Group 9:

HW 1.a

Strong Convexity

①

This is the definition of Strong Convexity.

$$(4) \quad f(x_2) \geq f(x_1) + \nabla f(x_1)^T (x_2 - x_1) + \frac{\mu}{2} \|x_2 - x_1\|_2^2$$

Prove that (4) is equivalent to

$$\nabla^2 f(x) \geq \mu I_d, \quad \forall x \in \mathcal{X} \qquad (4.1)$$

Start with (4) and take derivative of the inequality on both sides w.r.t x_2

$$f(x_2) \geq \overset{\to 0}{f(x_1)} + \nabla f(x_1)^T (x_2 - x_1) + \frac{\mu}{2} \|x_2 - x_1\|_2^2$$

$$\nabla f(x_2) \geq \overset{\to 0}{\cancel{\nabla f(x_1)^T}} + \cancel{2} \cdot \frac{\mu}{\cancel{2}} \cdot x_2$$

$$\nabla^2 f(x_2) \geq \mu I_d$$

✓

Prove that (4) is equivalent to

$$\left(\nabla f(x_2) - \nabla f(x_1)\right)^T (x_2 - x_1) \geq \mu \|x_2 - x_1\|_2^2 \qquad (4.2)$$

Start with (4)

$$f(x_2) \geq f(x_1) + \nabla f(x_1)^T (x_2 - x_1) + \frac{\mu}{2} \|x_2 - x_1\|_2^2$$

replace x_2 with x_1 in (4)

$$f(x_1) \geq f(x_2) + \nabla f(x_2)^T (x_1 - x_2) + \frac{\mu}{2} \|x_1 - x_2\|_2^2$$

Sum the 2 inequalities

$$\cancel{f(x_2)} + \cancel{f(x_1)} \geq \cancel{f(x_1)} + \cancel{f(x_2)} + \nabla f(x_1)^T (x_2 - x_1) + \nabla f(x_2)^T (x_1 - x_2) + \frac{\mu}{2} \|x_2 - x_1\|_2^2 + \frac{\mu}{2} \|x_1 - x_2\|_2^2$$

same

$$-\nabla f(x_1)^T (x_2 - x_1) - \nabla f(x_2)^T (x_1 - x_2) \geq \mu \|x_2 - x_1\|_2^2$$

$$\nabla f(x_2)^T (x_2 - x_1) - \nabla f(x_1)^T (x_2 - x_1) \geq \mu \|x_2 - x_1\|_2^2$$

✓ $$\left(\nabla f(x_2) - \nabla f(x_1)\right)^T (x_2 - x_1) \geq \mu \|x_2 - x_1\|_2^2$$

(4 b) Prove that (4) implies

$$\|x_2 - x_1\|_2 \leq \frac{1}{\mu} \|\nabla f(x_2) - \nabla f(x_1)\|_2 \quad \forall x_1, x_2$$

Start with 4·2 which has already been proved

$$(\nabla f(x_2) - \nabla f(x_1))^T (x_2 - x_1) \geq \mu \|x_2 - x_1\|_2^2$$

expand the norm and move the μ

$$\frac{1}{\mu}(\nabla f(x_2) - \nabla f(x_1))^T (x_2 - x_1) \geq (x_2 - x_1)^T (x_2 - x_1)$$

Take norm on both sides of inequality

$$\frac{1}{\mu} \|\nabla f(x_2) - \nabla f(x_1)\|_2 \geq \|x_2 - x_1\| \qquad (4.4)$$

(4c) Prove that (4) implies

$$(\nabla f(x_2) - \nabla f(x_1))^T (x_2 - x_1) \leq \frac{1}{\mu} \|\nabla f(x_2) - \nabla f(x_1)\|_2^2 \quad - (4.5) \qquad \forall x_1, x_2$$

write this in another equivalent form.

$$(\nabla f(x_2) - \cancel{\nabla f(x_1)})^T (x_2 - x_1) \leq \frac{1}{\mu} (\nabla f(x_2) - \nabla f(x_1))^T \cdot (\nabla f(x_2) \cancel{- \nabla f(x_1)})$$

another equivalent form by taking norm on both sides of the inequality to prove.

$$\|x_2 - x_1\|_2 \leq \frac{1}{\mu} \|\nabla f(x_2) - \nabla f(x_1)\|_2$$

$\hookrightarrow$ this is the same as (4·4) which we have already proved. Hence (4·5) is also proved.

HW1.b

Smoothness

A function f is L-smooth iff
- it is differentiable
- gradient is L-Lipschitz-continuous

(5) $\|\nabla f(x_2) - \nabla f(x_1)\|_2 \leq L\|x_2 - x_1\|_2 \quad \forall x_1, x_2 \in \mathcal{X}$

The physical meaning here seems to be that smoothness is how fast the curvature of the curve changes. So for it to be L-smooth the magnitude of the difference in gradient between 2 points on the curve $\|\nabla f(x_2) - \nabla f(x_1)\|_2$ should be less than a multiple L of the distance between the 2 points $L\|x_2 - x_1\|_2$. Slower the change in curvature, the smoother it is.

for twice differentiable f $\nabla^2 f(x) \leq L I_d$ from (5)

(5 a) Prove that (5) implies
$f(x_2) \leq f(x_1) + \nabla f(x_1)^T(x_2 - x_1) + \frac{L}{2}\|x_2 - x_1\|_2^2$
start with the statement to prove and modify it to show that it is the same as (5)
Sum the inequality for $f(x_2)$ and $f(x_1)$

$$f(x_2) + f(x_1) \leq f(x_1) + f(x_2) + \nabla f(x_1)^T(x_2 - x_1) + \nabla f(x_2)^T(x_1 - x_2) + \frac{L}{2}\|x_2 - x_1\|_2^2 + \frac{L}{2}\|x_1 - x_2\|_2^2$$

$$\nabla f(x_2)^T(x_2 - x_1) - \nabla f(x_1)^T(x_2 - x_1) \leq L\|x_2 - x_1\|_2^2$$

$$(\nabla f(x_2) - \nabla f(x_1))^T(x_2 - x_1) \leq L(x_2 - x_1)^T(x_2 - x_1)$$

Do norm on both sides

$$\|\nabla f(x_2) - \nabla f(x_1)\|_2 \leq L\|x_2 - x_1\|_2$$

this is (5) which is the condition for L-smooth.
Hence proved.

