Convex Optimization and Modeling

Duality Theory and Optimality Conditions

5th lecture, 12.05.2010

Jun.-Prof. Matthias Hein

Program of today/next lecture

Lagrangian and duality:

- the Lagrangian
- the dual Lagrange problem
- weak and strong duality, Slater's constraint qualification
- optimality conditions, complementary slackness, KKT conditions
- perturbation analysis of the constraints
- generalized inequalities

Duality theory is done for the general optimization problem !

Motivation of the Lagrange function: general optimization problem (MP)

$$\min_{x \in D} f(x),$$
$$\text{subject to: } g_i(x) \leq 0, \; i = 1, \ldots, r$$
$$h_j(x) = 0, \; j = 1, \ldots, s.$$

Idea:

- turn constrained problem into an unconstrained problem,
- the set of extremal points of the resulting unconstrained problem contains the extremal points of the original constrained problem (necessary condition). In some cases the two sets are equal (necessary and sufficient condition).

Definition 1. *The* ***Lagrangian*** *or* ***Lagrange function*** $L : \mathbb{R}^n \times \mathbb{R}^r_+ \times \mathbb{R}^s \to \mathbb{R}$ *associated with the MP is defined as*

$$L(x, \lambda, \mu) = f(x) + \sum_{j=0}^{r} \lambda_j \, g_j(x) + \sum_{i=0}^{s} \mu_i \, h_i(x),$$

with $\operatorname{dom} L = D \times \mathbb{R}^r_+ \times \mathbb{R}^s$ *where* D *is the domain of the optimization problem. The variables* λ_j *and* μ_i *are called* ***Lagrange multipliers*** *associated with the inequality and equality constraints.*

The Lagrange function III

Interpretation:

- The constrained problem is turned into an unconstrained problem using

$$I_-(x) = \begin{cases} 0 & \text{if } x \leq \infty, \\ \infty & \text{if } x > 0. \end{cases}, \qquad \text{and} \qquad I_0(x) = \begin{cases} 0 & \text{if } x = 0, \\ \infty & \text{if } x > 0. \end{cases}$$

Thus, we arrive at

$$\min_{x \in D} f(x) + \sum_{j=0}^{r} I_-(g_j(x)) + \sum_{i=0}^{s} I_0(h_i(x)).$$

Lagrangian: relaxation of the hard constraints to linear functions.
Note, that a linear function (which is positive in the positive orthant) is an underestimator of the indicator function $I_0(x)$ $(I_-(x))$.

- the extremal points of the Lagrangian are closely related to the extremal points of the optimization problem.

Definition 2. *The* ***dual Lagrange function*** $q : \mathbb{R}^r_+ \times \mathbb{R}^s \to \mathbb{R}$ *associated with the MP is defined as*

$$q(\lambda, \mu) = \inf_{x \in D} L(x, \lambda, \mu) = \inf_{x \in D} \Big(f(x) + \sum_{j=0}^{r} \lambda_j \, g_j(x) + \sum_{i=0}^{s} \mu_i \, h_i(x) \Big),$$

where $q(\lambda, \mu)$ *is defined to be* $-\infty$ *if* $L(x, \lambda, \mu)$ *is unbounded from below in* x*.*

Properties:

- the dual function is a pointwise infimum of a family of linear and thus concave functions (in λ and μ) and therefore concave.
  This holds irrespectively of the character of the MP, in particular this holds also for discrete optimization problems.

The dual function II

Proposition 1. *For any $\lambda \succeq 0$ and μ we have,*

$$q(\lambda, \mu) \leq p^*.$$

Proof. Suppose x' is a feasible point for the problem, then $g_i(x') \leq 0$, $i = 1, \ldots, r$ and $h_j(x') = 0$, $j = 1, \ldots, s$. Thus,

$$\sum_{j=0}^{r} \lambda_j \, g_j(x') + \sum_{i=0}^{s} \mu_i \, h_i(x') \leq 0.$$

Thus we have,

$$L(x', \lambda, \mu) = f(x') + \sum_{j=0}^{r} \lambda_j \, g_j(x') + \sum_{i=0}^{s} \mu_i \, h_i(x') \leq f(x').$$

$$\Longrightarrow \quad q(\lambda, \mu) = \inf_{x \in D} L(x, \lambda, \mu) \leq f(x').$$

Since this holds for any feasible point x', we get $q(\lambda, \mu) \leq p^*$. $\square$

The dual problem

For each pair (λ, μ) with $\lambda \succeq 0$ we have $q(\lambda, \mu) \leq p^*$.

What is the best possible lower bound ?

The dual problem

For each pair (λ, μ) with $\lambda \succeq 0$ we have $q(\lambda, \mu) \leq p^*$.

What is the best possible lower bound ?

Definition 4. *The* ***Lagrange dual problem*** *is defined as*

$$\max q(\lambda, \mu),$$
$$\textit{subject to: } \lambda \succeq 0.$$

Properties:

- For each MP the dual problem is **convex**.
- The original OP is called the **primal problem**.
- (λ, μ) is **dual feasible** if $q(\lambda, \mu) > -\infty$.
- (λ^*, μ^*) is called **dual optimal** if they are optimal for the dual problem.

The dual problem II

Making implicit constraints in the dual problem explicit:

The dimension of the domain of the dual function

$$\operatorname{dom} q = \{(\lambda, \mu) \,|\, q(\lambda, \mu) > -\infty\},$$

is often smaller than $r + s$ $(\lambda \in \mathbb{R}^r, \mu \in \mathbb{R}^s)$ $\Rightarrow$ identify "hidden constraints"

Making implicit constraints in the dual problem explicit:

The dimension of the domain of the dual function

$$\mathrm{dom}\, q = \{(\lambda, \mu) \,|\, q(\lambda, \mu) > -\infty\},$$

is often smaller than $r + s$ $(\lambda \in \mathbb{R}^r, \mu \in \mathbb{R}^s)$ $\Rightarrow$ identify "hidden constraints"

Example:

$$\min \; \langle c, x \rangle \qquad \text{(standard form LP)}$$

$$\text{subject to: } Ax = b, \quad x \succeq 0.$$

with dual function: $q(\lambda, \mu) = \begin{cases} -\langle b, \mu \rangle & \text{if } c + A^T\mu - \lambda = 0, \\ -\infty & \text{otherwise} \end{cases}.$

$$\max - \langle b, \mu \rangle \;,$$
$$\text{subject to: } \lambda \succeq 0,$$
$$A^T\mu - \lambda + c = 0.$$

$$\max - \langle b, \mu \rangle \;,$$
$$\text{subject to: } A^T\mu + c \succeq 0. \; .$$

