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

Michael Moeller
Thomas Möllenhoff
Emanuel Laude

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.

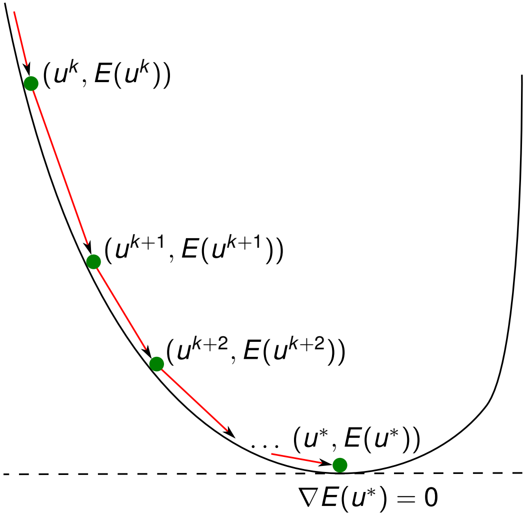
We start making our life easier:

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

Michael Moeller
Thomas Möllenhoff
Emanuel Laude

Descent methods

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



Descent methods

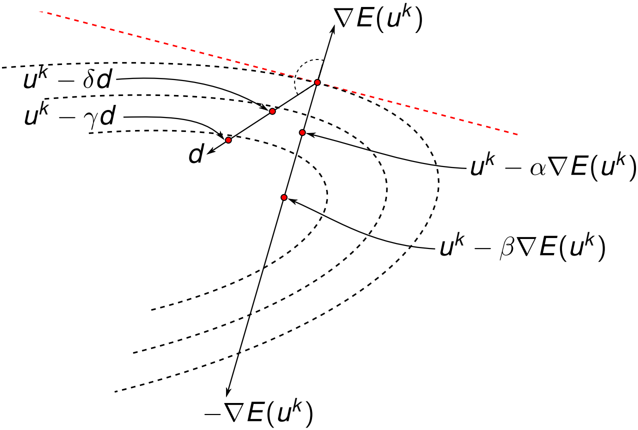
- Suppose we are at a point $u^k \in \mathbb{R}^n$ where $\nabla E(u^k) \neq 0$
- Consider the ray $u(\tau) = u^k + \tau d$ for some direction $d \in \mathbb{R}^n$
- Taylor expansion for E along ray

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

- 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

Descent methods



Descent methods

- 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 $\gamma : I \to \mathbb{R}^n$

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

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

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

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

is called *gradient descent*.

Philosophy:

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

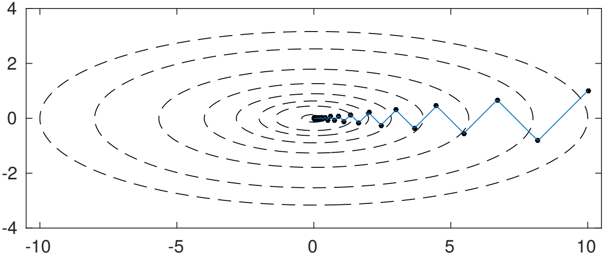
Choice of τ_k:

- $\tau_k = \tau$ 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

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

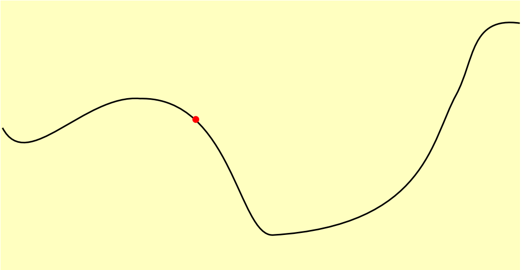
Lipschitz continuity

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*



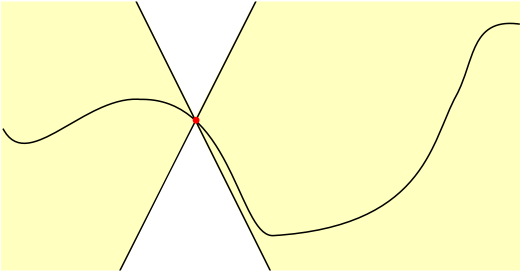
Lipschitz continuity

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*



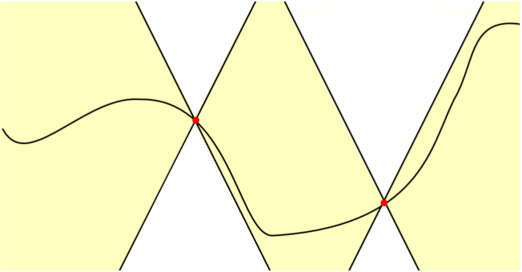
Lipschitz continuity

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

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

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!

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

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

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

Michael Moeller
Thomas Möllenhoff
Emanuel Laude

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

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

- For the minimum of the upper bound we have

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

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

Michael Moeller
Thomas Möllenhoff
Emanuel Laude

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

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

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.*

Michael Moeller
Thomas Möllenhoff
Emanuel Laude

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

Michael Moeller
Thomas Möllenhoff
Emanuel Laude

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 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 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

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

Proof: Board!

Michael Moeller
Thomas Möllenhoff
Emanuel Laude

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

Michael Moeller
Thomas Möllenhoff
Emanuel Laude

Worst-case complexities

- First order method:

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

- 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

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:

$$
\begin{aligned}
&\tau_k \leftarrow 1 \\
&\text{while } E\left(u^k - \tau_k \nabla E(u^k)\right) > E(u^k) - \alpha\tau_k \left\|\nabla E(u^k)\right\|^2 \\
&\qquad \tau_k \leftarrow \beta\tau_k \\
&\text{end}
\end{aligned}
$$

- 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

Image denoising

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

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

$$u^* \in \underset{u \in \mathbb{R}^N}{\text{argmax}}\; p(u|f) = \underset{u \in \mathbb{R}^N}{\text{argmax}}\; \frac{p(f|u)p(u)}{p(f)}$$

Image denoising

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

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

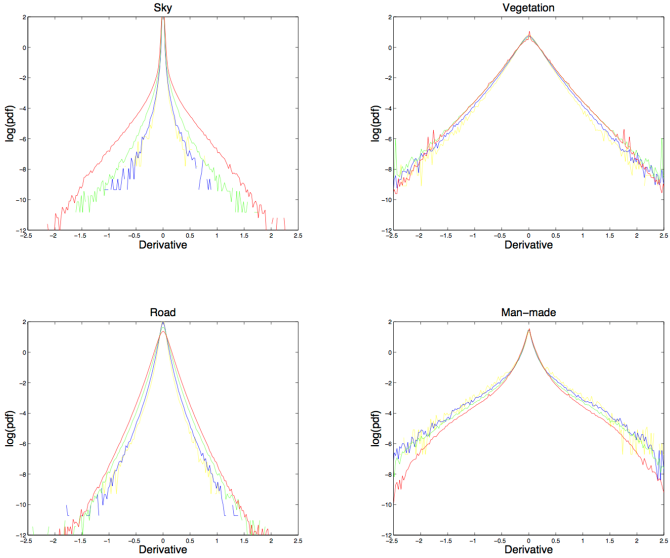
- Impose prior distribution on image gradient $Du \in \mathbb{R}^{2N}$

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

- Natural image statistics suggest the choice

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

Natural image statistics [6]



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

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 :-)

Image denoising

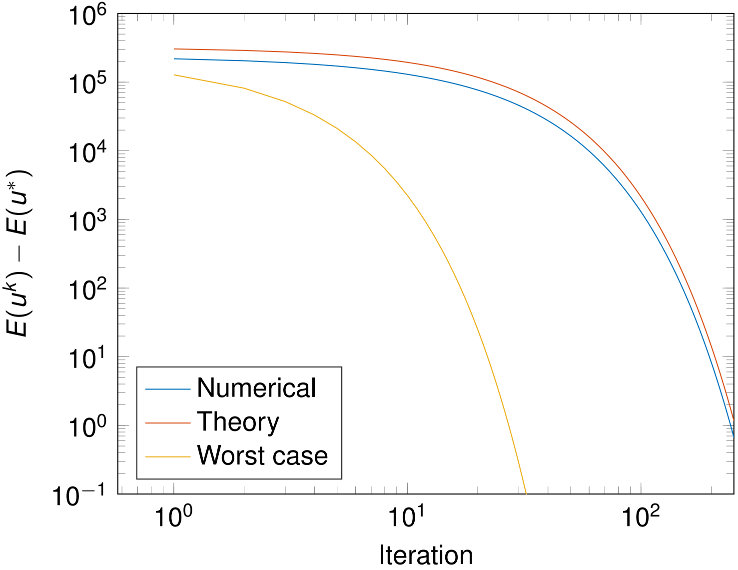
Evolution to global optimum via gradient descent

