Numerical Optimization

Lecture 6: Line Search Method

Hao Wang
Email: wanghao1@shanghaitech.edu.cn

Contents

Contents

Theorem 1 (Mean Value Theorem)

Given $f \in \mathcal{C}$, $x \in \mathbb{R}^n$, *and* $d \in \mathbb{R}^n$, *there exists* $\alpha \in (0,1)$ *such that*

$$f(x+d) = f(x) + \nabla f(x+\alpha d)^T d.$$

Theorem 2 (Taylor's Theorem)

Given $f \in \mathcal{C}^2$, $x \in \mathbb{R}^n$, *and* $d \in \mathbb{R}^n$, *there exists* $\alpha \in (0,1)$ *such that*

$$f(x+d) = f(x) + \nabla f(x)^T d + \tfrac{1}{2} d^T \nabla^2 f(x+\alpha d)^T d.$$

Definition 3 (Convex function)

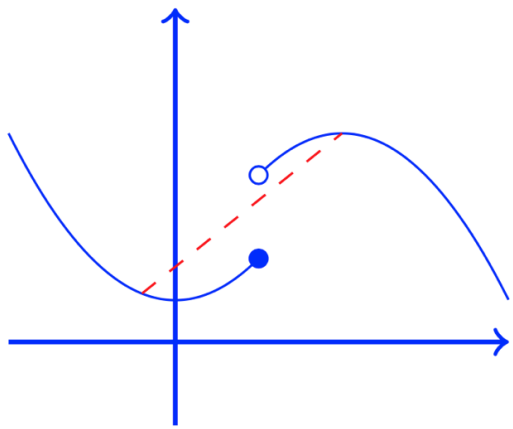
A function $f : \mathbb{R}^n \to \mathbb{R}$ is convex if for all $\{x_1, x_2\} \subset \mathbb{R}^n$ and $\alpha \in [0, 1]$ we have

$$f(\alpha x_1 + (1 - \alpha)x_2) \leq \alpha f(x_1) + (1 - \alpha) f(x_2).$$

- More generally, convexity of a function presumes convexity of its domain.
- A function f is strictly convex if for $x_1 \neq x_2$, the above inequality holds strictly.
- f is concave if $-f$ is convex
- What is its graphic intuition?

Continuity

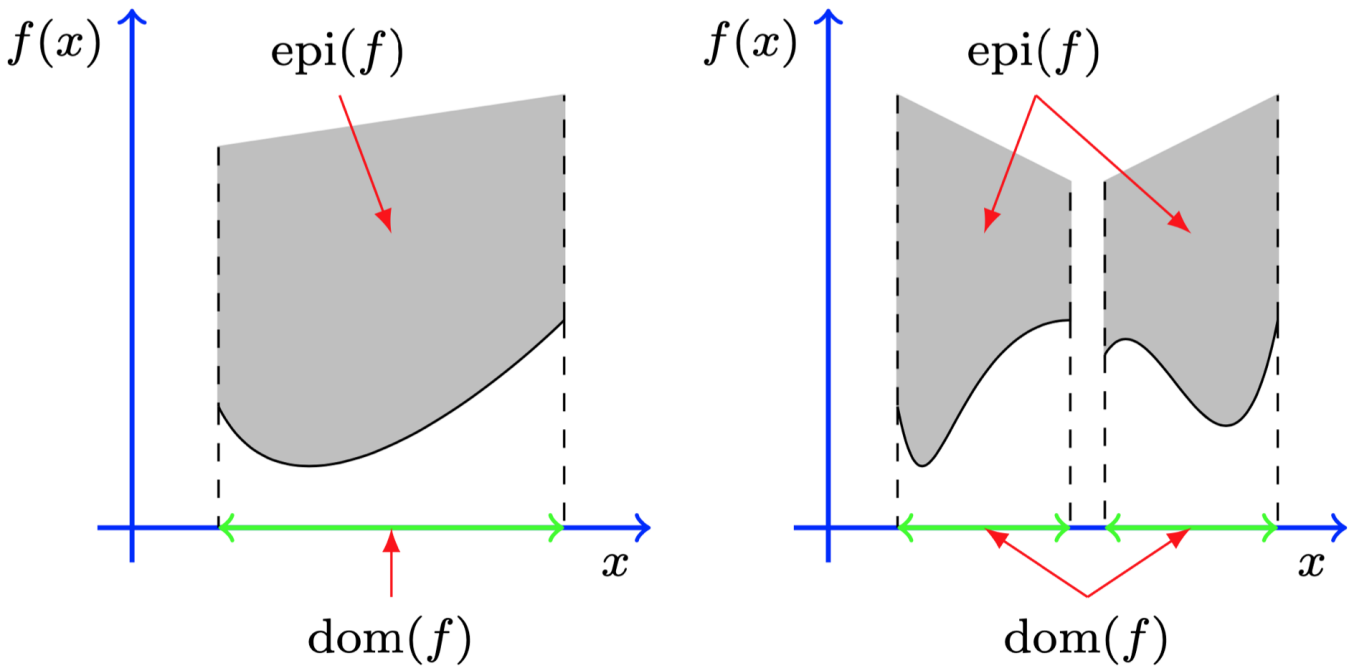
If $f : \mathbb{R}^n \to \mathbb{R}$ is convex, then it is continuous.



Definition 4

Epigraphs The epigraph of f is

$$\text{epi}(f) := \{(x, z) : x \in \mathcal{X}, z \in \mathbb{R}, \text{ and } f(x) \leq z\}.$$



Epigraphs of a convex (left) and a nonconvex (right) function

Convex functions:

- Affine: $a^T x + b$
- Powers: x^a for $x > 0$ and $a \in \mathbb{R} \setminus (0, 1)$
- Powers of absolute values: $|x|^a$ for $a \geq 1$
- Exponential: e^{ax} for $a \in \mathbb{R}$
- Negative entropy: $x \log x$ for $x > 0$
- p-norms: $\|x\|_p := (\sum_i |x_i|^p)^{1/p}$ for $p \geq 1$

Concave functions:

- Affine: $a^T x + b$
- Powers: x^a for $x > 0$ and $a \in [0, 1]$
- Logarithms: $\log x$ for $x > 0$

$\bar{\mathbb{R}}$ is the extended reals; i.e., $\bar{\mathbb{R}} := \mathbb{R} \cup \{-\infty, +\infty\}$

- Addition:
  If $f_1, ..., f_k : \mathbb{R}^n \to \bar{\mathbb{R}}$ are convex and $\alpha_1, ..., \alpha_k > 0$, then

$$f(x) = \sum_{i=1}^{k} \alpha_i f_i(x) \text{ is convex.}$$

- Maximization:
  If $f_1, ..., f_k : \mathbb{R}^n \to \bar{\mathbb{R}}$ are convex, then

$$f(x) = \max\{f_1(x), ..., f_k(x)\} \text{ is convex.}$$

- Composition:
  If $g : \mathbb{R} \to \bar{\mathbb{R}}$ is nondecreasing and convex, and $h : \mathbb{R}^n \to \bar{\mathbb{R}}$ is convex, then

$$f(x) = g(h(x)) \text{ is convex.}$$

Theorem 5

Let $\mathcal{X}$ be a nonempty convex subset of $\mathbb{R}^n$ and let $f : \mathbb{R}^n \to \mathbb{R}$ be differentiable over an open set containing $\mathcal{X}$. Then, the following hold:

(a) *f is convex over $\mathcal{X}$ if and only if, for all $\{x_1, x_2\} \subset \mathcal{X}$, we have*

$$f(x_2) \geq f(x_1) + \nabla f(x_1)^T (x_2 - x_1).$$

(b) *f is strictly convex over $\mathcal{X}$ if and only if the above inequality is strict when $x_1 \neq x_2$.*

What does this graphically mean?

Theorem 6

Let $\mathcal{X}$ be a nonempty convex subset of $\mathbb{R}^n$ and let $f : \mathbb{R}^n \to \mathbb{R}$ be twice continuously differentiable over an open set containing $\mathcal{X}$. Then, the following hold:

(a) *If $\nabla^2 f(x)$ is positive semidefnite for all $x \in \mathcal{X}$, then f is convex over $\mathcal{X}$.*

(b) *If $\nabla^2 f(x)$ is positive definite for all $x \in \mathcal{X}$, then f is convex over $\mathcal{X}$.*

(c) *If $\mathcal{X}$ is open and f is convex over $\mathcal{X}$, then $\nabla^2 f(x)$ is p.s.d. for all $x \in \mathcal{X}$.*

What does this graphically mean?

The effective domain of f is $\text{dom}(f) := \{x \mid x \in \mathcal{X} \text{ and } f(x) < \infty\}$.

Definition 7 (Subgradient and Subdifferential)

A vector $g \in \mathbb{R}^n$ is a subgradient of a proper convex f at $x \in \text{dom}(f)$ if

$$f(\bar{x}) \geq f(x) + g^T(\bar{x} - x) \text{ for all } \bar{x} \in \mathbb{R}^n.$$

The set of all subgradients of f at x, denoted $\partial f(x)$, is the subdifferential of f at x.

What does this graphically mean?

Theorem 8

If f is differentiable at $x \in \text{int dom}(f)$, then $\nabla f(x)$ is its unique subgradient at x.

Definition 9 (Directional derivative)

Given proper $f : \mathbb{R}^n \to \bar{\mathbb{R}}$, a point $x \in \operatorname{dom}(f)$, and a direction $d \in \mathbb{R}^n$, then directional derivative of f at x is the direction d (if it exits) is

$$f'(d; x) = \lim_{\alpha \searrow 0} \frac{f(x + \alpha d) - f(x)}{\alpha}.$$

- Note that this definition does not require f to be differentiable
- If f is convex, then for every $x \in \operatorname{dom}(f)$ and $d \in \mathbb{R}^n$, the limits exits
- If f is convex and $x \in \operatorname{int} \operatorname{dom}(f)$, then $f'(d, x)$ is finite
- If $f \in \mathcal{C}$, then $f'(d; x)$ exists and

$$f'(d; x) = \nabla f(x)^T d.$$

Consider the problem of minimizing a $\mathcal{C}^1$ function $f : \mathbb{R}^n \to \mathbb{R}$

- At a point $x \in \mathbb{R}^n$, a descent direction d is one for which we have

$$\nabla f(x)^T d = f'(d; x) < 0.$$

- We can decrease f by moving (a small distance) along such a direction d. (I expect you can prove it!)

