Q1

① $f(x+p) = f(x) + \nabla f(x)^T p + O(\|p\|^2)$

From Taylor's theorem, the k^{th} order Taylor polynomial as:

$$f(b) = f(a) + f'(a)(b-a) + \frac{f''a}{2!}(b-a)^2 + \cdots + \frac{f^{(k)}(a)}{k!}(b-a)^k + h_k(b)(b-a)^k, \lim_{b\to a} h_k(b) = 0$$

Therefore we can get:

The best linear approximation for $f(b)$ is:

$f(b) = f(a) + f'(a)(b-a) + R$, R represents the error term between approximation to real $f(b)$ value.

For a multivariable version,

$f(b) = f(b_1, b_2, b_3 \cdots b_n)$

then the linear approximation (first order) of $f(b)$ is:

$f(b) = f(a) + \nabla f(a)^T (b-a) + R$

where $\nabla f(a)$ represents a $1 \times n$ horizontal vector of partial derivatives of $f(a_1, a_2 \cdots a_n)$

subsitute b with $x+p$, a with x, we get:

$f(x+p) = f(x) + \nabla f(x)^T p + R$

Since $R = \frac{f''(x)}{2!}(p)^2 + \cdots \frac{f^k(x)}{k!} p^k$ and $\lim_{x+p\to x} R = 0$,

$R = O(\|p\|^2)$ (dominated by quadratic term),

$f(x+p) = f(x) + \nabla f(x)^T p + O(\|p\|^2)$

② $f(x+p) = f(x) + \nabla f(x)^T p + \frac{1}{2} p^T \nabla^2 f(x) p + O(\|p\|^3)$

Similar to ①, with second order Taylor polynomial.

or

$f(b) = f(a) + f'(a)(b-a) + \frac{f''a}{2}(b-a)^2 + R$, R represents the error term between approximation to real $f(b)$ value

For a multivariable version, $f(b) = f(b_1, b_2, b_3 \cdots b_n)$

From above, Since $f(b)$ is scalar, first derivative $f'(a)$ is a $1 \times n$ matrix denotes by $\nabla f(a)$, for the second derivative of $f(a)$, we can take the matrix of partial derivative of the vector $\nabla f(a)$ which writen as $\nabla^2 f(a)$, is an $n \times n$ matrix

Therefore, we can rewrite a single variable quadratic expression $\frac{1}{2}(b-a) f''(a)(b-a)$ into multivariable version $\frac{1}{2}(b-a)^T \nabla^2 f(a)(b-a)$

Combine the linear expression we derived from part ①, subsitute b with $x+p$, a with x, we get:

$$f(x+p) = f(x) + \nabla f(x)^T p + \frac{1}{2} p^T \nabla^2 f(x) p + R$$

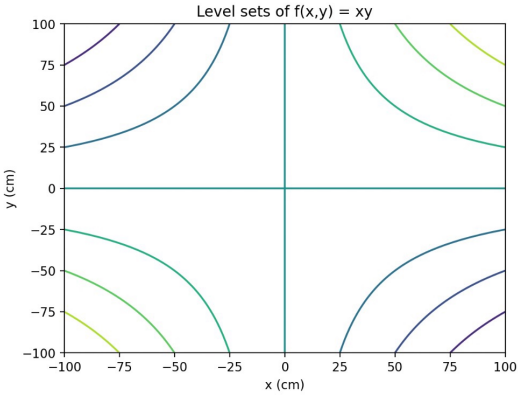
Since $R = \frac{f'''(x)}{3!}(p)^3 + \cdots \frac{f^k(x)}{k!} p^k$ and $\lim_{x+p \to x} R = 0$,

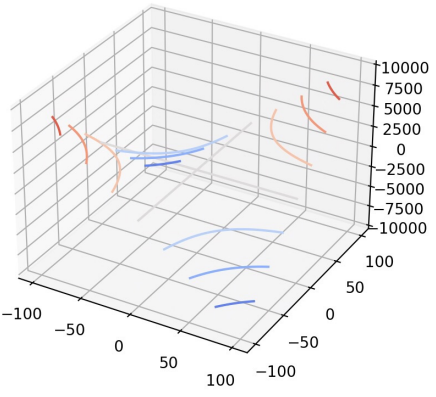
$R = O(\|p\|^3)$ (dominated by cubic term),

$$f(x+p) = f(x) + \nabla f(x)^T p + \frac{1}{2} p^T \nabla^2 f(x) p + O(\|p\|^3)$$

