Performance Guarantees for Random Fourier Features – Limitations and Merits

Zoltán Szabó

Joint work with Bharath K. Sriperumbudur (PSU)

ML@SITraN, University of Sheffield
June 25, 2015

Context

- Given:

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

- $\hat{k}(\mathbf{x},\mathbf{y})$: Monte-Carlo estimator of $k(\mathbf{x},\mathbf{y})$ using $(\boldsymbol{\omega}_j)_{j=1}^m \overset{i.i.d.}{\sim} \Lambda$ [Rahimi and Recht, 2007].
- Motivation:
  - Primal form – fast linear solvers.
  - Kernel function approximation: out-of-sample extension.
  - Online applications.

Performance measures

- Uniform ($r = \infty$):

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

- L^r ($1 \le r < \infty$):

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

Approximation of kernel derivatives

- One could also consider $\partial^{\mathbf{p},\mathbf{q}} k$.
- Motivation [Zhou, 2008, Shi et al., 2010, Rosasco et al., 2010, Rosasco et al., 2013, Ying et al., 2012, Sriperumbudur et al., 2014]:
  - semi-supervised learning with gradient information,
  - nonlinear variable selection,
  - fitting of infD exp. family distributions.
- Many of the presented results hold for derivatives ($[\mathbf{p}; \mathbf{q}] \neq \mathbf{0}$).

Goal

- Large deviation inequalities

$$\Lambda^m\left(\left\|k-\hat{k}\right\|_{\mathcal{S}}\le\epsilon\right)\ge f_1(\epsilon,d,m,|\mathcal{S}|),$$
$$\Lambda^m\left(\left\|k-\hat{k}\right\|_{L^r}\le\epsilon\right)\ge f_2(\epsilon,d,m,|\mathcal{S}|).$$

- Scaling of $|\mathcal{S}|$ and m ensuring a.s. convergence?

Existing results on the approximation quality

Notations: $X_n = \mathcal{O}_p(r_n)$ ($\mathcal{O}_{a.s.}(r_n)$) denotes $\frac{X_n}{r_n}$ boundedness in probability (almost surely).

- [Rahimi and Recht, 2007]:

$$\left\| \hat{k}(\mathbf{x}, \mathbf{y}) - k(\mathbf{x}, \mathbf{y}) \right\|_{\mathcal{S}} = \mathcal{O}_p\left(|\mathcal{S}| \sqrt{\frac{\log m}{m}} \right).$$

- [Sutherland and Schneider, 2015]: better constants.

Contents

- Uniform guarantee (empirical process theory),
- Two L^r guarantees (uniform consequence, direct).
- Kernel derivatives.

High-level proof

1. Empirical process form:

$$\left\|k - \hat{k}\right\|_{\mathcal{S}} = \sup_{g \in \mathcal{G}} |\Lambda g - \Lambda_m g| = \|\Lambda - \Lambda_m\|_{\mathcal{G}}.$$

2. $\|\Lambda - \Lambda_m\|_{\mathcal{G}}$ concentrates by its bounded difference property:

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

3. $\mathcal{G}$ is a uniformly bounded, separable Carathéodory family $\Rightarrow$

$$\mathbb{E}_{\boldsymbol{\omega}_{1:m}} \|\Lambda - \Lambda_m\|_{\mathcal{G}} \lesssim \mathbb{E}_{\boldsymbol{\omega}_{1:m}} \mathcal{R}\left(\mathcal{G}, \boldsymbol{\omega}_{1:m}\right).$$

High-level proof

4. Using Dudley's entropy integral:

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

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

$$\sqrt{\log \mathcal{N}(\mathcal{G}, L^2(\Lambda_m), r)} \le \sqrt{\log\left[\frac{C\left(\boldsymbol{\omega}_{1:m}\right)}{r}+1\right]} \Rightarrow$$

$$\mathbb{E}_{\boldsymbol{\omega}_{1:m}}\mathcal{R}\left(\mathcal{G},\boldsymbol{\omega}_{1:m}\right) \lesssim \frac{1}{\sqrt{m}}.$$

6. Putting together:

$$\left\|k-\hat{k}\right\|_{\mathcal{S}} \lesssim \frac{1}{\sqrt{m}}+\frac{1}{\sqrt{m}} = \mathcal{O}\left(\sqrt{\frac{\log|\mathcal{S}|}{m}}\right).$$

Step-1: empirical process form

- Notation: $\Lambda g = \int g(\omega)\mathrm{d}\Lambda(\omega)$, $\Lambda_m g = \int g(\omega)\mathrm{d}\Lambda_m(\omega) = \frac{1}{m}\sum_{j=1}^{m} g(\omega_j)$.

