Convex Optimization

Fallstudien der mathematische Modelbildung, Teil 2

20.10.2023 - 21.11.2023,

paul.catala@tum.de

Our Purpose

- Study the properties of
$$\min \frac{1}{2}\|Ax - y\|^2 + J(x)$$
where J is a convex function.
- Tikhonov has the advantage of having an analytical solution. That is in general not true.
- We need some tools from convex optimization.

- ***Definition (Convex Set).*** A set C is convex if

$$x, y \in C \implies \theta x + (1-\theta) y \in C \quad \forall \theta \in [0,1]$$

Example (Convex hull). The convex hull of a set D, denoted $\mathbf{conv}\, D$, is the set of all linear combinations of elements of D.

- ***Definition (Convex function).*** A function $f : \mathbb{R}^n \to \mathbb{R}$ is convex if

$$f(\theta x + (1-\theta) y) \leqslant \theta f(x) + (1-\theta) f(y), \quad \forall \theta \in [0,1]$$

The sublevel sets of a convex function, *i.e.* $\{x \,;\, f(x) \leqslant \alpha\}$, are convex sets.
Example. $\|\cdot\|_p$ is convex for $p \geqslant 1$. Quadratic functions

$$x^\top P x + q^\top x + r$$

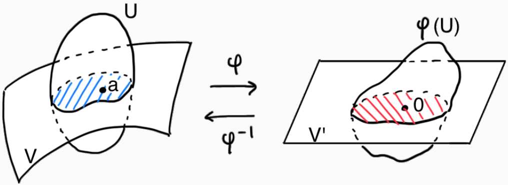
are convex.

- ***Proposition (Tangent).*** A differentiable $f : \mathbb{R}^n \to \mathbb{R}$ is convex if and only if

$$f(y) \geqslant f(x) + \langle \nabla f(x), y - x \rangle, \quad \forall x, y$$

Proposition (Minimum). x is a (global) minimizer of a differentiable convex function if and only if $\nabla f(x) = 0$

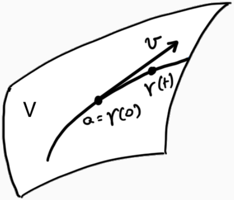
- ***Definition (Smooth Manifold).*** Let $V \subset \mathbb{R}^n$, $a \in V$ and $d \in \mathbb{N}$. We say that V is a smooth in a if there exists a C^1-diffeomorphism φ from an open neighborhood $U \subset \mathbb{R}^n$ of a to an open neighborhood $\varphi(U) \subset \mathbb{R}^n$ of 0 such that $\varphi(V)$ is a vector space of dimension d. We say that V is a (smooth) manifold if V is smooth at every point.



- ***Definition (Tangent Vector).*** Let $V \subset \mathbb{R}^n$ and $a \in V$. A vector $v \in \mathbb{R}^n$ is tangent to V at a if there exists a differentiable function $\gamma : I \to \mathbb{R}^n$ (where I is an open interval containing 0) such that

$$\gamma(I) \subset V, \quad \gamma(0) = a \quad \text{and} \quad \gamma'(0) = v.$$

  In other words, v is tangent to a path on V going going through a.



- ***Theorem (Tangent Space).*** If V is smooth in a, the tangent vectors span a d-dimensional vector space, called tangent space.

- Consider the generic optimization problem

$$\min f_0(x) \quad \text{s.t.} \quad x \in V$$

where V is a smooth manifold. If f_0 has a minimum in x_* on V, then for any differentiable path $\gamma : [0, 1] \to \mathbb{R}^n$ on V with $\gamma(t_*) = x_*$, the first-order optimality criterion gives

$$(f \circ \gamma)'(t_*) = f'(t_*) \cdot \gamma'(t_*) = 0.$$

In other words, $\nabla f(t_*)$ is orthogonal to the tangent space of V at t_*.

Optimization Under Equality Constraints

- Let $f_0, h_1, \ldots, h_m : \mathbb{R}^n \to \mathbb{R}$ be C^1-differentiable

$$\min f_0(x) \quad \text{s.t.} \quad h_i(x) = 0, i = 1, \ldots, m \tag{1}$$

$\Rightarrow$ admit that $h_i(x) = 0 \ \forall i$ define a smooth manifold, whose tangent space at a is given by

$$\text{Span}(\nabla h_1(a), \ldots, \nabla h_m(a))^{\perp}.$$

- Let $f_0, h_1, \ldots, h_m : \mathbb{R}^n \to \mathbb{R}$ be C^1-differentiable

