NTNU | Norwegian University of Science and Technology

TTK4135 – Lecture 13
Unconstrained optimization

Lecturer: Lars Imsland

Outline

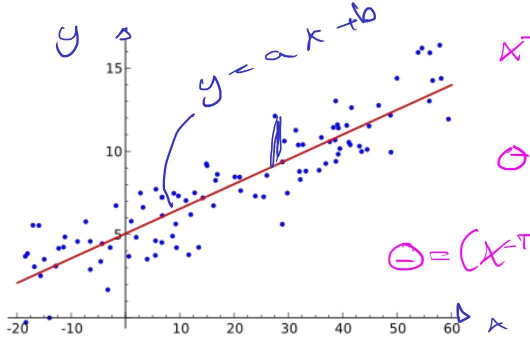
- Optimality conditions for unconstrained optimization
- Ingredients in gradient descent algorithms for unconstrained optimization
  - Descent directions (steepest descent, Newton, Quasi-Newton)
  - How far to walk in descent direction (line search, trust region)
  - Termination criteria
- Scaling

Reference: N&W Ch.2.1-2.2

Example: Machine Learning

$$\min_{a,b} \underbrace{\sum_{i=1}^{n} (y_i - a x_i - b)^2}_{(y - X\theta)^T (y - X\theta)}$$

- Learn, and make predictions, from data



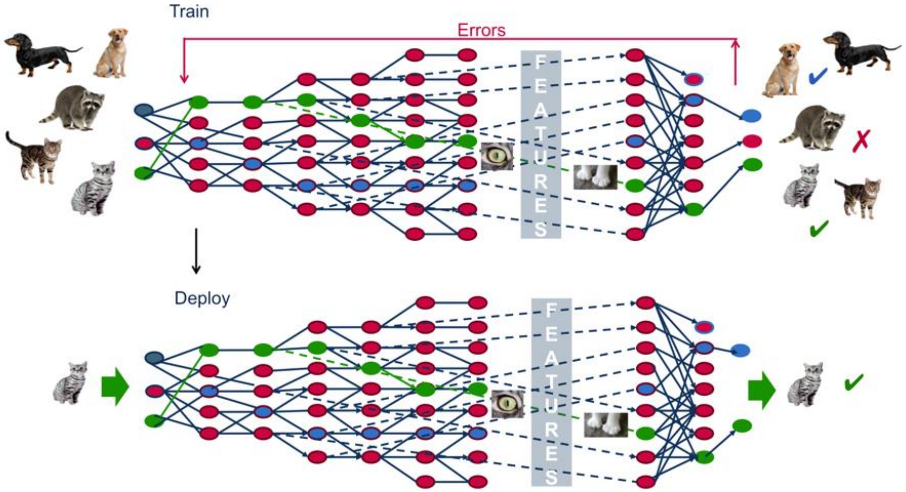
Wikipedia.org

$$X = \begin{bmatrix} x_1 & 1 \\ x_2 & 1 \\ \vdots & \vdots \end{bmatrix} \qquad \theta = \begin{bmatrix} a \\ b \end{bmatrix}$$

$$\theta = (X^T X)^{-1} X^T y$$

- Linear regression is the most basic ML algorithm, solved using optimization
  - Linear least squares: Explicit solution
  - Nonlinear least squares: Ch. 10, N&W

- In a similar fashion: ML, neural networks, deep learning etc. are "trained" using gradient descent algorithms
  - Gradient descent for unconstrained optimization is topic of Ch. 2-10, N&W

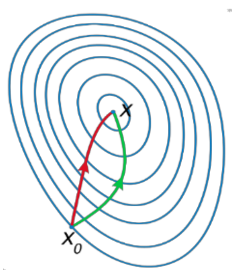


Learning goal Ch. 2, 3 and 6: Understand this slide

Line-search unconstrained optimization

$$\min_x f(x)$$

1. Initial guess x_0
2. While termination criteria not fulfilled
   a) Find descent direction p_k from x_k
   b) Find appropriate step length α_k ; set $x_{k+1} = x_k + \alpha_k p_k$
   c) $k = k+1$
3. $x_M = x^*$? (possibly check sufficient conditions for optimality)



A comparison of steepest descent and Newton's method. Newton's method uses curvature information to take a more direct route. (wikipedia.org)

Termination criteria:

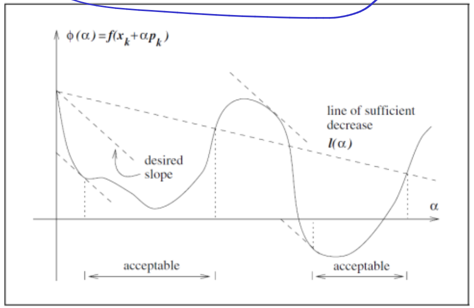
Stop when first of these become true:

- $\|\nabla f(x_k)\| \leq \epsilon$ (necessary condition)
- $\|x_k - x_{k-1}\| \leq \epsilon$ (no progress)
- $\|f(x_k) - f(x_{k-1})\| \leq \epsilon$ (no progress)
- $k \leq k_{\max}$ (kept on too long)

Descent directions:

- Steepest descent
  $p_k = -\nabla f(x_k)$
- Newton
  $p_k = -(\nabla^2 f(x_k))^{-1} \nabla f(x_k)$
- Quasi-Newton
  $p_k = -B_k^{-1} \nabla f(x_k)$
  $B_k \approx \nabla^2 f(x_k)$

Step length line search (Wolfe):



How to calculate derivatives – Ch. 8

How many iterations? (Convergence rates)

Unconstrained optimization

LP, QP

NLP

$$\min_{x \in \mathbb{R}^n} f(x) \text{ s.t. } \cancel{\begin{cases} c_i(x) = 0, & i \in \mathcal{E} \\ c_i(x) \geq 0, & i \in \mathcal{I} \end{cases}}$$

Now: $\mathcal{E} = \emptyset$, $\mathcal{I} = \emptyset$: Unconstrained opt.

$$\boxed{\min_{x \in \mathbb{R}^n} f(x)}$$

Note: $f(x)$ is "smooth"

$f \in C^1$ (or $f \in C^2$)

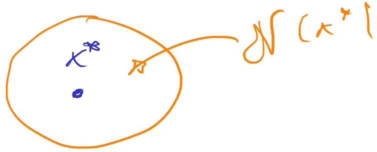
that is: ∇f exists (and $\nabla^2 f(x)$ exists)

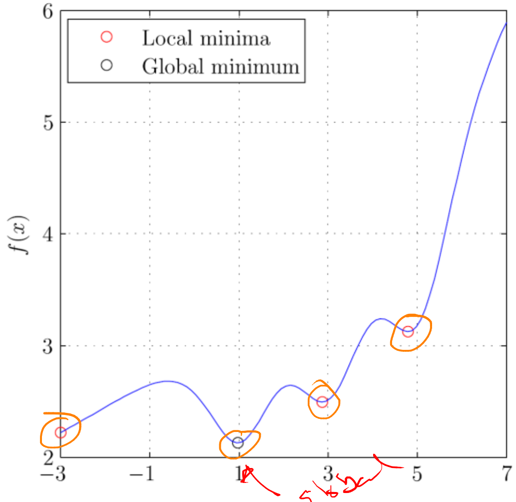
and they are continuous

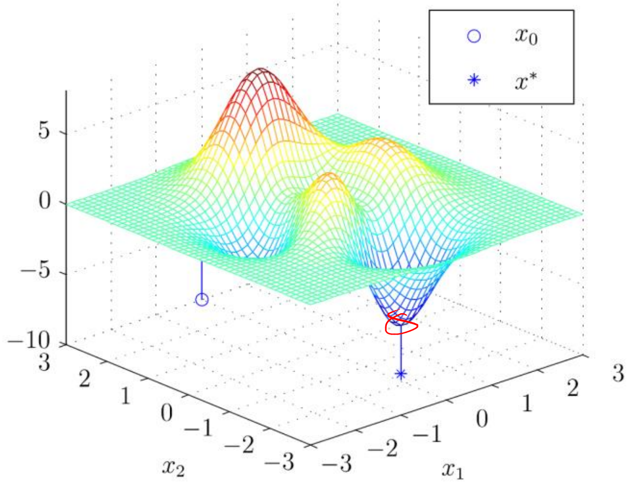
What is a solution? Local and global minimizers