Step-1: empirical process form

- Notation: $\Lambda g = \int g(\boldsymbol{\omega}) \mathrm{d}\Lambda(\boldsymbol{\omega})$, $\Lambda_m g = \int g(\boldsymbol{\omega}) \mathrm{d}\Lambda_m(\boldsymbol{\omega}) = \frac{1}{m} \sum_{j=1}^m g(\boldsymbol{\omega}_j)$.
- Reformulation of the objective:

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

where

$$\mathcal{G} = \{g_{\mathbf{z}} : \mathbf{z} \in \mathcal{S}_\Delta\},$$
$$\mathcal{S}_\Delta = \mathcal{S} - \mathcal{S} = \{\mathbf{x} - \mathbf{y} : \mathbf{x}, \mathbf{y} \in \mathcal{S}\},$$
$$g_{\mathbf{z}} : \boldsymbol{\omega} \mapsto \cos\left(\boldsymbol{\omega}^T \mathbf{z}\right).$$

Step-2: bounded difference property of $\|\Lambda - \Lambda_m\|_{\mathcal{G}}$

McDiarmid inequality: Let $\boldsymbol{\omega}_1, \ldots, \boldsymbol{\omega}_m \in D$ be independent r.v.-s, and $f : D^m \to \mathbb{R}$ satisfy the bounded diff. property $(\forall r)$:

$$\sup_{u_1,\ldots,u_m,u_r' \in D} \left| f(u_1, \ldots, u_m) - f(u_1, \ldots, u_{r-1}, u_r', u_{r+1}, \ldots, u_m) \right| \leq c_r.$$

Then for $\forall \beta > 0$

$$\mathbb{P}\left(f(\boldsymbol{\omega}_1, \ldots, \boldsymbol{\omega}_m) - \mathbb{E}\left[f(\boldsymbol{\omega}_1, \ldots, \boldsymbol{\omega}_m)\right] \geq \beta\right) \leq e^{-\frac{2\beta^2}{\sum_{r=1}^m c_r^2}}.$$

Step-2: bounded difference property of $\|\Lambda - \Lambda_m\|_{\mathcal{G}}$

Our choice: $f(\omega_1, \ldots, \omega_m) := \|\Lambda - \Lambda_m\|_{\mathcal{G}}$.

$$|f(\omega_1, \ldots, \omega_{r-1}, \omega_r, \omega_{r+1}, \ldots, \omega_m) - f(\omega_1, \ldots, \omega_{r-1}, \omega_r', \omega_{r+1}, \ldots, \omega_m)| =$$

$$= \left| \sup_{g \in \mathcal{G}} \left| \Lambda g - \frac{1}{m} \sum_{j=1} g(\omega_j) \right| - \sup_{g \in \mathcal{G}} \left| \Lambda g - \frac{1}{m} \sum_{j=1} g(\omega_j) + \frac{1}{m} \left[g(\omega_r) - g(\omega_r') \right] \right| \right|$$

Step-2: bounded difference property of $\|\Lambda - \Lambda_m\|_{\mathcal{G}}$

Our choice: $f(\omega_1, \ldots, \omega_m) := \|\Lambda - \Lambda_m\|_{\mathcal{G}}$.

$$|f(\omega_1, \ldots, \omega_{r-1}, \omega_r, \omega_{r+1}, \ldots, \omega_m) - f(\omega_1, \ldots, \omega_{r-1}, \omega'_r, \omega_{r+1}, \ldots, \omega_m)| =$$

$$= \left| \sup_{g \in \mathcal{G}} \left| \Lambda g - \frac{1}{m} \sum_{j=1} g(\omega_j) \right| - \sup_{g \in \mathcal{G}} \left| \Lambda g - \frac{1}{m} \sum_{j=1} g(\omega_j) + \frac{1}{m} \left[g(\omega_r) - g(\omega'_r) \right] \right| \right|$$

$$\overset{(*)}{\leq} \frac{1}{m} \sup_{g \in \mathcal{G}} \left| g(\omega_r) - g(\omega'_r) \right|$$

Step-2: bounded difference property of $\|\Lambda - \Lambda_m\|_{\mathcal{G}}$

Our choice: $f(\omega_1, \ldots, \omega_m) := \|\Lambda - \Lambda_m\|_{\mathcal{G}}$.

$$|f(\omega_1, \ldots, \omega_{r-1}, \omega_r, \omega_{r+1}, \ldots, \omega_m) - f(\omega_1, \ldots, \omega_{r-1}, \omega_r', \omega_{r+1}, \ldots, \omega_m)| =$$

$$= \left| \sup_{g\in\mathcal{G}} \left| \Lambda g - \frac{1}{m}\sum_{j=1} g(\omega_j) \right| - \sup_{g\in\mathcal{G}} \left| \Lambda g - \frac{1}{m}\sum_{j=1} g(\omega_j) + \frac{1}{m}\left[g(\omega_r) - g(\omega_r')\right] \right| \right|$$

$$\overset{(*)}{\leq} \frac{1}{m}\sup_{g\in\mathcal{G}} \left|g(\omega_r) - g(\omega_r')\right| \leq \frac{1}{m}\sup_{g\in\mathcal{G}} \left(\left|g(\omega_r)\right| + \left|g(\omega_r')\right|\right)$$

Step-2: bounded difference property of $\|\Lambda - \Lambda_m\|_{\mathcal{G}}$

Our choice: $f(\omega_1, \ldots, \omega_m) := \|\Lambda - \Lambda_m\|_{\mathcal{G}}$.

