Introduction to Neural Networks

Johns Hopkins University
Engineering for Professionals Program
605-447/625-438
Dr. Mark Fleischer

Module 6.1: Other Optimization Techniques—Theoretical Foundations of the Simulated Annealing Algorithm

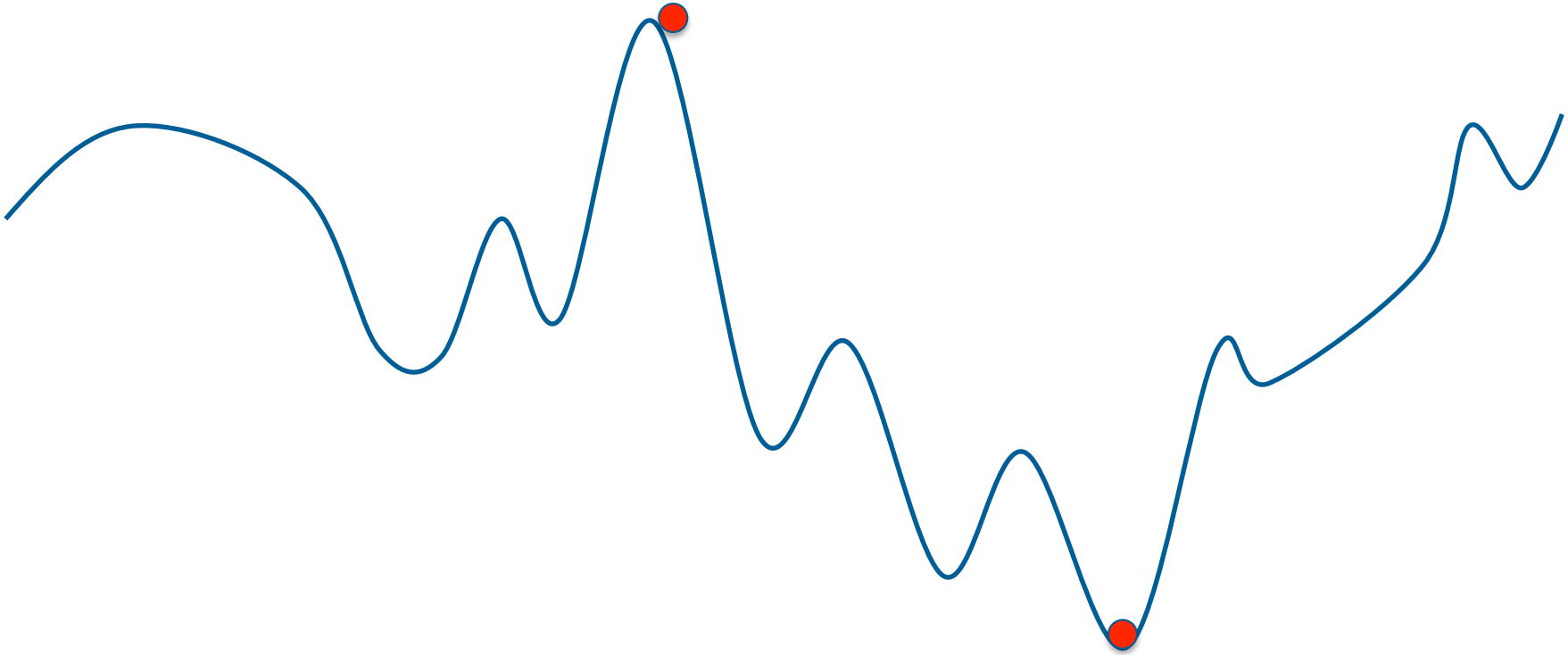
This Sub-Module Covers …

- Describe the history and mathematical foundations of the Simulated Annealing Program.
- Provides a foundation for study of stochastic neural networks which we will cover in later modules.
- Subsequent sub-modules will describe implementation issues and other optimization techniques.

