

# Chapter 2

## Optimization Algorithms

*Convex Optimization for Machine Learning & Computer Vision*  
SS 2018

Tao Wu  
Yuesong Shen  
Zhenzhang Ye

Computer Vision Group  
Department of Informatics  
TU Munich

Optimization  
Algorithms

Tao Wu  
Yuesong Shen  
Zhenzhang Ye



Gradient Methods

Proximal Algorithms

Convergence Theory

Acceleration



# Gradient-based Methods

Gradient Methods

Proximal Algorithms

Convergence Theory

Acceleration

## Overview of this section

### Unconstrained, differentiable, possibly nonconvex optimization

Problem setting:

$$\text{minimize } J(u) \quad \text{over } u \in \mathbb{E}.$$

Assume:

- 1  $J : \mathbb{E} \rightarrow \mathbb{R}$  is continuously differentiable.
- 2 There exists a global minimizer  $u^*$ . (Typically, an optimization algorithm seeks for a local minimizer s.t.  $\nabla J(u^*) = 0$ .)



## Overview of this section

### Unconstrained, differentiable, possibly nonconvex optimization

Problem setting:

$$\text{minimize } J(u) \quad \text{over } u \in \mathbb{E}.$$

Assume:

- 1  $J : \mathbb{E} \rightarrow \mathbb{R}$  is continuously differentiable.
- 2 There exists a global minimizer  $u^*$ . (Typically, an optimization algorithm seeks for a local minimizer s.t.  $\nabla J(u^*) = 0$ .)

Methods under consideration:

- 1 (Scaled) gradient descent.
- 2 Line search method.
- 3 Majorize-minimize method.



## Overview of this section

### Unconstrained, differentiable, possibly nonconvex optimization

Problem setting:

$$\text{minimize } J(u) \quad \text{over } u \in \mathbb{E}.$$

Assume:

- 1  $J : \mathbb{E} \rightarrow \mathbb{R}$  is continuously differentiable.
- 2 There exists a global minimizer  $u^*$ . (Typically, an optimization algorithm seeks for a local minimizer s.t.  $\nabla J(u^*) = 0$ .)

Methods under consideration:

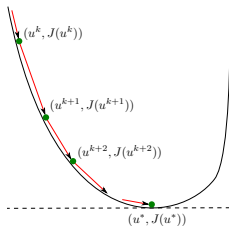
- 1 (Scaled) gradient descent.
- 2 Line search method.
- 3 Majorize-minimize method.

Analytical questions:

- 1 Convergence (or not); global vs. local convergence.
- 2 Convergence rate (in special cases).



## Descent method



## Descent method

Initialize  $u^0 \in \mathbb{E}$ . Iterate with  $k = 0, 1, 2, \dots$

- 1 If the stopping criteria  $\|\nabla J(u^k)\| \leq \epsilon$  is *not* satisfied, then continue; otherwise return  $u^k$  and stop.
- 2 Choose a **descent direction**  $d^k \in \mathbb{E}$  s.t.

$$\langle \nabla J(u^k), d^k \rangle < 0.$$

- 3 Choose an “appropriate” step size  $\tau^k > 0$ , and update

$$u^{k+1} = u^k + \tau^k d^k.$$



## Theorem

If  $\langle \nabla J(u^k), d^k \rangle < 0$ , then  $J(u^k + \tau d^k) < J(u^k)$  for all sufficiently small  $\tau > 0$ .



## Theorem

If  $\langle \nabla J(u^k), d^k \rangle < 0$ , then  $J(u^k + \tau d^k) < J(u^k)$  for all sufficiently small  $\tau > 0$ .

Proof: Use the Taylor expansion:

$$\begin{aligned} J(u^k + \tau d^k) &= J(u^k) + \tau \langle \nabla J(u^k), d^k \rangle + o(\tau) \\ &= J(u^k) + \tau \left( \langle \nabla J(u^k), d^k \rangle + o(1) \right) < J(u^k) \quad \text{as } \tau \rightarrow 0^+. \end{aligned}$$





### Theorem

If  $\langle \nabla J(u^k), d^k \rangle < 0$ , then  $J(u^k + \tau d^k) < J(u^k)$  for all sufficiently small  $\tau > 0$ .

