EXERCISES IN COMPUTATIONAL ANALYTICS

UNIVERSITY OF SAN FRANCISCO

Master of Science in Analytics

TERENCE PARR

parrt@cs.usfca.edu

http://parrt.cs.usfca.edu

ON

Contents

Part I

Introduction

Audience and Summary

Welcome to MSAN501, the computational analytics boot camp at the University of San Francisco! This exercise book collects all of the labs you must complete by the end of the boot camp in order to pass. The labs start out as very simple tasks or step-by-step recipes but then accelerate in difficulty, culminating with an interesting text analysis project. You will build all projects in Python (2.7.x).

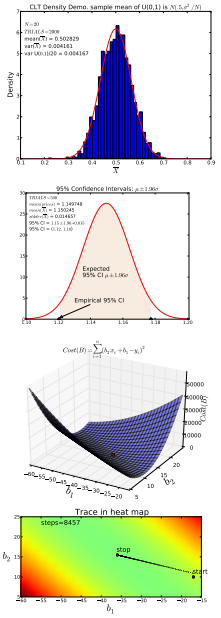
This course is specifically designed as an introduction to analytics programming for those who are not yet skilled programmers. The course also explores many concepts from math and statistics, but in an empirical fashion rather than symbolically as one would do in a math class. Consequently, this course is also useful to programmers who'd like to strengthen their understanding of numerical methods.

The exercises are grouped into parts. We begin with simple programs to compute statistics, build simple data structures, and use libraries to create visualizations and then move on to learning to use the UNIX command line, launch virtual computers in the cloud, and write simple Hadoop map-reduce programs. The empirical statistics part strives to give an intuitive feel for random variables, density functions, the central limit theorem, hypothesis testing, and confidence intervals. It's one thing to learn about their formal definitions, but to get a really solid grasp of these concepts, it really helps to observe statistics in action. All of the techniques we'll use in empirical statistics rely on the ability to generate random values from a particular distribution. We can do it all from a uniform random number generator, which is the first exercise in that part.

The next set of exercises deal with function optimization. Given a particular function, $f(x)$, optimizing it generally means finding its minimum or maximum, which occur when the derivative goes flat: $f'(x) = 0$. When the function's derivative cannot be derived symbolically, we're left with a general technique called *gradient descent* that searches for minima. It's like putting a marble on a hilly surface and letting gravity bring it to the nearest minimum.

Finally, you'll get an introduction to text analysis. We will compute something called *TFIDF* that indicates how well that word distinguishes a document from other documents in a corpus. That score is used broadly in text analytics, but our exercise uses it to summarize documents by listing the most important words.

You will work mostly on your own laptops, but you must get familiar with the UNIX command line. It's also important to learn how to install software and execute commands on a remote server; servers or what provide the websites you visit while browsing and they provide services to mobile apps on your phone. We've received an educational grant from Amazon to use their compute cloud called Amazon Web Services (AWS). We also have access to an IBM cluster, housed in the College of arts and sciences at USF.



As you progress through these exercises, you'll learn a great deal about Python and the following libraries: `matplotlib`, `numpy`, `scipy`, and `py.test`. I also recommend that you learn how to use a Python development environment called PyCharm, for which we have been granted a site license.

"Newbies say the darndest things"

in progress. warnings for newbies.

`-1/2` is `-1` and `1/2` is `0` in Python. There is no automatic promotion when you send an integer to a function that is expecting a floating-point number.

(-1/lambduh)*(np.log(1-u)) vs -(np.log(1-u))/lambduh. The latter is probably better because it does fewer floating-point operations and hence probably has fewer errors.

This doesn't do what you think: X = [[0] * N] * TRIALS

Part II

Python Programming and Data Structures

Computing Point Statistics

Discussion

The goal of this task is to get familiar with Python function definitions and looping structures, as well as to refresh your memory about a few point statistics.

Stats

This exercise involves writing functions to compute sample mean, variance, and covariance from a data set (list of values). In mathematics notation, the sample estimates are:

$$\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i \qquad \text{(Sample mean)}$$

$$s^2 = \frac{1}{N-1}\sum_{i=1}^{N}(x_i - m)^2 \qquad \text{(Unbiased sample variance)}$$

$$cov(x,y) = \frac{1}{N-1}\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y}) \qquad \text{(Unbiased sample covariance)}$$

In Python, you must define functions `mean(x)`, `var(x)`, `cov(x,y)` where `x` and `y` are objects that behave like a list or iterator. The functions return a floating-point value based on the above mathematics notation. If the length of the incoming vectors to `cov` are not the same, return 0. To test things out, use the `test_stats.py` file provided for you as part of this task.

Libraries

While we're at it let's learn about importing libraries. You'll notice that `test_stats.py` references your code like this:

```
from stats import *
```

That lets us directly access the functions defined in your `stats.py` file.

You can also test the correctness of the functions by using the `numpy` lib, make sure you ask for the sample population statistics by using parameter `ddof=1` for `var()` and `cov()`. E.g., `np.var(data, ddof=1)`. Be careful not to confuse function names; e.g., `numpy` has functions with the same names (although cov() returns a covariance matrix).

```
import numpy as np  # np is an alias for the numpy library
x = ...
y = ...
print np.cov(x,y)[0][1] # np.cov returns cov matrix
```

We now have the kernel of a small statistics library.

Deliverables

- `stats.py`

You may not use sum() or any other built-in functions for this project to compute the point statistics. The whole point of the exercise is to learn to build your own for loops. Obviously.

Approximating $\sqrt{n}$ with the Babylonian Method

Motivation

This lab is really a fancy way to learn about looping in Python and how to quickly prototype something in Excel (if warranted). It also gets you used to encoding mathematical expressions and recurrence relations in Python.

Discussion

To approximate square root, $\sqrt{n}$, the idea is to pick an initial estimate, x_0, and then iterate with better and better estimates, x_i, using the recurrence relation:

$$x_{i+1} = \frac{1}{2}(x_i + \frac{n}{x_i})$$

To see how this works, jump into Excel (yes, a spreadsheet) and crank through a few iterations by defining cells with n and your initial estimate x_0, which can be anything you want. (It's sometimes easier to play around without having to deal with a programming language.) Then you need to define a cell that computes the above better approximation using x_i as the cell above it. I hardcoded the names in column A and the first two rows of column B. Cell B3 should be a formula that computes B4 based upon B3. Then you can extend the formula down and watch it converge on the final (correct) value for $\sqrt{125348}$. My spreadsheet looks like this:

	A	B
1	n	125348
2	x_0	20
3	x_1	3143.7
4	x_2	1591.78638
5	...	835.266564
6		492.668011
7		373.547461
8		354.554285
9		354.04556
10		354.045195
11		354.045195

Try out any nonnegative number and you'll see that it still converges, though at different rates.

There's a great deal on the web you can read to learn more about why this process works but it relies on the average (midpoint) of x and n/x getting us closer to $\sqrt{n}$. It can be shown that if x is above $\sqrt{n}$ then n/x is below $\sqrt{n}$ and the reverse is true if x is below the root. The iteration converges and does so quickly. Informally, as shown in Wikipedia, we can represent the true square root by adding an error term to our estimate:

$$\sqrt{n} = x + \epsilon$$

or,

$$n = (x + \epsilon)^2$$

Expanding, we get:

$$n = x^2 + 2x\epsilon + \epsilon^2$$

Solving for ϵ:

$$n - x^2 = \epsilon(2x + \epsilon)$$

$$\epsilon = \frac{n - x^2}{2x + \epsilon}$$

Because ϵ is much smaller than x, we can drop it from the denominator leaving us with an estimate of epsilon:

$$\epsilon = \frac{n - x^2}{2x}$$

Then we can plug it back into $x + \epsilon$ and get:

$$x := x + \epsilon = x + \frac{n - x^2}{2x} = \frac{2x^2}{2x} + \frac{n - x^2}{2x} = \frac{1}{2}\frac{x^2 + n}{x} = \frac{1}{2}(x + \frac{n}{x})$$

Which gets its back to the Babylonian formula. Since we dropped an ϵ term, this formula for x is inexact but it gets us closer to $\sqrt{n}$.

Now that you understand how this estimate works, your goal is to implement a simple Python method called sqrt() that uses the Babylonian method to approximate the square root. Here is a starter kit for you. Please call the file `sqrt.py`.

```
import math

# Stop iterating when the new approximation is within
# PRECISION of the old value.
PRECISION = 0.000001

# compute square root of n
def sqrt(n):
    x_0 = 1.0 # pick any old value
    ... fill this in ...

def check(n):
    delta = math.sqrt(n) - sqrt(n)
    if math.abs(delta) > PRECISION:
        raise BaseException("Inaccurate sqrt(%f)=%f; estimate is %f" %
            (n, math.sqrt(n), sqrt(n)))
```

```
# check a range of values
check(125348)
check(100)
check(1)
check(0)
```

Deliverables

Please submit:

- a PDF showing a snapshot of your spreadsheet
- the formula you used in B3 **and** B4.
- your `sqrt.py` Python file

You may not use math.sqrt() for implementing your function, but you may use it for testing the results. Obviously.

Generating Uniform Random Numbers

Q: *How to generate pure random string*?
A: *Put a fresh student in front of vi editor and ask him to quit.*
— Emiliano Lourbet (@taitooz)

Discussion

To perform computer-based simulation we need to be able to generate random numbers. Generating random numbers following a uniform distribution are the easiest to generate and are what comes out of the standard programming language "give me a random number" function. Here's a sample Python session:

```
>>> import random
>>> print random.random()
0.758309680374
>>> print random.random()
0.790739160715
>>> print random.random()
0.892622796934
```

We could generate real random numbers by accessing, for example, noise on the ethernet network device but that noise might not be uniformly distributed. We typically generate pseudorandom numbers that aren't really random but look like they are. From Ross' *Simulation*, we see a very easy recursive mechanism that generators values in $[0, m-1]$:

$x_n = ax_{n-1}$ modulo m

That's recursive (or iterative and not *closed form*) because x_n is a function of a prior value:

$x_1 = ax_0$ modulo m
$x_2 = ax_1$ modulo m
$x_3 = ax_2$ modulo m
$x_4 = ax_3$ modulo m
...

To get random numbers between 0 and 1, we use x_n / m.

We must pick a value for a and m that make x_n seem random. Ross suggests choosing a large prime number for m that fits in our integer word size, e.g., $m = 2^{31} - 1$, and $a = 7^5 = 16807$.

Initially we set a value for x_0, called the *random seed* (it is not the first random number). Every seed leads to a different sequence of pseudorandom numbers. In Python, you can set the seed of the standard library by using `random.seed([x])`.

Your goal is to take that simple recursive formula and use it to generate the first 10 random numbers using a `for` loop in Python as part of the "main" code. Please use file `varunif.py` and make function `runif()` that returns a new random value per call. Use $m = 2^{31} - 1$, $a = 7^5 = 16807$, and an initial seed of $x_0 = 666$. Your output should look something like:

```
0.00521236192678
0.604166903349
0.233144581892
0.460987861017
0.822980116505
0.826818094508
0.331714398848
0.1239014343
0.411406287184
0.505468696591
```

Because we are all using the same seed, the sequence of numbers should be the same.

Next, make function `runif_(a,b)` that returns random values between `a` and `b`. Your function definitions should look like:

```
# U(0,1)
def runif():
    ...

# U(a,b)
def runif_(a,b):
    ...
```

Deliverables

You must submit your `varunif.py` containing a main loop and your functions runif and runif_. Also submit your output via Canvas as a text file.

You may not use random.random() or any other built-in random number generators for this project. Obviously.

Histograms Using `matplotlib`

TODO: reuse runif() from previous project for 2015.

Discussion

The goal of this lab is to teach you the basics of using matplotlib to display probability mass functions, otherwise known as histograms. In this lab we will use the uniform distribution. Use filename `hist.py`.

Steps

1. import the proper libraries

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

2. Get a sample of uniform random variables in $U(0, 1)$

```
N = 1000
X = [np.random.uniform(0,1) for i in range(N)] # U(0,1)
# or np.random.uniform(0,1,N)
```

3. Display a histogram using matplotlib (in a separate window)

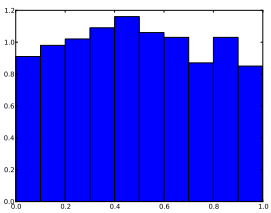
```
plt.hist(X, normed=1)
plt.show()
```

4. Run it.

5. Now, save the image as a PDF to the same directory by inserting a save command in between the histogram and the show method:

```
plt.hist(X, normed=1)
plt.savefig('unif-0-1-density.pdf', format="pdf")
plt.show()
```

6. Run it. Your pdf file should look like



7. Graphs should always have the axes labeled. Let's do that as well as add a title and set the range of the graph. Put this code right before the `savefig()`.

```
plt.title("U(0,1) Density Demo")
plt.xlabel("X", fontsize=16)
plt.ylabel("Density", fontsize=16)
plt.axis([0, 1, 0, 2])
```

8. Run it.

9. It's also common to add some annotations inside the graph to explain more about what we are seeing. First, we need to get access to the figure itself and then has to figure about its axes. (We need this in order to specify coordinates within the graph.) Put the following code before the `hist()` call.

```
fig = plt.figure()        # get a handle on the figure object itself
ax = fig.add_subplot(111) # weird stuff to get the Axes object within figure
```

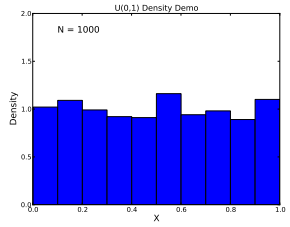
Then, before the `savefig()`, add the following to display some text above the histogram within the graph. The coordinates are from 0..1 where 0 is the left/bottom edge and 1 is the right/top edge.

```
# put N=... at top left
plt.text(.1,.9, 'N = %d' % N,
         fontsize=16,
         transform = ax.transAxes)
```

10. Also, let's change the file name slightly so we can keep our original graph plus our fancy one:

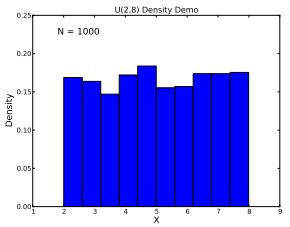
```
plt.savefig('unif-0-1-density-fancy.pdf', format="pdf")
```

11. Run it.



To understand distributions, it's a great idea to start messing around with the parameters of the density or mass function.

12. Change $U(0,1)$ to $U(2,8)$ and examine the results. You will have to alter the axis() to use different ranges. (Or let the plotting software do the work for you and get rid of the axis() call.) Run it. You should see something like the following.



Deliverables

Please submit:

- your `hist.py` Python file
- a PDF of your $U(2, 8)$ graph.

Graph Adjacency Lists and Matrices

Goal

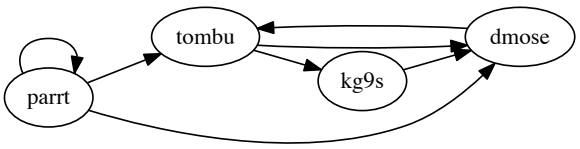
The goal of this task is to teach you about the implementation of graphs in Python, how to implement a few simple related algorithms, and do some simple data loading. As part of this exercise, you will also learn to transform data, which is an important data preparation skill you will need as an analyst.

Discussion

In this project, you will convert a string representation of a graph that looks like this:

```
parrt: tombu, dmose, parrt
tombu: dmose, kg9s
dmose: tombu
kg9s: dmose
```

to an adjacency list representation and ultimately generate a visual representation via graphviz/dot:



For fun, you will also create an edge matrix representation:

$$\begin{array}{c} \\ parrt \\ tombu \\ dmose \\ kg9s \end{array} \begin{array}{c} \begin{array}{cccc} parrt & tombu & dmose & kg9s \end{array} \\ \begin{pmatrix} 1 & 1 & 1 & 0 \\ 0 & 0 & 1 & 1 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{pmatrix} \end{array}$$

where the nodes have the following indexes (all we really care about here is the order):

$$\begin{bmatrix} 0 : parrt \\ 1 : tombu \\ 2 : dmose \\ 3 : kg9s \end{bmatrix}$$

The following sections describe the functions you must create in `graph.py`. See a `graph.py` starter kit I've built for you at github.

Processing an adjacency list string

First, you have to process a string representation of an adjacency list and create an internal data structure:

```
def adjlist(adj_list):
    """
    Read in adj list and store in form of dict mapping node
    name to list of outgoing edges. Preserve the order you find
    for the nodes.
    """
    ...
```

You will use an ordered dictionary (`OrderedDict`) that maps node name x to a list of target nodes. x will be a string and the target list will be a list of strings. For example, from line in string adj_list

`parrt: tombu, dmose, parrt`

you will create an entry in the dictionary with key `parrt` and value:

```
['tombu', 'dmose', 'parrt']
```

To process the text, you must split the incoming string into lines and then process them one at a time as each line represents an adjacency list. You will use string functions `split` and (likely) `strip` to process the text. The goal here is to learn how to process text so don't look for built-in functions that do all of this for you.
Printing the adjacency list dictionary from `adjlist`, we should all get the following output:

```
OrderedDict([('parrt', ['tombu', 'dmose', 'parrt']),
 ('tombu', ['dmose', 'kg9s']),
 ('dmose', ['tombu']),
 ('kg9s', ['dmose'])])
```

Adjacency list to adjacency matrix

Given an adjacency list stored as a dictionary per adjlist(), create a function that converts it to an adjacency matrix:

```
def adjmatrix(adj):
    """
    From an adjacency list, return the adjacency matrix with entries in {0,1}.
    The order of nodes in adj is assumed to be same as they were read in.
    """
    ...
```

The matrix should look like the one shown above.

