CSE543T: Algorithms for Nonlinear Optimization

Lecture 2: Unconstrained Optimization – Definitions and Conditions

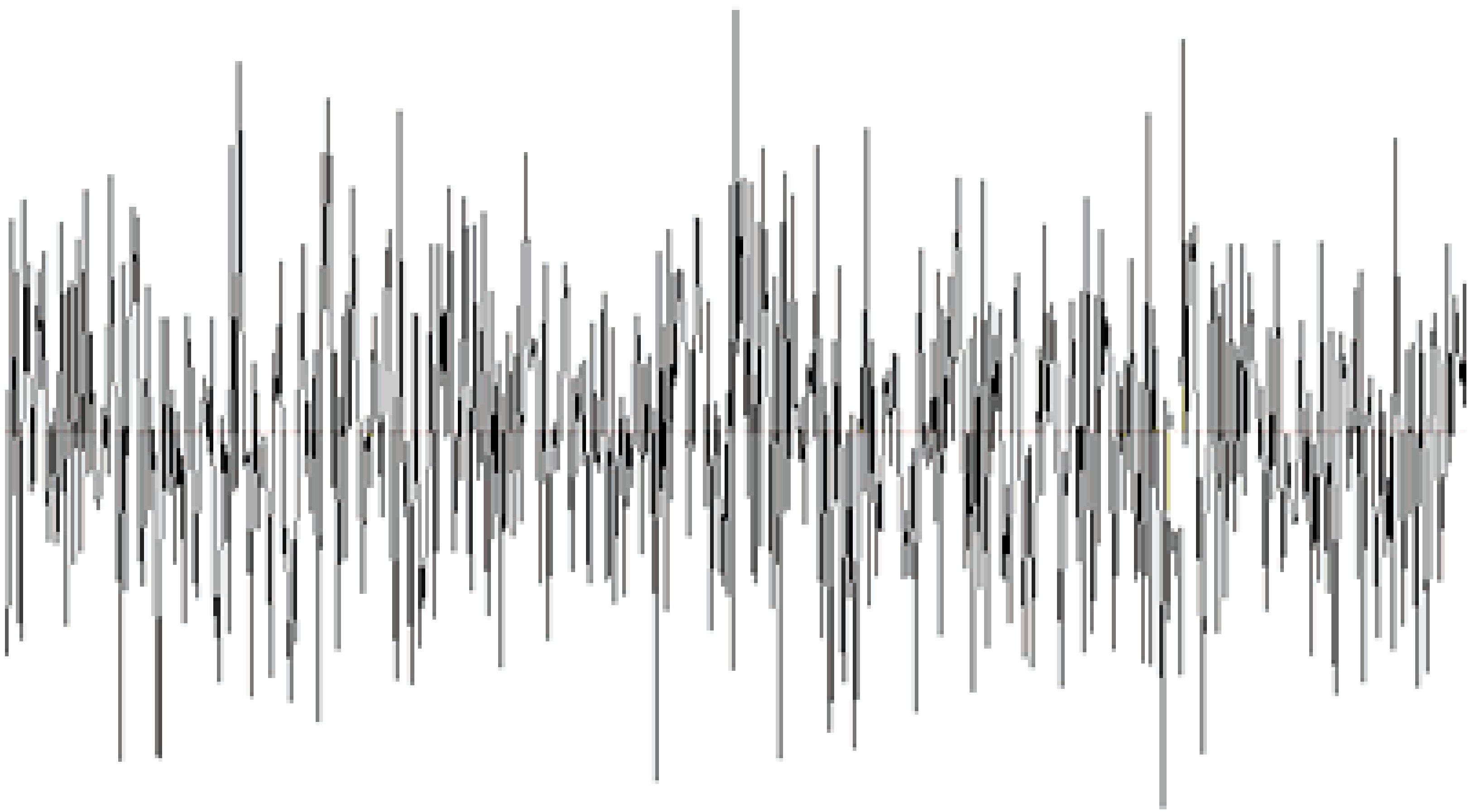
Unconstrained Optimization

Minimize f(x)
where x is defined in $\mathcal{X}$

A blackbox problem

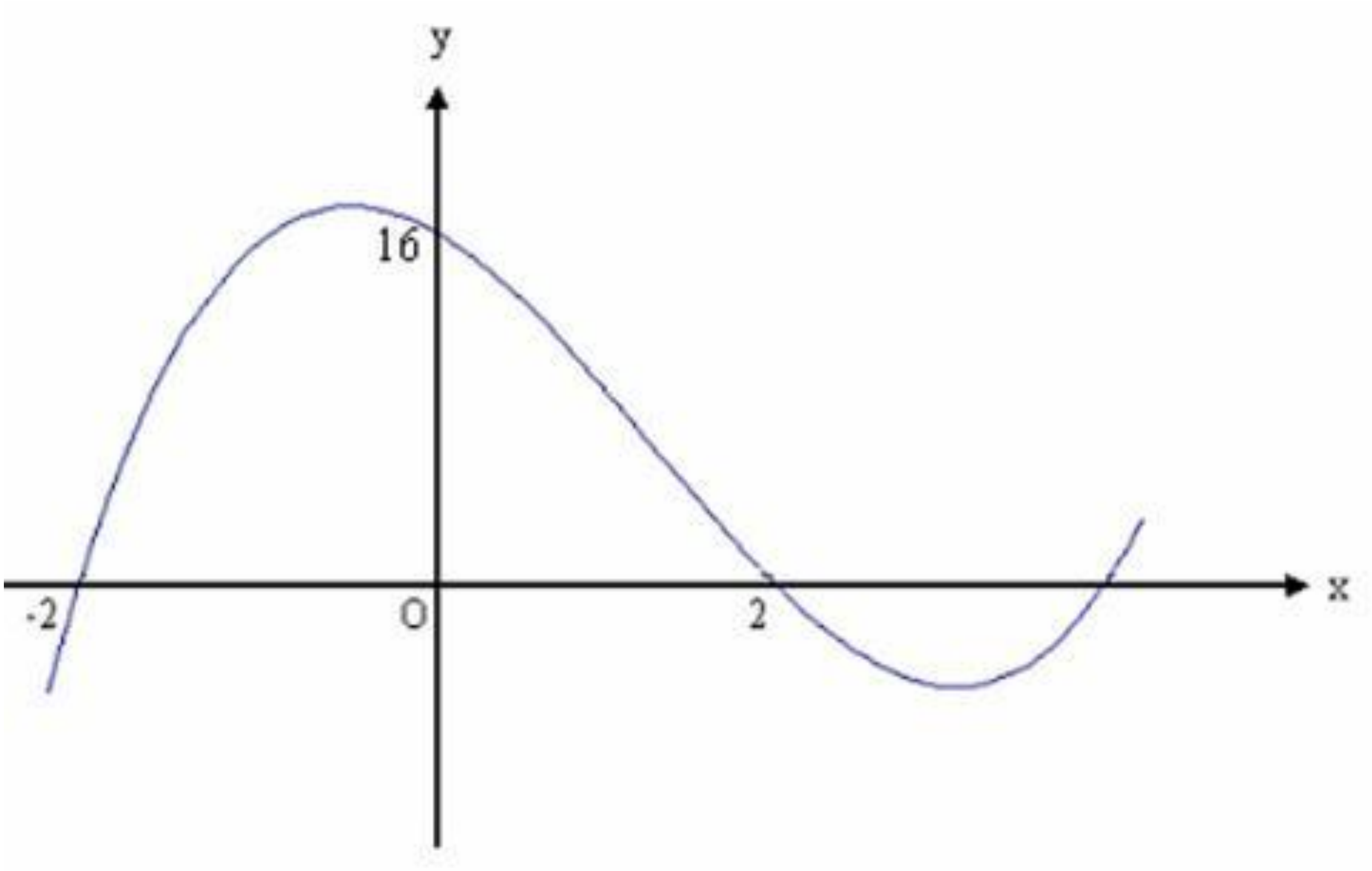
- Find x that minimizes E(x)
  - E(x) is a blackbox
  - x is defined over a finite discrete set X
  - there are n elements in X
- Are there more efficient ways than simple enumeration?
  - Deterministic vs. expected running time

Noise

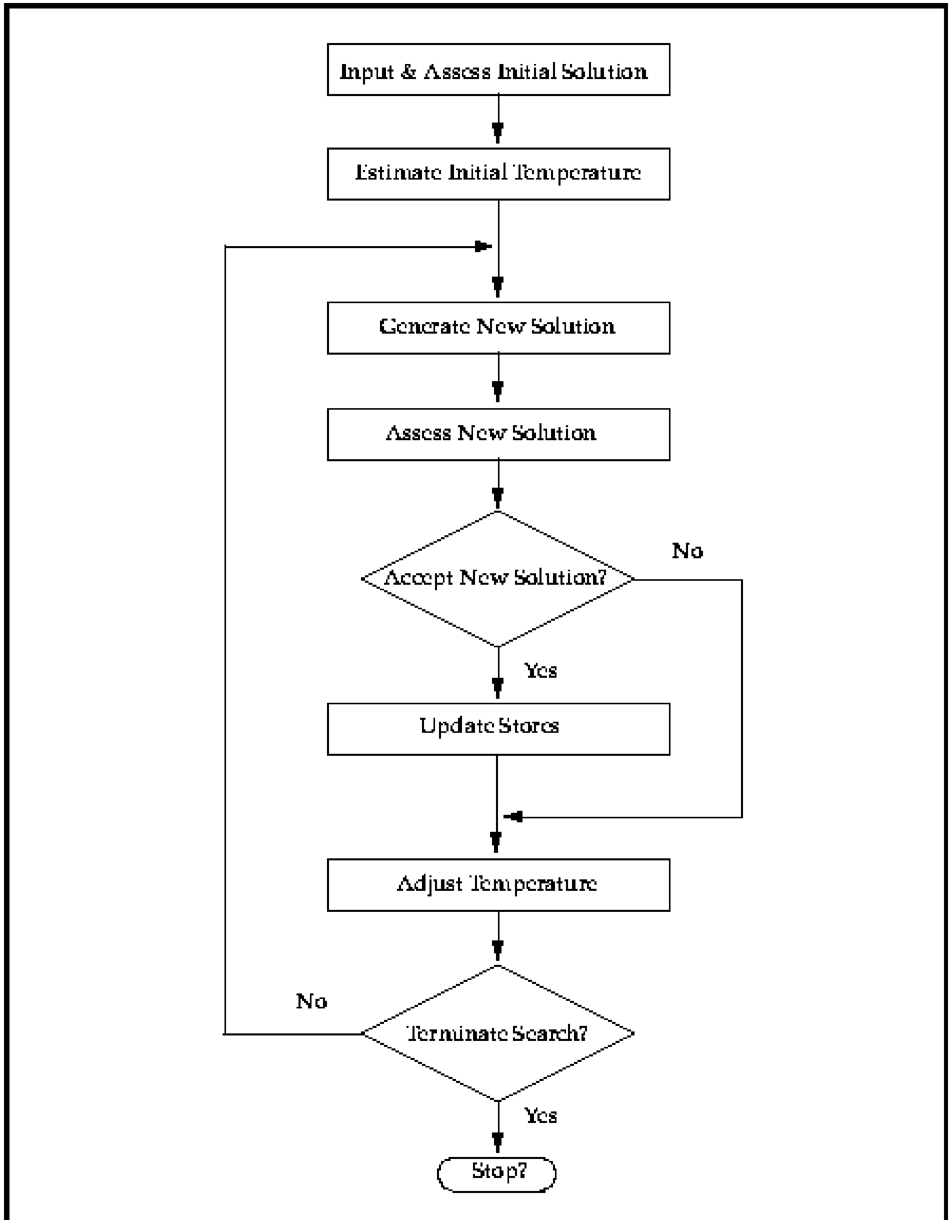


Highly nonlinear function

Nonlinear function



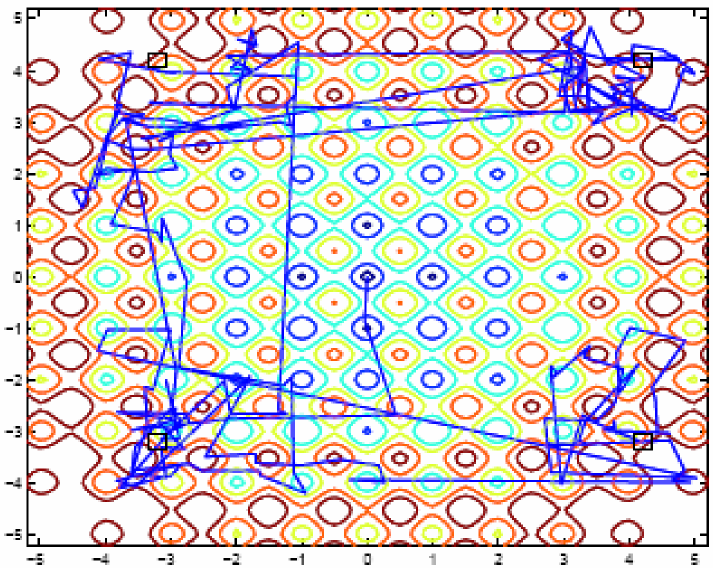
Simulated Annealing



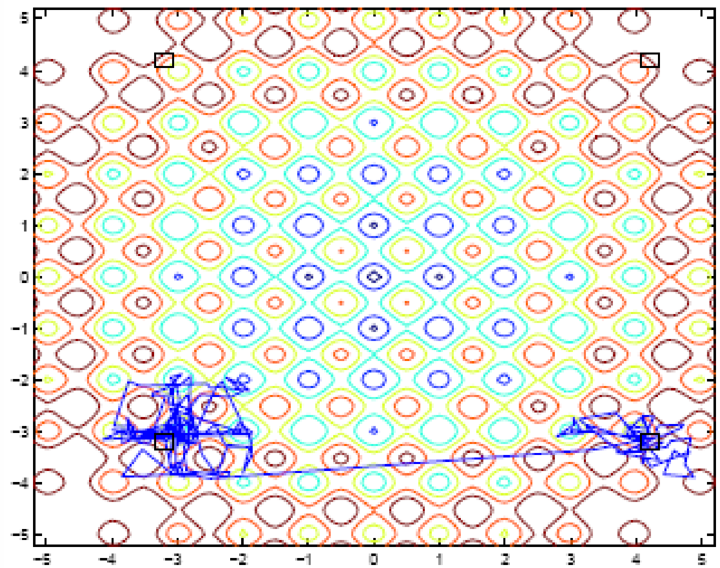
Simulated Annealing

Initialise x to x_0 and T to T_0

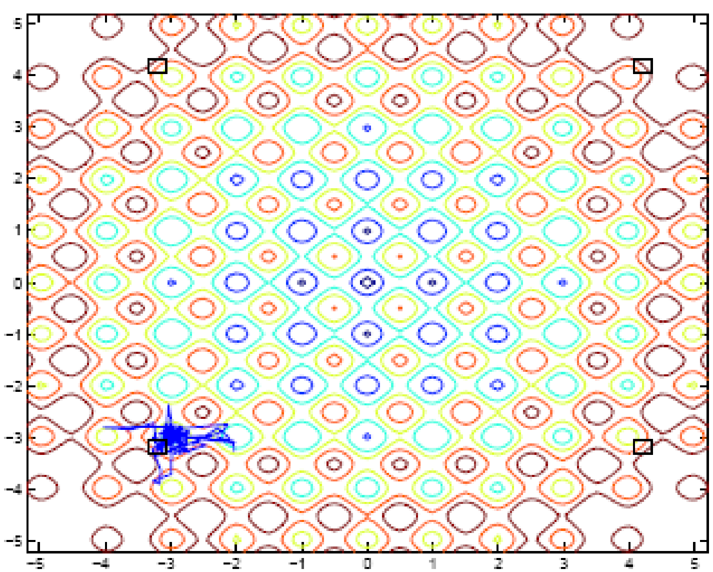
loop — Cooling
  loop — Local search
    Derive a neighbour, $x\prime$, of x
    $\Delta E := E(x\prime) - E(x)$
    if $\Delta E < 0$
    then $x := x\prime$
    else derive random number $r \in [0, 1]$
      if $r < e^{-\frac{\Delta E}{T}}$
      then $x := x\prime$
      end if
    end if
  end loop — Local search
  exit when the goal is reached or a pre-defined stopping condition is satisfied
  $T := C(T)$
end loop — Cooling



