Learning Latent Partial Matchings with Gumbel-IPF Networks

AISTATS(2024)

Hedda Cohen Indelman, Tamir Hazan
Reporter: Fengjiao Gong

May 9, 2024

Outline

1. Background
2. Partial Matching
3. Gumbel-IPF Networks
4. Experiments

Background

- **Matchings** is a fundamental building block in a variety of applications, such as alignment and sorting data.
- In principle, **deep networks** can learn arbitrarily sophisticated mappings from inputs to outputs.
- Operations on these discrete objects are approximated using **differentiable** operations on continuous relaxations of the objects.

However, the matching is not provided as supervision for learning models with *latent matchings*.

Analogy

Softmax can approximate the discrete category by continuous values:

$$\text{softmax}_\tau(x)_i = \frac{\exp(x_i/\tau)}{\sum_{j=1} \exp(x_j/\tau)} \tag{1}$$

- $\tau > 0$, a point in the **probability** simplex.
- $\tau \to 0$, converges to **a vertex of the simplex**.

A maximization problem:

$$v^* = \arg\max_i \langle x, v \rangle, \tag{2}$$

where $v^\star$ is a one-hot vector corresponding to the x_i.

Matching

Given two sets V^s and V^t,

- Number of elements, $|V^s| = |V^t| = n$,
- Data instance x^{st} — pair
- Corresponding label $y\left(x^{st}\right)$

$$\mathcal{M}_{st} = \left\{ y\left(x^{st}\right) \in \{0,1\}^{n \times n} : y\left(x^{st}\right) \mathbf{1}_n = y\left(x^{st}\right)^T \mathbf{1}_n = \mathbf{1}_n \right\} \tag{3}$$

Linear assignment problem

Matching

Permutation matrix P

$$M(X) = \arg\max_{P \in \mathcal{P}_N} \langle P, X \rangle_F, \tag{4}$$

where

- $\mathcal{P}_N$ — the set of permutation matrices
- $\langle A, B \rangle_F = \text{trace}\left(A^\top B\right)$ the (Frobenius) inner product of matrices
- $M(\cdot)$ the matching operator

Parameterize the hard choice of the permutation.

Sinkhorn operator

Define the **Sinkhorn operator S(X)**:

$$
\begin{aligned}
S^0(X) &= \exp(X), \\
S^l(X) &= \mathcal{T}_c\left(\mathcal{T}_r\left(S^{l-1}(X)\right)\right), \\
S(X) &= \lim_{l\to\infty} S^l(X).
\end{aligned}
\tag{5}
$$

Sinkhorn (1964) proved that $S(X)$ must belong to **the Birkhoff polytope**, the set of doubly stochastic matrices denoted as $\mathcal{B}_N$.

[Ref] Richard Sinkhorn. A relationship between arbitrary positive matrices and doubly stochastic matrices. The annals of mathematical statistics, 35(2):876–879, 1964.

Sinkhorn operator

Theorem 1. For a doubly-stochastic matrix P, define its entropy as

$$h(P) = -\sum_{i,j} P_{i,j} \log\left(P_{i,j}\right). \tag{6}$$

Then, one has,

$$S(X/\tau) = \arg\max_{P \in \mathcal{B}_N} \langle P, X \rangle_F + \tau h(P) \tag{7}$$

Assume the entries of X are drawn independently from a distribution that is **absolutely continuous** with respect to the Lebesgue measure in $\mathbb{R}$.

Then, almost surely, the following **convergence** holds:

$$M(X) = \lim_{\tau \to 0^+} S(X/\tau) \tag{8}$$

Outline

1. Background
2. **Partial Matching**
3. Gumbel-IPF Networks
4. Experiments

Partial Matching

Given two sets V^s and V^t,

- Number of elements

$$n = |V^s|, m = |V^t| \tag{9}$$

- Data instance x^{st} — pair
- Corresponding label $y\left(x^{st}\right)$

$$\mathcal{M}_{st} = \left\{ y\left(x^{st}\right) \in \{0,1\}^{n \times m} : y\left(x^{st}\right) \mathbf{1}_m \leq \mathbf{1}_n,\ y\left(x^{st}\right)^T \mathbf{1}_n \leq \mathbf{1}_m \right\} \tag{10}$$

In a realistic setting, some elements from V^s and V^t may not be matched.

