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

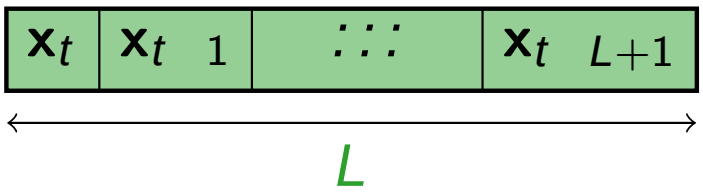
39th International Conference on Machine Learning
July 17th – 23rd, 2022
Baltimore, Maryland, USA

Random matrix framework for data stream clustering

- Observed data: $\mathbf{x}_1; \mathbf{x}_2; \dots; \mathbf{x}_t; \dots \in \mathbb{R}^p$ $\quad \mathbf{x}_t \overset{\text{i.i.d.}}{\sim} N(\quad ; \mathbf{I}_p)$

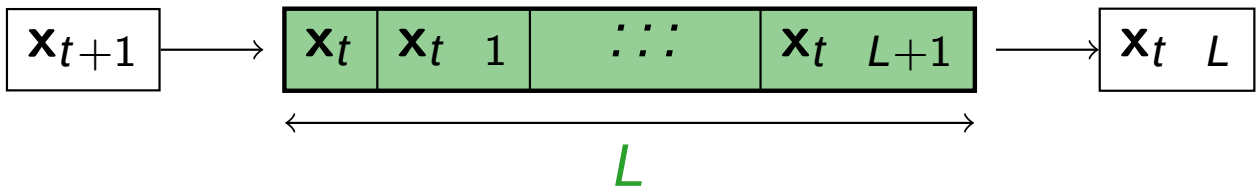
Random matrix framework for data stream clustering

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



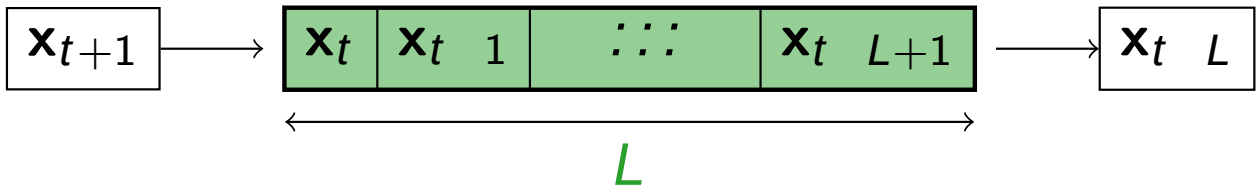
Random matrix framework for data stream clustering

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



Random matrix framework for data stream clustering

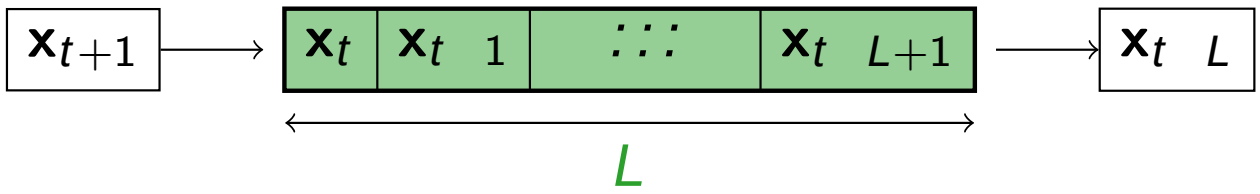
- Observed data: $\mathbf{x}_1; \mathbf{x}_2; \dots; \mathbf{x}_t; \dots \in \mathbb{R}^p$ $\quad \mathbf{x}_t \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\quad ; \mathbf{I}_p)$
- Memory



- **Clustering?**

Random matrix framework for data stream clustering

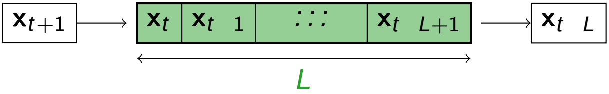
- Observed data: $\mathbf{x}_1; \mathbf{x}_2; \ldots; \mathbf{x}_t; \ldots \in \mathbb{R}^p$ $\quad \mathbf{x}_t \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\quad; \mathbf{I}_p)$
- Memory



- **Clustering** on the $n > L$ previous points
  $\mathbf{X} = \mathbf{x}_{t\ n+1}\ \mathbf{x}_{t\ n+2}\ \ldots\ \mathbf{x}_{t\ 2} \in \mathbb{R}^{p\ n}$

Random matrix framework for data stream clustering

- Observed data: $\mathbf{x}_1; \mathbf{x}_2; \ldots; \mathbf{x}_t; \ldots\ 2\ \mathbb{R}^p$ $\quad$ $\mathbf{x}_t \ \overset{\text{i.i.d.}}{}\ N(\quad; \mathbf{I}_p)$
- Memory



