Due Date : March 27th 23 :00, 2023

Instructions

- *For all questions, show your work !*
- *Use a document preparation system such as LaTeX.*
- *Submit your answers electronically via the course gradescope*
- *TA for this assignment are :* ***Dinghuai Zhang*** *(IFT6135B) and* ***Ghait Boukachab*** *(IFT6135A).*

Question 1 (8-9-8)**.** This question is about activation functions and vanishing/exploding gradients in recurrent neural networks (RNNs). Let $\sigma : \mathbb{R} \to \mathbb{R}$ be an activation function. When the argument is a vector, we apply σ element-wise. Consider the following recurrent unit :

$$\boldsymbol{h}_t = \boldsymbol{W}\sigma(\boldsymbol{h}_{t-1}) + \boldsymbol{U}\boldsymbol{x}_t + \boldsymbol{b}$$

1.1 Show that applying the activation function in this way results in an equivalent recurrence as the conventional way of applying the activation function : $\boldsymbol{g}_t = \sigma(\boldsymbol{W}\boldsymbol{g}_{t-1} + \boldsymbol{U}\boldsymbol{x}_t + \boldsymbol{b})$ (i.e. express $\boldsymbol{g}_t$ in terms of $\boldsymbol{h}_t$). More formally, you need to prove it using mathematical induction. You only need to prove the induction step in this question, assuming your expression holds for time step $t-1$.

1.2 Following the previous question and analyze property of $\boldsymbol{g}_t$. Let $||\boldsymbol{A}||$ denote the L_2 operator norm [1] of matrix $\boldsymbol{A}$ ($||\boldsymbol{A}|| := \max_{\boldsymbol{x}:||\boldsymbol{x}||=1} ||\boldsymbol{A}\boldsymbol{x}||$). Assume $\sigma(x)$ has bounded derivative, i.e. $|\sigma'(x)| \leq \gamma$ for some $\gamma > 0$ and for all x. We denote as $\lambda_1(\cdot)$ the largest eigenvalue of a symmetric matrix. Show that if the largest eigenvalue of the weights is bounded by $\frac{\delta^2}{\gamma^2}$ for some $0 \leq \delta < 1$, gradients of the hidden state will vanish over time, i.e.

$$\lambda_1(\boldsymbol{W}^\top \boldsymbol{W}) \leq \frac{\delta^2}{\gamma^2} \quad \Longrightarrow \quad \left|\left|\frac{\partial \boldsymbol{g}_T}{\partial \boldsymbol{g}_0}\right|\right| \to 0 \text{ as } T \to \infty$$

Use the following properties of the L_2 operator norm

$$||\boldsymbol{A}\boldsymbol{B}|| \leq ||\boldsymbol{A}||\,||\boldsymbol{B}|| \qquad \text{and} \qquad ||\boldsymbol{A}|| = \sqrt{\lambda_1(\boldsymbol{A}^\top \boldsymbol{A})}$$

1.3 What do you think will happen to the gradients of the hidden state if the condition in the previous question is reversed, i.e. if the largest eigenvalue of the weights is larger than $\frac{\delta^2}{\gamma^2}$? Is this condition *necessary* and/or *sufficient* for the gradient to explode ? (Answer in 1-2 sentences).

Question 2 (8-8-9)**.** In this question you will demonstrate that an estimate of the first moment of the gradient using an (exponential) running average is equivalent to using momentum, and is biased by a scaling factor. The goal of this question is for you to consider the relationship between different optimization schemes, and to practice noting and quantifying the effect (particularly in terms of bias/variance) of *estimating* a quantity.

Let $\boldsymbol{g}_t$ be an unbiased sample of gradient at time step t and $\Delta\boldsymbol{\theta}_t$ be the update to be made. Initialize $\boldsymbol{v}_0$ to be a vector of zeros.

2.1 For $t \geq 1$, consider the following update rules :

1. The L_2 operator norm of a matrix $\boldsymbol{A}$ is is an *induced norm* corresponding to the L_2 norm of vectors. You can try to prove the given properties as an exercise.

