CS229 - Problem Set 1

Or Haifler

November 2023

Exercise 1

(a)

$$J(\theta) = -\frac{1}{m}\left[\sum_{i=1}^{m} y^{(i)}\log(h_\theta(x^{(i)})) + (1-y^{(i)})\log(1-h_\theta(x^{(i)}))\right]$$

$$\frac{\partial}{\partial\theta_k\theta_j}\log(h_\theta(x)) = -x_k x_j h_\theta(x)(1-h_\theta(x)), \frac{\partial}{\partial\theta_k\theta_j}\log(1-h_\theta(x)) = -x_k x_j h_\theta(x)(1-h_\theta(x))$$

$$\frac{\partial}{\partial\theta_k\theta_j}J(\theta) = -\frac{1}{m}\left[\sum_{i=1}^{m} y^{(i)}\frac{\partial}{\partial\theta_k\theta_j}\log(h_\theta(x^{(i)})) + (1-y^{(i)})\frac{\partial}{\partial\theta_k\theta_j}\log(1-h_\theta(x^{(i)}))\right] = \frac{1}{m}\sum_{i=1}^{m} x_j x_k h_\theta(x^{(i)})(1-h_\theta(x^{(i)}))$$

$$\langle Hz, z\rangle = \sum_{k=1}^{n}\sum_{j=1}^{n} z_j z_k \frac{1}{m}\sum_{i=1}^{m} x_j x_k h_\theta(x^{(i)})(1-h_\theta(x^{(i)})) = \frac{1}{m}\left(\sum_{i=1}^{m} h_\theta(x^{(i)})(1-h_\theta(x^{(i)}))\right)\sum_{k=1}^{n}\sum_{j=1}^{n} z_j z_k x_j x_k = \frac{\lambda}{m}(x^T z)^2 \geq 0$$

(c)

$$p(y=1|x;\phi,\mu_0,\mu_1,\Sigma) = \frac{p(x|y=1)p(y=1)}{p(x|y=0)p(y=0)+p(x|y=1)p(y=1)}$$

$$= \frac{\frac{\phi}{(2\pi)^{n/2}\det(\Sigma)^{1/2}}\exp(-\frac{1}{2}(x-\mu_1)^T\Sigma^{-1}(x-\mu_1)}{\frac{1-\phi}{(2\pi)^{n/2}\det(\Sigma)^{1/2}}\exp(-\frac{1}{2}(x-\mu_0)^T\Sigma^{-1}(x-\mu_0)) + \frac{\phi}{(2\pi)^{n/2}\det(\Sigma)^{1/2}}\exp(-\frac{1}{2}(x-\mu_1)^T\Sigma^{-1}(x-\mu_1)}$$

$$= \frac{\exp(-\frac{1}{2}(x-\mu_1)^T\Sigma^{-1}(x-\mu_1))}{\frac{1-\phi}{\phi}\exp(-\frac{1}{2}(x-\mu_0)^T\Sigma^{-1}(x-\mu_0)) + \exp(-\frac{1}{2}(x-\mu_1)^T\Sigma^{-1}(x-\mu_1))}$$

$$= \frac{1}{1+\frac{1-\phi}{\phi}\exp(-\frac{1}{2}((x-\mu_0)^T\Sigma^{-1}(x-\mu_0) - (x-\mu_1)^T\Sigma^{-1}(x-\mu_1))}$$

$$= \frac{1}{1+\exp(-\frac{1}{2}(\mu_1-\mu_0)^T\Sigma^{-1}x + \frac{1}{2}(\mu_1^T\Sigma^{-1}\mu_1 - \mu_0^T\Sigma^{-1}\mu_0) + \log(\frac{1-\phi}{\phi}))} = \frac{1}{1+\exp(-(\theta^T x+\theta_0))}$$

$$\theta = (\Sigma^T)^{-1}(\mu_1-\mu_0) \overset{\Sigma^T=\Sigma}{=} \Sigma^{-1}(\mu_1-\mu_0), \theta_0 = \frac{1}{2}(\mu_1^T\Sigma^{-1}\mu_1 - \mu_0^T\Sigma^{-1}\mu_0) + \log(\frac{1-\phi}{\phi})$$

(d)

$$\frac{\partial}{\partial \phi}\ell(\phi,\mu_0,\mu_1,\sigma) = \frac{\partial}{\partial \phi}\log\prod_{i=1}^{m} p(x^{(i)}|y^{(i)};\mu_0,\mu_1,\Sigma)p(y^{(i)};\phi) = \sum_{i=1}^{m}\frac{\partial}{\partial \phi}\log p(x^{(i)}|y^{(i)};\mu_0,\mu_1,\sigma)p(y^{(i)};\phi)$$

$$\sum_{i=1}^{m}\frac{\partial}{\partial \phi}\log p(y^{(i)};\phi) = \sum_{i=1}^{m_0}\frac{\partial}{\partial \phi}\log\phi + \sum_{i=m_0}^{m}\frac{\partial}{\partial \phi}\log(1-\phi) = \frac{m_0}{\phi} + \frac{m-m_0}{1-\phi}$$

$$\frac{\partial}{\partial \phi}\ell(\phi,\mu_0,\mu_1,\sigma) = 0 \implies (1-\phi)m_0 + \phi(m_0 - m) = 0 \implies \phi = \frac{m_0}{m} = \frac{1}{m}\sum_{i=1}^{m} 1\{y^{(i)} = 1\}$$

$$\nabla_{\mu_0}\ell(\phi,\mu_0,\mu_1,\sigma) = \sum_{i=1}^{m}\nabla_{\mu_0}\log p(x^{(i)}|y^{(i)};\mu_0,\mu_1,\sigma)$$

$$= \sum_{i=m_0}^{m}\nabla_{\mu_0}\log\frac{1}{(2\pi)^{n/2}\det(\sigma)^{1/2}}\exp(-\frac{1}{2}(x^{(i)}-\mu_0)^T\sigma^{-1}(x^{(i)}-\mu_0)) =$$

$$= -\frac{1}{2}\sum_{i=m_0}^{m}\nabla_{\mu_0}(x^{(i)}-\mu_0)^T\sigma^{-1}(x^{(i)}-\mu_0) = \sum_{i=m_0}^{m}\sigma^{-1}(x^{(i)}-\mu_0)$$

$$\nabla_{\mu_0}\ell(\phi,\mu_0,\mu_1,\sigma) = 0 \implies \sum_{i=m_0}^{m}(x^{(i)}-\mu_0) = 0 \implies \mu_0 = \frac{\sum_{i=m_0}^{m} 1\{y^{(i)}=0\}x^{(i)}}{\sum_{i=m_0}^{m} 1\{y^{(i)}=0\}}$$