Getting a list of all nodes

A very useful function to have is the following that returns a list of all nodes visited starting at a particular node in a graph.

```
def nodes(adj, start_node):
    """
    Walk every node in graph described by adj list starting at start_node
    using a breadth-first search.  Return a list of all nodes found (in
    any order). Include the start_node.
    """
    ...
```

Do not build a recursive function as you must do a breadth-first search. (Recursive functions are much more useful when doing a depth-first search.) The basic algorithm looks like this:

```
visited = [];
add the start node to a work list;
while more work do
  node = remove a node from work list;
  add node to visited list;
  targets = adjacency_list[node];
  add all unvisited targets to work list;
end
return visited;
```

Generating DOT output

In order to visualize the graph you have read in, create the following function that dumps valid Graphviz DOT code, given an adjacency list. Then cut-and-paste the output and put it into Graphviz to display it.

```
def gendot(adj):
    """
    Return a string representing the graph in Graphviz DOT format
    with all p->q edges. Parameter adj is an adjacency list.
    """
    ...
```

Or, to amaze your family and friends, you can directly from the command line on a mac or unix box:

```
python test_dot.py | dot -Tpdf > /tmp/graph.pdf; open /tmp/graph.pdf
```

Here is a simple test rig, `test_dot.py`, that translates an input string description to DOT and prints it out.

```
from graph import *

# test dot
g = \
"""
parrt: tombu, dmose, parrt
tombu: dmose, kg9s
dmose: tombu
```

```
kg9s: dmose
"""
list = adjlist(g)
dot = gendot(list)
print dot
```

For the adjacency list shown at the start of this assignment, you should to generate the following DOT code:

```
digraph g {
  rankdir=LR;
  parrt -> tombu;
  parrt -> dmose;
  parrt -> parrt;
  tombu -> dmose;
  tombu -> kg9s;
  dmose -> tombu;
  kg9s -> dmose;
}
```

Testing

I have provided `test_graph.py` and `test_dot.py` test rigs that exercise the required functions using the sample adjacency list described above. Please make sure that your library works with this test rig at minimum.

Deliverables

Please submit the following via canvas:

- `graph.py` (the functions inside should emit no output at all, just return data as specified)
- a text file with the output of running `test_dot.py`, showing the DOT output.
- a PDF showing the visual representation of the graph as generated by graphviz/dot

Part III

A Taste of Distributed Computing

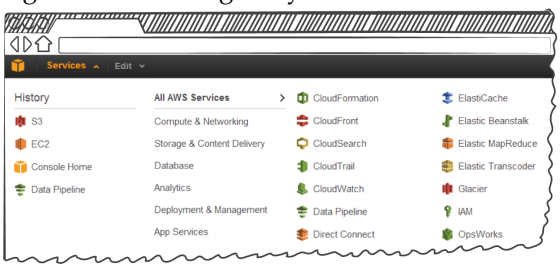
Launching a Virtual Machine at Amazon Web Services

Discussion

The goal of this lab is to teach you to create a Linux machine at Amazon Web Services, login, copy some data to that machine, and run a simple Python program on that data.

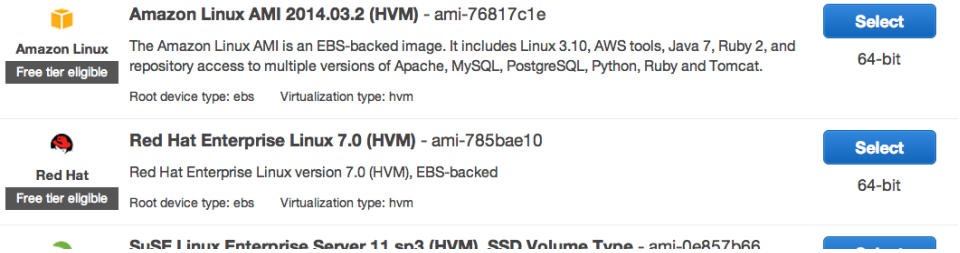
Steps

1. Login to AWS and go to your AWS console.

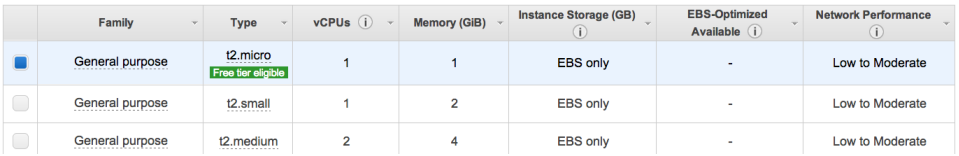


2. Click "Launch Instance", which will start the process to create a virtual machine in the cloud. An instance is just a virtual machine.

3. Select the "Amazon Linux AMI" server, which should be the first one. This is a commonly-used *image* that results in a Linux machine that contains lots of useful goodies as you can see from that list, such as Python and MySQL. An image is just a snapshot of the disk after someone carefully installs software properly on a Linux machine. This means we don't have to install software every time we create a new machine.

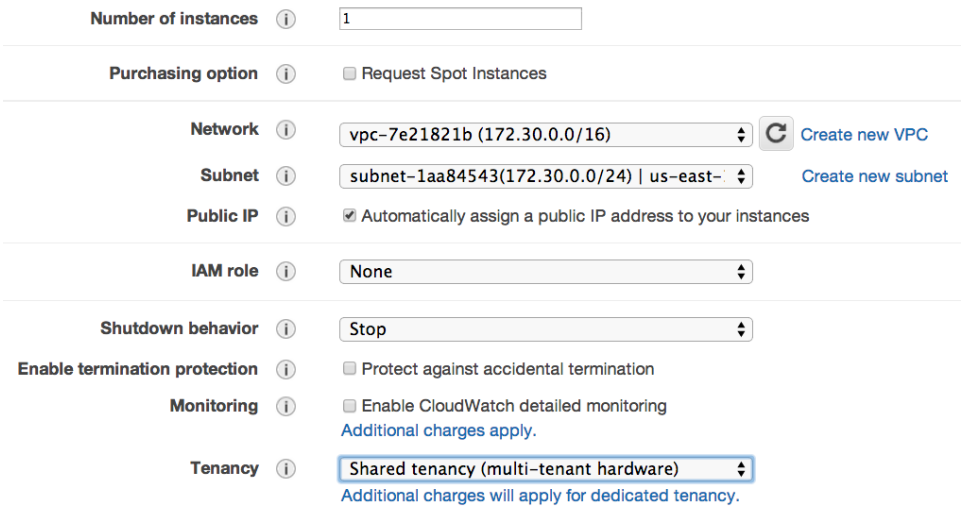


4. Select instance type "m1.micro," which should be the first machine type listed. This machine is very low powered but is sufficient for playing around. Click "Next: configure instance details."



	Family	Type	vCPUs	Memory (GiB)	Instance Storage (GB)	EBS-Optimized Available	Network Performance
☑	General purpose	t2.micro (Free tier eligible)	1	1	EBS only	-	Low to Moderate
☐	General purpose	t2.small	1	2	EBS only	-	Low to Moderate
☐	General purpose	t2.medium	2	4	EBS only	-	Low to Moderate

5. You can leave the configuration details as-is:



6. Ignore the network interface set up and advanced details. Click "Next: Add storage."

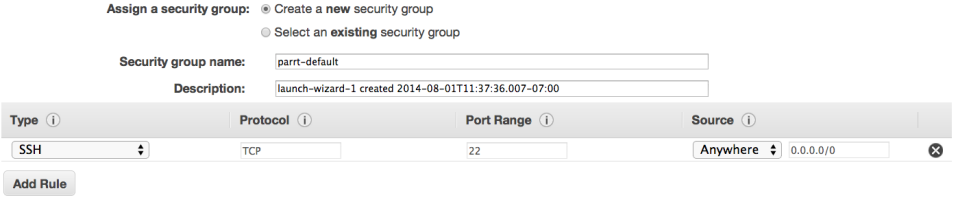
7. It shows that it will give us 8G of disk storage on a magnetic disk by default, which is good enough for our testing purposes. Click "Next: Tag instance."

Type	Device	Snapshot	Size (GiB)	Volume Type	IOPS	Delete on Termination	Encrypted
Root	/dev/xvda	snap-810ffc56	8	Magnetic	N/A	☑	Not Encrypted

8. For the key named "Name", change the value to something like *youruserid*-linux or something like that so that you can identify it later if you have multiple machines going. Click "Next: Configure security group." Then click on the group whose name is "default." Your list of security groups might not be the same.

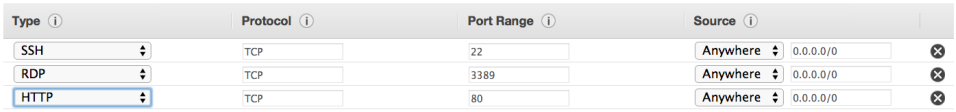
Key (127 characters maximum)	Value (255 characters maximum)
Name	parrt-linux

9. You want to create a new security group, so that you learn how to deal with firewalls. We want to allow SSH access, Windows RDP, and HTTP ports. Name it something like your userid-default. You should be able to reuse this the next time you create an instance just by selecting the name from the existing security groups pulldown. It initially shows just SSH port open so we have to add two more.

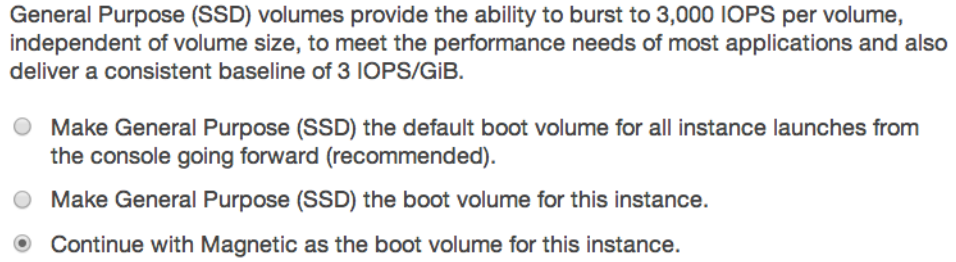


10. Click on the "Add rule" button and select RDP under the type and Anywhere under the source. That

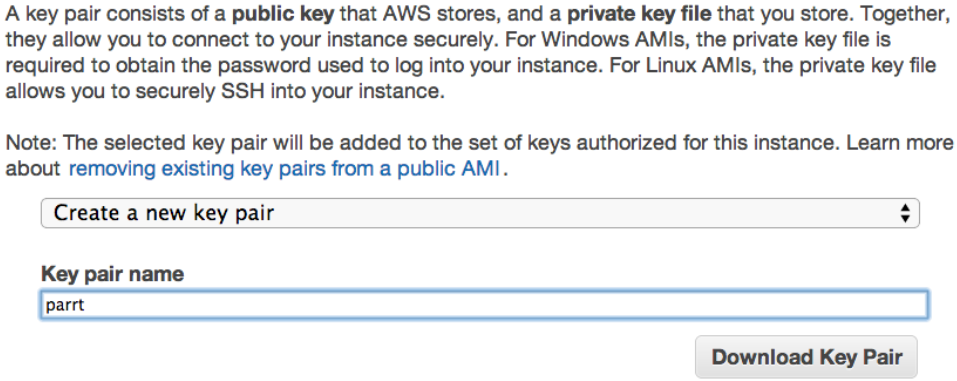
means we want anyone to be able to connect to this machine using the Windows Remote Desktop protocol and from any machine on the Internet. This is not a Windows machine but you will reuse the security group later. As we might want to start a Web server on our cloud computer, add a rule for HTTP.



11. Click "Review and launch," which will pull up a dialog box asking you to select whether you want SSD or old-school spinning magnetic disk. As we are just testing things and don't care about I/O speed, choose the magnetic disk and click "Next."

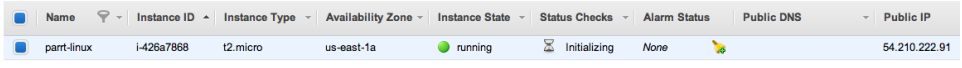


12. Click "Launch," which will bring a dialog box up to select a key pair. A key pair is what allows you to securely access the server and prevent unauthorized access. The first time, you will need to create a new key pair. Name it as your user ID then click on "Download key pair." It will download a *userid*.pem file, which are your security credentials for getting into the machine. Save that file in a safe spot. If you lose it you will not be able to get into the machine that you create.

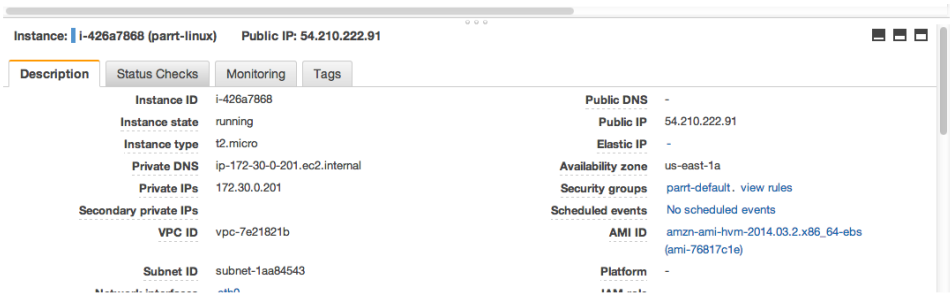


13. Click on the "I acknowledge that I have ..." checkbox then "Launch instances." You should see something like:

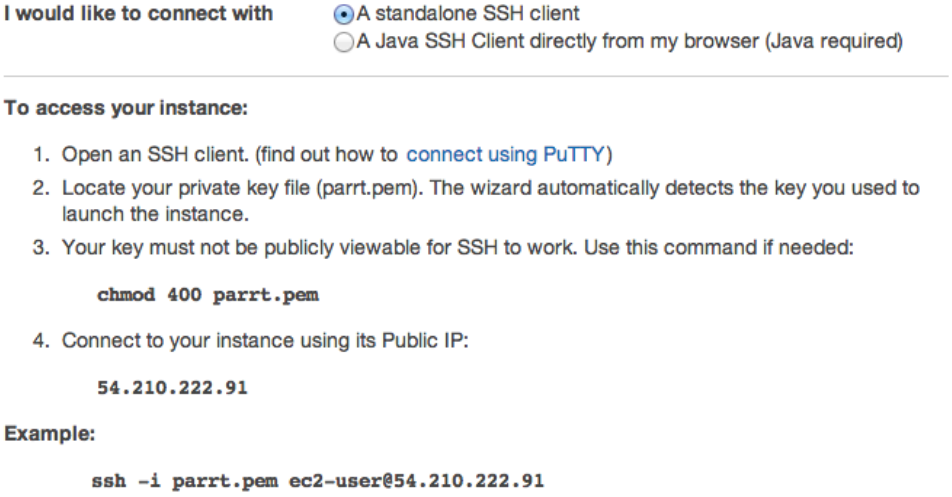
14. Click on the "i-..." link to go to the EC2 console showing your instance.



15. Click on your instance and you should see a description box at the bottom. Look for the "Public IP" address, which is 54.210.222.91 in this case:



16. Click on the "Connect" button at the top of the page and it will bring up a dialog box that tells you how to connect to the server. You want to connect with "A standalone SSH client" link (Java is now a security risk in the browser so we can't use that choice.) Inside you will see the `ssh` command necessary to connect to your machine. If you have Windows, there is a link to show you how to use an SSH client called PuTTY.



For mac and linux users, we will use the direct `ssh` command from the command line. It will be something like:

```
ssh -i ~/Dropbox/licenses/parrt.pem ec2-user@54.210.222.91
```

Naturally, you will have to provide the full pathname to your user.pem file.

17. Before we can connect, we have to make sure that the security file is not visible to everyone on the computer (other users). Otherwise ssh will not let us connect because the security file is not secure:

```
@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@
@         WARNING: UNPROTECTED PRIVATE KEY FILE!          @
@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@
Permissions 0644 for '/Users/parrt/Dropbox/licenses/parrt.pem' are too open.
It is required that your private key files are NOT accessible by others.
This private key will be ignored.
bad permissions: ignore key: /Users/parrt/Dropbox/licenses/parrt.pem
Permission denied (publickey).
```

Whoa! Do this:

```
$ cd ~/Dropbox/licenses
$ ls -l parrt.pem
-rw-r--r--@ 1 parrt  parrt  1696 Aug  4 15:15 /Users/parrt/Dropbox/licences/parrt.pem
```

To fix the permissions, we can use whatever "show information about file" GUI your operating system has or, from the command line, do this:

```
cd ~/Dropbox/licenses
chmod 600 parrt.pem
```

which changes the permissions like this:

```
$ ls -l parrt.pem
-rw-------@ 1 parrt  501  1696 Aug  1 12:12 /Users/parrt/Dropbox/licenses/parrt.pem
```

Don't worry if you don't understand exactly what's going on there. It's basically saying that the file is only read-write for me, the current user, with no permissions to anybody else.

18. Try to connect again and it will now warn you that you have never connected to that machine before. Again, this is a security measure. You can simply say "yes" here.

```
ssh -i ~/Dropbox/licenses/parrt.pem ec2-user@54.210.222.91
The authenticity of host '54.210.222.91 (54.210.222.91)' can't be established.
RSA key fingerprint is 49:1d:f6:ff:1a:19:5d:00:bb:cd:43:c1:84:ee:8e:a6.
Are you sure you want to continue connecting (yes/no)? yes
Warning: Permanently added '54.210.222.91' (RSA) to the list of known hosts.
```

Once you connect, you should see the following output from the terminal:

```
       __|  __|_  )
       _|  (     /   Amazon Linux AMI
      ___|___|___|

https://aws.amazon.com/amazon-linux-ami/2014.03-release-notes/
8 package(s) needed for security, out of 19 available
Run "sudo yum update" to apply all updates.
[ec2-user@ip-172-30-0-201 ~]$
```

The $ is your prompt just like you have on your local machine using the terminal / shell.

19. To get data up to the server, you can cut-and-paste if the file is small. For example, cut-and-paste the following data into a file called `coffee` in your home directory. First copy this data from the PDF:

```
3 parrt
2 jcoker
8 tombu
```

then type these commands and paste the data in the sequence:

```
$ cd ~ # get to my home directory
$ cat > coffee
3 parrt
2 jcoker
8 tombu
^D
$ cat coffee # print it back out
3 parrt
2 jcoker
8 tombu
$
```

The `^D` means control-D, which means end of file. `cat` is reading from standard input and writing to the file. The way it knows we are done is when we signal in the file with control-D.

