Model Order Reduction Summer School 2019,
Eindhoven, Sep. 23-27, 2019

A brief introduction to Randomized Linear Algebra

Anthony Nouy

Centrale Nantes, Laboratoire de Mathématiques Jean Leray

Motivation

Linear algebra is at the core of of scientific computing, statistics, data analysis, artificial intelligence...

For large scale numerical or statistical models or problems involving big data sets, classical linear algebra methods require elementary algebraic operations on matrices and large vectors (norm, dot product, matrix-vector product, matrix-matrix product), the complexity of which being often prohibitive.

Motivation

Randomized linear algebra aims at

- reducing the complexity of algorithms,
- improving stability,
- taking into account computational constraints (data not available from RAM, streamed data),
- or fully exploiting modern computational architectures (parallel computing, cloud computing).

The idea is to project vectors or matrices onto low dimensional spaces and perform algebraic operations there.

Random embeddings allow to perform algebraic operations with guaranteed precision with high probability.

Estimation of the euclidian norm

A fundamental problem is the estimation of the euclidian norm $\|x\|_2$ of a vector $x \in \mathbb{R}^n$,

$$\|x\|_2 = \left(\sum_{j=1}^{n} x_j^2 \right)^{1/2},$$

by the euclidian norm

$$\|Sx\|_2$$

of a vecteur $Sx \in \mathbb{R}^k$ of dimension $k \ll n$.

Estimation of the euclidian norm

This will allow to understand other classical algebraic operations such as

- the inner product between two vectors x et y,

$$(x, y) = \frac{1}{4}\left(\|x+y\|_2^2 - \|x-y\|_2^2\right),$$

- the product of a matrix $A \in \mathbb{R}^{m\times n}$ by a vector $x \in \mathbb{R}^n$

$$(Ax)_i = (a_i, x)$$

  where $a_i \in \mathbb{R}^n$ is the i-th row of A,
- the product of two matrices $A \in \mathbb{R}^{m\times n}$ and $B \in \mathbb{R}^{n\times p}$

$$(AB)_{ij} = (a_i, b^j)$$

  where $b^j \in \mathbb{R}^n$ is the j-th column of B,
- norms of a matrix $A \in \mathbb{R}^{m\times n}$

$$\|A\|_F^2 = \sum_{i=1}^{m} \|a_i\|_2^2, \quad \|A\|_2 = \max_{\|v\|_2=1} \|Av\|_2 = \sigma_1(A), \quad \dots$$

but also more complex operations such as factorizations of matrices (singular value decomposition, QR factorization...).

Estimation of the euclidian norm

- We would like S to be a quasi-isometry from $\mathbb{R}^n$ to $\mathbb{R}^k$, i.e.

$$(1-\epsilon)\|x\|_2 \leq \|Sx\|_2 \leq (1+\epsilon)\|x\|_2$$

  for some $\epsilon > 0$.

- This can not be satisfied for all x unless $k \geq n$.

- But using for S a random matrix with a well chosen distribution, we can expect

$$\mathbb{P}((1-\epsilon)\|x\|_2 \leq \|Sx\|_2 \leq (1+\epsilon)\|x\|_2) \geq 1-\delta$$

  for all x and a high probability $1-\delta$.

Outline

Random embeddings

Consider a random matrix $S \in \mathbb{R}^{k\times n}$ of the form

$$S = \frac{1}{\sqrt{k}} B$$

where B is a random matrix.

For x in $\mathbb{R}^n$, letting $v = \frac{x}{\|x\|_2}$, we have

$$\frac{\|Sx\|_2^2}{\|x\|_2^2} = \frac{1}{k}\sum_{i=1}^{k}(\sum_{j=1}^{n} B_{ij}v_j)^2 := \frac{1}{k}\sum_{i=1}^{k} Y_i$$

where

$$Y_i = (\sum_{j=1}^{n} B_{ij}v_j)^2.$$

If the matrix B has i.i.d. rows such that $\mathbb{E}(B_{ij}B_{il}) = \delta_{jl}$, then the Y_i are i.i.d. and such that

$$\mathbb{E}(Y_i) = \sum_{j,l}\mathbb{E}(B_{ij}B_{il})v_jv_l = \sum_j v_j^2 = \|v\|_2^2 = 1$$

and $\frac{1}{k}\sum_{i=1} Y_i$ converges almost surely to 1.

