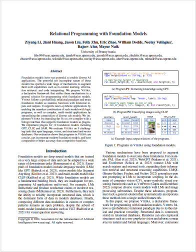
Training neural models using logic: results, challenges, and applications

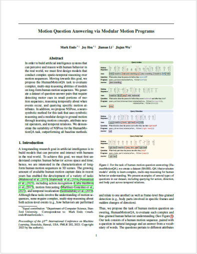
Efi Tsamoura

Why neurosymbolic AI?
(Some of) our success stories

Applications in foundational models

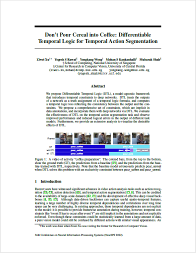

Improved Logical Reasoning of Language Models via Differentiable Symbolic Programming


Relational Programming with Foundation Models


Motion Question Answering via Modular Motion Programs

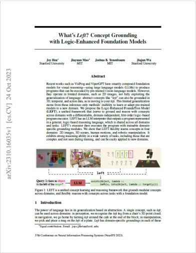
Ziyang Li, et al. **Relational Programming with Foundation Models**. In AAAI, 2024.
Hanlin Zhang, et al. **Improved Logical Reasoning of Language Models via Differentiable Symbolic Programming**. In ACL, 2023.
Joy Hsu, et al. **What's Left? Concept Grounding with Logic-Enhanced Foundation Models**. In NeurIPS 2023.

Applications in computer vision



Don't Pour Cereal into Coffee: Differentiable Temporal Logic for Temporal Action Segmentation

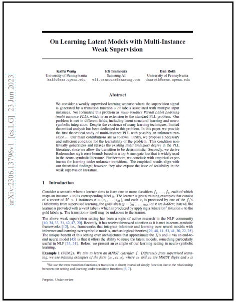
Scene Graph Generation with External Knowledge and Image Reconstruction



What's *Left*? Concept Grounding with Logic-Enhanced Foundation Models

Ziwei Xu, et al. **Don't Pour Cereal into Coffee: Differentiable Temporal Logic for Temporal Action Segmentation**. In NeurIPS, 2022.
Jiuxiang Gu, et al. **Scene Graph Generation with External Knowledge and Image Reconstruction**. In CVPR, 2019.
Mark Endo, et al. **Motion Question Answering via Modular Motion Programs**. In ICML, 2023.
Davide Buffelli and **Efthymia Tsamoura**. **Scalable Theory-Driven Regularization of Scene Graph Generation Models**. In AAAI, 2023.
Leon Jonathan Feldstein, Jurčius Modestas, and **Efthymia Tsamoura**. **Parallel neurosymbolic integration with Concordia**. In ICML, 2023.

About this talk


arXiv:2306.13796v1 [cs.LG] 23 Jun 2023

On Learning Latent Models with Multi-Instance Weak Supervision

Kaifu Wang University of Pennsylvania kaifu@sas.upenn.edu
Efi Tsamoura Samsung AI efi.tsamoura@samsung.com
Dan Roth University of Pennsylvania danroth@seas.upenn.edu

Abstract

We consider a weakly supervised learning scenario where the supervision signal is generated by a transition function σ of labels associated with multiple input instances. We formulate this problem as *multi-instance Partial Label Learning (multi-instance PLL)*, which is an extension to the standard PLL problem. Our problem is met in different fields, including latent structural learning and neuro-symbolic integration. Despite the existence of many learning techniques, limited theoretical analysis has been dedicated to this problem. In this paper, we provide the first theoretical study of multi-instance PLL with possibly an unknown transition σ. Our main contributions are as follows. Firstly, we propose a necessary and sufficient condition for the learnability of the problem. This condition non-trivially generalizes and relaxes the existing *small ambiguity degree* in the PLL literature, since we allow the transition to be deterministic. Secondly, we derive Rademacher-style error bounds based on a top-k surrogate loss that is widely used in the neuro-symbolic literature. Furthermore, we conclude with empirical experiments for learning under unknown transitions. The empirical results align with our theoretical findings; however, they also expose the issue of scalability in the weak supervision literature.

1 Introduction

Consider a scenario where a learner aims to learn one or more classifiers $f_1, \dots, f_n$, each of which maps an instance x to its corresponding label y. The learner is given training examples that consist of a *vector* of $M > 1$ instances $\mathbf{x} = (x_1, \dots, x_M)$, and each x_i is processed by one of the f_j's. Differently from supervised learning, the gold labels $\mathbf{y} = (y_1, \dots, y_M)$ of $\mathbf{x}$ are *hidden*; instead, the learner is provided with a *weak* label s which is produced by applying a *transition*[1] *function* σ to the gold labels $\mathbf{y}$. The transition σ itself may be unknown to the learner.

The above weak supervision setting has been a topic of active research in the NLP community [40, 34, 33, 31, 42, 47, 20]. Recently, it has received renewed attention as it is met in *neuro-symbolic* frameworks [12], i.e., frameworks that integrate inference and learning over neural models with inference and learning over symbolic models, such as logical theories [29, 48, 11, 53, 41, 30, 22, 25]. The unique benefit of this setting over architectures that approximate the f_i's and σ via an end-to-end neural model [45] is that it offers the ability to reuse the latent models, something particularly useful in NLP [33, 31]. Below, we present an example of our learning setting in neuro-symbolic learning.

Example 1 (SUM2). *We aim to learn an MNIST classifier f. Differently from supervised learning, we use training examples of the form (x_1, x_2, s), where x_1 and x_2 are MNIST digits and s is*

[1] We use the term transition function (or transition in short) instead of simply function due to the relationship between our setting and learning under transition functions [6, 7].

Preprint. Under review.

We will focus on <u>weakly supervised learning using logic.</u>

We will cover:

- Learnability.
  - That has been an *open problem*.
- New challenges that don't appear in traditional ML.

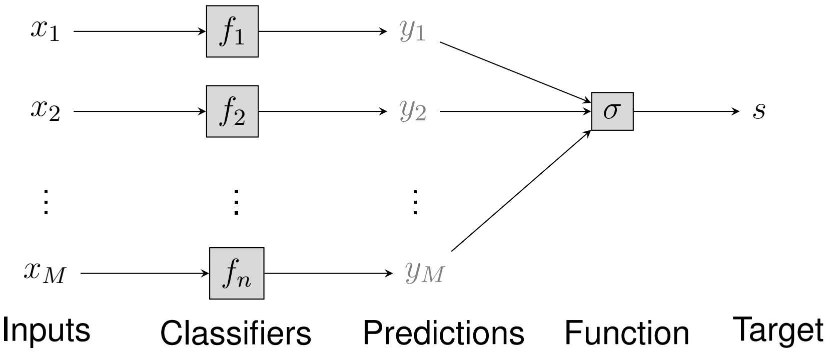
Kaifu Wang, **Efthymia Tsamoura**, and Dan Roth. **On learning latent models with multi-instance weak supervision**. In NeurIPS, 2023.

