



# Chapter 2

## Gradient Methods

*Convex Optimization for Computer Vision*  
SS 2016

### Gradient Descent

- Definition
- Convergence analysis
- Line search
- Applications
- Conclusion

### Subgradient Method

- Definition
- Convergence Analysis
- Applications

### Gradient Projection

- Projections
- Definition
- Convergence Analysis
- Applications

### Proximal Gradient

- Proximal Operator
- Convergence Analysis
- Applications

Michael Moeller  
Thomas Möllenhoff  
Emanuel Laude  
Computer Vision Group  
Department of Computer Science  
TU München

Michael Moeller

Thomas Möllenhoff

Emanuel Laude



## Gradient Descent

Definition

Convergence analysis

Line search

Applications

Conclusion

## Subgradient Method

Definition

Convergence Analysis

Applications

## Gradient Projection

Projections

Definition

Convergence Analysis

Applications

## Proximal Gradient

Proximal Operator

Convergence Analysis

Applications



# Unconstrained and smooth optimization

Recall what the lecture is all about:

$$u^* \in \arg \min_{u \in \mathbb{R}^n} E(u),$$

for  $E : \mathbb{R}^n \rightarrow \mathbb{R} \cup \{\infty\}$  proper, closed, convex.

We start making our life easier:

- $\text{dom } E = \mathbb{R}^n$
- $E \in \mathcal{C}^1(\mathbb{R}^n)$
- Even more assumptions later :-)

## Gradient Descent

## Definition

Convergence analysis

Line search

Applications

Conclusion

## Subgradient Method

## Definition

Convergence Analysis

Applications

## Gradient Projection

## Projections

## Definition

Convergence Analysis

Applications

## Proximal Gradient

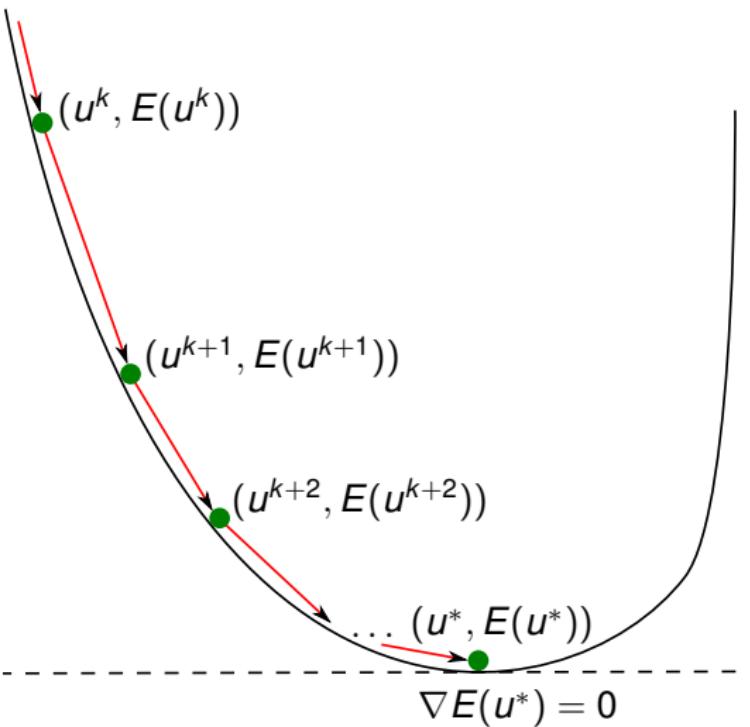
## Proximal Operator

Convergence Analysis

Applications



$$\min E(u), \quad u \in \mathbb{R}^n$$



## Gradient Descent

- Definition
- Convergence analysis
- Line search
- Applications
- Conclusion

## Subgradient Method

- Definition
- Convergence Analysis
- Applications

## Gradient Projection

- Projections
- Definition
- Convergence Analysis
- Applications

## Proximal Gradient

- Proximal Operator
- Convergence Analysis
- Applications



- Suppose we are at a point  $u^k \in \mathbb{R}^n$  where  $\nabla E(u^k) \neq 0$
- Consider the ray  $u(\tau) = u^k + \tau d$  for some direction  $d \in \mathbb{R}^n$
- Taylor expansion for  $E$  along ray

$$E(u(\tau)) = E(u^k + \tau d) = E(u^k) + \tau \langle \nabla E(u^k), d \rangle + o(\tau)$$

- The term  $\tau \langle \nabla E(u^k), d \rangle$  dominates  $o(\tau)$  for suff. small  $\tau$
- Pick  $d$  such that  $\langle \nabla E(u^k), d \rangle < 0$ , *descent direction*
- Then  $E(u(\tau)) < E(u)$  for suff. small  $\tau$

## Gradient Descent

## Definition

Convergence analysis

Line search

Applications

Conclusion

## Subgradient Method

## Definition

Convergence Analysis

Applications

## Gradient Projection

## Projections

## Definition

Convergence Analysis

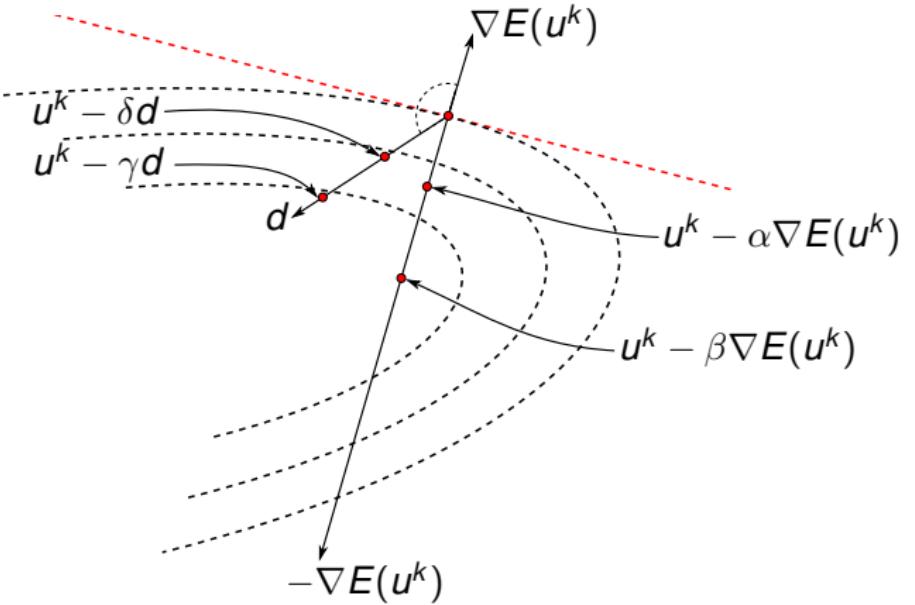
Applications

## Proximal Gradient

## Proximal Operator

Convergence Analysis

Applications



## Gradient Descent

## Definition

Convergence analysis

Line search

Applications

Conclusion

## Subgradient Method

## Definition

Convergence Analysis

Applications

## Gradient Projection

## Projections

## Definition

Convergence Analysis

Applications

## Proximal Gradient

## Proximal Operator

Convergence Analysis

Applications



- The negative gradient is the *steepest* descent direction

$$\operatorname{argmin}_{\|d\|=1} \left\{ \langle d, \nabla E(u^k) \rangle \right\} = -\frac{\nabla E(u^k)}{\|\nabla E(u^k)\|}$$

- The gradient is orthogonal to the iso-contours  $\gamma : I \rightarrow \mathbb{R}^n$

$$\nabla E(\gamma(t)) \perp \dot{\gamma}(t), \quad t \in I$$

- Possible choices of descent directions

- Scaled gradient:  $d^k = -D^k \nabla E(u^k)$ ,  $D^k \succeq 0$
- Newton:  $D^k = [\nabla^2 E(u^k)]^{-1}$
- Quasi-Newton:  $D^k \approx [\nabla^2 E(u^k)]^{-1}$
- Steepest descent:  $D^k = I$
- ...

## Gradient Descent

## Definition

Convergence analysis

Line search

Applications

Conclusion

## Subgradient Method

## Definition

Convergence Analysis

Applications

## Gradient Projection

Projections

Definition

Convergence Analysis

Applications

## Proximal Gradient

Proximal Operator

Convergence Analysis

Applications



## Gradient Descent

## Definition

Convergence analysis

Line search

Applications

Conclusion

## Subgradient Method

Definition

Convergence Analysis

Applications

## Gradient Projection

Projections

Definition

Convergence Analysis

Applications

## Proximal Gradient

Proximal Operator

Convergence Analysis

Applications

## Gradient descent

## Definition

Given a function  $E \in \mathcal{C}^1(\mathbb{R}^n)$ , an initial point  $u^0 \in \mathbb{R}^n$  and a sequence  $(\tau_k) \subset \mathbb{R}$  of step sizes, the iteration

$$u^{k+1} = u^k - \tau_k \nabla E(u^k), \quad k = 0, 1, 2, \dots,$$

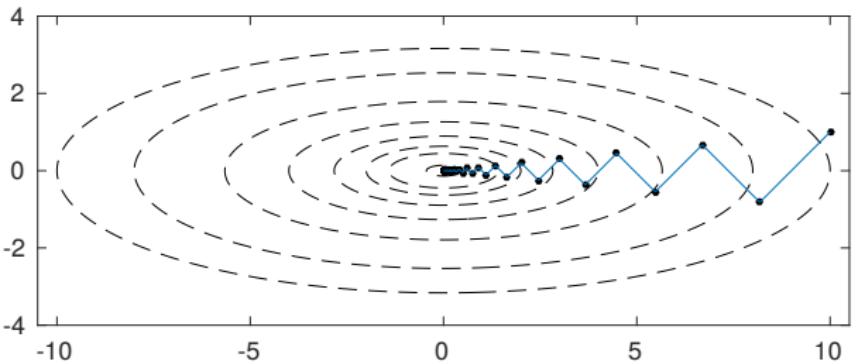
is called *gradient descent*.

Philosophy:

- Generate relaxation sequence  $\{E(u^k)\}_{k=0}^\infty$
- Each iteration is cheap, easy to code

Choice of  $\tau_k$ :

- $\tau_k = \tau$  for some constant  $\tau \in \mathbb{R}$  (this lecture)
- Exact line search  $\tau_k = \arg \min_\tau E(u^k - \tau \nabla E(u^k))$
- Inexact line search (more later)



- Convergence rate with exact line search <sup>1</sup>

$$\frac{\|u^k - u^*\|^2}{\|u^0 - u^*\|^2} \leq \left( \frac{\kappa - 1}{\kappa + 1} \right)^{2k}$$

<sup>1</sup>Nocedal and Wright, Numerical Optimization, Theorem 3.3

## Gradient Descent

Definition

Convergence analysis

Line search

Applications

Conclusion

## Subgradient Method

Definition

Convergence Analysis

Applications

## Gradient Projection

Projections

Definition

Convergence Analysis

Applications

## Proximal Gradient

Proximal Operator

Convergence Analysis

Applications



## Gradient Descent

Definition

Convergence analysis

Line search

Applications

Conclusion

## Subgradient Method

Definition

Convergence Analysis

Applications

## Gradient Projection

Projections

Definition

Convergence Analysis

Applications

## Proximal Gradient

Proximal Operator

Convergence Analysis

Applications

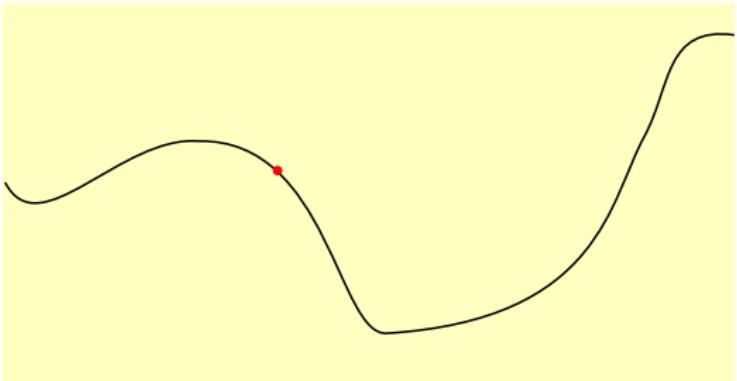
## Lipschitz continuity

## Definition

$f : \mathbb{R}^n \rightarrow \mathbb{R}^m$  is called Lipschitz continuous if for some  $L \geq 0$

$$\|f(x) - f(y)\| \leq L \|x - y\|, \quad \forall x, y \in \mathbb{R}^n.$$

- If  $L < 1$ , then  $f$  is a *contraction*
- If  $L \leq 1$ ,  $f$  is called *nonexpansive*





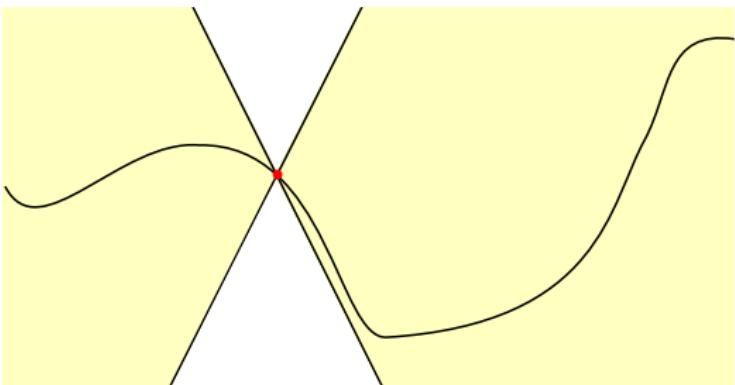
# Lipschitz continuity

## Definition

$f : \mathbb{R}^n \rightarrow \mathbb{R}^m$  is called Lipschitz continuous if for some  $L \geq 0$

$$\|f(x) - f(y)\| \leq L \|x - y\|, \quad \forall x, y \in \mathbb{R}^n.$$

- If  $L < 1$ , then  $f$  is a *contraction*
- If  $L \leq 1$ ,  $f$  is called *nonexpansive*



## Gradient Descent

### Definition

### Convergence analysis

### Line search

### Applications

### Conclusion

## Subgradient Method

### Definition

### Convergence Analysis

### Applications

## Gradient Projection

### Projections

### Definition

### Convergence Analysis

### Applications

## Proximal Gradient

### Proximal Operator

### Convergence Analysis

### Applications



## Gradient Descent

Definition

Convergence analysis

Line search

Applications

Conclusion

## Subgradient Method

Definition

Convergence Analysis

Applications

## Gradient Projection

Projections

Definition

Convergence Analysis

Applications

## Proximal Gradient

Proximal Operator

Convergence Analysis

Applications

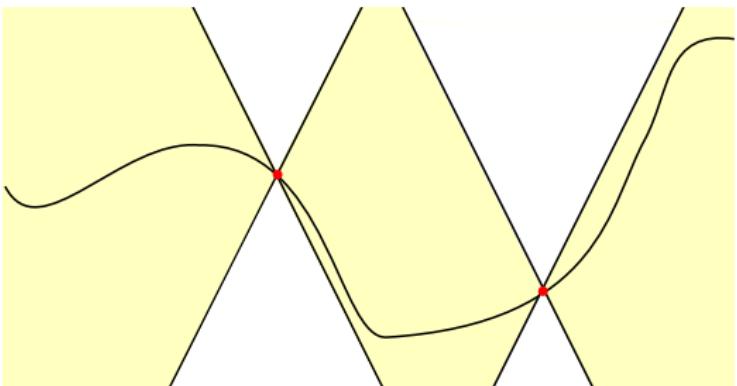
## Lipschitz continuity

## Definition

$f : \mathbb{R}^n \rightarrow \mathbb{R}^m$  is called Lipschitz continuous if for some  $L \geq 0$

$$\|f(x) - f(y)\| \leq L \|x - y\|, \quad \forall x, y \in \mathbb{R}^n.$$

- If  $L < 1$ , then  $f$  is a *contraction*
- If  $L \leq 1$ ,  $f$  is called *nonexpansive*





## Gradient Descent

Definition

Convergence analysis

Line search

Applications

Conclusion

## Subgradient Method

Definition

Convergence Analysis

Applications

## Gradient Projection

Projections

Definition

Convergence Analysis

Applications

## Proximal Gradient

Proximal Operator

Convergence Analysis

Applications

# Lipschitz continuity

- Important special case are linear functions  $f : \mathbb{R}^n \rightarrow \mathbb{R}^m$
- $f$  can be represented by matrix  $A \in \mathbb{R}^{m \times n}$
- Lipschitz constant of  $f$  is the *operator norm* or *spectral norm* of  $A$

$$\|A\| := \sup_{x \neq 0} \frac{\|Ax\|}{\|x\|} = \sup_{\|x\|=1} \|Ax\|$$

- A short calculation reveals

$$\|Ax\| \leq \|A\| \|x\|, \quad \forall x$$

- It can be shown that

$$\|A\| = \sqrt{\lambda_{\max}(A^T A)} = \sigma_{\max}(A)$$



## Gradient Descent

Definition

Convergence analysis

Line search

Applications

Conclusion

## Subgradient Method

Definition

Convergence Analysis

Applications

## Gradient Projection

Projections

Definition

Convergence Analysis

Applications

## Proximal Gradient

Proximal Operator

Convergence Analysis

Applications

# Lipschitz continuity

## Theorem: Lipschitz continuity for differentiable functions

A differentiable function  $E : \mathbb{R}^n \rightarrow \mathbb{R}^m$  is Lipschitz with parameter  $L$  if and only if  $\|\nabla E(x)\| \leq L$  for all  $x \in \mathbb{R}^n$ .

*Proof: Board!*



# Lipschitz continuity

Let  $Q \subset \mathbb{R}^n$ . We denote by  $\mathcal{C}_L^{k,p}(Q)$  the class of functions with the following properties:

- any  $f \in \mathcal{C}_L^{k,p}(Q)$  is  $k$  times continuously differentiable on  $Q$ .
- Its  $p$ -th derivative is Lipschitz continuous on  $Q$  with constant  $L$ .

## Definition: $L$ -smooth function

If  $E : \mathbb{R}^n \rightarrow \mathbb{R}$  and  $E \in \mathcal{C}_L^{1,1}(\mathbb{R}^n)$ , i.e.,

$$\|\nabla E(u) - \nabla E(v)\| \leq L \|u - v\|, \forall u, v \in \mathbb{R}^n,$$

it is called  $L$ -smooth (in some literature  $L$ -strongly smooth).

## Gradient Descent

- Definition
- Convergence analysis
- Line search
- Applications
- Conclusion

## Subgradient Method

- Definition
- Convergence Analysis
- Applications

## Gradient Projection

- Projections
- Definition
- Convergence Analysis
- Applications

## Proximal Gradient

- Proximal Operator
- Convergence Analysis
- Applications



# Convexity and Lipschitz continuity

## Reminder: Characterization of convex functions<sup>2</sup>

For  $E \in \mathcal{C}^1(\mathbb{R}^n)$  the following are equivalent

- $E(\theta u + (1 - \theta)v) \leq \theta E(u) + (1 - \theta)E(v), \forall u, v, \forall \theta \in [0, 1]$
- $E(v) \geq E(u) + \langle \nabla E(u), v - u \rangle$
- $\nabla^2 E(u) \succeq 0$ , if  $E \in \mathcal{C}^2(\mathbb{R}^n)$

## Definition: Convex functions with Lipschitz derivative

Let  $Q \subset \mathbb{R}^n$  be convex. The functions  $f \in \mathcal{C}_L^{k,p}(Q)$  which are also convex form the class  $\mathcal{F}_L^{k,p}(Q)$ .

