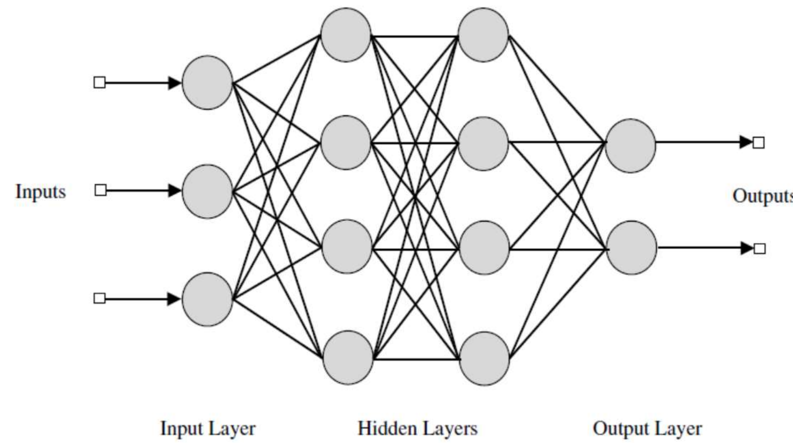
Neural Networks

Hopfield Nets and Auto Associators

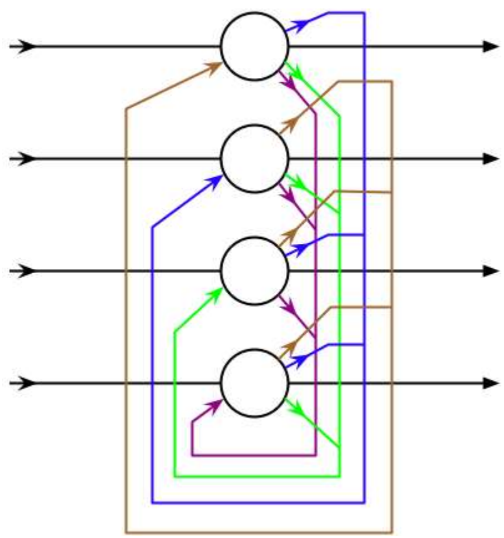
Fall 2021

Story so far

- **Neural networks for computation**
- All feedforward structures



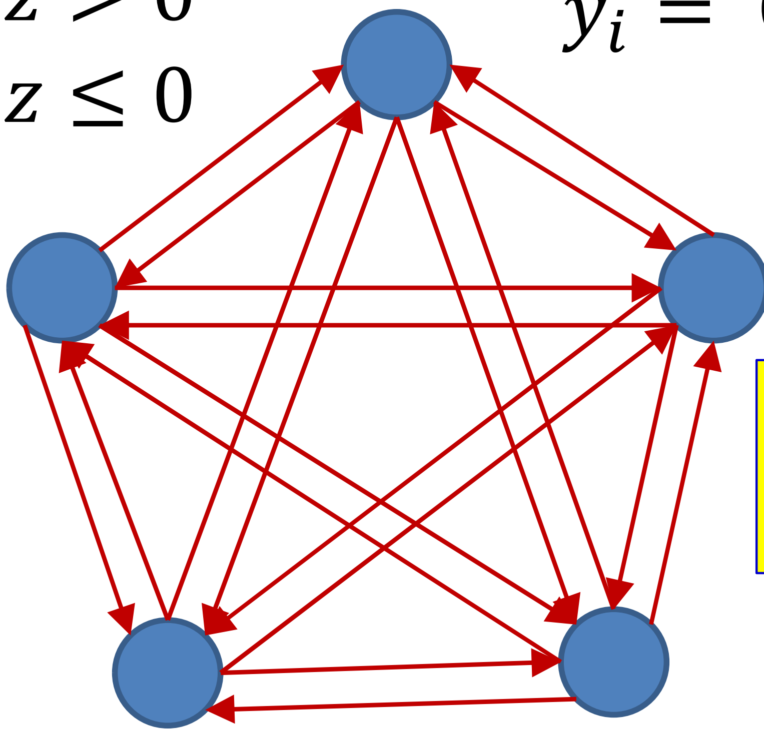
- But what about..



Consider this loopy network

$$\Theta(z) = \begin{cases} +1 \; if \; z > 0 \\ -1 \; if \; z \leq 0 \end{cases}$$

$$y_i = \Theta\left(\sum_{j \neq i} w_{ji} y_j + b_i\right)$$



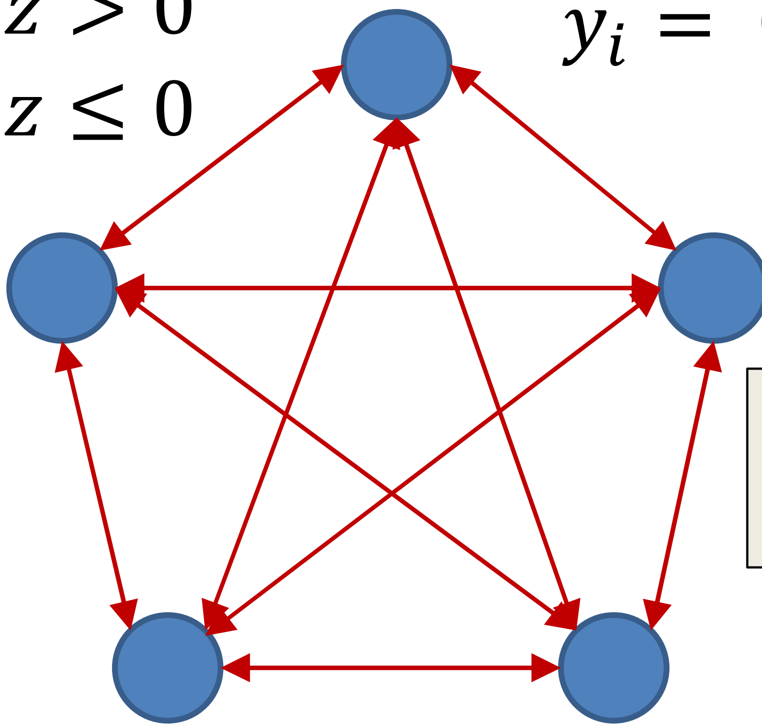
The output of a neuron affects the input to the neuron

- Each neuron is a perceptron with +1/-1 output
- Every neuron *receives* input from every other neuron
- Every neuron *outputs* signals to every other neuron

Consider this loopy network

$$\Theta(z) = \begin{cases} +1 \; if \; z > 0 \\ -1 \; if \; z \le 0 \end{cases}$$

$$y_i = \Theta\left(\sum_{j \neq i} w_{ji} y_j + b_i\right)$$



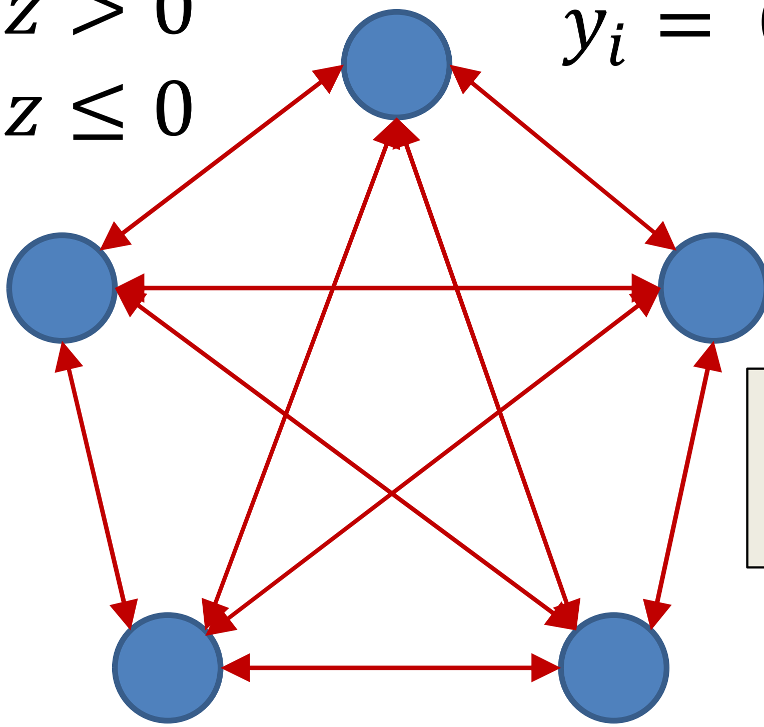
A symmetric network:
$w_{ij} = w_{ji}$

- Each neuron is a perceptron with +1/-1 output
- Every neuron *receives* input from every other neuron
- Every neuron *outputs* signals to every other neuron

Hopfield Net

$$\Theta(z) = \begin{cases} +1 \; if \; z > 0 \\ -1 \; if \; z \le 0 \end{cases}$$

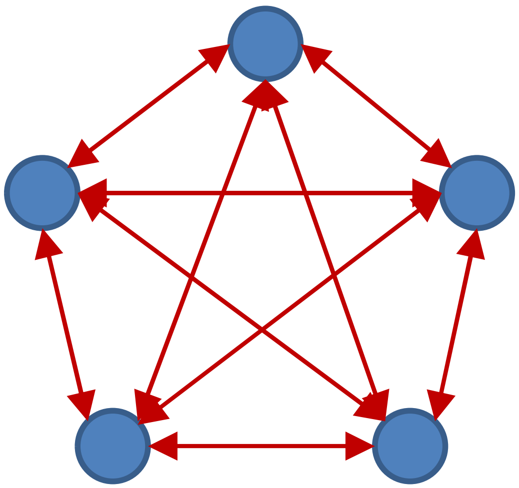
$$y_i = \Theta\left(\sum_{j \neq i} w_{ji} y_j + b_i\right)$$



A symmetric network:
$w_{ij} = w_{ji}$

- Each neuron is a perceptron with +1/-1 output
- Every neuron *receives* input from every other neuron
- Every neuron *outputs* signals to every other neuron

Loopy network

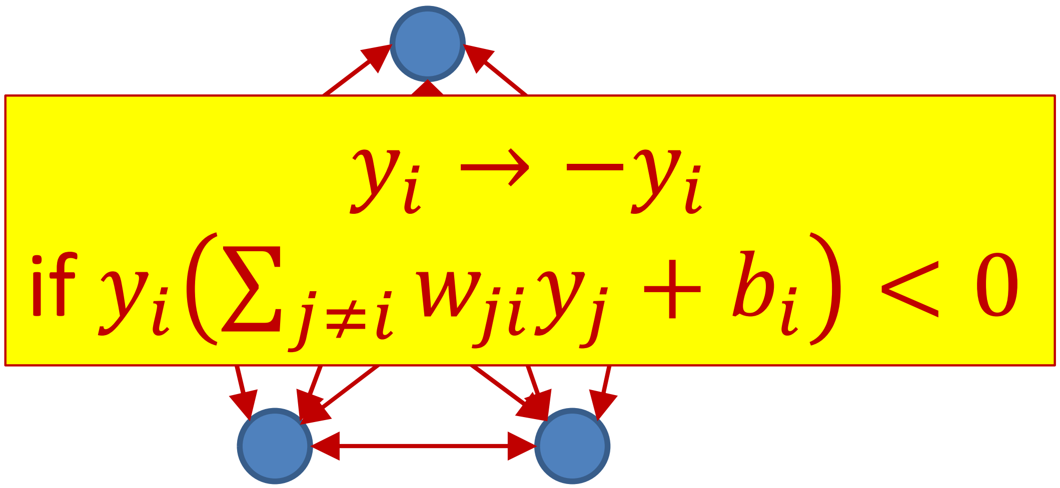


$$y_i = \Theta\left(\sum_{j \neq i} w_{ji} y_j + b_i\right)$$

$$\Theta(z) = \begin{cases} +1 \; if \; z > 0 \\ -1 \; if \; z \leq 0 \end{cases}$$

- At each time each neuron receives a “field” $\sum_{j \neq i} w_{ji} y_j + b_i$
- If the sign of the field matches its own sign, it does not respond
- If the sign of the field opposes its own sign, it “flips” to match the sign of the field

Loopy network

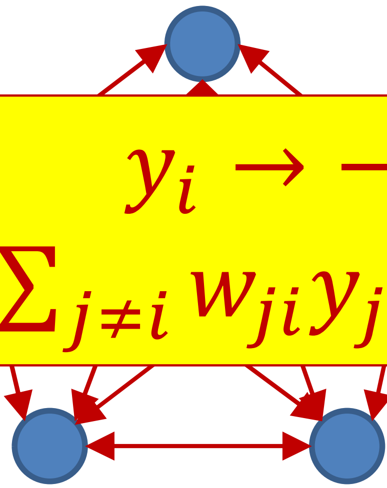


$$y_i = \Theta\left(\sum_{j\neq i} w_{ji}y_j + b_i\right)$$

$$\Theta(z) = \begin{cases} +1 \ if \ z > 0 \\ -1 \ if \ z \leq 0 \end{cases}$$

- At each time each neuron receives a "field" $\sum_{j\neq i} w_{ji}y_j + b_i$
- If the sign of the field matches its own sign, it does not respond
- If the sign of the field opposes its own sign, it "flips" to match the sign of the field

Loopy network



$$y_i = \Theta\left(\sum_{j \neq i} w_{ji} y_j + b_i\right)$$

$$\Theta(z) = \begin{cases} +1 \; if \; z > 0 \\ -1 \; if \; z \leq 0 \end{cases}$$

$$y_i \rightarrow -y_i$$

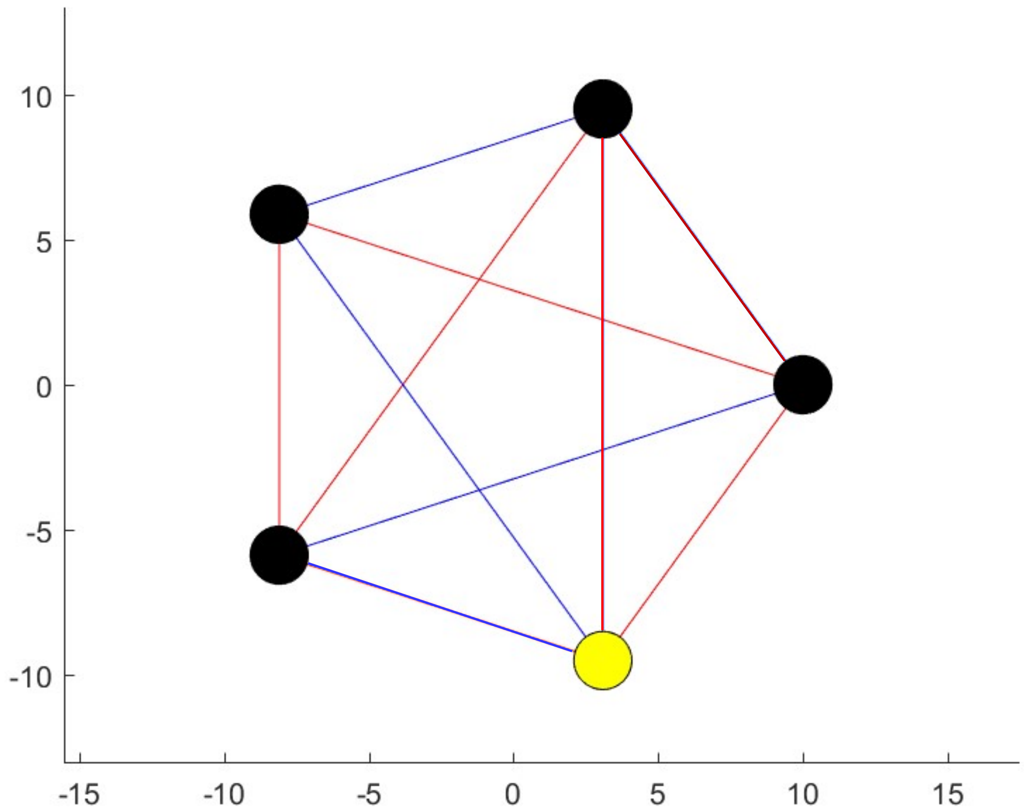
if $y_i\left(\sum_{j \neq i} w_{ji} y_j + b_i\right) < 0$

A neuron “flips” if weighted sum of other neurons’ outputs is of the opposite sign to its own current (output) value

But this may cause *other* neurons to flip!

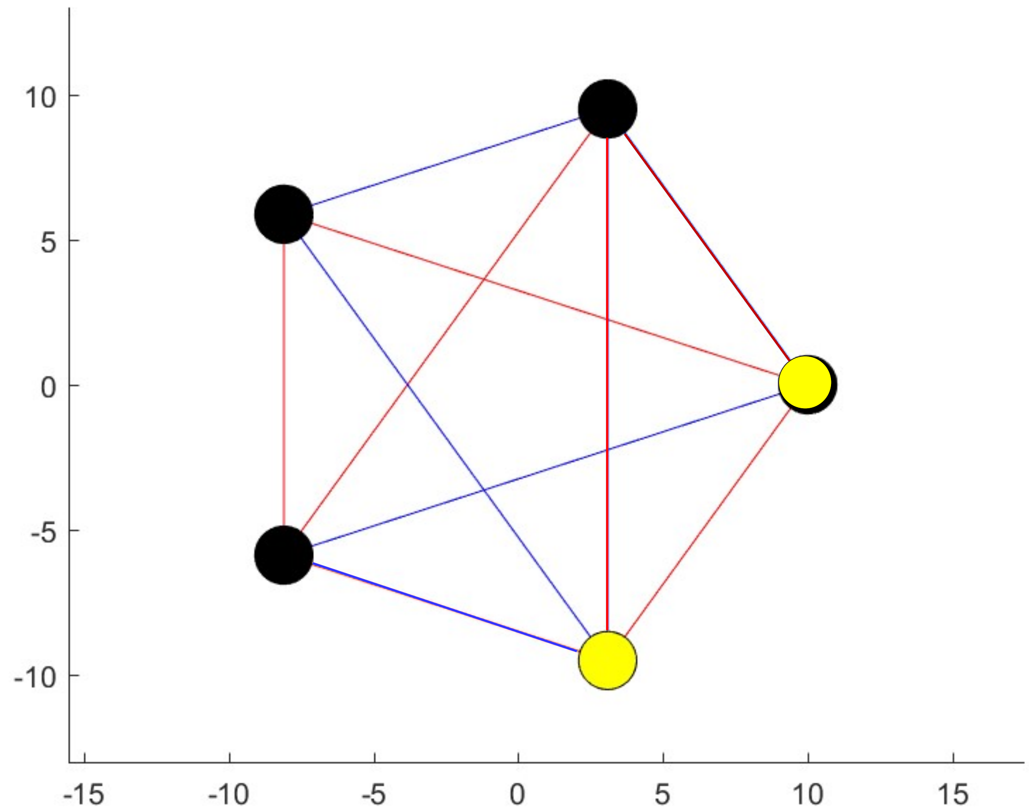
- …es a “field” $\sum_{j \neq i} w_{ji} y_j + b_i$
- …s own sign, it does not respond
- If the sign of the field opposes its own sign, it “flips” to match the sign of the field

Example



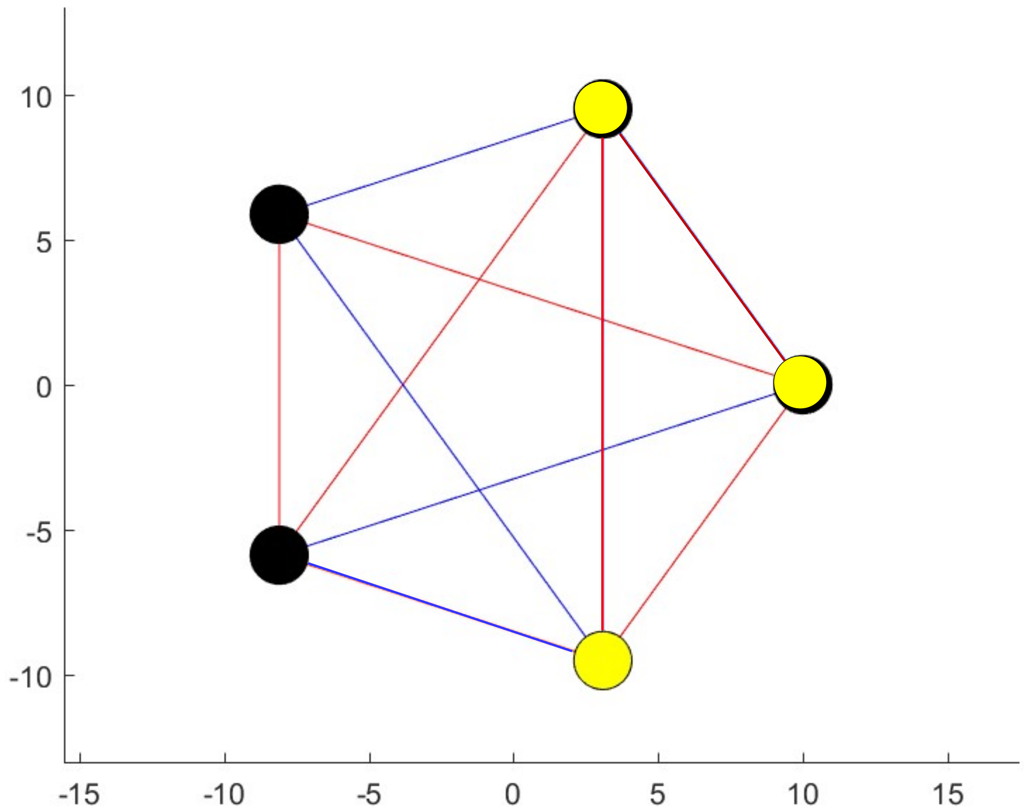
- Red edges are +1, blue edges are -1
- Yellow nodes are -1, black nodes are +1

Example



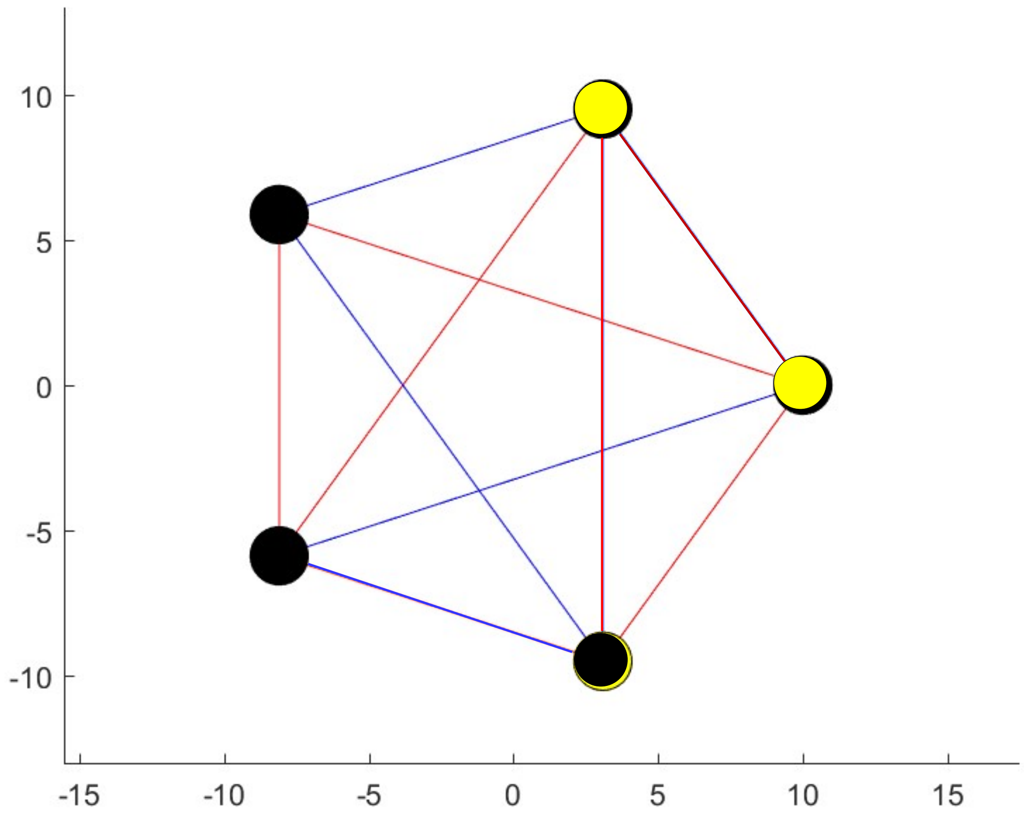
- Red edges are +1, blue edges are -1
- Yellow nodes are -1, black nodes are +1

