CS5540: Computational Techniques for Analyzing Clinical Data

Prof. Ramin Zabih (CS)
Prof. Ashish Raj (Radiology)

Today's topic

- Decisions based on densities
- Density estimation from data
  - Parametric (Gaussian)
  - Non-parametric
  - Non-parametric mode finding

Decisions from density

- Suppose we have a measure of an ECG
  - Different responses for shockable and normal
- Let's call our measure M (Measurement)
  - M(ECG) is a number
  - Tends to be 1 for normal sinus rhythm (NSR), 0 for anything else
    - We'll assume non-NSR needs shocking
  - It's reasonable to guess that M might be a Gaussian (why?)
  - Knowing how M looks allows decisions
    - Even ones with nice optimality properties

What densities do we need?

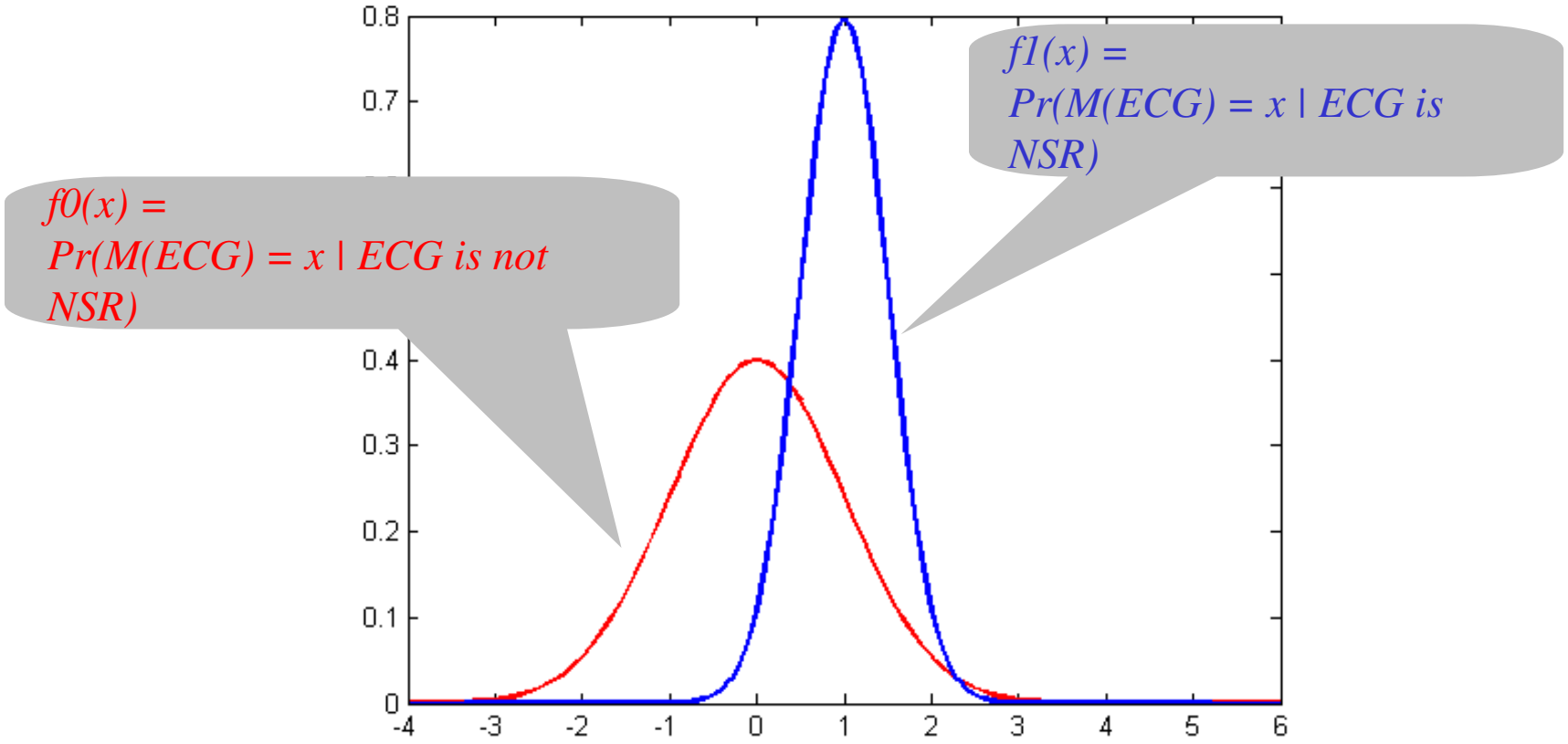
- Consider the values of M we get for patients with NSR
  - This density, peaked around 1:

    f1(x) = Pr(M(ECG) = x | ECG is NSR)
  - Similarly for not NSR, peaked around 0:

    f0(x) = Pr(M(ECG) = x | ECG is not NSR)
- What might these look like?

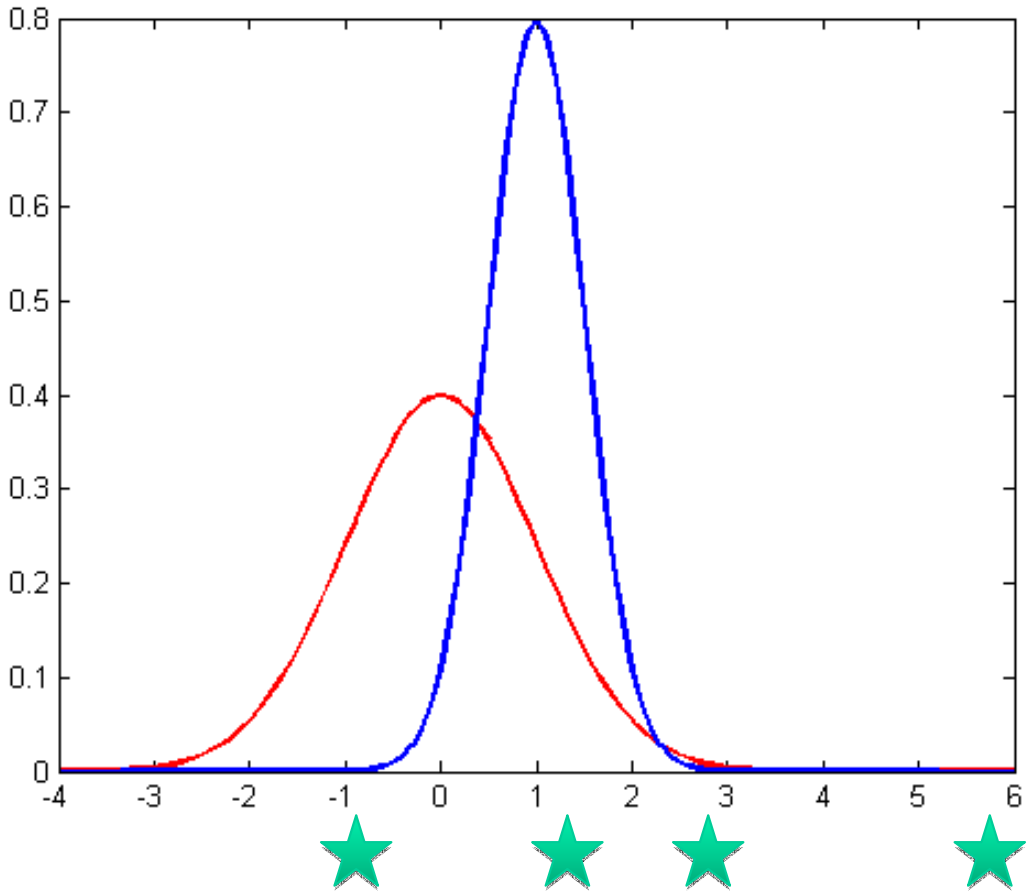
Densities

Probability



Value of M

Obvious decision rule



Choice of models

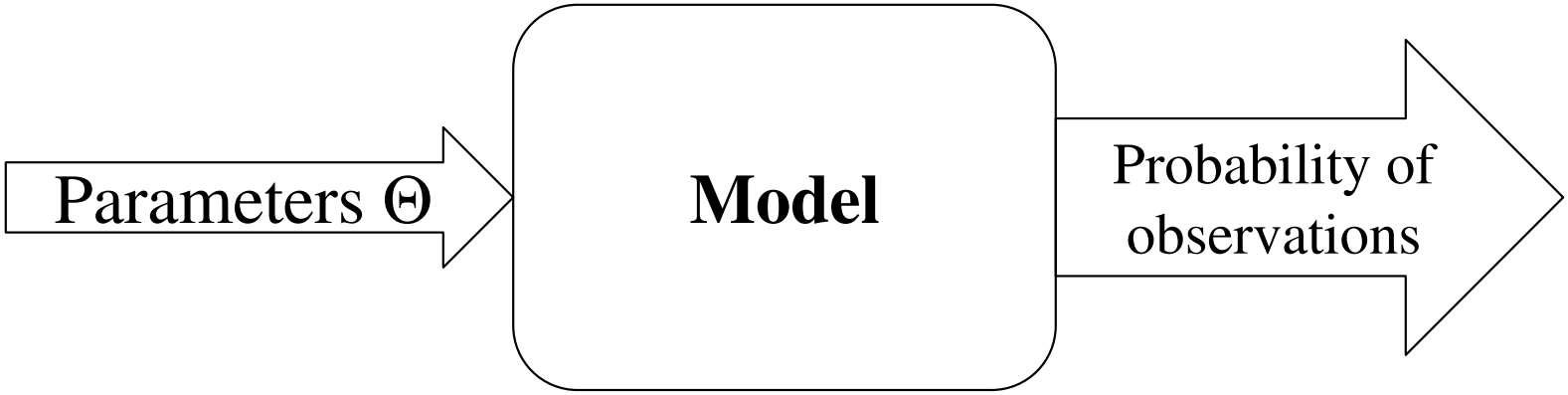
- Consider two possible models
  - E.g., Gaussian vs uniform
- Each one is a probability density
- More usefully, each one gives each possible observation a probability
  - E.g., for uniform model they are all the same
- This means that each model assigns a probability to what you actually saw
  - A model that almost never predicts what you saw is probably a bad one!

Maximum likelihood choice

- The likelihood of a model is the probability that it generated what you observed
- Maximum likelihood: pick the model where this is highest
- This justifies the obvious decision rule!
- Can derive this in terms of a loss function, which specifies the cost to be incorrect
  - If you guess X and the answer is Y, the loss typically depends on |X-Y|

