T-61.183 SUPPORT VECTOR MACHINES AND KERNEL METHODS

Learning with Kernels
Chapter 6: Optimization

Schölkopf, Smola

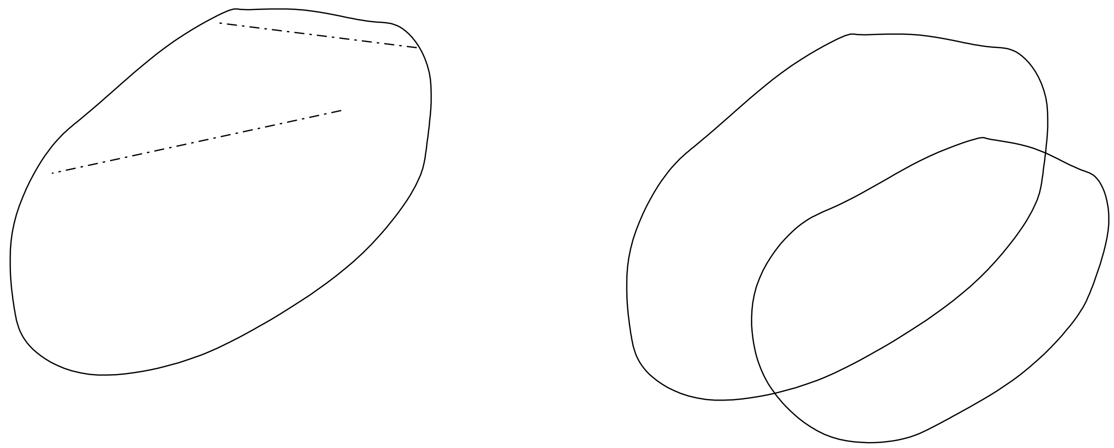
presented by Tapani Raiko

Mar 3, 2003

Introduction & Contents

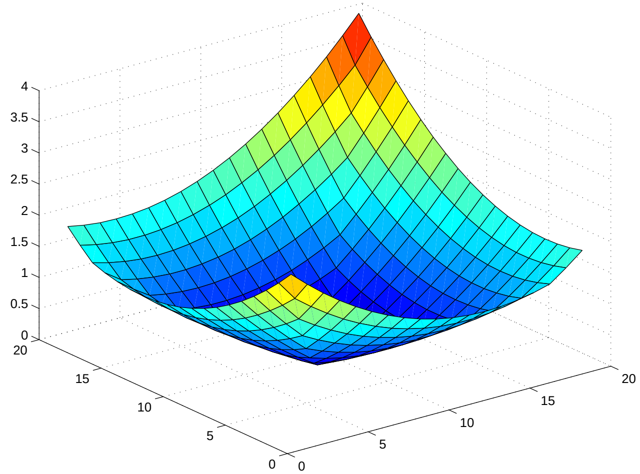
- Learning implies the minimization of some risk functional
- In general a difficult task (many local minima)
- In case of kernels: (typically) convex optimization
- 1-dimensional: Interval cutting, Newton method
- N-dimensional: Conjugate gradient descent, predictor corrector method
- Duality theory (Kuhn-Tucker (KKT) condition)

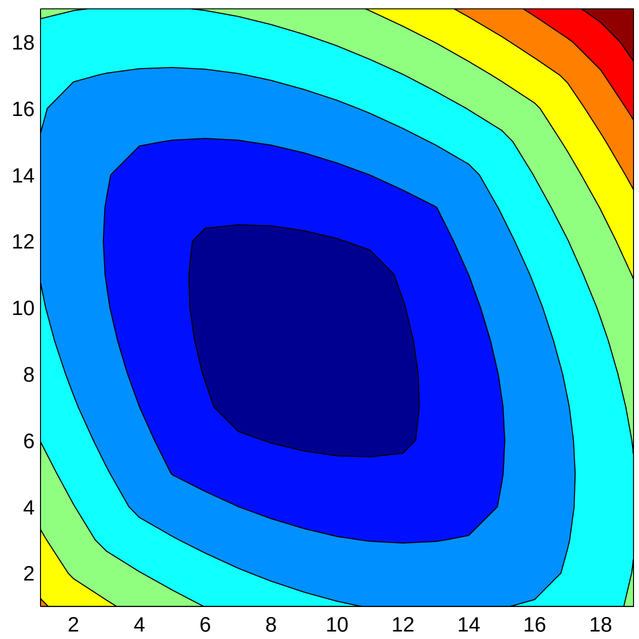
Convex Optimization (1/4): Convex Sets



- Lines with endpoints in the set are fully contained in the set
- Intersection of two convex sets is also convex

