Machine Learning	Prof. Matthias Hein
Exercise Sheet 4 - 20.5.2021 (16 Points)	due: 3.6.2021

Exercise 7 - Projected gradient descent and Lasso

Let $y \in \mathbb{R}^n$ be the n outputs and $\Phi \in \mathbb{R}^{n \times D}$ the design matrix of a regression problem (D basis functions $\phi_1, \ldots, \phi_D$, then $\Phi_{ij} = \phi_j(x_i)$, $i = 1, \ldots, n$, $j = 1, \ldots, D$).

In the lecture, introducing the new variables $w^+, w^- \in \mathbb{R}_+^D$ such that $w = w^+ - w^-$, the Lasso problem

$$\min_{w \in \mathbb{R}^D} \frac{1}{n} \|Y - \Phi w\|_2^2 + \lambda \|w\|_1 \tag{1}$$

has been rewritten into the optimization problem

$$\begin{aligned} \min_{w^+, w^- \in \mathbb{R}^D} \quad & \frac{1}{n} \left\|Y - \Phi w^+ + \Phi w^-\right\|_2^2 + \lambda \sum_{i=1}^D w_i^+ + \lambda \sum_{i=1}^D w_i^- \\ \text{subject to:} \quad & w_i^+ \geq 0, \ i = 1, \ldots, D, \\ & w_i^- \geq 0, \ i = 1, \ldots, D. \end{aligned} \tag{2}$$

a. The projection $P_C : \mathbb{R}^d \to \mathbb{R}$ onto a convex set C is defined for $x \in \mathbb{R}^d$ as

$$P_C(x) := \arg\min_{y \in C} \frac{1}{2} \|x - y\|_2^2 .$$

1. **(2 points)** Show that the projection onto a convex set is uniquely defined (Under which condition on the objective is the global minimum unique ?)
2. **(2 points)** Derive an analytical expression for the projection onto the convex set

$$C = \{x \in \mathbb{R}^d \,|\, x_i \geq 0\}, \quad \text{(positive orthant in } \mathbb{R}^d).$$

b. Instead of the interior point method of the lecture we use projected gradient descent which is a more simple method for constrained convex optimization. Let C be a convex, closed set and ϕ the differentiable, convex objective function, then the convex optimization problem $\min_{x \in C} \phi(x)$ can be solved via projected gradient descent which is defined as

$$x_{t+1} = P_C(x_t - \alpha_t \nabla \phi(x_t)),$$

where $\alpha_t > 0$ is the stepsize.

1. **(4 points)** Complete the functions needed for `Lasso` which has as arguments Y, Φ, λ and returns the weight vector w of the Lasso problem in Equation (1). Use projected gradient descent for the optimization problem (2).
2. **(2 points)** We use a real dataset, whose data is in `multidim_data_trainval.npy`, with 1300 training points (`Xtrain`, `Ytrain`), 200 test points (`Xval`, `Yval`). The task is to predict the total number of violent crimes per 100K population (output variable $Y_i \in \mathbb{R}$) from a set of features (input variables $X_i \in \mathbb{R}^{100}$) capturing all sorts of properties of the cities and their populations. Before you start do the following preprocessing - for each feature j you compute mean μ_j and standard deviation σ_j from the **training data** and you rescale **training and test data** as $X'_{ij} = \frac{X_{ij} - \mu_j}{\sigma_j}$. The idea behind is that we get rid of different scales of the variables.

   Run your Lasso implementation with linear design with offset, that is the model is $f(x) = \langle w, x \rangle + b$ (add a feature which is 1 for every data point, corresponding to the coefficient of the offset), for the training data with $\lambda = 10$.

   i. Report training and test loss of the obtained model.
   ii. Plot the predicted values by the computed model for the test set vs the true values. What do you observe?

3. **(1 point)** You are now free to use any set of basis functions. Write a function `Prediction(X)` which given a set of testpoints - a matrix $X \in \mathbb{R}^{n\times 100}$ outputs the predictions of your chosen learning method as a column vector $f \in \mathbb{R}^{n\times 1}$.
   For the best results on our hidden test set (which have to be better than the results of the method above) we have the following prizes
   i. **(10 Bonus Points)** for the winner
   ii. **(5 Bonus Points)** for the second best prediction
   iii. **(3 Bonus Points)** for the third best prediction

Hints:

- A sum $f + g$ of convex functions f, g is strictly convex if f or g is strictly convex.
- For the computation of the projection, note that each component can be minimized independently of the other ones. Why ?

Exercise 8 - Convex optimization

Lemma 1 *Let $X \in \mathbb{R}^{n\times d}$ and assume that X has rank n. Then the solution of*

$$\min_{w\in\mathbb{R}^d} \|w - w_0\|_2^2$$
$$\textit{such that: } Xw = b$$

has solution

$$w = w_0 - X^T(XX^T)^{-1}(Xw_0 - b).$$

- **(1 point)** Is the minimization problem above convex?
- **(4 points)** Provide a proof for the Lemma.

Hints:

- It might be useful to use the dual problem in the proof of the Lemma.