CMPUT 466 A2

Yuhan Ye 1463504

Q1.

(a) $\theta_{MAP} = \underset{\theta}{\operatorname{argmax}} \left(p(D|\theta)\, p(\theta) \right)$

$$p(x_i|\theta) = \frac{1}{\sqrt{2\pi\sigma_0^2}} e^{-\frac{(x_i-\theta)^2}{2\sigma_0^2}}$$

$$p(\theta) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(\theta-\mu)^2}{2\sigma^2}}$$

$$p(D|\theta) = \prod_{i=1}^{n} p(x_i|\theta) = \frac{1}{(2\pi)^{\frac{n}{2}}\sigma_0^n} \cdot e^{-\frac{\sum_{i=1}^{n}(x_i-\theta)^2}{2\sigma_0^2}}$$

$$p(D|\theta)\,p(\theta) = \frac{1}{(2\pi)^{\frac{n+1}{2}}\sigma_0^n\sigma} \cdot e^{-\frac{\sum_{i=1}^{n}(x_i-\theta)^2}{2\sigma_0^2} - \frac{(\theta-\mu)^2}{2\sigma^2}}$$

$$l(\theta) = \log\left(p(D|\theta)\,p(\theta)\right) = \log\left(\frac{1}{(2\pi)^{\frac{n+1}{2}}\sigma_0^n\sigma}\right) - \frac{\sum_{i=1}^{n}(x_i-\theta)^2}{2\sigma_0^2} - \frac{(\theta-\mu)^2}{2\sigma^2}$$

$$\frac{\partial l(\theta)}{\partial \theta} = \frac{\sum_{i=1}^{n}(x_i-\theta)}{\sigma_0^2} - \frac{\theta-\mu}{\sigma^2} = 0$$

$$\frac{\left(\sum_{i=1}^{n} x_i - n\theta\right)\cdot\sigma^2 - \sigma_0^2(\theta-\mu)}{\sigma_0^2\sigma^2} = 0$$

$$\theta\cdot(-n\sigma^2-\sigma_0^2) + \sum_{i=1}^{n} x_i\sigma^2 + \sigma_0^2\mu = 0$$

$$\theta_{MAP} = \frac{\sum_{i=1}^{n} x_i\sigma^2 + \sigma_0^2\mu}{n\sigma^2+\sigma_0^2}$$

(b) $p(D|\theta) = \frac{1}{(2\pi)^{\frac{n}{2}}\sigma_0^n} e^{\frac{-\sum_{i=1}^{n}(x_i-\theta)^2}{2\sigma_0^2}}$

$p(\theta) = \frac{1}{2b} e^{\frac{-|\theta|}{b}}$

$p(D|\theta)p(\theta) = \frac{1}{(2\pi)^{\frac{n}{2}}\sigma_0^n \cdot 2b} e^{\frac{-\sum_{i=1}^{n}(x_i-\theta)^2}{2\sigma_0^2} - \frac{|\theta|}{b}}$

$l(\theta) = \log(p(D|\theta)p(\theta)) = \log\left(\frac{1}{(2\pi)^{\frac{n}{2}}\sigma_0^n \cdot 2b}\right) - \frac{\sum_{i=1}^{n}(x_i-\theta)^2}{2\sigma_0^2} - \frac{|\theta|}{b}$

$\frac{dl(\theta)}{d\theta} = \frac{\sum_{i=1}^{n}(x_i-\theta)}{\sigma_0^2} - \frac{|\theta|}{b\theta} = 0$

$$\theta = \begin{cases} \frac{\sum_{i=1}^{n} x_i}{n} - \frac{\sigma_0^2}{bn}, & \theta > 0 \\ \frac{\sum_{i=1}^{n} x_i}{n} + \frac{\sigma_0^2}{bn}, & \theta < 0 \end{cases} \quad \text{or} \quad \theta = \frac{\sum_{i=1}^{n} x_i}{n} - \frac{\sigma_0^2|\theta|}{bn\theta}$$

We do not have a closed form for θ_{MAP}

Our objective is continuous but not smooth, it is not differentiable at $\theta = 0$, making whole expression non-differentiable.

