More data means less inference: A pseudo-max approach to structured learning

David Sontag  
Microsoft Research

Ofer Meshi  
Hebrew University

Tommi Jaakkola  
MIT

Amir Globerson  
Hebrew University

## Structured prediction

- **Multi-label prediction:**
  
  $x$: ![Multi-label prediction example](image)

  $y$: ✓ x ✓

- **Parsing of natural language:**
  
  $x$: *John hit the ball*

  $y$: ![Parsing example](image)

- **Protein side-chain placement:**
  
  $x$: *KDLMHNKCYHFFM*

  $y$: ![Protein example](image)
More data means less inference: A pseudo-max approach to structured learning

David Sontag
Microsoft Research

Ofer Meshi
Hebrew University

Tommi Jaakkola
MIT

Amir Globerson
Hebrew University

Structured prediction

- Multi-label prediction:
  \[ x : \]
  \[ y : \checkmark \quad x \quad \checkmark \]

- Each prediction task is specified by a feature function \( f(x, y) \) and a weight vector \( w \).

- Prediction is given by
  \[ y \leftarrow \arg\max_{\hat{y} \in Z} w \cdot f(x, \hat{y}) \]

- Typically decomposes as \( f(x, y) = \sum_c f_c(x, y_c) \), where \( c \) is a small set of variables
More data means less inference: A pseudo-max approach to structured learning

David Sontag  
Microsoft Research

Ofer Meshi  
Hebrew University

Tommi Jaakkola  
MIT

Amir Globerson  
Hebrew University

Learning problem (separable setting)

- Given training data \( \{x_m, y_m\}_{m=1}^M \)

- Assume that there exist “true” parameters \( w \) such that

\[
  y_m \leftarrow \text{argmax} \quad w \cdot f(x_m, \hat{y}) \quad \text{for all } m
\]

- Structured perceptron, stochastic subgradient, cutting-plane, ...

  All repeatedly do prediction during learning - very slow!

- Is there some way to circumvent prediction during learning?

- We give an efficient learning algorithm which, when distribution of training examples is sufficiently “nice”, is asymptotically consistent
More data means less inference: A pseudo-max approach to structured learning

David Sontag
Microsoft Research

Ofer Meshi
Hebrew University

Tommi Jaakkola
MIT

Amir Globerson
Hebrew University

The Pseudo-max Method

- **Exact:** \[ \{ w \cdot f(x^m, y^m) > w \cdot f(x^m, y), \forall m \text{ and } y \neq y^m \} \]

- **Pseudo-max:** \[ \{ w \cdot f(x^m, y^m) > w \cdot f(x^m, y^m_i, y_i), \forall m \text{ and } i, y_i \neq y^m_i \} \]

- Very small number of constraints: \( M \times \#\text{Vars} \times \#\text{Values} \)

- Does this ever work?
  - Yes, under some conditions on \( p(x) \).
  - When \( f \) corresponds to a pairwise Markov random field, these constraints suffice to identify \( w^* \).

- We also show how to apply to non-separable setting

- **Very fast**, and gives good results for multi-label prediction and protein side-chain placement