# Theory Assignment 2

COMP 451 - Fundamentals of Machine Learning

Prof. William L. Hamilton

Winter 2021

**Preamble** The assignment is due March 11 at 11:59pm via MyCourses. Late work will be automatically subject to a 20% penalty, and can be submitted up to 5 days after the deadline. You may scan written answers or submit a typeset assignment, as long as you submit a single pdf file with clear indication of what question each answer refers to. You may consult with other students in the class regarding solution strategies, but you must list all the students that you consulted with on the first page of your submitted assignment. You may also consult published papers, textbooks, and other resources, but you must cite any source that you use in a non-trivial way (except the course notes). You must write the answer in your own words and be able to explain the solution to the professor, if asked.

## Question 1 [6 points]

Recall that the logistic function is defined as

$$\sigma(z) = \frac{1}{1 + e^{-z}}. \tag{1}$$

Prove that the following two identities for the logistic function hold:

$$\frac{\partial}{\partial z}\sigma(z) = \sigma(z)(1 - \sigma(z)) \qquad \textbf{(1) derivative identity}$$

$$1 - \sigma(z) = \sigma(-z) \qquad \textbf{(2) symmetry identity}$$

## Question 2 [8 points]

Suppose we have a training dataset $\mathcal{D}_{\text{trn}} = \{(\mathbf{x}_i, y_i, t_i), i = 0, ..., n-1\}$ for linear regression, where each training point $(\mathbf{x}_i, y_i) \in \mathbb{R}^m \times \mathbb{R}$ is also associated with a importance weight $t_i \in \mathbb{R}^+$. We use these importance weights to modify the empirical risk minimization as follows:

$$\mathbf{w}^* = \arg\min_{\mathbf{w}} \sum_{(\mathbf{x}_i, y_i, t_i) \in \mathcal{D}_{\text{trn}}} t_i (\mathbf{w}^\top \mathbf{x}_i - y_i)^2. \tag{2}$$

In other words, we use $t_i$ to weigh the contribution of training point $i$ to the overall empirical risk.

- **[5 points]** First, derive the closed-form solution for the optimal $\mathbf{w}^*$ with these importance weights. *Hint: use* $\mathbf{T}$ *to denote a diagonal matrix that has the weights* $t_i, i = 0, .., n-1$ *along the diagonal.*
- **[3 points]** Next, mathematically justify how these importance weights can be motivated from a maximum likelihood perspective. *Hint: Consider a maximum likelihood interpretation of linear regression (as in Chapter 9 of the notes), but allow the variance* $\sigma$ *of the noise* $\epsilon \sim \mathcal{N}(0, \sigma)$ *to vary across training points, i.e., consider having a noise variance* $\sigma_i$ *that is specific to each training point.*

## Question 3 [4 points]

Suppose we are working on a binary classification problem with real-valued features, i.e., $\mathbf{x} \in \mathbb{R}^m$ and $y \in \{0, 1\}$. Consider the following empirical risk minimization function:

$$R(\mathbf{w}) = \sum_{(\mathbf{x},y)\in\mathcal{D}_{\text{trn}}} -\log\left(\sigma(\mathbf{w}^\top\mathbf{x})^y(1-\sigma(\mathbf{w}^\top\mathbf{x}))^{(1-y)}\right) + \|\mathbf{w}^\top\mathbf{x} - C\|^2, \tag{3}$$

where $C \in \mathbb{R}$ is a constant. Is this empirical risk minimization convex? Justify your answer. *Hint: you do not necessarily need to compute any Hessians! Consider how the function in Equation 3 relates to known loss functions from class.*

## Question 4 [short answers; 2 points each]

*Answer each question with 1-3 sentences for justification, potentially with equations/examples for support.*

**a)** True or false: Gradient descent for linear regression always returns a unique solution.

**b)** True or false: $k$-NNs will tend to overfit more as $k$ increases.

**c)** Suppose you have higher validation error than training error, name three things you could do to address this issue.