Problem 1. μ-strong convexity : for any $x,y \in X$ and $t \in [0,1]$, $\exists \mu > 0$ s.t.

$$f(tx+(1-t)y) \le t\cdot f(x) + (1-t)f(y) - \frac{\mu}{2}t(1-t)\|x-y\|^2 \cdots ①$$

C1 $f(y) \ge f(x) + \nabla f(x)^T(y-x) + \frac{\mu}{2}\|y-x\|^2$

C2. $(\nabla f(x) - \nabla f(y))^T(x-y) \ge \mu\|x-y\|^2$

C3. $\nabla^2 f(x) - \mu I \succ 0$

① $\Leftrightarrow$ C1. $f(tx+(1-t)y) \le tf(x) + f(y) - tf(y) - \frac{\mu}{2}t(1-t)\|x-y\|^2$

(calculate $\frac{d}{dt}\big|_{t=0}$) LHS: $\frac{d}{dt}f(tx+(1-t)y)\big|_{t=0} = \nabla f(y)^T(x-y)$

RHS: $\frac{d}{dt}\, tf(x)+(1-t)f(y) - \frac{\mu}{2}t(1-t)\|x-y\|^2\big|_{t=0} = \left[f(x)-f(y) - (1-2t)\cdot\frac{\mu}{2}\|x-y\|^2\right]_{t=0}$

$= f(x) - f(y) - \frac{\mu}{2}\|x-y\|^2$

$\Leftrightarrow \nabla f(y)^T(x-y) \le f(x) - f(y) - \frac{\mu}{2}\|x-y\|^2$

Rearrange info $\Leftrightarrow f(x) \ge f(y) + \nabla f(y)^T(x-y) + \frac{\mu}{2}\|x-y\|^2$

Interchange x, y $\Leftrightarrow f(y) \ge f(x) + \nabla f(x)^T(y-x) + \frac{\mu}{2}\|y-x\|^2$ ✱

$C_1 \Leftrightarrow C_2$. $(C_1)\begin{cases} f(x) \ge f(y) + \nabla f(y)^T(x-y) + \frac{\mu}{2}\|x-y\|^2 & \cdots \text{a.} \\ f(y) \ge f(x) + \nabla f(x)^T(y-x) + \frac{\mu}{2}\|y-x\|^2 & \cdots \text{b.}\end{cases}$

$(a.+b.) \Leftrightarrow \cancel{f(x)}+\cancel{f(y)} \ge \cancel{f(x)}+\cancel{f(y)} + f(y)^T(x-y) + \nabla f(x)^T(y-x) + \mu\|x-y\|^2$

(Rearrange) $\Leftrightarrow (\nabla f(x) - \nabla f(y))^T(x-y) \ge \mu\|x-y\|^2$ ✱

$C_1 \Leftrightarrow C_3$. (C_1) $f(y) \ge f(x) + \nabla f(x)^T(y-x) + \frac{\mu}{2}\|y-x\|^2 \cdots$ c.

(Taylor) $f(y) = f(x) + \nabla f(x)^T(y-x) + \frac{1}{2}(y-x)^T\nabla^2 f(x)(y-x) + o(\|y-x\|^2) \cdots$ d.

(by c, d) $\Leftrightarrow \frac{1}{2}(y-x)^T\nabla^2 f(x)(y-x) + o(\|y-x\|^2) \ge \frac{\mu}{2}\|y-x\|^2$

$(\times\frac{2}{\|y-x\|^2}$ on both side $) \Leftrightarrow \frac{(y-x)^T}{\|y-x\|}\nabla^2 f(x)\frac{(y-x)}{\|y-x\|} + \cancel{o(1)} \ge \mu$

$\Leftrightarrow \nabla^2 f(x) \succeq \mu I$

Problem 2.

$$f(x) = \begin{cases} 25x^2 & x<1 & \cdots f_1(x) \\ x^2+48x-24 & 1\le x\le 2 & \cdots f_2(x) \\ 25x^2-48x+72 & 2<x & \cdots f_3(x)\end{cases}$$

$f_1(1) = 25 = f_2(1)$, $f_2(2) = 76 = f_3(2)$

$f_1'(x) = 50x$, $f_2'(x) = 2x+48$, $f_3'(x) = 50x-48$

$f_1'(1) = 50 = f_2'(1)$, $f_2'(2) = 52 = f_3'(2)$

$f_1'' = 50$, $f_2'' = 2$, $f_3'' = 50$

Since the function pieces are continuous and differentiable at break points, so $f(x)$ is continuous and differentiable.

