Weighting Inputs by Sensitivity in Random Feature Expansions

John Darges

Department of Mathematics
Emory University

Joint work with Laura Weidensager
Chemnitz University of Technology

Partially supported by NSF Grant DMS-2038118

May 22, 2025

High-dimensional Approximation

Approximate a function $z = f(\boldsymbol{x})$, where $\boldsymbol{x}$ is a random vector taking values in $\mathbb{R}^n$ whose law is μ

- Approximate using a set of interpolation or sample points
- Budget for sample points is limited
- Input dimension is high enough that ordered basis methods (orthogonal polynomials, Fourier series) are not feasible
- Make no other assumptions about structure of f except it is in $L_2(\mu)$ *(this doesn't mean f has no structure)*
- Methods of choice often are kernels or neural networks

Kernel Methods and Kernel Operator

- Let $\kappa : X \times X \to \mathbb{C}$ be a reproducing kernel with associated RKHS $\mathcal{H}_\kappa$.
- Kernel approximation aims to find f^* that solves $\min_{\hat{f} \in \mathcal{H}_\kappa} \|f - \hat{f}\|_2$
- Note that the integral operator $T_\kappa : L_2(\mu) \to L_2(\mu)$ has $\mathcal{H}_\kappa$ as its image

$$(T_\kappa g)(\boldsymbol{x}) = \int_{\mathbb{R}^n} g(\boldsymbol{y}) \kappa(\boldsymbol{x}, \boldsymbol{y}) \, d\mu(\boldsymbol{y})$$

- There exists g^* such that $T_\kappa g^* = f^*$, we can describe an approximation

$$f^* \approx \sum_{l=1}^{M} \frac{g^*(\boldsymbol{x}_k)}{M} \kappa(\cdot, \boldsymbol{x}_k), \quad \boldsymbol{x}_k \sim \mu$$

- We don't know f^*, g^* – we find coefficients through least squares (kernel ridge regression)

Feature Map Representation

- Many favorite kernels come from a feature map ϕ and a measure τ, admitting a (non-unique) feature map representation

$$\kappa(\boldsymbol{x}, \boldsymbol{y}) = \int_{\mathbb{R}^n} \phi(\boldsymbol{x}^\top \boldsymbol{\omega}) \overline{\phi(\boldsymbol{y}^\top \boldsymbol{\omega})}\, d\tau(\boldsymbol{\omega})$$

- Bochner: shift invariant kernels are Fourier transforms of Borel measures
- Feature map induces an integral operator $T_\phi : L_2(\tau) \to L_2(\mu)$ whose image is $\mathcal{H}_\kappa$[1]

$$(T_\phi \alpha)(\boldsymbol{x}) = \int_{\mathbb{R}^n} \alpha(\boldsymbol{\omega}) \phi(\boldsymbol{x}^\top \boldsymbol{\omega})\, d\tau(\boldsymbol{\omega})$$

- There exists α^* such that $T_\phi \alpha^* = f^*$, leading to an approximation

$$f^* \approx \sum_{k=1}^{M} \frac{\alpha^*(\boldsymbol{\omega}_k)}{K} \phi(\boldsymbol{\omega}_k^\top \cdot), \quad \boldsymbol{\omega}_k \sim \tau$$

[1] F. Bach. On the Equivalence between Kernel Quadrature Rules and Random Feature Expansions. 2017.

Random Feature Expansions

- In practice, many RKHS dense in $L_2(\mu)$ – some f^* arbitrarily close to f exists[2]
- Class $\{\phi(\boldsymbol{\omega}^\top \cdot) : \boldsymbol{\omega} \sim \tau\}$ is dense in $\mathcal{H}_\kappa$, random feature expansion (RFE) is an expansion over random choices of basis functions from this class[3]
- Common choice is Fourier features $\{e^{i\boldsymbol{\omega}^\top \cdot)}\}$ or $\{\cos(\boldsymbol{\omega}^\top \cdot)\}$
- Entries of $\boldsymbol{\omega}$ often identical and independent, e.g., each $\omega^j \sim \mathcal{N}(0, 1)$ so τ is a product of identical Gaussian measures
- Coefficients β_k determined by solving linear least squares problem

$$\min_{\boldsymbol{\beta}} \sum_{i=1}^{N} \left(f(\boldsymbol{x}_i) - \sum_{k=1}^{K} \beta_k \phi(\boldsymbol{\omega}_k^\top \boldsymbol{x}_i) \right)^2, \quad \boldsymbol{\omega}_k \sim \tau$$

- Equivalent to single layer neural networks with random weights[4]

[2] C. Michelli, Y. Xu, H. Zhang. Universal Kernels. 2008.

