CS 104: Artificial Intelligence
Learning with Incomplete Data

Acknowledgement: Based on Prof. Collins (MIT) and Prof. Moore (CMU) lecture notes

An Experiment/Some Intuition

- I have one coin in my pocket,

  Coin 0 has probability λ of heads

- I toss the coin 10 times, and see the following sequence:

  `HHTTHHHTHH`

  (7 heads out of 10)

- What would you guess λ to be?

An Experiment/Some Intuition

- I have three coins in my pocket,

  Coin 0 has probability λ of heads;
  Coin 1 has probability p_1 of heads;
  Coin 2 has probability p_2 of heads

- For each trial I do the following:

  First I toss Coin 0
  If Coin 0 turns up **heads**, I toss **coin 1** three times
  If Coin 0 turns up **tails**, I toss **coin 2** three times

  I don't tell you whether Coin 0 came up heads or tails,
  or whether Coin 1 or 2 was tossed three times,
  but I do tell you how many heads/tails are seen at each trial

- You see the following sequence:

$$\langle HHH \rangle, \langle TTT \rangle, \langle HHH \rangle, \langle TTT \rangle, \langle HHH \rangle$$

What would you estimate as the values for λ, p_1 and p_2?

Maximum Likelihood Estimation

- We have data points $X_1, X_2, \ldots X_n$ drawn from some (finite or countable) set $\mathcal{X}$
- We have a parameter vector Θ
- We have a parameter space Ω
- We have a distribution $P(X \mid \Theta)$ for any $\Theta \in \Omega$, such that

$$\sum_{X \in \mathcal{X}} P(X \mid \Theta) = 1 \text{ and } P(X \mid \Theta) \geq 0 \text{ for all X}$$

- We assume that our data points $X_1, X_2, \ldots X_n$ are drawn at random (independently, identically distributed) from a distribution $P(X \mid \Theta^*)$ for some $\Theta^* \in \Omega$

Log-Likelihood

- We have data points $X_1, X_2, \ldots X_n$ drawn from some (finite or countable) set $\mathcal{X}$
- We have a parameter vector Θ, and a parameter space Ω
- We have a distribution $P(X \mid \Theta)$ for any $\Theta \in \Omega$
- The likelihood is

$$Likelihood(\Theta) = P(X_1, X_2, \ldots X_n \mid \Theta) = \prod_{i=1}^{n} P(X_i \mid \Theta)$$

- The log-likelihood is

$$L(\Theta) = \log Likelihood(\Theta) = \sum_{i=1}^{n} \log P(X_i \mid \Theta)$$

Maximum Likelihood Estimation

- Given a sample $X_1, X_2, \ldots X_n$, choose

$$\Theta_{ML} = \text{argmax}_{\Theta \in \Omega} L(\Theta) = \text{argmax}_{\Theta \in \Omega} \sum_i \log P(X_i \mid \Theta)$$

- For example, take the coin example:
  say $X_1 \ldots X_n$ has $Count(H)$ heads, and $(n - Count(H))$ tails

  $\Rightarrow$

$$\begin{aligned} L(\Theta) &= \log\left(\Theta^{Count(H)} \times (1-\Theta)^{n-Count(H)}\right) \\ &= Count(H)\log\Theta + (n - Count(H))\log(1-\Theta) \end{aligned}$$

- We now have

$$\Theta_{ML} = \frac{Count(H)}{n}$$

Models with Hidden Variables

- Now say we have two sets $\mathcal{X}$ and $\mathcal{Y}$, and a joint distribution $P(X, Y \mid \Theta)$

- If we had **fully observed data**, (X_i, Y_i) pairs, then

$$L(\Theta) = \sum_i \log P(X_i, Y_i \mid \Theta)$$

- If we have **partially observed data**, X_i examples, then

$$\begin{aligned} L(\Theta) &= \sum_i \log P(X_i \mid \Theta) \\ &= \sum_i \log \sum_{Y \in \mathcal{Y}} P(X_i, Y \mid \Theta) \end{aligned}$$

- The **EM (Expectation Maximization) algorithm** is a method for finding

$$\Theta_{ML} = \mathrm{argmax}_{\Theta} \sum_i \log \sum_{Y \in \mathcal{Y}} P(X_i, Y \mid \Theta)$$

The Three Coins Example

- e.g., in the three coins example:
  $\mathcal{Y} = \{\texttt{H,T}\}$
  $\mathcal{X} = \{\texttt{HHH,TTT,HTT,THH,HHT,TTH,HTH,THT}\}$
  $\Theta = \{\lambda, p_1, p_2\}$

- and

$$P(X, Y \mid \Theta) = P(Y \mid \Theta)P(X \mid Y, \Theta)$$

where

$$P(Y \mid \Theta) = \begin{cases} \lambda & \text{If } Y = \texttt{H} \\ 1 - \lambda & \text{If } Y = \texttt{T} \end{cases}$$

and

$$P(X \mid Y, \Theta) = \begin{cases} p_1^h(1 - p_1)^t & \text{If } Y = \texttt{H} \\ p_2^h(1 - p_2)t & \text{If } Y = \texttt{T} \end{cases}$$

where $h =$ number of heads in X, $t =$ number of tails in X

The Three Coins Example

- Various probabilities can be calculated, for example:

$$P(X = \mathtt{THT}, Y = \mathtt{H} \mid \Theta) \quad = \quad \lambda p_1 (1 - p_1)^2$$

