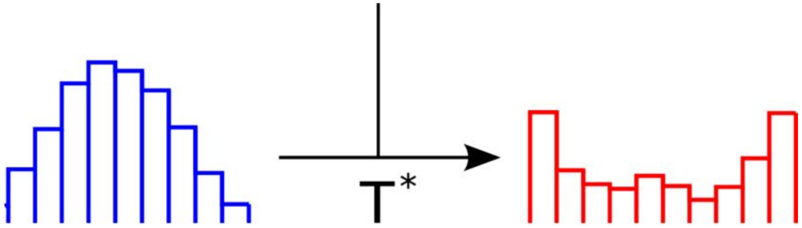
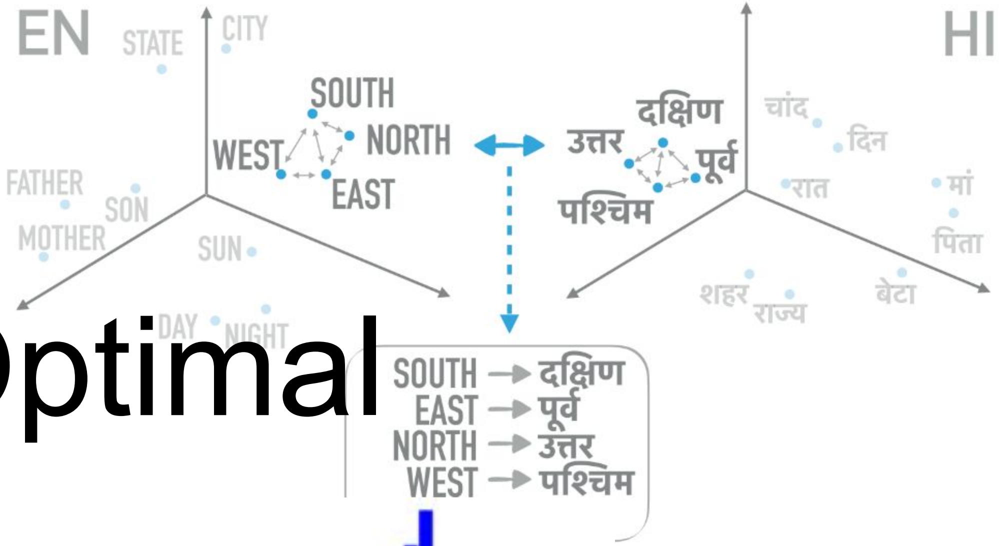
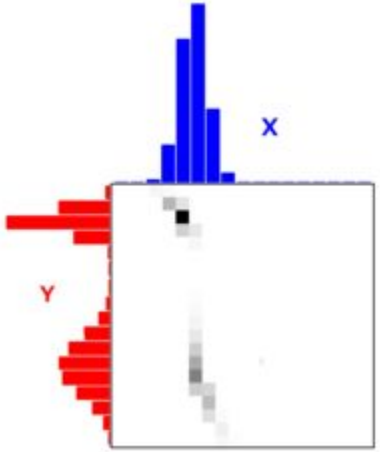


