Modulation Spaces and the Curse of Dimensionality

Rahul Parhi
Biomedical Imaging Group
École polytechnique fédérale de Lausanne

(joint work with Michael Unser)

Sampling Theory and Applications Conference

14 July 2023

What is the Curse of Dimensionality?

- The phrase the "curse of dimensionality" was (allegedly) coined by Bellman 1961.

$\Longrightarrow$ Optimization by exhaustive enumeration on product spaces.
$\Longrightarrow$ e.g., Cartesian grid of spacing, say, $1/5$ on the unit cube $[0,1]^d$.

  - $d = 5 \implies 5^5 \sim 3,000$
  - $d = 10 \implies 5^{10} \sim 10,000,000$
  - $d = 15 \implies 5^{15} \sim 30,000,000,000$

- Problems become intractable even in low ($d = 15$) dimensions!
- Many modern problems (data science/machine learning) are very high-dimensional.

Today's Fundamental Question

Is there a way to **avoid** the curse of dimensionality?

More Concretely...

Let $f \in W^{1,\infty}(\Omega)$, where $\Omega \subset \mathbb{R}^d$ is bounded (e.g., $\Omega = [0,1]^d$).

- Approximately optimize f to an error $\varepsilon > 0$.

  $\Longrightarrow$ Need $(1/\varepsilon)^d$ evaluations on a grid. (Bellman 1961)

- Approximate f to an error $\varepsilon > 0$ with, say, wavelets.

  $\Longrightarrow$ Need $N = (1/\varepsilon)^d$ wavelets. (DeVore 1998)

  $\Longrightarrow$ The best N-term L^2-approximation error rate is $N^{-\frac{1}{d}}$.

- Learn/estimate f from noisy measurements, say,

$$y_m = f(\boldsymbol{x}_m) + \varepsilon_m,\ m = 1, \dots, M.$$

$\Longrightarrow$ MISE rate from wavelet thresholding is $M^{-\frac{2}{2+d}}$. (Donoho and Johnstone 1998)

What's Going On?

- The assumption $f \in W^{1,\infty}(\Omega)$ is too general.

  $\Longrightarrow$ $W^{1,\infty}(\Omega)$ is too large of a **model class**.
- In fact, **all** model classes defined via classical notions of **smoothness** (say, s derivatives in L^p) suffer the curse of dimensionality.

  $\Longrightarrow$ The L^2-entropy number of the unit ball of $B^s_{p,q}(\Omega) \subset\subset L^2(\Omega)$ scales as

$$\varepsilon_N(\{f : \|f\|_{B^s_{p,q}} \leq 1\})_{L^2} \asymp N^{-\frac{s}{d}}$$

  - How precisely functions can be specified by N-bits.
  - Famous theorem of Birman and Solomyak 1967.

Question

Can we design model classes that are **immune** to the curse of dimensionality?

In 1993...

- ...Andrew Barron **broke** the curse of dimensionality.

930 IEEE TRANSACTIONS ON INFORMATION THEORY, VOL. 39, NO. 3, MAY 1993

Universal Approximation Bounds for Superpositions of a Sigmoidal Function

Andrew R. Barron, *Member, IEEE*

- If $\int_{\mathbb{R}^d} (1 + |\boldsymbol{\xi}|)^s |\widehat{f}(\boldsymbol{\xi})| \, \mathrm{d}\boldsymbol{\xi} < \infty$, then there exists a shallow neural network f_N with N neurons such that

$$\|f - f_N\|_{L^2(\Omega)} \lesssim N^{-\frac{1}{2}}$$

$\Longrightarrow$ This rate is **immune** to the curse of dimensionality!

A Key Observation

$\mathscr{B}^s(\mathbb{R}^d) = \{f \in \mathcal{S}'(\mathbb{R}^d) : \int_{\mathbb{R}^d} (1 + |\boldsymbol{\xi}|)^s |\widehat{f}(\boldsymbol{\xi})| \, \mathrm{d}\boldsymbol{\xi} < \infty\}$ is a Banach space defined by a measure of **sparsity** in the Fourier domain.

Breaking the Curse of Dimensionality with Sparsity

- The work of Barron 1993 spurred a lot of interest from the approximation theory community.

$\Longrightarrow$ Why are $\mathscr{B}^s$ functions “immune” to the curse of dimensionality?

