Online learning for audio clustering and segmentation

Alberto Bietti$^{1,2}$

$^1$Mines ParisTech
$^2$Ecole Normale Supérieure, Cachan

September 10, 2014

Supervisors: Arshia Cont, Francis Bach
Outline

1 Introduction

2 Representation, models, offline algorithms
   - Audio signal representation
   - Clustering with Bregman divergences
   - Hidden Markov Models (HMMs)
   - Hidden Semi-Markov Models (HSMMs)
   - Offline audio segmentation results

3 Online algorithms
   - Online EM
   - Non-probabilistic algorithm
   - Incremental EM
   - Online audio segmentation results
Audio segmentation

- **Goal**: segment audio signal into homogeneous chunks/segments
- Go from a signal representation to a symbolic representation
- Applications: music indexing, summarization, fingerprinting
Audio segmentation: approaches

- Most existing approaches: find change-points, compute similarities separately
- Change-point detection
  - Use audio features for detecting changes
  - Statistical model on the signal, likelihood ratio tests
- Issues: specific to the task, doesn’t use previous parts of the signal, often supervised (needs labeled data)
Audio segmentation: approaches

- Most existing approaches: find change-points, compute similarities separately
- Change-point detection
  - Use audio features for detecting changes
  - Statistical model on the signal, likelihood ratio tests
- Issues: specific to the task, doesn’t use previous parts of the signal, often supervised (needs labeled data)
- **Our goal**: unsupervised learning, joint segmentation and clustering. online/real-time
Audio segmentation: approaches

- Most existing approaches: find change-points, compute similarities separately
- Change-point detection
  - Use audio features for detecting changes
  - Statistical model on the signal, likelihood ratio tests
- Issues: specific to the task, doesn’t use previous parts of the signal, often supervised (needs labeled data)
- Our goal: unsupervised learning, joint segmentation and clustering. online/real-time
- Hidden (semi-)Markov Models
Online learning

- Learn a model incrementally, one observation at a time
- Very successful in machine learning, especially large-scale problems
- Usually independent observations, little work on sequential models
Online learning

- Learn a model incrementally, one observation at a time
- Very successful in machine learning, especially large-scale problems
- Usually independent observations, little work on sequential models
- **Our goal**: online algorithms for hidden (semi-)Markov models, applications to online audio segmentation and clustering
Outline

1 Introduction

2 Representation, models, offline algorithms
   - Audio signal representation
   - Clustering with Bregman divergences
   - Hidden Markov Models (HMMs)
   - Hidden Semi-Markov Models (HSMMs)
   - Offline audio segmentation results

3 Online algorithms
   - Online EM
   - Non-probabilistic algorithm
   - Incremental EM
   - Online audio segmentation results
Audio signal representation

- Discrete audio signal $x[t] \in \mathbb{R}$
- Short-time Fourier Transform

$$\hat{x}(t, e^{i\omega}) = \sum_{u=\pm\infty} x[u]g[u-t]e^{-i\omega u}$$

- Window $g$ (e.g., Hamming), compact support: FFT $\hat{x}_{t,1}, \ldots, \hat{x}_{t,p} \in \mathbb{C}$
- $x_t \in \mathbb{R}^p = (|\hat{x}_{t,1}|, \ldots, |\hat{x}_{t,p}|)^\top$
- Normalized $\sum_j x_{t,j} = 1$ for invariance to volume
Outline

1. Introduction

2. Representation, models, offline algorithms
   - Audio signal representation
   - Clustering with Bregman divergences
   - Hidden Markov Models (HMMs)
   - Hidden Semi-Markov Models (HSMMs)
   - Offline audio segmentation results

3. Online algorithms
   - Online EM
   - Non-probabilistic algorithm
   - Incremental EM
   - Online audio segmentation results
Bregman divergences

