Subsampling for Ridge Regression via Regularized Volume Sampling

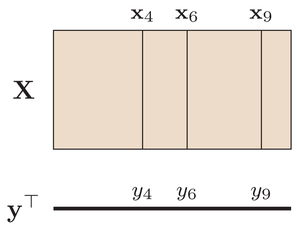
Michał Dereziński Manfred K. Warmuth

Regression with expensive labels

Linear model: $\mathbf{y} = \mathbf{X}^\top \mathbf{w}^* + \boldsymbol{\xi}, \quad \text{where } \mathbb{E}[\boldsymbol{\xi}] = 0, \quad \text{Var}[\boldsymbol{\xi}] \preceq \sigma^2 \mathbf{I}$

Given: n points $\mathbf{x}_i \in \mathbb{R}^d$
labels $y_i \in \mathbb{R}$ are hidden

- Select $S = \{4, 6, 9\}$
- Receive y_4, y_6, y_9
- Predict $\widehat{\mathbf{w}}(S)$



Mean Squared Prediction Error

Goal: Minimize $\text{MSPE}(\widehat{\mathbf{w}}(S)) = \mathbb{E}\,\frac{1}{n}\sum_{i=1}^{n}(\mathbf{x}_i^\top \widehat{\mathbf{w}}(S) - \mathbf{x}_i^\top \mathbf{w}^*)^2$

Optimal subsampling for ridge regression

Ridge estimator: $\widehat{\mathbf{w}}^*_\lambda(S) = \overbrace{\operatorname{argmin}_{\mathbf{w}} \|\mathbf{X}_S^\top \mathbf{w} - \mathbf{y}_S\|^2 + \lambda\|\mathbf{w}\|^2}^{(\mathbf{X}_S\mathbf{X}_S^\top+\lambda\mathbf{I})^{-1}\mathbf{X}_S\mathbf{y}_S}$

Question: How to best select subset S of size s?

Main Theorem
If $\lambda \leq \frac{\sigma^2}{\|\mathbf{w}^*\|^2}$, then there is a distribution over subsets S of size s

$$\text{s.t.} \qquad \text{MSPE}(\widehat{\mathbf{w}}^*_\lambda(S)) \leq \frac{\sigma^2 d_\lambda}{s - d_\lambda + 1} = O\left(\frac{\sigma^2 d_\lambda}{s}\right),$$

where $d_\lambda = \sum_{i=1}^{d} \frac{\lambda_i}{\lambda_i + \lambda} < d, \quad \{\lambda_i\}$ - eigenvalues of $\mathbf{X}\mathbf{X}^\top$

Lower-bound
There is $\mathbf{X}$ and $\mathbf{w}^*$ s.t. for all S of size s

$$\text{MSPE}(\widehat{\mathbf{w}}^*_\lambda(S)) \geq \frac{\sigma^2 d_\lambda}{s + d_\lambda} = \Omega\left(\frac{\sigma^2 d_\lambda}{s}\right).$$

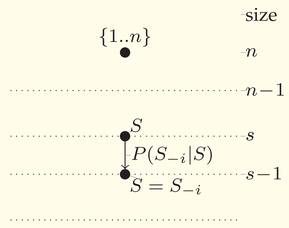
Any i.i.d. sampling is worse

How many labels needed to get $\text{MSPE}(\widehat{\mathbf{w}}^*_\lambda(S)) \leq \sigma^2$?

Our method:	$2\,d_\lambda$
Any i.i.d. sampling:	$\Omega(d_\lambda \ln(d_\lambda))$

Optimal subsampling distribution must be joint!

