EECS 182 Deep Neural Networks
Fall 2023 Anant Sahai

Homework 0

This homework is due on Tuesday, Aug 29, 2023, at 10:59PM.

1. Gradient Descent Doesn't Go Nuts with Ill-Conditioning

Consider a linear regression problem with n training points and d features. When $n = d$, the feature matrix $F \in \mathbb{R}^{n \times n}$ has some maximum singular value α and an extremely tiny minimum singular value. We have noisy observations $\mathbf{y} = F\mathbf{w}^* + \boldsymbol{\epsilon}$. If we compute $\hat{\mathbf{w}}_{inv} = F^{-1}\mathbf{y}$, then due to the tiny singular value of F and the presence of noise we observe that $\|\hat{\mathbf{w}}_{inv} - \mathbf{w}^*\|_2 = 10^{10}$.

Suppose instead of inverting the matrix we decide to use gradient descent instead. We run k iterations of gradient descent to minimize the loss $\ell(w) = \frac{1}{2}\|\mathbf{y} - F\mathbf{w}\|^2$ starting from $\mathbf{w}_0 = \mathbf{0}$. We use a learning rate η which is *small enough* that gradient descent cannot possibly diverge for the given problem. (**This is important. You will need to use this.**)

The gradient-descent update for $t > 0$ is:

$$\mathbf{w}_t = \mathbf{w}_{t-1} - \eta \left(F^\top (F\mathbf{w}_{t-1} - \mathbf{y}) \right).$$

We are interested in the error $\|\mathbf{w}_k - \mathbf{w}^*\|_2^2$. We want to show that in the worst case, this error can grow at most linearly with iterations k and in particular $\|\mathbf{w}_k - \mathbf{w}^*\|_2 \leq k\eta\alpha\|\mathbf{y}\|_2 + \|\mathbf{w}^*\|_2$.

i.e. The error cannot go "nuts," at least not very fast.

For the purposes of the homework, you only have to prove the key idea, since the rest follows by applying induction and the triangle inequality.

Show that for $t > 0, \|\mathbf{w}_t\|_2 \leq \|\mathbf{w}_{t-1}\|_2 + \eta\alpha\|\mathbf{y}\|_2$.

(HINT: What do you know about $(I - \eta F^\top F)$ if gradient descent cannot diverge? What are its eigenvalues like? Use this fact.)

2. Regularization from the Augmentation Perspective

Assume $\mathbf{w}$ is a d-dimensional Gaussian random vector $\mathbf{w} \sim \mathcal{N}(\mathbf{0}, \Sigma)$ and Σ is symmetric positive-definite. Our model for how the $\{y_i\}$ training data is generated is

$$y = \mathbf{w}^\top \mathbf{x} + Z, \quad Z \sim \mathcal{N}(0, 1), \tag{1}$$

where the noise variables Z are independent of $\mathbf{w}$ and iid across training samples. Notice that all the training $\{y_i\}$ and the parameters $\mathbf{w}$ are jointly normal/Gaussian random variables conditioned on the training inputs $\{\mathbf{x}_i\}$. Let us define the standard data matrix and measurement vector:

$$X = \begin{bmatrix} \mathbf{x}_1^T \\ \mathbf{x}_2^T \\ \vdots \\ \mathbf{x}_n^T \end{bmatrix}, \quad \mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}.$$

In this model, the MAP estimate of $\mathbf{w}$ is given by the Tikhonov regularization counterpart of ridge regression:

$$\widehat{\mathbf{w}} = (X^\top X + \Sigma^{-1})^{-1} X^\top \mathbf{y}, \tag{2}$$

In this question, we explore Tikhonov regularization from the data augmentation perspective.

Define the matrix Γ as a $d \times d$ matrix that satisfies $\Gamma^\top \Gamma = \Sigma^{-1}$. Consider the following augmented design matrix (data) $\hat{X}$ and augmented measurement vector $\hat{y}$:

$$\hat{X} = \begin{bmatrix} X \\ \Gamma \end{bmatrix} \in \mathbb{R}^{(n+d)\times d}, \quad \text{and} \quad \hat{\mathbf{y}} = \begin{bmatrix} \mathbf{y} \\ \mathbf{0}_d \end{bmatrix} \in \mathbb{R}^{n+d},$$

where $\mathbf{0}_d$ is the zero vector in $\mathbb{R}^d$. **Show that the ordinary least squares problem**

$$\underset{\mathbf{w}}{\operatorname{argmin}} \|\hat{\mathbf{y}} - \hat{X}\mathbf{w}\|_2^2$$

has the same solution as (2).