- Euclidian distance doesn’t perform well for audio
- Defines a different similarity measure
- Bregman divergence $D_\psi$ for $\psi$ strictly convex:

$$D_\psi(x, y) = \psi(x) - \psi(y) - \langle x - y, \nabla \psi(y) \rangle.$$

- Examples:
  - Squared Euclidian distance $\|x - y\|^2 = D_\psi$ with $\psi(x) = \|x\|^2$
  - KL divergence $D_{KL}(x \| y) = \sum_i x_i \log \frac{x_i}{y_i} = D_\psi(x, y)$ with $\psi(x) = \sum_i x_i \log x_i$
Bregman divergences

- Euclidian distance doesn’t perform well for audio
- Defines a different similarity measure
- Bregman divergence $D_\psi$ for $\psi$ strictly convex:

$$D_\psi(x, y) = \psi(x) - \psi(y) - \langle x - y, \nabla \psi(y) \rangle.$$  

- Examples:
  - Squared Euclidian distance $\|x - y\|^2 = D_\psi$ with $\psi(x) = \|x\|^2$
  - KL divergence $D_{KL}(x \| y) = \sum_i x_i \log \frac{x_i}{y_i} = D_\psi(x, y)$ with $\psi(x) = \sum_i x_i \log x_i$
- Right-type centroid = average (see e.g., (Nielsen and Nock, 2009))

$$\arg \min_c \sum_{i=1}^{n} D_\psi(x_i, c) = \frac{1}{n} \sum_{i=1}^{n} x_i$$
Hard clustering (K-means)

- $x_i$, $i = 1, \ldots, n$, centroids $\mu_1, \ldots, \mu_K$, assignments $z_i$

- **K-means**, replace $||x_i - \mu_{z_i}||^2$ with $D_\psi(x_i, \mu_{z_i})$
  
  - **E-step**
    
    $$z_i \leftarrow \arg\min_k D_\psi(x_i, \mu_k) \quad i = 1, \ldots, n$$
  
  - **M-step**
    
    $$\mu_k \leftarrow \frac{1}{|\{i : z_i = k\}|} \sum_{i : z_i = k} x_i \quad k = 1, \ldots, K$$
Hard clustering (K-means)

- \( x_i, i = 1, \ldots, n \), centroids \( \mu_1, \ldots, \mu_K \), assignments \( z_i \)
- **K-means**, replace \( \| x_i - \mu_{z_i} \|^2 \) with \( D_\psi(x_i, \mu_{z_i}) \)
  - **E-step**
    \[
    z_i \leftarrow \arg\min_k D_\psi(x_i, \mu_k) \quad i = 1, \ldots, n
    \]
  - **M-step**
    \[
    \mu_k \leftarrow \frac{1}{|\{i : z_i = k\}|} \sum_{i : z_i = k} x_i \quad k = 1, \ldots, K
    \]
- Decreases the (non-convex) objective
  \[
  \ell(\mu, z) = \sum_{i=1}^{n} D_\psi(x_i, \mu_{z_i}).
  \]
Bregman divergences and exponential families

- Exponential family:
  \[ p_{\theta}(x) = h(x) \exp(\langle \phi(x), \theta \rangle - a(\theta)) \]

- Regular exponential family: minimal, \( \Theta \) open
  \[ p_{\psi,\theta}(x) = h(x) \exp(\langle x, \theta \rangle - \psi(\theta)) \]

- Bijection between regular exponential families and regular Bregman divergences (Banerjee et al., 2005): \( \mu = \nabla \psi(\theta) = \mathbb{E}[X] \),
  \[ p_{\psi,\theta}(x) = h(x) \exp(-D_{\psi^*}(x, \mu)) \]

- Example: KL divergence \( \Leftrightarrow \) Multinomial distribution
  \[ h(x) \exp\left(-\sum_i x_i \log \frac{x_i}{\mu_i}\right) = h'(x) \prod_i \mu_i^{x_i} \]
Mixture models