2. a. 2-strongly convex :

for $x<1$, $\nabla^2 f(x) = 50 \ge 2$, for $1\le x\le 2$, $\nabla^2 f(x) = 2 \ge 2$, for $x>2$, $\nabla^2 f(x) = 50 \ge 2$

else, check $(\nabla f(x) - \nabla f(y))^T(x-y) \ge \mu\|x-y\|^2$

Since $x<y$

for $x<1,\ 1\le y\le 2$: $(50x-(2y+48))(x-y) = (50x-2y-48)\cancel{(x-y)} \overset{\le}{\cancel{\ge}} 2\|x-y\|$

$\Rightarrow\ 50x-2y-48 \le 2x-2y$ holds for $x<1$ #

for $x<1,\ 2<y$: $(50x-50y+48)\cancel{(x-y)} \overset{\le}{\cancel{\ge}} 2\|x-y\|$

$\Rightarrow\ 50x-50y+48 \le 2x-2y$

$\Rightarrow\ 48x+48 \le 48y$ holds for $x<1,\ y>2$ #

for $1\le x\le 2,\ y>2$: $(2x+48-50y+48)\cancel{(x-y)} \overset{\le}{\cancel{\ge}} 2\|x-y\|$

$\Rightarrow\ 2x-50y+96 \le 2x-2y$

$\Rightarrow\ 48y \ge 96$ holds for $y>2$ #

By all the cases above, $f(x)$ is 2-strongly convex.

L-smooth: $\nabla^2 f(x) \preceq LI$. $\|\nabla f(x)-\nabla f(y)\| \le L\|x-y\|$

for $x<1$, $\nabla^2 f(x) = 50 \le 50$, for $1\le x\le 2$, $\nabla^2 f(x) = 2 \le 50$, for $x\ge 2$, $\nabla^2 f(x) = 50 \le 50$

for $x<1,\ 1\le y\le 2$: $\|50x-2y-48\| \le 50\|x-y\|$

$\Rightarrow\ 50x-50y \le 50x-2y-48 \le 50y-50x$

$\Rightarrow \begin{cases} \cancel{50x}-50y \le \cancel{50x}-2y-48 & \text{holds for } y\ge 1 \\ 50y-50x \ge 50x-2y-48 \end{cases}$

$\Rightarrow\ 100x \le 52y+48$ holds for $x<1,\ 1\le y\le 2$ #

for $x<1,\ y>2$: $\|50x-50y+48\| \le 50\|x-y\|$

$\Rightarrow \begin{cases} 50x-50y+48 \ge 50x-50y & \text{always holds} \\ 50x-50y+48 \le 50y-50x \end{cases}$

$\Rightarrow\ 100x+48 \le 100y$ holds for $x<1,\ y>2$ #

for $1\le x\le 2,\ y>2$: $\|2x+48-50y+48\| \le 50\|x-y\|$

$\Rightarrow \begin{cases} 2x-50y+96 \ge 50x-50y & \text{holds for } 1\le x\le 2 \\ 2x-50y+96 \le 50y-50x \end{cases}$

$\Rightarrow\ 52x+96 \le 100y$ holds for $1\le x\le 2,\ y>2$ #

By all the six cases above, we know that $f(x)$ is L-smooth with $L=50$ #

2.b. Since $f(x)$ is convex, so x with $f'(x)=0$ implies it is a global optimizer.

