Numerical Optimization

Lecture 12: Line Search Method (线搜索方法)

王浩

信息科学与技术学院

Email: wanghao1@shanghaitech.edu.cn

Optimality conditions and stationary points

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 this is not guaranteed.
- Thus, we try to bias the iteration toward minimizers by requiring

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

  (We may converge to a non-minimizer, but at least the function has reduced.)
- Some algorithms guarantee that they converge to a point satisfying second-order conditions, but this isn't typical and requires more work.

Iterate updates

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

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

- Steps are taken based solely on local approximate information.
- There is no guarantee 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.

Role of d_k

Considering an unconstrained optimization problem

$$\min_x \ f(x),$$

the search direction d_k should fulfill two requirements:

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

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

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

$$f(x_k + \alpha_k d_k) < f(x_k).$$

Note:

- If f is bounded below, then $f(x_k) \to f(x_*)$ should mean $\nabla f(x_k) \to 0$. Thus, we can focus on 2 and 1 should follow.

Common choices for d_k (part 1)

- The simplest choice is the steepest descent direction

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

(Recall the proof of the first-order optimality conditions.)

- Another popular choice is the Newton direction

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

This direction is the solution to minimizing a local quadratic model

$$m_k(d) = f(x_k) + \nabla f(x_k)^T d + \tfrac{1}{2} d^T \nabla^2 f(x_k) d,$$

if we can assume that $\nabla^2 f(x_k) \succeq 0$. (More on this later.)

- Why Newton direction? It is the solution to the linear system

$$\nabla^2 f(x_k) d = -\nabla f(x_k)$$

which arises in a Newton iteration for the system of equations $\nabla f(x) = 0$.

Common choices for d_k (part 2)

- Building on the idea of a Newton-like iteration, we may choose

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

where $H_k \succeq 0$ approximates $\nabla^2 f(x_k)$. This is a quasi-Newton direction.

- Finally, we will also see another studied type of direction:

$$d_k = -\nabla f(x_k) + \beta_k d_{k-1},$$

where β_k is a scalar that ensures d_k and d_{k-1} are conjugate.

Role of α_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.

