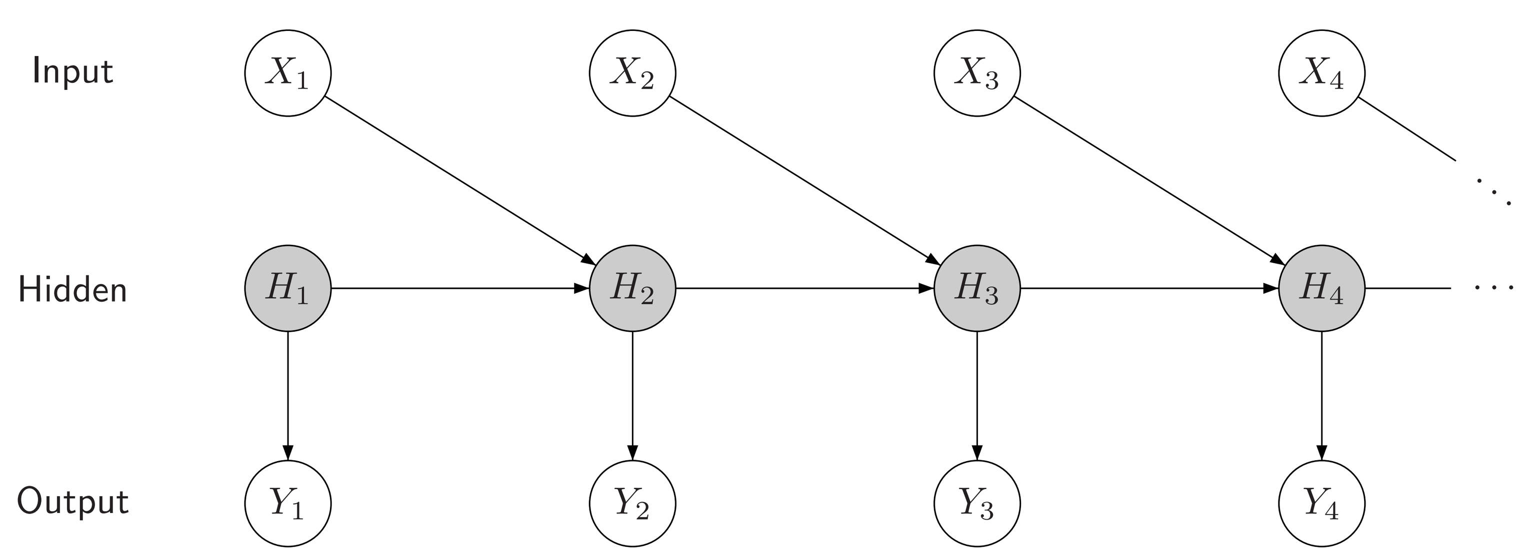
A Spectral Learning Algorithm for Finite State Transducers

Borja Balle, Ariadna Quattoni, Xavier Carreras
Universitat Politècnica de Catalunya

Summary

- FSTs model input-output relations with *hidden states*



- Main contribution: a spectral learning algorithm for FSTs
(Chang '96, Mossel-Roch '05, Hsu et al. '09, Siddiqi et al. '10)
- Key concept: Represent transition and emission structure with *Observable Operator Models*
- Advantages: fast and scalable, strong guarantees, beats EM

Observator Operator Models for FST

$\mathcal{X} = \{a_1, \dots, a_k\}$, $\mathcal{Y} = \{b_1, \dots, b_l\}$, $\mathcal{H} = \{c_1, \dots, c_m\}$
Given $(x, y) \in (\mathcal{X} \times \mathcal{Y})^t$, model computes a *conditional probability* as