Example



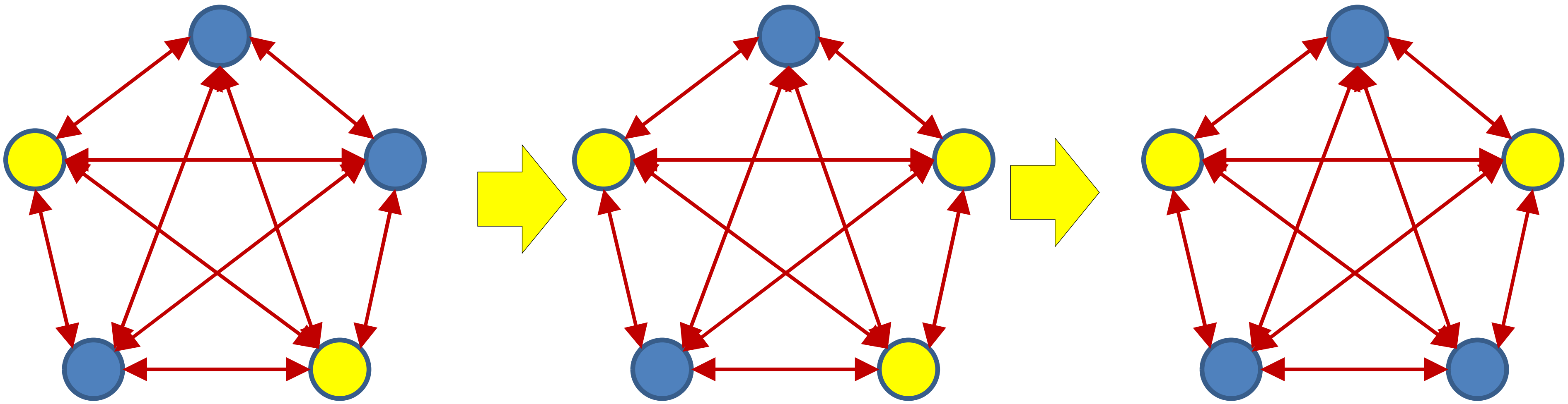
- Red edges are +1, blue edges are -1
- Yellow nodes are -1, black nodes are +1

Example



- Red edges are +1, blue edges are -1
- Yellow nodes are -1, black nodes are +1

Loopy network

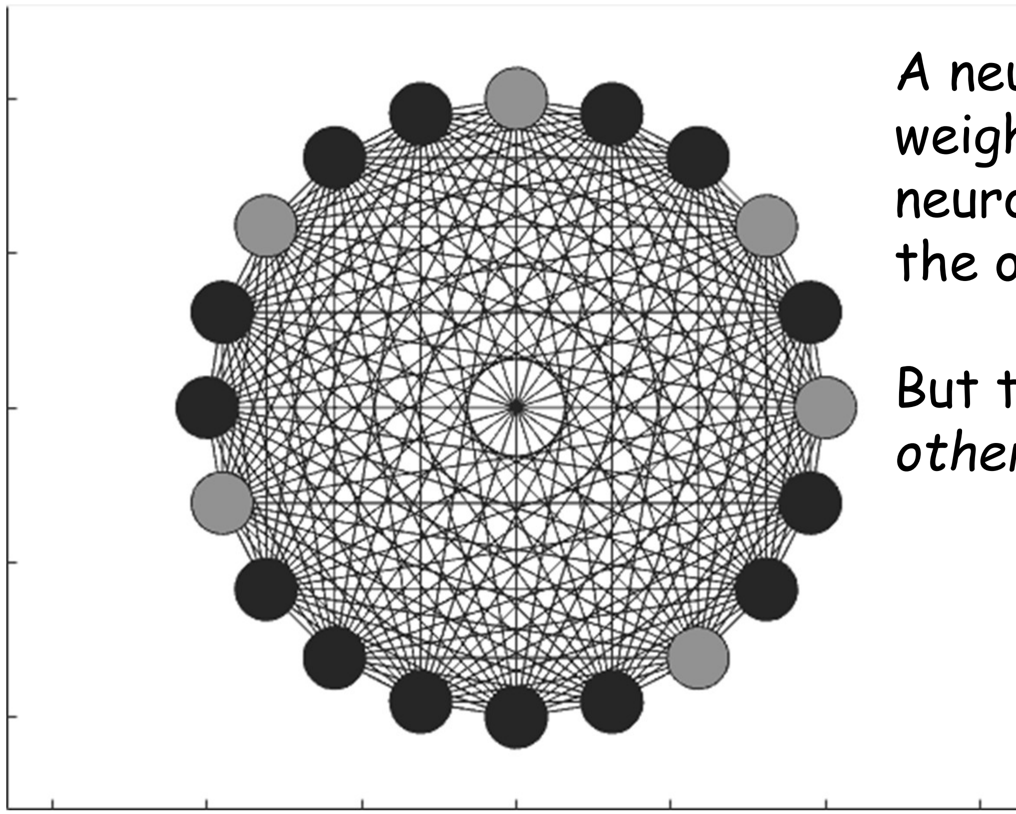


- If the sign of the field at any neuron opposes its own sign, it "flips" to match the field
  - Which will change the field at other nodes
    - Which may then flip
      - Which may cause other neurons including the first one to flip…
        - » And so on…

20 evolutions of a loopy net

$$\Theta(z) = \begin{cases} +1 \; if \; z > 0 \\ -1 \; if \; z \leq 0 \end{cases}$$

$$y_i = \Theta\left(\sum_{j \neq i} w_{ji} y_j + b_i\right)$$

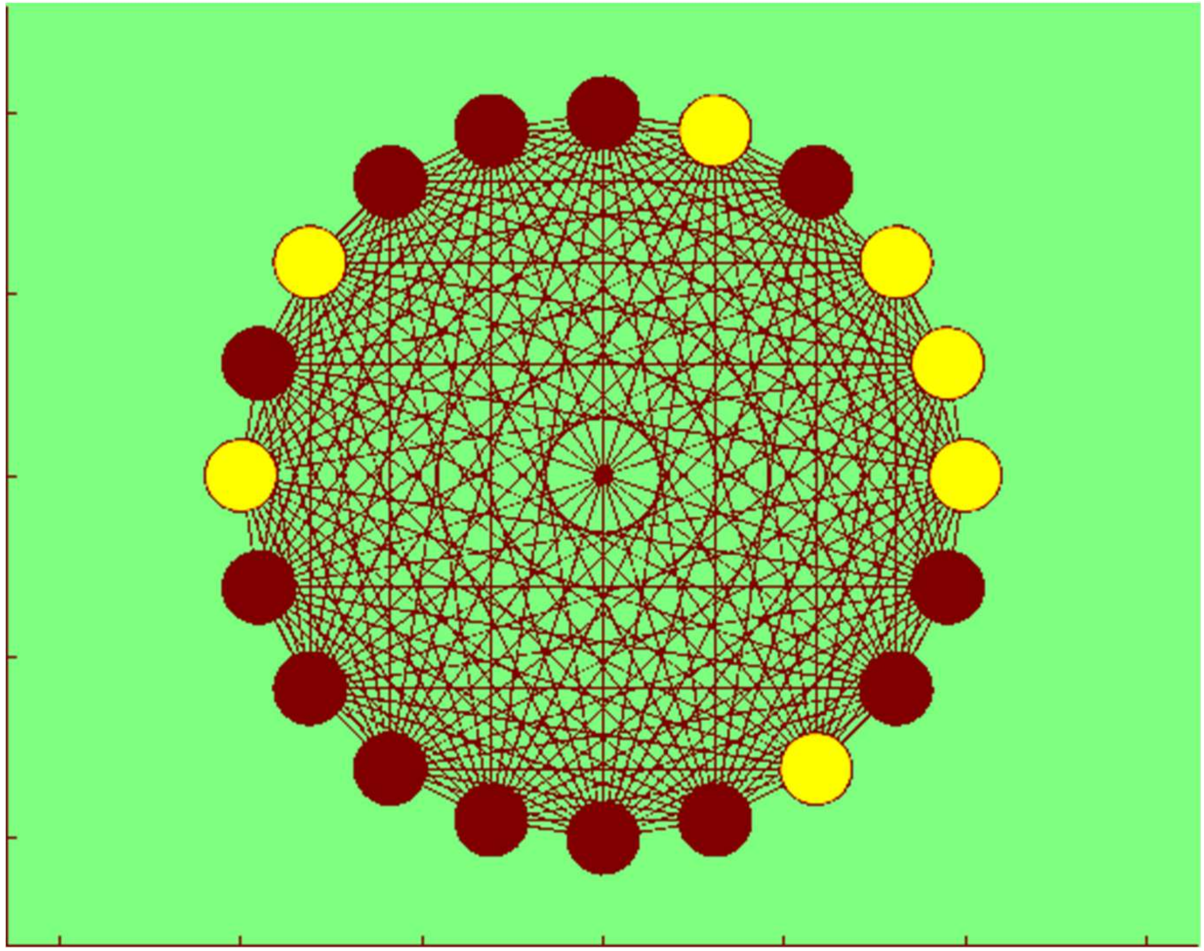


A neuron "flips" if weighted sum of other neuron's outputs is of the opposite sign

But this may cause *other* neurons to flip!

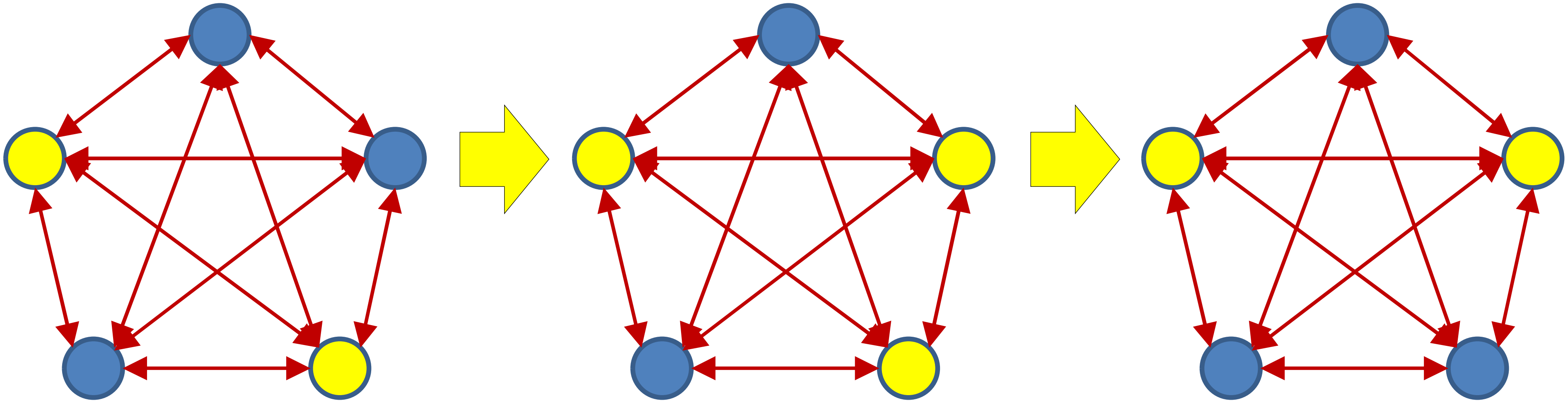
- All neurons which do not "align" with the local field "flip"

120 evolutions of a loopy net



- All neurons which do not “align” with the local field “flip”

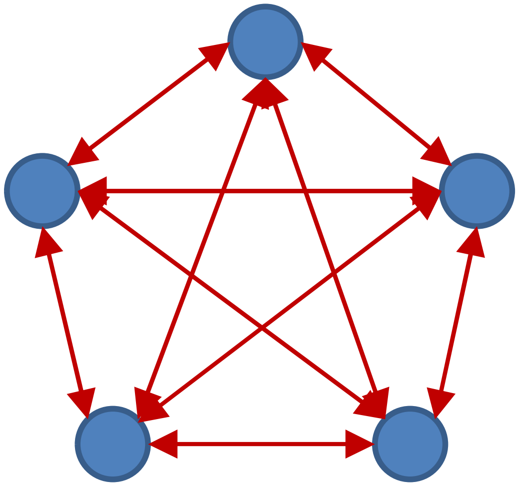
Loopy network



- If the sign of the field at any neuron opposes its own sign, it "flips" to match the field
  - Which will change the field at other nodes
    - Which may then flip
      - Which may cause other neurons including the first one to flip…

- *Will this behavior continue for ever??*

Loopy network



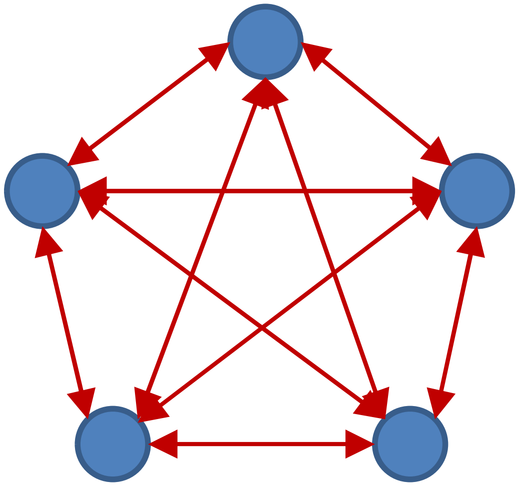
$$y_i = \Theta\left(\sum_{j \neq i} w_{ji} y_j + b_i\right)$$

$$\Theta(z) = \begin{cases} +1 \; if \; z > 0 \\ -1 \; if \; z \leq 0 \end{cases}$$

- Let y_i^- be the output of the *i*-th neuron just *before* it responds to the current field
- Let y_i^+ be the output of the *i*-th neuron just *after* it responds to the current field
- If $y_i^- = sign\left(\sum_{j \neq i} w_{ji} y_j + b_i\right)$, then $y_i^+ = y_i^-$
  - If the sign of the field matches its own sign, it does not flip

$$y_i^+\left(\sum_{j \neq i} w_{ji} y_j + b_i\right) - y_i^-\left(\sum_{j \neq i} w_{ji} y_j + b_i\right) = 0$$

Loopy network



$$y_i = \Theta\left(\sum_{j \neq i} w_{ji} y_j + b_i\right)$$

$$\Theta(z) = \begin{cases} +1 \; if \; z > 0 \\ -1 \; if \; z \leq 0 \end{cases}$$

- If $y_i^- \neq sign\left(\sum_{j \neq i} w_{ji} y_j + b_i\right)$, then $y_i^+ = -y_i^-$

$$y_i^+\left(\sum_{j \neq i} w_{ji} y_j + b_i\right) - y_i^-\left(\sum_{j \neq i} w_{ji} y_j + b_i\right) = 2y_i^+\left(\sum_{j \neq i} w_{ji} y_j + b_i\right)$$

  – This term is always positive!

- *Every flip of a neuron is guaranteed to locally increase*

$$y_i\left(\sum_{j \neq i} w_{ji} y_j + b_i\right)$$

Globally

- Consider the following sum across *all* nodes

$$D(y_1, y_2, \ldots, y_N) = \sum_i y_i \left(\sum_{j \neq i} w_{ji} y_j + b_i \right)$$

$$= \sum_{i, j \neq i} w_{ij} y_i y_j + \sum_i b_i y_i$$

  - Assume $w_{ii} = 0$

- For any unit k that "flips" because of the local field

$$\Delta D(y_k) = D(y_1, \ldots, y_k^+, \ldots, y_N) - D(y_1, \ldots, y_k^-, \ldots, y_N)$$

- This is strictly positive

$$\Delta D(y_k) = 2y_k^+ \left(\sum_{j \neq k} w_{jk} y_j + b_k \right)$$

Upon flipping a single unit

$$\Delta D(y_k) = D(y_1, \dots, y_k^+, \dots, y_N) - D(y_1, \dots, y_k^-, \dots, y_N)$$

- Expanding

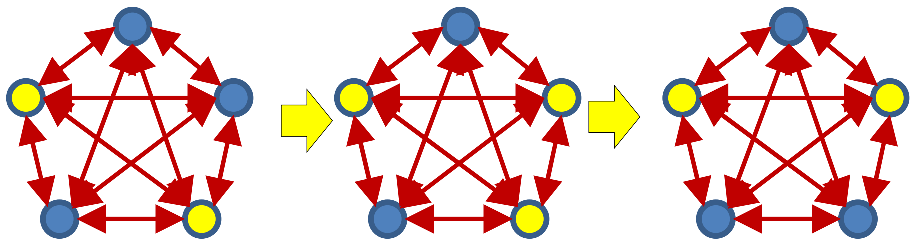
$$\Delta D(y_k) = (y_k^+ - y_k^-)\left(\sum_{j \neq k} w_{jk} y_j + b_k\right)$$

  - All other terms that do not include y_k cancel out

- This is always positive!

- *Every flip of a unit results in an increase in* D

Hopfield Net



- Flipping a unit will result in an increase (non-decrease) of

$$D = \sum_{i,j\neq i} w_{ij} y_i y_j + \sum_i b_i y_i$$

- D is bounded

$$D_{max} = \sum_{i,j\neq i} |w_{ij}| + \sum_i |b_i|$$

- The minimum increment of D in a flip is

$$\Delta D_{min} = \min_{i,\,\{y_i,\, i=1..N\}} 2 \left| \sum_{j\neq i} w_{ji} y_j + b_i \right|$$

- Any sequence of flips must converge in a finite number of steps

The Energy of a Hopfield Net

- Define the *Energy* of the network as

$$E = -\frac{1}{2}\left(\sum_{i,j\neq i} w_{ij} y_i y_j - \sum_i b_i y_i\right)$$

  – Just 0.5 times the negative of D
    - The 0.5 is only needed for convention
- The evolution of a Hopfield network constantly decreases its energy

Story so far

- A Hopfield network is a loopy binary network with symmetric connections
- Every neuron in the network attempts to “align” itself with the sign of the weighted combination of outputs of other neurons
  - The local “field”
- Given an initial configuration, neurons in the net will begin to “flip” to align themselves in this manner
  - Causing the field at other neurons to change, potentially making them flip
- Each evolution of the network is guaranteed to decrease the “energy” of the network
  - The energy is lower bounded and the decrements are upper bounded, so the network is guaranteed to converge to a stable state in a finite number of steps

Poll 1

Poll 1

Hopfield networks are loopy networks whose output activations “evolve” over time

- **True**
- False

Hopfield networks will evolve continuously, forever

- True
- **False**

Hopfield networks can also be viewed as infinitely deep shared parameter MLPs

- **True**
- False

The Energy of a Hopfield Net

- Define the *Energy* of the network as

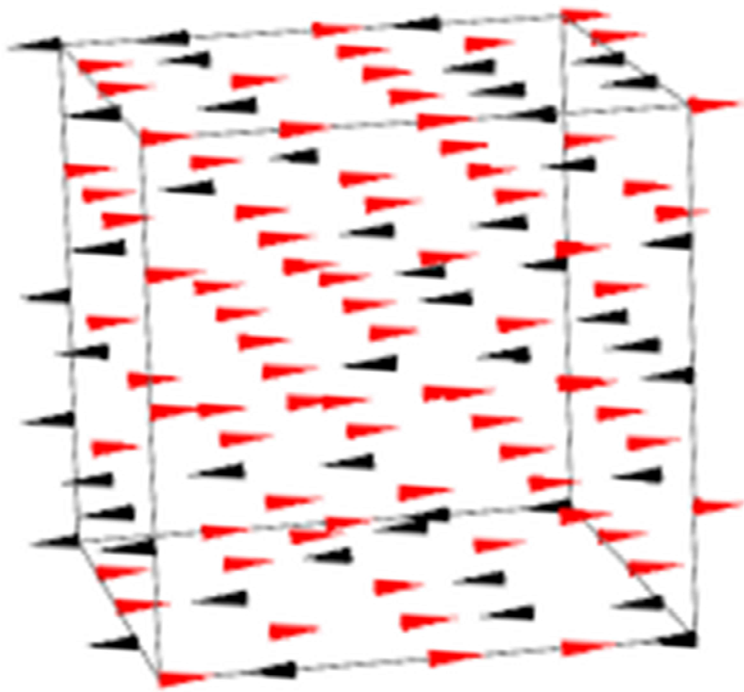
$$E = -\frac{1}{2}\left(\sum_{i,j\neq i} w_{ij} y_i y_j - \sum_i b_i y_i\right)$$

  – Just 0.5 times the negative of D

- The evolution of a Hopfield network constantly decreases its energy

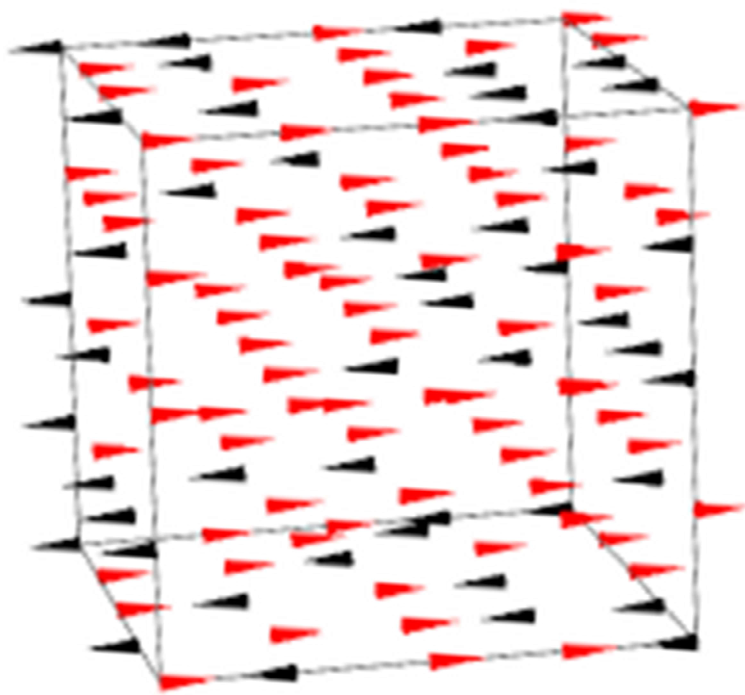
- Where did this "energy" concept suddenly sprout from?