$$\nabla_{\mu_1}\ell(\phi,\mu_0,\mu_1,\sigma) = 0 \implies \sum_{i=0}^{m_0}(x^{(i)}-\mu_1) = 0 \implies \mu_1 = \frac{\sum_{i=0}^{m_0} 1\{y^{(i)}=1\}x^{(i)}}{\sum_{i=0}^{m_0} 1\{y^{(i)}=1\}}$$

$$\frac{\partial}{\partial \sigma}\ell(\phi,\mu_0,\mu_1,\sigma) = \sum_{i=1}^{m}\frac{\partial}{\partial \sigma}\log p(x^{(i)}|y^{(i)};\mu_0,\mu_1,\sigma) = -\frac{m}{2\sigma} - \frac{1}{2}\frac{\partial}{\partial \sigma}\sigma^{-1}\sum_{i=1}^{m}(x^{(i)}-\mu_{y^{(i)}})^T(x^{(i)}-\mu_{y^{(i)}}) =$$

$$-\frac{m}{2\sigma} + \frac{1}{2\sigma^2}\sum_{i=1}^{m}(x^{(i)}-\mu_{y^{(i)}})^T(x^{(i)}-\mu_{y^{(i)}})$$

$$\frac{\partial}{\partial \sigma}\ell(\phi,\mu_0,\mu_1,\sigma) = 0 \implies m = \frac{1}{\sigma}\sum_{i=1}^{m}(x^{(i)}-\mu_{y^{(i)}})^T(x^{(i)}-\mu_{y^{(i)}}) \implies \sigma = \frac{1}{m}\sum_{i=1}^{m}(x^{(i)}-\mu_{y^{(i)}})^T(x^{(i)}-\mu_{y^{(i)}})$$

(g)

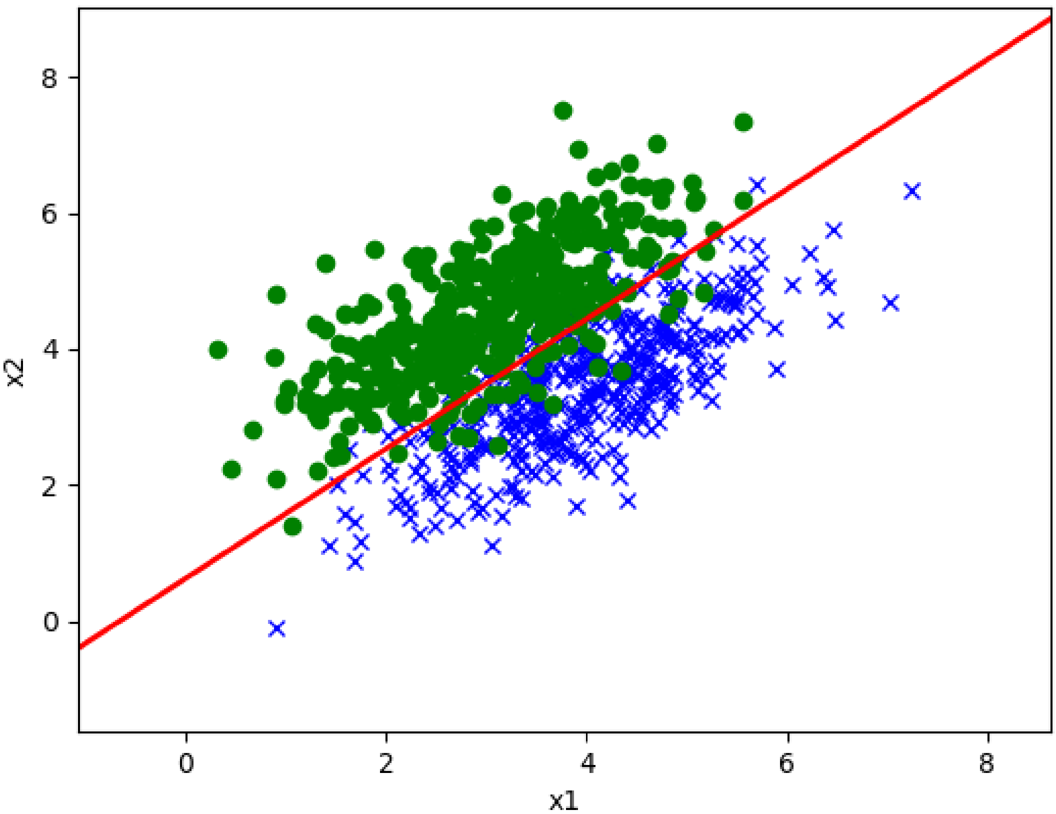


Figure 1: GDA

Exercise 2

(a)

$$p(y=1|t=1,x)p(t=1|x)p(x) = p(y=1,t=1,x) = p(t=1|y=1,x)p(y=1|x)p(x)$$
$$p(t=1|x) = p(y=1|x)\frac{p(t=1|y=1,x)}{p(y=1|t=1,x}$$
$$p(t=1|y=1,x) = 1, p(y=1|t=1,x) = p(y=1|t=1)$$
$$p(t=1|x) = \frac{p(y=1|x)}{p(y=1|t=1)}, p(y=1|t=1) = \alpha$$

(b)

$$h(x^{(i)}) \approx (y^{(i)}=1|x^{(i)}) \stackrel{(a)}{=} \alpha p(t^{(i)}=1|x^(i)) \approx \alpha$$