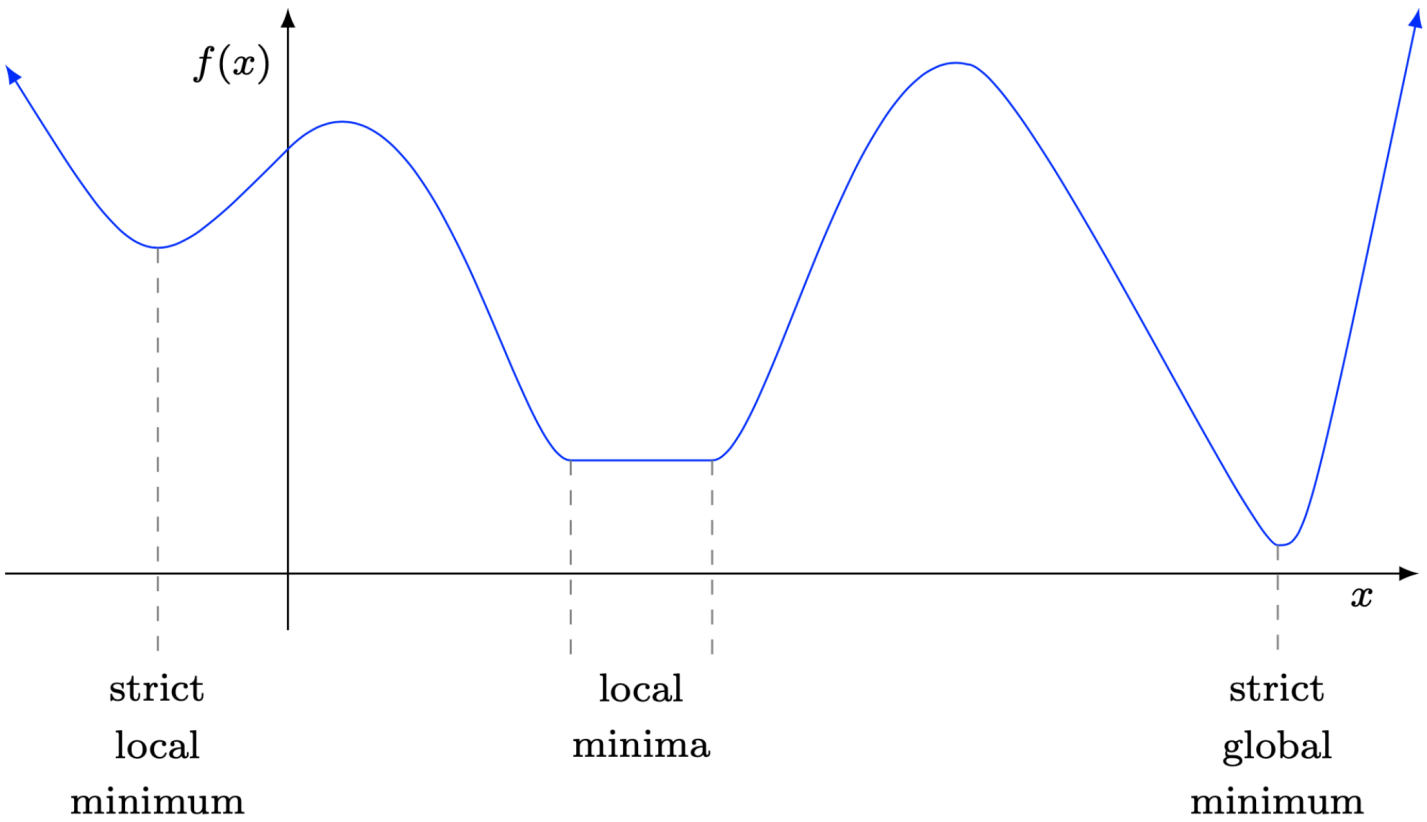
- Along which (normalized) direction is f decreasing at the fastest rate? The steepest descent direction is the solution of the optimization problem

$$\min_{d \in \mathbb{R}^n} f'(d; x) \quad \text{s.t. } \|d\|_2 \leq 1 \implies d = -\nabla f(x).$$

Contents

Consider the unconstrained optimization problem

$$\min_{x \in \mathbb{R}^n} f(x).$$



Ideal minima are those that minimize a function globally over its domain.

Definition 10 (Global minimum)

A vector x_* is a global minimum of f if

$$f(x_*) \leq f(x) \quad \text{for all } x \in \mathbb{R}^n.$$

Commonly, however, we are satisfied with a weaker form of minimum.

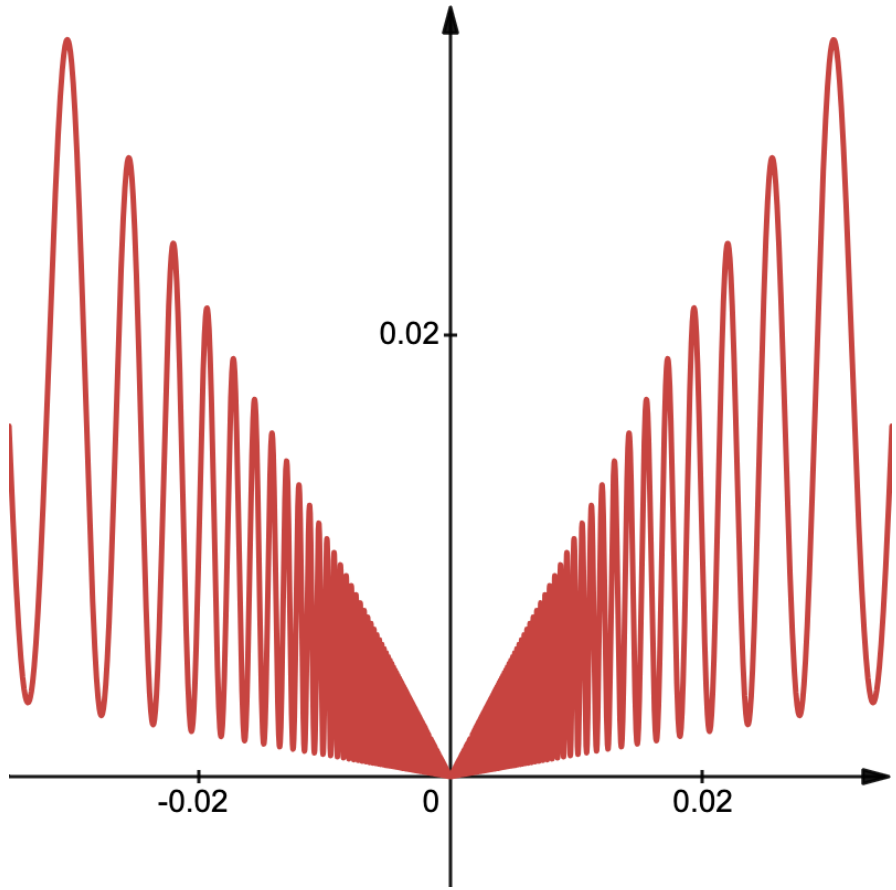
Definition 11 (Local minimum)

A vector x_* is a local minimum of f if there exists $\epsilon > 0$ such that

$$f(x_*) \leq f(x) \quad \text{for all } x \in \mathbb{B}(x_*, \epsilon) := \{x \in \mathbb{R}^n \mid \|x - x_*\|_2 \leq \epsilon\}.$$

- We also characterize a strict global/local minimizer if the $f(x_*) \leq f(x)$ holds strictly for $x \neq x_*$.
- x_* is an isolated global/local minimizer if, for some $\epsilon' > 0$, it is the only local minimizer in the neighborhood $\mathbb{B}(x_*, \epsilon')$.

Example: $0.5 \cdot x \cdot \sin\left(\frac{1}{x}\right) + 0.6 \cdot |x|$



A special fact in convex optimization is that all local minima are global minima.

Theorem 12

If $f : \mathbb{R}^n \to \mathbb{R}$ is convex, then a local minimum of f is a global minimum of f. If f is strictly convex, then there exits at most one global minimum of f.

Proof.

To derive a contradiction, suppose that x_* is a local minimum of f that is not a global minimum. Then, there exists $\bar{x} \in \mathbb{R}^n$ such that $f(\bar{x}) < f(x_*)$. By convexity of f, we have for all $\alpha \in (0, 1)$ that

$$f(\alpha x_* + (1-\alpha)\bar{x}) \leq \alpha f(x_*) + (1-\alpha) f(\bar{x}) < f(x_*).$$

This means that f has a value strictly lower than $f(x_*)$ at every point on the line segment $(x_*, \bar{x}]$, which violates the local minimality of x_*. (The statement about strictly convex f can be proved in a similar manner.) $\square$

- Unfortunately, for nonconvex optimization, the conditions in the definitions of global and local minima are not entirely useful.
- Unless we can verify strict quasiconvexity, we rarely have global information about f, and so have no way to verify if a point is a global minimizer.
- Thus, in nonconvex optimization, we often focus on finding a local minimizer.
- Using calculus, we can derive local optimality conditions that aid in determining if a point is a local minimizer.
- In this manner, we rarely (if ever) use the aforementioned definitions directly.

Theorem 13 (First-order necessary condition)

If $f \in \mathcal{C}$ and x_* is a local minimizer of f, then $\nabla f(x_*) = 0$.

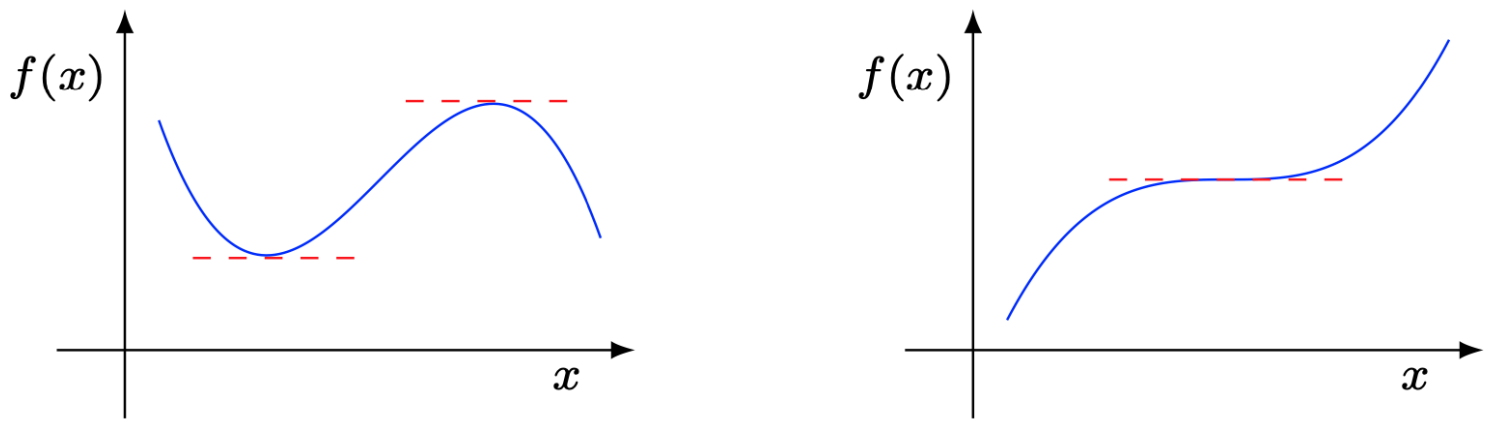
Proof.

For $x \in \mathbb{R}^n$ with $\nabla f(x) \neq 0$, let $d = -\nabla f(x)$ (with $\nabla f(x)^T d = -\|\nabla f(x)\|_2^2 < 0$). Since f is continuous, there exists $\alpha' > 0$ such that $d^T \nabla f(x + \alpha d) < 0$ for all $\alpha \in [0, \alpha']$, i.e., the directional derivative remains negative some way along d. By the Mean Value Theorem, for any $\alpha'' \in (0, \alpha']$ we have

$$f(x + \alpha'' d) = f(x) + \alpha'' d^T \nabla f(x + \alpha d) \quad \text{for some } \alpha \in (0, \alpha'').$$

Thus, $f(x + \alpha'' d) < f(x)$ for all $\alpha'' \in (0, \alpha']$. □

- We can limit our search to points where $\nabla f(x_*) = 0$.
- However, $\nabla f(x_*) = 0$ does not imply that we have a local minimizer!
- At least we know that if $\nabla f(x) = 0$, then x is not a local minimizer.



