ECE C247 — Homework 03

Chih-En Lin

1.

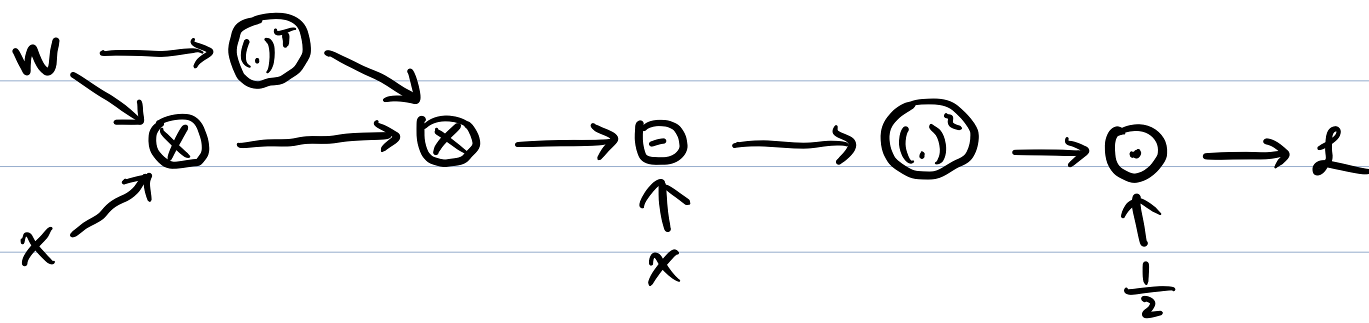
(a) Wx is to encode the information of x.

$W^T W x$ is to decode the information of x.

With the loss $\mathcal{L}$ be minimized, the difference between the reconstructed $W^T W x$ and the original x will be minimized.

Thus, Wx will preserve the information about x. #

(b)



#

(c)

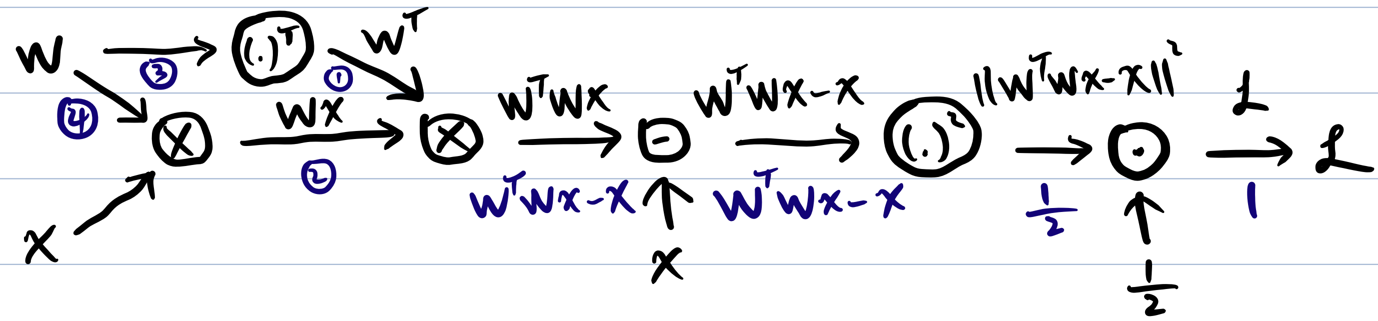
Let two paths are L_1 and L_2.

if $\begin{cases} L_1: a \to b \to d \\ L_2: a \to c \to d \end{cases}$

$$\nabla_W \mathcal{L} = \nabla_W \mathcal{L}_1 + \nabla_W \mathcal{L}_2 = \frac{\partial b}{\partial a} \cdot \frac{\partial d}{\partial b} + \frac{\partial c}{\partial a} \cdot \frac{\partial d}{\partial c}$$

#

(d) Use the computational graph from (b), Calculate the backpropragation



① $\frac{\partial \mathcal{L}}{\partial W^T} = (W^TWx - x)(Wx)^T$

② $\frac{\partial \mathcal{L}}{\partial Wx} = W(W^TWx - x)$

③ : ① Backpropragate to W : $(Wx)(W^TWx - x)^T$

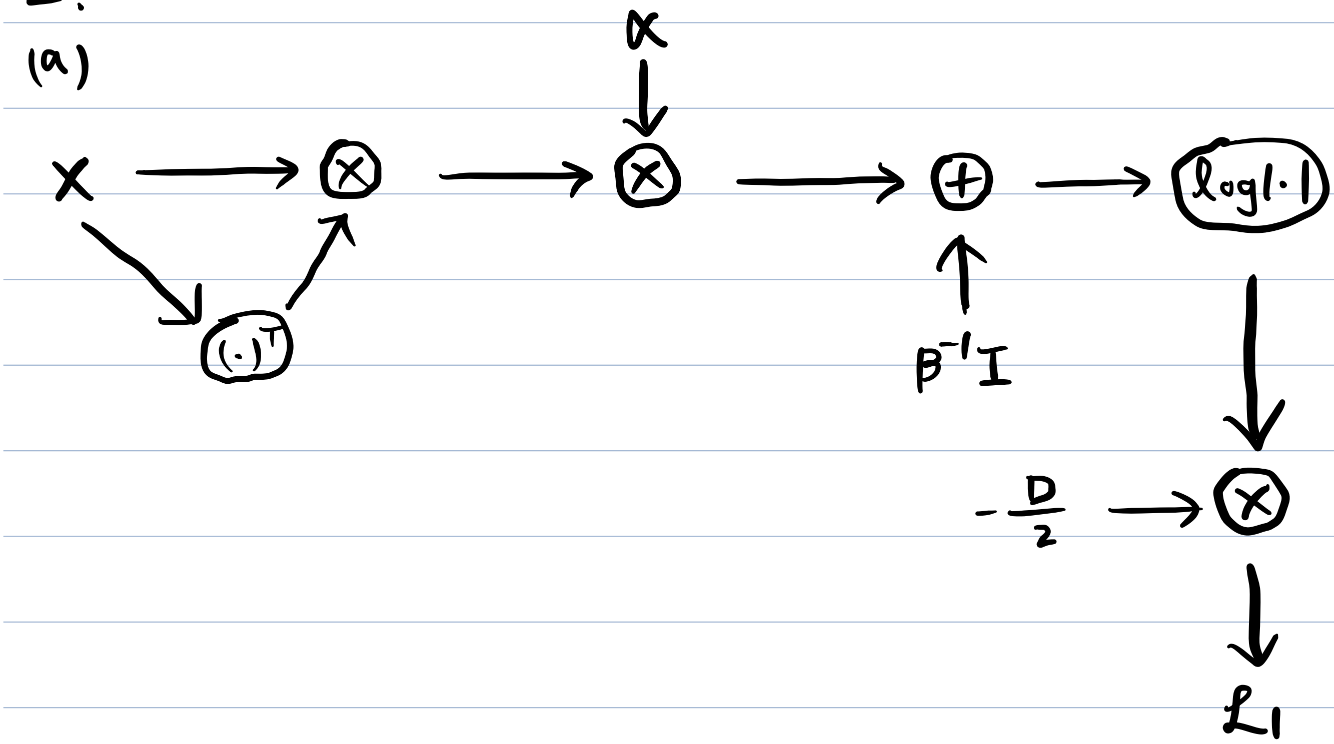
④ : ② Backpropragate to W : $W(W^TWx - x)x^T$

$$\nabla_W \mathcal{L} = ③ + ④ = (Wx)(W^TWx - x)^T + W(W^TWx - x)x^T$$

#

2.

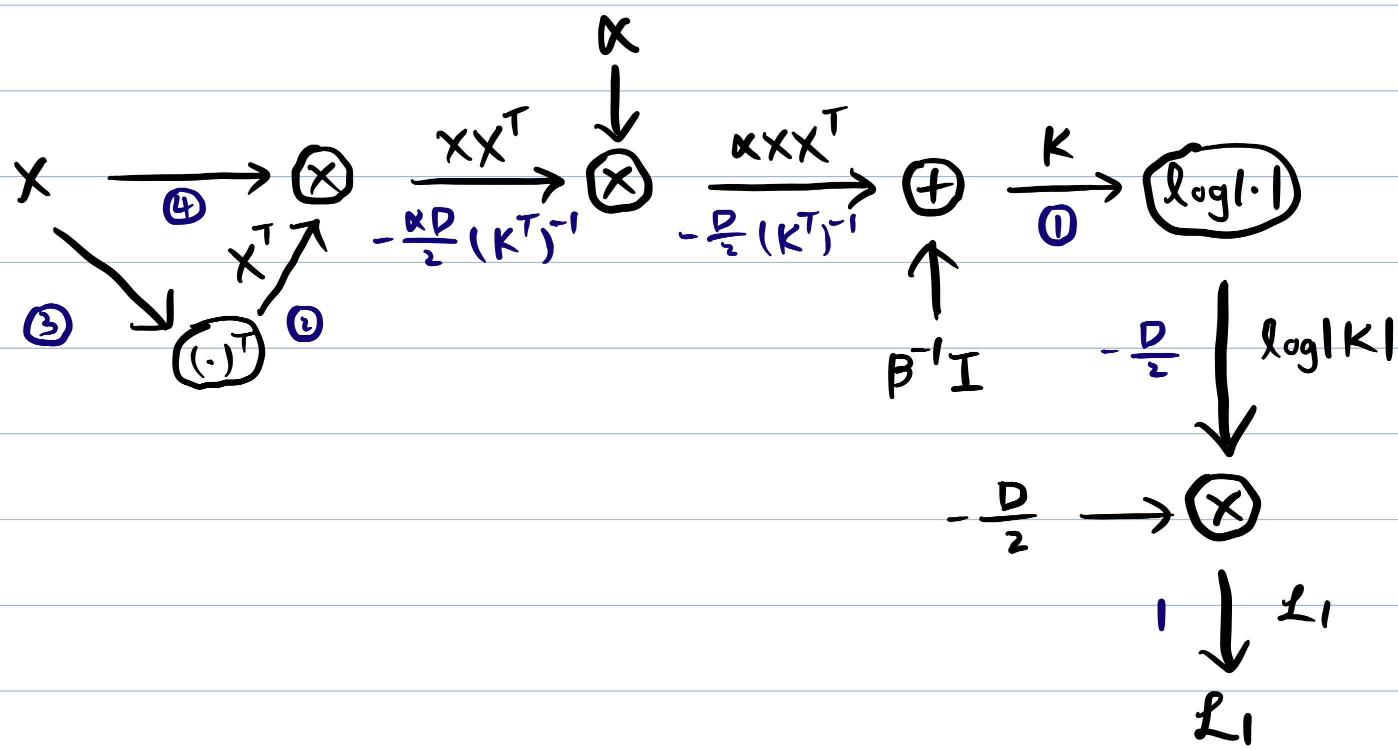
(a)



#

(b)

$$K = \alpha X X^T + \beta^{-1} I$$



$$① = \frac{\partial \mathcal{L}_1}{\partial K} = -\frac{D}{2}(K^T)^{-1}$$

$$② = \frac{\partial \mathcal{L}_1}{\partial X^T} = X^T\left(\frac{-\alpha D}{2}(K^T)^{-1}\right)$$

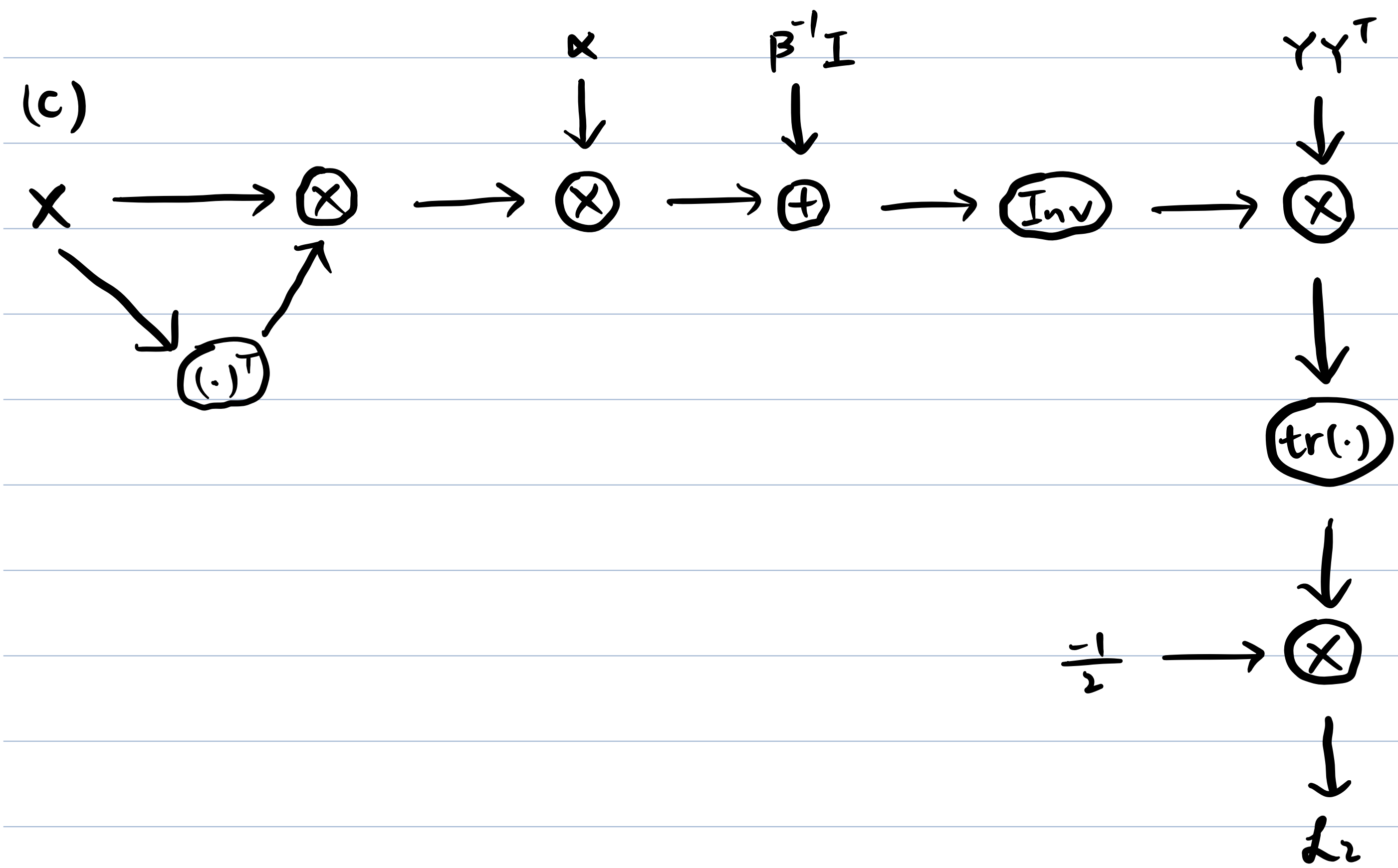
③ $= \frac{-\alpha D}{2} K^{-1} X$

④ $= \frac{-\alpha D}{2} (K^T)^{-1} X$ (K is symmetric)

$$\frac{\partial \mathcal{L}_1}{\partial X} = -\alpha D (K^T)^{-1} X = -\alpha D K^{-1} X$$ #