Weakly-supervised learning using logic

aka Multi-Instance Partial Label Learning (MI-PLL)

Kaifu Wang, **Efthymia Tsamoura**, and Dan Roth. **On learning latent models with multi-instance weak supervision**. In NeurIPS, 2023.

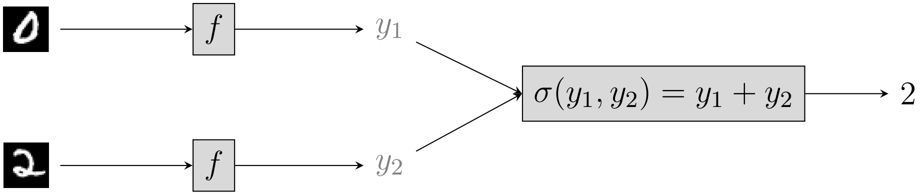
MI-PLL


x_1
x_2
x_M
f_1
f_2
f_n
y_1
y_2
y_M
σ
s
Inputs
Classifiers
Predictions
Function
Target

MI-PLL

- **Given**:
  - $x_1, \ldots, x_M$,
  - trainable classifiers $f_1, \ldots, f_n$,
  - a target $s = \sigma(y_1, \ldots, y_M)$, where y_i's are the predictions of the classifiers on x_i's,
- **learn** $f_1, \ldots, f_n$.

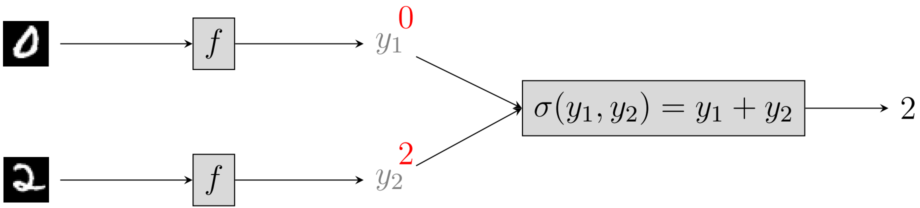
MI-PLL: 2SUM



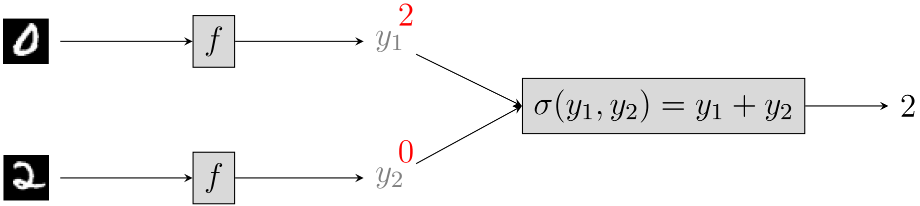
Challenges

- σ may not be 1-1.

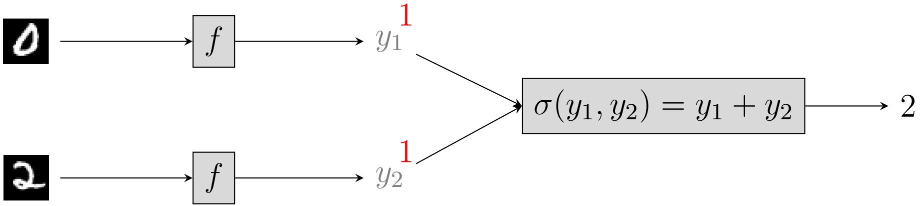
Challenges: σ may not be 1-1



Challenges: σ may not be 1-1



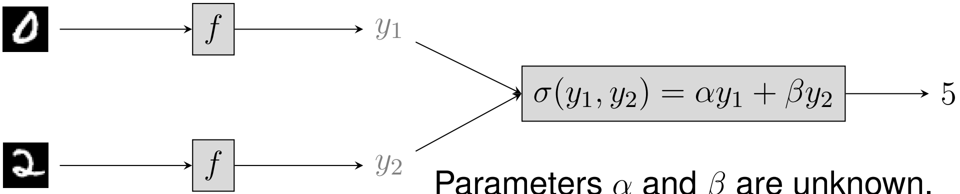
Challenges: σ may not be 1-1



Challenges

- σ may not be 1-1.
- σ may be unknown.

Challenges: σ may be unknown



Parameters α and β are unknown.

– Learning under unknown weighted sums allows us to simultaneously perform *theory induction* due to the relationship between integer linear programming and Boolean satisfiability [32].

Related work

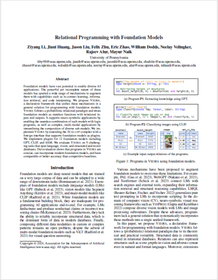
– MI-PLL is topic of active research in NLP [12, 25, 27, 28, 33, 38, 42].

– Renewed attention in *neurosymbolic* learning [9, 10, 13, 19, 23, 37, 43, 47].

– Applications in foundational models [20, 49].

Improved Logical Reasoning of Language Models via Differentiable Symbolic Programming

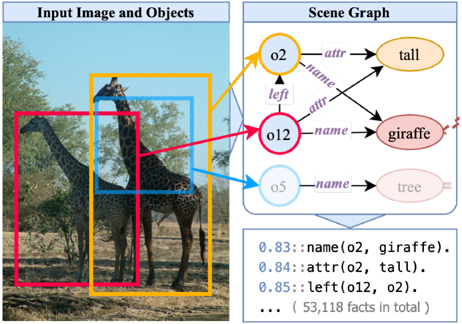
Hanlin Zhang, Jiani Huang, Ziyang Li, Mayur Naik, Eric Xing



Relational Programming with Foundation Models

Ziyang Li, Jiani Huang, Jason Liu, Felix Zhu, Eric Zhao, William Dodds, Neelay Velingker, Rajeev Alur, Mayur Naik

Visual QA (SIGMOD 2023)



$$\mathsf{Q}(O) \leftarrow \mathsf{NAME}(herbivore, O)$$

$$\mathsf{NAME}(N, O) \wedge \mathsf{NAME}(N', O) \rightarrow \mathsf{ISA}(N', N)$$

$$\rightarrow \mathsf{ISA}(giraffe, herbivore)$$

$$\rightarrow \mathsf{ISA}(dear, herbivore)$$

Table: Recall@5 on VQAR [13].

Testset	LXMERT [34]	RVC [11]	TG-Guided VQA
C5	64.05%	74.62%	**87.01%**
C6	56.51%	72.04%	**85.45%**

Efthymia Tsamoura, Jaehun Lee, and Jacopo Urbani. **Probabilistic Reasoning as Scale: Trigger Graphs to the Rescue**. In SIGMOD, 2023.

