CS142 ML: Lecture 21 Outline

- Expectation-Maximization for Learning HMMs
- Application: Robot Localization
- Principal Component Analysis (PCA)
- Application: Human shape modeling
EM for Hidden Markov Models

For an HMM with $K$ states:

$$
\begin{align*}
\pi_k &= P(z_1 = k) \\
A_{k\ell} &= P(x_{t+1} = \ell \mid x_t = k) \\
p(x_t \mid z_t = k) &= f(x_t \mid \theta_k)
\end{align*}
$$

- **E-Step:** Given parameters, infer posterior of hidden variables via Bayes rule:

$$
q^{(t)}(z) = p(z \mid x, \pi^{(t-1)}, A^{(t-1)}, \theta^{(t-1)})
$$

- **M-Step:** Choose parameters to maximize expected log-likelihood:

$$
\max_{\pi, A, \theta} \mathbb{E}_q[\log p(x, z \mid \pi, A, \theta)]
$$

- **Iterate.** $\log p(x \mid \pi^{(t)}, A^{(t)}, \theta^{(t)})$ monotonically increases to (local) maximum.
Learning HMMs from Multiple Sequences

\[ z_{ntk} = 1 \text{ if sequence } n \text{ is in state } k \text{ at time } t, \quad z_{ntk} = 0 \text{ otherwise.} \]
\[ x_{nt} = \text{observation at time } t \text{ in sequence } n. \]

\[ n = 1: \quad z_1 \rightarrow z_2 \rightarrow z_3 \rightarrow z_4 \]
\[ x_1 \quad x_2 \quad x_3 \quad x_4 \]

\[ n = 2: \quad z_1 \rightarrow z_2 \rightarrow z_3 \]
\[ x_1 \quad x_2 \quad x_3 \]

\[ n = 3: \quad z_1 \rightarrow z_2 \rightarrow z_3 \rightarrow z_4 \rightarrow z_5 \]
\[ x_1 \quad x_2 \quad x_3 \quad x_4 \quad x_5 \]

- Many audio recordings of people speaking words
- Many videos of people performing activities
- Many amino acid sequences from different proteins
- …
M-Step for HMMs

$z_{ntk} = 1$ if sequence $n$ is in state $k$ at time $t$, $z_{ntk} = 0$ otherwise.

Given $q(z)$ from last E-step:

$$\max_{\pi, A, \theta} \mathbb{E}_q[\log p(x, z \mid \pi, A, \theta)]$$

$$\mathbb{E}_q[\log p(x, z \mid \pi, A, \theta)] =$$

- **Initial states**
  $$\sum_{n=1}^{N} \sum_{k=1}^{K} \mathbb{E}_q[z_{n1k}] \log \pi_k$$

- **State transitions**
  $$\sum_{n=1}^{N} \sum_{t=2}^{T_n} \sum_{k=1}^{K} \sum_{\ell=1}^{K} \mathbb{E}_q[z_{nt-1,k} z_{nt\ell}] \log A_{k\ell}$$

- **Observed data**
  $$\sum_{n=1}^{N} \sum_{t=1}^{T_n} \sum_{k=1}^{K} \mathbb{E}_q[z_{ntk}] \log f(x_{nt} \mid \theta_k)$$
A Constrained Optimization Lemma

**Objective:**
\[ \hat{\theta} = \arg \max_{\theta} \sum_{k=1}^{K} a_k \log \theta_k \]

subject to
\[ \sum_{k=1}^{K} \theta_k = 1 \quad 0 \leq \theta_k \leq 1 \]

**Solution:**
\[ \hat{\theta}_k = \frac{a_k}{a_0} \]
\[ a_0 = \sum_{k=1}^{K} a_k \]

Proof for $K=2$: Change of variables to unconstrained problem

Proof for general $K$: Lagrange multipliers
HMM M-step: Initial States

$z_{ntk} = 1$ if sequence $n$ is in state $k$ at time $t$, $z_{ntk} = 0$ otherwise.

