Machine
Intelligence
&
Networked
Data
Science

PURDUE UNIVERSITY
Elmore Family School of Electrical and Computer Engineering

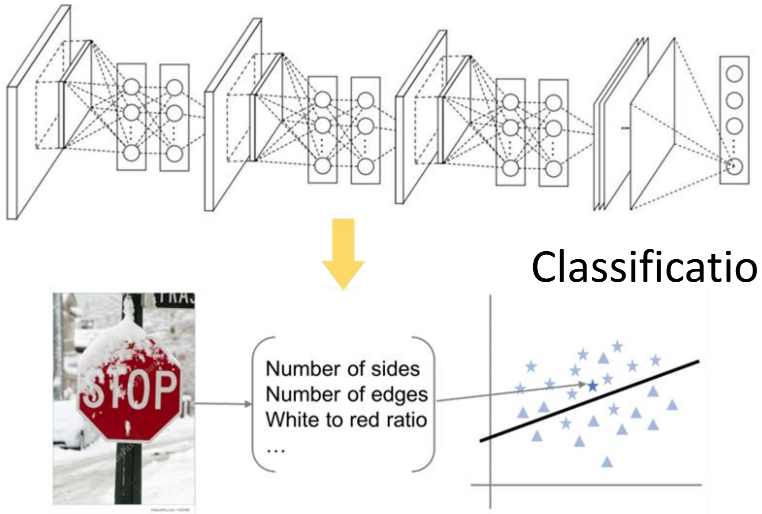
Generalization Bounds for Sparse Random Feature Expansions

Abolfazl Hashemi

SIAM MDS Conference

Sept 30th, 2022

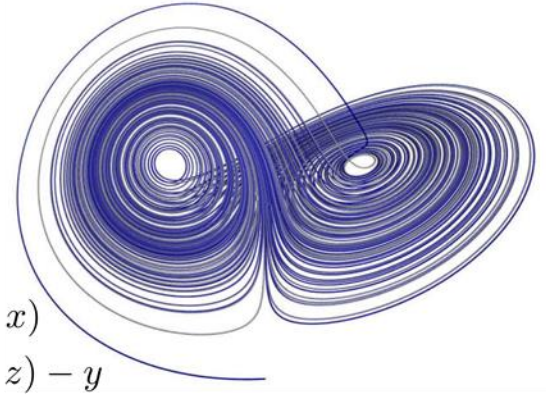
Function Approximation



Classification

Learn an unknown relation from data

Dynamical systems



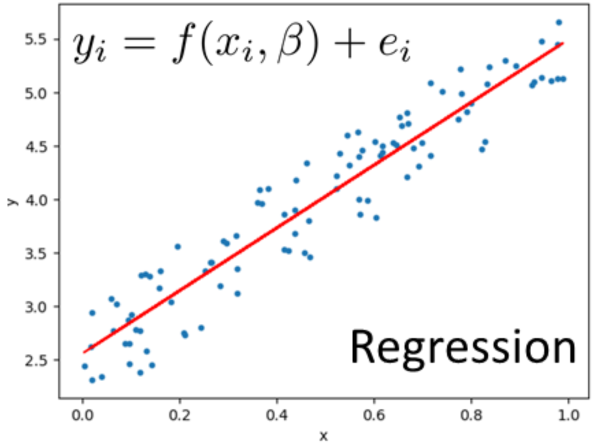
$$\dot{x} = \sigma(y - x)$$
$$\dot{y} = x(\rho - z) - y$$
$$\dot{z} = xy - \beta z$$

Access to a lot of data (e.g., neural networks)

