Mathematics of Learning – Worksheet 7

- The exercise sheets will be uploaded every Monday. Solution sketches will be uploaded one week later.
- You can hand in your own solutions via StudOn and we correct them - this is not mandatory. Please hand in small groups of 2-3 students.
- For questions, please use the forum on StudOn since other students may have similar questions. If you have a more personal question about the exercises please send an email to ehsan.waiezi@fau.de or lars.weidner@fau.de respectively.

Exercise 1 [KKT and Repetition Regression problems].
A constrained optimization problem has necessary optimality conditions: The KKT-conditions. Given a constrained optimization problem

$$\begin{aligned} &\min f(x) && \\ \text{s.t.}\quad &g_i(x) \leq 0 && \text{for all } i \in I \\ &h_j(x) = 0 && \text{for all } j \in J \\ &x \in \mathbb{R}^n && \end{aligned}$$

whereby $f, g_i, h_j : \mathbb{R}^n \to \mathbb{R}$ for all $i \in I, j \in J$, I and J some finite index sets.
Theorem. If a point $\hat{x}$ is an optimal solution of the constrained optimization problem, and the gradients of the functions defining the constraints fulfill some regularity conditions (you do not have to care about in the moment) in $\hat{x}$, then there exist some numbers $\mu_i \in \mathbb{R}$ and $\lambda_j \in \mathbb{R}$ for all $i \in I$ and $j \in J$, for which

$$\begin{aligned} \nabla f(\hat{x}) + \sum_{i \in I} \mu_i \nabla g_i(\hat{x}) + \sum_{j \in J} \lambda_j \nabla h_j(\hat{x}) = 0 \\ g_i(\hat{x}) \leq 0 \text{ for all } i \in I \\ h_j(\hat{x}) = 0 \text{ for all } j \in J \\ \mu_i \geq 0 \text{ for all } i \in I \\ \mu_i g_i(\hat{x}) = 0 \text{ for all } i \in I. \end{aligned}$$

This system is also called the KKT-System (x is then again considered as a variable).
Remark: This theorem generalizes the technique to find a critical point for unconstrained optimization problems, "gradient equals zero", to constrained optimization problems. Similarly to that case, the KKT conditions are only necessary optimality conditions. However, if the objective function f is convex, the inequality constraints g_i are convex, and the equality constraints h_j are affine linear, every KKT point is an optimal solution to the minimization problem. This is the case for the regression problem in the following.

a) Consider the constrained optimization problem (alternative ridge regression) for data $X \in \mathbb{R}^{N \times p}, Y \in \mathbb{R}^N$,

$$\begin{aligned} \min ||X\beta - Y||^2 \\ \text{s.t. } ||\beta||^2 \leq t \\ \beta \in \mathbb{R}^p \end{aligned}$$

for some shrinkage parameter $t \in \mathbb{R}$ and calculate the corresponding KKT-System.
b) Prove or disprove: for every $t \in \mathbb{R}_{>0}$ there exists $\lambda \in \mathbb{R}_{\geq 0}$ such that an optimal solution of the alternative ridge regression problem corresponding to t, $\hat{\beta}$, is an optimal solution of the classical (unconstrained) ridge regression problem with parameter λ, i.e.,

$$\min_{\beta \in \mathbb{R}^p} ||X\beta - Y||^2 + \lambda||\beta||^2.$$

c) Prove or disprove: for every $\lambda \in \mathbb{R}_{\geq 0}$ there exists $t \in \mathbb{R}_{>0}$ such that an optimal solution of the classical (unconstrained) ridge regression problem with parameter λ, $\hat{\beta}$, is an optimal solution of the alternative ridge regression problem.

Exercise 2 [Examples].
Let $x = (1, 2, 3, 4, 5)^T$ and $y = (4, 2, 5, 7, 2)^T$.
Calculate $\beta_0, \beta_1, \beta_2, \beta_3 \in \mathbb{R}$ with $||\beta||_2 \leq 1$ such that $\sum_{i=1}^{5}(\beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \beta_3 x_i^3 - y_i)^2$ is minimal. (You need knowledge about constrained optimization (KKT-conditions) for this exercise. This will be explained later in the semester.)

Exercise 3 [Regression, Regularization and Scaling].
Consider some data $X \in \mathbb{R}^{N \times p}$ and $Y \in \mathbb{R}^N$, and the classical linear regression problem

$$\min_{\beta \in \mathbb{R}^p} ||X\beta - Y||^2,$$

and let $\hat{\beta}$ be the optimal solution of the optimization problem.

a) Prove that for any positive definite diagonal matrix $\Theta \in \mathbb{R}^{p \times p}$, the linear regression problem corresponding to data $X\Theta$ and Y is solved by $\Theta^{-1}\hat{\beta}$.
b) Consider the ridge regression problem

$$\min_{\beta \in \mathbb{R}^p} ||X\beta - Y||^2 + \lambda||\beta||^2.$$

Give a concrete example (consisting of data X, Y, a matrix Θ and a number λ) for which the statement of part a) does not hold.
Remark: This means, ridge regression is not invariant to scaling, in contrary to classical linear regression - for the latter it does not matter, i.e., in which units the data comes, for the first it does. This is not desirable.
c) Find and describe a scaling invariant regularized regression method (optimally based on ridge regression). Prove that it has property a).