$n; p; L \quad ! \quad +1$

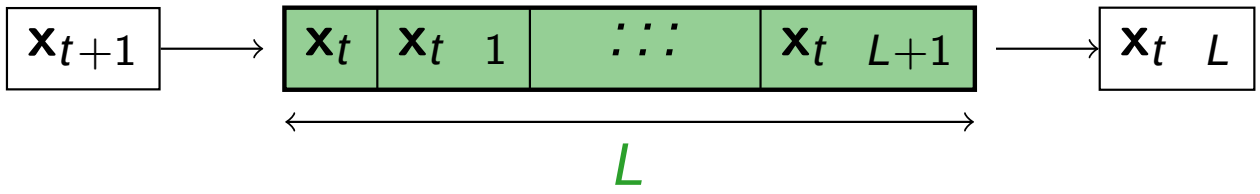
$p{=}n \quad ! \quad c$

$(2L \quad 1){=}n \quad ! \quad "$

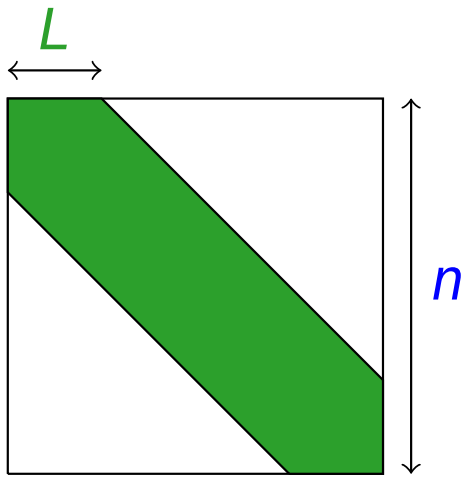
- **Clustering** on the $n > L$ previous points
  $\mathbf{X} = \quad \mathbf{x}_{t \quad n+1} \quad \mathbf{x}_{t \quad n+2} \quad \ldots \quad \mathbf{x}_t \quad 2\ \mathbb{R}^{p \quad n}$

Random matrix framework for data stream clustering

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

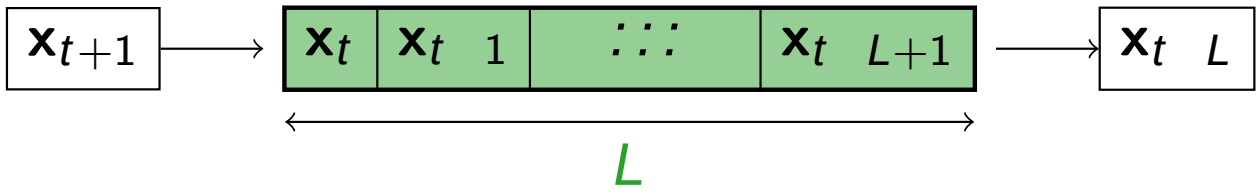


- **Clustering** on the $n > L$ previous points $\quad n; p; L \quad ! \quad +1$
  $p=n \quad ! \quad c$
  $\mathbf{X} = \mathbf{x}_{t\ n+1}\ \mathbf{x}_{t\ n+2}\ \dots\ \mathbf{x}_t \in \mathbb{R}^{p\ n} \quad (2L \quad 1)=n \quad ! \quad "$
- Kernel matrix

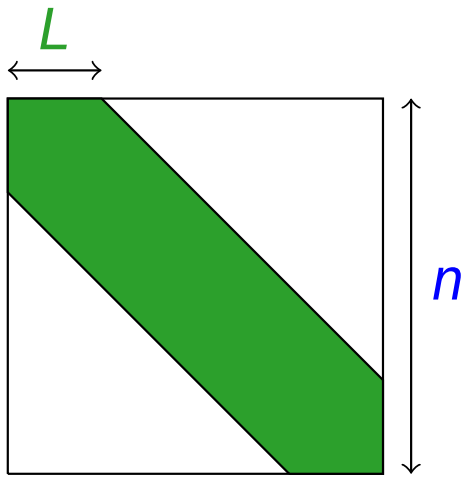


Random matrix framework for data stream clustering

- Observed data: $\mathbf{x}_1; \mathbf{x}_2; :::; \mathbf{x}_t; ::: \; 2 \; \mathrm{R}^p$ $\quad \mathbf{x}_t \overset{\text{i.i.d.}}{} N(\quad ; \mathbf{I}_p)$
- Memory



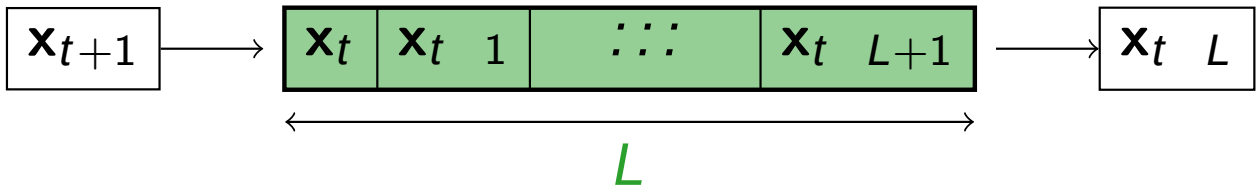
- **Clustering** on the $n > L$ previous points $\quad n; p; L \quad ! \quad +1$
  $p=n \quad ! \quad c$
  $\mathbf{X} = \mathbf{x}_t \;\; _{n+1} \quad \mathbf{x}_t \;\; _{n+2} \quad ::: \quad \mathbf{x}_t \;\; 2 \; \mathrm{R}^{p \;\; n} \quad (2L \quad 1)=n \quad ! \quad "$
- Kernel matrix



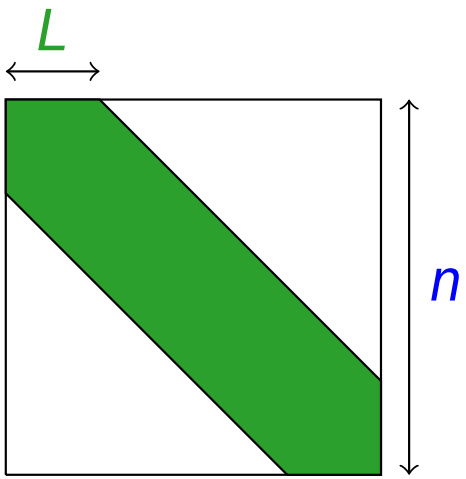
$$\mathbf{K} = \frac{\mathbf{X}^{>}\mathbf{X}}{p}$$

Random matrix framework for data stream clustering

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



- **Clustering** on the $n > L$ previous points $\quad n; p; L \quad ! \quad +1$
  $\mathbf{X} = \mathbf{x}_{t\ n+1}\ \mathbf{x}_{t\ n+2} \ldots \mathbf{x}_{t\ 2} \in \mathbb{R}^{p\ n}$ $\quad p{=}n \quad ! \quad c$
  $\quad (2L \quad 1){=}n \quad ! \quad "$
- Kernel matrix



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

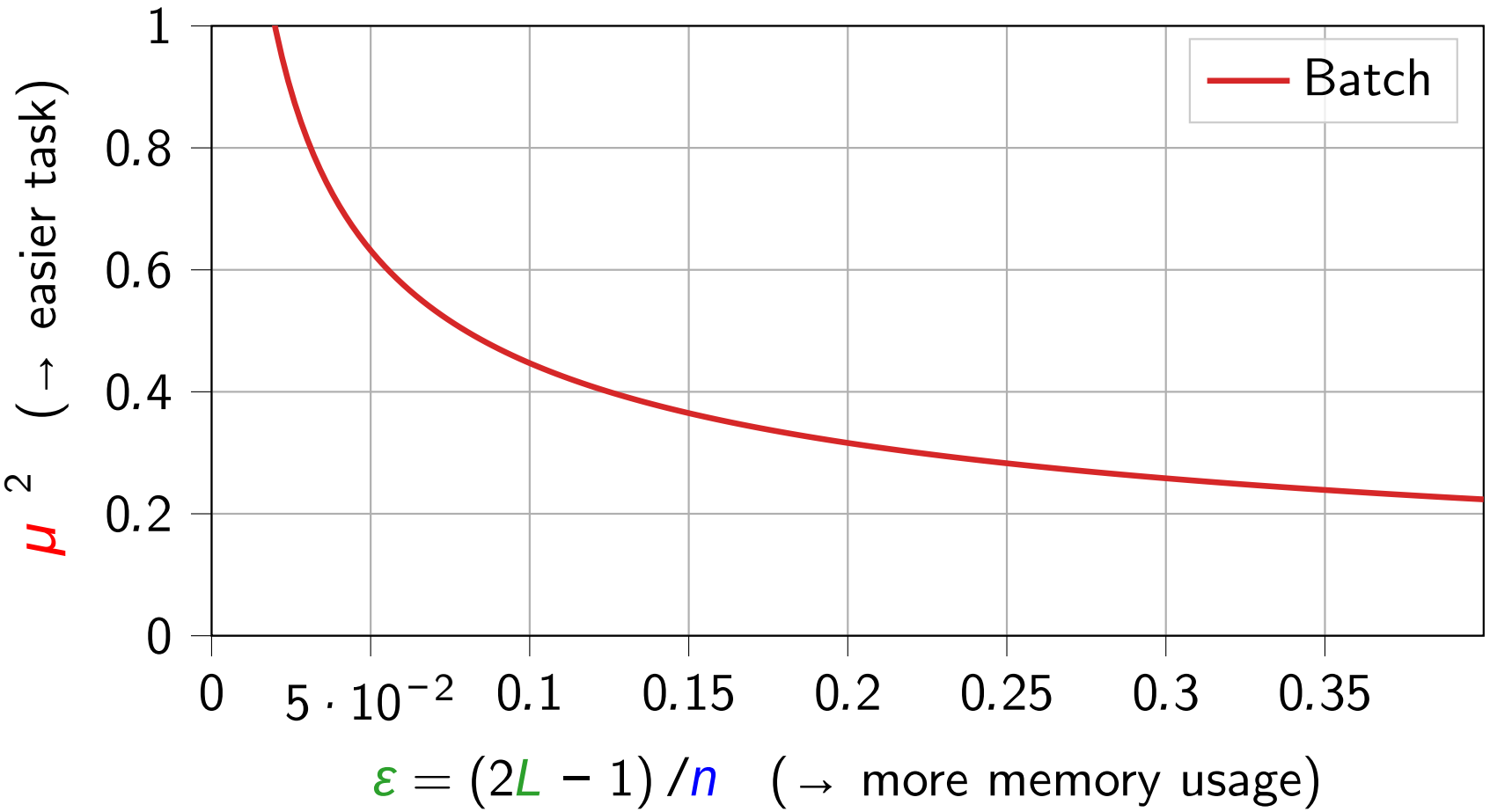
$$\begin{bmatrix} 2 & 1 & \cdots & 1 & 0 & 3 \\ 6 & \ddots & & & \ddots & 7 \\ 6 & & & & & 7 \\ 6 & & & & & 7 \\ 4 & 1 & & & 1 & 5 \\ & \ddots & & & \ddots & \vdots \\ | & \underbrace{0 \qquad 1 \quad \cdots \quad 1}_{T} & & & & \} \end{bmatrix}$$

