# Optimization for Control Engineering

## Plan

- Introduction (Javier): optimality conditions
- Part 1 (Javier): first-order methods
- Part 2 (Diego): sequential quadratic programming and interior-point methods

## Main references

- Beck, *First-order Methods in Optimization*, SIAM 2017
- Nocedal & Wright, *Numerical Optimization*, Springer 2006

# Optimization model

Problem of the form

$$\begin{array}{ll} \min\limits_{x} & f(x) \\ \text{s.t.} & x \in \mathcal{X}. \end{array}$$

We will concentrate on problems where $f : \mathbb{R}^n \to \mathbb{R}$ and $\mathcal{X} \subseteq \mathbb{R}^n$ is of the form

$$\mathcal{X} = \{x \in \mathbb{R}^n : c_i(x) = 0 \text{ for } i \in \mathcal{E} \text{ and } c_i(x) \geq 0 \text{ for } i \in \mathcal{I}\}$$

for some functions $c_i : \mathbb{R}^n \to \mathbb{R},\ i \in \mathcal{E} \cup \mathcal{I}$.

In this case it is customary to write the above problem as follows

$$\begin{array}{ll} \min\limits_{x} & f(x) \\ \text{s.t.} & c_i(x) = 0,\ i \in \mathcal{E} \\ & c_i(x) \geq 0,\ i \in \mathcal{I}. \end{array}$$

# Optimality conditions (unconstrained case)

Suppose $f : \mathbb{R}^n \to \mathbb{R}$ is differentiable. If $x^\star$ solves

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

then $\nabla f(x^\star) = 0$. The converse also holds if $f$ is convex.

Next: how the above optimality conditions extend to the case when we have constraints.

# Lagrangian function

The *Lagrangian function* of the problem

$$\begin{array}{ll} \min\limits_{x} & f(x) \\ \text{s.t.} & c_i(x) = 0,\ i \in \mathcal{E} \\ & c_i(x) \geq 0,\ i \in \mathcal{I}. \end{array}$$

is

$$L(x, \lambda) = f(x) - \lambda^{\mathsf{T}} c(x).$$

Observe that the above problem can be recast as follows

$$\min_{x} \max_{\substack{\lambda \\ \lambda_{\mathcal{I}} \geq 0}} \ L(x, \lambda)$$

# Optimality conditions (equality constraints only)

Suppose $f : \mathbb{R}^n \to \mathbb{R}$ and $c_i : \mathbb{R}^n \to \mathbb{R}, i \in \mathcal{E}$ are differentiable.
Suppose $x^\star$ solves

$$\begin{array}{ll} \min\limits_x & f(x) \\ \text{s.t.} & c_i(x) = 0,\ i \in \mathcal{E} \end{array}$$

and $\{\nabla c_i(x^\star),\ i \in \mathcal{E}\}$ is linearly independent. Then there exists $\lambda^\star$ such that $\nabla L(x^\star, \lambda^\star) = 0$ or equivalently

$$\begin{aligned} \nabla f(x^\star) - \sum_{i \in \mathcal{E}} \lambda_i^\star \nabla c_i(x^\star) &= 0 \\ c_i(x^\star) &= 0,\ i \in \mathcal{E}. \end{aligned}$$

The converse also holds if $f$ is convex and $c_i,\ i \in \mathcal{E}$ are affine.

# Optimality conditions (equality & inequality constraints)

Suppose $f : \mathbb{R}^n \to \mathbb{R}$ and $c_i : \mathbb{R}^n \to \mathbb{R}, i \in \mathcal{E} \cup \mathcal{I}$ are differentiable and consider the problem

$$\begin{array}{ll} \min\limits_{x} & f(x) \\ \text{s.t.} & c_i(x) = 0,\ i \in \mathcal{E} \\ & c_i(x) \geq 0,\ i \in \mathcal{I}. \end{array}$$

