CSCI B609:
"Foundations of Data Science"

Lecture 6/7: Best-Fit Subspaces and Singular Value Decomposition

Slides at http://grigory.us/data-science-class.html

Grigory Yaroslavtsev

http://grigory.us

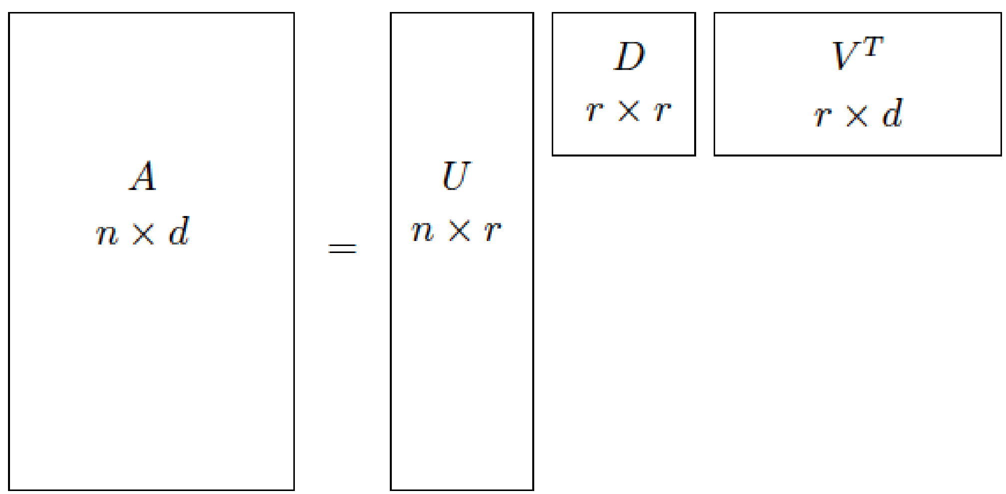
Singular Value Decomposition: Intro

- $\boldsymbol{n} \times \boldsymbol{d}$ data matrix A ($\boldsymbol{n}$ rows and $\boldsymbol{d}$ columns)
- Each row is a $\boldsymbol{d}$-dimensional vector
- Find best-fit k-dim. subspace S_k for rows of A?
- Minimize sum of squared distances from A_i to S_k

SVD: Greedy Strategy

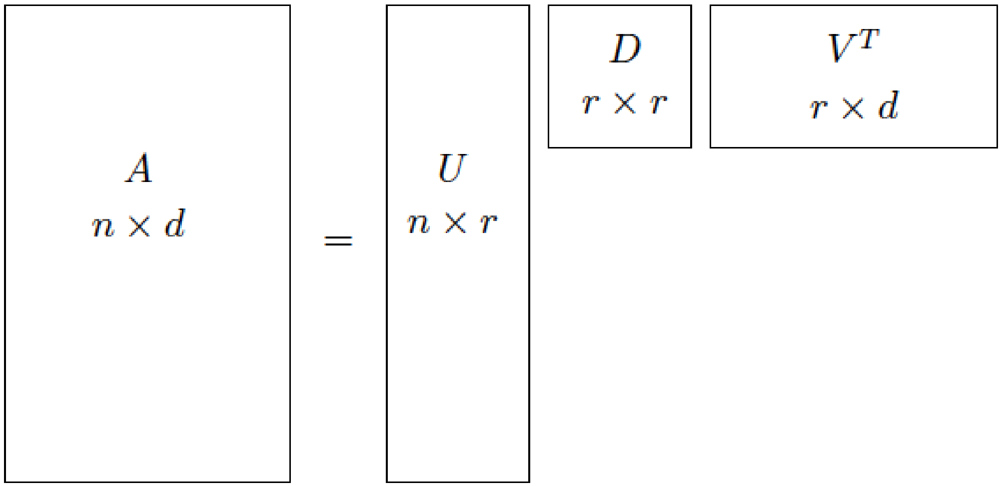
- Find best fit 1-dimensional line
- Repeat k times
- When $k = r = rank(A)$ we get the SVD:

$$A = UDV^T$$



$A = UDV^T$: Basic Properties

- $D =$ Diagonal matrix (positive real entries d_{ii})
- U, V: orthonormal columns:
  - $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_r \in \mathbb{R}^d$ (best fitting lines)
  - $\boldsymbol{u}_1, \ldots, \boldsymbol{u}_r \in \mathbb{R}^n$ (~projections of rows of A on $\boldsymbol{v}_i's$)
  - $\langle \boldsymbol{u}_i, \boldsymbol{u}_j \rangle = \delta_{ij}, \langle \boldsymbol{v}_i, \boldsymbol{v}_j \rangle = \delta_{ij}$
- $A = \sum_i d_{ii} \boldsymbol{u}_i \boldsymbol{v}_i^T$



Singular Values vs. Eigenvalues

- If A is a square matrix:
  - Vector $\boldsymbol{v}$ such that $A\boldsymbol{v} = \lambda\boldsymbol{v}$ is an eigenvector
  - λ = eigenvalue
  - For symmetric real matrices $\boldsymbol{v}$'s are orthonormal

$$A = VDV^T$$

  - $V's$ columns are eigenvectors of A
  - Diagonal entries of D are eigenvalues $\lambda_1, \ldots, \lambda_n$
- SVD is defined for all matrices (not just square)
  - Orthogonality of singular vectors is automatic

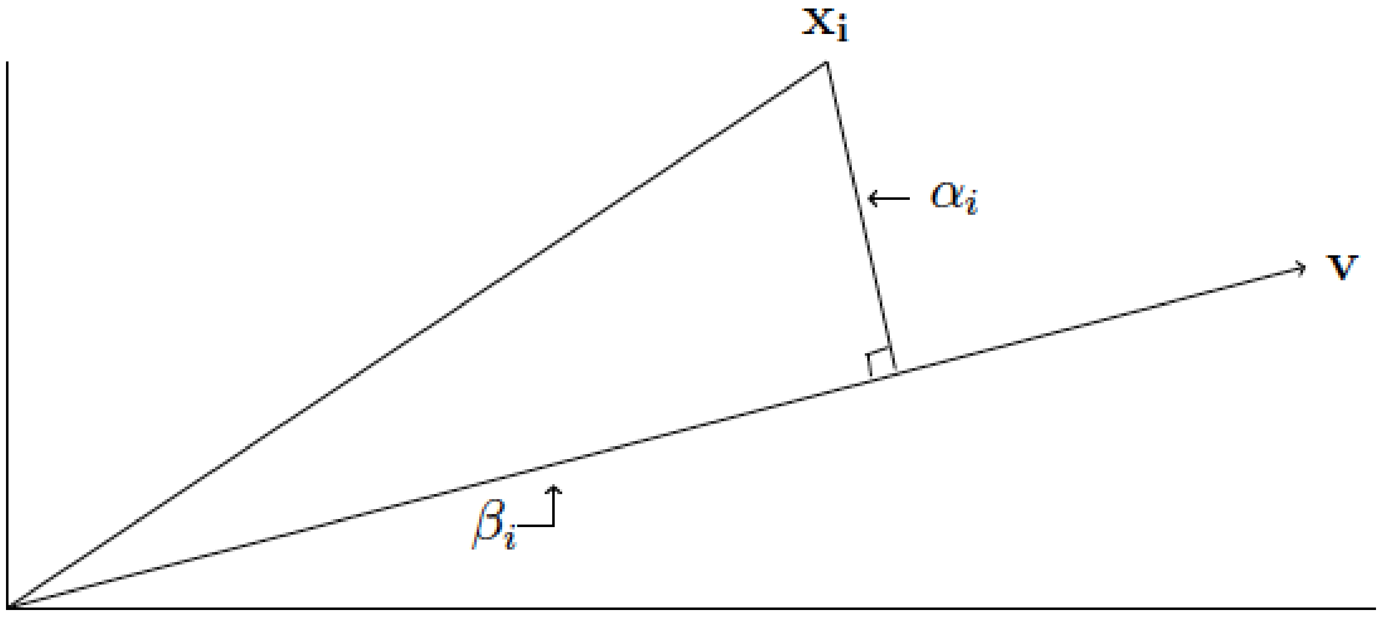
$$A\boldsymbol{v}_i = d_{ii}\boldsymbol{u}_i \text{ and } A^T\boldsymbol{u}_i = d_{ii}\boldsymbol{v}_i \text{ (will show)}$$

