Learnability with Indirect Supervision Signals

Kaifu Wang[1], Qiang Ning[2], Dan Roth[1]

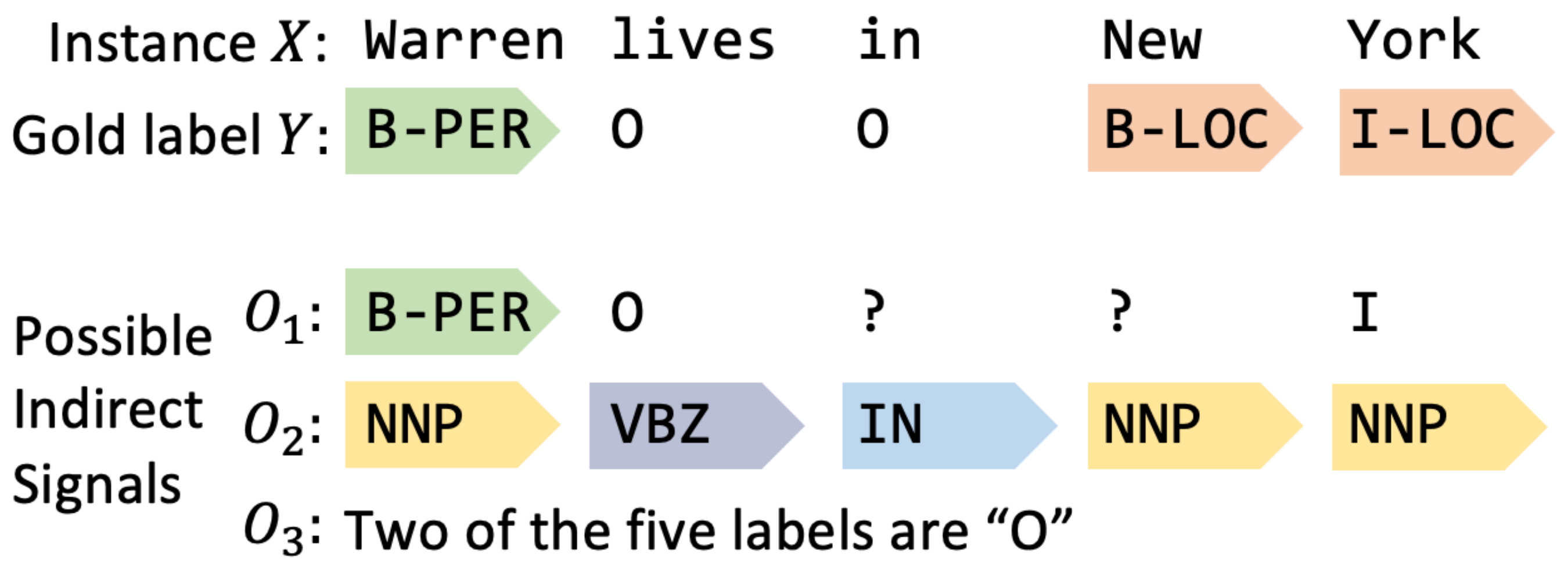
[1]University of Pennsylvania.

[2]Amazon, work done while at the Allen Institute for AI and at the University of Illinois at Urbana-Champaign

1. Motivation

In many machine learning problems, generating gold annotations for unlabeled instance is expensive or difficult. To alleviate this issue, the observation of a dependent variable (denoted by O) of the true label is often used as an **indirect supervision signal**.

