A Random Matrix Analysis of Data Stream Clustering: Coping With Limited Memory Resources

Hugo Lebeau [1] Romain Couillet [1] Florent Chatelain [2]

[1]Univ. Grenoble Alpes, CNRS, Inria, Grenoble INP, LIG, 38000 Grenoble, France

[2]Univ. Grenoble Alpes, CNRS, Grenoble INP, GIPSA-lab, 38000 Grenoble, France

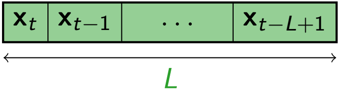
39th International Conference on Machine Learning
July 17th – 23rd, 2022
Baltimore, Maryland, USA

Random matrix framework for data stream clustering

- Observed data: $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_t, \ldots \in \mathbb{R}^p$ $\qquad \mathbf{x}_t \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\pm\boldsymbol{\mu}, \mathbf{I}_p)$

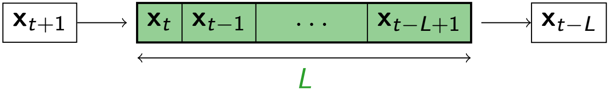
Random matrix framework for data stream clustering

- Observed data: $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_t, \ldots \in \mathbb{R}^p$ $\quad\quad \mathbf{x}_t \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\pm\boldsymbol{\mu}, \mathbf{I}_p)$
- Memory



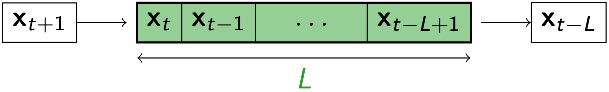
Random matrix framework for data stream clustering

- Observed data: $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_t, \ldots \in \mathbb{R}^p$ $\qquad \mathbf{x}_t \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\pm\boldsymbol{\mu}, \mathbf{I}_p)$
- Memory



Random matrix framework for data stream clustering

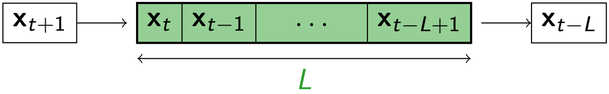
- Observed data: $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_t, \ldots \in \mathbb{R}^p$ $\qquad \mathbf{x}_t \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\pm\boldsymbol{\mu}, \mathbf{I}_p)$
- Memory



- **Clustering?**

Random matrix framework for data stream clustering

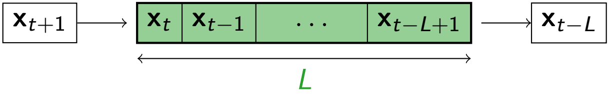
- Observed data: $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_t, \ldots \in \mathbb{R}^p$ $\qquad$ $\mathbf{x}_t \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\pm\boldsymbol{\mu}, \mathbf{I}_p)$
- Memory



- **Clustering** on the $n \geqslant L$ previous points
  $\mathbf{X} = \begin{bmatrix} \mathbf{x}_{t-n+1} & \mathbf{x}_{t-n+2} & \ldots & \mathbf{x}_t \end{bmatrix} \in \mathbb{R}^{p \times n}$

Random matrix framework for data stream clustering

- Observed data: $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_t, \ldots \in \mathbb{R}^p$ $\qquad \mathbf{x}_t \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\pm\boldsymbol{\mu}, \mathbf{I}_p)$
- Memory

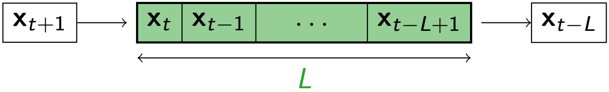


- **Clustering** on the $n \geqslant L$ previous points
  $\mathbf{X} = \left[\mathbf{x}_{t-n+1} \quad \mathbf{x}_{t-n+2} \quad \ldots \quad \mathbf{x}_t\right] \in \mathbb{R}^{p \times n}$

$$\begin{aligned} n, p, L &\to +\infty \\ p/n &\to c \\ (2L-1)/n &\to \varepsilon \end{aligned}$$

Random matrix framework for data stream clustering

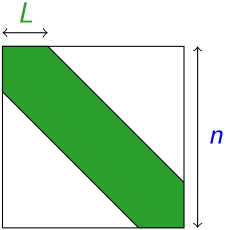
- Observed data: $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_t, \ldots \in \mathbb{R}^p$ $\qquad \mathbf{x}_t \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\pm\boldsymbol{\mu}, \mathbf{I}_p)$
- Memory



