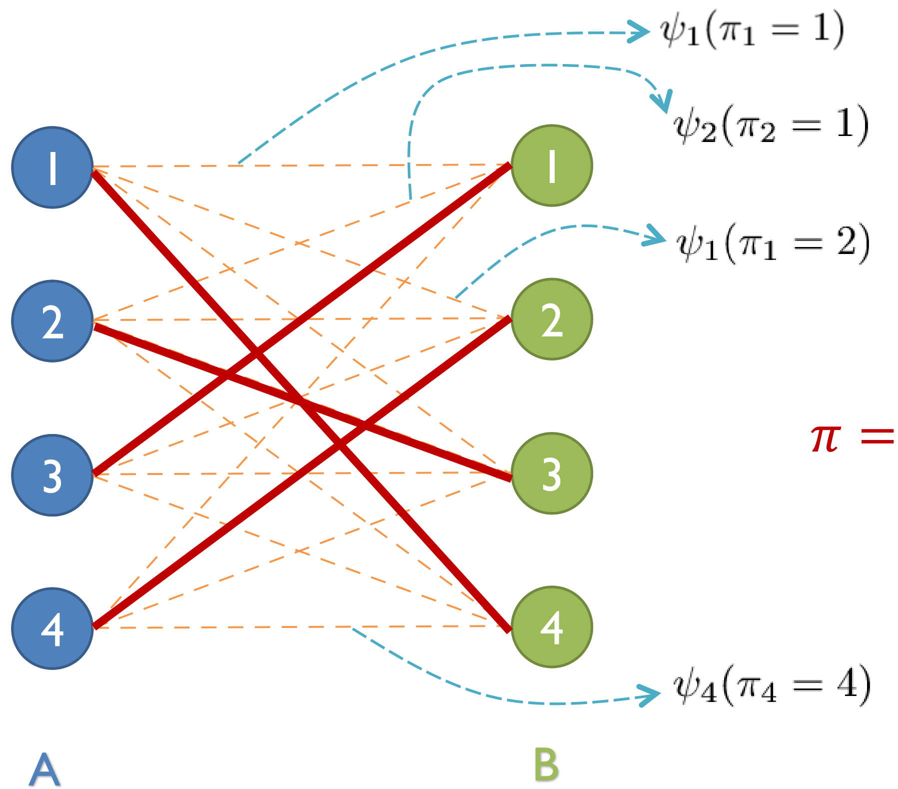
Efficient and **Consistent** Adversarial Bipartite Matching

Rizal Fathony*,#, Sima Behpour*,
Xinhua Zhang, Brian D. Ziebart

*) equal contribution
#) presenter

Bipartite Matching Tasks



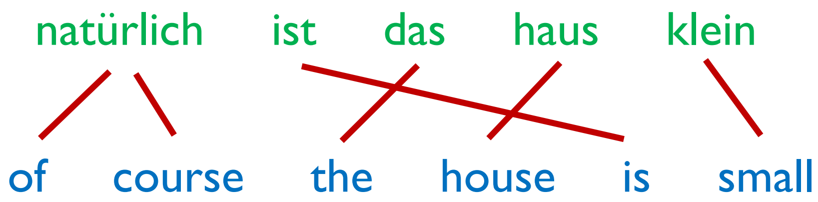
$\pi = [4, 3, 1, 2]$

Maximum weighted bipartite matching: $\max_{\pi \in \Pi} \psi(\pi) = \max_{\pi \in \Pi} \sum_i \psi_i(\pi_i)$

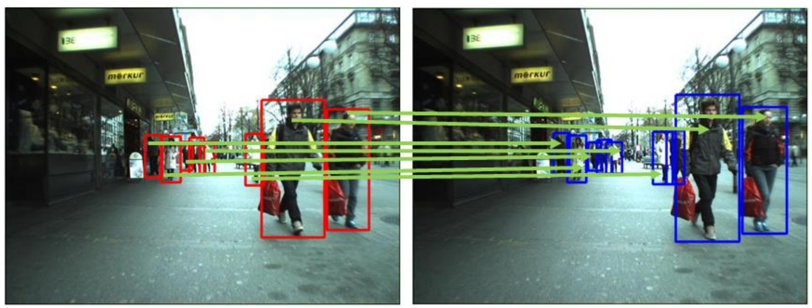
Machine learning task: Learn the appropriate weights $\psi_i(\cdot)$