Definition 14 (Stationary point)

A point $x \in \mathbb{R}^n$ is a stationary point for $f \in \mathcal{C}$ if $\nabla f(x) = 0$.

If $f : \mathbb{R}^n \to \mathbb{R}$ is convex (but not necessarily real-valued or differentiable), then we have the following stronger result.

Theorem 15 (First-order necessary and sufficient condition)

If $f : \mathbb{R}^n \to \bar{\mathbb{R}}$ is convex and $0 \in \partial f(x_*)$, then x_* is a global minimizer of f.

Theorem 16 (Second-order necessary condition)

If $f \in \mathcal{C}^2$ and x_* is a local minimizer of f, then $\nabla^2 f(x_*) \succeq 0$.

Proof.

For $x \in \mathbb{R}^n$ with $\nabla f(x) = 0$ but $\nabla^2 f(x) \not\succeq 0$, let $d \in \mathbb{R}^n$ satisfy $d^T \nabla f(x) d < 0$. (We call such a d a direction of negative curvature.) Since $\nabla^2 f$ is continuous, there exists $\alpha' > 0$ such that

$$d^T \nabla^2 f(x + \alpha d) d < 0 \quad \text{for all } \alpha \in [0, \alpha'],$$

i.e., the curvature remains negative some way along d. By Taylor's Theorem, for all $\alpha'' \in (0, \alpha']$ and some $\alpha \in (0, \alpha'')$ we have

$$\begin{aligned} f(x + \alpha'' d) &= f(x) + \alpha'' \nabla f(x)^T d + \tfrac{1}{2} \alpha''^2 d^T \nabla^2 f(x + \alpha d) d \\ &= f(x) + \tfrac{1}{2} \alpha''^2 d^T \nabla^2 f(x + \alpha d) d \\ &< f(x). \end{aligned}$$

Thus, x cannot be a minimizer. □

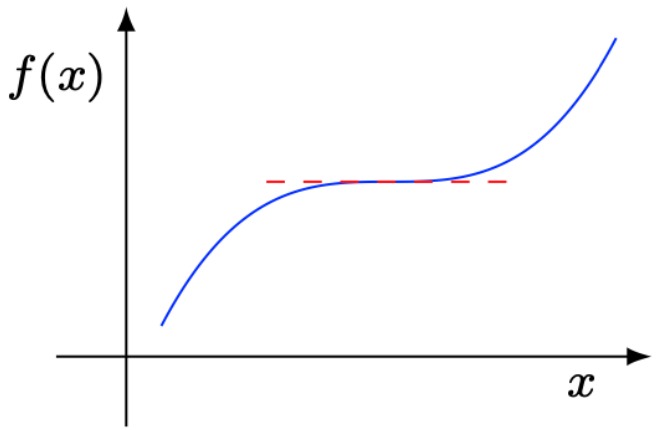
Thus, at a local minimizer x_*, the Hessian of f is positive semidefinite.

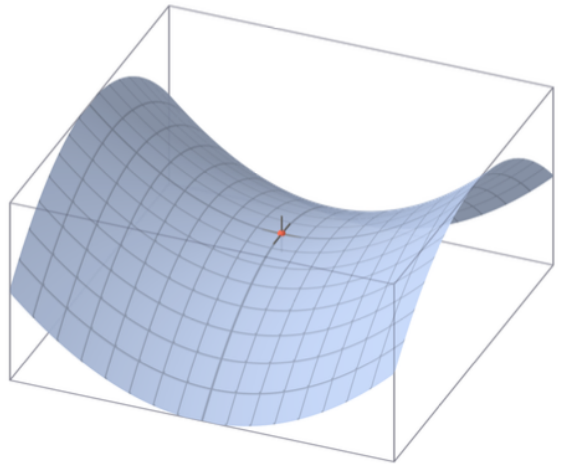
- We already know that at a minimizer x_*, we have $\nabla f(x_*) = 0$, so together

$$\nabla f(x_*) = 0 \text{ and } \nabla^2 f(x_*) \succeq 0$$

  must be true at any local minimizer x_* of f.
- We can limit our search to points with zero gradient, then throw out any points where the Hessian is not positive semidefinite.

Necessary, but not sufficient





- $f(x) = 1 + (x-4)^3$
- $f(x) = x_1^4 - x_2^4$
- $f(x) = -x_1^4 - x_2^4$

Theorem 17 (Second-order sufficient condition)

If $f \in \mathcal{C}^2$, $\nabla f(x_*) = 0$ and $\nabla^2 f(x_*) \succ 0$, then x_* is a strict local minimizer.

Proof.

Since $\nabla^2 f$ is continuous, it remains positive definite near x_*. Taylor's Theorem and $\nabla f(x_*) = 0$ then imply that, for some $\alpha \in (0, 1)$,

$$f(x_* + d) = f(x_*) + \tfrac{1}{2} d^T \nabla^2 f(x_* + \alpha d) d.$$

Hence, f must take larger value at other points near x_*. □

- A nice fact, when we can actually use it!
- By designing algorithms that find a sequence of points with decreasing function values, one hopes that maximizers and saddle points are avoided, i.e., one often focuses on finding a point with zero gradient. That being said, one can search over negative curvature directions to find a point satisfying the second-order necessary conditions, but, in general, a point satisfying the second-order sufficient conditions may not exist.

For a univariate function $f(x)$, the optimality conditions for identifying local extrema can be classified into three sufficient conditions based on the derivatives of the function:

- **First Sufficient Condition**: Suppose $f(x) \in \mathcal{C}^1$. For a $\delta > 0$,
  - If $f'(x) > 0$ on $x \in (x_* - \delta, x_*)$ and $f'(x) < 0$ on $x \in (x_*, x_* + \delta)$, then x_* is a local maximum.
  - If $f'(x) < 0$ on $x \in (x_* - \delta, x_*)$ and $f'(x) > 0$ on $x \in (x_*, x_* + \delta)$, then x_* is a local minimum.
  - If $f'(x)f'(z) > 0$ for $x \in (x_* - \delta, x_*)$ and $z \in (x_*, x_* + \delta)$, then x_* is not a minimum nor a maximum.

- **Second Sufficient Condition**: Suppose $f(x) \in \mathcal{C}^2$. If $f'(x_*) = 0$ and $f''(x_*) \neq 0$, then the following hold.
  - If $f''(x_*) > 0$, then x_* is a local minimum.
  - If $f''(x_*) < 0$, then x_* is a local maximum.

- **Third Sufficient Condition**: If $f(x)$ is continuously differentiable and $f'(x_*) = 0$, and if the first and second derivatives exist:
  - If there exists a neighborhood around x_* where $f'(x)$ does not change sign and $f''(x_*) = 0$, we can look at higher-order derivatives:
    - If $f^{(k)}(x_*) \neq 0$ for the smallest k (where k is odd), then x_* is not a local extremum.
    - If $f^{(k)}(x_*) \neq 0$ for the smallest k (where k is even), then x_* is a local extremum.

Contents

Ancient Origins and Early Developments

- 17th century:
  - Sir Isaac Newton (in 1697) and Joseph Raphson (1690): Newton-Raphson method

- 18th-19th centuries:
  - Euler and Lagrange: method of Lagrange multipliers (Euler-Lagrange equation), fundamental in physics and engineering.
  - Fourier, Cauchy, and Legendre: Gradient descent method.

Early 20th Century: Foundations of Convex Analysis

- Minkowski, Constantin Carathéodory, John von Neumann: convex analysis and duality.
- Marguerite Frank (1927, Harvard) and Philip Wolfe (1927, Berkeley): conditional gradient method for Quadratic Programming (Frank-Wolfe algorithm) in 1956.

Mid-20th Century: Optimization as a Discipline

- Linear Programming
- Gradient Methods, quasi-Newton (BFGS).
- Dynamic Programming: introduced by Bellman in the 1950s.
- Convex Optimization: optimality conditions (KKT) to constrained optimization.

Late 20th Century: Nonlinear Optimization and Computational Advances

- Interior-Point Methods: by Karmarkar (1984), later generalized for nonlinear optimization.
- Global Optimization: simulated annealing (1983) and genetic algorithms (1975).
- Semidefinite and Second-Order Cone Programming (SDP, SOCP): crucial techniques for convex optimization (CVX).

21st Century: High-Dimensional and Nonconvex and Nonsmooth Optimization

- Machine Learning and Deep Learning: Stochastic gradient descent and its variants (Adam, RMSProp, etc.)
- Optimization in Data Science: large-scale optimization, distributed optimization, sparse optimization
- Theory: variational analysis

History of solvers

- 1979: Lindo, Lingo, LINDO Systems Inc by Linus Schrage from University of Chicago.
- 1983: Xpress by Dash Optimization by Bob Daniel and Peter Grant from Edinburgh University, later acquired by FICO in 2008.
- 1987: Cplex by Robert E. Bixby and later commercialized by ILOG in 1988, which was then acquired by IBM in 2009.
- Gurobi by Robert Bixby, Zonghao Gu, and Ed Rothberg in 2008.
- MINOS, Bruce Murtagh and Michael Saunders, Stanford University, 1980s. LP, QP, NLP.
- SNOPT, Philip Gill, Walter Murray, and Michael Saunders, Stanford University and UC San Diego, 1990s. NLP.
- LANCELOT, Nick Gould, Philippe L. Toint, and Alexei Y. Zomitsky, Rutherford Appleton Laboratory (UK) and University of Namur (Belgium), 1980s-1990s.
- IPOPT, Andreas Wächter and Lorenz T. Biegler, Carnegie Mellon University, 2000s. NLP.
- KNITRO, Richard Byrd, Jorge Nocedal, and Todd Plantenga, Northwestern University (later commercialized by Ziena Optimization, which was co-founded by developers). 2000s. NLP.
- FilterSQP, Roger Fletcher and Sven Leyffer, University of Dundee, Late 1990s. NLP

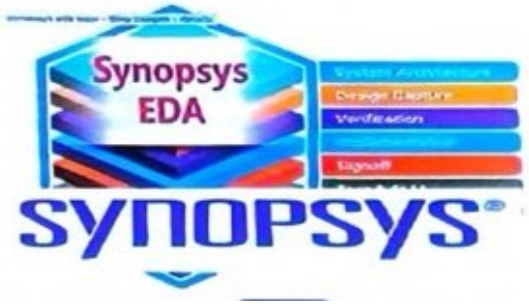

GUROBI
Industry Solution Sheet: Electric Power
Opportunities for Optimization
电网专用优化算法
Gurobi


SpaceX
轨迹控制
SPACEX


aspentech
Aspen PIMS
生产计划
The world's most trusted production planning software just got easier and faster – now with industrial AI.