20. For larger files, we need to use the secure copy `scp` command that has the same argument structure as

secure shell `ssh`. Get another shell up on your laptop. From the directory where you have the `coffee` file on your laptop, use the following similar command:

```
$ scp -i ~/Dropbox/licenses/parrt.pem access.log ec2-user@54.210.222.91:~ec2-user
access.log                                    100% 1363KB   1.3MB/s   00:00
$
```

Do not forget the `:~ec2-user` *on the end of that line.* The `access.log` file is at github under labs/data. From the shell that is connected to the remote server, ask for the directory listing and you will see the new file:

```
$ ls
access.log  coffee
$
```

21. Play around with your instance and then *TERMINATE YOUR INSTANCE WHEN YOU ARE DONE*, otherwise you will continue to get charged for the use of that machine. If you right-click on the instance and say "Stop", it will stop the machine and you still get charged but you can restart it without having to go through this whole procedure. If you say "Terminate", it will toss the machine out and you will have to go through this procedure again.

Deliverables

None. Please follow along in class.

Linux command line

UNIX shell is an interactive domain specific language used to control and monitor the UNIX operating system, which includes processes, devices, ram, cpus, disks etc. It is also a programming language, though we'll use it mostly to do scripting: lists of commands. If you have to use a Windows machine, the shell is useless so try to install a UNIX shell.

You need to get comfortable on the UNIX command line because many companies use Linux on their servers, which in my opinion, is best used from the command line for ultimate control over the server or cluster. For example, in the next lab we will launch Hadoop jobs from the shell. Facility with the show marks you as a more sophisticated programmer.

Everything is a stream

The first thing you need to know is that UNIX is based upon the idea of a stream. Everything is a stream, or appears to be. Device drivers look like streams, terminals look like streams, processes communicate via streams, etc... The input and output of a program are streams that you can redirect into a device, a file, or another program.

Here is an example device, the null device, that lets you throw output away. For example, you might want to run a program but ignore the output.

```
$ ls > /dev/null # ignore output of ls
```

where "# ignore output of ls" is a comment.

Most of the commands covered in this lecture process stdin and send results to stdout. In this manner, you can incrementally process a data stream by hooking the output of one tool to the input of another via a pipe. For example, the following piped sequence prints the number of files in the current directory modified in August.

```
$ ls -l | grep Aug | wc -l
```

Imagine how long it would take you to write the equivalent C or Java program. You can become an extremely productive UNIX programmer if you learn to combine the simple command-line tools.

The basics

UNIX disk structure: http://www.thegeekstuff.com/2010/09/linux-file-system-structure/

`~parrt` is my home directory, `/home/parrt`, as is ~.

```
$ls /
Applications
Library
Network
System
Users
Volumes
bin
```

```
cores
dev
etc
home
mach_kernel
net
opt
private
sbin
tmp
usr
var
```

Like when were typing in the Python shell, each command is terminated by newline. The first thing we type is the command followed by parameters (separated by whitespace):

```
$ foo arg1 arg2 ... argN
```

That is why whitespace in filenames sucks:

```
$ ls house\ of\ pancakes
```

But we can use this:

```
$ ls 'house of pancakes'
```

The commands can be built into the shell or they can be programs that we write and invoke. For example, here's how you ask which program is being executed when you type a command:

```
$which ls python
/bin/ls
/usr/local/bin/python
```

The Python interpreter is a program installed on our disk and when we say `python` at the shell, it finds the program using an ordered list of directories in `PATH` environment variable and executes it.

Next, we pass information around using streams and we can shunt that data into a file or pull data from a file using special operators. You can pretend these are like operators in a programming language like addition and multiplication. Each program has standard input, standard output, and standard error; three streams.

We can set the standard input of a process using > character:

```
$ls / > /tmp/foo
```

Here is how to type something directly into a text file:

```
cat > /tmp/foo
the first line of the file
the second line of the file
^D
$
```

The `^D` means control-D, which means end of file. `cat` is reading from standard input and writing to the file. The way it knows we are done is when we signal in the file with control-D.

We can set the standard input of a process to the contents of a file and redirect the output of a process to a file.

```
$wc < /tmp/foo
      19      19     118
```

or

```
$wc /tmp/foo
      19      19     118 /tmp/foo
```

We can connect to the output of one program to the input of another using pipes: `|`.

```
$ls / | wc # count files are in the root directory
      19      19     118
```

Here is a simple pipe (show first 5 lines of the text we stored in foo):

```
$cat /tmp/foo | head -5
Applications
Library
Network
System
Users
```

So, some programs take filenames on the command line and some expect standard input. For example, the `tr` translation command expects standard input and writes to standard output

```
$ls / | tr -d e # delete all 'e' char from output
Applications
Library
Ntwork
Systm
Usrs
Volums
bin
cors
dv
tc
hom
mach_krnl
nt
opt
privat
sbin
tmp
usr
var
```

Misc

man, help, apropos

ls, cd, pushd, popd

cp, scp

cat, more

head, tail

The most useful incantation of tail prints the last few lines of a file and then waits, printing new lines as they are appended to the file. This is great for watching a log file:

```
$ tail -f /var/log/mail.log
```

wc

Searching streams

One of the most useful tools available on UNIX and the one you may use the most is grep. This tool matches regular expressions (which includes simple words) and prints matching lines to stdout.

The simplest incantation looks for a particular character sequence in a set of files. Here is an example that looks for any reference to System in the java files in the current directory.

```
$ grep System *.java
```

You may find the dot '.' regular expression useful. It matches any single character but is typically combined with the star, which matches zero or more of the preceding item. Be careful to enclose the expression in single quotes so the command-line expansion doesn't modify the argument. The following example, looks for references to any a forum page in a server log file:

```
$ grep '/forum/.*' /home/public/cs601/unix/access.log
```

or equivalently:

```
$ cat /home/public/cs601/unix/access.log | grep '/forum/.*'
```

The second form is useful when you want to process a collection of files as a single stream as in:

```
cat /home/public/cs601/unix/access*.log | grep '/forum/.*'
```

If you need to look for a string at the beginning of a line, use caret '^':

```
$ grep '^195.77.105.200' /home/public/cs601/unix/access*.log
```

This finds all lines in all access logs that begin with IP address 195.77.105.200.

If you would like to invert the pattern matching to find lines that do not match a pattern, use -v. Here is an example that finds references to non image GETs in a log file:

```
$ cat /home/public/cs601/unix/access.log | grep -v '/images'
```

Now imagine that you have an http log file and you would like to filter out page requests made by nonhuman spiders. If you have a file called spider.IPs, you can find all nonspider page views via:

```
$ cat /home/public/cs601/unix/access.log | grep -v -f /tmp/spider.IPs
```

Finally, to ignore the case of the input stream, use -i.

Basics of file processing

cut, paste

```
$cat ../data/coffee
3 parrt
2 jcoker
8 tombu
```

cut grabs one or more fields according to a delimiter like strip in Python. It's also like SQL `select f1, f2 from file`.

```
$cut -d ' ' -f 1 ../data/coffee > /tmp/1
cut -d ' ' -f 2 ../data/coffee > /tmp/2
```

```
$cat /tmp/1
3
2
8
```

```
$cat /tmp/2
```

```
parrt
jcoker
tombu
```

paste combines files as if they were columns:

```
$paste /tmp/1 /tmp/2
3       parrt
2       jcoker
8       tombu

$paste -d ', ' /tmp/1 /tmp/2
3,parrt
2,jcoker
8,tombu
```

Get first and third column from names file

```
cut -d ' ' -f 1,3 names
```

`join` is like an INNER JOIN in SQL. (zip() in python) first column must be sorted.

```
$cat ../data/phones
parrt 5707
tombu 5001
jcoker 5099

$cat ../data/salary
parrt 50$
tombu  51$
jcoker 99$

$join ../data/phones ../data/salary
parrt 5707 50$
tombu 5001 51$
jcoker 5099 99$
```

Here is how I email around all the coupons for Amazon Web services without having to do it manually:

```
$ paste students aws-coupons
jim@usfca.edu X
kay@usfca.edu Y
sriram@usfca.edu Z
...
```

and here is a little Python script to process those lines:

```
import os
import sys
for line in sys.stdin.readlines():
    p = line.split('\t')
    p = (p[0].strip(), p[1].strip())
    print "echo '' | mail -s 'AWS coupon "+p[1]+"' "+p[0]
    os.system("echo '' | mail -s 'AWS coupon "+p[1]+"' "+p[0])
```

and here's how you run it:

```
$ paste students aws-coupons | python email_coupons.py
```

Processing log files

```
cut -d ' ' -f 1 access.log | sort | uniq -c | sort -r -n|head
```

get unique list of IPs. find out who is hitting your site by getting histogram. how many hits to the images directory? how many total hits to the website? histogram of URLs.

Python programs

```
$python resources/printargs.py hi mom
args: hi mom
```

That Python code:

```
import sys
print "args:", sys.argv[1], sys.argv[2]
```

We can use those arguments as filenames to open or we can read from standard input:

```
import sys
print sys.stdin.readlines()
```

Print coffee data out

```
$python resources/mycat.py < ../data/coffee
['3 parrt\n', '2 jcoker\n', '8 tombu\n']
```

Using the AWS Hadoop Streaming Interface

Goal

Your goal in this lab is to learn how to launch a simple map-reduce job at Amazon using their elastic map reduce mechanism and the shell/commandline. Our application is the trite "word counting," which we will use to find the most common words in a set of Google ads scarfed from the net in `ads1000.txt` at github/parrt/msan501. You'll use Python as in the other labs.

Discussion

Hadoop introduction

Hadoop is a distributed computing framework that supports a map-reduce computing paradigm. The *map* operation executes on multiple machines and gets partial results, which are then combined with the *reduce* operation.

Hadoop is written in Java and so, to use another language such as Python, we have to use the so-called *streaming interface*. That just means that we will write programs that read from standard input and write to standard output.

The *Hadoop file system* (HDFS) is a distributed file system that can handle massive amounts of data by distributing it across multiple machines and hard drives. Hadoop tries to keep map operations on the machines that store the associated data the mappers should run on. That is what typically is done, but we will be using Amazons S3 storage instead since it is the easiest thing to do.

Hadoop splits the input into chunks and splits the chunks into lines before feeding the lines to standard input of the mappers. It gives each chunk of lines to a separate mapper task, which generates partial results. The mappers generate partial results as a set of key-value pairs of the form:

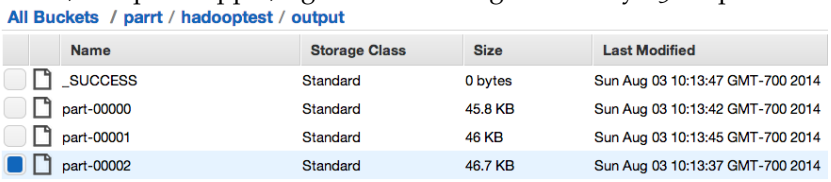
key `\t` *value* `\n`

Because partial results are created on a variety of machines where the map tasks run, hadoop has to collect this data from the machines of the cluster before giving it to the reducers. Hadoop sorts these partial results according to key (via merge sort) and distributes regions of the key space across one or more reducers. A specific key is only seen by a single reducer. A reducer reads these key-value pairs line by line from standard input and is responsible for generating a final result. That output can be whatever we want, but in our case we will use the same key-value output format. We don't have to have any reducers at all, if we just want to run mappers across all the data.

We will be using Amazon Elastic MapReduce (EMR) that will take care of all the details of launching a cluster, running our job, and creating the output files. A hadoop *job* is a chunk of work, which can have one or more tasks. If one of these tasks fails, hadoop tries to rerun them. One of Hadoop's big benefits is that it is fault-tolerant. In a cluster of 1000 computers, it's very possible a machine will go down or the system operator will kick a power cord out by mistake. AWS introduces the notion of a *step*, which is one of more jobs. We will not be using the step mechanism of the EMR GUI to launch jobs because, oddly

enough, it's much easier and quicker to do it from the command line of the master node of the cluster.

Hadoop streaming generates an output file per reducer, which can be handy if we are interested in partitioning, say, sales results per country. In that case, we would have one reducer per country. With three reducers, one per mapper, I got the following files in my S3 output folder:



To get a single output file, we need to specify a single reducer, which we will do below from the command line.

Testing map-reduce on single machine

Before spending money at Amazon to run your job, make sure that it works properly by simulating it from the command line on your laptop. To simulate hadoop collecting data and sending it as standard input to your mapper, we will use `cat`:

```
$ cat /tmp/ads1000.txt
"title" "blurb" "url" "target" "retrievetime"
"Exclusive Music For DJs" "DJ One Stop For Edits, Mash-Ups, Remixes. Browse, Listen, ...
...
```

Then, pipe that input into the mapper (which will appear like standard input to the mapper: `wcmap.py`):

```
$ cat /tmp/ads1000.txt | python wcmap.py
"title" 1
"blurb" 1
"url" 1
"target" 1
"retrievetime" 1
"Exclusive 1
Music 1
For 1
DJs" 1
"DJ 1
...
```

Hadoop always sorts the partial results coming out of each mapper before passing it to the reducer(s), which we can simulate bypassing the output of the mapper through `sort`:

```
$ cat /tmp/ads1000.txt | python wcmap.py | sort
! 1
! 1
! 1
! 1
! 1
!" 1
!" 1
!" 1
!" 1
...
```

Obviously the data is not very interesting because we have not stripped out punctuation, which you can do as an exercise. The point is that the data is sorted by key. Then, we can run the reduce job:

```
$ python wcmap.py < /tmp/ads1000.txt | sort | python wcreduce.py
...
```

The issue is that our Python program does not sort by keys when emitting key-value pairs, but we can use the command line to handle that. Here is our final command line that streams data using pipes between processes:

```
$ python wcmap.py < /tmp/ads1000.txt | sort | python wcreduce.py | sort
! 5
!" 6
"#3 3
"$189 3
"$200 1
...
```

We can also write that to file using:

```
$ python wcmap.py < /tmp/ads1000.txt | sort | python wcreduce.py | sort > output.txt
```

On a multicore machine, this process is virtually identical to what hadoop is doing for us, except of course on a smaller scale and without the network traffic.

S3 storage

AWS's elastic map reduce mechanism likes to process data out of its S3 storage. (It's tempting to try to process a local file with Hadoop from the cluster master, but that doesn't work as a local file is not available to slave nodes.) You will need to create a bucket in S3 that is unique across AWS so maybe use your user ID. My bucket is parrt. And I can access that with web address: parrt.s3.amazonaws.com. As long as I've made the folders underneath public, then I can add elements to the URL to get access to those files. For example, here is the data that I have updated for our lab in the msan501 bucket:

http://msan501.s3.amazonaws.com/data/ads1000.txt

You can load this data into your S3 bucket folder by downloading from msan501/data and uploading it into your own bucket.

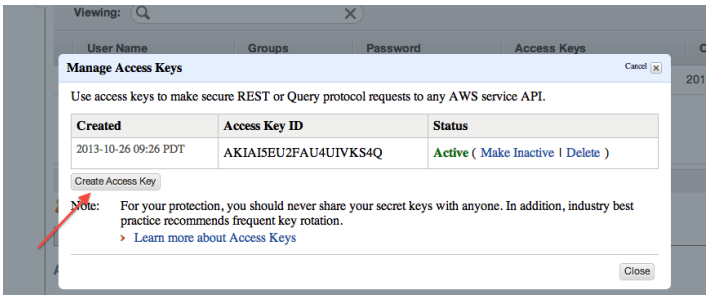
In order to access S3 from the command line of the master server of your cluster, we first need to configure the `aws` command with your access key and secret access key.

```
aws configure
aws s3 ls s3://parrt/hadooptest/
aws s3 cp s3://parrt/hadooptest/output localdir --recursive
```

Actually, I can't get this to work yet as I can't figure out how to give it my .pem file, or whatever it means to resolve the following issue: "`certificate routines:X509_load_cert_crl_file:system lib`"

Launching a cluster

Before creating a cluster, you have to create a root id. For good measure, create an access key as well:



Make sure both of them say they are active.

From the EMR console at Amazon, click on the "Create cluster" button. Choose all the defaults on the resulting page, except:

- Turn OFF "Termination protection" near the top.
- Set a log dir like `s3://parrt/hadooptest/logs` or something appropriate.
- You can delete Hive and Pig "Applications to be installed" as it slows down the launch and we aren't using them for this lab.
- Under "Security and access", select the EC2 key pair you created previously (and saved in userid.pem or whatever) or create a new one.

The default is to use three machines, one master and two core. That's fine.

At the bottom, click "Create cluster." It will take you to the status page showing you the cluster starting up. It will say Starting for a while and then it will eventually say "`Provisioning Amazon EC2 capacity`" and eventually Waiting, which means we can go to the server and launch jobs.

> If you see the error message "`No active keys found for user account`" after a few minutes, that means that you have not created the key correctly. See the stackoverflow Q/A.

After about 10 or 15 minutes, we will have a cluster up with everything installed properly to run Hadoop jobs!!!

Running a Hadoop job

On a cluster that has Hadoop installed, the easiest way to run a job is from the commandline shell. To demonstrate, we can use shell commands themselves as mappers and reducers:

```
$ hadoop jar /home/hadoop/contrib/streaming/hadoop-streaming.jar \
    -input s3://msan501/data/ads1000.txt \
    -output s3://parrt/hadooptest/output \
    -mapper /bin/cat \
    -reducer "/usr/bin/wc -l"
```

Once the job finishes, you can go to the S3 console at AWS and look in your directory for the output. Your output files, part-0000*, will have one line with just a number like 330 in it. That is the result of running wc. After downloading from S3's web interface, I can look at the results with a simple command on my local machine:

```
$ cat ~/Downloads/part-0000[0-2]
330
```

```
302
368
```

> **Warning!** If you launch a cluster and tell it to write output to an existing directory, it will fail with a permissions issue and the cluster will terminate. Consequently, use output directories with different names for each run.

