Uralp Ergin 5975013, İsmail Karabaş 7654321, Ahmet Yaşar Ayfer 5986167

1) $w = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$ $\quad v = [0 \; 0]$ $\quad \alpha = 0.2$

$\beta = 0.8$

$X = \left[\begin{bmatrix} 2 \\ -1 \end{bmatrix} \begin{bmatrix} -1 \\ 3 \end{bmatrix} \right]$ $\quad y = [3, 1]$ $\quad B = 2$

Iteration 1 :

$\rightarrow \hat{y}_i = w^T x_i = [1 \; 0] \begin{bmatrix} 2 \\ -1 \end{bmatrix} = \underline{2} \quad [1 \; 0] \begin{bmatrix} -1 \\ 3 \end{bmatrix} = \underline{-1}$

$MSE = \frac{1}{B} \sum_{i=1}^{B} (\hat{y_i} - y_i)^2 = \frac{1}{2}\left[(-2-3)^2 + (-1-1)^2 \right]$

$\hat{y} = Xw = \begin{bmatrix} 2 & -1 \\ -1 & 3 \end{bmatrix} \begin{bmatrix} 1 \\ 0 \end{bmatrix} = \begin{bmatrix} 2 \\ -1 \end{bmatrix}$

$\hat{y} - y = \begin{bmatrix} 2 \\ -1 \end{bmatrix} - \begin{bmatrix} 3 \\ 1 \end{bmatrix} = \begin{bmatrix} -1 \\ -2 \end{bmatrix}$

$\nabla w = \frac{2}{B} \frac{dL}{d\hat{y}} \frac{d\hat{y}}{dw} = \frac{1}{2} \cdot 2 \cdot X^T (\hat{y} - y)$

$\nabla w : \begin{bmatrix} 2 & -1 \\ -1 & 3 \end{bmatrix}^T \begin{bmatrix} -1 \\ -2 \end{bmatrix} = \begin{bmatrix} -2 \\ -5 \end{bmatrix}$

$v = \beta v - \alpha \nabla w = 0.8 \begin{bmatrix} 0 \\ 0 \end{bmatrix} - 0.2 \begin{bmatrix} -2 \\ -5 \end{bmatrix} = \begin{bmatrix} 0.4 \\ 1.0 \end{bmatrix}$

$w' = w + v = \begin{bmatrix} 1 \\ 0 \end{bmatrix} + \begin{bmatrix} 0.4 \\ 1.0 \end{bmatrix} = \begin{bmatrix} 1.4 \\ 1.0 \end{bmatrix}$

Iteratio 2 :

$$\hat{y} = Xw = \begin{bmatrix} 2 & -1 \\ -1 & 3 \end{bmatrix} \begin{bmatrix} 1.4 \\ 1 \end{bmatrix} = \begin{bmatrix} 1.8 \\ 1.6 \end{bmatrix}$$

$$\hat{y} - y = \begin{bmatrix} 1.8 \\ 1.6 \end{bmatrix} - \begin{bmatrix} 3 \\ 1 \end{bmatrix} = \begin{bmatrix} -1.2 \\ 0.6 \end{bmatrix}$$

$$\nabla_w = \frac{1}{B} 2 \cdot X^T \cdot (\hat{y} - y) = \frac{1}{2} \cdot 2 \cdot \begin{bmatrix} 2 & -1 \\ -1 & 3 \end{bmatrix}^T \cdot \begin{bmatrix} -1.2 \\ 0.6 \end{bmatrix}$$

$$= \begin{bmatrix} -3 \\ 2.4 \end{bmatrix}$$

$$v = 0.8 \begin{bmatrix} 0.4 \\ 1 \end{bmatrix} - 0.2 \begin{bmatrix} -3 \\ 2.4 \end{bmatrix} = \begin{bmatrix} 1 \\ 0.32 \end{bmatrix}$$

$$w = w + v = \begin{bmatrix} 1.4 \\ 1 \end{bmatrix} + \begin{bmatrix} 1 \\ 0.32 \end{bmatrix} = \underline{\underline{\begin{bmatrix} 2.4 \\ 1.32 \end{bmatrix}}}$$