Partial Matching Prediction

The label space is the **transportation polytope** in $\mathbb{R}^{n \times m}$, denoted $\mathcal{T}_{\bar{u}\bar{v}}$

- Marginals of rows and columns , $\bar{u} \in \{0,1\}^n$ and $\bar{v} \in \{0,1\}^m$.
- Points in $\mathcal{T}_{\bar{u}\bar{v}}$ are described by real $n \times m$ matrices.
- Its vertices $T_1, \ldots, T_r$ are $\{0,1\}^{n \times m}$ matrices with marginals $\bar{u}$ and $\bar{v}$.

Then,

$$\mathcal{T}_{\bar{u}\bar{v}} = \left\{ \sum_{r=1}^{R} \lambda_r T_r ; \sum_{r=1}^{R} \lambda_r = 1, \lambda_r \geq 0 \forall r \right\}. \tag{11}$$

Matching prediction

Learn to **predict** such structured labels,

- Fit a *parameterized* correspondence **scoring function** $\mu_w(x, y)$ of the training instance-label pairs (x, y)
- Minimize the loss $\ell(\cdot, \cdot)$ between label y and the **highest** scoring structure y^*

$$\min \ell(y, y^*) \tag{12}$$

with

$$y^* \left(\mu_w(x, y)\right) = \arg\max_{\hat{y} \in \mathcal{M}} \langle \mu_w(x, y), \hat{y} \rangle \,. \tag{13}$$

Partial Matching Prediction

Define the **highest-scoring partial matching prediction** as

$$y^* \left(\mu_w(x, y)\right) = \arg\max_{S \in \mathcal{T}_{\bar{u}\bar{v}}} \langle \mu_w(x, y), S \rangle \,, \tag{14}$$

which is the **hard choice** of a vertex in the $\mathcal{T}_{\bar{u}\bar{v}}$ polytope, with $\langle \cdot, \cdot \rangle$ the (Frobenius) inner product of matrices.

Partial Matching Prediction

Matching nature:

- At most one 1 in each row and each column.
- Polytope $\mathcal{T}_{\bar{u}\bar{v}}$ can be interpreted as a k-assignment polytope.

$$\sum_i u_i = \sum_j v_j = k. \tag{15}$$

Iterative Proportional Fitting(IPF)

The IPF is an iterative weighting method used to **biproportionally fit** an input matrix so that its row and column marginals agree with target marginals.

$$\begin{aligned} &\min_{\boldsymbol{S}} \sum_{i=1}^{n} \sum_{j=1}^{m} s_{ij} \log\left(\frac{s_{ij}}{z_{ij}}\right) \\ &\text{s.t. } \sum_{j=1}^{m} s_{ij} = u_i \forall i, \quad \sum_{i=1}^{n} s_{ij} = v_j \forall j, \end{aligned} \tag{16}$$

where

- input matrix $\boldsymbol{Z} \in \mathbb{R}^{n \times m}$ with positive entries
- target marginals $\boldsymbol{u} \in \mathbb{R}_{>0}^{n \times 1}$ and $\boldsymbol{v} \in \mathbb{R}_{>0}^{m \times 1}$.

It seeks to find matrix $\boldsymbol{S} \in \mathbb{R}^{n \times m}$ which is closest to $\boldsymbol{Z}$ w.r.t. the Kullback-Leibler distance and has the target marginals.

Iterative Proportional Fitting(IPF) Algorithm

Algorithm

- Initializing $s_{ij}^{(0)} = z_{ij}$
- Normalizing on rows and columns for $t > 0$ iterations

$$\begin{aligned} s_{ij}^{(2t-1)} &= s_{ij}^{(2t-2)} u_i / \textstyle\sum_{j=1}^{m} s_{ij}^{(2t-2)}, \\ s_{ij}^{(2t)} &= s_{ij}^{(2t-1)} v_j / \textstyle\sum_{i=1}^{n} s_{ij}^{(2t-1)}. \end{aligned} \tag{17}$$

Properties

- Converges to a limit matrix $\hat{\boldsymbol{S}} = \lim_{t\to\infty} \boldsymbol{S}^{(2t)}$.
- Simultaneously adheres to target marginals $\boldsymbol{v}, \boldsymbol{u}$.
- Reduces to the **Sinkhorn operator(normalization)** if target marginals equal one.

Outline

Relaxing Partial Matchings

Key insight — infer the **target marginals** from the highest-scoring partial matchings, i.e., marginals corresponding to predicted correspondences are one, and zero otherwise.

Positive marginals — Approximate *zero* target marginals by a small positive $\epsilon \to 0^+$.