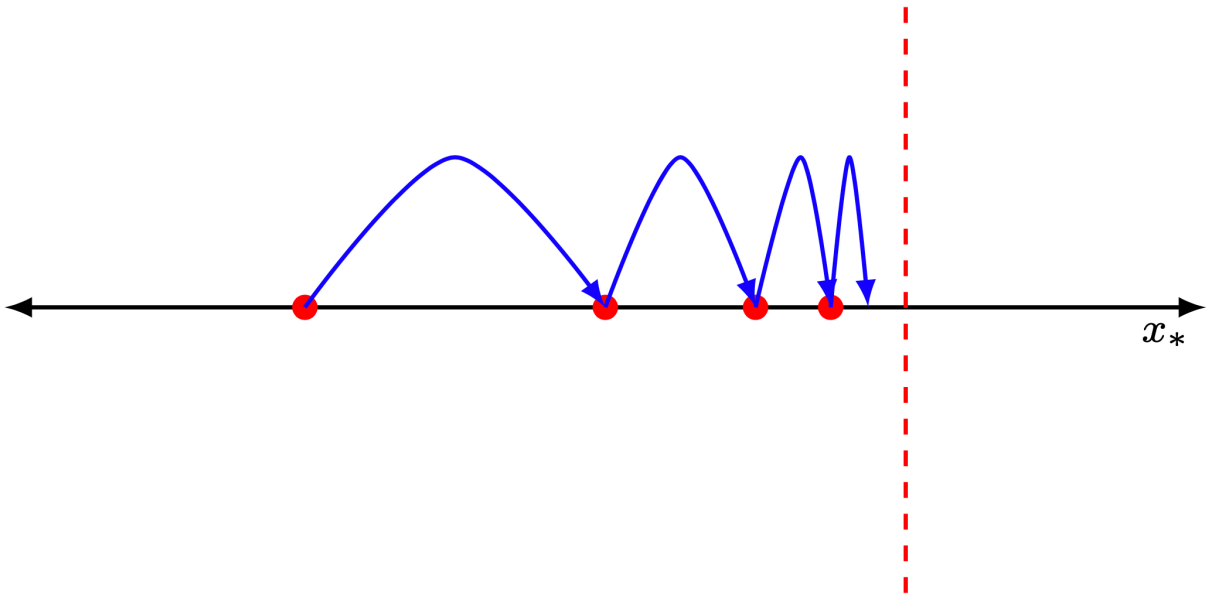
Motivating illustrations

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

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

and then choose $\alpha_k \geq 0$ so that

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



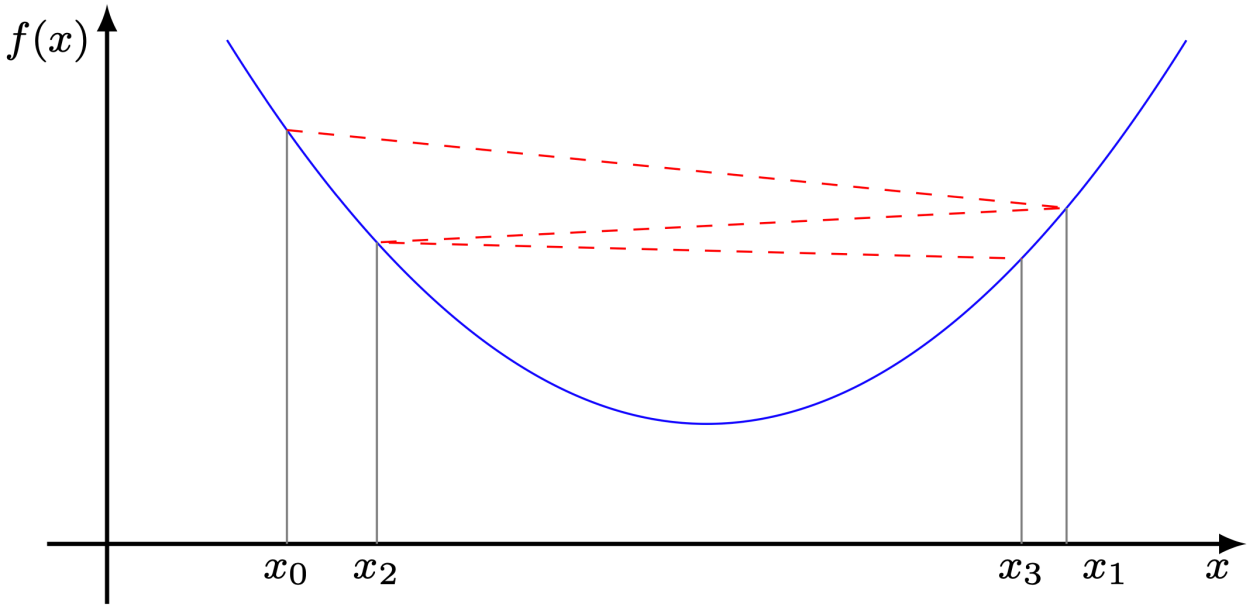
Motivating illustrations

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

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

and then choose $\alpha_k \geq 0$ so that

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



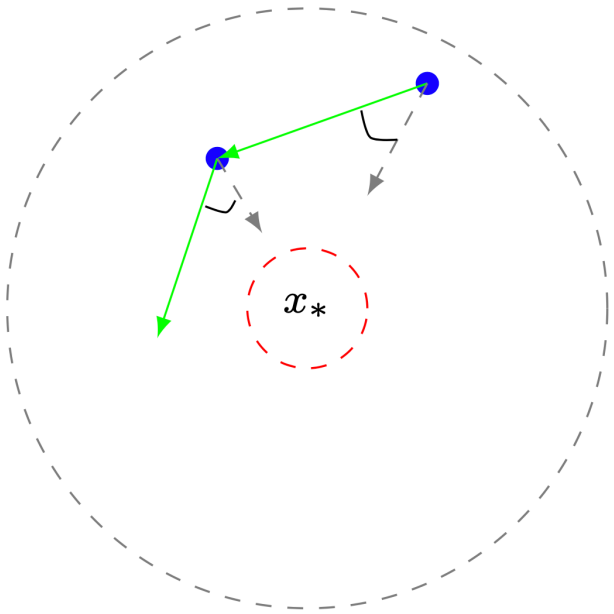