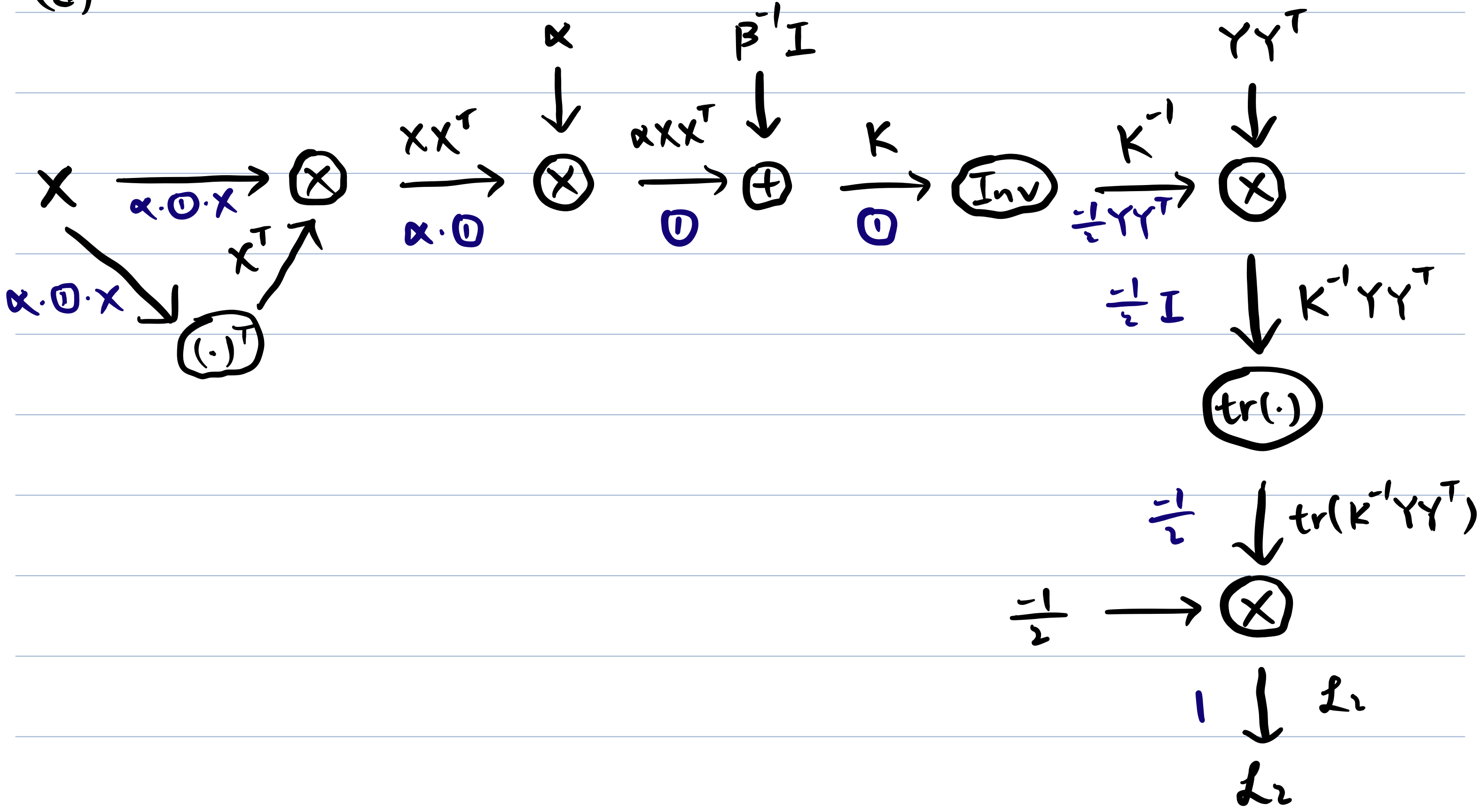
$$(K = \alpha X X^T + \beta^{-1} I)$$

(c)



#

(d)



$$\frac{\partial \mathcal{L}_2}{\partial K^{-1}YY^T} = -\frac{1}{2}I$$

$$① = \frac{\partial \mathcal{L}_2}{\partial K} = -(K^T)^{-1}\left(\frac{-1}{2}YY^T\right)(K^T)^{-1} = \frac{1}{2}K^{-1}YY^TK^{-1}$$

$$\frac{\partial \mathcal{L}_2}{\partial X} = 2\alpha \cdot \left(\frac{1}{2}K^{-1}YY^TK^{-1}\right) \cdot X = \alpha K^{-1}YY^TK^{-1}X$$

#

$$(K = \alpha XX^T + \beta^{-1}I)$$

(e)

$$\frac{\partial \mathcal{L}}{\partial X} = \frac{\partial \mathcal{L}_1}{\partial X} + \frac{\partial \mathcal{L}_2}{\partial X}$$

$$= \alpha K^{-1}YY^TK^{-1}X - \alpha D K^{-1}X$$

#

$$(K = \alpha XX^T + \beta^{-1}I)$$

This is the 2-layer neural network notebook for ECE C147/C247 Homework #3

Please follow the notebook linearly to implement a two layer neural network.

Please print out the notebook entirely when completed.

The goal of this notebook is to give you experience with training a two layer neural network.

In [1]:

```
import random
import numpy as np
from utils.data_utils import load_CIFAR10
import matplotlib.pyplot as plt

%matplotlib inline
%load_ext autoreload
%autoreload 2

def rel_error(x, y):
    """ returns relative error """
    return np.max(np.abs(x - y) / (np.maximum(1e-8, np.abs(x) + np.abs(y))))
```

Toy example

Before loading CIFAR-10, there will be a toy example to test your implementation of the forward and backward pass

In [2]:

```
from nndl.neural_net import TwoLayerNet
```

In [3]:

```
# Create a small net and some toy data to check your implementations.
# Note that we set the random seed for repeatable experiments.

input_size = 4
hidden_size = 10
num_classes = 3
num_inputs = 5

def init_toy_model():
    np.random.seed(0)
    return TwoLayerNet(input_size, hidden_size, num_classes, std=1e-1)

def init_toy_data():
    np.random.seed(1)
    X = 10 * np.random.randn(num_inputs, input_size)
    y = np.array([0, 1, 2, 2, 1])
    return X, y

net = init_toy_model()
X, y = init_toy_data()
```

Compute forward pass scores

In [4]:

```
## Implement the forward pass of the neural network.

# Note, there is a statement if y is None: return scores, which is why
```

```python
# the following call will calculate the scores.
scores = net.loss(X)
print('Your scores:')
print(scores)
print()
print('correct scores:')
correct_scores = np.asarray([
    [-1.07260209,  0.05083871, -0.87253915],
    [-2.02778743, -0.10832494, -1.52641362],
    [-0.74225908,  0.15259725, -0.39578548],
    [-0.38172726,  0.10835902, -0.17328274],
    [-0.64417314, -0.18886813, -0.41106892]])
print(correct_scores)
print()

# The difference should be very small. We get < 1e-7
print('Difference between your scores and correct scores:')
print(np.sum(np.abs(scores - correct_scores)))
```

```
Your scores:
[[-1.07260209  0.05083871 -0.87253915]
 [-2.02778743 -0.10832494 -1.52641362]
 [-0.74225908  0.15259725 -0.39578548]
 [-0.38172726  0.10835902 -0.17328274]
 [-0.64417314 -0.18886813 -0.41106892]]

correct scores:
[[-1.07260209  0.05083871 -0.87253915]
 [-2.02778743 -0.10832494 -1.52641362]
 [-0.74225908  0.15259725 -0.39578548]
 [-0.38172726  0.10835902 -0.17328274]
 [-0.64417314 -0.18886813 -0.41106892]]

Difference between your scores and correct scores:
3.381231233889892e-08
```

Forward pass loss

In [5]:

```python
loss, _ = net.loss(X, y, reg=0.05)
correct_loss = 1.071696123862817

# should be very small, we get < 1e-12
print("Loss:",loss)
print('Difference between your loss and correct loss:')
print(np.sum(np.abs(loss - correct_loss)))
```

```
Loss: 1.071696123862817
Difference between your loss and correct loss:
0.0
```

Backward pass

Implements the backwards pass of the neural network. Check your gradients with the gradient check utilities provided.

In [6]:

```python
from utils.gradient_check import eval_numerical_gradient

# Use numeric gradient checking to check your implementation of the backward pass.
# If your implementation is correct, the difference between the numeric and
# analytic gradients should be less than 1e-8 for each of W1, W2, b1, and b2.

loss, grads = net.loss(X, y, reg=0.05)
```

```
# these should all be less than 1e-8 or so
for param_name in grads:
    f = lambda W: net.loss(X, y, reg=0.05)[0]
    param_grad_num = eval_numerical_gradient(f, net.params[param_name], verbose=False)
    print('{} max relative error: {}'.format(param_name, rel_error(param_grad_num, grads[p
```

```
W2 max relative error: 2.9632227682005116e-10
b2 max relative error: 1.248270530283678e-09
W1 max relative error: 1.2832823337649917e-09
b1 max relative error: 3.172680092703762e-09
```

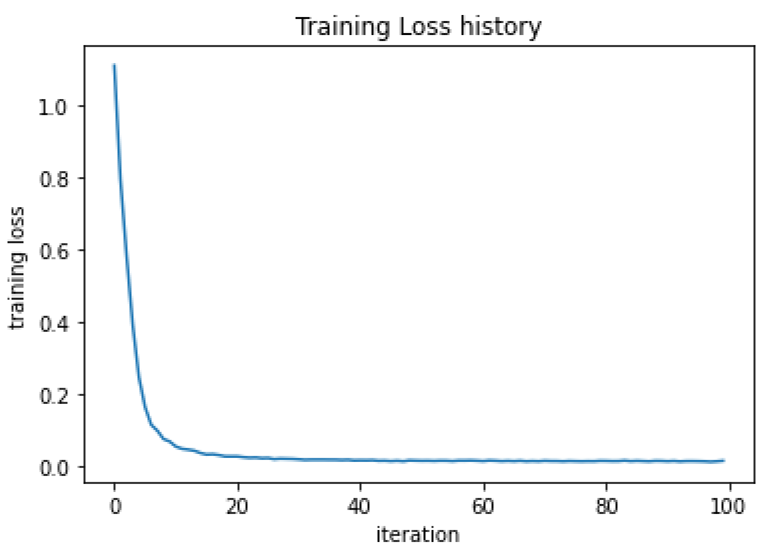
Training the network

Implement neural_net.train() to train the network via stochastic gradient descent, much like the softmax.

In [7]:

```
net = init_toy_model()
stats = net.train(X, y, X, y,
            learning_rate=1e-1, reg=5e-6,
            num_iters=100, verbose=False)

print('Final training loss: ', stats['loss_history'][-1])

# plot the loss history
plt.plot(stats['loss_history'])
plt.xlabel('iteration')
plt.ylabel('training loss')
plt.title('Training Loss history')
plt.show()
```

```
Final training loss:  0.014498902952971647
```



Classify CIFAR-10

Do classification on the CIFAR-10 dataset.