- $x_i, i = 1, \ldots, n$, $K$ mixture components, emission parameters $\mu_k$
- Model:

$$z_i \sim \pi, \quad i = 1, \ldots, n$$

$$x_i | z_i \sim p_{\mu z_i}, \quad i = 1, \ldots, n,$$
EM algorithm

- **x** observed variables, **z** hidden variables, **θ** parameter
- Goal: maximum likelihood \( \max_{\theta} p(x; \theta) \)

\[
\ell(\theta) = \log \sum_z p(x, z; \theta) = \log \sum_z q(z) \frac{p(x, z; \theta)}{q(z)} \\
\geq \sum_z q(z) \log \frac{p(x, z; \theta)}{q(z)}.
\]

- **E-step**: maximize w.r.t. \( q \). \( q(z) = p(z|x; \theta) \)
- **M-step**: maximize w.r.t. \( \theta \). \( \hat{\theta} = \arg \max_{\theta} \mathbb{E}_{z \sim q}[\log p(z, x; \theta)] \)

Alberto Bietti  
Online learning and audio segmentation  
September 10, 2014 13 / 55
Mixture models: EM (soft clustering)

- $x_i, i = 1, \ldots, n$, initial parameters $\pi, \mu_k$.

\[
\mathbb{E}_{z \sim q}[\log p(x, z; \pi, \mu)] \\
= \sum_{i} \sum_{k} \mathbb{E}_q[\mathbb{I}\{z_i = k\}] \log \pi_k + \sum_{i} \sum_{k} \mathbb{E}_q[\mathbb{I}\{z_i = k\}] \log p(x_i|k)
\]

- **E-step**

\[
\tau_{ik} \leftarrow p(z_i = k|x_i) = \frac{1}{Z_k} \pi_k e^{-D_\psi(x_i, \mu_k)}
\]

- **M-step**

\[
\pi_k \leftarrow \frac{1}{n} \sum_i \tau_{ik} \\
\mu_k \leftarrow \frac{\sum_i \tau_{ik} x_i}{\sum_i \tau_{ik}}
\]
Outline

1 Introduction

2 Representation, models, offline algorithms
   - Audio signal representation
   - Clustering with Bregman divergences
   - Hidden Markov Models (HMMs)
   - Hidden Semi-Markov Models (HSMMs)
   - Offline audio segmentation results

3 Online algorithms
   - Online EM
   - Non-probabilistic algorithm
   - Incremental EM
   - Online audio segmentation results
Hidden Markov Models (HMMs)

- Observed sequence $x_{1:T}$, hidden sequence $z_{1:T}$, parameters $\pi, A \in \mathbb{R}^{K \times K}, \mu_k$

  \[ z_1 \sim \pi \]

  \[ z_t | z_{t-1} = i \sim A_i, \quad t = 2, \ldots, T \]

  \[ x_t | z_t = i \sim p_{\mu_i}, \quad t = 1, \ldots, T \]

- Joint likelihood:

  \[ p(x_{1:T}, z_{1:T}; \pi, A, \mu) = p(z_1; \pi) \prod_{t=2}^{T} p(z_t | z_{t-1}; A) \prod_{t=1}^{T} p(x_t | z_t; \mu) \]
HMM inference: Forward-Backward algorithm

- Inference: compute \( p(\mathbf{z}_t = i | \mathbf{x}_{1:T}) \) (smoothing)
- Definitions:

  \[
  \alpha_t(i) = p(\mathbf{z}_t = i, \mathbf{x}_1, \ldots, \mathbf{x}_t) \\
  \beta_t(i) = p(\mathbf{x}_{t+1}, \ldots, \mathbf{x}_T | \mathbf{z}_t = i).
  \]

- Recursions, with \( \alpha_1(i) = \pi_i p(\mathbf{x}_1 | \mathbf{z}_1 = i), \beta_T(i) = 1 \):

  \[
  \alpha_{t+1}(j) = \sum_i \alpha_t(i) A_{ij} p(\mathbf{x}_{t+1} | \mathbf{z}_{t+1} = j) \\
  \beta_t(i) = \sum_j A_{ij} p(\mathbf{x}_{t+1} | \mathbf{z}_{t+1} = j) \beta_{t+1}(j)
  \]