Motivating illustrations

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

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

and then choose $\alpha_k \geq 0$ so that

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



Sufficient decrease condition

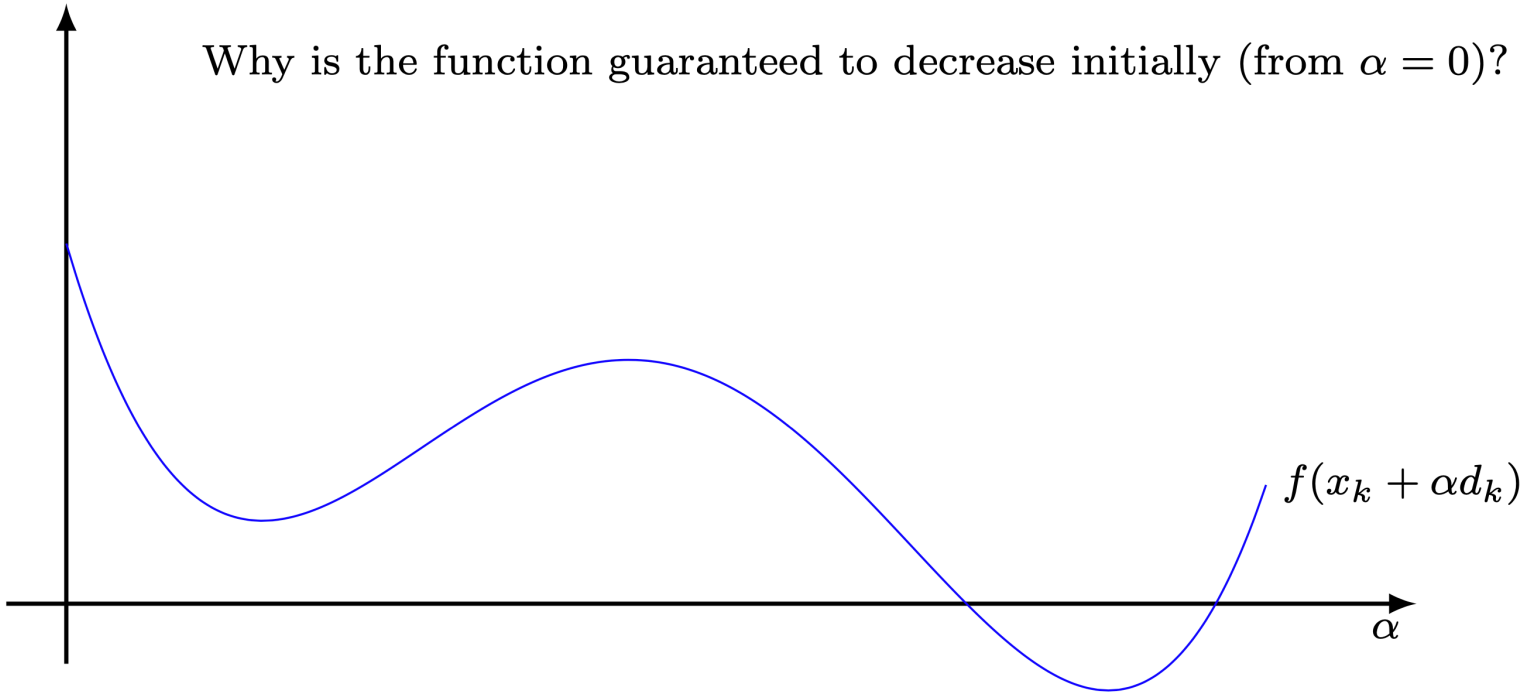
Perhaps the most common condition to place on α_k is the following sufficient decrease (also known as the Armijo) condition:

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

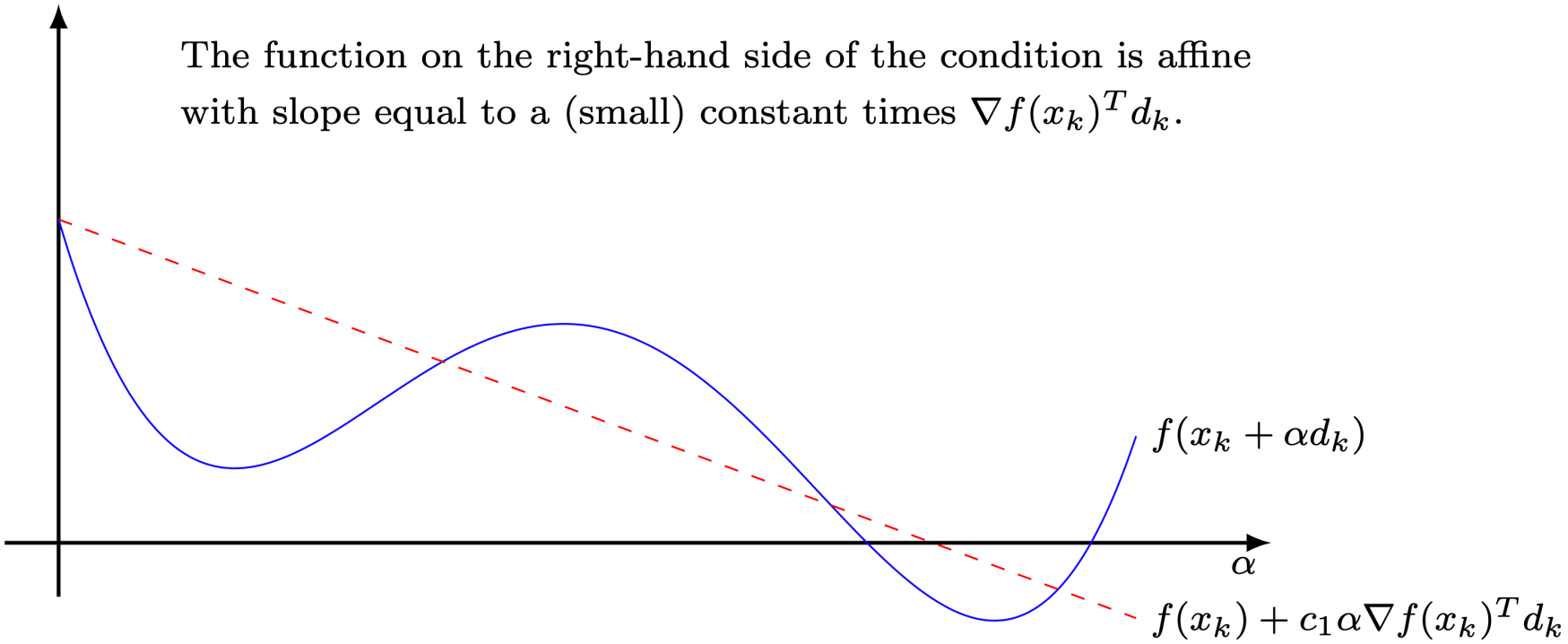
where $c_1 \in (0, 1)$ is a user-specified constant.

