Given: $\{(x_i, y_i)\}_{i=1}^{\ell}$, $H = H(k)$.
- Task: find $f \in H$ s.t. $f(x_i) \approx y_i$,

$$J(f) = \frac{1}{\ell}\sum_{i=1}^{\ell}[f(x_i) - y_i]^2 + \lambda \|f\|_H^2 \to \min_{f \in H} \quad (\lambda > 0).$$

Kernel: application example – ridge regression

- Given: $\{(x_i, y_i)\}_{i=1}^{\ell}$, $H = H(k)$.
- Task: find $f \in H$ s.t. $f(x_i) \approx y_i$,

$$J(f) = \frac{1}{\ell}\sum_{i=1}^{\ell}[f(x_i) - y_i]^2 + \lambda \|f\|_H^2 \to \min_{f \in H} \quad (\lambda > 0).$$

- Analytical solution, $\mathcal{O}(\ell^3)$ – expensive:

$$f(x) = [k(x_1, x), \dots, k(x_\ell, x)](\mathbf{G} + \lambda \ell I)^{-1}[y_1; \dots; y_\ell],$$
$$\mathbf{G} = [k(x_i, x_j)]_{i,j=1}^{\ell}.$$

Kernel: application example – ridge regression

- Given: $\{(x_i, y_i)\}_{i=1}^{\ell}$, $H = H(k)$.
- Task: find $f \in H$ s.t. $f(x_i) \approx y_i$,

$$J(f) = \frac{1}{\ell}\sum_{i=1}^{\ell}[f(x_i) - y_i]^2 + \lambda \|f\|_H^2 \to \min_{f \in H} \quad (\lambda > 0).$$

- Analytical solution, $\mathcal{O}(\ell^3)$ – expensive:

$$f(x) = [k(x_1, x), \ldots, k(x_\ell, x)](\mathbf{G} + \lambda \ell I)^{-1}[y_1; \ldots; y_\ell],$$
$$\mathbf{G} = [k(x_i, x_j)]_{i,j=1}^{\ell}.$$

- **Idea**: $\hat{\mathbf{G}}$, matrix-inversion lemma, fast primal solvers $\to$ RFF.

Kernels: more generally

- Requirement: inner product on the inputs ($k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$).
- Loss function ($\lambda > 0$):

$$J(f) = \sum_{i=1}^{\ell} V\left(y_i, f(x_i)\right) + \lambda \left\|f\right\|_{H(k)}^2 \to \min_{f \in H(k)}.$$

Kernels: more generally

- Requirement: inner product on the inputs ($k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$).
- Loss function ($\lambda > 0$):

$$J(f) = \sum_{i=1}^{\ell} V\left(y_i, f(x_i)\right) + \lambda \left\| f \right\|_{H(k)}^2 \to \min_{f \in H(k)}.$$

- By the representer theorem $[f(\cdot) = \sum_{i=1}^{\ell} \alpha_i k(\cdot, x_i)]$:

$$J(\boldsymbol{\alpha}) = \sum_{i=1}^{\ell} V\left(y_i, (\mathbf{G}\boldsymbol{\alpha})_i\right) + \lambda \boldsymbol{\alpha}^T \mathbf{G}\boldsymbol{\alpha} \to \min_{\boldsymbol{\alpha} \in \mathbb{R}^{\ell}}.$$

- $\Rightarrow$ $k(x_i, x_j)$ matters.

Kernel derivatives: application example

Motivation:

- fitting ∞-D exp. family distributions [Sriperumbudur et al., 2014],
- $k \leftrightarrow$ sufficient statistics,
- rich family,

Kernel derivatives: application example

Motivation:

- fitting ∞-D exp. family distributions [Sriperumbudur et al., 2014],
- $k \leftrightarrow$ sufficient statistics,
- rich family,
- fitting $=$ linear equation:
  - coefficient matrix: $(d\ell) \times (d\ell)$, $d = dim(x)$,
  - entries: kernel values and derivatives.

Kernel derivatives: more generally

- Objective:

$$J(f) = \frac{1}{\ell}\sum_{i=1}^{\ell} V\left(y_i, \{\partial^{\mathbf{p}} f(x_i)\}_{\mathbf{p}\in J_i}\right) + \lambda \left\|f\right\|_{H(k)}^2 \to \min_{f\in H(k)}.$$

Kernel derivatives: more generally

- Objective:

$$J(f) = \frac{1}{\ell}\sum_{i=1}^{\ell} V\left(y_i, \{\partial^{\mathbf{p}} f(x_i)\}_{\mathbf{p}\in J_i}\right) + \lambda \left\|f\right\|_{H(k)}^2 \to \min_{f\in H(k)}.$$

- [Zhou, 2008, Shi et al., 2010, Rosasco et al., 2010, Rosasco et al., 2013, Ying et al., 2012]:
    - semi-supervised learning with gradient information,
    - nonlinear variable selection.
- Kernel HMC [Strathmann et al., 2015].

