UniGe | MaLGa

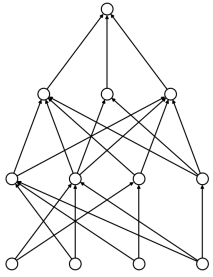
Learning Multi-Index Models with Hyper-Kernel Ridge Regression

Shuo Huang
MaLGa; Istituto Italiano di Tecnolog, Italy

Hippolyte Labarrière, Ernesto De Vito, Lorenzo Rosasco (MalGa), Tomaso Poggio (MIT)

Oct. 6, 2025 — NPTwins 2025, Messina

Why deep?

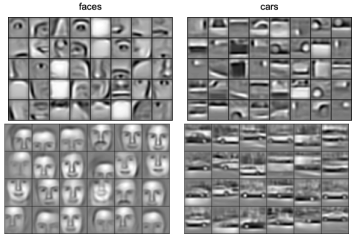


Exploring principles

Invariance

[Fukushima, 1980; LeCun et al., 1995],· · ·,[Mallat, 2016]

Compositionality



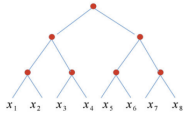
[Geman et al., 2002; Serre et al., 2005], · · ·,[Schmidt-Hieber, 2020]

(Sparse) Compositionality

$$f = f_\ell \circ \cdots f_2 \circ f_1, \quad f_j \in \mathcal{F}(\mathbb{R}^{d_j})$$

Spare tree structure

$$f(x^1, \cdots, x^8) = h_7\big(h_5((h_1(x^1, x^2), h_2(x^3, x^4))), h_6((h_3(x^5, x^6), h_4(x^7, x^8)))\big)$$



[Poggio et al., 2017; Mhaskar et al., 2017; Schmidt-Hieber, 2020]

Multi-index Model (MIM)

$$f(x) = g \circ B(x)$$

Outline

Learning theory and curses and blessings

Multi-index model and hyper-kernels

Optimization and numerical results

Statistical learning

Let P be a probability measure on $\mathbb{R}^D \times \mathbb{R}$, and for all $f \in \mathcal{M}(\mathbb{R}^D, \mathbb{R})$

$$L(f) = \int (y - f(x))^2 dP(x, y).$$

Problem: solve

$$\min_{f \in \mathcal{M}(\mathbb{R}^D, \mathbb{R})} L(f)$$

given $(x_1, y_1), \ldots, (x_n, y_n) \sim P^n$.

Empirical risk minimization (ERM)

$$\min_{f \in \mathcal{M}(\mathbb{R}^D, \mathbb{R})} L(f) \quad \mapsto \quad \min_{f \in \mathcal{H}} \widehat{L}(f)$$

where $\mathcal{H} \subset \mathcal{M}(\mathbb{R}^D, \mathbb{R})$ and

$$\widehat{L}(f) = \frac{1}{n} \sum_{i=1}^{n} (y_i - f(x_i))^2.$$

Choices for $\mathcal{H}$

- Neural nets

$$f(x) = \sum_{i=1}^{u} c_i \sigma(w_i^\top x), \qquad \sigma(a) = \max\{0, a\}.$$

[McCulloch, Pitts '43, Rosenblatt '58, Rumelhart et al. '86]

- Kernel methods

$$f(x) = \sum_{i=1}^{u} c_i K(x_i, x), \qquad K \text{ positive definite (PD)}.$$

[Smola, Schölkopf '02]

Kernel ridge regression (KRR)

$$\min_{f \in \mathcal{H}} \widehat{L}(f) + \lambda \|f\|_{\mathcal{H}}^2$$

Representer theorem: The unique minimizer can be written as

$$\widehat{f}(x) = \sum_{i=1}^{n} K(x, x_i) c_i$$

with

$$c = (\widehat{K} + \lambda I)^{-1} \widehat{y} \qquad \widehat{K}_{ij} = K(x_i, x_j) \qquad \widehat{y} = [y_1, \ldots, y_n]$$

[Kimeldorf, Wahba, '70]

Curse of dimensionality and blessing of smoothness

Assume $|y| \leq M$ a.s. and

$$L(f_*) = \min_{f \in \mathcal{M}(\mathbb{R}^D, \mathbb{R})} L(f)$$

with $f^* \in H^s(\mathbb{R}^D)$, $s \geq D/2 + 1/2$.