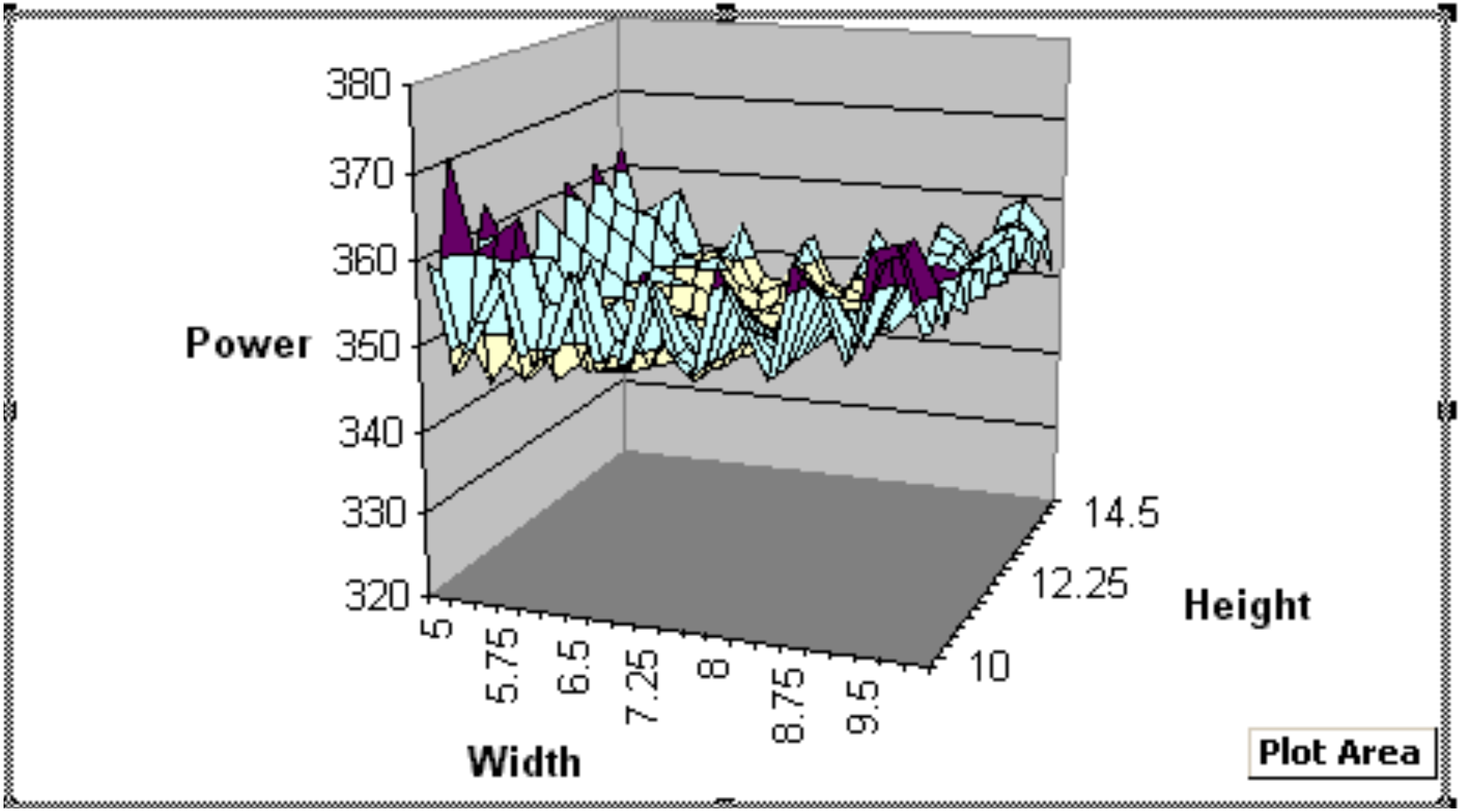
Overview

- Why Global Optimization?
- Metaheuristics
  - Simulated Annealing
  - Genetic Algorithms
  - Others as well
- Research Issues

A Nasty Landscape



A Really Nasty Landscape



What is Visible to an Algorithm?

- Can't see the entire landscape.
  - Doing so equivalent to total enumeration.
- Only a neighborhood is visible. Why?
  - A tractable computational effort required.

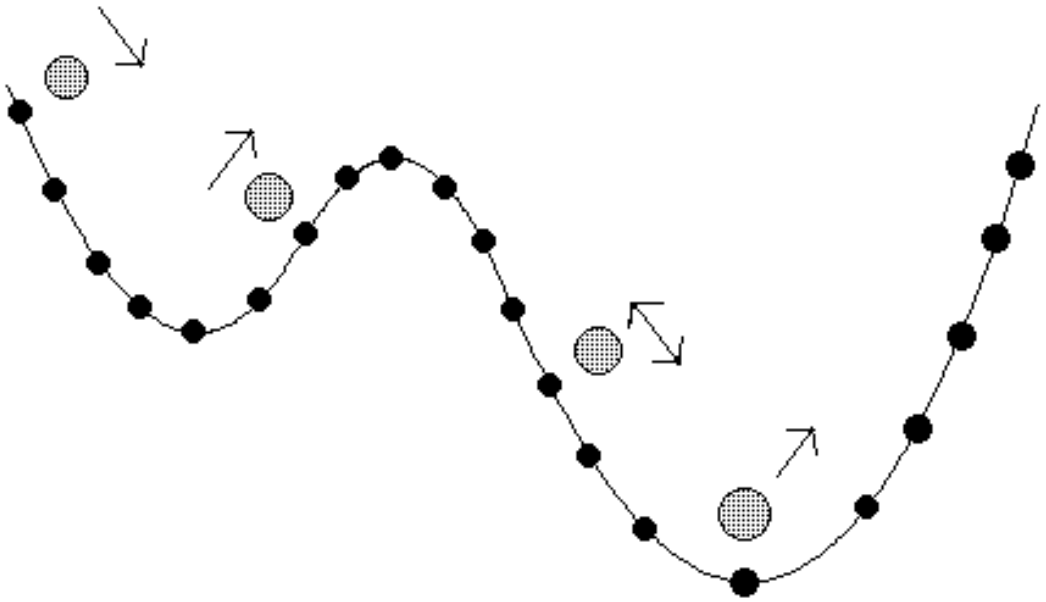
Search Themes

- Directed *e.g.* gradient based.
  - Not applicable to most real-world problems. Why?
- Deterministic.
- Random.
- Random with some determinism.
- Memory dependent – avoid where we have already been.

Search Paradigms
Rules for Path-Finding

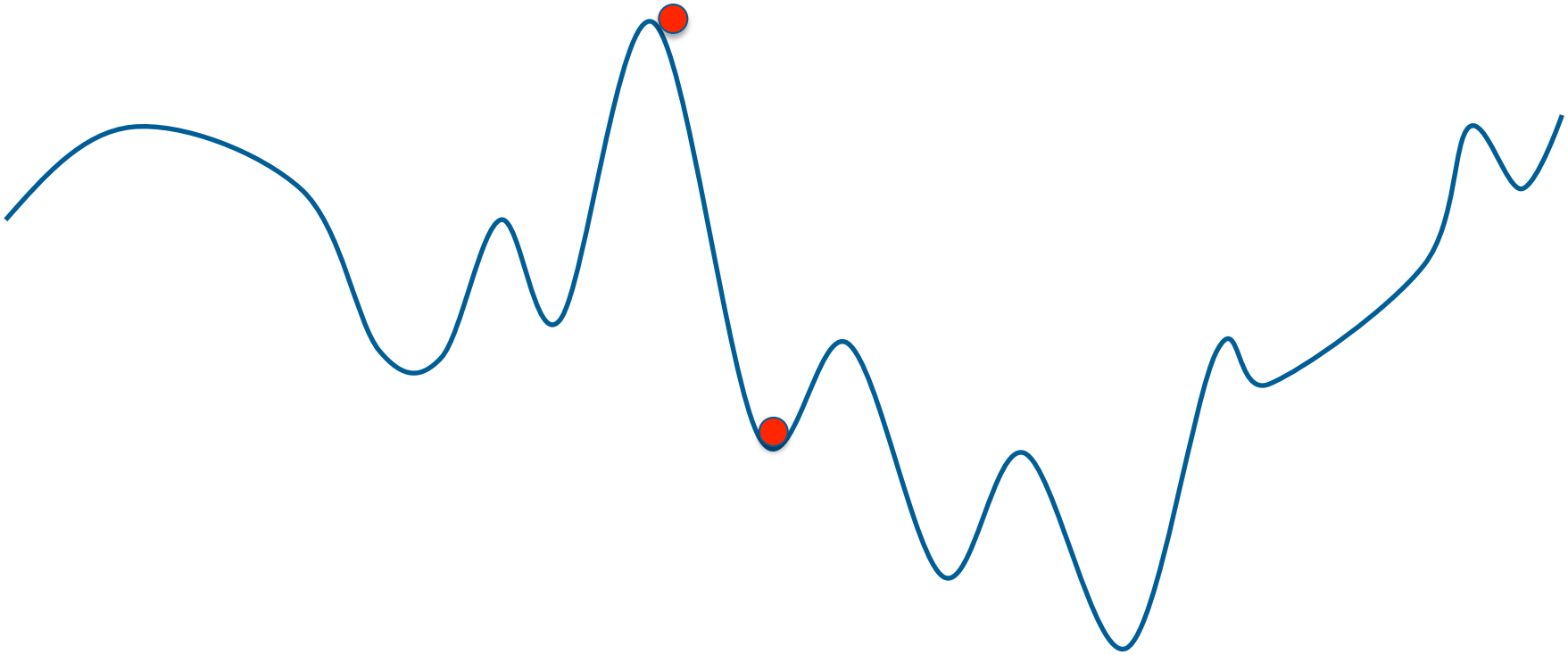
- Thermodynamics
  - Simulated Annealing is based on the laws of thermodynamics
- Biological
  - Genetic algorithms are based on the paradigms of evolution: natural selection, survival of the fittest, etc.
- Common-Sense/Intelligent

