EECS 182 Deep Neural Networks
Fall 2022 Anant Sahai

Homework 1

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

1. 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 $2\mathbf{x}^T A$?

2. Bias-Variance Tradeoff Review

(a) Suppose we have a randomly sampled training set $\mathcal{D}$ (drawn independently of our test data), and we select an estimator denoted $\theta = \hat{\theta}(\mathcal{D})$ (for example, via empirical risk minimization). **Show that we can decompose our expected mean squared error for a test input x into a bias, a variance and an irreducible error term as below:**

$$\mathbb{E}_{Y \sim p(y|x), \mathcal{D}}[(Y - f_{\hat{\theta}(\mathcal{D})}(x))^2] = \mathrm{Var}(f_{\hat{\theta}(\mathcal{D})}(x)) + \mathrm{Bias}(f_{\hat{\theta}(\mathcal{D})}(x))^2 + \sigma^2$$

You may find it helpful to recall the formulaic definitions of Variance and Bias, reproduced for you below:

$$\mathrm{Var}(f_{\hat{\theta}(\mathcal{D})}(x)) = \mathbb{E}_{\mathcal{D}}\left[(f_{\hat{\theta}(\mathcal{D})}(x) - \mathbb{E}[f_{\hat{\theta}(\mathcal{D})}(x)])^2\right]$$
$$\mathrm{Bias}(f_{\hat{\theta}(\mathcal{D})}(x)) = \mathbb{E}_{Y \sim p(Y|x), \mathcal{D}}[f_{\hat{\theta}(\mathcal{D})}(x) - Y]$$

(b) Suppose our training dataset consists of $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^n$ where the only randomness is coming from the label vector $Y = \mathbf{X}\theta^* + \boldsymbol{\varepsilon}$ where θ^* is the true underlying linear model and each noise variable ε_i is i.i.d. with zero mean and variance 1. We use ordinary least squares to estimate a $\hat{\theta}$ from this data. **Calculate the bias and covariance of the $\hat{\theta}$ estimate and use that to compute the bias and variance of the prediction at particular test inputs x.** Recall that the OLS solution is given by

$$\hat{\theta} = (X^\top X)^{-1} X^\top Y,$$

where $X \in \mathbb{R}^{n \times d}$ is our (nonrandom) data matrix, $Y \in \mathbb{R}^d$ is the (random) vector of training targets. For simplicity, assume that $X^\top X$ is diagonal.

3. The Many Interpretations of Ridge Regression

(a) Recall that ridge regression can be understood as the unconstrained optimization problem

$$\min_w \|\mathbf{y} - \mathbf{X}\mathbf{w}\|_2^2 + \lambda \|\mathbf{w}\|_2^2, \tag{1}$$

where $\mathbf{X} \in \mathbb{R}^{n \times d}$ is a design matrix (data), and $\mathbf{y} \in \mathbb{R}^n$ is the target vector of measurement values. One way to interpret "ridge regression" is as the ordinary least squares for an augmented data set — i.e. adding a bunch of fake data points to our data. Consider the following augmented measurement vector $\hat{\mathbf{y}}$ and data matrix $\hat{\mathbf{X}}$:

$$\hat{\mathbf{y}} = \begin{bmatrix} \mathbf{y} \\ \mathbf{0}_d \end{bmatrix} \quad \hat{\mathbf{X}} = \begin{bmatrix} \mathbf{X} \\ \sqrt{\lambda}\mathbf{I}_d \end{bmatrix},$$

where $\mathbf{0}_d$ is the zero vector in $\mathbb{R}^d$ and $\mathbf{I}_d \in \mathbb{R}^{d \times d}$ is the identity matrix. **Show that the optimization problem $\min_w \|\hat{\mathbf{y}} - \hat{\mathbf{X}}\mathbf{w}\|_2^2$ has the same minimizer as (1).**

(b) Perhaps more surprisingly, one can achieve the same effect as in the previous part by adding fake features to each data point instead of adding fake data points. Let's construct the augmented design matrix in the following way:

$$\hat{\mathbf{X}} = [\mathbf{X}\ \alpha\mathbf{I_n}]$$

i.e. we stack $\mathbf{X}$ with $\alpha\mathbf{I}_n$ horizontally. Here α is a scalar multiplier. Now our problem is underdetermined: the new dimension $d + n$ is larger than the number of points n. Therefore, there are infinitely many values $\boldsymbol{\eta} \in \mathbb{R}^{d+n}$ for which $\hat{\mathbf{X}}\boldsymbol{\eta} = \mathbf{y}$. Consider the following problem:

$$\min_{\boldsymbol{\eta}} \|\boldsymbol{\eta}\|_2^2 \text{ s.t. } \hat{\mathbf{X}}\boldsymbol{\eta} = \mathbf{y}. \tag{2}$$

Find the α such that if $\boldsymbol{\eta}^*$ is the minimizer of (2), then the first d coordinates of $\boldsymbol{\eta}^*$ form the minimizer of (1).

