CSOT: Curriculum and Structure-Aware Optimal Transport for Learning with Noisy Labels

Wanxing Chang; Ye Shi; and Jingya Wang

37th International Conference on Neural Information Processing Systems (NeurIPS 2023)

Learning with Noisy Labels, Optimal Transport, Curriculum Learning
https://dl.acm.org/doi/10.5555/3666122.3666495
https://github.com/changwxx/CSOT-for-LNL

Abstract

Learning with noisy labels (LNL) poses a significant challenge in training a well-generalized model while avoiding overfitting to corrupted labels.Recent advances have achieved impressive performance by identifying clean labels and correcting corrupted labels for training.However, the current approaches rely heavily on the model's predictions and evaluate each sample independently without considering either the global or local structure of the sample distribution.These limitations typically result in a suboptimal solution for the identification and correction processes, which eventually leads to models overfitting to incorrect labels.In this paper, we propose a novel optimal transport (OT) formulation, called Curriculum and Structure-aware Optimal Transport (CSOT). CSOT concurrently considers the inter- and intra-distribution structure of the samples to construct a robust denoising and relabeling allocator.During the training process, the allocator incrementally assigns reliable labels to a fraction of the samples with the highest confidence. These labels have both global discriminability and local coherence.Notably, CSOT is a new OT formulation with a nonconvex objective function and curriculum constraints, so it is not directly compatible with classical OT solvers. Here, we develop a lightspeed computational method that involves a scaling iteration within a generalized conditional gradient framework to solve CSOT efficiently. Extensive experiments demonstrate the superiority of our method over the current state-of-the-arts in LNL.

Problem Statement and Research Objectives

Mining large-scale labeled data based on a web search and user tags can provide a cost-effective way to collect labels, but this approach inevitably introduces noisy labels. Since DNNs can so easily overfit to noisy labels, such label noise can significantly degrade performance, giving rise to a challenging task: learning with noisy labels (LNL).

■ Learning with noisy labels

- **Identifying clean labels**: These methods often *model per-sample loss distributions* using a Beta Mixture Model or a Gaussian Mixture Model, **treating samples with smaller loss as clean ones**.
- **Label correction methods** : Typically adopt a pseudo-labeling strategy that **leverages the DNNs predictions to correct the labels**

  ➔ However, these approaches evaluate each sample independently **without considering the correlations among samples**, which leads to a suboptimal identification and correction solution.

■ Optimal transport-based pseudo-labeling

- OT-based PL optimizes the **mapping samples to class centroids**, while considering the global structure of the sample distribution in terms of marginal constraints instead of per-sample predictions.

  ➔ However, these approaches only consider the inter-distribution matching of samples and classes but **do not consider the intra-distribution coherence structure of samples**.

  - More specifically, the cost matrix in OT relies on pairwise metrics, so two nearby samples could be mapped to two far-away class centroids (Fig. 1).

