Problem Set 1:

1. (a) (i): $\theta^{(t)} = \theta^{(t-1)} - \alpha \nabla J(\theta^{(t-1)})$

$\nabla J(\theta^{(t-1)}) = \frac{\partial(\frac{1}{2}\beta\theta^2)}{\partial\theta} = \beta\theta$

$= \theta^{(t-1)} - \alpha\beta\theta^{(t-1)}$

$= (1-\alpha\beta)\cdot\theta^{(t-1)}$

$= (1-\alpha\beta)^t \theta^{(0)}$

assume $\lim_{t\to\infty} |\theta^{(t)} - \theta^*| = 0$

$\lim_{t\to\infty} |(1-\alpha\beta)^t \theta^{(0)} - \theta^*| = 0$

$\lim_{t\to\infty} (1-\alpha\beta)^t \theta^{(0)} = \theta^*$

$\because \theta^{(0)} \neq 0 \quad \therefore \lim_{t\to\infty} (1-\alpha\beta)^t = \frac{\theta^*}{\theta^{(0)}}$

~~$\theta^{(0)} = 0$~~ to let $\theta^* = 0$

$\lim_{t\to\infty} (1-\alpha\beta)^t = 0$

$-1 < (1-\alpha\beta) < 1$

$-2 < -\alpha\beta < 0$

$2 > \alpha\beta > 0$

$\because \beta > 0 \quad$ ~~$\alpha > 0$~~

$\therefore \frac{2}{\beta} > \alpha > 0$

close form: $\nabla J(\theta) = \beta\theta = 0$

$\because \beta > 0$

$\therefore \theta^* = 0$

$\therefore$ global minimum $\theta^* = 0 = \arg\min_\theta J(\theta)$

1.(a)(ii)

$$|\theta^{(T)} - \theta^*| \le \varepsilon$$

$\because \theta^* = 0$

$\therefore |\theta^{(T)}| \le \varepsilon$

$$|(1-\alpha\beta)^T \theta^{(0)}| \le \varepsilon$$

$$|(1-\alpha\beta)^T| \cdot |\theta^{(0)}| \le \varepsilon$$

$\because \theta^{(0)} \neq 0$

$\therefore |(1-\alpha\beta)^T| \le \frac{\varepsilon}{|\theta^{(0)}|} \Rightarrow |(1-\alpha\beta)|^T \le \underbrace{\frac{\varepsilon}{|\theta^{(0)}|}}_{\text{constant.}}$

$\because 1 \le (1-\alpha\beta) \le 1$

$\therefore 0 \le |1-\alpha\beta| < 1$

when $|1-\alpha\beta| = 0$, $T = 1$

when $0 < |1-\alpha\beta| < 1$, ~~then~~ when $|1-\alpha\beta| \nearrow$, $T \nearrow$

$\therefore$
- when $1-\alpha\beta = 0$, i.e. $\alpha = \frac{1}{\beta}$, $T = 1$
- when $1-\alpha\beta < 0$, i.e. $\alpha > \frac{1}{\beta}$, when $|1-\alpha\beta| \nearrow$, $1-\alpha\beta \searrow$, $\alpha \nearrow$, $T \nearrow$
- when $1-\alpha\beta > 0$, i.e. $\alpha < \frac{1}{\beta}$, when $|1-\alpha\beta| \nearrow$, $1-\alpha\beta \nearrow$, $\alpha \searrow$, $T \nearrow$

$\therefore$ when choose an inappropiate α (e.g. ~~0.00~~ 10^{-100}) can significantly increase T

1.(b)

$$\nabla J(\theta) = \begin{bmatrix} \beta_1 \theta_1 \\ \beta_2 \theta_2 \\ \vdots \\ \beta_d \theta_d \end{bmatrix}$$

let $\vec{\beta}$ be diagonal $(\beta_1, \beta_2, \cdots \beta_d)$

$$\theta^{(t)} = \theta^{(t-1)} - \alpha \nabla J(\theta^{(t-1)})$$

$$= \theta^{(t-1)} - \alpha \vec{\beta} \cdot \theta^{(t-1)}$$

$$= (\vec{I} - \alpha \vec{\beta}) \cdot \theta^{(t-1)}$$

$$= (\vec{I} - \alpha \vec{\beta})^t \cdot \theta^{(0)}$$

$$\lim_{t \to \infty} \| \theta^{(t)} - \theta^* \|_2 = 0$$

$$\lim_{t \to \infty} \| (\vec{I} - \alpha \vec{\beta})^t \theta^{(0)} - \theta^* \|_2 = 0$$

$\rightarrow$

$$\lim_{t \to \infty} (\vec{I} - \alpha \vec{\beta})^t \theta^{(0)} = \theta^*$$

when $\theta^* = \vec{0}$ and $|\vec{I} - \alpha \vec{\beta}| < 1$, it converges.

$\vec{I} - \alpha \vec{\beta}$

$\therefore -1 < 1 - \alpha \beta_i < 1$ for all i in $1, 2, \cdots, d$

$\because \beta_i > 0$

$\therefore \frac{2}{\beta_i} > \alpha > 0$

$\therefore 0 < \alpha < \min\left(\frac{2}{\beta_1}, \frac{2}{\beta_2}, \cdots \frac{2}{\beta_d}\right)$

(12(C)(iii)

Observation:

① for the same learning rate, both A and rotated A result in the same iterations to converge.

② once lr ≥ 0.5, it starts to overshoot during the training. It bounces between the global maximum.

③ the iterations number decreases then increases when lr increases

(d): $J(\theta^t) = J(\theta^{t-1}) + \nabla J(\theta^{t-1})^T (\theta^t - \theta^{t-1}) + \frac{1}{2}(\theta^t - \theta^{t-1})^T \nabla^2 J[\theta^{t-1} + c(\theta^t - \theta^{t-1})](\theta^t - \theta^{t-1})$

$\theta^t = \theta^{t-1} - \alpha \nabla J(\theta^{t-1})$

$\therefore \theta^t - \theta^{t-1} = -\alpha \nabla J(\theta^{t-1})$

$\therefore J(\theta^t) = J(\theta^{t-1}) - \alpha \nabla J(\theta^{t-1})^T \nabla J(\theta^{t-1}) + \frac{1}{2}(-\alpha \nabla J\theta^{t-1})^T \nabla^2 J[\theta^{t-1} + (-c\alpha \nabla J\theta^{t-1})] \cdot (-\alpha \nabla J(\theta^{t-1}))$

$= J(\theta^{t-1}) - \alpha \nabla J(\theta^{t-1})^T \left[I - \frac{1}{2}\alpha \nabla^2 J[\theta^{t-1} - c\alpha \nabla J\theta^{t-1})] \right] \nabla J(\theta^{t-1})$

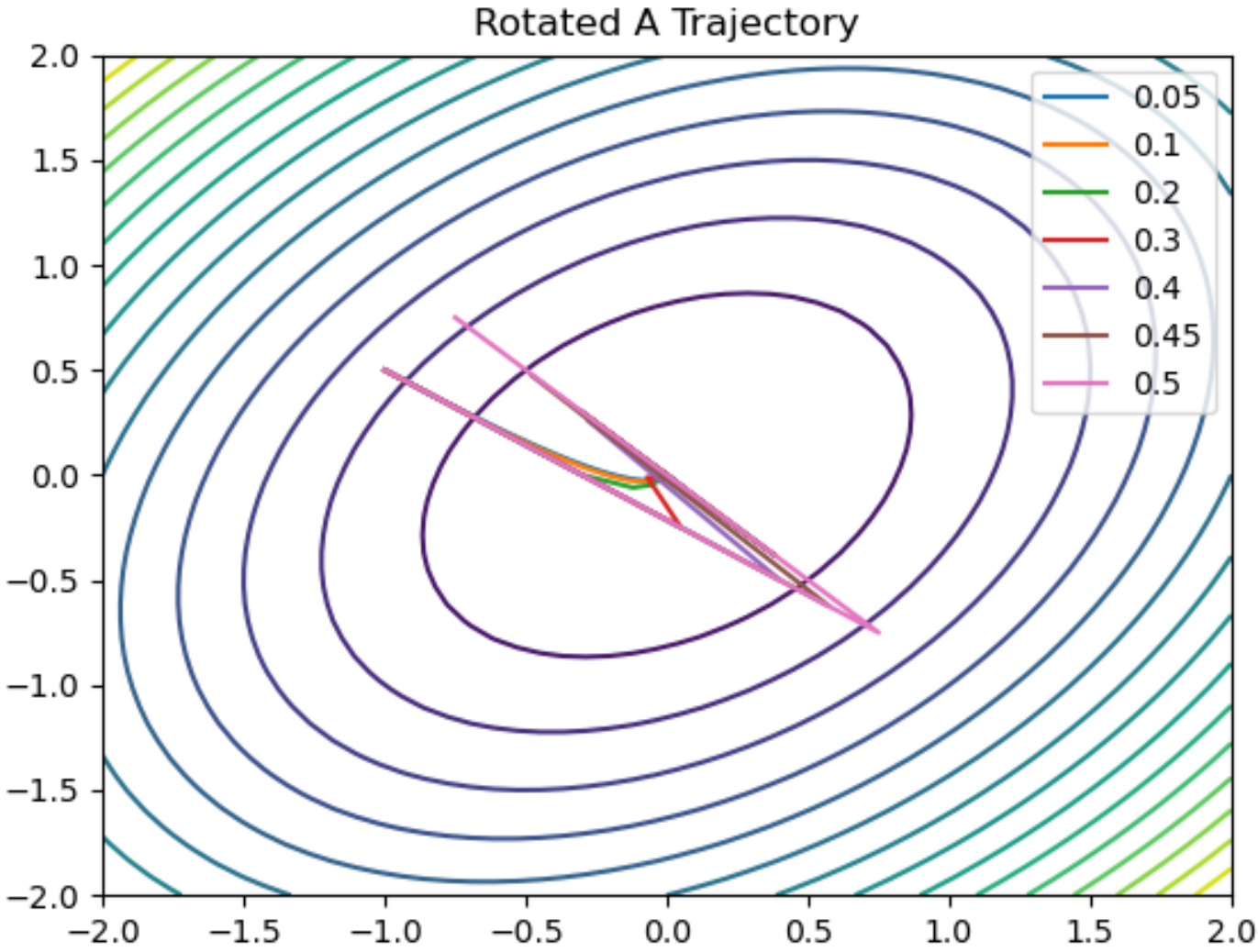
~~$\therefore \frac{1}{2}\alpha \nabla^2 J(\dots) \leq \frac{1}{2}\alpha \beta_{max}$ for all points θ.~~

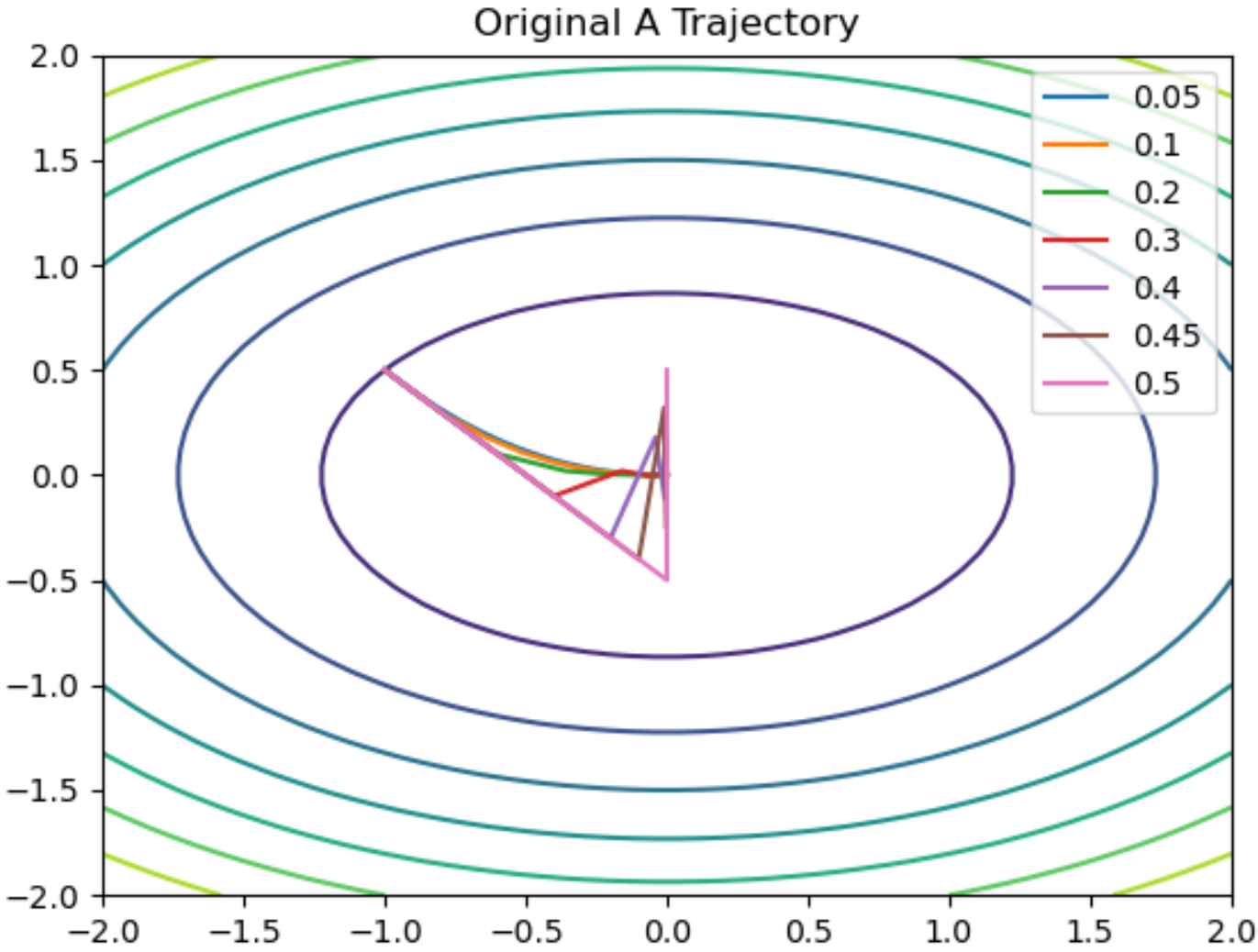
~~$\therefore J(\theta^t) - J(\theta^{t-1}) -$~~

~~When $I - \frac{1}{2}\alpha \nabla^2 J(\theta^{t-1} - \alpha \nabla J\theta^{t-1})$ is positive definite.~~

~~$J\theta^{(t)} < J\theta^{t-1}$~~

$\therefore 1 - \frac{1}{2}\alpha \cdot \beta \cdot z$


Rotated A Trajectory
0.05
0.1
0.2
0.3
0.4
0.45
0.5
2.0
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5
−2.0
−2.0
−1.5
−1.0
−0.5
0.0
0.5
1.0
1.5
2.0


Original A Trajectory
0.05
0.1
0.2
0.3
0.4
0.45
0.5
2.0
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5
−2.0
−2.0
−1.5
−1.0
−0.5
0.0
0.5
1.0
1.5
2.0

for $J\theta^t$ to converge

$$\alpha \nabla J\theta^{t-1T} \left[I - \frac{1}{2}\alpha \nabla^2 J\left(\theta^{t-1} - \alpha \nabla J\theta^{t-1}\right)\right] \nabla J\theta^{t-1} > 0$$

$\because H = \nabla^2 J(\theta) \leq \beta_{max}$ for all θ

$\therefore I - \frac{1}{2}\alpha \beta_{max} > 0$

$\therefore \frac{2}{\beta_{max}} > \alpha > 0$

$\therefore$ if $0 < \alpha < \frac{1}{\beta_{max}}$

$J\theta^t < J\theta^{t-1}$

(e)

2. (i)

$$J(\theta) = (X\theta - y)^T \cdot W (X\theta - y)$$

$$= \sum_{i=1}^{n} \sum_{j=1}^{n} W_{ij} (X\theta - y)_i (X\theta - y)_j$$

when $W = \frac{1}{2}\,\mathrm{diag}(w_1, w_2, w_3 \cdots, w_n)$

$$J(\theta) = \frac{1}{2} \cdot \sum_{i=1}^{n} W_{i,i} (X\theta - y)_i^2$$

$$= \frac{1}{2} \sum_{i=1}^{n} w^{(i)} (\theta^T x^{(i)} - y^{(i)})^2$$

$(X\theta)_i = \theta^T x^{(i)}$ is the ith data point

y_i is the ith observation.

2. (ii) $\nabla J(\theta) = \nabla_\theta (X\theta - y)^T W (X\theta - y)$

$$= \nabla_\theta [(X\theta)^T W X\theta - (X\theta)^T W y - y^T W X\theta + y^T W y]$$

$$= \nabla_\theta [\theta^T X^T W X\theta - \theta^T X^T W y - y^T W X\theta]$$

$$= \nabla_\theta [\theta^T X^T W X\theta - 2 y^T W X\theta]$$

$$= 2 X^T W X\theta - 2 X^T W y$$

$$= 0$$

$$X^T W X\theta = X^T W y$$

$$\theta = (X^T W X)^{-1} X^T W y$$

2.(a) iii:

$$\underset{w}{\operatorname{argmax}}\, P(y|X,w) = \underset{w}{\operatorname{argmax}} \prod_{i=1}^{N} p(y_i|x_i;w)$$

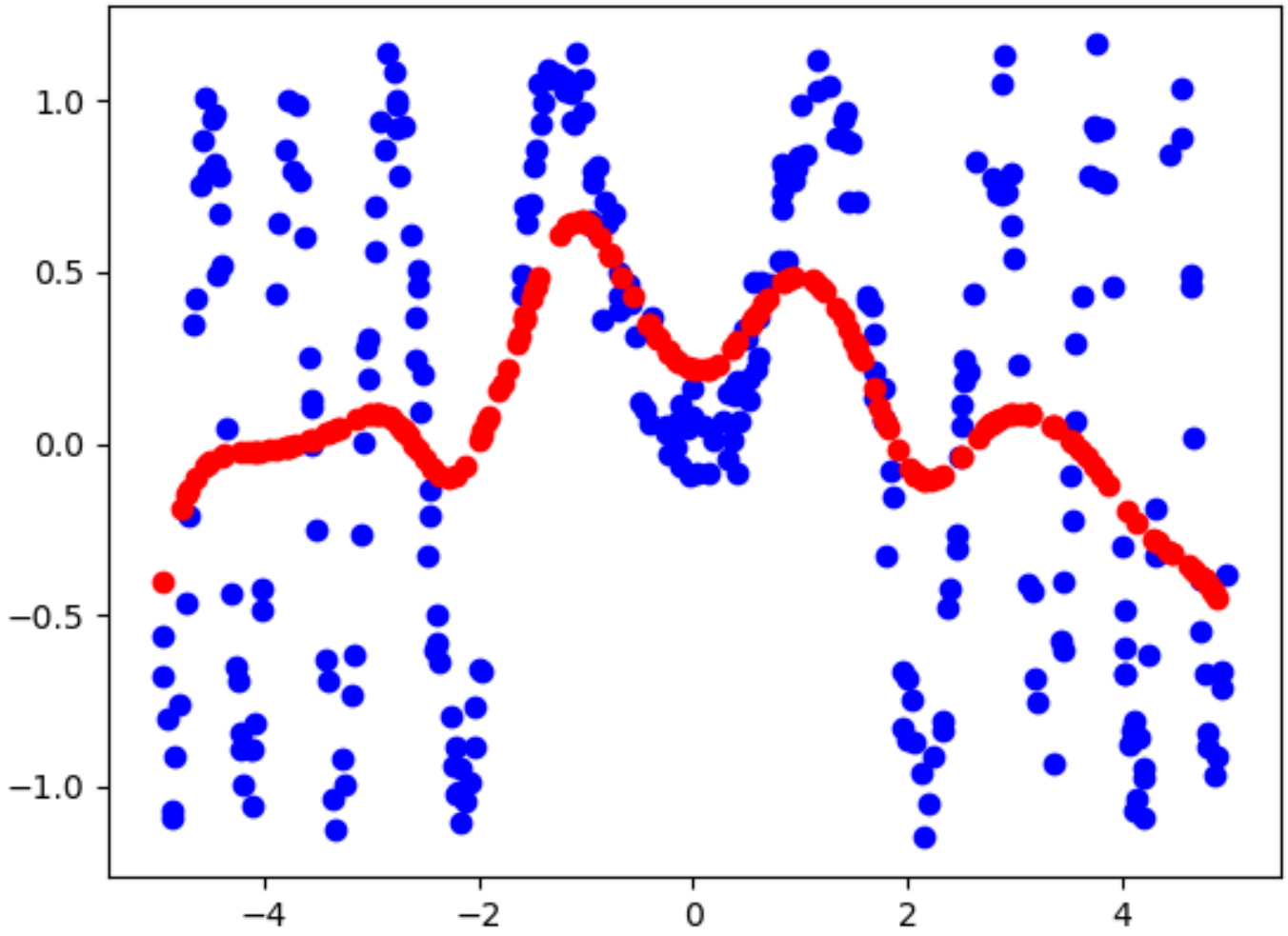
take log:

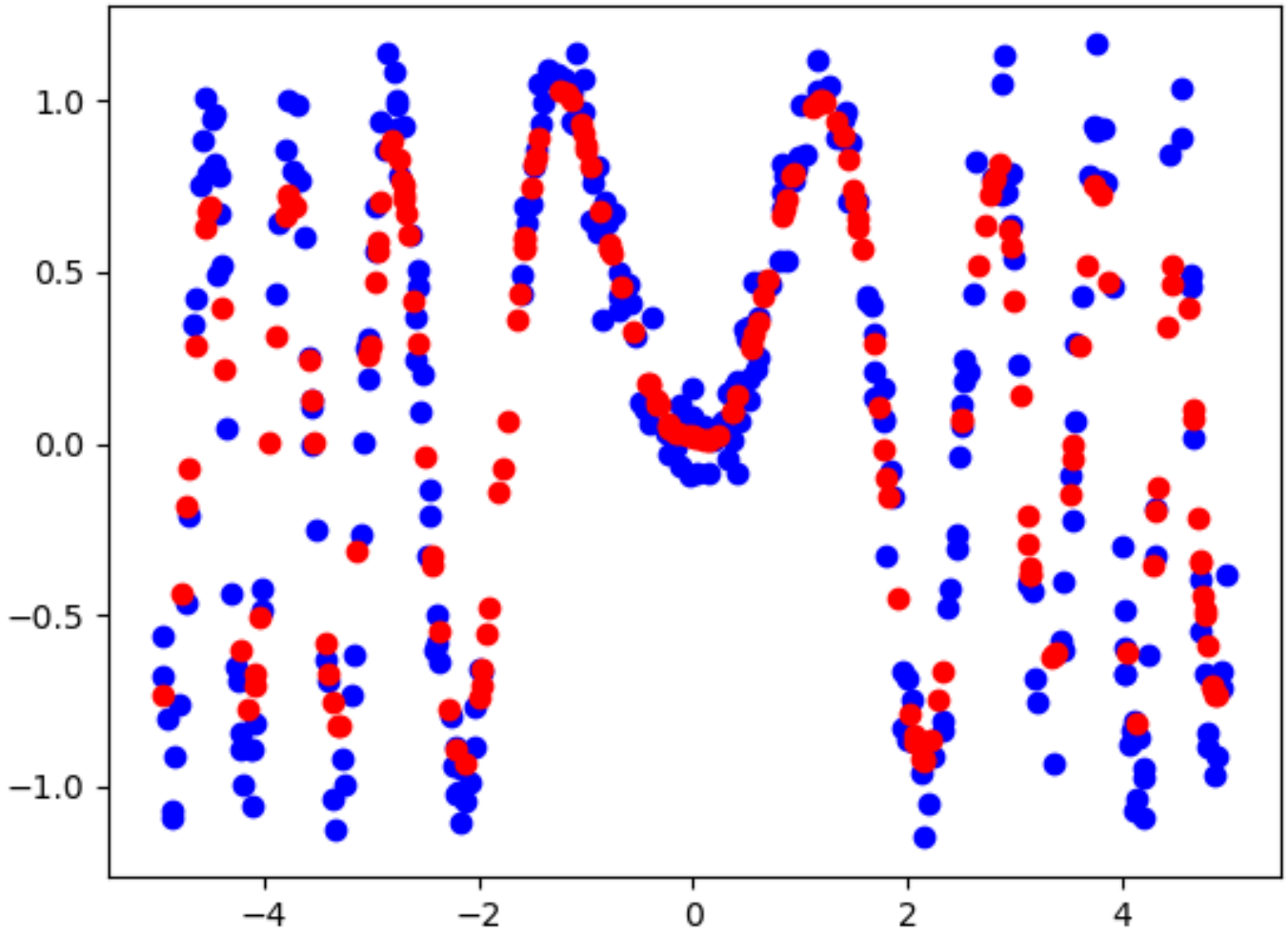
$$\underset{w}{\operatorname{argmax}} \ln\left(\prod_{i=1}^{N} p(y_i|x_i;w)\right)$$

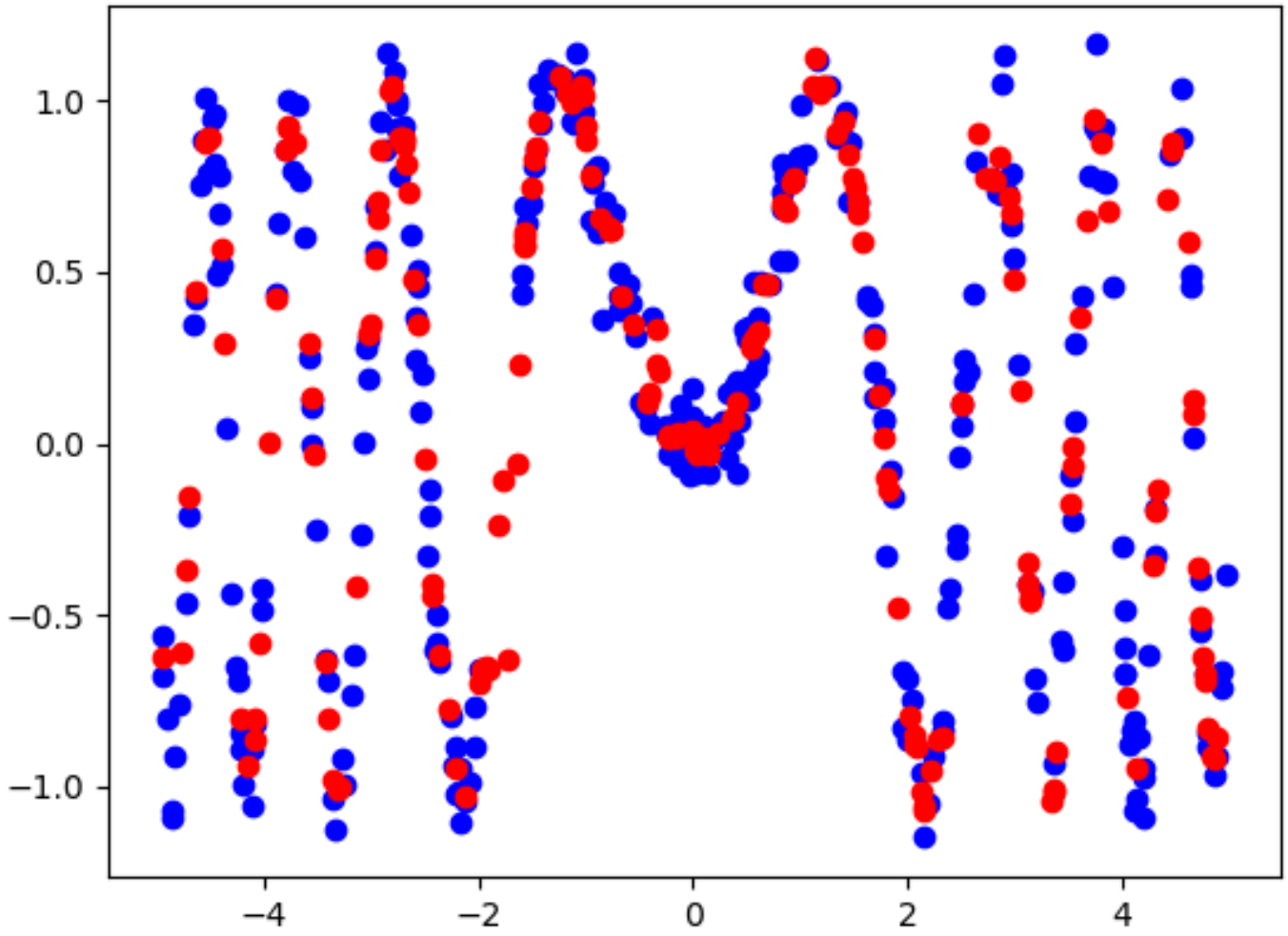
$$= \underset{w}{\operatorname{argmax}} \ln\left\{\left(\frac{1}{\sqrt{2\pi}}\right)^N \cdot \prod_{i=1}^{N} \frac{1}{\sigma^{(i)}} e^{-\frac{(y_i-\theta^T x_i)^2}{2\sigma^{(i)2}}}\right\}$$

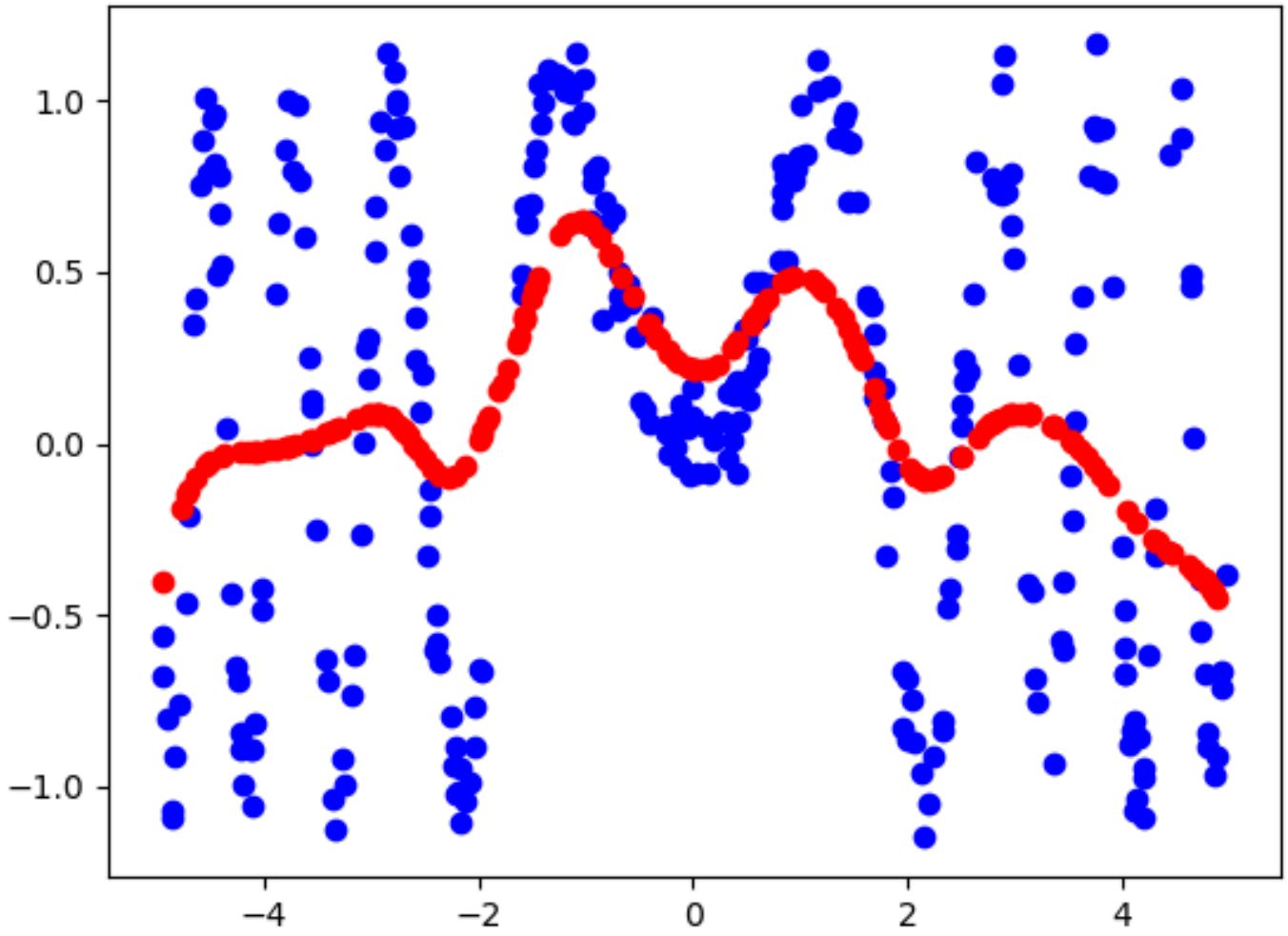
$$= N \ln\frac{1}{\sqrt{2\pi}} + \sum_{i=1}^{N}\left\{\ln\frac{1}{\sigma^2} - \frac{(y_i-\theta^T x_i)^2}{2\sigma^{(i)2}}\right\}$$

