Most Influential Subset Selection: Challenges, Promises, and Beyond

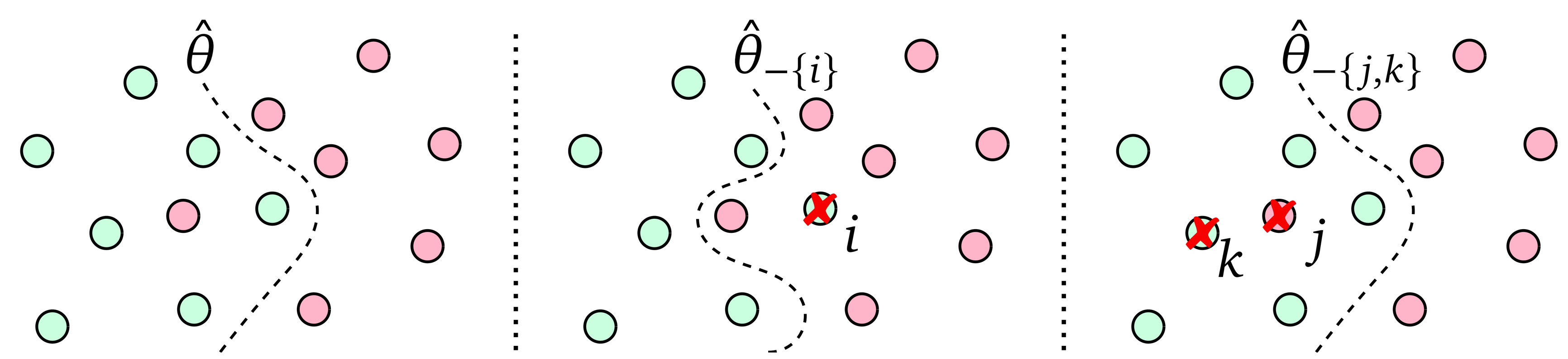
Yuzheng Hu Pingbang Hu Han Zhao Jiaqi W. Ma
University of Illinois Urbana-Champaign

Problem Formulation: Most Influential Subset Selection (MISS)

Given a train set $\{(x_i, y_i) \in \mathcal{X} \times \mathcal{Y}\}_{i \in [n]}$ and a loss $L(\cdot, \cdot)$, a prediction task aims to learn a predictor $f(\theta, \cdot) : \mathcal{X} \to \mathcal{Y}$ by ERM:

$$\hat{\theta} = \arg\min \frac{1}{n} \sum_{i=1}^{n} L(f(\theta, x_i), y_i).$$

For $S \subseteq [n]$, $\hat{\theta}_{-S}$:= opt sol. after excluding S from the train set.



Definition. The *actual effect* of S is defined as $A_{-S} = \phi(\hat{\theta}_{-S}) - \phi(\hat{\theta})$ w.r.t. a *target function* $\phi : \mathbb{R}^q \to \mathbb{R}$ (e.g., prediction).

Problem. The k-MISS problem: $S_{\text{opt},k} = \arg\max_{S \subseteq [n], |S| \le k} A_{-S}$.

Practical Relevance. MISS is a powerful diagnostic tool in social sciences (e.g., testing inferential *robustness*).

Dominant Approach: Influence-Based Greedy Heuristics

Procedure:
(1) Assign v_i per sample $\Rightarrow$ (2) Sort v_i's $\Rightarrow$ (3) Select top-k

Ex. Compute v_i's using *influence function*

$$\mathcal{I}_{-S} = \frac{1}{n} \nabla_\theta \phi(\hat{\theta})^\top H_{\hat{\theta}}^{-1} \sum_{i \in S} \nabla_\theta L(f(\hat{\theta}, x_i), y_i) =: \sum_{i \in S} v_i$$

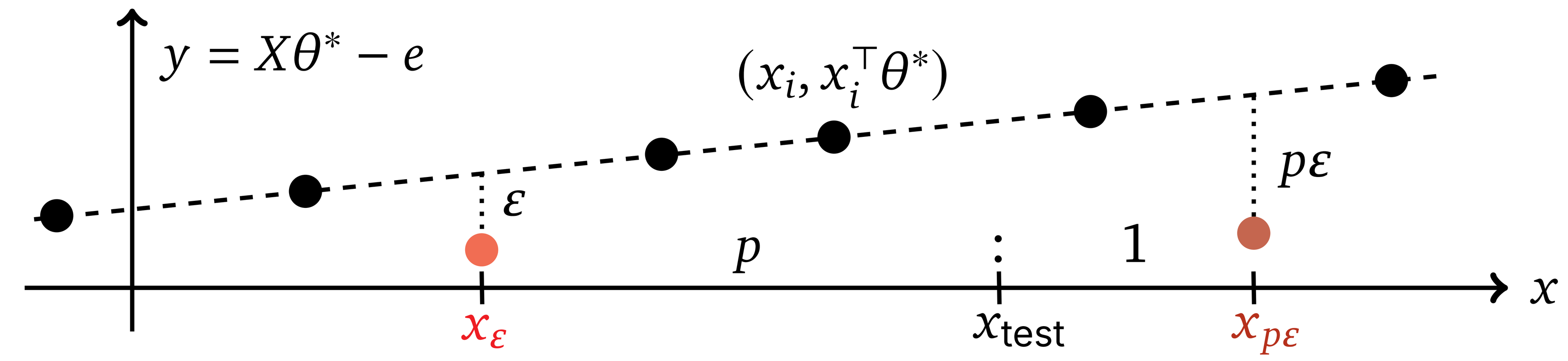
gives rise to ZAMinfluence [1].

Overview and Contributions

1. Analyze **failure modes** of influence-based greedy heuristics.
2. Prove theoretical **guarantees** of *adaptive greedy*, and
3. Demonstrating its empirical benefits.

Formal Analysis in Linear Regression

Consider a design matrix $X \in \mathbb{R}^{n \times d}$ and target vector $y \in \mathbb{R}^d$:



Target function: $\phi(\theta) = x_{\text{test}}^\top \theta$, where $x_{\text{test}} = (x_\varepsilon + p x_{p\varepsilon})/(p+1)$.

Greedy and its Pitfalls

- Pitfall 1: **Influence estimate is not accurate**

Actual Effect:
$$A_{-\{i\}} = x_{\text{test}}^\top \frac{N^{-1} x_i r_i}{1 - h_{ii}}$$

Influence Estimate:
$$\mathcal{I}_{-\{i\}} = x_{\text{test}}^\top N^{-1} x_i r_i$$

- $r_i = x_i^\top \hat{\theta} - y_i$
- $N = X^\top X$
- $H = X N^{-1} X^\top$
- $h_{ii} = H_{ii} > 0$, known as **leverage score** of x_i.

- Pitfall 2: **Sample influence is not additive**

Note: here we directly use the ground truth individual influence $A_{-\{i\}}$ instead of $\mathcal{I}_{-\{i\}}$ to perform greedy.

Theorem (Amplification). If there are c copies of x_ε and $x_{p\varepsilon}$, then there is some p s.t. greedy w.r.t. $A_{-\{i\}}$ fails in c-MISS .

Intuition: a group of samples with small individual effects, collectively can have larger effects. A special case:

$$\frac{A_{-\{i\}^c}}{A_{-\{i\}}} = \frac{c(1 - h_{ii})}{1 - c h_{ii}} > c \quad (\to \infty \text{ as } h_{ii} \to 1/c).$$

Theorem (Cancellation). There is some p s.t. x_ε and $x_{p\varepsilon}$ are the top 2 influential data (individually), but $A_{-\{x_\varepsilon, x_{p\varepsilon}\}} < A_{-\{x_{p\varepsilon}\}}$.

Intuition: effect of S < effect of $S' \subsetneq S$, i.e., removing $S \setminus S'$ induces a negative effect. This can happen for even $k = 2$.

Adaptive Greedy and its Promises

A natural extension: perform greedy **adaptively** [2]:

(1) Compute v_i's on current dataset
(2) Sort v_i's $\Rightarrow$ (3) Select (and remove) top-1

Theorem (Adaptivity & Cancellation). Suppose **cancellation** and $x_{p\varepsilon} \in S_{\text{opt},2}$. Then, the adaptive greedy solves 2-MISS.

Takeaway: Adaptivity captures more complex interactions between samples, while vanilla greedy only measures the contribution of each sample *solely* in relation to the full training set.

Experiments

Adaptive v.s. vanilla greedy on real data and non-linear models:

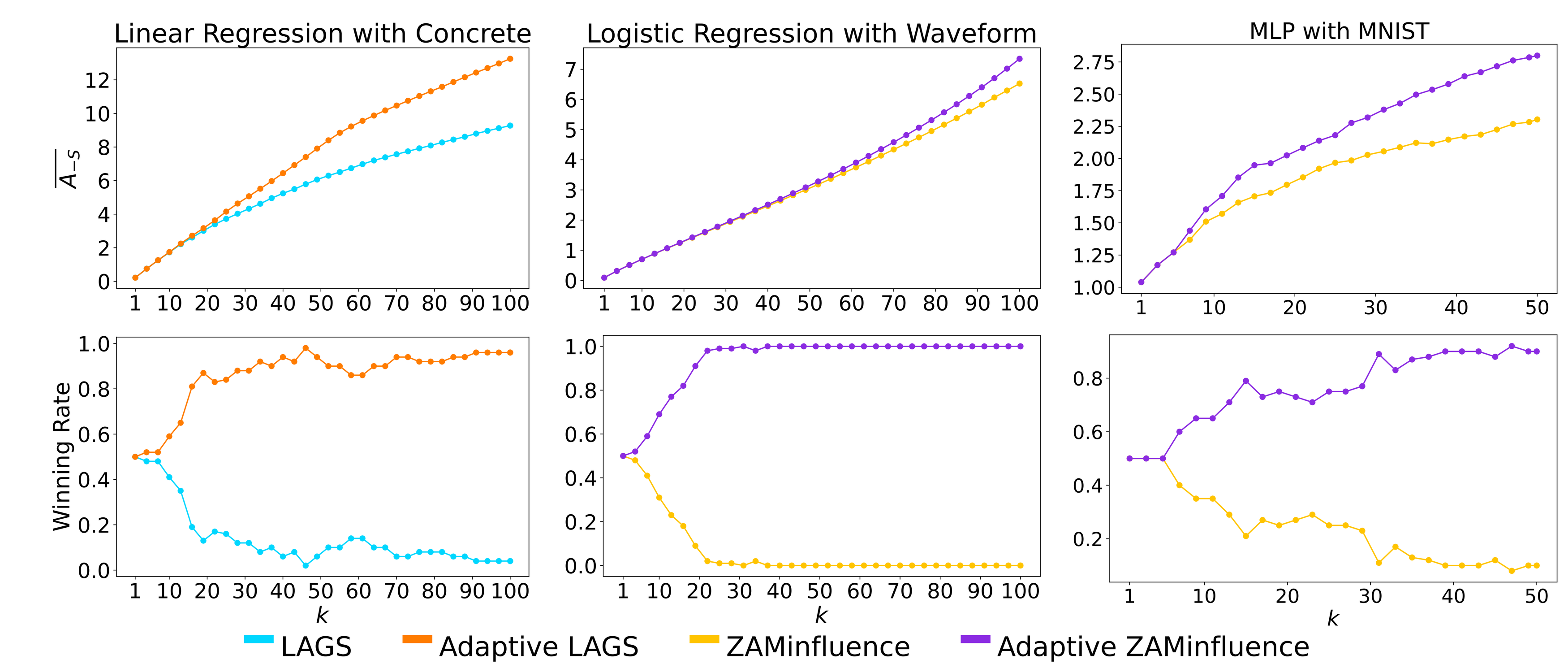


Figure 1. **Row 1:** $\overline{A_{-S}}$ measures the *averaged actual effect*. **Row 2:** *Winning rate* indicates the proportion of instances where one outperforms the other.

Further Discussions

1. Adaptive greedy is not a gold solution
2. Target function matters
3. Implications on linear datamodeling score (LDS) [3]

[1] Broderick, Giordano, and Meager. An automatic finite-sample robustness metric: when can dropping a little data make a big difference? *arXiv preprint arXiv:2011.14999* (2020).
[2] Kuschnig, Zens, and Cuaresma. Hidden in Plain Sight: Influential Sets in Linear Models. Tech. rep. CESifo, 2021.
[3] Park et al. TRAK: Attributing Model Behavior at Scale. ICML. 2023.