Double-Bounded Optimal Transport for Advanced Clustering and Classification

Liangliang Shi, Zhaoqi Shen, Junchi Yan

Department of Computer Science and Engineering, MoE Key Lab of Artificial Intelligence, Shanghai Jiao Tong University
{shiliangliang, shenzhaoqi2271621336, yanjunchi}@sjtu.edu.cn

Abstract

Optimal transport (OT) is attracting increasing attention in machine learning. It aims to transport a source distribution to a target one at minimal cost. In its vanilla form, the source and target distributions are predetermined, which contracts to the real-world case involving undetermined targets. In this paper, we propose Doubly Bounded Optimal Transport (DB-OT), which assumes that the target distribution is restricted within two boundaries instead of a fixed one, thus giving more freedom for the transport to find solutions. Based on the entropic regularization of DB-OT, three scaling-based algorithms are devised for calculating the optimal solution. We also show that our DB-OT is helpful for barycenter-based clustering, which can avoid the excessive concentration of samples in a single cluster. Then we further develop DB-OT techniques for long-tailed classification which is an emerging and open problem. We first propose a connection between OT and classification, that is, in the classification task, training involves optimizing the Inverse OT to learn the representations, while testing involves optimizing the OT for predictions. With this OT perspective, we first apply DB-OT to improve the loss, and the Balanced Softmax is shown as a special case. Then we apply DB-OT for inference in the testing process. Even with vanilla Softmax trained features, our extensive experimental results show that our method can achieve good results with our improved inference scheme in the testing stage.

Introduction

Optimal transport (OT) (Cuturi 2013) has been widely applied in machine learning (Cui et al. 2019a; Wang et al. 2013). For instance, Wasserstein distance (Arjovsky, Chintala, and Bottou 2017; Gulrajani et al. 2017) is applied with the dual form of OT to minimize the gap between the generated and real distributions via min-max adversarial optimization (Li et al. 2022). SwAV (Caron et al. 2020) employs the Sinkhorn algorithm for online clustering in self-supervised contrastive learning (Chen et al. 2020; Khosla et al. 2020; Wang and Liu 2021). OT-LDA (Huynh, Zhao, and Phung 2020) devises a Wasserstein barycenter-based topic model, while in (Xu et al. 2019), Gromov-Wasserstein distance is used for graph matching (Wang, Yan, and Yang 2019, 2021; Hu et al. 2020). These works all assume the source and target distributions are fixed.

However, in many real-world cases, the source or target distribution often varies over time and becomes uncertain. As shown in Fig. 1, in vanilla OT, it assumes that the supply from mines and the demand from factories are fixed and their total masses are equal to each other. While in our realistic setting we assume the demand from factories vary within a double-bounded range. When the demand exceeds the upper boundary, the factory reaches its production capacity limit. Conversely, when the demand falls below the lower boundary, the factory operates below its optimal capacity.1

Figure 1: Illustration for the difference between vanilla OT and our DB-OT using the example of mines and factories as source and target, respectively. Vanilla OT assumes the equivalence between the supply and demand. In our DB-OT, we assume that the demand of the factory is not a fixed value, but rather a certain range by upper and lower bounds.

---

1Correspondence author.
Copyright © 2024, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.
Unbalanced OT (Chizat et al. 2018) is an OT variant to mitigate this inconsistency between source and target distributions. It relaxes the equality by only penalizing marginal deviation using divergences e.g. \(l_1, l_2\), KL divergence, and etc. This method transforms the original hard constraints (equations) into penalty terms, allowing for a relaxation of the constraints and increasing the degree of freedom in transportation. However, the degree of relaxation can only be controlled via the penalty coefficient, which is not directly concerning the desired range of transportation results.

In this paper, we propose and term a new variant of OT called Double Bounded Optimal Transport (DB-OT), which allows the sum of coupling column to be in a range rather than a number, i.e., replacing the original equality constraint with two inequality constraints. Under this formulation, we further provide theoretical results with proof for DB-OT. Additionally, we propose three Sinkhorn-like algorithms based on entropic DB-OT. We also show that our results also inspire downstream applications.

Inspired by the OT-based topic model in OT-LDA (Huynh, Zhao, and Phung 2020), we propose a barycenter-based clustering method using DB-OT. This method allows for control of the sample quantity in each cluster, thereby avoiding the clustering of isolated or minority points within a cluster (Wang and Su 2011). Note that our barycenter-based clustering method is not limited to Euclidean space but can also be extended to other spaces, such as the Wasserstein space (Agueh and Carlier 2011). This extension allows our method to be generalized to handle more complex data, including text data.