(2/4): Convex Functions

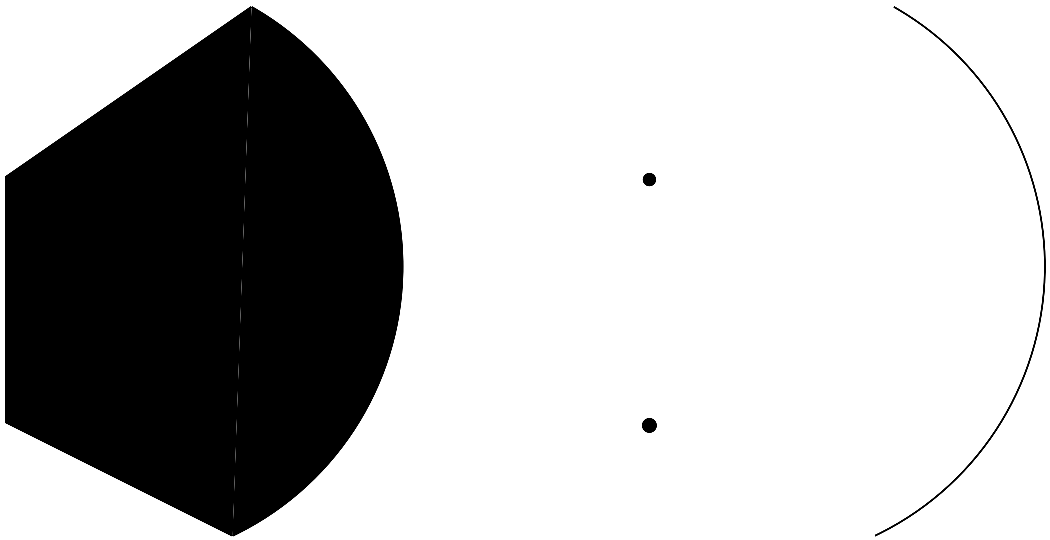




- Function $f : \mathcal{X} \to \mathbb{R}$ is convex iff below-sets are convex (assuming $\mathcal{X}$ convex)

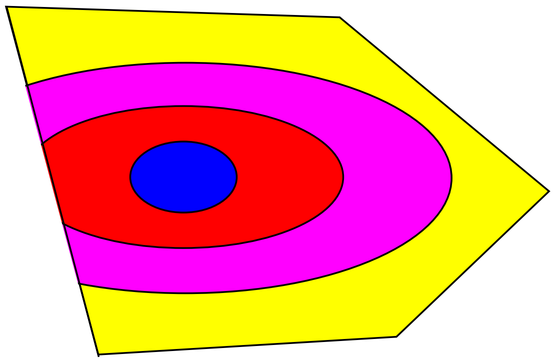
$$X_c := \{x \in \mathcal{X} \mid f(x) \leq c\} \tag{1}$$

(3/4): Vertex of a Set



- A point is a vertex, if it cannot be reconstructed from other points
- Line segments between vertices of a convex set reconstruct the whole set

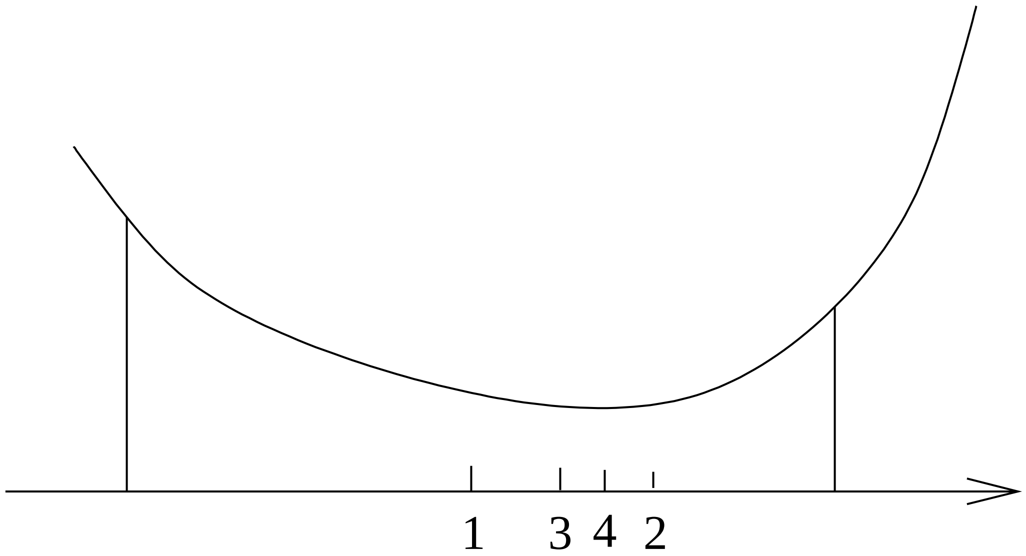
(4/4): Convex - Results



For convex functions on a convex set:

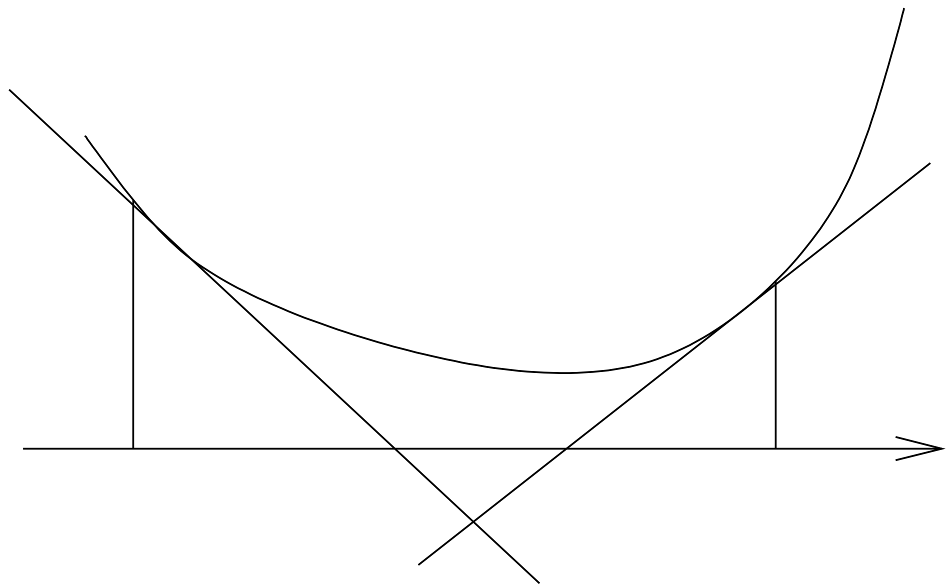
- Local minimum is a global minimum
- Maximum can be found at one of the vertices

