Optimization and Numerical Analysis: Nonlinear programming without constraints

Robert Gower

October 6, 2020

Table of Contents

- (1669) Invents simply version of Newton's method for finding roots of polynomials (no calculus!): *De analysi per aequationes numero terminorum infinitas*.
- (1740) Full Newton's method as we know it: Thomas Simpson

Figure: Isaac Newton

Figure: Augustin Louis Cauchy

- (1847) Invents gradient descent: *Compte Rendu á l'Académie des Sciences*
- Why? Solving algebraic equations of the orbit of heavenly bodies.
- École Polytechnique and he wrote almost 800 papers!

The Problem: Nonlinear programming

Minimize a nonlinear differentiable function $f : x \in \mathbb{R}^n \mapsto f(x) \in \mathbb{R}$

$$x^* = \arg\min_{x \in \mathbb{R}^n} f(x). \tag{1}$$

Warning: This problem is often impossible. First check there exists a minimum. Even linear programming does not always have a maximum! Develop iterative methods $x^1, \ldots, x^k, \ldots,$ such that

$$\lim_{k \to \infty} x^k = x^*.$$

Template method

$$x^{k+1} = x^k + s_k d^k,$$

where $s_k > 0$ is a *step size* and $d^k \in \mathbb{R}^n$ is *search direction*. Satisfy the *descent condition*

$$f(x^{k+1}) < f(x^k).$$

Local and Global Minima

Definition of Local Minima

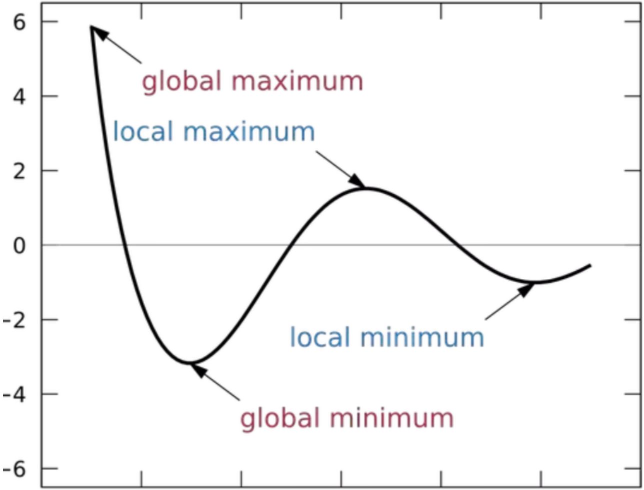
The point $x^* \in \mathbb{R}^n$ is a *local minima* of $f(x)$ if there exists $r > 0$ such that

$$f(x^*) \leq f(x), \quad \forall \|x - x^*\|_2 < r. \tag{2}$$

Definition of Global Minima

The point $x^* \in \mathbb{R}^n$ is a *global minima* of $f(x)$ if

$$f(x^*) \leq f(x), \quad \forall x. \tag{3}$$



In general finding global minima is impossible.

Multivariate Calculus

For a differentiable function $f : x \in \mathbb{R}^n \mapsto f(x) \in \mathbb{R}$, we refer to $\nabla f(x)$ as the gradient evaluated at x defined by

$$\nabla f(x) = \left[\frac{\partial f(x)}{\partial x_1}, \ldots, \frac{\partial f(x)}{\partial x_n}\right]^\top .$$

Note that $\nabla f(x)$ is a column-vector. For any vector valued function $F : x \in \mathbb{R}^n \to F(x) = [f_1(x), \ldots, f_n(x)]^\top \in \mathbb{R}^n$ define the *Jacobian matrix* by

$$\nabla F(x) \stackrel{\text{def}}{=} \begin{bmatrix} \frac{\partial f_1(x)}{\partial x_1} & \frac{\partial f_2(x)}{\partial x_1} & \frac{\partial f_3(x)}{\partial x_1} & \cdots & \frac{\partial f_n(x)}{\partial x_1} \\ \frac{\partial f_2(x)}{\partial x_1} & \frac{\partial f_2(x)}{\partial x_2} & \frac{\partial f_3(x)}{\partial x_2} & \cdots & \frac{\partial f_n(x)}{\partial x_2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \frac{\partial f_1(x)}{\partial x_n} & \frac{\partial f_2(x)}{\partial x_n} & \frac{\partial f_3(x)}{\partial x_n} & \cdots & \frac{\partial f_n(x)}{\partial x_n} \end{bmatrix}$$

$$= \left[\nabla f_1(x), \nabla f_2(x), \nabla f_3(x), \ldots, \nabla f_2(x)\right]$$

Multivariate Calculus

The gradient is useful because of 1st order Taylor expansion

$$f(x^0 + d) = f(x^0) + \nabla f(x^0)^\top d + \epsilon(d)\|d\|_2, \tag{4}$$

where $\epsilon(d)$ is a real valued such that

$$\lim_{d \to 0} \epsilon(d) = 0. \tag{5}$$

Multivariate Calculus

The gradient is useful because of 1st order Taylor expansion

$$f(x^0 + d) = f(x^0) + \nabla f(x^0)^\top d + \epsilon(d)\|d\|_2, \tag{4}$$

where $\epsilon(d)$ is a real valued such that

$$\lim_{d \to 0} \epsilon(d) = 0. \tag{5}$$

Definition of limit: given any constant $c > 0$ there exists $\delta > 0$ such that

$$\|d\| < \delta \quad \Rightarrow \quad |\epsilon(d)| < c. \tag{6}$$

Example (The $\epsilon(d)$ function)

If $f(x) = \|x\|_2^2$ and $f(x) = x^\top Ax$, where $A = A^\top$, what is $\epsilon(d)$?
Name three functions ϵ such that $\lim_{d \to 0} \epsilon(d) = 0$.

Example (The $\epsilon(d)$ function)

If $f(x) = \|x\|_2^2$ and $f(x) = x^\top Ax$, where $A = A^\top$, what is $\epsilon(d)$? Name three functions ϵ such that $\lim_{d\to 0} \epsilon(d) = 0$.

Solution:

$$f(x_0 + d) = (x_0 + d)^\top A(x_0 + d) = \underbrace{x_0^\top Ax_0}_{=f(x_0)} + \underbrace{2x_0^\top A}_{=\nabla f(x_0)^\top} d + d^\top Ad$$

Example (The $\epsilon(d)$ function)

If $f(x) = \|x\|_2^2$ and $f(x) = x^\top Ax$, where $A = A^\top$, what is $\epsilon(d)$? Name three functions ϵ such that $\lim_{d\to 0}\epsilon(d) = 0$.