Simulated Annealing



Hill Climbing ---> Getting to the Global Optima

A Nasty Landscape



Modeling Thermodynamic Systems

- a large (really large) population (ensemble) of physical entities, *e.g.,* molecules of a gas, liquid, or solid.
- obeys the laws of thermodynamics and statistical mechanics.
- entities subject to collisions.
- entities have mass, position, velocity (kinetic energy)

Statistical Mechanics

Diffusion, entropy, temperature

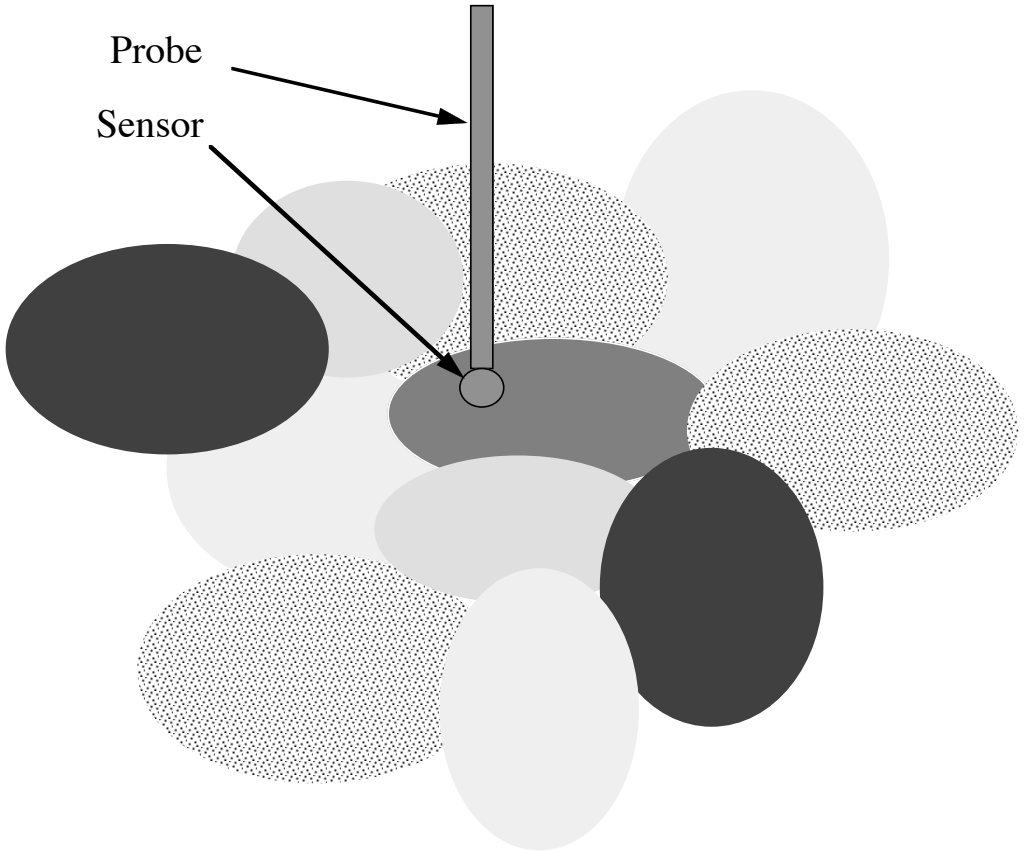
Thermodynamics/Statistical Mechanics

- Particles change energy levels sporadically because of collisions.
- Energy level of system proportional to the **mean distance** between particles.
- Want to facilitate a simulation of such a system!

Thermodynamics/Statistical Mechanics

- Kinetic energy of particles is virtually a **continuous function** because of the number of particles involved.
- Define a **finite** set of states *i* each with an associated energy level E_i .
- Consider the evolution of "states" at a given temperature.
  - Some regions become more/less dense randomly over time.
  - Some regions gain/lose energy over time.

Gas with Varying Densities



Temperature at sensor fluctuates.
Energy states change over time.

Definition of Terms

t = an estimate of the mean kinetic energy of a substance, gas, etc.

$\pi_i(t)$ the stationary probability of energy state *i* at temperature *t.*

E_i the energy level of partition *i*.

Modeling the System of Particles

- An ensemble of particles seeks the maximum state of disorder.
- Statistical Mechanics translates this to a probability distribution that maximizes entropy.

What is Entropy?

A measure of the total ‘uncertainty’ associated with an ensemble of possibilities.

A single, non-negative scalar value that is associated with uncertainty.

We want it to encompass many different possible events.

We want it to be additive.

$$S_i \propto \frac{1}{p_i}, \quad S_{ij} \propto S_i + S_j \qquad S_i = k \log \frac{1}{p_i} = -k \log p_i$$

$$S_{ij} = k \log \left(\frac{1}{p_i p_j} \right) = k \log \frac{1}{p_i} + k \log \frac{1}{p_j}$$

What is 'Uncertainty'