Given a feasible point $x^\star$, let $\mathcal{A}(x^\star)$ denote the set of *active constraints* at $x^\star$, that is,

$$\mathcal{A}(x^\star) := \mathcal{E} \cup \{i \in \mathcal{I} : c_i(x^\star) = 0\}.$$

# Optimality conditions (equality & inequality constraints)

Suppose $f : \mathbb{R}^n \to \mathbb{R}$ and $c_i : \mathbb{R}^n \to \mathbb{R}, i \in \mathcal{E} \cup \mathcal{I}$ are differentiable.
Suppose $x^\star$ solves

$$\begin{array}{ll} \min\limits_x & f(x) \\ \text{s.t.} & c_i(x) = 0,\ i \in \mathcal{E} \\ & c_i(x) \geq 0,\ i \in \mathcal{I} \end{array}$$

and the set $\{\nabla c_i(x^\star),\ i \in \mathcal{A}(x^\star)\}$ is linearly independent.
Then there exists $\lambda^\star$ such that

$$\begin{aligned} \nabla f(x^\star) - \sum_{i \in \mathcal{A}(x^\star)} \lambda_i^\star \nabla c_i(x^\star) = 0 \\ c_i(x^\star) = 0,\ i \in \mathcal{E} \\ \lambda_i^\star \geq 0,\ c_i(x^\star) \geq 0, i \in \mathcal{I} \\ \lambda_i^\star c_i(x^\star) = 0, i \in \mathcal{I}. \end{aligned}$$

The converse also holds if $f$ and $-c_i, i \in \mathcal{I}$ are convex and $c_i,\ i \in \mathcal{E}$ are affine.

# Special case: linear programming

Suppose $A \in \mathbb{R}^{m \times n}, c \in \mathbb{R}^n, b \in \mathbb{R}^m$ and consider a linear program in *standard form:*

$$\begin{array}{ll} \min\limits_{x} & c^\mathsf{T} x \\ \text{s.t.} & Ax = b \\ & x \geq 0. \end{array}$$

In this case $x^\star$ is an optimal solution if and only if there exist $\lambda^\star \in \mathbb{R}^m, s^\star \in \mathbb{R}^n$ such that

$$\begin{aligned} A^\mathsf{T} \lambda^\star + s^\star &= c \\ Ax^\star &= b \\ x^\star \geq 0,\ s^\star &\geq 0 \\ x_i^\star s_i^\star &= 0,\ i = 1, \ldots, n. \end{aligned}$$

Furthermore, the above holds if and only if $(\lambda^\star, s^\star)$ solves the dual problem

$$\begin{array}{ll} \max\limits_{\lambda, s} & b^\mathsf{T} \lambda \\ \text{s.t.} & A^\mathsf{T} \lambda + s = c \\ & s \geq 0. \end{array}$$

# Special case: quadratic programming (equality constraints)

Suppose $A \in \mathbb{R}^{m\times n}, c \in \mathbb{R}^n, b \in \mathbb{R}^m, G \in \mathbb{R}^{n\times n}$ and consider the quadratic program

$$\begin{array}{ll} \min\limits_x & \frac{1}{2}x^\mathsf{T}Gx + c^\mathsf{T}x \\ \text{s.t.} & Ax = b. \end{array}$$

If $x^\star$ is an optimal solution then there exist $\lambda^\star \in \mathbb{R}^m$ such that

$$\begin{aligned} Gx^\star - A^\mathsf{T}\lambda^\star &= -c \\ Ax^\star &= b. \end{aligned}$$

The converse also holds when $G$ is positive semidefinite.

