Convergence Theory for Vector-Valued Random Features

Nicholas H. Nelsen

Caltech

Email: nnelsen@caltech.edu
Web: nicholashnelsen.com

Supported By: Amazon, NSF, ONR, AFOSR, DOD

BIRS Scientific Machine Learning Workshop
Banff, Alberta, Canada

June 2023

N.H.N. and Andrew M. Stuart

The Random Feature Model for Input-Output Maps between Banach Spaces

SIAM Journal on Scientific Computing, Vol. 43, No. 5, pp. A3212–A3243, 2021.

Random Features

Random feature map ($\mathcal{X}$ and $\mathcal{Y}$ infinite-dimensional)

$\varphi\colon \mathcal{X} \times \Theta \to \mathcal{Y}$ and a probability measure μ

Parametric model (looks like Monte Carlo)

$$\Psi_{\mathrm{RFM}}(u;\alpha) \coloneqq \frac{1}{M}\sum_{m=1}^{M} \alpha_m \varphi(u;\theta_m), \quad \theta_m \overset{\mathrm{iid}}{\sim} \mu$$

Random Feature Ridge Regression

Data

$$u_n \overset{\text{iid}}{\sim} \nu \quad \text{and} \quad y_n = \Psi^\dagger(u_n) + \text{Noise} \quad \text{for} \quad n = 1, \ldots, N$$

Convex problem (RF-RR)

$$\hat{\alpha}^{(N,M,\lambda)} := \operatorname*{arg\,min}_{\alpha \in \mathbb{R}^M} \left\{ \frac{1}{N} \sum_{n=1}^{N} \big\| y_n - \Psi_{\mathrm{RFM}}(u_n; \alpha) \big\|_{\mathcal{Y}}^2 + \lambda \left(\frac{1}{M} \sum_{m=1}^{M} |\alpha_m|^2 \right) \right\}$$

Samuel Lanthaler and **N.H.N.**

Error Bounds for Learning with Vector-Valued Random Features

Submitted 2023 (arXiv:2305.17170 stat.ML)

Well-Specified Error Analysis (for any input and output dimension)

Trained RFM $\Psi_{\mathrm{RFM}}(\,\cdot\,; \hat{\alpha}^{(N,M,\lambda)})$, where $\hat{\alpha}^{(N,M,\lambda)} \in \mathbb{R}^M$ solves RF-RR

Assumptions (no spectral assumptions on K are needed due to matrix-free analysis)

- $y_n = \Psi^\dagger(u_n) + \eta_n$, where $\eta_n \overset{\text{iid}}{\sim} \eta$ is subexponential on $\mathcal{Y}$
- $\Psi^\dagger$ and φ are bounded a.s.

Theorem (Squared error: well-specified convergence rate)

If $\Psi^\dagger$ belongs to the RKHS of the RF pair (φ, μ) (relaxations too), then

$$\mathbb{E}^{u\sim\nu}\big\|\Psi^\dagger(u) - \Psi_{\mathrm{RFM}}(u; \hat{\alpha}^{(N,M,\lambda)})\big\|_{\mathcal{Y}}^2 \lesssim \lambda + \frac{1}{M} + \frac{1}{\sqrt{N}} \quad \textit{with high probability.}$$

Comparison to Existing Well-Specified Results

Recall notation

- Regularization parameter: λ
- Training data sample size: N
- Number of random features: M

Paper	Approach	λ	$\dim(\mathcal{Y})$	M	Squared Error
Rahimi & Recht '08	"kitchen sinks"	—	1	N	$\omega(N^{-1/2})$
Rudi & Rosasco '17	matrix concen.	$N^{-1/2}$	1	$\sqrt{N}\log(N)$	$O(N^{-1/2})$
Li et al. '21	matrix concen.	$N^{-1/2}$	1	$\sqrt{N}\log(O(N))$	$O(N^{-1/2})$
This Talk (SOTA)	**"kitchen sinks"**	$\boldsymbol{N^{-1/2}}$	$\boldsymbol{\infty}$	$\boldsymbol{\sqrt{N}}$	$\boldsymbol{O(N^{-1/2})}$

Core Proof Idea

Loss

$$L(u;\alpha) \coloneqq \left\|\Psi^\dagger(u) - \Psi_{\mathrm{RFM}}(u;\alpha)\right\|_{\mathcal{Y}}^2$$

Error decomposition (Rahimi & Recht '08)

$$\underbrace{\mathbb{E}^{u\sim\nu}\left[L(u;\hat{\alpha}^{(N,M,\lambda)})\right]}_{\text{squared error}} = \underbrace{\frac{1}{N}\sum_{n=1}^{N} L(u_n;\hat{\alpha}^{(N,M,\lambda)})}_{\text{empirical approximation error}}$$

(Monte Carlo)

$$+ \underbrace{\left[\mathbb{E}^{u\sim\nu}\left[L(u;\hat{\alpha}^{(N,M,\lambda)})\right] - \frac{1}{N}\sum_{n=1}^{N} L(u_n;\hat{\alpha}^{(N,M,\lambda)})\right]}_{\text{generalization gap}}$$

(linearity and empirical processes)

Unifying Sources of Error

Approximation, finite data, noise, and discretization (optimization error is zero)

Corollary (Stability to discretization error)

Let the data be discretized as

$$y_n = \Psi_h^\dagger(u_n) + \eta_n \,,$$

where $h > 0$ denotes a discretization parameter corresponding to bounded discretized operator $\Psi_h^\dagger$. If bounded $\Psi^\dagger$ belongs to the RKHS of (φ, μ), then $\lambda \asymp N^{-1/2} \asymp M^{-1}$ guarantees that

$$\mathbb{E}^{u\sim\nu}\big\|\Psi^\dagger(u) - \Psi_{\mathrm{RFM}}(u; \hat{\alpha}^{(N,M,\lambda)})\big\|_{\mathcal{Y}}^2 \lesssim \frac{1}{\sqrt{N}} + \varepsilon_h^2 \quad \textit{with high probability}\,,$$

where the discretization error is

$$\varepsilon_h := \operatorname*{ess\,sup}_{u\sim\nu} \|\Psi^\dagger(u) - \Psi_h^\dagger(u)\|_{\mathcal{Y}}\,.$$

Model Misspecification

Theorem (Strong statistical consistency)

If the number of features $M = \widetilde{\Omega}(\sqrt{N})$ *and the penalty strength* $\lambda = \widetilde{\Omega}(1/\sqrt{N})$, *then*

$$\lim_{N\to\infty} \mathbb{E}^{u\sim\nu}\big\|\Psi^\dagger(u) - \Psi_{\mathrm{RFM}}(u;\hat{\alpha}^{(N,M_N,\lambda_N)})\big\|_{\mathcal{Y}}^2 = 0 \quad \textit{with probability one.}$$

(Above, $\Psi^\dagger$ is just bounded)

Theorem (Squared error: slow rates)

If $\Psi^\dagger$ *does not belong to the RKHS of* (φ, μ) *but satisfies a "regularity source condition," and* $M \asymp \sqrt{N}$ *and* $\lambda \asymp 1/\sqrt{N}$, *then there exists* $0 < r \leq 1/2$ *such that*

$$\mathbb{E}^{u\sim\nu}\big\|\Psi^\dagger(u) - \Psi_{\mathrm{RFM}}(u;\hat{\alpha}^{(N,M,\lambda)})\big\|_{\mathcal{Y}}^2 \lesssim N^{-r} \quad \textit{with high probability.}$$

Model Misspecification

Theorem (Strong statistical consistency)

If the number of features $M = \widetilde{\Omega}(\sqrt{N})$ *and the penalty strength* $\lambda = \widetilde{\Omega}(1/\sqrt{N})$*, then*

$$\lim_{N\to\infty} \mathbb{E}^{u\sim\nu} \big\| \Psi^\dagger(u) - \Psi_{\mathrm{RFM}}(u; \hat{\alpha}^{(N, M_N, \lambda_N)}) \big\|_{\mathcal{Y}}^2 = 0 \quad \textit{with probability one.}$$

(Above, $\Psi^\dagger$ is just bounded)

Theorem (Squared error: slow rates)

If $\Psi^\dagger$ *does not belong to the RKHS of* (φ, μ) *but satisfies a "regularity source condition," and* $M \asymp \sqrt{N}$ *and* $\lambda \asymp 1/\sqrt{N}$*, then there exists* $0 < r \leq 1/2$ *such that*

$$\mathbb{E}^{u\sim\nu} \big\| \Psi^\dagger(u) - \Psi_{\mathrm{RFM}}(u; \hat{\alpha}^{(N, M, \lambda)}) \big\|_{\mathcal{Y}}^2 \lesssim N^{-r} \quad \textit{with high probability.}$$

Summary

Complete theory for vector-valued RF-RR algorithm

- Statistical consistency of RF for supervised learning and UQ
- SOTA rates in any dimension (matrix-free analysis)
- Includes error due to model misspecification and discretization