Learning Latent Partial Matchings with Gumbel-IPF Networks

Hedda Cohen Indelman
Technion

Tamir Hazan
Technion

Abstract

Learning to match discrete objects has been a central task in machine learning, often facilitated by a continuous relaxation of the matching structure. However, practical problems entail partial matchings due to missing correspondences, which pose difficulties to the one-to-one matching learning techniques that dominate the state-of-the-art. This paper introduces Gumbel-IPF networks for learning latent partial matchings. At the core of our method is the differentiable Iterative Proportional Fitting (IPF) procedure that biproportionally projects onto the transportation polytope of target marginals. Our theoretical framework also allows drawing samples from the temperature-dependent partial matching distribution. We investigate the properties of common-practice relaxations through the lens of biproportional fitting and introduce a new metric, the empirical prediction shift. Our method's advantages are demonstrated in experimental results on the semantic keypoints partial matching task on the Pascal VOC, IMC-PT-SparseGM, and CUB2001 datasets. The code is available at this url.

1 INTRODUCTION

Learning to match discrete objects has been a central task in machine learning and its applications. For example, matching molecular structures in biology (Kainmüller et al., 2014), and matching keypoints in computer vision (Zanfir and Sminchisescu, 2018). However, practical problems often entail partial matchings, which encode missing correspondences. Unfortunately, such problems pose difficulties to the current one-to-one matching learning techniques that dominate the state-of-the-art.

Deep matching techniques learn a parameterized scoring function fitted to minimize the loss of the instance-label pairs between the true label and the predicted matching. Since the highest scoring matching is a piece-wise constant function of the scoring function parameters, end-to-end learning is often facilitated by continuously relaxing the discrete structure. As such, a classification is continuously relaxed by applying the softmax operator, and the Sinkhorn operator may continuously relax a full matching. While the former proportionally projects onto the standard simplex, the latter biproportionally projects onto the Birkhoff polytope. Learning partial matchings requires a continuous operator that biproportionally projects onto the transportation polytope, the set of all non-negative rectangular matrices with prescribed row and column marginals. Previous research facilitated such relaxation by applying the Sinkhorn operator on a heuristically augmented rectangular scoring matrix. We investigate the properties of these common-practice relaxations through the lens of biproportional fitting, as well as suggest a new metric, the empirical prediction shift.

We introduce Gumbel-IPF networks for learning latent partial matchings, which generalize the Sinkhorn-networks (Mena et al., 2018; Cruz et al., 2019) for learning full matchings, and the method of Adams and Zemel (2011) for learning rankings. At the core of our method is the differentiable Iterative Proportional Fitting procedure (IPFP) that biproportionally projects onto the transportation polytope. We extend the temperature-dependent entropy relaxation to partial matchings and prove that the highest-scoring partial matching can be approximated as the limit of an entropy-regularized prediction, which can be obtained with IPFP. The same theoretical framework allows drawing samples from the partial matching distribution by re-reparameterizing the partial matching distribution based on the Gumbel-Max trick. Inspired by the Gumbel-Softmax (Jang et al., 2017; Maddison et al., 2017) and Gumbel-Sinkhorn (Mena et al., 2018), we coin this distribution Gumbel-IPF.

Proceedings of the 27th International Conference on Artificial Intelligence and Statistics (AISTATS) 2024, Valencia, Spain. PMLR: Volume 238. Copyright 2024 by the author(s).

In summary, our contributions are the following:

1. We prove that the highest-scoring partial matching can be approximated as the limit of an entropy-regularized prediction, which can be obtained with the IPF procedure.
2. We prove that samples from the partial matching distribution at a certain temperature can be approximated as the limit of entropy-regularized randomly perturbed prediction, which can be computed with the IPFP as well.
3. We investigate the properties of common-practice partial matching relaxations through the lens of biproportional fitting, and introduce a new metric, the empirical prediction shift.

2 RELATED WORK

Learning scoring-based models of structured distributions is challenging, as it requires computing an often intractable partition function. Instead, sampling from discrete distributions can be performed by the re-parameterization of the Gumbel-Max trick (Luce, 1959; Yellott, 1977; Papandreou and Yuille, 2011; Hazan et al., 2016) by solving a structured maximization of a randomly perturbed scoring function. However, since the arg max is a piece-wise constant function, sampling is often performed under a continuous relaxation of the corresponding structure.

The Gumbel-Softmax distribution (Jang et al., 2017; Maddison et al., 2017) allows drawing samples from a categorical distribution by computing a temperature-dependent softmax function. The softmax is a proportional continuous relaxation of the categorical arg max, as it preserves the rank order of its input score values. Sampling from (full-) matching distributions (Mena et al., 2018) is facilitated with the Sinkhorn operator (Sinkhorn, 1964), which biproportionally projects a positive square matrix to the Birkhoff polytope, the set of doubly-stochastic matrices. As the Gumbel-Max trick is unfeasible for the label set of permutations due to its factorial size, samples are drawn from the matching distribution arising from low-dimensional perturbations (Balog et al., 2017; Hazan and Jaakkola, 2012).

Continuously relaxing a partial matching is often performed on a dummy-padded rectangular matrix. Common when matching graphs with outliers, the Sinkhorn normalization (Sinkhorn, 1964) is performed on a square matrix produced by adding dummy rows or columns initialized to a small positive value (Cho et al., 2010; Wang et al., 2020a; Yew and Lee, 2020; Wang et al., 2020b; Yu et al., 2020; Wang et al., 2021). Alternatively, Sinkhorn normalization can be performed on a matrix padded with a 'dustbin' row and column, intended to absorb the probability of unmatched elements (Sarlin et al., 2020; Liu et al., 2021). Post-normalization, added elements are discarded. Yet, these heuristics do not guarantee a biproportional projection onto the transportation polytope. This difficulty is alternatively addressed by generating a balanced matching task with inliers only, which is a degenerate and often impractical setting.

Optimal Transport. Sinkhorn normalization (Sinkhorn, 1964; Sinkhorn and Knopp, 1967) often arises in optimal transport problems. The seminal work of Cuturi (2013) shows that one can efficiently solve an entropy-regularized problem of optimal transportation with the Sinkhorn normalization. In matching problems, an entropy-regularized optimal transport minimizes a distance map between elements of matched sets. Differently from our method, these methods often assume doubly-stochastic transport plans (Pai et al., 2021; Solomon et al., 2016, 2015), or add dummy elements to absorb unused probability mass in partial transport (Swanson et al., 2020).

3 BACKGOUND

3.1 Partial Matching

Let V^s and V^t be two sets of elements and denote by n the number of elements in V^s, i.e., $n = |V^s|$, and similarly $m = |V^t|$. The pair V^s and V^t form a data instance x^{st}. In a realistic setting, some elements from V^s and V^t may not be matched. This setting often arises in pairwise natural keypoint matching due to occlusions, deformations, and different points of view. In this case, the corresponding label is a partial permutation matrix $y(x^{st}) \in \{0,1\}^{n\times m}$ representing the partial matching between elements of V^s and elements in V^t. As such, $y(x^{st})_{ij} = 1$ if element $i \in V^s$ is matched to element $j \in V^t$, and $y(x^{st})_{ij} = 0$ otherwise. The set of partial matchings is defined as:

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

when $\mathbf{1}_m$ is a size m vector of 1s. Whenever the sets V^s, V^t consist of the same elements, $y(x^{st})$ is a full matching, and the constraints in Equation (1) hold with equality.

In order to learn to predict such structured labels, a parameterized correspondence scoring function $\mu_w(x,y)$ is typically fitted to minimize the loss $\ell(\cdot,\cdot)$ of the training instance-label pairs $(x,y) \in \mathcal{S}$ between the label y and the highest scoring structure

$$y^*(\mu_w(x,y)) = \arg\max_{\hat{y}\in\mathcal{M}} \langle \mu_w(x,y), \hat{y}\rangle. \tag{2}$$

The binary cross-entropy loss is often used in a supervised learning setting to measure the goodness of fit of a training-pair (x, y) and the predicted matching $y^*(\mu_w(x, y))$.

3.2 Reparameterizing Discrete Distributions

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 $\mu_w(x, y)$: $\mathbb{P}(y|(\mu_w(x, y)) \propto \exp(\mu_w(x, y))$. Unfortunately, computing the probability of a given structure y requires computing an intractable partition function. Instead, sampling from the discrete distribution can be performed by the re-parameterization of the Gumbel-Max trick (Luce, 1959; Yellott, 1977; Papandreou and Yuille, 2011; Hazan et al., 2016). When random perturbations $\gamma(y)$ follow the zero mean Gumbel distribution law, denoted by $\mathcal{G}$, one obtains the following identity:

$$\mathbb{P}_{\gamma \sim \mathcal{G}} \left(\arg\max_{y \in \mathcal{Y}} \{\mu_w(x, y) + \gamma(y)\} = y \right) \propto \exp(\mu_w(x, y)) \tag{3}$$

The Gumbel-Max trick (Eq. 3) allows 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.

