Local Loss Optimization in Operator Models: A New Insight into Spectral Learning

Borja Balle, Ariadna Quattoni, Xavier Carreras

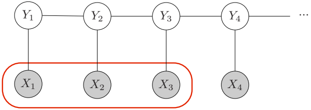
LARCA. Laboratory for Relational Algorithmics, Complexity and Learning
UNIVERSITAT POLITÈCNICA DE CATALUNYA

ICML 2012
June 2012, Edinburgh

This work is partially supported by the PASCAL2 Network and a Google Research Award

A Simple Spectral Method [HKZ09]

Discrete Homogeneous Hidden Markov Model



- n states – $Y_t \in \{1, \ldots, n\}$
- k symbols – $X_t \in \{\sigma_1, \ldots, \sigma_k\}$
- for now assume $n \leqslant k$
- Forward-backward equations with $A_\sigma \in \mathbb{R}^{n \times n}$:
$$\mathbb{P}[X_{1:t} = w] = \alpha_1^\top A_{w_1} \cdots A_{w_t} \vec{1}$$

- Probabilities arranged into matrices $H, H_{\sigma_1}, \ldots, H_{\sigma_k} \in \mathbb{R}^{k \times k}$
$$H(i,j) = \mathbb{P}[X_1 = \sigma_i,\ X_2 = \sigma_j]$$
$$H_\sigma(i,j) = \mathbb{P}[X_1 = \sigma_i,\ X_2 = \sigma,\ X_3 = \sigma_j]$$

- Spectral learning algorithm for $B_\sigma = QA_\sigma Q^{-1}$:
  1. Compute SVD $H = UDV^\top$ and take top n right singular vectors V_n
  2. $B_\sigma = (HV_n)^+ H_\sigma V_n$

(For simplicity, in this talk we ignore learning of initial and final vectors)

A Local Approach to Learning?

- *Maximum likelihood* uses the whole of the sample $S = \{w^1, \ldots, w^N\}$ and is *always consistent* in the realizable case

$$\max_{\alpha_1, \{A_\sigma\}} \frac{1}{N} \sum_{i=1}^{N} \log(\alpha_1^\top A_{w_1^i} \cdots A_{w_{t_i}^i} \vec{1})$$

- The *spectral method* only uses local information from the sample in $\widehat{H}, \widehat{H}_a, \widehat{H}_b$ and its consistency depends on properties of H

$$S = \{abbabba, aabaa, baaabbbabab, bbaaba, babab babbaaaba, abbb, \ldots\}$$

Questions

- Is the spectral method minimizing a "local" loss function?
- When does this minimization yield a consistent algorithm?

Outline

Spectral Learning as Local Loss Optimization

A Convex Relaxation of the Local Loss

Choosing a Consistent Local Loss

Loss Function of the Spectral Method

- Both ingredients in the spectral method have optimization interpretations

$$\text{SVD} \quad — \quad \min_{V_n^\top V_n = I} \|H V_n V_n^\top - H\|_F$$

$$\text{Pseudo-inverse} \quad — \quad \min_{B_\sigma} \|H V_n B_\sigma - H_\sigma V_n\|_F$$

- Can formulate a *joint optimization* for the spectral method

$$\min_{\{B_\sigma\}, V_n^\top V_n = I} \sum_{\sigma \in \Sigma} \|H V_n B_\sigma - H_\sigma V_n\|_F^2$$

Properties of the Spectral Optimization

$$\min_{\{B_\sigma\}, V_n^\top V_n = I} \sum_{\sigma \in \Sigma} \|H V_n B_\sigma - H_\sigma V_n\|_F^2$$

- Theorem The optimization is *consistent* under the same conditions of the spectral method
- The loss is *non-convex* due to $V_n B_\sigma$ and constraint $V_n^\top V_n = I$
- Spectral method equivalent to
  1. Choosing V_n using SVD
  2. Optimizing $\{B_\sigma\}$ with fixed V_n

Intuition about the Loss Function

- Minimize the ℓ_2 norm of the *unexplained (finite set of) futures* when a symbol σ is generated and the transition is explained using B_σ (*over a finite set of pasts*)
- Strongly based on the *markovianity* of the process – which generic ML does not exploit