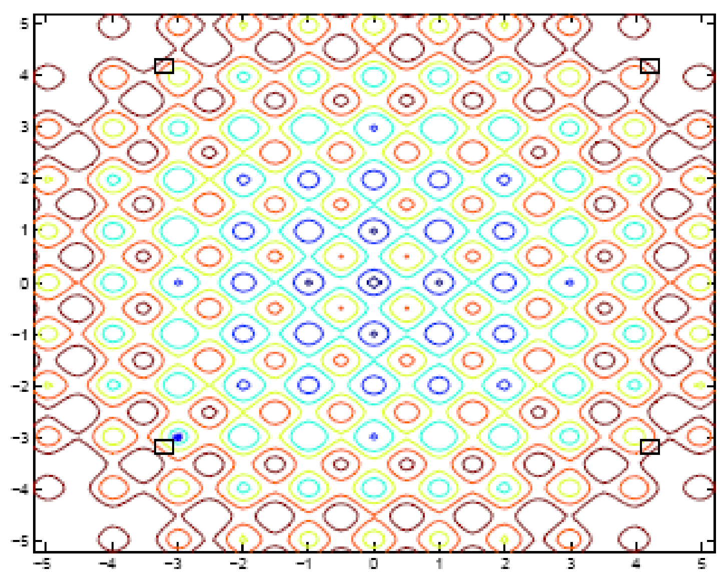
(a) T = 20



(b) T = 10.24

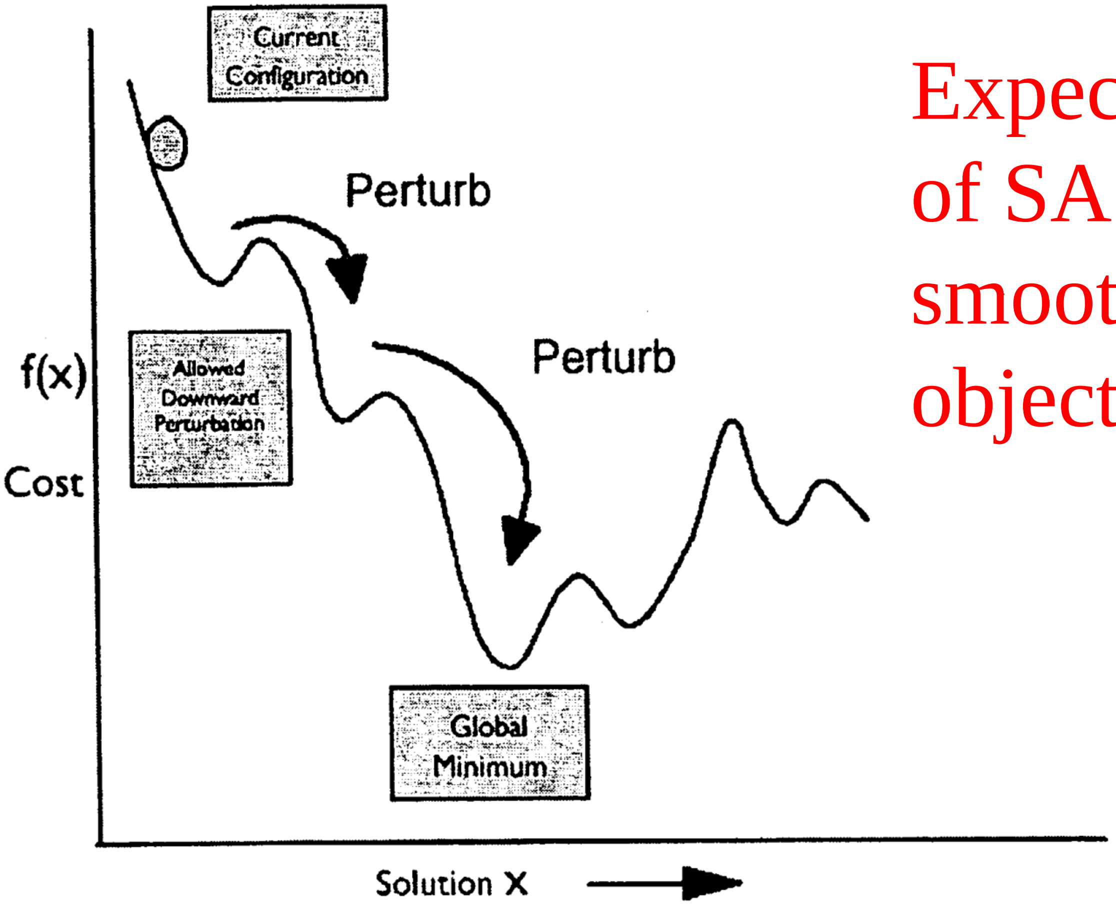


(c) T = 8.192



(d) T = 0.45

Simulated Annealing

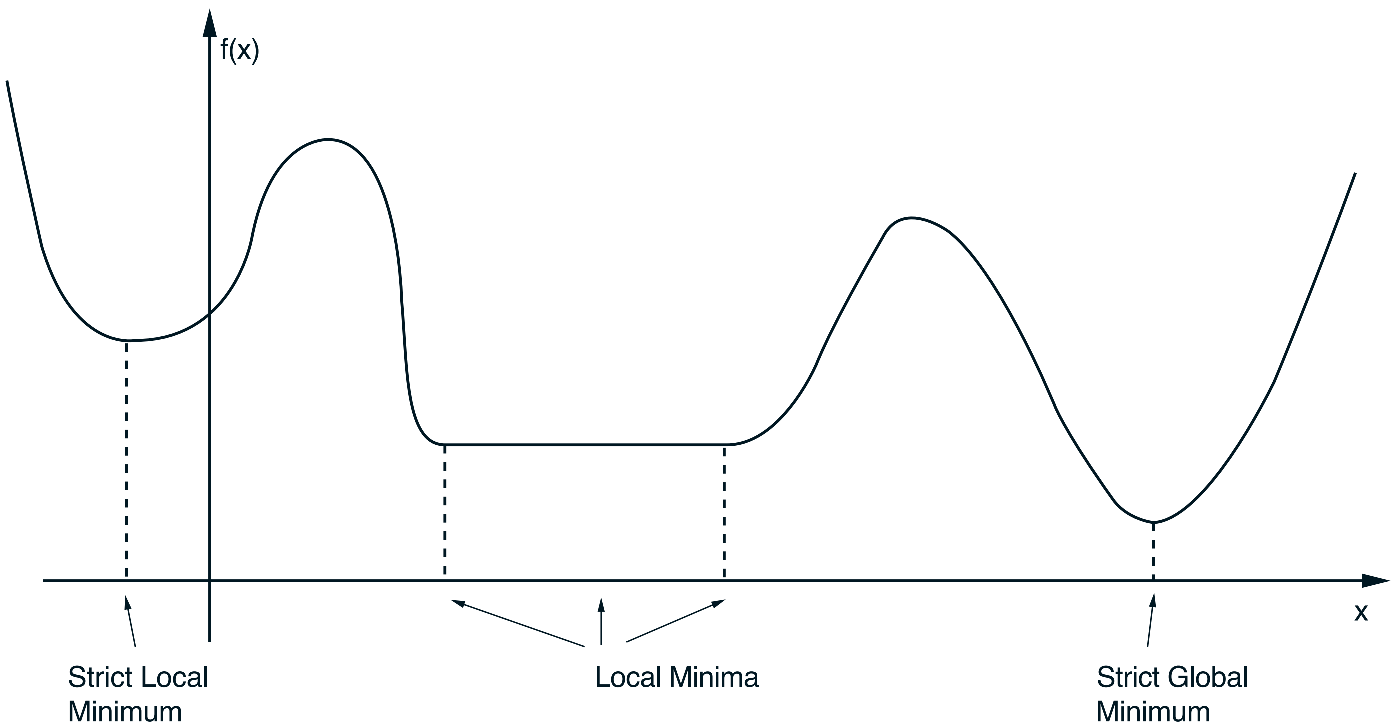


Expected complexity of SA depends on the smoothness of the objective function

Stochastic search algorithms

- Simulated annealing, genetic algorithms, ant colony algorithms, controlled random walk, etc.
- Much better performance than random sampling
  - Why?
- “slow” but valuable
  - Achieves global optimality
  - Very relaxed assumptions

LOCAL AND GLOBAL MINIMA



