# CS257 Linear and Convex Optimization

## Lecture 8

Bo Jiang

John Hopcroft Center for Computer Science
Shanghai Jiao Tong University

October 26, 2020

# Recap: QP, QCQP

Quadratic program (QP)

$$\begin{aligned} \min_{\boldsymbol{x}} \quad & \frac{1}{2}\boldsymbol{x}^T \boldsymbol{Q} \boldsymbol{x} + \boldsymbol{c}^T \boldsymbol{x} \\ \text{s.t.} \quad & \boldsymbol{B}\boldsymbol{x} \leq \boldsymbol{d} \\ & \boldsymbol{A}\boldsymbol{x} = \boldsymbol{b} \end{aligned}$$

Quadratically constrained quadratic program (QCQP)

$$\begin{aligned} \min_{\boldsymbol{x}} \quad & \frac{1}{2}\boldsymbol{x}^T \boldsymbol{Q} \boldsymbol{x} + \boldsymbol{c}^T \boldsymbol{x} \\ \text{s.t.} \quad & \frac{1}{2}\boldsymbol{x}^T \boldsymbol{Q}_i \boldsymbol{x} + \boldsymbol{c}_i^T \boldsymbol{x} + \boldsymbol{d}_i \leq 0, \quad i = 1, 2, \ldots, m \\ & \boldsymbol{A}\boldsymbol{x} = \boldsymbol{b} \end{aligned}$$

# Recap: LS, Lasso, and Ridge Regressions

Given $\boldsymbol{y} \in \mathbb{R}^n$, $\boldsymbol{X} \in \mathbb{R}^{n \times p}$.

## Linear least squares regression

$$\min_{\boldsymbol{w}} \ \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w}\|_2^2$$

## Lasso

$$\begin{aligned} \min_{\boldsymbol{w}} \quad & \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w}\|_2^2 \\ \text{s. t.} \quad & \|\boldsymbol{w}\|_1 \leq t \end{aligned}$$

## Ridge regression

$$\begin{aligned} \min_{\boldsymbol{w}} \quad & \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w}\|_2^2 \\ \text{s. t.} \quad & \|\boldsymbol{w}\|_2^2 \leq t \end{aligned}$$

# Recap: GP

## Posynomial form

$$
\begin{aligned}
\min_{\boldsymbol{x}} \quad & \sum_{k=1}^{p_0} \gamma_{0k} x_1^{a_{0k1}} x_2^{a_{0k2}} \cdots x_n^{a_{0kn}} \\
\text{s. t.} \quad & \sum_{k=1}^{p_i} \gamma_{ik} x_1^{a_{ik1}} x_2^{a_{ik2}} \cdots x_n^{a_{ikn}} \leq 1, \quad i = 1, \ldots, m \\
& \eta_j x_1^{c_{j1}} x_2^{c_{j2}} \cdots x_n^{c_{jn}} = 1, \quad j = 1, \ldots, r
\end{aligned}
$$

## Convex form

$$
\begin{aligned}
\min_{\boldsymbol{y}} \quad & \log\left(\sum_{k=1}^{p_0} e^{\boldsymbol{a}_{0k}^T \boldsymbol{y} + b_{0k}}\right) \\
\text{s. t.} \quad & \log\left(\sum_{k=1}^{p_i} e^{\boldsymbol{a}_{ik}^T \boldsymbol{y} + b_{ik}}\right) \leq 0, \quad i = 1, \ldots, m \\
& \boldsymbol{c}_j^T \boldsymbol{y} + d_j = 0, \quad j = 1, \ldots, r
\end{aligned}
$$

# Contents

# Unconstrained Optimization Problems

Consider an unconstrained, smooth convex optimization problem

$$\min_{\boldsymbol{x}} \; f(\boldsymbol{x})$$

where $f$ is convex and differentiable on $\mathbb{R}^n$.

