Lectures 8-9  CMS 165

Spectral Methods
Spectral Methods

• Utilize spectral decomposition of matrices (and tensors)
• Review of Eigen Decomposition

For a matrix $S$, $u$ is an eigenvector if $Su = \lambda u$ and $\lambda$ is eigenvalue.

- For symm. $S \in \mathbb{R}^{d \times d}$, there are $d$ eigen values.
- $S = \sum_{i \in [d]} \lambda_i u_i u_i^\top$. $U$ is orthogonal.

Rayleigh Quotient

For matrix $S$ with eigenvalues $\lambda_1 \geq \lambda_2 \ldots \lambda_d$ and corresponding eigenvectors $u_1, \ldots u_d$, then

$$\max_{\|z\|=1} z^\top Sz = \lambda_1, \quad \min_{\|z\|=1} z^\top Sz = \lambda_d,$$

and the optimizing vectors are $u_1$ and $u_d$.

Optimal Projection

$$\max_{P: P^2 = I, \text{Rank}(P) = k} \text{Tr}(P^\top SP) = \lambda_1 + \lambda_2 \ldots + \lambda_k$$
and $P$ spans $\{u_1, \ldots, u_k\}$. 
Simplest Spectral Method: PCA

Optimization problem

For (centered) points \( x_i \in \mathbb{R}^d \), find projection \( P \) with \( \text{Rank}(P) = k \) s.t.

\[
\min_{P \in \mathbb{R}^{d \times d}} \frac{1}{n} \sum_{i \in [n]} \| x_i - Px_i \|^2.
\]

Result: If \( S = \text{Cov}(X) \) and \( S = U \Lambda U^\top \) is eigen decomposition, we have \( P = U(k) U(k)^\top \), where \( U(k) \) are top-\( k \) eigen vectors.

Proof

- By Pythagorean theorem: \( \sum_i \| x_i - Px_i \|^2 = \sum_i \| x_i \|^2 - \sum_i \| Px_i \|^2 \).
- Maximize: \( \frac{1}{n} \sum_i \| Px_i \|^2 = \frac{1}{n} \sum_i \text{Tr} \left[ Px_i x_i^\top P^\top \right] = \text{Tr}[PS P^\top] \).
PCA on Gaussian Mixtures

- $k$ Gaussians: each sample is $x = Ah + z$.
- $h \in [e_1, \ldots, e_k]$, the basis vectors. $E[h] = w$.
- $A \in \mathbb{R}^{d \times k}$: columns are component means.
- Let $\mu := Aw$ be the mean.
- $z \sim \mathcal{N}(0, \sigma^2 I)$ is white Gaussian noise.

$$
E[(x - \mu)(x - \mu)^T] = \sum_{i \in [k]} w_i (a_i - \mu)(a_i - \mu)^T + \sigma^2 I.
$$

How the above equation is obtained

$$
E[(x - \mu)(x - \mu)^T] = E[(Ah - \mu)(Ah - \mu)^T] + E[zz^T]
$$

$$
= \sum_{i \in [k]} w_i (a_i - \mu)(a_i - \mu)^T + \sigma^2 I.
$$
PCA on Gaussian Mixtures Cont.

\[ \mathbb{E}[(x - \mu)(x - \mu)^T] = \sum_{i \in [k]} w_i (a_i - \mu)(a_i - \mu)^T + \sigma^2 I. \]

- The vectors \( \{a_i - \mu\} \) are linearly dependent: \( \sum_i w_i (a_i - \mu) = 0 \). The PSD matrix \( \sum_{i \in [k]} w_i (a_i - \mu)(a_i - \mu)^T \) has rank \( \leq k - 1 \).

- \((k - 1)\)-PCA on covariance matrix \( \cup \{\mu\} \) yields \( \text{span}(A) \).

