

Symmetric Networks

▶ Hertz, Krogh, Palmer: Introduction to the Theory of Neural Computation. Addison-Wesley Publishing Company (1991).

▶ How can we model an associative memory?

▷ Let $M = \{\bar{v}_1, \dots, \bar{v}_m\}$ be a set of patterns.

▷ Here, patterns are bit vectors of length l .

▷ Let \bar{x} be a bit vector of length l .

▷ Find $\bar{v}_j \in M$ which is most similar to \bar{x} .

▶ Possible solution:

▷ For all $j = 1 \dots m$ compute Hamming distance between \bar{v}_j and \bar{x} :

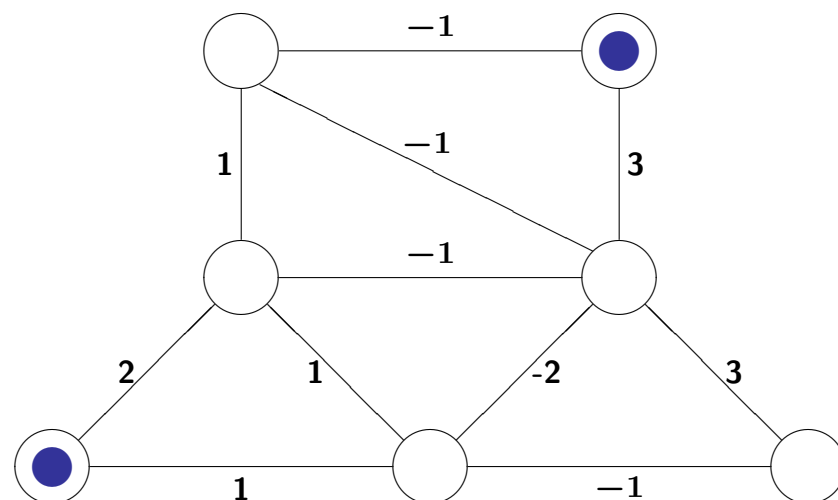
$$\sum_{i=1}^l (v_{ji}(1 - x_i) + (1 - v_{ji})x_i)$$

▷ Select \bar{v}_j , whose Hamming distance to \bar{x} is smallest.

McCulloch-Pitts Networks and Associative Memories

- ▶ Can we use McCulloch-Pitts networks as associative memories?
- ▶ Idea:
 - ▷ Activate a network externally by \bar{x} .
 - ▷ Search for weights such that after some time the network reaches a stable state which represents \bar{v}_j .

An Example Network



► **Update:**

```
 $t = 0;$   
do while current state is unstable;  
  select an arbitrary unit  $u_k$ ;  
  update  $u_k$ ;  
   $t := t + 1$ ;  
end;
```

► **Exercise:** Find the stable states of the network shown on this slide.

Attractors

- ▶ Consider space of states of a given network.
 - ▷ Stable states are called **attractors**.
 - ▷ Systems starts in one state corresponding to \bar{x} .
 - ▷ **Trajectories** lead to attractors.
 - ▷ **Basins of attractors**.
- ▶ **Exercise:**
Specify all basins of attractors and all trajectories for the network shown previously.

Notational Convention

▶ To simplify the mathematical model we assume:

▶ Threshold $\theta_k = 0$ for all units u_k .

▶ Output $v_j \in \{-1, 1\}$ for all units u_k .

▶ **Exercise:** Why is this not a restriction?

▶ Let l be the number of units in the network.

▶ Then

$$v_i = \operatorname{sgn}\left(\sum_{j=1}^l w_{ij}v_j\right),$$

where

$$\operatorname{sgn}(x) = \begin{cases} 1 & \text{if } x \geq 0, \\ -1 & \text{otherwise.} \end{cases}$$

▶ Note the case $x = 0$.