- For each row index $i \in \{1, ..n\}$, we set

$$u_i = \begin{cases} 1, \text{ if } \sum_{j=1}^{m} y^* \left(\mu_w(x, y)\right)_{ij} = 1 \\ \epsilon \to 0^+, \text{otherwise} \end{cases} \qquad \text{(Eq. 7)}$$

- Similarly for column marginals $v_j \in \{1, ..m\}$.

Theorem 2
Define the **entropy-regularized partial matching prediction** of a positive matrix μ_w as:

$$y^* (\mu_w(x, y)/\tau) = \arg\max_{S \in \mathcal{T}_w} \langle \mu_w(x, y), S \rangle + \tau \mathcal{H}(S), \tag{18}$$

for a regularization parameter $\tau \geq 0$, and $\mathcal{T}_w$ the transportation polytope in $\mathbb{R}^{n \times m}$, with **positive** rows and columns marginals, $u \in \{\epsilon, 1\}^n$ and $v \in \{\epsilon, 1\}^m$.

Then,

- $y^* (\mu_w(x, y)/\tau)$ exists, and is **unique**.
- For **small enough** τ and ϵ , it holds that

$$y^* (\mu_w(x, y)) \approx y^* (\mu_w(x, y)/\tau) \tag{19}$$

Sampling from Discrete Distribution

A Gibbs distribution on **any discrete set** of admissible structures, $\mathcal{Y}$, may be formulated based on a parameterized **scoring function** of the instance-label pair $\mu_w(x, y)$ as:

$$\mathbb{P}(y \mid (\mu_w(x, y)) \propto \exp(\mu_w(x, y)) \tag{20}$$

Unfortunately, computing the probability of a **given structure** y requires computing **an intractable partition function**.

In statistical mechanics and mathematics, a Boltzmann distribution (also called Gibbs distribution) is a probability distribution or probability measure that gives the probability that a system will be in a certain state as a function of that state's energy and the temperature of the system. Gibbs/Boltzmann's distribution is an exponential distribution.

Gumbel-Max trick

Gumbel-Max trick One obtains the following identity:

$$\mathbb{P}_{\gamma\sim\mathcal{G}}\left(\arg\max_{y\in\mathcal{Y}}\{\mu_w(x,y)+\gamma(y)\}=y\right)\propto\exp\left(\mu_w(x,y)\right) \tag{21}$$

when random perturbations $\gamma(y)$ follow the **zero mean Gumbel distribution** law, denoted by $\mathcal{G}$.

[Ref] Matej Balog, Nilesh Tripuraneni, Zoubin Ghahramani, and Adrian Weller. Lost relatives of the gumbel trick. arXiv preprint arXiv:1706.04161, 2017.
The cumulative distribution function of the Gumbel distribution with zero mean is $F(x;\beta)=e^{-e^{-x/\beta}}$.

Gumbel-Max trick

Gumbel-Max trick One obtains the following identity:

$$\mathbb{P}_{\gamma\sim\mathcal{G}}\left(\arg\max_{y\in\mathcal{Y}}\{\mu_w(x,y)+\gamma(y)\}=y\right)\propto\exp\left(\mu_w(x,y)\right) \tag{22}$$

- Allow drawing samples from a discrete distribution by solving a structured maximization problem of **a randomly perturbed scoring function**.
- To allow **end-to-end learning**, sampling in latent discrete probabilistic models is often performed by **continuously relaxing** the discrete structure.

Reparameterizing Partial Matching Distributions

Since the label set of $n \times m$ partial permutations with k ones is $\binom{n}{k}\binom{m}{k}k!$, we resort to low-dimensional perturbations

$$\gamma(y) = \sum_{i=1}^{n}\sum_{j=1}^{m} \gamma_{ij}\left(y_{ij}\right), \tag{23}$$

where

- $\gamma_{ij}\left(y_{ij}\right)$ is **independent** random perturbation for each index ij,
- y_{ij} follows the **zero mean Gumbel distribution** law.

With that, the number of random perturbations needed is **linear** in the matching dimension.

Reparameterizing Partial Matching Distributions

The corresponding **entropy-regularized randomly perturbed prediction** problem is

$$y^*\left(\mu'(x,y)/\tau\right) = \arg\max_{S\in\mathcal{T}_w} \left\langle \mu'_w(x,y), S\right\rangle + \tau\mathcal{H}(S), \tag{24}$$

with

$$\mu'(x,y) = \sum_{i=1}^{n}\sum_{j=1}^{m}\left(\mu_w(x,y)_{ij} + \gamma_{ij}\left(y_{ij}\right)\right). \tag{25}$$

Its solution $\boldsymbol{S}$ is of the form $\boldsymbol{A}\exp\left(\frac{1}{\tau}(\mu+\gamma(y))\right)\boldsymbol{B}$ for certain diagonal matrices $\boldsymbol{A},\boldsymbol{B}$ with positive diagonals.