Now let's run our Python code:

```
$ hadoop jar /home/hadoop/contrib/streaming/hadoop-streaming.jar \
      -files s3://parrt/hadooptest/wcmap.py,s3://parrt/hadooptest/wcreduce.py \
      -mapper wcmap.py \
      -reducer wcreduce.py \
      -input s3://msan501/data/ads1000.txt \
      -output s3://parrt/hadooptest/output2
```

We can also run with the Python files locally if we copy them up. On your local machine, do this with the appropriate pem file:

```
$ scp -i ~/Dropbox/licenses/parrt.pem wc*.py \
   hadoop@ec2-54-87-142-212.compute-1.amazonaws.com:~hadoop
```

You will also need that certificate file to communicate with the slave machines so copy that pem file to the master as well:

```
$ scp -i ~/Dropbox/licenses/parrt.pem ~/Dropbox/licenses/parrt.pem \
   hadoop@ec2-54-87-142-212.compute-1.amazonaws.com:~hadoop/.ssh
```

On the master, set the permissions of that file as we did in a previous lab:

```
$ chmod 600 .ssh/parrt.pem
```

Now that we have the Python code to the master, we need to copy the code to each of the slaves as the slaves will need to run the same code. To do that we need to know what the IP addresses of the slaves are:

```
$ hadoop dfsadmin -report | grep ^Name | cut -f2 -d:
DEPRECATED: Use of this script to execute hdfs command is deprecated.
Instead use the hdfs command for it.
10.110.197.20
10.206.58.80
$ scp -i .ssh/parrt.pem wc*.py hadoop@10.110.197.20:~hadoop
wcmap.py 100% 384 0.4KB/s 00:00
wcreduce.py 100% 677 0.7KB/s 00:00
$ scp -i .ssh/parrt.pem wc*.py hadoop@10.206.58.80:~hadoop
wcmap.py 100% 384 0.4KB/s 00:00
wcreduce.py 100% 677 0.7KB/s 00:00
```

Now, still on the master server, run the job without the `-files` option:

```
$ hadoop jar /home/hadoop/contrib/streaming/hadoop-streaming.jar \
      -mapper wcmap.py \
      -reducer wcreduce.py \
      -input s3://msan501/data/ads1000.txt \
      -output s3://parrt/hadooptest/output2
```

To run with a single reducer instead of 3, use:

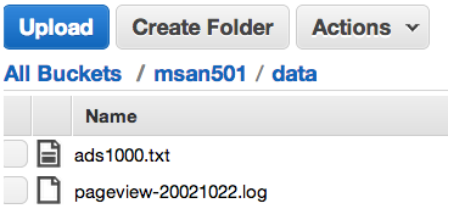
```
$ hadoop jar /home/hadoop/contrib/streaming/hadoop-streaming.jar \
   -D mapred.reduce.tasks=1 \
   -mapper wcmap.py \
   -reducer wcreduce.py \
   -input s3://msan501/data/ads1000.txt \
   -output s3://parrt/hadooptest/output3
```

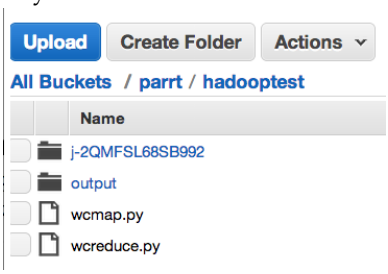
The output file will have a set of unsorted key-value pairs.

Running a job via AWS web interface

You can also run a job at Amazon without knowing the command line interface, but it's a lot of clicking. Here is the process:

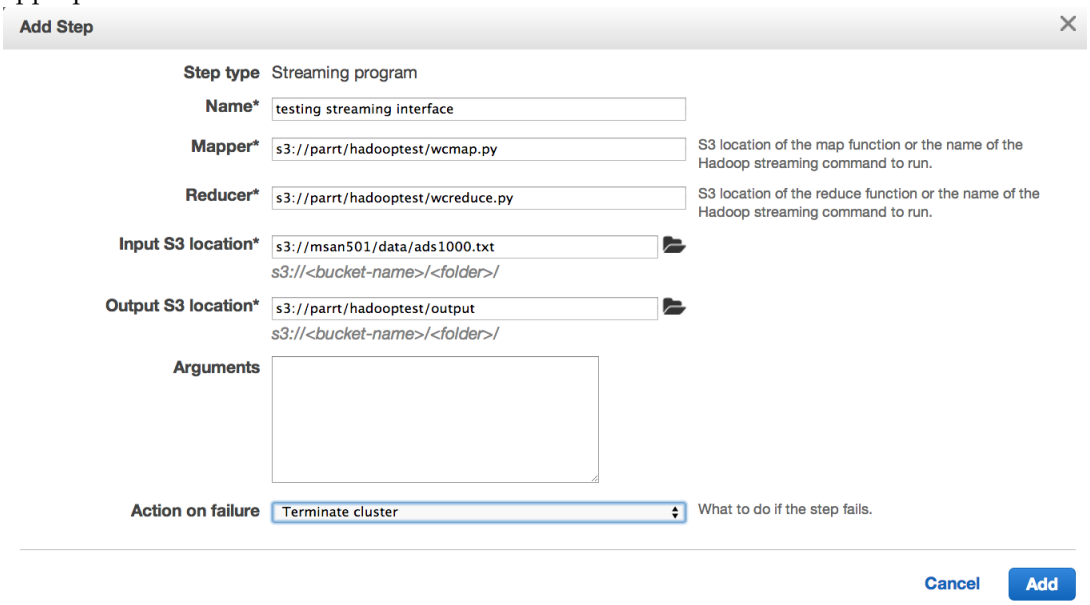
1. Load your data into an S3 bucket folder by downloading from msan501/data and uploading it into your own bucket. Load your code into S3, presumably in a different folder.





2. Create a cluster as shown previously and under "Steps" near the bottom, select "Streaming program" from the "Add step" drop-down. If you want to keep the cluster alive so that you can rerun jobs more quickly, set auto terminate to no. Otherwise set it to yes so that the cluster disappears after your job and you will not be charged further for it.

3. Enter the fields of the step dialogue as shown, substituting your user ID or your bucket/folder names as appropriate:



Warning! If you launch a cluster and tell it to write output to an existing directory, it will fail with a permissions issue and the cluster will terminate. Consequently, use output directories with different names for each run.

For convenience, here is the text so that you can cut-and-paste:

```
s3://parrt/hadooptest/wcmap.py
s3://parrt/hadooptest/wcreduce.py
s3://msan501/data/ads1000.txt
s3://parrt/hadooptest/output4
```

4. Wait about 15 minutes while Amazon creates the cluster and then wait a minute or so for your actual job to run.

5. Download or examine your data with the S3 interface.

Once our cluster is up, you can run another job "quickly" by adding another step. Go into EMR and select your cluster and then click "add step". That is a tiny link hidden down in the Steps area. Or, run a command-line job.

Extra credit

Now that you know how to run a job, it's a good idea to spend the time improving the mapper so that it strips punctuation. That way we'll get a much better set of keys. You can also strip characters not in the printable ASCII code which will automatically strip non-English characters.

Resources

- You will find `wcmap.py` and `wcreduce.py` at msan501 github. There is also the necessary data file, `ads1000.txt`.
- A helpful tutorial, from which we get our sample programs.

Deliverables

None. Please follow along in class.

Part IV

Empirical statistics

Generating Binomial Distributions

Goal

The goal of this lab is to simulate a binomial distribution using repeated Bernoulli trials and then compare it against the theoretical binomial distribution. Use filename `rbinomial.py`.

Discussion

1. First, import your uniform random number generator library and set the seed of the random number generator. (Otherwise you will always get the same Bernoulli trials.)

```
from varunif import runif
from varunif import setseed

# I defined a function in runif.py to hide implementation details
setseed( int(round(time.time() * 1000)) )
```

In this case, we're using the current time in milliseconds as the random seed so that it is different every time you run the program. (remember this trick.)

2. Next, define a function that performs n Bernoulli trials with probability p of success. It should return the number of successes out of n:

```
def binomial(n,p):
    "Sim with prob p, n bernoulli trials; return number of successes"
    ...
```

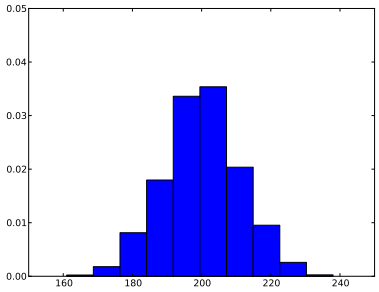
The pseudocode is just a loop that goes around n times and uses a variable from $U(0,1)$, using your `runif()` function, to check for success or failure. For example, my solution assumes failure if the uniform random variable is greater than p.

3. Now that we can know how to get a binomial random variable, we can examine the binomial distribution. All we have to do is grab a vector of, say, $SAMPLES$ binomial random values and the plot a histogram. The density function at k is just how many successes out of $SAMPLES$ there were ($k/SAMPLES$, actually).

Let me introduce you to something called a *list comprehension* in Python, which is a for loop that results in a list. It's also considered a *map* function ala *map-reduce*. Get list X as $SAMPLES$ binomial values with parameters $n = 500$ and $p = 0.4$. Do that by simply calling the `binomial` function N times.

```
X = [binomial(n,p) for t in range(SAMPLES)]
```

4. Plot the histogram normalized (normed=1) and run it. (You'll need `hist()`.) You should see a graph similar to the following:



5. We could use the built-in binomial mass function but let's define our own since it's easy:

$$\binom{n}{k} p^k (1-p)^{n-k} \qquad \text{(Binomial mass function)}$$

That's the probability that there are k successes in n trials with probability p of success. Define a function like this:

```
def binom(n, k, p):
    """
    If we run n trials with p prob for each trial of success,
    how many have k successes?
    """
    ...
```

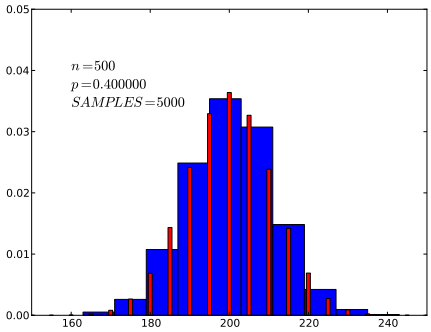
You may use function `scipy.misc.comb()` to compute $\binom{n}{k}$, but otherwise do the arithmetic yourself. (There is no loop in this function.)

6. To show the real distribution on top, we need to iterate k across the range $0..n$ used in our empirical simulation above. Since this is a mass function not a smooth density function, we can use every fifth value in the range. Let's also add some text to describe the parameters.

```
Y = [binom(n, k, p) for k in range(0,n+1,5)]
plt.bar(range(0,n+1,5), Y, color='red', align='center', width=1)
plt.axis([150,250,0,.05]) # set the axes so that we get a close-up
plt.text(160,0.04, '$n = %d$' % N, fontsize=16)
plt.text(160,0.037, '$p = %f$' % P, fontsize=16)
plt.text(160,0.034, '$SAMPLES = %d$' % SAMPLES, fontsize=16)
```

In this case I am not using 0..1 for the axes coordinates of the text; the default is the values of the graph itself. sometimes this is useful.

7. Run it and you should see something like the following:



Note that we use a bar chart for the binomial theoretical distribution and not a smooth graph because this is a mass function not a density function.

Deliverables

Please submit:

- your `rbinomial.py` file with values $n = 500, p = 0.4, SAMPLES = 5000$. Include the code to show the histogram but only run it as "main".
- submit a PDF of your final graph.
- your `varunif.py` used by your code

Generating Exponential Random Variables

Discussion

The goal of this lab is to generate random values from the exponential distribution using the inverse transform method. You will show the histogram of the random values and then show the theoretical exponential distribution on top to verify your results. You will reuse your exponential distribution random variable generator for the central limit theorem lab. Use filename `rexp.py`.

Steps

1. First, create a function called `rexp(lambduh)` that returns a random value from the exponential distribution using the inverse transform method. To do that, you need the inverse *cumulative distribution function* (CDF) for the exponential distribution $Exp(\lambda)$. The *probability density function* for the exponential distribution is:

$$p = F(x;\lambda) = \lambda e^{-\lambda x}$$

Therefore the inverse function to get the x value associated with a probability p, we use

$$x = F^{-1}(p;\lambda) = -\frac{ln(1-p)}{\lambda}$$

Your function should look like the following:

```
def rexp(lambduh): # lambduh mispelled to avoid clash with lambda in python
    # u = get value from U(0,1) then
    # return F^-1(u) for exp cdf F^-1
```

Use your `runif()` function from previous labs.

2. To plot things, you will need the usual libraries:

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

3. Get a sample of exponential random variables into variable `X` of size N from $Exp(1.5)$ using your `rexp()`. I usually define constants to make the code more readable:

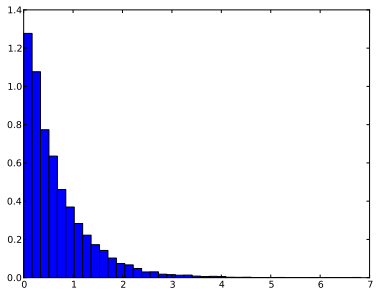
```
N = 1000
LAMBDUH = 1.5
```

then I can call `rexp(LAMBDUH)` and change LAMBDUH everywhere in my code by just changing the constant. In this case, there's no real need but it's good practice.

4. Verify that your exponential random variable behaves properly by displaying a histogram using matplotlib:

```
X = # N exponential random variables
plt.hist(X, bins=40, normed=1) # use bins option to get better resolution
plt.show()
```

You should see something like this:



How do we know that this accurately represents the exponential distribution? We plot the theoretical distribution on top with a red line.

5. Since it's easy, let's define our own exponential probability density function as follows:

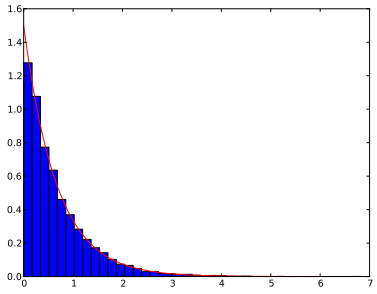
```
def exp_pdf(x, lambduh):
    ...
```

When you call it, make sure use the same lambduh.

6. Now, before the `show()`, plot the theoretical distribution so that we can see both at once:

```
# Show real distribution
x = np.arange(0,6, 0.01) # get a set of values from 0..6 stepping by 0.01
y = [exp_pdf(v, LAMBDUH) for v in x]
plt.plot(x,y, color='red')
```

You should see the following.



Deliverables

Please submit:

- your `rexp.py` file and please use the usual "if main" gate so that I can import your code for testing without creating the graph:

  ```
  if __name__ == '__main__':
  ```

- a PDF of your final graph.
- your `varunif.py` used by your code

The Central Limit Theorem in Action

Discussion

The goal of this lab is to observe how the sample means of uniform and exponential random variables have normal distributions with $N(\mu, \sigma^2/n)$ where σ^2 is the variance of the underlying distribution and n is the sample size whose mean we compute. Use filenames `clt_unif.py`, `clt_exp.py` for this lab.

Discussion

The CLT in a nutshell says that the sample mean, $X\_$, of samples X of size n from lots of distributions follows the normal distribution,specifically, $N(\mu, \sigma^2/n)$ for sample size n. In this lab will use $U(0,1)$ and the exponential distribution with $\lambda = 1.5$ and verify that using the mean as a random variable, the histogram shows a normal distribution of $N(0.5, 1/12)$ for the uniform and $N(\lambda^{-1}, \lambda^{-2}/n)$ for the exponential distribution. The mean of the uniform distribution is $\frac{a+b}{2}$ and the variance is $\frac{(b-a)^2}{12}$. The mean of the exponential distribution is $\mu = \lambda^{-1}$ and its variance is $\sigma^2 = \lambda^{-2}/n$.

The key thing here is to note that not only is the distribution of the mean random variable normal, but its variance gets tighter as we increase the sample size.

The law of large numbers says that the average of a large number of trials should approach the theoretical mean. That means that our sample mean:

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

will converge to μ as n approaches infinity with probability 1.

Also note that the number of trials we do improves the resolution of our normal distribution but doesn't change the variance.

CLT applied to uniform random variables

Steps

1. Import the usual libraries for plotting and then define these constants:

```
N = 4  # sample size (i.e, array size len(X))
TRIALS = 500 # how many samples we will take from the uniform distribution
```

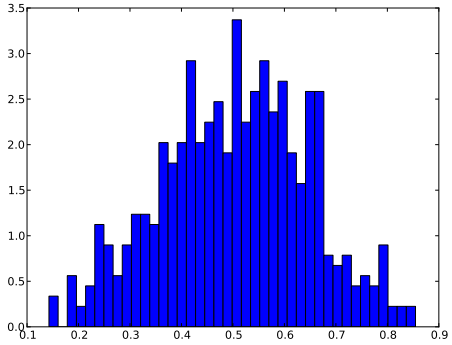
Now, we need to build a loop that gets `TRIALS` X vectors of size N with values from $U(0,1)$. Use your `runif()` function. Compute the mean of each X vector and add it to the end of a different array $X\_$.

2. Plot the histogram of `X_`:

```
# plot density of means (normalized histogram of means)
# WARNING: bins=40 is to show changes in resolution
```

```
#          where normally it's best to let the hist()
#          choose the bins for smoother view
plt.hist(X_, bins=40, normed=1)
```

3. Your histogram should look like this



Cool. It looks kind of like a normal distribution to me. Let's add the theoretical normal distribution on top. To do that we need the appropriate parameters of $Normal(\mu, \sigma^2/n)$. The mean μ of uniform samples should be midway between a and b from $U(a,b)$. In our case, that's 0.5 since we are doing $U(0,1)$. The variance of the uniform distribution is $(b-a)^2/12$ and we need the variance divided by sample size N. Define a function that returns the variance of uniform distribution $U(a,b)$:

```
def unifvar(a, b):
    ...
```

4. To get the theoretical distribution, let's define it ourselves:

```
def normpdf(x, mu, sigma): # sigma is the standard deviation, sigma^2 is the variance
    """
    Accept either a floating-point number or a numpy ndarray, such as what you get
    from arange().  You do not need a loop in the code does not change here
    because 2 * ndarray is another ndarray automatically. In this respect,
    numpy is very convenient and behaves like R.
    """
    ...
```

The function in math notation is:

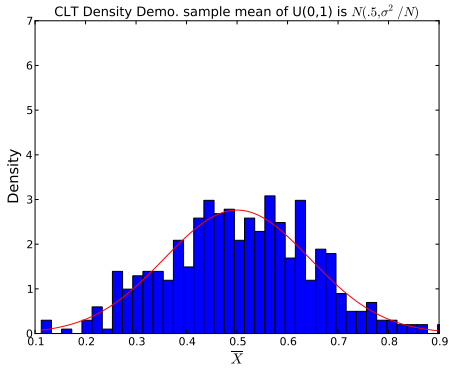
$$P(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2}$$

5. Then, plot the theoretical normal distribution on top of the histogram and set the axes so that we can use the same range throughout the next series of tests to see how the distribution changes. Note that the usual normal density function provided above expects the **standard deviation not the variance** and so we need to pass `normpdf()` the square root of the expected variance.

```
# plot real normal curve N(0.5, sigma^2/n)
x = np.arange(min(X_), max(X_), 0.01)
y = normpdf(x, 0.5, FILL THIS IN))
```

```
plt.axis([.1,.9,0,7])
plt.plot(x,y, color='red')
```

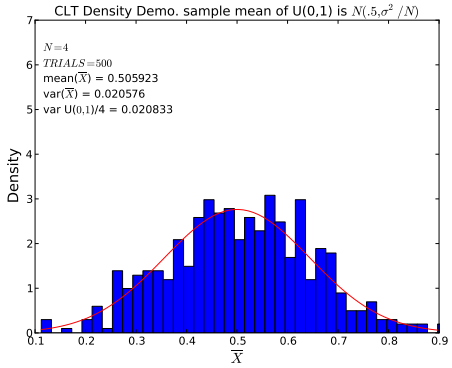
6. Run it. The resulting graph should look like this



7. Now, display some important parameters in the graph using `text()`. You will need to do that `fig.add_subplot(111)` thing again early in your script. The text in between the $ symbols is latex and lets us display nice math symbols (e.g., the title), although I'm not doing much with it here.