PDEs

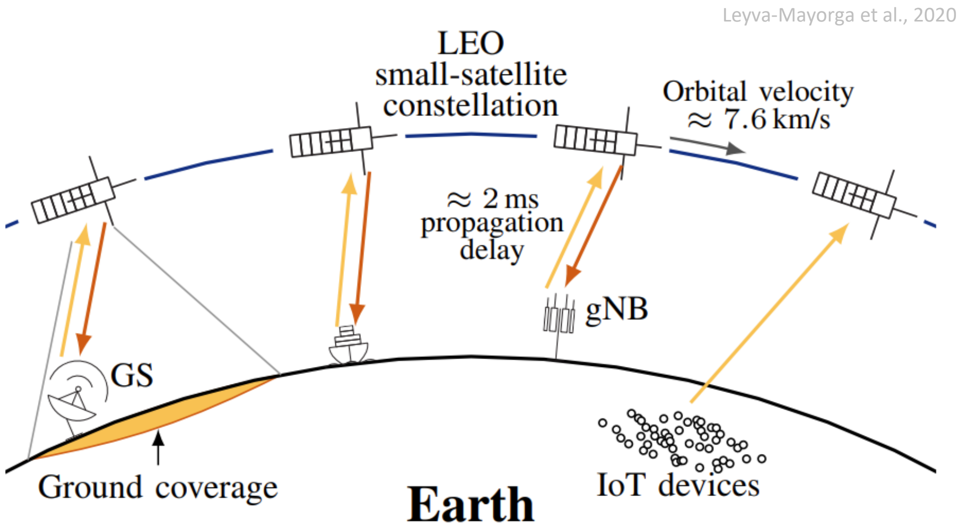
$$i\hbar\frac{\partial}{\partial t}|\Psi(t)\rangle = \hat{H}|\Psi(t)\rangle$$

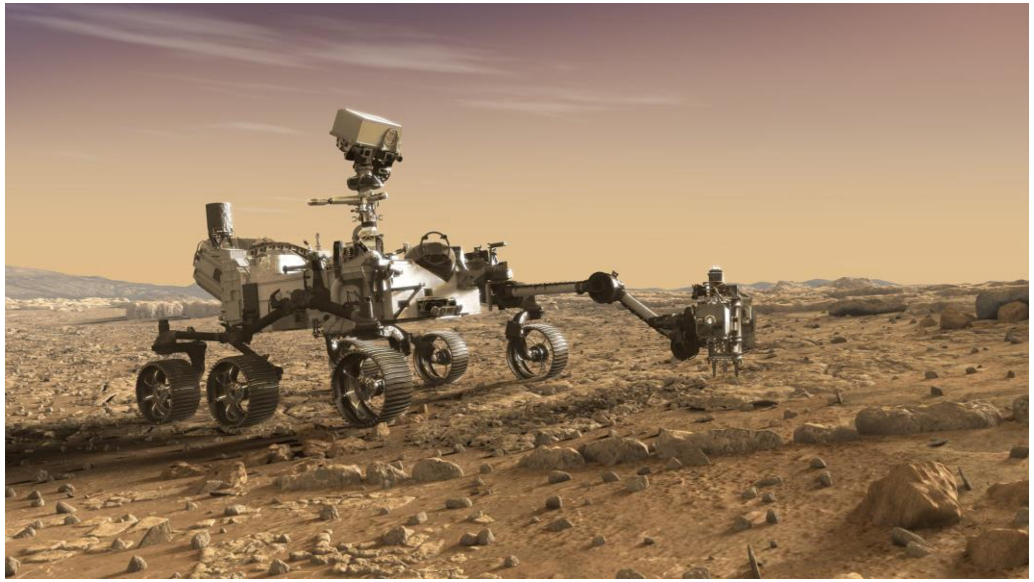
Targeted to specific function classes (e.g., polynomials)



Hashemi et al., 2022

Function Approximation Under Data-Scarcity





Limited energy budget

Costly observation gathering

Safety-critical operation

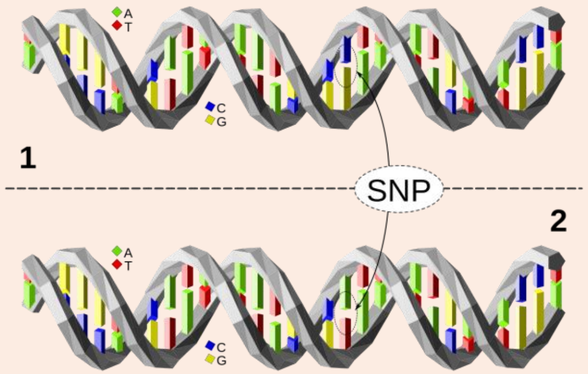
How can we design data-efficient algorithms with generalization guarantees in data-scarce settings?

Hashemi et al., 2022

Function Approximation with Latent, Parsimonious Structures

Low-rank structure

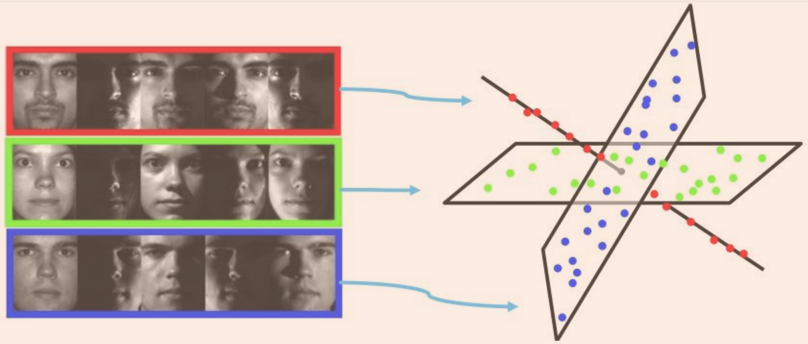
Genome sequencing



Wikipedia

Sparsity

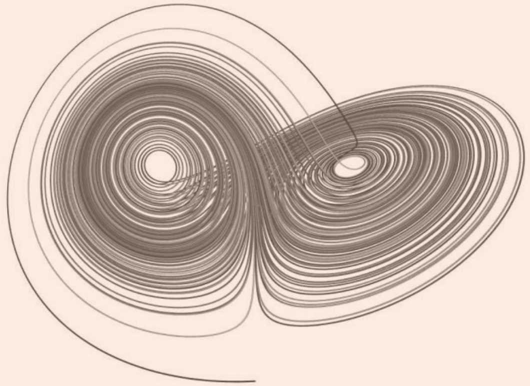
Clustering



You et al., 2018

Low-order structure

Dynamical systems



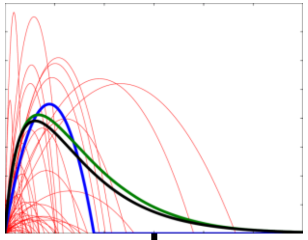
$$\dot{x}_i = x_{i+1}x_{i-1} - x_{i-1}x_{i-2} - x_i + 8, \quad i = 1, \ldots d$$

Wikipedia

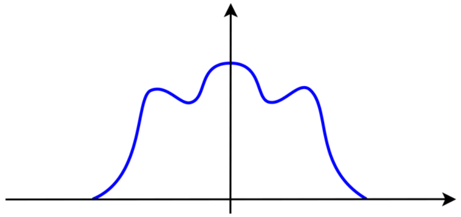
How can we leverage low-order structures for improved data-efficiency in data-scarce settings?

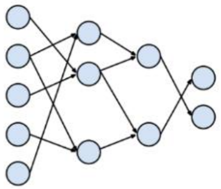
Sparse Random Feature Expansions

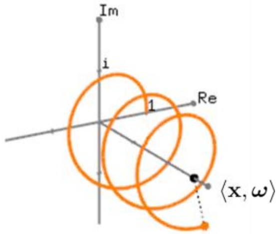
Limited data



Structures (bandlimited, low-order)







Parameterized basis via sparse random features

Sparse convex optimization

Contributions:

- Leveraging sparsity and low-order structures for improved data-efficiency
- Constructive accuracy guarantees and generalization bounds

Function Approximation: Basis Representation View

Learn the unknown function

$$f : \mathbb{R}^d \to \mathbb{C}$$

From the dataset

$$\{(y_1, \mathbf{x}_1), \ldots, (y_m, \mathbf{x}_m)\}$$

$$f(\mathbf{x}) \approx \sum_{j=1}^{N} c_j \phi(\mathbf{x}; \boldsymbol{\omega}_j) := \mathbf{A}(\mathbf{x}, \boldsymbol{\omega})\mathbf{c}$$

Representation coefficients

Candidate bases, e.g., Legendre polynomials

$1, x_1, x_2, 0.5(3x_1^2 - 1)x_2$

$$\text{Find } \mathbf{c} \text{ such that: } y_i \approx \langle \mathbf{a}(\mathbf{x}_i, \boldsymbol{\omega}), \mathbf{c} \rangle, \; i = 1, \ldots, m$$

Compressed sensing:

- Convex formulation
- Fixed and deterministic bases, e.g., orthonormal polynomials
- function need to be well-represented by polynomials

Neural networks:

- Nonlinear data-dependent basis functions
- Non-convex formulation
- Scarcity of theoretical guarantees
- Data intensive

Hashemi et al., 2022

Bridging the Gap: Random Feature for Compressive Sensing

$$f(\mathbf{x}) \approx \sum_{j=1}^{N} c_j \phi(\mathbf{x}; \boldsymbol{\omega}_j)$$

→ Two-layer, overparameterized neural network

Learn from data

draw at random and keep fixed

Option for basis function

Random Fourier features: $\phi(\mathbf{x}; \boldsymbol{\omega}) = \exp(i\langle \mathbf{x}, \boldsymbol{\omega} \rangle)$

Random trigonometric features: $\phi(\mathbf{x}; \boldsymbol{\omega}) = \cos(\langle \mathbf{x}, \boldsymbol{\omega} \rangle)$

Random ReLU features: $\phi(\mathbf{x}; \boldsymbol{\omega}) = \max(\langle \mathbf{x}, \boldsymbol{\omega} \rangle, 0)$

Leveraging sparsity

Data-scarcity

Low-order interactions

Sparsity in c_j

Sparsity in ω_j

Hashemi et al., 2022

Sparse Random Feature Expansions

Draw m data points $\mathbf{x}_k \sim \mathcal{D}_x$ and observe noisy measurements $y_k = f(\mathbf{x}_k) + e_k$
$|e_k| \leq E$

Draw a complete set of N q-sparse feature weights $\boldsymbol{\omega}_j \in \mathbb{R}^d$ sampled from density $\zeta : \mathbb{R}^q \rightarrow \mathbb{R}$ with variance σ^2

Construct a random feature matrix $\mathbf{A} \in \mathbb{C}^{m \times N}$ such that $a_{kj} = \phi(\mathbf{x}_k; \boldsymbol{\omega}_j)$

$$\mathbf{c}^\sharp = \arg\min_{\mathbf{c}} \quad \|\mathbf{c}\|_1 \qquad \text{s.t.} \qquad \|\mathbf{A}\mathbf{c} - \mathbf{y}\| \leq \eta\sqrt{m}$$

$$f^\sharp(\mathbf{x}) = \sum_{j=1}^{N} c_j^\sharp \phi(\mathbf{x}; \boldsymbol{\omega}_j)$$

Hashemi et al., 2022

Function Space

Bounded ρ-norm functions [Rahimi, Recht '08]

$$\underbrace{\mathcal{F}(\phi, \rho)} := \left\{ g(\mathbf{x}) = \int_{\boldsymbol{\omega} \in \mathbb{R}^d} \alpha(\boldsymbol{\omega}) \phi(\mathbf{x}; \boldsymbol{\omega}) \; d\boldsymbol{\omega} \; : \; \underbrace{\|g\|_\rho} := \sup_{\boldsymbol{\omega}} \left| \frac{\alpha(\boldsymbol{\omega})}{\rho(\boldsymbol{\omega})} \right| < \infty \right\}$$

Related to Barron function class in NN [Barron '92]

Strong dependence on dimension d

$$\|f\|_B = \int |\boldsymbol{\omega}| |\hat{f}(\boldsymbol{\omega})| \; d\boldsymbol{\omega}$$

$$\|f - f^\star\|_{L_2} \leq \frac{C \|f\|_B}{\sqrt{N}}$$

Smaller than

$\|f\|_\rho$

Order-q function

$$f(x_1, \ldots, x_d) = \frac{1}{K} \sum_{j=1}^{K} g_j(x_{j_1}, \ldots, x_{j_q})$$

$$|||f||| := \binom{d}{q}^{\frac{1}{2}} \left(\frac{1}{K} \sum_{j=1}^{K} \|g_j\|_\rho \right)$$