$\varepsilon = 0.1$

$\varepsilon = 0.01$

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

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



Convergence, backtracking line search

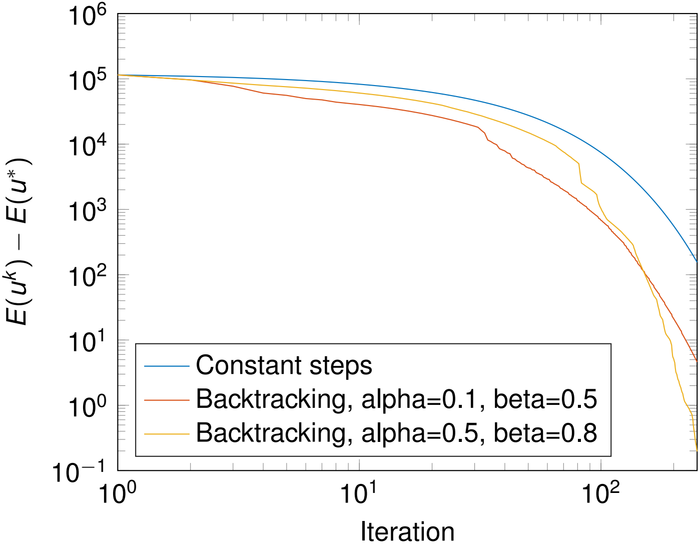
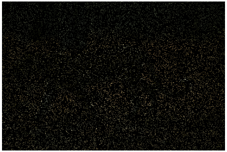
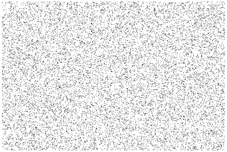