Taking the named entity recognition (NER) tagging as example:



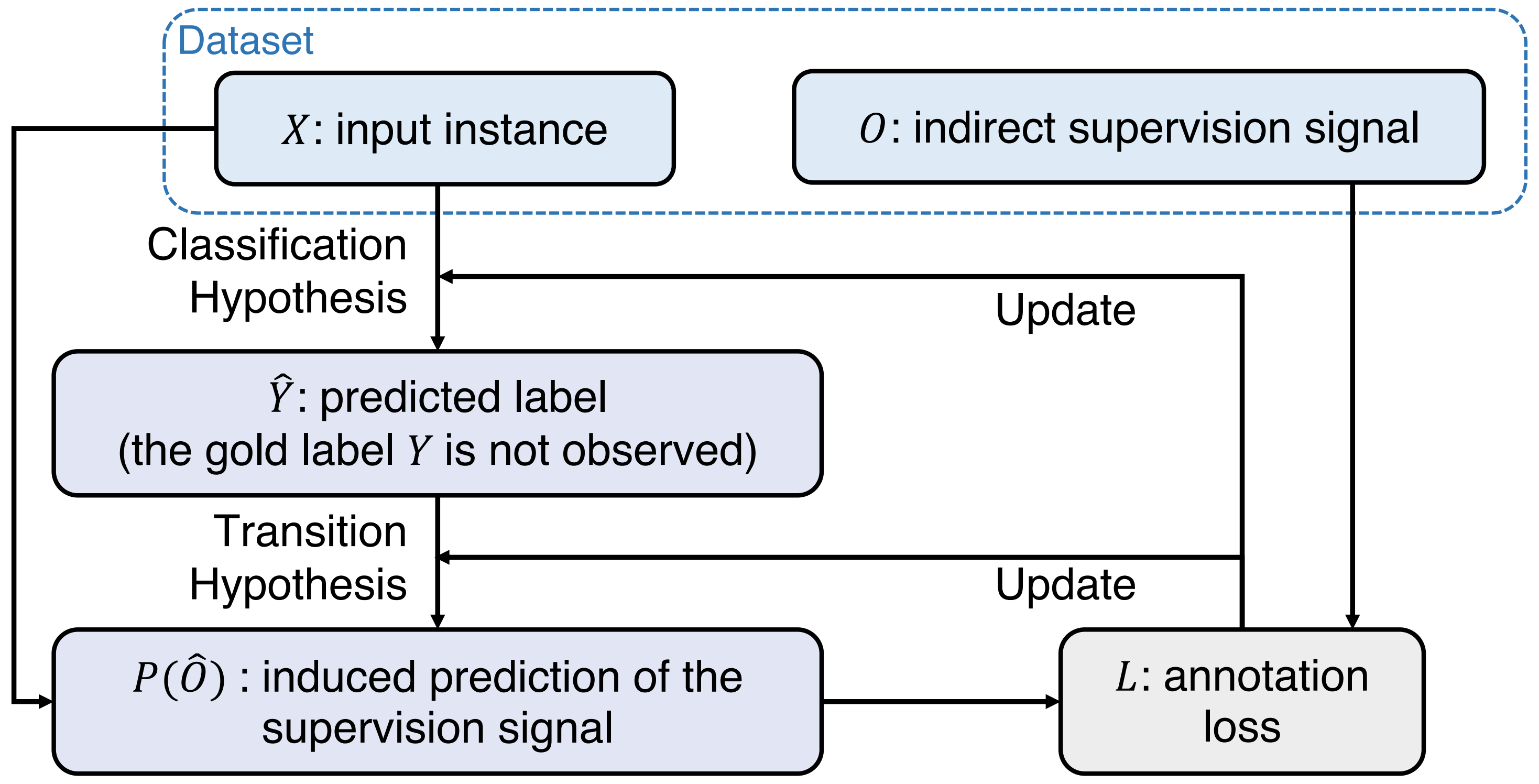
	Instance X:	Warren	lives	in	New	York
	Gold label Y:	B-PER	O	O	B-LOC	I-LOC
Possible Indirect Signals	O_1:	B-PER	O	?	?	I
	O_2:	NNP	VBZ	IN	NNP	NNP
	O_3:	Two of the five labels are “O”				

Our goal is to study **learnability** with indirect supervision. That is, whether the optimal hypothesis can be well approximated given a sufficiently large dataset of indirect signals.