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

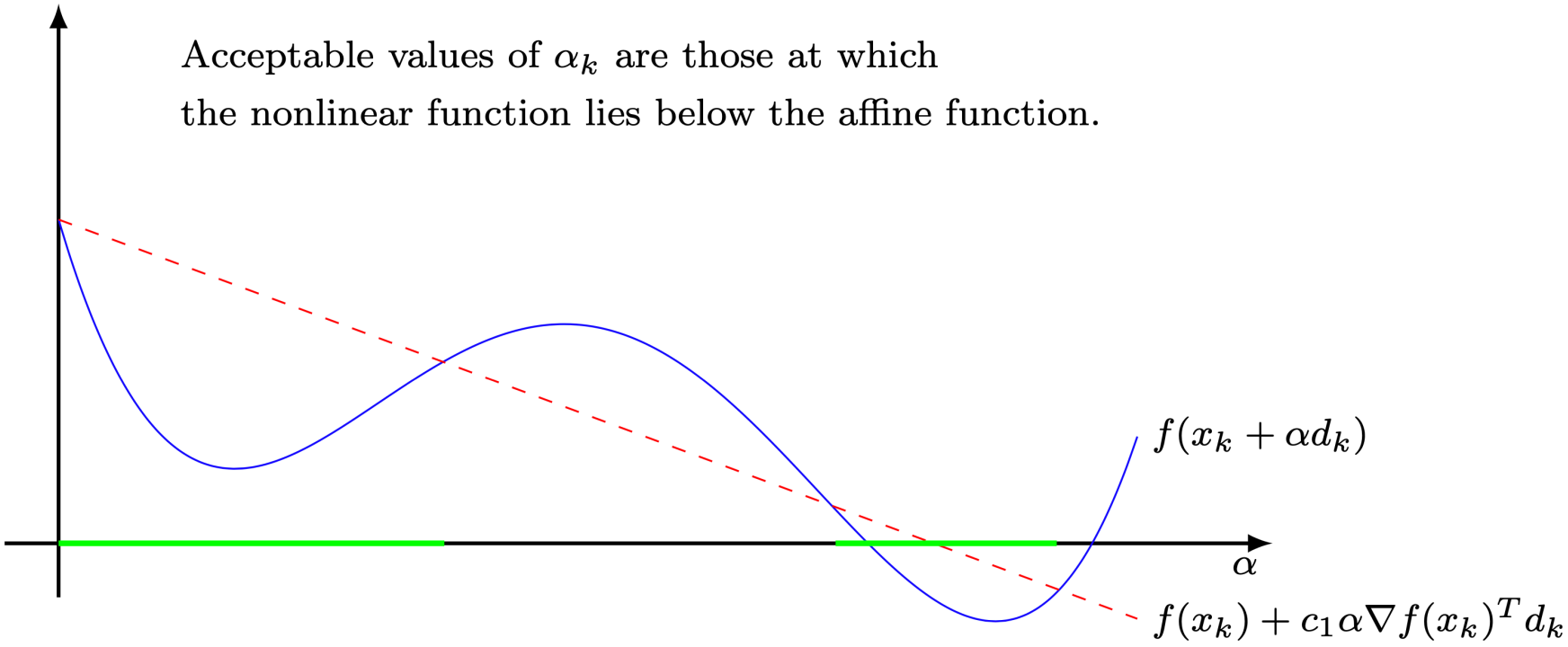
Sufficient decrease illustrated: $f(x_k + \alpha_k d_k) \leq f(x_k) + c_1 \alpha_k \nabla f(x_k)^T d_k$