SAP+Gurobi
供应链
SAP
Cloud Inventory Optimization, Demand Planning, Analytics, and Forecasting for SAP


DASSAULT
达索 CAD/CAM
DASSAULT SYSTEMES


Synopsys
EDA
SYNOPSYS

Contents

Algorithms for nonlinear optimization focus on stationary points, i.e., x_* with

$$\nabla f(x_*) = 0$$

This is a system of nonlinear equations (but solving it isn't our only goal)

- A stationary point may be a minimizer, but it is not guaranteed
- Thus, we try to bias the iteration toward minimizers by requiring

$$f(x_{k+1}) < f(x_k) \quad \text{for all} \quad k \in \mathbb{N}_+.$$

- A basic Newton method for nonlinear equations has iterate update:

$$x_{k+1} \leftarrow x_k + d_k.$$

- Steps are taken based on solely on local approximation information
- There is no guarantee that the next iterate is closer to a solution
- Iterations like this need to be modified to ensure global convergence
- Line search philosophy: compute d_k and then compute $\alpha_k > 0$ so that

$$x_{k+1} \leftarrow x_k + \alpha_k d_k$$

  is "better" than x_k in some way.

Considering an unconstrained optimization problem

$$\min_x f(x),$$

the search direction d_k should fulfill two requirements:

- It should "move" toward satisfying $\nabla f(x_k) \to 0$
- It should be a direction of descent, i.e., it should satisfy

$$\nabla f(x_k)^T d < 0$$

so that we can guarantee, for some $\alpha_k > 0$,

$$f(x_k + \alpha d_k) < f(x_k)$$

Common choices for d_k

- The simplest choice is the steepest descent direction (gradient descent)

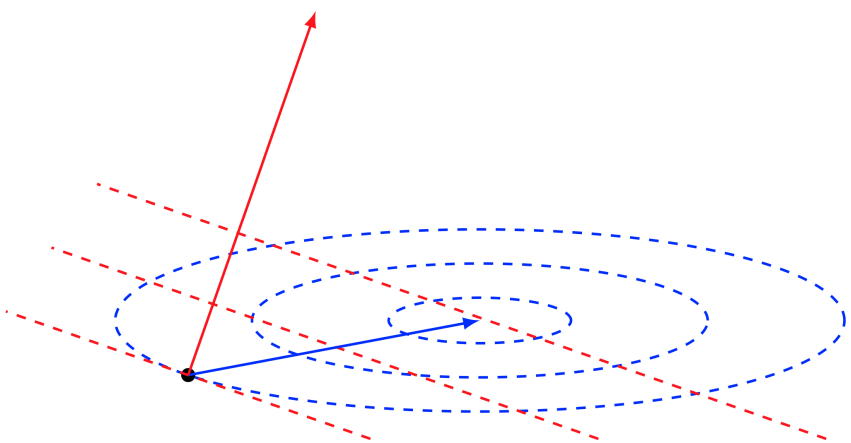
$$d_k = -\nabla f(x_k).$$

- Another popular choice is the Newton direction

$$d_k = -\nabla^2 f(x_k)^{-1} \nabla f(x_k).$$

- Some approximation of Newton direction

$$d_k = -H_k^{-1} \nabla f(x_k).$$



- Given a descent direction d_k, we try to find α_k so that we:
- at least ensure that for $x_{k+1} \leftarrow x_k + \alpha_k d_k$, we have $f(x_{k+1}) < f(x_k)$
- at most solve the one-dimensional (nonlinear) minimization problem

$$\min_{\alpha \geq 0} f(x_k + \alpha d_k).$$

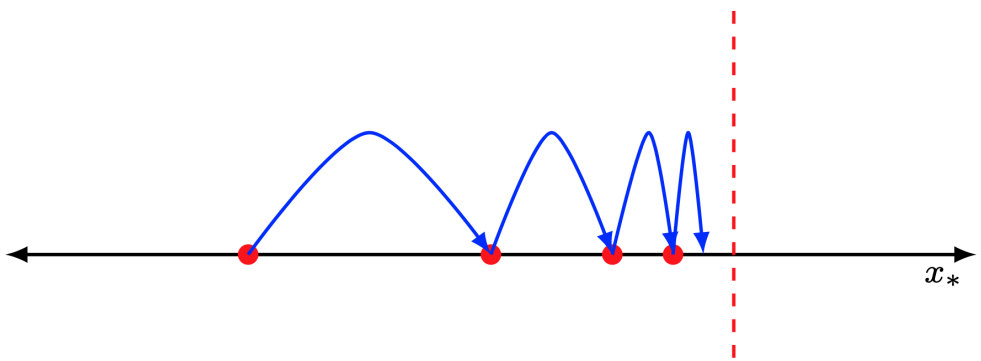
- Commonly, we choose α_k to satisfy conditions between these two extremes.

Why isn't it enough to simply compute a descent direction so that

$$\nabla f(x_k)^T d_k < 0$$

and then choose $\alpha_k > 0$ so that

$$f(x_{k+1}) < f(x_k)?$$

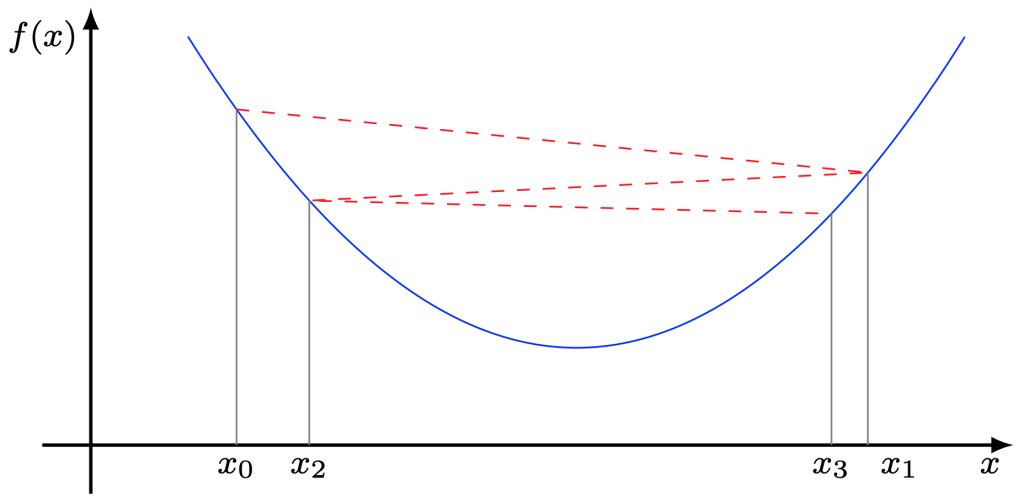


Why isn't it enough to simply compute a descent direction so that

$$\nabla f(x_k)^T d_k < 0$$

and then choose $\alpha_k > 0$ so that

$$f(x_{k+1}) < f(x_k)?$$

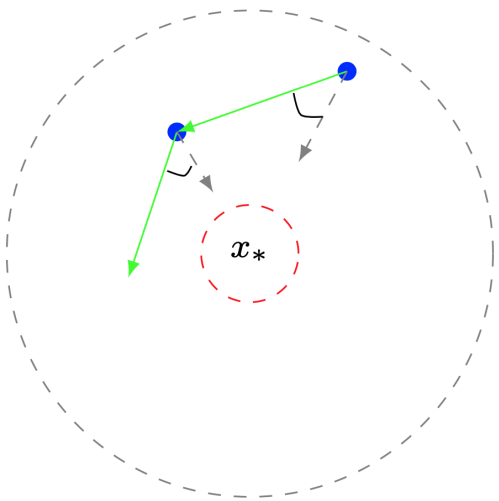


Why isn't it enough to simply compute a descent direction so that

$$\nabla f(x_k)^T d_k < 0$$

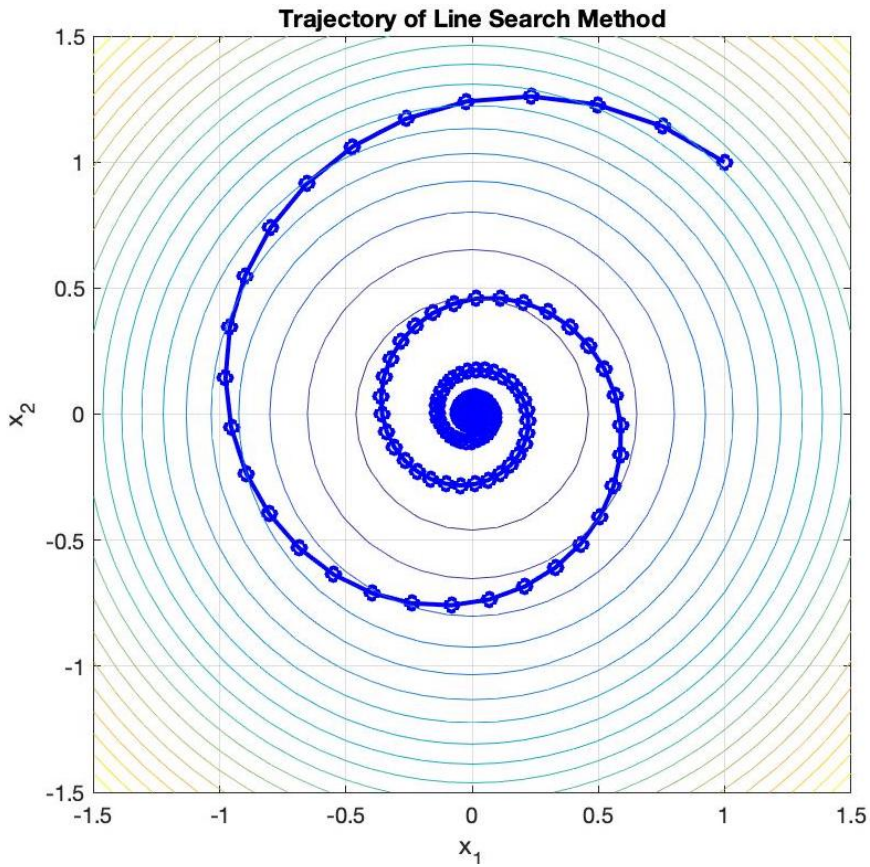
and then choose $\alpha_k > 0$ so that

$$f(x_{k+1}) < f(x_k)?$$

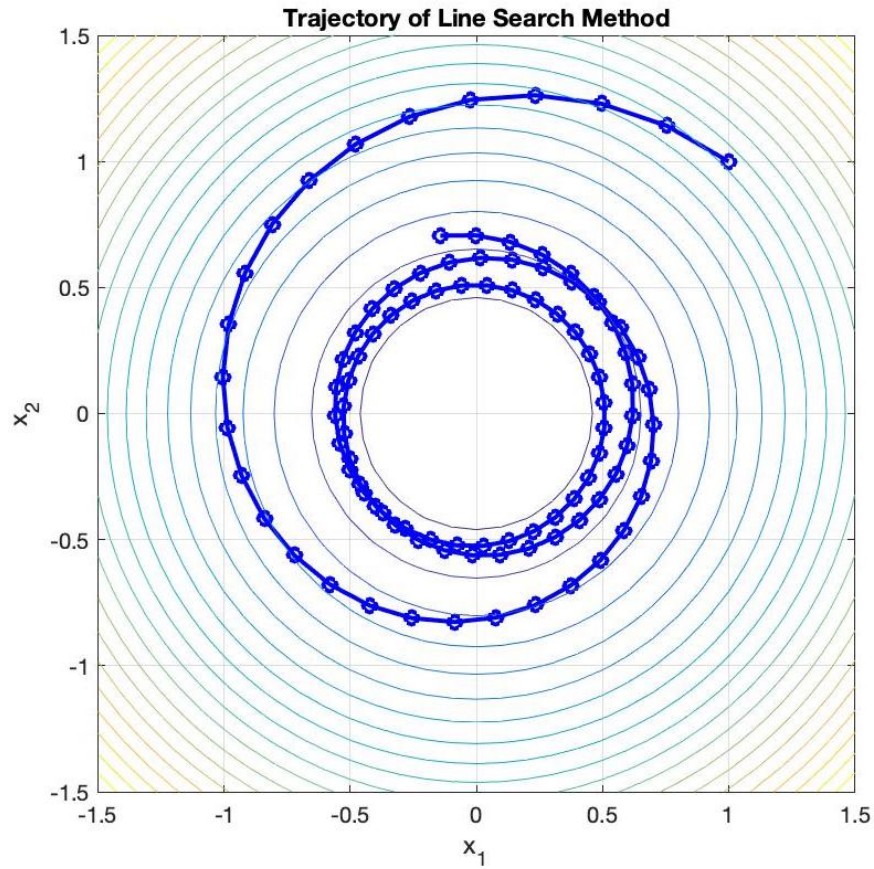


Consider minimizing $f(x) = x_1^2 + x_2^2$ and starting at $(0, 0)$. Direction is

