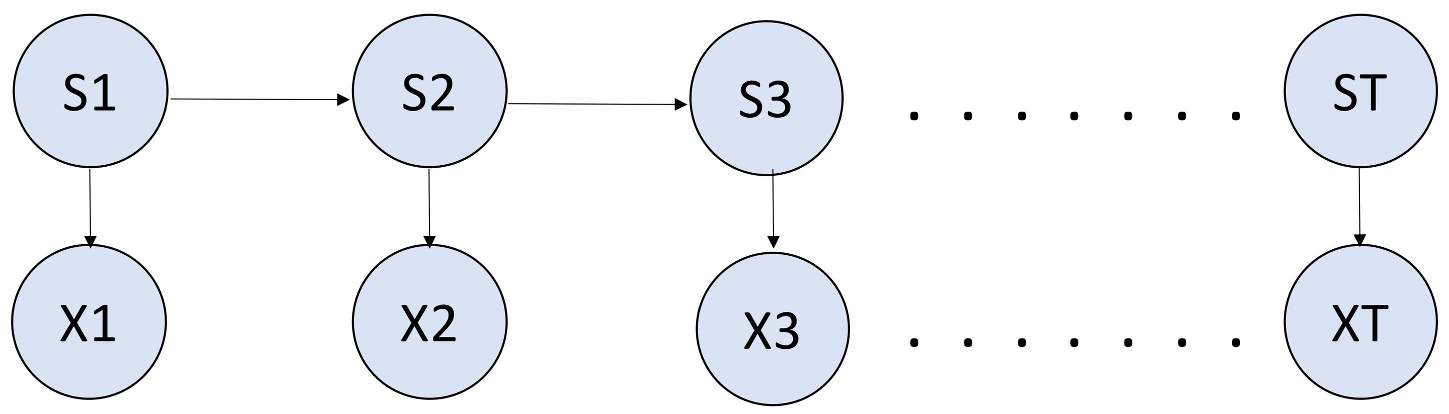
A Robust Implementation of the Expectation-Maximization Algorithm to a Hidden Markov Model

Massachusetts Institute of Technology

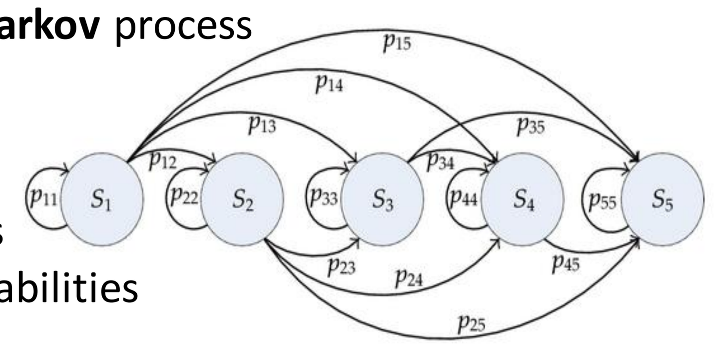
Aniruddha Deshpande[1], Alexander Birch[2]

[1]Technology & Policy Program
[2]Sloan School of Management

Introduction – Hidden Markov Model



- Statistical model that can be used to describe the evolution of observable events that depend on internal factors, which are not directly observable
- **Key Assumption**: State evolution is a **Markov** process
- **S**: Unobserved internal states
- **X**: Observable data
- $X_t \sim P_{S_t}(.)$: Set of Emission distributions
- $A_{ij} = P(S_{t+1}=j \mid S_t=i)$: State transition probabilities
- $\pi_i = P(S_1 = i)$: Initial State probabilities



- Given true model parameters, **Viterbi Algorithm** estimates underlying state sequence using data sequence S
- Applications: Speech recognition, Gene prediction, Computational Finance

Parameter Estimation – EM algorithm

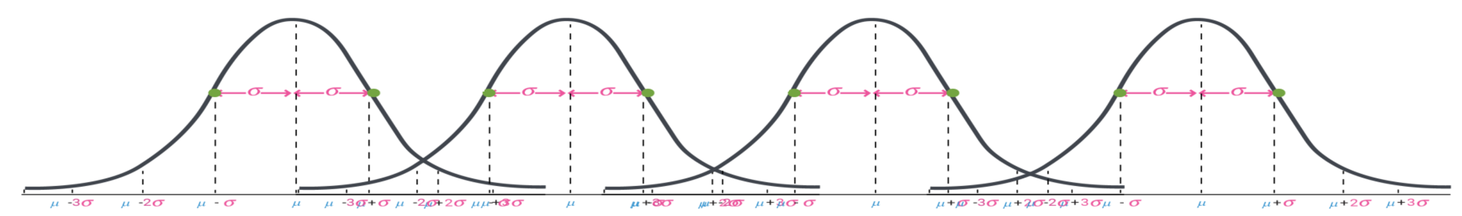
- **A, π , P** : Estimate transition probabilities, initial probabilities, emission distribution to be estimated
- Parameter estimation using Maximum Likelihood Estimation
- Non-trivial as full data is not observed, only X observed; states S unknown