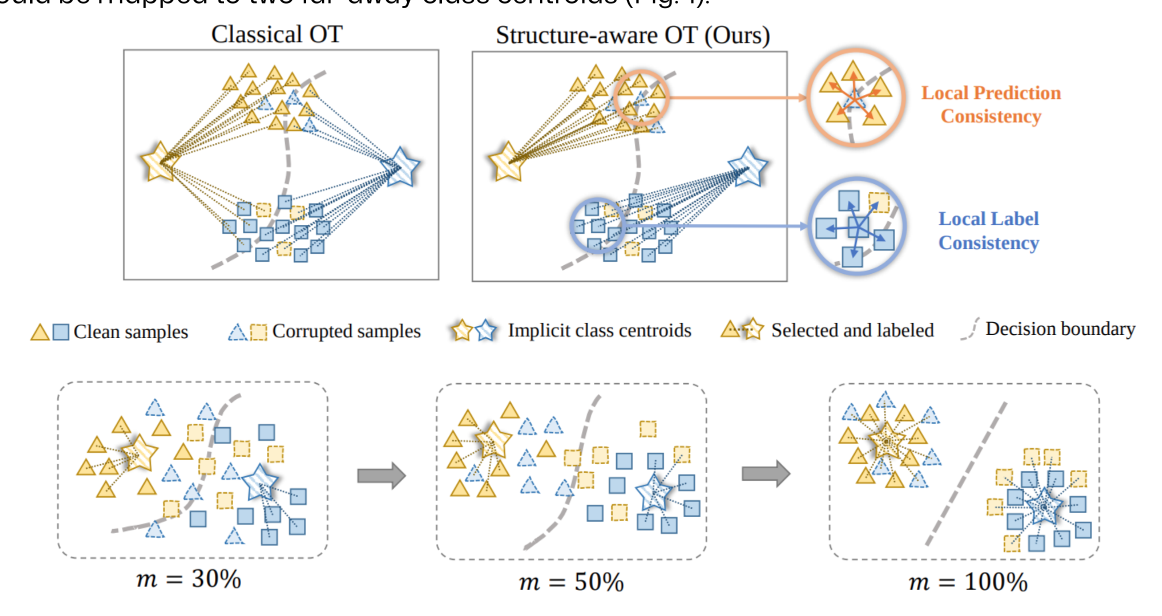


Figure 1: **(Top) Comparison between classical OT and our proposed Structure-aware OT.** Classical OT tends to mismatch two nearby samples to two far-away class centroids when the decision boundary is not accurate enough. To mitigate this, our SOT generates local consensus assignments for each sample by preserving prediction-level and label-level consistency. Notably, for vague samples located near the ambiguous decision boundary, SOT rectifies their assignments based on the neighborhood majority consistency. **(Bottom) The illustration of our curriculum denoising and relabeling based on proposed CSOT.** The decision boundary refers to the surface that separates two classes by the classifier. The m represents the curriculum budget that controls the number of selected samples and progressively increases during the training process.

Proposed Method

In this paper, to enhance intra-distribution coherence, we propose a new OT formulation for denoising and relabeling, called **Structure-aware Optimal Transport (SOT)**. This formulation fully considers the intra-distribution structure of the samples and produces robust assignments with both *global discriminability* and *local coherence*.

- Technically speaking, we introduce **local coherent regularized terms** to encourage both prediction- and label-level local consistency in the assignments.
- Furthermore, to avoid generating incorrect labels in the early stages of training or cases with high noise ratios, we devise **Curriculum and Structure-aware Optimal Transport (CSOT)** based on SOT.
  - CSOT constructs a robust denoising and relabeling allocator by relaxing one of the equality constraints to allow only a fraction of the samples with the highest confidence to be selected.
  - These samples are then assigned with reliable pseudo labels.
  - The allocator progressively selects and relabels batches of high-confidence samples based on an increasing budget factor that controls the number of selected samples.
  - Notably, CSOT is a new OT formulation with a nonconvex objective function and curriculum constraints, so it is significantly different from the classical OT formulations.

1. Structure-Aware Optimal Transport for Denoising and Relabeling

Our proposed SOT for denoising and relabeling is formulated by **adding two local coherent regularized terms** based on Problem (2). Given a cosine similarity $\mathbf{S} \in \mathbb{R}^{B\times B}$ among samples in feature space, a one-hot label matrix $\mathbf{L} \in \mathbb{R}^{B\times C}$ transformed from given noisy labels, SOT is formulated as follows

$$\min_{\mathbf{Q}\in\Pi(\frac{1}{B}\mathbb{1}_B,\frac{1}{C}\mathbb{1}_C)} \langle -\log \mathbf{P}, \mathbf{Q}\rangle + \kappa\left(\Omega^{\mathbf{P}}(\mathbf{Q}) + \Omega^{\mathbf{L}}(\mathbf{Q})\right) \qquad (3)$$

where the local coherent regularized terms $\Omega^{\mathbf{P}}$ and $\Omega^{\mathbf{L}}$ encourages prediction-level and label-level local consistency respectively, and are defined as follows (where $\odot$ indicates element-wise multiplication)

