STAT 535 Homework 5
Out November 8, 2022
Due November 15, 2022
©Marina Meilă
mmp@stat.washington.edu

Reminder: you are allowed and even encouraged to use results from previous homeworks, course notes, lectures withouth proof.

Problem 1 – Descent algorithms for training a neural network
This is **Problem 4** *from Homework 4. If you already submitted a solutions for this problem in Homework 4, you don't need to redo this problem.*

Problem 2 – Regularization is monotonic w.r.t. λ
Let $J_\lambda(w) = \hat{L}(w) + \frac{\lambda}{2}||w||^2$ be a regularized objective functions, where w are the parameters. For example, the linear ridge regression from Problem 2. Let $\lambda_1 > \lambda_2 > 0$ and denote $w_{1,2} = \text{argmin}_w J_{\lambda_{1,2}}$ the optimal solutions for λ_1, respectively λ_2, with $w_1 \neq w_2$, and assume further that $J_{\lambda_{1,2}}$ have *unique global minima.*

a. Prove that $||w_1|| < ||w_2||$ whenever $w_{1,2} \neq 0$.

b. Prove also that $\hat{L}(w_1) > \hat{L}(w_2)$.

In other words, imposing more regularization reduces the regularized quantity $||w||$, and increases the un-regularized one (i.e., the loss).

Problem 3 – Online linear regression by Stochastic gradient

Consider the linear regression problem with Least Square loss

$$\min_\beta E[(y - \beta^T x)^2] \;=\; \min_\beta L_{LS} \tag{1}$$

where $y \in \mathbb{R}$, $x \in \mathbb{R}^n$, $\beta \in \mathbb{R}^n$. For simplicity we consider the infinite sample version of the problem, but if you want a variation (ungraded) try also the finite sample version, where we optimize $\hat{L}_{LS}$ instead.

The function in (1) is a quadratic function that has a closed form solution, but we will pretend that we don't know this and investigate the use of (stochastic) gradient descent for this problem.

a. Find the expression of the gradient and Hessian of this problem, i.e $\nabla L_{LS}(\beta)$, $\nabla^2 L_{LS}(\beta)$. Express the Hessian as a function of some well known statistical descriptor(s) of the data distribution.

b. Assume that the covariates x are sampled from a Normal distribution with mean 0 and non-singular covariance Σ (known). Describe and motivate a reasonable way to find the λ parameter of the STOCHASTIC GRADIENT algorithm based on this assumption.

c. Write the expression of $d = \frac{\partial L_{LS}(y, \beta^T x)}{\partial \beta}$. Show that the direction of descent d is along x, i.e. $d = \alpha x$ for some scalar α, not necessarily positive. What does the scaling of x represent from a statistical modeling point of view?

e. Write the STOCHASTIC GRADIENT DESCENT algorithm to optimize this problem. Assume that λ is known.

For practice, ungraded Repeat the problem with an added regularization term $\frac{C}{2}||\beta||^2$.