HW1(c)

Newton's method is not optimal when N is large since it requires the computation of the second derivative which can be complicated or intractable when N is large. If N values are required to compute the first derivate, the N^2 values are required for the second derivative.

(a) When N = 1000 we can use the Newton's method since the second derivative will require 1,000,000
(b) If N = 10^9, the second derivative will require 10^{9x2} values which is very expensive
(c) When the Hessian matrix (second order gradient matrix) is too expensive to compute some Quasi Newton methods can be applied that simplify the computation of the Hessian matrix with some compromise in the speed of convergence. DFP and BFGS methods are a couple we came across in literature.
(d) Adding the twice diffrentiable function r(x) to the objective function will not change the answers for (a) - © while using the newton method since Newton method is affine and the gradient of the sum function can be found as the sum of the individual functions.

HW-2A

HW 2B

Given equations

(2a) - $\nabla f(w_k)^T \mathbb{E}_{\xi_k}[g(w_k; \xi_k)] \geq c \|\nabla f(w_k)\|_2^2$

(2b) - $\|\mathbb{E}_{\xi_k}[g(w_k; \xi_k)]\|_2 \leq C_0 \|\nabla f(w_k)\|_2$

existing scalar

$M \geq 0$ and $M_V \geq 0$ s.t for all $k \in \mathbb{N}$

$$\mathrm{Var}_{\xi_k}[g(w_k; \xi_k)] \leq M + M_V \|\nabla f(w_k)\|_2^2$$

For (2) and (3) ; this implies

$$\mathbb{E}_{\xi_k}[\|g(w_k; \xi_k)\|_2^2] \leq \alpha + \beta \|\nabla f(w_k)\|_2^2$$

$$\mathrm{Var}_{\xi_k}[g(w_k; \xi_k)] \leq \mathbb{E}_{\xi_k}[\|g(w_k); \xi_k\|_2^2] - [\mathbb{E}_{\xi_k}[g(w_k); \xi_k)]]^2$$
$$\leq M + M_V \|\nabla f(w_k)\|_2^2$$

$$\mathbb{E}_{\xi_k}[\|g(w_k; \xi_k)\|_2^2] \leq M + M_V \|\nabla f(w_k)\|_2^2 + [\mathbb{E}_{\xi_k}[g(w_k; \xi_k)]]^2$$

$\therefore$

$$\mathbb{E}_{\xi_k}[\|g(w_k; \xi_k)\|_2^2 \leq M + M_V \|\nabla f(w_k)\|_2^2 + C_0^2 \|\nabla f(w_k)\|_2^2$$
$$\leq M + (M_V + C_0^2) \|\nabla f(w_k)\|_2^2$$

where by $\alpha = M$ and $\beta = M_V + C_0^2$

for condition of unbiased gradient estimator

$c = C_0 = 1$ then

$\alpha = M$, $\beta = M_V + 1$

HW-2C

- Square summable but not summable step-size

suppose

$$||\alpha||_2^2 = \sum_{k=1}^{\infty} \alpha_k^2 < \infty \quad , \quad \sum_{k=1}^{\infty} \alpha_k = \infty$$

Then we have;

$$f_{best}^{(k)} - f^* \leq \frac{R^2 + G^2 ||\alpha||_2^2}{2 \sum_{i=1}^{k} \alpha_i}$$

This converges to zero as $k \to \infty$, since the numerator converges to $R^2 + G^2 ||\alpha||_2^2$, and the denominator grows without bound. Thus, the gradient method converges;

W.R.T.:

$$E\left[\frac{1}{\sum_{k \in [k]} a_k} \sum_{k \in [k]} \alpha_k ||\nabla f(w_k)||_2^2 \right] \xrightarrow[k]{\infty} 0$$

The denominator $\sum_{k \in [k]} a_k$ will continue to grow without bound and hence diminished;

this will cause the equation to be $\frac{1}{\infty} = 0$;

CA1, CA2:

CA1: Closed-form solution vs iterative approaches

Consider $\boldsymbol{x}^\star = \underset{\boldsymbol{w}\in\mathbb{R}^d}{\text{minimize}} \ \frac{1}{N} \sum_{i\in[N]} \|\boldsymbol{w}^T\boldsymbol{x}_i - \boldsymbol{y}_i\|^2 + \lambda\|\boldsymbol{w}\|_2^2$ for dataset $\{(\boldsymbol{x}_i, \boldsymbol{y}_i)\}$

1) Find a closed-form solution for this problem

2) Consider "Communities and Crime" dataset ($N = 1994$, $d = 128$) and find the optimal linear regressor from the closed-form expression

3) Repeat 2) for "Individual household electric power consumption" dataset ($N = 2075259$, $d = 9$) and observe the scalability issue of the closed-form expression

4) How would you address even bigger datasets?

CA2: Deterministic/stochastic algorithms in practice

Consider logistic ridge regression $f(\boldsymbol{w}) = \frac{1}{N}\sum_{i\in[N]} f_i(\boldsymbol{w}) + \lambda\|\boldsymbol{w}\|_2^2$ where $f_i(\boldsymbol{w}) = \log\left(1 + \exp\{-y_i\boldsymbol{w}^T\boldsymbol{x}_i\}\right)$ for "Individual household electric power consumption" dataset

1) Solve the optimization problem using GD, stochastic GD, SVRG, and SAG

2) Tune a bit hyper-parameters (including λ)

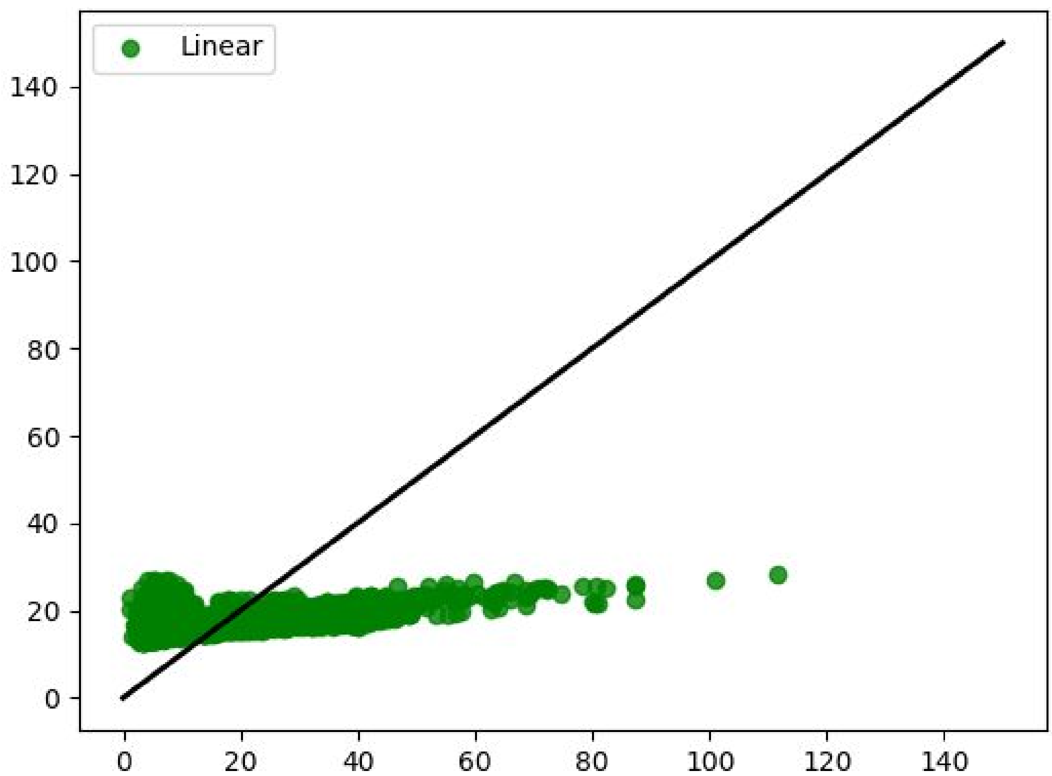
3) Compare these solvers in terms complexity of hyper-parameter tunning, convergence time, convergence rate (in terms of # outer-loop iterations), and memory requirement

CA1:

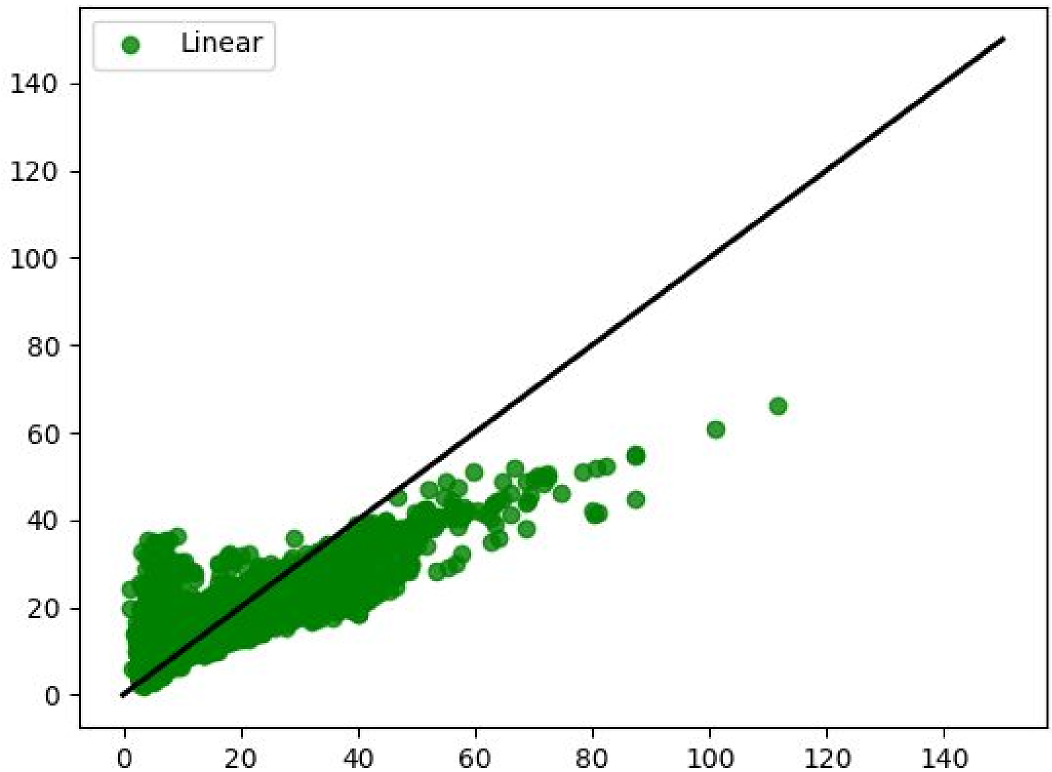
CA2:

1. The solution for optimization problem using GD, and stochastic GD are available on Github GD: https://github.com/mlongr9/MLONs-Assignments/blob/master/ridge_gradient_descent.py SGD: https://github.com/mlongr9/MLONs-Assignments/blob/master/ridge_stochastic_gradient_descent.py. The solutions are derived or adapted with modifications from the pyRidge https://github.com/vikasrtr/pyRidge (for education purpose).