- **Expectation-Maximization Algorithm:**
  - Iterative algorithm used for parameter estimation with partially observable data. In each iteration
  - **Initialize:** θ^0
  - **E-step:** Compute $U(\theta,\theta^i) = E_{P_{Z|Y}(.|y;\theta^i)}(P_Z(z;\theta))$
  - **M-Step:** $\theta^{i+1} = argmax_{\theta\in\Theta} U(\theta,\theta^i)$

- **KEY IDEA :** Transform M-step of the algorithm to a robust optimization step
  - **Robust M-step:** $\theta^{i+1} = argmax_{\theta\in\Theta} min_{y\in U} U(\theta,\theta^i)$
  - **Optimistic M-step:** $\theta^{i+1} = argmax_{\theta\in\Theta} max_{y\in U} U(\theta,\theta^i)$

Problem Formulation and Inner optimization

- Z = (S,X) ; Y = X
- **Emission distributions**: 1D Gaussian Emission distributions
  - Mean and Variance depending on the underlying state



- **m:** Number of states ; **T:** Number of timesteps ;
- **r:** Size of uncertainty set; $|\Delta x| < r$

- **Inner Optimization problem:**

$$\min_{|\Delta x_t|\le r;\forall t\in[T]} \sum_{i=1}^{m}\sum_{t=1}^{T}\left(-\frac{\alpha_i(t)\beta_i(t)}{2\sigma_i^2}\right)(x_t+\Delta x_t-\mu_i)^2$$

- **Robust: Non-convex** constrained optimization problem, with linear constraints
- **Optimistic: Convex** constrained optimization problem, with linear constraints
- α and β are quantities computed in the E-step

Algorithm

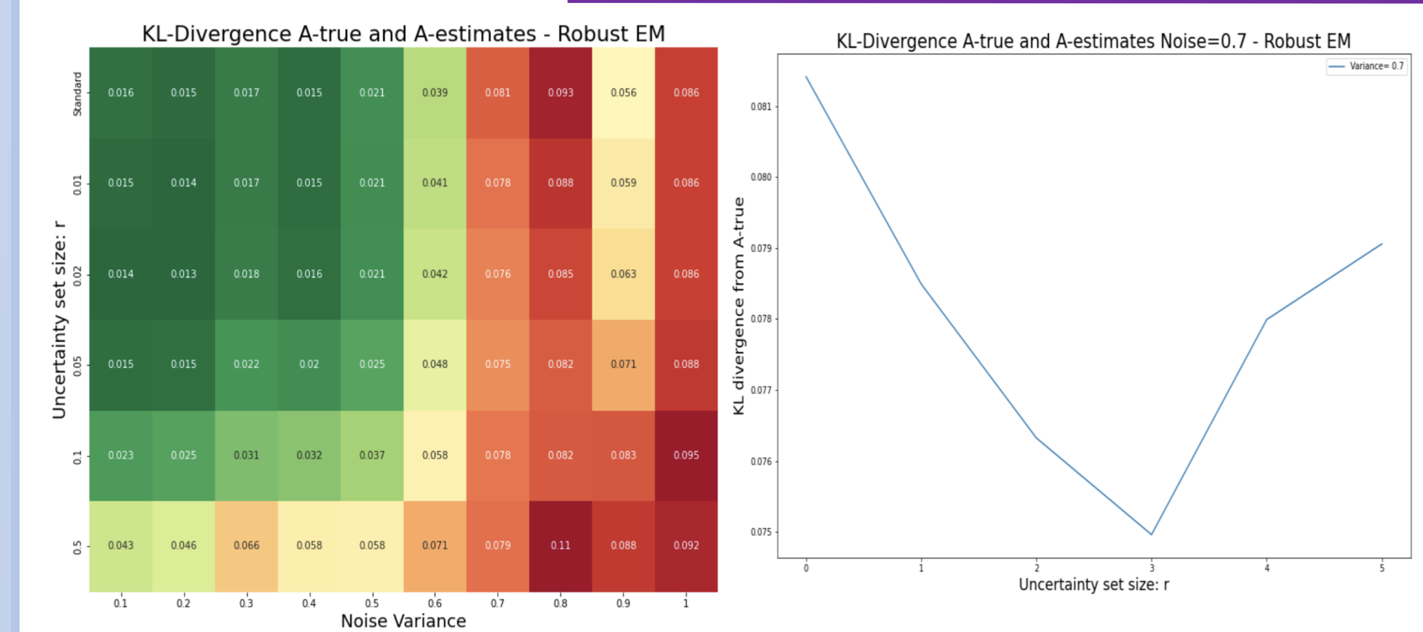
Initialize:	E-Step1:	M-Step1:	E-Step2:	M-Step2:
	Compute:	Solve inner optimization problem:	Compute:	Solve outer maximization:
θ^0 , Δx^0	$U(\theta^i, \Delta x^i)$	Δx^{i+1}	$U(\theta^i, \Delta x^{i+1})$	θ^{i+1}