On the power of σ

Our formulation is general enough to represent different languages, e.g.,

- non-linear functions.
- Systems of Boolean equations.
- Datalog.

Our formulation can express logical theories via backward reasoning [15].

Efthymia Tsamoura and Loizos Michael **Neural-Symbolic Integration: a Compositional Perspective**. In AAAI, pages 5051-5060, 2021.

Benefits of our learning setting

The unique benefit over end-to-end neural models [41] is that it offers the ability to reuse the latent models– particularly useful in NLP [25, 27].

Objective

- Develop **necessary** and **sufficient** conditions that ensure classifier *learnability*– will formally introduced the notion later.
- When σ is known, this condition is called M**-unambiguity**.

Neurosymbolic losses

- Losses based on weighted model counting [4, 45].
- Losses based on fuzzy logic semantics [31, 39].
- Learning based on expectation maximization [19, 29].
- Learning via differentiation through argmax [25, 27].

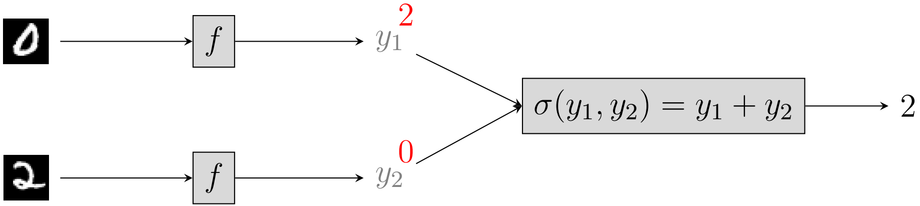
We will not cover this topic in this talk.

Kaifu Wang, **Efthymia Tsamoura**, and Dan Roth. **On learning latent models with multi-instance weak supervision**. In NeurIPS, 2023.
Kaifu Wang, **Efthymia Tsamoura**, and Dan Roth. **On characterizing and mitigating imbalances in multi-instance weak supervision**. CoRR, abs/2407.10000, 2024.

Notation

Supervised learning	MI-PLL	Meaning
x (given)	$\boldsymbol{x} = x_1, \ldots, x_M$ (given)	input(s)
y (given)	$\boldsymbol{y} = y_1, \ldots, y_M$ (unknown)	gold label(s)
-	$s = \sigma(\boldsymbol{y})$ (given)	partial label
-	σ (given)	transition function
$\mathcal{D}$	$\mathcal{D}_\mathsf{P}$	training distribution (drawing M independent samples from $\mathcal{D}$)
$[f](x)$	$[f](x)$	prediction
$\ell^{01}(y, y') := 1\{y \neq y'\}$	$\ell^{01}_\sigma(\boldsymbol{y}, s) := 1\{\sigma(\boldsymbol{y}) \neq s\}$	zero-one (partial) loss
$\mathcal{R}^{01}(f) :=$ $E_{(X,Y)\sim\mathcal{D}}[\ell^{01}([f](X), Y)]$	$\mathcal{R}^{01}_\mathsf{P}(f;\sigma) :=$ $E_{(\mathbf{X},S)\sim\mathcal{D}_\mathsf{P}}[\ell^{01}_\sigma([f](\mathbf{X}), S)]$	zero-one (partial) risk

Notation: 2SUM



- $\sigma(y_1, y_2) = y_1 + y_2$.
- $\ell^{01}_{\sigma}(y_1 = \mathbf{2}, y_2 = \mathbf{0}, s = \mathbf{2}) = \mathbf{0}$.
- $\ell^{01}_{\sigma}(y_1 = \mathbf{2}, y_2 = \mathbf{1}, s = \mathbf{2}) = \mathbf{1}$.
- The partial risk $\mathcal{R}^{01}_{\mathrm{P}}(f; \sigma)$ is the probability of predicting the wrong sum.

PAC-style learnability

An MI-PLL problem instance is *learnable*, if there exists an algorithm $\mathcal{A}$, that takes as **input** partial samples and **outputs** a classifier $\mathcal{A}(\mathcal{T}_{\mathsf{P}}) \in \mathcal{F}$, such that

- for any data distribution and
- any $\delta, \epsilon \in (0, 1)$

there is an integer $m_{\epsilon,\delta}$, such that $m_{\mathsf{P}} \geq m_{\epsilon,\delta}$, where m_{P} is the size of partial samples, implies $\mathcal{R}^{01}(\mathcal{A}(\mathcal{T}_{\mathsf{P}})) \leq \epsilon$ with probability at least $1 - \delta$.

PAC-style learnability: informal definition

An MI-PLL problem instance is *learnable*, if for any user-defined $\delta, \epsilon \in (0,1)$, it is highly likely (with probability at least $1-\delta$), that the learned classifier does few mistakes ($\mathcal{R}^{01}(\mathcal{A}(\mathcal{T}_{\mathsf{P}})) \leq \epsilon$), via using a large enough number of training samples ($m_{\mathsf{P}} \geq m_{\epsilon,\delta}$).

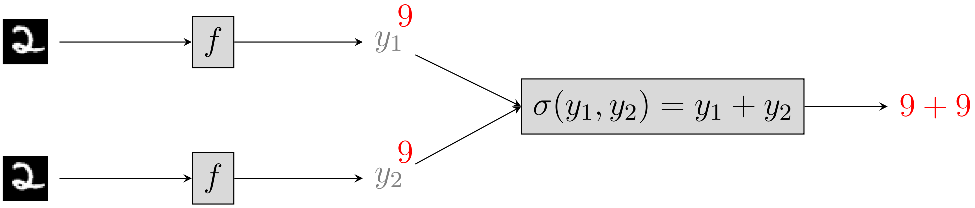
Learnability: intuition

- To prove learnability of an MI-PLL problem instance, we must bound $\mathcal{R}^{01}(f)$ (zero-one risk) with $\mathcal{R}^{01}_{\mathsf{P}}(f;\sigma)$ (zero-one **partial** risk), under *any* training distribution.

Learnability: intuition

– In other words, **mistakes** under the partial training samples, should be informative of the **classification errors** made by f, under *any* training distribution.

Learning under the spike distribution: intuition



- Suppose the mass is concentrated in a single element  (gold label is **2**).
- Suppose f misclassifies  as **9**.
- Then, the gold labels are **(2, 2)**, but the classifier outputs **(9, 9)**.
- If **2 + 2 = 9 + 9**, then $\mathcal{R}^{01}_{\mathrm{P}}(f; \sigma) = \mathbf{0}$, while $\mathcal{R}^{01}(f) \neq \mathbf{0}$.
- Hence, classifier errors are **concealed**.

