Dimensionality Reduction for Non-Linear Clustering Via Nystrom Approximation

Alex Gittens, Shusen Wang, Michael W. Mahoney

Overview

- Clustering: from linear k-means to kernel k-means.
- Scalable kernel k-means via the Nyström method .
- A novel $1 + \epsilon$ relative-error bound for Nyström.
- Kernel k-means VS spectral clustering.

K-Means Clustering

Clustering

Input: vectors $\mathbf{a}_1, \cdots, \mathbf{a}_n \in \mathbb{R}^d$ and cluster number k ($\ll n$).

Output: labels $y_1, \cdots, y_n \in \{1, 2, \cdots, k\}$.

Clustering

Input: vectors $\mathbf{a}_1, \cdots, \mathbf{a}_n \in \mathbb{R}^d$ and cluster number k ($\ll n$).

Output: labels $y_1, \cdots, y_n \in [k]$.

denote $[k] = \{1, 2, \cdots, k\}$

Clustering

Input: vectors $\mathbf{a}_1, \cdots, \mathbf{a}_n \in \mathbb{R}^d$ and desired cluster count k ($\ll n$).

Output: labels $y_1, \cdots, y_n \in [k]$.

Equivalently:

Clustering: partition $[n]$ into k disjoint sets $\mathcal{J}_1, \cdots, \mathcal{J}_k$,

- $\mathcal{J}_1 \cup \cdots \cup \mathcal{J}_k = [n]$,
- $\mathcal{J}_i \cap \mathcal{J}_j = \emptyset$, for all $i \neq j$.

K-Means Clustering

Clustering: partition $[n]$ into k disjoint sets $\mathcal{J}_1, \cdots, \mathcal{J}_k$,

- $\mathcal{J}_1 \cup \cdots \cup \mathcal{J}_k = [n]$,
- $\mathcal{J}_i \cap \mathcal{J}_j = \emptyset$, for all $i \neq j$.

K-Means Clustering:

$$\min_{\mathcal{J}_1, \cdots, \mathcal{J}_k} \sum_{i=1}^{k} \sum_{j \in \mathcal{J}_i} \left\| \boxed{\mathbf{a}_j} - \boxed{\frac{1}{|\mathcal{J}_i|} \sum_{l \in \mathcal{J}_i} \mathbf{a}_l} \right\|_2^2 .$$

- $\mathbf{a}_j$ belongs to $\mathcal{J}_i$

- Centroid of the i-th cluster.

Cluster Indicator Matrix

- Let sets $\mathcal{J}_1, \cdots, \mathcal{J}_k$ be an arbitrary partition of $[n]$.
- Define matrix $\mathbf{X} \in \mathbb{R}^{n \times k}$

$$x_{ji} = \begin{cases} |\mathcal{J}_i|^{-\frac{1}{2}}, & \text{if } \ j \in \mathcal{J}_i; \\ 0, & \text{else.} \end{cases}$$

- Example: partition $[12]$ to $k = 3$ disjoint sets.

$$\mathbf{X}^T = \begin{bmatrix} \frac{1}{\sqrt{4}} & 0 & \frac{1}{\sqrt{4}} & 0 & \frac{1}{\sqrt{4}} & 0 & \frac{1}{\sqrt{4}} & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & \frac{1}{\sqrt{3}} & 0 & 0 & 0 & 0 & \frac{1}{\sqrt{3}} & 0 & \frac{1}{\sqrt{3}} & 0 \\ 0 & \frac{1}{\sqrt{5}} & 0 & 0 & 0 & \frac{1}{\sqrt{5}} & 0 & \frac{1}{\sqrt{5}} & 0 & \frac{1}{\sqrt{5}} & 0 & \frac{1}{\sqrt{5}} \end{bmatrix}$$

J_1={1, 3, 5, 7}
J_2={4, 9, 11}
J_3={2, 6, 8, 10, 12}

Cluster Indicator Matrix

- Let sets $\mathcal{J}_1, \cdots, \mathcal{J}_k$ be an arbitrary partition of $[n]$.
- Define matrix $\mathbf{X} \in \mathbb{R}^{n \times k}$

$$x_{ji} = \begin{cases} |\mathcal{J}_i|^{-\frac{1}{2}}, & \text{if } \ j \in \mathcal{J}_i; \\ 0, & \text{else.} \end{cases}$$

- $\mathbf{X}$ is called a cluster indicator matrix.
  - Partition $\{\mathcal{J}_1, \cdots, \mathcal{J}_k\}$ $\longleftrightarrow$ $\mathbf{X}$ (one-to-one correspondence).
  - $\mathbf{X}$ has orthonormal columns.

Matrix Formulation of K-Means

- **K-Means Clustering:**

$$\min_{\mathcal{J}_1,\cdots,\mathcal{J}_k} \sum_{i=1}^{k} \sum_{j\in\mathcal{J}_i} \left\| \mathbf{a}_j - \frac{1}{|\mathcal{J}_i|} \sum_{l\in\mathcal{J}_i} \mathbf{a}_l \right\|_2^2 .$$

$$\Updownarrow$$

$$\min_{\mathbf{X}\in\mathcal{X}_{n,k}} \left\| \mathbf{A} - \mathbf{X}\mathbf{X}^{\mathrm{T}}\mathbf{A} \right\|_F^2 .$$

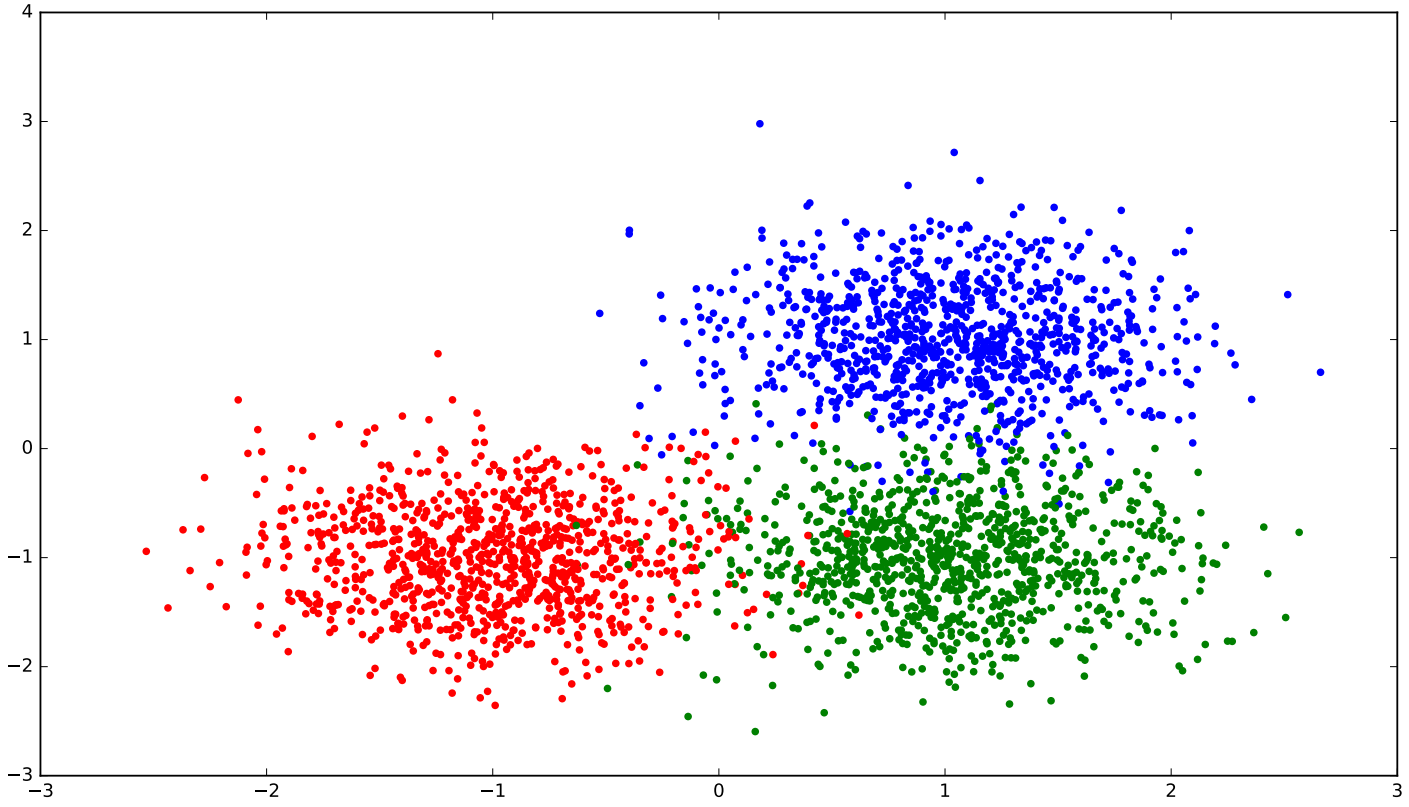
- $\mathbf{a}_1, \cdots, \mathbf{a}_n \in \mathbb{R}^d$ are the rows of $\mathbf{A} \in \mathbb{R}^{n\times d}$.
- $\mathcal{X}_{n,k} \subset \mathbb{R}^{n\times k}$ is the collection of all the $n\times k$ cluster indicator matrices.