Parameter estimation

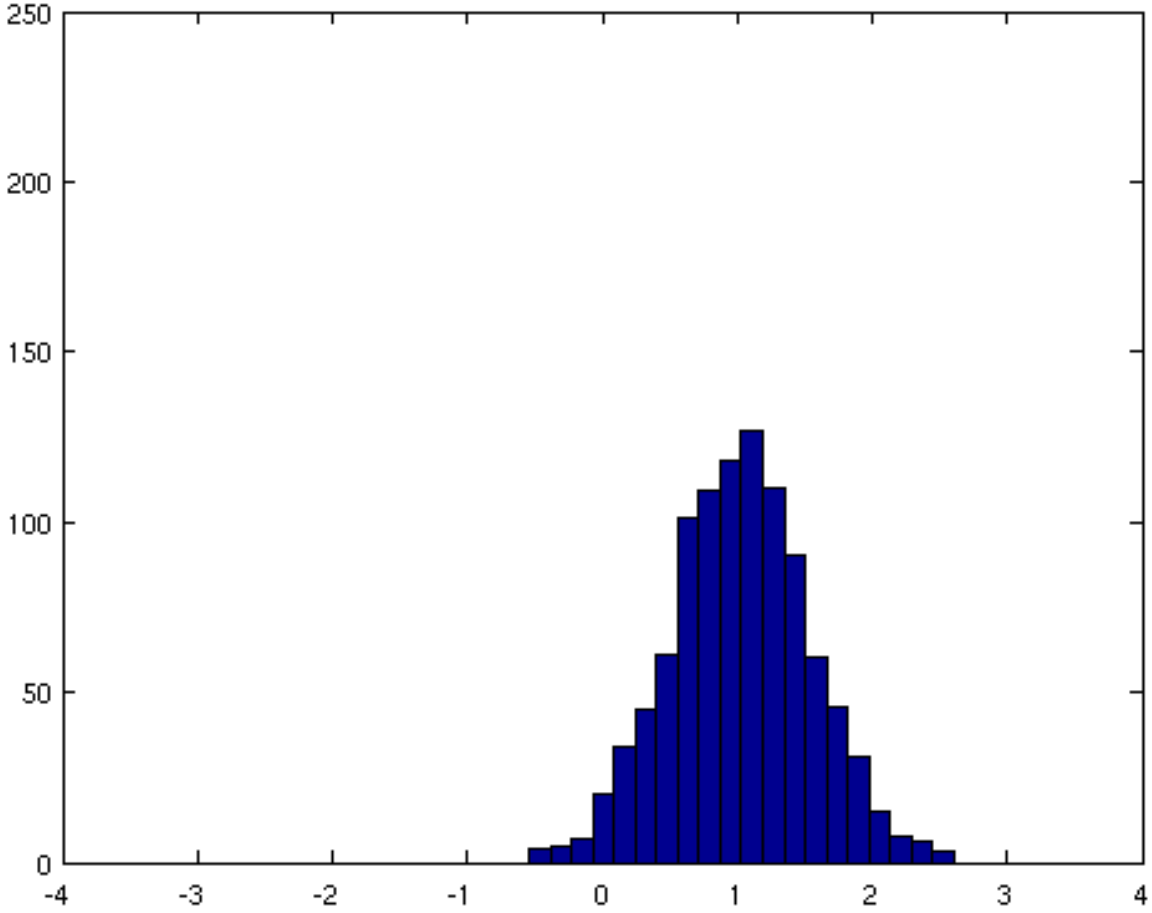
- Generalization of model choice
  - A Gaussian is an infinite set of models!



Summary

- We start with some training data (ECG's that we know are NSR or not)
  - We can compute M on each ECG
  - Get the densities f0, f1 from this data
    - How? We'll see shortly
- Based on these densities we can classify a new test ECG
  - Pick the hypothesis with the highest likelihood

Sample data from models



Density estimation

- How do we actually estimate a density from data?
- Loosely speaking, if we histogram "enough" data we should get the density
  - But this depends on bin size
  - Sparsity is a killer issue in high dimension
- What if we know the result is a Gaussian?
  - Only two parameters (center, width)
- Estimating the density means estimating the parameters of the Gaussian