Solution:

$$f(x_0 + d) = (x_0 + d)^\top A(x_0 + d) = \underbrace{x_0^\top Ax_0}_{=f(x_0)} + \underbrace{2x_0^\top A}_{=\nabla f(x_0)^\top} d + d^\top Ad$$

Thus $\epsilon(d)\|d\|_2 = d^\top Ad \Rightarrow \epsilon(d) = \dfrac{d^\top Ad}{\|d\|_2}$ and

$$\lim_{d\to 0}\epsilon(d) = 0$$

Example (The $\epsilon(d)$ function)

If $f(x) = \|x\|_2^2$ and $f(x) = x^\top Ax$, where $A = A^\top$, what is $\epsilon(d)$? Name three functions ϵ such that $\lim_{d\to 0}\epsilon(d) = 0$.

Solution:

$$f(x_0 + d) = (x_0 + d)^\top A(x_0 + d) = \underbrace{x_0^\top Ax_0}_{=f(x_0)} + \underbrace{2x_0^\top A}_{=\nabla f(x_0)^\top} d + d^\top Ad$$

Thus $\epsilon(d)\|d\|_2 = d^\top Ad \Rightarrow \epsilon(d) = \dfrac{d^\top Ad}{\|d\|_2}$ and

$$\lim_{d\to 0}\epsilon(d) = 0$$

Three examples:

$$\epsilon(d) = \log(d), \quad \epsilon(d) = \|d\|, \quad \epsilon(d) = \frac{a\|d\|^3 + b\|d\|^2}{c\|d\| + e}.$$

The Hessian Matrix

If $f \in C^2$, we refer to $\nabla^2 f(x)$ as the Hessian matrix:

$$\nabla^2 f(x) \quad \stackrel{\text{def}}{=} \quad \begin{bmatrix} \frac{\partial^2 f(x)}{\partial x_1 \partial x_1} & \frac{\partial^2 f(x)}{\partial x_1 \partial x_2} & \frac{\partial^2 f(x)}{\partial x_1 \partial x_3} & \cdots & \frac{\partial^2 f(x)}{\partial x_1 \partial x_n} \\ \frac{\partial^2 f(x)}{\partial x_1 \partial x_2} & \frac{\partial^2 f(x)}{\partial x_2 \partial x_2} & \frac{\partial^2 f(x)}{\partial x_2 \partial x_3} & \cdots & \frac{\partial^2 f(x)}{\partial x_2 \partial x_n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 f(x)}{\partial x_n \partial x_1} & \frac{\partial^2 f(x)}{\partial x_n \partial x_2} & \frac{\partial^2 f(x)}{\partial x_n \partial x_3} & \cdots & \frac{\partial^2 f(x)}{\partial x_n \partial x_n} \end{bmatrix}$$

If $f \in C^2$ then

$$\frac{\partial^2 f(x)}{\partial x_i \partial x_j} = \frac{\partial^2 f(x)}{\partial x_j \partial x_i}, \ \forall i, j \in \{1, \ldots, n\}, \quad \Leftrightarrow \quad \nabla^2 f(x) = \nabla^2 f(x)^\top .$$

Hessian matrix useful for 2nd order Taylor expansion.

$$f(x^0 + d) = f(x^0) + \nabla f(x^0)^\top d + \frac{1}{2} d^\top \nabla^2 f(x^0) d + \epsilon(d) \|d\|_2^2 . \qquad (7)$$

Exe: If $f(x) = x^3$ or $f(x) = x^\top A x$ what is $\epsilon(d)$?

The Hessian Matrix

If $f \in C^2$, we refer to $\nabla^2 f(x)$ as the Hessian matrix:

$$\nabla^2 f(x) \quad \overset{\text{def}}{=} \quad \begin{bmatrix} \frac{\partial^2 f(x)}{\partial x_1 \partial x_1} & \frac{\partial^2 f(x)}{\partial x_1 \partial x_2} & \frac{\partial^2 f(x)}{\partial x_1 \partial x_3} & \cdots & \frac{\partial^2 f(x)}{\partial x_1 \partial x_n} \\ \frac{\partial^2 f(x)}{\partial x_1 \partial x_2} & \frac{\partial^2 f(x)}{\partial x_2 \partial x_2} & \frac{\partial^2 f(x)}{\partial x_2 \partial x_3} & \cdots & \frac{\partial^2 f(x)}{\partial x_2 \partial x_n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 f(x)}{\partial x_n \partial x_1} & \frac{\partial^2 f(x)}{\partial x_n \partial x_2} & \frac{\partial^2 f(x)}{\partial x_n \partial x_3} & \cdots & \frac{\partial^2 f(x)}{\partial x_n \partial x_n} \end{bmatrix}$$

If $f \in C^2$ then

$$\frac{\partial^2 f(x)}{\partial x_i \partial x_j} = \frac{\partial^2 f(x)}{\partial x_j \partial x_i}, \ \forall i, j \in \{1, \ldots, n\}, \quad \Leftrightarrow \quad \nabla^2 f(x) = \nabla^2 f(x)^\top.$$

Hessian matrix useful for 2nd order Taylor expansion.

$$f(x^0 + d) = f(x^0) + \nabla f(x^0)^\top d + \frac{1}{2} d^\top \nabla^2 f(x^0) d + \epsilon(d) \|d\|_2^2. \qquad (7)$$

Exe: If $f(x) = x^3$ or $f(x) = x^\top A x$ what is $\epsilon(d)$?
Sol: $(x + d)^3 = x^3 + 3x^2 d + 3xd^2 + d^3$. Thus $\epsilon(d) = d$

The Product-rule

The vector valued version of the product rule

- For any function $F(x) : \mathbb{R}^n \to \mathbb{R}^n$ and matrix $A \in \mathbb{R}^{n\times n}$ we have
$$\nabla(F(x)^\top A) = \nabla F(x)^\top A. \tag{8}$$
- For any two vector valued functions F_1 and F_2 we have that
$$\nabla(F_1(x)^\top F_2(x)) = \nabla F_1(x) F_2(x) + \nabla F_2(x) F_1(x). \tag{9}$$

Example

Let $f(x) = \frac{1}{2}x^\top Ax$, where $A \in \mathbb{R}^{n\times n}$ is a symmetric positive definite matrix. Calculate the gradient and the Hessian of $f(x)$.

