1)

a) $L(x) = (2-x_1)^2 + 3(4x_2 - x_1^2)^2$

$$\nabla L(x) = \begin{bmatrix} -2\cdot(2-x_1) + 6(4x_2 - x_1^2)\cdot(-2x_1) \\ 6(4x_2 - x_1^2)\cdot 4 \end{bmatrix}$$

$$\nabla^2 L(x) = \begin{bmatrix} 2 - 48x_2 + 36x_1^2 & -48x_1 \\ -48x_1 & 96 \end{bmatrix}$$

when $\nabla L(x) = 0$, $\begin{cases} 4x_2 = x_1^2 \\ x_1 = 2 \end{cases} \Rightarrow \begin{cases} x_1 = 2 \\ x_2 = 1 \end{cases} \Rightarrow \nabla^2 L(x)\Big|_{x=(2,1)} = \begin{bmatrix} 98 & -96 \\ -96 & 96 \end{bmatrix}$

the eigenvalues $\lambda_1 \sim 193.005$ $\lambda_2 \approx 0.994$

Thus the Hessian is p.d at $(2,1)$, so $(2,1)$ is a Strict local minima

b) $L(u) = (u-3)(u+6)(u-2)$

$\nabla L(u) = (u-3)(u+6) + (u-3)(u-2) + (u+6)(u-2) = 3u^2 + 2u - 24$

when $\nabla L(u) = 0$ $u = \frac{-1 \pm \sqrt{73}}{3}$

$\nabla^2 L(u) = 6u + 2$ $\nabla^2 L(u)\Big|_{u = \frac{-1+\sqrt{73}}{3}} = 2\sqrt{73}$ $\nabla^2 L(u)\Big|_{u = \frac{-1-\sqrt{73}}{3}} = -2\sqrt{73}$

So when $u = \frac{-1+\sqrt{73}}{3}$, it is a strict local minima, when $u = \frac{-1-\sqrt{73}}{3}$, it is a strict local maxima

c) $L(u) = (u_1^2 + 2u_1 - 6)(3u_2^2 + 2u_2 + 1)$

$$\nabla L(u) = \begin{bmatrix} (2u_1+2)(3u_2^2+2u_2+1) \\ (u_1^2+2u_1-6)(6u_2+2) \end{bmatrix}$$ when $\nabla L(u) = 0$ $\begin{bmatrix} u_1 = -1 \\ u_2 = -\frac{1}{3} \end{bmatrix}$

$$\nabla^2 L(u) = \begin{bmatrix} 2(3u_2^2+2u_2+1) & (2u_1+2)(6u_2+2) \\ (2u_1+2)(6u_2+2) & 6(u_1^2+2u_1-6) \end{bmatrix}$$, at $(-1, -\frac{1}{3})$ $\nabla^2 L(u) = \begin{bmatrix} \frac{4}{3} & 0 \\ 0 & -7 \end{bmatrix}$

eigenvalues are $\{\frac{4}{3}, -7\}$, since the eigenvalues are both positive and negative at stationary point $(-1, -\frac{1}{3})$ the point is a saddle point

d) $L(x) = 3x_1^4 - 28x_1^3 + 84x_1^2 - 96x_1 + 64 + 27x_2^2$

$\nabla L(x) = \begin{bmatrix} 12x_1^3 - 84x_1^2 + 168x_1 - 96 \\ 54x_2 \end{bmatrix}$, when $\nabla L(x) = 0$, $x_1 = 1$ or 2 or 4, $x_2 = 0$

$\nabla^2 L(x) = \begin{bmatrix} 36x_1^2 - 168x_1 + 168 & 0 \\ 0 & 54 \end{bmatrix}$

at $(1,0)$ $\nabla^2 L(x) = \begin{bmatrix} 36 & 0 \\ 0 & 54 \end{bmatrix}$ is p.d., so $(1,0)$ is a strict local minima.

at $(2,0)$ $\nabla^2 L(x) = \begin{bmatrix} -24 & 0 \\ 0 & 54 \end{bmatrix}$, has both negative and positive eigenvalue so $(2,0)$ is a saddle point.

at $(4,0)$ $\nabla^2 L(x) = \begin{bmatrix} 72 & 0 \\ 0 & 54 \end{bmatrix}$, is p.d. so $(4,0)$ is a strict local minima.

$L(1,0) = 27 > L(4,0) = 0$, $(4,0)$ is a global minima.