$$A^TA\boldsymbol{v}_i = d_{ii}^2\boldsymbol{v}_i \Rightarrow \boldsymbol{v}_i's \text{ are eigenvectors of } A^TA$$

Projections and Distances

- Minimizing distance = maximizing projection

$$||\boldsymbol{x}||_2^2 = (projection)^2 + (distance\ to\ line)^2$$



SVD: First Singular Vector

- Find best fit 1-dimensional line
- $\boldsymbol{v}$ = unit vector along the best fit line
- $\boldsymbol{a}_i$= i-th row of A, length of its projection: $|\langle \boldsymbol{a}_i, \boldsymbol{v} \rangle|$
- Sum of squared projection lengths: $||A\boldsymbol{v}||_2^2$
- **First singular vector**:

$$\boldsymbol{v}_1 = \arg \max_{||\boldsymbol{v}||_2 = 1} ||A\boldsymbol{v}||_2$$

- If there are ties, break arbitrarily
- $\sigma_1(A) = ||A\boldsymbol{v}_1||_2$ is the **first singular value**

SVD: Greedy Construction

- Find best fit 1-dimensional line, repeat r times (until projection is 0)
- **Second singular vector and value:**

$$\boldsymbol{v}_2 = \arg \max_{\boldsymbol{v} \perp \boldsymbol{v}_1, ||\boldsymbol{v}||_2 = 1} ||A\boldsymbol{v}||_2$$

$$\sigma_2(A) = ||A\boldsymbol{v}_2||_2$$

- **k-th singular vector and value:**

$$\boldsymbol{v}_k = \arg \max_{\boldsymbol{v} \perp \boldsymbol{v}_1, \ldots \boldsymbol{v}_{k-1}, ||\boldsymbol{v}||_2 = 1} ||A\boldsymbol{v}||_2$$

$$\sigma_k(A) = ||A\boldsymbol{v}_k||_2$$

- Will show: $(\boldsymbol{v}_1, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_k)$ is best-fit subspace

Best-Fit Subspace Proof: $k = 2$

- W = best-fit 2-dimensional subspace
- Orthonormal basis $(\boldsymbol{w}_1, \boldsymbol{w}_2)$: $||A\boldsymbol{w}_1||_2^2 + ||A\boldsymbol{w}_2||_2^2$
- Key observation: choose $\boldsymbol{w}_2 \perp \boldsymbol{v}_1$
  - If $W \perp \boldsymbol{v}_1$ then any vector in W works
  - Otherwise $\boldsymbol{v}_1 = \boldsymbol{v}_1^{||} + \boldsymbol{v}_1^{\perp}$ for $\boldsymbol{v}_1^{||}$ = projection on W
  - Choose $\boldsymbol{w}_2 \perp \boldsymbol{v}_1^{||}$:

    $\langle \boldsymbol{w}_2, \boldsymbol{v}_1 \rangle = \langle \boldsymbol{w}_2, \boldsymbol{v}_1^{||} + \boldsymbol{v}_1^{\perp} \rangle = \langle \boldsymbol{w}_2, \boldsymbol{v}_1^{||} \rangle + \langle \boldsymbol{w}_2, \boldsymbol{v}_1^{\perp} \rangle = 0$
- $||A\boldsymbol{w}_1||_2^2 \leq ||A\boldsymbol{v}_1||_2^2$ and $||A\boldsymbol{w}_2||_2^2 \leq ||A\boldsymbol{v}_2||_2^2$

$$||A\boldsymbol{w}_1||_2^2 + ||A\boldsymbol{w}_2||_2^2 \leq ||A\boldsymbol{v}_1||_2^2 + ||A\boldsymbol{v}_2||_2^2$$

