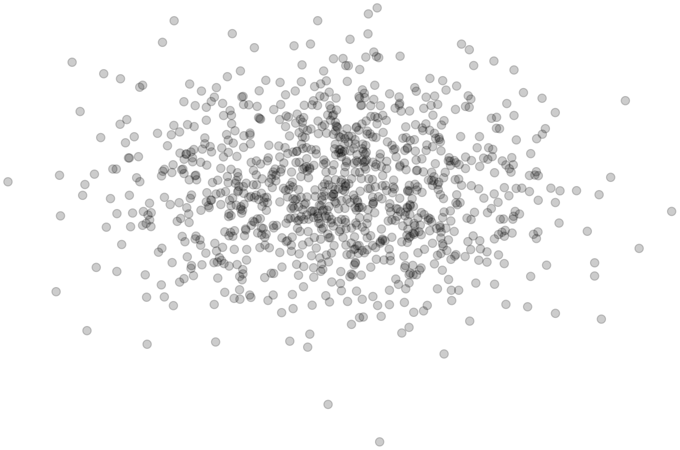
Determinantal Point Processes for Coresets

(Journées MAS, Rouen, August 2022)

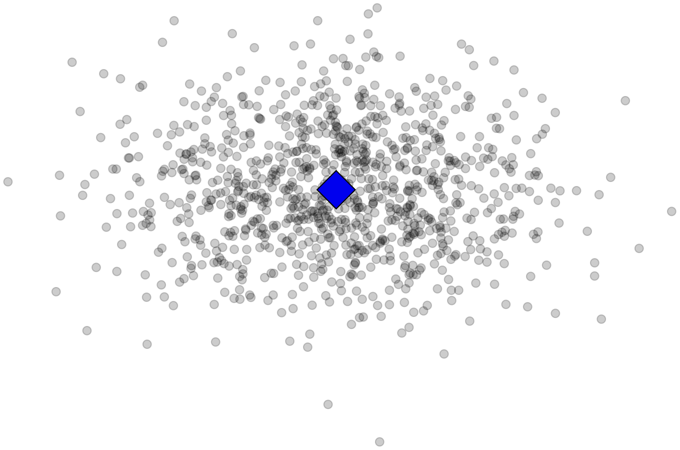
Nicolas Tremblay, Simon Barthelmé, Pierre-Olivier Amblard

CNRS, GIPSA-lab, Univ. Grenoble-Alpes, France

Task: find the mean



Task: find the mean



Task: find the mean

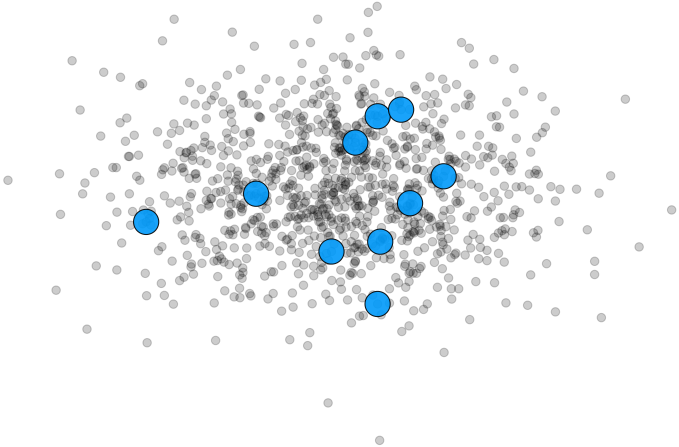


Figure: Example of iid uniform sampling

Task: find the mean

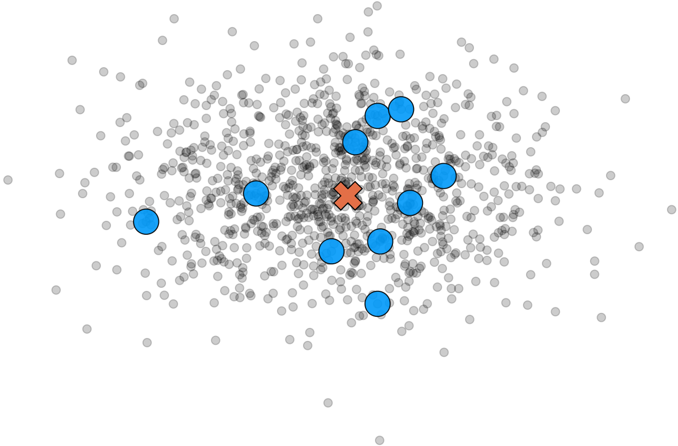


Figure: Example of iid uniform sampling

Task: find the mean

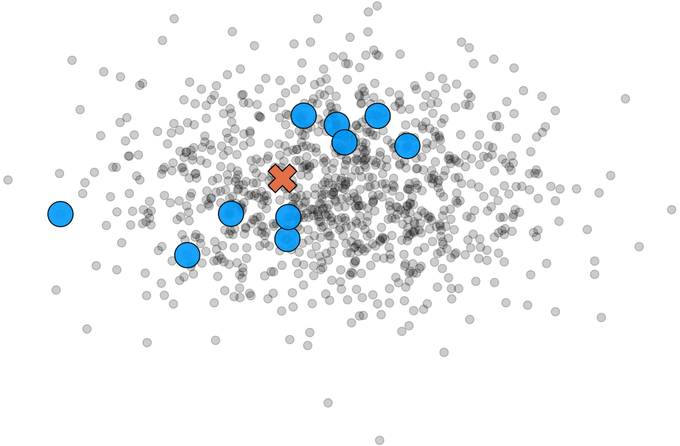


Figure: Example of iid uniform sampling

Task: find the mean

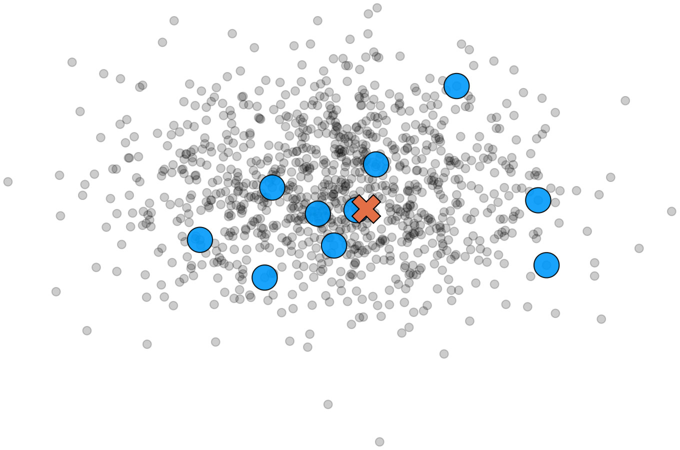


Figure: Example of iid uniform sampling

Task: find the mean

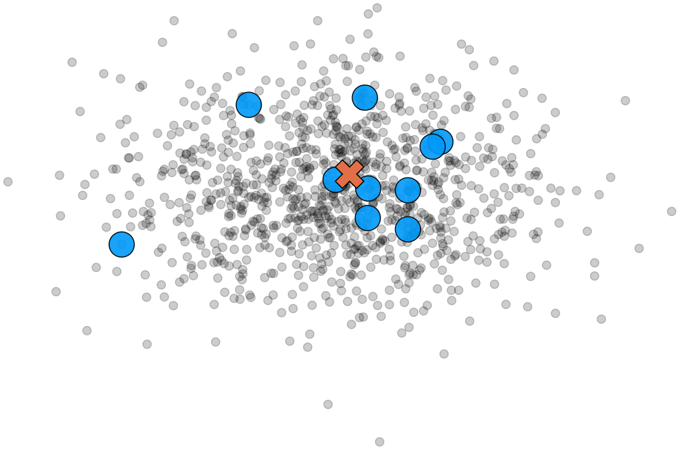


Figure: Example of iid uniform sampling

Task: find the mean

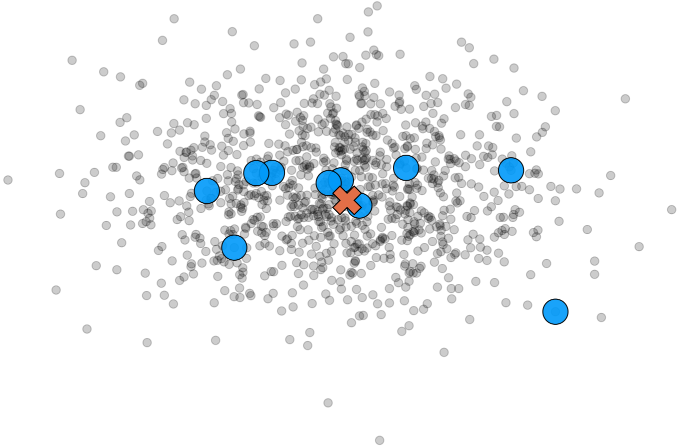


Figure: Example of iid uniform sampling

Task: find the mean

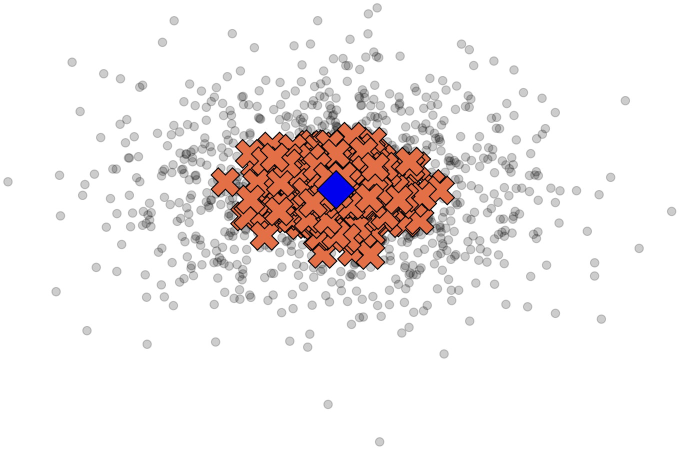


Figure: Uniform iid estimations of the mean

Task: find the mean

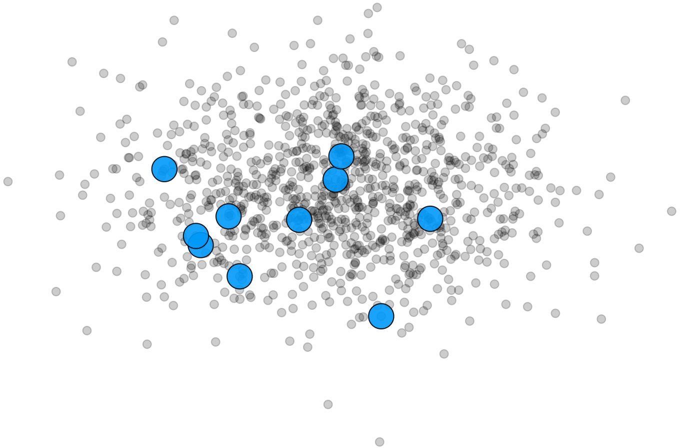


Figure: Example of a smarter iid sampling (sensitivity-based importance sampling)

