CS541 Homework 3

1. **Stochastic Gradient Descent**

Suppose F(w) is convex,

$$\eta_t = \frac{1}{\sqrt{t}\|\partial F(\omega^t)\|_2}$$

So $\lim\limits_{t\to\infty} \sqrt{t}\|\partial F(\omega^t)\|_2 \ = \ \infty$

$\lim\limits_{t\to\infty}\eta_t \ = \ 0$

$Also, \lim\limits_{t\to\infty} \omega^t = \omega^{t-1}$

$\eta_t \nabla f_{it}(\omega^{t-1}) = 0, \eta_t = 0 \ or \nabla f_{it}(\omega^{t-1}) \ = \ 0$

In SGD $\nabla f_{it}(\omega^{t-1}) \ \neq \ 0$

So $\lim\limits_{t\to\infty}\eta_t \ = \ 0$

2(a).

$$f_i(\omega) = (y_i - x_i \cdot \omega)^2$$

2. $(\boldsymbol{b})$.

$$f_i(\omega) = (y_i - x_i \cdot \omega)^2 + \lambda/2\|w\|_2^2$$

3.

$$F'_{(\omega)} = \frac{1}{n}\sum_{i=1}^{n} -y_i x_i + \lambda\omega \, , 1 - y_i x_i \omega > 0$$

$\frac{1}{n}\sum_{i=1}^{n} \ 0 + \lambda w$, else

$$F''(\omega) = \lambda$$

So, F(w) is λ – strongly convex, the solution is unique.

$$f_i(w) = \ \max\{1 - y_i x_i w, 0\} + \lambda/2\|\omega\|_2^2$$

Detailed update rule:

$$\eta_t = \frac{1}{t}$$

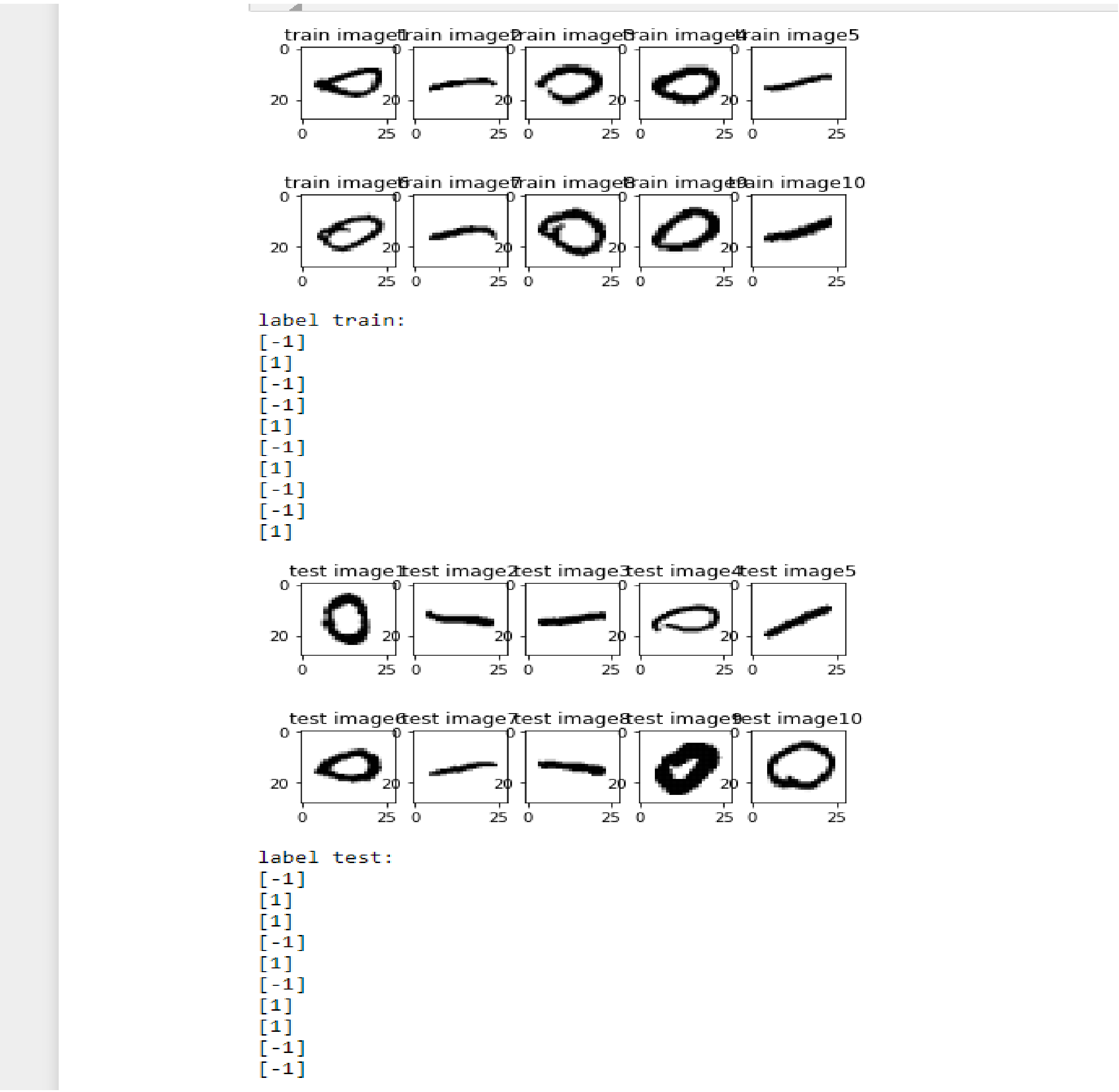
$$f_i'(\omega) = -y_i x_i + \lambda\omega \qquad , 1 - y_i x_i \omega > 0$$