- **Clustering** on the $n \geqslant L$ previous points
  $\mathbf{X} = \begin{bmatrix} \mathbf{x}_{t-n+1} & \mathbf{x}_{t-n+2} & \ldots & \mathbf{x}_t \end{bmatrix} \in \mathbb{R}^{p \times n}$

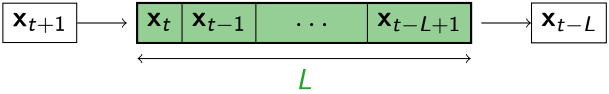
$$\begin{aligned} n, p, L &\to +\infty \\ p/n &\to c \\ (2L-1)/n &\to \varepsilon \end{aligned}$$

- Kernel matrix



Random matrix framework for data stream clustering

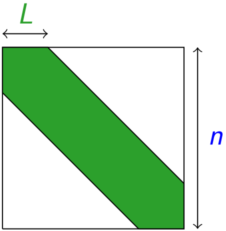
- Observed data: $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_t, \ldots \in \mathbb{R}^p$ $\qquad \mathbf{x}_t \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\pm\boldsymbol{\mu}, \mathbf{I}_p)$
- Memory



- **Clustering** on the $n \geqslant L$ previous points
  $\mathbf{X} = \left[\mathbf{x}_{t-n+1} \quad \mathbf{x}_{t-n+2} \quad \ldots \quad \mathbf{x}_t\right] \in \mathbb{R}^{p \times n}$

$$\begin{aligned} n, p, L &\to +\infty \\ p/n &\to c \\ (2L-1)/n &\to \varepsilon \end{aligned}$$

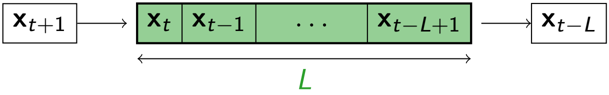
- Kernel matrix



$$\mathbf{K} = \frac{\mathbf{X}^\top \mathbf{X}}{p}$$

Random matrix framework for data stream clustering

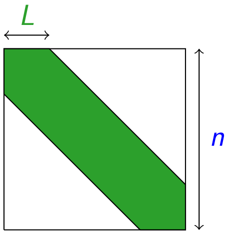
- Observed data: $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_t, \ldots \in \mathbb{R}^p$ $\qquad \mathbf{x}_t \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\pm\boldsymbol{\mu}, \mathbf{I}_p)$
- Memory



- **Clustering** on the $n \geqslant L$ previous points
  $\mathbf{X} = \begin{bmatrix} \mathbf{x}_{t-n+1} & \mathbf{x}_{t-n+2} & \ldots & \mathbf{x}_t \end{bmatrix} \in \mathbb{R}^{p \times n}$

$$\begin{aligned} n, p, L &\to +\infty \\ p/n &\to c \\ (2L-1)/n &\to \varepsilon \end{aligned}$$

- Kernel matrix



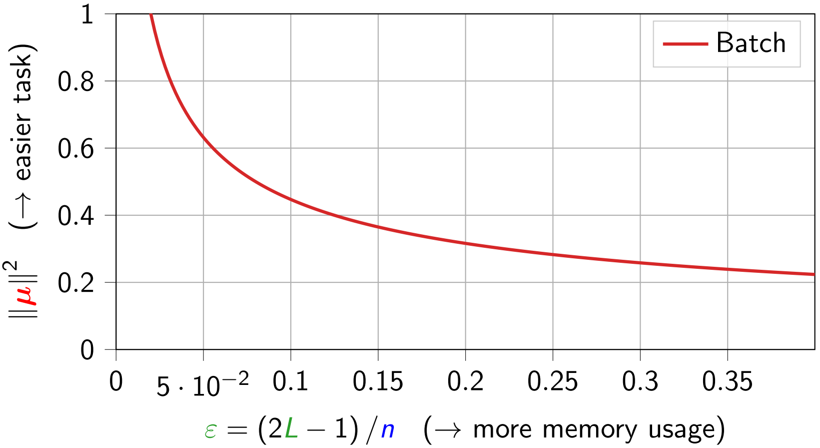
$$\mathbf{K}_L = \frac{\mathbf{X}^\top \mathbf{X}}{p} \odot \underbrace{\begin{bmatrix} 1 & \ldots & 1 & & 0 \\ \vdots & \ddots & & \ddots & \\ 1 & & \ddots & & 1 \\ & \ddots & & \ddots & \vdots \\ 0 & & 1 & \ldots & 1 \end{bmatrix}}_{\mathbf{T}}$$