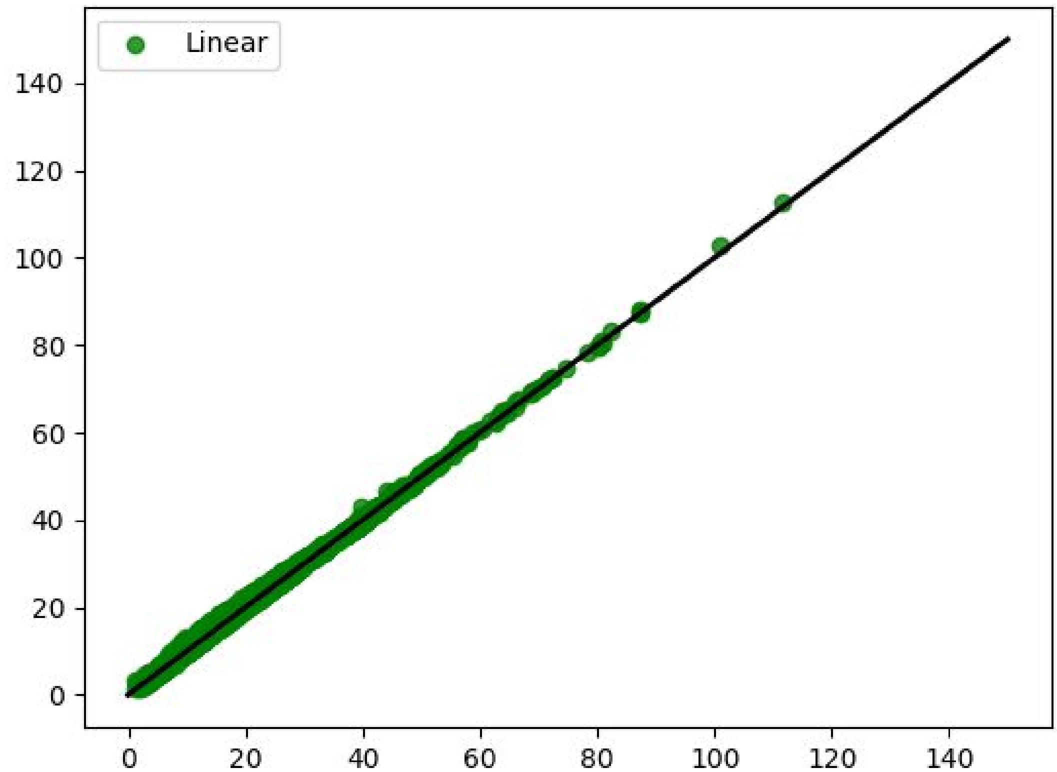
2. Tuning hyper-parameters:
   a. Type: GD; Number of iterations= 10; Alpha= 0.01; and λ(lambda)= 0.05;



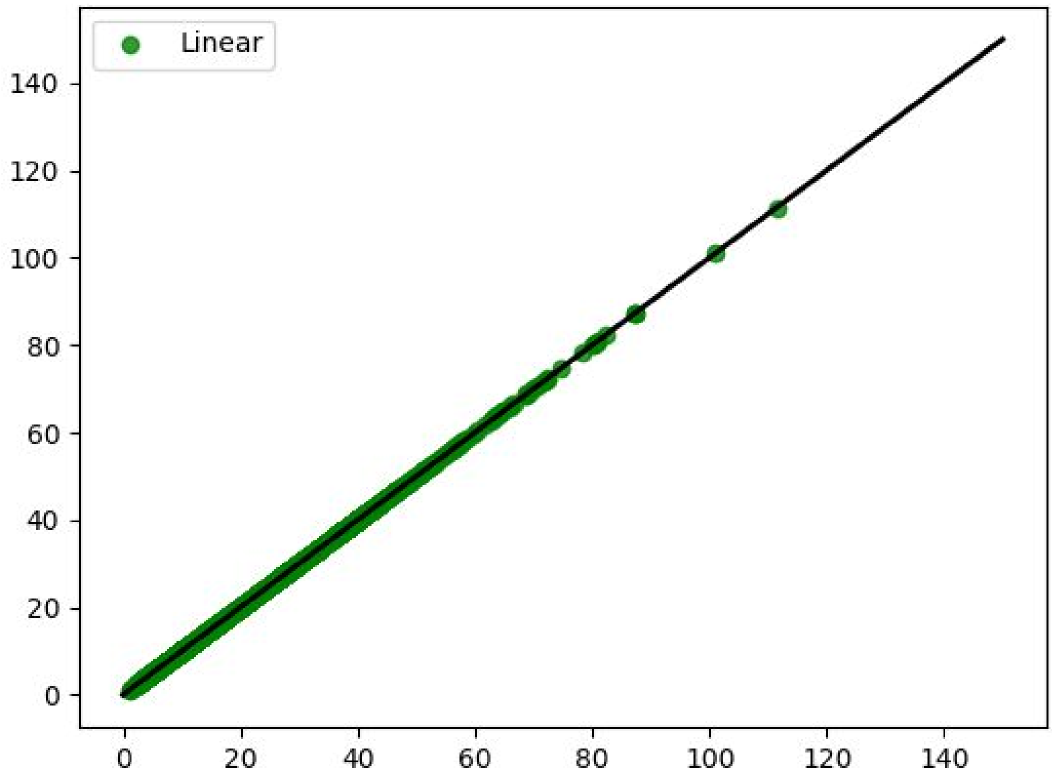
   b. Type: GD; Number of iterations= 100; Alpha= 0.01; and λ(lambda)= 0.05;



