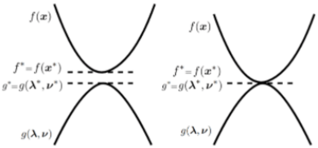
Optimization in Machine Learning

Nonlinear programs
Regularity Conditions

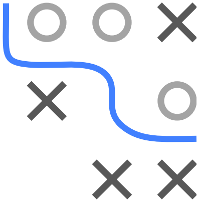

$f(\boldsymbol{x})$
$f^* = f(\boldsymbol{x}^*)$
$g^* = g(\boldsymbol{\lambda}^*, \boldsymbol{\nu}^*)$
$g(\boldsymbol{\lambda}, \boldsymbol{\nu})$
$f(\boldsymbol{x})$
$f^* = f(\boldsymbol{x}^*)$
$g^* = g(\boldsymbol{\lambda}^*, \boldsymbol{\nu}^*)$
$g(\boldsymbol{\lambda}, \boldsymbol{\nu})$

weak duality strong duality

Learning goals

- KKT conditions
- Regularity conditions
- Examples

STATIONARY POINT OF THE LAGRANGIAN



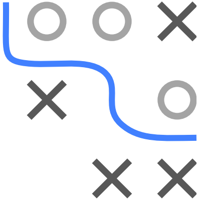
When we introduced the Lagrangian $\mathcal{L}$ from a geometrical perspective for the equality constraint problem, we realized that the geometrical conditions for the optimum coincided with finding a stationary point of $\mathcal{L}$:

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

For this and the general Lagrangian, this leads to the following question.

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

KKT CONDITIONS



In order to be able to formulate necessary and sufficient conditions for optimality, we need the **Karush-Kuhn-Tucker conditions** (KKT conditions).

A triple $(\mathbf{x}, \boldsymbol{\alpha}, \boldsymbol{\beta})$ satisfies the KKT conditions if

- $\nabla_x L(\mathbf{x}, \boldsymbol{\alpha}, \boldsymbol{\beta}) = 0$ (stationarity)
- $g_i(\mathbf{x}) \leq 0, h_j(\mathbf{x}) = 0$ for all i, j (primal feasibility)
- $\boldsymbol{\alpha} \geq 0$ (dual feasibility)
- $\alpha_i g_i(\mathbf{x}) = 0$ for all i (complementary slackness)

REGULARITY CONDITIONS

There are different regularity conditions (or constraint qualifications) that ensure that the KKT conditions apply (ACQ, LICQ, MFCQ, Slater condition, ...).

To be able 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 and refer to `https://docs.ufpr.br/~ademir.ribeiro/ensino/cm721/kkt.pdf`.

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?