OT-based Distillation

How to formulate LM-RNN distillation as an optimal transport problem,
more precisely, a fused gromov wasserstein 1-barycenter problem,
with a single distribution and no marginal constraint on the barycenter.

2021-11-16

Rémi Emonet

TAUDoS Meeting @ Saint-Étienne

LM-RNN Distillation Reminders

We start from a learned recurrent model *(presentation by Sri)*

- We can sample sequences on demand
- We gather an "infinity" of *(one for every token in every sequence we generate)*
  - **points** / latent vectors / hidden states: the internal representation of the LM-RNN
  - **edges** / transitions: from a point to another, annotated with a token/letter

Goal: use this dataset to "learn" an automata (PFA)

Remarks

- a good baseline is k-means + stats on transitions
- the actual graph is a tree (but we don’t use that)

(Wasserstein) Barycenter

Given B distributions $\{\mu^b\}_b$, and weights $\{\lambda_b\}_b$ (with $\sum_b \lambda_b \neq 0$)

$$\arg\min_{\nu} \sum_{b=1}^{B} \lambda_b W(\mu^b, \nu)$$

1-barycenter, B = 1

$$\arg\min_{\nu} W(\mu, \nu)$$

We can parametrize/constrain the form ν (e.g. few discrete diracs, small graph for GW, ...)

K-means

$$\arg\min_{\{c_k\}_k, \{z_i\}_i} \sum_{i=1}^{N} d(x_i, c_{z_i})^2$$

- c_k : position of the k^{th} cluster mean
- z_i : index of the center that is closest to point x_i

Wasserstein 1-Barycenter

$$\arg\min_{\{c_k\}_k, T\in\Pi} \sum_{i=1}^{N} \sum_{k=1}^{K} d(x_i, c_k)^2 T_{ik}$$

- c_k the position of the k^{th} cluster mean
- T_{ik} the mass of point i that is sent to center k
  - considering the vector $T_{i.}$
  - the optimal is

    to set the whole mass to the closest k

    i.e., $T_{ik} = 0, \;\forall k \neq z_i$
- Notes on Π
  - we do not constain/fix the marginal "on k"

    (the cluster mass/weight is not fixed)

Fused-GW 1-Barycenter

The formulation that does distillation.

Principle: a 1-barycenter formulation, with

- a Wasserstein term (k-means like)
  - data: $\{x_i\}_i$ in the latent space
  - barycenter: "cluster means" $\{c_k\}_k$ in the latent space
- a Gromov-Wasserstein term (graph reduction)
  - data: $\{d_{ii'}\}_{i,i'}$ observed transitions (token, one-hot encoded)
  - barycenter with edges between clusters described with $\{d_{kk'}\}_{k,k'}$ (distribution)
  - a loss $l_{\mathbf{comp}}$, to be defined, unperfectly set to l_2^2 for now
- a weighting of these two terms, controlled by α, an hyper-parameter

$$\underset{\{c_k\}_k,\{d_{kk'}\}_{k,k'},T\in\Pi}{\arg\min} \quad \alpha \sum_{i=1}^{N} \sum_{k=1}^{K} d(x_i, c_k)^2 T_{ik} + (1-\alpha) \sum_{i=1}^{N} \sum_{i'=1}^{N} \sum_{k=1}^{K} \sum_{k'=1}^{K} l_{\mathbf{comp}}(d_{ii'}, d_{kk'}) T_{ik} T_{i'k'}$$