$$d_k = -\begin{bmatrix} \cos\theta_k & \sin\theta_k \\ -\sin\theta_k & \cos\theta_k \end{bmatrix} \nabla f(x_k),\ \alpha = 0.1.$$



$\theta_k = 75^\circ + (90^\circ - 75^\circ)/1500$



$\theta_k = 75^\circ + (90^\circ - 75^\circ)/100$

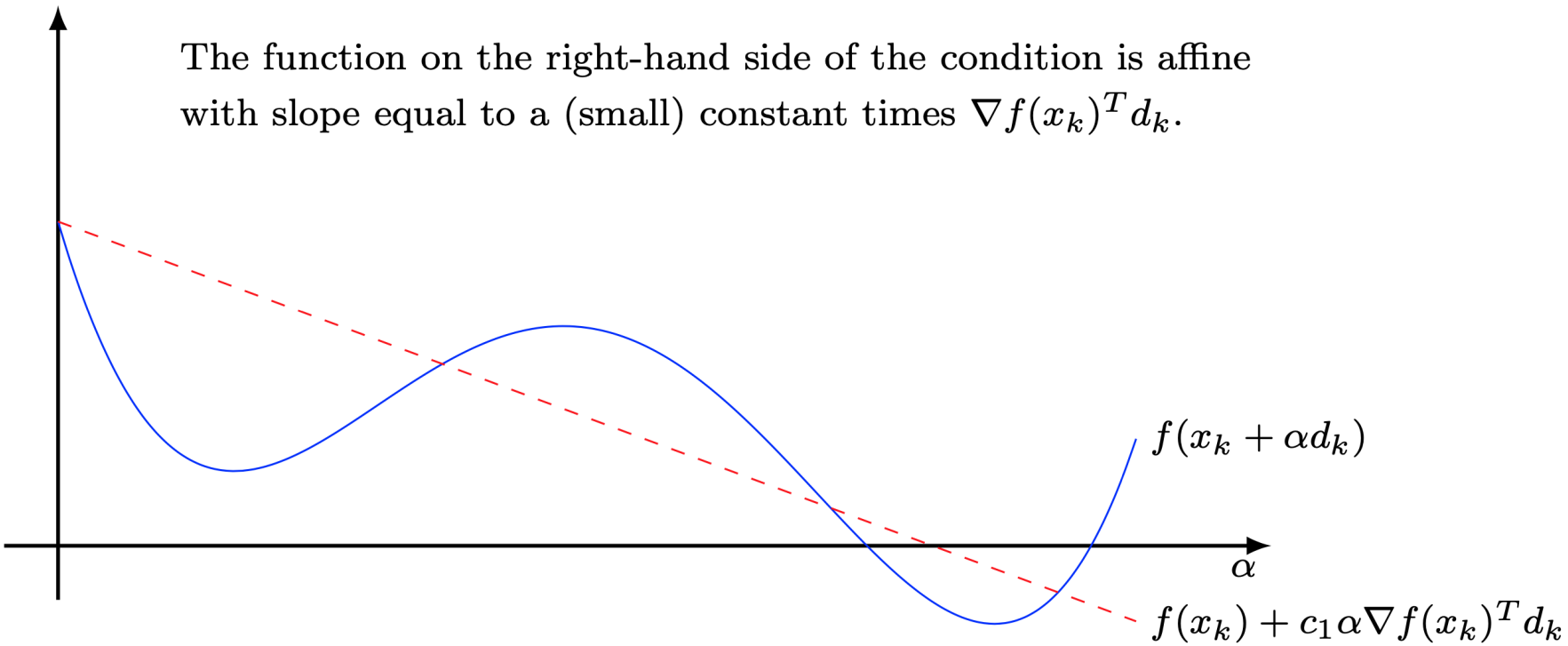
Perhaps the most common condition to place on α_k is the following sufficient decrease (also known as the Armijo) condition:

$$f(x_k + \alpha_k d_k) \leq f(x_k) + c_1 \alpha_k \nabla f(x_k)^T d_k,$$

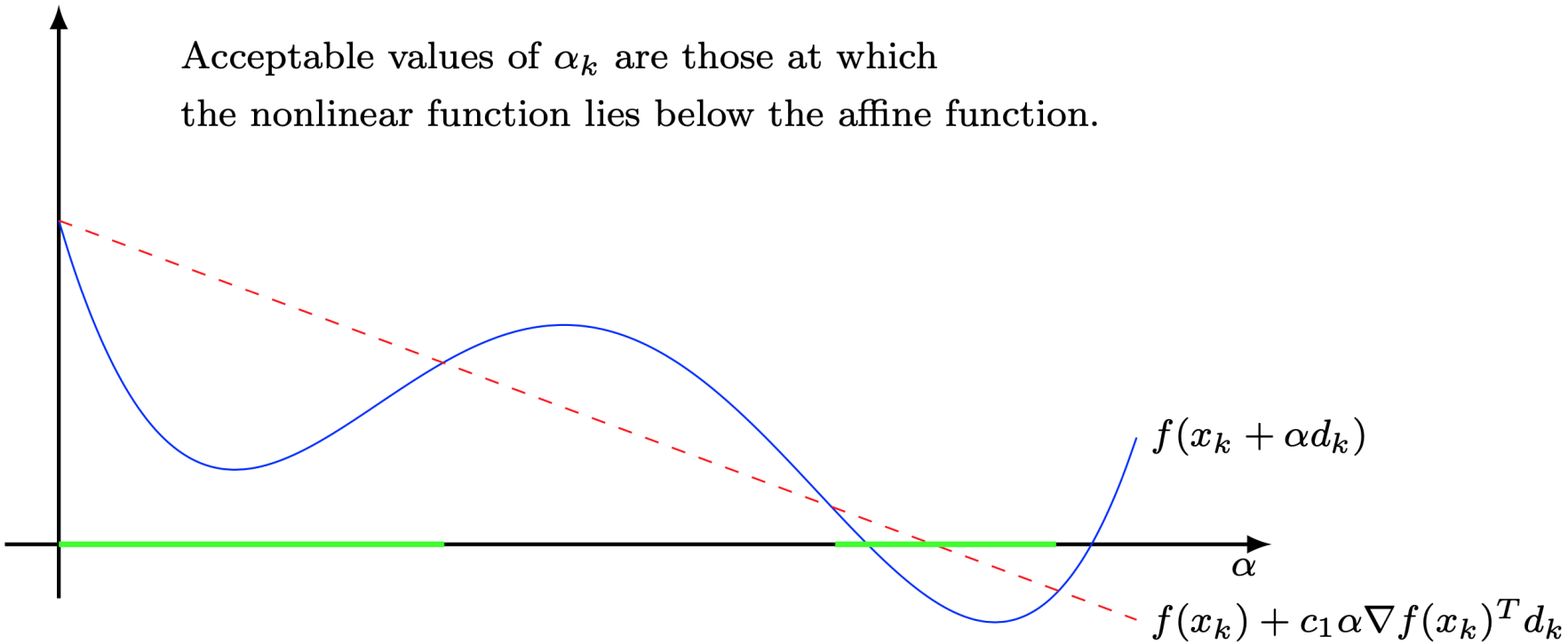
where $c_1 \in (0, 1)$ is a user-specified constant

- $c_1 = 0$ is too loose of a requirement
- $c_1 = 1$ is too strict, and may not be satisfiable if curvature is strictly positive.

Sufficient decrease: $f(x_k + \alpha_k d_k) \leq f(x_k) + c_1 \alpha_k \nabla f(x_k)^T d_k$



Sufficient decrease: $f(x_k + \alpha_k d_k) \leq f(x_k) + c_1 \alpha_k \nabla f(x_k)^T d_k$



- The Armijo condition is not enough by itself. ($\alpha_k = 0$ satisfies it!)
- There are generally two ways to "add" to the Armijo condition:
  1. Algorithmically, choose the largest value in the set

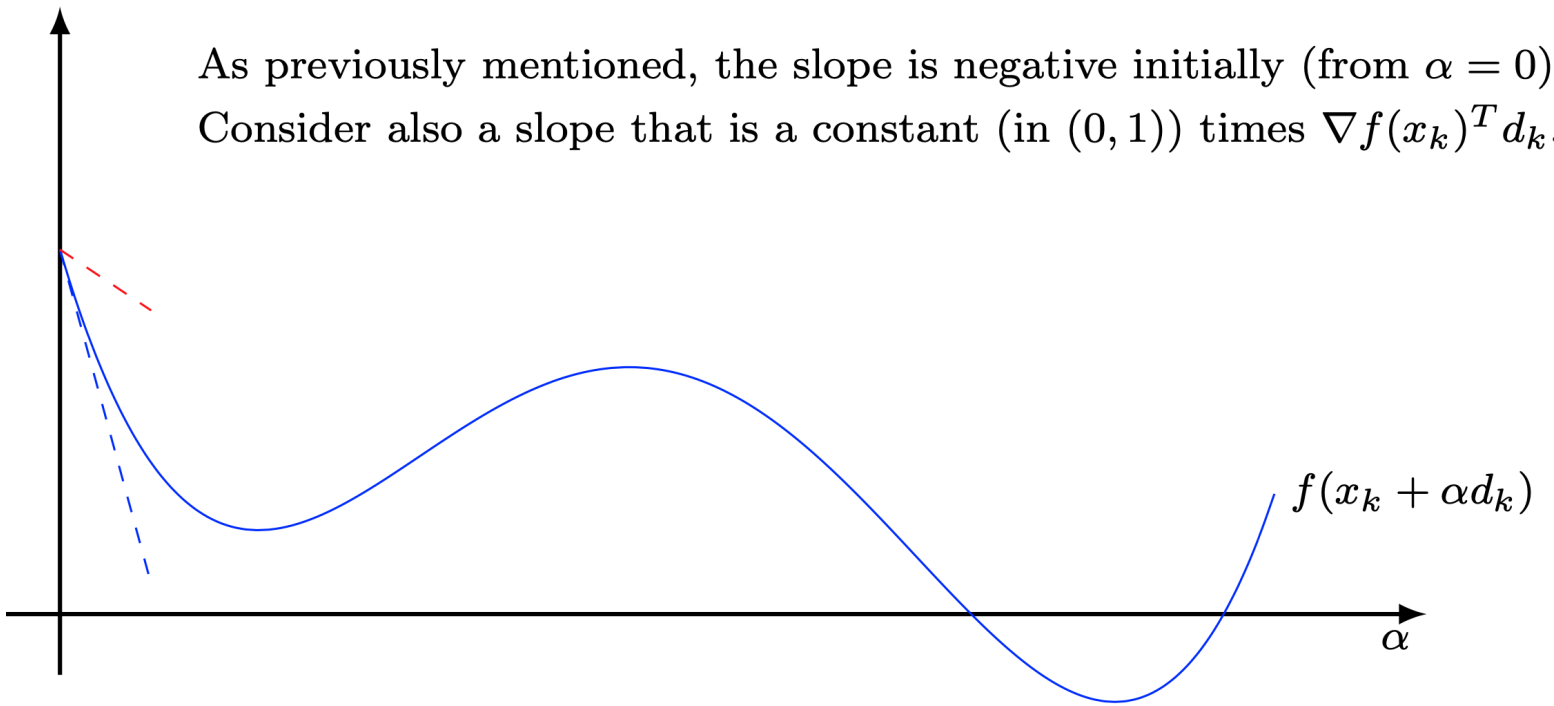
     $$\{\gamma^0, \gamma^1, \gamma^2, ...\}$$