Data stream clustering: improving over batch clustering

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

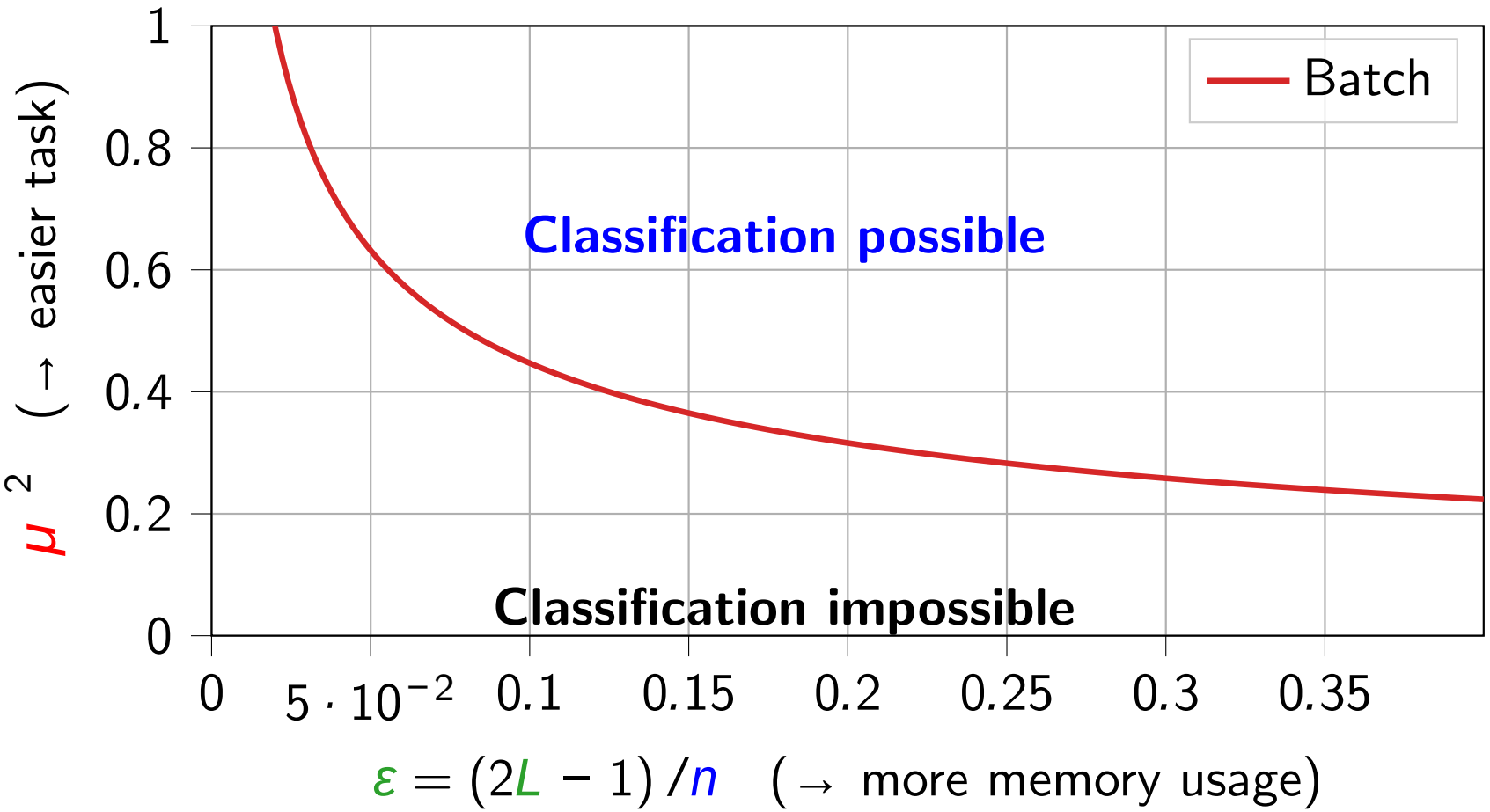
Data stream clustering: improving over batch clustering

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



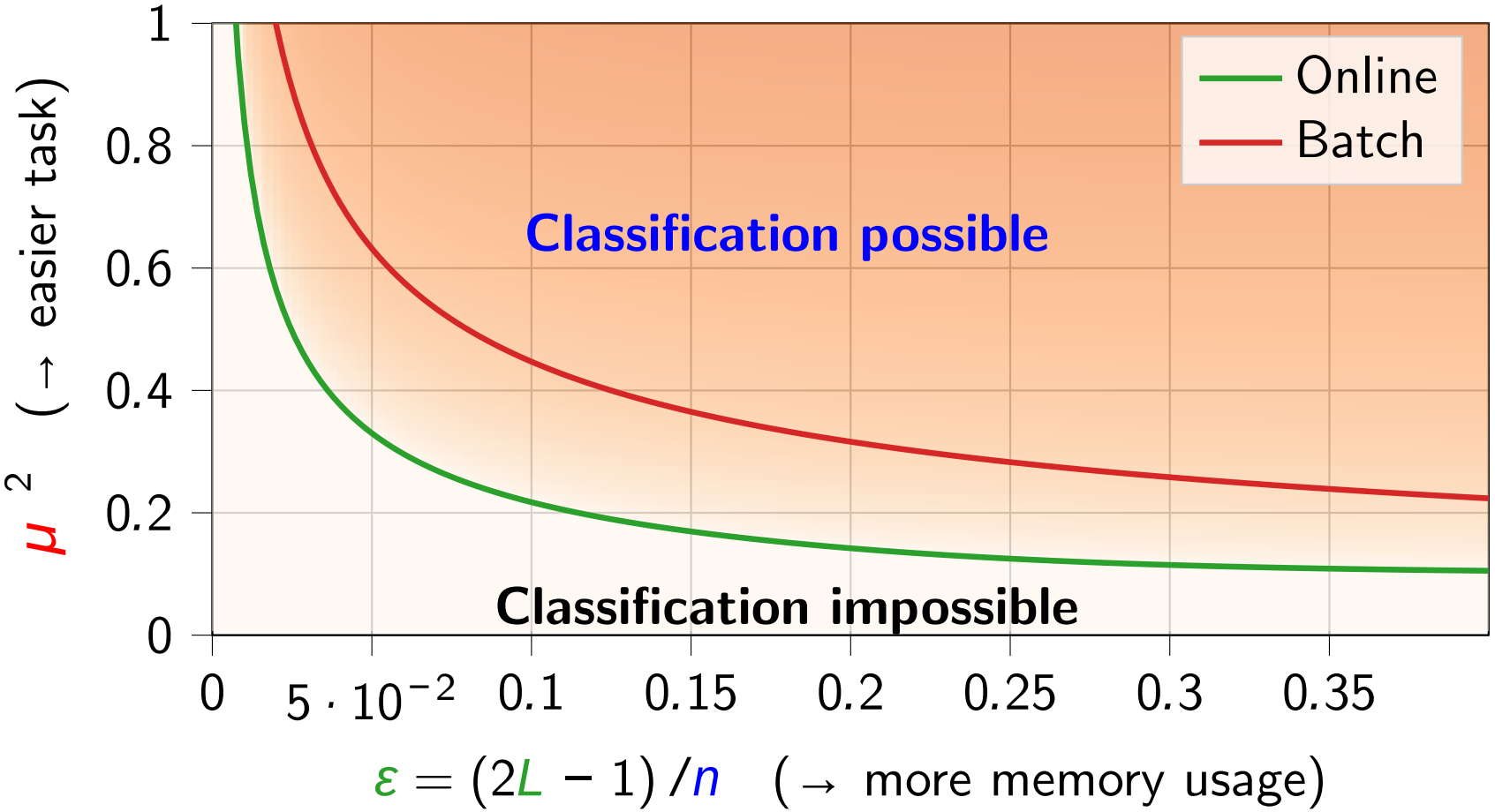
Data stream clustering: improving over batch clustering