$$|f(\omega_1, \ldots, \omega_{r-1}, \omega_r, \omega_{r+1}, \ldots, \omega_m) - f(\omega_1, \ldots, \omega_{r-1}, \omega_r', \omega_{r+1}, \ldots, \omega_m)| =$$

$$= \left| \sup_{g \in \mathcal{G}} \left| \Lambda g - \frac{1}{m} \sum_{j=1} g(\omega_j) \right| - \sup_{g \in \mathcal{G}} \left| \Lambda g - \frac{1}{m} \sum_{j=1} g(\omega_j) + \frac{1}{m} \left[g(\omega_r) - g(\omega_r') \right] \right| \right|$$

$$\overset{(*)}{\leq} \frac{1}{m} \sup_{g \in \mathcal{G}} \left| g(\omega_r) - g(\omega_r') \right| \leq \frac{1}{m} \sup_{g \in \mathcal{G}} \left(|g(\omega_r)| + |g(\omega_r')| \right)$$

$$\leq \frac{1}{m} \left[\sup_{g \in \mathcal{G}} |g(\omega_r)| + \sup_{g \in \mathcal{G}} |g(\omega_r')| \right]$$

Step-2: bounded difference property of $\|\Lambda - \Lambda_m\|_{\mathcal{G}}$

Our choice: $f(\omega_1, \ldots, \omega_m) := \|\Lambda - \Lambda_m\|_{\mathcal{G}}$.

$$|f(\omega_1, \ldots, \omega_{r-1}, \omega_r, \omega_{r+1}, \ldots, \omega_m) - f(\omega_1, \ldots, \omega_{r-1}, \omega'_r, \omega_{r+1}, \ldots, \omega_m)| =$$

$$= \left| \sup_{g \in \mathcal{G}} \left| \Lambda g - \frac{1}{m} \sum_{j=1} g(\omega_j) \right| - \sup_{g \in \mathcal{G}} \left| \Lambda g - \frac{1}{m} \sum_{j=1} g(\omega_j) + \frac{1}{m} \left[g(\omega_r) - g(\omega'_r) \right] \right| \right|$$

$$\overset{(*)}{\leq} \frac{1}{m} \sup_{g \in \mathcal{G}} \left| g(\omega_r) - g(\omega'_r) \right| \leq \frac{1}{m} \sup_{g \in \mathcal{G}} \left(|g(\omega_r)| + |g(\omega'_r)| \right)$$

$$\leq \frac{1}{m} \left[\sup_{g \in \mathcal{G}} |g(\omega_r)| + \sup_{g \in \mathcal{G}} |g(\omega'_r)| \right] \leq \frac{1+1}{m} = \frac{2}{m}.$$

Step-2: (*) = reverse triangle inequality with sup

- Lemma: $\mathcal{G}$: set of functions, $a, b : \mathcal{G} \to \mathbb{R}$ maps; then

$$\left| \sup_{g \in \mathcal{G}} |a(g)| - \sup_{g \in \mathcal{G}} |a(g) + b(g)| \right|$$

Step-2: (*) = reverse triangle inequality with sup

- Lemma: $\mathcal{G}$: set of functions, $a, b : \mathcal{G} \to \mathbb{R}$ maps; then

$$\left| \sup_{g\in\mathcal{G}} |a(g)| - \sup_{g\in\mathcal{G}} |a(g) + b(g)| \right| \leq \sup_{g\in\mathcal{G}} |b(g)|.$$

Step-2: (*) = reverse triangle inequality with sup

- Lemma: $\mathcal{G}$: set of functions, $a, b : \mathcal{G} \to \mathbb{R}$ maps; then

$$\left| \sup_{g\in\mathcal{G}} |a(g)| - \sup_{g\in\mathcal{G}} |a(g) + b(g)| \right| \leq \sup_{g\in\mathcal{G}} |b(g)|.$$

- Proof: combine

$$sup_{g\in\mathcal{G}} |a(g) + b(g)| \leq \sup_{g\in\mathcal{G}} (|a(g)| + |b(g)|) \leq \sup_{g\in\mathcal{G}} |a(g)| + \sup_{g\in\mathcal{G}} |b(g)|,$$

Step-2: (*) = reverse triangle inequality with sup

- Lemma: $\mathcal{G}$: set of functions, $a, b : \mathcal{G} \to \mathbb{R}$ maps; then

$$\left| \sup_{g\in\mathcal{G}} |a(g)| - \sup_{g\in\mathcal{G}} |a(g) + b(g)| \right| \leq \sup_{g\in\mathcal{G}} |b(g)|.$$

- Proof: combine

$$sup_{g\in\mathcal{G}} |a(g) + b(g)| \leq \sup_{g\in\mathcal{G}} \left(|a(g)| + |b(g)|\right) \leq \sup_{g\in\mathcal{G}} |a(g)| + \sup_{g\in\mathcal{G}} |b(g)|,$$

$$\begin{aligned} \sup_{g\in\mathcal{G}} |a(g)| &= \sup_{g\in\mathcal{G}} |a(g) + b(g) - b(g)| \\ &\leq \sup_{g\in\mathcal{G}} |a(g) + b(g)| + \sup_{g\in\mathcal{G}} |b(g)|. \end{aligned}$$

Step-2: (*) = reverse triangle inequality with sup

- Lemma: $\mathcal{G}$: set of functions, $a, b : \mathcal{G} \to \mathbb{R}$ maps; then

$$\left| \sup_{g\in\mathcal{G}} |a(g)| - \sup_{g\in\mathcal{G}} |a(g) + b(g)| \right| \leq \sup_{g\in\mathcal{G}} |b(g)|.$$