A	B
0.01	0.99

Uncertainty = Expected Surprisal

$$U = p_A S_A + p_B S_B$$

$$U = -k \sum_i p_i \log p_i$$

A	B
0.5	0.5

An NLP

$$\text{Max} \quad -k \sum_{i=1}^{n} \pi_i(t) \log \pi_i(t)$$

$$\text{s.t.} \quad \sum_{i=1}^{n} \pi_i(t) E_i = k\,t$$

$$\sum_{i=1}^{n} \pi_i(t) = 1$$

$$t, \pi_i(t), E_i \geq 0 \ \forall i$$

The Boltzmann Distribution

$$\pi_i(t) = \frac{e^{-E_i/kt}}{B(t)}$$

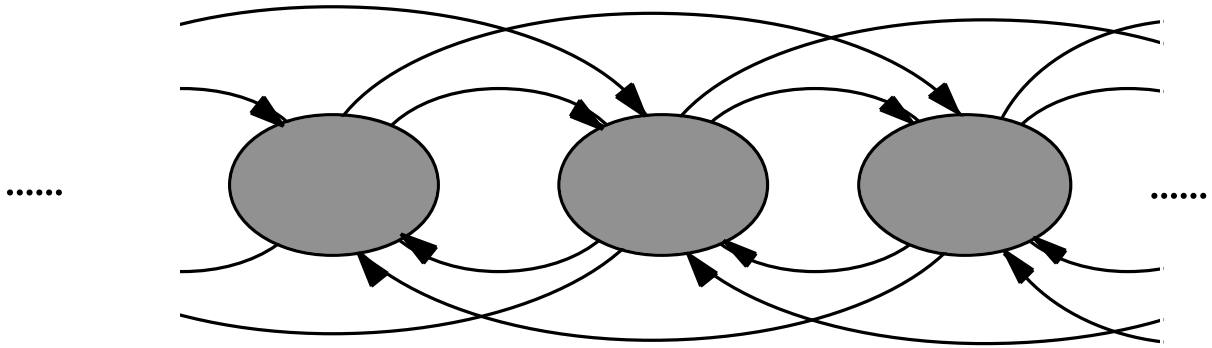
$$B(t) \equiv \sum_{i=1}^{n} e^{-E_i/kt}$$

State Transitions

(Metropolis, *et al.* 1953)

Detailed balance requires that

$$\pi_i(t)p_{ij}(t) = \pi_j(t)p_{ji}(t)$$

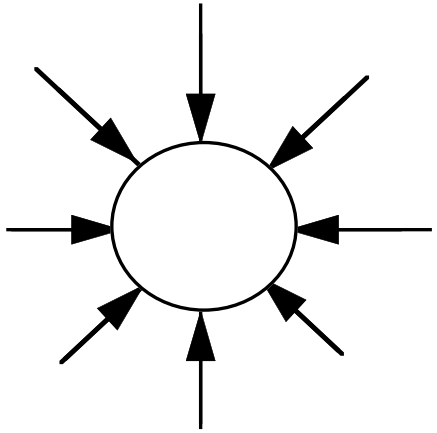


State Transitions

(Metropolis, *et al.* 1953)

Global balance requires that

$$\pi_i(t) = \sum_j \pi_j(t) p_{ji}(t)$$



Equations of State

(Metropolis Acceptance Criterion)

- Regions (states) move spontaneously to different energy levels.
- Given current energy level *i*,
  - if $\Delta E = E_j - E_i \leq 0$ then that region moves to the new energy level *j*.
  - if $\Delta E = E_j - E_i > 0$ then that region moves to the new energy level *j* with the following acceptance probability

$$\Pr\{\text{Accept state } j\} = e^{-\Delta E / t}$$

Transition Probability

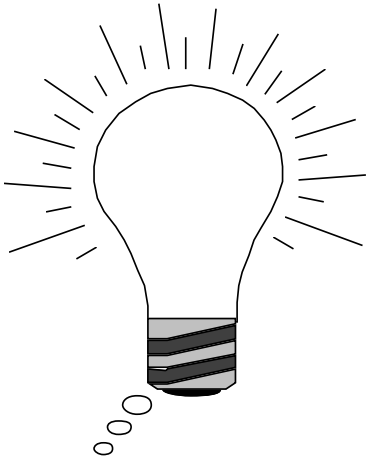
Balance conditions lead to the form for the transition probabilities:

$$p_{ij}(t) = \begin{cases} G_{ji} e^{-\Delta E^{+} / t} & \Delta E_{ji} > 0, j \neq i \\ G_{ji} & \Delta E_{ji} \leq 0, j \neq i \\ 1 - \sum_{k} p_{ik}(t) & j = i \\ 0 & \text{otherwise} \end{cases}$$

Kirkpatrick, et al. (1983), Cerny (1985)

The annealing process lowers
the energy content of a solid.

Energy content of a solid ⇔ Objective Function of an optimization problem



The Simulated Annealing Algorithm

Mathematical Properties of SA

Convergence in Probability (stationary probability)

$$\lim_{k \to \infty} \Pr\{S_k \in S_{\text{OPT}}\} = 1$$

Convergence in Distribution

$$\lim_{t \to 0} \pi_i(t) = \begin{cases} \dfrac{1}{|S_{\text{OPT}}|} & \text{if } i \text{ is in the set of optima} \\ 0 & \text{otherwise} \end{cases}$$

Summarizing Simulated Annealing

- The Simulated Annealing is based on modeling a thermodynamic system.
  - As in thermodynamics, we model large-scale behavior instead of modeling the dynamics of each involved entity.
  - Metropolis formulated 'Equations of State' as a non-linear mathematical program.
  - Led to the mathematical form of transition probabilities between states.
- Provides equivalent state probabilities as in thermodynamic systems.
- Kirkpatrick and Cerny found a way to apply these equations of state to simulate the annealing process.

Conclusion

- More metaphors exist
  - Tabu Search, Adaptive Memory Programming and Genetic Algorithms
- Active area of research
- Many overlooked/ignored possibilities
- Balance accuracy, search time
- Structure, properties, computer architecture.

Introduction to Neural Networks

Johns Hopkins University
Engineering for Professionals Program
605-447/625-438
Dr. Mark Fleischer

Module 6.2: Implementation of the Simulated Annealing Algorithm

This Sub-Module …

- describes approaches for implementing the Simulated Annealing Program.
- Implementation depends on the type of problem we're trying to solve.
  - We'll look at Combinatorial Optimization Problems (COPs)
  - Continuous variable problems.
  - Touch on approaches for parallizing SA.

Metropolis Acceptance Criterion

States i and j

Objective Function Values f_i and f_j

with $\Delta f_{ji} = f_j - f_i$

$$\Pr\{\text{Accept Cand. } J\} = \begin{cases} e^{-\Delta f_{ji}^{+}/t_k} & \Delta f_{ji} > 0, j \in N(i) \\ 1 & \Delta f_{ji} \le 0, j \in N(i) \\ 0 & \text{otherwise} \end{cases}$$

The Simulated Annealing Algorithm

1. Select a new state with a new objective function value.
2. Calculate the value of Δf.
3. If $\Delta f \leq 0$ (for a minimization problem) then the new state becomes the current state. If $\Delta f > 0$ then the new state becomes the current state via the Metropolis Acceptance Criterion. If state J is not accepted, keep state I as the current state.
4. Lower temperature and goto 1.

Requirements for SA

1. Define an appropriate set of states.
2. Define a suitable neighborhood structure.

   This entails the candidate generation mechanism.
3. Define a suitable cost/objective function.