Storing a Single Bit Vector

- ▶ How shall the weights look like?
- ▶ Let \bar{v} be a bit vector of length l .
- ▶ \bar{v} is a stable state if for all i we find

$$v_i = \text{sgn}\left(\sum_{j=1}^l w_{ij}v_j\right).$$

- ▶ This holds if the weights are proportional to $v_i v_j$, e.g., $w_{ij} = \frac{1}{l}v_i v_j$:

$$\begin{aligned}v_i &= \text{sgn}\left(\sum_{j=1}^l \frac{1}{l}v_i v_j v_j\right) \\ &= \text{sgn}\left(\sum_{j=1}^l \frac{1}{l}v_i\right) \\ &= \text{sgn } v_i \\ &= v_i\end{aligned}$$

- ▶ Errors in \bar{x} are corrected if $\#\text{errors}(\bar{x}) < \frac{l}{2}$.
- ▶ \bar{v} is an attractor.
- ▶ But there is another attractor: $-\bar{v}$
 - ▶ reached if $\#\text{errors}(\bar{x}) > \frac{l}{2}$.

Storing Several Bit Vectors

- ▶ Let m be the number of bit vectors and l the number of units in the network:

$$w_{ij} = \frac{1}{l} \sum_{k=1}^m v_{ki} v_{kj}.$$

- ▶ (Generalized) Hebb rule (Hebb 1949).
- ▶ Are all vectors $\bar{v}_r \in M$ stable states?

$$\begin{aligned} v_{ri} &= \operatorname{sgn}\left(\sum_{j=1}^l w_{ij} v_{rj}\right) \\ &= \operatorname{sgn}\left(\frac{1}{l} \sum_{j=1}^l \sum_{k=1}^m v_{ki} v_{kj} v_{rj}\right) \\ &= \operatorname{sgn}\left(v_{ri} + \frac{1}{l} \sum_{j=1}^l \sum_{k=1, k \neq r}^m v_{ki} v_{kj} v_{rj}\right). \end{aligned}$$

- ▶ Let $C_{ri} = \frac{1}{l} \sum_{j=1}^l \sum_{k=1, k \neq r}^m v_{ki} v_{kj} v_{rj}$.
- ▶ If $C_{ri} = 0$ for each i then each vector is stable state.
- ▶ If $|C_{ri}| < 1$ for each i then it cannot change sign of v_{ri} .
- ▶ Storage capacity: If vectors are stochastically independent and should be perfectly recalled then the maximum storage capacity is proportional to $\frac{l}{\log l}$.

Hopfield and Symmetric Networks

- ▶ A network realizing an associative memory as shown on the previous slide is often called **Hopfield network**.
- ▶ J.J. Hopfield: **Neural Networks and Physical Systems with Emergent Collective Computational Abilities**. In: **Proceedings of the National Academy of Sciences USA**, 2554-2558 (1982).
- ▶ Because $w_{ij} = w_{ji}$ it is also called **symmetric network**.
- ▶ **Exercise:** Suppose we want to store the vectors $(-1, -1, 1, -1, 1, -1)$ and $(1, 1, -1, -1, 1, 1)$ in a symmetric network with $N = 6$ units. Construct the network which solves this problem.

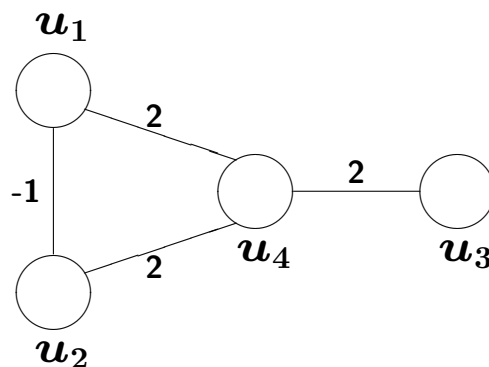
Energy Function

- ▶ What happens precisely when a symmetric network is updated?
- ▶ Consider the energy function

$$E(t) = -\frac{1}{2} \sum_{i,j=1}^l w_{ij} v_i(t) v_j(t)$$

describing the state of a symmetric network at time t .