The Three Coins Example

- Various probabilities can be calculated, for example:

$$P(X = \mathtt{THT}, Y = \mathtt{H} \mid \Theta) = \lambda p_1 (1 - p_1)^2$$

$$P(X = \mathtt{THT}, Y = \mathtt{T} \mid \Theta) = (1 - \lambda) p_2 (1 - p_2)^2$$

The Three Coins Example

- Various probabilities can be calculated, for example:

$$P(X = \texttt{THT}, Y = \texttt{H} \mid \Theta) = \lambda p_1 (1 - p_1)^2$$

$$P(X = \texttt{THT}, Y = \texttt{T} \mid \Theta) = (1 - \lambda) p_2 (1 - p_2)^2$$

$$\begin{aligned} P(X = \texttt{THT} \mid \Theta) &= P(X = \texttt{THT}, Y = \texttt{H} \mid \Theta) \\ &\quad + P(X = \texttt{THT}, Y = \texttt{T} \mid \Theta) \\ &= \lambda p_1 (1 - p_1)^2 + (1 - \lambda) p_2 (1 - p_2)^2 \end{aligned}$$

The Three Coins Example

- Various probabilities can be calculated, for example:

$$P(X = \mathtt{THT}, Y = \mathtt{H} \mid \Theta) = \lambda p_1 (1 - p_1)^2$$

$$P(X = \mathtt{THT}, Y = \mathtt{T} \mid \Theta) = (1 - \lambda) p_2 (1 - p_2)^2$$

$$\begin{aligned} P(X = \mathtt{THT} \mid \Theta) &= P(X = \mathtt{THT}, Y = \mathtt{H} \mid \Theta) \\ &\quad + P(X = \mathtt{THT}, Y = \mathtt{T} \mid \Theta) \\ &= \lambda p_1 (1 - p_1)^2 + (1 - \lambda) p_2 (1 - p_2)^2 \end{aligned}$$

$$\begin{aligned} P(Y = \mathtt{H} \mid X = \mathtt{THT}, \Theta) &= \frac{P(X = \mathtt{THT}, Y = \mathtt{H} \mid \Theta)}{P(X = \mathtt{THT} \mid \Theta)} \\ &= \frac{\lambda p_1 (1 - p_1)^2}{\lambda p_1 (1 - p_1)^2 + (1 - \lambda) p_2 (1 - p_2)^2} \end{aligned}$$

The Three Coins Example

- Fully observed data might look like:

$$(\langle HHH\rangle, H), (\langle TTT\rangle, T), (\langle HHH\rangle, H), (\langle TTT\rangle, T), (\langle HHH\rangle, H)$$

- In this case maximum likelihood estimates are:

$$\lambda = \frac{3}{5}$$

$$p_1 = \frac{9}{9}$$

$$p_2 = \frac{0}{6}$$

The Three Coins Example

- Partially observed data might look like:

$$\langle HHH\rangle, \langle TTT\rangle, \langle HHH\rangle, \langle TTT\rangle, \langle HHH\rangle$$

- How do we find the maximum likelihood parameters?

The Three Coins Example

- Partially observed data might look like:

$$\langle HHH \rangle, \langle TTT \rangle, \langle HHH \rangle, \langle TTT \rangle, \langle HHH \rangle$$

- If current parameters are λ, p_1, p_2

$$\begin{aligned} P(Y = \mathtt{H} \mid X = \langle \mathtt{HHH} \rangle) &= \frac{P(\langle \mathtt{HHH} \rangle, \mathtt{H})}{P(\langle \mathtt{HHH} \rangle, \mathtt{H}) + P(\langle \mathtt{HHH} \rangle, \mathtt{T})} \\ &= \frac{\lambda p_1^3}{\lambda p_1^3 + (1-\lambda) p_2^3} \end{aligned}$$

$$\begin{aligned} P(Y = \mathtt{H} \mid X = \langle \mathtt{TTT} \rangle) &= \frac{P(\langle \mathtt{TTT} \rangle, \mathtt{H})}{P(\langle \mathtt{TTT} \rangle, \mathtt{H}) + P(\langle \mathtt{TTT} \rangle, \mathtt{T})} \\ &= \frac{\lambda (1-p_1)^3}{\lambda (1-p_1)^3 + (1-\lambda)(1-p_2)^3} \end{aligned}$$

The Three Coins Example

- If current parameters are λ, p_1, p_2

$$P(Y = \texttt{H} \mid X = \langle \texttt{HHH} \rangle) = \frac{\lambda p_1^3}{\lambda p_1^3 + (1-\lambda) p_2^3}$$

$$P(Y = \texttt{H} \mid X = \langle \texttt{TTT} \rangle) = \frac{\lambda (1-p_1)^3}{\lambda (1-p_1)^3 + (1-\lambda)(1-p_2)^3}$$