Focus

- $\mathcal{X} = \mathbb{R}^d$. k: continuous, shift-invariant $[k(\mathbf{x},\mathbf{y}) = \tilde{k}(\mathbf{x}-\mathbf{y})]$.
- By Bochner's theorem:

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

Focus

- $\mathcal{X} = \mathbb{R}^d$. k: continuous, shift-invariant $[k(\mathbf{x},\mathbf{y}) = \tilde{k}(\mathbf{x}-\mathbf{y})]$.
- By Bochner's theorem:

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

- RFF trick [Rahimi and Recht, 2007] (MC): $\boldsymbol{\omega}_{1:m} := (\boldsymbol{\omega}_j)_{j=1}^m \stackrel{i.i.d.}{\sim} \Lambda$,

$$\hat{k}(\mathbf{x},\mathbf{y}) = \frac{1}{m}\sum_{j=1}^m \cos\left(\boldsymbol{\omega}_j^T(\mathbf{x}-\mathbf{y})\right) = \int_{\mathbb{R}^d} \cos\left(\boldsymbol{\omega}^T(\mathbf{x}-\mathbf{y})\right) \mathrm{d}\Lambda_m(\boldsymbol{\omega}).$$

RFF – existing guarantee, basically

- Hoeffding inequality + union bound:

$$\left\|k - \hat{k}\right\|_{L^\infty(\mathcal{S})} = \mathcal{O}_p\left(\underbrace{|\mathcal{S}|}_{\text{linear}}\sqrt{\frac{\log m}{m}}\right).$$

RFF – existing guarantee, basically

- Hoeffding inequality + union bound:

$$\left\| k - \hat{k} \right\|_{L^\infty(\mathcal{S})} = \mathcal{O}_p\left(\underbrace{|\mathcal{S}|}_{\text{linear}} \sqrt{\frac{\log m}{m}} \right).$$

- Characteristic function point of view [Csörgő and Totik, 1983] (asymptotic!):
  1. $|\mathcal{S}_m| = e^{o(m)}$ is the optimal rate for a.s. convergence,
  2. For faster growing $|\mathcal{S}_m|$: even convergence in probability fails.

Today: one-page summary

1. Finite-sample L^∞-guarantee $\xrightarrow{\text{specifically}}$

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

$\Rightarrow \mathcal{S}$ can grow exponentially $[|\mathcal{S}_m| = e^{o(m)}]$ – optimal!

Today: one-page summary

1. Finite-sample L^∞-guarantee $\xrightarrow{\text{specifically}}$

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

   $\Rightarrow$ $\mathcal{S}$ can grow exponentially $[|\mathcal{S}_m| = e^{o(m)}]$ – optimal!

2. Finite sample L^r guarantees, $r \in [1, \infty)$.

Today: one-page summary

1. Finite-sample L^∞-guarantee $\xrightarrow{\text{specifically}}$

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

   $\Rightarrow$ $\mathcal{S}$ can grow exponentially $[|\mathcal{S}_m|=e^{o(m)}]$ – optimal!
2. Finite sample L^r guarantees, $r\in[1,\infty)$.
3. Derivatives: $\partial^{\mathbf{p},\mathbf{q}}k$.

..., where

- Uniform ($r = \infty$), L^r ($1 \le r < \infty$) norm:

$$\|k - \hat{k}\|_{L^\infty(\mathcal{S})} := \sup_{\mathbf{x},\mathbf{y}\in\mathcal{S}} \left| k(\mathbf{x},\mathbf{y}) - \hat{k}(\mathbf{x},\mathbf{y}) \right|,$$

$$\|k - \hat{k}\|_{L^r(\mathcal{S})} := \left(\int_{\mathcal{S}} \int_{\mathcal{S}} |k(\mathbf{x},\mathbf{y}) - \hat{k}(\mathbf{x},\mathbf{y})|^r \,\mathrm{d}\mathbf{x}\,\mathrm{d}\mathbf{y} \right)^{\frac{1}{r}}.$$

