Optimizers

Gradient Descent

- Gradient Descent methods are used to find the minimum value of parameters to minimize the cost function.
- There are different methods to find the minimum values of parameters:

1. Stochastic Gradient Descent
2. Ada-grad Optimizer
3. RMS Prop Optimizer
4. Adam Optimizer

Stochastic Gradient Descent

$$b_{n+1} = b_n - \eta \frac{\partial L}{\partial b}$$

Updated weight

Initial weight

Learning rate

Derivative of Loss function

- Initial weight is chosen as a random number.
- Learning rate for data is constant which is initialized as 0.01.
- Loss function is the Sum Squared error of the data.

Momentum in SGD

Momentum is introduced in Stochastic gradient Descent to accelerate the values of parameters so that it can converge the values faster.

Beta value

$$\boxed{D_{n+1}(b) = \beta D_{avg(n)}(b) + (1-\beta) D_n(b)}$$

Updated derivative

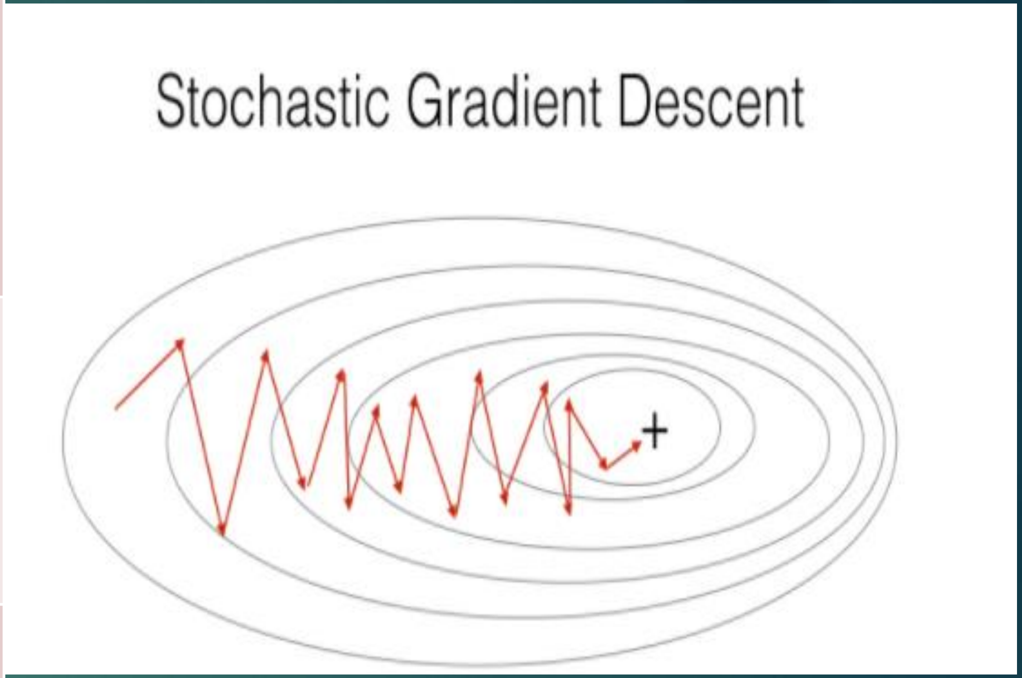
Average of initial derivates

Current derivative

Updated Weight formula

$$b_{n+1} = b_n - \eta D_{n+1}(b)$$

Advantages	Disadvantages
It is easier to fit into memory due to a single training sample being processed by the network	Due to frequent updates the steps taken towards the minima are very noisy. This can often lead the gradient descent into other directions.
It is computationally fast as only one sample is processed at a time	Also, due to noisy steps it may take longer to achieve convergence to the minima of the loss function
For larger datasets it can converge faster as it causes updates to the parameters more frequently	Frequent updates are computationally expensive due to using all resources for processing one training sample at a time



Ada-grad Optimizer

$$b_{n+1} = b_n - \frac{\eta}{\sqrt{\alpha_n + \varepsilon}} \frac{\partial L}{\partial b}$$

Notations	Description
b_{n+1}	Updated weight
b_n	Initial weight
η	Initialized leaning rate
α_n	Sum of squares of all past gradients
ε	Epsilon is small positive value which avoids division by zero