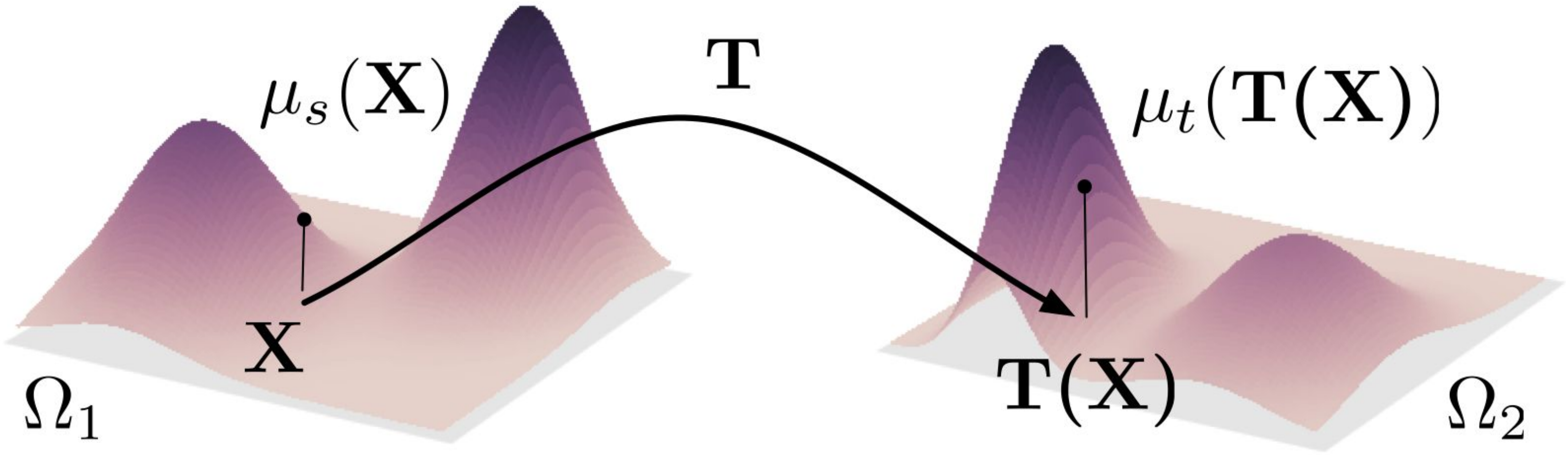


Structured Optimal Transport



~ Presented by

M S Nishanth
CS18MTECH11005



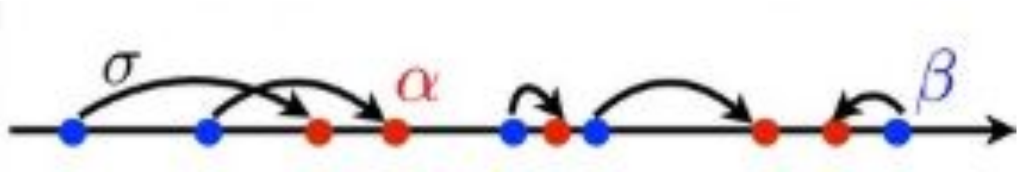
$\mu_s(\mathbf{X})$
$\mathbf{T}$
$\mu_t(\mathbf{T}(\mathbf{X}))$
$\mathbf{X}$
$\mathbf{T}(\mathbf{X})$
Ω_1
Ω_2

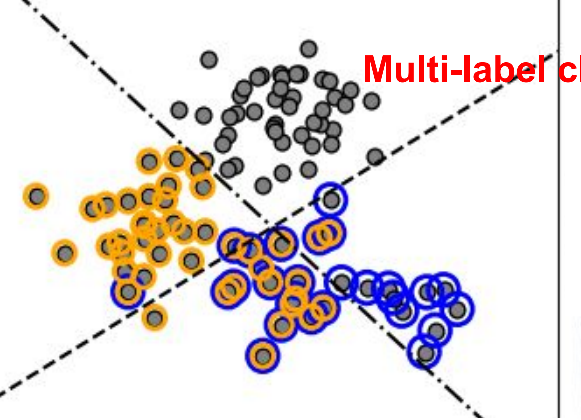
Optimal Transport Problem

Given the transportation of one unit of the commodity from Pi to Mj costs cij . How to transport the required quantity of the commodity at the lowest cost?

Optimal transport provides a natural, elegant framework for comparing probability distributions while respecting the underlying geometry [1]

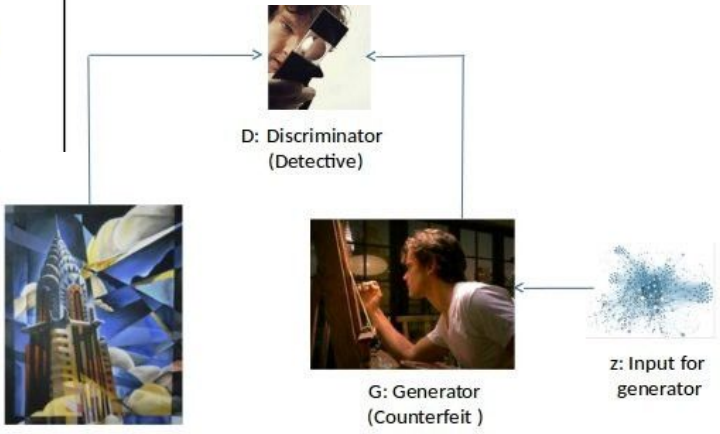
[1]: https://optimaltransport.github.io/book/


