Optimal Rates for Random Fourier Features*

Bharath K. Sriperumbudur†,1 Zoltán Szabó†,2

[1]Department of Statistics, Pennsylvania State University [2]Gatsby Unit, University College London

Brief Summary

- **Kernel methods** [1]:
  - **Pro**: flexible modelling toolkit.
  - **Contra**: computationally intensive, poor scalability.
- **Randomized algorithms**:
  - Low-D feature representation → **fast** linear methods.
  - Random Fourier features (RFF) [2]:
    - simple, popular, practically efficient, **but** theoretically not well-understood.
- **Contribution**: detailed theoretical analysis,
  - L^∞-optimal performance guarantees (RFF dimension, growing set size),
  - L^r $(1 \le r < \infty)$-guarantees,
  - RFF approximation for kernel derivatives + analysis.

RFF Idea (Kernel Approximation)

- $k : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$: continuous, bounded, translation-invariant kernel. Bochner's theorem $\Rightarrow$

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

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

Here: $(\boldsymbol{\omega}_j)_{j=1}^m \overset{i.i.d.}{\sim} \Lambda$, $\hat{k}(\mathbf{x},\mathbf{y}) = \langle \phi(\mathbf{x}), \phi(\mathbf{y})\rangle_{\mathbb{R}^{2m}}$ with

$$\phi(\mathbf{x}) = \frac{1}{\sqrt{m}}(\cos(\boldsymbol{\omega}_1^T\mathbf{x}),\ldots,\cos(\boldsymbol{\omega}_m^T\mathbf{x}),\sin(\boldsymbol{\omega}_1^T\mathbf{x}),\ldots,\sin(\boldsymbol{\omega}_m^T\mathbf{x})).$$

Existing RFF Guarantees

- [2]: $\hat{k}$ is consistent (compact convergence),

$$\|k-\hat{k}\|_{L^\infty(\mathcal{S}\times\mathcal{S})} := \sup_{(x,y)\in\mathcal{S}\times\mathcal{S}} |k(x,y) - \hat{k}(x,y)| = \mathcal{O}_p\left(|\mathcal{S}|\sqrt{m^{-1}\log m}\right).$$

- [3]: 3 RFF variants, better constants.

Theorem-1: k approximation, $L^\infty(\mathcal{S}\times\mathcal{S})$

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

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

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

Remark-1:

- A.s. convergence on compact sets: $\hat{k} \xrightarrow{m\to\infty} k$ at rate $\sqrt{\frac{\log|\mathcal{S}|}{m}}$ (B-C. lemma).
- Growing diameter $(\mathcal{S}_m)$:
  - $\frac{\log|\mathcal{S}_m|}{m} \xrightarrow{m\to\infty} 0$ is enough (i.e., $|\mathcal{S}_m| = e^{o(m)}$) $\leftrightarrow$ old result: $|\mathcal{S}_m| = o\left(\sqrt{m/\log(m)}\right)$.
- Specifically: *asymptotic* optimality [4, Theorem 2].

Theorem-2: k approximation, $L^r(\mathcal{S}\times\mathcal{S})$, $1 \le r < \infty$

For any $\tau > 0$, compact $\mathcal{S} \subset \mathbb{R}^d$

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

Remark-2:

- Consequence of Theorem-1.
- $L^r(\mathcal{S}\times\mathcal{S})$-consistency: $\|k-\hat{k}\|_{L^r(\mathcal{S}\times\mathcal{S})} = \mathcal{O}_{a.s.}\Big(\underbrace{m^{-1/2}|\mathcal{S}|^{2d/r}\sqrt{\log|\mathcal{S}|}}_{=:(*)_1;\ \text{if} \xrightarrow{m\to\infty} 0}\Big)$, i.e.
- Growing diameter: $(*)_1 \to 0 \Rightarrow |\mathcal{S}_m| = \tilde{o}(m^{\frac{r}{4d}})$; L^∞-case: $|\mathcal{S}_m| = e^{m^{\delta<1}}$.

Theorem-3: k approximation, $L^r(\mathcal{S}\times\mathcal{S})$, $1 < r < \infty$

Applying a direct reasoning: for any $\tau > 0$, compact $\mathcal{S} \subset \mathbb{R}^d$

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

Remark-3:

- $C_r = \mathcal{O}(\sqrt{r})$, universal constant.
- $L^r(\mathcal{S}\times\mathcal{S})$-consistency: if $2 \le r$, then $\|k-\hat{k}\|_{L^r(\mathcal{S}\times\mathcal{S})} = \mathcal{O}_{a.s.}\Big(\underbrace{m^{-1/2}|\mathcal{S}|^{2d/r}}_{\text{if} \xrightarrow{m\to\infty} 0}\Big)$,

  the $\sqrt{\log(\mathcal{S})}$ term disappeared ($\tilde{o} \to o$, see Remark-2).