A sufficient and necessary learnability condition

Definition (M-unambiguity)

Transition σ is M-*unambiguous* if for any two label vectors $\boldsymbol{y} = (y, \dots, y)$ and $\boldsymbol{y}' = (y', \dots, y')$, such that $\boldsymbol{y} \neq \boldsymbol{y}'$, we have $\sigma(\boldsymbol{y}) \neq \sigma(\boldsymbol{y}')$.

Let's map the definition to our example.

– Suppose the mass is concentrated in a single element  (gold label is **2**).

– Suppose f misclassifies  as **9**.

– Then, the gold labels are **(2, 2)**, but the classifier outputs **(9, 9)**.

– If **2 + 2 = 9 + 9**, then $\mathcal{R}^{01}_{\mathrm{P}}(f;\sigma) = \mathbf{0}$, while $\mathcal{R}^{01}(f) \neq \mathbf{0}$.

– Hence, classifier errors are **concealed**.

M-unambiguity: example

Example (Sum of two digits)

Transition $\sigma^*(y_1, y_2) \rightarrow y_1 + y_2$ **is** M-unambiguous, since for any two different integers y and y', we have:

$$y + y \neq y' + y'$$

M-unambiguity: example

Example (Product of two digits)

Transition $\sigma^*(y_1, y_2) \to y_1 \times y_2$ **is** M-unambiguous, since for any two different integers y and y', we have:

$$y \times y \neq y' \times y'$$

M-unambiguity: counter example

Example (XOR)

Transition $\sigma^*(y_1, y_2) \to y_1 \oplus y_2$ **is not** M-unambiguous, since we have:

$$0 \oplus 0 = 1 \oplus 1$$

Is M-unambiguity a good condition?

Definition (M-unambiguity)

Transition σ is M*-unambiguous* if for any two label vectors $\boldsymbol{y} = (y, \ldots, y)$ and $\boldsymbol{y}' = (y', \ldots, y')$, such that $\boldsymbol{y} \neq \boldsymbol{y}'$, we have green $\sigma(\boldsymbol{y}) \neq \sigma(\boldsymbol{y}')$.

- M-unambiguity: <u>invertibility only under inputs of the same class</u>.
- Looser conditions can be obtained when the input data distribution is not a spike.

Key result

Theorem. If σ is M-unambigous, then $\mathcal{R}^{01}(f) \leq \mathcal{O}(\mathcal{R}^{01}_{\mathsf{P}}(f;\sigma)^{1/M})$.

Learnability under M-unambguity

Theorem

Suppose $\mathcal{F}$ is realizable under ℓ_{P}^{01} and $[\mathcal{F}]$ has a finite Natarajan dimension $d_{[\mathcal{F}]}$. Then for any $\epsilon, \delta \in (0,1)$, there exists a universal constant $C_0 > 0$, such that with probability at least $1-\delta$, the empirical partial risk minimizer with $\widehat{\mathcal{R}}_{\mathsf{P}}^{01}(f;\sigma;\mathcal{T}_{\mathsf{P}}) = 0$ has a classification risk $\mathcal{R}^{01}(f) < \epsilon$, if

$$\boxed{m_{\mathsf{P}}} \geq C_0 \frac{c^{2M-2}}{\epsilon^M}\left(d_{[\mathcal{F}]} \log(6cMd_{[\mathcal{F}]}) \log\left(\frac{|\mathcal{Y}|^{2M-2}}{\epsilon^M}\right) + \log\left(\frac{1}{\delta}\right)\right)$$

Number of samples to ensure with probability $\geq 1-\delta$ that f does few mistakes ($\mathcal{R}^{01}(\mathcal{A}(\mathcal{T}_{\mathsf{P}})) \leq \epsilon$)

Summary of results

- Better convergence rates via forcing additional conditions.
- Learnability under **multiple classifiers**.
- Learnability under **non-deterministic** σ.
- Learnability under **unknown** σ.
- Rademacher error bounds under logic-based losses [45] based on weighted model counting [4, 45].
- Error bounds under approximations [13].

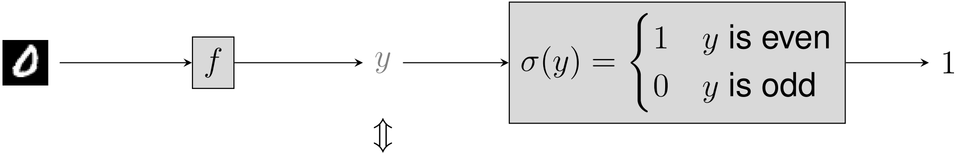
Kaifu Wang, **Efthymia Tsamoura**, and Dan Roth. **On learning latent models with multi-instance weak supervision**. In NeurIPS, 2023.

MI-PLL vs other ML problems

Relevant problems in ML

- **Partial label learning (PLL)** [2, 8, 14, 22, 30, 44, 46, 48].
- Learning via transition matrices [6, 7, 40, 50].

Example: PLL

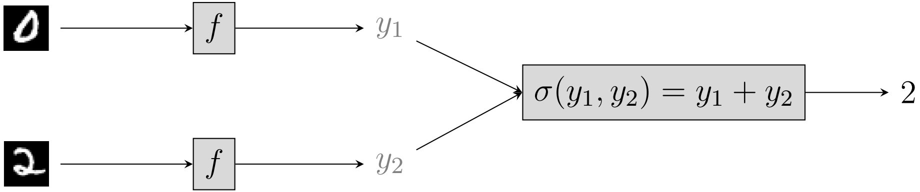


(, $\{0, 2, 4, 6, 8\}$) $\leftarrow$ PLL training example

$\uparrow$

Mutually exclusive candidate labels

Relationship of our problem to PLL



$$\Updownarrow$$

((,), $\{(0,2), (2,0), (1,1)\}$) $\leftarrow$ MI-PLL training example

$\uparrow$

Mutually exclusive candidate label vectors

Key differences with PLL

- Multiple vs single input.
- Deterministic vs stochastic σ.
  - Prior learnability results ([2, 8, 21]) rely on assumptions that are violated in our setting, i.e., that $\gamma < 1$, where

$$\gamma := \sup_{\underbrace{\mathcal{D}(x,y)}_{\text{density}} > 0, y' \neq y} \overbrace{\mathbb{P}_{(x,y)\sim\mathcal{D}}(\underbrace{y'}_{\text{noisy label}} \in \sigma(\underbrace{y}_{\text{gold label}}))}^{\text{probability noisy } y' \text{ co-occurs with } y}$$

Key differences with PLL

- Multiple vs single input.
- Deterministic vs stochastic σ.
  - Prior learnability results ([2, 8, 21]) rely on assumptions that are violated in our setting, i.e., that $\gamma < 1$, where

$$\gamma := \sup_{\underbrace{\mathcal{D}(x,y)}_{\text{density}}>0, y' \neq y} \overbrace{\mathbb{P}_{(x,y)\sim\mathcal{D}}(\underbrace{y'}_{\text{noisy label}} \in \sigma(\underbrace{y}_{\text{gold label}}))}^{\text{probability noisy } y' \text{ co-occurs with } y}$$

- M-unambiguity reduces to small ambiguity for single inputs ($M = 1$).
- Shown learnability under non-deterministic σ (proper extension of small ambiguity).

