

This homework is due **Friday, October 6 at 10pm.**

1 Getting Started

You may typeset your homework in latex or submit neatly handwritten and scanned solutions. Please make sure to start each question on a new page, as grading (with Gradescope) is much easier that way! Deliverables:

1. Submit a PDF of your writeup to assignment on Gradescope, “HW[n] Write-Up”
2. Submit all code needed to reproduce your results, “HW[n] Code”.
3. Submit your test set evaluation results, “HW[n] Test Set”.

After you've submitted your homework, be sure to watch out for the self-grade form.

- (a) Before you start your homework, write down your team. Who else did you work with on this homework? List names and email addresses. In case of course events, just describe the group. How did you work on this homework? Any comments about the homework?

None . work alone

Comments : usual

- (b) Please copy the following statement and sign next to it:

I certify that all solutions are entirely in my words and that I have not looked at another student's solutions. I have credited all external sources in this write up.

I certify that all solutions are entirely in my words and that I have not looked at another student's solutions. I have credited all external sources in this write up

Problem #2.

a.) $f(x) = \|x - b\|_2$

We know that $\|a\| + \|b\| \geq \|a+b\|$ for norm-2.

$$\Rightarrow \lambda \|x_1 - b\| + (1-\lambda) \|x_2 - b\|$$

$$= \|\lambda x_1 - \lambda b\| + \|(1-\lambda)x_2 - (1-\lambda)b\|$$

$$\geq \|\underbrace{\lambda x_1 + (1-\lambda)x_2 - \lambda b - (1-\lambda)b}\|$$

$$= f(\lambda x_1 + (1-\lambda)x_2) \quad \forall x_1, x_2 \in \mathbb{R}^d$$

i.e. $\lambda f(x_1) + (1-\lambda) f(x_2) \geq f(\lambda x_1 + (1-\lambda)x_2)$

$$\Rightarrow \text{convex} \quad \forall x_1, x_2 \in \mathbb{R}^d$$

b.) $f(x) = \|x - b\|_2 = \sqrt{(x-b)^T(x-b)}$

$$\Rightarrow \nabla f(x) = \frac{\frac{d}{dx}(x-b)^T(x-b)}{2\sqrt{(x-b)^T(x-b)}}$$

$$\text{numerator} = \frac{d}{dx}(x^T x - x^T b - b^T x + b^T b) = 2(x-b)$$

$$\text{denominator} = 2\|x-b\|$$

$$\Rightarrow \nabla f(x) = \frac{(x-b)}{\|x-b\|} \quad \text{unit vector}$$

$$x_0 = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \quad b = \begin{bmatrix} 4, 5 \\ 6 \end{bmatrix}$$

$$x_1 = x_0 - \nabla f(x_0) = [0.6, 0.8]^T, f(x_1) = 6.5$$

$$x_2 = x_1 - \nabla f(x_1) = [1.2, 1.6]^T, f(x_2) = 5.5$$

$$x_3 = x_2 - \nabla f(x_2) = [1.8, 2.4]^T, f(x_3) = 4.5$$

$$\dots$$

$$x_8 = x_7 - \nabla f(x_7) = [4.8, 6.4]^T, f(x_8) = 0.5$$

$$x_9 = x_8 - \nabla f(x_8) = [4.2, 5.6]^T, f(x_9) = 0.5$$

$$x_{10} = x_9 - \nabla f(x_9) = [4.8, 6.4]^T, f(x_{10}) = 0.5$$

The gradient descent does not find the optimal solution
(we know that opt. sol is $x = b = [4.5, 6]^T$ where $f(x) = 0$)

Reason: step is too large

Proof:

At $(i+1)$ -th iteration:

$$x_{i+1} = x_i - \nabla f(x_i) = x_i - \frac{x_i - b}{\|x_i - b\|}$$

we know that the true solution is $x = b$. Consider

$$\begin{aligned} x_{i+1} - b &= x_i - b - \frac{x_i - b}{\|x_i - b\|} \\ &= (x_i - b) \left(1 - \frac{1}{\|x_i - b\|} \right) \\ &= (x_i - b) \frac{\|x_i - b\| - 1}{\|x_i - b\|} \end{aligned}$$

