Learning and Inference in Structured Prediction

Tutorial AAAI

Gourab Kundu

IBM Research

Amortized Inference

Part 3: Amortized Inference

- ***Overview***
- **Amortization at Inference Time:**
  - **Theorems**
  - **Decomposition**
  - **Results**
- **Amortization during Learning:**
  - **Approximate Inference**
  - **Results**

Inference

S1	S2	POS
He	They	PRP
is	are	VBZ
reading	watching	VBG
a	a	DT
book	movie	NN

S1 & S2 look very different but their output structures are the same

The inference outcomes are the same

After inferring the POS structure for S1,
Can we speed up inference for S2 ?
Can we make the k-th inference problem cheaper than the first?

Amortized Inference [Kundu, Srikumar & Roth, EMNLP-12,ACL-13]

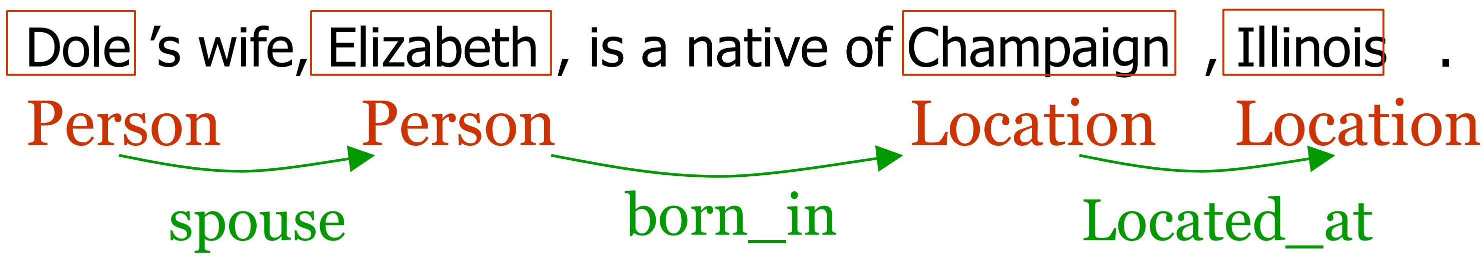
- We formulate the problem of amortized inference: reducing inference time over the lifetime of an NLP tool

- We develop conditions under which the solution of a new, previously unseen problem, can be exactly inferred from earlier solutions without invoking a solver.

- This results in a family of exact inference schemes
  - **Algorithms are invariant to the underlying solver; we simply reduce the number of calls to the solver**

- Significant improvements both in terms of solver calls and wall clock time in several structured prediction tasks

Entity Relation Extraction task



- The goal is to find a consistent assignment of entity types to all entities and relation types to all relations
  - Consistency constraint: A spouse relation can only hold between two person entities and cannot hold between two location entities

ILP Formulation for Entity Relation Task

Person Person Location Location

Dole 's wife, Elizabeth , is a native of Champaign , Illinois

spouse born_in Located_at

Dole

PER	0.5
LOC	0.3
ORG	0.2

Elizabeth

PER	0.6
LOC	0.1
ORG	0.3

Dole-Elizabeth

spouse	0.7
born_in	0.1
Located_at	0.1
No-relation	0.1

ILP Formulation

Dole

PER	0.5	y1
LOC	0.3	y2
ORG	0.2	y3

Elizabeth

PER	0.6	y4
LOC	0.1	y5
ORG	0.3	y6

Dole-Elizabeth

spouse	0.7	y7
born_in	0.1	y8
Located_at	0.1	y9
No-relation	0.1	y10

maximize

$0.5y1 + 0.3y2 + 0.2y3 +$

$0.6y4 + 0.1y5 + 0.3y6 +$

$0.7y7 + 0.1y8 + 0.1y9 + 0.1y10$

subj to yi ε {0,1}

$y1 + y2 + y3 = 1$

$y4 + y5 + y6 = 1$

$y7 + y8 + y9 + y10 = 1$

$2y7-y1-y4<=0$

A spouse relation can only hold between two person entities

Amortized Inference for ILP

- We can write the ILP as

  $\arg\max_{\mathbf{y}} \mathbf{c}\mathbf{y}$

  $A\mathbf{y} \leq \mathbf{b}$

  $\mathbf{y}_i \in \{0,1\}$

- Inference problems discussed in previous sections can be represented as 0-1 ILPs.

maximize

$0.5y1 + 0.3y2 + 0.2y3 +$

$0.6y4 + 0.1y5 + 0.3y6 +$

$0.7y7 + 0.1y8 + 0.1y9 + 0.1y10$

subj to $yi \in \{0,1\}$

$y1 + y2 + y3 = 1$

$y4 + y5 + y6 = 1$

$y7 + y8 + y9 + y10 = 1$

$2y7-y1-y4<=0$

Preliminary (1)

objective vector

P

$$\max\ 2y_1+3y_2+2y_3+1.0y_4$$
$$y_1 + y_2 \le 1$$
$$y_3 + y_4 \le 1$$

$$c_P = \begin{matrix} 2 \\ 3 \\ 2 \\ 1 \end{matrix}$$

Preliminary (2)

P

max $\mathbf{2}y_1+\mathbf{3}y_2+\mathbf{2}y_3+\mathbf{1.0}y_4$

$y_1 + y_2 \leq 1$

$y_3 + y_4 \leq 1$

objective vector

$c_P = \begin{bmatrix} 2 \\ 3 \\ 2 \\ 1 \end{bmatrix}$

optimal solution

$y^*_P = \begin{bmatrix} 0 \\ 1 \\ 1 \\ 0 \end{bmatrix}$

Preliminary (3)

objective vector

optimal solution

P

$$\max \; 2y_1 + 3y_2 + 2y_3 + 1.0y_4$$
$$y_1 + y_2 \le 1$$
$$y_3 + y_4 \le 1$$

$$c_P = \begin{bmatrix} 2 \\ 3 \\ 2 \\ 1 \end{bmatrix} \qquad y^*_P = \begin{bmatrix} 0 \\ 1 \\ 1 \\ 0 \end{bmatrix} \qquad c_P \cdot y^*_P = 5$$

score for optimal solution

Preliminary (4)

- We define an **equivalence class** as the set of ILPs that have:
  - the same number of inference variables
  - the same feasible set

    (same constraints modulo renaming)

P

$$\max\ 2y_1+3y_2+2y_3+y_4$$
$$y_1 + y_2 \leq 1$$
$$y_3 + y_4 \leq 1$$

Constraints are same

Q

$$\max\ 2y_1+4y_2+2y_3+0.5y_4$$
$$y_1 + y_2 \leq 1$$
$$y_3 + y_4 \leq 1$$

# of variables = 4

# of variables = 4

Same equivalence class

Recap: The Recipe

Given:

- A cache of solved ILPs and a new problem

```
If THEOREM_SATISFIED(cache, new problem)
then
     SOLUTION(new problem) = old solution
Else
     Call base solver and update cache
End
```

