Spectral Regularization for Max-Margin Sequence Tagging

Ariadna Quattoni ♥ Borja Balle ♦ Xavier Carreras ♥ Amir Globerson ▽

(♥) Universitat Politècnica de Catalunya
now at
Xerox Research Centre Europe

(♦) McGill University

(▽) The Hebrew University of Jerusalem

ICML 2014 — June 2014, Beijing

Supported by: XLike EU Project, VISEN EU Project, ISF Centers of Excellence, NSERC and James McGill Research Fund
Sequence Tagging

output: h l p - x p a t x m x s
input: h i p p o p o t a m u s

Fully Observable Models

Latent-variable Models

- Making predictions is tractable
- Learning is convex
- Performance crucially depends on features

+ Hidden layer provides more expressivity
- Making predictions is not tractable
- Learning is non-convex (this paper)
Learning Structured Predictors with Latent Variables

Desiderata:

- Expressive scoring functions
- Tractable prediction function
- Effective regularizer
- Convex training procedure
Main Idea: Change of Representation + Relaxation

- **Problem Formulation**
  - Scoring functions are Input-Output OOM (generalization of HMM)
  - Piecewise Prediction and Loss Function

- **Solving the Learning Problem**
  - Spectral trick:
    - Optimize over parameters of $f \rightarrow$ optimize low-rank matrix $H$
  - Relax low-rank constraint using nuclear norm of $H$
  - Recover parameters of $f$ from $H$ using the spectral method
Outline

- IO-OOM for Sequence Tagging
- A Convex Formulation for IO-OOM Learning
- Experiments
Scoring Functions Computed by IO-OOM

Latent Score $\theta(x, y, h)$:

$\alpha(h_0) \prod_{t=1}^{T} A_{y_t}^{x_t}(h_{t-1}, h_t) \beta(h_T)$

Scoring Function $F_A(x, y)$:

$\sum_{h} \theta(x, y, h) = \alpha^T A_{y_1}^{x_1} \ldots A_{y_T}^{x_T} \beta$

- Model: $A : \langle \alpha, \beta, \{A_b^a\} \rangle$
- Number of states: $n$
- Initial Weights: $\alpha \in \mathbb{R}^n$
- Final Weights: $\beta \in \mathbb{R}^n$
- Observable Operators $A_b^a \in \mathbb{R}^{n \times n}$

- Expressive Function Family $\rightarrow$ e.g. it includes HMM
- Making Predictions (i.e. maximizing $F_A(x, y)$) $\rightarrow$ NP-hard
Piecewise Prediction and Loss for IO-OOM

Approximation: $F^k_A(x, y)$:

$$
\sum_{t=1}^{T-(k-1)} F_A(x_{t:t+k-1}, y_{t:t+k-1})
$$

Loss $L_k(x, y, F_A)$:

$$
\max_z [F^k_A(x, z) - F^k_A(x, y) + l(y, z)]
$$

- Factor size: $k$
- Sum $k$-grams
- Task loss: $l(y, z)$
  e.g. hamming distance

- Prediction and Loss Function $\rightarrow$ computed in $O(T|Y|^k)$ using the Viterbi Algorithm
Discrete Regularizer for IO-OOM

Learning Problem:
\[
\arg\min_{A \in \mathcal{F}} \sum_{i=1}^{m} L_k(x^i, y^i, F_A) + \tau |A|
\]

- Function class (IO-OOM): $\mathcal{F}$
- Training Examples: $\langle x^i, y^i \rangle$
- Loss Function: $L_k$
- Regularizer $\rightarrow$ number of states: $|A|$
- Trade-off constant: $\tau$

- $k \geq 2 \rightarrow$ Non-convex dependence of $L_k$ on parameters of $A$
- $L_k$ involves polynomials of order $k + 3$
Optimization Strategy

- $L_k$ convex on values of $A$ $\rightarrow$ optimization over $(X \times Y)^k$ values

- Three challenges
  1. Table of values $\rightarrow$ must correspond to valid IO-OOM
  2. Regularizer over table $\rightarrow$ must correspond to $\#$ states of IO-OOM
  3. Recover parameters of $A$ from this table
Optimization Strategy

- $L_k$ convex on values of $A \rightarrow$ optimization over $(X \times Y)^k$ values
- Three challenges
  1. Table of values $\rightarrow$ must correspond to valid IO-OOM
  2. Regularizer over table $\rightarrow$ must correspond to #states of IO-OOM
  3. Recover parameters of $A$ from this table

Solution: the Spectral Trick
IO-OOM and Hankel Matrices

\[ X = \{a, b, c\} \quad Y = \{\diamond, \heartsuit\} \]

Hankel Structure:
- Equality constraints
- Low-rank constraints

Fundamental Theorem:
\( F \) is realized by an \( n \)-state IO-OOM
\( \mathcal{H} \) has rank at most \( n \) for every basis
IO-OOM and Hankel Matrices

\[ X = \{a, b, c\} \quad Y = \{\Diamond, \heartsuit\} \]

