Learning from long sequences of high-dimensional tokens, and extensive-width neural networks

Antoine Maillard

Inria

Les Houches – February 6th 2025

References

- Exact threshold for approximate ellipsoid fitting of random points (*arXiv:2310.05787)*
- Bayes-optimal learning of an extensive-width neural network from quadratically many samples (*arXiv:2408.03733, NeurIPS '24*)
- Bilinear Sequence Regression: A Model for Learning from Long Sequences of High-dimensional Tokens (*arXiv:2410.03733*)


Afonso Bandeira


Lenka Zdeborová


Florent Krzakala


Emanuele Troiani


Simon Martin


Vittorio Erba


Luca Biggio

EPFL ETH zürich Inria ENS | PSL

Solvable models of learning

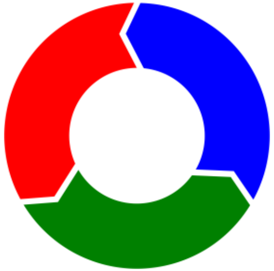
High-dimensional statistics / machine learning

Example: language modeling

Data

Sequences of tokens

$(\mathbf{x}_i, y_i) \sim p$



Model

RNN / Transformer / ...

$y = f(\boldsymbol{\theta}, \mathbf{x})$

Algorithms

Gradient descent (and variants)

$$\hat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}} \mathcal{R}(\{\mathbf{x}_i, y_i\}_{i=1}^n, \boldsymbol{\theta})$$

Why is it advantageous to present the data as long sequences of high-dimensional tokens ?

Objectives:

→ Build **solvable models** that retain the **key ingredients** 

→ A **rigorous** theory describing learning in these models

What is the **simplest** exactly solvable model that exhibits advantages in learning from long sequences of high-dimensional tokens ?

Bilinear Sequence Regression (BSR)

Erba, Troiani, Biggio, **M.** & Zdeborová '24

What is the **simplest** exactly solvable model that exhibits advantages in learning from long sequences of high-dimensional tokens ?

The simplest regression model for **vectorized data** $\mathbf{x} \in \mathbb{R}^d$ is **(generalized) linear regression** $y = \varphi(\mathbf{x} \cdot \mathbf{w})$

$\mathbf{X} \in \mathbb{R}^{L \times d}$: sequence of L d-dimensional tokens

L : **length of the sequences**

d : embedding **dimension of tokens**

$L, d \gg 1$

r **rank/width** of the model

Bilinear Sequence Regression (BSR)

The **most basic** regression model is

$$y = f_{\mathbf{M}}(\mathbf{X}) = \varphi(\mathrm{Tr}[\mathbf{X}\mathbf{M}^\intercal])$$

$\Rightarrow$

$$f_{\mathbf{U},\mathbf{V}}(\mathbf{X}) = \varphi\left(\mathrm{Tr}[\mathbf{X}\mathbf{U}\mathbf{V}^\intercal]\right) = \varphi\left(\sum_{a,i=1}^{L,d} X_{ai} \sum_{\gamma=1}^{r} U_{i\gamma} V_{\gamma a}\right)$$

U : token space

V : sequence space

Also called **Matrix Sensing** Recht&al '10, Gunasekar&al '17, …

Like linear regression is a base model for non-sequential data, BSR is a toy base model for sequences of tokens

Setting

Teacher-student

$$y_\mu = \mathrm{Tr}[\mathbf{X}_\mu \mathbf{U}^\star (\mathbf{V}^\star)^\intercal] = \sum_{a,i=1}^{L,d} X^\mu_{ai} \sum_{\gamma=1}^{r} U^\star_{i\gamma} V^\star_{\gamma a}$$

"Generic"

Ex: $\sim \mathcal{N}(0, \mathrm{I}_{L\times d})$

$\overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1)$

Learning from data

$$\mathcal{D} = \{(\mathbf{X}_1, y_1), \cdots, (\mathbf{X}_n, y_n)\} \Rightarrow \boxed{(\mathbf{U}^\star, \mathbf{V}^\star)} \, ?$$

Scalings

High-dimensional

$L, d \to \infty \, ; \, L = \Theta(d)$

Long sequences of **high-dimensional** tokens

Extensive-width

$r = \Theta(d)$

Related to **extensive-width** neural networks

Talks of Guilhem & Jean

Number of samples

$n = \Theta[r(L+d)]$

For **non-trivial** gen. error

Low-rank/width

$r = \Theta(1)$

[Schülke&al '16]
Optimal gen. error, and
optimal algorithm (AMP)

Objectives

$$L = \Theta(d)\,; r = \Theta(d)\; n = \Theta[r(L+d)]$$

$$y_\mu = \mathrm{Tr}[\mathbf{X}_\mu \mathbf{U}^\star (\mathbf{V}^\star)^\intercal] = \sum_{a,i=1}^{L,d} X^\mu_{ai} \sum_{\gamma=1}^{r} U^\star_{i\gamma} V^\star_{\gamma a}$$

"Generic"

Ex: $\sim \mathcal{N}(0, \mathrm{I}_{L\times d})$

$\overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1)$

Teacher-student model

Training dataset $\mathcal{D} = \{\mathbf{X}^\mu, y^\mu\}_{\mu=1}^n \Rightarrow \boxed{\mathbf{U}^\star, \mathbf{V}^\star}$?

