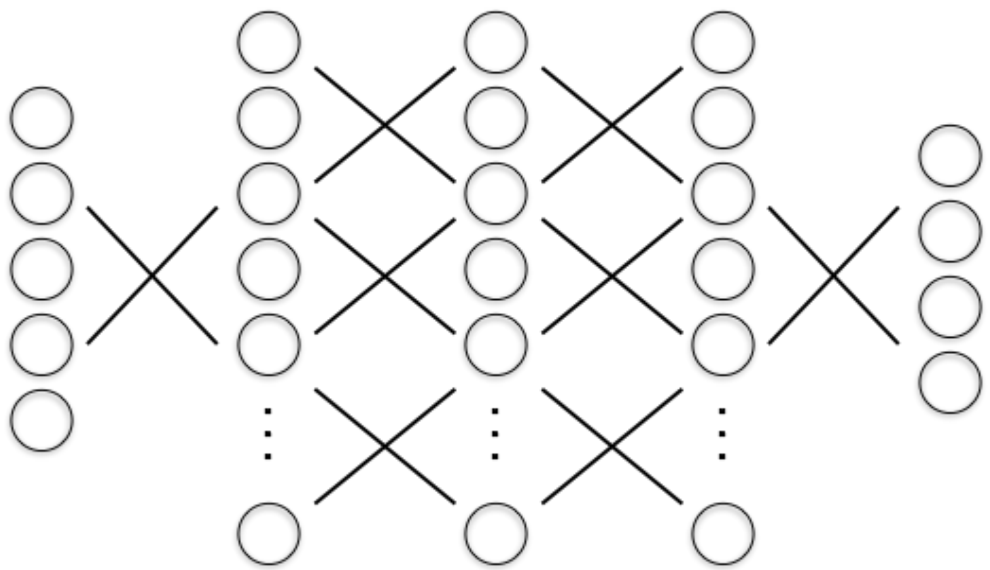
Feed-forward Neural Networks
(Part 2: learning)

Outline (part 2)

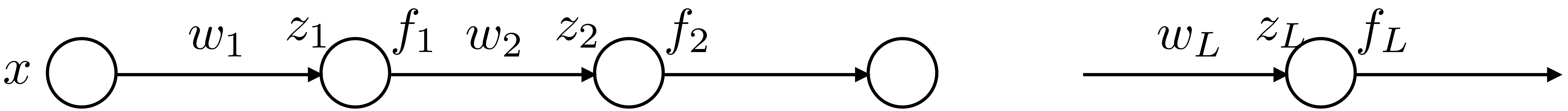
- Learning feed-forward neural networks
- SGD and back-propagation