- \( p(\mathbf{z}_t = i | \mathbf{x}_{1:T}) \propto \alpha_t(i) \beta_t(i) \)
HMM inference: Viterbi algorithm

- Compute *maximum a posteriori* (MAP) sequence:
  \[
  z_{1:T}^{\text{MAP}} = \arg \max_{z_{1:T}} p(z_{1:T} | x_{1:T})
  \]

- Define
  \[
  \gamma_t(i) = \max_{z_1, \ldots, z_{t-1}} p(z_1, \ldots, z_{t-1}, z_t = i, x_1, \ldots, x_t)
  \]

- Recursion, with \( \gamma_1(i) = \pi_i p(x_1 | z_1 = i; \mu_i) \):
  \[
  \gamma_{t+1}(j) = \max_i \gamma_t(i) A_{ij} p(x_{t+1} | z_{t+1} = j; \mu_j)
  \]

- Recover the sequence by storing back-pointers.
HMM learning: EM

- **E-step**

\[
\tau_t(i) \leftarrow p(z_t = i|x_1:T) \propto \alpha_t(i)\beta_t(i)
\]

\[
\tau_t(i, j) \leftarrow p(z_{t-1} = i, z_t = j|x_1:T) \propto \alpha_{t-1}(i)A_{ij}p(x_t|j)\beta_t(j)
\]

- **M-step**

\[
\pi_i \leftarrow \tau_1(i)
\]

\[
A_{ij} \leftarrow \frac{\sum_{t \geq 2} \tau_t(i, j)}{\sum_{j'} \sum_{t \geq 2} \tau_t(i, j')}
\]

\[
\mu_i \leftarrow \frac{\sum_{t \geq 1} \tau_t(i)x_i}{\sum_{t \geq 1} \tau_t(i)}
\]
Outline

1. Introduction

2. Representation, models, offline algorithms
   - Audio signal representation
   - Clustering with Bregman divergences
   - Hidden Markov Models (HMMs)
   - Hidden Semi-Markov Models (HSMMs)
   - Offline audio segmentation results

3. Online algorithms
   - Online EM
   - Non-probabilistic algorithm
   - Incremental EM
   - Online audio segmentation results
Duration distributions

- Probability of staying in state $i$ for $d$ time steps:
  \[ A_{ii}^{d-1}(1 - A_{ii}) \]

- i.e., segment lengths follow geometric distributions
- Duration distribution learned implicitly through $A_{ii}$
- HSMMs: model these duration distributions explicitly (explicit-duration HMM)
- Typical choices: Negative Binominal, Poisson
Segment = (state $z$, length $l$), with $l \sim p_z(d)$

(Markov) transitions $A_{ij}$ between segments

$l$ i.i.d. observations from cluster $z$ in each segment

$$x_t, \ldots, x_{t+l-1} \sim p_{\mu_z}, \quad \text{i.i.d.}$$
Hidden Semi-Markov Models (Murphy, 2002)

- Two hidden variables: state $z_t$, deterministic counter $z_t^D$
- $f_t = 1$ iff new segment starts at $t + 1$

\[
p(z_t = j | z_{t-1} = i, f_{t-1} = f) = \begin{cases} 
\delta(i, j), & \text{if } f = 0 \\
A_{ij}, & \text{if } f = 1 \text{ (transition)}
\end{cases}
\]

\[
p(z_t^D = d | z_t = i, f_{t-1} = 1) = p_i(d)
\]

\[
p(z_t^D = d | z_t = i, z_{t-1}^D = d' \geq 2) = \delta(d, d' - 1),
\]
HSMM inference: Forward-Backward algorithm

