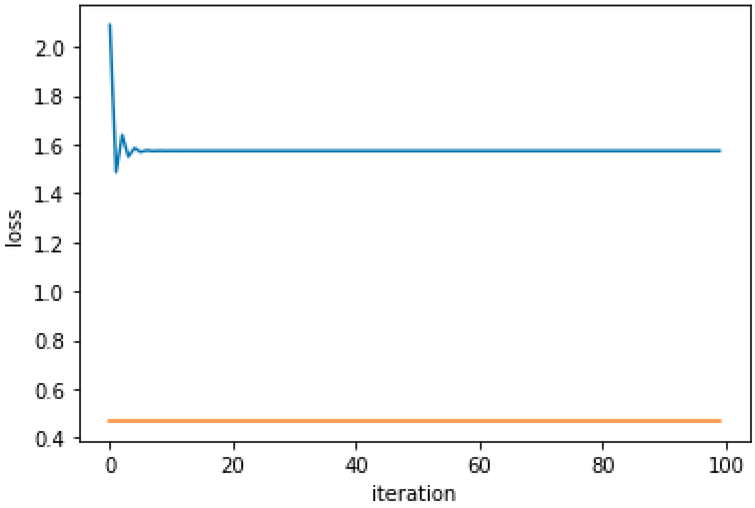
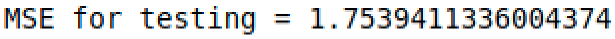
309553051
多媒體工程所
林畊廷

Part.1

1. & 2. The learning curve is below. The blue line is using the mini-batch gradient descent with mean square error. The orange line is using the mini-batch gradient descent with mean absolute error.

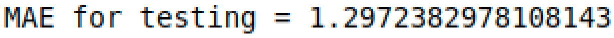


3. Mean square error is below

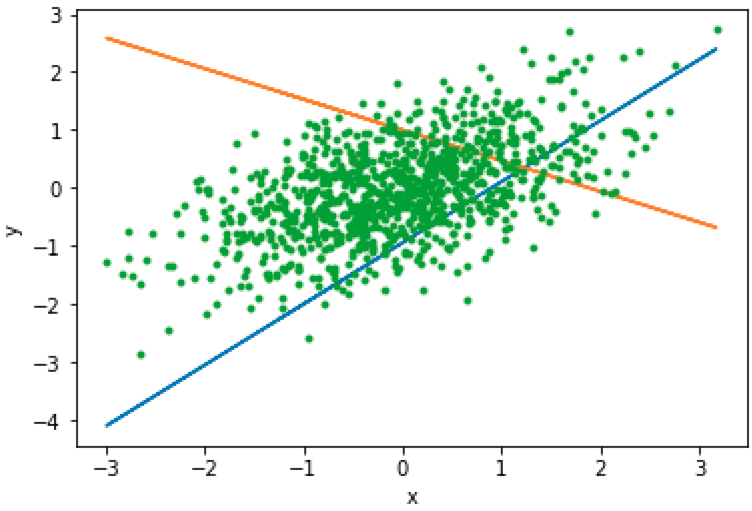


```
MSE for testing = 1.7539411336004374
```

Mean absolute error is below

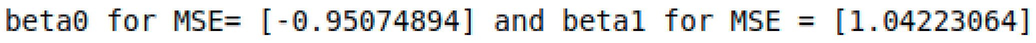


```
MAE for testing = 1.2972382978108143
```



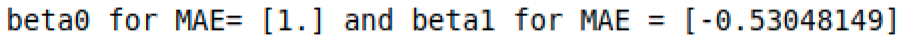
The green points are the ground truth.
The blue line is the linear regression of gradient descent which using mean square error as the objective function, and the orange line is the linear regression of gradient descent using mean absolute error as the objective function.

4. The weights (β1) and intercepts (β0) which using mean square error as the objective function is



```
beta0 for MSE= [-0.95074894] and beta1 for MSE = [1.04223064]
```

The weights (β1) and intercepts (β0) which using mean absolute error as the objective function is



```
beta0 for MAE= [1.] and beta1 for MAE = [-0.53048149]
```

5. The difference between gradient descent, mini-batch gradient descent, and stochastic gradient descent.

- Gradient Descent

The gradient descent can view as mini-batch gradient descent when the mini-batch size is equal to the training data size. The gradient descent used a fixed learning rate to whole training data. After an iteration, the weights will update.

- Mini-Batch Gradient Descent

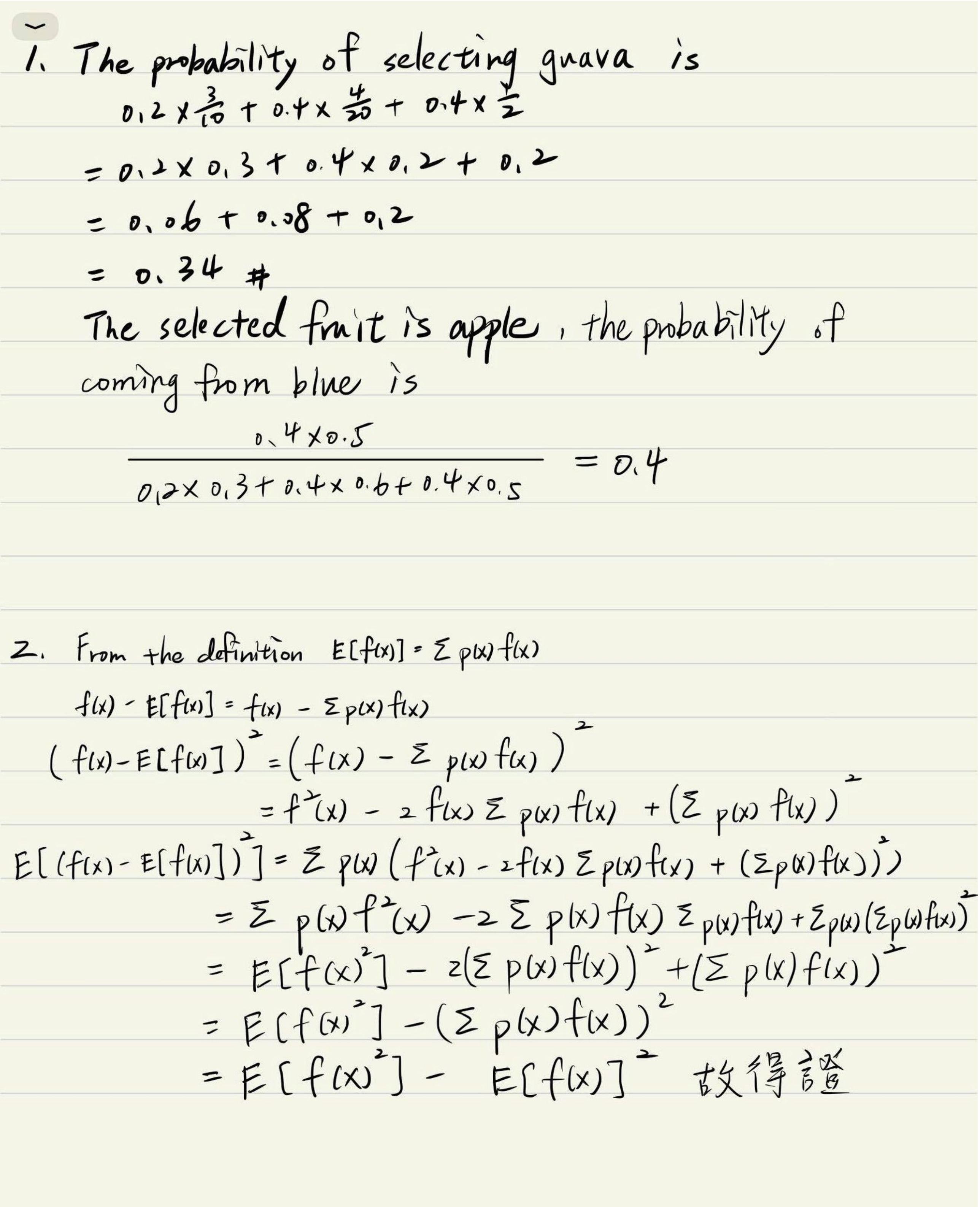
The mini-batch gradient descent divides whole training data into many batches with a fixed batch size. Assume the mini-batch size is 32, and it will update like gradient descent during an iteration. Unlike gradient descent, it will update the weights many times in an iteration and record the last weights in that iteration. The mini-batch gradient descent may be more robust than the stochastic gradient descent.

- Stochastic Gradient Descent

The Stochastic gradient descent is randomly update the learning rate after an iteration. The disadvantage of Stochastic gradient descent is that it may not converge in the end.

So I think the difference between the gradient descent and the mini-batch gradient is the batch-size and the frequency of updating weights in an iteration, and the difference between stochastic gradient descent and batch gradient descent is the changeable learning rate.

Part 2



1. The probability of selecting guava is

$$0.2 \times \frac{3}{10} + 0.4 \times \frac{4}{20} + 0.4 \times \frac{1}{2}$$

$$= 0.2 \times 0.3 + 0.4 \times 0.2 + 0.2$$

$$= 0.06 + 0.08 + 0.2$$

$$= 0.34 \ \#$$

The selected fruit is apple, the probability of coming from blue is

$$\frac{0.4 \times 0.5}{0.2 \times 0.3 + 0.4 \times 0.6 + 0.4 \times 0.5} = 0.4$$

2. From the definition $E[f(x)] = \Sigma p(x) f(x)$

$$f(x) - E[f(x)] = f(x) - \Sigma p(x) f(x)$$

$$(f(x) - E[f(x)])^2 = (f(x) - \Sigma p(x) f(x))^2$$

$$= f^2(x) - 2 f(x) \Sigma p(x) f(x) + (\Sigma p(x) f(x))^2$$

$$E[(f(x) - E[f(x)])^2] = \Sigma p(x) (f^2(x) - 2 f(x) \Sigma p(x) f(x) + (\Sigma p(x) f(x))^2)$$

$$= \Sigma p(x) f^2(x) - 2 \Sigma p(x) f(x) \Sigma p(x) f(x) + \Sigma p(x) (\Sigma p(x) f(x))^2$$

$$= E[f(x)^2] - 2(\Sigma p(x) f(x))^2 + (\Sigma p(x) f(x))^2$$

$$= E[f(x)^2] - (\Sigma p(x) f(x))^2$$

$$= E[f(x)^2] - E[f(x)]^2$$ 故得證

3. $E[x]=\sum x\, p(x)$

$$
\begin{aligned}
E_y E_x[x|y] &= E_y\Big[\sum_x x\cdot P(x|y)\Big] \\
&= \sum_y \Big[\sum_x x\cdot P(x|y)\Big]\cdot p(y) \\
&= \sum_y \sum_x x\cdot P(x,y) \\
&= \sum_x x \sum_y p(x,y) \\
&= \sum_x x\, P(x) \\
&= E[x]
\end{aligned}
$$