Learning neural networks

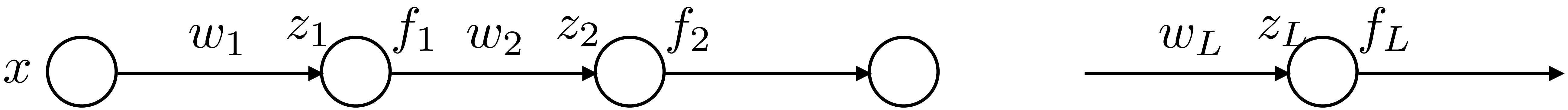


Simple example

- A long chain like neural network

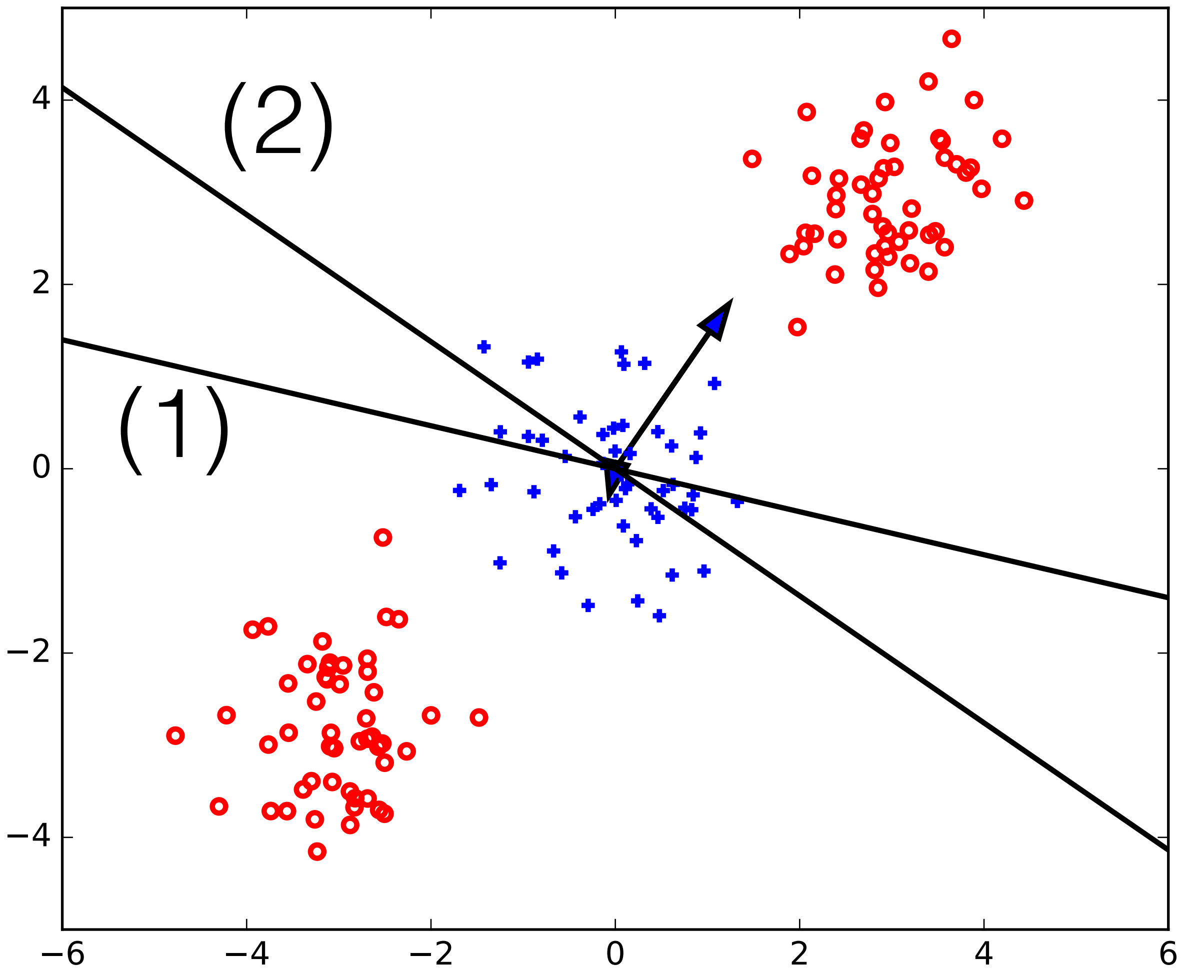


Back-propagation

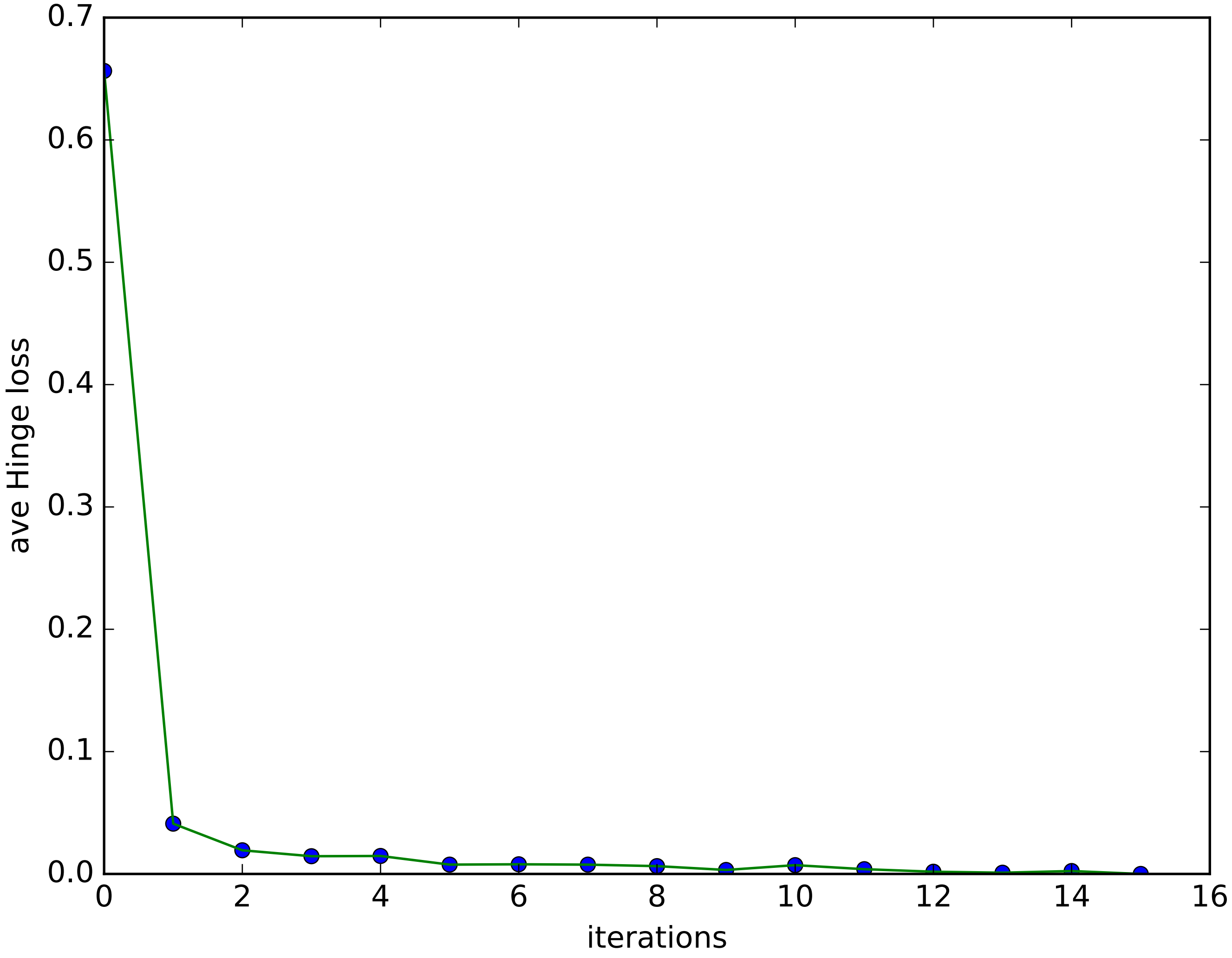


2 hidden units: training

Initial network (hidden units)