3.3 Iterative Proportional Fitting

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 (Deming and Stephan, 1940; Bacharach, 1970; Rote and Zachariasen, 2007; Fagan and Greenberg, 1987). For input matrix $Z \in \mathbb{R}^{n \times m}$ with positive entries, matrix $T \in \mathbb{R}^{n \times m}$ for which only the marginals are known (rows and columns sums), we seek to find matrix $S \in \mathbb{R}^{n \times m}$ which is closest to Z w.r.t. the Kullback-Leibler distance and has the same marginals as T's. Thus, the objective function is

$$\begin{aligned} &\min_S \textstyle\sum_{i=1}^n \sum_{j=1}^m s_{ij} \log(\frac{s_{ij}}{z_{ij}}) \\ &\text{s.t.} \;\; \textstyle\sum_{j=1}^m s_{ij} = u_i \; \forall i, \quad \sum_{i=1}^n s_{ij} = v_j \; \forall j, \end{aligned} \tag{4}$$

where $u \in \mathbb{R}_{>0}^{n \times 1}$ denotes marginal rows, and $v \in \mathbb{R}_{>0}^{m \times 1}$ the marginal columns. Biproportion fitting can be understood as preserving cross-product ratios (Mosteller, 1968). Solving the associated Lagrangian shows that the fitted matrix S is the unique solution of the form $S = PZQ$, where $P \in \mathbb{R}^{n \times n}$ and $Q \in \mathbb{R}^{m \times m}$ are diagonal matrices, and their main diagonal elements correspond to the Lagrange multipliers of the marginal constraints. Sinkhorn and Knopp (1967) prove similar results whenever $u \in \mathbb{R}_{>0}^{m \times 1}$, $v \in \mathbb{R}_{>0}^{n \times 1}$ and $\sum_i u_i = \sum_j v_j$.

The IPF procedure entails initializing $s_{ij}^{(0)} = z_{ij}$ and normalizing on rows and columns for $t > 0$ iterations $s_{ij}^{(2t-1)} = s_{ij}^{(2t-2)} u_i / \sum_{j=1}^m s_{ij}^{(2t-2)}, s_{ij}^{(2t)} = s_{ij}^{(2t-1)} v_j / \sum_{i=1}^n s_{ij}^{(2t-1)}$. When entries in the input matrix Z are positive, the procedure converges, under certain conditions, to a limit matrix $\hat{S} = lim_{t \to \infty} S^{(2t)}$ that simultaneously adheres to target marginals v, u. Further, $\hat{S}$ is the unique solution that is closest to the input matrix Z with respect to relative-entropy error (Ruschendorf, 1995). Whenever target marginals equal one, the procedure reduces to the Sinkhorn normalization (Sinkhorn, 1964).

The IPFP is often evaluated by 'internal' validation methods, which measure the differences between aggregate-level outputs and target marginals. Such metrics (e.g. mean absolute error, root mean squared error, and entropy-based measures) are also used to measure the procedure's iteration-wise progress (Pukelsheim et al., 2009; Legates and McCabe Jr., 1999).

4 GUMBEL-IPF NETWORKS

We introduce Gumbel-IPF networks for learning latent partial matchings, which are based on a continuous relaxation of partial matching performed by the IPF procedure. Further, we introduce a task-oriented evaluation metric, the empirical prediction shift.

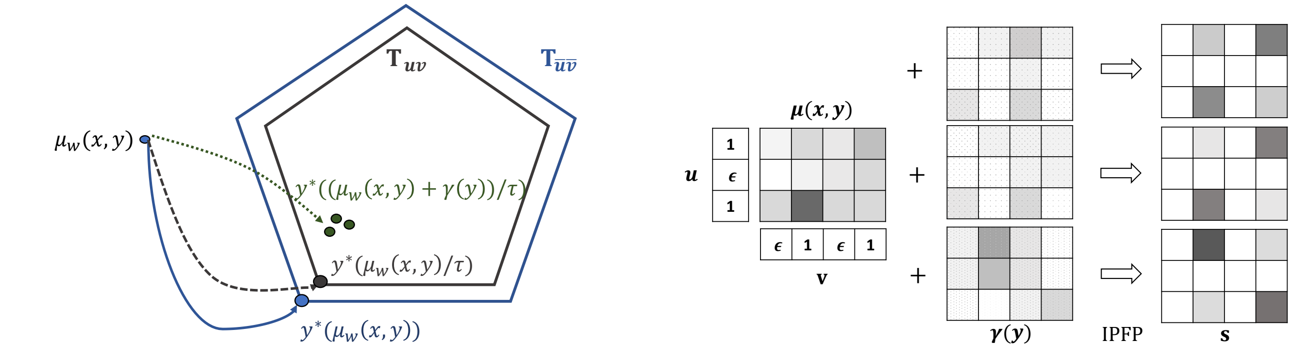
Partial Matching Predictions. Consider an unnormalized parametrized scoring function $\mu_w(x, y) \in \mathbb{R}^{n \times m}$ fitted to minimize the loss $\ell(y, y^*)$ between the label y and the highest scoring structure y^*. The partial matching label space is the transportation polytope in $\mathbb{R}^{n \times m}$, denoted $\mathcal{T}_{\bar{u}\bar{v}}$, with prescribed rows and columns marginals, $\bar{u} \in \{0, 1\}^n$ 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,

$$\mathcal{T}_{\bar{u}\bar{v}} = \{\sum_{r=1}^R \lambda_r T_r ; \sum_{r=1}^R \lambda_r = 1, \lambda_r \geq 0 \, \forall r\}. \tag{5}$$

As such, the highest-scoring partial matching prediction may be defined as

$$y^*(\mu_w(x, y)) \quad = \quad \arg\max_{S \in \mathcal{T}_{\bar{u}\bar{v}}} \langle \mu_w(x, y), S \rangle, \tag{6}$$

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


$\mathbf{T}_{uv}$
$\mathbf{T}_{\bar{u}\bar{v}}$
$\mu_w(x,y)$
$y^*((\mu_w(x,y)+\gamma(y))/\tau)$
$y^*(\mu_w(x,y)/\tau)$
$y^*(\mu_w(x,y))$
$\boldsymbol{\mu}(\boldsymbol{x},\boldsymbol{y})$
$\boldsymbol{u}$
1
ϵ
1
ϵ 1 ϵ 1
$\mathbf{v}$
+
+
+
$\boldsymbol{\gamma}(\boldsymbol{y})$
IPFP
$\mathbf{s}$

(a) The highest-scoring partial matching, $y^*(\mu_w(x,y))$ (Eq. 6) is a vertex of the transportation polytope $\mathcal{T}_{\bar{u}\bar{v}}$ (Eq. 5). At a small enough temperature τ, the relaxed partial matching $y^*(\mu_w(x,y)/\tau)$ (Eq. 8) is a vertex of the transportation polytope $\mathcal{T}_{uv}$, with rows and columns marginals, $u \in \{\epsilon, 1\}^n$ and $v \in \{\epsilon, 1\}^m$, which can be obtained with the IPFP. For $\epsilon \to 0^+$, the polytopes nearly coincide and $y^*(\mu_w(x,y)/\tau)$ approximates $y^*(\mu_w(x,y))$.

(b) Samples $y^*((\mu_w(x,y)+\gamma(y))/\tau)$ from the partial matching distribution corresponding to low-dimensional perturbations at temperature τ are solutions of the randomly perturbed entropy-regularized prediction problem (Eq. 10), which can be drawn with the IPFP. IPFP's outputs are biproportionally fit w.r.t. the input and adhere to the prediction's marginal constraints (Eq. 7).

Figure 1: Illustration of our Gumbel-IPF method for sampling from the partial matching distribution.

The problem of Equation (6) can be solved with linear assignment solvers (Jonker and Volgenant, 1987; Munkres, 1957).

The matching nature of the task dictates at most one 1 in each row and each column, thus the polytope $\mathcal{T}_{\bar{u}\bar{v}}$ can be interpreted as a k-assignment polytope, with $k = \sum \bar{u} = \sum \bar{v}$.