x^* is a global solution: $f(x^*) \le f(x)$, for all x

x^* is a local solution: $f(x^*) \le f(x)$, for all $x \in \mathcal{N}(x^*)$

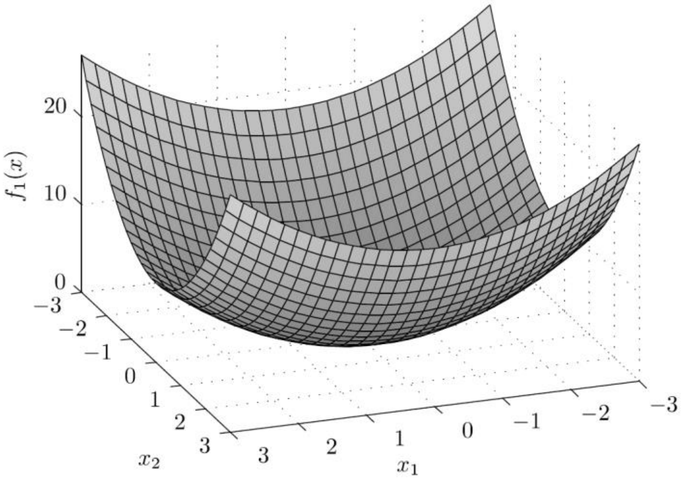




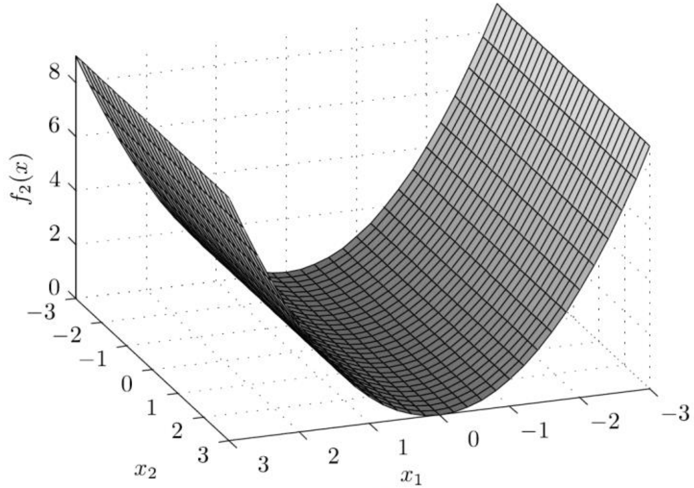


solutions

(Strict and non-strict optimizers)



$x^* = 0$ is a *strict minimizer.*



$x_1^* = 0$ is a *non-strict minimizer.*

Necessary condition for optimality

$$\min_x f(x)$$

Theorem 2.2: x^* local solution and $f \in C^1 \Rightarrow \nabla f(x^*) = 0$

Proof by contradiction: Assume x^* local solution and $\nabla f(x^*) \neq 0$.

- Select $p = -\nabla f(x^*) \Rightarrow p^T \nabla f(x^*) = -\|\nabla f(x^*)\|^2 < 0$.
- Since ∇f is continuous, there exists $T > 0$ s.t.

$$p^T \nabla f(x^* + tp) < 0, \text{ for all } t \in [0, T]$$

- Taylor: for any $\bar{t} \in [0, T]$

$$f(x^* + \underbrace{\bar{t}p}_{p}) = f(x^*) + \underbrace{\bar{t}p^T \nabla f(x^* + tp)}_{<0}, \quad \text{for some } t \in (0, \bar{t})$$

$$\Rightarrow f(x^* + \bar{t}p) < f(x^*)$$

$\Rightarrow x^*$ is not a local min. Contradiction! $\square$

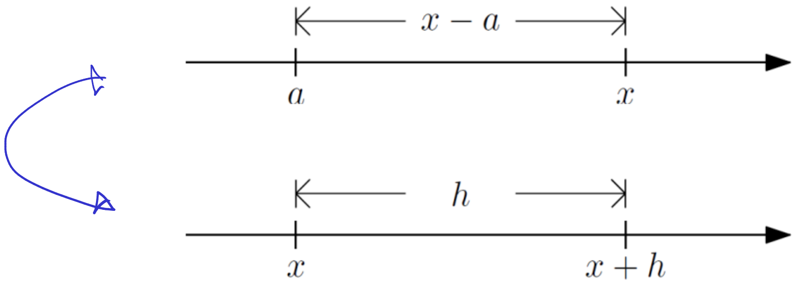
Taylor expansions

- From Calculus?