Unconstrained local and global minima in one dimension.

Why do we look for local optima

- Sometimes,
  - local optimum == global optimum
- If you can enumerate all the local optima, you can find the global optima

NECESSARY CONDITIONS FOR A LOCAL MIN

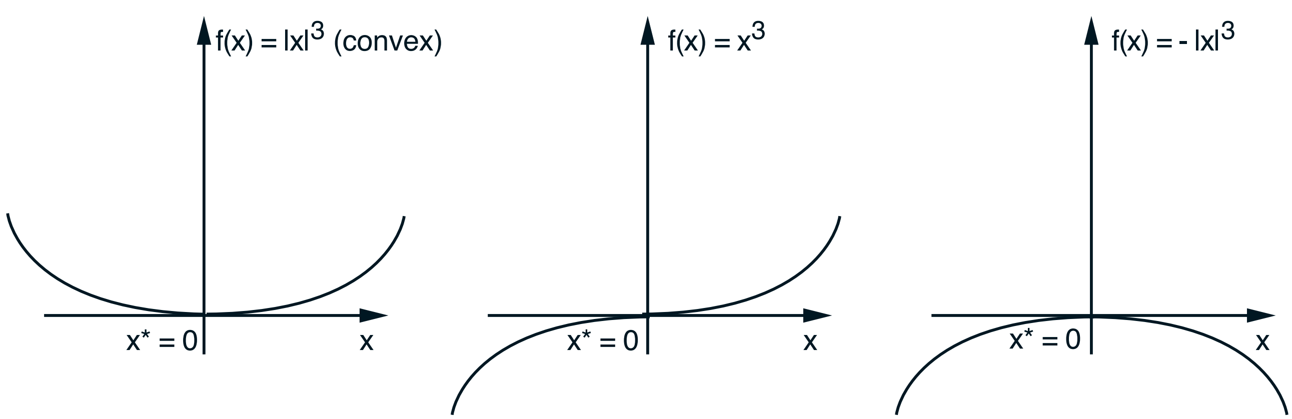
- **1st order condition:** Zero slope at a local minimum x^*

$$\nabla f(x^*) = 0$$

- **2nd order condition:** Nonnegative curvature at a local minimum x^*

$$\nabla^2 f(x^*) : \text{ Positive Semidefinite}$$

- There may exist points that satisfy the 1st and 2nd order conditions but are not local minima