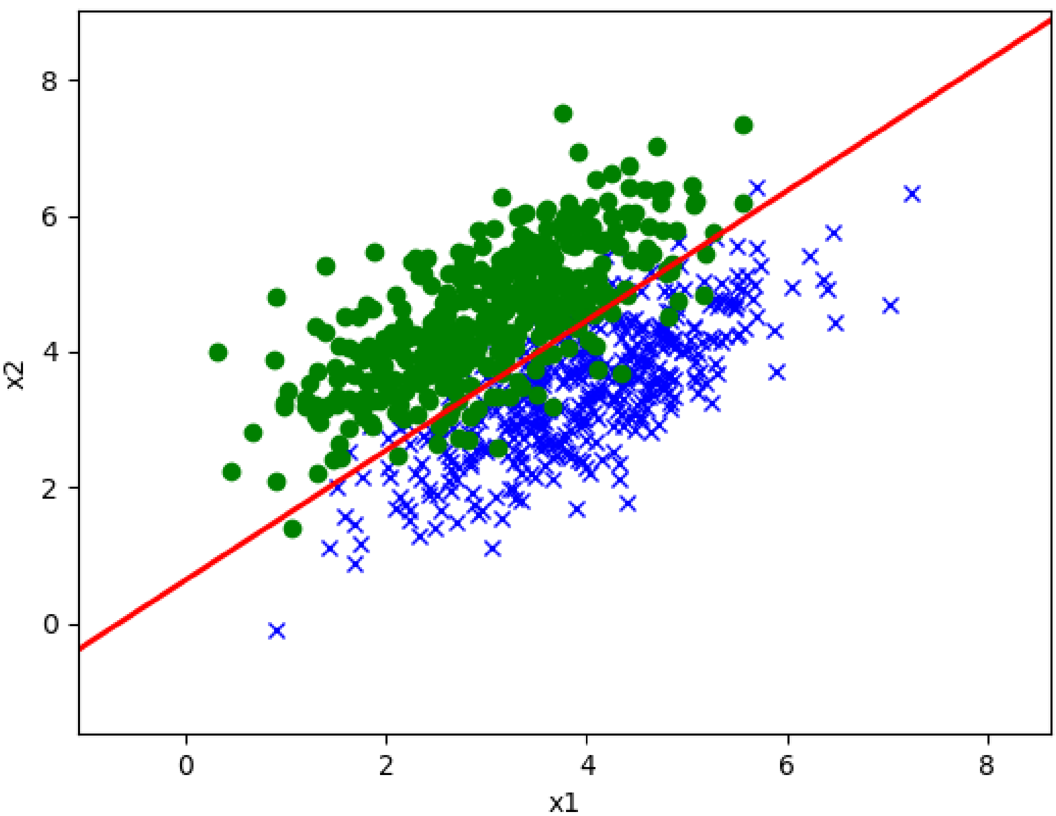


Figure 2: Logistic Regression

(c)

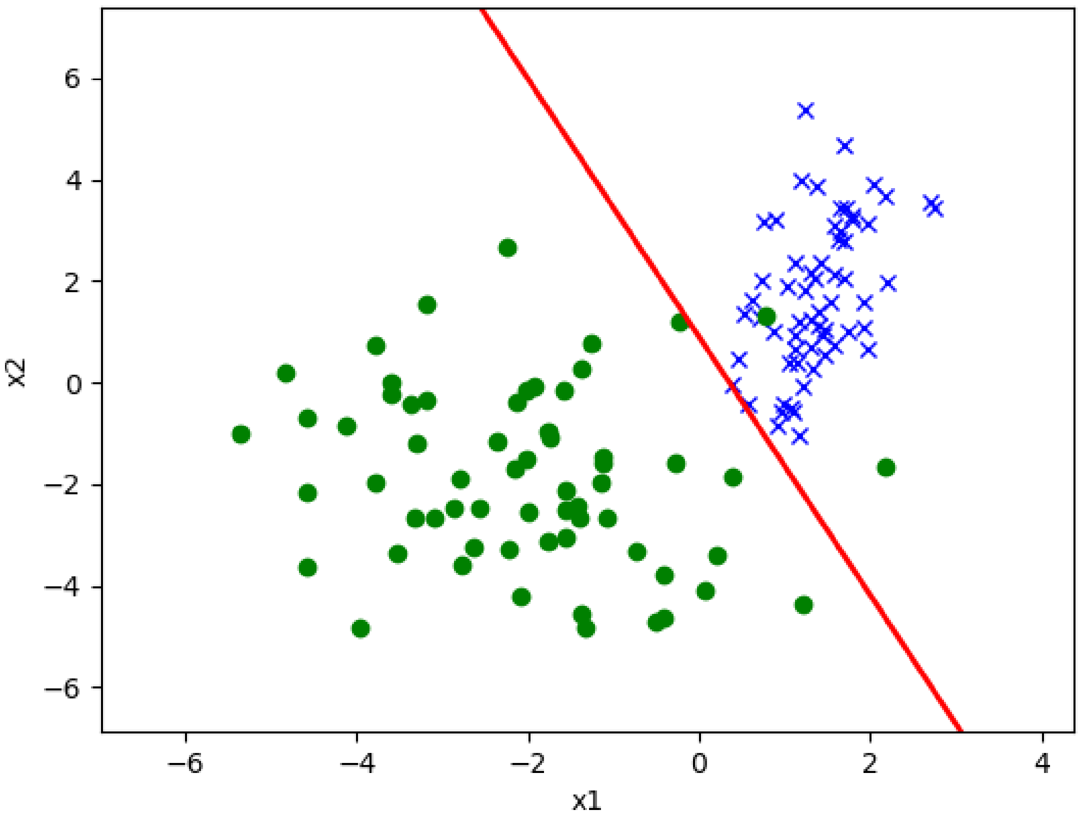


Figure 3: t-labels

(d)

