Optimization Problems 1

**Exercise 1: Regression**

(a) Show that ridge regression is a convex problem and compute its analytical solution (given the feature matrix $\mathbf{X} \in \mathbb{R}^{n\times d}$ and the target vector $\mathbf{y} \in \mathbb{R}^n$).

(b) When doing Bayesian regression we are interested in the posterior density $p_{\boldsymbol{\theta}|\,\mathbf{X},\mathbf{y}}(\boldsymbol{\theta}) \propto p_{\mathbf{y}|\,\mathbf{X},\boldsymbol{\theta}}(\mathbf{y})p_{\boldsymbol{\theta}}(\boldsymbol{\theta})$ where $p_{\mathbf{y}|\,\mathbf{X},\boldsymbol{\theta}}$ is the likelihood and $p_{\boldsymbol{\theta}}$ is the prior density. Assume the observations are i.i.d. with $y_i \sim \mathcal{N}(\mathbf{x}_i^\top\boldsymbol{\theta}, 1)$ and the parameters are also i.i.d. with $\boldsymbol{\theta}_j \sim \mathcal{N}(0, \sigma_w^2)$. Find the maximizer of the posterior density. What do you observe?

(c) Find the prior density that would result in Lasso regression in b).

(d) In the lecture you have learned that Ridge regression with regularization coefficient $\lambda$ can be equivalently stated as solving
$\min_{\boldsymbol{\theta}} \|(\mathbf{X}\theta - \mathbf{y})\|_2^2$ s.t. $\|\boldsymbol{\theta}\|_2 \leq t$.
This means we can associate with every $\lambda$ a $t$ and hence we can treat $t$ as a function of $\lambda$, i.e., $t : \mathbb{R}_{+,0} \to \mathbb{R}_{+,0}, \lambda \mapsto t(\lambda)$. Show that if $\lambda > 0$ and $\mathbf{X}^\top\mathbf{X}$ is non-singular then $\|\boldsymbol{\theta}^*_{\text{reg}}\|_2 = t(\lambda) < \|\boldsymbol{\theta}^*\|_2$ where $\boldsymbol{\theta}^*$ and $\boldsymbol{\theta}^*_{\text{reg}}$ are the minimzier of unregularized regression and the ridge regression, respectively.
*Hint 1*: For two non-singular matrices $\mathbf{A}, \mathbf{B}$ for which $\mathbf{A} + \mathbf{B}$ is invertible it holds that $(\mathbf{A} + \mathbf{B})^{-1} = \mathbf{A}^{-1} - \mathbf{A}^{-1}\mathbf{B}(\mathbf{A} + \mathbf{B})^{-1}$

**Exercise 2: Classification**

(a) In logistic regression, we model the conditional probability $\mathbb{P}(y = 1|\mathbf{x}^{(i)}) = \frac{1}{1+\exp(-\boldsymbol{\theta}^\top\mathbf{x}^{(i)})}$ of the target $y \in \{0, 1\}$ given a feature vector $\mathbf{x}^{(i)}$. From this it follows that $\mathbb{P}(y = y^{(i)}|\mathbf{x}^{(i)}) = \mathbb{P}(y = 1|\mathbf{x}^{(i)})^{y^{(i)}}(1 - \mathbb{P}(y = 1|\mathbf{x}^{(i)})^{1-y^{(i)}}$. With this derive the empirical risk $\mathcal{R}_{\text{emp}}$ as shown in the lecture following the maximum likelihood principle. (Assume the observations are independent)

(b) Show that $\mathcal{R}_{\text{emp}}$ of a) is convex.

(c) Show that the first primal form of the linear SVM with soft constraints
$\min_{\boldsymbol{\theta},\boldsymbol{\theta_0},\zeta^{(i)}} \frac{1}{2}\|\boldsymbol{\theta}\|_2^2 + C\sum_{i=1}^n \zeta^{(i)}$ s.t. $y^{(i)}\left(\boldsymbol{\theta}^\top\mathbf{x}^{(i)} + \boldsymbol{\theta}_0\right) \geq 1 - \zeta^{(i)} \quad \forall i \in \{1, \ldots, n\}$ and $\zeta^{(i)} \geq 0 \quad \forall i \in \{1, \ldots, n\}$ and its second primal form
$\min_{\boldsymbol{\theta},\boldsymbol{\theta_0}} \sum_{i=1}^n \max(1-y^{(i)}(\boldsymbol{\theta}^\top\mathbf{x}^{(i)}+\boldsymbol{\theta}_0), 0)+\lambda\|\boldsymbol{\theta}\|_2^2$ are equivalent. What is the functional relationship between $C$ and $\lambda$?
*Hint*: Try to insert the combined constraints into their associated objective.

(d) Show that the second primal form of the linear SVM is a convex problem