Optimality Conditions for Unconstrained Optimization

GIAN Short Course on Optimization: Applications, Algorithms, and Computation

Sven Leyffer

Argonne National Laboratory

September 12-24, 2016

Outline

1. Optimality Conditions for Unconstrained Optimization
   - Local and Global Minimizers

2. Iterative Methods for Unconstrained Optimization
   - Pattern Search Algorithms
   - Coordinate Descend with Exact Line Search
   - Line-Search Methods
   - Steepest Descend and Armijo Line Search

Unconstrained Optimization and Derivatives

Considering unconstrained optimization problem:

$$\underset{x \in \mathbb{R}^n}{\text{minimize}} \ f(x),$$

where $f : \mathbb{R}^n \to \mathbb{R}$ twice continuously differentiable.

Goal

Derive 1^{st} and 2^{nd} order optimality conditions.

Recall gradients and Hessian of $f : \mathbb{R}^n \to \mathbb{R}$:

- Gradient of $f(x)$:

$$\nabla f(x) := \left(\frac{\partial f}{\partial x_1}, \ldots, \frac{\partial f}{\partial x_n} \right)^T,$$

  exists $\forall x \in \mathbb{R}^n$.
- Hessian (2^{nd} derivative) matrix of $f(x)$:

$$\nabla^2 f(x) := \left[\frac{\partial^2 f}{\partial x_i \partial x_j} \right]_{i=1,\ldots,n, j=1,\ldots,n} \in \mathbb{R}^{n \times n}.$$

Derivatives: Simple Examples

Consider linear, $l(x)$, and quadratic function, $q(x)$:

$$l(x) = a^T x + b, \quad \text{and} \quad q(x) = \frac{1}{2} x^T G x + b^T x + c$$

Gradients and Hessians are given by:

- Gradient: $\nabla l(x) = a$ and Hessian $\nabla^2 l(x) = 0$, zero matrix.
- Gradient: $\nabla q(x) = Gx + b$ and Hessian $\nabla^2 q(x) = G$, constant matrix.

Remark

Linear unconstrained optimization, minimize $l(x)$, is unbounded.

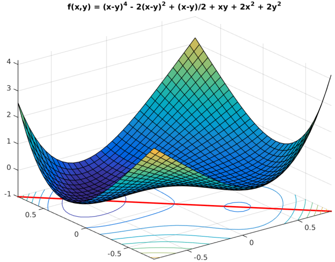
Lines and Restrictions along Lines

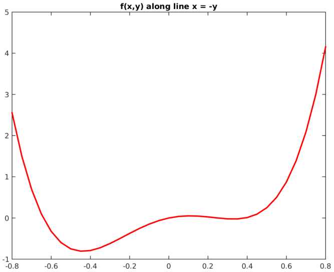
Consider restriction of nonlinear function along line, defined by:

$$\left\{x \in \mathbb{R}^n : x = x(\alpha) = x' + \alpha s,\ \forall \alpha \in \mathbb{R}\right\}$$

where α steplength for line through $x' \in \mathbb{R}^n$ in direction s.
Define restriction of $f(x)$ along line:

$$f(\alpha) := f(x(\alpha)) = f(x' + \alpha s).$$





$f(x, y) = (x - y)^4 - 2(x - y)^2 + (x - y)/2 + xy + 2x^2 + 2y^2$,
Contours and restriction along $x = -y$.

Deriving First-Order Conditions from Calculus

$$\underset{x \in \mathbb{R}^n}{\text{minimize}} \;\; f(x),$$

Use restriction of the objective $f(x)$ along a line ...
Recall sufficient conditions for local minimum of 1D function $f(\alpha)$:

$$\frac{df}{d\alpha} = 0, \quad \text{and} \quad \frac{d^2 f}{d\alpha^2} > 0$$

... first-order necessary condition is $\frac{df}{d\alpha} = 0$
Use chain rule (line $x = x' + \alpha s$) to derive operator

$$\frac{d}{dx} = \sum_{I=1}^{n} \frac{dx_i}{d\alpha} \frac{\partial}{\partial x_i} = \sum_{I=1}^{n} s_i \frac{\partial}{\partial x_i} = s^T \nabla .$$

Thus, slope of $f(\alpha) = f(x' + \alpha s)$ along direction s:

$$\frac{df}{dx} = s^T \nabla f(x') =: s^T g(x')$$

Deriving First-Order Conditions from Calculus

$$\underset{x \in \mathbb{R}^n}{\text{minimize}} \;\; f(x),$$

Thus, curvature of $f(\alpha) = f(x' + \alpha s)$ along s:

$$\frac{d^2 f}{dx^2} = \frac{d}{d\alpha} s^T g(x') = s^T \nabla g(x')^T s =: s^T H(x') s$$

Notation

- Gradient of $f(x)$ denoted as $g(x) := \nabla f(x)$
- Hessian of $f(x)$ denoted as $H(x) := \nabla^2 f(x)$

$$\Rightarrow \quad f(x' + \alpha s) = f(x') + \alpha s^T g(x') + \frac{1}{2}\alpha^2 s^T H(x') s + \ldots$$

ignoring higher-order terms $\mathcal{O}(|\alpha|^3)$.

Example: Powell's Function

Consider

$$f(x) = x_1^4 + x_1x_2 + (1 + x_2)^2$$

Gradient and Hessian are

$$\begin{pmatrix} 4x_1^3 + x_2 \\ x_1 + 2(1 + x_2) \end{pmatrix} \quad \text{and} \quad \begin{bmatrix} 12x_1^2 & 1 \\ 1 & 2 \end{bmatrix}$$

Local and Global Minimizers

$$\underset{x \in \mathbb{R}^n}{\text{minimize}} \; f(x),$$

Possible outcomes of optimization problem:

1. **Unbounded** if $\exists x^{(k)} \in \mathbb{R}^n$ such that $f^{(k)} = f(x^{(k)}) \to -\infty$,
2. **Minimizers may not exist**, or
3. **Local or Global Minimizer** defined below.

Definition

Let $x^* \in \mathbb{R}^n$, and $B(x^*, \epsilon) := \{x : \|x - x^*\| \leq \epsilon\}$ ball around x^*.

1. x^* *global minimizer*, iff $f(x^*) \leq f(x) \; \forall x \in \mathbb{R}^n$.
2. x^* *local minimizer*, iff $f(x^*) \leq f(x) \; \forall x \in B(x^*, \epsilon)$.
3. x^* *strict local minimizer*, iff $f(x^*) < f(x) \; \forall x^* \neq x \in B(x^*, \epsilon)$

Local and Global Minimizers

$$\underset{x \in \mathbb{R}^n}{\text{minimize}} \;\; f(x)$$

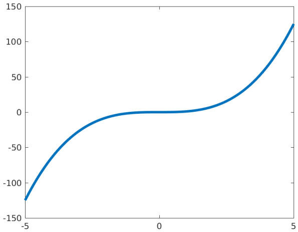
Global minimizer is local minimizer.

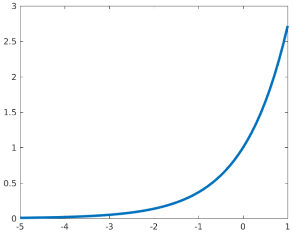
Examples: minimizer does not exist

- $f(x) = x^3$ unbounded below
  $\Rightarrow$ minimizer does not exist.

