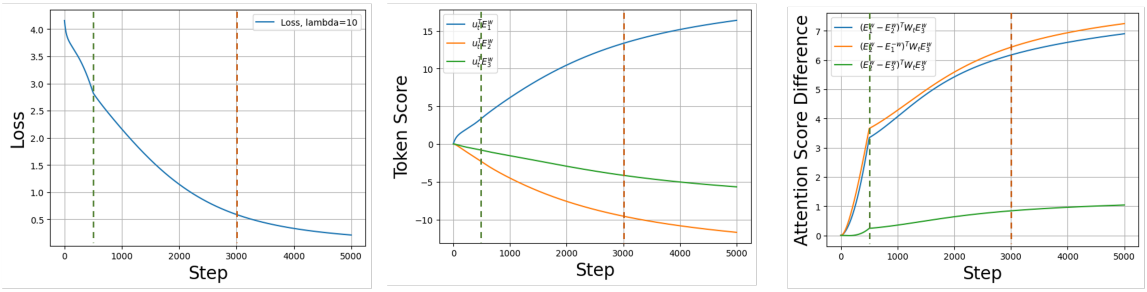

Figure 1: Results of choosing $\lambda = 10$

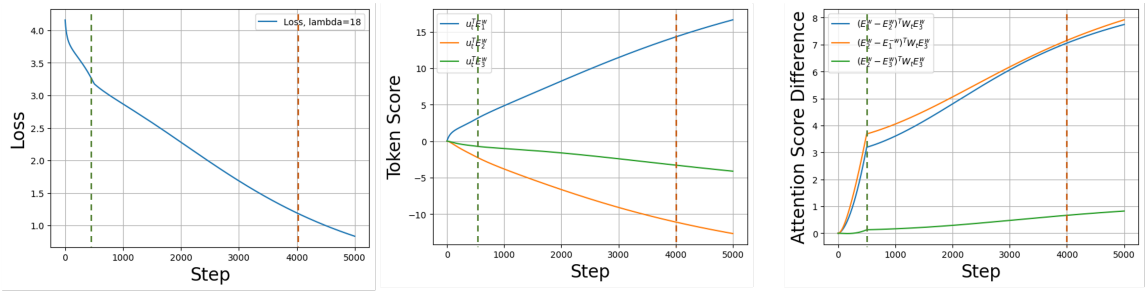

Figure 2: Results of choosing $\lambda = 18 = L_{\max}^2/2$

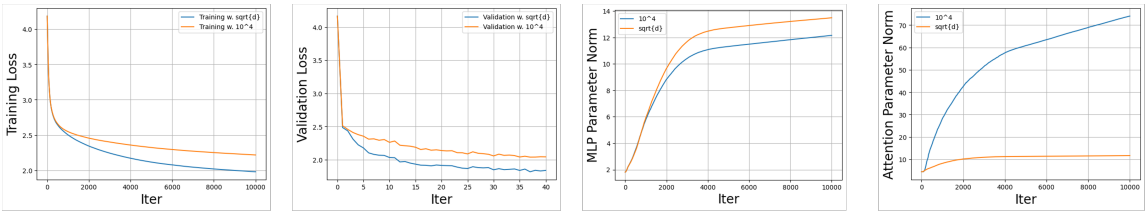

Figure 3: Results of nanoGPT, trained on 'shakespeare' dataset. Configuration: block-size=64, batch-size=12, n-layer=4, n-head=4, n-embd=128

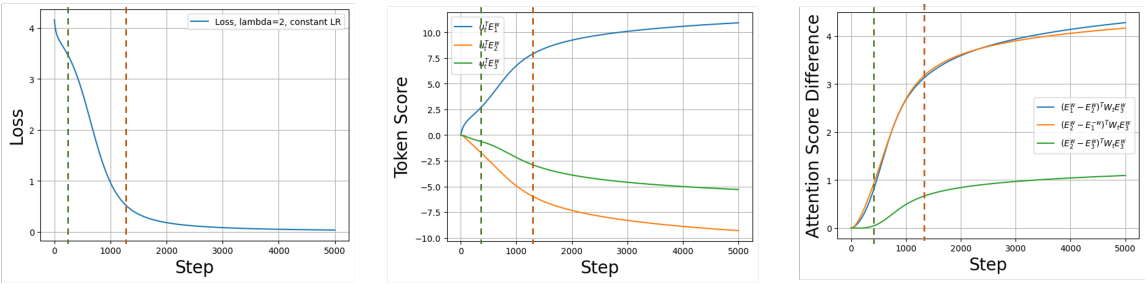

Figure 4: Results of training transformer on even pairs using vanilla GD (constant learning rate) and $\lambda = 2$.