1. a)

When $y = \text{Keep}$, $\mathbb{E}(\mathcal{L}(y,t)) = 0 + 1 \cdot 0.1 = \cancel{0.01}$ $0.1$

When $y = \text{Remove}$, $\mathbb{E}(\mathcal{L}(y,t)) = 100 \cdot 0.9 + 0 = 90$

1. b) $p(t|x) = \frac{p(x,t)}{p(x)}$

$$y_*(x) = \underset{y}{\arg\min} \sum_{t \in \{Spam, NonSpam\}} Pr(t|x)\,\mathcal{L}(y,t)$$

$$= \underset{y}{\arg\min}\left( Pr(t=Spam|x)\,\mathcal{L}(y,t=Spam) + (1 - Pr(t=Spam|x))\,\mathcal{L}(y,t=Nonspam)\right)$$

1. c) $Pr(x_1, x_2|t)$, $P(t)$ are given.

$$Pr(t|x_1,x_2) = \frac{Pr(x_1,x_2|t)\,P(t)}{Pr(x_1,x_2)} = \frac{Pr(x_1,x_2|t)\,P(t)}{\sum_t Pr(x_1,x_2|t)P(t)}$$

When $x_1 = x_2 = 0$,

$$Pr(t=Spam|x_1=0, x_2=0) = \frac{0.4 \cdot 0.1}{0.4 \cdot 0.1 + 0.998 \cdot 0.9} = 0.0426$$

$y = \text{Keep}$, $Pr(t=Spam|x_1=0,x_2=0)\,\mathcal{L}(y,t=Spam) + Pr(t=NonSpam|x_1=0,x_2=0)\,\mathcal{L}(y,t=NonSpam))$

$= 0.0426 \cdot 1 + 0.9573 \cdot 0$

$= 0.0426$

$y = \text{Remove}$, $Pr(t=Spam|x)\,\mathcal{L}(y,t=Spam) + Pr(t=NonSpam|x)\,\mathcal{L}(y,t=NonSpam)$

$= 0.0206 \cdot 0 + 0.9573 \cdot 100$

$= \cancel{97.94}\ 95.73$

So $y_*(x_1=0, x_2=0) = \text{Keep}$

When $x_1=0,\ x_2=1$

$$Pr(t=\text{Spam}\mid x_1=0,x_2=1)=\frac{0.3\cdot 0.1}{0.1\cdot 0.3+0.9\cdot 0.001}=0.9709.$$

If $y=\text{Keep}$, $\sum_t Pr(t\mid x)\mathcal{L}(y,t)$

$$=0.9709\cdot 1+0.0291\cdot 0$$

$$=\cancel{0.9708}\ 0.9709$$

If $y=\text{Remove}$, $\sum_t Pr(t\mid x)\mathcal{L}(y,t)$

$$=0.9708\cdot 0+0.0291\cdot 100$$

$$=2.91$$

$y_*(x_1=0,x_2=1)=\text{Keep}$

When $x_1=1,\ x_2=0$

$$Pr(t=\text{Spam}\mid x_1=1,x_2=0)=\frac{0.2\cdot 0.1}{0.2\cdot 0.1+0.001\cdot 0.9}=0.9569$$

$y=\text{Keep}$, $\sum_t Pr(t\mid x)\mathcal{L}(y,t)$

$$=0.9569\cdot 1+0.0431\cdot 0$$

$$=0.9569$$

$y=\text{Remove}$, $\sum_t Pr(t\mid x)\mathcal{L}(y,t)$

$$=0.9569\cdot 0+0.0431\cdot 100$$

$$=4.31$$

So $y_*(x_1=1,x_2=0)=\text{Keep}.$

When $x_1=x_2=1$

$$Pr(t=\text{Spam}\mid x_1=1,x_2=1)=\frac{0.1\cdot 0.1}{0.1\cdot 0.1+0.9\cdot 0}=1$$

