1(a) $f(x)=\frac{1}{2}\|Ax-b\|_2^2$, $\nabla f = A^TAx - A^Tb$

for $\forall x, y \in R^n$

$\|\nabla f(x) - \nabla f(y)\| = \|A^TA(x-y)\|$

let $\lambda_1 \le \cdots \le \lambda_n$ be the eigenvalues of (A^TA), then $\lambda_1\|z\|^2 \le z^T(A^TA)z \le \lambda_n\|z\|^2$ for all $z \in R^n$

Since $A^TA \succeq 0$, then $\lambda_1 = \lambda_{min} \ge 0$

so $\lambda_1^2\|z\|_2^2 \le z^T(A^TA \cdot A^TA)z \le \lambda_n^2\|z\|_2^2 = \lambda_{max}^2\|z\|_2^2$

so $\|A^TA(x-y)\|_2^2 \le \lambda_{max}^2(A^TA)\|x-y\|_2^2$

Since $\|A^TA(x-y)\|_2 \ge 0$ and $\lambda_{max}(A^TA)\|x-y\|_2 \ge 0$

so $\|A^TA(x-y)\|_2 \le \lambda_{max}(A^TA)\|x-y\|_2$ so, $\frac{1}{2}\|Ax-b\|_2$ has $\lambda_{max}(A^TA)$-Lipschitz gradient

to prove $\frac{1}{2}\|Ax-b\|_2^2$ is $\lambda_{min}(A^TA)$-strongly convex:

$F(x)=\frac{1}{2}\|Ax-b\|_2^2 - \frac{1}{2}u\|x\|_2^2$ should be a convex, $u=\lambda_{min}(A^TA)$

We have: $\nabla^2F(x) = A^TA - \lambda_{min}(A^TA)\cdot I$

Since $A^TA \succeq 0$ all eigenvalue of A^TA is ≥ 0

so $x^T(A^TA - \lambda_{min}(A^TA))x \le x^TA^TAx \le x^T\lambda_{max}(A^TA)x$

with (iv) condition for checking if $f(x)$ is a convex: (9-16 note P2)

if $\nabla^2F(x) \le \lambda_{max}(A^TA)\cdot I$ for $\forall x$, then $F(x)$ is a convex

so $\frac{1}{2}\|Ax-b\|_2^2$ is $\lambda_{min}(A^TA)$-stronly convex

1b) $f(x) = \dfrac{\exp(c^T x)}{1+\exp(c^T x)} = \dfrac{1}{\frac{1}{\exp(c^T x)} + 1}$,

$\nabla f(x) = \left(1 + \dfrac{1}{\exp(c^T x)}\right)^{-2} \cdot \dfrac{1}{\exp(c^T x)^2} \cdot \exp(c^T x) \cdot c = \dfrac{1}{e^{c^T x} + 2 + e^{-c^T x}}\, c$

for $\forall\, x, y \in \mathbb{R}^n$ $\|\nabla f(x) - \nabla f(y)\|$, $\exists\, \nabla^2 f(z)$ that $\|\nabla f(x) - \nabla f(y)\| = \|\nabla^2 f(z)\cdot(x-y)\|$ $z \in (x,y)$ ①

let $F(t) = \dfrac{1}{1+e^{-t}}$, $F''(t) = -(e^{t} + 2 + e^{-t})^{-2} \cdot (e^{t} - e^{-t})$ lagrange mean value theorem

$\|F''(t)\| < (e^{t} + 2 + e^{-t})^{-2} \cdot (e^{t} + e^{-t}) < (e^{t} + 2 + e^{-t})^{-2} (e^{t} + e^{-t} + 2) = (e^{t} + 2 + e^{-t})^{-1} \leq \frac{1}{4}$

