Random Fourier Features on Kernel Derivatives*

Zoltán Szabó[1], Bharath K. Sriperumbudur[2]

[1]CMAP, École Polytechnique [2]Department of Statistics, Pennsylvania State University

Quick Summary

- Context: supervised learning with RKHSs (kernels).
- Scalability: random Fourier features (RFF) [2]
  - one of the most influential approach,
  - 10-year test-of-time award at NIPS-2017.
- Our motivation: tasks with function derivatives,

$$\min_{f\in\mathcal{H}_k} J(f) = \frac{1}{n}\sum_{i=1}^n V_i\left(y_i, \{\partial^{\mathbf{p}} f(\mathbf{x}_i)\}_{\mathbf{p}\in I_i}\right) + \lambda \|f\|^2_{\mathcal{H}_k}.$$

  Representer theorem $\Rightarrow$ appr. of kernel derivatives.
- Goal: tight approximation guarantees on

$$\left\|\partial^{\mathbf{p},\mathbf{q}}k - \widehat{\partial^{\mathbf{p},\mathbf{q}}k}\right\|_{L^\infty(\mathcal{S}\times\mathcal{S})}.$$

Supervised Learning with Derivatives

- Task: given
  - samples $\{(x_i,y_i)\}_{i=1}^n \subset \mathcal{X}\times\mathbb{R}$,
  - $\mathcal{H}_k \subset \mathbb{R}^{\mathcal{X}}$ RKHS associated to kernel $k:\mathcal{X}\times\mathcal{X}\to\mathbb{R}$.
- Objective:

$$\min_{f\in\mathcal{H}_k} J(f) := \frac{1}{n}\sum_{i=1}^n V_i\left(y_i, \{\partial^{\mathbf{p}} f(\mathbf{x}_i)\}_{\mathbf{p}\in I_i}\right) + \lambda \|f\|^2_{\mathcal{H}_k}.$$

- Examples ($\mathbf{p}=\mathbf{q}=\mathbf{0}$):
  - kernel ridge regression (squared loss): $V(f(x_i),y_i) = [f(x_i)-y_i]^2$,
  - soft-classification (hinge loss): $V(f(x_i),y_i) = \max(1-y_i f(x_i), 0)$.
- Examples ($[\mathbf{p};\mathbf{q}]\neq 0$):
  - nonlinear variable selection,
  - semi-supervised or Hermite learning with gradient information,
  - fitting infinite-dimensional exponential families.
- Representer theorem [7]:

$$f(\cdot) = \sum_{j=1}^n \sum_{\mathbf{p}\in I_j} c_{j,\mathbf{p}} \partial^{\mathbf{p},\mathbf{0}} k(\cdot,\mathbf{x}_j), \quad (c_{j,\mathbf{p}}\in\mathbb{R}), \text{ and}$$

$$\min_{\mathbf{c}} \tilde{J}(\mathbf{c}) = \frac{1}{n}\sum_{i=1}^n V_i\left(y_i, \left\{\sum_{j=1}^n\sum_{\mathbf{p}\in I_j} c_{j,\mathbf{p}}\partial^{\mathbf{p},\mathbf{0}}k(\mathbf{x}_i,\mathbf{x}_j)\right\}_{\mathbf{p}\in I_i}\right) + \lambda\sum_{i=1}^n\sum_{\mathbf{p}\in I_i}\sum_{j=1}^n\sum_{\mathbf{q}\in I_j} c_{i,\mathbf{p}}c_{j,\mathbf{q}}\partial^{\mathbf{p},\mathbf{q}}k(\mathbf{x}_i,\mathbf{x}_j),$$

where $\mathbf{c} = (c_{i,\mathbf{p}})_{i\in\{1,\ldots,n\},\mathbf{p}\in I_i} \in \mathbb{R}^{\sum_{i=1}^n |I_i|}$.

Random Fourier Features

- $k:\mathbb{R}^d\times\mathbb{R}^d\to\mathbb{R}$ continuous bounded shift-invariant.
- By the Bochner theorem:

$$\begin{aligned}
k(\mathbf{x},\mathbf{y}) &= \int_{\mathbb{R}^d}\cos\left(\boldsymbol{\omega}^T(\mathbf{x}-\mathbf{y})\right)\mathrm{d}\Lambda(\boldsymbol{\omega})\\
&= \int_{\mathbb{R}^d}\langle\phi_{\boldsymbol{\omega}}(\mathbf{x}),\phi_{\boldsymbol{\omega}}(\mathbf{y})\rangle_{\mathbb{R}^2}\mathrm{d}\Lambda(\boldsymbol{\omega}), \text{ where}\\
\phi_{\boldsymbol{\omega}}(\mathbf{x}) &= \left[\cos\left(\boldsymbol{\omega}^T\mathbf{x}\right);\sin\left(\boldsymbol{\omega}^T\mathbf{x}\right)\right],\\
\partial^{\mathbf{p},\mathbf{q}}k(\mathbf{x},\mathbf{y}) &= \int_{\mathbb{R}^d}\langle\partial^{\mathbf{p}}\phi_{\boldsymbol{\omega}}(\mathbf{x}),\partial^{\mathbf{q}}\phi_{\boldsymbol{\omega}}(\mathbf{y})\rangle_{\mathbb{R}^2}\mathrm{d}\Lambda(\boldsymbol{\omega}).
\end{aligned}$$

- RFF idea [2] ($[\mathbf{p};\mathbf{q}]=\mathbf{0}$): change Λ to Λ_m

$$\begin{aligned}
\hat{k}(\mathbf{x},\mathbf{y}) &= \int_{\mathbb{R}^d}\langle\phi_{\boldsymbol{\omega}}(\mathbf{x}),\phi_{\boldsymbol{\omega}}(\mathbf{y})\rangle_{\mathbb{R}^2}\mathrm{d}\Lambda_m(\boldsymbol{\omega}),\\
\widehat{\partial^{\mathbf{p},\mathbf{q}}k}(\mathbf{x},\mathbf{y}) &= \int_{\mathbb{R}^d}\langle\partial^{\mathbf{p}}\phi_{\boldsymbol{\omega}}(\mathbf{x}),\partial^{\mathbf{q}}\phi_{\boldsymbol{\omega}}(\mathbf{y})\rangle_{\mathbb{R}^2}\mathrm{d}\Lambda_m(\boldsymbol{\omega}).
\end{aligned}$$

Related Work

- Kernel values [5]:

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

- Kernel ridge regression [3, 1], kernel PCA [4, 6].

Our Result

Finite sample guarantee for kernels satisfying the Bernstein condition: $\exists K\geq 1$

$$\int_{\mathbb{R}^d}\frac{|\boldsymbol{\omega}^{\mathbf{p}+\mathbf{q}}|^n}{(\sigma_{\mathbf{p},\mathbf{q}})^n}\mathrm{d}\Lambda(\boldsymbol{\omega}) \leq \frac{n!}{2}K^{n-2}, \quad n=2,3,\ldots,$$

where $\sigma_{\mathbf{p},\mathbf{q}} = \sqrt{\int_{\mathbb{R}^d}|\boldsymbol{\omega}^{\mathbf{p}+\mathbf{q}}|^2\mathrm{d}\Lambda(\boldsymbol{\omega})}$. Specifically,

$$\boxed{\left\|\partial^{\mathbf{p},\mathbf{q}}k - \widehat{\partial^{\mathbf{p},\mathbf{q}}k}\right\|_{L^\infty(\mathcal{S}_m\times\mathcal{S}_m)} = O_{a.s.}\left(\sqrt{\log(|\mathcal{S}_m|)}/\sqrt{m}\right).}$$

Our result implies that

$$\left\|\partial^{\mathbf{p},\mathbf{q}}k - \widehat{\partial^{\mathbf{p},\mathbf{q}}k}\right\|_{L^\infty(\mathcal{S}_m\times\mathcal{S}_m)} \xrightarrow{a.s.} 0 \text{ if } |\mathcal{S}_m| = e^{o(m)}.$$

Bernstein requirement: For $f_\Lambda(\omega)\propto e^{-\omega^{2\ell}}$, $p+q\leq 2\ell$: ✓

Acknowledgements. This work was started and partially carried out while ZSz was visiting BKS at the Department of Statistics, Pennsylvania State University; ZSz thanks for their generous support. BKS is supported by NSF-DMS-1713011.

References

[1] Zhu Li, Jean-François Ton, Dino Oglic, and Dino Sejdinovic. A unified analysis of random Fourier features. Technical report, 2018. (https://arxiv.org/abs/1806.09178).

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

[3] Alessandro Rudi and Lorenzo Rosasco. Generalization properties of learning with random features. In *NIPS*, pages 3218–3228, 2017.

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

[5] Bharath K. Sriperumbudur and Zoltán Szabó. Optimal rates for random Fourier features. In *NIPS*, pages 1144–1152, 2015.

[6] Enayat Ullah, Poorya Mianjy, Teodor V. Marinov, and Raman Arora. Streaming kernel PCA with $\tilde{O}(\sqrt{n})$ random features. Technical report, 2018. (https://arxiv.org/abs/1808.00934).

[7] Ding-Xuan Zhou. Derivative reproducing properties for kernel methods in learning theory. *Journal of Computational and Applied Mathematics*, 220:456–463, 2008.

*Data Learning and Inference (DALI), George, South Africa, 3-5 January, 2018.