Another application is using DB-OT for unbalanced image recognition. Specifically, to our best knowledge, we uncover the relation between classification and OT: training a classifier is equivalent to optimizing Inverse OT (Shi et al. 2023a; Stuart and Wolfram 2020; Chiu, Wang, and Shafto 2022), while inference in testing corresponds to optimizing OT problem. With this view, we first apply DB-OT to improve the loss for long-tailed data training. Different from traditional view using Softmax-based cross-entropy loss for learning, we increase the column-sum constraints for learning the representations and thus the Softmax generalizes to Sinkhorn-like iterations and the well-known balanced Softmax (Ren et al. 2020) can be a special case with only one iteration. Then we apply our DB-OT technique to inference in testing process and our approach is agnostic to the way how the classifier is trained. We successfully adopt our DB-OT inference technique to long-tailed, uniform, and reverse long-tailed datasets.

The contributions of our work are:

1) We propose DB-OT to handle the case that the targets of transportation fall in a double-bounded range, which is an important yet under-studied setting for practical machine learning, e.g. cluster size-controllable clustering and long-tailed classification. We formulate DB-OT with the Kantorovich form by replacing the column equality constraints with double-bounded inequalities.

2) We solve DB-OT using entropic regularization and provide theoretical results, such as the static Schrödinger form. Additionally, we develop matrix scaling-based methods using Lagrange methods, with two space-efficient variants derived from the primal and dual forms of the problem, respectively. Note that the motivation behind proposing these three algorithms is to align with the vanilla entropic OT for DB-OT.

3) We propose our barycenter-based method for clustering using DB-OT. It helps control the size of cluster and thereby avoiding the unwanted over-scattered clustering with isolated or very few samples in clusters. Besides, compared with the previous barycenter-based topic model, a reweighting modification is also proposed to avoid the bias for barycenter calculation especially in Euclidean space.

4) Finally, our DB-OT is employed in the long-tailed classification. From the OT perspective, we observe that training can be viewed as optimizing the Inverse Optimal Transport (IOT), while inference in the testing can be regarded as minimizing the OT itself. Building upon this insight, we apply DB-OT to enhance the classification loss during training and also utilize it in the testing process based on a trained model.

**Preliminaries and Related Work**

**Basics of Optimal Transport**

As originated from (Kantorovich 1942), the Kantorovich’s Optimal Transport is to solve a linear program, which is widely used for many classical problems such as matching (Wang et al. 2013). Specifically, given the cost matrix \(C\) and two histograms \((a, b)\), Kantorovich’s OT involves solving the coupling \(P\) (i.e., the joint probability matrix) by

\[
\min_{P \in \mathcal{U}(a,b)} < C, P >,\tag{1}
\]

where \(U(a,b) = \{ P \in \mathcal{P}_{mm}^{+} | P_{1n} = a, P^T 1_m = b \}\). Relaxing with the entropic regularization (Wilson 1969) is one of the simple yet efficient methods for solving OT, which can be formulated as (Liero, Mielke, and Savaré 2018):

\[
\min_{P \in \mathcal{U}(a,b)} < C, P > - \epsilon H(P),\tag{2}
\]

where \(\epsilon > 0\) is the coefficient for entropic regularization \(H(P)\), and the regularized \(H(P)\) can be specified as \(H(P) = - < P, \log P - 1_{\text{max}} >\). The objective in Eq. 2 is \(\epsilon\)-strongly convex, and thus it has a unique solution, which can be solved by Sinkhorn algorithms as discussed in (Cuturi 2013; Benamou et al. 2015).

In this paper, beyond vanilla OT, we present our formulation for DB-OT and also propose the Sinkhorn algorithm.

**Optimal Transport w/ Inequality Constraints**

The vanilla OT only considers the equality constraints in \(U(a, b)\) yet inequality constraints are also handled in a few studies (Caffarelli and McCann 2010; Benamou et al. 2015). For instance, the optimal partial transport (OPT) problem (Caffarelli and McCann 2010) assumes some mass variation or partial mass displacement should be handled for transportation. Thus OPT focuses on transporting only a fraction of mass \(s \in [0, \min(||a||_1, ||b||_1)]\) as cheaply as possible. Then the constraints in this case is specified as:

\[
C_{\text{OPT}} = \{ P \in \mathcal{P}_{mn}^{+} | P_{1n} \leq a, P^T 1_m \leq b, 1_m^T P_{1n} = s \}.\tag{3}
\]
One can get the Partial-Wasserstein distance as \( PW(a, b) = \min_{\pi \in \mathcal{P}_{OPT}} < C, P > \). This problem has been studied in (Caffarelli and McCann 2010; Figalli 2010; Chapel, Alaya, and Gasso 2020). In particular, (Chizat et al. 2018; Benamou et al. 2015) propose the numerical solutions of OPT with entropic regularization. Different from the above constraints which only give upper bound in the constraints for mass transportation, in this paper, both upper and lower bounds are assumed for transportation, which allows the transportation results to fluctuate within a certain range rather than an exact number.