Generalization Bounds

Theorem: (Generalization bound for low-order functions)

Let $\phi(\mathbf{x};\boldsymbol{\omega}) = \phi(\langle \mathbf{x}, \boldsymbol{\omega} \rangle) = \exp(i\langle \mathbf{x}, \boldsymbol{\omega} \rangle)$. For each subset $\mathcal{S} \subset [d]$ with $|\mathcal{S}| = q$, draw i.i.d. $\boldsymbol{\omega}_1, \ldots, \boldsymbol{\omega}_n \sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{I}_q)$. Let $\mathbf{x}_1, \ldots, \mathbf{x}_m \sim \mathcal{N}(\mathbf{0}, \gamma^2 \mathbf{I}_d)$, $\eta = \sqrt{2(\epsilon^2 |||f|||^2 + E^2)}$ and fix, *s*, δ, and ϵ

- γ-σ uncertainty principle $\quad \gamma^2\sigma^2 \geq \frac{1}{2}\left(\left(\frac{\sqrt{41}(2s-1)}{2}\right)^{\frac{2}{q}} - 1\right)$
- Number of features $\quad n \geq \frac{4}{\epsilon^2}\left(1 + 4\gamma\sigma d + \sqrt{\frac{q}{2}\log\left(\frac{d}{\delta}\right)}\right)^2$
- Number of measurement $\quad m \geq 4(2\gamma^2\sigma^2 + 1)^{2q} \log \frac{N^2}{\delta}$

With probability $1 - \delta$

$$\boxed{\sqrt{\int_{\mathbb{R}^d} |f(\mathbf{x}) - f^\star(\mathbf{x})|^2 \, d\mu} \leq \epsilon |||f||| + C' |||f||| + C\eta\sqrt{s}}$$

Improved Bound for Bandlimited Functions

$$\sqrt{\int_{\mathbb{R}^d} |f(\mathbf{x}) - f^\star(\mathbf{x})|^2 \, d\mu} \le \epsilon |||f||| + C' |||f||| + C\eta \sqrt{s}$$

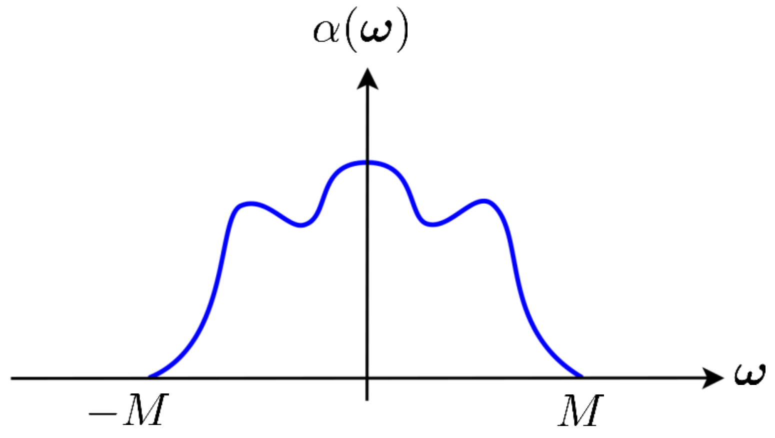
Can be improved for bandlimited functions

Theorem: (low-order and bandlimited functions)

Setting $s = \tilde{\mathcal{O}}(\sqrt{n})$, if there is no noise

$$\sqrt{\int_{\mathbb{R}^d} |f(\mathbf{x}) - f^\star(\mathbf{x})|^2 \, d\mu} \le \epsilon |||f||| + C |||f||| \sqrt{\epsilon}$$

$$\mathcal{F}_M(\phi, \rho) := \left\{ g(\mathbf{x}) = \int_{\boldsymbol{\omega} \in \mathbb{R}^q} \alpha(\boldsymbol{\omega}) \phi(\mathbf{x}; \boldsymbol{\omega}) \, d\boldsymbol{\omega} \; : \mathrm{supp}(\alpha) \subset \mathbb{B}(\mathbf{0}, M) \right\}$$



