GPT2-medium stacked weights (d=16384, n=1024)
Gram Matrix Approximation

Wall-clock sketch time (ms)

Gram Frobenius error (relative)

4.43x speedup

FlashSketch (Ours)
Dense Gaussian (cuBLAS)
SJLT (GraSS Kernel)
SJLT (cuSPARSE)
Subsampled Fast Hadamard Transform

Sketch dimension (k)
● 64
■ 256
▲ 512
◆ 1024
+ 2048
✕ 4096