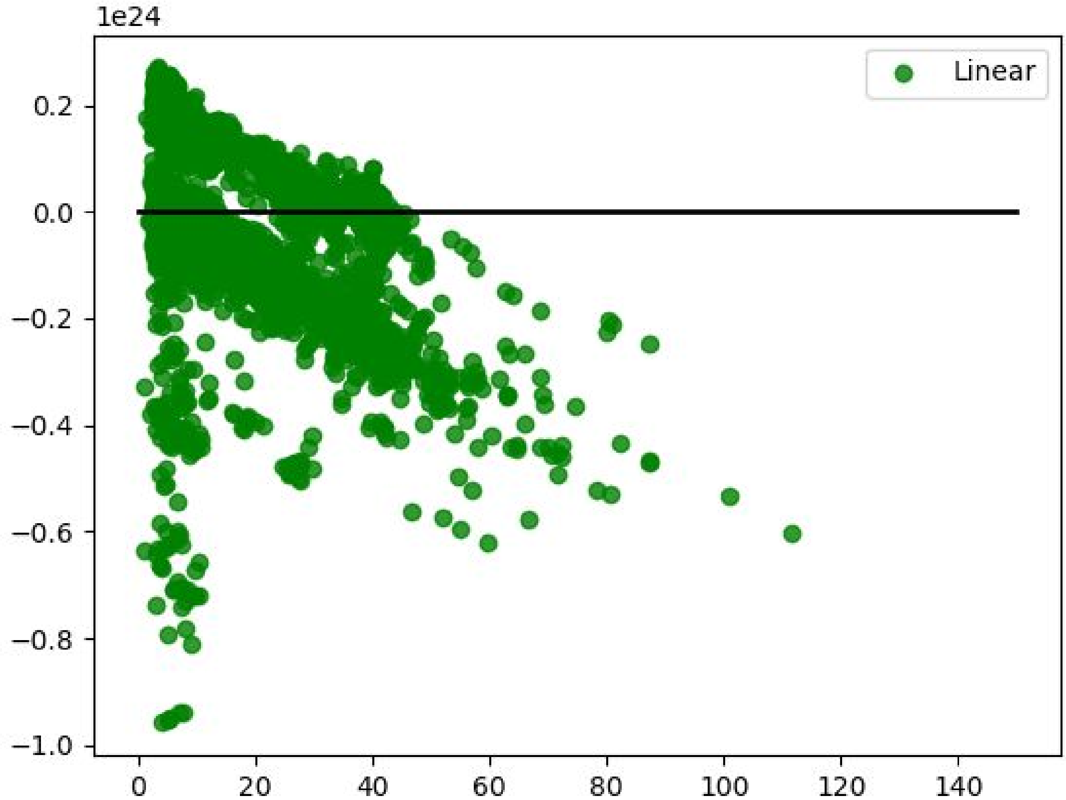
c. Type: GD; Number of iterations= 10000; Alpha= 0.01; and λ(lambda)= 0.05;



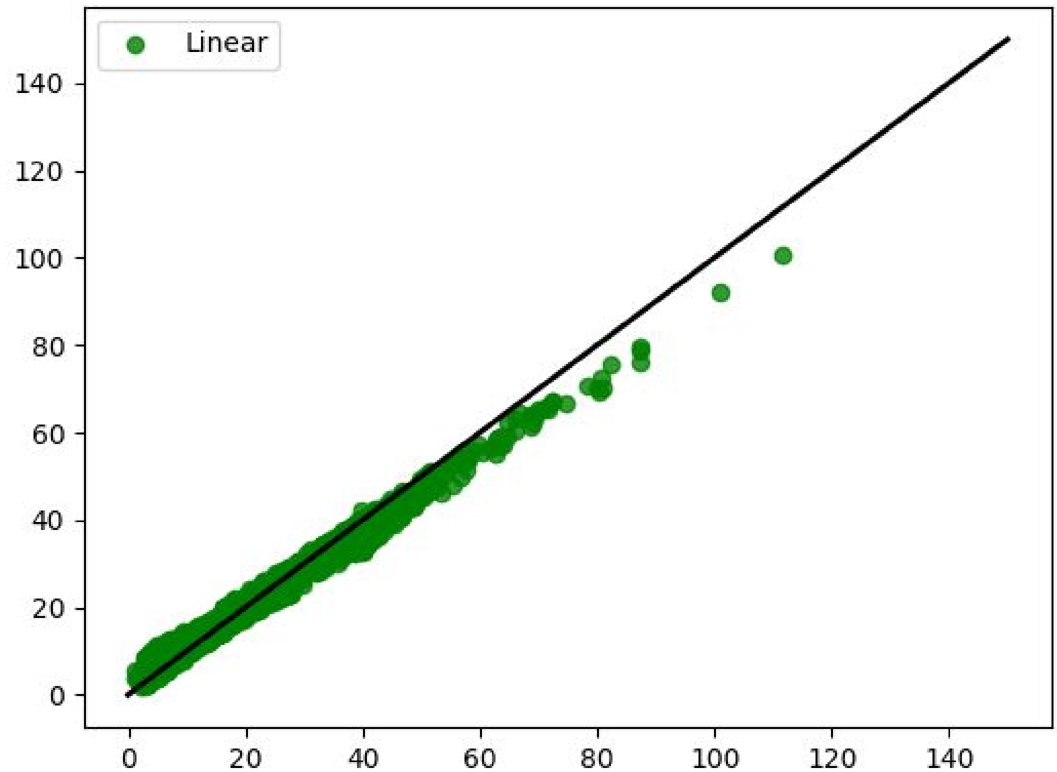
d. Type: GD; Number of iterations= 1000000; Alpha= 0.01; and λ(lambda)= 0.05;



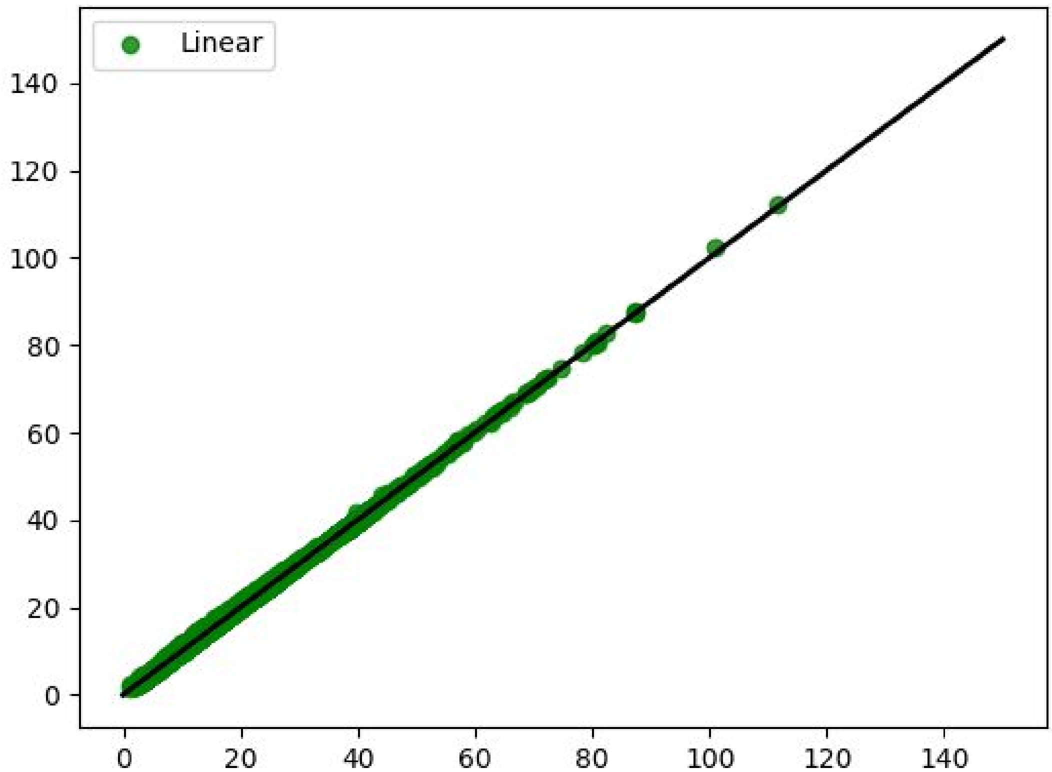
e. Type: SGD; Number of iterations= 10; Alpha= 0.01; and λ(lambda)= 0.05;