Concentration inequality

Non asymptotic results can be obtained by analyzing how fast the distribution of $\frac{1}{k}\sum_{i=1}^{k} Y_i$ concentrates around $\mathbb{E}(Y_1)$.

Assuming the concentration inequality

$$\mathbb{P}(|\frac{1}{k}\sum_{i=1}^{k} Y_i - \mathbb{E}(Y_1)| > \epsilon) \leq \eta(k,\epsilon)$$

and denoting by

$$k(\epsilon,\delta) = \min\{k : \eta(k,\epsilon) \leq \delta\},$$

we guarantee

$$\mathbb{P}(|\frac{1}{k}\sum_{i=1}^{k} Y_i - \mathbb{E}(Y_1)| > \epsilon) = \mathbb{P}(|\frac{\|Sx\|_2^2}{\|x\|_2^2} - 1| > \epsilon) \leq \delta$$

under the condition

$$k \geq k(\epsilon,\delta).$$

If $\eta(k,\epsilon)$ decays sufficiently fast with k, we can satisfy the quasi-isometry property with a high probability $1-\delta$ for moderate k.

Random matrices with sub-gaussian entries

Consider B with i.i.d. entries with zero mean and variance 1, so that

$$\mathbb{E}(Y_i) = \mathbb{V}(\sum_{j=1}^{n} B_{ij} v_j) = \sum_{j=1}^{n} v_j^2 = 1.$$

If we further assume that B_{ij} follows a sub-gaussian distribution $SG(\gamma^2)$, then

$$\sum_{j=1}^{n} B_{ij} v_j \sim SG(\gamma^2 \sum_{j=1}^{n} v_j^2) = SG(\gamma^2)$$

and for $\epsilon \leq \gamma^2$,

$$\mathbb{P}(|\frac{1}{k} \sum_{i=1}^{k} Y_i - \mathbb{E}(Y_1)| > \epsilon) \leq 2e^{-\frac{k\epsilon^2}{8\gamma^4}}$$

which gives

$$k(\epsilon, \eta) = 8\gamma^4 \epsilon^{-2} \log(2\delta^{-1}),$$

a condition independent of n and logarithmic in δ^{-1}. This allows to attain a very small probability δ with a moderate k.

Note that the normal distribution $\mathcal{N}(0,1)$ is $SG(1)$.

Outline

Embeddings of subsets of vectors

We would like that the quasi-isometry property is satisfied simultaneously for all vectors in a subset Σ of $\mathbb{R}^n$, with high probability, i.e.

$$(1-\epsilon)\|x\|_2^2 \leq \|Sx\|_2^2 \leq (1+\epsilon)\|x\|_2^2 \quad \forall x \in \Sigma, \tag{1}$$

with high probability.

Assume that $S \in \mathbb{R}^{k\times n}$ is such that

$$k \geq k(\epsilon, \delta) \quad \text{implies} \quad \mathbb{P}(\left|\frac{\|Sx\|_2^2}{\|x\|_2} - 1\right| > \epsilon) \leq \delta \quad \forall x. \tag{2}$$

If Σ is a finite set and

$$k \geq k(\epsilon, \delta \#\Sigma^{-1}),$$

then

$$\mathbb{P}(\exists x \in \Sigma \; s.t. \; \left|\frac{\|Sx\|_2^2}{\|x\|_2} - 1\right| > \epsilon) \leq \delta$$

Random embedding of non finite subsets

The set

$$K = \{\frac{x}{\|x\|_2} : x \in \Sigma\}$$

being compact, it can be covered by a finite union of balls. Satisfying a quasi-isometry property for the centers of the balls is sufficient for obtaining a quasi-isometry property for all vectors in K.

