K-shot Learning of Acoustic Context

Ivan Bocharov, Tjalling Tjalkens and Bert de Vries

Eindhoven University of Technology, the Netherlands

Email bert.de.vries@tue.nl

NIPS-2017 ML4AUDIO workshop, 8-Dec 2017

Use Case / Problem Statement

Approach: probabilistic modeling

ACOUSTIC MODEL SPECIFICATION

– Define a generative probabilistic model for acoustic signals that contains scenes as latent states.

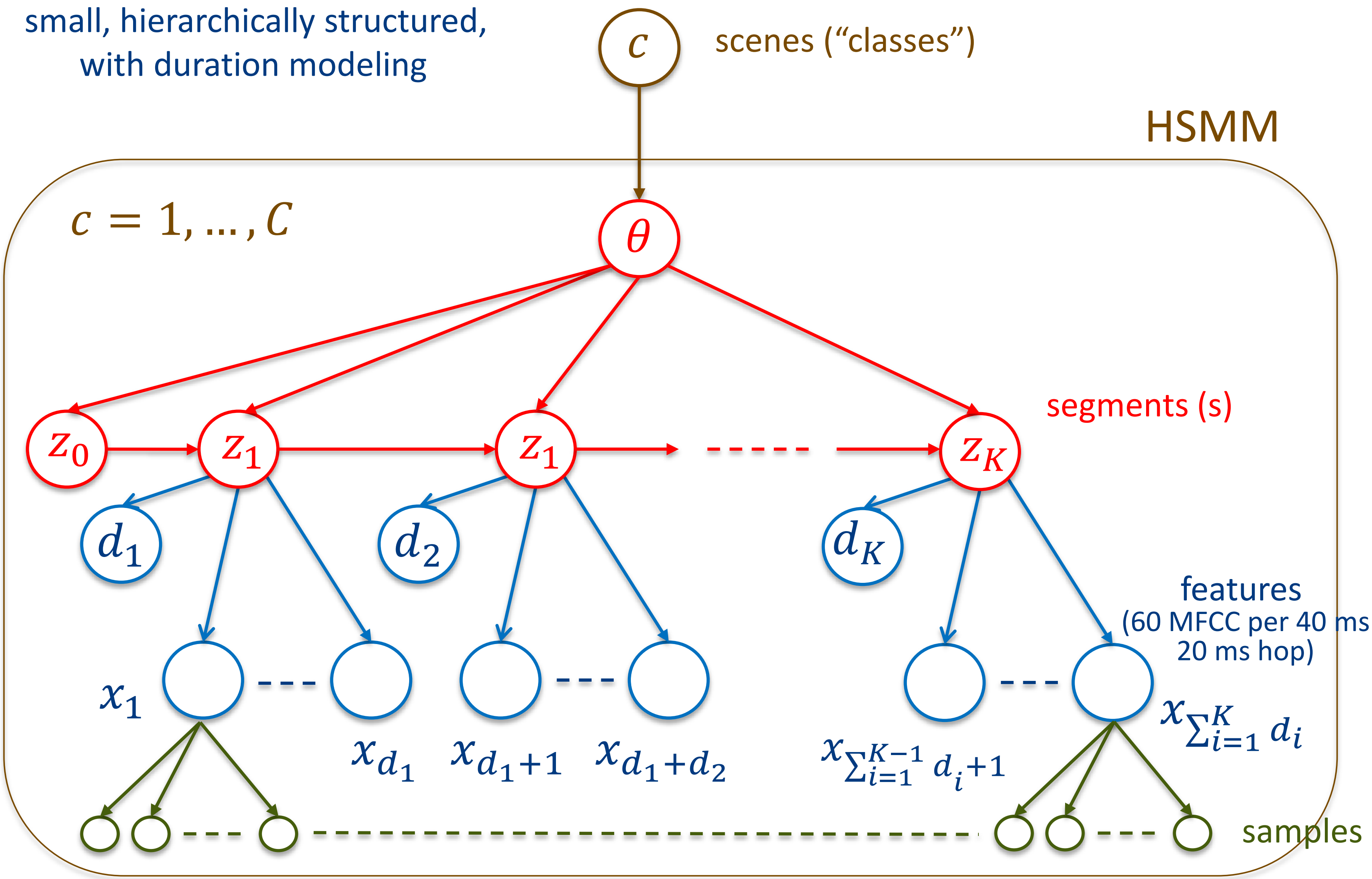
TRAINING

1. **“Representation training”:** Unsupervised offline training on a large database of acoustic signals across many scenes
2. **Train new scenes**: Continue with supervised training on an online recorded small set of scene-labeled waveforms