Functions of One Variable (1/4): Interval Cutting



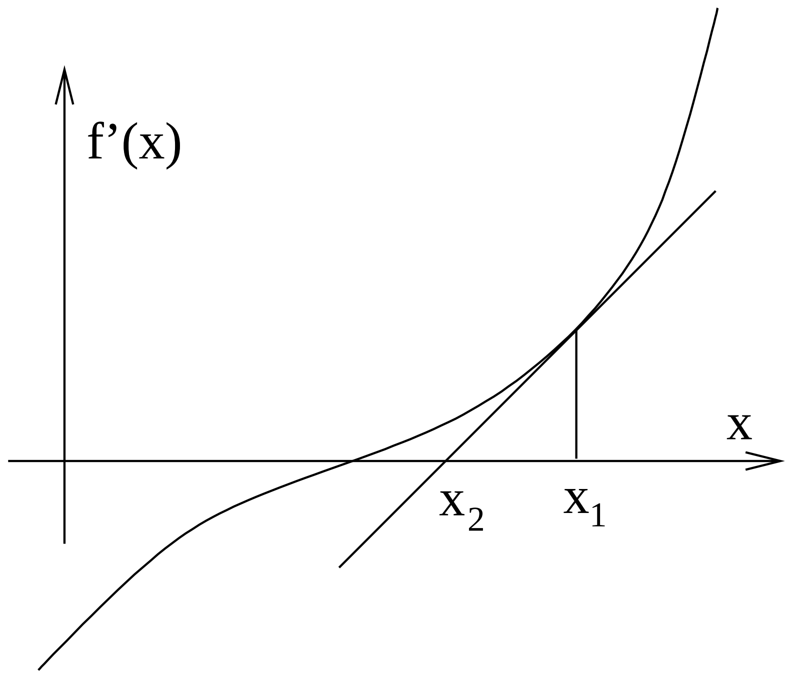
- Cut the interval in two halves
- Choose based on f'

(2/3): Error Bound and Convergence of Interval Cutting



- One can find a bound for the true minimum
- The convergence is linear with constant 0.5
  = The error is halved at each iteration

(3/4): Newton Method



- Fit a parabola to $f(x_1), f'(x_1), f''(x_1)$ and use it's minimum as x_2
- If the starting point is sufficiently close to the minimum:
  - Will converge at least quadratically

(4/4): 1-D Discussion

- If Newton method converges, we know the solution is correct
- If not, something must be done
- Sometimes the problem is unconstrained
  - One can guess an interval
  - If it was too small, enlarge it

Functions of Several Variables (1/6): Gradient Descent

- Find the direction of steepest descent
- Find the step size using one variable methods above

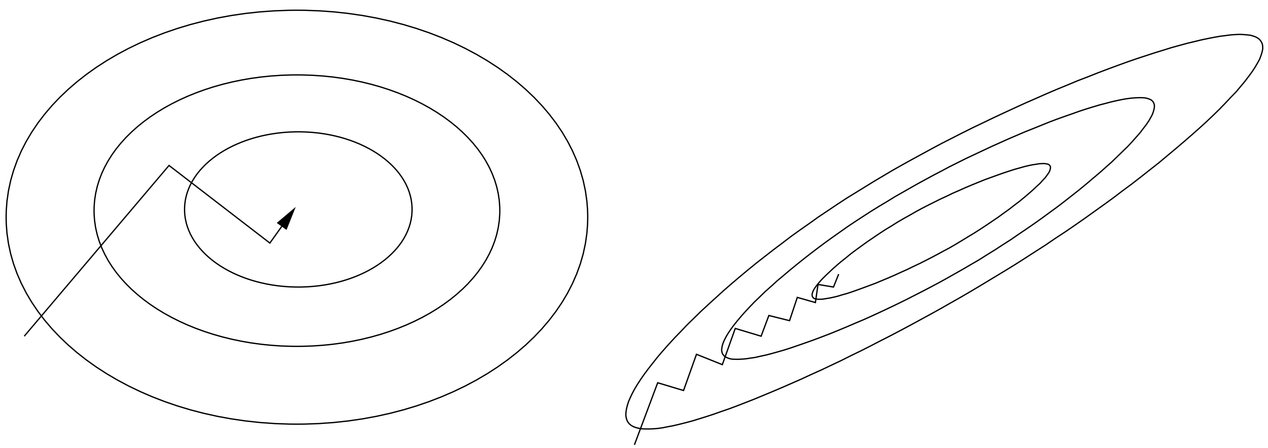
$$x_{n+1} = x_n - \gamma f'(x_n), \tag{2}$$

where $\gamma = \arg\min f(x_{n+1})$

- Gradient descent can be shown to converge
- Note: consecutive updates are orthogonal!