2) necessary condition: $\theta \in \mathbb{R}^n$ $F(\theta) \in \mathbb{R}^m$ $F'(\theta) \in \mathbb{R}^{m\times n}$ $F''(\theta) \in \mathbb{R}^{m\times n\times n}$

$\nabla L(\theta) \in \mathbb{R}^n$ $\nabla^2 L(\theta) \in \mathbb{R}^{n\times n}$

$L(\theta) = \frac{1}{2}\|F(\theta) - Y\|^2 = \frac{1}{2}(F(\theta) - Y)^T(F(\theta) - Y) = \frac{1}{2}F(\theta)^T F(\theta) - Y^T F(\theta) + \frac{1}{2}Y^T Y$

$\nabla L(\theta) = 0 \Rightarrow F'(\theta)^T(F(\theta) - Y) = 0 \Rightarrow F(\theta) = Y$ or $F'(\theta) = 0$ ①

$\nabla^2 L(\theta) \geq 0 \Rightarrow (F(\theta) - Y)^T F''(\theta) + F'(\theta)^T \cdot F'(\theta) \geq 0$. ②

when $F(\theta) = Y$, ① ② both satisfies,

but $F(\theta) = Y$ not always holds, since there are noises in Y

then, $F'(\theta)^T = 0$, $(F(\theta) - Y)^T F''(\theta)$ has to be p.s.d.

So the necessary condition is:

$F(\theta)=Y$ or $F'(\theta)=0$ and $(F(\theta)-Y)^T \cdot F''(\theta)$ is p.s.d

Sufficient condition:

$\nabla^2 L(\theta) > 0$, and $\nabla L(\theta)=0$

$\Rightarrow$ then $F(\theta)=Y$ and $F'(\theta)^T F'(\theta)$ is p.d

or $F'(\theta)=0$ and $(F(\theta)-Y)^T F''(\theta)$ is p.d

numerical method

stochastic gradient decent,

select a minibatch, and do gradient decent based on that minibatch

α: stepsize. Start with 1e-3 or other value.

```
for t in range(numsteps):
    minibatch = sample_data(data, batchsize)
    grad = compute_gradient(loss_fn, minibatch)
    if norm(grad) < 1e-8: break
    x = x - α·gradient
    α = α·β        β ∈ (0,1)
```

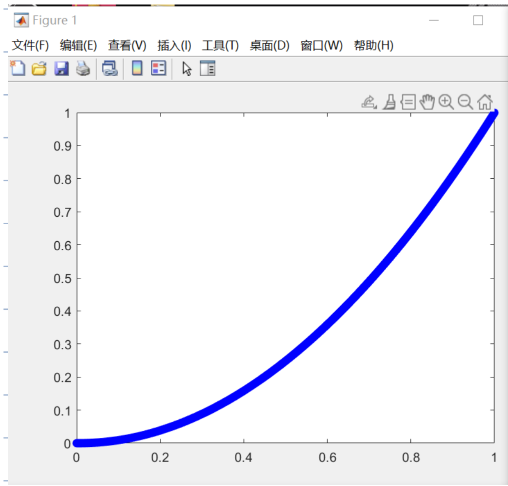
selecting batchsize data from sample data

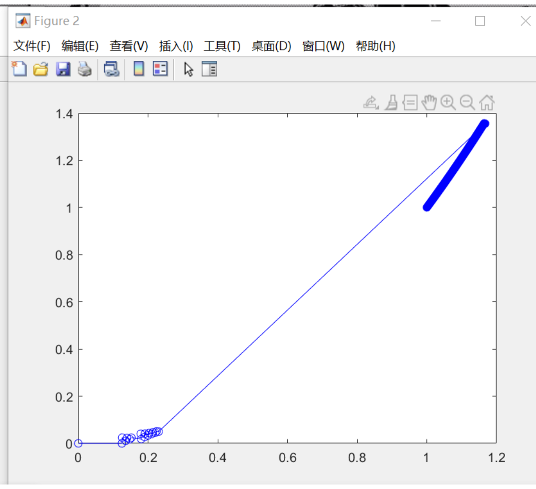
necessary assumptions:

the model needs to be a convex and $\nabla^2 L(\theta)$ at θ^* needs to be p.d.

and we need to have a large amount of relatively precise data (low noise) that the data could represent the real shape or distribution of model. then, we can find a unique minimizer