Best-Fit Subspace Proof: General k

- W = best-fit k-dimensional subspace
- $V_{k-1} = span(\boldsymbol{v}_1, \dots, \boldsymbol{v}_{k-1})$ best fit (k-1)-dimensional subspace
- Orthonormal basis $\boldsymbol{w}_1, \dots, \boldsymbol{w}_k$, where $\boldsymbol{w}_k \perp V_{k-1}$

$$\sum_{i=1}^{k-1} \left|\left|A\boldsymbol{w}_i\right|\right|_2^2 \leq \sum_{i=1}^{k-1} \left|\left|A\boldsymbol{v}_i\right|\right|_2^2$$

- $\boldsymbol{w}_k \perp V_{k-1} \Rightarrow$ by def. of $\boldsymbol{v}_k$ $\left|\left|A\boldsymbol{w}_k\right|\right|_2^2 \leq \left|\left|A\boldsymbol{v}_k\right|\right|_2^2$

$$\sum_{i=1}^{k} \left|\left|A\boldsymbol{w}_i\right|\right|_2^2 \leq \sum_{i=1}^{k} \left|\left|A\boldsymbol{v}_i\right|\right|_2^2$$

Singular Values and Frobenius Norm

- $\boldsymbol{v}_1, \dots, \boldsymbol{v}_r$ span the space of all rows of A
- $\langle \boldsymbol{a}_j, \boldsymbol{v} \rangle = 0$ for all $\boldsymbol{v} \perp \boldsymbol{v}_1, \dots, \boldsymbol{v}_r \Rightarrow$

$$\left|\left|\boldsymbol{a}_j\right|\right|_2^2 = \sum_{i=1}^{r} \langle \boldsymbol{a}_j, \boldsymbol{v}_i \rangle^2$$

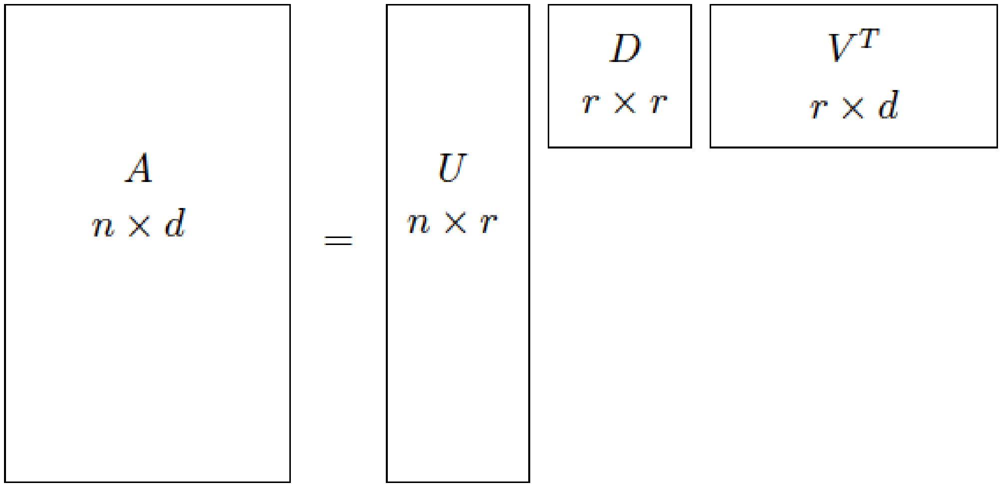
$$\sum_{j=1}^{n} \sum_{k=1}^{d} a_{jk}^2 = \sum_{j=1}^{n} \left|\left|\boldsymbol{a}_j\right|\right|_2^2 = \sum_{j=1}^{n} \sum_{i=1}^{r} \langle \boldsymbol{a}_j, \boldsymbol{v}_i \rangle^2 =$$

$$\sum_{i=1}^{r} \sum_{j=1}^{n} \langle \boldsymbol{a}_j, \boldsymbol{v}_i \rangle^2 = \sum_{i=1}^{r} \left|\left|A\boldsymbol{v}_i\right|\right|_2^2 = \sum_{i=1}^{r} \sigma_i^2(A)$$