### Gradient Descent

Definition

Convergence analysis

Line search

Applications

Conclusion

### Subgradient Method

Definition

Convergence Analysis

Applications

### Gradient Projection

Projections

Definition

Convergence Analysis

Applications

### Proximal Gradient

Proximal Operator

Convergence Analysis

Applications

<sup>2</sup>Boyd, Vandenberghe, Convex Optimization, Section 3.1.3



# Convexity and Lipschitz continuity

## Theorem: Characterization of convex $L$ -smooth functions<sup>3</sup>

For  $E \in \mathcal{F}_L^{1,1}(\mathbb{R}^n)$  the following are conditions equivalent:

- ①  $\|\nabla E(u) - \nabla E(v)\| \leq L \|u - v\|$
- ②  $\frac{L}{2} \|u\|^2 - E(u)$  is convex
- ③  $E(v) \leq E(u) + \langle \nabla E(u), v - u \rangle + \frac{L}{2} \|v - u\|^2$
- ④  $\langle \nabla E(u) - \nabla E(v), u - v \rangle \geq \frac{1}{L} \|\nabla E(u) - \nabla E(v)\|^2$
- ⑤  $\nabla^2 E(u) \preceq L \cdot I$ , if  $E \in \mathcal{C}^2(\mathbb{R}^n)$

### Gradient Descent

Definition

Convergence analysis

Line search

Applications

Conclusion

### Subgradient Method

Definition

Convergence Analysis

Applications

### Gradient Projection

Projections

Definition

Convergence Analysis

Applications

### Proximal Gradient

Proximal Operator

Convergence Analysis

Applications

<sup>3</sup>Nesterov, Introductory Lectures on Convex Optimization, Theorem 2.1.5



## Majorization minimization interpretation

- For  $E \in \mathcal{F}_L^{1,1}(\mathbb{R}^n)$  it holds for all  $u, v \in \mathbb{R}^n$

$$E(v) \leq E(u) + \langle \nabla E(u), v - u \rangle + \frac{L}{2} \|v - u\|^2$$

- Minimizing the quadratic upper bound at iterate  $u^k$  yields

$$\begin{aligned} u^{k+1} &= \underset{v}{\operatorname{argmin}} \quad E(u^k) + \langle \nabla E(u^k), v - u^k \rangle + \frac{L}{2} \|v - u^k\|^2 \\ &= u^k - \frac{1}{L} \nabla E(u^k) \end{aligned}$$

- For the minimum of the upper bound we have

$$\begin{aligned} E(u^*) &\leq \underset{v}{\operatorname{min}} \quad E(u^k) + \langle \nabla E(u^k), v - u^k \rangle + \frac{L}{2} \|v - u^k\|^2 \\ &= E(u^k) - \frac{1}{2L} \|\nabla E(u^k)\|^2 \end{aligned}$$

### Gradient Descent

Definition  
 Convergence analysis

Line search  
 Applications  
 Conclusion

### Subgradient Method

Definition  
 Convergence Analysis  
 Applications

### Gradient Projection

Projections  
 Definition  
 Convergence Analysis  
 Applications

### Proximal Gradient

Proximal Operator  
 Convergence Analysis  
 Applications



## Gradient Descent

Definition

Convergence analysis

Line search

Applications

Conclusion

## Subgradient Method

Definition

Convergence Analysis

Applications

## Gradient Projection

Projections

Definition

Convergence Analysis

Applications

## Proximal Gradient

Proximal Operator

Convergence Analysis

Applications

## Divergent example

- Minimize  $E(u) = u^4$  with gradient descent
- $\nabla E(u) = 4u^3$  is not Lipschitz
- Gradient descent iteration

$$u_{k+1} = u_k - \tau 4u_k^3 = u_k(1 - 4\tau u_k^2)$$

- For  $u_0 > \frac{1}{\sqrt{2\tau}}$  we have  $(1 - 4\tau u_0^2) < -1$  which implies

$$u_1 < -u_0$$

- Applying the above iteratively yields divergent sequence



# Strong convexity

## Definition: strong convexity

A function  $E : \mathbb{R}^n \rightarrow \overline{\mathbb{R}}$  is called *strongly convex* with constant  $m$  or  $m$ -strongly convex if  $E(u) - \frac{m}{2}\|u\|_2^2$  is still convex.

- Short exercise: strong convexity implies strict convexity
- Notation for cont. diff. and  $m$ -strongly convex:  $E \in \mathcal{S}_m^1(\mathbb{R}^n)$
- We will also consider the classes  $\mathcal{S}_{m,L}^{k,l}(\mathbb{R}^n)$  of  $m$ -strongly convex,  $k$ -times continuously differentiable functions with  $L$ -Lipschitz continuous  $l$ -th derivative

### Gradient Descent

Definition

Convergence analysis

Line search

Applications

Conclusion

### Subgradient Method

Definition

Convergence Analysis

Applications

### Gradient Projection

Projections

Definition

Convergence Analysis

Applications

### Proximal Gradient

Proximal Operator

Convergence Analysis

Applications



# Strong convexity

## Theorem: characterization of $m$ -strongly convex functions <sup>4</sup>

For  $E \in \mathcal{C}^1(\mathbb{R}^n)$  the following are equivalent:

- ①  $E(u) - \frac{m}{2} \|u\|^2$  is convex, i.e.,  $E \in \mathcal{S}_m^1(\mathbb{R}^n)$
- ②  $E(v) \geq E(u) + \langle \nabla E(u), v - u \rangle + \frac{m}{2} \|v - u\|^2$
- ③  $\langle \nabla E(u) - \nabla E(v), u - v \rangle \geq m \|u - v\|^2$
- ④  $\nabla^2 E(u) \succeq m \cdot I$ , if  $E \in \mathcal{C}^2(\mathbb{R}^n)$

### Gradient Descent

Definition

Convergence analysis

Line search

Applications

Conclusion

### Subgradient Method

Definition

Convergence Analysis

Applications

### Gradient Projection

Projections

Definition

Convergence Analysis

Applications

### Proximal Gradient

Proximal Operator

Convergence Analysis

Applications

<sup>4</sup>Ryu, Boyd, A Primer on Monotone Operator Methods, Appendix A



## Gradient Descent

Definition

Convergence analysis

Line search

Applications

Conclusion

## Subgradient Method

Definition

Convergence Analysis

Applications

## Gradient Projection

Projections

Definition

Convergence Analysis

Applications

## Proximal Gradient

Proximal Operator

Convergence Analysis

Applications

# Strong convexity and Lipschitz continuity

- The *condition number*  $\kappa$  of a function  $E \in \mathcal{S}_{m,L}^{1,1}(\mathbb{R}^n)$  is

$$\kappa = \frac{L}{m}$$

- If  $f$  is linear, i.e.,  $f(x) = Ax$  then

$$\kappa = \frac{\sqrt{\lambda_{\max}(A^T A)}}{\sqrt{\lambda_{\min}(A^T A)}} = \frac{\sigma_{\max}(A)}{\sigma_{\min}(A)}$$

- If  $f$  twice continuously differentiable, gives lower and upper bound on Hessian

$$m \cdot I \preceq \nabla^2 f(x) \preceq L \cdot I$$

→ *Online TED.*



- If initialized wrong, gradient descent doesn't converge when minimizing  $x^4$  for any fixed step size  $\tau > 0$
- Need additional structure beyond convexity for convergence analysis
- Lipschitz continuity of gradient,  $E \in \mathcal{F}_L^{1,1}(\mathbb{R}^n)$
- Strong convexity,  $E \in \mathcal{S}_m^1(\mathbb{R}^n)$
- Combination of both,  $E \in \mathcal{S}_{m,L}^{1,1}(\mathbb{R}^n)$
- **Today:** understand behaviour of gradient descent for these functions
- Some simple applications

## Gradient Descent

Definition

Convergence analysis

Line search

Applications

Conclusion

## Subgradient Method

Definition

Convergence Analysis

Applications

## Gradient Projection

Projections

Definition

Convergence Analysis

Applications

## Proximal Gradient

Proximal Operator

Convergence Analysis

Applications



# Strong convexity and Lipschitz continuity

## Theorem: strongly convex + $L$ -smooth bound

If  $E \in \mathcal{S}_{m,L}^{1,1}(\mathbb{R}^n)$ , then for any  $u, v \in \mathbb{R}^n$  we have

$$\begin{aligned} \langle \nabla E(u) - \nabla E(v), u - v \rangle &\geq \\ \frac{mL}{m+L} \|u - v\|^2 + \frac{1}{m+L} \|\nabla E(u) - \nabla E(v)\|^2 \end{aligned}$$

### Gradient Descent

Definition

Convergence analysis

Line search

Applications

Conclusion

### Subgradient Method

Definition

Convergence Analysis

Applications

### Gradient Projection

Projections

Definition

Convergence Analysis

Applications

### Proximal Gradient

Proximal Operator

Convergence Analysis

Applications



# Gradient descent convergence

## Theorem: Convergence ( $L$ -smooth + $m$ -strongly convex)

Let  $E \in \mathcal{S}_{m,L}^{1,1}(\mathbb{R}^n)$ . For the sequence  $(u^k)_k$  produced by gradient descent with step size  $0 < \tau \leq 2/(m + L)$  we have

$$\|u^k - u^*\|^2 \leq c^k \|u^0 - u^*\|^2,$$

$$E(u^k) - E(u^*) \leq \frac{Lc^k}{2} \|u^0 - u^*\|^2,$$

with  $c = 1 - \tau \frac{2mL}{m+L}$ .

*Proof: Board!*

Remarks:

- Optimal choice is  $\tau = 2/(m + L)$
- Results in factor  $c = \left(\frac{\kappa-1}{\kappa+1}\right)^2$ ,  $\kappa = L/m$

## Gradient Descent

Definition

Convergence analysis

Line search

Applications

Conclusion

## Subgradient Method

Definition

Convergence Analysis

Applications

## Gradient Projection

Projections

Definition

Convergence Analysis

Applications

## Proximal Gradient

Proximal Operator

Convergence Analysis

Applications



# Gradient descent convergence

## Theorem: Convergence ( $L$ -smooth)

Let  $E \in \mathcal{F}_L^{1,1}(\mathbb{R}^n)$ . For the sequence  $(u_k)_k$  produced by gradient descent with step size  $0 < \tau \leq 1/L$  we have

$$E(u^k) - E(u^*) \leq \frac{1}{2k\tau} \|u^0 - u^*\|^2.$$

*Proof: Board!*

### Gradient Descent

Definition

Convergence analysis

Line search

Applications

Conclusion

### Subgradient Method

Definition

Convergence Analysis

Applications

### Gradient Projection

Projections

Definition

Convergence Analysis

Applications

### Proximal Gradient

Proximal Operator

Convergence Analysis

Applications



# Rates of convergence

## Reminder: $\mathcal{O}$ -notation

$$\mathcal{O}(g) = \{f \mid \exists C \geq 0, \exists n_0 \in \mathbb{N}_0, \forall n \geq n_0 : |f(n)| \leq C|g(n)|\}$$

## Sublinear rate

- $r(k) = \mathcal{O}(\frac{1}{k^c})$ ,  $c > 0$
- New correct digit takes the amount of computations comparable with total amount of previous work.
- Constant factor in  $\mathcal{O}$ -notation plays a significant role

## Linear rate

- $r(k) = \mathcal{O}(c^k)$ ,  $c < 1$
- Each new correct digit takes a constant amount of computations

## Gradient Descent

Definition

Convergence analysis

Line search

Applications

Conclusion

## Subgradient Method

Definition

Convergence Analysis

Applications

## Gradient Projection

Projections

Definition

Convergence Analysis

Applications

## Proximal Gradient

Proximal Operator

Convergence Analysis

Applications



# Worst-case complexities

- First order method:

$$u^{k+1} \in u^0 + \text{span}\{\nabla E(u^0), \dots, \nabla E(u^k)\}$$

- We have shown the following for gradient descent:

- $E \in \mathcal{F}_L^{1,1}$  gives  $\mathcal{O}(1/k)$  convergence
- $E \in \mathcal{S}_{m,L}^{1,1}$  gives  $\mathcal{O}\left(\left(\frac{\kappa-1}{\kappa+1}\right)^{2k}\right)$  convergence

- Worst-case complexity of first-order methods <sup>5</sup>

- For  $E \in \mathcal{F}_L^{1,1}$  there is a  $\mathcal{O}(1/k^2)$  lower bound
- For  $E \in \mathcal{S}_{m,L}^{1,1}$  the lower bound is  $\mathcal{O}\left(\left(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\right)^{2k}\right)$

- It turns out that these lower bounds can be attained
- Theoretical convergence rates only tell half the story

## Gradient Descent

- Definition
- Convergence analysis
- Line search
- Applications
- Conclusion

## Subgradient Method

- Definition
- Convergence Analysis
- Applications

## Gradient Projection

- Projections
- Definition
- Convergence Analysis
- Applications

## Proximal Gradient

- Proximal Operator
- Convergence Analysis
- Applications

<sup>5</sup>Nesterov, Introductory Lectures on Convex Optimization, Theorem 2.1.7 and Theorem 2.1.13



## Gradient Descent

Definition

Convergence analysis

## Line search

Applications

Conclusion

## Subgradient Method

Definition

Convergence Analysis

Applications

## Gradient Projection

Projections

Definition

Convergence Analysis

Applications

## Proximal Gradient

Proximal Operator

Convergence Analysis

Applications

## Line search

- Sometimes Lipschitz constant  $L$  not known
- Use backtracking line search to estimate  $\tau_k$  each iteration
- Pick  $\alpha \in (0, 0.5)$ ,  $\beta \in (0, 1)$
- Then determine  $\tau_k$  each iteration by:

$$\tau_k \leftarrow 1$$

$$\text{while } E\left(u^k - \tau_k \nabla E(u^k)\right) > E(u^k) - \alpha \tau_k \|\nabla E(u^k)\|^2$$

$$\tau_k \leftarrow \beta \tau_k$$

end

- Often leads to improved convergence in practice
- (Slight) overhead each iteration
- Theory: same convergence rate as with constant steps

Observed image  $f \in \mathbb{R}^N$ Denoised image  $u^* \in \mathbb{R}^N$ 

$$u^* \in \operatorname{argmax}_{u \in \mathbb{R}^N} p(u|f) = \operatorname{argmax}_{u \in \mathbb{R}^N} \frac{p(f|u)p(u)}{p(f)}$$

## Gradient Descent

Definition

Convergence analysis

Line search

## Applications

Conclusion

## Subgradient Method

Definition

Convergence Analysis

Applications

## Gradient Projection

Projections

Definition

Convergence Analysis

Applications

## Proximal Gradient

Proximal Operator

Convergence Analysis

Applications



- Gaussian noise assumption  $f_i \sim \mathcal{N}(u_i, \sigma)$

$$p(f_i|u_i) \propto \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(u_i - f_i)^2}{2\sigma^2}\right)$$

- Impose prior distribution on image gradient  $Du \in \mathbb{R}^{2N}$

$$p(u) \propto \prod_{i=1}^{2N} \exp(-\varphi((Du)_i))$$

- Natural image statistics suggest the choice

$$\varphi(x) = c_\varepsilon(x) = \sqrt{x^2 + \varepsilon^2}$$

### Gradient Descent

Definition

Convergence analysis

Line search

**Applications**

Conclusion

### Subgradient Method

Definition

Convergence Analysis

Applications

### Gradient Projection

Projections

Definition

Convergence Analysis

Applications

### Proximal Gradient

Proximal Operator

Convergence Analysis

Applications

Michael Moeller

Thomas Möllenhoff

Emanuel Laude



## Gradient Descent

Definition

Convergence analysis

Line search

## Applications

Conclusion

## Subgradient Method

Definition

Convergence Analysis

Applications

## Gradient Projection

Projections

Definition

Convergence Analysis

Applications

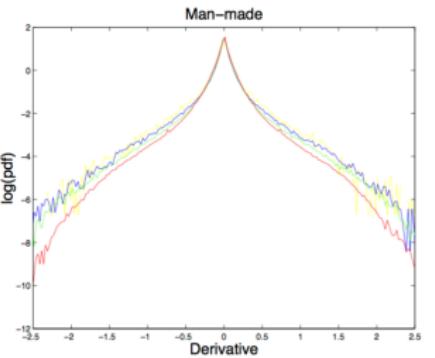
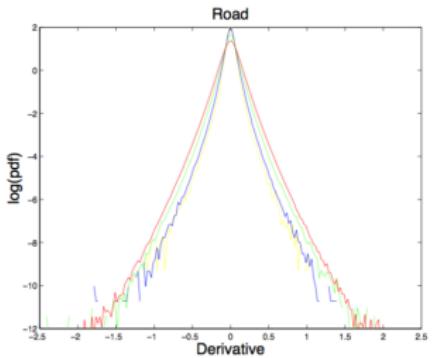
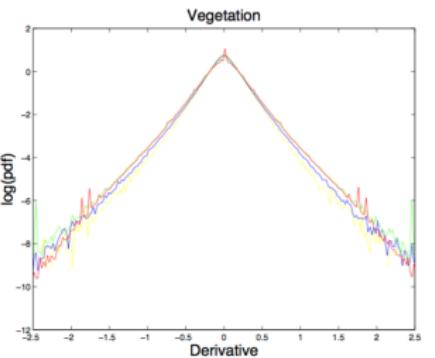
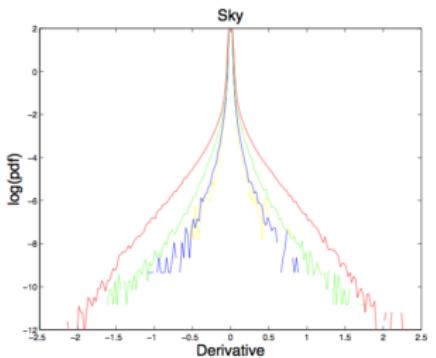
## Proximal Gradient

Proximal Operator

Convergence Analysis

Applications

updated 18.05.2016

Natural image statistics<sup>6</sup>

<sup>6</sup>J. Huang, D. Mumford, Statistics of Natural Images and Models, CVPR '99



- Minimize negative logarithm

$$\begin{aligned}
 u^* &\in \operatorname{argmin}_{u \in \mathbb{R}^N} -\log p(f|u)p(u) \\
 &= \operatorname{argmin}_{u \in \mathbb{R}^N} -\log p(f|u) - \log p(u) \\
 &= \operatorname{argmin}_{u \in \mathbb{R}^N} \underbrace{\frac{\lambda}{2} \|u - f\|^2 + \sum_{i=1}^{2N} c_\varepsilon((Du)_i)}_{=: E(u)}
 \end{aligned}$$

