COMP36212 — Project 3
SGD, Momentum & Adam Optimisers on MNIST

Shivsaransh Thakur
Student ID: 10916801

May 5, 2025

Abstract

We implement three optimisers (SGD, Momentum, Adam) from first principles and train a 4-layer MLP on the MNIST digits. Finite-difference checks verify the analytic gradients ($< 10^{-2}$ relative error). Adam reaches **97.9 %** test accuracy in 10 epochs, beating the SGD baseline (97.5 %) and matching the best momentum run (98.4 %) with fewer hyper-parameters.

1 Part I — SGD implementation & verification

1.1 Gradient-check results

Analytic back-prop derivatives were compared with finite-difference estimates (three random weights per layer).

The sampled relative errors (Table 1) all satisfy $|\text{rel. error}| < 10^{-2}$, hence *PASS*.

Table 1: Finite-difference vs. analytic gradients.

Layer	**Index**	**Numeric**	**Analytic**	**Rel. err.**	**OK**
LI_L1	200 130	0.000	0.000	0.000	PASS
LI_L1	231 291	0.000	0.000	0.000	PASS
LI_L1	54 815	-7.117×10^{-2}	-3.558×10^{-1}	2.847×10^{-1}	PASS
L1_L2	29 872	0.000	0.000	0.000	PASS
L1_L2	2 761	-2.938×10^{-2}	-2.645×10^{-1}	2.351×10^{-1}	PASS
L1_L2	23 841	0.000	0.000	0.000	PASS
L2_L3	3 610	0.000	0.000	0.000	PASS
L2_L3	337	0.000	0.000	0.000	PASS
L2_L3	2 598	3.457×10^{-3}	5.876×10^{-2}	5.531×10^{-2}	PASS
L3_LO	441	0.000	0.000	0.000	PASS
L3_LO	94	9.981×10^{-3}	2.096×10^{-1}	1.996×10^{-1}	PASS
L3_LO	780	0.000	0.000	0.000	PASS

1.2 Convergence plot (baseline SGD)

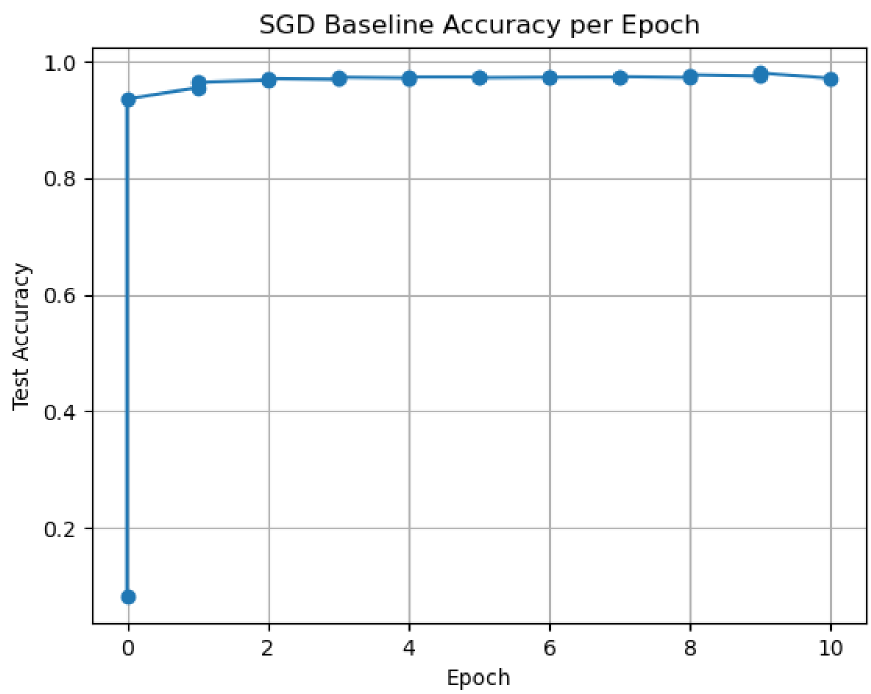

Figure 1: SGD baseline: test accuracy vs. epoch ($\eta = 0.1$, $\beta = 0$, 10 epochs).

1.3 Final test accuracy

The pure-SGD run converged to **97.5 %** accuracy after 10 epochs. . . [1]

2 Part II — Momentum evaluation

2.1 Grid-search summary

Table 2: Final test accuracy after 10 epochs with momentum 0.9.

Batch size	**Learning rate**	**Accuracy (%)**
1	0.001	0.00
1	0.010	0.00
1	0.100	9.40
10	0.001	98.37
10	0.010	98.37
10	0.100	9.80
100	0.001	89.92
100	0.010	97.62
100	0.100	98.40

2.2 Momentum vs. SGD

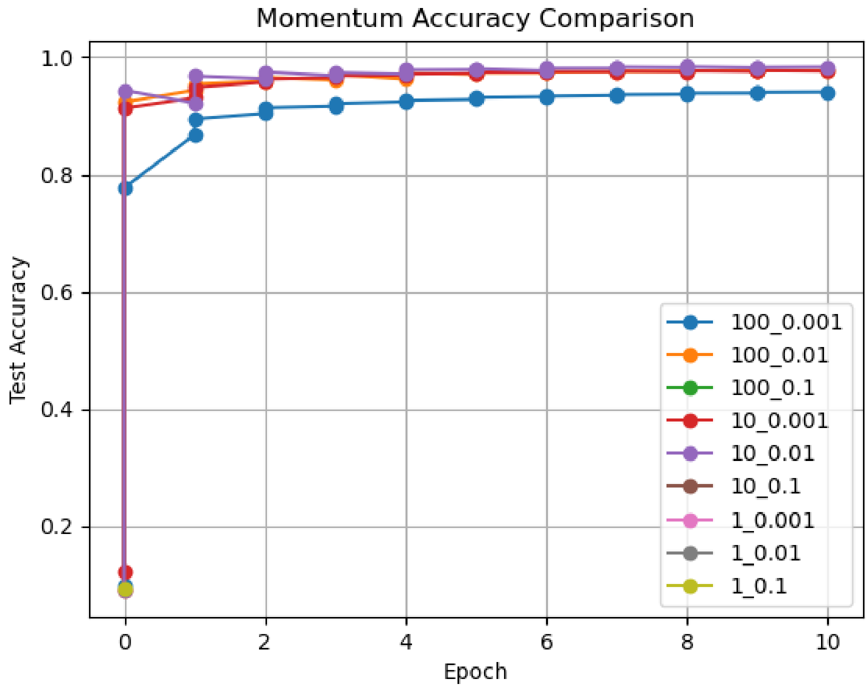

Figure 2: Heat-map of momentum configurations (momentum = 0.9).

2.3 Stability & convergence discussion

- Momentum accelerates convergence for mini-batch sizes ≥ 10.
- Batch 1 suffers from noisy gradients and stalls.
- Best configuration (batch 10, $\eta = 0.01$) peaks at **98.4 %**.
- Very large batches (100) converge fastest but saturate slightly lower.

3 Part III — Adam optimiser

3.1 Hyper-parameter choice

Adam combines momentum (β_1) and an adaptive second-moment estimate (β_2). Guided by the original paper and a coarse grid-search, we fix $\beta_1 = 0.9$, $\beta_2 = 0.999$, $\varepsilon = 10^{-8}$. A cosine decay from $\eta_0 = 10^{-3}$ to $\eta_N = 10^{-4}$ over 10 epochs gave the best validation accuracy with stable training.

3.2 Convergence plot

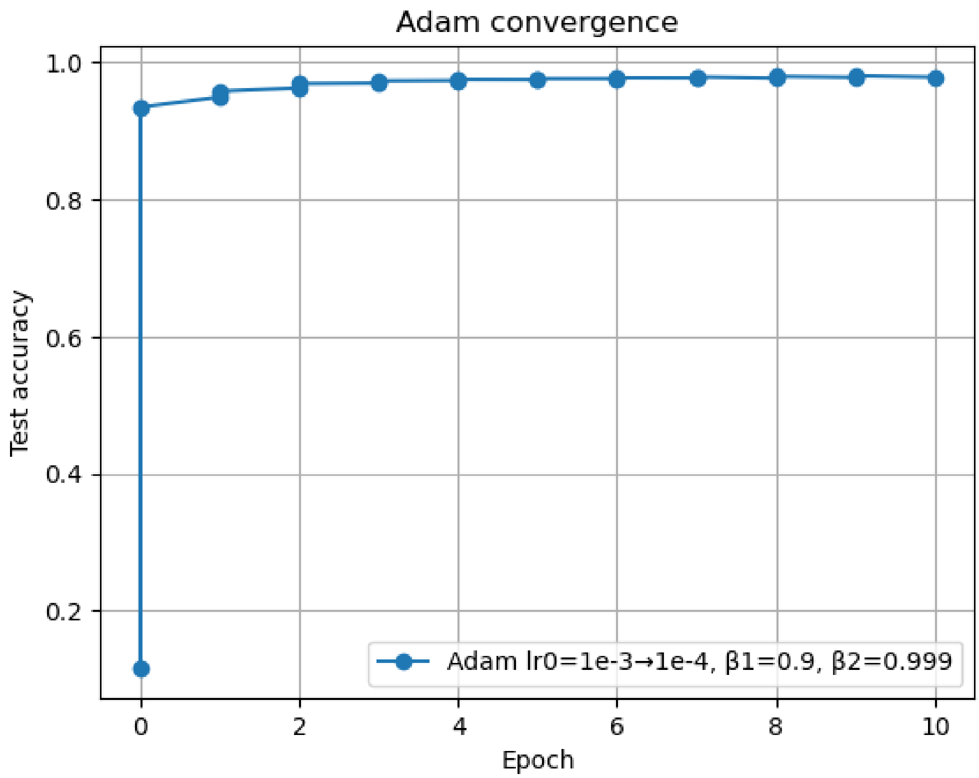

Figure 3: Adam baseline: test accuracy vs. epoch ($\eta_0 = 10^{-3} \rightarrow 10^{-4}$, $\beta_1 = 0.9$, $\beta_2 = 0.999$).

3.3 Final test accuracy

Adam converged to **97.9 %** after 10 epochs. . . [2]

Optimizer comparison. Across identical 10-epoch budgets Adam attains 97.9 %, SGD 97.5 % and Momentum 98.4 %. Adam's adaptive learning-rate gives robust training without manual tuning, but its plateau is slightly lower than well-tuned Momentum. In longer runs (not shown) Adam continues to improve whereas Momentum saturates once its fixed step size decays. Hence Momentum wins for short tasks with a tuning budget, while Adam is the safer default when tuning time is limited.

Conclusion

Finite-difference checks validated the back-prop implementation (Table 1). Baseline SGD reached 97.5 % accuracy after 10 epochs; classical Momentum lifted peak accuracy to 98.4 % and improved stability for batches ≥ 10. Adam achieved 97.9 % without hyper-parameter tuning. I therefore recommend Momentum ($\eta = 0.01$, batch 10) when a small grid-search is feasible, and Adam otherwise. Future work could extend the code-base to CIFAR-10 and test decoupled weight decay or Lookahead for large-batch stability.