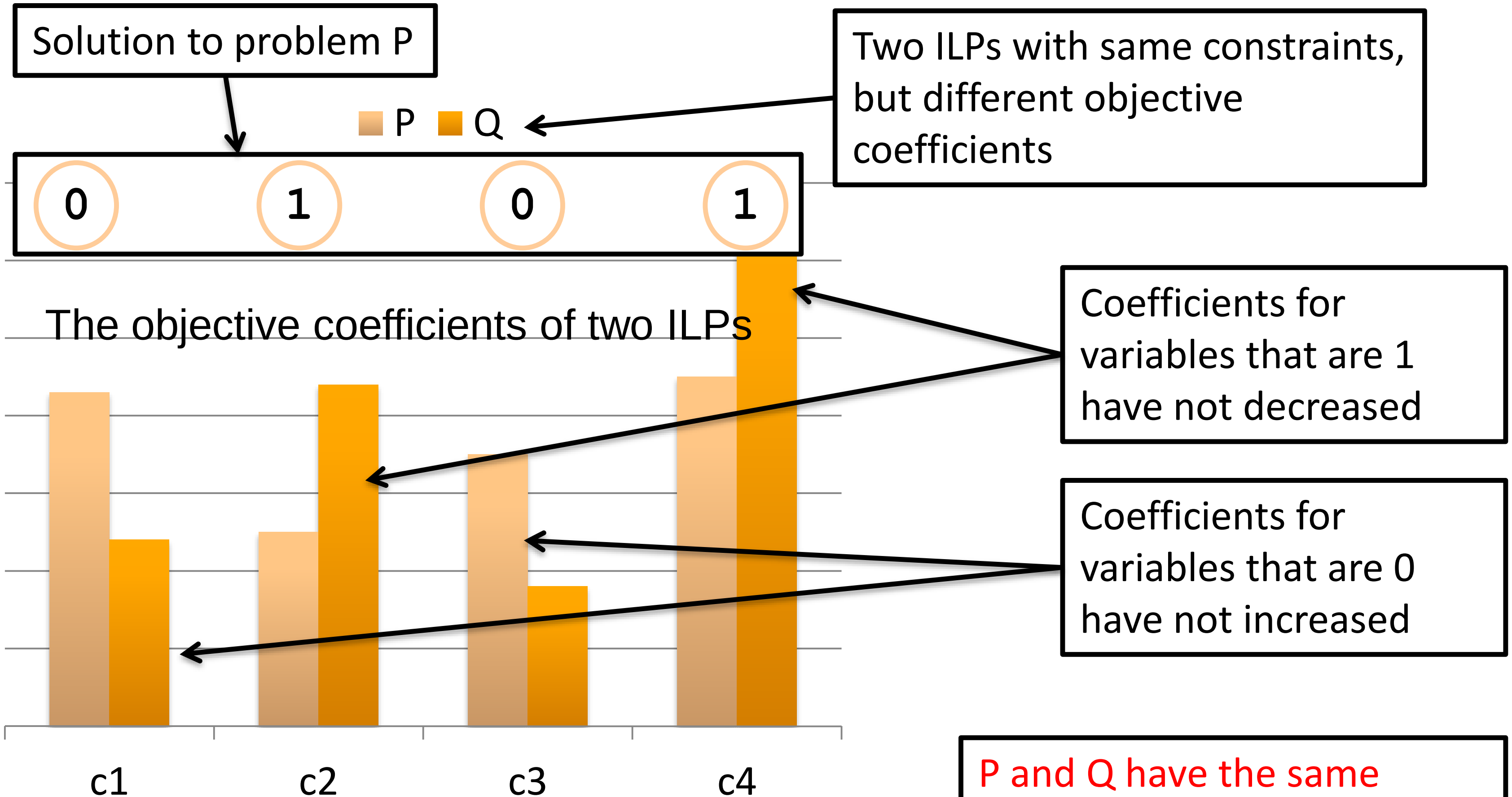
- We will show four different theorems.

Amortized Inference

Part 3: **Amortized Inference**

- ***Overview***
- **Amortization at Inference Time:**
  - **Theorems**
  - **Decomposition**
- **Amortization during Learning:**
  - **Approximate Inference**
  - **Results**

Intuition of Theorem I



P and Q have the same solution

Theorem I

- Denote: $\delta \mathbf{c} = \mathbf{c}_Q - \mathbf{c}_P$

$$\mathbf{y}^*_{P,i} = 0 \Rightarrow \mathbf{c}_{Q,i} \le \mathbf{c}_{P,i} \Rightarrow \delta \mathbf{c}_i \le \mathbf{0} \Rightarrow (2\mathbf{y}^*_{P,i} - 1)\delta \mathbf{c}_i \ge 0$$

$$\mathbf{y}^*_{P,i} = 1 \Rightarrow \mathbf{c}_{Q,i} \ge \mathbf{c}_{P,i} \Rightarrow \delta \mathbf{c}_i \ge \mathbf{0} \Rightarrow (2\mathbf{y}^*_{P,i} - 1)\delta \mathbf{c}_i \ge 0$$

Full Statement of Theorem I

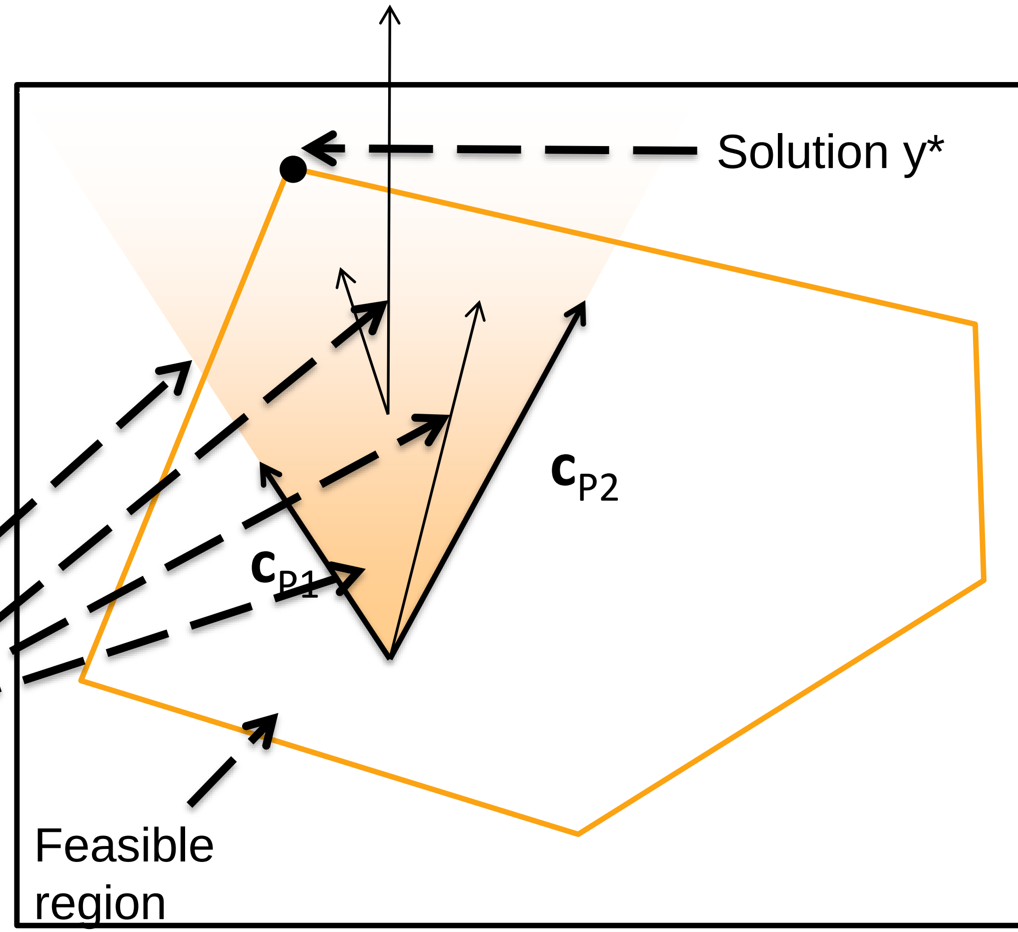
Theorem:

- Let $\mathbf{y}^{*}_{P}$ be the optimal solution of an ILP P. Assume that an ILP Q
  - Is in the same equivalence class as P
  - And, For each i ϵ {1, …, n_p } $(2\mathbf{y}^{*}_{P,i} - 1)\delta\mathbf{c}_i \geq 0$, where $\delta\mathbf{c} = \mathbf{c}_Q - \mathbf{c}_P$
- Then, without solving Q, we can guarantee that the optimal solution of Q is $\mathbf{y}^{*}_{Q} = \mathbf{y}^{*}_{P}$

Intuition of Theorem II (Geometric Interpretation)

All ILPs in the ***cone*** will share the maximizer

ILPs corresponding to all these objective vectors will share the same maximizer *for this feasible region*



Formal Statement of Theorem II

Theorem:

- Assume we have seen m ILP problems $\{P_1, P_2, \ldots, P_m\}$
  - All are in the same equivalence class
  - All have the same optimal solution
- Let ILP Q be a new problem s.t.
  - Q is in the same equivalence class as $P_1, P_2, \ldots, P_m$
  - There exists an $\mathbf{z} \geq \mathbf{0}$ such that $\mathbf{c}_Q = \sum \mathbf{z}_i \mathbf{c}_{P_i}$

- Then, without solving Q, we can guarantee that the optimal solution of Q is $\mathbf{y}^*_Q = \mathbf{y}^*_{P_i}$

Proof of Theorem II

P1

max $\mathbf{c_{P1}} \cdot \mathbf{y}$
subj to $A\mathbf{y} \leq b$

P2

max $\mathbf{c_{P2}} \cdot \mathbf{y}$
subj to $A\mathbf{y} \leq b$

- Let $\mathbf{y}^*$ be the optimal solution of both P1 and P2
  - $\mathbf{c_{P1}} \cdot \mathbf{y^*} \geq \mathbf{c_{P1}} \cdot \mathbf{y'}$ and $\mathbf{c_{P2}} \cdot \mathbf{y^*} \geq \mathbf{c_{P2}} \cdot \mathbf{y'}$
  - $(z_1\mathbf{c_{P1}} + z_2\mathbf{c_{P2}}) \cdot \mathbf{y^*} \geq (z_1\mathbf{c_{P1}} + z_2\mathbf{c_{P2}}) \cdot \mathbf{y'}$ if $z_1, z_2 \geq 0$
- $\mathbf{y^*}$ is optimal for any ILP with objective $(z_1\mathbf{c_{P1}} + z_2\mathbf{c_{P2}})$ with $z_1, z_2 \geq 0$ and same constraint set.

Formal Statement of Theorem III

Theorem:

- Assume we have seen m ILP problems $\{P_1, P_2, \ldots, P_m\}$
  - All are in the same equivalence class
  - All have the same optimal solution
- Let ILP Q be a new problem s.t.
  - Q is in the same equivalence class as $P_1, P_2, \ldots, P_m$
  - There exists an $\mathbf{z} \geq \mathbf{0}$ such that $\delta\mathbf{c} = \mathbf{c}_Q - \sum \mathbf{z}_i\, \mathbf{c}_{Pi}$ and $(2\mathbf{y}^*_{P,i} - 1)\, \delta\mathbf{c}_i \geq 0$
- Then, without solving Q, we can guarantee that the optimal solution of Q is $\mathbf{y}^*_Q = \mathbf{y}^*_{Pi}$

Proof of Theorem III

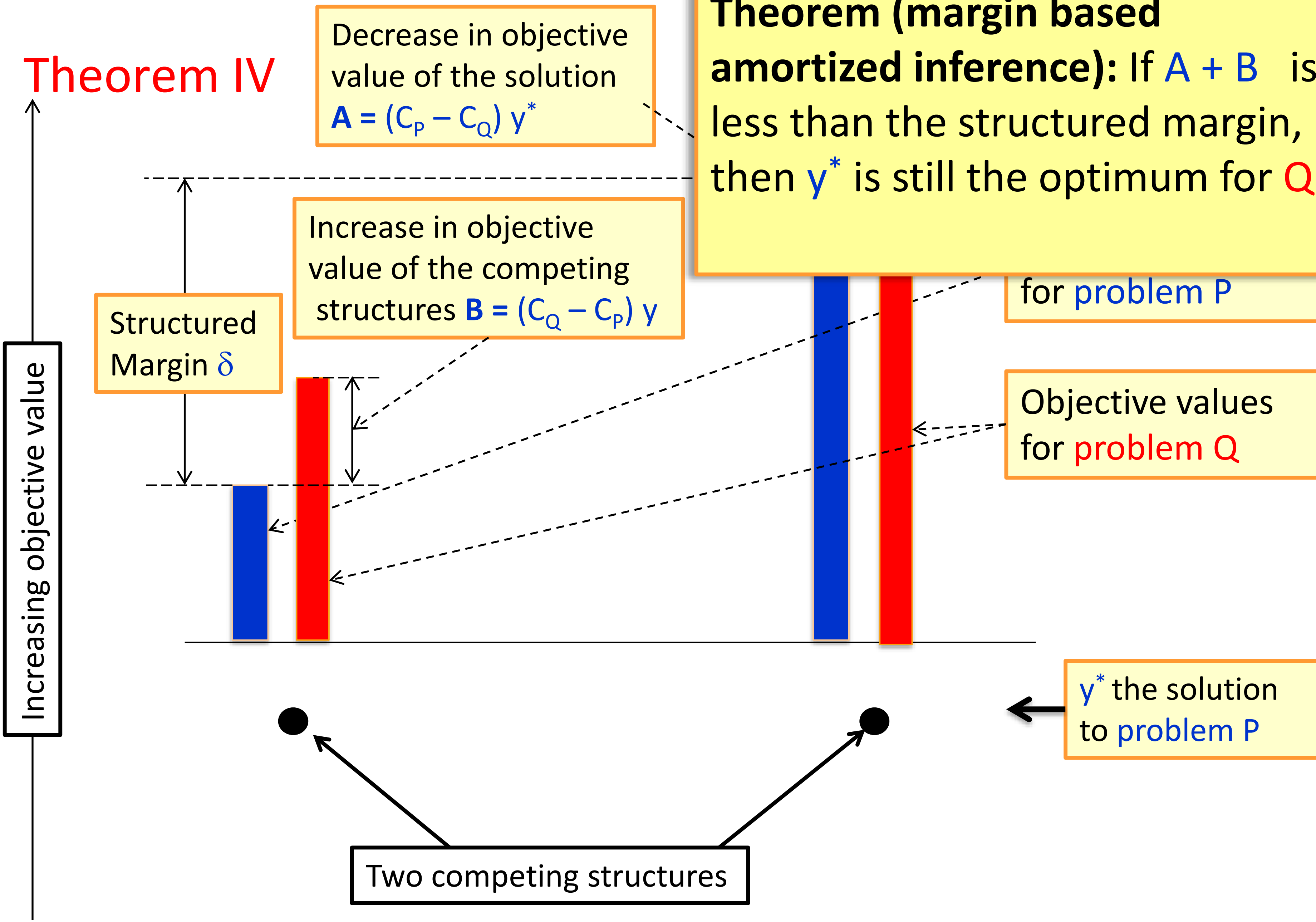
P1	P2	R	Q
max $\mathbf{c}_{P1} \cdot \mathbf{y}$	max $\mathbf{c}_{P2} \cdot \mathbf{y}$	max $\mathbf{c}_{R} \cdot \mathbf{y}$	max $\mathbf{c}_{Q} \cdot \mathbf{y}$
subj to $A\mathbf{y} \leq b$	subj to $A\mathbf{y} \leq b$	subj to $A\mathbf{y} \leq b$	subj to $A\mathbf{y} \leq b$