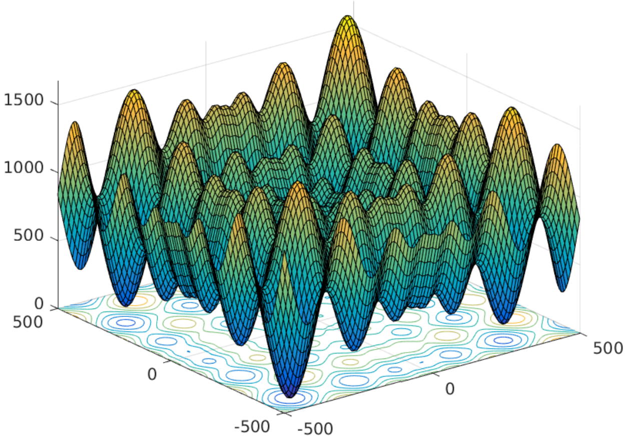
- $f(x) = \exp(x)$ bounded below,
  but minimizer does not exist.

... detected in practice monitoring $x^{(k)}$.





Global Optimization is Hard



Contours of Schefel function: $f(x) = 418.9829n + \sum_{i=1}^{n} x_i \sin(\sqrt{|x_i|})$

Global Optimization is Hard

Global Optimization is Much Harder

Finding (and verifying) a global minimizer is much harder:
... global optimization can be NP-hard or even undecidable!

Initially only consider local minimizers ...

Necessary Condition for Local Minimizers

$$\underset{x \in \mathbb{R}^n}{\text{minimize}} \; f(x)$$

At x^*, local minimizer

- Slope of $f(x)$ along s is zero

$$\Rightarrow s^T g(x^*) = 0 \quad \forall s \in \mathbb{R}^n$$

- Curvature of $f(x)$ along s is nonnegative

$$\Rightarrow s^T H(x^*) s \geq 0 \quad \forall s \in \mathbb{R}^n$$

Theorem (Necessary Conditions for Local Minimizer)

x^* local minimizer, then

$$g(x^*) := \nabla f(x^*) = 0, \;\; \textit{and} \;\; H(x^*) := \nabla^2 f(x^*) \succeq 0,$$

where $A \succeq 0$ means A positive semi-definite

Example: Powell's Function

Consider

$$f(x) = x_1^4 + x_1 x_2 + (1 + x_2)^2$$

Gradient and Hessian are

$$g(x) = \begin{pmatrix} 4x_1^3 + x_2 \\ x_1 + 2(1 + x_2) \end{pmatrix} \qquad \text{and} \qquad H(x) = \begin{bmatrix} 12x_1 & 1 \\ 1 & 2 \end{bmatrix}$$

At $x = (0.6959, -1.3479)$ get $g = 0$ and

$$H = \begin{bmatrix} 8.3508 & 1 \\ 1 & 2 \end{bmatrix} \quad \text{pos. def.}$$

... eigenvalue $= 1.8463, 8.5045$ ... using Matlab's `eig(H)` function

Sufficient Condition for Local Minimizers

$$\underset{x \in \mathbb{R}^n}{\text{minimize}} \; f(x)$$

Obtain sufficient condition by strengthening positive definiteness

Theorem (Sufficient Conditions for Local Min)

Assume that

$$g(x^*) := \nabla f(x^*) = 0, \quad \textit{and} \quad H(x^*) := \nabla^2 f(x^*) \succ 0,$$

then x^* is isolated local minimizer of $f(x)$.

Recall $A \succ 0$ positive definite, iff

- All eigenvalues of A are positive,
- $A = L^T DL$ factors exist with L lower triangular, $L_{ii} = 1$ and $D > 0$ diagonal,
- Cholesky factors, $A = L^T L$, exist with $L_{ii} > 0$, or
- $s^T As > 0$ for all $s \in \mathbb{R}^n, s \neq 0$.

Sufficient Condition for Local Minimizers

$$\underset{x \in \mathbb{R}^n}{\text{minimize}} \; f(x)$$

Gap between necessary and the sufficient conditions:
$H(x^*) := \nabla^2 f(x^*) \succeq 0$ versus $H(x^*) \succ 0$

Definition (Stationary Point)

x^* stationary point of $f(x)$, iff $g(x^*) = 0$ (aka 1^{st}-order condition).

Sufficient Condition for Local Minimizers

$$\underset{x \in \mathbb{R}^n}{\text{minimize}} \ f(x)$$

Definition (Stationary Point)

x^* stationary point of $f(x)$, iff $g(x^*) = 0$ (aka 1^{st}-order condition).

Classification of Stationary Points:

- **Local Minimizer:** $H(x^*) \succ 0$ then x^* is local minimizer.
- **Local Maximizer:** $H(x^*) \prec 0$ then x^* is local maximizer.
- **Unknown:** $H(x^*) \succeq 0$ cannot be classified.
- **Saddle Point:** $H(x^*)$ indefinite then x^* is a saddle point.

$H(x^*)$ indefinite, iff both positive and negative eigenvalues

Discussion of Optimality Conditions

Limitations of Optimality Conditions

Almost impossible to say anything about global optimality.

Why are optimality conditions important?

- Provide guarantees that candidate x^* is local min.
- Indicate when point is not optimal: necessary conditions.
- Provide termination condition for algorithms, e.g.

$$\|g(x^{(k)})\| \leq \epsilon \qquad \text{for tolerance} \quad \epsilon > 0$$

- Guide development of methods, e.g.

$$\underset{x}{\text{minimize}} \; f(x) \quad \text{“} \Leftrightarrow \text{”} \quad g(x) = \nabla f(x) = 0$$

... nonlinear system of equations ... use Newton's method

Outline

1. Optimality Conditions for Unconstrained Optimization
   - Local and Global Minimizers
2. Iterative Methods for Unconstrained Optimization
   - Pattern Search Algorithms
   - Coordinate Descend with Exact Line Search
   - Line-Search Methods
   - Steepest Descend and Armijo Line Search

Iterative Methods for Unconstrained Optimization

In general, cannot solve

$$\underset{x \in \mathbb{R}^n}{\text{minimize}} \; f(x)$$

analytically ... need iterative methods.

Iterative Methods for Optimization

- Start from initial guess of solution, $x^{(0)}$
- Given $x^{(k)}$, construct new (better) iterate $x^{(k+1)}$
- Construct sequence $x^{(k)}$, for $k = 1, 2, \dots$ converging to x^*

Key Question

Does $\|x^{(k)} - x^*\| \to 0$ hold? Speed of convergence?

Outline

1. Optimality Conditions for Unconstrained Optimization
   - Local and Global Minimizers
2. Iterative Methods for Unconstrained Optimization
   - Pattern Search Algorithms
   - Coordinate Descend with Exact Line Search
   - Line-Search Methods
   - Steepest Descend and Armijo Line Search

Pattern-Search Techniques

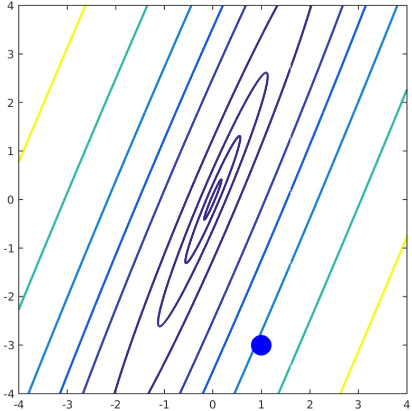
Class of methods that does not require gradients
... suitable for simulation-based optimization

Search for lower function value along coordinate directions: $\pm e_i$

Reduce "step-length' if no progress is made

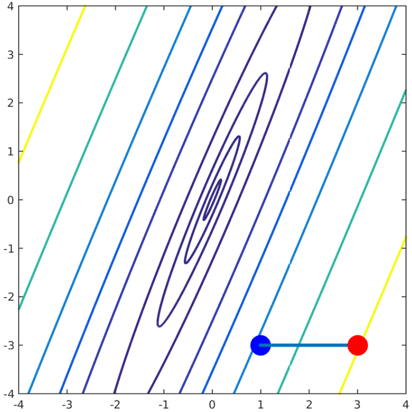
Weak convergence properties ... but can be effective for derivative-free optimization

Pattern-Search Techniques



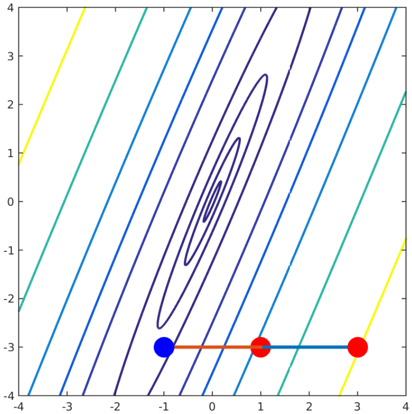
Starting from $x^{(0)}$ with $\Delta = 2$

Pattern-Search Techniques



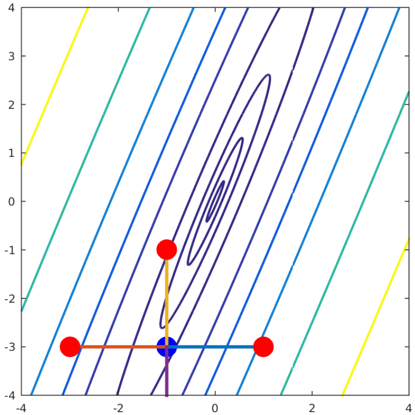
First search step reduces $f(x)$ ... new $x^{(1)}$

Pattern-Search Techniques



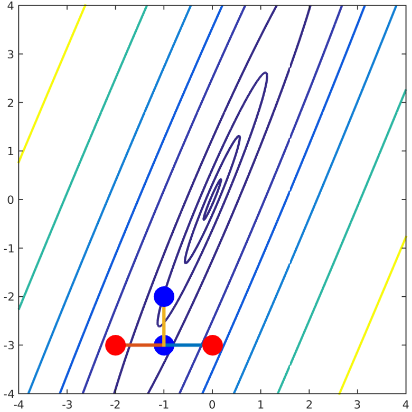
First search step reduces $f(x)$ ... new $x^{(2)}$

Pattern-Search Techniques



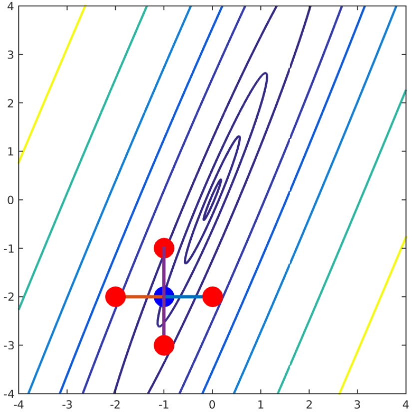
No polling step reduced $f(x) \Rightarrow$ shrink $\Delta = \Delta/2$

Pattern-Search Techniques



Second search step reduces $f(x)$ ... new $x^{(3)}$

Pattern-Search Techniques



No polling step reduced $f(x) \Rightarrow$ shrink $\Delta = \Delta/2$

Outline

1. Optimality Conditions for Unconstrained Optimization
   - Local and Global Minimizers
2. Iterative Methods for Unconstrained Optimization
   - Pattern Search Algorithms
   - Coordinate Descend with Exact Line Search
   - Line-Search Methods
   - Steepest Descend and Armijo Line Search

Coordinate Descend Algorithms [Wotao Yin, UCLA]

- Make progress by updating one (or a few) variables at a time.
- Regarded as inefficient and outdated since the 1960's
- Recently, found to work well for huge optimization problems
  ... arising in statistics, machine-learning, compressed sensing

$$\underset{x \in \mathbb{R}^n}{\text{minimize}} \; f(x)$$

Basic Coordinate Descend Method

Given $x^{(0)}$, set $k = 0$.

repeat

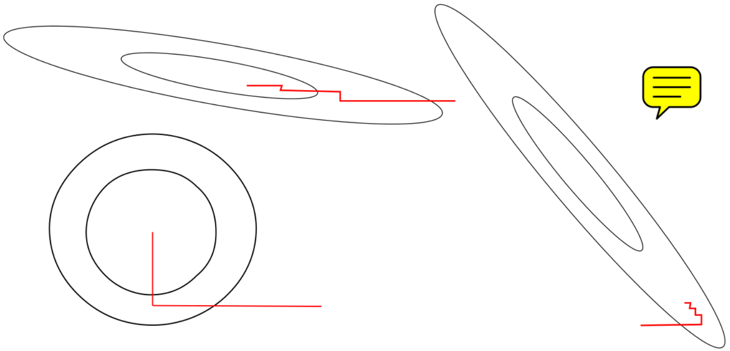
  Choose $i \in \{1, \ldots, n\}$ coordinate; set $x^{(k+1)} := x^{(k)}$.

$$x_i^{(k+1)} \leftarrow \underset{x_i \in \mathbb{R}}{\text{argmin}}\, f\left(x_1^{(k)}, \ldots, x_i, \ldots, x_n^{(k)}\right)$$

  Set $k = k + 1$

until $x^{(k)}$ *is (local) optimum*;

Coordinate Descend Algorithms [Wotao Yin, UCLA]



Outline

1. Optimality Conditions for Unconstrained Optimization
   - Local and Global Minimizers

2. Iterative Methods for Unconstrained Optimization
   - Pattern Search Algorithms
   - Coordinate Descend with Exact Line Search
   - Line-Search Methods
   - Steepest Descend and Armijo Line Search

Line-Search Method

Line-Search Methods

- Find descend direction, $s^{(k)}$, such that ${s^{(k)}}^T g(x^{(k)}) < 0$
- Search for reduction if $f(x)$ along line $s^{(k)}$

General Line-Search Method

Given $x^{(0)}$, set $k = 0$.

repeat

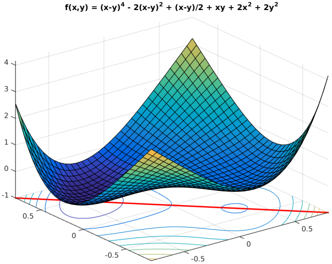
Find search direction $s^{(k)}$ with ${s^{(k)}}^T g(x^{(k)}) < 0$

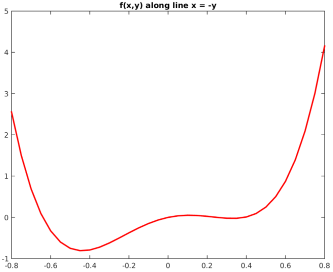
Find steplength α_k with $f(x^{(k)} + \alpha_k s^{(k)}) < f(x^{(k)})$

Set $x^{(k+1)} := x^{(k)} + \alpha_k s^{(k)}$ and $k = k + 1$

until $x^{(k)}$ *is (local) optimum*;

Line-Search Methods





$$f(x,y) = (x-y)^4 - 2(x-y)^2 + (x-y)/2 + xy + 2x^2 + 2y^2,$$

Contours and restriction along $x = -y$.

Line-Search Method

Remarks regarding general descend method:

- ${s^{(k)}}^T g(x^{(k)}) < 0$ not enough for convergence.
- Many possible choices for $s^{(k)}$.
- Exact line search:

$$\underset{\alpha}{\text{minimize}} \; f(x^{(k)} + \alpha s^{(k)})$$

  ... impractical $\Rightarrow$ consider approximate techniques.

- Simple descend,

$$f(x^{(k)} + \alpha_k s^{(k)}) < f(x^{(k)})$$

  ... not enough for convergence $\Rightarrow$ strengthen!