1-layer neural nets with appropriate width and KRR with appropriate λ both satisfy whp

$$L(\widehat{f}) - L(f_*) \lesssim n^{-\frac{2s}{2s+D}} \sim \epsilon.$$

or

$$n \sim \epsilon^{-\frac{2s+D}{2s}}.$$

[Barron '94, Caponetto. De Vito '05]

Remarks

$$n^{-\frac{2s}{2s+D}}$$

- Optimal for $f^* \in H^s(\mathbb{R}^D)$ [Stone '82]
- ...but does not distinguish kernel methods and NN.
- New smoothness classes? [Barron '93]

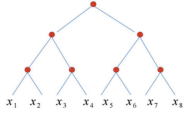
Compositional sparsity

$$f = h_L \circ \cdots h_2 \circ h_1, \quad h_j \in H^{s_j}(\mathbb{R}^{d_j})$$

e.g. [Mhaskar, Poggio '16, Schmidt-Hieber '20, Dhamen '22]

Tree sparse composition

$$f(x^1, \cdots, x^8) = h_7\big(h_5((h_1(x^1, x^2), h_2(x^3, x^4))), h_6((h_3(x^5, x^6), h_4(x^7, x^8)))\big)$$



Multi-Index model

$$f(x) = g(Bx)$$

Outline

Learning theory and curses and blessings

Multi-index model and hyper-kernels

Optimization and numerical results

Multi-index model (MIM)

$$f_*(x) = g_*(B_*x)$$

where

- $B_* \in \mathbb{R}^{d_* \times D}$, and
- $g_* : \mathbb{R}^{d_*} \to \mathbb{R}$.

Some context

$$f_*(x) = g_*(B_*x)$$

- Sufficient dimension reduction

  e.g. [Li, '91, Fukumizu, Bach, Jordan '09]

- Nonparametric statistics

  e.g. [Härdle et al. '93 , Klock et al. '00]

- Representation learning

  [Bach, '17, Bietti et al. '24, Damian et al. '24]

Hyper kernels

$\mathcal{H}_B$ RKHS with kernel $K_B : \mathbb{R}^D \times \mathbb{R}^D \to \mathbb{R}$

$$K_B(x, x') = K(Bx, Bx')$$

- $B \in \mathbb{R}^{d \times D}$.
- $K : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ pos. def. kernel, $\mathcal{H}_K$ RKHS on $\mathbb{R}^d$.

see also Hyper-RBF [Brunelli, Poggio '92] RFM [Radhakrishnan et al. '22]

Hyper-kernel ridge regression (HKRR)

$$\min_{\|B\| \leq 1} \min_{f \in \mathcal{H}_B} \frac{1}{n} \sum_{i=1}^{n} (f(x_i) - y_i)^2 + \lambda \|f\|^2_{\mathcal{H}_B}.$$

Q1 (Statistics): Sample complexity?

Q2 (Optimization): Compute the solution?

Sample complexity of HKRR

Assume $|y| \leq M$ a.s. and

$$L(f_*) = \min_{f \in \mathcal{M}(\mathbb{R}^D, \mathbb{R})} L(f)$$

with $g^* \in H_K$, $K_x \in C^s(\mathbb{R}^{d_*})$

HKRR with $\lambda = n^{-\frac{2s}{2s+2d_*}}$ satisfy whp

$$L(\widehat{f}) - L(f_*) \lesssim C_D n^{-\frac{2s}{2s+2d_*}}.$$

Remarks

$$C_D n^{-\frac{2s}{2s+2d_*}} \quad \text{vs} \quad n^{-\frac{2s}{2s+D}}$$

- Beyond the curse of dimensionality.
- Same result tuning by hold-out cross-validation.
- Suboptimal bound.
- Nyström …

