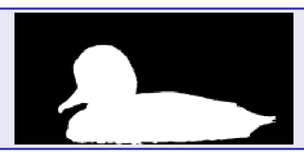
# Amortized Inference in Structured Learning

**Kai-Wei Chang, Shyam Upadhyay, Gourab Kundu, and Dan Roth**

COGNITIVE COMPUTATION GROUP UNIVERSITY OF ILLINOIS AT URBANA-CHAMPAIGN

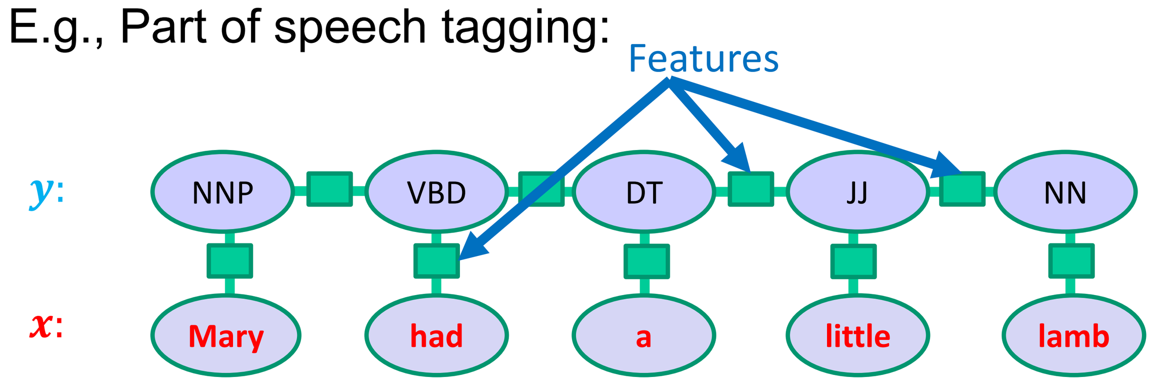
## Structured Learning

Learning how to make joint predictions

| Task | Input | Output |
|---|---|---|
| Part-of-speech Tagging | Mary had a little Lamb | NNP VBD DT JJ NN |
| Dependency Parsing | They operate ships and banks . | Root They operate ships and banks . |
| Segmentation |  |  |

Learning a score function $S(y; x, w) = w \cdot \phi(x, y)$

E.g., Part of speech tagging:




## Structured Learning Algorithms

➢ Structured Perceptron Training (Collins 02, McDonald et al 10)

Loop until stopping condition is met:

For each $(x_i, y_i)$ pair:

$$\bar{y} = \arg\max_{y} w^T \phi(x_i, y)$$

$$w \leftarrow w + \eta(\phi(x_i, y_i) - \phi(x_i, \bar{y}))$$ $\eta$: learning rate

➢ Structured SVMs

Given a set of training examples $D = \{x_i, y_i\}_{i=1}^{l}$

$$\min_{w,\xi} \frac{1}{2} w^T w + C \sum_i \xi_i^2$$

$$s.t.\ \ w^T \phi(x_i, y_i) - w^T \phi(x_i, y) \geq \Delta(y_i, y) - \xi_i \ \forall i, y \in Y_i$$

- ➢ Cutting plane: (Tsochantaridis+ 05, Joachims+ 09)
- ➢ Dual Coordinate Descent: (Shevade+ 11, Chang+ 13)
- ➢ Block-Coordinate Frank-Wolfe: (Lacoste-Julien+ 13)

They all share the following structure:

> Learning a Structured Prediction Model
>
> 1. Solving an inference problem on each sample
> 2. Update model based on the predictions

