Chapter 2

Gradient Methods

Convex Optimization for Computer Vision

SS 2016

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

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

Subgradient Method
- Definition
- Convergence Analysis
- Applications

Gradient Projection
- Projections
- Definition
- Convergence Analysis
- Applications

updated 05.05.2016

Gradient Descent

Gradient Descent
Definition
Convergence analysis
Line search
Applications
Conclusion

Subgradient Method
Definition
Convergence Analysis
Applications

Gradient Projection
Projections
Definition
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 $E : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ proper, closed, convex.

We start making our life easier:

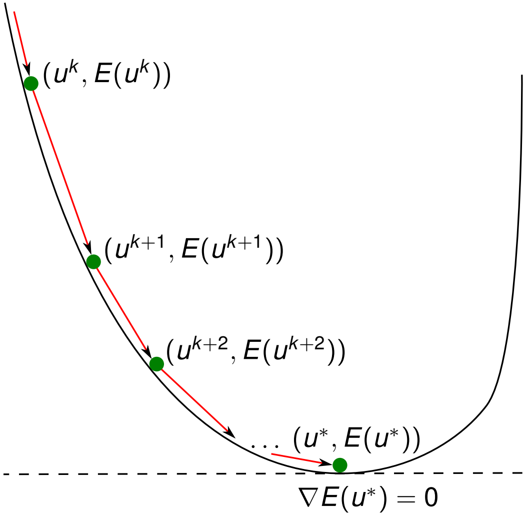
- $\operatorname{dom} E = \mathbb{R}^n$
- $E \in \mathcal{C}^1(\mathbb{R}^n)$
- Even more assumptions later :-)

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

Descent methods

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



Gradient Descent
Definition
Convergence analysis
Line search
Applications
Conclusion

Subgradient Method
Definition
Convergence Analysis
Applications

Gradient Projection
Projections
Definition
Convergence Analysis
Applications

Descent methods

- 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

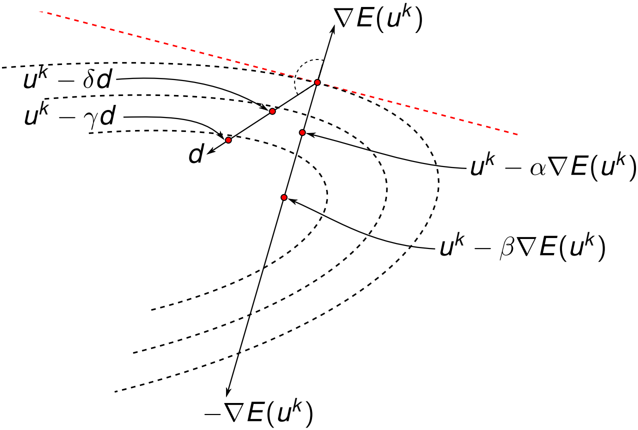
$$E(u(\tau)) = E(u^k + \tau d) = E(u^k) + \tau \langle \nabla E(u^k), d \rangle + o(\tau)$$

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

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

Descent methods



Gradient Descent
Definition
Convergence analysis
Line search
Applications
Conclusion
Subgradient Method
Definition
Convergence Analysis
Applications
Gradient Projection
Projections
Definition
Convergence Analysis
Applications

Descent methods

- The negative gradient is the *steepest* descent direction

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

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

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

- 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: $D^k = I$
  - ...

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

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

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

is called *gradient descent*.

Philosophy:

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

Choice of τ_k:

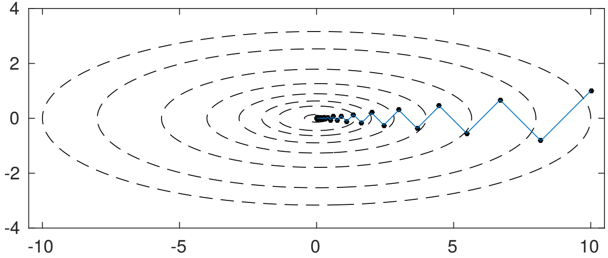
- $\tau_k = \tau$ for some constant $\tau \in \mathbb{R}$ (this lecture)
- Exact line search $\tau_k = \arg\min_\tau \; E\left(u^k - \tau \nabla E(u^k)\right)$
- Inexact line search (more later)

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

A first toy example

$$E(u) = \frac{1}{2}\left(u_1^2 + \kappa u_2^2\right) \qquad \kappa > 1$$



- Convergence rate with exact line search [1]

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

[1] 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

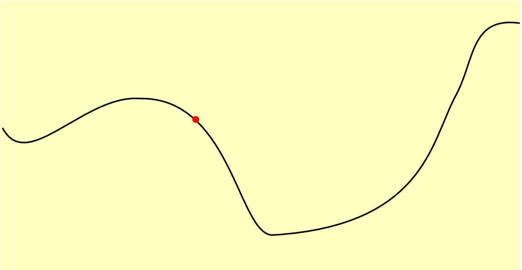
Lipschitz continuity

Definition

$f : \mathbb{R}^n \to \mathbb{R}^m$ 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*



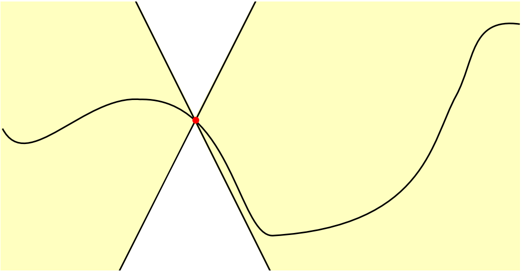
Lipschitz continuity

Definition

$f : \mathbb{R}^n \to \mathbb{R}^m$ 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