$$\qquad = 0 \; + \; \lambda\omega \qquad\qquad , else$$

$$\omega^t = \omega^{t-1} - \frac{1}{t} \cdot f_{it}'(\omega^{t-1})$$

2.Support Vector Machine

Data Preparation



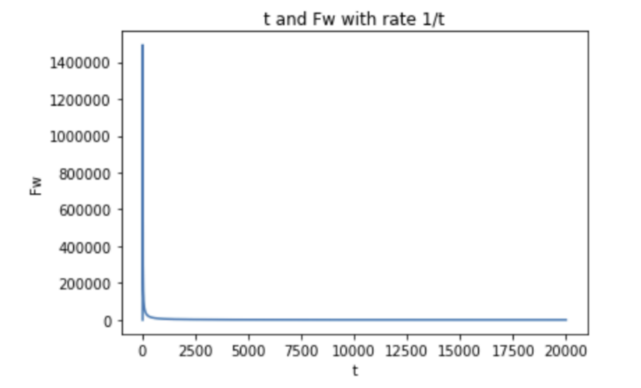
```
label train:
[-1]
[1]
[-1]
[-1]
[1]
[-1]
[1]
[-1]
[-1]
[1]
```

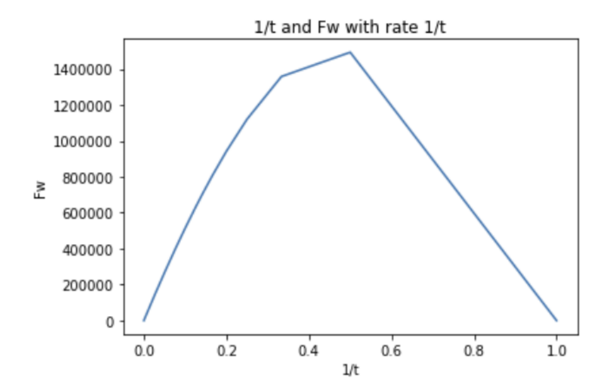
```
label test:
[-1]
[1]
[1]
[-1]
[1]
[-1]
[1]
[1]
[-1]
[-1]
```

