1. Gradient method

- gradient method, first-order methods
- convex functions
- Lipschitz continuity of gradient
- strong convexity
- analysis of gradient method

Gradient method

to minimize a convex differentiable function f: choose an initial point x_0 and repeat

$$x_{k+1} = x_k - t_k \nabla f(x_k), \qquad k = 0, 1, \ldots$$

step size t_k is constant or determined by line search

Advantages

- every iteration is inexpensive
- does not require second derivatives

Notation

- x_k can refer to kth element of a sequence, or to the kth component of vector x
- to avoid confusion, we sometimes use $x^{(k)}$ to denote elements of a sequence

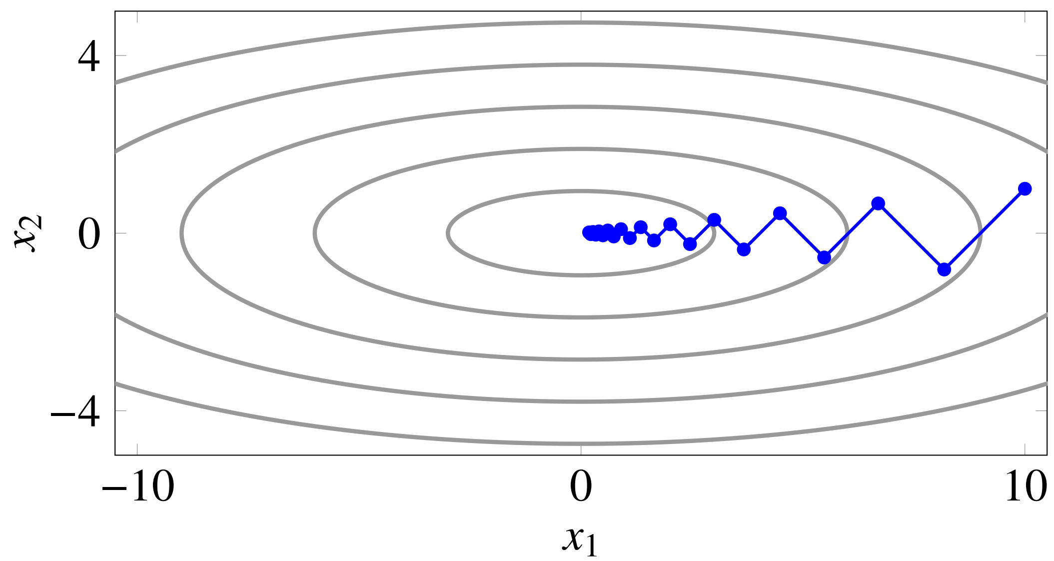
Quadratic example

$$f(x) = \frac{1}{2}(x_1^2 + \gamma x_2^2) \qquad \text{(with } \gamma > 1\text{)}$$

with exact line search and starting point $x^{(0)} = (\gamma, 1)$

$$\frac{\|x^{(k)} - x^\star\|_2}{\|x^{(0)} - x^\star\|_2} = \left(\frac{\gamma - 1}{\gamma + 1}\right)^k$$