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



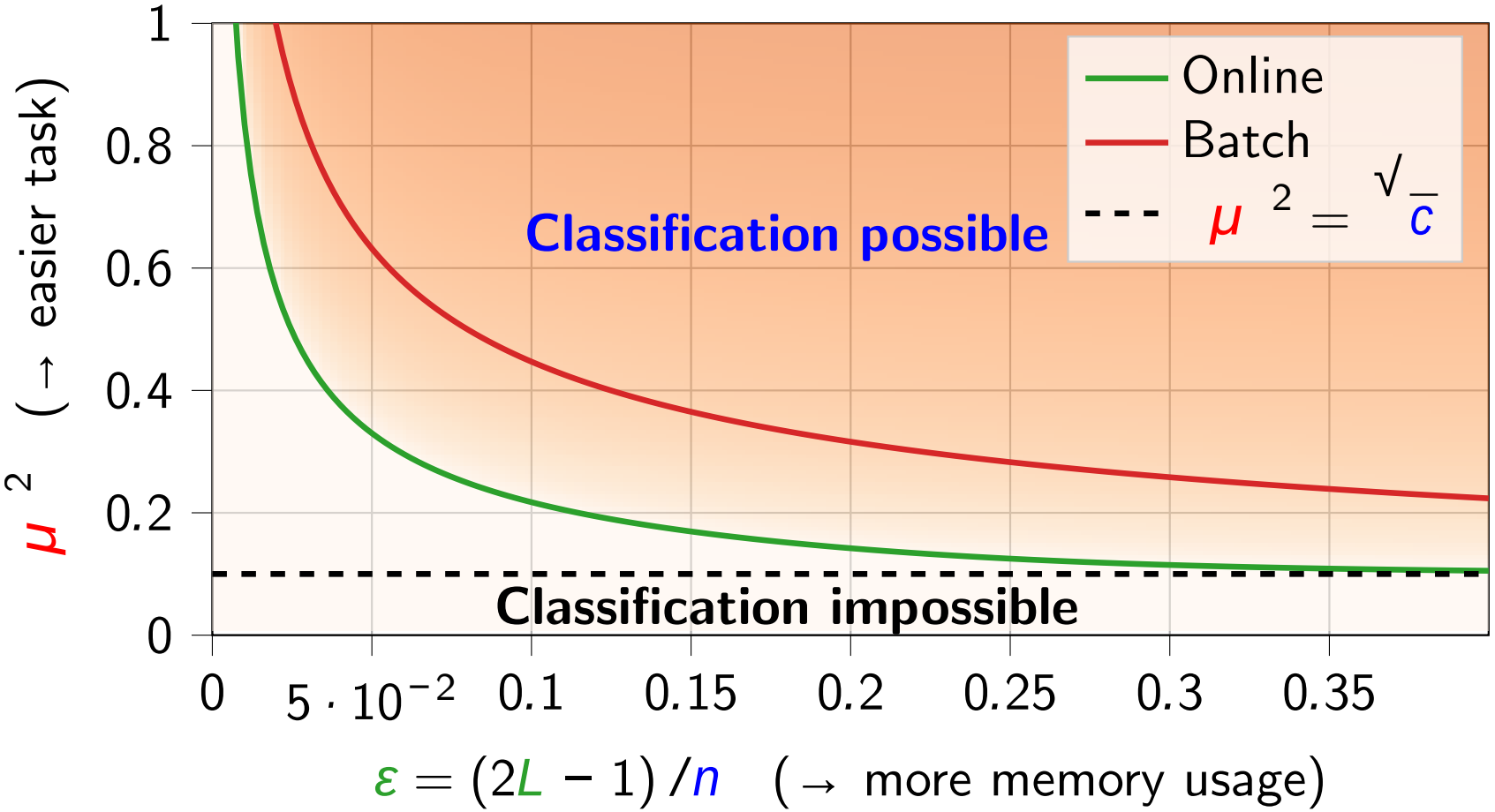
Data stream clustering: improving over batch clustering

