

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 . Given $0 < \varepsilon < 1$, we consider a linear map $\Phi : \mathbb{R}^n \mapsto \mathbb{R}^s$ with the property that

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

The map Φ is called a subspace embedding for \mathcal{L} with embedding dimension s and distortion ε .

Exercise: Prove that $s \geq d$.

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

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

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

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

is equivalent to the condition

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

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

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

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

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

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

2. Random subspace embeddings

- In many applications, it is imperative to construct a subspace embedding $\Phi : \mathbb{R}^n \mapsto \mathbb{R}^s$ 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:

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

where $\mathbf{R} \in \mathbb{R}^{s \times n}$ subsamples rows, $\mathbf{F} \in \mathbb{R}^{n \times n}$ is a DCT2 matrix, and $\mathbf{D} \in \mathbb{R}^{n \times n}$ is random diagonal. 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. $\text{uniform}\{\pm 1\}$ entries.

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

- 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

$$\Phi = [\varphi_1 \quad \cdots \quad \varphi_n] \in \mathbb{R}^{s \times n},$$

where $\varphi_i \in \mathbb{R}^s$ are i.i.d. sparse vectors. More precisely, each column φ_i contains exactly $\zeta < s$ nonzero entries, equally likely to be $\pm 1/\sqrt{\zeta}$, in uniformly positions. **Exercise:** $\mathbb{E}\|\Phi \mathbf{x}\|_2^2 = \|\mathbf{x}\|_2^2$.

- 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.

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 $\mathcal{O}(d^2n)$ operations.
- The *sketch-and-solve* approach can obtain a coarse solution to the least-squares problem efficiently ($\mathcal{O}(nd \log d + d^3/\varepsilon^2)$).
 - Construct a (random, fast) subspace embedding $\Phi \in \mathbb{R}^{s \times n}$ for $\text{range}([\mathbf{A} \ \mathbf{b}])$.
 - Reduce the dimension of the problem data: $\Phi\mathbf{A} \in \mathbb{R}^{s \times d}$ and $\Phi\mathbf{b} \in \mathbb{R}^s$. This step is commonly referred to as *sketching*.
 - Find a solution $\mathbf{x}_{\text{sk}} \in \mathbb{R}^d$ to the sketched least-squares problem:

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

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 $\Phi \in \mathbb{R}^{s \times n}$ for $\text{range}([\mathbf{A} \ \mathbf{b}])$ 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

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

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

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 $\text{range}(\mathbf{B}) = \text{range}(\mathbf{A})$.
- 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 $\Phi \in \mathbb{R}^{s \times n}$ for $\text{range}(\mathbf{A})$. $s = \mathcal{O}(d/\varepsilon^2)$
 - (2) Sketch the problem data: $\Phi \mathbf{A} \in \mathbb{R}^{s \times d}$. $\mathcal{O}(nd \log d)$
 - (3) Compute a (thin, pivoted) QR factorization of the sketched data: $\Phi \mathbf{A} = \mathbf{Q}\mathbf{R}$. $\mathcal{O}(d^3/\varepsilon^2)$
 - (4) (Implicitly) define well-conditioned $\mathbf{B} = \mathbf{A}\mathbf{R}^{-1}$ with $\text{range}(\mathbf{B}) = \text{range}(\mathbf{A})$. 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 $\Phi \in \mathbb{R}^{s \times d}$ for $\text{range}(\mathbf{A})$ with distortion ε . Form a QR factorization of the sketched matrix: $\Phi\mathbf{A} = \mathbf{Q}\mathbf{R}$ with $\mathbf{R} \in \mathbb{R}^{d \times d}$. Then \mathbf{R} has full rank, and the whitened matrix $\mathbf{B} = \mathbf{A}\mathbf{R}^{-1}$ satisfies

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

Proof. Since Φ is a subspace embedding for the d -dimensional subspace $\text{range}(\mathbf{A})$, the range of the sketched matrix $\Phi\mathbf{A}$ also has dimension d . Thus, \mathbf{R} must have full rank. For any $\mathbf{x} \in \mathbb{R}^d$, let $\mathbf{y} = \mathbf{R}^{-1}\mathbf{x}$. From

$$\|\mathbf{R}\mathbf{y}\|_2 = \|\Phi\mathbf{A}\mathbf{y}\|_2 \quad \text{and} \quad (1 - \varepsilon)\|\mathbf{A}\mathbf{y}\|_2 \leq \|\Phi\mathbf{A}\mathbf{y}\|_2 \leq (1 + \varepsilon)\|\mathbf{A}\mathbf{y}\|_2$$

we have

$$(1 - \varepsilon)\|\mathbf{A}\mathbf{R}^{-1}\mathbf{x}\|_2 \leq \|\mathbf{x}\|_2 \leq (1 + \varepsilon)\|\mathbf{A}\mathbf{R}^{-1}\mathbf{x}\|_2.$$

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

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 variational formulation of the problem:

$$\min_{\mathbf{X} \in \mathbb{R}^{d \times k}} \|\mathbf{A}\mathbf{X}\|_{\text{F}}^2 \quad \text{subject to} \quad \mathbf{X}^\top \mathbf{X} = \mathbf{I}_k.$$

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

- 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 $\Phi \in \mathbb{R}^{s \times n}$ for $\text{range}(\mathbf{A})$. $s = \mathcal{O}(d/\varepsilon^2)$
 - (2) Sketch the problem data: $\Phi \mathbf{A} \in \mathbb{R}^{s \times d}$. $\mathcal{O}(nd \log d)$
 - (3) Compute SVD of the sketched matrix: $\Phi \mathbf{A} = \mathbf{U} \Sigma \mathbf{V}^\top$. $\mathcal{O}(sd^2)$
 - (4) Set $\mathbf{W} = \mathbf{V}(:, (d - k + 1) : d) \in \mathbb{R}^{d \times k}$.

Proposition 5

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

$$\|\mathbf{AW}\|_{\text{F}}^2 \leq \frac{(1 + \varepsilon)^2}{(1 - \varepsilon)^2} \min_{\mathbf{X} \in \mathbb{R}^{d \times k}, \mathbf{X}^\top \mathbf{X} = \mathbf{I}_k} \|\mathbf{AX}\|_{\text{F}}^2.$$

In particular, if $\mathbf{AX} = \mathbf{0}$ for some k -dimensional subspace \mathcal{X} with $\mathcal{X} = \text{range}(\mathbf{X})$, then $\mathbf{AW} = \mathbf{0}$.

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

$$\|\mathbf{AW}\|_{\text{F}}^2 \leq \frac{1}{(1 - \varepsilon)^2} \|\Phi \mathbf{AW}\|_{\text{F}}^2 \leq \frac{1}{(1 - \varepsilon)^2} \|\Phi \mathbf{AX}_*\|_{\text{F}}^2 \leq \frac{(1 + \varepsilon)^2}{(1 - \varepsilon)^2} \|\mathbf{AX}_*\|_{\text{F}}^2.$$

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. □

Proposition 6

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

$$(1 - \varepsilon)\sigma_i(\mathbf{A}) \leq \sigma_i(\Phi \mathbf{A}) \leq (1 + \varepsilon)\sigma_i(\mathbf{A}) \quad \text{for } i = 1, \dots, d.$$

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

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

By Ostrowski's theorem for singular values, we have

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

By the subspace embedding property, we have

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