# Special case: quad programming (inequality constraints)

Suppose $A \in \mathbb{R}^{m\times n}, c \in \mathbb{R}^n, b \in \mathbb{R}^m, G \in \mathbb{R}^{n\times n}$ and consider the quadratic program

$$\begin{array}{ll} \min\limits_x & \frac{1}{2}x^\mathsf{T}Gx + c^\mathsf{T}x \\ \text{s.t.} & Ax \geq b. \end{array}$$

If $x^\star$ is an optimal solution then there exist $\lambda^\star \in \mathbb{R}^m$ such that

$$\begin{aligned} Gx^\star - A^\mathsf{T}\lambda^\star &= -c \\ Ax^\star &\geq b \\ \lambda^\star &\geq 0 \\ \lambda_i^\star(Ax^\star - b)_i &= 0,\ i = 1, \ldots, m. \end{aligned}$$

The converse also holds when $G$ is positive semidefinite.

# Algorithms for optimization

Consider the problem

$$\begin{array}{ll} \min\limits_{x} & f(x) \\ \text{s.t.} & x \in \mathcal{X}. \end{array} \tag{P}$$

Algorithm to "solve" (P):

- Construct a sequence $x_k \in \mathbb{R}^n, \; k = 0, 1, \ldots$ that hopefully converges to a solution to (P)
- Algorithm depends on what kind of "oracles" are available for $f$ and $\mathcal{X}$ and the type of operations that are performed at each main iteration
- "Simple" algorithms perform low-cost operations (memory and computation) but are usually slow.

  "Sophisticated" algorithms require more costly operations but are usually much faster. They also apply to a wider variety of problems.

# Two main ideas

Consider the unconstrained problem

$$\min_x f(x).$$

## Gradient descent

Given $x$ get a (hopefully better) new point $x_+$ via

$$\begin{aligned} x_+ &:= x - t \cdot \nabla f(x) \\ &= \operatorname{argmin}_y \left\{ f(x) + \langle \nabla f(x), y - x \rangle + \frac{1}{2t} \|y - x\|^2 \right\}. \end{aligned}$$

## Newton's method

Given $x$ get a (hopefully better) new point $x_+$ via

$$\begin{aligned} x_+ &:= x - \nabla^2 f(x)^{-1} \nabla f(x) \\ &= \operatorname{argmin}_y \left\{ f(x) + \langle \nabla f(x), y - x \rangle + \frac{1}{2} \langle \nabla^2 f(x)(y - x), y - x \rangle \right\}. \end{aligned}$$

(First step requires $\nabla^2 f(x)$ non-singular and second one requires $\nabla^2 f(x)$ positive definite.)

# Newton's method for root finding

Suppose $F : \mathbb{R}^n \to \mathbb{R}^n$ is differentiable and consider the system of $n$ equations and $n$ unknowns

$$F(x) = 0.$$

## Newton's method

Given $x$ get a (hopefully better) new point $x_+$ via

$$x_+ := x - F'(x)^{-1} F(x).$$

The latter is the solution for the variable $y$ of the linear system of equations

$$F(x) + F'(x)(y - x) = 0.$$

# Main agenda

### Part 1 (Javier)

Emphasis on unconstrained convex optimization. First-order methods: gradient descent and fast gradient descent.

### Part 2 (Diego)

General constrained optimization. Sequential quadratic programming. Interior-point methods.

*Gradient descent*

# Gradient descent (Cauchy 1847)

Let $f : \mathbb{R}^n \to \mathbb{R}$ be a convex differentiable function. Solve

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

via

$$x_{k+1} := x_k - t_k \cdot \nabla f(x_k).$$

Common shorthand: drop indices and write main update as

$$x_+ = x - t \cdot \nabla f(x).$$

Observe

$$x - t \cdot \nabla f(x) = \operatorname*{argmin}_y \left\{ f(x) + \langle \nabla f(x), y - x \rangle + \frac{1}{2t} \|y - x\|^2 \right\}.$$

Throughout our discussion: $\| \cdot \| = \| \cdot \|_2$.