Can you interpret what the final n coordinates of $\boldsymbol{\eta}^*$ represent?

(c) We know that the Moore-Penrose pseudo-inverse for an underdetermined system (wide matrix) is given by $A^\dagger = A^T(AA^T)^{-1}$, which corresponds to the min-norm solution for $A\boldsymbol{\eta} = \mathbf{z}$. That is, the optimization problem

$$\min \|\boldsymbol{\eta}\|^2 \text{s.t.} A\boldsymbol{\eta} = \mathbf{z}$$

is solved by $\boldsymbol{\eta} = A^\dagger\mathbf{z}$. Let $\hat{\mathbf{w}}$ be the minimizer of (1). **Use the pseudo-inverse to show that solving to the optimization problem in (2) yields**

$$\hat{\mathbf{w}} = \mathbf{X}^T(\mathbf{X}\mathbf{X}^T + \alpha^2\mathbf{I})^{-1}\mathbf{y}$$

Then, show that with the α you identified in the previous part, this is equivalent to the standard formula for Ridge Regression ($\hat{\mathbf{w}} = (\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^T\mathbf{y}$)

(Hint: For the last part, it might be useful to lookup Kernel Ridge Regression in your machine learning notes or textbook.)

(d) Suppose the SVD of $\mathbf{X}$ is $\mathbf{X} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top$. Let's make a change of coordinates in the feature space, such that $\mathbf{V}$ becomes identity: $\mathbf{X}_{\text{new}} = \mathbf{X}\mathbf{V}$ and $\mathbf{w}_{\text{new}} = \mathbf{V}^\top\mathbf{w}$. Denote the the solution to ridge regression (1) in these new coordinates as $\hat{\mathbf{w}}_{\text{new}}$. **Show that the i-th coordinate of $\hat{\mathbf{w}}_{\text{new}}$ can be obtained from the corresponding coordinate of $\mathbf{U}^\top\mathbf{y}$ by multiplication by $\frac{\sigma_i}{\sigma_i^2+\lambda}$, where σ_i is the i-th singular value of $\mathbf{X}$ (or zero if i is greater than the rank of X.)**

The remaining parts (e-h) of this question are optional.

(e) One reason why we might want to have small weights $\mathbf{w}$ has to do with the sensitivity of the predictor to its input. Let $\mathbf{x}$ be a d-dimensional list of features corresponding to a new test point. Our predictor is $\mathbf{w}^\top\mathbf{x}$. **What is an upper bound on how much our prediction could change if we added noise $\boldsymbol{\epsilon} \in \mathbb{R}^d$ to a test point's features $\mathbf{x}$?**

(f) We know that the solution to ridge regression (1) is given by $\hat{\mathbf{w}}_\mathbf{r} = (\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^\top\mathbf{y}$. **What happens when $\lambda \to \infty$?** It is for this reason that sometimes ridge regularization is referred to as "shrinkage."

(g) Another advantage of ridge regression can be seen for under-determined systems. Say we have the data drawn from a $d = 5$ parameter model, but only have $n = 4$ training samples of it, i.e. $\mathbf{X} \in \mathbb{R}^{4\times5}$. Now this is clearly an underdetermined system, since $n < d$. **Show that ridge regression with $\lambda > 0$ results in a unique solution, whereas ordinary least squares can have an infinite number of solutions.**

[Hint: To make this point, it may be helpful to consider $\mathbf{w} = \mathbf{w}_0 + \mathbf{w}^*$ where $\mathbf{w}_0$ is in the null space of $\mathbf{X}$ and $\mathbf{w}^*$ is a solution.]

[Alternative Hint: You might want to consider (2) as the way to interpret ridge regression.]

(h) For the previous part, **what will the answer be if you take the limit $\lambda \to 0$ for ridge regression?**

[Hint: You might want to consider (2) as the way to interpret ridge regression.]

4. General Case Tikhonov Regularization

Consider the optimization problem:

$$\min_{\mathbf{x}} ||W_1(A\mathbf{x} - \mathbf{b})||_2^2 + ||W_2(\mathbf{x} - \mathbf{c})||_2^2$$

Where W_1, A, and W_2 are matrices and $\mathbf{x}$, $\mathbf{b}$ and $\mathbf{c}$ are vectors. W_1 can be viewed as a generic weighting of the residuals and W_2 along with c can be viewed as a generic weighting of the parameters.

(a) Solve this optimization problem manually by expanding it out as matrix-vector products, setting the gradient to $\mathbf{0}$, and solving for $\mathbf{x}$.