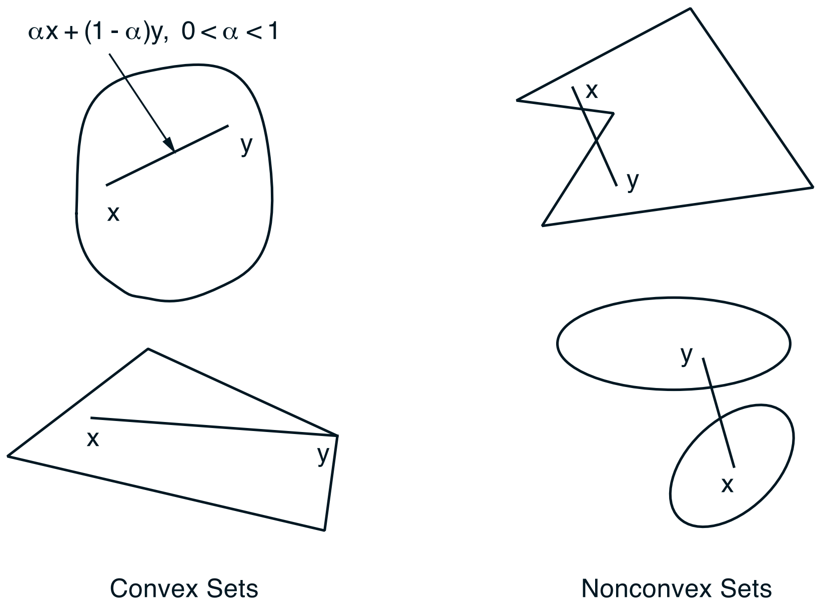


First and second order necessary optimality conditions for functions of one variable.

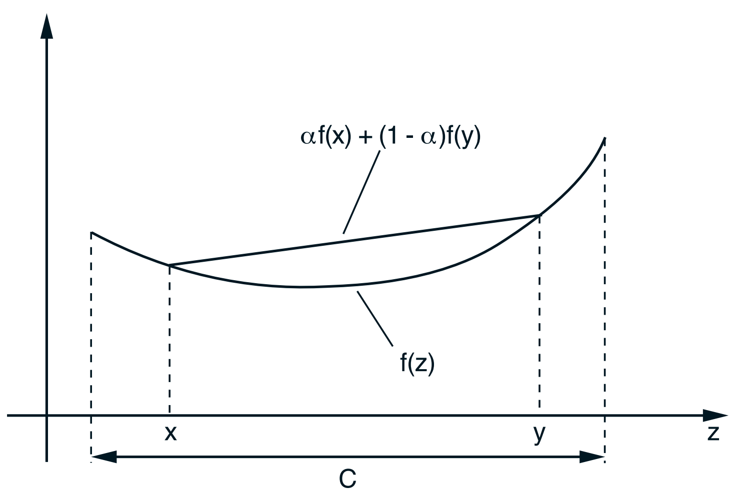
Why do we need optimality conditions

- Necessary vs. sufficient
- Ways that do not work
  - Enumerate and check
  - Enumerate and compare
  - Solving $\partial f = 0$ is nontrivial
- Provide guidance for design iterative search algorithms
  - Convergence
  - Conditions for the validity of points

CONVEXITY



Convex and nonconvex sets.



A convex function. Linear interpolation underestimates the function.

MINIMA AND CONVEXITY

- Local minima are also global under convexity

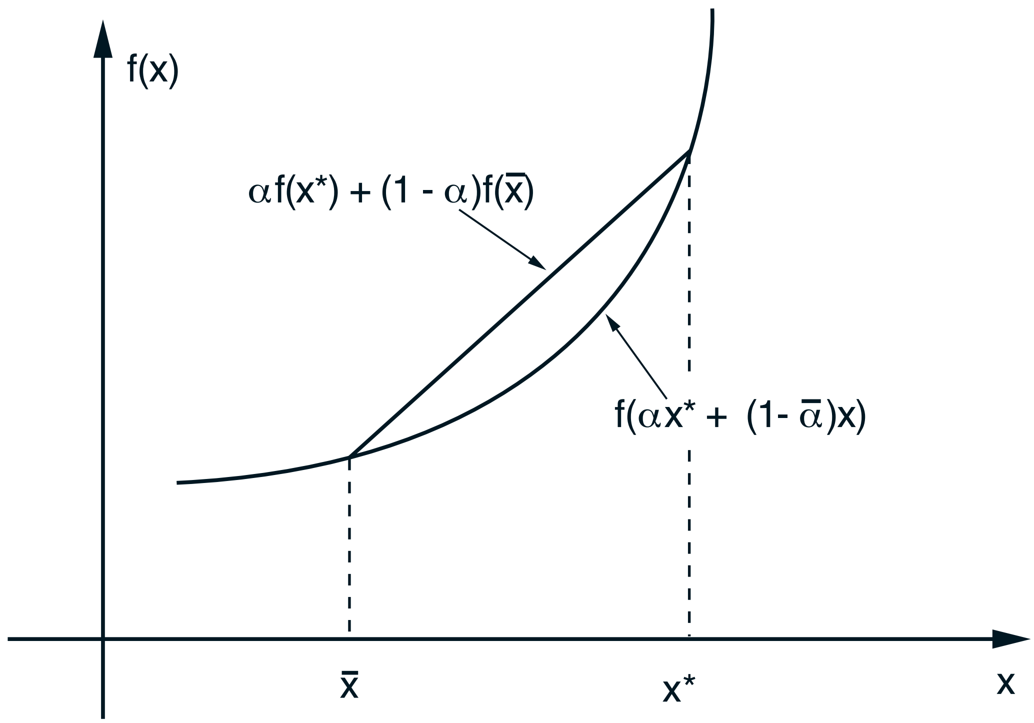


Illustration of why local minima of convex functions are also global. Suppose that f is convex and that x^* is a local minimum of f. Let $\overline{x}$ be such that $f(\overline{x}) < f(x^*)$. By convexity, for all $\alpha \in (0,1)$,

