15.095: Machine Learning Under a Modern Optimization Lens **Fall 2021**

Homework 2: Due September 29

Hand in: **pdf** upload to Canvas. Please append any Julia code **at the end** of the whole pdf.

Note: this homework covers lectures four-six and recitation three. We recommend attempting questions after the relevant content has been covered in class.

2.1 Question 1: True/False (20 marks)

Please classify the following statements as true or false and justify your answer. If the statement is false, please provide a counter example. We will assign 2 marks for correctly classifying the answer, and 3 marks for the validity of the justification/counterexample.

(a) When using a big M formulation for a mixed integer optimization problem, the value of M will not influence the speed of the algorithm.

(b) When applying the cutting plane method to an optimization problem, it is required that we have a closed form expression like the one we have for linear regression below:

$$\min_{\boldsymbol{z} \in \{0,1\}^p : \boldsymbol{e}^\top \boldsymbol{z} \leq k} \quad f(\boldsymbol{z}),$$

where:

$$f(\boldsymbol{z}) := \frac{1}{2} \boldsymbol{y}^\top \left(\mathbb{I}_n + \gamma \sum_{j=1}^{p} z_j \boldsymbol{x}_j \boldsymbol{x}_j^\top \right)^{-1} \boldsymbol{y}$$

(c) Given problem data $\boldsymbol{X} \in \mathbb{R}^{n \times p}$, $\boldsymbol{y} \in \mathbb{R}^n$, the ℓ_2^2-regularized sparse linear regression problem can be defined as:

$$\min_{\boldsymbol{\beta} \in \mathbb{R}^p} \quad \frac{1}{2\gamma} \|\boldsymbol{\beta}\|_2^2 + \frac{1}{2} \|\boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{y}\|_2^2 + \lambda \|\boldsymbol{\beta}\|_0,$$

for $\lambda, \gamma > 0$. A valid reformulation of this problem is given by:

$$\min_{\boldsymbol{\beta} \in \mathbb{R}^p, \boldsymbol{\theta} \in \mathbb{R}_+^p, \boldsymbol{z} \in \{0,1\}^p} \quad \frac{1}{2\gamma} \sum_{i=1}^{p} \theta_i + \frac{1}{2} \|\boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{y}\|_2^2 + \lambda \sum_i z_i \text{ s.t. } \theta_i z_i \geq \beta_i^2, \ \forall i \in [p].$$

Hint: at optimality in the above formulation, θ_i will be as small as possible, i.e., $\theta_i = x_i^2 / z_i$ whenever $z_i > 0$. Consider the cases where $z_i = 0$ and $z_i = 1$ separately. What happens to x_i and θ_i. In particular, does the constraint $\theta_i z_i \geq x_i^2$ impose the logical constraint $x_i = 0$ if $z_i = 0$?

(d) The goal of cross-validation is to improve the performance of the model on training data.

2.2 Question 2: Regularized Sparse Regression Revisited (55 marks)

In this question, we extend the sparse linear regression method introduced in lecture 4 to the following ℓ_0-ℓ_1-ℓ_2^2 regularized problem

$$\min_{\boldsymbol{\beta}\in\mathbb{R}^p} \quad \frac{1}{2\gamma}\|\boldsymbol{\beta}\|_2^2+\frac{1}{2}\|\boldsymbol{X\beta}-\boldsymbol{y}\|_2^2+\lambda_0\|\boldsymbol{\beta}\|_0+\lambda_1\|\boldsymbol{\beta}\|_1, \tag{2.1}$$

where $\gamma, \lambda_0, \lambda_1 > 0$.

(a) (5 marks) Argue that we can separate out the optimization problem into the following "outer" and "inner" problems, similarly to the sparse regression formulation in lecture 4:

$$\min_{\boldsymbol{z}\in\{0,1\}^p} \quad f(\boldsymbol{z}) + \lambda_0 \boldsymbol{e}^\top \boldsymbol{z}, \tag{2.2}$$

$$f(\boldsymbol{z}) := \min_{\boldsymbol{\beta}\in\mathbb{R}^p} \quad \frac{1}{2\gamma}\|\boldsymbol{\beta}\|_2^2+\frac{1}{2}\|\boldsymbol{X\beta}-\boldsymbol{y}\|_2^2+\lambda_1\|\boldsymbol{\beta}\|_1 \text{ s.t. } \beta_i = 0 \text{ if } z_i = 0,\ \forall i \in [p]. \tag{2.3}$$

(b) (5 marks) Argue that the formulation for $f(\boldsymbol{z})$ can be rewritten as

$$f(\boldsymbol{z}) := \min_{\boldsymbol{\beta}\in\mathbb{R}^p} \quad \frac{1}{2\gamma}\sum_{i=1}^n \frac{\beta_i^2}{z_i} + \frac{1}{2}\|\boldsymbol{X\beta}-\boldsymbol{y}\|_2^2+\lambda_1\|\boldsymbol{\beta}\|_1, \tag{2.4}$$

where we take $0/0 = 0$ and $a/0 = +\infty$ for $a > 0$.

(c) (20 marks) Take the dual of Problem (2.4) with respect to $\boldsymbol{\beta}$, and thereby establish that we can write

$$f(\boldsymbol{z}) := \max_{\boldsymbol{\alpha},\boldsymbol{u}} \quad \frac{1}{2}\|\boldsymbol{y}\|_2^2-\frac{1}{2}\|\boldsymbol{y}-\boldsymbol{u}\|_2^2-\frac{\gamma}{2}\sum_{i=1}^p z_i\alpha_i^2 \text{ s.t. } \|\boldsymbol{X}^\top\boldsymbol{u}+\boldsymbol{\alpha}\|_\infty \le \lambda_1. \tag{2.5}$$

Hint #1: You may use without proof the fact that if we have a problem of the form

$$f(\boldsymbol{z}) := \min_{\boldsymbol{\beta}} \quad \frac{1}{2\gamma}\sum_{i=1}^n \frac{\beta_i^2}{z_i} + g(\boldsymbol{\beta}),$$

for a convex function g then we can write

$$f(\boldsymbol{z}) := \max_{\boldsymbol{\alpha}} \quad -\frac{\gamma}{2}\sum_{i=1}^n z_i\alpha_i^2 + h(\boldsymbol{\alpha}),$$

where $h(\boldsymbol{\alpha}) := \min_{\boldsymbol{v}} g(\boldsymbol{v}) - \boldsymbol{v}^\top\boldsymbol{\alpha}$.

Hint #2: You may use without proof the fact that for a Lasso optimization problem

$$\min_{\boldsymbol{\beta}} \quad \frac{1}{2}\|\boldsymbol{y}-\boldsymbol{X\beta}\|_2^2+\lambda_1\|\boldsymbol{\beta}\|_1$$

Strong duality holds with its dual problem, which is,

$$\max_{\boldsymbol{u}} \quad \frac{1}{2}\|\boldsymbol{y}\|_2^2-\frac{1}{2}\|\boldsymbol{y}-\boldsymbol{u}\|_2^2 \text{ s.t. } \|\boldsymbol{X}^\top\boldsymbol{u}\|_\infty \le \lambda_1.$$

(d) (10 marks) Using the dual form of $f(\boldsymbol{z})$ derived in part (c), discuss how the cutting-plane method derived in lecture 4 can be applied to the $\ell_0 - \ell_1 - \ell_2^2$ regularized problem. In your answer, be sure to state what the (sub)gradient of $f(\boldsymbol{z})$ is at a given point.

(e) (5 marks) In one sentence, name an efficient numerical strategy which you could use to solve the inner dual problem in order to evaluate $f(\boldsymbol{z})$ and its subgradients.

(f) (5 marks) In one paragraph, argue that the $\ell_0 - \ell_2^2$ sparse linear regression problem we saw in lecture 4 is not harder to solve in the presence of an ℓ_1 regularization term.

(g) (5 marks) Name one advantage and one disadvantage of the $\ell_0 - \ell_1 - \ell_2^2$ formulation derived here over the $\ell_0 - \ell_2^2$ formulation from lecture 4 (Hint: how many hyperparameters are there?).

2.3 Question 3: Regularized Sparse Regression (Computational) (25 marks)

Note: we recommend attempting this question after recitation 3 has been released.

(a) (15 marks) Using the code from recitation 3 as a building block, implement the cutting-plane approach you described in question 2(d) in Julia. You may either use the subproblem strategy you described in 2(e) or solve the inner maximization problem as a quadratic program using Gurobi. **Hint:** you should only need to edit the subproblem strategy, the master problem strategy should not be different than the case without ℓ_1 regularization.

(b) (10 marks) For $k \in \{1, 2, 3, 4\}$, test out this dual formulation on the given dataset "Train lpga2008_opt.csv" whose first column is the vector $\boldsymbol{y} \in \mathbb{R}^n$ and whose remaining columns form the matrix $\boldsymbol{X} \in \mathbb{R}^{n \times p}$. Set $\lambda_0 = 0$ and impose a sparsity constraint $\boldsymbol{e}^\top \boldsymbol{z} \le k$ on $\boldsymbol{z}$ in the master problem. Pick any "reasonable" value of λ_1, γ such that the output makes sense to you (e.g. $\gamma = 1$). What are the indices of the non-zero β_i's for each k?