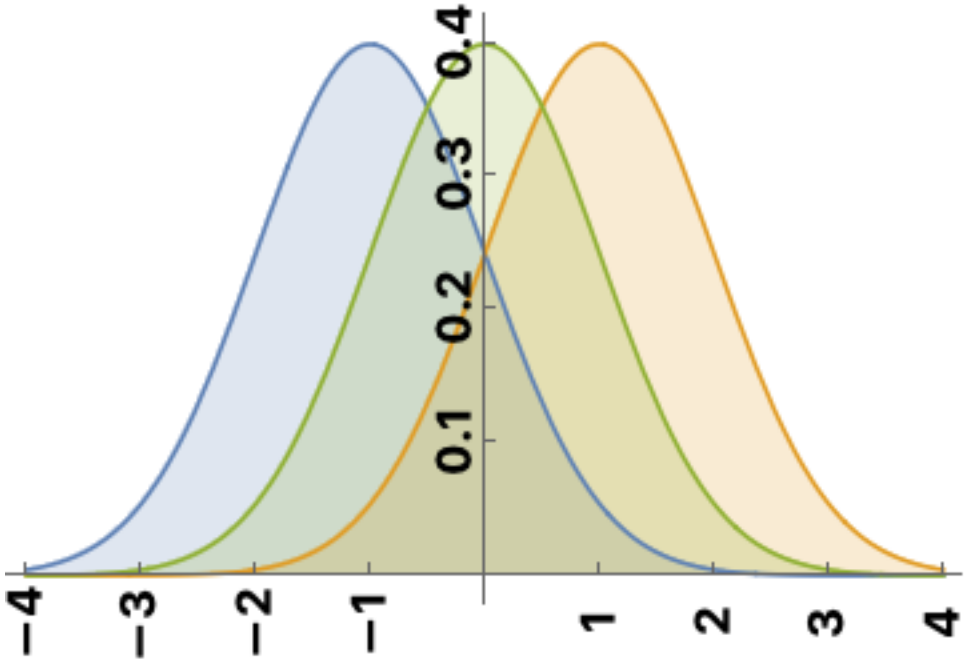
MultiArmedBandit

March 7, 2021

Multi-Armed Bandit Model

For this experiment there will be three bandits with differing statistical properties. The first bandit, bandit[0], will have a mean reward of zero, the second bandit, bandit[1], will have a mean reward of one, and the third bandit, bandit[2], will have a mean reward of minus one, which is a penalty. The standard deviation of the distributions are all set to one. The figure below shows these three probability distributions. Obviously, the green distribution corresponds to bandit[0], the yellow/orange distribution corresponds to bandit[1], and the blue distribution corresponds to bandit[2].



You are presented these three bandits not knowing the which bandit is which. Moreover, you do not know what the uderlying probability distributions are. You are given a fixed number of pulls of one of the arms. On each iteration, you select an arm, pull it, and get a reward. Your task is to maximize the reward, also known as the **value**, that you receive. To maximize your reward you must devise a strategy, also known as a **policy**, for deciding which are to pull on each iteration.

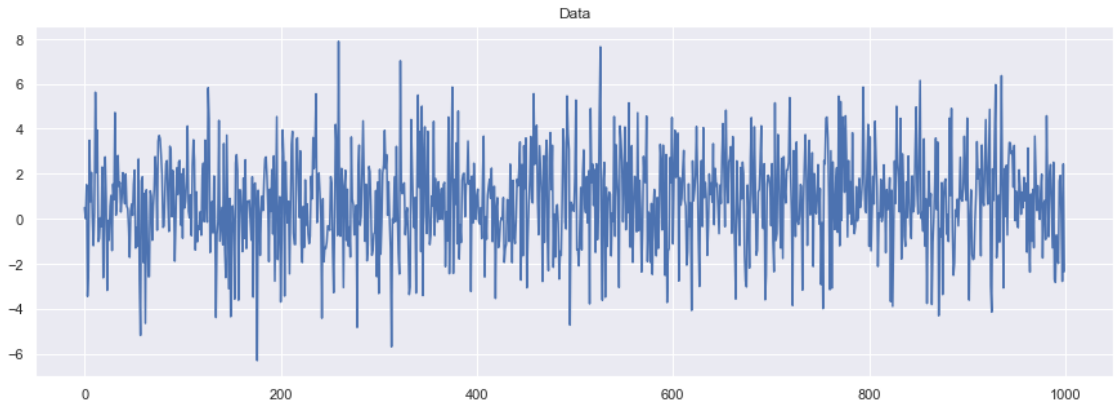
Here the policy will be what is called an epsilon greedy policy. Current mean values will be kept and updated with each pull of each bandit. With probability 1-eps, the bandit with the largest

mean will be selected on an iteration. This is exploitation. With probability eps, one of the three bandits is chosen at random. This is exploitation.

[1]:
```
import random
import numpy as np
%matplotlib inline

N = 1000
seed = 42 # Set to None to use system time
random.seed(seed)

eps = 0.25

mu = np.empty(3)
mu[0] = 0
mu[1] = 1
mu[2] = -1

sigma = np.empty(3)
sigma[0] = 2
sigma[1] = 2
sigma[2] = 2
```

[2]:
```
from run_experiment2 import run_experiment

run_experiment(mu, sigma, N, eps)
```

```
Bandit 0 - pulled 81 times. Estimated mean - 0.06636686332756062
Bandit 1 - pulled 832 times. Estimated mean - 1.0406786277940538
Bandit 2 - pulled 87 times. Estimated mean - -1.16344137841036
Total reward   -  769.4663567840236
Average reward -  0.7694663567840235
```




Average Reward
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
0
200
400
600
800
1000

[ ]: