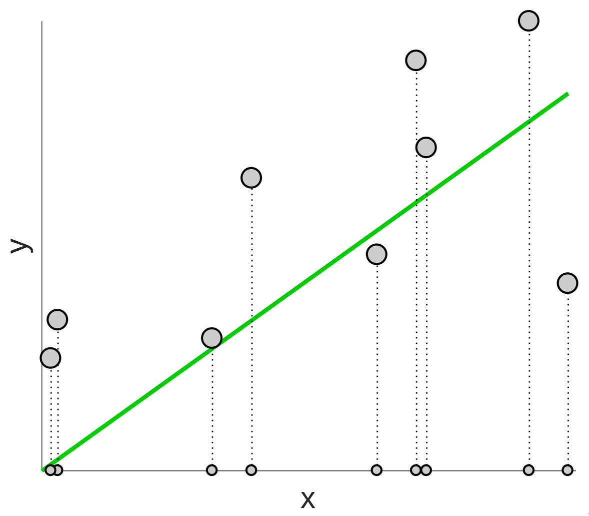
Foundations of Data Science
DATA SCIENCE INSTITUTE

Leveraged volume sampling for linear regression

Michał Dereziński — Berkeley, University of California
Manfred K. Warmuth — University of California Santa Cruz
Daniel Hsu — Columbia University in the City of New York

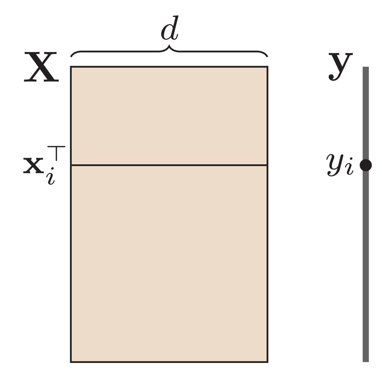
Data Science Institute, Columbia University

Setting: Linear regression



$$L(\mathbf{w}) = \sum_i (\mathbf{x}_i^\top \mathbf{w} - y_i)^2$$

$$\mathbf{w}^* = \underset{\mathbf{w}}{\operatorname{argmin}}\, L(\mathbf{w})$$



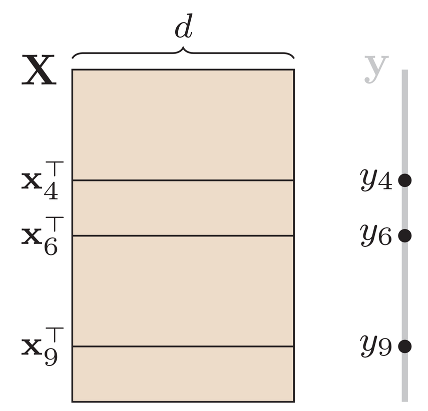
Problem: Expensive labels

- All $\mathbf{x}_i$ given
- Labels y_i unknown
- Learner can ask for a subset of labels

Goal: Unbiased estimators

Sample $S = \{4, 6, 9\}$

Receive y_4, y_6, y_9



Goal: Produce estimator $\widehat{\mathbf{w}}(S)$:

unbiased: $\mathbb{E}\big[\widehat{\mathbf{w}}(S)\big] = \mathbf{w}^*$

$1{+}\epsilon$-loss bound: $L\big(\widehat{\mathbf{w}}(S)\big) \le (1+\epsilon) L(\mathbf{w}^*)$

Label complexity $k(\epsilon)$:
smallest $|S|$ s.t. $\widehat{\mathbf{w}}(S)$ has a $1{+}\epsilon$*-loss bound*

Questions:

1. Which sampling distribution over S?
2. What estimator function $\widehat{\mathbf{w}}(S)$?
3. What is label complexity $k(\epsilon)$?