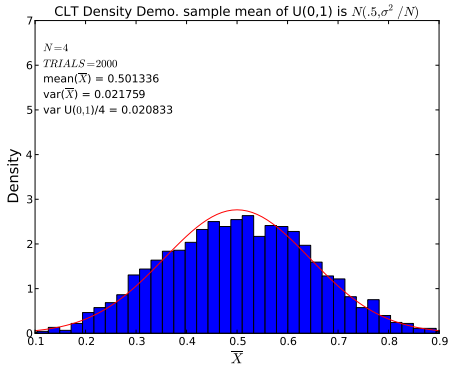
```
plt.text(.02,.9, '$N = %d$' % N, transform = ax.transAxes)
plt.text(.02,.85,'$TRIALS = %d$' % TRIALS, transform = ax.transAxes)
plt.text(.02,.8, 'mean($\\overline{X}$) = %f' % np.mean(X_), transform = ax.transAxes)
plt.text(.02,.75,'var($\\overline{X}$) = %f' % np.var(X_), transform = ax.transAxes)
plt.text(.02,.7, 'var U($0,1$)/%d = %f' % (N,varunif(0,1)/N), transform = ax.transAxes)

plt.title("CLT Density Demo. sample mean of U(0,1) is $N(.5, \sigma^2/N)$")
plt.xlabel("$\\overline{X}$", fontsize=16)
plt.ylabel("Density", fontsize=16)
```

8. Run it. The resulting graph should look like this

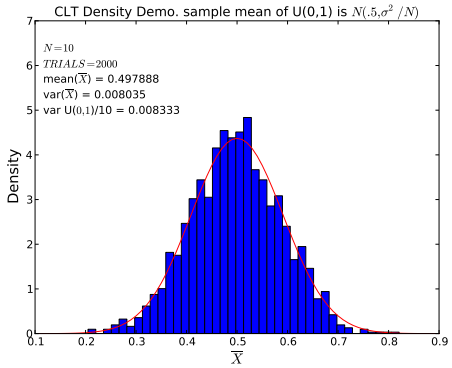


Notice how the mean is close to the expected 0.5 and that the variance of the sample mean is close to the theoretical variance.

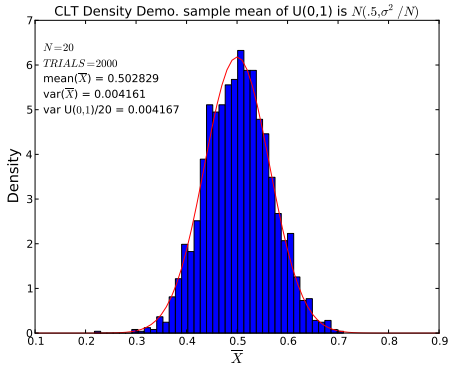
9. Increasing the number of trials two 2000 shows much higher resolution but does not change the variance/tightness of the distribution at all. Run it and see the following:



10. Now, if we increase the sample size to $N = 10$, we get a much tighter variance on the mean of $\overline{X}$. Run it:



11. Increasing to 20 we get:



CLT applied to exponential random variables

Now let's look at how the central limit theorem still gives us a normal distribution even when we pull random variables from a skewed distribution like the exponential. Create and edit a new file `clt_exp.py`.

Steps

12. Import the `rexp(lambduh)` function you wrote for the previous lab to get exponential random variables and start out with the following constants:

```
N = 10
TRIALS = 4000
LAMBDUH = 1.5
```

13. Repeat the loop we did above to get the mean of a bunch of samples into $X\_$, but this time from the exponential distribution `rexp(LAMBDUH)` instead of the uniform distribution function. Plot the histogram of $X\_$ as you did before, using a bin size of 40.

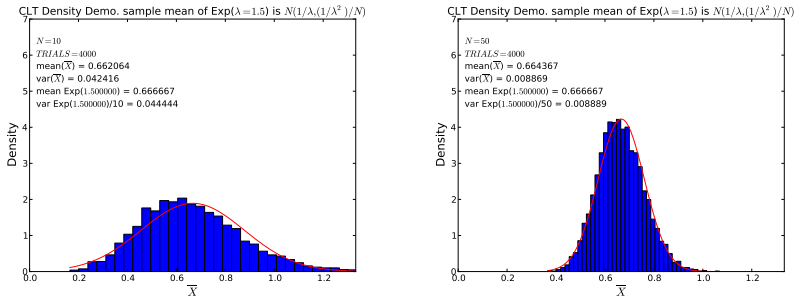
14. Plot the theoretical normal distribution on top using your `normpdf()` (you can cut/paste it into `clt_exp.py`). The mean of the exponential distribution is $\mu = \lambda^{-1}$ and its variance is $\sigma^2 = \lambda^{-2}$.

```
# plot real normal curve N(lambda^-1, sigma^-2 / N)
x = np.arange(min(X_), max(X_), 0.01)
y = normpdf(x, FILL IN MEAN, FILL IN STDDEV)
plt.plot(x,y, color='red')
```

15. Here are the appropriate text annotations:

```
plt.text(.02,.9, '$N = %d$' % N, transform = ax.transAxes)
plt.text(.02,.85, '$TRIALS = %d$' % TRIALS, transform = ax.transAxes)
plt.text(.02,.8,   'mean($\\overline{X}$) = %f' % np.mean(X_), transform = ax.transAxes)
plt.text(.02,.75, 'var($\\overline{X}$) = %f' % np.var(X_), transform = ax.transAxes)
plt.text(.02,.7, 'mean Exp($%f$) = %f' % (LAMBDUH,1/LAMBDUH), transform = ax.transAxes)
plt.text(.02,.65, 'var Exp($%f$)/%d = %f' % (LAMBDUH,N,(1/LAMBDUH**2)/N), transform = ax.transAxes)

plt.title("CLT Density Demo. sample mean of Exp($\lambda=1.5$) is $N(1/\lambda, (1/\lambda^2)/N)$")
plt.xlabel("$\\overline{X}$", fontsize=16)
plt.ylabel("Density", fontsize=16)
plt.axis([0,1.333,0,5])
plt.savefig('clt_exp-'+str(TRIALS)+'-'+str(N)+'.pdf', format="pdf")
```

16. Run it and you should see the following two graphs according to the value of N:



Notice that there is a slight leftward bias in that the normal distribution is a little bit to the right it looks like. This is to be expected. You really need to bump up N before you see it converge to the proper alignment.

17. Play around with other values of lambda and N.

Deliverables

Please submit:

- both `clt_unif.py`, `clt_exp.py` files
- your `varunif.py` used by your code
- a PDF for $N = 20$, $TRIALS = 2000$ for CLT $U(0, 1)$ demo
- a PDF for $N = 50$, $TRIALS = 4000$, $\lambda = 1.5$ for CLT $Exp(\lambda)$ demo

Generating Normal Random Variables

Discussion

The goal of this lab is to generate normal random variables but using the Central limit theorem instead of the inverse transform or the accept reject method. I'm not recommending this as the most efficient method, but it is a great practical application of the central limit theorem. The hard part about all of this is using the right variance and shifting from $N(0,1)$ to the general $N(\mu, \sigma^2)$. Use filename `rnorm.py`.

Steps

1. First, let's define some constants and the variance of a uniform variable (you should have this from the CLT lab already):

```
N = 100
TRIALS = 4000

def unifvar(a,b):
    return ((b-a)**2)/12.0
```

2. To define a function that generates normal random variables in $N(0,1)$, we rely on the fact that the sample mean, $\overline{X}$ from a sample, X, of uniform distribution values is normal. This gives us as many normal random values as we want, one per sample X. We just have to tweak things so that the mean of the distribution is zero-centered and has variance 1. That shifted and scaled value is what we return from `rnorm01()`:

```
def rnorm01():
    "return a value from N(0,1)"
    ...
```

The process looks like this:

A. Get N uniform random values from $U(0,1)$ into X using your `runif()` function.

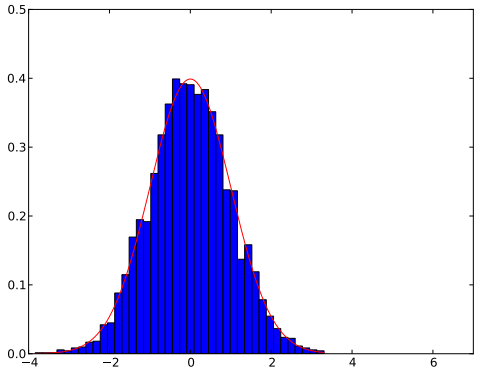
B. Then compute the mean $\overline{X}$.

C. Shift that value so that is zero-centered and call it rv.

D. We know from the CLT lab that the variance of random variable $\overline{X}$ is σ^2/N, where σ^2 is the variance of the underlying distribution $U(0,1)$, but we need the variance to be 1. Scale rv so that it has variance 1. Note that a "standard normal" variable can be created from an arbitrary normal X via $Z = (X - \mu)/\sigma$. Z is effectively a shifted and scaled version of the original. Interestingly, it really just measures how many standard deviations X is from $N(0,1)$.

3. Now, let's fill in the code we need to draw a histogram and the theoretical distribution on top using the `normpdf()` from the CLT labs:

```
# Get X taken from TRIALS trials, plot histogram normalized to density func
X = [rnorm01() for i in range(TRIALS)]
plt.axis([-4, 7, 0, 0.5]) # let's keep the same access across plot for this lab
plt.hist(X, bins=40, normed=1) # histogram should look standard normal

# plot real normal curve
x = np.arange(min(X),max(X), 0.01)
MEAN = 0
VARIANCE = 1
y = normpdf(x, MEAN, math.sqrt(VARIANCE)) # recall our normpdf takes standard deviation as variance
plt.plot(x,y, color='red')
plt.savefig('rnorm01-%d-%d.pdf' % (TRIALS,N), format="pdf")
plt.show()
```



4. Now define a more general method that accepts a desired mean and variance (*not the mean and the standard deviation*):

```
def rnorm(mean, variance):
    "return a value from N(mean,variance)"
    ...
```

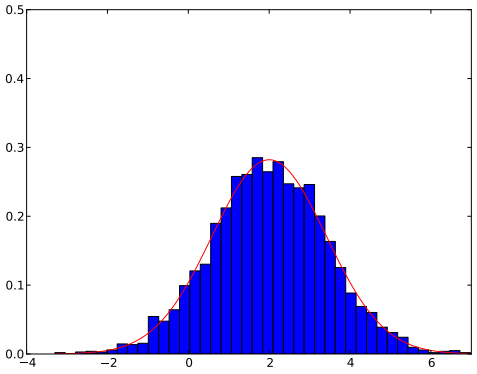
We know how to get a standard normal random variable, Z, as we just defined `rnorm01()`. To get a normal random variable with different mean and variance, we reverse the process we used to get a standard normal via $Z = (X - \mu)/\sigma$. Dust off your high school algebra and solve for X. That tells you how to shift and scale properly: $X = \mu + Z\sigma$.

5. And test as before but this time use $\mu = 2$ and $\sigma^2 = 2$:

```
MEAN = 2.0
VARIANCE = 2.0
# Get X taken from TRIALS trials, plot histogram normalized to density func
X = [rnorm(MEAN,VARIANCE) for i in range(TRIALS)]
plt.hist(X, bins=40, normed=1) # histogram should look gaussian

# plot real normal curve
```

```
x = np.arange(min(X),max(X), 0.01)
y = normpdf(x, MEAN, math.sqrt(VARIANCE))
plt.plot(x,y, color='red')
plt.savefig('rnorm-%d-%d-%d-%d.pdf' % (MEAN,VARIANCE,TRIALS,N), format="pdf")
plt.show()
```

You should get the following graph:



Deliverables

Please submit:

- your `rnorm.py` file and please use the usual "if main" gate so that I can import your code for testing without creating the graphs:

  ```
  if __name__ == '__main__':
  ```

- a PDF of the graphs shown above for $N(0,1)$ and $N(\mu = 2, \sigma^2 = 2)$.
- your `varunif.py` used by your code

Confidence Intervals for Price of Hostess Twinkies

Goal

The goal of this lab is to learn how to compute an empirical 95% confidence interval for sample means using an awesome technique called *bootstrapping*. As part of this lab, you will also learn to read in a file full of numbers. In this case, we are going to read in the price of Hostess Twinkies, a tasty snack recently returned from the dead, from around the US.

Discussion

A sample mean confidence interval of 95% tells us the range in which most (95% or 1.96σ) of the sample means fall. All we have to do is create a number of samples, X, and compute the means $\overline{X}$. If we do this lots of times (trials) then 95% of the time, we would expect the sample mean to fall within the range of 95% of the samples. We just have to order the $\overline{X}$ values and strip off the lower and top 2.5%. Then, the lowest and highest value in that stripped list represent the boundaries of the confidence interval. Cool, right?

From the central limit theorem, we know that the distribution of $\overline{X}$ is $N(\overline{X}, \sigma^2/n)$ for sample size n (not the number of trials). In this case, though we don't know what the underlying distribution is because we just got a bunch of prices from a file. We could assume that it's normally distributed, but there's no point. The central limit theorem works on any underlying distribution we care about here but we do need the variance. For that, we can use the sample variance as an estimate of the variation in the overall Hostess Twinkies price population.

The question is how do we get lots of trials from an underlying distribution that we cannot identify? By repeated sampling from our single sample *with replacement*. This is called *bootstrapping*, which you could also call *resampling*. The idea is to randomly select N values from our known data set of size N. That gives us one trial. We can then repeatedly compute our test statistic, the mean, on each sample.

To verify that we are doing the right thing, we will draw the theoretical normal distribution expected by the Central limit theorem and then shade in the 95% theoretical confidence interval, which we know is 1.96 standard deviations on either side of the mean: $\mu \pm 1.96\sigma$.

Please do your work in filename `conf.py`.

Steps

1. First, we have to get the data into a file called `prices.txt`:

```
prices = []
f = open("prices.txt")
for line in f:
    v = float(line.strip())
    prices.append(v)
```

When debugging or during development, you can print those numbers out to verify they look okay.

2. Now, we need a function that lets us sample *with replacement* from that raw data set. In other words, we need a function that gets n values at random from a data parameter (a list of numbers). It should allow repeated grabbing of the same value (that's what we call with replacement).

```
def sample(data):
    """
    Return a random sample of data values with replacement.
    The returned array has same length as data.
    """
```

The idea is to get an array of random numbers from $U(0, n)$ for `n=len(data)`. These then are a set of indices into the data array so just loop through this index array grabbing values from data according to the index. For example if you have indexes = [3,9] for a 2-element data array, then return a new array [data[3], data[9]. My solution has two lines in it.

3. Now define TRIALS=20 and perform that many samplings of prices. For each sample, create the sample mean and add it to an X_ list.

4. Sort that list and get the values from TRIALS*0.025..TRIALS*0.975 in X_ and call it `inside`.

5. Print the first and last value of the inside array as that will tell you what the bounds of your 95% confidence interval are

```
print inside[0], inside[-1]
```

You might get something like (there will be a lot of variation):

1.12295362319 1.16113333333

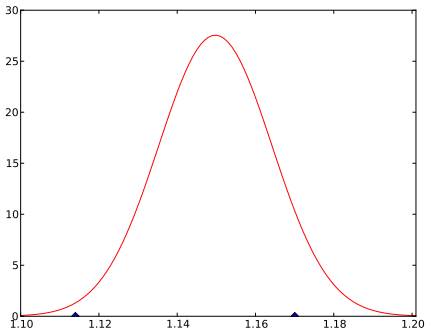
6. Add code to plot diamonds on the graph at those locations:

```
plt.plot(inside[0], 0, 'bD')
plt.plot(inside[-1], 0, 'bD')
```

7. Now plot the normal curve using your amazing new understanding of the central limit theorem. Use the following range and also set the overall graph range:

```
x = np.arange(1.05, 1.25, 0.001)
plt.axis([1.10, 1.201, 0, 30])
```

8. Run it and you should get the following graph:

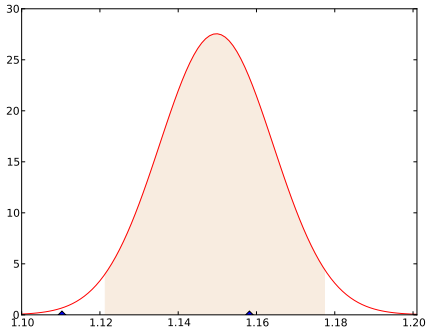


Ok, that's great but we have no idea if this is correct or not. Now, let's go nuts and show lots of stuff on the graph.

9. First, let's shade in the theoretical 95% confidence interval using your `normpdf()`.

```
mean = ...
stddev = ...
# redraw normal but only shade in 95% CI
left  = FILL IN
right = FILL IN

ci_x = np.arange(left, right, 0.001)
ci_y = normpdf(ci_x,mean,stddev)
# shade under (ci_x,ci_y) curve
plt.fill_between(ci_x,ci_y,color="#F8ECE0")
```

Run it again to see how it shades in the graph.