The optimal solution satisfies the first-order optimality condition

$$\nabla f(\boldsymbol{x}^*) = \boldsymbol{0}$$

In some rare cases, this yields closed-form solutions, e.g.

$$\min_{\boldsymbol{w}} \; \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w}\|_2^2$$

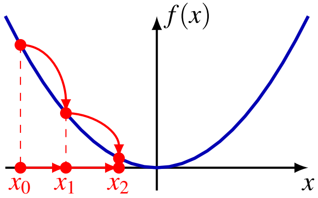
has closed-form solution

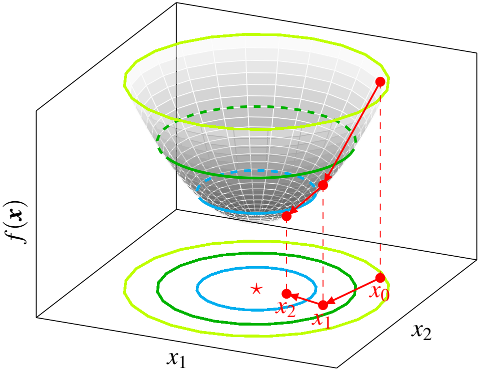
$$\boldsymbol{w}^* = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}$$

But in most cases we need numerical algorithms.

# Descent Method

choose initial point $\boldsymbol{x}_0 \in \mathbb{R}^n$
**repeat**
  choose descent direction $\boldsymbol{d}_k \in \mathbb{R}^n$ and step size $t_k > 0$
  $\boldsymbol{x}_{k+1} = \boldsymbol{x}_k + t_k \boldsymbol{d}_k \quad$ s.t. $\quad f(\boldsymbol{x}_{k+1}) < f(\boldsymbol{x}_k)$
**until** stopping criterion is satisfied




## Questions

- How to choose $\boldsymbol{d}_k$ and $t_k$?
- Does $\boldsymbol{x}_k$ converge to $\boldsymbol{x}^*$?

# Descent Direction

$\boldsymbol{d}_k$ is a descent direction at $\boldsymbol{x}_k$ if for all small enough $t > 0$

$$g(t) \triangleq f(\boldsymbol{x}_k + t\boldsymbol{d}_k) < f(\boldsymbol{x}_k) = g(0)$$

For differentiable $f$ (not necessarily convex),

- if $\boldsymbol{d}_k$ is a descent direction, then $g'(0) = \boldsymbol{d}_k^T \nabla f(\boldsymbol{x}_k) \leq 0$;
- if $g'(0) = \boldsymbol{d}_k^T \nabla f(\boldsymbol{x}_k) < 0$, then $\boldsymbol{d}_k$ is a descent direction.

For convex $f$, by the first-order condition for convexity,

$$f(\boldsymbol{x}_k) > f(\boldsymbol{x}_k + t\boldsymbol{d}_k) \geq f(\boldsymbol{x}_k) + t\boldsymbol{d}_k^T \nabla f(\boldsymbol{x}_k).$$

$\boldsymbol{d}_k^T \nabla f(\boldsymbol{x}_k) < 0$ is also necessary for $\boldsymbol{d}_k$ to be a descent direction.

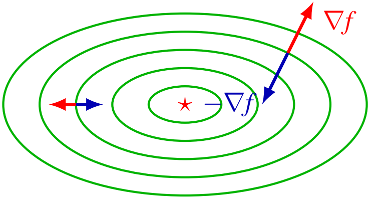
For convex differentiable $f$,

$$\boldsymbol{d}_k \text{ is a descent direction } \iff \boldsymbol{d}_k^T \nabla f(\boldsymbol{x}_k) < 0$$

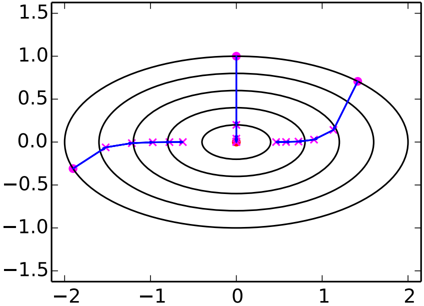
# Gradient Descent

Choose $\boldsymbol{d}_k = -\nabla f(\boldsymbol{x}_k)$, $\boldsymbol{d}_k^T \nabla f(\boldsymbol{x}_k) = -\|\nabla f(\boldsymbol{x}_k)\|_2^2 < 0$ unless $\nabla f(\boldsymbol{x}_k) = 0$.

$$\boxed{\boldsymbol{x}_{k+1} = \boldsymbol{x}_k - t_k \nabla f(\boldsymbol{x}_k)}$$




level curves of $f(x_1, x_2) = \frac{x_1^2}{4} + x_2^2$




Question. What happens if $\nabla f(\boldsymbol{x}_k) = \boldsymbol{0}$?