$y=\text{Keep}$, $\sum_t Pr(t\mid x)\mathcal{L}(y,t)=1\cdot 1+0\cdot 0=1$

$y=\text{Remove}$, $\sum_t Pr(t\mid x)\mathcal{L}(y,t)=1\cdot 0+0\cdot 100=0$

$y_*(x_1=1,x_2=1)=\text{Remove}.$

d) $\mathbb{E}(\mathcal{L}(y_*, t))$

$$= \sum_{x_1, x_2} Pr(x_1, x_2) \sum_t P(t|x_1, x_2)\, \mathcal{L}(y_*(x_1, x_2), t)$$

$$= Pr(x_1=0, x_2=0) \cdot 0.0426 + Pr(x_1=0, x_2=1) \cdot 0.9709$$

$$+ Pr(x_1=0, x_2=1) \cdot 0.9569 + Pr(x_1=1, x_2=1) \cdot 0$$

$$= (0.4 \cdot 0.1 + 0.998 \cdot 0.9) \cdot 0.0426 + (0.1 \cdot 0.3 + 0.9 \cdot 0.001) \cdot 0.9709$$

$$+ (0.2 \cdot 0.1 + 0.001 \cdot 0.9) \cdot 0.9569 + 0$$

$$= 0.04 + 0.03 + 0.02$$

$$= 0.09$$

2.a)

Suppose the given dataset is linearly separable.


Suppose there were feasible weights.

Since $x=-1, 3$ are positive examples, the segment connecting them must also be in the positive half-space.

However, $x=1$ is contained in the segment but it's a negative example.

Contradiction.

b) $z = \psi(x)^{\top} \vec{w}$ , $\vec{w} = (w_1, w_2)$

$= w_1 x + w_2 x^2$

$$y = \begin{cases} 0 & z < 0 \\ 1 & z \geq 0 \end{cases}$$

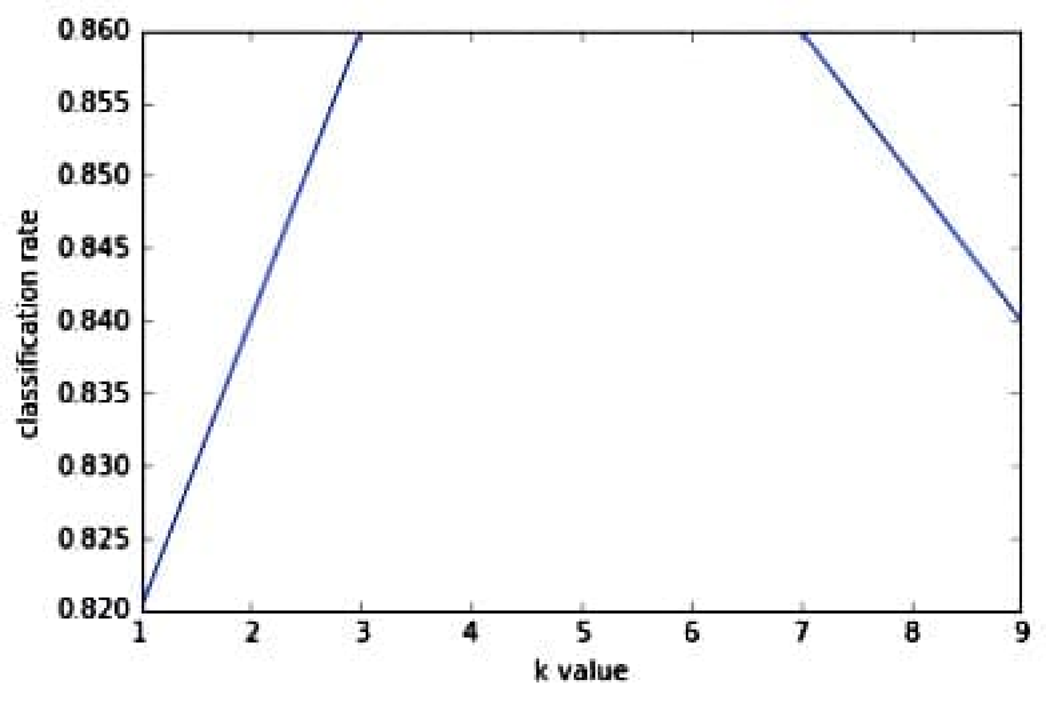
$$\Rightarrow \begin{cases} -w_1 + w_2 \geq 0 \\ w_1 + w_2 < 0 \\ 3w_1 + 9w_2 \geq 0 \end{cases}$$

One solution is $\begin{cases} w_1 = -3 \\ w_2 = 2 \end{cases}$

3.1 b)

The classification rate rises and the drops as k increases. I would choose $k=5$, which is not too big or small so that we can avoid overfit / underfit. The classification rate for $k=5$ is 0.86, (one of the highest) for $k=3$ and $k=7$ are also 0.86.

The test classification rates for $k=1,3,5,7,9$ are 0.88, 0.92, 0.94, 0.94, 0.88, which are in general higher than the validation rates but exhibit the same pattern of k as the validation rates. $k=5$ still gives the best performance.



0.860
0.855
0.850
0.845
0.840
0.835
0.830
0.825
0.820
classification rate
1
2
3
4
5
6
7
8
9
k value

3.2 b)

For mnist_train, set learning_rate = 0.005, num_iteration=1600

| | cross entropy | classification error |
|---|---|---|
| train | -2971 | 0 |
| valid | -1450 | 0.1 |
| test | -907 | 0.08 |

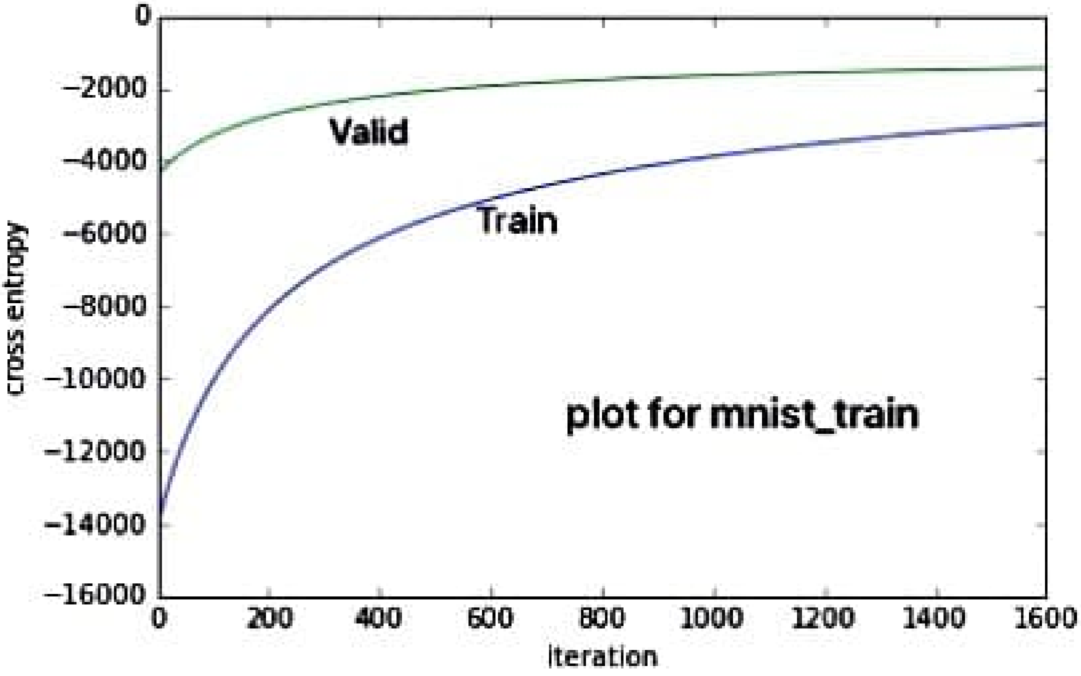
For mnist_train_small, set learning_rate=0.1, num_iteration= 10000

| | cross entropy | classification error |
|---|---|---|
| train | 0 | 0 |
| valid | -1425 | 0.26 |
| test | - 512 | 0.22 |