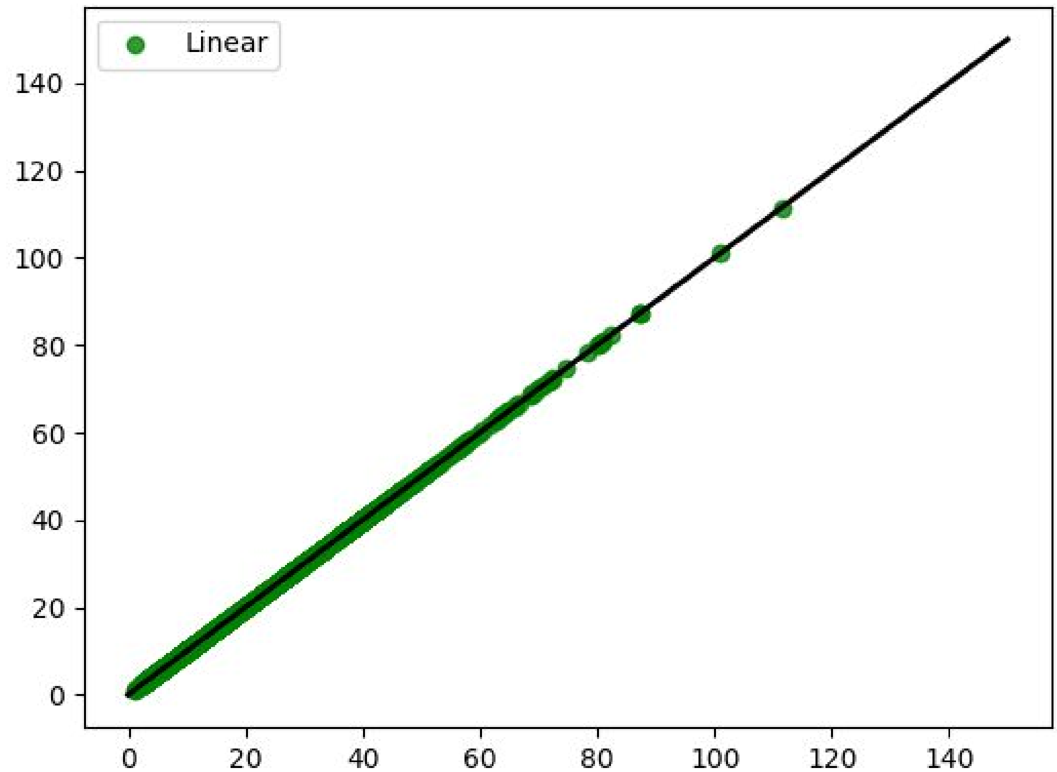
f. Type: SGD; Number of iterations= 10; Alpha= 0.0001; and λ(lambda)= 0.05;



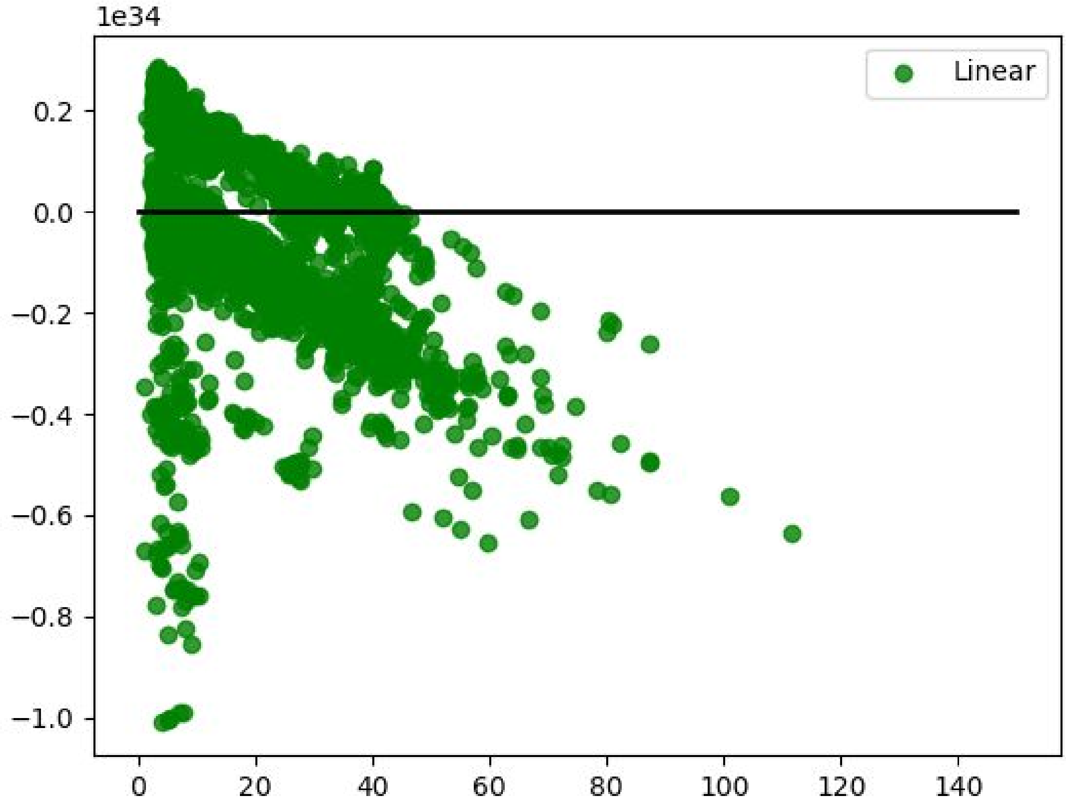
g. Type: SGD; Number of iterations= 10000; Alpha= 0.0001; and λ(lambda)= 0.05;