- $\sqrt{\sum_{j=1}^{n} \sum_{k=1}^{d} a_{jk}^2} = \left|\left|\mathrm{A}\right|\right|_{\mathrm{F}}$ (Frobenius norm) $= \sqrt{\sum_{i=1}^{r} \sigma_i^2(A)}$

Singular Value Decomposition

- $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_r$ are **right singular vectors**
- $\left|\left|A\boldsymbol{v}_i\right|\right|_2 = \sigma_i(A)$ are **singular values**
- $\boldsymbol{u}_1, \ldots, \boldsymbol{u}_r$ for $\boldsymbol{u}_i = \frac{A\boldsymbol{v}_i}{\sigma_i(A)}$ are **left singular vectors**



Singular Value Decomposition

- Will prove that $A = \sum_{i=1}^{r} \sigma_i \boldsymbol{u}_i \boldsymbol{v}_i^T$
- **Lem.** $A = B$ iff $\forall \boldsymbol{v}$: $A\boldsymbol{v} = B\boldsymbol{v}$
- $\sum_{i=1}^{r} \sigma_i \boldsymbol{u}_i \boldsymbol{v}_i^T \boldsymbol{v}_j = \sigma_j \boldsymbol{u}_j = A\boldsymbol{v}_j$
- $\boldsymbol{v}$ = linear combination of $\boldsymbol{v}_j's$ + orthogonal
- Duplicate singular values ⇒ singular values are not unique, but always can choose orthogonal

Best rank-k Approximation

- $A_k = \sum_{i=1}^{k} \sigma_i \boldsymbol{u}_i \boldsymbol{v}_i^T$
- A_k = best rank-k approx. in Frobenius norm
- **Lem**: rows of A_k = projections on span$(\boldsymbol{v}_1, \dots, \boldsymbol{v}_k)$
  - Projection of $\boldsymbol{a_i}$ = $\sum_{i=1}^{k} \langle \boldsymbol{a}_i, \boldsymbol{v}_i \rangle \boldsymbol{v}_i^T$
  - Projections of A: $\sum_{i=1}^{k} A \boldsymbol{v}_i \boldsymbol{v}_i^T = \sum_{i=1}^{k} \sigma_i \boldsymbol{u}_i \boldsymbol{v}_i^T = A_k$
- For any matrix B of rank $\leq k$ (convergence of greedy)

$$||A - A_k||_F \leq ||A - B||_F$$

- Recall: if $\boldsymbol{v}_i$ are orthonormal basis for column space:

$$||\mathrm{A}||_F^2 = \sum_{j=1}^{n} \sum_{i=1}^{k} \langle \boldsymbol{a}_j, \boldsymbol{v}_i \rangle^2 \Rightarrow \text{maximum for projections}$$

Rank-k Approximation and Similarity

- Database A: $\boldsymbol{n} \times \boldsymbol{d}$ matrix (document × term)
- Preprocess to answer similarity queries:
  - Query $\boldsymbol{x} \in \mathbb{R}^{\boldsymbol{d}}$ = new document
  - Output: $A\boldsymbol{x} \in \mathbb{R}^{\boldsymbol{n}}$ = vector of similarities
  - Naïve approach takes $O(\boldsymbol{nd})$ time
- If we construct $A_k = \sum_{i=1}^{k} \sigma_i \boldsymbol{u}_i \boldsymbol{v}_i^T$ first
  - $A_k \boldsymbol{x} = \sum_{i=1}^{k} \sigma_i \boldsymbol{u}_i (\boldsymbol{v}_i^T \boldsymbol{x}) \Rightarrow O(k\boldsymbol{d} + \boldsymbol{n}k)$ time
  - Error: $\max_{||x||_2 \leq 1} ||(A - A_k)\boldsymbol{x}|| \equiv ||(A - A_k)||_2$
  - $||(A - A_k)||_2 = \sigma_1(A - A_k) = \sigma_{k+1}(A)$

