Cost Parameterization and Constraint Relaxation for Inverse Optimal Transport with Applications to Contrastive Learning

Shanghai Jiao Tong University

April 25, 2023

IOT-CL

Contrastive Learning (CL)

Introduction

- Unsupervised Learning
- Transfer Knowledge
  - Pretrain + Downstream task
- Data Augmentation
  - Color Transformation
  - Geometric Transformation
  - Frame Order
- Architecture pipelines
  - End-to-End
  - Memory bank
  - Momentum Encoder
  - Clustering

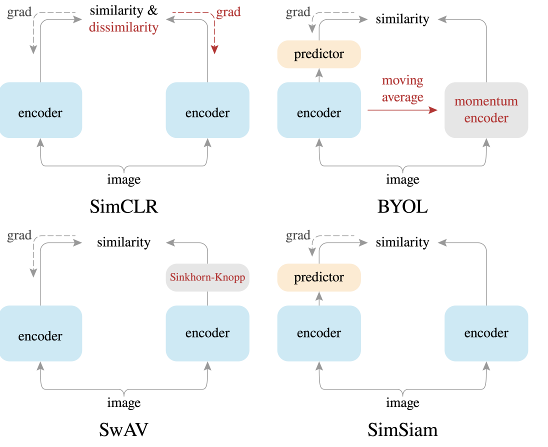

Figure. Comparison on CL architectures [textcitechen2021exploring].

Definition 1.1 (InfoNCE Loss)

The InfoNCE Loss Reads:

$$L_{\text{InfoNCE}} = \sum_{i=1}^{n} -\log \frac{\exp(s_{ii}/\tau)}{\exp(s_{ii}/\tau) + \sum_{k \neq i} \exp(s_{ik}/\tau)} \tag{1}$$

where $s_{ij} = \mathrm{sim}(z_i, z_j')$ *is a similarity between the feature* z_i *and* z_j' *from 2 semantically related data.*

Definition 1.2 (Alignment and Uniformity)

[Wang and Isola (2020)] view CL as enforcing 2 properties:

$$L_{\text{align}} = \sum_{i} ||z_i - z_i'||_2^2 \quad \textit{and} \quad L_{\text{uniform}} = \log \sum_{i,j} e^{2||z_i - z_j'||_2^2} \tag{2}$$

Optimal Transport (OT)

Introduction

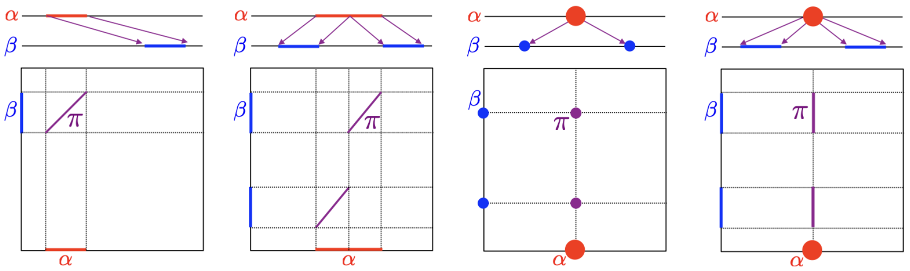

Figure. Four simple examples of optimal couplings between 1-D distributions, represented as maps above (arrows) and couplings below [From Peyré and Cuturi (2019)].

Definition 2.1 (Kantorovich's Optimal Transport Problem)

given the cost matrix $\mathbf{C}$, *Kantorovich's OT involves solving the coupling* $\mathbf{P}$

$$\min_{\mathbf{P}\in U(\mathbf{a},\mathbf{b})} \langle \mathbf{C}, \mathbf{P} \rangle = \sum_{i=1}^{n} \sum_{j=1}^{m} \mathbf{C}_{ij}\mathbf{P}_{ij} \tag{3}$$

where

$$U(\mathbf{a},\mathbf{b}) = \{\mathbf{P} \in \mathbb{R}_{+}^{n\times m} \mid \mathbf{P}\mathbf{1}_m = \mathbf{a},\ \mathbf{P}^T\mathbf{1}_n = \mathbf{b}\} \tag{4}$$

