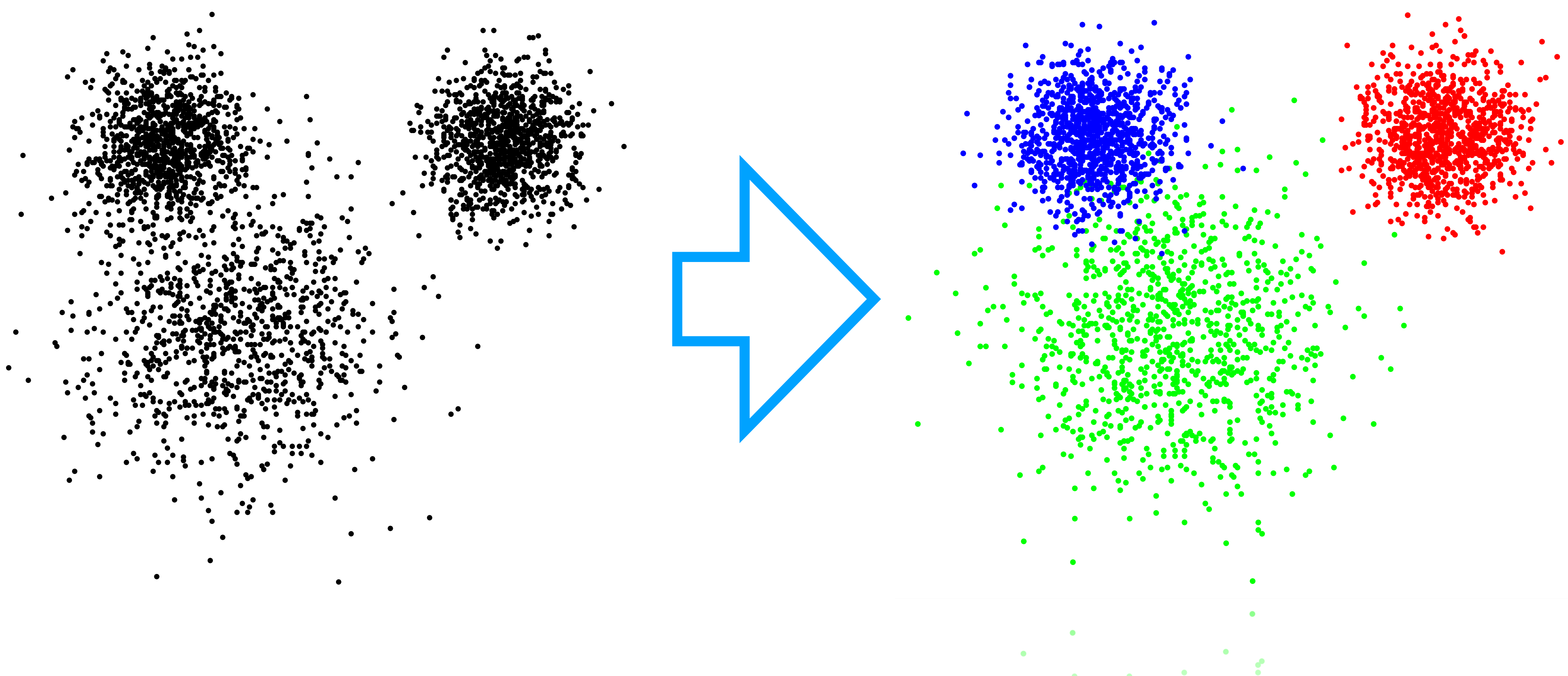
EM Algorithm and Mixture Models

Guojun Zhang
University of Waterloo

Unsupervised learning and clustering

- Learn the intrinsic representation of unlabeled data



- Other examples: density estimation, novelty detection

Mixture model

$$p(\mathbf{x}) = \sum_{c=1}^{m} \pi_c f(\mathbf{x}|\boldsymbol{\theta}_c),\ 0 \leqslant \pi_c \leqslant 1,\ \sum_{c=1}^{m} \pi_c = 1$$