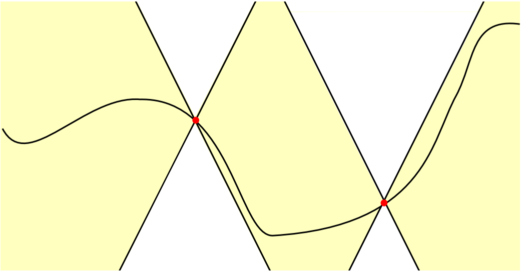
Lipschitz continuity

Definition

$f : \mathbb{R}^n \to \mathbb{R}^m$ 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

- Imporant special case are linear functions $f : \mathbb{R}^n \to \mathbb{R}^m$
- 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

Lipschitz continuity

Theorem: Lipschitz continuity for differentiable functions

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

Proof: Board!

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

Lipschitz continuity

Definition: Functions with Lipschitz derivative

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 $E : \mathbb{R}^n \to \mathbb{R}$ 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).

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

Convexity and Lipschitz continuity

Reminder: Characterization of convex functions [2]

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 $\mathcal{F}_L^{k,p}(Q)$.

[2] Boyd, Vandenberghe, Convex Optimization, Section 3.1.3

Gradient Descent
Definition
Convergence analysis
Line search
Applications
Conclusion
Subgradient Method
Definition
Convergence Analysis
Applications
Gradient Projection
Projections
Definition
Convergence Analysis
Applications

Convexity and Lipschitz continuity

Theorem: Characterization of convex L-smooth functions [3]

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 $E \in \mathcal{C}^2(\mathbb{R}^n)$

Proof: See notes!

[3] 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

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

- For the minimum of the upper bound we have

$$\begin{aligned} E(u^*) &\leq \min_v \; E(u^k) + \langle \nabla E(u^k), v - u^k \rangle + \frac{L}{2} \left\| v - u^k \right\|^2 \\ &= E(u^k) - \frac{1}{2L} \left\| \nabla E(u^k) \right\|^2 \end{aligned}$$

Gradient Descent
Definition
Convergence analysis
Line search
Applications
Conclusion

Subgradient Method
Definition
Convergence Analysis
Applications

Gradient Projection
Projections
Definition
Convergence Analysis
Applications

Divergent example

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

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

- 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

Gradient Descent
Definition
Convergence analysis
Line search
Applications
Conclusion
Subgradient Method
Definition
Convergence Analysis
Applications
Gradient Projection
Projections
Definition
Convergence Analysis
Applications

Strong convexity

Definition: strong convexity

A function $E : \mathbb{R}^n \to \overline{\mathbb{R}}$ 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

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

Strong convexity

Theorem: characterization of m-strongly convex functions [4]

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., $E \in \mathcal{S}^1_m(\mathbb{R}^n)$
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 $E \in \mathcal{C}^2(\mathbb{R}^n)$

Proof: See literature.

[4] 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

Strong convexity and Lipschitz continuity

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

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

$\rightarrow$ *Online TED.*

Gradient Descent
Definition
Convergence analysis
Line search
Applications
Conclusion

Subgradient Method
Definition
Convergence Analysis
Applications

Gradient Projection
Projections
Definition
Convergence Analysis
Applications

What we have seen so far...

- 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

Strong convexity and Lipschitz continuity

Theorem: strongly convex + L-smooth bound

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

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

Proof: Exercise!

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

Gradient descent convergence

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

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

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

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

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

Proof: Board!

Remarks:

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

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

Gradient descent convergence

Theorem: Convergence (L-smooth)

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

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

Proof: Board!

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

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

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 [5]
  - 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

[5] 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

Line search

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

$$\tau_k \leftarrow 1$$

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

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

$$\text{end}$$

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

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

Image denoising

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

Image denoising

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

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

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

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

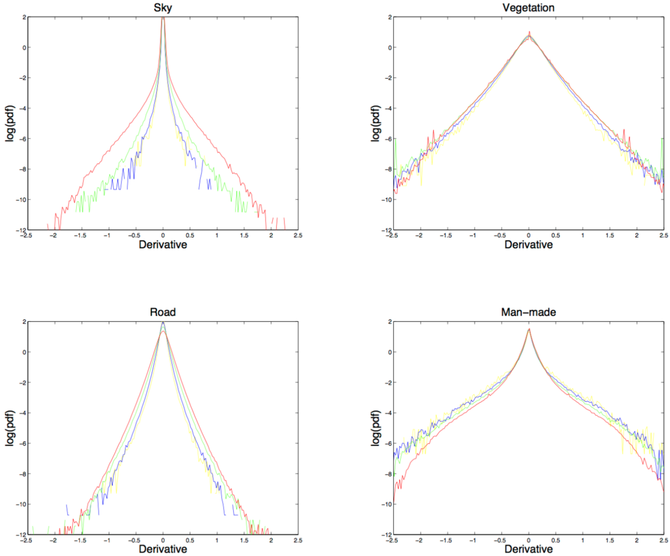
- Natural image statistics suggest the choice

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

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

Natural image statistics [6]



[6] J. Huang, D. Mumford, Statistics of Natural Images and Models, CVPR '99

Gradient Descent
Definition
Convergence analysis
Line search
Applications
Conclusion
Subgradient Method
Definition
Convergence Analysis
Applications
Gradient Projection
Projections
Definition
Convergence Analysis
Applications

