CS 229, Autumn 2016
Problem Set #1 Solutions: Supervised Learning

Qingxin6174

1. (a) *Proof.* $J(\theta) = -\frac{1}{m}\sum_{i=1}^{m} \log g\left(y^{(i)}\theta^T x^{(i)}\right)$,

$$H_\theta J(\theta) = \nabla_\theta \left(\nabla_\theta J(\theta)\right)^T = -\nabla_\theta \left(\frac{1}{m}\sum_{i=1}^{m} (1-g)\, y^{(i)} \left(x^{(i)}\right)^T\right) = \frac{1}{m}\sum_{i=1}^{m} g\,(1-g) \left(y^{(i)}\right)^2 x^{(i)} \left(x^{(i)}\right)^T,$$

$$\forall z,\ z^T H z = \frac{1}{m}\sum_{i=1}^{m} g\,(1-g) \left(y^{(i)}\right)^2 z^T x^{(i)} \left(x^{(i)}\right)^T z = \frac{1}{m}\sum_{i=1}^{m} g\,(1-g) \left(y^{(i)} z^T x^{(i)}\right)^2 \geqslant 0. \qquad \square$$

(b) $l(\theta) = \sum_{i=1}^{m} y^{(i)} \log g\left(\theta^T x^{(i)}\right) + \left(1-y^{(i)}\right) \log\left(1 - \left(\theta^T x^{(i)}\right)\right)$,

$$\nabla_\theta l(\theta) = \sum_{i=1}^{m} \left(y^{(i)} - g\right) x^{(i)},\quad H_\theta l(\theta) = \sum_{i=1}^{m} \nabla_\theta \left(y^{(i)} - g\right) \left(x^{(i)}\right)^T = \sum_{i=1}^{m} -g\,(1-g)\, x^{(i)} \left(x^{(i)}\right)^T,$$

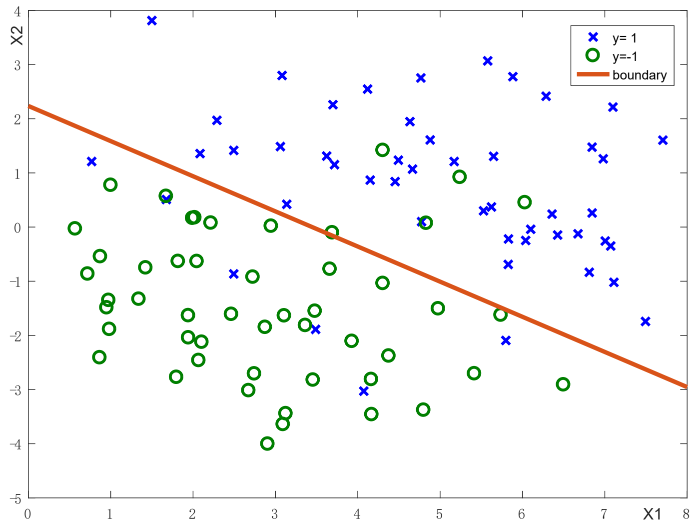
$\theta := \theta - H^{-1}\nabla_\theta l(\theta)$.

The MATLAB code are as follows:

```
[m,n] = size(logisticx); X = [logisticx,ones(m,1)]; Y = (logisticy > 0);
Logistic = @(theta,X) 1./(1+exp(-X*theta));
Gradient = @(theta,X,Y) X'*(Y-Logistic(theta,X));
Hessian = @(theta,X,Y) ...
    -X'*(Logistic(theta,X).*(1-Logistic(theta,X))*ones(1,n+1).*X);
err = 1; theta = zeros(n+1,1);
while err > 1e-28
    ERR = Hessian(theta,X,Y) \ Gradient(theta,X,Y);
    theta = theta - ERR; err = ERR'*ERR;
end
```

$\theta = (0.760371535897677, 1.17194674156714, -2.6205115971802)^T$.

(c)



2. (a) $p(y;\lambda) = \dfrac{e^{-\lambda}\lambda^y}{y!} = \dfrac{1}{y!}e^{y\log\lambda-\lambda}$. The Poisson distribution is in the exponential family, with $b(y) = \dfrac{1}{y!},\ \ \eta = \log\lambda,\ \ T(y) = y,\ \ a(\eta) = e^{\eta}$.

   (b) $E\left[y|x;\lambda\right] = \lambda = e^{\eta}$.

   (c) $l(\theta) = \sum_{i=1}^{m} \log p\left(y^{(i)}|x^{(i)};\theta\right) = \sum_{i=1}^{m} \log \dfrac{1}{y^{(i)}!}e^{y^{(i)}\theta^T x^{(i)} - e^{\theta^T x^{(i)}}} = \sum_{i=1}^{m} y^{(i)}\theta^T x^{(i)} - e^{\theta^T x^{(i)}} - \log y^{(i)}!\,,$

   $\nabla_\theta l(\theta) = \sum_{i=1}^{m} y^{(i)}\nabla_\theta\theta^T x^{(i)} - \nabla_\theta e^{\theta^T x^{(i)}} = \sum_{i=1}^{m} y^{(i)}x^{(i)} - e^{\theta^T x^{(i)}}x^{(i)} = \sum_{i=1}^{m}\left(y^{(i)} - e^{\theta^T x^{(i)}}\right)x^{(i)},$

   this gives the update rule: $\theta := \theta + \alpha\nabla_\theta l(\theta)$,

$$\theta := \theta - \alpha\sum_{i=1}^{m}\left(h_\theta(x^{(i)}) - y^{(i)}\right)x^{(i)}, \quad \text{or} \quad \theta_j := \theta_j - \alpha\sum_{i=1}^{m}\left(h_\theta(x^{(i)}) - y^{(i)}\right)x_j^{(i)}.$$

   (d) *Proof.* $p(y;\eta) = b(y)e^{\eta y - a(\eta)},\ \int_y p(y;\eta)\,\mathrm{d}y = 1 \Rightarrow$

   $0 = \dfrac{\mathrm{d}}{\mathrm{d}\eta}\int_y b(y)e^{\eta y-a(\eta)}\,\mathrm{d}y = \int_y b(y)e^{\eta y-a(\eta)}\left(y - a'(\eta)\right)\,\mathrm{d}y = E\left[y|x\right] - a'(\eta) = h(x) - a'(\eta),$

   $l(\theta) = \sum_{i=1}^{m}\log p\left(y^{(i)}|x^{(i)};\theta\right) = \sum_{i=1}^{m}\log b(y^{(i)})e^{\eta^{(i)}y^{(i)} - a(\eta^{(i)})} = \sum_{i=1}^{m}\eta^{(i)}y^{(i)} - a(\eta^{(i)}) - \log b(y^{(i)})\,,$

   $\nabla_\theta l(\theta) = \sum_{i=1}^{m} y^{(i)}x^{(i)} - a'(\eta^{(i)})x^{(i)} = \sum_{i=1}^{m}\left(y^{(i)} - h(x^{(i)})\right)x^{(i)},$

   this gives the update rule: $\theta := \theta + \alpha\nabla_\theta l(\theta)$: $\theta := \theta - \alpha\sum_{i=1}^{m}\left(h(x^{(i)}) - y^{(i)}\right)x^{(i)}$. □

