Convex Optimization and Modeling

Unconstrained minimization

7th lecture, 26.05.2010

Jun.-Prof. Matthias Hein

Program of today - Unconstrained Minimization

Descent Methods:

- strongly convex functions,
- descent methods,
- stopping criteria,
- the condition number and its influence,
- convergence analysis of gradient descent.

Steepest Descent:

- steepest with respect to what ?
- convergence analysis

Newton method:

- Newton's method
- convergence analysis of Newton
- self-concordant functions

Unconstrained Minimization:

- minimization: $f : \mathbb{R}^n \to \mathbb{R}$,
- f convex and $f \in C^2(\Omega) \implies$ where $\operatorname{dom} f = \Omega$ open,
- for **global** convergence analysis: strongly convex function.

What about non-convex functions ?

- convergence to global optimum not guaranteed,
- convergence to **local optimum** can be proven,
- analysis is basically the same !

Important note:

local/global minima could be at the boundary of $\operatorname{dom} f$! We will not treat this case here (usually $\operatorname{dom} f = \mathbb{R}^d$).

What do we do ?

- let x^* be global optimum of f,
- find iterative sequence x^k such that

$$f(x^k) \xrightarrow{k \to \infty} f(x^*) = p^*.$$

- for convex functions a necessary and sufficient condition for a global minimum x^* is given by

$$\nabla f(x^*) = 0.$$

$\Rightarrow$ iterative method for solving the equation $\nabla f = 0$.

Sequence $f(x^k)$ $|f(x^k) - p^*| \leq r_k$

- $r_k = \frac{1}{k}$, sub-linear, ε-solution in $\frac{1}{\varepsilon}$-steps
  Example: $\varepsilon = 10^{-15}$, 10^{15} steps

- $r_k = \frac{1}{k^2}$, sub-linear, ε-solution in $\frac{1}{\sqrt{\varepsilon}}$-steps
  Example: $\varepsilon = 10^{-15}$, $\approx 10^7$ steps

- $r_k = \beta^k$ for $\beta \in (0, 1)$, linear, ε-solution in $\frac{\log \varepsilon}{\log \beta}$ steps
  Example: $\varepsilon = 10^{-15}$, $\beta = 0.95$, 674 steps

- $r_k = \beta^{2^k}$ for $\beta \in (0, 1)$, quadratic, ε-solution in $\log\left(\frac{\log \varepsilon}{\log \beta}\right)$ steps
  Example: $\varepsilon = 10^{-15}$, $\beta = 0.95$, 7 steps

Initial starting point and initial level set

Assumptions:

1. the starting point x^0 lies in $\mathrm{dom}\, f$,
2. the sublevel set $S = \{x \in \mathrm{dom}\, f \mid f(x) \le f(x^0)\}$ is closed.

Reminder:

- S is closed when f is closed.
- if $\mathrm{dom}\, f = \mathbb{R}^n$ it is sufficient for f being closed, that f is continuous.

Descent Method

Requirement: f continuously differentiable.

Steps:

- find direction d^k at current point x^k, so that

$$\left\langle d, \nabla f(x^k) \right\rangle < 0, \quad \textbf{descent direction}.$$

- find a suitable step size.

Require: an initial starting point x^0.

1: **repeat**
2: find a descent direction d^k.
3: **Line Search:** choose a step size α^k.
4: **UPDATE:** $x^{k+1} = x^k + \alpha^k d^k$.
5: **until** stopping criterion is satisfied.

Motivation:

Lemma 1. *Let $\Omega \subseteq \mathbb{R}^n$ and suppose that f is $C^2(\Omega)$. Then let*

$$x^{k+1} = x^k + \alpha^k d^k, \qquad \textit{where} \qquad \left\langle d^k, \nabla f(x^k) \right\rangle < 0.$$

Then for sufficiently small $\alpha > 0$ one has

$$f(x^{k+1}) < f(x^k).$$

Proof. A first-order Taylor expansion of f at x^k yields,

$$f(x^{k+1}) = f(x^k) + \alpha \left\langle \nabla f(x^k), d^k \right\rangle + \alpha^2 \left\langle d^k, Hf|_z d^k \right\rangle.$$

Now, let $C = \sup_{z \in [x^k, x^{k+1}]} \langle d^k, Hf|_z d^k \rangle / \left\| d^k \right\|^2$. By assumption $\langle \nabla f(x^k), d^k \rangle < 0$. Then for $0 < \alpha < | \langle \nabla f(x^k), d^k \rangle | / (C \left\| d^k \right\|^2)$ we get

$$\left\langle \nabla f(x^k), d^k \right\rangle + \alpha \left\langle d^k, Hf|_z d^k \right\rangle \leq \left\langle \nabla f(x^k), d^k \right\rangle + \alpha C \left\| d^k \right\|^2 < 0.$$

Different choices for the descent direction:

Most descent directions are defined as

$$d^k = -D^k \nabla f(x^k),$$

where D^k is a positive definite matrix. Then

$$\left\langle \nabla f(x^k), d^k \right\rangle = -\left\langle \nabla f(x^k), D^k \nabla f(x^k) \right\rangle < 0.$$

- $D^k = \mathbb{1}$: gradient or steepest descent $d^k = -\nabla f(x^k)$,
- $D^k = (H f(x^k))^{-1}$: gives Newton's method,
- $D^k = \mathrm{diag}(\gamma)$, where $\gamma_i > 0$, diagonal approx. of Newton's method.
- $D^k = (\tilde{H} f(x^k))^{-1}$, where $\tilde{H} f$ is a discretized (finite difference) approximation of the true Hessian at x^k. This is used if either the Hessian can not be computed analytically or if it is too expensive.

How to choose the stepsize ?

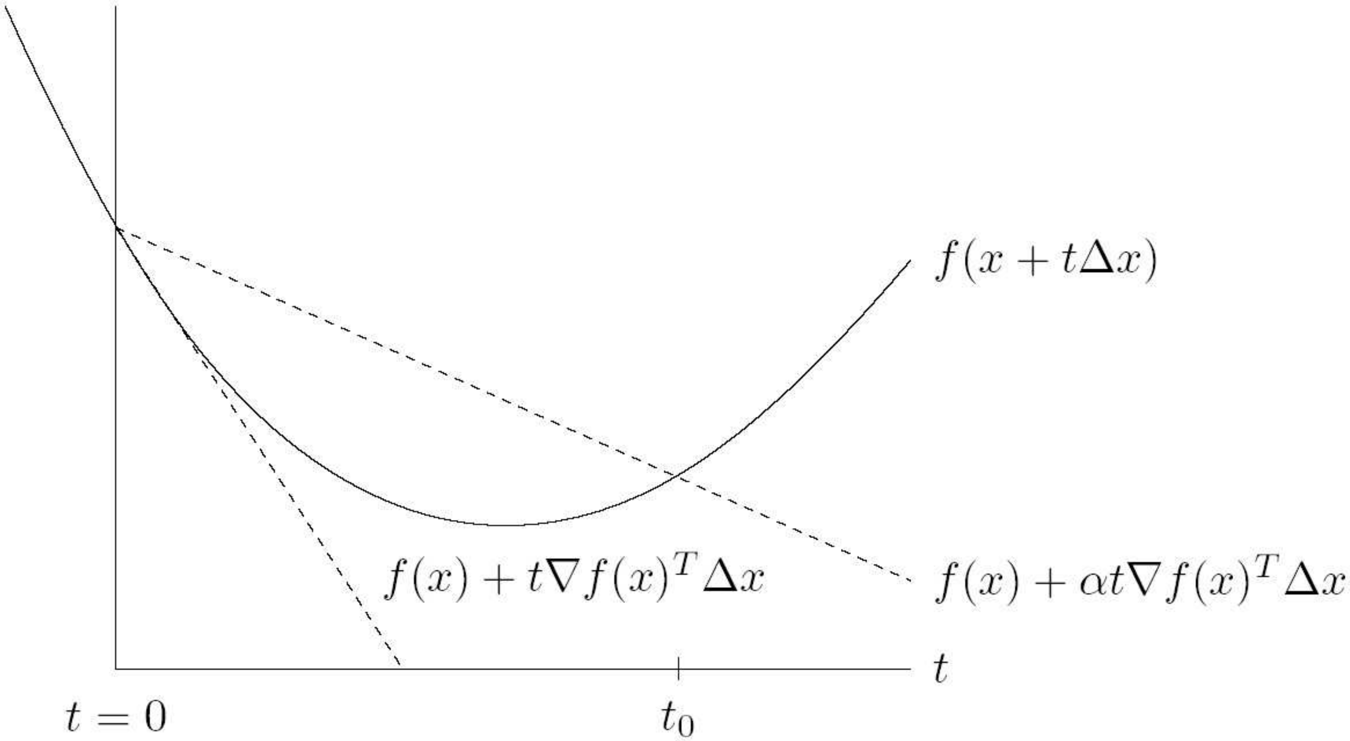
Different choices for the stepsize selection:

- **exact selection:** choose $\alpha^k = \arg\min_{\gamma \geq 0} f(x^k + \gamma d^k)$,
- **limited exact selection:** $\alpha_k = \arg\min_{\gamma \in [0,s]} f(x^k + \gamma d^k)$, for some $s > 0$.
- **Armijo rule or backtracking line search:** One chooses $\beta \in (0,1)$ and $\sigma \in (0,1)$ and $s > 0$. Then the stepsize α^k is defined as $\alpha^k = \beta^m s$, where m is the first non-negative integer such that

$$f(x^{k+1}) - f(x^k) = f(x^k + \beta^m s d^k) - f(x^k) \leq \sigma \beta^m s \left\langle \nabla f(x^k), d^k \right\rangle .$$

Note, that $\langle \nabla f(x^k), d^k \rangle < 0$ so that the stepsize is chosen such that $f(x^{k+1}) - f(x^k) < -K$ for $K > 0 \Longrightarrow$ sufficiently large descent per iteration.

Motivation for Armijo rule



first order approximation at x^k: $f(x^k) + \langle \nabla f(x^k), d^k \rangle$. The Armijo rule:

$$f(x^k) + \left\langle \nabla f(x^k), d^k \right\rangle \; < \; f(x^k) + \alpha t \left\langle \nabla f(x^k), d^k \right\rangle .$$

Since $\alpha < 1$ there will exist a stepsize t which fulfills the condition.

Strongly convex functions: needed for convergence analysis,

Definition 1. *A twice differentiable convex function f is said to be* ***strongly convex*** *if there exists $m > 0$ such that*

$$H f(x) \succeq m \mathbb{1}, \qquad \forall\, x \in \mathrm{dom}\, f.$$

$\Longrightarrow$ ensures that the global optimum of f is unique.

