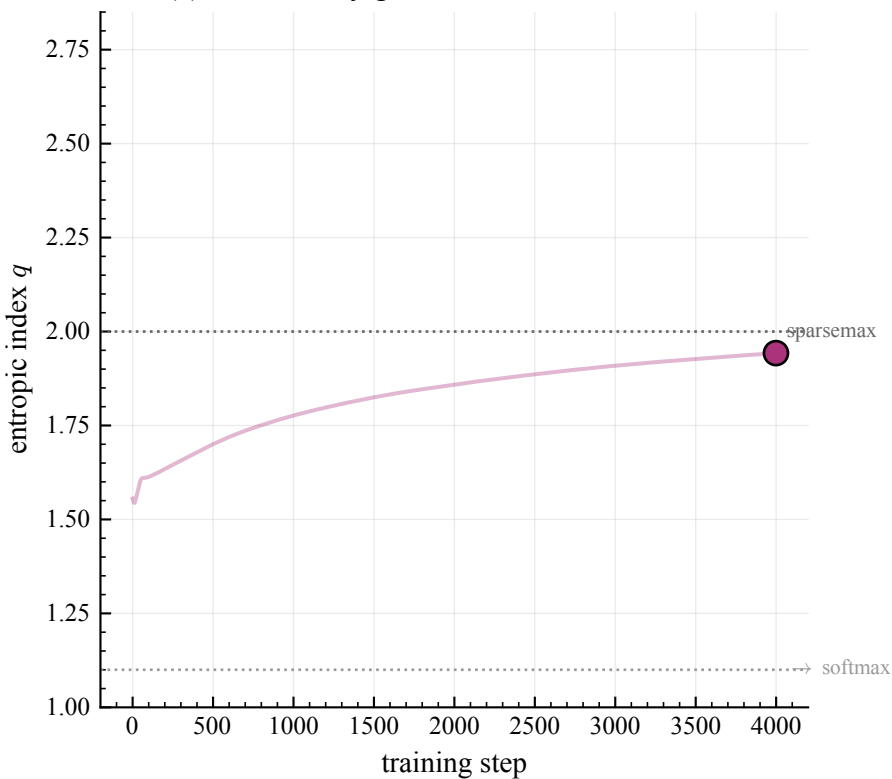


Learning the attention sparsity: q rises, attention sharpens

(a) q learned by gradient descent — $q = 1.94$



(b) attention at step 3999 ($q = 1.94$)