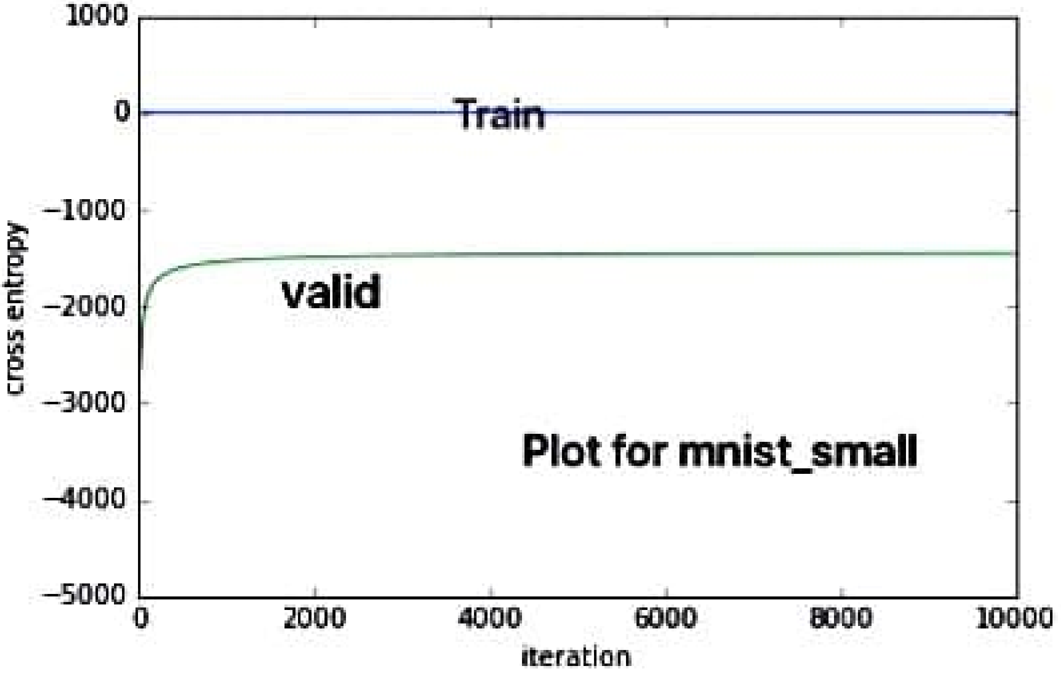
3.2. c)

The results don't change.

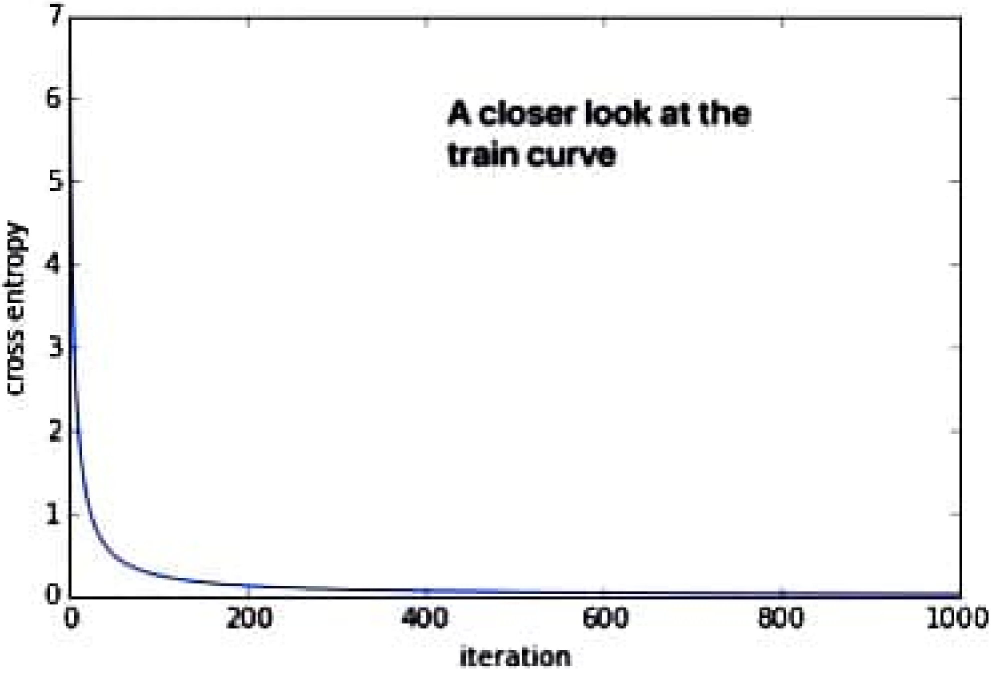
If they do, we should treat the intial weights as another hyperparameter.



