

# ML Project Proposal

Aniruddha Deshpande & Alexander Birch

## 1. Overview

Apply the robust optimization framework to the Expectation-Maximisation algorithm and test its performance as compared with the regular approach.

## 2. Theoretical framework

Setup:  $\rightarrow y$ : observed data  
 $y \sim p_y(y|x)$

Let  $z$ : complete data  $| z \sim p_z(z|x)$

- $z$  is not observed.
- Choice of  $z$  is made such that  $\max_{x,y} \log p_z(z|x)$  is easier than  $\max_{x,y} \log p_y(y|x)$ .
- $z$  maybe may not have a physical interpretation.
- $z$  is s.t  $y = g(z)$  deterministic function of  $z$

a.

EM alg.:

- 1) choose initialisation:  $\hat{x}(0)$
- 2) E-step: compute  
$$V(x, \hat{x}(t)) = \mathbb{E}_{P_{Z|Y}(z|y; \hat{x}(t))} [\log P_z(z|x)]$$
- 3) M-step:  
$$\hat{x}(t+1) = \arg \max_{x \in \mathcal{X}} V(x, \hat{x}(t))$$

b.

- c. Robustify the maximisation problem by considering different kinds of uncertainty sets and try to derive tractable formulations to solve the problem
- d. Will work specifically for the case of Mixture of Gaussian models, where each data point is chosen randomly from a given set of gaussian distributions with different parameters.

$z = (y, c) : Z$  is complete data containing the data point as well as the class it belongs to. Only  $y$  is observed

$$P(z ; x) = \sum_{c=1}^k p_c \cdot N(y ; \mu_c, \sigma_c) \text{ and given data want to estimate } p_c, \mu_c \text{ and } \sigma_c$$

### 3. Testing on data

- a. Begin by constructing a synthetic dataset with known statistical distributions and testing our model's performance on it
- b. Once more confident in our model, test in an area where the EM algorithm is currently used in literature, (e.g. Robotics AI, NLP, Speech Recognition, Computer Vision, Decoding Hidden Markov Models) and implement our robust algorithm to (hopefully) observe improvement.