Fundamental computational limits of weak learnability in high-dimensional multi-index models

EPFL

Emanuele Troiani[1], Yatin Dandi[1,2], Leonardo Defilippis[3], Lenka Zdeborová[1], Bruno Loureiro[3], Florent Krzakala[2]

[1] EPFL SPOC laboratory, [2] EPFL IDEPHICS laboratory, Lausanne

[3] Departement d'Informatique, École Normale Supérieure, PSL & CNRS, Paris

Model definition

Take a d - dimensional vector x of i.i.d. standard Gaussian random variables, project it onto a $p \times d$ weight matrix W and link the output with a generic function g

$$y = g\left(\begin{bmatrix} z_1 \\ z_2 \\ \vdots \\ z_p \end{bmatrix} = \begin{bmatrix} w_1 \\ w_2 \\ \vdots \\ w_p \end{bmatrix} \begin{bmatrix} x \end{bmatrix} \right)$$

Problem statement

What is the lowest number of samples $n = \alpha d$ needed to have **weak recovery** in the large d limit using first order algorithms?

In practice, can you propose a weight matrix $\hat{w}$ that will have non-zero projection on w?

$$M = \frac{W^T \widehat{W}}{d} = \mathcal{O}_d(1)$$

Hardness of directions

Some of the directions which span W can be harder to recover. We have three classes:

TRIVIAL

Recovery with arbitrarily small α in 1 step using AMP or a finite number of iterations with GD [1]

EASY

Recovery with $\alpha > \alpha_c$ given infinitesimal side information

HARD

Cannot be recovered with any α

Trivial directions

The trivial directions are all directions v such that

$$\mathbb{E}[z^T v | y] \neq 0$$

Examples include:

1. $v = 1$ for $p = 1$ and g is non-even
2. $v = 1_d$ for $g(z) = \sum f(z_i)$ and any non-even f
3. any v for $g(z) = z_1 + z_1 z_2 + z_1 z_2 z_3 + z_1 z_2 z_3 z_4 + \cdots$ [2]

All such v are recovered in **one step** by AMP and a finite number of steps by GD [1], as long as $\alpha > 0$.

If there are no trivial directions, then no direction is recovered unless α is bigger than a **critical α_c and infinitesimal side information**

