Lecture 4: Randomized linear dimension reduction

School of Mathematical Sciences, Xiamen University

1. Subspace embedding

Definition 1 (Subspace embedding)

Let $\mathcal{L} \subseteq \mathbb{R}^n$ be a linear subspace with dimension d. Consider a linear map $\mathbf{\Phi} : \mathbb{R}^n \mapsto \mathbb{R}^s$ with the property that

$$(1-\varepsilon)\|\mathbf{x}\|_2 \leq \|\mathbf{\Phi x}\|_2 \leq (1+\varepsilon)\|\mathbf{x}\|_2 \quad \text{for all} \quad \mathbf{x} \in \mathcal{L}.$$

The map $\mathbf{\Phi}$ is called a subspace embedding for $\mathcal{L}$ with embedding dimension $s \in \mathbb{N}$ and distortion $\varepsilon > 0$.

Exercise: Prove that $s \geq d$.

- By the linearity of $\mathbf{\Phi}$, for all $\mathbf{x}, \mathbf{y} \in \mathcal{L}$, it holds that

$$(1-\varepsilon)\|\mathbf{x}-\mathbf{y}\|_2 \leq \|\mathbf{\Phi x} - \mathbf{\Phi y}\|_2 \leq (1+\varepsilon)\|\mathbf{x}-\mathbf{y}\|_2.$$

- In real applications, the embedding dimension s is close to the subspace dimension d and much smaller than the ambient dimension n: $s \approx d \ll n$.

Proposition 2

Suppose that $\mathrm{range}(\mathbf{U}) = \mathcal{L}$ *where* $\mathbf{U} \in \mathbb{R}^{n\times d}$ *is a matrix with orthonormal columns. The subspace embedding property*

$$(1-\varepsilon)\|\mathbf{x}\|_2 \le \|\mathbf{\Phi x}\|_2 \le (1+\varepsilon)\|\mathbf{x}\|_2 \quad \textit{for all} \quad \mathbf{x} \in \mathcal{L}$$

is equivalent with the condition

$$1-\varepsilon \le \sigma_{\min}(\mathbf{\Phi U}) \le \sigma_{\max}(\mathbf{\Phi U}) \le 1+\varepsilon.$$

Proof. From $\mathcal{L} = \{\mathbf{Uy} : \mathbf{y} \in \mathbb{R}^d\}$, we have

$$(1-\varepsilon)\|\mathbf{Uy}\|_2 \le \|\mathbf{\Phi Uy}\|_2 \le (1+\varepsilon)\|\mathbf{Uy}\|_2 \quad \text{for all} \quad \mathbf{y} \in \mathbb{R}^d.$$

By $\|\mathbf{Uy}\|_2 = \|\mathbf{y}\|_2$, we have

$$1-\varepsilon \le \|\mathbf{\Phi Uz}\|_2 \le 1+\varepsilon \quad \text{for each unit vector} \quad \mathbf{z} \in \mathbb{R}^d.$$

The variational definition of $\sigma_{\min}$ and $\sigma_{\max}$ completes the proof. $\square$.

2. Random subspace embeddings

- In many applications, it is imperative to construct a subspace embedding $\mathbf{\Phi} : \mathbb{R}^n \mapsto \mathbb{R}^s$ without using prior knowledge about the subspace $\mathcal{L} \subseteq \mathbb{R}^n$. These are called *oblivious* subspace embeddings.
- By drawing a subspace embedding at random, we can ensure that the embedding property holds with high probability.

2.1 Subsampled randomized trigonometric transform (SRTT)

- Subsampled randomized trigonometric transform:

$$\mathbf{\Phi} := \sqrt{\frac{n}{s}}\mathbf{RDF} \in \mathbb{R}^{s \times n}$$

where $\mathbf{R} \in \mathbb{R}^{s \times n}$ subsamples rows, $\mathbf{D} \in \mathbb{R}^{n \times n}$ is random diagonal, and $\mathbf{F} \in \mathbb{R}^{n \times n}$ is a DCT2 matrix. More precisely, $\mathbf{R}$ is a uniformly random set of s rows drawn from the identity matrix $\mathbf{I}_n$, and the random diagonal matrix $\mathbf{D}$ has i.i.d. uniform$\{\pm 1\}$ entries.

Exercise: Prove that $\mathbb{E}\|\mathbf{\Phi x}\|_2^2 = \|\mathbf{x}\|_2^2$.