In [8]:

```
from utils.data_utils import load_CIFAR10

def get_CIFAR10_data(num_training=49000, num_validation=1000, num_test=1000):
    """
    Load the CIFAR-10 dataset from disk and perform preprocessing to prepare
    it for the two-layer neural net classifier.
    """
    # Load the raw CIFAR-10 data
```

```
    cifar10_dir = 'cifar-10-batches-py'
    X_train, y_train, X_test, y_test = load_CIFAR10(cifar10_dir)

    # Subsample the data
    mask = list(range(num_training, num_training + num_validation))
    X_val = X_train[mask]
    y_val = y_train[mask]
    mask = list(range(num_training))
    X_train = X_train[mask]
    y_train = y_train[mask]
    mask = list(range(num_test))
    X_test = X_test[mask]
    y_test = y_test[mask]

    # Normalize the data: subtract the mean image
    mean_image = np.mean(X_train, axis=0)
    X_train -= mean_image
    X_val -= mean_image
    X_test -= mean_image

    # Reshape data to rows
    X_train = X_train.reshape(num_training, -1)
    X_val = X_val.reshape(num_validation, -1)
    X_test = X_test.reshape(num_test, -1)

    return X_train, y_train, X_val, y_val, X_test, y_test


# Invoke the above function to get our data.
X_train, y_train, X_val, y_val, X_test, y_test = get_CIFAR10_data()
print('Train data shape: ', X_train.shape)
print('Train labels shape: ', y_train.shape)
print('Validation data shape: ', X_val.shape)
print('Validation labels shape: ', y_val.shape)
print('Test data shape: ', X_test.shape)
print('Test labels shape: ', y_test.shape)
```

```
Train data shape:  (49000, 3072)
Train labels shape:  (49000,)
Validation data shape:  (1000, 3072)
Validation labels shape:  (1000,)
Test data shape:  (1000, 3072)
Test labels shape:  (1000,)
```

Running SGD

If your implementation is correct, you should see a validation accuracy of around 28-29%.

In [9]:

```
input_size = 32 * 32 * 3
hidden_size = 50
num_classes = 10
net = TwoLayerNet(input_size, hidden_size, num_classes)

# Train the network
stats = net.train(X_train, y_train, X_val, y_val,
            num_iters=1000, batch_size=200,
            learning_rate=1e-4, learning_rate_decay=0.95,
            reg=0.25, verbose=True)

# Predict on the validation set
val_acc = (net.predict(X_val) == y_val).mean()
print('Validation accuracy: ', val_acc)
```

```
# Save this net as the variable subopt_net for later comparison.
subopt_net = net
```

```
iteration 0 / 1000: loss 2.302757518613176
iteration 100 / 1000: loss 2.302120159207236
iteration 200 / 1000: loss 2.2956136007408703
iteration 300 / 1000: loss 2.2518259043164135
iteration 400 / 1000: loss 2.188995235046776
iteration 500 / 1000: loss 2.1162527791897747
iteration 600 / 1000: loss 2.064670827698217
iteration 700 / 1000: loss 1.9901688623083942
iteration 800 / 1000: loss 2.002827640124685
iteration 900 / 1000: loss 1.94651768178565
Validation accuracy:  0.283
```

Questions:

The training accuracy isn't great.

(1) What are some of the reasons why this is the case? Take the following cell to do some analyses and then report your answers in the cell following the one below.

(2) How should you fix the problems you identified in (1)?

In [10]:

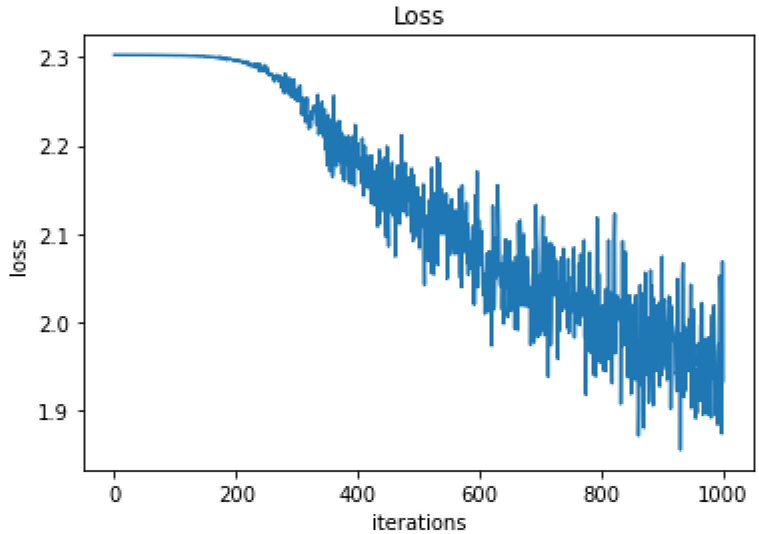
```
stats['train_acc_history']
```

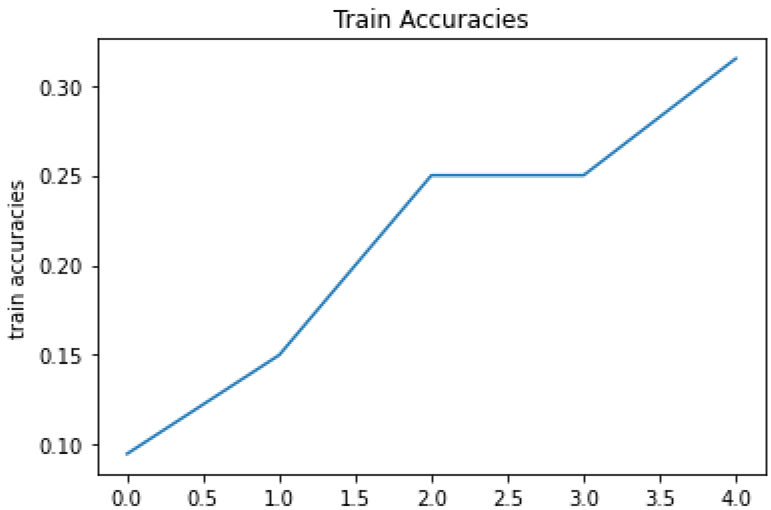
Out[10]:

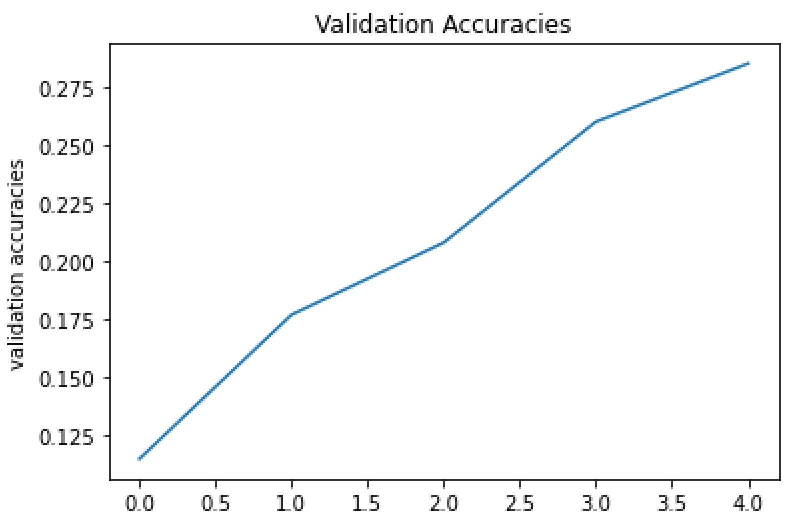
```
[0.095, 0.15, 0.25, 0.25, 0.315]
```

In [11]:

```
# ================================================================ #
# YOUR CODE HERE:
#   Do some debugging to gain some insight into why the optimization
#   isn't great.
# ================================================================ #

# Plot the loss function and train / validation accuracies

plt.plot(stats['loss_history'])
plt.xlabel('iterations')
plt.ylabel('loss')
plt.title('Loss')
plt.show()

plt.plot(stats['train_acc_history'])
plt.ylabel('train accuracies')
plt.title('Train Accuracies')
plt.show()

plt.plot(stats['val_acc_history'])
plt.ylabel('validation accuracies')
plt.title('Validation Accuracies')
plt.show()

# ================================================================ #
# END YOUR CODE HERE
# ================================================================ #
```







Answers:

(1) From the loss plot, the loss didn't decrease at the beginning. Therefore, we can guess that the learning rate is too small. From the two accuracy plots, we can see that the accuracies are sill rising. Hence, we can guess that 1000 iterations may not be enough.

(2) I will increase the learning rate and the number of iterations.

Optimize the neural network

Use the following part of the Jupyter notebook to optimize your hyperparameters on the validation set. Store your nets as best_net.

In [12]:

```
best_net = None # store the best model into this

# ================================================================ #
# YOUR CODE HERE:
#   Optimize over your hyperparameters to arrive at the best neural
#   network.  You should be able to get over 50% validation accuracy.
#   For this part of the notebook, we will give credit based on the
#   accuracy you get.  Your score on this question will be multiplied by:
#      min(floor((X - 28%)) / %22, 1)
#   where if you get 50% or higher validation accuracy, you get full
#   points.
#
#   Note, you need to use the same network structure (keep hidden_size = 50)!
# ================================================================ #

learning_rates = [1e-4, 1e-3, 3e-3, 5e-3, 1e-2, 3e-2, 5e-2, 1e-1]
accuracy = {}

best_learning_rate = 0
best_validation = 0

for learning_rate in learning_rates:
    net = TwoLayerNet(input_size, hidden_size, num_classes)
    stats = net.train(X_train, y_train, X_val, y_val,
            num_iters=3500, batch_size=200,
            learning_rate=learning_rate, learning_rate_decay=0.95,
            reg=0.55, verbose=True)

    y_train_pred = net.predict(X_train)
    train_accuracy = np.mean(np.equal(y_train, y_train_pred))

    y_val_pred = net.predict(X_val)
    val_accuracy = np.mean(np.equal(y_val, y_val_pred))

    accuracy[learning_rate] = (train_accuracy, val_accuracy)

    if best_validation < val_accuracy:
        best_learning_rate = learning_rate
        best_validation = val_accuracy
        best_net = net

for learning_rate in accuracy:
    print("Learning Rate: {}, Train Accuracy: {}, Validation: {}".format(learning_rate, ac

print("\nThe Best Learning Rate: {}\n".format(best_learning_rate))

# ================================================================ #
# END YOUR CODE HERE
# ================================================================ #
val_acc = (best_net.predict(X_val) == y_val).mean()
print('Validation accuracy: ', val_acc)
```

```
iteration 0 / 3500: loss 2.302997401109704
iteration 100 / 3500: loss 2.3025337778391846
iteration 200 / 3500: loss 2.298359174811698
iteration 300 / 3500: loss 2.264620168811341
iteration 400 / 3500: loss 2.2317722560179187
iteration 500 / 3500: loss 2.1473953573120252
iteration 600 / 3500: loss 2.078121763752046
iteration 700 / 3500: loss 2.045704612501719
iteration 800 / 3500: loss 1.965984706144976
```

```
iteration 900 / 3500: loss 2.014275621419568
iteration 1000 / 3500: loss 1.8921575161673971
iteration 1100 / 3500: loss 1.8598477867763987
iteration 1200 / 3500: loss 1.9298727425611475
iteration 1300 / 3500: loss 1.8962509476741416
iteration 1400 / 3500: loss 1.8794426275753993
iteration 1500 / 3500: loss 1.8120413442682877
iteration 1600 / 3500: loss 1.8291193208791063
iteration 1700 / 3500: loss 1.9029972517200149
iteration 1800 / 3500: loss 1.6872365256217474
iteration 1900 / 3500: loss 1.911582176437726
iteration 2000 / 3500: loss 1.7462618983914069
iteration 2100 / 3500: loss 1.757088388209533
iteration 2200 / 3500: loss 1.7543581198803018
iteration 2300 / 3500: loss 1.7144871769537742
iteration 2400 / 3500: loss 1.6471536306229593
iteration 2500 / 3500: loss 1.7821161425180871
iteration 2600 / 3500: loss 1.8573546497173177
iteration 2700 / 3500: loss 1.814129778440546
iteration 2800 / 3500: loss 1.7338615965988042
iteration 2900 / 3500: loss 1.7435264603180365
iteration 3000 / 3500: loss 1.699752449476504
iteration 3100 / 3500: loss 1.8324479342243916
iteration 3200 / 3500: loss 1.7612130378984054
iteration 3300 / 3500: loss 1.7218378406578894
iteration 3400 / 3500: loss 1.6993344012142404
iteration 0 / 3500: loss 2.302986411324343
iteration 100 / 3500: loss 1.9232232372666749
iteration 200 / 3500: loss 1.759347222135508
iteration 300 / 3500: loss 1.756488564219933
iteration 400 / 3500: loss 1.6727663289497905
iteration 500 / 3500: loss 1.633897580429175
iteration 600 / 3500: loss 1.678366978962524
iteration 700 / 3500: loss 1.5908599823897758
iteration 800 / 3500: loss 1.4762522293413398
iteration 900 / 3500: loss 1.53800150752706
iteration 1000 / 3500: loss 1.576294915354204
iteration 1100 / 3500: loss 1.488414553954279
iteration 1200 / 3500: loss 1.3870829853306392
iteration 1300 / 3500: loss 1.5789796596586665
iteration 1400 / 3500: loss 1.4333352374069301
iteration 1500 / 3500: loss 1.3348892324102062
iteration 1600 / 3500: loss 1.4882764475636145
iteration 1700 / 3500: loss 1.5386421917250106
iteration 1800 / 3500: loss 1.4865942997101596
iteration 1900 / 3500: loss 1.5243896200801552
iteration 2000 / 3500: loss 1.4723743667183031
iteration 2100 / 3500: loss 1.5392486542155455
iteration 2200 / 3500: loss 1.446332157494875
iteration 2300 / 3500: loss 1.4651521029458334
iteration 2400 / 3500: loss 1.4213970746406865
iteration 2500 / 3500: loss 1.5286906235240358
iteration 2600 / 3500: loss 1.4881782566544015
iteration 2700 / 3500: loss 1.3928417400520017
iteration 2800 / 3500: loss 1.3581560073462011
iteration 2900 / 3500: loss 1.3760299236515272
iteration 3000 / 3500: loss 1.3836823654682613
iteration 3100 / 3500: loss 1.3123833713906103
iteration 3200 / 3500: loss 1.462576568868863
iteration 3300 / 3500: loss 1.3868367750092365
iteration 3400 / 3500: loss 1.5591674124610215
iteration 0 / 3500: loss 2.303014357893828
iteration 100 / 3500: loss 1.7068276303772936
iteration 200 / 3500: loss 1.7414195800082797
iteration 300 / 3500: loss 1.7519837577366806
iteration 400 / 3500: loss 1.6578147651681054
```

```
iteration 500 / 3500: loss 1.7700501946492917
iteration 600 / 3500: loss 1.6070792781069048
iteration 700 / 3500: loss 1.7733688800892693
iteration 800 / 3500: loss 1.7319000906465882
iteration 900 / 3500: loss 1.5989403728534124
iteration 1000 / 3500: loss 1.5344029574444085
iteration 1100 / 3500: loss 1.9133985971316425
iteration 1200 / 3500: loss 1.68329616121571
iteration 1300 / 3500: loss 1.464549162163786
iteration 1400 / 3500: loss 1.5971767049298458
iteration 1500 / 3500: loss 1.5251982076559232
iteration 1600 / 3500: loss 1.5783437880344389
iteration 1700 / 3500: loss 1.5252112516453449
iteration 1800 / 3500: loss 1.5204179087252612
iteration 1900 / 3500: loss 1.6369273560697226
iteration 2000 / 3500: loss 1.7068115558302446
iteration 2100 / 3500: loss 1.6126205076589328
iteration 2200 / 3500: loss 1.582731622426254
iteration 2300 / 3500: loss 1.5754925872045902
iteration 2400 / 3500: loss 1.701496290072023
iteration 2500 / 3500: loss 1.526950587945974
iteration 2600 / 3500: loss 1.5014990142459348
iteration 2700 / 3500: loss 1.5347069926852663
iteration 2800 / 3500: loss 1.6185106216349627
iteration 2900 / 3500: loss 1.5708537119911778
iteration 3000 / 3500: loss 1.5657797880625761
iteration 3100 / 3500: loss 1.557372436470082
iteration 3200 / 3500: loss 1.307196726686845
iteration 3300 / 3500: loss 1.3896861393043467
iteration 3400 / 3500: loss 1.4343703894886384
iteration 0 / 3500: loss 2.303012541137858
/Users/jacky/My_Data/Data/UCLA/2022_Winter/ECE C247_Deep Learning/Homework/HW3/hw3-code/nn
dl/neural_net.py:115: RuntimeWarning: divide by zero encountered in log
  probs_log = -np.log(probs_row)
iteration 100 / 3500: loss inf
iteration 200 / 3500: loss inf
iteration 300 / 3500: loss inf
iteration 400 / 3500: loss inf
iteration 500 / 3500: loss inf
iteration 600 / 3500: loss inf
iteration 700 / 3500: loss inf
iteration 800 / 3500: loss inf
iteration 900 / 3500: loss inf
/Users/jacky/My_Data/Data/UCLA/2022_Winter/ECE C247_Deep Learning/Homework/HW3/hw3-code/nn
dl/neural_net.py:110: RuntimeWarning: overflow encountered in subtract
  scores -= np.max(scores, axis=1, keepdims=True)
/Users/jacky/My_Data/Data/UCLA/2022_Winter/ECE C247_Deep Learning/Homework/HW3/hw3-code/nn
dl/neural_net.py:110: RuntimeWarning: invalid value encountered in subtract
  scores -= np.max(scores, axis=1, keepdims=True)
iteration 1000 / 3500: loss nan
iteration 1100 / 3500: loss nan
iteration 1200 / 3500: loss nan
iteration 1300 / 3500: loss nan
iteration 1400 / 3500: loss nan
iteration 1500 / 3500: loss nan
iteration 1600 / 3500: loss nan
iteration 1700 / 3500: loss nan
iteration 1800 / 3500: loss nan
iteration 1900 / 3500: loss nan
iteration 2000 / 3500: loss nan
iteration 2100 / 3500: loss nan
iteration 2200 / 3500: loss nan
iteration 2300 / 3500: loss nan
iteration 2400 / 3500: loss nan
iteration 2500 / 3500: loss nan
```