Relaxing Partial Matchings. An analog of the Sinkhorn normalization for square permutations is the IPF procedure for partial permutations with marginal constraints (Sinkhorn and Knopp, 1967; Ruschendorf, 1995). In the following, we extend the temperature-dependent entropy relaxation to partial matching relaxation and prove that the highest scoring partial matching (Eq. 6) can be approximated as the limit of an entropy regularized prediction, which can be obtained with the IPF procedure.

Our key insight is that target marginals can be inferred from the highest-scoring partial matchings (Eq. 6), i.e., marginals corresponding to predicted correspondences are one, and zero otherwise. As the IPF procedure and its convergence properties are proved for positive target marginals, we approximate zero target marginals by a small positive $\epsilon \to 0^+$. For each row index $i, i = 1, ..n$, we set

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

and similarly for column marginals v_j, $j = 1, ..m$.

Theorem 4.1. *Denote the entropy of a matrix S as $\mathcal{H}(S) = -\sum_i \sum_j s_{ij} \log s_{ij}$. Denote by $\mathcal{T}_{uv}$ the transportation polytope in $\mathbb{R}^{n\times m}$, with positive rows and columns marginals, $u \in \{\epsilon, 1\}^n$ and $v \in \{\epsilon, 1\}^m$ (Eq. 7). For a matrix $S \in \mathbb{R}^{n\times m}$ in the transportation polytope $\mathcal{T}_{uv}$, define the entropy-regularized partial matching prediction of a positive matrix μ_w as:*

$$y^*(\mu_w(x,y)/\tau) = \arg\max_{S\in\mathcal{T}_{uv}} \langle \mu_w(x,y), S\rangle + \tau\mathcal{H}(S), \tag{8}$$

for a regularization parameter $\tau \geq 0$. Then, $y^*(\mu_w(x,y)/\tau)$ exists, and is unique. Further, for small enough τ and ϵ, it holds that $y^*(\mu_w(x,y)) \approx y^*(\mu_w(x,y)/\tau)$.

Proof. The Lagrangian of the entropy-regularized partial matching prediction problem (Eq. 8) is

$$\mathcal{L}(\mu, s, \tau) = \arg\max_{S\in\mathcal{T}_{uv}} \sum_{i=1}^{n}\sum_{j=1}^{m} s_{ij}(\mu_{ij} - \tau \log s_{ij}) \tag{9}$$
$$- \sum_{i=1}^{n} \alpha_i(u_i - \sum_{j=1}^{m} s_{ij}) - \sum_{j=1}^{m} \beta_j(v_j - \sum_{i=1}^{n} s_{ij}).$$

Its solution for each s_{ij} is given by $s_{ij} = \exp(\frac{1}{\tau}(\alpha_i - \tau))\exp(\frac{1}{\tau}\mu_{ij})\exp(\frac{1}{\tau}\beta_j)$ (proof in Appendix 1). As such, the partial matching solution S is of the form $\mathrm{A}\exp(\frac{1}{\tau}\mu)\mathrm{B}$ for certain diagonal matrices A, B with positive diagonals. Since the matrix $\exp(\frac{1}{\tau}\mu)$ is strictly positive, the solution $S \in \mathcal{T}_{uv}$ exists, is unique, and can be found with the IPF procedure (Ruschendorf, 1995). When $\tau \to 0$, the partial matching solution S is a vertex of the transportation polytope, $\mathcal{T}_{uv}$. Additionally, for $\epsilon \to 0^+$, $S \in \mathcal{T}_{uv}$ approximates the solution of $y^*(\mu_w(x,y))$. □

As the regularization parameter τ increases, the solution S is a point in the interior of $\mathcal{T}_{uv}$. Importantly,