# Gradient descent (continued)

Solve

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

via

$$\begin{aligned} x_{k+1} &:= x_k - t_k \cdot \nabla f(x_k) \\ &= \operatorname*{argmin}_y \left\{ f(x_k) + \langle \nabla f(x_k), y - x_k \rangle + \frac{1}{2t_k} \|y - x_k\|^2 \right\} \end{aligned}$$

## Natural questions

- How to choose $t_k$?
- Can we guarantee that the iterates $x_k,\ k = 0, 1, \ldots$ and/or iterate values $f(x_k)$ converge? If so, how fast?

# Choice of step-size in gradient descent

## Common approach

Pick $t > 0$ so that $x_+ = x - t \cdot \nabla f(x)$ satisfies

$$\begin{aligned} f(x_+) &\leq \min_y \left\{ f(x) + \langle \nabla f(x), y - x \rangle + \frac{1}{2t}\|y - x\|^2 \right\} \\ &= f(x) - \frac{t}{2}\|\nabla f(x)\|^2. \end{aligned} \tag{DC}$$

## Backtracking

We usually want $t$ as large as possible so that (DC) holds.

We can do that via "backtracking": pick initial $t > 0$ and scale it up or down until (DC) just holds.

# Convergence of gradient descent

## Notation & blanket assumption

Let $\bar{f} := \min_{x \in \mathbb{R}^n} f(x)$ is finite and $\bar{X} := \operatorname{argmin}_{x \in \mathbb{R}^n} f(x) \neq \emptyset$.

## Theorem

*Let $f : \mathbb{R}^n \to \mathbb{R}$ be convex and differentiable. If the step-sizes satisfy* (DC) *then the iterates $x_k, \; k = 0, 1, \ldots$ generated by the proximal gradient algorithm satisfy*

$$f(x_k) - \bar{f} \leq \frac{\operatorname{dist}(x_0, \bar{X})^2}{2 \sum_{i=0}^{k-1} t_i}, \;\; k = 1, 2, \ldots.$$

In particular if $t_k \geq \frac{1}{L} > 0$ for some $L > 0$ then

$$f(x_k) - \bar{f} \leq \frac{L \cdot \operatorname{dist}(x_0, \bar{X})^2}{2k}, \;\; k = 1, 2, \ldots.$$

# Proof of convergence of gradient descent

## Convex conjugate

Let $\phi : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$. Define $\phi^* : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ via

$$\phi^*(v) = \sup_{x \in \mathbb{R}^n} \{\langle v, x\rangle - \phi(x)\}.$$

## Lemma

*Let $f : \mathbb{R}^n \to \mathbb{R}$ be convex and differentiable. If the step-sizes satisfy* (DC) *then the gradient iterates satisfy*

$$f(x_k) \leq -f^*(v_k) - d_k^*(-v_k)$$

*where* $v_k := \frac{\sum_{i=0}^{k-1} t_i \nabla f(x_i)}{\sum_{i=0}^{k-1} t_i}$ *and* $d_k(x) := \frac{\|x - x_0\|^2}{2\sum_{i=0}^{k-1} t_i}$.

*Proof of previous Theorem.* For $\bar{x} \in \bar{X}$ we have

$$-f^*(v_k) - d_k^*(-v_k) \leq -\langle v_k, \bar{x}\rangle + \bar{f} + \langle v_k, \bar{x}\rangle + d_k(\bar{x}) = \bar{f} + \frac{\|\bar{x} - x_0\|^2}{2\sum_{i=0}^{k-1} t_i}.$$

# $L$-smoothness

When can we ensure that $t_k \geq 1/L$ for some constant $L > 0$?

$L$-smoothness
If $f : \mathbb{R}^n \to \mathbb{R}$ is convex, differentiable, and $\nabla f$ is $L$-Lipschitz continuous then $f$ satisfies the following $L$-smoothness condition

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

for all $x, y$.

In this case the (DC) condition holds for $t_k = \frac{1}{L}$ or possibly larger.

# Lower bound for *any* gradient method

Suppose $f : \mathbb{R}^n \to \mathbb{R}$ is convex and differentiable and consider an algorithm such that

$$x_{k+1} \in x_0 + \mathrm{span}\{\nabla f(x_0), \ldots, \nabla f(x_k)\}.$$

How good could that kind of algorithm be?