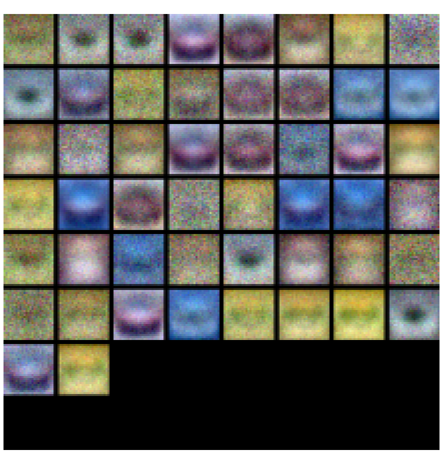
```
iteration 2600 / 3500: loss nan
iteration 2700 / 3500: loss nan
iteration 2800 / 3500: loss nan
iteration 2900 / 3500: loss nan
iteration 3000 / 3500: loss nan
iteration 3100 / 3500: loss nan
iteration 3200 / 3500: loss nan
iteration 3300 / 3500: loss nan
iteration 3400 / 3500: loss nan
iteration 0 / 3500: loss 2.303024951829114
iteration 100 / 3500: loss inf
iteration 200 / 3500: loss inf
iteration 300 / 3500: loss inf
iteration 400 / 3500: loss nan
iteration 500 / 3500: loss nan
iteration 600 / 3500: loss nan
iteration 700 / 3500: loss nan
iteration 800 / 3500: loss nan
iteration 900 / 3500: loss nan
iteration 1000 / 3500: loss nan
iteration 1100 / 3500: loss nan
iteration 1200 / 3500: loss nan
iteration 1300 / 3500: loss nan
iteration 1400 / 3500: loss nan
iteration 1500 / 3500: loss nan
iteration 1600 / 3500: loss nan
iteration 1700 / 3500: loss nan
iteration 1800 / 3500: loss nan
iteration 1900 / 3500: loss nan
iteration 2000 / 3500: loss nan
iteration 2100 / 3500: loss nan
iteration 2200 / 3500: loss nan
iteration 2300 / 3500: loss nan
iteration 2400 / 3500: loss nan
iteration 2500 / 3500: loss nan
iteration 2600 / 3500: loss nan
iteration 2700 / 3500: loss nan
iteration 2800 / 3500: loss nan
iteration 2900 / 3500: loss nan
iteration 3000 / 3500: loss nan
iteration 3100 / 3500: loss nan
iteration 3200 / 3500: loss nan
iteration 3300 / 3500: loss nan
iteration 3400 / 3500: loss nan
iteration 0 / 3500: loss 2.303007161998417
iteration 100 / 3500: loss inf
iteration 200 / 3500: loss nan
iteration 300 / 3500: loss nan
iteration 400 / 3500: loss nan
iteration 500 / 3500: loss nan
iteration 600 / 3500: loss nan
iteration 700 / 3500: loss nan
iteration 800 / 3500: loss nan
iteration 900 / 3500: loss nan
iteration 1000 / 3500: loss nan
iteration 1100 / 3500: loss nan
iteration 1200 / 3500: loss nan
iteration 1300 / 3500: loss nan
iteration 1400 / 3500: loss nan
iteration 1500 / 3500: loss nan
iteration 1600 / 3500: loss nan
iteration 1700 / 3500: loss nan
iteration 1800 / 3500: loss nan
iteration 1900 / 3500: loss nan
iteration 2000 / 3500: loss nan
iteration 2100 / 3500: loss nan
```

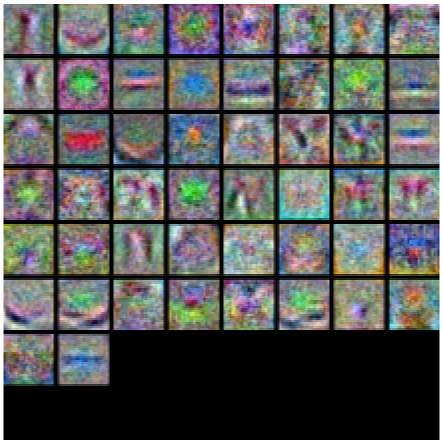
```
iteration 2200 / 3500: loss nan
iteration 2300 / 3500: loss nan
iteration 2400 / 3500: loss nan
iteration 2500 / 3500: loss nan
iteration 2600 / 3500: loss nan
iteration 2700 / 3500: loss nan
iteration 2800 / 3500: loss nan
iteration 2900 / 3500: loss nan
iteration 3000 / 3500: loss nan
iteration 3100 / 3500: loss nan
iteration 3200 / 3500: loss nan
iteration 3300 / 3500: loss nan
iteration 3400 / 3500: loss nan
iteration 0 / 3500: loss 2.3030106917108073
iteration 100 / 3500: loss inf
iteration 200 / 3500: loss nan
iteration 300 / 3500: loss nan
iteration 400 / 3500: loss nan
iteration 500 / 3500: loss nan
iteration 600 / 3500: loss nan
iteration 700 / 3500: loss nan
iteration 800 / 3500: loss nan
iteration 900 / 3500: loss nan
iteration 1000 / 3500: loss nan
iteration 1100 / 3500: loss nan
iteration 1200 / 3500: loss nan
iteration 1300 / 3500: loss nan
iteration 1400 / 3500: loss nan
iteration 1500 / 3500: loss nan
iteration 1600 / 3500: loss nan
iteration 1700 / 3500: loss nan
iteration 1800 / 3500: loss nan
iteration 1900 / 3500: loss nan
iteration 2000 / 3500: loss nan
iteration 2100 / 3500: loss nan
iteration 2200 / 3500: loss nan
iteration 2300 / 3500: loss nan
iteration 2400 / 3500: loss nan
iteration 2500 / 3500: loss nan
iteration 2600 / 3500: loss nan
iteration 2700 / 3500: loss nan
iteration 2800 / 3500: loss nan
iteration 2900 / 3500: loss nan
iteration 3000 / 3500: loss nan
iteration 3100 / 3500: loss nan
iteration 3200 / 3500: loss nan
iteration 3300 / 3500: loss nan
iteration 3400 / 3500: loss nan
iteration 0 / 3500: loss 2.3030054797882737
iteration 100 / 3500: loss inf
iteration 200 / 3500: loss nan
iteration 300 / 3500: loss nan
iteration 400 / 3500: loss nan
iteration 500 / 3500: loss nan
iteration 600 / 3500: loss nan
iteration 700 / 3500: loss nan
iteration 800 / 3500: loss nan
iteration 900 / 3500: loss nan
iteration 1000 / 3500: loss nan
iteration 1100 / 3500: loss nan
iteration 1200 / 3500: loss nan
iteration 1300 / 3500: loss nan
iteration 1400 / 3500: loss nan
iteration 1500 / 3500: loss nan
iteration 1600 / 3500: loss nan
iteration 1700 / 3500: loss nan
```

```
iteration 1800 / 3500: loss nan
iteration 1900 / 3500: loss nan
iteration 2000 / 3500: loss nan
iteration 2100 / 3500: loss nan
iteration 2200 / 3500: loss nan
iteration 2300 / 3500: loss nan
iteration 2400 / 3500: loss nan
iteration 2500 / 3500: loss nan
iteration 2600 / 3500: loss nan
iteration 2700 / 3500: loss nan
iteration 2800 / 3500: loss nan
iteration 2900 / 3500: loss nan
iteration 3000 / 3500: loss nan
iteration 3100 / 3500: loss nan
iteration 3200 / 3500: loss nan
iteration 3300 / 3500: loss nan
iteration 3400 / 3500: loss nan
Learning Rate: 0.0001, Train Accuracy: 0.397734693877551, Validation: 0.389
Learning Rate: 0.001, Train Accuracy: 0.5489387755102041, Validation: 0.496
Learning Rate: 0.003, Train Accuracy: 0.5375918367346939, Validation: 0.502
Learning Rate: 0.005, Train Accuracy: 0.10026530612244898, Validation: 0.087
Learning Rate: 0.01, Train Accuracy: 0.10026530612244898, Validation: 0.087
Learning Rate: 0.03, Train Accuracy: 0.10026530612244898, Validation: 0.087
Learning Rate: 0.05, Train Accuracy: 0.10026530612244898, Validation: 0.087
Learning Rate: 0.1, Train Accuracy: 0.10026530612244898, Validation: 0.087

The Best Learning Rate: 0.003

Validation accuracy:  0.502
```

In [13]:

```python
from utils.vis_utils import visualize_grid

# Visualize the weights of the network

def show_net_weights(net):
    W1 = net.params['W1']
    W1 = W1.T.reshape(32, 32, 3, -1).transpose(3, 0, 1, 2)
    plt.imshow(visualize_grid(W1, padding=3).astype('uint8'))
    plt.gca().axis('off')
    plt.show()

show_net_weights(subopt_net)
show_net_weights(best_net)
```





Question:

(1) What differences do you see in the weights between the suboptimal net and the best net you arrived at?

Answer:

(1) The images in the suboptimal net looks alike, and they contain much more noises. However, we can distinguish the differences between the images in the best net.