- If $\lambda = 0.3, p_1 = 0.3, p_2 = 0.6$:

$$P(Y = \texttt{H} \mid X = \langle \texttt{HHH} \rangle) = 0.0508$$

$$P(Y = \texttt{H} \mid X = \langle \texttt{TTT} \rangle) = 0.6967$$

The Three Coins Example

- After filling in hidden variables for each example, partially observed data might look like:

$$
\begin{array}{rl}
(\langle \texttt{HHH} \rangle, H) & P(Y = \texttt{H} \mid \texttt{HHH}) = 0.0508 \\
(\langle \texttt{HHH} \rangle, T) & P(Y = \texttt{T} \mid \texttt{HHH}) = 0.9492 \\
(\langle \texttt{TTT} \rangle, H) & P(Y = \texttt{H} \mid \texttt{TTT}) = 0.6967 \\
(\langle \texttt{TTT} \rangle, T) & P(Y = \texttt{T} \mid \texttt{TTT}) = 0.3033 \\
(\langle \texttt{HHH} \rangle, H) & P(Y = \texttt{H} \mid \texttt{HHH}) = 0.0508 \\
(\langle \texttt{HHH} \rangle, T) & P(Y = \texttt{T} \mid \texttt{HHH}) = 0.9492 \\
(\langle \texttt{TTT} \rangle, H) & P(Y = \texttt{H} \mid \texttt{TTT}) = 0.6967 \\
(\langle \texttt{TTT} \rangle, T) & P(Y = \texttt{T} \mid \texttt{TTT}) = 0.3033 \\
(\langle \texttt{HHH} \rangle, H) & P(Y = \texttt{H} \mid \texttt{HHH}) = 0.0508 \\
(\langle \texttt{HHH} \rangle, T) & P(Y = \texttt{T} \mid \texttt{HHH}) = 0.9492
\end{array}
$$

The Three Coins Example

- New Estimates:

$$
\begin{array}{rl}
(\langle \texttt{HHH} \rangle, H) & P(Y = \texttt{H} \mid \texttt{HHH}) = 0.0508 \\
(\langle \texttt{HHH} \rangle, T) & P(Y = \texttt{T} \mid \texttt{HHH}) = 0.9492 \\
(\langle \texttt{TTT} \rangle, H) & P(Y = \texttt{H} \mid \texttt{TTT}) = 0.6967 \\
(\langle \texttt{TTT} \rangle, T) & P(Y = \texttt{T} \mid \texttt{TTT}) = 0.3033 \\
\ldots &
\end{array}
$$

$$\lambda = \frac{3 \times 0.0508 + 2 \times 0.6967}{5} = 0.3092$$

$$p_1 = \frac{3 \times 3 \times 0.0508 + 0 \times 2 \times 0.6967}{3 \times 3 \times 0.0508 + 3 \times 2 \times 0.6967} = 0.0987$$

$$p_2 = \frac{3 \times 3 \times 0.9492 + 0 \times 2 \times 0.3033}{3 \times 3 \times 0.9492 + 3 \times 2 \times 0.3033} = 0.8244$$

The Three Coins Example: Summary

- Begin with parameters $\lambda = 0.3, p_1 = 0.3, p_2 = 0.6$
- Fill in hidden variables, using

$$
\begin{aligned}
P(Y = \texttt{H} \mid X = \langle \texttt{HHH} \rangle) &= 0.0508 \\
P(Y = \texttt{H} \mid X = \langle \texttt{TTT} \rangle) &= 0.6967
\end{aligned}
$$

- Re-estimate parameters to be $\lambda = 0.3092, p_1 = 0.0987, p_2 = 0.8244$

Iteration	λ	p_1	p_2	$\tilde{p}_1$	$\tilde{p}_2$	$\tilde{p}_3$	$\tilde{p}_4$
0	0.3000	0.3000	0.6000	0.0508	0.6967	0.0508	0.6967
1	0.3738	0.0680	0.7578	0.0004	0.9714	0.0004	0.9714
2	0.4859	0.0004	0.9722	0.0000	1.0000	0.0000	1.0000
3	0.5000	0.0000	1.0000	0.0000	1.0000	0.0000	1.0000

The coin example for $\mathbf{Y} = \{\langle HHH\rangle, \langle TTT\rangle, \langle HHH\rangle, \langle TTT\rangle\}$. The solution that EM reaches is intuitively correct: the coin-tosser has two coins, one which always shows up heads, the other which always shows tails, and is picking between them with equal probability ($\lambda = 0.5$). The posterior probabilities $\tilde{p}_i$ show that we are certain that coin 1 (tail-biased) generated Y_2 and Y_4, whereas coin 2 generated Y_1 and Y_3.