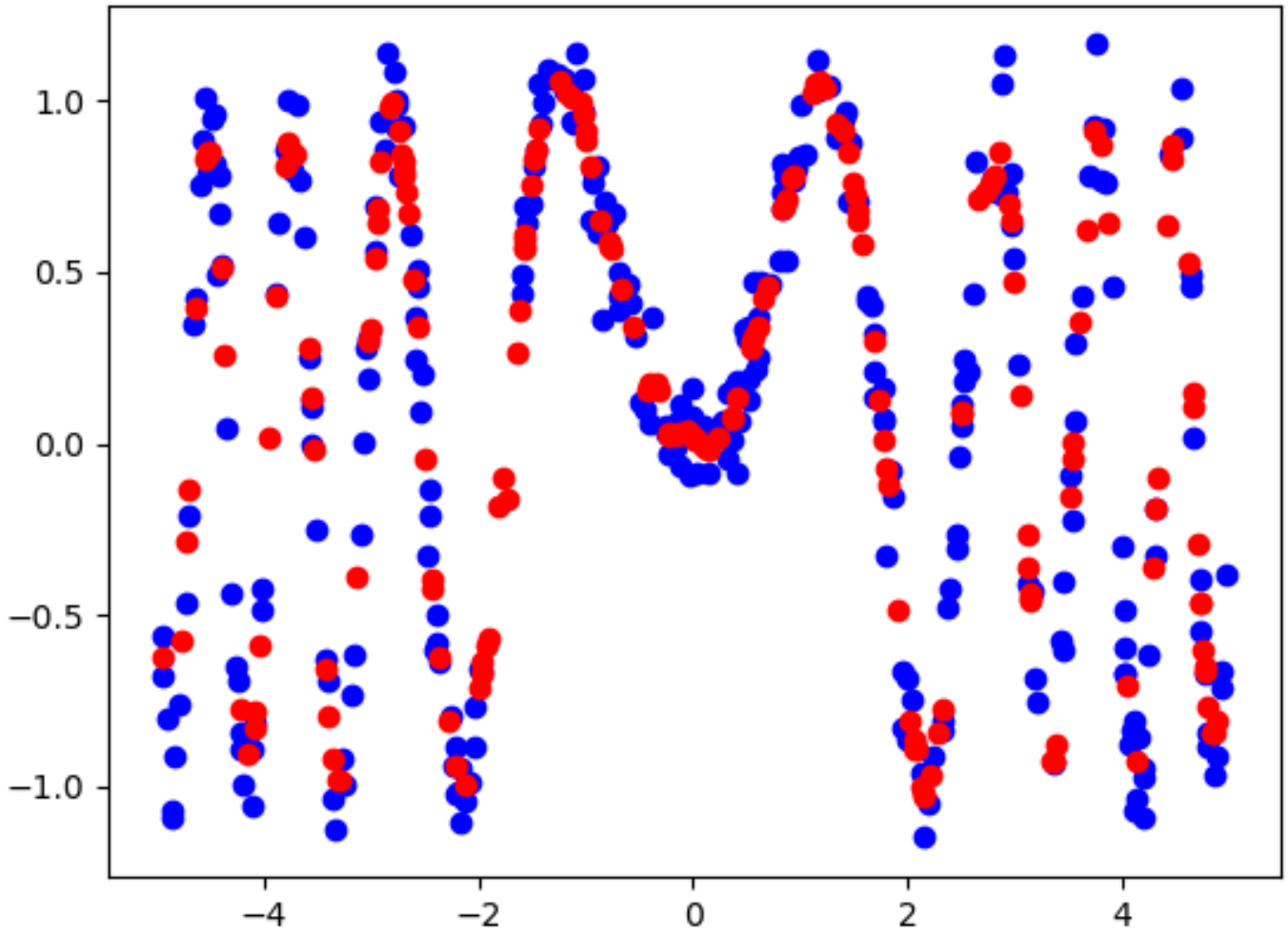
$$= N \ln\frac{1}{\sqrt{2\pi}} - \sum_{i=1}^{N}\left(\ln\sigma^2 + \frac{(y_i-\theta^T x_i)^2}{2\sigma^{(i)2}}\right\}$$