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



Data stream clustering: improving over batch clustering

- Size-L memory *batch clustering* vs. $\mathbf{K}_L = \frac{1}{p}\mathbf{X}^\top\mathbf{X}$ T
- **Spectral clustering phase transition** (n=$p = 100$ ( ) $c = 0{:}01$)

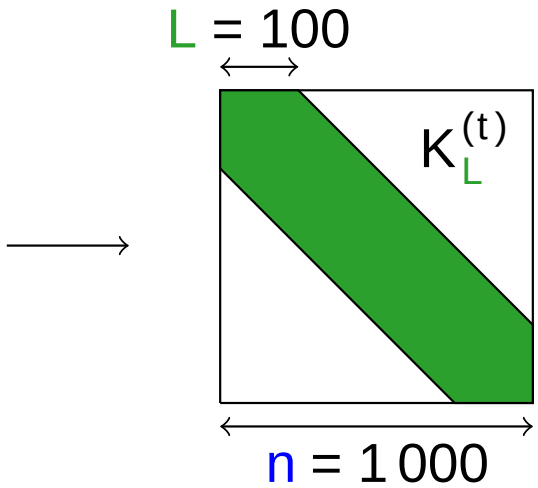


From theory to practice: online kernel spectral clustering

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

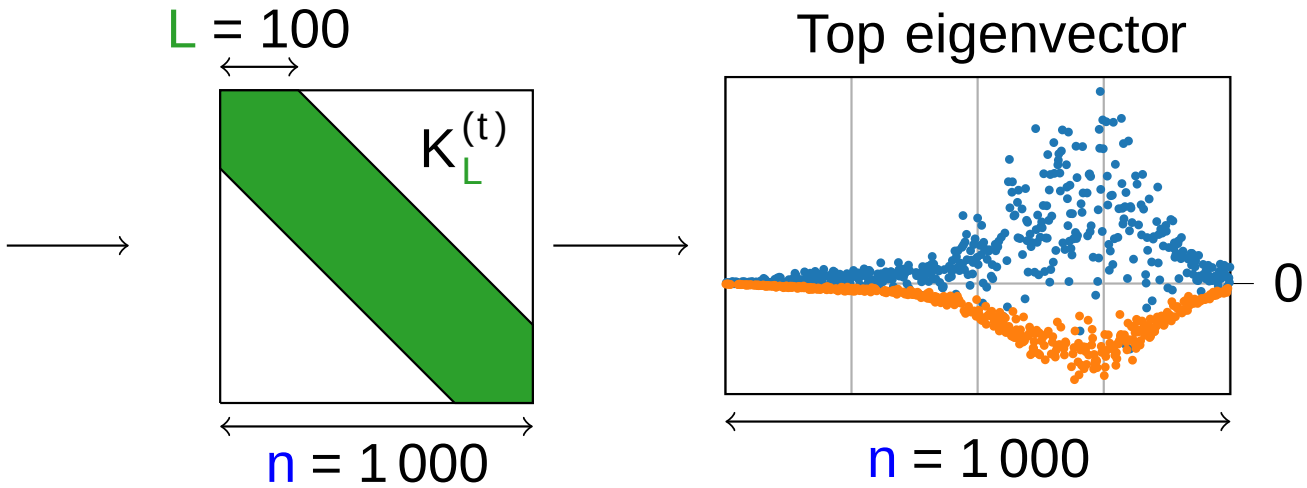
From theory to practice: online kernel spectral clusterin

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



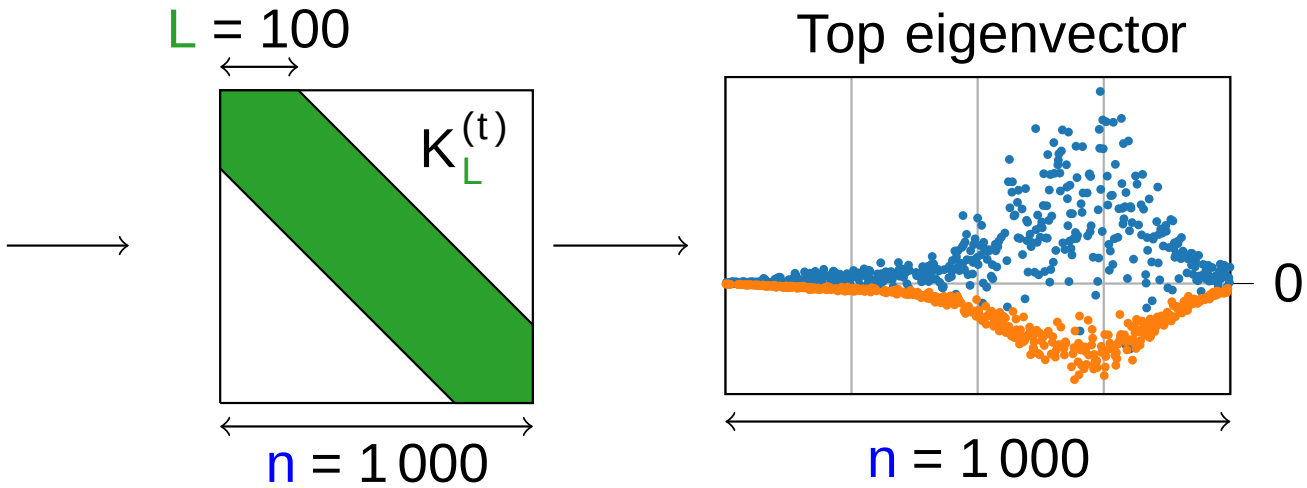
From theory to practice: online kernel spectral clusterin