Optimal Transport (OT)
Formalism

Definition 2.1 (Kantorovich's Optimal Transport Problem)

given the cost matrix **C***, Kantorovich's OT involves solving the coupling* **P**

$$\min_{\mathbf{P}\in U(\mathbf{a},\mathbf{b})} \langle \mathbf{C}, \mathbf{P} \rangle = \sum_{i=1}^{n}\sum_{j=1}^{m} \mathbf{C}_{ij}\mathbf{P}_{ij} \tag{3}$$

where

$$U(\mathbf{a},\mathbf{b}) = \{\mathbf{P} \in \mathbb{R}_{+}^{n\times m} \mid \mathbf{P}\mathbf{1}_m = \mathbf{a},\ \mathbf{P}^T\mathbf{1}_n = \mathbf{b}\} \tag{4}$$

- When $n = m$ and $a = b = \mathbf{1}/n$, the OT is equivalent to solve a **balanced matching problem**.
- Similarly, we introduce some relaxation of $U(\mathbf{a}, \mathbf{b})$:

$$U(1) = \{\mathbf{P} \in \mathbb{R}_{+}^{n\times m} \mid \mathbf{1}_n^T\mathbf{P}\mathbf{1}_m = 1\} \quad \text{and} \quad U(\mathbf{a}) = \{\mathbf{P} \in \mathbb{R}_{+}^{n\times m} \mid \mathbf{P}\mathbf{1}_m = \mathbf{a}\}$$

Here we introduce **Regularization** which is highly suited to execution of GPU.

Definition 2.2

The objective reads:

$$\min_{\mathbf{P}\in U}\langle \mathbf{C}, \mathbf{P}\rangle - \epsilon H(\mathbf{P}) \tag{5}$$

where

$$H(\mathbf{P}) = -\sum_{i,j}\mathbf{P}_{ij}(\log \mathbf{P}_{ij} - 1). \tag{6}$$

- We call H Entropic Regularization. The function H is 1-strongly concave.
- Other form of Regularization: $G(\mathbf{P}) = \sum_{i,j}\left(-\frac{1}{2}\mathbf{P}_{ij}^2 + \mathbf{P}_{ij}\right)$.It's also 1-strongly concave.

Optimal Transport (OT)

Inverse Optimal Transport

What if ...

- The cost matrix $\mathbf{C}$ is unknown but the coupling $\mathbf{P}$ is known?

What if ...

- The cost matrix **C** is unknown but the coupling **P** is known?

Definition 2.3 (IOT problem)

By construct a min-min problem:

$$\min_{\theta} KL(\tilde{\mathbf{P}}|\mathbf{P}^{\theta}) \quad \textit{where} \quad \mathbf{P}^{\theta} = \arg\min_{\mathbf{P}\in U} \langle \mathbf{C}^{\theta}, \mathbf{P} \rangle - \epsilon H(\mathbf{P}) \tag{7}$$

where $\tilde{P}$ is the ground truth and θ represents learnable parameters. Trivially, use Kullback–Leibler divergence to measure the distance between different distributions.

IOT Inspires CL

Example 2.1 (Equation 7 when $U = U(\mathbf{a})$)

The Lagrangian of the equation 7 reads:

$$L(\mathbf{P}, \lambda) = \langle \mathbf{C}^\theta, \mathbf{P} \rangle - \epsilon H(\mathbf{P}) - \sum_{i=1}^{n} \lambda_i \left(\sum_{j=1}^{m} \mathbf{P}_{ij} - \frac{1}{n} \right) \tag{8}$$

Through $\partial L / \partial P_{ij} = 0$, we have

$$\mathbf{P}_{ij}^\theta = \frac{\exp\left(-\mathbf{C}_{ij}^\theta / \epsilon\right)}{n \sum_{t=1}^{m} \exp\left(-\mathbf{C}_{it}^\theta / \epsilon\right)} \tag{9}$$

Setting $\tilde{\mathbf{P}} = \mathrm{diag}(1/n, \ldots, 1/n)$, the KL-divergence becomes:

$$KL(\tilde{\mathbf{P}} | \mathbf{P}^\theta) = -\frac{1}{n} \sum_{i=1}^{n} \log \left(\frac{\exp\left(-\mathbf{C}_{ii}^\theta / \epsilon\right)}{\sum_{t=1}^{m} \exp\left(-\mathbf{C}_{it}^\theta / \epsilon\right)} \right) + \mathrm{Const} \tag{10}$$