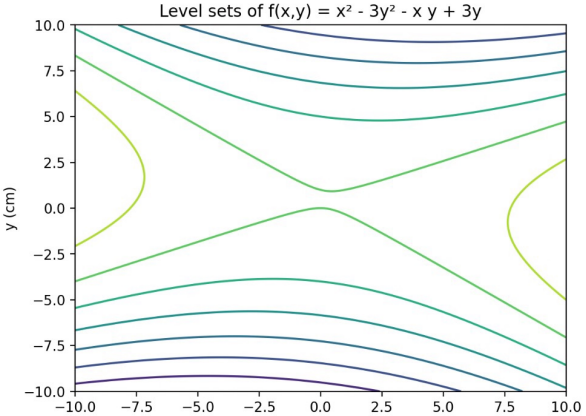
Q2

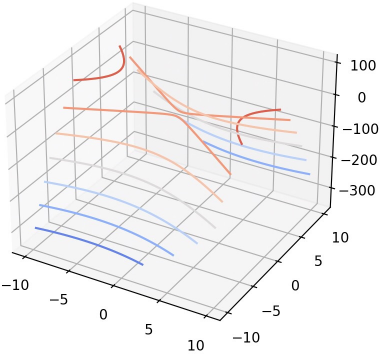
1



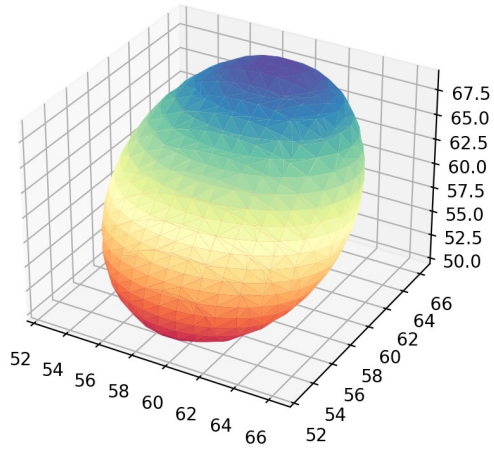


2





3



Q3 Convex function $f(x)$, by definition that on an interval $[a,b]$ if for any two points x_1 and x_2 in $[a,b]$ and with λ $0<\lambda<1$:

$$f[\lambda x_1+(1-\lambda)x_2) \leq \lambda f(x_1)+(1-\lambda)f(x_2) \quad (1)$$

Let x be a local minima at convex function $f(x)$, which means $f(x^*) \geq f(x)$ for every x^* in $B(x,\delta)$ $[\delta>0]$.

Suppose there is a point y, $f(y)<f(x)$

$$f(\lambda x+(1-\lambda)y) \leq \lambda f(x)+(1-\lambda)f(y)$$

for $0<\lambda<1$, if we pick a large enough λ, $\lambda x+(1-\lambda)y$ will fall in the inteval of $(x-\delta, x+\delta)$, we have:

$$\lambda f(x)+(1-\lambda)f(y) \geq f(\lambda x+(1-\lambda)y) \geq f(x)$$

$$\lambda f(x)+(1-\lambda)f(y) \geq f(x)$$

$f(y) \geq f(x)$, so $f(x)$ is global minima

Q4

① Given $f(x)$ being positive, since x^* maximize $f(x)$, $f(x^*) \geq f(x) > 0$, x being any input for $f(x)$.

since $-\log x$ monotonically decrease in $(0, +\infty)$, $-\log(f(x^*)) \leq f(x)$, which means x^* minimize $-\log(f(x))$.

if a, b on $f(x)$, b being any point other than a, and $-\log(a) \leq -\log(b)$, given $f(x)$ is positive, a must be the global maxima on $f(x)$.