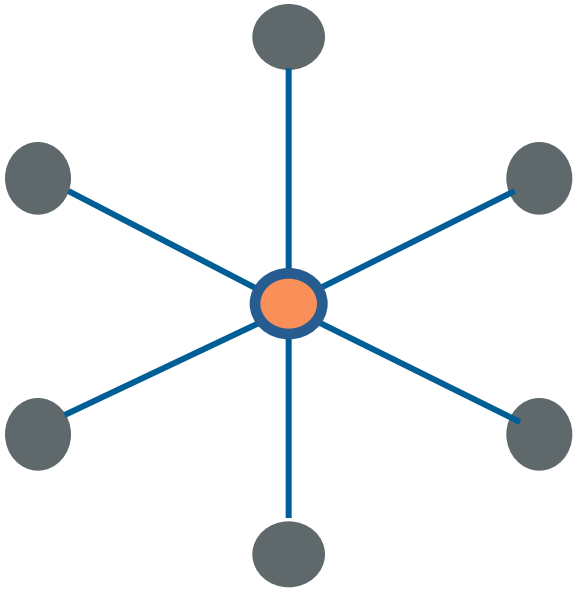
   This entails a method for calculating the change in objective function value.
4. Define a cooling schedule.
5. Define stopping criteria.

Cooling Schedules

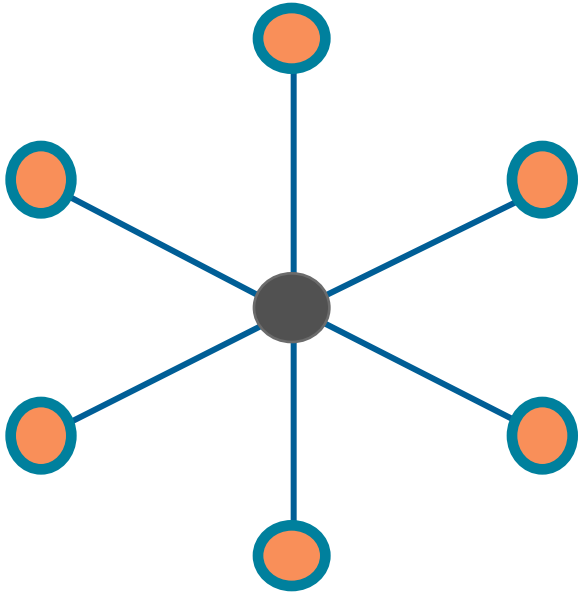
$$t_k = \frac{\gamma}{\log(c + k)}$$

$$t_k = \frac{\gamma}{c + k}$$

Combinatorial Optimization Problems (COPs)

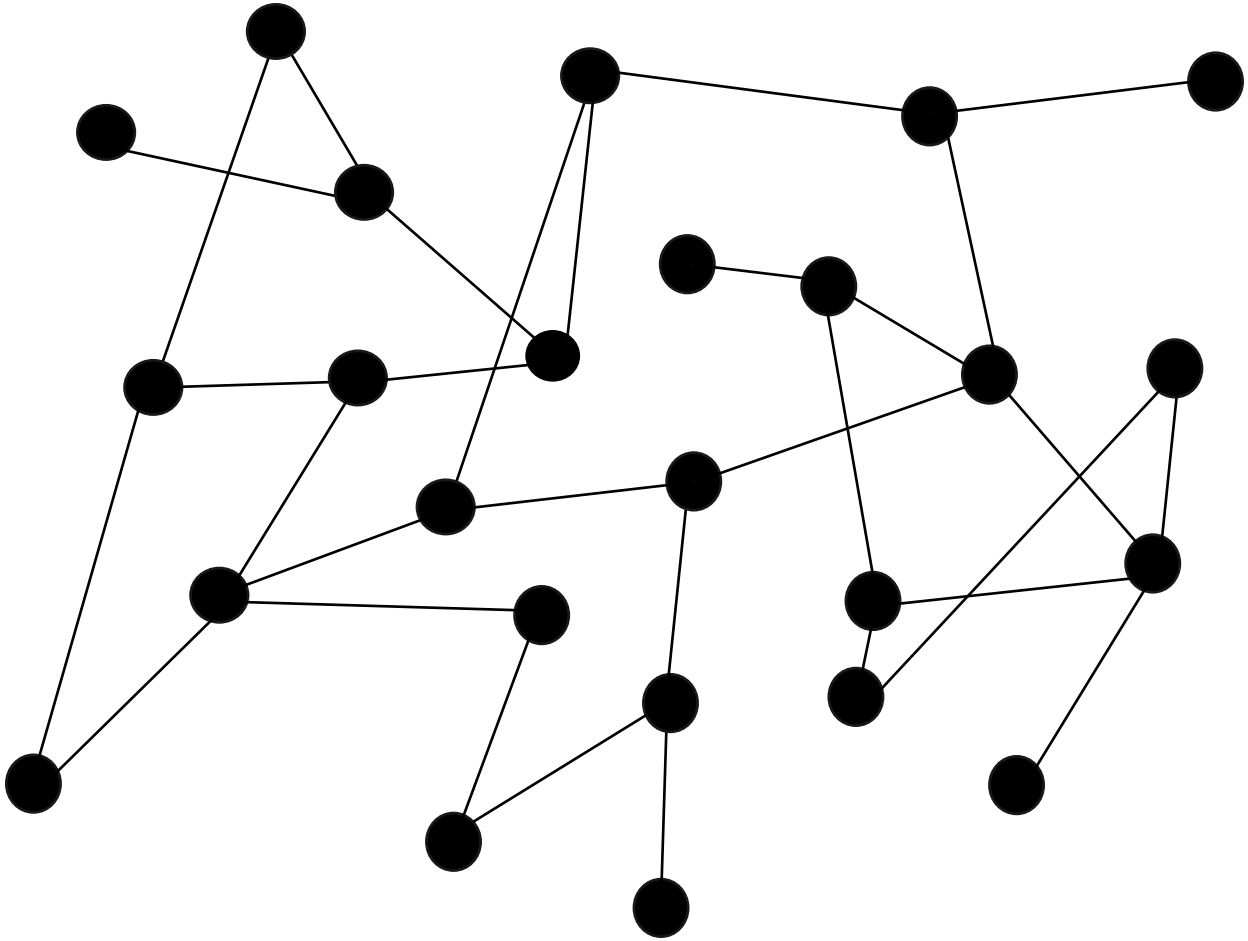


Minimum Vertex Cover



Maximum Independent Set

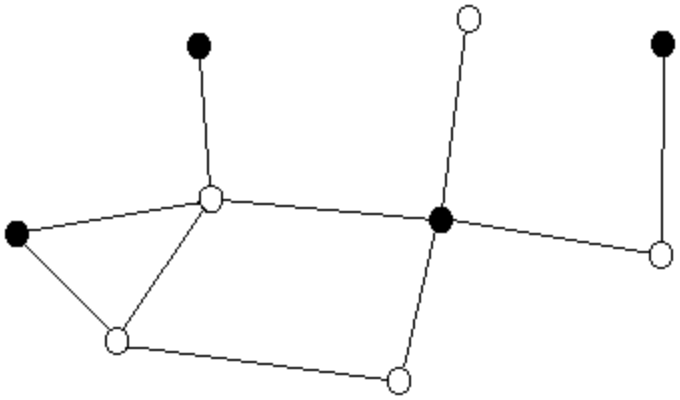
Not Always So Obvious . . .