2.2 Evaluation Metric (implement by code)

2.3 Convergence of SGD

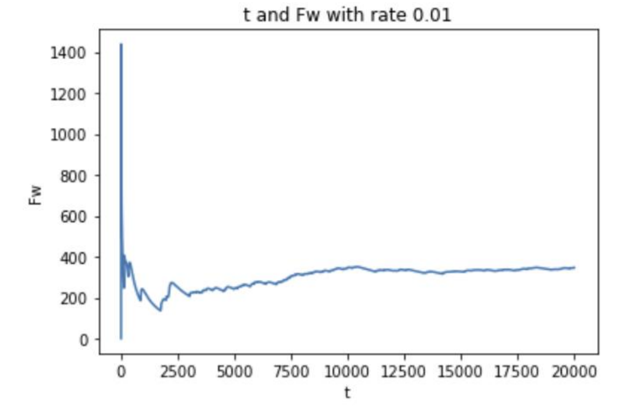
- **λ = 1 and rate = 1/t**

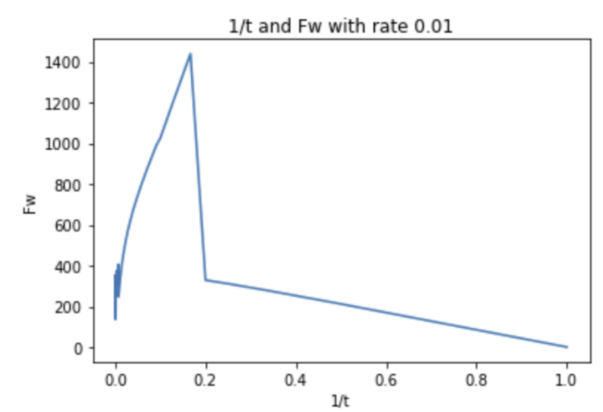




From the data in F(w) we can see that F(w) will close to 0 when the iteration value is about 500.

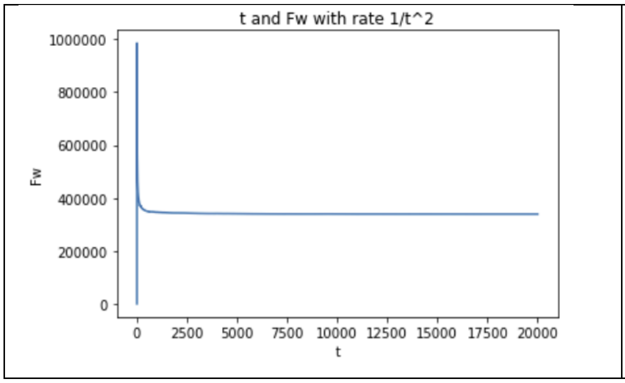
- **λ = 1 and rate = 0.01**

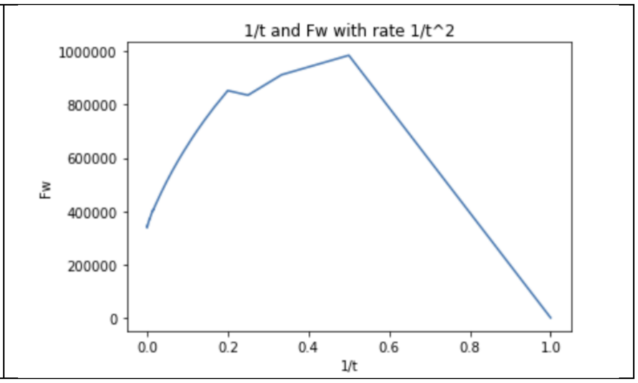




From the picture and data in F(w) we find that when learning rate is 0.01, F(w) won't close to 0.