where $x^\star = 0$

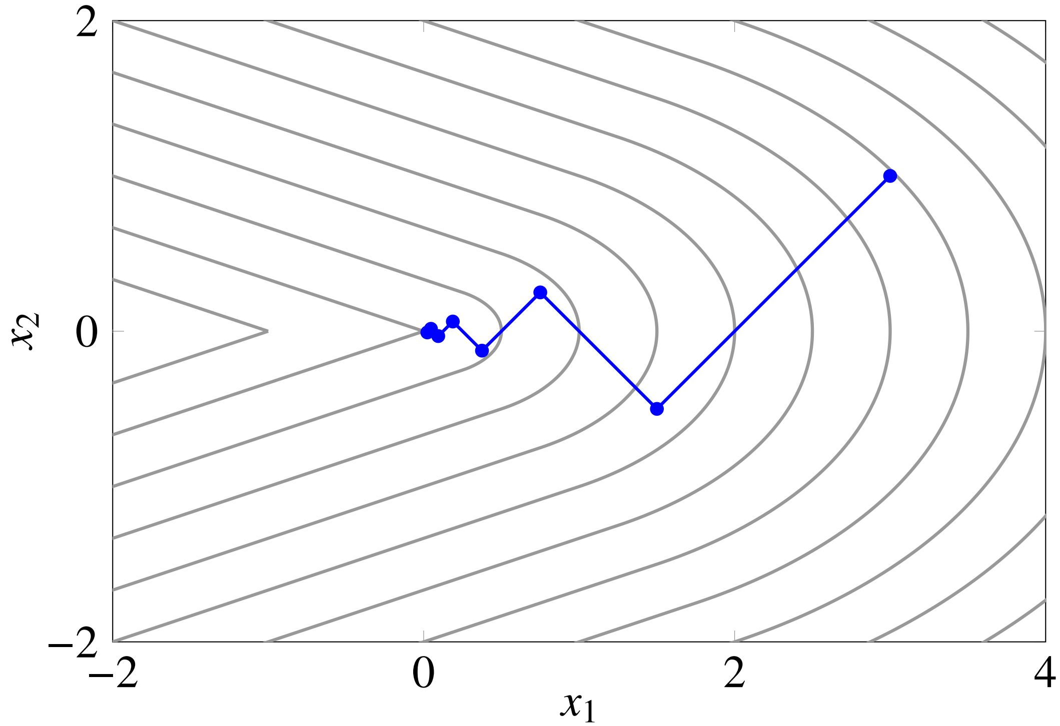


gradient method is often slow; convergence very dependent on scaling

Nondifferentiable example

$$f(x) = \sqrt{x_1^2 + \gamma x_2^2} \quad \text{if } |x_2| \le x_1, \qquad f(x) = \frac{x_1 + \gamma |x_2|}{\sqrt{1+\gamma}} \quad \text{if } |x_2| > x_1$$

with exact line search, starting point $x^{(0)} = (\gamma, 1)$, converges to non-optimal point



gradient method does not handle nondifferentiable problems

First-order methods

address one or both shortcomings of the gradient method

Methods for nondifferentiable or constrained problems

- subgradient method
- proximal gradient method
- smoothing methods
- cutting-plane methods

Methods with improved convergence

- conjugate gradient method
- accelerated gradient method
- quasi-Newton methods

Outline

- gradient method, first-order methods
- **convex functions**
- Lipschitz continuity of gradient
- strong convexity
- analysis of gradient method

Convex function

a function f is *convex* if $\mathrm{dom}\, f$ is a convex set and *Jensen's inequality* holds:

$$f(\theta x + (1-\theta)y) \leq \theta f(x) + (1-\theta) f(y) \quad \text{for all } x, y \in \mathrm{dom}\, f,\ \theta \in [0,1]$$

First-order condition

for (continuously) differentiable f, Jensen's inequality can be replaced with

$$f(y) \geq f(x) + \nabla f(x)^T (y - x) \quad \text{for all } x, y \in \mathrm{dom}\, f$$

Second-order condition

for twice differentiable f, Jensen's inequality can be replaced with

$$\nabla^2 f(x) \succeq 0 \quad \text{for all } x \in \mathrm{dom}\, f$$

Strictly convex function

f is *strictly convex* if $\mathrm{dom}\, f$ is a convex set and

$$f(\theta x + (1-\theta)y) < \theta f(x) + (1-\theta)f(y) \quad \text{for all } x, y \in \mathrm{dom}\, f,\ x \neq y, \text{ and } \theta \in (0,1)$$

strict convexity implies that if a minimizer of f exists, it is unique

First-order condition

for differentiable f, strict Jensen's inequality can be replaced with

$$f(y) > f(x) + \nabla f(x)^T(y-x) \quad \text{for all } x, y \in \mathrm{dom}\, f,\ x \neq y$$

Second-order condition

note that $\nabla^2 f(x) \succ 0$ is not necessary for strict convexity (*cf.*, $f(x) = x^4$)

Monotonicity of gradient

a differentiable function f is convex if and only if $\mathrm{dom}\, f$ is convex and

$$(\nabla f(x) - \nabla f(y))^T (x - y) \geq 0 \quad \text{for all } x, y \in \mathrm{dom}\, f$$

i.e., the gradient $\nabla f : \mathbf{R}^n \to \mathbf{R}^n$ is a *monotone* mapping

a differentiable function f is strictly convex if and only if $\mathrm{dom}\, f$ is convex and

$$(\nabla f(x) - \nabla f(y))^T (x - y) > 0 \quad \text{for all } x, y \in \mathrm{dom}\, f,\ x \neq y$$

i.e., the gradient $\nabla f : \mathbf{R}^n \to \mathbf{R}^n$ is a *strictly monotone* mapping

Proof

- if f is differentiable and convex, then

$$f(y) \geq f(x) + \nabla f(x)^T(y - x), \qquad f(x) \geq f(y) + \nabla f(y)^T(x - y)$$

  combining the inequalities gives $(\nabla f(x) - \nabla f(y))^T(x - y) \geq 0$

- if ∇f is monotone, then $g'(t) \geq g'(0)$ for $t \geq 0$ and $t \in \mathbf{dom}\, g$, where

$$g(t) = f(x + t(y - x)), \qquad g'(t) = \nabla f(x + t(y - x))^T(y - x)$$

  hence

$$\begin{aligned} f(y) = g(1) = g(0) + \int_0^1 g'(t)\,dt &\geq g(0) + g'(0) \\ &= f(x) + \nabla f(x)^T(y - x) \end{aligned}$$

  this is the first-order condition for convexity

Outline

- gradient method, first-order methods
- convex functions
- **Lipschitz continuity of gradient**
- strong convexity
- analysis of gradient method