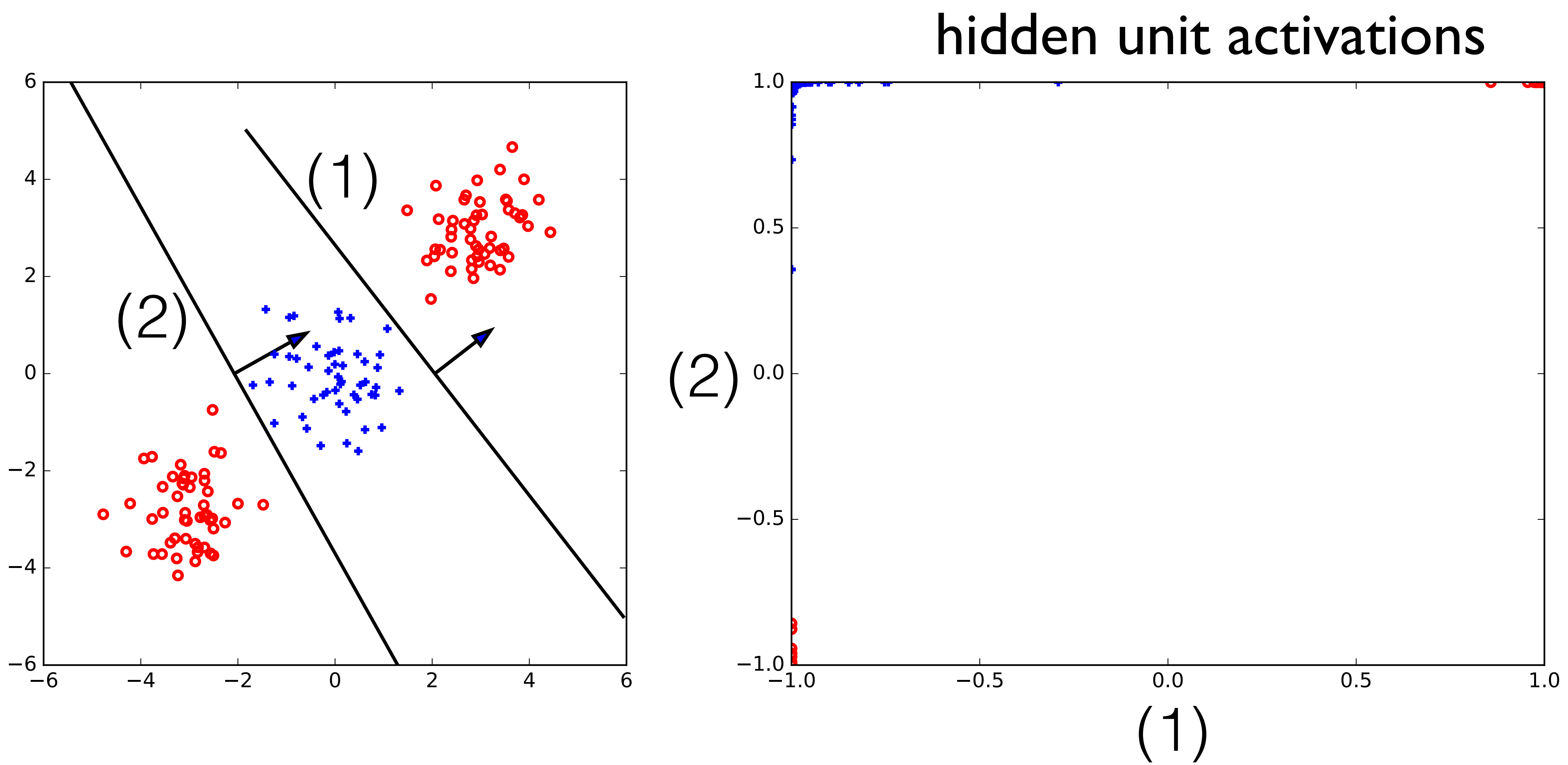


Average hinge loss per epoch



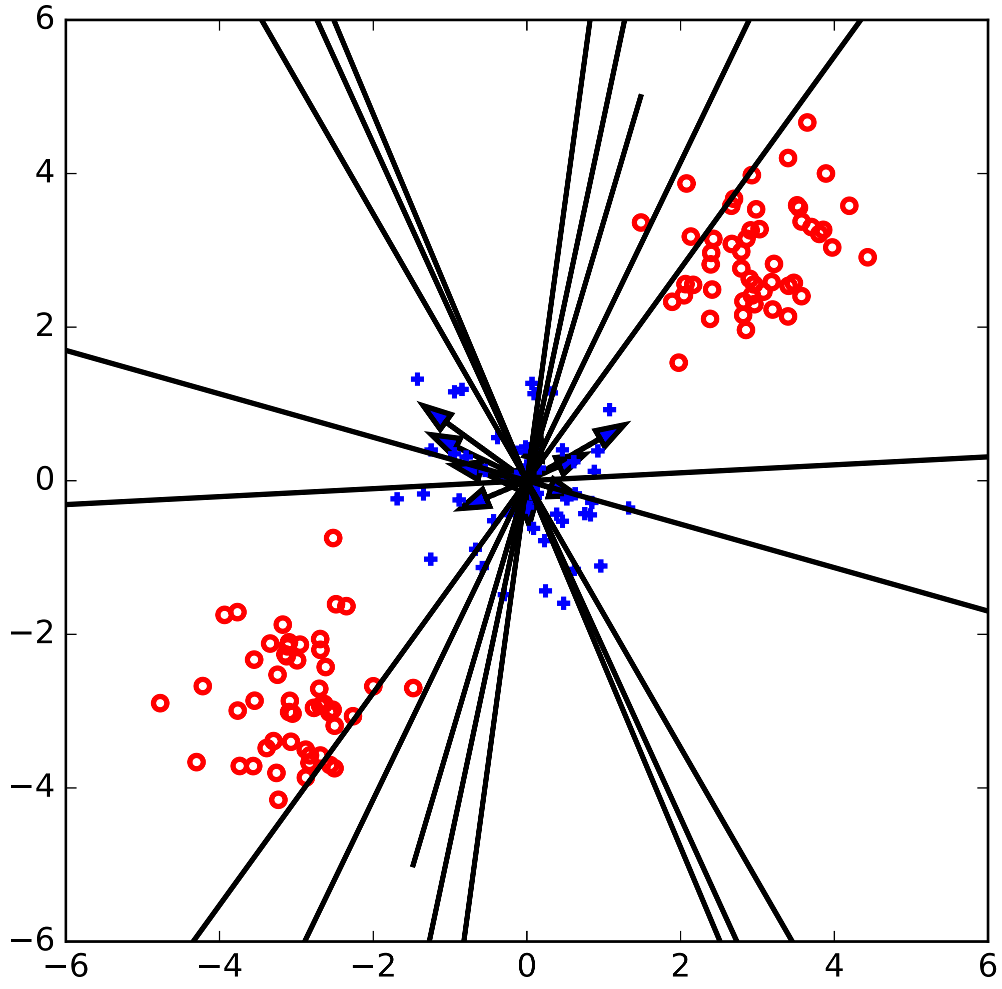
2 hidden units: training

- After ~10 passes through the data



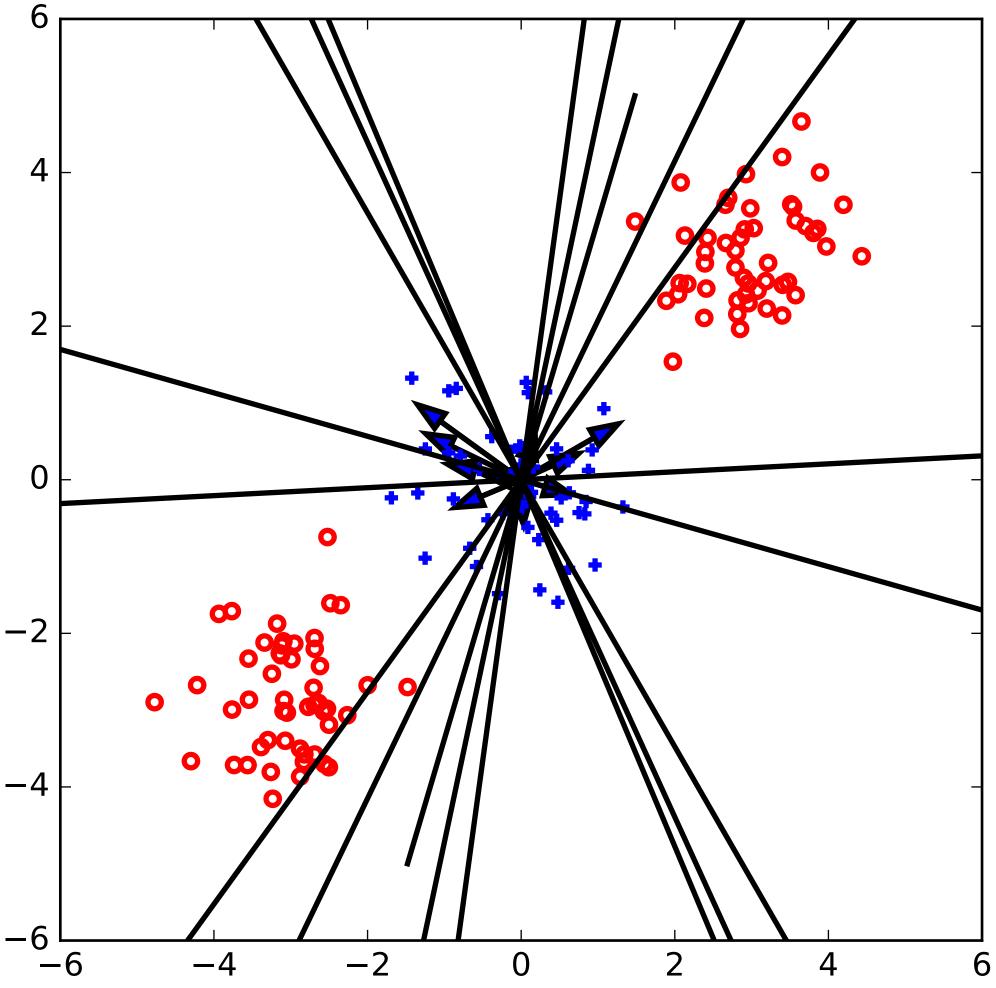
10 hidden units

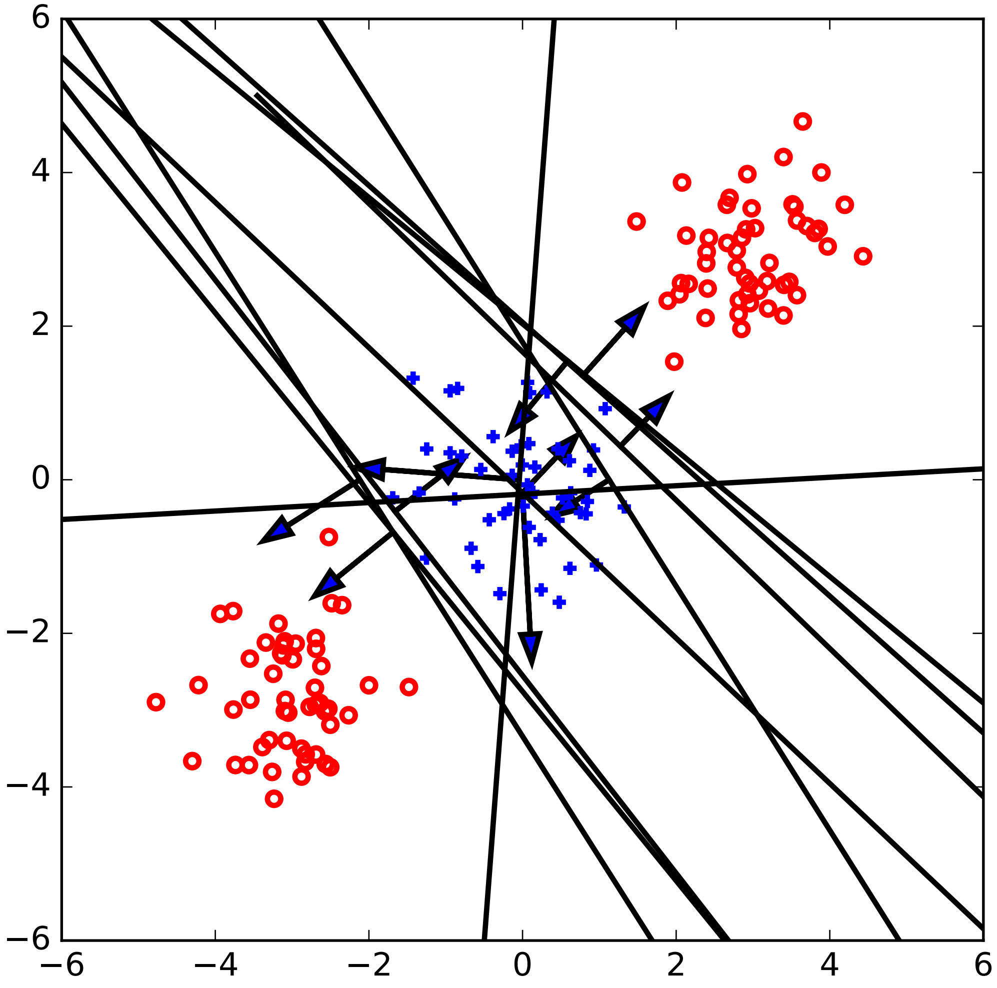
- Randomly initialized weights (zero offset) for the hidden units