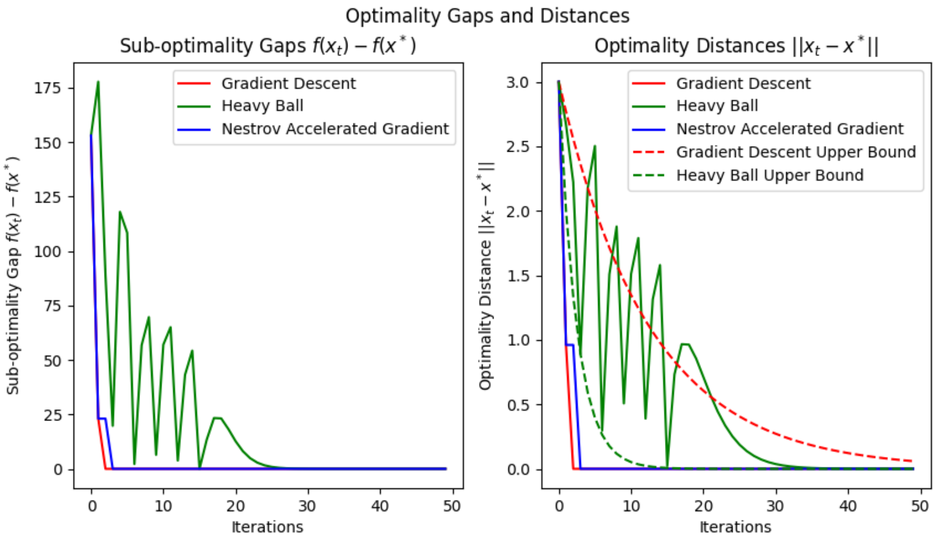
for $x<1$, $f'(x) = 50x = 0 \Leftrightarrow x=0$ (✓)

for $1\le x\le 2$, $f'(x) = 2x+48 = 0 \Leftrightarrow x=-24$ ($x\notin[1,2]$)

for $x>2$, $f'(x) = 50x-48 \Leftrightarrow x=0.96$ ($x\notin(2,\infty)$)

So the global optimizer $x^*=0$ with $f(x^*)=0$ #

2. (c) Gradient Descent method performs a really small optimality distance compares to the upper bound, while heavy ball method seems always swinging above the upper bound. I think it is because this function is not really a quadratic function.



Problem 3.a. $f(x) \geq f(y) + \nabla f(y)^T(x-y)$ by convexity $\cdots$ ①

$f(y - \frac{1}{L}\nabla f(y)) - f(y) \leq \nabla f(y)^T(-\frac{1}{L}\nabla f(y)) + \frac{L}{2}\|-\frac{1}{L}\nabla f(y)\|^2$ by L-smoothness

$= -\frac{1}{2L}\|\nabla f(y)\|^2 \cdots$ ②

By ①×(−1)+②: $f(y-\frac{1}{L}\nabla f(y)) - f(x) \leq -\frac{1}{2L}\|\nabla f(y)\|^2 - \nabla f(y)^T(x-y)$ # $\cdots$ ⓐ

3.b. $x_{t+1} = y_t - \frac{1}{L}\nabla f(y_t) \iff f(y_t - \frac{1}{L}\nabla f(y_t)) = f(x_{t+1})$, $\nabla f(y_t) = -L(x_{t+1} - y_t)$

Let $y = y_t$, $x = x_t$ in ⓐ: $f(y_t - \frac{1}{L}\nabla f(y_t)) - f(x_t) = f(x_{t+1}) - f(x_t)$

$\leq -\frac{1}{2L}\|L(x_{t+1}-y_t)\|^2 + L(x_{t+1}-y_t)^T(x_t - y_t)$

$= -\frac{L}{2}\|x_{t+1}-y_t\|^2 + L(x_{t+1}-y_t)^T(x_t-y_t)$ # $\cdots$ ③

Similarly, Let $y = y_t$, $x = x^*$ in ⓐ, we can get

$f(x_{t+1}) - f(x^*) \leq -\frac{L}{2}\|x_{t+1}-y_t\|^2 + L(x_{t+1}-y_t)^T(x^*-y_t)$ # $\cdots$ ④