$$f(x) = f(a) + (x-a)f'(a) + \frac{(x-a)^2}{2}f''(a) + \cdots$$

- In this course:

$$\boxed{f(x+h) = f(x) + hf'(x) + \frac{h^2}{2}f''(x) + \cdots}$$



Taylor's theorem

$$f : \mathbb{R}^n \to \mathbb{R},\ p \in \mathbb{R}^n$$

- First order: If *f* is continuously differentiable,

$$\boxed{f(x+p) = f(x) + \nabla f(x+tp)^\top p, \quad \text{for some } t \in (0,1)}$$

- Second order: If *f* is twice continuously differentiable

$$\boxed{f(x+p) = f(x) + \nabla f(x)^\top p + \frac{1}{2} p^\top \nabla^2 f(x+tp)^\top p, \quad \text{for some } t \in (0,1)}$$

Sufficient conditions for optimality

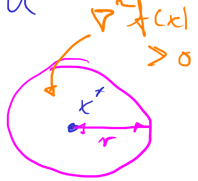
$+\ f(x) \in C^2$

Theorem 2.4: $\nabla f(x^*) = 0$ and $\nabla^2 f(x^*) > 0 \Rightarrow x^*$ strict local solution

Proof:

Note: $\underline{\nabla^2 f \text{ continuous}} \Rightarrow$ Exist $r > 0$ s.t. $\nabla^2 f(x) > 0$ for all

$x \in \{x \mid \|x - x^*\| < r\}$



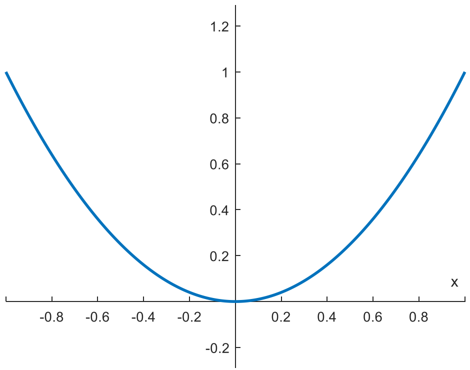
Taylor: For any $p \neq 0$, $\|p\| < r$

$$f(x^*+p) = f(x^*) + \underbrace{p^T \nabla f(x^*)}_{=0} + \frac{1}{2} \overbrace{p^T \nabla^2 f(x^*+tp)\, p}^{>0},$$

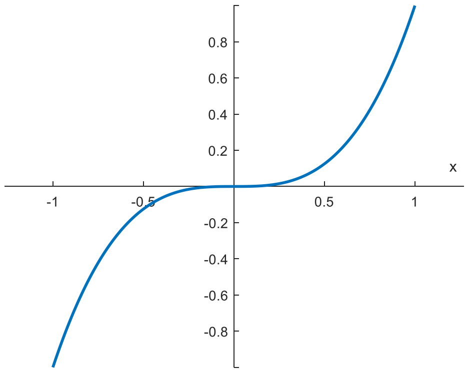
for some $t \in [0,1]$

$$\Rightarrow f(x^*+p) > f(x^*). \quad \square$$

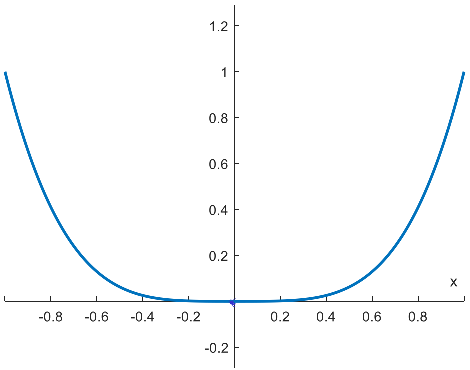
x^2



x^3



x^4



$$\nabla f(0) = 0$$
$$\nabla^2 f(0) > 0$$

$$\nabla f(0) = 0$$
$$\nabla^2 f(0) = 0$$

$$\nabla f(0) = 0$$
$$\nabla^2 f(0) = 0$$

General algorithm for solving $\min_x f(x)$

1) Initial guess x_0, $k = 0$
2) While termination criteria not fulfilled
   2a) Find descent direction p_k (for x_k)
   2b) Walk along p_k to x_{k+1}: $x_{k+1} = x_k + \alpha_k p_k$
   2c) $k = k+1$

   end
3) $x_M = x^*$ (?)