Given a is achieved by an input x^* to $f(x)$, therefore any input x^* that minimize $-\log(f(x))$ also maximize $f(x)$

② $L(\mu,\Sigma) = \prod_{i=1}^{k} N(d_i|\theta_i)$

$= \prod_{i=1}^{k} (2\pi)^{-\frac{n}{2}} |\Sigma|^{-\frac{1}{2}} \exp\left(-\frac{1}{2}(x^i-\mu)^T \Sigma^{-1}(x^i-\mu)\right)$

$= 2\pi^{-\frac{kn}{2}} |\Sigma|^{-\frac{k}{2}} \exp\left(-\frac{1}{2}\sum_{i=1}^{k}(x^i-\mu)^T\Sigma^{-1}(x^i-\mu)\right)$

$-\ln(L(\mu,\Sigma)) = -\ln 2\pi^{-\frac{kn}{2}} - \ln|\Sigma|^{-\frac{k}{2}} - \ln\exp\left(-\frac{1}{2}\sum_{i=1}^{k}x^i-\mu)^T\Sigma^{-1}(x^i-\mu)\right)$

$= \frac{k}{2} + \frac{1}{2}\sum_{i=1}^{k}(x^i-\mu)^T\Sigma^{-1}(x^i-\mu) + C$

$-\frac{\partial \ln L}{\partial \mu} = \frac{\partial}{\partial \mu}\left(\frac{1}{2}\sum_{i=1}^{k}(x^i-\mu)^T\Sigma^{-1}(x^i-\mu)\right)$

$(x^i-\mu)^T\Sigma^{-1}(x-\mu) = (x^T-\mu^T)\Sigma^{-1}(x-\mu)$

$= x^T\Sigma^{-1}x - x^T\Sigma^{-1}\mu - \mu^T\Sigma^{-1}x + \mu^T\Sigma^{-1}\mu$

$2\mu^T\Sigma^{-1}x$

$\sum_{i=1}^{k}(x^i-\mu)^T\Sigma^{-1}(x^i-\mu) = \sum_{i=1}^{k}x^{iT}\Sigma^{-1}x^i - 2\sum_{i=1}^{k}\mu^T\Sigma^{-1}x^i + k\mu^T\Sigma^{-1}\mu$

$$-\frac{\partial \ln L}{\partial \mu} = \frac{\partial}{\partial \mu}\frac{1}{2}\left(\sum_{i=1}^{k} x^{iT}\Sigma^{-1}x^{i} - 2\sum_{i=1}^{k}\mu^{T}\Sigma^{-1}x^{i} + k\mu^{T}\Sigma^{-1}\mu\right)$$

$$= \underbrace{\frac{\partial}{\partial \mu}\left(-\sum_{i=1}^{k} x^{T}\Sigma^{-1}\mu\right)}_{①} + \underbrace{\frac{1}{2}\frac{\partial}{\partial \mu} k\mu^{T}\Sigma^{-1}\mu}_{②}$$

$$① = -\sum_{i=1}^{k}\Sigma^{-1T}x^{i} \quad \left[f(x) = A^{T}x,\ \frac{df}{dx} = A\right]$$

$$\Sigma^{-1T} = \Sigma^{-1} \Rightarrow \quad ① = -\sum_{i=1}^{k}\Sigma^{-1}x^{i} = -\Sigma^{-1}\sum_{i=1}^{k}x^{i}$$

$$② = k\Sigma^{-1}\mu \quad \left[f(x) = x^{T}Ax,\ \frac{df}{dx} = Ax + A^{T}x = 2Ax \ (A = A^{T})\right]$$

$$\boxed{-\frac{\partial \ln L}{\partial \mu} = -\Sigma^{-1}\sum_{i=1}^{k}x^{i} + k\Sigma^{-1}\mu}$$

$$-\frac{\partial \ln L}{\partial \Sigma} = \underbrace{\frac{k}{2}\frac{\partial}{\partial \Sigma}\ln|\Sigma|}_{①} + \underbrace{\frac{1}{2}\frac{\partial}{\partial \Sigma}\sum_{i=1}^{k}(x-\mu)^T\Sigma^{-1}(x^i-\mu)}_{②}$$

$$① = \frac{k}{2}\Sigma^{-1}$$

$$② = \frac{1}{2}\sum_{i=1}^{k}\left(-\Sigma^{-1}(x^i-\mu)(x^i-\mu)^T\Sigma^{-1}\right)$$

$$= \frac{k}{2}\Sigma^{-1} + \frac{1}{2}\sum_{i=1}^{k}\left(-\Sigma^{-1}(x^i-\mu)(x^i-\mu)^T\Sigma^{-1}\right)$$

$$-\frac{\partial \ln L}{\partial \Sigma} = \frac{1}{2}\Sigma^{-1}\left(k\Sigma - \sum_{i=1}^{k}(x^i-\mu)(x^i-\mu)^T\right)\Sigma^{-1}$$