(HINT: Feel free to just use the formula you know for the OLS solution. You don't have to rederive that. This problem is not intended to be hard or time consuming.)

3. Vector Calculus Review

Let $\mathbf{x}, \mathbf{c} \in \mathbb{R}^n$ and $A \in \mathbb{R}^{n\times n}$. For the following parts, before taking any derivatives, identify what the derivative looks like (is it a scalar, vector, or matrix?) and how we calculate each term in the derivative. Then carefully solve for an arbitrary entry of the derivative, then stack/arrange all of them to get the final result. Note that the convention we will use going forward is that vector derivatives of a scalar (with respect to a column vector) are expressed as a row vector, i.e. $\frac{\partial f}{\partial \mathbf{x}} = [\frac{\partial f}{\partial x_1}, \frac{\partial f}{\partial x_2}, ..., \frac{\partial f}{\partial x_n}]$ since a row acting on a column gives a scalar. You may have seen alternative conventions before, but the important thing is that you need to understand the types of objects and how they map to the shapes of the multidimensional arrays we use to represent those types.

(a) Show $\frac{\partial}{\partial \mathbf{x}}(\mathbf{x}^T\mathbf{c}) = \mathbf{c}^T$

(b) Show $\frac{\partial}{\partial \mathbf{x}}||\mathbf{x}||_2^2 = 2\mathbf{x}^T$

(c) Show $\frac{\partial}{\partial \mathbf{x}}(A\mathbf{x}) = A$

(d) Show $\frac{\partial}{\partial \mathbf{x}}(\mathbf{x}^T A\mathbf{x}) = \mathbf{x}^T(A + A^T)$

(e) Under what condition is the previous derivative equal to $2\mathbf{x}^T A$?

4. ReLU Elbow Update under SGD

In this question we will explore the behavior of the ReLU nonlinearity with Stochastic Gradient Descent (SGD) updates. The hope is that this problem should help you build a more intuitive understanding for how SGD works and how it iteratively adjusts the learned function.

We want to model a 1D function $y = f(x)$ using a 1-hidden layer network with ReLU activations and no biases in the linear output layer. Mathematically, our network is

$$\hat{f}(x) = \mathbf{W}^{(2)}\Phi\left(\mathbf{W}^{(1)}x + \mathbf{b}\right)$$

where $x, y \in \mathbb{R}$, $\mathbf{b} \in \mathbb{R}^d$, $\mathbf{W}^{(1)} \in \mathbb{R}^{d \times 1}$, and $\mathbf{W}^{(2)} \in \mathbb{R}^{1 \times d}$. We define our loss function to be the squared error,

$$\ell\left(x, y, \mathbf{W}^{(1)}, \mathbf{b}, \mathbf{W}^{(2)}\right) = \frac{1}{2}\left\|\hat{f}(x) - y\right\|_2^2.$$

For the purposes of this problem, we define the gradient of a ReLU at 0 to be 0.

(a) Let's start by examining the behavior of a single ReLU with a linear function of x as the input,

$$\phi(x) = \begin{cases} wx + b, & wx + b > 0 \\ 0, & \text{else} \end{cases}.$$

Notice that the slope of $\phi(x)$ is w in the non-zero domain.
We define a loss function $\ell(x, y, \phi) = \frac{1}{2}\|\phi(x) - y\|_2^2$. **Find the following:**

(i) The location of the 'elbow' e of the function, where it transitions from 0 to something else.

(ii) The derivative of the loss w.r.t. $\phi(x)$, namely $\frac{d\ell}{d\phi}$

(iii) The partial derivative of the loss w.r.t. w, namely $\frac{\partial \ell}{\partial w}$

(iv) The partial derivative of the loss w.r.t. b, namely $\frac{\partial \ell}{\partial b}$

(b) Now suppose we have some training point (x, y) such that $\phi(x) - y = 1$. In other words, the prediction $\phi(x)$ is 1 unit above the target y — we are too high and are trying to pull the function downward.

Describe what happpens to the slope and elbow of $\phi(x)$ when we perform gradient descent in the following cases:

(i) $\phi(x) = 0$**.**

(ii) $w > 0$**,** $x > 0$**, and** $\phi(x) > 0$**. It is fine to check the behavior of the elbow numerically in this case.**

(iii) $w > 0$**,** $x < 0$**, and** $\phi(x) > 0$**.**

(iv) $w < 0$**,** $x > 0$**, and** $\phi(x) > 0$**. It is fine to check the behavior of the elbow numerically in this case.**

Additionally, draw and label $\phi(x)$, the elbow, and the qualitative changes to the slope and elbow after a gradient update to w and b. You should label the elbow location and a candidate (x, y) pair. Remember that the update for some parameter vector $\mathbf{p}$ and loss ℓ under SGD is

$$\mathbf{p}' = \mathbf{p} - \lambda \nabla_{\mathbf{p}}(\ell),\ \lambda > 0.$$

(c) Now we return to the full network function $\hat{f}(x)$. **Derive the location e_i of the elbow of the i'th elementwise ReLU activation.**

(d) **Derive the new elbow location e_i' of the i'th elementwise ReLU activation after one stochastic gradient update with learning rate λ.**

5. Using PyTorch to Learn the Color Organ

One of the greatest recent developments in easy-to-use software packages is the easy availability of automatic differentiation. Although the underlying technology had been well established for over four decades (originally developed for control and scientific modeling applications in the context of differential equations), today packages like PyTorch expose this power to us in an easy to use way. This means that we as human engineers no longer have to worry about manually computing derivatives for any purpose other than taking exams, doing proofs, and basically learning material. In practical applications, the computer can do it

for us without any bugs. As students whose careers will span the next four decades, we want you to consider this as a part of your engineering inheritance so that you can use it freely without thinking of it as being any more special than a hash table.

This problem and the accompanying Jupyter Notebook `color_organ_learning.ipynb` will show you how this power can be used to learn the value for the resistors in the color organ simply by having examples of where you want the LEDs to be on and off. This will build on the use of PyTorch that you will have seen in discussion.

(NOTE: A "color organ" is a fun hardware lab exercise where students build an analog circuit where different LEDs light up depending on whether a particular tone is a low frequency, high frequency, etc. This shows the intimate connection between filters and classifiers. Students in 16B often have to build such a circuit and manually tune it so that it recognizes the right kind of tones.)

(a) Let's start with an example low pass filter where, given a desired transfer function, we can determine a resistor value manually for our predicted transfer function such that the transfer functions match. In the interactive plot, we show a transfer function of a desired low pass filter (orange dotted line); we want to design our circuit (given a fixed capacitor value) such that predicted transfer function (blue solid line) is equivalent. For the following problems, we provide a function `evaluate_lp_circuit` that evaluates the transfer function magnitude given a resistor value. Note that these functions use `torch` functions instead of `numpy` as we will typically use `torch` tensors instead of `numpy` arrays in this notebook for training. **Use the slider to find a resistor value such that the predicted and desired transfer functions match and report the resistor value.**

The use of `torch` tensors instead of `numpy` arrays allows the package to do the bookkeeping behind the scenes that allows derivatives to be easily calculated. This will be important in later parts.

(b) Now, suppose that instead of seeing the entire transfer function that we want to match, we are only given some data about which frequencies lie in the pass band (i.e., which frequencies cause the LED on our color organ to be lit). Again, we can determine a resistor value manually. In the interactive plot, we show a transfer function of a low pass filter with its corresponding cutoff frequency. The table shows the desired behavior (red bars denote that the LED is on for a given frequency while black bars denote that the LED is off); we want to design the circuit such that the LEDs light up in the same way. **Use the slider to find a resistor value and corresponding cutoff frequency such that the predicted lights match the desired lights and report the corresponding resistor value and cutoff frequency.**

(c) Assume that we are able to query the desired transfer function directly (i.e., we can play a frequency and record the magnitude of the output). Can we *learn* the resistor value in the low pass circuit directly from this data (instead of manually designing the circuit as you did in the first part)?

The code for this part creates a model of the low pass filter in PyTorch, generates training data, and then trains the circuit using mean squared error loss until convergence (i.e., loss or gradients are very small) or the maximum number of training steps are reached. Here, we use the `torch.autograd.grad` function to automatically find the derivative of the loss with respect to the input (the resistor value of the low pass circuit). All we need to do is input the transfer function and define the loss!