Termination criteria

Given small "tolerance" $\varepsilon > 0$:

0. ~~$\| x_k - x^* \| < \varepsilon$ or $| f(x) - f(x^*) | < \varepsilon$~~

1. $\| \nabla f(x_k) \| < \varepsilon$ (nec. cond.)

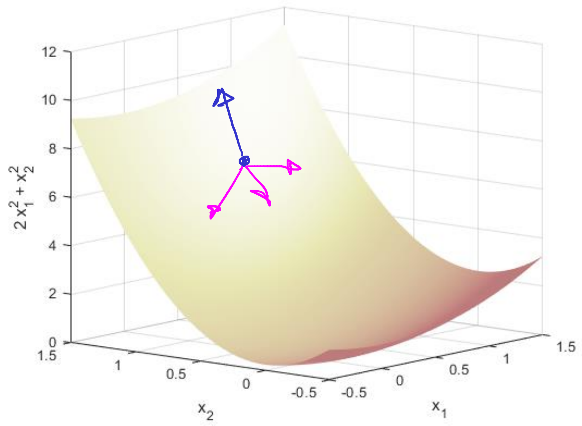
2. $\| x_k - x_{k-1} \| < \varepsilon$ (or $\| x_k - x_{k-1} \| < \varepsilon \| x_{k-1} \|$)

3. $| f(x_k) - f(x_{k-1}) | < \varepsilon$ (or $| f(x_k) - f(x_{k-1}) | < \varepsilon | f(x_{k-1}) |$)

4. $k > k_{max}$

In practice: Check all 1–4, terminate when first holds.

Descent (downhill) directions

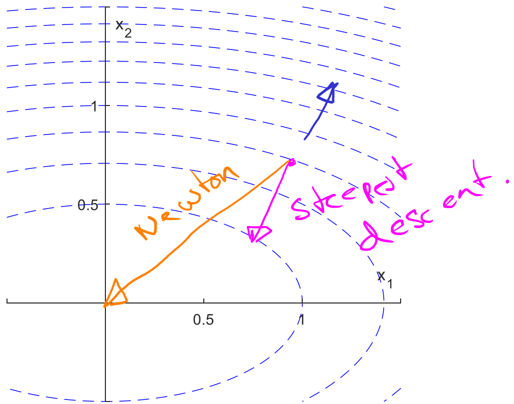


1. Steepest descent : $p = -\nabla f(x_k)$

2. Newton :

Approximate $f(x)$ around x_k

Taylor : $f(x_k+p) \approx \underbrace{f(x_k) + \nabla f(x_k)^T p + \frac{1}{2} p^T \nabla^2 f(x_k) p}_{:= m_k(p)}$

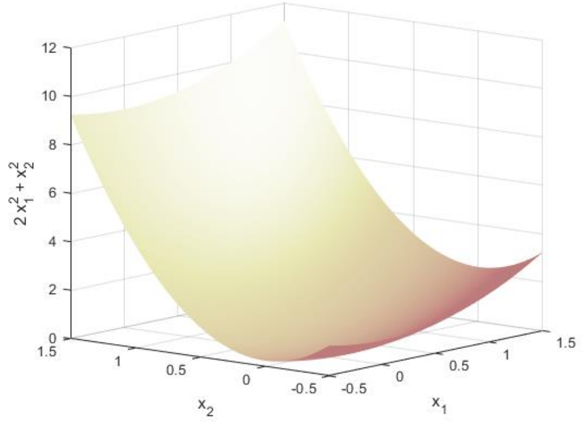


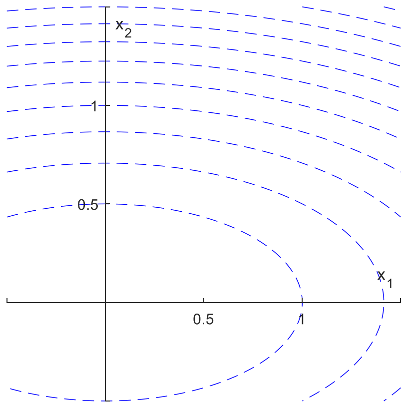
$\min_p \; m_k(p) \Rightarrow \nabla_p m_k(p) = 0$

$\Rightarrow \nabla f(x_k) + \nabla^2 f(x_k) p = 0$

Newton direction $\longrightarrow$ $\Rightarrow p = -\left[\nabla^2 f(x_k)\right]^{-1} \nabla f(x_k)$