     where $\gamma \in (0, 1)$ is a given constant, satsifying the Armijo condition. This is referred to as a backtracking line search
  2. Formulate an additional condition to ensure a productive step. For example, a popular choice is the curvature condition:

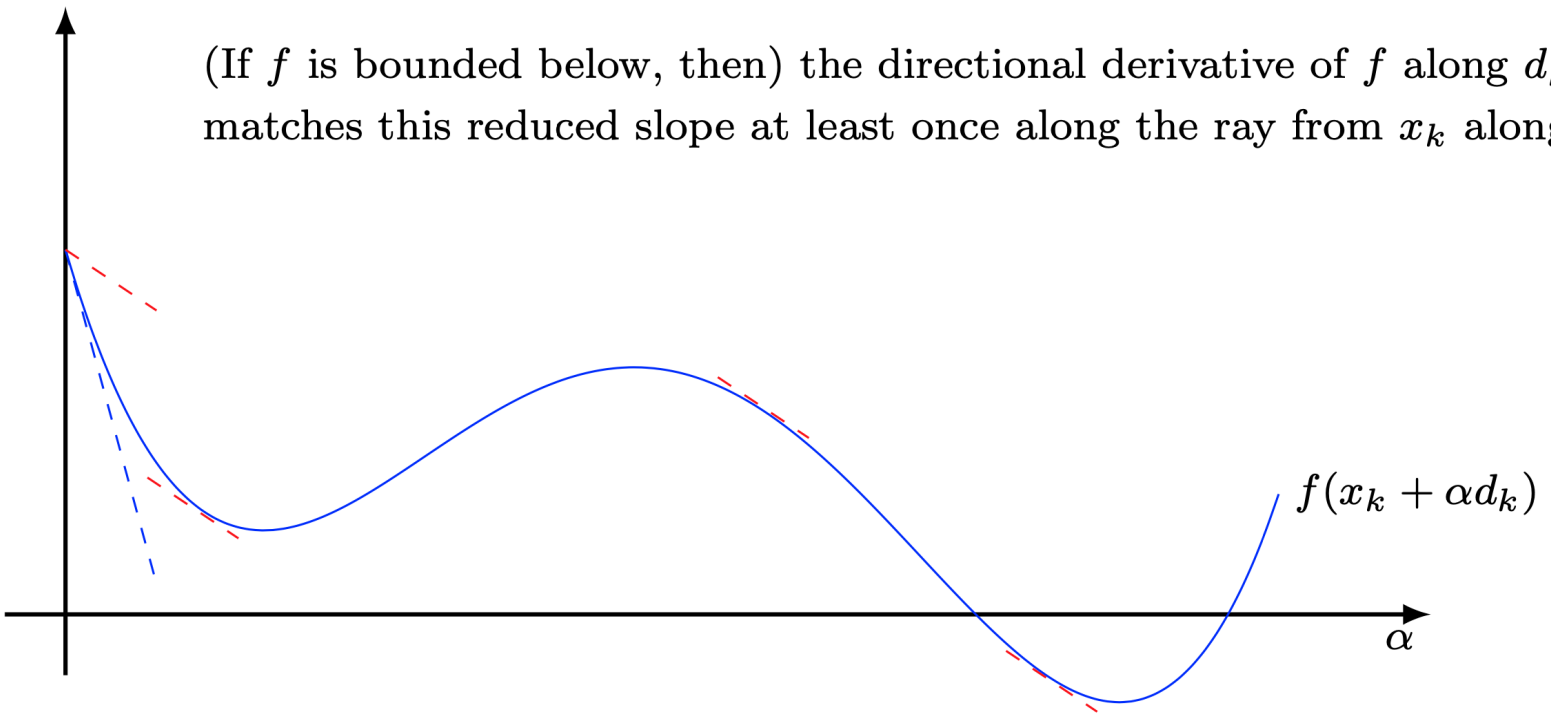
     $$\nabla f(x_k + \alpha_k d_k)^T d_k \geq c_2 \nabla f(x_k)^T d_k,$$

     where $c_2 \in (c_1, 1)$ is a user-specified constant.

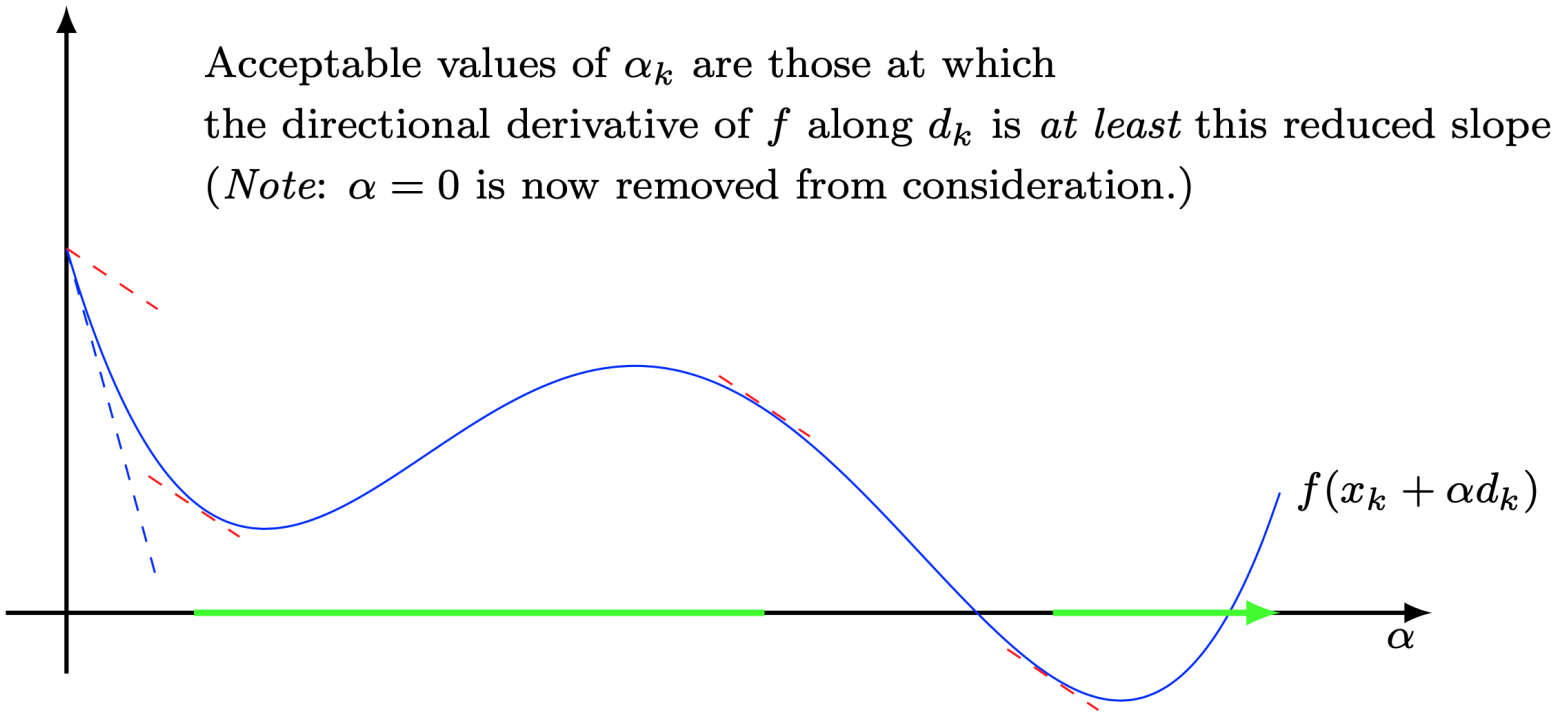
Curvature condition illustrated: $\nabla f(x_k + \alpha_k d_k)^T d_k \geq c_2 \nabla f(x_k)^T d_k$



Curvature condition illustrated: $\nabla f(x_k + \alpha_k d_k)^T d_k \geq c_2 \nabla f(x_k)^T d_k$



Curvature condition illustrated: $\nabla f(x_k + \alpha_k d_k)^T d_k \geq c_2 \nabla f(x_k)^T d_k$