- Proof: combine

$$sup_{g\in\mathcal{G}} |a(g) + b(g)| \leq \sup_{g\in\mathcal{G}} (|a(g)| + |b(g)|) \leq \sup_{g\in\mathcal{G}} |a(g)| + \sup_{g\in\mathcal{G}} |b(g)|,$$

$$\begin{aligned} \sup_{g\in\mathcal{G}} |a(g)| &= \sup_{g\in\mathcal{G}} |a(g) + b(g) - b(g)| \\ &\leq \sup_{g\in\mathcal{G}} |a(g) + b(g)| + \sup_{g\in\mathcal{G}} |b(g)|. \end{aligned}$$

$$\Rightarrow \pm \left[\sup_{g\in\mathcal{G}} |a(g)| - \sup_{g\in\mathcal{G}} |a(g) + b(g)| \right] \leq \sup_{g\in\mathcal{G}} |b(g)|.$$

- Our choice: $a(g) = \Lambda g - \frac{1}{m} \sum_{j=1} g(\omega_j)$, $b(g) = \frac{1}{m} [g(\omega_r) - g(\omega'_r)]$.

Step-2

Applying McDiarmid to f ($c_r = \frac{2}{m}$): with probability $1 - e^{-\tau}$

$$\|\Lambda - \Lambda_m\|_{\mathcal{G}} \leq \underbrace{\mathbb{E}_{\boldsymbol{\omega}_{1:m}} \|\Lambda - \Lambda_m\|_{\mathcal{G}}}_{\text{Step-3: bounding this term}} + \frac{\sqrt{2\tau}}{\sqrt{m}}.$$

Step-3: bounding $\mathbb{E}_{\omega_1,\ldots,\omega_m} \|\Lambda - \Lambda_m\|_{\mathcal{G}}$

$\mathcal{G} = \{g_{\mathbf{z}} : \mathbf{z} \in \mathcal{S}_\Delta\}$ is a separable Carathéodory family, i.e.

1. $\boldsymbol{\omega} \mapsto \cos\left(\boldsymbol{\omega}^T \mathbf{z}\right)$: measurable for $\forall \mathbf{z} \in \mathcal{S}_\Delta$.

Step-3: bounding $\mathbb{E}_{\omega_1,\ldots,\omega_m} \|\Lambda - \Lambda_m\|_{\mathcal{G}}$

$\mathcal{G} = \{g_{\mathbf{z}} : \mathbf{z} \in \mathcal{S}_\Delta\}$ is a separable Carathéodory family, i.e.

1. $\boldsymbol{\omega} \mapsto \cos\left(\boldsymbol{\omega}^T \mathbf{z}\right)$: measurable for $\forall \mathbf{z} \in \mathcal{S}_\Delta$.
2. $\mathbf{z} \mapsto \cos\left(\boldsymbol{\omega}^T \mathbf{z}\right)$: continuous for $\forall \boldsymbol{\omega}$.

Step-3: bounding $\mathbb{E}_{\omega_1,\ldots,\omega_m} \|\Lambda - \Lambda_m\|_{\mathcal{G}}$

$\mathcal{G} = \{g_{\mathbf{z}} : \mathbf{z} \in \mathcal{S}_\Delta\}$ is a separable Carathéodory family, i.e.

1. $\boldsymbol{\omega} \mapsto \cos\left(\boldsymbol{\omega}^T \mathbf{z}\right)$: measurable for $\forall \mathbf{z} \in \mathcal{S}_\Delta$.
2. $\mathbf{z} \mapsto \cos\left(\boldsymbol{\omega}^T \mathbf{z}\right)$: continuous for $\forall \boldsymbol{\omega}$.
3. $\mathbb{R}^d$ is separable, $\mathcal{S}_\Delta \subseteq \mathbb{R}^d \Rightarrow \mathcal{S}_\Delta$: separable.

Step-3: bounding $\mathbb{E}_{\omega_1,\ldots,\omega_m} \|\Lambda - \Lambda_m\|_{\mathcal{G}}$

$\mathcal{G} = \{g_{\mathbf{z}} : \mathbf{z} \in \mathcal{S}_\Delta\}$ is a separable Carathéodory family, i.e.

1. $\boldsymbol{\omega} \mapsto \cos\left(\boldsymbol{\omega}^T \mathbf{z}\right)$: measurable for $\forall \mathbf{z} \in \mathcal{S}_\Delta$.
2. $\mathbf{z} \mapsto \cos\left(\boldsymbol{\omega}^T \mathbf{z}\right)$: continuous for $\forall \boldsymbol{\omega}$.
3. $\mathbb{R}^d$ is separable, $\mathcal{S}_\Delta \subseteq \mathbb{R}^d \Rightarrow \mathcal{S}_\Delta$: separable.

Thus, by [Steinwart and Christmann, 2008, Prop. 7.10]

$$\mathbb{E}_{\boldsymbol{\omega}_{1:m}} \|\Lambda - \Lambda_m\|_{\mathcal{G}} \le 2\mathbb{E}_{\boldsymbol{\omega}_{1:m}}\Big[\underbrace{\mathcal{R}\left(\mathcal{G}, \boldsymbol{\omega}_{1:m}\right)}_{:=\mathbb{E}_\epsilon \sup_{g\in\mathcal{G}} \left|\frac{1}{m}\sum_{j=1}^m \epsilon_j g(\omega_j)\right|} \Big]$$

using the uniformly boundedness of $\mathcal{G}$ $(\sup_{g\in\mathcal{G}} \|g\|_\infty \le 1)$.