Objectives

Bayes-optimal generalization error

$$\mathcal{E}_{\text{gen.}} := \mathbb{E}_{\mathbf{U}^\star,\mathbf{V}^\star,\mathcal{D}} \min_{\hat{y}(\mathcal{D})} \mathbb{E}_{\mathbf{X}_{\text{test}}} [(\hat{y}(\mathbf{X}_{\text{test}}) - f_{\mathbf{U}^\star,\mathbf{V}^\star}(\mathbf{X}_{\text{test}}))^2]$$

- ❑ **Sharp thresholds** ? Phase transitions ?
- ❑ Efficiently achievable ? **Hard phases** ?

Comparison to **vectorized data** ? (linear regression)

$$\mathbf{S}^\star = \mathbf{U}^\star(\mathbf{V}^\star)^\intercal \Rightarrow \mathbf{S}^\star \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1)$$

Comparison to **minimal nuclear-norm** estimator

$$\min\{\|\mathbf{S}\|_{\text{NN}} \,:\, \mathbf{S} = \mathbf{U}\mathbf{V}^\intercal,\; y_\mu = \mathrm{Tr}[\mathbf{X}_\mu \mathbf{S}]\}$$

Recht&al '10

- ❑ Performance of **GD-based algorithms** ?
  Bhojanapalli&al '16

$$\mathcal{L}(\mathbf{U},\mathbf{V}) = \sum_{\mu=1}^{n} (y_\mu - \mathrm{Tr}[\mathbf{X}_\mu \mathbf{U}\mathbf{V}^\intercal])^2$$

- ❑ Is GD an **implicit nuclear-norm minimizer** ?
  Gunasekar&al '17

Result I : Bayes-optimal error

$$\left\{ y_\mu = \sum_{a,i=1}^{L,d} X_{ai}^\mu \sum_{\gamma=1}^{r} U_{i\gamma}^\star V_{\gamma a}^\star \right\}_{\mu=1}^{n}$$

Scaling regime

$$\rho = \frac{r}{d} \qquad \beta = \frac{L}{d} \qquad \alpha = \frac{n}{dL}$$

$$\mathcal{E}_{\text{gen.}} := \mathbb{E}_{\mathbf{U}^\star, \mathbf{V}^\star, \mathcal{D}} \min_{\hat{y}(\mathcal{D})} \mathbb{E}_{\mathbf{X}_{\text{test}}} [(\hat{y}(\mathbf{X}_{\text{test}}) - f_{\mathbf{U}^\star, \mathbf{V}^\star}(\mathbf{X}_{\text{test}}))^2]$$

$$\Downarrow$$

$$\boxed{\lim_{d \to \infty} \mathcal{E}_{\text{gen.}} = \alpha \zeta}$$

ζ solves the self-consistent equation

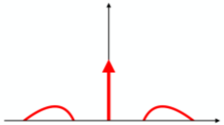
$$(1-\alpha) = \frac{2\zeta}{\beta^{3/2}} \int \mathrm{d}y \, \mu_\zeta(y) \left[\frac{(\beta-1)^2}{2x^2} + \frac{2\pi^2}{3} \mu_\zeta(y)^2 \right]$$

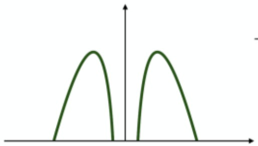
- **Easy-to-evaluate formula** for the Bayes-optimal generalization error
- Derivation via the **replica method** of stat. physics
- Paper with **proofs** in preparation
  [Xu, M., Krzakala, Zdeborová '25]

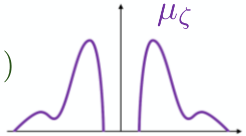
Analytical form using **free probability** tools

$$\mu_\zeta := \mathrm{SymSVD}(\mathbf{S}^\star + \sqrt{\zeta}\mathbf{Z}) \qquad \mathbf{Z} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1)$$

$$\mathbf{S}^\star = \mathbf{U}^\star (\mathbf{V}^\star)^\intercal$$







Benaych-Georges '09

Derivation: a symmetric variant to BSR

M., Troiani, Martin, Krzakala & Zdeborová '24

Bilinear Sequence Regression (BSR)

$$f_{\mathrm{U,V}}(\mathbf{X}) = \varphi\left(\mathrm{Tr}[\mathbf{X}\mathbf{U}\mathbf{V}^\intercal]\right) = \varphi\left(\sum_{a,i=1}^{L,d} X_{ai} \sum_{\gamma=1}^{r} U_{i\gamma} V_{\gamma a}\right)$$