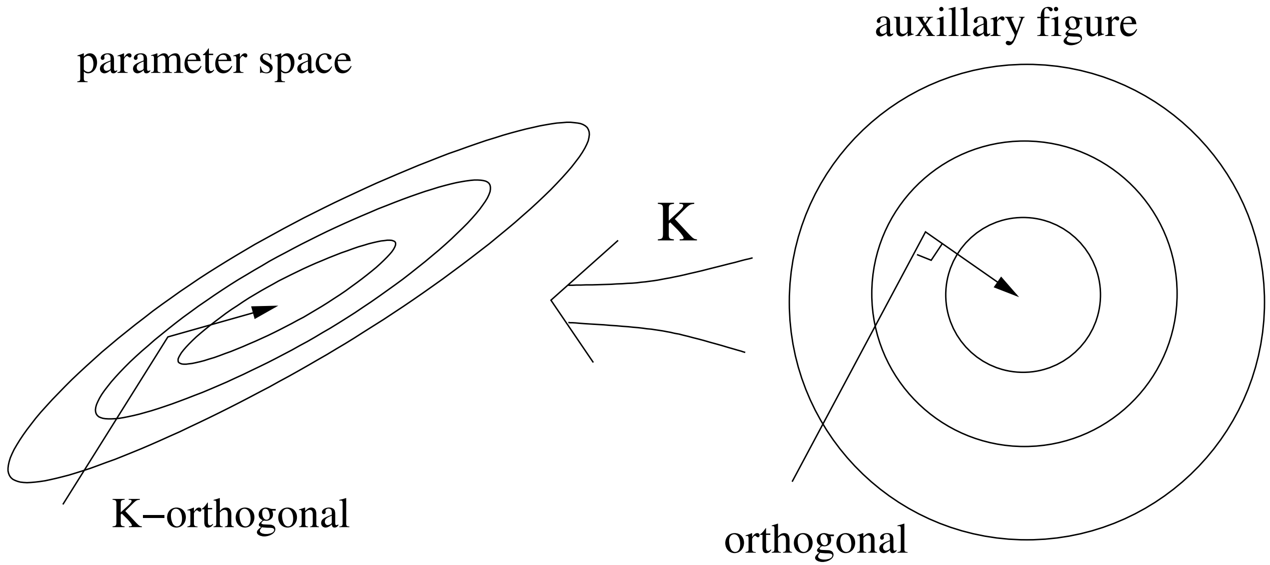
(2/6): Properties of Gradient Descent

- Assume that f is quadratic: $f(x) = \frac{1}{2}(x - x^*)^T K (x - x^*) + c$
- $\min f(x) = f(x^*) = c$, $f'(x) = K(x - x^*)$
- K assumed strictly positive definite and symmetric
- Kantorovich inequality tells:
  - Gradient descent performs poorly if some of the eigenvalues of K are small compared to the largest one



(3/6): Conjugate Gradient Descent

- x and y are K-orthogonal iff $x^T K y = 0$



- K-orthogonal updates do not disturb each other in the quadratic optimization problem
- Idea: fit a quadratic function to the object function
  That is, approximate K somehow (e.g. the Hessian of f)

(4/6): Conjugate Gradient Descent

Generic conjugate gradient descent vs. Polak-Ribiere

$$x_{i+1} = x_i - \frac{g_i^T v_i}{v_i^T f''(x_i) v_i} v_i \qquad x_{i+1} = x_i + \alpha v_i$$

$$v_{i+1} = -g_{i+1} + \frac{g_{i+1}^T f''(x_i) v_i}{v_i^T f''(x_i) v_i} v_i \quad v_{i+1} = -g_{i+1} + \frac{(g_{i+1} - g_i)^T g_{i+1}}{g_i^T g_i} v_i$$