3.c. $\theta_t(\theta_t - 1)\times$③ $+ \theta_t \times$④: LHS $\theta_t(\theta_t-1)(f(x_{t+1}) - f(x_t)) + \theta_t(f(x_{t+1}) - f(x^*))$

$= \theta_t^2 f(x_{t+1}) - \theta_t^2 f(x_t) + \theta_t f(x_t) - \theta_t f(x^*)$

(by $\theta_t^2 - \theta_t - \theta_{t-1}^2 = 0$) $= \theta_t^2 f(x_{t+1}) \cancel{- \theta_t^2 f(x_t)} + (\cancel{\theta_t^2} - \theta_{t-1}^2) f(x_t) - (\theta_t^2 - \theta_{t-1}^2) f(x^*)$

$= \theta_t^2(f(x_{t+1}) - f(x^*)) - \theta_{t-1}^2(f(x_t) - f(x^*))$

(Let $\Delta_t := f(x_t) - f(x^*)$) $= \theta_t^2 \Delta_{t+1} - \theta_{t-1}^2 \Delta_t$

RHS $\leq -\frac{L}{2}\theta_t^2\|x_{t+1}-y_t\|^2 + L(x_{t+1}-y_t)^T(\theta_t(\theta_t-1)(x_t-y_t) + \theta_t(x^*-y_t))$

$= -\frac{L}{2}\|\theta_t(x_{t+1}-y_t)\|^2 + \frac{L}{2}\cdot 2\theta_t(x_{t+1}-y_t)^T((\theta_t-1)x_t - \theta_t y_t + x^*)$

$= -\frac{L}{2}(\|\theta_t(x_{t+1}-y_t)\|^2 + 2\theta_t(x_{t+1}-y_t)^T(\theta_t y_t - (\theta_t-1)x_t - x^*))$

$$\Rightarrow \theta_t^2\Delta_{t+1} - \theta_{t-1}^2\Delta_t \leq -\frac{L}{2}\underbrace{(\|\theta_t(x_{t+1}-y_t)\|^2 + 2\theta_t(x_{t+1}-y_t)^T(\theta_t y_t - (\theta_t-1)x_t - x^*))}\ \# \cdots ⓒ$$

3.d. Complete the square in ⓒ: $\|\theta_t(x_{t+1}-y_t)\|^2 + 2\theta_t(x_{t+1}-y_t)^T(\theta_t y_t - (\theta_t-1)x_t - x^*)$

(Let $\phi_t := \theta_t y_t - (\theta_t-1)x_t - x^*$) $= \|\theta_t(x_{t+1}-y_t)\|^2 + 2\theta_t(x_{t+1}-y_t)^T\phi_t + \|\phi_t\|^2 - \|\phi_t\|^2$

$= \|\theta_t(x_{t+1}-y_t) + \phi_t\|^2 - \|\phi_t\|^2$

$= \|\theta_t x_{t+1} \cancel{- \theta_t y_t} \cancel{+ \theta_t y_t} - (\theta_t-1)x_t - x^*\|^2 - \|\phi_t\|^2 \cdots$ ⑤

Note that by the update rule (6.):

$y_{t+1} = x_{t+1} - \frac{(1-\theta_t)}{\theta_{t+1}}(x_{t+1} - x_t)$

both $\times\theta_{t+1}$ $\Rightarrow \theta_{t+1}y_{t+1} = \theta_{t+1}x_{t+1} + (\theta_t - 1)x_{t+1} - (\theta_t - 1)x_t$

$\Rightarrow -(\theta_t-1)x_t + \theta_t x_{t+1} = \theta_{t+1}y_{t+1} - (\theta_{t+1}-1)x_{t+1} \cdots$ ⑥

Plug ⑥ into ⑤:

⑤ $= \|\underline{\theta_t x_{t+1} - (\theta_t-1)x_t} - x^*\|^2 - \|\phi_t\|^2$

$= \|\theta_{t+1}y_{t+1} - (\theta_{t+1}-1)x_{t+1} - x^*\|^2 - \|\phi_t\|^2$

$= \|\phi_{t+1}\|^2 - \|\phi_t\|^2 \cdots$ ⑦

Therefore, by Ⓒ, ⑤, ⑥, ⑦, we can get:

$$\theta_t^2\Delta_{t+1} - \theta_{t-1}^2\Delta_t \le -\frac{L}{2}\left(\|\phi_{t+1}\|^2 - \|\phi_t\|^2\right) = \frac{L}{2}\left(\|\phi_t\|^2 - \|\phi_{t+1}\|^2\right) \#\quad \cdots Ⓓ$$