IOT Inspires CL

Example 2.1 (Equation 7 when $U = U(\mathbf{a})$)

The Lagrangian of the equation 7 reads:

$$L(\mathbf{P}, \lambda) = \langle \mathbf{C}^\theta, \mathbf{P} \rangle - \epsilon H(\mathbf{P}) - \sum_{i=1}^{n} \lambda_i \left(\sum_{j=1}^{m} \mathbf{P}_{ij} - \frac{1}{n} \right) \tag{8}$$

Through $\partial L / \partial P_{ij} = 0$, we have

$$\mathbf{P}_{ij}^\theta = \frac{\exp\left(-\mathbf{C}_{ij}^\theta / \epsilon\right)}{n \sum_{t=1}^{m} \exp\left(-\mathbf{C}_{it}^\theta / \epsilon\right)} \tag{9}$$

Setting $\tilde{\mathbf{P}} = \mathrm{diag}(1/n, \ldots, 1/n)$, the KL-divergence becomes:

$$KL(\tilde{\mathbf{P}} | \mathbf{P}^\theta) = -\frac{1}{n} \sum_{i=1}^{n} \log \left(\frac{\exp(-\mathbf{C}_{ii}^\theta / \epsilon)}{\sum_{t=1}^{m} \exp(-\mathbf{C}_{it}^\theta / \epsilon)} \right) + \text{Const} \tag{10}$$

- which is InfoNCE Loss!

Optimal Transport (OT)

IOT Inspires CL

Similarly, we could obtain the loss function under the circumstances in which $U = U(1)$.

$$L_H^{U(1)} = -\frac{1}{n}\sum_{i=1}^{n}\log\left(\frac{n\exp\left(-\mathbf{C}_{ii}^{\theta}/\epsilon\right)}{\sum_{t=-1}^{n}\sum_{s=1}^{m}\exp\left(-\mathbf{C}_{ts}^{\theta}\right)/\epsilon}\right) \tag{11}$$

However, if $U = U(\mathbf{a}, \mathbf{b})$, the closed-form coupling may not exists. We adopt Sinkhorn algorithm [Appendix 1] to approximate:

$$L_H^{\mathrm{U(a,b)}} = -\sum_{i=1}^{m}\sum_{j=1}^{n}\log\tilde{\mathbf{P}}_{ij}\left(\mathbf{P}_{ij}^{\theta}\right)^{K} \tag{12}$$

Use other regularization, like $G(\mathbf{P}) = \sum_{i,j}\left(-\frac{1}{2}\mathbf{P}_{ij}^2 + \mathbf{P}_{ij}\right)$

$$L_G^{U(a)} = -\frac{1}{n}\sum_{i=1}^{n}\log\left(\frac{1}{n}(1+\sum_{j=1}^{n}(\mathbf{C}_{ij}^{\theta}-\mathbf{C}_{ii}^{\theta}))\right) \quad \text{and} \quad L_G^{U(1)} = -\frac{1}{n}\sum_{i=1}^{n}\log\left(\frac{1}{n^2}(n+\sum_{s,t}(\mathbf{C}_{st}^{\theta}-\mathbf{C}_{ii}^{\theta}))\right)$$

Balanced Matching

Balanced Matching IOT-CL Loss

Given two collective points sets within 2N data points, we get features $\{z_i\}_{i=1}^N$ and $\{z_i^*\}_{i=1}^N$.
We denote $\tilde{z}_{2k-1} = z_k$, $\tilde{z}_{2k} = z_k^*$.