The Product-rule

The vector valued version of the product rule

- For any function $F(x) : \mathbb{R}^n \to \mathbb{R}^n$ and matrix $A \in \mathbb{R}^{n\times n}$ we have
$$\nabla(F(x)^\top A) = \nabla F(x)^\top A. \tag{8}$$
- For any two vector valued functions F_1 and F_2 we have that
$$\nabla(F_1(x)^\top F_2(x)) = \nabla F_1(x) F_2(x) + \nabla F_2(x) F_1(x). \tag{9}$$

Example

Let $f(x) = \frac{1}{2}x^\top Ax$, where $A \in \mathbb{R}^{n\times n}$ is a symmetric positive definite matrix. Calculate the gradient and the Hessian of $f(x)$.

Let $F_1(x) = A^\top x$ and $F_2(x) = x$ then
$\nabla f(x) = \frac{1}{2}\nabla(A^\top x)x + \frac{1}{2}\nabla(x)A^\top x = \frac{1}{2}(A + A^\top)x = Ax$ since
$\nabla(A^\top x) = A^\top \nabla(x) = A$. Differentiating again
$\nabla(\nabla f(x)) = \nabla(Ax) = \nabla(A)x + \nabla(x)A = A.$

Template method

$$x^{k+1} = x^k + s_k d^k,$$

where $s_k > 0$ is a *step size* and $d^k \in \mathbb{R}^n$ is *search direction*. Satisfy the *descent condition*

$$f(x^{k+1}) < f(x^k).$$

How to choose d?

How to find $d \in \mathbb{R}^n$ such that

$$f(x_k + s_k d) \le f(x_k).$$

Lemma (Steepest Descent)

For $d \in \mathbb{R}^n$ the local change of $f(x)$ around x_0 is

$$\Delta(d) \overset{def}{=} \lim_{s \to 0^+} \frac{f(x^0 + sd) - f(x^0)}{s}. \tag{10}$$

Let $v = -\nabla f(x^0) / \|\nabla f(x^0)\|_2$ be the normalized gradient. We have

$$v = \arg\min_{d \in \mathbb{R}^n} \Delta(d) \quad \text{subject to } \|d\|_2 = 1. \tag{11}$$

The negative normalized gradient is the direction that minimizes the *local change* of $f(x)$ around x^0. The normalized gradient

Proof.

Using 1st order Taylor we have that

$$f(x^0 + sd) - f(x^0) = s\nabla f(x^0)^\top d + \epsilon(sd)s.$$

Dividing by s and taking the limit $s \to 0$ we have

$$\Delta(d) = \lim_{s\to 0^+} \frac{f(x^0 + sd) - f(x^0)}{s} = \nabla f(x^0)^\top d + \lim_{s\to 0^+} \epsilon(sd) = \nabla f(x^0)^\top d.$$

Now using that $\|d\|_2 = 1$ together with the Cauchy inequality

$$-\|\nabla f(x^0)\|_2 \quad \leq \quad \Delta(d) = \nabla f(x^0)^\top d \quad \leq \quad \|\nabla f(x^0)\|_2. \tag{12}$$

The upper and lower bound is achieved when $d = \nabla f(x^0)/\left\|\nabla f(x^0)\right\|_2$ and $d = -\nabla f(x^0)/\left\|\nabla f(x^0)\right\|_2$, respectively. □

The search direction d is a *descent direction* if it has an obtuse angle with the gradient

Corollary (Descent Condition)

If $d^\top \nabla f(x_0) < 0$ *then there exists* $s > 0$ *such that*

$$f(x_0 + sd) < f(x_0).$$

The search direction d is a *descent direction* if it has an obtuse angle with the gradient

Corollary (Descent Condition)

If $d^\top \nabla f(x_0) < 0$ *then there exists* $s > 0$ *such that*

$$f(x_0 + sd) < f(x_0).$$

Proof.

From (12) we have that $\Delta(d) = \nabla f(x^0)^\top d < 0$.
Let $c = -\nabla f(x^0)^\top d > 0$.

The search direction d is a *descent direction* if it has an obtuse angle with the gradient

Corollary (Descent Condition)

If $d^\top \nabla f(x_0) < 0$ *then there exists* $s > 0$ *such that*

$$f(x_0 + sd) < f(x_0).$$

Proof.

From (12) we have that $\Delta(d) = \nabla f(x^0)^\top d < 0$.
Let $c = -\nabla f(x^0)^\top d > 0$.
Let $s > 0$ be such that $\epsilon(sd) < \frac{c}{2}$. (Because $\lim_{s\to 0} \epsilon(sd) = 0$)

The search direction d is a *descent direction* if it has an obtuse angle with the gradient

Corollary (Descent Condition)

If $d^\top \nabla f(x_0) < 0$ *then there exists* $s > 0$ *such that*

$$f(x_0 + sd) < f(x_0).$$

Proof.

From (12) we have that $\Delta(d) = \nabla f(x^0)^\top d < 0$.
Let $c = -\nabla f(x^0)^\top d > 0$.
Let $s > 0$ be such that $\epsilon(sd) < \frac{c}{2}$. (Because $\lim_{s\to 0} \epsilon(sd) = 0$)
Consequently from 1st order Taylor:

$$\frac{f(x^0 + sd) - f(x^0)}{s} = \nabla f(x^0)^\top d + \epsilon(sd) \leq -\frac{c}{2} < 0.$$

Re-arranging $f(x^0 + sd) \leq f(x^0) - s\frac{c}{2} < f(x^0)$ $\square$

Definition of Local Minima

The point $x^* \in \mathbb{R}^n$ is a *local minima* of $f(x)$ if there exists $r > 0$ such that

$$f(x^*) \leq f(x), \quad \forall \|x - x^*\|_2 < r. \tag{13}$$

Theorem (Necessary optimality conditions)

If x^* is a local minima of $f(x)$ then

1. $\nabla f(x^*) = 0$
2. $d^\top \nabla^2 f(x^*) d \geq 0, \quad \forall d \in \mathbb{R}^n.$

So it is necessary that $\nabla f(x^*) = 0$ and the d is positive curvature direction before we stop.

Proof.