Effectiveness on Low-order Function Approximation

Order of interactions: *q = 2*

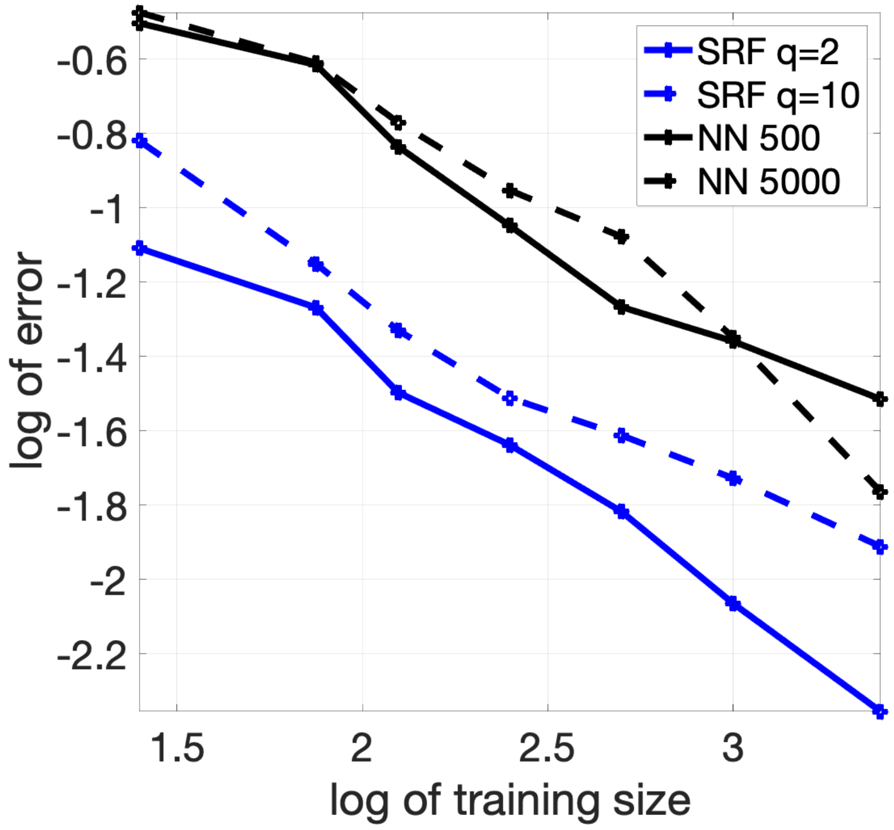
$$f(x_1,\ldots,x_{10}) = \frac{1}{10}\sum_{\ell=1}^{9}\frac{\exp(-x_\ell^2)}{1+x_{\ell+1}^2}$$