Example: Least squares with linear constraint

Consider the problem, with $x \in \mathbb{R}^n$, $A \in \mathbb{R}^{p \times n}$ and $b \in \mathbb{R}^p$,

$$\min_{x \in D} \; \|x\|_2^2 \,, \qquad\qquad \text{(Least squares with linear constraint)}$$

$$\text{subject to: } Ax = b, \;\; ,$$

- The Lagrangian is: $L(x, \mu) = \|x\|_2^2 + \langle \mu, Ax - b \rangle$.
- Lagrangian is a strictly convex function of x,

$$\nabla_x L(x, \mu) = 2x + A^T \mu = 0,$$

which gives $x = -\frac{1}{2} A^T \mu$ and we get the dual function

$$q(\mu) = \frac{1}{4} \|A^T \mu\|_2^2 + \left\langle \mu, -\frac{1}{2} A A^T \mu - b \right\rangle = -\frac{1}{4} \|A^T \mu\|_2^2 - \langle \mu, b \rangle.$$

The lower bound states for any $\mu \in \mathbb{R}^p$ we have

$$-\frac{1}{4} \|A^T \mu\|_2^2 - \langle \mu, b \rangle \leq p^* = \inf\{\|x\|_2^2 \mid Ax = b\}.$$

Example: A graph-cut problem

Problem:

- weighted, undirected graph with n vertices and weight matrix $W \in S^n$.
- split the graph (the vertex set) in two (disjoint) groups

$$\min \ \langle x, Wx \rangle \qquad \text{(graph cut criterion)}$$
$$\text{subject to: } x_i^2 = 1, \quad i = 1, \ldots, n,$$

$$L(x, \mu) = \langle x, Wx \rangle + \sum_{i=1}^{n} \mu_i (x_i^2 - 1) = \sum_{i,j=1}^{n} \Big(W_{ij} + \mu_i \delta_{ij} \Big) x_i x_j - \langle \mu, \mathbf{1} \rangle$$

and thus the dual function is: $q(\mu) = \begin{cases} -\langle \mu, \mathbf{1} \rangle, & \text{if } W + \mathrm{diag}(\mu) \succeq 0, \\ -\infty, & \text{otherwise} \end{cases}.$

For every μ which is feasible we get a lower bound e.g. $\mu = -\lambda_{\min}(W)\mathbf{1}$,

$$p^* \geq -\langle \mu, \mathbf{1} \rangle = n\, \lambda_{\min}(W).$$

Weak and strong duality

Corollary 1. *Let d^* and p^* be the optimal values of the dual and primal problem. Then*

$$d^* \leq p^*, \qquad \textbf{\textit{(weak duality)}}.$$

- The difference $p^* - d^*$ is the **optimal duality gap** of the MP.
- solving the convex dual problem provides lower bounds for any MP.

Corollary 2. *Let d^* and p^* be the optimal values of the dual and primal problem. Then*

$$d^* \leq p^*, \qquad \textbf{\textit{(weak duality)}}.$$

- The difference $p^* - d^*$ is the **optimal duality gap** of the MP.
- solving the convex dual problem provides lower bounds for any MP.

Definition 6. *We say that **strong duality** holds if*

$$d^* = p^*.$$

Constraint qualifications *are conditions under which strong duality holds.*

Strong duality **does not** hold in general !

But for convex problems strong duality holds quite often.

Geometric interpretation of strong duality:

- only inequality constraints: consider the set defined as

$$S = \{(g(x), f(x)) \in \mathbb{R}^r \times \mathbb{R} \,|\, x \in D\} \subseteq \mathbb{R}^{r+1}.$$

- Interpret the sum $L(x, \lambda) = f(x) + \langle \lambda, g(x) \rangle$ as

$$\langle \lambda, g(x) \rangle + f(x) = \langle (\lambda, 1), (g(x), f(x)) \rangle$$

  hyperplane H: normal vector $(\lambda, 1)$ going through the point $(g(x), f(x))$.

- We have $u = (\lambda, 1)$ and $x_0 = (g(x), f(x))$ and thus with $x = (z, w)$ we get

$$\langle u, x - x_0 \rangle = \langle \lambda, z - g(x) \rangle + (w - f(x)).$$

  $\Longrightarrow$ H intersects the vertical axis $\{(0, w) \,|\, w \in \mathbb{R}\}$ at $L(x, \lambda)$.

$$-\langle \lambda, g(x) \rangle + w - f(x) = 0 \quad \Longrightarrow \quad w = f(x) + \langle \lambda, g(x) \rangle \,.$$

Slater's constraint qualification II

- Among all hyperplanes with normal $u = (\lambda, 1)$ that have S in the positive half-space, the highest intersection point of the vertical axis will be

$$q(\lambda) = \inf_x L(x, \lambda).$$

- The hyperplane $u = (\lambda, 1)$ with offset c contains every $y \in S$ if and only if

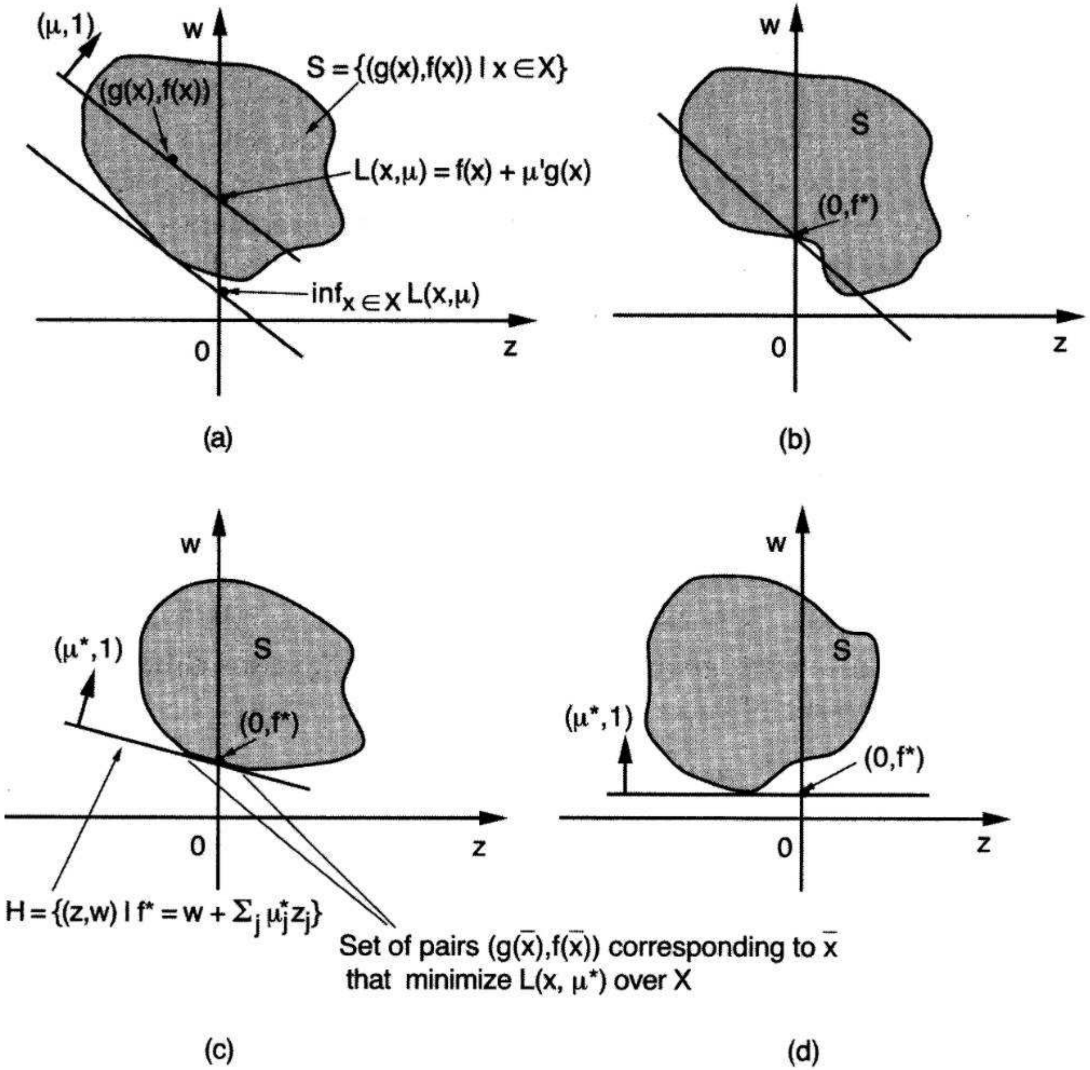
$$\forall y \in S, \quad \langle u, y \rangle \geq c \quad \Leftrightarrow \quad \langle \lambda, g(x) \rangle + f(x) \geq c, \quad \forall x \in D.$$

