Today’s goals

- Understanding deep neural networks as computational graphs.
  - Forward propagation of inputs to outputs.
  - Backward propagation of loss gradients to weights and biases.
- Understand numerical stability issues in training neural networks.
  - Vanishing or exploding gradients.
- Review of generalization how to use regularization for better generalization.
  - Overfitting, underfitting
  - Weight decay and dropout
Demo: Why multiple layers?

• https://playground.tensorflow.org/
Part I: Neural Networks as a Computational Graph
Review: neural networks with one hidden layer

- Input $\mathbf{x} \in \mathbb{R}^d$
- Hidden $W^{(1)} \in \mathbb{R}^{m \times d}$, $b^{(1)} \in \mathbb{R}^m$
- Intermediate output
  \[ h = \sigma(W^{(1)}\mathbf{x} + b^{(1)}) \]
  $h \in \mathbb{R}^m$
Review: neural networks with one hidden layer

\[ W \times m \times n = x \in \mathbb{R}^d \]
Review: neural networks with one hidden layer

Key elements: linear operations + Nonlinear activations

\[
W \in \mathbb{R}^{m \times d} \quad d \times 1 \quad m \times 1 \quad m \times 1
\]

\[
x \in \mathbb{R}^{d} \quad + \quad b \quad = \quad \text{Element-wise activation function}
\]

\[
\frac{1}{1 + e^{-x}}
\]
Deep neural networks (DNNs)

\[ h_1 = \sigma(W^{(1)}x + b^{(1)}) \]
\[ h_2 = \sigma(W^{(2)}h_1 + b^{(2)}) \]
\[ h_3 = \sigma(W^{(3)}h_2 + b^{(3)}) \]
\[ f = W^{(4)}h_3 + b^{(4)} \]
\[ p = \text{softmax}(f) \]

NNs are composition of nonlinear functions
Neural networks as variables + operations

\[ a = \text{sigmoid}(Wx + b) \]

- Can describe with a computational graph
- Decompose functions into atomic operations
- Separate data (variables) and computing (operations)
Neural networks as a computational graph

• A two-layer neural network
Neural networks as a computational graph

- A two-layer neural network
- Forward propagation vs. backward propagation
Neural networks: forward propagation

- A two-layer neural network
- Intermediate variables $Z$
Neural networks: backward propagation

- A two-layer neural network
- Assuming forward propagation is done
- Minimize a **loss function** $L$
Neural networks: backward propagation

- A two-layer neural network
- Assuming forward propagation is done
- Minimize a **loss function** $L$

\[
\frac{\partial L}{\partial z_5} = \frac{\partial L}{\partial a} \frac{\partial a}{\partial z_5}
\]
Neural networks: backward propagation

- A two-layer neural network
- Assuming forward propagation is done
- Minimize a **loss function** $L$

\[ \frac{\partial L}{\partial z_4} = \frac{\partial L}{\partial z_5} \frac{\partial z_5}{\partial z_4} \]

\[ \frac{\partial L}{\partial b^{(2)}} = \frac{\partial L}{\partial z_5} \frac{\partial z_5}{\partial b^{(2)}} \]
Neural networks: backward propagation

- A two-layer neural network
- Assuming forward propagation is done

\[
\frac{\partial L}{\partial z_3} = \frac{\partial L}{\partial z_4} \frac{\partial z_4}{\partial z_3}
\]

\[
\frac{\partial L}{\partial W^{(2)}} = \frac{\partial L}{\partial z_4} \frac{\partial z_4}{\partial W^{(2)}}
\]
Backward propagation: A modern treatment

- First, define a neural network as a computational graph
  - Nodes are variables and operations.
- Must be a directed graph
- All operations must be differentiable.
- Backpropagation computes partial derivatives starting from the loss and then working backwards through the graph.
Backward propagation: PyTorch

```python
for t in range(2000):
    # Forward pass: compute predicted y by passing x to the model
    # override the __call__ operator so you can call them like a function.
    # doing so you pass a Tensor of input data to the Module
    # a Tensor of output data.
    y_pred = model(xx)

    # Compute and print loss. We pass Tensors containing the
    # values of y, and the loss function returns a Tensor containing
    # loss.
    loss = loss_fn(y_pred, y)
    if t % 100 == 99:
        print(t, loss.item())

    # Zero the gradients before running the backward pass.
    model.zero_grad()

    # Backward pass: compute gradient of the loss with respect to parameters of the model. Internally, the parameters of
    # in Tensors with requires_grad=True, so this call will
    # all learnable parameters in the model.
    loss.backward()

    # Update the weights using gradient descent. Each parameter
    # we can access its gradients like we did before.
    with torch.no_grad():
        for param in model.parameters():
            param -= learning_rate * param.grad
```

Forward propagation

Backward propagation

Gradient Descent
Q1. Suppose we want to solve the following k-class classification problem with cross entropy loss
\[ \ell(y, \hat{y}) = -\sum_{j=1}^{k} y_j \log \hat{y}_j, \]
where the ground truth and predicted probabilities \( y, \hat{y} \in \mathbb{R}^k \). Recall that the softmax function turns output into probabilities:
\[ \hat{y}_j = \frac{\exp f_j(x)}{\sum_{i}^{k} \exp f_i(x)}. \]
What is the partial derivative \( \partial_{f_j} \ell(y, \hat{y}) \)?