Lipschitz continuous gradient

the gradient of f is *Lipschitz continuous* with parameter $L > 0$ if

$$\|\nabla f(x) - \nabla f(y)\|_* \leq L\|x - y\| \quad \text{for all } x, y \in \operatorname{dom} f$$

- functions f with this property are also called L*-smooth*
- the definition does not assume convexity of f (and holds for $-f$ if it holds for f)
- in the definition, $\|\cdot\|$ and $\|\cdot\|_*$ are a pair of dual norms:

$$\|u\|_* = \sup_{v \neq 0} \frac{u^T v}{\|v\|} = \sup_{\|v\|=1} u^T v$$

this implies a generalized Cauchy–Schwarz inequality

$$|u^T v| \leq \|u\|_* \|v\| \quad \text{for all } u, v$$

Choice of norm

Equivalence of norms

- for any two norms $\|\cdot\|_{\mathrm{a}}$, $\|\cdot\|_{\mathrm{b}}$, there exist positive constants c_1, c_2 such that

$$c_1\|x\|_{\mathrm{b}} \le \|x\|_{\mathrm{a}} \le c_2\|x\|_{\mathrm{b}} \quad \text{for all } x$$

- constants depend on dimension; for example, for $x \in \mathbf{R}^n$,

$$\|x\|_2 \le \|x\|_1 \le \sqrt{n}\,\|x\|_2, \qquad \frac{1}{\sqrt{n}}\|x\|_2 \le \|x\|_\infty \le \|x\|_2$$

Norm in definition of Lipschitz continuity

- without loss of generality we can use the Euclidean norm $\|\cdot\| = \|\cdot\|_* = \|\cdot\|_2$
- the parameter L depends on choice of norm
- in complexity bounds, choice of norm can simplify dependence on dimensions

Quadratic upper bound

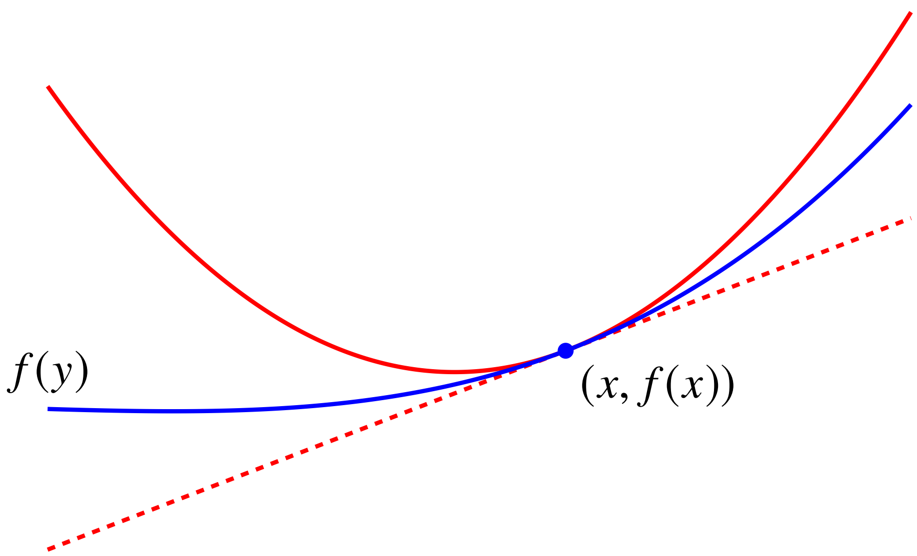
suppose ∇f is Lipschitz continuous with parameter L

- this implies (from the generalized Cauchy–Schwarz inequality) that

$$(\nabla f(x) - \nabla f(y))^T (x - y) \le L\|x - y\|^2 \quad \text{for all } x, y \in \operatorname{dom} f \tag{1}$$

- if $\operatorname{dom} f$ is convex, (1) is equivalent to

$$f(y) \le f(x) + \nabla f(x)^T (y - x) + \frac{L}{2}\|y - x\|^2 \quad \text{for all } x, y \in \operatorname{dom} f \tag{2}$$



Proof (of the equivalence of (1) and (2) if $\mathrm{dom}\, f$ is convex)