Task: find the mean

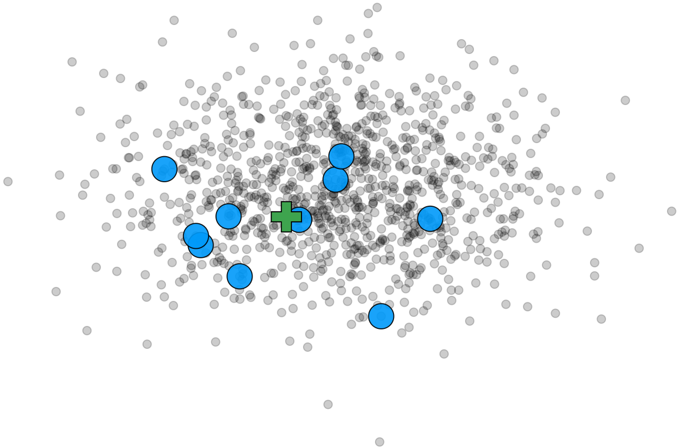


Figure: Example of a smarter iid sampling (sensitivity-based importance sampling)

Task: find the mean

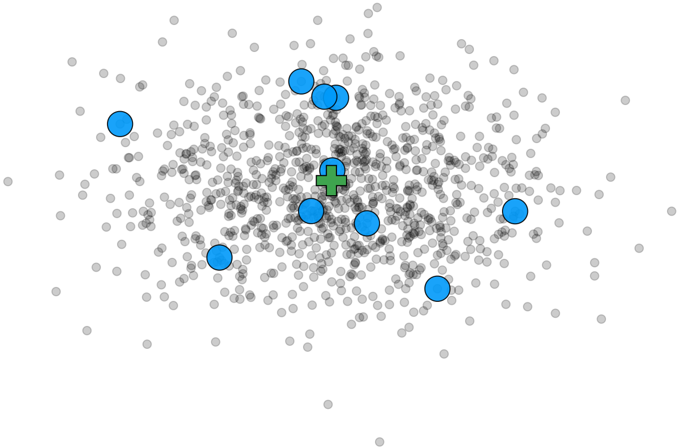


Figure: Example of a smarter iid sampling (sensitivity-based importance sampling)

Task: find the mean

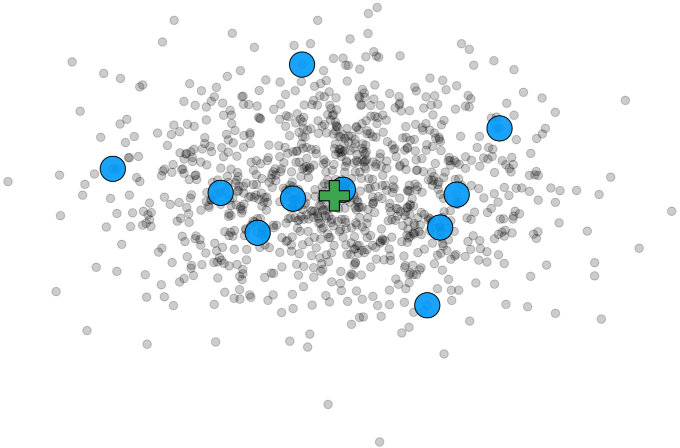


Figure: Example of a smarter iid sampling (sensitivity-based importance sampling)

Task: find the mean

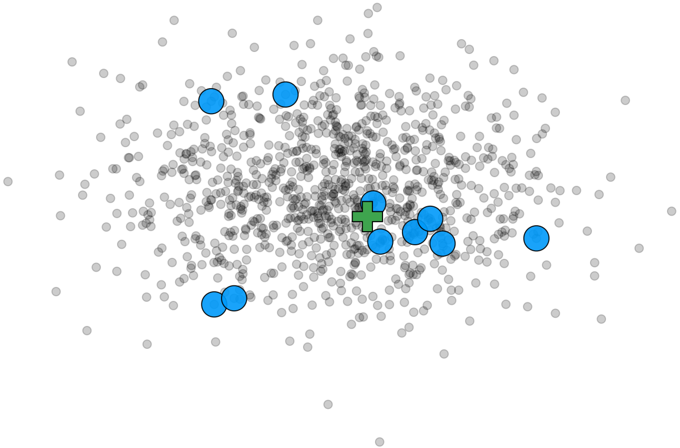


Figure: Example of a smarter iid sampling (sensitivity-based importance sampling)

Task: find the mean

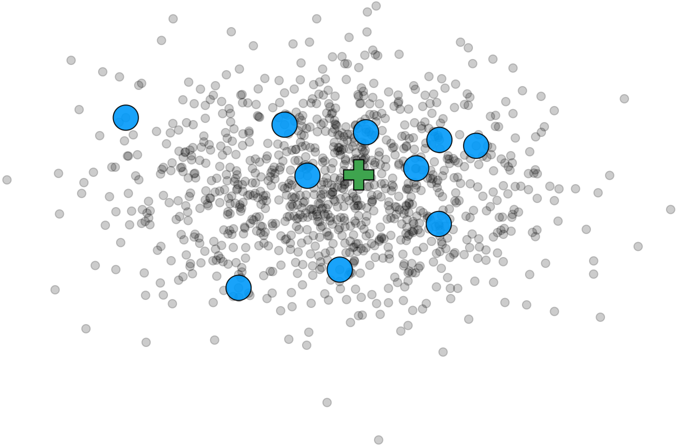


Figure: Example of a smarter iid sampling (sensitivity-based importance sampling)

Task: find the mean

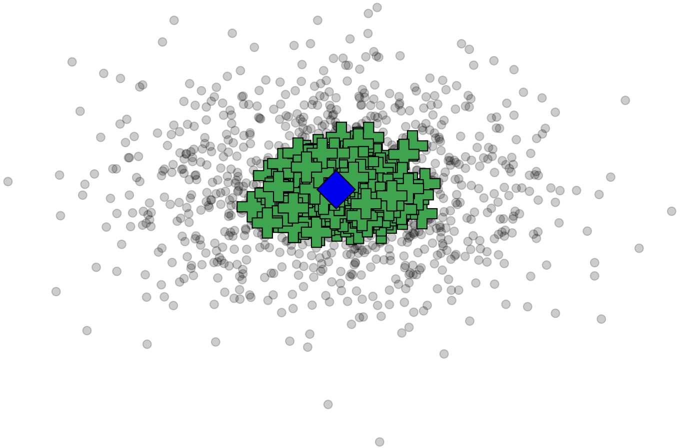


Figure: Smarter iid estimations of the mean (sensitivity-based importance sampling)

Task: find the mean

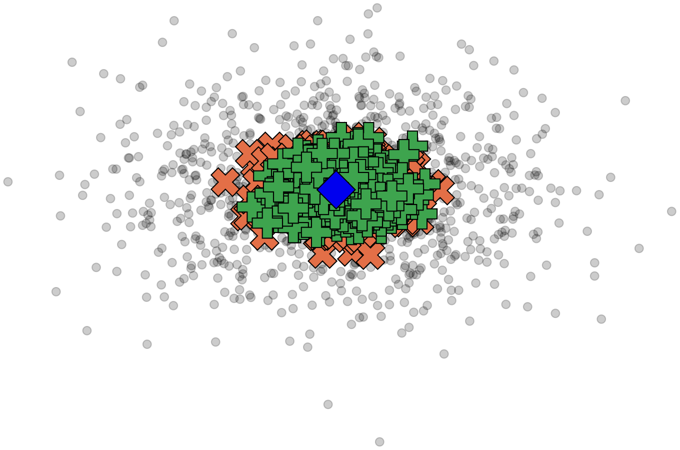


Figure: Comparison of both estimators: **variance reduction**

Task: find the mean

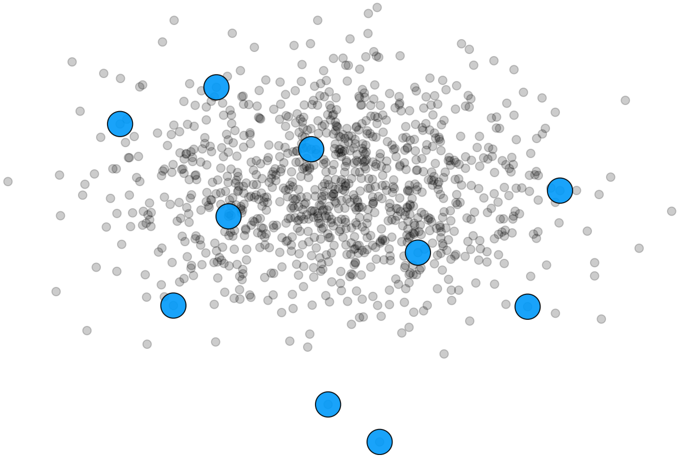


Figure: Example of DPP sampling (using Alg. 1 of the paper)

Task: find the mean

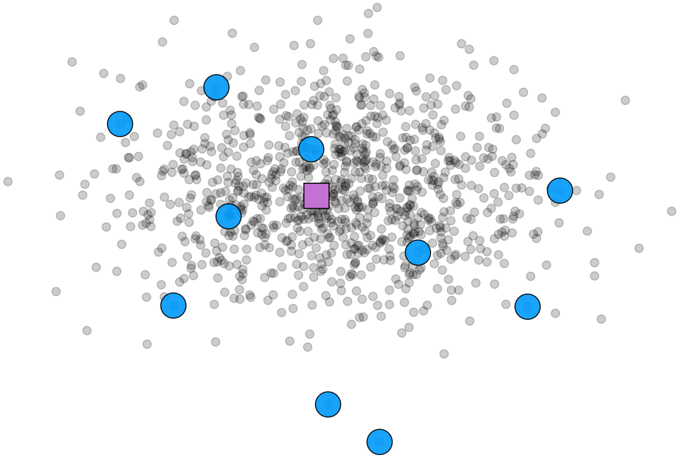


Figure: Example of DPP sampling (using Alg. 1 of the paper)