CLASSIFICATION

– Goal: assign future streaming acoustic data to the correct (or similar) scenes

(Mixture of) Hidden Semi-Markov Models



generative model:

$$p(x,d,z,c,\theta) = \underbrace{p_c(x,d,z|\theta)}_{\text{dynamics}} \; \underbrace{p(\theta|c)}_{\text{parameters}} \; \underbrace{p(c)}_{\text{class prior}} \; .$$

dynamics:

$$\begin{aligned}
p_c(x,d,z|\theta) &= p_c(x|z,d,\theta)\, p_c(d|z,\theta)\, p_c(z|\theta) \\
&= p_c(z_0) \prod_{k=1}^{K} p_c(x_{t_k:(t_k+d_k-1)}|z_k,d_k,\theta)\, p_c(d_k|z_k,\theta)\, p_c(z_k|z_{k-1},\theta) \\
&= p_c(z_0) \prod_{k=1}^{K} \left(\prod_{t=t_k}^{t_k+d_k-1} p_c(x_t|z_k,\theta) \right) \cdot p_c(d_k|z_k,\theta) \cdot p_c(z_k|z_{k-1},\theta) \\
&= p_c(z_0) \prod_{k=1}^{K} \left(\prod_{t=t_k}^{t_k+d_k-1} \underbrace{\mathcal{N}\left(x_t \,\middle|\, \mu^{(c,z_k)}, \Sigma^{(c,z_k)}\right)}_{\text{observation}} \right) \underbrace{\mathrm{Pois}\left(d_k \,\middle|\, \lambda^{(c,z_k)}\right)}_{\text{segment duration}} \underbrace{\mathrm{Cat}\left(z_k \,\middle|\, \pi^{(c,z_{k-1})}\right)}_{\text{segment transition}}
\end{aligned}$$

parameters:

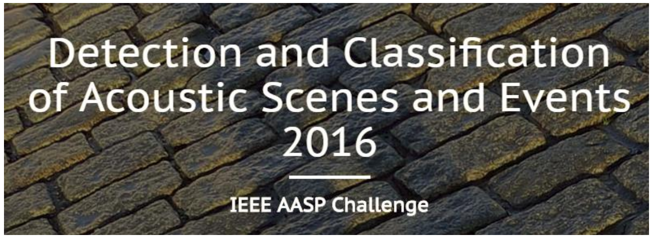
$$\lambda^{(c,z_k)} \sim \mathrm{Gam}\left(a^{(c,z_k)}, b^{(c,z_k)}\right), \quad \mu^{(c,z_k)} \sim \mathcal{N}\left(m^{(c,z_k)}, V^{(c,z_k)}\right)$$

$$\Sigma^{(c,z_k)} \sim \mathcal{W}^{-1}\left(\Psi^{(c,z_k)}, \xi^{(c,z_k)}\right), \quad \pi^{(c,z_{k-1})} \sim \mathrm{Dir}\left(\alpha^{(c)}\right)$$

class prior:

$$p(c) = \mathrm{Cat}\left(c \,\middle|\, \frac{1}{C}, \ldots, \frac{1}{C}\right).$$

Data set: TUT Acoustic Scenes 2016



- Collected by Tampere University of Technology
- 15 acoustic scenes
- ~40 min. of audio per class