Iteration	λ	p_1	p_2	$\tilde{p}_1$	$\tilde{p}_2$	$\tilde{p}_3$	$\tilde{p}_4$	$\tilde{p}_5$
0	0.3000	0.3000	0.6000	0.0508	0.6967	0.0508	0.6967	0.0508
1	0.3092	0.0987	0.8244	0.0008	0.9837	0.0008	0.9837	0.0008
2	0.3940	0.0012	0.9893	0.0000	1.0000	0.0000	1.0000	0.0000
3	0.4000	0.0000	1.0000	0.0000	1.0000	0.0000	1.0000	0.0000

The coin example for $\{\langle HHH\rangle, \langle TTT\rangle, \langle HHH\rangle, \langle TTT\rangle, \langle HHH\rangle\}$. λ is now 0.4, indicating that the coin-tosser has probability 0.4 of selecting the tail-biased coin.

Iteration	λ	p_1	p_2	$\tilde{p}_1$	$\tilde{p}_2$	$\tilde{p}_3$	$\tilde{p}_4$
0	0.3000	0.3000	0.6000	0.1579	0.6967	0.0508	0.6967
1	0.4005	0.0974	0.6300	0.0375	0.9065	0.0025	0.9065
2	0.4632	0.0148	0.7635	0.0014	0.9842	0.0000	0.9842
3	0.4924	0.0005	0.8205	0.0000	0.9941	0.0000	0.9941
4	0.4970	0.0000	0.8284	0.0000	0.9949	0.0000	0.9949

The coin example for $\mathbf{Y} = \{\langle HHT \rangle, \langle TTT \rangle, \langle HHH \rangle, \langle TTT \rangle\}$. EM selects a tails-only coin, and a coin which is heavily heads-biased ($p_2 = 0.8284$). It's certain that Y_1 and Y_3 were generated by coin 2, as they contain heads. Y_2 and Y_4 could have been generated by either coin, but coin 1 is far more likely.

Iteration	λ	p_1	p_2	$\tilde{p}_1$	$\tilde{p}_2$	$\tilde{p}_3$	$\tilde{p}_4$
0	0.3000	0.7000	0.7000	0.3000	0.3000	0.3000	0.3000
1	0.3000	0.5000	0.5000	0.3000	0.3000	0.3000	0.3000
2	0.3000	0.5000	0.5000	0.3000	0.3000	0.3000	0.3000
3	0.3000	0.5000	0.5000	0.3000	0.3000	0.3000	0.3000
4	0.3000	0.5000	0.5000	0.3000	0.3000	0.3000	0.3000
5	0.3000	0.5000	0.5000	0.3000	0.3000	0.3000	0.3000
6	0.3000	0.5000	0.5000	0.3000	0.3000	0.3000	0.3000

The coin example for $\mathbf{Y} = \{\langle HHH\rangle, \langle TTT\rangle, \langle HHH\rangle, \langle TTT\rangle\}$, with p_1 and p_2 initialised to the same value. EM is stuck at a saddle point