Proof: Use the Taylor expansion:

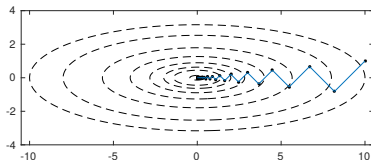
$$\begin{aligned} J(u^k + \tau d^k) &= J(u^k) + \tau \langle \nabla J(u^k), d^k \rangle + o(\tau) \\ &= J(u^k) + \tau \left( \langle \nabla J(u^k), d^k \rangle + o(1) \right) < J(u^k) \quad \text{as } \tau \rightarrow 0^+. \end{aligned}$$



### Choices of descent direction

- 1 Scaled gradient:  $d^k = -(H^k)^{-1} \nabla J(u^k)$ .
- 2 Gradient/Steepest descent:  $H^k = I$ .
- 3 Newton:  $H^k = \nabla^2 J(u^k)$ , assuming  $J$  is twice continuously differentiable and  $\nabla^2 J(u^k) \succ 0$ .
- 4 Quasi-Newton:  $H^k \approx \nabla^2 J(u^k)$ ,  $H^k$  is spd.

## Gradient descent with exact line search



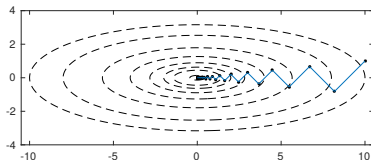
- Gradient descent with *exact* line search:

$$u^{k+1} = u^k - \tau^k \nabla J(u^k),$$

$$\tau^k = \arg \min_{\tau} J(u^k - \tau \nabla J(u^k)).$$



## Gradient descent with exact line search



- Gradient descent with *exact* line search:

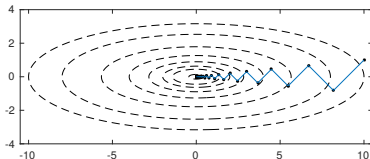
$$u^{k+1} = u^k - \tau^k \nabla J(u^k),$$

$$\tau^k = \arg \min_{\tau} J(u^k - \tau \nabla J(u^k)).$$

- Case study:  $J(u) = \frac{1}{2} \langle u, Qu \rangle - \langle b, u \rangle$ , matrix  $Q$  is spd.
  - $\nabla J(u) = Qu - b$ ,  $\|\cdot\|_Q^2 \equiv \langle \cdot, Q \cdot \rangle$ .



## Gradient descent with exact line search



- Gradient descent with *exact* line search:

$$\begin{aligned}u^{k+1} &= u^k - \tau^k \nabla J(u^k), \\ \tau^k &= \arg \min_{\tau} J(u^k - \tau \nabla J(u^k)).\end{aligned}$$

- Case study:  $J(u) = \frac{1}{2} \langle u, Qu \rangle - \langle b, u \rangle$ , matrix  $Q$  is spd.

- $\nabla J(u) = Qu - b, \quad \|\cdot\|_Q^2 \equiv \langle \cdot, Q \cdot \rangle.$

- $\tau^k = \arg \min_{\tau} J(u^k - \tau \nabla J(u^k)) = \frac{\|\nabla J(u^k)\|^2}{\|\nabla J(u^k)\|_Q^2} \Rightarrow$

$$\begin{aligned}\|u^{k+1} - u^*\|_Q^2 &= \left(1 - \frac{\|\nabla J(u^k)\|^4}{\|\nabla J(u^k)\|_Q^2 \|\nabla J(u^k)\|_{Q^{-1}}^2}\right) \|u^k - u^*\|_Q^2 \\ &\leq \left(\frac{\lambda_{\max}(Q) - \lambda_{\min}(Q)}{\lambda_{\max}(Q) + \lambda_{\min}(Q)}\right)^2 \|u^k - u^*\|_Q^2.\end{aligned}$$



## Backtracking line search

- Sufficient decrease condition (let  $c_1 \in (0, 1)$ ):

$$J(u^k + \tau d^k) \leq J(u^k) + c_1 \tau \langle \nabla J(u^k), d^k \rangle. \quad (\text{A})$$