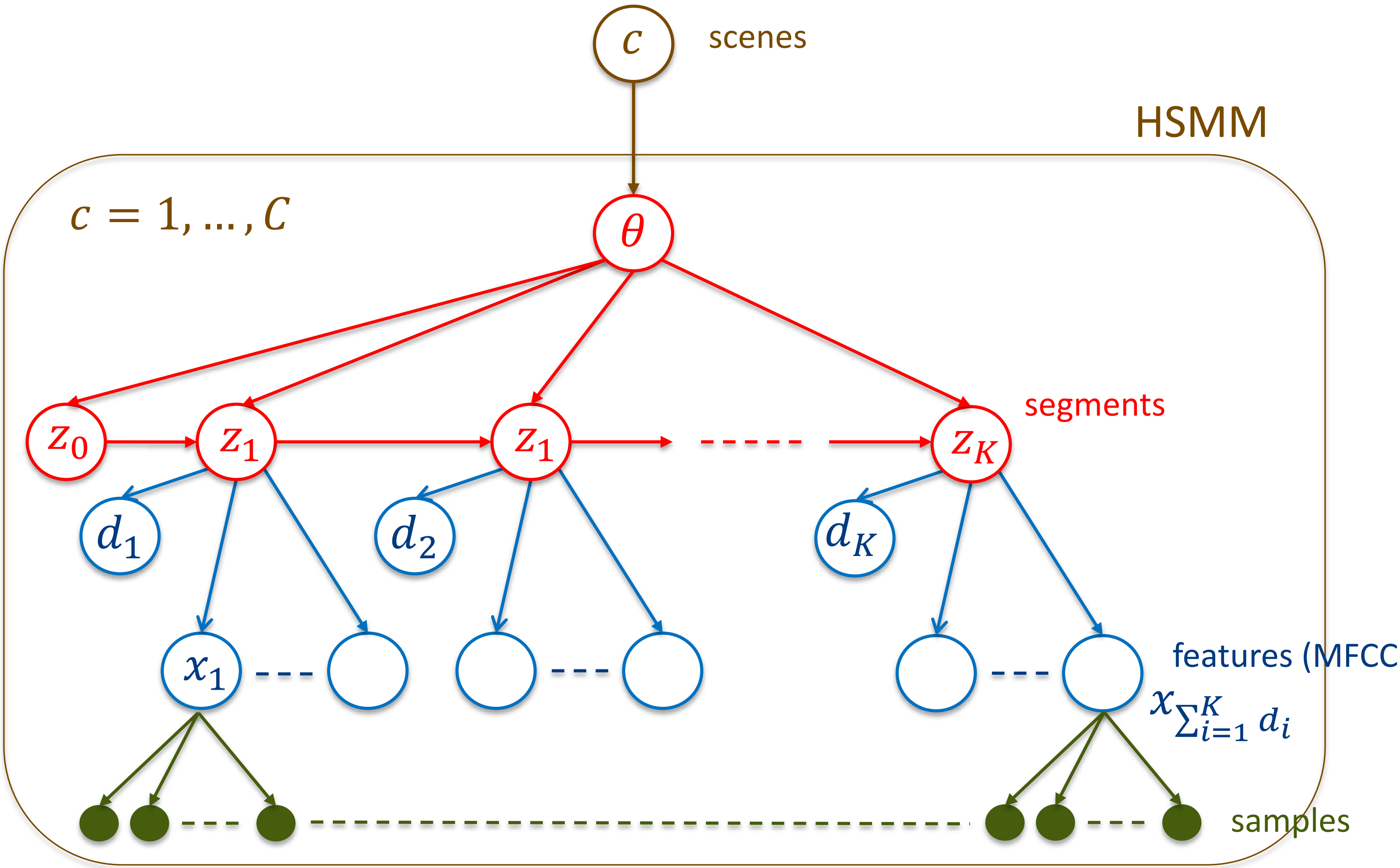




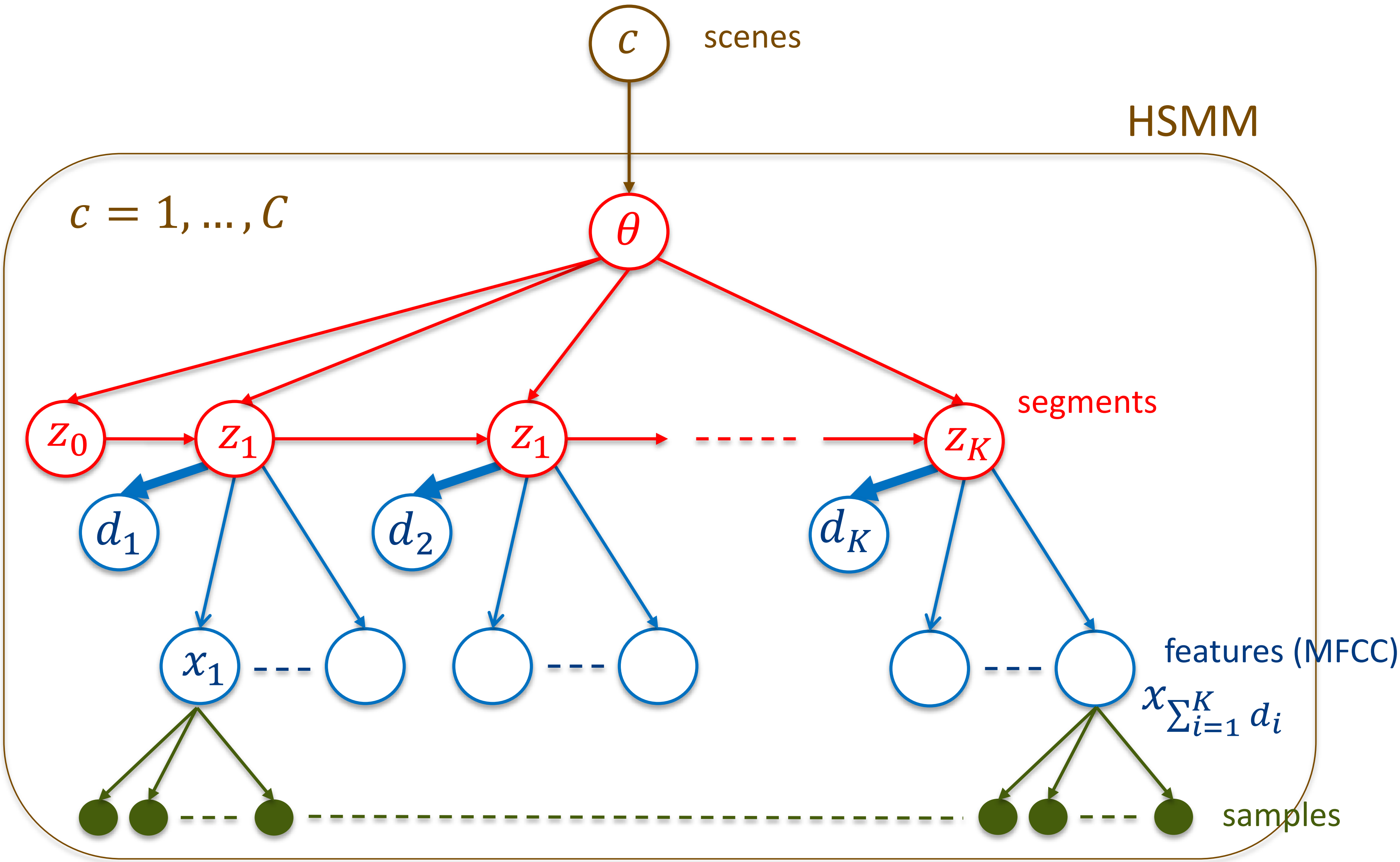

Data Preparation

- **Data set 1**: draw one example (30secs) from each of 11 randomly chosen scenes
- **Data set 2**: draw one example from remaining (4) classes.
- **Classify**: test on remaining examples of data set 2