**Unbalanced Optimal Transport**

OT in its vanilla form requires the two histograms \( a \) and \( b \) to have the same total mass. Many works have tried to go beyond this assumption and following the setting in Unbalanced OT (Liero, Mielke, and Savaré 2018; Bai et al. 2023), the Kantorovich formulation in Eq. 1 is “relaxed” by only penalizing marginal deviation using some divergence \( D_\phi \):

\[
\min_{P \in \mathcal{P}_{R_{mn}}} < C, P > + \tau_1 D_\phi(P|_n) + \tau_2 D_\phi(P|_m) \tag{4}
\]

where \( \tau_1 \) and \( \tau_2 \) control how much mass variations are penalized as opposed to transportation of the mass and when \( \tau_1 = \tau_2 = +\infty \). This formulation equals to the original Kantorovich for \( \sum_i a_i = \sum_j b_j \). When \( \tau_1 \rightarrow +\infty \) and \( \tau_2 < +\infty \), Unbalanced OT and DB-OT share a similar characteristic where the source distribution remains fixed while the target becomes unfixed. However, unlike DB-OT, it can only passively adjust the target distribution through \( \tau_2 \) lacking a direct mechanism to control the targets within a range.

**Unbalanced Image Recognition**

Unbalanced classification, particularly in the case of long-tailed recognition, is a well-known challenge that has garnered significant attention in vision and machine learning (Zhang et al. 2023; He and Garcia 2009; Lin et al. 2017).

Various approaches (Tan et al. 2020; Cui et al. 2019b; Lin et al. 2017; Kang et al. 2019; Zhang et al. 2021) have been proposed to address this issue. One popular strategy involves rebalancing the class distribution in the training data (Ren et al. 2020; Park et al. 2021). Some methods employ techniques such as re-sampling (Kang et al. 2019) or re-weighting (Cui et al. 2019b) to ensure that the model pays more attention to minority classes during training. Data augmentation techniques, such as synthesizing additional samples for underrepresented classes, have also been utilized (Kim, Jeong, and Shin 2020). In this paper, we solve the unbalanced recognition with the OT perspective. We apply DB-OT both for learning the representations and inference in testing process, which is different to previous Bayesian view for classification.

**Double-Bounded Optimal Transport**

Vanilla OT usually involves two sets of equality constraints. Our motivation is to modify one type of constraints to the double bounded form which can be applicable to advanced clustering and classification.

**Formulation of DB-OT**

We first assume the source measure \( \alpha = \sum_{i=1}^{m} a_i \delta_{x_i} \) and two bounds of target measure \( \beta^l = \sum_{j=1}^{n} b^l_j \delta_{y_j}, \beta^u = \sum_{j=1}^{n} b^u_j \delta_{y_j} \) where \( a \geq 0 \) and \( b^l \geq b^u \geq 0 \). Our purpose is to transport the source samples from \( \alpha \) to the target measure which is defined between \( \beta^l \) and \( \beta^u \). By defining the cost matrix \( C \) between \( \{x_i\} \) and \( \{y_j\} \) (e.g. \( C_{ij} = d(x_i, y_j) \) where \( d(\cdot, \cdot) \) is a distance), we can formulate the Double-Bounded Optimal Transport (DB-OT) as

\[
\min_{P \in \mathcal{C}(a,b^u,b^l)} < C, P > = \sum_{i,j} C_{ij} P_{ij}, \tag{5}
\]

where \( \mathcal{C}(a,b^u,b^l) \) is the coupling set defined by constraints:

\[
\mathcal{C}(a,b^u,b^l) = \{ P \in \mathbb{R}_{mn}^+ | \sum_{i} P_{in} = a, \sum_{j} P_{jm} \leq P \sum_{j} b^l_j \leq b^u \}. \tag{6}
\]

Similar to vanilla OT, the above model is still a linear program. From Eq. 6, we can see that we replace the original constraint \( P \sum_{j} b^l_j \leq b^u \) with \( b^l \leq P \sum_{j} b^l_j \leq b^u \). That is, we slice \( P \sum_{j} b^l_j \) between \( b^l \) and \( b^u \) to relax the constraint on \( P \sum_{j} b^l_j \). It is obvious that when \( b = b^u = b^l \), DB-OT degenerates to vanilla OT in Eq. 1. In this paper, we mainly focus on the entropic regularized formulation:

\[
\min_{P \in \mathcal{C}(a,b^u,b^l)} < C, P > - \epsilon H(P), \tag{7}
\]

where \( \epsilon \) is the regularization coefficient and \( H(P) \) is the entropic regularization. It is obvious that the objective in Eq. 7 and the constraint set \( \mathcal{C}(a,b^u,b^l) \) are both convex and thus Eq. 7 has a unique optimal solution. We will discuss its solution and the algorithm of entropic DB-OT in next subsection.

**Why not using double-bounded \( \alpha \)?** An intuitive question arises: why don’t we assume that the \( \alpha \) distribution is also constrained within upper and lower bounds? For instance, let’s assume \( \alpha^u = \sum_{i} a^u_i \delta_{x_i} \) and \( \alpha^l = \sum_{i} a^l_i \delta_{x_i} \), and the coupling satisfies \( a^u \leq P_{in} \leq a^l \). In practice, the optimal transportation tends to transport mass vertically and towards the smaller lower bound. Without loss of generality, let’s assume \( \sum_i a^l_i < \sum_i b^l_j \). In this case, the optimal solution must satisfy \( \sum_i P_{ij} = \sum_i a^l_i \) (proof provided in oneline Appendix). As a result, setting an upper bound for the source distribution does not hold much significance.

**Sinkhorn Algorithm Variants for DB-OT**

In this subsection, we introduce several properties of DB-OT and present three corresponding Sinkhorn algorithms, namely the Bregman iterative algorithm, Sinkhorn-Knopp algorithm, and Dual algorithm for DB-OT, respectively, which aims to be in line with the algorithms of vanilla OT. Although these algorithms are derived from different formulations or properties of DB-OT, they are fundamentally interconnected and converge to same solutions, which validates the effectiveness of the algorithms.

- **Variant-I: Bregman Iterations For DB-OT**

Similar to the vanilla entropic OT, our entropic DB-OT in Eq. 7 can also be reformulated to the “static Schrödinger form” (Léonard 2012), which exactly learns a projection under KL divergence.

14984
### Proposition 1 (Static Schrödinger Form)
Rede fine a general KL divergence in line with (Benamou et al. 2015)

\[ KL(P|K) = \sum_{ij} P_{ij} \log \frac{P_{ij}}{K_{ij}} - P_{ij} + K_{ij} \]

Let \( K_{ij} = e^{-C_{ij}/\epsilon} \), the optimization in Eq. 7 is equivalent to the minimization:

\[ P^* = \arg \min_{P \in C(a,b^d,b^g)} KL(P|K) \]

Following (Benamou et al. 2015), a new variant of Sinkhorn algorithm can be devised with iterative Bregman projections for DB-OT. By splitting the constraint set as \( C(a,b^d,b^g) = C_1 \cap C_2 \cap C_3 \) where \( C_1 = \{ P \in \mathbb{R}_{nm}^+ | P_{i*} = a \}, C_2 = \{ P \in \mathbb{R}_{nm}^+ | P^T 1_m \geq b^d \}, C_3 = \{ P \in \mathbb{R}_{nm}^+ | P^T 1_m \leq b^g \} \). We can get the KL projection for \( P \) under the constraints as:

\[
\begin{align*}
\text{Proj}_{C_1}(P) &= \text{diag} \left( \frac{a}{P^T 1_m} \right) P \\
\text{Proj}_{C_2}(P) &= \text{diag} \left( \max \left( \frac{b^d}{P^T 1_m}, 1_m \right) \right) \\
\text{Proj}_{C_3}(P) &= \text{diag} \left( \min \left( \frac{b^g}{P^T 1_m}, 1_m \right) \right)
\end{align*}
\]

where the above division operator between two vectors is to be understood entry-wise (see details in online Appendix).

Finally, as introduced by (Benamou et al. 2015), assuming \( C_{i} \) as a matrix with positive integer \( i \) as the index of Bregman iteration, the minimization in Eq. 8 can be solved with the iterative projection \( P^{(n)} = \text{Proj}_{C_n}(P^{(n-1)}) \), starting from \( P^{(0)} = K = e^{-C/\epsilon} \). Then we can get the solution by \( P^* = \lim_{n \to \infty} \text{Proj}_{C_n}(P^{(n-1)}) \), which exactly do the iterations for Eq. 9 and (Bregman 1967) shows its convergence guarantee.

**Variant-II:** Sinkhorn-Knopp algorithm for DB-OT

In addition to analyzing DB-OT with KL divergence and Bregman iterative projections, Lagrangian methods can also solve the problem. Following (Cuturi 2013), we show that the form of solution is specified as follows.