cuts the vertical axis at c and $c \leq f(x) + \langle \lambda, g(x) \rangle \,, \quad \forall\, x \in D$.
Thus: $c = \inf_{x \in D} f(x) + \langle \lambda, g(x) \rangle = q(\lambda)$.

- **dual problem:** find a hyperplane of form $u = (\lambda, 1)$ which has the highest interception with the vertical axis and contains S contained in its positive half-space.

- **strong duality:** there exists a supporting hyperplane of S which contains $(0, p^*)$.

Slater's constraint qualification III



a) supporting hyperplane of $S = \{(g(x), f(x)) \,|\, x \in D\}$ and the value $q(\lambda) = \inf_x L(x,\lambda)$ of the dual Lagrangian,

b) a set S with a duality gap,

c) no duality gap and the optimum is attained for an active inequality constraint,

d) no duality gap and the optimum is attained for an inactive inequality constraint.

Slater's constraint qualification:

Theorem 1. *Suppose that the primal problem is convex and there exists an* $x \in \operatorname{relint} D$ *such that*

$$g_i(x) < 0, \quad i = 1, \ldots, r,$$

then Slater's condition holds and strong duality holds. Strict inequality is not necessary if $g_i(x)$ *is an affine constraint.*

Proof:

- $A = \{(z_1, \ldots, z_r, w) \in \mathbb{R}^{r+1} \,|\, \exists x \in D,\ g_j(x) \le z_j,\ j = 1, \ldots, r,\ f(x) \le w\}$, contains the set S. Since f, g_j are convex, they have convex sublevel sets $\Longrightarrow$ A is convex.
- $(0, p^*)$ is a boundary point of A,
- supporting hyperplane which contains A in the positive half-space $\Longrightarrow$

$$\exists (\lambda, \beta) \neq (0, 0) \text{ such that } \langle \lambda, z \rangle + \beta(w - p^*) \ge 0, \quad \forall (z, w) \in A.$$

Proof continued:

- $(z, w + \gamma) \in A$ for $\gamma > 0$ and $(z_1, \ldots, z_j + \gamma, \ldots, z_r, w) \in A$ for $\gamma > 0$. If $\beta < 0$ we could find easily a γ in order to violate the above equation $\Longrightarrow$ $\lambda_j \geq 0$ for $j = 1, \ldots, r$ and $\beta \geq 0$.
- Assume: $\beta = 0 \quad \Longrightarrow \quad \langle \lambda, z \rangle \geq 0$ for all $(z, w) \in A$. Since $(g_1(x), \ldots, g_r(x), f(x)) \in A$ for all $x \in D$, $\sum_{j=1}^r \lambda_j g_j(x) \geq 0, \quad \forall\, x \in D$. **Assumption:** $\exists\, x$ such that $g_j(x) < 0$ for all $j = 1, \ldots, r$, thus with $\lambda \succeq 0$ we would get $\lambda = 0$ and thus $(\lambda, \beta) = (0, 0)$ ↯.
- Division of (λ, β) by β we get the standard form $(\lambda, 1)$. With $(g(x), f(x)) \in A$ for all $x \in D$ yields,

$$p^* \leq f(x) + \langle \lambda, g(x) \rangle, \quad \forall\, x \in D.$$

and thus $p^* \leq \inf_{x \in D} \Big(f(x) + \langle \lambda, g(x) \rangle \Big) = q(\lambda) \leq d^*$.
Using weak duality we get $d^* \leq p^*$ and thus $p^* = d^*$.

Remark:

- If the problem is convex, Slater's condition does not only imply that strong duality holds but also that the **dual optimal d^* is attained** given that $d^* > -\infty$, that means there exist (λ^*, μ^*) such that $q(\lambda^*, \mu^*) = d^* = p^*$.

Primal problem can be solved by solving the dual problem.

UNIVERSITÄT DES SAARLANDES

Minimax-Theorem for mixed strategy two-player matrix games

MACHINE LEARNING

Two Player matrix game:

- P1 has actions $\{1, \ldots, n\}$ and P2 has actions $\{1, \ldots, m\}$.
- payoff matrix: $P \in \mathbb{R}^{n \times m}$,
- P1 chooses k, P2 chooses l $\quad \Longrightarrow \quad$ P1 pays P2 an amount of P_{kl},
- **Goals:** P1 wants to minimize P_{kl}, P2 wants to maximize it
- In a mixed strategy game P1 and P2 make their choice using probability measures $\mathrm{P}_1, \mathrm{P}_2$,

$$\mathrm{P}_1(k = i) = u_i \geq 0, \qquad \text{and} \qquad \mathrm{P}_2(l = j) = v_j \geq 0,$$

  where $\sum_{i=1}^n u_i = \sum_{j=1}^m v_j = 1$.
- The expected amount player 1 has to pay to player 2 is given by

$$\langle u, Pv \rangle = \sum_{i=1}^{m} \sum_{j=1}^{n} u_i P_{ij} v_j .$$

Minimax-Theorem for mixed strategy two-player matrix games

- **Assumption:** strategy of player 1 is known to player 2.

$$\sup_{v \in \mathbb{R}^m, \sum_{j=1}^m v_j = 1} \langle u, Pv \rangle = \max_{i=1,\ldots,m} (P^T u)_i = \max_{i=1,\ldots,m} \sum_{j=1}^n P_{ji} u_j.$$

P1 has to choose u which minimizes the worst-case payoff to P2

$$\min \max_{i=1,\ldots,m} \sum_{j=1}^n P_{ji} u_j$$

$$\text{subject to: } \langle \mathbf{1}, u \rangle = 1, \quad u \succeq 0.$$

p_1^* is smallest payoff of P1 given that P2 knows the strategy of P1.

$$\max \min_{i=1,\ldots,n} \sum_{j=1}^m P_{ij} v_j$$

$$\text{subject to: } \langle \mathbf{1}, v \rangle = 1, \quad v \succeq 0.,$$

p_2^* is smallest payoff of P2 given that P1 knows the strategy of P2.