σ
α
β



Multi-label classification

Adversarial Neural Network



Domain Adaption

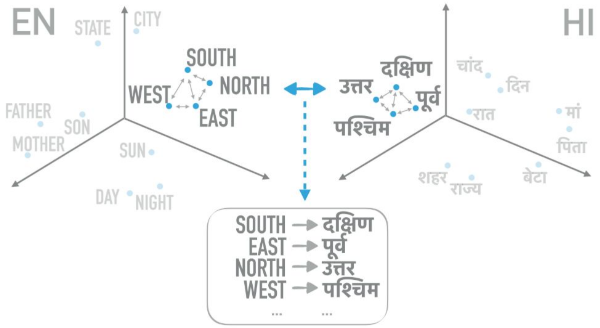


Image analysis

** All images from Google

Structured Optimal Transport

To capture the structure of the settings.

1. Intrinsic: if the distributions correspond to structured objects (e.g., images with segments, or sequences)
2. Extrinsic: if there is side information that induces structure (e.g. groupings).

example: Domain Adaptation.

Goal

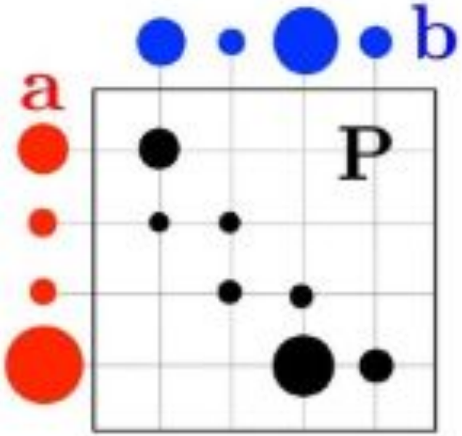
To incorporate structural information directly into the optimal transport problem.

Formulation of Optimal Transport

Monge's Optimal Transport Problem: *given* $\mu \in \mathcal{P}(X)$ *and* $\nu \in \mathcal{P}(Y)$,

$$\textit{minimise } \mathbb{M}(T) = \int_X c(x, T(x))\, \mathrm{d}\mu(x)$$

over μ*-measurable maps* $T : X \to Y$ *subject to* $\nu = T_{\#}\mu$.

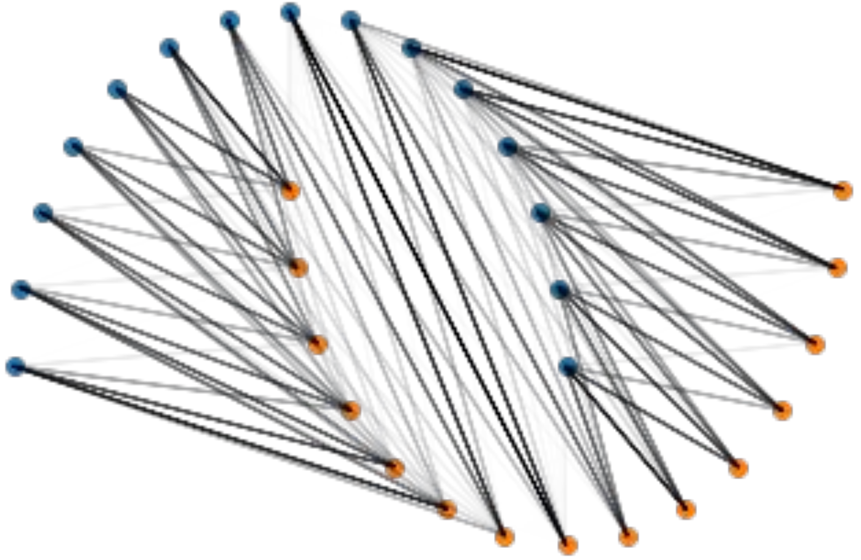


Kantorovich's Optimal Transport Problem: *given* $\mu \in \mathcal{P}(X)$ *and* $\nu \in \mathcal{P}(Y)$,

$$\textit{minimise } \mathbb{K}(\pi) = \int_{X \times Y} c(x, y)\, \mathrm{d}\pi(x, y) \qquad \textit{over } \pi \in \Pi(\mu, \nu).$$