- The cost of applying the SRTT to a vector is $\mathcal{O}(n \log n)$ operations using a standard fast DCT2 algorithm, and it can be reduced to $\mathcal{O}(n \log s)$ with a more careful implementation.
- What embedding dimension s does the SRTT require?
  In practice, $s \approx d/\varepsilon^2$ usually has 'satisfying' performance.

2.2 Sparse random matrices

- Consider a sparse random matrix of the form

$$\mathbf{\Phi} = \begin{bmatrix} \boldsymbol{\varphi}_1 & \cdots & \boldsymbol{\varphi}_n \end{bmatrix} \in \mathbb{R}^{s \times n},$$

  where $\boldsymbol{\varphi}_i \in \mathbb{R}^s$ are i.i.d. sparse vectors. More precisely, each column $\boldsymbol{\varphi}_i$ contains exactly ζ nonzero entries, equally likely to be $\pm 1/\sqrt{\zeta}$, in uniformly positions. Exercise: $\mathbb{E}\|\mathbf{\Phi x}\|_2^2 = \|\mathbf{x}\|_2^2$.
- We can apply this matrix to a vector in $\mathcal{O}(\zeta n)$ operations. The storage cost is at most ζn parameters. If $\zeta \ll s$, then we obtain a significant computational benefit.
- The existing theoretical results are not sufficiently precise that we can use them to set algorithm parameters a priori.

3. Approximate least-squares

- Consider the quadratic optimization problem

$$\min_{\mathbf{x}\in\mathbb{R}^d} \frac{1}{2}\|\mathbf{A}\mathbf{x}-\mathbf{b}\|_2^2 \quad \text{with} \quad \mathbf{A}\in\mathbb{R}^{n\times d},\ \mathbf{b}\in\mathbb{R}^n.$$

  We focus on the case where $d \ll n$ and $\mathbf{A}$ is dense and unstructured.
- The cost of solving the problem with a direct method, such as QR factorization, is $\mathcal{O}(d^2 n)$ operations.
- The *sketch-and-solve* approach can obtain a coarse solution to the least-squares problem efficiently ($\mathcal{O}(nd\log d + d^3/\varepsilon^2)$).

  (1) Construct a (random, fast) subspace embedding $\mathbf{\Phi}\in\mathbb{R}^{s\times n}$ for range($\begin{bmatrix}\mathbf{A} & \mathbf{b}\end{bmatrix}$).

  (2) Reduce the dimension of the problem data: $\mathbf{\Phi}\mathbf{A}\in\mathbb{R}^{s\times d}$ and $\mathbf{\Phi}\mathbf{b}\in\mathbb{R}^s$. This step is commonly referred to as *sketching*.

  (3) Find a solution $\mathbf{x}_{\text{sk}}\in\mathbb{R}^d$ to the sketched least-squares problem:

$$\min_{\mathbf{x}\in\mathbb{R}^d} \frac{1}{2}\|\mathbf{\Phi}(\mathbf{A}\mathbf{x}-\mathbf{b})\|_2^2.$$

Proposition 3

Suppose that $\mathbf{A} \in \mathbb{R}^{n\times d}$ *is a tall matrix and* $\mathbf{b} \in \mathbb{R}^{n}$. *Construct a subspace embedding* $\mathbf{\Phi} \in \mathbb{R}^{s\times n}$ *for* $\operatorname{range}(\begin{bmatrix}\mathbf{A} & \mathbf{b}\end{bmatrix})$ *with distortion* ε. *Let* $\mathbf{x}_\star \in \mathbb{R}^d$ *be a solution to the original least-squares problem, and let* $\mathbf{x}_{\text{sk}} \in \mathbb{R}^d$ *be a solution to the sketched problem. Then*

$$\|\mathbf{A}\mathbf{x}_{\text{sk}} - \mathbf{b}\|_2 \leq \frac{1+\varepsilon}{1-\varepsilon}\|\mathbf{A}\mathbf{x}_\star - \mathbf{b}\|_2.$$

Proof. Using the embedding property twice yields

$$\begin{aligned}\|\mathbf{A}\mathbf{x}_{\text{sk}} - \mathbf{b}\|_2 &\leq \frac{1}{1-\varepsilon}\|\mathbf{\Phi}(\mathbf{A}\mathbf{x}_{\text{sk}} - \mathbf{b})\|_2 \\ &\leq \frac{1}{1-\varepsilon}\|\mathbf{\Phi}(\mathbf{A}\mathbf{x}_\star - \mathbf{b})\|_2 \leq \frac{1+\varepsilon}{1-\varepsilon}\|\mathbf{A}\mathbf{x}_\star - \mathbf{b}\|_2.\end{aligned}$$