- The underlying idea was made precise by Donoho 2000:

$\Longrightarrow$ Let $\mathcal{F} := \mathcal{F}(\mathbb{R}^d)$ be a function space whose elements are representable by ℓ^1-**combinations of** L^∞**-atoms**, i.e., for every $f \in \mathcal{F}$, there exists a signed (Radon) measure μ such that

$$f(\cdot) = \int_\Omega \phi_\omega(\cdot)\,\mathrm{d}\mu(\omega),$$

where $\|\mu\|_{\mathcal{M}} < \infty$ and $\{\phi_\omega\}_{\omega\in\Omega}$ is a dictionary of L^∞-atoms.

$\Longrightarrow$ $f \in \mathcal{F}$ can be approximated (in L^2) with N-terms from the dictionary $\{\phi_\omega\}_{\omega\in\Omega}$ at a rate $N^{-\frac{1}{2}}$. (Maurey 1981)

$\Longrightarrow$ Such spaces are called **variation spaces**.

- The key idea here is **sparsity**.

$\Longrightarrow$ The $\mathcal{M}$-norm is the continuous-domain analogue of the ℓ^1-norm.

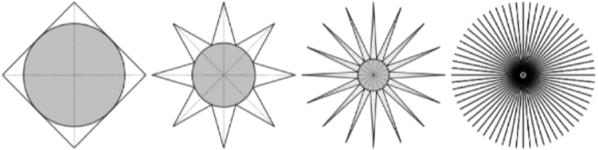
$\Longrightarrow$ Morally, $\mathcal{F}$ is an ℓ^1-type space and therefore has an interesting **geometry**.

The Geometry of Sparsity in High-Dimensions

- $d = 2$:

$\implies$ Misleading in high-dimensions!

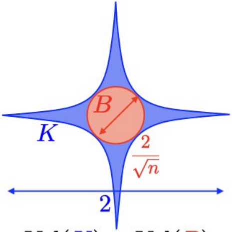
- ℓ^1-ball as d becomes large:



$\implies$ ℓ^1-balls become very "spikey" in high-dimensions.
$\implies$ High-dimensional ℓ^1-balls have **exponentially many tentacles** that grow in length as d becomes large.

The Geometry of Sparsity in High-Dimensions

Milman 1998 : high-dimensions $\Longrightarrow$ ℓ^1-balls look like hedgehogs.



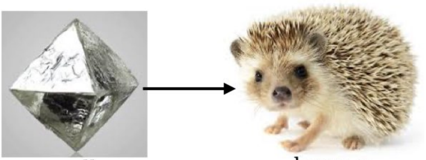
$\mathrm{Vol}(K) \sim \mathrm{Vol}(B)$

ℓ^1 ball:

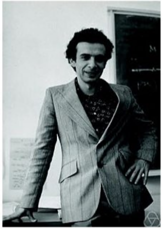
$$K \overset{\text{def.}}{=} \{x \in \mathbb{R}^n \,;\, \textstyle\sum_{i=1}^n |x_i| \leqslant 1\}$$

Bulk:

$$B \overset{\text{def.}}{=} \left\{x \in \mathbb{R}^n \,;\, \textstyle\sum_{i=1}^n |x_i|^2 \leqslant n^{-1}\right\}$$



small n large n



Approximation in Variation Spaces

- Let $\mathcal{F}$ be a variation space for the dictionary $\mathcal{D} := \{\phi_\omega\}_{\omega \in \Omega}$.
- Define

$$\Sigma_N := \Sigma_N(\mathcal{D}) := \left\{ \sum_{n=1}^{N} c_n \phi_{\omega_n} \, : \, \phi_{\omega_n} \in \mathcal{D} \right\}$$

- The **best N-term approximation** of $f \in \mathcal{F}$ from Σ_N is

$$\sigma_N(f)_{L^2} := \inf_{f_N \in \Sigma_N} \|f - f_N\|_{L^2}.$$

- This is **nonlinear approximation** since Σ_N is a nonlinear space:

$\Longrightarrow$ In general for $f, g \in \Sigma_N$, $f + g \in \Sigma_{2N}$.

Approximation in Variation Spaces

- From earlier, (Maurey 1981)

$$\sigma_N(f)_{L^2} \lesssim N^{-\frac{1}{2}}.$$

- This rate can be **improved** (Siegel and Xu 2022)

$$\sigma_N(f)_{L^2} \lesssim N^{-\frac{1}{2}-\frac{\alpha}{d}}.$$

$\implies$ $\alpha := \alpha(\mathcal{D})$ is the **smoothness constant** of $\mathcal{D}$.

- The improvement α/d captures the efficacy of **linear approximation methods**.

$\implies$ The best **linear approximation rate** typically scales as $N^{-\frac{\alpha}{d}}$.

$\implies$ Linear methods necessarily **suffer the curse of dimensionality**.