N = 5000 features

Varying size of the training dataset

Comparison with shallow NN

Measuring relative test error



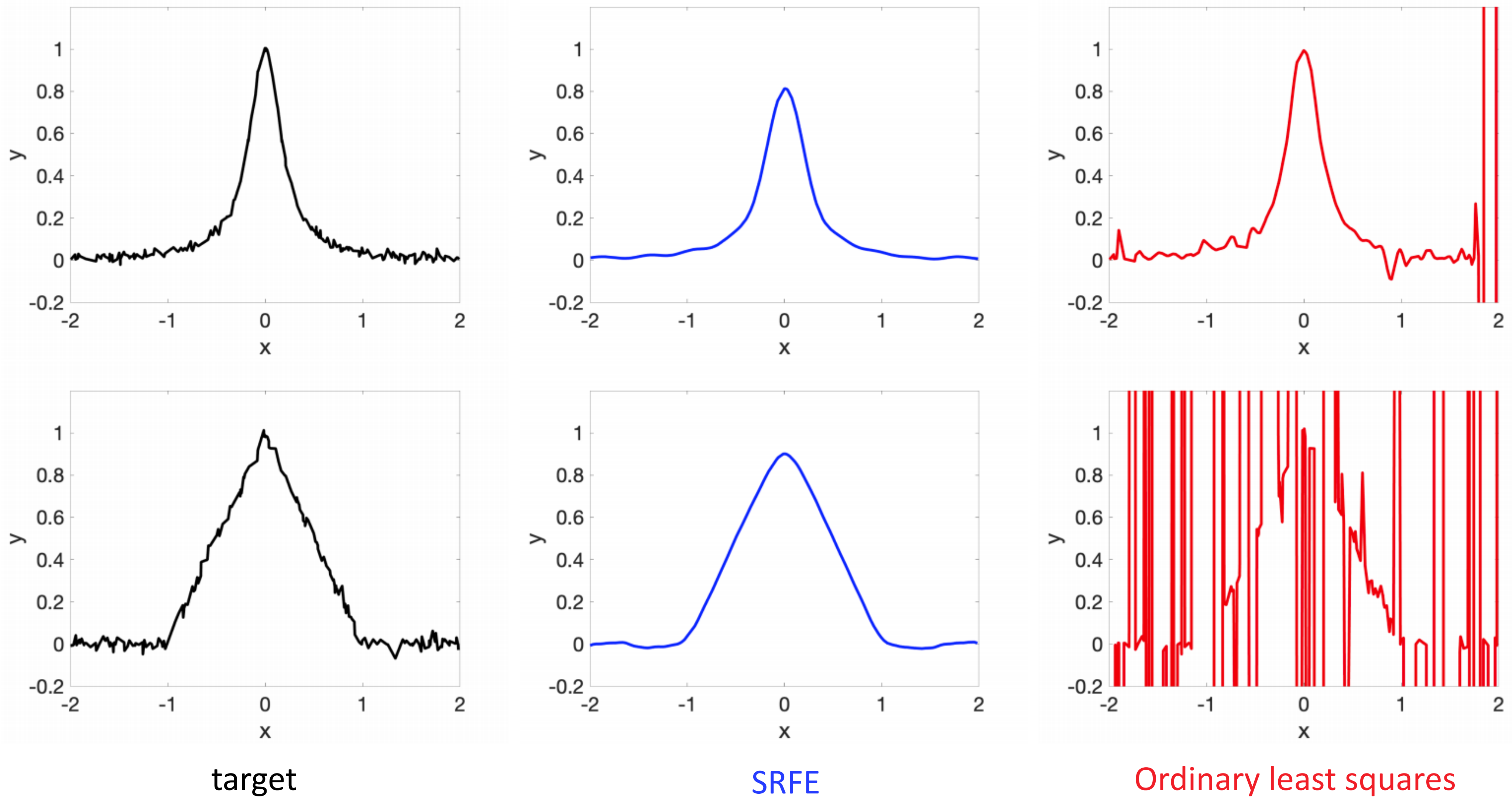
Hashemi et al., 2022

Impact of the Order of the Function

$f(\mathbf{x})$	σ	d	$q=1$	$q=2$	$q=3$	$q=5$
$\left(\sum_{i=1}^{d} x_i\right)^2$	0.1	1	0.82	5.71×10^{-6}	6.92×10^{-5}	8.3×10^{-4}
$(1+\lVert\mathbf{x}\rVert_2^2)^{-1/2}$	1	5	3.27	1.60	1.95	1.72
$\sqrt{1+\lVert\mathbf{x}\rVert_2^2}$	1	5	1.02	0.73	0.80	1.10
$\mathrm{sinc}(x_1)\mathrm{sinc}(x_3)^3 + \mathrm{sinc}(x_2)$	π	5	12.90	1.19	1.13	3.51
$\frac{x_1 x_2}{1+x_3^6}$	1	5	100.30	21.53	4.95	5.06
$\sum_{i=1}^{d} \exp(-\lvert x_i \rvert)$	1	100	0.91	1.43	1.57	1.96