Adding Insult to Injury. . .

$2^{1,000,000}$? Yikes!

Implementation Issues

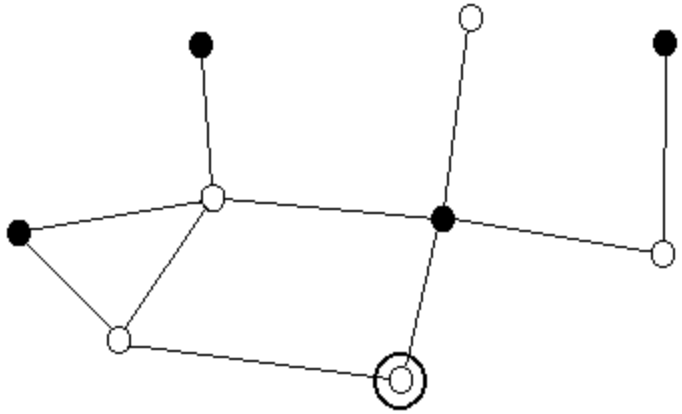


A good objective function is simply
the number of nodes in the current set.

Current Solution = 4
Maximum Set Size = 5

How do we get to better solutions
without doing a lot of work?

Select a node, any node…



But this violates independent set constraints!
Use some penalty function.
Penalty function should decrease
objective function value
and
be based on those elements which violate constraints.

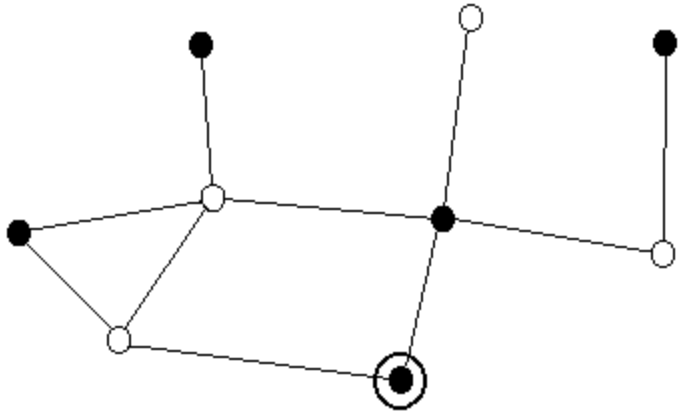
Objective Functions

What are the elements which violate constraints?

edges

f(G) = number of nodes – g(edges)

New Combination of Nodes, New Objective Function Value

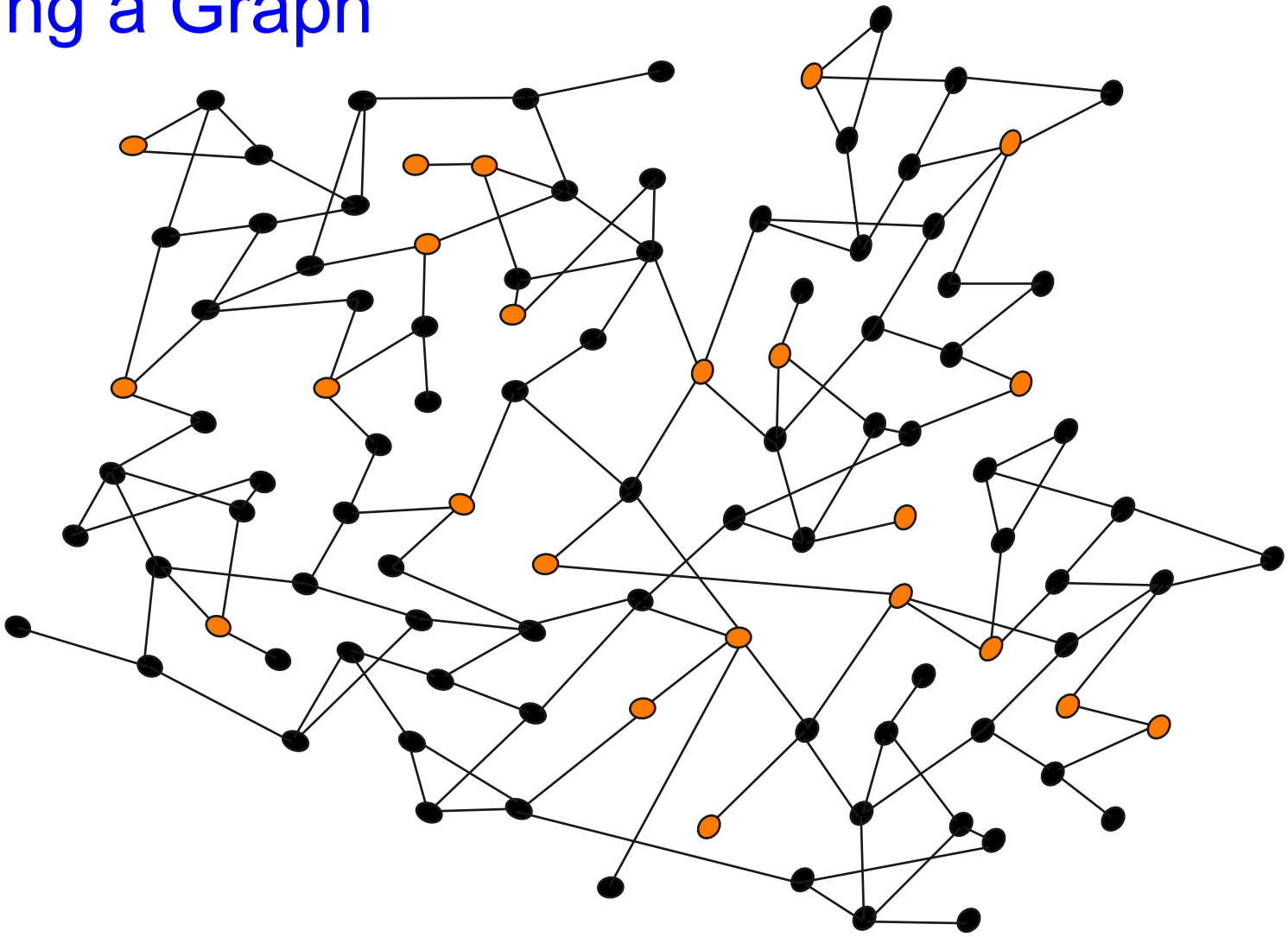


Current Objective Function Value:

From $f_{\text{indep set}}(V',G) = v' - \lambda E_1(V') = 4 - 0$

to $f_{\text{indep set}}(V',G) = v' - \lambda E_1(V') = 5 - \lambda 1$

Annealing a Graph



E.g. the Stationary Probability

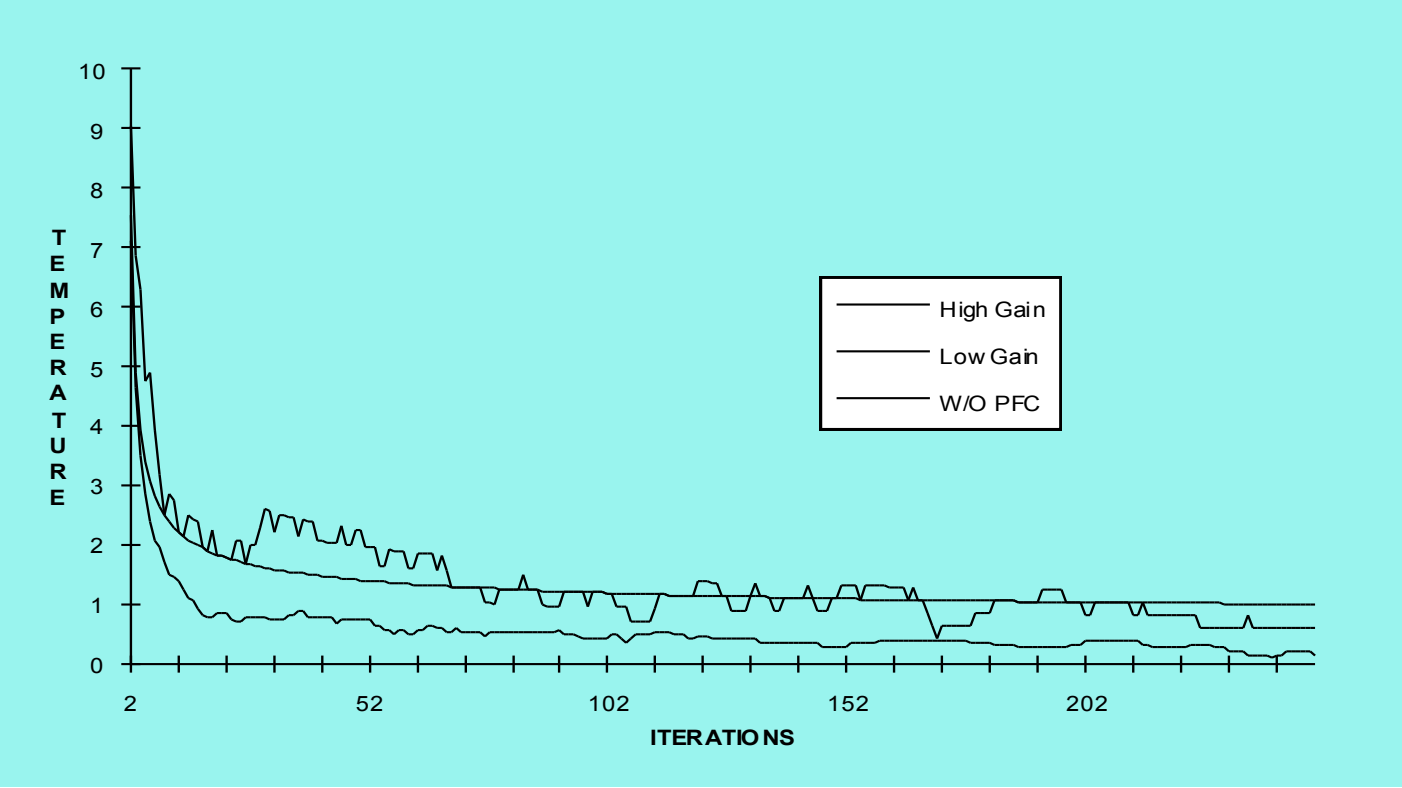
$$\pi_i(t) = \frac{e^{-f_i/t}}{\sum_{i=1}^{s} e^{-f_i/t}}.$$

Boltzmann Distribution Function

Joint Distribution

$$
\begin{aligned}
\pi_{i_1 \cdots i_p}(t) &= \prod_{m=1}^{p} \pi_{i_m}(t) \\
&= \prod_{m=1}^{p} \frac{e^{-f_{i_m}/t}}{\sum_{i_m=1}^{s} e^{-f_{i_m}/t}} \\
&= \frac{e^{-\left(\sum_{m=1}^{p} f_{i_m}\right)/t}}{\prod_{m=1}^{p} \sum_{i_m=1}^{s} e^{-f_{i_m}/t}} \\
&= \frac{e^{-f_{i_1 \cdots i_p}/t}}{\sum_{i_1 \cdots i_p}^{s^p} e^{-f_{i_1 \cdots i_p}/t}}
\end{aligned}
$$

Non-Monotonic Cooling



Journal of Heuristics Vol. 1 No. 2 p.245

Experimental Results and Observations

Number of Processors	Gain Setting	Objective Function Values With PFC $\overline{F}_{min}$	With PFC s^2_F	No PFC $\overline{G}_{min}$	No PFC s^2_G	z-value
5	1	0.683	0.303	0.541	0.057	1.290
5	5	0.212	0.057	0.536	0.536	-2.300
5	10	0.232	0.090	0.541	0.057	-4.419
10	1	0.488	0.560	0.388	0.040	0.709
10	5	0.130	0.027	0.324	0.047	-3.899
10	10	0.110	0.020	0.438	0.062	-6.261

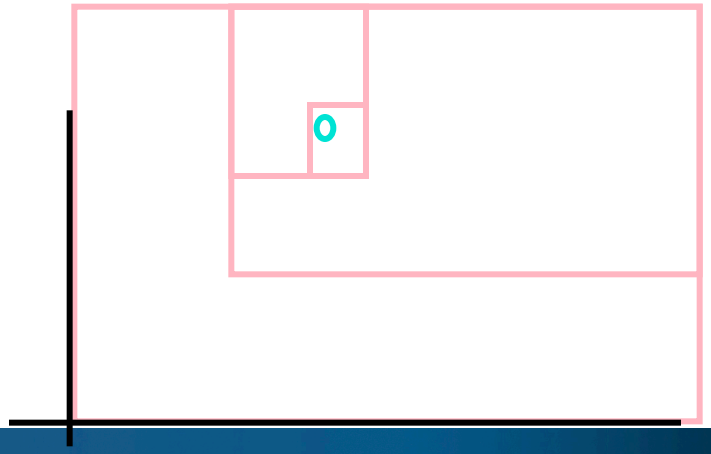
Table 5.1 : Efficacy of PFC by Candidate Generation

What About
Continuous Variable Problems?

- Picking candidate solutions is main issue.
- SA doesn't work well on continuous variable problems.

Recursive Intensification

- Accelerates convergence to optima, experimentally.
- Increases probability of never converging to the optima.



Introduction to Neural Networks

Johns Hopkins University
Engineering for Professionals Program
605-447/625-438
Dr. Mark Fleischer

Module 6.3: Genetic Algorithms

This Sub-Module …

- Briefly describes some of the history behind the development of genetic algorithms (GAs).
- Describes the basic ideas behind genetic algorithms.
- Implementation generally involves a few types of ‘operators’ inspired by biological systems

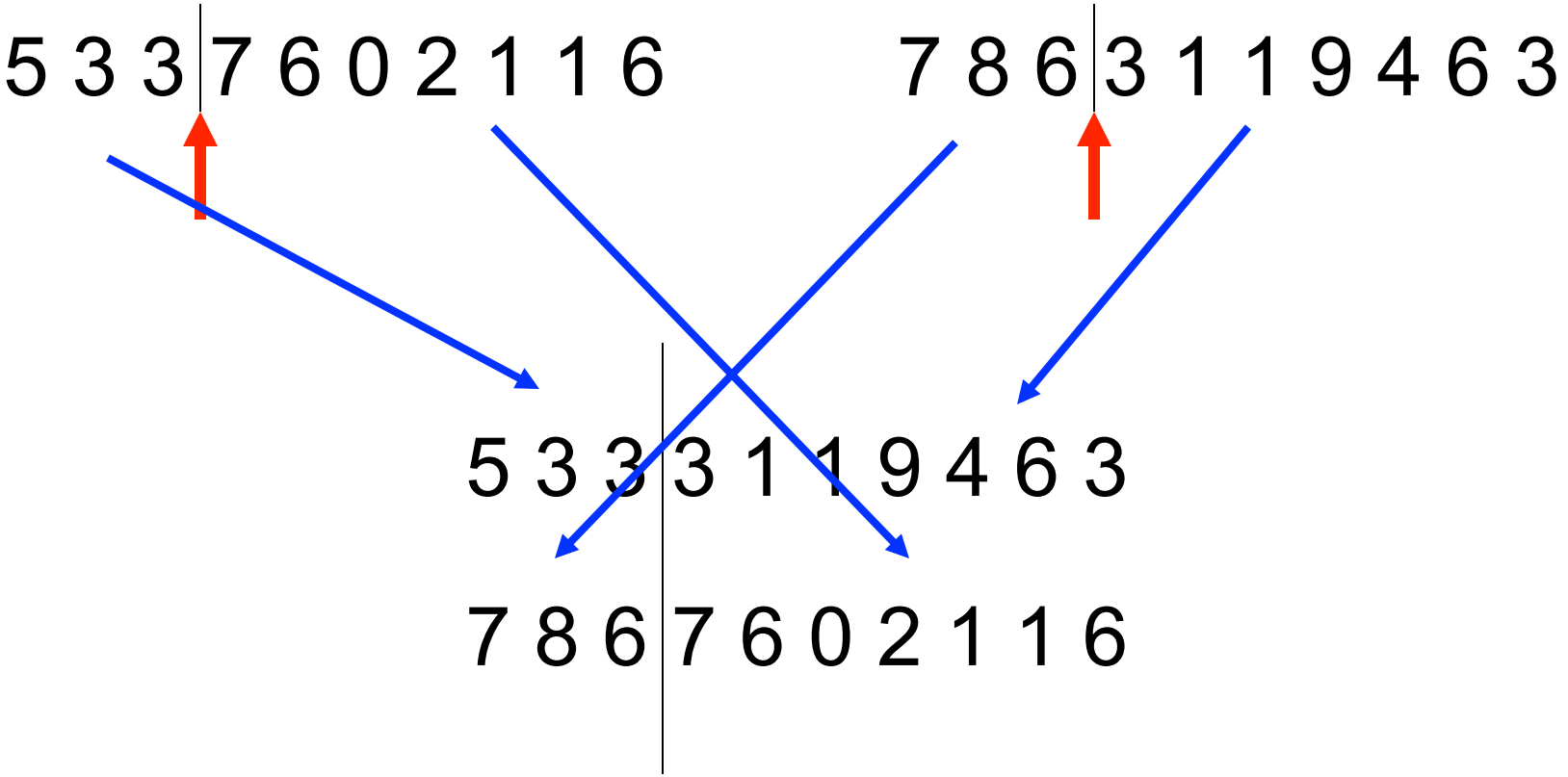
Genetic Algorithms

- Another metaphor for random search.
- Based on biological evolution.
- Analogies to biological mechanisms of sexual selection.
  - Chromosomes/Genes/DNA: solution schema
  - Combining DNA: crossover operation
  - Mutation: mutation operator
  - Natural Selection: fitness function

GA -- History

- Invented by Holland (1960s).
- Schema theorem developed in 1975.
- Many implementations, possibilities, and applications.
- One of the most useful global optimization technique.

Crossover Operator



Two new “children”(solutions).

Mutation Operator

5 3 3 3 1 1 9 4 6 3

5 3 3 3 1 1 5 4 6 3

Natural Selection via Fitness Operators

- Once new children are created, a *fitness operator* is applied to all population members.
- Those members that rank high enough in fitness value, are kept for another generation of crossover and mutation operations.
- Those members that do not rank sufficiently high are "deleted".

Implementation Issues

- Crossover operators
  - Many different operators possible
  - *E.g.,* 3 5 1 1 | 6 7 9 9 | 0 5 5 3 2
- Mutation operators
  - Different probability values
  - Different functional dependencies
- Fitness operators
  - Different scaling functions

Conclusion

- We've looked at two meta-heuristics inspired by nature.
  - Simulated Annealing: Thermodynamics and Statistical Mechanics
  - Genetic Algorithms: Evolution and Natural Selection
- More metaphors exist
- Many many research papers about on these topics.
- They can be applied to neural networks.