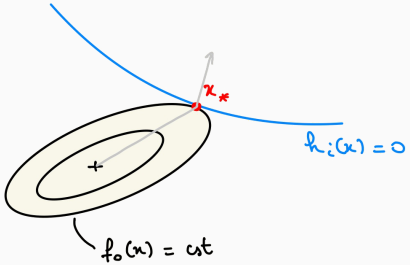
$$\min f_0(x) \quad \text{s.t.} \quad h_i(x) = 0, i = 1, \ldots, m \tag{1}$$

$\Rightarrow$ admit that $h_i(x) = 0\ \forall i$ define a smooth manifold, whose tangent space at a is given by

$$\text{Span}(\nabla h_1(a), \ldots, \nabla h_m(a))^\perp.$$

- Therefore, an optimal point for (1) must satisfy

$$\begin{cases} \nabla f_0(x_*) = \sum_{i=1}^m \nu_i \nabla h_i(x_*) \quad (\text{for some } \nu_i \in \mathbb{R}) & \text{(2a)} \\ h_1(x_*) = \ldots = h_m(x_*) = 0 & \text{(2b)} \end{cases}$$



Lagrangian

- We define

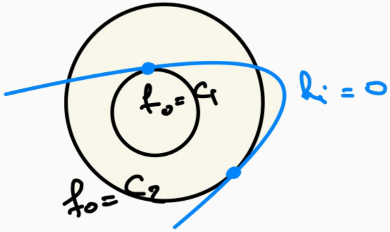
$$\mathcal{L}(x,\nu) = f_0(x) + \sum \nu_i h_i(x)$$

Then

$$\begin{aligned} \nabla_x \mathcal{L} = 0 \quad &\Leftrightarrow \quad (2a) \\ \nabla_\nu \mathcal{L} = 0 \quad &\Leftrightarrow \quad (2b) \end{aligned}$$

- *Definition (Lagrangian).*
  - $\mathcal{L}$ is the Lagrangian associated with problem (1)
  - ν_i are the Lagrange multipliers
  - (2a), (2b) are the first order optimality conditions (Karush-Kuhn-Tucker conditions)

Stationarity Conditons are not Sufficient



Dual problem

- The minimization/maximization of $\mathcal{L}$ (now unconstrained) with respect to x/ν yields information on the optimum. Consider:

$$g(\nu) = \min_x \mathcal{L}(x, \nu)$$

Note that for all feasible $\tilde{x}$, $g(\nu) \leqslant \mathcal{L}(\tilde{x}, \nu) = f_0(\tilde{x})$, hence $g(\nu) \leqslant f(x_*)$.
Note also that g is a minimum of affine functions, hence it is concave.

Dual problem

- The minimization/maximization of $\mathcal{L}$ (now unconstrained) with respect to x/ν yields information on the optimum. Consider:

$$g(\nu) = \min_x \mathcal{L}(x, \nu)$$

Note that for all feasible $\tilde{x}$, $g(\nu) \leqslant \mathcal{L}(\tilde{x}, \nu) = f_0(\tilde{x})$, hence $g(\nu) \leqslant f(x_*)$.
Note also that g is a minimum of affine functions, hence it is concave.

- We consider the following problem

$$\max_\nu g(\nu) \tag{3}$$

In particular, $g(\nu_*)$ provides a lower bound on the optimal value of the objective $f_0(x)$.

Dual problem

- The minimization/maximization of $\mathcal{L}$ (now unconstrained) with respect to x/ν yields information on the optimum. Consider:

$$g(\nu) = \min_x \mathcal{L}(x, \nu)$$

Note that for all feasible $\tilde{x}$, $g(\nu) \leqslant \mathcal{L}(\tilde{x}, \nu) = f_0(\tilde{x})$, hence $g(\nu) \leqslant f(x_*)$.
Note also that g is a minimum of affine functions, hence it is concave.

- We consider the following problem

$$\max_\nu g(\nu) \tag{3}$$

In particular, $g(\nu_*)$ provides a lower bound on the optimal value of the objective $f_0(x)$.

- ***Definition (Lagrange duality).***
  - $\max_\nu g(\nu)$ is the dual problem
  - $\min_x f_0(x)$ s.t. $h_i(x) = 0$ is the primal problem
  - $f_0(x_*) - g(\nu_*)$ is the duality gap
  - $f_0(x_*) > g(\nu_*)$ weak duality
  - $f_0(x_*) = g(\nu_*)$ strong duality, and then $\max_\nu \min_x \mathcal{L}(x, \nu) = \min_x \max_\nu \mathcal{L}(x, \nu)$

- In convex optimization, we only deal with affine constraints $\langle a_i, x\rangle = y_i$. The constraint set is then an affine space, with tangent space $\operatorname{Ker} A$. The optimality condition (2a) thus states that $\nabla f_0(x_*) \in (\operatorname{Ker} A)^\perp = \operatorname{Ran} A^\top$.

