Homework 2 - Version 1.0

Deadline: Monday, Feb.10, at 11:59pm.

Submission: You must submit your solutions as a PDF file through MarkUs[1]. You can produce the file however you like (e.g. LaTeX, Microsoft Word, scanner), as long as it is readable.

See the syllabus on the course website[2] for detailed policies. You may ask questions about the assignment on Piazza[3]. *Note that 10% of the homework mark may be removed for a lack of neatness.*

The teaching assistants for this assignment are Ian Shi and Andrew Jung.
mailto:csc413-2020-01-tas@cs.toronto.edu

1 Optimization

This week, we will continue investigating the properties of optimization algorithms, focusing on stochastic gradient descent and adaptive gradient descent methods. For a refresher on optimization, please refer to: https://csc413-2020.github.io/assets/readings/L04.pdf.

We will continue using the linear regression model established in Homework 1. Given n pairs of input data with d features and scalar labels $(\mathbf{x_i}, t_i) \in \mathbb{R}^d \times \mathbb{R}$, we wish to find a linear model $f(\mathbf{x}) = \hat{\mathbf{w}}^T\mathbf{x}$ with $\hat{\mathbf{w}} \in \mathbb{R}^d$ such that the squared error on training data is minimized. Given a data matrix $X \in \mathbb{R}^{n\times d}$ and corresponding labels $\mathbf{t} \in \mathbb{R}^n$, the objective function is defined as:

$$\mathcal{L} = \frac{1}{n}\|X\hat{\mathbf{w}} - \mathbf{t}\|_2^2 \tag{1}$$

1.1 Stochastic Gradient Descent (SGD)

SGD performs optimization by taking a stochastic estimate of the gradient from a single training example. This process is iterated until convergence is reached. Let $\mathbf{x}_i \in \mathbb{R}^d$, $1 \le i \le n$ be a single training datum taken from the data matrix X. $\mathcal{L}_i$ denotes the loss with respect to $\mathbf{x}_i$, the update for a single step of SGD at time t with scalar learning rate η is:

$$\mathbf{w}_{t+1} \leftarrow \mathbf{w}_t - \eta\nabla_{w_t}\mathcal{L}_i(\mathbf{x}_i, \mathbf{w}_t) \tag{2}$$

SGD iterates by randomly drawing training samples and updating model weights using the above equation until convergence is reached.

1.1.1 Minimum Norm Solution [2pt]

Recall Question 3.4 from Homework 1. For an overparameterized linear model, $d > n$, gradient descent (GD) starting from zero initialization finds the unique minimum norm solution $\mathbf{w}^*$ such that $X\mathbf{w}^* = \mathbf{t}$. Let $\mathbf{w}_0 = \mathbf{0}$, $d > n$. Assume SGD also converges to a solution $\hat{\mathbf{w}}$ such that $X\hat{\mathbf{w}} = \mathbf{t}$. Show that SGD solution is identical to the minimum norm solution $\mathbf{w}^*$ obtained by gradient descent, i.e., $\hat{\mathbf{w}} = \mathbf{w}^*$.

Hint: Reuse the properties shown in Homework 1 Q3.4. Is $\mathbf{x}_i$ contained in span of X? Do the update steps of SGD ever leave the span of X?

[1] https://markus.teach.cs.toronto.edu/csc413-2020-01
[2] https://csc413-2020.github.io/assets/misc/syllabus.pdf
[3] https://piazza.com/class/k58ktbdnt0h1wx?cid=1

1.1.2 Mini-batch SGD [0pt]

Recall that mini-batch SGD performs stochastic gradient descent by considering the gradient of mini-batches of data $B \in \mathbb{R}^{b \times d}$ where $1 < b << n$ and B is taken from the rows of X. Under the assumptions in Question 1.1.1, does mini-batch stochastic gradient descent obtain the minimum norm solution on convergence?