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



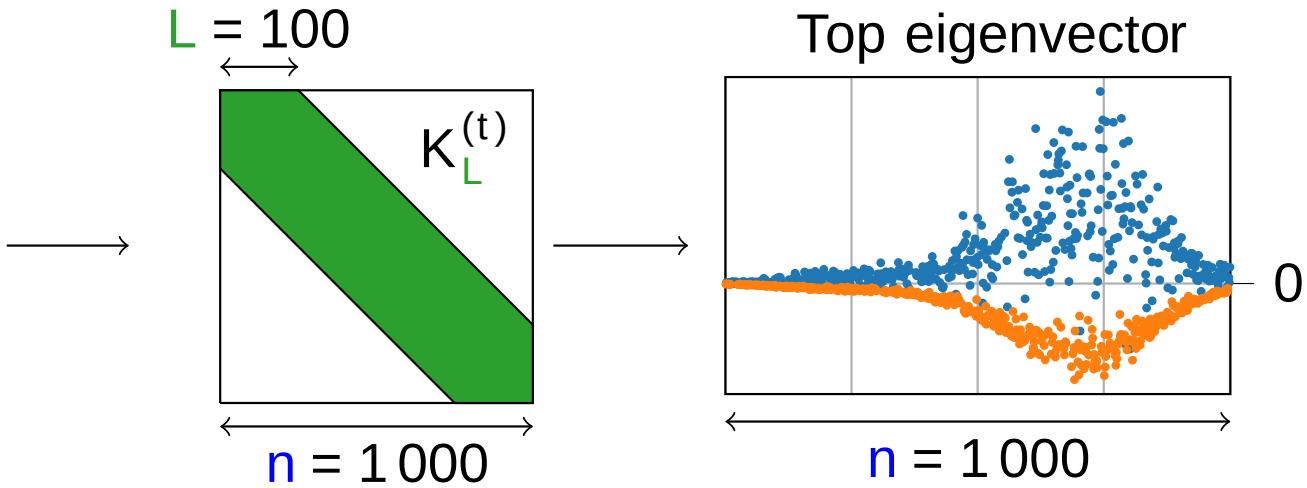
From theory to practice: online kernel spectral clusterin

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



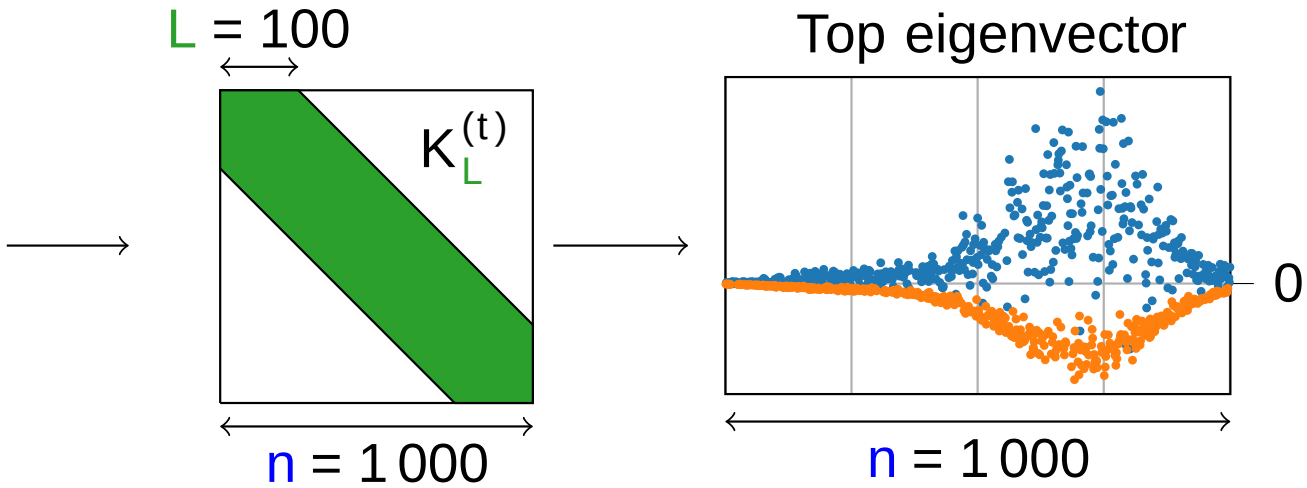
From theory to practice: online kernel spectral clusterin

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



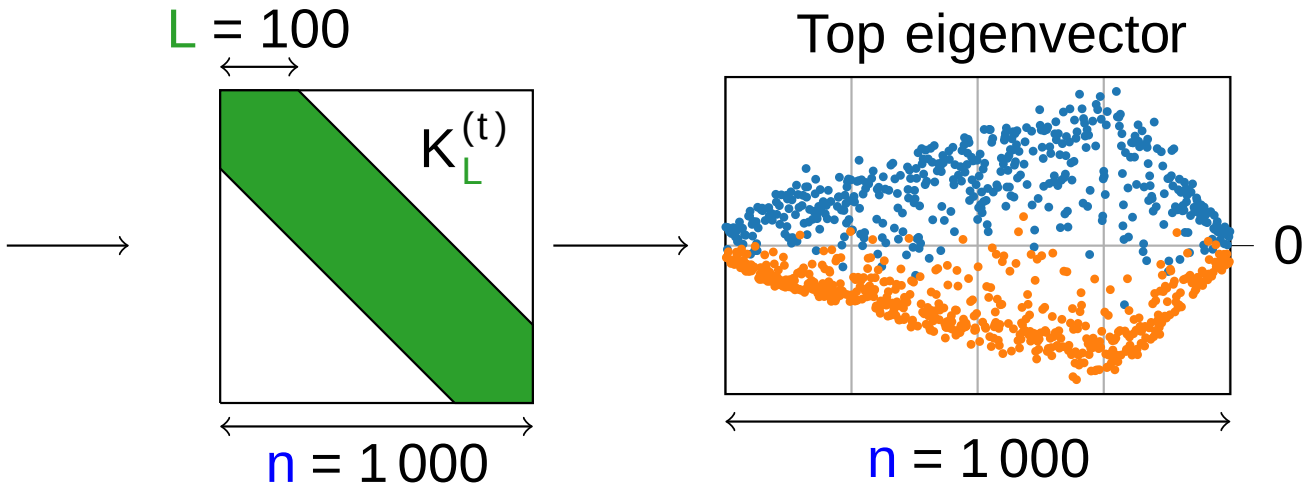
From theory to practice: online kernel spectral clusterin