- knowledge of the strategy of the opponent should help and $p_1^* \geq p_2^*$.
- difference $p_1^* - p_2^*$ could be interpreted as the advantage of P2 over P1.

- knowledge of the strategy of the opponent should help and $p_1^* \geq p_2^*$.
- difference $p_1^* - p_2^*$ could be interpreted as the advantage of P2 over P1.

But it turns out that:

$$p_1^* = p_2^*.$$

Proof in two steps:

- formulate both problems as LP's and show that they are dual to each other.
- since both are LP's we have strong duality and thus $p_1^* = p_2^*$.

Interpretation of weak and strong duality:

Lemma 1. *The primal and dual optimal value p^* and d^* can be expressed as*

$$p^* = \inf_{x \in X} \sup_{\lambda \succeq 0,\, \mu} L(x, \lambda, \mu), \qquad d^* = \sup_{\lambda \succeq 0,\, \mu} \inf_{x \in X} L(x, \lambda, \mu).$$

Remarks:

- Note: for any $f : \mathbb{R}^m \times \mathbb{R}^n \to \mathbb{R}$, we have for any $S_y \subset \mathbb{R}^n$ and $S_x \subset \mathbb{R}^m$,

$$\sup_{y \in S_y} \inf_{x \in S_x} f(x, y) \le \inf_{x \in S_x} \sup_{y \in S_y} f(x, y).$$

- strong duality: limit processes can be exchanged,
- In particular we have the saddle-point-interpretation

$$\sup_{\lambda \succeq 0, \mu} \inf_{x \in X} L(x, \lambda, \mu) \le L(x^*, \lambda^*, \mu^*) \le \inf_{x \in X} \sup_{\lambda \succeq 0, \mu} L(x, \lambda, \mu).$$

Measure of suboptimality

- every dual feasible point (λ, μ) provides a **certificate** that $p^* \geq q(\lambda, \mu)$,
- every feasible point $x \in X$ provides a **certificate** that $d^* \leq f(x)$,
- any primal/dual feasible pair x and (λ, μ) provides upper bound on the duality gap: $f(x) - q(\lambda, \mu)$, or

$$p^* \in [q(\lambda, \mu), f(x)], \qquad d^* \in [q(\lambda, \mu), f(x)].$$

- duality gap is zero $\Longrightarrow$ x and (λ, μ) is primal/dual optimal.
- **Stopping criterion:** for an optimization algorithm which produces a sequence of primal feasible x_k and dual feasible (λ_k, μ_k). If strong duality holds use:

$$f(x_k) - q(\lambda_k, \mu_k) \leq \varepsilon.$$

Complementary slackness

Corollary 3. *Suppose strong duality holds and let x^* be primal optimal and (λ^*, μ^*) be dual optimal. Then*

$$\lambda_i^* \, g_i(x^*) = 0, \quad i = 1, \ldots, r.$$

Proof. Under strong duality we have

$$f(x^*) = q(\lambda^*, \mu^*) = \inf_x \Big(f(x) + \sum_{j=1}^{r} \lambda_j^* g_j(x) + \sum_{i=1}^{s} \mu_i^* h_i(x) \Big)$$

$$\leq f(x^*) + \sum_{j=1}^{r} \lambda_j^* g_j(x^*) + \sum_{i=1}^{s} \mu_i^* h_i(x^*) \leq f(x^*),$$

which follows from $\lambda_j \geq 0$ together with $g_j(x) \leq 0$ and $h_i(x) = 0$. Thus $\lambda_j \, g_j(x^*) = 0, \quad i = 1, \ldots, r.$ □

$\Longrightarrow$ x^* is the minimizer of $L(x, \lambda^*, \mu^*)$!

KKT optimality conditions

Theorem

- f, g_i and h_j differentiable,
- strong duality holds.

Then **necessary** conditions for primal and dual optimal points x^* and (λ^*, μ^*) are the **Karush-Kuhn-Tucker(KKT) conditions**

$$g_i(x^*) \leq 0,\ i = 1, \ldots, r, \qquad h_j(x^*) = 0,\ j = 1, \ldots, s,$$

$$\lambda_i^* \geq 0,\ i = 1, \ldots, r \qquad \lambda_i^* g_i(x^*) = 0,\ i = 1, \ldots, r$$

$$\nabla f(x^*) + \sum_{i=1}^{r} \lambda_i^* \nabla g_i(x^*) + \sum_{j=1}^{s} \mu_j^* \nabla h_j(x^*) = 0.$$

If the primal problem is **convex**, then the KKT conditions are **necessary and sufficient** for primal and dual optimal points with zero duality gap.

Remarks

- The condition:

$$\nabla f(x^*) + \sum_{i=1}^{r} \lambda_i^* \nabla g_i(x^*) + \sum_{j=1}^{s} \mu_j^* \nabla h_j(x^*) = 0,$$

  is equivalent to $\nabla_x L(x, \lambda^*, \mu^*) = 0$.

- **convex problem:** any pair x, (λ, μ) which fulfills the KKT-conditions is primal and dual optimal. **Additionally:** Slater's condition holds $\Longrightarrow$ such a point exists.

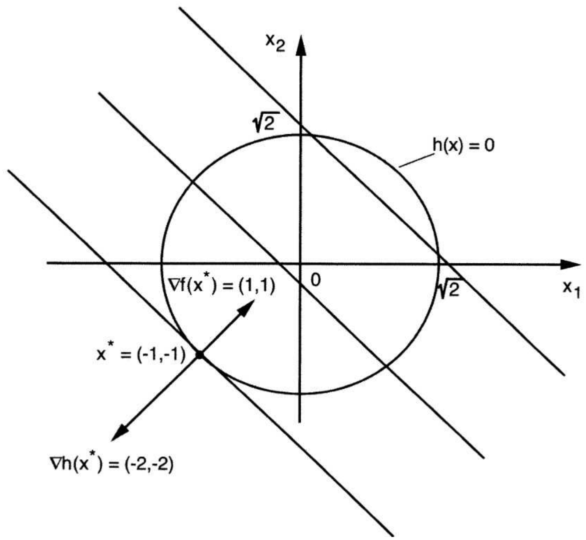
- Assume: strong duality and a dual optimal solution (λ^*, μ^*) is known and $L(x, \lambda^*, \mu^*)$ has a unique minimizer x^*

  1. x^* is primal optimal as long as x^* is primal feasible,
  2. If x^* is not primal feasible, then the primal optimal solution is not attained.