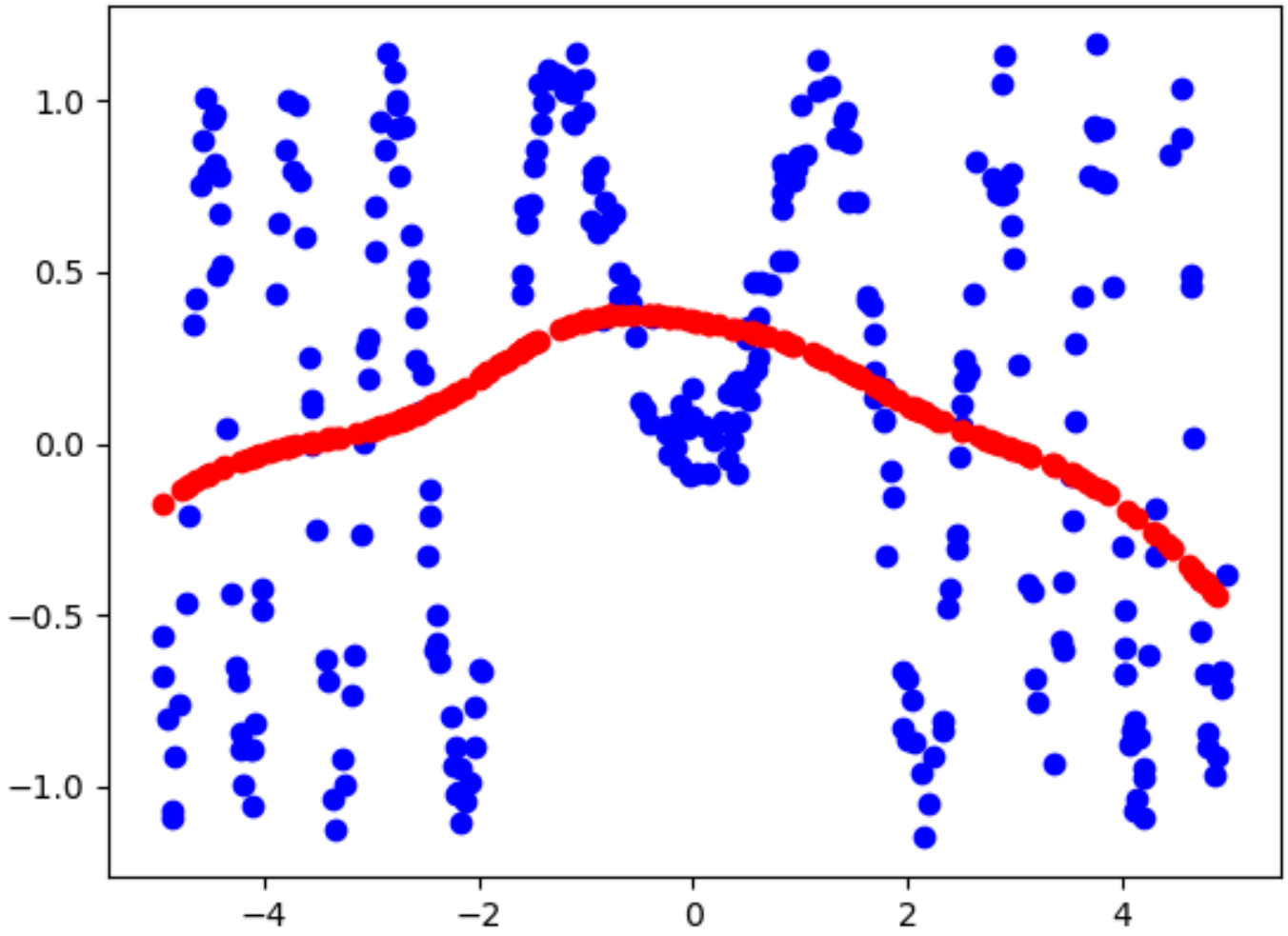

1.0
0.5
0.0
−0.5
−1.0
−4
−2
0
2
4

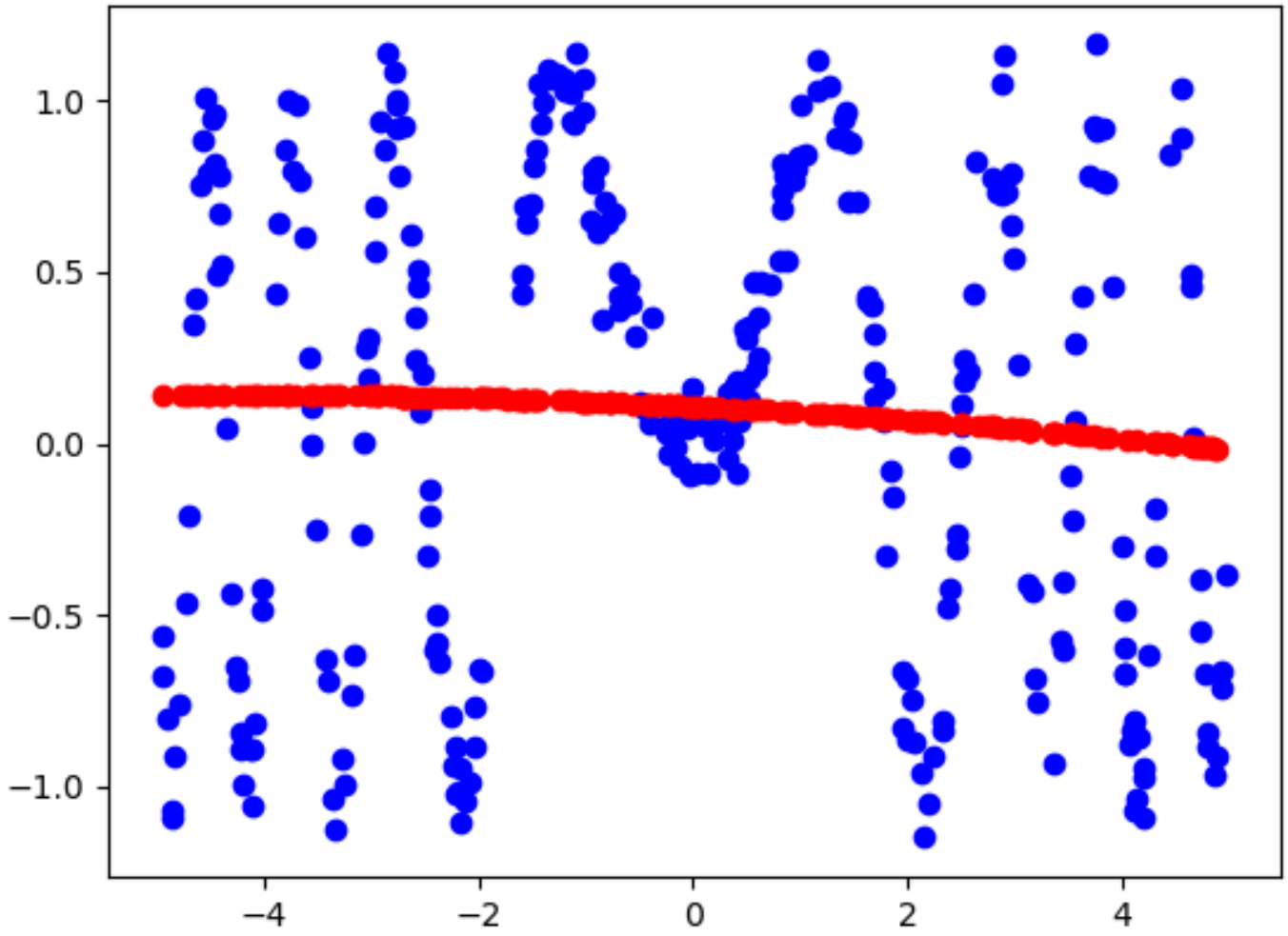

1.0
0.5
0.0
−0.5
−1.0
−4
−2
0
2
4


1.0
0.5
0.0
−0.5
−1.0
−4
−2
0
2
4


1.0
0.5
0.0
−0.5
−1.0
−4
−2
0
2
4


1.0
0.5
0.0
−0.5
−1.0
−4
−2
0
2
4


1.0
0.5
0.0
−0.5
−1.0
−4
−2
0
2
4


1.0
0.5
0.0
−0.5
−1.0
−4
−2
0
2
4

3) (a)

1) $J(\theta) = \frac{1}{2} \sum_{i=1}^{n} (\theta_i \hat{x}_i - y_i)^2$.