10 hidden units

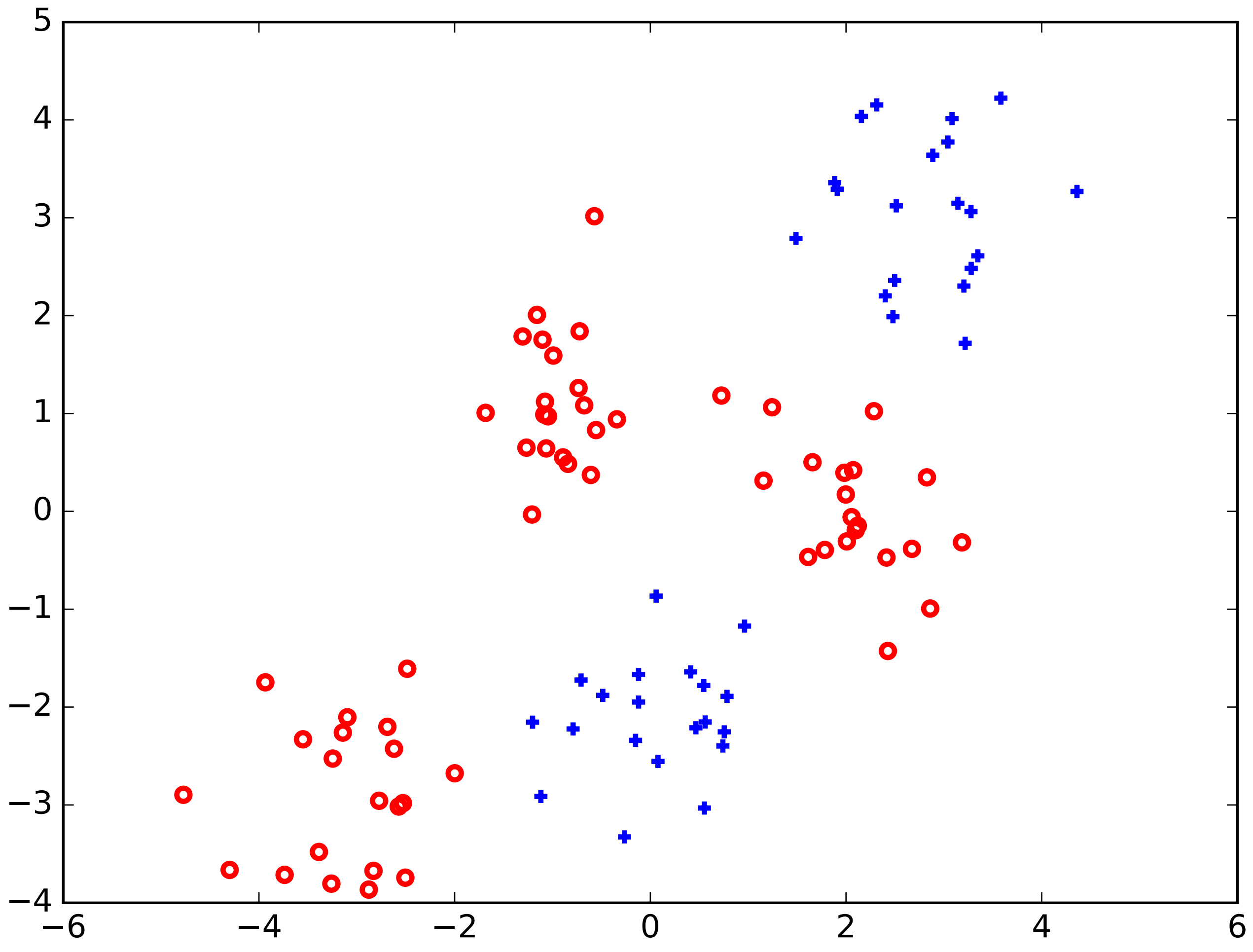
- After ~ 10 epochs the hidden units are arranged in a manner sufficient for the task (but not otherwise perfect)





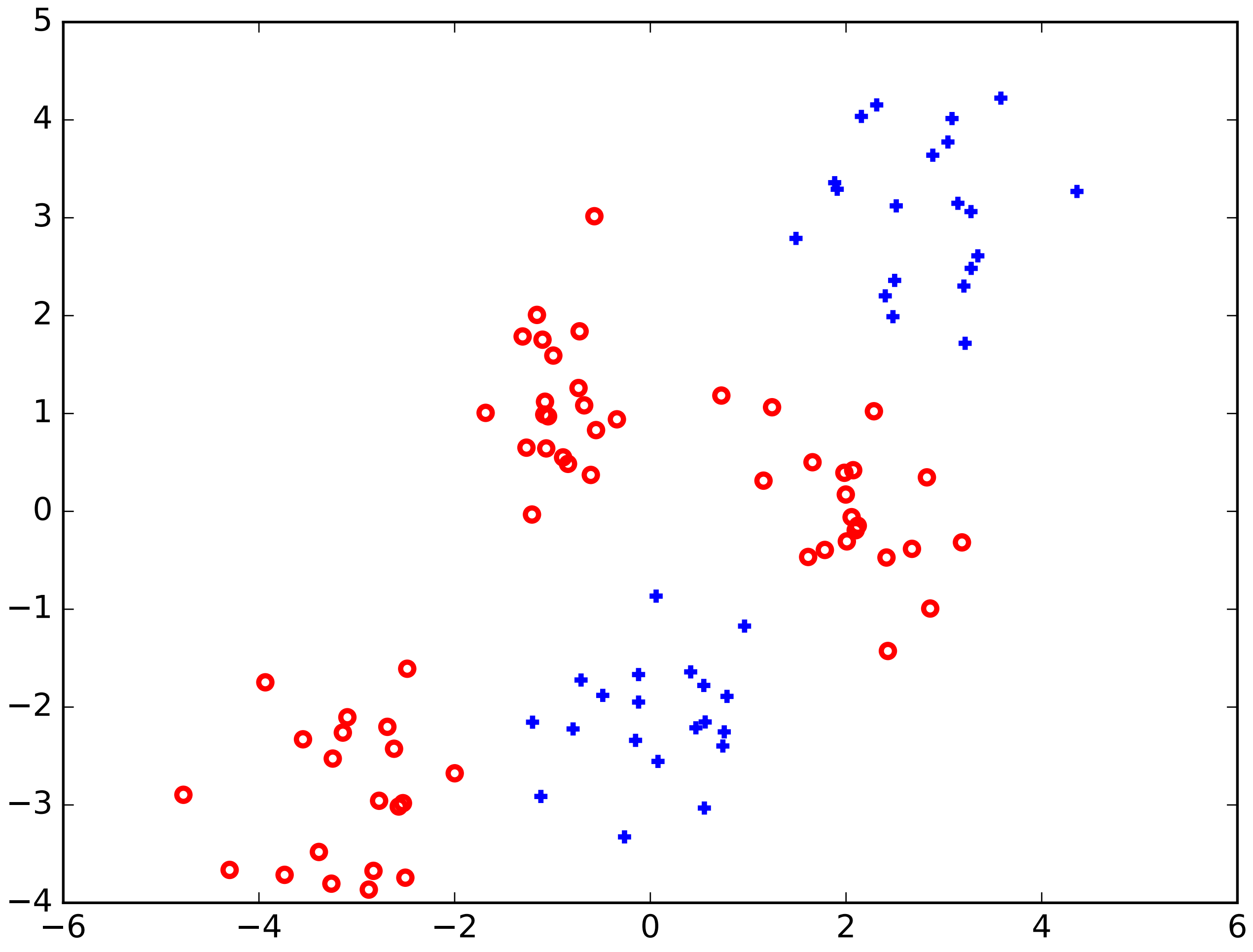
Decisions (and a harder task)