$$\Omega^{\mathbf{P}}(\mathbf{Q}) = -\sum_{i,j}\mathbf{S}_{ij}\sum_k \mathbf{P}_{ik}\mathbf{P}_{jk}\mathbf{Q}_{ik}\mathbf{Q}_{jk} = -\langle \mathbf{S}, (\mathbf{P}\odot\mathbf{Q})(\mathbf{P}\odot\mathbf{Q})^\top\rangle \qquad (4)$$

$$\Omega^{\mathbf{L}}(\mathbf{Q}) = -\sum_{i,j}\mathbf{S}_{ij}\sum_k \mathbf{L}_{ik}\mathbf{L}_{jk}\mathbf{Q}_{ik}\mathbf{Q}_{jk} = -\langle \mathbf{S}, (\mathbf{L}\odot\mathbf{Q})(\mathbf{L}\odot\mathbf{Q})^\top\rangle \qquad (5)$$

To be more specific, $\Omega^{\mathbf{P}}$ encourages assigning larger weight to $\mathbf{Q}_{ik}$ and $\mathbf{Q}_{jk}$ if the i-th sample is very close to the j-th sample, and their predictions $\mathbf{P}_{ik}$ and $\mathbf{P}_{jk}$ from the k-th class centroid are simultaneously high.

2. Curriculum and Structure-Aware Optimal Transport for Denoising and Relabeling

For the purpose of robust clean label identification and corrupted label correction, we further propose a Curriculum and Structure-aware Optimal Transport (CSOT), which constructs a robust curriculum allocator.

- This curriculum allocator gradually selects a fraction of the samples with high confidence from the noisy training set, controlled by a budget factor, then assigns reliable pseudo labels for them.
- Our proposed CSOT for denoising and relabeling is formulated by **introducing new curriculum constraints based on SOT** in Problem (3).
  - Given curriculum budget factor $m \in [0,1]$, **our CSOT seeks optimal coupling matrix $\mathbf{Q}$ by minimizing following objective**

$$\min_{\mathbf{Q}} \langle -\log \mathbf{P}, \mathbf{Q}\rangle + \kappa\left(\Omega^{\mathbf{P}}(\mathbf{Q}) + \Omega^{\mathbf{L}}(\mathbf{Q})\right)$$
$$\text{s.t.}\quad \mathbf{Q} \in \{\mathbf{Q}\in\mathbb{R}_+^{B\times C} \mid \mathbf{Q}\mathbb{1}_C \le \frac{1}{B}\mathbb{1}_B, \mathbf{Q}^\top\mathbb{1}_B \le \frac{m}{C}\mathbb{1}_C\} \qquad (6)$$

- Unlike SOT, which enforces an equality constraint on the samples, **CSOT relaxes this constraint and defines the total coupling budget** as $m \in [0,1]$, where m represents the expected total sum of Q.
  - Intuitively speaking, $m = 0.5$ indicates that top 50% confident samples are selected from all the classes, **avoiding only selecting the same class for all the samples within a mini-batch**.
- In addition, we define the general confident scores of samples as $\mathcal{W} = \{w_0, w_1, \cdots, w_{B-1}\}$, where $w_i = \mathbf{Q}_{i\hat{y}_i}/(m/C)$.
  - Since our curriculum allocator assigns weight to only a fraction of samples controlled by m, we use topK$(\mathcal{S}, k)$ operation (return top-k indices of input set $\mathcal{S}$) to identify selected samples denoted as δ_i

$$\delta_i = \begin{cases} 1, & \text{topK}(\mathcal{W}, \lfloor mB \rfloor) \\ 0, & \text{otherwise} \end{cases} \qquad (7)$$

3. Training Objectives

- To avoid error accumulation in the early stage of training, we adopt a two-stage training scheme.
  - In the first stage, the model is **supervised by progressively selected clean labels** and **self-supervised by unselected samples**.
  - In the second stage, the model is **semi-supervised by all denoised labels**.
- Notably, we construct our training objective mainly based on Mixup loss $\mathcal{L}^{mix}$ and Label consistency loss $\mathcal{L}^{lab}$ same as NCE [45], and a self-supervised loss $\mathcal{L}^{simsiam}$ proposed in SimSiam [15].

Our two-stage training objective can be constructed as follows

$$\mathcal{L}^{sup} = \mathcal{L}^{mix}_{\mathcal{D}_{clean}} + \mathcal{L}^{lab}_{\mathcal{D}_{clean}} + \lambda_1\mathcal{L}^{simsiam}_{\mathcal{D}_{corrupted}} \qquad (9)$$

$$\mathcal{L}^{semi} = \mathcal{L}^{mix}_{\mathcal{D}_{clean}} + \mathcal{L}^{lab}_{\mathcal{D}_{clean}} + \lambda_2\mathcal{L}^{lab}_{\mathcal{D}_{corrupted}} \qquad (10)$$

4. Lightspeed Computation for CSOT

- Solving Curriculum Optimal Transport
  - Problem (6) can be solved by performing iterative KL projection between $\mathcal{C}_1$ and $\mathcal{C}_2$, namely Dykstra's algorithm

$$\mathcal{C}_1 \overset{\text{def}}{=} \{\mathbf{Q}\in\mathbb{R}_+^{|\alpha|\times|\beta|} \mid \mathbf{Q}\mathbb{1}_\beta \ge \alpha$$
$$\mathcal{C}_2 \overset{\text{def}}{=} \{\mathbf{Q}\in\mathbb{R}_+^{|\alpha|\times|\beta|} \mid \mathbf{Q}^\top\mathbb{1}_\alpha \ge \beta$$

- Solving Curriculum and Structure-Aware Optimal Transport

Algorithm 1 Efficient scaling iteration for entropic regularized Curriculum OT

1: **Input:** Cost matrix $\mathbf{C}$, marginal constraints vectors $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$, entropic regularization weight ε
2: Initialize: $\mathbf{K} \leftarrow e^{-\mathbf{C}/\varepsilon}$, $\boldsymbol{v}^{(0)} \leftarrow \mathbb{1}_{|\boldsymbol{\beta}|}$
3: Compute: $\mathbf{K}_{\boldsymbol{\alpha}} \leftarrow \frac{\mathbf{K}}{\text{diag}(\boldsymbol{\alpha})\mathbb{1}_{|\boldsymbol{\alpha}|\times|\boldsymbol{\beta}|}}$, $\mathbf{K}_{\boldsymbol{\beta}}^\top \leftarrow \frac{\mathbf{K}^\top}{\text{diag}(\boldsymbol{\beta})\mathbb{1}_{|\boldsymbol{\beta}|\times|\boldsymbol{\alpha}|}}$ `// Saving computation`
4: **for** $n = 1, 2, 3, \ldots$ **do**
5: $\quad \boldsymbol{u}^{(n)} \leftarrow \min\left(\frac{\mathbb{1}_{|\boldsymbol{\alpha}|}}{\mathbf{K}_{\boldsymbol{\alpha}}\boldsymbol{v}^{(n-1)}}, \mathbb{1}_{|\boldsymbol{\alpha}|}\right)$
6: $\quad \boldsymbol{v}^{(n)} \leftarrow \frac{\mathbb{1}_{|\boldsymbol{\beta}|}}{\mathbf{K}_{\boldsymbol{\beta}}^\top\boldsymbol{u}^{(n)}}$
7: **end for**
8: **Return:** $\text{diag}(\boldsymbol{u}^{(n)})\mathbf{K}\text{diag}(\boldsymbol{v}^{(n)})$

Algorithm 2 Generalized conditional gradient algorithm for entropic regularized CSOT

1: **Input:** Cost matrix $\mathbf{C}$, marginal constraints vectors $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$, entropic regularization weight ε, local coherent regularization weight κ, local coherent regularization function $\Omega : \mathbb{R}^{|\boldsymbol{\alpha}|\times|\boldsymbol{\beta}|} \to \mathbb{R}$, and its gradient function $\nabla\Omega : \mathbb{R}^{|\boldsymbol{\alpha}|\times|\boldsymbol{\beta}|} \to \mathbb{R}^{|\boldsymbol{\alpha}|\times|\boldsymbol{\beta}|}$
2: Initialize: $\mathbf{Q}^{(0)} \leftarrow \boldsymbol{\alpha}\boldsymbol{\beta}^T$
3: **for** $i = 1, 2, 3, \ldots$ **do**
4: $\quad \mathbf{G}^{(i)} \leftarrow \mathbf{Q}^{(i)} + \kappa\nabla\Omega(\mathbf{Q}^{(i)})$ `// Gradient computation`
5: $\quad \widetilde{\mathbf{Q}}^{(i)} \leftarrow \text{argmin}_{\mathbf{Q}\in\boldsymbol{\Pi}^c(\boldsymbol{\alpha},\boldsymbol{\beta})} \left\langle \mathbf{Q}, \mathbf{G}^{(i)} \right\rangle + \varepsilon\left\langle \mathbf{Q}, \log\mathbf{Q}\right\rangle$
$\quad$ `// Linearization, solved efficiently by Algorithm 1`
6: $\quad$ Choose $\eta^{(i)} \in [0,1]$ so that it satisfies the Armijo rule `// Backtracking line-search`
7: $\quad \mathbf{Q}^{(i+1)} \leftarrow \left(1-\eta^{(i)}\right)\mathbf{Q}^{(i)} + \eta^{(i)}\widetilde{\mathbf{Q}}^{(i)}$ `// Update`
8: **end for**
9: **Return:** $\mathbf{Q}^{(i)}$

Evaluation and Results

1. Comparison with the State-of-the-Arts

Table 1: **Comparison with state-of-the-art methods in test accuracy (%) on CIFAR-10 and CIFAR-100.** The results are mainly copied from [45, 48]. We present the performance of our CSOT method using the "mean±variance" format, which is obtained from 3 trials with different seeds.

Dataset	CIFAR-10					CIFAR-100			
Noise type	Symmetric				Assymetric	Symmetric			
Method/Noise ratio	0.2	0.5	0.8	0.9	0.4	0.2	0.5	0.8	0.9
Cross-Entropy	86.8	79.4	62.9	42.7	85.0	62.0	46.7	19.9	10.1
F-correction [56]	86.8	79.8	63.3	42.9	87.2	61.5	46.6	19.9	10.2
Co-teaching+ [31]	89.5	85.7	67.4	47.9	-	65.6	51.8	27.9	13.7
PENCIL [76]	92.4	89.1	77.5	58.9	88.5	69.4	57.5	31.1	15.3
DivideMix [46]	96.1	94.6	93.2	76.0	93.4	77.3	74.6	60.2	31.5
ELR [50]	95.8	94.8	93.3	78.7	93.0	77.6	73.6	60.8	33.4
NGC [72]	95.9	94.5	91.6	80.5	90.6	79.3	75.9	62.7	29.8
RRL [48]	96.4	95.3	93.3	77.4	92.6	80.3	76.0	61.1	33.1
MOIT [55]	93.1	90.0	79.0	69.6	92.0	73.0	64.6	46.5	36.0
UniCon [41]	96.0	95.6	93.9	**90.8**	94.1	78.9	77.6	63.9	44.8
NCE [45]	96.2	95.3	93.9	88.4	94.5	**81.4**	76.3	64.7	41.1
OT Cleaner [73]	91.4	85.4	56.9	-	-	67.4	58.9	31.2	-
OT-Filter [23]	96.0	95.3	94.0	90.5	95.1	76.7	73.8	61.8	42.8
CSOT (Best)	**96.6±0.10**	**96.2±0.11**	**94.4±0.16**	90.7±0.33	**95.5±0.06**	80.5±0.28	**77.9±0.18**	**67.8±0.23**	**50.5±0.46**
CSOT (Last)	96.4±0.18	96.0±0.11	94.3±0.20	90.5±0.36	95.2±0.12	80.2±0.31	77.7±0.14	67.6±0.36	50.3±0.33

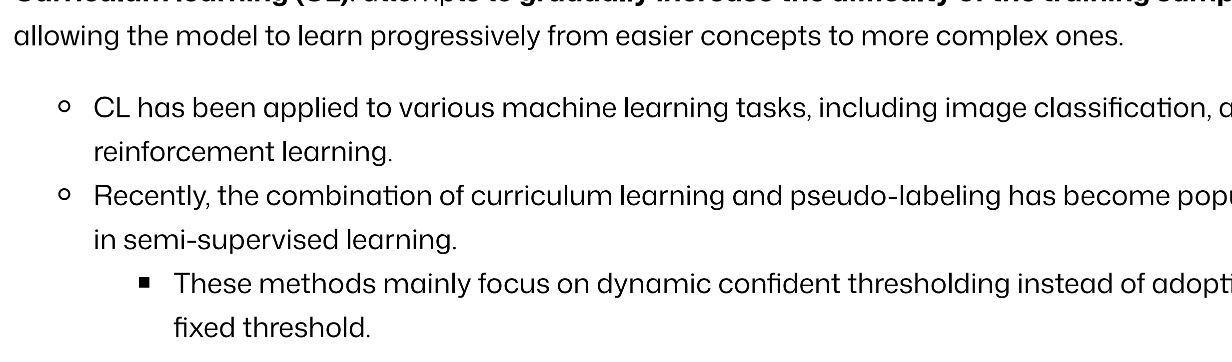
Table 2: **Comparison with SOTA methods in top-1 / 5 test accuracy (%) on the Webvision and ImageNet ILSVRC12 validation sets.**

Method	Webvision top-1	Webvision top-5	ILSVRC12 top-1	ILSVRC12 top-5
F-correction [56]	61.12	82.68	57.36	82.36
Decoupling [52]	62.54	84.74	58.26	82.26
MentorNet [39]	63.00	81.40	57.80	79.92
Co-teaching [31]	63.58	85.20	61.48	84.70
DivideMix [46]	77.32	91.64	75.20	90.84
ELR [50]	76.26	91.26	68.71	87.84
ELR+ [50]	77.78	91.68	70.29	89.76
NGC [72]	79.20	91.80	74.40	91.00
RRL [48]	77.80	91.30	74.40	90.90
UniCon [41]	77.60	93.44	75.29	93.72
MOIT [55]	77.90	91.90	73.80	91.70
NCE [45]	79.50	**93.80**	76.30	**94.10**
CSOT	**79.67**	91.95	**76.64**	91.67

2. Ablation Studies and Analysis

Table 4: **Ablation studies under multiple label noise ratios on CIFAR-10 and CIFAR-100.** "repl." is an abbreviation for "replaced", and $\mathcal{L}^{ce}$ represents a cross-entropy loss. GMM refers to the selection of clean labels based on small-loss criterion [46]. CT (confidence thresholding [62]) is a relabeling scheme where we set the CT value to 0.95.

	Dataset	CIFAR-10				CIFAR-100			
	Noise type	Sym.			Asym.	Sym.			Avg
	Method/Noise ratio	0.5	0.8	0.9	0.4	0.5	0.8	0.9	
Denoise Relabeling Technique	(a) Classical OT	95.45	91.95	82.35	95.04	75.96	62.46	43.28	78.07
	(b) Structure-aware OT	95.86	91.87	83.29	95.06	76.20	63.73	44.57	78.65
	(c) CSOT w/o $\Omega^{\mathbf{P}}$ and $\Omega^{\mathbf{L}}$	95.53	93.84	89.50	95.14	75.96	66.50	47.55	80.57
	(d) CSOT w/o $\Omega^{\mathbf{P}}$	95.77	94.08	89.97	95.35	76.09	66.79	48.13	80.88
	(e) CSOT w/o $\Omega^{\mathbf{L}}$	95.55	93.97	90.41	95.15	76.17	67.28	48.01	80.93
Learning Technique	(f) GMM + $\mathcal{L}^{sup}$	92.48	80.37	31.76	90.80	69.52	48.49	20.86	62.04
	(g) CSOT repl. $\mathcal{L}^{sup}$ with $\mathcal{L}^{ce}$	93.47	81.93	53.45	91.43	72.66	50.62	21.77	66.48
	(h) CSOT w/o $\mathcal{L}^{semi}$	95.34	93.04	88.9	94.11	75.16	61.13	36.94	77.80
	(i) CSOT repl. correction with CT (0.95)	95.46	90.73	89.09	95.21	75.85	64.28	48.76	79.91
	(j) CSOT w/o $\mathcal{L}^{simsiam}_{\mathcal{D}_{corrupted}}$	95.92	94.17	89.31	95.16	76.38	66.17	45.56	80.38
	CSOT	**96.20**	**94.39**	**90.65**	**95.50**	**77.94**	**67.78**	**50.50**	**81.85**



(a) Clean accuracy (b) Corrected accuracy (c) Clean recall rate

Figure 2: **Performance comparison for clean label identification and corrupted label correction.**

Notes

- https://github.com/lijichang/LNL-NCE
- **Curriculum learning (CL)**: attempts to **gradually increase the difficulty of the training samples**, allowing the model to learn progressively from easier concepts to more complex ones.
  - CL has been applied to various machine learning tasks, including image classification, and reinforcement learning.
  - Recently, the combination of curriculum learning and pseudo-labeling has become popular in semi-supervised learning.
    - These methods mainly focus on dynamic confident thresholding instead of adopting a fixed threshold.
- **Optimal Transport (OT)** vs **Pseudo-labeling based on optimal transport**

Optimal Transport (OT)	**Pseudo-labeling based on optimal transport**
$\min_{\mathbf{Q}\in\Pi(\alpha,\beta)}\langle \mathbf{C}, \mathbf{Q}\rangle \quad (1)$ $* <\cdot,\cdot>$: Frobenius dot-product	$\min_{\mathbf{Q}\in\Pi(\frac{1}{B}\mathbb{1}_B,\frac{1}{C}\mathbb{1}_C)}\langle -\log\mathbf{P}, \mathbf{Q}\rangle \quad (2)$ $* \mathbb{1}_d$: d-dimensional vector of ones
◦ α, β : probability vectors (indicating two distributions) ▪ $\lvert\alpha\rvert$: the dimension of α ◦ Cost matrix : $\mathbf{C} \in \mathbb{R}^{\lvert\alpha\rvert\times\lvert\beta\rvert}$ ◦ Coupling matrix $\mathbf{Q}$ ▪ Q_{ik} **indicates how the mass is moved** from the i of the α to the k of the β. ▪ Constrains: $\mathbf{Q}\mathbb{1}_{\lvert\beta\rvert} = \alpha, \quad \mathbf{Q}^\top\mathbb{1}_{\lvert\alpha\rvert} = \beta$	◦ B, C : the batch size of samples, and the number of classes ◦ Cost matrix : $\mathbf{C} = -\log\mathbf{P}$ ▪ $\mathbf{P} \in \mathbb{R}_+^{B\times C}$: **classifier softmax predictions** ◦ Coupling matrix $\mathbf{Q}$ ▪ Q_{ik} **indicates the probability** that the sample i will be mapped to a class k ▪ Constrains: $\mathbf{Q}\mathbb{1}_C = \frac{1}{B}\mathbb{1}_B, \quad \mathbf{Q}^\top\mathbb{1}_B = \frac{1}{C}\mathbb{1}_C$

  - Directly optimizing the exact OT problem would be time-consuming, and an entropic regularization term is introduced:

$$\min_{\mathbf{Q}\in\Pi(\alpha,\beta)} \langle \mathbf{C}, \mathbf{Q}\rangle + \epsilon\langle \mathbf{Q}, \log\mathbf{Q}\rangle, \quad \text{where } \epsilon > 0$$