New stochastic sketching methods for Big Data Ridge Regression

Cheikh Saliou Touré

Student at ENS Cachan

Tutor : Robert Gower

Inria Paris (Sierra department)

July, 2017

Abstract

//

Contents

1. *Randomized Newton Method*

1.1 Algorithm

1.2 Convergence rate (draft)

1.2.1 General case

A is a $n \times n$ positive definite matrix representing our problem.

For C any subset of $\{1, \ldots, n\}$ of length s, we denote by I_C the $s \times n$ matrix which rows are $\left\{ e_i^T \right\}_{i \in C}$ up to a permutation, where $\{e_i\}_{i=1,\ldots,n}$ is a canonical basis of $\mathbb{R}^n$.

Throughout the computations, we denote by $Z = A I_C^T (I_C A I_C^T)^{-1} I_C A$. That is a quantity that intervenes in the computation of the convergence rate.

The convergence rate is defined by $\rho = 1 - \lambda_{min}(A^{-\frac{1}{2}} E[Z] A^{-\frac{1}{2}})$.

By defiition, $A^{-\frac{1}{2}} E[Z] A^{-\frac{1}{2}} = \sum_i p_i A^{\frac{1}{2}} I_{C_i}^T (I_{C_i} A I_{C_i}^T)^{-1} I_{C_i} A^{\frac{1}{2}}$

for any $i \in \{1, \ldots, n\}$, $A^{\frac{1}{2}} I_{C_i}^T (I_{C_i} A I_{C_i}^T)^{-1} I_{C_i} A^{\frac{1}{2}}$ is a projection matrix and then its eigenvalues are a nonempty subset of $\{0, 1\}$.

Since λ_{max} is convex, we obtain that :

$$0 \leqslant \lambda_{min}(A^{-\frac{1}{2}} E[Z] A^{-\frac{1}{2}}) \leqslant \lambda_{max}(A^{-\frac{1}{2}} E[Z] A^{-\frac{1}{2}}) \leqslant \sum_i p_i \lambda_{max}(A^{\frac{1}{2}} I_{C_i}^T (I_{C_i} A I_{C_i}^T)^{-1} I_{C_i} A^{\frac{1}{2}}) \leqslant 1.$$

Denote by $\mathbf{C} = (I_{C_1}^T, \ldots, I_{C_r}^T)$ which is of size $n \times rs$.

$A^{-\frac{1}{2}} E[Z] A^{-\frac{1}{2}} = (A^{\frac{1}{2}} \mathbf{C} D)(D \mathbf{C}^T A^{\frac{1}{2}})$ where
$D = \text{diag}(\sqrt{p_1}(I_{C_1} A I_{C_1}^T)^{-\frac{1}{2}}, \ldots, \sqrt{p_r}(I_{C_r} A I_{C_r}^T)^{-\frac{1}{2}}) \in \mathcal{M}_{rs}(\mathbb{R})$

> **Proposition 1.2.1**
>
> $$\lambda_{min}(A^{-\frac{1}{2}} E[Z] A^{-\frac{1}{2}}) \geqslant \binom{n-1}{s-1} \frac{\lambda_{min}(A)}{\lambda_{max}(A)} \min_i p_i$$

Proof :

$\lambda_{min}(A^{-\frac{1}{2}}E[Z]A^{-\frac{1}{2}}) \geqslant \lambda_{min}(\mathbf{C}^T A \mathbf{C})\lambda_{min}(D^2)$

$\lambda_{min}(D^2) = \min_i \dfrac{p_i}{\lambda_{max}(I_{C_i}AI_{C_i}^T)} \geqslant \min_i \dfrac{p_i}{\lambda_{max}(I_{C_i}^T I_{C_i})\lambda_{max}(A)} \geqslant \min_i \dfrac{p_i}{\lambda_{max}(A)}$, since for any $i \in$

$\{1, \ldots, n\}$, for any x in $\mathbb{R}^n$ $\left\langle I_{C_i}^T I_{C_i} x \,|\, x \right\rangle = \|I_{C_i}x\|^2 \leqslant \|x\|^2$ and then $\lambda_{max}(I_{C_i}^T I_{C_i}) \leqslant 1$.

Therefore, $\lambda_{min}(A^{-\frac{1}{2}}E[Z]A^{-\frac{1}{2}}) \geqslant \min_i p_i \dfrac{\lambda_{min}(\mathbf{C}^T A \mathbf{C})}{\lambda_{max}(A)} = \min_i p_i \dfrac{\lambda_{min}(A)\lambda_{min}(\mathbf{C}\mathbf{C}^T)}{\lambda_{max}(A)}$.

$\mathbf{C}\mathbf{C}^T = \sum_i I_{C_i}^T I_{C_i} = \binom{n-1}{s-1} I_n$ and then we obtain that :

$\lambda_{min}(A^{-\frac{1}{2}}E[Z]A^{-\frac{1}{2}}) \geqslant \binom{n-1}{s-1}\dfrac{\lambda_{min}(A)}{\lambda_{max}(A)}\min_i p_i$

1.2.2   Uniform case

For any i, $p_i = \dfrac{1}{\binom{n}{s}}$ is the uniform probability of choosing s rows uniformly on $\{1, \ldots, n\}$, knowing that s is the sketch size. That leads towards that corollary of **Proposition** 1.2.1 :

> **Corollary 1.2.2**
>
> $\lambda_{min}(A^{-\frac{1}{2}}E[Z]A^{-\frac{1}{2}}) \geqslant \dfrac{s}{n}\dfrac{\lambda_{min}(A)}{\lambda_{max}(A)}$

Robert: This is already pretty interesting! It shows an improvement for using bigger bachsize! We should try to push this further, for instance, when $s = n$ we know the method converges in one step. It would be great if we have a convergence rate that shows this phenomena. In other words, when $s = n$ we have $\lambda_{min}(A^{-1/2}E[Z]A^{-1/2}) = 1$! Also, please have a look at the paper "paving_kaczmarz.pdf" which I've just added to our repo.