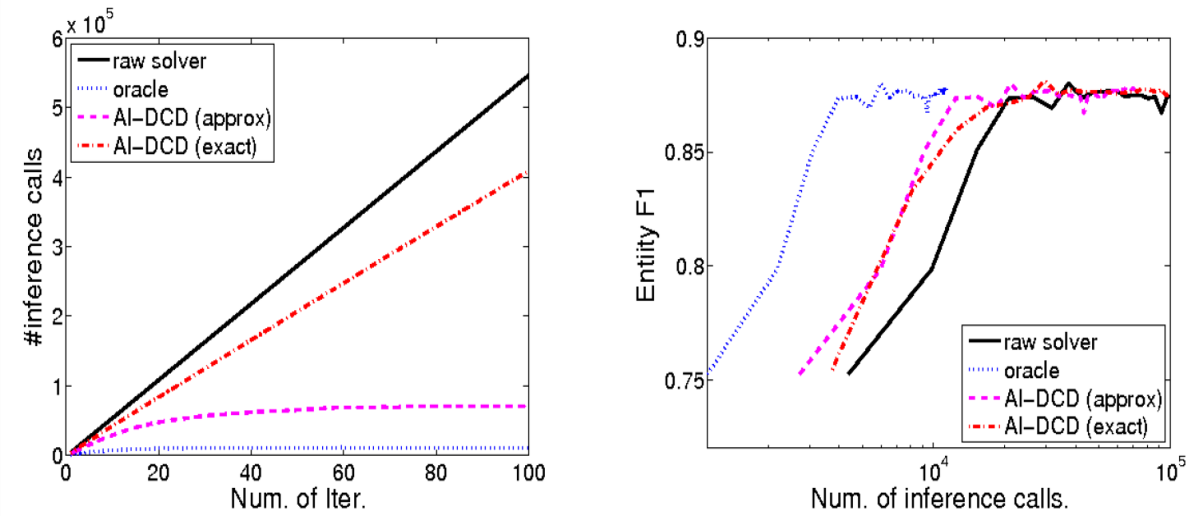
## Abstract

Training a structured prediction model involves performing many inference steps. Many of these inference problems, although different, share the same solution. We propose AI-DCD (**A**mortized **I**nference framework for **D**ual **C**oordinate **D**escent method) that accelerates the training process of Structured SVM by exploiting this redundancy of solutions.

Our results build on the ability to formulate the inference step as an integer linear program. ILP formulations can be shown to have the following property: given an ILP formulation of Inference problem A, there are conditions that determine if A has the same solution as a previously observed problem B. Our method reduces 76% of the inference calls while converging to the same model. We observe similar gains on a multi-label classification task and with a Structured Perceptron model.

## Learning with Amortization

- ➢ Many inference problems share the same solution
  - ➢ # valid outputs is small
  - ➢ Models converges after a few iterations.
- ➢ Exploit this redundancy by caching old inferences




- ➢ We need two components:
  - ➢ General framework to represent inferences
  - ➢ Conditions that determine if A has the same solution as a previously observed problem (Srikumar+ 12*; Kundu+ 13** )

```
If CONDITION(problem cache, new problem)
  then (no need to call the solver)
   SOLUTION(new problem) = old solution
Else
    Call base solver and update cache
End
```

Looking for a solution in cache takes 0.04 ms while solving an inference (e.g., Gurobi) takes 2 ms.

## General Inference Framework

- ➢ Formulating the inference as an Integer Linear Programming (ILP) (Roth & Yih 04)

$$\max_{y} \sum_c S_c y_c$$
$$Ay \leq b$$
$$y_c \in \{0,1\}$$

- ➢ Inference using ILP has been successful in NLP & Vision tasks
  - ➢ Dependency Parsing, Sentence Compression
  - ➢ Any MPE problem w.r.t. any probabilistic model, can be formulated as an ILP(Roth & Yih 04, Sontag 10)
- ➢ The inferences can be solved by any approach

## Amortized Inference Theorem

Given two inference problems P and Q. $c_P$ and $c_Q$ are vectors that represent the coefficients of the objective functions of C and Q, respectively.

$x_P^*$: the solution of P

**Theorem**: If the following conditions are satisfied

1. Same #variables & same constraints (same equivalence class)
2. $\forall i,\ (2x_{p,i}^* - 1)(c_{Q,i} - c_{P,i}) \geq 0$

then the optimal solution of Q is $x_p^*$

- ➢ There are only a few equivalence classes
  - ➢ E.g., in POS tagging task, sentences with the same #words are in the same equivalence class.