Geometric Interpretation for an equality constraint:

- The set, $h_i(x) = 0$, $i = 1, \ldots, m$, determines a constraint surface in $\mathbb{R}^d$.
- First order variations of the constraints (tangent space of the constraint surface)

$$h(x) = h(x^*) + \langle \nabla h(x^*), x - x^* \rangle \approx 0 \qquad \Longrightarrow \qquad \langle \nabla h(x^*), x - x^* \rangle = 0.$$

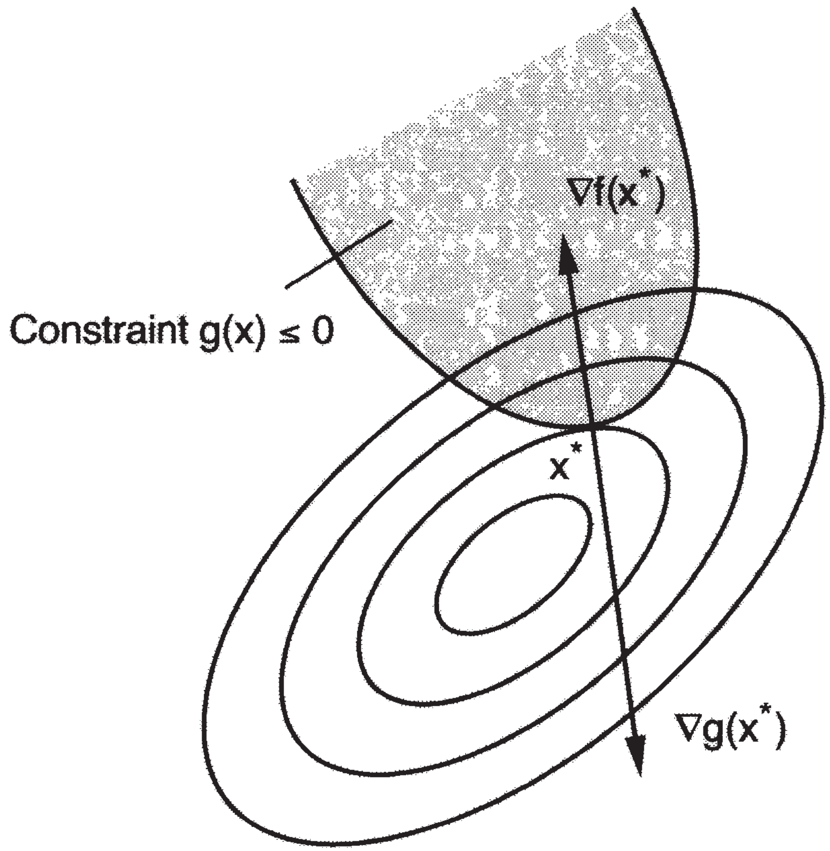


- at a local minima x^* the gradient ∇f is orthogonal to the subspace of first order variations

$$V(x^*) = \{w \in \mathbb{R}^d \mid \langle w, \nabla h_i(x^*) \rangle = 0,\ i = 1, \ldots, m\}$$

- Equivalently,
$\nabla f(x^*) + \sum_{i=1}^m \lambda_i \nabla h_i(x^*) = 0.$

Geometric Interpretation for an inequality constraint:



Two cases:

- constraint active: $g(x^*) = 0$:

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

- constraint inactive: $g(x^*) < 0$,

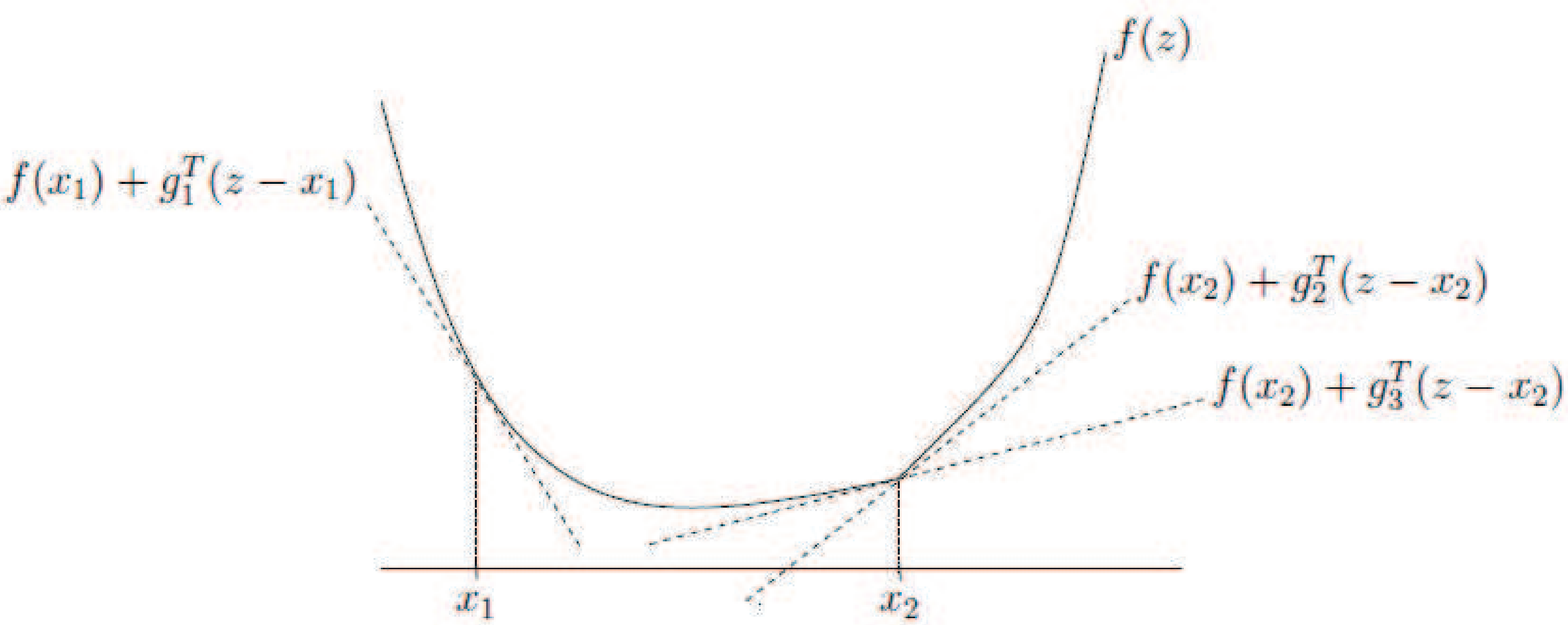
$$\nabla f(x^*) = 0.$$

Subgradient and Subdifferential: Let $f : \mathbb{R}^n \to \mathbb{R}$ be convex.