### Proposition 2 (Solution Property)
The optimal solution of the optimization in Eq. 7 is unique and has the form

\[ P^* = \text{diag}(u)K \text{diag}(q \odot v) \]

where \( q \odot v = (q_i v_i) \in \mathbb{R}_n \), and the three scaling variables \( (u,v) \) satisfy \( u \in \mathbb{R}_m^+, 0_n \leq v \leq 1_n, \) and \( q \geq 1_n \).

The proof is given in uploaded online appendix. Comparing to vanilla OT with equality constraints, three scaling variables need to be iterated. We then also derive the Sinkhorn-Knopp algorithm of DB-OT given the iteration number \( l \):

\[
\begin{align*}
u^{(l+1)} &= \frac{a}{K(v^{(l)} \odot q^{(l)})}, \\
q^{(l+1)} &= \max \left( \frac{b^d}{(u^{(l+1)}K) \odot v^{(l)}}, 1_n \right), \\
v^{(l+1)} &= \min \left( \frac{b^g}{(u^{(l+1)}K) \odot q^{(l+1)}}, 1_n \right)
\end{align*}
\]

By initializing the scaling variables \( 0_n \leq v^{(0)} \leq 1_n \) and \( q^{(0)} \geq 1_n \), we can iterate by Eq. 11 to compute \( (u^{(1)}, v^{(1)}), (u^{(2)}, v^{(2)}), \ldots \) until convergence to \( (u^*, q^*, v^*) \). Then the solution can be obtained as \( P^* = \text{diag}(u^*)K \text{diag}(q^* \odot v^*) \).

**Variant-III:** The Dual Sinkhorn algorithm

We also analyze entropic DB-OT with dual formulation, which is devised as follows.

### Proposition 3 (Dual Formulation)
From the optimization in Eq. 7, we can get its dual formulation by maximizing

\[ L = < f, a > + < g, b^d > + < h, b^g > - \epsilon B(f, g, h) \]

where \( f \in \mathbb{R}^n, g \preceq 0 \) and \( h \succeq 0 \) are the corresponding dual variables and \( B(f, g, h) = < \epsilon e^{f/\epsilon}, K e^{(g+h)/\epsilon} > \).

Then we can get the new algorithm with a block coordinate ascent on the dual:

\[
\begin{align*}
\Delta_f L &= a - \epsilon e^{f/\epsilon} \odot K e^{(g+h)/\epsilon} \\
\Delta_g L &= b^d - \epsilon e^{(g+h)/\epsilon} \odot K^T e^{f/\epsilon} \\
\Delta_h L &= b^g - \epsilon e^{(g+h)/\epsilon} \odot K^T e^{f/\epsilon}
\end{align*}
\]

Starting from arbitrary \( g^{(0)} \leq 0 \) and \( h^{(0)} \geq 0 \), we get the dual algorithm by iterating the following with iteration number \( l \):

\[
\begin{align*}
f^{(l+1)} &= \epsilon \log a - \epsilon \log \left( K e^{(g^{(l)}+h^{(l)})/\epsilon} \right) \\
g^{(l+1)} &= \max \left( \epsilon \log b^d - \epsilon \log \left( K^T e^{f^{(l)}/\epsilon} \right) - h^{(l)}, 0 \right) \\
h^{(l+1)} &= \min \left( \epsilon \log b^g - \epsilon \log \left( K^T e^{f^{(l)}/\epsilon} \right) - g^{(l+1)}, 0 \right)
\end{align*}
\]

The iteration continues till convergence to \( (f^*, g^*, h^*) \) and the optimal solution equals \( P^* = \text{diag}(e^{f^*/\epsilon})K \text{diag}(e^{(g^*+h^*)/\epsilon}) \).

Exactly, Eq. 14 are mathematically equivalent to the Sinkhorn iterations in Eq. 11 when considering the primal-dual relations. Indeed, we recover that at any iteration: \( (f^{(l)}, g^{(l)}, h^{(l)}) = \epsilon \log (u^{(l)}, q^{(l)}, v^{(l)}) \). In this paper, we utilize the Sinkhorn-Knopp algorithm for clustering tasks and the Bregman iterations for long-tail classification.