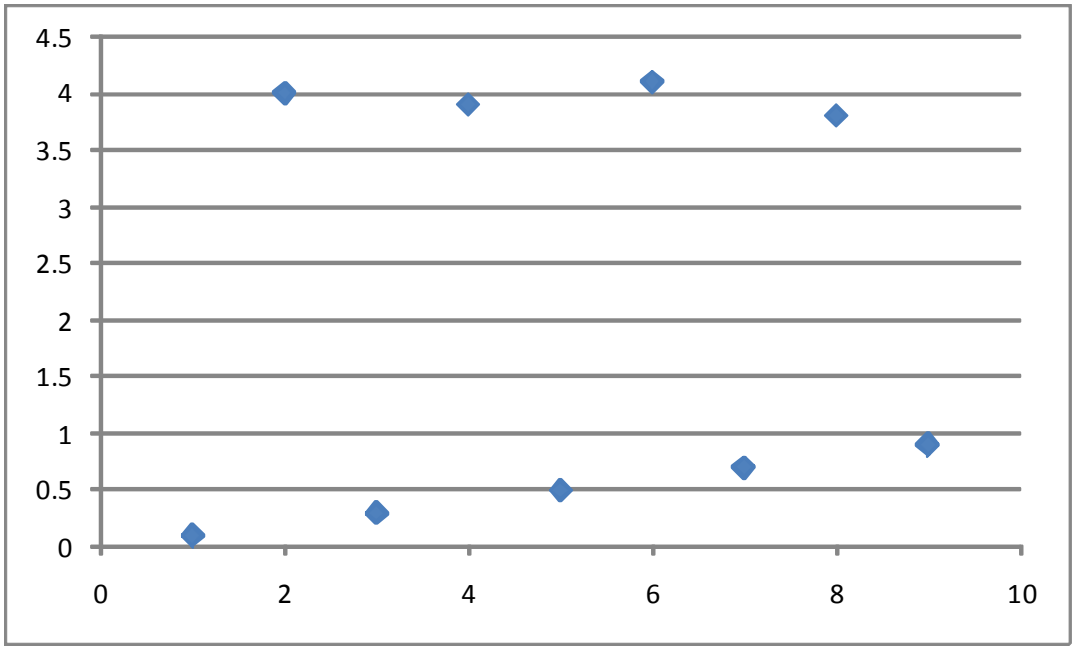
ML density estimation

- Among all the infinite number of Gaussians that might have generated the observed data, we can easily compute the one with the highest likelihood!
  - What are the right parameters?
  - There is actually a nice closed-form solution!

Harder cases

- What if your data came from two Gaussians with unknown parameters?
  - This is like our ECG example if the training data was not labeled
- If you knew which data point came from which Gaussian, could solve exactly
  - See previous slide!
- Similarly, if you knew the Gaussian parameters, you could tell which data point came from which Gaussian

Simplify: intermixed lines



- What about cases like this?
  - Can we find both lines?
  - "Hidden" variable: which line owns a point