Coefficient a = 0.45

Coefficient b = 0.75

Normalization

$$X = \frac{x - x_{\min}}{x_{\max} - x_{\min}}$$

SqFt	X
1100	0.00
1400	0.22
1425	0.58
1550	0.20
1600	0.55
1700	0.39
1700	0.54
1875e	0.53
2350	1.00
2450	0.61
Min = 1100	Min = 0
Max = 2450	Max = 1

Normalization

$$X = \frac{x - x_{\min}}{x_{\max} - x_{\min}}$$

Prices	y
199000	0.00
245000	0.22
319000	0.24
240000	0.33
312000	0.37
279000	0.44
310000	0.44
308000	0.57
405000	0.93
324000	1.00
Min = 199000	Min = 0
Max = 405000	Max = 1

Coefficient a = 0.45

+

Coefficient b = 0.75

X

X
0.00
0.22
0.58
0.20
0.55
0.39
0.54
0.53
1.00
0.61

$$y_p = a + bX$$

Predicted y
0.45
0.62
0.63
0.70
0.73
0.78
0.78
0.88
1.14
1.20
Min = 0
Max = 1

y
0.00
0.22
0.24
0.33
0.37
0.44
0.44
0.57
0.93
1.00
Min = 0
Max = 1
Range = 1

-

Predicted y
0.45
0.62
0.63
0.70
0.73
0.78
0.78
0.88
1.14
1.20
Min = 0
Max = 1
Range = 1

$$\frac{d(SSE)}{da} = -(y - y_p)$$

d(SSE)/da
0.45
0.39
0.05
0.50
0.18
0.39
0.24
0.35
0.14
0.59
Total = 3.30

X	y	Predicted y
0.00	0.00	0.45
0.22	0.22	0.62
0.58	0.24	0.63
0.20	0.33	0.70
0.55	0.37	0.73
0.39	0.44	0.78
0.54	0.44	0.78
0.53	0.57	0.88
1.00	0.93	1.14
0.61	1.00	1.20
Min = 0	Min = 0	Min = 0
Max = 1	Max = 1	Max = 1

X * y - Predicted y

$$\frac{d(SSE)}{db} = -X(y - y_p)$$

d(SSE)/db
0.00
0.09
0.01
0.17
0.07
0.18
0.11
0.20
0.13
0.59
Total = 1.55

First Epoch

Step 1

$$\alpha_a = \left(\frac{d(SSE)}{da}\right)^2$$

$$\alpha_b = \left(\frac{d(SSE)}{db}\right)^2$$

$\frac{d(SSE)}{da}$=3.30

$\frac{d(SSE)}{da}$=1.55

$\alpha_a = (3.30)^2$=10.89

$\alpha_b = (1.55)^2$=2.4025

Step 2

$$a_{n+1} = a_n - \frac{\eta}{\sqrt{\alpha_a + \varepsilon}} \frac{d(SSE)}{da}$$

$$b_{n+1} = b_n - \frac{\eta}{\sqrt{\alpha_b + \varepsilon}} \frac{d(SSE)}{db}$$

$a_n = 0.45$

$\eta = 0.01$

$\varepsilon = 0.00000001$

$b_n = 0.75$

$a_{n+1} = 0.45 - \frac{0.01}{\sqrt{10.89+0.00000001}}$(3.30)
=0.45-(0.00303)(3.30)
=0.44

$b_{n+1} = 0.75 - \frac{0.01}{\sqrt{2.4025+0.00000001}}$(1.55)
=0.75-(0.00645)(1.55)
=0.74

Second Epoch

L3 =SUM(L14^2+M14^2)