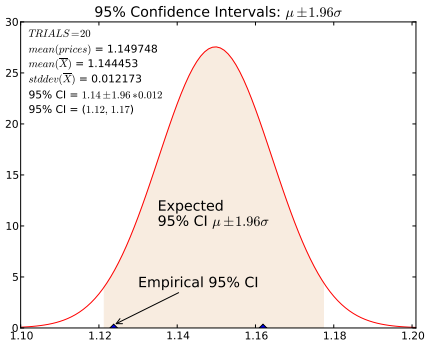


10. Now let's annotate with lots of information. Please read through and figure out what all of that stuff does to draw the nice arrows and so on.

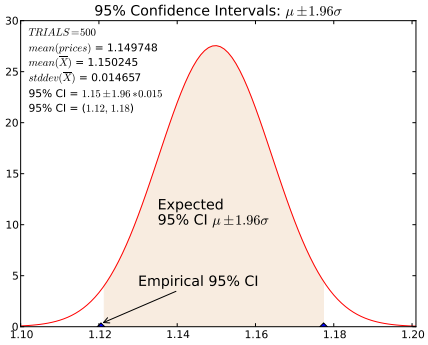
```
plt.text(.02,.95, '$TRIALS = %d$' % TRIALS, transform = ax.transAxes)
plt.text(.02,.9,  '$mean(prices)$ = %f' % np.mean(prices), transform = ax.transAxes)
plt.text(.02,.85, '$mean(\\overline{X})$ = %f' % np.mean(X_), transform = ax.transAxes)
plt.text(.02,.80, '$stddev(\\overline{X})$ = %f' %
    np.std(X_), transform = ax.transAxes)
plt.text(.02,.75, '95%% CI = $%1.2f \\pm 1.96*%1.3f$' %
   (np.mean(X_),np.std(X_)), transform = ax.transAxes)
plt.text(.02,.70, '95%% CI = ($%1.2f,\\ %1.2f$)' %
                  (np.mean(X_)-1.96*np.std(X_),
                   np.mean(X_)+1.96*np.std(X_)),
         transform = ax.transAxes)

plt.text(1.135, 11.5, "Expected", fontsize=16)
plt.text(1.135, 10, "95% CI $\\mu \\pm 1.96\\sigma$", fontsize=16)
plt.title("95% Confidence Intervals: $\\mu \\pm 1.96\\sigma$", fontsize=16)
```

```
ax.annotate("Empirical 95% CI",
            xy=(inside[0], .3),
            xycoords="data",
            xytext=(1.13,4), textcoords='data',
            arrowprops=dict(arrowstyle="->",
                            connectionstyle="arc3"),
            fontsize=16)
```

11. Run it and you should get the following graph:



12. We don't have to increase the number of trials very much before the confidence interval tightens up nicely. Try 500:



Deliverables

Please submit:

- your `conf.py` file with TRIALS=500
- a PDF of the graph with TRIALS=500 shown above.

Is Free Beer Good For Tips?

Goal

The goal of this lab is to test a hypothesis using a variety of techniques: "eyeball" test, t-test, and bootstrapping. Use filename `hyp.py`.

Discussion

Here is a typical statistics question (derived from one by Jeff Hamrick) that we will solve in multiple ways.

Q. *Psychologists studied the size of the tip in a restaurant when the waiter/waitress gave the patron a free beer. Here are tips from 20 patrons, measured in percent of the total bill: 20.8, 18.7, 19.1, 20.6, 21.9, 20.4, 22.8, 21.9, 21.2, 20.3, 21.9, 18.3, 21.0, 20.3, 19.2, 20.2, 21.1, 22.1, 21.0, and 21.7. Does a beer-inspired tip exceed 20 percent or perhaps dip below 20 percent (maybe patrons get drunk and can't do math)? Use a significance level equal to $\alpha = 0.06$.*

Side note: Always pick the significance level before you run your experiment. It is really bad mojo to pick your significance after you know what the p-value is.

Before starting on this, let's interpret that question: It asks whether the mean of the specified sample differs significantly from the usual 20% tip. By "significantly" we refer to the likelihood that the usual population (with mean 20.0) could yield a sample with the observed sample mean. By "usual" we mean our control of approximately: $N(20.0, s^2/n)$ where s is the sample variance of the sample tips and $n = len(tips)$. (We can reasonably assume that tips follow a normal distribution.)

While the population mean is 20.0, the means of any resamples we take will bounce around left and right of 20.0. The question is: does this particular test sample's mean, $m = 20.725$, fall outside of the typical variability of the sample means?

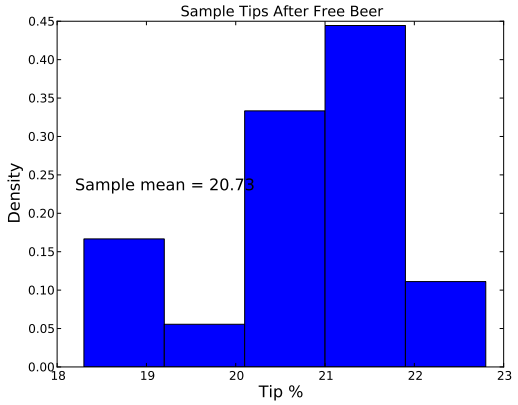
More formally, we would say the following: The **null hypothesis** is that the mean for the specified sample does not differ significantly from $\mu = 20.0$. I think of this as the *control* in my experiment. The **alternate hypothesis** is that the sample mean differs significantly above or below the population mean. Formally,

$H_0 : m = 20.0$ (non-free beer situation)
$H_1 : m \neq 20.0$ (free beer situation; two-sided alternative hypothesis)
We could also say that $H_0 : m - \mu = 0$ and $H_1 : |m - \mu| > 0$.

Steps

Eyeballing it

1. First, just draw a histogram of the tips to see what it looks like. For this exercise, create a file called `hyp.py`.

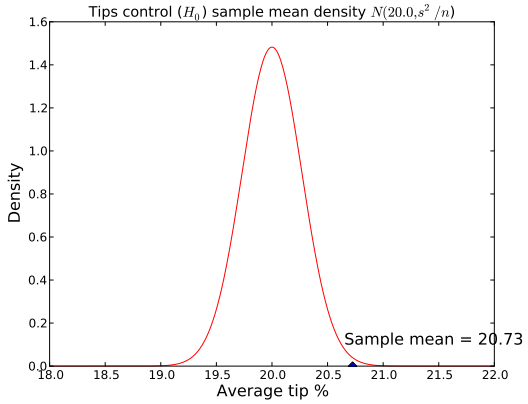


For your convenience, here are the tips in python format:

```
tips = [20.8, 18.7, 19.1, 20.6, 21.9, 20.4, 22.8,
        21.9, 21.2, 20.3, 21.9, 18.3, 21.0, 20.3,
        19.2, 20.2, 21.1, 22.1, 21.0, 21.7]
```

(Use your awesome new skills from previous labs to generate the histogram.) To me, there is a lot of "mass" to the right of the usual 20% tip but my eyeball is not a rigorous significance mechanism.

2. To get a better idea, let's simply plot the distribution of the sample means given our H_0 assumption: $N(20.0, s^2/n)$. We need to use the sample variance s^2 from our test sample because we don't know the variance of the original distribution. It safe to assume that the variance is similar. This is our "control" or the usual tipping distribution: the distribution of the set of average tips per day if H_0, the control, is true.



Looking at that graph, it seems that a sample mean of 20.73 is pretty far in the right tail of a normal curve centered at the control average 20% tip. It looks to be a few standard deviations away from the mean. My gut says that it's pretty likely that giving people a free beer increases tips significantly.

t-test

1. Let's use a *t-test* now to test for significance, just like we would do in statistics class. The t value measures the number of standard deviations a sample mean, m, is away from our presumed population mean μ:

$$t = \frac{m - \mu}{s/\sqrt{N}} \quad \text{(t-value)}$$

It's just the difference between the means scaled to be in units of standard deviations. Write some code to compute the t-value. When computing s, the sample standard deviation, note that the numpy std() function returns a biased estimate of the standard deviation. Use `np.std(tips, ddof=1)` instead of just `std(tips)`. Print out the value of t.

I get $t = 2.69417199392$. That means that m is about 2.7 standard deviations away from μ, which is a very significant departure.

2. To get a p-value, likelihood that we would see such a t value in the nonfree beer situation, look up t in a t-distribution c.d.f. using `1-scipy.stats.t.cdf(t,N-1)`. You should get 0.0071844. Since we need to check both tails, the probability is actually $2x$ that, or, p-value=0.014369 (1.4%). The definition of significance is $\alpha = 0.06$, which means that our sample mean is definitely significant since $1.4\% < 6\%$. There is only a 1.4% chance that the control could generate a value that extreme or beyond.

We must conclude that m differs significantly from $\mu = 20.0$ based upon the significance of $\alpha = 0.06$ and, therefore, we reject H_0 in favor of H_1. Giving out free beers is extremely likely to have increased the average tip in that experiment.

Boostrapping for empirical hypothesis testing

Ok, now, let's use bootstrapping to estimate a *p-value*. A p-value for some point statistic or value is the probability that the control (null hypothesis H_0) could generate that statistic or value. In our case, a p-value can tell us the likelihood that a normal distribution centered around $\mu = 20.0$ with $s^2 = var(tips)$ could generate a sample mean of 20.725. (We approximate the population variance with our sample variance.) *Note and we are sampling from $N(\mu = 20.0, s^2)$ to conjure up samples from the control situation. We are not resampling from the tips list as we are trying to see how the observed sample mean, 20.725, fits within the control distribution not the test distribution. We are also not generating samples from the distribution of a mean random variable, $N(\mu = 20.0, s^2/n)$.*

1. Bootstrap TRIALS=5000 samples of size $n = len(tips)$ from $N(\mu, s^2)$. It's very important that we use the same sample size as $len(tips)$ so we are comparing the same thing. Compute the mean of each sample, X, an add to $\overline{X}$ as you generate samples from the normal distribution.

2. Compute how many means in $\overline{X}$ are greater than or equal to mean(tips):

```
greater = np.sum(X_ >= np.mean(tips))
```

or

```
greater = sum([x>=np.mean(tips) for x in X_]) # the number of true values
```

3. The (one-sided) p-value is just the ratio of values above the observed mean, mean(tips), to the number of trials. Double that because we're doing a two-sided test. With 5000 trials, I see just 13 values greater than $m = 20.725$. That gives us a p-value of 2*13/5000 = 0.0052 or .52%. That means that, empirically, we find that there is an extremely small probability that the control could generate an extreme value like $m = 20.725$. Certainly the likelihood is less than the required 6% significant value.

Note: we would expect the empirical p-value (.52%) and the p-value derived from the t-test (1.4%) to be very close to each other when the number of trials is large with bootstrapping. Our resident statistician,

Jeff Hamrick, explains that the difference is not a problem with our bootstrapping solution and is ok.

"*A student t distribution with dof=19, is pretty close to a normal. But the differences are most greatly felt in the tails, and we're in the tails (rejection H_0), thus casting a little bit of sketchiness or your choice to draw the simulated raw data from a normal random variable. If we were performing this exact same operation on a data set with reasonably large size (say, 40 or 50 or 75) the differences would still exist but be even more minute.*"

Again, we easily reject the control and conclude that giving out free beers increases tips.

Deliverables

Please submit:

- `hyp.py`
- A text file that gives your t-value, and p-value from the t-test. Also give your empirical p-value from bootstrapping with TRIALS=5000.

Part V

Optimization and Prediction

Iterative Optimization Via Gradient Descent

Goal

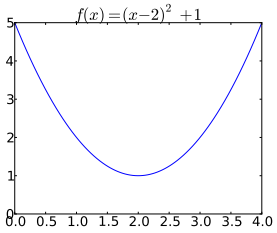
The goal of this task is to increase your programming skill by solving an iterative computation problem with nontrivial iteration and termination conditions: *gradient descent function minimization*.

Discussion

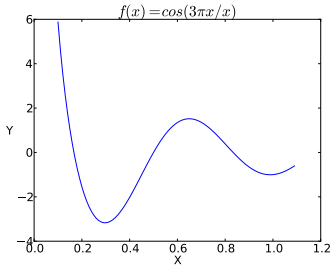
Finding x that minimizes function $f(x)$ (usually over some range) is an incredibly important operation as we use it to minimize risk and, for machine learning, to learn the parameters of our classifiers or predictors. Generally x will be a vector but we will assume x is a scalar to learn the basics. If we know that the function is convex like a quadratic polynomial, there is a unique solution and we can simply set the derivative equal to zero and solve for x:

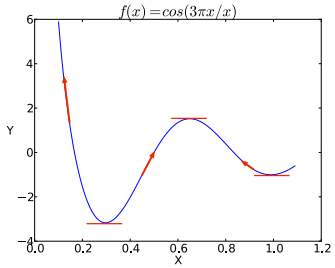
$$f'(x) = 0 \qquad \text{(Analytic solution to optimization)}$$

For example, the function $f(x) = (x-2)^2 + 1$ has $f'(x) = 2x - 4$ whose zero is $x = 2$.



We prefer to find the *global minimum* but generally have to be satisfied with a *local minimum*, which we hope is close to the global minimum. A decent approach to finding the global minimum is to find a number of local minima via random starting x_0 and just choose the minimum local minimum discovered. For example, the function $f(x) = cos(3\pi x)/x$ has two minima in $[0, 1.1]$, with one obvious global minimum:





If the function has lots of minima/maxima or is very complicated, there may be no easy analytic solution. There are many approaches to finding function minima iteratively (i.e., non-analytically), but we will use a well-known technique called *gradient descent* or *method of steepest descent*.

Gradient descent

This technique can be used to train everything from *linear regression* models (see next lab) to *neural networks*. Gradient descent requires a starting position, x_0, the function to optimize, $f(x)$, and its derivative $f'(x)$. Recall that the derivative is just the slope of a function at a particular point. In other words, as x shifts away from a specific position, does y go up or down, and by how much? E.g., the derivative of x^2 is $2x$, which gives us a positive slope when $x > 0$ and a negative slope when $x < 0$. Gradient descent uses the derivative to iteratively pick a new value of x that gets us closer and closer to the minimum of $f(x)$. The negative of the derivative tells us the direction of the nearest minimum. For example, the graph to the right above shows a number of vectors representing derivatives at particular points. Note that the derivative is zero, i.e. flat, at the minima (same is true for maxima). The recurrence relation for updating our estimate of x that minimizes $f(x)$ is then just:

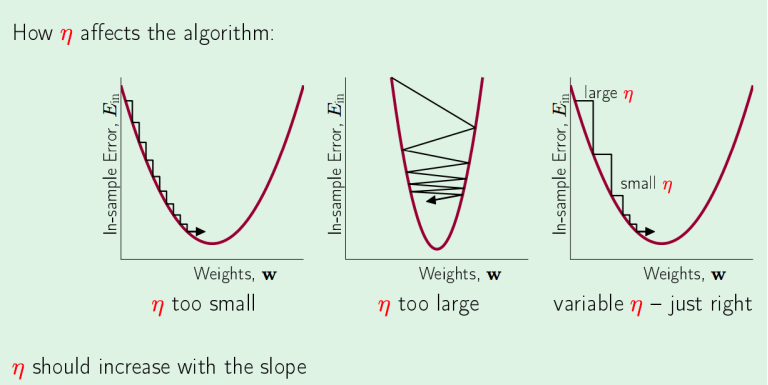
$$x_{i+1} = x_i - \eta f'(x_i)$$

where η is called the *learning rate*, which we'll discuss below. The $\eta f'(x_i)$ term represents the size of the step we take towards the minimum. The basic algorithm is:

1. Pick an initial x_0, let $x = x_0$
2. Let $x_{i+1} = x_i - \eta f'(x_i)$ until $f'(x_i) = 0$

That algorithm is extremely simple but knowing when to stop the algorithm is problematic when dealing with the finite precision of computers. Specifically, no two floating-point numbers are ever equal really. So $f'(x) = 0$ is always false. Usually we do something like $abs(x_{i+1} - x_i) < precision$ or when $abs(f(x_{i+1}) - f(x_i)) < precision$ where precision is some very small number like 0.0000001. Personally, I like the concept of stopping when there is a very small vertical change **and** x_{i+1} is heading back up.

The steps we take are scaled by the learning rate η. Yaser S. Abu-Mostafa has some great slides and videos that you should check out. Here is his description on slide 21 of how the learning rate can affect convergence:



The domain of x also affects the learning rate magnitude. This is all a very complicated finicky business and those experienced in the field tell me it's very much an art picking the learning rate, starting positions, precision, and so on. You can start out with a low learning rate and crank it up to see if you still converge without oscillating around the minimum. An excellent description of gradient descent and other minimization techniques can be found in Numerical Recipes.

Approximating derivatives with finite differences

Sometimes, the derivative is hard, expensive, or impossible to find analytically (symbolically). For example, some functions are themselves iterative in nature or even simulations that must be optimized. There might be no closed form for $f(x)$. To get around this and to reduce the input requirements, we can approximate the derivative in the neighborhood of a particular x value. That way we can optimize any reasonably well behaved function (left and right continuity would be nice). Our minimizer then only requires a starting location and $f(x)$ but not $f'(x)$, which makes the lives of our users much simpler and our minimizer much more flexible.

To approximate the derivative, we can take several approaches. The simplest involves a comparison. Since we really just need a direction, all we have to do is compare the current $f(x_i)$ with values a small step, h, away in either direction: $f(x_i - h)$ and $f(x_i + h)$. If $f(x_i - h) < f(x_i)$, we should move x_{i+1} to the left of x_i. If $f(x_i + h) < f(x_i)$, we should move x_{i+1} to the right. This is called the forward difference but there is also backward difference and a central difference. The excellent article Stochastic Gradient Descent Tricks has a lot of practical information on computing gradients etc...