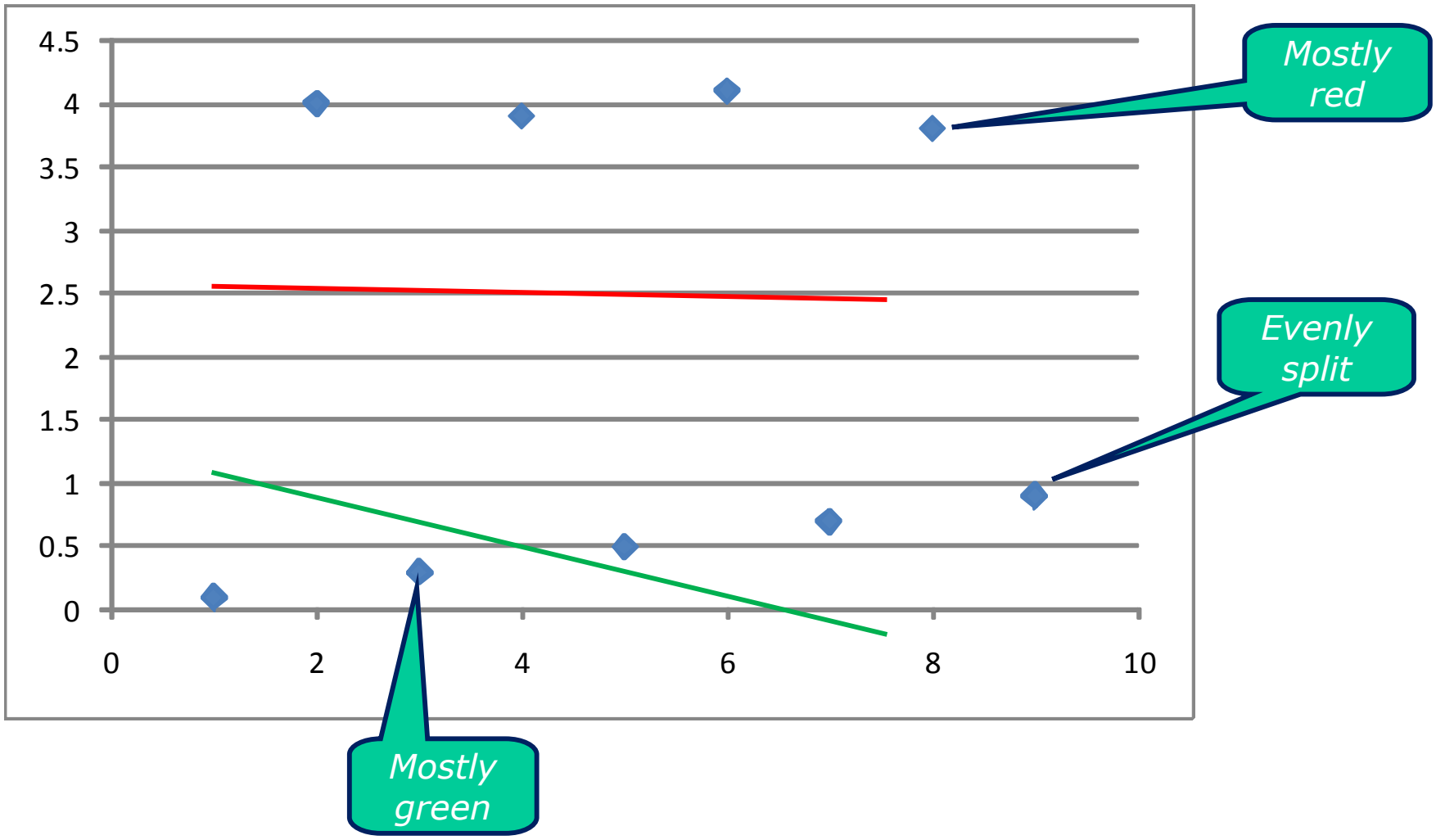
Lines and ownership

- Need to know lines and "ownership"
  - Which point belongs to which lines
  - We'll use partial ownership (non-binary)
- If we knew the lines we could apportion the ownership
  - A point is primarily owned by closest line
- If we knew ownership we could find the line via weighted LS
  - Weights from ownership