- $E(u)$  is  $\lambda$ -strongly convex and  $L$ -smooth with  $L = \lambda + \frac{\|D\|^2}{\varepsilon}$
- Proof and implementation: last week's exercises :-)

## Gradient Descent

Definition

Convergence analysis

Line search

## Applications

Conclusion

## Subgradient Method

Definition

Convergence Analysis

Applications

## Gradient Projection

Projections

Definition

Convergence Analysis

Applications

## Proximal Gradient

Proximal Operator

Convergence Analysis

Applications

Michael Moeller

Thomas Möllenhoff

Emanuel Laude



## Gradient Descent

Definition

Convergence analysis

Line search

## Applications

Conclusion

## Subgradient Method

Definition

Convergence Analysis

Applications

## Gradient Projection

Projections

Definition

Convergence Analysis

Applications

## Proximal Gradient

Proximal Operator

Convergence Analysis

Applications

# Image denoising



# Evolution to global optimum via gradient descent

Gradient Methods

Michael Moeller

Thomas Möllenhoff

Emanuel Laude



## Gradient Descent

Definition

Convergence analysis

Line search

### Applications

Conclusion

## Subgradient Method

Definition

Convergence Analysis

Applications

## Gradient Projection

Projections

Definition

Convergence Analysis

Applications

## Proximal Gradient

Proximal Operator

Convergence Analysis

Applications

$\varepsilon = 0.1$

Gradient Methods

Michael Moeller

Thomas Möllenhoff

Emanuel Laude



Gradient Descent

Definition

Convergence analysis

Line search

Applications

Conclusion

Subgradient Method

Definition

Convergence Analysis

Applications

Gradient Projection

Projections

Definition

Convergence Analysis

Applications

Proximal Gradient

Proximal Operator

Convergence Analysis

Applications

$\varepsilon = 0.01$



## Gradient Methods

Michael Moeller  
Thomas Möllenhoff  
Emanuel Laude



### Gradient Descent

Definition  
Convergence analysis  
Line search  
**Applications**  
Conclusion

### Subgradient Method

Definition  
Convergence Analysis  
Applications

### Gradient Projection

Projections  
Definition  
Convergence Analysis  
Applications

### Proximal Gradient

Proximal Operator  
Convergence Analysis  
Applications

→ *Motivation for non-smooth optimization!*



## Gradient Descent

- Definition
- Convergence analysis
- Line search
- Applications**
- Conclusion

## Subgradient Method

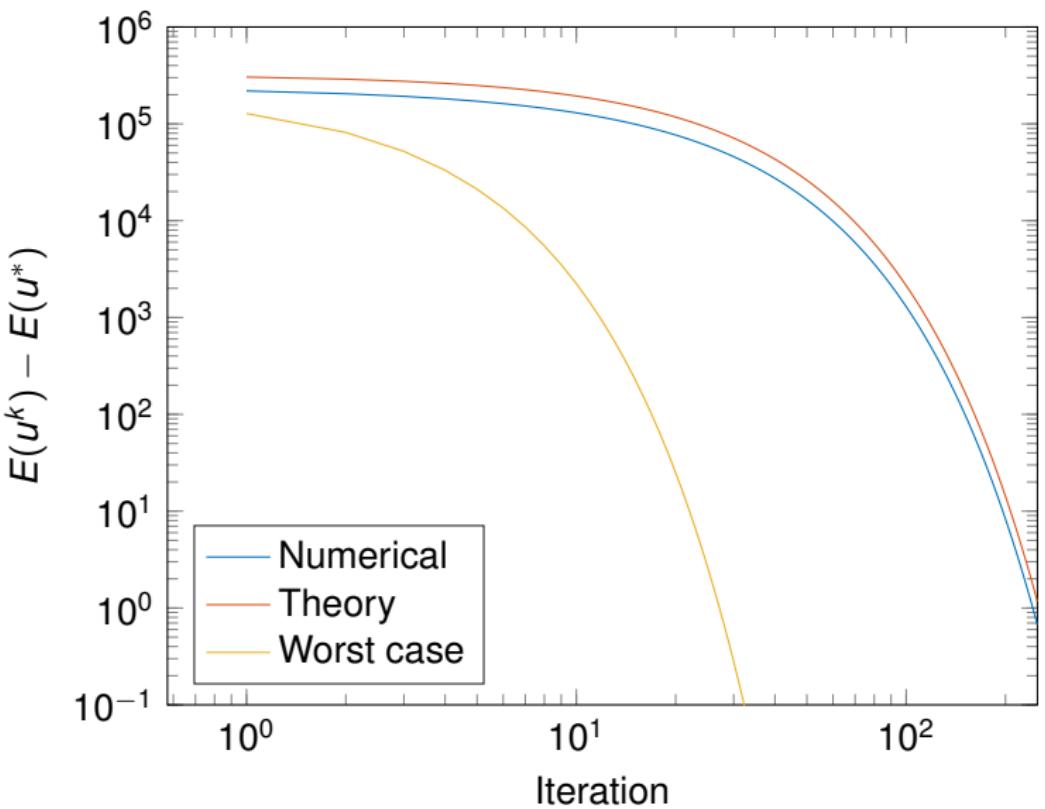
- Definition
- Convergence Analysis
- Applications

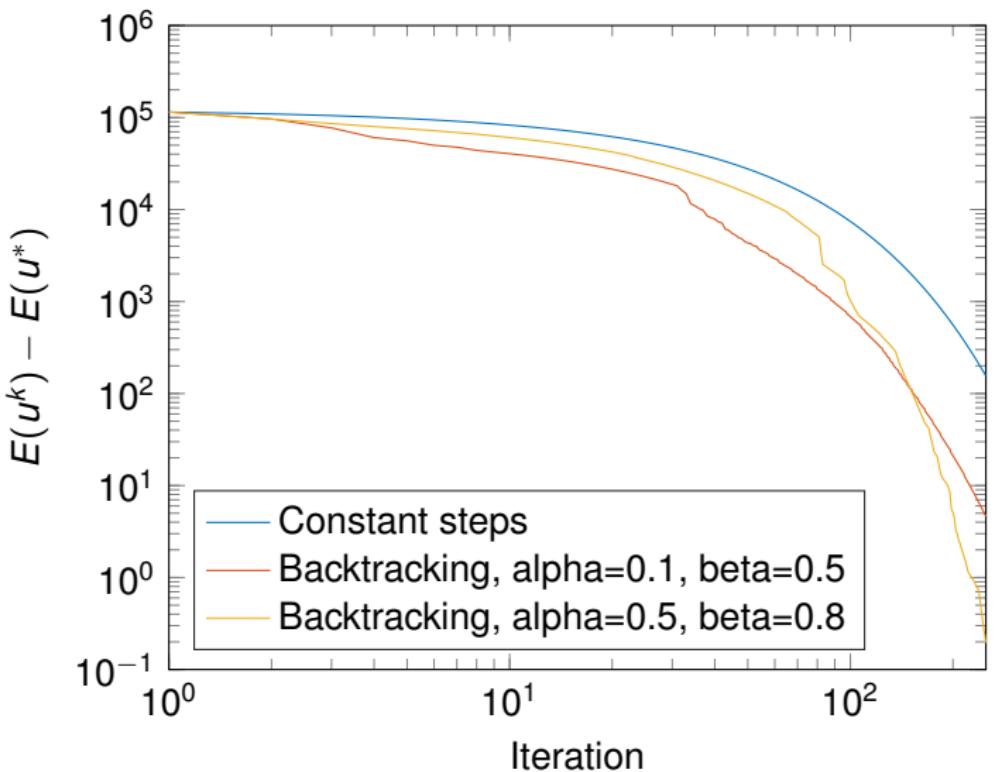
## Gradient Projection

- Projections
- Definition
- Convergence Analysis
- Applications

## Proximal Gradient

- Proximal Operator
- Convergence Analysis
- Applications

Convergence,  $\tau = 2/(m + L)$ 



## Gradient Descent

- Definition
- Convergence analysis
- Line search
- Applications**
- Conclusion

## Subgradient Method

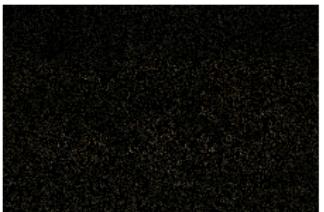
- Definition
- Convergence Analysis
- Applications

## Gradient Projection

- Projections
- Definition
- Convergence Analysis
- Applications

## Proximal Gradient

- Proximal Operator
- Convergence Analysis
- Applications



$$f \in \mathbb{R}^N$$

$$1 - m \in \mathbb{R}^N$$

$$u^* \in \mathbb{R}^N$$

$$u^* \in \operatorname{argmin}_u \frac{\lambda}{2} \|m \cdot (u - f)\|^2 + \sum_{i=1}^{2N} c_\varepsilon ((\nabla u)_i)$$

- Energy is not strongly convex, but  $L$ -smooth
- Sublinear  $\mathcal{O}(1/k)$  upper bound on convergence speed

## Gradient Descent

Definition

Convergence analysis

Line search

## Applications

Conclusion

## Subgradient Method

Definition

Convergence Analysis

Applications

## Gradient Projection

Projections

Definition

Convergence Analysis

Applications

## Proximal Gradient

Proximal Operator

Convergence Analysis

Applications

Michael Moeller

Thomas Möllenhoff

Emanuel Laude



## Gradient Descent

Definition

Convergence analysis

Line search

### Applications

Conclusion

## Subgradient Method

Definition

Convergence Analysis

Applications

## Gradient Projection

Projections

Definition

Convergence Analysis

Applications

## Proximal Gradient

Proximal Operator

Convergence Analysis

Applications

Michael Moeller

Thomas Möllenhoff

Emanuel Laude



## Gradient Descent

Definition

Convergence analysis

Line search

### Applications

Conclusion

## Subgradient Method

Definition

Convergence Analysis

Applications

## Gradient Projection

Projections

Definition

Convergence Analysis

Applications

## Proximal Gradient

Proximal Operator

Convergence Analysis

Applications

Michael Moeller

Thomas Möllenhoff

Emanuel Laude



## Gradient Descent

Definition

Convergence analysis

Line search

## Applications

Conclusion

## Subgradient Method

Definition

Convergence Analysis

Applications

## Gradient Projection

Projections

Definition

Convergence Analysis

Applications

## Proximal Gradient

Proximal Operator

Convergence Analysis

Applications

Michael Moeller

Thomas Möllenhoff

Emanuel Laude



## Gradient Descent

Definition

Convergence analysis

Line search

### Applications

Conclusion

## Subgradient Method

Definition

Convergence Analysis

Applications

## Gradient Projection

Projections

Definition

Convergence Analysis

Applications

## Proximal Gradient

Proximal Operator

Convergence Analysis

Applications

Michael Moeller

Thomas Möllenhoff

Emanuel Laude



## Gradient Descent

Definition

Convergence analysis

Line search

### Applications

Conclusion

## Subgradient Method

Definition

Convergence Analysis

Applications

## Gradient Projection

Projections

Definition

Convergence Analysis

Applications

## Proximal Gradient

Proximal Operator

Convergence Analysis

Applications

# 70% missing pixels



Michael Moeller

Thomas Möllenhoff

Emanuel Laude



## Gradient Descent

Definition

Convergence analysis

Line search

## Applications

Conclusion

## Subgradient Method

Definition

Convergence Analysis

Applications

## Gradient Projection

Projections

Definition

Convergence Analysis

Applications

## Proximal Gradient

Proximal Operator

Convergence Analysis

Applications

Michael Moeller

Thomas Möllenhoff

Emanuel Laude



## Gradient Descent

Definition

Convergence analysis

Line search

Applications

Conclusion

## Subgradient Method

Definition

Convergence Analysis

Applications

## Gradient Projection

Projections

Definition

Convergence Analysis

Applications

## Proximal Gradient

Proximal Operator

Convergence Analysis

Applications



## Gradient Descent

Definition

Convergence analysis

Line search

## Applications

Conclusion

## Subgradient Method

Definition

Convergence Analysis

Applications

## Gradient Projection

Projections

Definition

Convergence Analysis

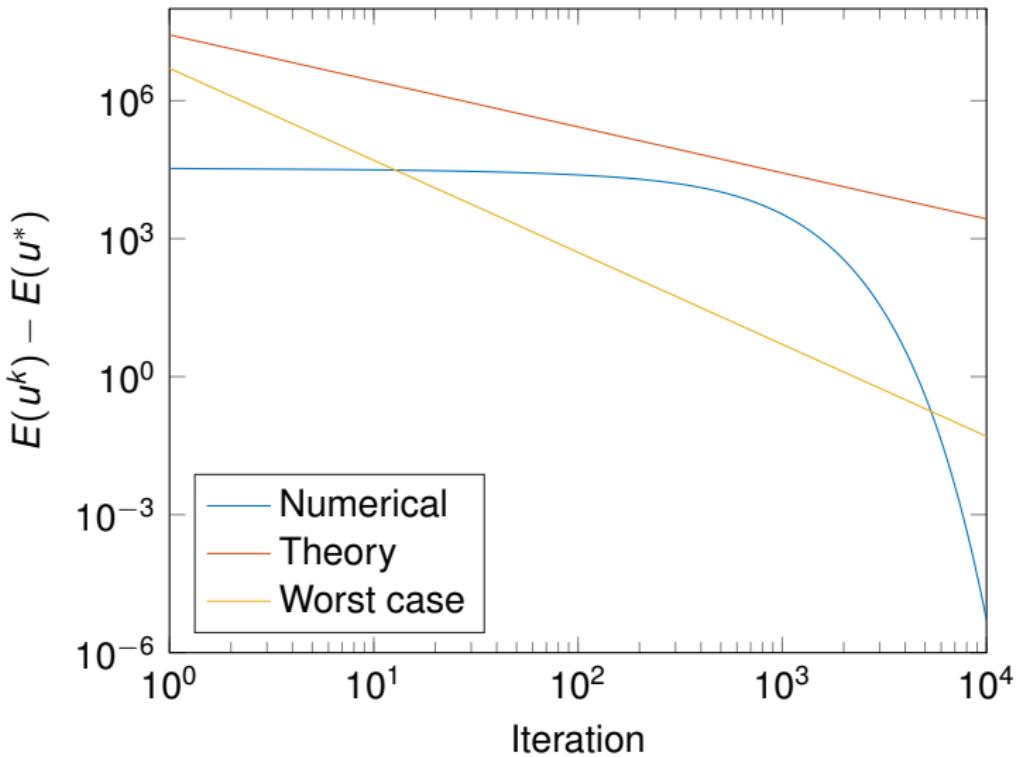
Applications

## Proximal Gradient

Proximal Operator

Convergence Analysis

Applications

Convergence,  $\tau = 1/L$ 



- Minimize the inpainting energy

$$E(u) = \frac{\lambda}{2} \|m \cdot (u - f)\|^2 + \sum_{i=1}^{2N} h_\varepsilon((Du)_i) + \beta \|u\|^2$$

- Huber penalty  $h_\varepsilon(x) = \begin{cases} \frac{x^2}{2\varepsilon} & \text{if } |x| \leq \varepsilon, \\ |x| - \frac{\varepsilon}{2} & \text{otherwise.} \end{cases}$
- Given all the parameters, return the solution once

$$\frac{E(u^k) - E(u^*)}{E(u^*)} < \delta$$

- See template `challenge_hubert_inpainting.m`
- Live leaderboard on homepage
- Fastest solution at end of semester receives a prize

## Gradient Descent

Definition  
 Convergence analysis  
 Line search  
**Applications**  
 Conclusion

## Subgradient Method

Definition  
 Convergence Analysis  
 Applications

## Gradient Projection

Projections  
 Definition  
 Convergence Analysis  
 Applications

## Proximal Gradient

Proximal Operator  
 Convergence Analysis  
 Applications



- MNIST dataset<sup>7</sup>, handwritten digit recognition
- $K = 10$  digits,  $28 \times 28$  grayscale images
- $n = 60000$  training images  $X \in \mathbb{R}^{n \times 768}$ , with ground-truth labels  $Y \in \{1, \dots, 10\}^n$
- Learn simple *linear* model  $W \in \mathbb{R}^{10 \times 768}$  on raw pixel data
- Softmax regression (multinomial logistic regression)

$$p(y_i = k | x_i, W) = \frac{\exp(\langle w_k, x_i \rangle)}{\sum_{j=1}^K \exp(\langle w_j, x_i \rangle)}$$

<sup>7</sup><http://yann.lecun.com/exdb/mnist/>

## Gradient Descent

Definition  
 Convergence analysis  
 Line search  
**Applications**  
 Conclusion

## Subgradient Method

Definition  
 Convergence Analysis  
 Applications

## Gradient Projection

Projections  
 Definition  
 Convergence Analysis  
 Applications

## Proximal Gradient

Proximal Operator  
 Convergence Analysis  
 Applications



- Minimize negative log-likelihood

$$E(W) = -\log \frac{1}{n} \prod_{i=1}^n \prod_{k=1}^K p(y_i = k | x_i, W)^{1\{y_i=k\}} p(W)$$

$$= -\frac{1}{n} \sum_{i=1}^n \sum_{k=1}^K 1\{y_i = k\} \log p(y_i = k | x_i, W) + \lambda \|W\|_F^2$$

- It can be shown that  $E(W)$  is  $\lambda$ -strongly convex
- $E(W)$  is also  $L$ -smooth (bound:  $\lambda + \frac{\|X\|^2}{4n}$ )
- Minimize using gradient descent with  $\tau = \frac{2}{2\lambda + \|X\|^2/4n}$
- Gradient computation expensive  $\rightarrow$  *stochastic* methods!  
(we won't cover them)

## Gradient Descent

Definition  
 Convergence analysis  
 Line search  
**Applications**  
 Conclusion