③

$$\nabla \text{-} \ln L(\mu) = \frac{\partial \ln L}{\partial \mu} = -\Sigma^{-1} \sum_{i=1}^{k} x^i + k\Sigma^{-1}\mu$$

$$\nabla \text{-} \ln L(\Sigma) = \frac{\partial \ln L}{\partial \Sigma} = \frac{1}{2}\Sigma^{-1}\left(k\Sigma - \sum_{i=1}^{k} (x^i-\mu)(x^i-\mu)^T\right)\Sigma^{-1}$$

$$-\nabla \ln L(\mu) = 0$$

$$\Rightarrow \frac{\Sigma^{-1}\sum_{i=1}^{k} x^i}{k\Sigma^{-1}\mu} = 0 \qquad \mu = \frac{1}{k}\sum_{i=1}^{k} x = \bar{x}$$

$$-\nabla \ln L(\Sigma) = 0$$

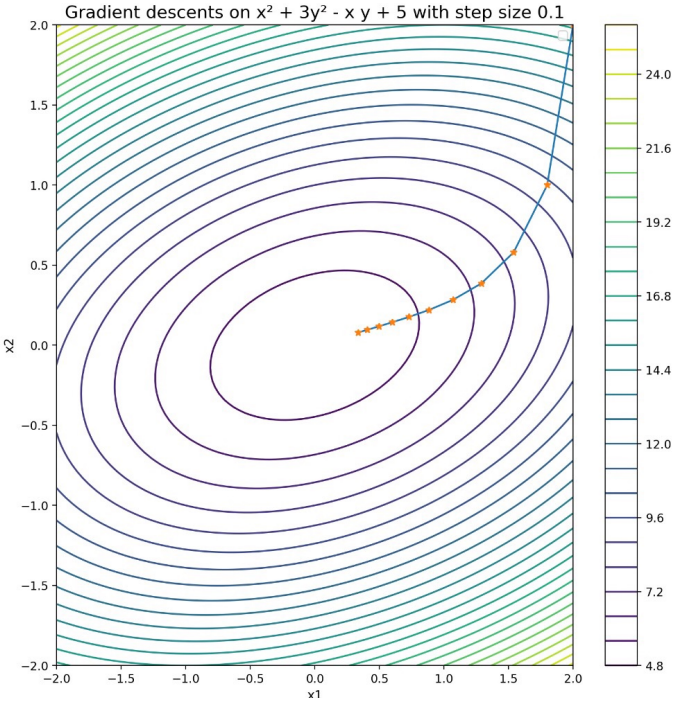
$$\Rightarrow \left(k\Sigma - \sum_{i=1}^{k} (x-\mu)(x-\mu)^T\right) = 0$$