A. \( \hat{y}_j - y_j \)
B. \( \exp(y_j) - y_j \)
C. \( y_j - \hat{y}_j \)
Q1. Suppose we want to solve the following $k$-class classification problem with cross entropy loss

$$\ell(y, \hat{y}) = - \sum_{j=1}^{k} y_j \log \hat{y}_j,$$

where $y, \hat{y} \in \mathbb{R}^k$. Recall that the softmax function turns output into probabilities: 

$$\hat{y}_j = \frac{\exp f_j(x)}{\sum_i^{k} \exp f_i(x)}.$$

What is the partial derivative $\partial_j \ell(y, \hat{y})$?

Rewrite

\[ \ell(y, \hat{y}) = - \sum_{j=1}^{k} y_j \log \left( \frac{\exp (f_j)}{\sum_{i=1}^{k} \exp (f_i)} \right) \]

\[ = \sum_{j=1}^{k} y_j \log \sum_{i=1}^{k} \exp (f_i) - \sum_{j=1}^{k} y_j f_j \]

\[ = \log \sum_{i=1}^{k} \exp (f_i) - \sum_{j=1}^{k} y_j f_j. \]

We have

$$\partial_j \ell(y, \hat{y}) = \frac{\exp (f_j)}{\sum_{i=1}^{k} \exp (f_i)} y_j - \hat{y}_j = \hat{y}_j - y_j.$$
Part II: Numerical Stability
Gradients for Neural Networks

• Compute the gradient of the loss $\ell$ w.r.t. $W_t$

$$\frac{\partial \ell}{\partial W^t} = \frac{\partial \ell}{\partial h^d} \frac{\partial h^d}{\partial h^{d-1}} \ldots \frac{\partial h^{t+1}}{\partial h^t} \frac{\partial h^t}{\partial W^t}$$

Multiplication of many matrices
Two Issues for Deep Neural Networks

Gradient Exploding

\[ 1.5^{100} \approx 4 \times 10^{17} \]

Gradient Vanishing

\[ 0.8^{100} \approx 2 \times 10^{-10} \]
Issues with Gradient Exploding

• Value out of range: infinity value (NaN)
• Sensitive to learning rate (LR)
  • Not small enough LR -> larger gradients
  • Too small LR -> No progress
  • May need to change LR dramatically during training
Gradient Vanishing

- Use sigmoid as the activation function

\[ \sigma(x) = \frac{1}{1 + e^{-x}} \quad \sigma'(x) = \sigma(x)(1 - \sigma(x)) \]
Issues with Gradient Vanishing

- Gradients with value 0
- No progress in training
  - No matter how to choose learning rate
- Severe with bottom layers
  - Only top layers are well trained
- No benefit to make networks deeper
How to stabilize training?
Stabilize Training: Practical Considerations

• Goal: make sure gradient values are in a proper range
  • E.g. in $[1e-6, 1e3]$  
  • Multiplication $\rightarrow$ plus  
  • Architecture change (e.g., ResNet)  
• Normalize  
  • Batch Normalization, Gradient clipping  
• Proper activation functions
Quiz. Which of the following are TRUE about the vanishing gradient problem in neural networks? Multiple answers are possible?

A. Deeper neural networks tend to be more susceptible to vanishing gradients.

B. Using the ReLU function can reduce this problem.

C. If a network has the vanishing gradient problem for one training point due to the sigmoid function, it will also have a vanishing gradient for every other training point.

D. Networks with sigmoid functions don’t suffer from the vanishing gradient problem if trained with the cross-entropy loss.
Quiz. Which of the following are TRUE about the vanishing gradient problem in neural networks? Multiple answers are possible?

A. Deeper neural networks tend to be more susceptible to vanishing gradients.

B. Using the ReLU function can reduce this problem.

C. If a network has the vanishing gradient problem for one training point due to the sigmoid function, it will also have a vanishing gradient for every other training point.

D. Networks with sigmoid functions don’t suffer from the vanishing gradient problem if trained with the cross-entropy loss.
Quiz. Let’s compare sigmoid with rectified linear unit (ReLU). Which of the following statement is NOT true?

A. Sigmoid function is more expensive to compute

B. ReLU has non-zero gradient everywhere

C. The gradient of Sigmoid is always less than 0.3

D. The gradient of ReLU is constant for positive input
Quiz. Let’s compare sigmoid with rectified linear unit (ReLU). Which of the following statement is NOT true?

A. Sigmoid function is more expensive to compute

B. ReLU has non-zero gradient everywhere

C. The gradient of Sigmoid is always less than 0.3

D. The gradient of ReLU is constant for positive input
Q5. A Leaky ReLU is defined as \( f(x) = \max(0.1x, x) \). Let \( f'(0) = 1 \). Does it have non-zero gradient everywhere??