Task: find the mean

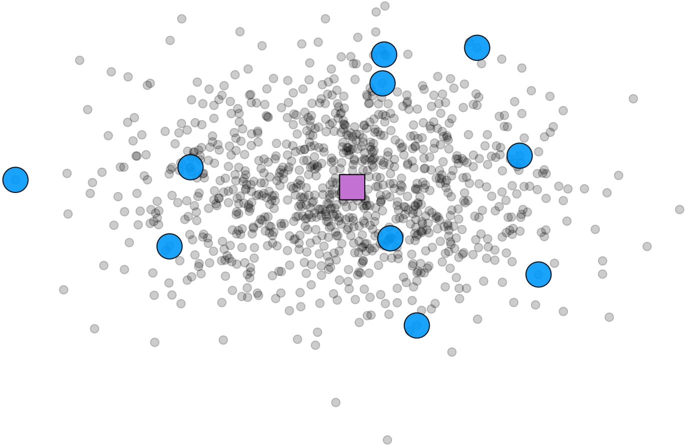


Figure: Example of DPP sampling (using Alg. 1 of the paper)

Task: find the mean

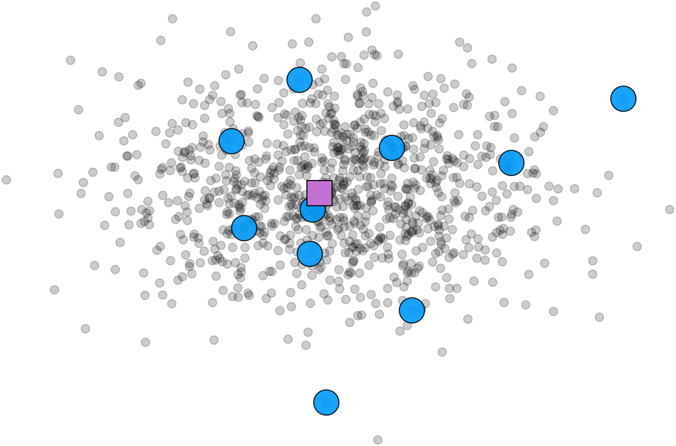


Figure: Example of DPP sampling (using Alg. 1 of the paper)

Task: find the mean

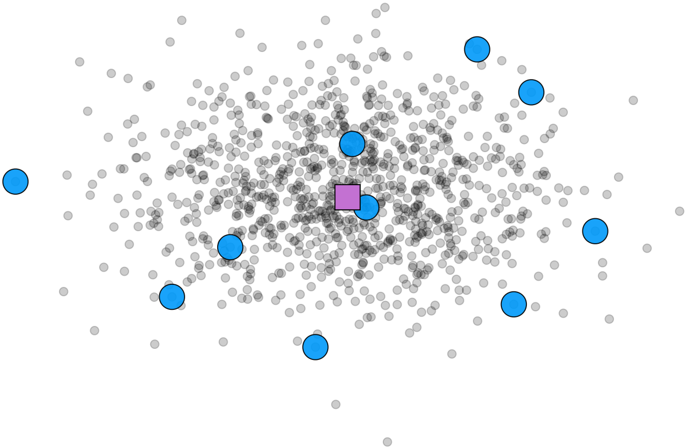


Figure: Example of DPP sampling (using Alg. 1 of the paper)

Task: find the mean

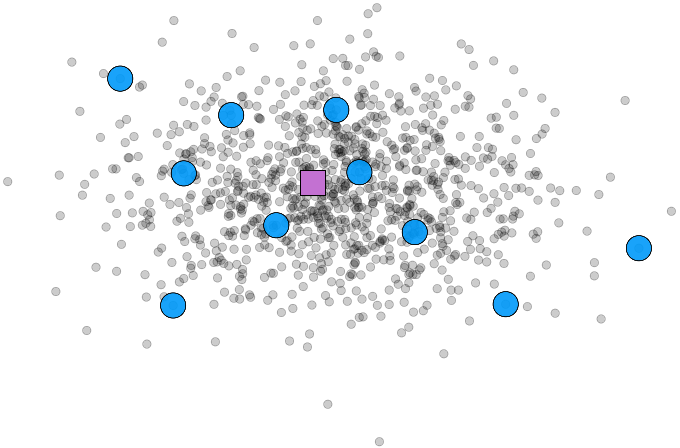


Figure: Example of DPP sampling (using Alg. 1 of the paper)

Task: find the mean

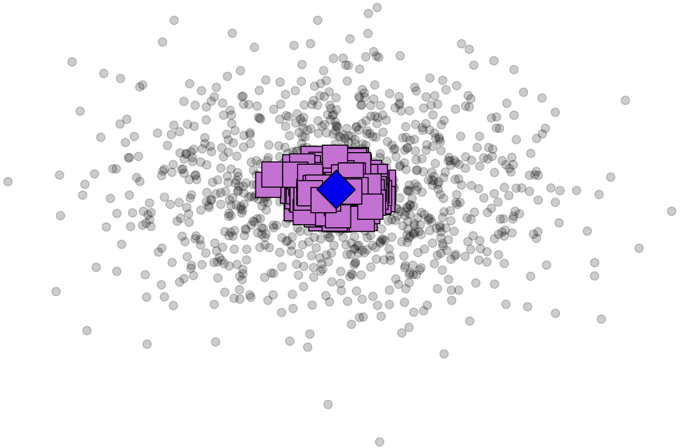


Figure: DPP estimations of the mean (using Alg. 1 of the paper)

Task: find the mean

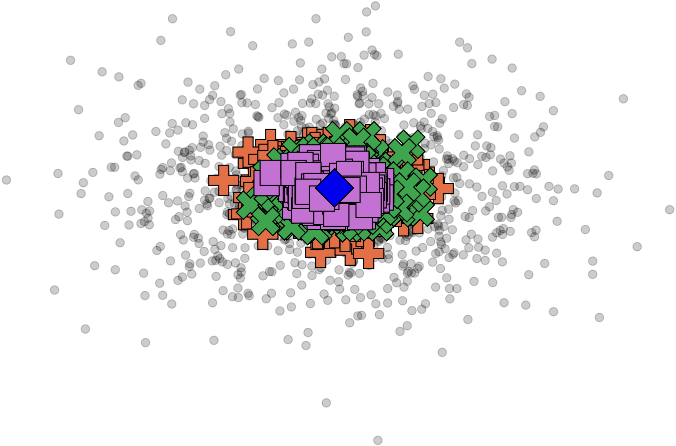
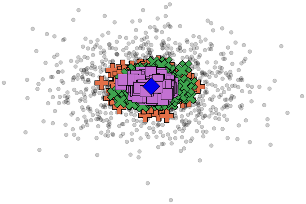


Figure: Comparison of all estimators: **more variance reduction**

Context and goal of coresets

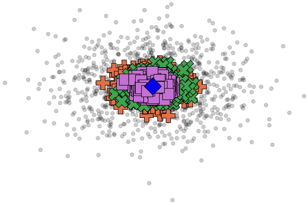


Context

- (very) large n (size of dataset)
- a precise task at hand

[1]Munteanu and Schwiegelshohn, *Coresets-Methods and History: A Theoreticians Design Pattern...*, KI, 2017

Context and goal of coresets



Context

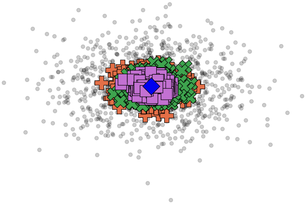
- (very) large n (size of dataset)
- a precise task at hand

Goal

- a coreset is a **tiny** (size indep./polylog of n) sample of the data **for the task** at hand
- a coreset has **provable guarantees** on the error made
- a coreset should be **sampled faster** than solving the task on the original data (!)

[1] Munteanu and Schwiegelshohn, *Coresets-Methods and History: A Theoreticians Design Pattern...*, KI, 2017

Context and goal of coresets



Context

- (very) large n (size of dataset)
- a precise task at hand

Goal

- a coreset is a **tiny** (size indep./polylog of n) sample of the data **for the task** at hand
- a coreset has **provable guarantees** on the error made
- a coreset should be **sampled faster** than solving the task on the original data (!)

State-of-the-art[1]

- Verifying all 3 points is very challenging. The state-of-the-art usually comes with
  - a provable algorithm but very expensive
  - some work-arounds more affordable, still provable, but (much) less efficient
  - some heuristics inspired by these provable algorithms
- coresets under research: deterministic, iid random, multi-task, streaming, ...

[1] Munteanu and Schwiegelshohn, *Coresets-Methods and History: A Theoreticians Design Pattern...*, KI, 2017

A generic class of problems

- Consider a dataset $\mathcal{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)$, say: n points in dimension d.
- Let Θ be a parameter space and consider cost functions of the form:

$$L(\mathcal{X}, \theta) = \sum_{i=1}^{n} f(\boldsymbol{x}_i, \theta)$$

  where $f : \mathcal{X}, \Theta \to \mathbb{R}^+$.
- A classical ML objective: find

$$\theta^* = \operatorname*{argmin}_{\theta \in \Theta} L(\mathcal{X}, \theta).$$

Examples falling in this class of problems

$\boxed{\text{Find } \theta^* = \operatorname{argmin}_{\theta \in \Theta} L(\mathcal{X}, \theta) \text{ where } L(\mathcal{X}, \theta) = \sum_{i=1}^n f(\boldsymbol{x}_i, \theta), \text{ and } f : \mathcal{X}, \Theta \to \mathbb{R}^+.}$

Examples falling in this class of problems

Find $\theta^* = \text{argmin}_{\theta\in\Theta} L(\mathcal{X}, \theta)$ where $L(\mathcal{X}, \theta) = \sum_{i=1}^{n} f(\boldsymbol{x}_i, \theta)$, and $f : \mathcal{X}, \Theta \rightarrow \mathbb{R}^+$.

