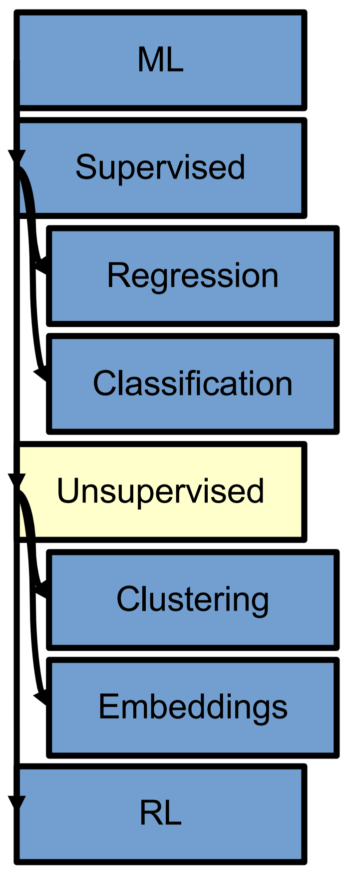
CS181: Introduction to Machine Learning

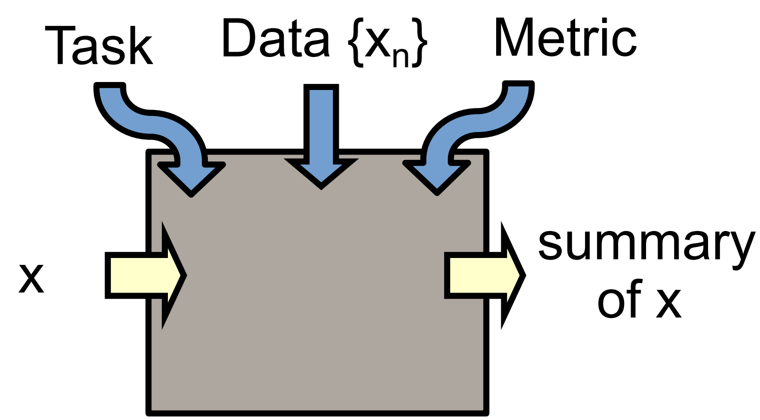
Lecture 14 (Mixture models)

