	13B		30B	
	Per step	Total GPU hours	Per step	Total GPU hours
MeZO	1× (2.72s)	1×	1× (5.90s)	1×
Full-parameter fine-tuning	5.04×	0.63×	7.74×	1.94×

Table 1: Wallclock time per step and total GPU hours of different training methods. The statistics are measured on 80GB A100s with NVLink and InfiniteBand connections. The wallclock time is averaged over 100 steps. It is measured on the MultiRC task with the OPT family. 13B and 30B OPT require 1 A100 with MeZO and 4 and 8 A100s with fine-tuning respectively. We use the hyperparameters in our main experiments (5 epochs or 625 steps for fine-tuning and 20K steps for MeZO; a batch size of 8 for fine-tuning and 16 for MeZO). Total GPU hours are calculated as **wallclock time per step** × **number of steps** × **number of GPUs**.

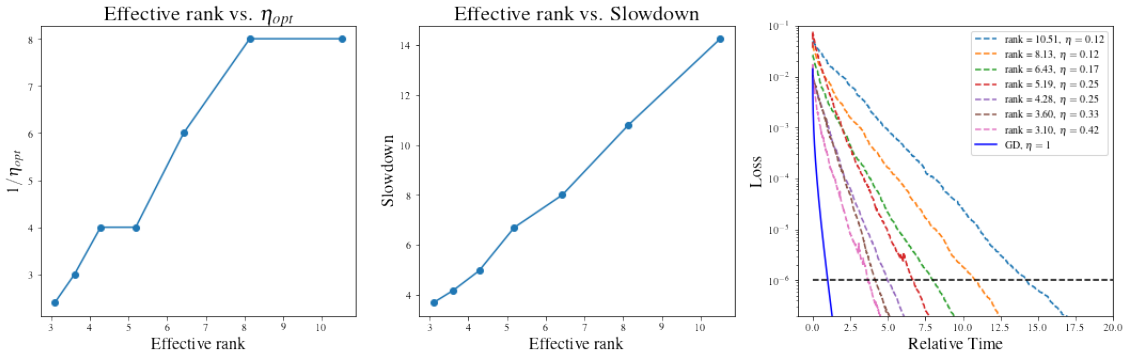

Figure 1: To verify the dependence of MeZO on effective rank, we run MeZO on the toy target function $f^*(x) = \frac{1}{2}x^T Ax$, where $x \in \mathbb{R}^{100}$. We consider various candidate problems where the eigenvalues of A are $\lambda_i = i^{-\alpha}$ for $i \in [d]$, where the value of α changes for each problem; this yields a suite of problems each with differing effective ranks. **Left:** For each problem, we search over multiple learning rates and compute the optimal learning rate η_{opt}, measured by fewest number of iterations to reach a target threshold (10^{-6}). We plot η_{opt}^{-1} against the effective rank, and see that the optimal learning rate decreases by a factor of the effective rank. **Middle:** For each problem we compute the slowdown, measured by the number of steps MeZO with η_{opt} requires to reach the threshold divided by the number of steps GD with $\eta = 1$ (the optimal learning rate for GD) to reach the threshold. Again, we observe that slowdown scales with effective rank. **Right:** For each task, we plot the loss of MeZO with learning rate η_{opt} as a function of relative time. Relative time rescales time for each task so that GD reaches the threshold loss at time $t = 1$. Again, we observe that slowdown scales with the effective rank.