Approximately order *q = 2* *functions*

Robustness Against Overfitting and Noise



Data-Scarcity and Generalization Error

HyShot 30 data

M= 26 data points

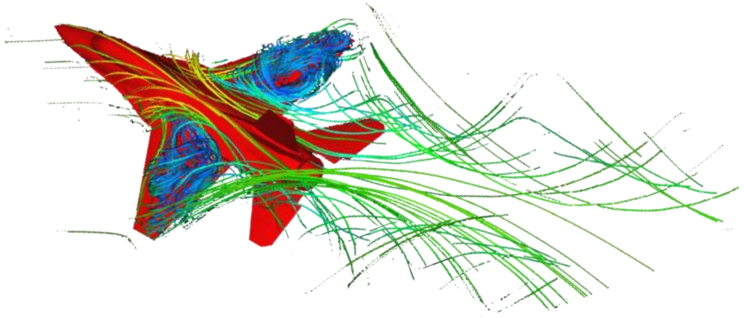
Full order: *d = q = 7*

NACA sound data

M= 1202 data points

Full order: *d = q = 5*

HyShot 30	$N = 100$	$N = 200$	$N = 400$	$N = 800$
SRFE with Sine	6.95	6.23	5.76	5.64
SRFE with ReLU	1.40	1.45	1.51	1.59
Random Fourier Features	84.23	89.99	95.17	97.84
Two-layer ReLU Network	7.29	11.50	11.19	11.33
NACA Sound	$N = 250$	$N = 1500$	$N = 5000$	$N = 10000$
SRFE (Train)	3.22	2.30	2.30	2.31
SRFE (Test)	3.22	3.04	2.77	2.78
SRFE (Average Sparsity)	250	364.4	185.7	185.7
Random Fourier Features (Train)	3.22	0.25	0.20	0.19
Random Fourier Features (Test)	7.45	2.13×10^8	1.69×10^8	1.48×10^8



Low generalization error due to coefficient sparsity

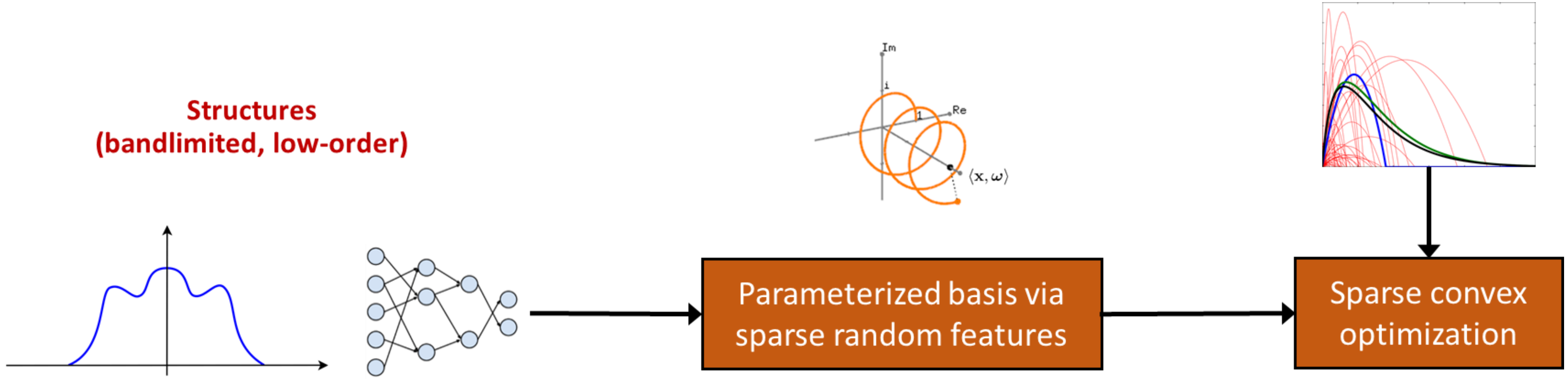
Hashemi et al., 2022

Conclusion and Future Work

- Sparse random feature expansions
- Leveraging coefficient sparsity in data-scarce setting
- Leveraging feature sparsity for exploiting low-order structures
- Constructive generalization bounds for function approximation

Future Directions

- Approximately low-order structure
- Tuning the feature weights
- Incorporate additional functional structures such as Lipschitzness



Generalization Bounds for Sparse Random Feature Expansions

Abolfazl Hashemi, Hayden Schaeffer, Robert Shi, Giang Tran, Rachel Ward, Ufuk Topcu.

Appeared at "Applied and Computational Harmonic Analysis," 2022.