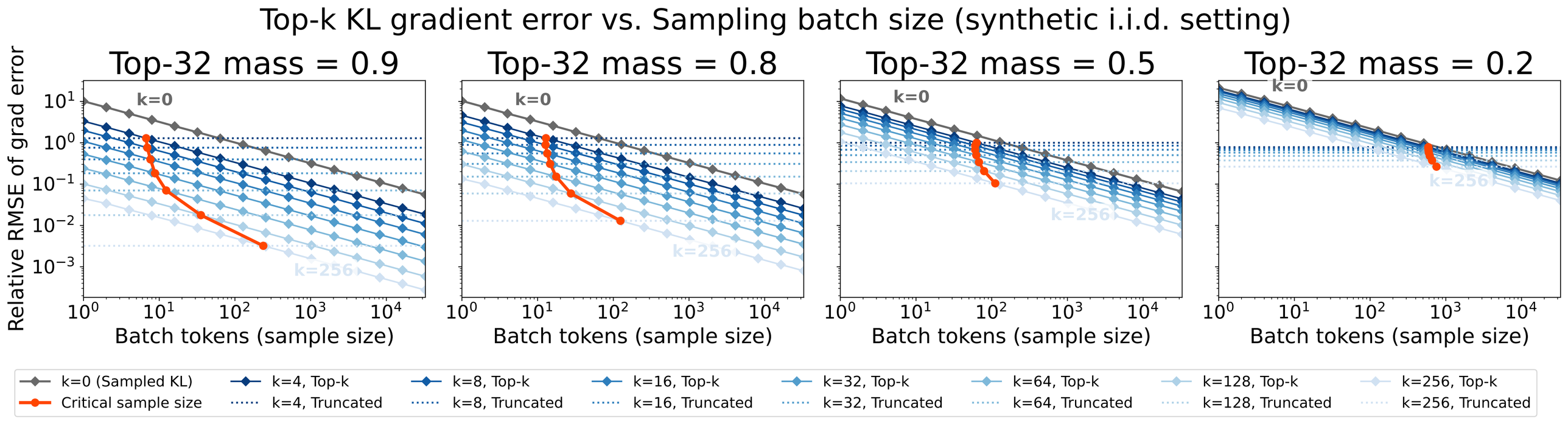

Top-k KL gradient error vs. Sampling batch size (synthetic i.i.d. setting)
Top-32 mass = 0.9
Top-32 mass = 0.8
Top-32 mass = 0.5
Top-32 mass = 0.2
Relative RMSE of grad error
Batch tokens (sample size)
k=0
k=256
k=0 (Sampled KL)
Critical sample size
k=4, Top-k
k=4, Truncated
k=8, Top-k
k=8, Truncated
k=16, Top-k
k=16, Truncated
k=32, Top-k
k=32, Truncated
k=64, Top-k
k=64, Truncated
k=128, Top-k
k=128, Truncated
k=256, Top-k
k=256, Truncated