A symmetric variant

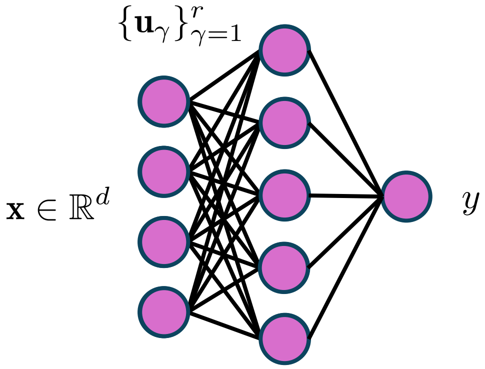
$L = d$

$\mathbf{X} = \mathbf{X}^\intercal$

$+$

$$f_{\mathbf{S}}(\mathbf{X}) = \varphi\left(\mathrm{Tr}[\mathbf{X}\mathbf{S}]\right)$$

Cf Jean's talk on Tuesday



Example

"Extensive-width sign retrieval"

⇕

Extensive-width 2-layer NN with a quadratic activation

$\mathbf{X} = \mathbf{x}\mathbf{x}^\intercal$ rank-1

$$\mathbf{S} = \sum_{\gamma=1}^{r} \mathbf{u}_\gamma \mathbf{u}_\gamma^\intercal \succeq 0$$

$$y_\mu = \mathbf{x}_\mu^\intercal \mathbf{S}^\star \mathbf{x}_\mu = \sum_{\gamma=1}^{r} \left[(\mathbf{u}_\gamma^\star)^T \cdot \mathbf{x}_\mu\right]^2$$

$\sim \mathcal{N}(0, \mathrm{I}_d)$

$\mathbf{u}_\gamma^\star \sim \mathcal{N}(0, \mathrm{I}_d)$

High-dimensional limit

$$d \to \infty\,;\, r = \Theta(d)\,;\, n = \Theta(d^2)$$

Same objectives as BSR

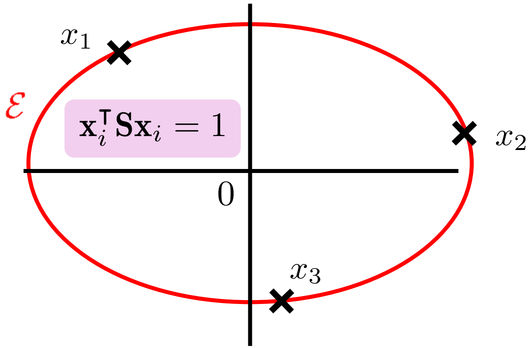
Proof ideas (1)

$$y_\mu = \varphi\left(\mathrm{Tr}[\mathbf{X}_\mu \mathbf{S}^\star]\right) \qquad \mathbf{S}^\star = \mathbf{U}^\star(\mathbf{U}^\star)^\intercal \sim \text{ Wishart matrix}$$

Step 1 : "Gaussian universality"

Universality of Bayes-optimal generalization error

Developed for the analysis of ellipsoid fitting [**M.** & Bandeira '23]



The "generic" assumption on $\mathbb{P}(\mathbf{X}_\mu)$

Gaussian matrix

Uniform Central Limit Theorem for 1-dimensional projections $\mathrm{Tr}[\mathbf{XS}]$

$$\sup_{\mathbf{S}} |\mathbb{E}\phi(\mathrm{Tr}[\mathbf{XS}]) - \mathbb{E}\phi(\mathrm{Tr}[\mathbf{GS}])| \to 0$$

Satisfied e.g. by $\mathbf{X} = \mathbf{x}\mathbf{x}^\intercal$, $\mathbf{x} \sim \mathcal{N}(0, \mathrm{I}_d)$ (extensive-width sign retrieval)

$$\min \mathcal{E}_{\mathrm{gen.}}(\hat{\mathbf{w}}_k) = \min \|\mathbf{y}(\hat{\mathbf{S}}) - \mathbf{y}(\mathbf{S}^\star)\|^2 \quad = \quad \min \widetilde{\mathcal{E}}_{\mathrm{gen.}}(\hat{\mathbf{S}}) = \min \|\widetilde{\mathbf{y}}(\hat{\mathbf{S}}) - \widetilde{\mathbf{y}}(\mathbf{S}^\star)\|^2 \times (1 + o(1))$$

from $\{y_\mu = \varphi\left(\mathrm{Tr}[\mathbf{S}^\star \mathbf{X}_\mu]\right)\}_{\mu=1}^n$

from $\{\tilde{y}_\mu = \varphi(\mathrm{Tr}[\mathbf{S}^\star \mathbf{G}_\mu])\}_{\mu=1}^n$