**Learning \( A \) through Spectral Clustering**

- Project samples \( x \) on to \( \text{span}(A) \).
- Distance-based clustering (e.g. \( k \)-means).
- A series of works, e.g. Vempala & Wang.
Hidden Markov Models

- Why HMMs?
  - Handle temporally-dependent data
  - Succinct “factored” representation when state space is low-dimensional (c.f. autoregressive model)

- Some uses of HMMs:
  - Monitor “belief state” of dynamical system
  - Infer latent variables from time series
  - Density estimation

Source: slides from Daniel Hsu
Discrete Hidden Markov Models

- $\mathbb{P}[h_{t+1} = i | h_t = j] = T_{i,j}.$
- $\mathbb{E}[x_t | h_t = j] = Oe_j.$
- $\pi$: Initial distribution (of $x_1$).
- Three view model. $w := T\pi.$

\[ \mathbb{E}[x_1 | h_2] = O\text{Diag}(\pi)T^\top\text{Diag}(w)^{-1}h_2 \]
\[ \mathbb{E}[x_2 | h_2] = Oh_2 \]
\[ \mathbb{E}[x_3 | h_2] = OTh_2. \]

Condition for non-degeneracy
- $O \in \mathbb{R}^{d \times k}$ has full column rank.
- $T$ is invertible, $\pi$ and $T\pi$ have positive entries.
Observable operator in HMM

Discrete HMMs: observation operators

For $x \in \{1, \ldots, n\}$: define

$$A_x \triangleq \begin{bmatrix} T \end{bmatrix} \begin{bmatrix} O_{x,1} & \cdots & 0 \\ 0 & \ddots & 0 \\ \vdots & \ddots & O_{x,m} \end{bmatrix} \in \mathbb{R}^{m \times m}$$

$$[A_x]_{i,j} = \Pr[h_{t+1} = i \land x_t = x \mid h_t = j].$$

The $\{A_x\}$ are observation operators (Schützenberger, '61; Jaeger, '00).

Source: slides from Daniel Hsu
Observable operator in HMM contd.

**Using observation operators**

Matrix multiplication handles “local” marginalization of hidden variables: e.g.

\[
\Pr[x_1, x_2] = \sum_{h_1} \Pr[h_1] \cdot \sum_{h_2} \Pr[h_2|h_1] \Pr[x_1|h_1] \cdot \sum_{h_3} \Pr[h_3|h_2] \Pr[x_2|h_2]
\]

\[
= \vec{1}_m^T A_{x_2} A_{x_1} \vec{\pi}
\]

where \( \vec{1}_m \in \mathbb{R}^m \) is the all-ones vector.

**Upshot:** The \( \{A_x\} \) contain the same information as \( T \) and \( O \).
Learning Observable Operators in HMM

Key rank condition: require $T \in \mathbb{R}^{m \times m}$ and $O \in \mathbb{R}^{n \times m}$ to have rank $m$
(rules out pathological cases from hardness reductions)

Define $P_1 \in \mathbb{R}^n$, $P_{2,1} \in \mathbb{R}^{n \times n}$, $P_{3,x,1} \in \mathbb{R}^{n \times n}$ for $x = 1, \ldots, n$ by

$$
[P_1]_i = \Pr[x_1 = i] \\
[P_{2,1}]_{i,j} = \Pr[x_2 = i, x_1 = j] \\
[P_{3,x,1}]_{i,j} = \Pr[x_3 = i, x_2 = x, x_1 = j]
$$

(probabilities of singletons, doubles, and triples).

Claim: Can recover equivalent HMM parameters from $P_1$, $P_{2,1}$, $\{P_{3,x,1}\}$, and these quantities can be estimated from data.

Source: slides from Daniel Hsu
"Thin" SVD: \( P_{2,1} = U \Sigma V^\top \) where \( U = [\bar{u}_1 \ldots |\bar{u}_m] \in \mathbb{R}^{n \times m} \)

Guaranteed \( m \) non-zero singular values by rank condition.

New parameters (based on \( U \)) implicitly transform hidden states

\[
\tilde{h}_t \mapsto (U^\top O)\tilde{h}_t = U^\top \mathbb{E}[\tilde{x}_t|\tilde{h}_t]
\]

(i.e. change to coordinate representation of \( \mathbb{E}[\tilde{x}_t|\tilde{h}_t] \) w.r.t. \( \{\bar{u}_1, \ldots, \bar{u}_m\} \)).
Learning Observable Operators in HMM cont.