Using the direction of the slope works, but does not converge very fast. What we really want is to use the magnitude of the slope to make the algorithm go fast where it's steep and slow where it's shallow because it will be approaching a minima. So, rather than just using the sign of the finite difference, we should use the magnitude or rate of change. Using finite differences then, we get a similar formula but replace the derivative with the finite (forward) difference:

$$x_{i+1} = x_i - \eta \frac{f(x_i + h) - f(x_i)}{h} \text{ where } f'(x) \approx \frac{f(x_i + h) - f(x_i)}{h}$$

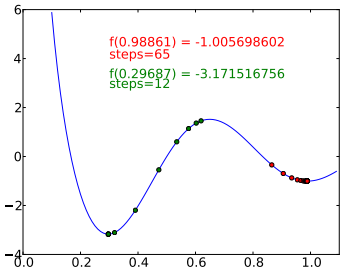
To simplify things, we can roll the step size h into the learning rate η constant as we are going to pick that anyway.

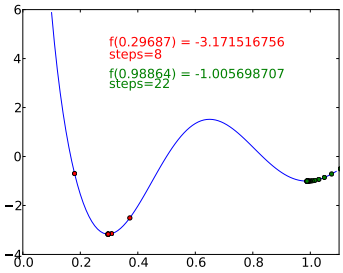
$$x_{i+1} = x_i - \eta(f(x_i + h) - f(x_i))$$

The step size is bigger when the slope is bigger and is smaller as we approach the minimum (since the region is flatter). Abu-Mostafa indicates in his slides that η should increase with the slope whereas we are keeping it fixed and allowing the finite difference to increase the step size. We are not normalizing the derivative/difference to a unit vector like he does (see his slides).

Your task

You will use gradient descent to minimize $f(x) = cos(3\pi x)/x$. To increase chances of finding the global minimum, pick **two** random locations in the range [0.1,1.2] using your `runif_` function (*don't forget to set the seed or you will get the same starting points every time*) and perform gradient descent with both of them. As part of your final submission, you must provide a plot of $f(x)$ with traces that indicate the steps taken by your gradient descent; use a different color for each descent. Here are two sample descents where the x and $f(x)$ values are displayed as well as the minimum of those two:





To create the dots you just need to add the x values to an array as you search for the minimum and then plot the x and $f(x)$ values with red or green dots:

```
tracey = [f(x) for x in tracex]
plt.plot(tracex, tracey, 'ro')
```

Please show the information as I have shown in the graphs to make it easier to compare results and for me to grade.

Define a function called `minimize` that takes the indicated parameters and returns a trace of all x values visited including the initial guess:

```
def minimize(f, x0, eta, h, precision):
    tracex = []
    tracex.append(x0)  # add starting position
    ...
    return tracex
```

Hide all of your other junk inside of the usual "main" area.

As an example, I call that function like this:

```
tracex = minimize(f, x0, ETA, STEP, PRECISION)
```

for an appropriate `f()` definition per the above cosine function. Note that Python allows us to pass a function just like any other object. For parameter `f`, we can call that function from within `minimize()` with the usual syntax `f()`.

So that we all have the same graph structure, please use the following code to plot the cosine function:

```
import matplotlib.pyplot as plt

graphx = np.arange(.1,1.1,0.01)
graphy = f(graphx)
plt.plot(graphx,graphy)
plt.axis([0,1.1,-4,6])
```

You will have to pick an appropriate step value h to get a decent approximation of the derivative through finite differences but that is large enough to avoid faulty results from lack of precision (subtracting two floating-point numbers in the computer results in a number with much less precision than the original numbers). You want that number to be small enough so that your algorithm does not oscillate around the minimum. If the number is too big it will compute a finite difference that makes x_{i+1} leap across the minimum to the other wall of the function. You must pick a learning rate η that allows you to go as fast as you can but not so fast that it overruns the minimum back and forth. When I crank up my learning rate too far, I also see the algorithm oscillate:

...

```
f(0.491296576641) = -0.166774773584 , delta = 2.0576303375622805821
f(0.296744439739) = -3.171512867583 , delta = -3.0047380939991360556
f(0.297092626880) = -3.171512816769 , delta = 0.00000005081414267138
...
```

To help you understand what your program is doing, print out x, $f(x)$, and any other value you think is helpful to see how your program explores the curve. BUT, your code shouldn't print that out in your final submission.

To give you some idea about how fast your minimization function should converge my implementation seems to converge in less than 70 steps.

Deliverables

Please submit the following via canvas:

- A PDF of your graph with two visible traces (sometimes they will overlap and you can't see one of them). It doesn't matter if they both are converging to the same minimum or two different ones. The graph should include the text I have on mine for x, $f(x)$, number of steps, etc...
- Please put all of your "main" program inside of the usual: `if __name__ == '__main__':`.
- Your `descent.py` code and `varunif.py`.

Predicting Murder Rates With Gradient Descent

Goal

The goal of this exercise is to extend the techniques you learned in the one-dimensional gradient descent task to a two-dimensional domain space, solving a *linear regression problem*. This problem is also known as *curve fitting*. As part of this lab, you will learn how to compute with vectors instead of scalars.

Discussion

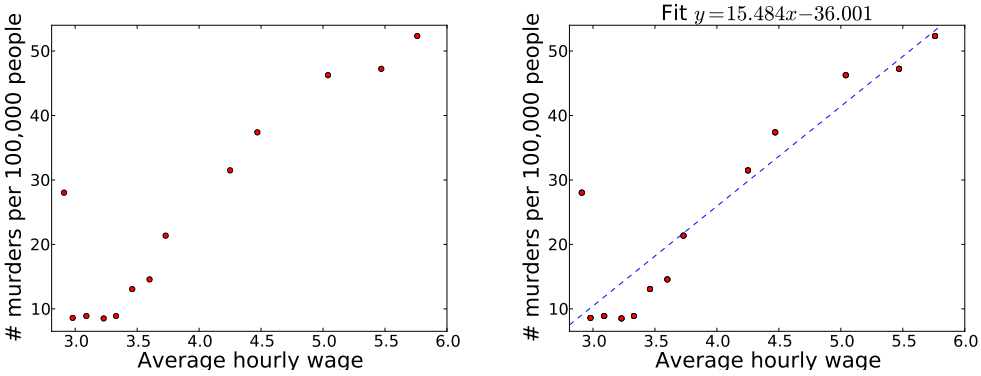
Problem statement

Given training data (x_i, y_i) for $i = 1..n$ samples with dependent variable y_i, we would like to predict y for some x's not in our training set. x_i is generally a vector of independent variables but we'll use a scalar. If we assume there is a linear relationship between x and y, then we can draw a line through the data and predict future values with that line function. To do that, we need to compute the two parameters of our model: a slope and a y intercept. (We will see the model below.)

For example, if we compare the number of murders per 100,000 people in Detroit to the average hourly wage, our eyeballs easily detect a correlation. Here is data suitable to copy and paste into Python:

```
HOURLY_WAGE = [2.98, 3.09, 3.23, 3.33, 3.46, 3.6, 3.73, 2.91, 4.25, 4.47, 5.04, 5.47, 5.76]
MURDERS = [8.6, 8.9, 8.52, 8.89, 13.07, 14.57, 21.36, 28.03, 31.49, 37.39, 46.26, 47.24, 52.33]
```

and here is a scatter plot and best fit line as determined by numpy (using `np.polyfit(HOURLY_WAGE, MURDERS,1)`).



Here, for example, $x_0 = 2.98$ and $y_0 = 8.6$.

This might be a good point to remind everyone that correlation does not equal causation. I hardly think that paying people more makes them murderous, although I could see the opposite. ;) Correlation is a *necessary* but not *sufficient* condition for causation. When you find a correlation, that gives you a candidate to check for cause-and-effect.

Best fit line that minimizes squared error

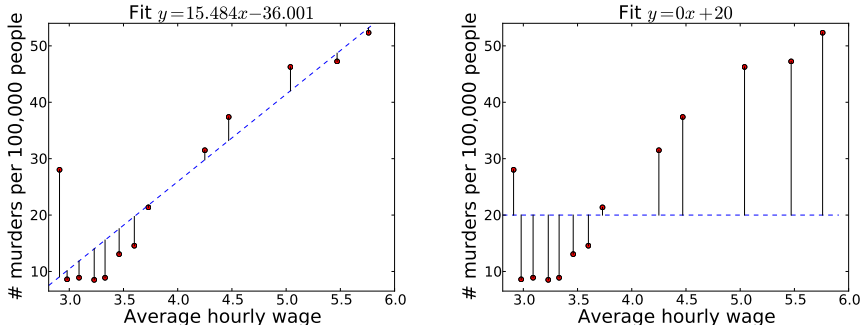
Recall the formula for a line from high school: $y = mx + b$. We normally rewrite that using elements of vector $\vec{B}$ in preparation for describing it with vector notation from linear algebra. For simplicity, though, we'll stick with scalar coefficients for now:

$$y = b_1 + b_2 x$$

The "best line" is one that minimizes some cost function that compares the known y values at x to the predicted y of the linear model that we conjure up using parameters b_1, b_2. A good measure is the *sum of squared errors*. The cost function adds up all of these squared errors to tell us how good of a fit our linear model is:

$$Cost(B) = \sum_{i=1}^{n} (\underbrace{b_1 + b_2 x_i}_{\text{linear model}} - \overbrace{y_i}^{\text{true value}})^2$$

As we change the linear model parameters, the value of the cost function will change. The following graphs shows the errors/residuals that are squared and summed to get the overall cost for two different "curve fits."



The costs are 533.82 for the left and 3563.50 for the right.

The good news is that we know the cost function is a quadratic, which is convex and has an exact solution. All we have to do is figure out where the partial derivatives of the cost function are both zero; i.e., where the cost function flattens out (at the bottom).

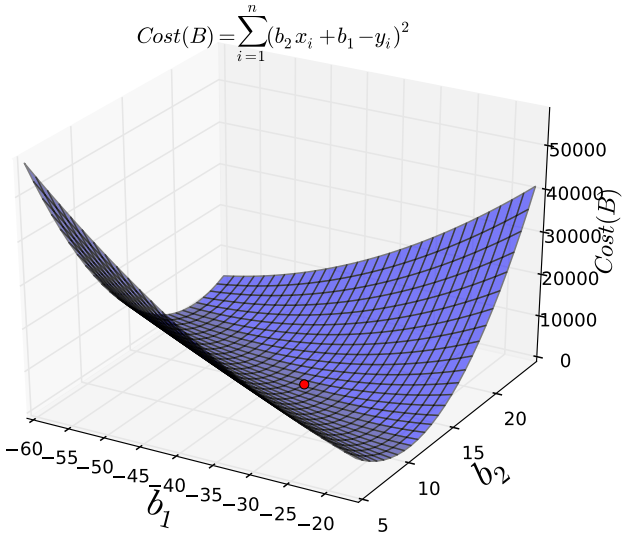
$$\nabla Cost(B) = 0 \qquad \text{(Analytic solution to optimization)}$$

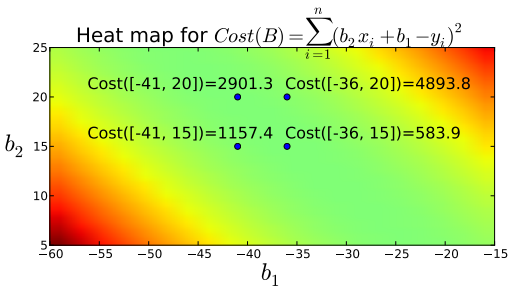
For our purposes, though, we'll use gradient descent to minimize the cost function.

To show our prediction model in action, we can ask how many murders there would be in Detroit if the average salary were \$4.7? (Obviously, these wages are from 30 years ago.) To make a prediction, all we have to do is plug $x = 4.7$ into $y = -36.001 + 14.484x$, which gives us 32.074 murders.

Gradient descent in 3D

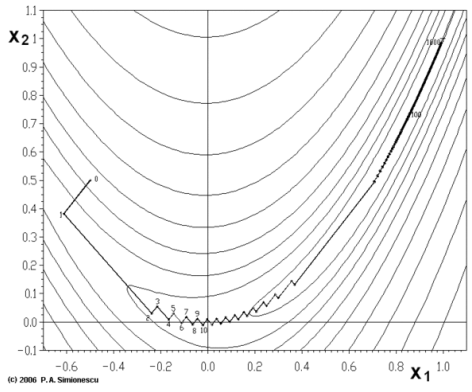
Before trying to minimize the cost function, it's helpful to study what the surface looks like in three dimensions, as shown in the following two graphs. The x and y dimensions are the coefficients of our linear model and the z coordinate is the cost function.





What surprised me is that changes to the slope of the linear model's coefficient b_2, away from the optimal $b_2 = 15.484$, cost much more than tweaks to the y intercept, b_1. Regardless, the surface is convex and a unique solution exists.

Unfortunately, based upon the deep trough that grows slowly along the diagonal of (b_1, b_2), gradient descent takes a while to converge on the minimum. We will examine the path of gradient descent for a few initial starting point. Wikipedia says that the Rosenbrock function is a pathological case for traditional gradient descent and it looks pretty similar to our surface with its shallow valley:



The recurrence relation for updating our estimate of $\vec{B} = [b_1, b_2]$ that minimizes $Cost(\vec{B})$ is the same as our previous lab but with a vector instead of a scalar:

$$\vec{B}_{i+1} = \vec{B}_i - \eta \nabla Cost(\vec{B}_i)$$

where we will approximate vectors of partial derivatives with partial finite differences defined generically as:

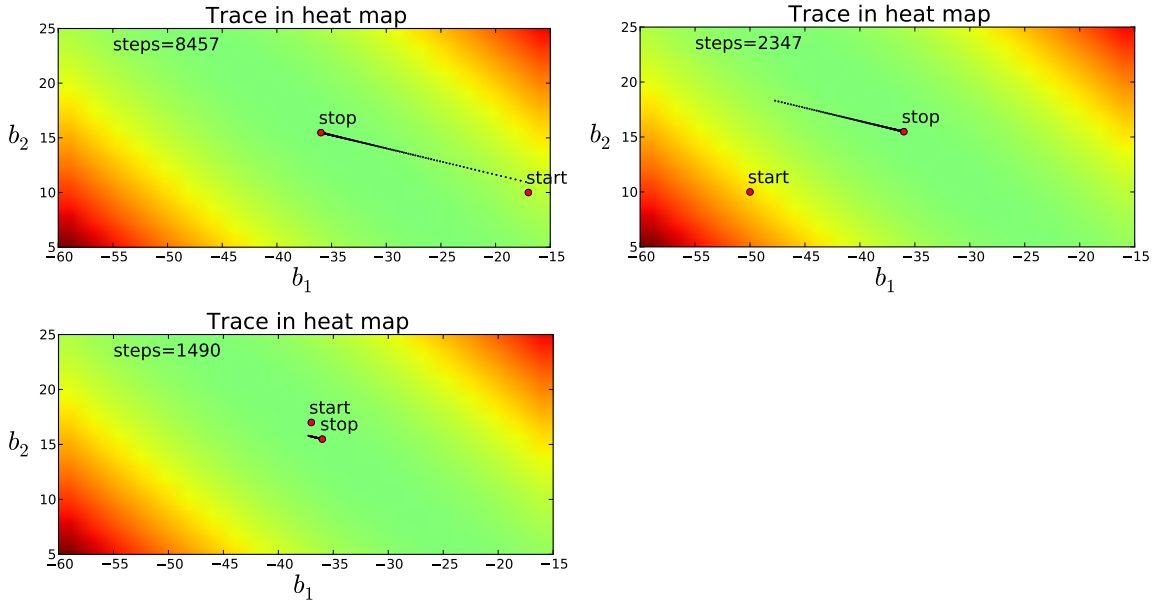
$$\nabla F(\vec{X}) = \begin{bmatrix} \frac{\delta}{x_1} F(\vec{X}) \\ \frac{\delta}{x_2} F(\vec{X}) \end{bmatrix} \approx \begin{bmatrix} \frac{F(\begin{bmatrix} x_1+h \\ x_2 \end{bmatrix}) - F(\vec{X})}{h} \\ \frac{F(\begin{bmatrix} x_1 \\ x_2+h \end{bmatrix}) - F(\vec{X})}{h} \end{bmatrix}$$

In our case, we will compute the components of a finite difference vector C' ignoring the division by the step h.

The minimization algorithm looks like:

1. Pick an initial B_0
2. let $B = B_0$
3. $C' = \begin{bmatrix} Cost(\begin{bmatrix} b_1+h \\ b_2 \end{bmatrix}) - Cost(B_i) \\ Cost(\begin{bmatrix} b_1 \\ b_2+h \end{bmatrix}) - Cost(B_i) \end{bmatrix}$
4. Let $B_{i+1} = B_i - \eta C'$
5. goto step 3 until $abs(Cost(B_{i+1}) - Cost(B_i)) < precision$ or "close enough" by some measure

Using a low learning rate, my solution takes 4843 steps starting from coordinate (-45,10) using a very small step size, which gives me a fairly decent approximation of the minimum: [-36.00066933 15.48414587] compared to the analytic solution [-36.000625 15.484375]. Starting from about the same distance away in the shallow valley at (-45,25), my solution takes 5142 steps. Cutting my step size by 5x, takes 30463 steps but gives a slightly more precise result [-36.00063497 15.48432944]. Starting at (-45,25) takes 31958 steps. Surprisingly, when I start very close to the minimum at (-36,15), my solution takes 47350 steps and does not exactly give a more accurate result. "*Your mileage may vary.*"



When I crank up the learning rate and use a very small step size, my solution converges must faster and with the same accuracy. For example, with 10 times the learning rate as before, (-45,25) converges in 3193 steps instead of 31958 steps.

Getting fancy

For the most part, the algorithm described above converges on the analytic solution with a high degree of precision, assuming that your h is small enough. In some cases however it only gets 3 digits of precision versus about six, depending on where the starting point is. To improve that accuracy, we can improve our termination condition from simply comparing previous and new value of the cost function to use the Euclidean distance between B_{i+1} and B_i. To increase convergence rate, it's a good idea to have different eta learning rates per dimension, particularly in this case because we want to speed along the valley of b_1, but not jump too far on b_2 as it is already steep; the finite difference will cause it to move along quickly.