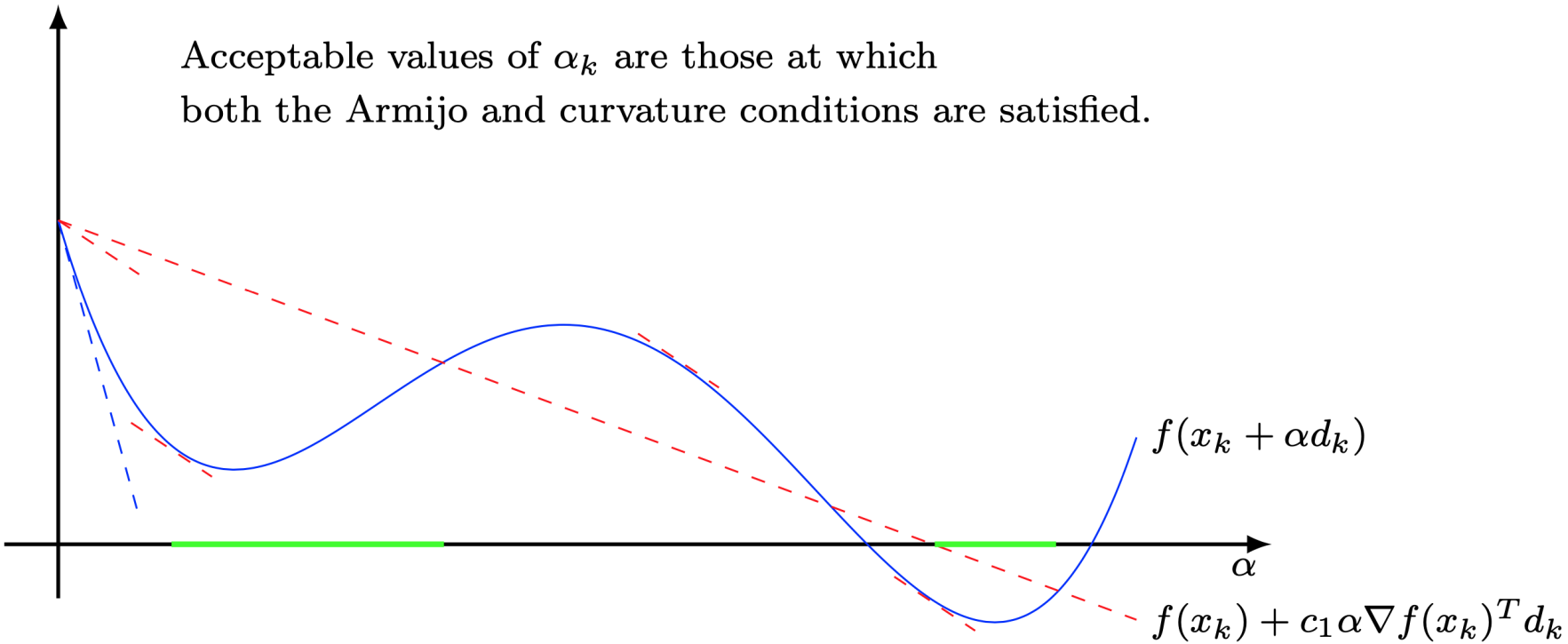
- The Armijo and curvature conditions together compose the Wolfe conditions:

$$\begin{aligned} f(x_k + \alpha_k d_k) &\leq f(x_k) + c_1 \alpha_k \nabla f(x_k)^T d_k, \quad && c_1 \in (0, 1); \\ \nabla f(x_k + \alpha_k d_k)^T d_k &\geq c_2 \nabla f(x_k)^T d_k, && c_2 \in (c_1, 1). \end{aligned}$$

  See Lemma 3.1 in the textbook[1] to see why we need $c_2 > c_1$.

[1] Numerical Optimization

Wolfe conditions illustrated



- Along with the theorem of the quadratic convergence of Newton's method, the other absolutely fundamental result we will cover is the one we consider next. (Indeed, the assumptions we make are similar to those for Newton's method.)
- Within the theorem and the proof, the critical properties of the search direction d_k, the steplength α_k, and the line search conditions are revealed.
- Recall that the angle θ_k between d_k and $-\nabla f(x_k)$ is defined by

$$\cos\theta_k = \frac{-\nabla f(x_k)^T d_k}{\|\nabla f(x_k)\|\|d_k\|}.$$

Theorem 18 (Zoutendijk's Theorem)

Suppose that f is bounded below and continuously differentiable in an open set $\mathcal{N}$ containing the sublevel set $\mathcal{L} := \{x \mid f(x) \leq f(x_0)\}$. Suppose also that ∇f is Lipschitz continuous on $\mathcal{N}$ with constant L. Consider any iteration of the form

$$x_{k+1} \leftarrow x_k + \alpha_k d_k \quad \text{for all } k \in \mathbb{N}_+,$$

where, for all k,

- *d_k is a descent direction, and*
- *α_k satisfies the Wolfe conditions.*

Then,

$$\sum_{k=0}^{\infty} \cos^2 \theta_k \|\nabla f(x_k)\|^2 < \infty.$$

Proof, part 1

First, we show that the curvature condition and Lipschitz continuity of the gradient imply that we have a lower bound on α_k.

- The curvature condition can be rewritten as

$$(\nabla f(x_{k+1}) - \nabla f(x_k))^T d_k \geq (c_2 - 1)\nabla f(x_k)^T d_k.$$

- Lipschitz continuity implies that we have

$$(\nabla f(x_{k+1}) - \nabla f(x_k))^T d_k \leq \alpha_k L \|d_k\|^2.$$

- Thus, together we find (note that $c_2 - 1 < 0$ and $\nabla f(x_k)^T d < 0$)

$$\alpha_k L \|d_k\|^2 \geq (c_2 - 1)\nabla f(x_k)^T d_k \implies \alpha_k \geq \frac{(c_2 - 1)\nabla f(x_k)^T d_k}{L\|d_k\|^2}.$$

What can affect the stepsize?

Proof, part 2

Second, we show that due to the sufficient decrease condition and our lower bound on α_k, each iteration reduces $f(x)$ monotonically.

- From the previous slide, we have

$$\alpha_k \geq \frac{(c_2 - 1)\nabla f(x_k)^T d_k}{L\|d_k\|^2}.$$

- Substituting this expression in for α_k in the Armijo condition, we find

$$\begin{aligned} f(x_{k+1}) &\leq f(x_k) + c_1 \alpha_k \nabla f(x_k)^T d_k \\ &\leq f(x_k) + \frac{c_1(c_2 - 1)(\nabla f(x_k)^T d_k)^2}{L\|d_k\|^2} \\ &= f(x_k) - c\cos^2\theta_k \|\nabla f(x_k)\|^2, \end{aligned}$$

where $c = c_1(1 - c_2)/L > 0$ is a (not necessarily known) constant.

Intuitively, notice that when is the reduction in $f(x)$ large?

Proof, part 3

Finally, we show that since f is bounded below, the reductions in f get squeezed down to zero over the course of the optimization.

- We have shown already that

$$f(x_{k+1}) \leq f(x_k) - c\cos^2\theta_k\|\nabla f(x_k)\|^2$$

- Summing over iterations $k = 0, 1, 2, ..., K$, we obtain

$$f(x_{k+1}) \leq f(x_k) - c\sum_{k=0}^{K}\cos^2\theta_k\|\nabla f(x_k)\|^2$$

- Thus, since f is bounded below,

$$\cos^2\theta_k\|\nabla f(x_k)\|^2 \to 0.$$

- What we really want is $\|\nabla f(x_k)\| \to 0$.
- What we have proved is

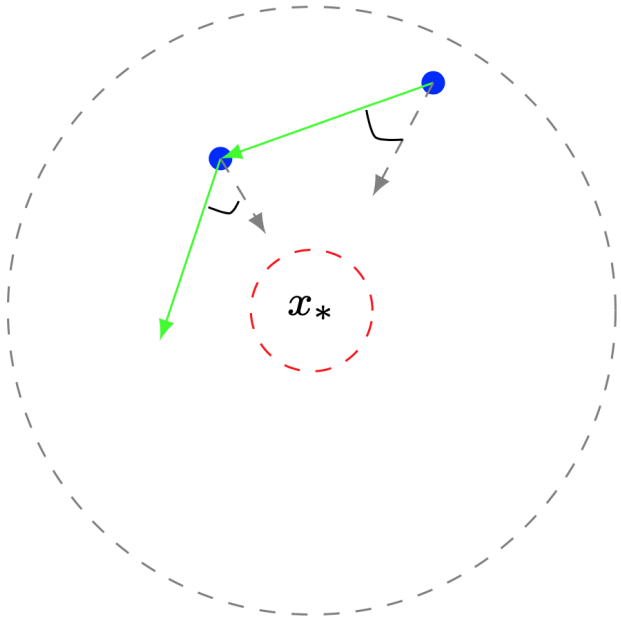
$$\cos^2 \theta_k \|\nabla f(x_k)\|^2 \to 0$$

- The latter does not imply the former.
- This means that simply having a descent direction is not enough!
- What can we do?
- If we can guarantee that eventually the angle between d_k and $-\nabla f(x_k)$ is bounded away from 90°, i.e.,

$$\cos \theta_k \geq \delta > 0$$

for all sufficiently large k, then we immediately have $\|\nabla f(x_k)\| \to 0$

Implications of the angle between d_k and $-\nabla f(x_k)$



Consider the following iteration from x_0:

1. Let $d_k = -\nabla f(x_k)$
2. Compute α_k satisfying the Wolfe conditions.
3. Update $x_{k+1} \leftarrow x_k + \alpha_k d_k$, return to step 1.

- Obviously, the angle between d_k and $-\nabla f(x_k)$ is always $0°$, so

$$\cos\theta_k = 1 > 0 \quad \text{for all } k.$$

- Thus, we have global convergence due to Zoutendijk's theorem.

Strong convexity of f means for some $\mu > 0$,

$$\nabla^2 f(x) \succeq \mu I \quad \text{for any } x$$

Better lower bound than that from usual convexity

$$f(y) \geq f(x) + \nabla f(x)^T (y - x) + \frac{\mu}{2}\|y - x\|^2 \quad \forall x, y$$

Under Lipschitz assumption as before, and also strong convexity

Theorem 19

For L-smooth and μ-strongly convex f, Gradient descent with fixed step size $\alpha \leq 2/(\mu + L)$ satisfies

$$f(x^k) - f(x^*) \leq \left(\frac{L/\mu - 1}{L/\mu + 1}\right)^{2k} \frac{L}{2}\|x^0 - x^*\|^2.$$

Since $\exp(-x) \geq 1 - x$ for every x, we get:

$$\left(\frac{L/\mu - 1}{L/\mu + 1}\right)^2 = \left(1 - \frac{4L\mu}{(L+\mu)^2}\right) \implies \left(\frac{L/\mu - 1}{L/\mu + 1}\right)^2 \leq \exp\left(-\frac{4L\mu}{(L+\mu)^2}\right)$$

$$\implies f(x^k) - f(x^*) \leq \frac{L}{2} \exp\left(-\frac{4L\mu k}{(L+\mu)^2}\right) \|x^0 - x^*\|^2$$

i.e., rate with strong convexity is exponentially fast!

i.e., to get $f(x^k) - f(x^*) \leq \epsilon$, need $O(\log(1/\epsilon))$ iterations

Called linear convergence, because looks linear on a semi-log plot:
Rate constant depends adversely on condition number L/μ (higher condition number) $\implies$ slower rate)

Lemma 20

Let f be L-smooth and μ-strongly convex. Then for all $x, y \in \mathbb{R}^n$, one has

$$(\nabla f(x) - \nabla f(y))^T (x - y) \geq \frac{\mu L}{\mu + L} \|x - y\|^2 + \frac{1}{\mu + L} \|\nabla f(x) - \nabla f(y)\|^2.$$

Proof.

Note that $\phi(x) := f(x) - \frac{\mu}{2}\|x\|^2$ is convex and $(L - \mu)$-smooth, so that (using a previous lemma)

$$[\nabla \phi(x) - \nabla \phi(y)]^T (x - y) \geq \frac{1}{L - \mu} \|\nabla \phi(x) - \nabla \phi(y)\|^2$$

implying

$$[\nabla f(x) - \mu x - (\nabla f(y) - \mu y)]^T (x - y) \geq \frac{1}{L - \mu} \|\nabla f(x) - \mu x - (\nabla f(y) - \mu y)\|^2.$$

Rearranging we have the desired result.