Step-4: bounding $\mathcal{R}$

$$\mathcal{R}\left(\mathcal{G},(\boldsymbol{\omega}_j)_{j=1}^m\right) \le \frac{8\sqrt{2}}{\sqrt{m}} \int_0^{|\mathcal{G}|_{L^2(\Lambda_m)}} \sqrt{\log \mathcal{N}(\mathcal{G}, L^2(\Lambda_m), r)}\mathrm{d}r,$$

Step-4: bounding $\mathcal{R}$

$$\mathcal{R}\left(\mathcal{G},(\boldsymbol{\omega}_j)_{j=1}^m\right) \le \frac{8\sqrt{2}}{\sqrt{m}} \int_0^{|\mathcal{G}|_{L^2(\Lambda_m)}} \sqrt{\log \mathcal{N}(\mathcal{G}, L^2(\Lambda_m), r)}\mathrm{d}r,$$

where

- $L^2(\Lambda_m) = L^2(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d), \Lambda_m)$, $\|g\|_{L^2(\Lambda_m)} = \sqrt{\frac{1}{m}\sum_{j=1}^m g^2(\omega_j)}$,

Step-4: bounding $\mathcal{R}$

$$\mathcal{R}\left(\mathcal{G},(\boldsymbol{\omega}_j)_{j=1}^m\right) \le \frac{8\sqrt{2}}{\sqrt{m}} \int_0^{|\mathcal{G}|_{L^2(\Lambda_m)}} \sqrt{\log \mathcal{N}(\mathcal{G}, L^2(\Lambda_m), r)}\mathrm{d}r,$$

where

- $L^2(\Lambda_m) = L^2(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d), \Lambda_m)$, $\|g\|_{L^2(\Lambda_m)} = \sqrt{\frac{1}{m}\sum_{j=1}^m g^2(\omega_j)}$,
- $|\mathcal{G}|_{L^2(\Lambda_m)} = \sup_{g_1,g_2\in\mathcal{G}} \|g_1 - g_2\|_{L^2(\Lambda_m)}$,

Step-4: bounding $\mathcal{R}$

$$\mathcal{R}\left(\mathcal{G},(\boldsymbol{\omega}_j)_{j=1}^m\right) \le \frac{8\sqrt{2}}{\sqrt{m}} \int_0^{|\mathcal{G}|_{L^2(\Lambda_m)}} \sqrt{\log \mathcal{N}(\mathcal{G}, L^2(\Lambda_m), r)}\mathrm{d}r,$$

where

- $L^2(\Lambda_m) = L^2(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d), \Lambda_m)$, $\|g\|_{L^2(\Lambda_m)} = \sqrt{\frac{1}{m}\sum_{j=1}^m g^2(\omega_j)}$,
- $|\mathcal{G}|_{L^2(\Lambda_m)} = \sup_{g_1,g_2\in\mathcal{G}} \|g_1 - g_2\|_{L^2(\Lambda_m)}$,
- $\mathcal{N}(\mathcal{G}, L^2(\Lambda_m), r)$: r-covering number.
  - r-net: $S \subseteq \mathcal{G}$, for $\forall g \in \mathcal{G}$ $\exists s \in S$ such that $\|g - s\|_{L^2(\Lambda_m)} \le r$.
  - $\mathcal{N}$: size of the smallest r-net of $\mathcal{G}$.

Step-5: bound on $|\mathcal{G}|_{L^2(\Lambda_m)}$

$$|\mathcal{G}|_{L^2(\Lambda_m)} = \sup_{g_1,g_2\in\mathcal{G}} \|g_1 - g_2\|_{L^2(\Lambda_m)} \leq \sup_{g_1,g_2\in\mathcal{G}} \left(\|g_1\|_{L^2(\Lambda_m)} + \|g_2\|_{L^2(\Lambda_m)}\right)$$

$$\leq \sup_{g_1\in\mathcal{G}} \|g_1\|_{L^2(\Lambda_m)} + \sup_{g_1\in\mathcal{G}} \|g_2\|_{L^2(\Lambda_m)} \overset{*}{\leq} 2 \times 1,$$

Step-5: bound on $|\mathcal{G}|_{L^2(\Lambda_m)}$

$$|\mathcal{G}|_{L^2(\Lambda_m)} = \sup_{g_1,g_2\in\mathcal{G}} \|g_1 - g_2\|_{L^2(\Lambda_m)} \le \sup_{g_1,g_2\in\mathcal{G}} \left(\|g_1\|_{L^2(\Lambda_m)} + \|g_2\|_{L^2(\Lambda_m)}\right)$$

$$\le \sup_{g_1\in\mathcal{G}} \|g_1\|_{L^2(\Lambda_m)} + \sup_{g_1\in\mathcal{G}} \|g_2\|_{L^2(\Lambda_m)} \overset{*}{\le} 2 \times 1,$$

$$\sup_{g\in\mathcal{G}} \|g\|_{L^2(\Lambda_m)} = \sup_{\mathbf{z}\in\mathcal{S}_\Delta} \sqrt{\frac{1}{m}\sum_{j=1}^m g_{\mathbf{z}}^2(\boldsymbol{\omega}_j)}$$

$$= \sup_{\mathbf{z}\in\mathcal{S}_\Delta} \sqrt{\frac{1}{m}\sum_{j=1}^m \cos^2\left(\boldsymbol{\omega}_j^T\mathbf{z}\right)} \le \sup_{\mathbf{z}\in\mathcal{S}_\Delta} \sqrt{\frac{1}{m}\sum_{j=1}^m 1} = 1.$$

