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

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.

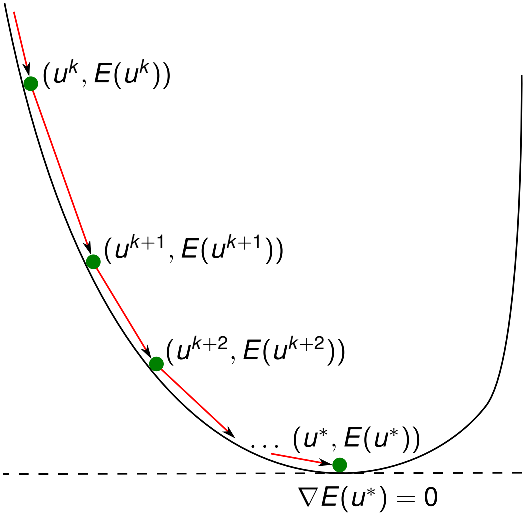
We start making our life easier:

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

Michael Moeller
Thomas Möllenhoff
Emanuel Laude

Descent methods

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



Michael Moeller
Thomas Möllenhoff
Emanuel Laude

Descent methods

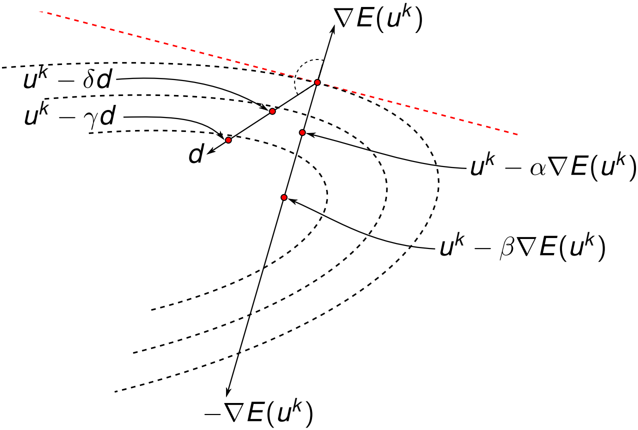
- Suppose we are at a point $u^k \in \mathbb{R}^n$ where $\nabla E(u^k) \neq 0$
- Consider the ray $u(\tau) = u^k + \tau d$ for some direction $d \in \mathbb{R}^n$
- Taylor expansion for E along ray

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

- 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

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

- 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

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

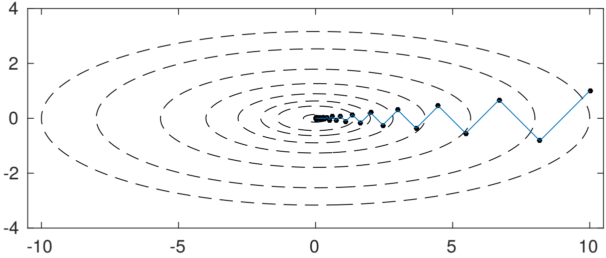
Choice of τ_k:

- $\tau_k = \tau$ 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

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

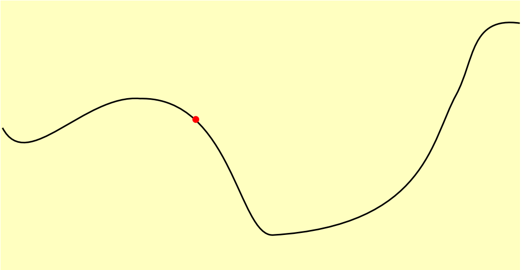
Lipschitz continuity

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*



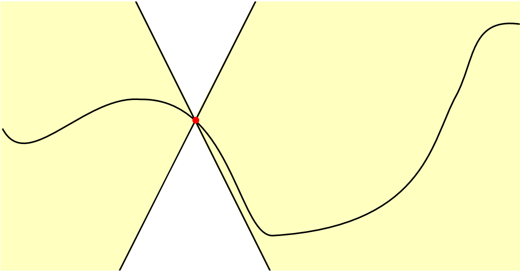
Lipschitz continuity

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*



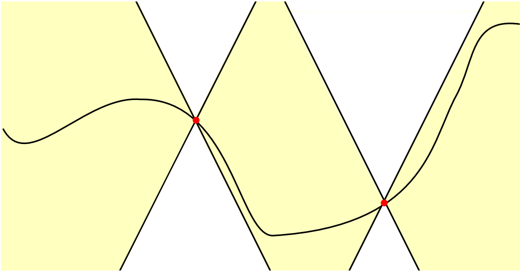
Lipschitz continuity

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*



Michael Moeller
Thomas Möllenhoff
Emanuel Laude

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

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

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!

Michael Moeller
Thomas Möllenhoff
Emanuel Laude

Lipschitz continuity

Michael Moeller
Thomas Möllenhoff
Emanuel Laude

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).

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

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

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

- For the minimum of the upper bound we have

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

Michael Moeller
Thomas Möllenhoff
Emanuel Laude

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

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: $E \in \mathcal{S}^1_m(\mathbb{R}^n)$
- We will also consider the classes $\mathcal{S}^{k,l}_{m,L}(\mathbb{R}^n)$

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

Michael Moeller
Thomas Möllenhoff
Emanuel Laude

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

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

- 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