IPFP

Gumbel-IPF Networks

Sampling from a partial matching distribution:

Algorithm 1 Gumbel-IPF

Input: unnormalized scoring function $\mu_w(x, y) \in \mathbb{R}^{n \times m}$, $\gamma(y) \in \mathbb{R}^{n \times m}$, temperature $\tau \geq 0$, row and column target marginals v, u respectively (Eq. 7).

Initialize:

$S = \exp^{(\mu(x,y)+\gamma(y))/\tau}$

$s_{ij}^{(0)} = s_{ij}$

for $t = 1$ **to** T **do**

$$s_{ij}^{(2t-1)} = \frac{s_{ij}^{(2t-2)} u_i}{\sum_{j=1}^{m} s_{ij}^{(2t-2)}} \;,\; s_{ij}^{(2t)} = \frac{s_{ij}^{(2t-1)} v_j}{\sum_{i=1}^{n} s_{ij}^{(2t-1)}}$$

end for

Return: $S^{(2T)}$

Outline

1. Background
2. Partial Matching
3. Gumbel-IPF Networks
4. **Experiments**

Experiments

Two parts:

- Semantic Keypoint Partial Matching
- Properties Of Continuous Partial Matching Relaxation Techniques

Datasets

Dataset	**Mean**	**std**
Pascal VOC	1.37	0.46
MC-PT-SparseGM	1.66	0.97
CUB2011	1.14	0.19

Figure 1: Statistics of imbalance between the number of keypoints measured on samples of in-class image pairs.

Semantic Keypoint Partial Matching

Method	**Accuracy**	F_1 **score**
CIE-H	48.8%	45.9%
qc-DGM	-	52.6%
BB-GM	59.5%	57.3%
NGM-v2	57.5%	53.7%
Ours: Gumbel-IPF	**62.9%**	**58.8%**

Figure 2: Partial matching average accuracy and $F1$ score on the Pascal VOC dataset.

Semantic Keypoint Partial Matching

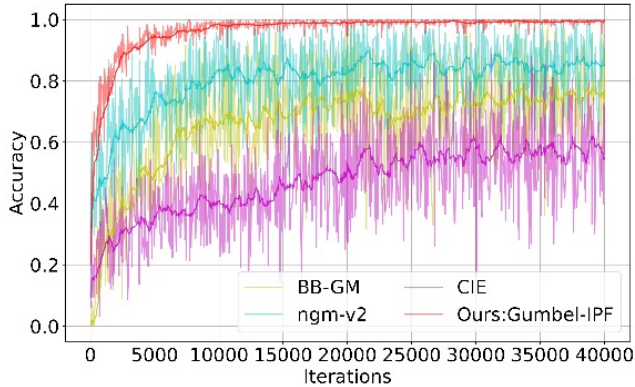


Figure 3: A comparison of average training set accuracy over learning iterations of the partial matching experiment on Pascal VOC.

Gumbel-IPF is the most stable and reaches the highest average training set accuracy

Semantic Keypoint Partial Matching

Method	**Accuracy**	F_1 **score**
IPCA-GM	44.9%	42.7%
Ours: Gumbel-IPF	**46.6%**	**44.6%**

Figure 4: Partial matching average accuracy and average $F1$ score on the IMC-PT-SparseGM dataset.

Semantic Keypoint Partial Matching

Method	**Accuracy**	F_1 **score**
GANN-MGM	-	82.6%
PCA-GM	84.8%	79.7%
IPCA-GM	88.5%	83.2%
Ours: Gumbel-IPF	**89.3%**	**84.1%**

Figure 5: Partial matching average accuracy and $F1$ score on the CUB2011 dataset.

Properties Of Continuous Partial Matching Relaxation Techniques

1. **Mean Absolute Error(MAE)**

$$\frac{1}{2n}\sum_{i=1}^{n}\left|\sum_{j=1}^{m}\tilde{s}_{ij}-u_i\right|+\frac{1}{2m}\sum_{j=1}^{m}\left|\sum_{i=1}^{n}\tilde{s}_{ij}-v_j\right| \tag{26}$$