Data stream clustering: improving over batch clustering

- Size-L memory $\rightsquigarrow$ *batch clustering* vs. $\mathbf{K}_L = \frac{1}{p}\mathbf{X}^\top\mathbf{X} \odot \mathbf{T}$

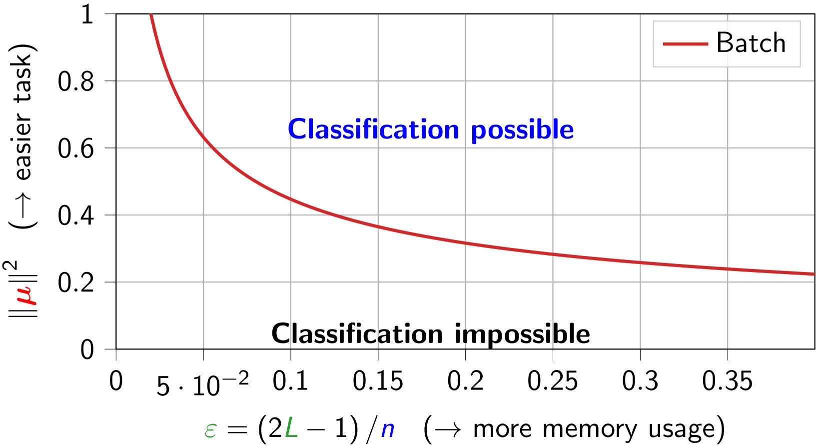
Data stream clustering: improving over batch clustering

- Size-L memory $\rightsquigarrow$ *batch clustering* vs. $\mathbf{K}_L = \frac{1}{p}\mathbf{X}^\top\mathbf{X} \odot \mathbf{T}$
- **Spectral clustering phase transition** ($n/p = 100 \iff c = 0.01$)



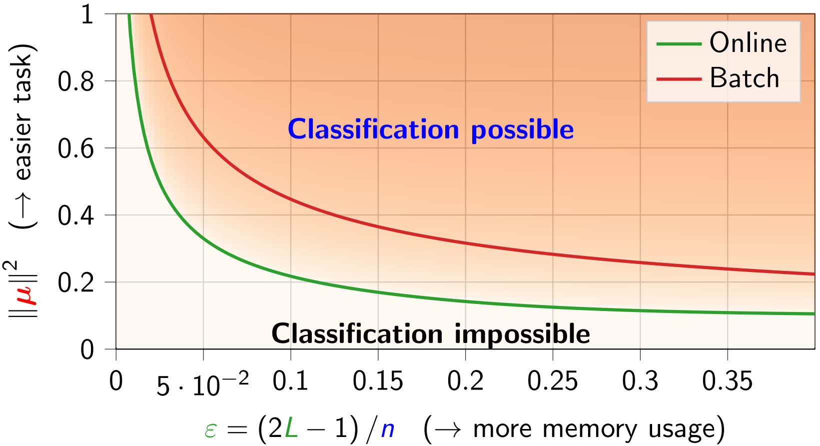
Data stream clustering: improving over batch clustering

- Size-L memory $\rightsquigarrow$ *batch clustering* vs. $\mathbf{K}_L = \frac{1}{p}\mathbf{X}^\top\mathbf{X} \odot \mathbf{T}$
- **Spectral clustering phase transition** ($n/p = 100 \iff c = 0.01$)



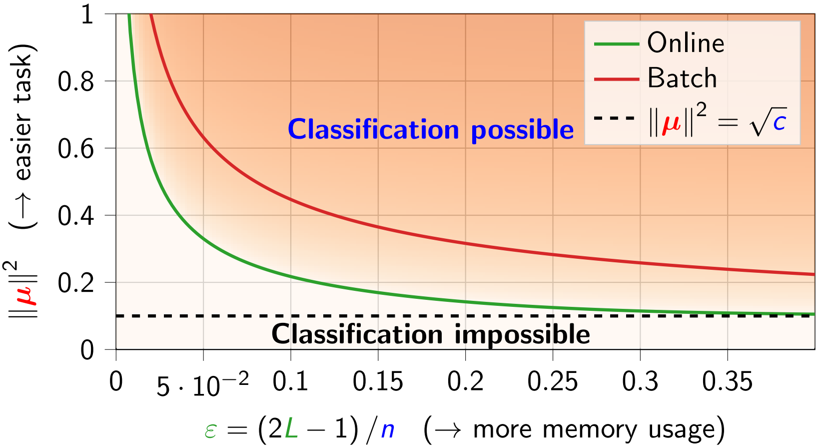
Data stream clustering: improving over batch clustering