Learning Bipartite Matching | Applications

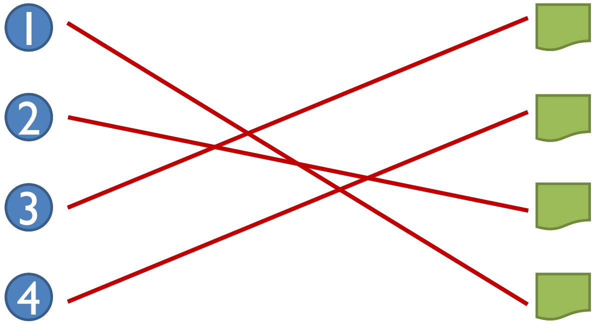
❶ **Word alignment** (Taskar et. al., 2005; Pado & Lapta, 2006; Mac-Cartney et. al., 2008)

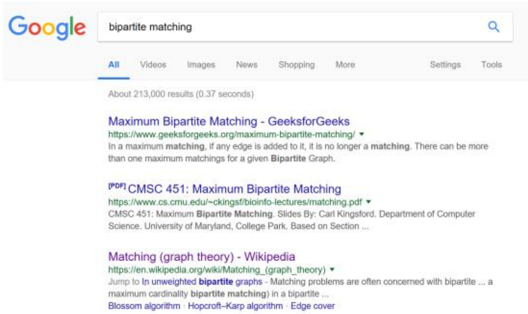


❷ **Correspondence between images** (Belongie et. al., 2002; Dellaert et. al., 2003)



❸ **Learning to rank documents** (Dwork et. al., 2001; Le & Smola, 2007)





Desiderata for a Predictor

Learning objective:

seek pairwise potentials that most compatible with training data

Challenge:

loss functions (e.g. Hamming loss): non-continuous & non convex

Desiderata for predictor:

❶ **Efficiency**

runtime: (low degree) polynomial time

❷ **Consistency**

must also minimize Hamming loss under ideal condition
(given the true distribution and fully expressive model parameters)

Exponential Family Random Field Approach

(Petterson et. al., 2009; Volkovs & Zemel, 2012)

Probabilistic model:

$$P_\psi(\pi) = \frac{1}{Z_\psi} \exp\left(\sum_{i=1}^{n} \psi_i(\pi_i)\right)$$

$$Z_\psi = \sum_\pi \prod_{i=1}^{n} \exp\left(\psi_i(\pi_i)\right) = \text{perm}(\mathbf{M})$$

$$\text{where} M_{i,j} = \exp\left(\psi_i(j)\right)$$

#P-hard

(Valiant, 1979)

✓ Consistent?

produce Bayes optimal prediction in an ideal condition

✗ Efficient?

normalization term Z_ψ involves matrix permanent computation

impractical even for modestly-size $n = 20$

Maximum Margin Approach

(Tsochantaridis et. al., 2005)

Max-margin model:

$$\min_{\psi} \mathbb{E}_{\pi \sim \tilde{P}} \left[\max_{\pi'} \{ \text{loss}(\pi, \pi') + \psi(\pi') \} - \psi(\pi) \right]$$

$\tilde{P}$ is the empirical distribution

✓ Efficient?

polynomial algorithm for computing maximum violated constraint: (Hungarian algorithm)

✗ Consistent?

- based on Crammer & Singer multiclass SVM formulation
- is not consistent for distribution with no majority label (Liu, 2007)

Adversarial Bipartite Matching

(our approach)

Seek a predictor that robustly minimize Hamming loss

against the worst-case permutation mixture probability

$$\min_{\hat{P}(\hat{\pi}|x)} \max_{\check{P}(\check{\pi}|x)} \mathbb{E}_{x\sim\tilde{P};\hat{\pi}|x\sim\hat{P};\check{\pi}|x\sim\check{P}} [\text{loss}(\hat{\pi}, \check{\pi})]$$

$$\text{s.t.} \ \mathbb{E}_{x\sim\tilde{P};\check{\pi}|x\sim\check{P}} \left[\sum_{i=1}^{n} \phi_i(x, \check{\pi}_i) \right] = \mathbb{E}_{(x,\pi)\sim\tilde{P}} \left[\sum_{i=1}^{n} \phi_i(x, \pi_i) \right]$$

Predictor:

- makes a probabilistic prediction $\hat{P}(\hat{\pi}|x)$
- aims to minimize the loss
- is pitted with an adversary instead of the empirical distribution

Adversary:

- makes a probabilistic prediction $\check{P}(\check{\pi}|x)$
- aims to maximize the loss
- constrained to select probability that match the statistics of empirical distribution ($\tilde{P}$) via moment matching on the features $\phi(x,\pi) = \sum_{i=1}^{n} \phi_i(x,\pi_i)$

Adversarial Bipartite Matching | Dual

Dual Formulation of the Adversarial Bipartite Matching

(methods of Lagrange multipliers, Von Neumann & Sion minimax duality)

$$\min_{\theta} \mathbb{E}_{x,\pi\sim\tilde{P}} \min_{\hat{P}(\hat{\pi}|x)} \max_{\check{P}(\check{\pi}|x)} \mathbb{E}_{\substack{\hat{\pi}|x\sim\hat{P} \\ \check{\pi}|x\sim\check{P}}} \left[\text{loss}(\hat{\pi},\check{\pi}) + \theta \cdot \sum_{i=1}^{n} \left(\phi_i(x,\check{\pi}_i) - \phi_i(x,\pi_i)\right) \right]$$

Lagrangian term

where θ is the dual variable for moment matching constraints

Hamming loss

Augmented Hamming loss matrix for $n = 3$ permutations

	$\check{\pi}=123$	$\check{\pi}=132$	$\check{\pi}=213$	$\check{\pi}=231$	$\check{\pi}=312$	$\check{\pi}=321$
$\hat{\pi}=123$	$0+\delta_{123}$	$2+\delta_{132}$	$2+\delta_{213}$	$3+\delta_{231}$	$3+\delta_{312}$	$2+\delta_{321}$
$\hat{\pi}=132$	$2+\delta_{123}$	$0+\delta_{132}$	$3+\delta_{213}$	$2+\delta_{231}$	$2+\delta_{312}$	$3+\delta_{321}$
$\hat{\pi}=213$	$2+\delta_{123}$	$3+\delta_{132}$	$0+\delta_{213}$	$2+\delta_{231}$	$2+\delta_{312}$	$3+\delta_{321}$
$\hat{\pi}=231$	$3+\delta_{123}$	$2+\delta_{132}$	$2+\delta_{213}$	$0+\delta_{231}$	$3+\delta_{312}$	$2+\delta_{321}$
$\hat{\pi}=312$	$3+\delta_{123}$	$2+\delta_{132}$	$2+\delta_{213}$	$3+\delta_{231}$	$0+\delta_{312}$	$2+\delta_{321}$
$\hat{\pi}=321$	$2+\delta_{123}$	$3+\delta_{132}$	$3+\delta_{213}$	$2+\delta_{231}$	$2+\delta_{312}$	$0+\delta_{321}$

size:

$n! \times n!$

Intractable

for modestly-sized n

Efficient Algorithms

1. Double Oracle Method (Constraint Generations)

2. Marginal Distribution Formulation

Double Oracle Method

Based on the observation:

equilibrium is usually supported by small number of permutations

Iterative procedure:

	$\check{\pi}$=123	$\check{\pi}$=213	$\check{\pi}$=312
$\hat{\pi}$=123	$0+\delta_{123}$	$2+\delta_{213}$	$3+\delta_{213}$
$\hat{\pi}$=213	$2+\delta_{123}$	$0+\delta_{213}$	$2+\delta_{213}$
$\hat{\pi}$=312	$3+\delta_{123}$	$2+\delta_{213}$	$0+\delta_{213}$

Adversary's best response: $\check{\pi}$=213
Adversary's best response: $\check{\pi}$=312

Predictor's best
Predictor's best

- no formal polynomial bound is known
- runtime: cannot be characterized as polynomial

Marginal Distribution Formulation

Marginal Distribution Matrices:

Predictor

$\mathbf{P} =$

	1	2	3
$\hat{\pi}_1$	$p_{1,1}$	$p_{1,2}$	$p_{1,3}$
$\hat{\pi}_2$	$p_{2,1}$	$p_{2,2}$	$p_{2,3}$
$\hat{\pi}_3$	$p_{3,1}$	$p_{3,2}$	$p_{3,3}$

$$p_{i,j} = \hat{P}(\hat{\pi}_i = j)$$

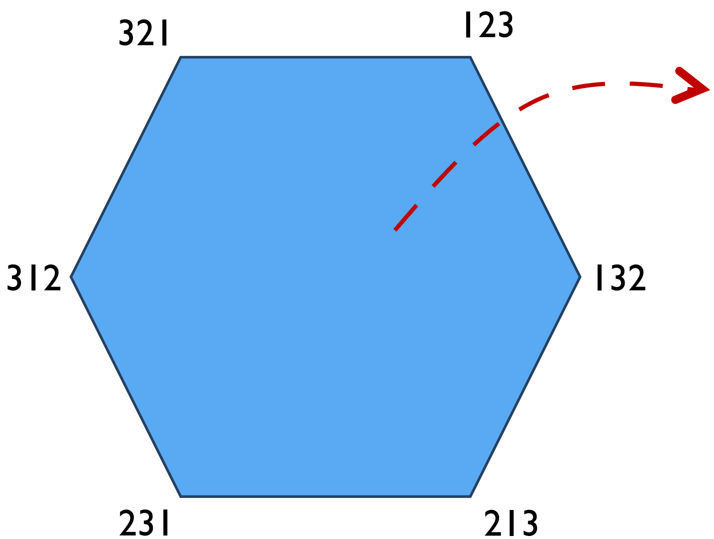
Adversary

$\mathbf{Q} =$

	1	2	3
$\check{\pi}_1$	$q_{1,1}$	$q_{1,2}$	$q_{1,3}$
$\check{\pi}_2$	$q_{2,1}$	$q_{2,2}$	$q_{2,3}$
$\check{\pi}_3$	$q_{3,1}$	$q_{3,2}$	$q_{3,3}$

$$q_{i,j} = \check{P}\,(\check{\pi}_i = j)$$

Birkhoff – Von Neumann theorem:



convex polytope whose points are doubly stochastic matrix

$$\mathbf{P1} = \mathbf{P}^{\top}\mathbf{1} = \mathbf{Q1} = \mathbf{Q}^{\top}\mathbf{1} = \mathbf{1}$$

reduce the space of optimization:
from $O(n!)$ to $O(n^2)$

Marginal Formulation | Optimization

Optimization:

add regularization and smoothing penalty

$$\max_{\mathbf{Q}\geq\mathbf{0}} \min_{\theta} \frac{1}{m} \sum_{i=1}^{m} \min_{\mathbf{P}_i\geq\mathbf{0}} \left[\langle \mathbf{Q}_i - \mathbf{Y}_i, \sum_k \theta_k \mathbf{X}_{i,k} \rangle - \langle \mathbf{P}_i, \mathbf{Q}_i \rangle + \frac{\mu}{2} \|\mathbf{P}_i\|_F^2 - \frac{\mu}{2} \|\mathbf{Q}_i\|_F^2 \right] + \frac{\lambda}{2} \|\theta\|_2^2$$

$$\text{s.t. : } \mathbf{P}_i\mathbf{1} = \mathbf{P}_i^\top\mathbf{1} = \mathbf{Q}_i\mathbf{1} = \mathbf{Q}_i^\top\mathbf{1} = \mathbf{1}, \quad \forall i$$

Techniques:

- Outer (Q): * projected Quasi-Newton (Schmidt, et.al., 2009)
  * projection to doubly-stochastic matrix
- Inner (θ): closed-form solution
- Inner (P): projection to doubly-stochastic matrix
- Projection to doubly-stochastic matrix : ADMM

Consistency

Empirical Risk Perspective of Adversarial Bipartite Matching

$$\min_{\theta} \mathbb{E}_{\substack{x\sim P \\ \pi|x\sim \tilde{P}}} \left[AL^{\text{perm}}_{f_\theta}(x,\pi) \right]$$

$$\text{where: } AL^{\text{perm}}_{f_\theta}(x,\pi) \triangleq \min_{\hat{P}(\hat{\pi}|x)} \max_{\check{P}(\check{\pi}|x)} \mathbb{E}_{\substack{\hat{\pi}|x\sim \hat{P} \\ \check{\pi}|x\sim \check{P}}} \left[\text{loss}(\hat{\pi},\check{\pi}) + f_\theta(x,\check{\pi}) - f_\theta(x,\pi) \right]$$

Consistency:

our method also minimize the Hamming loss in ideal case.

arg-max of f is in the Bayes optimal responses

$$f^* \in \mathcal{F}^* \triangleq \operatorname*{argmin}_{f} \mathbb{E}_{\pi|x\sim P} \left[AL^{\text{perm}}_{f}(x,\pi) \right]$$

$$\Rightarrow \operatorname*{argmax}_{\pi} f^*(x,\pi) \subseteq \Pi^\diamond \triangleq \operatorname*{argmin}_{\pi} \mathbb{E}_{\bar{\pi}|x\sim P}[\text{loss}(\pi,\bar{\pi})]$$

Experiment Setup

Application: Video Tracking

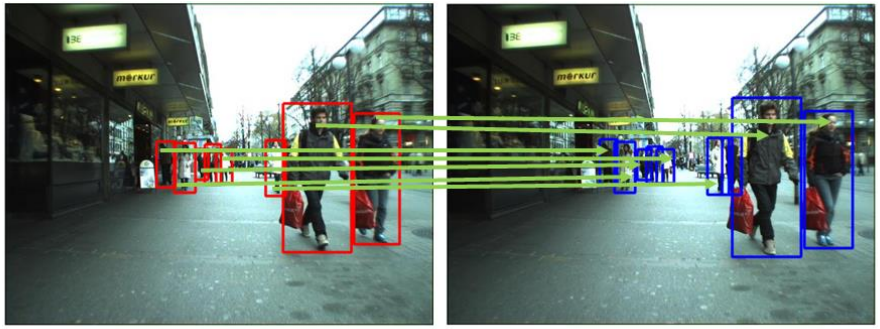


Table 3. Dataset properties

DATASET	# ELEMENTS	# EXAMPLES
TUD-CAMPUS	12	70
TUD-STADTMITTE	16	178
ETH-SUNNYDAY	18	353
ETH-BAHNHOF	34	999
ETH-PEDCROSS2	30	836

Empirical runtime (until convergence)

Table 5. Running time (in seconds) of the model for various number of elements n with fixed number of samples ($m = 50$)

DATASET	# ELEMENTS		ADV MARG.		SSVM
CAMPUS	12	1.0	1.96	1.0	0.22
STADTMITTE	16	1.3	2.46	1.2	0.25
SUNNYDAY	18	1.5	2.75	1.4	0.15
PEDCROSS2	30	2.5	8.18	4.2	0.26
BAHNHOF	34	2.8	9.79	5.0	0.31

relative: 12=1.0 relative: 1.96=1.0

Adv. Marginal Form.: grows (roughly) quadratically in n

CRF: impractical even for $n = 20$
(Petterson et. al., 2009)

Experiment Results

Table 4. The mean and standard deviation (in parenthesis) of the average accuracy (1 - the average Hamming loss) for the adversarial bipartite matching model compared with Structured-SVM.

Training/ Testing	Adv DO	Adv Marg.	SSVM	Adv DO #Perm.
Campus/ Stadtmitte	0.662 (0.09)	0.662 (0.08)	0.662 (0.08)	11.4
Stadtmitte/ Campus	0.672 (0.12)	0.667 (0.11)	0.660 (0.12)	7.4
Bahnhof/ Sunnyday	**0.758** (0.12)	**0.754** (0.10)	0.729 (0.15)	5.8
Pedcross2/ Sunnyday	**0.760** (0.08)	**0.750** (0.10)	0.736 (0.13)	8.2
Sunnyday/ Bahnhof	**0.755** (0.20)	**0.751** (0.18)	0.739 (0.20)	9.8
Pedcross2/ Bahnhof	**0.760** (0.12)	**0.763** (0.16)	0.731 (0.21)	10.8
Bahnhof/ Pedcross2	**0.718** (0.16)	**0.714** (0.16)	0.701 (0.18)	8.5
Sunnyday/ Pedcross2	**0.719** (0.18)	**0.712** (0.17)	0.700 (0.18)	14.4

6 pairs of dataset significantly outperforms SSVM

2 pairs of dataset competitive with SSVM

Adv. Double Oracle: small number of permutations

Conclusions

	Efficient?	Consistent?	Perform well?
Exponential Family Random Field (Petterson et. al., 2009; Volkovs & Zemel, 2012)	✗	✓	?
Maximum Margin (Tsochantaridis et. al., 2005)	✓	✗	??
Adversarial Bipartite Matching (our approach)	✓	✓	✓

THANK YOU