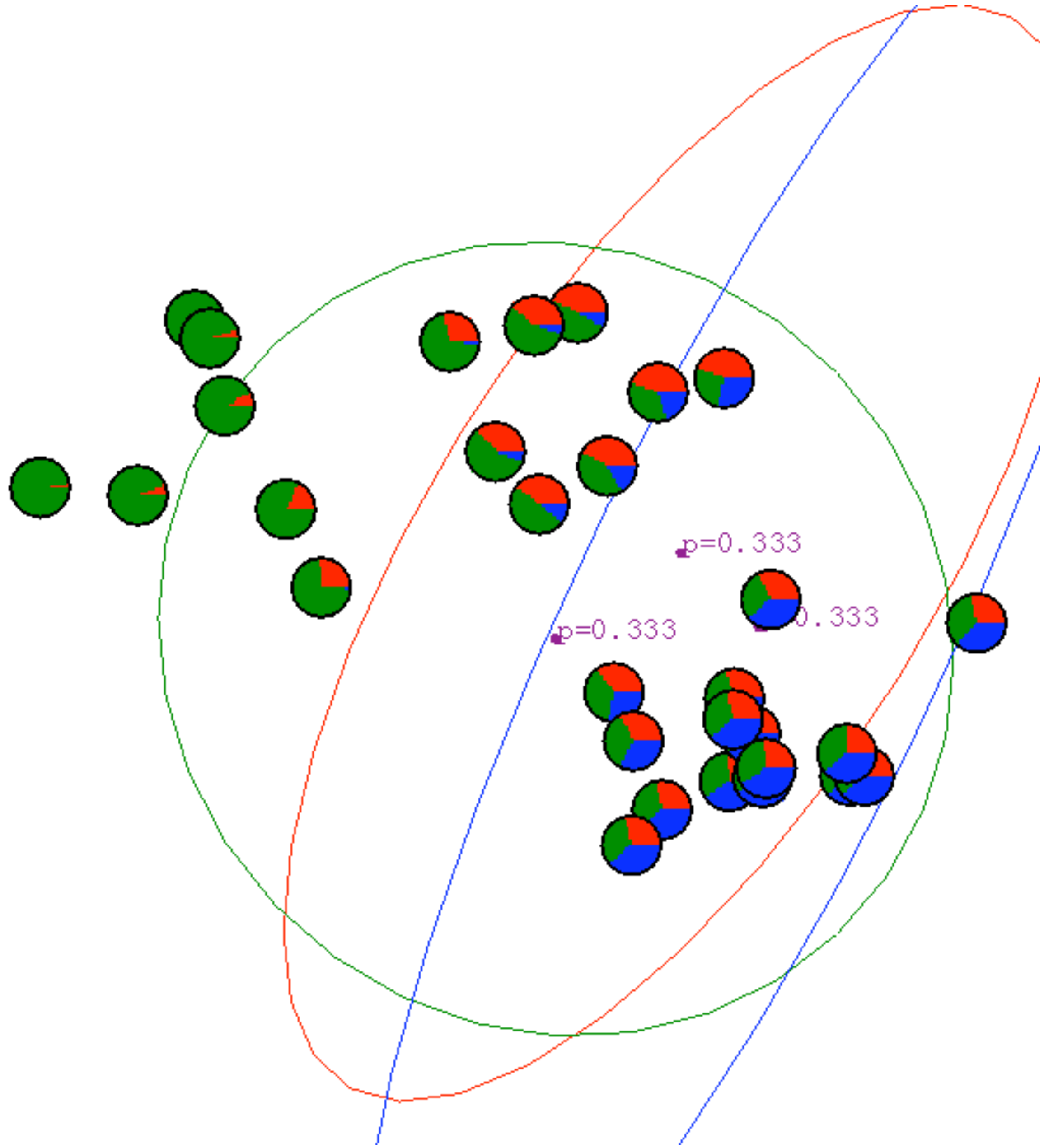
Iteration	λ	p_1	p_2	$\tilde{p}_1$	$\tilde{p}_2$	$\tilde{p}_3$	$\tilde{p}_4$
0	0.3000	0.7001	0.7000	0.3001	0.2998	0.3001	0.2998
1	0.2999	0.5003	0.4999	0.3004	0.2995	0.3004	0.2995
2	0.2999	0.5008	0.4997	0.3013	0.2986	0.3013	0.2986
3	0.2999	0.5023	0.4990	0.3040	0.2959	0.3040	0.2959
4	0.3000	0.5068	0.4971	0.3122	0.2879	0.3122	0.2879
5	0.3000	0.5202	0.4913	0.3373	0.2645	0.3373	0.2645
6	0.3009	0.5605	0.4740	0.4157	0.2007	0.4157	0.2007
7	0.3082	0.6744	0.4223	0.6447	0.0739	0.6447	0.0739
8	0.3593	0.8972	0.2773	0.9500	0.0016	0.9500	0.0016
9	0.4758	0.9983	0.0477	0.9999	0.0000	0.9999	0.0000
10	0.4999	1.0000	0.0001	1.0000	0.0000	1.0000	0.0000
11	0.5000	1.0000	0.0000	1.0000	0.0000	1.0000	0.0000

The coin example for $\mathbf{Y} = \{\langle HHH\rangle, \langle TTT\rangle, \langle HHH\rangle, \langle TTT\rangle\}$. If we initialise p_1 and p_2 to be a small amount away from the saddle point $p_1 = p_2$, the algorithm diverges from the saddle point and eventually reaches the global maximum.

