A Graph Matching Approach to Balanced Data Sub-Sampling for Self-Supervised Learning

Hugues Van Assel
ENS Lyon
hugues.van_assel@ens-lyon.fr

Randall Balestriero
Brown University
rbalestr@brown.edu

Abstract

Real-world datasets often display inherent imbalances in the distribution of classes or concepts. Recent studies indicate that such imbalances can lead to suboptimal performances of Self-Supervised Learning (SSL) models when evaluated across the full spectrum of concepts. To address this issue, we propose a data curation method that automatically selects a balanced subset of the data. This problem is approached as a graph matching task, where the goal is to identify a data subset that is most distinct in terms of pairwise similarities. We achieve this by mapping an isolated graph onto the similarity graph of the input data, leveraging the optimal transport semi-unbalanced Gromov-Wasserstein problem. We demonstrate that this problem can be solved with linear complexity and is well-suited for GPU acceleration. The effectiveness of our method is validated through experiments on small datasets, setting the stage for future exploration on larger-scale problems.

1 Introduction

Data curation is a critical component in modern Self-Supervised Learning (SSL) workflows [4]. This process typically includes the preparation and refinement of the training data to enhance the model's performance and mitigate existing biases. A recent study [15] emphasized three critical factors for constructing an effective dataset in SSL: dataset size, diversity, and balance. Among these, achieving balance is particularly challenging yet crucial. Commonly used SSL datasets are generally balanced in the sense that each concept or class label is represented by an equal number of samples. This contrasts with real-world data where concepts often follow a long-tailed power-law distribution [12]. Research has shown that training SSL models on long-tailed datasets dominated by a few common concepts leads to significant performance drops [10], which challenges the broader application of these methods. Ensuring a balanced distribution of concepts (or classes) in the training data is therefore essential to mitigate this bias. Most existing balancing techniques, however, rely heavily on some form of labeling or supervision [17, 3]. Since this level of supervision is typically unavailable in most real world scenarios, there is a pressing need for the development of unsupervised or minimally supervised approaches to attain balanced datasets. Addressing this challenge is the primary objective of this work.

Background. A common method for balanced data subsampling is to use clustering algorithms, where the subsampled data can be represented by centroids from k-means or k-medoids for instance. However, it is well known that centroids are biased toward dominant concepts, as these concepts tend to occupy more centroids than less frequent ones. A significant result by [18] demonstrates that in high dimensions, k-means centroids asymptotically follow the data distribution, thereby preserving the same imbalance. Some approaches have been proposed to counteract this bias. For example, [5] introduced radius upper bounds on the centroids, while [15] suggested a multi-stage clustering and resampling scheme. In this work, we take a different path by not relying on clustering. In fact, our

NeurIPS 2024 Workshops on Self-supervised Learning - Theory and Practice.