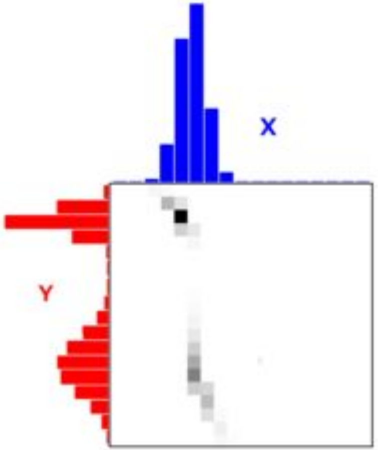
** Introduction to Optimal Transport Matthew Thorpe

Img source: Google



Img source: Google

$$\inf_{\gamma} \left\{ \int_{\mathcal{X}\times\mathcal{Y}} c(x,y)\, d\gamma(x,y) \mid \gamma \in \Gamma(\mu,\nu) \right\},$$



The cost function only needs to be specified for every pair $(\mathbf{x}_i^s, \mathbf{x}_j^t)$, i.e., it is a matrix $C \in \mathbb{R}^{n\times m}$, and the total transportation cost incurred by γ is $\sum_{ij} \gamma_{ij} c_{ij}$. Thus, the discrete optimal transport (DOT) problem consists of finding a transport plan that solves

$$\min_{\gamma \in \mathcal{M}_{\mu,\nu}} \langle \gamma, C \rangle.$$

** https://optimaltransport.github.io/book/

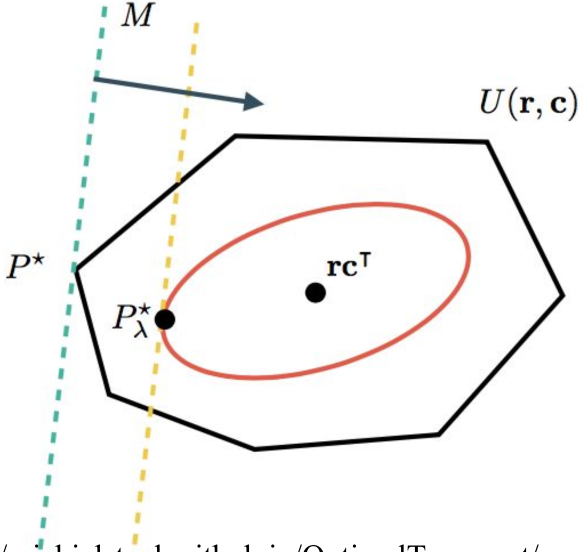
Img source: Google

Entropic Regularization

A matrix with low entropy will be sparser, Conversely, a matrix with high entropy will be smoother [with the maximum entropy achieved with a uniform distribution of values across its elements].

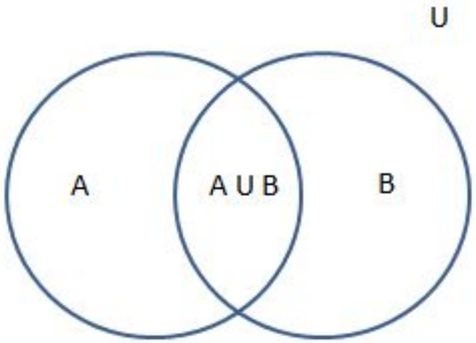
With a regularization coefficient, we can include this in the optimal transport problem to encourage smoother coupling matrices.

$$\min_{\gamma \in \mathcal{M}} \langle \gamma, C \rangle - \frac{1}{\lambda} H(\gamma).$$



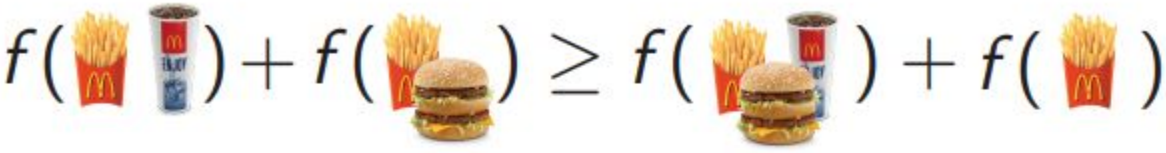
Img source: https://michielstock.github.io/OptimalTransport/

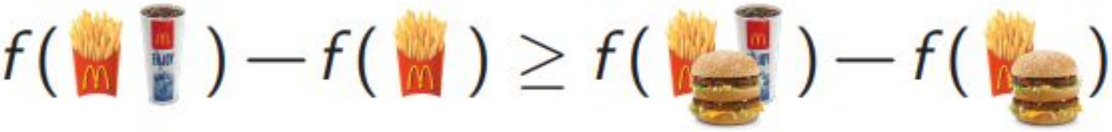
Modular function



$$n(A \cup B) = n(A) + n(B) - n(A \cap B)$$

Submodularity





Why Submodularity?

- It allows us to encode various types of structural information in the cost function.

- Ex: Let $V = \{v_1, v_2\}$ be a set of actions with:

v_1 = "buy milk at the store" v_2 = "buy honey at the store"


- For $A \subseteq V$, let $f(A)$ be the consumer cost of set of items A.
- $f(\{v_1\})$ = cost to drive to and from store c_d, and cost to purchase milk c_m, so $f(\{v_1\}) = c_d + c_m$.
- $f(\{v_2\})$ = cost to drive to and from store c_d, and cost to purchase honey c_h, so $f(\{v_2\}) = c_d + c_h$.
- But $f(\{v_1, v_2\}) = c_d + c_m + c_h < 2c_d + c_m + c_h$ since c_d (driving) is a shared fixed cost.
- Shared fixed costs are submodular: $f(v_1) + f(v_2) \geq f(v_1, v_2) + f(\emptyset)$

How?

We try a matching of variables in U (source) and V (target) with minimal cost. Here any matching can be expressed as a set of edges S = {(u1, v1), . . . ,(uk , vk)}, and its cost as a set function F : 2|U|×|V| → R+.

We divide the support of the source and target distributions μ and v into regions Uk ⊂ U and Vl ⊂ V.

$$E_{kl} := \{(u, v) \mid u \in U_k,\ v \in V_l\}$$

And calculate the cost as:

$$F(S) := \sum_{kl} F_{kl}(S \cap E_{kl}),$$

Here, each Fkl is a Submodular function

Formula: Structured Optimal Transport David Alvarez-Melis, Tommi S. Jaakkola, Stefanie Jegelka

Computation time for this?

{a,b,c}

$$\left\{ \begin{array}{l} \{\}, \\ \{a\}, \{b\}, \{c\}, \\ \{a,b\}, \{a,c\}, \{b,c\}, \\ \{a,b,c\} \end{array} \right\}$$

Lovasz Extension and Submodularity

If F is submodular, the Lovasz extension is equivalent to the support function.

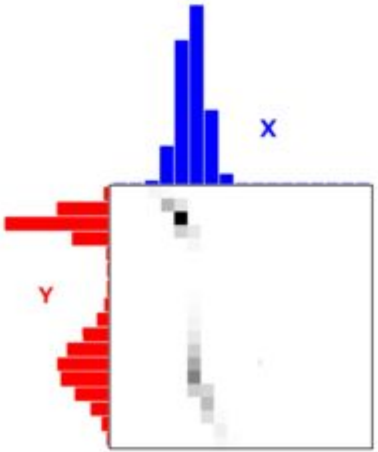
And this is convex. In Fact a conic.

$$f(w) = \max_{x \in \mathcal{B}_F} w^T x,$$

** https://optimaltransport.github.io/book/

Previously, our optimization function was

$$\inf_{\gamma} \left\{ \int_{\mathcal{X}\times\mathcal{Y}} c(x,y)d\gamma(x,y) \mid \gamma \in \Gamma(\mu,\nu) \right\}$$ Converted to $$\min_{\gamma\in\mathcal{M}_{\mu,\nu}} \langle \gamma, C \rangle.$$



We know that this is a convex function we can apply lovasz extension and get,

$$\min_{\gamma\in\mathcal{M}} f(\gamma) \equiv \min_{\gamma\in\mathcal{M}} \max_{\kappa\in\mathcal{B}_F} \langle \gamma, \kappa \rangle.$$

Mirror Descent Algorithm

The algorithm has 2 steps. Which is iterated to get an optimal solution.

1. Calculate gradient for lovasz extension.
2. Sinkhorn projection of the Gradient.

Calculating gradient for lovasz extension.