We can use proximal methods. Use gradient descent for the smooth component and for θ that are close to 0, set them to 0.

with step size η. $\theta = \theta_i + \frac{\sum_{i=1}^{n} x_i}{n}$

$$\text{Prox}_\eta(\theta_i) = \begin{cases} \theta_i - \eta\frac{\sigma_0^2}{bn}, & \theta_i > \eta\frac{\sigma_0^2}{bn} \\ 0, & |\theta_i| \le \eta\frac{\sigma_0^2}{bn} \\ \theta_i + \eta\frac{\sigma_0^2}{bn}, & \theta_i < -\eta\frac{\sigma_0^2}{bn} \end{cases}$$

(c) $p(x_i|\theta) = \frac{1}{(2\pi)^{\frac{d}{2}}} e^{-\frac{1}{2}(x_i-\theta)^T \cdot I \cdot (x_i-\theta)}$

$p(\theta) = \frac{1}{(2\pi)^{\frac{d}{2}}\sigma^2} e^{-\frac{1}{2}\theta^T \Sigma^{-1}\theta}$ $\quad \Sigma^{-1} = \sigma^{-2} I$

$p(D|\theta) = \frac{1}{(2\pi)^{\frac{nd}{2}}} e^{-\frac{1}{2}\sum_{i=1}^{n}[(x_i-\theta)^T \cdot I \cdot (x_i-\theta)]}$

$l(\theta) = \log(p(D|\theta)p(\theta)) = \log\left(\frac{1}{(2\pi)^{\frac{d(n+1)}{2}}\sigma^2}\right) - \frac{1}{2}\sum_{i=1}^{n}[(x_i-\theta)^T \cdot I \cdot (x_i-\theta)]$

$- \frac{1}{2}\theta^T \cdot \sigma^{-2} \cdot I \cdot \theta$

$\frac{dl(\theta)}{d(\theta)} = \sum_{i=1}^{n}(x_i-\theta) - \sigma^{-2}\theta = 0$

$\sum_{i=1}^{n} x_i - n\theta - \sigma^{-2}\theta = 0$

$\theta_{MAP} = \frac{\sum_{i=1}^{n} x_i}{n+\sigma^{-2}}$

(0-70)

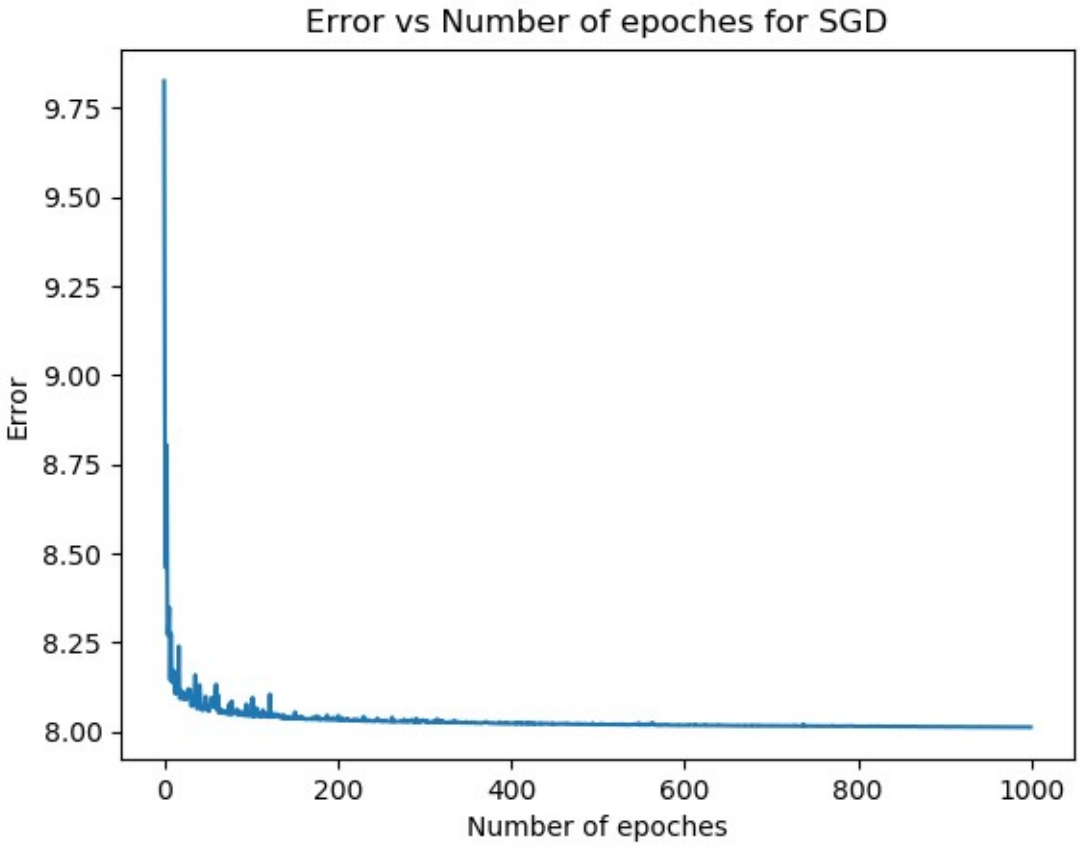
2. (a) Increasing number of features to more than 69 raise numpy.linalg.LinAlgError: Singular matrix
It is because that X is at low-rank, X^TX is not invertible. To solve this, we can use psendo-inverse numpy.linalg.pinv