Gaussian matrix

To compute $\mathcal{E}_{\mathrm{gen.}}$ we **pretend the data is a Gaussian matrix**

Proof ideas (2)

A (generalized) **linear model on** $\mathbf{S}^\star$, with…

Step 2 : $\{\tilde{y}_\mu = \varphi(\mathrm{Tr}[\mathbf{S}^\star \mathbf{G}_\mu])\}_{\mu=1}^n$

- Gaussian data $\mathbf{G} \coloneqq \begin{pmatrix} \mathrm{flatt}(\mathbf{G}_1) \\ \vdots \\ \mathrm{flatt}(\mathbf{G}_n) \end{pmatrix}$ ✚
- Wishart prior $\mathbf{S}^\star = \mathbf{U}^\star(\mathbf{U}^\star)^\intercal \sim \mathcal{W}_{r,d}$

[Barbier & al '19]: **i.i.d. priors**

Generalizing [Barbier & al '19]

An explicit "replica-symmetric" formula for $\widetilde{\mathcal{E}}_{\mathrm{gen.}}$

Involves …

Scalar estimation problem involving φ ✔

Step 3 : **Denoising** problem : $\mathbf{Y} = \sqrt{\lambda}\mathbf{S}^\star + \mathbf{Z} \longrightarrow \boxed{\mathbf{S}^\star}$? ✔

Guilhem's talk

Gaussian (GOE) matrix

[Bun & al '16 ; **M.**, Krzakala & al '22 ; Pourkamali & al '23 ; Semerjian '24 ; …]

- The optimal estimator is **spectral :** $\mathbf{Y} = \mathbf{ODO}^\intercal \Rightarrow \hat{\mathbf{S}}(\mathbf{Y}) = \mathbf{O} f_{\mathrm{opt.}}(\mathbf{D})\mathbf{O}^\intercal$
- Analytical expressions for $f_{\mathrm{opt.}}$ and the **asymptotic MMSE** $\lim_{d\to\infty} \|\hat{\mathbf{S}}(\mathbf{Y}) - \mathbf{S}^\star\|_F^2$

Combining all steps… $\Rightarrow$ $\boxed{\lim_{d\to\infty} \mathcal{E}_{\mathrm{gen.}} = \cdots}$ ■

Result II : The GAMP-RIE algorithm

$y_\mu \sim \varphi\left(\cdot|\mathrm{Tr}[\mathbf{S}^\star \mathbf{X}_\mu]\right)$

Example

$\mathbf{X}_\mu = \mathbf{x}_\mu \mathbf{x}_\mu^\intercal$

$\mathbf{S}^\star \sim \mathcal{W}_{r,d}$ (Wishart)

Informal hypothesis

Universality $\mathbf{X}_\mu \Rightarrow \mathbf{G}_\mu$ also holds "at the level of algorithms"

Generalized linear model w. Gaussian data → MSE-optimal algorithm → **Generalized Approximate Message Passing (GAMP)** [Donoho&al '09 ; Rangan '11 ; ...]

Each GAMP iteration solves

$\mathbf{Y} = \sqrt{\lambda}\mathbf{S}^\star + \mathbf{Z} \longrightarrow \boxed{\mathbf{S}^\star}$?

Guilhem's talk

Rotationally-Invariant Estimator (RIE)

$\mathbf{Y} = \mathbf{ODO}^\intercal \Rightarrow \hat{\mathbf{S}}(\mathbf{Y}) = \mathbf{O} f_{\mathrm{opt.}}(\mathbf{D})\mathbf{O}^\intercal$

[Bun & al '16 ; ...]

Known "optimal shrinkage" function

An explicit easy-to-implement polynomial-time algorithm

GAMP — **RIE**

+

Similar algorithm in non-symmetric model (**Bilinear Sequence Regression**)

Non-symmetric denoising in [Troiani, Erba, Krzakala, **M.** & Zdeborová '22]

Objectives

Objectives

Bayes-optimal generalization error

$$\mathcal{E}_{\text{gen.}} := \mathbb{E}_{\mathbf{U}^\star,\mathbf{V}^\star,\mathcal{D}} \min_{\hat{y}(\mathcal{D})} \mathbb{E}_{\mathbf{X}_{\text{test}}}[(\hat{y}(\mathbf{X}_{\text{test}}) - f_{\mathbf{U}^\star,\mathbf{V}^\star}(\mathbf{X}_{\text{test}}))^2]$$

- ❑ **Sharp thresholds** ? Phase transitions ?
- ❑ Efficiently achievable ? **Hard phases** ?