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



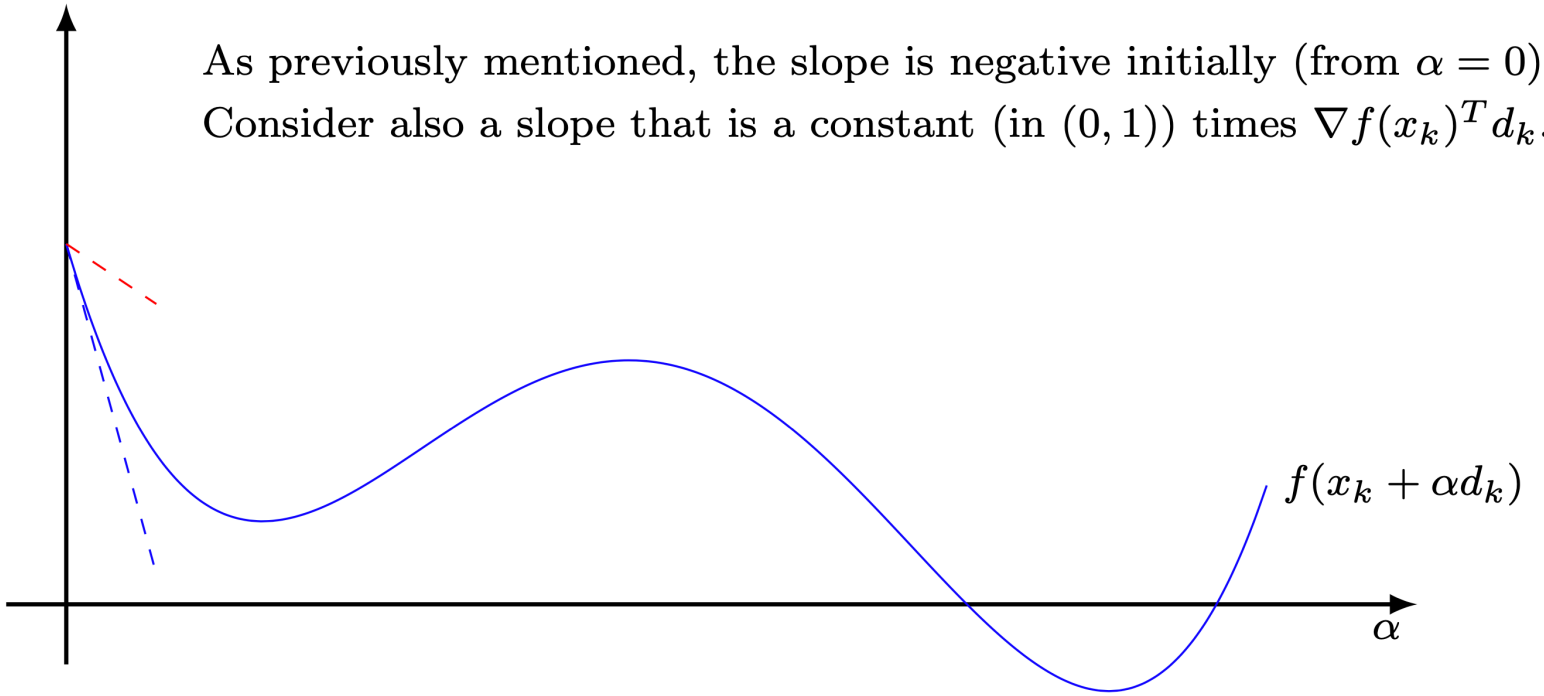
Sufficient decrease illustrated: $f(x_k + \alpha_k d_k) \leq f(x_k) + c_1 \alpha_k \nabla f(x_k)^T d_k$