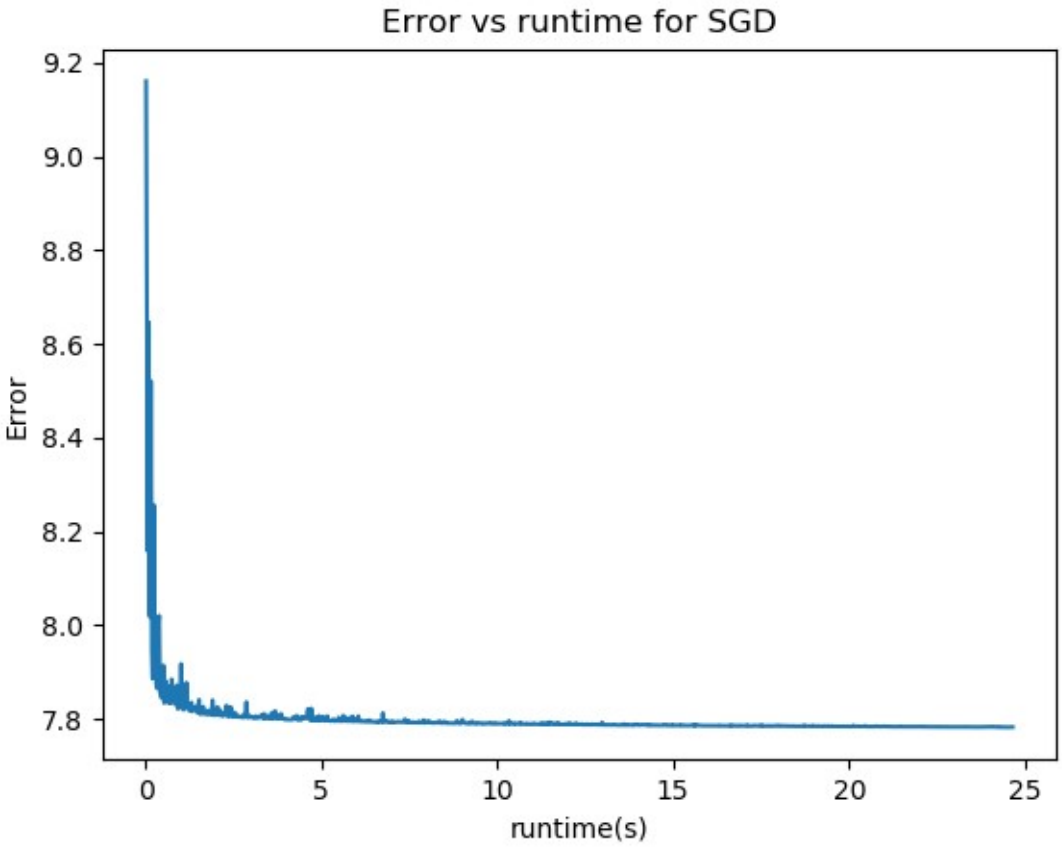
(c) Training error of RidgeRegression is larger than that of FSLRegression. While Test error of Ridge Regression is smaller. It indicates that FSLRegression is 'overfit', while ridge balances value of weights (penalize big values in w) and performs better on test set.