The main challenge is how to quantify the *usable* information contained in the indirect signal, which not only depends on the joint distribution of the gold label and the indirect signal, but also depends on the learner’s **prior information** about this distribution.

2. Learning Framework

We propose a general learning framework which is shown in the figure. The learner not only models the classifier $X \to Y$, but also models the conditional distribution of $O|Y$ (called **transition hypothesis**). Then, the predicted label will induce predictions about O. This prediction is evaluated by the observed dataset.



Dataset

X: input instance

O: indirect supervision signal

Classification Hypothesis

Update

$\hat{Y}$: predicted label (the gold label Y is not observed)

Transition Hypothesis

Update

$P(\hat{O})$: induced prediction of the supervision signal

L: annotation loss

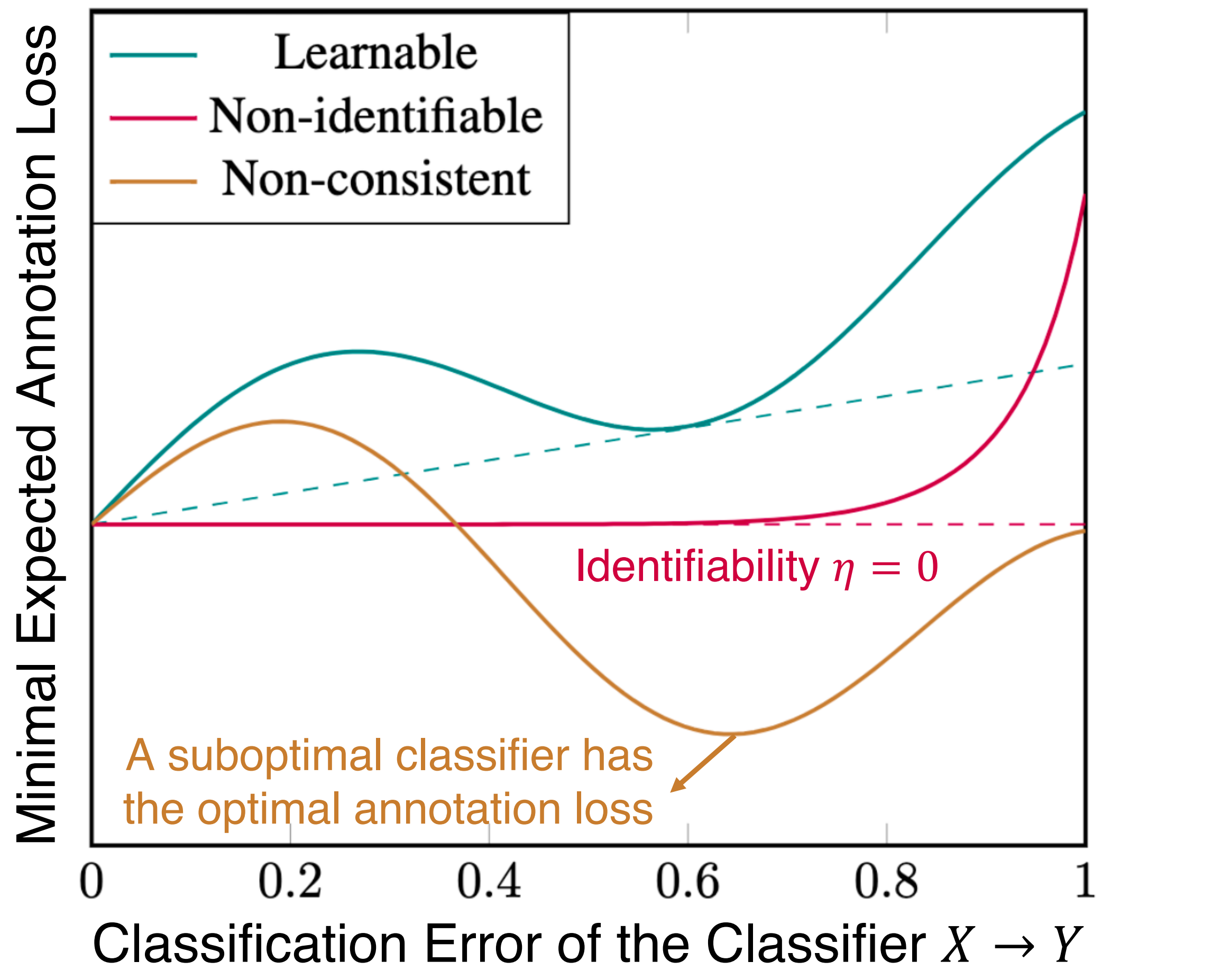
3. General Learnability Conditions