	A	B	C	D	E	F	G	H	I	J	K	L	M
1		a = **0.44**		b= **0.74**				del SSE/del(a) =-(Y-YP)	del SSE/del(b) =-(Y-YP)X				
2		*Sq Ft*	*Price$*	*X*	*Y*	*YP*	*(1/2)SSE*						
3		1100	199000	0.00	0.00	0.44	0.0968	0.44	0.00		alpha(a)=	20.8125	
4		1400	245000	0.22	0.22	0.60	0.072635	0.38	0.08		eta'(a)=	0.002191986	
5		1425	319000	0.24	0.58	0.62	0.000635	0.04	0.01		a2=	0.433085146	
6		1550	240000	0.33	0.20	0.69	0.118895	0.49	0.16				
7		1600	312000	0.37	0.55	0.71	0.0137	0.17	0.06				
8		1700	279000	0.44	0.39	0.77	0.072405	0.38	0.17		alpha(b)=	4.56340000000000	
9		1700	310000	0.44	0.54	0.77	0.026462	0.23	0.10		eta'(b)=	0.004681184	
10		1875	308000	0.57	0.53	0.86	0.056343	0.34	0.19		b2=	0.733117119	
11		2350	405000	0.93	1.00	1.13	0.007836	0.13	0.12				
12		2450	324000	1.00	0.61	1.18	0.164281	0.57	0.57				
13	MIN	1100	199000	0	0	**Total SSE=**	**0.629993**	**3.15**	**1.47**			**1st**	**2nd**
14	MAX	2450	405000	1	1						D(a) values	3.3	3.15
15	RANGE	1350	206000	1	1						D(b) values	1.55	1.47
16													
17													

Third Epoch

SUM | =SUM(L14^2,M14^2,N14^2)

	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1		a = **0.433085**		b= **0.733117**				del SSE/del(a) =-(Y-YP)	del SSE/del(b) =-(Y-YP)X					
2		*Sq Ft*	*Price$*	*X*	*Y*	*YP*	*(1/2)SSE*							
3		1100	199000	0.00	0.00	0.43	0.093781	0.43	0.00					
4		1400	245000	0.22	0.22	0.60	0.069452	0.37	0.08		alpha(a)=	M14^2,N14^2)		
5		1425	319000	0.24	0.58	0.61	0.000366	0.03	0.01		eta'(a)=	0.001822253		
6		1550	240000	0.33	0.20	0.68	0.114447	0.48	0.16		a3=	0.4275198		
7		1600	312000	0.37	0.55	0.70	0.012178	0.16	0.06					
8		1700	279000	0.44	0.39	0.76	0.068659	0.37	0.16		alpha(b)=	5.64363081742593		
9		1700	310000	0.44	0.54	0.76	0.024218	0.22	0.10		eta'(b)=	0.004209405		
10		1875	308000	0.57	0.53	0.85	0.052755	0.32	0.19		b3=	0.727145619		
11		2350	405000	0.93	1.00	1.11	0.006261	0.11	0.10					
12		2450	324000	1.00	0.61	1.17	0.156468	0.56	0.56					
13	MIN	1100	199000	0	0	**Total SSE=**	**0.598585**	**3.05**	**1.42**			**1st**	**2nd**	**3rd**
14	MAX	2450	405000	1	1						D(a) values	3.3	3.15	3.05
15	RANGE	1350	206000	1	1						D(b) values	1.55	1.47	1.42
16														
17														
18														

Disadvantage of Ada-grad

- As the learning rate is decreasing drastically so in deep neural networks, at a particular time the weight up-dation will be so small that it will stop moving towards global minima.

RMS Prop Optimizer

$$w_{n+1} = \gamma w_n + (1-\gamma) D_n^2(b)$$

Notations	Description
w_{n+1}	Moving weighted average
γ	Moving average parameter
w_n	Initial weighted average
$D_n(b)$	Derivative of loss function

Updated Weight

$$b_{n+1} = b_n - \frac{\eta}{\sqrt{w_{n+1}} + \varepsilon} \frac{\partial L}{\partial b}$$

First Epoch

Step 1

$$w_{a(n+1)} = \gamma w_{a(n)} + (1-\gamma)\left(\frac{d(SSE)}{da}\right)^2$$

$$w_{b(n+1)} = \gamma w_{b(n)} + (1-\gamma)\left(\frac{d(SSE)}{db}\right)^2$$

$$\boxed{\begin{aligned} &\frac{d(SSE)}{da} = 3.30 \\ &w_{a(n)} = 0 \\ &\gamma = 0.9 \\ &\frac{d(SSE)}{da} = 1.55 \\ &w_{b(n)} = 0 \end{aligned}}$$

$$\begin{aligned} w_{n+1} &= (0.9)(0) + (1-0.9)(3.30)^2 \\ &= 0 + 1.089 \\ &= 1.089 \end{aligned}$$

$$\begin{aligned} w_{b(n+1)} &= (0.9)(0) + (1-0.9)(1.55)^2 \\ &= 0 + 0.240 \\ &= 0.240 \end{aligned}$$

Step 2

$$a_{n+1} = a_n - \frac{\eta}{\sqrt{w_{a(n+1)} + \varepsilon}} \frac{d(SSE)}{da}$$

$$b_{n+1} = b_n - \frac{\eta}{\sqrt{w_{b(n+1)} + \varepsilon}} \frac{d(SSE)}{db}$$

$$\boxed{\begin{aligned} &a_n = 0.45 \\ &\eta = 0.01 \\ &\varepsilon = 0.00000001 \\ &b_n = 0.75 \end{aligned}}$$

$$\begin{aligned} a_{n+1} &= 0.45 - \frac{0.01}{\sqrt{1.089 + 0.00000001}}(3.30) \\ &= 0.45 - 0.03162 \\ &= 0.41838 \end{aligned}$$

$$\begin{aligned} b_{n+1} &= 0.75 - \frac{0.01}{\sqrt{0.240 + 0.00000001}}(1.55) \\ &= 0.75 - 0.03164 \\ &= 0.72 \end{aligned}$$

Second Epoch

I15 | =M10*M11+(1-M10)*(H13)^2

	A	B	C	D	E	F	G	H	I	J	K	L	M
1		a = **0.418377**		b= **0.72**				del SSE/del(a) =-(Y-YP)	del SSE/del(b) =-(Y-YP)X				
2		*Sq Ft*	*Price$*	*X*	*Y*	*YP*	*(1/2)SSE*						
3		1100	199000	0.00	0.00	0.42	0.08752	0.42	0.00				
4		1400	245000	0.22	0.22	0.58	0.06304	0.36	0.08				
5		1425	319000	0.24	0.58	0.59	4.22E-05	0.01	0.00				
6		1550	240000	0.33	0.20	0.66	0.1055	0.46	0.15				
7		1600	312000	0.37	0.55	0.69	0.009316	0.14	0.05				
8		1700	279000	0.44	0.39	0.74	0.06126	0.35	0.16				
9		1700	310000	0.44	0.54	0.74	0.019909	0.20	0.09				
10		1875	308000	0.57	0.53	0.83	0.045779	0.30	0.17		gamma=		0.9
11		2350	405000	0.93	1.00	1.09	0.003616	0.09	0.08		initial weight_a=		**1.089106798735**
12		2450	324000	1.00	0.61	1.14	0.141289	0.53	0.53		initial weight_b=		0.23783001
13	MIN	1100	199000	0	0	**Total SSE=**	**0.53727**	**2.85**	**1.31**				
14	MAX	2450	405000	1	1								
15	RANGE	1350	206000	1	1			w_avg(a)	**1.790889275389**		w_avg(b)=	0.386463	
16								eta'(a)=	**0.007472495**		eta'(b)=	0.016086	
17								a2=	0.397101033		b2=	0.70	
18													

Third Epoch

I16 | =0.01/SQRT(I15+0.00000001)

	A	B	C	D	E	F	G	H	I	J	K	L	M
1		a = **0.397101**		b= **0.7**				del SSE/del(a) =-(Y-YP)	del SSE/del(b) =-(Y-YP)X				
2		*Sq Ft*	*Price$*	*X*	*Y*	*YP*	*(1/2)SSE*						
3		1100	199000	0.00	0.00	0.40	0.078845	0.40	0.00				
4		1400	245000	0.22	0.22	0.55	0.054238	0.33	0.07				
5		1425	319000	0.24	0.58	0.57	0.000143	-0.02	0.00				
6		1550	240000	0.33	0.20	0.63	0.093055	0.43	0.14				
7		1600	312000	0.37	0.55	0.66	0.005812	0.11	0.04				
8		1700	279000	0.44	0.39	0.71	0.051156	0.32	0.14				
9		1700	310000	0.44	0.54	0.71	0.014344	0.17	0.08				
10		1875	308000	0.57	0.53	0.80	0.036403	0.27	0.15		gamma=		0.9
11		2350	405000	0.93	1.00	1.05	0.001024	0.05	0.04		initial_weight_a=		1.790889
12		2450	324000	1.00	0.61	1.10	0.120199	0.49	0.49		initial_weight_b=		0.386463
13		1100	199000	0	0	**Total SSE=**	**0.455219**	**2.54**	**1.16**				
14	MAX	2450	405000	1	1								
15	RANGE	1350	206000	1	1			w_avg(a)	2.258685837197		w_avg(b)=	0.481773	
16								eta'(a)=	0.006653836		eta'(b)=	0.014407	
17								a3=	0.380177704		b3=	0.68	
18													