3. (a) *Proof.* According to the bayesian formula

$$\begin{aligned}
p\left(y=1\mid x\right) &= \frac{p\left(x\mid y=1\right)p\left(y=1\right)}{p\left(x\mid y=1\right)p\left(y=1\right) + p\left(x\mid y=-1\right)p\left(y=-1\right)}\\
&= \frac{\phi\,\frac{1}{(2\pi)^{\frac{n}{2}}|\Sigma|^{\frac{1}{2}}}\,e^{-\frac{1}{2}(x-\mu_1)^T\Sigma^{-1}(x-\mu_1)}}{\phi\,\frac{1}{(2\pi)^{\frac{n}{2}}|\Sigma|^{\frac{1}{2}}}\,e^{-\frac{1}{2}(x-\mu_1)^T\Sigma^{-1}(x-\mu_1)} + (1-\phi)\,\frac{1}{(2\pi)^{\frac{n}{2}}|\Sigma|^{\frac{1}{2}}}\,e^{-\frac{1}{2}(x-\mu_{-1})^T\Sigma^{-1}(x-\mu_{-1})}}\\
&= \frac{1}{1 + \frac{(1-\phi)}{\phi}\,e^{\frac{1}{2}(x-\mu_1)^T\Sigma^{-1}(x-\mu_1) - \frac{1}{2}(x-\mu_{-1})^T\Sigma^{-1}(x-\mu_{-1})}}\\
&= \frac{1}{1 + e^{-\left[\left(\mu_1^T - \mu_{-1}^T\right)\Sigma^{-1}x + \log\frac{\phi}{1-\phi} - \frac{1}{2}\left(\mu_1^T\Sigma^{-1}\mu_1 - \mu_{-1}^T\Sigma^{-1}\mu_{-1}\right)\right]}}\\
p\left(y=-1\mid x\right) &= \frac{1}{1 + e^{+\left[\left(\mu_1^T - \mu_{-1}^T\right)\Sigma^{-1}x + \log\frac{\phi}{1-\phi} - \frac{1}{2}\left(\mu_1^T\Sigma^{-1}\mu_1 - \mu_{-1}^T\Sigma^{-1}\mu_{-1}\right)\right]}},
\end{aligned}$$

   assume that

$$\begin{aligned}
\theta &= \Sigma^{-1}\left(\mu_1 - \mu_{-1}\right)\\
\theta_0 &= \log\frac{\phi}{1-\phi} - \frac{1}{2}\left(\mu_1^T\Sigma^{-1}\mu_1 - \mu_{-1}^T\Sigma^{-1}\mu_{-1}\right),
\end{aligned}$$

   $p\left(y\mid x;\phi,\Sigma,\mu_{-1},\mu_1\right) = \dfrac{1}{1 + e^{-y(\theta^T x + \theta_0)}}$. □

(b) Together with (c).

(c) $l(\phi, \mu_{-1}, \mu_1, \Sigma)$

$$
\begin{aligned}
&= \log \prod_{i=1}^{m} p\left(x^{(i)}, y^{(i)}; \phi, \mu_{-1}, \mu_1, \Sigma\right) \\
&= \log \prod_{i=1}^{m} p\left(x^{(i)} \mid y^{(i)}; \mu_{-1}, \mu_1, \Sigma\right) p\left(y^{(i)}; \phi\right) \\
&= \log \prod_{i=1}^{m} \frac{1}{(2\pi)^{\frac{n}{2}} |\Sigma|^{\frac{1}{2}}} \left[\phi\, e^{-\frac{1}{2}\left(x^{(i)}-\mu_1\right)^T \Sigma^{-1}\left(x^{(i)}-\mu_1\right)}\right]^{1\{y^{(i)}=1\}} \left[(1-\phi)\, e^{-\frac{1}{2}\left(x^{(i)}-\mu_{-1}\right)^T \Sigma^{-1}\left(x^{(i)}-\mu_{-1}\right)}\right]^{1\{y^{(i)}=-1\}} \\
&= \sum_{i=1}^{m} 1\{y^{(i)}=1\} \left[\log\phi - \frac{1}{2}\left(x^{(i)}-\mu_1\right)^T \Sigma^{-1}\left(x^{(i)}-\mu_1\right)\right] + \\
&\qquad 1\{y^{(i)}=-1\} \left[\log(1-\phi) - \frac{1}{2}\left(x^{(i)}-\mu_{-1}\right)^T \Sigma^{-1}\left(x^{(i)}-\mu_{-1}\right)\right] - \frac{n}{2}\log 2\pi - \frac{1}{2}\log|\Sigma|,
\end{aligned}
$$