Descent (downhill) directions





3. Quasi-Newton:

$$P_k = -B_k^{-1} \nabla f(x_k)$$

$$B_k \approx \nabla^2 f(x_k)$$

(later)

Quadratic approximation to objective function

$$f(x_k + p) \approx m_k(p) = f(x_k) + p^\top \nabla f(x_k) + \frac{1}{2} p^\top \nabla^2 f(x_k) p$$

Minimize approximation:

$$\nabla_p m_k(p) = 0 \Rightarrow p_k = -\left(\nabla^2 f(x_k)\right)^{-1} \nabla f(x_k)$$

"Newton step":

$$x_{k+1} = x_k + p_k = x_k - \left(\nabla^2 f(x_k)\right)^{-1} \nabla f(x_k)$$

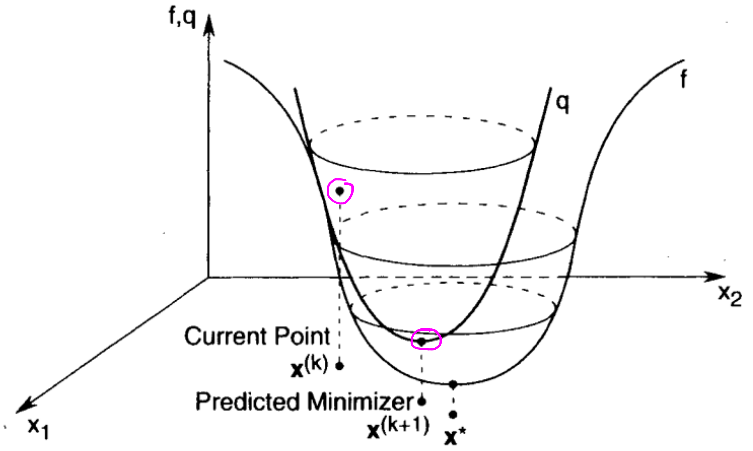
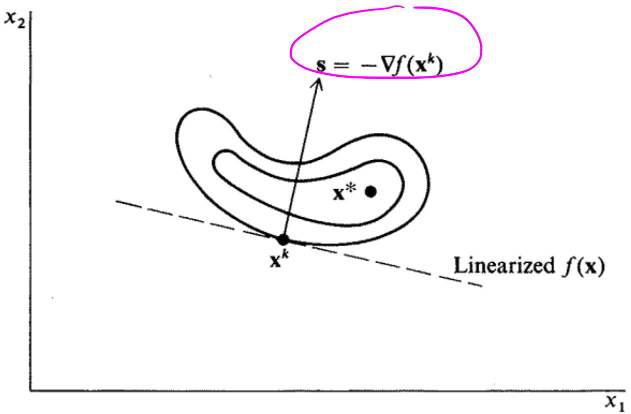


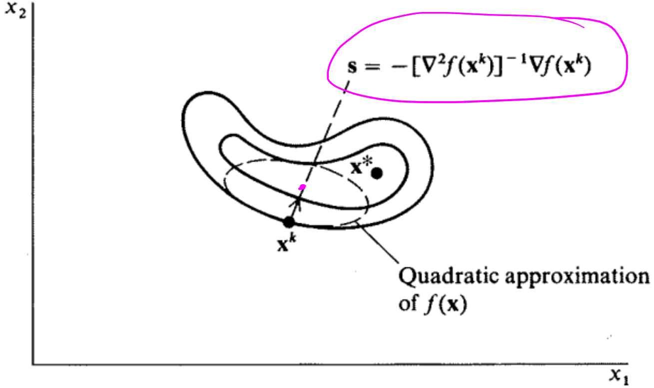
Figure 9.1 Quadratic approximation to the objective function using first and second derivatives.