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:** $S^{(2T)}$
```

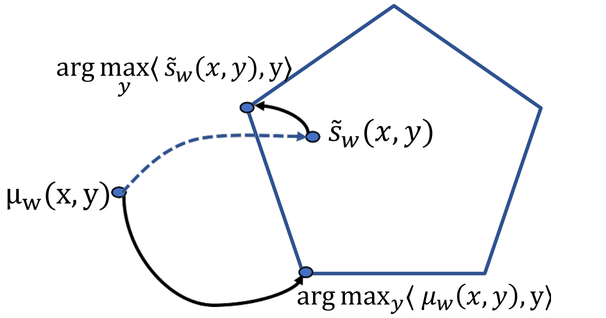


Figure 2: Illustration of the prediction shift phenomenon. The highest scoring structure of an unnormalized scoring function $\mu_w(x,y)$ differs from the highest scoring structure of the corresponding normalized scoring function $\tilde{s}_w(x,y)$.

for any τ, the IPFP guarantees that the relaxed partial matching S adheres to marginal constraints. The solution S may not be precisely attained as the IPFP is run for a fixed number of iterations, or until an iteration-wise evaluation metric is sufficiently small.

Reparameterizing Partial Matching Distributions. Unfortunately, it is infeasible to draw samples from the partial matching distribution with the Gumbel-Max trick (Eq. 3) as the label set of $n \times m$ partial permutations with k ones is $\binom{n}{k}\binom{m}{k}k!$ (Gill and Linusson, 2009). Since the label set of $n \times m$ permutations factorizes, we resort to low-dimensional perturbations $\gamma(y) = \sum_{i=1}^{n}\sum_{j=1}^{m}\gamma_{ij}(y_{ij})$, where $\gamma_{ij}(y_{ij})$ is independent random perturbation for each index ij, and each y_{ij} that follows the zero mean Gumbel distribution law. With that, the number of random perturbations needed is linear in the matching dimension instead of factorial. A random partial matching follows the partial matching distribution induced by low-dimensional perturbations of the scoring function (Hazan and Jaakkola, 2012; Balog et al., 2017). The corresponding entropy-regularized randomly perturbed prediction problem is

$$y^*(\mu'(x,y)/\tau) = \arg\max_{S\in\mathcal{T}_{uv}} \langle \mu'_w(x,y), S\rangle + \tau\mathcal{H}(S), \quad (10)$$

with $\mu'(x,y) = \sum_{i=1}^{n}\sum_{j=1}^{m}(\mu_w(x,y)_{ij} + \gamma_{ij}(y_{ij}))$. Its solution S is of the form $\mathrm{A}\exp(\frac{1}{\tau}(\mu+\gamma(y)))\mathrm{B}$ for certain diagonal matrices A, B with positive diagonals (proof in Appendix 7.2). Thus, samples from the bespoke distribution at temperature τ are solutions of the randomly perturbed entropy-regularized problem (Eq. 10), which can be drawn with the IPFP.

With that, we present our method Gumbel-IPF (illustrated in Figure 1) for sampling from a partial matching distribution in Algorithm 1. In practice, we find that it sufficiently converges after a few dozen of iterations. While learning the perturbation variance is possible (Cohen Indelman and Hazan, 2021), fine-tuning it as a hyper-parameter leads to satisfactory results.

Complexity Analysis. Our relaxation method's complexity is $\mathcal{O}(n*m)$ for a $n \times m$ matrix at each step, which is no greater than the complexity of other methods based on the Sinkhorn normalization having $\mathcal{O}(n^2)$ complexity for a square $n \times n$ matrix. This analysis is validated in experimental results (Table 5).

Evaluation Metrics. We introduce a task-oriented evaluation metric: the empirical prediction shift, in addition to revisiting the well-known mean absolute error (MAE).

The empirical prediction shift metric is motivated by observing that ideally, any continuous partial matching relaxation should preserve the ordering of structure scores, i.e., maintain the monotonicity of scores. In such case, it will hold that

$$\arg\max_{S\in\mathcal{T}_{\bar{u}\bar{v}}} \langle \mu_w(x,y), S\rangle = \arg\max_{S\in\mathcal{T}_{\bar{u}\bar{v}}} \langle \tilde{s}_w(x,y), S\rangle, \quad (11)$$

where $\mu_w(x,y)$ denotes the unnormalized scoring function and $\tilde{s}_w(x,y)$ the normalized scoring function obtained by a normalization technique. While this property is not guaranteed in biproportional fitting techniques, it is central for obtaining low-bias gradients of relaxed discrete structures. Thus, we propose the empirical prediction shift metric, which measures the degree by which the highest unnormalized and normalized scoring structures differ (illustrated in Fig. 2):

$$\frac{1}{2\min(m,n)} \sum_{i=1}^{n}\sum_{j=1}^{m} \left| y^*(\mu(x,y))_{i,j} - y^*(\tilde{s}_w(x,y))_{i,j} \right|. \quad (12)$$

The lower the empirical prediction shift of a normalization technique, the more structure scores order-preserving it is.

The MAE measures the mean absolute difference between marginals of a normalized matrix $\tilde{s}$ and target marginals in the units of the variable itself,

$$\frac{1}{2n}\sum_{i=1}^{n}|\sum_{j=1}^{m}\tilde{s}_{ij}-u_i|+\frac{1}{2m}\sum_{j=1}^{m}|\sum_{i=1}^{n}\tilde{s}_{ij}-v_j|. \quad (13)$$

It's an effective metric for comparing the goodness of fit of normalization techniques and analyzing the interplay between matching imbalance and the error measure. In our context, a high MAE of a normalization technique suggests that it did not project onto the transportation polytope of the target marginals. We define the matching imbalance by

$$M_{imb}(n,m)=\max(n,m)/\min(n,m). \quad (14)$$

5 EXPERIMENTS

Missing correspondence approximation (ϵ) setting. Following the conditions of the IPF procedure convergence analysis (Pukelsheim et al., 2009), we approximate zero target marginals by a small positive $\epsilon \to 0^+$. Experiments with various ϵ value settings showed little variation, therefore we heuristically set $\epsilon = 1e-6$ in all experiments.

Entropy regularization parameter (τ) effect and setting. As the entropy regularization increases, the relaxed partial matching produced by our method is a point in the interior of the transportation polytope with inferred marginals. Put differently, for $\tau \to \infty$, our method produces a sample from a uniform distribution over partial matchings. To support this analysis, we follow the semantic keypoint partial matching experiment on the Pascal VOC dataset (detailed in Section 5.1.1) and sample 250 unnormalized keypoint partial matching scoring matrices. Our method is applied to each matrix with varying $\tau \in [0.01, 1, 10, 100]$. As expected, the average relaxed partial matching matrices entropy increases as τ increases (Figure 3).

In experiments, $\tau = 1$ was set to compare relaxation methods. Related work suggests that an entropy regularization annealing scheme during training is beneficial, though we haven't employed it in this work.

5.1 Semantic Keypoint Partial Matching

To demonstrate our method's advantage across architecture backbones and datasets we carry two semantic keypoint partial matching experiments: NGM-v2 (Wang et al., 2021) backbone on the Pascal VOC dataset (Everingham et al., 2010), and IPCA-GM (Wang et al., 2020b) backbone on the IMC-PT-SparseGM dataset (Wang et al., 2023) and the CUB2011 dataset (Wah et al., 2011). These natural image datasets translate to a partial matching problem as images have a varying number of keypoints, due to occlusions, different points of view, etc. Table 1 summarizes the average and std keypoint imbalance measured on sampled pairs of images (per-class analysis in Appendix 7.3).

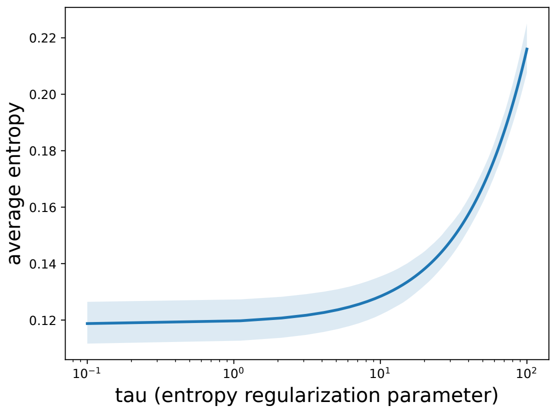


Figure 3: The effect of the entropy regularization parameter on the average Gumbel-IPF normalized matrices entropy, measured on 250 unnormalized keypoint partial matching scoring matrices following the semantic keypoint partial matching experiment detailed in Section 5.1.1.

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

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

Keypoint Filtering. To bring to light the partial matching task we follow the 'unfiltered' setup in all experiments - keypoints are *neither* filtered to reach intersection *nor* to reach inclusion in image pairs.