3.e. Telescope summing Ⓓ from $t=1$ to $T-1$

$$t=1: \cancel{\theta_1^2\Delta_2} - \theta_0^2\Delta_1 \le \frac{L}{2}\left(\|\phi_1\|^2 - \cancel{\|\phi_2\|^2}\right)$$

$$t=2: \cancel{\theta_2^2\Delta_3} - \cancel{\theta_1^2\Delta_2} \le \frac{L}{2}\left(\cancel{\|\phi_2\|^2} - \cancel{\|\phi_3\|^2}\right)$$

$$\vdots$$

$$+)\ t=T-1: \theta_{T-1}^2\Delta_T - \cancel{\theta_{T-2}^2\Delta_{T-1}} \le \frac{L}{2}\left(\cancel{\|\phi_{T-1}\|^2} - \|\phi_T\|^2\right)$$

$$\Rightarrow \theta_{T-1}^2\Delta_T - \theta_0^2\Delta_1 \le \frac{L}{2}\left(\|\phi_1\|^2 - \|\phi_T\|^2\right) \le \frac{L}{2}\|\phi_1\|^2$$

$$\Rightarrow \theta_{T-1}^2\Delta_T \le \frac{L}{2}\|\phi_1\|^2 + \cancel{\theta_0^2\Delta_1} \quad \theta_0 = 0$$

$$= \frac{L}{2}\|\theta_1 y_1 - (\theta_1 - 1)x_1 - x^*\|^2$$

$$= \frac{L}{2}\|x_1 - x^* - \cancel{\theta_1 x_1 + \theta_1 y_1}\|^2 \quad x_1 = y_1$$

$$\Rightarrow \Delta_T \le \frac{L}{2\theta_{T-1}^2}\|x_1 - x^*\|^2 \quad \cdots ⑧$$

To Prove: $\theta_{t-1} \ge \frac{t}{2}$

Case $t=1$, $\theta_0 = 0 \ge 0$

Suppose it holds up to $t=n-1$, $\theta_{n-2} \ge \frac{n-1}{2}$

Case $t=n$, $\theta_{n-1} = \frac{1}{2}\left(1+\sqrt{1+4\theta_{n-2}^2}\right)$

$$\ge \frac{1}{2}\left(1+\sqrt{1+(n-1)^2}\right) \quad \text{by hypothesis}$$

$$> \frac{1}{2}\left(1+\sqrt{n^2-2n+1}\right)$$

$$= \frac{1}{2}(1+n-1) = \frac{n}{2} \#$$

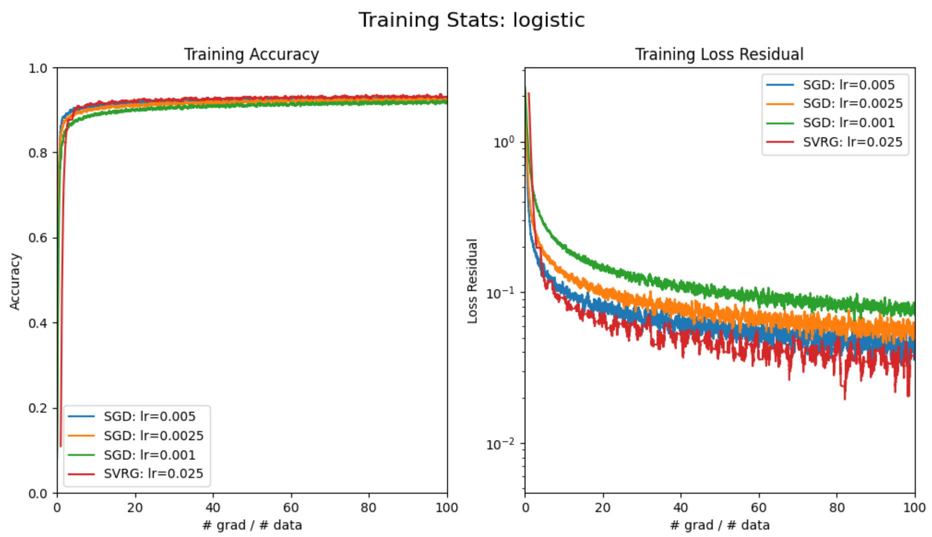
by induction, $\theta_{t-1} \ge \frac{t}{2}$ holds for $t \ge 1 \quad \cdots ⑨$