A. Yes

B. No
Q5. A Leaky ReLU is defined as $f(x) = \max(0.1x, x)$. Let $f'(0) = 1$. Does it have non-zero gradient everywhere?

A. Yes

B. No
Part III: Generalization & Regularization
How good are the models?
Training Error and Generalization Error

• Training error: model error on the training data
• **Generalization error**: model error on new data
• Example: practice a future exam with past exams
  • Doing well on past exams (training error) doesn’t guarantee a good score on the future exam (generalization error)
Underfitting

Overfitting

Image credit: hackernoon.com
Model Capacity

- The ability to fit variety of functions
- Low capacity models struggles to fit training set
  - Underfitting
- High capacity models can memorize the training set
  - Overfitting
Influence of Model Complexity

* Recent research has challenged this view for some types of models.

Also known as “Test loss”
Estimate Neural Network Capacity

• It’s hard to compare complexity between different families of models.
• e.g. K-NN vs neural networks
• Given a model family, two main factors matter:
  • The number of parameters
  • The values taken by each parameter

\[(d + 1)(d + 1) + (m + 1)k\]
Data Complexity

- Multiple factors matters
  - # of examples
  - # of features in each example
  - time/space structure
  - # of labels
Quiz Break: When training a neural network, which one below indicates that the network has overfit the training data?

A. Training loss is low and generalization loss is high.
B. Training loss is low and generalization loss is low.
C. Training loss is high and generalization loss is high.
D. Training loss is high and generalization loss is low.
E. None of these.
Quiz Break: When training a neural network, which one below indicates that the network has overfit the training data?

A. Training loss is low and generalization loss is high.
B. Training loss is low and generalization loss is low.
C. Training loss is high and generalization loss is high.
D. Training loss is high and generalization loss is low.
E. None of these.
Quiz Break: Adding more layers to a multi-layer perceptron may cause ______.

A. Vanishing gradients during back propagation.
B. A more complex decision boundary.
C. Underfitting.
D. Lower test loss.
E. None of these.
Quiz Break: Adding more layers to a multi-layer perceptron may cause ______.

A. Vanishing gradients during back propagation.
B. A more complex decision boundary.
C. Underfitting.
D. Higher test loss.
E. None of these.
How to regularize the model for better generalization?
Weight Decay
Squared Norm Regularization as Hard Constraint

• Reduce model complexity by limiting value range

\[ \min L(w, b) \text{ subject to } \|w\|^2 \leq B \]

• Often do not regularize bias \( b \)
  • Doing or not doing has little difference in practice
  • A small \( B \) means more regularization
Squared Norm Regularization as Soft Constraint

• We can rewrite the hard constraint version as

\[
\min L(w, b) + \frac{\lambda}{2} ||w||^2
\]
Squared Norm Regularization as Soft Constraint

- We can rewrite the hard constraint version as

$$\min L(w, b) + \frac{\lambda}{2} \|w\|^2$$

- Hyper-parameter $\lambda$ controls regularization importance
- $\lambda = 0$: no effect
- $\lambda \to \infty$, $w^* \to 0$
Illustrate the Effect on Optimal Solutions

\[
\hat{w}^* = \arg \min \ L(w, b) + \frac{\lambda}{2} ||w||^2
\]

\[
w^* = \arg \min \ L(w, b)
\]
Dropout

Hinton et al.
Apply Dropout

- Often apply dropout on the output of hidden fully-connected layers

\[ h = \sigma(W^{(1)}x + b^{(1)}) \]
\[ h' = \text{dropout}(h) \]
\[ o = W^{(2)}h' + b^{(2)} \]
\[ p = \text{softmax}(o) \]
Figure 2: **Left:** A unit at training time that is present with probability $p$ and is connected to units in the next layer with weights $w$. **Right:** At test time, the unit is always present and the weights are multiplied by $p$. The output at test time is same as the expected output at training time.
Figure 4: Test error for different architectures with and without dropout. The networks have 2 to 4 hidden layers each with 1024 to 2048 units.
Q3. In standard dropout regularization, with dropout probability $p$, the each intermediate activation $h$ is replaced by a random variable $h'$ as: $h' = \begin{cases} 0 & \text{with probability } p \\ ? & \text{otherwise} \end{cases}$.

To make $E[h'] = h$. What is “?”

A. $h$
B. $h/p$
C. $h/(1-p)$
D. $h(1-p)$
Q3. In standard dropout regularization, with dropout probability $p$, the each intermediate activation $h$ is replaced by a random variable $h'$ as: 

$$h' = \begin{cases} 
0 & \text{with probability } p \\
? & \text{otherwise}
\end{cases}$$

To make $E[h'] = h$. What is “?”

A. $h$

B. $h/p$

C. $h/(1-p)$

D. $h(1-p)$
What we’ve learned today...

• Deep neural networks
  • Computational graph (forward and backward propagation)
• Numerical stability in training
  • Gradient vanishing/exploding
• Generalization and regularization
  • Overfitting, underfitting
  • Weight decay and dropout