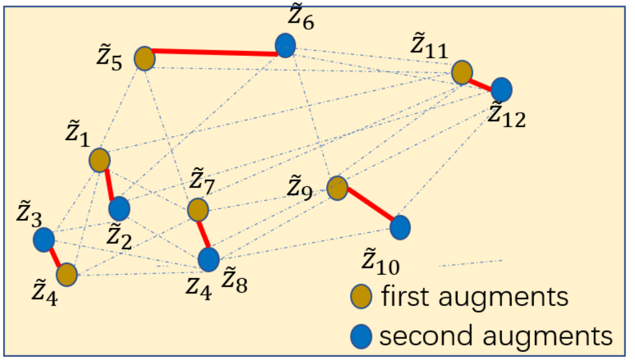

Figure. Balanced Matching

On the 2N data points,the cost matrix and $\tilde{\mathbf{P}}_{ij}$ are defined as follows:

$$\mathbf{C}^{\theta}_{ij} = \begin{cases} +\infty & \text{if} \quad i = j \\ 1 - \tilde{s}_{ij} & \text{otherwise} \end{cases} \quad \text{and} \quad \tilde{\mathbf{P}}_{ij} = \begin{cases} 1/2N & \text{if} \quad (i,j) \in S \\ 0 & \text{otherwise} \end{cases} \tag{13}$$

where $S = S_1 \cup S_2$, S_1 and S_2 are defined the cost matrix and coupling as below:

$$S_1 = \Big\{(i,j) \mid i = 2k,\ j = 2k-1\Big\}\ S_2 = \Big\{(i,j) \mid i = 2k-1,\ j = 2k\Big\} \quad \text{where} \quad k = 1, ..., N \tag{14}$$

Here we introduce different IOT-CL loss function for balanced matching problem under different constraint relaxations .

IOT-CL Loss Under U(a)

$$L_{\mathrm{IOT-CL}}^{\mathrm{U(a)}} = -\frac{1}{2N} \sum_{(i,j)\in S} \log \left(\frac{\exp\left(-\mathbf{C}_{ij}^{\theta}/\epsilon\right)}{\sum_{s=1}^{2N} \mathbb{1}_{i\neq s} \exp\left(-\mathbf{C}_{is}^{\theta}/\epsilon\right)} \right) \tag{15}$$

This is the same as NT-Xent loss in SimCLR [Chen et al. (2020)].

IOT-CL Loss Under U(1)

$$L_{\mathrm{IOT-CL}}^{\mathrm{U(1)}} = -\frac{1}{2N} \sum_{(i,j)\in S} \log \left(\frac{2N \exp\left(-\mathbf{C}_{ij}^{\theta}/\epsilon\right)}{\sum_{s=1}^{2N} \sum_{t=1}^{2N} \mathbb{1}_{s\neq t} \exp\left(-\mathbf{C}_{st}^{\theta}/\epsilon\right)} \right) \tag{16}$$

IOT-CL Loss Under U(a,b)

$$L_{\mathrm{IOT-CL}}^{\mathrm{U(a,b)}} = -\frac{1}{2N} \sum_{(i,j)\in S} \log \left(\mathbf{P}_{ij}^{\theta} \right)^{K} \tag{17}$$

where $(\mathbf{P}_{ij}^{\theta})^{K}$ is solved by Sinkhorn algorithm, K is the iteration number.

IOT-CL Loss Under Gradient Constraint Relaxation

Under different constraint relaxations, $\mathbf{P}_{ij}^{\theta}$ will learn to approximate $\tilde{\mathbf{P}}_{ij}$ in different ways. We designed two gradient losses as below.

$$\begin{aligned} L_{\mathrm{IOT-CL}}^{\mathrm{Tighten}} &= L^{\mathrm{U(a)}} \rightarrow L^{\mathrm{U(a,b),K=1}} \rightarrow L^{\mathrm{U(a,b),K=2}} \rightarrow L^{\mathrm{U(a,b),K=4}} \rightarrow L^{\mathrm{U(a,b),K=8}} \\ L_{\mathrm{IOT-CL}}^{\mathrm{Relax}} &= L^{\mathrm{U(a,b),K=8}} \rightarrow L^{\mathrm{U(a,b),K=4}} \rightarrow L^{\mathrm{U(a,b),K=2}} \rightarrow L^{\mathrm{U(a,b),K=1}} \rightarrow L^{\mathrm{U(a)}} \end{aligned} \tag{18}$$

where the changing interval is fixed.

