000
001
002
003
004
005
006
007
008
009
010
011
012
013
014
015
016
017
018
019
020
021
022
023
024
025
026
027
028
029
030
031
032
033
034
035
036
037
038
039
040
041
042
043
044
045
046
047
048
049
050
051
052
053
054

Figure 1: **Under the suggested weight decay scaling, the optimal learning rate is stable across training length.** Mirroring Fig.1 in the main text, we trained the model for 100 *epochs* with different dataset sizes under a fixed batch size. Using a fixed weight decay (top row), the optimal learning rate *decreases* with the dataset size. Under our suggested weight decay scaling, where $\lambda \propto \frac{1}{\text{dataset size}}$, the optimal learning rate becomes more stable across dataset sizes. Note that we select the values for λ as 0.1 as they were close-to-optimal for the experiments in Fig.1.