MICAS901 - Introduction to Optimization: Homework 2

Please submit your solution before Nov 27th, 2022

Michèle Wigger, michele.wigger@telecom-paris.fr

All the steps and derivations are needed for full credit. The solution for each homework should be submitted individually by each student (one student per homework). Regarding the questions that need programming or coding, feel free to use any programing language you are comfortable with. The homework is worth $22,5\%$ of total course grade.

Part I: Closed-form solutions vs iterative algorithms

Consider the following linear regression problem:

$$\boldsymbol{w}^\star = \min_{\boldsymbol{w}\in\mathbb{R}^d} \frac{1}{N} \sum_{i\in[N]} \|\boldsymbol{w}^T \boldsymbol{x}_i - y_i\|_2^2 + \lambda \|\boldsymbol{w}\|_2^2 \tag{1}$$

where $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^N$ is the training set.

a) Derive a closed-form solution for $\boldsymbol{w}^\star$ in (1).
Derive an approximation of the computational complexity (in terms of $\mathcal{O}(-)$ analysis) for this solution.

b) Consider "Communities and Crime" dataset ($N = 1994$, $d = 128$) and find the optimal linear regressor from the closed-form expression.
What is the complexity in that case for the considered values of N, and d ?

c) Derive a simple gradient algorithm to solve (1), iteratively.
How do you select the 'best' stepsize ? argue your choice
Derive an approximation of the computational complexity (in terms of $\mathcal{O}(-)$ analysis) for this method. Compare it with that of the closed form solution.
Implement this algorithm using the "Individual household electric power consumption"
Plot the training error (using ℓ_2 loss function) vs the number of iterations, against the optimal closed-form solution.
Verify that the gradient descent eventually converges to the optimal solution. Argue the reason for that.

Part II: Deterministic vs Stochastic algorithms

Now consider logistic ridge regression

$$\min_{\boldsymbol{w}\in\mathbb{R}^d} f(\boldsymbol{w}) = \frac{1}{N}\sum_{i\in[N]} f_i(\boldsymbol{w}) + \lambda\|\boldsymbol{w}\|_2^2, \quad \text{where} \quad f_i(\boldsymbol{w}) = \ (\ln(2))^{-1}\log\left(1+\exp\{-y_i\boldsymbol{w}^T\boldsymbol{x}_i\}\right)$$

for the "Individual household electric power consumption" dataset ($N = 2075259$, $d = 9$). Please read the slides on logistic regression (Lect 3) before solving the problem. This is supervised learning problem, specifically, a classification problem, i.e., the training set is defined as: $\{\ (\boldsymbol{x}_i, y_i)\ \}_{i=1}^N$, where $\boldsymbol{x}_i \in \mathbb{R}^d$ is a feature vector, and $y_i \in \{-1,+1\}$ is a *binary label.* Solve the optimization problem using GD, and plain stochastic GD. For all the questions below, use $\lambda = 1$.

0-a) show that $f(\boldsymbol{w})$ is L-smooth and μ strongly convex.

0-b) derive the gradient, $\nabla f(\boldsymbol{w})$

a) Write/Derive the update equations for each algorithm, and implement them

b) Compare all these methods in terms complexity of hyper-parameter tuning, convergence rate, and computational complexity (per iteration of each algorithm).

c) Compare the training error, by plotting the training loss/error vs iterations, for all these algorithms. Comment on these results: are they expected or is there any discrepancy?

Part III: Nonlinear regression

Consider a non-linear regression (NLR), with a non-negative constraint:

$$(NLR) \qquad \underset{\boldsymbol{w}\in\mathbb{R}^d\ ,\ \boldsymbol{w}\geq\boldsymbol{0}}{\operatorname{argmin}}\ f(\boldsymbol{w}) = \frac{1}{N}\sum_{i\in[N]} f_i(\boldsymbol{w}) = \frac{1}{N}\sum_{i\in[N]} (\ y_i - \sigma(\boldsymbol{w})\ )^2$$

where $\sigma(\boldsymbol{w}) := (1+\exp^{-\boldsymbol{w}^T\boldsymbol{x}_i})^{-1}$ is the non-linear sigmoid activation function, and $(\ \boldsymbol{x}_i \in \mathbb{R}^d\ ,\ y_i \in \mathbb{R}\ )$ is the feature vector and label for sample $i \in [N]$.

1. Derive the gradient of $f(\boldsymbol{w})$, $\nabla f(\boldsymbol{w})$.
   hint: $f(\boldsymbol{w})$ is a composition of several functions. Then, use the chain rule for differentiation to compute $\nabla f(\boldsymbol{w})$

2. Solve NLR using the successive linear approximation (SLA) variant of successive approx methods.
   use the gradient (derived in prev question) to formulate the surrogate cost func (SLA upperbound) and constraint of the resulting sub-problem. Show all the steps that lead to the surrogate problem formulation.