[3] A. Rahimi, B. Recht. Uniform approximation of functions with random bases. 2008.

[4] W. Cao, X. Wang, Z. Ming, J. Gao. A review on neural networks with random weights. 2018.

Choices in Kernels and Choosing Kernels

- Even if kernel is universal, choice must be catered to the function or data
- If kernel chosen from a a parametric class, must make choice of parameters, e.g., shape parameter
- Class of kernels might be unsuitable for f
- Choosing kernel with a feature map representation comes down to choosing the measure τ
- One should choose kernel based on smoothness and **multivariable interaction structure** of f
- Even if inputs are high-dimensional, f may be dominated by only a few inputs or a few lower-order interaction terms

Decompositions of Multivariable Functions

- Multivariable functions can be decomposed (in different ways) into terms depending on groups of inputs[5]

$$f = \sum_{\boldsymbol{u} \subseteq [n]} f_{\boldsymbol{u}}, \quad f_{\boldsymbol{u}} = T_{\boldsymbol{u}} f - \sum_{v \subset u} f_{\boldsymbol{v}}$$

- Terms $f_{\boldsymbol{u}}$ depends only on subset of inputs x_j, $j \in \boldsymbol{u}$
- $T_{\boldsymbol{u}}$ is a linear transformation that determines the type of decomposition
- Anchored: $(T_{\boldsymbol{u}} f)(\boldsymbol{x}_{\boldsymbol{u}}) = f(\boldsymbol{x}_{\boldsymbol{u}}, \boldsymbol{x}_{\boldsymbol{u}^{\complement}} = 0)$ fixes values of inputs not in indices $\boldsymbol{u}$
- Hoeffding-Sobol' (aka ANOVA)[6]: $(T_{\boldsymbol{u}} f)(\boldsymbol{x}_{\boldsymbol{u}}) = \int f \, d\boldsymbol{x}_{\boldsymbol{u}^{\complement}}$ integrates out inputs not in indices $\boldsymbol{u}$
- Allow us to project f into 2^n orthogonal components

[5] F. Kuo, I. Sloan, G. Wasilkowski, H. Woźniakowski. On Decompositions of Multivariable Functions. 2010.

[6] I.M. Sobol'. Multidimensional Quadrature Formulas and Haar Functions. 1969.

Multivariable Structure Matters for the Sampling Distribution

- Suppose f is an additive function, meaning it has no higher order interaction terms
- Then $f = \sum_{i=1}^{n} f_i$ and $f_{\boldsymbol{u}} = 0$ if $|\boldsymbol{u}| > 1$
- Let $f^N = \sum_{k=1}^{K} \beta_k e^{i\omega_k^\top \cdot}$ be a random Fourier approximation to f where ω_k is sampled from standard normal distribution
- The error in the approximation is

$$\|f - f^N\| = \sum_{\boldsymbol{u} \subseteq [n]} \|f_{\boldsymbol{u}} - f_{\boldsymbol{u}}^N\| \tag{1}$$

$$= \sum_{i=1}^{n} \|f_i - f_i^N\| + \sum_{|\boldsymbol{u}| \geq 2} \|f_{\boldsymbol{u}}^N\| \tag{2}$$

- This happens because we almost surely sample basis functions from the class $\{e^{i\omega^\top \cdot}\}$ containing every order of interaction

Sparse Sampling

- Recent work proposed imposing interaction structure on RFEs through sparse weights
- Sparse features $\phi(\boldsymbol{\omega}^\top \boldsymbol{x})$ have weights where most entries in $\boldsymbol{\omega}$ are 0
- Robust sparse reconstruction method proposed in[7]
- Sparse random features proposed in[8] assume target model only has low order interactions
- Weight sparsity in[9] determined by hyperparameter tuning
- Sparse RFEs in[10] sample weights based on ANOVA decomposition of target model

[7] E. Saha, H. Schaeffer, G. Tran. HARFE: hard-ridge random feature expansion. 2023.

[8] A. Hashemi, H. Schaeffer, R. Shi, U. Topcu, G. Tran, R. Ward. Generalization bounds for sparse random feature expansions. 2023.

[9] J.D., A. Alexanderian, P.A. Gremaud. Extreme Learning Machines for Variance-Based Global Sensitivity Analysis. 2024.

[10] L. Weidensager, D. Potts ANOVA-boosting for Random Fourier Features. 2024.