Examples of Variation Spaces

- $\mathscr{B}^s(\Omega)$ is a variation space for the dictionary

$$\{\boldsymbol{x} \mapsto (1 + |\boldsymbol{\xi}|)^{-s} e^{\mathrm{j}2\pi\boldsymbol{\xi}^\mathsf{T}\boldsymbol{x}}\}_{\boldsymbol{\xi}\in\mathbb{R}^d}$$

$\implies$ P. and Nowak 2022; Siegel and Xu 2023
$\implies$ $\sigma_N(f)_{L^2} \lesssim N^{-\frac{1}{2}-\frac{s}{d}}$ (i.e., $\alpha = s$).

- $\mathscr{R}\,\mathrm{BV}^k(\Omega)$ (BV-type space defined in the Radon domain) is a variation space for the dictionary

$$\{\boldsymbol{x} \mapsto (\boldsymbol{w}^\mathsf{T}\boldsymbol{x} - b)_+^{k-1}\}_{(\boldsymbol{w},b)\in\mathbb{S}^{d-1}\times\mathbb{R}}$$

$$\Updownarrow$$

ReLU^{k-1} neurons.

$\implies$ Ongie et al. 2020; P. and Nowak 2021, 2022, 2023
$\implies$ $\sigma_N(f)_{L^2} \lesssim N^{-\frac{1}{2}-\frac{2k-1}{2d}}$ (i.e., $\alpha = (2k-1)/2$).

- For these two examples, the rates are **sharp**.

Modulation Spaces

- Modulation spaces are smoothness spaces defined in the **short-time Fourier transform domain**.
- $M^s_{p,q}(\mathbb{R}^d)$ is the subspace of $\mathcal{S}'(\mathbb{R}^d)$ such that

$$\|f\|_{M^s_{p,q}} := \left(\int_{\mathbb{R}^d}\left(\int_{\mathbb{R}^d}|\mathrm{V}_g\{f\}(\boldsymbol{x},\boldsymbol{\xi})|^p(1+|(\boldsymbol{x},\boldsymbol{\xi})|)^{sp}\,\mathrm{d}\boldsymbol{x}\right)^{q/p}\mathrm{d}\boldsymbol{\xi}\right)^{1/q}$$

is finite.

$\Longrightarrow$ $\mathrm{V}_g\{f\}$ is the STFT of f with respect to the window $g \in \mathcal{S}(\mathbb{R}^d)$.

Modulation Spaces

- Modulation spaces stemmed from the work of Feichtinger 1981.

$\Longrightarrow$ $M^0_{1,1}(\mathbb{R}^d)$ is the smallest Segal algebra isometrically invariant under modulations.

- **Gabor/local Fourier/Wilson-type bases** are unconditional bases for the modulation spaces. (Feichtinger et al. 1992)

$\Longrightarrow$ $M^s_{1,1}(\mathbb{R}^d)$ is formed from functions that are ℓ^1-combinations of Gabor atoms.

$\Longrightarrow$ $M^s_{1,1}(\mathbb{R}^d)$ is a variation space!

Nonlinear Approximation in Modulation Spaces

- Define

$$\Sigma_N := \left\{ \sum_{n=1}^{N} c_n \psi_n : \ \psi_n \text{ is an element of a Gabor frame} \right\}$$

- Define the **best N-term approximation** of $f \in M^s_{1,1}(\mathbb{R}^d)$ from Σ_N as

$$\sigma_N(f)_{L^2} = \inf_{f_N \in \Sigma_N} \|f - f_N\|_{L^2}$$

- Again, this is **nonlinear approximation**.
- Many existing results on approximating $M^s_{p,q}(\mathbb{R}^d)$ functions with Gabor atoms.
  - $\Longrightarrow$ Gröchenig and Samarah 2000; Borup and Nielsen 2006; Borup and Nielsen 2007
  - $\Longrightarrow$ Many unresolved questions as well.
  - $\Longrightarrow$ Today, we will find several new results in the context of dimension-free nonlinear approximation rates in modulation spaces.

Main Results: Approximation Upper Bound

Theorem (P. and Unser 2023)

Let $s \geq 0$. For every $f \in M^s_{1,1}(\mathbb{R}^d)$,

$$\sigma_N(f)_{L^2} = \inf_{f_N \in \Sigma_N} \|f - f_N\|_{L^2} \lesssim N^{-\frac{1}{2}-\frac{s}{2d}}.$$

Furthermore, the approximant f_N that achieves this rate is found by thresholding the Gabor coefficients of f.