Chong & Zak, "An introduction to optimization"

Steepest descent directions vs Newton directions from objective function approximations



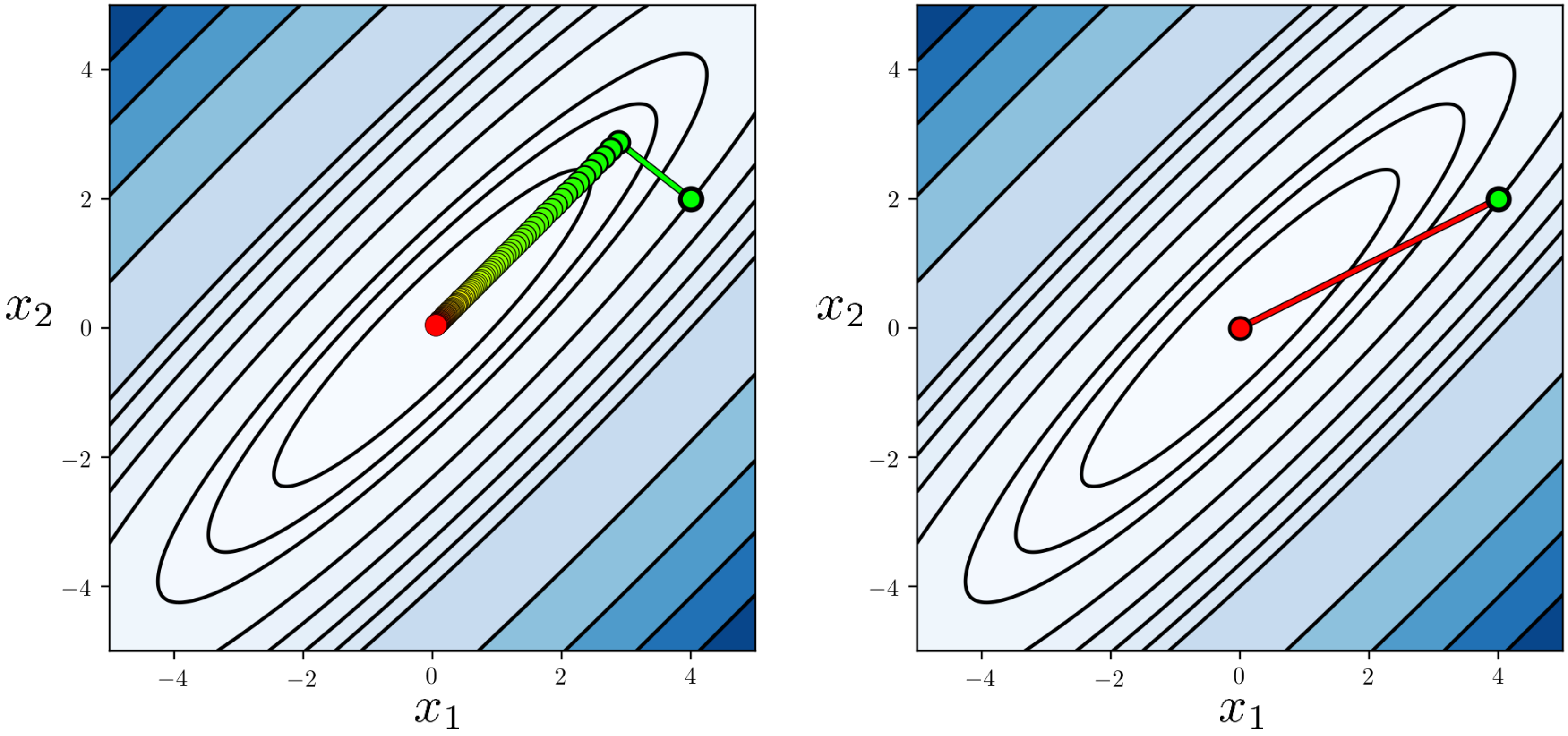
(a) Steepest descent: first-order approximation (linearization) of $f(\mathbf{x})$ at $\mathbf{x}^k$



(b) Newton's method: second-order (quadratic) approximation of $f(\mathbf{x})$ at $\mathbf{x}^k$

From Edgar, Himmelblau, Lasdon: "Optimization of Chemical Processes"

Steepest descent vs Newton



https://jermwatt.github.io/

How far should we walk along p_k?