- *k*-**means** The k-means objective is to find k centroids $\{\boldsymbol{c}_\ell\}_{\ell=1,\dots,k}$ in $\mathbb{R}^d$ such that $L(\mathcal{X}, \{\boldsymbol{c}_\ell\}) = \sum_{\boldsymbol{x}\in\mathcal{X}} f(\boldsymbol{x}, \{\boldsymbol{c}_\ell\})$ is minimal, with

$$f(\boldsymbol{x}, \{\boldsymbol{c}_\ell\}) = \min_{\boldsymbol{c}_\ell} ||\boldsymbol{x} - \boldsymbol{c}_\ell||^2$$

Examples falling in this class of problems

$$\boxed{\text{Find } \theta^* = \operatorname{argmin}_{\theta\in\Theta} L(\mathcal{X},\theta) \text{ where } L(\mathcal{X},\theta) = \sum_{i=1}^n f(\mathbf{x}_i,\theta), \text{ and } f : \mathcal{X},\Theta \to \mathbb{R}^+.}$$

- *k*-**means** The k-means objective is to find k centroids $\{\mathbf{c}_\ell\}_{\ell=1,\dots,k}$ in $\mathbb{R}^d$ such that $L(\mathcal{X},\{\mathbf{c}_\ell\}) = \sum_{\mathbf{x}\in\mathcal{X}} f(\mathbf{x},\{\mathbf{c}_\ell\})$ is minimal, with

$$f(\mathbf{x},\{\mathbf{c}_\ell\}) = \min_{\mathbf{c}_\ell} ||\mathbf{x}-\mathbf{c}_\ell||^2 \qquad (\geqslant 0).$$

Examples falling in this class of problems

Find $\theta^* = \text{argmin}_{\theta\in\Theta} L(\mathcal{X},\theta)$ where $L(\mathcal{X},\theta) = \sum_{i=1}^n f(\boldsymbol{x}_i,\theta)$, and $f : \mathcal{X},\Theta \rightarrow \mathbb{R}^+$.

- *k*-**means** The *k*-means objective is to find k centroids $\{\boldsymbol{c}_\ell\}_{\ell=1,\dots,k}$ in $\mathbb{R}^d$ such that $L(\mathcal{X},\{\boldsymbol{c}_\ell\}) = \sum_{\boldsymbol{x}\in\mathcal{X}} f(\boldsymbol{x},\{\boldsymbol{c}_\ell\})$ is minimal, with

$$f(\boldsymbol{x},\{\boldsymbol{c}_\ell\}) = \min_{\boldsymbol{c}_\ell} ||\boldsymbol{x}-\boldsymbol{c}_\ell||^2 \qquad (\geqslant 0).$$

- **Other examples:** linear regression, logistic regression, *k*-median, low-rank approx. of matrices, etc.

Coresets: definition

- Consider $\{\mathcal{S}, \{\omega_s > 0\}\}$ a weighted sample of $\mathcal{X}$ and its associated estimated cost

$$\hat{L}(\mathcal{S}, \theta) = \sum_{\boldsymbol{s} \in \mathcal{S}} \omega_s f(\boldsymbol{s}, \theta)$$

Coresets: definition

- Consider $\{\mathcal{S}, \{\omega_s > 0\}\}$ a weighted sample of $\mathcal{X}$ and its associated estimated cost

$$\hat{L}(\mathcal{S}, \theta) = \sum_{\boldsymbol{s} \in \mathcal{S}} \omega_s f(\boldsymbol{s}, \theta)$$

- **Definition** (ϵ-coreset) *A weighted sample $\mathcal{S}$ is an ϵ-coreset of $\mathcal{X}$ wrt L if:*

$$\forall \theta \in \Theta \qquad (1-\epsilon) L(\mathcal{X}, \theta) \leqslant \hat{L}(\mathcal{S}, \theta) \leqslant (1+\epsilon) L(\mathcal{X}, \theta)$$

Coresets: definition

- Consider $\{\mathcal{S}, \{\omega_s > 0\}\}$ a weighted sample of $\mathcal{X}$ and its associated estimated cost

$$\hat{L}(\mathcal{S}, \theta) = \sum_{\boldsymbol{s} \in \mathcal{S}} \omega_s f(\boldsymbol{s}, \theta)$$

- **Definition** (ϵ-coreset) *A weighted sample $\mathcal{S}$ is an ϵ-coreset of $\mathcal{X}$ wrt L if:*

$$\forall \theta \in \Theta \qquad (1-\epsilon)L(\mathcal{X}, \theta) \leqslant \hat{L}(\mathcal{S}, \theta) \leqslant (1+\epsilon)L(\mathcal{X}, \theta)$$

Coresets: definition

- Consider $\{\mathcal{S}, \{\omega_s > 0\}\}$ a weighted sample of $\mathcal{X}$ and its associated estimated cost

$$\hat{L}(\mathcal{S}, \theta) = \sum_{\boldsymbol{s} \in \mathcal{S}} \omega_s f(\boldsymbol{s}, \theta)$$

- **Definition** (ϵ-coreset) *A weighted sample $\mathcal{S}$ is an ϵ-coreset of $\mathcal{X}$ wrt L if:*

$$\forall \theta \in \boldsymbol{\Theta} \qquad (1-\epsilon) L(\mathcal{X}, \theta) \leqslant \hat{L}(\mathcal{S}, \theta) \leqslant (1+\epsilon) L(\mathcal{X}, \theta)$$

- Denote by $\hat{\theta}^*$ the argmin of $\hat{L}$: $\quad \hat{\theta}^* = \operatorname*{argmin}_{\theta \in \Theta} \hat{L}(\mathcal{S}, \theta).$

Coresets: definition

- Consider $\{\mathcal{S}, \{\omega_s > 0\}\}$ a weighted sample of $\mathcal{X}$ and its associated estimated cost

$$\hat{L}(\mathcal{S}, \theta) = \sum_{\boldsymbol{s} \in \mathcal{S}} \omega_s f(\boldsymbol{s}, \theta)$$

- **Definition** (ϵ-coreset) *A weighted sample $\mathcal{S}$ is an ϵ-coreset of $\mathcal{X}$ wrt L if:*

$$\forall \theta \in \Theta \qquad (1-\epsilon)L(\mathcal{X}, \theta) \leqslant \hat{L}(\mathcal{S}, \theta) \leqslant (1+\epsilon)L(\mathcal{X}, \theta)$$

- Denote by $\hat{\theta}^*$ the argmin of $\hat{L}$: $\qquad \hat{\theta}^* = \operatorname{argmin}_{\theta \in \Theta} \hat{L}(\mathcal{S}, \theta)$.
- Why are coresets interesting?
If L has a clear global minimum (up to an ϵ factor) in θ^*, then $\hat{\theta}^* \simeq \theta^*$:

Coresets: definition

- Consider $\{\mathcal{S}, \{\omega_s > 0\}\}$ a weighted sample of $\mathcal{X}$ and its associated estimated cost

$$\hat{L}(\mathcal{S}, \theta) = \sum_{s \in \mathcal{S}} \omega_s f(\boldsymbol{s}, \theta)$$

- **Definition** (ϵ-coreset) *A weighted sample $\mathcal{S}$ is an ϵ-coreset of $\mathcal{X}$ wrt L if:*

$$\forall \theta \in \Theta \qquad (1-\epsilon) L(\mathcal{X}, \theta) \leqslant \hat{L}(\mathcal{S}, \theta) \leqslant (1+\epsilon) L(\mathcal{X}, \theta)$$

- Denote by $\hat{\theta}^*$ the argmin of $\hat{L}$: $\qquad \hat{\theta}^* = \underset{\theta \in \Theta}{\operatorname{argmin}} \hat{L}(\mathcal{S}, \theta).$
- Why are coresets interesting?
  If L has a clear global minimum (up to an ϵ factor) in θ^*, then $\hat{\theta}^* \simeq \theta^*$:

$$\hat{L}(\mathcal{S}, \hat{\theta}^*)$$

Coresets: definition

- Consider $\{\mathcal{S}, \{\omega_s > 0\}\}$ a weighted sample of $\mathcal{X}$ and its associated estimated cost

$$\hat{L}(\mathcal{S}, \theta) = \sum_{\boldsymbol{s} \in \mathcal{S}} \omega_s f(\boldsymbol{s}, \theta)$$

- **Definition** (ϵ-coreset) *A weighted sample $\mathcal{S}$ is an ϵ-coreset of $\mathcal{X}$ wrt L if:*

$$\forall \theta \in \Theta \qquad (1-\epsilon) L(\mathcal{X}, \theta) \leqslant \hat{L}(\mathcal{S}, \theta) \leqslant (1+\epsilon) L(\mathcal{X}, \theta)$$

- Denote by $\hat{\theta}^*$ the argmin of $\hat{L}$: $\qquad \hat{\theta}^* = \underset{\theta \in \Theta}{\operatorname{argmin}} \hat{L}(\mathcal{S}, \theta).$
- Why are coresets interesting?
  If L has a clear global minimum (up to an ϵ factor) in θ^*, then $\hat{\theta}^* \simeq \theta^*$:

$$\hat{L}(\mathcal{S}, \hat{\theta}^*) \leqslant \hat{L}(\mathcal{S}, \theta^*)$$

Coresets: definition

- Consider $\{\mathcal{S}, \{\omega_s > 0\}\}$ a weighted sample of $\mathcal{X}$ and its associated estimated cost

$$\hat{L}(\mathcal{S}, \theta) = \sum_{\boldsymbol{s} \in \mathcal{S}} \omega_s f(\boldsymbol{s}, \theta)$$

- **Definition** (ϵ-coreset) *A weighted sample $\mathcal{S}$ is an ϵ-coreset of $\mathcal{X}$ wrt L if:*

$$\forall \theta \in \Theta \qquad (1-\epsilon) L(\mathcal{X}, \theta) \leqslant \hat{L}(\mathcal{S}, \theta) \leqslant (1+\epsilon) L(\mathcal{X}, \theta)$$

- Denote by $\hat{\theta}^*$ the argmin of $\hat{L}$: $\qquad \hat{\theta}^* = \operatorname*{argmin}_{\theta \in \Theta} \hat{L}(\mathcal{S}, \theta).$
- Why are coresets interesting?
  If L has a clear global minimum (up to an ϵ factor) in θ^*, then $\hat{\theta}^* \simeq \theta^*$:

$$\hat{L}(\mathcal{S}, \hat{\theta}^*) \leqslant \hat{L}(\mathcal{S}, \theta^*) \leqslant (1+\epsilon) L(\mathcal{X}, \theta^*)$$

Coresets: definition

- Consider $\{\mathcal{S}, \{\omega_s > 0\}\}$ a weighted sample of $\mathcal{X}$ and its associated estimated cost

$$\hat{L}(\mathcal{S}, \theta) = \sum_{\boldsymbol{s} \in \mathcal{S}} \omega_s f(\boldsymbol{s}, \theta)$$

- **Definition** (ϵ-coreset) *A weighted sample $\mathcal{S}$ is an ϵ-coreset of $\mathcal{X}$ wrt L if:*

$$\forall \theta \in \Theta \qquad (1-\epsilon)L(\mathcal{X}, \theta) \leqslant \hat{L}(\mathcal{S}, \theta) \leqslant (1+\epsilon)L(\mathcal{X}, \theta)$$

- Denote by $\hat{\theta}^*$ the argmin of $\hat{L}$: $\qquad \hat{\theta}^* = \operatorname*{argmin}_{\theta \in \Theta} \hat{L}(\mathcal{S}, \theta)$.