by ⑧ and ⑨, $\Delta_T = f(x_t) - f(x^*) \le \frac{L}{2\theta_{t-1}^2}\|x_1 - x^*\|^2 \le \frac{L}{2}\left(\frac{2}{t}\right)^2\|x_1 - x^*\|^2 = \frac{2L}{t^2}\|x_1 - x^*\|^2 \#$

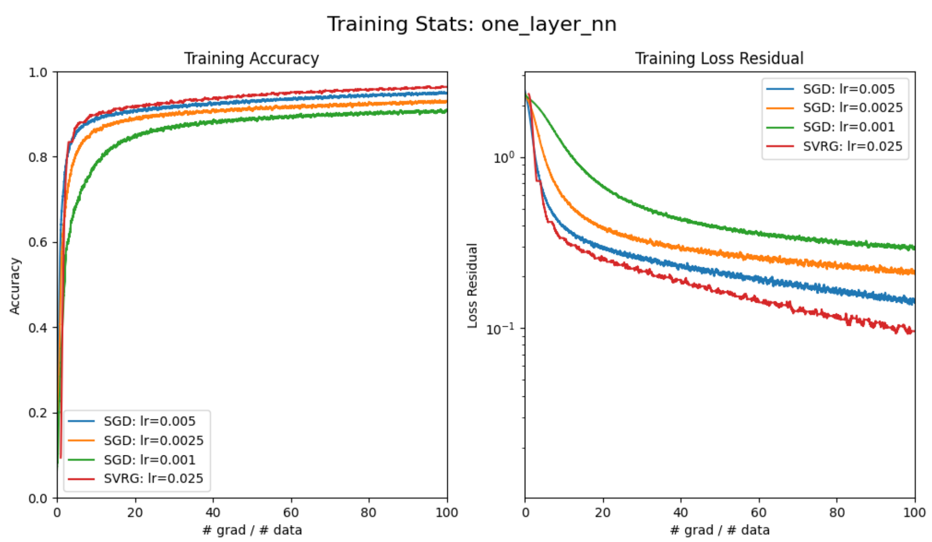
Problem 4.

μ	$f(x^*)$	x^*			
0	-0.12	1	0	0	0
0.1	-0.104694	$6.38 * 10^{-1}$	$3.61 * 10^{-1}$	$2.76 * 10^{-9}$	$2.02 * 10^{-10}$
1.0	-0.0715251	$2.36 * 10^{-1}$	$3.97 * 10^{-1}$	$3.66 * 10^{-1}$	$1.06 * 10^{-9}$
2.0	-0.0524063	0.16754123	0.31296852	0.35757121	0.16191904
5.0	-0.0232699	0.09715142	0.19490255	0.24962519	0.45832084
10.0	0.0126792	0.07368816	0.15554723	0.21364318	0.55712144

Problem 5. (a) logistic: SVRG loss outperforms SGD a sometimes.



Problem 5. (b) one-layer NN: SVRG outperforms SGD (in loss residual) more than pervious experiment.



Hyperparameters:

Batch size: 64
Learning rate: SGD = [0.005, 0.0025, 0.001], SVRG = 0.025, GD = 0.05
Iterations: 100

Logistic primal optimal (loss=0.21524): GD for 1000 iterations
NN primal optimal (loss=0.034793): SGD with lr = 0.005 for 1000 iterations

NN hidden layer: 128 nodes