$$
\begin{aligned}
\frac{\partial l}{\partial \phi} &= \sum_{i=1}^{m} \frac{1\{y^{(i)}=1\}}{\phi} - \frac{1\{y^{(i)}=-1\}}{1-\phi}, \\
\frac{\partial l}{\partial \mu_1} &= \sum_{i=1}^{m} -\frac{1}{2} 1\{y^{(i)}=1\} \frac{\partial}{\partial \mu_1} \mathrm{tr}\left(x^{(i)}-\mu_1\right)^T \Sigma^{-1}\left(x^{(i)}-\mu_1\right) = \Sigma^{-1}\sum_{i=1}^{m} 1\{y^{(i)}=1\}\left(\mu_1 - x^{(i)}\right), \\
\frac{\partial l}{\partial \mu_{-1}} &= \Sigma^{-1}\sum_{i=1}^{m} 1\{y^{(i)}=-1\}\left(\mu_{-1} - x^{(i)}\right), \\
\frac{\partial l}{\partial \Sigma} &= \sum_{i=1}^{m} 1\{y^{(i)}=1\}\left(-\frac{1}{2}\right) \frac{\partial}{\partial \Sigma} \mathrm{tr}\left(x^{(i)}-\mu_1\right)^T \Sigma^{-1}\left(x^{(i)}-\mu_1\right) + \\
&\qquad 1\{y^{(i)}=-1\}\left(-\frac{1}{2}\right) \frac{\partial}{\partial \Sigma} \mathrm{tr}\left(x^{(i)}-\mu_{-1}\right)^T \Sigma^{-1}\left(x^{(i)}-\mu_{-1}\right) - \frac{1}{2}\frac{\partial}{\partial \Sigma}\log|\Sigma| \\
&= \frac{1}{2}\sum_{i=1}^{m} 1\{y^{(i)}=1\}\Sigma^{-1}\left(x^{(i)}-\mu_1\right)\left(x^{(i)}-\mu_1\right)^T \Sigma^{-1} + \\
&\qquad 1\{y^{(i)}=-1\}\Sigma^{-1}\left(x^{(i)}-\mu_{-1}\right)\left(x^{(i)}-\mu_{-1}\right)^T \Sigma^{-1} - \left(\Sigma^{-1}\right)^T \\
&= \frac{1}{2}\Sigma^{-1}\left[\sum_{i=1}^{m}\left(x^{(i)}-\mu_{y^{(i)}}\right)\left(x^{(i)}-\mu_{y^{(i)}}\right)^T\right]\Sigma^{-1} - \frac{m}{2}\Sigma^{-1},
\end{aligned}
$$

let $\dfrac{\partial l}{\partial \phi} = 0,\ \dfrac{\partial l}{\partial \mu_1} = 0,\ \dfrac{\partial l}{\partial \mu_{-1}} = 0,\ \dfrac{\partial l}{\partial \Sigma} = 0,$

$$
\begin{aligned}
\phi &= \frac{1}{m}\sum_{i=1}^{m} 1\{y^{(i)}=1\} \\
\mu_1 &= \frac{\sum_{i=1}^{m} 1\{y^{(i)}=1\}x^{(i)}}{\sum_{i=1}^{m} 1\{y^{(i)}=1\}} \\
\mu_{-1} &= \frac{\sum_{i=1}^{m} 1\{y^{(i)}=-1\}x^{(i)}}{\sum_{i=1}^{m} 1\{y^{(i)}=-1\}} \\
\Sigma &= \frac{1}{m}\sum_{i=1}^{m}\left(x^{(i)}-\mu_{y^{(i)}}\right)\left(x^{(i)}-\mu_{y^{(i)}}\right)^T.
\end{aligned}
$$

Remark: $\dfrac{\partial\,\mathrm{tr}(X^{-1}A)}{\partial X} = -X^{-1}A^TX^{-1}$, $\dfrac{\partial \log|X|}{\partial X} = (X^{-1})^T$.