(b) Construct an appropriate matrix $\mathbf{C}$ and vector $\mathbf{d}$ that allows you to rewrite this problem as

$$\min_x \|C\mathbf{x} - \mathbf{d}\|^2$$

and use the OLS solution ($\mathbf{x}^* = (C^TC)^{-1}C^T\mathbf{d}$ to solve. Confirm your answer is in agreement with the previous part.

(c) Under which case would this reduce to the simple case of ridge regression that you've seen in the previous problem $\mathbf{x}^* = (A^TA + \lambda I)^{-1}A^Tb$?

5. System ID for Continuous Systems: Understanding how PyTorch can be used to estimate underlying dynamics given observed samples of a trajectory

(a) Suppose we have a system that we believe is of the form

$$\frac{\mathrm{d}}{\mathrm{d}t}x(t) = \lambda x(t) + bu(t) \tag{3}$$

where $x(t) \in \mathbb{R}$ and $u(t) \in \mathbb{R}$ is a scalar input.

Given an initial condition $x(t_0)$, and that $u(t)$ is some constant input $\bar{u}$ over the interval $[t_0, t_f)$, then for all $t \in [t_0, t_f)$, we know that this differential equation eq. (3) has the unique solution

$$x(t) = x(t_0)e^{\lambda(t-t_0)} + \frac{e^{\lambda(t-t_0)} - 1}{\lambda} b\bar{u}. \tag{4}$$

Assume that we record the state at known times $\tau_0, \tau_1, \ldots, \tau_n$ as having corresponding state values $x_0 = x(\tau_0), x_1 = x(\tau_1), \ldots, x_n = x(\tau_n)$. The continuous-time input is known to be piecewise constant $u(t) = u_i$ for $t \in [\tau_i, \tau_{i+1})$, where we know the sequence of inputs $u_0, u_1, \ldots, u_{n-1}$.

We now want to formulate this as a system ID question by relating the unknown parameters λ, b to the data we have. However, the relationship between the parameters and the data we collected is now non-linear. **For the data point x_{i+1}, use eq. (4) to write out how x_{i+1} should be related to λ and b in the form**

$$x_{i+1} = f(\lambda, x_i, \tau_i, \tau_{i+1}) + bg(\lambda, x_i, u_i, \tau_i, \tau_{i+1}). \tag{5}$$

What are the functions f and g?

(b) The previous part gave rise to a sequence of n equations of the form eq. (5). Because of observation noise and imperfection in our model, we are going to assume that these equations hold only approximately and hope to find values for the two parameters λ, b that minimize the cost function:

$$c(\lambda, b) = \sum_{i=0}^{n-1} \ell\left(x_{i+1}, f(\lambda, x_i, \tau_i, \tau_{i+1}) + bg(\lambda, x_i, u_i, \tau_i, \tau_{i+1})\right) \tag{6}$$

where $f(\lambda, x, \sigma, \tau)$ and $g(\lambda, x, u, \sigma, \tau)$ are given nonlinear scalar functions, and $\ell(s, p)$ is a loss function that penalizes how far the prediction p is from the measured state s. For example, you could use squared loss $\ell(s, p) = (s - p)^2$.

To find the best possible λ_*, b_*, you observe that you want to solve the nonlinear system of equations:

$$\frac{\partial}{\partial\lambda}c(\lambda, b)\Big|_{\lambda_*, b_*} = 0 \tag{7}$$

$$\frac{\partial}{\partial b}c(\lambda, b)\Big|_{\lambda_*, b_*} = 0 \tag{8}$$

and decide to do so using Newton's method starting with an initial guess λ_0, b_0 and linearizing the system of equations eq. (7) and eq. (8) to get a system of linear equations to solve at each step.

The system of linear equations at each iteration $j + 1$ can be expressed in vector form as:

$$A \begin{bmatrix} \lambda - \lambda_j \\ b - b_j \end{bmatrix} = \mathbf{y}. \tag{9}$$

What are the entries of the matrix A and the vector $\mathbf{y}$ in terms of the appropriate partial derivatives of $c(\lambda, b)$ evaluated at the appropriate arguments?

Assume that you can use PyTorch to compute whatever derivatives of $c(\lambda, b)$ that you want — all given functions are sufficiently differentiable. You don't have to take the derivatives by hand. You just need to tell us what derivatives and what arguments to evaluate them at.

6. Movie Ratings with Missing Entries: understanding validation in a way that points towards self-supervision

In a matrix R, you have users' movie ratings. However, not all users watched all the movies.

$$R = \left[\begin{array}{cccc|cc} 0.50 & 0.00 & 0.50 & 0.50 & 0.20 & 1.0 \\ 0.60 & 0.20 & 0.40 & 0.50 & ? & ? \\ 0.50 & 0.50 & 0.00 & 0.25 & 0.60 & 1.0 \\ 0.60 & 0.10 & 0.50 & 0.55 & ? & ? \\ \hline 1.00 & 0.40 & 0.60 & ? & ? & ? \end{array}\right] \tag{10}$$

where the element at the ith row and jth column indicates the rating of movie i by user j. A "?" means that there's no rating for that movie.