Weighted ANOVA Kernel

- To account for multivariable structure, use a weighted ANOVA kernel

$$\kappa_W(\boldsymbol{x}, \boldsymbol{y}) = \sum_{\boldsymbol{u} \subseteq [n]} \gamma_{\boldsymbol{u}} \kappa_{\boldsymbol{u}}(\boldsymbol{x}_{\boldsymbol{u}}, \boldsymbol{y}_{\boldsymbol{u}})$$

- If we use a kernel $\kappa(\boldsymbol{x}, \boldsymbol{y}) = \int e^{i\boldsymbol{x}^\top \boldsymbol{\omega}} \overline{e^{i\boldsymbol{y}^\top \boldsymbol{\omega}}}\, d\tau(\boldsymbol{\omega})$ and replace with a measure that places weights on inputs

$$\tau = \bigotimes_{i=1}^{n} \nu \quad \rightarrow \quad \tau_W = \bigotimes_{i=1}^{n} (\rho_i \nu + (1 - \rho_i)\delta_0), \quad \rho_i \in (0, 1)$$

- This results in a weighted kernel derived from the original

$$\kappa_W(\boldsymbol{x}, \boldsymbol{y}) = \int e^{i\boldsymbol{x}^\top \boldsymbol{\omega}} \overline{e^{i\boldsymbol{y}^\top \boldsymbol{\omega}}}\, d\tau_W(\boldsymbol{\omega}) = \sum_{\boldsymbol{u} \subseteq [n]} \gamma_{\boldsymbol{u}} \int e^{i\boldsymbol{x}_{\boldsymbol{u}}^\top \boldsymbol{\omega}_{\boldsymbol{u}}} \overline{e^{i\boldsymbol{y}_{\boldsymbol{u}}^\top \boldsymbol{\omega}_{\boldsymbol{u}}}}\, d\tau(\boldsymbol{\omega})$$

- Kernel weights are $\gamma_{\boldsymbol{u}} = \prod_{i \in \boldsymbol{u}} \rho_i \prod_{j \notin \boldsymbol{u}} (1 - \rho_j)$

Weighted Random Features

- Weighted ANOVA kernel is too cumbersome to work with – use feature representation instead

$$T_\phi \alpha = \int \alpha(\boldsymbol{\omega}) e^{i \boldsymbol{x}^\top \boldsymbol{\omega}} \, d\tau_W(\boldsymbol{\omega}) = \sum_{\boldsymbol{u} \subseteq [n]} \int \alpha(\boldsymbol{\omega}_u) e^{i \boldsymbol{x}_u^\top \boldsymbol{\omega}_u} \, d\tau(\boldsymbol{\omega})$$

- Random weighted feature expansion is

$$f^N = \sum_{k=1}^{K} \hat{\beta}_k e^{i \boldsymbol{x}^\top \boldsymbol{\omega}_k}, \quad \boldsymbol{\omega}_k \sim \tau_W$$

- This means the entry in each $\boldsymbol{\omega}$ is $\omega^i = YZ_i$, $Y \sim \nu$, $Z_i \sim \text{Bern}(1 - \rho_i)$

Proposed Approach

- Introduce random sparsity to feature weights based on sensitivity of inputs they correspond to
- If f is not sensitive to x_i, all its corresponding weights should be sparse (set to 0)
- Sample entries in weight vector $\boldsymbol{\omega}^i$ corresponding to θ_i as random variable $\omega^i = Z_i X$, where $X \sim \mathcal{N}(0, 1)$, $Z_i \sim \text{Bern}(1 - \rho_i)$,
- ρ_i gives sensitivity of θ_i (**total Sobol' index**, derivative-based measures, Shapley value, etc.)
- **Begin by training a rough approximation to determine estimates for ρ_i, then build new approximation with updated sampling distribution**

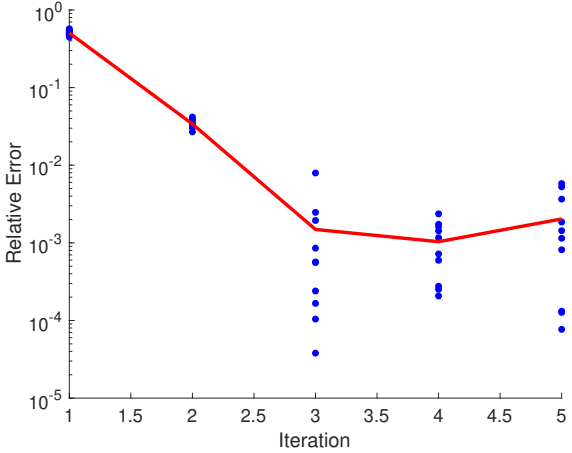
Algorithm

1.
2. Set sampling distribution to standard normal distribution, $\boldsymbol{\omega}^i \sim \mathcal{N}(0, 1)$ for all i
3. Sample feature weights and compute RFE approximation
4. Estimate input sensitivities ρ_i using the approximation
5. Update sampling distribution with random sparsity, $\omega^i = Z_i X$, where $X \sim \mathcal{N}(0, 1)$, $Z_i \sim \text{Bern}(1 - \rho_i)$
6. Iteratively repeat 2–5

How does this work in practice?