$$\Sigma = \frac{1}{k}\sum_{i=1}^{k} (x-\mu)(x-\mu)^T$$

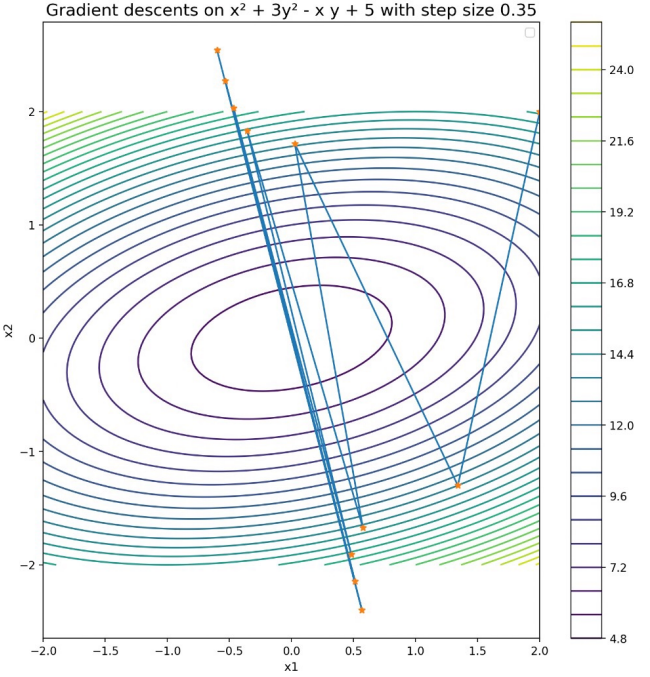
(4) Since $L(\theta)$ is convex function and θ is the point that locally maximized $L(\theta)$ by doing gradient descent, so θ globally maximizes $L(\theta)$

Q5

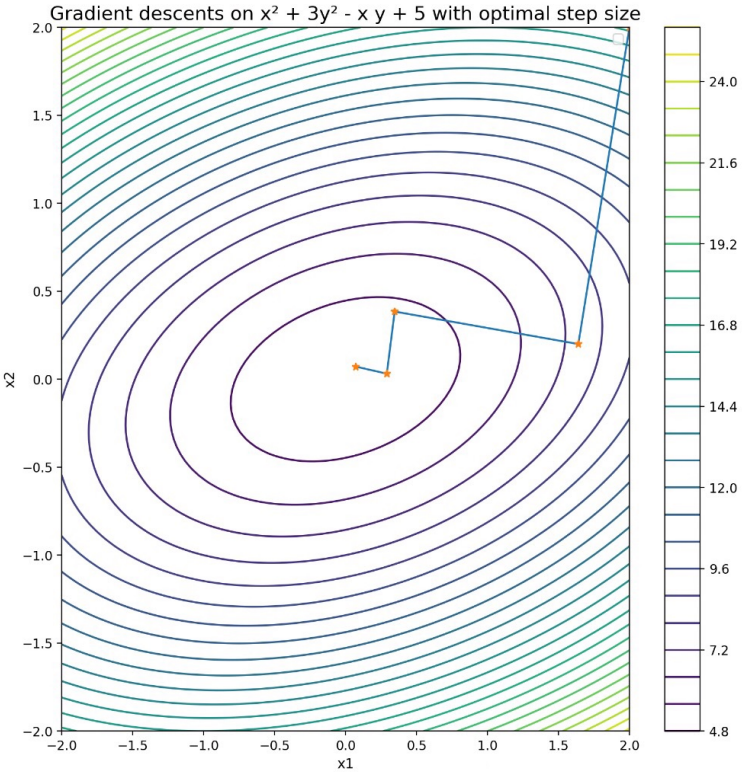
(1)


Gradient descents on x² + 3y² - x y + 5 with step size 0.1
x1
x2

(2)


Gradient descents on x² + 3y² - x y + 5 with step size 0.35
x1
x2

(3)



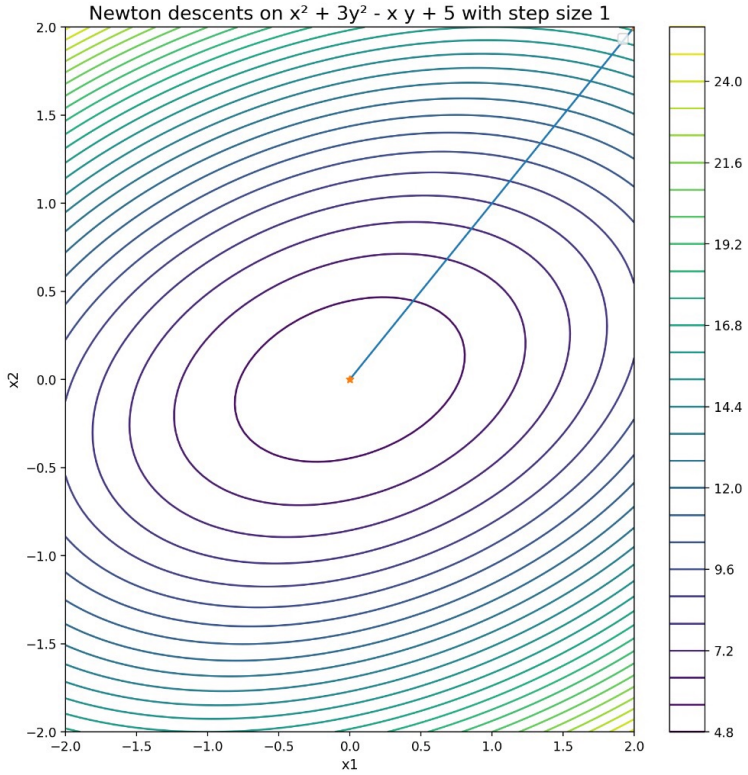
To find optimal step size, I loop through all step value between 0.01 to 10, find the step value α such:

$$\nabla f(x^{(k)} + p^{(k)} \cdot \alpha)^T \cdot p^{(k)} \text{ closest to } 0$$

$p^{(k+1)}$

which means the current step direction and next step direction is closest to orthogonal

(4)


Newton descents on x² + 3y² - x y + 5 with step size 1
x1
x2
2.0
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5
−2.0
24.0
21.6
19.2
16.8
14.4
12.0
9.6
7.2
4.8

Q6

① if $Q \in \mathbb{R}^n$ and Q is a symmetric positive semi-definite matrix, then eigenvalues of Q are real and positive

Proof of Q's eigenvalues are real:

if Q is a positive, semi-definite matrix (Also as Hermitian matrix), then

$\boxed{Q^{-T} = Q}$, $\boxed{Qx = \lambda x}$, λ is an eigenvalue of Q, x is a eigenvector to eigenvalue λ

$$\Rightarrow \bar{x}^T Q x = \bar{x}^T \lambda x = \lambda \bar{x}^T x = \lambda \|x\| \quad (1)$$

Also, $$x^T Q x = (Qx)^T \bar{x} = \bar{x}^T Q^T \bar{x} \quad (2)$$

$$\Rightarrow x^T Q^T x = \lambda \|x\| \quad (3)$$

with complete conjugate on (3), we have

$$x^T Q^T \bar{x} = \bar{\lambda} \|x\|$$

Since $Q^{-T} = Q$

$$x^T Q \bar{x} = \bar{\lambda} \|x\| = x^T \lambda x = \lambda \|x\|, \quad \lambda = \bar{\lambda}, \quad \|x\| \neq 0$$

$\therefore$ λ is a real number

Proof of λ is positive :

$Qx = \lambda x$ from (1)

$x^T Q x = \lambda x^T x = \lambda \|x\|^2$

$x^T Q x$ ⇓ positive because Q is positive semi definite, x is also positive bc they are eigenvector

Q. Which means $\lambda \|x\|^2$ λ is positive since $\|x\|^2$ is positive as well

② From spectrum theorm, a symmetric, positive definite matrix is diagnizable, which means there exists a matrix D, which

$D = \begin{bmatrix} \lambda_1 & 0 & 0 & \cdots \\ 0 & \lambda_2 & 0 & 0 \\ 0 & 0 & \lambda_3 & \ddots \\ 0 & 0 & \ddots & \lambda_n \end{bmatrix}$ that $\lambda_1 \cdots \lambda_n$ are Q's eigenvalue.

Therefore, to pair with these n eigenvalue, there are n eigenvectors.

Let v_1, v_2 be the two of arbitrary eigenvectors with eigen value λ_1, λ_2

$$Qv_1 = \lambda v_1, \quad Qv_2 = \lambda v_2, \quad Q = Q^T$$

$$Qv_1 \cdot v_2 = v_1 \cdot Q^T v_2 = v_1 \cdot Q v_2$$

$$\lambda_1 v_1 \cdot v_2 = v_1 \cdot \lambda_2 v_2$$

$$(\lambda_1 - \lambda_2) v_1 \cdot v_2 = 0, \quad \lambda_1 \neq \lambda_2$$

$\Rightarrow v_1 \cdot v_2 = 0 \Rightarrow v_1$ and v_2 are orthoganol to each other