HOMEWORK 7. GRADIENT CLIPPING

Kun Yuan

November 23, 2023

Attention: Turn in your homework at the beginning of our lecture on Nov. 28, 2023

1 Backpropagation in RNN

Consider the following RNN formulation

$$h_t = w_x x_t + w_h h_{t-1} \tag{1}$$

$$\hat{y}_t = w_o h_t \tag{2}$$

for $t = 1, \cdots, T$, where $w_x \in \mathbb{R}, w_h \in \mathbb{R}$, and $w_o \in \mathbb{R}$ are parameters to learn, $x_t \in \mathbb{R}$ is the input data at iteration t, $h_t \in \mathbb{R}$ is the hidden state at iteration t with initialization h_0, and $\hat{y}_t \in \mathbb{R}$ is the output at iteration t. Given the samples $\{x_t, y_t\}_{t=1}^T$, we consider the following loss function

$$L(w_x, w_h, w_o; \{x_t, y_t\}_{t=1}^T) = \frac{1}{2T} \sum_{t=1}^T (\hat{y}_t - y_t)^2 \tag{3}$$

Please derive $\frac{\partial L}{\partial w_x}$ and $\frac{\partial L}{\partial w_o}$.

2 (L_0, L_1)-smooth condition

Prove the following statement: Let f be the univariate polynomial $f(x) = \sum_{i=1}^d a_i x^i$. When $d \geq 3$, then $f(x)$ is (L_0, L_1)-smooth for some L_0 and L_1 but not L-smooth.

Hint: Since $f(x)$ is twice differentiable, the (L_0, L_1)-smooth condition can be simplified as $\|\nabla^2 f(x)\| \leq L_0 + L_1 \|\nabla f(x)\|$.