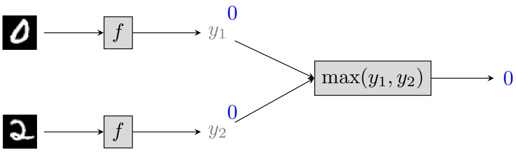
Learning imbalances in MI-PLL

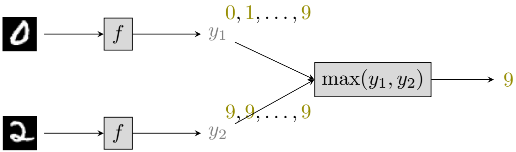
Kaifu Wang, **Efthymia Tsamoura**, and Dan Roth. **On characterizing and mitigating imbalances in multi-instance weak supervision**. CoRR, abs/2407.10000, 2024.

Learning imbalances: what are they?

– Major differences in the errors occurring when classifying instances of different classes (aka *class-specific risks*).

Learning imbalances: 2MAX

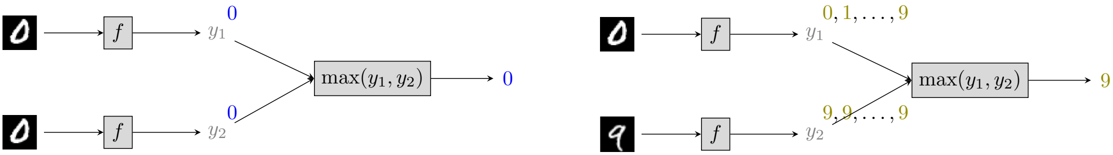




Question: Which class is <u>easier</u> to learn if *the number of ((**,**),* 0 *) samples equals the number of ((**,**),* 9 *) samples.*

Answer: Intuitively, class 0, as learning 0 reduces to *supervised learning*.

Learning imbalances: 2MAX



Question: Which class is easier to learn if *the number of* *'s equals that of* *'s.*

Answer: We have **more** samples of the form ((,), 9) than of the form ((,), 0). Hence, class 9 might be easier to learn?

Learning imbalances: 2MAX

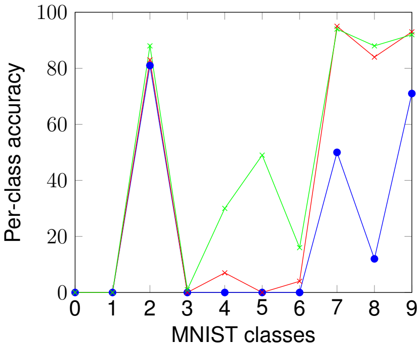


Figure: Accuracy of the MNIST classifier. Blue, red and green curves show accuracy at 20, 40 and 100 epochs. Learning converges in 100 epochs.

Learning imbalances in traditional machine learning

– Core ML problem [1, 3, 5, 16, 24, 26, 35, 36], as real data is imbalanced.

– ML techniques cannot characterize imbalances in our setting:
  – Work for long-tail data only– we also have imbalances due to σ.

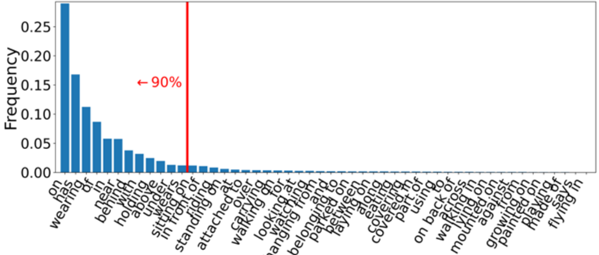
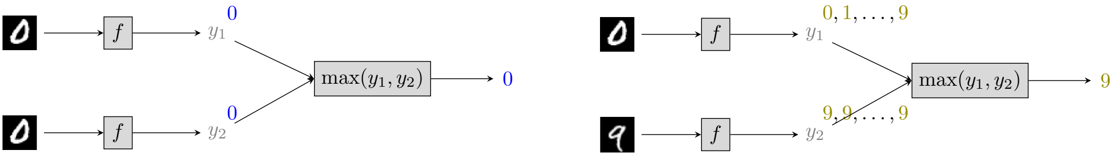


Figure: Distribution of classes in Visual Genome [17].

Learning imbalances: 2MAX



ML characterizations would naively say: it is equally difficult to learn classes 0 and 9 if the instance distributions are uniform.

Learning imbalances: theoretical characterization

– We bounded the class-specific risk $R_j(f)$ via function:

Transition, e.g., $\max$ Class

$$\Phi_{\sigma,j}(R_{\mathsf{P}}(f;\sigma))$$

Probability of wrong overall output, e.g., wrong maximum

– This bound is computed by solving a quadratic program.

Extends our previous results!

– Bound $\Phi_{\sigma,j}(R_{\mathsf{P}}(f;\sigma))$ does not rely on M-unambiguity.

– Tighter bounds than what we discussed already.

Learning imbalances: theoretical characterization

- We bounded the class-specific risk $R_j(f)$ via function:

Transition, e.g., $\max$ Class

$$\Phi_{\sigma, j}\left(R_{\mathsf{P}}(f;\sigma) \right)$$

Probability of wrong overall output, e.g., wrong maximum

- We can derive a computable bound for $R_j(f)$ using a dataset of partial sample and tools, such as VC-dimension and the Rademacher complexity.

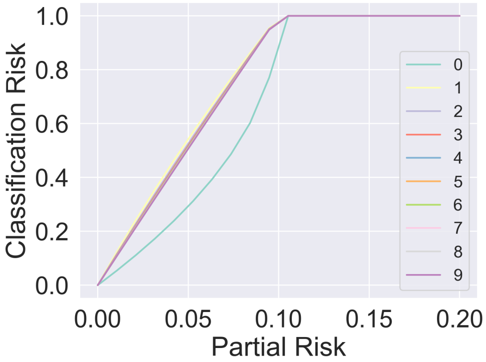
Learning imbalances: theoretical characterization

Theorem

Let $d_{[\mathcal{F}]}$ be the Natarajan dimension of $[\mathcal{F}]$, $c = |\mathcal{Y}|$, and m_{P} be the number of partial samples. Given a confidence level $\delta \in (0, 1)$, we have that $R_j(f) \leq \Phi_{\sigma,j}(\widetilde{R}_{\mathsf{P}}(f; \sigma, \mathcal{T}_{\mathsf{P}}, \delta))$ with probability $1 - \delta$ for any label $j \in [c]$, where

$$\underbrace{\widetilde{R}_{\mathsf{P}}(f; \sigma, \mathcal{T}_{\mathsf{P}}, \delta)}_{\text{Generalization bound}} = \underbrace{\widehat{R}_{\mathsf{P}}(f; \sigma, \mathcal{T}_{\mathsf{P}})}_{\text{Empirical partial risk}} + \sqrt{\frac{2\log(\mathrm{e}m_{\mathsf{P}}/2d_{[\mathcal{F}]}\log(6Mc^2d_{[\mathcal{F}]}/\mathrm{e}))}{m_{\mathsf{P}}/2d_{[\mathcal{F}]}\log(6Mc^2d_{[\mathcal{F}]}/\mathrm{e})}} + \sqrt{\frac{\log(1/\delta)}{2m_{\mathsf{P}}}}$$

Learning imbalances: 2MAX



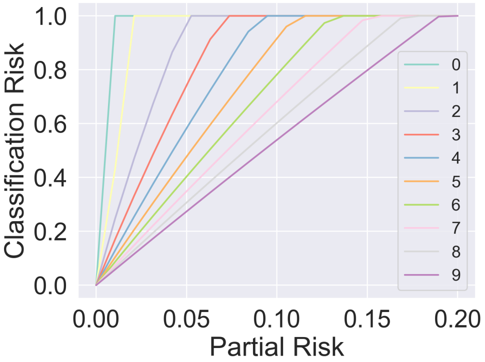


Figure: Class-specific upper bounds. (left) Partial labels are uniform. (right) Hidden labels are uniform.

Learning imbalances: testing-time mitigation

Classifier's predictions P



$$\begin{array}{c} y=0 \cdots\cdots y=9 \\ \begin{pmatrix} 0.1 & \cdots\cdots & 0.05 \\ \vdots & & \vdots \\ 0.7 & \cdots\cdots & 0.01 \\ \vdots & & \vdots \\ 0.01 & \cdots\cdots & 0.8 \end{pmatrix} \end{array}$$

Predictions for the i-th test sample

Rationale.
Given a (gold) hidden label distribution $\widehat{\boldsymbol{r}}$, correct the predictions $\boldsymbol{P}$ to $\boldsymbol{P}'$, so that $\boldsymbol{P}'$ adheres to $\widehat{\boldsymbol{r}}$.

Learning imbalances: testing-time mitigation

Classifier's predictions P

$$\begin{array}{c} y=0 \cdots\cdots y=9 \\ \begin{pmatrix} 0.1 \cdots\cdots\cdots 0.05 \\ \vdots \qquad\qquad \vdots \\ 0.7 \cdots\cdots\cdots 0.01 \\ \vdots \qquad\qquad \vdots \\ 0.01 \cdots\cdots\cdots 0.8 \end{pmatrix} \end{array}$$

Predictions for the i-th test sample

Rationale.
Given a (gold) hidden label distribution $\widehat{\boldsymbol{r}}$, correct the predictions $\boldsymbol{P}$ to $\boldsymbol{P}'$, so that $\boldsymbol{P}'$ adheres to $\widehat{\boldsymbol{r}}$.

Challenges.

- The developed technique should be lightweight.
- $\boldsymbol{P}'$ should be close *enough* to $\boldsymbol{P}$.
- $\boldsymbol{P}'$ should not strictly abide to $\widehat{\boldsymbol{r}}$ (to tolerate noise).

Learning imbalances: testing-time mitigation

Rationale. Given a (gold) hidden label distribution $\widehat{\boldsymbol{r}}$, correct the predictions $\boldsymbol{P}$ to $\boldsymbol{P}'$, so that $\boldsymbol{P}'$ adheres to $\widehat{\boldsymbol{r}}$.

Challenges:

- The developed technique should be lightweight.
- $\boldsymbol{P}'$ should be close *enough* to $\boldsymbol{P}$.
- $\boldsymbol{P}'$ should not strictly abide to $\widehat{\boldsymbol{r}}$ (to tolerate noise).

$$\min_{\boldsymbol{P}'\in\mathbb{R}_+^{n\times c},\ \underbrace{\boldsymbol{P}'\mathbf{1}_c=\mathbf{1}_n}_{\boldsymbol{P}' \text{ induces a valid distribution}}} \overbrace{\langle -\log(\boldsymbol{P}), \boldsymbol{P}'\rangle}^{\text{Closeness to original predictions}} + \overbrace{\tau KL(\boldsymbol{P}'\mathbf{1}_n || n\widehat{\boldsymbol{r}})}^{\text{Robustness to } \widehat{\boldsymbol{r}}} \tag{1}$$

Learning imbalances: testing-time mitigation

- This formulation is a *robust semi-constrained optimal transport* (RSOT) problem instance [18].
- Approximate the optimal solution using the robust semi-Sinkhorn algorithm [18].

Closeness to original predictions

Robustness to $\widehat{\boldsymbol{r}}$

$$\min_{\boldsymbol{P}' \in \mathbb{R}_+^{n \times c},\ \boldsymbol{P}'\mathbf{1}_c = \mathbf{1}_n} \langle -\log(\boldsymbol{P}), \boldsymbol{P}' \rangle + \tau KL(\boldsymbol{P}'^\top \mathbf{1}_n \| n\widehat{\boldsymbol{r}}) - \eta H(\boldsymbol{P}')$$

$\boldsymbol{P}'$ induces a valid distribution

Entropic regularization to find solutions in PTIME.

Learning imbalances: more results

- Statistically consistent technique to compute the hidden label ratios $\widehat{\boldsymbol{r}}$.
- Technique to mitigate learning imbalances at training-time.
- Improved the accuracy on multiple benchmarks by $> 20\%$.

Kaifu Wang, **Efthymia Tsamoura**, and Dan Roth. **On characterizing and mitigating imbalances in multi-instance weak supervision**. CoRR, abs/2407.10000, 2024.

Conclusions

Keywords (instead of conclusions)

- Applications.
- Scalability.
- Uncertainty– many proposals, what is the right semantics?
- Formal guarantees.

Efthymia Tsamoura, et al. **Probabilistic Reasoning at Scale: Trigger Graphs to the Rescue**. In SIGMOD, 2023.
Efthymia Tsamoura, et al. **Materializing Knowledge Bases via Trigger Graphs**. In VLDB, 2021.
Efthymia Tsamoura, et al. **Beyond the Grounding Bottleneck: Datalog Techniques for Inference in Probabilistic Logic Programs**. In AAAI, 2020.
Michael Benedikt, Boris Motik, and Efthymia Tsamoura. **Goal-Driven Query Answering for Existential Rules With Equality**. In AAAI, 2018.

Thanks!

Contact info: `efthymia.tsamoura@gmail.com`.

References I

[1] Mateusz Buda, Atsuto Maki, and Maciej A. Mazurowski. A systematic study of the class imbalance problem in convolutional neural networks. *Neural Networks*, 106:249–259, 2018.

[2] Vivien Cabannes, Alessandro Rudi, and Francis Bach. Structured prediction with partial labelling through the infimum loss. In *ICML*, page 1230–1239, 2020.

[3] Kaidi Cao, Colin Wei, Adrien Gaidon, Nikos Arechiga, and Tengyu Ma. Learning imbalanced datasets with label-distribution-aware margin loss. In *NeurIPS*, pages 1567–1578, 2019.

[4] Mark Chavira and Adnan Darwiche. On probabilistic inference by weighted model counting. *Artificial Intelligence*, 172(6):772 – 799, 2008.

[5] Nitesh V. Chawla, Kevin W. Bowyer, Lawrence O. Hall, and W. Philip Kegelmeyer. Smote: synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, 16(1):321–357, 2002.

References II

[6] Jesús Cid-Sueiro. Proper losses for learning from partial labels. In *NeurIPS*, pages 1574–1582, 2012.

[7] Jesús Cid-Sueiro, Darío García-García, and Raúl Santos-Rodríguez. Consistency of losses for learning from weak labels. In *ECML PKDD*, pages 197–210, 2014.

[8] Timothee Cour, Ben Sapp, and Ben Taskar. Learning from partial labels. *Journal of Machine Learning Research*, 12:1501–1536, 2011.

[9] Wang-Zhou Dai, Qiuling Xu, Yang Yu, and Zhi-Hua Zhou. Bridging Machine Learning and Logical Reasoning by Abductive Learning. In *NeurIPS*, pages 2815–2826, 2019.

[10] Artur S. d'Avila Garcez, Krysia Broda, and Dov M. Gabbay. *Neural-symbolic learning systems: foundations and applications*. Perspectives in neural computing. Springer, 2002.

References III

[11] Difei Gao, Ruiping Wang, Shiguang Shan, and Xilin Chen. From two graphs to N questions: A VQA dataset for compositional reasoning on vision and commonsense. *CoRR*, abs/1908.02962, 2019.

[12] Nitish Gupta, Sameer Singh, Matt Gardner, and Dan Roth. Paired examples as indirect supervision in latent decision models. In *EMNLP*, pages 5774–5785, 2021.

[13] Jiani Huang, Ziyang Li, Binghong Chen, Karan Samel, Mayur Naik, Le Song, and Xujie Si. Scallop: From probabilistic deductive databases to scalable differentiable reasoning. In *NeurIPS*, pages 25134–25145, 2021.

[14] Rong Jin and Zoubin Ghahramani. Learning with multiple labels. In *NeurIPS*, pages 897–904, 2002.

References IV

[15] Antonis C. Kakas. Abduction. In Claude Sammut and Geoffrey I. Webb, editors, *Encyclopedia of Machine Learning and Data Mining*, pages 1–8. Springer US, Boston, MA, 2017.

[16] Bingyi Kang, Saining Xie, Marcus Rohrbach, Zhicheng Yan, Albert Gordo, Jiashi Feng, and Yannis Kalantidis. Decoupling representation and classifier for long-tailed recognition. In *ICLR*, 2020.

[17] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A. Shamma, Michael S. Bernstein, and Li Fei-Fei. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *Int. J. Comput. Vis.*, 123(1):32–73, 2017.

[18] Khang Le, Huy Nguyen, Quang M Nguyen, Tung Pham, Hung Bui, and Nhat Ho. On robust optimal transport: Computational complexity and barycenter computation. In *Advances in Neural Information Processing Systems*, pages 21947–21959, 2021.

References V

[19] Zenan Li, Yuan Yao, Taolue Chen, Jingwei Xu, Chun Cao, Xiaoxing Ma, and Jian Lu. Softened symbol grounding for neurosymbolic systems. In *ICLR*, 2023.

[20] Ziyang Li, Jiani Huang, Jason Liu, Felix Zhu, Eric Zhao, William Dodds, Neelay Velingker, Rajeev Alur, and Mayur Naik. Relational programming with foundational models. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(9):10635–10644, 2024.

[21] Li-Ping Liu and Thomas G. Dietterich. Learnability of the superset label learning problem. In *ICML*, pages 1629—-1637, 2014.

[22] Jiaqi Lv, Miao Xu, Lei Feng, Gang Niu, Xin Geng, and Masashi Sugiyama. Progressive identification of true labels for partial-label learning. In *ICML*, page 6500–6510, 2020.

References VI

[23] Robin Manhaeve, Sebastijan Dumancic, Angelika Kimmig, Thomas Demeester, and Luc De Raedt. Deepproblog: Neural probabilistic logic programming. In *NeurIPS*, pages 3749–3759, 2018.

[24] Aditya Krishna Menon, Sadeep Jayasumana, Ankit Singh Rawat, Himanshu Jain, Andreas Veit, and Sanjiv Kumar. Long-tail learning via logit adjustment. In *ICLR*, 2021.

[25] Tsvetomila Mihaylova, Vlad Niculae, and André F. T. Martins. Understanding the mechanics of SPIGOT: Surrogate gradients for latent structure learning. In *EMNLP*, pages 2186–2202, 2020.

[26] Hanyu Peng, Mingming Sun, and Ping Li. Optimal transport for long-tailed recognition with learnable cost matrix. In *ICLR*, 2022.

References VII

[27] Hao Peng, Sam Thomson, and Noah A. Smith. Backpropagating through structured argmax using a SPIGOT. In *ACL*, pages 1863–1873, 2018.

[28] Aditi Raghunathan, Roy Frostig, John Duchi, and Percy Liang. Estimation from indirect supervision with linear moments. In *ICML*, volume 48, pages 2568–2577, 2016.

[29] Rajhans Samdani, Ming-Wei Chang, and Dan Roth. Unified expectation maximization. In *ACL*, pages 688–698, 2012.

[30] Junghoon Seo and Joon Suk Huh. On the power of deep but naive partial label learning. In *ICASSP*, pages 3820–3824, 2021.

[31] Luciano Serafini and Artur S. d'Avila Garcez. Logic tensor networks: Deep learning and logical reasoning from data and knowledge. *CoRR*, abs/1606.04422, 2016.

References VIII

[32] Vivek Srikumar and Dan Roth. The integer linear programming inference cookbook. *ArXiv*, abs/2307.00171, 2023.

[33] Jacob Steinhardt and Percy S Liang. Learning with relaxed supervision. In *NeurIPS*, volume 28, 2015.

[34] Hao Tan and Mohit Bansal. LXMERT: Learning cross-modality encoder representations from transformers. In *EMNLP*, pages 5100–5111, 2019.

[35] Jingru Tan, Xin Lu, Gang Zhang, Changqing Yin, and Quanquan Li. Equalization loss v2: A new gradient balance approach for long-tailed object detection. In *CVPR*, pages 1685–1694, June 2021.

[36] Jingru Tan, Changbao Wang, Buyu Li, Quanquan Li, Wanli Ouyang, Changqing Yin, and Junjie Yan. Equalization loss for long-tailed object recognition. In *CVPR*, pages 11662–11671, 2020.

References IX

[37] Efthymia Tsamoura, Timothy Hospedales, and Loizos Michael. Neural-symbolic integration: A compositional perspective. In *AAAI*, pages 5051–5060, 2021.

[38] Shyam Upadhyay, Ming-Wei Chang, Kai-Wei Chang, and Wen-tau Yih. Learning from explicit and implicit supervision jointly for algebra word problems. In *EMNLP*, pages 297–306, 2016.

[39] Emile van Krieken, E. Acar, and Frank van Harmelen. Semi-supervised learning using differentiable reasoning. *IFCoLog Journal of Logic and its Applications*, 6(4):633–653, 2019.

[40] Brendan van Rooyen and Robert C. Williamson. A theory of learning with corrupted labels. *Journal of Machine Learning Research*, 18(228):1–50, 2018.

References X

[41] Haoyu Peter Wang, Nan Wu, Hang Yang, Cong Hao, and Pan Li. Unsupervised learning for combinatorial optimization with principled objective relaxation. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *NeurIPS*, volume 35, pages 31444–31458, 2022.

[42] Lei Wang, Dongxiang Zhang, Jipeng Zhang, Xing Xu, Lianli Gao, Bing Tian Dai, and Heng Tao Shen. Template-based math word problem solvers with recursive neural networks. In *AAAI*, pages 7144–7151, 2019.

[43] Po-Wei Wang, Priya L. Donti, Bryan Wilder, and J. Zico Kolter. Satnet: Bridging deep learning and logical reasoning using a differentiable satisfiability solver. In *ICML*, 2019.

[44] Hongwei Wen, Jingyi Cui, Hanyuan Hang, Jiabin Liu, Yisen Wang, and Zhouchen Lin. Leveraged weighted loss for partial label learning. *CoRR*, abs/2106.05731, 2021.

References XI

[45] Jingyi Xu, Zilu Zhang, Tal Friedman, Yitao Liang, and Guy Van den Broeck. A semantic loss function for deep learning with symbolic knowledge. In *ICML*, pages 5502–5511, 2018.

[46] Ning Xu, Congyu Qiao, Xin Geng, and Min-Ling Zhang. Instance-dependent partial label learning. In *NeurIPS*, volume 34, pages 27119–27130, 2021.

[47] Zhun Yang, Adam Ishay, and Joohyung Lee. NeurASP: Embracing neural networks into answer set programming. In *IJCAI*, pages 1755–1762, 2020.

[48] Peilin Yu, Tiffany Ding, and Stephen H. Bach. Learning from multiple noisy partial labelers. In *PMLR*, volume 151, pages 11072–11095, 2022.

[49] Hanlin Zhang, Jiani Huang, Ziyang Li, Mayur Naik, and Eric Xing. Improved logical reasoning of language models via differentiable symbolic programming. In *ACL*, pages 3062–3077, 2023.

References XII

[50] Mingyuan Zhang, Jane Lee, and Shivani Agarwal. Learning from noisy labels with no change to the training process. In *ICML*, volume 139 of *Proceedings of Machine Learning Research*, pages 12468–12478, 2021.

Relevant problems in ML

- Partial label learning (PLL) [2, 8, 14, 22, 30, 44, 46, 48].
- **Learning via transition matrices** [6, 7, 40, 50].

Learning via transition matrices

Definition (Transition matrix [6, 40])

A *transition matrix* $\mathbf{T}$ for a learning problem with hidden label $Y \in \mathcal{Y}$ and observed label $S \in \mathcal{S}$ is a stochastic matrix, where the element in its i^{th} column and j^{th} row is the conditional probability $P(S = j|Y = i)$.

$$\mathbf{T} = \begin{array}{c} \\ s=1 \\ \vdots \\ s=j \\ \vdots \\ s=|\mathcal{S}| \end{array} \begin{array}{c} \begin{array}{ccc} y=1 & y=i & y=|\mathcal{Y}| \end{array} \\ \begin{bmatrix} \mathbb{P}(S=1|Y=1) & & \\ & & \\ & \mathbb{P}(S=j|Y=i) & \\ & & \\ & & \mathbb{P}(S=|\mathcal{S}||Y=|\mathcal{Y}|) \end{bmatrix} \end{array}$$

Learning via transition matrices

– If the transition is **invertible**, then we can compute the hidden data distribution via its association with the observed data distribution. Hence, we construct an unbiased estimator for the **classification loss**.

$$\underbrace{\boldsymbol{o}(\boldsymbol{x})}_{[\mathbb{P}(S=1|\boldsymbol{x}),\ldots,\mathbb{P}(S=|\mathcal{S}||\boldsymbol{x})]} = \overbrace{\boldsymbol{T}(\boldsymbol{x})}^{\text{Transition matrix}} \underbrace{h(\boldsymbol{x})}_{[\mathbb{P}(Y=1|\boldsymbol{x}),\ldots,\mathbb{P}(Y=|\mathcal{Y}||\boldsymbol{x})]}$$

$$\underbrace{h(\boldsymbol{x})}_{[\mathbb{P}(Y=1|\boldsymbol{x}),\ldots,\mathbb{P}(Y=|\mathcal{Y}||\boldsymbol{x})]} = \overbrace{\boldsymbol{T}^{+}(\boldsymbol{x})}^{\text{Transition matrix left inverse}} \underbrace{\boldsymbol{o}(\boldsymbol{x})}_{[\mathbb{P}(S=1|\boldsymbol{x}),\ldots,\mathbb{P}(S=|\mathcal{S}||\boldsymbol{x})]}$$

Results

– Reduction of MI-PLL to learning via transition matrices is not straightforward: naive reductions lead to non-invertible matrices :)

– For non-naive reductions, we have:
  – M-unambiguity $\not\Rightarrow$ matrix invertibility.
  – Matrix inveritbility $\not\Rightarrow$ M-unambiguity.

Kaifu Wang, **Efthymia Tsamoura**, and Dan Roth. **On learning latent models with multi-instance weak supervision**. In NeurIPS, 2023.