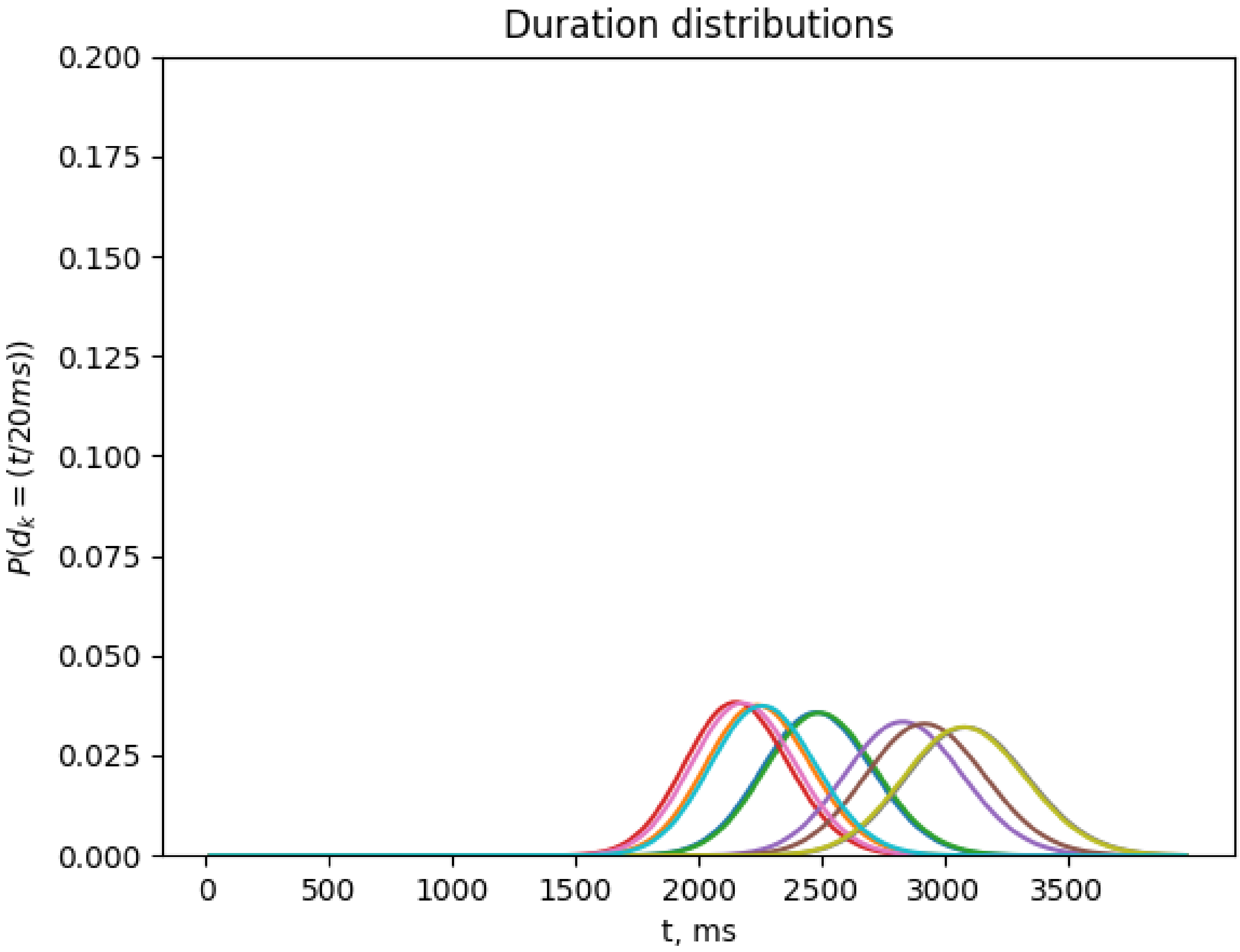
Step 1: Train Duration Models



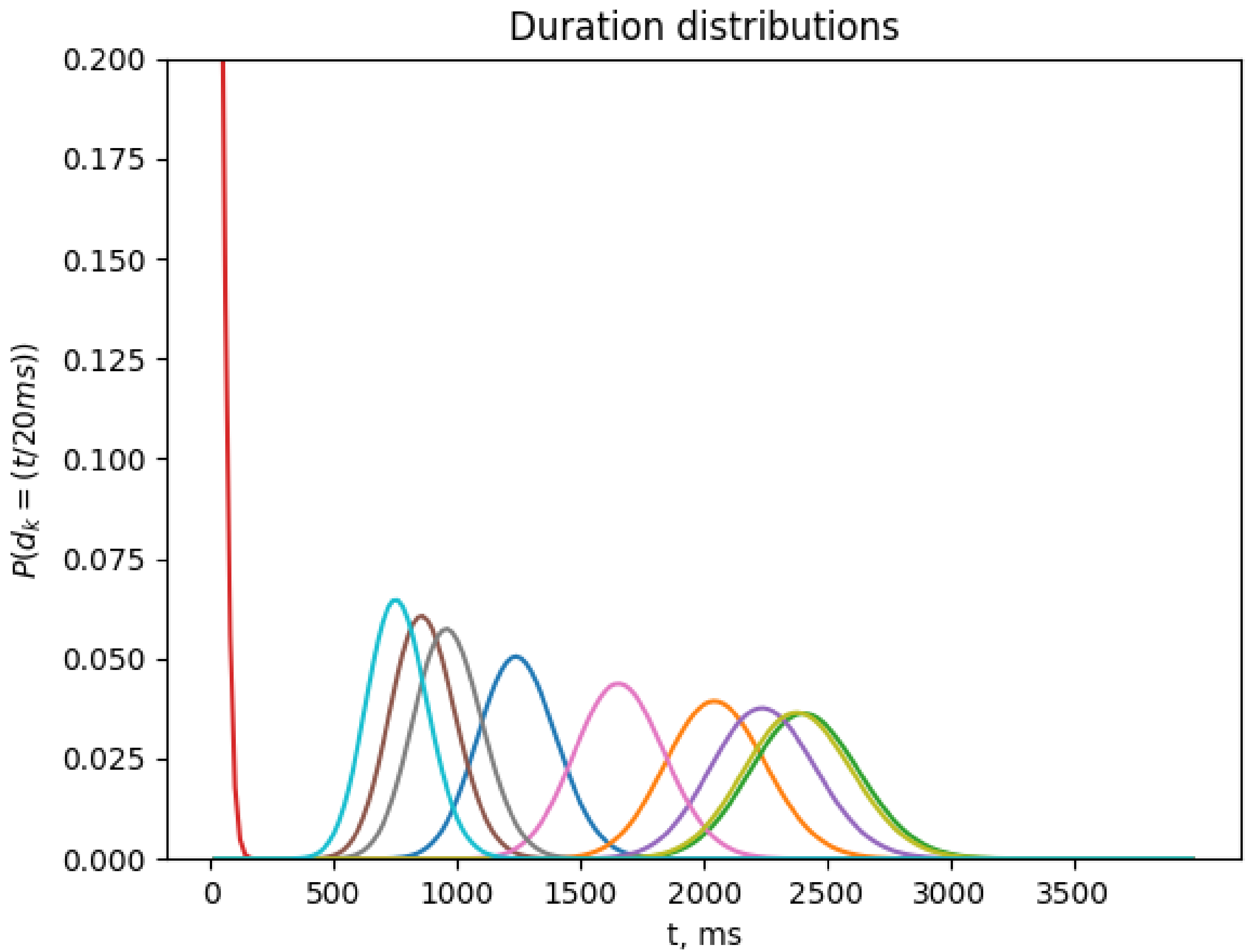