Valid
Train
plot for mnist_train
cross entropy
Iteration
0
−2000
−4000
−6000
−8000
−10000
−12000
−14000
−16000
0
200
400
600
800
1000
1200
1400
1600



1000
0
−1000
−2000
−3000
−4000
−5000
cross entropy
Train
valid
Plot for mnist_small
0
2000
4000
6000
8000
10000
iteration


A closer look at the
train curve
cross entropy
iteration
0
1
2
3
4
5
6
7
0
200
400
600
800
1000

4.a) $J(\vec{w}) = \frac{1}{2}\sum_i a^{(i)}(y^{(i)} - w^T x^{(i)})^2 + \frac{\lambda}{2}\|\vec{w}\|^2$

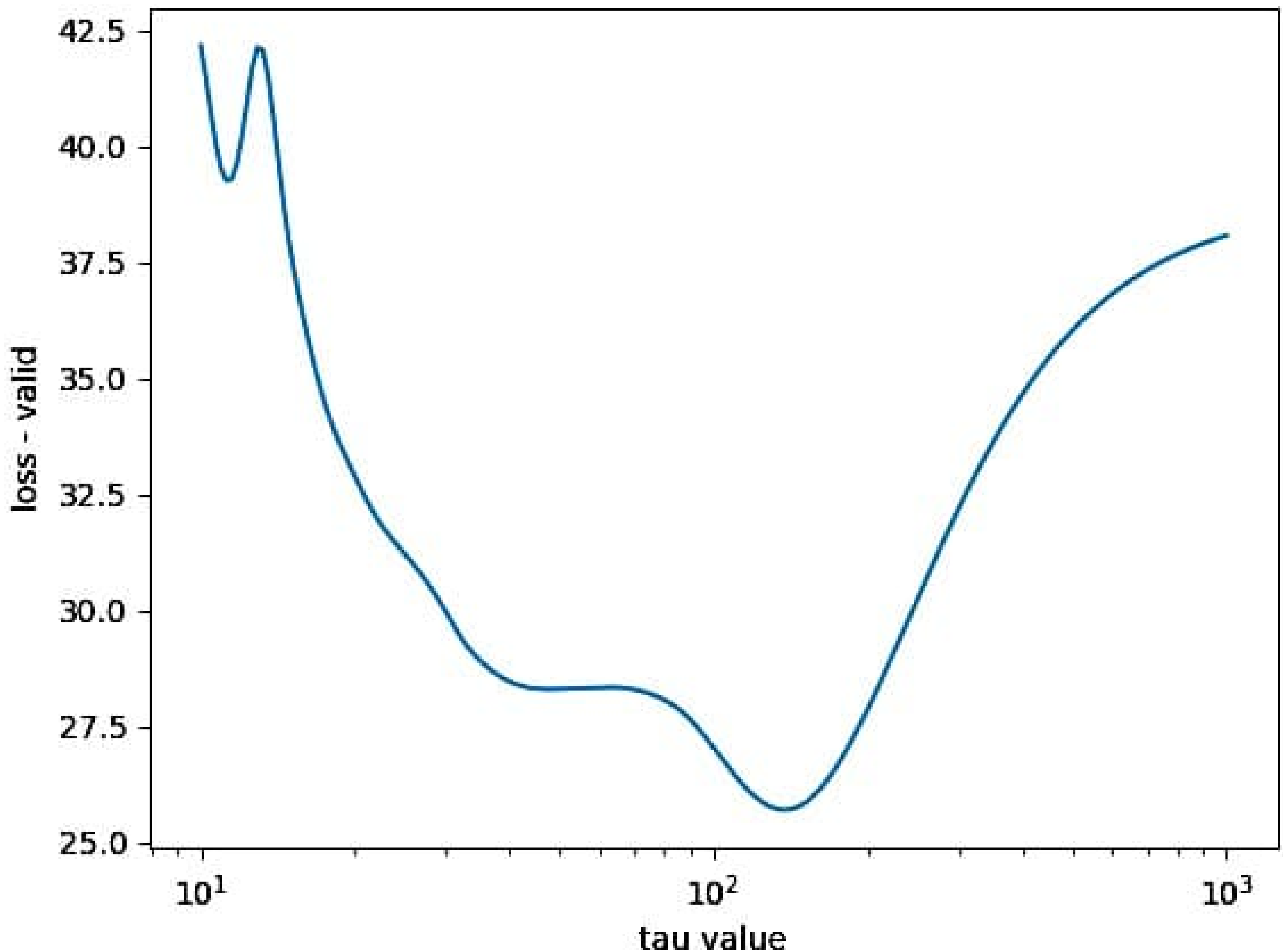
$$\frac{\partial J}{\partial w_j} = \sum_i a^{(i)}(y^{(i)} - \vec{w}^T x^{(i)}) \cdot \left(-x^{(i)(j)}\right) + \lambda \|\vec{w}\| w_j = 0$$