Spring 2021

Finale Doshi-Velez and David C. Parkes
Harvard Computer Science

Unsupervised Learning

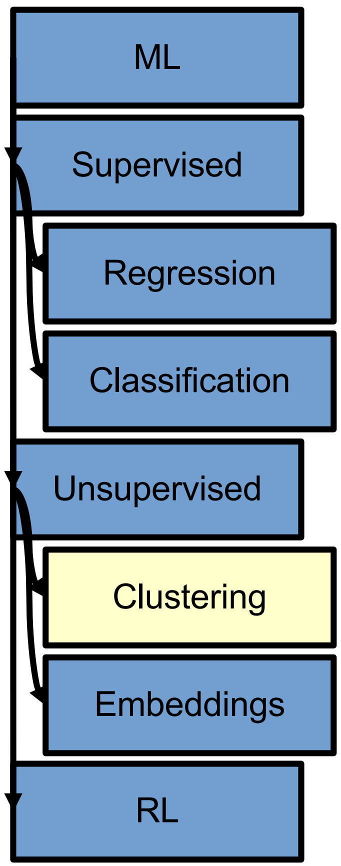


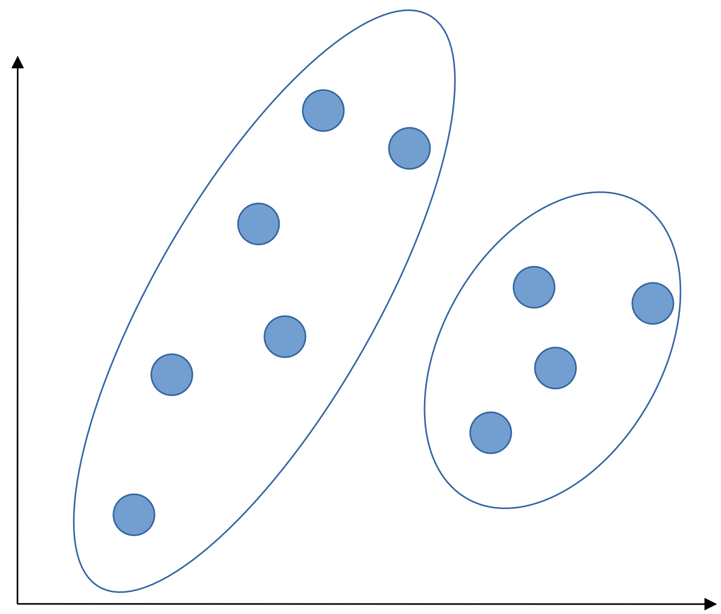


Data D = $\{x_1, x_2, \ldots x_N\}$

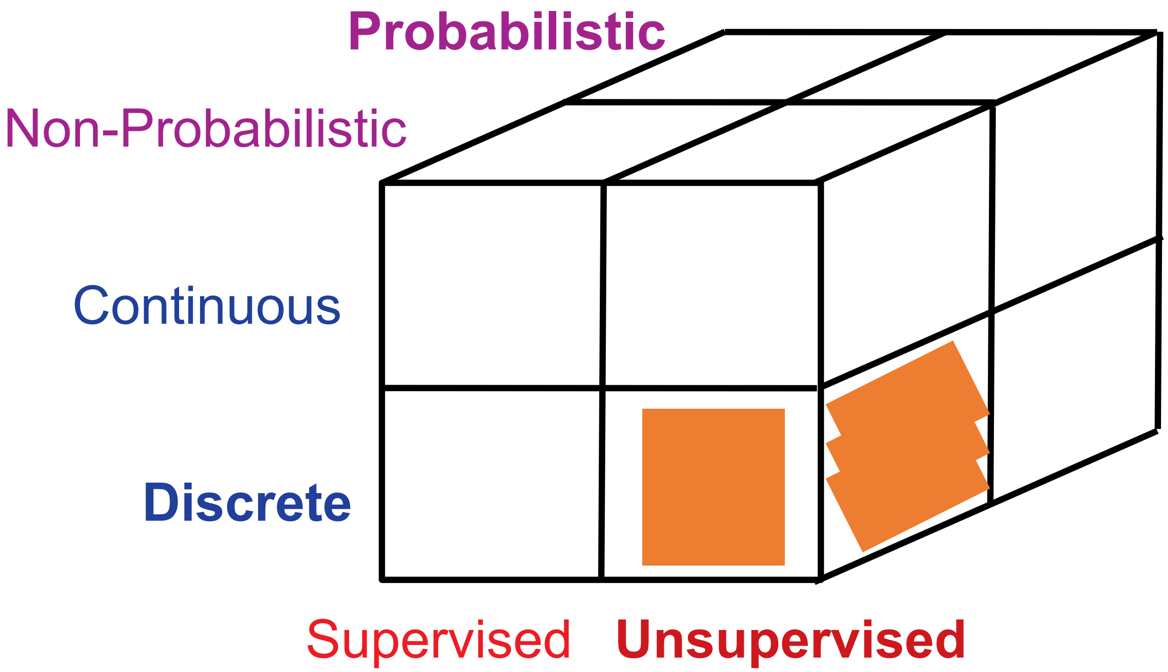
Typical goals: understand, summarize, identify concepts