- Why are coresets interesting?
  If L has a clear global minimum (up to an ϵ factor) in θ^*, then $\hat{\theta}^* \simeq \theta^*$:

$$(1-\epsilon)L(\mathcal{X}, \hat{\theta}^*) \leqslant \hat{L}(\mathcal{S}, \hat{\theta}^*) \leqslant \hat{L}(\mathcal{S}, \theta^*) \leqslant (1+\epsilon)L(\mathcal{X}, \theta^*)$$

Coresets: definition

- Consider $\{\mathcal{S}, \{\omega_s > 0\}\}$ a weighted sample of $\mathcal{X}$ and its associated estimated cost

$$\hat{L}(\mathcal{S}, \theta) = \sum_{\boldsymbol{s} \in \mathcal{S}} \omega_s f(\boldsymbol{s}, \theta)$$

- **Definition** (ϵ-coreset) *A weighted sample $\mathcal{S}$ is an ϵ-coreset of $\mathcal{X}$ wrt L if:*

$$\forall \theta \in \Theta \qquad (1-\epsilon) L(\mathcal{X}, \theta) \leqslant \hat{L}(\mathcal{S}, \theta) \leqslant (1+\epsilon) L(\mathcal{X}, \theta)$$

- Denote by $\hat{\theta}^*$ the argmin of $\hat{L}$: $\qquad \hat{\theta}^* = \underset{\theta \in \Theta}{\operatorname{argmin}} \hat{L}(\mathcal{S}, \theta)$.
- Why are coresets interesting?
  If L has a clear global minimum (up to an ϵ factor) in θ^*, then $\hat{\theta}^* \simeq \theta^*$:

$$(1-\epsilon) L(\mathcal{X}, \theta^*) \leqslant (1-\epsilon) L(\mathcal{X}, \hat{\theta}^*) \leqslant \hat{L}(\mathcal{S}, \hat{\theta}^*) \leqslant \hat{L}(\mathcal{S}, \theta^*) \leqslant (1+\epsilon) L(\mathcal{X}, \theta^*)$$

Coresets: definition

- Consider $\{\mathcal{S}, \{\omega_s > 0\}\}$ a weighted sample of $\mathcal{X}$ and its associated estimated cost

$$\hat{L}(\mathcal{S}, \theta) = \sum_{\boldsymbol{s} \in \mathcal{S}} \omega_s f(\boldsymbol{s}, \theta)$$

- **Definition** (ϵ-coreset) *A weighted sample $\mathcal{S}$ is an ϵ-coreset of $\mathcal{X}$ wrt L if:*

$$\forall \theta \in \Theta \qquad (1-\epsilon)L(\mathcal{X}, \theta) \leqslant \hat{L}(\mathcal{S}, \theta) \leqslant (1+\epsilon)L(\mathcal{X}, \theta)$$

- Denote by $\hat{\theta}^*$ the argmin of $\hat{L}$: $\qquad \hat{\theta}^* = \underset{\theta \in \Theta}{\text{argmin}}\, \hat{L}(\mathcal{S}, \theta).$
- Why are coresets interesting?
  If L has a clear global minimum (up to an ϵ factor) in θ^*, then $\hat{\theta}^* \simeq \theta^*$:

$$(1-\epsilon)L(\mathcal{X}, \theta^*) \leqslant (1-\epsilon)L(\mathcal{X}, \hat{\theta}^*) \leqslant \hat{L}(\mathcal{S}, \hat{\theta}^*) \leqslant \hat{L}(\mathcal{S}, \theta^*) \leqslant (1+\epsilon)L(\mathcal{X}, \theta^*)$$
$$\Downarrow$$
$$(1-\epsilon)L(\mathcal{X}, \theta^*) \leqslant \hat{L}(\mathcal{S}, \hat{\theta}^*) \leqslant (1+\epsilon)L(\mathcal{X}, \theta^*)$$

Coresets: illustration on the 1-means problem

- Data $\mathcal{X}$

Coresets: illustration on the 1-means problem

- Data $\mathcal{X}$
- Cost function:

$$L(\mathcal{X}, \theta) = \sum_{i=1}^{n} \|\boldsymbol{x}_i - \theta\|^2$$

Coresets: illustration on the 1-means problem

- Data $\mathcal{X}$
- Cost function:

$$L(\mathcal{X}, \theta) = \sum_{i=1}^{n} \|\boldsymbol{x}_i - \theta\|^2$$

- Optimal θ:

$$\theta^* = \underset{\theta \in \Theta}{\operatorname{argmin}}\, L(\mathcal{X}, \theta)$$

Coresets: illustration on the 1-means problem

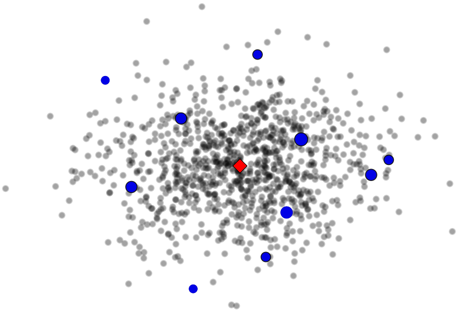
- Data $\mathcal{X}$
- Cost function:

$$L(\mathcal{X}, \theta) = \sum_{i=1}^{n} \|\boldsymbol{x}_i - \theta\|^2$$

- Optimal θ:

$$\theta^* = \underset{\theta \in \Theta}{\operatorname{argmin}}\, L(\mathcal{X}, \theta)$$

- A weighted subset $\mathcal{S}$



Coresets: illustration on the 1-means problem

- Data $\mathcal{X}$
- Cost function:

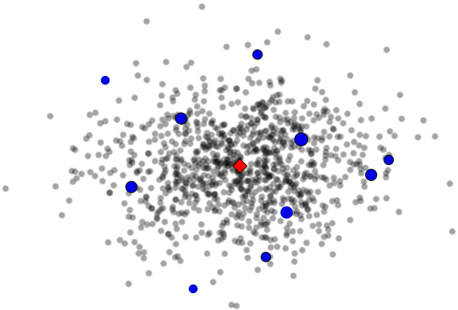
$$L(\mathcal{X}, \theta) = \sum_{i=1}^{n} \|\boldsymbol{x}_i - \theta\|^2$$

- Optimal θ:

$$\theta^* = \underset{\theta \in \Theta}{\operatorname{argmin}}\, L(\mathcal{X}, \theta)$$

- A weighted subset $\mathcal{S}$
- Estimated cost function:

$$\hat{L}(\mathcal{S}, \theta) = \sum_{\boldsymbol{s} \in \mathcal{S}} \omega_s \|\boldsymbol{s} - \theta\|^2$$



Coresets: illustration on the 1-means problem

- Data $\mathcal{X}$
- Cost function:

$$L(\mathcal{X}, \theta) = \sum_{i=1}^{n} \|\boldsymbol{x}_i - \theta\|^2$$

- Optimal θ:

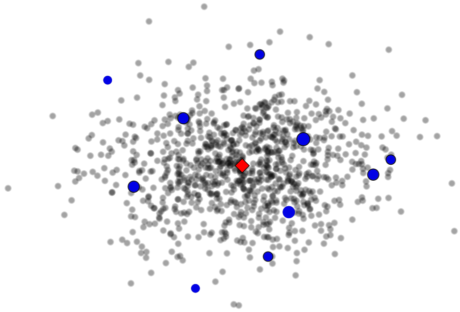
$$\theta^* = \operatorname*{argmin}_{\theta \in \Theta} L(\mathcal{X}, \theta)$$

- A weighted subset $\mathcal{S}$
- Estimated cost function:

$$\hat{L}(\mathcal{S}, \theta) = \sum_{\boldsymbol{s} \in \mathcal{S}} \omega_s \|\boldsymbol{s} - \theta\|^2$$

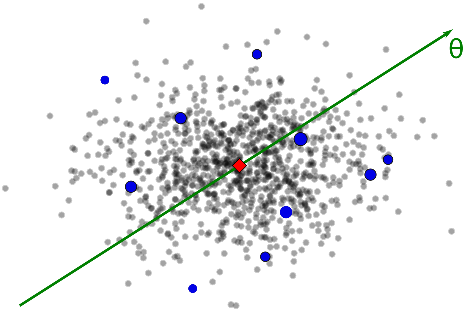
- $\mathcal{S}$ is a ϵ-coreset if:

$$\forall \theta \quad \left| \frac{\hat{L}}{L} - 1 \right| \leqslant \epsilon$$



Coresets: illustration on the 1-means problem

- Data $\mathcal{X}$
- Cost function:
$$L(\mathcal{X}, \theta) = \sum_{i=1}^{n} \|\boldsymbol{x}_i - \theta\|^2$$
- Optimal θ:
$$\theta^* = \underset{\theta \in \Theta}{\operatorname{argmin}} \, L(\mathcal{X}, \theta)$$
- A weighted subset $\mathcal{S}$
- Estimated cost function:
$$\hat{L}(\mathcal{S}, \theta) = \sum_{\boldsymbol{s} \in \mathcal{S}} \omega_s \|\boldsymbol{s} - \theta\|^2$$
- $\mathcal{S}$ is a ϵ-coreset if:
$$\forall \theta \quad \left| \frac{\hat{L}}{L} - 1 \right| \leqslant \epsilon$$