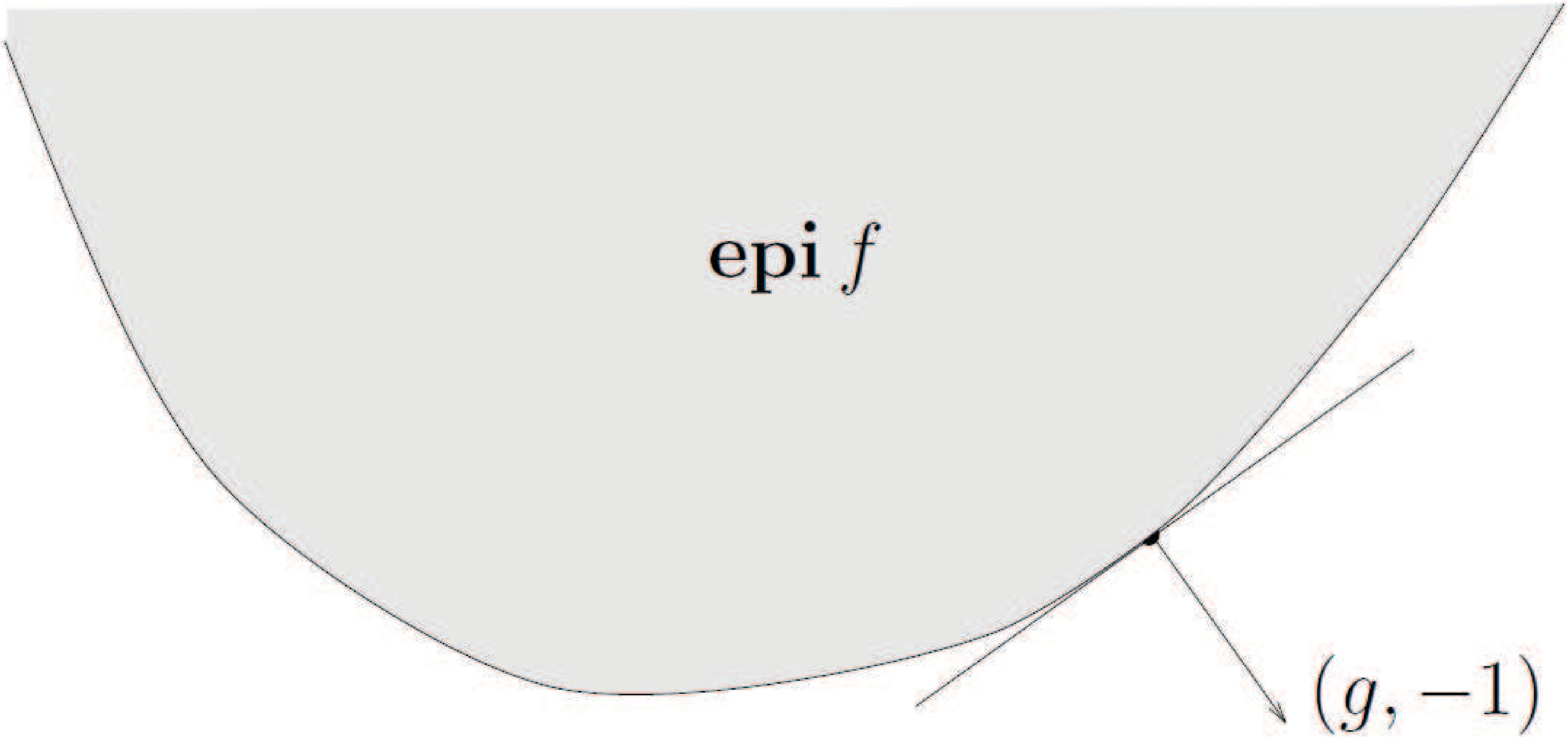
Definition 7. *A vector v is a* ***subgradient*** *of f at x if*

$$f(z) \geq f(x) + \langle v, z - x \rangle, \qquad \forall\, z \in \mathbb{R}^n.$$

The subdifferential $\partial f(x)$ of f at x is the set of all subgradients of f at x.



Subdifferential as supporting hyperplane of the epigraph



Optimality conditions - non-differentiable case

No constraints: f is convex.

- f is differentiable everywhere,

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

- f is not differentiable everywhere

$$f(x^*) = \inf_x f(x) \qquad \Longleftrightarrow \qquad 0 \in \partial f(x^*).$$

Proof: for all $x \in \operatorname{dom} f$,

$$f(x) \geq f(x^*) = f(x^*) + \langle 0, x - x* \rangle \,.$$

$\Longrightarrow$ statement is simple - checking $0 \in \partial f(x)$ can be quite difficult !

Example for optimality condition

Regression: Squared loss + L_1-regularization for orthogonal design,

$$\Psi(w) = \frac{1}{2} \|y - w\|_2^2 + \lambda \|w\|_1 \,,$$

where $Y \in \mathbb{R}^n$ and $\lambda \geq 0$ is the regularization parameter.

Subdifferential of the objective Ψ

$$\partial \Psi(w) = \big\{ w - y + \lambda u \mid u \in \partial \|w\|_1 \big\}.$$

At the optimum w^*, $0 \in \Psi(w^*)$, that is there exists $u \in \partial \|w^*\|_1$ such that

$$w_i^* = y_i - \lambda u_i.$$

This yields the so-called soft shrinkage solution:

$$w_i^* = \mathrm{sign}(y_i)\,(|y_i| - \lambda)_+.$$

KKT optimality conditions (non-smooth case)

Theorem

- f, g_i are convex and $h_j(x) = \langle a_j, x\rangle - b_j$.
- strong duality holds.

Then **necessary and sufficient** conditions for primal and dual optimal points x^* and (λ^*, μ^*) are the **Karush-Kuhn-Tucker(KKT) conditions**

$$g_i(x^*) \leq 0,\ i = 1, \ldots, r, \qquad h_j(x^*) = 0,\ j = 1, \ldots, s,$$
$$\lambda_i^* \geq 0,\ i = 1, \ldots, r \qquad \lambda_i^* g_i(x^*) = 0,\ i = 1, \ldots, r$$
$$0 \in \partial f(x^*) + \sum_{i=1}^{r} \lambda_i^* \partial g_i(x^*) + A^T \mu^*.$$

Optimization problem with perturbed constraints:

$$\min_{x \in D} f(x),$$
$$\text{subject to: } g_i(x) \leq u_i, \quad i = 1, \ldots, r$$
$$h_j(x) = v_j, \quad j = 1, \ldots, s.$$

How sensitive is p^* to a slight variation of the constraints ?

Optimization problem with perturbed constraints:

$$\min_{x \in D} f(x),$$

$$\text{subject to: } g_i(x) \leq u_i, \quad i = 1, \ldots, r$$

$$h_j(x) = v_j, \quad j = 1, \ldots, s.$$

- $p^*(u,v)$ is the primal optimal value of the perturbed problem, where $p^* = p^*(0,0)$,
- If the original problem is convex, then the function $p^*(u,v)$ is convex in u and v.