Left Singular Values and Spectral Norm

See Section 3.6 for proofs

- Left singular vectors $\boldsymbol{u}_1, \dots, \boldsymbol{u}_k$ or orthogonal
- $\left|\left|(A - A_k)\right|\right|_2 = \sigma_{k+1}$
- For any rank $\leq k$ matrix B

$$\left|\left|A - A_k\right|\right|_2 \leq \left|\left|A - B\right|\right|_2$$

- $A\boldsymbol{v}_i = d_{ii}\boldsymbol{u}_i$ and $A^T\boldsymbol{u}_i = d_{ii}\boldsymbol{v}_i$

Power Method

- $B = A^T A$ is a $\boldsymbol{d} \times \boldsymbol{d}$ matrix
- $B = \left(\sum_{i=1}^{r} \sigma_i \boldsymbol{u}_i \boldsymbol{v}_i^T\right)^T \left(\sum_{i=1}^{r} \sigma_i \boldsymbol{u}_i \boldsymbol{v}_i^T\right) =$

$$= \left(\sum_{i=1}^{r} \sigma_i \boldsymbol{v}_i \boldsymbol{u}_i^T\right)\left(\sum_{j=1}^{r} \sigma_i \boldsymbol{u}_i \boldsymbol{v}_i^T\right) =$$

$$\sum_{i,j=1}^{r} \sigma_i \sigma_j \boldsymbol{v}_i (\boldsymbol{u}_i^T \boldsymbol{u}_j)\, \boldsymbol{v}_j^T = \sum_{i=1}^{r} \sigma_i^2 \boldsymbol{v}_i \boldsymbol{v}_i^T$$

- $B^2 = \left(\sum_{i=1}^{r} \sigma_i^2 \boldsymbol{v}_i \boldsymbol{v}_i^T\right)^T \left(\sum_{j=1}^{r} \sigma_j^2 \boldsymbol{v}_j \boldsymbol{v}_j^T\right) = \sum_{i=1}^{r} \sigma_i^4\, \boldsymbol{v}_i \boldsymbol{v}_i^T$
- $B^k = \sum_{i=1}^{r} \sigma_i^{2k}\, \boldsymbol{v}_i \boldsymbol{v}_i^T \Rightarrow$ if $\sigma_1 > \sigma_2$ take scaled 1st row

Faster Power Method

- PM drawback: $A^T A$ is dense even for sparse A
- Pick random Gaussian $\boldsymbol{x}$ and compute $B^k \boldsymbol{x}$
- $\boldsymbol{x} = \sum_{i=1}^{\boldsymbol{d}} c_i \boldsymbol{v}_i$(augment $\boldsymbol{v_i}$'s to o.n.b. if $r < \boldsymbol{d}$)
- $B^k \boldsymbol{x} \approx \left(\sigma_1^{2k} \boldsymbol{v}_1 \boldsymbol{v}_1^T\right)\left(\sum_{i=1}^{d} c_i \boldsymbol{v}_i\right) = \sigma_1^{2k} c_1 \boldsymbol{v}_1$

$$B^k \boldsymbol{x} = (A^T A)(A^T A) \ldots (A^T A)\boldsymbol{x}$$

- **Theorem:** If $\boldsymbol{x}$ is unit $\mathbb{R}^{\boldsymbol{d}}$-vector, $|\boldsymbol{x}^T \boldsymbol{v}_1| \geq \boldsymbol{\delta}$:
  - V = subspace spanned by $\boldsymbol{v}_i' s$ for $\sigma_j \geq (1 - \boldsymbol{\epsilon})\sigma_1$
  - $\boldsymbol{w}$ = unit vector after $k = \frac{1}{2\boldsymbol{\epsilon}} \ln\left(\frac{1}{\boldsymbol{\epsilon\delta}}\right)$ iterations of PM