Evaluate on test set

In [14]:

```
test_acc = (best_net.predict(X_test) == y_test).mean()
print('Test accuracy: ', test_acc)
```

```
Test accuracy:  0.513
```

```python
import numpy as np
import matplotlib.pyplot as plt


class TwoLayerNet(object):
  """
  A two-layer fully-connected neural network. The net has an input dimension of
  N, a hidden layer dimension of H, and performs classification over C classes.
  We train the network with a softmax loss function and L2 regularization on the
  weight matrices. The network uses a ReLU nonlinearity after the first fully
  connected layer.

  In other words, the network has the following architecture:

  input - fully connected layer - ReLU - fully connected layer - softmax

  The outputs of the second fully-connected layer are the scores for each class.
  """

  def __init__(self, input_size, hidden_size, output_size, std=1e-4):
    """
    Initialize the model. Weights are initialized to small random values and
    biases are initialized to zero. Weights and biases are stored in the
    variable self.params, which is a dictionary with the following keys:

    W1: First layer weights; has shape (H, D)
    b1: First layer biases; has shape (H,)
    W2: Second layer weights; has shape (C, H)
    b2: Second layer biases; has shape (C,)

    Inputs:
    - input_size: The dimension D of the input data.
    - hidden_size: The number of neurons H in the hidden layer.
    - output_size: The number of classes C.
    """
    self.params = {}
    self.params['W1'] = std * np.random.randn(hidden_size, input_size)
    self.params['b1'] = np.zeros(hidden_size)
    self.params['W2'] = std * np.random.randn(output_size, hidden_size)
    self.params['b2'] = np.zeros(output_size)


  def loss(self, X, y=None, reg=0.0):
    """
    Compute the loss and gradients for a two layer fully connected neural
    network.

    Inputs:
    - X: Input data of shape (N, D). Each X[i] is a training sample.
    - y: Vector of training labels. y[i] is the label for X[i], and each y[i] is
      an integer in the range 0 <= y[i] < C. This parameter is optional; if it
      is not passed then we only return scores, and if it is passed then we
      instead return the loss and gradients.
```

```
    - reg: Regularization strength.

    Returns:
    If y is None, return a matrix scores of shape (N, C) where scores[i, c] is
    the score for class c on input X[i].

    If y is not None, instead return a tuple of:
    - loss: Loss (data loss and regularization loss) for this batch of training
      samples.
    - grads: Dictionary mapping parameter names to gradients of those parameters
      with respect to the loss function; has the same keys as self.params.
    """
    # Unpack variables from the params dictionary
    W1, b1 = self.params['W1'], self.params['b1']
    W2, b2 = self.params['W2'], self.params['b2']
    N, D = X.shape

    # Compute the forward pass
    scores = None

    # ================================================================ #
    # YOUR CODE HERE:
    #   Calculate the output scores of the neural network.  The result
    #   should be (N, C). As stated in the description for this class,
    #   there should not be a ReLU layer after the second FC layer.
    #   The output of the second FC layer is the output scores. Do not
    #   use a for loop in your implementation.
    # ================================================================ #

    relu = lambda x: np.maximum(x, 0)

    h1 = np.dot(X, W1.T) + b1
    scores = np.dot(relu(h1), W2.T) + b2

    # ================================================================ #
    # END YOUR CODE HERE
    # ================================================================ #


    # If the targets are not given then jump out, we're done
    if y is None:
      return scores

    # Compute the loss
    loss = None

    # ================================================================ #
    # YOUR CODE HERE:
    #   Calculate the loss of the neural network.  This includes the
    #   softmax loss and the L2 regularization for W1 and W2. Store the
    #   total loss in teh variable loss.  Multiply the regularization
    #   loss by 0.5 (in addition to the factor reg).
    # ================================================================ #

    # scores is num_examples by num_classes

    scores -= np.max(scores, axis=1, keepdims=True)
```

```
    scores_exp = np.exp(scores)

    probs = scores_exp / np.sum(scores_exp, axis=1, keepdims=True)
    probs_row = probs[range(N), y]
    probs_log = -np.log(probs_row)
    softmax_loss = np.sum(probs_log) / N

    reg_loss = 0.5 * reg * (np.linalg.norm(W1, 'fro')**2 + np.linalg.norm(W2, 'fro')**2)

    loss = softmax_loss + reg_loss

    # ================================================================ #
    # END YOUR CODE HERE
    # ================================================================ #

    grads = {}

    # ================================================================ #
    # YOUR CODE HERE:
    #   Implement the backward pass.  Compute the derivatives of the
    #   weights and the biases.  Store the results in the grads
    #   dictionary.  e.g., grads['W1'] should store the gradient for
    #   W1, and be of the same size as W1.
    # ================================================================ #

    probs[range(N), y] -= 1

    dLdb = probs / N
    dLdW2 = np.maximum(np.dot(W1, X.T)+b1.reshape([W1.shape[0], 1]), 0)

    grads['W2'] = np.dot(dLdb.T, dLdW2.T) + reg * W2
    grads['b2'] = np.sum(dLdb, axis=0, keepdims=True)

    dbdh = W2.T
    dLda = np.dot(dbdh, dLdb.T) * (np.dot(W1, X.T) > 0)

    grads['W1'] = np.dot(dLda, X) + reg * W1
    grads['b1'] = np.sum(dLda, axis=1, keepdims=True).T

    # ================================================================ #
    # END YOUR CODE HERE
    # ================================================================ #

    return loss, grads

  def train(self, X, y, X_val, y_val,
            learning_rate=1e-3, learning_rate_decay=0.95,
            reg=1e-5, num_iters=100,
            batch_size=200, verbose=False):
    """
    Train this neural network using stochastic gradient descent.

    Inputs:
    - X: A numpy array of shape (N, D) giving training data.
    - y: A numpy array f shape (N,) giving training labels; y[i] = c means that
      X[i] has label c, where 0 <= c < C.
    - X_val: A numpy array of shape (N_val, D) giving validation data.
    - y_val: A numpy array of shape (N_val,) giving validation labels.
```

```
    - learning_rate: Scalar giving learning rate for optimization.
    - learning_rate_decay: Scalar giving factor used to decay the learning rate
      after each epoch.
    - reg: Scalar giving regularization strength.
    - num_iters: Number of steps to take when optimizing.
    - batch_size: Number of training examples to use per step.
    - verbose: boolean; if true print progress during optimization.
    """
    num_train = X.shape[0]
    iterations_per_epoch = max(num_train / batch_size, 1)

    # Use SGD to optimize the parameters in self.model
    loss_history = []
    train_acc_history = []
    val_acc_history = []

    for it in np.arange(num_iters):
      X_batch = None
      y_batch = None

      # ================================================================ #
      # YOUR CODE HERE:
      #   Create a minibatch by sampling batch_size samples randomly.
      # ================================================================ #

      idx = np.random.choice(num_train, batch_size)
      X_batch = X[idx]
      y_batch = y[idx]

      # ================================================================ #
      # END YOUR CODE HERE
      # ================================================================ #

       # Compute loss and gradients using the current minibatch
      loss, grads = self.loss(X_batch, y=y_batch, reg=reg)
      loss_history.append(loss)

      # ================================================================ #
      # YOUR CODE HERE:
      #   Perform a gradient descent step using the minibatch to update
      #   all parameters (i.e., W1, W2, b1, and b2).
      # ================================================================ #

      self.params['W1'] = self.params['W1'] - learning_rate * grads['W1']
      self.params['b1'] = self.params['b1'] - learning_rate * grads['b1']

      self.params['W2'] = self.params['W2'] - learning_rate * grads['W2']
      self.params['b2'] = self.params['b2'] - learning_rate * grads['b2']

      # ================================================================ #
      # END YOUR CODE HERE
      # ================================================================ #

      if verbose and it % 100 == 0:
        print('iteration {} / {}: loss {}'.format(it, num_iters, loss))

      # Every epoch, check train and val accuracy and decay learning rate.
      if it % iterations_per_epoch == 0:
        # Check accuracy
```

```python
        train_acc = (self.predict(X_batch) == y_batch).mean()
        val_acc = (self.predict(X_val) == y_val).mean()
        train_acc_history.append(train_acc)
        val_acc_history.append(val_acc)

        # Decay learning rate
        learning_rate *= learning_rate_decay

    return {
      'loss_history': loss_history,
      'train_acc_history': train_acc_history,
      'val_acc_history': val_acc_history,
    }

  def predict(self, X):
    """
    Use the trained weights of this two-layer network to predict labels for
    data points. For each data point we predict scores for each of the C
    classes, and assign each data point to the class with the highest score.

    Inputs:
    - X: A numpy array of shape (N, D) giving N D-dimensional data points to
      classify.

    Returns:
    - y_pred: A numpy array of shape (N,) giving predicted labels for each of
      the elements of X. For all i, y_pred[i] = c means that X[i] is predicted
      to have class c, where 0 <= c < C.
    """
    y_pred = None

    # ================================================================ #
    # YOUR CODE HERE:
    #   Predict the class given the input data.
    # ================================================================ #

    relu = lambda x: np.maximum(x, 0)

    h1 = np.dot(X, self.params['W1'].T) + self.params['b1']
    scores = np.dot(relu(h1), self.params['W2'].T) + self.params['b2']

    y_pred = np.argmax(scores, axis=1)

    # ================================================================ #
    # END YOUR CODE HERE
    # ================================================================ #

    return y_pred

```

Fully connected networks

In the previous notebook, you implemented a simple two-layer neural network class. However, this class is not modular. If you wanted to change the number of layers, you would need to write a new loss and gradient function. If you wanted to optimize the network with different optimizers, you'd need to write new training functions. If you wanted to incorporate regularizations, you'd have to modify the loss and gradient function.

Instead of having to modify functions each time, for the rest of the class, we'll work in a more modular framework where we define forward and backward layers that calculate losses and gradients respectively. Since the forward and backward layers share intermediate values that are useful for calculating both the loss and the gradient, we'll also have these function return "caches" which store useful intermediate values.

The goal is that through this modular design, we can build different sized neural networks for various applications.

In this HW #3, we'll define the basic architecture, and in HW #4, we'll build on this framework to implement different optimizers and regularizations (like BatchNorm and Dropout).

Modular layers

This notebook will build modular layers in the following manner. First, there will be a forward pass for a given layer with inputs (`x`) and return the output of that layer (`out`) as well as cached variables (`cache`) that will be used to calculate the gradient in the backward pass.

```
def layer_forward(x, w):
  """ Receive inputs x and weights w """
  # Do some computations ...
  z = # ... some intermediate value
  # Do some more computations ...
  out = # the output

  cache = (x, w, z, out) # Values we need to compute gradients

  return out, cache
```

The backward pass will receive upstream derivatives and the `cache` object, and will return gradients with respect to the inputs and weights, like this:

```
def layer_backward(dout, cache):
  """
  Receive derivative of loss with respect to outputs and cache,
  and compute derivative with respect to inputs.
  """
  # Unpack cache values
  x, w, z, out = cache

  # Use values in cache to compute derivatives
  dx = # Derivative of loss with respect to x
  dw = # Derivative of loss with respect to w

  return dx, dw
```

In [1]:

```
## Import and setups
```

```python
import time
import numpy as np
import matplotlib.pyplot as plt
from nndl.fc_net import *
from utils.data_utils import get_CIFAR10_data
from utils.gradient_check import eval_numerical_gradient, eval_numerical_gradient_array
from utils.solver import Solver

%matplotlib inline
plt.rcParams['figure.figsize'] = (10.0, 8.0) # set default size of plots
plt.rcParams['image.interpolation'] = 'nearest'
plt.rcParams['image.cmap'] = 'gray'

# for auto-reloading external modules
# see http://stackoverflow.com/questions/1907993/autoreload-of-modules-in-ipython
%load_ext autoreload
%autoreload 2

def rel_error(x, y):
  """ returns relative error """
  return np.max(np.abs(x - y) / (np.maximum(1e-8, np.abs(x) + np.abs(y))))
```

In [2]:

```python
# Load the (preprocessed) CIFAR10 data.

data = get_CIFAR10_data()
for k in data.keys():
  print('{}: {} '.format(k, data[k].shape))
```

```
X_train: (49000, 3, 32, 32)
y_train: (49000,)
X_val: (1000, 3, 32, 32)
y_val: (1000,)
X_test: (1000, 3, 32, 32)
y_test: (1000,)
```

Linear layers

In this section, we'll implement the forward and backward pass for the linear layers.

The linear layer forward pass is the function `affine_forward` in `nndl/layers.py` and the backward pass is `affine_backward` .

After you have implemented these, test your implementation by running the cell below.

Affine layer forward pass

Implement `affine_forward` and then test your code by running the following cell.

In [3]:

```python
# Test the affine_forward function

num_inputs = 2
input_shape = (4, 5, 6)
output_dim = 3

input_size = num_inputs * np.prod(input_shape)
weight_size = output_dim * np.prod(input_shape)

x = np.linspace(-0.1, 0.5, num=input_size).reshape(num_inputs, *input_shape)
w = np.linspace(-0.2, 0.3, num=weight_size).reshape(np.prod(input_shape), output_dim)
```

```
b = np.linspace(-0.3, 0.1, num=output_dim)

out, _ = affine_forward(x, w, b)
correct_out = np.array([[ 1.49834967,  1.70660132,  1.91485297],
                        [ 3.25553199,  3.5141327,   3.77273342]])

# Compare your output with ours. The error should be around 1e-9.
print('Testing affine_forward function:')
print('difference: {}'.format(rel_error(out, correct_out)))
```

```
Testing affine_forward function:
difference: 9.769849468192957e-10
```

Affine layer backward pass

Implement `affine_backward` and then test your code by running the following cell.

In [4]:

```
# Test the affine_backward function

x = np.random.randn(10, 2, 3)
w = np.random.randn(6, 5)
b = np.random.randn(5)
dout = np.random.randn(10, 5)

dx_num = eval_numerical_gradient_array(lambda x: affine_forward(x, w, b)[0], x, dout)
dw_num = eval_numerical_gradient_array(lambda w: affine_forward(x, w, b)[0], w, dout)
db_num = eval_numerical_gradient_array(lambda b: affine_forward(x, w, b)[0], b, dout)

_, cache = affine_forward(x, w, b)
dx, dw, db = affine_backward(dout, cache)

# The error should be around 1e-10
print('Testing affine_backward function:')
print('dx error: {}'.format(rel_error(dx_num, dx)))
print('dw error: {}'.format(rel_error(dw_num, dw)))
print('db error: {}'.format(rel_error(db_num, db)))
```

```
Testing affine_backward function:
dx error: 4.429582091703194e-10
dw error: 1.7674091675743669e-10
db error: 3.275488993710175e-12
```

Activation layers

In this section you'll implement the ReLU activation.