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



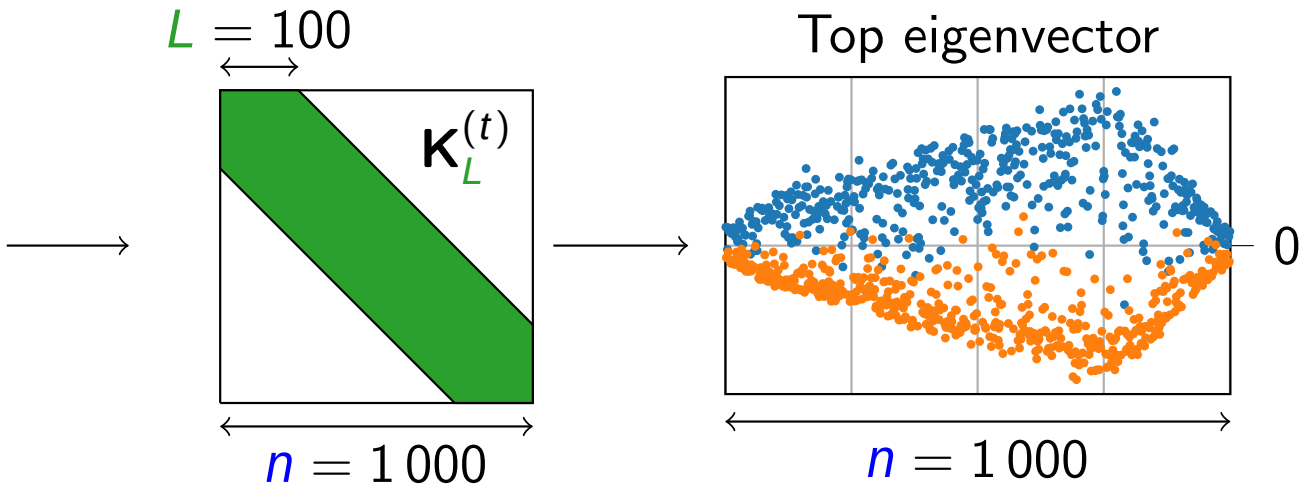
From theory to practice: online kernel spectral clusterin

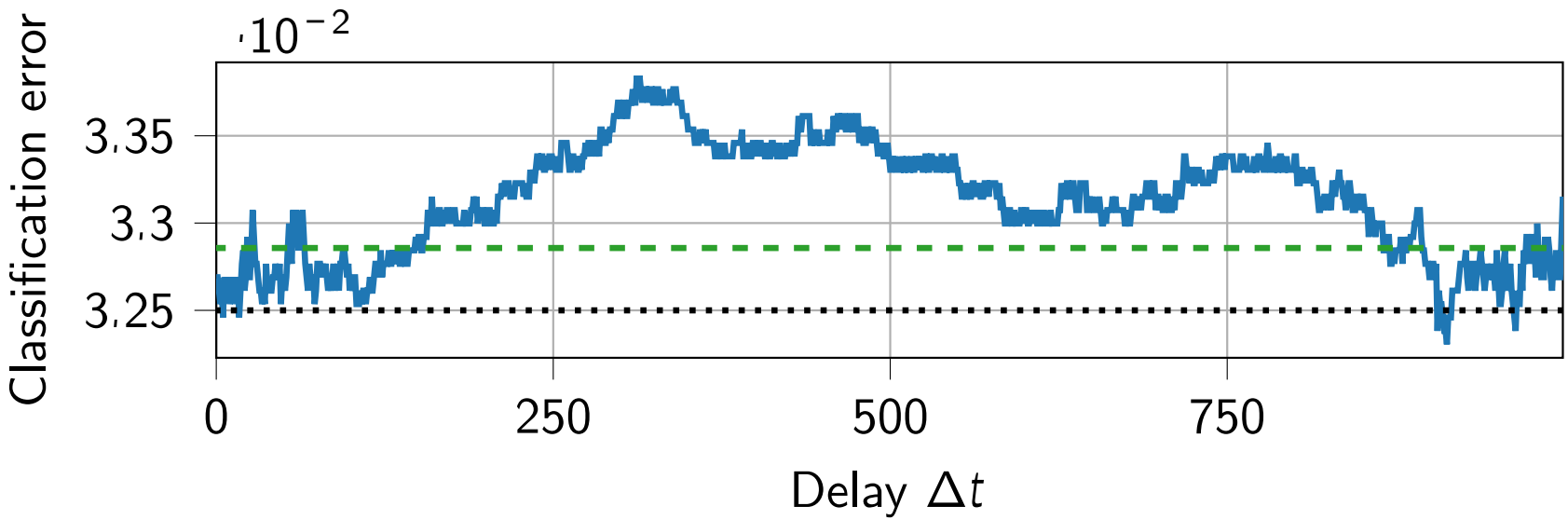
Fashion-MNIST images (raw, $p = 784$) $T = 14\,000$



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