Denoting by $\mathcal{N}_\epsilon(K)$ the covering number of K (the minimal number of balls of radius ϵ for covering K), we have that

$$\mathbb{P}((1-\epsilon)\|x\|_2^2 \leq \|Sx\|_2^2 \leq (1+\epsilon)\|x\|_2^2 \quad \forall x \in \Sigma) \leq \delta$$

under the condition

$$k \geq k(\frac{\epsilon}{4}, \delta \mathcal{N}_{\epsilon/4}(K)^{-1})$$

Assuming $k(\epsilon, \delta) = C\epsilon^{-2}\log(D\delta^{-1})$, then the condition is

$$k \geq 16C\epsilon^{-2}(\log(D\delta^{-1}) + \log \mathcal{N}_{\epsilon/4}(K)).$$

For V_m a subspace of dimension m, $K = \{x \in V_m : \|x\|_2 = 1\}$ is such that

$$\mathcal{N}_\epsilon(K) \leq (1 + \frac{2}{\epsilon})^m,$$

and the condition becomes

$$k \geq 16C\epsilon^{-2}(\log(D\delta^{-1}) + m\log(9\epsilon^{-1})).$$

This result can be improved by better exploiting the geometry of the unit sphere.

Outline

Random embeddings with good computational properties

In practice, it is interesting to construct random matrices S with good computational properties: reduced storage, efficient matrix-vector multiplication...

Random matrices with discrete distributions

Consider

$$S = \frac{1}{\sqrt{k}} B$$

- Choosing a matrix B whose entries are i.i.d. Rademacher random variables,

$$B_{ij} = \begin{cases} 1 & \text{with probability } 1/2 \\ -1 & \text{with probability } 1/2 \end{cases}$$

  yields a reduced storage and applying B to a vector only requires changes of signs and additions.

- Choosing a matrix B whose entries are i.i.d. and such that

$$B_{ij} = \begin{cases} \sqrt{r} & \text{with probability } \frac{1}{2r} \\ 0 & \text{with probability } 1 - \frac{1}{r} \\ -\sqrt{r} & \text{with probability } \frac{1}{2r} \end{cases} \quad (3)$$

  yields a sparse matrix whose average sparsity ratio $\frac{1}{r}$. The r.v. B_{ij} defined by (3) has zero mean, unit variance and $B_{ij} \in SG(r)$.

Subsampling

An sparser matrix can be defined by

$$S_{ij} = \sqrt{\frac{n}{k}}\delta_{J_i,j} \tag{4}$$

where the J_i are i.i.d. and uniformly distributed over $\{1, \ldots, n\}$.

The entries of S are not independent but the rows are independent. The row i contains $\pm\sqrt{\frac{n}{k}}$ in the column J_i drawn randomly and 0 in the other columns.

$$S = \sqrt{\frac{n}{k}} \begin{pmatrix} & 1 & & & \\ & & & & 1 \\ -1 & & & & \\ & & & -1 & \\ & & 1 & & \end{pmatrix}$$

Then

$$\frac{\|Sx\|_2^2}{\|x\|_2^2} = \frac{1}{k}\sum_{i=1}^{k} Y_i$$

where the Y_i are independent and

$$Y_i = n v_{J_i}^2,$$

such that $\mathbb{E}(Y_i) = 1$.

Subsampling

Since Y_i is bounded by $n\|v\|_\infty^2$, we obtain from Bernstein inequality that

$$\mathbb{P}(|\frac{1}{k}\sum_{i=1}^{k} Y_i - 1| > \epsilon) \leq 2e^{-\frac{k\epsilon^2}{2(1+\epsilon)(n\|v\|_\infty^2 - 1)}}$$

so that

$$\mathbb{P}(|\frac{\|Sx\|_2^2}{\|x\|_2^2} - 1| > \epsilon) \leq \delta$$

provided

$$k \geq k(\epsilon, \delta) = 2(1+\epsilon)\log(2\delta^{-1})\epsilon^{-2}(n\|v\|_\infty^2 - 1) \quad (5)$$

- For all v, $\|v\|_\infty \leq \|v\|_2 = 1$ but the condition (5) makes the sampling approach useless compared to a classical approach.
- For (homogeneous) vectors with components of equal magnitude, $|v_j| = \frac{1}{\sqrt{n}}$ for all j and $Y_i = 1$, so that $\|Sx\|_2^2 = \|x\|_2^2$ almost surely.
- For vectors such that $\sqrt{n}\|v\|_\infty \leq \beta$ (with $\beta \geq 1$), a sufficient condition is

  $$k \geq k(\epsilon, \delta) \leq 2(1+\epsilon)\log(2\delta^{-1})\epsilon^{-2}\beta^2,$$

  which is independent of n.