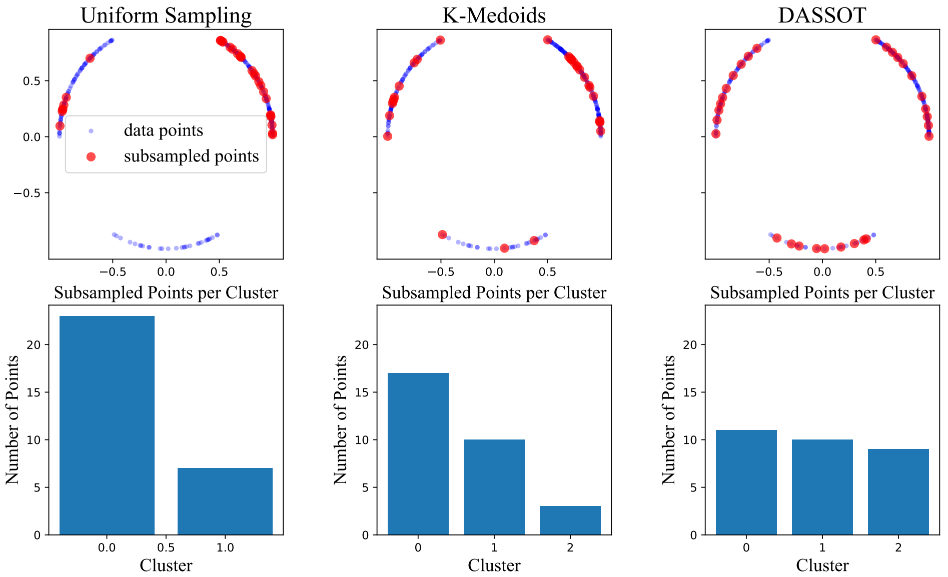

Figure 1: Comparison of different subsampling strategies on a simulated dataset with 3 imbalanced clusters on the $2D$ sphere. Left: Points are selected uniformly at random from the dataset, preserving the initial imbalance. Middle: The K-Medoids clustering method shows that medoids retain a similar imbalance to the original data. Right: Our proposed method (Eq. DASSOT) selects points that are maximally separated, leading to a more balanced representation of the data distribution. Note that a slight imbalance remains, primarily due to the larger support of the more populated clusters.

method can be viewed as a reverse clustering strategy (see Remark 2.1), where instead of merging points with high similarity, we focus on identifying points that are maximally separated.

Contributions. In this work, we propose a novel method for selecting a subset of n data points that optimally captures the diversity of the dataset. Our approach is agnostic to downstream tasks and solely requires a feature space with a meaningful similarity measure. It leverages a graph matching formulation to identify a subset whose similarity structure approximates that of a disconnected graph. Doing so, the method aims to select points that are maximally separated, ensuring the subset uniformly represents the underlying data distribution. We introduce a single-loop, GPU-friendly algorithm with linear time and memory complexity relative to the number of samples. We then show on CIFAR that our method effectively re-balances imbalanced datasets and improves the performance of the self-supervised learning SimCLR model.

2 Method

We consider a similarity matrix $\mathbf{S}_\mathbf{X} \in \mathbb{R}^{N \times N}$, which captures the pairwise similarities between the samples $(\mathbf{x}_1, \dots, \mathbf{x}_N)^\top \in \mathbb{R}^{N \times p}$, each of dimensionality p. In this work, we use cosine similarity, meaning that for any pair (i, j), the similarity is defined as $[\mathbf{S}_\mathbf{X}]_{ij} = \langle \tilde{\mathbf{x}}_i, \tilde{\mathbf{x}}_j \rangle$, where $\tilde{\mathbf{x}}_i = \frac{\mathbf{x}_i}{\|\mathbf{x}_i\|_2}$ denotes the normalized sample $\mathbf{x}_i$. Our goal is to select a subset of n samples such that their corresponding pairwise similarity matrix, which is a submatrix of $\mathbf{S}_\mathbf{X}$, closely approximates the disconnected similarity matrix $\mathbf{D}_n$. The diagonal elements of $\mathbf{D}_n$ are set to 1 (indicating maximum similarity), while its off-diagonal elements are set to -1 (indicating minimum similarity). The key intuition behind this is to make the selected subset as diverse as possible by ensuring that the pairwise similarities between the selected points are low.

Problem formulation. Our method, coined *DASSOT* for *Data Sub-Sampling with Optimal Transport*, is based on the following optimization problem:

$$\min_{\mathbf{T}\mathbf{1}_N=\mathbf{1}_n;\mathbf{T}\geq 0} \quad \mathcal{L}(\mathbf{T}) := \sum_{ijkl}([\mathbf{D}_n]_{ij} - [\mathbf{S}_\mathbf{X}]_{kl})^2 T_{ik}T_{jl} + \gamma \mathrm{KL}(\mathbf{T}^\top \mathbf{1}_n|\mathbf{h}) \qquad \text{(DASSOT)}$$

where $\mathbf{h} = \frac{n}{N}\mathbf{1}_N$, $\gamma > 0$ is a hyperparameter and $\mathrm{KL}(\mathbf{a}|\mathbf{b}) = \sum_i \mathrm{KL}(a_i|b_i)$ is the Kullback-Leibler divergence with $\mathrm{KL}(a|b) = a\log(a/b) - a + b$ for two positive scalars a and b. In the above, $\mathbf{T} \in \mathbb{R}_+^{n\times N}$ can be seen as a soft mapping between the rows of $\mathbf{D}_n$ and $\mathbf{S}_\mathbf{X}$. The final selected points are then given by $\left(\max_{j\in[\![N]\!]} T^\star_{i,j}\right)_{i\in[\![n]\!]}$ where $\mathbf{T}^\star$ solves DASSOT.

Interestingly, the first term on the right-hand side of the above objective vanishes if there exists a subset I of $[\![N]\!]$ with cardinality n, such that the submatrix $\mathbf{S}_\mathbf{X}^I \in \mathbb{R}^{n\times n}$, restricted to the indices in I, is exactly equal to $\mathbf{D}_n$. Indeed, in this case for any bijection $\sigma : [\![n]\!] \to I$, one can construct $\mathbf{T}^\star$ such that for any i, $T^\star_{i,\sigma(i)} = 1$ and $T^\star_{i,j} = 0$ for all $j \neq \sigma(i)$. Optimization over the constraints $\{\mathbf{T}\mathbf{1}_N = \mathbf{1}_n; \mathbf{T} \geq 0\}$ can therefore be viewed as a relaxation of this subgraph identification problem, where the goal is to find a probabilistic matching, or *coupling*, between the rows of the two similarity matrices. This relaxation allows the development of efficient heuristics and is well-known in the optimal transport literature. In fact, this problem is a specific case of the unbalanced Gromov-Wasserstein problem [8, 7, 9]. More specifically, this is a *semi-unbalanced Gromov-Wasserstein* problem, as the first marginal $\mathbf{T}\mathbf{1}_N$ is fixed, since we aim to select exactly n points.

The second term, $\gamma\mathrm{KL}(\mathbf{T}^\top\mathbf{1}_n|\mathbf{h})$, is the unbalanced relaxation of the second marginal $\mathbf{T}^\top\mathbf{1}_n$. As $\gamma \to +\infty$, this term enforces that all data points receive equal mass, which conflicts with our objective of selecting a subset of the data. However, for smaller values of γ, this regularizer prevents the coupling matrix $\mathbf{T}$ from collapsing onto only a few data points. This collapse can happen in certain constrained geometries where the first term pushes to maximize large-scale similarities at the expense of smaller ones.

Remark 2.1. (Relation to clustering) [16] demonstrated that state-of-the-art performance in clustering can be achieved by transporting via Gromov-Wasserstein the data similarity onto the isolated similarity, which serves as an ideally clustered template graph. Additionally, [2] and [11] have uncovered theoretical connections between this approach and spectral clustering. In contrast, DASSOT follows the opposite strategy by transporting the isolated similarity onto the data similarity, allowing us to uniformly sample points from the underlying data distribution.

Algorithm. We propose to solve DASSOT using the mirror-descent algorithm with respect to the KL geometry. This approach offers non-asymptotic convergence guarantees towards a stationary point of the objective [13]. The method iteratively updates $\mathbf{T}$ according to the following optimization step: $\mathbf{T}^{(i+1)} \leftarrow \arg\min_{\mathbf{T}\mathbf{1}_N=\mathbf{1}_n,\mathbf{T}\geq 0}\langle \mathbf{T}, \nabla_\mathbf{T}\mathcal{L}(\mathbf{T}^{(i)})\rangle + \varepsilon\mathrm{KL}(\mathbf{T}|\mathbf{T}^{(i)})$ where $\varepsilon > 0$ is a tunable hyperparameter and $\mathrm{KL}(\mathbf{A}|\mathbf{B}) = \sum_{ij}\mathrm{KL}(A_{ij}|B_{ij})$. The latter simplifies to the following explicit update rule, as described in [14], which is particularly suited for GPU computations

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

where $\mathbf{K}^{(i)} = \exp\left(\nabla_\mathbf{T}\mathcal{L}(\mathbf{T}^{(i)}) - \varepsilon\log(\mathbf{T}^{(i)})\right)$ and the expression of $\nabla_\mathbf{T}\mathcal{L}$ is given in equation 3.