Peer Methods. The blackbox differentiation method (Pogančić et al., 2020) was recently applied to neural graph matching (Rol'inek et al., 2020), denoted BB-GM, based on graph matching solvers (Swoboda et al., 2019). The blackbox differentiation method's gradients are of a surrogate linearized loss. BB-GM's feature extraction was adopted in the NGM-V2 architecture (Wang et al., 2021), which based on Lawler's Quadratic Assignment Problem casts a constrained neural graph matching into learning a vertex classification (matching) based on association graph embedding. The qc-DGM model (Gao et al., 2021a) also presents a differentiable approach to the quadratic constraints of pairwise structural discrepancy between graphs. The IPCA-GM (Wang et al., 2020a) network is based on the PCA-GM (Wang et al., 2019) network for embedding graph structure features with iterative cross-graph convolution and relaxing the matching structure by Sinkhorn normalization. The CIE-H model (Yu et al., 2020) builds upon the PCA-GM model and adds a channel-independent edge embedding module and a Hungarian attention layer over the loss function, such that the most contributing matching pairs are attended to. These methods, except for BB-GM, either apply Sinkhorn normalization with dummy elements for partial matching or employ keypoint intersection filtering to reduce the task to balanced matching. We build upon peer implementations in the ThinkMatch project.

In the following experiments, the backbone's Sinkhorn normalization with dummy elements on the partial matching prediction head is replaced with our Gumbel-IPF method (Algorithm 1). Further details are in the appendix.

5.1.1 Pascal VOC

We perform a semantic keypoint partial matching experiment on the Pascal VOC dataset (Everingham et al., 2010) with Berkeley annotations (Bourdev and Malik, 2009). Our architecture is based on the NGM-v2 (Wang et al., 2021) backbone. Following prior research (Wang et al., 2021; Rol'inek et al., 2020; Wang et al., 2019), poorly annotated images and keypoints annotated as 'truncated', 'occluded', and 'difficult' are filtered. As mentioned in Wang et al. (2021), Rol'inek et al. (2020) has a slightly favorable setting by filtering keypoints outside of the bounding box.

Our Gumbel-IPF method is compared to the following two-graph unbalanced matching peer methods: CIE-H (Yu et al., 2020), qc-DGM (Gao et al., 2021b), BB-GM (Rol'inek et al., 2020), NGM-v2 (Wang et al., 2021).

Results. The test set's average and per-class average matching prediction accuracy (recall) accuracies are reported, as well as F_1 score. Our method outperforms peer methods considerably on the average test set results (Table 4). A per-class average accuracy comparison reveals that our method outperformed the NGM-v2 baseline in 18 classes (Table 8 in the Appendix). Moreover, as the class average imbalance increases (refer to Table 6 in the Appendix) our method's improvement over the baseline increases (Figure 8 in the Appendix), further demonstrating our method's advantage in relaxing partial matchings.

Our method is also the most stable and reaches the highest average training set accuracy (Figure 4).

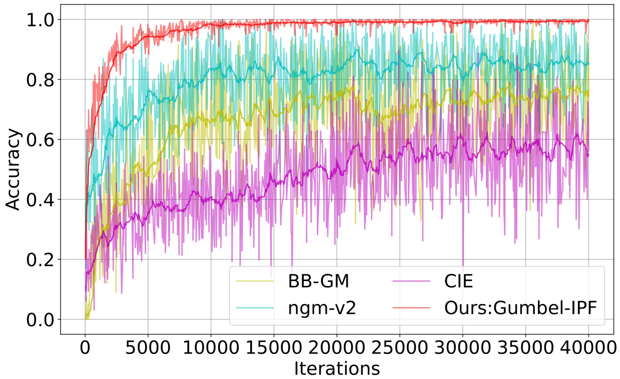


Figure 4: A comparison of average training set accuracy over learning iterations of the partial matching experiment on Pascal VOC.

5.1.2 IMC-PT-SparseGM

We explore the recently introduced IMC-PT-SparseGM dataset (Wang et al., 2023) originated from the stereo benchmark Image Matching Challenge PhotoTourism (Jin et al., 2020) in combination with the IPCA-GM (Wang et al., 2020b) backbone. This dataset is characterized by a high degree of keypoints imbalance (Table 1) and a large number of keypoints.

Results. Our method improves the baseline method's average test set accuracy and F_1 score (Table 2). A per-class comparison is detailed in Table 9 in the Appendix.

Table 2: Partial matching average accuracy and average F_1 score on the IMC-PT-SparseGM dataset. The keypoint filtering preserves outlier keypoints in both images. Best results are in bold.

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

5.1.3 CUB2011

We repeat the experiment with the IPCA-GM (Wang et al., 2020b) architecture backbone and the CUB2011 dataset (Wah et al., 2011).

Results. Our method moderately improves the baseline method's average test set accuracy and F_1 score (Table 3). Indeed, among the datasets tested, CUB2011 has the lowest average matching imbalance.

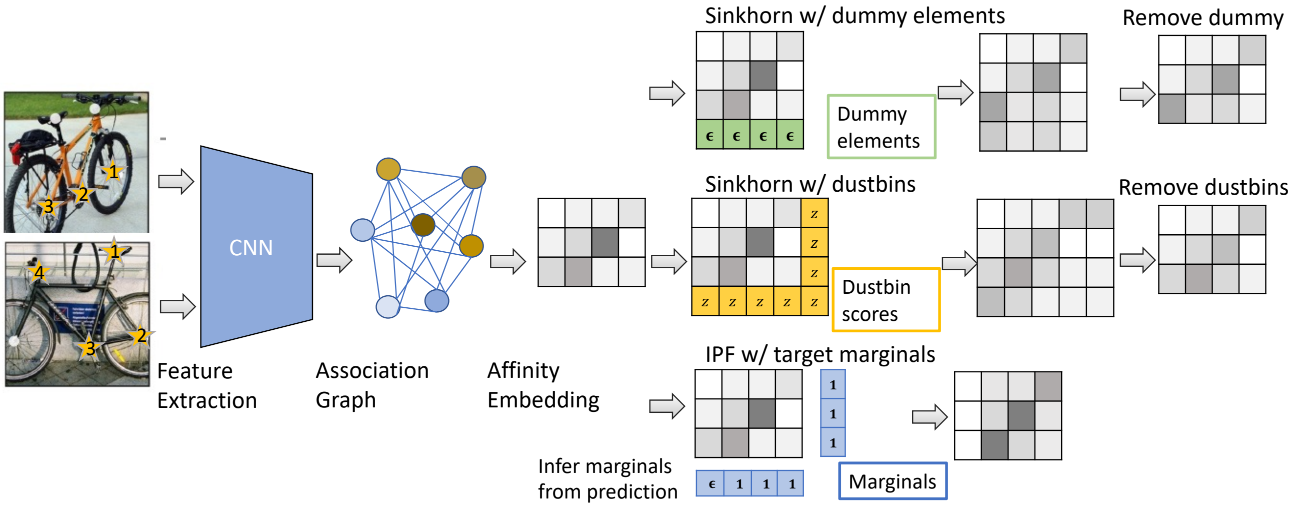


Figure 5: Partial matching relaxation methods illustration. Sinkhorn with dummy elements pads a scoring matrix with dummy rows or columns, initialized to a small ϵ, to form a square matrix. Sinkhorn with dustbins pads a scoring matrix with a row and column. The point-to-bin and bin-to-bin scores are filled with a single learnable parameter z. Post normalization, added elements are removed in both methods. Our IPF method is performed on the rectangular scoring matrix and biproportionally normalizes it while adhering to the prediction's marginals.

Table 3: Partial matching average accuracy and F_1 score on the CUB2011 dataset. The keypoint filtering preserves outlier keypoints in both images. Best results are in bold.

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%**

5.2 Properties Of Continuous Partial Matching Relaxation Techniques

This experiment aims to analyze the properties of continuous partial matching relaxation techniques. Thus, we focus on experiments on the two most imbalanced datasets: Pascal VOC and IMC-PT-SparseGM. Results for the CUB2011 dataset are detailed in Appendix 7.3.3. The scoring function isn't randomly perturbed, to eliminate the effect of stochasticity.

Based on experiments in Sections 5.1.1 and 5.1.2, the keypoints partial matching head of the baseline backbone is adjusted per: i our method (Algorithm 1), denoted 'IPF', ii Sinkhorn relaxation with dummy elements denoted 'd_Sinkhorn', iii Sinkhorn relaxation with 'dustbins' accounting for missing correspondences, denoted 'dustbin' (Sarlin et al., 2020). These methods are illustrated in Figure 5 in a pairwise keypoint correspondence prediction task. Metrics for the 'dustbin' method are collected when the network training has plateaued, to allow learning of the point/bin-to-bin parameter.

Mean Absolute Error. All methods reach nearly zero MAE in the case of full matchings. However, while peer methods suffer high MAE for all matching imbalances, our method reaches nearly constant zero MAE (Eq. 13) (i.e., ϵ error by construction). Furthermore, the lower the matching imbalance, the higher the MAE of peer methods (Figure 6). These results validate that other relaxation methods tend to assign a non-negligible probability mass to missing correspondences, while our method doesn't.

Empirical Prediction Shift. All methods exhibit a negligible empirical prediction shift (Eq. 2) in normalizing full matchings. However, in both partial matching experiments, the method of Sinkhorn normalization with dummy elements suffers a high prediction shift. Compared to our method, the 'dustbin' method exhibits more dispersed statistics, with a significantly higher upper 3^{rd} quantile (Figure 7) and maximum. Thus, our method is empirically the most order-preserving partial relaxation technique.

Our method is also beneficial as a hidden correspondence layer normalization technique (Appendix 7.3.4).

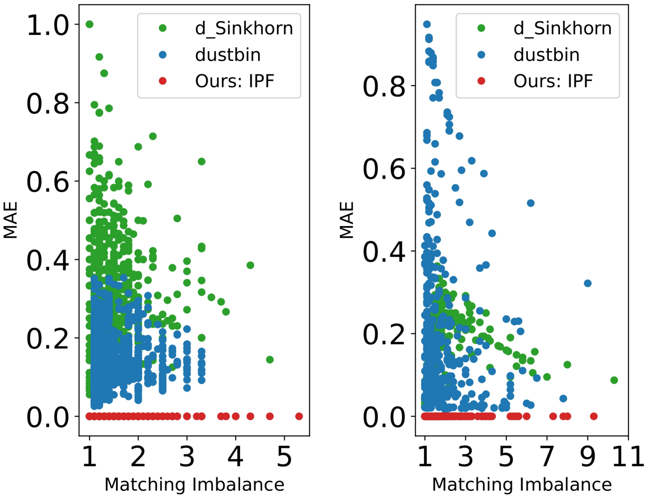
Time Complexity. Measurements of the training average samples per second support our complexity analysis, as the differences between the relaxation methods' computation times are negligible (Table 5). The number of iterations is consistent across all methods.

Table 4: Partial matching average accuracy and F_1 score on the Pascal VOC dataset. The keypoint filtering preserves outlier keypoints in both images. Best results are in bold.

