Random Fourier Features: Optimal Uniform Bounds

Zoltán Szabó*

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

'Statistics with coffee' seminar
École Polytechnique
October 5, 2016

Outline

- Kernel.
- Random Fourier features (RFFs).
- Optimal uniform guarantee on RFF approximation.

Kernel

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,
  - $k(a, b) = e^{-\theta \|a-b\|_2^2}$: 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,
  - $k(a,b) = e^{-\theta \|a-b\|_2^2}$: Gaussian,
  - $k(a,b) = e^{-\theta \|a-b\|_2}$: Laplacian.
- In the $H = H(k)$ RKHS ($\exists!$): $\varphi(b) = k(\cdot, b)$.

RKHS: evaluation point of view

- Let $H \subset \mathbb{R}^{\mathcal{X}}$ be a Hilbert space.
- Consider for fixed $x \in \mathcal{X}$ the $\delta_x : f \in H \mapsto f(x) \in \mathbb{R}$ map.
- The evaluation functional is linear:

$$\delta_x(\alpha f + \beta g) = \alpha \delta_x(f) + \beta \delta_x(g).$$

RKHS: evaluation point of view

- Let $H \subset \mathbb{R}^{\mathcal{X}}$ be a Hilbert space.
- Consider for fixed $x \in \mathcal{X}$ the $\delta_x : f \in H \mapsto f(x) \in \mathbb{R}$ map.
- The evaluation functional is linear:

$$\delta_x(\alpha f + \beta g) = \alpha \delta_x(f) + \beta \delta_x(g).$$

- Def.: H is called *RKHS* if δ_x is continuous for $\forall x \in \mathcal{X}$.

RKHS: reproducing point of view

- Let $H \subset \mathbb{R}^{\mathcal{X}}$ be a Hilbert space.
- $k : \mathcal{X} \times \mathcal{X} \to$ is called a *reproducing kernel of H* if for $\forall x \in \mathcal{X}, f \in H$
  1. $k(\cdot, x) \in H$,
  2. $\langle f, k(\cdot, x) \rangle_H = f(x)$ (reproducing property).

RKHS: reproducing point of view

- Let $H \subset \mathbb{R}^{\mathcal{X}}$ be a Hilbert space.
- $k : \mathcal{X} \times \mathcal{X} \to$ is called a *reproducing kernel of H* if for $\forall x \in \mathcal{X}, f \in H$
  1. $k(\cdot, x) \in H$,
  2. $\langle f, k(\cdot, x)\rangle_H = f(x)$ (reproducing property).

  Specifically, $\forall x, y \in \mathcal{X}$

$$k(x, y) = \langle k(\cdot, x), k(\cdot, y)\rangle_H .$$

RKHS: positive-definite point of view

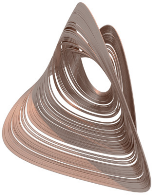
- Let us given a $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ symmetric function.
- k is called *positive definite* if $\forall n \geq 1$, $\forall (a_1, \ldots, a_n) \in \mathbb{R}^n$, $(x_1, \ldots, x_n) \in \mathcal{X}^n$

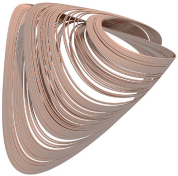
$$\sum_{i,j=1}^{n} a_i a_j k(x_i, x_j) = \mathbf{a}^T \mathbf{G} \mathbf{a} \geq 0,$$

where $\mathbf{G} = [k(x_i, x_j)]_{i,j=1}^{n}$.

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:

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

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 $\rightarrow$ RFF.

Random Fourier features

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.

Uniform finite-sample bound for RFFs

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. Dissemination: NIPS-2015 [spotlight - 3.65%],
   - H. Strathmann, D. Sejdinovic, S. Livingston, Z. Szabó, A. Gretton. Gradient-free Hamiltonian Monte Carlo with Efficient Kernel Exponential Families. In NIPS-2015, pages 955-963.
   - W. Jitkrittum, A. Gretton, N. Heess, A. Eslami, B. Lakshminarayanan, D. Sejdinovic, Z. Szabó. Kernel-Based Just-In-Time Learning for Passing Expectation Propagation Messages. In UAI-2015, pages 405-414.