- 2 hidden units can no longer solve this task

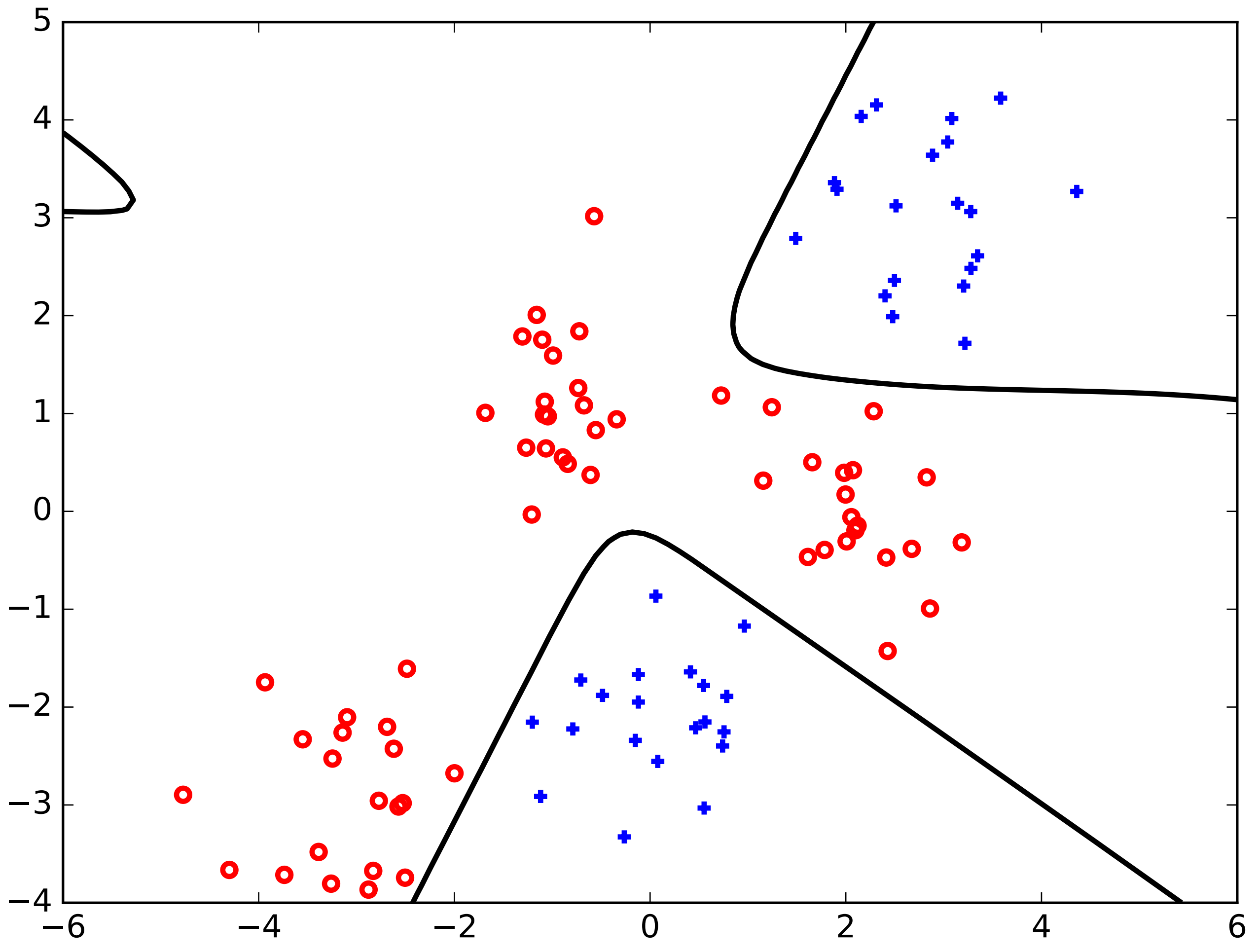


Decisions (and a harder task)

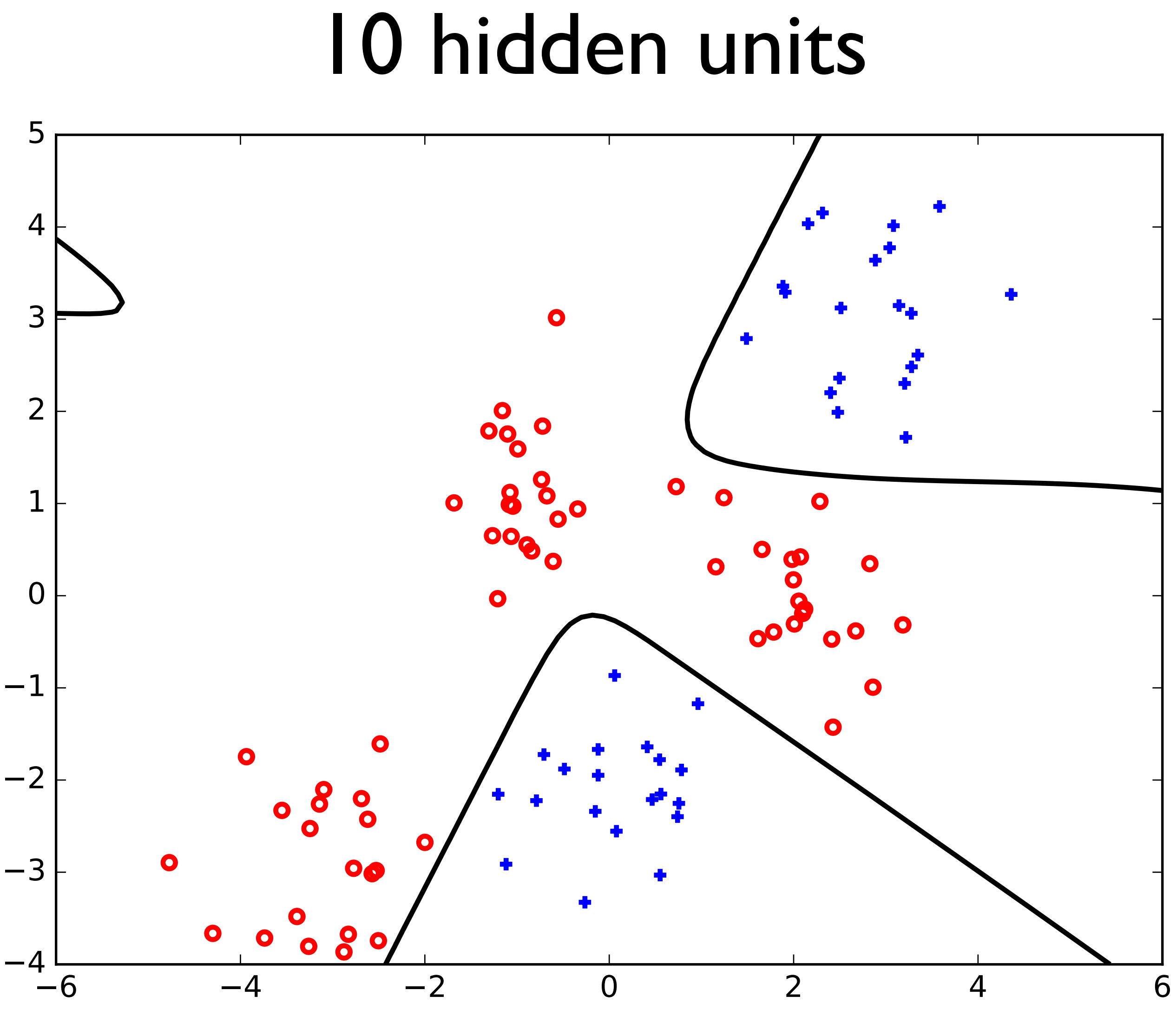
- 2 hidden units can no longer solve this task