Our goal is to predict ratings for the missing entries, so we can recommend movies to users. In order to do this, you want to find the hidden goodness vectors for the movies, and the hidden sensitivity vectors of the users. However, due to missing entries, it is not possible to run an SVD on the entire rating matrix R.

It turns out that we have a submatrix R' in R that does not have any missing entries.

$$R' = \begin{bmatrix} 0.50 & 0.00 & 0.50 & 0.50 \\ 0.60 & 0.20 & 0.40 & 0.50 \\ 0.50 & 0.50 & 0.00 & 0.25 \\ 0.60 & 0.10 & 0.50 & 0.55 \end{bmatrix} \tag{11}$$

We provide a decomposition of this matrix:

$$R' = \underbrace{\begin{bmatrix} 0.5 & 0.0 \\ 0.4 & 0.2 \\ 0.0 & 0.5 \\ 0.5 & 0.1 \end{bmatrix}}_{G} \underbrace{\begin{bmatrix} 4.0 & 0.0 \\ 0.0 & 2.0 \end{bmatrix}}_{D} \underbrace{\begin{bmatrix} 0.25 & 0.0 & 0.25 & 0.25 \\ 0.5 & 0.5 & 0.0 & 0.25 \end{bmatrix}}_{S} \tag{12}$$

where the ith row of the matrix G represents the goodness row vector $\mathbf{g}_i^\top$ of the movie i, the jth column of the matrix S represents the sensitivity vector $\mathbf{s}_j$ of user j, and each diagonal entry of the matrix D shows the weight each attribute has in determining the rating of a movie by a user.

NOTE: This decomposition in eq. (12) is not an SVD; G and S do not have orthonormal vectors.

(a) Suppose $\mathbf{s}_6$ (the sensitivity vector of user 6) is:

$$\mathbf{s}_6 = \begin{bmatrix} 0.5 \\ 1.0 \end{bmatrix} \tag{13}$$

Use this to **estimate the rating of movie 2 as rated by user 6.** (Show work that uses $\mathbf{s}_6$. Unjustified answers will get no credit.)

(b) For the 5th movie, we have three ratings and want to find two parameters of goodness. **Formulate a least squares problem $A\mathbf{g}_5 \approx \mathbf{b}$ to estimate $\mathbf{g}_5$ (goodness vector of movie 5).** You need to tell us A

explicitly as a matrix with numerical entries. We give you that $\mathbf{b} = \begin{bmatrix} 1.00 \\ 0.40 \\ 0.60 \end{bmatrix}$ since those are the three ratings we know for this movie.

(c) We now consider a ratings matrix R without missing entries (that is different from the previous R) where the matrix is partitioned into four blocks $R_{11}, R_{12}, R_{21}, R_{22}$ as below.

$$R = \begin{bmatrix} R_{11} & R_{12} \\ R_{21} & R_{22} \end{bmatrix} \tag{14}$$

In order to find the optimal number of principal components, we compute a PCA model from the SVD of R_{11} with k principal components, with $k = 2, 3, 4, 5$. We then use the chosen components and the singular values of R_{11} together with the information in R_{12} and R_{21} to create an estimate $\widehat{R}_{22}$ for the held-out ratings in R_{22}. We can also use the first k terms of the SVD of R_{11} to reconstruct $\widehat{R}_{11}$ as the best rank-k approximation to R_{11}.
The training errors $\|R_{11} - \widehat{R}_{11}\|_F^2$ and validation errors $\|R_{22} - \widehat{R}_{22}\|_F^2$ for each candidate choice for k are given in the table below.

Select	k	Training error	Validation error
◯	1	1.428	3.104
◯	2	0.414	2.494
◯	3	0.093	0.462
◯	4	0.017	0.090
◯	5	0.011	0.132
◯	6	0.006	0.161

Find the optimal number of principal components k we should use. (No need to give any justification.)

7. ReLU Elbow Update under SGD (Optional)

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.**

8. Coding Question: Fully Connected Networks

In this last question, you'll implement Fully Connected Networks with ReLU nonlinearity. Look at the Jupyter notebook FullyConnectedNets.ipynb in the hw1 folder to see what you have to do!

For this question, please submit a .zip file your completed work to the Gradescope assignment titled "HW 1 (Code)". No written portion.

9. 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:

- Saagar Sanghavi.
- Brandon Trabucco.
- Alexander Tsigler.
- Anant Sahai.
- Jane Yu.
- Philipp Moritz.
- Soroush Nasiriany.
- Ashwin Vangipuram.
- Jichan Chung.
- Druv Pai.
- Josh Sanz.

