Q3

1): $\mathbb{E}_I(\frac{1}{m}\sum_{i\in I} a_i)$

$= \frac{1}{m}\mathbb{E}(a_{i1} + a_{i2} + \cdots a_{im})$

$= \frac{1}{m}\left[\mathbb{E}(a_{i1}) + \cdots + \mathbb{E}(a_{im})\right]$

$= \frac{1}{m}\left[\sum_{i=1}^{n} a_i(\frac{1}{n}) + \cdots + \sum_{i=1}^{n} a_i(\frac{1}{n})\right]$

$= \frac{1}{n}\sum_{i=1}^{n} a_i$

2): $\mathbb{E}_I(\nabla L_I(x,y,\theta))$

$= \mathbb{E}_I\left[\frac{1}{m}\nabla\sum_{i\in I} L(x,y,\theta))\right]$

$= \frac{1}{n}\sum_{i=1}^{n}\nabla L(x^{(i)}, y^{(i)}, \theta)$ (by (1))

$= \nabla\left[\frac{1}{n}\sum_{i=1}^{n} L(x^{(i)}, y^{(i)}, \theta)\right]$

$= \nabla L(x,y,\theta)$

3): Mini-batch method produces an unbiased estimator of true gradient

4): a): $L(x,y,w) = (y - w^T x)^2$

Assume $w: d\times 1$; $x: d\times N$; $y: 1\times N$

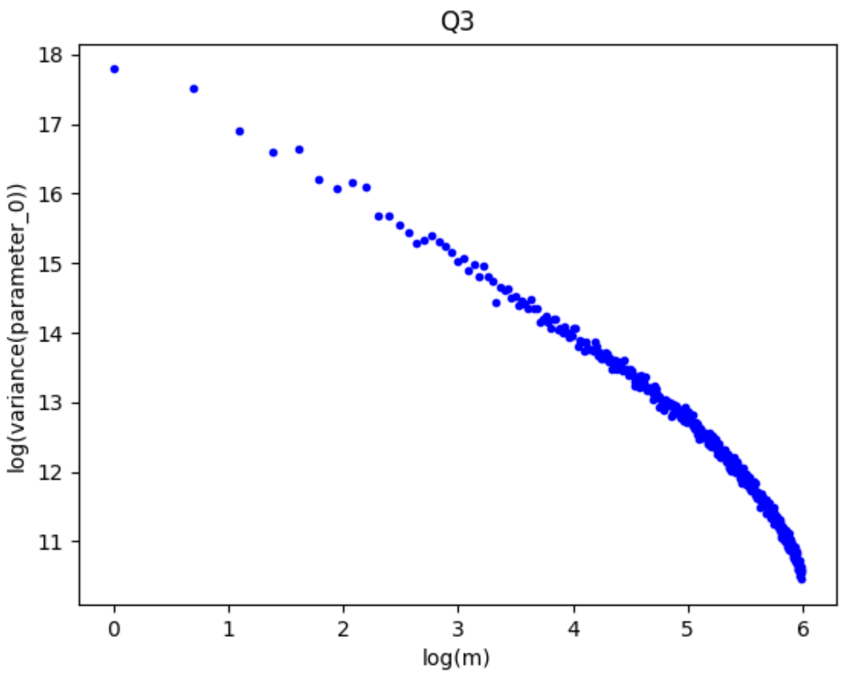
$\nabla_w L(x,y,w)$

$= \nabla_w (yy^T + w^T x x^T w - 2w^T x y^T)$

$= 2xx^T w - 2xy^T$

5): cos_similarity is more meanful

Since gradient computed has large values. Even with cos_similarity over 99%, square distance can still be quite large b): plot:



Q3
log(variance(parameter_0))
log(m)
18
17
16
15
14
13
12
11
0
1
2
3
4
5
6