EECS 182 Deep Neural Networks
Fall 2022 Anant Sahai

Homework 3

This homework is due on Saturday, September 24, 2022, at 10:59PM.

1. Accelerating Gradient Descent with momentum

Consider the problem finding the minimizer of the following objective:

$$\mathcal{L}(w) = ||y - Xw||_2^2 \tag{1}$$

In the previous homework, we proved that gradient descent (GD) algorithm can converge and derive the convergence rate. In this homework, we will add the momentum term and how it affects to the convergence rate. The optimization procedure of gradient descent+momentum is given below:

$$\begin{aligned} w_{t+1} &= w_t - \eta z_{t+1} \\ z_{t+1} &= (1-\beta) z_t + \beta g_t, \end{aligned} \tag{2}$$

where $g_t = \nabla \mathcal{L}(w_t)$, η is learning rate and β defines how much averaging we want for the gradient. Note that when $\beta = 1$, the above procedure is just original gradient descent.

Let's investigate the effect of this change. We'll see that this modification can actually 'accelerate' the convergence by allowing larger learning rates.

(a) Recall that the gradient descent update of 1 is

$$w_{t+1} = \left(I - 2\eta(X^T X)\right) w_t + 2\eta X^T y \tag{3}$$

and the minimizer is

$$w^* = (X^T X)^{-1} X^T y \tag{4}$$

The geometric convergence rate (in the sense of what base is there for convergence as rate^t) of this procedure is

$$\text{rate} = \max_i |1 - 2\eta\sigma_i^2| \tag{5}$$

You saw on the last homework that if we choose the learning rate that maximizes Eq. 5, the optimal learning rate, η^* is

$$\eta^* = \frac{1}{\sigma_{\text{min}}^2 + \sigma_{\text{max}}^2}, \tag{6}$$

where σ_{max} and σ_{min} are the maximum and minimum singular value of the matrix X. The correspond-

ing optimal convergence rate is

$$\text{optimal rate} = \frac{(\sigma_{\max}/\sigma_{\min})^2 - 1}{(\sigma_{\max}/\sigma_{\min})^2 + 1} \tag{7}$$

Therefore, how fast ordinary gradient descent converges is determined by the ratio between the maximum singular value and the minimum singular value as above.

Now, let's consider using momentum to smooth the gradients before taking a step in Eq.2.

$$\begin{aligned} w_{t+1} &= w_t - \eta z_{t+1} \\ z_{t+1} &= (1-\beta) z_t + \beta(2X^T X w_t - 2X^T y) \end{aligned} \tag{8}$$

We can use the SVD of the matrix $X = U\Sigma V^T$, where $\Sigma = \text{diag}(\sigma_{\max}, \sigma_2, \ldots, \sigma_{\min})$ with the same (potentially rectangular) shape as X. This allows us to reparameterize the parameters w_t and averaged gradients z_t as below:

$$\begin{aligned} x_t &= V^T(w_t - w^*) \\ a_t &= V^T z_t. \end{aligned} \tag{9}$$

Please rewrite Eq. 8 with the reparameterized variables, $x_t[i]$ and $a_t[i]$. ($x_t[i]$ and $a_t[i]$ are i-th components of x_t and a_t respectively.)

(b) Notice that the above 2×2 vector/matrix recurrence has no external input. We can derive the 2×2 system matrix R_i from above such that

$$\begin{bmatrix} a_{t+1}[i] \\ x_{t+1}[i] \end{bmatrix} = R_i \begin{bmatrix} a_t[i] \\ x_t[i] \end{bmatrix} \tag{10}$$

Derive R_i.

(c) **Use the computer to symbolically find the eigenvalues of the matrix R_i.**
When are they purely real? When are they repeated and purely real? When are they complex?