- Size-L memory $\rightsquigarrow$ *batch clustering* vs. $\mathbf{K}_L = \frac{1}{p}\mathbf{X}^\top\mathbf{X} \odot \mathbf{T}$
- **Spectral clustering phase transition** $(n/p = 100 \iff c = 0.01)$



Data stream clustering: improving over batch clustering

- Size-L memory $\rightsquigarrow$ *batch clustering* vs. $\mathbf{K}_L = \frac{1}{p}\mathbf{X}^\top\mathbf{X} \odot \mathbf{T}$
- **Spectral clustering phase transition** $(n/p = 100 \iff c = 0.01)$

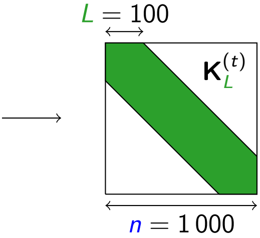


From theory to practice: online kernel spectral clustering

- **BigGAN images** (VGG features, $p = 4\,096$) $T = 20\,000$

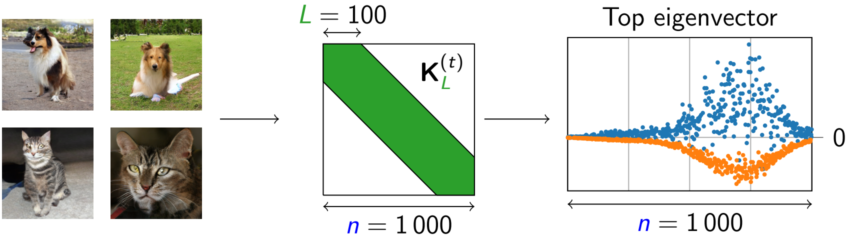
From theory to practice: online kernel spectral clustering

- **BigGAN images** (VGG features, $p = 4\,096$) $T = 20\,000$



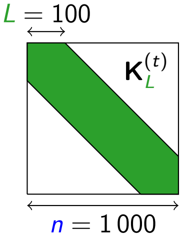
From theory to practice: online kernel spectral clustering

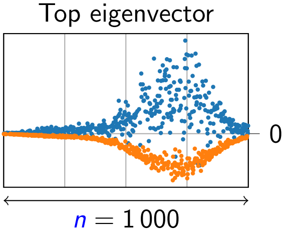
- **BigGAN images** (VGG features, $p = 4\,096$) $T = 20\,000$

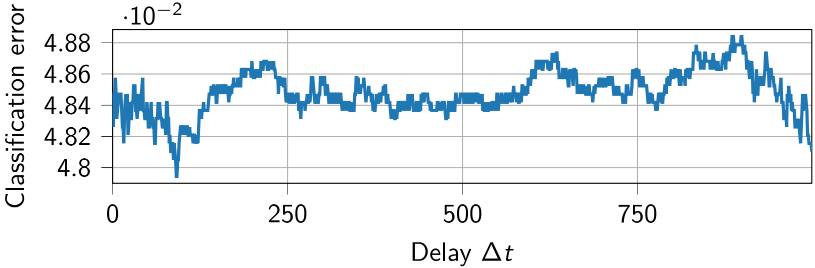


From theory to practice: online kernel spectral clustering

- **BigGAN images** (VGG features, $p = 4\,096$) $T = 20\,000$

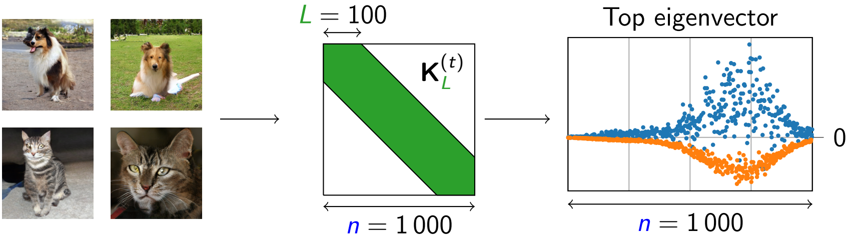


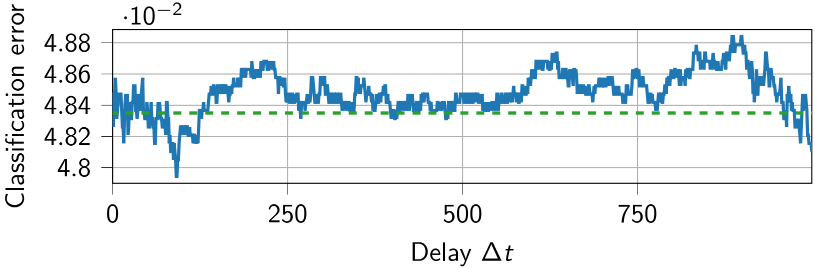




From theory to practice: online kernel spectral clustering

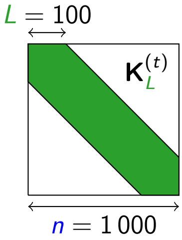
- **BigGAN images** (VGG features, $p = 4\,096$) $T = 20\,000$

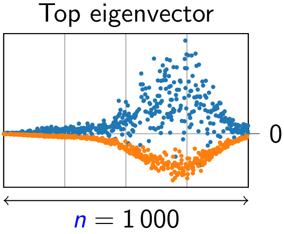


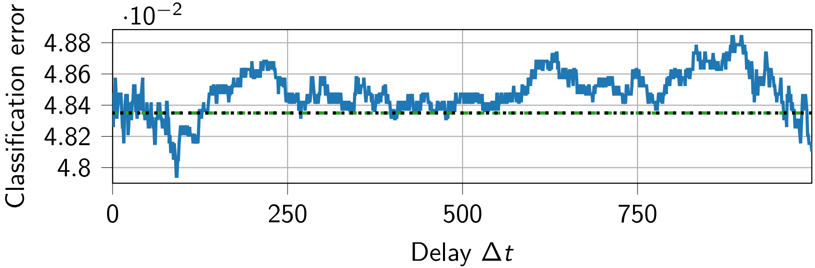


From theory to practice: online kernel spectral clustering

- **BigGAN images** (VGG features, $p = 4\,096$) $T = 20\,000$

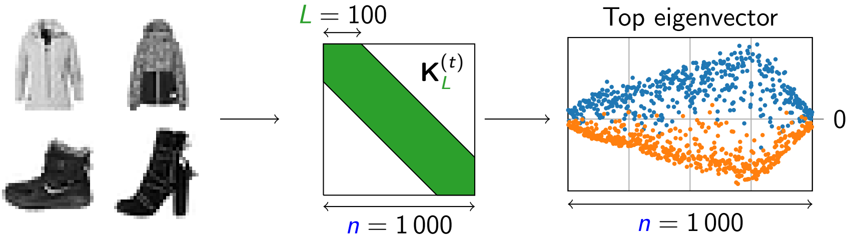






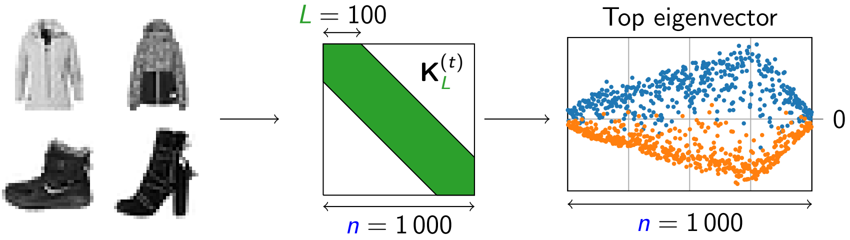
From theory to practice: online kernel spectral clustering

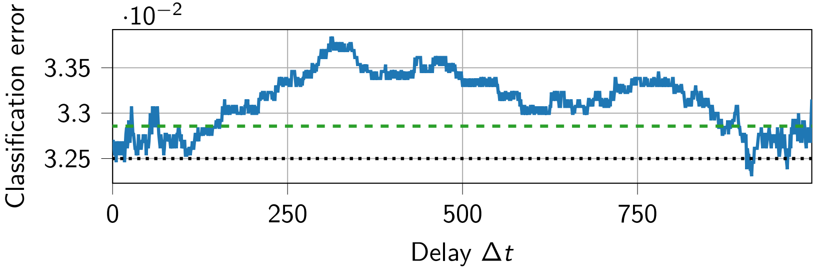
- **Fashion-MNIST images** (raw, $p = 784$) $T = 14\,000$



From theory to practice: online kernel spectral clustering

- **Fashion-MNIST images** (raw, $p = 784$) $T = 14\,000$





A Random Matrix Analysis of Data Stream Clustering: Coping With Limited Memory Resources

Hugo Lebeau [1] Romain Couillet [1] Florent Chatelain [2]

[1]Univ. Grenoble Alpes, CNRS, Inria, Grenoble INP, LIG, 38000 Grenoble, France

[2]Univ. Grenoble Alpes, CNRS, Grenoble INP, GIPSA-lab, 38000 Grenoble, France

39th International Conference on Machine Learning
July 17th – 23rd, 2022
Baltimore, Maryland, USA