That $\nabla f(x^*) = 0$ follows from Descent Condition. Suppose there exists $d \in \mathbb{R}^n$ such that $d^\top \nabla^2 f(x^*) d < 0$. Suppose w.l.o.g that $\|d\|_2 = 1$. Using the 2nd order Taylor we have that

$$f(x^* + sd) \quad = \quad f(x^*) + \frac{s^2}{2} d^\top \nabla^2 f(x^*) d + \epsilon(sd) s^2.$$

Proof.

That $\nabla f(x^*) = 0$ follows from Descent Condition. Suppose there exists $d \in \mathbb{R}^n$ such that $d^\top \nabla^2 f(x^*) d < 0$. Suppose w.l.o.g that $\|d\|_2 = 1$. Using the 2nd order Taylor we have that

$$f(x^* + sd) \;=\; f(x^*) + \frac{s^2}{2} d^\top \nabla^2 f(x^*) d + \epsilon(sd) s^2.$$

Let $\delta > 0$ be such that for $s \leq \delta$ we have that $\epsilon(sd) < \left| d^\top \nabla^2 f(x^*) d \right| / 4$. Dividing the above by s^2, for $s \leq \delta$ we have that

$$\begin{aligned} \frac{f(x^* + sd)}{s^2} &= \frac{f(x^*)}{s^2} + \frac{1}{2} d^\top \nabla^2 f(x^*) d + \epsilon(sd) \\ &< \frac{f(x^*)}{s^2} + \frac{1}{4} d^\top \nabla^2 f(x^*) d, \end{aligned}$$

thus $f(x^* + sd) < f(x^*)$ for all $s \leq \delta$ which contradicts the definition of local minima. □

With a slight modification, same conditions they are also sufficient.

Theorem (Sufficient Local Optimality conditions)

If $x^* \in \mathbb{R}^n$ is such that

1. $\nabla f(x^*) = 0$
2. $d^\top \nabla^2 f(x^*) d > 0, \quad \forall d \in \mathbb{R}^n$ *with* $d \neq 0$,

then x^* is a local minima.

We can use this theorem to find local minima!

Proof: Let $d \in \mathbb{R}^n$. Because $\nabla^2 f(x^*)$ is positive definite, the smallest non-zero eigenvalue must be strictly positive. Consequently

$$\|d\|^2 \lambda_{\min}(\nabla^2 f(x^*)) \leq d^\top \nabla^2 f(x^*) d.$$

Using the second-order Taylor expansion, we have that

$$\begin{aligned} f(x^* + d) &= f(x^*) + \frac{1}{2} d^\top \nabla^2 f(x^*) d + \epsilon(d) \|d\|_2^2 \\ &\geq f(x^*) + \frac{\|d\|_2^2}{2} \lambda_{\min}(\nabla^2 f(x^*)) + \epsilon(d) \|d\|_2^2. \end{aligned}$$

Let $r > 0$ be such that every d with $\|d\| \leq r$ we have that

$$|\epsilon(d)| < \lambda_{\min}(\nabla^2 f(x^*))/4 \quad \Rightarrow \quad \epsilon(d) > -\lambda_{\min}(\nabla^2 f(x^*))/4.$$

Thus for $\|d\| \leq r$ we have

$$\begin{aligned} f(x^* + d) &\geq f(x^*) + \frac{\|d\|_2^2}{2} \lambda_{\min}(\nabla^2 f(x^*)) + \epsilon(d) \|d\|_2^2 \\ &\geq f(x^*) + \frac{\|d\|_2^2}{4} \lambda_{\min}(\nabla^2 f(x^*)) > f(x^*). \quad \square \end{aligned}$$

Exercise

Let $f(x) = \frac{1}{2}x^\top Ax - x^\top b + c$, with A symmetric positive definite. How many local/global minimas can $f(x)$ have? Find a formula for the minima using only the *data* A and b.

Proof.

By the sufficient conditions x^* is a local minima if

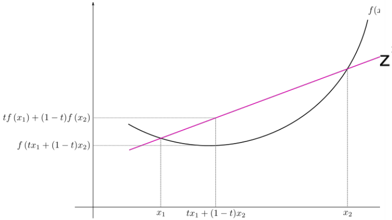
$$\nabla f(x^*) = 0 \;\Leftrightarrow\; Ax^* = b,$$

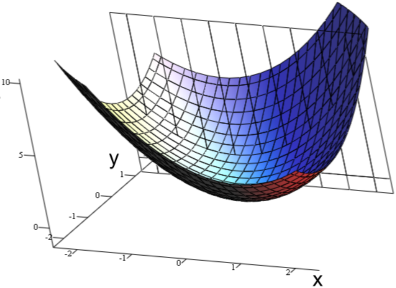
and

$$\nabla^2 f(x^*) = A \succ 0.$$

Since $Ax = b$ has only one solution there exists only one local minima which must be the global minima. $\square$

Convex Functions





$$f(tx + (1-t)y) \leq tf(x) + (1-t)f(y), \quad \forall x, y \in \mathbb{R}^d,\ t \in [0, 1].$$

Theorem

If f is a convex function, then every local minima of f is also a global minima. We only need to check 1st order $\nabla f(x^*) = 0$!

Proof.

Let x^* be a local minima and suppose there exists $\bar{x} \in \mathbb{R}^n$ such that $f(\bar{x}) < f(x^*)$. Let $z_t = t\bar{x} + (1-t)x^*$ for $t \in [0, 1]$. By the definition of convexity we have that

$$\begin{aligned} f(z_t) &= f((1-t)\bar{x} + tx^*) \leq (1-t)f(\bar{x}) + tf(x^*) \\ &< (1-t)f(x^*) + tf(x^*) = f(x^*). \end{aligned} \tag{14}$$

Thus x^* cannot be a local minima. Indeed, for any $r > 0$ with $r \leq \|\bar{x} - x^*\|_2$, we have that by choosing $t = 1 - r/\|\bar{x} - x^*\|_2$ we have that

$$\|z_t - x^*\|_2 = (1-t)\|\bar{x} - x^*\|_2 \leq r.$$

Yet from (14) we have that $f(z_t) < f(x^*)$. A contraction. Thus there exists no $\bar{x}$ with $f(\bar{x}) < f(x^*)$. $\square$

Theorem

If f is twice continuously differentiable, then the following three statements are equivalent

$$\begin{aligned} f(tx+(1-t)y) &\leq tf(x)+(1-t)f(y), \quad \forall x,y,\, t\in[0,\,1]. && \text{(0th)} \\ f(y) &\geq f(x)+\nabla f(x)^\top (y-x), \quad \forall x,y. && \text{(1st)} \\ 0 &\leq d^\top \nabla^2 f(x) d, \quad \forall x,d. && \text{(2nd)} \end{aligned}$$

Proof.

We prove (0th)$\Rightarrow$ (1st)$\Rightarrow$ (2nd).
The remaing (2nd)$\Rightarrow$ (0th) is left as an exercise.
(0th)$\Rightarrow$ (1st): Dividing (0th) by t and re-arranging

$$\frac{f(y+t(x-y))-f(y)}{t} \leq f(x)-f(y).$$

Now taking the limit $t \to 0$ gives (1st).

Proof.

(1st)$\Rightarrow$ (2nd): First we prove this holds for 1–dimensional functions $f : \mathbb{R} \to \mathbb{R}$. From (1st) we have that

$$\begin{aligned} f(y) &\geq f(x) + f'(x)(y-x), \\ f(x) &\geq f(y) + f'(y)(x-y). \end{aligned}$$

Combining the above two we have that

$$f'(x)(y-x) \leq f(y) - f(x) \leq f'(y)(y-x).$$

Dividing by $(y-x)^2$ we have

$$\frac{f'(y)-f'(x)}{y-x} \geq 0, \quad \forall x, y, x \neq y.$$

It remains to take the limit. Extend to every n–dimensional function using

$$\left.\frac{d^2 f(x+tv)}{dv^2}\right|_{t=0} = v^\top \nabla^2 f(x) v \geq 0, \forall v \neq 0. \qquad \square$$

Move in negative gradient direction iteratively

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

where $s^k > 0$ is the step size. How to choose s^k the stepsize?
Sometimes constant step size works

Theorem

Let $A \in \mathbb{R}^{n\times n}$ is symmetric positive definite. $f(x) = \frac{1}{2}x^\top Ax - x^\top b + c$. If we choose a fixed stepsize of $s^k = 1/\sigma_{\max}(A)$ then GD converges

$$\|\nabla f(x^{k+1})\|_2 \leq \left(1 - \frac{\sigma_{\min}(A)}{\sigma_{\max}(A)}\right)^k \|\nabla f(x^0)\|_2. \tag{15}$$

Proof part I:

$$\begin{aligned} \nabla f(x^{k+1}) &= Ax^{k+1} - b \\ &= A(x^k - s\nabla f(x^k)) - b \\ &= A(x^k - s(Ax^k - b)) - b \\ &= Ax^k - b - sA(Ax^k - b) \;=\; (I - sA)\nabla f(x^k). \end{aligned}$$

Proof part II: From $\nabla f(x^{k+1}) = (I - sA)\nabla f(x^k)$ taking norms

$$\|\nabla f(x^{k+1})\|_2 \leq \|I - sA\|_2 \|\nabla f(x^k)\|_2.$$

Choosing $s = 1/\sigma_{\max}(A)$ we have that $I - sA$ is symmetric positive definite and

$$\|I - sA\|_2 = 1 - s\sigma_{\min}(A) = 1 - \frac{\sigma_{\min}(A)}{\sigma_{\max}(A)} < 1.$$

Homework: Prove this last step! Thus finally

$$\begin{aligned} \|\nabla f(x^{k+1})\|_2 \leq & \left(1 - \frac{\sigma_{\min}(A)}{\sigma_{\max}(A)}\right) \|\nabla f(x^k)\|_2 \\ \leq & \left(1 - \frac{\sigma_{\min}(A)}{\sigma_{\max}(A)}\right)^k \|\nabla f(x^0)\|_2. \quad \square \end{aligned}$$

What to do for non-quadratic functions? Choose the best s^k ?

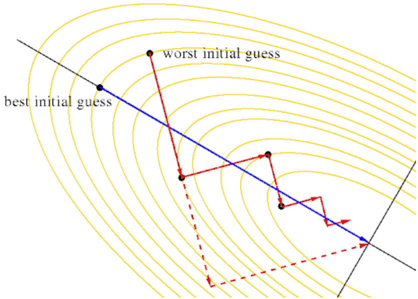
$$s^k = \arg\min_{s \geq 0} f(x^k + sd^k).$$

What to do for non-quadratic functions? Choose the best s^k ?

$$s^k = \arg\min_{s \geq 0} f(x^k + sd^k).$$

Seems good, but leads to zig-zagging convergence because

$$\nabla f(x^{k+1})^\top \nabla f(x^k) = 0.$$



To prove this

$$\frac{d}{ds} f(x^k - s\nabla f(x^k))\Big|_{s=s^k} = 0.$$

Using the chain-rule we have that

$$\frac{d}{ds} f(x^k - s\nabla f(x^k))\Big|_{s=s^k} = -s^k \nabla f(x^k - s^k \nabla f(x^k))^\top \nabla f(x^k) = 0.$$

Backtracking Line search

Instead of *best* step size, find a good one.

Algorithm 1 Backtracking Line Search(α, ρ, c)

1: Choose $\alpha > 0, \rho, c \in (0,\ 1)$.
2: **while** $f(x^k + \alpha d^k) \leq f(x^k) + c\,\alpha\,\nabla f(x^k)^\top d^k$ **do**
3: Update $\alpha = \rho\alpha$.

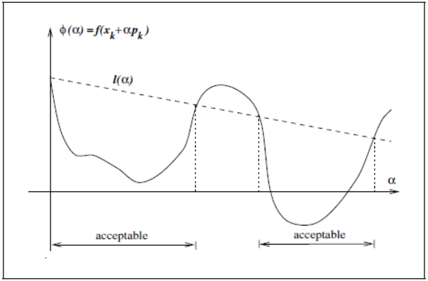
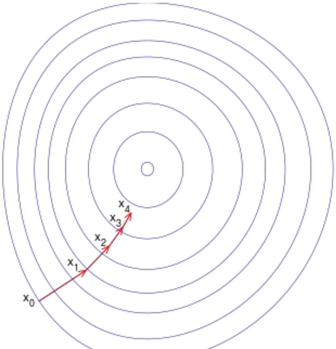


Figure: Where $\phi(\alpha) = f(x^k + \alpha d^k)$ and $l(\alpha) = f(x^k) + c\,\alpha\,\nabla f(x^k)^\top d^k$

Putting everything together with a stopping criteria

Algorithm 2 Gradient Descent

1: Choose $x^0 \in \mathbb{R}^n$.
2: **while** $\|\nabla f(x^k)\|_2 > \epsilon$ or $f(x^{k+1}) - f(x^k) \leq \epsilon$ **do**
3: Calculate $d^k = -\nabla f(x^k)$
4: Calculate s^k using Backtracking Line Search.
5: Update $x^{k+1} = x^k + s^k d^k$.



Gradient uses 1st order approximation. What about 2nd order?

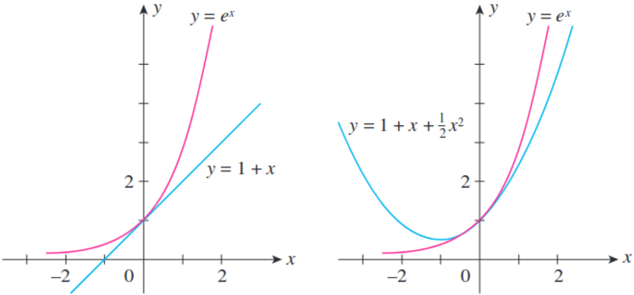


Figure: Comparing 1st order and 2nd Taylor of $f(x) = e^x$.

Local quadratic approximation using 2nd Taylor

$$q_k(x) = f(x^k) + \nabla f(x^k)^\top (x - x^k) + \frac{1}{2}(x - x^k)^\top \nabla^2 f(x^k)(x - x^k).$$

Newton's Method

Newton's method minimizes the local quadratic approximation.

$$q_k(x) = f(x^k) + \nabla f(x^k)^\top (x - x^k) + \frac{1}{2}(x - x^k)^\top \nabla^2 f(x^k)(x - x^k).$$

Assume that $\nabla^2 f(x^k)$ is invertible. Let x^{k+1} be the point that solves

$$\nabla_x q_k(x) = \nabla f(x^k) + \nabla^2 f(x^k)(x^{k+1} - x^k) = 0.$$

Isolating x^{k+1} we have

$$\boxed{x^{k+1} = x^k - \nabla^2 f(x^k)^{-1} \nabla f(x^k)}.$$

Newton's method can converge at a quadratic speed. Much faster than Gradient Descent.

Theorem

Let $f(x)$ be a μ–strongly convex function:

$$v^\top \nabla^2 f(x) v \;\geq\; \mu \|v\|^2, \quad \forall x, v \in \mathbb{R}^n. \tag{16}$$

If the Hessian is also Lipschitz

$$\|\nabla^2 f(x) - \nabla^2 f(y)\|_2 \;\leq\; L\|x - y\|_2 \tag{17}$$

then Newton's method converges according to

$$\|x^{k+1} - x^*\|_2 \leq \frac{L}{2\mu}\|x^k - x^*\|_2^2. \tag{18}$$

In particular if $\|x^0 - x^*\|_2 \leq \frac{\mu}{L}$, then for $k \geq 1$ we have that

$$\|x^k - x^*\|_2 \leq \frac{1}{2^{2^k}}\frac{\mu}{L}. \tag{19}$$

Proof:

$$
\begin{aligned}
x^{k+1} - x^* &= x^k - x^* - \nabla^2 f(x^k)^{-1} \left(\nabla f(x^k) - \nabla f(x^*)\right) \\
&= x^k - x^* - \nabla^2 f(x^k)^{-1} \int_{s=0}^{1} \nabla^2 f(x^k + s(x^* - x^k))(x^k - x^*) ds \\
&= \nabla^2 f(x^k)^{-1} \int_{s=0}^{1} \left(\nabla^2 f(x^k) - \nabla^2 f(x^k + s(x^* - x^k))\right)(x^k - x^*) ds
\end{aligned}
$$

Let $\delta^k := x^k - x^*$. Taking norms we have that

$$
\begin{aligned}
\|\delta^{k+1}\| &\leq \|\nabla^2 f(x^k)^{-1}\| \int_{s=0}^{1} \|\nabla^2 f(x^k) - \nabla^2 f(x^k + s(x^* - x^k))\| \, \|\delta^k\| ds \\
&\leq \frac{L}{\mu} \int_{s=0}^{1} s\|\delta^k\|^2 ds \\
&= \frac{L}{2\mu}\|\delta^k\|^2.
\end{aligned}
$$

Proof Part II: So now we have shown

$$\|x^{k+1} - x^*\| \ \leq ; \frac{L}{2\mu}\|x^k - x^*\|^2.$$

If $\|x^0 - x^*\| \leq \frac{\mu}{L}$, then by induction that

$$\|x^k - x^*\| \leq \frac{1}{2^{2^k}}\frac{\mu}{L}, \tag{20}$$

then we have that

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

which concludes the induction proof. □

Constrained Nonlinear Optimization

Let f, g_i and h_j be C^1 continuous functions, for $i = 1, \ldots, m$ and $j = 1, \ldots, p$. Consider the *constrained* optimization problem

$$\begin{aligned} \min_{x \in \mathbb{R}^n} \quad & f(x) \\ \text{subject to} \quad & g_i(x) \leq 0, \quad \text{for } i \in I. \\ & h_j(x) = 0, \quad \text{for } j \in J, \end{aligned} \tag{21}$$

where $I = \{1, \ldots, m\}$ and $J = \{1, \ldots, p\}$. Some notation:

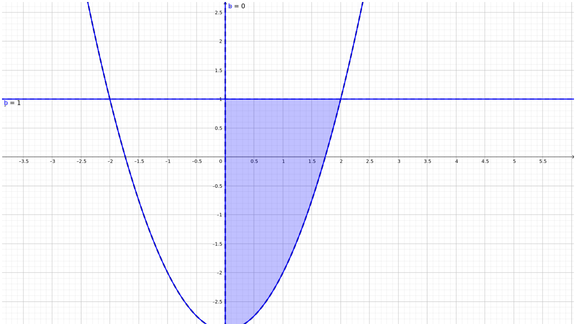
- Inequality constraints: $g_i(x) \leq 0$, for $i \in I$
- Equality constraints: $h_j(x) = 0$, for $j \in J$
- Feasible point x: Satisfies all inequality and equality constraints.
- Feasible set X: All the feasible points

$$X \stackrel{\text{def}}{=} \{x \in \mathbb{R}^n \,:\, g_i(x) \leq 0,\, h_j(x) = 0, \quad \text{for } i \in I, \text{ and } j \in J\}.$$

- Abbreviated form: $\min_{x \in X} f(x)$.

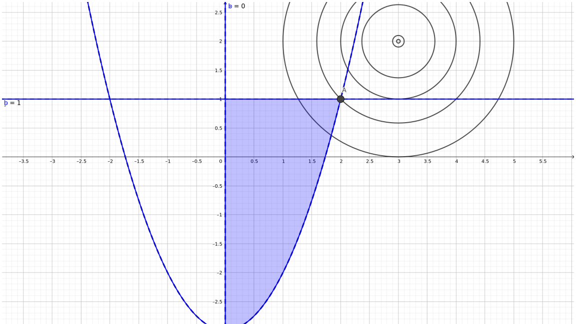
Exercise: Solve the following constrained nonlinear optimization problem graphically.

$$
\begin{aligned}
\min_{x \in \mathbb{R}^n} \quad & (x_1 - 3)^2 + (x_2 - 2)^2 \\
\text{subject to} \quad & x_1^2 - x_2 - 3 \leq 0, \\
& x_2 - 1 \leq 0, \\
& -x_1 \leq 0.
\end{aligned}
$$

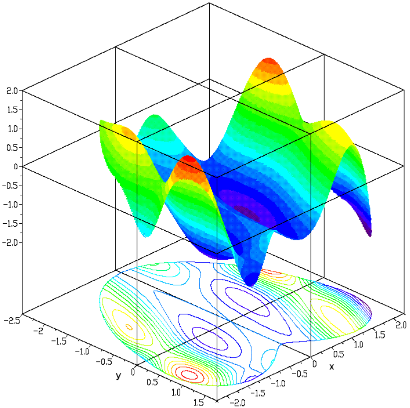


Exercise: Solve the following constrained nonlinear optimization problem graphically.

$$
\begin{aligned}
\min_{x\in\mathbb{R}^n} \quad & (x_1-3)^2+(x_2-2)^2 \\
\text{subject to} \quad & x_1^2 - x_2 - 3 \le 0, \\
& x_2 - 1 \le 0, \\
& -x_1 \le 0.
\end{aligned}
$$



Adding constraints can make the problem easy.
Easy example: If $X = \{x_0\}$ is a single point, we are done. If $X = \{x_0 + td_0, \quad \forall t \in \mathbb{R}\}$ it is easier.
But constraints can also make the problem harder (specially conceptually). Also even if g_i and h_j are smooth, the feasible set can be non-smooth. Hard example:



Theorem (Existence)

If the feasible set X is bounded and non-empty, then there exists a solution to $\min_{x \in X} f(x)$.

Proof.

Given that the sets $\mathbb{R}_- = [-\infty, 0]$ and $\{0\}$ are closed, by the continuity of g_i and h_j we have that X is closed. Indeed,

$$X = \left(\bigcap_{i=1}^{m} g_i^{-1}([-\infty, 0]) \right) \cap \left(\bigcap_{j=1}^{p} h_j^{-1}(\{0\}) \right),$$

and thus is a finite intersection of closed sets. By assumption X is bounded, thus it is compact. By the continuity of f we have that $f(X)$ is also compact (The Extreme value theorem). Consequently there exists a minimum in $f(X)$. □ □

Definition

We say that $f : \mathbb{R}^n \to \mathbb{R}$ is coercive if $\lim_{\|x\| \to \infty} f(x) = \infty$.

Theorem

If X is non-empty and f is coercive, then there exists a solution to $\min_{x \in X} f(x)$.

Proof.

Let $x_0 \in X$. Define $B_r := \{x : \|x\| \leq r\}$. Since f is coercive, there exists r such that for each x with $\|x\| \geq r$ we have that $f(x_k) \geq f(x_0)$.
Otherwise we would be able to construct a sequence x_k with $\|x_k\| \to \infty$ such that $f(x) \leq f(x_0)$, which contracts the coercivity of f.
Thus clearly the minimum of f is in B_r. Since B_r is bounded and closed, we have that $x_0 \in B_r \cap X$ thus it is bounded, closed and nonempty.
Again by the extreme value theorem, $f(x)$ attains its minimum in $B_r \cap X$, which is also the minimum in X. □ □

Given $x_0 \in X$ how can me move and still stay inside X?
If X was a polyhedra then d is a *feasible* or an *admissible* direction at $x_0 \in X$ if there exists $\epsilon > 0$ such that $x_0 + td \in X$ for all $0 \leq t \leq \epsilon$.

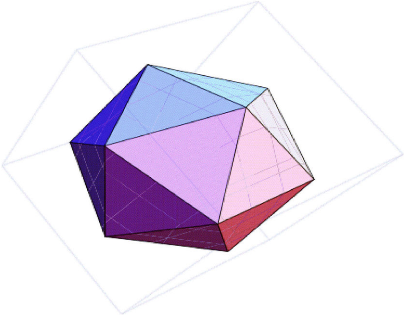


Figure: Difficult feasible set with objective function

For the case that the frontier of the feasible set is nonlinear, we need to consider a more general notion of feasible directions.

Definition

We say that d is an *admissible* direction at $x_0 \in X$ if there exists a C^1 differentiable curve $\phi : \mathbb{R}_+ \to \mathbb{R}^n$ such that

1. $\phi(0) = x_0$
2. $\phi'(0) = d$
3. There exists $\epsilon > 0$ such that $t \leq \epsilon$ we have $\phi(t) \in X$

We denote by $A(x_0)$ the set of admissable directions at x_0.