Proposition 2. *Suppose that strong duality holds and the dual optimum is attained. Let* (λ^*, μ^*) *be dual optimal for the unperturbed problem, that is* $u = 0$ *and* $v = 0$. *Then*

$$p^*(u, v) \geq p(0, 0) - \langle \lambda^*, u \rangle - \langle \mu^*, v \rangle, \quad \forall\, u \in \mathbb{R}^r,\; v \in \mathbb{R}^s.$$

If additionally $p^*(u, v)$ *is* ***differentiable in*** u ***and*** v, *then*

$$\lambda_i^* = -\frac{\partial p^*}{\partial u_i}, \qquad \mu_j^* = -\frac{\partial p^*}{\partial v_j}, \qquad at \quad (u, v) = (0, 0).$$

Interpretation:

- 1. λ_i^* is large and $u_i < 0$ then $p^*(u, v)$ will increase strongly,
  2. λ_i^* is small and $u_i > 0$ then $p^*(u, v)$ will not decrease too much,
  3. $|\mu_i^*|$ is large and $\operatorname{sign} v_i = -\operatorname{sign} \mu_i$ then $p^*(u, v)$ will increase strongly,
  4. $|\mu_i^*|$ is small and $\operatorname{sign} v_i = -\operatorname{sign} \mu_i$ then $p^*(u, v)$ will decrease little,

Proof: Let x be any feasible point for the perturbed problem, that is $g_i(x) \leq u_i$, $i = 1, \ldots, r$ and $h_j(x) = v_j$, $j = 1, \ldots, s$. Then by strong duality,

$$p^*(0,0) = q(\lambda^*, \mu^*) \leq f(x) + \sum_{i=1}^{r} \lambda_i^* g_i(x) + \sum_{j=1}^{s} \mu_j^* h_j(x)$$
$$\leq f(x) + \langle \lambda^*, u \rangle + \langle \mu^*, v \rangle ,$$

using the definition of $q(\lambda, \mu)$ and $\lambda^* \succeq 0$. Thus

$$\forall \text{ feasible } x : \quad f(x) \geq p(0,0) - \langle \lambda^*, u \rangle - \langle \mu^*, v \rangle ,$$

$$\Longrightarrow \quad p^*(u,v) \geq p(0,0) - \langle \lambda^*, u \rangle - \langle \mu^*, v \rangle .$$

The derived inequality states that, $p^*(te_i, 0) - p^* \geq -t\,\lambda_i^*$, and thus

$$\forall t > 0, \qquad \frac{p^*(te_i,0) - p^*}{t} \geq -\lambda_i^*, \qquad \forall t < 0, \qquad \frac{p^*(te_i,0) - p^*}{t} \leq -\lambda_i^*,$$

and thus since $p^*(u,v)$ is differentiable by assumption we have $\frac{\partial p^*}{\partial u_i} = -\lambda_i^*$.

Dependency of the dual on the primal problem: The norm approximation problem

$$\min_x \|Ax - b\|_p \,,$$

where $b \in \mathbb{R}^n$, $x \in \mathbb{R}^m$ and $A \in \mathbb{R}^{n \times m}$.

Interpretation:

- find the solution to the linear system $Ax = b$ if such a solution exists, if not find the best approximation with respect to the chosen p-norm,
- find the projection of b onto the subspace S spanned by the columns of A,

$$S = \Big\{ y = \sum_{i=1}^{m} a_i y_i \,|\, a_i \in \mathbb{R}^n, \quad A = (a_1, \ldots, a_m) \Big\},$$

with respect to the p-norm.

Dependency of the dual on the primal problem:

$$\min_x \|Ax - b\|_p$$

a) Lagrangian: $L(x) = \|Ax - b\| \Longrightarrow$ dual function $q = \inf_{x \in \mathbb{R}^m} \|Ax - b\|_p$.

Dependency of the dual on the primal problem:

b) Introduction of a new equality constraint:

$$\min_{x\in\mathbb{R}^m,\, y\in\mathbb{R}^n} \ \|y\|_p \,,$$

$$\text{subject to: } Ax - b = y.$$

Lagrangian: $L(x, y, \mu) = \|y\|_p + \langle \mu, Ax - b - y\rangle \,,$

$$\inf_{x\in\mathbb{R}^m} \langle \mu, Ax\rangle = 0, \ \text{ if } A^T\mu = 0, \ \text{otherwise } -\infty.$$

Hölder's ineq.: $(\frac{1}{q} + \frac{1}{p} = 1)$: $\langle \mu, y\rangle \leq \|\mu\|_q \|y\|_p \,,$ equality is attained for y^*,

$$\inf_{y\in\mathbb{R}^n} \|y\|_p - \langle \mu, y\rangle = \|y^*\|_p (1 - \|\mu\|_q) = 0, \ \text{ if } \|\mu\|_q \leq 1, \ \text{otherwise } -\infty.$$

$$\max_{\mu\in\mathbb{R}^n} \ \langle \mu, b\rangle$$

$$\text{subject to: } \|\mu\|_q \leq 1, \qquad A^T\mu = 0.$$

Dependency of the dual on the primal problem:

c) Strictly monotonic transformation of the objective of the primal problem

$$\min_{x\in\mathbb{R}^m,\, y\in\mathbb{R}^n} \quad \|y\|_p^2\,,$$
$$\text{subject to: } Ax - b = y.$$

Lagrangian: $L(x, y, \mu) = \|y\|_p^2 + \langle \mu, Ax - b - y\rangle\,,$

$$\text{Hölder's inequality: } \inf_{y\in\mathbb{R}^n} \|y\|_p^2 - \langle \mu, y\rangle = \|y^*\|_p^2 - \|\mu\|_q \|y^*\|_p\,.$$

Minimum of quadratic function: $\|y^*\| = \frac{1}{2}\|\mu\|_q$ and the value is:
$\frac{1}{4}\|\mu\|_q^2 - \frac{1}{2}\|\mu\|_q^2 = -\frac{1}{4}\|\mu\|_q^2,$

$$\max_{\mu\in\mathbb{R}^n} \; -\frac{1}{4}\|\mu\|_q^2 + \langle \mu, b\rangle$$
$$\text{subject to: } A^T\mu = 0.$$

Weak alternatives: Feasibility problem for general optimization problem:

$$\min_{x \in \mathbb{R}^n} 0$$
$$\text{subject to: } g_i(x) \leq 0, \quad i = 1, \ldots, r,$$
$$h_j(x) = 0, \quad j = 1, \ldots, s.$$

Primal optimal value: $p^* = \begin{cases} 0, & \text{if optimization problem is feasible} , \\ \infty, & \text{else.} \end{cases}$.

Dual function: $q(\lambda, \mu) = \inf_{x \in D} \Big(\sum_{i=1}^r \lambda_i g_i(x) + \sum_{j=1}^s \mu_j h_j(x) \Big)$.

$$\text{Dual problem: } \max_{\lambda \in \mathbb{R}^r, \mu \in \mathbb{R}^s} q(\lambda, \mu)$$
$$\text{subject to: } \lambda \succeq 0.$$

Dual optimal value: $d^* = \begin{cases} \infty, & \text{if } \lambda \succeq 0 \text{ and } q(\lambda, \mu) > 0 \text{ is feasible} , \\ 0, & \text{else.} \end{cases}$

By weak duality: $d^* \leq p^*$.

- If the dual problem is feasible ($d^* = \infty$) then the primal problem must be infeasible,
- If the primal problem is feasible ($p^* = 0$) then the dual problem is infeasible.