Unsupervised Learning: Clustering



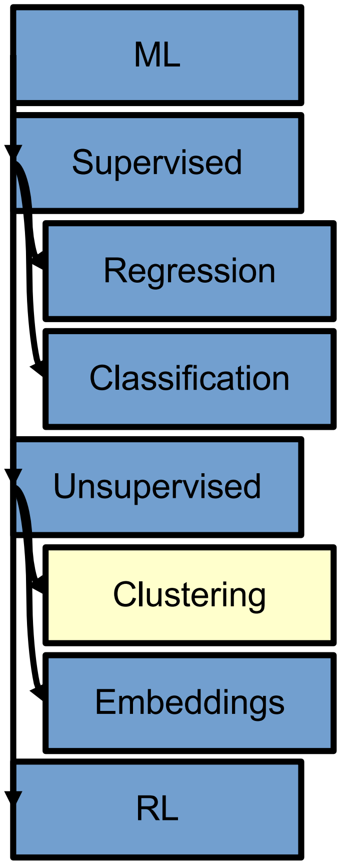


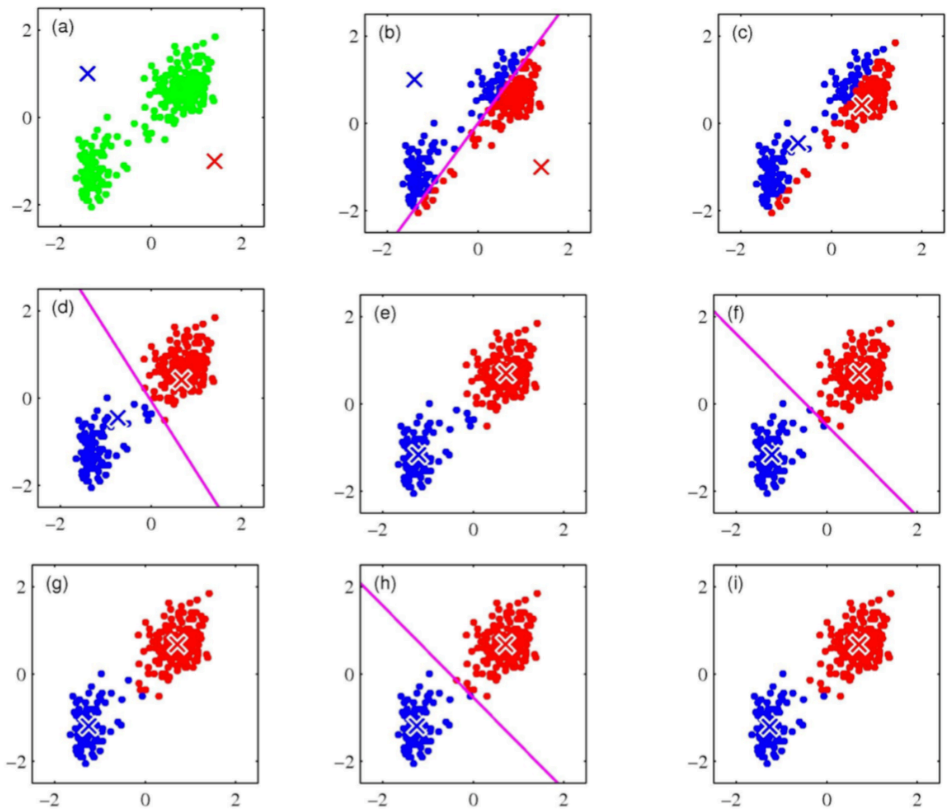
Last Class (1 of 3): The Cube



+ graphical models, reinforcement learning

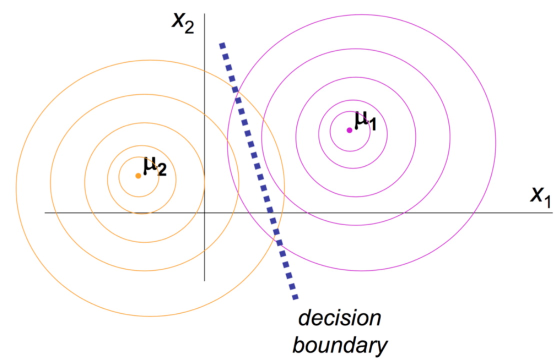
Last Class (2 of 3): K-Means



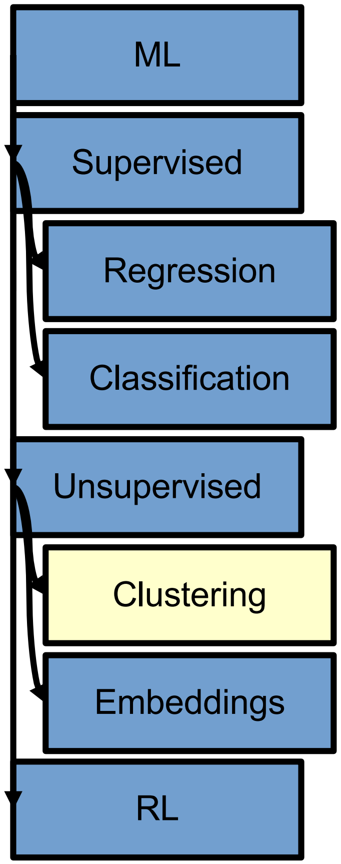


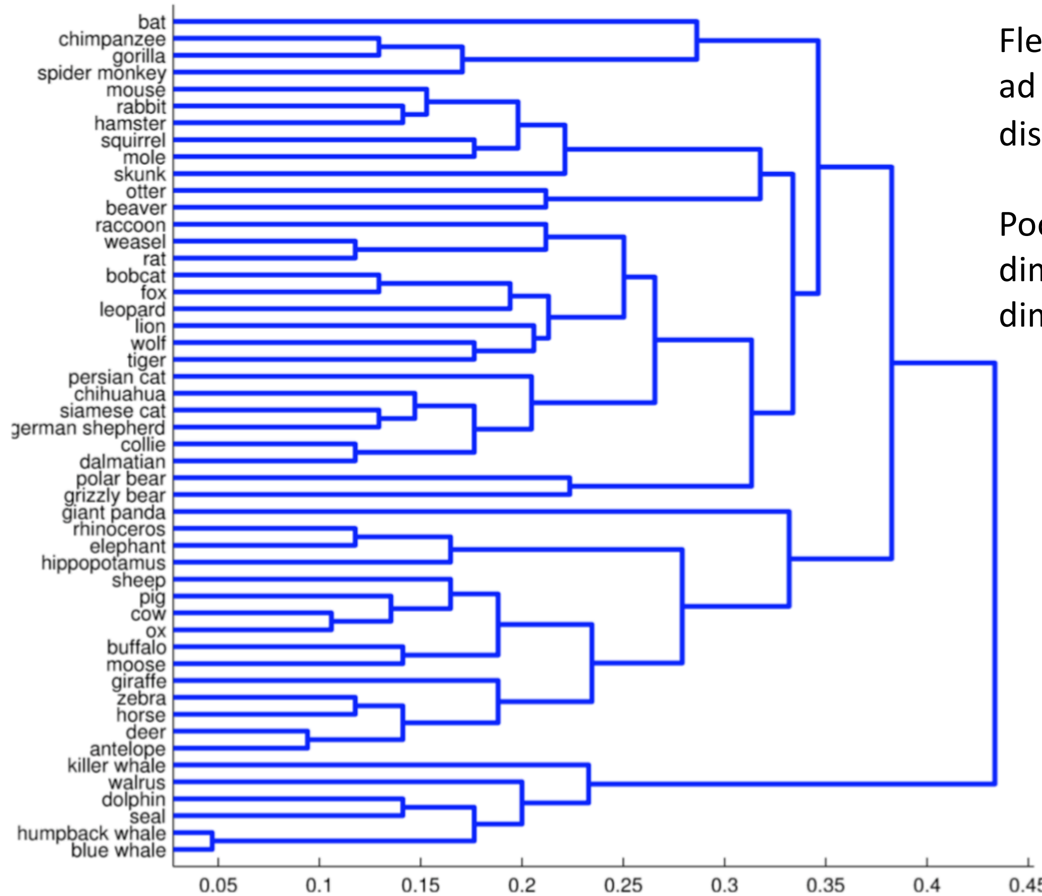
“Old Faithful” Geyser Eruptions (Bishop)

Simple, but inflexible (linear decision boundaries)



Last Class (3 of 3): HAC



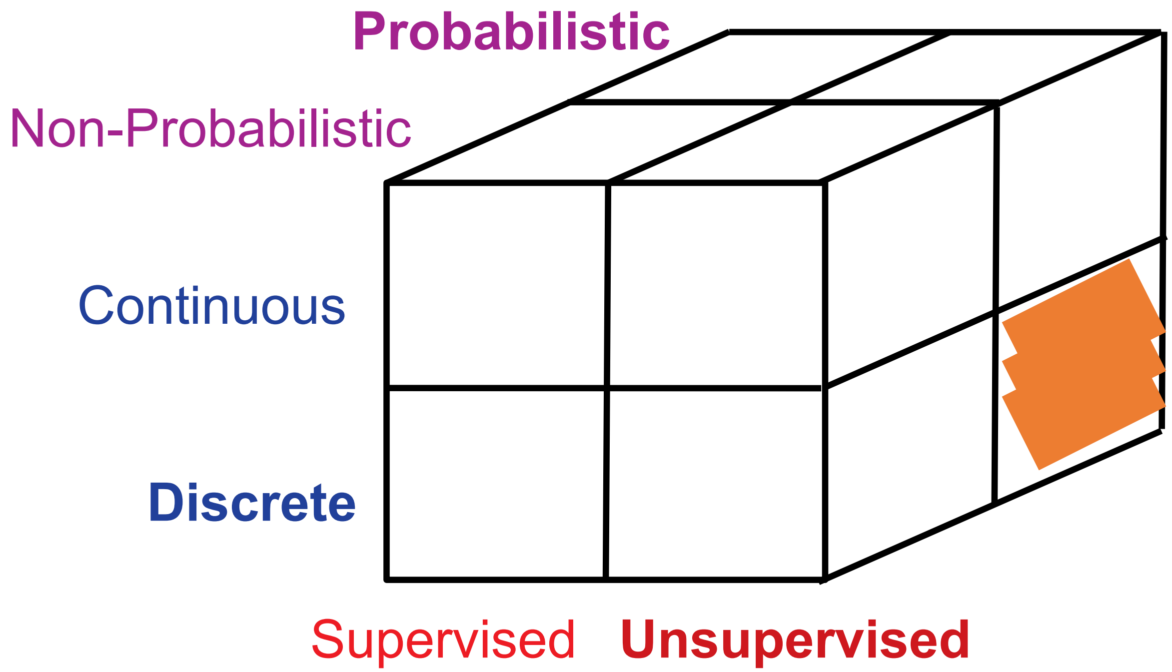


Flexible, but somewhat ad hoc (distance, linkage distance).