Optimization Algorithm

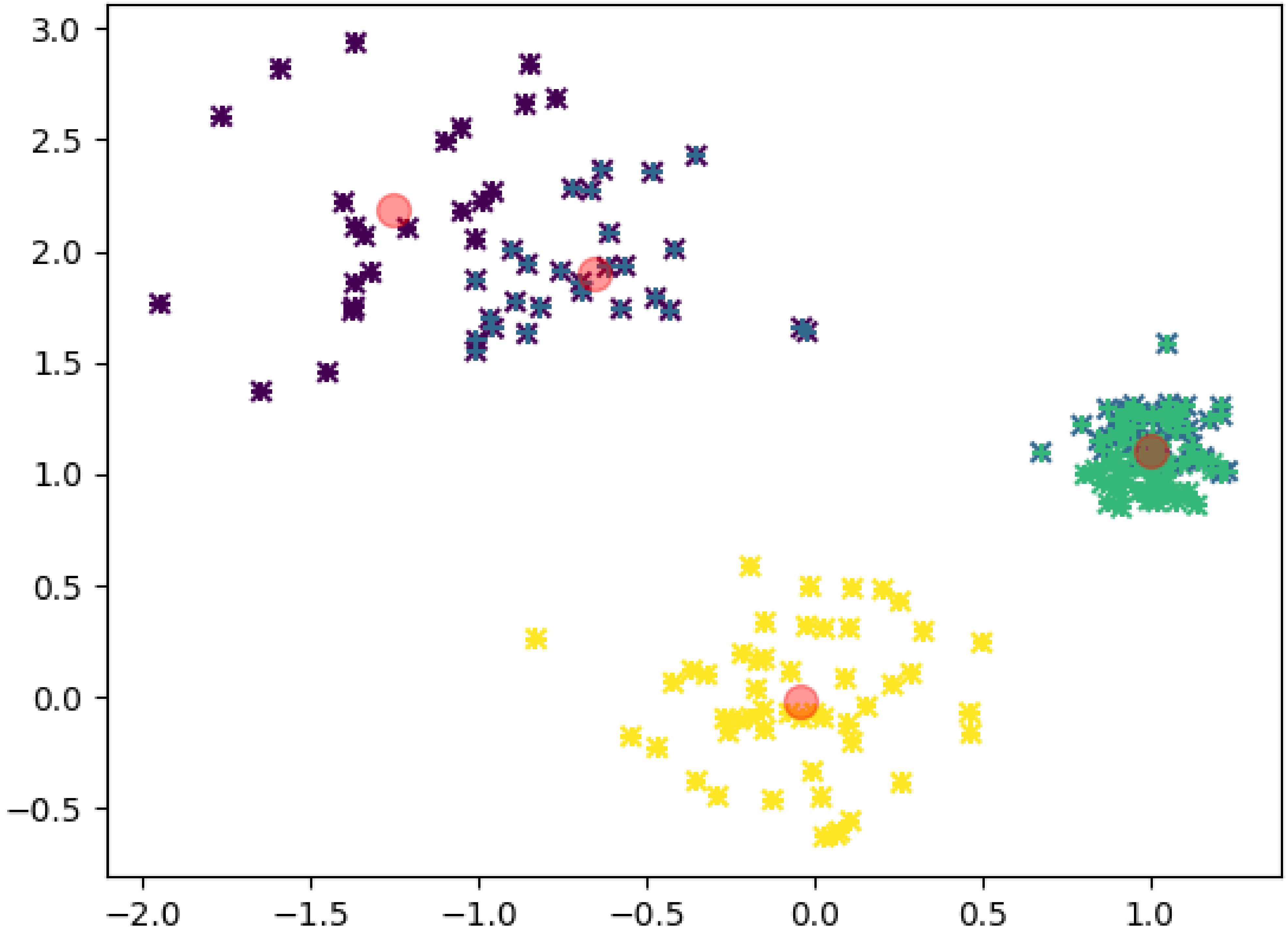
Alternating estimation of T and $\{c_k, d_{kk'}\}$ *Credit: Tanguy Kerdoncuff*

$$\underset{\{c_k\}_k, \{d_{kk'}\}_{k,k'}, T \in \Pi}{\arg\min} \quad \alpha \sum_{i=1}^{N} \sum_{k=1}^{K} d(x_i, c_k)^2 T_{ik} + (1-\alpha) \sum_{i=1}^{N} \sum_{i'=1}^{N} \sum_{k=1}^{K} \sum_{k'=1}^{K} l_{\mathbf{comp}}(d_{ii'}, d_{kk'}) T_{ik} T_{i'k'}$$

- Initialize with a random T
- Repeat
  - update, with T fixed
    - c_k as T-weighted means (1)
    - $d_{kk'}$ as in GW 1-barycenter (2)
  - update T with the rest fixed (3)
    - using Frank-Wolfe

(repeat with several initializations)

Illustration with $\alpha = 1.000$ (k-means)



Issues / TO DO

- Used l_2^2 for $l_{\mathbf{comp}}$ -> use a KL
- Scalability -> stochastic version
- Tested on synthetic data -> move to real data
- PFA -> sparsity-inducing $l_{\mathbf{comp}}$ to have a DFA
- more ideas? suggestions?


Discussion, Questions?