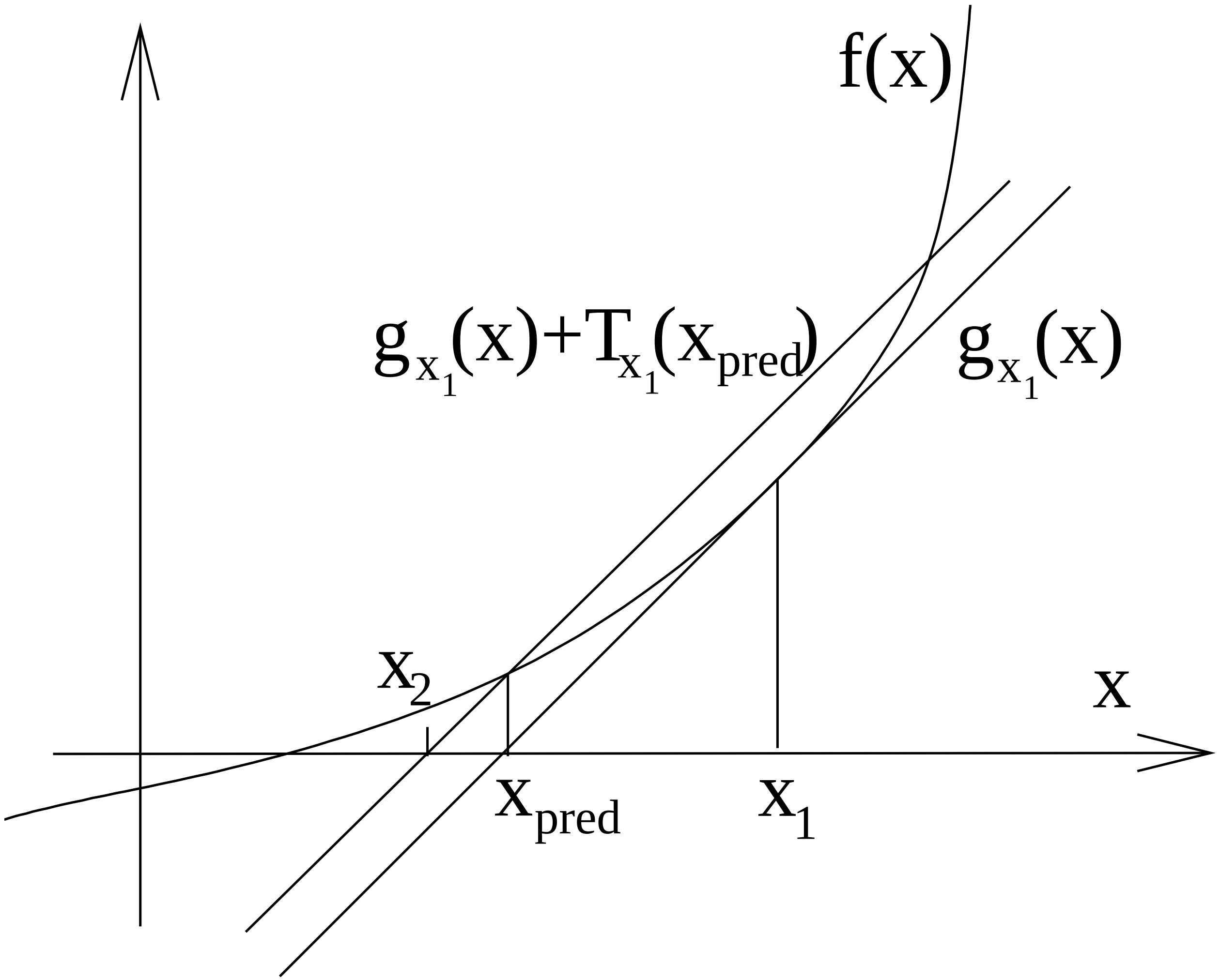
where g_i is shorthand for $f'(x_i)$

- The computation of the Hessian $f''(x_i)$ is a costly operation
- Since it is an approximation anyway, some variants avoid it