- Curvature condition (let  $c_2 \in (c_1, 1)$ ):

$$\langle \nabla J(u^k + \tau d^k), d^k \rangle \geq c_2 \langle \nabla J(u^k), d^k \rangle. \quad (\text{C})$$



## Backtracking line search

- Sufficient decrease condition (let  $c_1 \in (0, 1)$ ):

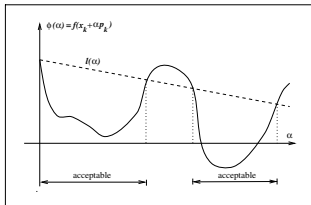
$$J(u^k + \tau d^k) \leq J(u^k) + c_1 \tau \langle \nabla J(u^k), d^k \rangle. \quad (\text{A})$$

- Curvature condition (let  $c_2 \in (c_1, 1)$ ):

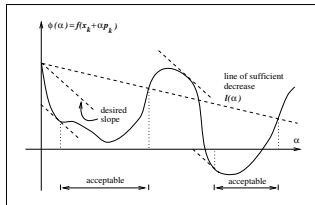
$$\langle \nabla J(u^k + \tau d^k), d^k \rangle \geq c_2 \langle \nabla J(u^k), d^k \rangle. \quad (\text{C})$$

- (A)  $\rightsquigarrow$  **Armijo** line search; (A) & (C)  $\rightsquigarrow$  **Wolfe-Powell** I.s.

Armijo I.s.



Wolfe-Powell I.s.



# Convergence of backtracking line search

## Lemma (feasibility of line search)

Assume that  $J : \mathbb{E} \rightarrow \mathbb{R}$  is continuously differentiable,  $\langle \nabla J(u^k), d^k \rangle < 0 \forall k$ , and  $0 < c_1 < c_2 < 1$ . Then there exists an open interval in which the step size  $\tau$  satisfies (A) and (C).

Proof: on board.



## Convergence of backtracking line search

### Lemma (feasibility of line search)

Assume that  $J : \mathbb{E} \rightarrow \mathbb{R}$  is continuously differentiable,  $\langle \nabla J(u^k), d^k \rangle < 0 \forall k$ , and  $0 < c_1 < c_2 < 1$ . Then there exists an open interval in which the step size  $\tau$  satisfies (A) and (C).

Proof: on board.

### Theorem (Zoutendijk)

Assume that  $J : \mathbb{E} \rightarrow \mathbb{R}$  is cont'ly differentiable, and (A) and (C) are both satisfied with  $0 < c_1 < c_2 < 1$  for each  $k$ . In addition,  $J$  is  $\mu$ -Lipschitz differentiable on  $\{u \in \mathbb{E} : J(u) \leq J(u^0)\}$ . Then

$$\sum_{k=0}^{\infty} \frac{|\langle \nabla J(u^k), d^k \rangle|^2}{\|d^k\|^2} < \infty.$$

Proof: on board.

### Remark

If  $\frac{|\langle \nabla J(u^k), d^k \rangle|}{\|\nabla J(u^k)\| \|d^k\|} \geq \text{constant} > 0$ , then  $\lim_{k \rightarrow \infty} \|\nabla J(u^k)\| = 0$ .





# Majorize-minimize method

## Majorizing function

A function  $\hat{J}(\cdot; u)$  is a **majorant** of  $J$  at  $u \in \mathbb{E}$  if

$$\begin{cases} \hat{J}(u; u) = J(u), \\ \hat{J}(\cdot; u) \geq J(\cdot). \end{cases}$$



# Majorize-minimize method

## Majorizing function

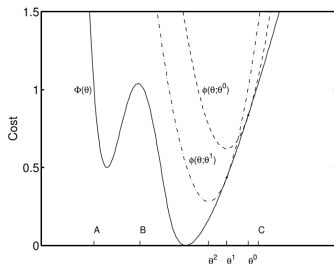
A function  $\hat{J}(\cdot; u)$  is a **majorant** of  $J$  at  $u \in \mathbb{E}$  if

$$\begin{cases} \hat{J}(u; u) = J(u), \\ \hat{J}(\cdot; u) \geq J(\cdot). \end{cases}$$

