STAT3007/7007 Deep Learning, Tutorial 9
2022 Semester 2

1. (Input transformation) Consider training a parametric model $f_{\mathbf{w}}(\mathbf{x})$ by minimizing its MSE. Let $f_{\mathbf{w}^*}(\mathbf{x})$ be the learned model on a training set $(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n) \in \mathbf{R}^{d+1} \times \mathbf{R}$ (the extra 1 in $d+1$ means we have a dummy feature in $\mathbf{x}$), and $f_{\mathbf{w}'}(\mathbf{x}')$ be the learned model on the training set $(\mathbf{x}'_1, y_1), \ldots, (\mathbf{x}'_n, y_n)$ obtained by normalizing the d non-dummy and non-constant features to have zero mean and unit variance. We assume that both $\mathbf{w}^*$ and $\mathbf{w}'^*$ are unique global minimizers in this question.

- **(a)** If $f_{\mathbf{w}}(\mathbf{x}) = \mathbf{w}^\top \mathbf{x}$, can you express $\mathbf{w}^*$ in terms of $\mathbf{w}'^*$? Justify your answer.
- **(b)** If $f_{\mathbf{w}}(\mathbf{x}) = \sigma(\mathbf{w}^\top \mathbf{x})$, can you express $\mathbf{w}^*$ in terms of $\mathbf{w}'^*$? Justify your answer.

2. (Adaptive learning rate) Consider minimizing the function $f(x, y) = ax^2 + by^2$, where $a, b > 0$, using gradient-based method, starting from an initial solution $(x_0, y_0) \in \mathbf{R}^2$.

- **(a)** Let (x_t, y_t) be the solution obtained by applying gradient descent with a learning rate of η on this function. For what values of the learning rate does gradient descent converge to the minimizer for an arbitrary initial solution? Justify your answer.
- **(b)** Does Newton's method converge to the minimizer? Justify your answer.
- **(c)** Assume that $x_0, y_0 > 0$, and $\eta = \min(x_0, y_0)$ Show that AdaGrad converges to the minimizer $(0, 0)$. Compare the convergence rates for AdaGrad and gradient descent.