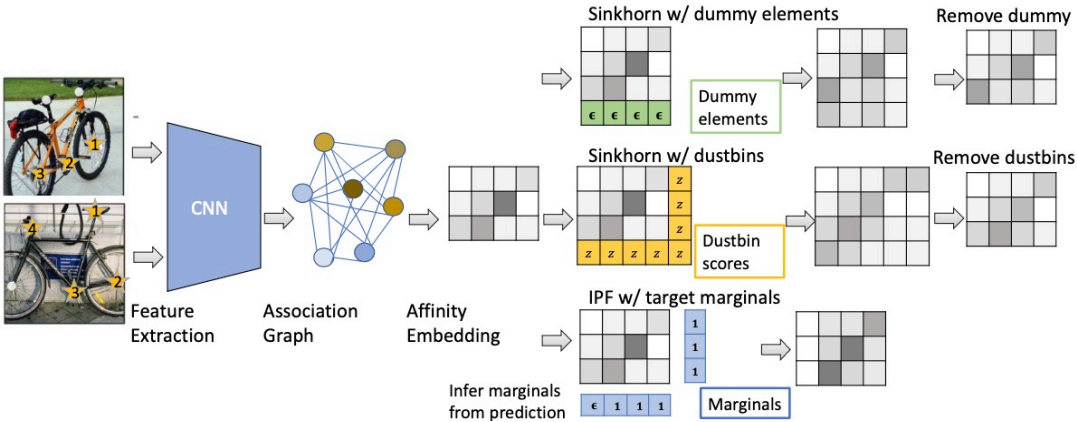
It measures the mean absolute difference between marginals of a normalized matrix's and target marginals.

2. **Empirical Prediction Shift**(EPS)

$$\frac{1}{2\min(m,n)}\sum_{i=1}^{n}\sum_{j=1}^{m}\left|y^*(\mu(x,y))_{i,j}-y^*\left(\tilde{s}_w(x,y)\right)_{i,j}\right|. \tag{27}$$

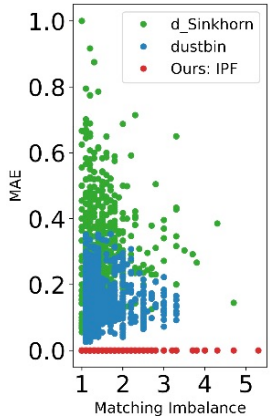
It measures the degree by which the highest unnormalized and normalized scoring structures differ.

Properties Of Continuous Partial Matching Relaxation Techniques

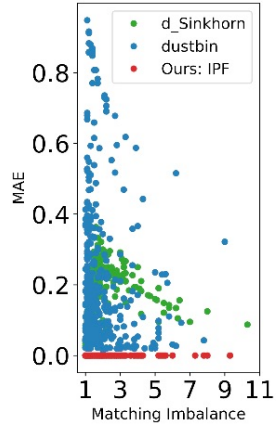


- d_Sinkhorn — Sinkhorn relaxation with dummy elements
- dustbin — Sinkhorn relaxation with dustbins accounting for missing correspondences

MAE



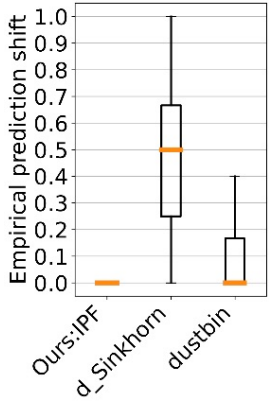
(a) The NGM-v2 backbone on the Pascal VOC dataset.



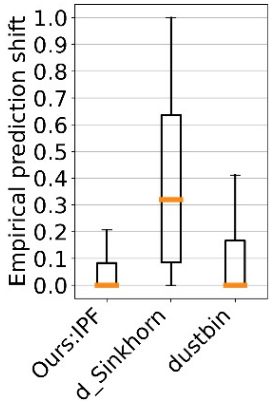
(b) The IPCA-GM backbone on the IMC-PT-SparseGM dataset.

While peer methods suffer high MAE for all matching imbalances, IPF reaches nearly constant zero MAE, which validates that other relaxation methods tend to assign a non-negligible probability mass to missing correspondences.

Empirical Prediction Shift



(c) The NGM-v2 backbone on the Pascal VOC dataset.



(d) The IPCA-GM backbone on the IMC-PT-SparseGM dataset.

IPF is empirically the most order-preserving partial relaxation technique.

Time Complexity

Method	**Training average samples/s ↑**
d_Sinkhorn	3.77
dustbin	3.27
Ours: IPF	3.42

Figure 6: Time complexity analysis of the partial matching relaxation methods.

Conclusion

Address the challenges of learning partial matching structures by

- Biproportional fitting,
- Structured distribution parameterization.

Allow sampling from a partial matching distribution in an end-to-end manner.

Thanks!