For extremely large data sets or very high dimension data sets, you can use *stochastic gradient descent* which takes a single random or a small random subset of the samples when computing the cost function. Surprisingly, this still converges and takes much more time. I believe that the math says that the algorithm will converge no matter how many samples (don't quote me on that). Further, picking a random subset adds noise to the process, which has been shown to pull gradient descent out of local minima and get it moving again towards a lower minimum. In our case, there is an exact solution and therefore a single minimum so we would only get a speed benefit from stochastic gradient descent.

Your task

You will use gradient descent to solve the linear regression problem above, using the same data. As part of your final submission, you must provide heat maps with traces that indicate the steps taken by your gradient descent as I have shown above. Have your program choose two random starting B_0 vectors to produce your heat maps, as always using your awesome and amazing `runif_()`. Define a function called `minimize` that takes the indicated parameters and returns the minimum B parameters of your linear model, the number of steps, and the trace array of intermediate B_i values.

```
def minimize(f, B0, eta, h, precision):
    trace = []
    B = B0
    steps = 0
    while True:
        steps += 1
        if steps % 10 == 0: # only capture every 10th value
            trace.append(B)
        ...
    ...
    return (B, steps, trace)
```

As an example, I call that function like this:

```
def f(B): # a helper function that simply adds in the default arguments of the data
    return Cost(B, HOURLY_WAGE, MURDERS)
# or, if you are one of the cool kids:
f = lambda B : Cost(B, HOURLY_WAGE, MURDERS)
(m,steps,trace) = minimize(f, B0, LEARNING_RATE, h, PRECISION)
heatmap(HOURLY_WAGE, MURDERS, trace=trace)
```

Use `pylab.imshow()` to draw the heat map whose b_1 are the y intercepts, b_2 coordinates are the slopes, and heat value is the cost of x, y. In other words, create two ranges for the b_1 and b_2 coordinates and then simply compute the cost, of storing it in a matrix such as `C[b1][b2] = Cost([b1,b2],...)`. The trace is plotted on top of the map once you create it. The coordinate ranges of the heat map are not related to the trace at all. It took me a while to figure out all of the crazy methods to draw the heat maps. Plan on some frustration here. Please show the information as I have shown in the graphs to make it easier to compare results and for me to grade. Hide all of your heat map construction and function like this:

```
def heatmap(X, Y, trace): # trace is a list of [b1, b2] pairs
    ...
```

I plot the trace using:

```
plot(p[0], p[1], "ko", markersize=1)
```

You will have to pick an appropriate step value h to get a decent approximation of the derivative through finite differences that is large enough to avoid faulty results from lack of precision (subtracting two floating-point numbers in the computer results in a number with much less precision than the original numbers). You want that number to be small enough that your algorithm does not oscillate around the minimum. If the number is too big it will compute a finite difference that leads to B_{i+1} leaping across the minimum to the other wall of the function. You must pick a learning rate η that allows you to go as fast as you can but not so fast that it overruns the minimum back and forth. When I crank up my learning rate too far, I also see the algorithm go off into the weeds and stops with a minimum of [-4.86000929e+10 -2.85744570e+12].

Resources

There is a lot of material out there on the web that can be helpful.

- Finite difference at Wikipedia
- Stochastic Gradient Descent Tricks
- Numerical recipes (See Chap 10 on minimization of functions)
- Single verbal minimization in line searches
- Andrew Ng's CS229 Lecture notes
- Data analysis with Python

Deliverables

You must tweak the step size and other parameters so that your results agree with the first three decimal points of the analytic solution [-36.000625000000007, 15.484375]. (My solution is much better than that, except for a couple of weird starting positions where it only gets three decimal places.)

Please submit the following via canvas:

- A PDF of your graph with two visible traces on two heat maps or the same heat map if the traces are clear. The graph should include the start, stop location and the number of steps as I have done on mine. As part of your PDF, please indicate the B parameters you compute with your minimize function.

- Please put all of your "main" program inside of the usual: `if __name__ == '__main__':`.
- Your `regression_descent.py` code and `varunif.py`.

Part VI

Text Analysis

Summarizing Reuters Articles with TFIDF

Goal

The goal of this task is to learn a core technique used in text analysis called *TFIDF* or *term frequency, inverse document frequency*. We will use what is called a *bag-of-words* representation where the order of words in a document doesn't matter—we care only about the words and how often they are present. A word's TFIDF value is often used as a feature for document clustering or classification. We will use it simply as a document summary mechanism. The more a term helps to distinguish its enclosing document from other documents, the higher its TFIDF score.

This task is also an opportunity to practice organizing your Python code as a set of functions rather than an unstructured script (blob) with a bunch of global variables. You will also learn how to translate some simple algorithms written in pseudocode to Python code. As a practical matter, you will learn how to process XML files in Python.

Discussion

One way to summarize a text document is to list, say, the top 25 words that seem most important. That could also be used to compare documents to see if they're talking about the same thing. For example, I had to solve a problem 15 years ago to reduce noise in the forums of a Java developer's website. Users were posting stupid posts about movies and were also putting database questions in the forum on GUIs. The goal was to detect non-Java posts and also to detect misplaced posts. What does it mean to "talk about Java"? How do I know when someone is talking about databases versus GUIs? My solution was to identify the words important to Java as a whole ("Java-speak"), database, and GUI posts. Any posts that did not have words important to Java, were tossed out as irrelevant after giving them a mild smack on the snout. Similarly, posts without words relevant to databases were compared to vocabularies associated with other topics to see if another forum would be more appropriate. To make this work, I needed a precise definition of "important words." As I did for that project, you will use a classic text analysis technique called TFIDF in this project.

Certainly a word is important to a document if it's used a lot, but that would also include words like "the" so we need to discount words used frequently among our *corpus* (set of documents). So, we boost words used frequently in a document but attenuate words that are used in a lot of documents. For more on this topic, see Introduction to Information Retrieval.

The *term frequency* is just the term count within a document divided by the number of words in that document (some people use "frequency" to mean "count" but that is an affront to the gods):

$$tf(t,d) = \frac{count(t), t \in d}{|d|} \qquad \text{(Term frequency of term } t\text{, document } d\text{)}$$

A term's *document frequency* is the count of documents containing that term divided by the total number of documents:

$$df(t, N) = \frac{|\{d_i : t \in d_i,\ i = 1..N\}|}{N} \qquad \text{(Document frequency of } t \text{ in } N \text{ documents)}$$

We can think of the document frequency as the probability of seeing t in a document.

In order to attenuate the TFIDF scores for terms with high document frequencies, we need the document frequency in the denominator:

$$tfidf(t, d, N) = \frac{tf(t, d)}{df(t, N)} \qquad \text{(First approximation to TFIDF)}$$

This formula is meaningful but gives a poor weight because the document frequency tends to overwhelm the term frequency in the numerator so we take the log of the denominator first. Here's the formula slightly rewritten as it is normally shown:

$$tfidf(t, d, N) = tf(t, d) \times log(\frac{1}{df(t, N)}) \qquad \text{(TFIDF with attenuated document frequency)}$$

When t is in every document, idf is $log(1/df(t, N)) = log(1) = 0$. When t is in very few documents, such as $1/10^8$, idf is $log(10^8)$, which is about 18.4.

> **Aside.** To prevent division by 0 errors when a term does not exist in a corpus (e.g., $df(t, N) = 0$ in search applications where we pass unknown term(s) t), we can simply add 1 to the denominator. This is similar to *additive smoothing* that you will see when estimating term probabilities in document classifiers. The technique is like pretending there is an imaginary document with every unknown word (and, indeed, every possible word). To keep document frequencies in $[0..1]$, we can to bump the document count, N, as well.
>
> $$df(t, N) = \frac{|\{d_i : t \in d_i,\ i = 1..N\}| + 1}{N + 1} \qquad (df \text{ with smoothing})$$
>
> For example, if we have a vector of words in a search query $[the, apple, cat, foo]$ aggregated from 100 documents, we might get a set of document frequencies like this:
>
> $$[\frac{100}{100}, \frac{4}{100}, \frac{9}{100}, \frac{0}{100}]$$
>
> With smoothing, we would get:
>
> $$[\frac{101}{101}, \frac{5}{101}, \frac{10}{101}, \frac{1}{101}]$$
>
> This is like converting zeros to some really small number (assuming N is large) and adding that same small number to the other document frequencies.

To summarize a document, we can order its terms by *tfidf* in reverse order and look at the top 20 words, for example. To get the lexicon of a topic like databases, we can collect a known set of database posts into a single document and compute the tfidf in association with the aggregated documents of the other topics. Any word below a certain threshold, that we find by eyeballing it, is considered not relevant to that particular topic.

For example, here is a set of terms and the associated *tfidf* computed from a sample Reuters article. It's clear that it's talking about Nielsen ratings for news programs, without even looking at the original article.

Term	*tfidf*
rating	0.12332931962551781
fox	0.11911646171233138
nbc	0.11911646171233138
homes	0.11408838230482544
cbs	0.0794109744748876
audiences	0.0794109744748876
neilsen	0.0794109744748876
evening	0.06678324701893232
abc	0.06678324701893232
watching	0.0634765565309808

To compute TFIDF, we need an overall index that maps term t to document frequency $df(t, N)$ for all t in all N documents and an index that maps document d to another index that maps each $t \in d$ to $tf(t, d)$. From that, we can compute all of the TFIDF scores. That is what you will do for this project, as described in the next section.

Your task

To implement this project, you have six key functions to implement. Four of them were in the pseudocode shown in floating boxes interspersed below. You must translate them to Python in a file called `tfidf.py`.

Function: *words*(document d)
Input: Document d
Result: non-unique list of words $wordlist$
  Replace numbers, punctuation, tab, carriage return, newline with space
  $wordlist$ = Split d into words
  Strip out $w \in wordlist$ smaller than 3 letters
  Normalize $w \in wordlist$ to lowercase
  return *wordlist*

You must also provide a function called `filelist(pathspec)` that returns a list of all files that match `pathspec` and that *have non-zero file sizes*:

```
def filelist(pathspec):
    ...
    return files
```

For example, I might pass in string `../data/reuters-vol1-disk1/*.xml`. Naturally, if this doesn't work, then the rest of the code will not work as it won't get the proper data. That function should only consider the files in the specified directory, not subdirectories. You will probably want to use Python function `glob.glob()`.

To process XML files, you should also create a function called `get_text()` to open a file, load it as XML, find the title and text elements and return that combines text as a string. It's important that we all follow the same text normalization so that we get the same word list and hence TFIDF scores for comparison.

```
Function: create_indexes(list of files)
Input: List of filenames files
Result: (Map document name to Counter object mapping term to
         frequency map tf_map, Counter object mapping term to
         document count df)
  df = Counter(); tf_map = {}
  foreach f in files do
      d = get_text(f)
      words = words(d)
      n = len(words)
      tf = Counter(words)
      # walk unique word list
      foreach t ∈ tf do
          tf[t] = tf[t]/n # convert to a term freq from count
          df[t] += 1      # not currently a frequency; it's a count
      end
      tf_map[f] = tf
  end
  return (tf_map,df)
```

```
Function: doc_tfidf(tf, df, N)
Input: Term to frequency map tf
Input: Term to document count map df
Input: Number of documents N
Result: Map of each term in doc (tf) to TFIDF score
  tfidf = {}
  foreach t ∈ tf do
      df_t = df[t]/N
      idf_t = 1/df_t
      tfidf[t] = tf[t] × log(idf_t)
  end
  return tfidf
```

```
Function: create_tfidf_map(files)
Input: List of xml filenames files
Result: Map from file to map of term to TFIDF score
  (tf_map, df) = create_indexes(files)
  tfidf_map = {}
  foreach f ∈ files do
      tfidf = doc_tfidf(tf_map[f], df)
      tfidf_map[f] = tfidf
  end
  return tfidf_map
```

```
def get_text(fileName):
    """
    Read an xml file and return the text from <title> and <text>.
    Concatenate those two elements, putting a space in between so it doesn't
    form an incorrect compound word.
    """
    ...
    return text
```

As part of your development work you will use lots of maps that look like {dog: 36, cat: 19, ...}. Those integers, such as term counts, are easy to compute yourself but Python has an object that is effectively a histogram called `Counter`. For example, if you give it a list of words, it will return an object that maps terms to their count. When you print them out, it will do so in reverse order of term count, which is very handy for testing. Further, the unit tests I provide expect Counter objects.

For what it's worth, my implementation is just 60 lines including the `import` statements. This is not a huge project but it is tricky when messing around with all of these maps of maps and lists of things. Start by understanding the problem and working a few TFIDF examples manually. Then, build a simple functions and test them individually before moving on to the more complex functions. For example, you should start by building `filelist()` and then probably `get_text()`. My typical strategy is to design from the top down and test from the bottom up.

XML Input

As part of this project, I will provide you with a set of Reuters news articles in XML format, which will be the input to your program. From it, you will create the appropriate indexes and I will test those values against what I computed with my solution.

The format of the files doesn't matter much except that you need to pull out the `title` and `text` tags. The `p` paragraph tags inside `text` need to be collected. All of this text is what you will return from `get_text()`.

```
<?xml version="1.0" encoding="iso-8859-1" ?>
<newsitem itemid="131701" ...>
<title>German consumer confidence rises in Aug/Sept</title>
<text>
<p>German consumer confidence rose...</p>
<p>The Icon index, which...</p>
</text>
...
</newsitem>
```

The collection of Reuters articles is considered proprietary to Reuters and, to get access to the data, the faculty had to promise Reuters the data would not be made available on a public website or given to anyone else. Please treat this data with care, do not posted to github, etc.

Testing

In computer science, programmers recognize two primary kinds of tests: *unit tests* and *functional tests*. A unit test is really just testing a function or a few functions whereas functional tests test the overall functionality of the program. In file `test_tfidf.py`, I have provided a set of unit and functional tests that you can use for basic sanity checking of your TFIDF project. I would typically test your projects with a different set of unit tests but, in this case, we will define success as getting the correct answers for the large corpus of Reuters articles that I will provide to you.

To make the unit tests work, make sure that you install py.test, which is usually just a matter of:

```
easy_install -U pytest
```

I will test your code using the following command line (with your `tfidf.py` is in the same directory):

```
$ python -m pytest test_tfidf.py
============================= test session starts ==============================
platform darwin -- Python 2.7.7 -- py-1.4.20 -- pytest-2.5.2
collected 5 items

test_tfidf.py .....

========================== 5 passed in 0.04 seconds ===========================
```

If you don't see all tests passing, and there is a problem at a basic level with your software.

Note that the test file imports your file with:

```
from tfidf import *
```

If you name it incorrectly, the program won't work.

To test the entire corpus of Reuters articles, I will run your program as follows, potentially with a different path specification.

```
python test_corpus.py '../data/reuters-vol1-disk1/*.xml'
```

Note that the quotations around the path specification are required to prevent the command line from expanding `*.xml`. You want that path specification to go in as a single argument, not a list of files.

The core of `test_corpus.py` is:

```
(file_to_histo, word_to_numdocs) = create_indexes(files)
for f in files:
    pairs = doc_tfidf(file_to_histo[f], word_to_numdocs, N)
    # convert map to a Counter object so we can use most_common()
    term_pair = Counter(tfmap).most_common(1)[0]
    print os.path.basename(f), "(\%s, \%1.4f)" \% (term_pair[0], term_pair[1])
```

The output, which I have provided in file `corpus_output.txt.7z`, starts with (there are actually more files than 81880, but they are mysteriously empty):

```
81880 files
131674newsML.xml (ewe, 0.7714)
131675newsML.xml (tisa, 0.2909)
131676newsML.xml (ingenico, 0.4040)
131677newsML.xml (lisbon, 0.1876)
```

```
131678newsML.xml (drachmas, 0.0844)
131679newsML.xml (satisfying, 0.1774)
13167newsML.xml (cents, 0.3350)
131680newsML.xml (tightness, 0.0891)
131681newsML.xml (tisa, 0.3626)
131682newsML.xml (intervention, 0.1766)
131683newsML.xml (nordic, 0.1249)
131684newsML.xml (oilseeds, 0.1030)
131685newsML.xml (crowns, 0.1299)
131686newsML.xml (trelleborg, 0.2399)
131687newsML.xml (nni, 0.4351)
131688newsML.xml (nantes, 0.3898)
131689newsML.xml (advances, 0.4745)
13168newsML.xml (utilicorp, 0.3165)
131690newsML.xml (sas, 0.2201)
131691newsML.xml (austria, 0.0869)
131692newsML.xml (wage, 0.1244)
131693newsML.xml (herzog, 0.2330)
...
```

My implementation takes about 1 minute 30 seconds to compute TFIDF scores for 81880 XML files loaded from an SSD on a fast machine. Loading those files takes just 5 seconds.

Resources

I provide for you the following files:

- `test_tfidf.py`: some simple tests using py.test.
- `test_corpus.py`: prints out the file, term, and TFIDF score for the highest scoring term in each file.
- `corpus_output.txt.7z`: compressed output from running test_corpus.py
- `reuters-vol1-disk1-subset.7z`: compressed directory full of XML files—the corpus. It is 385M when uncompressed.

Deliverables

Please submit the following file via canvas:

- Your `tfidf.py` file. *I will deduct a full point if your library is not executable exactly in the fashion mentioned in this project; that is, method names and filename must be exactly right. For you PC folks, note that case is significant for file names on unix! All projects must run properly under linux or OS X.* Please make sure that there is no extraneous output generated by your code.

Extra credit — Search Engine

In this project, we created an index from term to the number of documents that contain that term. If we extend that to be an index from term to the list of documents containing the term, we can get the same

results as we did before. The benefit would be that we could also create a search engine.

Given a query such as "consumer confidence," we could merge the list of files containing those two terms and display those to a user. It's fast and works great! The only problem is that we might get 1000 documents back and we'd really like to show the most relevant documents first. Using the tf_map index, we can compute a relevance score for a query, relative to a document, by summing the TFIDF scores for each term in the query that is present in the document. The document with the highest two TFIDF scores for "consumer confidence," would be the first document we displayed.

You need to modify the df map from above, use additive smoothing to handle unknown words, and then implement the following function.

```
def search(query): # query is a string with a list of words
    docs = []
    # find list of documents for each term in query
    # docs = intersection of these files
    # compute sum of TFIDF scores for each term in query relative to each document in docs
    # sort documents by reverse score
    # Returns a list of document filenames in reverse TFIDF order
    return docs
```