The first (third) inequality is the lower (upper) bound in the embedding property. The second inequality holds because $\mathbf{x}_{\text{sk}}$ is the optimal solution to the sketched least-squares problem. □

4. Approximate orthogonalization

- Problem: Consider a matrix $\mathbf{A} \in \mathbb{R}^{n\times d}$ with full column rank. The task is to find a well-conditioned matrix $\mathbf{B} \in \mathbb{R}^{n\times d}$ with $\mathrm{range}(\mathbf{B}) = \mathrm{range}(\mathbf{A})$.
- A direct method for orthogonalizing the columns of the matrix $\mathbf{A}$ requires $\mathcal{O}(nd^2)$ arithmetic.
- Randomized Gram–Schmidt:

  (1) Construct a (random, fast) subspace embedding $\mathbf{\Phi} \in \mathbb{R}^{s\times n}$ for $\mathrm{range}(\mathbf{A})$. $\quad s = \mathcal{O}(d/\varepsilon^2)$

  (2) Sketch the problem data: $\mathbf{\Phi A} \in \mathbb{R}^{s\times d}$. $\quad \mathcal{O}(nd \log d)$

  (3) Compute a (thin, pivoted) QR factorization of the sketched data: $\mathbf{\Phi A} = \mathbf{QR}$. $\quad \mathcal{O}(d^3/\varepsilon^2)$

  (4) (Implicitly) define well-conditioned $\mathbf{B} = \mathbf{A}\mathbf{R}^{-1}$ with $\mathrm{range}(\mathbf{B}) = \mathrm{range}(\mathbf{A})$. If we wish to form the matrix $\mathbf{B}$ explicitly, we must spend $\mathcal{O}(nd^2)$ operations.

Proposition 4

Let $\mathbf{A} \in \mathbb{R}^{n\times d}$ *be a tall matrix with full column rank. Construct a subspace embedding* $\mathbf{\Phi} \in \mathbb{R}^{s\times d}$ *for* $\mathrm{range}(\mathbf{A})$ *with distortion* ε*. Form a QR factorization of the sketched matrix:* $\mathbf{\Phi A} = \mathbf{QR}$ *with* $\mathbf{R} \in \mathbb{R}^{d\times d}$*. Then* $\mathbf{R}$ *has full rank, and the whitened matrix* $\mathbf{B} = \mathbf{AR}^{-1}$ *satisfies*

$$\frac{1}{1+\varepsilon} \le \sigma_{\min}(\mathbf{B}) \le \sigma_{\max}(\mathbf{B}) \le \frac{1}{1-\varepsilon}.$$

Proof. Since $\mathbf{\Phi}$ is a subspace embedding for the d-dimensional subspace $\mathrm{range}(\mathbf{A})$, the range of the sketched matrix $\mathbf{\Phi A}$ also has dimension d. Thus, $\mathbf{R}$ must have full rank. From $\|\mathbf{Ry}\|_2 = \|\mathbf{\Phi Ay}\|_2$, $\mathbf{y} = \mathbf{R}^{-1}\mathbf{x}$, and

$$(1-\varepsilon)\|\mathbf{Ay}\|_2 \le \|\mathbf{\Phi Ay}\|_2 \le (1+\varepsilon)\|\mathbf{Ay}\|_2,$$

we have

$$(1-\varepsilon)\|\mathbf{AR}^{-1}\mathbf{x}\|_2 \le \|\mathbf{x}\|_2 \le (1+\varepsilon)\|\mathbf{AR}^{-1}\mathbf{x}\|_2.$$

The variational definition of $\sigma_{\min}$ and $\sigma_{\max}$ completes the proof. □

5. Approximate null space

- Problem: Consider a tall matrix $\mathbf{A} \in \mathbb{R}^{n\times d}$. The task is to find an orthonormal matrix $\mathbf{W} \in \mathbb{R}^{d\times k}$ whose range aligns with the k trailing right singular vectors of $\mathbf{A}$.
- A full SVD of the input matrix $\mathbf{A}$ requires $\mathcal{O}(nd^2)$ arithmetic.
- The *sketch-and-solve* approach: $\mathcal{O}(nd\log d + d^3/\varepsilon^2)$

  (1) Construct a (random, fast) subspace embedding $\mathbf{\Phi} \in \mathbb{R}^{s\times n}$ for range($\mathbf{A}$). $s = \mathcal{O}(d/\varepsilon^2)$