Approximate K-Means Algorithms

- K-means $\min_{\mathbf{X} \in \mathcal{X}_{n,k}} \left|\left|\mathbf{A} - \mathbf{X}\mathbf{X}^T\mathbf{A}\right|\right|_F^2$ is NP-Hard
- γ-approximate algorithm:
  - output cluster indicator matrix $\widetilde{\mathbf{X}} \in \mathcal{X}_{n,k}$
  - satisfies $\left|\left|\mathbf{A} - \widetilde{\mathbf{X}}\widetilde{\mathbf{X}}^T\mathbf{A}\right|\right|_F^2 \leq \gamma \cdot \min_{\mathbf{X} \in \mathcal{X}_{n,k}} \left|\left|\mathbf{A} - \mathbf{X}\mathbf{X}^T\mathbf{A}\right|\right|_F^2$
- There are constant-factor and $1 + \epsilon$ algorithms.

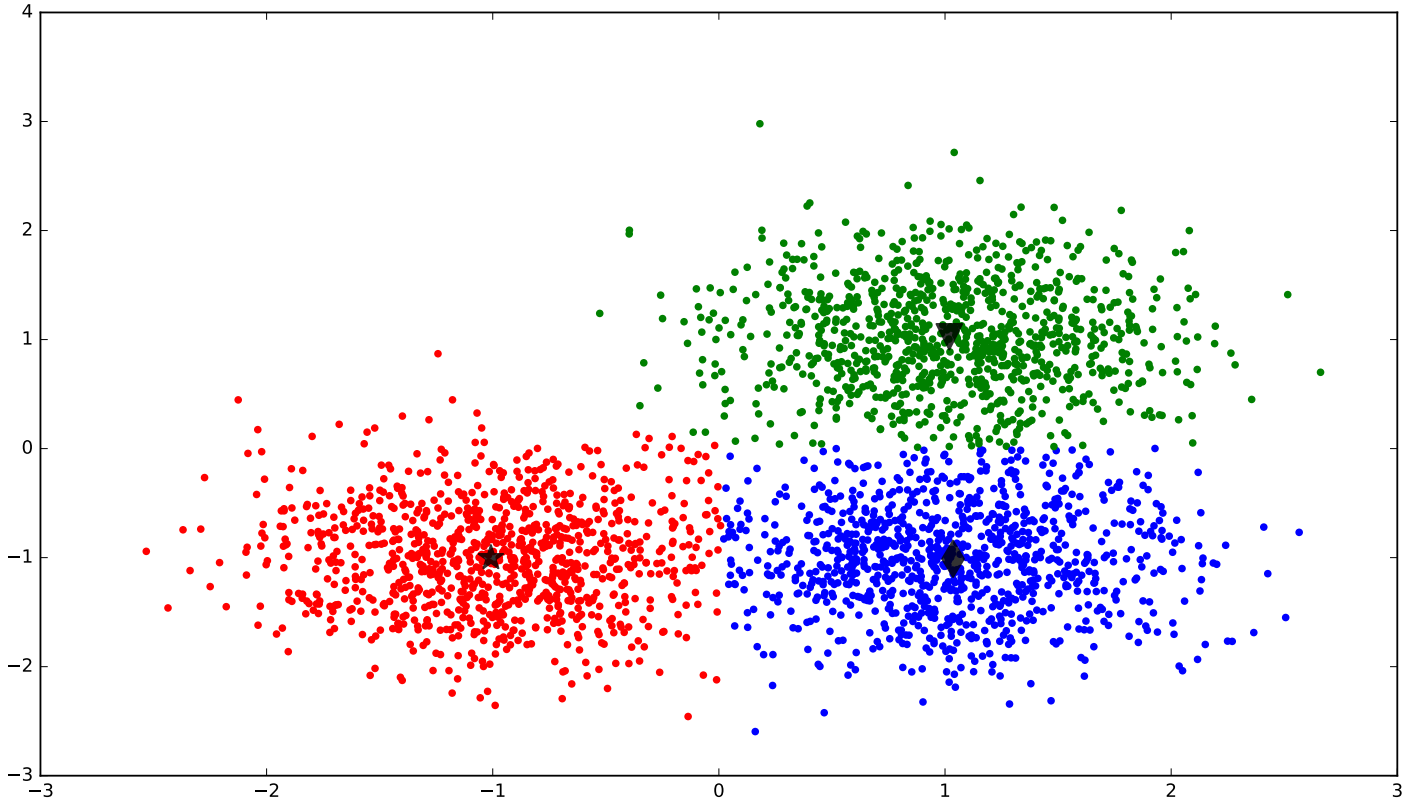
K-Means Clustering

- Example: blobs



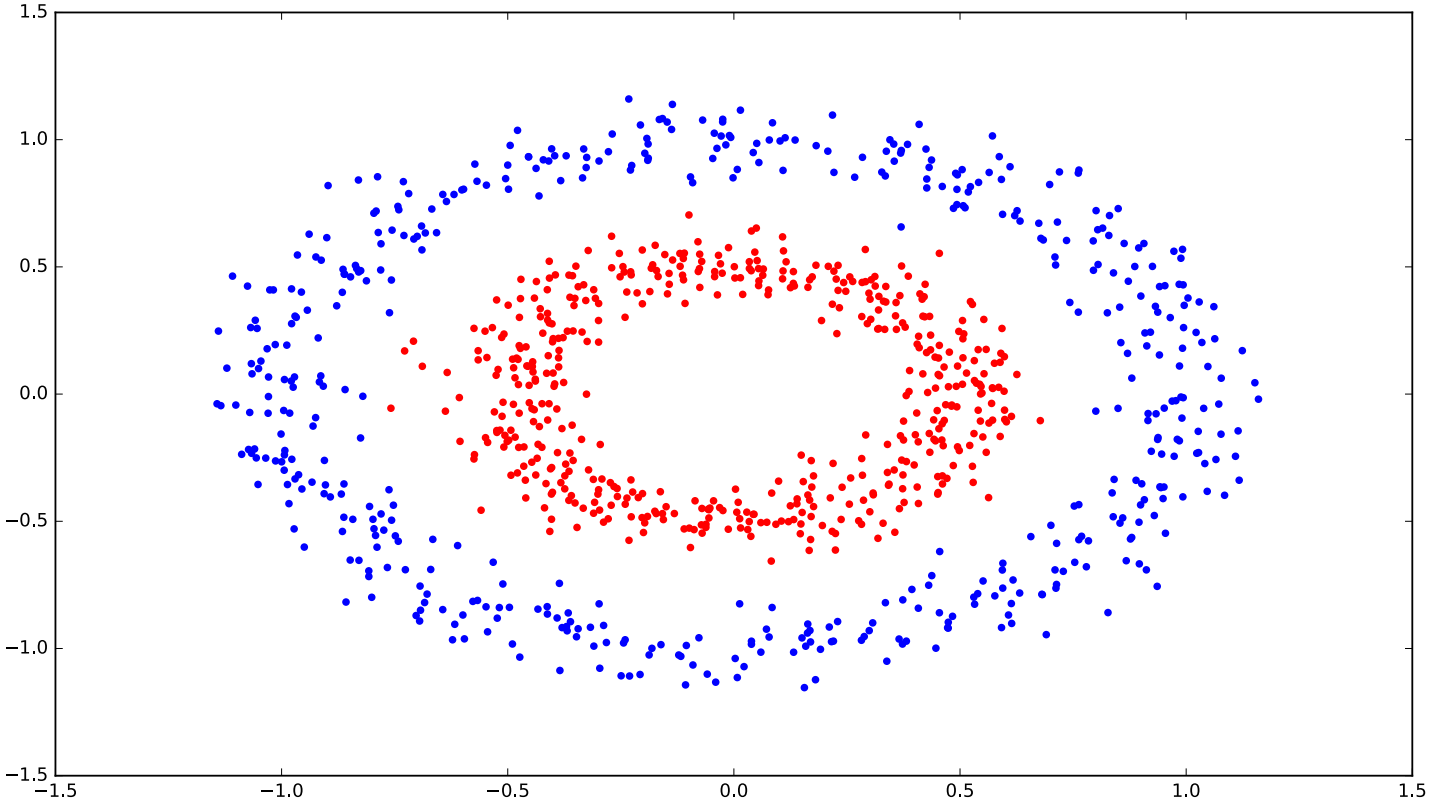
K-Means Clustering

- Output of Lloyd's algorithm



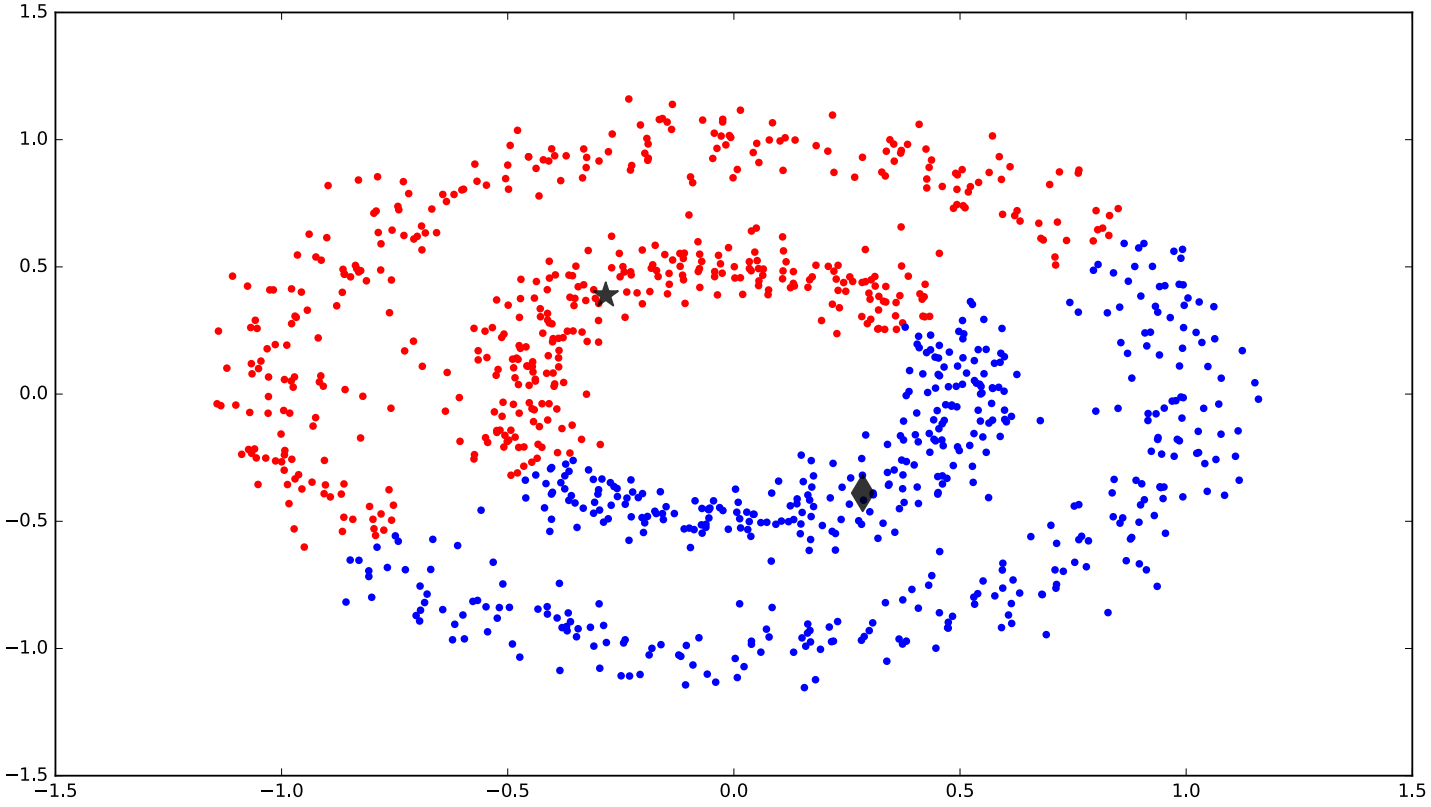
K-Means Clustering

- Example: circles



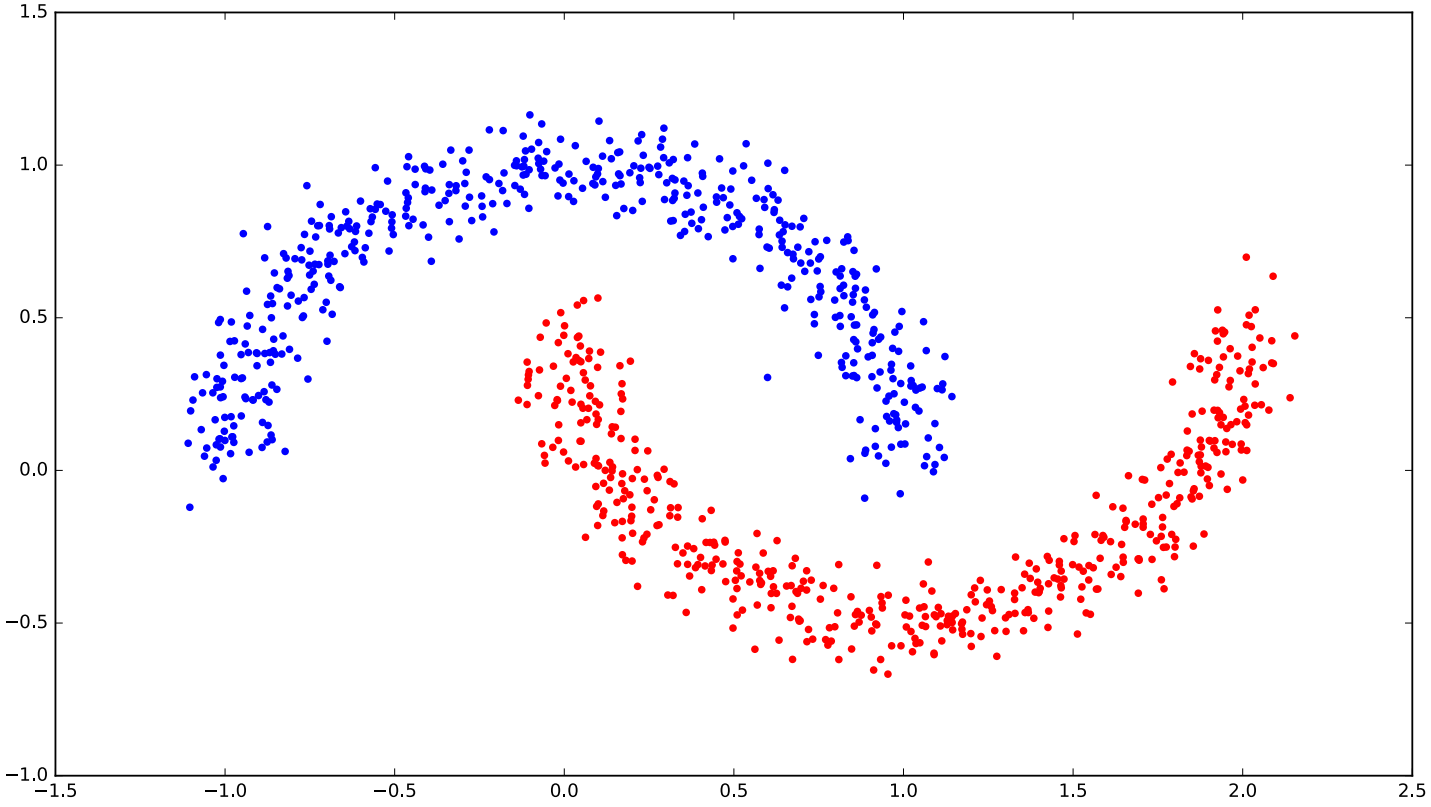
K-Means Clustering

- Output of Lloyd's algorithm



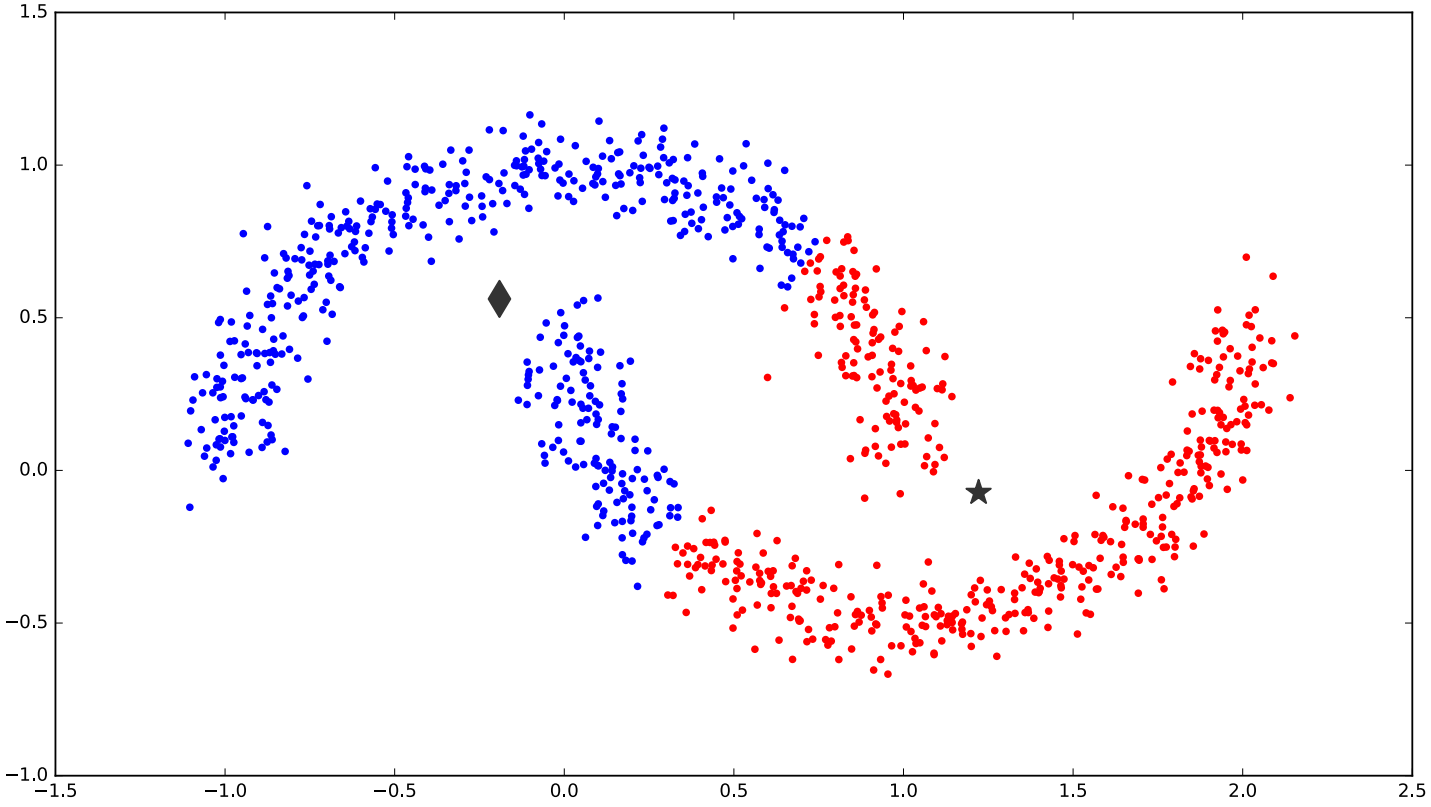
K-Means Clustering

- Example: moons



K-Means Clustering

- Output of Lloyd's algorithm



Kernel K-Means

Kernel Method

- Feature mapping $\boldsymbol{\phi}: \mathbb{R}^d \mapsto \mathcal{F}$.
  - Denote $\boldsymbol{\phi}_i = \boldsymbol{\phi}(\mathbf{a}_i)$.
