EM for Mixtures of Gaussians

0. Initialize parameters
\[ \Theta = \{ \pi, \{ \mu^{(k)}, \Sigma^{(k)} \} \} \]

1. E-Step
Fix soft responsibilities:
\[ r_k^{(n)} = P(z^{(n)} = k | x^{(n)}, \Theta) \]

2. M-Step
Fit \( \Theta \) using \( \{ r_k^{(n)} \} \)
\[ r_k = \sum_n r_k^{(n)}, \quad \pi_k = \frac{r_k}{N}, \quad \mu^{(k)} = \frac{1}{r_k} \sum_n r_k^{(n)} x^{(n)} \]

3. Go to 1. or stop:
\[ \Sigma^{(k)} = \frac{1}{r_k} \sum_n r_k^{(n)} (x^{(n)} - \mu^{(k)})(x^{(n)} - \mu^{(k)}) \]
Use in "Stanley" car

Component 1 (road)
Component 2 (background)

Feature space describing pixels around a given location
Interpretation

We are using an approx. posterior

\[ Q(z^{(n)} = k) = \Gamma_k^{(n)} = P(z^{(n)} = k | x^{(n)}, \Theta^{(old)}) \]

Only correct when \( \Theta = \Theta^{(old)} \)

Compare distributions

\[ D_{KL}(Q || P) = \sum_z Q(z) \log \frac{Q(z)}{P(z|x, \Theta)} \geq 0 \]

Use \( P(z|x, \Theta) = \frac{P(z, x|\Theta)}{P(x|\Theta)} \leftarrow \text{Likelihood} \)

\[ \Rightarrow \sum_z Q(z) \log \frac{Q(z)}{P(x, z|\Theta)} \geq - \log P(x|\Theta) \]

\[ \Rightarrow \text{Bound on log-likelihood} \]

Tight if \( \Theta = \Theta^{(old)} \) used to set \( Q \)
Bound-based optimizer

Can't have loose bound if want to guarantee improvement

Lower bound at time \( t \)

\[
\log p(x|\theta)
\]
Newton's Method

Cost function $E(w)$

Gradients $g = \nabla_w E(w)$

Hessian $H_{ij} = \frac{\partial^2 E}{\partial w_i \partial w_j}$

Initialize $w(0)$

\[ w(t+1) = w(t) - H^{-1} g \]

If we have a quadratic cost:

\[ E(w) = \frac{1}{2} (w - w^*)^T H (w - w^*) + \text{const.} \]

Here $g = H (w - w^*)$

\[ w(t+1) = w(t) - H^{-1} H (w(t) - w^*) \]

\[ = w^* \]
Why use other optimizers?

- Convergence?
- Not needing to tune...
- SGD can’t give “sparse” solutions
  some $w_a = 0$
L1 Regularization

\[ c(w) = E(w) + \frac{\lambda \sum d |w_d|}{\lambda \|w\|_1} \]

Training error

L2 Regularizer

\[ c(w) = E(w) + \frac{\lambda \sum d |w_d|}{\lambda \|w\|_2} \]
Contours do not cross at optimum:

- Contours of $L_2$ and $E(w)$
- Contours of regularization

$\text{argmin } c(w)$
There are many ways to fit $L^1$. One way are proximal methods:

\[ \tilde{E}(w) = E(w) + \lambda ||w||_1 \]

Can fit in closed form, some $w_d = 0$. 

\[ w_1 = 0 \]
From Bayesian view, predictions are never sparse

\[ p(y|x, D) = \int p(y|x, w) p(w|D) dw \]

\[ p(w_a \neq 0|D) > 0 \]