- Abstract result of Maurey 1981, gives the rate $N^{-\frac{1}{2}}$ for free.
- With some extra work, we get the **improved rate** $N^{-\frac{1}{2}-\frac{s}{2d}}$.
  - $\Longrightarrow$ Improved rate uncovers the role of s.
  - $\Longrightarrow$ Functions in $M^s_{1,1}(\mathbb{R}^d)$ for large s are **smoother** and hence **easier** to approximate.
- This rate is **immune to the curse of dimensionality**.

Main Results: Approximation Lower Bound

Theorem (P. and Unser 2023)

Let $s > 0$. For every $f \in M^s_{1,1}(\mathbb{R}^d)$,

$$\sigma_N(f)_{L^2} = \inf_{f_N \in \Sigma_N} \|f - f_N\|_{L^2} \gtrsim N^{-\frac{1}{2}-\frac{s}{2d}}.$$

- The requirement $s > 0$ arises since the result is proved using a technique based on **entropy**. (Carl 1981; Cohen et al. 2022)
  $\implies$ $M^s_{1,1}(\mathbb{R}^d) \subset\subset L^2(\mathbb{R}^d)$ iff $s > 0$. (Hinrichs et al. 2008)
- Rate achieved by thresholding is **sharp**: $\sigma_N(f)_{L^2} \asymp N^{-\frac{1}{2}-\frac{s}{2d}}$

Main Results: Suboptimality of Linear Methods

Theorem (P. and Unser 2023)

Let $s > 0$. Given $f \in M^s_{1,1}(\mathbb{R}^d)$. The best N-term **linear approximation** of f cannot achieve an approximation error that decays faster than $N^{-\frac{s}{2d}}$.

- Technically, we showed that the linear N-width of the unit ball in $M^s_{1,1}(\mathbb{R}^d)$ scales as $\asymp N^{-\frac{s}{2d}}$.

Transform-Domain Sparsity Breaks the Curse?

- $\mathscr{B}^s$: sparsity in the Fourier domain.
  $\Longrightarrow$ Nonlinear approximation rate: $N^{-\frac{1}{2}-\frac{s}{d}}$.
  $\Longrightarrow$ Linear approximation rate: $N^{-\frac{s}{d}}$.
- $\mathscr{R}\,\mathrm{BV}^k$: sparsity in the Radon domain.
  $\Longrightarrow$ Nonlinear approximation rate: $N^{-\frac{1}{2}-\frac{2k-1}{2d}}$.
  $\Longrightarrow$ Linear approximation rate: $N^{-\frac{2k-1}{2d}}$.
- $M^s_{1,1}$: sparsity in the STFT domain.
  $\Longrightarrow$ Nonlinear approximation rate: $N^{-\frac{1}{2}-\frac{s}{2d}}$.
  $\Longrightarrow$ Linear approximation rate: $N^{-\frac{s}{2d}}$.

Observations

- Sparsity in a transform domain "breaks" the curse of dimensionality for **nonlinear** approximation rates.
- **Linear** approximation methods **always** "suffer" the curse of dimensionality.
- Nonlinear methods are **required** to break the curse.

A Recipe for Breaking the Curse of Dimensionality

- Explicitly define a variation space $\mathcal{F}$ with respect to a dictionary $\mathcal{D}$.

  $\implies$ Best N-term nonlinear approximation rate from $\Sigma_N(\mathcal{D})$ is immune to the curse of dimensionality.

  $\implies$ Best N-term linear approximation rate from $\Sigma_N(\mathcal{D})$ suffers the curse of dimensionality.

Caveat

The space $\mathcal{F}$ is **already constructed** with the property that its N-term approximation rates are immune to the curse. Therefore, this result can be viewed as **boring**.

- Define a function space based on different kinds of **smoothness** and show that the spaces are equivalent to certain variation spaces.

  $\implies$ This was the story for $\mathscr{B}^s$, $\mathscr{R}\,\mathrm{BV}^k$, and $M^s_{1,1}$.

  $\implies$ Transform-domain sparsity often seems to work. Can we make this a **precise mathematical statement**?

Open Problems

- Further understanding of what analytic properties of functions leads to breaking the curse.
- Having a complete story for approximation theory with Gabor atoms.
  - $\Longrightarrow$ A complete characterization of the **approximation spaces** for Gabor frames.
- Bridging the gap between **mathematical statistics** and **Gabor analysis**.
  - $\Longrightarrow$ Some preliminary work in this direction: Dahlke et al. 2022.
  - $\Longrightarrow$ A complete understanding of approximation theory in the Gabor analysis setting is the first step towards bringing **finite data** to the problem.
  - $\Longrightarrow$ A story similar to wavelets, Besov spaces, and nonparametric statistics in the Gabor analysis setting would be nice.

