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 $Z = AI_C^T(I_C A I_C^T)^{-1} I_C A$. 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
$D =$ 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 $i \in \{1,\dots,n\}$, 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 $\{1,\dots,n\}$, knowing that s is the sketch size. That leads towards that corollary of **Proposition** 3.2.1 :

Corollary 1.2.2

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

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 $i = 1,\dots,r$.

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 :

$$H_{2^p} = \begin{pmatrix} H_{2^{p-1}} & -H_{2^{p-1}} \\ H_{2^{p-1}} & H_{2^{p-1}} \end{pmatrix} \text{ for } p = 1, 2, \ldots \text{ and } H_1 = 1.$$

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

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

$S = I_C HD$ where C is a uniform random subset of $\{1, \ldots, n\}$ of size s, as defined in the $Randomized\ Newton$ **section** 1, H is the $Hadamard$ matrix ($HH^T = nI_n$) and D is a diagonal random matrix which values are uniformly distributed in $\{-1, 1\}$

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

Lemma 2.2.1

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

Proof :

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 :

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

Hence :

$$\lambda_{min}(A^{-\frac{1}{2}}E[Z]A^{-\frac{1}{2}}) = \lambda_{min}\left(E_D\left[\tilde{A}_D^{\frac{1}{2}}E[I_C^T(I_C\tilde{A}_DI_C^T)^{-1}I_C]\tilde{A}_D^{\frac{1}{2}}\right]\right).$$

Denote by $(D_i)_{i=1,\dots,2^n}$ the 2^n possible values of the random matrix D.
We obtain that :

$$\lambda_{min}(A^{-\frac{1}{2}}E[Z]A^{-\frac{1}{2}}) = \lambda_{min}\left(\sum_{i=1}^{2^n}\frac{1}{2^n}\tilde{A}_{D_i}^{\frac{1}{2}}E[I_C^T(I_C\tilde{A}_{D_i}I_C^T)^{-1}I_C]\tilde{A}_{D_i}^{\frac{1}{2}}\right).$$

And thanks to the concavity of λ_{min}, we obtain that :

$$\lambda_{min}(A^{-\frac{1}{2}}E[Z]A^{-\frac{1}{2}}) \geqslant \sum_{i=1}^{2^n}\frac{1}{2^n}\lambda_{min}\left(\tilde{A}_{D_i}^{\frac{1}{2}}E[I_C^T(I_C\tilde{A}_{D_i}I_C^T)^{-1}I_C]\tilde{A}_{D_i}^{\frac{1}{2}}\right).$$

We recognize least eigenvalues of Newton Sketches and then by **Corollary** 1.2.2, we obtain that :

$$\lambda_{min}(A^{-\frac{1}{2}}E[Z]A^{-\frac{1}{2}}) \geqslant \sum_{i=1}^{2^n}\frac{1}{2^n}\frac{s}{n}\frac{\lambda_{min}(\tilde{A}_{D_i})}{\lambda_{max}(\tilde{A}_{D_i})}.$$

Since for all $i = 1,\dots,2^n$, $\tilde{A}_{D_i}$ is similar to A, we obtain that :

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

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 is the i^{th} coloumn of the identity matrix.

Denote by $\{e_i\}_{i=1,\ldots,n}$ a canonical basis of $\mathbb{R}^n$ and $\{f_i\}_{i=1,\ldots,s}$ a canonical basis of $\mathbb{R}^s$.
Then we obtain that every count-min random matrix is of the form :
$S = \sum_{i=1}^{n} \epsilon(i) f_{\pi(i)} e_i^T \in \mathcal{M}_{s,n}(\mathbb{R})$, where $\epsilon : \{1, \ldots, n\} \to \{1, -1\}$ and $\pi : \{1, \ldots, n\} \to \{1, \ldots, s\}$.

We therefore can rewrite S as :

$$S = \Big(\epsilon(1) f_{\pi(1)}, \epsilon(2) f_{\pi(2)}, \ldots, \epsilon(n) f_{\pi(n)}\Big) \begin{pmatrix} e_1^T \\ \vdots \\ e_n^T \end{pmatrix} = \Big(f_{\pi(1)}, f_{\pi(2)}, \ldots, f_{\pi(n)}\Big)\, \mathrm{diag}\,(\epsilon(1), \ldots, \epsilon(n)).$$

For any $\pi : \{1, \ldots, n\} \to \{1, \ldots, s\}$, denote by I_π the $s \times n$ matrix which columns are $\Big\{f_{\pi(i)}\Big\}_{i=1,\ldots,n}$.

Let S be a random count-min sketch matrix.
$S = I_\pi D$ where π is a uniform random element of $\{1, \ldots, s\}^{\{1,\ldots,n\}}$ and D is a $n \times n$ diagonal random matrix which values are uniformly distributed in $\{-1, 1\}$

Denote again by $Z = AS^T(SAS^T)^{-1}SA$, where S is our count-min random matrix.
Recall that the convergence rate is $\rho = 1 - \lambda_{min}(A^{-\frac{1}{2}} E[Z] A^{-\frac{1}{2}})$.