- Let $\mathbf{y}^*$ be the optimal solution of both P1 and P2
- Theorem II: P1,P2 => R
  - if $\mathbf{c}_R = z_1\mathbf{c}_{P1}+z_2\mathbf{c}_{P2},\ z_1,z_2>=0 \Rightarrow \mathbf{y}^*_R = \mathbf{y}^*_{P1} = \mathbf{y}^*_{P2}$

- Theorem I: R => Q
  - if $(2\mathbf{y}^*_{R,i} - 1)\delta\mathbf{c}_i \geq 0,\ \delta\mathbf{c} = \mathbf{c}_Q - \mathbf{c}_R \Rightarrow \mathbf{y}^*_Q = \mathbf{y}^*_R$
  - if $(2\mathbf{y}^*_{P1,i} - 1)\delta\mathbf{c}_i \geq 0,\ \delta\mathbf{c} = \mathbf{c}_Q - \sum \mathbf{z}_i\, \mathbf{c}_{Pi} \Rightarrow \mathbf{y}^*_Q = \mathbf{y}^*_{P1}$

Theorem IV

Theorem (margin based amortized inference): If $A + B$ is less than the structured margin, then y^* is still the optimum for Q



Formally

P

max $\mathbf{c}_P \cdot \mathbf{y}$

subj to $A_1\mathbf{y} \leq \mathbf{b}_1, A_2\mathbf{y} \leq \mathbf{b}_2$

Q

max $\mathbf{c}_Q \cdot \mathbf{y}$

subj to $A_1\mathbf{y} \leq \mathbf{b}_1, A_2\mathbf{y} \leq \mathbf{b}_2$

- Let $\mathbf{y}^*$ be optimal for P with structured margin δ
  - $\mathbf{c}_P.\mathbf{y}^* \geq \mathbf{c}_P.\mathbf{y} + \delta$ for all $\mathbf{y}$, $A_1\mathbf{y} \leq \mathbf{b}_1, A_2\mathbf{y} \leq \mathbf{b}_2$
- Objective increase for $\mathbf{y}$ from P to Q is $(\mathbf{c}_Q - \mathbf{c}_P).\mathbf{y}$
- Objective decrease for $\mathbf{y}^*$ from P to Q is $(\mathbf{c}_P - \mathbf{c}_Q).\mathbf{y}^*$

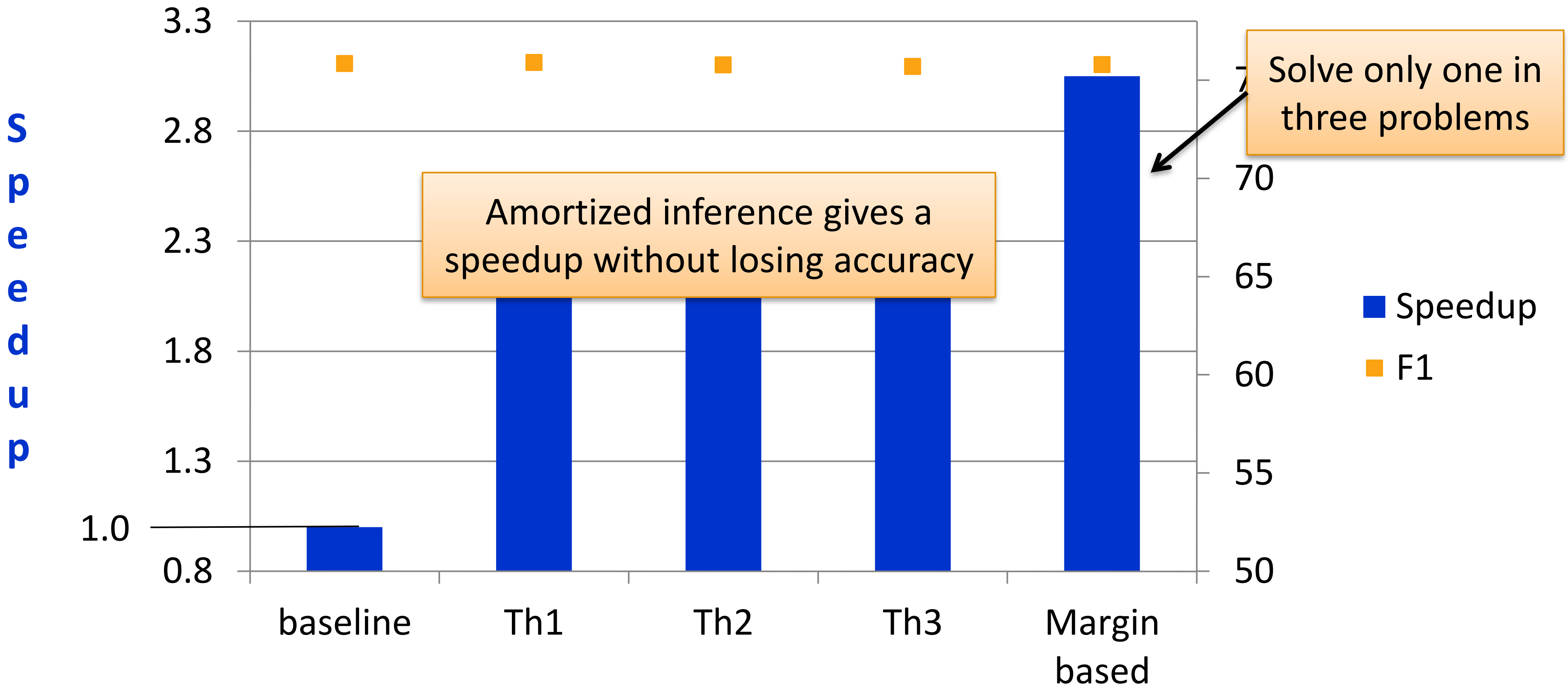
- $\mathbf{y}^*$ is optimal for Q if
  - $(\mathbf{c}_Q - \mathbf{c}_P).\mathbf{y} + (\mathbf{c}_P - \mathbf{c}_Q).\mathbf{y}^* \leq \delta$ for all $\mathbf{y}$, $A_1\mathbf{y} \leq \mathbf{b}_1, A_2\mathbf{y} \leq \mathbf{b}_2$
  - $(\mathbf{c}_Q - \mathbf{c}_P).\mathbf{y} + (\mathbf{c}_P - \mathbf{c}_Q).\mathbf{y}^* \leq \delta$ for all $\mathbf{y}$, $A_1\mathbf{y} \leq \mathbf{b}_1$

Amortized Inference Experiments

- Setup
  - Verb semantic role labeling
    - **Other results also at the end of the section**
  - Speedup & Accuracy are measured over WSJ test set (Section 23)
  - Baseline is solving ILP using Gurobi solver.
- For amortization
  - Cache 250,000 SRL inference problems from Gigaword
  - For each problem in test set, invoke an amortized inference algorithm

Speedup & Accuracy

$$\text{Speedup} = \frac{\text{number of inference calls without amortization}}{\text{number of inference calls with amortization}}$$



Amortization schemes [EMNLP'12, ACL'13]