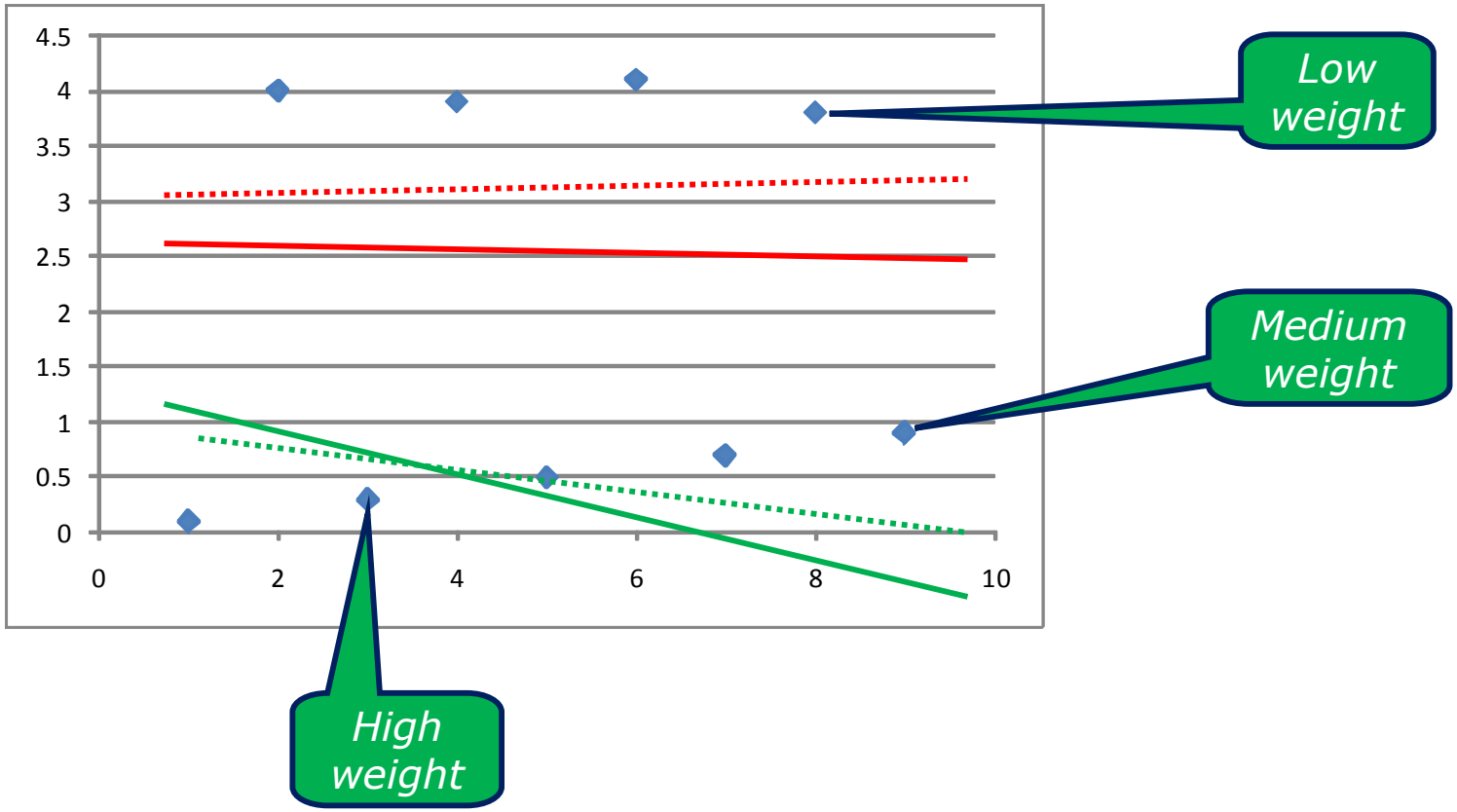
Solution: guess and iterate

- Put two lines down at random
- Compute ownership given lines
  - Each point divides its ownership among the lines, mostly to the closest line
- Compute lines given ownership
  - Weighted LS: weights rise with ownership
- Iterate between these