Given $q(z)$ from last E-step:

$$\max_{\pi, A, \theta} \mathbb{E}_q[\log p(x, z | \pi, A, \theta)]$$

initial states

$$\sum_{n=1}^{N} \sum_{k=1}^{K} \mathbb{E}_q[z_{n1k}] \log \pi_k$$

- Applying constrained optimization lemma:
  $$\hat{\pi}_k = \frac{1}{N} \sum_{n=1}^{N} \mathbb{E}_q[z_{n1k}]$$

- Requirement from E-step: Marginal distribution of state at time 1 in each of $N$ training sequences, given observations at all times
HMM M-step: Observations

\( z_{ntk} = 1 \) if sequence \( n \) is in state \( k \) at time \( t \), \( z_{ntk} = 0 \) otherwise.

Given \( q(z) \) from last E-step:

\[
\max_{\pi, A, \theta} \mathbb{E}_q \left[ \log p(x, z | \pi, A, \theta) \right]
\]

\[
\sum_{n=1}^{N} \sum_{t=1}^{T_n} \sum_{k=1}^{K} \mathbb{E}_q [z_{ntk}] \log f(x_{nt} | \theta_k)
\]

- Independent weighted maximum likelihood problems for each state \( k \):
  \[
  \hat{\theta}_k = \arg \max_{\theta_k} \sum_{n=1}^{N} \sum_{t=1}^{T_n} \mathbb{E}_q [z_{ntk}] \log f(x_{nt} | \theta_k)
  \]

- Requirement from E-step: Marginal distribution of state at all times in each of \( N \) training sequences, given observations at all times

- Objective is identical to M-step for mixture models
Forward-Backward for HMMs

\[
p(z_t \mid x_1, \ldots, x_T) \propto \alpha_t(z_t) \beta_t(z_t)
\]

**Marginal:** Posterior dist. of state given all data

**Forward Recursion:** Distribution of State Given Past Data

\[
\alpha_t(z_t) = p(z_t \mid x_t, x_{t-1}, \ldots, x_1) \quad \alpha_1(z_1) \propto p(z_1)p(x_1 \mid z_1)
\]

\[
\alpha_{t+1}(z_{t+1}) \propto p(x_{t+1} \mid z_{t+1}) \sum_{z_t=1}^{K} p(z_{t+1} \mid z_t) \alpha_t(z_t)
\]

**Backward Recursion:** Likelihood of Future Data Given State

\[
\beta_t(z_t) \propto p(x_{t+1}, \ldots, x_T \mid z_t) \quad \beta_T(z_T) = 1
\]

\[
\beta_t(z_t) \propto \sum_{z_{t+1}=1}^{K} p(x_{t+1} \mid z_{t+1})p(z_{t+1} \mid z_t) \beta_{t+1}(z_{t+1})
\]
HMM M-step: State Transitions

\[ z_{ntk} = 1 \text{ if sequence } n \text{ is in state } k \text{ at time } t, \ z_{ntk} = 0 \text{ otherwise.} \]

Given \( q(z) \) from last E-step:

\[
\max_{\pi, A, \theta} \mathbb{E}_q[\log p(x, z | \pi, A, \theta)]
\]

state transitions

\[
\sum_{n=1}^{N} \sum_{t=2}^{T_n} \sum_{k=1}^{K} \sum_{\ell=1}^{K} \mathbb{E}_q[z_{n,t-1,k}, z_{nt\ell}] \log A_{k\ell}
\]

- Applying constrained optimization lemma separately to each state:

\[
\hat{A}_{k\ell} = \frac{1}{N_k} \sum_{n=1}^{N} \sum_{t=2}^{T_n} \mathbb{E}_q[z_{n,t-1,k}, z_{nt\ell}]
\]

\[
N_k = \sum_{n=1}^{N} \sum_{t=1}^{T_n-1} \mathbb{E}_q[z_{ntk}]
\]

- Requirement from E-step: Marginal distribution of pairs of sequential states, for all sequences and times, given all observations.
Forward-Backward for HMMs

**Forward Recursion:** Distribution of State Given Past Data

\[ \alpha_t(z_t) = p(z_t \mid x_t, x_{t-1}, \ldots, x_1) \]