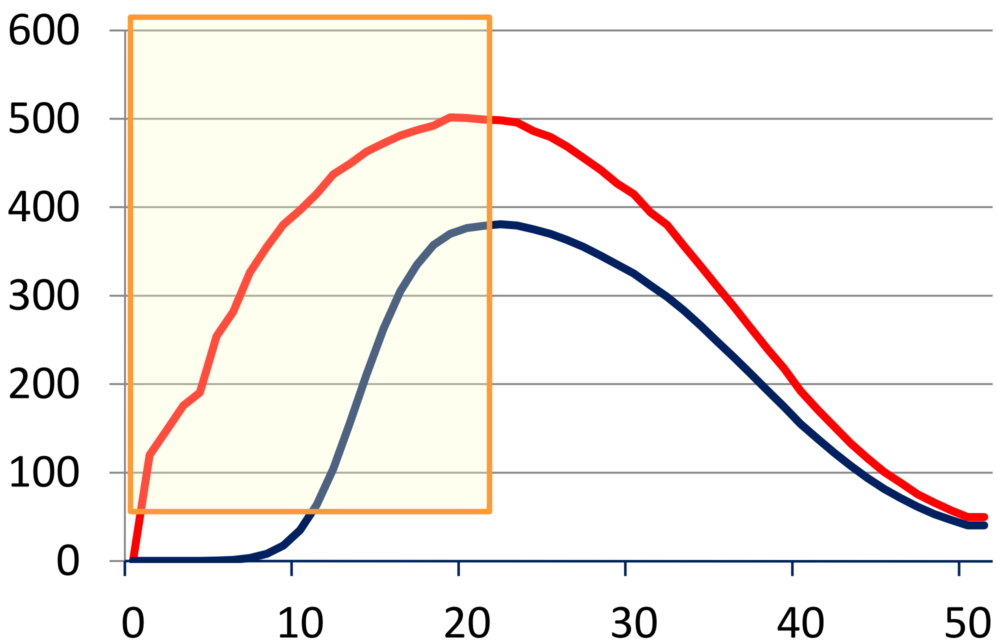
Amortized Inference

Part 3: Amortized Inference

- **Overview**
- ***Amortization at Inference Time:***
  - **Theorems**
  - ***Decomposition***
  - **Results**
- **Amortization during Learning:**
  - **Approximate Inference**
  - **Results**

So far…

- Amortized inference
  - Making inference faster by re-using previous computations
- Techniques for amortized inference
- But these are not useful if the full structure is not redundant!



Smaller Structures are more redundant

Decomposed amortized inference

- Taking advantage of redundancy in components of structures
  - Extend amortization techniques to cases where the full structured output may not be repeated
  - Store partial computations of “components” for use in future inference problems

Entity Relation Extraction task

Dole 's wife, Elizabeth , is a native of Champaign , Illinois .

Person — spouse → Person — born_in → Location — Located_at → Location

maximize

$0.5y1 + 0.3y2 + 0.2y3 +$

$0.6y4 + 0.1y5 + 0.3y6 +$

$0.7y7 + 0.1y8 + 0.1y9 + 0.1y10$

+ additional variable

subj to $yi \in \{0,1\}$

$y1 + y2 + y3 = 1$

$y4 + y5 + y6 = 1$

$y7 + y8 + y9 + y10 = 1$

$2y7-y1-y4<=0$

+ additional constraints

Decomposed inference for ER task

Dole 's wife, Elizabeth , is a native of Champaign , Illinois .

Person Person Location Location

spouse born_in Located_at

Consistent assignment of entity types to first two entities (Dole, Elizabeth) and relation types to relations among these entities

Consistent assignment of entity types to last two entities (Champaign, Illinois) and relation types to relations among these entities

~~Joint constraints~~

Re-introduce constraints using Lagrangian Relaxation

Rush & Collins, *A Tutorial on Dual Decomposition and Lagrangian Relaxation for Inference in Natural Language Processing,*JAIR, 2011.

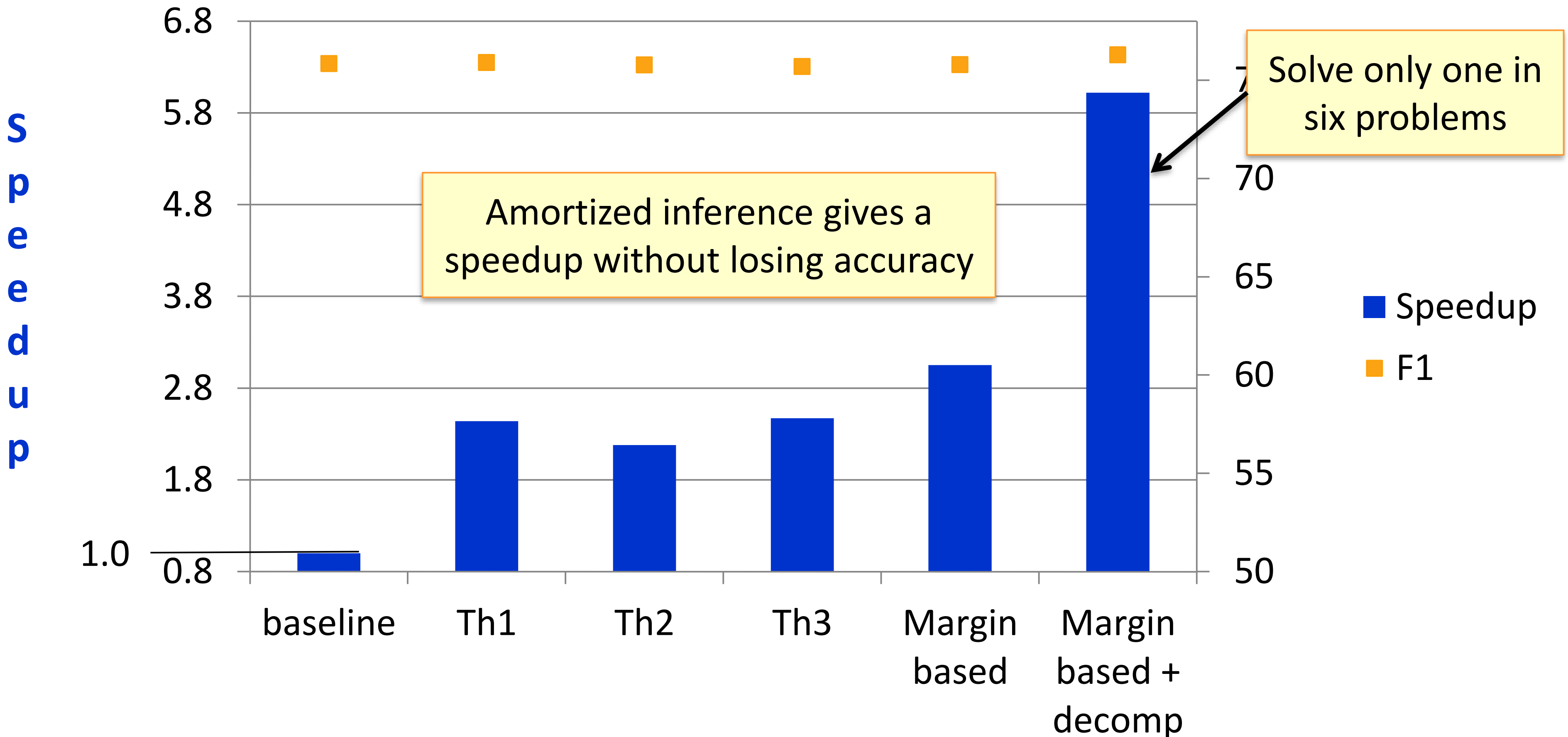
Amortized Inference

Part 3: Amortized Inference

- **Overview**
- ***Amortization at Inference Time:***
  - **Theorems**
  - ***Decomposition***
  - **Results**
- **Amortization during Learning:**
  - **Approximate Inference**
  - **Results**