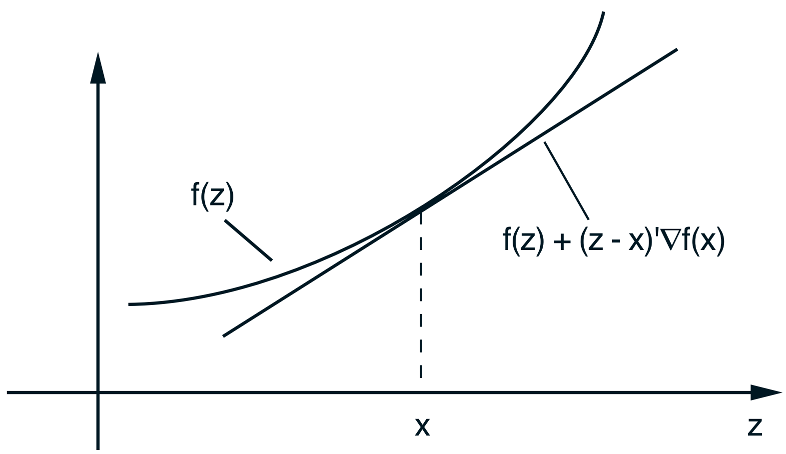
$$f\Big(\alpha x^* + (1-\alpha)\overline{x}\Big) \leq \alpha f(x^*) + (1-\alpha) f(\overline{x}) < f(x^*).$$

Thus, f takes values strictly lower than $f(x^*)$ on the line segment connecting x^* with $\overline{x}$, and x^* cannot be a local minimum which is not global.

OTHER PROPERTIES OF CONVEX FUNCTIONS

- f is convex if and only if the linear approximation at a point x based on the gradient, underestimates f:

$$f(z) \geq f(x) + \nabla f(x)'(z - x), \qquad \forall\, z \in \Re^n$$



  – Implication:

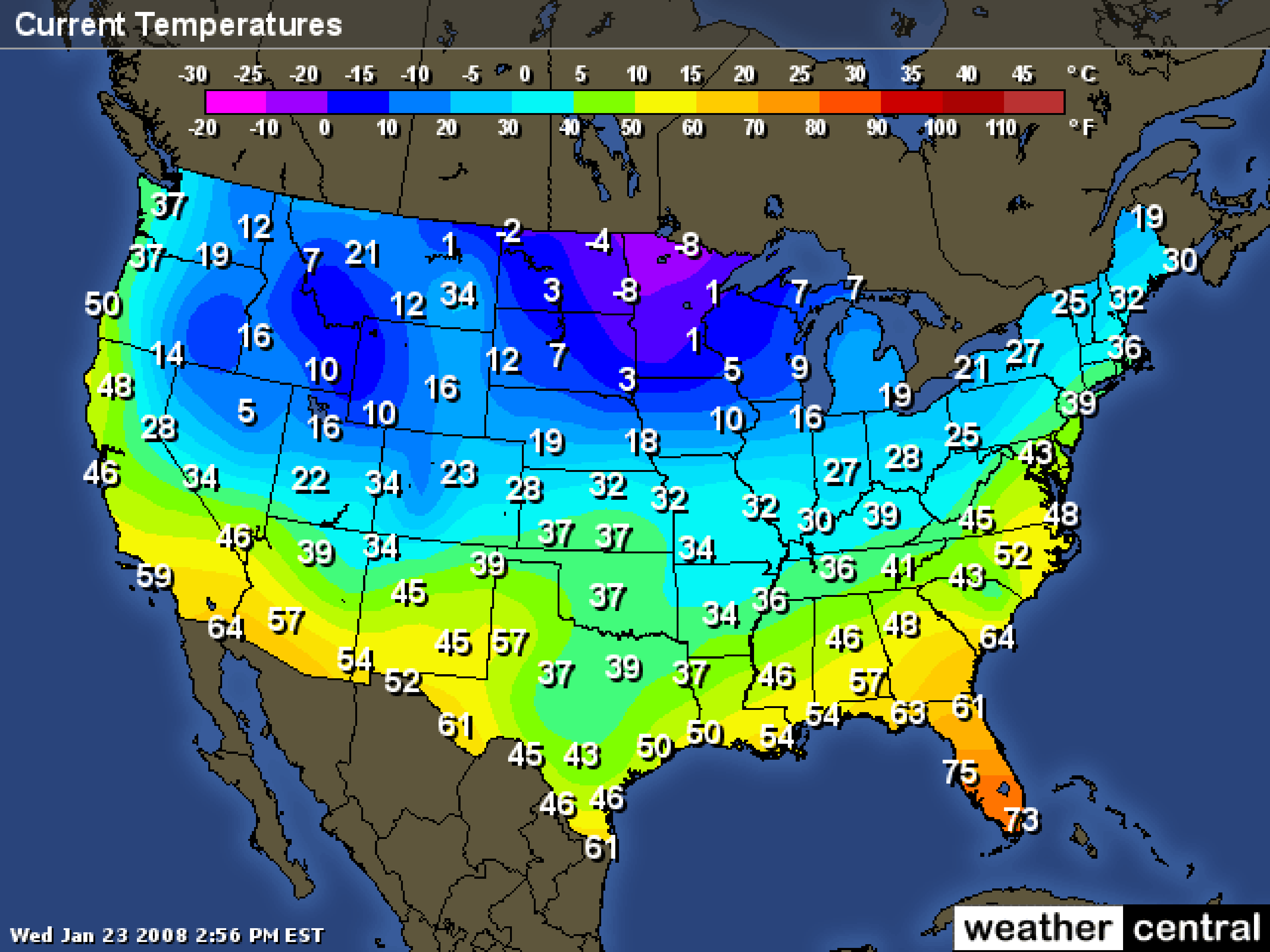
$$\nabla f(x^*) = 0 \quad \Rightarrow x^* \text{ is a global minimum}$$

