Fundamental limits of learning with equivariant algorithms

Theodor Misiakiewicz

Yale University

August 8th, 2025

Statistical Physics & Machine Learning: moving forward (Cargese 2025)

Joint work with **Hugo Koubbi** (ENS/Yale), **Hugo Latourelle-Vigeant** (Yale), **Nirmit Joshi** (TTIC), and **Nati Srebro** (TTIC).

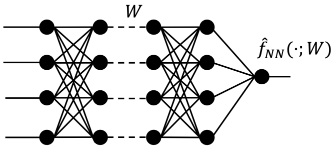
Complexity of gradient-based learning

- Source distribution $(y, x) \sim \mathcal{D}$ over $\mathcal{Y} \times \mathcal{X}$:

  Goal: fit a predictor $f : \mathcal{X} \to \mathbb{R}$ that minimizes a *population loss*

$$\mathcal{L}_{\mathcal{D}}(f) := \mathbb{E}_{(y,x)\sim\mathcal{D}}[\ell(y, f(x))].$$

- **Modern approach:** SGD (or variants) on parametrized model $f : \mathcal{X} \times \mathcal{W} \to \mathbb{R}$



$$W^{t+1} = W^t - \nabla_W \ell(y_t, f(x_t; W^t))$$

High-dim dynamics

- A major theme of modern ML/statistics: computational bottlenecks

Computational-to-statistical gaps.
Sample-runtime trade-offs.

What are the fundamental limits of learning with gradient-based algorithms?

- Goal: make some (modest) progress on this question. Ideally, the theory should:
  - explain some of the empirical phenomenology
  - describe some of the stat/computational trade-offs of gradient algo...
  - ...while capturing some fundamental hardness properties (not be too sensitive to design choices or hyperparameters).

- Here, we focus on a specific property of 'generic' gradient-type algorithms:

  Equivariance with respect to a large symmetry group.

Equivariant algorithms

- Source distribution $(y, x) \sim \mathcal{D}$ over $\mathcal{Y} \times \mathcal{X}$. **Goal:** fit a predictor $f : \mathcal{X} \to \mathbb{R}$ that minimizes a *population loss*

$$\mathcal{L}_{\mathcal{D}}(f) := \mathbb{E}_{(y,x)\sim\mathcal{D}}[\ell(y, f(x))].$$

- Learning algorithm $\mathcal{A}$ takes source $\mathcal{D}$ and outputs a predictor $\mathcal{A}(\mathcal{D}) : \mathcal{X} \to \mathbb{R}$

$$\mathcal{R}_{\mathcal{D}}(\mathcal{A}) = \mathbb{E}_{\mathcal{A}}[\mathcal{L}_{\mathcal{D}}(\mathcal{A}(\mathcal{D}))] = \mathbb{E}_{\mathcal{A}}\Big[\ \mathbb{E}_{(y,x)\sim\mathcal{D}}[\ell(y, \mathcal{A}(\mathcal{D})(x))]\ \Big].$$

- Group $\mathcal{G}$ of transformations $g : \mathcal{X} \to \mathcal{X}$

$$\mathcal{D}^g \text{ distribution of } (y, g \cdot x) \text{ with } (y, x) \sim \mathcal{D}.$$

- $\mathcal{A}$ is $\mathcal{G}$-equivariant if for all $g \in \mathcal{G}$

$$\mathcal{A}(\mathcal{D}^g) \circ g \overset{\mathrm{d}}{=} \mathcal{A}(\mathcal{D}).$$

  E.g., SGD on FCNNs with Gaussian initialization: rotationally equivariant.
  Adam/AdaGrad/ℓ_1-norm: permutation equivariant.

Orbit class of distributions

- If $\mathcal{A}$ is $\mathcal{G}$-equivariant:

$$\mathcal{R}_{\mathcal{D}^g}(\mathcal{A}) = \mathbb{E}_{\mathcal{A}}\Big[\, \mathbb{E}_{(y,x)\sim\mathcal{D}}[\ell(y, \mathcal{A}(\mathcal{D}^g)(g \cdot x))] \,\Big] = \mathcal{R}_{\mathcal{D}}(\mathcal{A}), \qquad \forall g \in \mathcal{G}.$$

- $\mathcal{A}$ learns $\mathcal{D}$ iff it learns the entire orbit

$$\mathcal{D}[\mathcal{G}] := \{\mathcal{D}^g \;:\; g \in \mathcal{G}\}.$$

- Learning $\mathcal{D}[\mathcal{G}]$ $\iff$ Learning $\mathcal{D}$ with $\mathcal{G}$-equivariant algos.
  - Lower bound $\implies$ lower bound on learning $\mathcal{D}$ with $\mathcal{G}$-equivariant algo.
  - Upper bound $\implies$ algo can be randomized to make it $\mathcal{G}$-equivariant.

What is the complexity of learning $\mathcal{D}[\mathcal{G}]$?

- Previous works have exploited equivariance to show LBs on optimization algo [Ng, '04], [Shamir, '18], [Li, Zhang, Arora, '21], [Abbe, Boix-Adsera, '22]

Our work

Group-theoretic characterization of the complexity of learning $\mathcal{D}[\mathcal{G}]$.

- Outline:
  - Most of the talk: the example of learning single-index models.
  - Learning multi-index models.
  - Weak learning of $\mathcal{D}[\mathcal{G}]$.
  - Strong learning of $\mathcal{D}[\mathcal{G}]$.

1 Learning Single-Index Models

Gaussian Single-Index Models

- Distribution $\mathcal{D} := \mathcal{D}_{\boldsymbol{w}_*}$ indexed by $\boldsymbol{w}_* \in \mathbb{S}^{d-1}$

$$(y, \boldsymbol{x}) \sim \mathcal{D} : \quad \boldsymbol{x} \sim \mathsf{N}(0, \mathbf{I}_d), \quad y|\boldsymbol{x} \sim \rho(\cdot|\langle \boldsymbol{w}_*, \boldsymbol{x}\rangle).$$

- Consider $\mathcal{G} = \mathcal{O}_d$ the orthogonal group in d-dimension: for $g \in \mathcal{O}_d$,

$$(y, \boldsymbol{x}) \sim \mathcal{D}^g : \qquad \boldsymbol{x} \sim \mathsf{N}(0, \mathbf{I}_d), \quad y|\boldsymbol{x} \sim \rho(\cdot|\langle g \cdot \boldsymbol{w}_*, \boldsymbol{x}\rangle),$$

  so that $\mathcal{D}^g_{\boldsymbol{w}_*} = \mathcal{D}_{g\cdot \boldsymbol{w}_*}$.

- $\mathcal{O}_d$-equivariant algorithms learn $\mathcal{D}_{\boldsymbol{w}_*}$ if and only if they learn

$$\mathcal{D}[\mathcal{O}_d] = \{\mathcal{D}_{\boldsymbol{w}} \ : \ \boldsymbol{w} \in \mathbb{S}^{d-1}\}.$$

..., [Barbier, Krzakala, Macris, Miolane, Zdeborova,'19], [Mondelli, Montanari,'18], [Lu, Li,'20], [Ben Arous, Gheissari, Jagannath,'21], [Mousavi-Hossein, Park, Girotti, Mitliagkas, Erdogdu,'22], [Bietti, Bruna, Sanford, Song, '22], [Veiga, Stephan, Loureiro, Krzakala, Zdeborova,'22], [Damian, Nichani, Ge, Lee, '23], [Damian, Pillaud-Vivien, Lee, Bruna, '24], [Lee, Oko, Suzuki, Wu,'24], [Arnaboldi, Dandi, Krzakala, Loureiro, Pesce, Stephan,'24], [Chen, Wu, Lu, Yang, Wang, '24], ...