References I

Barron, Andrew R. (1993). "Universal approximation bounds for superpositions of a sigmoidal function". In: **IEEE Transactions on Information theory** 39.3, pp. 930–945.

Bellman, Richard E. (1961). **Adaptive Control Processes: A Guided Tour**. Princeton University Press.

Birman, Mikhail Shlemovich and Mikhail Zakharovich Solomyak (1967). "Piecewise-polynomial approximations of functions of the classes W_p^α". In: **Matematicheskii Sbornik** 115.3, pp. 331–355.

Borup, Lasse and Morten Nielsen (2006). "Nonlinear approximation in α-modulation spaces". In: **Mathematische Nachrichten** 279.1-2, pp. 101–120.

— (2007). "Frame decomposition of decomposition spaces". In: **Journal of Fourier Analysis and Applications** 13.1, pp. 39–70.

Carl, Bernd (1981). "Entropy numbers, s-numbers, and eigenvalue problems". In: **Journal of Functional Analysis** 41.3, pp. 290–306.

Cohen, Albert, Ronald DeVore, Guergana Petrova, and Przemyslaw Wojtaszczyk (2022). "Optimal stable nonlinear approximation". In: **Foundations of Computational Mathematics** 22.3, pp. 607–648.

References II

- Dahlke, Stephan, Sven Heuer, Hajo Holzmann, and Pavel Tafo (2022). "Statistically optimal estimation of signals in modulation spaces using Gabor frames". In: **IEEE Transactions on Information Theory** 68.6, pp. 4182–4200.
- DeVore, Ronald A. (1998). "Nonlinear approximation". In: **Acta numerica** 7, pp. 51–150.
- Donoho, David L. (2000). "High-dimensional data analysis: The curses and blessings of dimensionality". In: **AMS math challenges lecture** 1.2000, p. 32.
- Donoho, David L. and Iain M. Johnstone (1998). "Minimax estimation via wavelet shrinkage". In: **The Annals of Statistics** 26.3, pp. 879–921.
- Feichtinger, Hans G. (1981). "On a new Segal algebra". In: **Monatshefte für Mathematik** 92, pp. 269–289.
- Feichtinger, Hans G, Karlheinz Gröchenig, and D Walnut (1992). "Wilson bases and modulation spaces". In: **Mathematische Nachrichten** 155.1, pp. 7–17.
- Gröchenig, Karlheinz and Salti Samarah (2000). "Nonlinear approximation with local Fourier bases". In: **Constructive Approximation** 16, pp. 317–331.

References III

- Hinrichs, Aicke, Iwona Piotrowska, and Mariusz Piotrowski (2008). "On the degree of compactness of embeddings between weighted modulation spaces". In: **Journal of Function Spaces and Applications** 6.3, pp. 303–317.
- Milman, Vitali (1998). "Surprising geometric phenomena in high-dimensional convexity theory". In: **European Congress of Mathematics: Budapest, July 22–26, 1996 Volume II**. Springer, pp. 73–91.
- Ongie, Greg, Rebecca Willett, Daniel Soudry, and Nathan Srebro (2020). "A Function Space View of Bounded Norm Infinite Width ReLU Nets: The Multivariate Case". In: **International Conference on Learning Representations**.
- Parhi, Rahul and Robert D. Nowak (2021). "Banach space representer theorems for neural networks and ridge splines". In: **Journal of Machine Learning Research** 22.43, pp. 1–40.
- — (2022a). "On Continuous-Domain Inverse Problems with Sparse Superpositions of Decaying Sinusoids as Solutions". In: **IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)**, pp. 5603–5607.

References IV

Parhi, Rahul and Robert D. Nowak (2022b). "What kinds of functions do deep neural networks learn? Insights from variational spline theory". In: **SIAM Journal on Mathematics of Data Science** 4.2, pp. 464–489.

— (2023). "Near-Minimax Optimal Estimation With Shallow ReLU Neural Networks". In: **IEEE Transactions on Information Theory** 69.2, pp. 1125–1140.

Pisier, Gilles (1981). "Remarques sur un résultat non publié de B. Maurey". In: **Séminaire d'Analyse Fonctionnelle (dit "Maurey-Schwartz")**, pp. 1–12.

Siegel, Jonathan W. and Jinchao Xu (2022). "Sharp bounds on the approximation rates, metric entropy, and n-widths of shallow neural networks". In: **Foundations of Computational Mathematics**, pp. 1–57.

— (2023). "Characterization of the variation spaces corresponding to shallow neural networks". In: **Constructive Approximation**, pp. 1–24.