Kernel Derivative Approximation

$$\partial^{\mathbf{p},\mathbf{q}}k(\mathbf{x},\mathbf{y}) = \int_{\mathbb{R}^d} \boldsymbol{\omega}^{\mathbf{p}}(-\boldsymbol{\omega})^{\mathbf{q}} h_{|\mathbf{p}+\mathbf{q}|}\left(\boldsymbol{\omega}^T(\mathbf{x}-\mathbf{y})\right)\mathrm{d}\Lambda(\boldsymbol{\omega}),\ \ h_n = \cos^{(n)}, n \in \mathbb{N}$$

$$\widehat{\partial^{\mathbf{p},\mathbf{q}}k}(\mathbf{x},\mathbf{y}) = \frac{1}{m}\sum_{j=1}^m \boldsymbol{\omega}_j^{\mathbf{p}}(-\boldsymbol{\omega}_j)^{\mathbf{q}} h_{|\mathbf{p}+\mathbf{q}|}\left(\boldsymbol{\omega}_j^T(\mathbf{x}-\mathbf{y})\right) = \langle \phi^{\mathbf{p}}(\mathbf{x}), \phi^{\mathbf{q}}(\mathbf{y})\rangle_{\mathbb{R}^{2m}}.$$

Th.-4: $\partial^{\mathbf{p},\mathbf{q}}k(\mathbf{x},\mathbf{y})$ appr., $supp(\Lambda)$: bounded, $L^\infty(\mathcal{S}\times\mathcal{S})$

Let $\mathbf{p},\mathbf{q} \in \mathbb{N}^d$, $T_{\mathbf{p},\mathbf{q}} := \sup_{\boldsymbol{\omega}\in supp(\Lambda)}|\boldsymbol{\omega}^{\mathbf{p}+\mathbf{q}}|$, $C_{\mathbf{p},\mathbf{q}} := \mathbb{E}_{\boldsymbol{\omega}\sim\Lambda}\left[|\boldsymbol{\omega}^{\mathbf{p}+\mathbf{q}}|\,\|\boldsymbol{\omega}\|_2^2\right]$. Assume: $C_{2\mathbf{p},2\mathbf{q}} < \infty$; $supp(\Lambda)$ is bounded if $[\mathbf{p};\mathbf{q}] \neq \mathbf{0}$. Then for $\forall\tau > 0$, compact $\mathcal{S} \subset \mathbb{R}^d$

$$\Lambda^m\left(\|\partial^{\mathbf{p},\mathbf{q}}k - \widehat{\partial^{\mathbf{p},\mathbf{q}}k}(\mathbf{x},\mathbf{y})\|_{L^\infty(\mathcal{S}\times\mathcal{S})} \ge \frac{H(d,\mathbf{p},\mathbf{q},|\mathcal{S}|)+T_{\mathbf{p},\mathbf{q}}\sqrt{2\tau}}{\sqrt{m}}\right) \le e^{-\tau},$$

where

$$\frac{H(d,\mathbf{p},\mathbf{q},|\mathcal{S}|)}{32\sqrt{2d\,T_{2\mathbf{p},2\mathbf{q}}}} = \left[\sqrt{U(\mathbf{p},\mathbf{q},|\mathcal{S}|)} + \frac{1}{2\sqrt{U(\mathbf{p},\mathbf{q},|\mathcal{S}|)}} + \sqrt{\log(\sqrt{C_{2\mathbf{p},2\mathbf{q}}}+1)}\right],$$

$$U(\mathbf{p},\mathbf{q},|\mathcal{S}|) = \log\left(\frac{2|\mathcal{S}|}{\sqrt{T_{2\mathbf{p},2\mathbf{q}}}}+1\right).$$

Remark-4:

- Theorem-4 $\xrightarrow{\text{spec. } \mathbf{p}=\mathbf{q}=0}$ Theorem-1, $T_{\mathbf{p},\mathbf{q}} = T_{2\mathbf{p},2\mathbf{q}} = 1$. Else: $supp(\Lambda)$: bounded $\Rightarrow T_{\mathbf{p},\mathbf{q}} < \infty$ and $T_{2\mathbf{p},2\mathbf{q}} < \infty$.
- Growth of $|\mathcal{S}_m|$: A la Remarks 1-2
  - $\|\partial^{\mathbf{p},\mathbf{q}}k - \widehat{\partial^{\mathbf{p},\mathbf{q}}k}(\mathbf{x},\mathbf{y})\|_{L^\infty(\mathcal{S}_m\times\mathcal{S}_m)} \xrightarrow{a.s.} 0$ if $|\mathcal{S}_m| = e^{o(m)}$.
  - $\|\partial^{\mathbf{p},\mathbf{q}}k - \widehat{\partial^{\mathbf{p},\mathbf{q}}k}(\mathbf{x},\mathbf{y})\|_{L^r(\mathcal{S}_m\times\mathcal{S}_m)} \xrightarrow{a.s.} 0$ if $m^{-1/2}|\mathcal{S}_m|^{2d/r}\sqrt{\log|\mathcal{S}_m|} \xrightarrow{m\to\infty} 0$ $(1 \le r < \infty)$.