1. sort

$$x_{\pi(1)} \geq x_{\pi(2)} \geq \ldots \geq x_{\pi(n)}$$

2. chain of sets

$$S_0 = \emptyset,\ S_i = \{\pi(1), \ldots, \pi(i)\}$$

3. assign values

$$y_{\pi(i)} = F(S_i) - F(S_{i-1})$$

$$F(S) = \max\{|S|, 1\}$$

$$\overset{x}{\begin{bmatrix} 0.5 \\ 1.0 \end{bmatrix}} = 0.5 \begin{bmatrix} 1.0 \\ 1.0 \end{bmatrix} + 0.5 \begin{bmatrix} 0 \\ 1.0 \end{bmatrix}$$

sort: $x_2 \geq x_1 \quad \Rightarrow \quad S_1 = \{2\},\ S_2 = \{2, 1\}$

$$y_2 = F(2) = 1$$

$$y_1 = F(2,1) - F(2) = 1 - 1 = 0$$

$$f(x) = y^\top x = 1 \cdot x_1 + 0 \cdot x_2 = \max_i x_i$$

Sinkhorn projection

- The solution is easily derivable.
- Solution by adding Lagrange’s multipliers and solving.

Results from paper

Submod OT

Shortcomings of paper

- The original OT is a linear program, whereas the structured OT is NOT.
- The claim is that we can introduce structure and handle the same applications as SOT without changing the form of the OT to a non-linear program

Our method

We propose to change the cost function to that of an induced norm from the paper.

- structured sparsity-inducing norms through submodular functions.

** Structured sparsity-inducing norms through submodular functions Francis Bach

references

1. Cuturi, Marco. "Sinkhorn distances: Lightspeed computation of optimal transport." *Advances in neural information processing systems*. 2013.
2. Alvarez-Melis, David, Tommi S. Jaakkola, and Stefanie Jegelka. "Structured optimal transport." *arXiv preprint arXiv:1712.06199* (2017).
3. Bach, Francis R. "Structured sparsity-inducing norms through submodular functions." *Advances in Neural Information Processing Systems*. 2010.
4. Some online materials:
   a. https://www.youtube.com/watch?v=ZZT3bQ8BgV4
   b. https://www.youtube.com/watch?v=sMWQkl0p6XM
   c. https://dfdazac.github.io/sinkhorn.html
   d. https://michielstock.github.io/OptimalTransport/
   e. https: //github.com/rflamary/POT
   f. Submodular Functions – Part II ML Summer School Cádiz Stefanie Jegelka

Thank you

$$\inf_{\gamma} \left\{ \int_{\mathcal{X}\times\mathcal{Y}} c(x,y) d\gamma(x,y) \mid \gamma \in \Gamma(\mu,\nu) \right\}, \tag{2}$$

where $\Gamma(\mu,\nu)$ is the set of *transportation plans*, i.e., joint distributions with marginals μ and ν. If μ and ν are only available through discrete samples $\{\mathbf{x}_i^s\}_{i=1}^n$, $\{\mathbf{x}_i^t\}_{i=1}^m$, the empirical distributions can be written as

$$\mu = \sum_{i=1}^{n} p_i^s \delta_{\mathbf{x}_i^s}, \quad \nu = \sum_{i=1}^{m} p_i^t \delta_{\mathbf{x}_i^t} \tag{3}$$

where p_i^s, p_i^t are the probabilities associated with the samples. It is easy to adapt Kantorovich's formulation to this discrete setting. In this case, the space of transportation plans is a polytope:

$$\mathcal{M}_{\mu,\nu} = \{\gamma \in \mathbb{R}_+^{n\times m} \mid \gamma\mathbf{1} = \mu,\ \gamma^T\mathbf{1} = \nu\} \tag{4}$$

The cost function only needs to be specified for every pair $(\mathbf{x}_i^s, \mathbf{x}_j^t)$, i.e., it is a matrix $C \in \mathbb{R}^{n\times m}$, and the total transportation cost incurred by γ is $\sum_{ij} \gamma_{ij} c_{ij}$. Thus, the discrete optimal transport (DOT) problem consists of finding a transport plan that solves

$$\min_{\gamma\in\mathcal{M}_{\mu,\nu}} \langle \gamma, C \rangle. \tag{5}$$

Applications of optimal transport distances: Shape analysis [Gangbo and McCann, 2000], image registration and interpolation [Solomon et al., 2015], domain adaptation [Courty et al., 2017], adversarial neural networks [Arjovsky et al., 2017], and multi-label prediction [Frogner et al., 2015]... etc