1.2 Adaptive Methods

We will next consider the behavior of adaptive gradient descent methods. In particular, we will investigate the AdaGrad[4] method. Let w_i denote the i-th parameter. A scalar learning rate η is used. At time t for parameter i, the update step for AdaGrad is shown by:

$$w_{i,t+1} = w_{i,t} - \frac{\eta}{\sqrt{G_i, t + \epsilon}} \nabla_{w_{i,t}} \mathcal{L}(w_{i,t}) \tag{3}$$

$$G_{i,t} = G_{i,t-1} + (\nabla_{w_{i,t}} \mathcal{L}(w_{i,t}))^2 \tag{4}$$

The term ϵ is a fixed small scalar used for numerical stability. Intuitively, Adagrad can be thought of as adapting the learning rate in each dimension to efficiently move through badly formed curvatures (see lecture slides/notes).

1.2.1 Minimum Norm Solution [0pt]

Consider the overparameterized linear model ($d > n$) for the loss function defined in Section 1. Assume the AdaGrad optimizer converges to a solution. Provide a proof or counterexample for whether AdaGrad always obtains the minimum norm solution.

Hint: Compute the 2D case from HW1. Let $\mathbf{x}_1 = [2, 1]$, $w_0 = [0, 0]$, $t = [2]$.

1.2.2 General case [0pt]

Consider the result from the previous section. Does this result hold true for other adaptive methods (RMSprop, Adam) in general? Why might making learning rates independent per dimension be desireable?

2 Gradient-based Hyper-parameter Optimization

In this problem, we will implement a simple toy example of *gradient-based hyper-parameter optimization*, introduced in Lecture 3 (slides 21).

Often in practice, hyper-parameters are chosen by trial-and-error based on a model evaluation criterion. Instead, *gradient-based hyper-parameter optimization* computes gradient of the evaluation criterion w.r.t. the hyper-parameters and use this gradient to directly optimize for the best set of hyper-parameters. For this problem, we will optimize for the learning rate of gradient descent in a linear regression problem, like in homework 1.

Similar to homework 1, a linear model will be used for this problem. Specifically, given n pairs of input data with d features and scalar label $(\mathbf{x}_i, t_i) \in \mathbb{R}^d \times \mathbb{R}$, we wish to find a linear model $f(\mathbf{x}) = \hat{\mathbf{w}}^\top \mathbf{x}$ with $\hat{\mathbf{w}} \in \mathbb{R}^d$ that minimizes the squared error of prediction on the training samples.

[4] `http://www.jmlr.org/papers/volume12/duchi11a/duchi11a.pdf`

Using the concise notation for the data matrix $X \in \mathbb{R}^{n \times d}$ and the corresponding label vector $\mathbf{t} \in \mathbb{R}^n$, the squared error loss can be written as:

$$\mathcal{L} = \frac{1}{n} \|X\hat{\mathbf{w}} - \mathbf{t}\|_2^2.$$

Starting with the initial weights $\mathbf{w}_0$, gradient descent (GD) updates $\mathbf{w}_0$ with a learning rate η for t number of iterations. Let's denote the weights after t GD iterations as $\mathbf{w}_t$, the loss as $\mathcal{L}_t$, and its gradient as $\nabla_{\mathbf{w}_t}$. The goal is to find the optimal learning rate by following the gradient of $\mathcal{L}_t$ w.r.t. the learning rate η.

2.1 Computation Graph of Learning Rates [2pt]

2.1.1

Consider a case of 2 GD iterations. Draw the computation graph to obtain the final loss $\mathcal{L}_2$ in terms of $\mathbf{w}_0, \mathcal{L}_0, \nabla_{\mathbf{w}_0}\mathcal{L}_0, \mathbf{w}_1, \mathcal{L}_1, \nabla_{\mathbf{w}_1}\mathcal{L}_1, \mathbf{w}_2$, and η.

2.1.2

Then, consider a case of t GD updates. What is the memory complexity for the forward-propagation to compute $\mathcal{L}_t$ in terms of t? What is the memory complexity for using the standard back-propagation to compute the gradient w.r.t. the learning rate, $\nabla_\eta \mathcal{L}_t$ in terms of t?