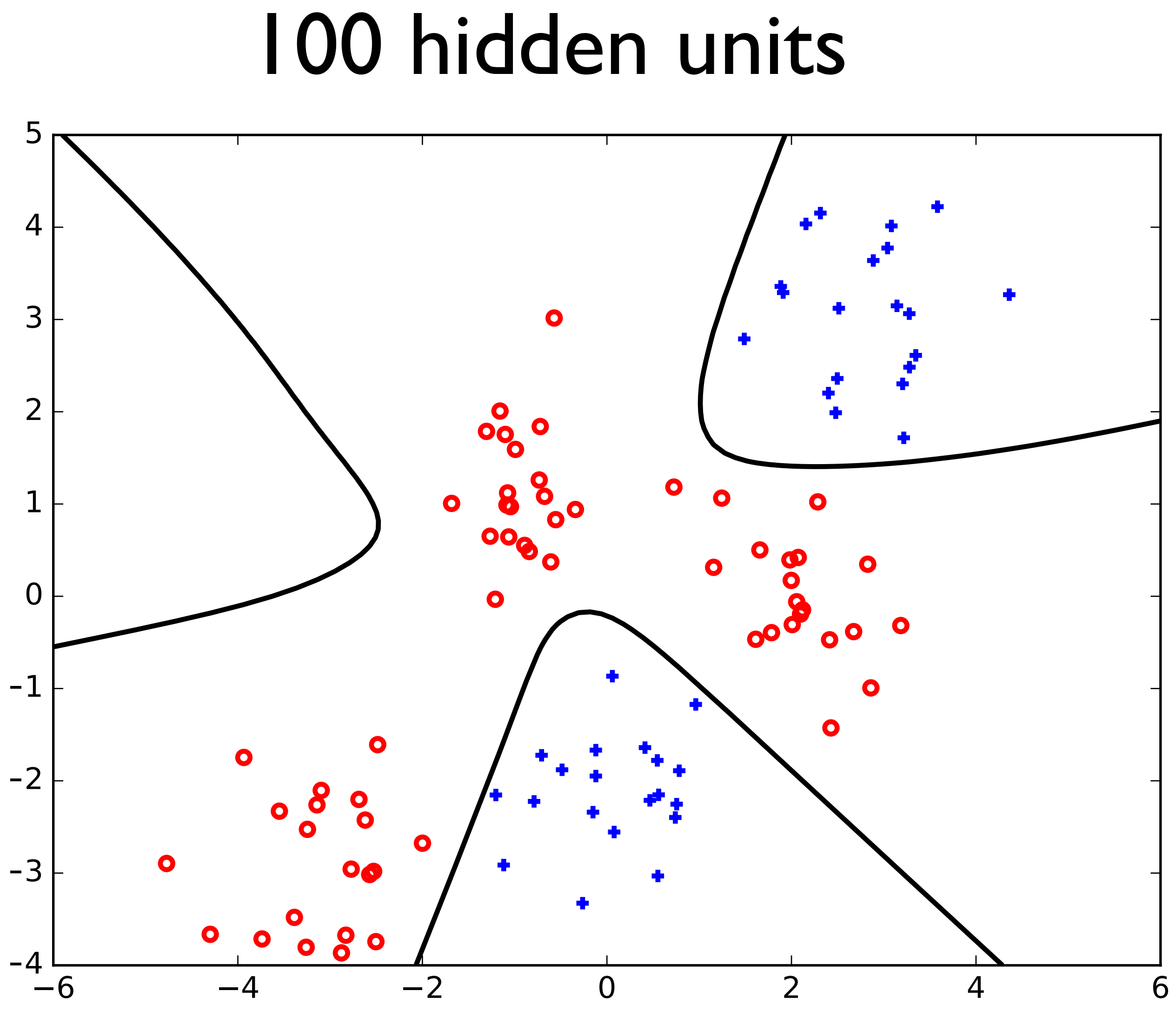


10 hidden units



Decisions (and a harder task)