ReLU forward pass

Implement the `relu_forward` function in `nndl/layers.py` and then test your code by running the following cell.

In [5]:

```
# Test the relu_forward function

x = np.linspace(-0.5, 0.5, num=12).reshape(3, 4)

out, _ = relu_forward(x)
correct_out = np.array([[ 0.,          0.,          0.,          0.,         ],
                        [ 0.,          0.,          0.04545455,  0.13636364,],
                        [ 0.22727273,  0.31818182,  0.40909091,  0.5,       ]])

# Compare your output with ours. The error should be around 1e-8
```

```
print('Testing relu_forward function:')
print('difference: {}'.format(rel_error(out, correct_out)))
```

```
Testing relu_forward function:
difference: 4.999999798022158e-08
```

ReLU backward pass

Implement the `relu_backward` function in `nndl/layers.py` and then test your code by running the following cell.

In [6]:
```
x = np.random.randn(10, 10)
dout = np.random.randn(*x.shape)

dx_num = eval_numerical_gradient_array(lambda x: relu_forward(x)[0], x, dout)

_, cache = relu_forward(x)
dx = relu_backward(dout, cache)

# The error should be around 1e-12
print('Testing relu_backward function:')
print('dx error: {}'.format(rel_error(dx_num, dx)))
```

```
Testing relu_backward function:
dx error: 3.2756320859180464e-12
```

Combining the affine and ReLU layers

Often times, an affine layer will be followed by a ReLU layer. So let's make one that puts them together. Layers that are combined are stored in `nndl/layer_utils.py` .

Affine-ReLU layers

We've implemented `affine_relu_forward()` and `affine_relu_backward` in `nndl/layer_utils.py` . Take a look at them to make sure you understand what's going on. Then run the following cell to ensure its implemented correctly.

In [7]:
```
from nndl.layer_utils import affine_relu_forward, affine_relu_backward

x = np.random.randn(2, 3, 4)
w = np.random.randn(12, 10)
b = np.random.randn(10)
dout = np.random.randn(2, 10)

out, cache = affine_relu_forward(x, w, b)
dx, dw, db = affine_relu_backward(dout, cache)

dx_num = eval_numerical_gradient_array(lambda x: affine_relu_forward(x, w, b)[0], x, dout)
dw_num = eval_numerical_gradient_array(lambda w: affine_relu_forward(x, w, b)[0], w, dout)
db_num = eval_numerical_gradient_array(lambda b: affine_relu_forward(x, w, b)[0], b, dout)

print('Testing affine_relu_forward and affine_relu_backward:')
print('dx error: {}'.format(rel_error(dx_num, dx)))
print('dw error: {}'.format(rel_error(dw_num, dw)))
print('db error: {}'.format(rel_error(db_num, db)))
```

```
Testing affine_relu_forward and affine_relu_backward:
dx error: 2.70400194212495e-10
```

```
dw error: 2.821713532998111e-10
db error: 8.736539594688683e-11
```

Softmax losses

You've already implemented it, so we have written it in `layers.py`. The following code will ensure its working correctly.

In [8]:
```python
num_classes, num_inputs = 10, 50
x = 0.001 * np.random.randn(num_inputs, num_classes)
y = np.random.randint(num_classes, size=num_inputs)



dx_num = eval_numerical_gradient(lambda x: softmax_loss(x, y)[0], x, verbose=False)
loss, dx = softmax_loss(x, y)

# Test softmax_loss function. Loss should be 2.3 and dx error should be 1e-8
print('\nTesting softmax_loss:')
print('loss: {}'.format(loss))
print('dx error: {}'.format(rel_error(dx_num, dx)))
```

```
Testing softmax_loss:
loss: 2.302442595204509
dx error: 7.794705234130109e-09
```

Implementation of a two-layer NN

In `nndl/fc_net.py`, implement the class `TwoLayerNet` which uses the layers you made here. When you have finished, the following cell will test your implementation.

In [9]:
```python
N, D, H, C = 3, 5, 50, 7
X = np.random.randn(N, D)
y = np.random.randint(C, size=N)

std = 1e-2
model = TwoLayerNet(input_dim=D, hidden_dims=H, num_classes=C, weight_scale=std)

print('Testing initialization ... ')
W1_std = abs(model.params['W1'].std() - std)
b1 = model.params['b1']
W2_std = abs(model.params['W2'].std() - std)
b2 = model.params['b2']
assert W1_std < std / 10, 'First layer weights do not seem right'
assert np.all(b1 == 0), 'First layer biases do not seem right'
assert W2_std < std / 10, 'Second layer weights do not seem right'
assert np.all(b2 == 0), 'Second layer biases do not seem right'

print('Testing test-time forward pass ... ')
model.params['W1'] = np.linspace(-0.7, 0.3, num=D*H).reshape(D, H)
model.params['b1'] = np.linspace(-0.1, 0.9, num=H)
model.params['W2'] = np.linspace(-0.3, 0.4, num=H*C).reshape(H, C)
model.params['b2'] = np.linspace(-0.9, 0.1, num=C)
X = np.linspace(-5.5, 4.5, num=N*D).reshape(D, N).T
scores = model.loss(X)
correct_scores = np.asarray(
  [[11.53165108,  12.2917344,   13.05181771,  13.81190102,  14.57198434, 15.33206765,  16.
   [12.05769098,  12.74614105,  13.43459113,  14.1230412,   14.81149128, 15.49994135,  16.
   [12.58373087,  13.20054771,  13.81736455,  14.43418138,  15.05099822, 15.66781506,  16.
scores_diff = np.abs(scores - correct_scores).sum()
assert scores_diff < 1e-6, 'Problem with test-time forward pass'
```

```
print('Testing training loss (no regularization)')
y = np.asarray([0, 5, 1])
loss, grads = model.loss(X, y)
correct_loss = 3.4702243556
assert abs(loss - correct_loss) < 1e-10, 'Problem with training-time loss'

model.reg = 1.0
loss, grads = model.loss(X, y)
correct_loss = 26.5948426952
assert abs(loss - correct_loss) < 1e-10, 'Problem with regularization loss'

for reg in [0.0, 0.7]:
  print('Running numeric gradient check with reg = {}'.format(reg))
  model.reg = reg
  loss, grads = model.loss(X, y)

  for name in sorted(grads):
    f = lambda _: model.loss(X, y)[0]
    grad_num = eval_numerical_gradient(f, model.params[name], verbose=False)
    print('{} relative error: {}'.format(name, rel_error(grad_num, grads[name])))
```

```
Testing initialization ...
Testing test-time forward pass ...
Testing training loss (no regularization)
Running numeric gradient check with reg = 0.0
W1 relative error: 1.521570416306979e-08
W2 relative error: 3.2068321167375225e-10
b1 relative error: 8.368200428642256e-09
b2 relative error: 4.3291360264321544e-10
Running numeric gradient check with reg = 0.7
W1 relative error: 2.527915175868136e-07
W2 relative error: 2.8508510893102143e-08
b1 relative error: 1.564680516145745e-08
b2 relative error: 7.759095355706557e-10
```

Solver

We will now use the utils Solver class to train these networks. Familiarize yourself with the API in `utils/solver.py` . After you have done so, declare an instance of a TwoLayerNet with 200 units and then train it with the Solver. Choose parameters so that your validation accuracy is at least 50%.

In [11]:

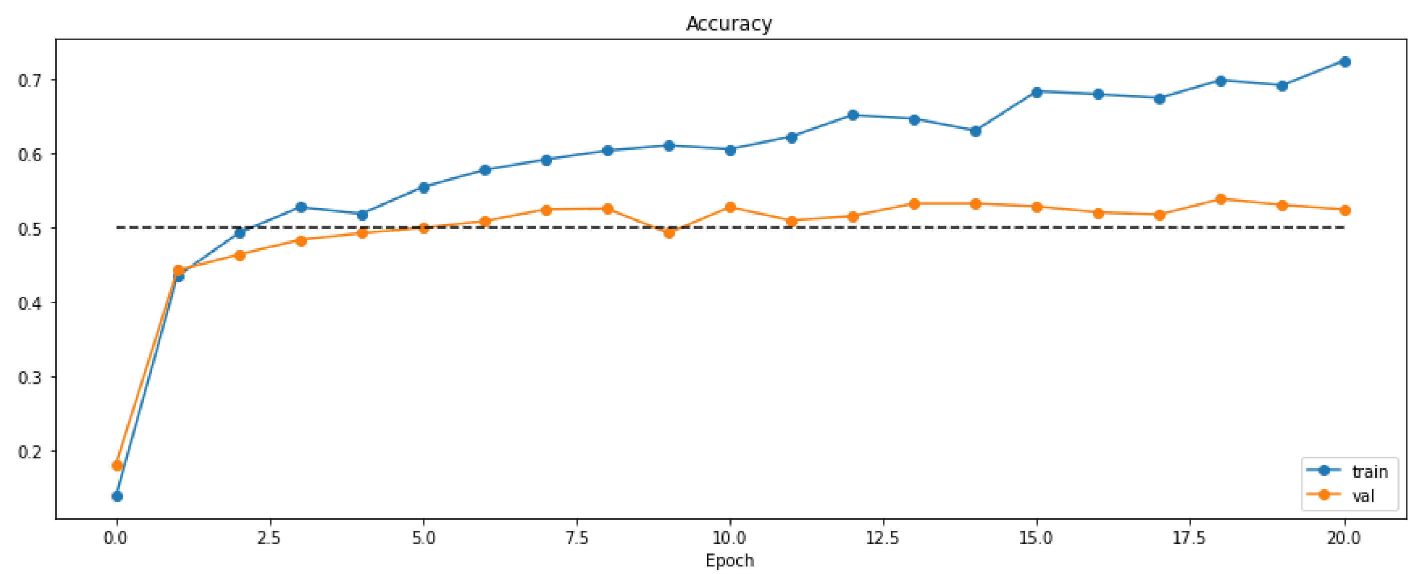
```
model = TwoLayerNet()
solver = None

# ================================================================ #
# YOUR CODE HERE:
#   Declare an instance of a TwoLayerNet and then train
#   it with the Solver. Choose hyperparameters so that your validation
#   accuracy is at least 50%.  We won't have you optimize this further
#   since you did it in the previous notebook.
#
# ================================================================ #
model = TwoLayerNet(hidden_dims = 200)

solver = Solver(model=model, data=data, update_rule='sgd', optim_config={'learning_rate':
                lr_decay=0.95, batch_size=200, num_epochs=20, print_every=100)

solver.train()

# ================================================================ #
# END YOUR CODE HERE
# ================================================================ #
```

```
(Iteration 1 / 4900) loss: 2.306057
(Epoch 0 / 20) train acc: 0.139000; val_acc: 0.182000
(Iteration 101 / 4900) loss: 1.684505
(Iteration 201 / 4900) loss: 1.551681
(Epoch 1 / 20) train acc: 0.435000; val_acc: 0.443000
(Iteration 301 / 4900) loss: 1.488144
(Iteration 401 / 4900) loss: 1.509027
(Epoch 2 / 20) train acc: 0.493000; val_acc: 0.464000
(Iteration 501 / 4900) loss: 1.312502
(Iteration 601 / 4900) loss: 1.423017
(Iteration 701 / 4900) loss: 1.441494
(Epoch 3 / 20) train acc: 0.528000; val_acc: 0.484000
(Iteration 801 / 4900) loss: 1.445859
(Iteration 901 / 4900) loss: 1.347744
(Epoch 4 / 20) train acc: 0.519000; val_acc: 0.493000
(Iteration 1001 / 4900) loss: 1.216793
(Iteration 1101 / 4900) loss: 1.396402
(Iteration 1201 / 4900) loss: 1.276835
(Epoch 5 / 20) train acc: 0.555000; val_acc: 0.500000
(Iteration 1301 / 4900) loss: 1.380393
(Iteration 1401 / 4900) loss: 1.277726
(Epoch 6 / 20) train acc: 0.578000; val_acc: 0.509000
(Iteration 1501 / 4900) loss: 1.168672
(Iteration 1601 / 4900) loss: 1.213925
(Iteration 1701 / 4900) loss: 1.308235
(Epoch 7 / 20) train acc: 0.592000; val_acc: 0.525000
(Iteration 1801 / 4900) loss: 1.084997
(Iteration 1901 / 4900) loss: 1.128507
(Epoch 8 / 20) train acc: 0.604000; val_acc: 0.526000
(Iteration 2001 / 4900) loss: 1.126776
(Iteration 2101 / 4900) loss: 1.342422
(Iteration 2201 / 4900) loss: 1.347454
(Epoch 9 / 20) train acc: 0.611000; val_acc: 0.493000
(Iteration 2301 / 4900) loss: 0.974861
(Iteration 2401 / 4900) loss: 1.100155
(Epoch 10 / 20) train acc: 0.606000; val_acc: 0.528000
(Iteration 2501 / 4900) loss: 1.083061
(Iteration 2601 / 4900) loss: 1.031448
(Epoch 11 / 20) train acc: 0.623000; val_acc: 0.510000
(Iteration 2701 / 4900) loss: 1.144850
(Iteration 2801 / 4900) loss: 1.073761
(Iteration 2901 / 4900) loss: 1.102385
(Epoch 12 / 20) train acc: 0.652000; val_acc: 0.516000
(Iteration 3001 / 4900) loss: 1.239039
(Iteration 3101 / 4900) loss: 0.998472
(Epoch 13 / 20) train acc: 0.647000; val_acc: 0.533000
(Iteration 3201 / 4900) loss: 1.027785
(Iteration 3301 / 4900) loss: 0.941700
(Iteration 3401 / 4900) loss: 0.995375
(Epoch 14 / 20) train acc: 0.631000; val_acc: 0.533000
(Iteration 3501 / 4900) loss: 1.016337
(Iteration 3601 / 4900) loss: 0.979582
(Epoch 15 / 20) train acc: 0.684000; val_acc: 0.529000
(Iteration 3701 / 4900) loss: 0.942912
(Iteration 3801 / 4900) loss: 1.118059
(Iteration 3901 / 4900) loss: 0.970065
(Epoch 16 / 20) train acc: 0.680000; val_acc: 0.521000
(Iteration 4001 / 4900) loss: 1.024775
(Iteration 4101 / 4900) loss: 0.806615
(Epoch 17 / 20) train acc: 0.675000; val_acc: 0.518000
(Iteration 4201 / 4900) loss: 0.976157
(Iteration 4301 / 4900) loss: 1.011377
(Iteration 4401 / 4900) loss: 0.839201
(Epoch 18 / 20) train acc: 0.699000; val_acc: 0.539000
```

```
(Iteration 4501 / 4900) loss: 0.842462
(Iteration 4601 / 4900) loss: 0.878570
(Epoch 19 / 20) train acc: 0.692000; val_acc: 0.531000
(Iteration 4701 / 4900) loss: 0.775290
(Iteration 4801 / 4900) loss: 0.665518
(Epoch 20 / 20) train acc: 0.725000; val_acc: 0.525000
```