. . . , where

- Uniform ($r = \infty$), L^r ($1 \le r < \infty$) norm:

$$\|k - \hat{k}\|_{L^\infty(\mathcal{S})} := \sup_{\mathbf{x},\mathbf{y}\in\mathcal{S}} \left| k(\mathbf{x},\mathbf{y}) - \hat{k}(\mathbf{x},\mathbf{y}) \right|,$$

$$\|k - \hat{k}\|_{L^r(\mathcal{S})} := \left(\int_{\mathcal{S}} \int_{\mathcal{S}} |k(\mathbf{x},\mathbf{y}) - \hat{k}(\mathbf{x},\mathbf{y})|^r \,\mathrm{d}\mathbf{x}\,\mathrm{d}\mathbf{y} \right)^{\frac{1}{r}}.$$

- Kernel derivatives:

$$\partial^{\mathbf{p},\mathbf{q}} k(\mathbf{x},\mathbf{y}) = \frac{\partial^{|\mathbf{p}|+|\mathbf{q}|} k(\mathbf{x},\mathbf{y})}{\partial x_1^{p_1} \cdots \partial x_d^{p_d} \partial y_1^{q_1} \cdots \partial y_d^{q_d}}, \qquad |\mathbf{p}| = \sum_{j=1}^{d} |p_j|.$$

$\|k - \hat{k}\|_{L^\infty(\mathcal{S})}$: proof idea

1. Empirical process form [$\mathbb{P}g := \int g \mathrm{d}\mathbb{P}$]:

$$\sup_{\mathbf{x},\mathbf{y}\in\mathcal{S}} \left| k(\mathbf{x},\mathbf{y}) - \hat{k}(\mathbf{x},\mathbf{y}) \right| = \sup_{g\in\mathcal{G}} |\Lambda g - \Lambda_m g| = \|\Lambda - \Lambda_m\|_{\mathcal{G}}.$$

$\|k - \hat{k}\|_{L^\infty(\mathcal{S})}$: proof idea

1. Empirical process form [$\mathbb{P}g := \int g \mathrm{d}\mathbb{P}$]:

$$\sup_{\mathbf{x},\mathbf{y}\in\mathcal{S}} \left| k(\mathbf{x},\mathbf{y}) - \hat{k}(\mathbf{x},\mathbf{y}) \right| = \sup_{g\in\mathcal{G}} |\Lambda g - \Lambda_m g| = \|\Lambda - \Lambda_m\|_{\mathcal{G}}.$$

2. $f(\boldsymbol{\omega}_{1:m}) = \|\Lambda - \Lambda_m\|_{\mathcal{G}}$ concentrates (bounded difference):

$$\|\Lambda - \Lambda_m\|_{\mathcal{G}} \lesssim \mathbb{E}_{\boldsymbol{\omega}_{1:m}} \|\Lambda - \Lambda_m\|_{\mathcal{G}} + \frac{1}{\sqrt{m}}.$$

$\|k - \hat{k}\|_{L^\infty(\mathcal{S})}$: proof idea

1. Empirical process form [$\mathbb{P}g := \int g \mathrm{d}\mathbb{P}$]:

$$\sup_{\mathbf{x},\mathbf{y}\in\mathcal{S}} \left| k(\mathbf{x},\mathbf{y}) - \hat{k}(\mathbf{x},\mathbf{y}) \right| = \sup_{g\in\mathcal{G}} |\Lambda g - \Lambda_m g| = \|\Lambda - \Lambda_m\|_{\mathcal{G}}.$$

2. $f(\boldsymbol{\omega}_{1:m}) = \|\Lambda - \Lambda_m\|_{\mathcal{G}}$ concentrates (bounded difference):

$$\|\Lambda - \Lambda_m\|_{\mathcal{G}} \lesssim \mathbb{E}_{\boldsymbol{\omega}_{1:m}} \|\Lambda - \Lambda_m\|_{\mathcal{G}} + \frac{1}{\sqrt{m}}.$$

3. $\mathcal{G}$ is 'nice' (uniformly bounded, separable Carathéodory) $\Rightarrow$

$$\mathbb{E}_{\boldsymbol{\omega}_{1:m}} \|\Lambda - \Lambda_m\|_{\mathcal{G}} \lesssim \mathbb{E}_{\boldsymbol{\omega}_{1:m}} \underbrace{\mathcal{R}\left(\mathcal{G}, \boldsymbol{\omega}_{1:m}\right)}_{\mathbb{E}_{\epsilon} \sup_{g\in\mathcal{G}} \left| \frac{1}{m} \sum_{j=1}^{m} \epsilon_j g(\omega_j) \right|}.$$

Proof idea

4. Using Dudley's entropy bound:

$$\mathcal{R}\left(\mathcal{G}, \boldsymbol{\omega}_{1:m}\right) \precsim \frac{1}{\sqrt{m}} \int_0^{|\mathcal{G}|_{L^2(\Lambda_m)}} \sqrt{\log \mathcal{N}(\mathcal{G}, L^2(\Lambda_m), u)} \mathrm{d}u.$$

Proof idea

4. Using Dudley's entropy bound:

$$\mathcal{R}\left(\mathcal{G},\boldsymbol{\omega}_{1:m}\right) \lesssim \frac{1}{\sqrt{m}} \int_0^{|\mathcal{G}|_{L^2(\Lambda_m)}} \sqrt{\log \mathcal{N}(\mathcal{G}, L^2(\Lambda_m), u)}\mathrm{d}u.$$

5. $\mathcal{G}$ is smoothly parameterized by a compact set $\Rightarrow$

$$\mathcal{N}(\mathcal{G}, L^2(\Lambda_m), u) \le \left(\frac{4|\mathcal{S}|A}{u}+1\right)^d, \ A(\boldsymbol{\omega}_{1:m}) = \sqrt{\frac{1}{m}\sum_{j=1}^m \|\boldsymbol{\omega}_j\|_2^2}.$$

Proof idea

4. Using Dudley's entropy bound:

$$\mathcal{R}\left(\mathcal{G}, \boldsymbol{\omega}_{1:m}\right) \lesssim \frac{1}{\sqrt{m}} \int_0^{|\mathcal{G}|_{L^2(\Lambda_m)}} \sqrt{\log \mathcal{N}(\mathcal{G}, L^2(\Lambda_m), u)}\mathrm{d}u.$$

5. $\mathcal{G}$ is smoothly parameterized by a compact set $\Rightarrow$

$$\mathcal{N}(\mathcal{G}, L^2(\Lambda_m), u) \leq \left(\frac{4|\mathcal{S}|A}{u} + 1\right)^d, \; A(\boldsymbol{\omega}_{1:m}) = \sqrt{\frac{1}{m}\sum_{j=1}^m \|\boldsymbol{\omega}_j\|_2^2}.$$

6. Putting together $[|\mathcal{G}|_{L^2(\Lambda_m)} \leq 2$, Jensen inequality] we get ...

L^∞ result for k

Let k be continuous, $\sigma^2 := \int \|\boldsymbol{\omega}\|^2 \,\mathrm{d}\Lambda(\boldsymbol{\omega}) < \infty$. Then for $\forall \tau > 0$ and compact set $\mathcal{S} \subset \mathbb{R}^d$

$$\Lambda^m \left(\|\hat{k} - k\|_{L^\infty(\mathcal{S})} \geq \frac{h(d, |\mathcal{S}|, \sigma) + \sqrt{2\tau}}{\sqrt{m}} \right) \leq e^{-\tau},$$

$$h(d, |\mathcal{S}|, \sigma) := 32\sqrt{2d \log(2|\mathcal{S}| + 1)} + 16\sqrt{\frac{2d}{\log(2|\mathcal{S}| + 1)}} + 32\sqrt{2d \log(\sigma + 1)}.$$

Consequence-1 (Borel-Cantelli lemma)

- A.s. convergence on compact sets: $\hat{k} \xrightarrow{m\to\infty} k$ at rate $\sqrt{\frac{\log|\mathcal{S}|}{m}}$.

Consequence-1 (Borel-Cantelli lemma)

- A.s. convergence on compact sets: $\hat{k} \xrightarrow{m\to\infty} k$ at rate $\sqrt{\frac{\log|\mathcal{S}|}{m}}$.
- Growing diameter:
  - $\frac{\log|\mathcal{S}_m|}{m} \xrightarrow{m\to\infty} 0$ is enough, i.e. $|\mathcal{S}_m| = e^{o(m)}$.

Consequence-1 (Borel-Cantelli lemma)

- A.s. convergence on compact sets: $\hat{k} \xrightarrow{m\to\infty} k$ at rate $\sqrt{\frac{\log|\mathcal{S}|}{m}}$.
- Growing diameter:
  - $\frac{\log|\mathcal{S}_m|}{m} \xrightarrow{m\to\infty} 0$ is enough, i.e. $|\mathcal{S}_m| = e^{o(m)}$.
- Specifically:
  - *asymptotic* optimality [Csörgő and Totik, 1983, Theorem 2] (if $k(\mathbf{z})$ vanishes at ∞).

Consequence-2: L^r result for k $(1 \le r)$

Idea:

- Note that

$$\|\hat{k} - k\|_{L^r(\mathcal{S})} = \left(\int_{\mathcal{S}} \int_{\mathcal{S}} |\hat{k}(\mathbf{x}, \mathbf{y}) - k(\mathbf{x}, \mathbf{y})|^r \, \mathrm{d}\mathbf{x} \, \mathrm{d}\mathbf{y} \right)^{\frac{1}{r}} \le \|\hat{k} - k\|_{L^\infty(\mathcal{S})} \mathrm{vol}^{2/r}(\mathcal{S}).$$

Consequence-2: L^r result for k $(1 \le r)$

Idea:

- Note that

$$\|\hat{k} - k\|_{L^r(\mathcal{S})} = \left(\int_{\mathcal{S}}\int_{\mathcal{S}} |\hat{k}(\mathbf{x},\mathbf{y}) - k(\mathbf{x},\mathbf{y})|^r \,\mathrm{d}\mathbf{x}\,\mathrm{d}\mathbf{y}\right)^{\frac{1}{r}}$$
$$\le \|\hat{k} - k\|_{L^\infty(\mathcal{S})} \mathrm{vol}^{2/r}(\mathcal{S}).$$

- $\mathrm{vol}(\mathcal{S}) \le \mathrm{vol}(B)$, where $B := \left\{\mathbf{x} \in \mathbb{R}^d : \|\mathbf{x}\|_2 \le \frac{|\mathcal{S}|}{2}\right\}$,

Consequence-2: L^r result for k $(1 \le r)$

Idea:

- Note that

$$\|\hat{k} - k\|_{L^r(\mathcal{S})} = \left(\int_{\mathcal{S}} \int_{\mathcal{S}} |\hat{k}(\mathbf{x}, \mathbf{y}) - k(\mathbf{x}, \mathbf{y})|^r \, \mathrm{d}\mathbf{x} \, \mathrm{d}\mathbf{y}\right)^{\frac{1}{r}} \le \|\hat{k} - k\|_{L^\infty(\mathcal{S})} \mathrm{vol}^{2/r}(\mathcal{S}).$$

- $\mathrm{vol}(\mathcal{S}) \le \mathrm{vol}(B)$, where $B := \left\{\mathbf{x} \in \mathbb{R}^d : \|\mathbf{x}\|_2 \le \frac{|\mathcal{S}|}{2}\right\}$,
- $\mathrm{vol}(B) = \frac{\pi^{d/2}|\mathcal{S}|^d}{2^d \Gamma\left(\frac{d}{2}+1\right)}$, $\Gamma(t) = \int_0^\infty u^{t-1} e^{-u} \, \mathrm{d}u$. $\Rightarrow$

L^r result for k

Under the previous assumptions, and $1 \le r < \infty$:

$$\Lambda^m \left(\|\hat{k} - k\|_{L^r(\mathcal{S})} \ge \left(\frac{\pi^{d/2} |\mathcal{S}|^d}{2^d \Gamma(\frac{d}{2}+1)} \right)^{2/r} \frac{h(d, |\mathcal{S}|, \sigma) + \sqrt{2\tau}}{\sqrt{m}} \right) \le e^{-\tau}.$$

L^r result for k

Under the previous assumptions, and $1 \le r < \infty$:

$$\Lambda^m \left(\|\hat{k} - k\|_{L^r(\mathcal{S})} \ge \left(\frac{\pi^{d/2} |\mathcal{S}|^d}{2^d \Gamma(\frac{d}{2} + 1)} \right)^{2/r} \frac{h(d, |\mathcal{S}|, \sigma) + \sqrt{2\tau}}{\sqrt{m}} \right) \le e^{-\tau}.$$

Hence,

$$\|\hat{k} - k\|_{L^r(\mathcal{S})} = \mathcal{O}_{a.s.} \Bigg(\underbrace{\frac{|\mathcal{S}|^{2d/r} \sqrt{\log |\mathcal{S}|}}{\sqrt{m}}}_{L^r(\mathcal{S})\text{-consistency if } \xrightarrow{m \to \infty} 0} \Bigg).$$