- consider arbitrary $x, y \in \mathrm{dom}\, f$ and define $g(t) = f(x + t(y - x))$
- $g(t)$ is defined for $t \in [0, 1]$ because $\mathrm{dom}\, f$ is convex
- if (1) holds, then

$$g'(t) - g'(0) = (\nabla f(x + t(y - x)) - \nabla f(x))^T (y - x) \leq tL\|x - y\|^2$$

integrating from $t = 0$ to $t = 1$ gives (2):

$$\begin{aligned} f(y) = g(1) = g(0) + \int_0^1 g'(t)\, dt &\leq g(0) + g'(0) + \frac{L}{2}\|x - y\|^2 \\ &= f(x) + \nabla f(x)^T (y - x) + \frac{L}{2}\|x - y\|^2 \end{aligned}$$

- conversely, if (2) holds, then (2) and the same inequality with x, y switched, *i.e.*,

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

can be combined to give $(\nabla f(x) - \nabla f(y))^T (x - y) \leq L\|x - y\|^2$

Consequence of quadratic upper bound

if $\operatorname{dom} f = \mathbf{R}^n$ and f has a minimizer $x^\star$, then

$$\frac{1}{2L}\|\nabla f(z)\|_*^2 \le f(z) - f(x^\star) \le \frac{L}{2}\|z - x^\star\|^2 \quad \text{for all } z$$

- right-hand inequality follows from upper bound property (2) at $x = x^\star$, $y = z$
- left-hand inequality follows by minimizing quadratic upper bound for $x = z$

$$\begin{aligned}
\inf_y f(y) \quad &\le \quad \inf_y \left(f(z) + \nabla f(z)^T(y - z) + \frac{L}{2}\|y - z\|^2 \right) \\
&= \quad \inf_{\|v\|=1} \inf_t \left(f(z) + t\nabla f(z)^T v + \frac{Lt^2}{2} \right) \\
&= \quad \inf_{\|v\|=1} \left(f(z) - \frac{1}{2L}(\nabla f(z)^T v)^2 \right) \\
&= \quad f(z) - \frac{1}{2L}\|\nabla f(z)\|_*^2
\end{aligned}$$

Co-coercivity of gradient

if f is convex with $\mathrm{dom}\, f = \mathbf{R}^n$ and ∇f is L-Lipschitz continuous, then

$$(\nabla f(x) - \nabla f(y))^T (x - y) \geq \frac{1}{L} \|\nabla f(x) - \nabla f(y)\|_*^2 \quad \text{for all } x, y$$

- this property is known as *co-coercivity* of ∇f (with parameter $1/L$)
- co-coercivity in turn implies Lipschitz continuity of ∇f (by Cauchy–Schwarz)
- hence, for differentiable convex f with $\mathrm{dom}\, f = \mathbf{R}^n$

$$\begin{aligned}
\text{Lipschitz continuity of } \nabla f \quad &\Rightarrow \quad \text{upper bound property (2) (equivalently, (1))} \\
&\Rightarrow \quad \text{co-coercivity of } \nabla f \\
&\Rightarrow \quad \text{Lipschitz continuity of } \nabla f
\end{aligned}$$

therefore the three properties are equivalent

Proof of co-coercivity: define two convex functions f_x, f_y with domain $\mathbf{R}^n$

$$f_x(z) = f(z) - \nabla f(x)^T z, \qquad f_y(z) = f(z) - \nabla f(y)^T z$$

- the two functions have L-Lipschitz continuous gradients
- $z = x$ minimizes $f_x(z)$; from the left-hand inequality on page 1.14,

$$\begin{aligned} f(y) - f(x) - \nabla f(x)^T(y - x) &= f_x(y) - f_x(x) \\ &\geq \frac{1}{2L}\|\nabla f_x(y)\|_*^2 \\ &= \frac{1}{2L}\|\nabla f(y) - \nabla f(x)\|_*^2 \end{aligned}$$

- similarly, $z = y$ minimizes $f_y(z)$; therefore

$$f(x) - f(y) - \nabla f(y)^T(x - y) \geq \frac{1}{2L}\|\nabla f(y) - \nabla f(x)\|_*^2$$

combining the two inequalities shows co-coercivity

Lipschitz continuity with respect to Euclidean norm

supose f is convex with $\operatorname{dom} f = \mathbf{R}^n$, and L-smooth for the Euclidean norm:

$$\|\nabla f(x) - \nabla f(y)\|_2 \leq L\|x - y\|_2 \quad \text{for all } x, y$$

- the equivalent property (1) states that

$$(\nabla f(x) - \nabla f(y))^T (x - y) \leq L(x - y)^T (x - y) \quad \text{for all } x, y$$

- this is monotonicity of $Lx - \nabla f(x)$, *i.e.*, equivalent to the property that

$$\frac{L}{2}\|x\|_2^2 - f(x) \quad \text{is a convex function}$$

- if f is twice differentiable, the Hessian of this function is $LI - \nabla^2 f(x)$:

$$\lambda_{\max}(\nabla^2 f(x)) \leq L \quad \text{for all } x$$

  is an equivalent characterization of L-smoothness

Outline

- gradient method, first-order methods
- convex functions
- Lipschitz continuity of gradient
- **strong convexity**
- analysis of gradient method

Strongly convex function

f is *strongly convex* with parameter $m > 0$ if $\operatorname{dom} f$ is convex and

$$f(\theta x + (1-\theta)y) \le \theta f(x) + (1-\theta)f(y) - \frac{m}{2}\theta(1-\theta)\|x-y\|^2$$

holds for all $x, y \in \operatorname{dom} f$, $\theta \in [0,1]$

- this is a stronger version of Jensen's inequality
- it holds if and only if it holds for f restricted to arbitrary lines:

$$f(x + t(y-x)) - \frac{m}{2}t^2\|x-y\|^2 \qquad (3)$$

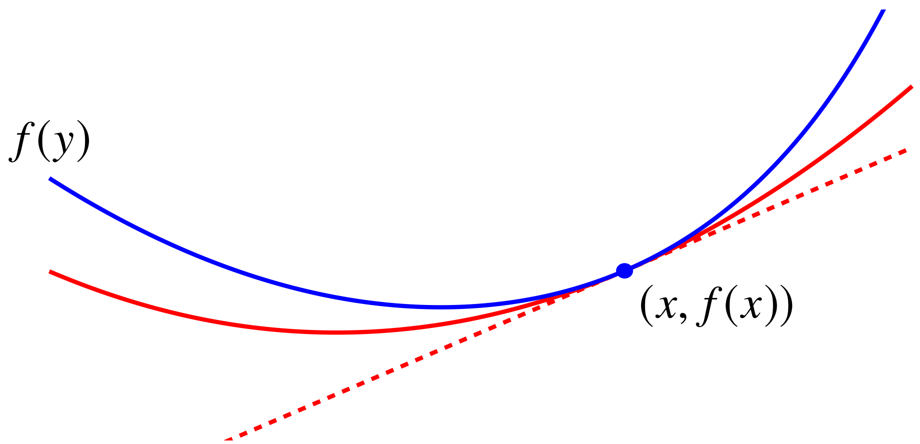
  is a convex function of t, for all $x, y \in \operatorname{dom} f$

- without loss of generality, we can take $\|\cdot\| = \|\cdot\|_2$
- however, the strong convexity parameter m depends on the norm used

Quadratic lower bound

if f is differentiable and m-strongly convex, then

$$f(y) \geq f(x) + \nabla f(x)^T(y-x) + \frac{m}{2}\|y-x\|^2 \quad \text{for all } x, y \in \operatorname{dom} f \tag{4}$$



- follows from the 1st order condition of convexity of (3)
- this implies that the sublevel sets of f are bounded
- if f is closed (has closed sublevel sets), it has a unique minimizer $x^\star$ and

$$\frac{m}{2}\|z-x^\star\|^2 \leq f(z) - f(x^\star) \leq \frac{1}{2m}\|\nabla f(z)\|_*^2 \quad \text{for all } z \in \operatorname{dom} f$$

(proof as on page 1.14)

Strong monotonicity

differentiable f is strongly convex if and only if $\mathrm{dom}\, f$ is convex and

$$(\nabla f(x) - \nabla f(y))^T(x - y) \geq m\|x - y\|^2 \quad \text{for all } x, y \in \mathrm{dom}\, f$$

this is called *strong monotonicity (coercivity)* of ∇f

Proof

- one direction follows from (4) and the same inequality with x and y switched
- for the other direction, assume ∇f is strongly monotone and define

$$g(t) = f(x + t(y - x)) - \frac{m}{2}t^2\|x - y\|^2$$

  then $g'(t)$ is nondecreasing, so g is convex

Strong convexity with respect to Euclidean norm

suppose f is m-strongly convex for the Euclidean norm:

$$f(\theta x + (1-\theta)y) \leq \theta f(x) + (1-\theta)f(y) - \frac{m}{2}\theta(1-\theta)\|x-y\|_2^2$$

for $x, y \in \operatorname{dom} f$, $\theta \in [0,1]$

- this is Jensen's inequality for the function

$$h(x) = f(x) - \frac{m}{2}\|x\|_2^2$$

- therefore f is strongly convex if and only if h is convex
- if f is twice differentiable, h is convex if and only if $\nabla^2 f(x) - mI \succeq 0$, or

$$\lambda_{\min}(\nabla^2 f(x)) \geq m \quad \text{for all } x \in \operatorname{dom} f$$

Extension of co-coercivity

suppose f is m-strongly convex and L-smooth for $\|\cdot\|_2$, and $\mathrm{dom}\, f = \mathbf{R}^n$

- then the function

$$h(x) = f(x) - \frac{m}{2}\|x\|_2^2$$

  is convex and $(L-m)$-smooth:

$$\begin{aligned} 0 &\le (\nabla h(x) - \nabla h(y))^T(x-y) \\ &= (\nabla f(x) - \nabla f(y))^T(x-y) - m\|x-y\|_2^2 \\ &\le (L-m)\|x-y\|_2^2 \end{aligned}$$

- co-coercivity of ∇h can be written as

$$(\nabla f(x) - \nabla f(y))^T(x-y) \ge \frac{mL}{m+L}\|x-y\|_2^2 + \frac{1}{m+L}\|\nabla f(x) - \nabla f(y)\|_2^2$$

  for all $x, y \in \mathrm{dom}\, f$

Outline

- gradient method, first-order methods
- convex functions
- Lipschitz continuity of gradient
- strong convexity
- **analysis of gradient method**

Analysis of gradient method

$$x_{k+1} = x_k - t_k \nabla f(x_k), \qquad k = 0, 1, \ldots$$

with fixed step size or backtracking line search

Assumptions

1. f is convex and differentiable with $\operatorname{dom} f = \mathbf{R}^n$
2. $\nabla f(x)$ is L-Lipschitz continuous with respect to the Euclidean norm, with $L > 0$
3. optimal value $f^\star = \inf_x f(x)$ is finite and attained at $x^\star$

Basic gradient step

- from quadratic upper bound (page 1.12) with $y = x - t\nabla f(x)$:

$$f(x - t\nabla f(x)) \leq f(x) - t(1 - \frac{Lt}{2}) \, \|\nabla f(x)\|_2^2$$

- therefore, if $x^+ = x - t\nabla f(x)$ and $0 < t \leq 1/L$,

$$f(x^+) \leq f(x) - \frac{t}{2}\|\nabla f(x)\|_2^2 \qquad (5)$$

- from (5) and convexity of f,

$$\begin{aligned} f(x^+) - f^\star &\leq \nabla f(x)^T(x - x^\star) - \frac{t}{2}\|\nabla f(x)\|_2^2 \\ &= \frac{1}{2t}\left(\|x - x^\star\|_2^2 - \left\|x - x^\star - t\nabla f(x)\right\|_2^2\right) \\ &= \frac{1}{2t}\left(\|x - x^\star\|_2^2 - \|x^+ - x^\star\|_2^2\right) \qquad (6) \end{aligned}$$

Descent properties

assume $\nabla f(x) \neq 0$ and $0 < t \leq 1/L$

- the inequality (5) shows that

$$f(x^+) < f(x)$$

- the inequality (6) shows that

$$\|x^+ - x^\star\|_2 < \|x - x^\star\|_2$$

in the gradient method, function value *and* distance to the optimal set decrease

Gradient method with constant step size

$$x_{k+1} = x_k - t\nabla f(x_k), \quad k = 0, 1, \ldots$$

- take $x = x_{i-1}$, $x^+ = x_i$ in (6) and add the bounds for $i = 1, \ldots, k$:

$$\begin{aligned}
\sum_{i=1}^{k} (f(x_i) - f^\star) &\leq \frac{1}{2t} \sum_{i=1}^{k} \left(\|x_{i-1} - x^\star\|_2^2 - \|x_i - x^\star\|_2^2 \right) \\
&= \frac{1}{2t} \left(\|x_0 - x^\star\|_2^2 - \|x_k - x^\star\|_2^2 \right) \\
&\leq \frac{1}{2t} \|x_0 - x^\star\|_2^2
\end{aligned}$$

- since $f(x_i)$ is non-increasing (see (5))