In [12]:

```
# Run this cell to visualize training loss and train / val accuracy

plt.subplot(2, 1, 1)
plt.title('Training loss')
plt.plot(solver.loss_history, 'o')
plt.xlabel('Iteration')

plt.subplot(2, 1, 2)
plt.title('Accuracy')
plt.plot(solver.train_acc_history, '-o', label='train')
plt.plot(solver.val_acc_history, '-o', label='val')
plt.plot([0.5] * len(solver.val_acc_history), 'k--')
plt.xlabel('Epoch')
plt.legend(loc='lower right')
plt.gcf().set_size_inches(15, 12)
plt.show()
```



Multilayer Neural Network

Now, we implement a multi-layer neural network.

Read through the `FullyConnectedNet` class in the file `nndl/fc_net.py` .

Implement the initialization, the forward pass, and the backward pass. There will be lines for batchnorm and dropout layers and caches; ignore these all for now. That'll be in HW #4.

In [15]:

```
N, D, H1, H2, C = 2, 15, 20, 30, 10
X = np.random.randn(N, D)
y = np.random.randint(C, size=(N,))

for reg in [0, 3.14]:
  print('Running check with reg = {}'.format(reg))
  model = FullyConnectedNet([H1, H2], input_dim=D, num_classes=C,
                            reg=reg, weight_scale=5e-2, dtype=np.float64)

  loss, grads = model.loss(X, y)
  print('Initial loss: {}'.format(loss))

  for name in sorted(grads):
    f = lambda _: model.loss(X, y)[0]
    grad_num = eval_numerical_gradient(f, model.params[name], verbose=False, h=1e-5)
    print('{} relative error: {}'.format(name, rel_error(grad_num, grads[name])))
```

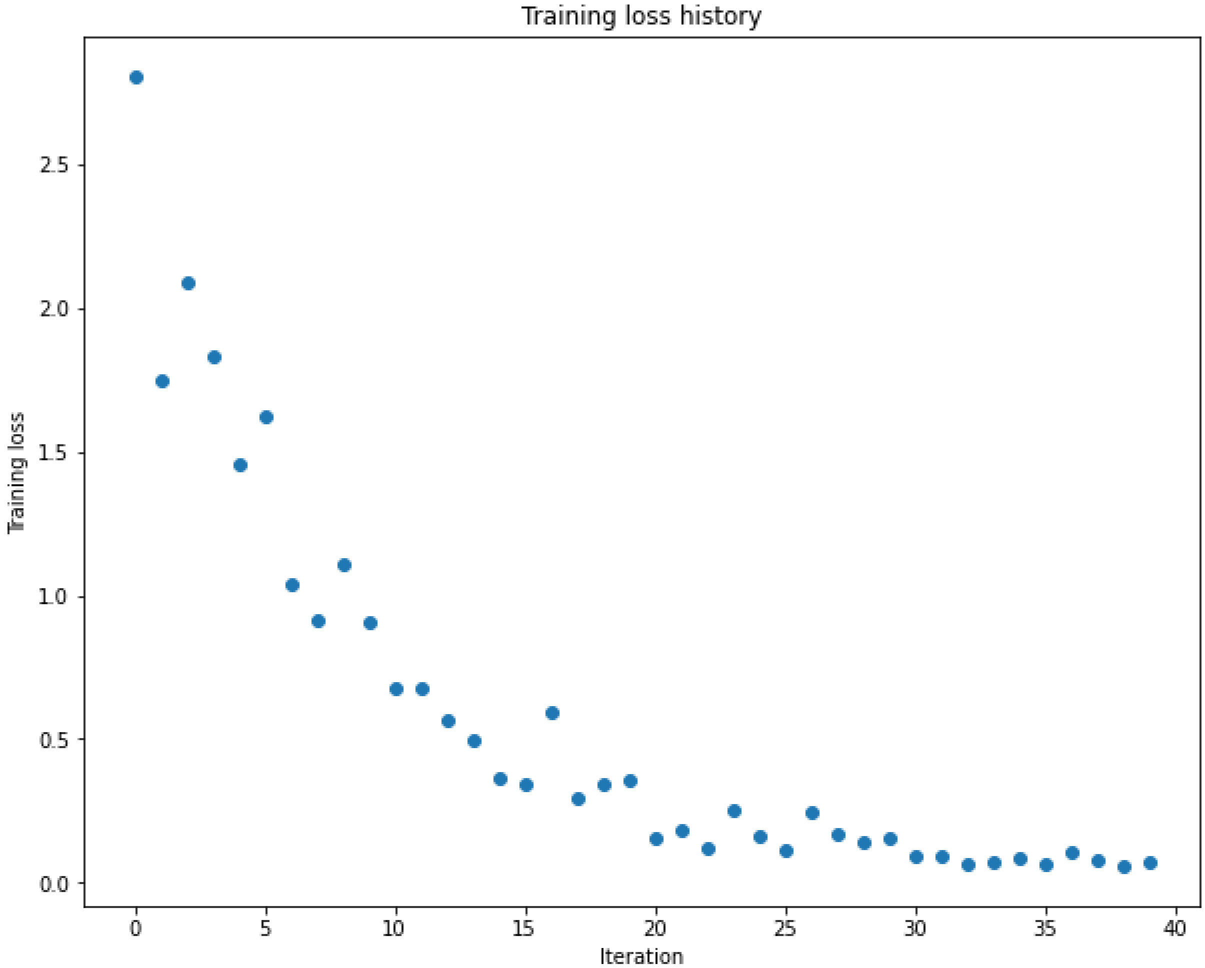
```
Running check with reg = 0
Initial loss: 2.29987087328397
W1 relative error: 1.0767409320850991e-07
W2 relative error: 1.535895657617354e-07
W3 relative error: 1.2183217354960783e-07
b1 relative error: 6.306523546533626e-09
b2 relative error: 1.5443067428863164e-09
b3 relative error: 1.3940105778846514e-10
Running check with reg = 3.14
Initial loss: 6.985047275077401
W1 relative error: 9.003035977130088e-08
W2 relative error: 2.238394860126933e-08
W3 relative error: 1.0143496565336713e-08
b1 relative error: 3.428469487512458e-08
b2 relative error: 3.7162697538043037e-09
b3 relative error: 1.1024429933261265e-10
```

In [16]:

```
# Use the three layer neural network to overfit a small dataset.

num_train = 50
small_data = {
  'X_train': data['X_train'][:num_train],
  'y_train': data['y_train'][:num_train],
  'X_val': data['X_val'],
  'y_val': data['y_val'],
}


#### !!!!!!
# Play around with the weight_scale and learning_rate so that you can overfit a small data
# Your training accuracy should be 1.0 to receive full credit on this part.
weight_scale = 2e-2
learning_rate = 3e-3

model = FullyConnectedNet([100, 100],
              weight_scale=weight_scale, dtype=np.float64)
solver = Solver(model, small_data,
                print_every=10, num_epochs=20, batch_size=25,
                update_rule='sgd',
```

```python
                optim_config={
                  'learning_rate': learning_rate,
                }
         )
solver.train()

plt.plot(solver.loss_history, 'o')
plt.title('Training loss history')
plt.xlabel('Iteration')
plt.ylabel('Training loss')
plt.show()
```

```
(Iteration 1 / 40) loss: 2.809405
(Epoch 0 / 20) train acc: 0.240000; val_acc: 0.087000
(Epoch 1 / 20) train acc: 0.360000; val_acc: 0.099000
(Epoch 2 / 20) train acc: 0.540000; val_acc: 0.122000
(Epoch 3 / 20) train acc: 0.560000; val_acc: 0.129000
(Epoch 4 / 20) train acc: 0.720000; val_acc: 0.152000
(Epoch 5 / 20) train acc: 0.860000; val_acc: 0.151000
(Iteration 11 / 40) loss: 0.677575
(Epoch 6 / 20) train acc: 0.920000; val_acc: 0.154000
(Epoch 7 / 20) train acc: 0.940000; val_acc: 0.161000
(Epoch 8 / 20) train acc: 0.940000; val_acc: 0.164000
(Epoch 9 / 20) train acc: 1.000000; val_acc: 0.158000
(Epoch 10 / 20) train acc: 0.960000; val_acc: 0.147000
(Iteration 21 / 40) loss: 0.151375
(Epoch 11 / 20) train acc: 0.980000; val_acc: 0.165000
(Epoch 12 / 20) train acc: 1.000000; val_acc: 0.163000
(Epoch 13 / 20) train acc: 1.000000; val_acc: 0.156000
(Epoch 14 / 20) train acc: 1.000000; val_acc: 0.161000
(Epoch 15 / 20) train acc: 1.000000; val_acc: 0.157000
(Iteration 31 / 40) loss: 0.090922
(Epoch 16 / 20) train acc: 1.000000; val_acc: 0.164000
(Epoch 17 / 20) train acc: 1.000000; val_acc: 0.170000
(Epoch 18 / 20) train acc: 1.000000; val_acc: 0.166000
(Epoch 19 / 20) train acc: 1.000000; val_acc: 0.160000
(Epoch 20 / 20) train acc: 1.000000; val_acc: 0.169000
```



Training loss history
Training loss
Iteration
0.0
0.5
1.0
1.5
2.0
2.5
0
5
10
15
20
25
30
35
40

```python
import numpy as np
import pdb




def affine_forward(x, w, b):
  """
  Computes the forward pass for an affine (fully-connected) layer.

  The input x has shape (N, d_1, ..., d_k) and contains a minibatch of N
  examples, where each example x[i] has shape (d_1, ..., d_k). We will
  reshape each input into a vector of dimension D = d_1 * ... * d_k, and
  then transform it to an output vector of dimension M.

  Inputs:
  - x: A numpy array containing input data, of shape (N, d_1, ..., d_k)
  - w: A numpy array of weights, of shape (D, M)
  - b: A numpy array of biases, of shape (M,)

  Returns a tuple of:
  - out: output, of shape (N, M)
  - cache: (x, w, b)
  """

  # ================================================================ #
  # YOUR CODE HERE:
  #   Calculate the output of the forward pass.  Notice the dimensions
  #   of w are D x M, which is the transpose of what we did in earlier
  #   assignments.
  # ================================================================ #

  x_reshape = x.reshape((x.shape[0], -1)) #N x D
  out = np.dot(x_reshape, w) + b.reshape((1, b.shape[0])) #N x M

  # ================================================================ #
  # END YOUR CODE HERE
  # ================================================================ #

  cache = (x, w, b)
  return out, cache


def affine_backward(dout, cache):
  """
  Computes the backward pass for an affine layer.

  Inputs:
  - dout: Upstream derivative, of shape (N, M)
  - cache: Tuple of:
    - x: Input data, of shape (N, d_1, ... d_k)
    - w: Weights, of shape (D, M)

  Returns a tuple of:
  - dx: Gradient with respect to x, of shape (N, d1, ..., d_k)
  - dw: Gradient with respect to w, of shape (D, M)
  - db: Gradient with respect to b, of shape (M,)
  """
  x, w, b = cache
```

```
    dx, dw, db = None, None, None

    # ================================================================ #
    # YOUR CODE HERE:
    #   Calculate the gradients for the backward pass.
    # ================================================================ #

    # dout is N x M
    # dx should be N x d1 x ... x dk; it relates to dout through multiplication with w, which is D x M
    # dw should be D x M; it relates to dout through multiplication with x, which is N x D after reshaping
    # db should be M; it is just the sum over dout examples

    x_reshape = np.reshape(x, (x.shape[0], -1)) #N x D
    dx_reshape = np.dot(dout, w.T)

    dx = np.reshape(dx_reshape, x.shape) #N x D
    dw = np.dot(x_reshape.T, dout) #D x M
    db = np.sum(dout.T, axis=1, keepdims=True).T  #M x 1

    # ================================================================ #
    # END YOUR CODE HERE
    # ================================================================ #

    return dx, dw, db

def relu_forward(x):
    """
    Computes the forward pass for a layer of rectified linear units (ReLUs).

    Input:
    - x: Inputs, of any shape

    Returns a tuple of:
    - out: Output, of the same shape as x
    - cache: x
    """
    # ================================================================ #
    # YOUR CODE HERE:
    #   Implement the ReLU forward pass.
    # ================================================================ #

    out = np.maximum(x, 0)

    # ================================================================ #
    # END YOUR CODE HERE
    # ================================================================ #

    cache = x
    return out, cache


def relu_backward(dout, cache):
    """
    Computes the backward pass for a layer of rectified linear units (ReLUs).

    Input:
    - dout: Upstream derivatives, of any shape
    - cache: Input x, of same shape as dout
```

```python

  Returns:
  - dx: Gradient with respect to x
  """
  x = cache

  # =========================================================================== #
  # YOUR CODE HERE:
  #   Implement the ReLU backward pass
  # =========================================================================== #

  # ReLU directs linearly to those > 0

  dx = dout * (x > 0)

  # =========================================================================== #
  # END YOUR CODE HERE
  # =========================================================================== #

  return dx

def svm_loss(x, y):
  """
  Computes the loss and gradient using for multiclass SVM classification.

  Inputs:
  - x: Input data, of shape (N, C) where x[i, j] is the score for the jth class
    for the ith input.
  - y: Vector of labels, of shape (N,) where y[i] is the label for x[i] and
    0 <= y[i] < C

  Returns a tuple of:
  - loss: Scalar giving the loss
  - dx: Gradient of the loss with respect to x
  """
  N = x.shape[0]
  correct_class_scores = x[np.arange(N), y]
  margins = np.maximum(0, x - correct_class_scores[:, np.newaxis] + 1.0)
  margins[np.arange(N), y] = 0
  loss = np.sum(margins) / N
  num_pos = np.sum(margins > 0, axis=1)
  dx = np.zeros_like(x)
  dx[margins > 0] = 1
  dx[np.arange(N), y] -= num_pos
  dx /= N
  return loss, dx


def softmax_loss(x, y):
  """
  Computes the loss and gradient for softmax classification.

  Inputs:
  - x: Input data, of shape (N, C) where x[i, j] is the score for the jth class
    for the ith input.
  - y: Vector of labels, of shape (N,) where y[i] is the label for x[i] and
    0 <= y[i] < C