Comparison to **minimal nuclear-norm** estimator

$$\min\{\|\mathbf{S}\|_{\text{NN}} \ : \ \mathbf{S} = \mathbf{U}\mathbf{V}^\intercal, \ y_\mu = \text{Tr}[\mathbf{X}_\mu \mathbf{S}]\}$$

Recht&al '10

Comparison to **linear regression** on **vectorized data** ?

- ❑ Performance of **GD-based algorithms** ?
  Bhojanapalli&al '16

$$\mathcal{L}(\mathbf{U},\mathbf{V}) = \sum_{\mu=1}^{n} (y_\mu - \text{Tr}[\mathbf{X}_\mu \mathbf{U}\mathbf{V}^\intercal])^2$$

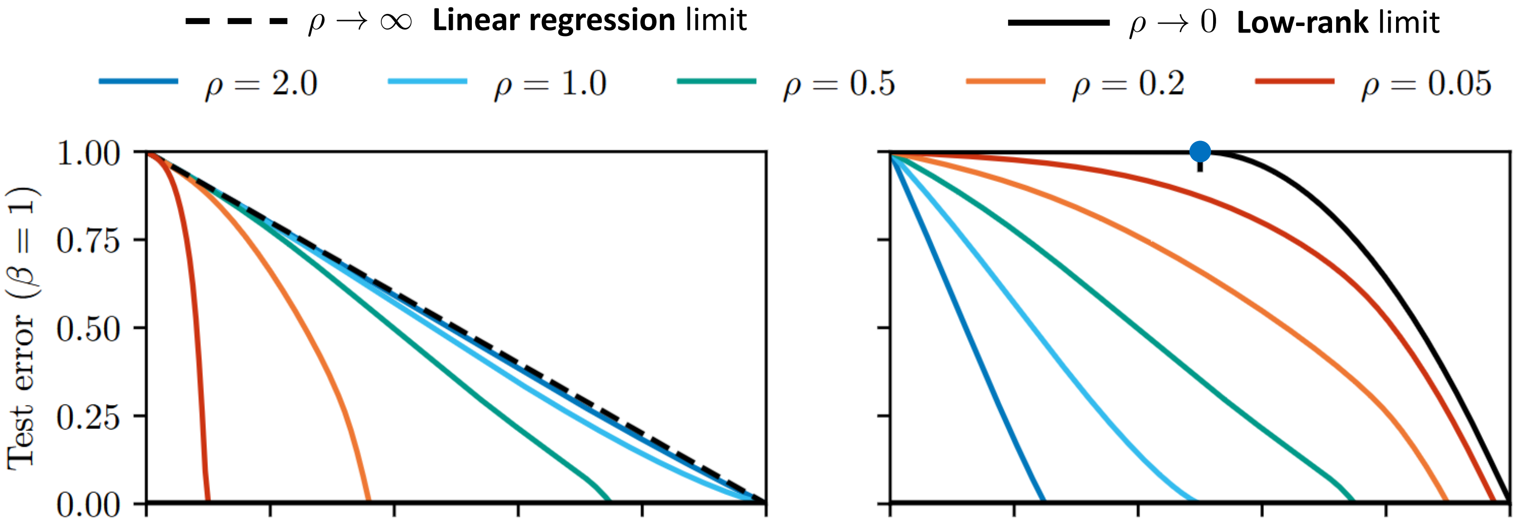
- ❑ Is GD an **implicit nuclear-norm minimizer** ?
  Gunasekar&al '17

Bilinear Sequence Regression (BSR) /
Extensive-rank matrix sensing

Bayes-optimal error

$$\left\{ y_\mu = \sum_{a,i=1}^{L,d} X^\mu_{ai} \sum_{\gamma=1}^{r} U^\star_{i\gamma} V^\star_{\gamma a} \right\}_{\mu=1}^{n}$$

$$\rho = \frac{r}{d} \qquad \beta = \frac{L}{d} \qquad \alpha = \frac{n}{dL}$$



$$\alpha = \frac{n}{dL}$$

Rescaled $\overline{\alpha} = \dfrac{n}{r(d+L)} = \dfrac{\alpha\beta}{\rho(1+\beta)}$

Perfect recovery threshold

$$\alpha_{\mathrm{PR}} = \begin{cases} \frac{\rho}{\beta}(1+\beta-\rho) & 0 < \rho < 1, \\ 1 & \rho \geq 1 \end{cases}$$

$$\overline{\alpha}_{\mathrm{WR}} \overset{\rho\to 0}{=} \frac{\sqrt{\beta}}{1+\beta}$$