For each $x = 1, \ldots, n$,

$$B_x \triangleq (U^T P_{3,x,1}) (U^T P_{2,1})^+ \quad (X^+ \text{ is pseudoinv. of } X)$$
$$= (U^T O) A_x (U^T O)^{-1} \quad \text{(algebra)}$$

The $B_x$ operate in the coord. system defined by $\{\bar{u}_1, \ldots, \bar{u}_m\}$ (columns of $U$).

Pr[$x_{1:t}$] = $\bar{1}_m^T A_{x_t} \cdots A_{x_1} \bar{\pi} = \bar{1}_m^T (U^T O)^{-1} B_{x_t} \cdots B_{x_1} (U^T O) \bar{\pi}$

**Upshot:** Suffices to learn $\{B_x\}$ instead of $\{A_x\}$.
Learning Algorithm for HMM

1. Look at triples of observations $(x_1, x_2, x_3)$ in data; estimate frequencies $\hat{P}_1$, $\hat{P}_{2,1}$, and $\{\hat{P}_{3,x,1}\}$

2. Compute SVD of $\hat{P}_{2,1}$ to get matrix of top $m$ singular vectors $\hat{U}$ ("subspace identification")

3. Compute $\hat{B}_x \triangleq (\hat{U}^\top \hat{P}_{3,x,1})(\hat{U}^\top \hat{P}_{2,1})^+$ for each $x$ ("observation operators")

4. Compute $\hat{b}_1 \triangleq \hat{U}^\top \hat{P}_1$ and $\hat{b}_\infty \triangleq (\hat{P}_{2,1}^\top \hat{U})^+ \hat{P}_1$

- Joint probability calculations:
  \[
  \hat{\Pr}[x_1, \ldots, x_t] \triangleq \hat{b}_\infty^\top \hat{B}_{x_1} \ldots \hat{B}_{x_t} \hat{b}_1.
  \]

- Conditional probabilities: Given $x_{1:t-1}$,
  \[
  \hat{\Pr}[x_t|x_{1:t-1}] \triangleq \hat{b}_\infty^\top \hat{B}_{x_t} \hat{b}_t
  \]

  where
  \[
  \hat{b}_t \triangleq \frac{\hat{B}_{x_{t-1}} \ldots \hat{B}_{x_1} \hat{b}_1}{\hat{b}_\infty \hat{B}_{x_{t-1}} \ldots \hat{B}_{x_1} \hat{b}_1} \approx (U^\top O)\mathbb{E}[\tilde{h}_t|x_{1:t-1}].
  \]

"Belief states" $\hat{b}_t$ linearly related to conditional hidden states. ($b_t$ live in hypercube $[-1, +1]^m$ instead simplex $\Delta^m$)
Learning Guarantees

Sample complexity bound

Joint probability accuracy: with probability $\geq 1 - \delta$,

$$O\left(\frac{t^2}{\epsilon^2} \cdot \left(\frac{m}{\sigma_m(O)^2 \sigma_m(P_{2,1})^4} + \frac{m \cdot n_0}{\sigma_m(O)^2 \sigma_m(P_{2,1})^2}\right) \cdot \log \frac{1}{\delta}\right)$$

observation triples sampled from the HMM suffices to guarantee

$$\sum_{x_1, \ldots, x_t} |\Pr[x_1, \ldots, x_t] - \hat{\Pr}[x_1, \ldots, x_t]| \leq \epsilon.$$

- $m$: number of states
- $n_0$: number of observations that account for most of the probability mass
- $\sigma_m(M)$: $m$th largest singular value of matrix $M$

Also have a sample complexity bound for conditional probability accuracy.
Lots of other applications of spectral methods

- Extending HMMs to Partially observed Markov decision processes (POMDP) and Predictive state representations (PSR): passive vs active.
- POMDP: Action based on each observation and can influence Markovian evolution of hidden state
- PSR: No explicit Markovian assumption on hidden state. Directly predicts future (tests) based on past observations and actions (For linear PSR, similar to spectral updates in HMM)
- Stochastic bandits in a low rank subspace (ask TA Sahin about it)
References

• Matrix computations (textbook) by Golub and Van Loan
• A spectral algorithm for learning hidden Markov models by Hsu, Kakade and Zhang.
• Spectral Approaches to Learning Predictive Representations by Byron Boots (PhD thesis)