- Definitions:

\[ \alpha_t(j) = p(z_t = j, f_t = 1, x_{1:t}) \]
\[ \alpha_t^*(j) = p(z_{t+1} = j, f_t = 1, x_{1:t}) \]
\[ \beta_t(i) = p(x_{t+1:T} | z_t = i, f_t = 1) \]
\[ \beta_t^*(i) = p(x_{t+1:T} | z_{t+1} = i, f_t = 1). \]

- Recursions, with \( \alpha_0^*(j) = \pi_j \) and \( \beta_T(i) = 1 \):

\[ \alpha_t(j) = \sum_d p(x_{t-d+1:t} | j, d)p(d | j)\alpha_{t-d}^*(j) \]
\[ \alpha_t^*(j) = \sum_i \alpha_t(i)A_{ij} \]
\[ \beta_t(i) = \sum_j \beta_t^*(j)A_{ij} \]
\[ \beta_t^*(i) = \sum_d \beta_{t+d}(i)p(d | i)p(x_{t+1:t+d} | i, d). \]
HSMM: EM

- Define:
  \[ \gamma_t(i) = p(z_t = i, f_t = 1| x_{1:T}) \propto \alpha_t(i) \beta_t(i) \]
  \[ \gamma^*_t(i) = p(z_{t+1} = i, f_t = 1| x_{1:T}) \propto \alpha^*_t(i) \beta^*_t(i). \]

- E-step
  \[ p(z_t = i| x_{1:T}) = \sum_{\tau < t} (\gamma^*_\tau(i) - \gamma_\tau(i)) \]
  \[ p(z_t = i, z_{t+1} = j| f_t = 1, x_{1:T}) \propto \alpha_t(i) A_{ij} \beta^*_t(j) \]

- M-step
  \[ \pi_i = p(z_1 = i| x_{1:T}) \]
  \[ A_{ij} = \frac{\sum_t p(z_t = i, z_{t+1} = j| f_t = 1, x_{1:T})}{\sum_{j'} \sum_t p(z_t = i, z_{t+1} = j'| f_t = 1, x_{1:T})} \]
  \[ \mu_i = \frac{\sum_t p(z_t = i| x_{1:T}) x_t}{\sum_t p(z_t = i| x_{1:T})} \]
Outline

1 Introduction

2 Representation, models, offline algorithms
   - Audio signal representation
   - Clustering with Bregman divergences
   - Hidden Markov Models (HMMs)
   - Hidden Semi-Markov Models (HSMMs)
   - Offline audio segmentation results

3 Online algorithms
   - Online EM
   - Non-probabilistic algorithm
   - Incremental EM
   - Online audio segmentation results
Examples

Ravel, *Ma Mère l’Oye*

Bach, Violin sonata n. 2, *Allegro*
Different K-means initializations. $K = 9$. HSMM duration distributions fixed to $\text{NegBin}(5, 0.95)$. 
HMM and HSMM randomly initialized (uniform spectrum + noise).

\[ K = 10. \] HSMM durations: \( NB(5, 0.2) \) (mean 20).
Outline

1. Introduction

2. Representation, models, offline algorithms
   - Audio signal representation
   - Clustering with Bregman divergences
   - Hidden Markov Models (HMMs)
   - Hidden Semi-Markov Models (HSMMs)
   - Offline audio segmentation results

3. Online algorithms
   - Online EM
   - Non-probabilistic algorithm
   - Incremental EM
   - Online audio segmentation results
Online EM for i.i.d. data (Cappé and Moulines, 2009)

- Complete-data model:
  \[
p(x, z; \theta) = h(x, z) \exp(\langle s(x, z), \eta(\theta) \rangle - a(\theta))
  \]

- Batch EM can be written as:
  \[
  S_t = \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}_z[s(x_i, z_i)|x_i; \theta_{t-1}]
  \]
  \[
  \theta_t = \bar{\theta}(S_t)
  \]