Coresets: illustration on the 1-means problem

- Data $\mathcal{X}$
- Cost function:

$$L(\mathcal{X}, \theta) = \sum_{i=1}^{n} \|\boldsymbol{x}_i - \theta\|^2$$

- Optimal θ:

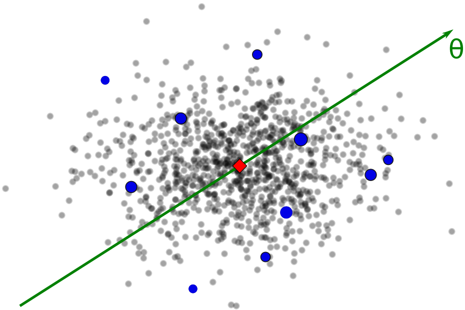
$$\theta^* = \underset{\theta \in \Theta}{\operatorname{argmin}} \, L(\mathcal{X}, \theta)$$

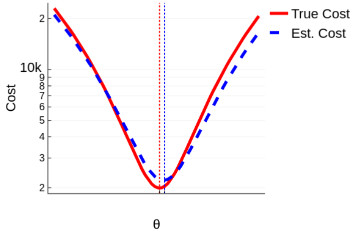
- A weighted subset $\mathcal{S}$
- Estimated cost function:

$$\hat{L}(\mathcal{S}, \theta) = \sum_{\boldsymbol{s} \in \mathcal{S}} \omega_s \|\boldsymbol{s} - \theta\|^2$$

- $\mathcal{S}$ is a ϵ-coreset if:

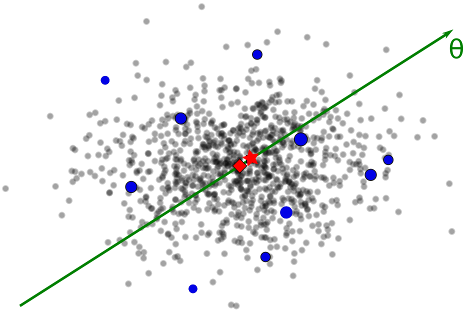
$$\forall \theta \quad \left| \frac{\hat{L}}{L} - 1 \right| \leqslant \epsilon$$

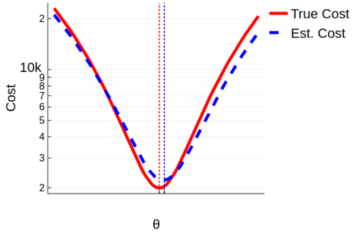




Coresets: illustration on the 1-means problem

- Data $\mathcal{X}$
- Cost function:
$$L(\mathcal{X}, \theta) = \sum_{i=1}^{n} \|\mathbf{x}_i - \theta\|^2$$
- Optimal θ:
$$\theta^* = \underset{\theta \in \Theta}{\operatorname{argmin}} \, L(\mathcal{X}, \theta)$$
- A weighted subset $\mathcal{S}$
- Estimated cost function:
$$\hat{L}(\mathcal{S}, \theta) = \sum_{\mathbf{s} \in \mathcal{S}} \omega_s \, \|\mathbf{s} - \theta\|^2$$
- $\mathcal{S}$ is a ϵ-coreset if:
$$\forall \theta \quad \left| \frac{\hat{L}}{L} - 1 \right| \leqslant \epsilon$$
- Estimated optimal θ:
$$\hat{\theta}^* = \underset{\theta \in \Theta}{\operatorname{argmin}} \, \hat{L}(\mathcal{S}, \theta)$$





Sensitivity: a central object

- The **sensitivity** of a datapoint $\boldsymbol{x}_i \in \mathcal{X}$ with respect to $f : \mathcal{X}, \Theta \to \mathbb{R}^+$ is:

$$\sigma_i = \max_{\theta \in \Theta} \frac{f(\boldsymbol{x}_i, \theta)}{L(\mathcal{X}, \theta)} \qquad \in [0, 1].$$

  The total sensitivity is denoted $\mathfrak{S} = \sum_{i=1}^n \sigma_i$.

Sensitivity: a central object

- The **sensitivity** of a datapoint $\boldsymbol{x}_i \in \mathcal{X}$ with respect to $f : \mathcal{X}, \Theta \rightarrow \mathbb{R}^+$ is:

$$\sigma_i = \max_{\theta \in \Theta} \frac{f(\boldsymbol{x}_i, \theta)}{L(\mathcal{X}, \theta)} \qquad \in [0, 1].$$

  The total sensitivity is denoted $\mathfrak{S} = \sum_{i=1}^n \sigma_i$.
- In general, the sensitivity is unknown analytically. In the paper, we managed to compute the analytic sensitivities for two simples cases: 1-means and linear regression

Random coresets

- Random context: suppose $\mathcal{S}$ is a *random* subset $\mathcal{S} \subset \mathcal{X}$ (possibly with repetitions)
- Importance sampling notations:
  - Define ϵ_i the *random variable* counting the number of times $\boldsymbol{x}_i$ is in $\mathcal{S}$
  - To each element x_i associate the weight $\omega_i = \frac{1}{\mathbb{E}(\epsilon_i)}$
- One has:

$$\hat{L}(\mathcal{S}, \theta) = \sum_{i=1}^{n} f(\boldsymbol{x}_i, \theta) \frac{\epsilon_i}{\mathbb{E}(\epsilon_i)}.$$

- By construction, $\hat{L}$ is an unbiased estimator of L:

$$\mathbb{E}\left(\hat{L}(\mathcal{S}, \theta)\right) = \sum_{i=1}^{n} f(\boldsymbol{x}_i, \theta) = L(\mathcal{X}, \theta).$$

A classical iid coreset theorem[1]

- Let $\boldsymbol{p} \in [0,1]^n$ be a probability distribution over all datapoints $\mathcal{X}$ with p_i the probability of sampling $\boldsymbol{x}_i$ and $\sum_i p_i = 1$.

[1] Langberg and Schulman, *Universal ϵ-approximators for integrals*, SIAM, 2010

A classical iid coreset theorem[1]

- Let $\boldsymbol{p} \in [0,1]^n$ be a probability distribution over all datapoints $\mathcal{X}$ with p_i the probability of sampling $\boldsymbol{x}_i$ and $\sum_i p_i = 1$.
- Draw $\mathcal{S}$: m iid samples with replacement according to $\boldsymbol{p}$.

[1] Langberg and Schulman, *Universal ϵ-approximators for integrals*, SIAM, 2010

A classical iid coreset theorem[1]

- Let $\boldsymbol{p} \in [0,1]^n$ be a probability distribution over all datapoints $\mathcal{X}$ with p_i the probability of sampling $\boldsymbol{x}_i$ and $\sum_i p_i = 1$.
- Draw $\mathcal{S}$: m iid samples with replacement according to $\boldsymbol{p}$.
- Associate importance sampling weights to each element of $\mathcal{S}$.

[1] Langberg and Schulman, *Universal ϵ-approximators for integrals*, SIAM, 2010

A classical iid coreset theorem[1]

- Let $\boldsymbol{p} \in [0,1]^n$ be a probability distribution over all datapoints $\mathcal{X}$ with p_i the probability of sampling $\boldsymbol{x}_i$ and $\sum_i p_i = 1$.
- Draw $\mathcal{S}$: m iid samples with replacement according to $\boldsymbol{p}$.
- Associate importance sampling weights to each element of $\mathcal{S}$.
- **Theorem** The weighted sample $\mathcal{S}$ is a ϵ-coreset with high probability if:

$$m \geqslant \mathcal{O}\left(\frac{d'}{\epsilon^2}\left(\max_i \frac{\sigma_i}{p_i}\right)^2\right),$$

where d' is the pseudo-dimension of Θ (a generalization of the Vapnik-Chervonenkis dimension).

[1] Langberg and Schulman, *Universal ϵ-approximators for integrals*, SIAM, 2010

A classical iid coreset theorem[1]

- Let $\boldsymbol{p} \in [0,1]^n$ be a probability distribution over all datapoints $\mathcal{X}$ with p_i the probability of sampling $\boldsymbol{x}_i$ and $\sum_i p_i = 1$.
- Draw $\mathcal{S}$: m iid samples with replacement according to $\boldsymbol{p}$.
- Associate importance sampling weights to each element of $\mathcal{S}$.
- **Theorem** The weighted sample $\mathcal{S}$ is a ϵ-coreset with high probability if:

$$m \geqslant \mathcal{O}\left(\frac{d'}{\epsilon^2}\left(\max_i \frac{\sigma_i}{p_i}\right)^2\right),$$

  where d' is the pseudo-dimension of Θ (a generalization of the Vapnik-Chervonenkis dimension).
- The optimal probability distribution minimizing the rhs is $p_i = \sigma_i/\mathfrak{S}$.

[1] Langberg and Schulman, *Universal ϵ-approximators for integrals*, SIAM, 2010

A classical iid coreset theorem[1]

- Let $\boldsymbol{p} \in [0,1]^n$ be a probability distribution over all datapoints $\mathcal{X}$ with p_i the probability of sampling $\boldsymbol{x}_i$ and $\sum_i p_i = 1$.
- Draw $\mathcal{S}$: m iid samples with replacement according to $\boldsymbol{p}$.
- Associate importance sampling weights to each element of $\mathcal{S}$.
- **Theorem** The weighted sample $\mathcal{S}$ is a ϵ-coreset with high probability if:

$$m \geqslant \mathcal{O}\left(\frac{d'}{\epsilon^2}\left(\max_i \frac{\sigma_i}{p_i}\right)^2\right),$$

  where d' is the pseudo-dimension of Θ (a generalization of the Vapnik-Chervonenkis dimension).
- The optimal probability distribution minimizing the rhs is $p_i = \sigma_i/\mathfrak{S}$.
- In this case, $\mathcal{S}$ is a ϵ-coreset with high probability if:

$$m \geqslant \mathcal{O}\left(\frac{d'\mathfrak{S}^2}{\epsilon^2}\right).$$

[1] Langberg and Schulman, *Universal ϵ-approximators for integrals*, SIAM, 2010

A classical iid coreset theorem[1]

- Let $\boldsymbol{p} \in [0,1]^n$ be a probability distribution over all datapoints $\mathcal{X}$ with p_i the probability of sampling $\boldsymbol{x}_i$ and $\sum_i p_i = 1$.
- Draw $\mathcal{S}$: m iid samples with replacement according to $\boldsymbol{p}$.
- Associate importance sampling weights to each element of $\mathcal{S}$.
- **Theorem** The weighted sample $\mathcal{S}$ is a ϵ-coreset with high probability if:

$$m \geqslant \mathcal{O}\left(\frac{d'}{\epsilon^2}\left(\max_i \frac{\sigma_i}{p_i}\right)^2\right),$$

  where d' is the pseudo-dimension of Θ (a generalization of the Vapnik-Chervonenkis dimension).
- The optimal probability distribution minimizing the rhs is $p_i = \sigma_i/\mathfrak{S}$.
- In this case, $\mathcal{S}$ is a ϵ-coreset with high probability if:

$$m \geqslant \mathcal{O}\left(\frac{d'\mathfrak{S}^2}{\epsilon^2}\right).$$

- In the k-means setting, $\mathfrak{S} = \mathcal{O}(k)$, $d' = kd\log k$, yielding $m \geqslant \mathcal{O}\left(\frac{dk^3 \log k}{\epsilon^2}\right)$.