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%**

Table 5: Time complexity analysis of the partial matching relaxation methods.

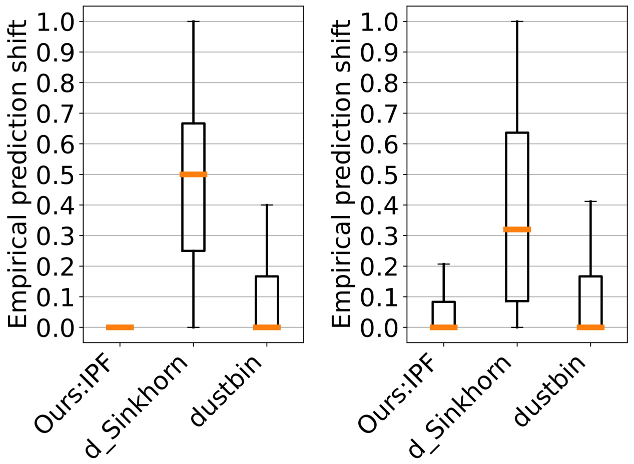
Method	**Training average samples/s** ↑
d_Sinkhorn	3.77
dustbin	3.27
Ours: IPF	3.42



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

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

Figure 6: Mean absolute error from target marginals versus matching imbalance of partial matching relaxation techniques on 1k random samples of image pairs. Cases of full matchings are not displayed.



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

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

Figure 7: Empirical prediction shift (Eq. 12) of partial matching relaxation methods on partial matching tasks. Cases of full matchings are not displayed.

6 CONCLUSIONS

Our focus has been addressing the challenges of learning partial matching structures. By combining methods of biproportional fitting and structured distribution parameterization, our method allows sampling from a partial matching distribution in an end-to-end manner. In summary, we draw the following conclusions from experimental results:

(i) Other methods tend to assign non-negligible probability mass to missing correspondences, which elucidates that they do not project onto the transportation polytope of the prediction's target marginals (Figures 6 and 9a),

(ii) Our method is empirically the most order-preserving, as it exhibits significantly lower values of empirical prediction shift (Figures 7 and 9b).

(iii) Our method improves the baseline method's test set metrics in various backbone architectures and datasets (Section 5.1)

References

Adams, R. P. and Zemel, R. S. (2011). Ranking via sinkhorn propagation. *ArXiv*, abs/1106.1925.

Bacharach, M. (1970). Biproportional matrices & input-output change. In *Cambridge University Press*.

Balog, M., Tripuraneni, N., Ghahramani, Z., and Weller, A. (2017). Lost relatives of the gumbel trick. In *International Conference on Machine Learning*.

Bourdev, L. D. and Malik, J. (2009). Poselets: Body part detectors trained using 3d human pose annotations. *2009 IEEE 12th International Conference on Computer Vision*, pages 1365–1372.

Cho, M., Lee, J., and Lee, K. M. (2010). Reweighted random walks for graph matching. In *CVPR 2011*, volume 6315, pages 492–505.

Cohen Indelman, H. and Hazan, T. (2021). Learning randomly perturbed structured predictors for direct loss minimization. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139. PMLR.

Cruz, R. S., Fernando, B., Cherian, A., and Gould, S. (2019). Visual permutation learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41:3100–3114.

Cuturi, M. (2013). Sinkhorn distances: Lightspeed computation of optimal transport. In Burges, C., Bottou, L., Welling, M., Ghahramani, Z., and Weinberger, K., editors, *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc.

Deming, W. E. and Stephan, F. F. (1940). On a least squares adjustment of a sampled frequency table when the expected marginal totals are known. *The Annals of Mathematical Statistics*, 11(4):427–444.

Everingham, M., Gool, L. V., Williams, C. K. I., Winn, J. M., and Zisserman, A. (2010). The pascal visual object classes (voc) challenge. *International Journal of Computer Vision*, 88:303–338.

Fagan, J. and Greenberg, B. (1987). Making tables additive in the presence of zeros. *American Journal of Mathematical and Management Sciences*, 7.

Gao, Q., Wang, F., Xue, N., Yu, J.-G., and Xia, G.-S. (2021a). Deep graph matching under quadratic constraint. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5069–5078.

Gao, Q., Wang, F., Xue, N., Yu, J.-G., and Xia, G.-S. (2021b). Deep graph matching under quadratic constraint. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5067–5074.

Gill, J. and Linusson, S. (2009). The k-assignment polytope. *Discret. Optim.*, 6:148–161.

Hazan, T. and Jaakkola, T. (2012). On the partition function and random maximum a-posteriori perturbations. In *International Conference on Machine Learning*.

Hazan, T., Papandreou, G., and Tarlow, D. (2016). *Perturbations, Optimization, and Statistics*. The MIT Press.

Jang, E., Gu, S., and Poole, B. (2017). Categorical reparameterization with gumbel-softmax. In *International Conference on Learning Representations*.

Jin, Y., Mishkin, D., Mishchuk, A., Matas, J., Fua, P., Yi, K. M., and Trulls, E. (2020). Image matching across wide baselines: From paper to practice. *International Journal of Computer Vision*, 129(2):517–547.

Jonker, R. and Volgenant, A. (1987). A shortest augmenting path algorithm for dense and sparse linear assignment problems. *Computing*, 38(4):325–340.

Kainmüller, D., Jug, F., Rother, C., and Myers, E. W. (2014). Active graph matching for automatic joint segmentation and annotation of c. elegans. *Medical image computing and computer-assisted intervention : MICCAI ... International Conference on Medical Image Computing and Computer-Assisted Intervention*, 17 Pt 1:81–8.

Legates, D. R. and McCabe Jr., G. J. (1999). Evaluating the use of "goodness-of-fit" measures in hydrologic and hydroclimatic model validation. *Water Resources Research*, 35(1):233–241.

Liu, L., Hughes, M. C., Hassoun, S., and Liu, L. (2021). Stochastic iterative graph matching. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 6815–6825. PMLR.

Luce, R. D. (1959). *Individual Choice Behavior: A Theoretical analysis*. Wiley, New York, NY, USA.

Maddison, C. J., Mnih, A., and Teh, Y. W. (2017). The concrete distribution: A continuous relaxation of discrete random variables. In *International Conference on Learning Representations*.

Mena, G. E., Belanger, D., Linderman, S. W., and Snoek, J. (2018). Learning latent permutations with gumbel-sinkhorn networks. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*.

Mosteller, F. (1968). Association and estimation in contingency tables. *Journal of the American Statistical Association*, 63(321):1–28.

Munkres, J. (1957). Algorithms for the assignment and transportation problems. *Journal of the Society for Industrial and Applied Mathematics*, 5(1):32–38.

Pai, G., Ren, J., Melzi, S., Wonka, P., and Ovsjanikov, M. (2021). Fast Sinkhorn Filters: Using Matrix Scaling for Non-Rigid Shape Correspondence with Functional Maps. In *CVPR*.

Papandreou, G. and Yuille, A. L. (2011). Efficient variational inference in large-scale bayesian compressed sensing. *2011 IEEE International Conference on Computer Vision Workshops (ICCV Workshops)*, pages 1332–1339.

Pogančić, M. V., Paulus, A., Musil, V., Martius, G., and Rolinek, M. (2020). Differentiation of blackbox combinatorial solvers. In *International Conference on Learning Representations*.

Pukelsheim, F., Simeone, B., and Sapienza Università di Roma, Dipartimento di Statistica, P. e. S. A. (2009). *On the Iterative Proportional Fitting Procedure: Structure of Accumulation Points and L1-Error Analysis*. Preprints des Instituts für Mathematik der Universität Augsburg. Universität Augsburg.

Rol'inek, M., Swoboda, P., Zietlow, D., Paulus, A., Musil, V., and Martius, G. (2020). Deep graph matching via blackbox differentiation of combinatorial solvers. In *ECCV*.

Rote, G. and Zachariasen, M. (2007). Matrix scaling by network flow. In *Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms: (SODA 07)*, pages 848–854, United States. Association for Computing Machinery.

Ruschendorf, L. (1995). Convergence of the iterative proportional fitting procedure. *The Annals of Statistics*, 23(4):1160 – 1174.

Sarlin, P.-E., DeTone, D., Malisiewicz, T., and Rabinovich, A. (2020). SuperGlue: Learning feature matching with graph neural networks. In *CVPR*.