Learning Gaussian SIMs

Given m iid data $(y_i, \boldsymbol{x}_i) \sim \mathcal{D}$:

$$(y, \boldsymbol{x}) \sim \mathcal{D} : \quad \boldsymbol{x} \sim \mathsf{N}(0, \mathbf{I}_d), \quad y|\boldsymbol{x} \sim \rho(\cdot|\langle \boldsymbol{w}_*, \boldsymbol{x}\rangle),$$

for some unknown $\boldsymbol{w}_*$, compute $\hat{\boldsymbol{w}}$ such that with probability at least $1 - \delta$,

$$|\langle \boldsymbol{w}_*, \hat{\boldsymbol{w}}\rangle| \geq 1 - \varepsilon. \tag{$\star$}$$

- What are the optimal

  m : sample-size and T : runtime

  to solve $(\star)$?

  - Information theoretically $\mathsf{m} = \Theta(d/\varepsilon)$ is always optimal. In this talk:

    sample-optimal = optimal sample-size to solve $(\star)$ in polynomial time.

Sharp characterization

- [Barbier et al.,'17], [Lu, Li,'17], [Mondelli, Montanari,'18] ($\mathsf{k}_\star = 1, 2$)
  [Damian, Pillaud-Vivien, Lee, Bruna,'24] ($\mathsf{k}_\star \geq 3$)

  $$\mathsf{m} = \Theta_d(d^{\max(\mathsf{k}_\star/2,1)}), \qquad \mathsf{T} = \widetilde{\Theta}_d(d^{\max(\mathsf{k}_\star/2,1)+1}).$$

  where $\mathsf{k}_\star$ = "generative exponent" of ρ. (SQ and LDP lower bounds.)

- Several works have progressively close the gap to these optimal rates ($\mathsf{k}_\star \geq 2$):
  - Online SGD [Ben Arous, Gheissari, Jagannath, '21]:

    $$\mathsf{m} = \widetilde{\Theta}_d(d^{\mathsf{k}_\star - 1}), \qquad \mathsf{T} = \widetilde{\Theta}_d(d^{\mathsf{k}_\star}).$$

  - Landscape smoothing [Damian, Nichani, Ge, Lee,'23]:

    $$\mathsf{m} = \widetilde{\Theta}_d(d^{\mathsf{k}_\star/2}), \qquad \mathsf{T} = \widetilde{\Theta}_d(d^{\mathsf{k}_\star/2+1}).$$

  - Partial trace estimator [Damian, Pillaud-Vivien, Lee, Bruna,'24]:

    $$\mathsf{m} = \Theta_d(d^{\mathsf{k}_\star/2}), \qquad \mathsf{T} = \widetilde{\Theta}_d(d^{\mathsf{k}_\star/2+1}).$$

Online SGD algorithm

- [Ben Arous, Gheissari, Jagannath, '21] Online SGD on population loss

$$\mathcal{L}(\boldsymbol{w}) = \frac{1}{2}\mathbb{E}_{(y,\boldsymbol{x})\sim\mathbb{P}_{\boldsymbol{w}*}}\left[\Big(y - \sigma(\langle \boldsymbol{w}, \boldsymbol{x}\rangle)\Big)^2\right]$$

- Information exponent:

$$\mathsf{k}_\mathsf{I} := \arg\min\{k \geq 1 \ : \ \mu_k(y) = \mathbb{E}_\rho[Y\mathrm{He}_k(G)] \neq 0\}.$$

  So that $\mathcal{L}(\boldsymbol{w}) = \mathcal{L}_* - \Theta(\langle \boldsymbol{w}, \boldsymbol{w}_*\rangle^{\mathsf{k}_\mathsf{I}})$.

- Initialization $\boldsymbol{w}_0 \sim \mathrm{Unif}(\mathbb{S}^{d-1})$, we have $\langle \boldsymbol{w}_*, \nabla\mathcal{L}(\boldsymbol{w}_0)\rangle = \Theta_{d,\mathbb{P}}(d^{-(\mathsf{k}_\mathsf{I}-1)/2})$.

- [Ben Arous, Gheissari, Jagannath, '21] # of SGD iterations (= # of samples)

$$\mathsf{m} = \begin{cases} \Theta(d) & \text{if } \mathsf{k}_\mathsf{I} = 1, \\ \widetilde{\Theta}(d^{\mathsf{k}_\mathsf{I}-1}) & \text{if } \mathsf{k}_\mathsf{I} > 1. \end{cases}$$

  Total runtime: $\mathsf{T} = \Theta_d(\mathsf{m}d) = \widetilde{\Theta}_d(d^{\max(\mathsf{k}_\mathsf{I},2)})$.

Generative exponent

- Are $\mathsf{m} = \widetilde{\Theta}_d(d^{\max(\mathsf{k}_\mathsf{I}-1,1)})$ and $\mathsf{T} = \widetilde{\Theta}_d(d^{\max(\mathsf{k}_\mathsf{I},2)})$ optimal to learn SIM?
- We can do much better if we:
  - Reuse samples [Dandi, Troiani, Arnaboldi, Pesce, Zdeborová, Krzakala,'24], [Lee, Oko, Suzuki, Wu,'24], [Arnaboldi, Dandi, Krzakala, Loureiro, Pesce, Stephan,'24]
  - Change loss function [Joshi, **M.**, Srebro, '24]
  - Apply a transformation $\mathcal{T}(y)$ to the label [Damian, Pillaud-V, Lee, Bruna,'24]
- [Damian, Pillaud-Vivien, Lee, Bruna,'24] Generative exponent of ρ:

$$\mathsf{k}_\star := \arg\min\{k \geq 1 : \exists \mathcal{T} : \mathcal{Y} \to \mathbb{R},\ \mu_k(\mathcal{T}(y)) = \mathbb{E}_\rho[\mathcal{T}(Y)\mathrm{He}_k(G)] \neq 0\},$$

and showed that (optimal within SQ and LDP):

$$\mathsf{m} = \Theta_d(d^{\max(\mathsf{k}_\star/2,1)}), \qquad \mathsf{T} = \widetilde{\Theta}_d(d^{\max(\mathsf{k}_\star/2,1)+1}).$$

Online SGD algorithm suboptimal

- Online SGD on $\mathcal{L}(\boldsymbol{w}) = \frac{1}{2}\mathbb{E}[(\mathcal{T}(y) - \sigma(\langle \boldsymbol{w}, \boldsymbol{x} \rangle))^2]$:

$$\mathsf{m} = \widetilde{\Theta}_d(d^{\max(\mathsf{k}_\star - 1, 1)}), \qquad \mathsf{T} = \widetilde{\Theta}_d(d^{\max(\mathsf{k}_\star, 2)}).$$

- Suboptimal compared to $\mathsf{m} = \Theta_d(d^{\max(\mathsf{k}_\star/2, 1)})$ or $\mathsf{T} = \widetilde{\Theta}_d(d^{\max(\mathsf{k}_\star/2, 1)+1})$.
  - Changing loss will not help.
  - Reusing samples unlikely to help (bad local minima [M., Saeed, Zhu,'25]).

Why is SGD suboptimal here?

Landscape smoothing

- [Damian, Nichani, Ge, Lee,'23] modified this algo using *landscape smoothing*, from tensor PCA [Biroli, Cammarota, Ricci-Tersenghi,'20]

- Online SGD on population loss

$$\mathcal{L}_\lambda(\boldsymbol{w}) := \mathbb{E}_{\boldsymbol{u}\sim \mathrm{Unif}(\mathbb{S}^{d-1})}\left[\mathcal{L}\left(\frac{\boldsymbol{w}+\lambda\boldsymbol{u}}{\|\boldsymbol{w}+\lambda\boldsymbol{u}\|_2}\right)\right]$$

where $\lambda = d^{1/4}$ and $\mathcal{L}(\boldsymbol{w}) = \frac{1}{2}\mathbb{E}[(\mathcal{T}(y) - \sigma(\langle \boldsymbol{w}, \boldsymbol{x}\rangle))^2]$.