Lemma 1. *Let f be a strongly convex function. Then for all $x, y \in \mathrm{dom}\, f$,*

$$f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle + \frac{m}{2} \|y - x\|^2 .$$

Proof. A second-order Taylor expansion of f yields that for all $y, x \in \mathrm{dom}\, f$,

$$f(y) = f(x) + \langle \nabla f(x), y - x \rangle + \frac{1}{2} \langle y - x, H f(z) y - x \rangle ,$$

for some $z = \theta x + (1 - \theta) y$ with $\theta \in [0, 1]$. Using the property of a strongly convex function $\langle w, H f(z) w \rangle \geq m \|w\|^2$ we get directly the result. $\square$

Proposition 1. *Let f be a strongly convex function. Denote by p^* the global minimum of f attained at x^*. Then we have*

$$\|\nabla f\|_2^2 \leq 2m\varepsilon \quad \Longrightarrow \quad f(x) - p^* \leq \varepsilon,$$

and it holds $\|x - x^*\| \leq \frac{2}{m} \|\nabla f\|_2$.

Proof. We have: $f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle + \frac{m}{2} \|y - x\|^2$.
Minimizing over y on the rhs yields: $\nabla f + m(y - x) \implies y = x - \frac{1}{m}\nabla f(x)$.
Minimizing over both sides of the above inequality yields

$$p^* \geq f(x) - \frac{1}{2m} \|\nabla f\|^2 \quad \Longrightarrow \quad f(x) - p^* \leq \frac{1}{2m} \|\nabla f\|^2,$$

For the second result plug the optimal point $y = x^*$ into the above inequality

$$p^* = f(x^*) \geq f(x) + \langle \nabla f(x), x^* - x \rangle + \frac{m}{2} \|x^* - x\|^2$$
$$\geq f(x) - \|\nabla f\| \|x^* - x\| + \frac{m}{2} \|x^* - x\|^2.$$

Stopping criterion: $\|\nabla f\|_2^2 \leq 2m\varepsilon$.

$\Rightarrow$ similar bound for a C^2-function for all $x \in B(x^*, r)$, where $Hf(x) \succeq m\mathbb{1}$ for all $x \in B(x^*, r)$.

Lemma 1. *Let* $S = \{x \in \mathrm{dom}\, f \mid f(x) \leq f(x^0)\}$ *and* f *a strongly convex function, then*

- *the sublevel sets of* f *contained in* S *are bounded,*
- S *itself is bounded,*
- *there exists a constant* M *such that* $Hf \preceq M\mathbb{1}$.
- *for all* $x, y \in S$, $f(y) \leq f(x) + \langle \nabla f(x), y - x \rangle + \frac{M}{2} \|y - x\|^2$,
- $p^* \leq f(x) - \frac{1}{2M} \|\nabla f\|_2^2$.

Definition 2. *The* ***condition number*** $\kappa(A)$ *of a matrix* $A \in \mathbb{R}^{m\times n}$ *is defined as*

$$\kappa(A) = \frac{\sup_{\|x\|=1} \|Ax\|}{\inf_{\|x\|=1} \|Ax\|}.$$

For a matrix A *of full rank and the Euclidean norm* $\|\cdot\|_2$*, we have*

$$\kappa(A) = \frac{\sigma_{\max}(A)}{\sigma_{\min}(A)},$$

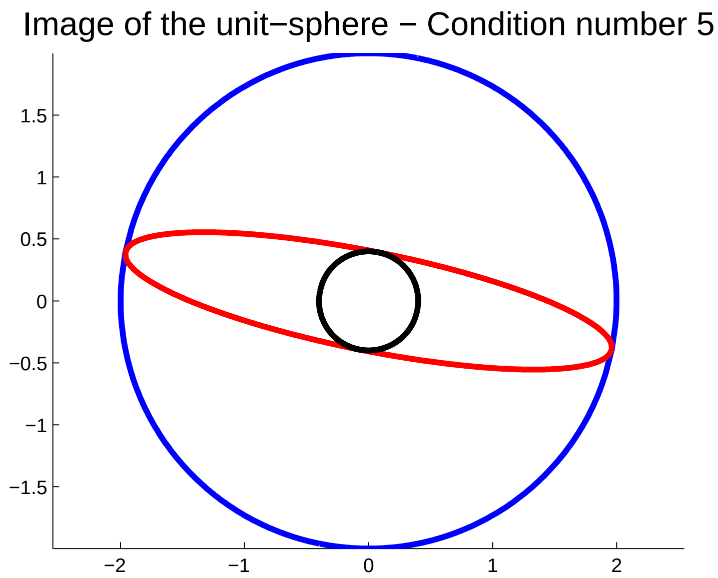
where $\sigma_{\min}, \sigma_{\max}$ *are the smallest and largest singular values of* A*. If* A *has full rank and is symmetric, positive definite we get*

$$\kappa(A) = \frac{\lambda_{\max}(A)}{\lambda_{\min}(A)},$$

where $\lambda_{\min}, \lambda_{\max}$ *are the smallest and largest eigenvalues of* A*.*

Condition Number II

Interpretation: condition number characterizes distortion of the unit-sphere under the matrix A.



The condition number can be seens as measuring the distortion of the unit sphere under the mapping of A. The higher the condition number the more elongated become the level sets of the second-order approximation.