- f is convex if and only if $\nabla^2 f(x)$ is positive semidefinite for all x

Convex Set and Convex Function

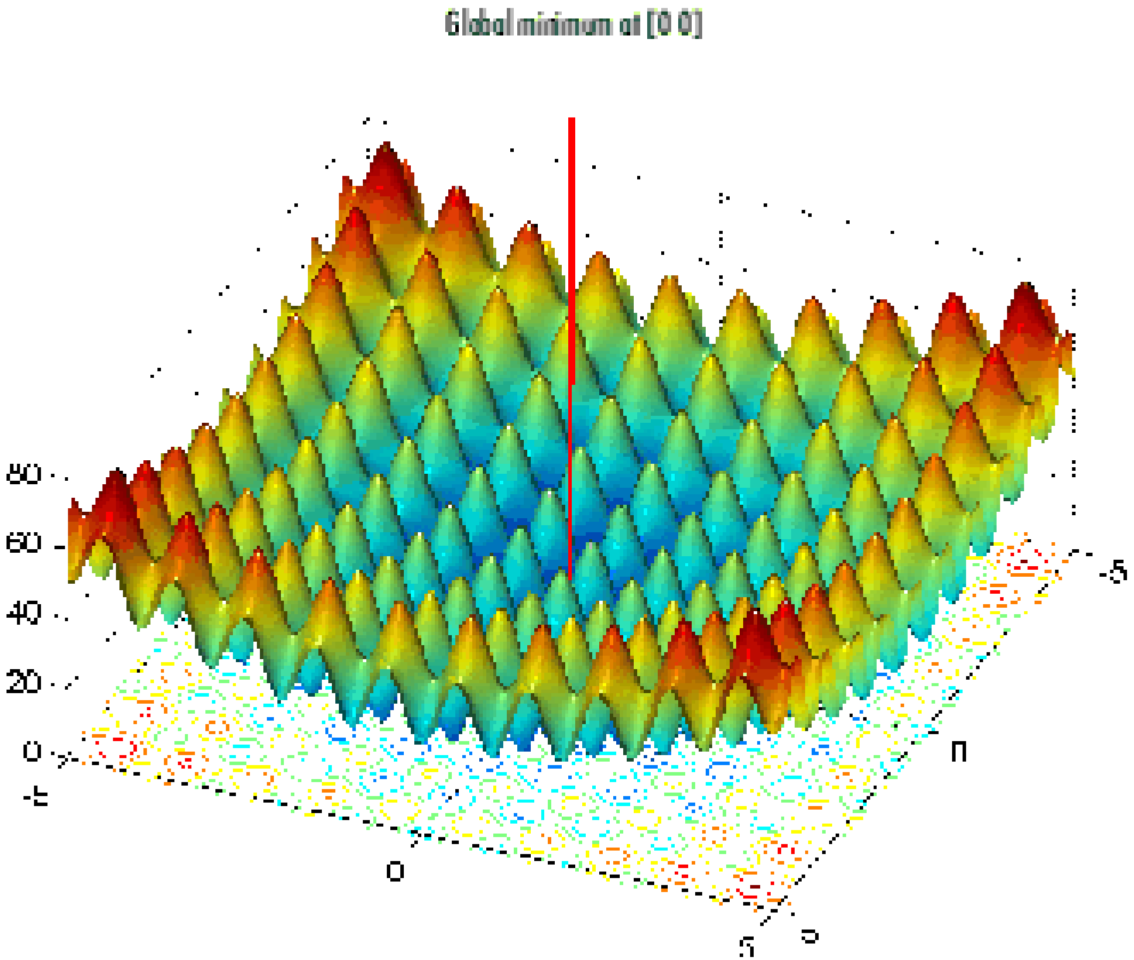
Motivation for Gradient Methods

- We have the closed-form information of f(x)
  - How to utilize this information
- What information can we get?
  - Why is gradient so important?
  - Easy to find the descending direction locally
  - Hard globally



Current Temperatures
-30 -25 -20 -15 -10 -5 0 5 10 15 20 25 30 35 40 45 °C
-20 -10 0 10 20 30 40 50 60 70 80 90 100 110 °F
37
12
37
19
7
21
1
-2
-4
-8
19
30
50
12
34
3
-8
1
7
7
25
32
16
14
10
12
7
1
36
48
16
3
5
9
21
27
5
16
10
19
10
16
39
28
19
18
25
46
34
22
34
23
28
32
32
32
27
28
43
30
39
45
48
46
37
37
34
39
34
39
36
41
52
59
43
45
37
36
64
57
34
48
46
64
45
57
54
52
37
39
37
46
57
61
63
54
61
50
54
50
45
43
75
46
46
73
61
Wed Jan 23 2008 2:56 PM EST
weather central

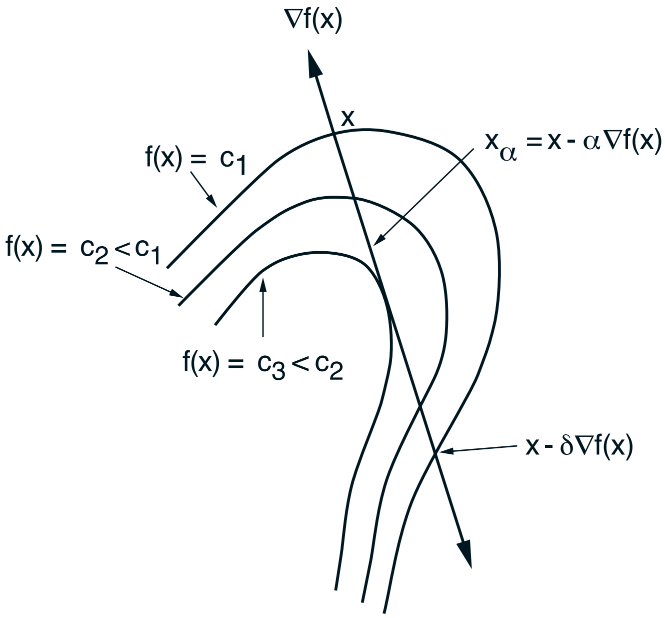
The Rastrigin Function Terrain



Idea of Gradient Methods

- Form an approximation to the function
  – Minimize the approximated function
- Two ingredients:
  – Direction
  – Stepsize

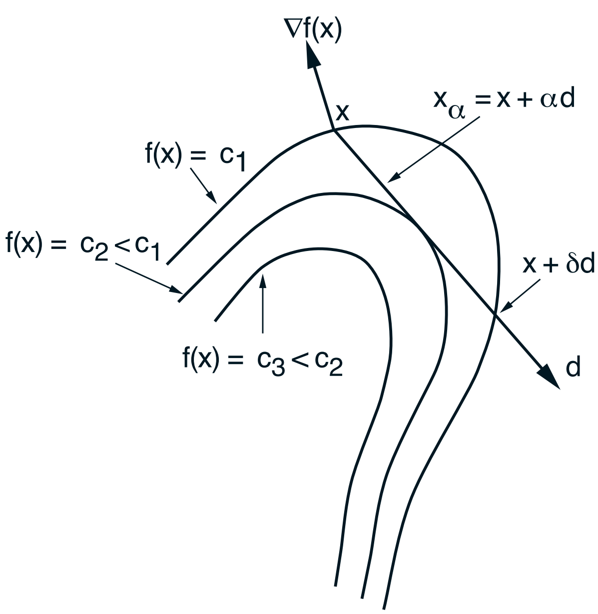
GRADIENT METHODS - MOTIVATION



If $\nabla f(x) \neq 0$, there is an interval $(0, \delta)$ of stepsizes such that

$$f\big(x - \alpha\nabla f(x)\big) < f(x)$$

for all $\alpha \in (0, \delta)$.