L^r result for k

Under the previous assumptions, and $1 \le r < \infty$:

$$\Lambda^m \left(\|\hat{k} - k\|_{L^r(\mathcal{S})} \geq \left(\frac{\pi^{d/2}|\mathcal{S}|^d}{2^d \Gamma(\frac{d}{2}+1)} \right)^{2/r} \frac{h(d, |\mathcal{S}|, \sigma) + \sqrt{2\tau}}{\sqrt{m}} \right) \le e^{-\tau}.$$

Hence,

$$\|\hat{k} - k\|_{L^r(\mathcal{S})} = \mathcal{O}_{a.s.}\Bigg(\underbrace{\frac{|\mathcal{S}|^{2d/r}\sqrt{\log|\mathcal{S}|}}{\sqrt{m}}}_{L^r(\mathcal{S})\text{-consistency if } \xrightarrow{m\to\infty} 0} \Bigg).$$

Uniform guarantee: $|\mathcal{S}_m| = e^{m^{\delta<1}}$; now: $\frac{|\mathcal{S}_m|^{2d/r}}{\sqrt{m}} \to 0 \Rightarrow |\mathcal{S}_m| = \tilde{o}(m^{\frac{r}{4d}})$.

Direct L^r result for k (proof idea after discussion)

Under the previous assumptions, and $1 < r < \infty$:

$$\Lambda^m \left(\|\hat{k} - k\|_{L^r(\mathcal{S})} \geq \left(\frac{\pi^{d/2}|\mathcal{S}|^d}{2^d \Gamma(\frac{d}{2}+1)} \right)^{2/r} \left(\frac{C'_r}{m^{1-\max\{\frac{1}{2},\frac{1}{r}\}}} + \frac{\sqrt{2\tau}}{\sqrt{m}} \right) \right) \leq e^{-\tau},$$

$C'_r = \mathcal{O}(\sqrt{r})$: universal constant; only r-dependent (not $|\mathcal{S}|$ or m-dep.).

Direct L^r result for k (proof idea after discussion)

Under the previous assumptions, and $1 < r < \infty$:

$$\Lambda^m\left(\|\hat{k}-k\|_{L^r(\mathcal{S})} \geq \left(\frac{\pi^{d/2}|\mathcal{S}|^d}{2^d\Gamma(\frac{d}{2}+1)}\right)^{2/r}\left(\frac{C'_r}{m^{1-\max\{\frac{1}{2},\frac{1}{r}\}}}+\frac{\sqrt{2\tau}}{\sqrt{m}}\right)\right) \leq e^{-\tau},$$

$C'_r = \mathcal{O}(\sqrt{r})$: universal constant; only r-dependent (not $|\mathcal{S}|$ or m-dep.).
Note: if $2 \leq r$, then

1. $m^{1-\max\{\frac{1}{2},\frac{1}{r}\}} = \sqrt{m}$,
2. $\|\hat{k}-k\|_{L^r(\mathcal{S}_m)} \xrightarrow{a.s.} 0$ if $|\mathcal{S}_m| = o\left(m^{\frac{r}{4d}}\right)$ as $m \to \infty$.
3. In short, we got rid of $\sqrt{\log(|\mathcal{S}|)}$: $\tilde{o} \to o$.

Direct L^r result for k: proof idea

1. $f(\boldsymbol{\omega}_1, \ldots, \boldsymbol{\omega}_m) = \|k - \hat{k}\|_{L^r(\mathcal{S})}$ concentrates (bounded difference):

$$\|k - \hat{k}\|_{L^r(\mathcal{S})} \le \mathbb{E}_{\boldsymbol{\omega}_{1:m}} \|k - \hat{k}\|_{L^r(\mathcal{S})} + \mathrm{vol}^{2/r}(\mathcal{S}) \sqrt{\frac{2\tau}{m}}.$$

Direct L^r result for k: proof idea

1. $f(\boldsymbol{\omega}_1, \ldots, \boldsymbol{\omega}_m) = \|k - \hat{k}\|_{L^r(\mathcal{S})}$ concentrates (bounded difference):

$$\|k - \hat{k}\|_{L^r(\mathcal{S})} \leq \mathbb{E}_{\boldsymbol{\omega}_{1:m}} \|k - \hat{k}\|_{L^r(\mathcal{S})} + \mathrm{vol}^{2/r}(\mathcal{S}) \sqrt{\frac{2\tau}{m}}.$$