- ▶ We assume $w_{ii} = 0$ for all $1 \leq i \leq N$.
- ▶ Example:



$$E(t) = v_1(t)v_2(t) - 2v_1(t)v_4(t) - 2v_2(t)v_4(t) - 2v_3(t)v_4(t).$$

Properties of the Energy Function

- ▶ **Theorem:** E is monoton decreasing, i.e., $E(t + 1) \leq E(t)$.
- ▶ **Exercise:** How does an update change the energy of a symmetric network if we do not assume that $w_{ii} = 0$?
- ▶ **Exercise:** Is the energy function still monoton decreasing if we do not assume that $w_{ij} = w_{ji}$? Prove your answer.
- ▶ **How plausible is the assumption that $w_{ij} = w_{ji}$?**
- ▶ **Exercise:** Consider symmetric networks, where the threshold of the units need not be 0. Define a monoton decreasing energy function for these networks. Prove your claim.

Relation to Ising Models

- ▶ **Spins**, i.e., magnetic atoms with directions 1 and -1 .
- ▶ Suppose there are l atoms.
- ▶ For each atom v_i a magnetic field h_i is defined by

$$h_i = \sum_{j=1}^l w_{ij} v_j + h^e$$

where h^e is external field.

- ▶ At low temperatures spins follow magnetic field. This is described by energy function

$$H = -\frac{1}{2} \sum_{i,j=1}^l w_{ij} v_i v_j - h^e \sum_{i=1}^l v_i.$$

More on Ising Models

- ▶ At high temperatures spins do not follow magnetic field.
- ▶ Thermal fluctuations depending on the temperature.
- ▶ Mathematical model: **Glauber dynamics**

$$v_i = \begin{cases} 1 & \text{with probability } g(h_i), \\ -1 & \text{with probability } 1 - g(h_i), \end{cases}$$

where

$$g(h) = \frac{1}{1 + \exp(-2\beta h)}, \quad \beta = \frac{1}{k_B T}, \quad k_B \text{ Boltzmann's constant.}$$

- ▶ $1 - g(h) = g(-h)$.
- ▶ Behaviour of spins:

$$\text{prob}(v_i = \pm 1) = \frac{1}{1 + \exp(\mp 2\beta h_i)}.$$

- ▶ In equilibrium states with low energy are more likely than states with higher energy.

Stochastic Networks

- ▶ Hinton, Sejnowski: Optimal Perceptual Inference. In: Proceedings of the IEEE Conference on Computer Vision and Recognition, 448-453 (1983).

- ▶ They applied these results to symmetric networks:

$$\text{prob}(v_i = 1) = \frac{1}{1 + \exp(-\beta \sum_{j=1}^l w_{ij} v_j)} \quad \text{where} \quad \beta = \frac{1}{T}.$$

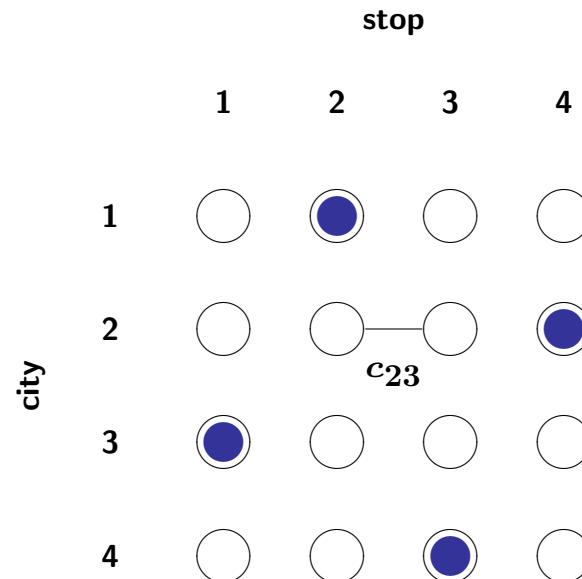
- ▶ Those networks are called **Boltzmann machines** or **stochastic networks**.
- ▶ $T = 0$: symmetric networks.
- ▶ Kirkpatrick, Gelatt, Vecchi: Optimization by simulated annealing. Science 220, 671-680 (1983)
 - ▶ **Simulated annealing**.
- ▶ Geman, Geman: Stochastic Relaxation, Gibbs Distribution, and the Bayesian Restoration of Images. IEEE Transactions on Pattern Analysis and Machine Intelligence, 6, 721-741 (1984).
 - ▶ Simulated annealing is guaranteed to find a global minima of the energy function if temperature is lowered in infinitesimal small steps.