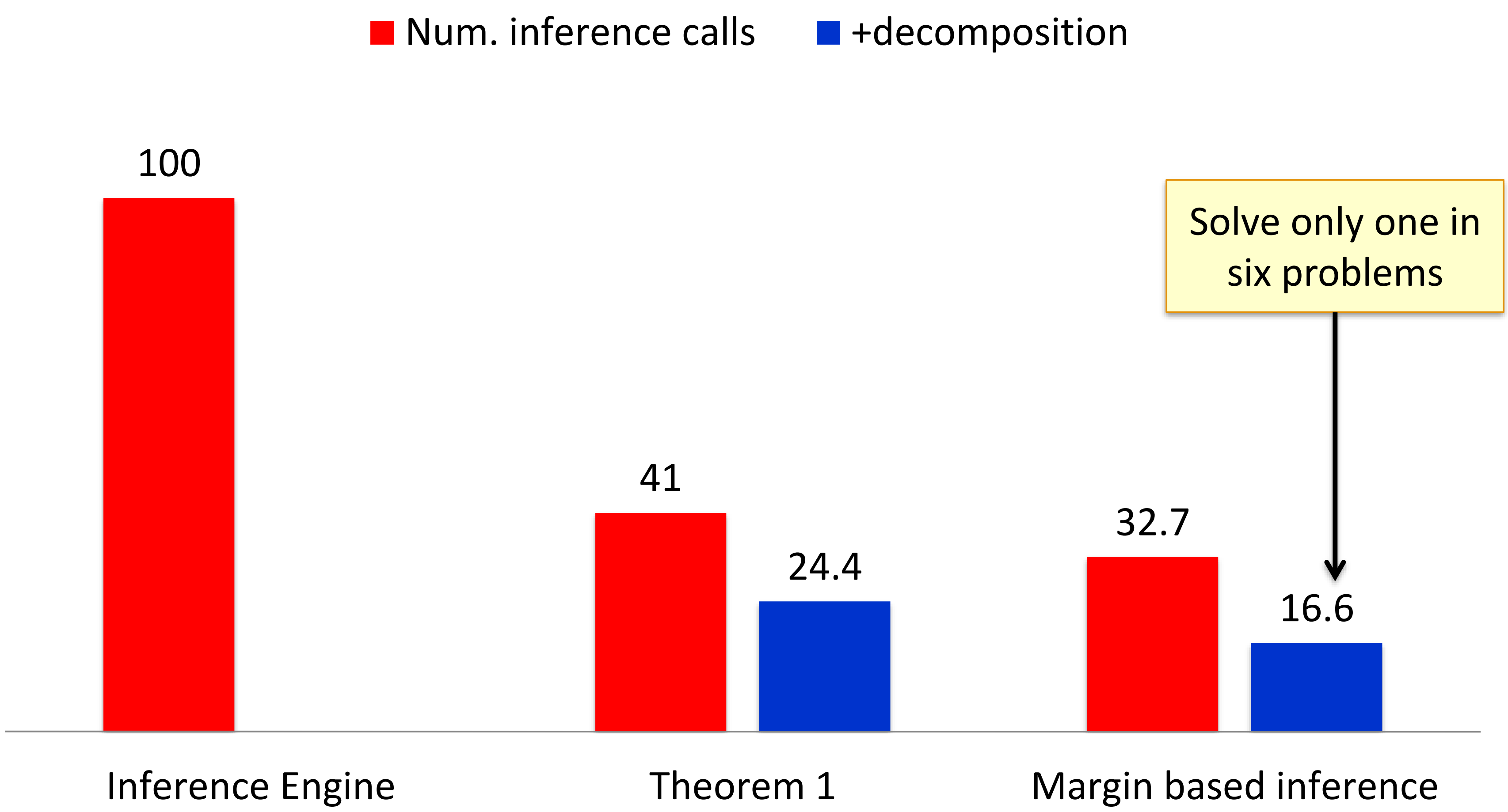
Speedup & Accuracy

$$\text{Speedup} = \frac{\text{number of inference calls without amortization}}{\text{number of inference calls with amortization}}$$

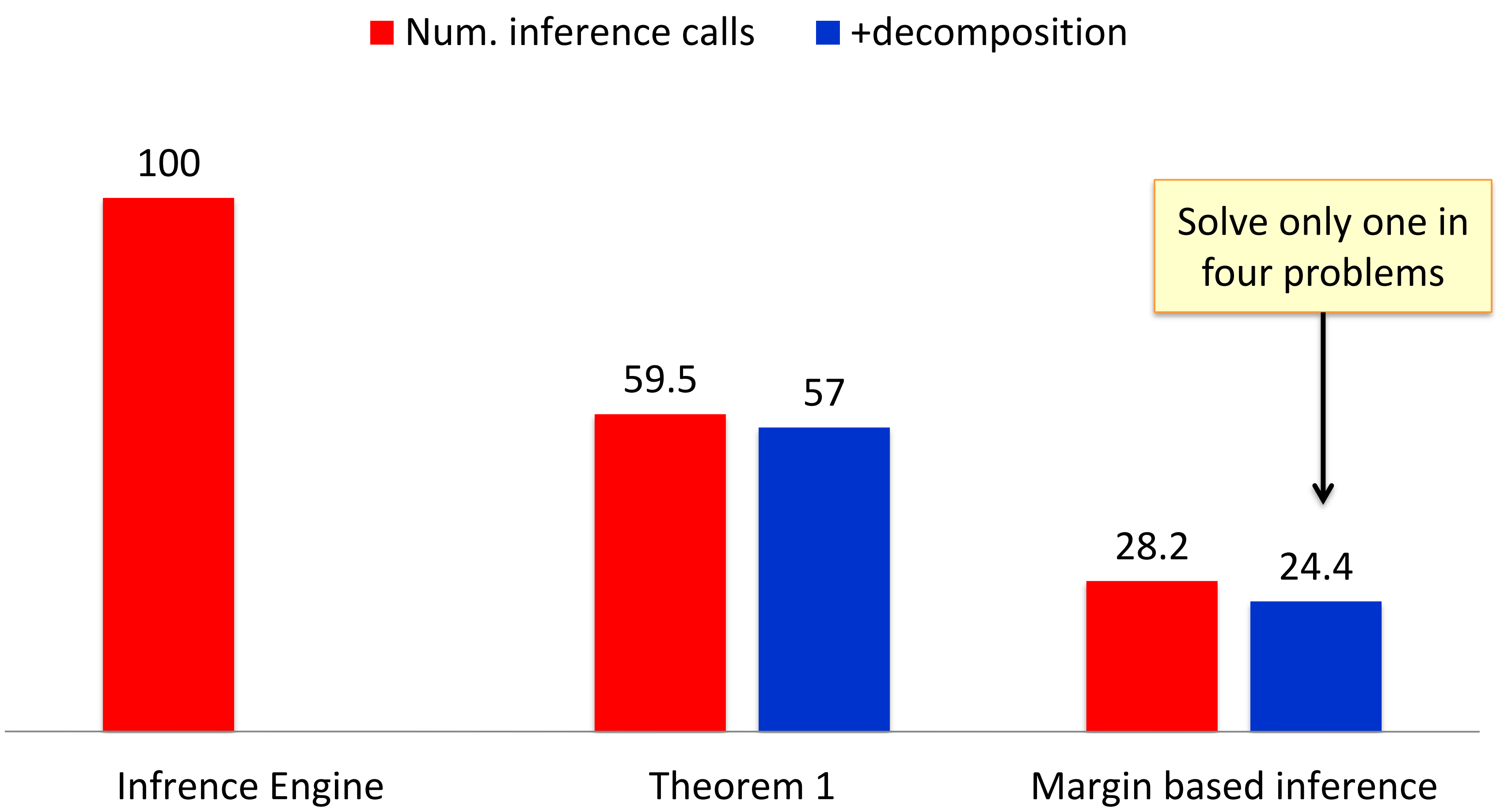


Amortization schemes [EMNLP'12, ACL'13]