### DB-OT for Clustering and Classification

#### Barycenter-based Clustering

Inspired by the OT-based topic model (Huynh, Zhao, and Pang 2020), we show the application of DB-OT to an advanced clustering task whereby the number of samples in each cluster is softly required to fall into a certain range. Specifically given samples \( \{x_i\} \), it learns the coupling \( P \) to match the samples to centroids, and the clustered centroids \( \{\beta_i\} \) are viewed as barycenters which needs to optimize. The optimization is specified as:

\[
\min_{P \in C(a,b^d,b^g), \{\beta_i\}} \sum_{st} P_{st} D(\alpha_{st}, \beta_t) - \epsilon H(P),
\]

where \( a = 1 \) (here we break the probability measure assumption as done in (Saad-Eldin et al. 2021)) and \( b^d, b^g \) are used to denote the given bounds of clustered centroids i.e. the minimum-maximum “quantity” of samples in each cluster. \( D(\alpha_{st}, \beta_t) \) is the distance between \( \alpha_s \) and \( \beta_t \) (e.g. Euclidean or Wasserstein distance), and \( H(P) \) is the entropic regularization which aims to relax the solution of coupling. This problem can be solved by alternating optimization. For each iteration, we first fix the centroid distributions \( \{\beta_t\} \) to

14985
compute the distance $D(\alpha_s, \beta_t)$ between sample $i$ and centroid $j$ and then transportation matrix $P$ can be learned from the source samples to clustered centroids.

Note in our method, we have $b^d \leq P^t \mathbf{1}_n \leq b^v$, which means the coupling between samples and centroids are controlled within a desired range, avoiding clusters with isolated or very few samples, as well as dominating clusters.

Specifically, for solving the optimization to clustering in Eq. 15, we first initialize $\{\beta_t\}$ with initialization of Kmeans++ or set them as randomly from $\{\alpha_s\}$. Then we can learn by iterating the following two steps: 1) Fixing $\{\beta_t\}$ to compute $P$. If the barycenter $\{\beta_t\}$ is known, then one can calculate the matrix $D$ where $D_{st} = D(\alpha_s, \beta_t)$. Viewing the matrix $D$ as the cost function, the optimization of $P$ in Eq. 15 is equal to DB-OT optimization in Eq. 7 and the algorithm proposed in Sec. can be used to get the result of $P$. Here we adopt the Sinkhorn-Knopp algorithm as given in Eq. 11.

2) Fixing $P$ to compute $\{\beta_t\}$. Assuming the transportation $P$ is known, our goal is now to update $\{\beta_t\}$ as the solution of the following optimization problem

$$\beta_t = \arg \min \beta \sum_s \frac{P_{st} \cdot R_{st}}{\sum_s P_{st} \cdot R_{st}} D(\alpha_s, \beta)$$

where $R_{st} = 1$ if $t = \arg \max_r P_{st}$, otherwise $R_{st} = 0$ given the sample $\alpha_s$. Here $R_{st}$ are used for re-weighting the weights of barycenters, which aims to debias the influence for samples from different clusters. Fig. 2 is the clustering results and we can find a large bias for centroids without re-weighting the barycenter weights.

**Under Different Metric Space for $D(\cdot, \cdot)$.** Clustering can be performed in different metric spaces by setting the distance function $D(\cdot, \cdot)$ to adapt to various metric settings or datasets. For example, when $D(\cdot, \cdot)$ represents the Euclidean distance and $\alpha_s$ corresponds to feature points in Euclidean space, we aim to cluster $\alpha_s$ in the Euclidean space. The centroids $\beta_t$ can be directly computed as follows:

$$\beta_t = \sum_s \frac{P_{st} \cdot R_{st}}{\sum_s P_{st} \cdot R_{st}} \alpha_s.$$  \hspace{1cm} (17)

Exactly if we set $R_{st} = 1$ for all $s$ and $t$ (i.e. without reweighting as shown in the second row of Fig. 2), the barycenters can be easily influenced by samples from other clusters. This can lead to a bias in the barycenter calculation, particularly when the value of $\epsilon$ is not sufficiently small.

In addition to Euclidean space, clustering can also be performed in Wasserstein space. In this case, we assume that $\alpha_s$ is no longer a set of feature points but a collection of probability measures, i.e., $\alpha_s = \sum_a \alpha^a_s \delta_a$. In this setting, the computation of the barycenter in Eq. 16 does not have an analytical solution but requires iterative methods, as proposed in (Benamou et al. 2015) for barycenter calculation.

**Advantages for our Barycenter-based Clustering.** Our barycenter-based clustering is essentially a unified clustering framework. Compare to the previous clustering works (e.g. Kmeans (Ahmed, Seraj, and Islam 2020) or OT-LDA (Huynh, Zhao, and Phung 2020)), the advantages can be summarized as follows: 1) **Our** clustering methods apply DB-OT to calculate the matching between samples and centroids, providing better controllability. As shown in Fig. 3, by setting $b^d$ and $b^v$, we can constrain the number of samples in each cluster to a certain extent, thereby avoiding clusters with isolated or very few samples, as well as clusters that dominate the majority of the data; 2) **Our** clustering methods can be applied not only to clustering feature points but also to clustering probability measures. For example, as shown in Fig. 3, we can perform clustering on the MNIST dataset in the Wasserstein space. In this case, we no
DB-OT by setting $U$. Training via Inverse DB-OT. Classification inference is learning OT. Learning can be viewed as optimizing Inverse OT and Classification. In the following, we show our main idea that classification can also be used for text clustering (e.g., topic modeling), making it more adaptable to different types of datasets; 3) Compared to previous work in OT-LDA (Huynh, Zhao, and Phung 2020), we introduce a reweighting of the barycenter weights, which helps to mitigate the biases in centroid computation, as shown in Fig. 2.

DB-OT for Long-tailed Classification

Motivation of using DB-OT in classification. Here we apply DB-OT for Long-tailed unbalanced classification. Following the works (Shi et al. 2023a,b), the Inverse Optimal Transport (Li et al. 2019; Stuart and Wolfram 2020; Chiu, Wang, and Shafto 2022) can be set as a bi-level optimization method for calculating $\tilde{P}$ with $K$ iteration. We find that the Balanced Softmax (Ren et al. 2020) is a special case of our loss, which validates the effectiveness of our theory.

Testing-time Inference with DB-OT. Here we focus on using DB-OT for inference in testing process. Note our inference method is orthogonal to the training method proposed in Sec., which means our inference can be used for different classifiers. For the testing inference, we treat all the training as a process of feature learning and then with the feature learned, we can define the cost matrix given each batch data and match the features and labels with OT. For the long-tailed classification, the inference is as follows:

$$\min_{P \in C(a, b^u, b^d)} < C, P > -\epsilon H(P).$$

Here $C_{ij} = c - l_{ij}$ with large enough value $c$ to guarantee the positiveness of the the cost. Meanwhile, we set $b^u$ and $b^d$ as Eq. 20, which enables predictions to fluctuate both upward and downward, taking into account the randomness of batch data sampling. During the testing process, $r$ can be configured as a known long-tailed distribution, uniform distribution, reverse long-tailed distribution, or any other appropriate distribution, which is a main advantage of our method.

Experiments

We conduct experiments on both clustering and classification. To showcase the advantage of DB-OT on fine-grained controlling the behavior of clustering under bounded sizes of clusters, and classification in long-tail cases.

Experiments on Size-controlled Clustering

We first study our clustering setting which each cluster’s size is bounded by a certain range, using our DB-OT function with Barycenter. The experimental results on Gaussian mixture synthetic datasets with 2D 150 points are shown in...
Table 1: Top-1 accuracy (%) for long-tailed image classification with 200 imbalanced factor on three popular LT datasets.

<table>
<thead>
<tr>
<th>Method</th>
<th>CIFAR10-LT</th>
<th>CIFAR100-LT</th>
<th>ImageNet-LT</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Many</td>
<td>Few</td>
<td>All</td>
</tr>
<tr>
<td>Vanilla Softmax</td>
<td>77.4</td>
<td>68.9</td>
<td>74.9</td>
</tr>
<tr>
<td>LDAM (Cao et al. 2019)</td>
<td>80.5</td>
<td>65.2</td>
<td>75.9</td>
</tr>
<tr>
<td>Balanced Softmax</td>
<td>82.2</td>
<td>71.6</td>
<td>79.0</td>
</tr>
<tr>
<td>Focal Loss (Lin et al. 2017)</td>
<td>79.6</td>
<td>58.4</td>
<td>73.3</td>
</tr>
<tr>
<td>LogitAdjust (Menon et al. 2020)</td>
<td>80.0</td>
<td>35.3</td>
<td>66.6</td>
</tr>
<tr>
<td>CB-CE (Cui et al. 2019b)</td>
<td>76.6</td>
<td>70.7</td>
<td>74.8</td>
</tr>
<tr>
<td>CB-FC (Cui et al. 2019b)</td>
<td>76.6</td>
<td>70.7</td>
<td>74.8</td>
</tr>
<tr>
<td>DB-OT Loss (ours)</td>
<td>82.4</td>
<td>80.8</td>
<td>81.9</td>
</tr>
</tbody>
</table>

Table 2: Top-1 accuracy (%) of CIFAR-100 for the comparison of different testing inference methods given four trained models.

<table>
<thead>
<tr>
<th>Testing Inference</th>
<th>Vanilla Softmax Loss</th>
<th>Balanced Softmax Loss</th>
<th>Focal Loss</th>
<th>Logit Adjustment Loss</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>LT Uniform</td>
<td>RLT</td>
<td>LT Uniform</td>
<td>RLT</td>
</tr>
<tr>
<td>Vanilla Softmax</td>
<td>71.6</td>
<td>42.0</td>
<td>18.7</td>
<td>66.5</td>
</tr>
<tr>
<td>classifier normalize</td>
<td>70.0</td>
<td>45.2</td>
<td>16.8</td>
<td>54.9</td>
</tr>
<tr>
<td>Class-aware bias</td>
<td>68.0</td>
<td>41.3</td>
<td>14.5</td>
<td>41.5</td>
</tr>
<tr>
<td>DB-OT inference (ours)</td>
<td>71.8</td>
<td>48.5</td>
<td>36.6</td>
<td>71.4</td>
</tr>
</tbody>
</table>

Fig. 2. We can find that the OT-LDA and our method without reweighting exhibit a noticeable bias and re-weighting the barycenter weights can overcome this issue with Eq. 16. We also evaluate the top-1 accuracy within the cluster and the results of OT-LDA, ours (without and with reweighting) are 82.67%, 88.00%, and 100.00% respectively, which shows the superiority of our method. We also do the clustering experiments on MNIST with three different methods (i.e. K-means, DB-OT under Euclidean space and DB-OT under Wasserstein space). We select parts of MNIST data (12 images in each class) and set 16 clusters for the experiments. The results are shown in Fig. 3, and we can find that our method can be more controllable for the number of samples in each cluster. Note the bounds are to control the column probability sum of the coupling and thus it does not strictly meet the bound range for sample quantity.

Experiments on Long-tailed Classification

We evaluate our two methods – both for using DB-OT to improve the loss and adopting the DB-OT for classification inference in the testing process. These two methods are parallel to each other as the former aims to learn a better representation and the latter is to get a more accurate prediction based on a known testing label ratio. We do the experiments on CIFAR10-LT, CIFAR100-LT (Krizhevsky, Hinton et al. 2009), ImageNet-LT (Liu et al. 2019) for image classification. For a fair comparison, all methods share the same network backbone and hyperparameters. More details about the experimental settings is given in the online Appendix.

Improvements of DB-OT based loss. To evaluate the performance of our DB-OT based loss, following (Ren et al. 2020), we start with a vanilla Softmax pretrained model and train with our DB-OT based loss to improve the representations. We use the corresponding balanced testing dataset for evaluation, where its labels are uniformly distributed. We report top-1 accuracy as the evaluation metric. Specifically, for CIFAR10-LT, we report accuracy on two sets of classes in detail: Many-shot (more than 100 images) and Few-shot (less than 100 images). For CIFAR00-LT and ImageNet-LT, we report accuracy on three sets: Many-shot (more than 100 images), Medium-shot (20 ~ 100 images), and Few-shot (less than 20 images). The experiments for unbalanced image classification are all conducted with an imbalanced factor of less than 200, which is defined as the ratio of the number of training instances in the largest class to the smallest (Ren et al. 2020). The results for long-tailed classification are presented in Table 1. From a comprehensive perspective, our approach achieves the best average accuracy across the entire dataset. When examining the results for different data splits, our method particularly outperforms the others on subsets with fewer images (i.e., median or few-shot).

Performance of testing inference. To evaluate the DB-OT based inference, we use the corresponding long-tailed (LT), uniform, and Reverse long-tailed (RLT) testing dataset for evaluation. We compare our method with vanilla Softmax, classifier normalization (Kang et al. 2019) and class-aware bias (Menon et al. 2020), which are discussed in (Wu et al. 2021). Table 2 shows the comparison results for all the models and different testing data in CIFAR100. Our inference method outperforms and can achieve a great improvement when the testing data is reverse long-tailed distributed and the model trained by vanilla softmax performs the best using our DB-OT testing inference though it may fail with vanilla softmax prediction in testing. More experimental results (e.g. ablation study) are given in online appendix.

Conclusion

We have presented the so-called double-bounded optimal transport, with theoretical analysis and further derive three variants of algorithms to solve the problem. We test our technique for the challenging yet realistic tasks of cluster-size bounded clustering, as well as long-tailed image recognition.
Acknowledgements
Junchi Yan is the corresponding author, and this work is supported by the National Natural Science Foundation of China (grant numbers 92370201 and 62222607).

References
Arjovsky, M.; Chintala, S.; and Bottou, L. 2017. Wasserstein GAN. ICML.
Gulrajani, I.; Ahmed, F.; Arjovsky, M.; Dumoulin, V.; and Courville, A. 2017. Improved Training of Wasserstein GANs. In NIPS.


Shi, L.; Zhang, G.; Zhen, H.; Fan, J.; and Yan, J. 2023a. Understanding and Generalizing Contrastive Learning from the Inverse Optimal Transport Perspective. in ICML.