We want $\|x_{i+1} - b\| \leq 0.01$

$$\frac{\|x_i - b\|}{\|x_i - b\|} \frac{\|x_i - b\| - 1}{\|x_i - b\|} \leq 0.01$$

$$\|x_{i+1} - b\| = \sqrt{\|x_i - b\|^2 - 1} \leq 0.01$$

let $e_{i+1} = \|x_{i+1} - b\|$

$$\Rightarrow e_{i+1} = \sqrt{e_i^2 - 1}$$

$$e_{i+1} \leq 0.01$$

$$1 - 0.01 \leq e_i \leq 1 + 0.01$$

$$i - 0.01 \leq e_i \leq i + 0.01$$

$$i - 0.01 \leq \sqrt{\begin{bmatrix} 0 \\ 0 \end{bmatrix} - \begin{bmatrix} 7.5 \\ 6 \end{bmatrix}} \leq i + 0.01$$

$$i - 0.01 \leq 7.5 \leq i + 0.01$$

\rightarrow no integer i \equiv

if $b = \begin{bmatrix} 6 \\ 8 \end{bmatrix} \Rightarrow$ we can find $i = 10$

In general for $b \neq \vec{0}$, the gradient descent converges with step size 1 iff e_i is integer. If we take

$x_0 = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \Rightarrow e_i = \sqrt{b_1^2 + b_2^2}$ is integer (where $b = \begin{bmatrix} b_1 \\ b_2 \end{bmatrix}$)

iff b_1 & b_2 are two smaller numbers of Pythagorean triple.

c.) Similar to question (b), we have:

$$e_{i+1} = \left| e_i - \left(\frac{5}{6}\right)^i \right|$$

$$e_{i+1} \leq 0.01$$

$$-0.01 \leq e_i - \left(\frac{5}{6}\right)^i \leq 0.01$$

$$\left(\frac{5}{6}\right)^i - 0.01 \leq e_i \leq \left(\frac{5}{6}\right)^i + 0.01$$

$$\left(\frac{5}{6}\right)^{i-1} - 0.01 \leq \left| e_{i-1} - \left(\frac{5}{6}\right)^{i-1} \right| \leq \left(\frac{5}{6}\right)^{i-1} + 0.01$$

$$\left(\frac{5}{6}\right)^{i-1} \left(\frac{5}{6}\right)^i - 0.01 \leq e_{i-1} \leq \left(\frac{5}{6}\right)^{i-1} + \left(\frac{5}{6}\right)^i + 0.01$$

$$\frac{5}{6} \frac{1 - \left(\frac{5}{6}\right)^i}{1 - \frac{5}{6}} - 0.01 \leq e_1 \leq \frac{5}{6} \frac{1 - \left(\frac{5}{6}\right)^i}{1 - \frac{5}{6}} + 0.01$$

$$5 \times \underbrace{\left(1 - \left(\frac{5}{6}\right)^i\right)}_{\text{around } 5} - 0.01 \leq 7.5 \leq 5 \times \underbrace{\left(1 - \left(\frac{5}{6}\right)^i\right)}_{\text{around } 5} + 0.01$$

around 5

\rightarrow no such i integer with $b = \left[\frac{6.5}{6} \right] =$

But if $e_1 \approx 5$, i.e. $\sqrt{b_1^2 + b_2^2} \approx 5$

say $b_1 = 3$, $b_2 = 4$, we can find

$$5 \left(1 - \left(\frac{5}{6}\right)^i\right) - 0.01 \leq 5 \leq 5 \times \left(1 - \left(\frac{5}{6}\right)^i\right) + 0.01$$

↑
always $\Rightarrow i = 35$

(d) Again, we have:

$$e_{i+1} = \left| e_i - \frac{1}{i+1} \right|$$

$$e_1 = 7.5 \Rightarrow e_2 = \left| 7.5 - \frac{1}{2} \right| = 7$$

$$e_3 = \left| 7 - \frac{1}{3} \right| = \frac{20}{3}$$

...
write a function to test:

```
def test(k):
```

$$e_1 = 7.5$$

```
for i in range(2, k+1):
```

$$e_1 -= \frac{1}{i}$$

```
return e1
```

```
>>> test(2731)
```

0.0107

```
>> test(2732)
```

0.0098

$\Rightarrow i = 2732$

For general $b \neq 0$, we have arbitrary error e_i . Since the harmonic series diverges, we can always find i such that after i iteration, the error