Adam Optimizer (Combination of Ada-grad and RMS Prop)

First moment vector

$$m_{n+1} = \beta_1 m_n + (1 - \beta_1) D_n(b)$$

First decay rate Initial moment vector Derivative of loss function

Second moment vector

$$v_{n+1} = \beta_2 v_n + (1 - \beta_2) D_n^2(b)$$

Second decay rate Initial moment vector Square of derivative of loss function

Step 1

$$m_{a(n+1)} = \beta_1 m_{a(n)} + (1-\beta_1)\frac{d(SSE)}{da}$$

$$m_{b(n+1)} = \beta_1 m_{b(n)} + (1-\beta_1)\frac{d(SSE)}{db}$$

Step 2

$$v_{a(n+1)} = \beta_2 v_{a(n)} + (1-\beta_2)\left(\frac{d(SSE)}{da}\right)^2$$

$$v_{b(n+1)} = \beta_2 v_{b(n)} + (1-\beta_2)\left(\frac{d(SSE)}{db}\right)^2$$

$$m_{a(n)} = 0$$

$$m_{b(n)} = 0$$

$$\frac{d(SSE)}{da} = 3.30$$

$$\frac{d(SSE)}{db} = 1.55$$

$$\beta_1 = 0.9$$

$$\beta_2 = 0.999$$

$$v_{a(n)} = 0$$

$$v_{b(n)} = 0$$

$$\begin{aligned} m_{a(n+1)} &= (0.9)(0) + (1-0.9)(3.30) \\ &= 0 + 0.33 \\ &= 0.33 \end{aligned}$$

$$\begin{aligned} m_{b(n+1)} &= (0.9)(0) + (1-0.9)(1.55) \\ &= 0 + 0.155 \\ &= 0.155 \end{aligned}$$

$$\begin{aligned} v_{a(n+1)} &= (0.999)(0) + (1-0.999)(3.30)^2 \\ &= 0 + 0.01089 \\ &= 0.01089 \end{aligned}$$

$$\begin{aligned} v_{b(n+1)} &= (0.999)(0) + (1-0.999)(1.55)^2 \\ &= 0 + 0.0024 \\ &= 0.0024 \end{aligned}$$

First moment bias correction

$$\boxed{\hat{m}_{n+1} = \frac{m_{n+1}}{1-\beta_1}}$$

Second moment bias correction

$$\boxed{\hat{v}_{n+1} = \frac{v_{n+1}}{1-\beta_2}}$$

Updated weight

$$\boxed{b_{n+1} = b_n - \eta \frac{\hat{m}_{n+1}}{\sqrt{\hat{v}_{n+1}} + \varepsilon}}$$

Note:

β_1 is initialized to 0.9

β_2 is initialized to 0.999

Step 3

$$\hat{m}_{a(n+1)} = \frac{m_{a(n+1)}}{1-\beta_1}$$

$$\hat{m}_{b(n+1)} = \frac{m_{b(n+1)}}{1-\beta_1}$$

$$\beta_1 = 0.9$$

$$m_{a(n+1)} = 0.33$$

$$m_{b(n+1)} = 0.155$$

$$\hat{m}_{a(n+1)} = \frac{0.33}{1-0.9} = 3.3$$

$$\hat{m}_{b(n+1)} = \frac{0.155}{1-0.9} = 1.55$$

Step 4

$$\hat{v}_{a(n+1)} = \frac{v_{a(n+1)}}{1-\beta_2}$$

$$\hat{v}_{b(n+1)} = \frac{v_{b(n+1)}}{1-\beta_2}$$

$$\beta_2 = 0.999$$

$$v_{a(n+1)} = 0.01089$$

$$v_{b(n+1)} = 0.0024$$

$$\hat{v}_{a(n+1)} = \frac{0.01089}{1-0.999} = 10.89$$

$$\hat{v}_{b(n+1)} = \frac{0.0024}{1-0.999} = 2.4$$

Step 5

$$a_{n+1} = a_n - \eta \frac{\hat{m}_{a(n+1)}}{\sqrt{\hat{v}_{a(n+1)}} + \varepsilon}$$

$$b_{n+1} = b_n - \eta \frac{\hat{m}_{b(n+1)}}{\sqrt{\hat{v}_{b(n+1)}} + \varepsilon}$$