Armijo Line-Search

Armijo Line-Search Method at x in Direction s

$\alpha =$ **function** Armijo$(f(x), x, s)$

Let $t > 0$, $0 < \beta < 1$, and $0 < \sigma < 1$ constants

Set $\alpha_0 := t$, and $j := 0$.

while $f(x) - f(x + \alpha s) < -\alpha\sigma g(x)^T s$ **do**

| Set $\alpha_{j+1} := \beta\alpha_j$ and $j := j + 1$.

end

- Simple back-tracking line-search.
- Typically start at $t = 1$.
- Tightens simple descend condition $f(x) - f(x + \alpha s) < 0$:
  - $g(x)^T s$ *predicted reduction* from linear model
  - $f(x) - f(x + \alpha s)$ *actual reduction*
  - Step must achieve factor $\sigma < 1$ of predicted reduction.

$\Rightarrow$ allows convergence proofs!

Outline

1. Optimality Conditions for Unconstrained Optimization
    - Local and Global Minimizers
2. Iterative Methods for Unconstrained Optimization
    - Pattern Search Algorithms
    - Coordinate Descend with Exact Line Search
    - Line-Search Methods
    - Steepest Descend and Armijo Line Search

Steepest Descend Method

Search direction that maximizes the descend,

$$s^{(k)} := -g(x^{(k)}) \quad \text{steepest descend direction}$$

Steepest descend satisfies descend property:

$$s^{(k)^T} g(x^{(k)}) = -g^{(k)^T} g^{(k)} = -\|g^{(k)}\|_2^2 < 0$$

$s^{(k)} := -g^{(k)}/\|g^{(k)}\|$ normalized direction of most negative slope
Let θ be angle between direction, s, gradient g, then:

$$s^T g = \|s\| \cdot \|g\| \cdot \cos(\theta),$$

and get min when $\cos(\theta) = -1$, or $\theta = \pi$, i.e. $s = -g$.

Steepest Descend with Armijo Line-Search

Steeped Descend Armijo Line-Search Method

Given $x^{(0)}$, set $k = 0$.

repeat

Find steepest descend direction $s^{(k)} := -g(x^{(k)})$

Armijo Line search: $\alpha_k :=$ Armijo$(f(x), x^{(k)}, s^{(k)})$

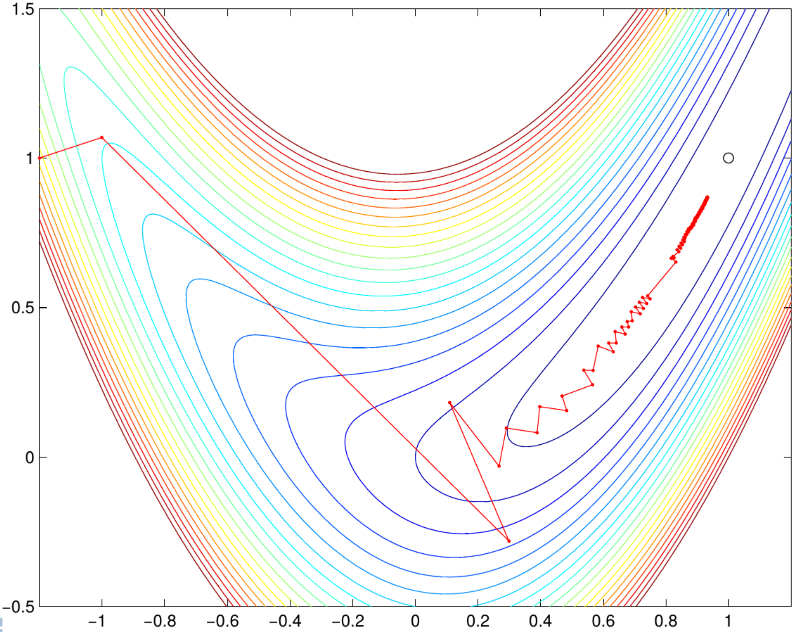
Set $x^{(k+1)} := x^{(k)} + \alpha_k s^{(k)}$ and $k = k + 1$.

until $x^{(k)}$ *is (local) optimum*;

Theorem

If $f(x)$ bounded below, then converge to stationary point.

Steepest Descend can be Inefficient in Practice



Main Take-Aways from Lecture

- Optimality conditions
- General structure of methods
- Line Search
- Steepest Descend