We show that learnability can be guaranteed if the following three conditions are satisfied:

- **Complexity**: the model should have a finite VC-dimension.
- **Consistency**: the optimal classifier h_0 should have the lowest annotation loss on average. Formally: $h_0 \in \underset{h\in\mathcal{H},T\in\mathcal{T}}{\operatorname{argmin}} R_{\mathcal{O}}(T \circ h)$.
- **Identifiability**: a suboptimal classifier should induce higher annotation loss than the lowest annotation loss on average. Formally, we define and require

$$\eta \overset{\text{def}}{=\!=} \inf_{h\in\mathcal{H},T\in\mathcal{T}:R(h)>0} \frac{R_{\mathcal{O}}(T \circ h) - \inf_{T\in\mathcal{T}} R_{\mathcal{O}}(T \circ h_0)}{R(h)} > 0.$$

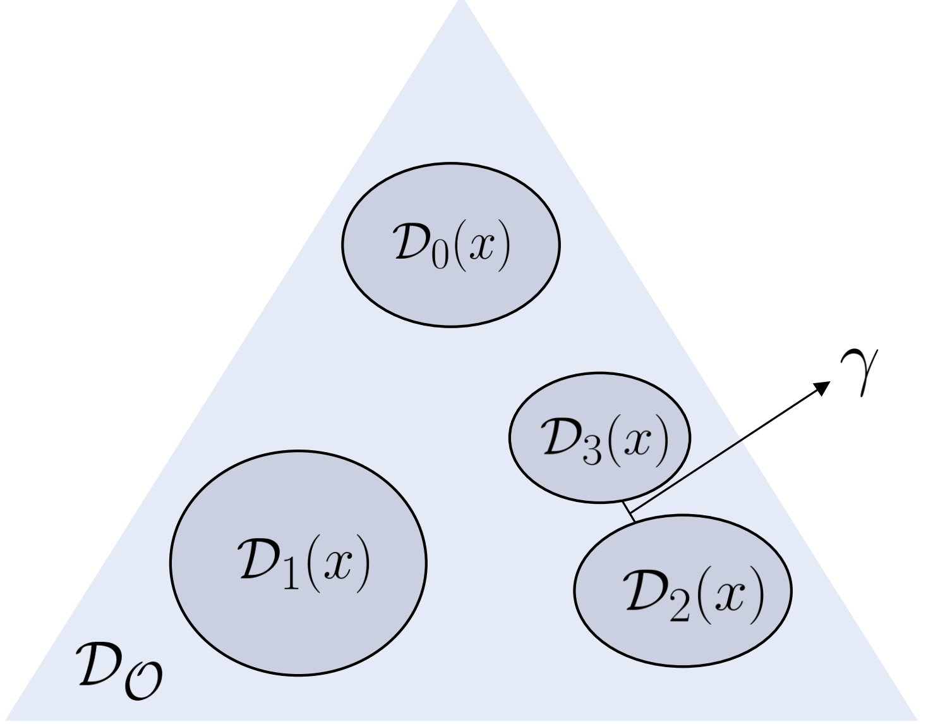
The last two conditions are visualized in the following plot of annotation loss and classification error for some artificial learning scenarios. The dashed lines’ slopes represent the identifiability η.