Sinkhorn, R. (1964). A relationship between arbitrary positive matrices and doubly stochastic matrices. *Ann. Math. Statist.*, 35(2):876–879.

Sinkhorn, R. and Knopp, P. (1967). Concerning nonnegative matrices and doubly stochastic matrices. *Pacific Journal of Mathematics*, 21:343–348.

Solomon, J., de Goes, F., Peyré, G., Cuturi, M., Butscher, A., Nguyen, A., Du, T., and Guibas, L. (2015). Convolutional wasserstein distances: Efficient optimal transportation on geometric domains. *ACM Trans. Graph.*, 34(4).

Solomon, J., Peyré, G., Kim, V. G., and Sra, S. (2016). Entropic metric alignment for correspondence problems. *ACM Trans. Graph.*, 35(4).

Swanson, K., Yu, L., and Lei, T. (2020). Rationalizing text matching: Learning sparse alignments via optimal transport. *ArXiv*, abs/2005.13111.

Swoboda, P., Kainm"uller, D., Mokarian, A., Theobalt, C., and Bernard, F. (2019). A convex relaxation for multi-graph matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Wah, C., Branson, S., Welinder, P., Perona, P., and Belongie, S. (2011). The Caltech-UCSD Birds-200-2011 Dataset. Technical Report CNS-TR-2011-001, California Institute of Technology.

Wang, R., Guo, Z., Jiang, S., Yang, X., and Yan, J. (2023). Deep learning of partial graph matching via differentiable top-k. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6272–6281.

Wang, R., Yan, J., and Yang, X. (2019). Learning combinatorial embedding networks for deep graph matching. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 3056–3065.

Wang, R., Yan, J., and Yang, X. (2020a). Combinatorial learning of robust deep graph matching: an embedding based approach. *IEEE Transactions on Pattern Analysis & Machine Intelligence*.

Wang, R., Yan, J., and Yang, X. (2020b). Graduated assignment for joint multi-graph matching and clustering with application to unsupervised graph matching network learning. In *Advances in Neural Information Processing Systems*, volume 33.

Wang, R., Yan, J., and Yang, X. (2021). Neural graph matching network: Learning lawlerś quadratic assignment problem with extension to hypergraph and multiple-graph matching. *IEEE Transactions on Pattern Analysis & Machine Intelligence*.

Yellott, J. I. (1977). The relationship between luce's choice axiom, thurstone's theory of comparative judgment, and the double exponential distribution. *Journal of Mathematical Psychology*, 15:109–144.

Yew, Z. J. and Lee, G. H. (2020). RPM-net: Robust point matching using learned features. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE.

Yu, T., Wang, R., Yan, J., and Li, B. (2020). Learning deep graph matching with channel-independent embedding and hungarian attention. In *ICLR*.

Zanfir, A. and Sminchisescu, C. (2018). Deep learning of graph matching. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Checklist

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. [Yes, in Section 4]
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [Yes, applicable for time complexity (Table 5).]
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. [Yes]

2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. [Yes, in Section 4]
   (b) Complete proofs of all theoretical results. [Yes]
   (c) Clear explanations of any assumptions. [Yes]

3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). [Yes in supplementary material and code]
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [Yes in supplementary material]
   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). [Yes]
   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [Yes]

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:
   (a) Citations of the creator If your work uses existing assets. [Yes]
   (b) The license information of the assets, if applicable. [Not Applicable]
   (c) New assets either in the supplemental material or as a URL, if applicable. [Not Applicable]
   (d) Information about consent from data providers/curators. [Not Applicable]
   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. [Not Applicable]

5. If you used crowdsourcing or conducted research with human subjects, check if you include:
   (a) The full text of instructions given to participants and screenshots. [Not Applicable]
   (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. [Not Applicable]
   (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. [Not Applicable]

7 APPENDIX

7.1 Solution Of The Entropy-Regularized Problem

The Lagrangian of the entropy-regularized problem is

$$\mathcal{L}(\mu, s, \tau) = \arg\max_{s \in \mathcal{T}_{uv}} \sum_i \sum_j s_{ij}(\mu_{ij} - \tau \log s_{ij}) - \sum_{i=1}^{n} \alpha_i (u_i - \sum_{j=1}^{m} s_{ij}) - \sum_{j=1}^{m} \beta_j (v_j - \sum_{i=1}^{n} s_{ij}). \tag{15}$$

Its solution for each s_{ij} is given by

$$\frac{\partial \mathcal{L}(\mu, s, \tau)}{\partial s_{ij}} = \mu_{ij} - \tau \log s_{ij} - \tau + \alpha_i + \beta_j = 0 \tag{16}$$

$$s_{ij} = \exp(\frac{1}{\tau}(\alpha_i - \tau)) \exp(\frac{1}{\tau}\mu_{ij}) \exp(\frac{1}{\tau}\beta_j) \tag{17}$$

As such, the solution S is of the form $\mathrm{A}\exp(\frac{1}{\tau}\mu)\mathrm{B}$ for certain diagonal matrices A, B with positive diagonals. Then, since the matrix $\exp(\frac{1}{\tau}\mu)$ is strictly positive, the solution $S \in \mathcal{T}_{uv}$ exists, is unique, and can be found with the IPF procedure (Ruschendorf, 1995).

7.2 Solution of the Randomly Perturbed Entropy-Regularized Problem

The Lagrangian of the randomly perturbed entropy-regularized problem is

$$\mathcal{L}(\mu, \gamma, s, \tau) = \arg\max_{s \in \mathcal{T}_{uv}} \sum_i \sum_j s_{ij}(\mu_{ij} + \gamma_{ij} - \tau \log s_{ij}) - \sum_{i=1}^{n} \alpha_i (u_i - \sum_{j=1}^{m} s_{ij}) - \sum_{j=1}^{m} \beta_j (v_j - \sum_{i=1}^{n} s_{ij}). \tag{18}$$

Its solution for each s_{ij} is given by

$$\frac{\partial \mathcal{L}(\mu, \gamma, s, \tau)}{\partial s_{ij}} = \mu_{ij} + \gamma_{ij} - \tau \log s_{ij} - \tau + \alpha_i + \beta_j = 0 \tag{19}$$

$$s_{ij} = \exp(\frac{1}{\tau}(\alpha_i - \tau)) \exp(\frac{1}{\tau}(\mu_{ij} + \gamma_{ij})) \exp(\frac{1}{\tau}\beta_j) \tag{20}$$

As such, the solution S is of the form $\mathrm{A}\exp(\frac{1}{\tau}(\mu + \gamma(y)))\mathrm{B}$ for certain diagonal matrices A, B with positive diagonals. Then, since the matrix $\exp(\frac{1}{\tau}(\mu + \gamma(y)))$ is strictly positive, the solution $S \in \mathcal{T}_{uv}$ exists, is unique, and can be found with the IPF procedure (Ruschendorf, 1995).

7.3 Experiments

We use unchanged peer methods implementations in the unified ThinkMatch project as much as possible. Our code was written in adherence with the setting of the ThinkMatch project to allow comparison.

Datasets. Datasets should be downloaded and organized as instructed in the ThinkMatch project.

Average imbalance in sampled pairs of images from the Pascal VOC and the CUB2011 dataset. We measure the datasets' average keypoints imbalance (max/min) in randomly sampled pairs of images of the same category.

The CUB2011 dataset has an average keypoint imbalance of 1.137 with a standard deviation 0.16 based on 1,548 sampled pairs of images.

The Pascal VOC dataset has an average keypoint imbalance of 1.4 based on 6,385 sampled pairs of images. Results in Table 6 show that the average imbalance over all sampled pairs per class. The class with the lowest imbalance is 'bottle' (1.14) and the class with the highest imbalance is 'sheep' (1.96).

Settings. Partial matching is formed by setting both problem configurations $TGT\_OUTLIER$ and $SRC\_OUTLIER$ to $TRUE$ and set in the configuration file: $MATCHING\_TYPE =' Unbalanced'$ and $filter\_type =' NoFilter'$. We run experiments on an Nvidia Tesla K80 12GB GPU.