Convergence speed

Definition 3. *Let x^k, $k \in \mathbb{N}$ with $x^k \in \mathbb{R}$ be a convergent sequence with limit x^*. Then x^k* ***converges with order*** *p if there exists a $\mu \in (0,1)$ such that*

$$\lim_{k\to\infty} \frac{|x^{k+1} - x^*|}{|x^k - x^*|^p} = \mu.$$

Remarks:

- If $p = 1$ we have **linear convergence**. If $\mu = 0$ for $p = 1$ we say that x^k converges **superlinearly**, whereas if the limit does not hold for any $\mu < 1$ then we say that x^k converges **sublinearly**.
- If $p = 2$ we have **quadratic convergence**.

Proof idea for the convergence analysis:

convergence analysis for the gradient descent method with exact line search.

The basic steps in the proof are,

- we derive a lower bound on the stepsize taken by the exact line search,
- this yields an upper bound on the difference $f(x^{k+1}) - p^*$ in term of $f(x^k) - p^*$.

Proposition 2. *Let f be strongly convex with*

$$m\mathbb{1} \preceq Hf(x) \preceq M\mathbb{1}, \quad \forall\, x \in \operatorname{dom} f.$$

The gradient descent method with exact line search fulfills,

$$f(x^{k+1}) - p^* \ \leq \ \left(1 - \frac{m}{M}\right)(f(x^k) - p^*),$$

so that with $c = 1 - \frac{m}{M}$ the number of steps required for $f(x^k) - p^* \leq \varepsilon$ is

$$k \ \leq \ \frac{\log\left(\frac{f(x^0)-p^*}{\varepsilon}\right)}{\log\left(\frac{1}{c}\right)}.$$

Linear convergence of gradient descent II

Proof: Gradient descent, that is $x^{k+1} = x^k - t\nabla f(x^k)$ or $d^k = -\nabla f(x^k)$. The stepsize t is found by exact line search. We have,

$$f(x^k - t\nabla f) \le f(x^k) - t \left\|\nabla f(x^k)\right\|_2^2 + t^2 \frac{M}{2} \left\|\nabla f(x^k)\right\|_2^2 .$$

The exact line search minimizes the left-hand side with respect to t and gives $f(x^{k+1})$. The right hand side is minimized for $t^* = \frac{1}{M}$ and we get,

$$f(x^{k+1}) \le f(x^k - t^*\nabla f) = f(x^k) + \frac{1}{2M} \left\|\nabla f(x^k)\right\|_2^2 .$$

Subtraction of p^* from both sides yields

$$f(x^{k+1}) - p^* \le f(x^k) - p^* - \frac{1}{2M} \left\|\nabla f(x^k)\right\|_2^2 .$$

From a previous bound: $-\left\|\nabla f\right\|_2^2 \ge 2m(p^* - f(x))$ for all $x \in S$ and thus,

$$f(x^{k+1}) - p^* \le f(x^k) - p^* + \frac{m}{M}(p^* - f(x^k)) = \left(1 - \frac{m}{M}\right)(f(x^k) - p^*).$$

Convergence analysis for gradient descent with Armijo rule:

Proposition 3. *Let f be strongly convex with*

$$m\mathbb{1} \preceq Hf(x) \preceq M\mathbb{1}, \quad \forall\, x \in \operatorname{dom} f.$$

The gradient descent method with Armijo rule with parameters (α, β) fulfills,

$$f(x^{k+1}) - p^* \;\leq\; c(f(x^k) - p^*),$$

where $c = 1 - \alpha \min\{2m, \frac{\beta(1-\alpha)m}{M}\} < 1$.

Discussion:

- convergence determined by condition number $\frac{m^*}{M^*}$ at optimum x^*
- at least linear convergence (also empirical observation)
- empirically: Armijo rule vs. exact line search for stepsize selection makes minor difference

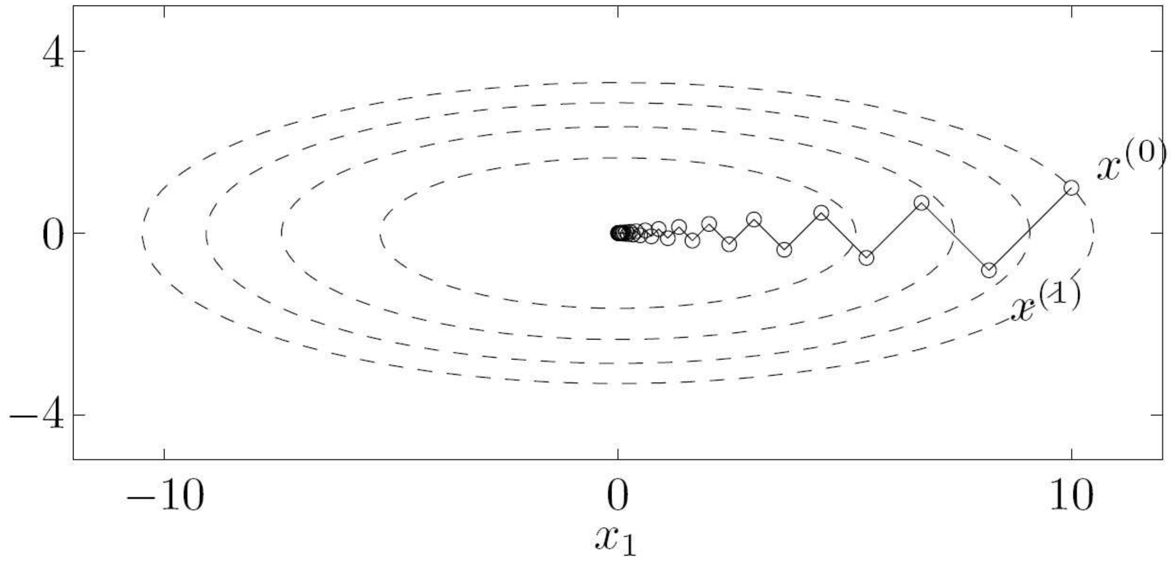
Discussion of gradient descent

Pro:

- very cheap computations
- can easily solve large-scale systems

Contra:

- sensitive to the condition number
- only linear convergence $\Longrightarrow$ slow !