Denote $r \overset{def}{=} s^n$ and $(\pi_1, \ldots, \pi_r)$ the different elements of $\{1, \ldots, s\}^{\{1,\ldots,n\}}$.
Denote by $\mathbf{C} = (I_{\pi_1}^T, \ldots, I_{\pi_r}^T)$ which is of size $n \times rs$.

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

Proposition 3.2.1

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

Proof :

Denote by $\tilde{A} = DAD$.

$$\begin{aligned} A^{-\frac{1}{2}}E[Z|D]A^{-\frac{1}{2}} &= E[A^{\frac{1}{2}}S^T(SAS^T)^{-1}SA^{\frac{1}{2}}|D] \\ &= \sum_i p_i A^{\frac{1}{2}}DI_{\pi_i}^T(I_{\pi_i}DADI_{\pi_i}^T)^{-1}I_{\pi_i}DA^{\frac{1}{2}} \\ &= A^{\frac{1}{2}}DE[I_\pi^T(I_\pi\tilde{A}_DI_\pi^T)^{-1}I_\pi]DA^{\frac{1}{2}} \end{aligned}$$

Then : $A^{-\frac{1}{2}}E[Z|D]A^{-\frac{1}{2}} = (A^{\frac{1}{2}}D\mathbf{C}\Delta)(\Delta\mathbf{C}^TDA^{\frac{1}{2}})$ where $\Delta = \text{diag}(\sqrt{p_1}(I_{\pi_1}\tilde{A}_DI_{\pi_1}^T)^{-\frac{1}{2}}, \dots, \sqrt{p_r}(I_{\pi_r}\tilde{A}_DI_{\pi_r}^T)^{-\frac{1}{2}}) \in \mathcal{M}_{rs}(\mathbb{R})$.

By concavity of λ_{min}, we obtain that :

$$\lambda_{min}(A^{-\frac{1}{2}}E[Z]A^{-\frac{1}{2}}) = \lambda_{min}\left(E_D\left[A^{-\frac{1}{2}}E[Z|D]A^{-\frac{1}{2}}\right]\right) \geqslant E_D\left(\lambda_{min}(A^{\frac{1}{2}}D\mathbf{C}\Delta)(\Delta\mathbf{C}^TDA^{\frac{1}{2}})\right).$$

Then :
$\lambda_{min}(A^{-\frac{1}{2}}E[Z]A^{-\frac{1}{2}}) \geqslant E_D\left(\lambda_{min}(\mathbf{C}^TDAD\mathbf{C}\Delta^2)\right) \geqslant E_D\left(\lambda_{min}(\mathbf{C}^TDAD\mathbf{C})\lambda_{min}(\Delta^2)\right).$

$\lambda_{min}(\Delta^2) = \min_i \dfrac{p_i}{\lambda_{max}(I_{\pi_i}\tilde{A}_DI_{\pi_i}^T)} \geqslant \min_i \dfrac{p_i}{\lambda_{max}(I_{\pi_i}^TI_{\pi_i})\lambda_{max}(\tilde{A}_D)} \geqslant \min_i \dfrac{p_i}{n\,\lambda_{max}(\tilde{A}_D)}$, since for any $i \in \{1, \dots, n\}$, for any x in $\mathbb{R}^n$

$\left\langle I_{\pi_i}^TI_{\pi_i}x \,|\, x\right\rangle = \|I_{\pi_i}x\|^2 = \|\sum_{i=1}^n x_if_{\pi(i)}\|^2 \leqslant \left(\sum_{i=1}^n |x_i|\right)^2 \leqslant n\|x\|^2$ and then $\lambda_{max}(I_{\pi_i}^TI_{\pi_i}) \leqslant n$.

Therefore, $\lambda_{min}(\mathbf{C}^TDAD\mathbf{C}\Delta^2) \geqslant \min_i p_i \dfrac{\lambda_{min}(\mathbf{C}^T\tilde{A}_D\mathbf{C})}{n\,\lambda_{max}(\tilde{A}_D)} = \min_i p_i \dfrac{\lambda_{min}(\tilde{A}_D)\lambda_{min}(\mathbf{C}\mathbf{C}^T)}{n\,\lambda_{max}(\tilde{A})}.$

$\mathbf{C}\mathbf{C}^T = \sum_i I_{\pi_i}^TI_{\pi_i} = ...I_n.$

Plus, $\tilde{A}_D$ is similar to A, thereby we obtain that :

$$\lambda_{min}(A^{-\frac{1}{2}}E[Z]A^{-\frac{1}{2}}) \geqslant E_D\left[...\frac{\lambda_{min}(A)}{n\,\lambda_{max}(A)}\min_i p_i\right] = \frac{\lambda_{min}(A)}{n\,\lambda_{max}(A)}\min_i p_i$$

4. *Conclusion*

References

[1] Robert Gower and Peter Richtarik, Randomized iterative methods for linear systems, SIAM, (2015).