Poor performance in high dimensions (curse of dimensionality!)

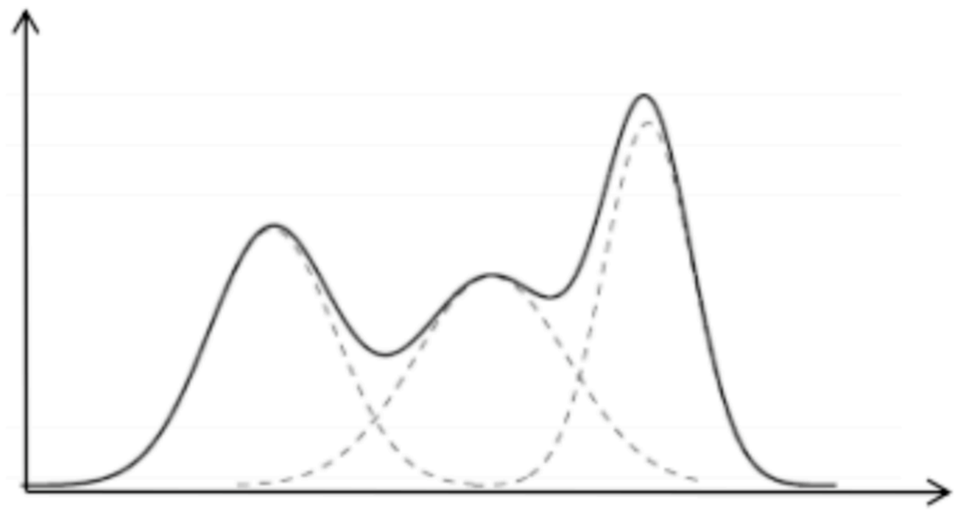
50 animals, 85 binary features (e.g., long neck, water, smelly)

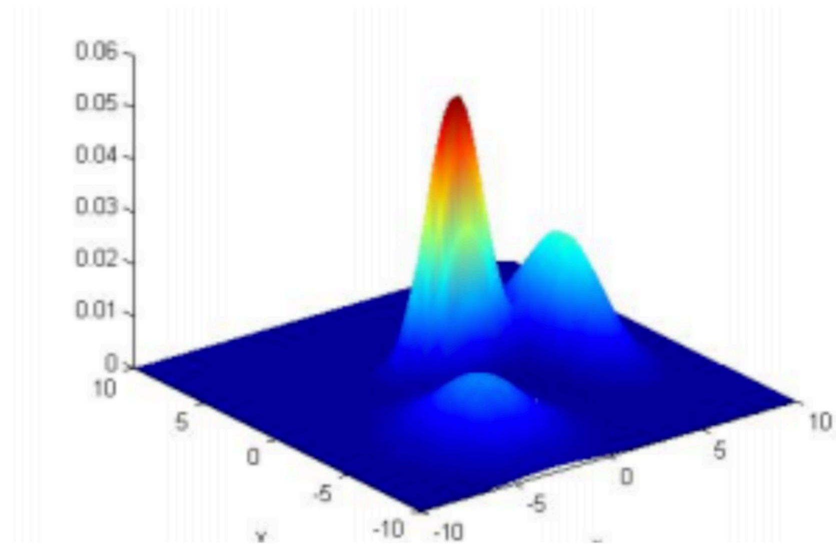
Today (1 of 2): The Cube



+ graphical models, reinforcement learning

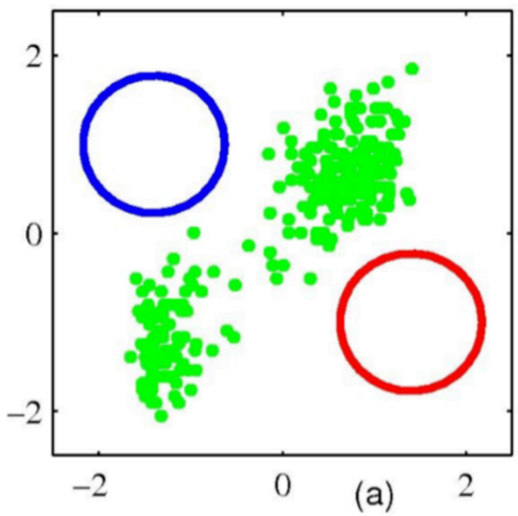
Today (2 of 2): Model data through a mixture of components





Run-through of GMM with E-M for estimation on the Old Faithful Geyser Data

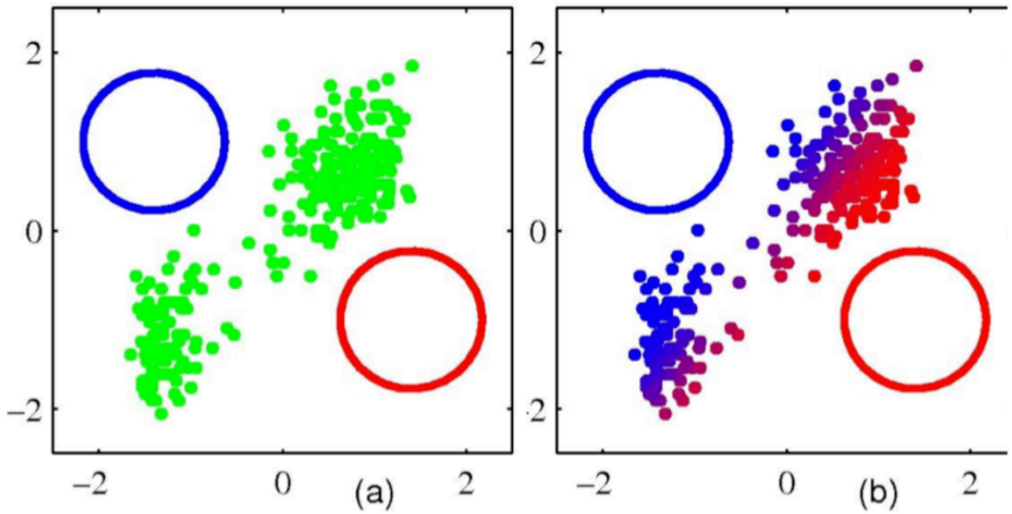
(Bishop. Old Faithful data. 1 st. dev. contours. (b) E step. (c) M step; 2, 5, 20 interations.)



