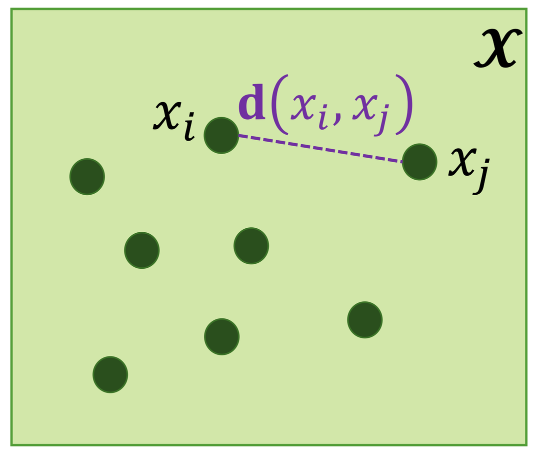
Sample-Optimal Low-Rank Approximation of Distance Matrices

Piotr Indyk — MIT
Ali Vakilian — MIT
Tal Wagner — MIT
David Woodruff — CMU

Distance Matrices

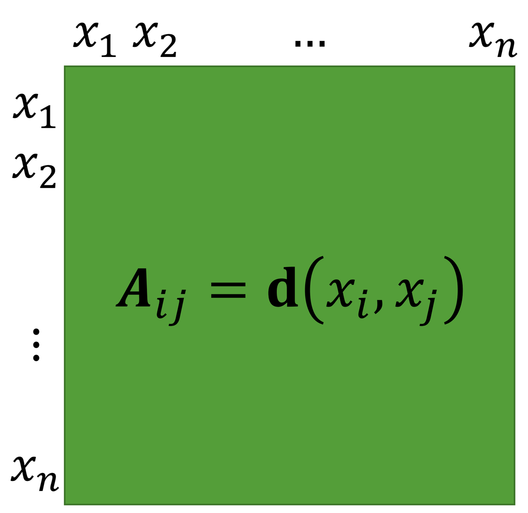
Let $(\mathcal{X}, \mathbf{d})$ be a metric space

- $\mathcal{X} = \{x_1, \ldots, x_n\}$
- $\mathbf{d}: \mathcal{X} \times \mathcal{X} \to \mathbb{R}_{\geq 0}$
- Symmetric
- Triangle inequality



Distance matrix:

$$A_{ij} = \mathbf{d}(x_i, x_j)$$



Many applications:

E.g. image learning, image understanding, protein structure analysis, **see survey:** [DPRV15]

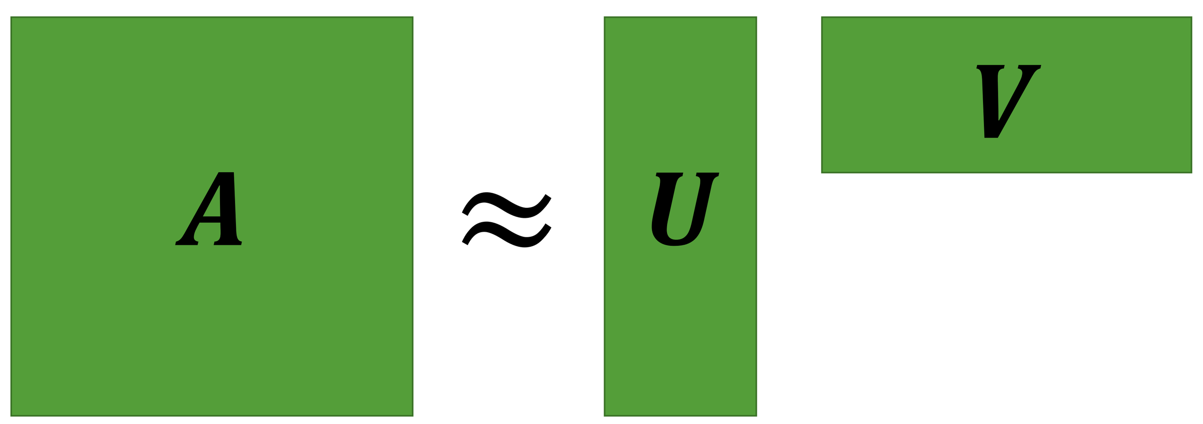
Low-Rank Approximation

Input: $A \in \mathbb{R}^{n \times n}$, integer $0 < k \ll n$

Output: Rank-k approximation of A:

$$U, V^T \in \mathbb{R}^{n \times k} \text{ such that } A \approx UV$$

Why? Matrices are space and time intensive



This work: Low-rank approximation of distance matrices in $\tilde{O}(n)$ time

Result: Algorithm

Input: distance matrix $A \in \mathbb{R}^{n \times n}$, k, and $\epsilon > 0$.