$e_i = |e_{i-1} - \frac{1}{i}|$ converges to 0

e.) See appendix

for step size = 1 and $(\frac{5}{6})^i$ it is very slowly converge

But there are some case it converge, for example

$$A = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \quad b = \begin{bmatrix} 3 \\ 4 \end{bmatrix}$$

step size = 1 \rightarrow converge after 5 time step

step size = $(\frac{5}{6})^i$ \rightarrow converge after 35 time steps

...

But if step size = $\frac{1}{i+1}$ \rightarrow always converge after some finite steps.

Explanation:

For step sizes 1 & $(\frac{5}{6})^i$ it is either too big or too small step size to converge

for step size $\frac{1}{i}$ \rightarrow we use harmonic series whose alternate series can converge to any number to make

error - (alternate harmonic series at some point)

converge to 0

Problem #3

a.) $f(x) = \frac{1}{2} \|Ax - b\|_2^2$

$$b = \vec{0} \Rightarrow f(x) = \frac{1}{2} \|Ax\|_2^2 = \frac{1}{2} x^T A^T Ax$$

$$\Rightarrow \nabla f(x) = A^T Ax$$

\Rightarrow the state evolution of iterations of gradient descent

$$x_{i+1} = x_i - \gamma \nabla f(x_i)$$

$$x_{i+1} = x_i - \gamma A^T A x_i = x_i (I - \gamma A^T A)$$

(b) From above, we can write

$$x_{i+1} = (I - \gamma A^T A)^{i+1} x_0$$

The scheme does not blow up if all eigen values of $(I - \gamma A^T A)$ less than or equal 1.

(c) $\varphi(x) = x - \gamma \nabla f(x)$

$$\text{LHS} = \|\varphi(x) - \varphi(x')\| = \|x - \gamma \nabla f(x) - x' + \gamma \nabla f(x')\|$$

$$f(x) = \frac{1}{2} \|Ax - b\|_2^2 \Rightarrow \nabla f(x) = A^T Ax - b$$

$$\Rightarrow \text{LHS} = \|x - x' - \gamma A^T Ax + \gamma A^T Ax' + \gamma b - \gamma b\|$$

$$= \|x - x' - \gamma A^T A(x - x')\|$$

$$= \|(I - \gamma A^T A)(x - x')\|$$

We know that $(I - \gamma A^T A)(x - x') = \lambda (I - \gamma A^T A)(x - x')$
 \uparrow
eigen value

$$\Rightarrow \| (I - \gamma A^T A)(x - x') \| = \| \lambda (I - \gamma A^T A)(x - x') \| \\ = \| x - x' \| | \lambda (I - \gamma A^T A) | \\ \leq \| x - x' \| | \lambda_{\max}(I - \gamma A^T A) |$$

We know that $\lambda(I + A) = 1 + \lambda(A)$

$$(\text{proof: } (I + A)x = \lambda(I + A)x)$$

$$Ix + Ax = x + \lambda(A)x$$

$$\Rightarrow \lambda(I + A)x = x(1 + \lambda(A))$$

$$\Rightarrow \lambda(I + A) = 1 + \lambda(A)$$

$$\Rightarrow |\lambda_{\max}(I - \gamma A^T A)| = \max \{ |\lambda(I - \gamma A^T A)| \}$$

$$= \max \{ |1 - \lambda(A^T A)| \}$$

$$= \max \{ |1 - \gamma \lambda_{\max}(A^T A)|, |1 - \gamma \lambda_{\min}(A^T A)| \}$$

$$= \beta$$

$$\Rightarrow \| (I - \gamma A^T A)(x - x') \| \leq \| x - x' \| \beta$$

$$\text{i.e. } \| \varphi(x) - \varphi(x') \| \leq \beta \| x - x' \|$$

From previous part, we do this to avoid blow up.