Proof sketch

Error decomposition:

$$
\begin{aligned}
& L(\widehat{f}_{\lambda,B}) - L(f_*) \\
& \leq \left\{ L(\widehat{f}_{\lambda,B}) - L(f_*) - (\hat{L}(\widehat{f}_{\lambda,B}) - \hat{L}(f_*)) \right\} + \left\{ \hat{L}(f_{\lambda,B_*}) - \hat{L}(f_*) - (L(f_{\lambda,B_*}) - L(f_*)) \right\} \\
& \quad + \left\{ L(f_{\lambda,B_*}) - L(f_*) + \lambda \|f_\lambda^{B_*}\|_{B_*}^2 \right\}
\end{aligned}
$$

Sample error estimate: Key technique L_∞-covering of the composite classes class $g \circ B$.
Approximation error estimate

[Cucker, Zhou '07]

Outline

Learning theory and curses and blessings

Multi-index model and hyper-kernels

Optimization and numerical results

Hyper-kernel ridge regression (HKRR)

$$\min_{\|B\| \leq 1} \min_{f \in \mathcal{H}_B} \frac{1}{n} \sum_{i=1}^{n} (f(x_i) - y_i)^2 + \lambda \|f\|^2_{\mathcal{H}_B}.$$

Q1 (Statistics): Sample complexity?

Q2 (Optimization): Compute the solution?

Representer theorem

$$\min_{\|B\|\leq 1} \min_{f\in\mathcal{H}_B} \frac{1}{n}\sum_{i=1}^{n}(f(x_i)-y_i)^2+\lambda\|f\|^2_{\mathcal{H}_B}.$$

$$\Downarrow$$

$$\min_{\|B\|\leq 1} \min_{c\in\mathbb{R}^n} \frac{1}{n}\left\|\widehat{K}_B c-\mathbf{y}\right\|^2+\lambda c^T\widehat{K}_B c$$

$$\min_{\|B\|\leq 1} \min_{c\in\mathbb{R}^M} \underbrace{\frac{1}{n}\left\|\widehat{K}_B c - \mathbf{y}\right\|^2 + \lambda c^T \widehat{K}_B c}_{\widehat{L}_\lambda(B,\alpha)},$$

$$\min_{\|B\|\leq 1} \min_{c\in\mathbb{R}^M} \underbrace{\frac{1}{n}\left\|\widehat{K}_B c - \mathbf{y}\right\|^2 + \lambda c^T \widehat{K}_B c}_{\widehat{L}_\lambda(B,\alpha)},$$

VarPro

$$\begin{cases} c_t = \arg\min_{c\in\mathbb{R}^M} \widehat{L}_\lambda(B_{t-1}, c) \\ B_t = \mathrm{P}_{\|B\|\leq 1}\left(B_{t-1} - \eta_B \nabla_B \widehat{L}_\lambda(B_{t-1}, c_t)\right) \end{cases}$$

AGD

$$\begin{cases} c_t = c_{t-1} + \eta_c \nabla_c \widehat{L}_\lambda(B_{t-1}, c_{t-1}) \\ B_t = \mathrm{P}_{\|B\|\leq 1}\left(B_{t-1} - \eta_B \nabla_B \widehat{L}_\lambda(B_t, c_t)\right) \end{cases}$$

Algorithm 1: VarPro (informal)

Require: B^0, $s_B > 0$

$\alpha^0 = \arg\min_{\alpha\in\mathbb{R}^m} \mathcal{L}(B^0, \alpha)$

for $i = 0, 1, \ldots$ **do**

  $B^{i+1} = B^i - s_B \nabla_B \hat{\mathcal{L}}(B^i, \alpha^i)$

  $\alpha^{i+1} = \arg\min_{\alpha\in\mathbb{R}^m} \hat{\mathcal{L}}(B^{i+1}, \alpha)$

end for

return (B^{i+1}, α^{i+1})

Algorithm 2: AGD (informal)