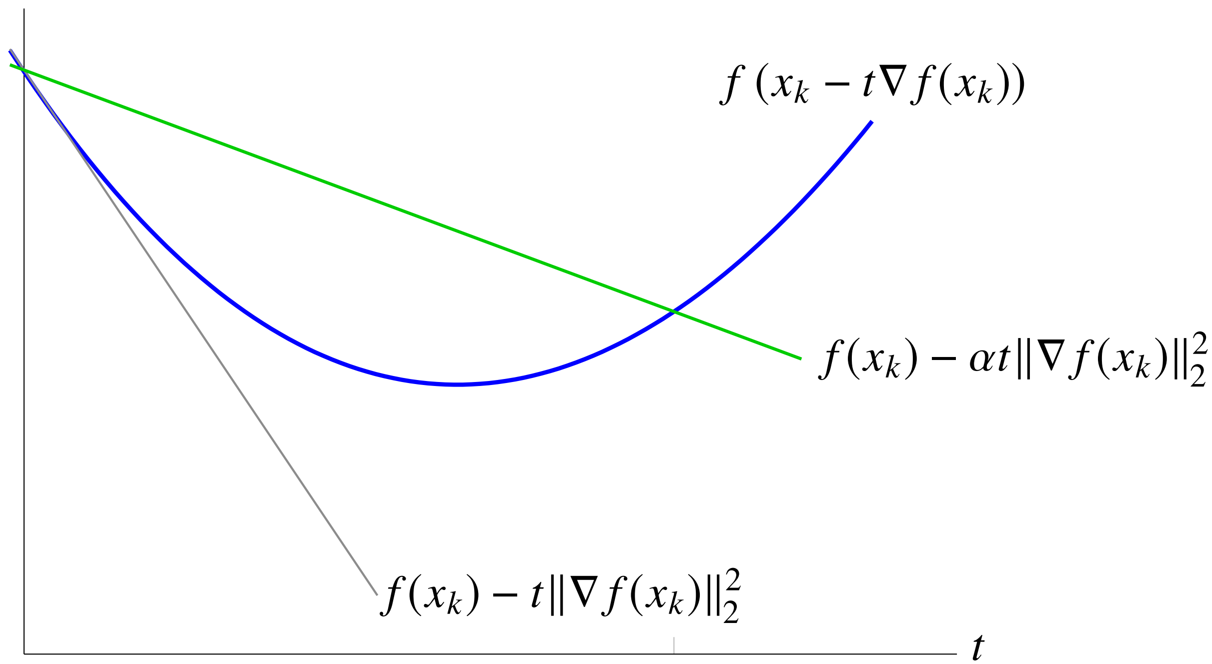
$$f(x_k) - f^\star \leq \frac{1}{k} \sum_{i=1}^{k} (f(x_i) - f^\star) \leq \frac{1}{2kt} \|x_0 - x^\star\|_2^2$$

Conclusion: number of iterations to reach $f(x_k) - f^\star \leq \epsilon$ is $O(1/\epsilon)$

Backtracking line search

initialize t_k at $\hat{t} > 0$ (for example, $\hat{t} = 1$) and take $t_k := \beta t_k$ until

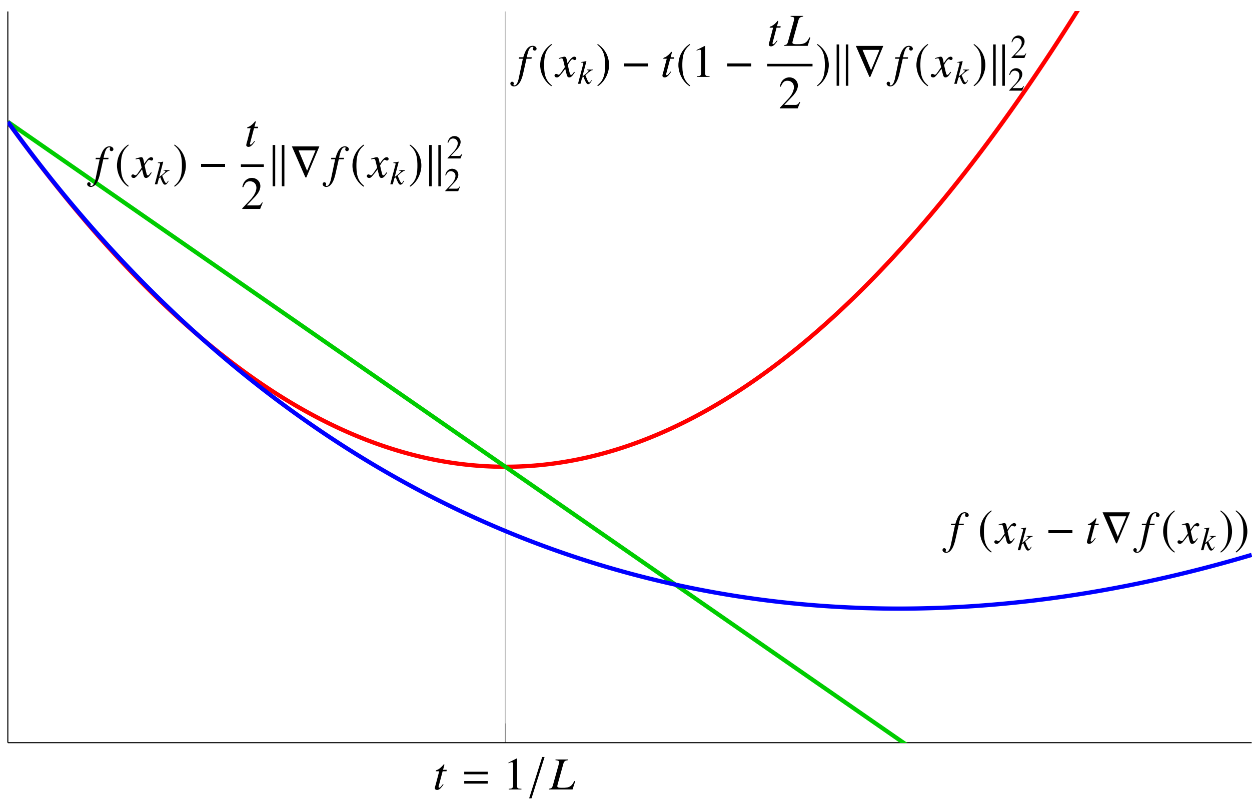
$$f(x_k - t_k \nabla f(x_k)) < f(x_k) - \alpha t_k \|\nabla f(x_k)\|_2^2$$