Step 1: Train Duration Models



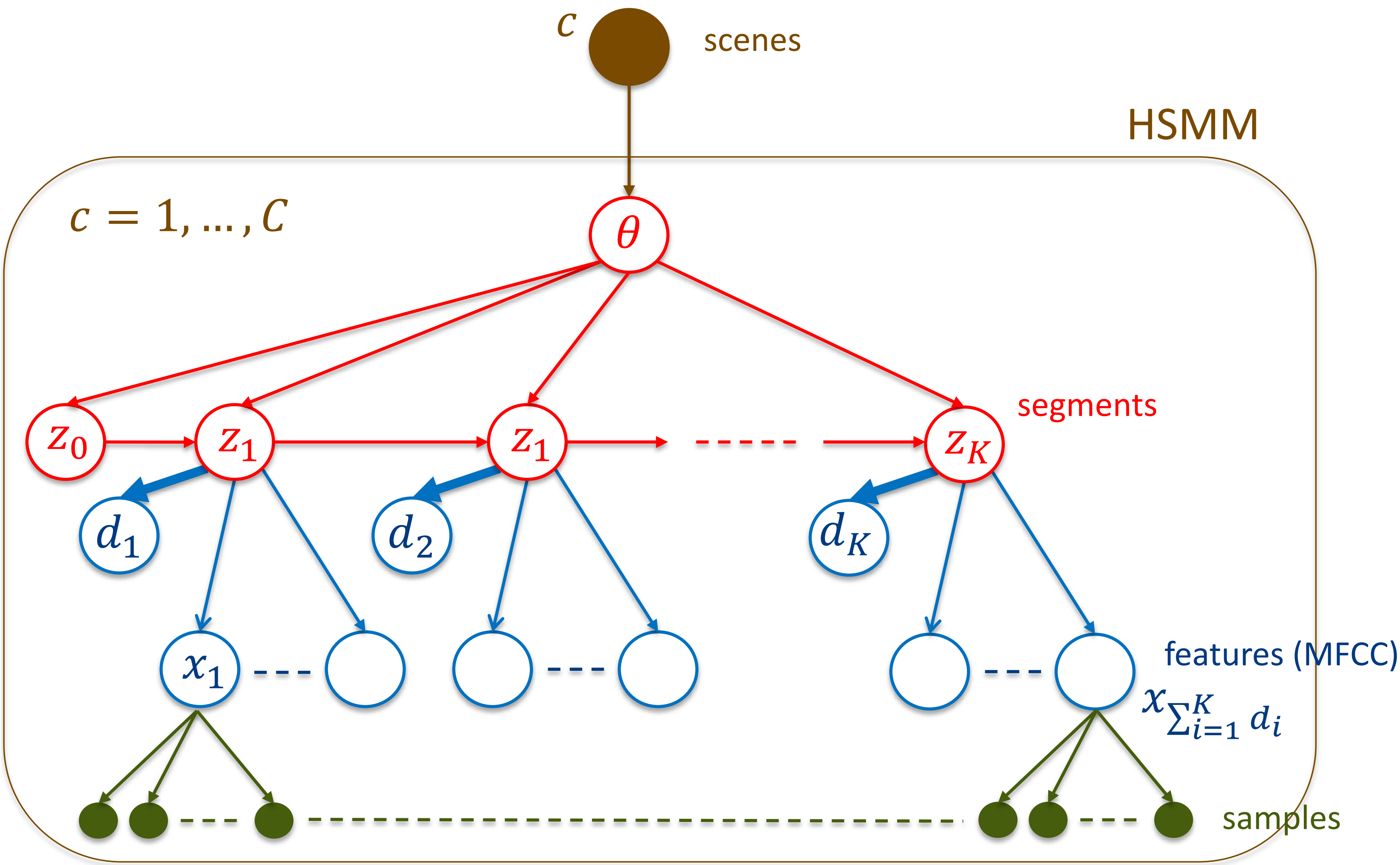
Duration distributions (initialization Pois(.))



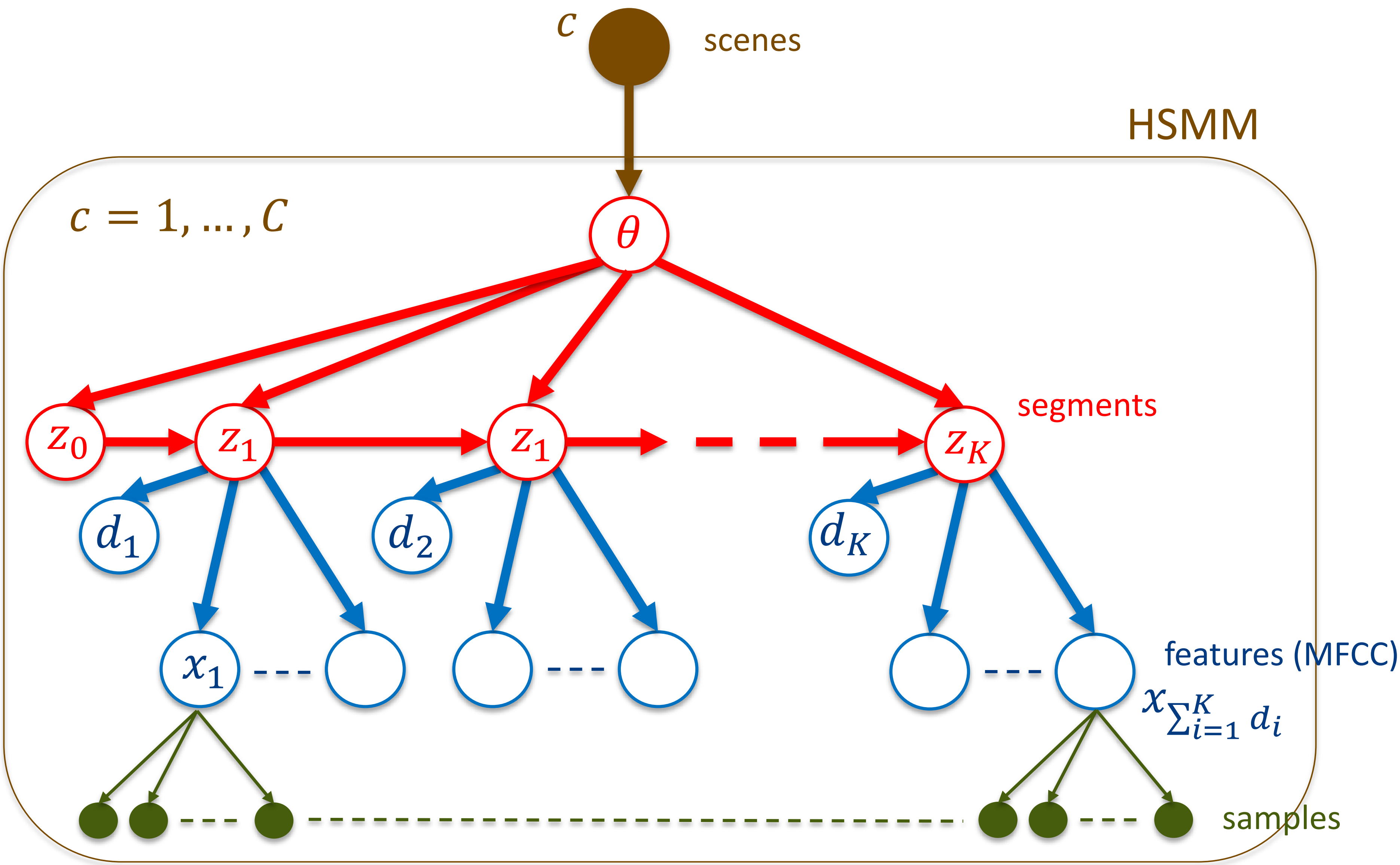
Duration distributions (after training)