2.1.3

Explain one potential problem for applying gradient-based hyper-parameter optimization in more realistic examples where models often take many iterations to converge.

2.2 Learning Learning Rates [2pt]

In this section, we will take a closer look at the gradient w.r.t. the learning rate. Let's start with the case with only one GD iteration, where GD updates the model weight from $\mathbf{w}_0$ to $\mathbf{w}_1$.

2.2.1

Write down the expression of $\mathbf{w}_1$ in terms of $\mathbf{w}_0$, η, $\mathbf{t}$ and X. Then, using the expression to derive the loss $\mathcal{L}_1$ after single GD iteration in terms of η.
Hint: if the expression gets too messy, introduce a constant vector $\mathbf{a} = X\mathbf{w}_0 - \mathbf{t}$

2.2.2

Determine if this $\mathcal{L}_1$ is convex w.r.t. the learning rate η.
Hint: a function is convex if its second order derivative is positive

2.2.3

Write down the derivative of $\mathcal{L}_1$ w.r.t. η and use it to find the optimal learning rate η^* that minimizes the loss after one GD iteration.

2.3 Multiple Inner-loop Iterations [0pt]

2.3.1

Derive the expression of the loss $\mathcal{L}_t$ after t gradient descent updates.
Hint: proof by induction and binomial coefficients can be useful

2.3.2

Determine if this $\mathcal{L}_t$ is in general convex w.r.t. the learning rate η?

2.3.3

For the previous 2D over-parameterized case from homework 1, describe the convexity of $\mathcal{L}_t$ w.r.t. η. What is the optimal η?

3 Convolutional Neural Networks

The last set of questions aims to build basic familiarity with Convolutional Neural Networks. To refresh your knowledge, please refer to the reading on CNNs: `https://csc413-2020.github.io/assets/readings/L05.pdf`

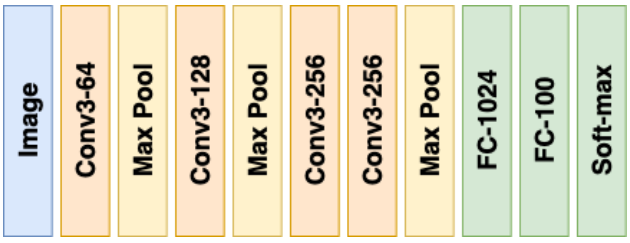
3.1 Convolutional Filters [1pt]

Given the input matrix $\mathbf{I}$ and filter $\mathbf{J}$ shown below, 1) Write down the values of the resulting matrix $(\mathbf{I} * \mathbf{J})$ (the convolution operation as defined in the Lec 5 slides). Assume we have use zero padding around the input. 2) What feature does this convolutional filter detect?

$$\mathbf{I} = \begin{bmatrix} 0 & 1 & 1 & 0 & 0 \\ 0 & 1 & 1 & 1 & 0 \\ 1 & 1 & 1 & 1 & 0 \\ 0 & 1 & 1 & 1 & 0 \\ 0 & 0 & 1 & 0 & 0 \end{bmatrix} \quad \mathbf{J} = \begin{bmatrix} 0 & -1 & 0 \\ -1 & 4 & -1 \\ 0 & -1 & 0 \end{bmatrix} \quad \mathbf{I} * \mathbf{J} = \begin{bmatrix} ? & ? & ? & ? & ? \\ ? & ? & ? & ? & ? \\ ? & ? & ? & ? & ? \\ ? & ? & ? & ? & ? \\ ? & ? & ? & ? & ? \end{bmatrix}$$

3.2 Size of ConvNets [1pt]

Consider a conv net with 6 conv layers like in the diagram below. All 6 conv layers have kernel size of 3×3. The number after the hyphen specifies the number of output channels or units of a layer (e.g. *Conv3-64* layer has 64 output channels and *FC-1024* has 1024 output units). All the *Max Pool* in the diagram has size of 2×2.



Size of the RGB input image is 112×112 (3 channels).
Calculate the number of parameters for this conv net including the bias units.