The plots show how the transfer function of the learned circuit evolved during training, the loss surface, the derivative with respect to each training point at each iteration, and the total gradient at each training iteration. Note that the learned transfer function and resistor value change more slowly when the gradient is small, and more quickly when the gradient is large, but if we continue to iterate, we will converge to a local minimum in the loss surface. Try initializing the circuit with different resistor values (there is an optional argument for the circuit class constructor; if you leave it blank it will be initialized to a random value between 0 and 1000). **How long does the circuit take to converge with**

different initializations? Does it converge to the same value you found in the previous parts? Try changing the learning rate (this parameter controls how far we step at each iteration in the direction of the negative gradient). **What values of `lr` cause training to diverge? What values cause the circuit to converge quickly?**

(d) Now, using the same loss function as in the previous part, let's try to learn the resistor value using only the binary data we have (LED on the color organ being on or off). **Change the code to use binary data instead of the transfer function magnitudes. What is the learned resistor value?** (Hints: Some potentially useful constants are defined at the start of the notebook. The loss function must take in floats (not boolean values).)

(e) What happened in the previous part? We did not converge to the same resistor value as we did in the previous part. Why? Because in trying to fit to the binary data, we can see that the positive and negative samples in our training data are pulling the resistor value in opposite directions (bottom left plot) and the final solution we end up at is where this "tug of war" situation achieves a balance. This balance need not end up where we would like it to. Can we fix this problem by adjusting the loss function? **Adjust the loss function such that the negative samples (where the LEDs should be off) are demanding an output other than 0 in a way that helps get the balance to end up where you want it. Can you find a loss function that yields the same resistor value as when training with the full transfer function?**

(f) Let's use what we learned in the previous parts to also learn our high pass filter from binary data. **Input the high pass filter transfer function into the high pass circuit module (use the same `loss_fn` as you found in the previous part) and fill in the code that updates the value of the resistor at each training step** (Hint: use the low pass filter as an example). **What is the learned resistor value?**

(g) Now let's extend the problem to a circuit with multiple parameters and learn both resistors for a band pass filter. **Input the band pass filter transfer function into the band pass circuit module.** (Hint: you can use the functions previously defined for high and low pass circuits). **Then complete the code for updating both resistors (note that `torch.autograd.grad` returns a tuple of gradients corresponding to each input). What are the learned resistor values?** What happens if the initial resistor values are far from the solution? **Try training with initial resistor values of 900 Ohms each. Does the circuit converge within the maximum number of training steps? How much longer does it take to converge? How large are the gradients and what does the loss surface look like when the resistor values are very far from the correct solution?**

(h) Now that we have tried learning the resistor values for a band pass filter directly from binary data, let's explore a different parametrization of our filter: the Bode Plot. Let's again start by trying to learn cutoff frequencies from samples of a transfer function using two ReLU functions. **Run the code. Does the resulting Bode plot match what you would draw given the underlying transfer function data? Do the cutoff frequencies match those corresponding to the resistor values that were found in the previous part?**

(i) Let's put all of these together to try and learn a color organ circuit from a low pass, high pass, and band pass circuits. Here, we train with a vector of size `(3, n)` which is *one-hot encoded*, meaning that for each of the n datapoints, one of the three values is 1 and the rest are 0. This encoding corresponds to our LEDs being on or off for one and only one of the three filters at a given frequency. **Train the color organ circuit and verify that the learned resistor values match those from the previous parts. Try initializing the resistors to different values; does it take longer or shorter to learn the entire color organ circuit than a single one of the filters (low pass, high pass, or band pass)?**

The final part of the notebook visualizes the computation graph that PyTorch is using to compute the derivatives of each transfer function with respect to the resistor values. See if you can match each operation in the graph to the transfer functions for each filter. Hopefully, this graph gives you an idea of

how PyTorch can determine the partial derivatives that you have been using throughout the notebook. Congratulations, you now know how to use the considerable power of PyTorch to automatically differentiate arbitrary functions and find the corresponding local minima of these functions via gradient descent!

6. 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:

- Yaodong Yu.
- Anant Sahai.
- Saagar Sanghavi.
- Josh Sanz.
- Michael Danielczuk.
- Kumar Krishna Agrawal.