- Taking the limit \( n \to \infty \) (limiting EM):
  \[
  S_t = \mathbb{E}_{x \sim P}[\mathbb{E}_z[s(x, z)|x; \theta_{t-1}]]
  \]
  \[
  \theta_t = \bar{\theta}(S_t).
  \]
Online EM for i.i.d. data (Cappé and Moulines, 2009)

- Stochastic approximation (Robbins-Monro) procedure to solve
  \[ S_{t+1} = \mathbb{E}_{x \sim p}[\mathbb{E}_z[s(x, z)|x; \bar{\theta}(S_t)]] \]

- Online EM algorithm:
  \[
  \hat{s}_t = (1 - \gamma_t)\hat{s}_{t-1} + \gamma_t \mathbb{E}_z[s(x_t, z)|x_t; \hat{\theta}_{t-1}] \\
  \hat{\theta}_t = \bar{\theta}(\hat{s}_t).
  \]

- \( \gamma_t = t^{-\alpha}, \alpha \in (0.5, 1] \)
Online EM for HMMs (Cappé, 2011)

- Complete-data model:

\[ p(x_t, z_t | z_{t-1}; \theta) = h(z_t, x_t) \exp(\langle s(z_{t-1}, z_t, x_t), \eta(\theta) \rangle - a(\theta)) \]

- Batch EM can be written as:

\[
S_k = \frac{1}{T} \mathbb{E}_z \left[ \sum_{t=1}^{T} s(z_{t-1}, z_t, x_t) \bigg| x_0:T; \theta_{k-1} \right] \\
\theta_k = \bar{\theta}(S_k)
\]

- Limiting EM \((T \to \infty, \text{with strong assumptions})\):

\[
S_k = \mathbb{E}_{x \sim P} [\mathbb{E}_z [s(z_{-1}, z_0, x_0) | x_{-\infty:0}; \theta_{k-1}]] \\
\theta_k = \bar{\theta}(S_k),
\]
Online EM for HMMs

- Based on the *forward smoothing* recursion
- Define

\[
S_t = \frac{1}{t} \mathbb{E}_z \left[ \sum_{t' = 1}^{t} s(z_{t' - 1}, z_{t'}, x_{t'}) \bigg| x_{0:t}; \theta \right]
\]

\[
\phi_t(i) = p(z_t = i | x_{0:t})
\]

\[
\rho_t(i) = \frac{1}{t} \mathbb{E}_z \left[ \sum_{t' = 1}^{t} s(z_{t' - 1}, z_{t'}, x_{t'}) \bigg| x_{0:t}, z_t = i; \theta \right]
\]

- We have \(S_t = \sum_i \rho_t(i) \phi_t(i)\).
Online EM for HMMs

- Smoothing recursion

\[ \phi_{t+1}(j) = \frac{1}{Z} \sum_i \phi_t(i) A_{ij} p(x_{t+1}|z_{t+1} = j) \]

\[ \rho_{t+1}(j) = \sum_i \left( \frac{1}{t+1} s(i, j, x_{t+1}) + \left( 1 - \frac{1}{t+1} \right) \rho_t(i) \right) r_{t+1}(i|j), \]

with \( r_{t+1}(i|j) = p(z_t = i|z_{t+1} = j, x_{0:t}) \). Complexity \( O(K^4 + K^3 p) \).
Online EM for HMMs

- Smoothing recursion

\[ \phi_{t+1}(j) = \frac{1}{Z} \sum_i \phi_t(i) A_{ij} p(x_{t+1}|z_{t+1} = j) \]

\[ \rho_{t+1}(j) = \sum_i \left( \frac{1}{t+1} s(i, j, x_{t+1}) + \left(1 - \frac{1}{t+1}\right) \rho_t(i) \right) r_{t+1}(i|j), \]

with \( r_{t+1}(i|j) = p(z_t = i|z_{t+1} = j, x_{0:t}) \). Complexity \( O(K^4 + K^3 p) \).