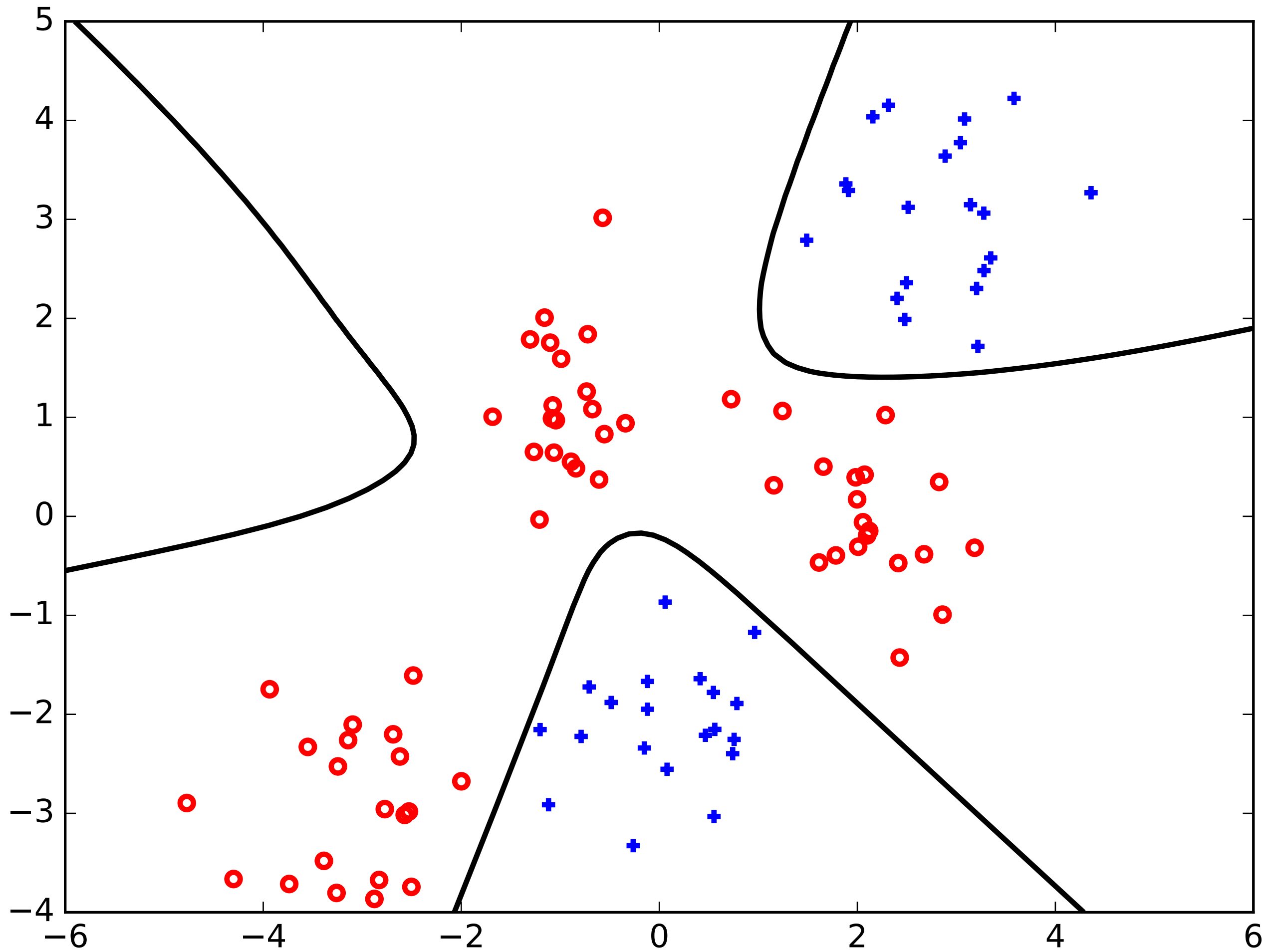




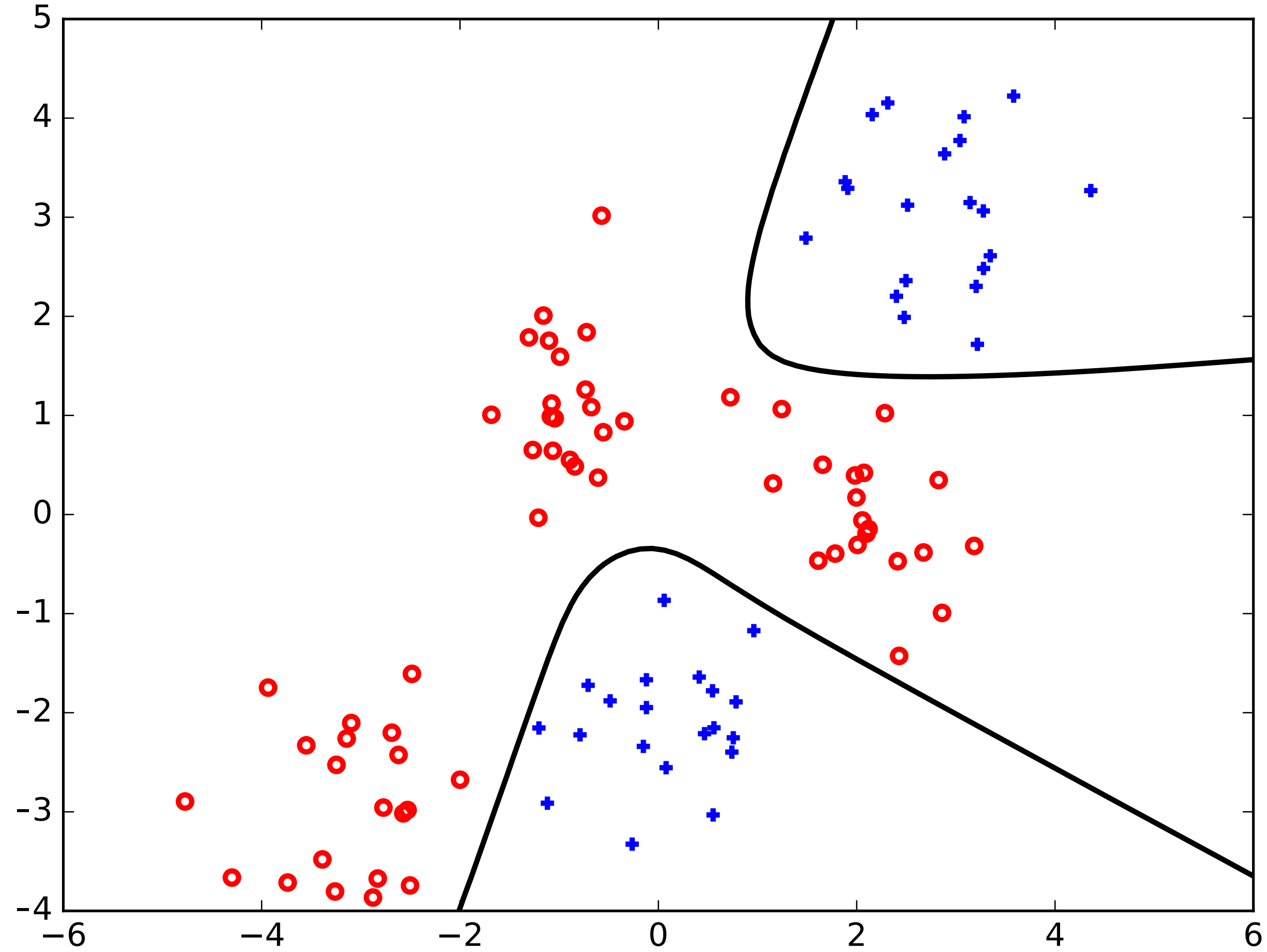
Decision boundaries

- Symmetries introduced in initialization can persist…

100 hidden units
(zero offset initialization)

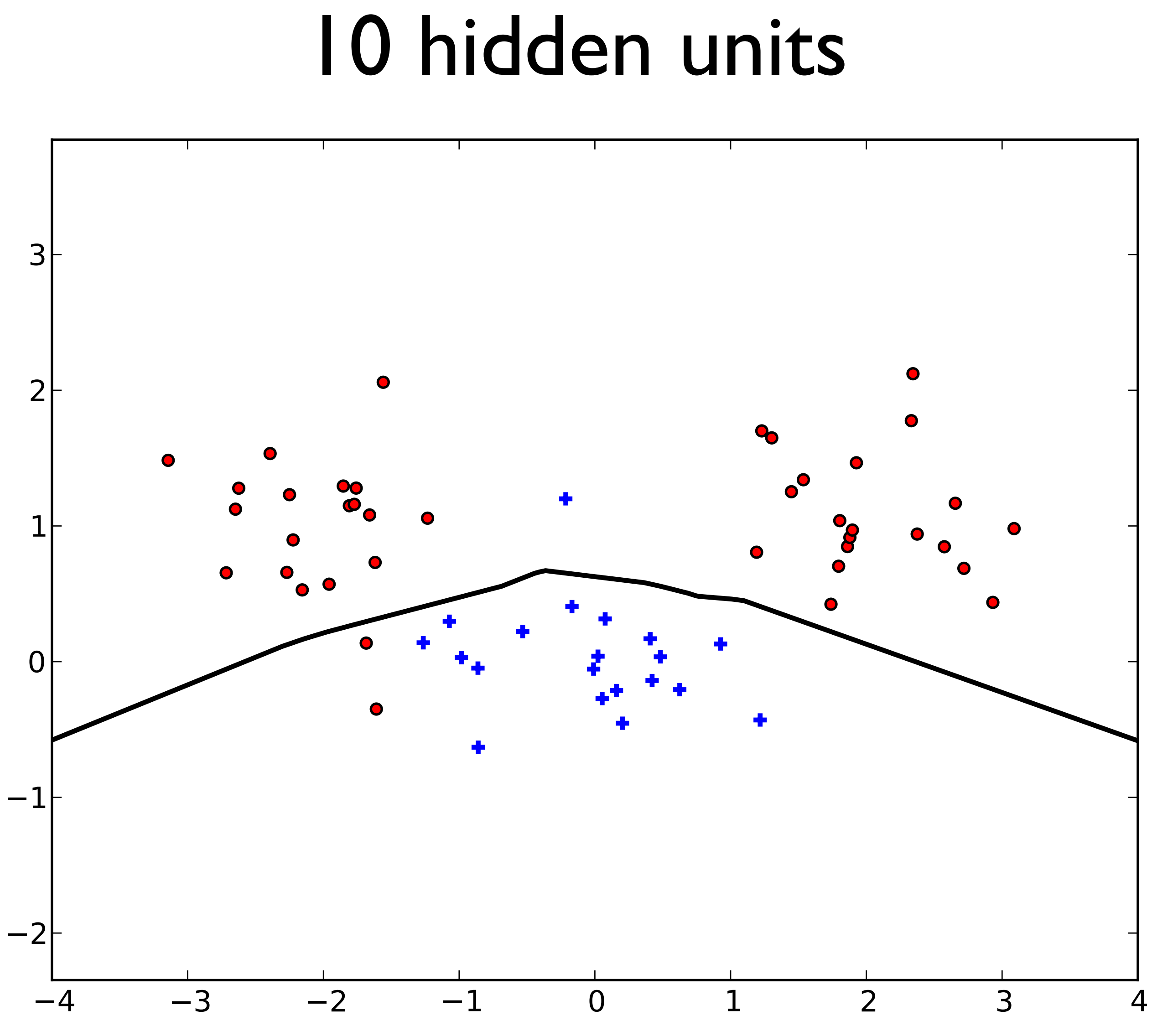


100 hidden units
(random offset initialization)



Size, optimization

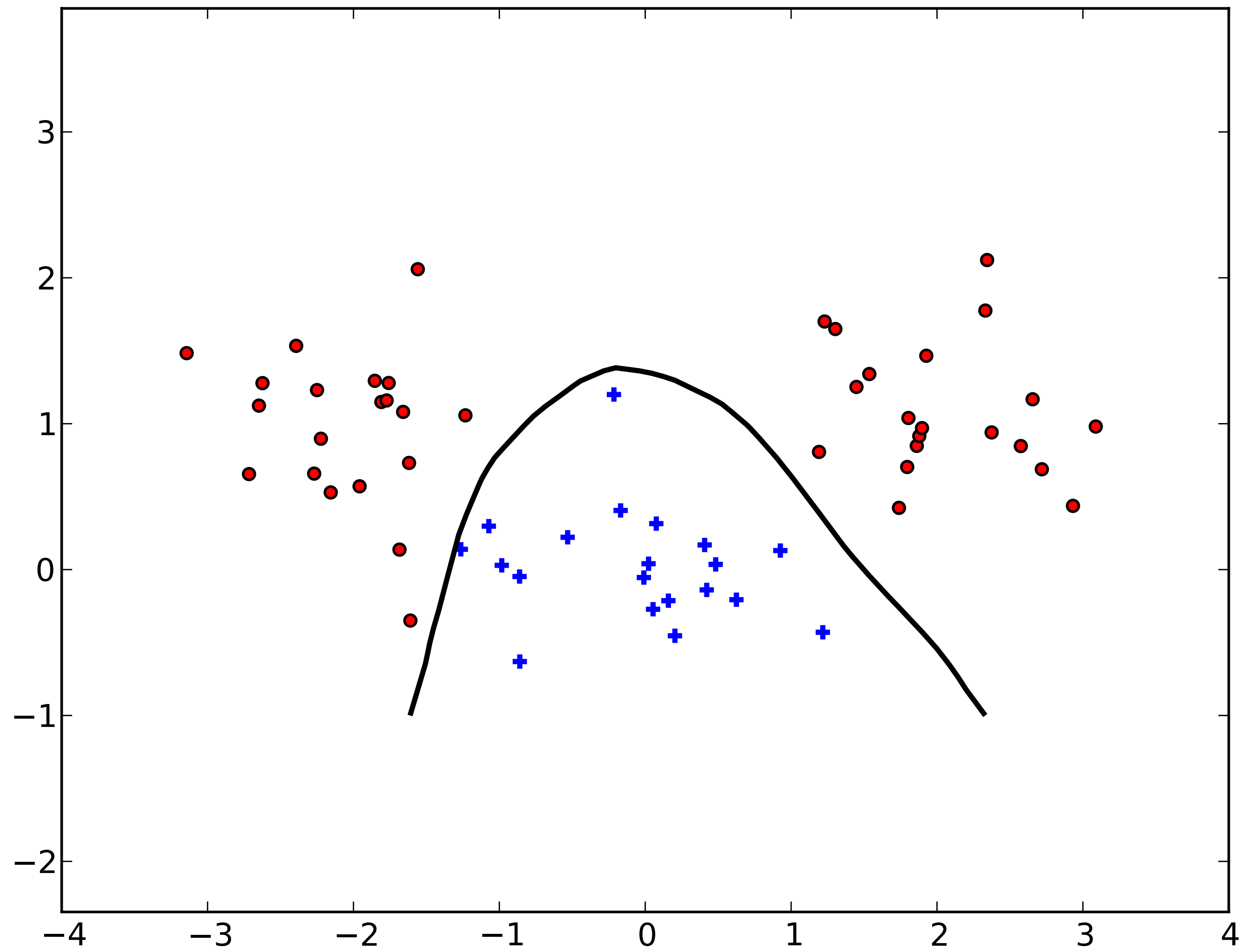
- Many recent architectures use ReLU units (cheap to evaluate, sparsity)
- Easier to learn as large models…



Size, optimization

- Many recent architectures use ReLU units (cheap to evaluate, sparsity)
- Easier to learn as large models…

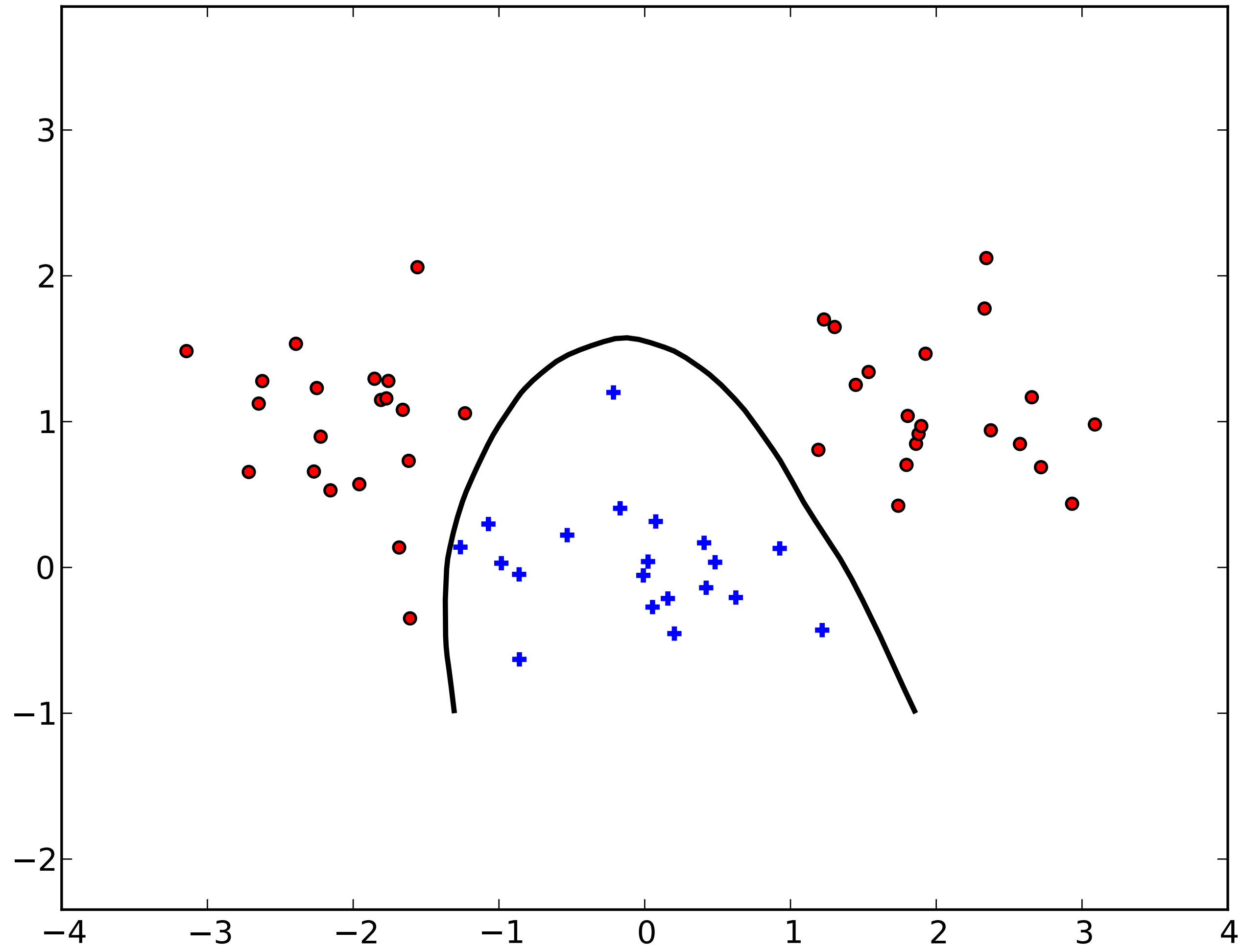
100 hidden units



Size, optimization

- Many recent architectures use ReLU units (cheap to evaluate, sparsity)
- Easier to learn as large models…

500 hidden units



Summary (part 2)

- Neural networks can be learned with SGD similarly to linear classifiers
- The derivatives necessary for SGD can be evaluated effectively via back-propagation
- Multi-layer neural network models are complicated… we are no longer guaranteed to reach global (only local) optimum with SGD
- Larger models tend to be easier to learn … units only need to be adjusted so that they are, collectively, sufficient to solve the task