  - Map $\mathbf{a}_1, \cdots, \mathbf{a}_n \in \mathbb{R}^d$ to $\boldsymbol{\phi}_1, \cdots, \boldsymbol{\phi}_n \in \mathcal{F}$.
- Kernel trick
  - Kernel function: $\kappa(\mathbf{a}_i, \mathbf{a}_j) = \langle \boldsymbol{\phi}(\mathbf{a}_i), \boldsymbol{\phi}(\mathbf{a}_j) \rangle = \langle \boldsymbol{\phi}_i, \boldsymbol{\phi}_j \rangle$.
  - Kernel matrix $\mathbf{K} \in \mathbb{R}^{n \times n}$, where $k_{ij} = \kappa(\mathbf{a}_i, \mathbf{a}_j)$.
- Let $\boldsymbol{\phi}_1, \cdots, \boldsymbol{\phi}_n$ be the rows of $\boldsymbol{\Phi}$. Then $\mathbf{K} = \boldsymbol{\Phi}\boldsymbol{\Phi}^T$.

Kernel K-Means

- K-means: $$\min_{\mathbf{X}\in\mathcal{X}_{n,k}} \left|\left|\left(\mathbf{I}_n - \mathbf{X}\mathbf{X}^{\mathbf{T}}\right)\mathbf{A}\right|\right|_F^2$$ K-means in input space.

- Kernel k-means: $$\min_{\mathbf{X}\in\mathcal{X}_{n,k}} \left|\left|\left(\mathbf{I}_n - \mathbf{X}\mathbf{X}^{\mathbf{T}}\right)\mathbf{\Phi}\right|\right|_F^2$$ K-means in feature space.

$$\Updownarrow$$

$$\min_{\mathbf{X}\in\mathcal{X}_{n,k}} \left|\left|\left(\mathbf{I}_n - \mathbf{X}\mathbf{X}^{\mathbf{T}}\right)\mathbf{K}^{1/2}\right|\right|_F^2$$

- Let $\mathbf{K} = \mathbf{V}\mathbf{\Sigma}\mathbf{V}^T$ be the SVD.