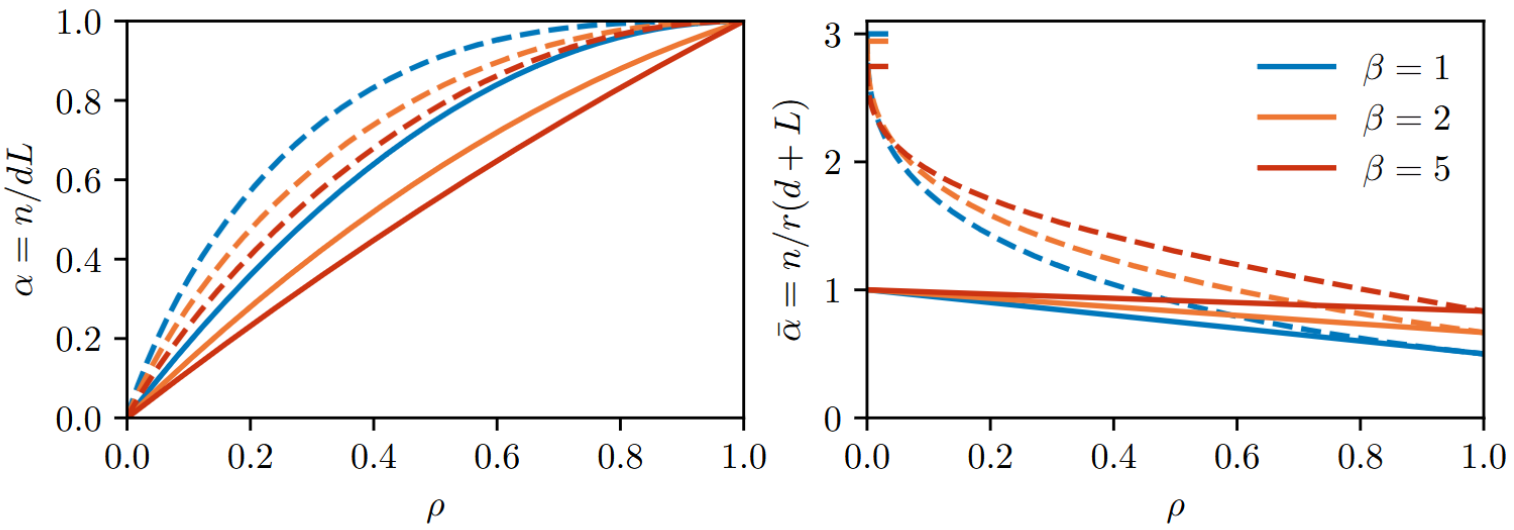
Matches a naïve "counting argument" $\mathrm{DOF}[\mathbf{S} = \mathbf{U}\mathbf{V}^\intercal] \simeq \alpha_{\mathrm{PR}} \cdot d^2$

- **Weak recovery** threshold in the low-rank limit.
- Matches the $r = \mathcal{O}_d(1)$, then $r \to \infty$ limiting curve.

Minimal nuclear-norm estimator

$$\left\{ y_\mu = \sum_{a,i=1}^{L,d} X^\mu_{ai} \sum_{\gamma=1}^{r} U^\star_{i\gamma} V^\star_{\gamma a} \right\}_{\mu=1}^{n}$$

$$\rho = \frac{r}{d} \qquad \beta = \frac{L}{d} \qquad \alpha = \frac{n}{dL}$$



Bayes-optimal **perfect recovery** $\alpha_{\rm PR}$

Perfect recovery of the **minimal nuclear norm** estimator $\min\{\|\mathbf{S}\|_{\rm NN} \,:\, \mathbf{S} = \mathbf{U}\mathbf{V}^\intercal,\, y_\mu = \mathrm{Tr}[\mathbf{X}_\mu \mathbf{S}]\}$

[Donoho, Gavish & Montanari '13]

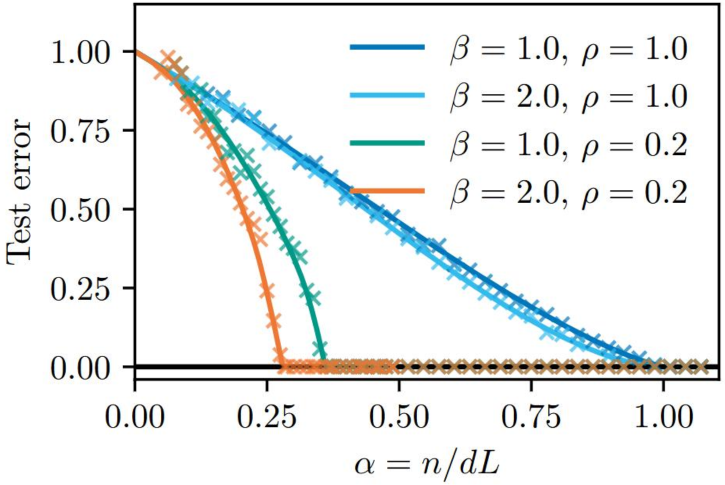
Suboptimality of the min-nuclear norm estimator