```

```
    Returns a tuple of:
    - loss: Scalar giving the loss
    - dx: Gradient of the loss with respect to x
    """

    probs = np.exp(x - np.max(x, axis=1, keepdims=True))
    probs /= np.sum(probs, axis=1, keepdims=True)
    N = x.shape[0]
    loss = -np.sum(np.log(probs[np.arange(N), y])) / N
    dx = probs.copy()
    dx[np.arange(N), y] -= 1
    dx /= N
    return loss, dx
```

```python
import numpy as np

from .layers import *
from .layer_utils import *


class TwoLayerNet(object):
  """
  A two-layer fully-connected neural network with ReLU nonlinearity and
  softmax loss that uses a modular layer design. We assume an input dimension
  of D, a hidden dimension of H, and perform classification over C classes.

  The architecure should be affine - relu - affine - softmax.

  Note that this class does not implement gradient descent; instead, it
  will interact with a separate Solver object that is responsible for running
  optimization.

  The learnable parameters of the model are stored in the dictionary
  self.params that maps parameter names to numpy arrays.
  """

  def __init__(self, input_dim=3*32*32, hidden_dims=100, num_classes=10,
               dropout=0, weight_scale=1e-3, reg=0.0):
    """
    Initialize a new network.

    Inputs:
    - input_dim: An integer giving the size of the input
    - hidden_dims: An integer giving the size of the hidden layer
    - num_classes: An integer giving the number of classes to classify
    - dropout: Scalar between 0 and 1 giving dropout strength.
    - weight_scale: Scalar giving the standard deviation for random
      initialization of the weights.
    - reg: Scalar giving L2 regularization strength.
    """
    self.params = {}
    self.reg = reg

    # ================================================================ #
    # YOUR CODE HERE:
    #   Initialize W1, W2, b1, and b2.  Store these as self.params['W1'],
    #   self.params['W2'], self.params['b1'] and self.params['b2']. The
    #   biases are initialized to zero and the weights are initialized
    #   so that each parameter has mean 0 and standard deviation weight_scale.
    #   The dimensions of W1 should be (input_dim, hidden_dim) and the
    #   dimensions of W2 should be (hidden_dims, num_classes)
    # ================================================================ #

    self.params['W1'] = weight_scale * np.random.randn(input_dim, hidden_dims) + 0
    self.params['W2'] = weight_scale * np.random.randn(hidden_dims, num_classes) + 0

    self.params['b1'] = np.zeros((hidden_dims, 1))
    self.params['b2'] = np.zeros((num_classes, 1))

    # ================================================================ #
```

```python
    # END YOUR CODE HERE
    # ================================================================ #

  def loss(self, X, y=None):
    """
    Compute loss and gradient for a minibatch of data.

    Inputs:
    - X: Array of input data of shape (N, d_1, ..., d_k)
    - y: Array of labels, of shape (N,). y[i] gives the label for X[i].

    Returns:
    If y is None, then run a test-time forward pass of the model and return:
    - scores: Array of shape (N, C) giving classification scores, where
      scores[i, c] is the classification score for X[i] and class c.

    If y is not None, then run a training-time forward and backward pass and
    return a tuple of:
    - loss: Scalar value giving the loss
    - grads: Dictionary with the same keys as self.params, mapping parameter
      names to gradients of the loss with respect to those parameters.
    """
    scores = None

    # ================================================================ #
    # YOUR CODE HERE:
    #   Implement the forward pass of the two-layer neural network. Store
    #   the class scores as the variable 'scores'.  Be sure to use the layers
    #   you prior implemented.
    # ================================================================ #

    out_l1, cache_l1 = affine_forward(X, self.params['W1'], self.params['b1'])
    out_relu, cache_relu = relu_forward(out_l1)
    scores, cache_l2 = affine_forward(out_relu, self.params['W2'], self.params['b2'])

    # ================================================================ #
    # END YOUR CODE HERE
    # ================================================================ #

    # If y is None then we are in test mode so just return scores
    if y is None:
      return scores

    loss, grads = 0, {}
    # ================================================================ #
    # YOUR CODE HERE:
    #   Implement the backward pass of the two-layer neural net.  Store
    #   the loss as the variable 'loss' and store the gradients in the
    #   'grads' dictionary.  For the grads dictionary, grads['W1'] holds
    #   the gradient for W1, grads['b1'] holds the gradient for b1, etc.
    #   i.e., grads[k] holds the gradient for self.params[k].
    #
    #   Add L2 regularization, where there is an added cost 0.5*self.reg*W^2
    #   for each W.  Be sure to include the 0.5 multiplying factor to
    #   match our implementation.
    #
    #   And be sure to use the layers you prior implemented.
    # ================================================================ #
```

```python

    loss, dx2 = softmax_loss(scores, y)
    reg_loss = 0.5 * self.reg * (np.linalg.norm(self.params['W1'], 'fro')**2 + np.linalg.norm(self.params['W2'], 'fro')**2)
    loss += reg_loss

    dh1, dW2, db2 = affine_backward(dx2, cache_l2)
    da = relu_backward(dh1, cache_relu)
    dx1, dW1, db1 = affine_backward(da, cache_l1)

    grads['W1'] = dW1 + self.reg * self.params['W1']
    grads['b1'] = db1.T

    grads['W2'] = dW2 + self.reg * self.params['W2']
    grads['b2'] = db2.T

    # ================================================================ #
    # END YOUR CODE HERE
    # ================================================================ #

    return loss, grads


class FullyConnectedNet(object):
  """
  A fully-connected neural network with an arbitrary number of hidden layers,
  ReLU nonlinearities, and a softmax loss function. This will also implement
  dropout and batch normalization as options. For a network with L layers,
  the architecture will be

  {affine - [batch norm] - relu - [dropout]} x (L - 1) - affine - softmax

  where batch normalization and dropout are optional, and the {...} block is
  repeated L - 1 times.

  Similar to the TwoLayerNet above, learnable parameters are stored in the
  self.params dictionary and will be learned using the Solver class.
  """

  def __init__(self, hidden_dims, input_dim=3*32*32, num_classes=10,
               dropout=0, use_batchnorm=False, reg=0.0,
               weight_scale=1e-2, dtype=np.float32, seed=None):
    """
    Initialize a new FullyConnectedNet.

    Inputs:
    - hidden_dims: A list of integers giving the size of each hidden layer.
    - input_dim: An integer giving the size of the input.
    - num_classes: An integer giving the number of classes to classify.
    - dropout: Scalar between 0 and 1 giving dropout strength. If dropout=0 then
      the network should not use dropout at all.
    - use_batchnorm: Whether or not the network should use batch normalization.
    - reg: Scalar giving L2 regularization strength.
    - weight_scale: Scalar giving the standard deviation for random
      initialization of the weights.
    - dtype: A numpy datatype object; all computations will be performed using
      this datatype. float32 is faster but less accurate, so you should use
```

```
      float64 for numeric gradient checking.
    - seed: If not None, then pass this random seed to the dropout layers. This
      will make the dropout layers deteriminstic so we can gradient check the
      model.
    """
    self.use_batchnorm = use_batchnorm
    self.use_dropout = dropout > 0
    self.reg = reg
    self.num_layers = 1 + len(hidden_dims)
    self.dtype = dtype
    self.params = {}

    # ================================================================ #
    # YOUR CODE HERE:
    #   Initialize all parameters of the network in the self.params dictionary.
    #   The weights and biases of layer 1 are W1 and b1; and in general the
    #   weights and biases of layer i are Wi and bi. The
    #   biases are initialized to zero and the weights are initialized
    #   so that each parameter has mean 0 and standard deviation weight_scale.
    # ================================================================ #

    dims = []
    dims.append(input_dim)
    dims.extend(hidden_dims)
    dims.append(num_classes)

    for i in np.arange(self.num_layers):
      num = str(i+1)
      self.params['W'+num] = weight_scale * np.random.randn(dims[i], dims[i+1]) + 0
      self.params['b'+num] = np.zeros((dims[i+1], 1))

    # ================================================================ #
    # END YOUR CODE HERE
    # ================================================================ #

    # When using dropout we need to pass a dropout_param dictionary to each
    # dropout layer so that the layer knows the dropout probability and the mode
    # (train / test). You can pass the same dropout_param to each dropout layer.
    self.dropout_param = {}
    if self.use_dropout:
      self.dropout_param = {'mode': 'train', 'p': dropout}
      if seed is not None:
        self.dropout_param['seed'] = seed

    # With batch normalization we need to keep track of running means and
    # variances, so we need to pass a special bn_param object to each batch
    # normalization layer. You should pass self.bn_params[0] to the forward pass
    # of the first batch normalization layer, self.bn_params[1] to the forward
    # pass of the second batch normalization layer, etc.
    self.bn_params = []
    if self.use_batchnorm:
```

```
    self.bn_params = [{'mode': 'train'} for i in np.arange(self.num_layers - 1)]

    # Cast all parameters to the correct datatype
    for k, v in self.params.items():
      self.params[k] = v.astype(dtype)


  def loss(self, X, y=None):
    """
    Compute loss and gradient for the fully-connected net.

    Input / output: Same as TwoLayerNet above.
    """
    X = X.astype(self.dtype)
    mode = 'test' if y is None else 'train'

    # Set train/test mode for batchnorm params and dropout param since they
    # behave differently during training and testing.
    if self.dropout_param is not None:
      self.dropout_param['mode'] = mode
    if self.use_batchnorm:
      for bn_param in self.bn_params:
        bn_param[mode] = mode

    scores = None

    # ================================================================ #
    # YOUR CODE HERE:
    #   Implement the forward pass of the FC net and store the output
    #   scores as the variable "scores".
    # ================================================================ #

    outs = {}
    h = {}
    h[0] = [X]

    for i in np.arange(self.num_layers):
      num = str(i+1)
      outs[i+1] = affine_forward(h[i][0], self.params['W'+num], self.params['b'+num])
      if i != (self.num_layers-1):
        h[i+1] = relu_forward(outs[i+1][0])

    scores = outs[self.num_layers][0]

    # ================================================================ #
    # END YOUR CODE HERE
    # ================================================================ #

    # If test mode return early
    if mode == 'test':
      return scores

    loss, grads = 0.0, {}
    # ================================================================ #
    # YOUR CODE HERE:
    #   Implement the backwards pass of the FC net and store the gradients
    #   in the grads dict, so that grads[k] is the gradient of self.params[k]
    #   Be sure your L2 regularization includes a 0.5 factor.
```

```
    # ================================================================ #

    loss, dx = softmax_loss(scores, y)
    reg_loss_sum = 0
    for i in np.arange(self.num_layers):
        num = str(i+1)
        reg_loss_sum += np.linalg.norm(self.params['W'+num], 'fro')**2

    loss += 0.5 * self.reg * reg_loss_sum

    dict_dW = {}
    dict_db = {}

    dict_da = {}
    dict_da[self.num_layers] = dx

    for i in np.arange(self.num_layers, 0, -1):
      dh, dW, db = affine_backward(dict_da[i], outs[i][1])
      dict_dW[i] = dW
      dict_db[i] = db

      if i != 1:
          dict_da[i-1] = relu_backward(dh, h[i-1][1])

    for i in np.arange(self.num_layers):
      num = str(i+1)
      grads['W'+num] = dict_dW[i+1] + self.reg * self.params['W'+num]
      grads['b'+num] = dict_db[i+1].T

    # ================================================================ #
    # END YOUR CODE HERE
    # ================================================================ #
    return loss, grads
```