$0 < \beta < 1$; we will take $\alpha = 1/2$ (mostly to simplify proofs)

Analysis for backtracking line search

line search with $\alpha = 1/2$, if f has a Lipschitz continuous gradient



selected step size satisfies $t_k \geq t_{\min} = \min\{\hat{t}, \beta/L\}$

Gradient method with backtracking line search

- from line search condition and convexity of f,

$$\begin{aligned}
f(x_{i+1}) &\leq f(x_i) - \frac{t_i}{2}\|\nabla f(x_i)\|_2^2 \\
&\leq f^\star + \nabla f(x_i)^T(x_i - x^\star) - \frac{t_i}{2}\|\nabla f(x_i)\|_2^2 \\
&= f^\star + \frac{1}{2t_i}\left(\|x_i - x^\star\|_2^2 - \|x_{i+1} - x^\star\|_2^2\right)
\end{aligned}$$

- this implies $\|x_{i+1} - x^\star\|_2 \leq \|x_i - x^\star\|$, so we can replace t_i with $t_{\min} \leq t_i$:

$$f(x_{i+1}) - f^\star \leq \frac{1}{2t_{\min}}\left(\|x_i - x^\star\|_2^2 - \|x_{i-1} - x^\star\|_2^2\right)$$

- adding the upper bounds gives same $1/k$ bound as with constant step size

$$f(x_k) - f^\star \leq \frac{1}{k}\sum_{i=1}^{k}(f(x_i) - f^\star) \leq \frac{1}{2kt_{\min}}\|x_0 - x^\star\|_2^2$$

Gradient method for strongly convex functions

better results exist if we add strong convexity to the assumptions on p. 1.23

Analysis for constant step size

if $x^+ = x - t\nabla f(x)$ and $0 < t \leq 2/(m+L)$:

$$
\begin{aligned}
\|x^+ - x^\star\|_2^2 &= \|x - t\nabla f(x) - x^\star\|_2^2 \\
&= \|x - x^\star\|_2^2 - 2t\nabla f(x)^T(x - x^\star) + t^2\|\nabla f(x)\|_2^2 \\
&\leq (1 - t\frac{2mL}{m+L})\|x - x^\star\|_2^2 + t(t - \frac{2}{m+L})\|\nabla f(x)\|_2^2 \\
&\leq (1 - t\frac{2mL}{m+L})\|x - x^\star\|_2^2
\end{aligned}
$$

(step 3 follows from result on page 1.22)

Distance to optimum

$$\|x_k - x^\star\|_2^2 \le c^k \|x_0 - x^\star\|_2^2, \qquad c = 1 - t\frac{2mL}{m+L}$$

- implies (linear) convergence
- for $t = 2/(m+L)$, get $c = \left(\frac{\gamma - 1}{\gamma + 1}\right)^2$ with $\gamma = L/m$

Bound on function value (from page 1.14)

$$f(x_k) - f^\star \le \frac{L}{2}\|x_k - x^\star\|_2^2 \le \frac{c^k L}{2}\|x_0 - x^\star\|_2^2$$

Conclusion: number of iterations to reach $f(x_k) - f^\star \le \epsilon$ is $O(\log(1/\epsilon))$

Limits on convergence rate of first-order methods

First-order method: any iterative algorithm that selects x_{k+1} in the set

$$x_0 + \mathrm{span}\{\nabla f(x_0), \nabla f(x_1), \ldots, \nabla f(x_k)\}$$

Problem class: any function that satisfies the assumptions on page 1.23

Theorem (Nesterov): for every integer $k \leq (n-1)/2$ and every x_0, there exist functions in the problem class such that for any first-order method

$$f(x_k) - f^\star \geq \frac{3}{32} \frac{L\|x_0 - x^\star\|_2^2}{(k+1)^2}$$

- suggests $1/k$ rate for gradient method is not optimal
- more recent accelerated gradient methods have $1/k^2$ convergence (see later)

References

- A. Beck, *First-Order Methods in Optimization* (2017), chapter 5.
- Yu. Nesterov, *Lectures on Convex Optimization* (2018), section 2.1. (The result on page 1.32 is Theorem 2.1.7 in the book.)
- B. T. Polyak, *Introduction to Optimization* (1987), section 1.4.
- The example on page 1.4 is from N. Z. Shor, *Nondifferentiable Optimization and Polynomial Problems* (1998), page 37.