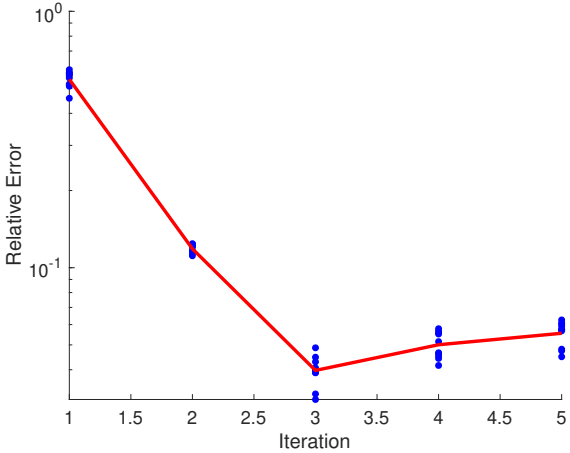
Example: Grienwank Function

$$F(\boldsymbol{x}) = \sum_{k=1}^{50} \frac{x_k^2}{4000} + \prod_{k=1}^{50} \left(\cos\left(\frac{x}{\sqrt{k}}\right) + 1\right), \quad \boldsymbol{x} \sim \mathcal{U}([-600, 600]^{50}).$$



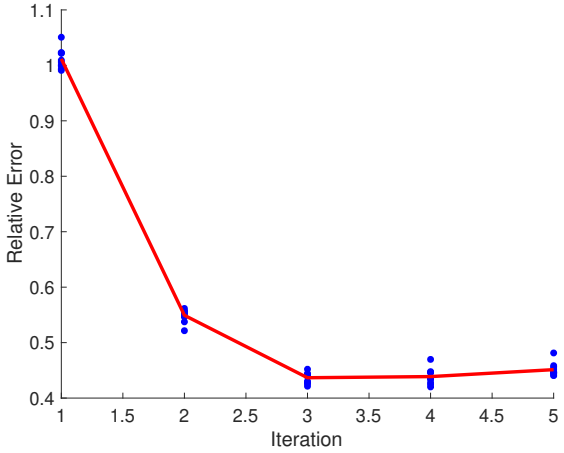
Example: Dixon-Price Function

$$f(\boldsymbol{x}) = (x_1 - 1)^2 + \sum_{k=2}^{50} k(2x_k - x_{k-1})^2, \quad \boldsymbol{x} \sim \mathcal{U}([-10, 10]^{50}).$$



Example: Sobol' g-function

$$F(\boldsymbol{x}) = \prod_{k=1}^{10} \frac{|4x_k - 2| + a_k}{1 + a_k}, \quad a_k = \frac{k-2}{2}, \quad \boldsymbol{x} \sim \mathcal{U}([0,1]^{10})$$



Avenues for Theory

- Does altering the sampling distribution based on weight sensitivity provide benefits for certain classes of functions?
- Does $\|f - f_W^N\|$ converge faster than $\|f - f^N\|$ for certain classes of f dominated by a subset of inputs or interaction terms?
- This analysis relies on showing $\|f^* - f_W^N\|$ converges faster than $\|f^* - f^N\|$, recall

$$\|f^* - f^N\| = \left\| \int \alpha^*(\boldsymbol{\omega}) e^{i\boldsymbol{\omega}^\top \mathbf{x}} \, d\tau(\boldsymbol{\omega}) - \sum_{k=1}^{K} \frac{\alpha^*(\boldsymbol{\omega}_k)}{K} e^{i\boldsymbol{\omega}_k^\top \mathbf{x}} \right\|$$

-

Summary and Goals

- Proposed an algorithm for random feature expansions that modifes weight sampling by taking relative input sensitivities into account
- Incorporate input weighting into the state-of-the-art RFE algorithms
- Identify ideal applications and adverserial examples
- Are there better approaches to input sensitivity based on the choice of kernel?
- Should input weighting be implemented in different fashion, e.g., by scaling inputs or scaling variances in the sampling distribution?