$$\Pr[\, y \mid x \,] = 1^\top A_{x_t}^{y_t} \cdots A_{x_1}^{y_1} \alpha$$

$$A_a^b = T_a D_b \in \mathbb{R}^{m \times m} \quad \text{(factorized operator)}$$
$$T_a(i,j) = \Pr[H_s = c_i | X_{s-1} = a, H_{s-1} = c_j] \in \mathbb{R}^{m \times m} \quad \text{(state transition)}$$
$$D_b(i,j) = \delta_{i,j} \Pr[Y_s = b | H_s = c_j] \in \mathbb{R}^{m \times m} \quad \text{(observation emission)}$$
$$O(i,j) = \Pr[Y_s = b_i | H_s = c_j] \in \mathbb{R}^{l \times m} \quad \text{(collected emissions)}$$
$$\alpha(i) = \Pr[H_1 = c_i] \in \mathbb{R}^m \quad \text{(initial probabilites)}$$

Choice of operator A_a^b depends only on *observable* symbols ...

... but operator *parameters* are conditioned by *hidden* states

Learnable Set of Observable Operators

Idea

(*subspace identification methods for linear systems, '80s*)

Find a basis for the state space such that operators in the new basis are related to observable quantities

Find a basis Q where operators can be expressed in terms of unigram, bigram and trigram probabilities

$$\rho(i) = \Pr[Y_1 = b_i] \in \mathbb{R}^l$$
$$P(i,j) = \Pr[Y_1 = b_j, Y_2 = b_i] \in \mathbb{R}^{l \times l}$$
$$P_a^b(i,j) = \Pr[Y_1 = b_j, Y_2 = b, Y_3 = b_i | X_2 = a] \in \mathbb{R}^{l \times l}$$

Theorem (ρ, P and P_a^b are sufficient statistics)

Let $P = U \Sigma V^*$ be a thin SVD decomposition, then $Q = U^\top O$ yields (under certain assumptions)

$$Q\,\alpha = U^\top \rho$$
$$1^\top Q^{-1} = \rho^\top (U^\top P)^+$$
$$Q A_a^b Q^{-1} = (U^\top P_a^b)(U^\top P)^+$$

Spectral Learning Algorithm

Input: number of states m and sample $S = \{(x^1, y^1), \dots, (x^n, y^n)\}$

1. Compute unigram $\widehat{\rho}$, bigram $\widehat{P}$ and trigram $\widehat{P}_a^b$ relative frequencies in S
2. Perform SVD on $\widehat{P}$ and take $\widehat{U}$ with top m left singular vectors
3. Compute operators using matrix operations on $\widehat{\rho}$, $\widehat{P}$, $\widehat{P}_a^b$ and $\widehat{U}$

Time complexity: $O(n + |\mathcal{Y}|^3)$

PAC-style Result

- X random variable over $\mathcal{X}^*$ with $\lambda = \mathrm{E}[|X|]$, $\mu = \min_a \Pr[X_1 = a]$
- Y random variable over $\mathcal{Y}^*$ whose distribution conditioned on X is given by an FST with m states
- Sampling i.i.d. from (X, Y)

Theorem

For any $0 < \varepsilon, \delta < 1$, if the algorithm receives a sample of size

$$n \geq O\left(\frac{\lambda^2 m |\mathcal{Y}|}{\varepsilon^4 \mu \sigma_O^2 \sigma_P^4} \log \frac{|\mathcal{X}|}{\delta} \right) ,$$

(σ_O and σ_P are m-th singular values of O and P in target)

then with probability at least $1 - \delta$ the hypothesis $\widehat{\Pr}$ satisfies