First order Taylor:

$$f(x+v) \approx f(x) + \langle \nabla f, v \rangle \,.$$

What is the direction of steepest descent ?

First order Taylor:

$$f(x+v) \approx f(x) + \langle \nabla f, v \rangle \, .$$

What is the direction of steepest descent ?

Answer: depends on how we measure distances !

Definition 5. *The* ***normalized steepest descent direction*** d *with respect to the norm* $\|\cdot\|$ *is defined as*

$$d_{\mathrm{norm}} = \arg\min\{\langle \nabla f, v \rangle \;|\; \|v\| = 1\}.$$

Let $\|\cdot\|_*$ be the dual norm of $\|\cdot\|$. Then

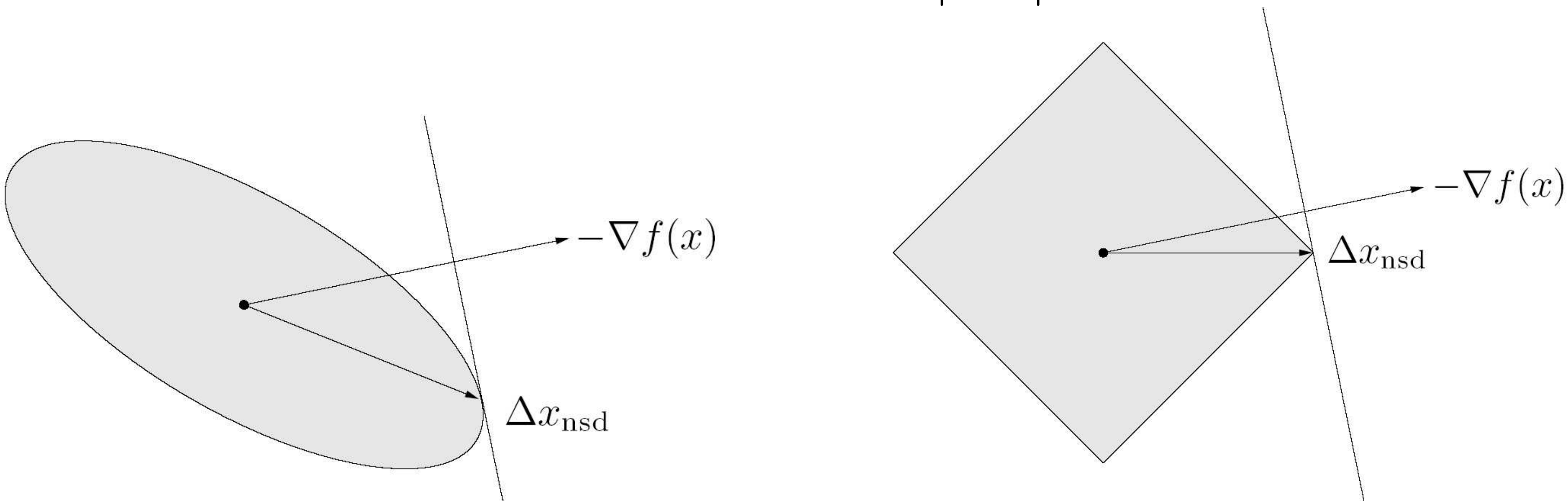
$$d_{\mathrm{unnorm}} = -\,\|\nabla f\|_*\, d_{\mathrm{norm}}.$$

Examples of steepest descent:

- Euclidean norm: $d_{\text{unnorm}} = -\nabla f$,
- Modified Euclidean norm: $\|z\|_P = \sqrt{\langle z, Pz\rangle} = \left\| P^{\frac{1}{2}} z \right\|$ where $P \in S^n_{++}$.

$$d_{\text{unnorm}} = -P^{-1}\nabla f.$$

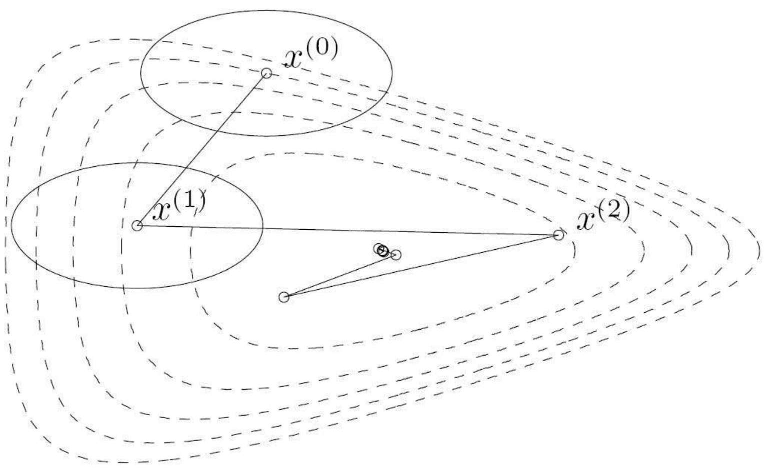
- L_1-norm: $d_{\text{unnorm}} = -\|\nabla f\|_\infty e_i$, where $\left|\frac{\partial f}{\partial x_i}\right| = \|\nabla f\|_\infty$.

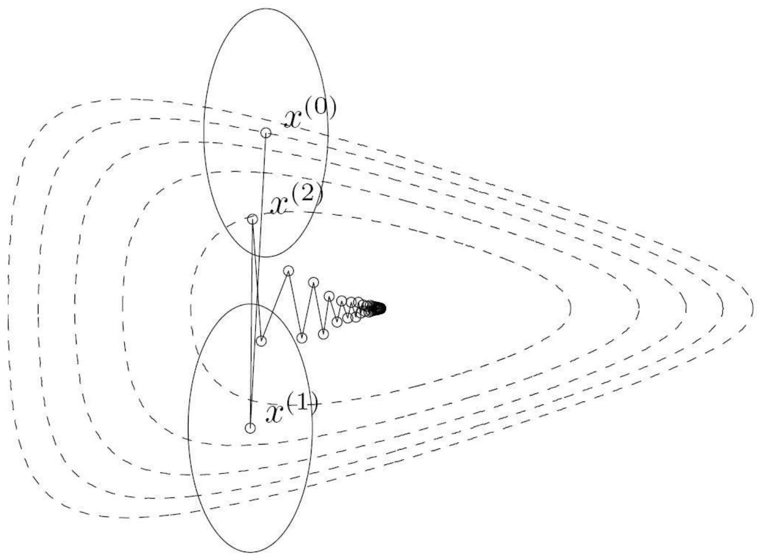


Left: descent direction for modified Euclidean norm. Right: for the L_1-norm.

Discussion:

- similar convergence proof: linear rate,
- find P such that the condition number becomes smaller
- ideally: $P \approx Hf(x^*) \Rightarrow$ condition number ≈ 1 at the optima !

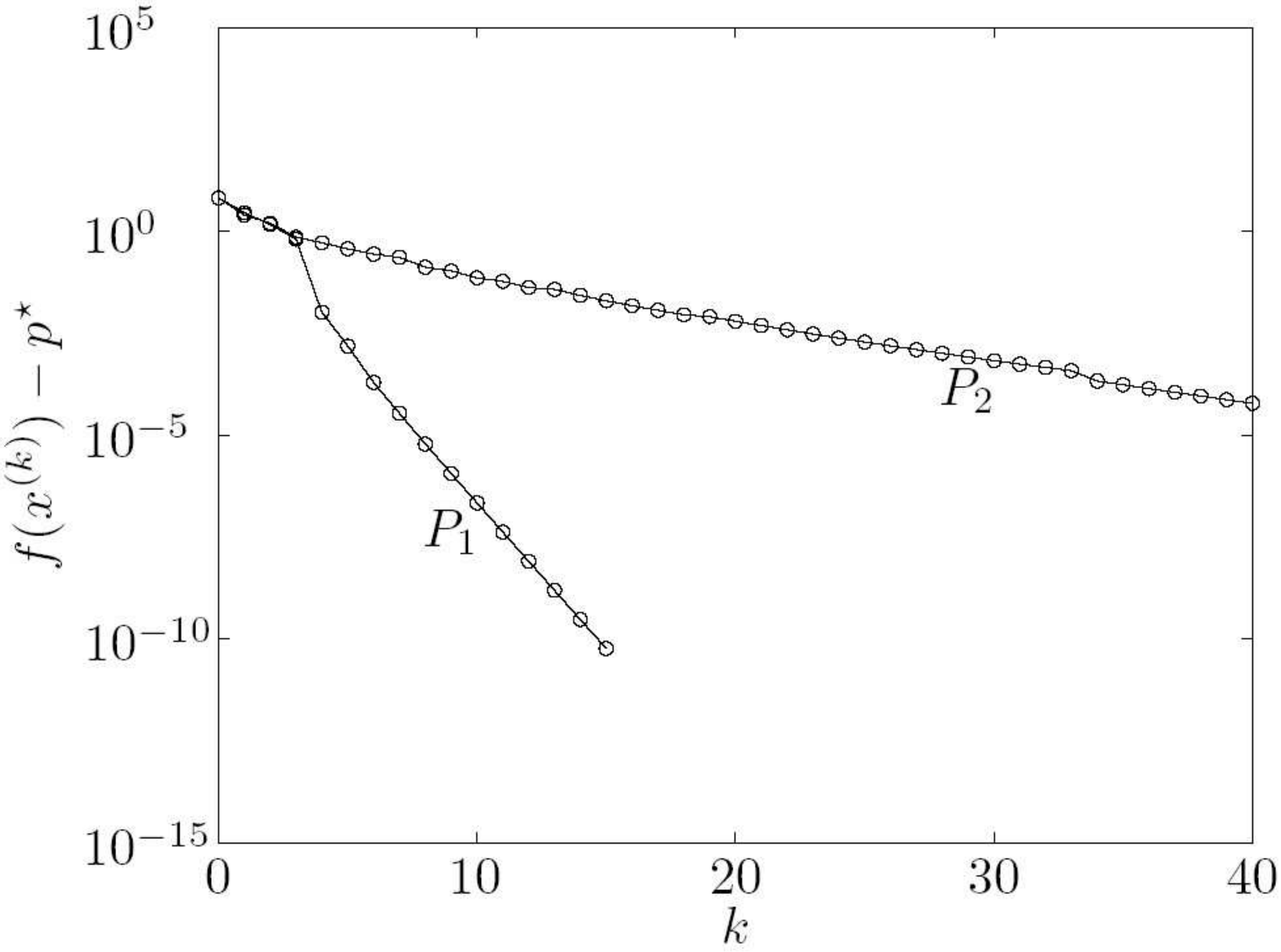




Two examples how the change of the norm/coordinates affects the convergence of gradient descent.

Steepest descent IV

Convergence rate:



Differences in convergence rates for two modified Euclidean norms.

Newton method

Descent direction:

$$d = -Hf(x)^{-1}\,\nabla f(x).$$

Motivation:

Minimization of second-order approximation

$$d = \arg\min_{v} \Big(f(x) + \langle \nabla f(x), v \rangle + \frac{1}{2}\,\langle v, Hf(x)v \rangle \Big).$$

Local coordinate change such that the condition number is minimal

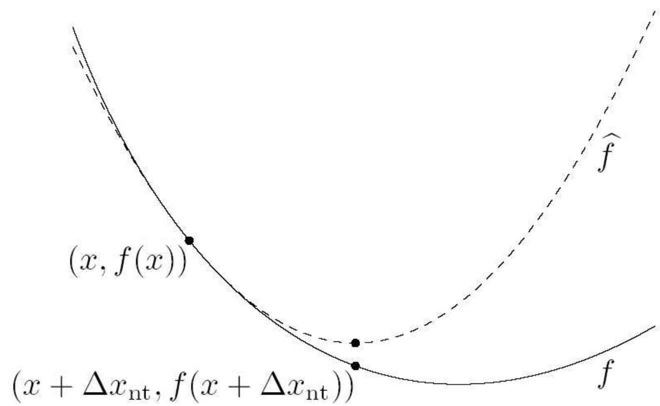
$$x \to x' = Hf^{\frac{1}{2}}x, \quad f(x'+v) = f(x') + \Big\langle Hf(x)^{-\frac{1}{2}}\nabla_x f, v \Big\rangle + \frac{1}{2}\,\langle v, v \rangle\,.$$

in new coordinates:

$$d' = -Hf(x)^{-\frac{1}{2}}\nabla_x f.$$

in old coordinates:

$$d = Hf(x)^{-\frac{1}{2}} d = -Hf^{-1}\nabla f.$$



The Newton method is affine invariant:

Let $A \in \mathbb{R}^{n \times n}$,where A has full rank.

Define: $f'(y) = f(Ay)$, with $x = Ay$, where y are new coordinates. We have

$$\nabla f'(y) = A^T \nabla f(x), \qquad H f'(y) = A^T H f(x) A.$$

and thus

$$\Big(H f'(y)\Big)^{-1} \nabla f'(y) = (A^T H f(x) A)^{-1} A^T \nabla f(x) = A^{-1} H f(x)^{-1} \nabla f(x).$$

which gives

$$d' = A^{-1} d \quad \text{or} \quad y + \alpha d' = A^{-1}(x + \alpha d).$$

Is the stepsize α also invariant with respect to affine transformations ?

Newton decrement: descent $d = -(Hf(x))^{-1}\nabla f$,

$$\lambda(x)^2 = \left\langle \nabla f(x), (Hf(x))^{-1}\nabla f \right\rangle = \langle d, Hf(x)d \rangle .$$

- $\hat{f}$ second order approximation of f at x, then

$$f(x) - \inf_y \hat{f}(y) = f(x) - \hat{f}(x+d) = \frac{1}{2}\lambda(x)^2.$$

- $\lambda(x)$ is affine invariant,
- $\lambda(x)$ is the norm of d in the modified Euclidean norm with $P = Hf(x)$
  $\Rightarrow$ $\lambda(x)$ can be used as an **affine invariant stopping criterion.**
- note that $\langle \nabla f(x), d\rangle = -\lambda(x)^2 \Rightarrow$ **stepsize selection is affine invariant !**

Newton's method:

Require: an initial starting point x^0.

repeat

compute the Newton step and decrement

$$d^k = -(Hf(x^k))^{-1}\nabla f(x^k), \quad \lambda(x^k)^2 = -\left\langle d^k, \nabla f(x^k)\right\rangle.$$

Line Search: choose a step size α^k with the Armijo rule.

UPDATE: $x^{k+1} = x^k + \alpha^k d^k$.

until $\lambda(x^k)^2 \leq 2\varepsilon$.

The stopping criterion is sometimes put directly after the computation of the Newton decrement.

Sketch of convergence analysis of Newton's method

Assumption: Lipschitz condition on Hf,

$$\|Hf(x) - Hf(y)\| \leq L\,\|x - y\|\,.$$

Two phases: $0 < \eta < \frac{m^2}{L}$

- **damped Newton phase:** $\left\|\nabla f(x^k)\right\|_2 \geq \eta$

$$\gamma > 0, \qquad f(x^{k+1}) - f(x^k) \leq -\gamma.$$

- **pure Newton phase:** $\left\|\nabla f(x^l)\right\|_2 \leq \eta$

$$\left\|\nabla f(x^{l+1})\right\|_2 \leq \frac{L}{2m^2}\left\|\nabla f(x^l)\right\|_2^2.$$

stepsize $\alpha^k = 1 \Rightarrow$ pure Newton step for $l \geq k$

$$f(x^l) - p^* \leq \frac{1}{2m}\left\|\nabla f(x^l)\right\|_2^2 \leq \frac{2m^3}{L^2}\Big(\frac{1}{2}\Big)^{2^{l-k+1}}.$$