Complexity. We now present the details of computing the gradient $\nabla_\mathbf{T}\mathcal{L}$, which constitutes the most computationally intensive part of the algorithm. Removing constant terms the above problem DASSOT can be rewritten as

$$\min_{\mathbf{T}\mathbf{1}_N=\mathbf{1}_n;\mathbf{T}\geq 0} \quad \langle \mathbf{1}_n\mathbf{1}_n^\top\mathbf{T}\mathbf{S}_\mathbf{X}^{\odot 2} - 2\mathbf{D}_n\mathbf{T}\mathbf{S}_\mathbf{X}, \mathbf{T}\rangle + \left(\mathbf{T}^\top\mathbf{1}_n\right)\log\left(\left(\mathbf{T}^\top\mathbf{1}_n\right)\oslash\mathbf{h}\right) - \left(\mathbf{T}^\top\mathbf{1}_n\right) \ . \tag{2}$$

The gradient of the above w.r.t $\mathbf{T}$ is given by:

$$\nabla_\mathbf{T}\mathcal{L}(\mathbf{T}) = \mathbf{1}_n\mathbf{1}_n^\top\mathbf{T}\mathbf{S}_\mathbf{X}^{\odot 2} - 2\mathbf{D}_n\mathbf{T}\mathbf{S}_\mathbf{X} + \mathbf{1}_n\log\left(\left(\mathbf{T}^\top\mathbf{1}_n\right)\oslash\mathbf{h}\right)^\top \ . \tag{3}$$

The key components in the above gradient are the matrix-matrix products $\mathbf{T}\mathbf{S}_\mathbf{X}^{\odot 2}$ and $\mathbf{T}\mathbf{S}_\mathbf{X}$, which, at first glance, seem to require $\mathcal{O}(N^2 \times n)$ operations. Fortunately, we can exploit the low-rank structure of $\mathbf{S}_\mathbf{X}$ to reduce this computational cost. Specifically, since $\mathbf{S}_\mathbf{X}$ is a cosine similarity matrix, it can be decomposed as $\mathbf{S}_\mathbf{X} = \widetilde{\mathbf{X}}\widetilde{\mathbf{X}}^\top$, where $\widetilde{\mathbf{X}} = (\tilde{\mathbf{x}}_1, \ldots, \tilde{\mathbf{x}}_N)^\top \in \mathbb{R}^{N\times p}$. This allows us to compute $\mathbf{T}\mathbf{S}_\mathbf{X}$ as $(\mathbf{T}\widetilde{\mathbf{X}})\widetilde{\mathbf{X}}^\top$ in $\mathcal{O}(N \times n \times p)$ operations. Similarly, we can derive a low-rank factorization for $\mathbf{S}_\mathbf{X}^{\odot 2}$. Note that for any vectors $\mathbf{a}$ and $\mathbf{b}$ in $\mathbb{R}^p$, it holds that $\langle \mathbf{a}, \mathbf{b}\rangle^2 = \langle \mathbf{a}\mathbf{a}^\top, \mathbf{b}\mathbf{b}^\top\rangle$. Using this property, we can express $\mathbf{S}_\mathbf{X}^{\odot 2}$ as $\mathbf{\Psi}_\mathbf{X}\mathbf{\Psi}_\mathbf{X}^\top$, where $\mathbf{\Psi}_\mathbf{X} = (\mathrm{vec}(\tilde{\mathbf{x}}_1\tilde{\mathbf{x}}_1^\top), \ldots, \mathrm{vec}(\tilde{\mathbf{x}}_N\tilde{\mathbf{x}}_N^\top))^\top \in \mathbb{R}^{N\times p^2}$. This formulation allows us to compute $\mathbf{T}\mathbf{S}_\mathbf{X}^{\odot 2}$ as $(\mathbf{T}\mathbf{\Psi}_\mathbf{X})\mathbf{\Psi}_\mathbf{X}^\top$ in $\mathcal{O}(N \times n \times p^2)$ operations. The overall computational complexity of the gradient is $\mathcal{O}(N \times n \times p^2)$. Consequently, it scales linearly with the number of samples N, under the assumption that $p^2 \ll N$.