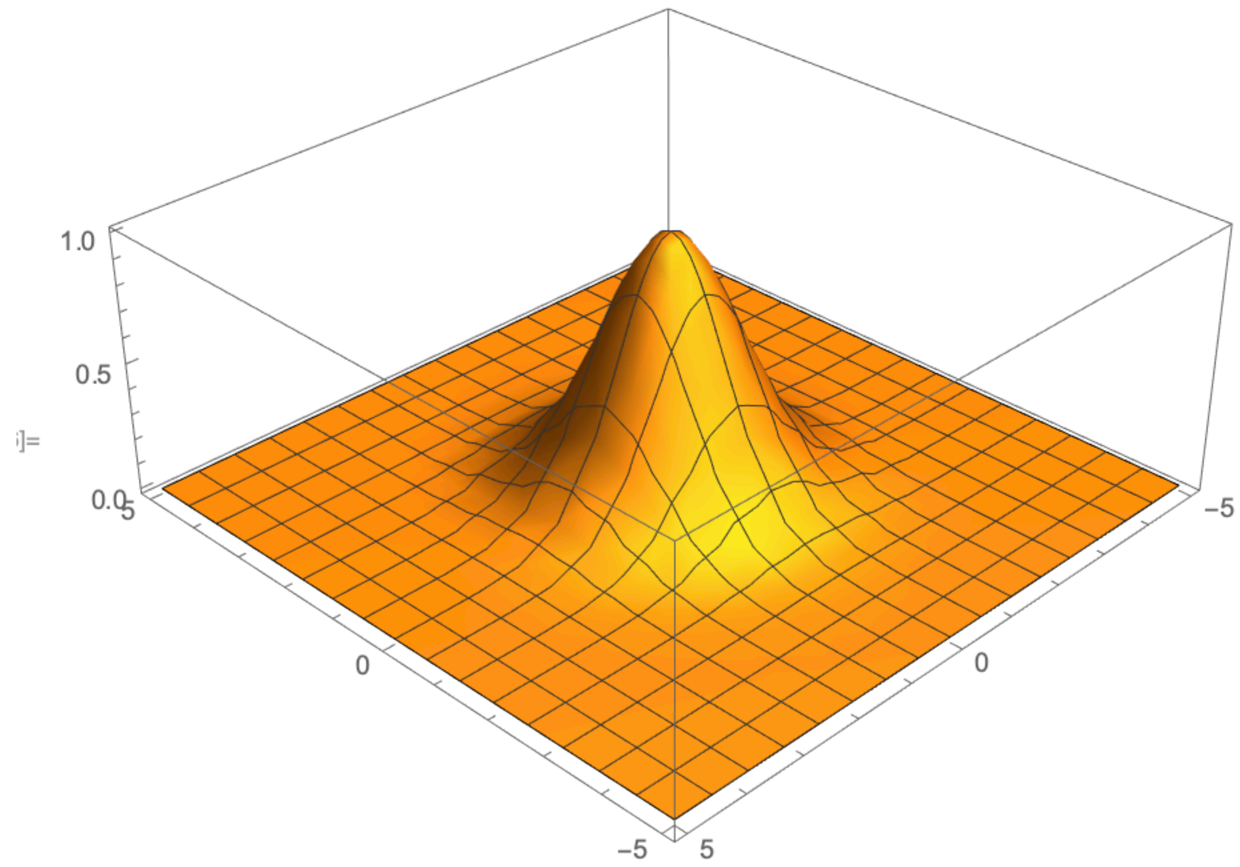
- Continuous: mixture of Gaussians

$$f(\mathbf{x}|\boldsymbol{\theta}_c) = \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_c, \boldsymbol{\Sigma}_c)$$

- Discrete: mixture of Bernoullis

$$f(\mathbf{x}|\boldsymbol{\theta}_c) = B(\mathbf{x}|\boldsymbol{\mu}_c) = \prod_{i=1}^{D} B(x_i|\mu_{c,i})$$

Gaussian



Bernoulli: flipping a coin

$$B(x|\mu) = \mu^x (1-\mu)^{1-x},\ x = 0, 1$$

Optimization algorithms

- Loss function: negative log likelihood

$$\ell = -\mathbb{E}[\log p(\mathbf{x})] = -\sum_{i=1}^{N} \log p(\mathbf{x})$$

- Expectation-Maximization (DLR 1977):

E step Evaluate $p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta}^{\text{old}})$.