- **λ = 1 and rate = 1/t^2**

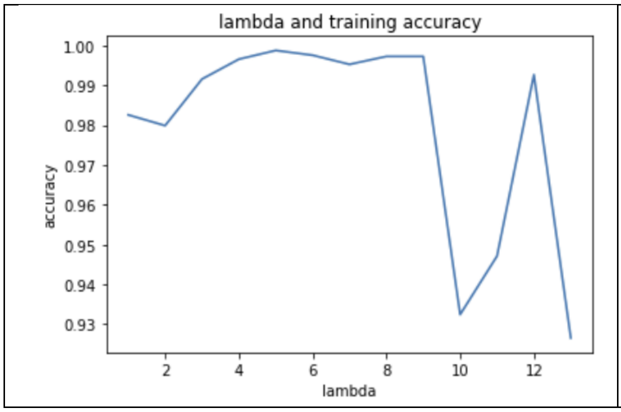


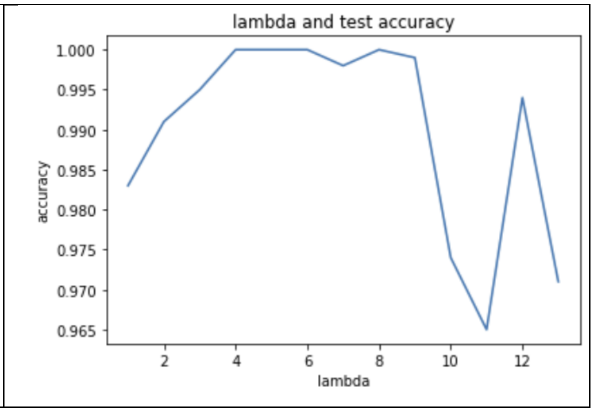


From the picture we can find that F(w) won't close to 0 when learning rate is 1/t^2.

2.4 Hyper-Parameter

λ in [0.000001, 0.001, 0.1, 0.5, 1, 2, 5, 10, 20, 50, 100, 500, 1000, 10000]



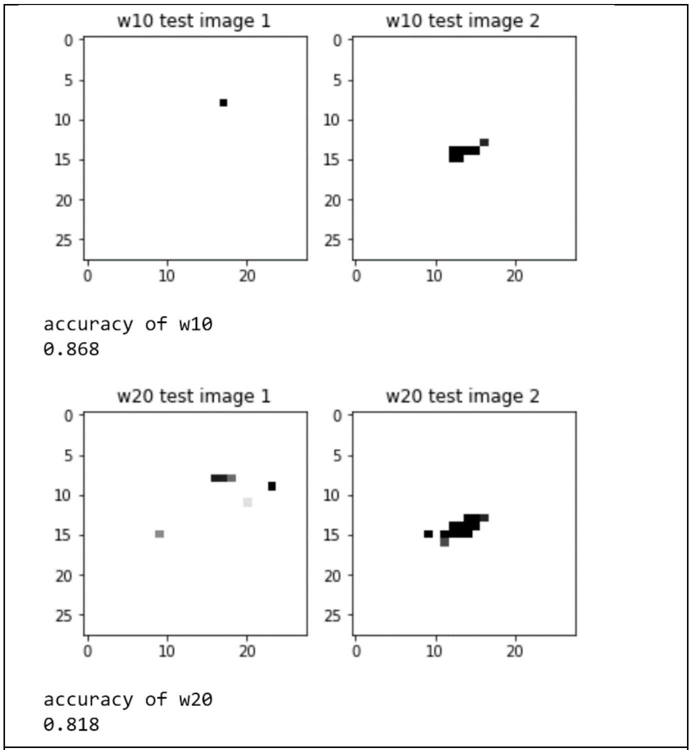


λ	0.000001	0.001	0.1	0.5	1	2	5	10	20	50	100	500	1000	10000
||w10000||2/d	3.8033702	2.79457159	1.0543638	0.02884166	0.00158474	0.000982787	0.000784602	0.0005003	0.00010653	0.0005128	0.00029024	6.2906E-06	4.59301E-06	2.71913E-06

When λ increase, the value of$\|\omega^{10000}\|_2 \;/\; d$ become smaller, the accuracy of training and testing data also decrease when λ become too big.

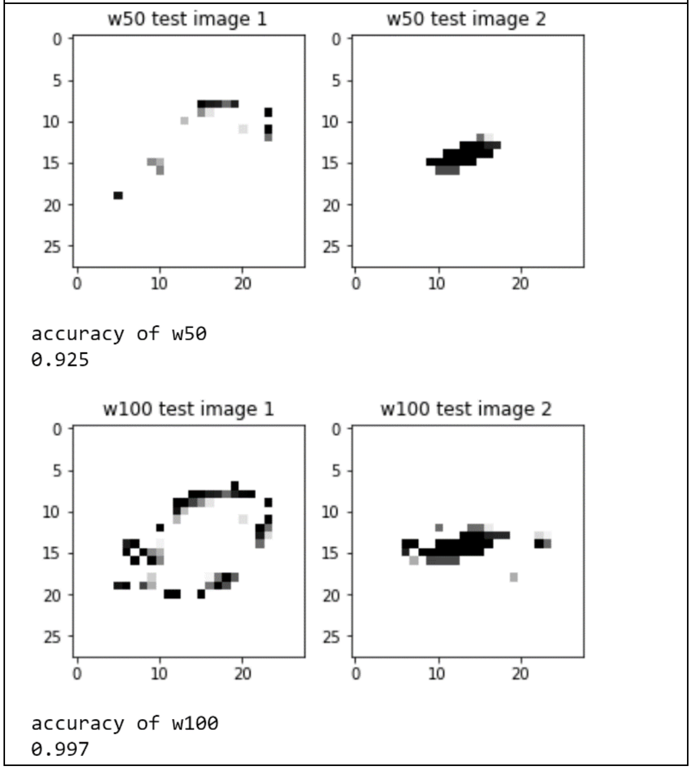
I select the best w when λ = 0.5 for further work.

Visualization and accuracy for sparsity s = [10, 20, 50, 100, 200, 400]



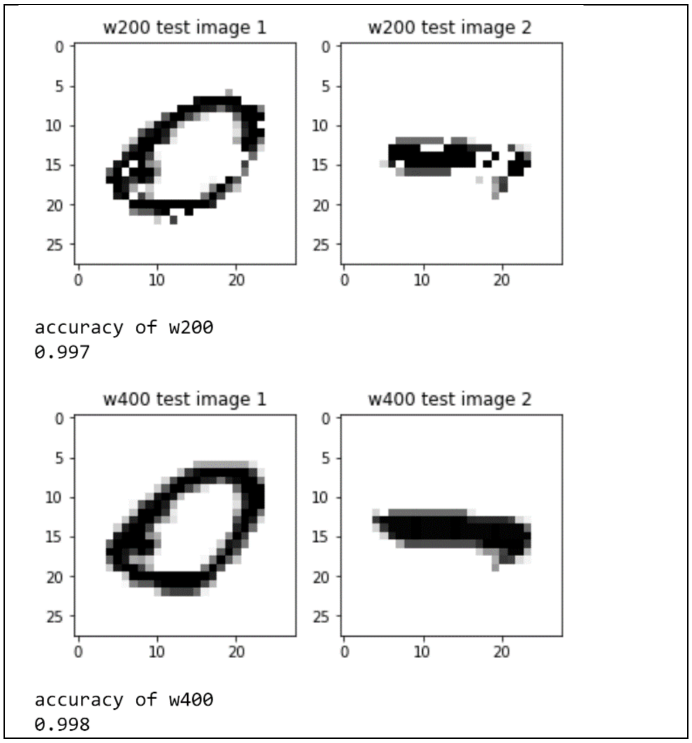
```
accuracy of w10
0.868
```

```
accuracy of w20
0.818
```



```
accuracy of w50
0.925
```

```
accuracy of w100
0.997
```



```
accuracy of w200
0.997
```

```
accuracy of w400
0.998
```

From the accuracy data we can find that the value of accuracy is well when s is equal to or bigger than 100, but the accuracy is not so good when s small than 100. For our picture data, the edge of each picture is blank and only data in the middle of the picture is useful for training and testing. So when s is big enough to contain the data at the middle of the picture, the model performs as well as the whole ω^*.

2.5 Noisy Labels

ρ	λ	Training Acc	Testing Acc	||w||2 / d
0	1	0.9987	0.999	0.00268381
0.01	5	0.9891	0.998	0.001081923
0.1	0.001	0.9305	0.987	2.412479679
0.2	1	0.9058	0.995	0.003510566
0.3	20	0.8239	0.965	0.001100146
0.5	5	0.7987	0.982	0.001801701
0.7	5	0.7487	0.971	0.002051681

When the noisy level increase, the value of λ that get the best testing accuracy is become bigger, the Training accuracy decrease with the increasing of noisy level and λ , $\|\omega_\rho\|_2 \,/\, d$ decrease when λ become bigger.