$|x^T \nabla^2 f(x) x| = \left| -(e^{c^T x} + 2 + e^{-c^T x})^{-2} \cdot (e^{c^T x} - e^{-c^T x}) x^T c \cdot c^T x \right| < \frac{1}{4} |x^T c \cdot c^T x|$

so $|\nabla^2 f(x)| < \frac{1}{4} c \cdot c^T$ we can make $\nabla^2 f(x)$ to be symmetric

then, $(x-y)^T \nabla^2 f(x)^T \cdot \nabla^2 f(x)(x-y) < (x-y)^T \dfrac{c \cdot c^T \cdot c \cdot c^T}{16} \cdot (x-y)$

so $\|\nabla^2 f(x)(x-y)\| < \left\| \dfrac{c \cdot c^T}{4} (x-y) \right\| \leq \frac{1}{4} c^T c \|x-y\|$

we can have $\|\nabla f(x) - \nabla f(y)\| < \underbrace{\frac{1}{4} c^T c}_{\text{Lipschitz. gradient}} \cdot \|x-y\|$

2) assume $g=\rho(\bar{x}-\bar{x}^*)$ is not a subgradient of f at $\bar{x}^*$

for simplicity, we change $\bar{x}$ to c, $\bar{x}^*$ to x^*, this is just a change of symbol

then, $\exists z\in\mathbb{R}^d$ that $f(z)<f(x^*)+g^T(z-x^*)$

since f is a convex:

$$f(x^*+\lambda(z-x^*))\le(1-\lambda)f(x^*)+\lambda f(z)<f(x^*)+\lambda g^T(z-x^*) \quad \lambda\in(0,1]$$

since x^* is a minimizer:

$$f(x^*+\lambda(z-x^*))+\frac{\rho}{2}\|x^*+\lambda(z-x^*)-c\|_2^2>f(x^*)+\frac{\rho}{2}\|x^*-c\|_2^2$$

so $f(x^*+\lambda(z-x^*))>f(x^*)+\frac{\rho}{2}\|x^*-c\|_2^2-\frac{\rho}{2}\|x^*+\lambda(z-x^*)-c\|_2^2$

so $\lambda g^T(z-x^*)>\frac{\rho}{2}\|x^*-c\|_2^2-\frac{\rho}{2}\|x^*+\lambda(z-x^*)-c\|_2^2$

divide ρ on both sides:

$$\lambda(c-x^*)^T(z-x^*)>\frac{1}{2}\|c-x^*\|_2^2-\frac{1}{2}\|c-\lambda(z-x^*)-x^*\|_2^2$$

so $(c-x^*)^T(z-x^*)>\dfrac{\frac{1}{2}\|c-x^*\|_2^2-\frac{1}{2}\|c-\lambda(z-x^*)-x^*\|_2^2}{\lambda}$

when λ is close to 0, right hand side we have:

let $F(x)=\frac{1}{2}\|c-x\|_2^2$

$$\lim_{\lambda\to 0}\frac{\frac{1}{2}\|c-x^*\|_2^2-\frac{1}{2}\|c-\lambda(z-x^*)-x^*\|_2}{\lambda}=\frac{F(x^*)-F(x^*+\lambda(z-x^*))}{\lambda}=\frac{dF}{d\lambda}\Big|_{x=x^*}$$

$$=(c-x^*)^T(z-x^*)$$

So we get $(c-x^*)^T(z-x^*)>(c-x^*)^T(z-x^*)$ □

which doesn't hold.

so g^T is a subgradient of f at $\bar{x}^*$

3) Since f is a μ-strongly convex, $\Rightarrow h(x) = f(x) - \frac{\mu}{2}\|x\|_2^2$ is a convex

a) so $f(x) = h(x) + \frac{\mu}{2}\|x\|_2^2$

since x^* minimizes f, so x^* minimizes $h(x) + \frac{\mu}{2}\|x\|_2^2$.

use the solution of Q2, let $\rho = \mu$, $\bar{x} = 0$

we have: $h(x) \geq h(x^*) + \mu(0 - x^*)^T(x - x^*)$

$$= h(x^*) - \mu x^{*T}x + \mu x^{*T}x^*$$

then, $\underbrace{h(x) + \frac{\mu}{2}\|x\|_2^2}_{f(x)} \geq h(x^*) - \mu x^{*T}x + \mu x^{*T}x^* + \frac{\mu}{2}\|x\|_2^2$

$$= h(x^*) + \frac{\mu}{2}\|x^*\|_2^2 + \frac{\mu}{2}(\|x\|_2^2 - 2x^T x^* + \|x^*\|_2^2)$$