- This modification achieves (near-)optimal complexity:

$$\mathsf{m} = \widetilde{\Theta}_d(d^{\mathsf{k}_\star/2}), \qquad \mathsf{T} = \widetilde{\Theta}_d(d^{\mathsf{k}_\star/2+1}).$$

Why does this modification achieve optimal complexity ?*

Why $d^{\mathsf{k}_\star/2+1}$ versus $d^{\mathsf{k}_\star}$ runtime complexity?

(*Note that this algo fails on a slightly modified SIM)

Partial trace of Hermite tensor

- [Damian, Pillaud-Vivien, Lee, Bruna,'24] achieved $\mathsf{m} = \Theta(d^{\mathsf{k}_\star/2})$ using partial trace of an Hermite tensor (again from tensor PCA [Hopkins et al.,'16]).

- Construct an empirical tensor

$$\hat{\boldsymbol{T}} := \frac{1}{\mathsf{m}} \sum_{i \in [\mathsf{m}]} \mathcal{T}(y_i) \mathbf{He}_{\mathsf{k}_\star}(\boldsymbol{x}_i) \in (\mathbb{R}^d)^{\otimes \mathsf{k}_\star} \qquad (\text{s.t.}, \ \mathbb{E}[\hat{\boldsymbol{T}}] = c_{\mathcal{T},\mathsf{k}_\star} \cdot \boldsymbol{w}_*^{\otimes \mathsf{k}_\star})$$

  and take partial trace (here for $\mathsf{k}_\star$ even):

$$\hat{\boldsymbol{w}} = \arg\min_{\|\boldsymbol{u}\|_2 = 1} \boldsymbol{u}^\mathsf{T} \hat{\boldsymbol{M}} \boldsymbol{u}, \quad \hat{\boldsymbol{M}} = \hat{\boldsymbol{T}}[\mathbf{I}_d^{\otimes(\mathsf{k}_\star/2-1)}] \in \mathbb{R}^{d \times d}.$$

- Achieves

$$\mathsf{m} = \Theta_d(d^{\mathsf{k}_\star/2}), \qquad \mathsf{T} = \widetilde{\Theta}_d(d^{\mathsf{k}_\star/2+1}).$$

Why does partial trace achieve optimal complexity ?*

Why $d^{\mathsf{k}_\star/2}$ sample complexity?

(*Note that this algo fails on a slightly modified SIM)

Summary

- [Damian, Pillaud-Vivien, Lee, Bruna,'24] sharp characterization of complexity of learning Gaussian SIMs:

$$\mathsf{m} = \Theta_d(d^{\mathsf{k}_\star/2}), \qquad \mathsf{T} = \widetilde{\Theta}_d(d^{\mathsf{k}_\star/2+1}),$$

  where $\mathsf{k}_\star$ is the generative exponent of ρ.

- Several conceptual gaps:
  - Why is SGD algorithm suboptimal with runtime $d^{\mathsf{k}_\star}$ instead of $d^{\mathsf{k}_\star/2+1}$?
  - Why do landscape smoothing and partial trace estimators (both borrowed from tensor PCA) achieve optimal complexity?
  - What role does the Gaussian assumption play in these results?

- **Goal:** see how our general equivariance framework which focuses on the symmetry group clarifies these questions.

Our framework

- Gaussian SIMs correspond to the orbit class
$$\mathcal{D}[\mathcal{O}_d] = \{\mathcal{D}_{\boldsymbol{w}} : \boldsymbol{w} \sim \mathbb{S}^{d-1}\}.$$

- Natural basis associated to $\mathcal{O}_d$ symmetry are spherical harmonics and not Hermite polynomials (harmonic subspaces = irreducible representations of $\mathcal{O}_d$.)

- Adopting spherical harmonic basis:
  - Clarify above questions.
  - Uncover new phenomena.
  - Extends Gaussian setting to arbitrary spherically symmetric distributions.

1'a Learning single-index models via harmonic decomposition

[Joshi, Koubbi, **M.**, Srebro, arXiv:2506.09887]

Spherical Single-Index models

- $\boldsymbol{x} \sim \mu$ rotationally invariant

$$\boldsymbol{x} = r\boldsymbol{z} : \qquad r = ||\boldsymbol{x}||_2 \sim \mu_R \quad \perp \quad \boldsymbol{z} = \boldsymbol{x}/||\boldsymbol{x}||_2 \sim \tau_d := \mathrm{Unif}(\mathbb{S}^{d-1}).$$

- Spherical single-index models: unknown $\boldsymbol{w}_* \in \mathbb{S}^{d-1}$ and

$$(y, \boldsymbol{x}) \sim \mathbb{P}_{\boldsymbol{w}_*, \nu_d} : \quad \boldsymbol{x} = (r, \boldsymbol{z}) \sim \mu = \mu_R \otimes \tau_d \quad \text{and} \quad y|(r, \boldsymbol{z}) \sim \nu_d(\cdot|r, \langle \boldsymbol{w}_*, \boldsymbol{z}\rangle).$$

  - Link fct $\nu_d \in \mathcal{P}(\mathcal{Y} \times \mathbb{R}_{\geq 0} \times [-1, 1])$

$$(Y, R, Z) \sim \nu_d : \quad R \sim \nu_{d,R} \;\perp\; Z \sim \tau_{d,1} \quad \text{and} \quad Y|(R, Z) \sim \nu_d(\cdot|R, Z).$$

  - Gaussian SIMs: $\mu_R = \chi_d$ and $\nu_d(\cdot|r, \langle \boldsymbol{w}_*, \boldsymbol{z}\rangle) = \rho(\cdot|r \cdot \langle \boldsymbol{w}_*, \boldsymbol{z}\rangle)$.

- Given m iid data $(y_i, r_i, \boldsymbol{z}_i) \sim \mathbb{P}_{\nu_d, \boldsymbol{w}_*}$ with unknown $\boldsymbol{w}_*$, compute $\hat{\boldsymbol{w}}$ such that

$$|\langle \hat{\boldsymbol{w}}, \boldsymbol{w}_* \rangle| \geq 1 - \varepsilon,$$

with proba $1 - \delta$.

Harmonic decomposition

- Harmonic decomposition of $L^2(\mathbb{S}^{d-1})$ into

$$L^2(\mathbb{S}^{d-1}) = \bigoplus_{\ell=0}^{\infty} V_{d,\ell}, \qquad n_{d,\ell} = \dim(V_{d,\ell}) = \Theta_d(d^\ell),$$

  where $V_{d,\ell}$ denotes the space of degree-ℓ spherical harmonics.

- $\mathbb{P}_{\nu_d,0}$ distribution of $(y, r) \sim \nu_{d,Y,R}$ and $z \sim \tau_d$ independent.
- Decomposition of likelihood ratio:

$$\frac{\mathrm{d}\mathbb{P}_{\nu_d, \boldsymbol{w}_*}}{\mathrm{d}\mathbb{P}_{\nu_d,0}}(y, r, \boldsymbol{z}) = 1 + \sum_{\ell=1}^{\infty} \xi_{d,\ell}(y, r) Q_\ell(\langle \boldsymbol{w}_*, \boldsymbol{z} \rangle),$$

$$\xi_{d,\ell}(y, r) := \mathbb{E}_{(Y,R,Z)\sim\nu_d}\left[Q_\ell(Z) | Y = y, R = r\right],$$

  where Q_ℓ are the orthonormal Gegenbauer polynomials (in $L^2([-1, 1], \tau_{d,1})$)

$$\mathbb{E}_{\boldsymbol{z}\sim\tau_d}[Q_\ell(\langle \boldsymbol{e}_1, \boldsymbol{z} \rangle) Q_k(\langle \boldsymbol{e}_1, \boldsymbol{z} \rangle)] = \delta_{\ell=k}.$$

Complexity lower bounds

- Lower bounds (within SQ and LDP):

$$\text{Sample:} \quad \mathsf{m} \gtrsim \inf_{\ell \geq 1} \frac{\sqrt{n_{d,\ell}}}{||\xi_{d,\ell}||_{L^2}^2}, \qquad \text{Runtime:} \quad \mathsf{T} \gtrsim \inf_{\ell \geq 1} \frac{n_{d,\ell}}{||\xi_{d,\ell}||_{L^2}^2},$$

where $n_{d,\ell} = \dim(V_{d,\ell})$ and $\xi_{d,\ell}(y,r) = \mathbb{E}_{(Y,R,Z)\sim\nu_d}\left[Q_\ell(Z)|Y=y, R=r\right]$.