(a)

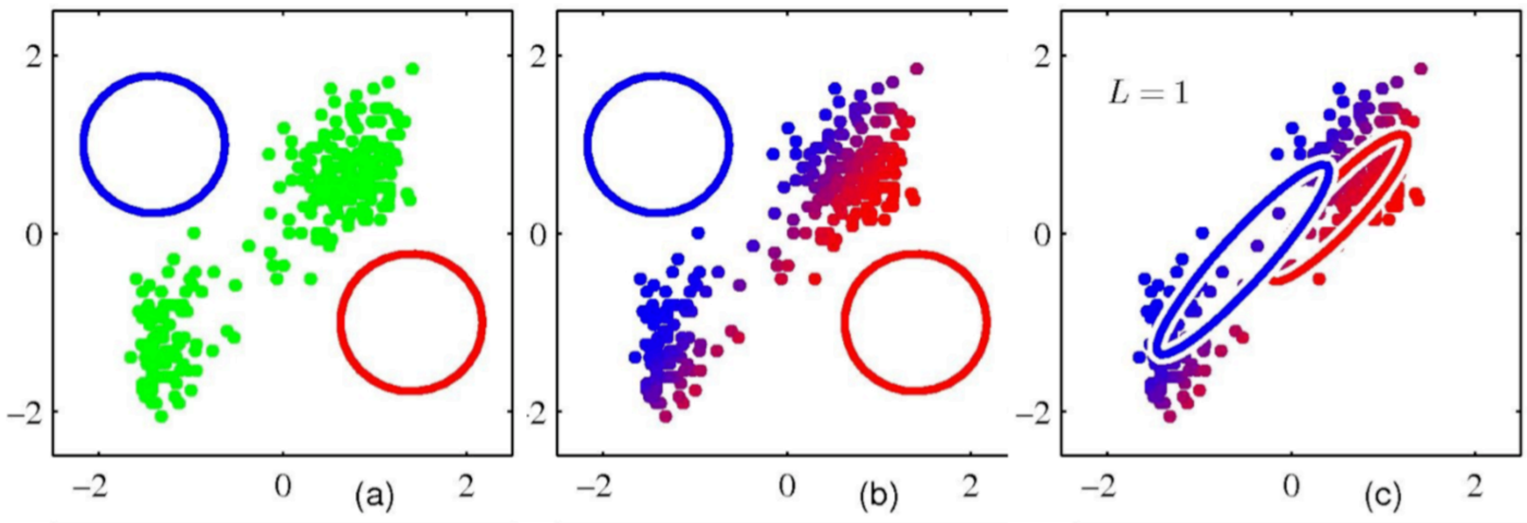
Initialization

(Bishop. Old Faithful data. 1 st. dev. contours. (b) E step. (c) M step; 2, 5, 20 interations.)