Analogy: Spin Glass



- Magnetic diploes in a disordered magnetic material
- Each dipole tries to *align* itself to the local field
  - In doing so it may flip
- This will change fields at *other* dipoles
  - Which may flip
- Which changes the field at the current dipole…

Analogy: Spin Glasses

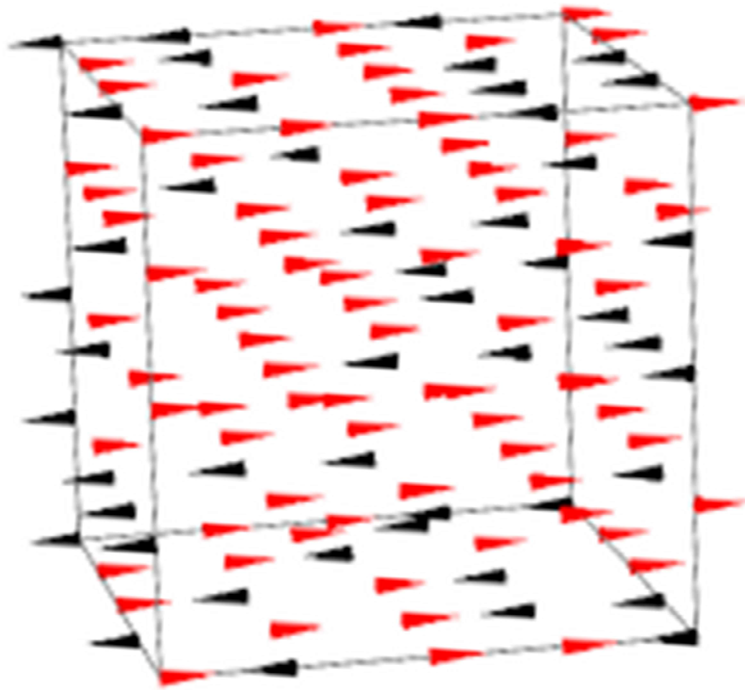


Total field at current dipole:

$$f(p_i) = \underbrace{\sum_{j \neq i} J_{ji} x_j}_{\text{intrinsic}} + \underbrace{b_i}_{\text{external}}$$

- p_i is vector position of i-th dipole
- The field at any dipole is the sum of the field contributions of all other dipoles
- The contribution of a dipole to the field at any point depends on interaction J
  - Derived from the "Ising" model for magnetic materials (Ising and Lenz, 1924)

Analogy: Spin Glasses



Total field at current dipole:

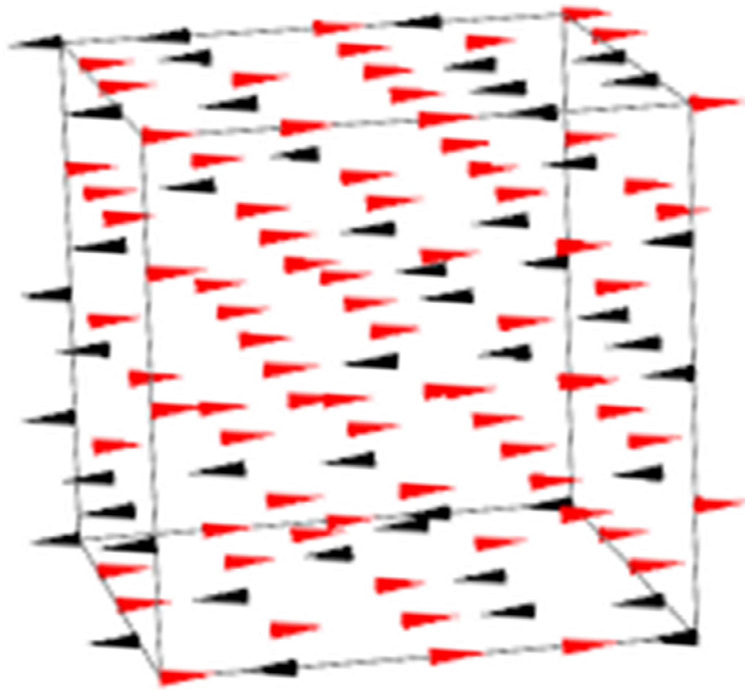
$$f(p_i) = \sum_{j \neq i} J_{ji} x_j + b_i$$

Response of current dipole

$$x_i = \begin{cases} x_i \; if \; sign\big(x_i \, f(p_i)\big) = 1 \\ -x_i \; otherwise \end{cases}$$

- A Dipole flips if it is misaligned with the field in its location

Analogy: Spin Glasses



Total field at current dipole:

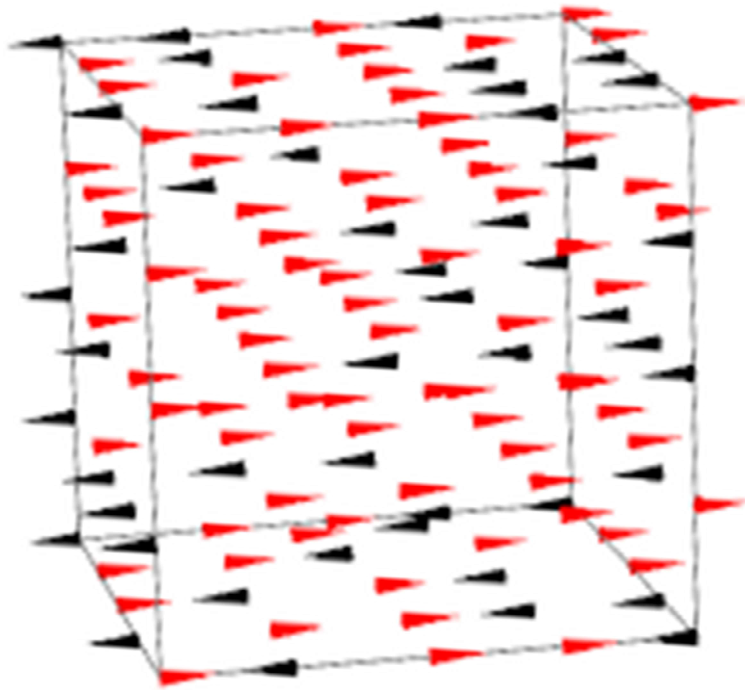
$$f(p_i) = \sum_{j \neq i} J_{ji} x_j + b_i$$

Response of current dipole

$$x_i = \begin{cases} x_i \ if \ sign\big(x_i \ f(p_i)\big) = 1 \\ -x_i \ otherwise \end{cases}$$

- Dipoles will keep flipping
  - A flipped dipole changes the field at other dipoles
    - Some of which will flip
  - Which will change the field at the current dipole
    - Which may flip
  - Etc..

Analogy: Spin Glasses



Total field at current dipole:

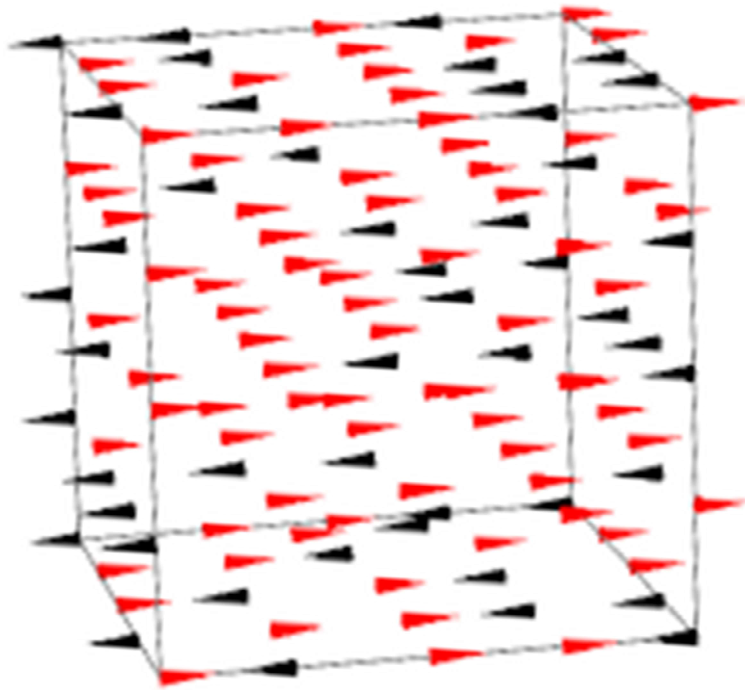
$$f(p_i) = \sum_{j \neq i} J_{ji} x_j + b_i$$

Response of current dipole

$$x_i = \begin{cases} x_i \text{ } if \text{ } sign\left(x_i \, f(p_i)\right) = 1 \\ -x_i \text{ } otherwise \end{cases}$$

- When will it stop???

Analogy: Spin Glasses



Total field at current dipole:

$$f(p_i) = \sum_{j \neq i} J_{ji} x_j + b_i$$

Response of current dipole

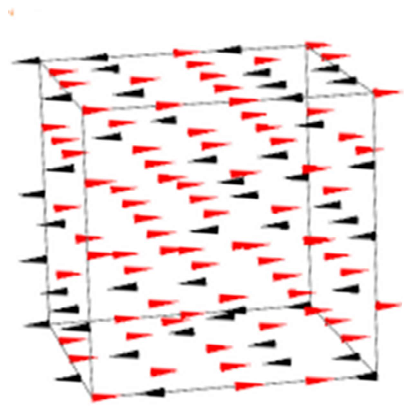
$$x_i = \begin{cases} x_i \ if \ sign\left(x_i \ f(p_i)\right) = 1 \\ -x_i \ otherwise \end{cases}$$

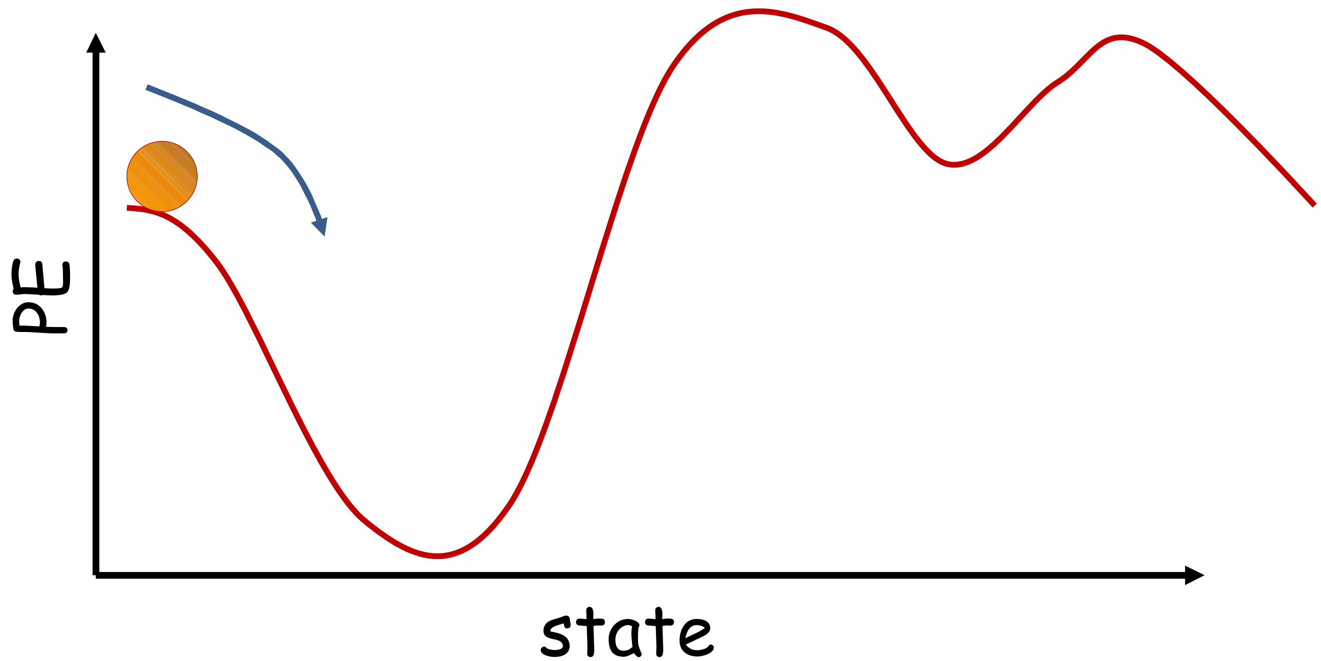
- The “Hamiltonian” (total energy) of the system

$$E = -\frac{1}{2}\sum_i x_i f(p_i) = -\sum_i \sum_{j>i} J_{ji} x_i x_j - \sum_i b_i x_i$$

- The system *evolves* to minimize the energy
  - Dipoles stop flipping if any flips result in increase of energy

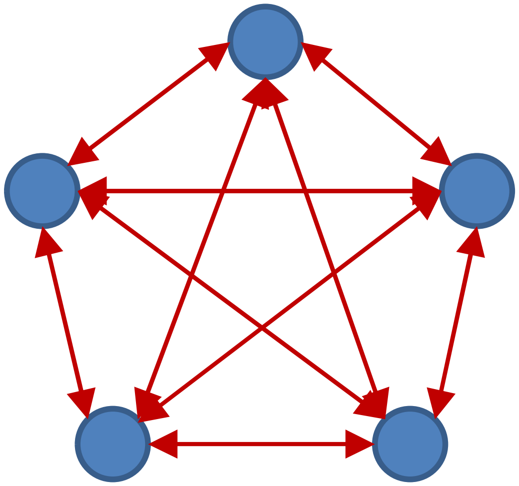
Spin Glasses





- The system stops at one of its *stable* configurations
  - Where energy is a local minimum
- Any small jitter from this stable configuration *returns it* to the stable configuration
  - I.e. the system *remembers* its stable state and returns to it

Hopfield Network



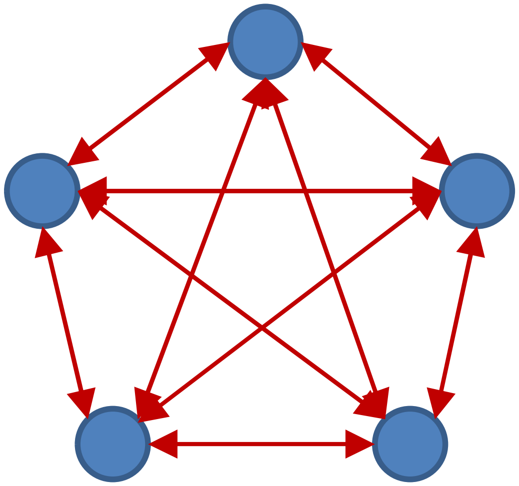
$$y_i = \Theta\left(\sum_{j \neq i} w_{ji} y_j + b_i\right)$$

$$\Theta(z) = \begin{cases} +1 \; if \; z > 0 \\ -1 \; if \; z \leq 0 \end{cases}$$

$$E = -\frac{1}{2}\left(\sum_{i,j \neq i} w_{ij} y_i y_j - \sum_i b_i y_i\right)$$

- This is analogous to the potential energy of a spin glass
  - The system will evolve until the energy hits a local minimum

Hopfield Network



$$y_i = \Theta\left(\sum_{j \neq i} w_{ji} y_j + b_i\right)$$

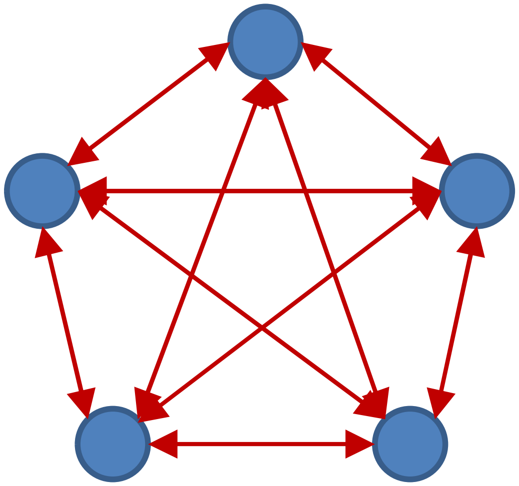
$$\Theta(z) = \begin{cases} +1 \ if \ z > 0 \\ -1 \ if \ z \leq 0 \end{cases}$$

The bias is similar to having a single extra neuron that is pegged to 1.0

We may not explicitly represent it in the slides

– The system will evolve until the energy hits a local minimum

Hopfield Network



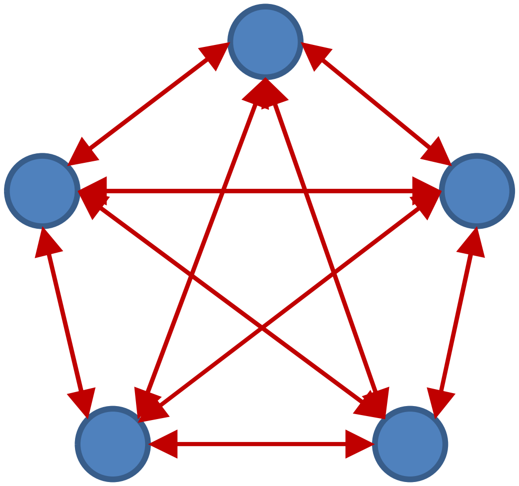
$$y_i = \Theta\left(\sum_{j \neq i} w_{ji} y_j\right)$$

$$\Theta(z) = \begin{cases} +1 \; if \; z > 0 \\ -1 \; if \; z \leq 0 \end{cases}$$

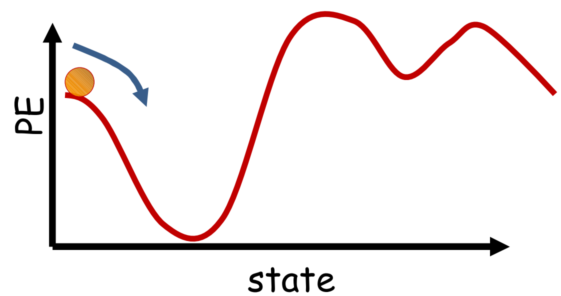
$$E = -\frac{1}{2}\sum_{i,j} w_{ij} y_i y_j$$

- This is analogous to the potential energy of a spin glass
  - The system will evolve until the energy hits a local minimum

Evolution

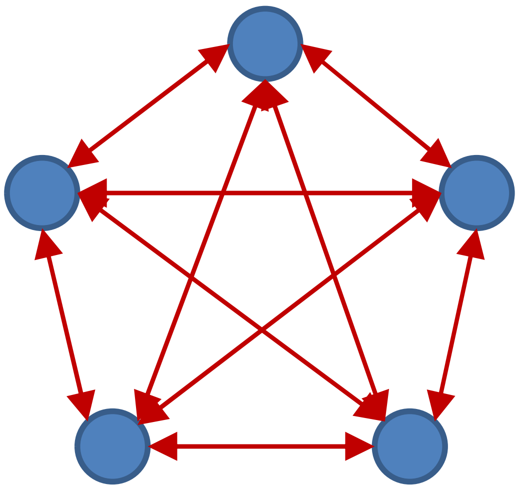


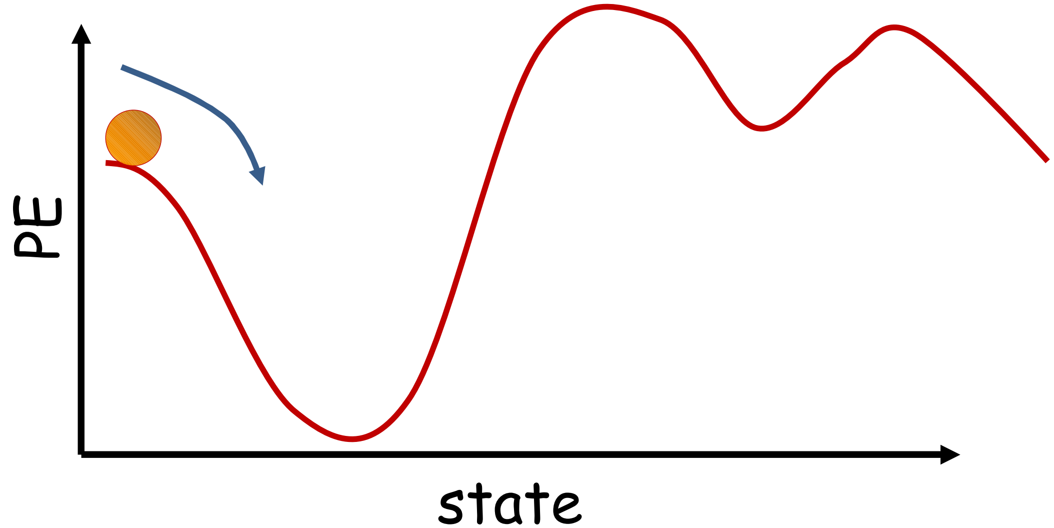
$$E = -\frac{1}{2}\sum_{i,j} w_{ij} y_i y_j$$



- The network will evolve until it arrives at a local minimum in the energy contour

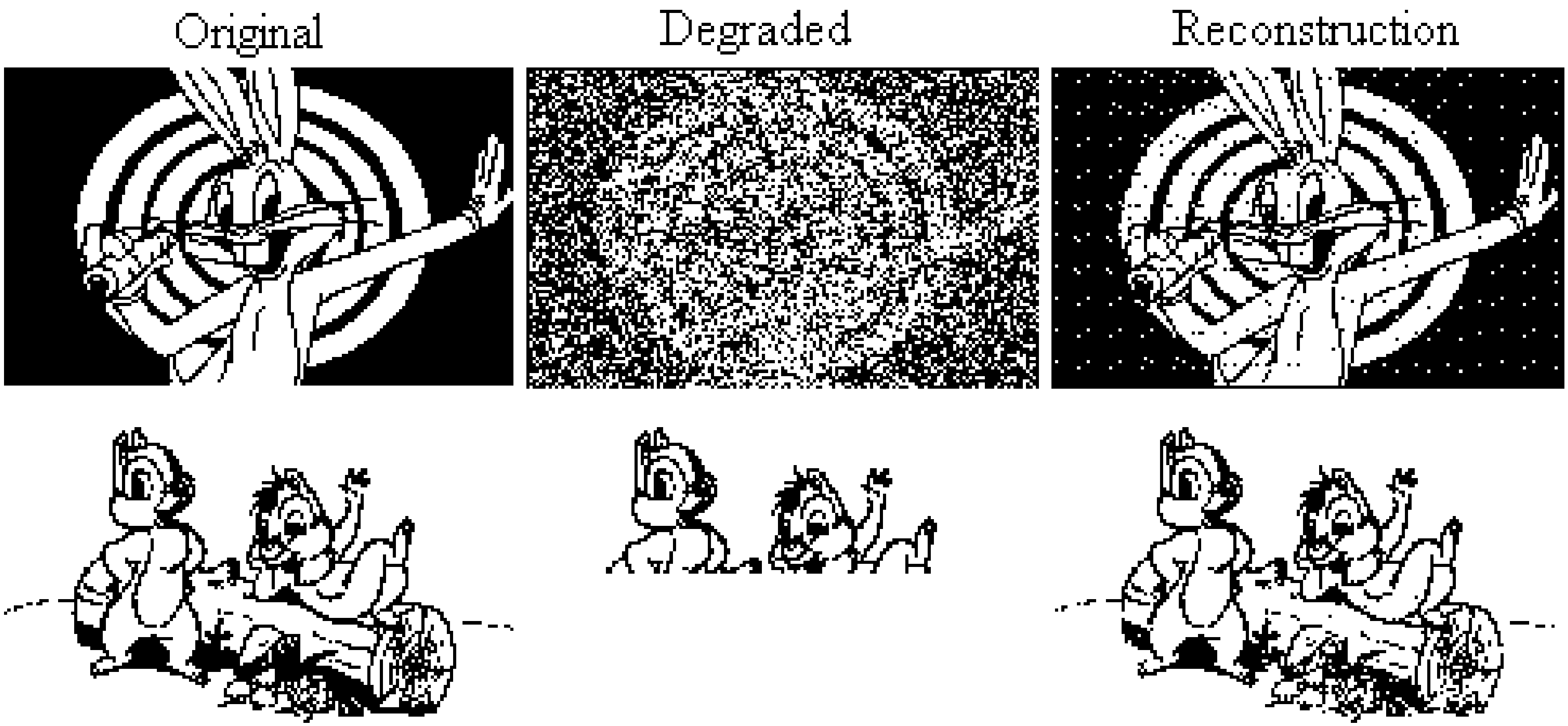
Content-addressable memory





- Each of the minima is a "stored" pattern
  - If the network is initialized close to a stored pattern, it will inevitably evolve to the pattern
- **This is a *content addressable memory***
  - Recall memory content from partial or corrupt values
- Also called ***associative memory***

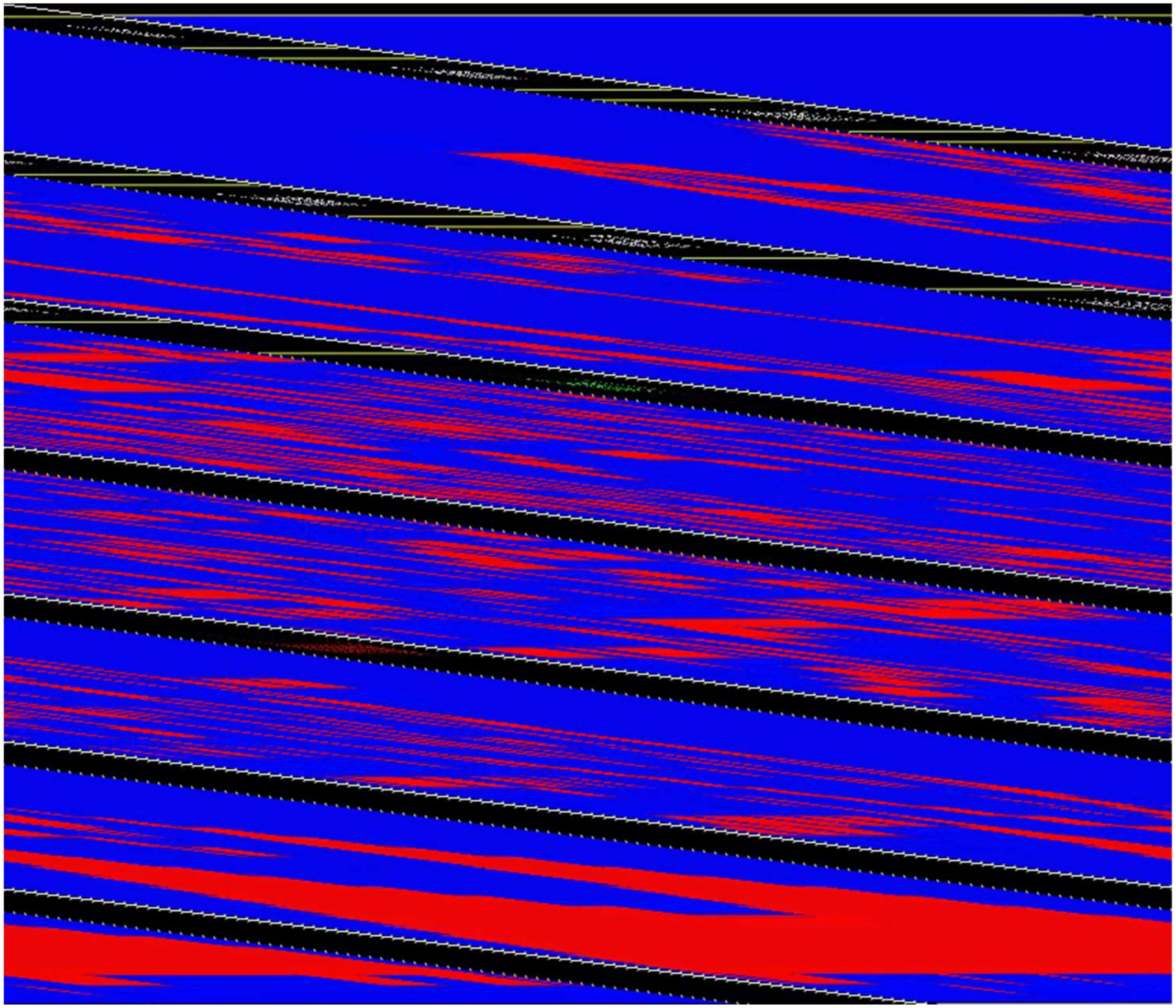
Examples: Content addressable memory



Hopfield network reconstructing degraded images from noisy (top) or partial (bottom) cues.