**Backward Recursion:** Likelihood of Future Data Given State

\[ \beta_t(z_t) \propto p(x_{t+1}, \ldots, x_T \mid z_t) \]

**Marginal Distributions:** Posterior of states, and state pairs, given all data

\[ p(z_t \mid x_1, \ldots, x_T) \propto \alpha_t(z_t)\beta_t(z_t) \]

\[ p(z_t, z_{t+1} \mid x_1, \ldots, x_T) \propto \alpha_t(z_t)p(z_{t+1} \mid z_t)p(x_{t+1} \mid z_{t+1})\beta_{t+1}(z_{t+1}) \]

Example of “belief propagation” for inference in graphical models.
E-Step for HMMs

$$\pi_k = P(z_1 = k)$$

$$A_{k\ell} = P(z_{t+1} = \ell \mid z_t = k)$$

$$p(x_t \mid z_t = k) = f(x_t \mid \theta_k)$$

$$q^{(t)}(z) \propto p(z_1 \mid \pi^{(t-1)})p(x_1 \mid z_1, \theta^{(t-1)}) \prod_{n=2}^{T} p(z_n \mid A_{z_{n-1}}^{(t-1)})p(x_n \mid z_n, \theta^{(t-1)})$$

- Hidden states are \textit{conditionally dependent} given parameters
- Naïve representation of full posterior has size $O(K^T)$

But, the posterior distribution has Markov chain structure:
- Can use \textit{forward-backward algorithm} to efficiently compute the single-state and pair-state marginals needed by the M-step

$$q(z_t) = p(z_t \mid x, \pi, A, \theta) \quad q(z_t, z_{t+1}) = p(z_t, z_{t+1} \mid x, \pi, A, \theta)$$
CS142 ML: Lecture 21 Outline

- Expectation-Maximization for Learning HMMs
- Application: Robot Localization
- Principal Component Analysis (PCA)
- Application: Human shape modeling
HMM: States & Transitions

\[ z_t = (x, y, \theta) \]

\[ p(z_t | z_{t-1}) \]
HMM: Laser Range Finder Observations

(a) RHINO

(c) probability

(expected distance [cm])

(d) probability

(expected distance [cm])
HMM Localization for Mobile Robots
CS142 ML: Lecture 21 Outline

- Expectation-Maximization for Learning HMMs
- Application: Robot Localization
- Principal Component Analysis (PCA)
- Application: Human shape modeling
### Dimensionality Reduction

<table>
<thead>
<tr>
<th>Supervised Learning</th>
<th>Unsupervised Learning</th>
</tr>
</thead>
<tbody>
<tr>
<td>Discrete</td>
<td></td>
</tr>
<tr>
<td></td>
<td>classification or categorization</td>
</tr>
<tr>
<td></td>
<td>clustering</td>
</tr>
<tr>
<td>Continuous</td>
<td></td>
</tr>
<tr>
<td></td>
<td>regression</td>
</tr>
<tr>
<td></td>
<td>dimensionality reduction</td>
</tr>
</tbody>
</table>

- **Goal:** Infer low-dimensional embedding $t$ of raw data $x$
- **Classical:** Find latent variables $t$ good for compression of $x$
- **Probabilistic:** Distribution $p(t,x)$ that maximizes marginal probability $p(x)$
Dimensionality Reduction

Principal Components Analysis (PCA)

- Observed feature vectors:
  \[ x_n \in \mathbb{R}^D, \quad n = 1, 2, \ldots, N \]

- Hidden manifold coordinates \((M < D)\):
  \[ z_n \in \mathbb{R}^M, \quad n = 1, 2, \ldots, N \]

- Hidden linear mapping:
  \[ \tilde{x}_n = W z_n + b \]

- Compression objective:
  \[ J(z, W, b \mid x, M) = \sum_{n=1}^{N} \| x_n - \tilde{x}_n \|^2 = \sum_{n=1}^{N} \| x_n - W z_n - b \|^2 \]