## Majorize-minimize (MM) algorithm

Let  $\hat{J}(\cdot; u)$  majorize  $J \forall u \in \mathbb{E}$ . Then the MM iteration reads:

$$u^{k+1} \in \arg \min_u \hat{J}(u; u^k).$$



## Remark

- 1 Monotonic decrease of objectives:

$$J(u^{k+1}) \leq \widehat{J}(u^{k+1}; u^k) \leq \widehat{J}(u^k; u^k) = J(u^k).$$

- 2 Efficiency of MM relies on the choice of the majorant  $\widehat{J}(\cdot; u)$ , i.e.,  $\widehat{J}(\cdot; u)$  is easy to minimize.
- 3 Common choices of  $\widehat{J}(\cdot; u)$  are quadratics.



### Remark

- 1 Monotonic decrease of objectives:

$$J(u^{k+1}) \leq \widehat{J}(u^{k+1}; u^k) \leq \widehat{J}(u^k; u^k) = J(u^k).$$

- 2 Efficiency of MM relies on the choice of the majorant  $\widehat{J}(\cdot; u)$ , i.e.,  $\widehat{J}(\cdot; u)$  is easy to minimize.
- 3 Common choices of  $\widehat{J}(\cdot; u)$  are quadratics.



### Gradient descent as MM

- Observe that  $u^{k+1} = u^k - \tau \nabla J(u^k)$  iff

$$u^{k+1} = \arg \min_u J(u^k) + \left\langle \nabla J(u^k), u - u^k \right\rangle + \frac{1}{2\tau} \|u - u^k\|^2.$$

- When  $J(u^k) + \left\langle \nabla J(u^k), \cdot - u^k \right\rangle + \frac{1}{2\tau} \|\cdot - u^k\|^2 \geq J(\cdot)$  holds?

## Lemma

Assume that  $J : \mathbb{E} \rightarrow \mathbb{R}$  is  $\mu$ -Lipschitz differentiable. Then  
 $\forall u, v \in \mathbb{E}$  :

$$|J(v) - J(u) - \langle \nabla J(u), v - u \rangle| \leq \frac{\mu}{2} \|v - u\|^2.$$

Proof: on board.



## Gradient descent as MM

### Lemma

Assume that  $J : \mathbb{E} \rightarrow \mathbb{R}$  is  $\mu$ -Lipschitz differentiable. Then  $\forall u, v \in \mathbb{E}$ :

$$|J(v) - J(u) - \langle \nabla J(u), v - u \rangle| \leq \frac{\mu}{2} \|v - u\|^2.$$

Proof: on board.

### Theorem (convergence of gradient descent)

Assume that  $J : \mathbb{E} \rightarrow \mathbb{R}$  is  $\mu$ -Lipschitz differentiable. Then the gradient descent iteration

$$u^{k+1} = u^k - \tau \nabla J(u^k)$$

with  $\tau \in (0, 1/\mu]$  yields  $\lim_{k \rightarrow \infty} \nabla J(u^k) = 0$ .

Proof: on board.



## Gradient descent as MM

### Lemma

Assume that  $J : \mathbb{E} \rightarrow \mathbb{R}$  is  $\mu$ -Lipschitz differentiable. Then  $\forall u, v \in \mathbb{E}$  :

$$|J(v) - J(u) - \langle \nabla J(u), v - u \rangle| \leq \frac{\mu}{2} \|v - u\|^2.$$

Proof: on board.

### Theorem (convergence of gradient descent)

Assume that  $J : \mathbb{E} \rightarrow \mathbb{R}$  is  $\mu$ -Lipschitz differentiable. Then the gradient descent iteration

$$u^{k+1} = u^k - \tau \nabla J(u^k)$$

with  $\tau \in (0, 1/\mu]$  yields  $\lim_{k \rightarrow \infty} \nabla J(u^k) = 0$ .

Proof: on board.

### Recipe of convergence

By solving the surrogate problem in MM, we achieve: (1) sufficient decrease in the objective; (2) inexact optimality condition which matches the exact OC in the limit.