Combinatorial Optimization Problem

- ▶ We consider problems of size n having typically e^n or $n!$ many possible solutions, of which we want to find the “optimal” one.
- ▶ Class \mathcal{P} :
there exists a deterministic algorithm solving the problem in polynomial time.
- ▶ Class \mathcal{NP} (non-deterministic polynomial):
one can test in polynomial time whether any “guess” of the solution is right.
- ▶ \mathcal{NP} -complete problem:
if one could find a deterministic algorithm solving the problem in polynomial time, then all other \mathcal{NP} problems could be solved in polynomial time.
- ▶ $\mathcal{P} \neq \mathcal{NP}$.
- ▶ Garey, Johnson: *Computers and Intractability*. H. Freeman and Company (1979).

The Travelling Salesman Problem

- ▶ **Given:** n cities and costs c_{ij} for traveling from city j to city i .
- ▶ **Problem:** Find a tour visiting each city exactly once and return to the start city such that the accumulated costs of the tour are minimal.



Modelling the Travelling Salesman Problem (1)

- ▶ Binary threshold units:

$$v_{ik} = \begin{cases} 1 & \text{if the } k\text{th stop is in the } i\text{th city,} \\ 0 & \text{otherwise.} \end{cases}$$

- ▶ Costs of the tour:

$$\frac{1}{2} \sum_{i,j,k=1}^n c_{ij} v_{ik} (v_{j,k+1} + v_{j,k-1}),$$

where indices are taken modulo n .

- ▶ Each city occurs only once on the tour:

$$(\forall i) \sum_{k=1}^n v_{ik} = 1.$$

- ▶ Each stop on the tour is at just one city:

$$(\forall k) \sum_{i=1}^n v_{ik} = 1.$$

Modelling the Travelling Salesman Problem (2)

- ▶ Altogether, we obtain

$$\frac{1}{2} \sum_{i,j,k=1}^n c_{ij} v_{ik} (v_{j,k+1} + v_{j,k-1}) + \frac{\gamma}{2} \left(\sum_{i=1}^n (1 - \sum_{k=1}^n v_{ik})^2 + \sum_{k=1}^n (1 - \sum_{i=1}^n v_{ik})^2 \right).$$

- ▶ This corresponds to an energy function.
- ▶ **Exercise:** Construct the corresponding symmetric network.
- ▶ **Exercise:** What is the role of γ ?
- ▶ See (Hertz et al. 1991) for a discussion of different solutions for the travelling salesman problem.

Graph Bipartitioning

- **Exercise:** Consider the following optimization problem: A graph with an even number of vertices shall be bipartitioned with a minimal cut. I.e., the vertices of the graph shall be split into two sets A and B having the same cardinality such that the number of edges going from A to B is minimal.
- (a) Construct an energy function such that the minima of the energy function correspond precisely to the minimal cut through the graph.
- (b) Specify an additional term ensuring $|A| = |B|$.

Hint: Let

$$v_i = \begin{cases} 1 & \text{if vertex } i \in A, \\ -1 & \text{otherwise} \end{cases}$$

and

$$c_{ij} = \begin{cases} 1 & \text{if there is an edge from } j \text{ to } i, \\ 0 & \text{otherwise.} \end{cases}$$

Propositional Logic

- ▶ Pinkas: Symmetric Networks and Logic Satisfiability. *Neural Computation* 3, 282-291 (1991).
- ▶ **Propositional variables:** $\{V_1, \dots, V_n\}$.
- ▶ **Connectives:** $\neg, \wedge; \vee, \rightarrow, \leftrightarrow$ as abbreviations.
- ▶ **Formulas:** $F(\bar{V}), G(\bar{V})$.
 - ▷ We will omit the variables if they can be inferred from the context.
- ▶ **Interpretation:** mapping from $\{V_1, \dots, V_n\}$ to $\{0, 1\}$ encoded as \bar{v} .