2) $s_t = p_1 s_{t-1} + (1 - p_1) g$

$g \rightarrow$ constant gradient $\qquad s = 0$

$p_1 \rightarrow$ decay rate for first moment

$$\hat{s_t} = \frac{s_t}{1 - p_1^t} \quad \text{(bias correction)}$$

prove $s_t = g$ at every step

$t=1$

$$s_1 = p_1 \cdot s_0 + (1-p_1)g = 0 + (1-p_1)g$$

$$\hat{s_1} = \frac{s_1}{1-p_1^{\,t}} = \frac{\cancel{(1-p_1)}\,g}{\cancel{1-p_1}} = g$$

Inductive Step:

Assume $\hat{s_t} = g$ for $t=k$ and prove for $t=k+1$

$\rightarrow$ step $t=k$ we have $\hat{s_k} = g$

$$\hat{s_k} = \frac{s_k}{1-p_1^{\,k}} = g$$

$$s_{k+1} = p_1 s_k + (1-p_1)g$$

$$= p_1 (1-p_1^{\,k})g + (1-p_1)g$$

$$= (p_1(1-p_1^{\,k}) + (1-p_1))g$$

$$s_{k+1} = (p_1 - p_1^{\,k+1} + 1 - p_1)g = (1-p_1^{\,k+1})g$$

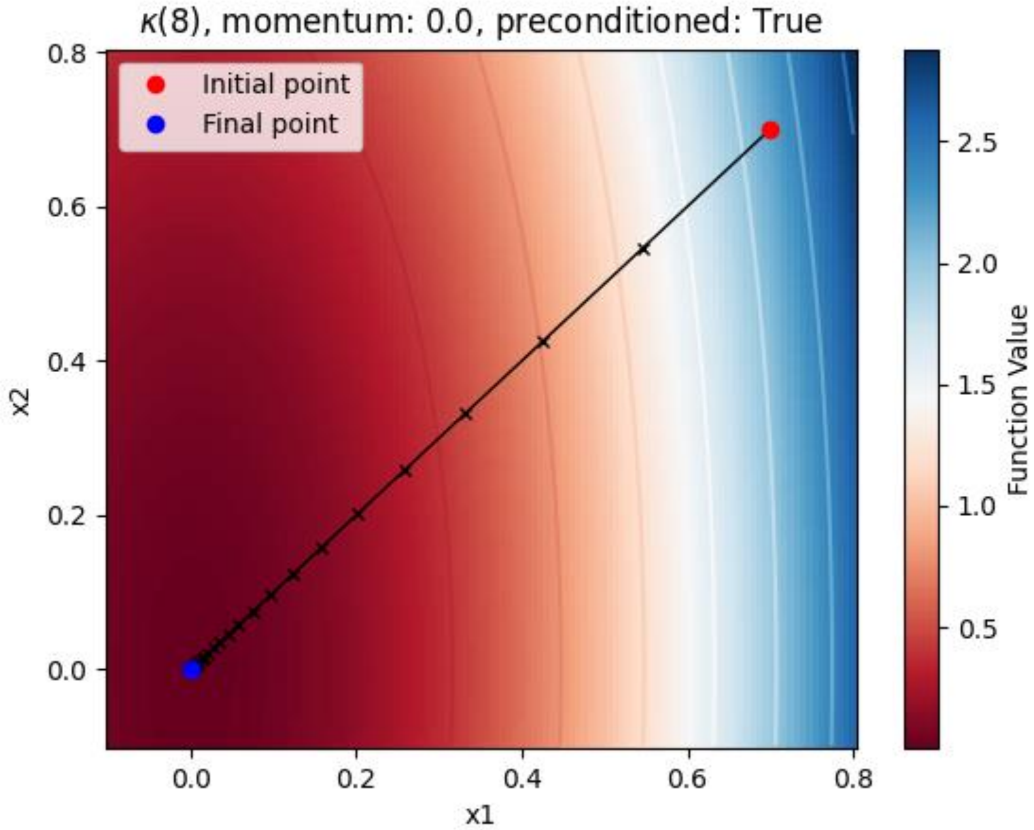
$$\hat{s_{k+1}} = \frac{s_{k+1}}{1-p^{k+1}} = \frac{\cancel{(1-p_1^{\,k+1})}\,g}{\cancel{1-p_1^{\,k+1}}} = g$$

Answer for the reason of zig-zag behaviour at 2nd ill-conditioned plot:

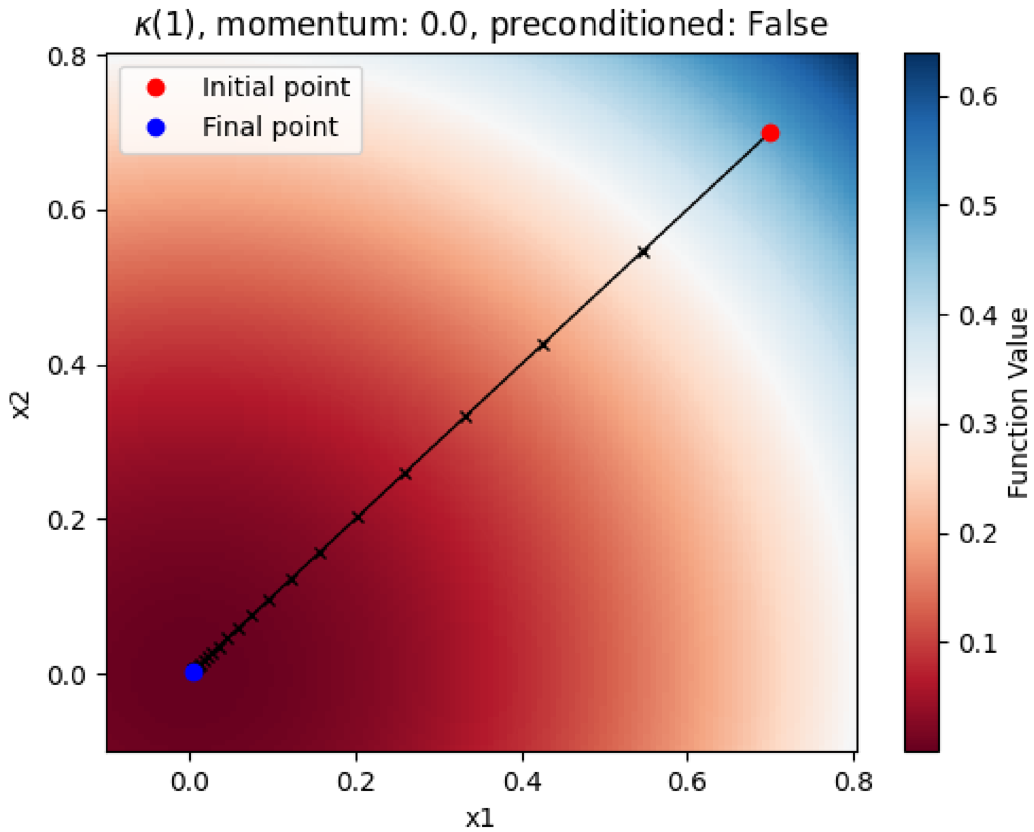
Because along the directions that have large eigenvalues, gradient takes big steps and overshoot which causes the zig-zag behaviour.

Momentum: Accumulates past gradients and smooth the zig-zag behaviour a little bit, enabling faster convergence (fewer steps taken)

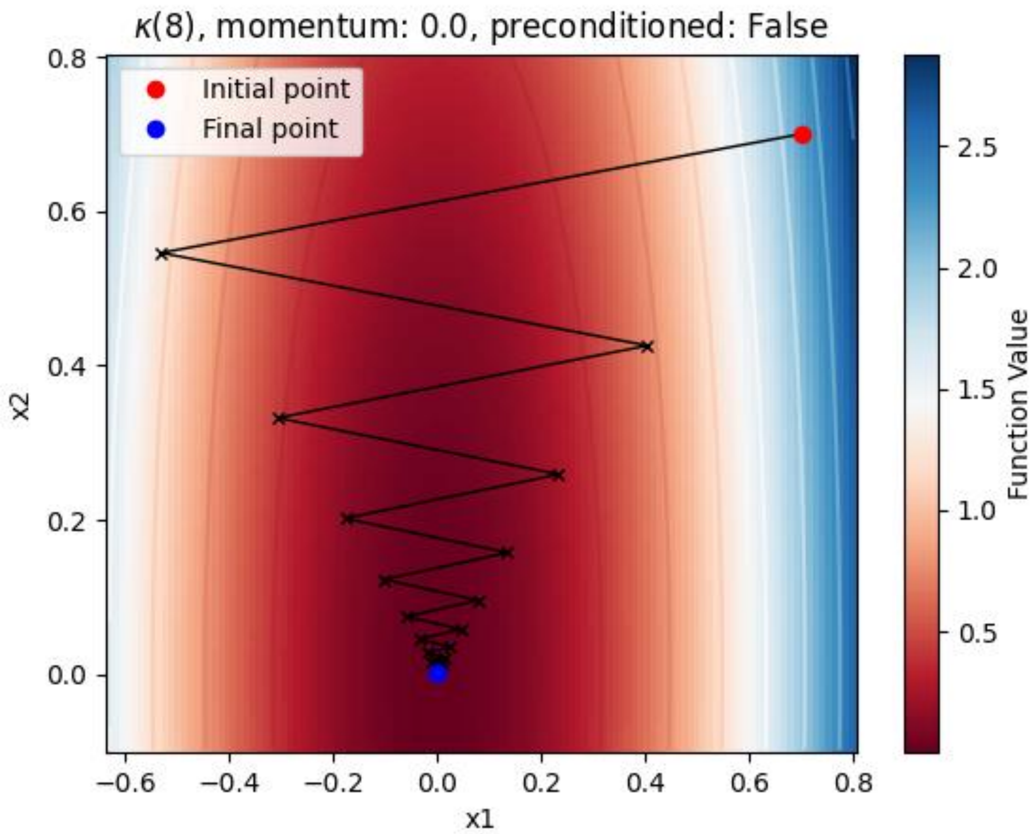
Preconditioning: Reduces the difference between eigenvalues which help preventing the ill-condition and zig-zag behaviour. Steps are more consistent and directly to the minimum.



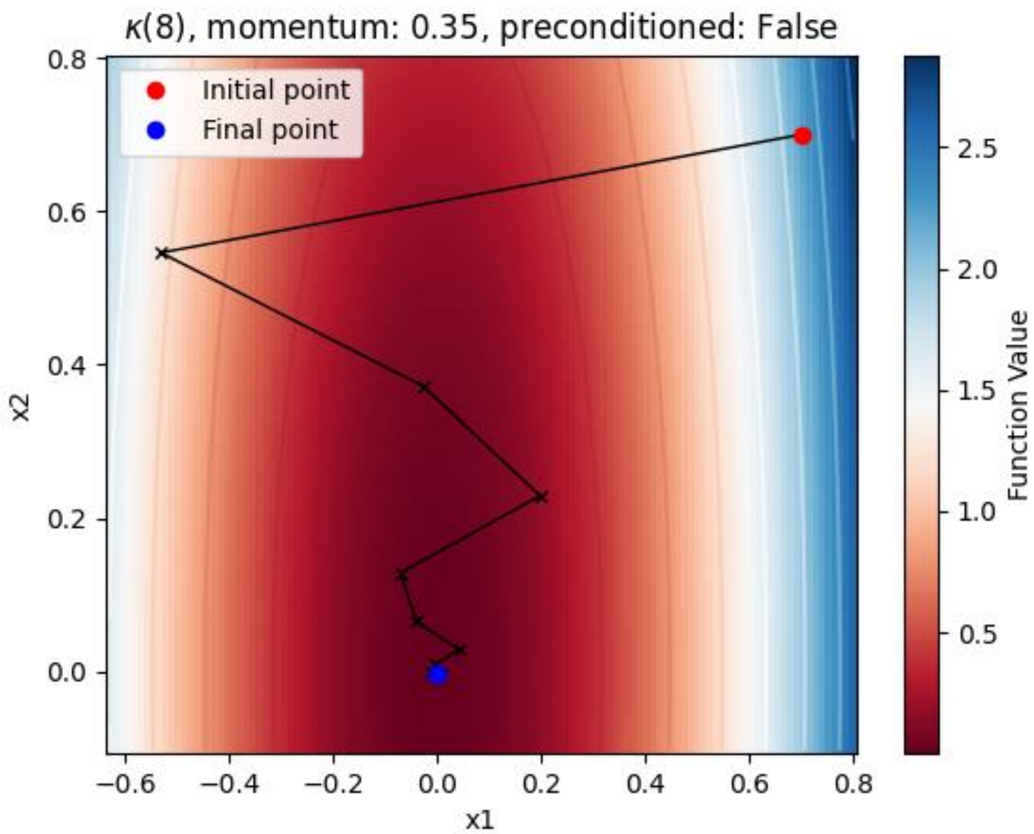
κ(8), momentum: 0.0, preconditioned: True
Initial point
Final point
x1
x2
Function Value
0.0
0.2
0.4
0.6
0.8
0.5
1.0
1.5
2.0
2.5



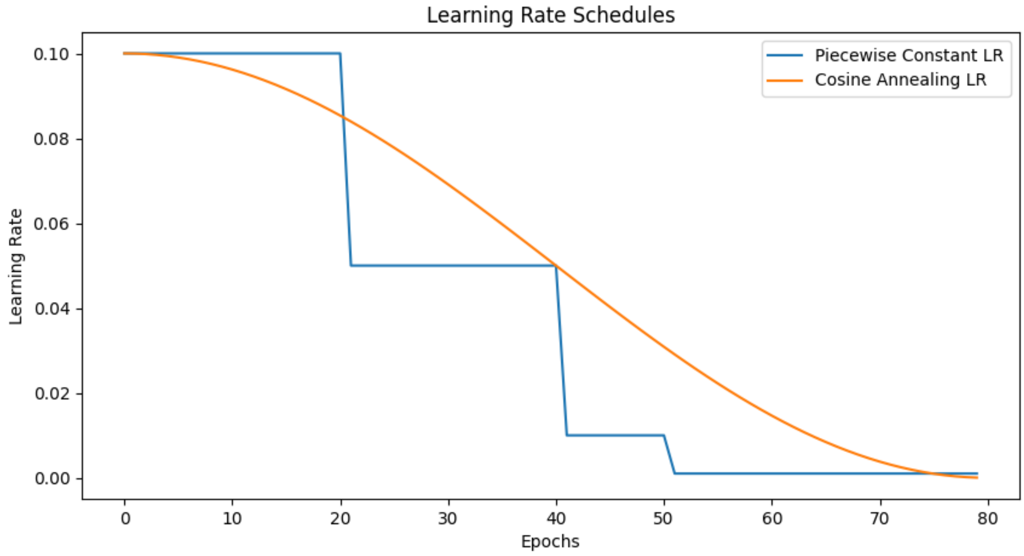
κ(1), momentum: 0.0, preconditioned: False
Initial point
Final point
x1
x2
Function Value
0.0
0.2
0.4
0.6
0.8
0.1
0.2
0.3
0.4
0.5
0.6

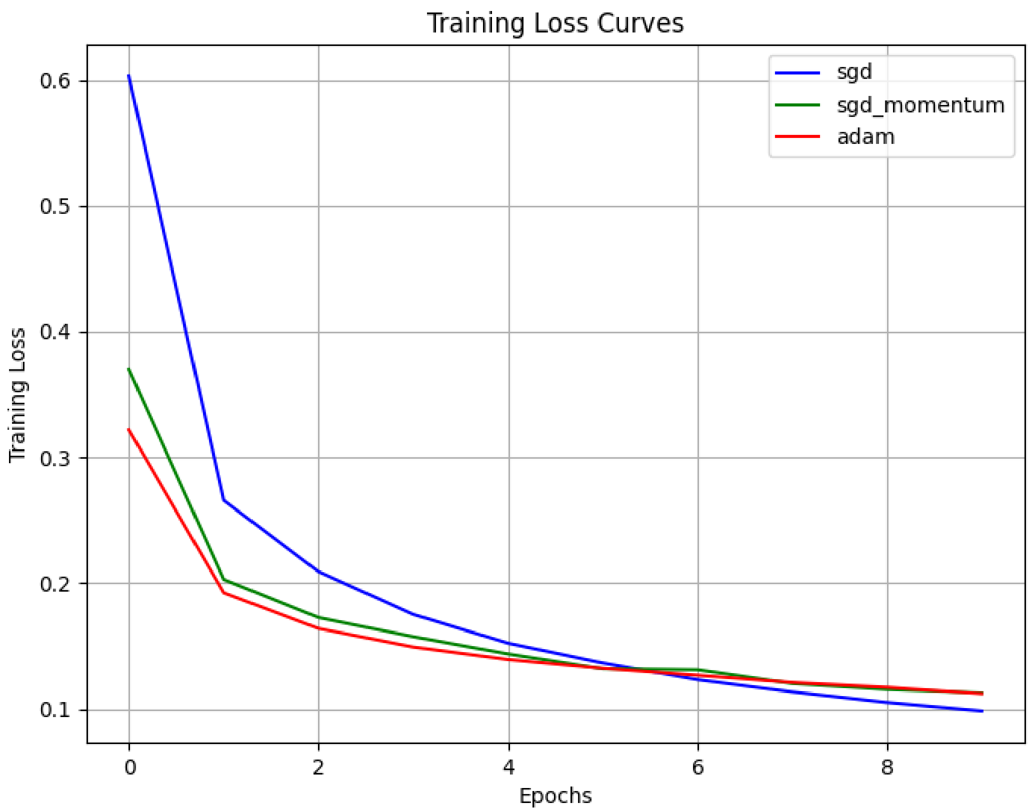


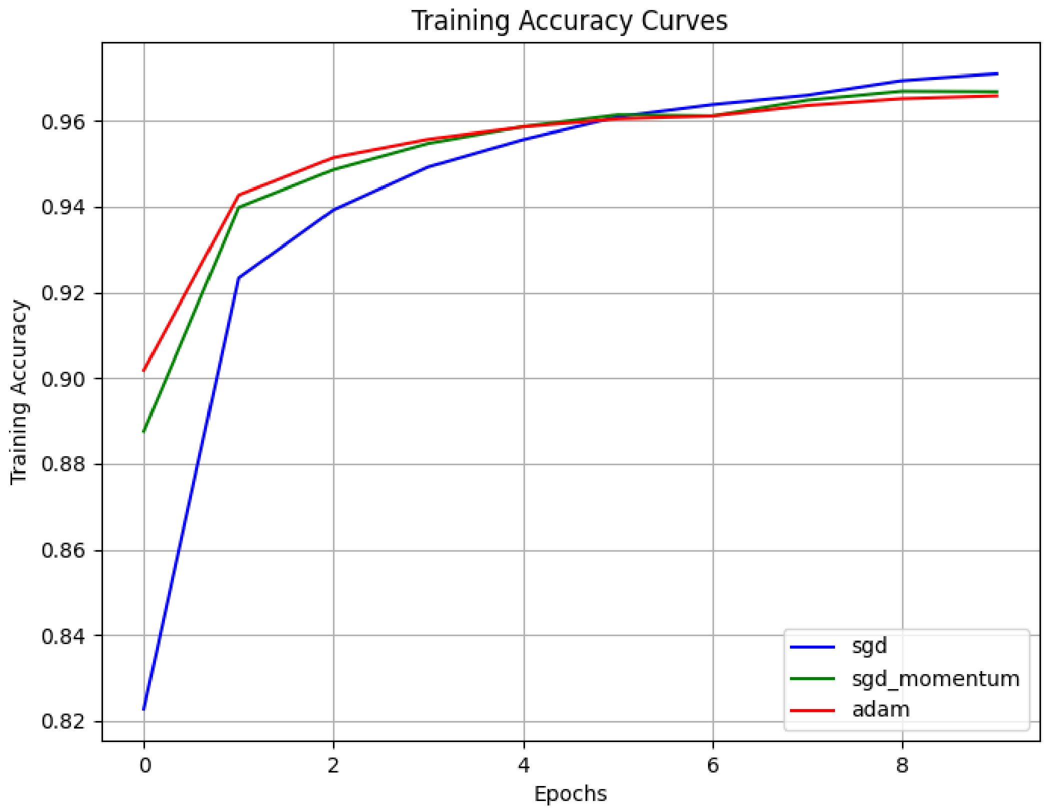
κ(8), momentum: 0.0, preconditioned: False
Initial point
Final point
x1
x2
Function Value



κ(8), momentum: 0.35, preconditioned: False
Initial point
Final point
x1
x2
Function Value


Learning Rate Schedules
Piecewise Constant LR
Cosine Annealing LR
Learning Rate
Epochs
0.10
0.08
0.06
0.04
0.02
0.00
0
10
20
30
40
50
60
70
80


Training Loss Curves
sgd
sgd_momentum
adam
Training Loss
Epochs
0.6
0.5
0.4
0.3
0.2
0.1
0
2
4
6
8


Training Accuracy Curves
Training Accuracy
Epochs
0.82
0.84
0.86
0.88
0.90
0.92
0.94
0.96
0
2
4
6
8
sgd
sgd_momentum
adam