Subsampling

For sparse vectors (or with very heterogeneous components), subsampling approach has bad performances.

However, uncertainty principle states that a vector x and its discrete Fourier transform Fx can not be sparse simultaneously. Then we can expect using subsampling on Fx. This is also true for the Hadamard transform.

This yields embeddings called Subsampled Randomized Fourier Transform (SRFT) or Subsampled Randomized Hadamard Transform (SRHT).

Subsampled Randomized Hadamard Transform

Definition (Hadamard matrix)

Let $n = 2^d$. The Hadamard matrix H_d is defined recursively by

$$H_1 = \frac{1}{\sqrt{2}} \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix} \quad \text{and} \quad H_d = H_{d-1} \otimes H_1 = \frac{1}{\sqrt{2}} \begin{pmatrix} H_{d-1} & H_{d-1} \\ H_{d-1} & -H_{d-1} \end{pmatrix}.$$

The component (i, j) of H_d is

$$(H_d)_{ij} = \frac{1}{2^{d/2}} (-1)^{\sum_{l=0}^{d-1} i_l j_l}$$

where $i = \sum_{l=0}^{d-1} i_l 2^l$ et $j = \sum_{l=0}^{d-1} j_l 2^l$.

The complexity of applying H_d to a vector is in $O(dn) = O(n \log(n))$.

The Hadamard matrix H_d is symmetric, orthogonal and defines an isometry from $\mathbb{R}^{2^d}$ to $\mathbb{R}^{2^d}$. Its components verify $|(H_d)_{ij}| \leq \frac{1}{\sqrt{n}}$ with $n = 2^d$.

Subsampled Randomized Hadamard Transform

Definition (SRHT)

For $n = 2^d$, the SRHT is defined by

$$S = PHD$$

where

- $D \in \mathbb{R}^{n \times n}$ is a diagonal matrix whose diagonal entries are i.i.d. and uniform on $\{-1, 1\}$,
- $H \in \mathbb{R}^{n \times n}$ is a Hadamard matrix,
- $P \in \mathbb{R}^{k \times n}$ is a random matrix implementing subsampling (i.e., $P_{ij} = \sqrt{\frac{n}{k}} \delta_{J_i, j}$ where the J_i are i.i.d. and uniform on $\{1, \ldots, n\}$)

The complexity of applying H is in $O(n \log(n))$, of applying D is $O(n)$ and applying P is $O(k)$. The complexity of applying S is then $O(n \log(n))$.

Matrices alÃ©atoires structurÃ©es: SRHT

For v a vector with norm $\|v\|_2 = 1$,

$$\mathbb{P}(n\|HDv\|_\infty^2 - 1 > t) \leq ne^{-t^2/8}. \tag{6}$$

From results on subsampling, we then deduce that the quasi-isometry property is satisfied with probability higher than $1 - \delta$ if

$$k \geq k(\epsilon, \delta) = 4\sqrt{2}(1+\epsilon)\epsilon^{-2} \log(2n\delta^{-1})^{1/2} \log(4\delta^{-1}),$$

References I

David P Woodruff.
Sketching as a tool for numerical linear algebra.
Foundations and Trends® in Theoretical Computer Science, 10(1–2):1–157, 2014.

Michael Mahoney.
Randomized algorithms for matrices and data.
Foundations and Trends® in Machine Learning, 3(2):123–224, 2011.

Roman Vershynin.
Introduction to the non-asymptotic analysis of random matrices.
arXiv preprint arXiv:1011.3027, 2010.

Ravindran Kannan and Santosh Vempala.
Randomized algorithms in numerical linear algebra.
Acta Numerica, 26:95–135, 2017.

References II

O. Balabanov and A. Nouy.
Randomized linear algebra for model reduction. Part I: Galerkin methods and error estimation.
ArXiv e-prints, March 2018.

O. Balabanov and A. Nouy.
Randomized linear algebra for model reduction. Part II: minimal residual methods and dictionary-based approximation.
Available soon.