1.2.3   A convenient probability

Suppose here that $p_i = \dfrac{Tr(I_{C_i}AI_{C_i}^T)}{\|A^{\frac{1}{2}}\mathbf{C}\|_F^2}$, for any $i = 1, \ldots, r$.

2. *Randomized orthonormal systems*

This type of randomized system is well-suited for big data regression, thanks to the efficiency of matrix multiplication used in this method.

When the dimension of our matrix A is n, we denote by H_n the Hadamard matrix (well defined if the dimension of the problem n is a power of 2) defined recursively as :

$H_p = ..$ and $H_1 = 1$.

The Hadamard sketch consists of choosing a sketch matrix $S \in \mathcal{M}_{s,n}$ where s is called the sketch size of the problem, as follows :

we sample s *i.i.d.* rows of the form $s^T = e_j^T H_n D$ with probability $\frac{1}{n}$ for $j = 1, \ldots, n$,where $(e_j)_j$ forms a canonical base of $\mathbb{R}^n$, and $D = diag(\nu)$ is a diagonal matrix of *i.i.d.* Rademacher variables $\nu \in \{-1, 1\}^n$.

2.1 Algorithm

2.2 Convergence rate (draft)

Now we denote by $Z = AS^T(SAS^T)^{-1}SA$, where S is our Hadamard random matrix.

The convergence rate is then $\rho = 1 - \lambda_{min}(A^{-\frac{1}{2}}E[Z]A^{-\frac{1}{2}})$

Notice that $S_i = I_{C_i}HD$. where $(C_i)_i$ are uniform random subsets of $\{1, \ldots, n\}$ of size s, as defined in the *Randomized Newton* section 1.

Let's condition on the Rademacher diagonal matrix D.

Define by $\tilde{A}_D = \frac{H}{\sqrt{n}}DAD\frac{H^T}{\sqrt{n}}$. We obtain that :

$$
\begin{aligned}
A^{-\frac{1}{2}}E[Z|D]A^{-\frac{1}{2}} &= E[A^{\frac{1}{2}}S^T(SAS^T)^{-1}SA^{\frac{1}{2}}|D] \\
&= \sum_i p_i A^{\frac{1}{2}}DH^T I_{C_i}^T (I_{C_i}HDADH^T I_{C_i}^T)^{-1}I_{C_i}HDA^{\frac{1}{2}} \\
&= A^{\frac{1}{2}}DH^T E[I_C^T(I_C\tilde{A}I_C^T)^{-1}I_C]HDA^{\frac{1}{2}} \\
&= nDH^{-1}\tilde{A}^{\frac{1}{2}}E[I_C^T(I_C\tilde{A}I_C^T)^{-1}I_C]\tilde{A}^{\frac{1}{2}}n(H^T)^{-1}D \\
&= DH^T\tilde{A}^{\frac{1}{2}}E[I_C^T(I_C\tilde{A}I_C^T)^{-1}I_C]\tilde{A}^{\frac{1}{2}}HD.
\end{aligned}
$$

(following to be changed)

Hence :

$$\rho = 1 - \lambda_{min}(A^{-\frac{1}{2}}E[Z]A^{-\frac{1}{2}}) = 1 - \lambda_{min}(\tilde{A}^{\frac{1}{2}}E[I_C^T(I_C\tilde{A}I_C^T)^{-1}I_C]\tilde{A}^{\frac{1}{2}})$$

We recognize the convergence rate in the Randomized Newton Method and then, denoting by $\rho_{Newton}(M)$ the convergence rate of the Newton method associated with the definite positive matrix M, we obtain that :

$\rho = 1 - \lambda_{min}(\tilde{A}^{\frac{1}{2}} E[I_C^T (I_C \tilde{A} I_C^T)^{-1} I_C] \tilde{A}^{\frac{1}{2}}) = 1 - (1 - \rho_{Newton}(\tilde{A})) = \rho_{Newton}(\tilde{A}) = \rho_{Newton}(A)$, since A and $\tilde{A}$ have the same eigenvalues.

Proposition 2.2.1

$$\lambda_{min}(A^{-\frac{1}{2}} E[Z] A^{-\frac{1}{2}}) \geqslant \frac{s}{n} \frac{\lambda_{min}(A)}{\lambda_{max}(A)}$$

3. *Count-min Sketches*

3.1 Algorithm

3.2 Convergence rate

S is constructed as follows :
For every $i \in \{1, \ldots, n\}$, l is chosen uniformly on $\{1, \ldots, n\}$ and ϵ uniformly on $\{-1, 1\}$, then S is updated in his l^{th} row as :
$S(l, :) := S(l, :) + \epsilon\, e_i^T$, where e_i^T is the ith coloumn of the identity matrix.

$$\mathbf{C} = (S_1^T, \ldots, S_r^T) \text{ and } \lambda_{max}(S_i^T S_i) = \lambda_{max}(S_i S_i^T).$$
$$S_i S_i^T = \sum_{i,k} f_{\pi(j)} e_j^T e_k f_{\pi(k)}^T.$$

4. *Conclusion*

References

[1] ROBERT GOWER AND PETER RICHTARIK, Randomized iterative methods for linear systems, SIAM, (2015).