

1 Neurips Rebuttal for MomentumSMoE: Integrating Momentum into Sparse 2 Mixture of Experts.

Table 1: Top-1, top-5 accuracy (%) and mean corruption error (mCE) of MomentumV-MoE and Robust MomentumV-MoE vs. the V-MoE baseline on ImageNet-1K and popular robustness benchmarks for image classification.

Model	Params	Train IN-1K		Valid IN-1K		IN-R	IN-A	IN-C	
		Top-1 $\uparrow$	Top-5 $\uparrow$	Top-1 $\uparrow$	Top-5 $\uparrow$	Top-1 $\uparrow$	Top-1 $\uparrow$	Top-1 $\uparrow$	mCE $\downarrow$
<i>V-MoE (baseline)</i>	297M	76.49	92.27	73.16	90.42	36.10	5.25	46.98	67.14
MomentumV-MoE	297M	76.92	92.19	73.26	90.30	37.45	6.48	48.11	65.77
Robust MomentumV-MoE	297M	76.66	92.27	73.20	90.36	37.57	6.37	48.82	64.92

Table 2: Perplexity (PPL) of NegativeMomentumSMoE and ComplexMomentumSMoE vs. AdamSMoE, MomentumSMoE (Positive) and SMoE baseline on clean/attacked WikiText-103.

Momentum Type	Momentum Coefficient	Clean WikiText-103		Attacked WikiText-103	
		Valid PPL ($\downarrow$)	Test PPL ($\downarrow$)	Valid PPL ($\downarrow$)	Test PPL ($\downarrow$)
<i>None (baseline)</i>	-	33.76	35.55	42.24	44.19
Positive	0.7	32.29	33.46	40.94	42.33
Adam	0.9, 0.999	31.59	33.25	39.27	41.11
Complex	$0.7 + 0.1i$	32.08	33.34	40.24	41.66
Negative	-0.2	33.48	35.09	41.68	43.62

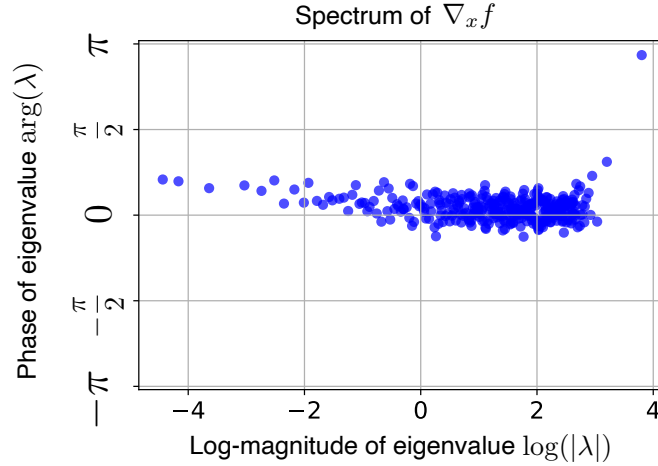


Figure 1: A log-polar coordinate visualization of the spectrum of $\nabla_x f$. As the phase of the eigenvalues are close to 0, the imaginary part is small in comparison to the real part.