$\Longrightarrow$ **at most one** of the system of inequalities is feasible,

- $g_i(x) \leq 0, \quad i = 1, \ldots, r, \qquad h_j(x) = 0, \quad j = 1, \ldots, s,$
- $\lambda \succeq 0, \qquad q(\lambda, \mu) > 0.$

Definition 8. *An inequality system where at most one of the two holds is called* ***weak alternatives****.*

Note: the case $d^* = \infty$ and $p^* = \infty$ can also happen

Strong alternatives:

- optimization problem is convex (g_i convex and h_j affine),
- there exists an $x' \in \operatorname{relint} D$ such that $Ax' = b$.

Two sets of inequalities

- $g_i(x) \leq 0, \quad i = 1, \ldots, r, \qquad Ax = b$
- $\lambda \succeq 0, \qquad q(\lambda, \mu) > 0.$

Under the above condition exactly one of them holds:

Strong alternatives

Replace inequality constraint:

$$g_i(x) \leq 0 \quad \Longrightarrow \quad g_i(x) \preceq_K 0.$$

Optimization problem with generalized inequality constraint:

$$\begin{aligned} &\min_{x \in D} f(x), \\ \text{subject to: } & g_i(x) \preceq_{K_i} 0, \quad i = 1, \ldots, r, \\ & h_j(x) = 0, \quad j = 1, \ldots, s, \end{aligned}$$

where $K_i \subset \mathbb{R}^{k_i}$ are proper cones.

Almost all properties carry over with only minor changes !

Dual cone K^*

$$K^* := \{y \mid \langle x, y\rangle \geq 0, \, \forall\, x \in K\}.$$

Dual cone of S_+^n

$$y \in (S_+^n)^* \iff \operatorname{tr}(XY) \geq 0, \; \forall\, X \in S_+^n.$$

Now with $X = \sum_i \lambda_i u^i (u^i)^T$,

$$\operatorname{tr}(XY) = \operatorname{tr}\Big(\sum_i \lambda_i u^i (u^i)^T\Big) = \sum_i \lambda_i \operatorname{tr}\big(u^i (u^i)^T Y\big)$$
$$= \sum_i \lambda_i \sum_{r,s} u_r^i u_s^i Y_{rs} = \sum_i \lambda_i \langle u_i, Y u_i\rangle$$

If $Y \notin S_+^n$ there exists q such that $\langle q, Yq\rangle < 0 \implies X = qq^T$, $\operatorname{tr}(XY) < 0$.

The dual cone of S_+^n is S_+^n (self-dual).

Generalizes inequalities II

- **Lagrangian:** for, $g_i(x) \preceq_{K_i} 0$, we get a Lagrange multiplier $\lambda \in \mathbb{R}^{k_i}$.

$$L(x, \lambda, \mu) = f(x) + \sum_{i=1}^{r} \langle \lambda_i, g_i(x) \rangle + \sum_{j=1}^{s} \mu_j h_j(x).$$

- **Dual function:** $\lambda_i \geq 0 \quad \Longrightarrow \quad \lambda_i \succeq_{K_i^*} 0, \quad (K_i^*$ dual cone of $K_i)$.
  Note: $\lambda_i \succeq_{K_i^*} 0$ and $g_i(x) \preceq_{K_i} 0 \Longrightarrow \langle \lambda_i, g_i(x) \rangle \leq 0$,

$$x \text{ feasible }, \lambda_i \succeq_{K_i^*} 0 \quad \Longrightarrow \quad f(x) + \sum_{i=1}^{r} \langle \lambda_i, g_i(x) \rangle + \sum_{j=1}^{s} \mu_j h_j(x) \leq f(x).$$

- **Dual problem:** The dual problem becomes

$$\max_{\lambda, \mu} q(\lambda, \mu),$$

$$\text{subject to: } \lambda_i \succeq_{K_i^*} 0, \quad i = 1, \ldots, r.$$

We have **weak duality**: $d^* \leq p^*$.

- **Slater's condition and strong duality:** for a convex primal problem

$$\min_{x \in D} f(x),$$

$$\text{subject to: } g_i(x) \preceq_{K_i} 0, \quad i = 1, \ldots, r$$

$$Ax = b, \ ,$$

where f is convex, g_i is K_i-convex.

Proposition 3. *If there exists an* $x \in \operatorname{relint} D$ *with* $Ax = b$ *and* $g_i(x) \prec_{K_i} 0$*, then **strong duality**,* $d^* = p^*$*, holds.*

Example: Lagrange dual of a semidefinite program:

$$\min_{x \in D} \langle c, x \rangle ,$$

$$\text{subject to: } \sum_{i=1}^{n} x_i F_i + G \preceq_{S_+^k} 0,$$

where $F_1, \ldots, F_n, G \in S_+^k$. The Lagrangian is

$$L(x, \lambda) = \langle c, x \rangle + \sum_{i=1}^{n} x_i \operatorname{tr}(\lambda F_i) + \operatorname{tr}(\lambda G) = \sum_{i=1}^{n} x_i \Big(c_i + \operatorname{tr}(\lambda F_i) \Big) + \operatorname{tr}(\lambda G),$$

where $\lambda \in S^k$ and thus the dual problem becomes

$$\max_{\lambda,\, \mu} \operatorname{tr}(\lambda G),$$

$$\text{subject to: } c_i + \operatorname{tr}(\lambda F_i) = 0, \quad i = 1, \ldots, n.$$

Generalizes inequalities V

- **Complementary slackness:** One has

$$\langle \lambda_i^*, g_i(x^*) \rangle = 0, \quad i = 1, \dots, r.$$

From this we deduce

$$\lambda_i^* \succ_{K_i^*} 0 \quad \Longrightarrow \quad g_i(x^*) = 0, \qquad\qquad g_i(x^*) \prec_{K_i} 0 \quad \Longrightarrow \quad \lambda_i^* = 0.$$

Important: the condition $\langle \lambda_i^*, g_i(x^*) \rangle = 0$ can be fulfilled if $\lambda_i^* \neq 0$ **and** $g_i(x^*) \neq 0$.

UNIVERSITÄT DES SAARLANDES

Generalizes inequalities VI

MACHINE LEARNING

- **KKT conditions:** f, g_i and h_j are differentiable:

Proposition 4. *If strong duality holds, the following KKT-conditions are necessary conditions for primal x^* and dual optimal (λ^*, μ^*) points,*

$$g_i(x^*) \leq 0, \quad i = 1, \ldots, r, \qquad h_j(x^*) = 0, \quad j = 1, \ldots, s,$$

$$\lambda_i^* \succeq_{K_i^*} 0, \quad i = 1, \ldots, r \qquad \langle \lambda_i^*, g_i(x^*) \rangle = 0, \quad i = 1, \ldots, r$$

$$\nabla f(x^*) + \sum_{i=1}^{r} D g_i(x^*)^T \lambda_i^* + \sum_{j=1}^{s} \mu_j^* \nabla h_j(x^*) = 0.$$

If the problem is convex, then the KKT-conditions are necessary and sufficient for optimality of λ^*, μ^*.