M step Evaluate $\boldsymbol{\theta}^{\text{new}}$ given by

$$\boldsymbol{\theta}^{\text{new}} = \arg\max_{\boldsymbol{\theta}} \mathcal{Q}(\boldsymbol{\theta}, \boldsymbol{\theta}^{\text{old}})$$

where

$$\mathcal{Q}(\boldsymbol{\theta}, \boldsymbol{\theta}^{\text{old}}) = \sum_{\mathbf{Z}} p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta}^{\text{old}}) \ln p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta}).$$

Optimization algorithms

- Loss function: negative log likelihood

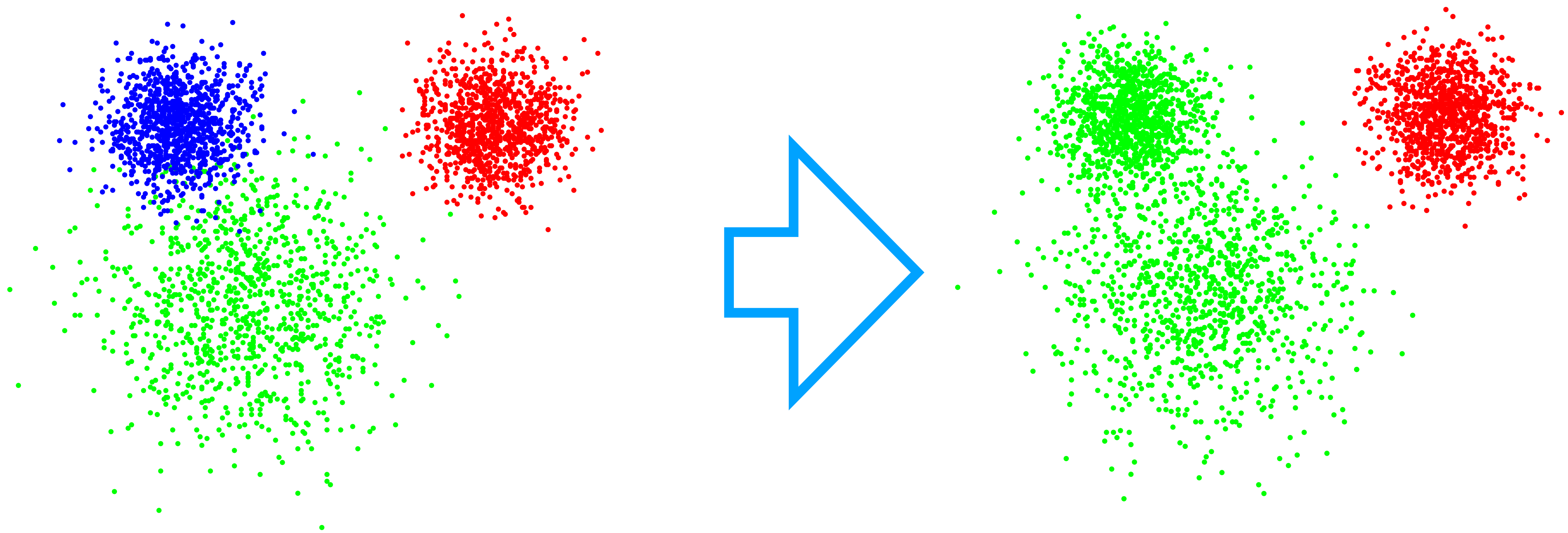
$$\ell = -\mathbb{E}[\log p(\mathbf{x})] = -\sum_{i=1}^{N} \log p(\mathbf{x})$$

- Gradient descent:

$$\boldsymbol{\pi} \leftarrow P_{\boldsymbol{\pi}}\left(\boldsymbol{\pi} - \alpha \frac{\partial \ell}{\partial \boldsymbol{\pi}}\right),\ \boldsymbol{\mu}_c \leftarrow P_{\boldsymbol{\mu}_c}\left(\boldsymbol{\mu}_c - \alpha \frac{\partial \ell}{\partial \boldsymbol{\mu}_c}\right)$$

k-cluster region

- What if just some clusters are used? Has the algorithm learned the ground truth? How bad are these regions?



Potential project

- To study how EM and GD (or any other algorithm) behave in learning mixture models

- Can they avoid some bad local minima, such as the k-cluster regions?

- Some Results/Guesses: 1) EM does but GD does not (on BMMs) 2) EM escapes exponentially faster than GD (on GMMs)

- Ultimate goal: to understand their convergence property and the limit of each algorithm; to propose better algorithms

- Need strong mathematical background: linear algebra, advanced calculus, probability theory and statistics, continuous optimization, (maybe) dynamical systems…

References

- Christopher Bishop, “Pattern Recognition and Machine Learning” (2006).

- Guojun Zhang, Pascal Poupart and George Trimponias, “Comparing EM with GD in Mixtures of Two Components,” to appear in UAI 2019.

- Dempster, Arthur P., Nan M. Laird and Donald B. Rubin. “Maximum likelihood from incomplete data via the EM algorithm.” Journal of the Royal Statistical Society: Series B (1977).