Table 6: The average overall and per-class imbalance (max/min) of the number of keypoints in $6,385$ pairs of images sampled from the Pascal VOC dataset.

Class	✈	🚲	🐦	🚢	🍾	🚌	🚗	🐈	🪑	🐄	🍽	🐕	🐎	🏍	🚶	🪴	🐑	🛋	🚆	🖥	average
imbalance	1.32	1.19	1.464	1.32	1.14	1.22	1.38	1.458	1.4	1.364	1.708	1.4	1.42	1.42	1.45	1.267	1.96	1.52	1.42	1.26	1.40

Table 7: The average overall and per-class imbalance (max/min) of the number of keypoints in $2,785$ pairs of images sampled from the IMC-PT-SparseGM dataset.

Class	**Average Imbalance**
brandenburg_gate	2.00
buckingham_palace	1.88
colosseum_exterior	1.76
grand_place_brussels	1.72
hagia_sophia_interior	1.85
notre_dame_front_facade	1.92
palace_of_westminster	1.80
pantheon_exterior	1.69
prague_old_town_square	1.99
reichstag	1.41
taj_mahal	1.54
temple_nara_japan	1.45
trevi_fountain	1.62
westminster_abbey	1.78

7.3.1 Pascal VOC

Following prior research, each image is cropped to its bounding box and scaled to 256×256 px.

General hyper-parameters and settings. Batches consist of 26 pairs of sampled images from the same category class. Training is set for 20 epochs. All relaxation techniques were run for 25 iterations. Optimization is carried with Adam optimizer, with a learning rate of 0.002 and momentum of 0.9 and a scheduled decay of 0.5 at epochs $2, 4, 6, 8, 10$. An element-wise Sigmoid function is applied to the scoring function prior to the IPF procedure to ensure its positive values. We set a temperature τ=1 in all our experiments.

Results.

The test set's per-class average matching prediction accuracy (recall) accuracies are reported in Table 8.

Table 8: Partial matching average accuracy (in %) per class on the Pascal VOC test set. The same keypoint filtering was applied on all methods (preserving outlier keypoints in both images). Method names are abbreviated. Best results are in bold.

Class	✈	🚲	🐦	🚢	🍾	🚌	🚗	🐈	🪑	🐄	🍽	🐕	🐎	🏍	🚶	🪴	🐑	🛋	🚆	🖥
CIE-H	34.0	59.1	47.0	33.7	81.5	54.1	31.9	47.1	28.3	46.2	52.7	45.0	45.4	50.0	29.3	82.9	39.2	35.4	56.1	76.5
qc-DGN	30.9	59.8	48.8	40.5	79.6	51.7	32.5	55.8	27.5	52.1	48.0	50.7	57.3	60.3	28.1	90.8	51.0	35.5	71.5	79.9
BB-GM	42.9	64.3	54.9	48.0	84.7	65.9	45.9	59.9	**40.1**	63.6	49.1	60.2	58.7	**62.3**	39.0	92.7	56.0	40.6	75.9	**86.4**
NGM-v2	41.9	**65.9**	54.7	47.4	83.5	**68.9**	59.2	53.8	37.6	56.3	34.6	55.1	52.7	55.3	41.8	87.7	47.0	39.0	71.0	78.4
Ours: Gumbel-IPF	**47.4**	65.5	**62.3**	**47.9**	**88.9**	64.3	**65.4**	**62.3**	40.0	**64.8**	**50.9**	**66.5**	**63.0**	61.8	**46.8**	**94.9**	**57.7**	**42.4**	**81.7**	83.6

Further, as the class average imbalance increases, our method's improvement over the baseline increases (Table 8)

7.3.2 IMC-PT-SparseGM

General hyper-parameters and settings. Batches consist of 8 pairs of sampled images from the same category class. Training is set for 20 epochs. All relaxation techniques were run for 25 iterations. Optimization is carried with Adam optimizer, with a learning rate of 0.001 and momentum of 0.9 and a scheduled decay of

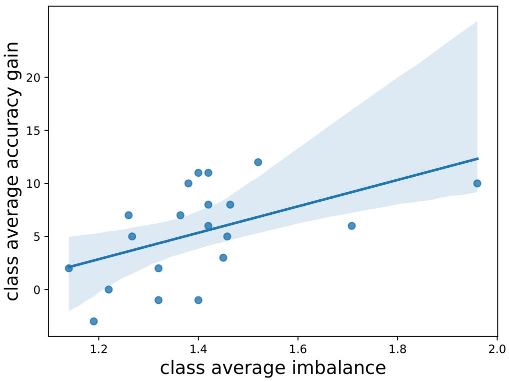


Figure 8: Average accuracy improvement of our method over the baseline of 20 Pascal VOC classes versus average class imbalance. A linear regression model fit is depicted.

0.1 at epochs $2, 6, 10$. An element-wise exponential function is applied to the scoring function prior to the IPF procedure to ensure its positive values. We set a temperature $\tau=1$ in all our experiments.

We follow prior research, such that training is carried out on samples from the 13 classes, and evaluation is performed on the remaining three classes: the Reichstag, the Sacré-Coeur, and St. Peter's square.

Results. The test set's per-class average matching prediction accuracy (recall) accuracies are reported in Table 9.

Table 9: Partial matching per-class average accuracy and average F_1 score on the IMC-PT-SparseGM dataset. The keypoint filtering preserves outlier keypoints in both images. Best results are in bold.

Method	Reichstag	Sacré-Coeur	St. Peter's square
IPCA-GM	64.5%	28.2%	42.0%
Ours: Gumbel-IPF	62.2%	31.2%	46.3%

7.3.3 CUB2011

General hyper-parameters and settings. Batches consist of 8 pairs of sampled images from the same category class. Training is set for 20 epochs. All relaxation techniques were run for 25 iterations. Optimization is carried with Adam optimizer, with a learning rate of 0.001 and momentum of 0.9 and a scheduled decay of 0.1 at epochs $2, 4, 6, 10$. An element-wise sigmoid function is applied to the scoring function prior to the IPF procedure to ensure its positive values. We set a temperature $\tau=1$ in all our experiments.

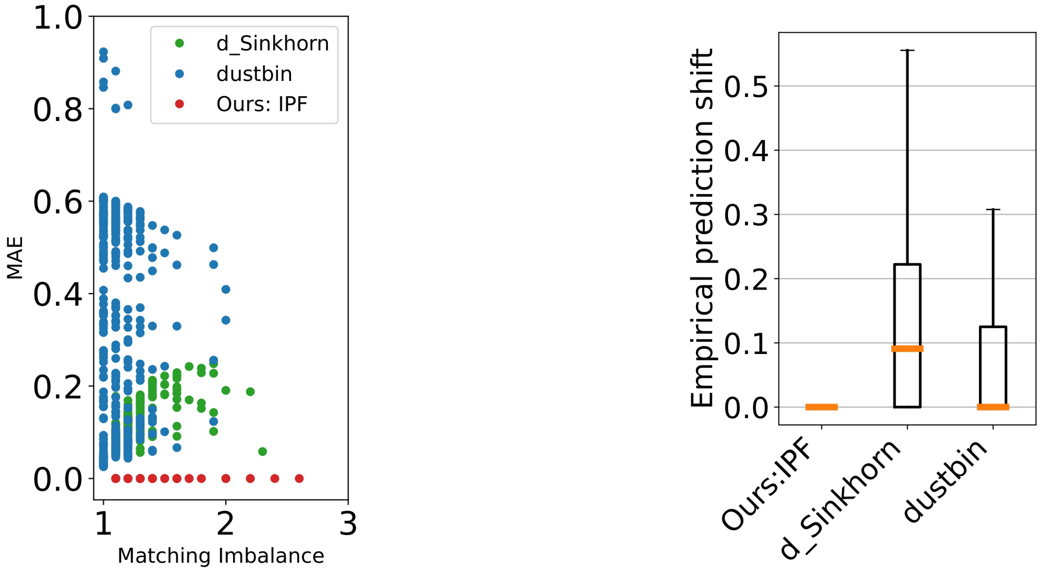
We carry out the analysis of the properties of continuous relaxation methods on the IPCA-GM architecture backbone and the CUB2011 dataset.

Mean Absolute Error. Figure 9a demonstrates that peer methods suffer high MAE for all matching imbalances, while our method reaches nearly constant zero MAE.

Prediction Shift. Figure 9b shows that the method of Sinkhorn normalization with dummy elements suffers a high prediction shift. Compared to our method, the 'dustbin' method exhibits more dispersed statistics, with a significantly higher upper 3^{rd} quantile (Figure 9b) and maximum.

7.3.4 Properties Of Continuous Partial Matching Relaxation Techniques

Peer Methods The 'dustbin' method of performing Sinkhorn normalization is based on (Sarlin et al., 2020) and authors' published code https://github.com/magicleap/SuperGluePretrainedNetwork. The common 'dummy-elements' method of performing Sinkhorn normalization is based on the implementation of Wang et al. (2021) and authors' project code ThinkMatch.



(a) Mean absolute error from target marginals versus matching imbalance of partial matching relaxation techniques on 1k random samples of image pairs on the IPCA-GM backbone on the CUB2011 dataset. Cases of full matchings are not displayed.

(b) Empirical prediction shift of partial matching relaxation methods on the IPCA-GM backbone on the CUB2011 dataset. Cases of full matchings are not displayed.

Empirical Prediction Shift Of An Intermediate Layer To measure the empirical prediction shift of an intermediate layer in the NGM-v2 backbone, we construct an affinity matrix, denoted μ^K, from pairs of node embeddings $f_i^s \in \mathbb{R}^{1024}, f_j^t \in \mathbb{R}^{1024}$, $i = 1, .., n$,$j = a, ..m$, where f_i^s denotes the i^{th} node embedding of image I^s and similarly for f_j^t. These node embeddings are an intermediate representation, used to construct the association graph. We experiment with two affinity measures: Gaussian kernel and cosine similarity. The matching that maximizes the affinity matrix prior to relaxation is predicted, denoted $y^*(\mu^K)$. Then, the affinity matrix is normalized by each technique, and the empirical prediction shift is measured. Our method in this experiment derives the target marginals solely from $y^*(\mu^K)$. All methods exhibit a negligible empirical prediction shift in full matchings. However, in partial matching, the method of Sinkhorn normalization with dummy elements suffers a high prediction shift in both affinities (medians 31% and 28%). In comparison, the prediction shift is lower with the 'dustbin' method (medians 8% and 6%), while our method exhibits a negligible empirical prediction shift in both affinities (Figure 10).

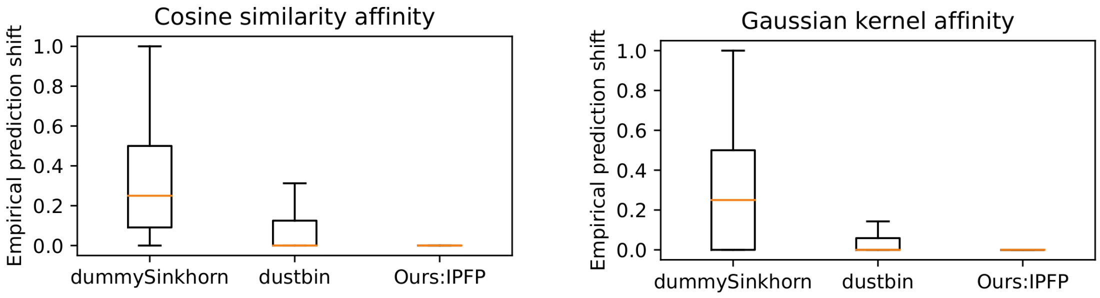


Figure 10: Empirical prediction shift of partial matching normalization methods measured on pairwise node cosine similarity and Gaussian kernel affinities between intermediate node embeddings. Cases of full matchings are not displayed.

In all methods, the scoring function is not perturbed with random noise, to allow comparison that isn't influenced by the stochastic nature of random perturbation.