- http://staff.itee.uq.edu.au/janetw/cmc/chapters/Hopfield/

Hopfield net examples



Computational algorithm

1. Initialize network with initial pattern

$$y_i(0) = x_i, \qquad 0 \le i \le N-1$$

2. Iterate until convergence

$$y_i(t+1) = \Theta\left(\sum_{j \neq i} w_{ji} y_j\right), \qquad 0 \le i \le N-1$$

- Very simple
- Updates can be done sequentially, or all at once
- Convergence

$$E = -\sum_i \sum_{j>i} w_{ji} y_j y_i$$

does not change significantly any more

Computational algorithm

1. Initialize network with initial pattern

$$\mathbf{y} = \mathbf{x}, \qquad 0 \leq i \leq N-1$$

2. Iterate until convergence

$$\mathbf{y} = \Theta(\mathbf{W}\mathbf{y})$$

Writing $\mathbf{y} = [y_1, y_2, y_3, \cdots, y_N]^\top$
and arranging the weights as a matrix $\mathbf{W}$

- Very simple
- Updates can be done sequentially, or all at once
- Convergence

$$E = -0.5\mathbf{y}^\top \mathbf{W}\mathbf{y}$$

does not change significantly any more

Story so far

- A Hopfield network is a loopy binary network with symmetric connections
  - Neurons try to align themselves to the local field caused by other neurons

- Given an initial configuration, the patterns of neurons in the net will evolve until the “energy” of the network achieves a local minimum
  - The evolution will be monotonic in total energy
  - The dynamics of a Hopfield network mimic those of a spin glass
  - The network is symmetric: if a pattern Y is a local minimum, so is $-Y$

- The network acts as a *content-addressable* memory
  - If you initialize the network with a somewhat damaged version of a local-minimum pattern, it will evolve into that pattern
  - Effectively “recalling” the correct pattern, from a damaged/incomplete version

Poll 2

Poll 2

Mark all that are correct about Hopfield nets

- **The network activations evolve until the energy of the net arrives at a local minimum**
- **Hopfield networks are a form of content addressable memory**
- It is possible to analytically determine the stored memories by inspecting the weights matrix

Issues

- How do we make the network store *a specific* pattern or set of patterns?

- How many patterns can we store?

- How to “retrieve” patterns better..

Issues

- How do we make the network store *a specific* pattern or set of patterns?

- How many patterns can we store?

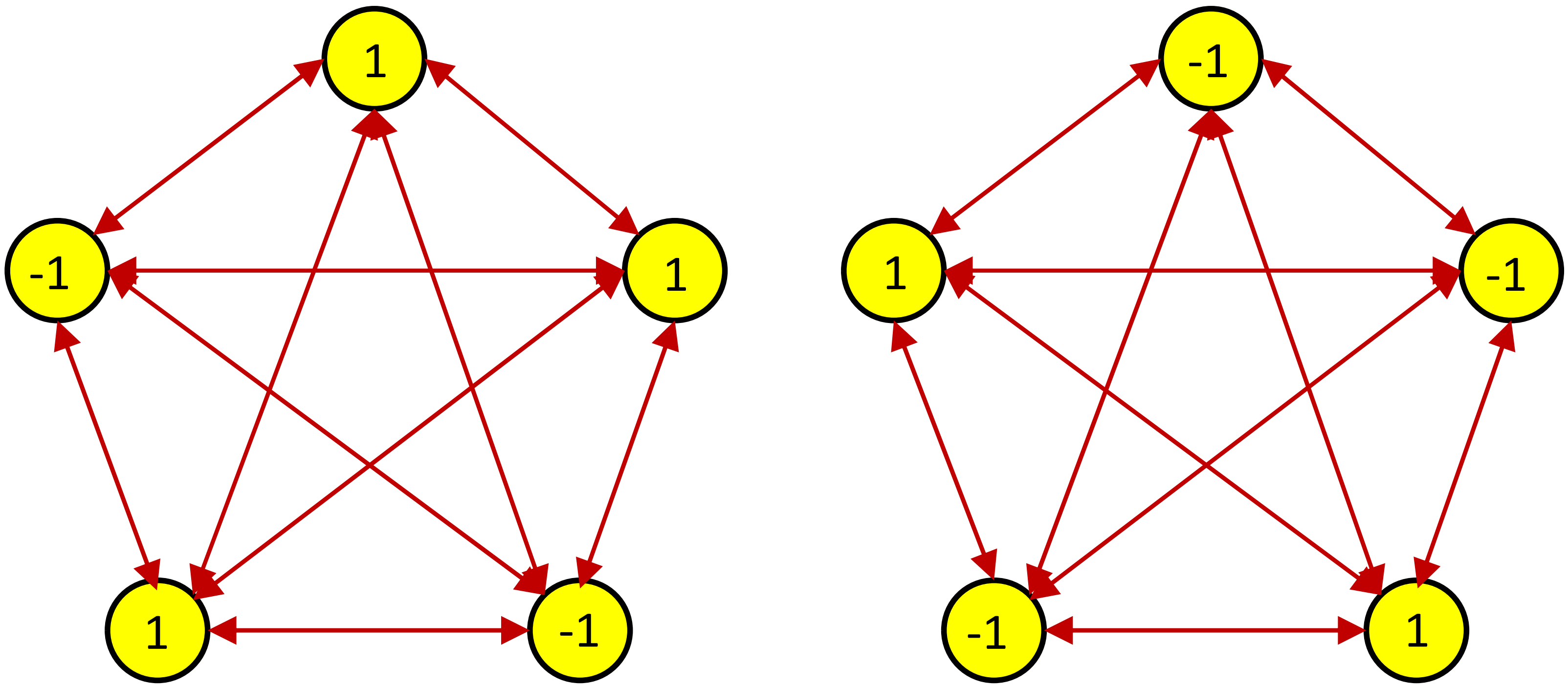
- How to "retrieve" patterns better..

How do we remember a *specific* pattern?

- How do we teach a network to "remember" this image

- For an image with N pixels we need a network with N neurons
- Every neuron connects to every other neuron
- Weights are symmetric (not mandatory)
- $\frac{N(N-1)}{2}$ weights in all

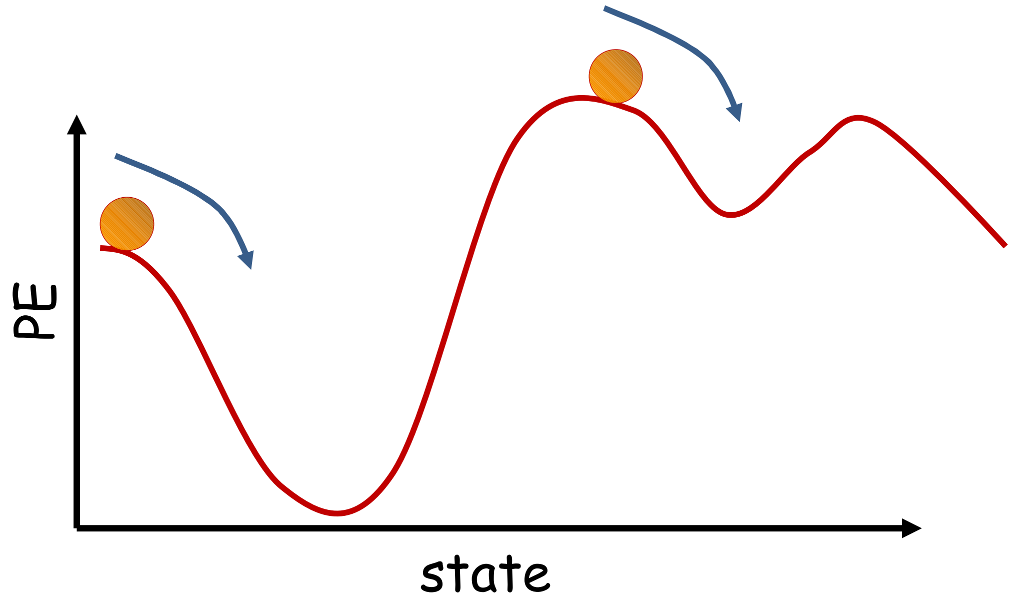
Storing patterns: Training a network

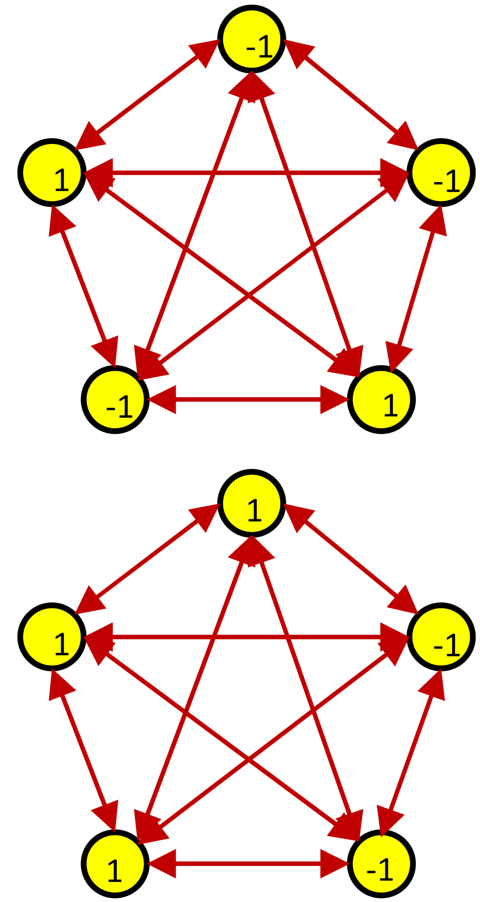


- A network that stores pattern P also naturally stores $-P$
  - Symmetry $E(P) = E(-P)$ since E is a function of $y_i y_j$

$$E = -\sum_i \sum_{j<i} w_{ji} y_j y_i$$

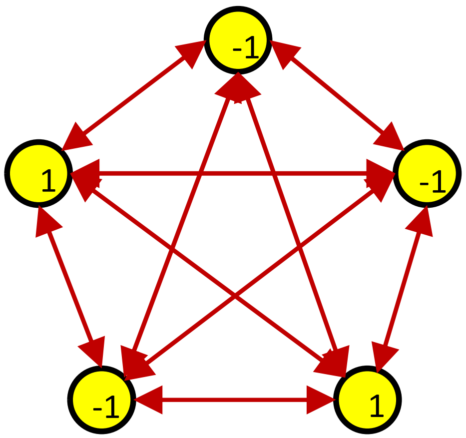
A network can store *multiple* patterns

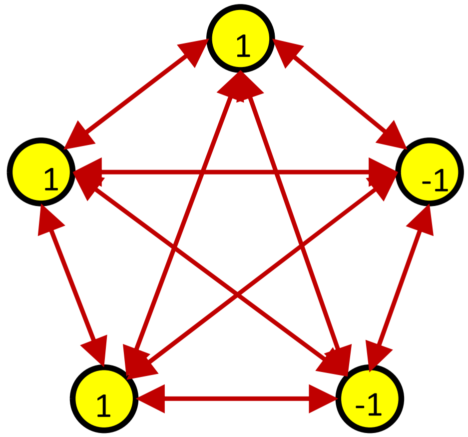




- Every stable point is a stored pattern
- So we could design the net to store multiple patterns
  – Remember that every stored pattern P is actually *two* stored patterns, P and $-P$

Storing a pattern

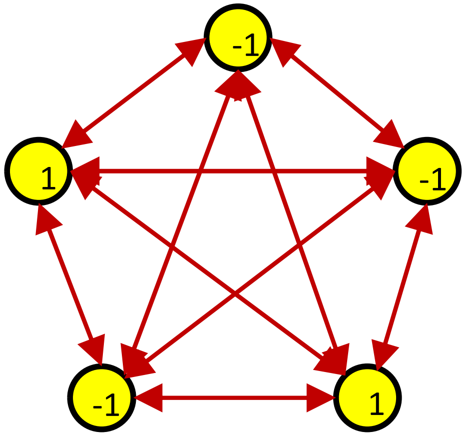




$$E = -\sum_{i}\sum_{j<i} w_{ji} y_j y_i$$

- Design $\{w_{ij}\}$ such that the energy is a local minimum at the desired $P = \{y_i\}$

Storing specific patterns

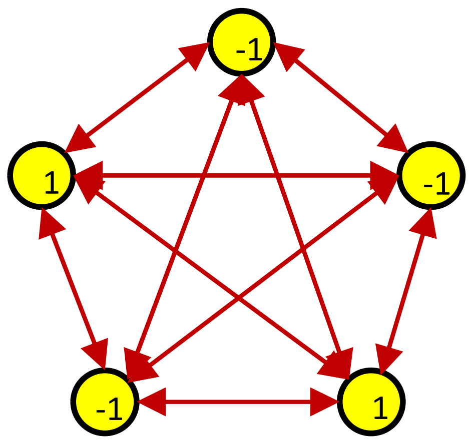


- Storing 1 pattern: We want

$$sign\left(\sum_{j \neq i} w_{ji} y_j\right) = y_i \quad \forall\, i$$

- This is a stationary pattern

Storing specific patterns



HEBBIAN LEARNING:

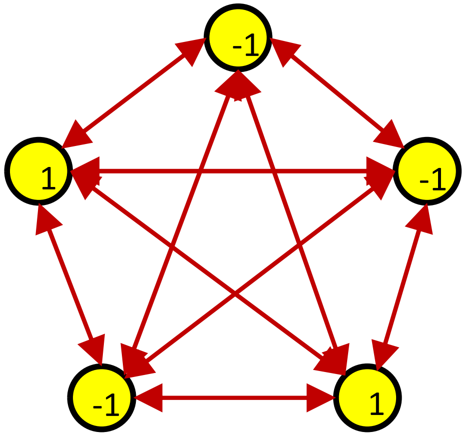
$$w_{ji} = y_j y_i$$

- Storing 1 pattern: We want

$$sign\left(\sum_{j \neq i} w_{ji} y_j\right) = y_i \quad \forall\, i$$

- This is a stationary pattern

Storing specific patterns



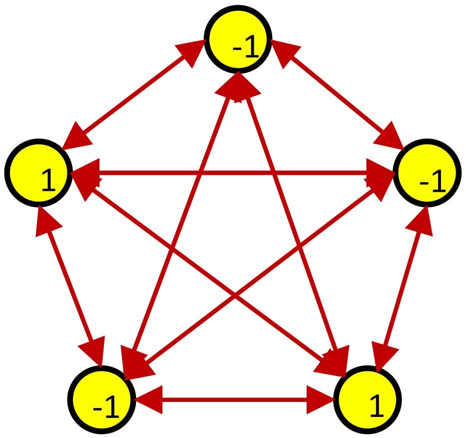
HEBBIAN LEARNING:

$$w_{ji} = y_j y_i$$

- $sign\left(\sum_{j\neq i} w_{ji} y_j\right) = sign\left(\sum_{j\neq i} y_j y_i y_j\right)$

$$= sign\left(\sum_{j\neq i} y_j^2 y_i\right) = sign(y_i) = y_i$$

Storing specific patterns



HEBBIAN LEARNING:

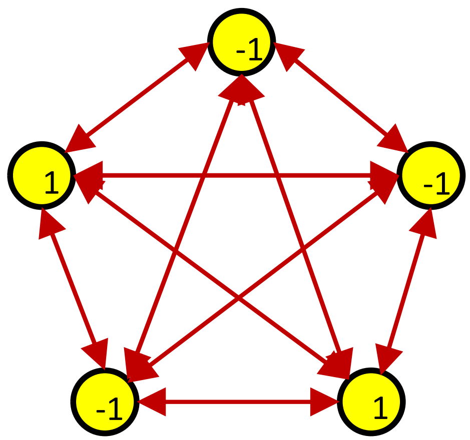
$$w_{ji} = y_j y_i$$

The pattern is stationary

- $sign\left(\sum_{j \neq i} w_{ji} y_j\right) = sign\left(\sum_{j \neq i} y_j y_i y_j\right)$

$$= sign\left(\sum_{j \neq i} y_j^2 y_i\right) = sign(y_i) = y_i$$

Storing specific patterns



HEBBIAN LEARNING:

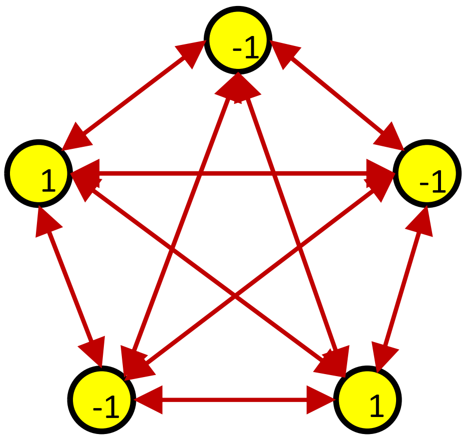
$$w_{ji} = y_j y_i$$

$$E = -\sum_i \sum_{j<i} w_{ji} y_j y_i = -\sum_i \sum_{j<i} y_i^2 y_j^2$$

$$= -\sum_i \sum_{j<i} 1 = -0.5N(N-1)$$

- This is the lowest possible energy value for the network

Storing specific patterns



HEBBIAN LEARNING:

$$w_{ji} = y_j y_i$$

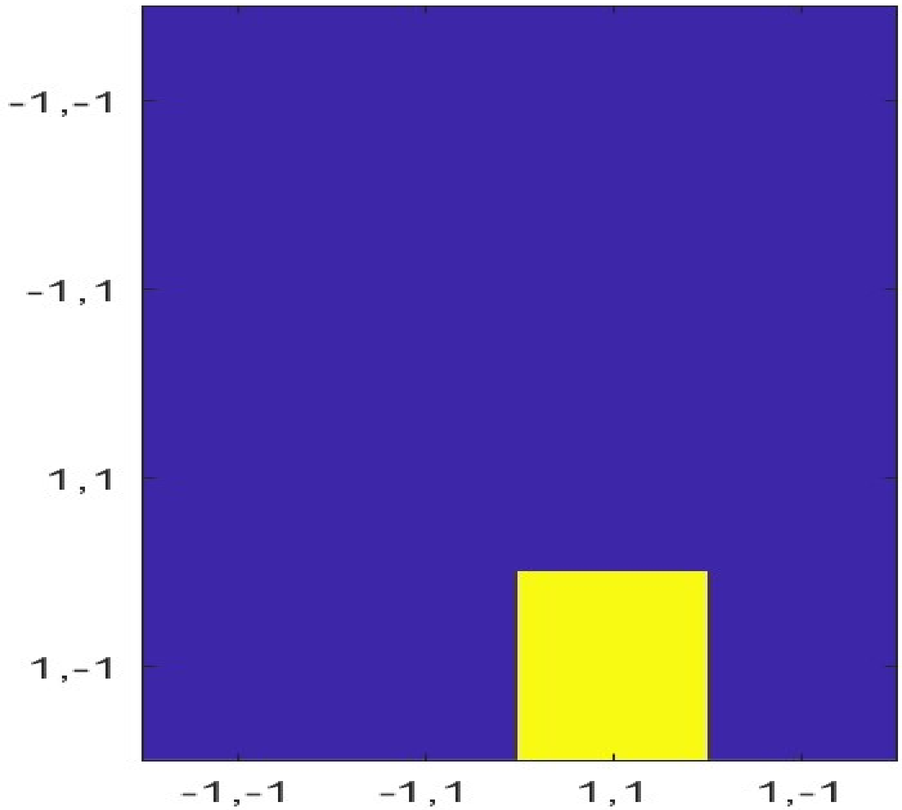
The pattern is *STABLE*

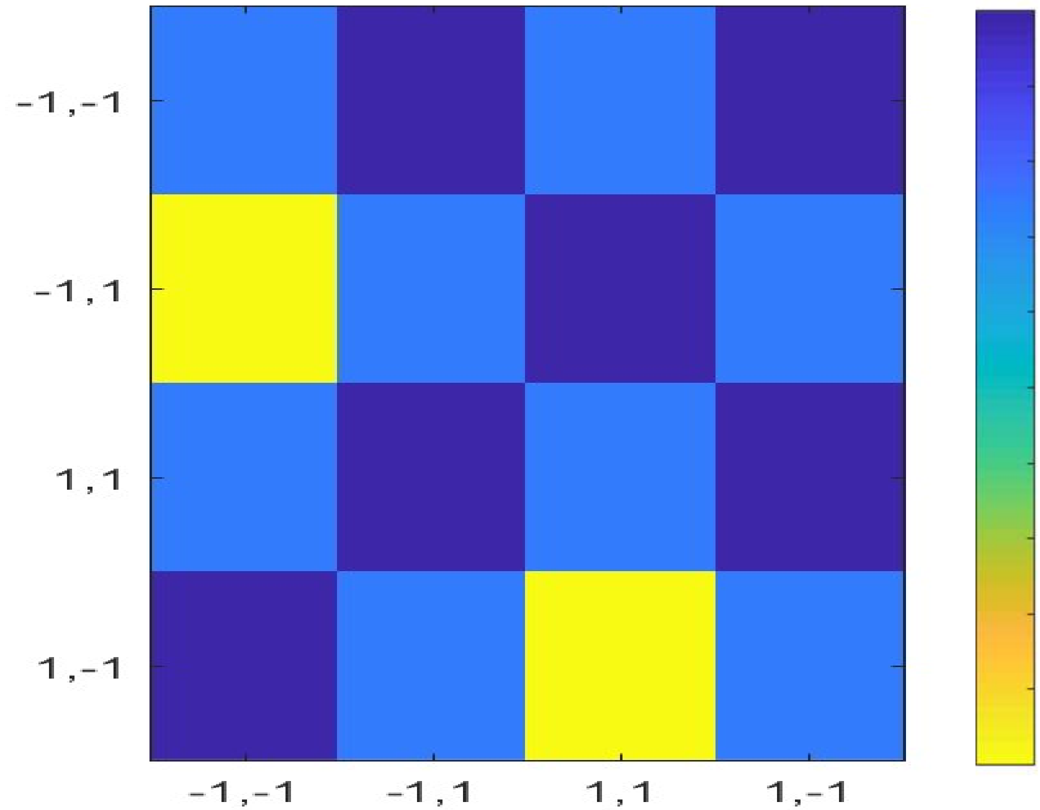
$$E = -\sum_i \sum_{j<i} w_{ji} y_j y_i = -\sum_i \sum_{j<i} y_i^2 y_j^2$$

$$= -\sum_i \sum_{j<i} 1 = -0.5N(N-1)$$

- This is the lowest possible energy value for the network

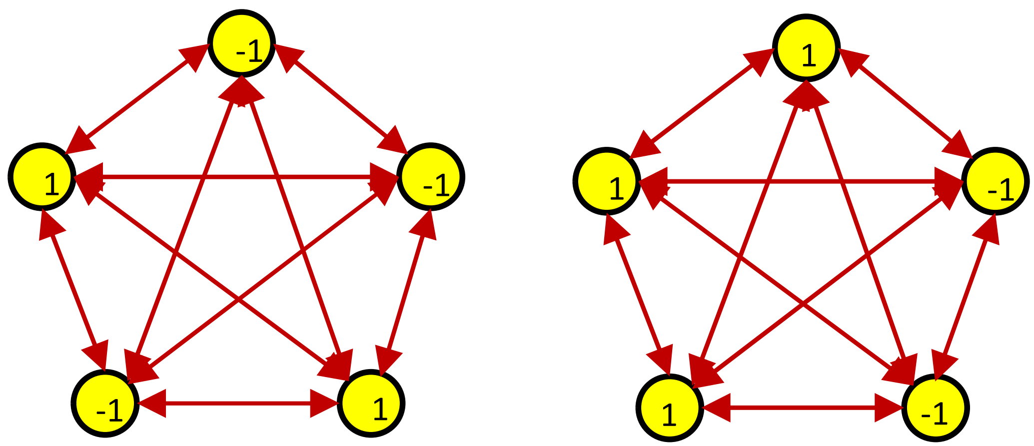
Hebbian learning: Storing a 4-bit pattern





- Left: Pattern stored.   Right: Energy map
- Stored pattern has lowest energy
- Gradation of energy ensures stored pattern (or its ghost) is recalled from everywhere

Storing multiple patterns



- To store *more* than one pattern

$$w_{ji} = \sum_{\mathbf{y}_p \in \{\mathbf{y}_p\}} y_i^p y_j^p$$

- $\{\mathbf{y}_p\}$ is the set of patterns to store
- Super/subscript p represents the specific pattern

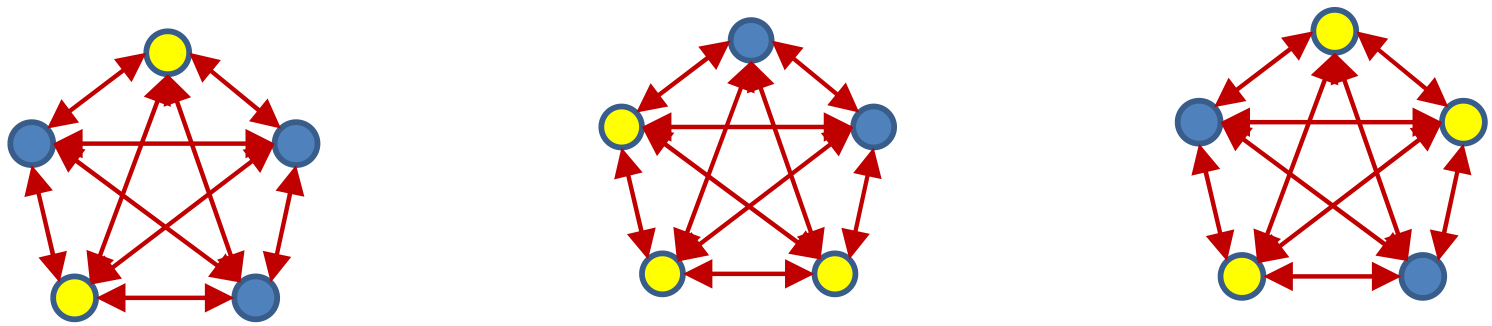
Issues

- How do we make the network store *a specific* pattern or set of patterns?