Image denoising

- Minimize negative logarithm

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

- $E(u)$ is λ-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

Image denoising

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

Evolution to global optimum via gradient descent

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

Subgradient Method
- Definition
- Convergence Analysis
- Applications

Gradient Projection
- Projections
- Definition
- Convergence Analysis
- Applications

$\varepsilon = 0.1$

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

Subgradient Method
- Definition
- Convergence Analysis
- Applications

Gradient Projection
- Projections
- Definition
- Convergence Analysis
- Applications

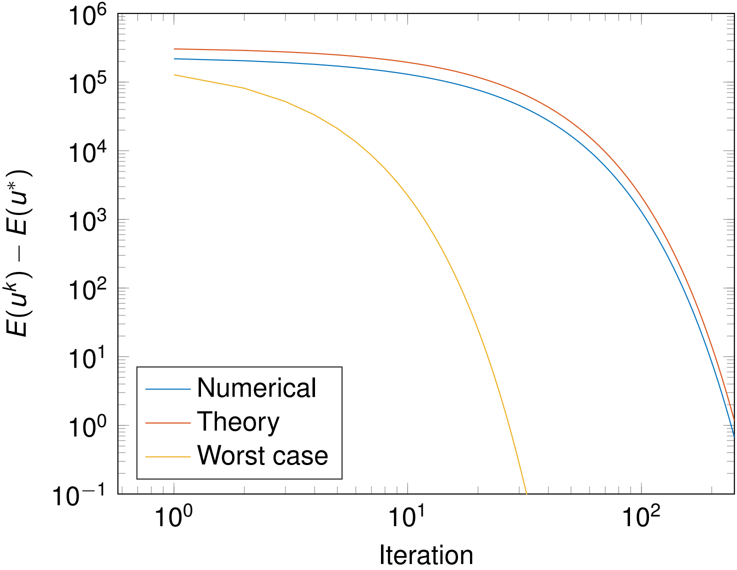
$\varepsilon = 0.01$

$\rightarrow$ *Motivation for non-smooth optimization!*

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

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

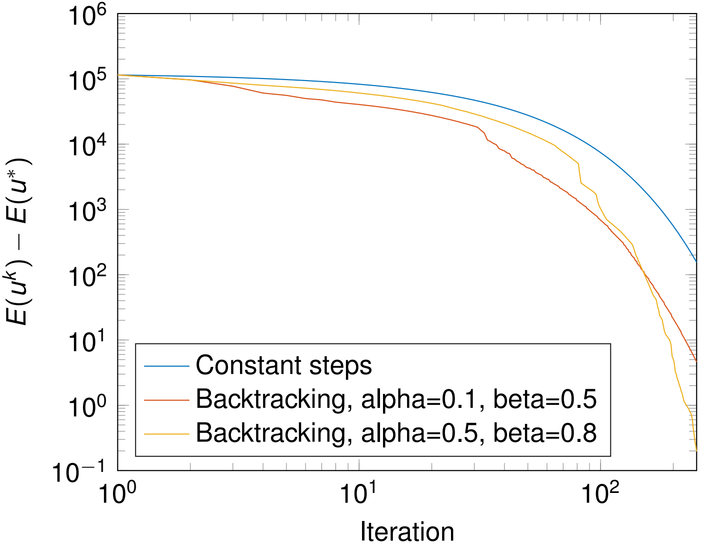


Gradient Descent
Definition
Convergence analysis
Line search
Applications
Conclusion

Subgradient Method
Definition
Convergence Analysis
Applications

Gradient Projection
Projections
Definition
Convergence Analysis
Applications

Convergence, backtracking line search

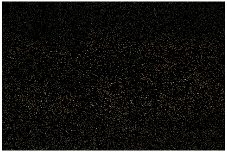


Gradient Descent
Definition
Convergence analysis
Line search
Applications
Conclusion

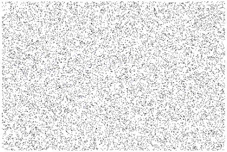
Subgradient Method
Definition
Convergence Analysis
Applications

Gradient Projection
Projections
Definition
Convergence Analysis
Applications

Image inpainting



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



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

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

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

- 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

Image Inpainting

Gradient Descent
Definition
Convergence analysis
Line search
Applications
Conclusion

Subgradient Method
Definition
Convergence Analysis
Applications

Gradient Projection
Projections
Definition
Convergence Analysis
Applications

50% missing pixels

Gradient Descent
Definition
Convergence analysis
Line search
Applications
Conclusion
Subgradient Method
Definition
Convergence Analysis
Applications
Gradient Projection
Projections
Definition
Convergence Analysis
Applications

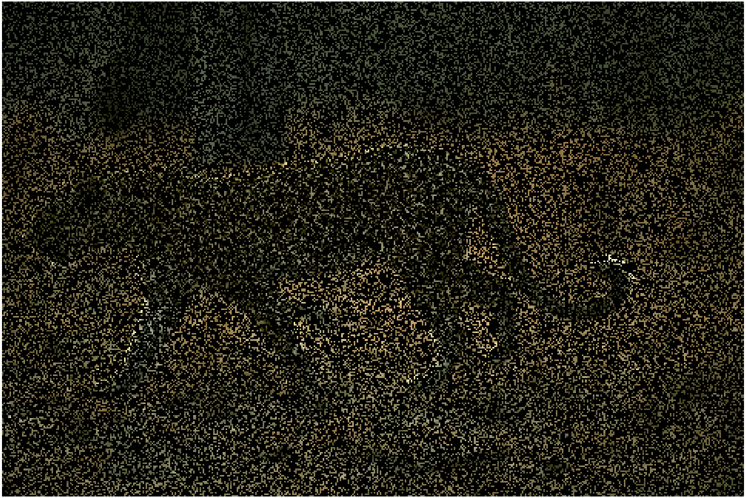
50% missing pixels

Gradient Descent
Definition
Convergence analysis
Line search
Applications
Conclusion

Subgradient Method
Definition
Convergence Analysis
Applications

Gradient Projection
Projections
Definition
Convergence Analysis
Applications

70% missing pixels



Gradient Descent
Definition
Convergence analysis
Line search
Applications
Conclusion

Subgradient Method
Definition
Convergence Analysis
Applications

Gradient Projection
Projections
Definition
Convergence Analysis
Applications

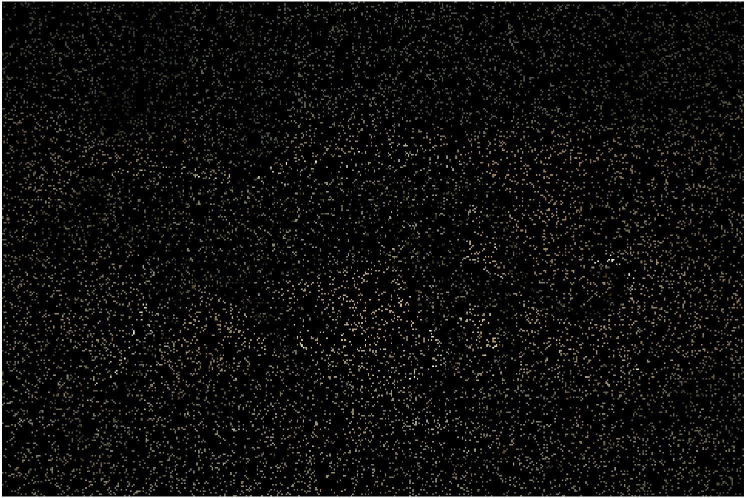
70% missing pixels

Gradient Descent
Definition
Convergence analysis
Line search
Applications
Conclusion

Subgradient Method
Definition
Convergence Analysis
Applications

Gradient Projection
Projections
Definition
Convergence Analysis
Applications

90% missing pixels



Gradient Descent
Definition
Convergence analysis
Line search
Applications
Conclusion

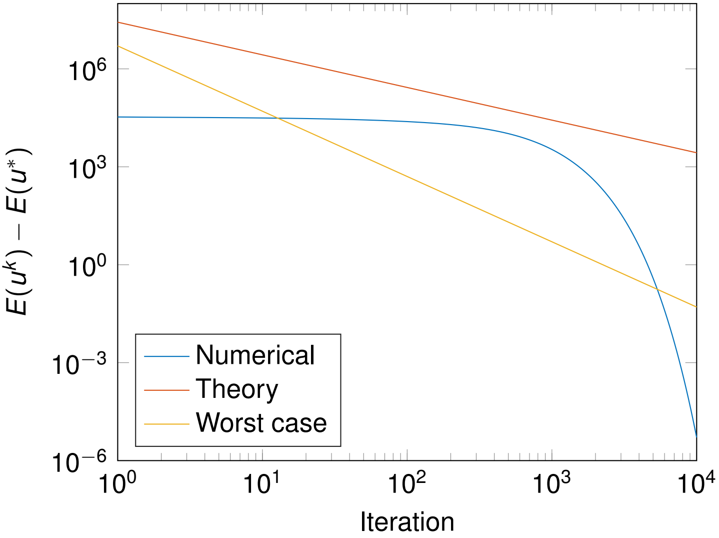
Subgradient Method
Definition
Convergence Analysis
Applications

Gradient Projection
Projections
Definition
Convergence Analysis
Applications

90% missing pixels

Gradient Descent
Definition
Convergence analysis
Line search
Applications
Conclusion
Subgradient Method
Definition
Convergence Analysis
Applications
Gradient Projection
Projections
Definition
Convergence Analysis
Applications

Convergence, $\tau = 1/L$



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

Subgradient Method
- Definition
- Convergence Analysis
- Applications

Gradient Projection
- Projections
- Definition
- Convergence Analysis
- Applications

Fast optimization challenge I

- Minimize the inpainting energy

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

- 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_huber_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

Handwritten digit recognition

- MNIST dataset[7], handwritten digit recognition
- $K = 10$ digits, 28×28 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)

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

[7] 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

Multinomial logistic regression

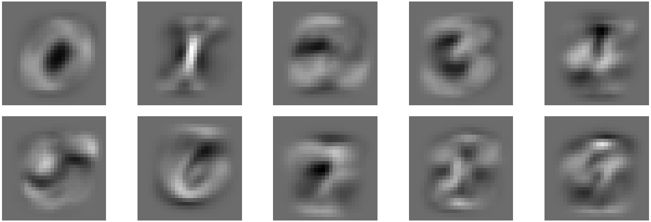
- Minimize negative log-likelihood

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

- It can be shown that $E(W)$ is λ-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

Multinomial logistic regression

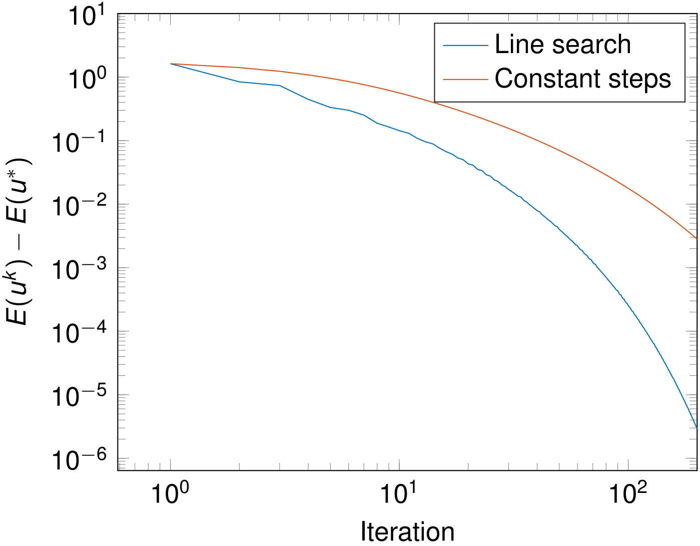


- Classifier gives around 10% error on test set
- Can be easily improved to around $1 - 2\%$ with a few additional lines of MATLAB code (use features instead of raw pixels)
- Current best: 0.23% (convolutional neural networks)
- Learn more about learning:
  `https://vision.in.tum.de/teaching/ss2016/mlcv16`

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

Multinomial logistic regression



Gradient Descent
Definition
Convergence analysis
Line search
Applications
Conclusion

Subgradient Method
Definition
Convergence Analysis
Applications

Gradient Projection
Projections
Definition
Convergence Analysis
Applications

Concluding remarks and outlook

- 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

Subgradient Method

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

Non-smooth optimization

- 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 ∂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

Subgradient descent

Definition

Given a convex function $E : \mathbb{R}^n \to \mathbb{R}$, 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

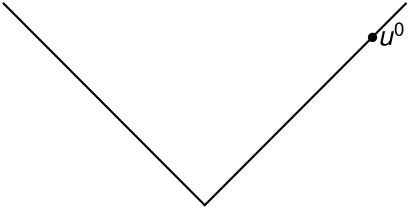
Michael Moeller
Thomas Möllenhoff
Emanuel Laude

Gradient Descent
Definition
Convergence analysis
Line search
Applications
Conclusion
Subgradient Method
Definition
Convergence Analysis
Applications
Gradient Projection
Projections
Definition
Convergence Analysis
Applications

First example

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

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



Gradient Descent
Definition
Convergence analysis
Line search
Applications
Conclusion

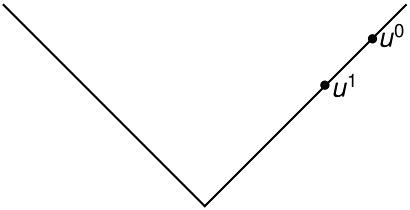
Subgradient Method
Definition
Convergence Analysis
Applications

Gradient Projection
Projections
Definition
Convergence Analysis
Applications

First example

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

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



Gradient Descent
Definition
Convergence analysis
Line search
Applications
Conclusion

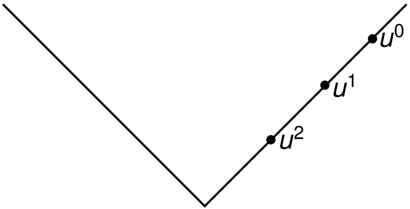
Subgradient Method
Definition
Convergence Analysis
Applications

Gradient Projection
Projections
Definition
Convergence Analysis
Applications

First example

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

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

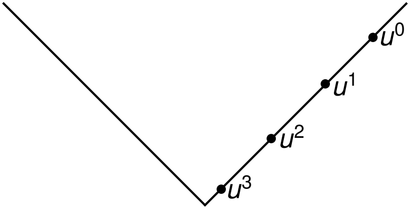


Gradient Descent
Definition
Convergence analysis
Line search
Applications
Conclusion
Subgradient Method
Definition
Convergence Analysis
Applications
Gradient Projection
Projections
Definition
Convergence Analysis
Applications

First example

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

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



Gradient Descent
Definition
Convergence analysis
Line search
Applications
Conclusion

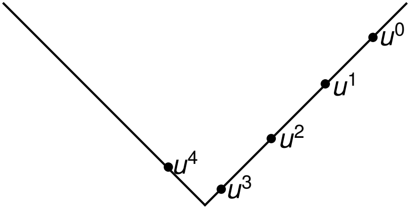
Subgradient Method
Definition
Convergence Analysis
Applications

Gradient Projection
Projections
Definition
Convergence Analysis
Applications

First example

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

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

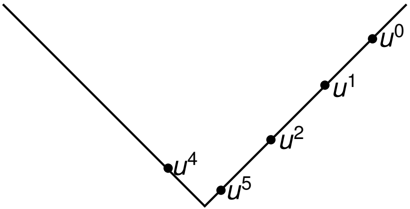


Gradient Descent
Definition
Convergence analysis
Line search
Applications
Conclusion
Subgradient Method
Definition
Convergence Analysis
Applications
Gradient Projection
Projections
Definition
Convergence Analysis
Applications

First example

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

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

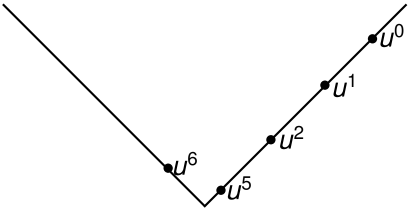


Gradient Descent
Definition
Convergence analysis
Line search
Applications
Conclusion
Subgradient Method
Definition
Convergence Analysis
Applications
Gradient Projection
Projections
Definition
Convergence Analysis
Applications

First example

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

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

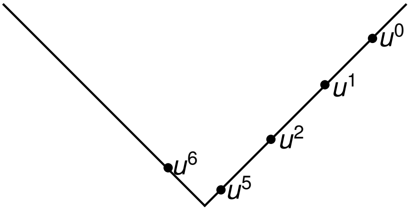


Gradient Descent
Definition
Convergence analysis
Line search
Applications
Conclusion
Subgradient Method
Definition
Convergence Analysis
Applications
Gradient Projection
Projections
Definition
Convergence Analysis
Applications

First example

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

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

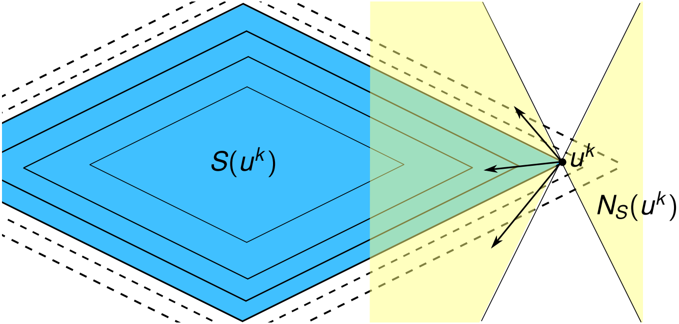


- Doesn't converge to optimum for constant step sizes

Gradient Descent
Definition
Convergence analysis
Line search
Applications
Conclusion
Subgradient Method
Definition
Convergence Analysis
Applications
Gradient Projection
Projections
Definition
Convergence Analysis
Applications

Not a descent method!

- 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

Some assumptions

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

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

Proof: Board!

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

Convergence analysis

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

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

- Rearranging the above yields:

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

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

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

Gradient Descent
Definition
Convergence analysis
Line search
Applications
Conclusion

Subgradient Method
Definition
Convergence Analysis
Applications

Gradient Projection
Projections
Definition
Convergence Analysis
Applications

Convergence analysis for fixed step size

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

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

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

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

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

Gradient Descent
Definition
Convergence analysis
Line search
Applications
Conclusion
Subgradient Method
Definition
Convergence Analysis
Applications
Gradient Projection
Projections
Definition
Convergence Analysis
Applications

Diminishing step sizes

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

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

- For such a sequence it holds that

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

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

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

Polyak step size

- Recall the inequality we started with

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

- Right hand side is minimized for

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

- Plugging this back in yields

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

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

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

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

Worst-case complexity

- Problem class: convex functions $E : \mathbb{R}^n \to \mathbb{R}$
- 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 concievable methods is optimal
- Indicates that the problem class of general convex functions is too complicated to be solved efficiently

[8] Nesterov, Introductory Lectures on Convex Optimization, Theorem 3.2.1

Gradient Descent
Definition
Convergence analysis
Line search
Applications
Conclusion
Subgradient Method
Definition
Convergence Analysis
Applications
Gradient Projection
Projections
Definition
Convergence Analysis
Applications

The total variation

- 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:

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

- $D : \mathbb{R}^{Nc} \to \mathbb{R}^{2Nc}$ 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

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

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

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

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

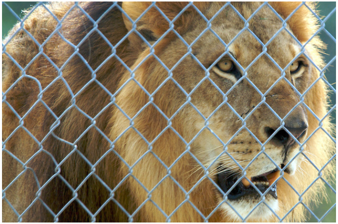
$$\mathbb{R}^{2c} \supset I_k = \begin{cases} \left\{ \frac{g_k}{\|g_k\|_2} \right\}, & \text{if } 0 \neq g_k, \\ B(0,1), & \text{otherwise.} \end{cases}$$

Gradient Descent
Definition
Convergence analysis
Line search
Applications
Conclusion

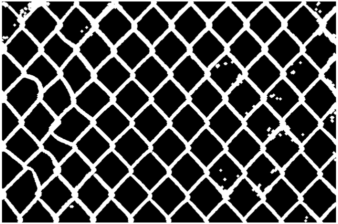
Subgradient Method
Definition
Convergence Analysis
Applications

Gradient Projection
Projections
Definition
Convergence Analysis
Applications

TV Inpainting



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



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

- 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

TV Inpainting

Gradient Descent
Definition
Convergence analysis
Line search
Applications
Conclusion

Subgradient Method
Definition
Convergence Analysis
Applications

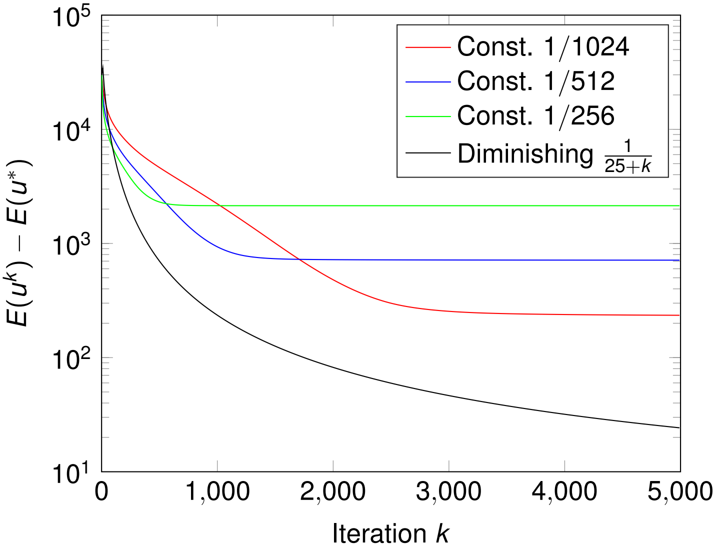
Gradient Projection
Projections
Definition
Convergence Analysis
Applications

TV Inpainting

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

Numerical convergence results



Gradient Descent
Definition
Convergence analysis
Line search
Applications
Conclusion

Subgradient Method
Definition
Convergence Analysis
Applications

Gradient Projection
Projections
Definition
Convergence Analysis
Applications

The TV-ℓ_1 model

- A robust model for image denoising is given by

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

- The ℓ_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

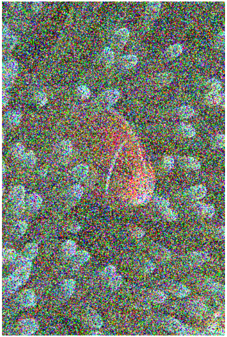
$$g^k = \lambda \text{sign}(u^k - f) + p^k, \text{ with } p^k \in \partial TV(u^k)$$

Michael Moeller
Thomas Möllenhoff
Emanuel Laude

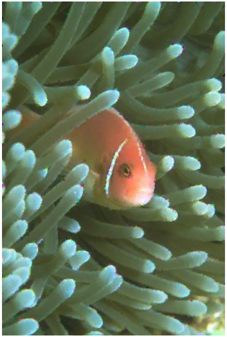
Gradient Descent
Definition
Convergence analysis
Line search
Applications
Conclusion
Subgradient Method
Definition
Convergence Analysis
Applications
Gradient Projection
Projections
Definition
Convergence Analysis
Applications

Robust denoising

Original



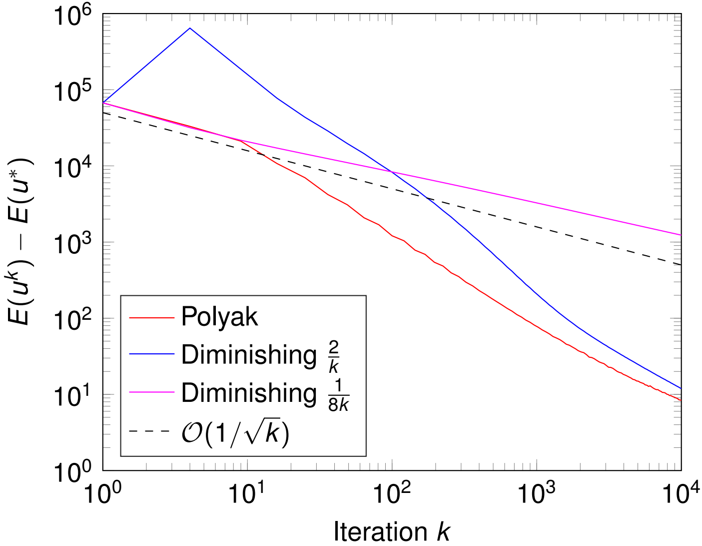
Noisy input



$TV - \ell_1$ denoised

Gradient Descent
Definition
Convergence analysis
Line search
Applications
Conclusion
Subgradient Method
Definition
Convergence Analysis
Applications
Gradient Projection
Projections
Definition
Convergence Analysis
Applications

Numerical convergence results



Gradient Descent
Definition
Convergence analysis
Line search
Applications
Conclusion

Subgradient Method
Definition
Convergence Analysis
Applications

Gradient Projection
Projections
Definition
Convergence Analysis
Applications

Concluding remarks

- 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

Michael Moeller
Thomas Möllenhoff
Emanuel Laude

Gradient Projection

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

Gradient and subgradient descent

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

for $E : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ 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

Gradient projection

Type of problem:

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

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

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.

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

Example projections

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

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

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

Properties of the projection

Firm Nonexpansiveness

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

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

By Cauchy-Schwartz, this implies the nonexpansiveness

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

Proof: Board

Gradient Descent
Definition
Convergence analysis
Line search
Applications
Conclusion

Subgradient Method
Definition
Convergence Analysis
Applications

Gradient Projection
Projections
Definition
Convergence Analysis
Applications

Idea of gradient projection

Consider a problem

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

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.

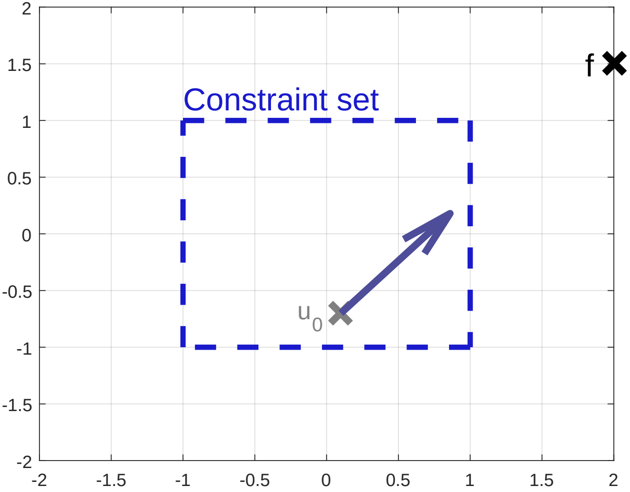
$$u^{k+1} = \pi_C(u^k - \tau^k \nabla E(u^k))$$

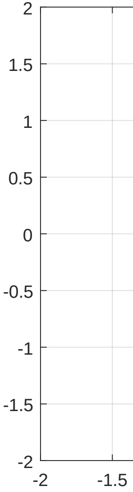
Michael Moeller
Thomas Möllenhoff
Emanuel Laude

Gradient Descent
Definition
Convergence analysis
Line search
Applications
Conclusion
Subgradient Method
Definition
Convergence Analysis
Applications
Gradient Projection
Projections
Definition
Convergence Analysis
Applications

Idea of gradient projection

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





Gradient projection algorithm

Gradient projection algorithm

Let $C \subset \mathbb{R}^n$ be a nonempty closed convex set and let $E : \mathbb{R}^n \to \mathbb{R} \in C^1(\mathbb{R}^n)$. 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.

When, how, why, and for which E and τ does it work?

Remember: Gradient descent

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

Same convergence for gradient projection?

Gradient Descent
Definition
Convergence analysis
Line search
Applications
Conclusion

Subgradient Method
Definition
Convergence Analysis
Applications

Gradient Projection
Projections
Definition
Convergence Analysis
Applications

Gradient projection algorithm

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

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

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

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:

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

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

Gradient projection algorithm

Gradient Projection Estimate

For $E \in \mathcal{S}^{1,1}_{m,L}(\mathbb{R}^n)$, $\tau = 1/L$, 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

Gradient projection algorithm

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

For $E \in \mathcal{S}^{1,1}_{m,L}(\mathbb{R}^n)$ 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

Gradient projection algorithm

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

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

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

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

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

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

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

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

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?

Gradient Descent
Definition
Convergence analysis
Line search
Applications
Conclusion

Subgradient Method
Definition
Convergence Analysis
Applications

Gradient Projection
Projections
Definition
Convergence Analysis
Applications

Example Application: Solving a SUDOKU

In the 4×4 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 $\boldsymbol{u} \in \{0, 1\}^{4\times 4\times 4}$, where $\boldsymbol{u}_{i,j,k} = 1$ means $u_{i,j} = k$.

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

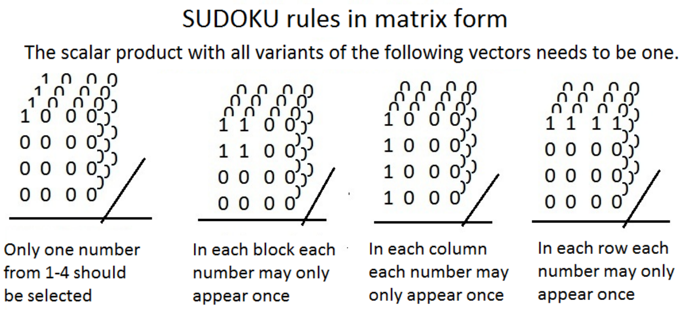
Find $\boldsymbol{u}$ with $\boldsymbol{u}_{i,j,k} \in \{0, 1\}$ subject to the above constraints!

Gradient Descent
Definition
Convergence analysis
Line search
Applications
Conclusion
Subgradient Method
Definition
Convergence Analysis
Applications
Gradient Projection
Projections
Definition
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

Example Application: Solving a SUDOKU

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



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

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

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

Example Application: Solving a SUDOKU

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

This means we may solve

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

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

Remarks:

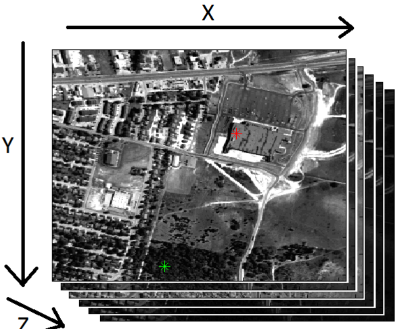
- Exact recovery guarantees (when is $\hat{\boldsymbol{u}}_{i,j,k} \in \{0, 1\}$) are an active field of research.
- Similar constructions can be done for many computer vision problems! Look for *labeling problems*, *segmentation*, *graph cuts*, or *functional lifting*.

Michael Moeller
Thomas Möllenhoff
Emanuel Laude

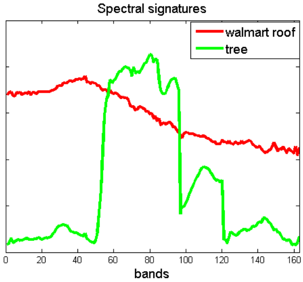
Gradient Descent
Definition
Convergence analysis
Line search
Applications
Conclusion
Subgradient Method
Definition
Convergence Analysis
Applications
Gradient Projection
Projections
Definition
Convergence Analysis
Applications

Example application: Unmixing and sparse recovery

Hyperspectral imagery



Hyperspectral cube



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

Example application: Unmixing and sparse recovery

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

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

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

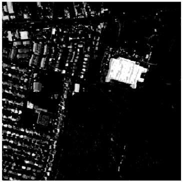
Michael Moeller
Thomas Möllenhoff
Emanuel Laude

Gradient Descent
Definition
Convergence analysis
Line search
Applications
Conclusion
Subgradient Method
Definition
Convergence Analysis
Applications
Gradient Projection
Projections
Definition
Convergence Analysis
Applications

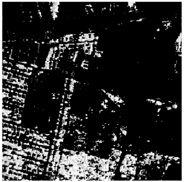
Example application: Unmixing and sparse recovery

color image illustration

endmember "road"



endmember "roof"



endmember "trees"

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