Step-5: bound on $\mathcal{N}(\mathcal{G}, L^2(\Lambda_m), r)$

Let $g_{\mathbf{z}_1}, g_{\mathbf{z}_2} \in \mathcal{G}$. We want to bound $\|g_{\mathbf{z}_1} - g_{\mathbf{z}_2}\|_{L^2(\Lambda_m)}$. One term:

$$\begin{aligned}
&\left|\cos\left(\boldsymbol{\omega}^T \mathbf{z}_1\right) - \cos\left(\boldsymbol{\omega}^T \mathbf{z}_2\right)\right| \\
&\quad = \left\|\nabla_{\mathbf{z}} \cos\left(\boldsymbol{\omega}^T \mathbf{z}_c\right)\right\|_2 \|\mathbf{z}_1 - \mathbf{z}_2\|_2 \\
&\quad = \left\|-\sin\left(\boldsymbol{\omega}^T \mathbf{z}_c\right)\boldsymbol{\omega}\right\|_2 \|\mathbf{z}_1 - \mathbf{z}_2\|_2 \\
&\quad \leq \|\boldsymbol{\omega}\|_2 \|\mathbf{z}_1 - \mathbf{z}_2\|_2,
\end{aligned}$$

where $\mathbf{z}_c \in (\mathbf{z}_1, \mathbf{z}_2)$.

Step-5: bound on $\mathcal{N}(\mathcal{G}, L^2(\Lambda_m), r)$

- Smooth parameterization:

$$\|g_{\mathbf{z}_1} - g_{\mathbf{z}_2}\|_{L^2(\Lambda_m)} \leq \sqrt{\frac{1}{m}\sum_{j=1}^{m}\left(\|\boldsymbol{\omega}_j\|_2 \|\mathbf{z}_1 - \mathbf{z}_2\|_2\right)^2} = \|\mathbf{z}_1 - \mathbf{z}_2\|_2 \underbrace{\sqrt{\frac{1}{m}\sum_{j=1}^{m}\|\boldsymbol{\omega}_j\|_2^2}}_{=:A}.$$

- r-net on $(\mathcal{S}_\Delta, \|\cdot\|_2) \Rightarrow r' = Ar$-net on $(\mathcal{G}, L^2(\Lambda_m))$.
- In other words, $\mathcal{N}\left(\mathcal{G}, L^2(\Lambda_m), r\right) \leq \mathcal{N}\left(\mathcal{S}_\Delta, \|\cdot\|_2, \frac{r}{A}\right)$.

Step-5: bound on $\mathcal{N}(\mathcal{G}, L^2(\Lambda_m), r)$

- Note that $\mathcal{S}_\Delta \subseteq B_{\|\cdot\|_2}\left(\mathbf{t}, \frac{|\mathcal{S}_\Delta|}{2}\right)$ for some $\mathbf{t} \in \mathbb{R}^d$.
- $\mathcal{N}(B_{\|\cdot\|_2}(\mathbf{s}, R), \|\cdot\|_2, \epsilon) \leq \left(\frac{2R}{\epsilon} + 1\right)^d$ for $\forall \mathbf{s} \in \mathbb{R}^d$.
- Thus

$$\mathcal{N}\left(\mathcal{G}, L^2(\Lambda_m), r\right) \leq \left(\frac{2|\mathcal{S}|A}{r} + 1\right)^d$$

by $|\mathcal{S}_\Delta| \leq 2|\mathcal{S}|$ and the compactness of $\mathcal{S}_\Delta$.

Combining the obtained

$$\mathcal{R}\left(\mathcal{G},\boldsymbol{\omega}_{1:m}\right) \le \frac{8\sqrt{2}}{\sqrt{m}}\int_0^{|\mathcal{G}|_{L^2(\Lambda_m)}} \sqrt{\log \mathcal{N}(\mathcal{G},L^2(\Lambda_m),r)}\mathrm{d}r,$$

$$|\mathcal{G}|_{L^2(\Lambda_m)} \le 2,$$

$$\log\left[\mathcal{N}\left(\mathcal{G},L^2(\Lambda_m),r\right)\right] \le d\log\left(\frac{2|\mathcal{S}|A}{r}+1\right)$$

results

Step-5: bound on $\mathcal{R}$

Combining the obtained

$$\mathcal{R}\left(\mathcal{G},\boldsymbol{\omega}_{1:m}\right) \le \frac{8\sqrt{2}}{\sqrt{m}} \int_0^{|\mathcal{G}|_{L^2(\Lambda_m)}} \sqrt{\log \mathcal{N}(\mathcal{G}, L^2(\Lambda_m), r)}\mathrm{d}r,$$

$$|\mathcal{G}|_{L^2(\Lambda_m)} \le 2,$$

$$\log\left[\mathcal{N}\left(\mathcal{G}, L^2(\Lambda_m), r\right)\right] \le d \log\left(\frac{2|\mathcal{S}|A}{r} + 1\right)$$

results, we have ($r \le 2$)

$$\mathcal{R}\left(\mathcal{G},\boldsymbol{\omega}_{1:m}\right) \le \frac{8\sqrt{2d}}{\sqrt{m}} \int_0^2 \sqrt{\log\left(\frac{2|\mathcal{S}|A+2}{r}\right)}\,\mathrm{d}r.$$