- **Interpretation:** consider an algorithm that only uses statistics in $V_{d,\ell}$:

$$\text{Sample:} \quad \mathsf{m} \gtrsim \frac{\sqrt{n_{d,\ell}}}{||\xi_{d,\ell}||_{L^2}^2}, \qquad \text{Runtime:} \quad \mathsf{T} \gtrsim \frac{n_{d,\ell}}{||\xi_{d,\ell}||_{L^2}^2}.$$

For each $V_{d,\ell}$: matching algorithm (next slide).

Problem decouples across irreducible subspaces with optimal algo on each $V_{d,\ell}$.

Matching algorithms

$$\text{Sample:} \quad \mathsf{m} \gtrsim \inf_{\ell \geq 1} \frac{d^{\ell/2}}{||\xi_{d,\ell}||_{L^2}^2}, \qquad \text{Runtime:} \quad \mathsf{T} \gtrsim \inf_{\ell \geq 1} \frac{d^{\ell}}{||\xi_{d,\ell}||_{L^2}^2},$$

Subspace $V_{d,\ell}$	Sample optimal	Runtime optimal
$\ell = 1$	Spectral algorithm: $\mathsf{m} \asymp \frac{d^{1/2}}{\|\|\xi_{d,1}\|\|_{L^2}^2}$,	$\mathsf{T} \asymp \frac{d^{3/2}}{\|\|\xi_{d,1}\|\|_{L^2}^2}$
$\ell = 2$	Spectral algorithm: $\mathsf{m} \asymp \frac{d}{\|\|\xi_{d,2}\|\|_{L^2}^2}$,	$\mathsf{T} \asymp \frac{d^2 \log(d)}{\|\|\xi_{d,2}\|\|_{L^2}^2}$.
$\ell \geq 3$	Harmonic tensor unfolding ℓ even: $\mathsf{m} \asymp \frac{d^{\ell/2}}{\|\|\xi_{d,\ell}\|\|_{L^2}^2}$, $\mathsf{T} \asymp \frac{d^{\ell} \log(d)}{\|\|\xi_{d,\ell}\|\|_{L^2}^2}$ ℓ odd: $\mathsf{m} \asymp \frac{d^{\ell/2}}{\|\|\xi_{d,\ell}\|\|_{L^2}^2}$, $\mathsf{T} \asymp \frac{d^{\ell+\frac{1}{2}} \log(d)}{\|\|\xi_{d,\ell}\|\|_{L^2}^2}$	Online SGD $\mathsf{m} \asymp \frac{d^{\ell-1}}{\|\|\xi_{d,\ell}\|\|_{L^2}^2}$, $\mathsf{T} \asymp \frac{d^{\ell}}{\|\|\xi_{d,\ell}\|\|_{L^2}^2}$

Spectral/Online SGD algorithm

- 'Spectral algorithm': ($\ell = 2$ case) [Lu,Li,'17], [Mondelli, Montanari,'18]

$$\hat{\boldsymbol{w}} = \arg\min_{\|\boldsymbol{w}\|_2=1} \boldsymbol{w}^{\mathsf{T}} \hat{\boldsymbol{M}} \boldsymbol{w}, \qquad \hat{\boldsymbol{M}} = \frac{1}{\mathsf{m}} \sum_{i \in [\mathsf{m}]} \mathcal{T}(y_i, r_i) \left[d \cdot \boldsymbol{z}_i \boldsymbol{z}_i^{\mathsf{T}} - \mathbf{I}_d \right] \in \mathbb{R}^{d \times d}$$

achieves

$$\mathsf{m} \asymp \frac{d}{\|\xi_{d,2}\|_{L^2}^2}, \qquad \mathsf{T} \asymp \frac{d^2}{\|\xi_{d,2}\|_{L^2}^2} \log(d).$$

- 'Online SGD algorithm' for $\ell \geq 3$: online SGD on loss

$$\min_{\boldsymbol{w} \in \mathbb{S}^{d-1}} \mathbb{E}\left[\left(\mathcal{T}(y, r) - Q_\ell(\langle \boldsymbol{w}, \boldsymbol{z} \rangle) \right)^2 \right]$$

achieves

$$\mathsf{m} \asymp \frac{d^{\ell-1}}{\|\xi_{d,\ell}\|_{L^2}^2}, \qquad \mathsf{T} \asymp \frac{d^{\ell}}{\|\xi_{d,\ell}\|_{L^2}^2}.$$

Harmonic tensor unfolding

- Harmonic tensor: $\mathcal{H}_\ell(\boldsymbol{z}) \in (\mathbb{R}^d)^{\otimes \ell}$ defined such that

$$Q_\ell(\langle \boldsymbol{w}, \boldsymbol{z} \rangle) = \langle \mathcal{H}_\ell(\boldsymbol{z}), \boldsymbol{w}^{\otimes \ell} \rangle, \qquad \text{for all } \boldsymbol{w} \in \mathbb{S}^{d-1}.$$

  Explicit formula:

$$\mathcal{H}_\ell(\boldsymbol{z}) = \sum_{j=0}^{\lfloor \ell/2 \rfloor} (-1)^j 2^{\ell-2j} \frac{\ell!}{j!(\ell-2j)!} \frac{(d/2-1)_{\ell-j}}{(d-2)_\ell} \sqrt{n_{d,\ell}} \cdot \mathrm{Sym}(\boldsymbol{z}^{\otimes(\ell-2j)} \otimes \mathbf{I}_d^{\otimes j}).$$

- Reproducing property:

$$\mathbb{E}[Q_k(\langle \boldsymbol{w}_*, \boldsymbol{z} \rangle) \mathcal{H}_\ell(\boldsymbol{z})] = \frac{\delta_{k\ell}}{\sqrt{n_{d,\ell}}} \mathcal{H}_\ell(\boldsymbol{w}_*) \approx \boldsymbol{w}_*^{\otimes \ell} + o_{d,\|\cdot\|_F}(d^{-1/2}).$$

  Second moment:

$$\mathbb{E}\Big[\mathcal{H}_\ell(\boldsymbol{z}) \otimes \mathcal{H}_\ell(\boldsymbol{z})\Big] = \sum_{j=0}^{\lfloor \ell/2 \rfloor} c_{\ell,j} \cdot \mathrm{Sym}_A \left(\mathbf{I}_d^{\otimes(\ell-2j)} \otimes (\mathbf{I}_d \otimes \mathbf{I}_d)^{\otimes j} \right)$$

Harmonic tensor unfolding

Tensor unfolding algorithm (below the even case $\ell = 2p$)

- Compute empirical tensor:

$$\hat{\boldsymbol{T}} = \frac{1}{\mathsf{m}} \sum_{i \in [\mathsf{m}]} \mathcal{T}(y_i, r_i) \mathcal{H}_\ell(\boldsymbol{z}_i) \in (\mathbb{R}^d)^{\otimes \ell}, \qquad \mathbb{E}[\hat{\boldsymbol{T}}] = c_{\mathcal{T}} \cdot \boldsymbol{w}_*^{\otimes \ell} + o_{d,\|\cdot\|_{\mathrm{op}}}(d^{-1/2}).$$

- Unfold the tensor [Richard, Montanari,'14]:

$$\hat{\boldsymbol{M}} = \mathbf{Mat}_{p,p}(\hat{\boldsymbol{T}}) \in \mathbb{R}^{d^p \times d^p}.$$

and compute top eigenvector $\boldsymbol{s}_1 \in \mathbb{R}^{d^p}$ of $\hat{\boldsymbol{M}}$.