$$\Rightarrow \quad \lambda \|\vec{w}\| w_j = \sum_i a^{(i)}(y^{(i)} - \vec{w}^T x^{(i)})\, x^{(i)(j)}$$

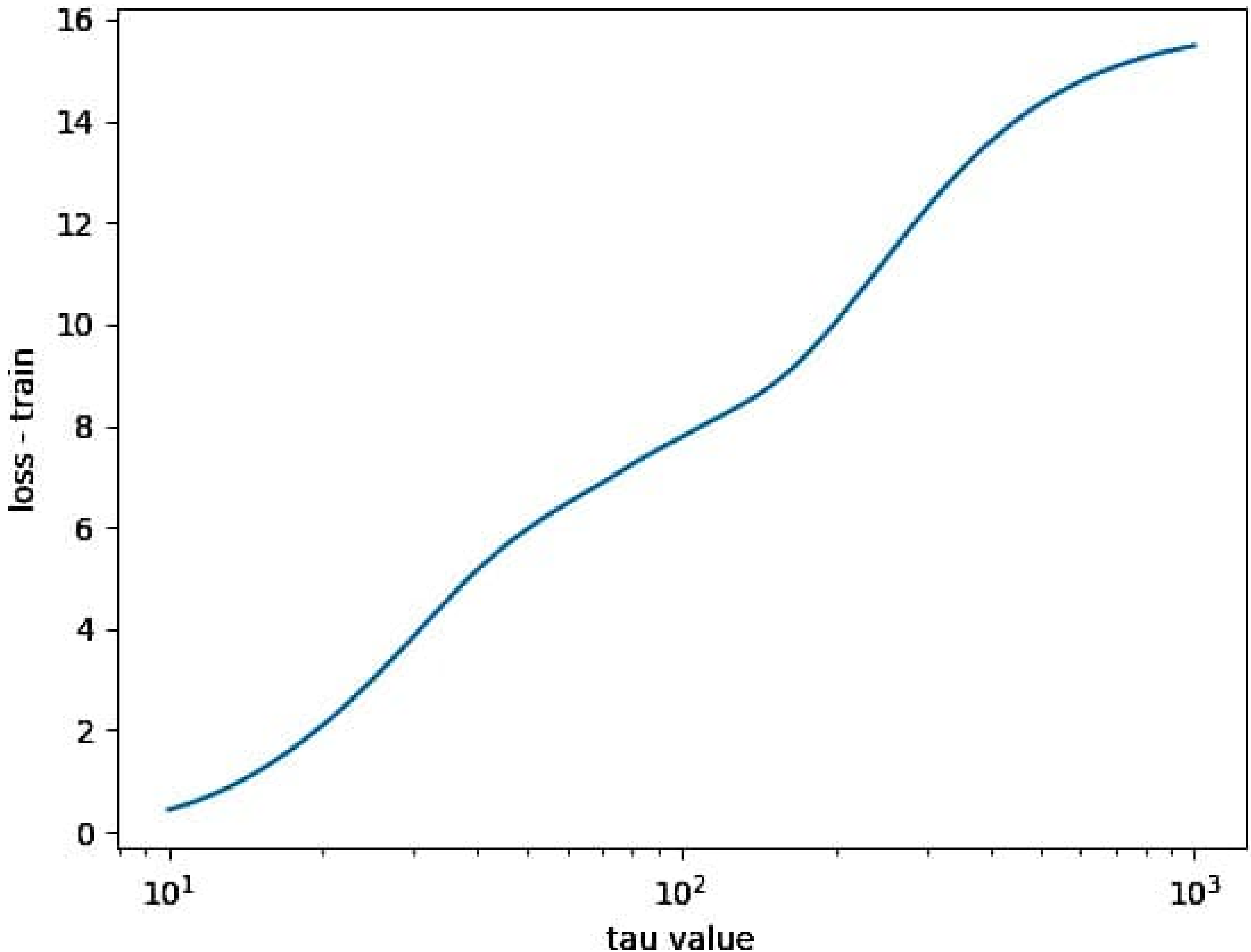
$$\Rightarrow \quad \sum_i a^{(i)}(\vec{w}^T x^{(i)})\, x^{(i)(j)} + \lambda \|\vec{w}\| w_j = \sum_i a^{(i)} y^{(i)} x^{(i)(j)}$$