- ***Theorem.*** Assume f_0 differentiable, convex, and $h_i(x)$ affine. Then KKT conditions are sufficient, and strong duality holds.

Proof. Assume x_*, ν_* satisfy the KKT conditions. Since all functions are convex and differentiable, $\mathcal{L}(x, \nu_*)$ is convex differentiable in x and condition (2a) states that x_* minimizes it. Hence,

$$g(\nu_*) = \mathcal{L}(x_*, \nu_*) = f_0(x_*) + \sum \nu_i^* h_i(x_*) = f_0(x_*)$$

where the last equality holds because x_* is feasible from (2b). Hence f_0 is minimal at x_* and strong duality holds.

- *Example.* For $A \in \mathbb{R}^{m \times n}$ and $y \in \operatorname{Ran} A$ we consider the problem

$$\min \frac{1}{2}\|x\|_2^2 \quad \text{s.t.} \quad Ax = y$$

Then $\mathcal{L}(x, \nu) = \frac{1}{2}\|x\|^2 + \nu^\top (y - Ax)$, and the optimality conditions read

$$\begin{cases} x = A^\top \nu \\ Ax = y \end{cases}$$

The dual problem is

$$\max_{\nu} \nu^\top y - \frac{1}{2}\|A^\top \nu\|^2$$

- *Example (Least-squares).* For $A \in \mathbb{R}^{m \times n}$ and $y \in \mathbb{R}^m$, consider

$$\min \frac{1}{2}\|Ax - y\|^2 + \frac{\lambda}{2}\|x\|^2$$

then (see exercise)

- A convex optimization problem is of the form

$$\min f_0(x) \quad \text{s.t.} \quad \begin{cases} f_i(x) \leqslant 0 & i = 1, \ldots, p \\ a_i^\top x = y_i, & i = 1, \ldots, m \end{cases}$$

where $f_0, f_1, \ldots, f_p : \mathbb{R}^n \to \mathbb{R}$ are convex.

- The Lagrangian is given by

$$\mathcal{L}(x, \mu, \nu) = f_0(x) + \sum \mu_i f_i(x) + \sum \nu_i h_i(x)$$

We have that $g(\mu, \nu) = \inf_x \mathcal{L}(x, \mu, \nu) \leqslant \mathcal{L}(\tilde{x}, \mu, v) \leqslant f_0(\tilde{x})$ provided $\mu_i \geqslant 0$.

- A convex optimization problem is of the form

$$\min f_0(x) \quad \text{s.t.} \quad \begin{cases} f_i(x) \leqslant 0 & i = 1, \dots, p \\ a_i^\top x = y_i, & i = 1, \dots, m \end{cases}$$

where $f_0, f_1, \dots, f_p : \mathbb{R}^n \to \mathbb{R}$ are convex.

- The Lagrangian is given by

$$\mathcal{L}(x, \mu, \nu) = f_0(x) + \sum \mu_i f_i(x) + \sum \nu_i h_i(x)$$

We have that $g(\mu, \nu) = \inf_x \mathcal{L}(x, \mu, \nu) \leqslant \mathcal{L}(\tilde{x}, \mu, v) \leqslant f_0(\tilde{x})$ provided $\mu_i \geqslant 0$.

- The dual problem is

$$\max_{\mu, \nu} g(\mu, \nu) \quad \text{s.t.} \quad \mu \succeq 0$$

KKT Conditions

- Assume strong duality holds. Then

$$f_0(x_*) = g(\mu_*, \nu_*) \leqslant \mathcal{L}(x_*, \mu_*, \nu_*) = f_0(x_*) + \sum (\mu_*)_i f_i(x_*).$$

  Hence $0 \leqslant \sum (\mu_*)_i f_i(x_*)$, which is only possible if

$$(\mu_*)_i f_i(x_*) = 0, \quad \forall i.$$

  This is called complementary slackness.

KKT Conditions

- Assume strong duality holds. Then

$$f_0(x_*) = g(\mu_*, \nu_*) \leqslant \mathcal{L}(x_*, \mu_*, \nu_*) = f_0(x_*) + \sum (\mu_*)_i f_i(x_*).$$

  Hence $0 \leqslant \sum (\mu_*)_i f_i(x_*)$, which is only possible if

$$(\mu_*)_i f_i(x_*) = 0, \quad \forall i.$$

  This is called complementary slackness.

- **KKT conditions** must hold at optimality when strong duality holds.

  - $\nabla f_0(x) + \sum \mu_i \nabla f_i(x) + \sum \nu_i \nabla h_i(x) = 0 \quad$ (stationarity)
  - $h_i(x) = 0 \quad$ (primal feasability)
  - $f_i(x) \leqslant 0 \quad$ (primal feasability)
  - $\mu_i \geqslant 0 \quad$ (dual feasability)
  - $\mu_i f_i(x) = 0 \quad$ (complementary slackness)

- In many practical situations, one does not need to check these conditions

Theorem (Slater's condition). Let $f_0, f_1, \ldots, f_p$ convex and h_i affine. If there exists x such that $f_i(x) < 0$ and $Ax = b$ (strict feasability), then strong duality holds.

- Perturbed problem

$$\min f_0(x) \quad \text{s.t.} \quad f_i(x) \leqslant u_i, \quad h_i(x) = v_i$$

with optimal value $p(u, v)$.
Assume that the problem is feasible for small u_i, v_i, and that strong duality holds.

- ***Proposition.*** Globally, $p(u, v) \geqslant p(0, 0) - \langle \mu_*, u \rangle - \langle \nu_*, v \rangle$

Proof. We have by strong duality, for any feasible x

$$p(0,0) = g(\mu_*, \nu_*) \leqslant f_0(x) + \sum \mu_{*i} f_i(x) + \sum \nu_{*i} h_i(x) \leqslant f_0(x) + \langle \mu_*, u \rangle + \langle \nu_*, v \rangle$$

and hence

$$p(u, v) \geqslant f_0(x) \geqslant p(0,0) - \langle \mu_*, u \rangle - \langle \nu_*, v \rangle. \qquad \square$$

- Perturbed problem

$$\min f_0(x) \quad \text{s.t.} \quad f_i(x) \leqslant u_i, \quad h_i(x) = v_i$$

with optimal value $p(u, v)$.
Assume that the problem is feasible for small u_i, v_i, and that strong duality holds.

- ***Proposition.*** Globally, $p(u, v) \geqslant p(0, 0) - \langle \mu_*, u \rangle - \langle \nu_*, v \rangle$

Proof. We have by strong duality, for any feasible x

$$p(0, 0) = g(\mu_*, \nu_*) \leqslant f_0(x) + \sum \mu_{*i} f_i(x) + \sum \nu_{*i} h_i(x) \leqslant f_0(x) + \langle \mu_*, u \rangle + \langle \nu_*, v \rangle$$

and hence

$$p(u, v) \geqslant f_0(x) \geqslant p(0, 0) - \langle \mu_*, u \rangle - \langle \nu_*, v \rangle. \qquad \square$$

- ***Proposition.*** Locally, $\frac{\partial p}{\partial u_i}(0, 0) = -\mu_*$ and $\frac{\partial p}{\partial v_i}(0, 0) = -\nu_*$

Proof. Setting $u = te_i$ and $v = 0$ in the previous inequality, we obtain

$$\frac{p(te_i, 0) - p(0, 0)}{t} \geqslant -\mu_{*i} \quad \text{if} \quad t > 0 \quad \text{and} \quad \frac{p(te_i, 0) - p(0, 0)}{t} \leqslant -\mu_{*i} \quad \text{if} \quad t < 0.$$

Taking the limit yields the desired limit. The same reasoning applies for ν. $\square$

Sensitivity Analysis and Dual

- Perturbed problem

$$\min f_0(x) \quad \text{s.t.} \quad f_i(x) \leqslant u_i, \quad h_i(x) = v_i$$

with optimal value $p(u, v)$.
Assume that the problem is feasible for small u_i, v_i, and that strong duality holds.

- ***Proposition.*** Globally, $p(u, v) \geqslant p(0, 0) - \langle \mu_*, u \rangle - \langle \nu_*, v \rangle$

Proof. We have by strong duality, for any feasible x

$$p(0, 0) = g(\mu_*, \nu_*) \leqslant f_0(x) + \sum \mu_{*i} f_i(x) + \sum \nu_{*i} h_i(x) \leqslant f_0(x) + \langle \mu_*, u \rangle + \langle \nu_*, v \rangle$$

and hence

$$p(u, v) \geqslant f_0(x) \geqslant p(0, 0) - \langle \mu_*, u \rangle - \langle \nu_*, v \rangle. \qquad \square$$

- ***Proposition.*** Locally, $\frac{\partial p}{\partial u_i}(0, 0) = -\mu_* \quad$ and $\quad \frac{\partial p}{\partial v_i}(0, 0) = -\nu_*$

Proof. Setting $u = te_i$ and $v = 0$ in the previous inequality, we obtain

$$\frac{p(te_i, 0) - p(0, 0)}{t} \geqslant -\mu_{*i} \quad \text{if} \quad t > 0 \quad \text{and} \quad \frac{p(te_i, 0) - p(0, 0)}{t} \leqslant -\mu_{*i} \quad \text{if} \quad t < 0.$$

Taking the limit yields the desired limit. The same reasoning applies for ν. $\square$

- The dual variable gives indication on the sensitivity of the problem