$$(d) \quad \|\varphi(x_k) - \varphi(x^*)\| = \|x_k - \gamma \nabla f(x_k) - x^* + \gamma \nabla f(x^*)\|$$

$$x_{k+1} = x_k - \gamma \nabla f(x_k)$$

\nearrow \searrow

$$\nabla f(x^*) = 0 \quad \text{since } x^* \text{ is minimum point}$$

$$\Rightarrow \|\varphi(x_k) - \varphi(x^*)\| = \|x_{k+1} - x^*\|$$

$$\begin{aligned} \|x_{k+1} - x^*\| &= \|\varphi(x_k) - \varphi(x^*)\| \leq \beta \|x_k - x^*\| \\ &= \beta \|\varphi(x_{k-1}) - \varphi(x^*)\| \\ &\leq \beta^2 \|x_{k-1} - x^*\| \\ &= \beta^2 \|\varphi(x_{k-2}) - \varphi(x^*)\| \\ &\leq \beta^3 \dots \end{aligned}$$

$$\Rightarrow \|x_{k+1} - x^*\| \leq \beta^{k+1} \|x_0 - x^*\|$$

$$\begin{aligned} e) \quad f(x) - f(x^*) &= \frac{1}{2} \|Ax - b\|^2 - \frac{1}{2} \|Ax^* - b\|^2 \\ &= \frac{1}{2} \|Ax - b\|^2 - (Ax - b)^T (Ax^* - b) + (Ax - b)^T (Ax^* - b) \\ &\quad + \frac{1}{2} \|Ax^* - b\|^2 - \|Ax^* - b\|^2 \\ &= \frac{1}{2} (Ax - b)^T (Ax - b) - (Ax - b)^T (Ax^* - b) + \frac{1}{2} (Ax^* - b)^T (Ax - b) \\ &\quad + (Ax - b)^T (Ax^* - b) - (Ax^* - b)^T (Ax^* - b) \\ &= \frac{1}{2} (\underbrace{Ax - b - Ax^* + b}_A(x - x^*))^T \underbrace{(Ax - b - Ax^* + b)}_{A(x - x^*)} + (Ax - b - Ax^* + b)^T (Ax^* - b) \end{aligned}$$

$$\begin{aligned}
 &= \frac{1}{2} (A(x-x^*))^T (A(x-x^*)) + (Ax-b-Ax^*+b)^T (Ax^*-b) \\
 &= \frac{1}{2} \|A(x-x^*)\|^2
 \end{aligned}$$

0 when $f(x)$ converges
to $f(x^*)$

Thus

$$f(x) - f(x^*) = \frac{1}{2} \|A(x-x^*)\|_2^2.$$

when $f(x) \rightarrow f(x^*)$

(f) From (e) :

$$\begin{aligned}
 f(x_k) - f(x^*) &= \frac{1}{2} \|A(x_k - x^*)\|_2^2 \\
 &= \frac{1}{2} (x_k - x^*)^T A^T A (x_k - x^*) \\
 &\quad \underbrace{\lambda(A^T A)}_{\lambda(A^T A)} (x_k - x^*) \\
 &= \frac{1}{2} \lambda(A^T A) \underbrace{(x_k - x^*)^T (x_k - x^*)}_{\|x_k - x^*\|^2} \\
 &= \frac{1}{2} \lambda(A^T A) \|x_k - x^*\|^2 \\
 &\leq \frac{1}{2} \underbrace{\lambda_{\max}(A^T A)}_{\alpha} \|x_k - x^*\|^2
 \end{aligned}$$

$$\text{From (d)} : \|x_k - x^*\| \leq \beta^k \|x_0 - x^*\|$$

$$\Rightarrow \|x_k - x^*\|^2 \leq \beta^{2k} \|x_0 - x^*\|^2$$

$$\Rightarrow f(x_k) - f(x^*) \leq \frac{\alpha}{2} \beta^{2k} \|x_0 - x^*\|^2$$

$$f(x_k) - f(x^*) = \frac{\alpha}{2} \beta^{2k} \|x_0 - x^*\|_2^2$$

iff $f(x_k)$ converges to $f(x^*)$

(g) Pick γ such that

$$1 - \gamma \lambda_{\max}(A^T A) = \gamma \lambda_{\min}(A^T A) - 1$$

$$\Rightarrow \gamma = \frac{2}{\lambda_{\min}(A^T A) + \lambda_{\max}(A^T A)}$$

$$\Rightarrow \beta = |1 - \gamma \lambda_{\max}(A^T A)|$$

$$= \left| 1 - \frac{2 \lambda_{\max}}{\lambda_{\min} + \lambda_{\max}} \right|$$

$$= \left| 1 - \frac{2 \varrho}{1 + \varrho} \right| = \frac{|1 - \varrho|}{1 + \varrho}$$