Convergence analysis

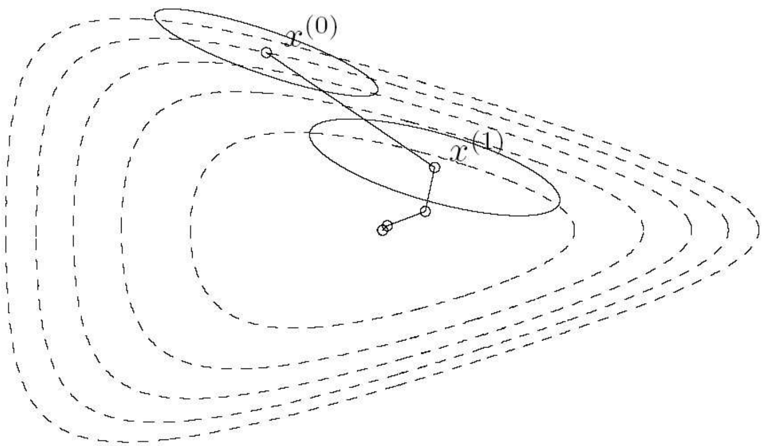
Two phases:

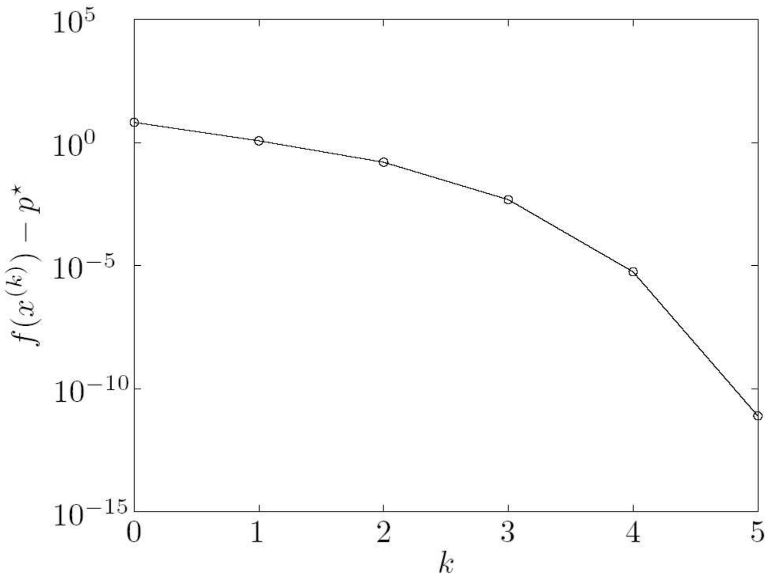
- **damped Newton phase:** linear convergence,
- **pure Newton phase:** quadratic convergence.

Required number of steps: for $f(x) - p^* \leq \varepsilon$,

$$k \leq \frac{f(x^{(0)}) - p^*}{\gamma} + \log_2 \log_2 \frac{2m^3}{L^2 \varepsilon}.$$

second term **grows extremely slow** $\Rightarrow$ can be seen as constant !





Newton's method

Pro:

- fast convergence of Newton's method,
- Newton's method is affine invariant,
- much less dependent on the choice of the parameters than gradient descent.

Contra:

- requires second derivative,
- does not scale easily to large problems if Hessian has no special structure (e.g. sparse, banded etc.) $\Longrightarrow$ one needs a fast way of solving

$$Hf(x)d = \nabla f.$$

Problems of classical convergence analysis

- depends on unknown constants $(m, L, \cdots)$,
- Newtons method is affine invariant but not the bound.

Convergence analysis via self-concordance (Nesterov and Nemirovski)

- does not depend on any unknown constants
- gives affine-invariant bound
- applies to special class of convex functions (self-concordant functions)
- developed to analyze polynomial-time interior-point methods for convex optimization

Self-concordance II

Self-concordant functions:

Definition 6. *A function $f : \mathbb{R} \to \mathbb{R}$ is* ***self-concordant*** *if*

$$|f'''(x)| \leq 2 f''(x)^{\frac{3}{2}}.$$

A function $f : \mathbb{R}^n \to \mathbb{R}$ is ***self-concordant*** *if $t \mapsto f(x + tv)$ is self-concordant for every $x, v \in \mathbb{R}^n$.*

Examples:

- linear and quadratic functions,
- negative logarithm $f(x) = -\log x$.

Properties:

- If f self-concordant, then also γf where $\gamma > 0$.
- If f is self-concordant then $f(Ax + b)$ is also self-concordant.

Convergence analysis for a strictly convex self-concordant function:

Two phases: $0 < \eta < \frac{1}{4}$, $\gamma > 0$,

- **damped Newton phase:** $\lambda(x^k) > \eta$,

$$\gamma > 0, \qquad f(x^{k+1}) - f(x^k) \le -\gamma.$$

- **pure Newton phase:** $\lambda(x^k) \le \eta$,

$$2\lambda(x^{k+1}) \le \Big(2\lambda(x^k)\Big)^2.$$

stepsize $\alpha^k = 1 \Rightarrow$ pure Newton step for $l \ge k$

$$f(x^l) - p^* \le \lambda(x^l)^2 \le \left(\frac{1}{2}\right)^{2^{l-k+1}}.$$

$\Longrightarrow$ **complexity bound only depends on known constants !**

$\Longrightarrow$ **does not imply that Newton's method works better for self-concordant functions !**