$$a_n = 0.45$$

$$b_n = 0.75$$

$$\varepsilon = 0.00000001$$

$$\eta = 0.01$$

$$a_{n+1} = 0.45 - (0.01)\frac{3.3}{\sqrt{10.89} + 0.00000001} = 0.449$$

$$b_{n+1} = 0.75 - (0.01)\frac{1.55}{\sqrt{2.4} + 0.0000001} = 0.749$$

Second Epoch

	A	B	C	D	E	F	G	H	I	J	K	L	M
1	a = **0.449**		b= **0.749353**					del SSE/del(a) =-(Y-YP)	del SSE/del(b) =-(Y-YP)X				
2		*Sq Ft*	*Price$*	*X*	*Y*	*YP*	*(1/2)SSE*						
3		1100	199000	0.00	0.00	0.45	0.100801	0.45	0.00				
4		1400	245000	0.22	0.22	0.62	0.076919	0.39	0.09				
5		1425	319000	0.24	0.58	0.63	0.001099	0.05	0.01			m(a)_n=	0.330016
6		1550	240000	0.33	0.20	0.70	0.124878	0.50	0.17			v(a)_n=	0.010891
7		1600	312000	0.37	0.55	0.73	0.015841	0.18	0.07			m(b)_n=	0.154526
8		1700	279000	0.44	0.39	0.78	0.077498	0.39	0.17			v(b)_n=	0.002388
9		1700	310000	0.44	0.54	0.78	0.029576	0.24	0.11			beta_1=	0.9
10		1875	308000	0.57	0.53	0.88	0.06127	0.35	0.20			beta_2=	0.999
11		2350	405000	0.93	1.00	1.14	0.010202	0.14	0.13			eta=	0.001
12		2450	324000	1.00	0.61	1.20	0.17497	0.59	0.59			epsilon=	0.00000001
13	MIN	1100	199000	0	0	**Total SSE=**	**0.673053**	**3.29**	**1.54**				
14	MAX	2450	405000	1	1								
15	RANGE	1350	206000	1	1		m(a)=	0.625736	m(b)=	0.292954			
16							v(a)=	0.021686	v(b)=	0.004754			
17							m^(a)=	6.257358	m^(b)=	2.929538			
18							v^(a)=	21.68589	v^(b)=	4.753531			
19							w(a)=	0.447656	w(b)=	0.748894			

Third Epoch

	A	B	C	D	E	F	G	H	I	J	K	L	M
1		a = **0.447656**		b= **0.748894**				del SSE/del(a) =-(Y-YP)	del SSE/del(b) =-(Y-YP)X				
2		***Sq Ft***	***Price$***	***X***	***Y***	***YP***	***(1/2)SSE***						
3		1100	199000	0.00	0.00	0.45	0.100198	0.45	0.00				
4		1400	245000	0.22	0.22	0.61	0.076353	0.39	0.09				
5		1425	319000	0.24	0.58	0.63	0.001032	0.05	0.01			m(a)_n=	0.625736
6		1550	240000	0.33	0.20	0.70	0.124131	0.50	0.17			v(a)_n=	0.021686
7		1600	312000	0.37	0.55	0.73	0.015573	0.18	0.07			m(b)_n=	0.292954
8		1700	279000	0.44	0.39	0.78	0.07689	0.39	0.17			v(b)_n=	0.004754
9		1700	310000	0.44	0.54	0.78	0.0292	0.24	0.11			beta_1=	0.9
10		1875	308000	0.57	0.53	0.88	0.060709	0.35	0.20			beta_2=	0.999
11		2350	405000	0.93	1.00	1.14	0.009951	0.14	0.13			eta=	0.001
12		2450	324000	1.00	0.61	1.20	0.173905	0.59	0.59			epsilon=	0.00000001
13	MIN	1100	199000	0	0	**Total SSE=**	**0.667941**	**3.27**	**1.53**				
14	MAX	2450	405000	1	1								
15	RANGE	1350	206000	1	1		m(a)=	0.890331	m(b)=	0.416792			
16							v(a)=	0.032368	v(b)=	0.007094			
17							m^(a)=	8.903307	m^(b)=	4.167919			
18							v^(a)=	32.36823	v^(b)=	7.094228			
19							w(a)=	0.446091	w(b)=	0.748519			

Thank You