Require: B^0, α^0, $s_\alpha > 0$, $s_B > 0$, $n_\alpha \in \mathbb{N}^*$

for $i = 0, 1, \ldots$ **do**

  $B^{i+1} = B^i - s_B \nabla_B \hat{\mathcal{L}}(B^i, \alpha^i)$

  $\alpha^{i,0} = \alpha^i$

  for $j = 0, 1, \ldots, n_\alpha - 1$ **do**

    $\alpha^{i,j+1} = \alpha^{i,j} - s_\alpha \nabla_\alpha \hat{\mathcal{L}}(B^{i+1}, \alpha^{i,j})$

  end for

  $\alpha^{i+1} = \alpha^{i,n_\alpha}$

end for

return (B^{i+1}, α^{i+1})

$$\min_{\|B\|\leq 1} \min_{c\in\mathbb{R}^M} \underbrace{\frac{1}{n}\left\|\widehat{K}_B c - \mathbf{y}\right\|^2 + \lambda c^T \widehat{K}_B c}_{\widehat{L}_\lambda(B,\alpha)},$$

VarPro

$$\begin{cases} c_t = \arg\min_{c\in\mathbb{R}^M} \widehat{L}_\lambda(B_{t-1}, c) \\ B_t = \mathrm{P}_{\|B\|\leq 1}\left(B_{t-1} - \eta_B \nabla_B \widehat{L}_\lambda(B_{t-1}, c_t)\right) \end{cases}$$

AGD

$$\begin{cases} c_t = c_{t-1} + \eta_c \nabla_c \widehat{L}_\lambda(B_{t-1}, c_{t-1}) \\ B_t = \mathrm{P}_{\|B\|\leq 1}\left(B_{t-1} - \eta_B \nabla_B \widehat{L}_\lambda(B_t, c_t)\right) \end{cases}$$

Algorithm 1: VarPro (informal)

Require: B^0, $s_B > 0$

$\alpha^0 = \arg\min_{\alpha\in\mathbb{R}^{n_h}} \mathcal{L}(B^0, \alpha)$

for $i = 0, 1, \ldots$ **do**

  $B^{i+1} = B^i - s_B \nabla_B \hat{\mathcal{L}}(B^i, \alpha^i)$

  $\alpha^{i+1} = \arg\min_{\alpha\in\mathbb{R}^{n_h}} \hat{\mathcal{L}}(B^{i+1}, \alpha)$

end for

return (B^{i+1}, α^{i+1})

Algorithm 2: AGD (informal)

Require: B^0, α^0, $s_\alpha > 0$, $s_B > 0$, $n_\alpha \in \mathbb{N}^*$

for $i = 0, 1, \ldots$ **do**

  $B^{i+1} = B^i - s_B \nabla_B \hat{\mathcal{L}}(B^i, \alpha^i)$

  $\alpha^{i,0} = \alpha^i$

  for $j = 0, 1, \ldots, n_\alpha - 1$ **do**

    $\alpha^{i,j+1} = \alpha^{i,j} - s_\alpha \nabla_\alpha \hat{\mathcal{L}}(B^{i+1}, \alpha^{i,j})$

  end for

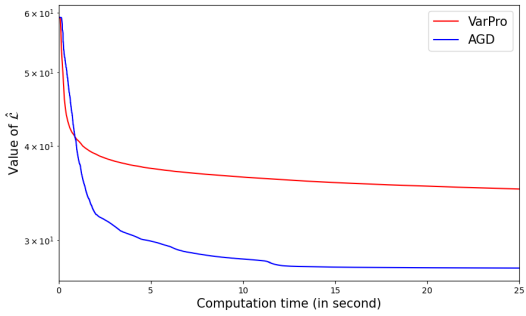
  $\alpha^{i+1} = \alpha^{i,n_\alpha}$

end for

return (B^{i+1}, α^{i+1})

Both algorithms converge to a critical point.