# Max-rate Descending Direction

$-\nabla f(\boldsymbol{x}_k)$ is the max-rate descending direction

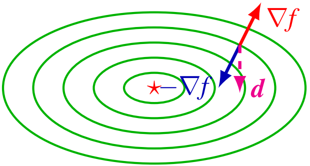
- If $\|\boldsymbol{d}_k\|_2 = 1$, the rate of change of $f$ at $\boldsymbol{x}_k$ along the direction $\boldsymbol{d}_k$ is

$$\nabla_{\boldsymbol{d}_k} f(\boldsymbol{x}_k) = g'(0) = \lim_{t \downarrow 0} \frac{f(\boldsymbol{x}_k + t\boldsymbol{d}_k) - f(\boldsymbol{x}_k)}{t} = \boldsymbol{d}_k^T \nabla f(\boldsymbol{x}_k)$$

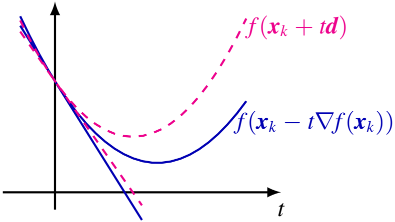
- by the Cauchy-Schwarz inequality,

$$\boldsymbol{d}_k^T \nabla f(\boldsymbol{x}_k) \geq -\|\boldsymbol{d}_k\|_2 \cdot \|\nabla f(\boldsymbol{x}_k)\|_2 = -\|\nabla f(\boldsymbol{x}_k)\|_2$$

  with equality iff $\boldsymbol{d}_k = -\nabla f(\boldsymbol{x}_k)/\|\nabla f(\boldsymbol{x}_k)\|$




level curves of $f(x_1, x_2) = \frac{x_1^2}{4} + x_2^2$



# Gradient Descent Algorithm

```
initialization x ← x_0 ∈ R^n
while ‖∇f(x)‖ > δ do
    x ← x − t∇f(x)
end while
return x
```

Step size (aka learning rate in machine learning)

- the above algorithm uses constant step size $t$ for all iterations
- there are other methods for choosing $t$ for each iteration, e.g. exact line search, backtracking line search

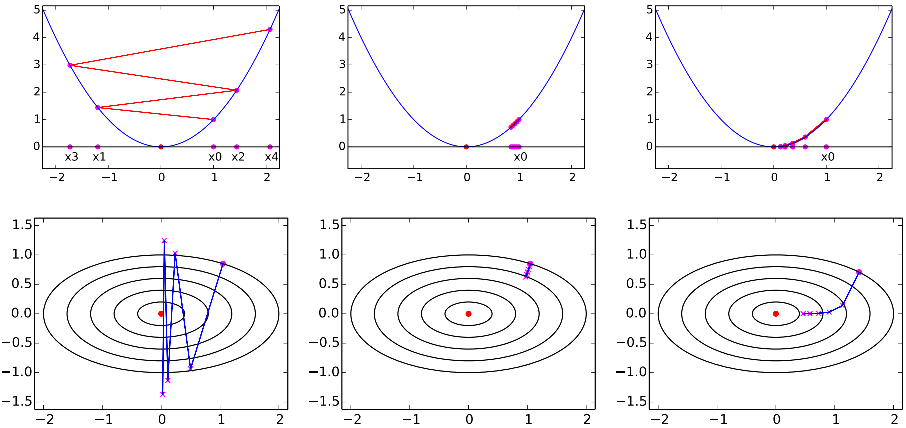
Stopping criterion

- ideally, stop if $\nabla f(\boldsymbol{x}) = 0$ (optimality condition), but impractical
- more practical: stop when $\|\nabla f(\boldsymbol{x})\| \leq \delta$ for some small $\delta$
- other criteria: $|f(\boldsymbol{x}_{\text{new}}) - f(\boldsymbol{x}_{\text{old}})| \leq \delta$, $\frac{|f(\boldsymbol{x}_{\text{new}}) - f(\boldsymbol{x}_{\text{old}})|}{|f(\boldsymbol{x}_{\text{old}})|} \leq \delta$, ...
- in practice, also stop if maximum # of iterations is reached

# Large vs. Small Step Size

Consider constant step size. How large should the step size be?