Step-5: bound on $\mathcal{R}$

Using $|\mathcal{S}|A + 1 \le (|\mathcal{S}| + 1)(A + 1)$

$$\begin{aligned}
\mathcal{R}\left(\mathcal{G}, \omega_{1:m}\right) &\le \frac{8\sqrt{2d}}{\sqrt{m}} \int_0^2 \sqrt{\log\left(\frac{2|\mathcal{S}|A + 2}{r}\right)} \mathrm{d}r \\
&\le \frac{8\sqrt{2d}}{\sqrt{m}} \left[\int_0^2 \sqrt{\log \frac{2\left(|\mathcal{S}| + 1\right)}{r}} \mathrm{d}r + 2\sqrt{\log(A+1)}\right] \\
&= \frac{16\sqrt{2d}}{\sqrt{m}} \left[\int_0^1 \sqrt{\log \frac{|\mathcal{S}| + 1}{r}} \mathrm{d}r + \sqrt{\log(A+1)}\right].
\end{aligned}$$

Applying $\int_0^1 \sqrt{\log \frac{a}{r}}\, dr \le \sqrt{\log a} + \frac{1}{2\sqrt{\log a}}$ $(a > 1)$

Step-5: bound on $\mathcal{R}$

we get

$$\mathcal{R}\left(\mathcal{G}, \boldsymbol{\omega}_{1:m}\right) \leq \tag{1}$$
$$\frac{16\sqrt{2d}}{\sqrt{m}}\left[\sqrt{\log(|\mathcal{S}|+1)} + \frac{1}{2\sqrt{\log(|\mathcal{S}|+1)}} + \sqrt{\log(A+1)}\right].$$

Step-5: bound on $\mathcal{R}$

we get

$$\mathcal{R}\left(\mathcal{G}, \boldsymbol{\omega}_{1:m}\right) \leq \frac{16\sqrt{2d}}{\sqrt{m}}\left[\sqrt{\log(|\mathcal{S}|+1)} + \frac{1}{2\sqrt{\log(|\mathcal{S}|+1)}} + \sqrt{\log(A+1)}\right]. \tag{1}$$

By the Jensen inequality

Step-5: bound on $\mathcal{R}$

we get

$$\mathcal{R}\left(\mathcal{G}, \boldsymbol{\omega}_{1:m}\right) \leq \tag{1}$$
$$\frac{16\sqrt{2d}}{\sqrt{m}}\left[\sqrt{\log(|\mathcal{S}|+1)} + \frac{1}{2\sqrt{\log(|\mathcal{S}|+1)}} + \sqrt{\log(A+1)}\right].$$

By the Jensen inequality

$$\mathbb{E}_{\boldsymbol{\omega}_{1:m}}\sqrt{\log(A+1)} \leq \sqrt{\mathbb{E}_{\boldsymbol{\omega}_{1:m}}\log(A+1)} \leq \sqrt{\log(\mathbb{E}_{\boldsymbol{\omega}_{1:m}}A+1)},$$

Step-5: bound on $\mathcal{R}$

we get

$$\mathcal{R}\left(\mathcal{G}, \boldsymbol{\omega}_{1:m}\right) \leq \tag{1}$$
$$\frac{16\sqrt{2d}}{\sqrt{m}}\left[\sqrt{\log(|\mathcal{S}|+1)} + \frac{1}{2\sqrt{\log(|\mathcal{S}|+1)}} + \sqrt{\log(A+1)}\right].$$

By the Jensen inequality

$$\mathbb{E}_{\boldsymbol{\omega}_{1:m}}\sqrt{\log(A+1)} \leq \sqrt{\mathbb{E}_{\boldsymbol{\omega}_{1:m}}\log(A+1)} \leq \sqrt{\log(\mathbb{E}_{\boldsymbol{\omega}_{1:m}}A+1)},$$
$$\mathbb{E}_{\boldsymbol{\omega}_{1:m}}A \leq \sqrt{\frac{1}{m}\sum_{j=1}^{m}\mathbb{E}_{\boldsymbol{\omega}_j}\left[\|\boldsymbol{\omega}_j\|_2^2\right]} =: \sigma. \ \Rightarrow$$

Step-5: bound on $\mathcal{R}$

we get

$$\mathcal{R}(\mathcal{G}, \boldsymbol{\omega}_{1:m}) \leq \tag{1}$$
$$\frac{16\sqrt{2d}}{\sqrt{m}}\left[\sqrt{\log(|\mathcal{S}|+1)} + \frac{1}{2\sqrt{\log(|\mathcal{S}|+1)}} + \sqrt{\log(A+1)}\right].$$

