Prof. Dr. Jan Rolfes, Ehsan Waiezi, Lars Weidne Winter term 23/24

Mathematics of Learning – Worksheet 7 - – Discussion on Nov. 30th/Dec. 01st, 2023

- 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.

Solution. a) The KKT-System is the following:

$$\begin{aligned} \nabla ||X\beta - Y||^2 + \mu \nabla (||\beta||^2 - t) = 0 \\ ||\beta||^2 - t \leq 0 \\ \mu \geq 0 \\ \mu(||\beta||^2 - t) = 0 \end{aligned}$$

The first equality evaluates to

$$X^T(X\beta - Y) + \mu\beta = 0.$$

b) Let $\hat{\beta}$ be an optimal solution of the alternative ridge regression problem for parameter t. Then there exists a multiplier $\hat{\mu}$ such that $(\hat{\beta}, \hat{\mu})$ solves the KKT-System. In case we choose $\lambda = \hat{\mu}$ for the classical (unconstrained) ridge regression problem, we get as necessary optimality condition the following:

$$\nabla(||X\beta - Y||^2 + \hat{\mu}||\beta||^2) = 0,$$

which evaluates to

$$X^T(X\beta - Y) + \hat{\mu}\beta = 0,$$

which is exactly the first equation in the KKT-System and is hence solved by $\hat{\beta}$. This is the desired result.

c) Let $\hat{\beta}$ be an optimal solution of the classical ridge regression problem for parameter λ. In case we choose $t = ||\hat{\beta}||^2$, $(\hat{\beta}, \lambda)$ solves the KKT-System of the alternative ridge regression problem and hence $\hat{\beta}$ is an optimal solution of it, which is the desired result.

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.)
Solution. We know that

$$\min_{\beta:||\beta||_2^2\leq 1} ||X\beta - y||_2^2$$

is a convex constrained, optimization problem with a convex objective (and the constraint is convex and Slater's condition is fulfilled), hence, a KKT-point is a global optimum. Hence, we are looking for (β, μ) solving the KKT conditions

$$\begin{aligned} X^T X\beta + \mu\mathbb{1}\beta - X^T y &= 0 \\ ||\beta||_2^2 &\leq 1 \\ \mu(||\beta||_2^2 - 1) &= 0 \\ \mu &\geq 0 \end{aligned}$$

We can distinguish between two cases: Either, this system has a solution with $||\beta||_2^2 = 1$, then μ can be whatever positive; or a solution with $||\beta||_2^2 < 1$, then μ has to be equal to 0, simplifying the first equation considerably. We can exclude the second case, since we have already calculated in a) that equation number 1 is (uniquely, since it is a full rank linear equation system) solved by $\beta = (17, -\frac{41}{2}, \frac{17}{2}, -1)^T$ in case $\mu = 0$, with a norm clearly too large.

So we are in the first case with $||\beta|| = 1$ and μ arbitary positive. We consider the function (which is defined for all μ such that $(X^T X + \mu\mathbb{1})$ has full rank)

$$f : \mu \mapsto ||(X^T X + \mu\mathbb{1})^{-1} X^T y||_2^2.$$

that maps μ to the corresponding $||\beta||_2^2$ that solves the first KKT-condition. We also realize, that this β solves the penalized problem $\min_\beta ||X\beta - y||^2 + \mu||\beta||$ as seen in the previous part. We show that f is non-increasing: Consider the opposite, i.e., we have for $\mu_1 < \mu_2$ two solutions of the corresponding penalized linear regression problem β_1 and β_2 with $||\beta_1||_2^2 < ||\beta_2||_2^2$. Then we could derive the following:

$$\begin{aligned} ||X\beta_1 - y||_2^2 + \mu_2||\beta_1||_2^2 &\geq ||X\beta_2 - y||_2^2 + \mu_2||\beta_2||_2^2 \text{ (since } \beta_2 \text{ optimal for } \mu_2) \\ &= ||X\beta_2 - y||_2^2 + \mu_1||\beta_2||_2^2 + (\mu_2 - \mu_1)||\beta_2||_2^2 \\ &\geq ||X\beta_1 - y||_2^2 + \mu_1||\beta_1||_2^2 + (\mu_2 - \mu_1)||\beta_2||_2^2 \\ &> ||X\beta_1 - y||_2^2 + \mu_1||\beta_1||_2^2 + (\mu_2 - \mu_1)||\beta_1||_2^2 \\ &= ||X\beta_1 - y||_2^2 + \mu_2||\beta_1||_2^2 \end{aligned}$$

and this is a contradiction. Hence, we can look for the μ for which the function gets 1, by, e.g., an appropriate bisection search approach (we find a point for which the function gets below 1, and a point for which the function gets over 1, take the middle, iterate. This leads to $\mu = 7.965593$.
The corresponding β is determined by solving the normal equation

$$(X^T X + \mu\mathbb{1})\beta = X^T y,$$

hence, $\beta = (0.3252, 0.4787, 0.8001, -0.1574)$.

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).

Solution. a) A solution of the linear regression problem for data $X\Theta$, Y has to fulfill the equation

$$\Theta X^T X\Theta\beta = \Theta X^T Y.$$

We insert $\Theta^{-1}\hat{\beta}$ and multiply Θ^{-1} from left to get

$$\Theta^{-1}\Theta X^T X\Theta\Theta^{-1}\hat{\beta} = \Theta^{-1}\Theta X^T Y \Leftrightarrow X^T X\hat{\beta} = X^T Y$$

The last equation holds by prerequisite.

b) We can do the following to demonstrate that ridge regression is not scaling invariant. We investigate the weight of some object measured in kg with the same weight of the same object in kg. This is perfectly linear. We consider two data points: $(X_1, Y_1) = (0, 0)$ and $(X_2, Y_2) = (1, 1)$. We choose $\lambda = 1$. Then the solution of the regression problem is the solution of

$$X^T X\beta + 1\beta = X^T Y \Leftrightarrow \beta + \beta = 1,$$

Hence the optimal solution is $\hat{\beta} = 0.5$. On the other hand, in case we scale X by factor 1000 (basically, measuring X, but not Y, in g instead of kg), we get, inserting $\frac{1}{1000}\hat{\beta}$

$$1000X^T X 1000\frac{1}{1000}\hat{\beta} + \frac{1}{1000}\hat{\beta} = 1000X^T Y$$

which evaluates to

$$1000.001\hat{\beta} = 1000$$

which is not solved by 0.5. This is the desired result.

c) There are most certain many ways to implement a scaling invariant method which works like ridge regression. One (certainly not the best) is: Before doing the ridge regression, scale each data feature by its second moment,

$$\tilde{X}_i = \frac{X_i}{||X_i||_2}$$

which is feasible for each column not being the zero column, which you would sort out beforehand anyways. Calculate $\tilde{X}_i$ for some scaled X_i, i.e., θX_i for some $\theta > 0$, we get

$$\frac{\theta X_i}{||\theta X_i||_2} = \frac{\theta X_i}{\theta ||X_i||_2} = \tilde{X}_i.$$

Afterwards we calculate the ridge regression. This procedure delivers the same result, no matter how the initial data is scaled, which is our desired result.