2) $\theta^t = \theta^{t-1} - \eta \nabla J(\theta)^{t-1}$

(b) underfitting

mse = 0.33053126821375245

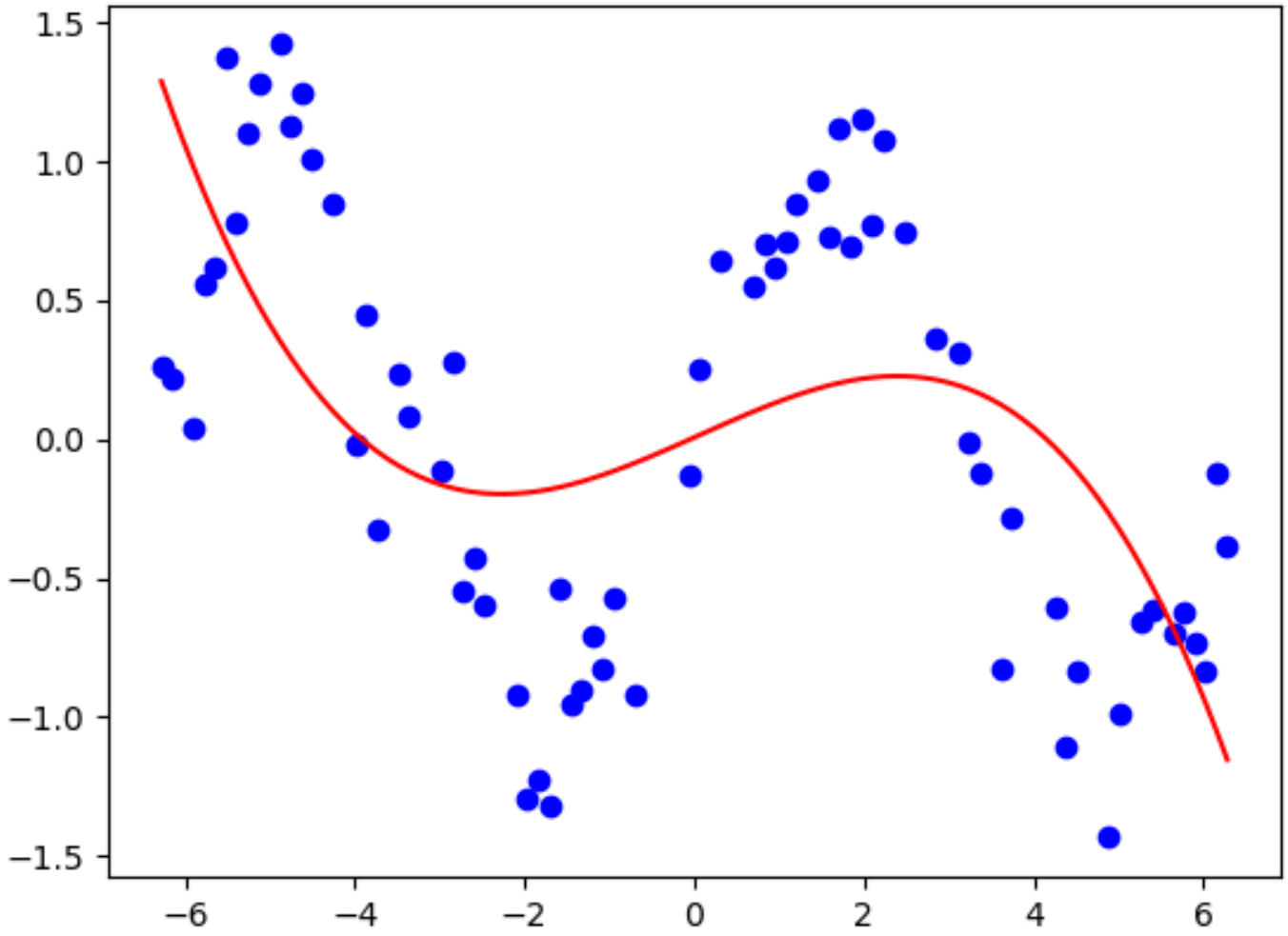
(C) best tau: 0.05

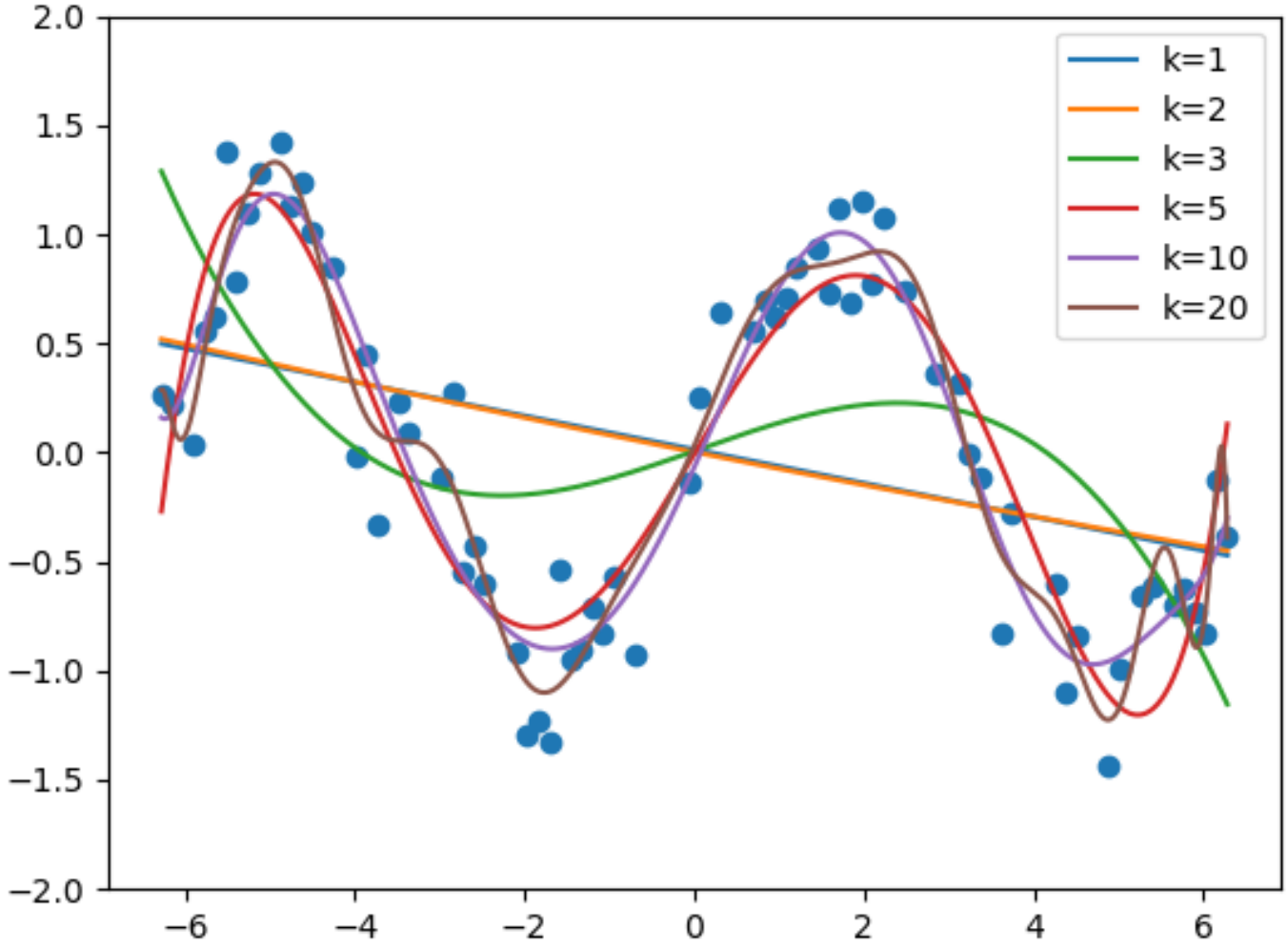
mse: 0.016906174558

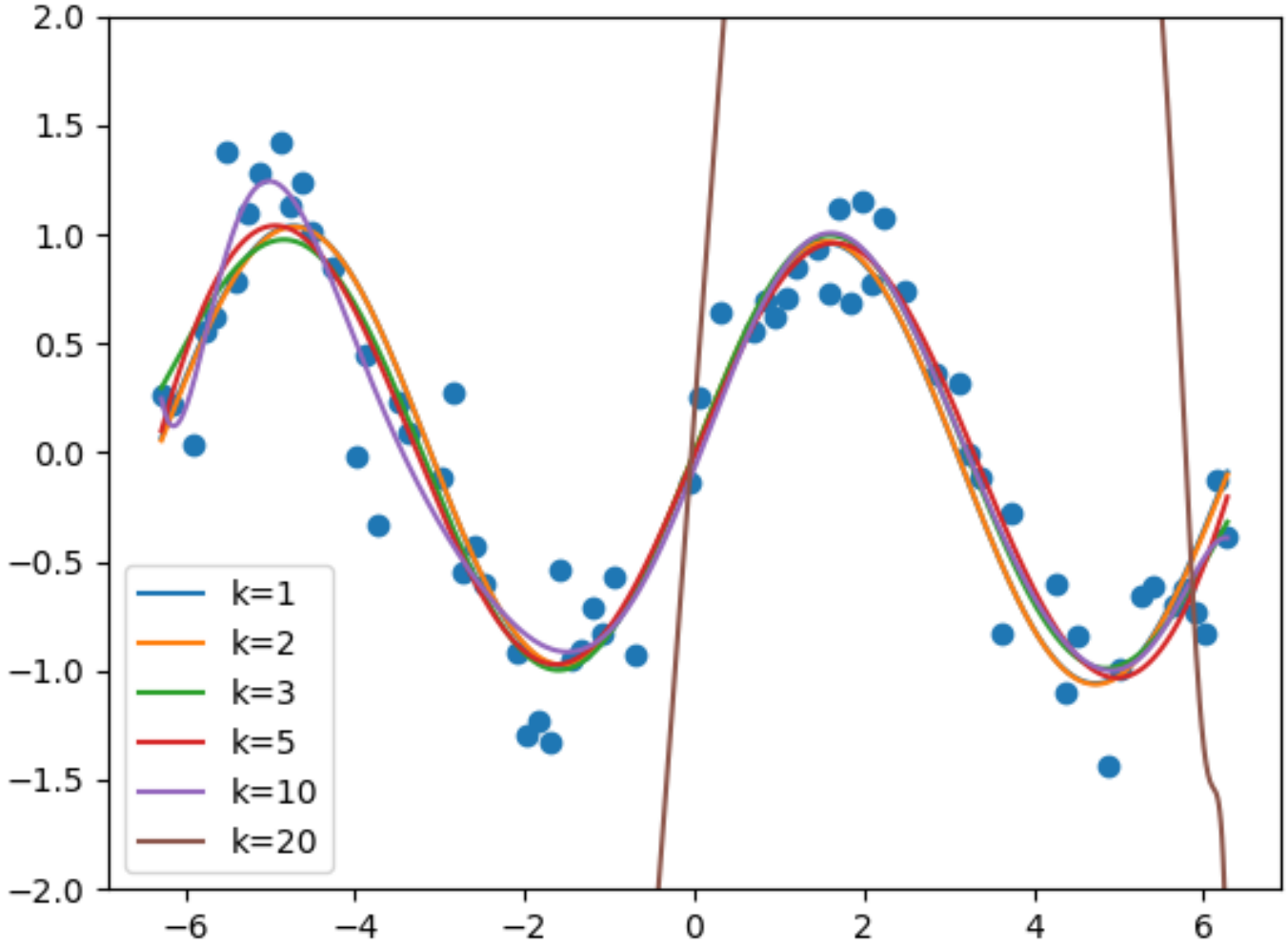
(Q) for small dataset

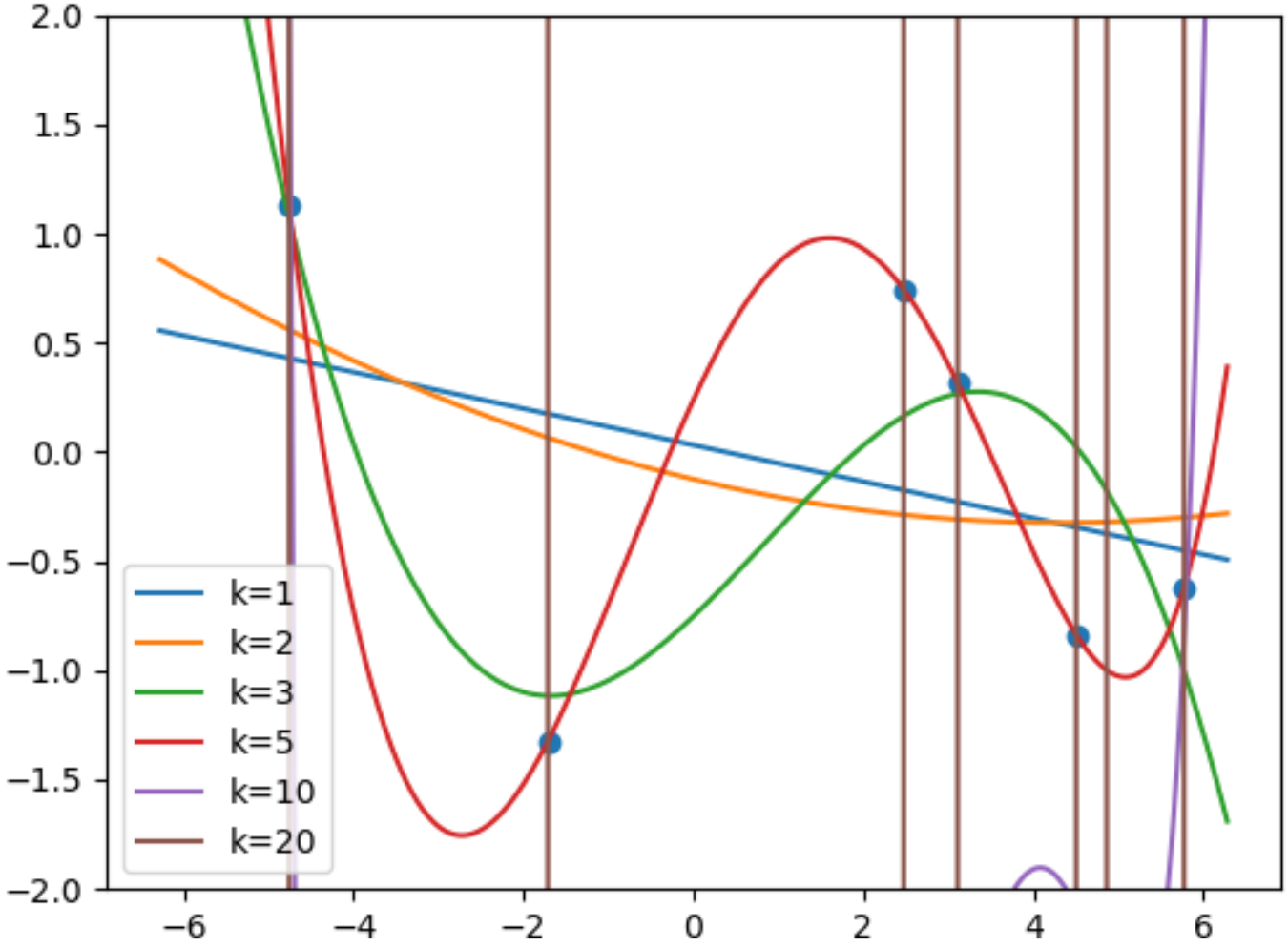
fitting change drastically when k changes

it goes from underfit to overfit as k ↗.


1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5
−6
−4
−2
0
2
4
6


k=1
k=2
k=3
k=5
k=10
k=20
2.0
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5
−2.0
−6
−4
−2
0
2
4
6


2.0
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5
−2.0
−6
−4
−2
0
2
4
6
k=1
k=2
k=3
k=5
k=10
k=20


2.0
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5
−2.0
−6
−4
−2
0
2
4
6
k=1
k=2
k=3
k=5
k=10
k=20

4. (a) $J(\beta) = 0$

$\rightarrow \quad X\beta - y = \vec{0}$

$X\beta = y.$

$X^T(XX^T)^{-1}X\beta = X^T(XX^T)^{-1}y$

~~$X^T(X^T)^{-1}(X^{-1})\cdot X\beta = X^T(XX^T)^{-1}y$~~

$\beta = X^T(XX^T)^{-1}y$

~~further $X(\beta + N(X)) = X\beta + X\cdot N(X)$~~

~~when C^T~~

let C to be the null space of X.

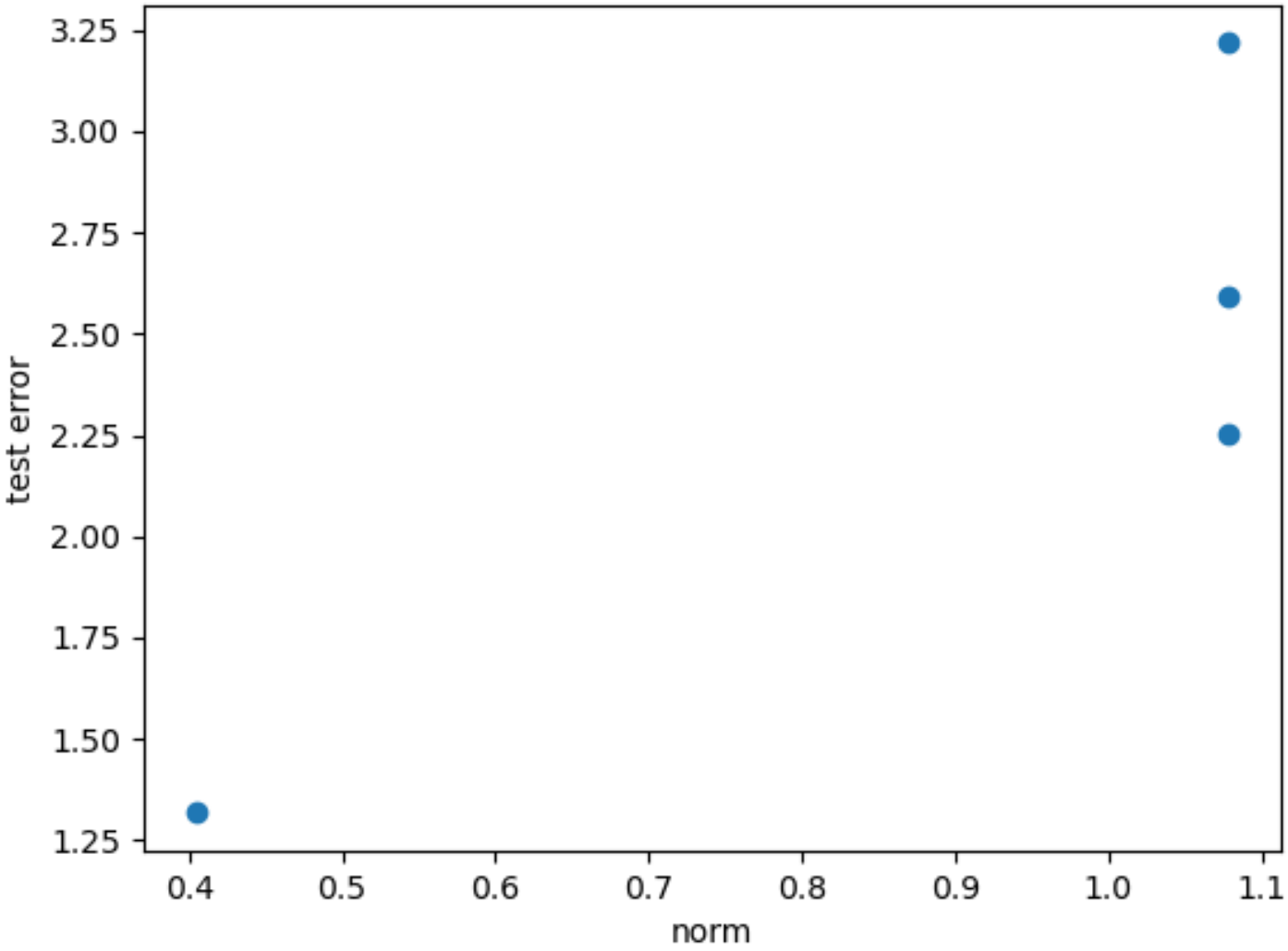
$X\cdot(\beta + C) = X\beta + X\cdot C = X\beta.$

$\therefore \beta = X^T(XX^T)^{-1}y + N(X)$

b)

$\|\beta\|_2 = \|X^T(XX^T)^{-1}y\|_2^2 + \|C\|_2^2$

$= \|P\|_2^2 + \|C\|_2^2$


3.25
3.00
2.75
2.50
2.25
2.00
1.75
1.50
1.25
test error
0.4
0.5
0.6
0.7
0.8
0.9
1.0
1.1
norm

4.(e) $J(\theta, \phi) = \frac{1}{4n} \sum_{i=1}^{n} (x^{(i)T}(\theta^{\circ 2} - \phi^{\circ 2}) - y^{(i)})^2$

by definition

$$\hookrightarrow = \frac{1}{4n} \| X \cdot (\theta^{\circ 2} - \phi^{\circ 2}) - Y \|_2^2$$

$\therefore = \frac{1}{2} J(\beta)$ where $\beta = \theta^{\circ 2} - \phi^{\circ 2}$.

$= \frac{1}{2} \cdot \frac{1}{2n} \| X\beta - y \|_2^2$

base on (a) $J(\beta)$ has ~~many~~ infinite many optimal solution

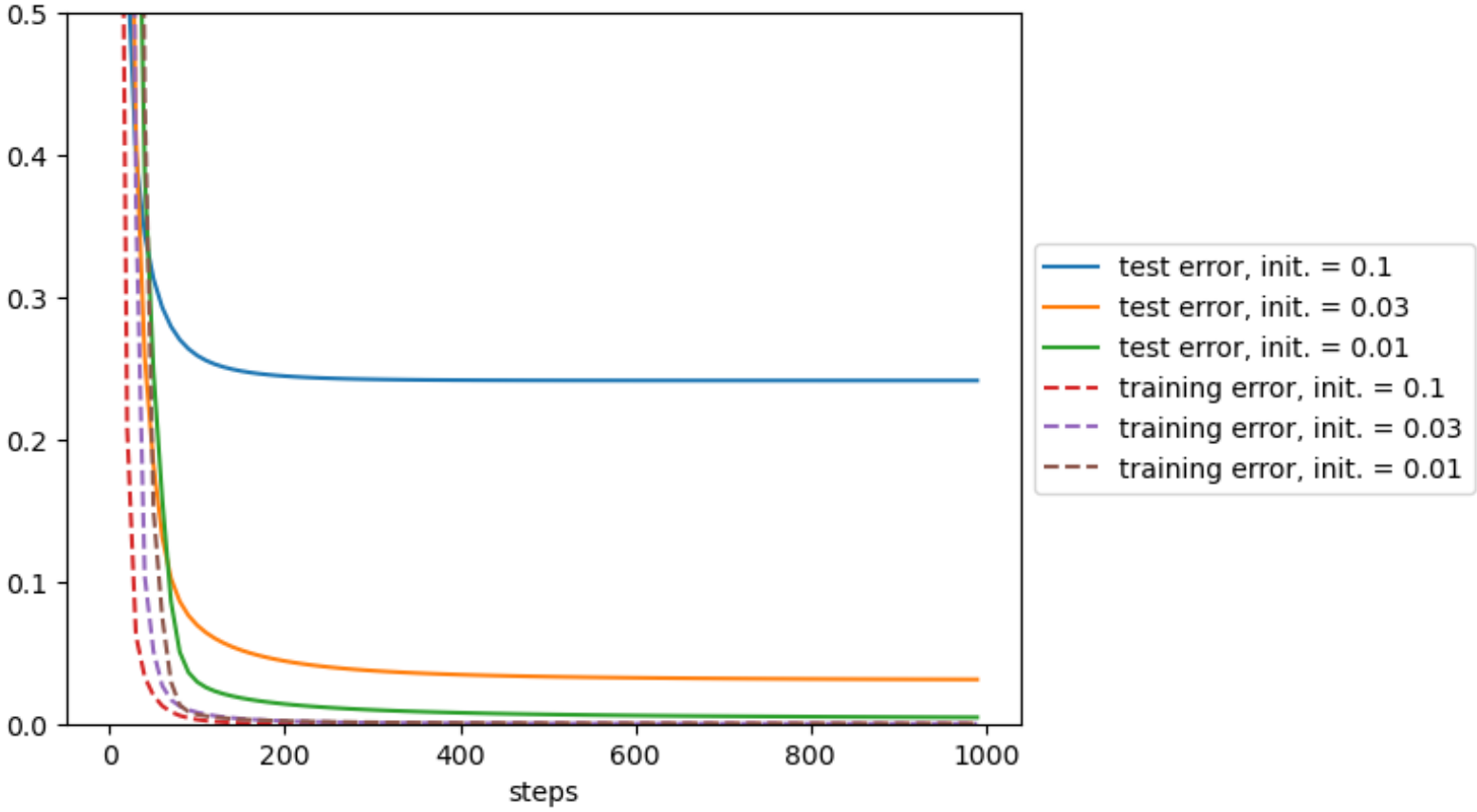
$\therefore \theta^{\circ 2} - \phi^{\circ 2}$ has " " " "

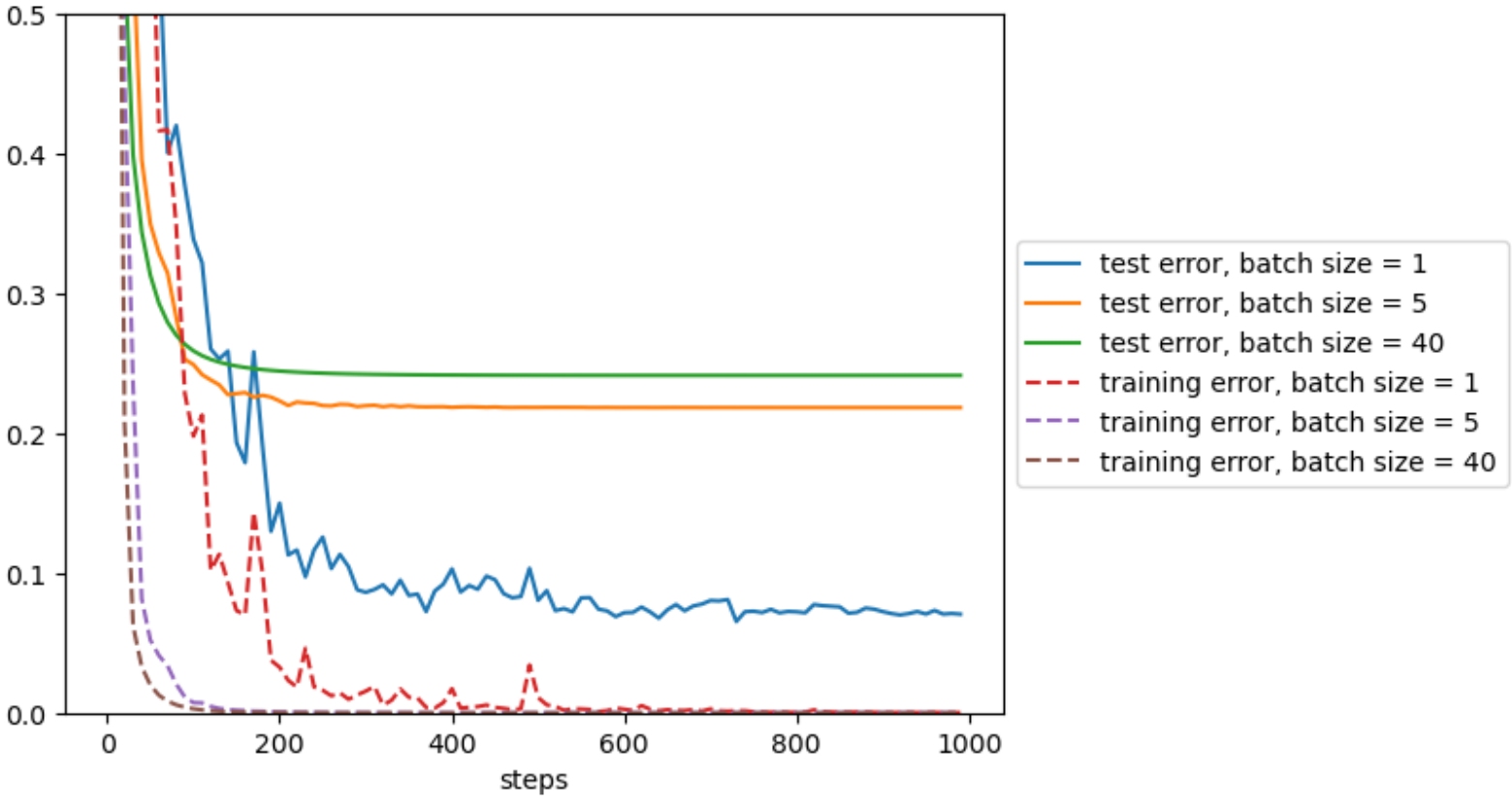
$\therefore \theta^{\circ 2}, \phi^{\circ 2}$ " " " " "

4.(f) all can fit training set very well but
init = 0.01 has the best test error.