By the Jensen inequality

$$\mathbb{E}_{\boldsymbol{\omega}_{1:m}}\sqrt{\log(A+1)} \leq \sqrt{\mathbb{E}_{\boldsymbol{\omega}_{1:m}}\log(A+1)} \leq \sqrt{\log(\mathbb{E}_{\boldsymbol{\omega}_{1:m}}A+1)},$$

$$\mathbb{E}_{\boldsymbol{\omega}_{1:m}}A \leq \sqrt{\frac{1}{m}\sum_{j=1}^{m}\mathbb{E}_{\boldsymbol{\omega}_j}\left[\|\boldsymbol{\omega}_j\|_2^2\right]} =: \sigma. \;\Rightarrow$$

$\mathbb{E}_{\boldsymbol{\omega}_{1:m}}\mathcal{R}(\mathcal{G}, \boldsymbol{\omega}_{1:m}) \leq$ (1), but with $A \to \sigma$.

Step-6: putting together

Result: k continuous, shift-invariant kernel; for any $\tau > 0$, $\mathcal{S} \neq \emptyset$ compact set,

$$\Lambda^m \left(\sup_{\mathbf{x},\mathbf{y} \in \mathcal{S}} |\hat{k}(\mathbf{x},\mathbf{y}) - k(\mathbf{x},\mathbf{y})| \geq \frac{h(d,|\mathcal{S}|,\sigma) + \sqrt{2\tau}}{\sqrt{m}} \right) \leq e^{-\tau},$$

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

Step-6: putting together

Result: k continuous, shift-invariant kernel; for any $\tau > 0$, $\mathcal{S} \neq \emptyset$ compact set,

$$\Lambda^m \left(\sup_{\mathbf{x},\mathbf{y} \in \mathcal{S}} |\hat{k}(\mathbf{x},\mathbf{y}) - k(\mathbf{x},\mathbf{y})| \geq \underbrace{\frac{h(d,|\mathcal{S}|,\sigma) + \sqrt{2\tau}}{\sqrt{m}}}_{:=\epsilon} \right) \leq e^{-\tau},$$

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

Equivalently

$$\Lambda^m \left(\left\| \hat{k} - k \right\|_{\mathcal{S}} \geq \epsilon \right) \leq e^{-\frac{\left[\epsilon\sqrt{m} - h(d,|\mathcal{S}|,\sigma)\right]^2}{2}}.$$

Discussion (Borel-Cantelli lemma)

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

Discussion (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., $|\mathcal{S}_m| = e^{o(m)}$) $\leftrightarrow$
  - Old: $|\mathcal{S}_m| = o\left(\sqrt{m/\log m}\right)$.

Discussion (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., $|\mathcal{S}_m| = e^{o(m)}$) $\leftrightarrow$
  - Old: $|\mathcal{S}_m| = o\left(\sqrt{m/\log m}\right)$.
- Specifically:
  - *asymptotically* optimal result [Csörgő and Totik, 1983, Theorem 2] (if ψ vanishes at ∞),
  - at faster rate $\Rightarrow$ even conv. in prob. would fail.

Direct consequence: L^r guarantee $(1 < r)$

Idea:

- Note that

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

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

L^r large deviation inequality

Under the previous assumptions:

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

In other words,

$$\|\hat{k}-k\|_{L^r(\mathcal{S})} = O_{a.s.}\left(m^{-1/2}|\mathcal{S}|^{2d/r}\sqrt{\log|\mathcal{S}|}\right).$$

For $2 \leq r$: direct L^r proof $\Rightarrow$ $\sqrt{\log(|\mathcal{S}|)}$ factor can be discarded.

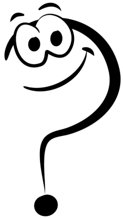
Kernel derivatives

- If $supp(\Lambda)$ is bounded
  - k-proof can be extended (L^r as well), but
  - Gaussian kernel:(
- [Rahimi and Recht, 2007]'s proof:
  - Hoeffding inequality (boundedness!) + Lipschitzness,
- Bernstein + Lipschitzness: handles $\partial^{\mathbf{p},\mathbf{q}}k$ with
  - moment constraints on Λ (example: Gaussian kernel).
  - slightly worse rates.

Conclusion

- Kernel + derivative approximations.
- Performance: uniform, L^r.
- Detailed finite-sample analysis, optimal rates.
- Paper (submitted to NIPS):
  - RFF: `http://arxiv.org/abs/1506.02155`,
  - infD exp. fitting: `http://arxiv.org/abs/1506.02564`.

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 Sci. Math. (Szeged), 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.

Steinwart, I. and Christmann, A. (2008).
Support Vector Machines.
Springer.

Sutherland, D. and Schneider, J. (2015).
On the error of random fourier features.
In *Conference on Uncertainty in Artificial Intelligience (UAI)*.

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.

Support of a measure

- Ingredients:
  - (X, τ): topological space with a countable basis.
  - $\mathcal{B} = \sigma(\tau)$: sigma-algebra generated by τ.
  - Λ: measure on $(X, \mathcal{B})$.

  Then

$$supp(\Lambda) = \overline{\cup\{A \in \tau : \Lambda(A) = 0\}},$$

  i.e., the complement of the union of all open Λ-null sets.
- Our choice: $X = \mathbb{R}^d$.