## Subgradient Method

Definition  
 Convergence Analysis  
 Applications

## Gradient Projection

Projections  
 Definition  
 Convergence Analysis  
 Applications

## Proximal Gradient

Proximal Operator  
 Convergence Analysis  
 Applications



## Gradient Descent

Definition  
Convergence analysis  
Line search  
**Applications**  
Conclusion

## Subgradient Method

Definition  
Convergence Analysis  
Applications

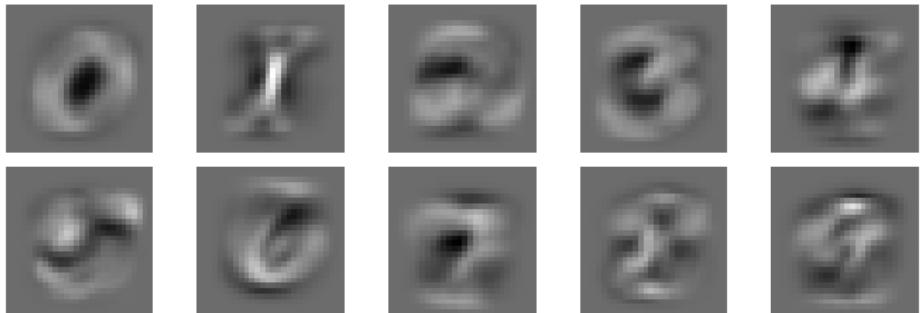
## Gradient Projection

Projections  
Definition  
Convergence Analysis  
Applications

## Proximal Gradient

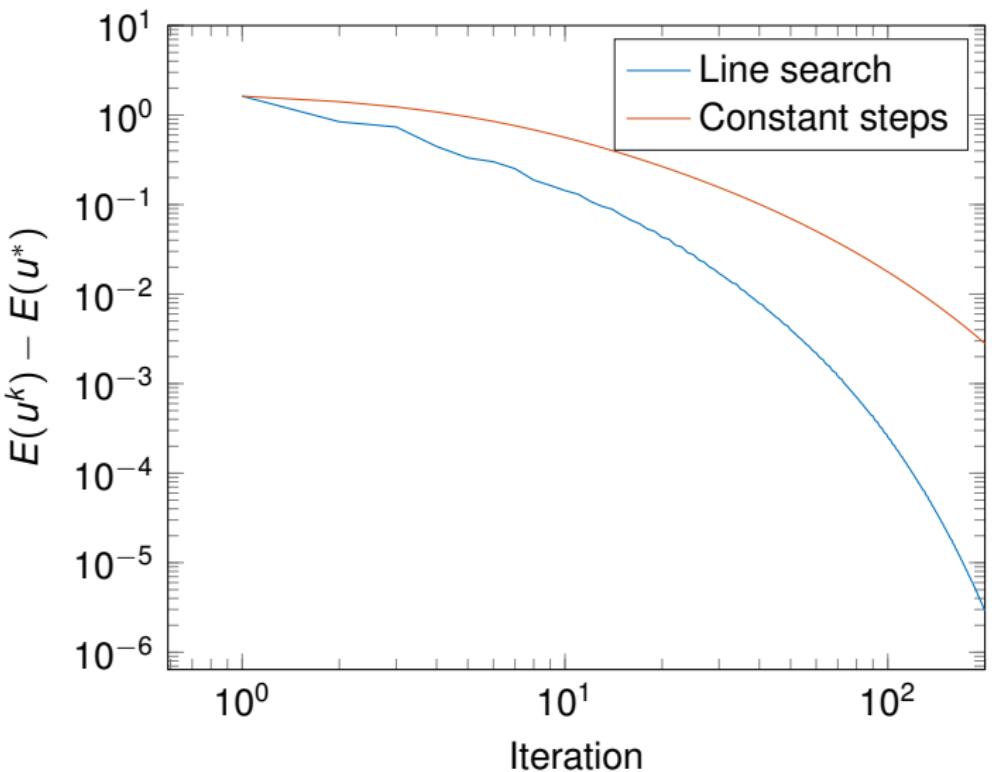
Proximal Operator  
Convergence Analysis  
Applications

# Multinomial logistic regression



- Classifier gives around 10% error on test set
- Can be easily improved to around 1 – 2% with a few additional lines of MATLAB code (use features instead of raw pixels)
- Current best: 0.23% (convolutional neural networks)
- Learn more about learning:

<https://vision.in.tum.de/teaching/ss2016/mlcv16>



## Gradient Descent

- Definition
- Convergence analysis
- Line search
- Applications**
- Conclusion

## Subgradient Method

- Definition
- Convergence Analysis
- Applications

## Gradient Projection

- Projections
- Definition
- Convergence Analysis
- Applications

## Proximal Gradient

- Proximal Operator
- Convergence Analysis
- Applications



- GD is still popular to date due to its simplicity and flexibility
- Various theoretically optimal extensions (Heavy-ball acceleration, Nesterov momentum) exist
- *Envelope approach*: many advanced algorithms for non-smooth optimization are just gradient descent on a particular (albeit complicated) energy
- Endless of variants and modifications of descent methods
- conjugate, accelerated, preconditioned, projected, conditional, mirrored, stochastic, coordinate, continuous, online, variable metric, subgradient, proximal, ...

## Gradient Descent

Definition

Convergence analysis

Line search

Applications

Conclusion

## Subgradient Method

Definition

Convergence Analysis

Applications

## Gradient Projection

Projections

Definition

Convergence Analysis

Applications

## Proximal Gradient

Proximal Operator

Convergence Analysis

Applications

Michael Moeller

Thomas Möllenhoff

Emanuel Laude



## Gradient Descent

Definition

Convergence analysis

Line search

Applications

Conclusion

## Subgradient Method

Definition

Convergence Analysis

Applications

## Gradient Projection

Projections

Definition

Convergence Analysis

Applications

## Proximal Gradient

Proximal Operator

Convergence Analysis

Applications



- Last lecture: analysis of gradient descent method
- Assumption: energy  $E(u)$  is  $L$ -smooth
- However, many energies in practice are not even differentiable
- Smoothing the energy leads to poor approximation and high condition number
- Last week: subdifferential  $\partial E$  as a generalization of the gradient for nonsmooth functions
- Can we use it to construct an algorithm?

## Gradient Descent

Definition  
Convergence analysis  
Line search  
Applications  
Conclusion

## Subgradient Method

Definition  
Convergence Analysis  
Applications

## Gradient Projection

Projections  
Definition  
Convergence Analysis  
Applications

## Proximal Gradient

Proximal Operator  
Convergence Analysis  
Applications



# Subgradient descent

## Definition

Given a convex function  $E : \mathbb{R}^n \rightarrow \mathbb{R}$ , an initial point  $u^0 \in \mathbb{R}^n$  and a sequence  $(\tau_k) \subset \mathbb{R}$  of step sizes, the iteration

$$u^{k+1} = u^k - \tau_k g^k, \text{ where } g^k \in \partial E(u^k), \quad k = 0, 1, 2, \dots,$$

is called *subgradient descent*.

Some remarks:

- $g^k$  can be *any* subgradient of  $E$  at  $u^k$
- Simple to implement
- Typically low per iteration complexity
- We'll see later: not a descent method

## Gradient Descent

Definition

Convergence analysis

Line search

Applications

Conclusion

## Subgradient Method

Definition

Convergence Analysis

Applications

## Gradient Projection

Projections

Definition

Convergence Analysis

Applications

## Proximal Gradient

Proximal Operator

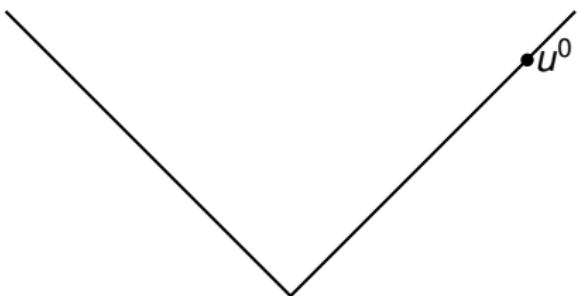
Convergence Analysis

Applications



- Let's use it to minimize  $E(u) = |u|$  with  $\tau_k = \tau$
- Iteration is given by

$$u^{k+1} = u^k - \tau \text{sign}(u^k)$$



## Gradient Descent

Definition

Convergence analysis

Line search

Applications

Conclusion

## Subgradient Method

Definition

Convergence Analysis

Applications

## Gradient Projection

Projections

Definition

Convergence Analysis

Applications

## Proximal Gradient

Proximal Operator

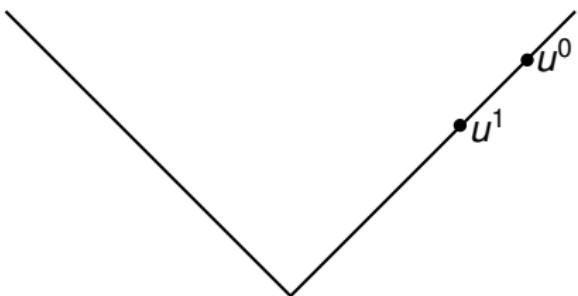
Convergence Analysis

Applications



- Let's use it to minimize  $E(u) = |u|$  with  $\tau_k = \tau$
- Iteration is given by

$$u^{k+1} = u^k - \tau \text{sign}(u^k)$$



## Gradient Descent

Definition  
 Convergence analysis  
 Line search  
 Applications  
 Conclusion

## Subgradient Method

Definition  
 Convergence Analysis  
 Applications

## Gradient Projection

Projections  
 Definition  
 Convergence Analysis  
 Applications

## Proximal Gradient

Proximal Operator  
 Convergence Analysis  
 Applications



## Gradient Descent

Definition  
 Convergence analysis  
 Line search  
 Applications  
 Conclusion

## Subgradient Method

Definition  
 Convergence Analysis  
 Applications

## Gradient Projection

Projections  
 Definition  
 Convergence Analysis  
 Applications

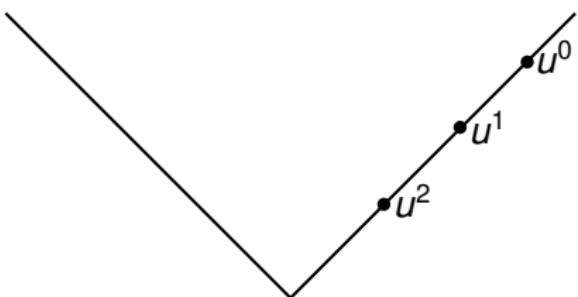
## Proximal Gradient

Proximal Operator  
 Convergence Analysis  
 Applications

## First example

- Let's use it to minimize  $E(u) = |u|$  with  $\tau_k = \tau$
- Iteration is given by

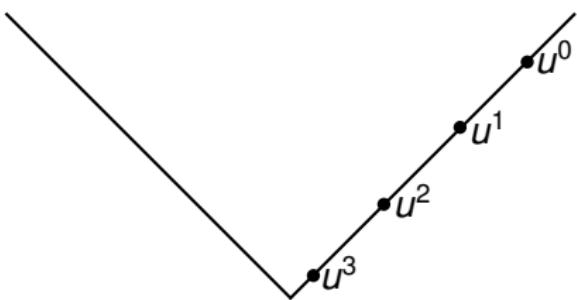
$$u^{k+1} = u^k - \tau \text{sign}(u^k)$$





- Let's use it to minimize  $E(u) = |u|$  with  $\tau_k = \tau$
- Iteration is given by

$$u^{k+1} = u^k - \tau \text{sign}(u^k)$$



## Gradient Descent

Definition  
Convergence analysis  
Line search  
Applications  
Conclusion

## Subgradient Method

Definition  
Convergence Analysis  
Applications

## Gradient Projection

Projections  
Definition  
Convergence Analysis  
Applications

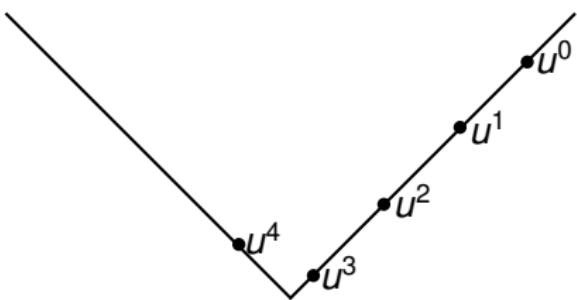
## Proximal Gradient

Proximal Operator  
Convergence Analysis  
Applications



- Let's use it to minimize  $E(u) = |u|$  with  $\tau_k = \tau$
- Iteration is given by

$$u^{k+1} = u^k - \tau \text{sign}(u^k)$$



## Gradient Descent

Definition

Convergence analysis

Line search

Applications

Conclusion

## Subgradient Method

Definition

Convergence Analysis

Applications

## Gradient Projection

Projections

Definition

Convergence Analysis

Applications

## Proximal Gradient

Proximal Operator

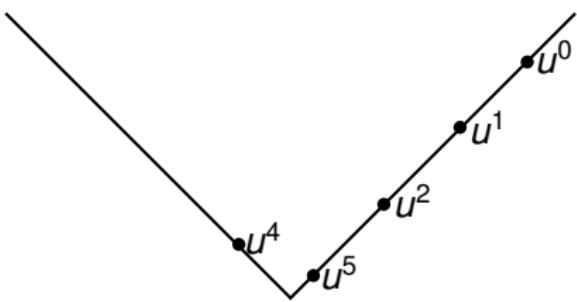
Convergence Analysis

Applications



- Let's use it to minimize  $E(u) = |u|$  with  $\tau_k = \tau$
- Iteration is given by

$$u^{k+1} = u^k - \tau \text{sign}(u^k)$$



## Gradient Descent

Definition  
 Convergence analysis  
 Line search  
 Applications  
 Conclusion

## Subgradient Method

Definition  
 Convergence Analysis  
 Applications

## Gradient Projection

Projections  
 Definition  
 Convergence Analysis  
 Applications

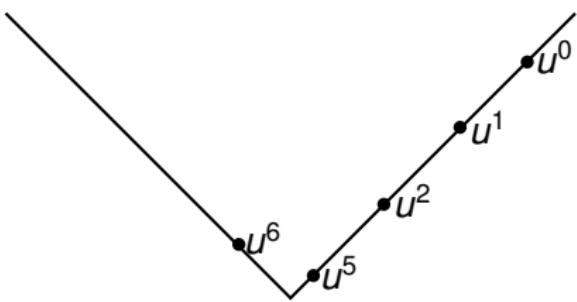
## Proximal Gradient

Proximal Operator  
 Convergence Analysis  
 Applications



- Let's use it to minimize  $E(u) = |u|$  with  $\tau_k = \tau$
- Iteration is given by

$$u^{k+1} = u^k - \tau \text{sign}(u^k)$$



## Gradient Descent

Definition  
 Convergence analysis  
 Line search  
 Applications  
 Conclusion

## Subgradient Method

Definition  
 Convergence Analysis  
 Applications

## Gradient Projection

Projections  
 Definition  
 Convergence Analysis  
 Applications

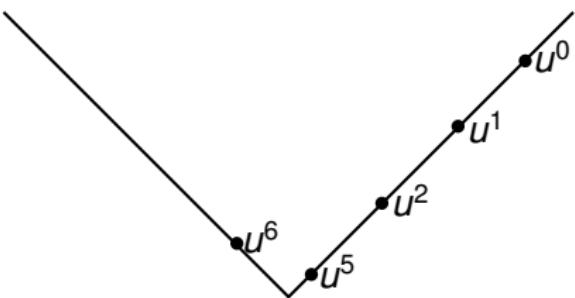
## Proximal Gradient

Proximal Operator  
 Convergence Analysis  
 Applications



- Let's use it to minimize  $E(u) = |u|$  with  $\tau_k = \tau$
- Iteration is given by

$$u^{k+1} = u^k - \tau \text{sign}(u^k)$$



- Doesn't converge to optimum for constant step sizes

### Gradient Descent

Definition  
Convergence analysis  
Line search  
Applications  
Conclusion

### Subgradient Method

Definition  
Convergence Analysis  
Applications

### Gradient Projection

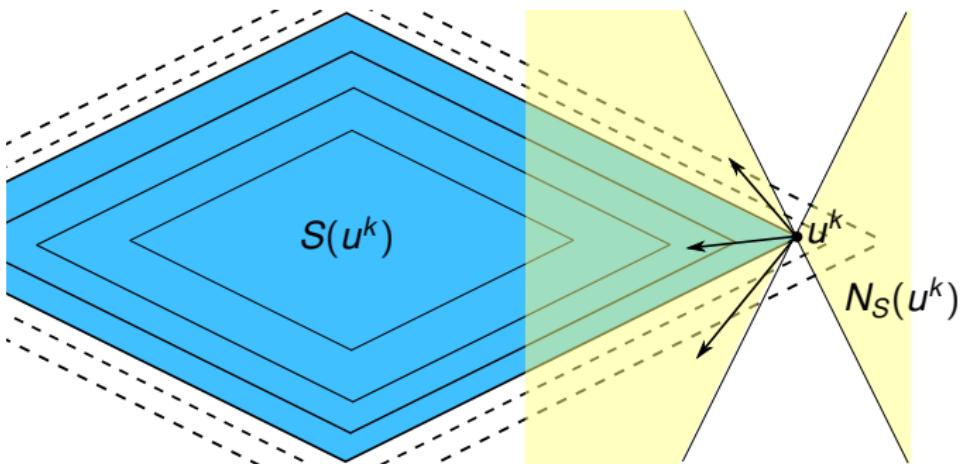
Projections  
Definition  
Convergence Analysis  
Applications

### Proximal Gradient

Proximal Operator  
Convergence Analysis  
Applications



- Minimize  $E(u) = |u_1| + 2|u_2|$



- Consider sub level sets at point  $u^k$  (shown in blue)

$$S(u^k) = \{u \in \mathbb{R}^n \mid E(u) \leq E(u^k)\}$$

- Subgradient method: move along vector from  $-N_S(u^k)$
- These are not necessarily descent directions

### Gradient Descent

- Definition
- Convergence analysis
- Line search
- Applications
- Conclusion

### Subgradient Method

- Definition
- Convergence Analysis
- Applications

### Gradient Projection

- Projections
- Definition
- Convergence Analysis
- Applications

### Proximal Gradient

- Proximal Operator
- Convergence Analysis
- Applications



- $E$  has a minimizer  $u^*$
- $E \in \mathcal{F}_G^0(\mathbb{R}^n)$ , i.e.,  $E$  is convex and Lipschitz continuous with constant  $G$ , and  $\text{dom}(E) = \mathbb{R}^n$

### Theorem: Bounded subdifferential

If  $E : \mathbb{R}^n \rightarrow \mathbb{R}$  is convex and Lipschitz continuous with constant  $G > 0$ , then this is equivalent to

$$\|g\|_2 \leq G, \quad \forall g \in \partial E(u), \forall u \in \mathbb{R}^n.$$

Proof: Board!