[1] Langberg and Schulman, *Universal ϵ-approximators for integrals*, SIAM, 2010

In practice?

- In practice, computing the sensitivity is either i/ impossible, or ii/ costs more than solving the initial problem on the whole data set.

[1] see Feldman and Langberg, *A unified framework for...*, ACM symp. on Theory of computing, 2011 or Bachem et al., *Practical Coreset Constructions for Machine Learning*, Arxiv, 2017

[2] Braverman et al., *New frameworks for offline and streaming coreset constructions*, Arxiv, 2016

In practice?

- In practice, computing the sensitivity is either i/ impossible, or ii/ costs more than solving the initial problem on the whole data set.
- To circumvent this problem, upper bounds (easier to estimate) are used[1].
- Even then, finding algorithms that estimate useful upper bounds faster than the time needed to solve the problem on the whole dataset, remains a challenge.
- *N.B.* Those are not the current best sensitivity-based iid theorems[2]

[1] see Feldman and Langberg, *A unified framework for...*, ACM symp. on Theory of computing, 2011 or Bachem et al., *Practical Coreset Constructions for Machine Learning*, Arxiv, 2017

[2] Braverman et al., *New frameworks for offline and streaming coreset constructions*, Arxiv, 2016

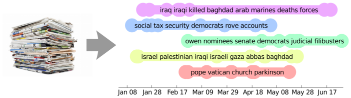
DPPs in a nutshell

Determinantal point processes (or DPP) are:

- *random* processes that *induce diversity*.
- *tractable*.
- used for three main purposes:
  - i/ *produce diverse samples* of a large database
  - ii/ *use as a tool* in a variety of SP/ML contexts
  - iii/ *characterize* various observed phenomena.

i/ This sample diversity can be directly useful[1,2]:

summary generation:

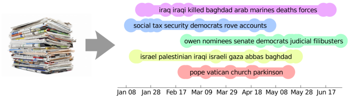


[1] left figure: from Kulesza and Taskar, *DPPs for machine learning*, Found. and Trends in ML, 2013

[2] right figure: from G. Gautier's slides `guilgautier.github.io/pdfs/GaBaVa17_slides.pdf`

i/ This sample diversity can be directly useful[1,2]:

summary generation:



search engines / recommendation:

'bolt' query

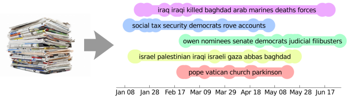
[1] left figure: from Kulesza and Taskar, *DPPs for machine learning*, Found. and Trends in ML, 2013

[2] right figure: from G. Gautier's slides guilgautier.github.io/pdfs/GaBaVa17_slides.pdf

i/ This sample diversity can be directly useful[1,2]:

search engines / recommendation:

summary generation:



'bolt' query

ii/ DPP samples can also be used as a tool in several SP/ML contexts:

- Monte Carlo integration
- Feature selection problems
- Coresets!
- etc.

[1] left figure: from Kulesza and Taskar, *DPPs for machine learning*, Found. and Trends in ML, 2013

[2] right figure: from G. Gautier's slides `guilgautier.github.io/pdfs/GaBaVa17_slides.pdf`

Determinantal Point Processes: formal definition and notations

- Let $\mathbf{L} \in \mathbb{R}^{n \times n}$ be a positive semi-definite matrix, where $\mathbf{L}_{ij}$ encodes some kind of interaction between $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$; *e.g.*, the Gaussian kernel $\mathsf{L}_{ij} = \exp^{-\frac{\left\|\boldsymbol{x}_i - \boldsymbol{x}_j\right\|^2}{2\tau^2}}$.

Determinantal Point Processes: formal definition and notations

- Let $\mathbf{L} \in \mathbb{R}^{n \times n}$ be a positive semi-definite matrix, where $\mathbf{L}_{ij}$ encodes some kind of interaction between $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$; *e.g.*, the Gaussian kernel $\mathsf{L}_{ij} = \exp^{-\frac{\|\boldsymbol{x}_i - \boldsymbol{x}_j\|^2}{2\tau^2}}$.
- Let m be a fixed integer and $\mathcal{S}$ a random subset of $\mathcal{X}$ of size m.
- We say that $\mathcal{S}$ is distributed according to a m-DPP, and write $\mathcal{S} \sim mDPP(\mathbf{L})$, if:

$$\mathbb{P}(\mathcal{S} = S) = \begin{cases} 0 & \text{if } |S| \neq m \\ \frac{1}{Z} \det(\mathsf{L}_S) & \text{if } |S| = m \end{cases}$$

where Z is a normalization constant.

Determinantal Point Processes: formal definition and notations

- Let $\mathbf{L} \in \mathbb{R}^{n \times n}$ be a positive semi-definite matrix, where $\mathbf{L}_{ij}$ encodes some kind of interaction between $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$; *e.g.*, the Gaussian kernel $\mathsf{L}_{ij} = \exp^{-\frac{\|\boldsymbol{x}_i - \boldsymbol{x}_j\|^2}{2\tau^2}}$.
- Let m be a fixed integer and $\mathcal{S}$ a random subset of $\mathcal{X}$ of size m.
- We say that $\mathcal{S}$ is distributed according to a m-DPP, and write $\mathcal{S} \sim mDPP(\mathbf{L})$, if:

$$\mathbb{P}(\mathcal{S} = S) = \begin{cases} 0 & \text{if } |S| \neq m \\ \frac{1}{Z} \det(\mathsf{L}_S) & \text{if } |S| = m \end{cases}$$

  where Z is a normalization constant.
- Denote by π_i the inclusion probability of $\boldsymbol{x}_i$:

$$\pi_i = \mathbb{P}(\boldsymbol{x}_i \in \mathcal{S}) = \frac{1}{Z} \sum_{S \text{ s.t. } i \in S, |S|=m} \det(\mathsf{L}_S).$$

- By construction, $\sum_i \pi_i = m$.

Questions

Consider the same class of minimization problems as previously. Say $\mathcal{S} \sim mDPP(\mathbf{L})$.

On the theoretical side (forgetting numerical efficiency for now):

- Under what conditions on $\mathbf{L}$ is $\mathcal{S}$ an ϵ-coreset with high probability?
- Can we *do better* than the iid case? (as strong coresets, but with smaller m?)
- What is the *optimal* $\mathbf{L}$?

Questions

Consider the same class of minimization problems as previously. Say $\mathcal{S} \sim mDPP(\mathbf{L})$.

On the theoretical side (forgetting numerical efficiency for now):

- Under what conditions on $\mathbf{L}$ is $\mathcal{S}$ an ϵ-coreset with high probability?
- Can we *do better* than the iid case? (as strong coresets, but with smaller m?)
- What is the *optimal* $\mathbf{L}$?

On the practical side (back IRL where we look for a practical implementation):

- In the iid world, the (sub-optimal but more realistic) strategies based on upper-bounding the sensitivity have a cost linear in n. For instance in the k-means context, they have a cost in $\mathcal{O}(nkd)$.
- Can we design a coreset algorithm based on DPPs that
  - outperforms in practice its iid couterpart
  - is not *ridiculously* longer than its iid counterpart?

In theory: two theorems

A first theorem (#9 in the paper) states that if $\pi_i = m\frac{\sigma_i}{\sigma}$ then we recover the iid performance. Frustratingly, this thm

i/ only proves that DPPs do not fare worse than iid

ii/ only provides conditions on $\{\pi_i\}$, nothing on higher order marginals (concentration tools well adapted to this case are hard to come by)

In theory: two theorems

A first theorem (#9 in the paper) states that if $\pi_i = m\frac{\sigma_i}{\sigma}$ then we recover the iid performance. Frustratingly, this thm

i/ only proves that DPPs do not fare worse than iid

ii/ only provides conditions on $\{\pi_i\}$, nothing on higher order marginals (concentration tools well adapted to this case are hard to come by)

A second theorem (#14 in the paper)

- Recall the ideal iid probability distribution: $p_i = \sigma_i/\mathfrak{S}$.
- Consider a PSD matrix L verifying:
  1. L is projective of rank m: $\mathsf{L} = \mathsf{U}\mathsf{U}^t$ with $\mathsf{U} \in \mathbb{R}^{n\times m}$ and $\mathsf{U}^t\mathsf{U} = \mathsf{I}_m$.
  2. $\forall i, \mathsf{L}_{ii} = mp_i$.