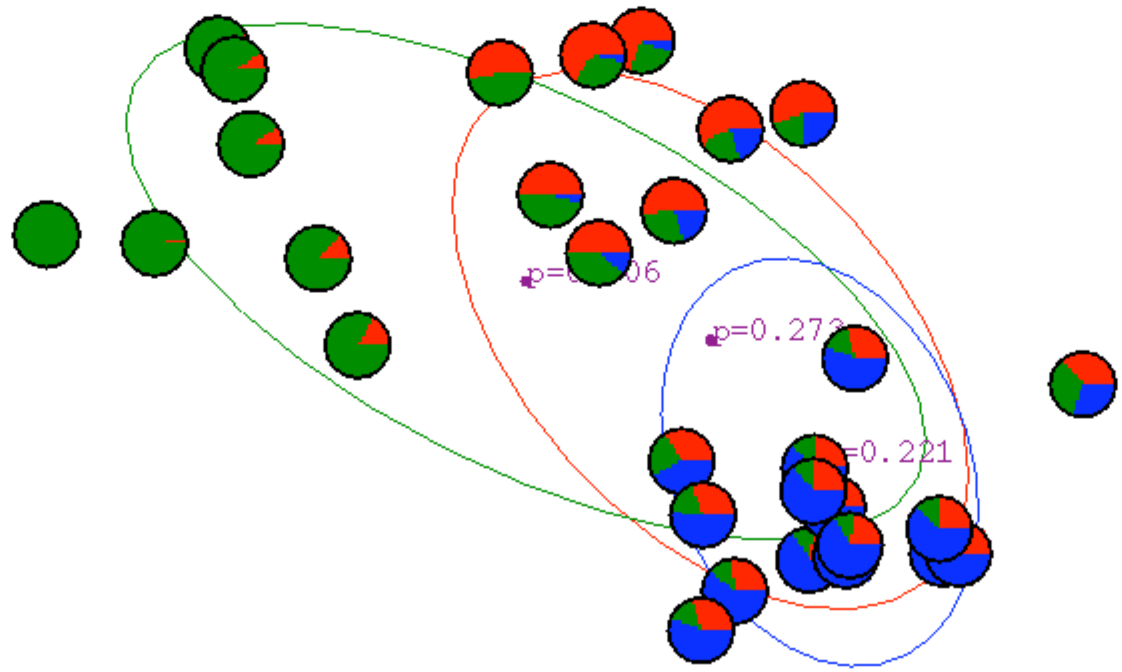
Iteration	λ	p_1	p_2	$\tilde{p}_1$	$\tilde{p}_2$	$\tilde{p}_3$	$\tilde{p}_4$
0	0.3000	0.6999	0.7000	0.2999	0.3002	0.2999	0.3002
1	0.3001	0.4998	0.5001	0.2996	0.3005	0.2996	0.3005
2	0.3001	0.4993	0.5003	0.2987	0.3014	0.2987	0.3014
3	0.3001	0.4978	0.5010	0.2960	0.3041	0.2960	0.3041
4	0.3001	0.4933	0.5029	0.2880	0.3123	0.2880	0.3123
5	0.3002	0.4798	0.5087	0.2646	0.3374	0.2646	0.3374
6	0.3010	0.4396	0.5260	0.2008	0.4158	0.2008	0.4158
7	0.3083	0.3257	0.5777	0.0739	0.6448	0.0739	0.6448
8	0.3594	0.1029	0.7228	0.0016	0.9500	0.0016	0.9500
9	0.4758	0.0017	0.9523	0.0000	0.9999	0.0000	0.9999
10	0.4999	0.0000	0.9999	0.0000	1.0000	0.0000	1.0000
11	0.5000	0.0000	1.0000	0.0000	1.0000	0.0000	1.0000

The coin example for $\mathbf{Y} = \{\langle HHH\rangle, \langle TTT\rangle, \langle HHH\rangle, \langle TTT\rangle\}$. If we initialise p_1 and p_2 to be a small amount away from the saddle point $p_1 = p_2$, the algorithm diverges from the saddle point and eventually reaches the global maximum.