  (2) Sketch the problem data: $\mathbf{\Phi A} \in \mathbb{R}^{s\times d}$. $\mathcal{O}(nd\log d)$

  (3) Compute SVD of the sketched matrix: $\mathbf{\Phi A} = \mathbf{U\Sigma V}^*$. $\mathcal{O}(sd^2)$

  (4) Set $\mathbf{W} = \mathbf{V}(:, (d-k+1):d) \in \mathbb{R}^{d\times k}$.
- A variational formulation of the null space problem:

$$\min_{\mathbf{V}\in\mathbb{R}^{d\times k}} \|\mathbf{AV}\|_{\mathrm{F}}^2 \quad \text{subject to} \quad \mathbf{V}^*\mathbf{V} = \mathbf{I}_k.$$

  The solution is a matrix of k trailing right singular vectors.

Proposition 5

Let $\mathbf{A} \in \mathbb{R}^{n\times d}$ *be a tall matrix, and let* $\mathbf{\Phi} \in \mathbb{R}^{s\times d}$ *be a subspace embedding for* $\operatorname{range}(\mathbf{A})$ *with distortion* ε. *The* $\mathbf{W}$ *generated by the sketch-and solve approach satisfies*

$$\|\mathbf{AW}\|_{\mathrm{F}}^2 \leq \frac{(1+\varepsilon)^2}{(1-\varepsilon)^2} \min_{\mathbf{V}\in\mathbb{R}^{d\times k}, \mathbf{V}^*\mathbf{V}=\mathbf{I}_k} \|\mathbf{AV}\|_{\mathrm{F}}^2.$$

In particular, if $\mathbf{AV} = \mathbf{0}$ *for some* k*-dimensional subspace* $\mathbf{V}$, *then* $\mathbf{AW} = \mathbf{0}$.

Proof. Fix an orthonormal matrix $\mathbf{V}_\star \in \mathbb{R}^{d\times k}$ that solves the null space problem. Since v is a subspace embedding for $\operatorname{range}(\mathbf{A})$,

$$\|\mathbf{AW}\|_{\mathrm{F}}^2 \leq \frac{1}{(1-\varepsilon)^2}\|\mathbf{\Phi AW}\|_{\mathrm{F}}^2 \leq \frac{1}{(1-\varepsilon)^2}\|\mathbf{\Phi AV}_\star\|_{\mathrm{F}}^2 \leq \frac{(1+\varepsilon)^2}{(1-\varepsilon)^2}\|\mathbf{AV}_\star\|_{\mathrm{F}}^2.$$

The first (third) inequality is the lower (upper) bound in the embedding property. The second inequality holds because $\mathbf{W}$ is the optimal solution to the sketched problem. $\square$

Proposition 6

Let $\mathbf{A} \in \mathbb{R}^{n\times d}$ *be a tall matrix, and let* $\mathbf{\Phi} \in \mathbb{R}^{s\times d}$ *be a subspace embedding for* $\operatorname{range}(\mathbf{A})$ *with distortion* ε. *The singular values of the sketched matrix* $\mathbf{\Phi A}$ *satisfy*

$$(1-\varepsilon)\sigma_i(\mathbf{A}) \le \sigma_i(\mathbf{\Phi A}) \le (1+\varepsilon)\sigma_i(\mathbf{A}) \quad \textit{for} \quad i = 1, \ldots, d.$$

Proof. Let $\mathbf{A} = \mathbf{U\Sigma V}^*$ be an SVD. Then $\mathbf{\Phi}$ is a subspace embedding for $\operatorname{range}(\mathbf{U})$. For each index $i = 1, \ldots, d$, by the rotational invariance of singular values,

$$\sigma_i(\mathbf{\Phi A}) = \sigma_i(\mathbf{\Phi}(\mathbf{U\Sigma V}^*)) = \sigma_i(\mathbf{\Phi U \Sigma}).$$

By Ostrowski's relative perturbation theorem, we have

$$\sigma_d(\mathbf{\Phi U})\sigma_i(\mathbf{\Sigma}) \le \sigma_i(\mathbf{\Phi A}) \le \sigma_1(\mathbf{\Phi U})\sigma_i(\mathbf{\Sigma}).$$

By the subspace embedding property, we have

$$(1-\varepsilon)\sigma_i(\mathbf{A}) \le \sigma_i(\mathbf{\Phi A}) \le (1+\varepsilon)\sigma_i(\mathbf{A}). \quad \square$$