Dataset & Experiments

- **Dataset -** Synthetic data generated from:
  - HMM with 5 states, Known 1D Gaussian emissions -
  - **S, X:** 10 sequences of length 500 for training, 5 sequences for testing
    - X_{noisy} : Add zero-mean noise with variances 0.1 to 1.0
- **Evaluation metrics:**
  - **$D(A_{true} || A_{est})$** - Average KL-Divergence b/w rows of true and estimated transition probability matrix
  - **State Prediction accuracy**: Fraction of states predicted correctly using estimated parameters
- **Experiments:**
  - Run robust and optimistic EM with varying uncertainty set sizes on entire range of noisy data
  - Uncertainty set: r in [0.01, 0.02, 0.05, 0.1, 0.5]
  - Observe how performance metrics vary with changing r and noise
  - Timing analysis

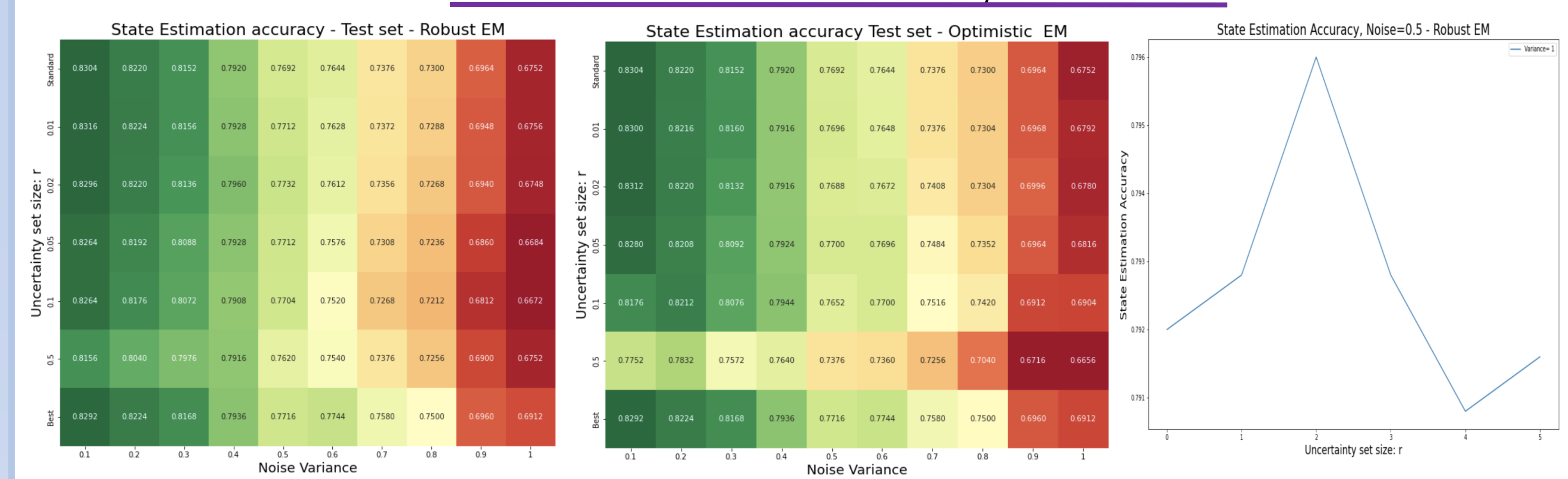
Results

Metric 1: Absolute Error in Parameter Estimations



Metric1:
Improvements for some cases of robust EM
No improvements with optimistic EM
Metric2:
Both robust and optimistic EM given improved performance for some cases
Trend: Best performance for intermediate r value

Metric 2: State Prediction Accuracy on Test set



Timing Analysis

Scalability

# States	Sequence Length	# Sequences	Standard (s)	Robust(s)	Optimistic (s)
5	500	10	8.62	134.42	33.32
10	500	10	20.75	186.68	59.35
20	500	10	96.63	369.91	150.81
2	100	5	2.69	32.56	2.12
2	200	5	3.34	56.07	3.36
2	400	5	1.57	70.89	5.08
2	50	10	0.51	19.72	2.20
2	50	20	0.83	38.48	3.78
2	50	40	1.40	73.05	6.64

Areas for future work

- Generalize formulation for n-dimensional Gaussian emissions
- Generalize formulation for mixture of Gaussian emissions – Applications in speech recognitions
- Explore theoretical performance bounds, or bound on uncertainty set size for improved performance