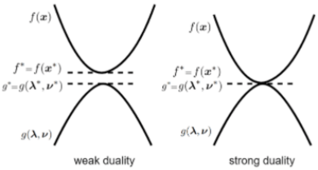
Optimization in Machine Learning

Nonlinear programs
Regularity Conditions



Learning goals

- KKT conditions
- Regularity conditions
- Examples

STATIONARY POINT OF THE LAGRANGIAN

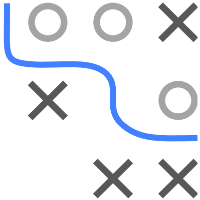
- When we introduced the Lagrangian $\mathcal{L}$ from a geometrical perspective for the equality constraint problem, we realized that the geometrical conditions for the optimum coincided with finding a stationary point of $\mathcal{L}$:

$$\begin{pmatrix} \nabla_{\boldsymbol{x}} \mathcal{L}(\boldsymbol{x}^*, \beta) \\ \nabla_{\beta} \mathcal{L}(\boldsymbol{x}^*, \beta) \end{pmatrix} = \begin{pmatrix} \nabla f(\boldsymbol{x}^*) + \beta \nabla h(\boldsymbol{x}^*) \\ h(\boldsymbol{x}) \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

- For the general Lagrangian, this leads to the following question:

Is $\nabla L(\boldsymbol{x}, \alpha, \boldsymbol{\beta}) = 0$ a **necessary** / **sufficient condition for the optimum**?

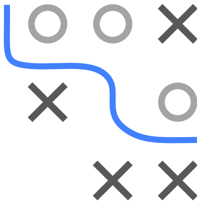
KKT CONDITIONS



- To formulate necessary and sufficient conditions for optimality, we need the **Karush-Kuhn-Tucker conditions** (KKT conditions)

- A triple $(\boldsymbol{x}, \boldsymbol{\alpha}, \boldsymbol{\beta})$ satisfies the KKT conditions if
  - $\nabla_x L(\boldsymbol{x}, \boldsymbol{\alpha}, \boldsymbol{\beta}) = 0$ (stationarity)
  - $g_i(\boldsymbol{x}) \leq 0, h_j(\boldsymbol{x}) = 0$ for all i, j (primal feasibility)
  - $\boldsymbol{\alpha} \geq 0$ (dual feasibility)
  - $\alpha_i g_i(\boldsymbol{x}) = 0$ for all i (complementary slackness)

KKT CONDITIONS



timality: Let $\boldsymbol{x}^*$ be a local minimum.
If certain regularity conditions are fulfilled, there are $\boldsymbol{\alpha}^*, \boldsymbol{\beta}^*$ such that $(\boldsymbol{x}^*, \boldsymbol{\alpha}^*, \boldsymbol{\beta}^*)$ fulfill the KKT conditions

- Under certain conditions, KKT conditions are also sufficient for optimality

timality: Given a **convex problem** (f convex, $\mathcal{S}$ convex) and $(\boldsymbol{x}^*, \boldsymbol{\alpha}^*, \boldsymbol{\beta}^*)$ satisfies the KKT conditions.
Then $\boldsymbol{x}^*$ is a global solution to the problem

REGULARITY CONDITIONS

- Different regularity conditions (or constraint qualifications) ensure that the KKT conditions apply (ACQ, LICQ, MFCQ, Slater condition, ...)
- To use the above results, at least one regularity condition must be examined to prove that the function behaves “regular”
- We do not go further into these regularity conditions here

RIDGE REGRESSION

- The following two formulas are common for ridge regression:

  Formula 1:

$$\min_{\boldsymbol{\theta}} \quad f_\lambda(\boldsymbol{\theta}) := \|\boldsymbol{y} - \mathbf{X}\boldsymbol{\theta}\|_2^2 + \lambda\|\boldsymbol{\theta}\|_2^2 \tag{1}$$

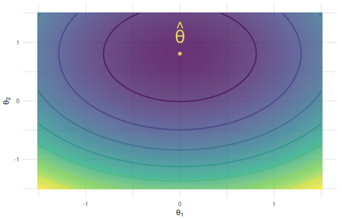
  Formula 2:

$$\begin{aligned} \min_{\boldsymbol{\theta}} \quad & \|\boldsymbol{y} - \mathbf{X}\boldsymbol{\theta}\|_2^2 \\ \text{s.t.} \quad & \|\boldsymbol{\theta}\|_2^2 - t \leq 0 \end{aligned} \tag{2}$$

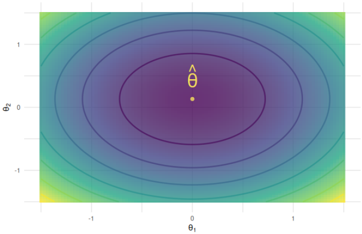
- Why are these two formulas (for appropriate values t, λ) equivalent?

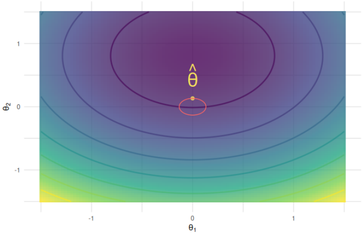
RIDGE REGRESSION – VISUALIZATION

- **Visualization:** see additional material



- Quadratic loss for the `cars` dataset without penalty





- Left: loss for ridge regression with penalty term
  Right: loss for ridge regression with corresponding constraint

RIDGE REGRESSION – EQUIVALENCE

- Consider (1). If $\boldsymbol{\theta}^*$ is our minimum, the necessary condition applies:

$$\nabla f_\lambda(\boldsymbol{\theta}^*) = -2\mathbf{y}^T\mathbf{X} + 2(\boldsymbol{\theta}^*)^T\mathbf{X}^T\mathbf{X} + 2\lambda(\boldsymbol{\theta}^*)^T = 0$$

- We show that we can find a t so that $\boldsymbol{\theta}^*$ is also solution for (2)
- We calculate the Lagrange function of (2):

$$L(\boldsymbol{\theta}, \alpha) = \|\mathbf{y} - \mathbf{X}\boldsymbol{\theta}\|_2^2 + \alpha(\|\boldsymbol{\theta}\|_2^2 - t)$$

- The first KKT condition (stationarity) is:

$$\nabla_\theta L(\boldsymbol{\theta}, \alpha) = -2\mathbf{y}^T\mathbf{X} + 2\boldsymbol{\theta}^T\mathbf{X}^T\mathbf{X} + 2\alpha\boldsymbol{\theta}^T = 0$$

- Since $\nabla f_\lambda(\boldsymbol{\theta}^*) = 0$, this is fulfilled if we set $\boldsymbol{\theta} = \boldsymbol{\theta}^*$ and $\alpha = \lambda$

RIDGE REGRESSION – EQUIVALENCE

- However, complementary slackness must still apply for the KKT conditions:
$$\alpha(\|\boldsymbol{\theta}\|_2^2 - t) = 0$$
- This is the case if we choose $t = \|\boldsymbol{\theta}^*\|^2$
- Vice versa it can be shown that a solution of (2) is a solution of (1) if we set $\lambda = \alpha$