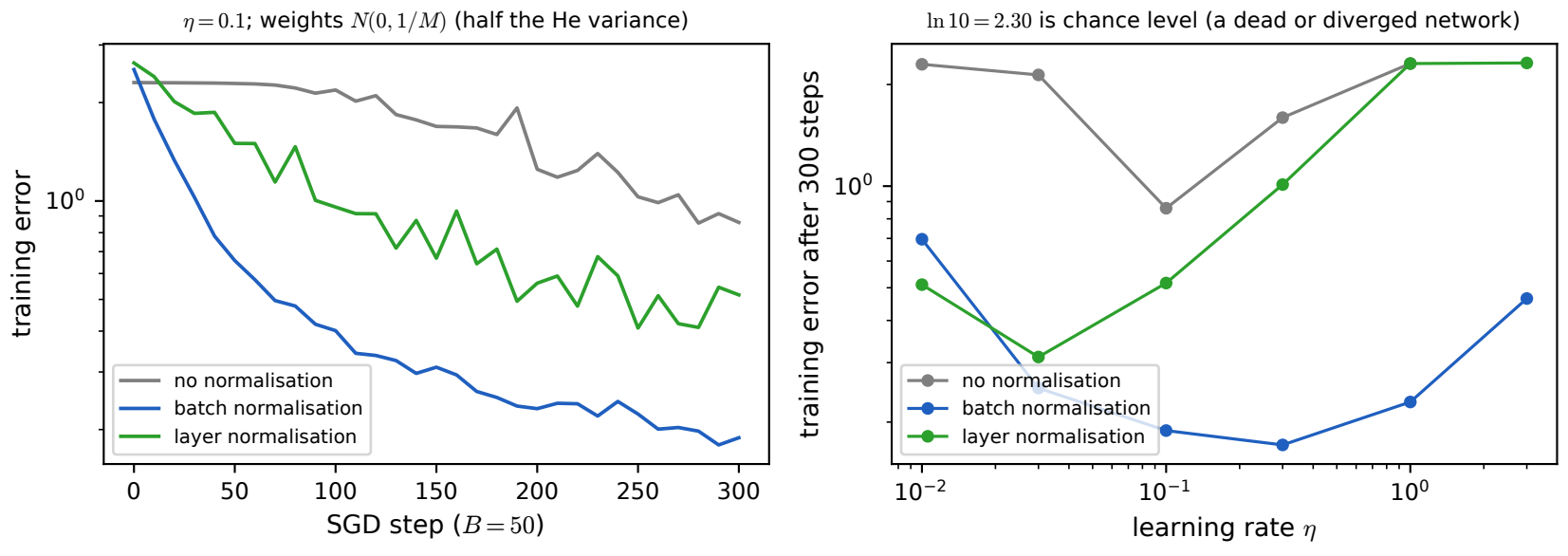

$\eta = 0.1$; weights $N(0, 1/M)$ (half the He variance)
training error
10^0
0
50
100
150
200
250
300
SGD step ($B = 50$)
no normalisation
batch normalisation
layer normalisation
$\ln 10 = 2.30$ is chance level (a dead or diverged network)
training error after 300 steps
10^0
10^{-2}
10^{-1}
10^0
learning rate η
no normalisation
batch normalisation
layer normalisation