$$\begin{aligned}V_i[\bar{v}] &= v_i \\(\neg F)[\bar{v}] &= 1 - F[\bar{v}] \\(F \wedge G)[\bar{v}] &= F[\bar{v}] \times G[\bar{v}]\end{aligned}$$

- ▶ **Example:**

$$\begin{aligned}(\neg(V_1 \wedge V_2) \vee V_3)[101] &= (\neg((V_1 \wedge V_2) \wedge \neg V_3))[101] \\&= 1 - (1 \times 0) \times (1 - 1) \\&= 1.\end{aligned}$$

Visible and Hidden Variables

- ▶ We will distinguish between **visible** (\bar{V}) and **hidden** (\bar{H}) variables: $F(\bar{V}, \bar{H})$.
- ▶ **Definition:** \bar{v} is a **visible model** for $F(\bar{V}, \bar{H})$
iff there exists \bar{h} such that $F[\bar{v}, \bar{h}] = 1$.
- ▶ **Definition:** $F_1[\bar{X}, \bar{Y}] \equiv_F F_2[\bar{X}, \bar{Z}]$
iff the visible models for F_1 and F_2 are equal.
- ▶ **Definition:** A formula F is in **conjunctive triple** (or **CTF-**) **form** **iff** it is of the form

$$F = \bigwedge_{i=1}^m F_i$$

and in each F_i there occur at most three variables.

Transformation into Conjunctive Triple Form

- ▶ **Proposition:** Each formula F can be transformed into a formula G in CTF-form by introducing new hidden variables such that $F \equiv_F G$.
- ▶ **Example:**

$$\begin{aligned} & ((V_1 \vee \neg V_2) \rightarrow (V_3 \vee V_4)) \\ \equiv_F & ((V_1 \vee \neg V_2) \leftrightarrow H_1) \wedge ((V_3 \vee V_4) \leftrightarrow H_2) \wedge (H_1 \rightarrow H_2). \end{aligned}$$

- ▶ **Exercise:** Proof this proposition.
- ▶ **Exercise:** Let l be the length of a formula F . Show that the transformation into CTF-form requires at most $O(l)$ steps and $O(l)$ many new hidden variables.
- ▶ **Exercise:** Show that there is a linear algorithm for transforming formulas into conjunctive normal form. How much time does the standard transformation (as, for example, presented in the ICL lecture) require in the worst case?

Energy Functions Revisited

- ▶ **Definition:** An **energy function** is a mapping $\{0, 1\}^n \rightarrow \mathbb{R}$. It has a **degree** of k and will be written $E^k(\bar{V})$ **iff** it can be written as sum of product terms each consisting of at most k variables.
- ▶ So far we have considered only energy functions of degree 2.
- ▶ We will distinguish between visible and hidden variables: $E^k(\bar{V}, \bar{H})$.
- ▶ The evaluation of an energy function $E^k[\bar{v}, \bar{h}]$ is defined as usual.
- ▶ **Definition:** \bar{v} is a **visible solution** of E^k **iff** there exists \bar{h} such that

$$E^k[\bar{v}, \bar{h}] = \min_{\bar{v}', \bar{h}'} \{E^k[\bar{v}', \bar{h}']\}.$$

- ▶ **Example:** Consider

$$E^2 = V_1V_2 - 2V_1H - 2V_2H - 2V_3H + 5H.$$

Visible solutions are: [000], [001], [111].

- ▶ **Definition:** $E_1^k(\bar{V}, \bar{H}_1) \equiv_E E_2^l(\bar{V}, \bar{H}_1)$ **iff** the sets of visible solutions for E_1^k and E_2^l are identical.