A Convex Relaxation of the Local Loss

- For algorithmic purposes a *convex local loss* function is more desirable
- A relaxation can be obtained by *replacing* the projection V_n with a *regularization* term

$$\min_{\{B_\sigma\}, V_n^\top V_n = I} \sum_{\sigma \in \Sigma} \|H V_n B_\sigma - H_\sigma V_n\|_F^2$$

$$\downarrow$$

1. fix $n = |\mathcal{S}|$ and take $V_n = I$
2. $B_\Sigma = [B_{\sigma_1} | \cdots | B_{\sigma_k}]$ and $H_\Sigma = [H_{\sigma_1} | \cdots | H_{\sigma_k}]$
3. regularize via nuclear norm to *emulate* V_n

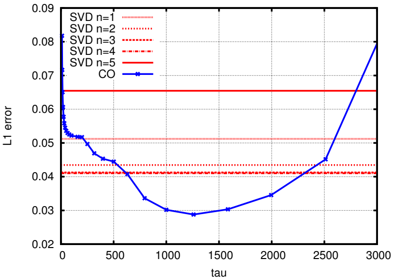
$$\min_{B_\Sigma} \|H B_\Sigma - H_\Sigma\|_F^2 + \tau \|B_\Sigma\|_*$$

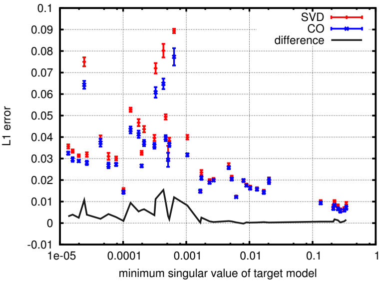
- This optimization is *convex* and has some interesting theoretical (see paper) and empirical properties

Experimental Results with the Convex Local Loss

Performing experiments with synthetic targets the following facts are observed

- Tuning the regularization parameter τ a better trade-off between generalization and model complexity can be achieved
- The largest gains when using the convex relaxation are attained for targets suposedly hard to the spectral method





The Hankel Matrix

For any function $f : \Sigma^\star \to \mathbb{R}$ its *Hankel matrix* $H_f \in \mathbb{R}^{\Sigma^\star \times \Sigma^\star}$ is defined as $H_f(p, s) = f(p \cdot s)$

$\Sigma^\star$	λ	a	b	aa	ab	...
λ	1	0.3	0.7	0.05	0.25	...
a	0.3	0.05	0.25	0.02	0.03	...
b	0.7	0.6	0.1	0.03	0.2	...
aa	0.05	0.02	0.03	0.017	0.003	...
ab	0.25	0.23	0.02	0.11	0.12	...
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$

H $\quad$ H_a

- Blocks defined by sets of rows (prefixes $\mathcal{P}$) and columns (suffixes $\mathcal{S}$)
- Can *parametrize* the spectral method by $\mathcal{P}$ and $\mathcal{S}$ taking $H \in \mathbb{R}^{\mathcal{P} \times \mathcal{S}}$
- Each pair $(\mathcal{P}, \mathcal{S})$ defines a different *local loss* function

Consistency of the Local Loss

Theorem (Schützenberger '61) $\mathrm{rank}(H_f) = n$ iff f can be computed with operators $A_\sigma \in \mathbb{R}^{n \times n}$

Consequences

- The spectral method is consistent iff $\mathrm{rank}(H) = \mathrm{rank}(H_f) = n$
- There always exist $|\mathcal{P}| = |\mathcal{S}| = n$ with $\mathrm{rank}(H) = n$

Trade-off

- Larger $\mathcal{P}$ and $\mathcal{S}$ more likely to have $\mathrm{rank}(H) = n$, but also require larger samples for good estimation $\widehat{H}$

Question

- Given a sample, how to choose *good* $\mathcal{P}$ and $\mathcal{S}$?

Answer

- Random sampling succeeds w.h.p. with $|\mathcal{P}|$ and $|\mathcal{S}|$ depending polynomially on the complexity of the target

Visit us at poster 53

Local Loss Optimization in Operator Models: A New Insight into Spectral Learning

Borja Balle, Ariadna Quattoni, Xavier Carreras

LARCA. Laboratory for Relational Algorithmics, Complexity and Learning
UNIVERSITAT POLITÈCNICA DE CATALUNYA

ICML 2012
June 2012, Edinburgh

This work is partially supported by the PASCAL2 Network and a Google Research Award