- Too large: may oscillate and diverge
- Too small: may be too slow
- “Just right”: fast convergence



# 1D Example

Consider $f(x) = \frac{1}{2}ax^2$, where $a > 0$.

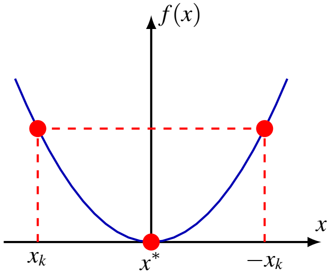
- gradient at $x_k$ is $f'(x_k) = ax_k$, so $d_k = -ax_k$ in gradient descent
- update

$$x_{k+1} = x_k - tf'(x_k) = (1 - at)x_k$$

- in order for $f(x_{k+1}) < f(x_k)$, the step size should satisfy

$$|x_{k+1}| < |x_k| \implies |1 - at| < 1 \implies 0 < t < \frac{2}{a}$$

- $x_k \to x^* = 0$ geometrically for such $t$




Note $f$ satisfies

- $|f'(x) - f'(y)| = a|x - y|$
- $f''(x) = a$

$f'$ is Lipschitz continuous

# Lipschitz Continuity

A function $f : \mathbb{R}^n \to \mathbb{R}^m$ is Lipschitz continuous with Lipschitz constant $L > 0$, or simply $L$-Lipschitz, if

$$\|f(\boldsymbol{x}) - f(\boldsymbol{y})\| \leq L\|\boldsymbol{x} - \boldsymbol{y}\|, \quad \forall \boldsymbol{x}, \boldsymbol{y}$$

Note. Lipschitz continuity can be defined with respect to any norms. But we will assume the norms in the above definition are the 2-norms in $\mathbb{R}^n$ and $\mathbb{R}^m$, respectively, unless stated otherwise.

Note. Lipschitz continuity implies uniform continuity.

Example. $f(x) = ax$ is $|a|$-Lipschitz, $|f(x) - f(y)| = |a| \cdot |x - y|$

Example. $f(x) = |x|$ is $1$-Lipschitz, $|f(x) - f(y)| = \big||x| - |y|\big| \leq |x - y|$

Example. $f(\boldsymbol{x}) = \boldsymbol{a}^T\boldsymbol{x}$ is $\|\boldsymbol{a}\|$-Lipschitz, $|\boldsymbol{a}^T\boldsymbol{x} - \boldsymbol{a}^T\boldsymbol{y}| \leq \|\boldsymbol{a}\| \cdot \|\boldsymbol{x} - \boldsymbol{y}\|$ by the Cauchy-Schwarz inequality.

# Lipschitz Continuity (cont'd)

Example. Let $\boldsymbol{Q} = \begin{pmatrix} 1 & 0 \\ 0 & 2 \end{pmatrix}$. $f(\boldsymbol{x}) = \boldsymbol{Q}\boldsymbol{x} = (x_1, 2x_2)^T$ is 2-Lipschitz.

$$f(\boldsymbol{x}) - f(\boldsymbol{y}) = (x_1 - y_1, 2x_2 - 2y_2)^T = (d_1, 2d_2)^T$$

$$\|f(\boldsymbol{x}) - f(\boldsymbol{y})\| = \sqrt{d_1^2 + 4d_2^2} \le 2\sqrt{d_1^2 + d_2^2} = 2\|\boldsymbol{x} - \boldsymbol{y}\|$$

More generally, $f(\boldsymbol{x}) = \boldsymbol{Q}\boldsymbol{x}$ with $\boldsymbol{Q} \succeq \boldsymbol{O}$ is $\lambda_{\max}(\boldsymbol{Q})$-Lipschitz, where $\lambda_{\max}(\boldsymbol{Q})$ is the largest eigenvalue of $\boldsymbol{Q}$[1].

Proof. Let $\boldsymbol{d} = \boldsymbol{x} - \boldsymbol{y}$.

$$\|f(\boldsymbol{x}) - f(\boldsymbol{y})\| = \|\boldsymbol{Q}\boldsymbol{d}\| = \sqrt{\boldsymbol{d}^T\boldsymbol{Q}^2\boldsymbol{d}} \le \sqrt{\lambda_{\max}(\boldsymbol{Q}^2)\|\boldsymbol{d}\|^2} = \lambda_{\max}(\boldsymbol{Q})\|\boldsymbol{x} - \boldsymbol{y}\|$$

The last equality uses the fact $\lambda_{\max}(\boldsymbol{Q}^2) = \lambda_{\max}^2(\boldsymbol{Q})$.