**Theorem**

*For all $x_0 \in \mathbb{R}^n$ there exists a strictly convex and differentiable $f : \mathbb{R}^n \to \mathbb{R}$ with $L$-Lipschitz $\nabla f$ and such that*

$$f(x_k) - \bar{f} \geq \frac{3L}{32(k+1)^2}\|x_0 - \bar{x}\|^2 \quad \textit{for} \;\; k \leq n/2.$$

Observe: for $L$-Lipschitz $\nabla f$ gradient descent iterates satisfy

$$f(x_k) - \bar{f} \leq \frac{L}{2k}\|x_0 - \bar{x}\|^2.$$

Is it possible to do better?

*Fast gradient descent*

# Fast gradient descent (Nesterov 1983)

## Main idea

Generate two different sequences that have the same initial point

$$y_0 = x_0$$

and are updated via

$$x_{k+1} = y_k - t_k \cdot \nabla f(y_k)$$

and

$$y_{k+1} = x_{k+1} + \beta_k \cdot (x_{k+1} - x_k)$$

for some $\beta_k \geq 0,\ k = 0, 1, \ldots$.

## Observe

The sequence $y_k,\ k = 0, 1, \ldots$ includes some "momentum".

# Convergence of fast gradient descent

## Recall notation & blanket assumption

Let $\bar{f} := \min_{x \in \mathbb{R}^n} f(x)$ is finite and $\bar{X} := \underset{x \in \mathbb{R}^n}{\text{argmin}}\, f(x) \neq \emptyset$.

## Theorem

*Suppose* $\beta_k = \frac{k}{k+3}$ *and* $t_k \geq 1/L,\ k = 0, 1, 2, \ldots$ *are non-increasing and satisfy* (DC), *that is,*

$$f(x_{k+1}) \leq f(y_k) - \frac{t_k}{2} \|\nabla f(y_k)\|^2.$$

*Then the iterates generated by the fast gradient algorithm satisfy*

$$f(x_k) - \bar{f} \leq \frac{2L}{(k+1)^2} \cdot \operatorname{dist}(x_0, \bar{X})^2.$$

## Another popular choice for $\beta_k$

Take $\beta_k := \frac{\theta_{k+1}(1-\theta_k)}{\theta_k}$, where $\theta_0 = 1$ and $\theta_{k+1}^2 = \theta_k^2(1 - \theta_{k+1})$.

# Proof of convergence of fast gradient descent (simplified)

Consider the special case when $t_k = 1/L$ and $\beta_k$ is chosen via $\theta_k$.

In this case we can rewrite the updates as follows

$$y_k = (1 - \theta_k) \cdot x_k + \theta_k \cdot z_k$$
$$x_{k+1} = (1 - \theta_k) \cdot x_k + \theta_k \cdot z_{k+1}$$

where

$$z_{k+1} = z_k - \frac{1}{\theta_k L} \cdot \nabla f(y_k).$$

## Properties of $\theta_k$

If $\theta_0 = 1$ and $\theta_{k+1}^2 = \theta_k^2(1 - \theta_{k+1}),\ k = 0, 1, \ldots$ then

$$\sum_{i=0}^{k} \frac{1}{\theta_i} = \theta_k^2 \leq \frac{4}{(k+2)^2},\ k = 0, 1, \ldots.$$

# Proof of convergence of fast gradient descent (simplified)

## Lemma

*Under the above assumptions the fast gradient iterates satisfy*

$$f(x_k) \leq -f^*(v_k) - d_k^*(-v_k)$$

*where* $v_k := \theta_{k-1}^2 \cdot \sum_{i=0}^{k-1} \frac{\nabla f(y_i)}{\theta_i}$ *and* $d_k(x) := \frac{L\theta_{k-1}^2}{2}\|x - x_0\|^2$.