## Approximate Amortized Inference

Approximate inference by relaxing the condition.

**Theorem**: If the following conditions are satisfied

1. Same # variables & same constraints (same equivalence class)
2. $\forall i,\ (2x_{p,i}^* - 1)(c_{Q,i} - c_{P,i}) \geq -\epsilon |c_{Q,i}|$

then $x_P^*$ is a $(\frac{1}{1+M\epsilon})$-approximate solution to Q (M: a constant)

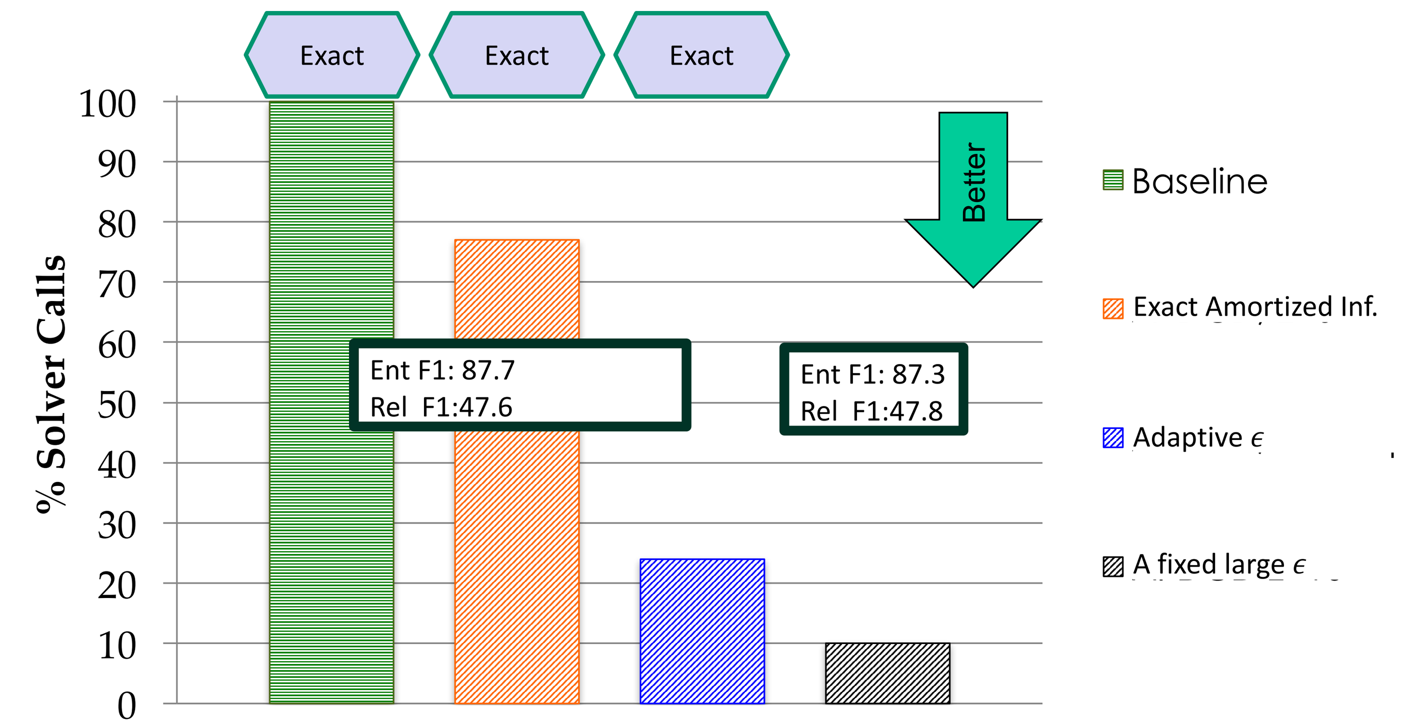
DCD with fixed $\epsilon$

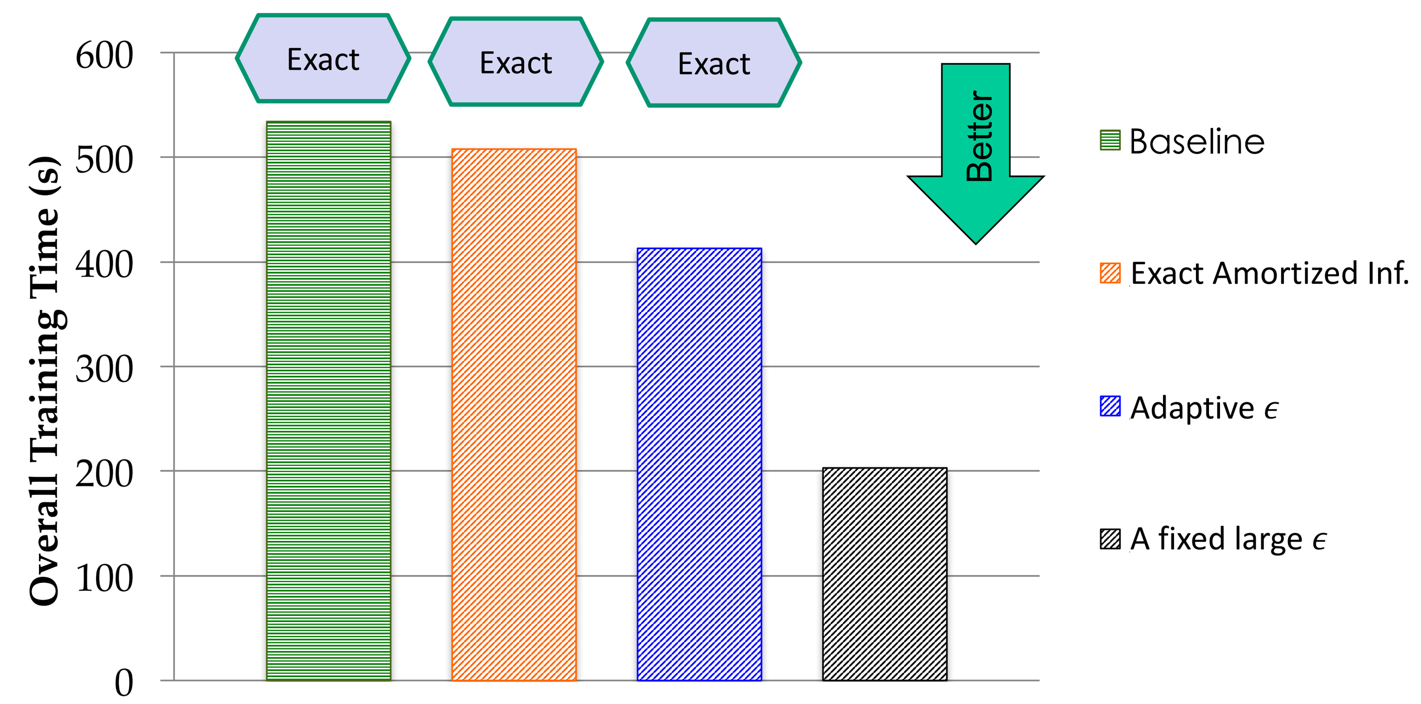
- Inference with an undegenerating approximate method
- DCD algorithm will stop and the empirical risk of the trained model is bounded (related to Finley & Joachims 08)

DCD with adaptive $\epsilon$

- Guarantees to return an exact model

## Performance on Entity-Relation Task




* Srikumar, V.; Kundu, G.; and Roth, D. 2012. On amortizing inference cost for structured prediction. In EMNLP.

** Kundu, G.; Srikumar, V.; and Roth, D. 2013. Margin-based decomposed amortized inference. In ACL.

This research is sponsored by ARL, DARPA, and an ONR award