---

[1] if we do not assume $\boldsymbol{Q} \succeq \boldsymbol{O}$, then we should replace $\lambda_{\max}(\boldsymbol{Q})$ by $\sigma_{\max}(\boldsymbol{Q})$, the largest singular value of $\boldsymbol{Q}$

# Appendix: Bounds on Quadratic Forms

**Proposition.** For a symmetric matrix $\boldsymbol{Q} \in \mathbb{R}^{n\times n}$,

$$\lambda_{\min}\|\boldsymbol{x}\|_2^2 \le \boldsymbol{x}^T\boldsymbol{Q}\boldsymbol{x} \le \lambda_{\max}\|\boldsymbol{x}\|_2^2, \quad \forall \boldsymbol{x} \in \mathbb{R}^n$$

where $\lambda_{\max}$ and $\lambda_{\min}$ are the largest and the smallest eigenvalues of $\boldsymbol{Q}$, respectively.

**Proof.** Recall that $\boldsymbol{Q}$ can be orthogonally diagonalized, i.e. $\boldsymbol{Q} = \boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^T$, where $\boldsymbol{\Lambda} = \text{diag}\{\lambda_1, \ldots, \lambda_n\}$ and $\boldsymbol{U}^T\boldsymbol{U} = \boldsymbol{I}$ (Lecture 2, slide 24). Let $\boldsymbol{x} = \boldsymbol{U}\boldsymbol{y}$.

$$\boldsymbol{x}^T\boldsymbol{Q}\boldsymbol{x} = \boldsymbol{y}^T(\boldsymbol{U}^T\boldsymbol{Q}\boldsymbol{U})\boldsymbol{y} = \boldsymbol{y}^T\boldsymbol{\Lambda}\boldsymbol{y} = \sum_{i=1}^{n}\lambda_i y_i^2 \le \sum_{i=1}^{n}\lambda_{\max} y_i^2 = \lambda_{\max}\|\boldsymbol{y}\|_2^2$$

Then use the fact that orthogonal transformations preserves 2-norms, i.e.

$$\|\boldsymbol{x}\|_2^2 = \boldsymbol{x}^T\boldsymbol{x} = (\boldsymbol{U}\boldsymbol{y})^T(\boldsymbol{U}\boldsymbol{y}) = \boldsymbol{y}^T(\boldsymbol{U}^T\boldsymbol{U})\boldsymbol{y} = \boldsymbol{y}^T\boldsymbol{y} = \|\boldsymbol{y}\|_2^2.$$

Similarly for $\boldsymbol{x}^T\boldsymbol{Q}\boldsymbol{x} \ge \lambda_{\min}\|\boldsymbol{x}\|_2^2$.

# Lipschitz Continuity of Gradient

A function is $L$-smooth if it is differentiable and its gradient is $L$-Lipschitz, i.e.

$$\|\nabla f(\boldsymbol{x}) - \nabla f(\boldsymbol{y})\| \le L\|\boldsymbol{x} - \boldsymbol{y}\|, \quad \forall \boldsymbol{x}, \boldsymbol{y}$$

Note. $L$ upper bounds the rate of change of $\nabla f$

Example. $f(x) = \frac{1}{2}ax^2$ is $|a|$-smooth, since $f'(x) = ax$ is $|a|$-Lipschitz

Example. $f(\boldsymbol{x}) = \frac{1}{2}\boldsymbol{x}^T\boldsymbol{Q}\boldsymbol{x}$ with $\boldsymbol{Q} \succeq \boldsymbol{O}$ is $\lambda_{\max}(\boldsymbol{Q})$-smooth, since $\nabla f(\boldsymbol{x}) = \boldsymbol{Q}\boldsymbol{x}$ is $\lambda_{\max}(\boldsymbol{Q})$-Lipschitz.

With $\boldsymbol{Q} = \begin{pmatrix} 1 & 0 \\ 0 & 2 \end{pmatrix}$, we obtain $f(\boldsymbol{x}) = \frac{1}{2}x_1^2 + x_2^2$ is 2-smooth.