- How many patterns can we store?

- How to “retrieve” patterns better..

How many patterns can we store?



- **Hopfield**: For a network of N neurons can store up to ~$0.15N$ random patterns through Hebbian learning
  - Provided they are “far” enough
- Where did this number come from?

The limits of Hebbian Learning

- Consider the following: We must store K N-bit patterns of the form
$$\mathbf{y}_k = [y_1^k, y_2^k, \dots, y_N^k], k = 1 \dots K$$

- Hebbian learning (scaling by $\frac{1}{N}$ for normalization, this does not affect actual pattern storage):
$$w_{ij} = \frac{1}{N} \sum_k y_i^k y_j^k$$

- **For any pattern $\mathbf{y}_p$ to be stable:**

$$y_i^p \sum_j w_{ij} y_j^p > 0 \quad \forall i$$

$$y_i^p \frac{1}{N} \sum_j \sum_k y_i^k y_j^k \, y_j^p > 0 \quad \forall i$$

The limits of Hebbian Learning

- **For any pattern $\mathbf{y}_p$ to be stable:**

$$y_i^p \frac{1}{N} \sum_j \sum_k y_i^k y_j^k \, y_j^p > 0 \quad \forall i$$

$$y_i^p \frac{1}{N} \sum_j y_i^p y_j^p y_j^p + y_i^p \frac{1}{N} \sum_j \sum_{k \neq p} y_i^k y_j^k \, y_j^p > 0 \quad \forall i$$

- Note that the first term equals 1 (because $y_j^p y_j^p = y_i^p y_i^p = 1$)
  - i.e. for $\mathbf{y}_p$ to be stable the requirement is that the second *crosstalk term*:

$$y_i^p \frac{1}{N} \sum_j \sum_{k \neq p} y_i^k y_j^k \, y_j^p > -1 \quad \forall i$$

- The pattern will *fail* to be stored if the *crosstalk*

$$y_i^p \frac{1}{N} \sum_j \sum_{k \neq p} y_i^k y_j^k \, y_j^p < -1 \quad for\ any\ i$$

The limits of Hebbian Learning

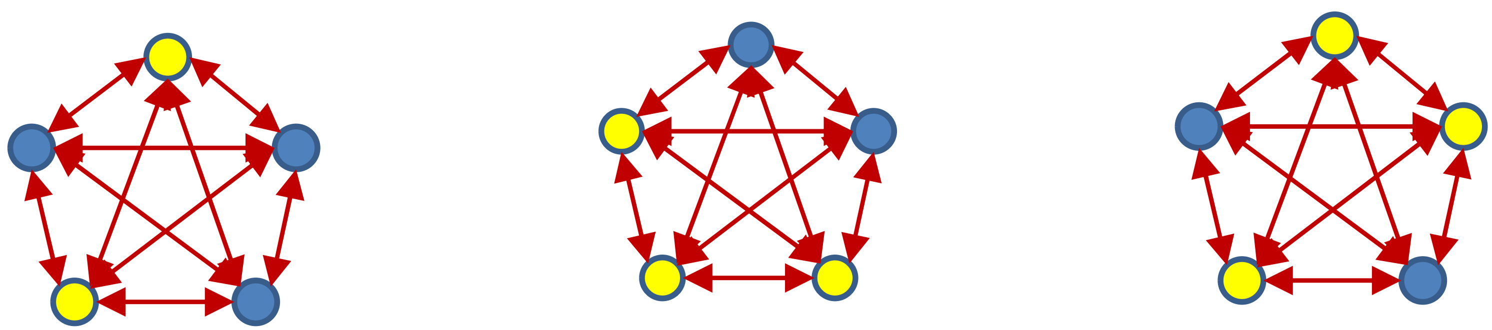
- For any random set of K patterns to be stored, the probability of the following must be low

$$\left(C_i^p = \frac{1}{N}\sum_j \sum_{k\neq p} y_i^p y_i^k y_j^k\, y_j^p\right) < -1$$

- For large N and K the probability distribution of C_i^p approaches a Gaussian with 0 mean, and variance K/N
  - Considering that individual bits $y_i^l \in \{-1, +1\}$ and have variance 1
- For a Gaussian, $C \sim N(0, K/N)$
  - $P(C < -1 \mid \mu = 0, \sigma^2 = K/N) < 0.004$ for $K/N < 0.14$

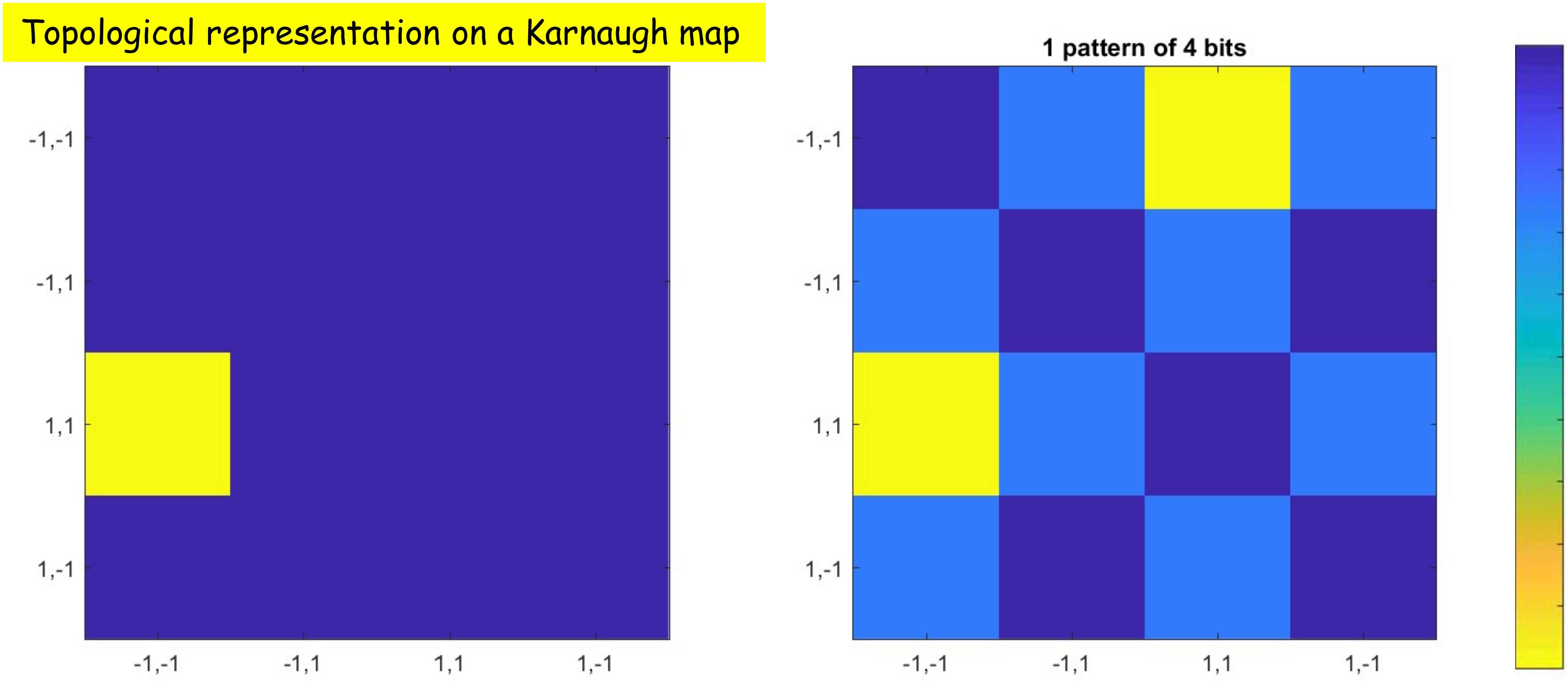
- I.e. To have less than 0.4% probability that stored patterns will *not* be stable, $K < 0.14N$

How many patterns can we store?



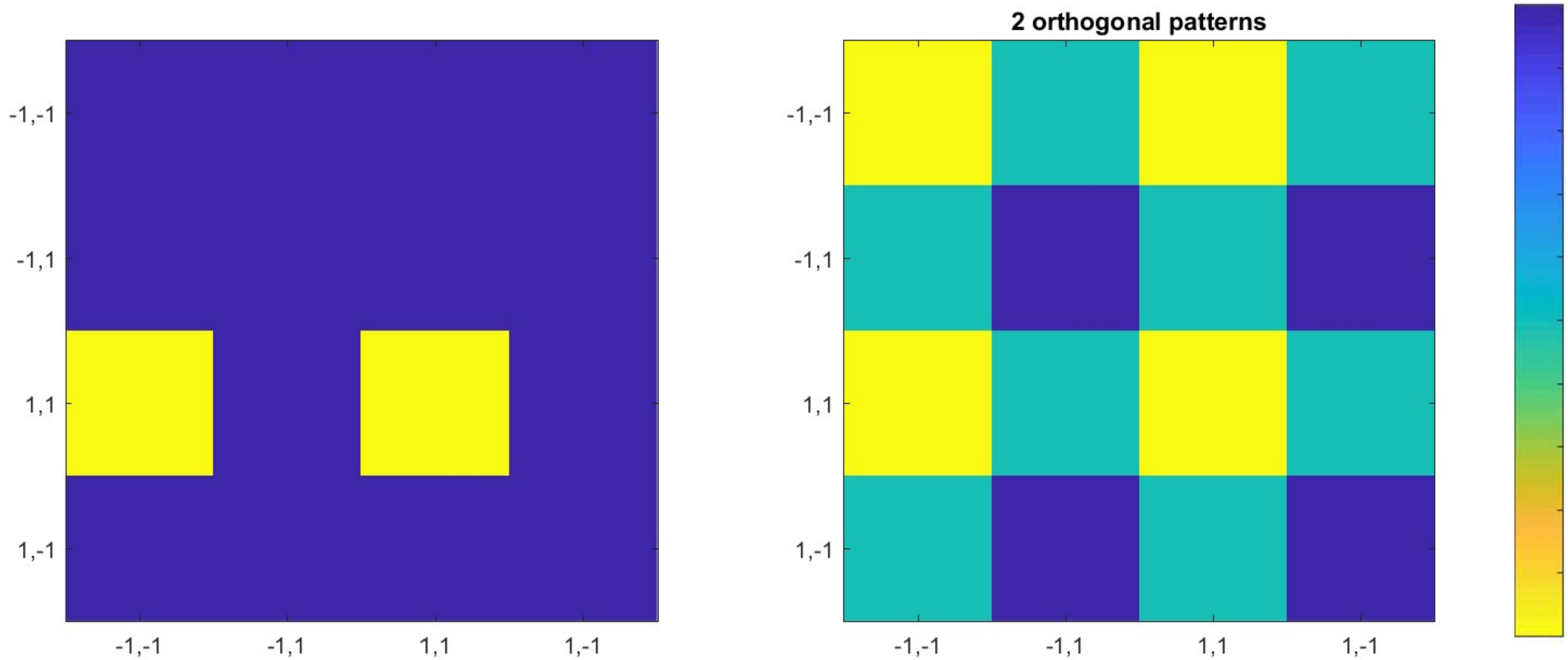
- A network of N neurons trained by Hebbian learning can store up to ~$0.14N$ random patterns with low probability of error
  - Computed assuming $prob(bit = 1) = 0.5$
    - On average no. of matched bits in any pair = no. of mismatched bits
      - Patterns are "orthogonal" – maximally distant – from one another
  - Expected behavior for *non-orthogonal* patterns?
- To get some insight into what is stored, lets see some examples

Hebbian learning: One 4-bit pattern



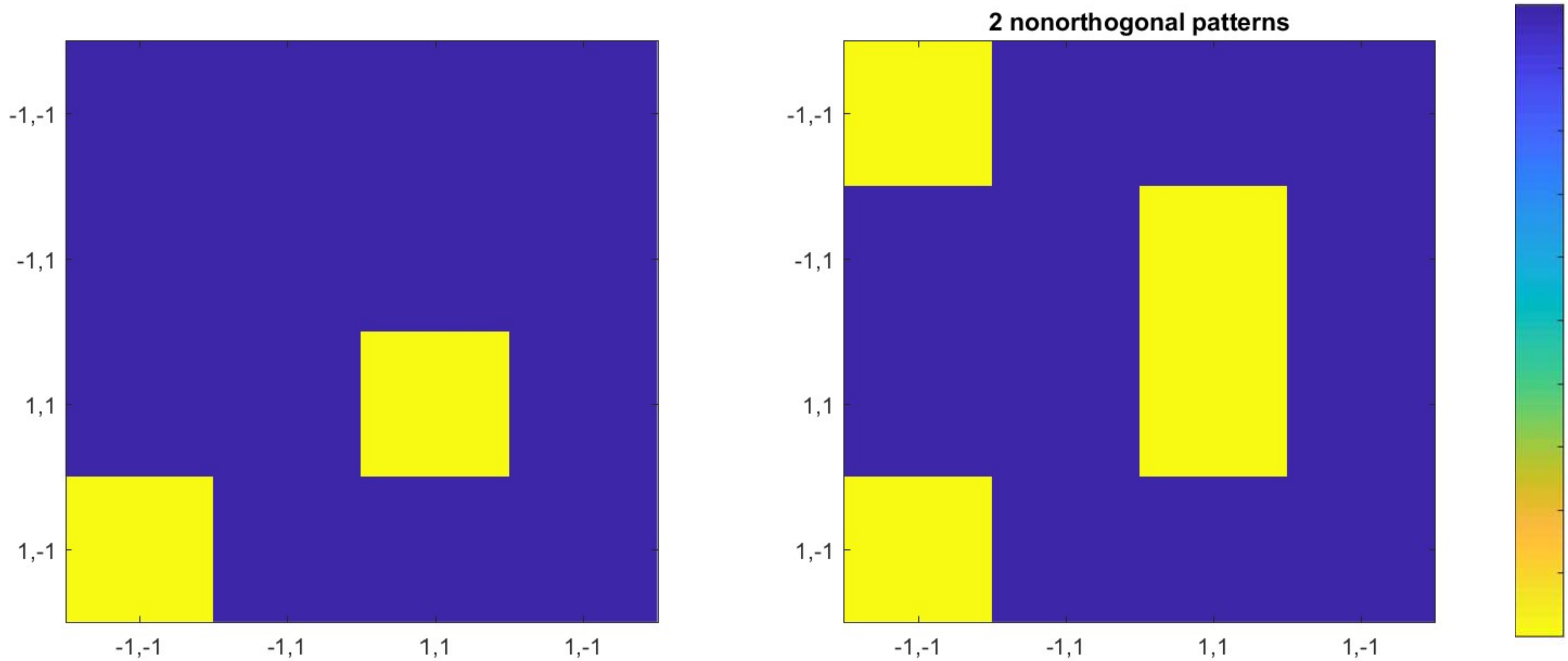
- Left: Pattern stored. Right: Energy map
- Note: Pattern is an energy well, but there are other local minima
  - Where?
  - Also note "shadow" pattern

Two orthogonal 4-bit patterns



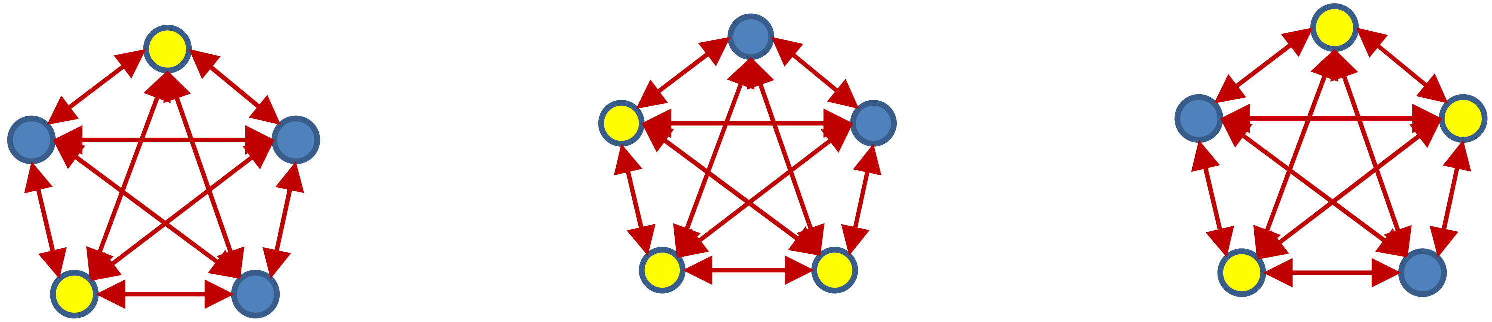
- Patterns are local minima (stationary and stable)
  - No other local minima exist
  - But patterns perfectly confusable for recall

Two *non*-orthogonal 4-bit patterns



- Patterns are local minima (stationary and stable)
  - No other local minima exist
  - Actual *wells* for patterns
    - Patterns may be perfectly recalled!
  - Note K > 0.14 N

How many patterns can we store?

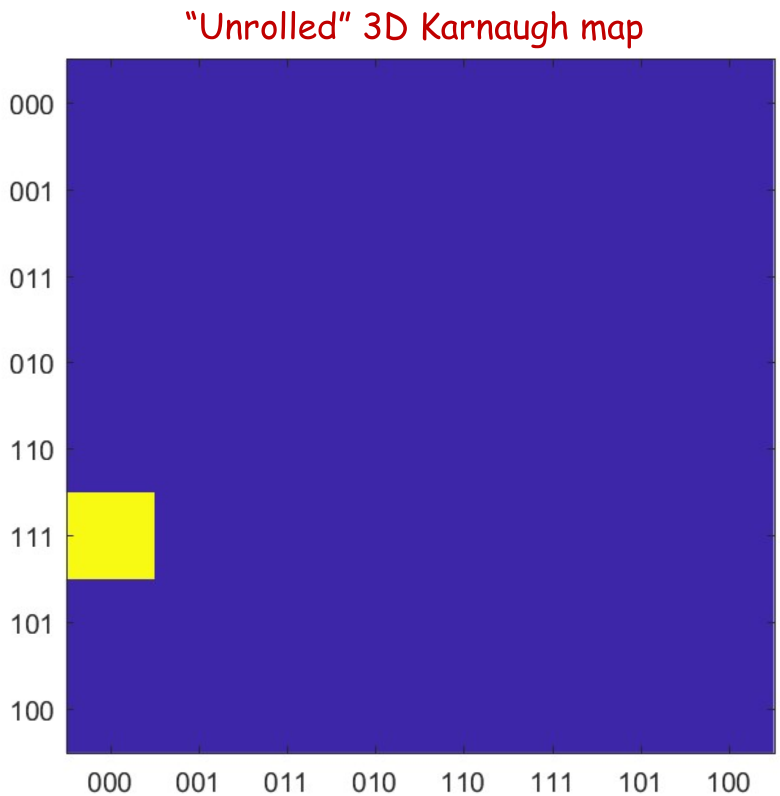


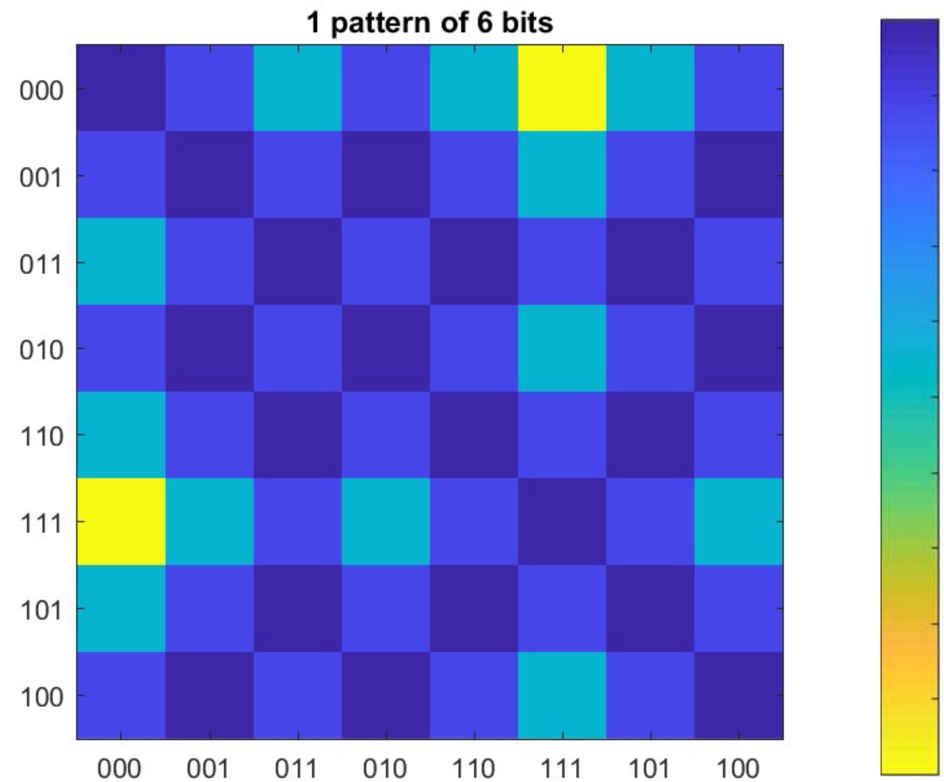
- Hopfield: For a network of N neurons can store up to $0.14N$ random patterns

- Apparently a fuzzy statement
  - What does it really mean to say “stores” 0.14N random patterns?
    - Stationary? Stable? No other local minima?
  - What if the patterns to store are not random?

- N=4 may not be a good case (N too small)

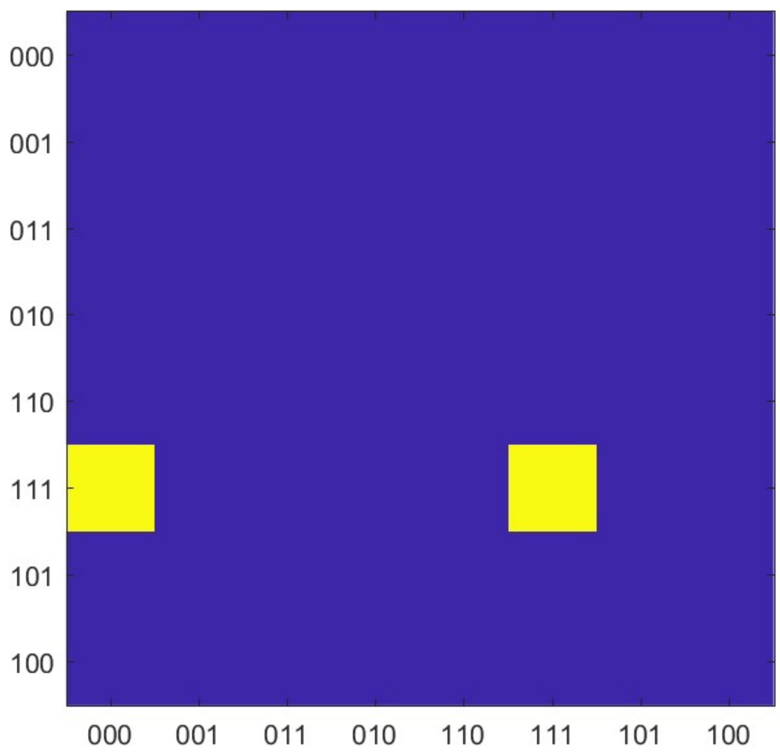
A 6-bit pattern

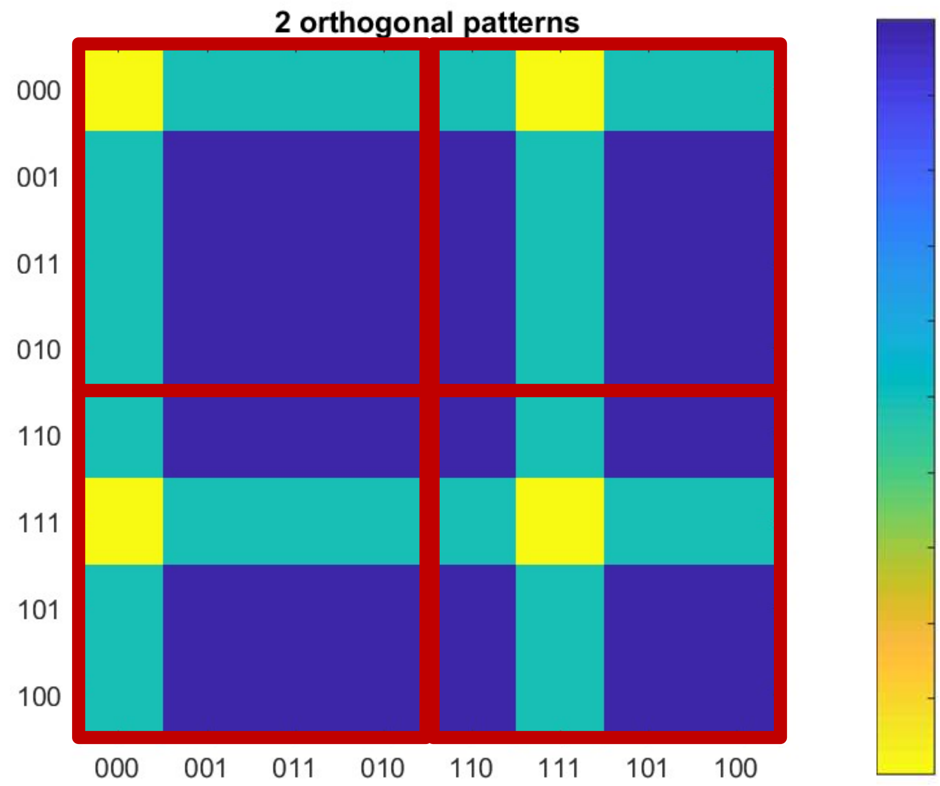




- Perfectly stationary and stable
- But many spurious local minima..
  - Which are "fake" memories

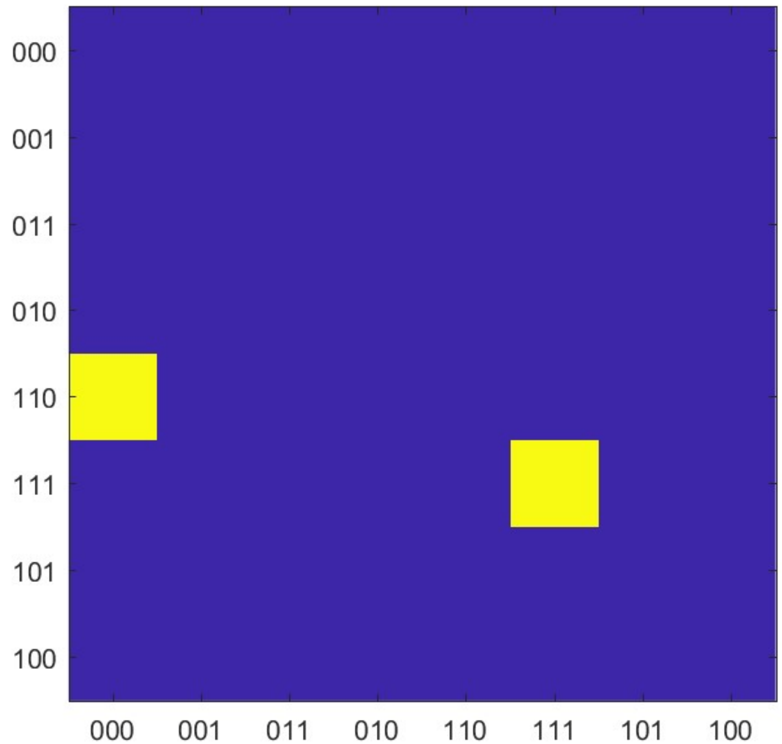
Two orthogonal 6-bit patterns

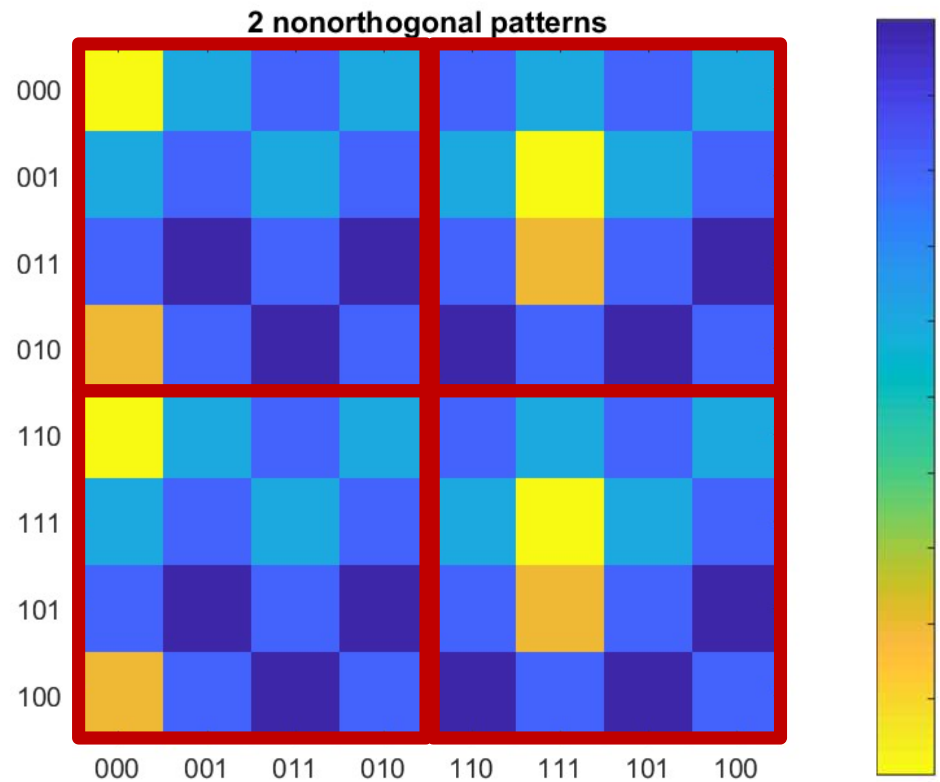




- Perfectly stationary and stable
- Several spurious “fake-memory” local minima..
  – Figure overstates the problem: actually a 3-D Kmap

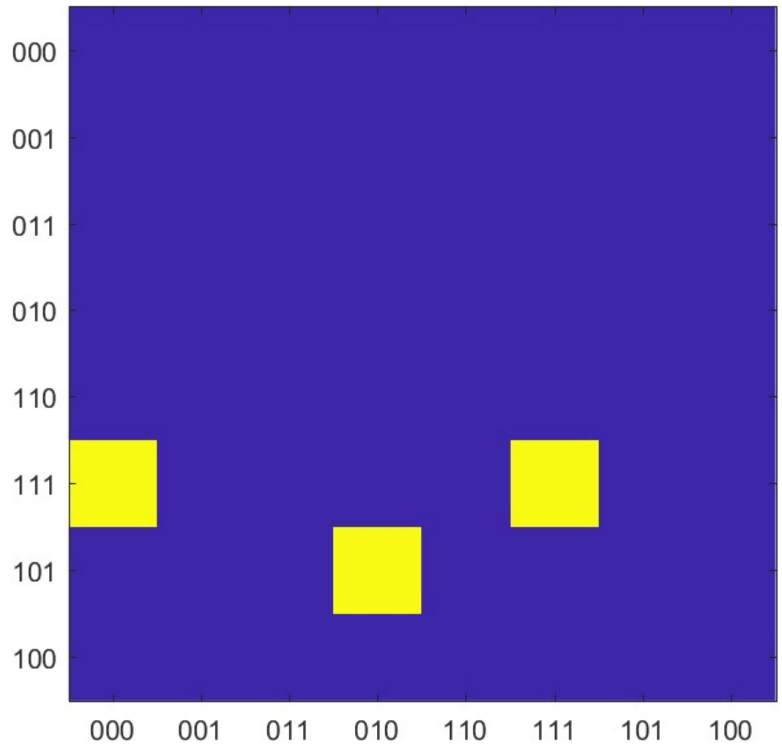
Two non-orthogonal 6-bit patterns

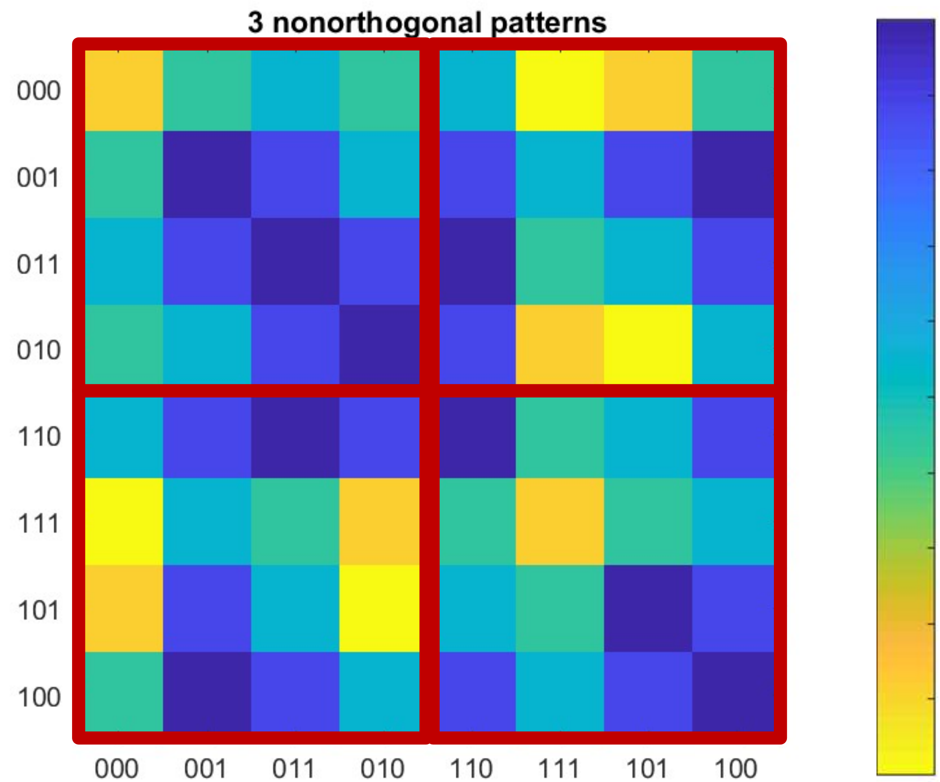




- Perfectly stationary and stable
- Some spurious “fake-memory” local minima..
  - But every stored pattern has “bowl”
  - *Fewer* spurious minima than for the orthogonal case

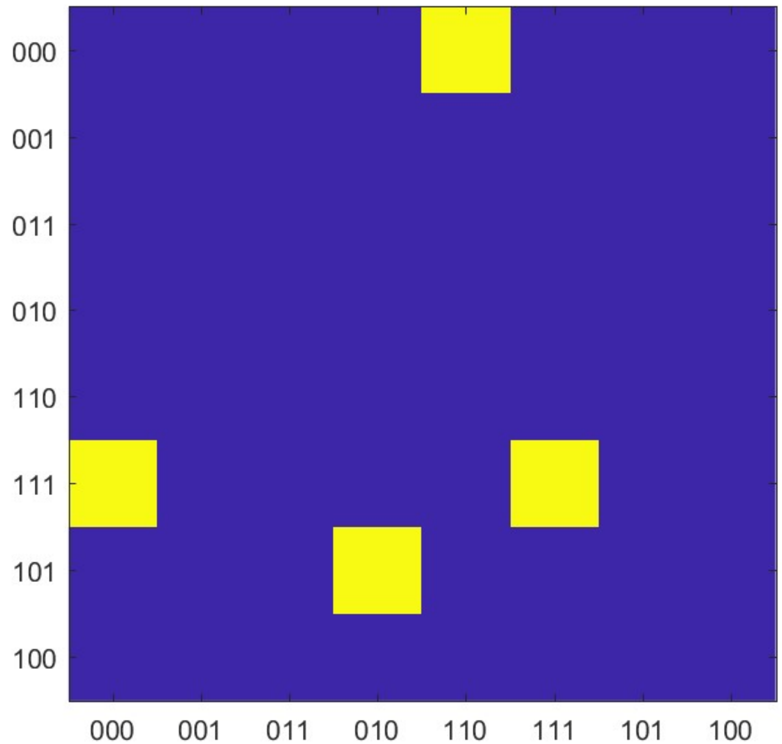
Three *non*-orthogonal 6-bit patterns

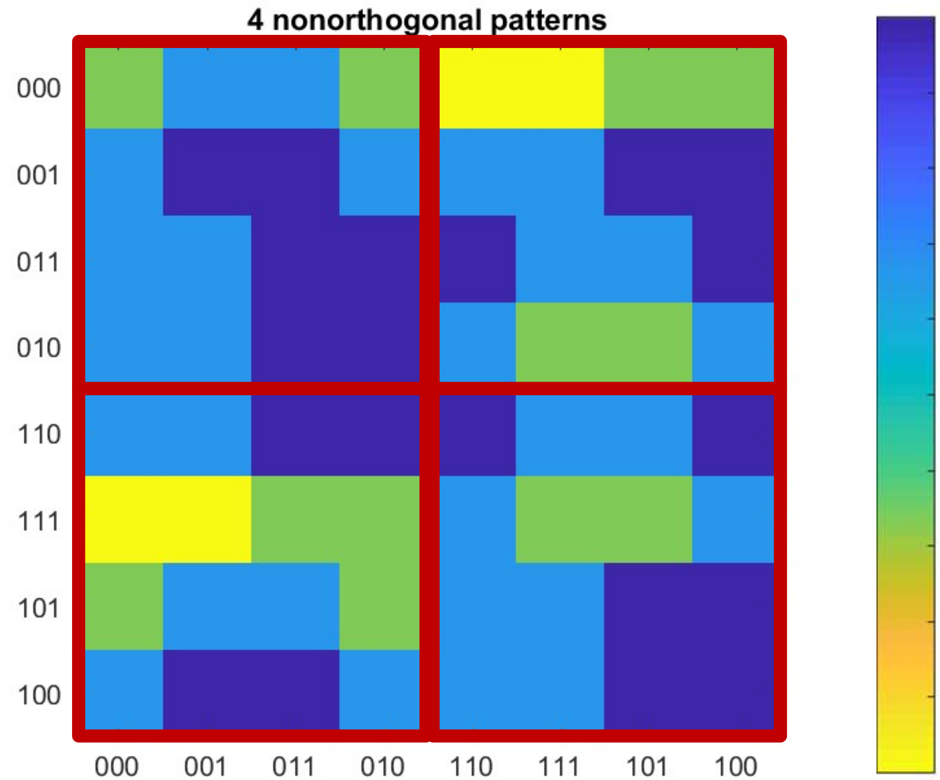




- Note: Cannot have 3 or more orthogonal 6-bit patterns..
- Patterns are perfectly stationary and stable (K > 0.14N)
- Some spurious "fake-memory" local minima..
  - But every stored pattern has "bowl"
  - *Fewer* spurious minima than for the orthogonal 2-pattern case

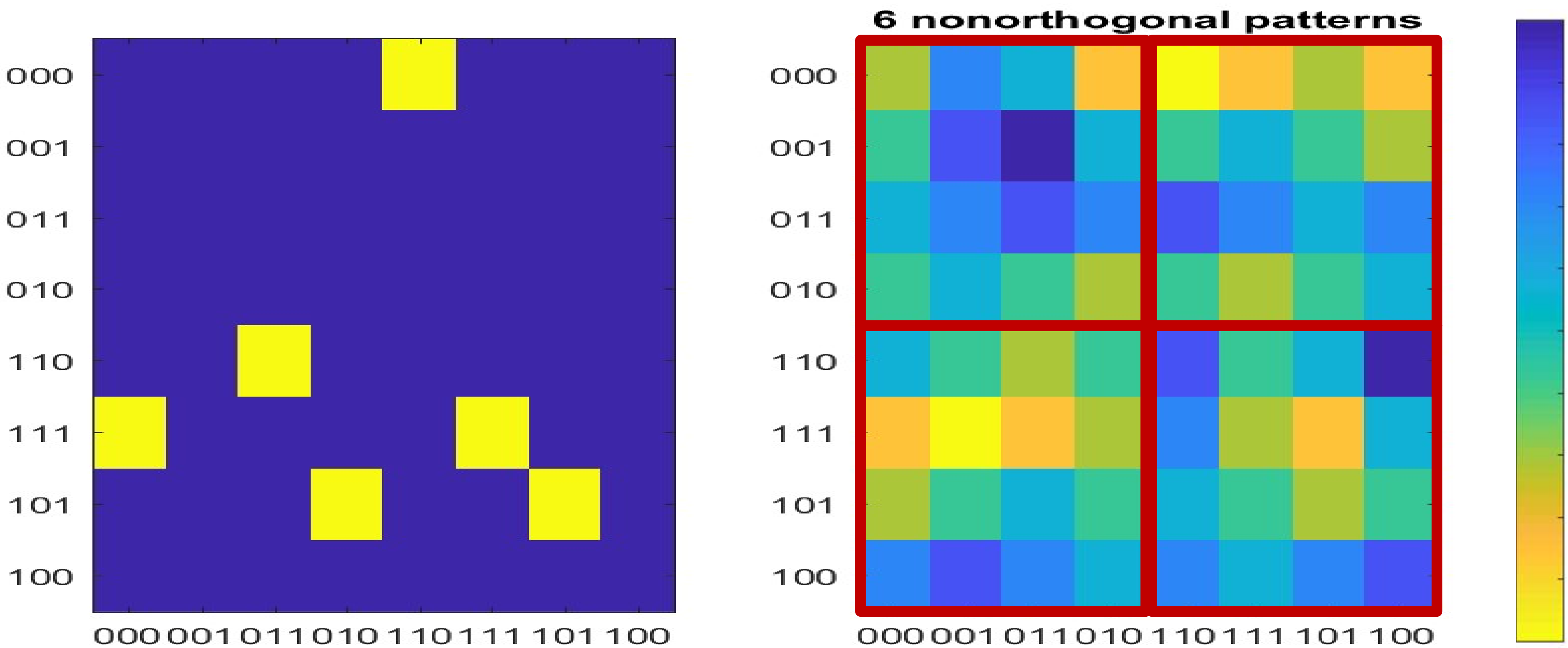
Four *non*-orthogonal 6-bit patterns





- Patterns are perfectly stationary for K > 0.14N
- *Fewer* spurious minima than for the orthogonal 2-pattern case
  – Most fake-looking memories are in fact ghosts..

Six *non*-orthogonal 6-bit patterns



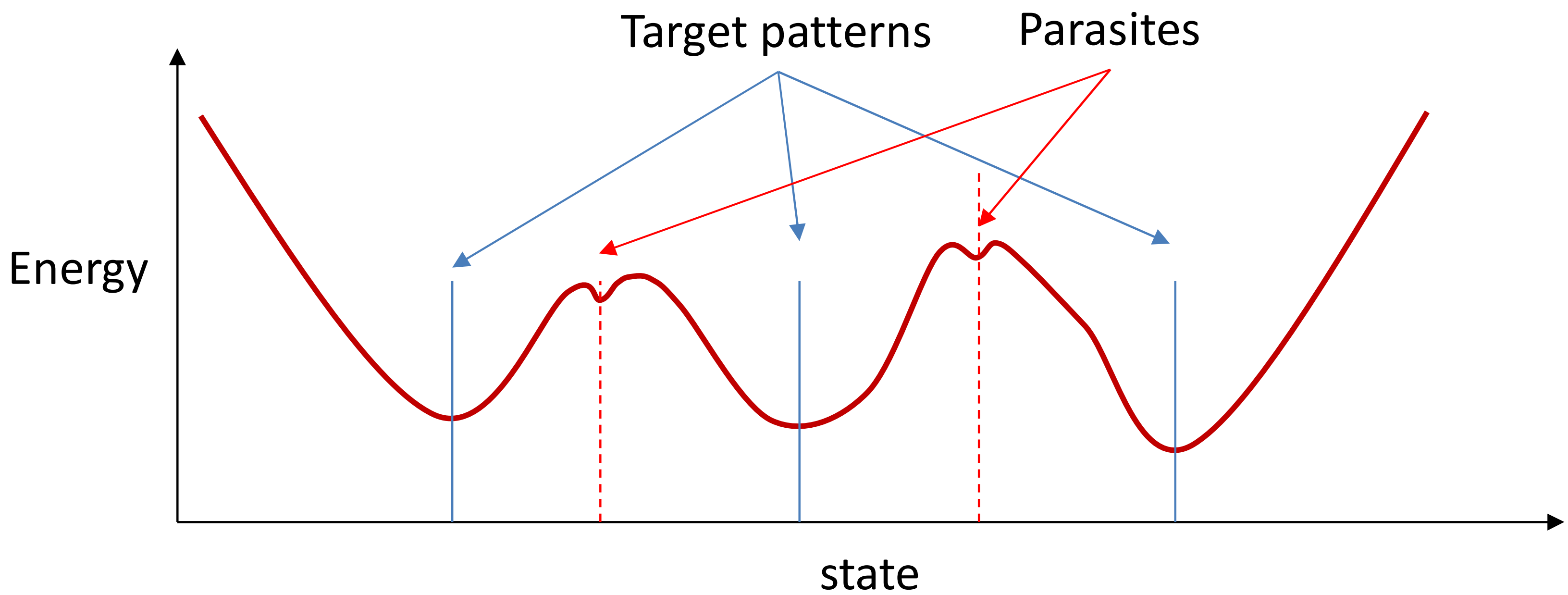
- Breakdown largely due to interference from “ghosts”
- But multiple patterns are stationary, and often stable
  - For K >> 0.14N

Observations

- Many “parasitic” patterns
  - Undesired patterns that also become stable or attractors

- Apparently, a capacity to store *more* than 0.14N patterns

Parasitic Patterns



- Parasitic patterns can occur because sums of odd numbers of stored patterns are also stable for Hebbian learning:
  - $\mathbf{y}_{parasite} = sign(\mathbf{y}_a + \mathbf{y}_b + \mathbf{y}_c)$
- They are also from other random local energy minima from the weights matrices themselves

Capacity

- Seems possible to store K > 0.14N patterns
  - i.e. obtain a weight matrix W such that K > 0.14N patterns are stationary
  - Possible to make more than 0.14N patterns at-least 1-bit stable

- Patterns that are *non-orthogonal* easier to remember
  - I.e. patterns that are *closer* are easier to remember than patterns that are farther!!

- Can we attempt to get greater control on the process than Hebbian learning gives us?
  - Can we do *better* than Hebbian learning?
    - Better capacity and fewer spurious memories?

Story so far

- A Hopfield network is a loopy binary net with symmetric connections
  - Neurons try to align themselves to the local field caused by other neurons
- Given an initial configuration, the patterns of neurons in the net will evolve until the "energy" of the network achieves a local minimum
  - The network acts as a *content-addressable* memory
    - Given a damaged memory, it can evolve to recall the memory fully
- The network must be designed to store the desired memories
  - Memory patterns must be *stationary* and *stable* on the energy contour
- Network memory can be trained by Hebbian learning
  - Guarantees that a network of N bits trained via Hebbian learning can store 0.14N random patterns with less than 0.4% probability that they will be unstable
- However, empirically it appears that we may sometimes be able to store *more than* 0.14N patterns

Poll 3

Poll 3

Mark all that are true

- **We can try to "assign" memories to a Hopfield network through Hebbian learning of the weights matrix**
- All patterns learned through Hebbian learning will be "remembered"
- The N-bit Hopfield network has the capacity to remember up to 0.14N patterns

Bold Claim

- I can *always* store (upto) N orthogonal patterns such that they are stationary!

  – Why?

- I can avoid spurious memories by adding some noise during recall!

The bottom line

- With a network of N units (i.e. N-bit patterns)
- The maximum number of stationary patterns is actually *exponential* in N
  - McElice and Posner, 84'
  - E.g. when we had the Hebbian net with N orthogonal base patterns, *all* patterns are stationary

- For a *specific* set of K patterns, we can *always* build a network for which all K patterns are stable provided $K \leq N$
  - Mostafa and St. Jacques 85'
    - For large N, the upper bound on K is actually N/4logN
      - McElice et. Al. 87'
  - **But this may come with many "parasitic" memories**

The bottom line

- With a network of N units (i.e. N-bit patterns)
- The maximum number of stationary patterns is actually *exponential* in N
  - McElice and Posner, 84'
  - E.g. when we had the H… …se patterns, *all* patterns are stable

How do we find this network?

- For a *specific* set of K patterns, we can *always* build a network for which all K patterns are stable provided $K \leq N$
  - Mostafa and St. Jacques 85'
    - For large N, the upper bound on K is actually N/4logN
      - McElice et. Al. 87'
  - **But this may come with many "parasitic" memories**

The bottom line

- With a network of N units (i.e. N-bit patterns)
- The maximum number of stationary patterns is actually *exponential* in N
  - McElice and Posner, 84'
  - E.g. when we had the H... ...se patterns, *all* patterns are stable

How do we find this network?

- For a *specific* set of K patterns, we can *always* build a network for which all K patterns are stable provided $K \leq N$
  - Mostafa and St. Jacques 85'
    - For large N, the upper bound on K is actually N...
      - McElice et. Al. 87'
  - **But this may come with many "parasitic" memories**

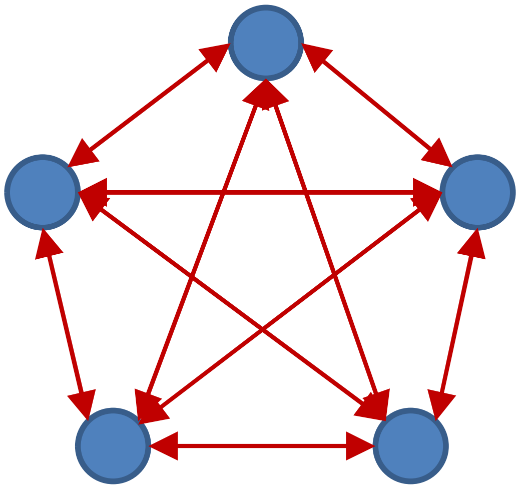
Can we do something about this?

A different tack

- How do we make the network store *a specific* pattern or set of patterns?
  - Hebbian learning
  - Optimization

- Secondary question
  - How many patterns can we store?

Consider the energy function



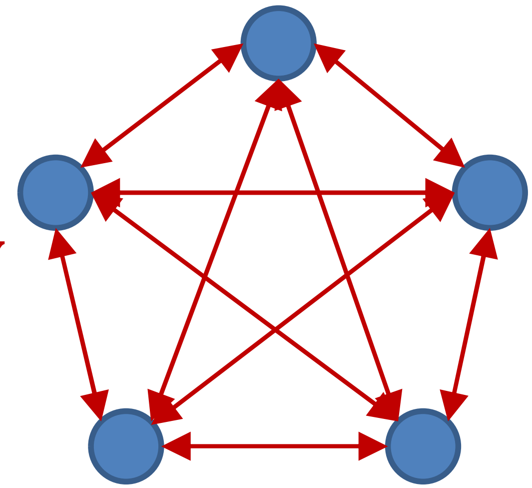
$$E = -\frac{1}{2}\sum_{i,j} w_{ij} y_i y_j - \sum_i b_i y_i$$

$$E = -\frac{1}{2}\mathbf{y}^T\mathbf{W}\mathbf{y} - \mathbf{b}^T\mathbf{y}$$

- This must be *maximally* low for target patterns
- Must be *maximally* high for *all other patterns*
  – So that they are unstable and evolve into one of the target patterns

Alternate Approach to Estimating the Network

$$E(\mathbf{y}) = -\frac{1}{2}\mathbf{y}^T\mathbf{W}\mathbf{y} - \mathbf{b}^T\mathbf{y}$$



- Estimate $\mathbf{W}$ (and $\mathbf{b}$) such that
  - E is minimized for $\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_P$
  - E is maximized for all other $\mathbf{y}$
- Caveat: Unrealistic to expect to store more than N patterns, but can we make those N patterns *memorable*

Optimizing W (and b)

$$E(\mathbf{y}) = -\frac{1}{2}\mathbf{y}^T\mathbf{W}\mathbf{y} \qquad \widehat{\mathbf{W}} = \underset{\mathbf{W}}{\operatorname{argmin}} \sum_{\mathbf{y}\in\mathbf{Y}_P} E(\mathbf{y})$$

The bias can be captured by another fixed-value component

- Minimize total energy of target patterns
  - Problem with this?

Optimizing W

$$E(\mathbf{y}) = -\frac{1}{2}\mathbf{y}^T\mathbf{W}\mathbf{y}$$

$$\widehat{\mathbf{W}} = \underset{\mathbf{W}}{\operatorname{argmin}} \sum_{\mathbf{y}\in\mathbf{Y}_P} E(\mathbf{y}) - \sum_{\mathbf{y}\notin\mathbf{Y}_P} E(\mathbf{y})$$

- Minimize total energy of target patterns
- Maximize the total energy of all *non-target* patterns

Optimizing W

$$E(\mathbf{y}) = -\frac{1}{2}\mathbf{y}^T\mathbf{W}\mathbf{y} \qquad \widehat{\mathbf{W}} = \underset{\mathbf{W}}{\operatorname{argmin}} \sum_{\mathbf{y}\in\mathbf{Y}_P} E(\mathbf{y}) - \sum_{\mathbf{y}\notin\mathbf{Y}_P} E(\mathbf{y})$$

- Simple gradient descent:

$$\mathbf{W} = \mathbf{W} + \eta\left(\sum_{\mathbf{y}\in\mathbf{Y}_P} \mathbf{y}\mathbf{y}^T - \sum_{\mathbf{y}\notin\mathbf{Y}_P} \mathbf{y}\mathbf{y}^T\right)$$

Set diagonal terms to 0 to eliminate self-edges

Hebbian learning
(which is why minimizing energy of target patterns is not enough)

"Anti" Hebbian learning

Optimizing W

$$\mathbf{W} = \mathbf{W} + \eta \left(\sum_{\mathbf{y} \in \mathbf{Y}_P} \mathbf{y}\mathbf{y}^T - \sum_{\mathbf{y} \notin \mathbf{Y}_P} \mathbf{y}\mathbf{y}^T \right)$$

- Can “emphasize” the importance of a pattern by repeating
  - More repetitions → greater emphasis

Optimizing W

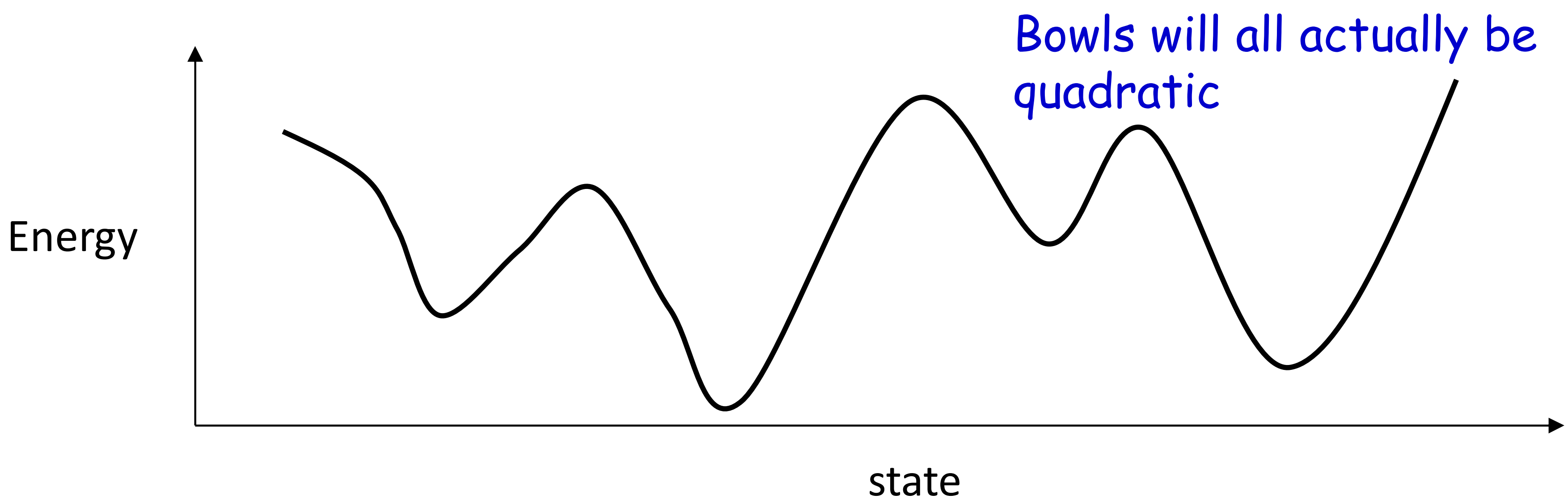
$$\mathbf{W} = \mathbf{W} + \eta \left(\sum_{\mathbf{y} \in \mathbf{Y}_P} \mathbf{y}\mathbf{y}^T - \sum_{\mathbf{y} \notin \mathbf{Y}_P} \mathbf{y}\mathbf{y}^T \right)$$

- Can "emphasize" the importance of a pattern by repeating
  - More repetitions → greater emphasis
- How many of these?
  - Do we need to include *all* of them?
  - Are all equally important?

The training again..

$$\mathbf{W} = \mathbf{W} + \eta \left(\sum_{\mathbf{y} \in \mathbf{Y}_P} \mathbf{y}\mathbf{y}^T - \sum_{\mathbf{y} \notin \mathbf{Y}_P} \mathbf{y}\mathbf{y}^T \right)$$

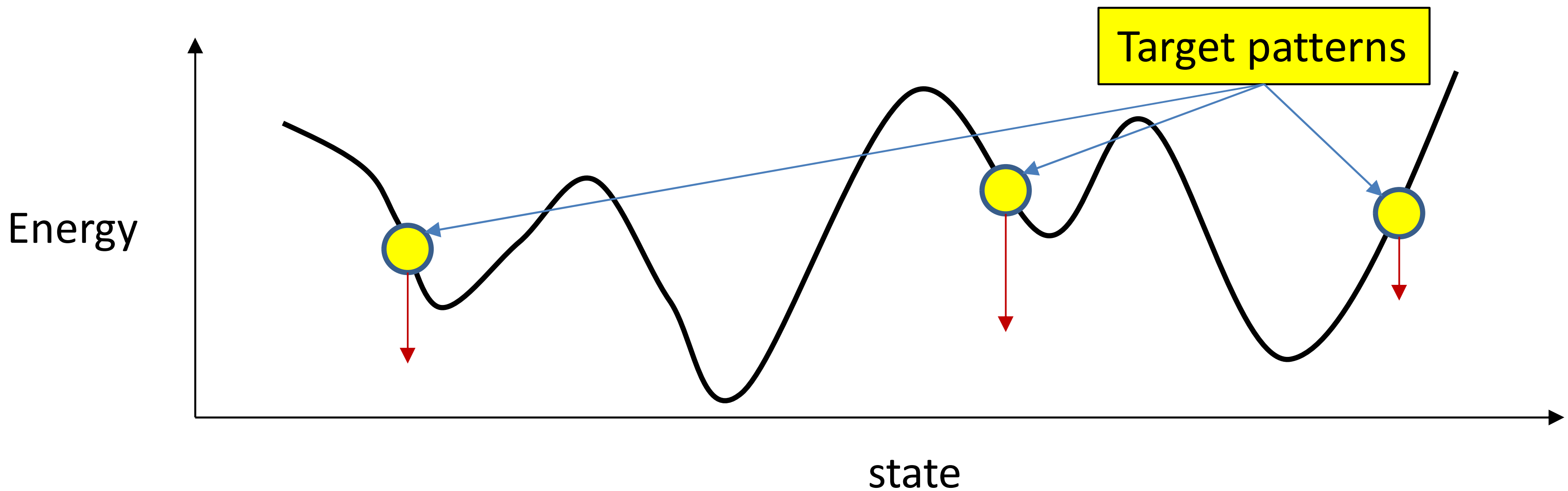
- Note the energy contour of a Hopfield network for any weight $\mathbf{W}$



The training again

$$\mathbf{W} = \mathbf{W} + \eta\left(\sum_{\mathbf{y}\in\mathbf{Y}_P} \mathbf{y}\mathbf{y}^T - \sum_{\mathbf{y}\notin\mathbf{Y}_P} \mathbf{y}\mathbf{y}^T\right)$$

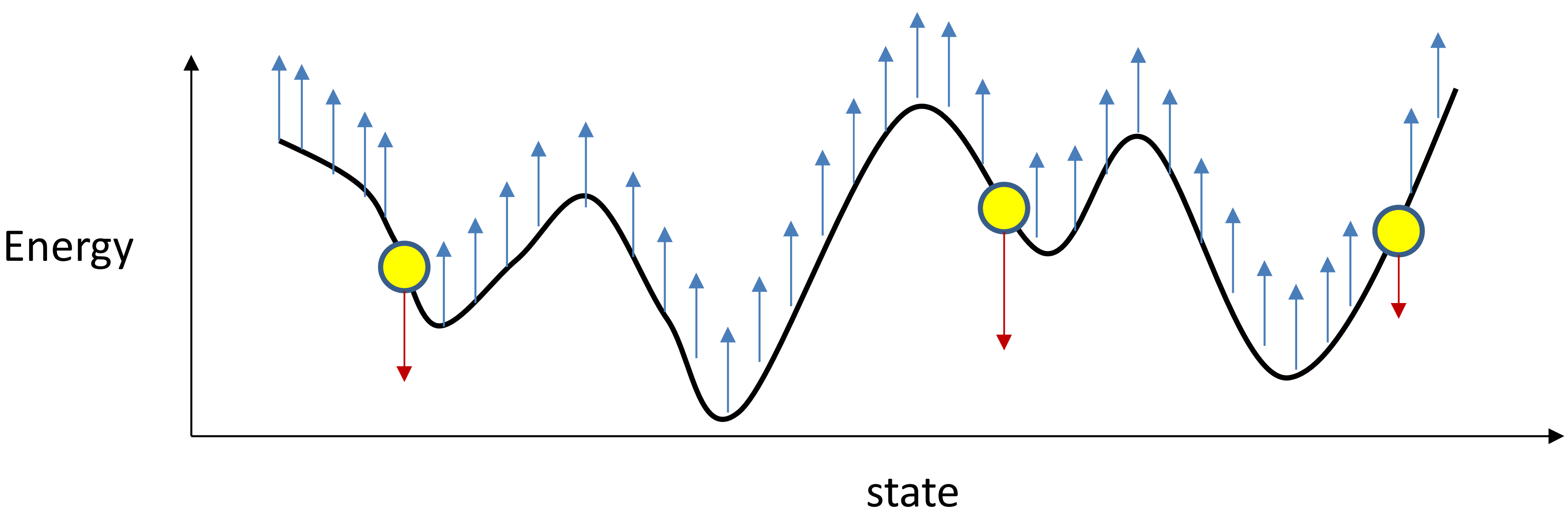
- The first term tries to *minimize* the energy at target patterns
  - Make them local minima
  - Emphasize more "important" memories by repeating them more frequently



The negative class

$$\mathbf{W} = \mathbf{W} + \eta\left(\sum_{\mathbf{y}\in\mathbf{Y}_P} \mathbf{y}\mathbf{y}^T - \sum_{\mathbf{y}\notin\mathbf{Y}_P} \mathbf{y}\mathbf{y}^T\right)$$

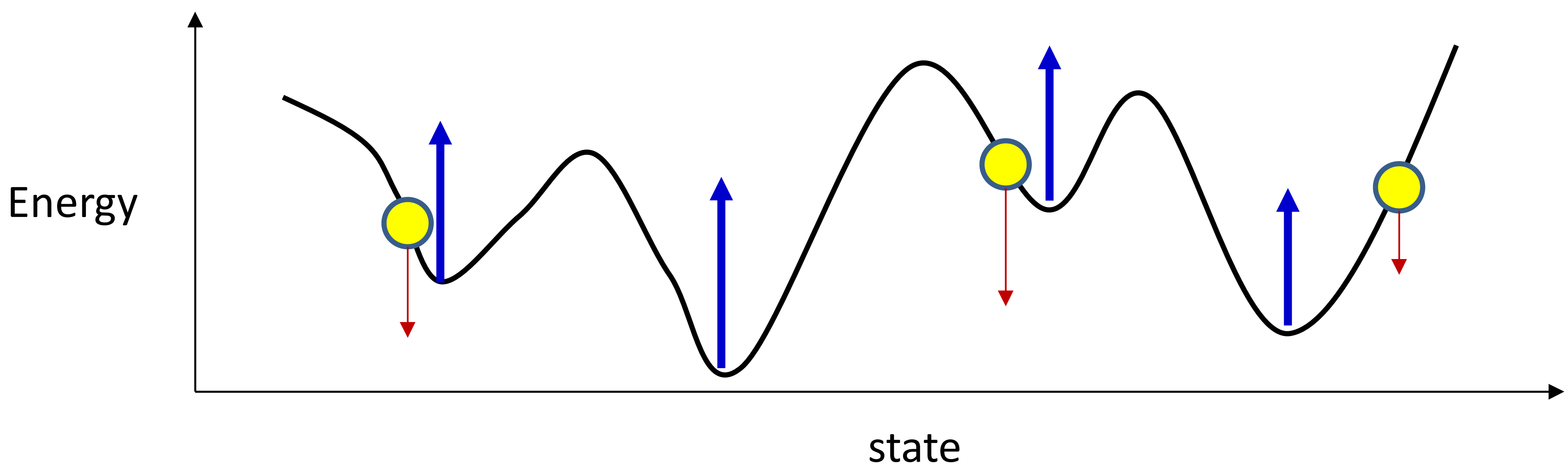
- The second term tries to "raise" all non-target patterns
  - Do we need to raise *everything*?



Option 1: Focus on the valleys

$$\mathbf{W} = \mathbf{W} + \eta\left(\sum_{\mathbf{y}\in\mathbf{Y}_P}\mathbf{y}\mathbf{y}^T - \sum_{\mathbf{y}\notin\mathbf{Y}_P \& \mathbf{y}=valley}\mathbf{y}\mathbf{y}^T\right)$$

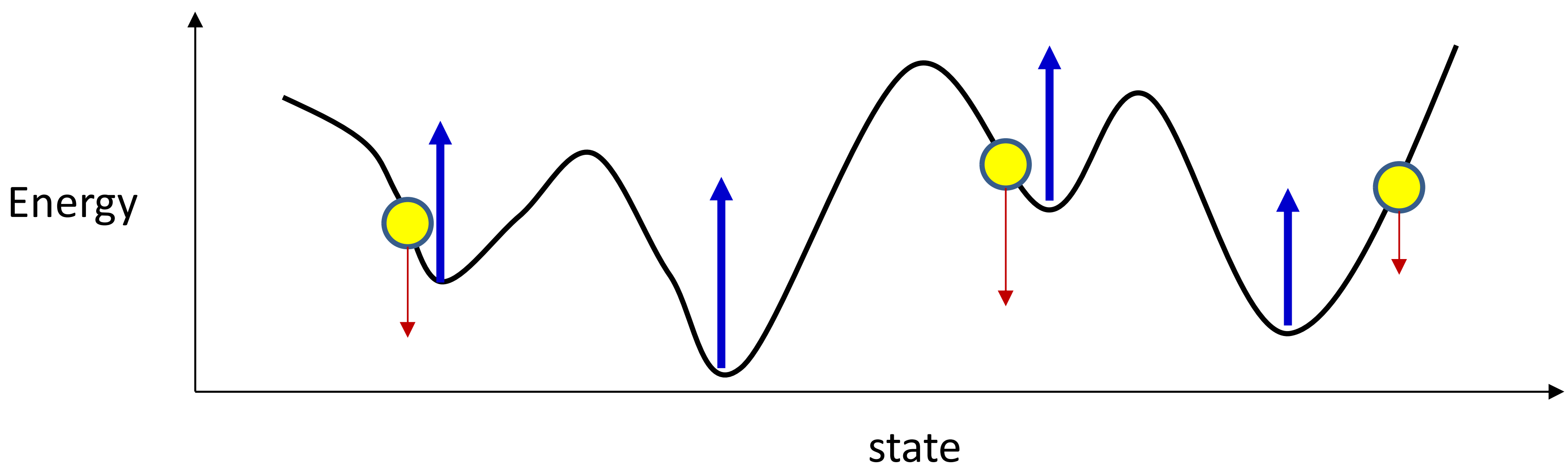
- Focus on raising the valleys
  - If you raise *every* valley, eventually they'll all move up above the target patterns, and many will even vanish



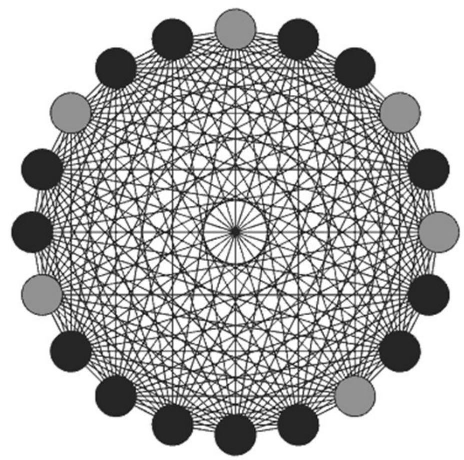
Identifying the valleys..

$$\mathbf{W} = \mathbf{W} + \eta\left(\sum_{\mathbf{y}\in\mathbf{Y}_P} \mathbf{y}\mathbf{y}^T - \sum_{\mathbf{y}\notin\mathbf{Y}_P \& \mathbf{y}=valley} \mathbf{y}\mathbf{y}^T\right)$$

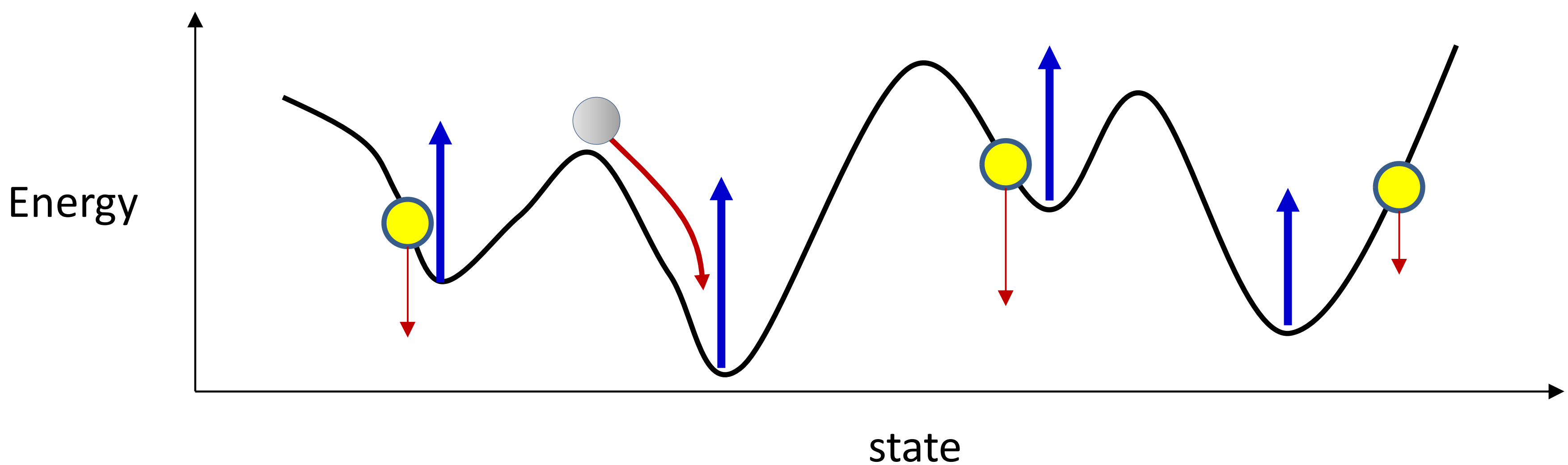
- Problem: How do you identify the valleys for the current $\mathbf{W}$?



Identifying the valleys..



- Initialize the network randomly and let it evolve
  - It will settle in a valley



Training the Hopfield network

$$\mathbf{W} = \mathbf{W} + \eta\left(\sum_{\mathbf{y}\in\mathbf{Y}_P} \mathbf{y}\mathbf{y}^T - \sum_{\mathbf{y}\notin\mathbf{Y}_P \& \mathbf{y}=valley} \mathbf{y}\mathbf{y}^T\right)$$

- Initialize $\mathbf{W}$
- Compute the total outer product of all target patterns
  - More important patterns presented more frequently
- Randomly initialize the network several times and let it evolve
  - And settle at a valley
- Compute the total outer product of valley patterns
- Update weights

Training the Hopfield network: SGD version

$$\mathbf{W} = \mathbf{W} + \eta \left(\sum_{\mathbf{y} \in \mathbf{Y}_P} \mathbf{y}\mathbf{y}^T - \sum_{\mathbf{y} \notin \mathbf{Y}_P \& \mathbf{y} = valley} \mathbf{y}\mathbf{y}^T \right)$$

- Initialize $\mathbf{W}$
- Do until convergence, satisfaction, or death from boredom:
  - Sample a target pattern $\mathbf{y}_p$
    - Sampling frequency of pattern must reflect importance of pattern
  - Randomly initialize the network and let it evolve
    - And settle at a valley $\mathbf{y}_v$
  - Update weights
    - $\mathbf{W} = \mathbf{W} + \eta \left(\mathbf{y}_p \mathbf{y}_p^T - \mathbf{y}_v \mathbf{y}_v^T \right)$

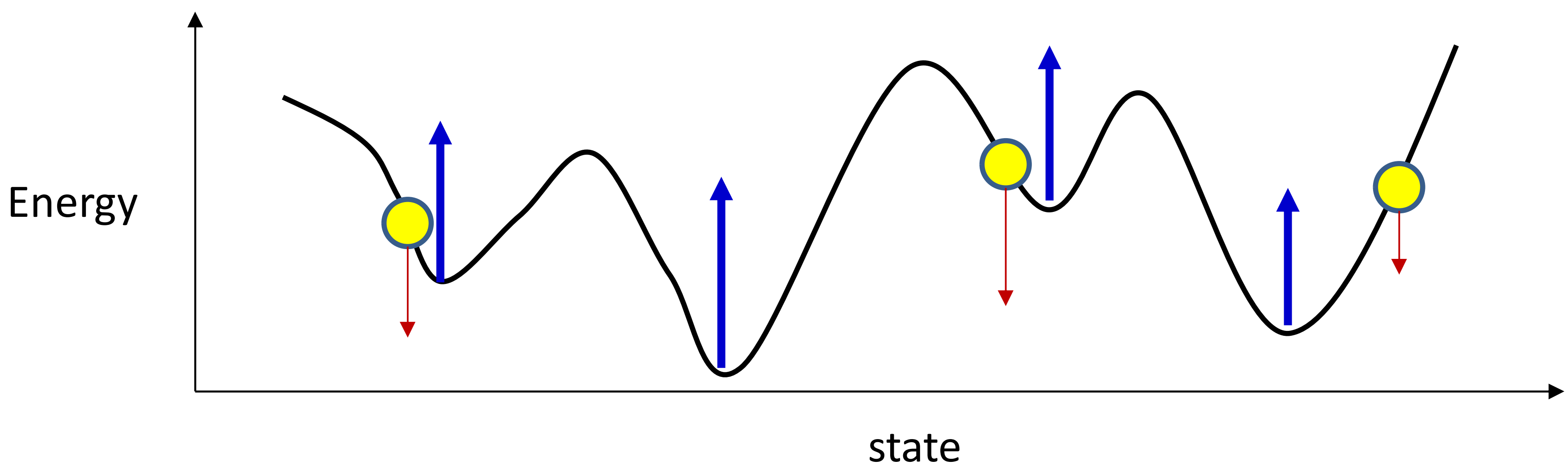
Training the Hopfield network

$$\mathbf{W} = \mathbf{W} + \eta \left(\sum_{\mathbf{y} \in \mathbf{Y}_P} \mathbf{y}\mathbf{y}^T - \sum_{\mathbf{y} \notin \mathbf{Y}_P \& \mathbf{y} = valley} \mathbf{y}\mathbf{y}^T \right)$$

- Initialize $\mathbf{W}$
- Do until convergence, satisfaction, or death from boredom:
  - Sample a target pattern $\mathbf{y}_p$
    - Sampling frequency of pattern must reflect importance of pattern
  - Randomly initialize the network and let it evolve
    - And settle at a valley $\mathbf{y}_v$
  - Update weights
    - $\mathbf{W} = \mathbf{W} + \eta(\mathbf{y}_p \mathbf{y}_p^T - \mathbf{y}_v \mathbf{y}_v^T)$

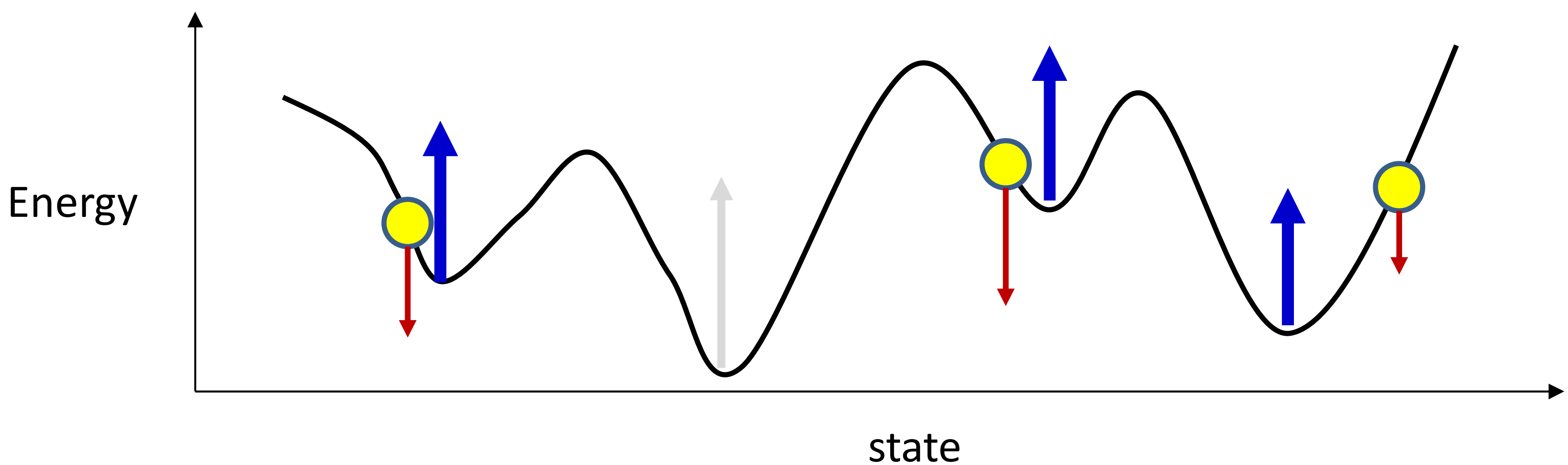
Which valleys?

- Should we *randomly* sample valleys?
  - Are all valleys equally important?

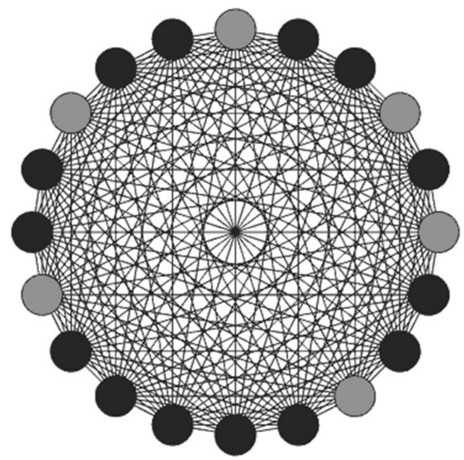


Which valleys?

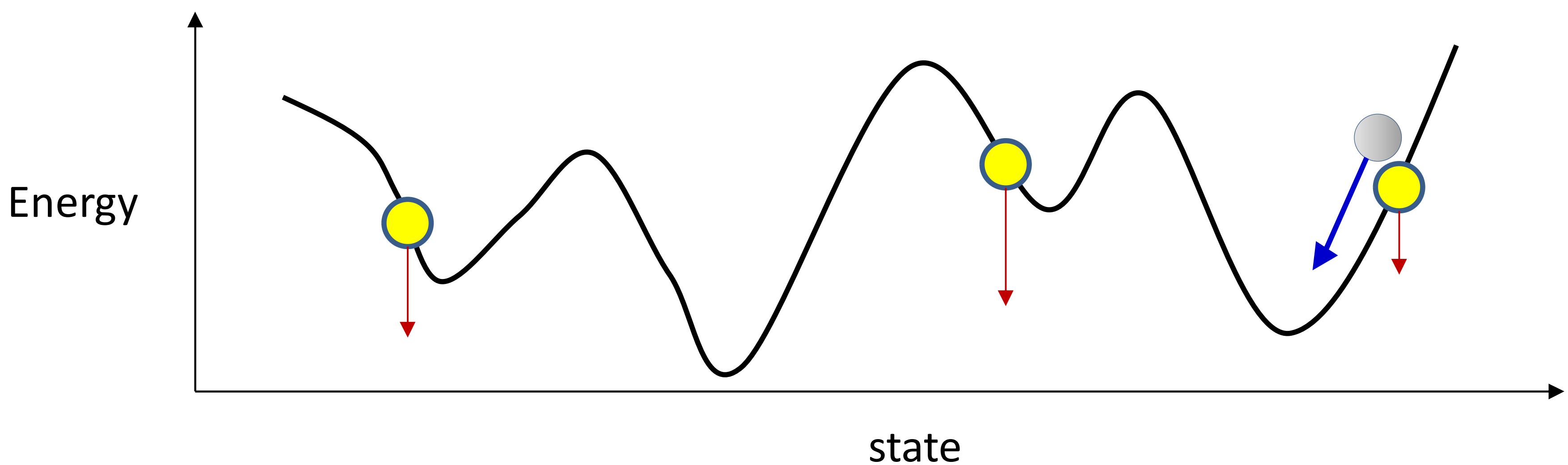
- Should we *randomly* sample valleys?
  – Are all valleys equally important?
- Major requirement: memories must be stable
  – They *must* be broad valleys
- Spurious valleys in the neighborhood of memories are more important to eliminate



Identifying the valleys..



- Initialize the network at valid memories and let it evolve
  - It will settle in a valley. If this is not the target pattern, raise it



Training the Hopfield network

$$\mathbf{W} = \mathbf{W} + \eta\left(\sum_{\mathbf{y}\in\mathbf{Y}_P} \mathbf{y}\mathbf{y}^T - \sum_{\mathbf{y}\notin\mathbf{Y}_P \& \mathbf{y}=valley} \mathbf{y}\mathbf{y}^T\right)$$

- Initialize $\mathbf{W}$
- Compute the total outer product of all target patterns
  - More important patterns presented more frequently
- Initialize the network with each target pattern and let it evolve
  - And settle at a valley
- Compute the total outer product of valley patterns
- Update weights

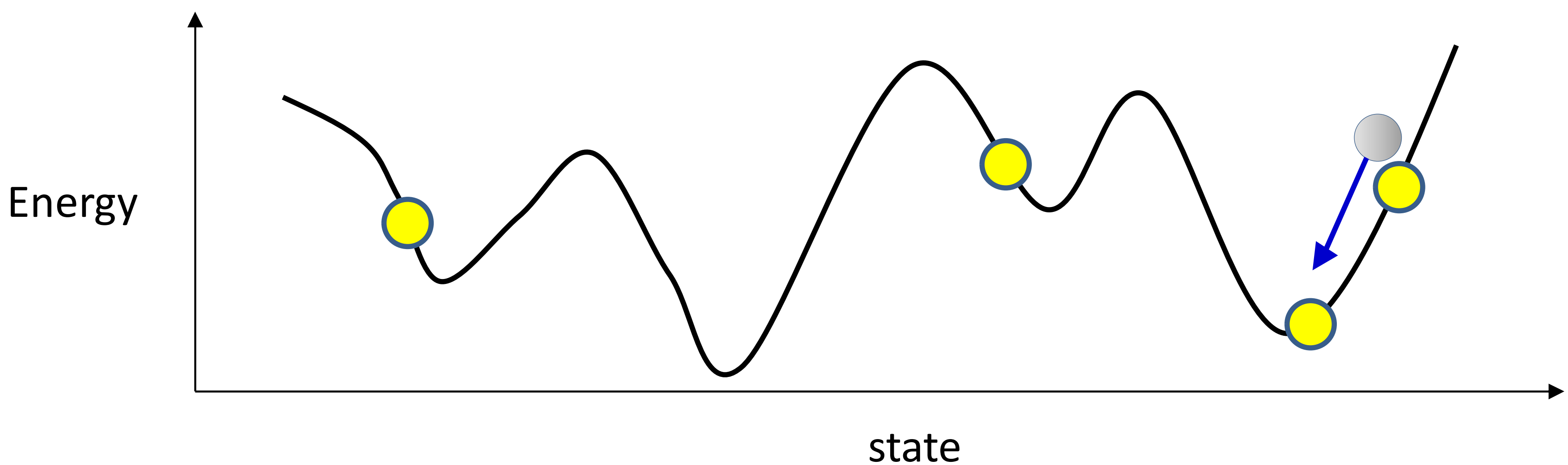
Training the Hopfield network: SGD version

$$\mathbf{W} = \mathbf{W} + \eta\left(\sum_{\mathbf{y}\in\mathbf{Y}_P}\mathbf{y}\mathbf{y}^T - \sum_{\mathbf{y}\notin\mathbf{Y}_P\&\mathbf{y}=valley}\mathbf{y}\mathbf{y}^T\right)$$

- Initialize $\mathbf{W}$
- Do until convergence, satisfaction, or death from boredom:
  – Sample a target pattern $\mathbf{y}_p$
    - Sampling frequency of pattern must reflect importance of pattern
  – Initialize the network at $\mathbf{y}_p$ and let it evolve
    - And settle at a valley $\mathbf{y}_v$
  – Update weights
    - $\mathbf{W} = \mathbf{W} + \eta\left(\mathbf{y}_p\mathbf{y}_p^T - \mathbf{y}_v\mathbf{y}_v^T\right)$

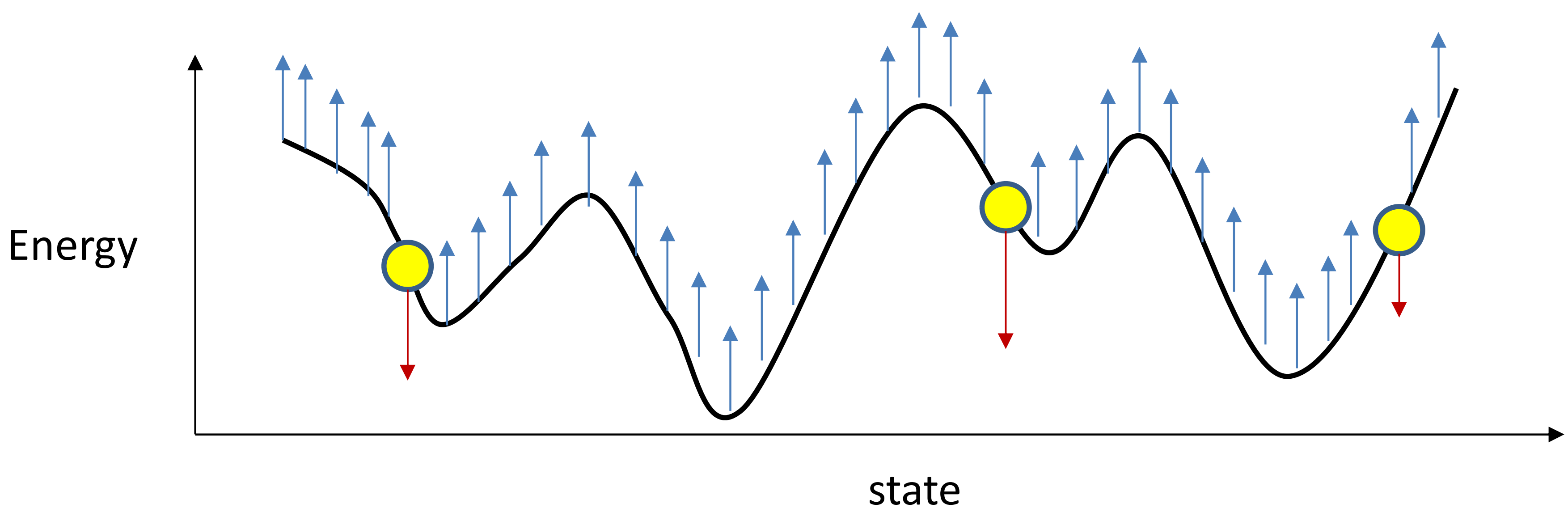
A possible problem

- What if there's another target pattern downvalley
  - Raising it will destroy a better-represented or stored pattern!



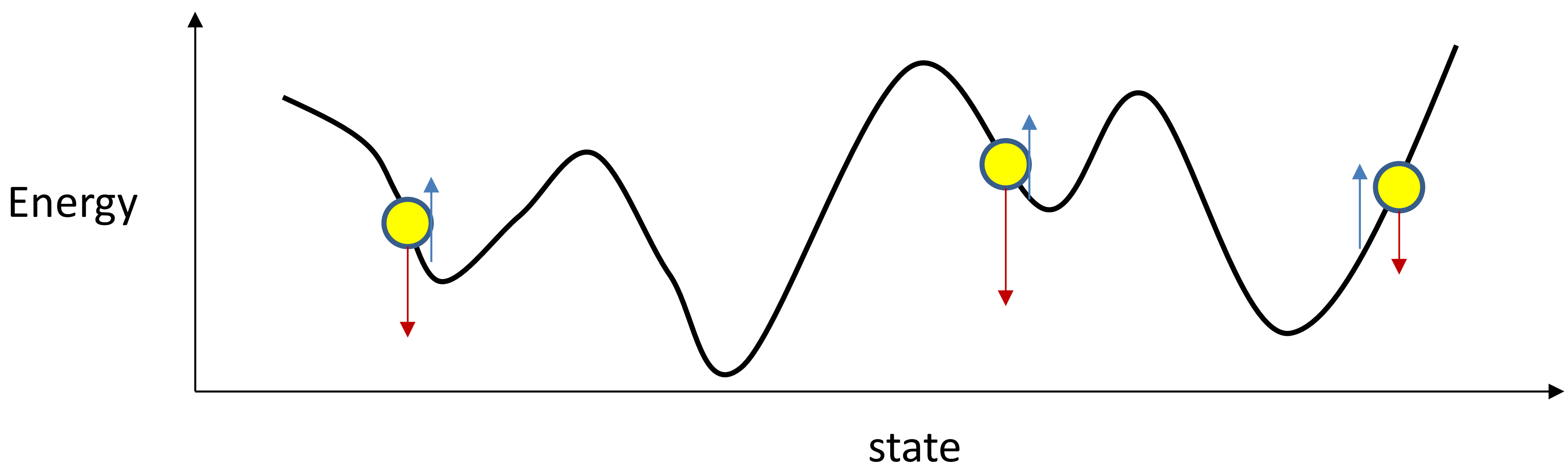
A related issue

- Really no need to raise the entire surface, or even every valley



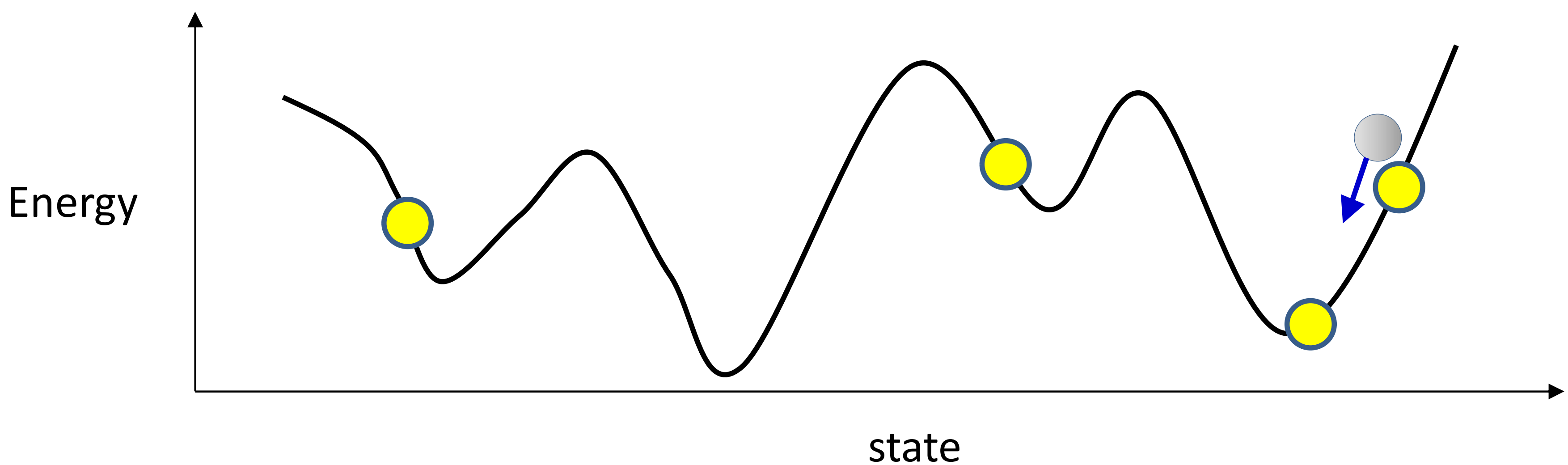
A related issue

- Really no need to raise the entire surface, or even every valley
- Raise the *neighborhood* of each target memory
  - Sufficient to make the memory a valley
  - The broader the neighborhood considered, the broader the valley



Raising the neighborhood

- Starting from a target pattern, let the network evolve only a few steps
  - Try to raise the resultant location
- Will raise the neighborhood of targets
- Will avoid problem of down-valley targets



Training the Hopfield network: SGD version

$$\mathbf{W} = \mathbf{W} + \eta\left(\sum_{\mathbf{y}\in\mathbf{Y}_P}\mathbf{y}\mathbf{y}^T - \sum_{\mathbf{y}\notin\mathbf{Y}_P \& \mathbf{y}=valley}\mathbf{y}\mathbf{y}^T\right)$$

- Initialize $\mathbf{W}$
- Do until convergence, satisfaction, or death from boredom:
  – Sample a target pattern $\mathbf{y}_p$
    - Sampling frequency of pattern must reflect importance of pattern
  – Initialize the network at $\mathbf{y}_p$ and let it evolve ***a few steps (2-4)***
    - And arrive at a down-valley position $\mathbf{y}_d$
  – Update weights
    - $\mathbf{W} = \mathbf{W} + \eta\left(\mathbf{y}_p\mathbf{y}_p^T - \mathbf{y}_d\mathbf{y}_d^T\right)$

Poll 4

Poll 4

Mark all that are true about the optimization-based method to store memories in a Hopfield net

- **It finds weights that minimize the energy of target patterns**
- **It maximizes the energy of non-target patterns**
- It is an exact gradient descent formulation
- **It minimizes the energy of target patterns through Hebbian learning**
- **It maximizes the energy of non-target patterns through Hebbian learning**

Story so far

- Hopfield nets with N neurons can store up to $0.14N$ patterns through Hebbian learning
  - Issue: Hebbian learning assumes all patterns to be stored are equally important

- In theory the number of *intentionally* stored patterns (stationary *and* stable) can be as large as N
  - But comes with many parasitic memories

- Networks that store $O(N)$ memories can be trained through optimization
  - By minimizing the energy of the target patterns, while increasing the energy of the neighboring patterns