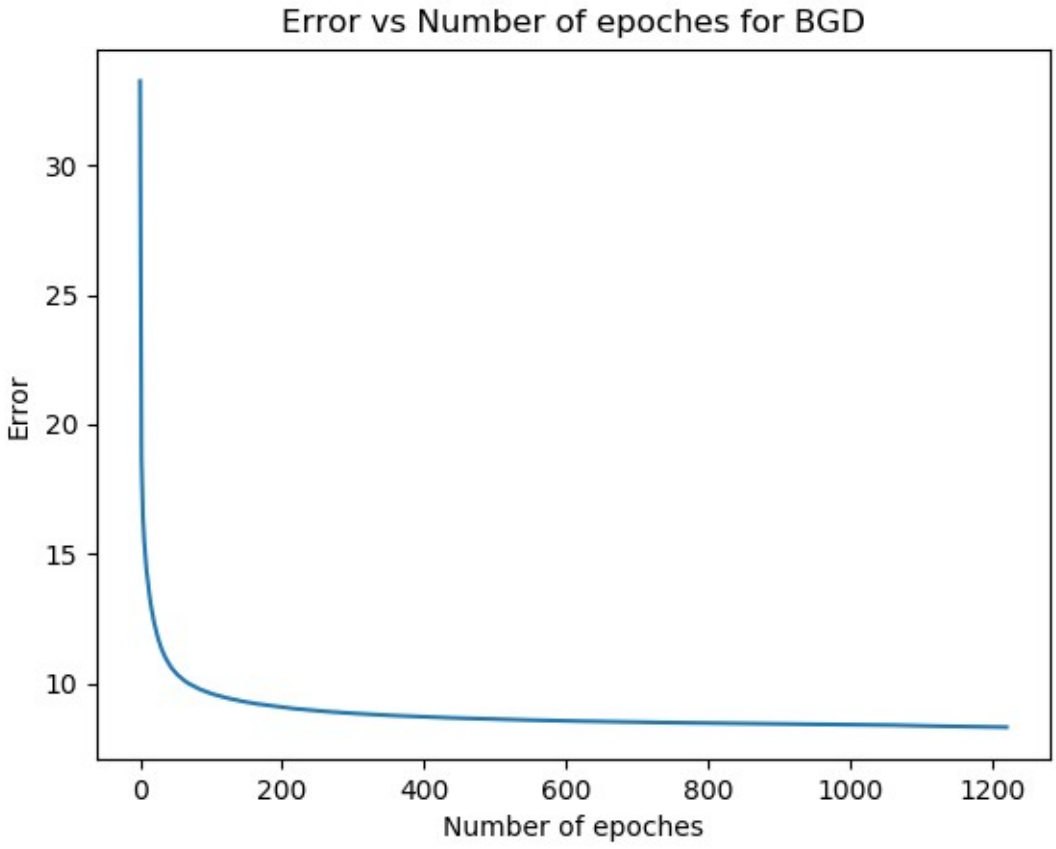
(f) SGD is faster to approach local minima, that weights is converged although it is 'noisy' at the beginning.
BGD plot is smooth, since it decreases in a right direction each epoch. but takes longer time.