Dataset	α	n	k-Means	k-Medoid	DASSOT
CIFAR-10	1.2	5000	279.5 (0.4)	296.0 (0.3)	**239.3** (8.3)
-	1.2	10000	594.7 (0.5)	595.4 (0.6)	**522.4** (11.2)
-	1.5	5000	561.3 (0.4)	577.3 (0.3)	**463.5** (7.7)
-	1.5	10000	1141.0 (0.5)	1148.4 (0.5)	**1013.2** (12.8)
CIFAR-100	1.2	5000	260.2 (0.3)	251.8 (0.3)	**243.4** (6.7)
-	1.2	10000	505.6 (0.7)	496.7 (0.8)	**451.4** (8.7)
-	1.5	5000	444.1 (0.4)	443.9 (0.4)	**436.2** (8.1)
-	1.5	10000	666.5 (0.5)	666.5 (0.8)	**628.5** (9.2)

Table 1: Standard deviation of the number of points per class in the subsampled dataset for each configuration. Lower standard deviation values suggest a more balanced distribution of points across classes. The runtimes are reported between parentheses in seconds.

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**

Table 2: Top-1 accuracy of the SimCLR model trained on the subsampled dataset for the different configurations.

3 Experiments

To validate the effectiveness of our method, we conduct experiments on the CIFAR-10 and CIFAR-100 [6] datasets. We preprocess the data to create an exponential class imbalance by sampling as follows: for the k-th class, we sample $N_k = N \times e^{-k\log(\alpha)}$ samples where N is the number of samples per class. We test two levels of imbalance for each dataset: $\alpha = 1.2$ and $\alpha = 1.5$. We then apply DASSOT to select a balanced subset of n samples. For DASSOT, we validate both ε and γ in the set $\{1, 10, 100, 1000\}$ and initialize $\mathbf{T}$ with the uniform plan $\frac{1}{N}\mathbf{1}_n\mathbf{1}_N^\top$. We experiment with $n = 5000$ and $n = 10000$ samples, and compare our method to the k-Means and k-Medoids clustering approaches, each taken by considering the best runs among 10 random initializations. The latter two approaches are applied using the cosine similarity matrix $\mathbf{S_X}$ as metric. It is important to note that k-Medoid generates a subset of n samples directly from the original dataset, whereas k-Means computes barycenter points, which do not correspond to actual samples in the dataset. Instead of working with raw pixel data, we utilize features extracted by a randomly initialized ResNet50 model.

We assess both the ability of DASSOT to achieve a balanced dataset and the downstream performance of the subsampled data when used with the SimCLR model [1]. For our SimCLR experiments, we follow the same hyperparameters as outlined in the original SimCLR paper. We use a ResNet-50 model, a batch size of 256, a learning rate of 0.5, and apply a cosine learning rate schedule. The model is trained for 500 epochs, and we report the top-1 accuracy on the regular balanced test set. The results, shown in Table 1 and Table 2, demonstrate that DASSOT achieves superior class balancing compared to k-Means and k-Medoids, leading to higher test accuracy for the SimCLR model in most considered scenarios.

4 Opening Remarks

In this paper, we present an initial exploration of a graph matching strategy designed to select a subset of data points that exhibit high diversity based on pairwise similarities. Moving forward, we aim to further investigate the empirical benefits of this method. We anticipate that it will be especially useful in scenarios where data clusters are not clearly defined, such as when there is a gradual transition between different concepts. Additionally, we plan to explore improved techniques for constructing the data similarity matrix by leveraging features from the SSL model itself, allowing for iterative refinement of the data selection process.

References

[1] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020.

[2] S. Chowdhury and T. Needham. Generalized spectral clustering via gromov-wasserstein learning. In *International Conference on Artificial Intelligence and Statistics*, pages 712–720. PMLR, 2021.

[3] M. Dehghani, J. Djolonga, B. Mustafa, P. Padlewski, J. Heek, J. Gilmer, A. P. Steiner, M. Caron, R. Geirhos, I. Alabdulmohsin, et al. Scaling vision transformers to 22 billion parameters. In *International Conference on Machine Learning*, pages 7480–7512. PMLR, 2023.

[4] E. Grave, P. Bojanowski, P. Gupta, A. Joulin, and T. Mikolov. Learning word vectors for 157 languages. *arXiv preprint arXiv:1802.06893*, 2018.

[5] A. I. Humayun, R. Balestriero, A. Kyrillidis, and R. Baraniuk. No more than 6ft apart: Robust k-means via radius upper bounds. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4433–4437. IEEE, 2022.

[6] A. Krizhevsky, G. Hinton, et al. Learning multiple layers of features from tiny images. 2009.

[7] F. Mémoli. Gromov–wasserstein distances and the metric approach to object matching. *Foundations of computational mathematics*, 11:417–487, 2011.

[8] T. Séjourné, F.-X. Vialard, and G. Peyré. The unbalanced gromov wasserstein distance: Conic formulation and relaxation. *Advances in Neural Information Processing Systems*, 34:8766–8779, 2021.

[9] K.-T. Sturm. On the geometry of metric measure spaces. 2006.

[10] Y. Tian, O. J. Henaff, and A. Van den Oord. Divide and contrast: Self-supervised learning from uncurated data. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10063–10074, 2021.

[11] H. Van Assel, C. Vincent-Cuaz, N. Courty, R. Flamary, P. Frossard, and T. Vayer. Distributional reduction: Unifying dimensionality reduction and clustering with gromov-wasserstein projection. *arXiv preprint arXiv:2402.02239*, 2024.

[12] G. Van Horn, O. Mac Aodha, Y. Song, Y. Cui, C. Sun, A. Shepard, H. Adam, P. Perona, and S. Belongie. The inaturalist species classification and detection dataset. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8769–8778, 2018.

[13] C. Vincent-Cuaz. *Optimal transport for graph representation learning*. PhD thesis, Université Côte d'Azur, 2023.

[14] C. Vincent-Cuaz, R. Flamary, M. Corneli, T. Vayer, and N. Courty. Semi-relaxed gromov-wasserstein divergence with applications on graphs. *arXiv preprint arXiv:2110.02753*, 2021.

[15] H. V. Vo, V. Khalidov, T. Darcet, T. Moutakanni, N. Smetanin, M. Szafraniec, H. Touvron, C. Couprie, M. Oquab, A. Joulin, et al. Automatic data curation for self-supervised learning: A clustering-based approach. *arXiv preprint arXiv:2405.15613*, 2024.

[16] H. Xu, D. Luo, and L. Carin. Scalable gromov-wasserstein learning for graph partitioning and matching. *Advances in neural information processing systems*, 32, 2019.

[17] H. Xu, S. Xie, X. E. Tan, P.-Y. Huang, R. Howes, V. Sharma, S.-W. Li, G. Ghosh, L. Zettlemoyer, and C. Feichtenhofer. Demystifying clip data. *arXiv preprint arXiv:2309.16671*, 2023.

[18] P. Zador. Asymptotic quantization error of continuous signals and the quantization dimension. *IEEE Transactions on Information Theory*, 28(2):139–149, 1982.