Lemma. A twice continuously differentiable convex $f : \mathbb{R}^n \to \mathbb{R}$ is $L$-smooth iff $\nabla^2 f(\boldsymbol{x}) \preceq L\boldsymbol{I}$, meaning $L\boldsymbol{I} - \nabla^2 f(\boldsymbol{x}) \succeq \boldsymbol{O}$, or equivalently $\lambda_{\max}(\nabla^2 f(\boldsymbol{x})) \le L$.

# Lipschitz Continuity of Gradient (cont'd)

Lemma. A twice continuously differentiable $f : \mathbb{R}^n \to \mathbb{R}$ is $L$-smooth iff for any $\boldsymbol{x}$, $-L\boldsymbol{I} \preceq \nabla^2 f(\boldsymbol{x}) \preceq L\boldsymbol{I}$, or equivalently $|\lambda| \leq L$ for all eigenvalues $\lambda$ of $\nabla^2 f(\boldsymbol{x})$.

Proof. "$\Leftarrow$". Assume $-L\boldsymbol{I} \preceq \nabla^2 f(\boldsymbol{x}) \preceq L\boldsymbol{I}$ for all $\boldsymbol{x}$.

By the Mean Value Theorem, there exists $\boldsymbol{z} = \boldsymbol{y} + t(\boldsymbol{x} - \boldsymbol{y})$ for some $t \in [0, 1]$ s.t.

$$\nabla f(\boldsymbol{x}) - \nabla f(\boldsymbol{y}) = \nabla^2 f(\boldsymbol{z})(\boldsymbol{x} - \boldsymbol{y})$$

Since $-L\boldsymbol{I} \preceq \nabla^2 f(\boldsymbol{x}) \preceq L\boldsymbol{I}$,

$$\|\nabla f(\boldsymbol{x}) - \nabla f(\boldsymbol{y})\| = \|\nabla^2 f(\boldsymbol{z})(\boldsymbol{x} - \boldsymbol{y})\| \leq L\|\boldsymbol{x} - \boldsymbol{y}\|$$

by the proof on slide 15.

# Lipschitz Continuity of Gradient (cont'd)

Proof (cont'd). "$\Rightarrow$". Assume $f$ is $L$-smooth. Fix a direction $\boldsymbol{d}$. Let $h(t) = \nabla f(\boldsymbol{x} + t\boldsymbol{d})$. Since $f$ is $L$-smooth,

$$\|g(t) - g(0)\| = \|\nabla f(\boldsymbol{x} + t\boldsymbol{d}) - \nabla f(\boldsymbol{x})\| \leq L\|t\boldsymbol{d}\|$$

so

$$\left\|\frac{g(t) - g(0)}{t}\right\| \leq L\|\boldsymbol{d}\|$$

Letting $t \to 0$ and using the chain rule

$$\|\nabla^2 f(\boldsymbol{x})\boldsymbol{d}\| = \|g'(0)\| \leq L\|\boldsymbol{d}\|$$

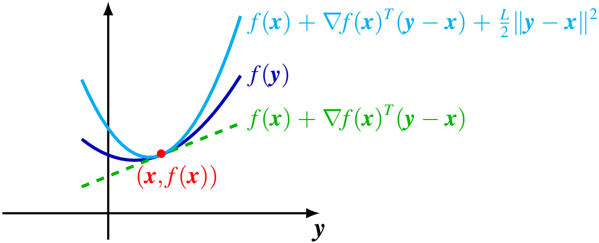
Let $\boldsymbol{d}$ be an eigenvector of $\nabla^2 f(\boldsymbol{x})$ with associated eigenvalue $\lambda$,

$$|\lambda| \cdot \|\boldsymbol{d}\| = \|\lambda \boldsymbol{d}\| \leq L\|\boldsymbol{d}\| \implies |\lambda| \leq L$$

# Quadratic Upper Bound

Lemma. If $f$ is $L$-smooth, then

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




Note. The upper bound does not assume the convexity of $f$.