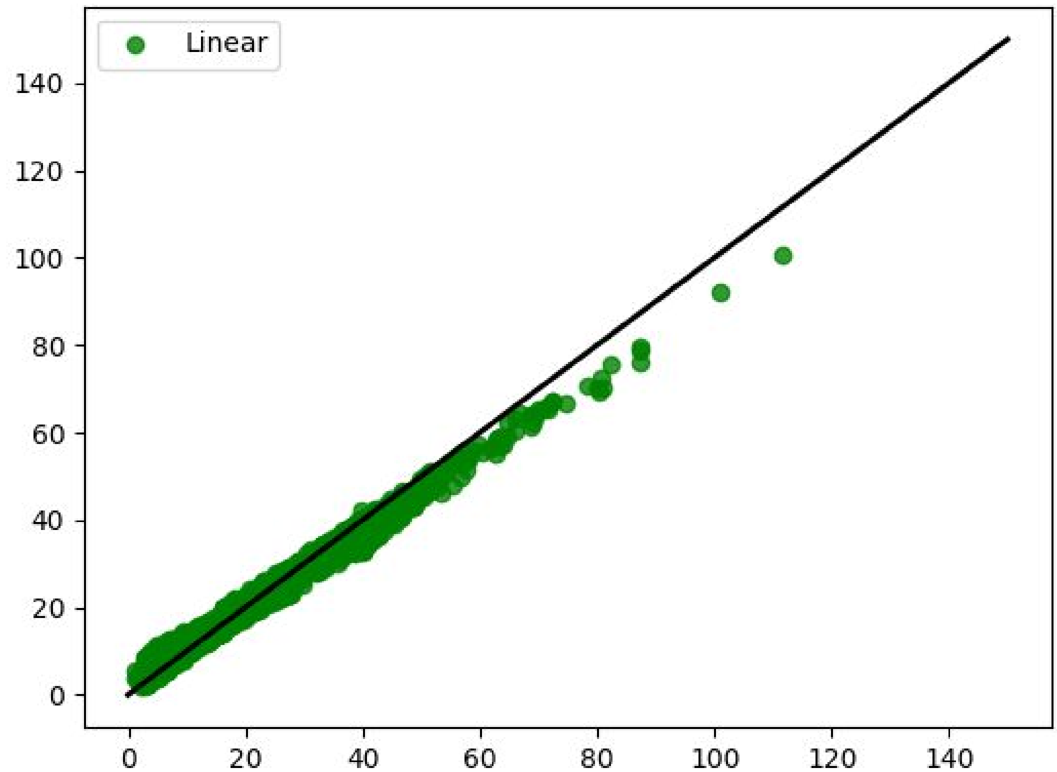
h. Type: SGD; Number of iterations= 1000000; Alpha= 0.0001; and λ(lambda)= 0.05;



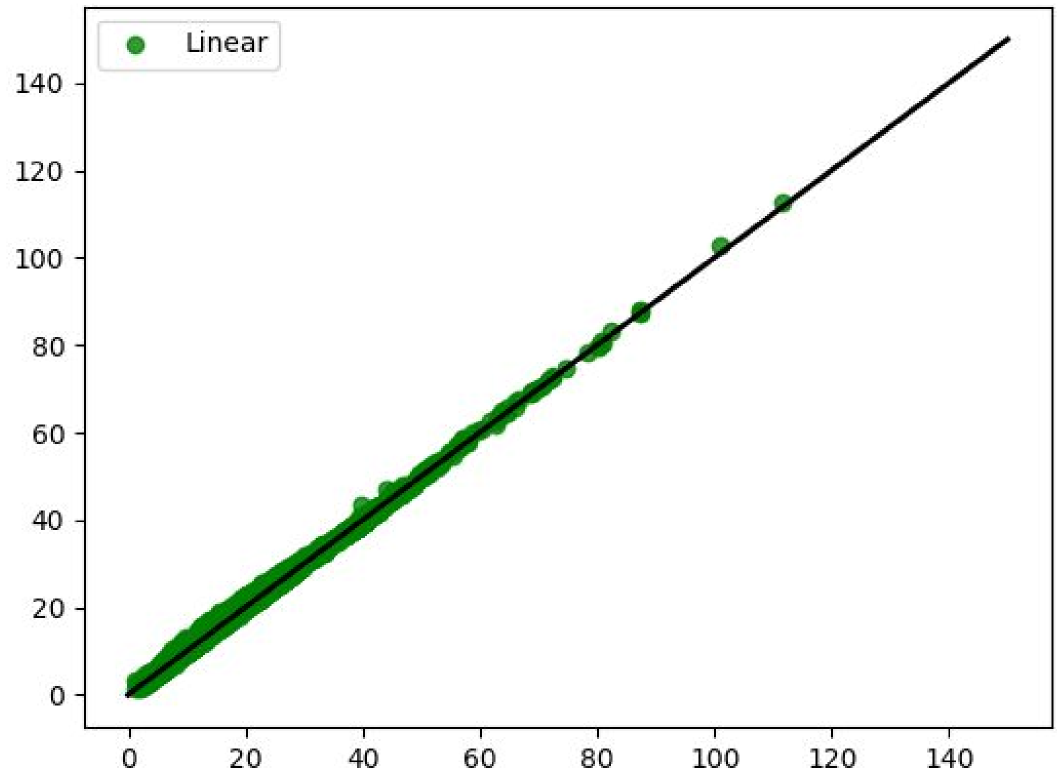
i. Type: SGD; Number of iterations= 10; Alpha= 0.1; and λ(lambda)= 0.1;



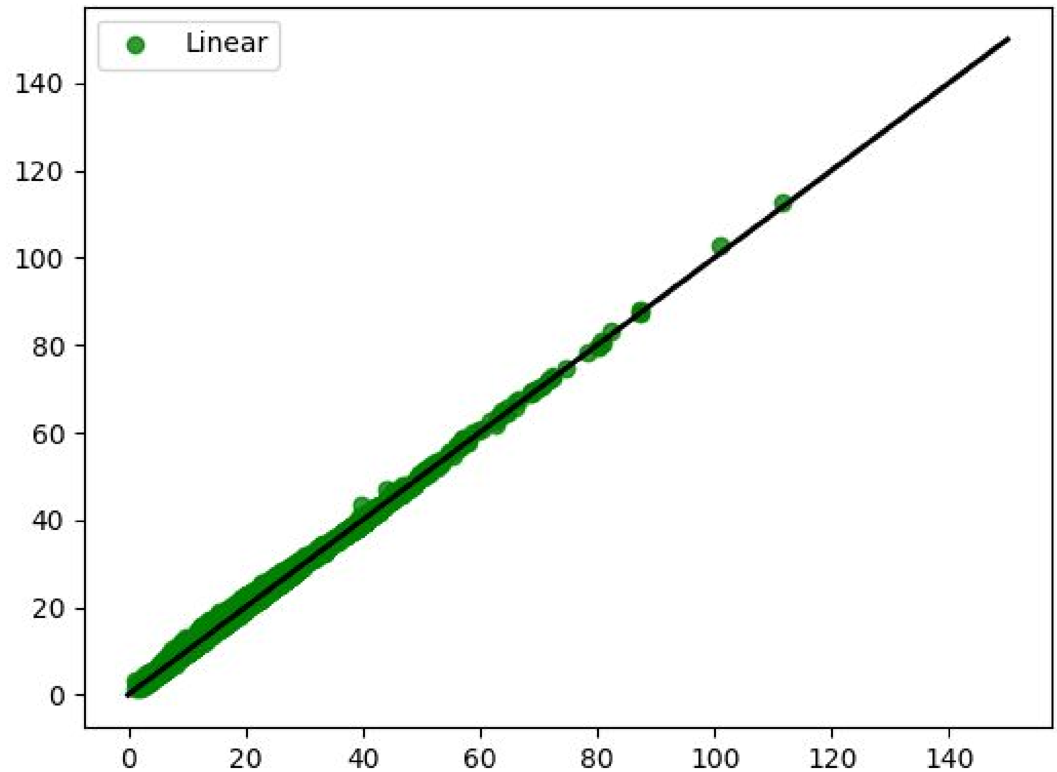
j. Type: SGD; Number of iterations= 10; Alpha= 0.0001; and λ(lambda)= 0.1;



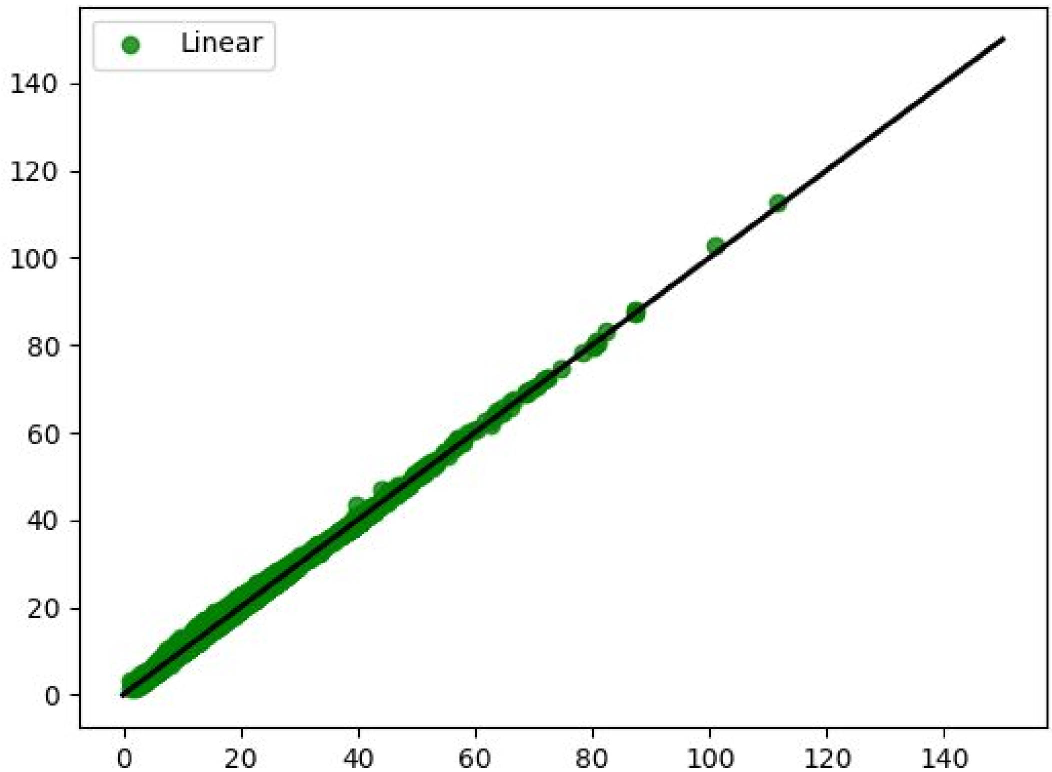
k. Type: SGD; Number of iterations= 100; Alpha= 0.0001; and λ(lambda)= 0.1;



I. Type: SGD; Number of iterations= 100; Alpha= 0.0001; and λ(lambda)= 0.5;



m. Type: SGD; Number of iterations= 100; Alpha= 0.0001; and λ(lambda)= 0.9;



3.

Type	Samples (N)	Number of Iterations	λ(lambda)	Alpha	Convergence time	Memory management
Gd	5000	1000000	0.05	0.01	Slower	NA
SGD	5000	100	0.9	0.00001	Faster	NA