4. The Separation Condition

We further propose a sufficient condition for consistency and identifiability, called "separation”, which characterizes the learner’s prior information about the indirect signal. (see figure below for a 4-class classification problem)

With the learner’s transition hypothesis class, the ith label will induce a *family* of distributions of O, denoted by $\mathcal{D}_i(x)$. The separation condition requires these families to be separated by a minimum KL-divergence γ.



Probability Simplex

5. Applications

To show the application of separation, we study several cases:

Known Transition

The simplest case is when the conditional distribution of $O|Y$ is fully known to the learner. In this case, $\mathcal{D}_i(x)$ reduces to a point and separation only requires these points to be different.

Concentration

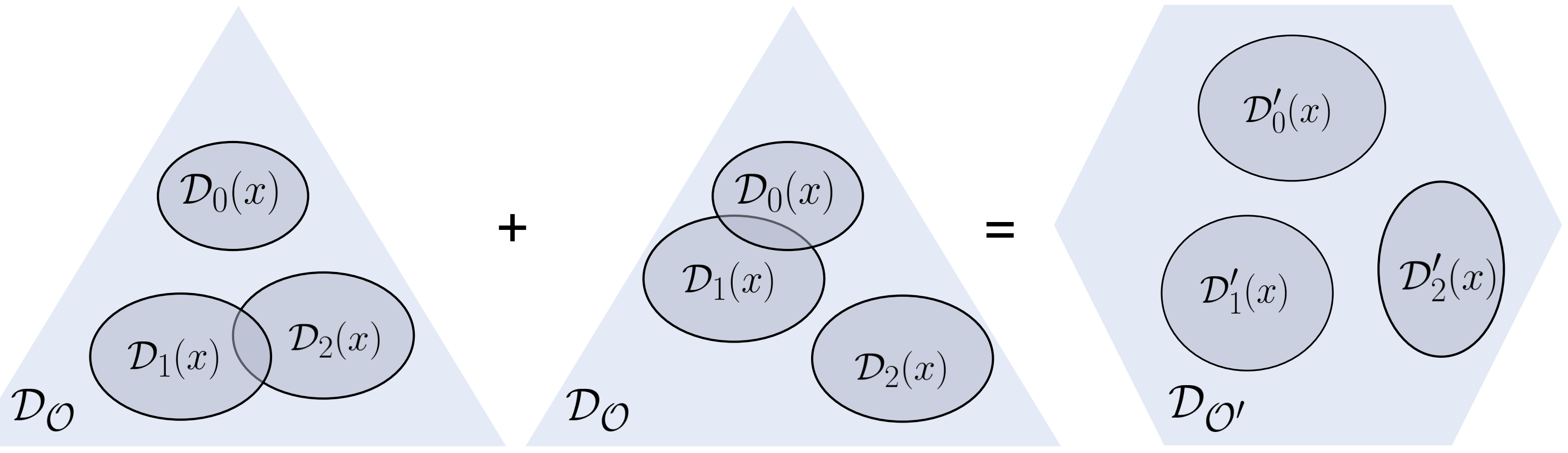
The *concentration* condition requires the distributions in different $\mathcal{D}_i(x)$ is concentrated on some different sets of O.

This condition helps to recover and extend many previously known results about learnability with partial and noisy labels.

Joint Supervision

If a single source of supervision signal cannot ensure learnability, it should be used jointly with other signals. We show that a joint supervision can:

- Harm the separation if supervision signals are simply mixed. This is due to the convexity of the KL-divergence.
- Preserve the pairwise separation if modelled properly. This effect is visualized in the following figure, where each signal cannot separate one pair of labels, but can be combined to ensure global separation.
- Create new separation: If there are additional constraints between different signals, these constraints can be utilized to supervise the learning.



Acknowledgments

This work was supported by the Army Research Office under Grant Number W911NF-20-1-0080 and by contract FA8750-19-2-0201 with the US Defense Advanced Research Projects Agency (DARPA).