- SGD with momentum :

$$\boldsymbol{v}_t = \alpha \boldsymbol{v}_{t-1} + \epsilon \boldsymbol{g}_t \qquad \Delta\boldsymbol{\theta}_t = -\boldsymbol{v}_t$$

  where $\epsilon > 0$ and $\alpha \in (0, 1)$.

- SGD with running average of $\boldsymbol{g}_t$:

$$\boldsymbol{v}_t = \beta \boldsymbol{v}_{t-1} + (1-\beta)\boldsymbol{g}_t \qquad \Delta\boldsymbol{\theta}_t = -\delta \boldsymbol{v}_t$$

  where $\beta \in (0, 1)$ and $\delta > 0$.

Express the two update rules recursively ($\Delta\boldsymbol{\theta}_t$ as a function of $\Delta\boldsymbol{\theta}_{t-1}$). Show that these two update rules are equivalent ; i.e. express (α, ϵ) as a function of (β, δ).

2.2 Unroll the running average update rule, i.e. express $\boldsymbol{v}_t$ as a linear combination of $\boldsymbol{g}_i$'s $(1 \leq i \leq t)$.

2.3 Assume $\boldsymbol{g}_t$ has a stationary distribution independent of t. Derive an estimator of $\mathbb{E}[\boldsymbol{g}_t]$ using $\mathbb{E}[\boldsymbol{v}_t]$.

Question 3 (8-8-9)**.** The following equation is the second order Taylor expansion of a function f at the point x_0 :

$$\hat{f}_{x_0}(x) = f(x_0) + (x - x_0)^T g + \frac{1}{2}(x - x_0)^T H (x - x_0), \tag{1}$$

where $g = \frac{\partial f}{\partial x}(x_0)$ and $H = \frac{\partial^2 f}{\partial^2 x}(x_0)$. Here $f(x) \in \mathbb{R}$, $x, x_0, g \in \mathbb{R}^n, H \in \mathbb{R}^{n \times n}$.

3.1 Say we start from x_0 and perform one-step gradient descent with the gradient g and learning rate equal to ϵ, what is the value of $\hat{f}_{x_0}(\cdot)$ after the update ?

3.2 Analyze the obtained terms from the previous question. Explain under what conditions of ϵ would gradient descent work (i.e., reduce the value of target function $\hat{f}_{x_0}(\cdot)$).

3.3 Setting the gradient of $\hat{f}_{x_0}(\cdot)$ to zero. Could you derive a new optimization algorithm based on this ?

Question 4 (8-9-8)**.** Weight Normalization (WN) is a reparameterization technique inspired by Batch normalization (BN), aiming at improving the conditioning of the gradient. Instead of having a regular weight parameter (denoted as w) for each neuron, i.e. $y = \sigma(w^T x + b)$, we decouple the weight vector into two terms :

$$w = \frac{g}{\|u\|} u \tag{2}$$

where $g \in \mathbf{R}$ is a scaling factor and u is normalized by the Euclidean $\|u\|$. Doing so has similar effects as implementing BN, but has a lower computational overhead.

4.1 Consider the simplest model, where we only have one single output layer conditioned on one input feature x. Assume x is whitened to be independently distributed with zero mean and unit variance. A standard BN operation is defined by Equation (8.35) of the deep learning book. Show that in this simple case WN is equivalent to BN (ignoring the learned scale and shift terms for both BN and WN) that normalizes the linearly transformed feature $w^T x + b$.

4.2 Show that the gradient of a loss function L with respect to the new parameters u can be expressed in the form $sW^* \cdot \nabla_w L$, where s is a scalar and W^* is the orthogonal complement projection matrix. As a side note : W^* projects the gradient away from the direction of w, which is usually (empirically) close to a dominant eigenvector of the covariance of the gradient. This helps to condition the landscape over which we want to optimize.

4.3 Researchers have found that for weight normalization, the norm of parameters ($\|u\|$) keeps increasing. Show that $\|u\|$ becomes equal or larger after one gradient update step with the Pythogorean theorem and the gradient update equation from the previous question.