Theorem-5: $\partial^{\mathbf{p},\mathbf{q}}k(\mathbf{x},\mathbf{y})$ approximation, $supp(\Lambda)$: unbounded, $L^\infty(\mathcal{S}\times\mathcal{S})$

Assume: (i) $\mathbf{z} \mapsto \nabla_{\mathbf{z}}[\partial^{\mathbf{p},\mathbf{q}}k(\mathbf{z})]$: continuous; (ii) $\mathcal{S} \subset \mathbb{R}^d$: compact, (iii) $E_{\mathbf{p},\mathbf{q}} := \mathbb{E}_{\boldsymbol{\omega}\sim\Lambda}|\boldsymbol{\omega}^{\mathbf{p}+\mathbf{q}}|\,\|\boldsymbol{\omega}\|_2 < \infty$, (iv) $\exists L > 0, \sigma > 0$ such that with $\mathcal{S}_\Delta := \mathcal{S} - \mathcal{S}$

$$\mathbb{E}_{\boldsymbol{\omega}\sim\Lambda}|f(\mathbf{z};\boldsymbol{\omega})|^M \overset{(*)_2}{\le} \frac{M!\,\sigma^2 L^{M-2}}{2} \quad (\forall M \ge 2, \forall\mathbf{z} \in \mathcal{S}_\Delta),$$

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

Then with $F_d := d^{-\frac{d}{d+1}} + d^{\frac{1}{d+1}}$, $D_{\mathbf{p},\mathbf{q},\mathcal{S}} := \sup_{\mathbf{z}\in conv(\mathcal{S}_\Delta)}\|\nabla_{\mathbf{z}}[\partial^{\mathbf{p},\mathbf{q}}k(\mathbf{z})]\|_2$

$$\Lambda^m\left(\|\partial^{\mathbf{p},\mathbf{q}}k - \widehat{\partial^{\mathbf{p},\mathbf{q}}k}(\mathbf{x},\mathbf{y})\|_{L^\infty(\mathcal{S}\times\mathcal{S})} \ge \epsilon\right) \le$$

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

Remark-5:

- F_d: monotonically decreasing in d, $F_1 = 2$.
- $(*)_2$ holds if $|f(\mathbf{z};\boldsymbol{\omega})| \le \frac{L}{2}$ and $\mathbb{E}_{\boldsymbol{\omega}\sim\Lambda}[|f(\mathbf{z};\boldsymbol{\omega})|^2] \le \sigma^2$ $(\forall\mathbf{z} \in \mathcal{S}_\Delta)$.
- $\|\partial^{\mathbf{p},\mathbf{q}}k - \widehat{\partial^{\mathbf{p},\mathbf{q}}k}(\mathbf{x},\mathbf{y})\|_{L^\infty(\mathcal{S}\times\mathcal{S})} = \mathcal{O}_{a.s.}\left(|\mathcal{S}|\sqrt{\frac{\log m}{m}}\right)$:
  - slightly worse than Theorem-4, but it handles unbounded functions.

Future Research Directions

(i) Kernel derivatives: tighter guarantees, (ii) prediction using kernel (derivative) estimates, (iii) analysis of smart RFF approximations [5].

References

[1] Bernhard Schölkopf and Alexander J. Smola. *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond.* MIT Press, 2002.

[2] Ali Rahimi and Benjamin Recht. Random features for large-scale kernel machines. In *Neural Information Processing Systems (NIPS)*, pages 1177–1184, 2007.

[3] Dougal J. Sutherland and Jeff Schneider. On the error of random Fourier features. In *Conference on Uncertainty in Artificial Intelligence (UAI)*, pages 862–871, 2015.

[4] Sándor Csörgő and Vilmos Totik. On how long interval is the empirical characteristic function uniformly consistent? *Acta Scientiarum Mathematicarum*, 45:141–149, 1983.

[5] Quoc Le, Tamás Sarlós, and Alexander Smola. Fastfood - computing Hilbert space expansions in loglinear time. *JMLR W&CP - International Conference on Machine Learning (ICML)*, 28:244–252, 2013.

†Contributed equally. *NIPS-2015, Montréal, Canada, December 7-12, 2015. Acknowledgments: This work was supported by the Gatsby Charitable Foundation.