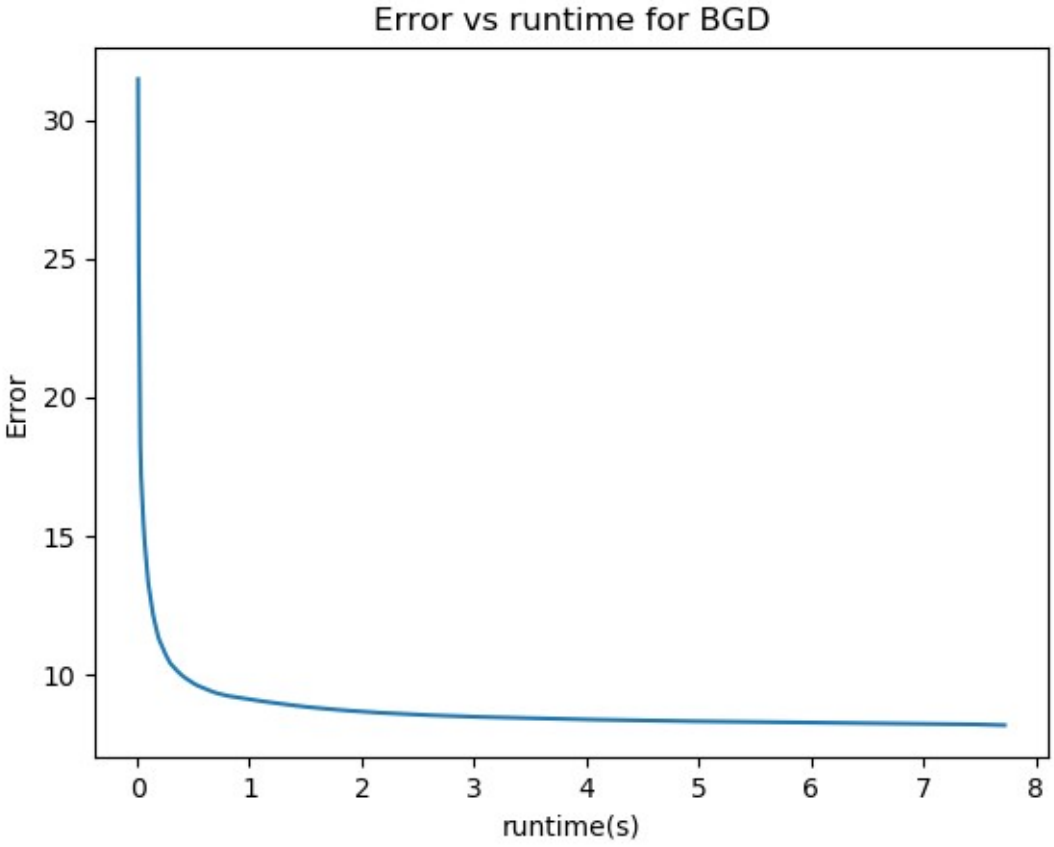
Bonus

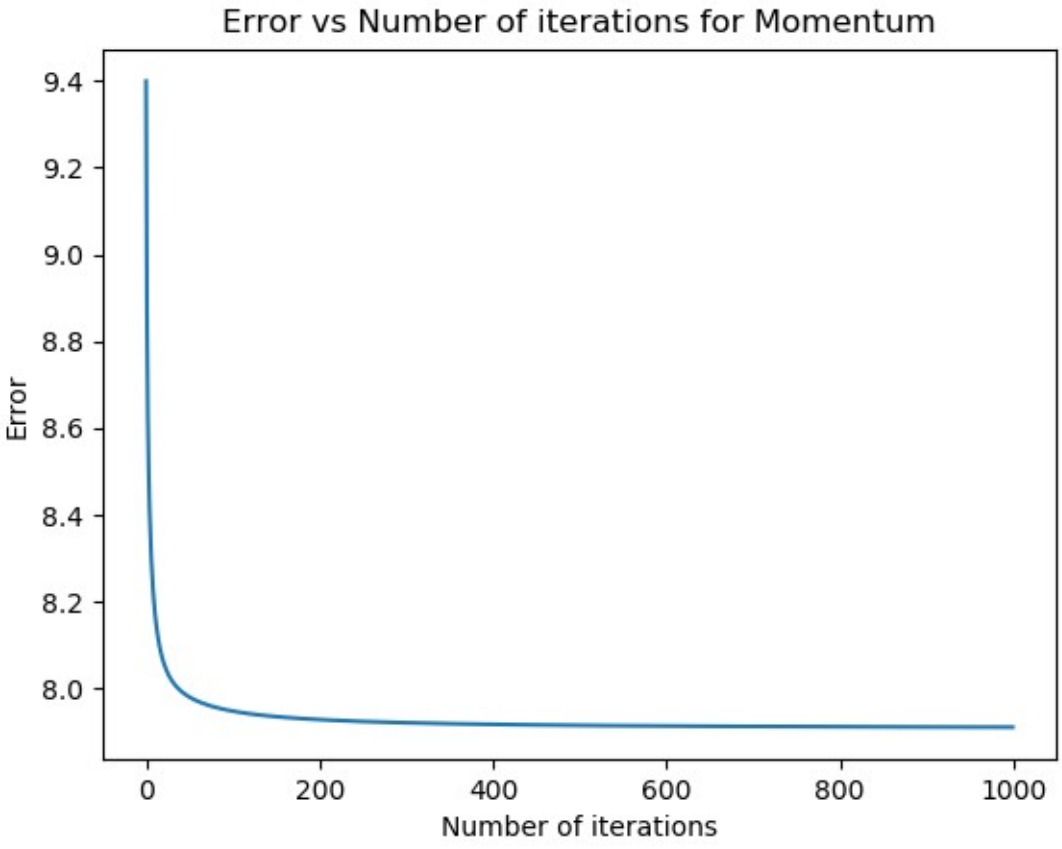
cc) Note that in the initial phase of iterations Momentum has an offset to the initial value ($m_0 = 0$, ~~$v_0 = 0$~~ $v_0 = 0$, so there are bias toward 0) Thus Momentum without bias correction is actually a biased estimate of the expected gradient