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

1. Empirical process form [$\mathbb{P}g := \int g \mathrm{d}\mathbb{P}$; $g(\omega) = \cos\left(\omega^T(\mathbf{x}-\mathbf{y})\right)$]:

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

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

1. Empirical process form [$\mathbb{P}g := \int g \mathrm{d}\mathbb{P}$; $g(\boldsymbol{\omega}) = \cos\left(\boldsymbol{\omega}^T(\mathbf{x} - \mathbf{y})\right)$]:

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

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

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

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

1. Empirical process form [$\mathbb{P}g := \int g \mathrm{d}\mathbb{P}$; $g(\boldsymbol{\omega}) = \cos\left(\boldsymbol{\omega}^T(\mathbf{x} - \mathbf{y})\right)$]:

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

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

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

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

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

4. Using Dudley's entropy bound:

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

Proof idea

4. Using Dudley's entropy bound:

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

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

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

Proof idea

4. Using Dudley's entropy bound:

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

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

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

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 $\mathcal{S} \subset \mathbb{R}^d$

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

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

Open questions

- We also have finite-sample bounds in $L^p(\mathcal{S})$: optimal?
- Kernel derivatives:
  - Applications, e.g.
    1. fitting ∞-D exp. family distributions [Sriperumbudur et al., 2014],
    2. nonlinear variable selection [Rosasco et al., 2013].
  - Challenge: *non-uniformly* bounded functions.

Open questions-2

- FiniteD features with $\|\cdot\|_{L^\infty(\mathbb{R}^d)}$-guarantee?
- For operator-valued kernels
  - $k(x,y) \in \mathbb{R} \quad \leftrightarrow H(k) = \{f : \mathcal{X} \to \mathbb{R} \text{ functions}\}$
  - $k(x,y) \in \mathcal{L}(Z) \leftrightarrow H(k) = \{f : \mathcal{X} \to Z \text{ functions}\}$.

  RFFs exist [Brault et al., 2016, Minh, 2016], but with loose bounds.

Thank you for the attention!

Uniformly bounded, separable Carathéodory family

$\mathcal{G} = \{\boldsymbol{\omega} \mapsto \cos\left(\boldsymbol{\omega}^T \mathbf{z}\right) : \mathbf{z} = \mathbf{x} - \mathbf{y} \in \Delta_{\mathcal{S}} := \mathcal{S} - \mathcal{S}\}$ is a separable Carathéodory family, i.e.

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

Uniformly bounded, separable Carathéodory family

$\mathcal{G} = \{\boldsymbol{\omega} \mapsto \cos\left(\boldsymbol{\omega}^T \mathbf{z}\right) : \mathbf{z} = \mathbf{x} - \mathbf{y} \in \Delta_{\mathcal{S}} := \mathcal{S} - \mathcal{S}\}$ is a separable Carathéodory family, i.e.

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

Uniformly bounded, separable Carathéodory family

$\mathcal{G} = \{\boldsymbol{\omega} \mapsto \cos\left(\boldsymbol{\omega}^T \mathbf{z}\right) : \mathbf{z} = \mathbf{x} - \mathbf{y} \in \Delta_{\mathcal{S}} := \mathcal{S} - \mathcal{S}\}$ is a separable Carathéodory family, i.e.

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

and $\mathcal{G}$ is uniformly bounded ($\sup_{g \in \mathcal{G}} \|g\|_\infty \le 1 < \infty$).

Infinite-dimensional exponential family

- Exponential family:

$$p_{\boldsymbol{\theta}}(\mathbf{x}) \propto e^{\langle \boldsymbol{\theta}, T(\mathbf{x}) \rangle},$$

  where $\boldsymbol{\theta}$: natural parameter, $T(\mathbf{x})$: sufficient statistics.
- InfiniteD generalization:

$$p_f(\mathbf{x}) \propto e^{f(\mathbf{x})} = e^{\langle f, k(\cdot,\mathbf{x}) \rangle_{H(k)}}.$$

Infinite-dimensional exponential family

- Exponential family:

$$p_{\boldsymbol{\theta}}(\mathbf{x}) \propto e^{\langle \boldsymbol{\theta}, T(\mathbf{x}) \rangle},$$

where $\boldsymbol{\theta}$: natural parameter, $T(\mathbf{x})$: sufficient statistics.

- InfiniteD generalization:

$$p_f(\mathbf{x}) \propto e^{f(\mathbf{x})} = e^{\langle f, k(\cdot,\mathbf{x}) \rangle_{H(k)}}.$$

Fitting idea (score matching, Fischer divergence):

$$J(p_*, p_f) := \int p^*(\mathbf{x}) \left\| \frac{\partial p_*(\mathbf{x})}{\partial \mathbf{x}} - \frac{\partial p_f(\mathbf{x})}{\partial \mathbf{x}} \right\|_2^2 \mathrm{d}\mathbf{x} \to \min_f .$$

Brault, R., d'Alché Buc, F., and Heinonen, M. (2016).
Random Fourier features for operator-valued kernels.
Technical report.
https://arxiv.org/abs/1605.02536.

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

Minh, H. Q. (2016).
Operator-valued Bochner theorem, Fourier feature maps for operator-valued kernels, and vector-valued learning.
Technical report, Istituto Italiano di Tecnologia.
http://128.84.21.199/abs/1608.05639v1.

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

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

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.