#### Gradient Descent

- Definition
- Convergence analysis
- Line search
- Applications
- Conclusion

#### Subgradient Method

- Definition
- Convergence Analysis
- Applications

#### Gradient Projection

- Projections
- Definition
- Convergence Analysis
- Applications

#### Proximal Gradient

- Proximal Operator
- Convergence Analysis
- Applications



## Gradient Descent

Definition  
 Convergence analysis  
 Line search  
 Applications  
 Conclusion

## Subgradient Method

Definition  
**Convergence Analysis**  
 Applications

## Gradient Projection

Projections  
 Definition  
 Convergence Analysis  
 Applications

## Proximal Gradient

Proximal Operator  
 Convergence Analysis  
 Applications

## Convergence analysis

- Consider distance to optimal set,  $u^+ = u - \tau g$ ,  $g \in \partial E(u)$

$$\begin{aligned}\|u^+ - u^*\|^2 &= \|u - \tau g - u^*\|^2 \\ &= \|u - u^*\|^2 - 2\tau \langle g, u - u^* \rangle + \tau^2 \|g\|^2 \\ &\leq \|u - u^*\|^2 - 2\tau (E(u) - E(u^*)) + \tau^2 \|g\|^2\end{aligned}$$

- Rearranging the above yields:

$$2\tau (E(u) - E(u^*)) \leq \|u - u^*\|^2 - \|u^+ - u^*\|^2 + \tau^2 \|g\|^2$$

- Set  $u^+ = u^k$ ,  $u^- = u^{k-1}$ ,  $\widehat{E}_N = \min_{0 \leq k \leq N} E(u^k)$ :

$$2 \left( \sum_{k=1}^N \tau_k \right) (\widehat{E}_N - E(u^*)) \leq \|u^0 - u^*\|^2 + \sum_{k=1}^N \tau_k^2 \|g^k\|^2$$



## Gradient Descent

Definition

Convergence analysis

Line search

Applications

Conclusion

## Subgradient Method

Definition

## Convergence Analysis

Applications

## Gradient Projection

Projections

Definition

Convergence Analysis

Applications

## Proximal Gradient

Proximal Operator

Convergence Analysis

Applications

## Convergence analysis for fixed step size

- For fixed step size  $\tau_k = \tau$  we have

$$\hat{E}_N - E(u^*) \leq \frac{\|u^0 - u^*\|^2}{2N\tau} + \frac{G^2\tau}{2}$$

- Does not guarantee convergence
- $\hat{E}_N$  is  $(G^2\tau/2)$ -suboptimal for large  $N$
- For step size  $\tau_k = \tau / \|g^k\|$  we have

$$\hat{E}_N - E(u^*) \leq \frac{G\|u^0 - u^*\|^2}{2N\tau} + \frac{G\tau}{2}$$

- Also does not guarantee convergence
- $\hat{E}_N$  is  $(G\tau/2)$ -suboptimal for large  $N$



## Diminishing step sizes

- Choose sequence  $\tau_k \rightarrow 0$ ,  $\sum_{k=1}^{\infty} \tau_k = \infty$
- Example: harmonic series  $\tau_k = 1/k$
- For non-constant steps we have the following bound

$$\hat{E}_N - E(u^*) \leq \frac{\|u^0 - u^*\|^2 + G^2 \sum_{k=1}^N \tau_k^2}{2 \sum_{k=1}^N \tau_k}$$

- For such a sequence it holds that

$$\frac{\sum_{k=1}^N \tau_k^2}{\sum_{k=1}^N \tau_k} \rightarrow 0, \quad \text{for } N \rightarrow \infty$$

- Thus  $\hat{E}_N$  converges to the optimal  $E(u^*)$  for  $N \rightarrow \infty$

### Gradient Descent

Definition

Convergence analysis

Line search

Applications

Conclusion

### Subgradient Method

Definition

**Convergence Analysis**

Applications

### Gradient Projection

Projections

Definition

Convergence Analysis

Applications

### Proximal Gradient

Proximal Operator

Convergence Analysis

Applications



## Polyak step size

- Recall the inequality we started with

$$\|u^+ - u^*\|^2 \leq \|u - u^*\|^2 - 2\tau(E(u) - E(u^*)) + \tau^2 \|g\|^2$$

- Right hand side is minimized for

$$\tau = \frac{E(u) - E(u^*)}{\|g\|^2}$$

- Plugging this back in yields

$$\|u^+ - u^*\|^2 \leq \|u - u^*\|^2 - \frac{(E(u) - E(u^*))^2}{\|g\|^2}$$

- A short calculation ( $\rightarrow$  board!) shows:

$$\hat{E}_N - E(u^*) \leq \frac{G \|u^0 - u^*\|}{\sqrt{N}}$$

### Gradient Descent

Definition

Convergence analysis

Line search

Applications

Conclusion

### Subgradient Method

Definition

### Convergence Analysis

Applications

### Gradient Projection

Projections

Definition

Convergence Analysis

Applications

### Proximal Gradient

Proximal Operator

Convergence Analysis

Applications



# Worst-case complexity

- Problem class: convex functions  $E : \mathbb{R}^n \rightarrow \mathbb{R}$
- First-order method:

$$u^{k+1} \in u^0 + \text{span}\{g^0, g^1, \dots, g^k\}, \quad g^k \in \partial E(u^k)$$

- Worst-case complexity:  $E(u^k) - E(u^*) = \mathcal{O}(1/\sqrt{k})^8$
- The subgradient method, which is amongst the simplest conceivable methods is optimal
- Indicates that the problem class of general convex functions is too complicated to be solved efficiently

## Gradient Descent

- Definition
- Convergence analysis
- Line search
- Applications
- Conclusion

## Subgradient Method

- Definition
- Convergence Analysis
- Applications

## Gradient Projection

- Projections
- Definition
- Convergence Analysis
- Applications

## Proximal Gradient

- Proximal Operator
- Convergence Analysis
- Applications



- Consider an image  $f \in \mathbb{R}^{Nc}$  with  $N$  pixels and  $c$  channels
- Many possible ways of defining the total variation for color images, one choice:

$$TV(f) = \varphi(Df)$$

- $D : \mathbb{R}^{Nc} \rightarrow \mathbb{R}^{2Nc}$  is the usual finite differencing matrix
- $\varphi(g) = \sum_{i=1}^N \|g_i\|_2$  is the sum of consecutive  $g_i \in \mathbb{R}^{2c}$
- It is non-differentiable, since  $\|\cdot\|_2$  is not differentiable at 0

## Gradient Descent

Definition  
 Convergence analysis  
 Line search  
 Applications  
 Conclusion

## Subgradient Method

Definition  
 Convergence Analysis  
 Applications

## Gradient Projection

Projections  
 Definition  
 Convergence Analysis  
 Applications

## Proximal Gradient

Proximal Operator  
 Convergence Analysis  
 Applications



# Subdifferential of the total variation

- The subdifferential follows from the chain rule

$$\partial TV(u) = (D^T \circ \partial\varphi \circ D)(u)$$

- The subdifferential of  $\varphi$  for  $g \in \mathbb{R}^{2Nc}$  is given as a product of  $N$  sets

$$\partial\varphi(g) = I_1 \times I_2 \times \dots \times I_N \subset \mathbb{R}^{2Nc}$$

- The individual sets are the subdifferentials of  $\|\cdot\|_2$

$$\mathbb{R}^{2c} \supset I_k = \begin{cases} \left\{ \frac{g_k}{\|g_k\|_2} \right\}, & \text{if } 0 \neq g_k, \\ B(0, 1), & \text{otherwise.} \end{cases}$$

## Gradient Descent

Definition  
 Convergence analysis  
 Line search  
 Applications  
 Conclusion

## Subgradient Method

Definition  
 Convergence Analysis  
Applications

## Gradient Projection

Projections  
 Definition  
 Convergence Analysis  
 Applications

## Proximal Gradient

Proximal Operator  
 Convergence Analysis  
 Applications



$$f \in \mathbb{R}^{Nc}$$



$$m \in \mathbb{R}^{Nc}$$

- Non-differentiable energy due to  $TV$  term

$$E(u) = \frac{\lambda}{2} \|m \cdot (u - f)\|^2 + TV(u)$$

- Subgradient can be easily computed:

$$g^k = \lambda(m \cdot (u^k - f)) + p^k, \text{ with } p^k \in \partial TV(u^k)$$

## Gradient Descent

Definition  
 Convergence analysis  
 Line search  
 Applications  
 Conclusion

## Subgradient Method

Definition  
 Convergence Analysis  
Applications

## Gradient Projection

Projections  
 Definition  
 Convergence Analysis  
 Applications

## Proximal Gradient

Proximal Operator  
 Convergence Analysis  
 Applications

Michael Moeller

Thomas Möllenhoff

Emanuel Laude



## Gradient Descent

Definition

Convergence analysis

Line search

Applications

Conclusion

## Subgradient Method

Definition

Convergence Analysis

Applications

## Gradient Projection

Projections

Definition

Convergence Analysis

Applications

## Proximal Gradient

Proximal Operator

Convergence Analysis

Applications

Michael Moeller

Thomas Möllenhoff

Emanuel Laude



## Gradient Descent

Definition

Convergence analysis

Line search

Applications

Conclusion

## Subgradient Method

Definition

Convergence Analysis

Applications

## Gradient Projection

Projections

Definition

Convergence Analysis

Applications

## Proximal Gradient

Proximal Operator

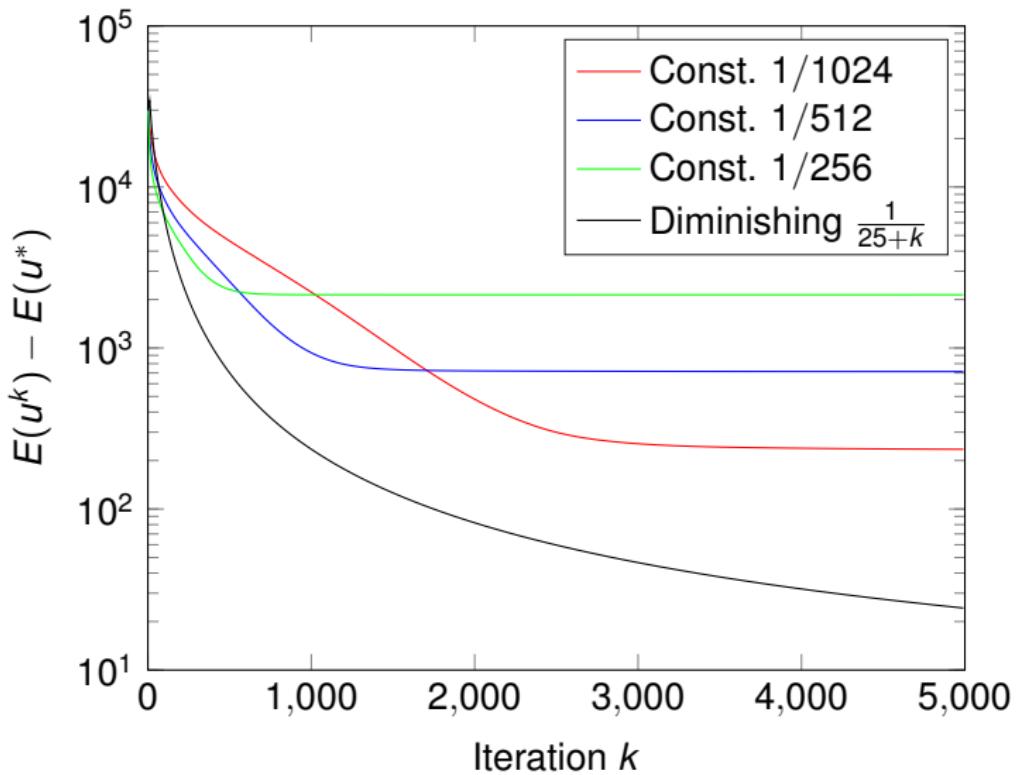
Convergence Analysis

Applications

# Numerical convergence results

Gradient Methods

Michael Moeller  
Thomas Möllenhoff  
Emanuel Laude



Gradient Descent

Definition  
Convergence analysis  
Line search  
Applications  
Conclusion

Subgradient Method

Definition  
Convergence Analysis  
Applications

Gradient Projection

Projections  
Definition  
Convergence Analysis  
Applications

Proximal Gradient

Proximal Operator  
Convergence Analysis  
Applications



- A robust model for image denoising is given by

$$E(u) = \lambda \|u - f\|_1 + TV(u)$$

- The  $\ell_1$ -data term is less sensitive to outliers than the previous quadratic data term
- Both data term and regularizer are non-smooth
- Getting a subgradient  $g^k$  is straightforward

$$g^k = \lambda \text{sign}(u^k - f) + p^k, \text{ with } p^k \in \partial TV(u^k)$$

## Gradient Descent

Definition

Convergence analysis

Line search

Applications

Conclusion

## Subgradient Method

Definition

Convergence Analysis

Applications

## Gradient Projection

Projections

Definition

Convergence Analysis

Applications

## Proximal Gradient

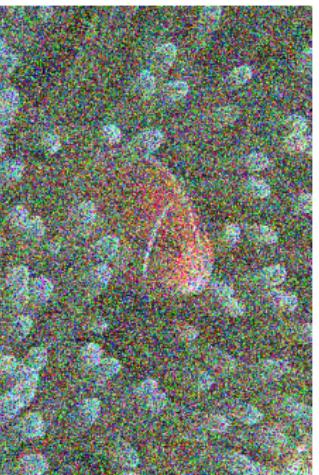
Proximal Operator

Convergence Analysis

Applications



Original



Noisy input

 $TV - \ell_1$  denoised

## Gradient Descent

Definition

Convergence analysis

Line search

Applications

Conclusion

## Subgradient Method

Definition

Convergence Analysis

Applications

## Gradient Projection

Projections

Definition

Convergence Analysis

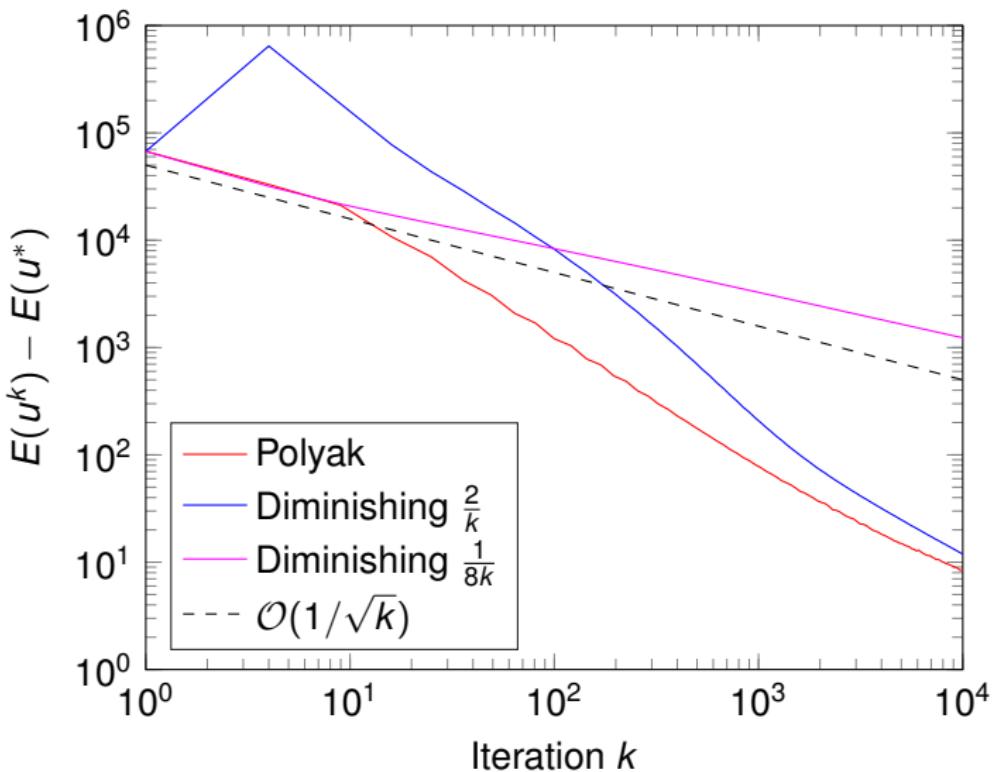
Applications

## Proximal Gradient

Proximal Operator

Convergence Analysis

Applications



## Gradient Descent

- Definition
- Convergence analysis
- Line search
- Applications
- Conclusion

## Subgradient Method

- Definition
- Convergence Analysis
- Applications

## Gradient Projection

- Projections
- Definition
- Convergence Analysis
- Applications

## Proximal Gradient

- Proximal Operator
- Convergence Analysis
- Applications



- Why care about subgradient method?
  - Simple
  - Each iteration fast
  - Low memory requirements
- We covered only the absolute basics
- Many extensions to the subgradient method exist (acceleration, constraints, stochastic, ...)
- Next week: solving constrained problems, duality

## Gradient Descent

Definition  
Convergence analysis  
Line search  
Applications  
Conclusion

## Subgradient Method

Definition  
Convergence Analysis  
**Applications**

## Gradient Projection

Projections  
Definition  
Convergence Analysis  
Applications

## Proximal Gradient

Proximal Operator  
Convergence Analysis  
Applications

Michael Moeller  
Thomas Möllenhoff  
Emanuel Laude



# Gradient Projection

## Gradient Descent

Definition  
Convergence analysis  
Line search  
Applications  
Conclusion

## Subgradient Method

Definition  
Convergence Analysis  
Applications

## Gradient Projection

Projections  
Definition  
Convergence Analysis  
Applications

## Proximal Gradient

Proximal Operator  
Convergence Analysis  
Applications



$$u^* \in \arg \min_{u \in \mathbb{R}^n} E(u),$$

for  $E : \mathbb{R}^n \rightarrow \mathbb{R} \cup \{\infty\}$  proper, closed, convex.

## Gradient descent:

- $\text{dom } E = \mathbb{R}^n$
- For  $E \in \mathcal{F}_L^{1,1}(\mathbb{R}^n)$  energy convergence in  $\mathcal{O}(1/k)$
- For  $E \in \mathcal{S}_{m,L}^{1,1}(\mathbb{R}^n)$  energy and iterate convergence in  $\mathcal{O}(c^k)$

## Subgradient descent:

- $\text{dom}(E) = \mathbb{R}^n$
- Applicable to any Lipschitz-continuous convex energy
- Usually rather slow

**Gradient projection:** Generalizes gradient descent to arbitrary (nonempty, closed, convex)  $\text{dom}(E)$ .

## Gradient Descent

- Definition
- Convergence analysis
- Line search
- Applications
- Conclusion

## Subgradient Method

- Definition
- Convergence Analysis
- Applications

## Gradient Projection

- Projections
- Definition
- Convergence Analysis
- Applications

## Proximal Gradient

- Proximal Operator
- Convergence Analysis
- Applications



# Gradient projection

Type of problem:

$$u^* \in \arg \min_{u \in C} E(u), \quad (1)$$

for  $E \in \mathcal{F}_L^{1,1}(\mathbb{R}^n)$ , and a nonempty, closed, convex set  $C$ .

What is the *projection* onto the set  $C$ ?

## Definition: Projection

For a (nonempty) closed convex set  $C \subset \mathbb{R}^n$ ,

$$\pi_C(v) = \operatorname{argmin}_{u \in C} \|u - v\|_2^2$$

is called the projection of  $v$  onto the set  $C$ .

## Gradient Descent

Definition

Convergence analysis

Line search

Applications

Conclusion

## Subgradient Method

Definition

Convergence Analysis

Applications

## Gradient Projection

Projections

Definition

Convergence Analysis

Applications

## Proximal Gradient

Proximal Operator

Convergence Analysis

Applications



# Projections

## Existence and Uniqueness of the Projection

For any (nonempty) closed convex set  $C \subset \mathbb{R}^n$  and any  $v$  the projection  $\pi_C(v)$  exists and is single valued.