- Unlike K-means clustering, can find the global minimum efficiently:

  \[ b = \frac{1}{N} \sum_{n=1}^{N} x_n \]

  Construct \( W \) from the top eigenvectors of the sample covariance matrix (the directions of largest variance)
Example: PCA for Images of Digit 3

\[ J(z, W, b \mid x, M) = \sum_{n=1}^{N} \| x_n - \tilde{x}_n \|^2 = \sum_{n=1}^{N} \| x_n - W z_n - b \|^2 \]

- PCA models all translations of data equally well (by shifting \( b \))
- PCA models all rotations of data equally well (by rotating \( W \))
PCA Derivation: Centering the Data

- Observed feature vectors:
  \[ x_n \in \mathbb{R}^D, \quad n = 1, 2, \ldots, N \]

- Hidden manifold coordinates (\( M < D \)): 
  \[ z_n \in \mathbb{R}^M, \quad n = 1, 2, \ldots, N \]

- Hidden linear mapping:
  \[ \tilde{x}_n = Wz_n + b \]

  \[ W \in \mathbb{R}^{D \times M} \]
  \[ b \in \mathbb{R}^{D \times 1} \]

  \[ J(z, W, b \mid x, M) = \sum_{n=1}^{N} \|x_n - \tilde{x}_n\|^2 = \sum_{n=1}^{N} \|x_n - Wz_n - b\|^2 \]

- By analyzing this quadratic function, one can show that the optimal value of the offset vector \( b \) is the mean of the observed data:
  \[ b = \bar{x} = \frac{1}{N} \sum_{n=1}^{N} x_n \]

- To simplify later derivations, we center the data as a preprocessing step:
  \[ x_n \leftarrow x_n - \bar{x} \]

  \[ \tilde{x}_n = Wz_n, \quad W \in \mathbb{R}^{D \times M} \]
PCA Derivation: Scale Ambiguity

- Observed feature vectors:
  \[ x_n \in \mathbb{R}^D, \quad n = 1, 2, \ldots, N \]

- Hidden manifold coordinates \((M < D)\):
  \[ z_n \in \mathbb{R}^M, \quad n = 1, 2, \ldots, N \]

- Hidden linear mapping:
  \[ \tilde{x}_n = W z_n, \quad W \in \mathbb{R}^{D \times M} \]
  \[ J(z, W \mid x, M) = \sum_{n=1}^{N} \|x_n - \tilde{x}_n\|^2 = \sum_{n=1}^{N} \|x_n - W z_n\|^2 \]

- The manifold coordinate system is not uniquely defined, for example:
  \[ W' = \alpha W, \quad z' = \alpha^{-1} W, \quad \alpha \neq 0 \]
  \[ J(z, W) = J(z', W') \]

- To remove the scale ambiguity, we constrain \(W\) to be orthogonal:
  \[ W^T W = I_M \]

- Ambiguities remain: can rotate and/or reflect both \(W\) and \(z\)
PCA Derivation: Dimension $M=1$

- Observed feature vectors: $x_n \in \mathbb{R}^D$, $n = 1, 2, \ldots, N$
- Hidden manifold coordinates: $z_n \in \mathbb{R}$, $n = 1, 2, \ldots, N$
- Hidden linear mapping: $\tilde{x}_n = wz_n$ where $w \in \mathbb{R}^{D \times 1}$, $\|w\|^2 = 1$

$$J(z, w \mid x) = \frac{1}{N} \sum_{n=1}^{N} \|x_n - \tilde{x}_n\|^2 = \frac{1}{N} \sum_{n=1}^{N} \|x_n - wz_n\|^2$$

**Step 1:** Optimal manifold coordinate is orthogonal projection:

$$\hat{z}_n = w^T x_n$$

**Step 2:** Optimal mapping maximizes variance: $\Sigma = \frac{1}{N} \sum_{n=1}^{N} x_n x_n^T$

$$J(\hat{z}, w \mid x) = C - \frac{1}{N} \sum_{n=1}^{N} (w^T x_n)^2 = C - w^T \Sigma w$$