

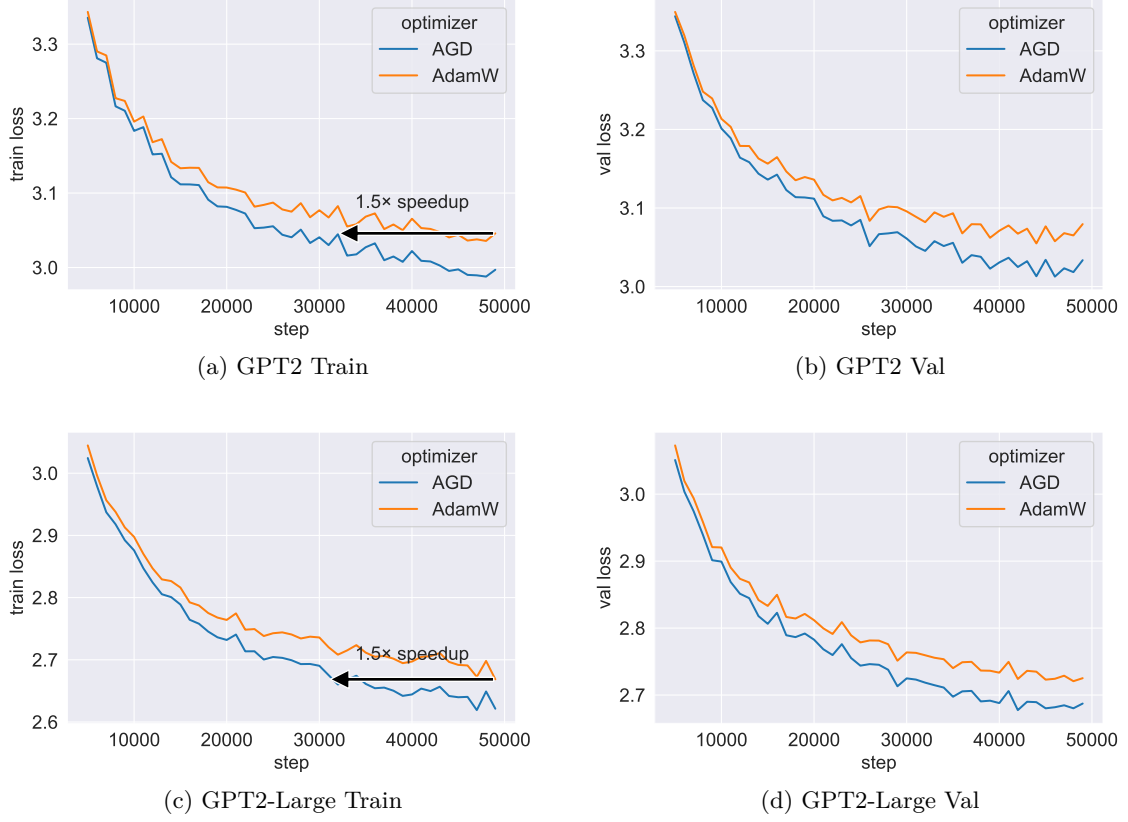


Figure 1: Training/validation loss of GPT2 (124 M) and GPT2-Large (774 M) on OpenWebText

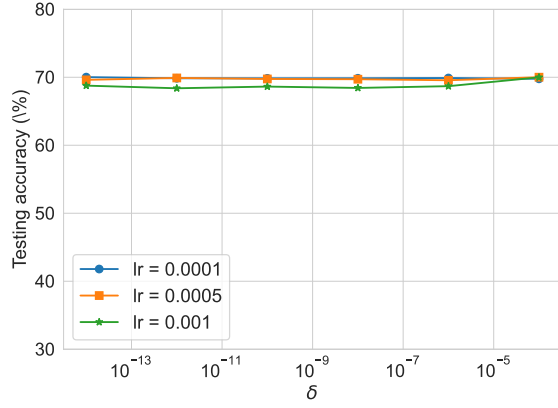


Figure 2: Hyperparameter δ robustness for ResNet18 on ImageNet.

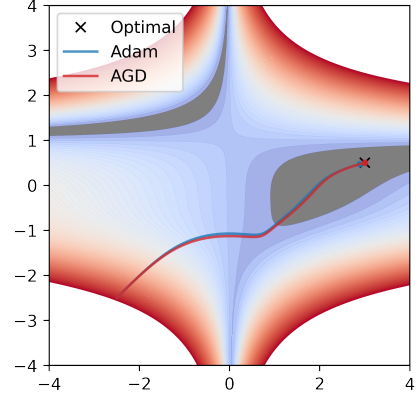


Figure 3: The whole optimization trajectories of AGD and Adam of Beale function.

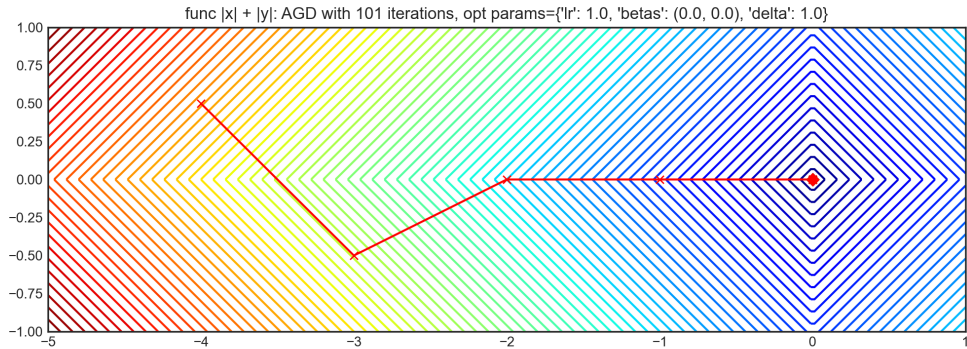


Figure 4: Optimize the function $f(x, y) = |x| + |y|$ using AGD from the starting point $(-4, 0.5)$.