*Proof: Board.*

*Abuse of notation: Although  $\pi_C(v)$  is (by definition) a set, we also identify  $\pi_C(v)$  with the single element in the set.*

### Gradient Descent

- Definition
- Convergence analysis
- Line search
- Applications
- Conclusion

### Subgradient Method

- Definition
- Convergence Analysis
- Applications

### Gradient Projection

- Projections
- Definition
- Convergence Analysis
- Applications

### Proximal Gradient

- Proximal Operator
- Convergence Analysis
- Applications



What is the projection of  $v \in \mathbb{R}^n$  onto

- $C = \{u \in \mathbb{R}^n \mid \|u\|_2 \leq 1\}$ ?
- $C = \{u \in \mathbb{R}^n \mid \|u\|_\infty := \max_i |u_i| \leq 1\}$ ?
- $C = \{u \in \mathbb{R}^n \mid u_i \in [a, b]\}$ ?
- $C = \{u \in \mathbb{R}^n \mid u_i \geq a\}$ ?

### Gradient Descent

- Definition
- Convergence analysis
- Line search
- Applications
- Conclusion

### Subgradient Method

- Definition
- Convergence Analysis
- Applications

### Gradient Projection

- Projections**
- Definition
- Convergence Analysis
- Applications

### Proximal Gradient

- Proximal Operator
- Convergence Analysis
- Applications



# Properties of the projection

## Firm Nonexpansiveness

The projection  $\pi_C$  onto a nonempty closed convex set  $C \subset \mathbb{R}^n$  is *firmly nonexpansive* or *co-coercive*, i.e. it meets

$$\langle u - v, \pi_C(u) - \pi_C(v) \rangle \geq \|\pi_C(u) - \pi_C(v)\|^2 \quad \forall u, v \in \mathbb{R}^n.$$

By Cauchy-Schwarz, this implies the nonexpansiveness

$$\|\pi_C(u) - \pi_C(v)\| \leq \|u - v\|, \quad \forall u, v \in \mathbb{R}^n.$$

*Proof: Board*

## Gradient Descent

- Definition
- Convergence analysis
- Line search
- Applications
- Conclusion

## Subgradient Method

- Definition
- Convergence Analysis
- Applications

## Gradient Projection

- Projections
- Definition
- Convergence Analysis
- Applications

## Proximal Gradient

- Proximal Operator
- Convergence Analysis
- Applications



## Idea of gradient projection

Consider a problem

$$u^* \in \arg \min_{u \in C} E(u), \quad (2)$$

for  $E \in \mathcal{F}_L^{1,1}(\mathbb{R}^n)$ , and a nonempty, closed, convex set  $C$ .

We know how gradient descent works, but updating  
 $u^{k+1} = u^k - \tau^k \nabla E(u^k)$  may lead to  $u^{k+1} \notin C$ .

**Idea: Project every iteration back to the feasible set, i.e.**

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

### Gradient Descent

Definition

Convergence analysis

Line search

Applications

Conclusion

### Subgradient Method

Definition

Convergence Analysis

Applications

### Gradient Projection

Projections

Definition

Convergence Analysis

Applications

### Proximal Gradient

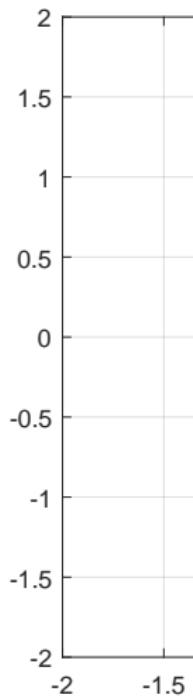
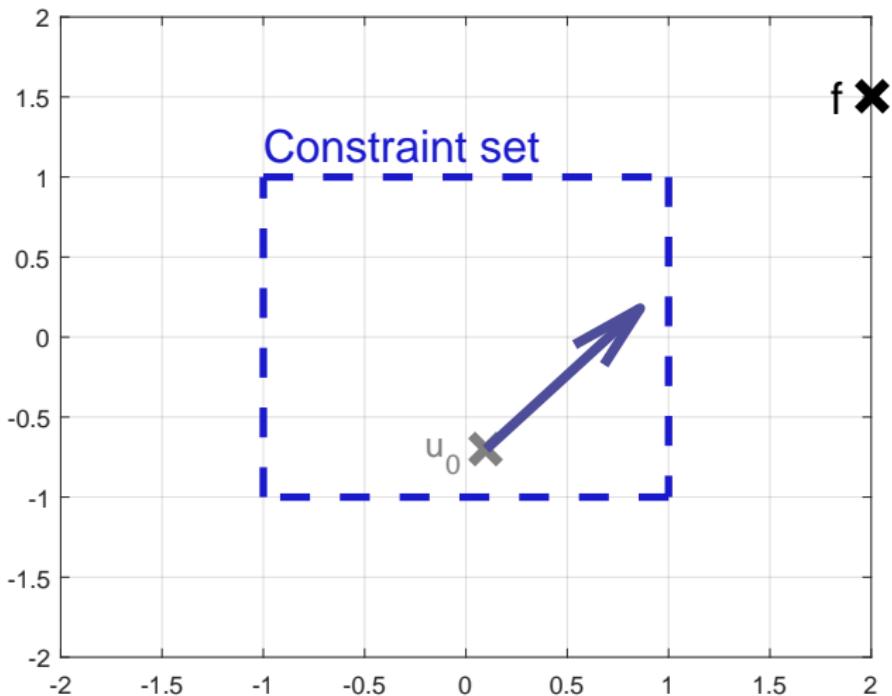
Proximal Operator

Convergence Analysis

Applications

## Idea of gradient projection

Toy problem  $\min_{|u_i| \leq 1} \|u - f\|_2^2$





# Gradient projection algorithm

Let  $C \subset \mathbb{R}^n$  be a nonempty closed convex set and let  $E : \mathbb{R}^n \rightarrow \mathbb{R}$  be  $C^1(\mathbb{R}^n)$ . Then, for  $u^0 \in C$

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

is called the *gradient projection* algorithm.

## Gradient Descent

- Definition
- Convergence analysis
- Line search
- Applications
- Conclusion

## Subgradient Method

- Definition
- Convergence Analysis
- Applications

## Gradient Projection

- Projections
- Definition
- Convergence Analysis
- Applications

## Proximal Gradient

- Proximal Operator
- Convergence Analysis
- Applications

*When, how, why, and for which  $E$  and  $\tau$  does it work?*

Remember: Gradient descent

- $E \in \mathcal{S}_{m,L}^{1,1}(\mathbb{R}^n)$  leads to a convergence of  $\mathcal{O}(c^k)$ ,  $c < 1$ .
- $E \in \mathcal{F}_L^{1,1}(\mathbb{R}^n)$  leads to a convergence of  $\mathcal{O}(1/k)$ .

Same convergence for gradient projection?



## Gradient projection algorithm

First:  $E \in \mathcal{S}_{m,L}^{1,1}(\mathbb{R}^n)$ . Convergence proof of gradient descent

$$\begin{aligned}
 \|u^{k+1} - u^*\|^2 &= \|u^{k+1} - u^k + u^k - u^*\|^2 \\
 &= \|u^k - u^*\|^2 + \underbrace{2\langle u^{k+1} - u^k, u^k - u^* \rangle}_{\text{bound from above by something negative} \cdot \|u^k - u^*\|^2} + \|u^{k+1} - u^k\|^2 \\
 &\leq c\|u^k - u^*\|^2
 \end{aligned}$$

To carry out a similar proof we need an upper bound on

$$\langle u^{k+1} - u^k, u^k - u^* \rangle + \|u^{k+1} - u^k\|^2$$

### Gradient Descent

Definition

Convergence analysis

Line search

Applications

Conclusion

### Subgradient Method

Definition

Convergence Analysis

Applications

### Gradient Projection

Projections

Definition

**Convergence Analysis**

Applications

### Proximal Gradient

Proximal Operator

Convergence Analysis

Applications



# Gradient projection algorithm

We will need (at least) three things:

- ①  $E$  is  $L$ -smooth, i.e. for all  $u, v$  it holds that

$$E(v) - E(u) - \langle \nabla E(u), v - u \rangle - \frac{L}{2} \|v - u\|^2 \leq 0$$

- ②  $E$  is  $m$ -strongly convex, i.e. for all  $u, v$  it holds that

$$E(v) - E(u) - \langle \nabla E(u), v - u \rangle - \frac{m}{2} \|v - u\|^2 \geq 0$$

- ③ Gradient projection equation:

$$0 = u^{k+1} - u^k + \tau \nabla E(u^k) + p^{k+1} \quad p^{k+1} \in \partial \iota_C(u^{k+1})$$

*Continue on the board.*

## Gradient Descent

Definition
Convergence analysis
Line search
Applications
Conclusion

## Subgradient Method

Definition
Convergence Analysis
Applications

## Gradient Projection

Projections
Definition
Convergence Analysis
Applications

## Proximal Gradient

Proximal Operator
Convergence Analysis
Applications



# Gradient projection algorithm

## Gradient Projection Estimate

For  $E \in \mathcal{S}_{m,L}^{1,1}(\mathbb{R}^n)$ ,  $\tau = 1/L$ , and  $u \in C$  arbitrary it holds that

$$0 \leq E(u) - E(u^{k+1}) - \frac{L}{2} \|u - u^{k+1}\|^2 + \frac{L-m}{2} \|u - u^k\|^2$$

## Corollary

In the above setting it holds that

$$0 \leq E(u^k) - E(u^{k+1}) - \frac{L}{2} \|u^{k+1} - u^k\|^2$$

$$0 \leq -\frac{L}{2} \|u^* - u^{k+1}\|^2 + \frac{L-m}{2} \|u^* - u^k\|^2$$

### Gradient Descent

- Definition
- Convergence analysis
- Line search
- Applications
- Conclusion

### Subgradient Method

- Definition
- Convergence Analysis
- Applications

### Gradient Projection

- Projections
- Definition
- Convergence Analysis
- Applications

### Proximal Gradient

- Proximal Operator
- Convergence Analysis
- Applications



# Gradient projection algorithm

## Convergence of gradient projection for $E \in \mathcal{S}_{m,L}^{1,1}(\mathbb{R}^n)$

For  $E \in \mathcal{S}_{m,L}^{1,1}(\mathbb{R}^n)$  the gradient projection algorithm with constant stepsize  $\tau = \frac{1}{L}$  converges with

$$\|u^k - u^*\|^2 \leq \left(1 - \frac{m}{L}\right)^k \|u^0 - u^*\|^2.$$

What happens if we do not have strong convexity?

### Gradient Descent

Definition

Convergence analysis

Line search

Applications

Conclusion

### Subgradient Method

Definition

Convergence Analysis

Applications

### Gradient Projection

Projections

Definition

### Convergence Analysis

Applications

### Proximal Gradient

Proximal Operator

Convergence Analysis

Applications



## Gradient projection algorithm

Our gradient projection estimate for  $E \in \mathcal{S}_{m,L}^{1,1}(\mathbb{R}^n)$  with  $m = 0$  and  $u \in C$  arbitrary, yields

$$0 \leq E(u) - E(u^{k+1}) - \frac{L}{2} \|u - u^{k+1}\|^2 + \frac{L}{2} \|u - u^k\|^2$$

Picking  $u = u^*$  we find

$$\begin{aligned} E(u^{k+1}) - E(u^*) &\leq \frac{L}{2} \|u - u^k\|^2 - \frac{L}{2} \|u - u^{k+1}\|^2 \\ \Rightarrow \sum_{k=0}^{K-1} (E(u^{k+1}) - E(u^*)) &\leq \frac{L}{2} \|u - u^0\|^2 - \frac{L}{2} \|u - u^K\|^2 \end{aligned}$$

Similar to the gradient descent case, the monotonicity of the energy yields the convergence.

### Gradient Descent

Definition

Convergence analysis

Line search

Applications

Conclusion

### Subgradient Method

Definition

Convergence Analysis

Applications

### Gradient Projection

Projections

Definition

Convergence Analysis

Applications

### Proximal Gradient

Proximal Operator

Convergence Analysis

Applications



# Gradient projection algorithm

Did we really show convergence of  $E(u)$  or did we implicitly make an additional assumption?

## Convergence of gradient projection for $E \in \mathcal{F}_L^{1,1}(\mathbb{R}^n)$

Let  $E \in \mathcal{F}_L^{1,1}(\mathbb{R}^n)$  have a global minimizer  $u^*$ . Then the gradient projection algorithm with constant stepsize  $\tau = \frac{1}{L}$  yields

$$E(u^k) - E(u^*) \leq \frac{L}{2k} \|u^0 - u^*\|^2.$$

### Gradient Descent

- Definition
- Convergence analysis
- Line search
- Applications
- Conclusion

### Subgradient Method

- Definition
- Convergence Analysis
- Applications

### Gradient Projection

- Projections
- Definition
- Convergence Analysis
- Applications

### Proximal Gradient

- Proximal Operator
- Convergence Analysis
- Applications



## Gradient Descent

Definition

Convergence analysis

Line search

Applications

Conclusion

## Subgradient Method

Definition

Convergence Analysis

Applications

## Gradient Projection

Projections

Definition

Convergence Analysis

Applications

## Proximal Gradient

Proximal Operator

Convergence Analysis

Applications

# Example Application: Solving a SUDOKU

Find the missing numbers such that each block, each row, and each column contains each number 1– 4 only once!

2			3
1	3		
		3	2
	2	4	

2	4	1	3
1	3	2	4
4	1	3	2
3	2	4	1

How can we do this with convex optimization?

Ideas: Identify the constraints with



## Gradient Descent

Definition  
 Convergence analysis  
 Line search  
 Applications  
 Conclusion

## Subgradient Method

Definition  
 Convergence Analysis  
 Applications

## Gradient Projection

Projections  
 Definition  
 Convergence Analysis  
 Applications

## Proximal Gradient

Proximal Operator  
 Convergence Analysis  
 Applications

## Example Application: Solving a SUDOKU

In the  $4 \times 4$  case we look for a matrix  $u \in \{1, 2, 3, 4\}^{4 \times 4}$  such that  $u_{i,j} = f_{i,j}$  for those entries  $f_{i,j}$  which are given.

Reformulation: We look for a matrix  $u \in \{0, 1\}^{4 \times 4 \times 4}$ , where  $u_{i,j,k} = 1$  means  $u_{i,j} = k$ .

Rule	Implication
One number for each blank spot	$\sum_k u_{i,j,k} = 1 \quad \forall i, j$
Respect given entries	$u_{i,j,k} = 1 \text{ if } f_{i,j} = k$
Numbers occur in a row once	$\sum_j u_{i,j,k} = 1 \quad \forall i, k$
Numbers occur in a column once	$\sum_i u_{i,j,k} = 1 \quad \forall j, k$
Numbers occur in a block once	$\sum_{(i,j) \in B_l} u_{i,j,k} = 1 \quad \forall B_l, k$

Find  $u$  with  $u_{i,j,k} \in \{0, 1\}$  subject to the above constraints!



## Example Application: Solving a SUDOKU

All constraints are linear, i.e. can be expressed as  $A\vec{u} = \vec{1}$ .

### SUDOKU rules in matrix form

The scalar product with all variants of the following vectors needs to be one.

$$\begin{pmatrix} 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}$$

Only one number from 1-4 should be selected

$$\begin{pmatrix} 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}$$

In each block each number may only appear once

$$\begin{pmatrix} 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \end{pmatrix}$$

In each column each number may only appear once

$$\begin{pmatrix} 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}$$

In each row each number may only appear once

Find  $\mathbf{u}$  with  $u_{i,j,k} \in \{0, 1\}$  is a nonconvex constraint!

**Convex relaxation:** Use the smallest convex set that contains the nonconvex one,  $u_{i,j,k} \in [0, 1]$ .

If the result meets  $u_{i,j,k} \in \{0, 1\}$ , we solved the nonconvex problem.

### Gradient Descent

- Definition
- Convergence analysis
- Line search
- Applications
- Conclusion

### Subgradient Method

- Definition
- Convergence Analysis
- Applications

### Gradient Projection

- Projections
- Definition
- Convergence Analysis
- Applications

### Proximal Gradient

- Proximal Operator
- Convergence Analysis
- Applications



## Example Application: Solving a SUDOKU

Nice thing for SUDOKU: There exists a solution to  $A\vec{u} = \vec{1}$ !

This means we may solve

$$\hat{\mathbf{u}} \in \operatorname{argmin}_{\mathbf{u}_{i,j,k} \in [0,1]} \|A\vec{\mathbf{u}} - \vec{1}\|_2^2$$

Hope that  $\hat{\mathbf{u}}_{i,j,k} \in \{0, 1\}$  in which case we solved the SUDOKU!

Remarks:

- Exact recovery guarantees (when is  $\hat{\mathbf{u}}_{i,j,k} \in \{0, 1\}$ ) are an active field of research.
- Similar constructions can be done for many computer vision problems! Look for *labeling problems, segmentation, graph cuts, or functional lifting*.

### Gradient Descent

Definition

Convergence analysis

Line search

Applications

Conclusion

### Subgradient Method

Definition

Convergence Analysis

Applications

### Gradient Projection

Projections

Definition

Convergence Analysis

Applications

### Proximal Gradient

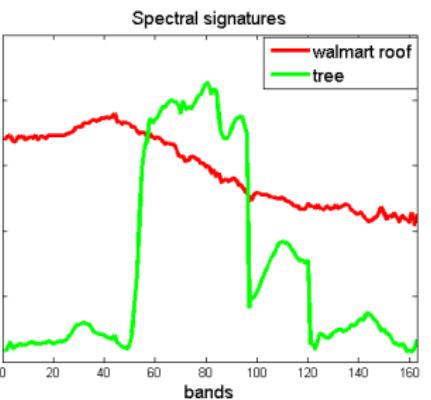
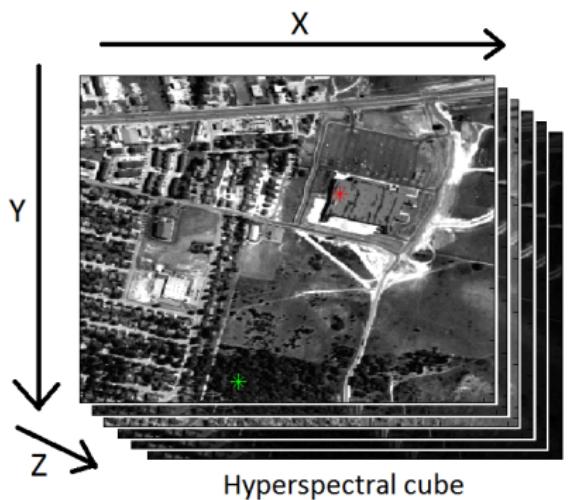
Proximal Operator

Convergence Analysis

Applications



## Hyperspectral imagery



**z-direction:** Material specific reflected energy depending on the wavelength of the incoming light

### Gradient Descent

- Definition
- Convergence analysis
- Line search
- Applications
- Conclusion

### Subgradient Method

- Definition
- Convergence Analysis
- Applications

### Gradient Projection

- Projections
- Definition
- Convergence Analysis
- Applications

### Proximal Gradient

- Proximal Operator
- Convergence Analysis
- Applications



Measured signals  $f$

Find decomposition  $f = Au + n$

Dictionary of materials  $A$ , mixing coefficients  $u$  (sparse) and noise  $n$

### Gradient Descent

- Definition
- Convergence analysis
- Line search
- Applications
- Conclusion

### Subgradient Method

- Definition
- Convergence Analysis
- Applications

### Gradient Projection

- Projections
- Definition
- Convergence Analysis
- Applications

### Proximal Gradient

- Proximal Operator
- Convergence Analysis
- Applications



## Example application: Unmixing and sparse recovery

General setup: Minimize a data fidelity term  $H_f(v)$  which is  $L$ -smooth, such that  $v$  can be represented in a dictionary  $A$ , i.e.  $v = Au$ , and the representing coefficients  $u$  are sparse.

Energy minimization approach:

$$\min_u H_f(Au) + \alpha \|u\|_1.$$

Can we apply gradient descent/ gradient projection?

Not directly, but the problem is equivalent to

$$\min_u H_f(A(u_1 - u_2)) + \alpha \langle u_1, \mathbf{1} \rangle + \alpha \langle u_2, \mathbf{1} \rangle, \quad u_1 \geq 0, u_2 \geq 0!$$

### Gradient Descent

Definition

Convergence analysis

Line search

Applications

Conclusion

### Subgradient Method

Definition

Convergence Analysis

Applications

### Gradient Projection

Projections

Definition

Convergence Analysis

Applications

### Proximal Gradient

Proximal Operator

Convergence Analysis

Applications



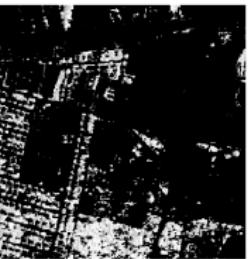
color image illustration



endmember "road"



endmember "roof"



endmember "trees"

## Gradient Descent

Definition

Convergence analysis

Line search

Applications

Conclusion

## Subgradient Method

Definition

Convergence Analysis

Applications

## Gradient Projection

Projections

Definition

Convergence Analysis

Applications

## Proximal Gradient

Proximal Operator

Convergence Analysis

Applications



# Proximal Gradient

## Gradient Descent

- Definition
- Convergence analysis
- Line search
- Applications
- Conclusion

## Subgradient Method

- Definition
- Convergence Analysis
- Applications

## Gradient Projection

- Projections
- Definition
- Convergence Analysis
- Applications

## Proximal Gradient

- Proximal Operator
- Convergence Analysis
- Applications



## What has happened so far...?

- Convex functions and sets
- Existence and uniqueness of minimizers
- Smooth optimization: gradient descent
- Non-smooth optimization: subgradient descent
- Dealing with constraints: projected gradient descent
- Last two lectures: theory of convex duality
- TV minimization: dual problem has favorable structure

$$\min_u \frac{\lambda}{2} \|u - f\|^2 + \|Du\|_{2,1}$$

↑↓

$$\min_{\|q\|_{2,\infty} \leq 1} \frac{1}{2\lambda} \|D^* q\|^2 - \langle q, Df \rangle$$

### Gradient Descent

- Definition
- Convergence analysis
- Line search
- Applications
- Conclusion

### Subgradient Method

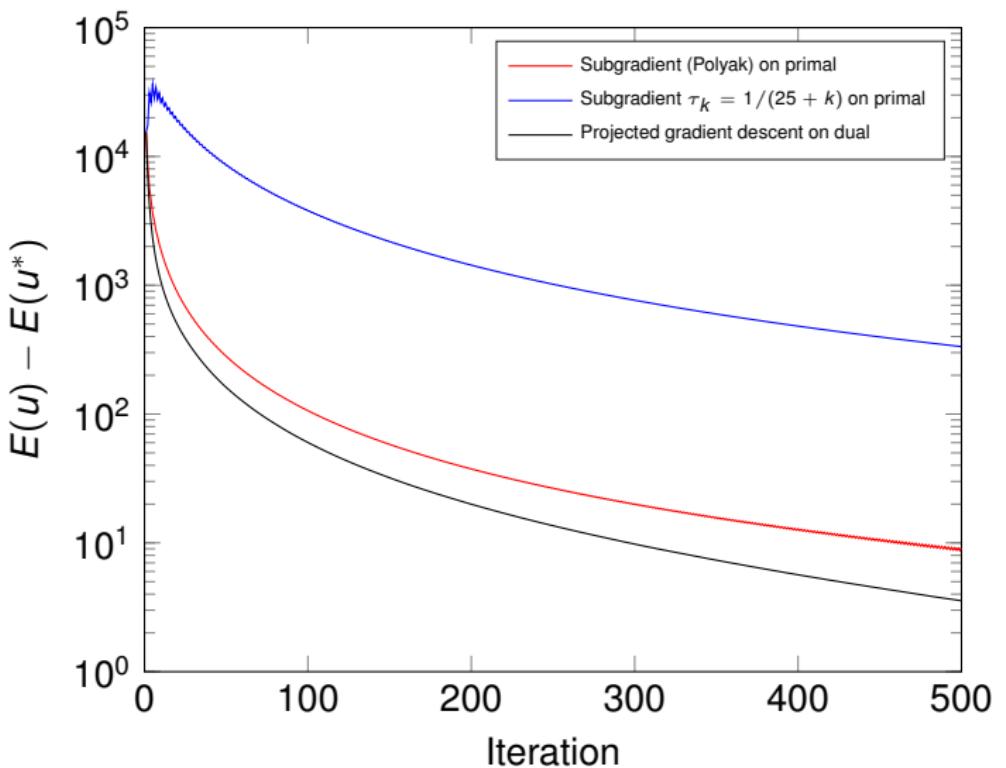
- Definition
- Convergence Analysis
- Applications

### Gradient Projection

- Projections
- Definition
- Convergence Analysis
- Applications

### Proximal Gradient

- Proximal Operator
- Convergence Analysis
- Applications



## Gradient Descent

- Definition
- Convergence analysis
- Line search
- Applications
- Conclusion

## Subgradient Method

- Definition
- Convergence Analysis
- Applications

## Gradient Projection

- Projections
- Definition
- Convergence Analysis
- Applications

## Proximal Gradient

- Proximal Operator
- Convergence Analysis
- Applications



- Recall gradient projection method for finding minimizer of

$$\operatorname{argmin}_{u \in \mathbb{R}^n} E(u) + \iota_C(u), \quad \iota_C(u) := \begin{cases} 0, & \text{if } u \in C, \\ \infty, & \text{if } u \notin C. \end{cases}$$

- Idea: replace  $\iota_C$  by any closed, proper, convex function
- It turns out the gradient projection algorithm can be generalized to such a setting
- Today:** an algorithm for minimizing the sum of one smooth and one (possibly non-smooth) convex function

$$\operatorname{argmin}_{u \in \mathbb{R}^n} \underbrace{G(u)}_{\text{convex}} + \underbrace{F(u)}_{L\text{-smooth + convex}}$$

## Gradient Descent

Definition  
 Convergence analysis  
 Line search  
 Applications  
 Conclusion

## Subgradient Method

Definition  
 Convergence Analysis  
 Applications

## Gradient Projection

Projections  
 Definition  
 Convergence Analysis  
 Applications

## Proximal Gradient

Proximal Operator  
 Convergence Analysis  
 Applications



# Key ingredient: the proximal operator

## Definition

Given a closed, proper, convex function  $E : \mathbb{R}^n \rightarrow \mathbb{R} \cup \{\infty\}$ , the mapping  $\text{prox}_E : \mathbb{R}^n \rightarrow \mathbb{R}^n$  defined as

$$\text{prox}_E(v) := \underset{u \in \mathbb{R}^n}{\operatorname{argmin}} E(u) + \frac{1}{2} \|u - v\|^2$$

is called the *proximal operator* or *proximal mapping* of  $E$ .

- Trade-off between minimizing  $E$  and staying close to  $v$
- For proper, closed, convex  $E$ ,  $\text{prox}_E(v)$  exists and is unique for all  $v$
- **Existence:**  $E(u) + (1/2) \|u - v\|^2$  is closed and has bounded sublevel sets
- **Uniqueness:**  $E(u) + (1/2) \|u - v\|^2$  is strictly (and also strongly) convex

## Gradient Descent

Definition  
 Convergence analysis  
 Line search  
 Applications  
 Conclusion

## Subgradient Method

Definition  
 Convergence Analysis  
 Applications

## Gradient Projection

Projections  
 Definition  
 Convergence Analysis  
 Applications

## Proximal Gradient

Proximal Operator  
 Convergence Analysis  
 Applications



## Properties of proximal operator

- We will often encounter the proximal operator of the scaled function  $\tau E : \mathbb{R}^n \rightarrow \mathbb{R} \cup \{\infty\}$

$$\text{prox}_{\tau E}(v) = \operatorname{argmin}_{u \in \mathbb{R}^n} E(u) + \frac{1}{2\tau} \|u - v\|^2$$

- Special case of projection: for closed, nonempty, convex set  $C \subset \mathbb{R}^n$ , it holds

$$\text{prox}_{\iota_C}(v) = \pi_C(v)$$

- Separable sum: for  $f(x, y) = \varphi(x) + \psi(y)$  we have

$$\text{prox}_f(v, w) = (\text{prox}_\varphi(v), \text{prox}_\psi(w))$$

- Non-separable sum: difficult, in some cases<sup>9</sup> it holds

$$\text{prox}_{f+g}(u) = \text{prox}_f(\text{prox}_g(u))$$

- Orthogonal transform:  $f(x) = \varphi(Qx)$ ,  $QQ^T = Q^TQ = I$

$$\text{prox}_f(v) = Q^T \text{prox}_\varphi(Qv)$$

<sup>9</sup>Yu, Yao-Liang. On decomposing the proximal map. NIPS 2013

### Gradient Descent

Definition
Convergence analysis
Line search
Applications
Conclusion

### Subgradient Method

Definition
Convergence Analysis
Applications

### Gradient Projection

Projections
Definition
Convergence Analysis
Applications

### Proximal Gradient

Proximal Operator
Convergence Analysis
Applications
updated 18.05.2016



## Gradient Descent

Definition  
 Convergence analysis  
 Line search  
 Applications  
 Conclusion

## Subgradient Method

Definition  
 Convergence Analysis  
 Applications

## Gradient Projection

Projections  
 Definition  
 Convergence Analysis  
 Applications

## Proximal Gradient

Proximal Operator  
 Convergence Analysis  
 Applications

# Properties of the proximal operator

## Firm nonexpansiveness of proximal mapping

For proper, closed, convex  $E : \mathbb{R}^n \rightarrow \mathbb{R} \cup \{\infty\}$ , the proximal operator  $\text{prox}_E$  is *firmly nonexpansive* or *1-co-coercive*, i.e. it meets

$$\langle u - v, \text{prox}_E(u) - \text{prox}_E(v) \rangle \geq \|\text{prox}_E(u) - \text{prox}_E(v)\|^2,$$

for all  $u, v \in \mathbb{R}^n$ . By Cauchy-Schwarz, this implies the nonexpansiveness (Lipschitz continuity with constant 1)

$$\|\text{prox}_E(u) - \text{prox}_E(v)\| \leq \|u - v\|, \quad \forall u, v \in \mathbb{R}^n.$$

*Proof: Board!*

- Remark: For strongly convex  $E$ ,  $\text{prox}_E$  is a contraction (Lipschitz with constant  $< 1$ )



- Quadratic functions ( $A \succeq 0$ , symmetric)

$$f(x) = \frac{1}{2} \langle x, Ax \rangle + \langle b, x \rangle + c, \quad \text{prox}_{\tau f}(v) = (I + \tau A)^{-1}(v - \tau b)$$

- Euclidean norm

$$f(x) = \|x\|, \quad \text{prox}_{\tau f}(v) = \begin{cases} (1 - \tau / \|v\|)v & \text{if } \|v\| \geq \tau \\ 0 & \text{otherwise.} \end{cases}$$

- $\ell_1$ -norm (cf. exercise sheet 4), “soft thresholding”

$$f(x) = \|x\|_1, \quad (\text{prox}_{\tau f}(v))_i = \begin{cases} v_i + \tau & \text{if } v_i < -\tau \\ 0 & \text{if } |v_i| \leq \tau \\ v_i - \tau & \text{if } v_i > \tau. \end{cases}$$

## Gradient Descent

Definition  
 Convergence analysis  
 Line search  
 Applications  
 Conclusion

## Subgradient Method

Definition  
 Convergence Analysis  
 Applications

## Gradient Projection

Projections  
 Definition  
 Convergence Analysis  
 Applications

## Proximal Gradient

Proximal Operator  
 Convergence Analysis  
 Applications



# Moreau decomposition

## Theorem: Moreau decomposition

For closed, proper, convex  $E : \mathbb{R}^n \rightarrow \mathbb{R} \cup \{\infty\}$  the following decomposition holds:

$$u = \text{prox}_E(u) + \text{prox}_{E^*}(u), \quad \forall u \in \mathbb{R}^n.$$

### *Proof: Board!*

- Generalizes decomposition result from linear algebra for subspace  $V \subset \mathbb{R}^n$ :

$$u = \Pi_V(u) + \Pi_{V^\perp}(u)$$

- $V^\perp$  denotes the orthogonal complement:

$$V^\perp = \{u \in \mathbb{R}^n \mid \langle u, v \rangle = 0, \forall v \in V\}$$

- Extended Moreau decomposition for  $\tau > 0$

$$u = \text{prox}_{\tau E}(u) + \tau \text{prox}_{(1/\tau)E^*}(u/\tau), \quad \forall u \in \mathbb{R}^n.$$

### Gradient Descent

- Definition
- Convergence analysis
- Line search
- Applications
- Conclusion

### Subgradient Method

- Definition
- Convergence Analysis
- Applications

### Gradient Projection

- Projections
- Definition
- Convergence Analysis
- Applications

### Proximal Gradient

- Proximal Operator
- Convergence Analysis
- Applications



## Gradient Descent

Definition

Convergence analysis

Line search

Applications

Conclusion

## Subgradient Method

Definition

Convergence Analysis

Applications

## Gradient Projection

Projections

Definition

Convergence Analysis

Applications

## Proximal Gradient

## Proximal Operator

Convergence Analysis

Applications

# Interpretation: Resolvent operator

- Let's start with the definition of proximal operator

$$u^* = \operatorname{argmin}_u E(u) + \frac{1}{2\tau} \|u - v\|^2$$

- Assuming smooth  $E$ , optimality conditions of optimization problem are given by

$$\begin{aligned} 0 &= \nabla E(u^*) + \frac{u^* - v}{\tau} \\ \Leftrightarrow u^* &= \underbrace{(I + \tau \nabla E)^{-1}}_{\text{resolvent operator}}(v) \end{aligned}$$

- Later in the semester: for non-smooth functions the resolvent is formulated using the subdifferential

$$\operatorname{prox}_{\tau E}(v) = (I + \tau \partial E)^{-1}(v)$$



- Consider a time-continuous gradient descent for differentiable  $h : \mathbb{R}^n \rightarrow \mathbb{R}$

$$\frac{d}{dt}x(t) = -\nabla h(x(t))$$

- Explicit (Forward) Euler time discretization:

$$x^{t+1} = x^t - \tau \nabla h(x^t) = (I - \tau \nabla h)x^t$$

- Implicit (Backward) Euler time discretization:

$$x^{t+1} = x^t - \tau \nabla h(x^{t+1}) = (I + \tau \nabla h)^{-1}(x^t) = \text{prox}_{\tau h}(x^t)$$

## Gradient Descent

Definition  
 Convergence analysis  
 Line search  
 Applications  
 Conclusion

## Subgradient Method

Definition  
 Convergence Analysis  
 Applications

## Gradient Projection

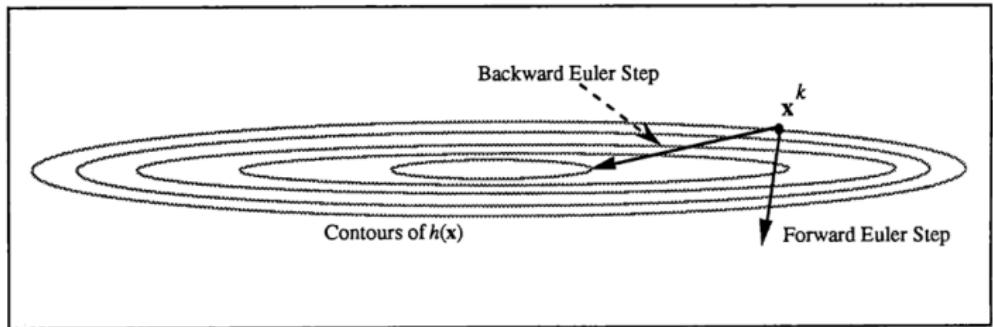
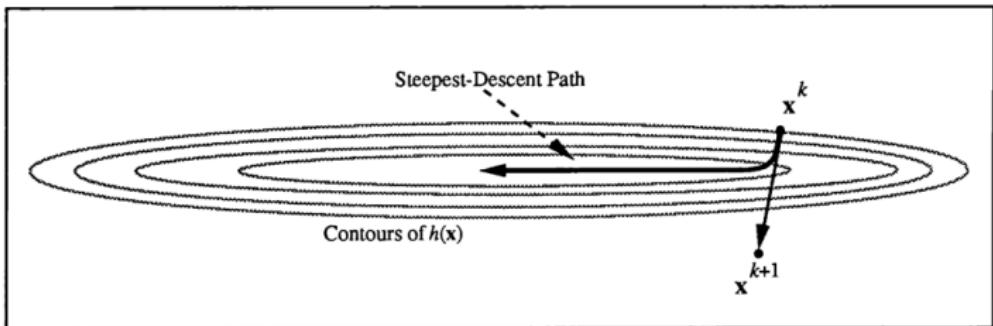
Projections  
 Definition  
 Convergence Analysis  
 Applications

## Proximal Gradient

Proximal Operator  
 Convergence Analysis  
 Applications



# Interpretation: Gradient flow<sup>10</sup>



## Gradient Descent

- Definition
- Convergence analysis
- Line search
- Applications
- Conclusion

## Subgradient Method

- Definition
- Convergence Analysis
- Applications

## Gradient Projection

- Projections
- Definition
- Convergence Analysis
- Applications

## Proximal Gradient

- Proximal Operator
- Convergence Analysis
- Applications

<sup>10</sup>Eckstein, Splitting methods for monotone operators with applications to parallel optimization, 1989, pp. 37-42



# Proximal gradient algorithm

## Definition

For a closed, proper, convex function  $G : \mathbb{R}^n \rightarrow \mathbb{R} \cup \{\infty\}$  and a function  $F \in \mathcal{C}^1(\mathbb{R}^n)$ , given an initial point  $u^0 \in \mathbb{R}^n$  and a step size sequence  $(\tau_k) \subset \mathbb{R}$ , the algorithm

$$u^{k+1} = \text{prox}_{\tau_k G}\left(u^k - \tau_k \nabla F(u^k)\right), \quad k = 0, 1, 2, \dots,$$

is called the *proximal gradient method*.