Changing the Degree of Energy Functions

► **Theorem:** Let E^k be an energy function. There exists \overline{H}' such that:

$$(1) \quad E^k(\overline{V}, \overline{H}) \equiv_E E^{k-1}(\overline{V}, \overline{H} \cdot \overline{H}')$$

$$(2) \quad E^k(\overline{V}, \overline{H} \cdot \overline{H}') \equiv_E E^{k+1}(\overline{V}, \overline{H})$$

► **Example:**

$$V_1V_2 - V_1V_2V_3 \equiv_E V_1V_2 - 2V_1H - 2V_2H - 2V_3H + 5H.$$

► **Exercise:** Complete the proof of the theorem.

► **Exercise:** How many hidden variables have to be introduced in the worst case if an energy function of degree $k \geq 3$ and with n variables is transformed into a quadratic one?

The Satisfiability Problem

▶ **Satisfiability Problem:**

Given $F(\bar{V})$. Does there exist an interpretation \bar{v} such that $F[\bar{v}] = 1$?

▶ **Idea:** Find an energy function E corresponding to F such that the global minimal of E correspond precisely to the models of F .

▶ **Definition:** $F(\bar{V}, \bar{H}) \equiv_T E(\bar{V}, \bar{H})$

iff the set of visible models of F is equal to set of visible solutions for E .

▶ **Example:**

$$(\neg(V_1 \wedge V_2) \vee V_3) \equiv_T V_1V_2 - 2V_1H - 2V_2H - 2V_3H + 5H.$$

From Propositional Logic To Energy Functions

▶ Given $F = \bigwedge_{i=1}^m F_i$ in CTF-form.

▶ **Definition:** The **penalty function** corresponding to $F(\bar{V})$ is defined as:

$$P_{F(\bar{V})} = \sum_{i=1}^m (1 - F_i[\bar{V}]).$$

▶ **Example:**

$$\begin{aligned} P_{(\neg(V_1 \wedge V_2) \vee V_3)} &= V_1 V_2 - V_1 V_2 V_3 \\ &= V_1 V_2 - 2V_1 H - 2V_2 H - 2V_3 H + 5H. \end{aligned}$$

▶ **Exercise:** Consider $F = ((V_1 \wedge V_2) \rightarrow V_3) \wedge (V_4 \rightarrow \neg V_1) \wedge (V_1 \vee \neg V_3)$.

(a) Specify the energy function E corresponding directly to F .

(b) Transform E into a quadratic energy function.

(c) Construct the corresponding symmetric network.

▶ **Exercise:** What is gained by using the CTF-form?

From Energy Functions to Propositional Logic

- ▶ Given $E(V_1, \dots, V_n)$ be an energy function.
- ▶ Compute the set S of visible solutions.
- ▶ For each $\bar{v} \in S$ and $1 \leq i \leq n$ let

$$L_{\bar{v}}^i = \begin{cases} V_i & \text{if } v_i = 1, \\ \neg V_i & \text{otherwise.} \end{cases}$$

- ▶ Construct the formula

$$F_{E(\bar{V})} = \bigvee_{\bar{v} \in S} \left(\bigwedge_{i=1}^n L_{\bar{v}}^i \right).$$

Propositional Non-Monotonic Reasoning

- ▶ Pinkas: Propositional Non-Monotonic Reasoning and Inconsistency in Symmetrical Neural Networks. In: Proceedings IJCAI, 525-530 (1991).
- ▶ Consider formulas F of the form $F = \{\langle F_1, n_1 \rangle, \dots, \langle F_m, n_m \rangle\}$, where $n_i \in \mathbb{N}$, $1 \leq i \leq m$.

- ▶ **Example:**

$$F = \left\{ \begin{array}{ll} \langle O \rightarrow M, & 1 \rangle, \\ \langle S \rightarrow \neg M, & 2 \rangle, \\ \langle C \rightarrow S, & 4 \rangle, \\ \langle C \rightarrow M, & 4 \rangle, \\ \langle V \rightarrow \neg M, & 4 \rangle \end{array} \right\}$$

- ▶ **Definition:** Penalty of \bar{v} for $\langle F_i, n_i \rangle$:

$$P(\bar{v}, F_i) = \begin{cases} n_i & \text{if } F_i[\bar{v}] = 0, \\ 0 & \text{otherwise.} \end{cases}$$

Penalty of \bar{v} for F : $P(\bar{v}, F) = \sum_{i=1}^m P(\bar{v}, F_i)$.

- ▶ **Definition:** \bar{v}_1 is preferred over \bar{v}_2 wrt F iff $P(\bar{v}_1, F) < P(\bar{v}_2, F)$.
 \bar{v}_1 is most preferred wrt F iff $\neg(\exists \bar{v}_2) P(\bar{v}_2, F) < P(\bar{v}_1, F)$.

Propositional Non-Monotonic Reasoning: An Example

- ▶ Consider again:

$$F = \left\{ \begin{array}{ll} \langle O \rightarrow M, & 1 \rangle, \\ \langle S \rightarrow \neg M, & 2 \rangle, \\ \langle C \rightarrow S, & 4 \rangle, \\ \langle C \rightarrow M, & 4 \rangle, \\ \langle V \rightarrow \neg M, & 4 \rangle \end{array} \right\}$$

- ▶ Assume $O \prec M \prec S \prec C \prec V$.
- ▶ Consider $\bar{v}_1 = 11000$ and $\bar{v}_2 = 00000$.
- ▶ $P(\bar{v}_1, F) = P(\bar{v}_2, F) = 0$.
 - ▶ \bar{v}_1 and \bar{v}_2 are most preferred.
- ▶ Let $F_1 = F \cup \{\langle O, \infty \rangle\}$.
 - ▶ \bar{v}_1 is the only most preferred interpretation.
- ▶ Let $F_2 = F_1 \cup \{\langle S, \infty \rangle\}$.
 - ▶ 10100 and 10101 are the most preferred interpretations.
- ▶ Nonmonotonicity.

Mapping to Energy Function

► Modified penalty function

$$P_{F(\bar{V})} = \sum_{i=1}^m n_i (1 - F_i[\bar{V}]).$$

► **Example:** Reconsider

$$F = \left\{ \begin{array}{ll} \langle O \rightarrow M, & 1 \rangle, \\ \langle S \rightarrow \neg M, & 2 \rangle, \\ \langle C \rightarrow S, & 4 \rangle, \\ \langle C \rightarrow M, & 4 \rangle, \\ \langle V \rightarrow \neg M, & 4 \rangle \end{array} \right\}$$

We obtain $P_{F(\bar{V})} = 4VM - 4CM - 4CS + 2SM - OM + 8C + O$.

Simulated Annealing vs. Greedy Satisfiability Testing

GWSat (see ICL-manuscript):

- ▶ only variables in unsatisfied clauses are flipped
- ▶ probability to escape local minimum is not changed over time

N	variant	forced	unforced
100	gsat	76.1%	23.5%
	gwsat	100%	82.2%
	SA	99.8%	70%
200	gsat	75.3%	8.4%
	gwsat	99%	48.7%
	SA	99%	35.9%
300	gsat	77.0%	2.7%
	gwsat	100%	26.3%
	SA	99.4%	16.3%
500	gsat	70.2%	—
	gwsat	100%	—
	SA	95.4%	—

Some Exercises

- ▶ Extend symmetric networks such that a unit becomes active as soon as a model for the corresponding propositional logic formula has been found. Hint: The extension may contain other units than logical threshold ones and the additional connections need not to be symmetric.
- ▶ Will a symmetric network converge to a local minima if the units are updated in parallel? Proof your claim.
- ▶ Why can the constant term occurring in an energy function be neglected? What is the effect of the elimination of the constant term wrt the models of the corresponding propositional logic formulas and the search for global minima?
- ▶ Let $E = V_1V_2 + HV_1V_3 + HV_1 + 2HV_2 - H$
 - (a) Eliminate the hidden variable H .
 - (b) Which propositional logic formula is represented by the energy function obtained from E in (a)?