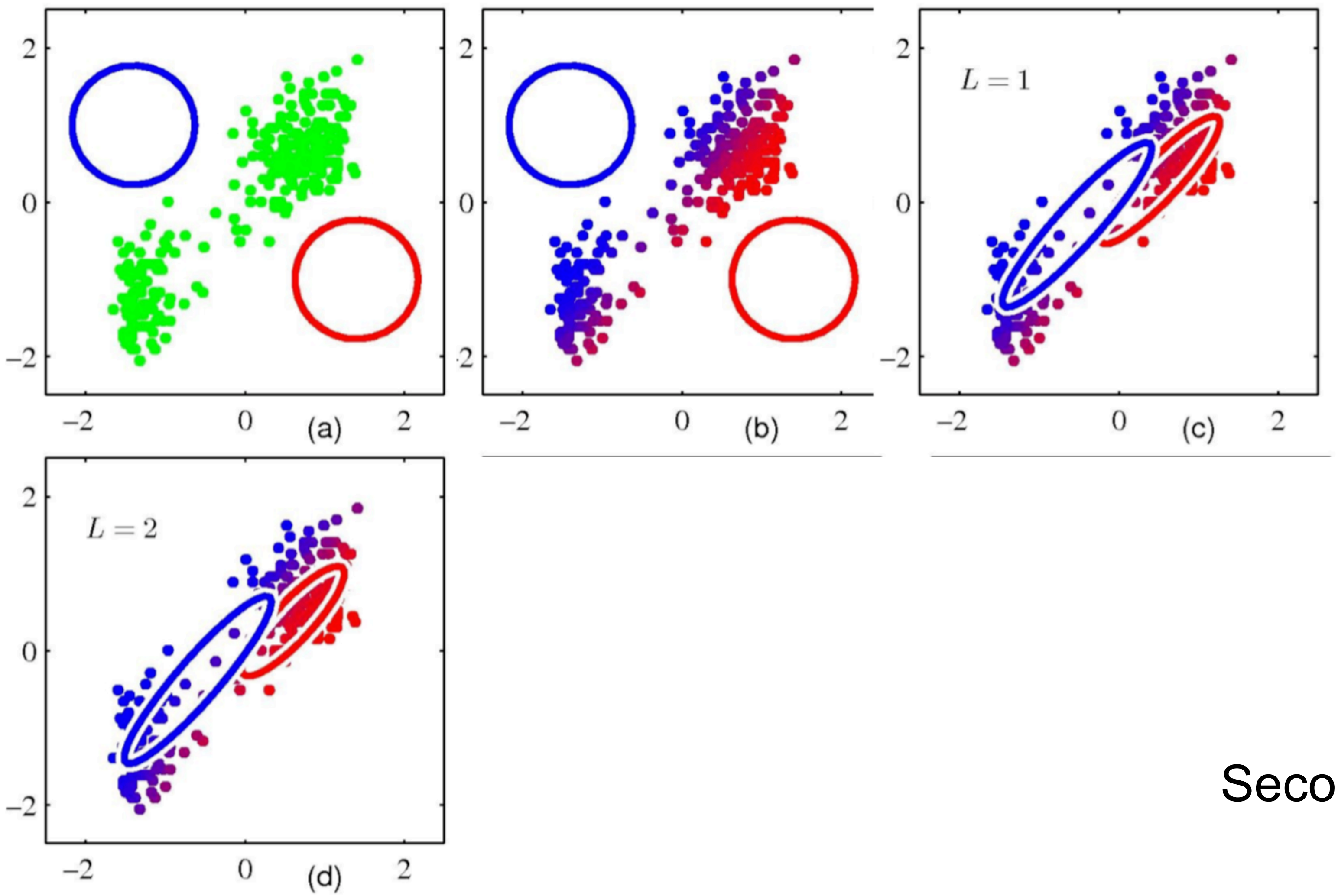
First E-Step

(Bishop. Old Faithful data. 1 st. dev. contours. (b) E step. (c) M step; 2, 5, 20 interations.)



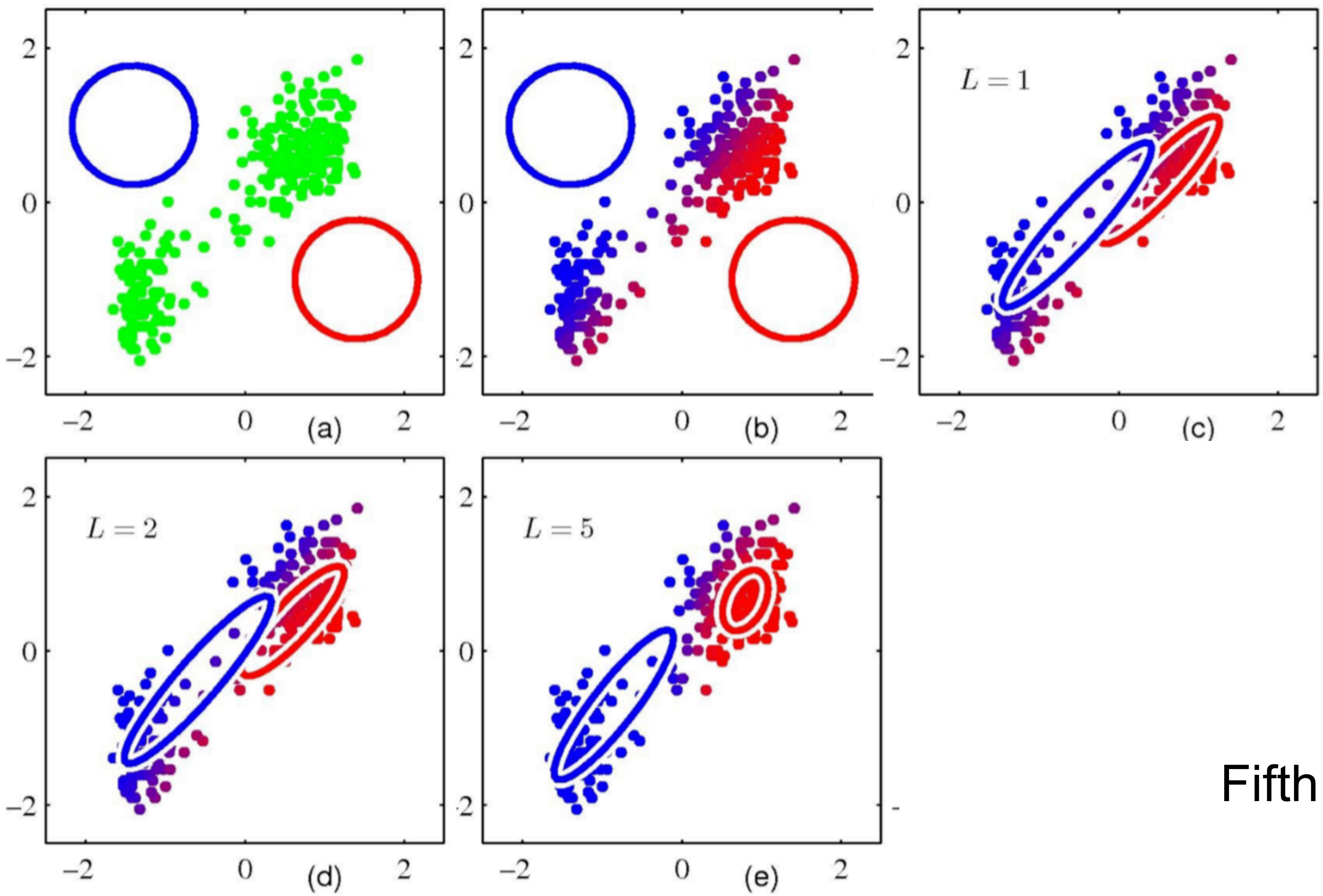
First M-step

(Bishop. Old Faithful data. 1 st. dev. contours. (b) E step. (c) M step; 2, 5, 20 interations.)