- Idea: Minimize sum  $E(u) = F(u) + G(u)$  by taking gradient step on smooth part and prox-step on possibly nonsmooth part
- Often referred to as *forward-backward splitting* or ISTA
- For constant  $G$ , it reduces to *gradient descent*
- For constant  $F$ , it is called *proximal point algorithm*
- For  $G = \iota_C$ , it reduces to *projected gradient descent*

## Gradient Descent

Definition
Convergence analysis
Line search
Applications
Conclusion

## Subgradient Method

Definition
Convergence Analysis
Applications

## Gradient Projection

Projections
Definition
Convergence Analysis
Applications

## Proximal Gradient

Proximal Operator
Convergence Analysis
Applications



- Recall majorization minimization interpretation of gradient descent (for  $L$ -smooth energies  $E$ )

$$u^{k+1} = \operatorname{argmin}_u E(u^k) + \langle \nabla E(u^k), u - u^k \rangle + \frac{L}{2} \|u - u^k\|^2$$

- Similarly it holds for the proximal gradient method:

$$\begin{aligned} u^{k+1} &= \operatorname{argmin}_u G(u) + F(u^k) + \langle \nabla F(u^k), u - u^k \rangle + \frac{L}{2} \|u - u^k\|^2 \\ &= \operatorname{prox}_{(1/L)G} \left( u^k - (1/L) \nabla F(u^k) \right) \end{aligned}$$

- Interpretation: replace the  $L$ -smooth function by a quadratic upper bound to obtain bound on whole function
- Iteratively minimize sequence of upper bounds

## Gradient Descent

Definition

Convergence analysis

Line search

Applications

Conclusion

## Subgradient Method

Definition

Convergence Analysis

Applications

## Gradient Projection

Projections

Definition

Convergence Analysis

Applications

## Proximal Gradient

Proximal Operator

Convergence Analysis

Applications



## Convergence analysis

- Assume  $G$  is closed, proper and convex, so that  $\text{prox}_G$  is well-defined
- Assume  $F$  is  $L$ -smooth and that  $E = G + F$  has at least one global minimizer  $u^*$
- Central ingredient for analysis is the **gradient map**

$$\varphi_\tau(u) = \frac{1}{\tau} (u - \text{prox}_{\tau G}(u - \tau \nabla F(u)))$$

- One iteration of the proximal gradient algorithm is then

$$u^+ = u - \tau \varphi_\tau(u)$$

- A short calculation shows that

$$\varphi_\tau(u) \in \nabla F(u) + \partial G(u - \tau \varphi_\tau(u))$$

- From the above it immediately follows:

$$\varphi_\tau(u) = 0 \Leftrightarrow -\nabla F(u) \in \partial G(u) \Leftrightarrow u \text{ is optimal.}$$

### Gradient Descent

- Definition  
 Convergence analysis  
 Line search  
 Applications  
 Conclusion

### Subgradient Method

- Definition  
 Convergence Analysis  
 Applications

### Gradient Projection

- Projections  
 Definition  
 Convergence Analysis  
 Applications

### Proximal Gradient

- Proximal Operator  
 Convergence Analysis  
 Applications



- Since  $F$  is  $L$ -smooth we have the quadratic upper bound:

$$F(v) \leq F(u) + \langle \nabla F(u), v - u \rangle + \frac{L}{2} \|v - u\|^2, \forall v \in \mathbb{R}^n$$

- For special choice  $v = u - \tau \varphi_\tau(u)$  we have

$$\begin{aligned} F(u - \tau \varphi_\tau(u)) &\leq F(u) - \tau \langle \nabla F(u), \varphi_\tau(u) \rangle + \frac{L\tau^2}{2} \|\varphi_\tau(u)\|^2 \\ &\stackrel{\tau \leq 1/L}{\leq} F(u) - \tau \langle \nabla F(u), \varphi_\tau(u) \rangle + \frac{\tau}{2} \|\varphi_\tau(u)\|^2 \end{aligned}$$

- From that we can conclude ( $\rightarrow$  board!) the following **global estimate**,  $\forall w$ :

$$E(u - \tau \varphi_\tau(u)) \leq E(w) + \langle \varphi_\tau(u), u - w \rangle - \frac{\tau}{2} \|\varphi_\tau(u)\|^2$$

## Gradient Descent

Definition

Convergence analysis

Line search

Applications

Conclusion

## Subgradient Method

Definition

Convergence Analysis

Applications

## Gradient Projection

Projections

Definition

Convergence Analysis

Applications

## Proximal Gradient

Proximal Operator

Convergence Analysis

Applications



## Convergence analysis

- Proximal gradient update

$$u^+ = u - \tau \varphi_\tau(u)$$

- From the global estimate with  $w = u$  it follows that proximal gradient is a descent method ( $\tau < 1/L$ )

$$E(u^+) \leq E(u) - \frac{\tau}{2} \|\varphi_\tau(u)\|^2$$

- With the choice  $w = u^*$  it follows

$$E(u^+) - E(u^*) \leq \frac{1}{2\tau} \left( \|u - u^*\|^2 - \|u^+ - u^*\|^2 \right)$$

- Summing the above estimate yields  $\mathcal{O}(1/k)$  convergence rate for energy

$$E(u^k) - E(u^*) \leq \frac{1}{2k\tau} \|u^0 - u^*\|^2$$

### Gradient Descent

Definition

Convergence analysis

Line search

Applications

Conclusion

### Subgradient Method

Definition

Convergence Analysis

Applications

### Gradient Projection

Projections

Definition

Convergence Analysis

Applications

### Proximal Gradient

Proximal Operator

Convergence Analysis

Applications



# Convergence analysis

## Convergence of proximal gradient

Let  $E = G + F$ , with proper, closed, convex  $G : \mathbb{R}^n \rightarrow \mathbb{R} \cup \{\infty\}$ ,  $F \in \mathcal{F}_L^{1,1}(\mathbb{R}^n)$  have a global minimizer  $u^*$ . Then the proximal gradient algorithm with constant stepsize  $\tau = \frac{1}{L}$  yields

$$E(u^k) - E(u^*) \leq \frac{L}{2k} \|u^0 - u^*\|^2.$$

Some remarks:

- If either  $G$  or  $F$  is additionally strongly convex, can be shown to converge with linear rate (not covered here)
- Similar to gradient descent, proximal gradient descent can be combined with a line search ( $\rightarrow$  exercise!)

### Gradient Descent

- Definition
- Convergence analysis
- Line search
- Applications
- Conclusion

### Subgradient Method

- Definition
- Convergence Analysis
- Applications

### Gradient Projection

- Projections
- Definition
- Convergence Analysis
- Applications

### Proximal Gradient

- Proximal Operator
- Convergence Analysis
- Applications



Measured signals  $f$

Find decomposition  $f = Au + n$

Dictionary of materials  $A$ , mixing coefficients  $u$  (sparse) and noise  $n$

### Gradient Descent

- Definition
- Convergence analysis
- Line search
- Applications
- Conclusion

### Subgradient Method

- Definition
- Convergence Analysis
- Applications

### Gradient Projection

- Projections
- Definition
- Convergence Analysis
- Applications

### Proximal Gradient

- Proximal Operator
- Convergence Analysis
- Applications



## Example application: Unmixing and sparse recovery

- Recall the setup of sparse recovery from two weeks ago:
- Minimize a data fidelity term  $H_f(v)$  such that  $v$  can be represented in a dictionary  $A$ , i.e.  $v = Au$ , and the representing coefficients  $u$  are sparse
- Energy minimization approach:

$$\min_u H_f(Au) + \alpha \|u\|_1.$$

- To make the problem amenable to gradient projection we rewrote it as

$$\min_u H_f(A(u_1 - u_2)) + \alpha \langle u_1, \mathbf{1} \rangle + \alpha \langle u_2, \mathbf{1} \rangle, \quad u_1 \geq 0, u_2 \geq 0$$

- With proximal gradient, the  $\ell_1$ -norm formulation can be optimized directly:

$$u^{k+1} = \text{prox}_{\tau \|\cdot\|_1}(u^k - \tau A^T \nabla H_f(Au^k)).$$

### Gradient Descent

- Definition
- Convergence analysis
- Line search
- Applications
- Conclusion

### Subgradient Method

- Definition
- Convergence Analysis
- Applications

### Gradient Projection

- Projections
- Definition
- Convergence Analysis
- Applications

### Proximal Gradient

- Proximal Operator
- Convergence Analysis
- Applications



## Gradient Descent

Definition

Convergence analysis

Line search

Applications

Conclusion

## Subgradient Method

Definition

Convergence Analysis

Applications

## Gradient Projection

Projections

Definition

Convergence Analysis

Applications

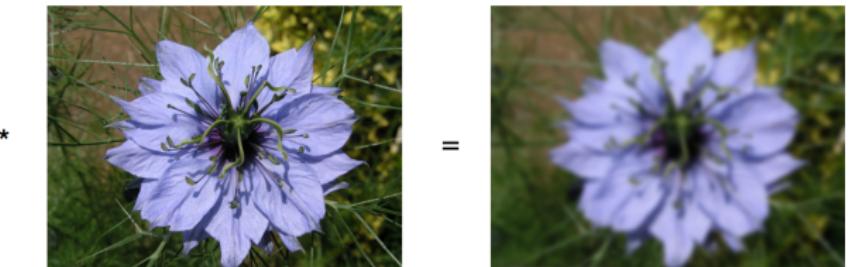
## Proximal Gradient

Proximal Operator

Convergence Analysis

Applications

## Example application: Deconvolution



- Convolution computes a weighted sum of the image values.
- The goal of image deconvolution is to restore the original sharp image.



- Given a *kernel*  $k : \mathbb{R}^2 \rightarrow \mathbb{R}$  and an image  $u : \Omega \rightarrow \mathbb{R}^n$ :

$$(k * u)(x, y) = \int_{\mathbb{R}^2} k(a, b) u(x - a, y - b) da db$$

- This sums up the  $u$  values around  $(x, y)$ , weighted by  $K$ .
- The formula needs values of  $u$  outside of the domain  $\Omega$ .
- Possible choices
  - Clamping* of  $(x, y)$  back to  $\Omega$
  - Periodic* boundary conditions
  - Mirroring* boundary conditions

### Gradient Descent

Definition  
 Convergence analysis  
 Line search  
 Applications  
 Conclusion

### Subgradient Method

Definition  
 Convergence Analysis  
 Applications

### Gradient Projection

Projections  
 Definition  
 Convergence Analysis  
 Applications

### Proximal Gradient

Proximal Operator  
 Convergence Analysis  
 Applications



# Convolution

- Finite weighted sum:

$$(k * u)(x, y) = \sum_{(a,b) \in S_K} k(a, b) \cdot u(x - a, y - b)$$

- $S_k$  is the *support* of  $k$ : positions  $(a, b)$  with  $k(a, b) \neq 0$ .
- It is often assumed to lie entirely in a *small window* of size  $(2r_x + 1) \times (2r_y + 1)$ :

$$(k * u)(x, y) = \sum_{a=-r_x}^{r_x} \sum_{b=-r_y}^{r_y} k(a, b) \cdot u(x - a, y - b)$$

- One often deals with small-support kernels  $k$ , or the kernel is truncated artificially (e.g. *Gaussian kernel*).
- Convolution is *linear* and can be represented by a matrix

## Gradient Descent

Definition  
 Convergence analysis  
 Line search  
 Applications  
 Conclusion

## Subgradient Method

Definition  
 Convergence Analysis  
 Applications

## Gradient Projection

Projections  
 Definition  
 Convergence Analysis  
 Applications

## Proximal Gradient

Proximal Operator  
 Convergence Analysis  
 Applications



- Assumption: observation  $f$  is convolution of unknown image  $u$  with known kernel  $k$  plus Gaussian noise

$$f = k * u + \eta, \quad \eta \sim \mathcal{N}(0, \sigma)$$

- Leads to convex optimization problem,  $A \in \mathbb{R}^{n \times n}$  represents the convolution:

$$\operatorname{argmin}_{u \in \mathbb{R}^n} \underbrace{(1/2) \|Au - f\|^2}_{\text{convex + } L\text{-smooth}} + \underbrace{\lambda TV(u)}_{\text{convex}}$$

- Energy is sum of  $L$ -smooth and non-smooth term
- One possibility: solve using proximal gradient algorithm

## Gradient Descent

Definition  
 Convergence analysis  
 Line search  
 Applications  
 Conclusion

## Subgradient Method

Definition  
 Convergence Analysis  
 Applications

## Gradient Projection

Projections  
 Definition  
 Convergence Analysis  
 Applications

## Proximal Gradient

Proximal Operator  
 Convergence Analysis  
 Applications



- Proximal gradient algorithm for TV deconvolution:

$$u^{k+1} = \operatorname{argmin}_u \tau\lambda TV(u) + \frac{1}{2} \left\| u - (u^k - \tau A^T(Au^k - f)) \right\|^2$$

- Solve a sequence of total variation denoising problems (using projected gradient on dual problem)
- Warm-start with solution from previous outer iteration
- Potential difficulty: inexact solution of inner problem
- In practice, proximal gradient is quite robust
- Later in the semester we will learn how to solve this problem without the need for inner iterations

## Gradient Descent

Definition  
 Convergence analysis  
 Line search  
 Applications  
 Conclusion

## Subgradient Method

Definition  
 Convergence Analysis  
 Applications

## Gradient Projection

Projections  
 Definition  
 Convergence Analysis  
 Applications

## Proximal Gradient

Proximal Operator  
 Convergence Analysis  
 Applications

Michael Moeller

Thomas Möllenhoff

Emanuel Laude



## Gradient Descent

Definition

Convergence analysis

Line search

Applications

Conclusion

## Subgradient Method

Definition

Convergence Analysis

Applications

## Gradient Projection

Projections

Definition

Convergence Analysis

Applications

## Proximal Gradient

Proximal Operator

Convergence Analysis

Applications

Michael Moeller

Thomas Möllenhoff

Emanuel Laude



## Gradient Descent

Definition

Convergence analysis

Line search

Applications

Conclusion

## Subgradient Method

Definition

Convergence Analysis

Applications

## Gradient Projection

Projections

Definition

Convergence Analysis

Applications

## Proximal Gradient

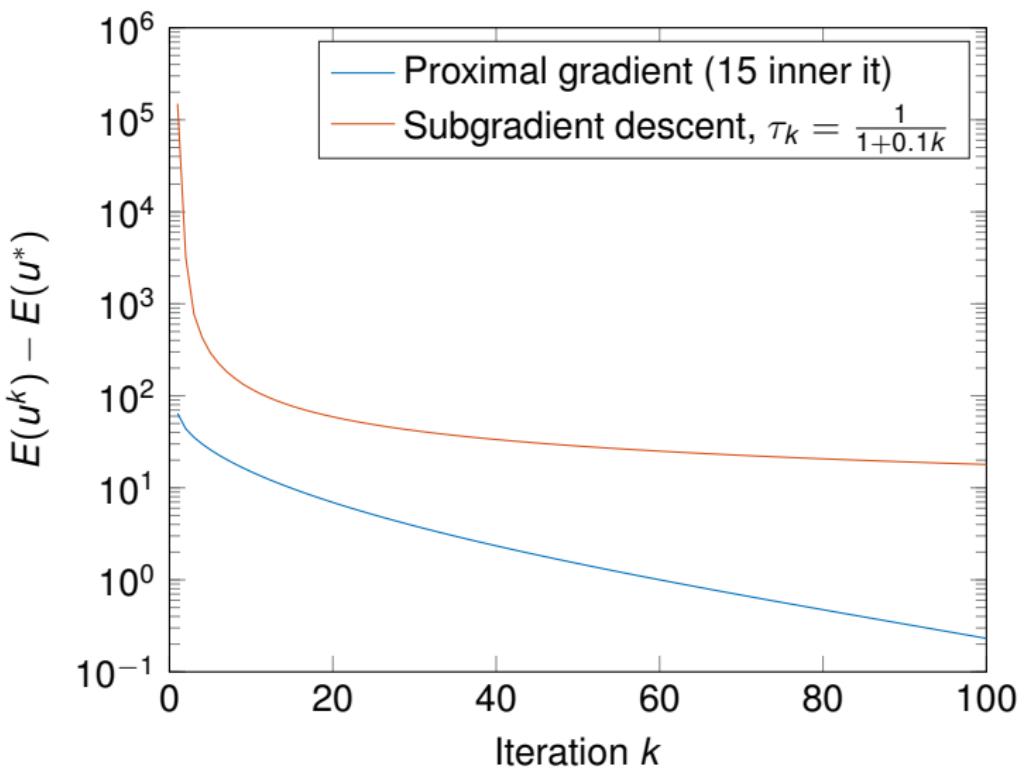
Proximal Operator

Convergence Analysis

Applications

# Deblurred result





## Gradient Descent

- Definition
- Convergence analysis
- Line search
- Applications
- Conclusion

## Subgradient Method

- Definition
- Convergence Analysis
- Applications

## Gradient Projection

- Projections
- Definition
- Convergence Analysis
- Applications

## Proximal Gradient

- Proximal Operator
- Convergence Analysis
- Applications



- Downsampling operator  $S : \mathbb{R}^n \rightarrow \mathbb{R}^m$ ,  $m \ll n$
- $Su \in \mathbb{R}^m$  downsampled version of  $u \in \mathbb{R}^n$
- Image zooming:

$$\operatorname{argmin}_u \frac{1}{2} \|Su - f\|^2 + \lambda TV(u)$$

- Solution with proximal gradient method same as in deblurring example (sequence of TV denoising problems)

## Gradient Descent

Definition  
 Convergence analysis  
 Line search  
 Applications  
 Conclusion

## Subgradient Method

Definition  
 Convergence Analysis  
 Applications

## Gradient Projection

Projections  
 Definition  
 Convergence Analysis  
 Applications

## Proximal Gradient

Proximal Operator  
 Convergence Analysis  
 Applications

Michael Moeller

Thomas Möllenhoff

Emanuel Laude



## Gradient Descent

Definition

Convergence analysis

Line search

Applications

Conclusion

## Subgradient Method

Definition

Convergence Analysis

Applications

## Gradient Projection

Projections

Definition

Convergence Analysis

Applications

## Proximal Gradient

Proximal Operator

Convergence Analysis

Applications

Michael Moeller

Thomas Möllenhoff

Emanuel Laude



## Gradient Descent

Definition

Convergence analysis

Line search

Applications

Conclusion

## Subgradient Method

Definition

Convergence Analysis

Applications

## Gradient Projection

Projections

Definition

Convergence Analysis

Applications

## Proximal Gradient

Proximal Operator

Convergence Analysis

Applications



## Concluding remarks

- This concludes the chapter on gradient methods
- We have introduced and analyzed four algorithms:
  - Gradient descent
  - Subgradient descent
  - Projected (sub-)gradient descent
  - Proximal gradient descent
- The proximal mapping “prox” will be the central building block for most further methods
- After the break: **primal-dual algorithms!**

### Gradient Descent

Definition  
Convergence analysis  
Line search  
Applications  
Conclusion

### Subgradient Method

Definition  
Convergence Analysis  
Applications

### Gradient Projection

Projections  
Definition  
Convergence Analysis  
Applications

### Proximal Gradient

Proximal Operator  
Convergence Analysis  
Applications