(5/6): Predictor Corrector Method

- Predictor corrector method obtains the performance of higher order methods without actually implementing them
- To find $f(x^*) = 0$
- Expand $f(x) = g_{x_i}(x) + T_{x_i}(x)$,
  where g_{x_i} is a simple function fitted to f at x_i
- Predictor: Solve $g_{x_i}(x_{\text{pred}}) = 0$ for x_{pred}
- Corrector: Solve $g_{x_i}(x_{i+1}) + T_{x_i}(x_{\text{pred}}) = 0$ for x_{i+1}
- Eliminates lower order terms

(6/6): Predictor Corrector Method - Example



Constrained Problems (1/5): Problem Statement

- The typical problem with kernel machines is:
- Minimize $f(x)$
  subject to $c_i(x) \leq 0$ for all $i = 1, 2, \ldots, n$
- Equality constraints $e_j(x) = 0$ can be handled analogously
- Note 1: If c_i are convex functions,
  the feasible region $\{x \mid \forall i : c_i(x) \leq 0\}$ is convex
- Note 2: Optimality of x^* does not require $f'(x^*) = 0$

(2/5): Kuhn-Tucker Saddle Point Condition

- Define a Lagrangian:

$$L(x, \alpha) := f(x) + \sum_{i=1}^{n} \alpha_i c_i(x) \tag{3}$$

- Restrict $\alpha_i \geq 0$ for all i
- If there is such an (x^*, α^*) that for every (x, α)

$$L(x^*, \alpha) \leq L(x^*, \alpha^*) \leq L(x, \alpha^*) \tag{4}$$

- Then x^* is a solution and $\forall i : \alpha_i^* c_i(x^*) = 0$
- This KKT criterion is also necessary if f and c_i are convex

(3/5): KKT for Differentiable Problems

- The KKT condition can be rewritten as:

$$\partial_x L(x^*, \alpha^*) = 0 \tag{5}$$

$$\forall i : \partial_{\alpha_i} L(x^*, \alpha^*) \leq 0 \tag{6}$$

$$\sum_{i=1}^{n} \alpha_i^* c_i(x^*) = 0 \tag{7}$$

- Optimization problem transformed into a set of equations
- Error bound: $f(x) \geq f(x^*) \geq f(x) + \sum_{i=1}^{n} \alpha_i c_i(x)$ (KKT-gap) assuming that (x, α) satisfies (5) and (6)

(4/5): Wolfe's Dual Optimization Problem

- It is possible to eliminate x from the differentiated KKT condition if the functions are simple enough
- The resulting optimization problem with α is called the Wolfe's dual
- Primal has m variables and n constraints
  Dual has n variables and m constraints
  $\Rightarrow$ If $n < m$, the dimensionality of the problem is smaller
- Constraints become simpler ($\alpha_i \geq 0$)

(5/5): Primal and Dual of Linear and Quadratic Problems

primal (in x)	dual (in α)
solution exists	solution exists
no solution	unbounded or infeasible
unbounded or infeasible	no solution
inequality constraint	inequality constraint
equality constraint	free variable
free variable	equality constraint

Summary

- Machine learning $\approx$ optimization of a risk functional
- Optimization step can be divided into
  1) finding a direction and 2) finding a step size
- Typical idea: Fit a simpler function to the current hypothesis
- Convexity is a useful property
  - Local minimum $\Rightarrow$ global minimum
  - Maximum can be found on the vertices
  - Kuhn-Tucker condition becomes equivalent to finding the solution $\rightarrow$ duality theory

Exercise 6.4

Denote by f a convex function on $[a, b]$. Show that the algorithm below finds the minimum of f. What is the rate of convergence in x to $\arg\min_x f(x)$? Can you obtain a bound in $f(x)$ wrt. $\min_x f(x)$?

input: a, b, f and threshold ϵ

$x_1 = a$, $x_2 = \frac{a+b}{2}$, $x_3 = b$

repeat

  if $x_3 - x_2 > x_2 - x_1$ then $x_4 = \frac{x_2+x_3}{2}$ else $x_4 = \frac{x_1+x_2}{2}$

  Keep the two points closest to the point with the minimum value of $f(x_i)$ and rename them such that $x_1 < x_2 < x_3$

until $x_3 - x_1 \geq \epsilon$