If $\nabla^2 f(\boldsymbol{x}) \preceq L\boldsymbol{I}$, this is intuitive from the second-order Taylor expansion

$$f(\boldsymbol{y}) = f(\boldsymbol{x}) + \nabla f(\boldsymbol{x})^T(\boldsymbol{y}-\boldsymbol{x}) + \frac{1}{2}(\boldsymbol{y}-\boldsymbol{x})^T\nabla^2 f(\boldsymbol{z})(\boldsymbol{y}-\boldsymbol{x})$$

for some $\boldsymbol{z}$ on the line segment between $\boldsymbol{x}$ and $\boldsymbol{y}$. (Check $f(\boldsymbol{x}) = \frac{1}{2}\boldsymbol{x}^T\boldsymbol{Q}\boldsymbol{x}$)

# Proof

Let $z(t) = \boldsymbol{x} + t(\boldsymbol{y} - \boldsymbol{x})$ and $g(t) = f(z(t))$. Then $g(0) = f(\boldsymbol{x})$, $g(1) = f(\boldsymbol{y})$, $g'(t) = \nabla f(z(t))^T(\boldsymbol{y} - \boldsymbol{x})$, $g'(0) = \nabla f(\boldsymbol{x})^T(\boldsymbol{y} - \boldsymbol{x})$.

1. By the Newton-Leibniz formula,

$$g(1) - g(0) - g'(0) = \int_0^1 [g'(t) - g'(0)]dt \leq \int_0^1 |g'(t) - g'(0)|dt$$

2. By the Cauchy-Schwarz inequality and $L$-smoothness

$$\begin{aligned} |g'(t) - g'(0)| &= \left|[\nabla f(z(t)) - \nabla f(\boldsymbol{x})]^T(\boldsymbol{y} - \boldsymbol{x})\right| \\ &\leq \|\nabla f(z(t)) - \nabla f(\boldsymbol{x})\| \cdot \|\boldsymbol{y} - \boldsymbol{x}\| \\ &\leq L\|z(t) - \boldsymbol{x}\| \cdot \|\boldsymbol{y} - \boldsymbol{x}\| \\ &= tL\|\boldsymbol{x} - \boldsymbol{y}\|^2 \end{aligned}$$

3. Plugging $|g'(t) - g'(0)| \leq tL\|\boldsymbol{x} - \boldsymbol{y}\|^2$ into step 1,

$$f(\boldsymbol{y}) - f(\boldsymbol{x}) - \nabla f(\boldsymbol{x})^T(\boldsymbol{y} - \boldsymbol{x}) = g(1) - g(0) - g'(0) \leq \frac{L}{2}\|\boldsymbol{x} - \boldsymbol{y}\|^2$$

# Consequence of Quadratic Upper Bound

For $L$-smooth $f$, the sequence $\{\boldsymbol{x}_k\}$ produced by gradient descent satisfies

$$f(\boldsymbol{x}_{k+1}) \leq f(\boldsymbol{x}_k) - t\left(1 - \frac{Lt}{2}\right)\|\nabla f(\boldsymbol{x}_k)\|^2$$

Proof. Plugging in $\boldsymbol{x} = \boldsymbol{x}_k$ and $\boldsymbol{y} = \boldsymbol{x}_{k+1} = \boldsymbol{x}_k - t\nabla f(\boldsymbol{x}_k)$ in the quadratic upper bound,

$$\begin{aligned} f(\boldsymbol{x}_{k+1}) &\leq f(\boldsymbol{x}_k) - t\|\nabla f(\boldsymbol{x}_k)\|^2 + \frac{L}{2}t^2\|\nabla f(\boldsymbol{x}_k)\|^2 \\ &= f(\boldsymbol{x}_k) - t\left(1 - \frac{Lt}{2}\right)\|\nabla f(\boldsymbol{x}_k)\|^2 \end{aligned}$$

Note. If $\nabla f(\boldsymbol{x}_k) \neq 0$ and $0 < t < \frac{2}{L}$, then $f(\boldsymbol{x}_{k+1}) < f(\boldsymbol{x}_k)$, so gradient descent with step size $t \in (0, 2/L)$ is indeed a descent method.

Note. We can lower bound the decrease in function value in each step. In particular, for $0 < t \leq \frac{1}{L}$,

$$f(\boldsymbol{x}_k) - f(\boldsymbol{x}_{k+1}) \geq \frac{t}{2}\|\nabla f(\boldsymbol{x}_k)\|^2$$

# Convergence Analysis

Theorem. If $f$ is convex and $L$-smooth, and $\boldsymbol{x}^*$ is a minimum of $f$, then for step size $t \in (0, \frac{1}{L}]$, the sequence $\{\boldsymbol{x}_k\}$ produced by the gradient descent algorithm satisfies

$$f(\boldsymbol{x}_k) - f(\boldsymbol{x}^*) \leq \frac{\|\boldsymbol{x}_0 - \boldsymbol{x}^*\|^2}{2tk}$$

Notes.