3. a

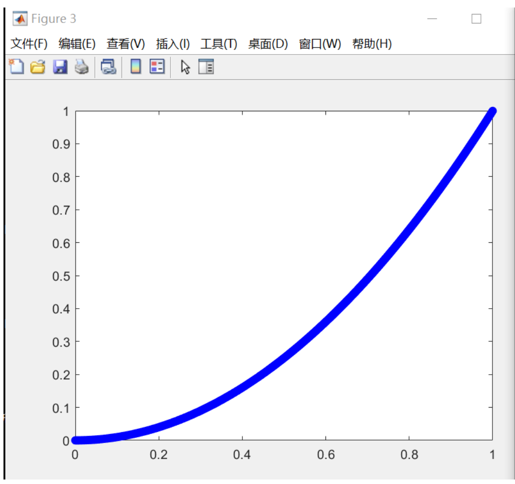


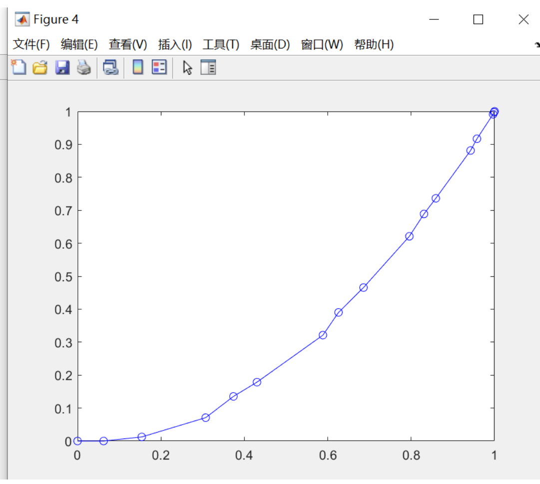



```
############################################## a
gradient decent with fixed stepsize:
x=[1          1] f=1.2509e-16
Steps=428070
gradient decent with variable stepsize:
x=[1          1] f=9.3833e-17
Steps=41987
############################################## b
newton method with fixed stepsize:
x=[1          1] f=2.7061e-17
Steps=190930
newton method with variable stepsize:
x=[1          1] f=3.7948e-18
Steps=17
############################################## c
gradient decent with variable stepsize:
x=[-0.63613   2.6262e-09] f=-0.38396
Steps=48
newton method with variable stepsize:
x=[0.22048      0.29503] f=0.21464
Steps=1000000
>>
```

Variable stepsize converges in fewer steps

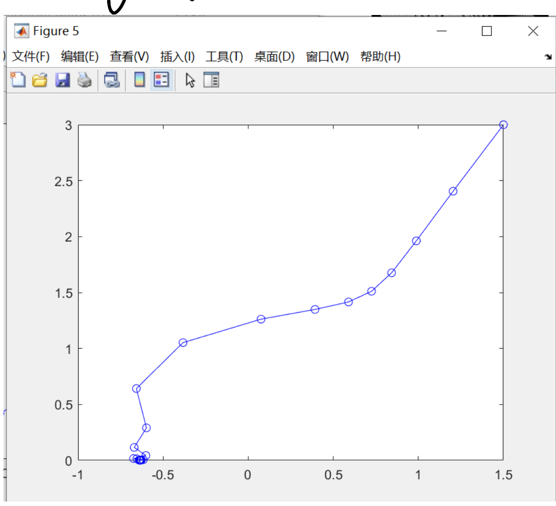
b)

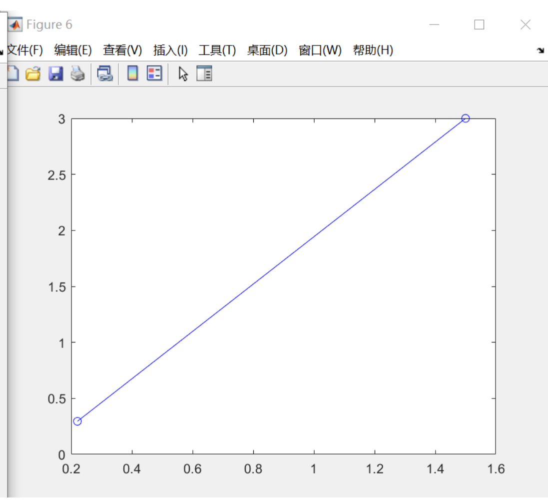




newton's methods converges in fewer steps

c) gradient decent newton's method.





newton's method got stucked, but gradient decent get close to the solution