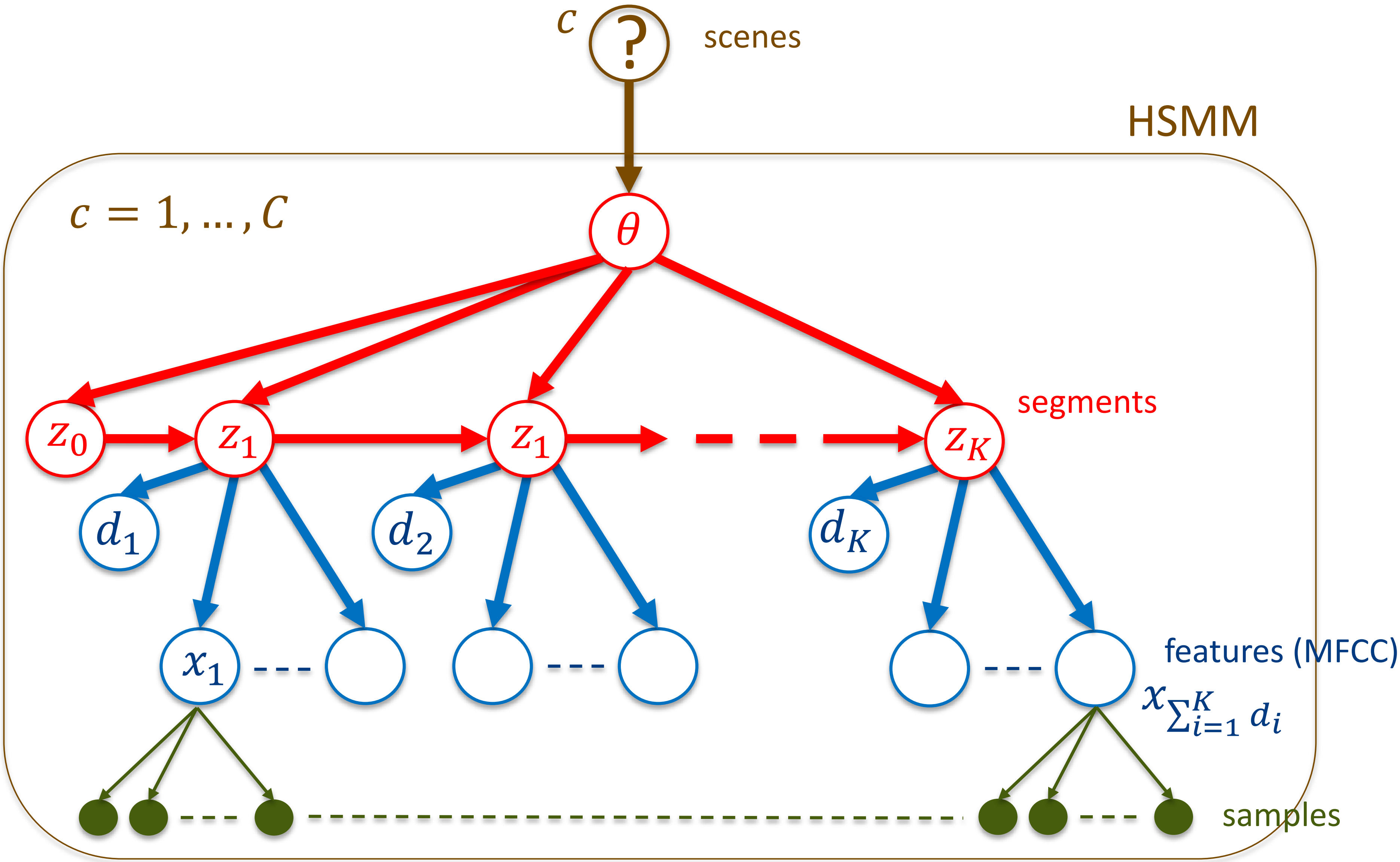
Step 2: One-shot Training



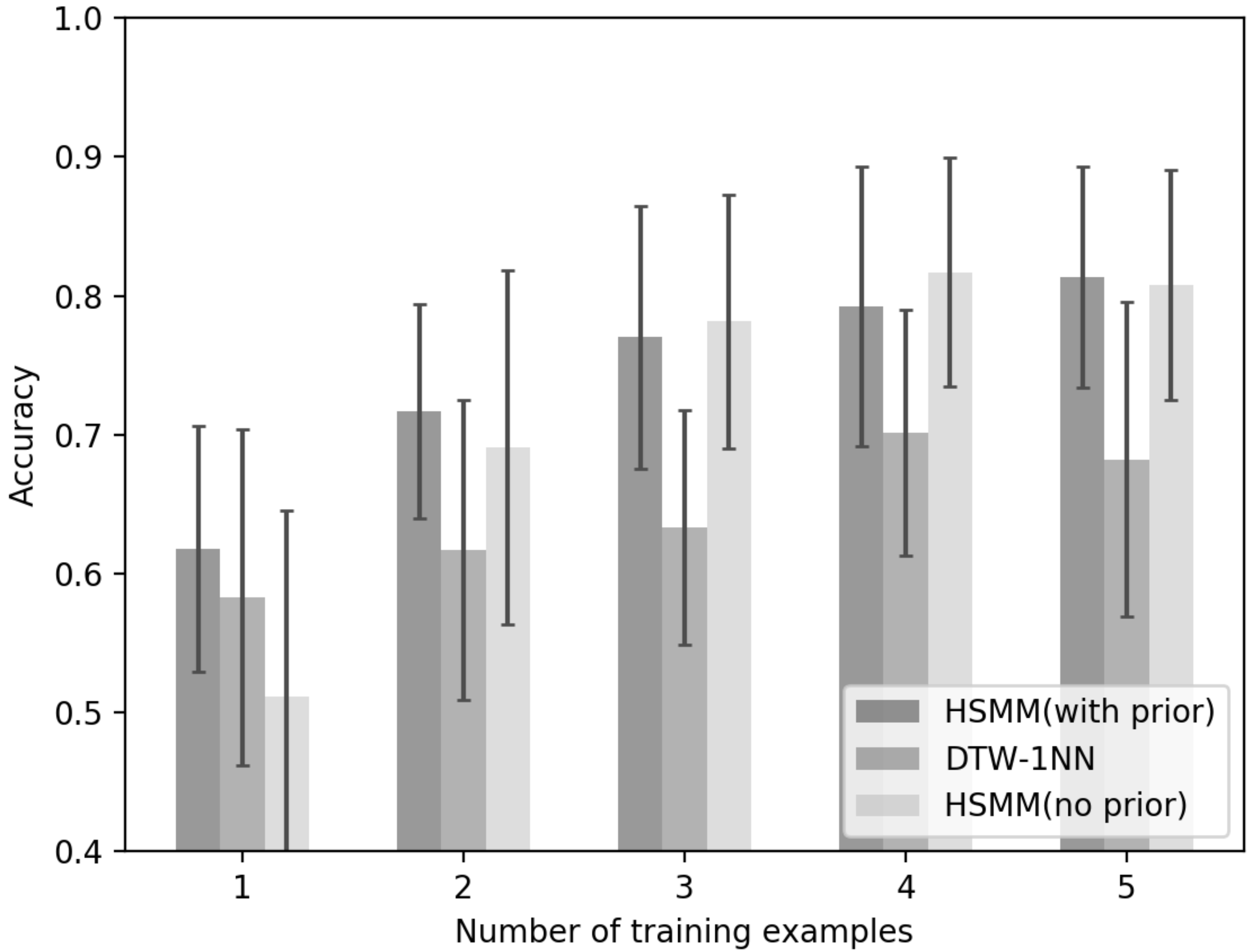
Step 2: One-shot Training



Classification



Results



1.0
0.9
0.8
0.7
0.6
0.5
0.4
Accuracy
1
2
3
4
5
Number of training examples
HSMM(with prior)
DTW-1NN
HSMM(no prior)

Summary and Future Plans

- Ongoing research on in-situ one-shot learning of a personalized acoustic scene classifier
- Use case is hearing aids personalization, but also applicable to urban monitoring, elderly care, etc.
- Generative modeling approach, inspired by one-shot learning work of (a.o.) Brendan Lake et al (2014), Matthew Johnson et al. (2013)
- An HSMM-based probabilistic classifier shows promising performance on one-shot learning task compared to 1NN-DTW.
- Specifically, learned priors for segment duration models parameters helps the classifier to recognize new classes from a single example.
- Future work includes more thorough analysis and exploration of competing models.

Acknowledgements

- Matthew Johnson et al. for Package **Pyhsmm** (@ https://github.com/mattjj/pyhsmm)

Thank you