- $\mathbf{K}^{1/2}$ can be $\mathbf{V}\mathbf{\Sigma}^{1/2}$ or $\mathbf{V}\mathbf{\Sigma}^{1/2}\mathbf{V}^T$.

Kernel K-Means

- Kernel k-means:

$$\min_{\mathbf{X}\in\mathcal{X}_{n,k}} \left|\left|\left(\mathbf{I}_n - \mathbf{X}\mathbf{X}^{\mathbf{T}}\right)\mathbf{\Phi}\right|\right|_F^2$$

$$\Updownarrow$$

$$\min_{\mathbf{X}\in\mathcal{X}_{n,k}} \left|\left|\left(\mathbf{I}_n - \mathbf{X}\mathbf{X}^{\mathbf{T}}\right)\mathbf{K}^{1/2}\right|\right|_F^2$$

Proof.

$$\left|\left|\left(\mathbf{I}_n - \mathbf{X}\mathbf{X}^{\mathbf{T}}\right)\mathbf{\Phi}\right|\right|_F^2 = \mathrm{Tr}\left(\left(\mathbf{I}_n - \mathbf{X}\mathbf{X}^{\mathbf{T}}\right)\mathbf{\Phi}\mathbf{\Phi}^{\mathbf{T}}\left(\mathbf{I}_n - \mathbf{X}\mathbf{X}^{\mathbf{T}}\right)\right)$$

$$= \mathrm{Tr}\left(\left(\mathbf{I}_n - \mathbf{X}\mathbf{X}^{\mathbf{T}}\right)\mathbf{K}^{1/2}\mathbf{K}^{1/2}\left(\mathbf{I}_n - \mathbf{X}\mathbf{X}^{\mathbf{T}}\right)\right) = \left|\left|\left(\mathbf{I}_n - \mathbf{X}\mathbf{X}^{\mathbf{T}}\right)\mathbf{K}^{1/2}\right|\right|_F^2$$

Kernel K-Means

- Kernel k-means:

$$\min_{\mathbf{X} \in \mathcal{X}_{n,k}} \left\| \left(\mathbf{I}_n - \mathbf{X}\mathbf{X}^{\mathbf{T}} \right) \mathbf{\Phi} \right\|_F^2$$

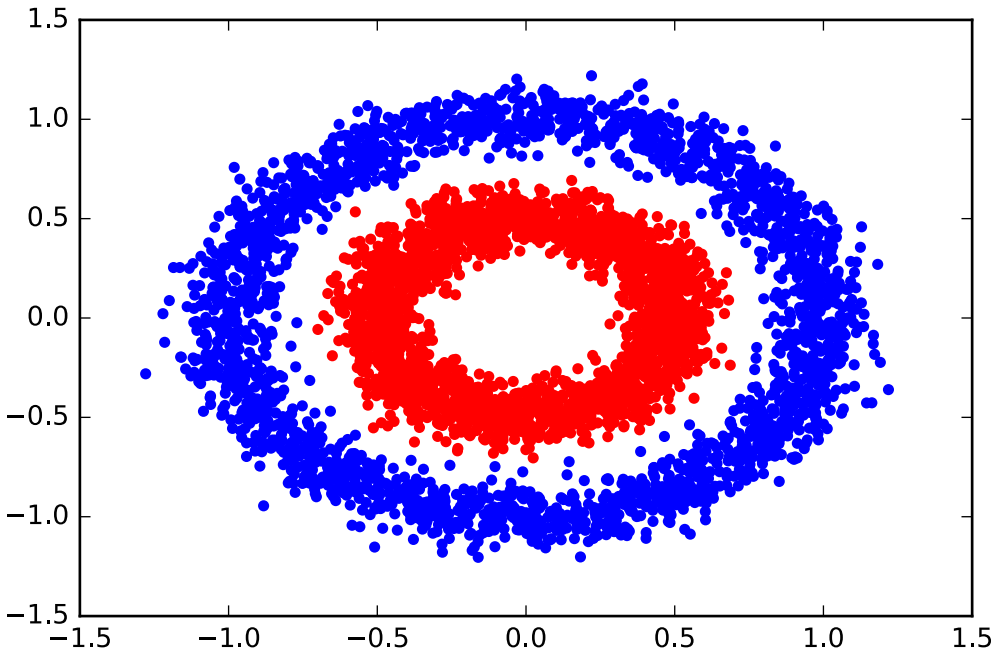
$$\Updownarrow$$

$$\min_{\mathbf{X} \in \mathcal{X}_{n,k}} \left\| \left(\mathbf{I}_n - \mathbf{X}\mathbf{X}^{\mathbf{T}} \right) \mathbf{K}^{1/2} \right\|_F^2$$

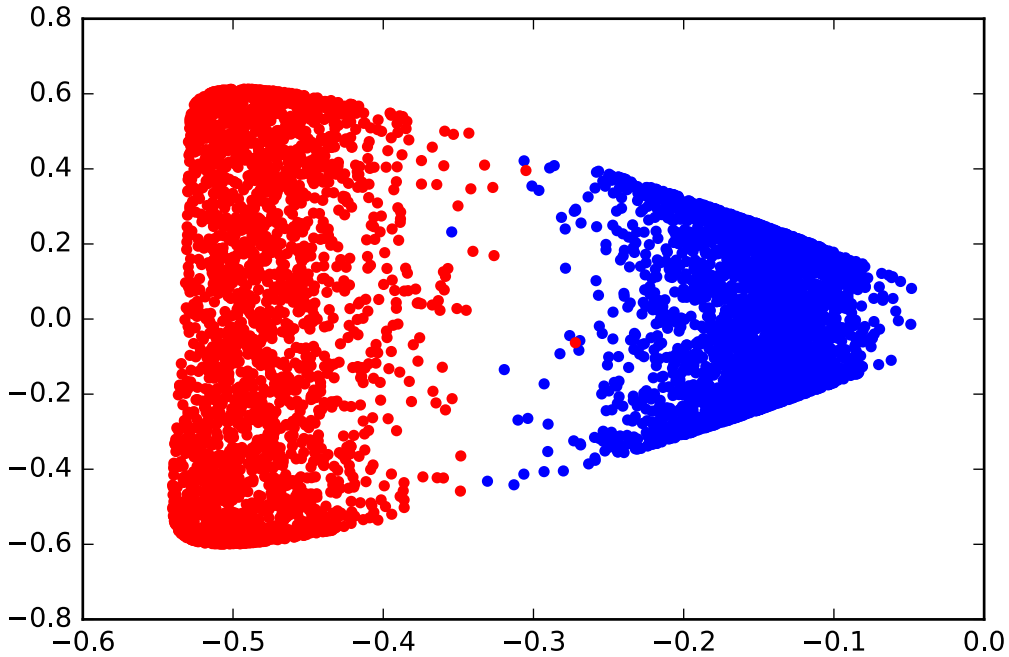
Kernel k-means is **standard k-means** over the rows of $\mathbf{K}^{1/2} \in \mathbb{R}^{n \times n}$.

Kernel K-Means

Raw 2-dim input vectors.



Top 2 principal components of $\mathbf{K}^{1/2}$



Kernel K-Means

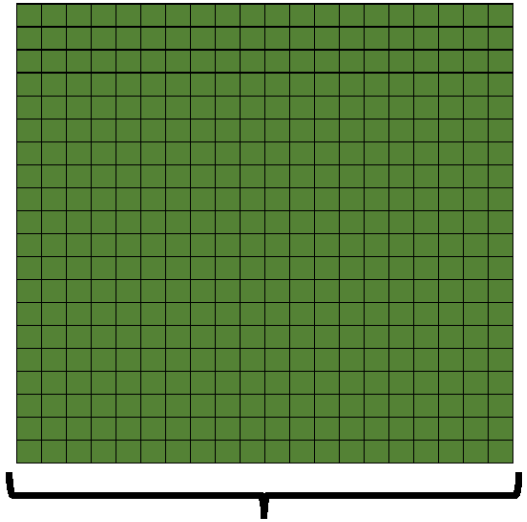
- Computational costs
  1. Form kernel matrix $\mathbf{K} \in \mathbb{R}^{n \times n}$: $O(n^2 d)$ time.
  2. Decomposition $\mathbf{K} = \mathbf{K}^{1/2} \ \mathbf{K}^{1/2}$: $O(n^3)$ time.
  3. Lloyd's algorithm over the rows of $\mathbf{K}^{1/2}$: $O(n^2 k)$ time per iteration.

Naïve kernel k-means is impractical!

Nyström for Kernel K-Means

Alex Gittens

Nyström Method

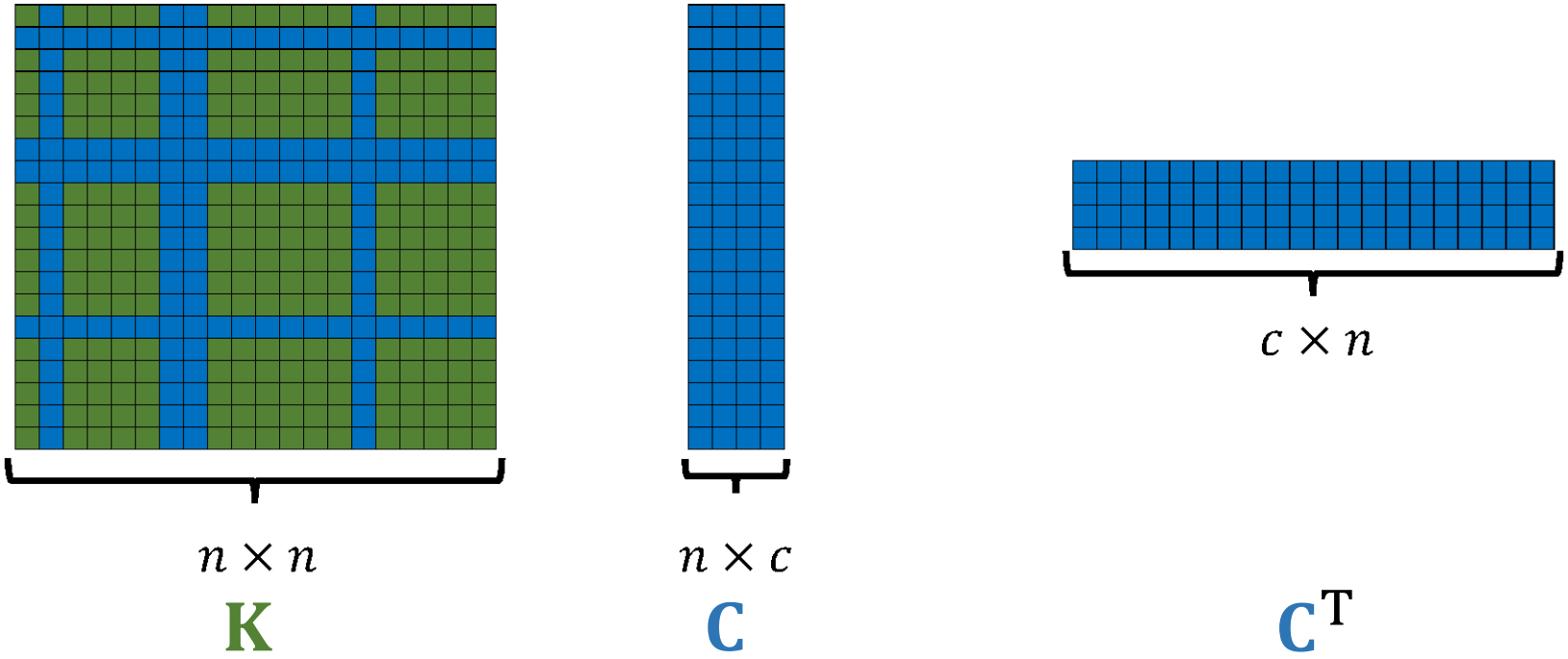


$n \times n$

K

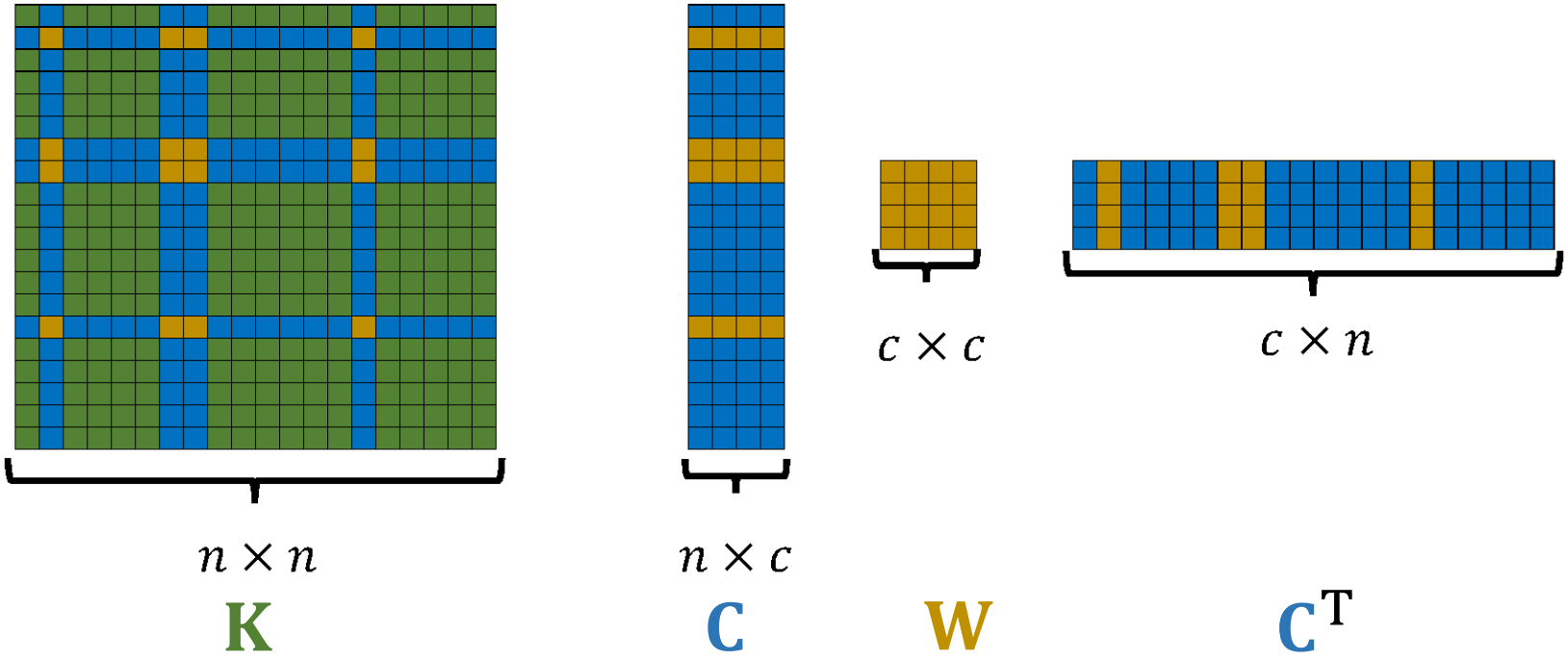
Nyström Method

- $\mathbf{P} \in \mathbb{R}^{n \times c}$: column selection matrix
- $\mathbf{C} = \mathbf{K}\mathbf{P} \in \mathbb{R}^{n \times c}$



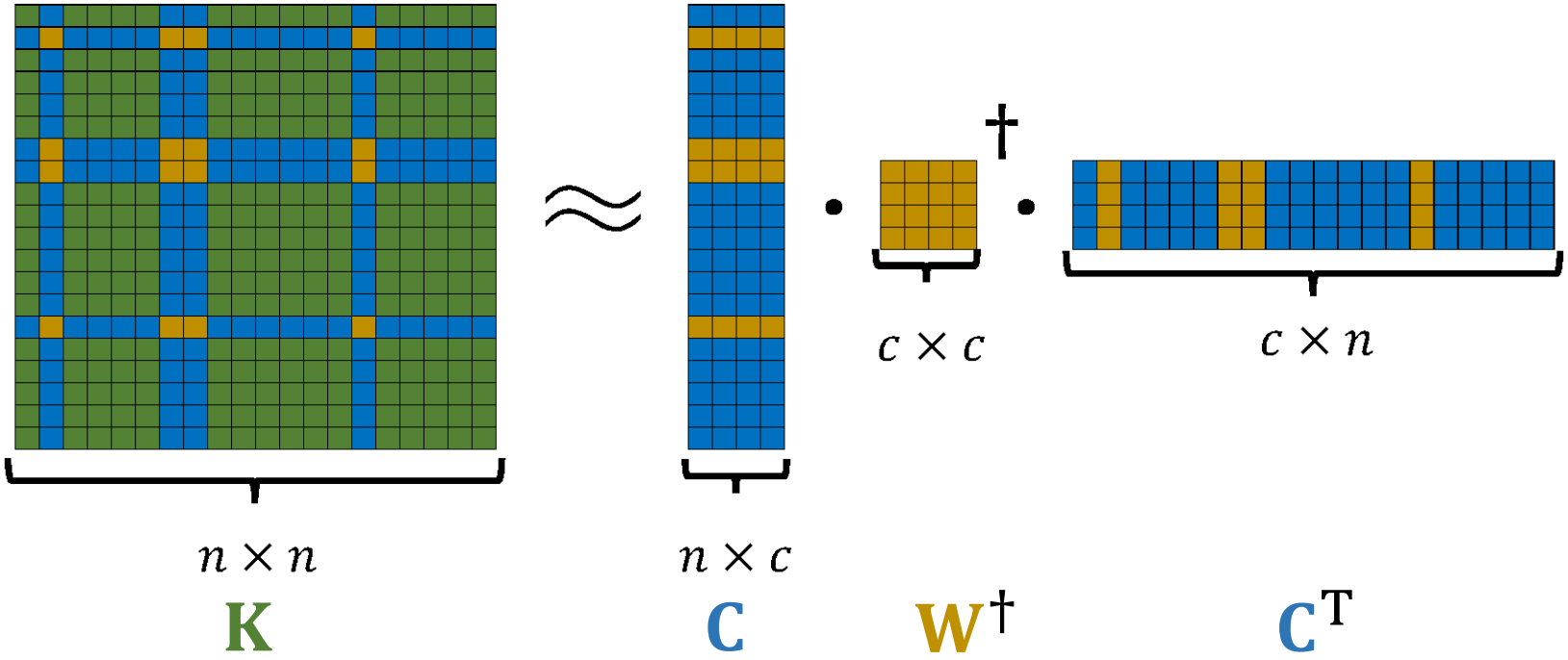
Nyström Method

- $\mathbf{P} \in \mathbb{R}^{n \times c}$: column selection matrix
- $\mathbf{C} = \mathbf{K}\mathbf{P} \in \mathbb{R}^{n \times c}$ and $\mathbf{W} = \mathbf{P}^T\mathbf{K}\mathbf{P} \in \mathbb{R}^{c \times c}$



Nyström Method

- $\mathbf{P} \in \mathbb{R}^{n \times c}$: column selection matrix
- $\mathbf{C} = \mathbf{K}\mathbf{P} \in \mathbb{R}^{n \times c}$ and $\mathbf{W} = \mathbf{P}^T \mathbf{K}\mathbf{P} \in \mathbb{R}^{c \times c}$
- Nyström approximation: $\mathbf{K} \approx \mathbf{C}\,\mathbf{W}^{\dagger}\,\mathbf{C}^{\mathrm{T}}$



Nyström for Kernel K-Means

1. Nyström: $\mathbf{K} \approx \mathbf{C}\mathbf{W}^{\dagger}\mathbf{C}^{T}$ (sketch size is c).
2. Rank s ($< c$) truncated SVD: $\left(\mathbf{C}\mathbf{W}^{\dagger}\mathbf{C}^{\mathrm{T}}\right)_{s} = \widetilde{\mathbf{V}}_{s}\widetilde{\mathbf{\Lambda}}_{s}\widetilde{\mathbf{V}}_{s}^{T}$.
3. Let $\mathbf{B} = \widetilde{\mathbf{V}}_{s}\widetilde{\mathbf{\Lambda}}_{s} \in \mathbb{R}^{n\times s}$ be the feature matrix.
4. Apply γ-approximate k-means algorithm to the rows of $\mathbf{B}$ and output $\mathbf{X}_{\mathbf{B}} \in \mathcal{X}_{n,k}$.

Theorem. (this work)

Let $s = \frac{k}{\epsilon}$ and $c = \tilde{O}\left(\frac{s}{\epsilon}\right)$, then with probability 0.9:

$$\left|\left|\mathbf{K}^{1/2} - \mathbf{X}_{\mathbf{B}}\,\mathbf{X}_{\mathbf{B}}^{T}\mathbf{K}^{1/2}\right|\right|_{F}^{2} \leq (1+2\epsilon)\cdot\gamma\cdot\min_{\mathbf{X}\in\mathcal{X}_{n,k}}\left|\left|\mathbf{K}^{1/2} - \mathbf{X}\,\mathbf{X}^{T}\mathbf{K}^{1/2}\right|\right|_{F}^{2}.$$

Sketch of Proof

This Work: novel bound for Nyström (sketch size $c = \tilde{O}(s/\epsilon)$):

1. Trace norm bound: $\left\| \mathbf{K} - \left(\mathbf{C}\mathbf{W}^{\dagger}\mathbf{C}\right)_s \right\|_* \leq (1+\epsilon) \left\| \mathbf{K} - \mathbf{K}_s \right\|_*$.

2. $\left(\mathbf{C}\mathbf{W}^{\dagger}\mathbf{C}\right)_s = \mathbf{K}^{1/2}\mathbf{M}\mathbf{K}^{1/2}$ for a rank s orthogonal projector $\mathbf{M}$.

Sketch of Proof

This Work: novel bound for Nyström (sketch size $c = \tilde{O}(s/\epsilon)$):

1. Trace norm bound: $\left\| \mathbf{K} - \left(\mathbf{C}\mathbf{W}^{\dagger}\mathbf{C}\right)_s \right\|_* \leq (1+\epsilon) \left\| \mathbf{K} - \mathbf{K}_s \right\|_*$.
2. $\left(\mathbf{C}\mathbf{W}^{\dagger}\mathbf{C}\right)_s = \mathbf{K}^{1/2}\mathbf{M}\mathbf{K}^{1/2}$ for a rank s orthogonal projector $\mathbf{M}$.

Previous Work (Gittens & Mahoney, 2013)

Trace norm bound: $\left\| \mathbf{K} - \mathbf{C}\mathbf{W}^{\dagger}\mathbf{C} \right\|_* \leq (1+\epsilon) \left\| \mathbf{K} - \mathbf{K}_s \right\|_*$.

Sketch of Proof

This Work: novel bound for Nyström (sketch size $c = \tilde{O}(s/\epsilon)$):

1. Trace norm bound: $\left|\left|\mathbf{K} - \mathbf{B}\mathbf{B}^{\mathbf{T}}\right|\right|_* \leq (1+\epsilon)\left|\left|\mathbf{K} - \mathbf{K}_s\right|\right|_*$.
2. $\mathbf{B}\mathbf{B}^{\mathbf{T}} = \mathbf{K}^{1/2}\mathbf{M}\mathbf{K}^{1/2}$ for a rank s orthogonal projector $\mathbf{M}$.

Sketch of Proof

This Work: novel bound for Nyström (sketch size $c = \tilde{O}(s/\epsilon)$):

1. Trace norm bound: $\left|\left|\mathbf{K} - \mathbf{B}\mathbf{B}^{\mathbf{T}}\right|\right|_* \leq (1+\epsilon)\left|\left|\mathbf{K} - \mathbf{K}_s\right|\right|_*$.
2. $\mathbf{B}\mathbf{B}^{\mathbf{T}} = \mathbf{K}^{1/2}\mathbf{M}\mathbf{K}^{1/2}$ for a rank s orthogonal projector $\mathbf{M}$.

Cohen et al. 2015: Projection-Cost Preservation

- $\mathbf{\Pi}$ ($n \times n$, rank k): any fixed orthogonal projector.
- There exists a positive α (independent of $\mathbf{\Pi}$) such that

$$\left|\left|\mathbf{K}^{1/2} - \mathbf{\Pi}\mathbf{K}^{1/2}\right|\right|_F^2 \leq \left|\left|\mathbf{B} - \mathbf{\Pi}\mathbf{B}\right|\right|_F^2 + \alpha \leq \left(1 + \epsilon + \frac{k}{s}\right)\left|\left|\mathbf{K}^{1/2} - \mathbf{\Pi}\mathbf{K}^{1/2}\right|\right|_F^2.$$

Sketch of Proof

Cohen et al. 2015: Projection-Cost Preservation

$$\left|\left|\mathbf{K}^{1/2} - \mathbf{\Pi}\mathbf{K}^{1/2}\right|\right|_F^2 \leq \left|\left|\mathbf{B} - \mathbf{\Pi}\mathbf{B}\right|\right|_F^2 + \alpha \leq \left(1 + \epsilon + \frac{k}{s}\right)\left|\left|\mathbf{K}^{1/2} - \mathbf{\Pi}\mathbf{K}^{1/2}\right|\right|_F^2.$$

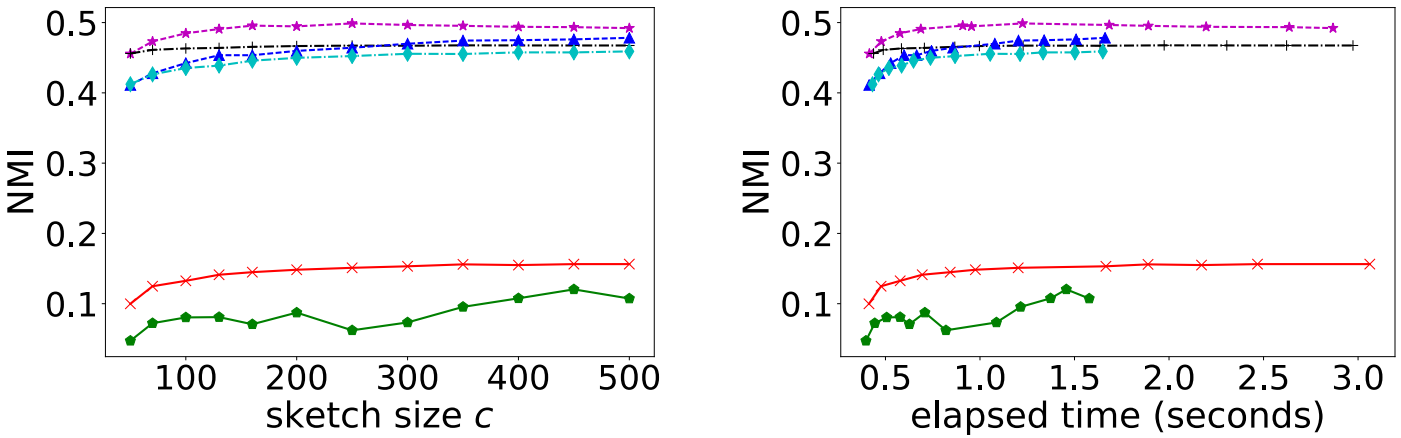
Cohen et al. 2015: Projection-Cost Preservation → K-Means