2. By $L^r \cong (L^{r'})^*$ $\left(\frac{1}{r} + \frac{1}{r'} = 1\right)$, the separability of $L^{r'}(\mathcal{S})$ $(r > 1)$ and symmetrization:

$$\mathbb{E}_{\boldsymbol{\omega}_{1:m}} \|k - \hat{k}\|_{L^r(\mathcal{S})} \leq \frac{2}{m} \mathbb{E}_{\boldsymbol{\omega}_{1:m}} \underbrace{\mathbb{E}_{\boldsymbol{\varepsilon}} \left\| \sum_{i=1}^{m} \varepsilon_i \cos(\langle \boldsymbol{\omega}_i, \cdot - \cdot \rangle) \right\|_{L^r(\mathcal{S})}}_{=:(*)}.$$

Direct L^r result for k: proof idea

1. $f(\boldsymbol{\omega}_1,\ldots,\boldsymbol{\omega}_m) = \|k-\hat{k}\|_{L^r(\mathcal{S})}$ concentrates (bounded difference):

$$\|k-\hat{k}\|_{L^r(\mathcal{S})} \le \mathbb{E}_{\boldsymbol{\omega}_{1:m}}\|k-\hat{k}\|_{L^r(\mathcal{S})} + \mathrm{vol}^{2/r}(\mathcal{S})\sqrt{\frac{2\tau}{m}}.$$

2. By $L^r \cong (L^{r'})^*$ $(\frac{1}{r}+\frac{1}{r'}=1)$, the separability of $L^{r'}(\mathcal{S})$ $(r>1)$ and symmetrization:

$$\mathbb{E}_{\boldsymbol{\omega}_{1:m}}\|k-\hat{k}\|_{L^r(\mathcal{S})} \le \frac{2}{m}\mathbb{E}_{\boldsymbol{\omega}_{1:m}} \underbrace{\mathbb{E}_{\boldsymbol{\varepsilon}}\left\|\sum_{i=1}^m \varepsilon_i \cos(\langle \boldsymbol{\omega}_i, \cdot - \cdot\rangle)\right\|_{L^r(\mathcal{S})}}_{=:(*)}.$$

3. Since $L^r(\mathcal{S})$ is of type $\min(2,r)$ ['$\diamond$-rule'] $\exists C_r'$ such that

$$(*) \le C_r' \left(\sum_{i=1}^m \|\cos(\langle \boldsymbol{\omega}_i, \cdot - \cdot\rangle)\|_{L^r(\mathcal{S})}^{\min(2,r)}\right)^{\frac{1}{\min(2,r)}}.$$

Kernel derivatives: $\mathbb{N}^{2d} \ni [\mathbf{p}; \mathbf{q}] \neq 0$

Goal: $\widehat{k^{\mathbf{p},\mathbf{q}}}$. If

1. *supp*(Λ) is bounded:
   - $C_{k,\mathbf{p},\mathbf{q}} := \mathbb{E}_{\boldsymbol{\omega}\sim\Lambda}\left[\left|\boldsymbol{\omega}^{\mathbf{p}+\mathbf{q}}\right| \|\boldsymbol{\omega}\|_2^2\right] < \infty$: L^{∞}, L^r✓, but
   - Gaussian, Laplacian, inverse multiquadratic, Matern:(
   - c_0 universality $\Leftrightarrow$ *supp*$(\Lambda) = \mathbb{R}^d$, if $k(\mathbf{z}) \in C_0(\mathbb{R}^d)$.

Kernel derivatives: $\mathbb{N}^{2d} \ni [\mathbf{p};\mathbf{q}] \neq 0$

Goal: $\widehat{k^{\mathbf{p},\mathbf{q}}}$. If

1. *supp*(Λ) is bounded:
   - $C_{k,\mathbf{p},\mathbf{q}} := \mathbb{E}_{\boldsymbol{\omega}\sim\Lambda}\left[\left|\boldsymbol{\omega}^{\mathbf{p}+\mathbf{q}}\right| \|\boldsymbol{\omega}\|_2^2\right] < \infty$: $L^\infty, L^r \checkmark$, but
   - Gaussian, Laplacian, inverse multiquadratic, Matern:(
   - c_0 universality $\Leftrightarrow$ *supp*$(\Lambda) = \mathbb{R}^d$, if $k(\mathbf{z}) \in C_0(\mathbb{R}^d)$.
2. *supp*(Λ) is unbounded:
   - $\mathcal{G}$: becomes unbounded.
   - [Rahimi and Recht, 2007]: 'Hoeffding $\rightarrow$ Bernstein', but

Kernel derivatives: unbounded $supp(\Lambda)$

Assumptions [$h_a = cos^{(a)}$, $\mathcal{S}_\Delta = \mathcal{S} - \mathcal{S}$]:

1. $\mathbf{z} \mapsto \nabla_{\mathbf{z}} \left[\partial^{\mathbf{p},\mathbf{q}} k(\mathbf{z})\right]$: continuous; $\mathcal{S} \subset \mathbb{R}^d$: compact, $E_{\mathbf{p},\mathbf{q}} := \mathbb{E}_{\boldsymbol{\omega}\sim\Lambda} |\boldsymbol{\omega}^{\mathbf{p}+\mathbf{q}}| \, \|\boldsymbol{\omega}\|_2 < \infty$.
2. $\exists L > 0, \sigma > 0$

$$\mathbb{E}_{\boldsymbol{\omega}\sim\Lambda} |f(\mathbf{z}; \boldsymbol{\omega})|^M \leq \frac{M!\, \sigma^2 L^{M-2}}{2} \quad (\forall M \geq 2, \forall \mathbf{z} \in \mathcal{S}_\Delta),$$

$$f(\mathbf{z}; \boldsymbol{\omega}) = \partial^{\mathbf{p},\mathbf{q}} k(\mathbf{z}) - \boldsymbol{\omega}^{\mathbf{p}} (-\boldsymbol{\omega})^{\mathbf{q}} h_{|\mathbf{p}+\mathbf{q}|}(\boldsymbol{\omega}^T \mathbf{z}).$$

Kernel derivatives: unbounded $supp(\Lambda)$

Then with $F_d := d^{-\frac{d}{d+1}} + d^{\frac{1}{d+1}}$

$$\Lambda^m \left(\|\partial^{\mathbf{p},\mathbf{q}} k - \widehat{\partial^{\mathbf{p},\mathbf{q}} k}\|_{L^\infty(\mathcal{S})} \geq \epsilon \right) \leq$$

$$\leq 2^{d-1} e^{-\frac{m\epsilon^2}{8\sigma^2\left(1+\frac{\epsilon L}{2\sigma^2}\right)}} + F_d 2^{\frac{4d-1}{d+1}} \left[\frac{|\mathcal{S}|(D_{\mathbf{p},\mathbf{q},\mathcal{S}} + E_{\mathbf{p},\mathbf{q}})}{\epsilon} \right]^{\frac{d}{d+1}} e^{-\frac{m\epsilon^2}{8(d+1)\sigma^2\left(1+\frac{\epsilon L}{2\sigma^2}\right)}},$$

where $D_{\mathbf{p},\mathbf{q},\mathcal{S}} := \sup_{\mathbf{z}\in conv(\mathcal{S}_\Delta)} \|\nabla_{\mathbf{z}} [\partial^{\mathbf{p},\mathbf{q}} k(\mathbf{z})]\|_2$.

Summary

Finite sample

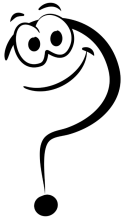
- $L^\infty(\mathcal{S})$ guarantees $\xrightarrow{\text{spec.}} |\mathcal{S}_m| = e^{o(m)}$ – asymp. optimal!
- $L^r(\mathcal{S})$ results ($\Leftarrow$ uniform, type of L^r).

Summary

Finite sample

- $L^\infty(\mathcal{S})$ guarantees $\xrightarrow{\text{spec.}}$ $|\mathcal{S}_m| = e^{o(m)}$ – asymp. optimal!
- $L^r(\mathcal{S})$ results ($\Leftarrow$ uniform, type of L^r).
- derivative approximation guarantees:
  - bounded spectral support: ✓
  - unbounded spectral support: trickier – to be continued;)

Thank you for the attention!



Acknowledgments: This work was supported by the Gatsby Charitable Foundation.

Csörgő, S. and Totik, V. (1983).
On how long interval is the empirical characteristic function uniformly consistent?
Acta Scientiarum Mathematicarum, 45:141–149.

Rahimi, A. and Recht, B. (2007).
Random features for large-scale kernel machines.
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.
JMLR W&CP – International Conference on Artificial Intelligence and Statistics (AISTATS), 9:653–660.

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

Shi, L., Guo, X., and Zhou, D.-X. (2010).
Hermite learning with gradient data.
Journal of Computational and Applied Mathematics, 233:3046–3059.

Sriperumbudur, B. K., Fukumizu, K., Gretton, A., Hyvärinen, A., and Kumar, R. (2014).
Density estimation in infinite dimensional exponential families.
Technical report.
http://arxiv.org/pdf/1312.3516.pdf.

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

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

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