- **Sublinear runtime:** $\tilde{O}(n) \cdot poly(k, \epsilon^{-1})$
- **Approximation:** returns $U, V^T \in \mathbb{R}^{n \times k}$ s.t.

$$\|A - UV\|_F^2 \leq \underbrace{\|A - A_k\|_F^2}_{\text{Optimal (SVD) error}} + \underbrace{\epsilon \|A\|_F^2}_{\text{Additional error}}$$

- **Simple and practical**

Prior work [BW18]: $\tilde{O}(n^{1+\gamma}) \cdot poly(k, \epsilon^{-1})$.

Result: Lower Bound

Tight query complexity:

Our algorithm reads $O(nk\epsilon^{-1})$ entries of A.

Theorem: Any algorithm with the same guarantee must read $\Omega(nk\epsilon^{-1})$ entries of A.

Method

Theorem (Frieze, Kannan, Vempala 2004): For any matrix,

Sampling rows proportionally to ℓ_2-norms $\Rightarrow$ **Low-rank approximation**

Thus for distance matrices, our goal is, for all $x \in \mathcal{X}$:

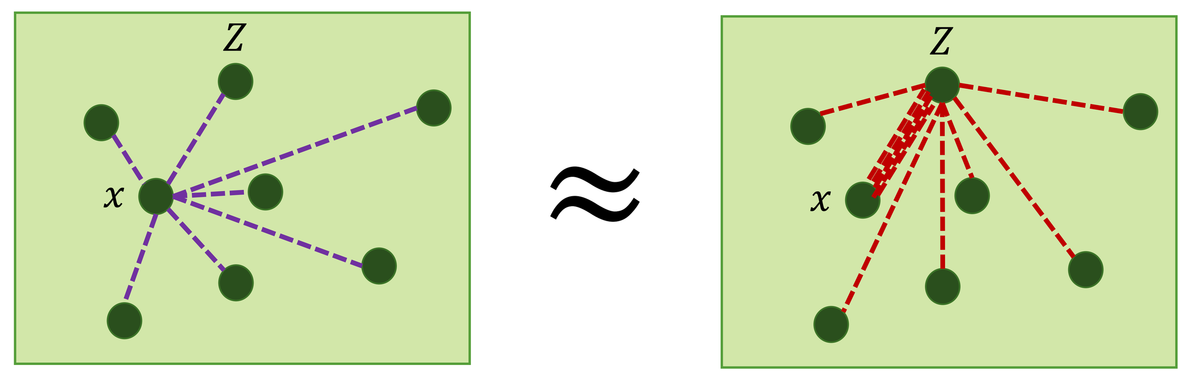
Estimate $\|A_{x,:}\|_2^2 = \sum_y \mathbf{d}(x, y)^2$

Our method:

- Pick $Z \sim \mathcal{X}$ uniformly at random
- Estimate each distance by the detour through Z:

$$\mathbf{d}(x, y)^2 \approx \mathbf{d}(x, Z)^2 + \mathbf{d}(Z, y)^2$$

- Thus, $\sum_y \mathbf{d}(x, y)^2 \approx n \cdot \mathbf{d}(x, Z)^2 + \sum_y \mathbf{d}(Z, y)^2$



This involves only distances from Z -- hence, $\tilde{O}(n)$ time.

Experiment: MNIST with Euclidean Distance

Method	Comments	Analytic Time	Empirical Time
SVD	Optimal error	$O(n^3)$	398.50
[CW13]	Input-sparsity time for arbitrary matrices	$\tilde{O}(n^2)$	34.32
[BW18]	Prior work on distance matrices	$\tilde{O}(n^{1+\gamma})$	4.17
Ours		$\tilde{O}(n)$	1.23

(secs, $k = 40$)