\Rightarrow The convergence rate $\frac{|1 - \varrho|}{1 + \varrho}$

Problem #4

$$(a) L(x_1, y_1) = - \sum_{i=1}^7 \left(\sqrt{(a_i - x_1)^2 + (b_i - y_1)^2} - d_i \right)^2$$

$$\frac{\partial L}{\partial x_1} = - \sum_{i=1}^7 \frac{\partial}{\partial x_1} \left((a_i - x_1)^2 + (b_i - y_1)^2 - 2d_i \sqrt{(a_i - x_1)^2 + (b_i - y_1)^2} + d_i^2 \right)$$

$$= - \sum_{i=1}^7 \left(2x_1 - 2a_i - 2d_i \frac{2x_1 - 2a_i}{2\sqrt{(a_i - x_1)^2 + (b_i - y_1)^2}} \right)$$

$$= - \sum_{i=1}^7 2(x_1 - a_i - d_i \frac{x_1 - a_i}{\sqrt{(a_i - x_1)^2 + (b_i - y_1)^2}})$$

$$\frac{\partial}{\partial y_1} = - \sum_{i=1}^7 \frac{\partial}{\partial y_1} \left((a_i - x_1)^2 + (b_i - y_1)^2 - 2d_i \sqrt{(a_i - x_1)^2 + (b_i - y_1)^2} + d_i^2 \right)$$

$$= - \sum_{i=1}^7 \left(2y_1 - 2b_i - 2d_i \frac{2y_1 - 2b_i}{2\sqrt{(a_i - x_1)^2 + (b_i - y_1)^2}} \right)$$

$$= - \sum_{i=1}^7 2(y_1 - b_i - d_i \frac{y_1 - b_i}{\sqrt{(a_i - x_1)^2 + (b_i - y_1)^2}})$$

(b) See appendix

Problem #15

done

Problem # 6

Given :

$$A = \begin{bmatrix} 3 & -1 & 1 \\ -1 & 3 & -1 \\ 1 & -1 & 3 \end{bmatrix} \quad b = \begin{bmatrix} -1 \\ 7 \\ -7 \end{bmatrix}$$

a) Can the method gradient descent can be used to solve the equation $Ax = b$?

Solution

a.) The sufficient condition is A psd.

$$\det([3]) = 3 > 0, \quad \det\left(\begin{bmatrix} 3 & -1 \\ -1 & 3 \end{bmatrix}\right) = 8 > 0$$

$$\det\left(\begin{bmatrix} 3 & -1 & 1 \\ -1 & 3 & -1 \\ 1 & -1 & 3 \end{bmatrix}\right) = 20 > 0 \quad (\text{using python})$$

$\Rightarrow A$ is PSD \Rightarrow can use gradient descent.

b.) Given a quadratic function :

$$F(x) = \frac{1}{2} x^T A x - x^T b$$

Derive the line search for this function using gradient descent

Solution

$$\nabla F(x) = Ax - b$$

Assume $x = x_0 + \alpha v$ is the line search w/ $v, \alpha \in \mathbb{R}^n$ given

$$\begin{aligned} f(\alpha) &= J(x_0 + \alpha v) = \frac{1}{2} (x_0 + \alpha v)^T A (x_0 + \alpha v) \\ &\quad - (x_0 + \alpha v)^T b \\ &= \frac{1}{2} [x_0^T A x_0 + 2\alpha v^T A x_0 + \alpha^2 v^T A v] \\ &\quad - x_0^T b - \alpha v^T b \end{aligned}$$

$f(\alpha)$ should be minimum

$$\begin{aligned} \Rightarrow \frac{\partial f(\alpha)}{\partial \alpha} &= v^T A x_0 + \alpha v^T A v - v^T b = 0 \\ \Rightarrow \alpha &= \frac{v^T (b - A x_0)}{v^T A v} \end{aligned}$$