Some examples of admissable sets

- As a straight forward example, given $d \in \mathbb{R}^n$ let $X = \{x \mid \forall\alpha \in \mathbb{R}, x = \alpha d\}$. For any $x_0 \in X$ we have that $A(x_0) = X$.
- Consider the circle $X = \{(\cos(\theta), \sin(\theta)) \mid 0 \leq \theta \leq 2\pi\} \subset \mathbb{R}^2$. Then for every $x_0 = ((\cos(\theta_0), \sin(\theta_0)))$ we have that

$$A(x_0) = \{(-\alpha \sin(\theta), \alpha \cos(\theta)), \forall\alpha \in \mathbb{R}\}.$$

Taylor for Composition with Curve

Lemma

Let $\phi : \mathbb{R}_+ \to \mathbb{R}^n$ *be a* C^1 *curve as defined in Definition 15. Let* $f : \mathbb{R}^n \to \mathbb{R}$ *be continuously differentiable. Then the first order Taylor expansion of the composition* $f(\phi(t))$ *around* x_0 *can be written as*

$$f(\phi(t)) = f(x_0) + td^\top \nabla f(x_0) + t\hat{\epsilon}(t), \tag{22}$$

where $\lim_{t\to 0} \hat{\epsilon}(t) = 0$.

Proof: Since both f and ϕ are C^1, their composition is also C^1. Thus $f(\phi(t))$ first order Taylor expansion around $t = 0$ gives

$$f(\phi(t)) = f(\phi(0)) + t\frac{df(\phi(t))}{dt}|_{t=0} + t\epsilon(t).$$

Now plugging in $\phi(0) = x_0$ and using the chain-rule

$$\frac{df(\phi(t))}{dt}|_{t=0} \; = \; (\phi'(t)^\top \nabla f(\phi(t)))|_{t=0} \; = \; (d^\top \nabla f(x_0)). \quad \square$$

Theorem (Necessary Condition for Admissable Direction)

Let $I_0(x_0) = \{i \,:\, g_i(x_0) = 0,\ i \in I\}$ be the indexes of saturated inequalities. If $d \in A(x_0)$ is an admissable direction then

1. *For every $i \in I(x_0)$ we have that $d^\top \nabla g_i(x^0) \leq 0$.*
2. *For every $j \in J$ we have that $d^\top \nabla h_j(x^0) = 0$.*

Let $B(x_0)$ be the set of directions that satisfy the above two conditions. Thus $A(x_0) \subset B(x_0)$.

Proof 1. Let $i \in I(x_0)$. Let $\phi(t)$ be the curve associated to d. The 1st order Taylor expansion of g_i around x_0 in the d direction which is

$$\begin{aligned} g_i(\phi(t)) &\overset{(22)}{=} g_i(x_0) + td^\top \nabla g_i(x_0) + t\epsilon(t) \\ &= td^\top \nabla g_i(x_0) + t\epsilon(t) \ \leq \ 0, \end{aligned}$$

where we used $g_i(\phi(t)) \leq 0$ for t sufficiently small. Dividing by t gives

$$d^\top \nabla g_i(x^0) + \epsilon(t) \leq 0.$$

Letting $t \to 0$ we have that $d^\top \nabla g_i(x^0) \leq 0$.

Theorem (Necessary Condition for Admissable Direction)

Let $I_0(x_0) = \{i : g_i(x_0) = 0, \ i \in I\}$ be the indexes of saturated inequalities. If $d \in A(x_0)$ is an admissable direction then

1. *For every $i \in I(x_0)$ we have that $d^\top \nabla g_i(x^0) \leq 0$.*
2. *For every $j \in J$ we have that $d^\top \nabla h_j(x^0) = 0$.*

Let $B(x_0)$ be the set of directions that satisfy the above two conditions. Thus $A(x_0) \subset B(x_0)$.

Proof 2. Using the first order Taylor expansion of h_j around x_0 we have that

$$h_j(\phi(t)) \overset{(22)}{=} h_j(x_0) + td^\top \nabla h_j(x_0) + t\epsilon(t) = td^\top \nabla h_j(x_0) + t\epsilon(t) = 0.$$

Dividing by t and then taking the limit as $t \to 0$ gives $d^\top \nabla h_j(x^0) = 0$. $\square$

Cone of Feasible Directions

We refer to $B(x_0)$ as the cone of feasible directions.
Cones are easy to work with. We would like to use $B(x_0)$ instead $A(x_0)$.
But sometimes $B(x_0)$ and to $A(x_0)$ are not the same.

Example (Degeneracy)

Consider the constraint given by

$$h_1(x) = (x_1^2 + x_2^2 - 2)^2 = 0.$$

Thus

$$\nabla h_1(x) = 2(x_1^2 + x_2^2 - 2) \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} .$$

Every feasible point satisfies $\nabla h_1(x) = 0$. Consequently $B(x) = \mathbb{R}^2$ for every feasible point. Yet $h_1(x) = 0$ describes a circle, and clearly $A(x)$ is the tangent line at x. Thus we cannot use $\nabla h_1(x)$ to describe feasible directions. We would not have this problem if instead we used instead

$$h_1(x) = (x_1^2 + x_2^2 - 2) = 0.$$

To exclude these degeneracies, we impose the *Constraint qualifications*.

Definition

We say that the constraint qualifications hold at x_0 if for every $d \in B(x_0)$ there exists a sequence $(d_t)_{t=1}^{\infty} \in A(x_0)$ such that $d_t \to d$.

Constraint qualifications makes things easier.

Theorem (Necessary conditions)

Let x^* be a local minimum. If the constraint qualification holds at x^* then for every $d \in B(x^*)$ we have that $\nabla f(x^*)^\top d \geq 0$. Every direction in the feasible cone is not descent directions.

So we can check if x^* is a local minima by testing the directions in the feasible cone!

Theorem (Necessary conditions)

Let x^* be a local minimum. If the constraint qualification holds at x^* then for every $d \in B(x^*)$ we have that $\nabla f(x^*)^\top d \geq 0$. Every direction in the feasible cone is not descent directions.

Proof: Let $d_k \in A(x_*)$ be a sequence such that $d_k \to d$. Let ϕ_k be the curve associated to d_k. Using the first order Taylor expansion we have

$$f(\phi_k(t)) = f(x_*) + t\nabla f(x_*)^\top d_k + t\epsilon_k(t).$$

Since x_* is a local minima, there exists T for which $t \leq T$ we have that $f(x_*) \leq f(\phi_k(t))$. Consequently

$$t\nabla f(x_*)^\top d_k + t\epsilon_k(t) \;=\; f(\phi_k(t)) - f(x_*) \;\geq\; 0, \quad \text{for } t \leq T.$$

Dividing by t and taking the limit we have

$$\lim_{t\to 0} \nabla f(x_*)^\top d_k + \epsilon_k(t) = \nabla f(x_*)^\top d_k \geq 0.$$

Taking the limit in k concludes the proof. □