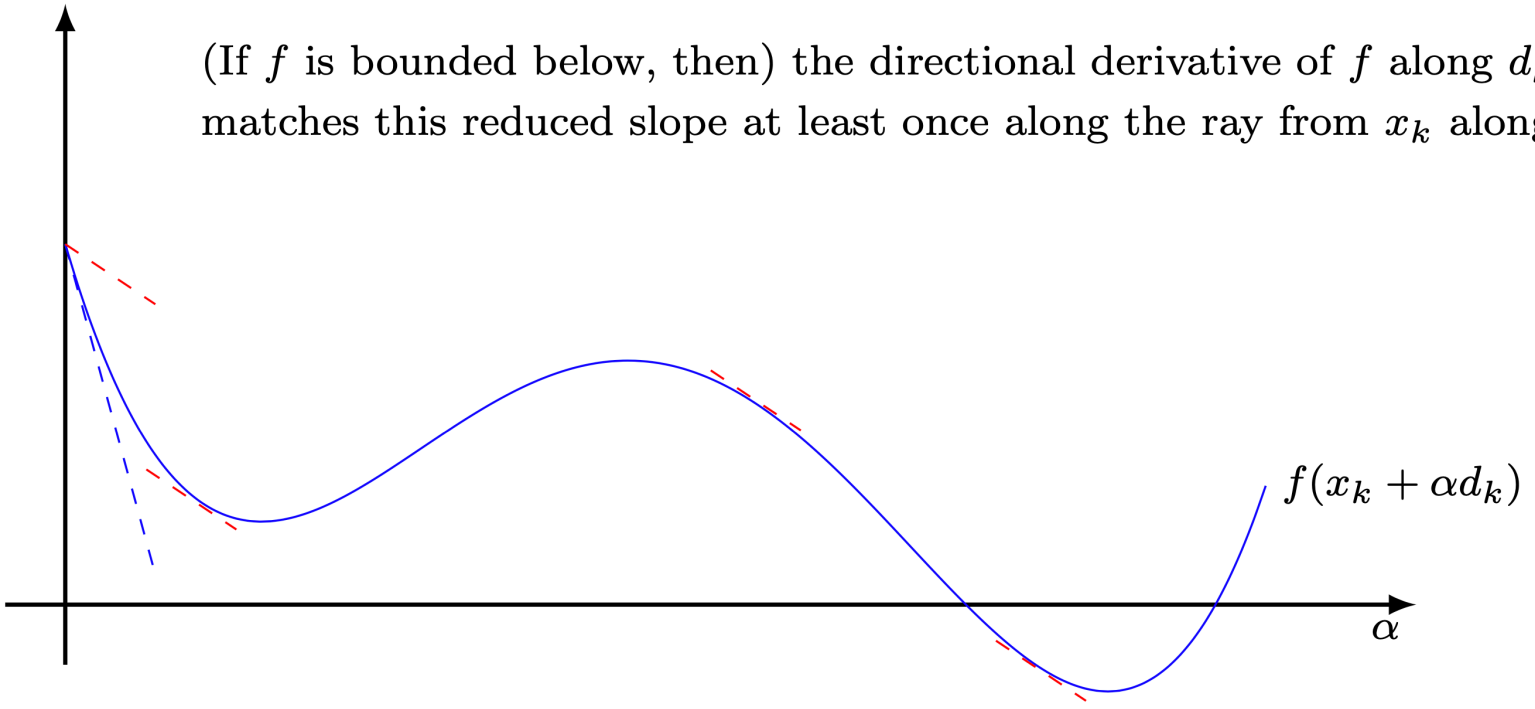
Curvature condition

- 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
  $$\{\gamma^0, \gamma^1, \gamma^2, \dots\}$$
  where $\gamma \in (0, 1)$ is a given constant, satisfying 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.

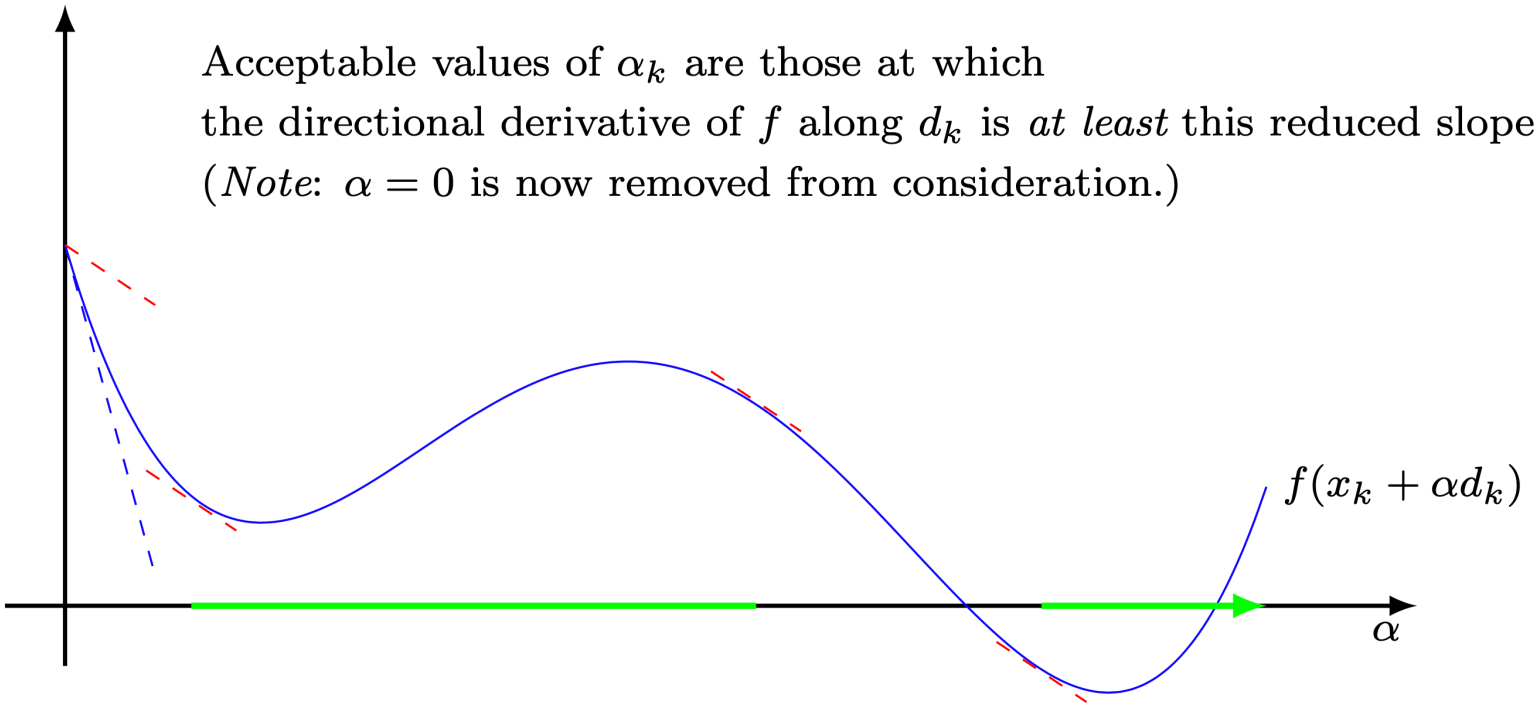
Curvature condition illustrated: $\nabla f(x_k + \alpha_k d_k)^T d_k \geq c_2 \nabla f(x_k)^T d_k$



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



Wolfe conditions

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

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

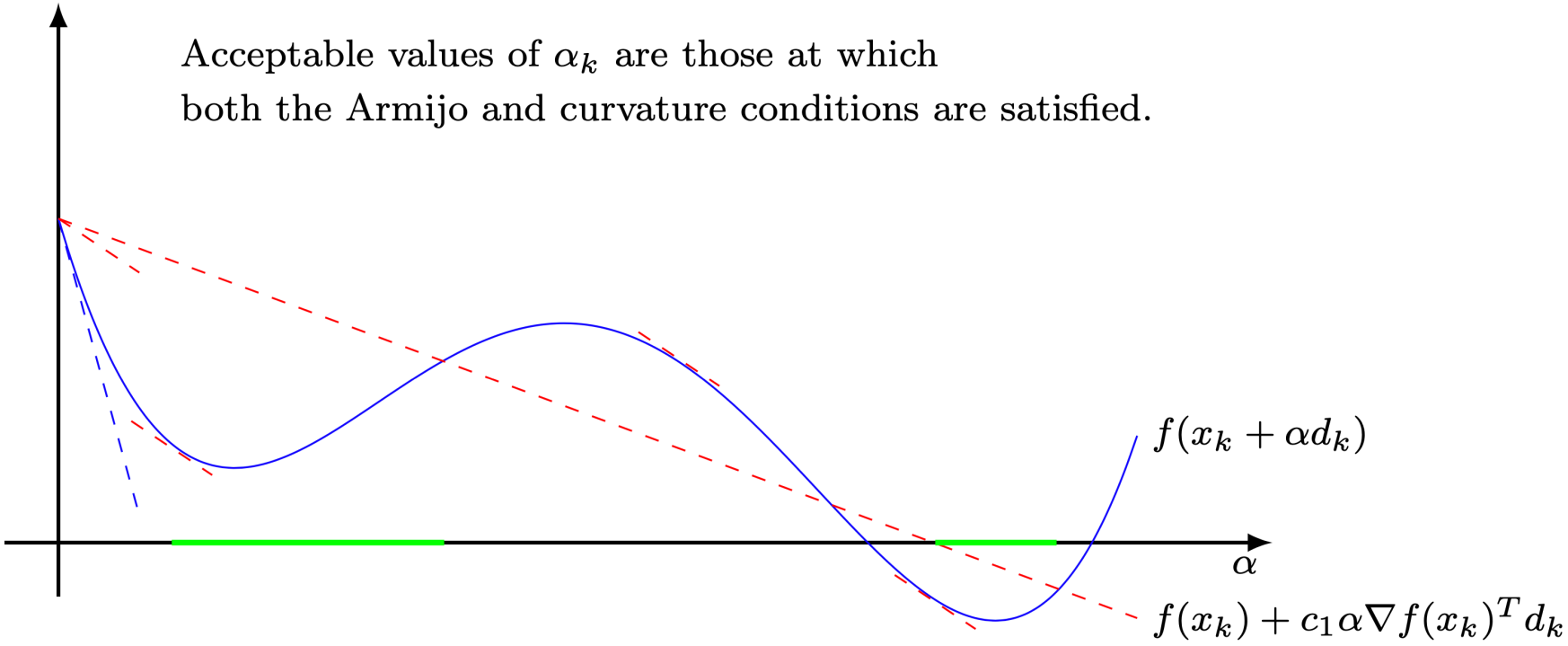
- Harder to satisfy conditions that may bring us closer to exact minimizers are the related strong Wolfe conditions:

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

- See Lemma 3.1 in the textbook* to see that an α_k satisfying the (strong) Wolfe conditions is always computable, and to see why we need $c_2 > c_1$.

*Textbook: Numerical Optimization

Wolfe conditions illustrated



Goldstein准则

定义 6.2 (Goldstein 准则) 设 d^k 是点 x^k 处的下降方向，若

$$f(x^k+\alpha d^k) \leqslant f(x^k)+c\alpha\nabla f(x^k)^{\mathrm{T}}d^k, \tag{6.1.3a}$$

$$f(x^k+\alpha d^k) \geqslant f(x^k)+(1-c)\alpha\nabla f(x^k)^{\mathrm{T}}d^k, \tag{6.1.3b}$$

则称步长 α 满足 **Goldstein 准则**，其中 $c \in \left(0,\frac{1}{2}\right)$.