Reduction in inference calls (SRL)



Reduction in inference calls (Entity-relation extraction)



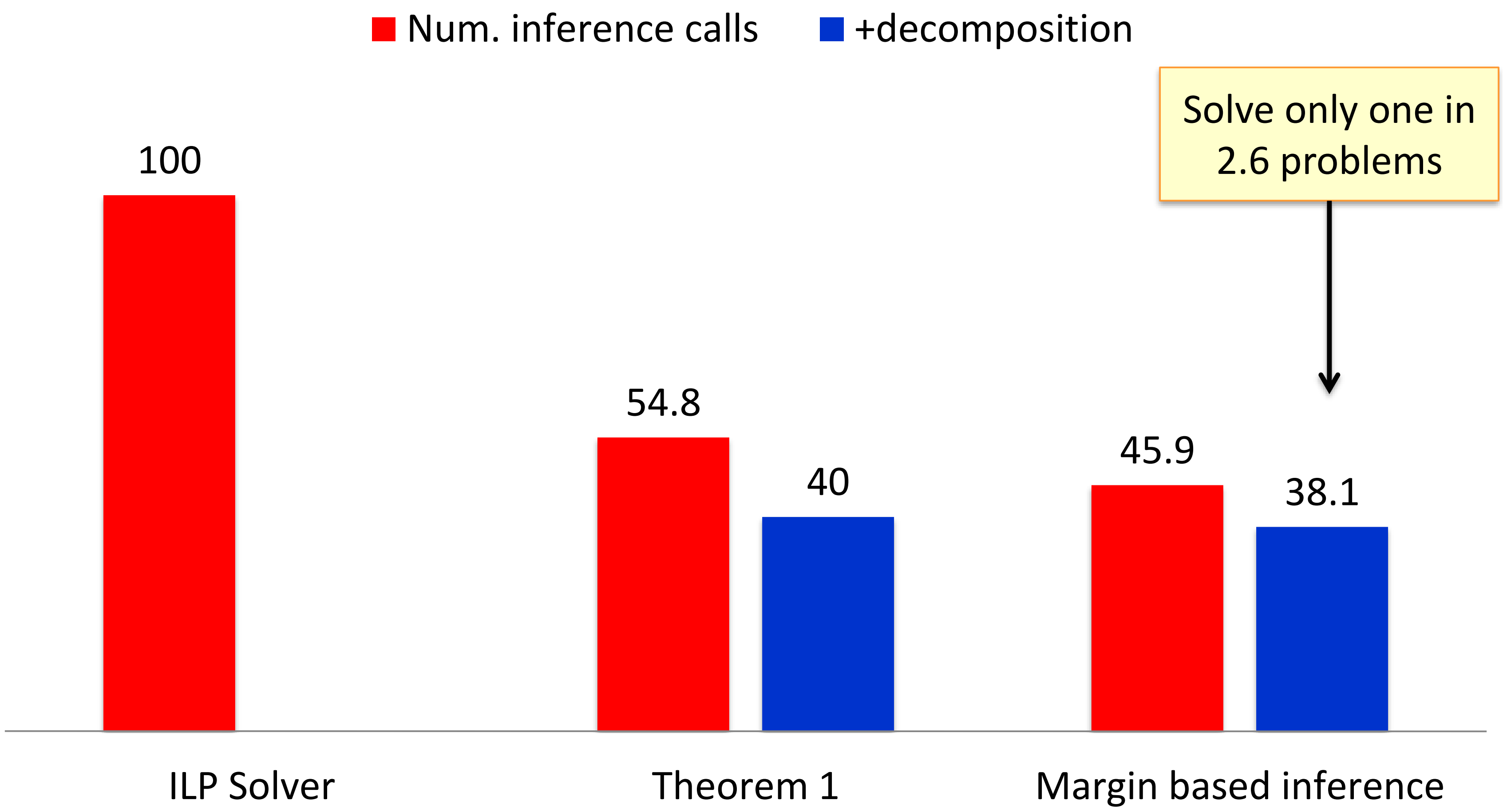
So far…

- We have given theorems that allow savings of 5/6 of the calls to your favorite inference engine.

- But, there is some cost in
  - Checking the conditions of the theorems
  - Accessing the cache

- Our implementations are clearly not state-of-the-art but….

Reduction in wall-clock time (SRL)