- Online EM recursion replaces quantities by estimates, e.g.

\[ \hat{\rho}_{t+1}(j) = \sum_i \left( \gamma_{t+1} s(i, j, x_{t+1}) + (1 - \gamma_{t+1}) \hat{\rho}_t(i) \right) \hat{r}_{t+1}(i|j) \]

- and updates parameters after each observation.
Online EM for HSMMs

- Parameterize HSMM as HMM with 2 hidden variables, $z_t$ and an increasing counter $z_t^D$

$$p(z_t = j | z_{t-1} = i, z_t^D = d) = \begin{cases} A_{ij}, & \text{if } d = 1 \\ \delta(i, j), & \text{otherwise} \end{cases}$$

$$p(z_t^D = d' | z_{t-1} = i, z_{t-1}^D = d) = \begin{cases} \frac{D_i(d+1)}{D_i(d)}, & \text{if } d' = d + 1 \\ 1 - \frac{D_i(d+1)}{D_i(d)}, & \text{if } d' = 1 \\ 0, & \text{otherwise} \end{cases}$$

- Complexity per observation increased to $O(K^4D + K^3Dp)$ instead of $O(K^4D^2 + K^3D^2p)$ thanks to deterministic transitions.
Outline

1 Introduction

2 Representation, models, offline algorithms
   - Audio signal representation
   - Clustering with Bregman divergences
   - Hidden Markov Models (HMMs)
   - Hidden Semi-Markov Models (HSMMs)
   - Offline audio segmentation results

3 Online algorithms
   - Online EM
   - Non-probabilistic algorithm
   - Incremental EM
   - Online audio segmentation results
Objective function from probabilistic models

- Mixture model (with $p_i k = 1/K$)
  - Complete-data likelihood
    \[ p(x, z; \mu) = \prod_{i=1}^{n} p(z_i) p(x_i | z_i; \mu) \]
  - Objective ($= - \log p(x, z; \mu) + C$)
    \[ \ell(z, \theta) = \sum_{i=1}^{n} D_\psi(x_i, \mu z_i) \]

- HMM
  - Complete-data likelihood
    \[ p(x_{1:T}, z_{1:T}; \mu) = p(z_1) \prod_{t=2}^{T} p(z_t | z_{t-1}) \prod_{t=1}^{T} p(x_t | z_t; \mu) \]
  - Objective
    \[ \ell(z_{1:T}, \mu) = \frac{1}{T} \sum_{t \geq 1} D_\psi(x_t, \mu z_t) + \frac{\lambda_1}{T} \sum_{t \geq 2} d(z_{t-1}, z_t) \]
Online objective

- Online objective:
  \[ f_T(\mu) := \min_{z_1:T} \ell(z_1:T, \mu) \]

- New upper bound (majorizing surrogate) at time \( t \):
  \[ \hat{f}_t(\mu) := \frac{1}{t} \sum_{i=1}^{t} D_\psi(x_i, \mu z_i) + \frac{\lambda_1}{t} \sum_{i=2}^{t} d(z_{i-1}, z_i) \]

- At time \( t \):
  - \( z_{1:t-1} \) fixed from past
  - E-step: \( z_t = j = \arg \min_k D_\psi(x_t, \mu_k) + \lambda_1 d(z_{t-1}, k) \)
  - M-step: update cluster \( \mu_j = \mu_j + \frac{1}{n_j} (x_t - \mu_j) \)
Outline

1. Introduction

2. Representation, models, offline algorithms
   - Audio signal representation
   - Clustering with Bregman divergences
   - Hidden Markov Models (HMMs)
   - Hidden Semi-Markov Models (HSMMs)
   - Offline audio segmentation results

3. Online algorithms
   - Online EM
   - Non-probabilistic algorithm
   - Incremental EM
   - Online audio segmentation results
Incremental EM for i.i.d. data (Neal and Hinton, 1998)

