ECS289 Deep Learning Homework

Fangzhou Li

November 22, 2020

1 Vanishing Gradients

1.1 The Gradient of the Overall Loss

$$
\begin{aligned}
\frac{\partial \epsilon}{\partial \theta} &= \frac{\partial}{\partial \theta}(\sum_{t=1}^{T} \epsilon_t) \\
&= \sum_{t=1}^{T} \frac{\partial \epsilon_t}{\partial \theta}
\end{aligned} \tag{1}
$$

1.2 Chain Rule 1

$$
\begin{aligned}
\frac{\partial \epsilon}{\partial \theta} &= \sum_{t=1}^{T} \frac{\partial \epsilon_t}{\partial \theta} \\
&= \sum_{t=1}^{T} \frac{\partial \epsilon_t}{\partial \mathbf{x}_k} \frac{\partial \mathbf{x}_k}{\partial \theta}
\end{aligned} \tag{2}
$$

1.3 Chain Rule 2

$$
\begin{aligned}
\frac{\partial \mathbf{x}_t}{\partial \mathbf{x}_k} &= \frac{\partial \mathbf{x}_t}{\partial \mathbf{W}_{rec}} \frac{\partial \mathbf{W}_{rec}}{\partial \mathbf{x}_k} \\
&= \frac{\partial \mathbf{x}_t}{\partial \mathbf{W}_{rec}} \frac{\partial \mathbf{W}_{rec}}{\partial \mathbf{x}_{t-1}} \frac{\partial \mathbf{x}_{t-1}}{\partial \mathbf{x}_k} \\
&= \prod_{i=0}^{t-k-1} \frac{\partial \mathbf{x}_{t-i}}{\partial \mathbf{W}_{rec}} \frac{\partial \mathbf{W}_{rec}}{\partial \mathbf{x}_{t-i-1}}
\end{aligned} \tag{3}
$$

1.4 Eigenvalues and Gradient Vanishing

Let's observe how a hidden state over time steps influence the loss,

$$
\begin{aligned}
\frac{\partial \epsilon_t}{\partial \mathbf{x}_k} &= \frac{\partial \epsilon_t}{\partial \mathbf{x}_t} \frac{\partial \mathbf{x}_t}{\partial \mathbf{x}_k} \\
&= \frac{\partial \epsilon_t}{\partial \mathbf{x}_t} \prod_{i=0}^{t-k-1} \frac{\partial \mathbf{x}_{t-i}}{\partial \mathbf{x}_{t-i-1}}
\end{aligned} \tag{4}
$$

The term $\frac{\partial \mathbf{x}_{t-i}}{\partial \mathbf{x}_{t-i-1}}$ can be calculated from hidden state equation,

$$
\frac{\partial \mathbf{x}_{t-i}}{\partial \mathbf{x}_{t-i-1}} = \mathbf{W}_{rec} \sigma^{'}(\mathbf{x}_{t-i-1}) \leq \mathbf{W}_{rec} \gamma, \tag{5}
$$

Take (5) into (4), we have,

$$
\begin{aligned}
\frac{\partial \epsilon_t}{\partial \mathbf{x}_k} &= \frac{\partial \epsilon_t}{\partial \mathbf{x}_t} \prod_{i=0}^{t-k-1} \mathbf{W}_{rec} \sigma^{'}(\mathbf{x}_{t-i-1}) \\
&\leq \frac{\partial \epsilon_t}{\partial \mathbf{x}_t} \prod_{i=0}^{t-k-1} \mathbf{W}_{rec} \gamma
\end{aligned}
\tag{6}
$$

By eigendecomposition, $W_{rec}^m = Q\Lambda^m Q^{-1}$, where Q is an orthonormal matrix and Λ is a diagonal matrix with each non-zero element being an eigenvalue of W_{rec}. Let $\mathbf{q}_{max}$ be the row vector of Q corresponding to λ_{max}, With $m = t - k \rightarrow \infty$, then the product in (6) results in,

$$
\begin{aligned}
\mathbf{q}_{max} \lambda_{max}^m \mathbf{q}_{max} \gamma^m &= \mathbf{q}_{max} (\lambda_{max} \gamma)^m \mathbf{q}_{max} \\
&\rightarrow 0
\end{aligned}
\tag{7}
$$

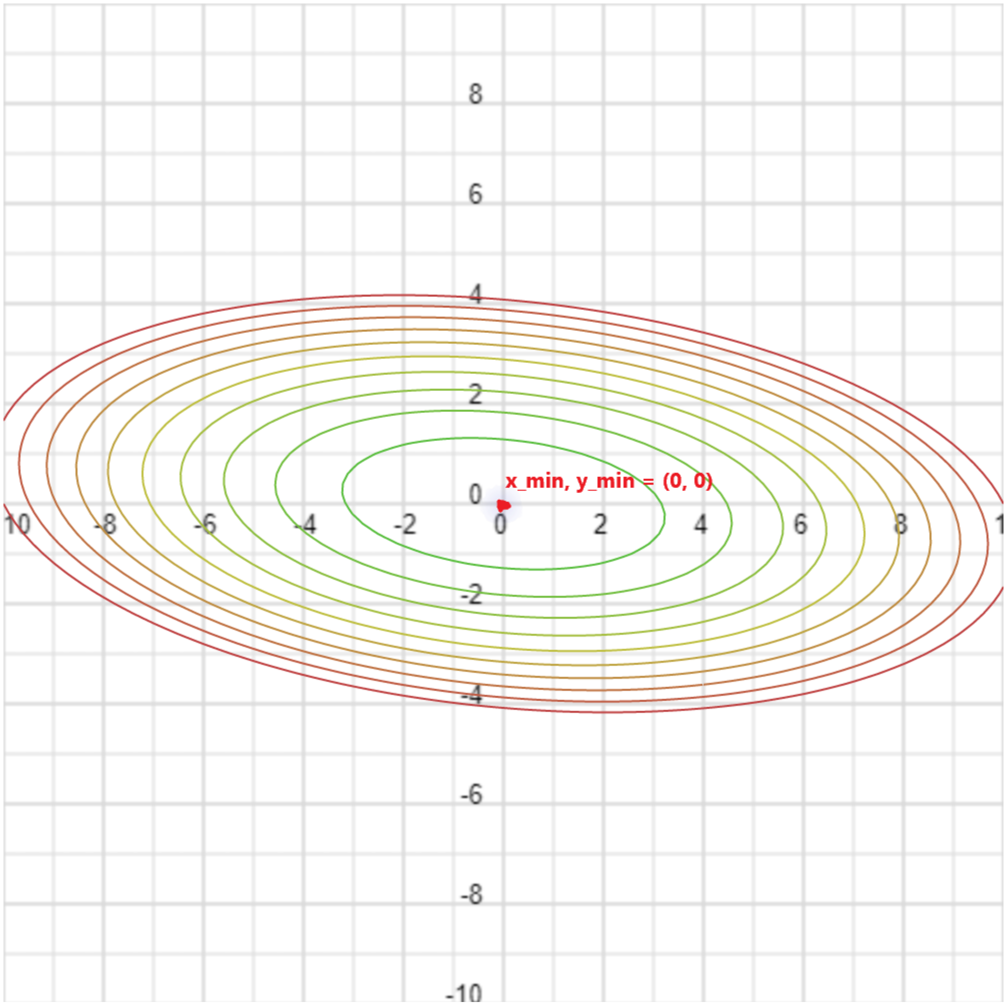
since $\lambda_{max} < \frac{1}{\gamma}$. This cause the loss related to $\mathbf{x}_k$ be trivial.

1.5 Eigenvalues and Gradient Vanishing 2

The vanishing gradient requires the largest eigenvalue of $\mathbf{W}_{rec}$ to be smaller than $\frac{1}{\gamma}$. The exploding gradient requires it to be larger than $\frac{1}{\gamma}$.

2 Optimizers

2.1 Contour Plot



2.2 Weight Calculation

$$
\begin{aligned}
\frac{\partial f}{\partial x} &= 2x + y \\
\frac{\partial f}{\partial y} &= 12y + x \\
\nabla f &= [2x + y, x + 12y] \\
x_0 &= 4 \\
y_0 &= 1 \\
\alpha &= 0.05 \\
\gamma &= 0.9
\end{aligned} \tag{8}
$$

2.2.1 Stochastic Gradient Descent

$$
\begin{aligned}
x_1 &= x_0 - \alpha \frac{\partial f(x_0, y_0)}{\partial x} \\
&= 3.55 \\
y_1 &= y_0 - \alpha \frac{\partial f(x_0, y_0)}{\partial y} \\
&= 0.2 \\
x_2 &= x_1 - \alpha \frac{\partial f(x_1, y_1)}{\partial x} \\
&= 3.19 \\
y_2 &= y_1 - \alpha \frac{\partial f(x_1, y_1)}{\partial y} \\
&= -0.1 \\
x_3 &= x_2 - \alpha \frac{\partial f(x_2, y_2)}{\partial x} \\
&= 2.88 \\
y_3 &= y_2 - \alpha \frac{\partial f(x_2, y_2)}{\partial y} \\
&= -0.2
\end{aligned}
\tag{9}
$$

2.2.2 SGD with Nesterov's

$$
\begin{aligned}
v_0 &= [0, 0] \\
v_1 &= \gamma v_0 - \alpha \nabla f([x_0, y_0] + \gamma v_0) \\
&= [-0.45, -0.8] \\
[x_1, y_1] &= [x_0, y_0] + v_1 \\
&= [3.55, 0.2] \\
v_2 &= \gamma v_1 - \alpha \nabla f([x_1, y_1] + \gamma v_1) \\
&= [-0.6935, -0.56525] \\
[x_2, y_2] &= [x_1, y_1] + v_2 \\
&= [2.8565, -0.36525] \\
v_3 &= \gamma v_2 - \alpha \nabla f([x_2, y_2] + \gamma v_2) \\
&= [-0.80368625, -0.0959575] \\
[x_3, y_3] &= [x_2, y_2] + v_3 \\
&= [2.05281375, -0.4612075]
\end{aligned}
\tag{10}
$$