1. Line search: Finding α that approximately solves

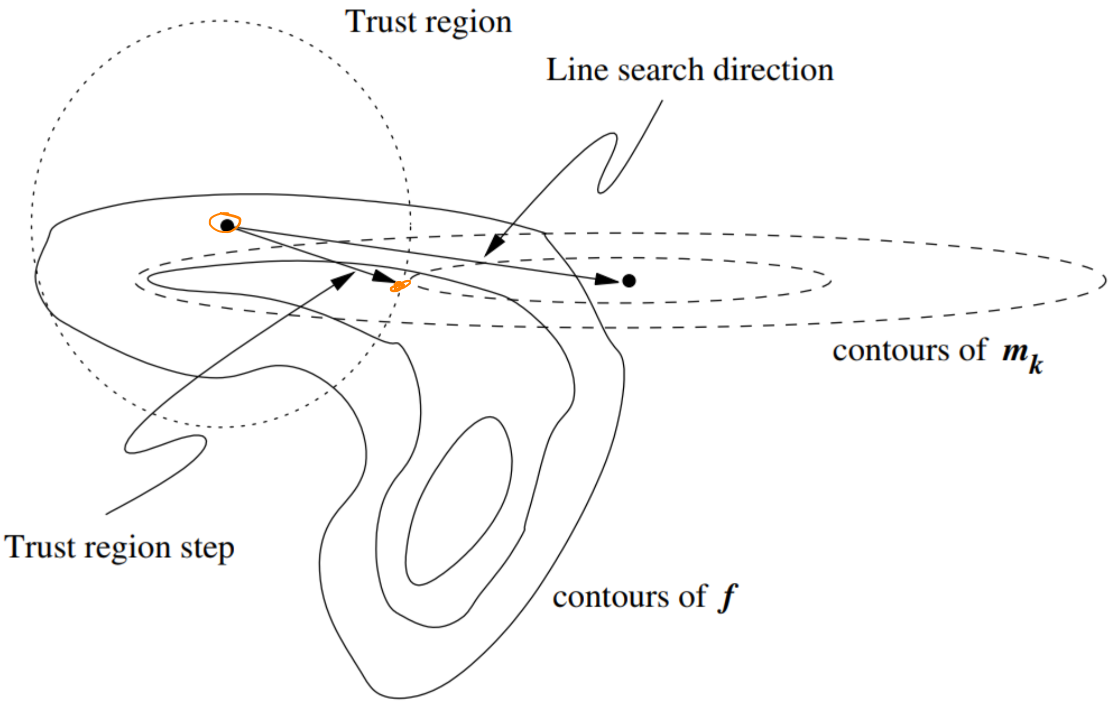
$$\min_{\alpha} f(x_k + \alpha p_k) \rightarrow \alpha_k^*$$

set $x_{k+1} = x_k + \alpha_k^* p_k$

Next time!

2. Trust region (not curriculum)

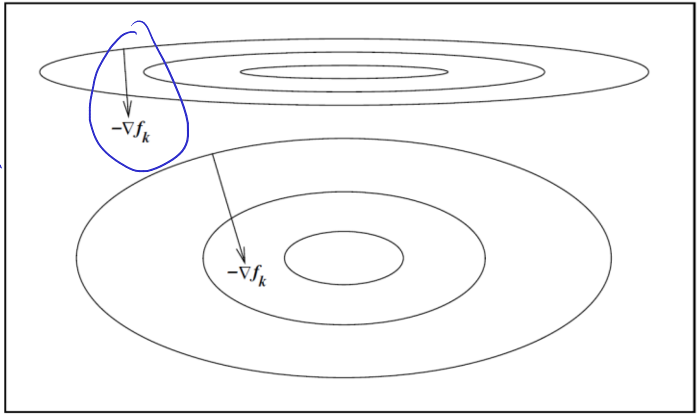
Newton line search and trust region steps



Scaling, scale invariance

Poorly scaled obj. fun. $f(x)$:

$f(x)$ changes faster in some directions than other.



← poorly scaled

← better scaled

Figure 2.7 Poorly scaled and well scaled problems, and performance of the steepest descent direction.

$$\underbrace{\min_{x\in\mathbb{R}^2} \frac{1}{2}\left(x_1^2 + \gamma x_2^2\right)}$$

$\gamma >> 1$ or $\gamma << 1$:

poorly scaled

steepest descent $p_k = -\begin{pmatrix} x_1 \\ \gamma x_2 \end{pmatrix}$

Newton: $p_k = -\begin{bmatrix} 1 & 0 \\ 0 & \gamma \end{bmatrix}^{-1} \cdot \begin{bmatrix} x_1 \\ \gamma x_2 \end{bmatrix} = -\begin{pmatrix} x_1 \\ x_2 \end{pmatrix}$

scale invariant!