VarPro V.S. AGD

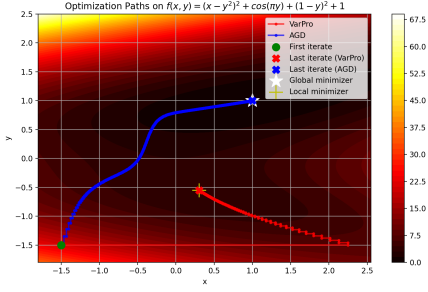


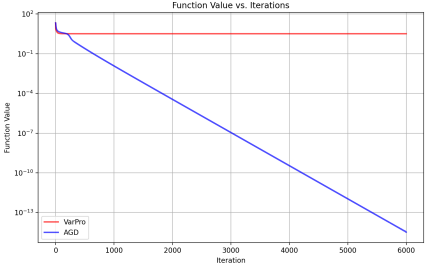
VarPro: **faster convergence, trapped in local minima**

2D visualization

$$f : x, y \mapsto (x - y^2)^2 + \cos(\pi y) + (1 - y)^2.$$

$(x_0, y_0) = (-1.5, -1.5)$

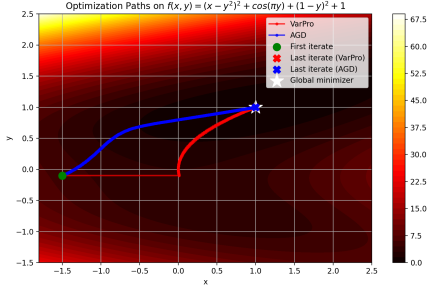


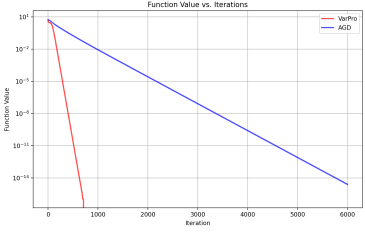


2D visualization

$$f : x, y \mapsto (x - y^2)^2 + \cos(\pi y) + (1 - y)^2.$$

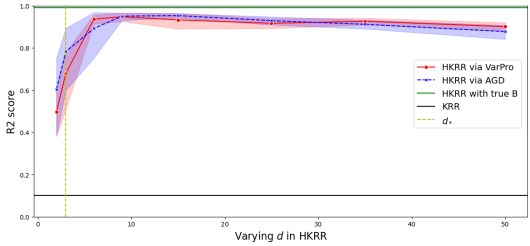
$(x_0, y_0) = (-1.5, -0.1)$





The role of latent dimension d

$$d_* = 3, D = 50$$



Wrapping up

- Depth and compositional sparsity.
- MIM with HKRR: sample complexity.
- HKRR optimization.

What next?

- Better bounds? SGD?
- Compositional sparsity: non linear MIM, deeeeep.

Thanks for listening!

Bibliography I

Kunihiko Fukushima. Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biological cybernetics*, 36(4):193–202, 1980.

Yann LeCun, Yoshua Bengio, et al. Convolutional networks for images, speech, and time series. 1995.

Stéphane Mallat. Understanding deep convolutional networks. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 374(2065):20150203, 2016.

Stuart Geman, Daniel F Potter, and Zhiyi Chi. Composition systems. *Quarterly of Applied Mathematics*, 60(4): 707–736, 2002.

Thomas Serre, Lior Wolf, and Tomaso Poggio. Object recognition with features inspired by visual cortex. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, volume 2, pages 994–1000. Ieee, 2005.

Johannes Schmidt-Hieber. Nonparametric regression using deep neural networks with relu activation function. *The Annals of Statistics*, 48(4):1875–1897, 2020.

Tomaso Poggio, Hrushikesh Mhaskar, Lorenzo Rosasco, Brando Miranda, and Qianli Liao. Why and when can deep-but not shallow-networks avoid the curse of dimensionality: a review. *International Journal of Automation and Computing*, 14(5):503–519, 2017.

Hrushikesh Mhaskar, Qianli Liao, and Tomaso Poggio. When and why are deep networks better than shallow ones? In *Proceedings of the AAAI conference on artificial intelligence*, volume 31, 2017.