A Deterministic Streaming Sketch for Ridge Regression

Benwei Shi and Jeff M. Phillips

THE UNIVERSITY OF UTAH®

Background: How to calculate Hessian inverse matrix efficiently?

- Gradient-based pruning archived great performance than other methods
- The target of pruning a network:

$$\min_{\delta\mathbf{w}\in\mathbb{R}^d} \delta L = \min_{\delta\mathbf{w}\in\mathbb{R}^d} \Big(L(\mathbf{w}+\delta\mathbf{w}) - L(\mathbf{w})\Big)$$

- Approximate the object function after pruning by a Taylor series

$$L(\mathbf{w}+\delta\mathbf{w}) = L(\mathbf{w}) + \nabla_{\mathbf{w}} L^\top \delta\mathbf{w} + \tfrac{1}{2}\delta\mathbf{w}^\top \mathbf{H}\, \delta\mathbf{w} + O(||\delta\mathbf{w}||^3)$$

$$\delta L \approx \frac{1}{2}\delta\mathbf{w}^\top \mathbf{H}\, \delta\mathbf{w} \quad \rightarrow \quad \delta\mathbf{w}^* = \frac{-w_q \mathbf{H}^{-1}\mathbf{e}_q}{[\mathbf{H}^{-1}]_{qq}}$$

Background: How to calculate Hessian inverse matrix efficiently?

- Estimate the Hessian with Fisher matrix:

$$\bar{H} = \frac{1}{N}\sum_{n=1}^{N} \underbrace{\nabla \ell\left(\mathbf{y}_n, f\left(\mathbf{x}_n; \mathbf{w}\right)\right)}_{\nabla \ell_n} \nabla \ell\left(\mathbf{y}_n, f\left(\mathbf{x}_n; w\right)\right)^{\top} \quad = \frac{1}{N} G^T G, G \in R^{N \times D}$$

- Calculate its inverse with WoodBurry methods:

$$\widehat{F}_{n+1} = \widehat{F}_n + \frac{1}{N} \nabla \ell_{n+1} \nabla \ell_{n+1}{}^{\top}, \quad \text{where} \quad \widehat{F}_0 = \lambda I_d.$$

$$\widehat{F}_{n+1}^{-1} = \widehat{F}_n^{-1} - \frac{\widehat{F}_n^{-1} \nabla \ell_{n+1} \nabla \ell_{n+1}^{\top} \widehat{F}_n^{-1}}{N + \nabla \ell_{n+1}^{\top} \widehat{F}_n^{-1} \nabla \ell_{n+1}}, \quad \text{where} \quad \widehat{F}_0^{-1} = \lambda^{-1} I_d.$$

- Time and space complexity:

$$O(N \times D^2)$$

Ridge Regression

Why regularization is needed?

Given : $\mathbf{A} \in \mathbb{R}^{n\times d}$, $\mathbf{b} \in \mathbb{R}^{n}$ in rows.

Goal : $\mathbf{x}_\gamma \equiv \underset{\mathbf{x}\in\mathbb{R}^d}{\operatorname{argmin}}(\|\mathbf{A}\mathbf{x}-\mathbf{b}\|^2 + \boxed{\gamma\|\mathbf{x}\|^2})$

$$= \boxed{(\mathbf{A}^\top\mathbf{A}+\gamma\mathbf{I})^{-1}}\mathbf{A}^\top\mathbf{b} \qquad n \gg d$$

Space: $O(d^2)$ $O(d)$

Time: $O(d^3+nd^2)$ $O(nd)$

$(1\pm\varepsilon)$-relative error

with only roughly $\ell = O(1/\varepsilon)$ rows

$O(d\ell) = O(d/\varepsilon) = o(d^2)$ space

Running time: $O(nd\ell)$

Our approach: Use Frequent Directions (FD) (Liberty, 2013) to estimate $\mathbf{A}^\top\mathbf{A}$ in stream.