- Apply γ-approximate k-means algorithm to the rows of $\mathbf{B}$ and outputs $\mathbf{X_B} \in \mathcal{X}_{n,k}$.
- It holds that

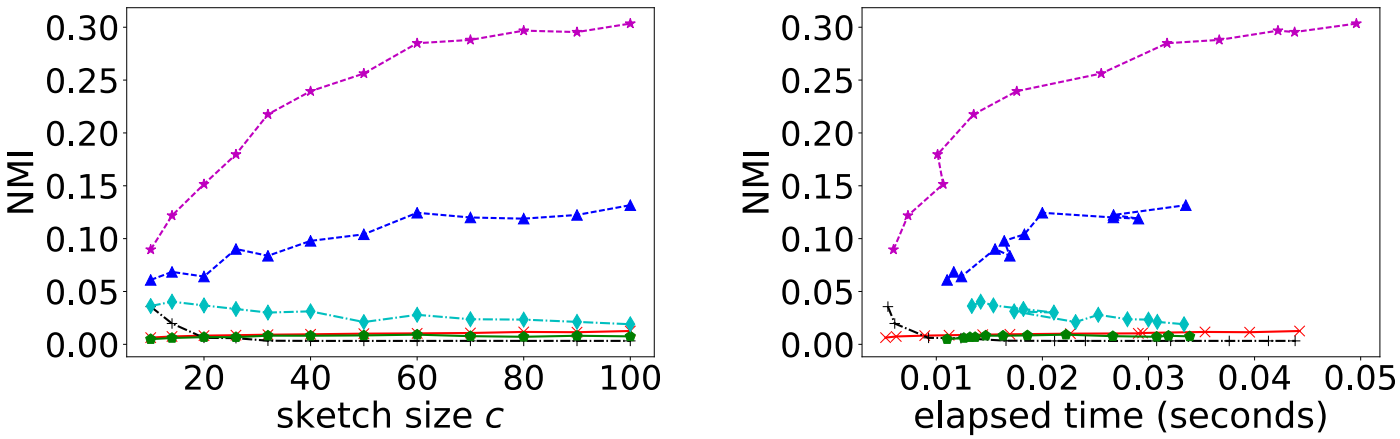
$$\left|\left|\mathbf{K}^{1/2} - \mathbf{X_B}\,\mathbf{X_B}^T\mathbf{K}^{1/2}\right|\right|_F^2 \leq \left(1 + \epsilon + \frac{k}{s}\right) \cdot \gamma \cdot \min_{\mathbf{X} \in \mathcal{X}_{n,k}} \left|\left|\mathbf{K}^{1/2} - \mathbf{X}\,\mathbf{X}^T\mathbf{K}^{1/2}\right|\right|_F^2.$$

Comparison to Spectral Clustering

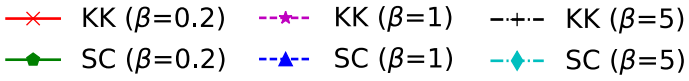
- both clustering methods use the Gaussian kernel with bandwidth β
- Evaluated using normalized mutual information with true clustering (higher is better)



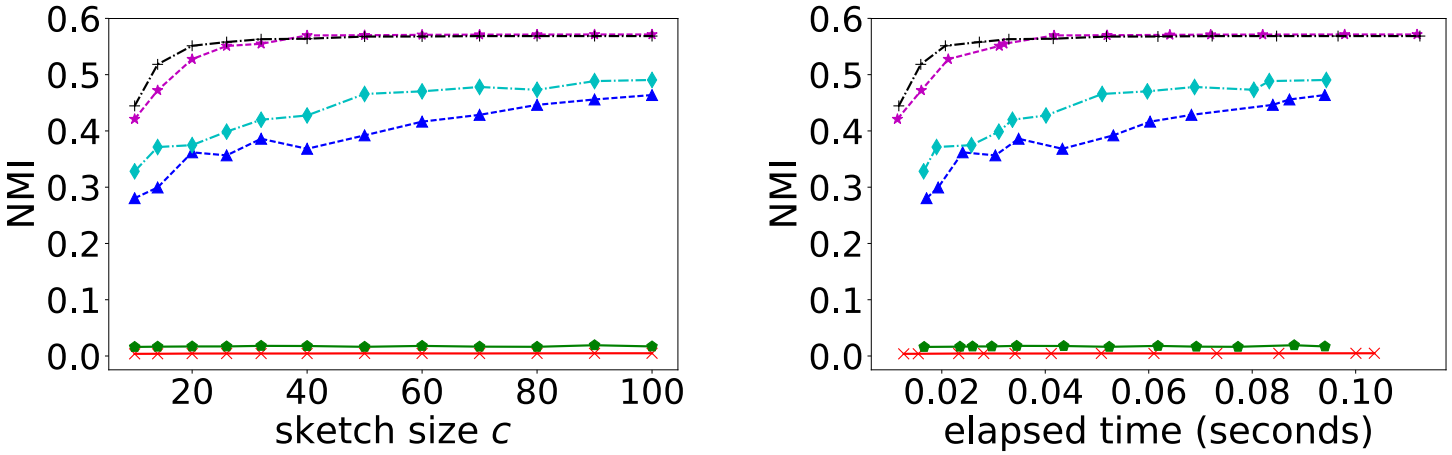
(a) MNIST ($n = 60,000$, $d = 780$, $k = 10$).



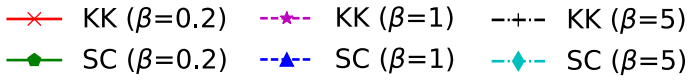
(b) Phishing ($n = 11,055$, $d = 68$, $k = 2$).

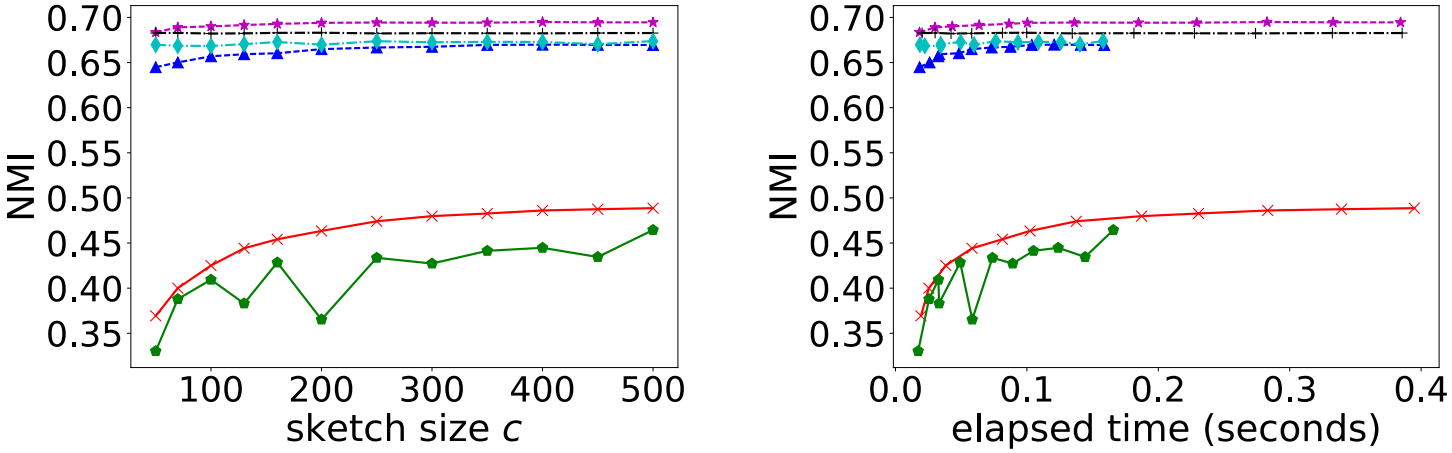


Comparison to Spectral Clustering



(c) Mushroom ($n = 8,124$, $d = 112$, $k = 2$).





(d) PenDigits ($n = 7,494$, $d = 16$, $k = 10$).

Thank You!

For more details see the preprint "*Scalable Kernel K-Means Clustering with Nystrom Approximation: Relative-Error Bounds*" on arXiv (accepted to JMLR)

Questions? *gittea@rpi.edu*