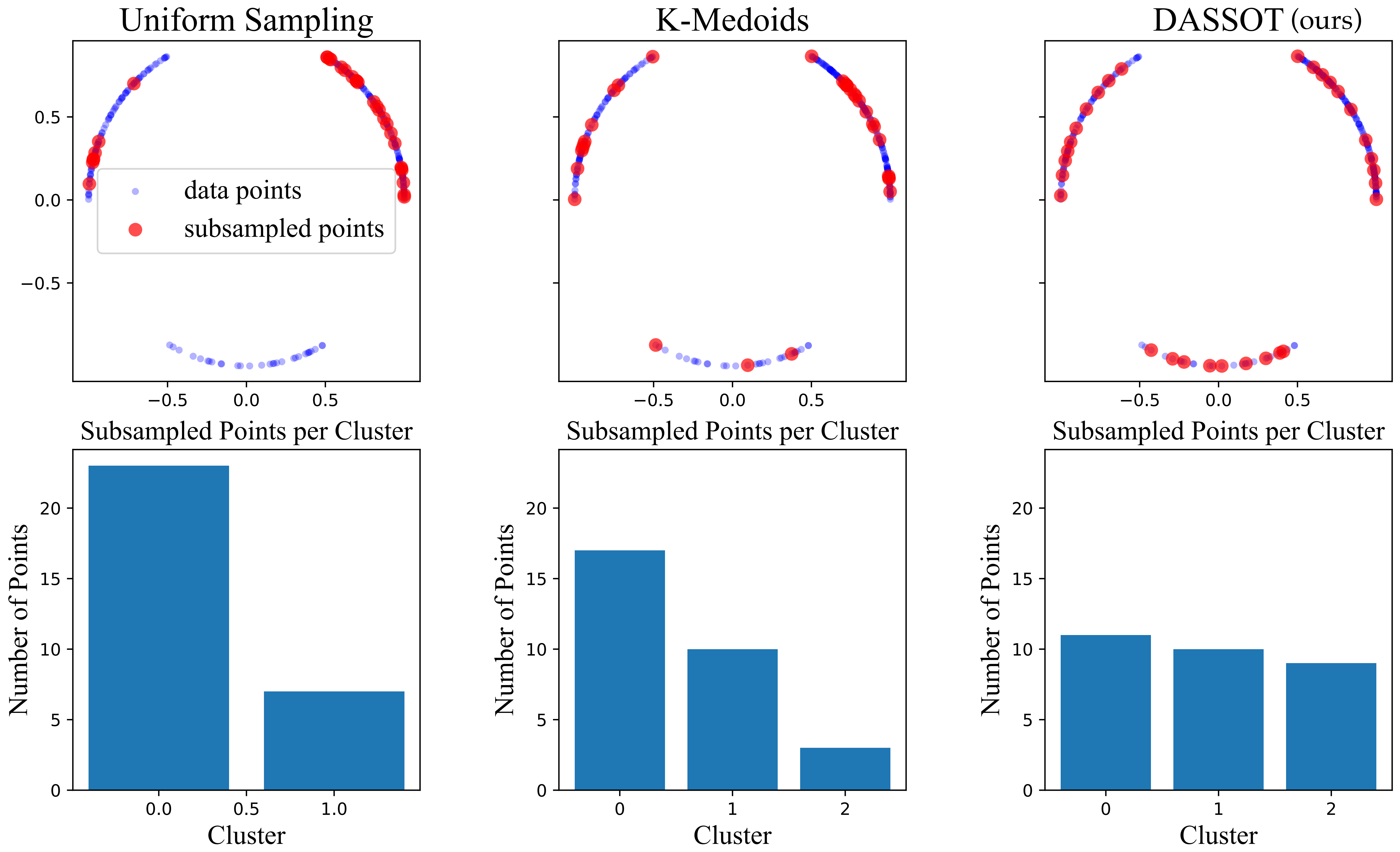
A Graph Matching Approach to Balanced Data Sub-Sampling for SSL

ENS DE LYON

BROWN

Hugues Van Assel, Randall Balestriero

Problem: ensuring that each concept is represented by a similar number of samples is crucial for SSL performance. Balanced data sub-sampling aims to extract a subset of data where concepts are evenly represented. **Existing work:** common approaches rely on clustering algorithms (k-means, k-medoids etc.) and their centroids. However, these centroids favor dominant concepts and are not suited for balanced data-subsampling.



Intuition: our method selects points with low pairwise similarities to make the selected subset as diverse as possible.

DASSOT

$$\min_{\mathbf{T}\mathbf{1}_N=\mathbf{1}_n,\ \mathbf{T}\geq 0} \mathcal{L}(\mathbf{T}) := \sum_{ijkl} \left([\mathbf{D}_n]_{ij} - [\mathbf{S}_x]_{kl}\right)^2 T_{ik}T_{jl} + \gamma\,\mathrm{KL}\left(\mathbf{T}^\top \mathbf{1}_n \middle\| \frac{n}{N}\mathbf{1}_N\right)$$

Select n points among N — **Graph matching term** — **Ensure diversity**

$$[\mathbf{D}_n]_{ij} = \begin{cases} 1 & \text{if } i = j \text{ (maximum similarity)}, \\ -1 & \text{if } i \neq j \text{ (minimum similarity)}. \end{cases} \quad \& \quad [\mathbf{S}_\mathbf{x}]_{ij} = \langle \tilde{\mathbf{x}}_i, \tilde{\mathbf{x}}_j \rangle, \text{ where } \tilde{\mathbf{x}}_i = \frac{\mathbf{x}_i}{\|\mathbf{x}_i\|_2}$$

Mirror descent

Until convergence:

$$\mathbf{K}^{(i)} \leftarrow \exp\left(\nabla_\mathbf{T}\mathcal{L}(\mathbf{T}^{(i)}) - \varepsilon \log(\mathbf{T}^{(i)})\right)$$

$$\mathbf{T}^{(i+1)} \leftarrow \mathrm{diag}\left(\mathbf{1}_n \oslash (\mathbf{K}^{(i)}\mathbf{1}_N)\right)\mathbf{K}^{(i)}$$

- ◆ Complexity O(N) in time and complexity.
- ◆ About 100 iterations to reach convergence.
- ◆ GPU friendly.

Experiments: balancing input data improves the performances of SimCLR when evaluated on a balanced test dataset.

Dataset	α	n	k-Means	k-Medoid	DASSOT
CIFAR-10	1.2	5000	81.8	81.9	**82.7**
-	1.2	10000	**85.7**	85.5	85.5
-	1.5	5000	59.3	58.7	**62.9**
-	1.5	10000	71.6	71.6	**73.2**
CIFAR-100	1.2	5000	55.2	55.5	**56.2**
-	1.2	10000	60.8	**61.3**	61.1
-	1.5	5000	43.9	44.2	**48.6**
-	1.5	10000	51.9	**52.7**	**52.7**

strength of Imbalance