A possible solution: approximate A with a much smaller matrix!

computes the SVD of A and approximates it using the first k singular vectors x

that $\|Ax\| \geq t$

Frequent-items[1]

- m items $a_1, \ldots, a_m$ and a stream $A_1, \ldots, A_n$ of item appearances
- The frequency f_i of item a_i stands for the number of times a_i appears in the stream
- Goal: approximate frequencies g_j such that $|f_j - g_j| \leq n/\ell$

  use $O(\ell)$ space

```
x[1]...x[N] is the input sequence
K is a set of symbols initially empty
count is an array of integers indexed by K
for i:= 1,...,N do
 {if x[i] is in K then count[x[i]] := count[x[i]] + 1
    else  {insert x[i] in K, set count[x[i]] := 1}
   if |K| > 1/theta then
      for all a in K do
         { count[a] := count[a] - 1,
          if count[a] = 0 then delete a from K}}
output K
```

periodically deletes ℓ different elements

if one sets $\ell > 1/\varepsilon$, then any item that appears more than εn times in the stream must appear in the final sketch

Ref: [1] A Simple Algorithm for Finding Frequent Elements in Streams and Bags, 2003

Frequent-items[2]: extend to matrix

- Let A be a matrix as a stream of its rows.
- let us constrain the rows of A to be basis vector $A_i \in \{e_1, \ldots, e_m\}$;
- $A_i = e_j$ If the i'th element in the stream is a_j
- $f_j = \|Ae_j\|^2$
- Goal: $g_j = \|Be_j\|^2$ is a good approximation to f_j $\quad B \in \mathbb{R}^{\ell \times m}$

$$|f_j - g_j| \le n/\ell \quad n = \|A\|_f^2 \longrightarrow |\|Ae_j\|^2 - \|Be_j\|^2| \le \|A\|_f^2/\ell.$$

Ref: [1] Simple and Deterministic Matrix Sketching, KDD13

Frequent-directions[2]

- Given any matrix $A \in \mathbb{R}^{n \times m}$ the algorithm processes the rows of A one by one and produces a sketch matrix $B \in \mathbb{R}^{\ell \times m}$, such that

$$B^T B \prec A^T A \text{ and } \|A^T A - B^T B\| \leq 2\|A\|_f^2/\ell \,.$$

- periodically 'shrinks' ℓ orthogonal vectors by roughly the same amount
- Goal: to uncover any unit vector (direction) in space x for which

$$\|Ax\|^2 \geq \varepsilon\|A\|_2^2 \text{ by taking } \ell > 2r/\varepsilon$$

$$r = \|A\|_f^2/\|A\|_2^2 \text{ denotes the numeric rank of } A$$

Ref: [1] Simple and Deterministic Matrix Sketching, KDD13

Frequent-directions[1]

Algorithm 1 *Frequent-directions*

Input: ℓ, $A \in \mathbb{R}^{n \times m}$
$B \leftarrow$ all zeros matrix $\in \mathbb{R}^{\ell \times m}$
for $i \in [n]$ **do**
  Insert A_i into a zero valued row of B
  if B has no zero valued rows **then**
    $[U, \Sigma, V] \leftarrow \text{SVD}(B)$
    $C \leftarrow \Sigma V^T$ # Only needed for proof notation
    $\delta \leftarrow \sigma^2_{\ell/2}$
    $\check{\Sigma} \leftarrow \sqrt{\max(\Sigma^2 - I_\ell \delta, 0)}$
    $B \leftarrow \check{\Sigma} V^T$ # At least half the rows of B are all zero
  end if
end for
Return: B

Ref: [1] Simple and Deterministic Matrix Sketching, KDD13

Frequent Direction for Ridge Regression

Given : $\mathbf{A} \in \mathbb{R}^{n \times d}, \mathbf{b} \in \mathbb{R}^{n}$ in rows.