Absence of hard phase

$$\left\{ y_\mu = \sum_{a,i=1}^{L,d} X^\mu_{ai} \sum_{\gamma=1}^{r} U^\star_{i\gamma} V^\star_{\gamma a} \right\}_{\mu=1}^{n}$$

$\rho = \frac{r}{d}$ $\beta = \frac{L}{d}$ $\alpha = \frac{n}{dL}$

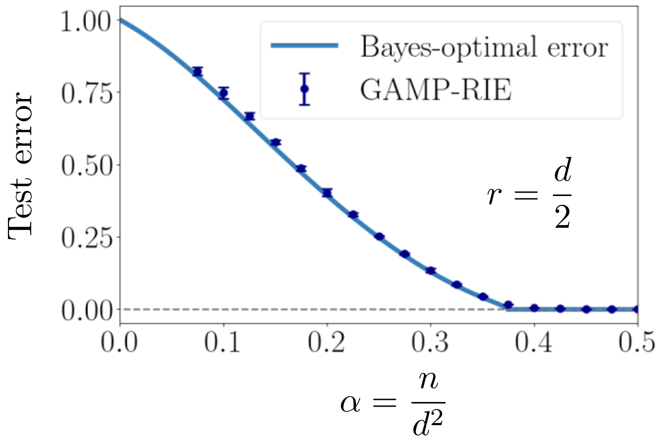
Bilinear Sequence Regression



✖ GAMP-RIE Algorithm

— Bayes-optimal error

Extensive-width sign retrieval



For $r = \mathcal{O}(1)$ there is a hard phase !

cf. [Barbier & al '19] for $r = 1$

No computational-to-statistical gap / hard phase in extensive-width regime

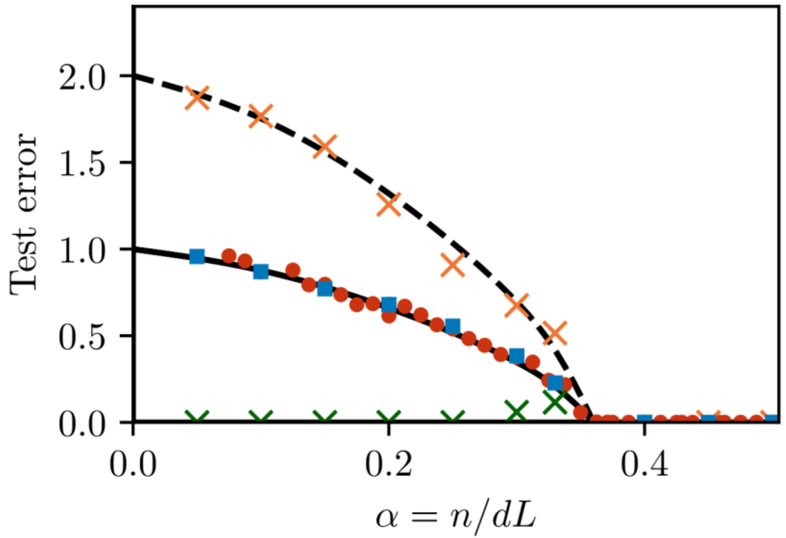
Empirical performance of gradient descent

$\rho = \frac{r}{d} \qquad \beta = \frac{L}{d} \qquad \alpha = \frac{n}{dL}$

$$\mathcal{L}(\mathbf{U}, \mathbf{V}) = \sum_{\mu=1}^{n} \left(y_\mu - \mathrm{Tr}[\mathbf{X}_\mu \mathbf{U}\mathbf{V}^\intercal]\right)^2$$

- $d = 100$ for GAMP/GD runs
- Random initialization
- Cross-validated learning rate

$\beta = 1 \quad \rho = 0.2$



GAMP-RIE

Error of the Bayes-optimal (BO) estimator $\mathbb{E}[\mathbf{U}, \mathbf{V}|\mathcal{D}]$

$2\times$ BO = Error of posterior sampler $\sim \mathbb{P}(\mathbf{U}, \mathbf{V}|\mathcal{D})$

Gradient descent

Gradient descent (final training loss value)

Averaged Gradient descent

Averaged over many initializations

For **any** ρ, AGD seems to reach the Bayes-optimal MMSE

Despite **non-convexity** of the problem!

- Similar results in **extensive-width phase retrieval.**
- No longer true when adding noise to y_μ.

Does GD do implicit nuclear norm regularization ?