4. (a) *Proof.* $x = Az$, and assume that $x^k = Az^k$,

$$
\begin{aligned}
x^{k+1} &= x^k - H_x^{-1}\nabla_x f(x)\Big|_{x=x^k}\\
z^{k+1} &= z^k - H_z^{-1}\nabla_z g(z)\Big|_{z=z^k} = z^k - H_z^{-1}\nabla_z f(Az)\Big|_{z=z^k} = z^k - H_z^{-1}\nabla_z f(x)\Big|_{z=z^k},\\
\nabla_z f(x) &= \begin{pmatrix} f_1' a_{11} + f_2' a_{21} + \cdots + f_n' a_{n1}\\ f_1' a_{12} + f_2' a_{22} + \cdots + f_n' a_{n2}\\ \vdots \\ f_1' a_{1n} + f_2' a_{2n} + \cdots + f_n' a_{nn}\end{pmatrix} = A^T\nabla_x f(x),\\
H_z f(x) &= \nabla_z(\nabla_z f(x))^T = \nabla_z(\nabla_x f(x))^T A = A^T\nabla_x(\nabla_x f(x))^T A = A^T H_x f(x) A,\\
z^{k+1} &= z^k - H_z^{-1}\nabla_z f(x)\Big|_{z=z^k} = z^k - A^{-1}H_x f(x)(A^T)^{-1}A^T\nabla_x f(x)\Big|_{z=z^k} = A^{-1}x^{k+1}.
\end{aligned}
$$

So Newton's method is invariant to linear reparameterizations. □

(b) $x = Az$, and assume that $x^k = Az^k$,

$$
\begin{aligned}
x^{k+1} &= x^k - \alpha\,\nabla_x f(x)\Big|_{x=x^k}\\
z^{k+1} &= z^k - \alpha\,\nabla_z\, g(z)\Big|_{z=z^k} = z^k - \alpha\,\nabla_z\, f(x)\Big|_{z=z^k} = A^{-1}x^k - \alpha\, A^T\nabla_x f(x)\Big|_{z=z^k},
\end{aligned}
$$

if A is an orthogonal matrix, the gradient descent is invariant to linear reparameterizations.

5. (a) i. $J(\theta)$ is a quadratic form:

$$
J(\theta) = \sum_{i=1}^m \frac{1}{2}\,w^{(i)}\left(\theta^T x^{(i)} - y^{(i)}\right)^2 = (X\theta - \vec{y}\,)^T \begin{pmatrix} \frac{1}{2}w^{(1)} & & & \\ & \frac{1}{2}w^{(2)} & & \\ & & \ddots & \\ & & & \frac{1}{2}w^{(m)}\end{pmatrix} (X\theta - \vec{y}\,),
$$

$W = \dfrac{1}{2}\,\mathrm{diag}\left[w^{(1)}, w^{(2)}, \cdots, w^{(m)}\right]$.

ii. $J(\theta) = (X\theta - \vec{y}\,)^T\, W\,(X\theta - \vec{y}\,) = \theta^T X^T W X\theta - \theta^T X^T W\vec{y} - \vec{y}^T W X\theta + \vec{y}^T W\vec{y}$,
$\nabla_\theta J(\theta) = \nabla_\theta\,\mathrm{tr}\,\theta^T X^T W X\theta - \nabla_\theta\,\mathrm{tr}\,\theta^T X^T W\vec{y} - \nabla_\theta\,\mathrm{tr}\,\vec{y}^T W X\theta = 2X^T W X\theta - 2X^T W\vec{y}$,
apparently $\theta = (X^T W X)^{-1}X^T W\vec{y}$.

iii. Likelihood function:

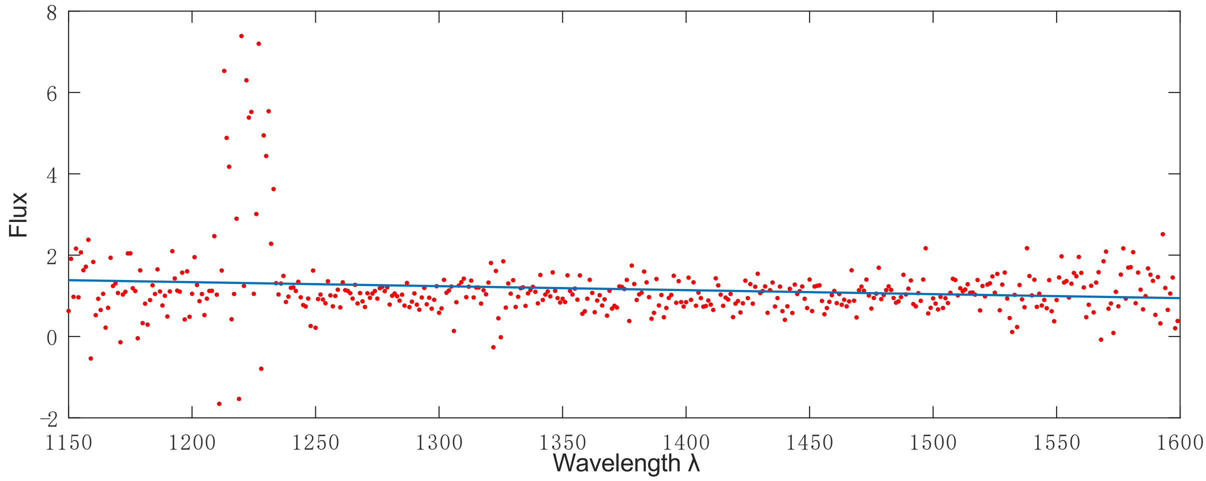
$$
l(\theta) = \log\prod_{i=1}^m \frac{1}{\sqrt{2\pi}\sigma^{(i)}}\, e^{-\frac{(y^{(i)}-\theta^T x^{(i)})^2}{2(\sigma^{(i)})^2}} = \sum_{i=1}^m - \log\sqrt{2\pi}\sigma^{(i)} - \frac{\left(y^{(i)} - \theta^T x^{(i)}\right)^2}{2\left(\sigma^{(i)}\right)^2},
$$

hence, maximizing $l(\theta)$ gives the same answer as minimizing

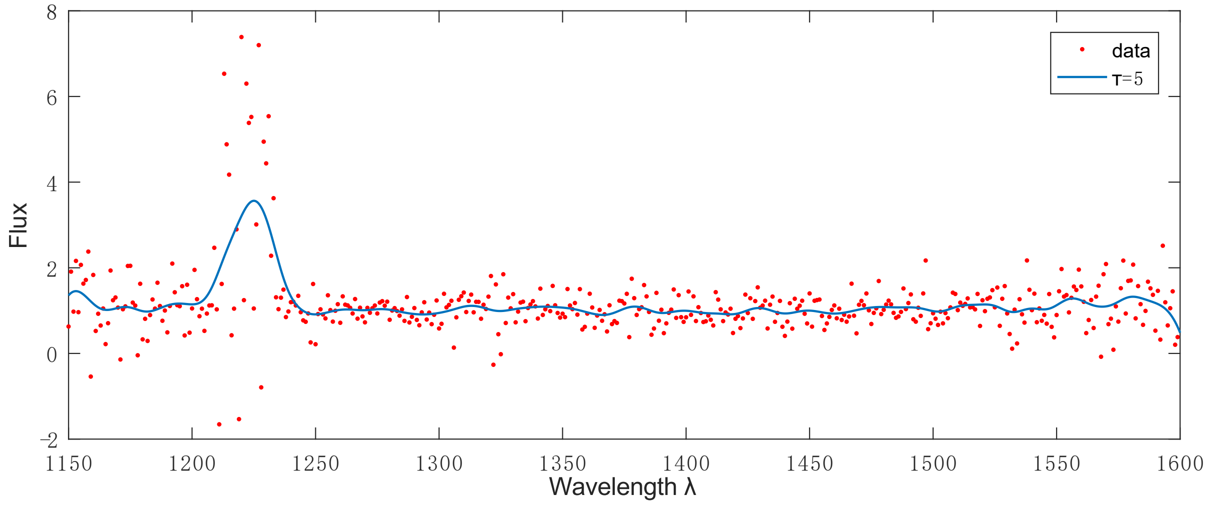
$$
J(\theta) = \sum_{i=1}^m \frac{1}{2\left(\sigma^{(i)}\right)^2}\left(y^{(i)} - \theta^T x^{(i)}\right)^2,
$$