Goal : $\mathbf{x}_{\gamma} \equiv \underset{\mathbf{x} \in \mathbb{R}^{d}}{\operatorname{argmin}}(\|\mathbf{A}\mathbf{x} - \mathbf{b}\|^2 + \boxed{\gamma\|\mathbf{x}\|^2})$

$$= (\mathbf{A}^\top \mathbf{A} + \gamma \mathbf{I})^{-1} \mathbf{A}^\top \mathbf{b} \qquad n \gg d$$

Our approach: Use Frequent Directions (FD) (Liberty, 2013) to estimate $\mathbf{A}^\top\mathbf{A}$ in stream.

$(1 \pm \varepsilon)$-relative error with only roughly $\ell = O(1/\varepsilon)$ rows

$O(d\ell) = O(d/\varepsilon) = o(d^2)$ space

Running time: $O(nd\ell)$

Frequent Directions Ridge Regression

Algorithm FDRR (Based on Frequent Directions)

Input: $\mathbf{A} \in \mathbb{R}^{n \times d}, \mathbf{b} \in \mathbb{R}^{n}, \ell, \gamma$

$\Sigma \leftarrow \mathbf{0}^{\ell \times \ell}, \mathbf{V}^\top \leftarrow \mathbf{0}^{\ell \times d}, \mathbf{c} \leftarrow \mathbf{0}^{d}$

$\mathbf{C} = \Sigma \mathbf{V}^\top$

Initialization

for batch $\mathbf{A}_\ell \in \mathbf{A}, \mathbf{b}_\ell \in \mathbf{b}$ do

size ℓ batch $\mathbf{A}_\ell$ and $\mathbf{b}_\ell$

$\quad \_, \Sigma', \mathbf{V}'^\top \leftarrow \mathbf{svd}\left(\left[\mathbf{C}^\top; \mathbf{A}_\ell^\top\right]^\top\right)$

$\quad \Sigma \leftarrow \sqrt{\Sigma_\ell'^2 - \sigma_{\ell+1}^2 \mathbf{I}_\ell}$

$\quad \mathbf{V} \leftarrow \mathbf{V}_\ell'$

$\quad \mathbf{C} = \Sigma \mathbf{V}^\top$

Frequent Directions

$\quad \mathbf{c} \leftarrow \mathbf{c} + \mathbf{A}_\ell^\top \mathbf{b}_\ell$

Compute $\mathbf{A}^\top \mathbf{b}$ on the fly

end for

$\mathbf{c}' = \mathbf{V}^\top \mathbf{c}$

$\hat{\mathbf{x}}_\gamma \leftarrow \mathbf{V}(\Sigma^2 + \gamma \mathbf{I})^{-1} \mathbf{c}' + \gamma^{-1}(\mathbf{c} - \mathbf{V}\mathbf{c}')$

return $\hat{\mathbf{x}}_\gamma$

Return the solution $\hat{\mathbf{x}}_\gamma = (\mathbf{C}^\top \mathbf{C} + \gamma \mathbf{I})^{-1} \mathbf{A}^\top \mathbf{b} = (\mathbf{V}\Sigma^2\mathbf{V}^\top + \gamma \mathbf{I})^{-1} \mathbf{c}$

Recall the RR solution $\mathbf{x}_\gamma = (\mathbf{A}^\top \mathbf{A} + \gamma \mathbf{I})^{-1} \mathbf{A}^\top \mathbf{b}$

$O(d\ell) = O(d/\varepsilon) = o(d^2)$ space

Running time: $O(nd\ell)$

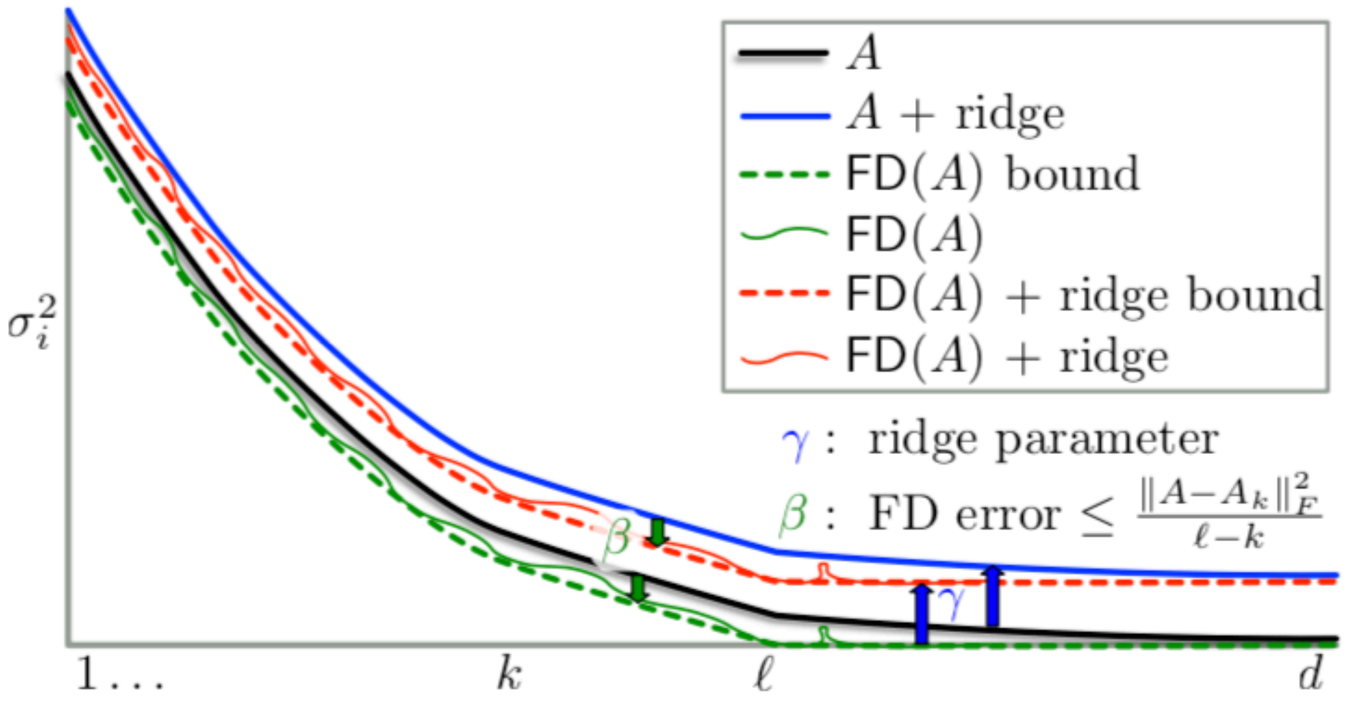


Figure 1: A figurative illustration of possible eigenvalues (σ_i^2) of a covariance matrices $\mathbf{A}^\top \mathbf{A}$ and variants when approximated by FD or adding a ridge term $\gamma \mathbf{I}$, along sorted eigenvectors.

Frequent Directions Ridge Regression

Running time: $O(nd\ell)$, required space: $O(d\ell)$. Note that $\ell \le d$.