*Proof of previous Theorem.* For $\bar{x} \in \bar{X}$ we have

$$\begin{aligned} -f^*(v_k) - d_k^*(-v_k) &\leq -\langle v_k, \bar{x}\rangle + \bar{f} + \langle v_k, \bar{x}\rangle + d_k(\bar{x}) \\ &= \bar{f} + \frac{L\theta_{k-1}^2}{2}\|\bar{x} - x_0\|^2 \\ &\leq \bar{f} + \frac{2L}{(k+1)^2}\|\bar{x} - x_0\|^2. \end{aligned}$$

---

Recall $\phi^*(v) = \sup_x\{\langle v, x\rangle - \phi(x)\}$.

# Examples

## Least squares

Let $A \in \mathbb{R}^{m \times n},\ b \in \mathbb{R}^m$ and consider the *loss* function

$$f(x) = \frac{1}{2}\|Ax - b\|^2.$$

## Logistic regression

Let $X \in \mathbb{R}^{N \times p},\ y \in \{0,1\}^N$ and consider the *logistic loss* function

$$\begin{aligned} f(\beta) &= -\sum_{i=1}^{N} \left\{ y_i \log\left(\frac{1}{1+e^{-X_i\beta}}\right) + (1-y_i)\log\left(1 - \frac{1}{1+e^{-X_i\beta}}\right)\right\} \\ &= -y^\mathsf{T} X\beta + \mathbf{1}^\mathsf{T}\log(1+e^{X\beta}). \end{aligned}$$

See Python code: `Algorithms.py`, `Functions.py`, `Example1.py`

## *Proximal gradient methods*

# Projected gradient method

Let $f : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ be a differentiable convex function and $\mathcal{X} \subseteq \mathrm{dom}(f)$ be a closed convex set. Solve

$$\min_{x \in \mathcal{X}} f(x)$$

via

$$\begin{aligned} x_{k+1} &:= \mathsf{Proj}_{\mathcal{X}}(x_k - t_k \cdot \nabla f(x_k)) \\ &= \operatorname*{argmin}_{y \in \mathcal{X}} \|y - (x_k - t_k \cdot \nabla f(x_k))\|^2 \\ &= \operatorname*{argmin}_{y \in \mathcal{X}} \left\{ f(x_k) + \langle \nabla f(x_k), y - x_k \rangle + \frac{1}{2t_k} \|y - x_k\|^2 \right\} \end{aligned}$$

Projected gradient is a special case of proximal gradient (next).

# Proximal gradient method (Lions & Mercier 1979)

a.k.a. forward-backward splitting

Let $f : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ be differentiable and convex, and $\psi : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ be closed and convex with $\mathrm{dom}(\psi) \subseteq \mathrm{dom}(f)$.

Solve

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

via

$$x_{k+1} := \mathsf{Prox}_{t_k}(x_k - t_k \cdot \nabla f(x_k))$$

where $\mathsf{Prox}_t$ is the following *proximal map*

$$\mathsf{Prox}_t(x) := \operatorname*{argmin}_{y \in \mathbb{R}^n} \left\{ \frac{1}{2t}\|y - x\|^2 + \psi(y) \right\}.$$

Observe:

$$\mathsf{Prox}_t(x - t \cdot \nabla f(x)) = \operatorname*{argmin}_{y \in \mathbb{R}^n} \left\{ f(x) + \langle \nabla f(x), y - x \rangle + \frac{1}{2t}\|y - x\|^2 + \psi(y) \right\}.$$

# Fast proximal gradient algorithm

(Beck-Teboulle 2009, Nesterov 2013)

Again generate two sequences and incorporate momentum.

Solve

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

via

$$x_{k+1} = \mathsf{Prox}_{t_k}(y_k - t_k \cdot \nabla f(y_k))$$

and

$$y_{k+1} = x_{k+1} + \frac{\theta_{k+1}(1-\theta_k)}{\theta_k} \cdot (x_{k+1} - x_k)$$

where $\theta_0 = 1$ and $\theta_{k+1}^2 = \theta_k^2(1-\theta_{k+1}),\ k = 0, 1, \ldots.$

Same convergence properties as fast gradient.

# Example: $\ell_1$ regularization

Consider the problem

$$\min_x f(x) + \lambda\|x\|_1.$$

This type of problem is the crux of lasso and compressive sensing.

For $\psi(x) = \lambda\|x\|_1$ the proximal map is (componentwise)

$$\mathsf{Prox}_t(g)_i = \begin{cases} g_i - \lambda t & \text{if} \quad g_i > \lambda t \\ 0 & \text{if} \quad |g_i| \leq \lambda t \\ g_i + \lambda t & \text{if} \quad g_i < -\lambda t \end{cases}$$

See Python code: `Algorithms.py`, `Functions.py`, `Example2.py`

*OPTIONAL: strong convexity*

# Strong convexity

Let $f : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ be convex and differentiable.

## Definition

$f$ is strongly convex with modulus $\mu > 0$ if

$$f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle + \frac{\mu}{2} \|y - x\|^2$$

for all $x, y$.

## Recall

$f$ is $L$-smooth if

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

for all $x, y$.

# Linear convergence of gradient descent

### Theorem

*Let $f : \mathbb{R}^n \to \mathbb{R}$ be $\mu$-strongly convex and $L$-smooth. Then $\bar{f} := \min_{x \in \mathbb{R}^n} f(x) < \infty$ and $f$ has a unique minimizer $\bar{x}$.*

*If the step-sizes satisfy $t_k \geq \frac{1}{L} > 0$ then the iterates generated by the gradient descent algorithm satisfy*

$$f(x_k) - \bar{f} \leq \left(1 - \frac{\mu}{L}\right)^k (f(x_0) - \bar{f}),$$

*and*

$$\|x_k - \bar{x}\|^2 \leq \frac{L}{\mu} \left(1 - \frac{\mu}{L}\right)^k \|x_0 - \bar{x}\|^2.$$

*In particular, the algorithm finds $x \in \mathbb{R}^n$ with $f(x) - \bar{f} \leq \epsilon(f(x_0) - \bar{f})$ in $\mathcal{O}\left(\frac{L}{\mu} \cdot \log\left(\frac{1}{\epsilon}\right)\right)$ iterations.*

# Lower bound for *any* gradient method

Suppose $f : \mathbb{R}^n \to \mathbb{R}$ is convex and differentiable and consider an algorithm such that

$$x_{k+1} \in x_0 + \mathsf{span}\{\nabla f(x_0), \ldots, \nabla f(x_k)\}.$$

How good could that kind of algorithm be?

**Theorem**

*For all $x_0 \in \mathbb{R}^n$ there exists a $\mu$-strongly convex, $L$-smooth, and differentiable function $f : \mathbb{R}^n \to \mathbb{R}$ such that*

$$f(x_k) - \bar{f} \geq \frac{\mu}{2} \left( \frac{\sqrt{L/\mu} - 1}{\sqrt{L/\mu} + 1} \right)^{2k} \|x_0 - \bar{x}\|^2 \quad \textit{for} \;\; k \leq n/2.$$

# Fast (linear) gradient descent

Suppose $f$ is $\mu$-strongly convex and $L$-smooth. Generate two sequences that start at the same initial point $y_0 = x_0$ and are updated via

$$x_{k+1} = x_k - \frac{1}{L} \cdot \nabla f(y_k)$$

$$y_{k+1} = x_{k+1} + \frac{1 - \sqrt{\mu/L}}{1 + \sqrt{\mu/L}} \cdot (x_{k+1} - x_k)$$

## Theorem

*The iterates generated by the above algorithm satisfy*

$$f(x_k) - \bar{f} \leq L \left( \frac{\sqrt{L/\mu} - 1}{\sqrt{L/\mu} + 1} \right)^k \|x_0 - \bar{x}\|^2.$$