- $f(\boldsymbol{x}_k) \downarrow f^*$ as $k \to \infty$.
- Any limiting point of $\boldsymbol{x}_k$ is an optimal solution.
- The rate of convergence is $O(1/k)$, i.e. # of iterations to guarantee $f(\boldsymbol{x}_k) - f(\boldsymbol{x}^*) \leq \epsilon$ is $O(1/\epsilon)$. For $\epsilon = 10^{-p}$, $k = O(10^p)$, exponential in the number of significant digits!
- Faster convergence with larger $t$; best $t = \frac{1}{L}$, but $L$ is unknown.
- Good initial guess helps.

# Proof

1. By the basic gradient step $\boldsymbol{x}_{k+1} = \boldsymbol{x}_k - t\nabla f(\boldsymbol{x}_k)$,

$$\begin{aligned}\|\boldsymbol{x}_{k+1} - \boldsymbol{x}^*\|^2 &= \|\boldsymbol{x}_k - t\nabla f(\boldsymbol{x}_k) - \boldsymbol{x}^*\|^2 \\ &= \|\boldsymbol{x}_k - \boldsymbol{x}^*\|^2 + t^2\|\nabla f(\boldsymbol{x}_k)\|^2 + 2t\nabla f(\boldsymbol{x}_k)^T(\boldsymbol{x}^* - \boldsymbol{x}_k)\end{aligned}$$

2. By the first-order condition for convexity,

$$\nabla f(\boldsymbol{x}_k)^T(\boldsymbol{x}^* - \boldsymbol{x}_k) \le f(\boldsymbol{x}^*) - f(\boldsymbol{x}_k)$$

3. Plugging 2 into 1,

$$\|\boldsymbol{x}_{k+1} - \boldsymbol{x}^*\|^2 \le \|\boldsymbol{x}_k - \boldsymbol{x}^*\|^2 + t^2\|\nabla f(\boldsymbol{x}_k)\|^2 + 2t[f(\boldsymbol{x}^*) - f(\boldsymbol{x}_k)]$$

4. Plugging in $\frac{t}{2}\|\nabla f(\boldsymbol{x}_k)\|^2 \le f(\boldsymbol{x}_k) - f(\boldsymbol{x}_{k+1})$ from slide 21,

$$\|\boldsymbol{x}_{k+1} - \boldsymbol{x}^*\|^2 \le \|\boldsymbol{x}_k - \boldsymbol{x}^*\|^2 + 2t[f(\boldsymbol{x}^*) - f(\boldsymbol{x}_{k+1})]$$

# Proof (cont'd)

5. Rearranging,

$$f(\boldsymbol{x}_{k+1}) - f(\boldsymbol{x}^*) \leq \frac{\|\boldsymbol{x}_k - \boldsymbol{x}^*\|^2 - \|\boldsymbol{x}_{k+1} - \boldsymbol{x}^*\|^2}{2t}$$

6. Summing over $k$ from $0$ to $N-1$,

$$\sum_{k=0}^{N-1} [f(\boldsymbol{x}_{k+1}) - f(\boldsymbol{x}^*)] \leq \frac{\|\boldsymbol{x}_0 - \boldsymbol{x}^*\|^2 - \|\boldsymbol{x}_N - \boldsymbol{x}^*\|^2}{2t} \leq \frac{\|\boldsymbol{x}_0 - \boldsymbol{x}^*\|^2}{2t}$$

7. Recalling the descent property $f(\boldsymbol{x}_{k+1}) \leq f(\boldsymbol{x}_k)$,

$$f(\boldsymbol{x}_N) - f(\boldsymbol{x}^*) \leq \frac{1}{N} \sum_{k=0}^{N-1} [f(\boldsymbol{x}_{k+1}) - f(\boldsymbol{x}^*)] \leq \frac{\|\boldsymbol{x}_0 - \boldsymbol{x}^*\|^2}{2tN}$$