ITCS 6101/8101 Homework 2: Theory (60 points)

February 3, 2026

1 Differentiation exercises

1.1 Chain rule

1. (5p) Compute the derivative of $h(x) = (3x + 2)^2 + 2$ by writing $h(x) = f(g(x))$ where $g(x) = 3x + 2$ and $f(g) = g^2 + 2$.
   - (5p) Implement the function $h(x)$ in PyTorch and use the automatic differentiation capability to compute and report $h'(-1)$.
2. (5p) Compute the derivative of $h(x) = (2x + 1)^2 \ln(x + 1)$ by writing $h(x) = f(g_1(x), g_2(x))$ where $g_1(x) = 2x + 1$, $g_2(x) = \ln(x + 1)$ and $f(g) = g_1^2 g_2$.
   - (5p) Implement the function $h(x)$ in PyTorch and use the automatic differentiation capability to compute and report $h'(+1)$.
3. (5p) Show that the derivative of the sigmoid function can be written as $\sigma'(x) = \sigma(x)(1 - \sigma(x))$
   - (5p) Implement the sigmoid in PyTorch and use the automatic differentiation capability to compute and report $\sigma'(+5)$.

1.2 Gradient based optimization

By setting the gradient to 0, find the solutions to the following optimization problems:

1. (5p) $\hat{x} = \arg\min_x J(x)$, where $J(x) = x^2 - 2x + 3$.
2. (5p) $\hat{x}, \hat{y} = \arg\min_{x,y} J(x, y)$, where $J(x, y) = 2x^2 + 3y^2 - 4x + 12y + 15$.
3. (Bonus 5p) $\hat{x}, \hat{y} = \arg\min_{x,y} J(x, y)$, where $J(x, y) = x^2 + 4y^2 - 4xy + 2x - 4y + 4$.

(Bonus 10p) Prove that all objective functions above are convex.

2 Skip-gram model: Loss and gradients

The Skip-gram model computes for every target token the Noise Contrastive Estimation (NCE) loss below:

$$L(\mathbf{w}) = -\left[\log \sigma(\mathbf{w}^T \mathbf{c}_{pos}) + \sum_{i=1}^{k} \log \sigma(-\mathbf{w}^T \mathbf{c}_{neg_i})\right] \tag{1}$$

1. (10p) Show that the gradient of the loss with respect to the target embedding $\mathbf{w}$ is:

$$\frac{\partial L}{\partial \mathbf{w}} = \left(\sigma(\mathbf{w}^T \mathbf{c}_{pos}) - 1\right) \mathbf{c}_{pos} + \sum_{i=1}^{k} \sigma(\mathbf{w}^T \mathbf{c}_{neg_i}) \mathbf{c}_{neg_i} \tag{2}$$

2. (10p) Assume that we train a Skip-gram model to use just one embedding matrix W for both target and context embeddings.
   (a) Rewrite the loss and the gradient above to indicate that the context embeddings $\mathbf{c}$ are the same as the target embeddings $\mathbf{w}$.
   (b) Show an example where the gradient above would not be correct: show the sentence, the target word, the context, and explain why the gradient formula would be wrong for that example.

3 Submission

Submit your theory responses on Canvas as one file named `hw02-theory.pdf` and the PyTorch solutions as one file named `hw02-pytorch.py`. It is important that you show clearly all the derivation steps. We recommend using an editor such as Overleaf for Latex, Word, or Jupyter-Notebook that allows proper formatting of equations. Alternatively, if you choose to write your solutions on paper, submit an electronic scan / photo of it exported to **PDF**. Make sure that your writing is **legible** and the scan has good quality (we will not grade solutions that we struggle to read).