*OPTIONAL: conditional gradient method*

# Conditional gradient (Frank-Wolfe) method

Let $\mathcal{X} \subseteq \mathbb{R}^n$ be a closed convex set, $f : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ be a differentiable convex function and suppose the following *linear oracle* is available:

$$g \mapsto \operatorname*{argmin}_{y \in \mathcal{X}} \langle g, y \rangle.$$

Solve

$$\min_{x \in \mathcal{X}} f(x)$$

via

$$\begin{aligned} s_k &:= \operatorname*{argmin}_{y \in \mathcal{X}} \langle \nabla f(x_k), y \rangle \\ x_{k+1} &:= x_k + \theta_k (s_k - x_k) \text{ for } \theta_k \in [0, 1] \end{aligned}$$

## Observe

Conditional gradient relies on linear oracle (no projection) for $\mathcal{X}$.

# Conditional gradient (Frank-Wolfe) method

Curvature constant (Jaggi 2013)

$$C_f := \sup_{\substack{x,s\in\mathcal{X}\\ \theta\in(0,1]}} \frac{2}{\theta^2}\left(f(x+\theta(s-x)) - f(x) - \langle \nabla f(x), \theta(s-x)\rangle\right).$$

Theorem

*Suppose* $C_f < \infty$ *and* $\theta_k := \frac{2}{k+2},\ k = 0, 1, \ldots$. *Then the conditional gradient iterates satisfy*

$$f(x_k) - \bar{f} \leq \frac{2C_f}{k+2}.$$

*OPTIONAL: subgradient methods*

# Subgradient method

### Fact

Let $f : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ be a convex function. Then for all $x \in \mathrm{ri}(\mathrm{dom}(f))$ there exists $g \in \mathbb{R}^n$ such that

$$f(y) \geq f(x) + \langle g, y - x \rangle \;\; \forall y \in \mathbb{R}^n.$$

### Definition

Let $f : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$. The subdifferential of $f$ at $x \in \mathrm{dom}(f)$ is

$$\partial f(x) := \{g \in \mathbb{R}^n : f(y) \geq f(x) + \langle g, y - x \rangle \;\; \forall y \in \mathbb{R}^n\}.$$

### Fact

Let $f : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ be a convex function. Then $f$ is differentiable at $x \in \mathrm{int}(\mathrm{dom}(f))$ if and only if

$$\partial f(x) = \{\nabla f(x)\}.$$

# Subgradient method

Let $f : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ be a convex function and $\mathcal{X} \subseteq \mathrm{dom}(f)$ be a closed convex set. Solve

$$\min_{x \in \mathcal{X}} \; f(x)$$

via

$$x_{k+1} := \Pi_{\mathcal{X}}(x_k - t_k \cdot g_k), \text{ for } \; g_k \in \partial f(x_k).$$

## Observe

Subgradient method subsumes projected gradient descent when $f$ is convex and differentiable.

# Proximal subgradient algorithm

Let $f : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ and $\psi : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ be convex functions such that $\mathrm{dom}(\psi) \subseteq \mathrm{ri}(\mathrm{dom}(f))$. Let $\phi := f + \psi$.

Solve

$$\min_{x \in \mathbb{R}^n} \{f(x) + \psi(x)\} \;\Leftrightarrow\; \min_{x \in \mathbb{R}^n} \phi(x)$$

via

$$x_{k+1} := \mathrm{Prox}_{t_k}(x_k - t_k \cdot g_k), \text{ for } g_k \in \partial f(x_k)$$

## Theorem

*If $\phi$ is $G$-Lipschitz then the proximal subgradient iterates satisfy*

$$\min_{i=0,1,\ldots,k} \phi(x_i) - \bar{\phi} \leq \frac{\mathrm{dist}(x_0, \bar{X})^2 + G^2 \sum_{i=0}^{k} t_i^2}{2 \sum_{i=0}^{k} t_i}$$