FAST RIDGE REGRESSION WITH SUBSAMPLED RANDOMIZED HADAMARD TRANSFORM

Yichao Lu, Paramveer Dhillon, Dean P Foster and Lyle Ungar

{yichaolu | foster }@wharton.upenn.edu , {dhillon | ungar}@cis.upenn.edu

Penn
University of Pennsylvania

Contribution

We propose a fast randomized algorithm **SRHT-DRR** for solving large scale Ridge Regression when the number of features is much larger than the number of observations ($p \gg n \gg 1$). The exact solution in this case costs $O(n^2p)$ FLOPS. Our algorithm costs only $O(np\log(p_{\text{subs}})+n^2p_{\text{subs}})$ FLOPS with a performance guarantee under the fixed design setting. Here $p_{\text{subs}} \ll p$ is a parameter which controls the trade-off between accuracy and efficiency.

Motivation: Approximate $\mathbf{XX}^\top$

Consider the dual formulation of Ridge Regression:

$$\hat{\alpha}_\lambda = \arg\min_{\alpha\in n\times 1} \frac{1}{n}\|Y-\mathbf{K}\alpha\|^2 + \lambda\alpha^\top\mathbf{K}\alpha$$

where $\mathbf{K}=\mathbf{XX}^\top$. The exact solution is

$$\hat{\alpha}_\lambda = (\mathbf{K}+n\lambda I_n)^{-1}Y$$

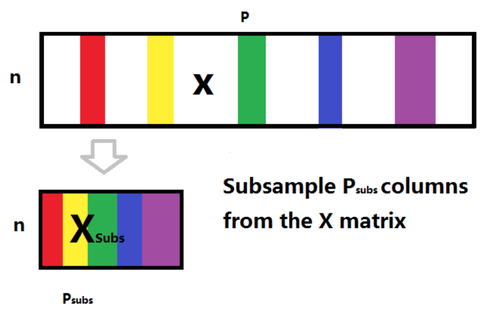
Problem: Computing $\mathbf{XX}^\top$ costs $O(n^2p)$, slow for large n, p.
Solution: Construct $\mathbf{X}_{\text{subs}} \in n \times p_{\text{subs}}$ by subsampling p_{subs} columns from $\mathbf{X}$.
Use $\mathbf{K}_{\text{subs}} = \frac{p}{p_{\text{subs}}}\mathbf{X}_{\text{subs}}\mathbf{X}_{\text{subs}}^\top$ as an approximation of $\mathbf{K}$.

Why Precondition

Good Case: $\mathbf{X} = \begin{pmatrix} 1 & 1 & 1 & \dots & 1 \\ 1 & 1 & 1 & \dots & 1 \end{pmatrix}$

Bad Case: $\mathbf{X} = \begin{pmatrix} 100 & 0 & 0 & \dots & 0 \\ 100 & 0 & 0 & \dots & 0 \end{pmatrix}$

An extra preconditioning step is necessary before subsampling.



Algorithm Sketch

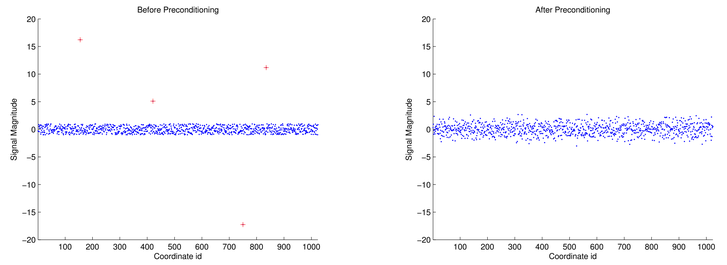
SRHT-DRR
Input: data $\mathbf{X}, Y$, hyperparameter λ and subsample size p_{subs}.
Output: $\hat{\alpha}_{\mathbf{SRHT-DRR}}$, the dual weight vector.

- Precondition: Right multiply $\mathbf{X}$ by a $p \times p$ structured random matrix called a Randomized Hadamard Transform.
- Subsampling: Subsample p_{subs} columns from the preconditioned matrix and compute $\mathbf{K}_{\text{subs}}$.
- Compute $\hat{\alpha}_{\mathbf{SRHT-DRR}} = (\mathbf{K}_{\text{subs}}+n\lambda I_n)^{-1}Y$

Properties of Randomized Hadamard Transform

- Randomized Hadamard Transform smears energy among all columns.

A Toy Example



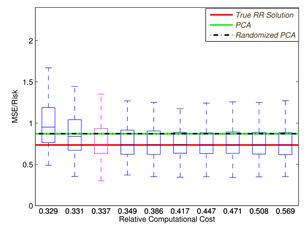
Left: Before preconditioning there are some high energy coordinates (red).
Right: The energy becomes widely spread after preconditioning.

- Randomized Hadamard Transform multiplies fast due to its recursive structure. In **SRHT-DRR** preconditioning costs is only $O(np\log(p_{\text{subs}}))$ FLOPS.

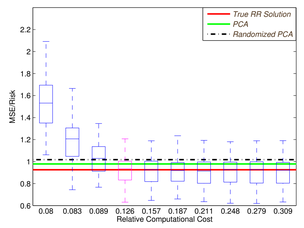
Experiments

Simulation

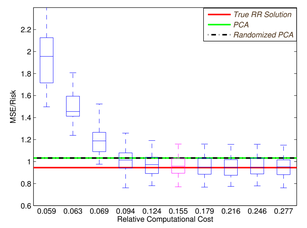
$n = 20, p = 8192$



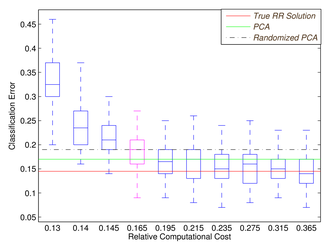
$n = 100, p = 8192$



$n = 200, p = 8192$



Real Data
n=100,p=10000



- **SRHT-DRR** is implemented on both simulated and real datasets with different p_{subs}. The corresponding relative computational cost and prediction accuracy are recorded. Here relative computational cost $= \frac{\text{FLOPS of } \mathbf{SRHT-DRR}}{\text{FLOPS of the exact ridge solution}}$.
- Simulation data: $\mathbf{X}$ of different sizes and Y are generated from the fixed design model. We use MSE to evaluate prediction accuracy.
- Real data (ARCENE): Distinguish cancer versus normal patterns from mass-spectrometric data. Prediction accuracy is evaluated by classification error on test data.
- The exact ridge solution, PCA and randomized PCA are considered as baselines. Under $p \gg n \gg 1$ assumption all these algorithms are slow.
- We suggest to set $p_{\text{subs}} \approx 5n$ (pink boxes in the plots).