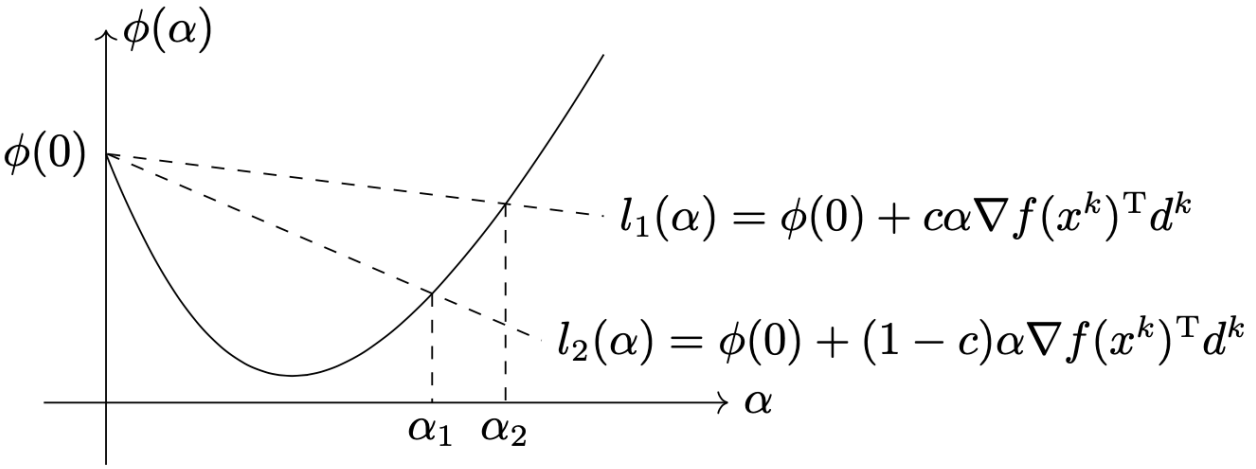


图 6.2 Goldstein 准则

Nonmonotone线搜索

定义 6.4 (Grippo [91]) 设 d^k 是点 x^k 处的下降方向，$M>0$ 为给定的正整数. 以下不等式可作为一种线搜索准则：

$$f(x^k+\alpha d^k) \leqslant \max_{0\leqslant j\leqslant \min\{k,M\}} f(x^{k-j}) + c_1\alpha \nabla f(x^k)^{\mathrm{T}} d^k, \tag{6.1.5}$$

其中 $c_1 \in (0,1)$ 为给定的常数.

Zoutendijk's theorem

- 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

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

Zoutendijk's Theorem

Theorem 7.3.1 (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

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

where, for all $k \in \mathbb{N}_+$,

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

Then,

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

Zoutendijk's Theorem

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

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

Intuitively, we can interpret this by noting that α_k can be small if:

- L is large, i.e., the gradient may change very quickly;
- $\|d_k\|$ is large, i.e., the full step moves far away from local information;
- $\nabla f(x_k)^T d_k \approx 0$, i.e., the step is one of weak descent.

Zoutendijk's Theorem

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 the reduction in $f(x)$ is large when:

- $\cos^2\theta_k$ is large, i.e., when d_k and $-\nabla f(x_k)$ are (nearly) parallel;
- $\|\nabla f(x_k)\|$ is large, i.e., when the gradient is large in norm.

Zoutendijk's Theorem

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 $k = 0, 1, 2, \ldots, K$, we obtain

$$f(x_{k+1}) \leq f(x_0) - \sum_{k=0}^{K} \cos^2\theta_k \|\nabla f(x_k)\|^2.$$

- Thus, since f is bounded below,

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

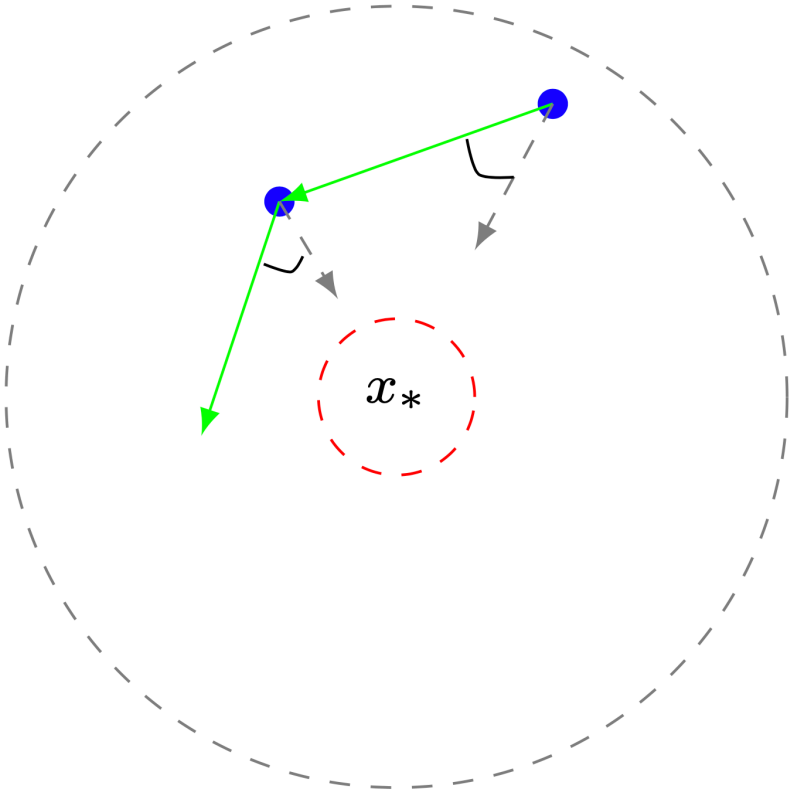
Intuitively, this means that for large k we either have:

- $\cos^2\theta_k \approx 0$, i.e., d_k is (nearly) perpendicular to $-\nabla f(x_k)$;
- $\|\nabla f(x_k)\| \approx 0$, i.e., the gradient is small in norm.

Implications

- 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.,
$$\cos \theta_k \geq \delta > 0 \text{ for all sufficiently large } k,$$
then we immediately have
$$\|\nabla f(x_k)\| \to 0.$$

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



Nonconvexity and line search methods