In theory: two theorems

A first theorem (#9 in the paper) states that if $\pi_i = m\frac{\sigma_i}{\sigma}$ then we recover the iid performance. Frustratingly, this thm

i/ only proves that DPPs do not fare worse than iid

ii/ only provides conditions on $\{\pi_i\}$, nothing on higher order marginals (concentration tools well adapted to this case are hard to come by)

A second theorem (#14 in the paper)

- Recall the ideal iid probability distribution: $p_i = \sigma_i/\mathfrak{S}$.
- Consider a PSD matrix L verifying:
  1. L is projective of rank m: $\mathrm{L} = \mathrm{U}\mathrm{U}^t$ with $\mathrm{U} \in \mathbb{R}^{n\times m}$ and $\mathrm{U}^t\mathrm{U} = \mathrm{I}_m$.
  2. $\forall i, \mathrm{L}_{ii} = mp_i$.
- **Lemma [via Schur-Horn]** *Such a matrix necessarily exists. In general, there are many degrees of freedom left to define* U*.*

In theory: two theorems

A first theorem (#9 in the paper) states that if $\pi_i = m\frac{\sigma_i}{\sigma}$ then we recover the iid performance. Frustratingly, this thm

i/ only proves that DPPs do not fare worse than iid

ii/ only provides conditions on $\{\pi_i\}$, nothing on higher order marginals (concentration tools well adapted to this case are hard to come by)

A second theorem (#14 in the paper)

- Recall the ideal iid probability distribution: $p_i = \sigma_i/\mathfrak{S}$.
- Consider a PSD matrix L verifying:
  1. L is projective of rank m: $\mathrm{L} = \mathrm{U}\mathrm{U}^t$ with $\mathrm{U} \in \mathbb{R}^{n\times m}$ and $\mathrm{U}^t\mathrm{U} = \mathrm{I}_m$.
  2. $\forall i, \mathrm{L}_{ii} = mp_i$.
- **Lemma [via Schur-Horn]** *Such a matrix necessarily exists. In general, there are many degrees of freedom left to define* U*.*
- **Theorem [Variance reduction theorem]** *Sample $\mathcal{S}_{iid}$ by drawing m samples iid from $\boldsymbol{p}$. Sample $\mathcal{S}_{dpp} \sim mDPP(\mathbf{L})$. One has:*

$$\forall \theta \in \Theta \qquad Var\left[\hat{L}(\mathcal{S}_{dpp}, \theta)\right] = Var\left[\hat{L}(\mathcal{S}_{iid}, \theta)\right] - \frac{m-1}{m} Y$$

where $Y \geqslant 0$ depends on intricate frame properties of the lines of U*.*

$\Rightarrow$ Such a DPP necessarily provides a better coreset than its iid counterpart.

$\Rightarrow$ Finding the *best* projective DPP is however out of (our) reach theoretically.

In practice: heuristics

Sampling from a DPP requires a worst-case $\mathcal{O}(n^3)$ number of operations. Low-rank DPPs have a more reasonable $\mathcal{O}(nm^2)$ complexity. We propose two DPP heuristics based on low-rank kernels:

Alg. 1: **Approximate Gaussian kernel (with parameter $\tau > 0$)**

- Compute $r \geqslant \mathcal{O}(m)$ Random Fourier Features[1] and obtain $\Psi \in \mathbb{R}^{n\times r}$ s.t. $\Psi\Psi^t \in \mathbb{R}^{n\times n}$ approximates the Gaussian kernel
- Sample an m-DPP from $\mathbf{L} = \Psi\Psi^t$
- ✓ This runs in $\mathcal{O}(nm^2 + nmd)$
- × τ is a (annoying) hyper-parameter.

Alg. 2: **Vandermonde kernel (here for $d = 1$, can be extended to $d \geqslant 2$)**

- Compute $\mathrm{V} \in \mathbb{R}^{n\times m}$ the partial Vandermonde matrix $\mathrm{V}_{ij} = x_i^{j-1}$
- Sample an m-DPP from $\mathbf{L} = \mathrm{V}\mathrm{V}^t$ (it is a projective DPP)
- ✓ This runs in $\mathcal{O}(nm^2)$
- ✓ No hyper-parameter tuning
- × For $d \geqslant 2$, not all values of m are admissible.

[1] Rahimi et al., *Random features for large-scale kernel machines*, NIPS, 2008

[3] Barthelmé, Tremblay, Usevich, Amblard. *Determinantal point processes in the flat limit*, Bernoulli, 2022

In practice: heuristics

Sampling from a DPP requires a worst-case $\mathcal{O}(n^3)$ number of operations. Low-rank DPPs have a more reasonable $\mathcal{O}(nm^2)$ complexity. We propose two DPP heuristics based on low-rank kernels:

Alg. 1: **Approximate Gaussian kernel (with parameter $\tau > 0$)**

- Compute $r \geqslant \mathcal{O}(m)$ Random Fourier Features[1] and obtain $\mathbf{\Psi} \in \mathbb{R}^{n\times r}$ s.t. $\mathbf{\Psi}\mathbf{\Psi}^t \in \mathbb{R}^{n\times n}$ approximates the Gaussian kernel
- Sample an m-DPP from $\mathbf{L} = \mathbf{\Psi}\mathbf{\Psi}^t$
- ✓ This runs in $\mathcal{O}(nm^2 + nmd)$
- × τ is a (annoying) hyper-parameter.

Alg. 2: **Vandermonde kernel (here for $d = 1$, can be extended to $d \geqslant 2$)**

- Compute $\mathrm{V} \in \mathbb{R}^{n\times m}$ the partial Vandermonde matrix $\mathrm{V}_{ij} = x_i^{j-1}$
- Sample an m-DPP from $\mathbf{L} = \mathrm{V}\mathrm{V}^t$ (it is a projective DPP)
- ✓ This runs in $\mathcal{O}(nm^2)$
- ✓ No hyper-parameter tuning
- × For $d \geqslant 2$, not all values of m are admissible.

- Advertisement: both kernels become equivalent in the flat limit $(\tau \to \infty)$[3].

[1] Rahimi et al., *Random features for large-scale kernel machines*, NIPS, 2008

[3] Barthelmé, Tremblay, Usevich, Amblard. *Determinantal point processes in the flat limit*, Bernoulli, 2022

In practice: illustration on 1-means

- Data $\mathcal{X}$, parameter θ
- Cost func.

$$L(\mathcal{X}, \theta) = \sum_{i=1}^{n} \|\boldsymbol{x}_i - \theta\|^2$$

In practice: illustration on 1-means

- Data $\mathcal{X}$, parameter θ
- Cost func.

$$L(\mathcal{X}, \theta) = \sum_{i=1}^{n} \|\boldsymbol{x}_i - \theta\|^2$$

Compare:

- `uniform iid` sampling
- `sensitivity iid`: ideal iid sampling based on exact sensitivities
- m-`DPP` (heuristic) based on RFFs of the Gaussian L-ensemble
- `Proj Poly DPP` (heuristic) based on the partial Vandermonde matrix

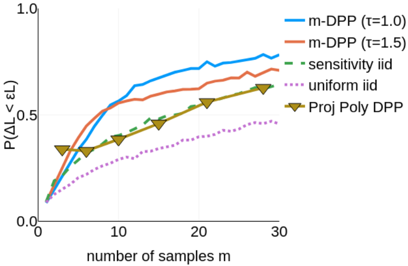
In practice: illustration on 1-means

- Data $\mathcal{X}$, parameter θ
- Cost func.

$$L(\mathcal{X},\theta)=\sum_{i=1}^{n}\|\boldsymbol{x}_i-\theta\|^2$$

Compare:

- `uniform iid` sampling
- `sensitivity iid`: ideal iid sampling based on exact sensitivities
- m-DPP (heuristic) based on RFFs of the Gaussian L-ensemble
- `Proj Poly DPP` (heuristic) based on the partial Vandermonde matrix

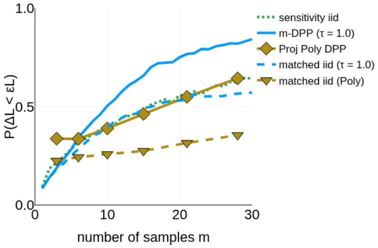


In practice: illustration on 1-means

- `sensitivity iid`: as before
- m-DPP ($\tau = 1$): as before
- `Proj Poly DPP`: as before
- `matched iid` ($\tau = 1$): iid version of m-DPP ($\tau = 1$)
- `matched iid (Poly)`: iid version of `Proj Poly DPP`

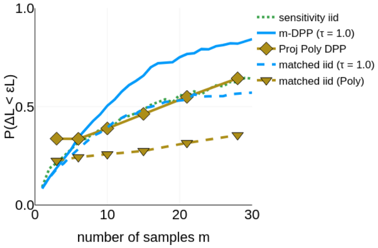
In practice: illustration on 1-means

- `sensitivity iid`: as before
- m-`DPP` ($\tau = 1$): as before
- `Proj Poly DPP`: as before
- `matched iid` ($\tau = 1$): iid version of m-`DPP` ($\tau = 1$)
- `matched iid (Poly)`: iid version of `Proj Poly DPP`



In practice: illustration on 1-means

- `sensitivity iid`: as before
- m-`DPP` $(\tau = 1)$: as before
- `Proj Poly DPP`: as before
- `matched iid` $(\tau = 1)$: iid version of m-`DPP` $(\tau = 1)$
- `matched iid (Poly)`: iid version of `Proj Poly DPP`



As the dimension d increases:

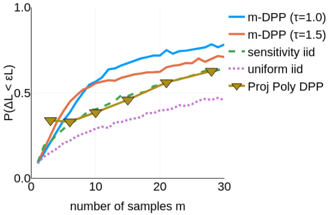


Figure: $d = 2$

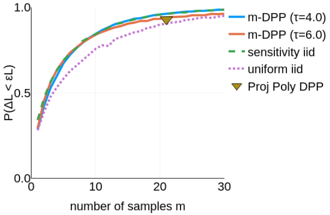


Figure: $d = 20$

Conclusion: take home messages

- This is exploratory work on the simple question: can DPPs help create better coresets? If so, how?
- We have a few (mainly frustrating) theorems stating that DPPs do not fare worse than its iid counterpart. The strongest result is the variance reduction theorem.
- We propose 2 DPP-based heuristics (= no provable guarantees), running in $\mathcal{O}(nm^2)$
- In the k-means and linear regression problems, these heuristics outperform (= better coresets for a similar computation time) the iid scheme especially:
  - for small m : to keep the $\mathcal{O}(nm^2)$ DPP sampling cost under control
  - and small d : to keep the DPP's repulsiveness significant.
- For (many) more theoretical and experimental details, the paper is available at: `http://jmlr.org/papers/volume20/18-167/18-167.pdf`
- The DPP4Coresets Julia toolbox is available at: `https://gricad-gitlab.univ-grenoble-alpes.fr/tremblan/dpp4coresets.jl` or on my website.