- $\hat{\boldsymbol{w}}$ top left singular vector of $\mathbf{Mat}_{1,p-1}(\boldsymbol{s}_1) \approx \boldsymbol{w}_* [\boldsymbol{w}_*^{\otimes p-1}]^{\mathsf{T}} \in \mathbb{R}^{d \times d^{p-1}}$.

- Tensor unfolding achieves

$$\mathsf{m} \asymp \frac{d^{\ell/2}}{\|\xi_{d,\ell}\|_{L^2}^2}, \qquad \mathsf{T} \asymp \frac{d^{\ell}}{\|\xi_{d,\ell}\|_{L^2}^2} \log(d).$$

Algorithms

$$\text{Sample:} \quad \mathsf{m} \gtrsim \inf_{\ell \geq 1} \frac{d^{\ell/2}}{\|\xi_{d,\ell}\|_{L^2}^2}, \qquad \text{Runtime:} \quad \mathsf{T} \gtrsim \inf_{\ell \geq 1} \frac{d^{\ell}}{\|\xi_{d,\ell}\|_{L^2}^2},$$

Subspace $V_{d,\ell}$	Sample optimal	Runtime optimal
	Spectral algorithm	
$\ell = 1$	$\mathsf{m} \asymp \frac{d^{1/2}}{\|\xi_{d,1}\|_{L^2}^2}$,	$\mathsf{T} \asymp \frac{d^{3/2}}{\|\xi_{d,1}\|_{L^2}^2}$
$\ell = 2$	$\mathsf{m} \asymp \frac{d}{\|\xi_{d,2}\|_{L^2}^2}$,	$\mathsf{T} \asymp \frac{d^2 \log(d)}{\|\xi_{d,2}\|_{L^2}^2}$.
$\ell \geq 3$	Harmonic tensor unfolding ℓ even: $\mathsf{m} \asymp \frac{d^{\ell/2}}{\|\xi_{d,\ell}\|_{L^2}^2}$, $\mathsf{T} \asymp \frac{d^{\ell} \log(d)}{\|\xi_{d,\ell}\|_{L^2}^2}$ ℓ odd: $\mathsf{m} \asymp \frac{d^{\ell/2}}{\|\xi_{d,\ell}\|_{L^2}^2}$, $\mathsf{T} \asymp \frac{d^{\ell+\frac{1}{2}} \log(d)}{\|\xi_{d,\ell}\|_{L^2}^2}$	Online SGD $\mathsf{m} \asymp \frac{d^{\ell-1}}{\|\xi_{d,\ell}\|_{L^2}^2}$, $\mathsf{T} \asymp \frac{d^{\ell}}{\|\xi_{d,\ell}\|_{L^2}^2}$

Runtime-optimal vs sample-optimal

- Optimal algorithm to estimate $\boldsymbol{w}_*$: compute degree

$$\mathsf{l}_{\mathsf{m},\star} = \arg\min_{\ell \geq 1} \frac{\sqrt{n_{d,\ell}}}{\|\xi_{d,\ell}\|_{L^2}^2}, \qquad \mathsf{l}_{\mathsf{T},\star} = \arg\min_{\ell \geq 1} \frac{n_{d,\ell}}{\|\xi_{d,\ell}\|_{L^2}^2},$$

  and use associated algorithm on $V_{d,\mathsf{l}_{\mathsf{m},\star}}$ or $V_{d,\mathsf{l}_{\mathsf{T},\star}}$.

  Competition between $\dim(V_{d,\ell})$ and signal strength $\|\xi_{d,\ell}\|_{L^2}^2$ on that subspace.

- If $\mathsf{l}_{\mathsf{m},\star} = \mathsf{l}_{\mathsf{T},\star}$, then tensor algo is both sample- and runtime-optimal (nearly).

- In general, we can have $\mathsf{l}_{\mathsf{m},\star} \gg \mathsf{l}_{\mathsf{T},\star}$: we expect no algorithm can simultaneously achieve optimal sample and runtime complexity.

  $\neq$ Gaussian SIMs where both complexities are always jointly achievable.

Additional sample-runtime trade-offs when learning SIMs beyond the Gaussian setting.

Example

- Fix $k \in \mathbb{N}$. Consider $Y|R, Z \sim \nu_d$ mixture of

$$Y|R, Z \sim \nu_{1,d}(\cdot|R, Z) \quad \text{w. p. } 1 - d^{-2k}, \qquad Y|R, Z \sim \nu_{2,d}(\cdot|R, Z) \quad \text{w. p. } d^{-2k}.$$

- SIMs are chosen such that $\mathsf{l}_{\star,\mathsf{m}} = 10k$ thanks to $\nu_{d,1}$ and $\mathsf{l}_{\star,\mathsf{T}} = 4k$ thanks to $\nu_{d,2}$.

- Optimal algorithms:
  - **Sample-optimal:** harmonic tensor unfolding at $\mathsf{l}_{\star,\mathsf{m}} = 10k$

$$\mathsf{m} \asymp d^{5k}, \qquad \mathsf{T} \asymp d^{10k}.$$

  - **Runtime-optimal:** harmonic tensor unfolding at $\mathsf{l}_{\star,\mathsf{T}} = 4k$

$$\mathsf{m} \asymp d^{6k}, \qquad \mathsf{T} \asymp d^{8k}.$$

Summary

- Harmonic decomposition:

$$L^2(\mathbb{S}^{d-1}) = \bigoplus_{\ell=0}^{\infty} V_{d,\ell}, \qquad n_{d,\ell} = \dim(V_{d,\ell}) = \Theta_d(d^\ell).$$

SIM coefficients: $\xi_{d,\ell}(y,r) = \mathbb{E}_{(Y,R,Z)\sim\nu_d}[Q_\ell(Z)|Y=y, R=r]$.

- Lower bounds decouple across these harmonic subspaces:

$$\text{Sample:} \quad \mathsf{m} \gtrsim \inf_{\ell\geq 1} \frac{\sqrt{n_{d,\ell}}}{||\xi_{d,\ell}||_{L^2}^2}, \qquad \text{Runtime:} \quad \mathsf{T} \gtrsim \inf_{\ell\geq 1} \frac{n_{d,\ell}}{||\xi_{d,\ell}||_{L^2}^2}.$$

- Matching algo for each $V_{d,\ell}$ (spectral, online SGD, harmonic tensor unfolding).

- Optimal algo: take algo on $V_{d,\ell}$ with ℓ taken either

$$\mathsf{l}_{\mathsf{m},\star} = \arg\min_{\ell\geq 1} \frac{\sqrt{n_{d,\ell}}}{||\xi_{d,\ell}||_{L^2}^2}, \qquad \mathsf{l}_{\mathsf{T},\star} = \arg\min_{\ell\geq 1} \frac{n_{d,\ell}}{||\xi_{d,\ell}||_{L^2}^2}.$$

1’b Learning Gaussian Single-Index Models

Harmonic decomposition

- Gaussian SIMs: $r \sim \chi_d$ and $\nu_d(\cdot|r, \langle \boldsymbol{w}_*, \boldsymbol{z}\rangle) = \rho(\cdot|r\langle \boldsymbol{w}_*, \boldsymbol{z}\rangle)$ with gen. exp.

$$\mathsf{k}_\star = \arg\min_{k\geq 1}\{k : \|\zeta_k\|_{L^2} > 0 \text{ where } \zeta_k = \mathbb{E}_{(Y,G)\sim\rho}[\mathrm{He}_k(G)|Y]\}.$$

- Hermite to Gegenbauer decomposition:

$$\mathrm{He}_k(r\cdot\langle \boldsymbol{w}_*, \boldsymbol{z}\rangle) = \sum_{\ell\leq k} c_{k,\ell}(r) Q_\ell(\langle \boldsymbol{w}_*, \boldsymbol{z}\rangle), \qquad \|c_{k,\ell}\|_{L^2}^2 \asymp \delta_{\ell\equiv \mathsf{k}_\star[2]} d^{-(k-\ell)/2}.$$

- Vanishing projection on lower degree harmonics: $\|\mathsf{P}_{V_{d,\ell}}\mathrm{He}_k\|_{L^2}^2 \asymp d^{-(k-\ell)/2}$. However, it will have important algorithmic consequences!