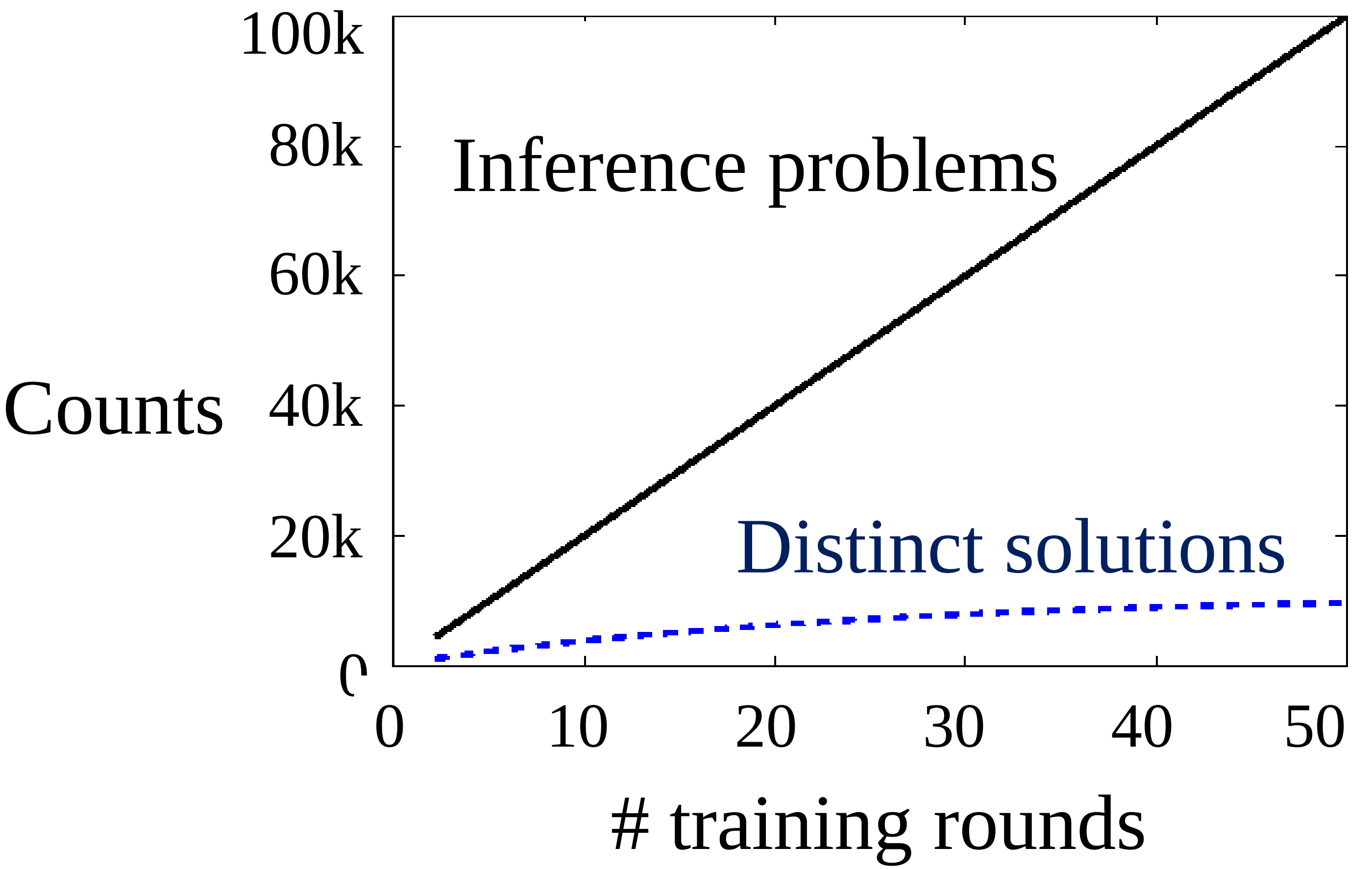
Amortized Inference

Part 3: Amortized Inference

- **Overview**
- **Amortization at Inference Time:**
  - **Theorems**
  - **Decomposition**
  - **Results**
- ***Amortization during Learning:***
  - **Approximate Inference**
  - **Results**

Redundancy in Learning Phase

- [AAAI 15]: Structural Learning with Amortized Inference



Amortization during Learning w/ Theorem I

- We can apply Theorem I to amortize inference calls during learning.

- Recall: Condition of Theorem 1:
  - For each $i \in \{1, \ldots, n_p\}$ $(2\mathbf{y}^*_{P,i} - 1)\delta\mathbf{c}_i \geq 0$, where $\delta\mathbf{c} = \mathbf{c}_Q - \mathbf{c}_P$

- Guarantee of exactness: $\mathbf{y}^*_Q = \mathbf{y}^*_P$

Amortization during Learning w/ Approximate Solution

- Approximate solutions to inference problems can be good enough to guide learning.

- New Condition:
  - For each $i \in \{1, \ldots, n_p\}$ $(2\mathbf{y}^*_{P,i} - 1)\delta\mathbf{c}_i \geq -\varepsilon\ |c_{Q,i}|$, where $\delta\mathbf{c} = \mathbf{c}_Q - \mathbf{c}_P$

- Guarantee of Approximation
  - $\mathbf{y}^*_{\mathbf{P}}$ is a $1 / (1 + M\ \varepsilon)$ approximate solution to Q.

Learning with Approximate Amortized Inference

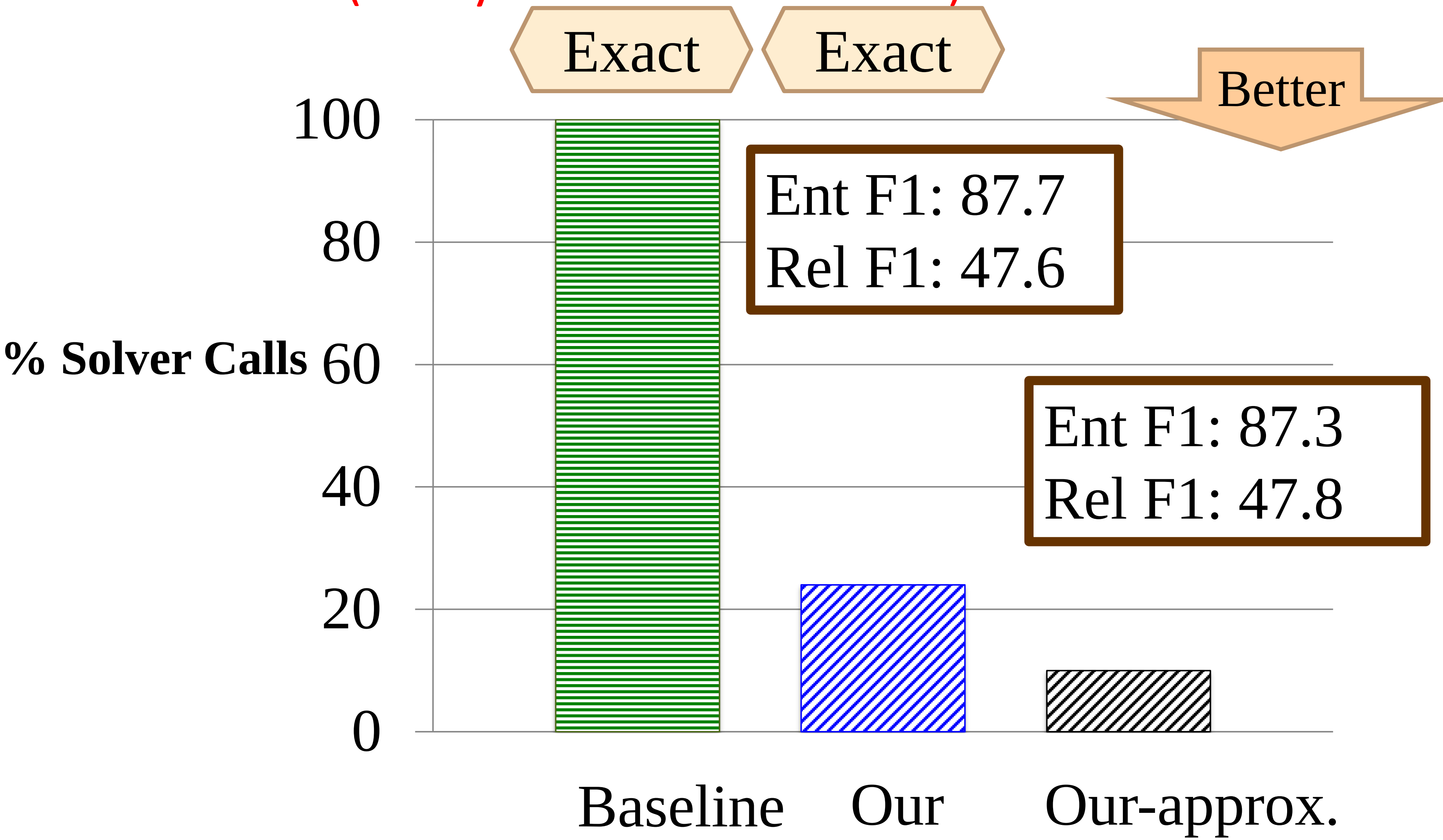
- Learning Structured SVM with approximate amortized inference gives a model with bounded empirical risk
  - Finley, T., and Joachims, T. 2008. *Training structural SVMs when exact inference is intractable*. In ICML 2008
  - our formulation is an under-generating approximation with approximation ratio $1 / (1 + M\varepsilon)$
- Dual coordinate descent for structured SVM can still return an exact model even if approx. amortized inference is used.
  - call exact inference after every τ iterations

Amortized Inference

Part 3: **Amortized Inference**

- **Overview**
- **Amortization at Inference Time:**
  - **Theorems**
  - **Decomposition**
  - **Results**
- ***Amortization during Learning:***
  - **Approximate Inference**
  - **Results**

Solver Calls (Entity-Relation Extraction)



Amortized Inference

Part 3: **Amortized Inference**

- **Amortization at Inference Time:**
  - **Theorems**
  - **Decomposition**
  - **Results**
- **Amortization during Learning:**
  - **Approximate Inference**
  - **Results**