Key Steps of the Proof

Now we are ready to prove the linear convergence of GD. Indeed recall that smoothness can be defined via the inequality:

$$f(x) - f(y) \leq \nabla f(y)^T (x - y) + \frac{L}{2}\|x - y\|^2,$$

$$\implies f(x^k) - f(x^*) \leq \nabla f(x^*)^T (x^k - x^*) + \frac{L}{2}\|x^k - x^*\|^2$$

By strongly convexity,

$$\begin{aligned}
\|x^k - x^*\|^2 &= \|x^{k-1} - \alpha \nabla f(x^{k-1}) - x^*\|^2 \\
&= \|x^{k-1} - x^*\|^2 - 2\alpha \nabla f(x^{k-1})^T (x^{k-1} - x^*) + \alpha^2 \|\nabla f(x^{k-1})\|^2 \\
&\leq \left(1 - 2\frac{\alpha\mu L}{L+\mu}\right) \|x^{k-1} - x^*\|^2 + \left(\alpha^2 - 2\frac{\alpha}{L+\mu}\right) \|\nabla f(x^{k-1}) - \nabla f(x^*)\|^2 \\
&\leq \left(1 - 2\frac{\alpha\mu L}{L+\mu}\right) \|x^{k-1} - x^*\|^2 + \left(\alpha^2 - 2\frac{\alpha}{L+\mu}\right) L^2 \|x^{k-1} - x^*\|^2 \\
&= \left(\frac{L + \mu - 2\alpha L\mu + \alpha^2 L^2 (L+\mu) - 2\alpha L^2}{L+\mu}\right) \|x^{k-1} - x^*\|^2 \\
&= \left(\frac{\alpha^2 L^2 (L+\mu) - 2\alpha L(L+\mu) + L + \mu}{L+\mu}\right) \|x^{k-1} - x^*\|^2 \\
&= \left(\alpha^2 L^2 - 2\alpha L + 1\right) \|x^{k-1} - x^*\|^2 \\
&= (\alpha L - 1)^2 \|x^{k-1} - x^*\|^2 = 0, \quad \text{if we set } \alpha = 1/L, \text{ so what is wrong??}
\end{aligned}$$

- If $0 \leq \alpha \leq \frac{2}{L+\mu}$,

$$\|x^k - x^*\|^2 \leq \left(1 - \frac{2\alpha\mu L}{L+\mu} + (\alpha^2 - \frac{2\alpha}{L+\mu})L^2\right)\|x^{k-1} - x^*\|^2 = (\alpha L - 1)^2\|x^k - x^*\|^2$$

$$\implies \alpha^* = \min(2/(L+\mu), 1/L) = 2/(L+\mu)$$

- If $\alpha \geq \frac{2}{L+\mu}$,

$$\|x^k - x^*\|^2 \leq \left(1 - \frac{2\alpha\mu L}{L+\mu}\right)\|x^{k-1} - x^*\|^2$$

$$\implies \alpha^* = 2/(L+\mu)$$

-

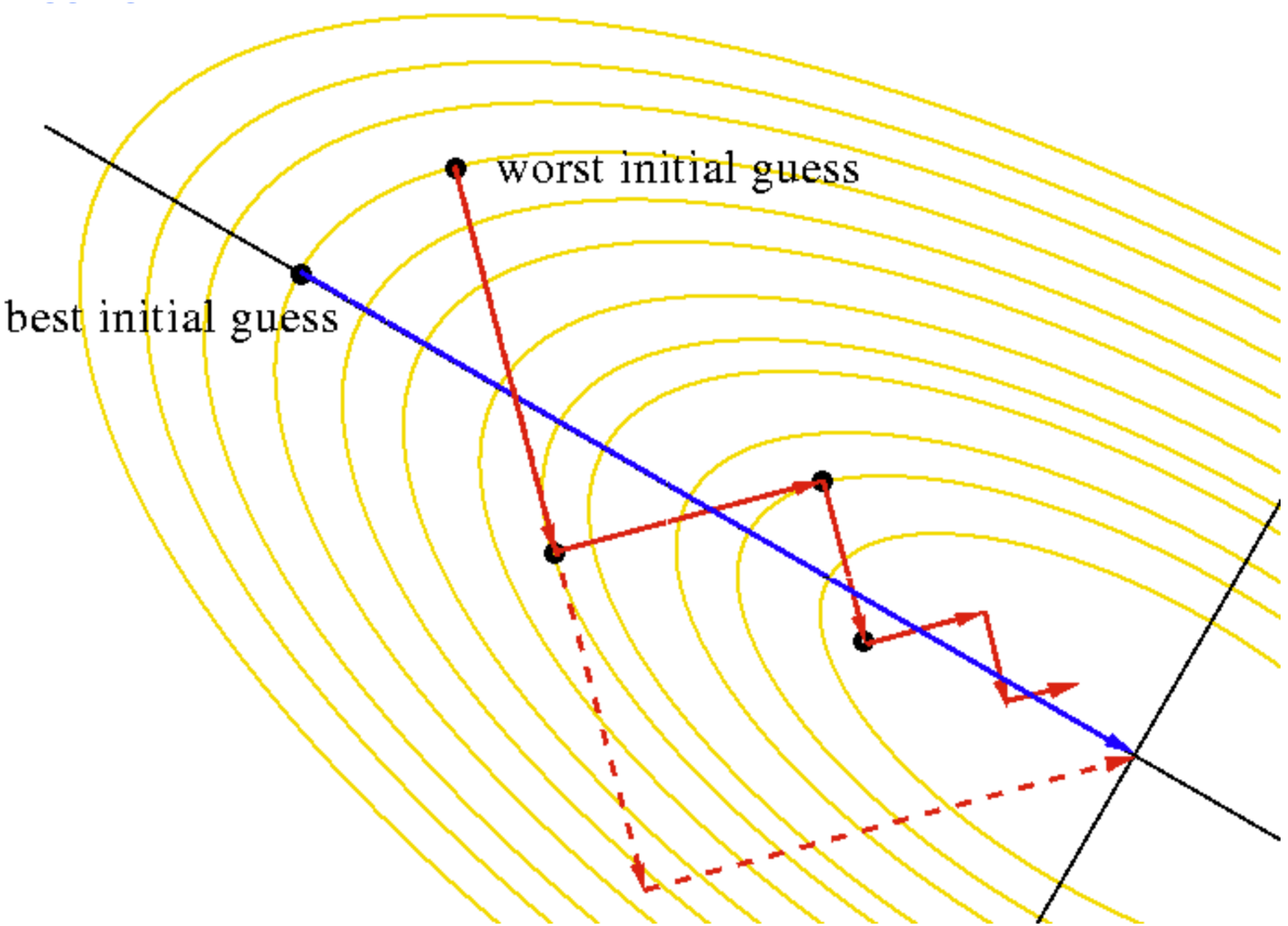
$$\implies \|x^k - x^*\|^2 \leq \left(\frac{L/\mu - 1}{L/\mu + 1}\right)^2 \|x^{k-1} - x^*\|^2 \quad \text{with } \alpha = 2/(L+\mu)$$

$$\implies \|x^k - x^*\|^2 \leq \left(\frac{L/\mu - 1}{L/\mu + 1}\right)^{2k} \|x^0 - x^*\|^2.$$

$$\implies f(x^k) - f(x^*) \leq \frac{L}{2}\left(\frac{L/\mu - 1}{L/\mu + 1}\right)^{2k} \|x^0 - x^*\|^2$$

- The number L/μ is called the condition number of f.

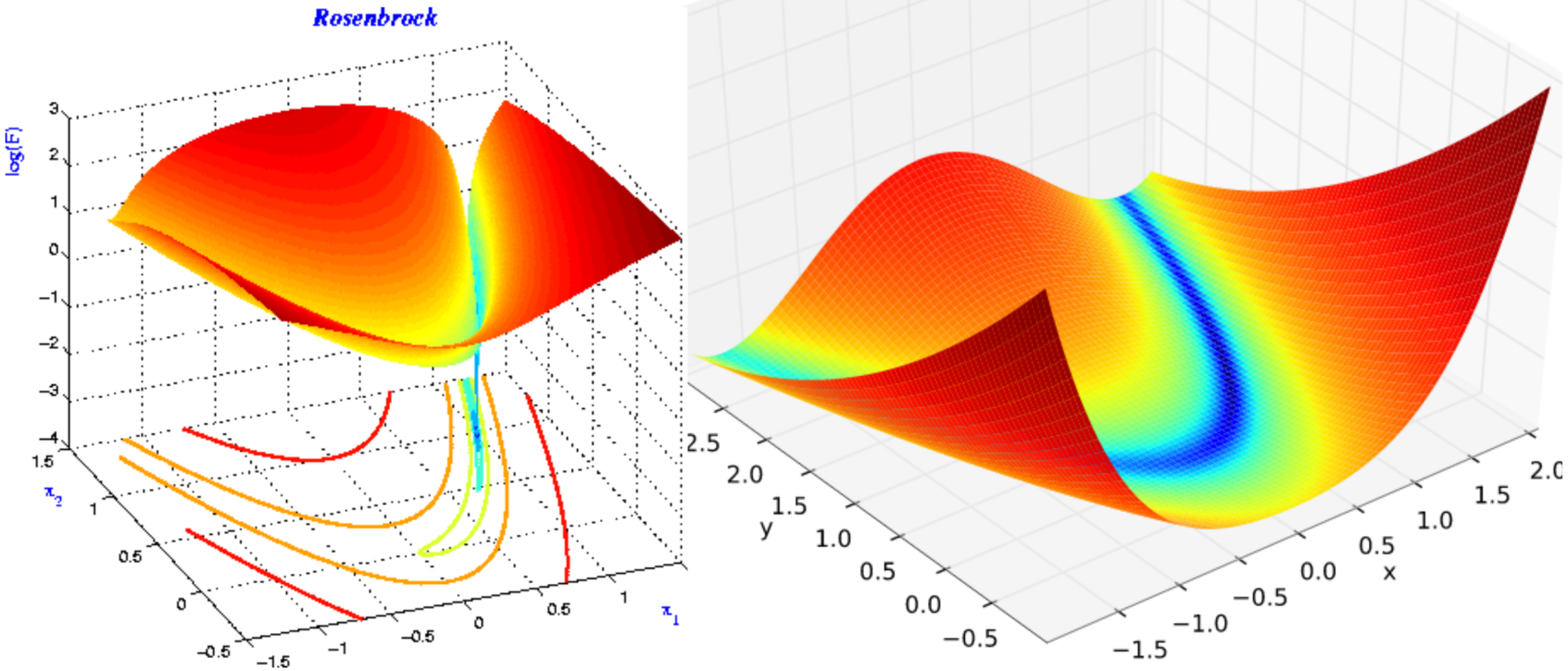
Zigzagging



Rosenbrock function

$$f(x) = \sum_{i=1}^{n/2} \left[\alpha(x_{2i} - x_{2i-1}^2)^2 + (1 - x_{2i-1})^2\right]$$

$$f(x) = 100(x_2 - x_1^2)^2 + (1 - x_1)^2$$



Rosenbrock function

steepest descent	BFGS	Newton
1.827e-04	1.70e-03	3.48e-02
1.826e-04	1.17e-03	1.44e-02
1.824e-04	1.34e-04	1.82e-04
1.823e-04	1.01e-06	1.17e-08