Regularized volume sampling



Start with $S = \{1..n\}$

Sample index $i \in S$
Go to set $S_{-i} = S - \{i\}$

Repeat until desired size

$$P(S_{-i}|S) \sim \frac{\det(\mathbf{X}_{S_{-i}}\mathbf{X}_{S_{-i}}^\top + \lambda\mathbf{I})}{\det(\mathbf{X}_S\mathbf{X}_S^\top + \lambda\mathbf{I})} = 1 - \mathbf{x}_i^\top(\mathbf{X}_S\mathbf{X}_S^\top + \lambda\mathbf{I})^{-1}\mathbf{x}_i$$

Proof of Main Theorem

Key Expectation Bound
For λ-regularized volume sampling of set S of size s

$$\mathbb{E}_S\,(\mathbf{X}_S\mathbf{X}_S^\top + \lambda\mathbf{I})^{-1} \preceq \frac{n - d_\lambda + 1}{s - d_\lambda + 1}(\mathbf{X}\mathbf{X}^\top + \lambda\mathbf{I})^{-1}.$$

Next, we do bias-variance decomposition for fixed S:

$$\text{MSPE}(\widehat{\mathbf{w}}^*_\lambda(S)) = (\text{bias})^2 + \text{variance} \leq \frac{\sigma^2}{n} tr(\mathbf{X}^\top(\mathbf{X}_S\mathbf{X}_S^\top + \lambda\mathbf{I})^{-1}\mathbf{X}).$$

Then, we take expectation over S and apply the bound.

Fast volume sampling algorithm

Simple algorithm: Update distribution $P(S_{-i}|S)$ at every step

Runtime: $\overbrace{\text{steps}}^{n-s} \times \overbrace{\text{update}}^{O(n\,d)} = O(n^2 d)$

Problem: Quadratic dependence on n

Idea: Rejection sampling from distribution $P(S_{-i}|S)$

1. Sample i uniformly from set S,
2. Compute $h_i = 1 - \mathbf{x}_i^\top(\mathbf{X}_S\mathbf{X}_S^\top + \lambda\mathbf{I})^{-1}\mathbf{x}_i$,
3. With probability $1 - h_i$ reject and go back to 1.

one trial

We show: Number of rejection trials per step is constant w.h.p.

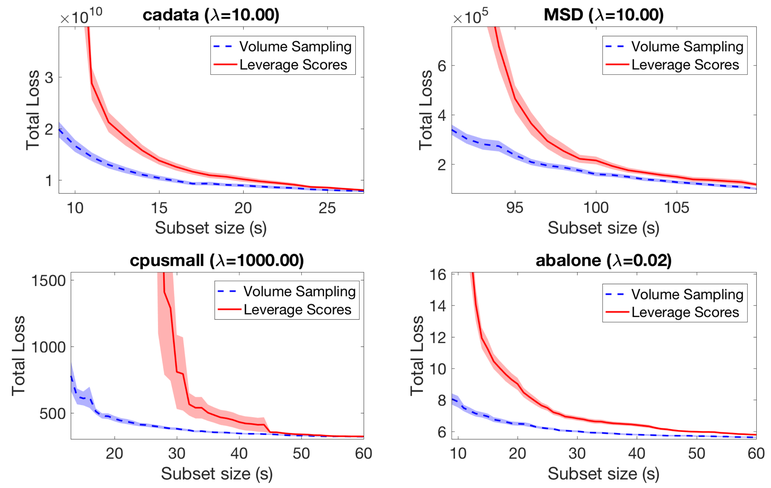
Runtime: $\overbrace{\text{steps}}^{n-s} \times \overbrace{\text{trials per step}}^{O(1)} \times \overbrace{\text{compute } h_i}^{O(d^2)} = O(nd^2)$

Result: Linear dependence on n

Volume sampling vs i.i.d. leverage scores

Leverage score sampling: examples are selected i.i.d.
w.p. $P(i) = (\mathbf{x}_i^\top(\mathbf{X}\mathbf{X}^\top)^{-1}\mathbf{x}_i)/d$.

For small subsets, volume sampling is better than leverage scores.



Total Loss: $L(\widehat{\mathbf{w}}^*_\lambda(S)) \overset{\text{def}}{=} \frac{1}{n}\|\mathbf{X}^\top\widehat{\mathbf{w}}^*_\lambda(S) - \mathbf{y}\|^2.$

Fast volume sampling vs state-of-the-art

Two implementations of regularized volume sampling:

1.	**RegVol**	based on existing method	runtime: $O(n^2 d)$
2.	**FastRegVol**	our new algorithm	runtime: $O(nd^2)$

Our algorithm is much faster when $n \gg d$.