$w^{(i)} = \dfrac{1}{\left(\sigma^{(i)}\right)^2}$.

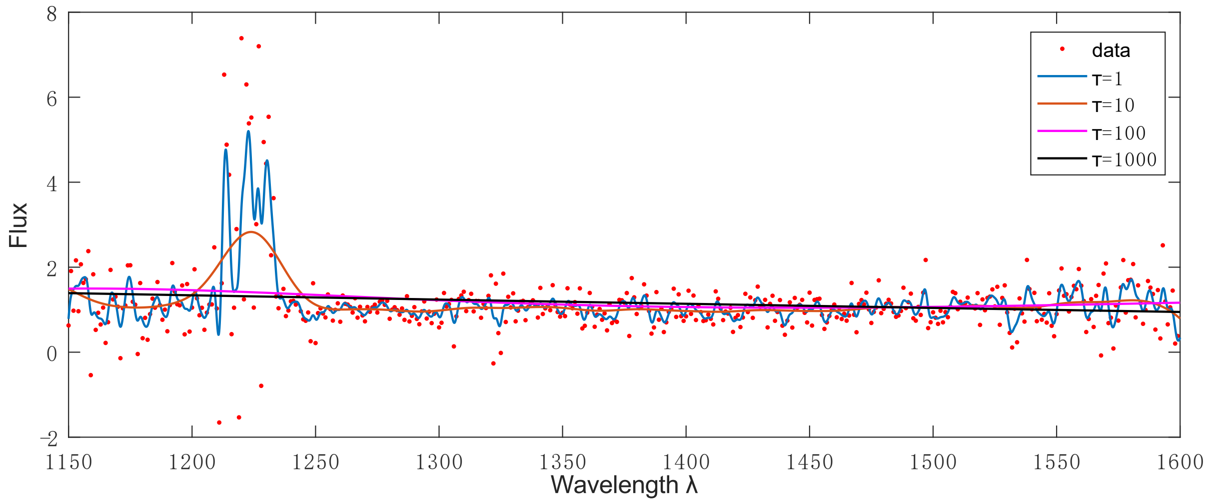
(b) i. $\theta = (-0.000981122145459,\ 2.5133990556)^T$.



ii. $\tau = 5$.



iii. $\tau = 1,\ 10,\ 100$ and 1000.
The curve $\tau = 1$ appeared oscillation, witch is overfitting. As the growth of τ, curve becomes smooth. The curves $\tau \geqslant 100$ are underfitting.



(c) i. My implementation of `lwlinreg.m`:

```
function qso_smoothed = lwlinreg (lambdas, qso, tau)
[x1,x2] = ndgrid(lambdas,lambdas);
W = exp(-(x1-x2).^2/(2*tau^2))/2;
qso_smoothed = zeros(size(qso));
for idx = 1:size(qso,1)
    y = qso(idx,:)';
    for lambda_idx = 1:length(lambdas)
        theta = [lambdas'*(W(:,lambda_idx).*lambdas),  ...
                                lambdas'*W(:,lambda_idx);  ...
                     sum(W(:,lambda_idx).*lambdas),        ...
                               sum(W(:,lambda_idx))      ]\ ...
                    [lambdas'*(W(:,lambda_idx).*y);        ...
                     sum(W(:,lambda_idx).*y)      ];
        qso_smoothed(idx,lambda_idx) = ...
                    theta(1)*lambdas(lambda_idx) + theta(2);
    end
end
```