|   | \(\epsilon\) | \(a\) | \(b\) | \(c\) | \(a\) | \(a\) | \(a\) | \(b\) | \(a\) | \(b\) | \(a\) | \(c\) | \(a\) | \(a\) | \(b\) | \(a\) | \(b\) | \(a\) | \(c\) | \(a\) | \(a\) | \(b\) | \(c\) | \(a\) | \(a\) | \(b\) | \(c\) |
| \(\epsilon\) | \(\epsilon\) | \(\Diamond\) | \(\Diamond\) | \(\Diamond\) | \(\Diamond\) | \(\Diamond\) | \(\Diamond\) | \(\Diamond\) | \(\Diamond\) | \(\Diamond\) | \(\Diamond\) | \(\Diamond\) | \(\Diamond\) | \(\Diamond\) | \(\Diamond\) | \(\Diamond\) | \(\Diamond\) | \(\Diamond\) | \(\Diamond\) | \(\Diamond\) | \(\Diamond\) | \(\Diamond\) | \(\Diamond\) | \(\Diamond\) | \(\Diamond\) |
| \(e\) | \(\epsilon\) | \(a\) | \(b\) | \(c\) | \(a\) | \(a\) | \(b\) | \(a\) | \(a\) | \(b\) | \(a\) | \(c\) | \(a\) | \(a\) | \(b\) | \(a\) | \(b\) | \(a\) | \(c\) | \(a\) | \(a\) | \(b\) | \(c\) | \(a\) | \(a\) | \(b\) | \(c\) |
| \(\Diamond\) | \(\Diamond\) | \(a\) | \(b\) | \(c\) | \(a\) | \(a\) | \(b\) | \(a\) | \(a\) | \(b\) | \(a\) | \(c\) | \(a\) | \(a\) | \(b\) | \(a\) | \(b\) | \(a\) | \(c\) | \(a\) | \(a\) | \(b\) | \(c\) | \(a\) | \(a\) | \(b\) | \(c\) |
| \(\heartsuit\) | \(\heartsuit\) | \(a\) | \(b\) | \(c\) | \(a\) | \(a\) | \(b\) | \(a\) | \(a\) | \(b\) | \(a\) | \(c\) | \(a\) | \(a\) | \(b\) | \(a\) | \(b\) | \(a\) | \(c\) | \(a\) | \(a\) | \(b\) | \(c\) | \(a\) | \(a\) | \(b\) | \(c\) |

Hankel Structure:

- Equality constraints
- Low-rank constraints

Fundamental Theorem:

\(F_k = 3(c\ a\ a\ b\ c) = F(c\ a\ a) + F(a\ a\ b) + F(a\ b\ c)\)
IO-OOM and Hankel Matrices

\[ X = \{a, b, c\} \quad Y = \{\diamondsuit, \heartsuit\} \]

Hankel Structure:
- Equality constraints
- Low-rank constraints

Fundamental Theorem:
\[ F \text{ is realized by an } n\text{-state IO-OOM} \]
\[ H \text{ has rank at most } n \text{ for every basis} \]
Max-Margin Completion of Hankel Matrices

Optimization with rank regularization:

\[
\arg\min_{H \in \mathbb{H}(P, S)} \sum_{i=1}^{m} L_k(x^i, y^i, H) + \tau \text{ rank}(H)
\]

Convex relaxation:

\[
\arg\min_{H \in \mathbb{H}(P, S)} \sum_{i=1}^{m} L_k(x^i, y^i, H) + \tau \|H\|_*
\]

- Set of Hankel Matrices over some basis: \( \mathbb{H}(P, S) \)
- Rank regularizer: \( \text{rank}(H) \)
- Nuclear norm relaxation: \( \|H\|_* \)

- Optimization almost equivalent \( \rightarrow \) we search over IO-OOM that can be recovered from \( H \in \mathbb{H}(P, S) \)
- Once we solve for \( H \) we can recover parameters using the spectral technique
FOBOS Algorithm: Minimization of $L(H) + \tau \|H\|^*$

- Initialize: $H_0 = 0$
- while $t \leq \text{MaxIter}$ do
  - Set $G_t$ to a subgradient of $L(H)$ at $H_t$
  - Set $H_{t+0.5} = H_t - \frac{c}{\sqrt{t}} G_t$
  - Calculate the SVD of $H_{t+0.5} = U\Sigma V^\top$
  - Define a diagonal matrix $\Sigma'$ such that $\sigma'_i = \max[\sigma_i - \nu_t \tau, 0]$
  - set $H_{t+1} = U\Sigma' V^\top$
- end while
Spectral Recovery
using the method by (Hsu et al. 2009)

- **Spectral Algorithm for IO-OOM**
  - Assume $F$ is realized by a minimal $n$-state IO-OOM $A$
  - We are given a basis $(P, S)$ such that $H$ has rank $n$
  - We are given corresponding $H_a$
  - To recover parameters of $A$:
    - Perform SVD to get $H = UV^T$
    - Define $A^a_b = (HV)^+H_b V$

- Typical spectral algorithms assume that we can estimate $H$
- In contrast, we regard $H$ as an optimization variable in a loss minimization procedure
Experiments

- Task: Phonetic Transcription (UCI "Nettalk" Dataset)

@ p - L - h l p - x p a t x m x s
a p p l e h i p p o p o t a m u s

— Regularization Path —

— Learning Curve —
Conclusion

- Convex formulation for learning structured prediction models with latent variables and max-sum predictions
- The spectral trick seen as a linearization
  Polynomial optimization
  \[\rightarrow\]
  linear optimization over low-rank Hankel matrices
- Generalizable to other losses and structured prediction settings
- Take-home message: Fundamental ideas behind spectral learning have a wide range of applicability for structured prediction
Conclusion

- Convex formulation for learning structured prediction models with latent variables and max-sum predictions
- The spectral trick seen as a linearization
  Polynomial optimization
  \[\rightarrow\]
  linear optimization over low-rank Hankel matrices
- Generalizable to other losses and structured prediction settings
- Take-home message: Fundamental ideas behind spectral learning have a wide range of applicability for structured prediction

For more information:
- Come tonight to our poster S63
- On Wednesday, Workshop on Method of Moments and Spectral Learning