IOT-CL Loss Under Gradient Epsilon

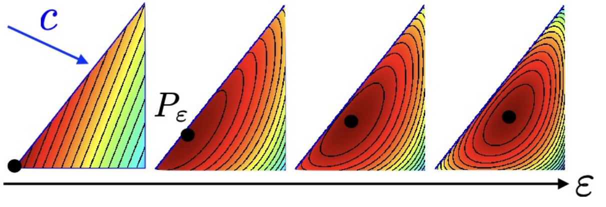

Figure. Entropic Regularization [Peyré and Cuturi (2019)].

Above figure illustrates the effect of the entropy to regularize a linear program over the simplex $\sum_3$. We can see the entropy pushes the original LP solution away from the boundary of the triangle.

Thus the ϵ will control the "sharpness" of $\mathbf{P}_{ij}^{\theta}$. A gradient setting will help $\mathbf{P}_{ij}^{\theta}$ move to $\tilde{\mathbf{P}}_{ij}$ faster.

Balanced Matching

Alignment And Uniformity

We rethink the alignment and uniformity [Wang and Isola (2020)] from the matching prospective. By adding the uniformity penalty, we can get alignment and uniformity loss with matching view:

$$\min_{\theta} L_{\mathrm{IOT-CL}}^{\mathrm{Uniform}} = L_{\mathrm{IOT-CL}} + \lambda_p KL(\bar{\mathbf{Q}}^{\theta} | \mathbf{P}^{\theta}) \tag{19}$$

where

$$\bar{\mathbf{Q}}_{ij}^{\theta} = \begin{cases} \mathbf{P}_{ij}^{\theta}, & \text{positive pair} \\ \underset{\text{negative pair}}{\mathrm{mean}} \mathbf{P}_{ij}^{\theta}, & \text{negative pair} \end{cases} \tag{20}$$

The uniformity penalty will enhance our IOT-CL loss. The detailed experiment results will be available in our final paper.

Balanced Matching

Visualization

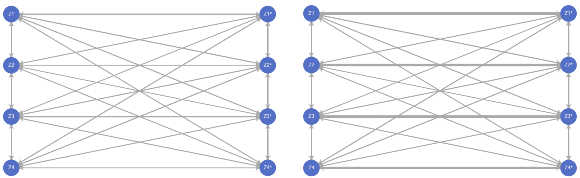

Figure. Visualization on balanced pair matching.

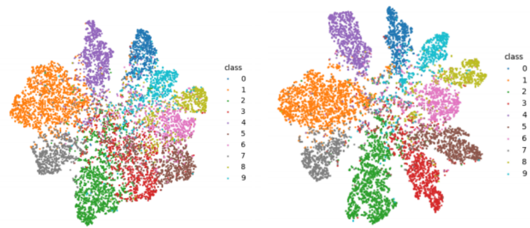

Figure. The effect of the Uniformity loss

References I

Chen, Ting et al. (2020). "A Simple Framework for Contrastive Learning of Visual Representations". In: *arXiv preprint arXiv:2002.05709*.

Peyré, Gabriel and Marco Cuturi (2019). "Computational Optimal Transport". In: *Foundations and Trends in Machine Learning* 11.5-6, pp. 355–607.

Wang, Tongzhou and Phillip Isola (2020). "Understanding contrastive representation learning through alignment and uniformity on the hypersphere". In: *International Conference on Machine Learning*. PMLR, pp. 9929–9939.

Appendix

Sinkhorn Algorithm

Firstly, initialize $\mathbf{P}^\theta$:

$$(\mathbf{P}^\theta)^0 = \exp\left(-\mathbf{C}^\theta/\epsilon\right) \tag{21}$$

Then update it step by step:

$$\begin{aligned} \left(\mathbf{P}^\theta\right)^k &\leftarrow \frac{1}{n}\left(\mathbf{P}^\theta\right)^{k-1} \oslash \left(\left(\mathbf{P}^\theta\right)^{k-1}\mathbf{1}_{m\times m}\right) \\ \left(\mathbf{P}^\theta\right)^k &\leftarrow \frac{1}{m}\left(\mathbf{P}^\theta\right)^{k} \oslash \left(\left(\mathbf{1}_{n\times n}\mathbf{P}^\theta\right)^{k}\right) \end{aligned} \tag{22}$$

where the symbol $\oslash$ represents element-wise division.