Solving Sparse & High-Dimensional-Output Regression via Compression

Renyuan Li[1] Zhehui Chen[2] Guanyi Wang [1]

NeurIPS 2024

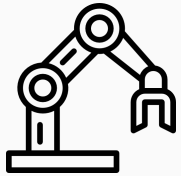
National University of Singapore[1] Google[2]

Motivation

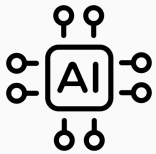
Regression for very high-dim output

- Strong interpretability,
- Computational scalability,
- Provable accuracy.

algorithmic trading



RL for high-dim decision



grounding in LLM

Problem Setting

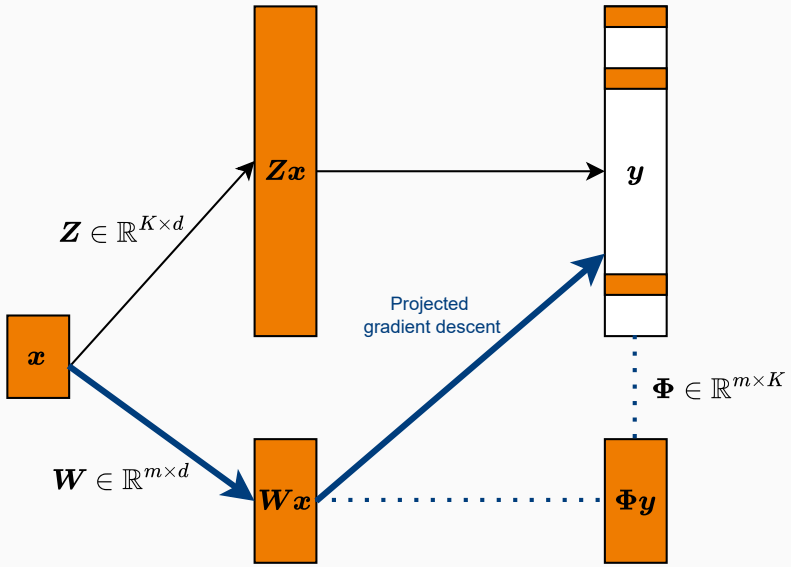
Given input $x \in \mathbb{R}^d$, predict high-dim output $y \in \mathbb{R}^K$,

$$y := \operatorname*{arg\,min}_{u \in \mathcal{Y}} \ \mathtt{dist}(u, \widehat{g}(x))$$

$$\text{with} \ \ \widehat{g} := \operatorname*{arg\,min}_{g \in \mathcal{G}} \ \frac{1}{n} \sum_{i=1}^{n} \ell(y^i, g(x^i)),$$

- high-dimensional output, i.e., $K \gg d$.
- sparsity constraint for interpretability, i.e., s-sparse output: $\mathcal{Y} \subseteq \{y \in \mathbb{R}^K \mid \|y\|_0 \leq s\}$.
- Using linear models $\mathcal{G}$, i.e., $g(x) = Zx$, for simplicity.

Framework



- Color of the boxes: orange/white – nonzero/zero components.
- Height of boxes: dimension of corresponding vector.
- Black arrows: the common approach. Blue arrows: our approach.
- $\boldsymbol{\Phi y}$: underlying feature for the output.

Contribution

Propose an **two-stage framework via compression** with

- improved computational efficiency
- the same order of generalization bounds before and after compression

Numerical Experiments: improved efficiency

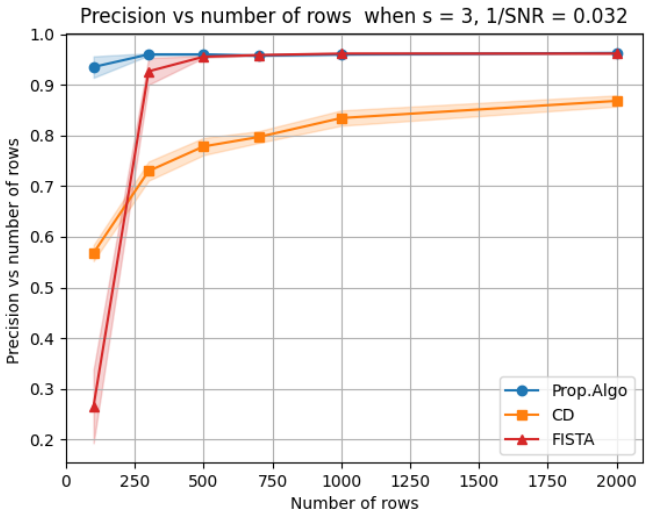
Running Time: Time Complexity Comparison for each prediction.

Dataset	Prop.Algo.	OMP	CD	FISTA
Synthetic Data	≈1 second	200-400 seconds	<1 second	<3 seconds
EURLex-4K	<1 second	20-80 seconds	<1 second	≈ 1 second
Wiki10-31K	<5 seconds	500-700 seconds	<5 seconds	5-10 seconds

- Enjoys a better running time for large scale instances

Numerical Experiments: better accuracy

Solution Accuracy



Experiment on synthetic data with $d = 10^4$, $K = 2 \times 10^4$, $s = 3$, and $n = 3 \times 10^4$.

- Outperforms the baselines on precision