Easy directions

The easy directions are all the non-trivial directions with finite critical α_c

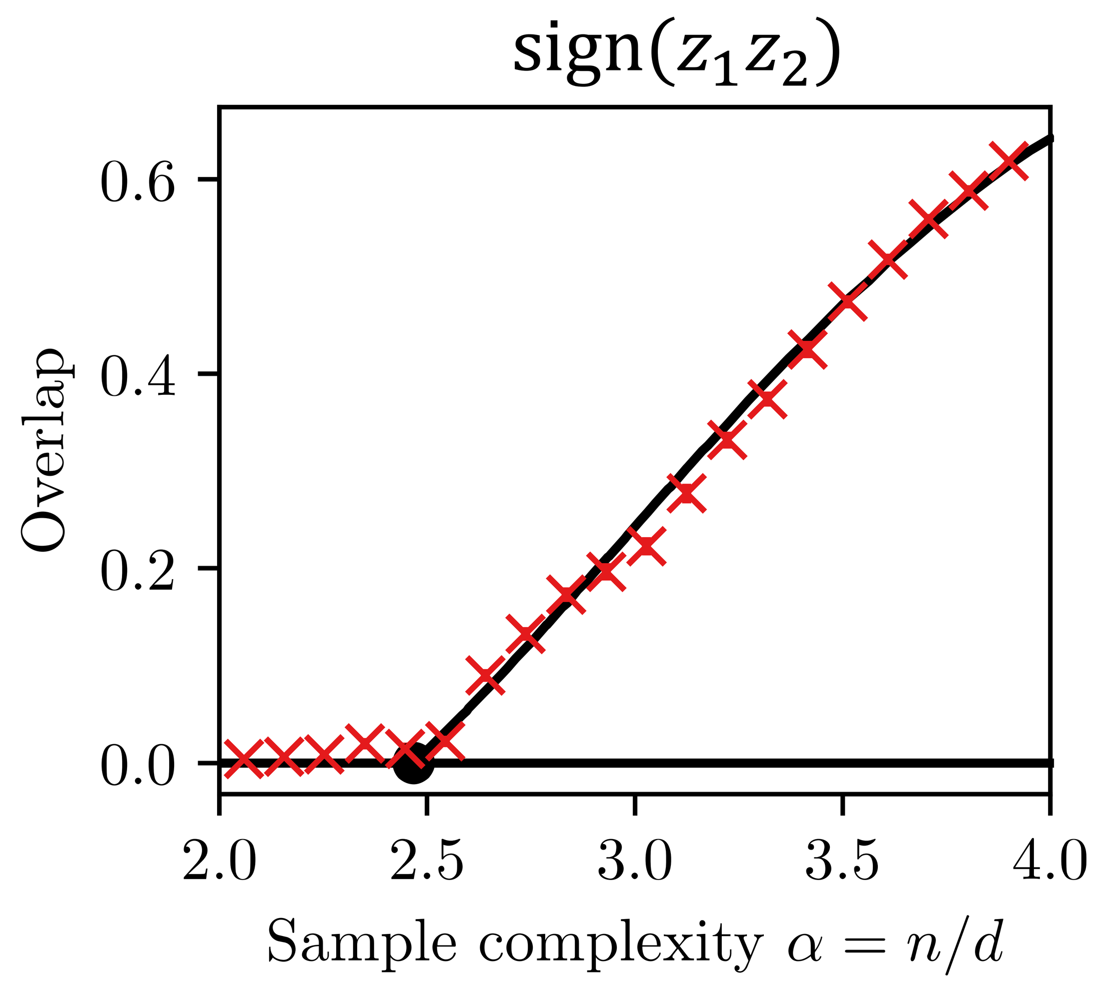
Examples include:

1. any v for $g(z) = z_1 z_2 \dots z_p$
2. any v for $g(z) = \text{sign}(z_1 z_2 \dots z_p)$ if $p \leq 2$

The recovery requires infinitesimal side information.

Easy directions are linked to symmetries in the argument of g [1], such as permutation invariance for 1 and change of sign in 2

If the critical α_c is infinite, that direction is hard and there is no transformation on the output that would make it easy



Iterative denoising

If we run AMP we have three possible scenarios at the start of the run:

1. If all directions are hard, nothing is recovered
2. If some directions are trivial, they will be recovered in one step
3. If there are easy directions and no trivial ones, nothing will be recovered until $\mathcal{O}(\log d)$ iterations, when the recovery will happen

After this stage, if we are in scenarios 2 or 3 we reduce to the subproblem of recovering the remaining directions

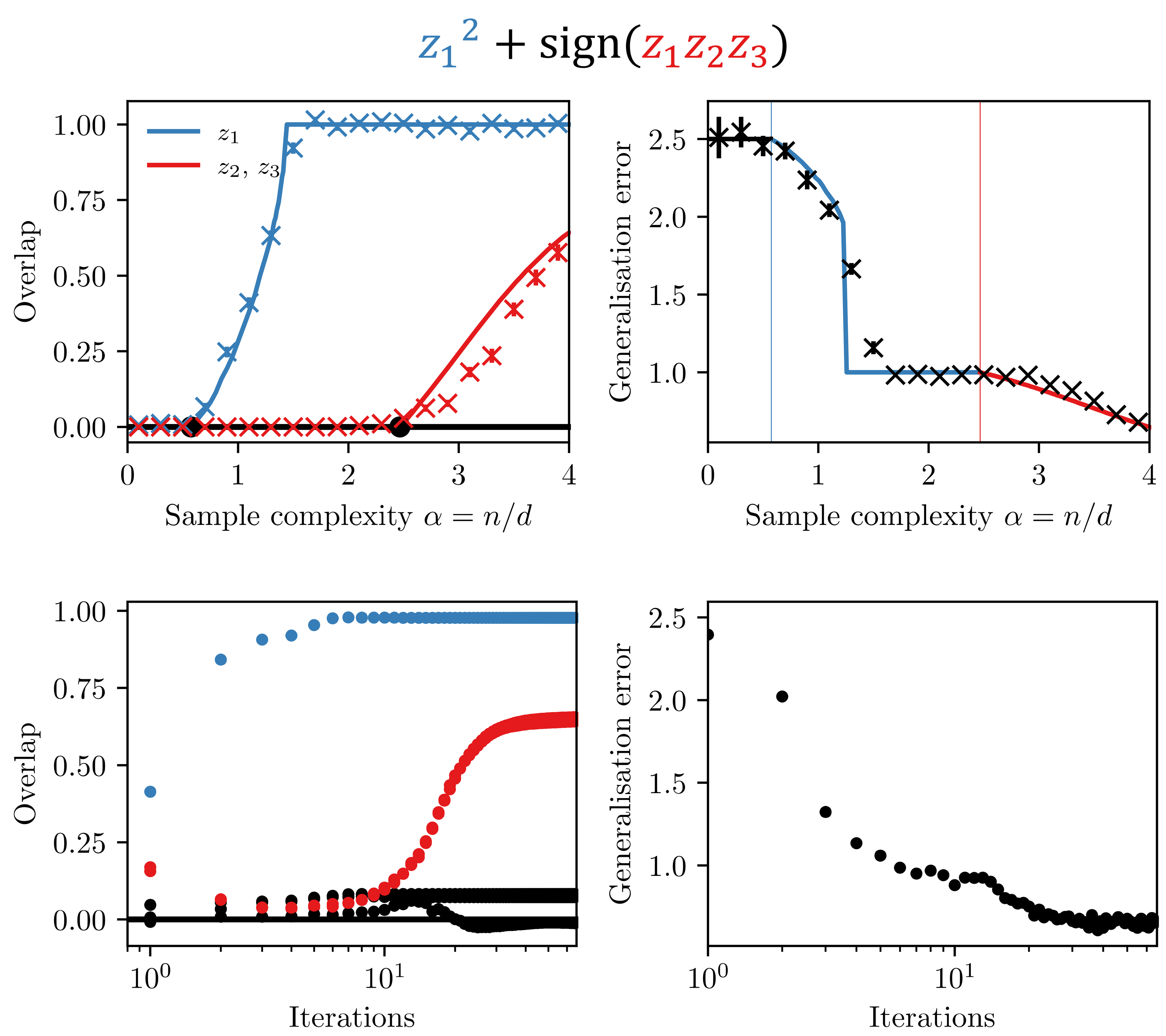
Some hard directions may become easy or trivial

As an example:

$$z_1{}^2 + \text{sign}(z_1 z_2 z_3) + \text{sign}(z_3 z_4 z_5) + \text{sign}(z_5 z_6 z_7) + \cdots$$

There are no trivial directions, and z_1 is the only easy one. After it is recovered z_2, z_3 become easy, then z_4, z_5, etc.

Because of this AMP will have a saddle to saddle dynamic, where the algorithm remains stuck for $\mathcal{O}(\log d)$ plateaus and quickly recovers new directions between them.



References

[1] Yatin Dandi, Emanuele Troiani, Luca Arnaboldi, Luca Pesce, Lenka Zdeborová, and Florent Krzakala. "The benefits of reusing batches for gradient descent in two-layer networks: Breaking the curse of information and leap exponents" *ICML* (2024)

[2] Emmanuel Abbe, Enric Boix Adserà, and Theodor Misiakiewicz. "Sgd learning on neural networks: leap complexity and saddle-to-saddle dynamics" *The Annals of Statistics, Ann. Statist.* (2018)

SPOC Statistical Physics Of Computation · IdePHICS Information, Learning & Physics Lab. · PSL Université Paris · ENS École Normale Supérieure