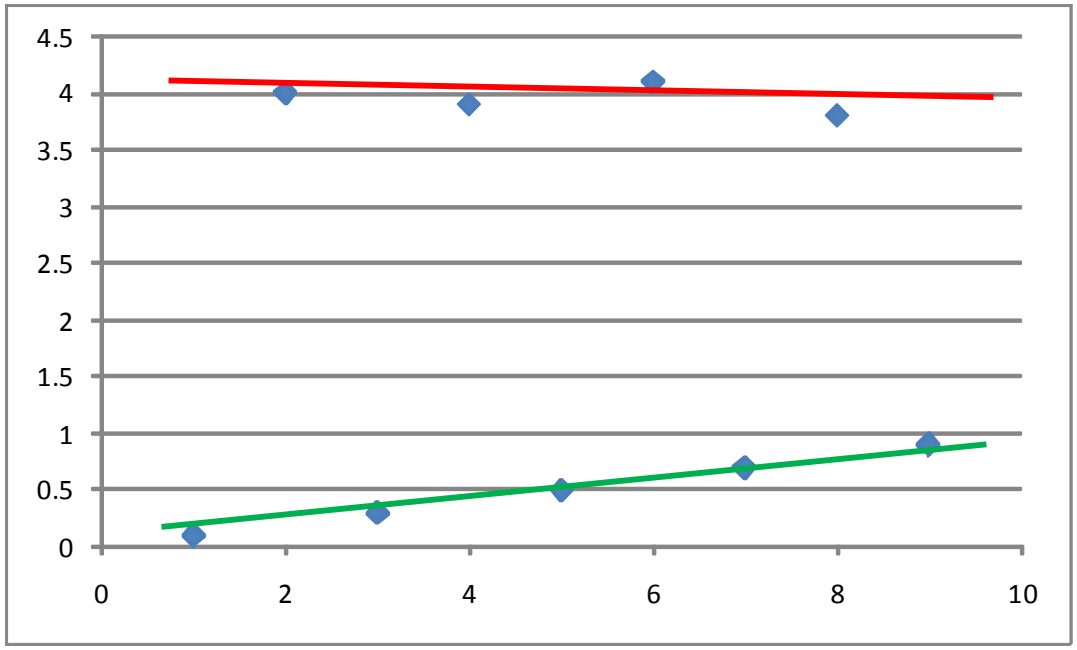
Ownership from lines



Lines from ownership



Convergence



Questions about EM

- Does this converge?
  - Not obvious that it won't cycle between solutions, or run forever
- Does the answer depend on the starting guess?
  - I.e., is this non-convex?