(4)(g) for the same init = ~~0~~ 0.1

SGD has a lower test error
thus generalice better.


0.5
0.4
0.3
0.2
0.1
0.0
0
200
400
600
800
1000
steps
test error, init. = 0.1
test error, init. = 0.03
test error, init. = 0.01
training error, init. = 0.1
training error, init. = 0.03
training error, init. = 0.01


0.5
0.4
0.3
0.2
0.1
0.0
0
200
400
600
800
1000
steps
test error, batch size = 1
test error, batch size = 5
test error, batch size = 40
training error, batch size = 1
training error, batch size = 5
training error, batch size = 40

5. (a) $J_\lambda(\beta) = \frac{1}{2}\|X\beta - y\|_2^2 + \frac{\lambda}{2}\|\beta\|_2^2$

$= \frac{1}{2}\left((X\beta - y)^T (X\beta - y) + \lambda \beta^T \beta \right)$

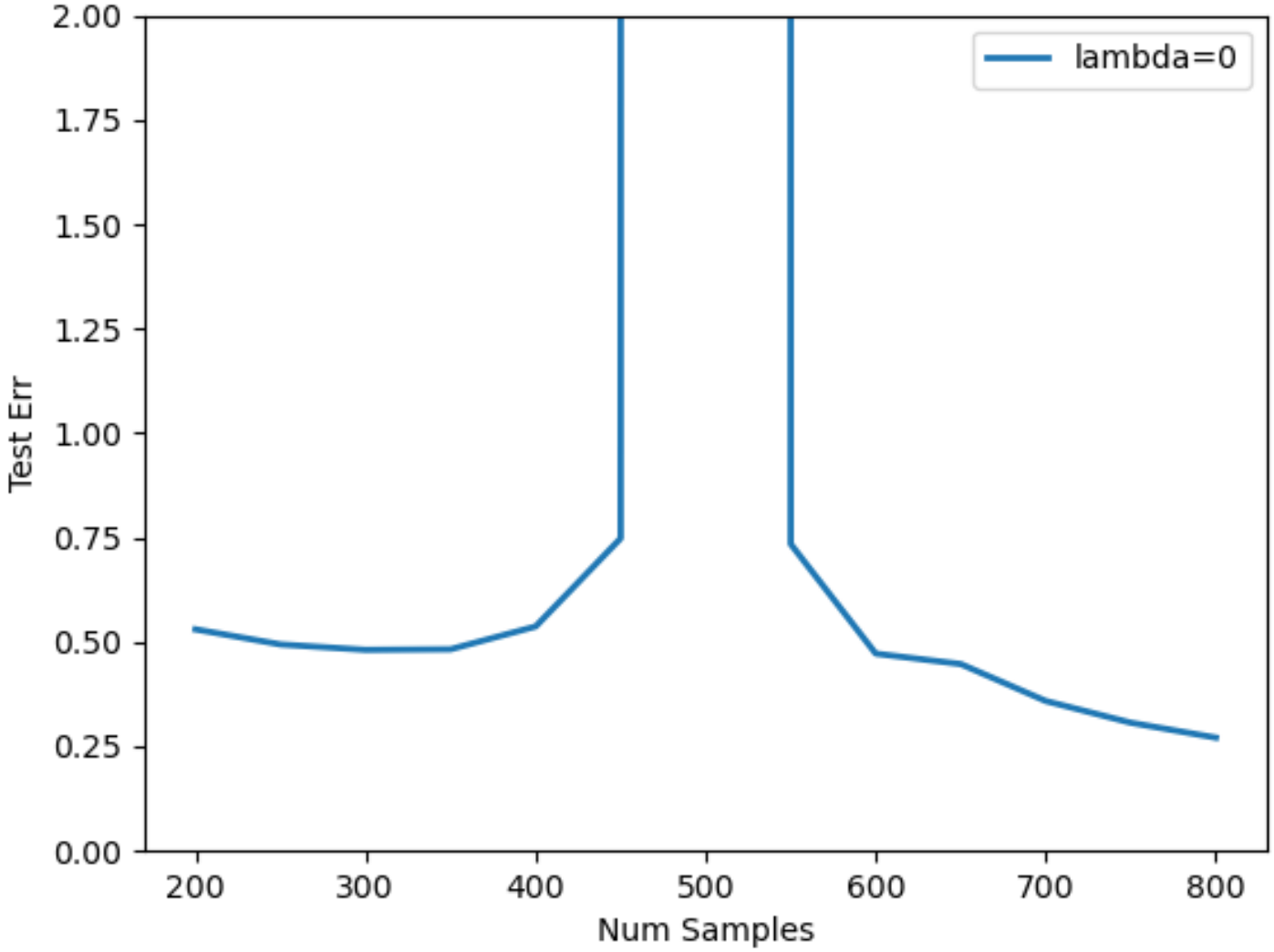
$= \frac{1}{2}\left(\beta^T X^T X\beta - \beta^T X^T y - y^T X\beta + y^T y + \lambda \beta^T \beta \right)$

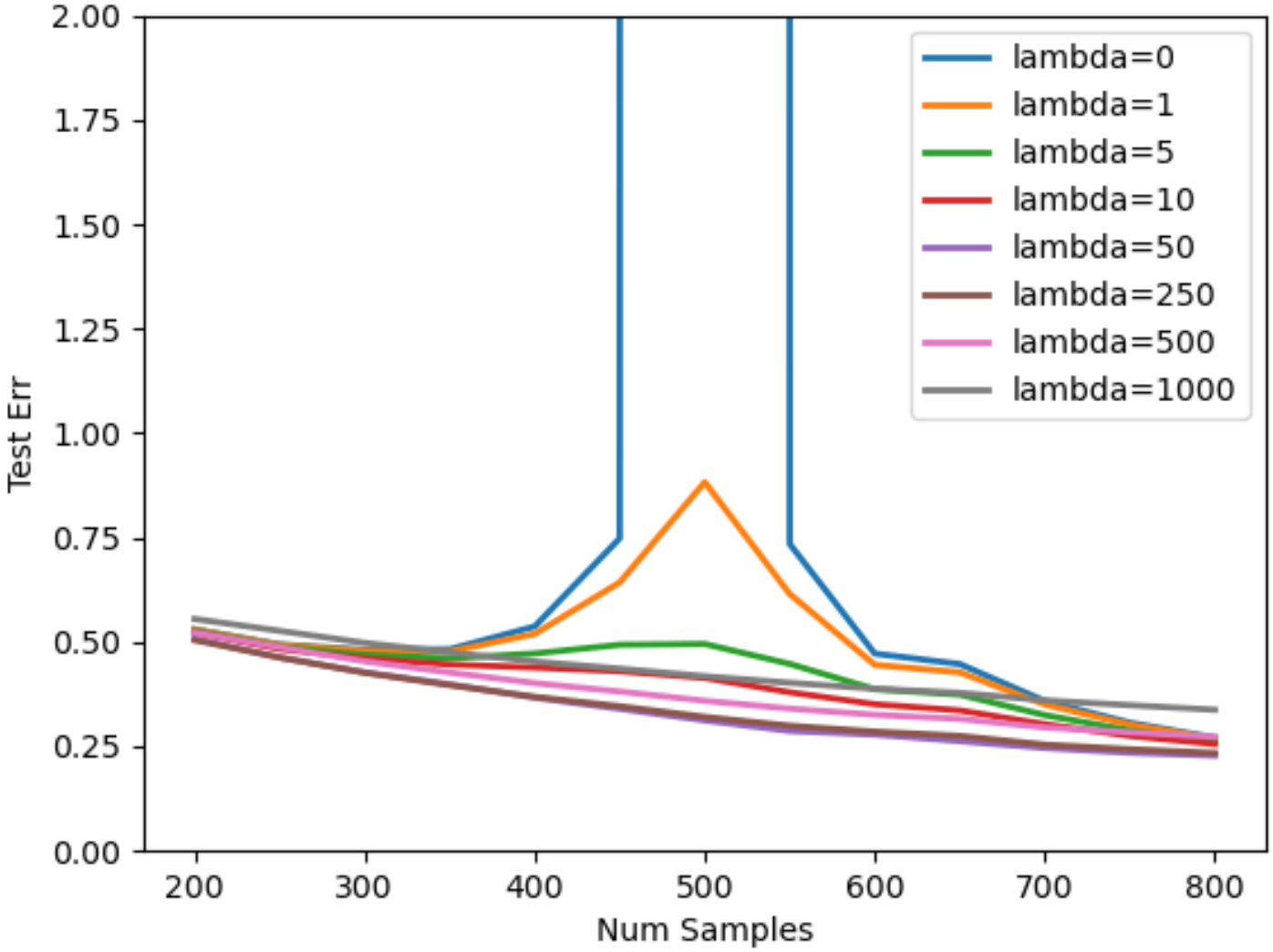
$= \frac{1}{2}\left(\beta^T X^T X\beta - 2y^T X\beta + \lambda \beta^T \beta \right)$

$\nabla J_\lambda(\beta) = \frac{1}{2}\left(2(X^T X)\beta - 2(y^T X)^T + 2\lambda I\beta \right) = 0$

$(X^T X + \lambda I)\beta = X^T y$

$\beta = (X^T X + \lambda I)^{-1} X^T y$


lambda=0
2.00
1.75
1.50
1.25
1.00
0.75
0.50
0.25
0.00
Test Err
200
300
400
500
600
700
800
Num Samples


2.00
1.75
1.50
1.25
1.00
0.75
0.50
0.25
0.00
Test Err
200
300
400
500
600
700
800
Num Samples
lambda=0
lambda=1
lambda=5
lambda=10
lambda=50
lambda=250
lambda=500
lambda=1000