Image inpainting



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



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

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

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

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

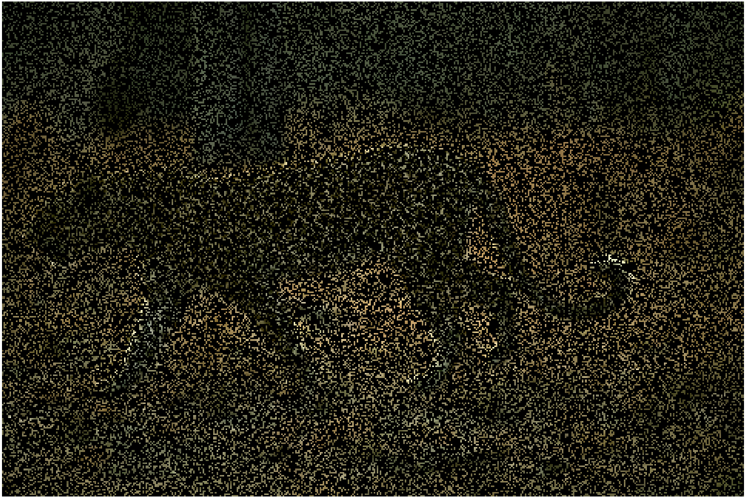
Image Inpainting

50% missing pixels

Michael Moeller
Thomas Möllenhoff
Emanuel Laude

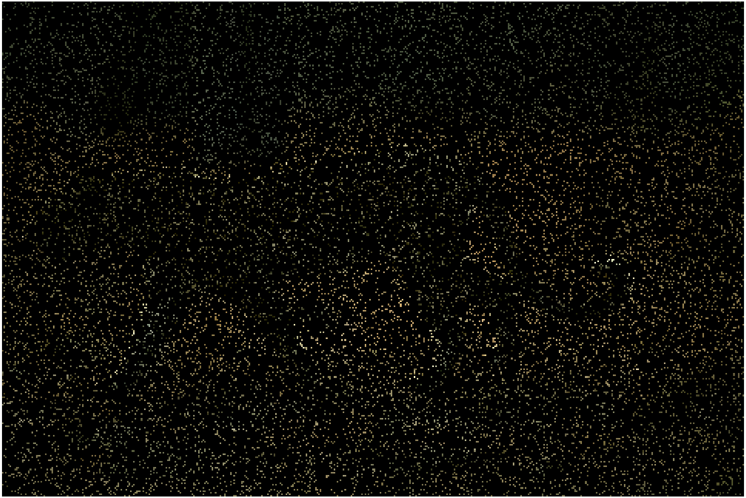
50% missing pixels

70% missing pixels



70% missing pixels

90% missing pixels

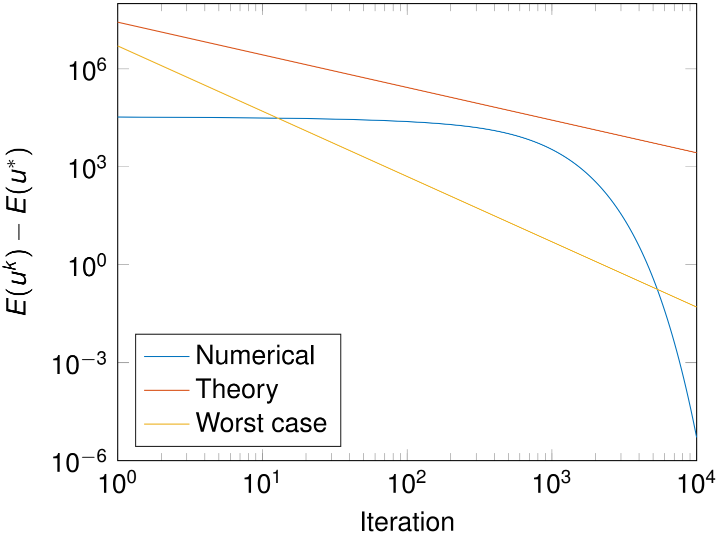


Michael Moeller
Thomas Möllenhoff
Emanuel Laude

90% missing pixels

Michael Moeller
Thomas Möllenhoff
Emanuel Laude

Convergence, $\tau = 1/L$



Michael Moeller
Thomas Möllenhoff
Emanuel Laude

Fast optimization challenge I

- Minimize the inpainting energy

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

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

$$\frac{E(u^k) - E(u^*)}{E(u^0)} < \delta$$

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

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/

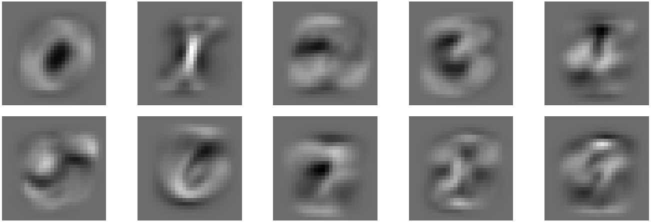
Multinomial logistic regression

- 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)

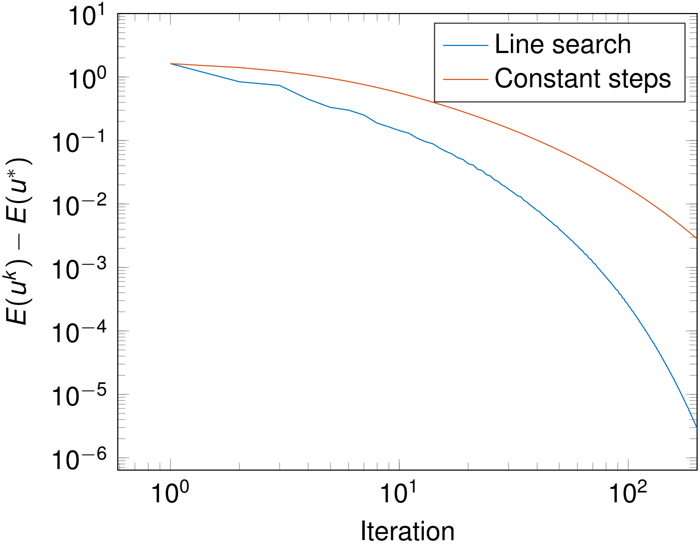
Multinomial logistic regression



- 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

Multinomial logistic regression



Michael Moeller
Thomas Möllenhoff
Emanuel Laude

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, ...

Michael Moeller
Thomas Möllenhoff
Emanuel Laude