$\Rightarrow$ $\boldsymbol{w}$ has a component at most $\boldsymbol{\epsilon}$ orthogonal to V

Faster Power Method: Analysis

- $A = \sum_{i=1}^{r} \sigma_i \boldsymbol{u}_i \boldsymbol{v}_i^T$ and $\boldsymbol{x} = \sum_{i=1}^{\boldsymbol{d}} c_i \boldsymbol{v}_i$
- $B^{k}\boldsymbol{x} = \sum_{i=1}^{\boldsymbol{d}} \sigma_i^{2k} \boldsymbol{v}_i \boldsymbol{v}_i^T \sum_{j=1}^{\boldsymbol{d}} c_j \boldsymbol{v}_j = \sum_{i=1}^{\boldsymbol{d}} \sigma_i^{2k} c_i \boldsymbol{v}_i$

$$\left\|B^{k}\boldsymbol{x}\right\|_2^2 = \left\|\sum_{i=1}^{\boldsymbol{d}} \sigma_i^{2k} c_i \boldsymbol{v}_i\right\|_2^2 = \sum_{i=1}^{\boldsymbol{d}} \sigma_i^{4k} c_i^2 \geq \sigma_1^{4k} c_1^2 \geq \sigma_i^{4k} \delta^2$$

- (Squared) component orthogonal to V is

$$\sum_{i=m+1}^{\boldsymbol{d}} \sigma_i^{4k} c_i^2 \leq (1-\epsilon)^{4k} \sigma_1^{4k} \sum_{i=m+1}^{\boldsymbol{d}} c_i^2 \leq (1-\epsilon)^{4k} \sigma_1^{4k}$$

- Component of $\boldsymbol{w} \perp V \leq (1-\epsilon)^{2k} / \boldsymbol{\delta} \leq \epsilon$

Choice of $\boldsymbol{x}$

- $\boldsymbol{y}$ random spherical Gaussian with unit variance
- $\boldsymbol{x} = \frac{\boldsymbol{y}}{||\boldsymbol{y}||_2}$:

$$Pr\left[\left|\boldsymbol{x}^T\boldsymbol{v}\right| \leq \frac{1}{20\sqrt{\boldsymbol{d}}}\right] \leq \frac{1}{10} + 3e^{-\boldsymbol{d}/64}$$

- $Pr\left[||\boldsymbol{y}||_2 \geq \boldsymbol{2}\sqrt{\boldsymbol{d}}\right] \leq 3e^{-\boldsymbol{d}/64}$ (Gaussian Annulus)
- $\boldsymbol{y}^T\boldsymbol{v} \sim N(0,1) \Rightarrow \Pr\left[\left|\left|\boldsymbol{y}^T\boldsymbol{v}\right|\right|_2 \leq \frac{1}{10}\right] \leq \frac{1}{10}$
- Can set $\boldsymbol{\delta} = \frac{1}{20\sqrt{\boldsymbol{d}}}$ in the “faster power method”

Singular Vectors and Eigenvectors

- Right singular vectors are eigenvectors of $A^T A$
- σ_i^2 are eigenvalues of $A^T A$
- Left singular vectors are eigenvectors of AA^T
- $A^T A$ satisfies $\forall \boldsymbol{x}$: $\boldsymbol{x}^T B \boldsymbol{x} \geq 0$
  - $B = \sum_i \sigma_i^2 \boldsymbol{v}_i \boldsymbol{v}_i^T$
  - $\forall \boldsymbol{x}$: $\boldsymbol{x}^T \boldsymbol{v}_i \boldsymbol{v}_i^T \boldsymbol{x} = (\boldsymbol{x}^T \boldsymbol{v}_i)^2 \geq 0$
  - Such matrices are called positive semi-definite
- Any p.s.d matrix can be decomposed as $A^T A$