- The Gegenbauer coeffs of ν_d: $\|\xi_{d,\ell}\|_{L^2}^2 \asymp d^{-(\mathsf{k}_\star-\ell+\delta_{\ell\neq\mathsf{k}_\star[2]})/2}$

$$\mathsf{m} \gtrsim \inf_{\ell\geq 1} \frac{\sqrt{n_{d,\ell}}}{\|\xi_{d,\ell}\|_{L^2}^2} \asymp d^{\mathsf{k}_\star/2}, \qquad \mathsf{T} \gtrsim \inf_{\ell\geq 1} \frac{n_{d,\ell}}{\|\xi_{d,\ell}\|_{L^2}^2} \asymp d^{\mathsf{k}_\star/2+1}.$$

Always achieved at $\mathsf{l}_{\mathsf{m},\star} = \mathsf{l}_{\mathsf{T},\star} = 1$ if $\mathsf{k}_\star$ odd and $\mathsf{l}_{\mathsf{m},\star} = \mathsf{l}_{\mathsf{T},\star} = 2$ if $\mathsf{k}_\star$ even.

Optimal algorithms for Gaussian SIMs

- Optimal algorithms on $V_{d,1}$ and $V_{d,2}$: spectral algorithm

$$\mathsf{m} \asymp d^{\mathsf{k}_*/2}, \qquad \mathsf{T} \asymp d^{\mathsf{k}_\star/2+1}\log(d).$$

- For any $\mathsf{k}_\star$: uses degree-1 or 2 spherical harmonics (depending on parity of $\mathsf{k}_\star$).

- For $\ell = 2$ (all Gaussian SIMs with even information exponent):

$$\hat{\boldsymbol{w}} = \arg\min_{\boldsymbol{w}\in\mathbb{S}^{d-1}} \boldsymbol{w}^\mathsf{T}\hat{\boldsymbol{M}}\boldsymbol{w}, \qquad \hat{\boldsymbol{M}} = \frac{1}{m}\sum_{i\in[m]} \mathcal{T}(y_i, r_i)[d\cdot \boldsymbol{z}_i\boldsymbol{z}_i^\mathsf{T} - \mathbf{I}_d].$$

This is simply the algo for phase retrieval [Lu, Li,'20], [Mondelli, Montanari,'18].

Without using the norm

- Consider algo that only uses directional information $\boldsymbol{z}_i = \boldsymbol{x}_i / \|\boldsymbol{x}_i\|_2$.
  E.g., common practice in stats/ML of normalizing input vectors to unit norm.

- Indeed: $\|\boldsymbol{x}\|_2$ does not contain any information about $\boldsymbol{w}_*$ and $\|\boldsymbol{x}\|_2/\sqrt{d} \to 1$ a.s.

- However: for Gaussian SIMs with info exponent $\mathsf{k}_\star$, the complexity becomes

$$\mathsf{m} \asymp d^{\mathsf{k}_\star/2}, \qquad \mathsf{T} \asymp d^{\mathsf{k}_\star}, \qquad \text{(optimal algo now at } \mathsf{l}_{\mathsf{m},\star} = \mathsf{l}_{\mathsf{T},\star} = \mathsf{k}_\star\text{)}.$$

To get from $\Theta(d^{\mathsf{k}_\star})$ to $\Theta(d^{\mathsf{k}_\star/2+1})$ runtime, one has to exploit the norm $\|\boldsymbol{x}\|_2$.

Online SGD

- [Ben Arous, Gheissari, Jagannath, '21] Online SGD on population loss

$$\min_{\boldsymbol{w} \in \mathbb{S}^{d-1}} \mathcal{L}(\boldsymbol{w}) = \frac{1}{2} \mathbb{E}_{(y,\boldsymbol{x}) \sim \mathbb{P}_{\boldsymbol{w}_*}} \left[\Big(\mathcal{T}(y) - \sigma(\langle \boldsymbol{w}, \boldsymbol{x} \rangle) \Big)^2 \right] \qquad (\star)$$

requires suboptimal $\mathsf{m} = \tilde{\Theta}(d^{\mathsf{k}_\star - 1})$ and $\mathsf{T} = \tilde{\Theta}(d^{\mathsf{k}_\star})$.

- Dynamics stay essentially the same if $\boldsymbol{x}$ is replaced by $\sqrt{d}\boldsymbol{x}/||\boldsymbol{x}||_2$: dynamics does not exploit the norm of the Gaussian vector.

- From our results, estimators only using $\boldsymbol{z} = \boldsymbol{x}/||\boldsymbol{x}||_2$ incur $\mathsf{T} = \Omega(d^{\mathsf{k}_\star})$.

- In this sense, $(\star)$ is runtime optimal among algo that only use directional info.

Landscape smoothing

- [Damian, Nichani, Ge, Lee,'23] Online SGD on 'smoothed landscape':

$$\min_{\boldsymbol{w}\in\mathbb{S}^{d-1}} \mathbb{E}_{\boldsymbol{u}\sim\mathrm{Unif}(\mathbb{S}^{d-1})}\mathbb{E}_{(y,\boldsymbol{x})}\left[\left(\mathcal{T}(y)-\sigma\left(\frac{\boldsymbol{w}+\lambda\boldsymbol{u}}{\|\boldsymbol{w}+\lambda\boldsymbol{u}\|_2}\cdot\boldsymbol{x}\right)\right)^2\right]$$

achieves $\mathsf{m}=\tilde{\Theta}(d^{\mathsf{k}_\star/2})$ and runtime $\tilde{\Theta}(d^{\mathsf{k}_\star/2+1})$.

- Frequency decomposition of the loss:

$$\mathbb{E}_{\boldsymbol{u}}\mathbb{E}_{y,\boldsymbol{x}}\left[\mathcal{T}(y)\mathrm{He}_{\mathsf{k}_\star}\left(\frac{\boldsymbol{w}+\lambda\boldsymbol{u}}{\|\boldsymbol{w}+\lambda\boldsymbol{u}\|_2}\cdot\boldsymbol{x}\right)\right]=\sum_{\ell\leq\mathsf{k}_\star} m_\ell(\lambda)\cdot\mathbb{E}\left[\mathcal{T}(y)c_{\mathsf{k}_\star,\ell}(r)Q_\ell(\langle\boldsymbol{w},\boldsymbol{z}\rangle)\right]$$

  - No smoothing: $m_\ell(0)=1$, dominated by $V_{d,\mathsf{k}_\star}\longrightarrow\tilde{\Theta}(d^{\mathsf{k}_\star})$ runtime.
  - Smoothing: $m_\ell(d^{\frac{1}{4}})\asymp d^{-\frac{\ell}{2}}$, dominated by $V_{d,1}/V_{d,2}\longrightarrow\tilde{\Theta}(d^{\frac{\mathsf{k}_\star}{2}+1})$ runtime.

Smoothing reweights the landscape towards smaller frequencies $(V_{d,1}/V_{d,2})$.