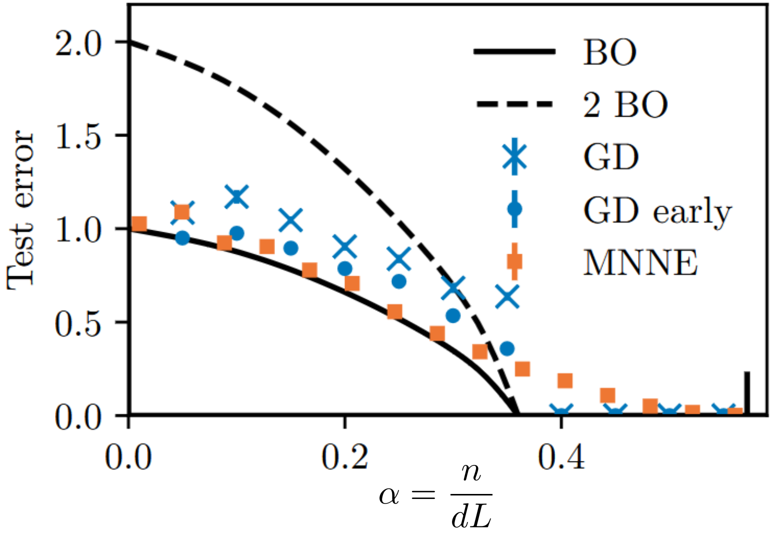
$$\left\{ y_\mu = \sum_{a,i=1}^{L,d} X^\mu_{ai} \sum_{\gamma=1}^{r} U^\star_{i\gamma} V^\star_{\gamma a} \right\}_{\mu=1}^{n}$$

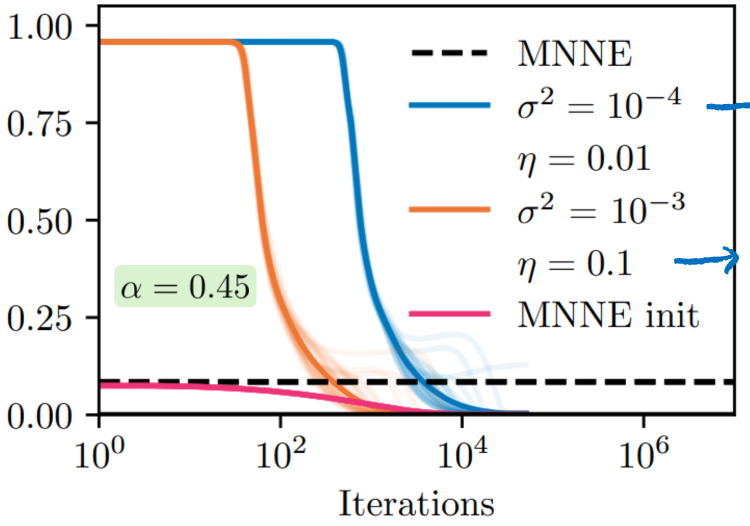
Suggestion: GD with small initialization has an implicit bias towards the minimal nuclear norm estimator (MNNE)

Gunasekar, Woodworth, Bhojanapalli, Neyshabur & Srebro ‘17

$$\min\{\|\mathbf{S}\|_{\mathrm{NN}} \;:\; \mathbf{S} = \mathbf{U}\mathbf{V}^\intercal,\; y_\mu = \mathrm{Tr}[\mathbf{X}_\mu \mathbf{S}]\}$$

$\beta = 1 \quad \rho = 0.2$





Norm of initialization

Learning rate

CONCLUSION

GD does not reach the MNNE

What about **overparametrized** settings $(r^\star < r)$

Summary

Bilinear Sequence Regression (BSR)

Most basic model for learning from **long** sequences of **high-dimensional** tokens

$$y_\mu = \mathrm{Tr}[\mathbf{X}_\mu \mathbf{U}^\star (\mathbf{V}^\star)^\intercal] = \sum_{a,i=1}^{L,d} X^\mu_{ai} \sum_{\gamma=1}^{r} U^\star_{i\gamma} V^\star_{\gamma a}$$

$$L = \Theta(d)\,;\, r = \Theta(d)\; n = \Theta[r(L+d)]$$

1. Analytical formula for the **Bayes-optimal generalization error.**
2. Optimal algorithm (GAMP-RIE), **no computational-statistical gap.**
3. Gap between BO error and **linear regression and MNNE**
4. (Averaged) **Gradient descent seems to sample from the posterior in the noiseless setting**, despite non-convexity !

- Theoretical analysis of GD properties / implicit regularization
- For extensive-width 2-layer NNs: beyond **quadratic activation** ? Cf Jean's talk on Tuesday
- Overparametrization $(r^\star < r)$?
- Correlations between tokens ?

THANK YOU !