Prior work:

- Best unbiased estimator: $k(\epsilon) = O(\frac{d^2}{\epsilon})$
- Best biased estimator: $k(\epsilon) = O(\frac{d}{\epsilon})$

Our result:
Unbiased estimator with $k(\epsilon) = O(d \log d + \frac{d}{\epsilon})$

Prior work: Volume sampling

Distribution over all k-element subsets S:

$$\Pr(S) = \frac{\det(\mathbf{X}_S^\top \mathbf{X}_S)}{\binom{n-d}{k-d}\det(\mathbf{X}^\top \mathbf{X})}$$

Unbiased estimator:

$$\mathbb{E}\big[\widehat{\mathbf{w}}(S)\big] = \mathbf{w}^*$$

where $\widehat{\mathbf{w}}(S) = \overbrace{\underset{w}{\operatorname{argmin}} \sum_{i\in S} (\mathbf{x}_i^\top \mathbf{w} - y_i)^2}^{(\mathbf{X}_S)^+ \mathbf{y}_S}$

New lower bound

$$\mathbf{X} = \begin{pmatrix} \mathbf{I}_{d\times d} \\ \gamma\, \mathbf{I}_{d\times d} \\ \vdots \\ \gamma\, \mathbf{I}_{d\times d} \end{pmatrix} \qquad \mathbf{y} = \begin{pmatrix} \mathbf{1}_d \\ \mathbf{0}_d \\ \vdots \\ \mathbf{0}_d \end{pmatrix}$$

Theorem If $S \sim$ size k volume sampling, then for any $k \le \frac{n}{2}$, w.p. $\ge \frac{1}{4}$,

$$L(\widehat{\mathbf{w}}(S)) \ge \frac{3}{2} L(\mathbf{w}^*).$$

New algorithm

Determinantal rejection sampling

1: **Input:** $\mathbf{X} \in \mathbb{R}^{n\times d}$, $k \ge d$
2: $s \leftarrow \max\{k, 4d^2\}$
3: **repeat**
4: Sample $\pi_1, \dots, \pi_s$ i.i.d. $\sim (l_1, \dots, l_n)$
5: $Acc \sim$ Bernoulli$\Big(\frac{\det(\frac{1}{s}\sum_i \frac{d}{l_{\pi_i}} \mathbf{x}_{\pi_i}\mathbf{x}_{\pi_i}^\top)}{\det(\mathbf{X}^\top\mathbf{X})}\Big)$
6: **until** $Acc =$ true
7: $S \leftarrow$ VolumeSample$\Big(\big\{\frac{1}{\sqrt{l_{\pi_i}}}\mathbf{x}_{\pi_i}\big\}_i, k\Big)$
8: **return** π_S

Our approach: Leverage score rescaled volume sampling

Idea: Use i.i.d. and joint sampling

$\pi \in \{1..n\}^k$

$$\Pr(\pi) \sim \overbrace{\Big(\prod_i l_{\pi_i}\Big)}^{\text{leverage scores}} \overbrace{\det\Big(\sum_i \frac{1}{l_{\pi_i}} \mathbf{x}_{\pi_i}\mathbf{x}_{\pi_i}^\top\Big)}^{\text{volume sampling}}$$

$$\widehat{\mathbf{w}}(\pi) = \underset{\mathbf{w}}{\operatorname{argmin}} \sum_i \frac{1}{l_{\pi_i}} \big(\mathbf{x}_{\pi_i}^\top \mathbf{w} - y_{\pi_i}\big)^2$$

leverage scores: $l_i = \mathbf{x}_i^\top (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{x}_i$

Theorem For size $k(\epsilon) = O(d \log d + \frac{d}{\epsilon})$ leveraged volume sampling:

$$\mathbb{E}\big[\widehat{\mathbf{w}}(\pi)\big] = \mathbf{w}^* \quad \text{and}$$

w.h.p. $L\big(\widehat{\mathbf{w}}(\pi)\big) \le (1+\epsilon) L(\mathbf{w}^*)$

Open: unbiased estimator with $k(\epsilon) = O(\frac{d}{\epsilon})$

Time complexity

Theorem Sampling time is

$$O\Big((d^4 + kd^2)\ln(1/\delta)\Big) \quad \text{w.p. } \ge 1-\delta$$

New: Sampling time does not depend on n

Preprocessing: computing leverage scores

1. *Exact:* $O(nd^2)$
2. *Approximate:* $O(nd \log n + d^4 \log d)$

Loss comparison on Libsvm datasets