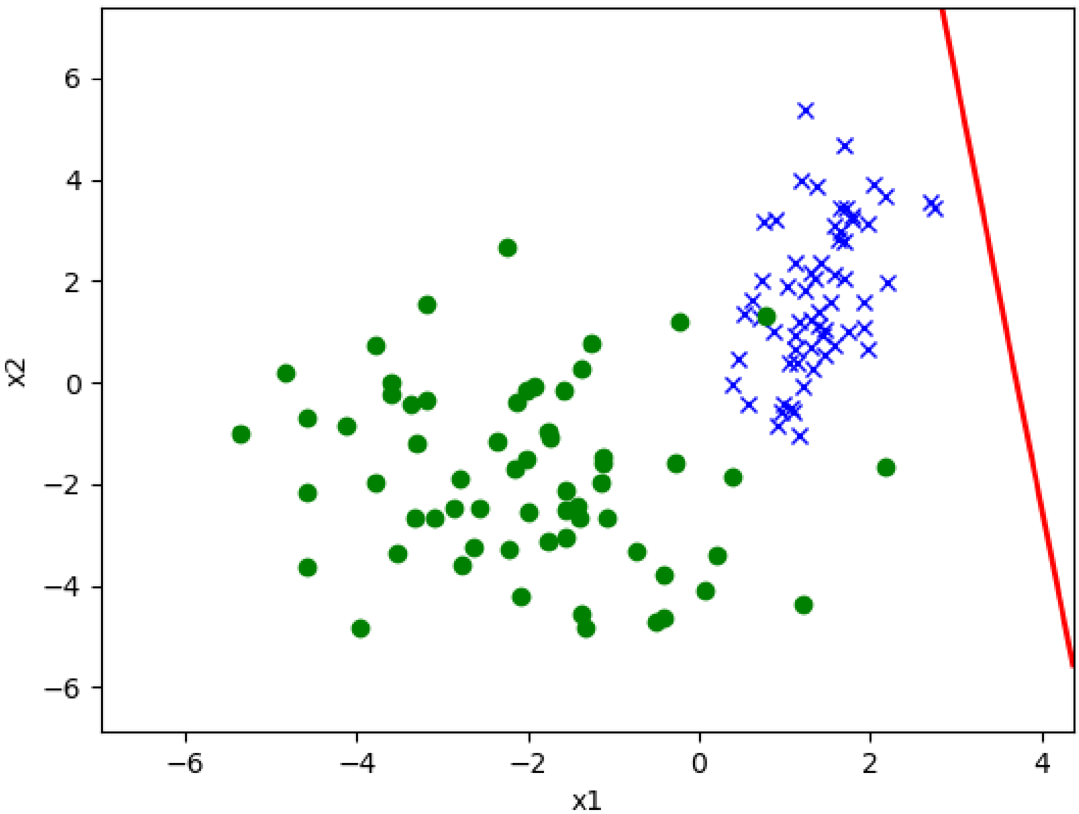


Figure 4: y-labels

Exercise 3

(a)

$$p(y;\lambda) = \frac{1}{y!}\exp(y\ln(\lambda) - \lambda) = b(y)\exp(T(y)\eta - a(\eta))$$
$$b(y) = \frac{1}{y!}, T(y) = y, \eta = \ln(\lambda), a(\eta) = \exp(\eta)$$

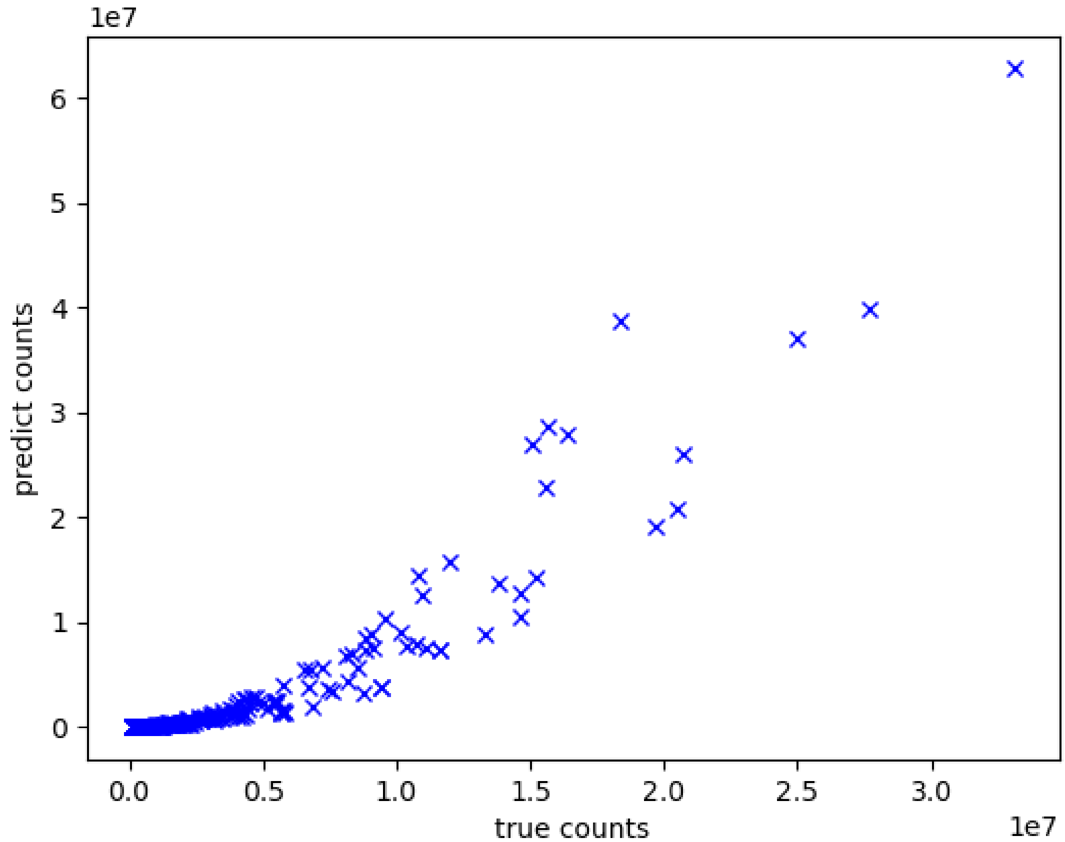
(b)

$$g(\eta) = \mathbb{E}[T(y);\eta] = \mathbb{E}[y;\eta], y \sim \text{Poisson}(\lambda) \implies g(\eta) = \lambda = \exp(\eta)$$

(c)

$$p(y^{(i)}|x^{(i)};\theta) = \frac{1}{y!}\exp(y^{(i)}\theta^T x^{(i)} - \exp(\theta^T x^{(i)}))$$
$$\frac{\partial}{\partial\theta_j}\log p(y^{(i)}|x^{(i)};\theta) = \frac{\partial}{\partial\theta_j}\left(y^{(i)}\theta^T x^{(i)} - \exp(\theta^T x^{(i)})\right)$$
$$= x_j^{(i)}(y^{(i)} - \exp(\theta^T x^{(i)})), \theta_j := \theta_j + \alpha x_j^{(i)}(y^{(i)} - \exp(\theta^T x^{(i)}))$$

(d)



1e7
6
5
4
3
2
1
0
predict counts
0.0
0.5
1.0
1.5
2.0
2.5
3.0
true counts
1e7

Exercise 4

(a)

$$
\begin{aligned}
&\mathbb{E}[y|x;\theta] = \mathbb{E}[b(y)\exp(\eta y - a(\eta))] = \int_{-\infty}^{\infty} yb(y)\exp(\eta y - a(\eta)) \\
&\frac{\partial a(\eta)}{\partial \eta}\int_{-\infty}^{\infty} p(y;\eta) = \int_{-\infty}^{\infty}\frac{\partial a(\eta)}{\partial \eta}p(y;\eta) = \int_{-\infty}^{\infty} b(y)(y - \frac{\partial a(\eta)}{\partial \eta}\exp(\eta y - a(\eta)) \\
&= \int_{-\infty}^{\infty} yb(y)\exp(\eta y - a(\eta)) - \frac{\partial a(\eta)}{\partial \eta}\int_{-\infty}^{\infty} b(y)\exp(\eta y - a(\eta)) = \mathbb{E}[y|x;\theta] - \frac{\partial a(\eta)}{\partial \eta} \\
&\mathbb{E}[y|x;\theta] - \frac{\partial a(\eta)}{\partial \eta} = \frac{\partial a(\eta)}{\partial \eta}\int_{-\infty}^{\infty} p(y;\eta) = 0 \implies \mathbb{E}[y|x;\theta] = \frac{\partial a(\eta)}{\partial \eta}
\end{aligned}
$$

(b)

$$
\begin{aligned}
&\frac{\partial^2 a(\eta)}{\partial \eta^2} = \frac{\partial}{\partial \eta}\mathbb{E}[y|x;\theta] = \int_{-\infty}^{\infty}\frac{\partial}{\partial \eta}yp(y|x;\theta) = \int_{-\infty}^{\infty} yb(y)(y - \frac{\partial a(\eta)}{\partial \eta}\exp(\eta y - a(\eta)) \\
&= \int_{-\infty}^{\infty} y^2 b(y)\exp(\eta y - a(\eta)) - \frac{\partial a(\eta)}{\partial \eta}\int_{-\infty}^{\infty} yb(y)\exp(\eta y - a(\eta)) = \mathbb{E}[y^2|x;\theta] - \mathbb{E}[y|x;\theta]^2 = \mathrm{Var}[y|x;\theta]
\end{aligned}
$$

(c)

$$
\begin{aligned}
&\frac{\partial}{\partial \theta_k \theta_j}\ell(\theta) = -\sum_{i=1}^{m}\frac{\partial}{\partial \theta_k \theta_j}\log p(y^{(i)}|x^{(i)};\theta) = -\sum_{i=1}^{m}\frac{\partial}{\partial \theta_k \theta_j}(y^{(i)}\eta - a(\theta^T x^{(i)})) \\
&= \sum_{i=1}^{m}\frac{\partial}{\partial \theta_k \theta_j}a(\theta^T x^{(i)}) = \sum_{i=1}^{m} x_j^{(i)}x_k^{(i)}\frac{\partial^2 a(\eta)}{\partial \eta^2}(\theta^T x^{(i)}) = \sum_{i=1}^{m} x_j^{(i)}x_k^{(i)}\mathrm{Var}[Y|X;\theta] \\
&\langle Hz, z\rangle = \sum_{i=1}^{m}\sum_{j=1}^{n}\sum_{k=1}^{n} z_j z_k x_j^{(i)}x_k^{(i)}\mathrm{Var}[y|x^{(i)};\theta] = \sum_{i=1}^{m}(z^T x^{(i)})^2\mathrm{Var}[y|x^{(i)};\theta] \geq 0
\end{aligned}
$$

1 Exercise 5

(a)

(i).

$$W_{ij} = \begin{cases} 0, & \text{if } i \neq j \text{ odd} \\ \frac{1}{2}w^{(i)}, & \text{if } i = j \end{cases}$$

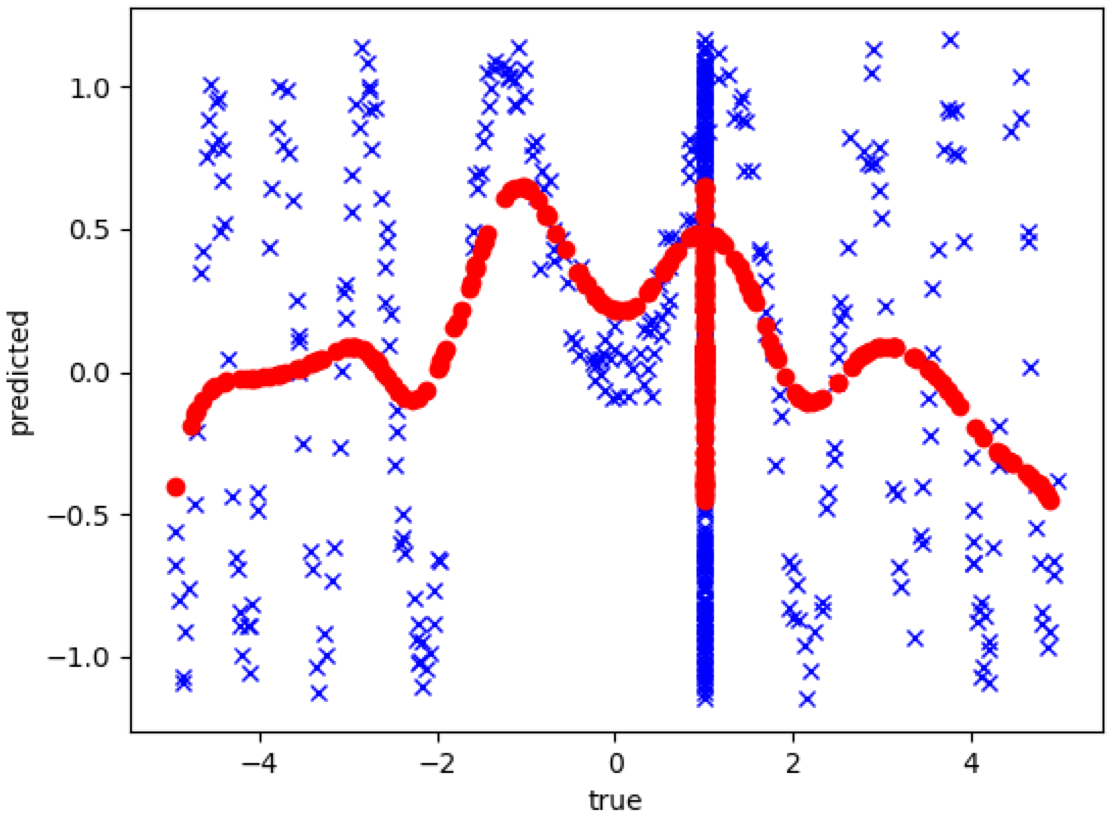
(ii).

$$\begin{aligned}
&\nabla_\theta J(\theta) = \nabla_\theta (X\theta - y)^T W (X\theta - y) = \nabla_\theta \Big((X\theta)^T W X\theta - (X\theta)^T W y - y^T W X\theta + y^T W y) \Big) \\
&= \nabla_\theta (X\theta)^T W X\theta - \nabla_\theta (X\theta)^T W y - \nabla_\theta y^T W X\theta = \nabla_\theta (X\theta)^T W X\theta - 2\nabla_\theta y^T W X\theta \\
&= \nabla_\theta (X\theta)^T W X\theta - 2\nabla_\theta (X^T W y)^T \theta = 2X^T W X\theta - 2X^T W y \\
&\nabla_\theta J(\theta) = 0 \implies X^T W X\theta = X^T W y \implies \theta = (X^T W X)^{-1} X^T W y
\end{aligned}$$

(iii).

$$\begin{aligned}
&\ell(\theta) = \log \prod_{i=1}^{m} p(y^{(i)}|x^{(i)};\theta) = \sum_{i=1}^{m} \log \frac{1}{\sqrt{2\pi}\sigma^{(i)}} \exp\Big(- \frac{(y^{(i)} - \theta^T x^{(i)})^2}{2(\sigma^{(i)})^2} \Big) \\
&= \sum_{i=1}^{m} \log \frac{1}{\sqrt{2\pi}\sigma^{(i)}} - \sum_{i=1}^{m} \frac{(y^{(i)} - \theta^T x^{(i)})^2}{2(\sigma^{(i)})^2}, \frac{\partial \ell(\theta)}{\partial \theta_j} = \sum_{i=1}^{m} w^{(i)} (y^{(i)} - \theta^T x^{(i)}) x_j^{(i)}, w^{(i)} = -\frac{1}{(\sigma^{(i)})^2}
\end{aligned}$$

(b)



Seems to be underfitting, the MSE is 0.331

(c)

$\tau = 0.05$ got the lowest MSE on the validation set(0.012), and 0.017 on the test set

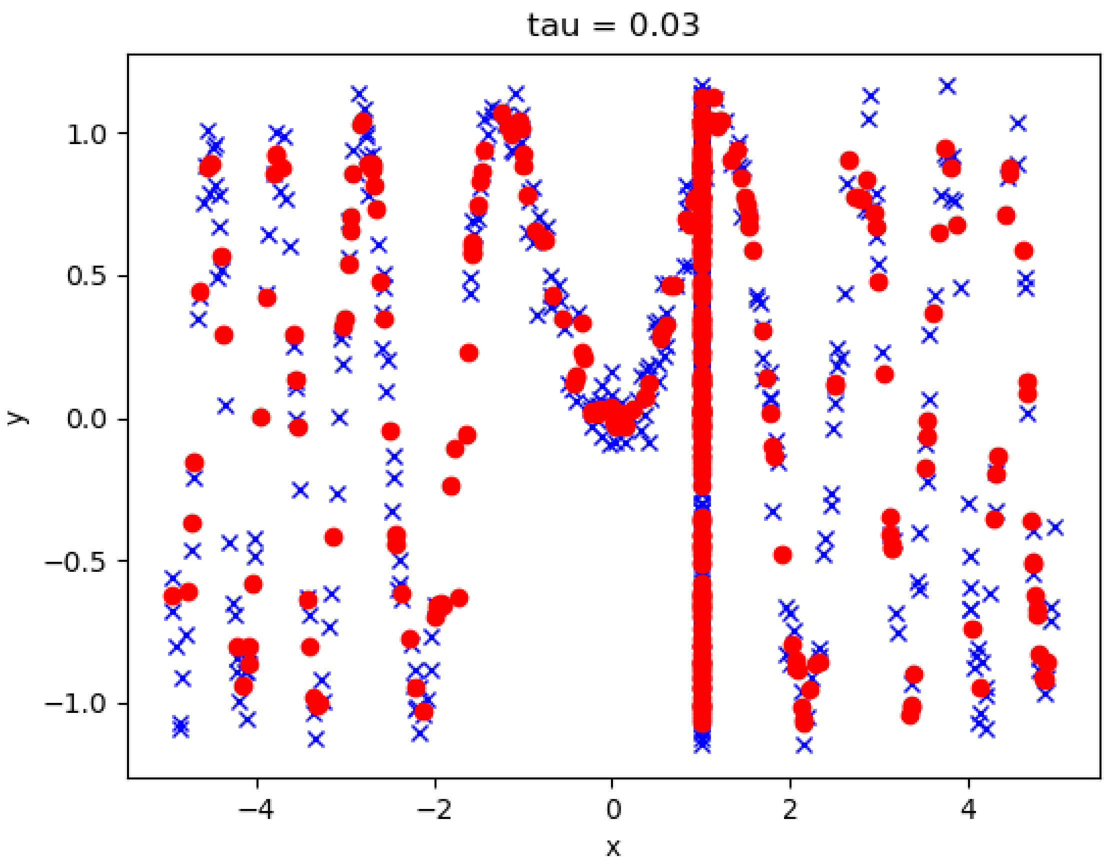


Figure 5: $\tau = 0.03$

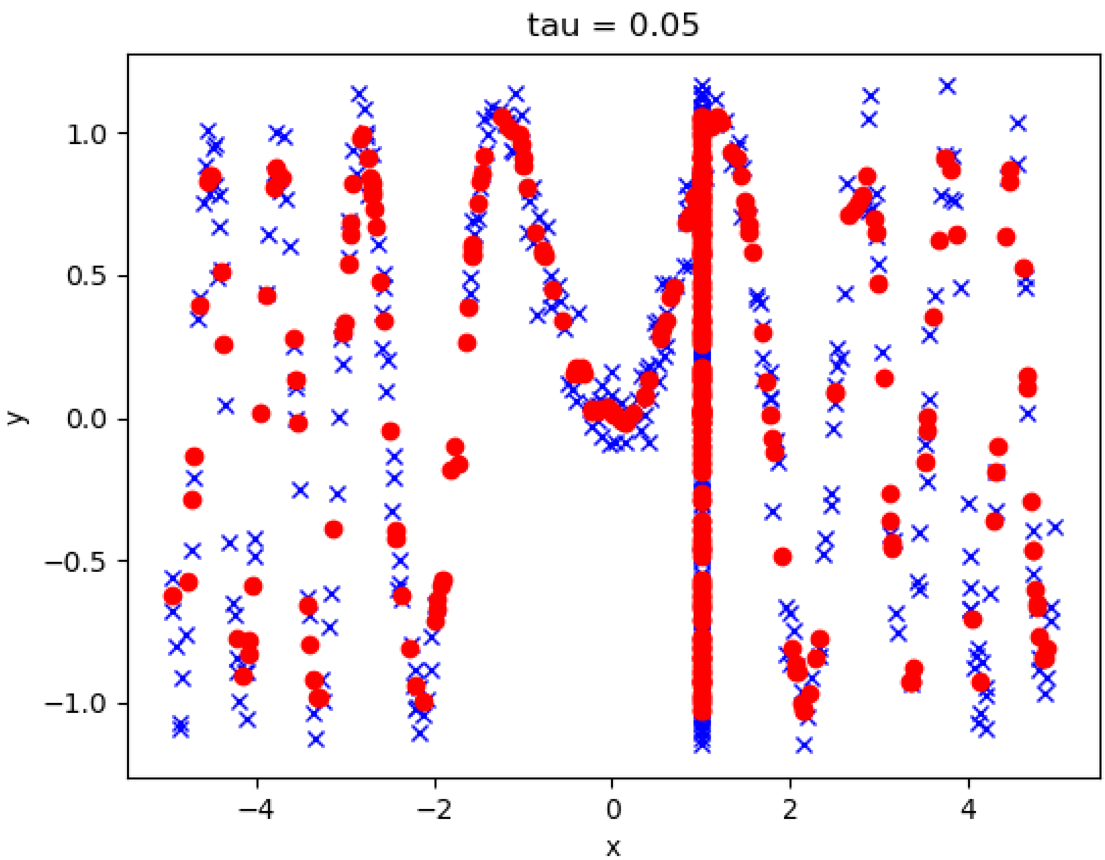


Figure 6: $\tau = 0.05$

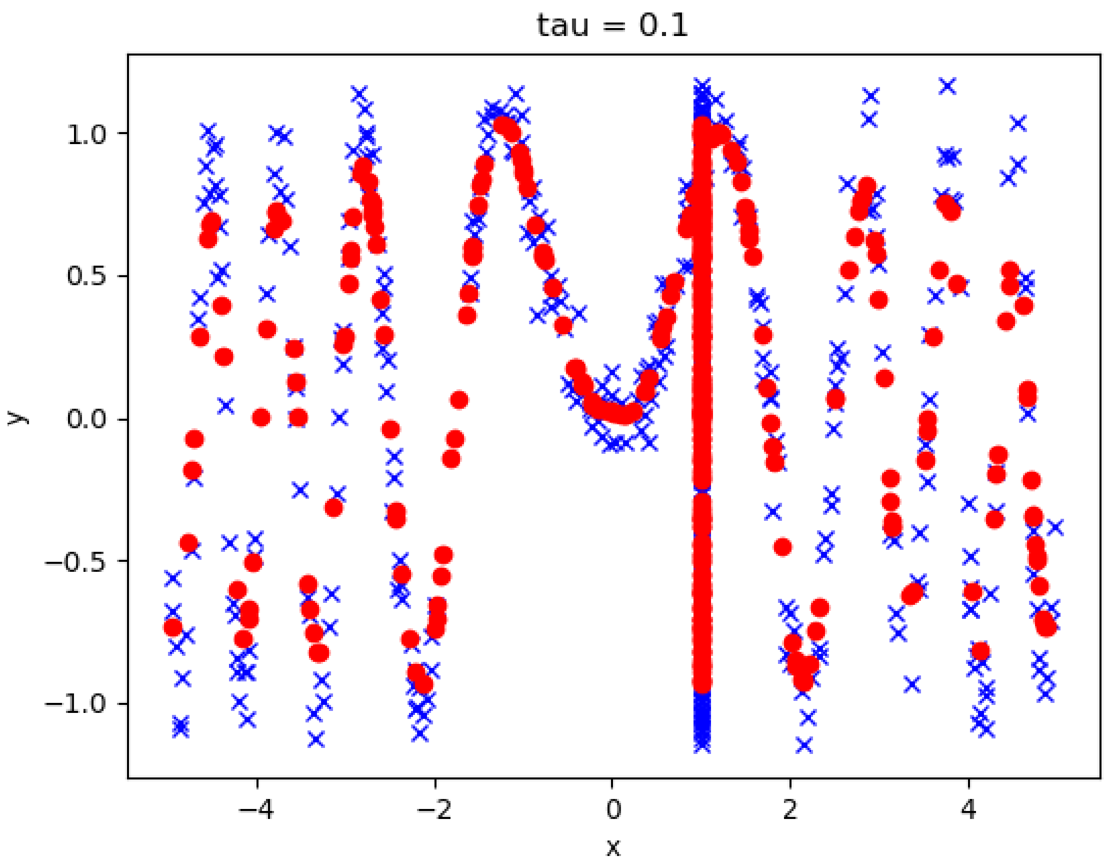


Figure 7: $\tau = 0.1$

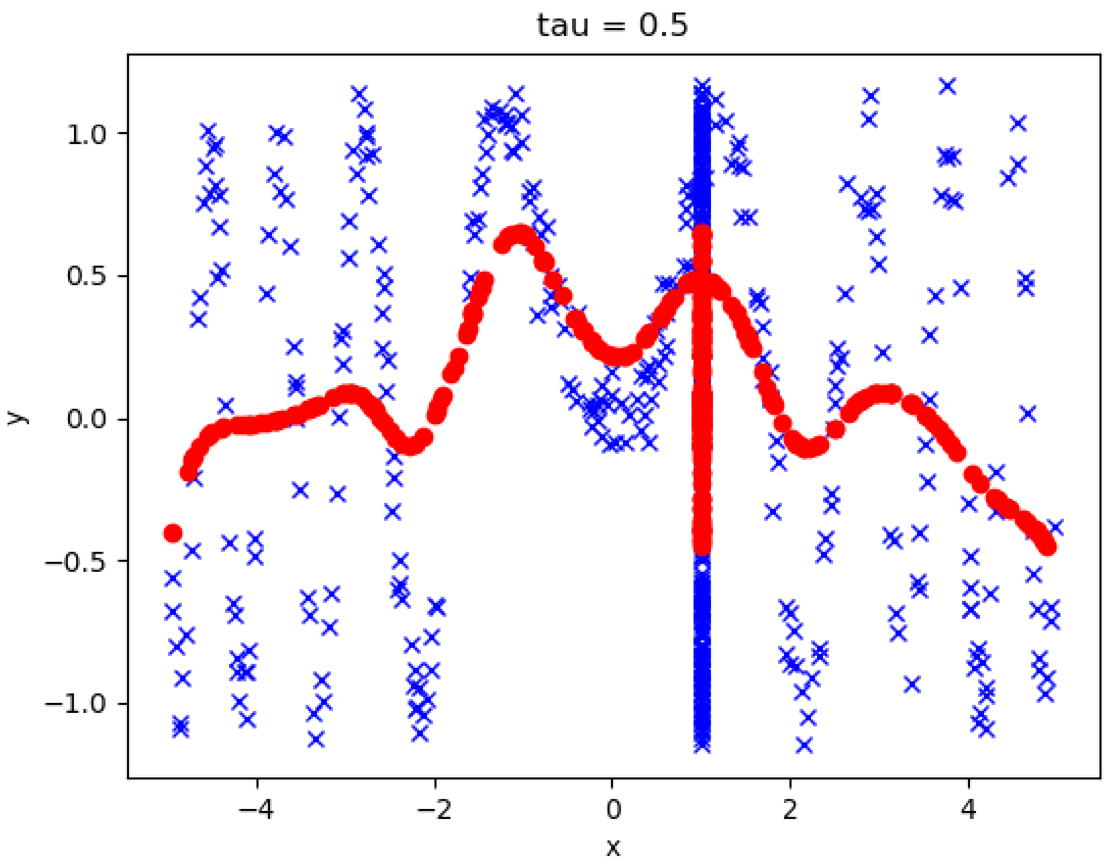


Figure 8: $\tau = 0.5$

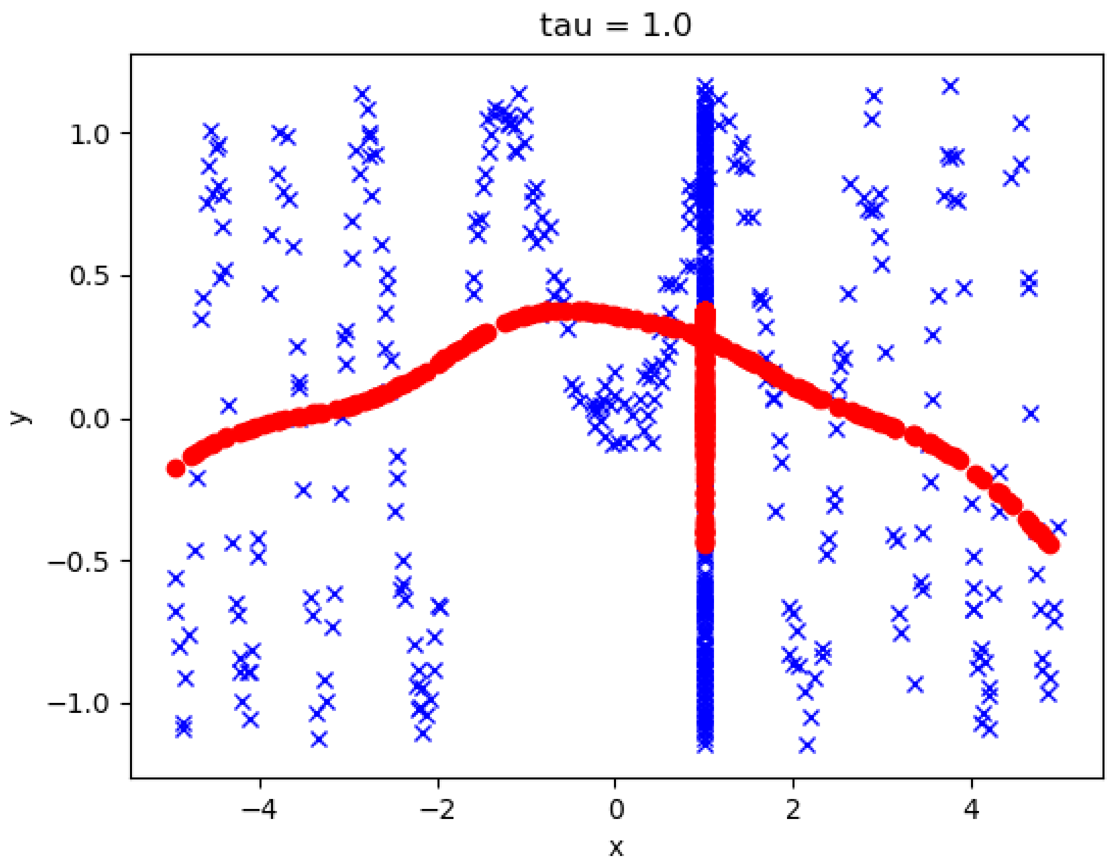


Figure 9: $\tau = 1.0$

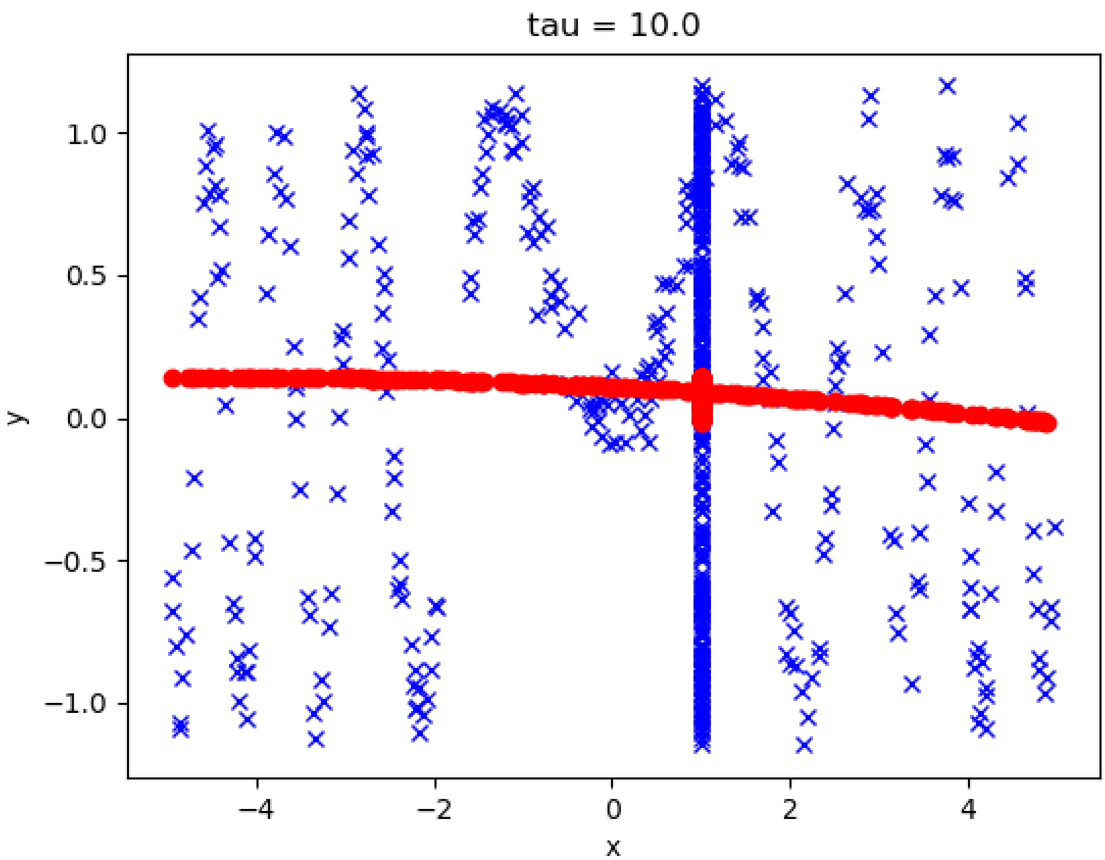


Figure 10: $\tau = 10.0$