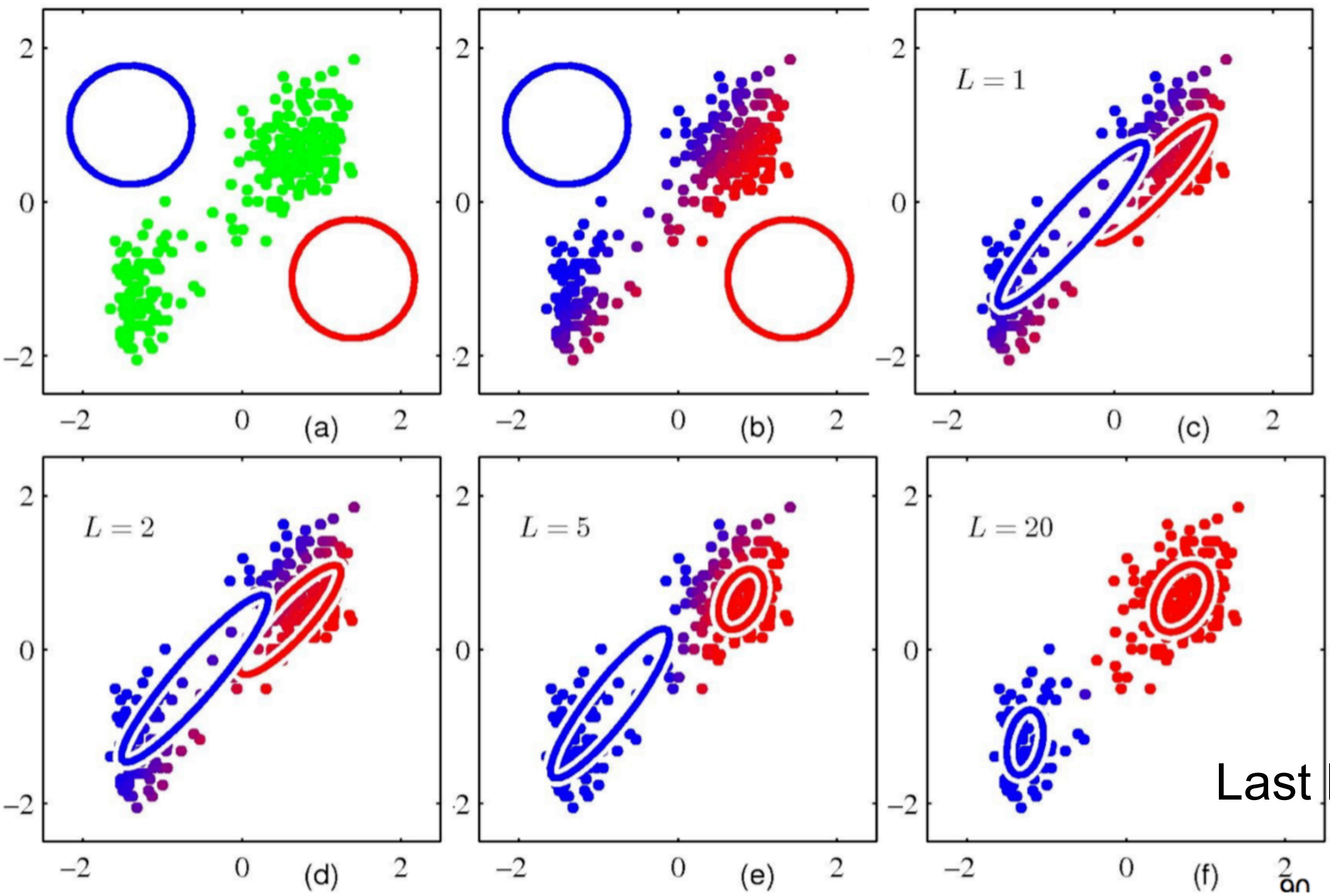
Second M-step

(Bishop. Old Faithful data. 1 st. dev. contours. (b) E step. (c) M step; 2, 5, 20 interations.)



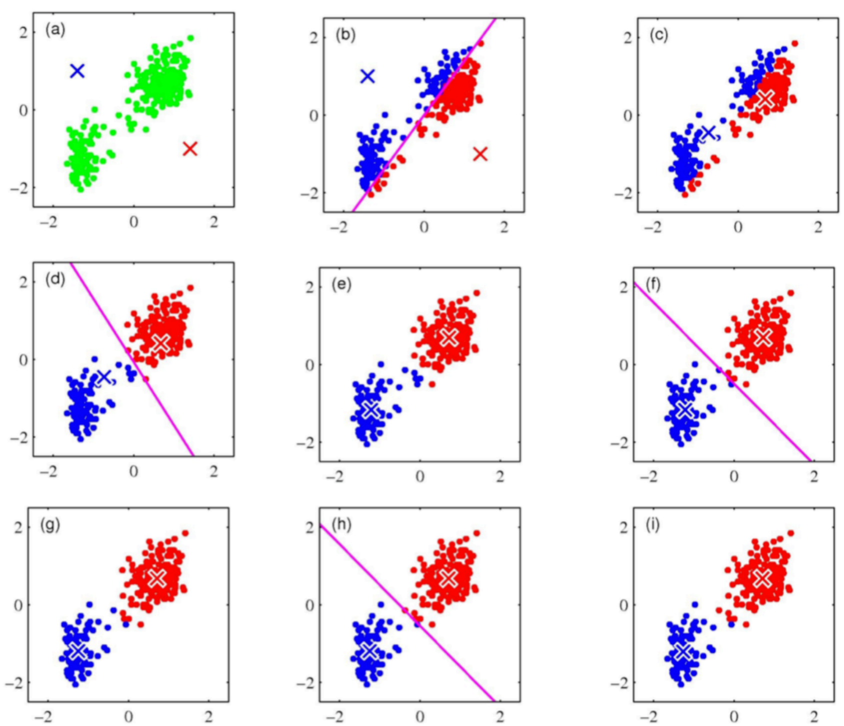
Fifth M-step

(Bishop. Old Faithful data. 1 st. dev. contours. (b) E step. (c) M step; 2, 5, 20 interations.)

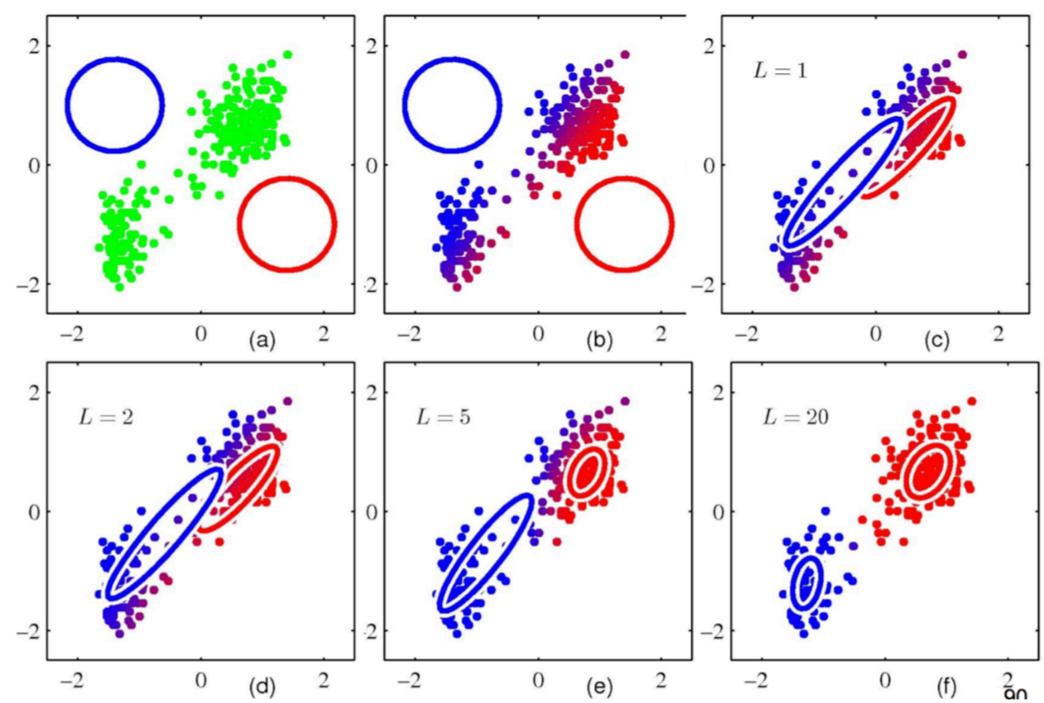


Last M-step

K-Means



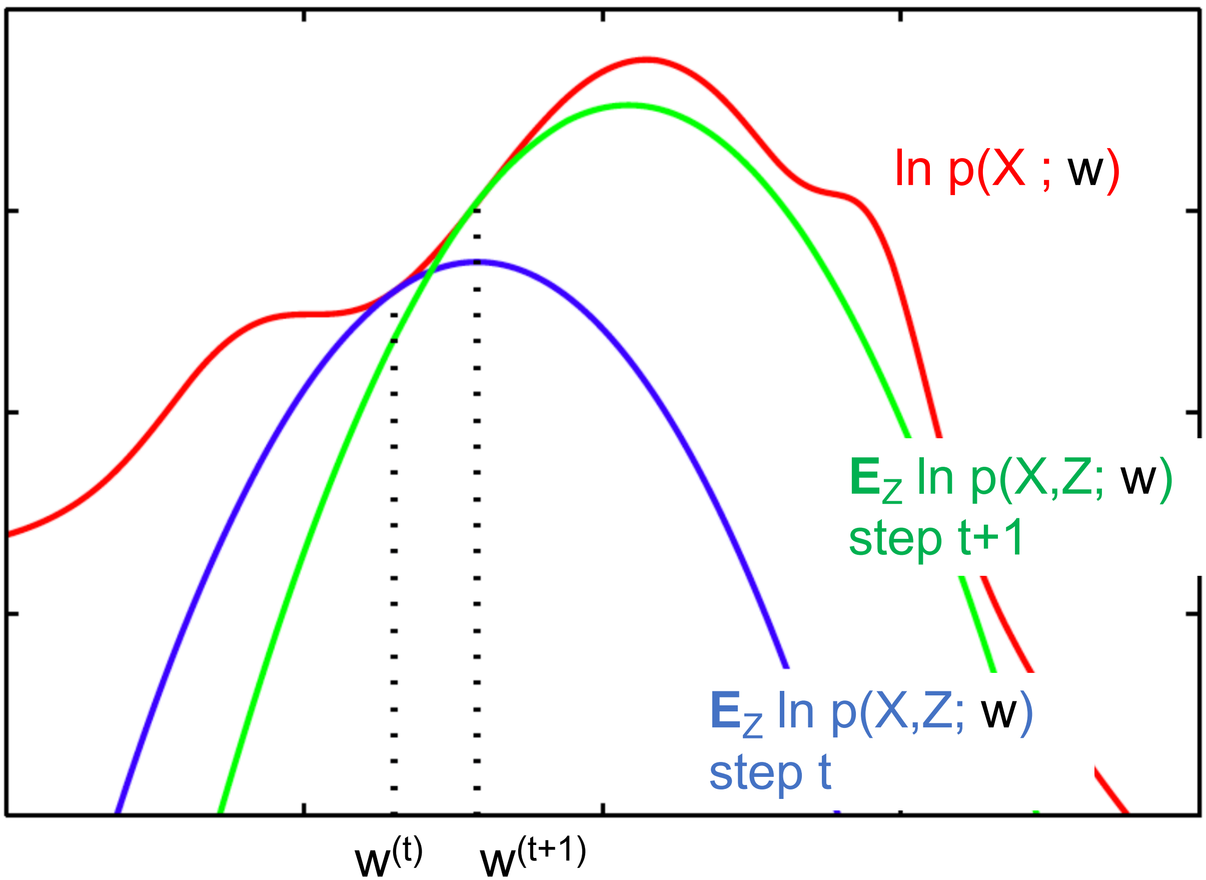
GMM / E-M algorithm



(Bishop)

What is E-M doing? (*advanced, and not covered in lecture)

(Zemel, Urtasun, Fidler)



p(X; w) is non-convex

At each iteration t

- **E**$_Z$[.] is a lower bound on ln p(X;w), and convex
- E-step: choose q s.t. **E**$_Z$[.] = p(X;w) ("pull up to likelihood curve")
- M-step: optimize lower bound