Partial trace estimator

- [Damian, Pillaud-Vivien, Lee, Bruna, '24] compute empirical tensor

$$\widehat{\boldsymbol{T}} = \frac{1}{m} \sum_{i \in [m]} \mathcal{T}(y_i) \mathbf{He}_{\mathsf{k}_\star}(\boldsymbol{x}_i) \in (\mathbb{R}^d)^{\otimes \mathsf{k}_\star},$$

and $\hat{\boldsymbol{w}} =$ top eigenvector of partial trace $\hat{\boldsymbol{M}} = \hat{\boldsymbol{T}}[\mathbf{I}_d^{\otimes(\mathsf{k}_\star/2-1)}] \in \mathbb{R}^{d \times d}$

$$\mathsf{k}_\star \text{ even:} \quad \widehat{\boldsymbol{M}} = \frac{1}{m} \sum_{i \in [m]} \mathcal{T}(y_i) P_{\mathsf{k}_\star}(\|\boldsymbol{x}_i\|_2) \left[\boldsymbol{x}_i \boldsymbol{x}_i^\mathsf{T} - c_k \mathbf{I}_d\right]$$

$$\approx \frac{1}{m} \sum_{i \in [m]} \widetilde{\mathcal{T}}(y_i, \|\boldsymbol{x}_i\|_2) \left[\frac{\boldsymbol{x}_i \boldsymbol{x}_i^\mathsf{T}}{\|\boldsymbol{x}_i\|_2^2} - \frac{\mathbf{I}_d}{d}\right] \quad \text{(spectral estimator).}$$

Partial trace projects on optimal subspace $V_{d,2}$ (and $V_{d,1}$ for odd).

- Landscape smoothing and partial trace: if we normalize $\boldsymbol{x}$, then sample complexity becomes $d^{\mathsf{k}_\star - 1}$ for both.
  (The low frequencies $V_{d,1}/V_{d,2}$ are not optimal anymore.)

Summary: Gaussian single-index model

Advantages of this "harmonic analysis" perspectives:

- Natural basis to study single index-models:
  - It explicitly exploits the spherical symmetry of the problem.
  - Explicitly decompose function space by delineating (r, z) and harmonic degree. This has crucial algorithmic consequences.
  - More transparent derivation of optimal algorithms in the Gaussian setting.
- Recover generative exponent. Interpretation $d^{k/2+1}$ vs d^k runtime:
  $\longrightarrow$ harmonic subspaces $V_{d,1}, V_{d,2}$/whether exploit the norm or not.
- Success of landscape smoothing/partial trace estimator:
  $\longrightarrow$ effectively project on optimal $V_{d,1}/V_{d,2}$ subspaces.

  (These algo come from tensor PCA, with similar gap $d^{k/2+1}$ vs d^k .... ???)
- Does not use Gaussianity, only spherical invariance
  $\longrightarrow$ applies to general spherically symmetric distribution μ.
  $\longrightarrow$ there are new phenomena beyond Gaussian setting.

2 Learning multi-index models

[Koubbi, Latourelle-Vigeant, **M.**,???'25]

Multi-index models

- Label y now depends on a s-dimensional subspace $\boldsymbol{W}_*^\top \boldsymbol{x}$ with $\boldsymbol{W}_*^\top \boldsymbol{W}_* = \mathbf{I}_s$.
- Spherical multi-index models: unknown $\boldsymbol{W}_* \in O(d,s)$ and

$$(y, \boldsymbol{x}) \sim \mathbb{P}_{\boldsymbol{W}_*, \nu_d} : \quad \boldsymbol{x} = (r, \boldsymbol{z}) \sim \mu = \mu_R \otimes \tau_d \quad \text{and} \quad y|(r, \boldsymbol{z}) \sim \nu_d(\cdot | r, \langle \boldsymbol{W}_*, \boldsymbol{z} \rangle).$$

- Lower bounds for detection (within SQ and LDP):

$$\text{Sample:} \quad \mathsf{m} \gtrsim \inf_{\ell \geq 1} \frac{\sqrt{n_{d,\ell}}}{\|\xi_{d,\ell}\|_{L^2}^2}, \qquad \text{Runtime:} \quad \mathsf{T} \gtrsim \inf_{\ell \geq 1} \frac{n_{d,\ell}}{\|\xi_{d,\ell}\|_{L^2}^2},$$

where $\xi_{d,\ell} := \mathsf{P}_{L^2(\nu_{Y,R}) \otimes V_{d,\ell}} \frac{\mathrm{d}\mathbb{P}_{\boldsymbol{W}_*, \nu_d}}{\mathrm{d}\mathbb{P}_{0,\nu_d}}$.

An example

$$y = \underbrace{\langle \boldsymbol{w}_1, \boldsymbol{x} \rangle}_{\in V_{d,1}} + \underbrace{\text{sign}(\langle \boldsymbol{w}_1, \boldsymbol{x} \rangle \langle \boldsymbol{w}_2, \boldsymbol{x} \rangle \cdots \langle \boldsymbol{w}_k, \boldsymbol{x} \rangle)}_{\in V_{d,k}} .$$

- Algos:
  - On $V_{d,1}$: $\mathsf{m} \asymp d$ and $\mathsf{T} \asymp d^2$ and recover $\boldsymbol{w}_1$.
  - On $V_{d,k}$: $\mathsf{m} \asymp d^{k/2}$ and $\mathsf{T} \asymp d^k$ and recover $[\boldsymbol{w}_1, \ldots, \boldsymbol{w}_k]$.
- Optimal detection: it is enough to consider $V_{d,1}$. But can only recover $\boldsymbol{w}_1$.

  Full support recovery in one step using $V_{d,k}$.
- Optimal recovery algorithm: sequential adaptive learning of support
  - **Step 1:** on $V_{d,1}$ recover $\langle \hat{\boldsymbol{w}}_1, \boldsymbol{x} \rangle$: $\mathsf{m} \asymp d$, $\mathsf{T} \asymp d^2$.
  - **Step 2:** conditional on $\langle \hat{\boldsymbol{w}}_1, \boldsymbol{x} \rangle$, on $V_{d-1,k-1}$: $\mathsf{m} \asymp d^{(k-1)/2}$, $\mathsf{T} \asymp d^{k-1}$.

  Total complexity: $\mathsf{m} \asymp d^{(k-1)/2}$, $\mathsf{T} \asymp d^{k-1}$.

Sequential learning

- Optimal algorithms recover the support $\boldsymbol{W}_*$ sequentially:

$$\{0\} \subset \boldsymbol{U}_1 \subset \boldsymbol{U}_2 \subset \cdots \subset \boldsymbol{U}_{q-1} \subset \boldsymbol{U}_q = \boldsymbol{W}_* \in \mathcal{O}(d,s)$$

- Conditional on having recovered $\boldsymbol{U}^\mathsf{T}\boldsymbol{x}$, we can decompose $(y,\boldsymbol{x}) \sim \mathbb{P}_{\boldsymbol{W}_*,\nu_d}$:

$$\boldsymbol{x} = \boldsymbol{U}^\mathsf{T}\boldsymbol{x} + (\|\boldsymbol{x}\|_2^2 - \|\boldsymbol{U}^\mathsf{T}\boldsymbol{x}\|_2^2)^{1/2}(\mathbf{I}_d - \boldsymbol{U}\boldsymbol{U}^\mathsf{T})^{1/2}\boldsymbol{z}, \qquad \boldsymbol{z} \sim \mathrm{Unif}(\mathbb{S}^{d-s_0-1}).$$

- Lower bounds for next step:

$$\text{Sample:} \quad \mathsf{m} \gtrsim \inf_{\ell \geq 1} \frac{\sqrt{n_{d-s_0,\ell}}}{\|\xi_{d,\ell,\boldsymbol{U}}\|_{L^2}^2}, \qquad \text{Runtime:} \quad \mathsf{T} \gtrsim \inf_{\ell \geq 1} \frac{n_{d-s_0,\ell}}{\|\xi_{d,\ell,\boldsymbol{U}}\|_{L^2}^2},$$

where $\xi_{d,\ell,\boldsymbol{U}} := \mathsf{P}_{V_{d-s_0,\ell}} \frac{\mathrm{d}\mathbb{P}_{\boldsymbol{W}_*,\nu_d}}{\mathrm{d}\mathbb{P}_{\boldsymbol{U},\nu_d}}$.

- Using optimal ℓ: learn new directions $\tilde{\boldsymbol{U}}$ and $\boldsymbol{U} \to \boldsymbol{U}' = [\boldsymbol{U}, \tilde{\boldsymbol{U}}]$.

Leap complexities

- [Abbe, Boix-Adsera, **M.**,'23], [Bietti, Bruna, Pillaud-Vivien,'23], [Damian, Lee, Bruna,'25] "complexity of the worst subspace to recover"

$$\mathsf{m} \gtrsim \mathsf{Leap}_{\mathsf{m}}(\nu_d), \qquad \mathsf{T} \gtrsim \mathsf{Leap}_{\mathsf{T}}(\nu_d),$$

  where

$$\text{sample-optimal leap:} \qquad \mathsf{Leap}_{\mathsf{m}}(\nu_d) = \sup_{\boldsymbol{U} \subset \boldsymbol{W}_*} \inf_{\ell \geq 1} \frac{d^{\ell/2}}{\|\xi_{d,\ell,\boldsymbol{U}}\|_{L^2}^2},$$

$$\text{runtime-optimal leap:} \qquad \mathsf{Leap}_{\mathsf{T}}(\nu_d) = \sup_{\boldsymbol{U} \subset \boldsymbol{W}_*} \inf_{\ell \geq 1} \frac{d^{\ell}}{\|\xi_{d,\ell,\boldsymbol{U}}\|_{L^2}^2}.$$

- Matching algorithm on each $V_{d',\ell}$ using harmonic tensor unfolding.
  Both sample and (near-)runtime optimal on $V_{d',\ell}$.

- Whether we are sample or compute-constrained, might choose different ℓ.

  Sample-optimal and runtime-optimal algorithms will recover the support with different sequences $\{0\} \subset \boldsymbol{U}_1 \subset \cdots \subset \boldsymbol{U}_{q-1} \subset \boldsymbol{W}_*$ and match these LBs.

2 General Framework

[Joshi, Koubbi, **M.**, Nati,???'25]

Summary (I)

- Learning $\mathcal{D}$ with $\mathcal{G}$-equivariant algos $\iff$ Learning orbit $\mathcal{D}[\mathcal{G}] = \{\mathcal{D}^g : g \in \mathcal{G}\}$.

- Lower bounds within SQ and LDP:
  - "Weak learning": Alignment complexities

$$\mathsf{m} \gtrsim \mathsf{Align}_{\mathsf{m}}(\mathcal{D};\mathcal{G}) := \inf_{\hat{\rho}\in\hat{\mathcal{G}}_0} \frac{\sqrt{n_{\hat{\rho}}}}{\mathsf{Q}_{\hat{\rho}}(\mathcal{D};\mathcal{G})},$$

$$\mathsf{T} \gtrsim \mathsf{Align}_{\mathsf{T}}(\mathcal{D};\mathcal{G}) := \inf_{\hat{\rho}\in\hat{\mathcal{G}}_0} \frac{n_{\hat{\rho}}}{\mathsf{M}_{\hat{\rho}}(\mathcal{D};\mathcal{G})}.$$

  - "Strong learning": Leap complexities

$$\mathsf{m} \gtrsim \mathsf{Leap}_{\mathsf{m}}(\mathcal{D};\mathcal{G}) := \sup_{\mathcal{H}\in\mathcal{S}_\varepsilon} \mathsf{Align}_{\mathsf{m}}(\mathcal{D};\mathcal{H}),$$

$$\mathsf{T} \gtrsim \mathsf{Leap}_{\mathsf{T}}(\mathcal{D};\mathcal{G}) := \sup_{\mathcal{H}\in\mathcal{S}_\varepsilon} \mathsf{Align}_{\mathsf{T}}(\mathcal{D};\mathcal{H}).$$

  Worst-case complexity of learning subgroup $\mathcal{H}$.

Summary (II)

- Optimal algorithms chosen at each step

$$\hat{\rho}_{\mathsf{m},\star} := \arg\min_{\hat{\rho}\in\hat{\mathcal{H}}_0} \frac{\sqrt{n_{\hat{\rho}}}}{\mathsf{Q}_{\hat{\rho}}(\mathcal{D};\mathcal{H})}, \qquad \hat{\rho}_{\mathsf{T},\star} := \arg\min_{\hat{\rho}\in\hat{\mathcal{H}}_0} \frac{n_{\hat{\rho}}}{\mathsf{M}_{\hat{\rho}}(\mathcal{D};\mathcal{H})}.$$

- Sequential adaptive learning of the group:
  - Nested sequence of subgroups:

$$\mathcal{G} =: \mathcal{H}^{(0)} \supset \mathcal{H}^{(1)} \supset \mathcal{H}^{(2)} \supset \cdots \supset \mathcal{H}^{(t+1)} = \{e\}.$$

  - Factorization of the group:

$$\mathcal{G} = (\mathcal{H}^{(0)}/\mathcal{H}^{(1)}) \times (\mathcal{H}^{(1)}/\mathcal{H}^{(2)}) \times \cdots \times (\mathcal{H}^{(t)}/\mathcal{H}^{(t+1)}).$$

  - To learn $g_* = (h_1^*, \ldots, h_t^*) \in \mathcal{G}$, learn sequentially $\hat{g} = (\hat{h}_1, \hat{h}_2, \ldots, \hat{h}_t)$.

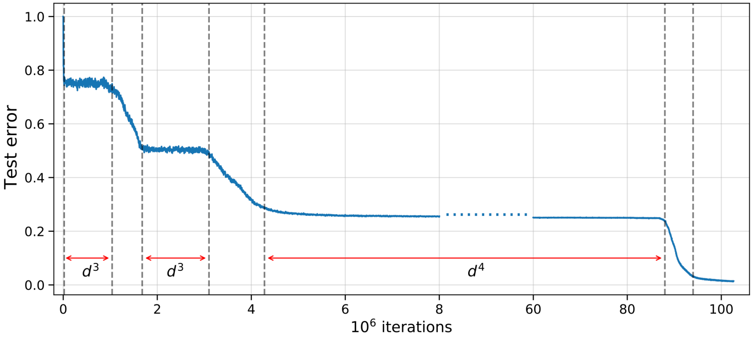
- Lower bounds in terms of generic properties of the group.

  Upper bounds: case by case analysis.

Example

$$f_*(x) = x_1 + x_1x_2x_3x_4 + x_1x_2\cdots x_7 + x_1x_2x_3\cdots x_{11}.$$

$$\mathfrak{S}_d \xrightarrow{d} \mathrm{Id}_1 \oplus \mathfrak{S}_{d-1} \xrightarrow{d^3} \mathrm{Id}_4 \oplus \mathfrak{S}_{d-4} \xrightarrow{d^3} \mathrm{Id}_7 \oplus \mathfrak{S}_{d-7} \xrightarrow{d^4} \mathrm{Id}_{11} \oplus \mathfrak{S}_{d-11}$$



[Abbe, Boix-Adsera, M.,'23]

Open questions

- This framework 'compactly' captures a number of phenomena, but it is far from a complete picture:
  - Systematic procedure to design optimal equivariant algorithms?
  - When do gradient-trained neural networks match these lower bounds?
  - Leap captures complexity of breaking a symmetry. How to capture other aspects? (e.g., μ that is non $\mathcal{G}$-invariant).
- Harmonic analysis: useful tool to decompose function spaces and finding optimal statistics of the data.
- Orbit classes $\mathcal{D}[\mathcal{G}]$ appear in many planted models:
  sparse PCA, tensor PCA, planted subgraphs, planted submatrix...
  - Many complexity gaps $d^{k/2}$ vs d^k between classes of algos in these models
  - For Gaussian SIMs, e.g., depends on using optimal harmonics + $\|x\|_2$.