$$= f(x^*) + \frac{\mu}{2}(x - x^*)^T(x - x^*)$$

$$= f(x^*) + \frac{\mu}{2}\|x - x^*\|_2^2$$

So we have $f(x) \geq f(x^*) + \frac{\mu}{2}\|x - x^*\|_2^2$

b) we know that $f(y_k) - f(x^*) \leq \frac{2L}{k^2}\|x_0 - x^*\|_2^2$ (9.23 note page 2)

after $k = \lceil 4\sqrt{L/\mu} \rceil$,

the inequality becomes: $f(y_k) - f(x^*) \leq \frac{2L}{16\frac{L}{\mu}}\|x_0 - x^*\|_2^2 = \frac{\mu}{8}\|x_0 - x^*\|_2^2$

since $f(x_0) - f(x^*) \geq \frac{\mu}{2}\|x_0 - x^*\|_2^2 \Rightarrow \frac{1}{4}(f(x_0) - f(x^*)) \geq \frac{\mu}{8}\|x_0 - x^*\|_2^2$

let $x_0 = y_0$

then, we have: $f(y_k) - f(x^*) \leq \frac{1}{4}(f(y_0) - f(x^*))$

since $f(y_0) - f(x^*) \geq 0$ (x^* is the minimizer)

so we have $f(y_k) - f(x^*) \leq \frac{1}{2}(f(y_0) - f(x^*))$

c) from b) we know $f(y_k) - f(x^*) \le (f(y_0) - f(x^*))/2$

after $k = \lceil 4\sqrt{L/\mu} \rceil$ steps,

then, we set 2nd $y_0 = f(y_k)$

then, after another k steps.

$$f(y_{2k}) \le (f(y_k) - f(x^*))/2 \le (f(y_0) - f(x^*))/4$$

so every k steps, the difference of current $f(y_{nk})$ to $f(x^*)$ is halved

then, it takes $\log_2\left(\frac{f(x_0) - f(x^*)}{\varepsilon}\right)$ turns for the $f(y) - f(x^*) < \varepsilon$

so it takes at most $\underbrace{4\sqrt{L/\mu}}_{k} \cdot \log_2\left(\frac{f(x_0) - f(x^*)}{\varepsilon}\right)$ steps for $f(y) - f(x^*) < \varepsilon$

4)

a)

```
ATA = A.T.dot(A)
eigenvalue, eigenvector = np.linalg.eig(ATA)
lambda_max = np.max(eigenvalue)
lambda_min = np.min(eigenvalue)
print("Lipschitz gradient constant is: ",lambda_max)
print("Strong convexity constant is: ",lambda_min)
```

```
Lipschitz gradient constant is:  5777.227211559757
Strong convexity constant is:  170.00457117919555
```

b)

```
iteration:  95
Gradient L2-norm is:  3.055200058826525

iteration:  96
Gradient L2-norm is:  2.9462920380119475

iteration:  97
Gradient L2-norm is:  2.8414846247256036

iteration:  98
Gradient L2-norm is:  2.740611751331693

iteration:  99
Gradient L2-norm is:  2.6435147379476227

iteration:  100
Gradient L2-norm is:  2.550041923482435
```

GD

c) Accelerated gradient method

```
iteration 95
Gradient norm is:  0.057167971437197485

iteration 96
Gradient norm is:  0.05252309518785786

iteration 97
Gradient norm is:  0.049417378137607014

iteration 98
Gradient norm is:  0.04779653613973019

iteration 99
Gradient norm is:  0.04738631378079718

iteration 100
Gradient norm is:  0.04772994314976655
```

yes, it outperforms (b)

d) Restarted gradient method.

```
iteration:  95
Gradient L2-norm is:  0.0007474335376580011

iteration:  96
Gradient L2-norm is:  0.000608456021036935

iteration:  97
Gradient L2-norm is:  0.0004803620977788164

iteration:  98
Gradient L2-norm is:  0.00037276592901337393

iteration:  99
Gradient L2-norm is:  0.00027705422401241197

iteration:  100
Gradient L2-norm is:  0.00020111211125739493
```

yes, it outperform (c)