(d) **For the case when they are repeated, what is the condition on η, β, σ_i that keeps them stable (strictly inside the unit circle)? What is the highest learning rate η as a function of β and σ_i that results in repeated eigenvalues?**

(e) **For the case when the eigenvalues are real, what is the condition on η, β, σ_i that keeps them stable (strictly inside the unit circle)? What is the upper bound of learning rate? Express with β, σ_i**

(f) **For the case when the eigenvalues are complex, what is the condition on η, β, σ_i that keeps them stable (strictly inside the unit circle)? What is the highest learning rate η as a function of β and σ_i that results in complex eigenvalues?**

(g) Now, apply what you have learned to the following problem. Assume that $\beta = 0.1$ and we have a problem with two singular values $\sigma^2_{\max} = 5$ and $\sigma^2_{\min} = 0.05$. **What learning rate η should we choose to get the fastest convergence for gradient descent with momentum? Compare how many iterations it will take to get within 99.9% of the optimal solution (starting at 0) using this learning rate and momentum with what it would take using ordinary gradient descent.**

2. Understanding Convolution as Finite Impulse Response Filter

For the discrete time signal, the output of linear time invariant system is defined as:

$$y[n] = x[n] * h[n] = \sum_{i=-\infty}^{\infty} x[n-i] \cdot h[i] = \sum_{i=-\infty}^{\infty} x[i] \cdot h[n-i] \tag{11}$$

, where x is the input signal, h is impulse response (also referred to as the filter). Please note that the convolution operations is to 'flip and drag'. But for neural networks, we simply implement the convolutional layer without flipping and such operation is called correlation. Interestingly, in CNN those two operations are equivalent because filter weights are initialized and updated. Even though you implement 'true' convolution, you just ended up with getting the flipped kernel. In this question, we will follow the definition.

Now let's consider rectangular signal with the length of L (sometimes also called the "rect" for short, or, alternatively, the "boxcar" signal). This signal is defined as:

$$x(n) = \begin{cases} 1 & n = 0, 1, 2, ..., L-1 \\ 0 & \text{otherwise} \end{cases}$$

Here's an example plot for $L = 7$, with time indices shown from -2 to 8 (so some implicit zeros are shown):

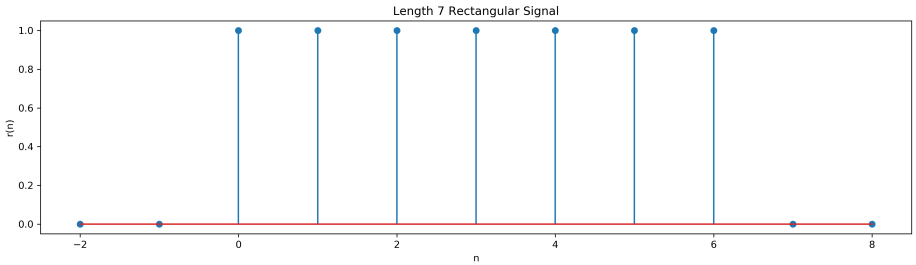

Figure 1: The rectangular signal with the length of 7

(a) The impulse response is define as:

$$h(n) = (\frac{1}{2})^n u(n) = \begin{cases} (\frac{1}{2})^n & n = 0, 1, 2, ... \\ 0 & \text{otherwise} \end{cases}$$

. **Compute and plot the convolution of $x(n)$ and $h(n)$.**

(b) Now let's shift $x(n)$ by N, i.e. $x_2(n) = x(n-N)$. Let's put $N = 5$ **Then, compute** $y_2(n) = h(n) * x_2(n)$**. Which property of the convolution can you find?**

Now, let's extend 1D to 2D. The example of 2D signal is the image. The operation of 2D convolution is defined as follows:

$$y[m,n] = x[m,n] * h[m,n] = \sum_{i,j=-\infty}^{\infty} x[m-i, n-j] \cdot h[i,j] = \sum_{i,j=-\infty}^{\infty} x[i,j] \cdot h[m-i, n-j] \tag{12}$$

, where x is input signal, h is FIR filter and y is the output signal.

(c) 2D matrices, x and h are given like below:

$$x = \begin{bmatrix} 1 & 2 & 3 & 4 & 5 \\ 6 & 7 & 8 & 9 & 10 \\ 11 & 12 & 13 & 14 & 15 \\ 16 & 17 & 18 & 19 & 20 \\ 21 & 22 & 23 & 24 & 25 \end{bmatrix} \tag{13}$$

$$h = \begin{bmatrix} -1 & -2 & -1 \\ 0 & 0 & 0 \\ 1 & 2 & 1 \end{bmatrix} \tag{14}$$

Then, evaluate y. Assume that there is no pad and stride is 1.

(d) Now let's consider striding and padding. Evaluate y for following cases:

i. stride, pad = 1, 1

ii. stride, pad = 2, 1

3. **Feature Dimensions of Convolutional Neural Network** In this problem, we compute output feature shape of convolutional layers and pooling layers, which are building blocks of CNN. Let's assume that input feature shape is $W \times H \times C$, where W is the width, H is the height and C is the number of channels of input feature.

   (a) A convolutional layer has 4 hyperparameters: the filter size(K), the padding size (P), the stride step size (S) and the number of filters (F). How many weights and biases in this convolutional layer? And what is the shape of output feature that this convolutional layer produces?

   (b) A pooling layer has 2 hyperparameters: the stride step size(S) and the filter size (K). What is the output feature shape that this pooling layer produces?

   (c) Let's assume that we have the CNN model which consists of L successive convolutional layers and the filter size is K and the stride step size is 1 for every convolutional layer. Then what is the receptive field size?

   (d) Consider a downsampling layer (e.g. pooling layer and strided convolution layer). In this problem, we investigate pros and cons of downsampling layer. This layer reduces the output feature resolution and this implies that the output features loose the certain amount of spatial information. Therefore when we design CNN, we usually increase the channel length to compensate this loss. For example, if we apply the max pooling layer with kernel size of 2 and stride size of 2, we increase the output feature size by a factor of 2. **If we apply this max pooling layer, how much the receptive field increases? Explain the advantage of decreasing the output feature resolution with the perspective of reducing the amount of computation.**

4. Coding Question: BatchNorm

Look at the BatchNormalization.ipynb. In this notebook, you'll implement batch normalization layer. For this question, please submit a .zip file your completed work to the Gradescope assignment titled "HW 3 (Code)". No written portion.

(a) Implement forward operation of batch normalization layer.

(b) Implement backward operation of batch normalization layer.

(c) Implement Fully Connected Nets with Batch Normalization

5. Coding Question: Desigining 2D Filter

Look at the HandDesignFilters.ipynb. In this notebook, you'll design 2D image filter by hand. For this question, please submit a .zip file your completed work to the Gradescope assignment titled "HW 3 (Code)". No written portion.

(a) Design averaging filter.

(b) Design edge detection filter.

6. Coding Question: CNN

Look at the ConvolutionalNetworks.ipynb. In this notebook, you'll implement Convolutional Neural Networks. For this question, please submit a .zip file your completed work to the Gradescope assignment titled "HW 3 (Code)". No written portion.

(a) Implement forward operation of convolutional layer and max pooling layer.

(b) Implement Three-layer ConvNet

(c) Implement forward operation of spatial batch normalization layer.

7. Homework Process and Study Group

Citing sources and collaborators are an important part of life, including being a student!
We also want to understand what resources you find helpful and how much time homework is taking, so we can change things in the future if possible.

(a) **What sources (if any) did you use as you worked through the homework?**

(b) **If you worked with someone on this homework, who did you work with?**
List names and student ID's. (In case of homework party, you can also just describe the group.)

(c) **Roughly how many total hours did you work on this homework? Write it down here where you'll need to remember it for the self-grade form.**

Contributors:

- Suhong Moon.
- Gabriel Goh.
- Anant Sahai.

- Dominic Carrano.
- Babak Ayazifar.
- Sukrit Arora.
- Fei-Fei Li.
- Sheng Shen.
- Jake Austin.
- Kevin Li.