If d makes an angle with $\nabla f(x)$ that is greater than 90 degrees,

$$\nabla f(x)' d < 0,$$

there is an interval $(0, \delta)$ of stepsizes such that $f(x + \alpha d) < f(x)$ for all $\alpha \in (0, \delta)$.

Steepest Descent Example

- Minimize

$$X^2 + 100\ y^2$$

Where x,y are real numbers

PRINCIPAL GRADIENT METHODS

$$x^{k+1} = x^k + \alpha^k d^k, \qquad k = 0, 1, \ldots$$

where, if $\nabla f(x^k) \neq 0$, the direction d^k satisfies

$$\nabla f(x^k)' d^k < 0,$$

and α^k is a positive stepsize. Principal example:

$$x^{k+1} = x^k - \alpha^k D^k \nabla f(x^k),$$

where D^k is a positive definite symmetric matrix

- Simplest method: **Steepest descent**

$$x^{k+1} = x^k - \alpha^k \nabla f(x^k), \qquad k = 0, 1, \ldots$$

- Most sophisticated method: **Newton's method**

$$x^{k+1} = x^k - \alpha^k \big(\nabla^2 f(x^k)\big)^{-1} \nabla f(x^k), \qquad k = 0, 1, \ldots$$

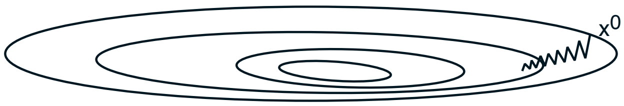
Newton’s Method Example

- Minimize

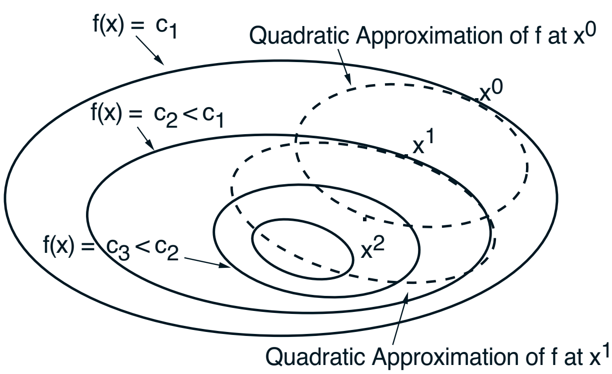
$$X^2 + 100\ y^2$$

Where x,y are real numbers

STEEPEST DESCENT AND NEWTON'S METHOD



Slow convergence of steepest descent



Fast convergence of Newton's method w/ $\alpha^k = 1$.

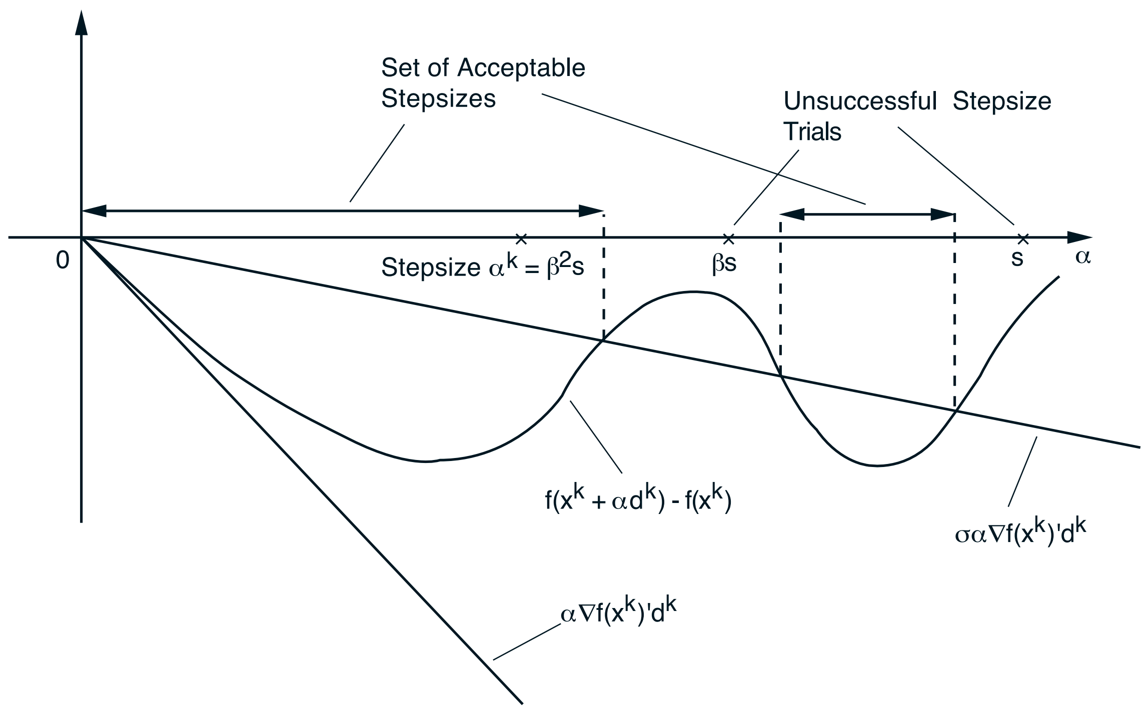
Given x^k, the method obtains x^{k+1} as the minimum of a quadratic approximation of f based on a second order Taylor expansion around x^k.

CHOICES OF STEPSIZE I

- Minimization Rule: α^k is such that

$$f(x^k + \alpha^k d^k) = \min_{\alpha \geq 0} f(x^k + \alpha d^k).$$

- Limited Minimization Rule: Min over $\alpha \in [0,\ s]$
- Armijo rule:



Start with s and continue with $\beta s, \beta^2 s, ...,$ until $\beta^m s$ falls within the set of α with

$$f(x^k) - f(x^k + \alpha d^k) \geq -\sigma \alpha \nabla f(x^k)' d^k.$$

CHOICES OF STEPSIZE II

- Constant stepsize: α^k is such that

$$\alpha^k = s : \text{ a constant}$$

- Diminishing stepsize:

$$\alpha^k \to 0$$

but satisfies the infinite travel condition

$$\sum_{k=0}^{\infty} \alpha^k = \infty$$