- EM = maximize lower bounds

\[ f(\theta) = p(x; \theta) \geq \sum_z q(z) \log \frac{p(x, z; \theta)}{q(z)}. \]

- Maximizer \( q(z) = \prod_i p(z_i|x_i; \theta) \), limit to \( \prod_i q_i(z_i) \)

- Minorizing surrogates:

\[ \hat{f}_n(\theta) = \frac{1}{n} \sum_{i=1}^{n} \sum_{z_i} q_i(z_i) \log \frac{p(x_i, z_i; \theta)}{q_i(z_i)} \]

- Repeat: update single \( q_i \) (E-step), maximize \( (1/n) \mathbb{E}_q[\log p(x, z)] \)

- Can be expressed in terms of sufficient statistics
Incremental EM for HMMs

- Only consider lower bounds with $q(z_1:T) = q_1(z_1) \prod_{t \geq 2} q_t(z_t|z_{t-1})$
- Surrogates:

$$\hat{f}_T(\theta) = \frac{1}{T} \sum_{t=1}^{T} \left[ \sum_{z_{t-1},z_t} \phi_{t-1}(z_{t-1}) q_t(z_t|z_{t-1}) \log \frac{p(x_t, z_t|z_{t-1}; \theta)}{q_t(z_t|z_{t-1})} \right],$$

with $\phi_t(z_t) := \sum_{z_{t-1}} \phi_{t-1}(z_{t-1}) q(z_t|z_{t-1})$.

- At time $T$:
  - $q_{1:T-1}$, $\phi_{1:T}$ fixed from past
  - E-step: $q_T(z_T|z_{T-1}) = p(z_T|z_{T-1}, x_T; \theta)$
  - M-step: $\hat{\theta} = \arg \max_{\theta} \hat{f}_T(\theta)$
Experiments on synthetic data

Squared Euclidean distance (left) and KL divergence (right).

\( K = 4, \ p = 5. \)
Experiments on synthetic data

Squared Euclidean distance (left) and KL divergence (right).

\( K = 20, \ p = 5. \)
Experiments on synthetic data

Squared Euclidian distance (left) and KL divergence (right).

\[ K = 20, \ p = 100. \]
Outline

1 Introduction

2 Representation, models, offline algorithms
   - Audio signal representation
   - Clustering with Bregman divergences
   - Hidden Markov Models (HMMs)
   - Hidden Semi-Markov Models (HSMMs)
   - Offline audio segmentation results

3 Online algorithms
   - Online EM
   - Non-probabilistic algorithm
   - Incremental EM
   - Online audio segmentation results
Online EM for HMM vs HSMM

Online EM for HMM/HSMM on Bach. $K = 10$, $NB(30, 0.6)$ (mean 20).
Online EM for HMM vs HSMM

Online EM for HMM/HSMM on Bach. $K = 10$, $NB(30, 0.6)$ (mean 20).
Online vs incremental EM for HMM

![Graph showing HMM online EM filter, HMM online EM smoothing, HMM incremental EM filter, HMM incremental EM smoothing, and ground truth.]
Online vs incremental EM for HMM

HMM online EM filter

HMM online EM smoothing

HMM incremental EM filter

HMM incremental EM smoothing

ground truth

Spectrogram
Scenes segmentation

Dropping keys and closing doors (from office live dataset). $K = 10$
Scenes segmentation

Telephone ringing and coughing sounds (from office live dataset). $K = 10$
Scenes segmentation

Telephone ringing and coughing sounds (from office live dataset). $K = 10$
Conclusion

- Joint segmentation and clustering: challenging task
- Offline algorithms perform well
- Harder task for online algorithms, but results improve over time
- Can be used for adaptive estimation (e.g., note templates in Antescofo score-following system)

Main contributions:
- Extension of online EM algorithm to HSMMs thanks to new parameterization
- Incremental optimization algorithms for HMMs (EM and non-probabilistic)
- Applications to audio segmentation, potential improvements in Antescofo.
References