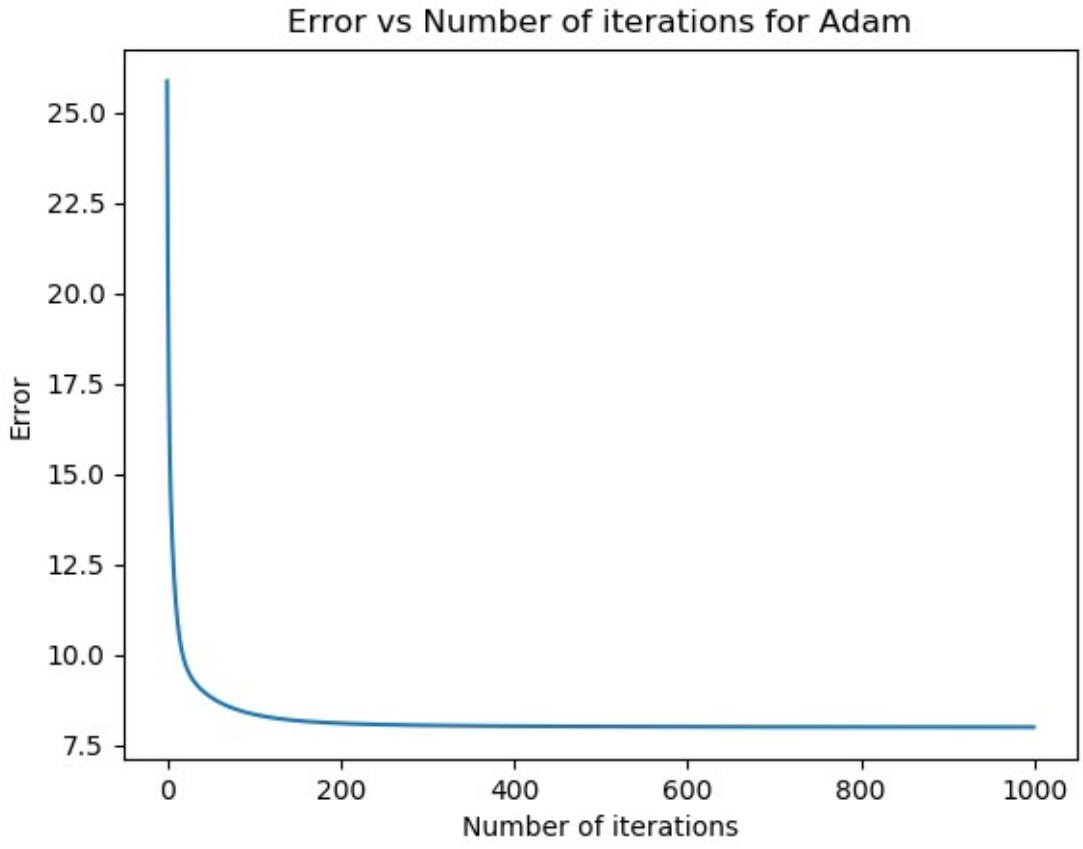

Error vs Number of epoches for SGD
9.75
9.50
9.25
9.00
8.75
8.50
8.25
8.00
Error
0
200
400
600
800
1000
Number of epoches


Error vs runtime for SGD
9.2
9.0
8.8
8.6
8.4
8.2
8.0
7.8
Error
0
5
10
15
20
25
runtime(s)


Error vs Number of epoches for BGD
Error
30
25
20
15
10
0
200
400
600
800
1000
1200
Number of epoches


Error vs runtime for BGD
Error
runtime(s)
30
25
20
15
10
0
1
2
3
4
5
6
7
8


Error vs Number of iterations for Momentum
9.4
9.2
9.0
8.8
8.6
8.4
8.2
8.0
Error
0
200
400
600
800
1000
Number of iterations


Error vs Number of iterations for Adam
25.0
22.5
20.0
17.5
15.0
12.5
10.0
7.5
Error
0
200
400
600
800
1000
Number of iterations