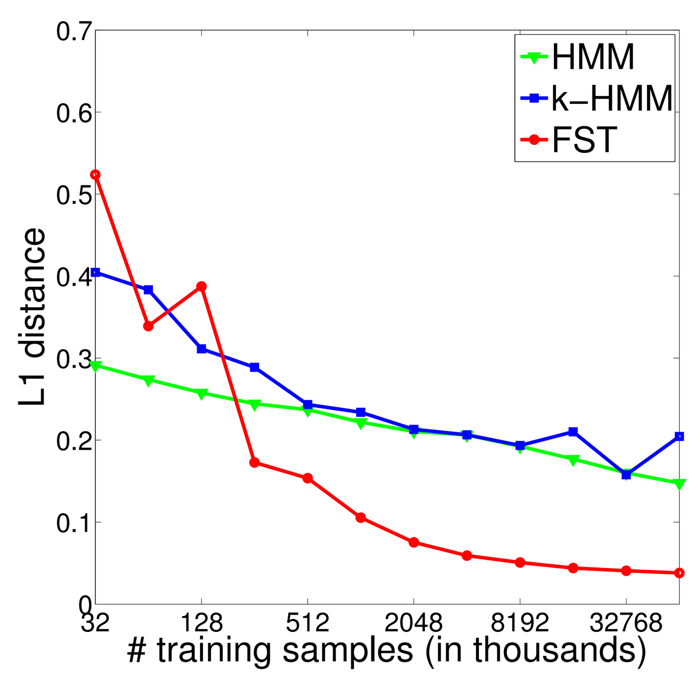
$$\mathrm{E}\left[\sum_{y \in \mathcal{Y}^*} \left| \Pr[y|X] - \widehat{\Pr}[y|X] \right| \right] \leq \varepsilon \ .$$

(L_1 distance between joint distributions $D_{X,Y}$ and $D_{X,\hat{Y}}$)

Synthetic Experiments

Goal: Compare against baselines when learning hypothesis hold

Target: Randomly generated with $|\mathcal{X}| = 3$, $|\mathcal{Y}| = 3$, $|\mathcal{H}| = 2$

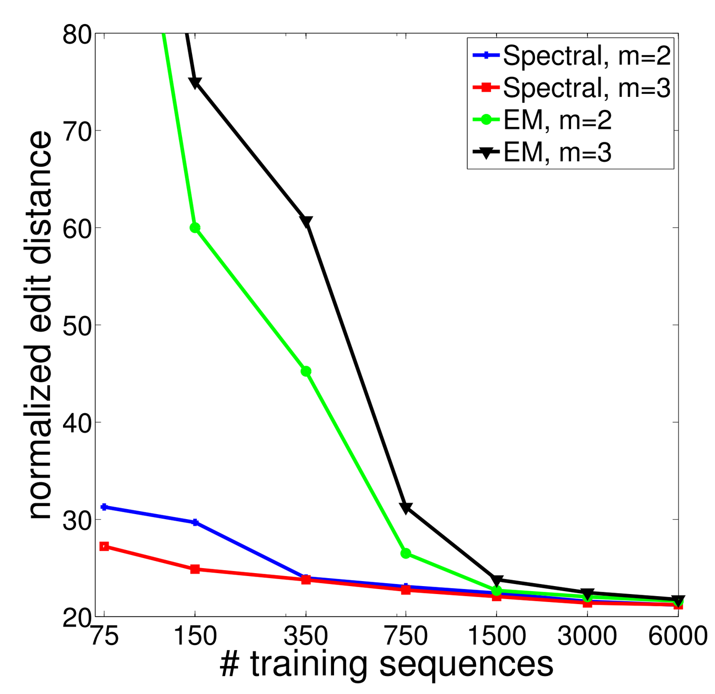


- HMM: model input-output jointly
- k-HMM: one model for each input symbol
- Results averaged over 5 runs

Transliteration Experiments

Goal: Compare against EM in a real task (where modeling assumptions fail)

Task: English to Russian transliteration (brooklyn → бруклин)



Training times	
Spectral	26 s
EM (iteration)	37 s
EM (best)	1133 s

- *Alignment* and *inference* dealt with standard techniques
- Test size: 943, $|\mathcal{X}| = 82$, $|\mathcal{Y}| = 34$

Acknowledgements: This work was partially supported by the EU PASCAL2 Network of Excellence (FP7-ICT-216886), and by a Google Research Award. Borja Balle was supported by an FPU fellowship (AP2008-02064) of the Spanish Ministry of Education. The Spanish Ministry of Science and Innovation supported Ariadna Quattoni (JCI-2009-04240) and Xavier Carreras (RYC-2008-02223 and "KNOW2" TIN2009-14715-C04-04).