Use the following two commands to smooth all spectra in the training set:

```
>> train_smoothed = lwlinreg (lambdas, train_qso, 5);
>> test_smoothed = lwlinreg (lambdas, test_qso, 5);
```

ii. Create data:

```
>> train_f_left = train_smoothed(:,1:50);
>> train_f_right = train_smoothed(:,151:end);
>> test_f_left = test_smoothed(:,1:50);
>> test_f_right = test_smoothed(:,151:end);
```

The estimator $\widehat{f_{\text{left}}}$ can be created using **funreg.m**

```
function f_left_estimated = funreg (f_left,f_right,k)
Dist = pdist2(f_right,f_right);
[Dist,Idx] = sort(Dist,1);
ker = @(t) (t<1).*(1-t);
KER = ker(Dist(1:k,:)./(ones(k,1)*Dist(end,:)));
n = size(f_left,2);
f_left_estimated = ...
    reshape(sum( ...
              reshape(reshape(KER,numel(KER),1)*ones(1,n).* ...
              f_left(Idx(1:k,:),:), k,size(Idx,2),n),1), ...
            size(Idx,2),n) ./ ...
    (sum(KER)'*ones(1,n));
```

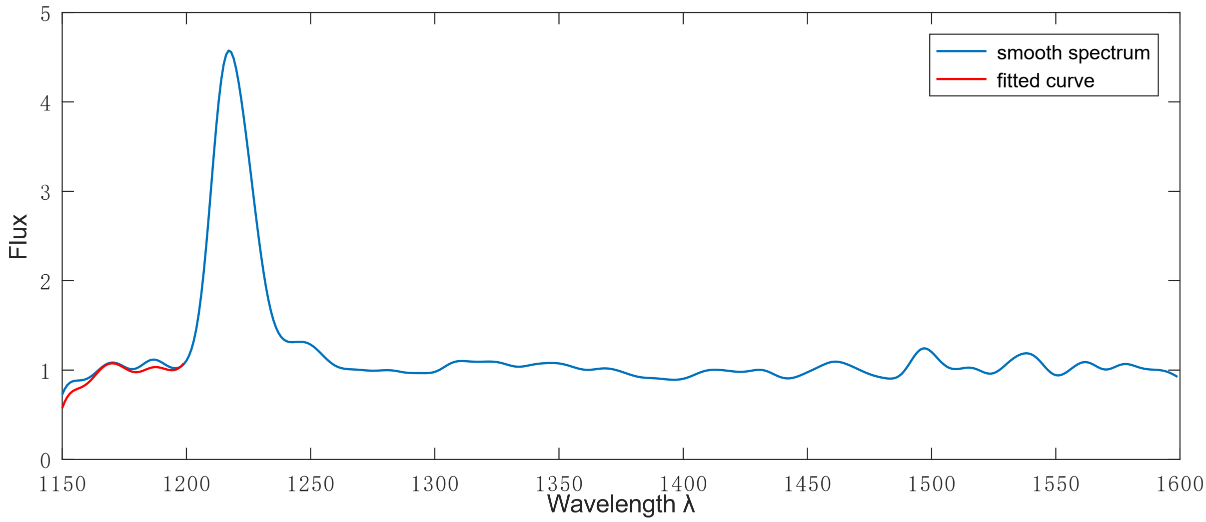
The average training error:

```
>> train_f_left_estimated = funreg(train_f_left, train_f_right, 3);
>> train_err = sum((train_f_left_estimated - train_f_left).^2, 2);
>> mean(train_err)
ans =
    0.7838
```

iii. The average testing error:

```
>> test_f_left_estimated = funreg(test_f_left, test_f_right, 3);
>> test_err = sum((test_f_left_estimated - test_f_left).^2, 2);
>> mean(test_err)
ans =
    0.6596
```

Test example 1:



Test example 6: