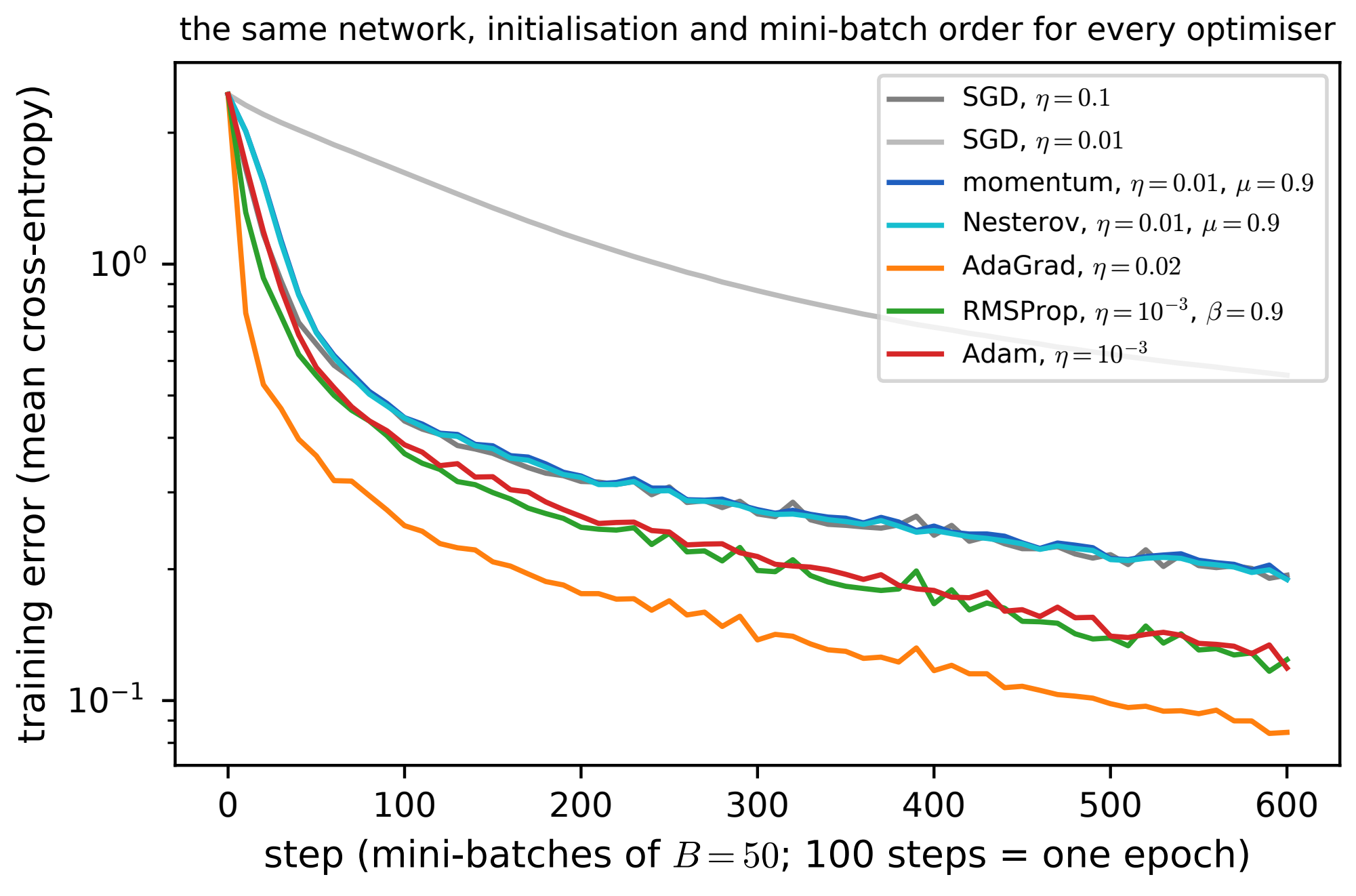

the same network, initialisation and mini-batch order for every optimiser
training error (mean cross-entropy)
10^0
10^{-1}
0
100
200
300
400
500
600
step (mini-batches of $B = 50$; 100 steps = one epoch)
SGD, $\eta = 0.1$
SGD, $\eta = 0.01$
momentum, $\eta = 0.01$, $\mu = 0.9$
Nesterov, $\eta = 0.01$, $\mu = 0.9$
AdaGrad, $\eta = 0.02$
RMSProp, $\eta = 10^{-3}$, $\beta = 0.9$
Adam, $\eta = 10^{-3}$