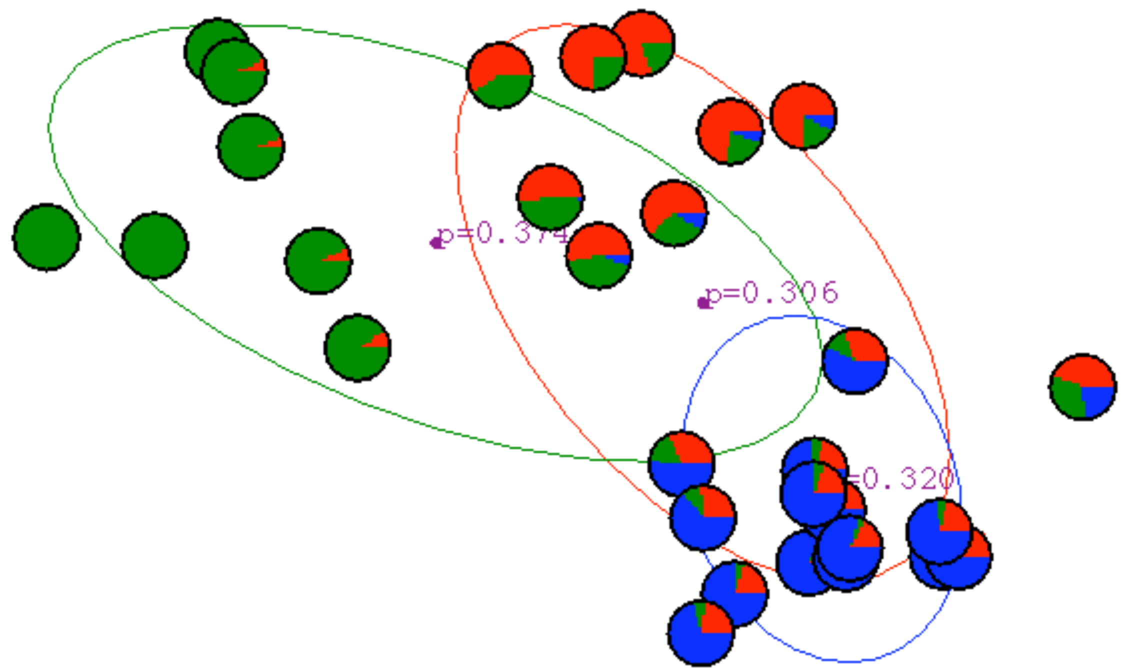
Gaussian Mixture Example: Start



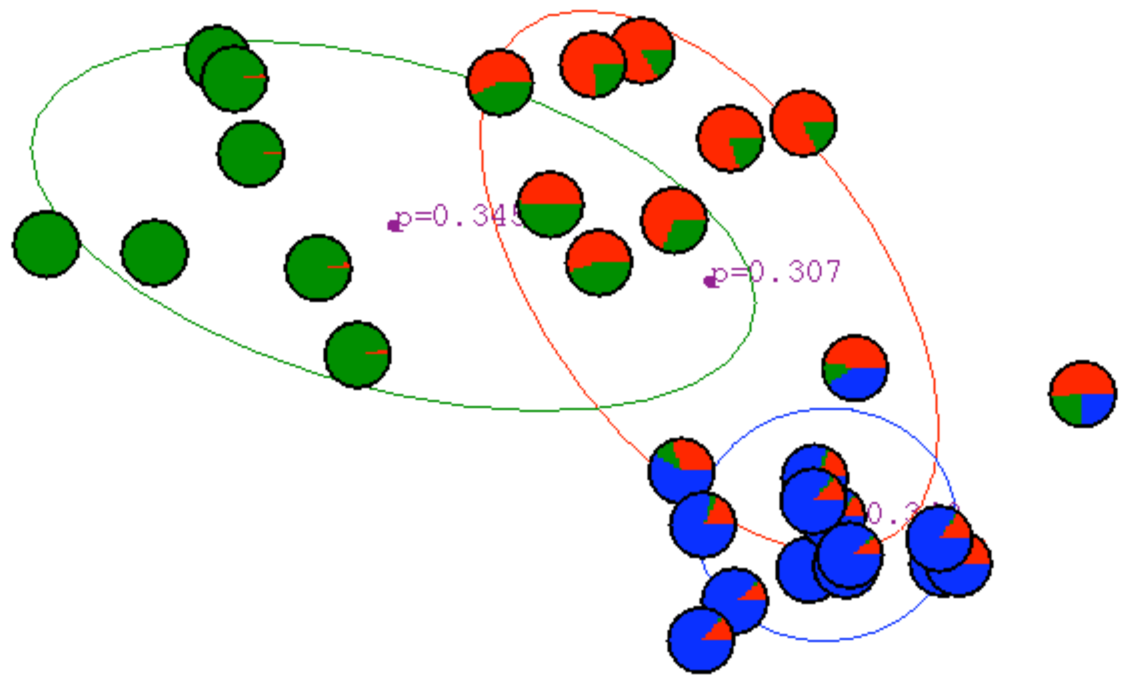
After first iteration



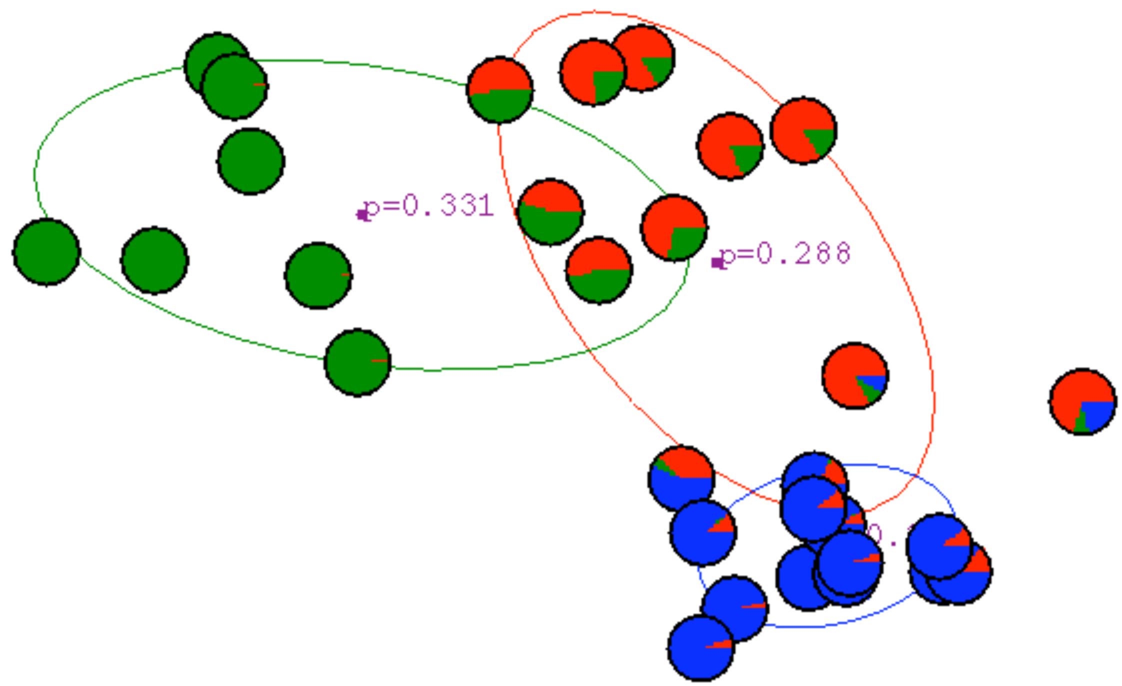
After 2nd iteration