$$\Rightarrow \quad X^T A X \vec{w} + \lambda \vec{w} = X^T A \vec{y}$$

$$\Rightarrow \quad \vec{w} = (X^T A X + \lambda I)^{-1} X^T A \vec{y}$$



42.5
40.0
37.5
35.0
32.5
30.0
27.5
25.0
loss - valid
$10^1$
$10^2$
$10^3$
tau value



16
14
12
10
8
6
4
2
0
loss - train
$10^1$
$10^2$
$10^3$
tau value

4.d)

$$\lim_{\tau\to\infty} a^{(i)} = \lim_{\tau\to\infty} \frac{e^{-\|x-x^{(i)}\|^2/2\tau^2}}{\sum_j e^{-\|x-x^{(j)}\|^2/2\tau^2}}$$

$$= \lim_{\tau\to\infty} \frac{1}{\sum_j e^{\frac{-\|x-x^{(j)}\|^2 + \|x-x^{(i)}\|^2}{2\tau^2}}}$$

$$= \frac{1}{\sum_j e^0}$$

$$= \frac{1}{N}$$

Therefore, as $\tau\to\infty$, $A \to \frac{1}{N} I$ and the predicted values and losses will be stable (tend to an asymptote).

$$\lim_{\tau\to 0} a^{(i)} = \lim_{\tau\to 0} \frac{1}{\sum_j e^{\frac{-\|x-x^{(j)}\|^2 + \|x-x^{(i)}\|^2}{2\tau^2}}}$$