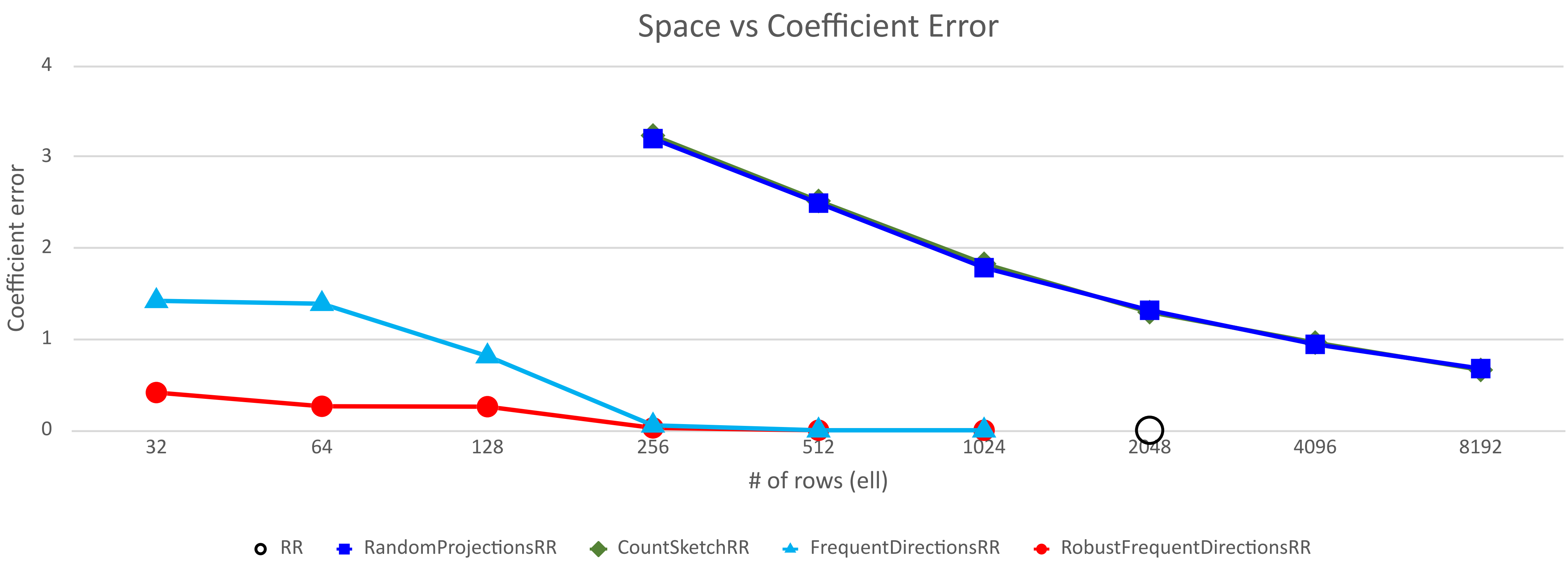
If

$$\ell \ge \frac{\|\mathbf{A}-\mathbf{A}_k\|_F^2}{\varepsilon\gamma} + k, \qquad \text{or} \qquad \gamma \ge \frac{\|\mathbf{A}-\mathbf{A}_k\|_F^2}{\varepsilon(\ell-k)}$$

Then

- $\left\|\hat{\mathbf{x}}_\gamma - \mathbf{x}_\gamma\right\| \le \varepsilon \left\|\mathbf{x}_\gamma\right\|$, or the coefficient error $\frac{\left\|\hat{\mathbf{x}}_\gamma - \mathbf{x}_\gamma\right\|}{\left\|\mathbf{x}_\gamma\right\|} \le \varepsilon$
- $\left|\hat{\mathbf{x}}_\gamma^\top \mathbf{a} - \mathbf{x}_\gamma^\top \mathbf{a}\right| \le \varepsilon \left\|\mathbf{x}_\gamma\right\| \|\mathbf{a}\|$ for any $\mathbf{a} \in \mathbb{R}^d$
- $\mathcal{B}^2\left(\hat{\mathbf{x}}_\gamma\right) \le \left(1 + \frac{\varepsilon^2}{\gamma^2}\|\mathbf{A}\|_2^4\right)\mathcal{B}^2\left(\mathbf{x}_\gamma\right)$
- $\mathcal{V}\left(\hat{\mathbf{x}}_\gamma\right) \le \left(1 + \frac{1}{\gamma}\|\mathbf{A}\|_2^2\right)\mathcal{V}\left(\mathbf{x}_\gamma\right)$

Experiments



Experiments