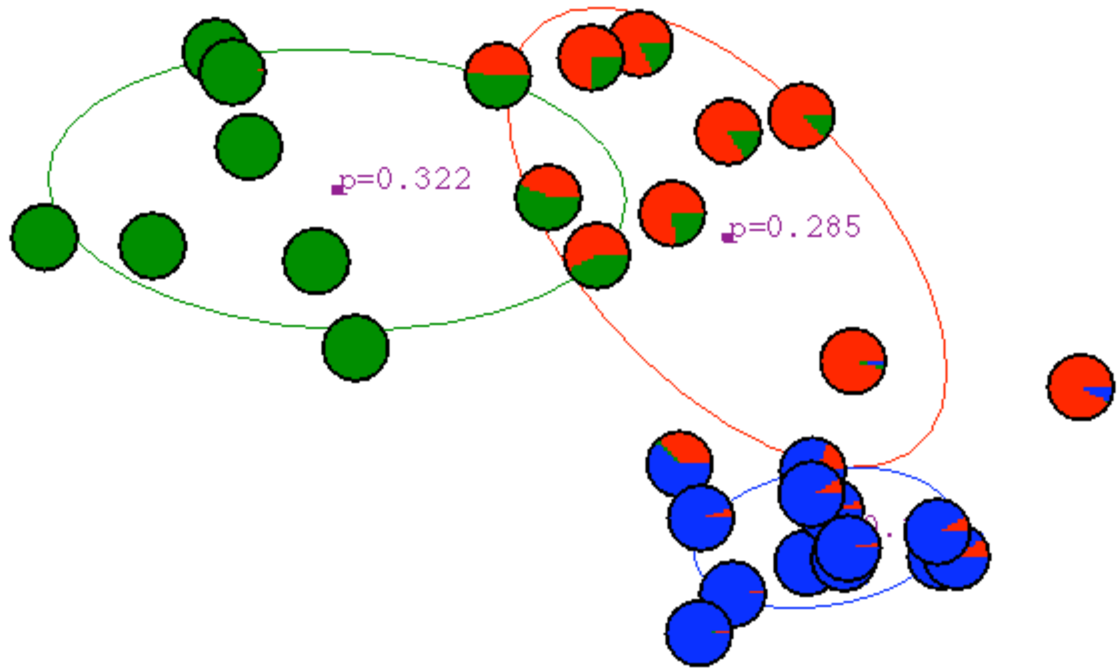
After 3rd iteration



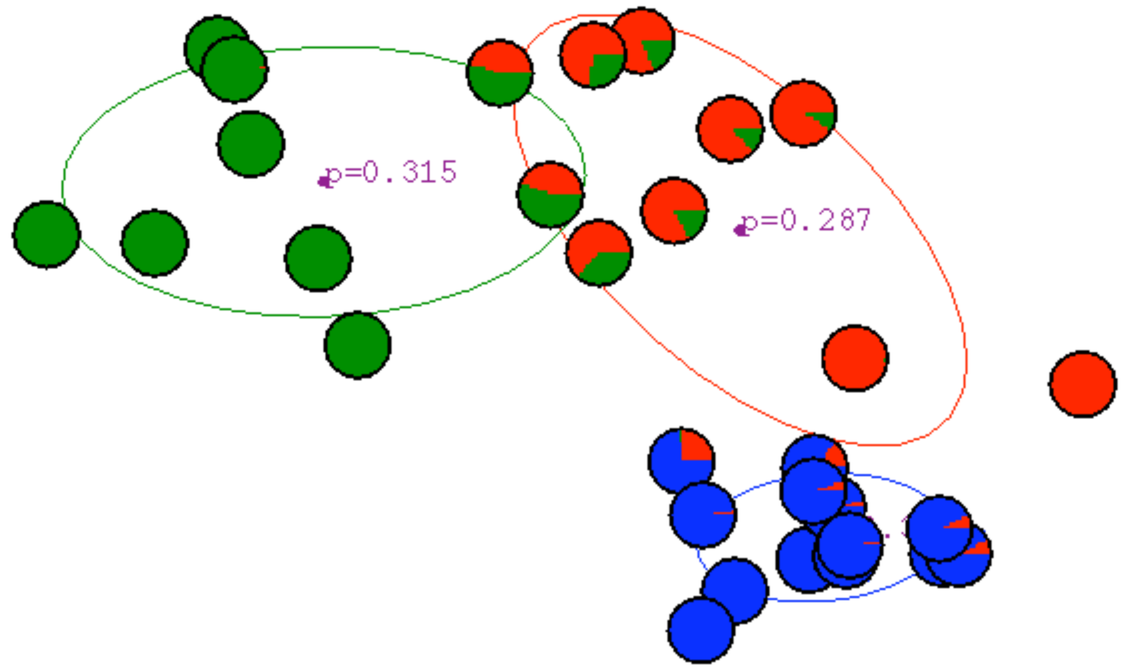
After 4th iteration



After 5th iteration



After 6th iteration



After 20th iteration