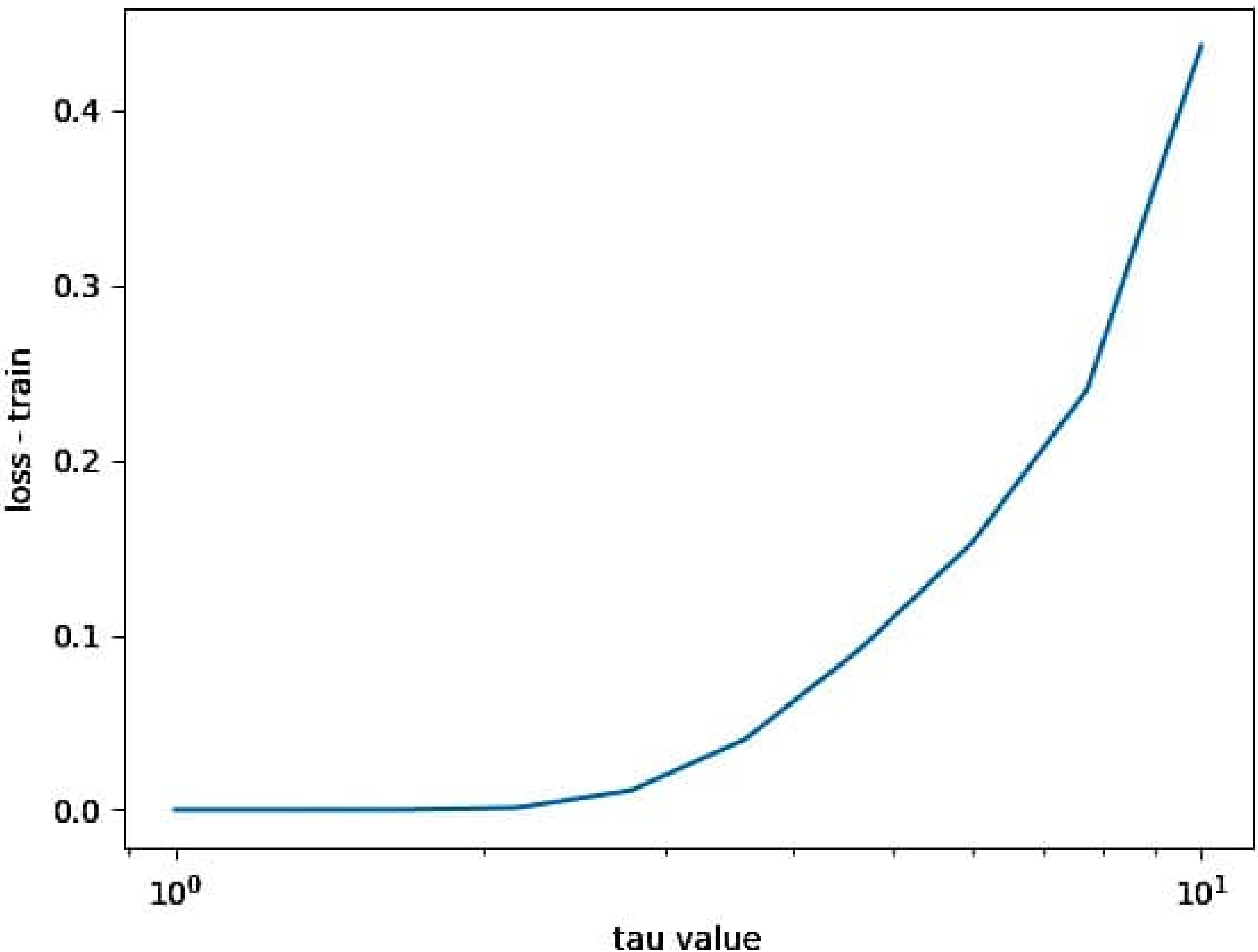
$$= \begin{cases} 1 & \text{if } \|\vec{x}^{(i)} - \vec{x}\|^2 \geq \|\vec{x}^{(j)} - \vec{x}\|^2 \text{ for all } j. \\ 0 & \text{otherwise} \end{cases}$$

So $A \to \begin{pmatrix} 0 & & & 0 \\ & \ddots & & \\ & & 1 & \\ 0 & & \underset{a^{(i)}}{\uparrow} & \ddots & 0 \end{pmatrix}$

So $w^* \to \arg\min \frac{1}{2} (y^{(i)} - \vec{w}^T \vec{x}^{(i)})^2 + \frac{\lambda}{2}\|\vec{w}\|^2$ where $i$ is such that $\|\vec{x}^{(i)} - \vec{x}\|^2$ is max.

As the predicted value gets closer and closer to $y^{(i)}$, it gets farther apart from other $y^{(j)}$'s.

So the losses


0.4
0.3
0.2
0.1
0.0
loss - train
$10^0$
$10^1$
tau value