Updated experiments

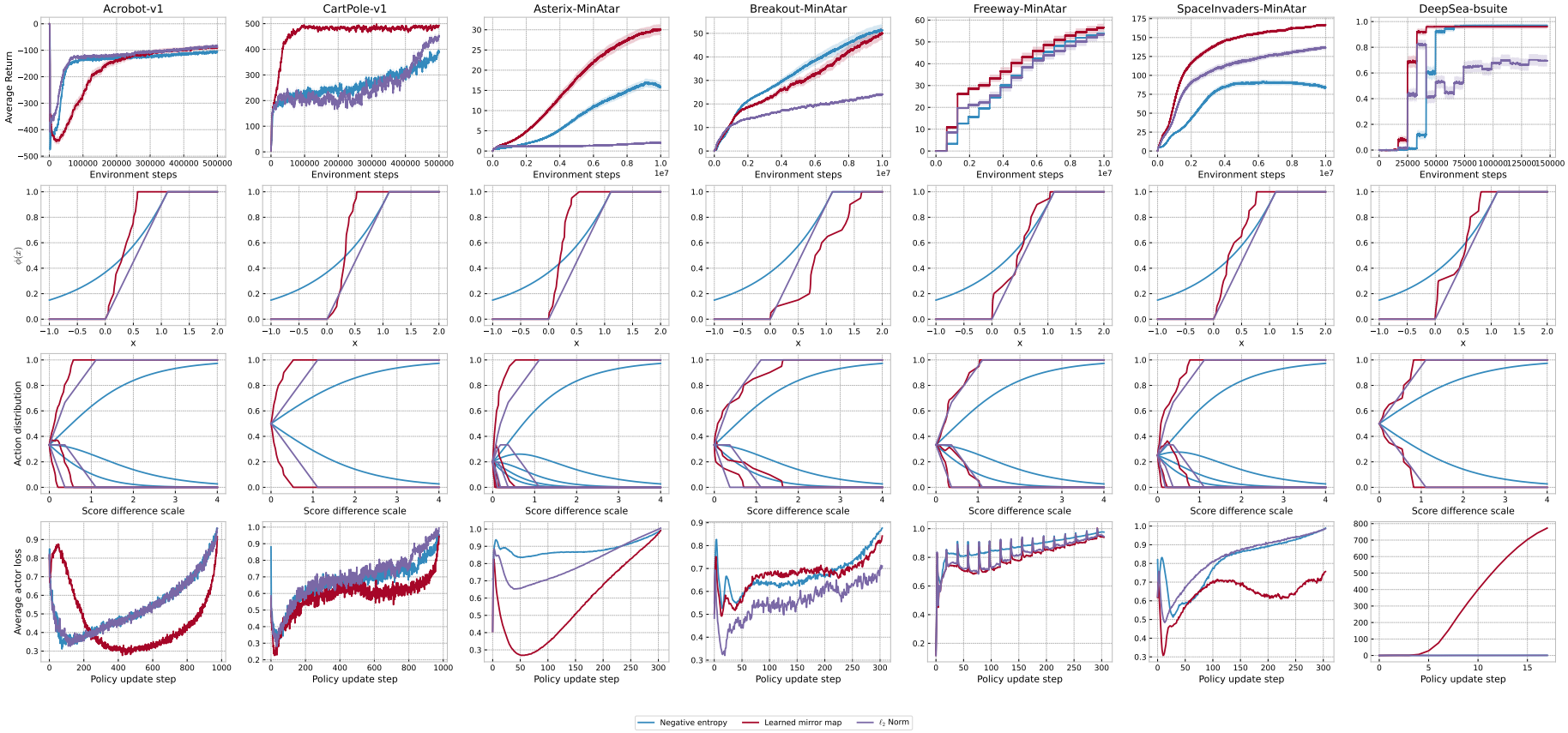

Figure 1: Comparison between the meta-learned mirror map, the negative entropy, and the ℓ_2 norm across a range of standard environments. The top plots display a shaded region denoting the standard error around the mean lines (averaged over 100 realizations). The second and third row show the potentials that induce the mirror maps and the associated Bregman projections, where the action scores are set to $[1, \dots, |\mathcal{A}|] *$ Score difference scale. The forth row shows the actor error along the path of the algorithm, that is $\mathbb{E}\big[||f^{t+1} - \widehat{Q}^t - \eta_t^{-1}\nabla h(\pi^t)||^2_{L_2(v^t)}\big]$, averaged over 100 runs.

	Acrobot	CartPole	Asterix	Breakout	Freeway	SpaceInvaders	DeepSea	Negative entropy
Acrobot	-88.49	-83.76	-103.55	-121.50	-82.51	-78.29	-488.42	-105.63
CartPole	476.41	499.93	490.86	359.06	457.47	489.56	37.15	359.14
Asterix	24.03	27.26	30.22	7.95	3.20	4.36	0.44	17.80
Breakout	9.63	7.77	7.94	51.85	19.96	11.15	0.32	49.34
Freeway	56.00	52.26	58.56	28.37	58.21	22.27	15.69	53.69
SpaceInvaders	144.01	100.07	122.00	85.68	143.93	170.24	3.00	81.77
DeepSea	0.94	0.92	0.96	0.97	0.75	0.74	0.96	0.99

Table 1: In this table we report, for each entry, the performance of the final policy outputted by AMPO trained on the environment corresponding to the row with the mirror map learned on the environment corresponding to the column. The last column represents the performance of AMPO with the negative entropy for the corresponding row environments. The performance is averaged over 100 runs. The displayed results attest how learned mirror maps can generalize well to different environments, as they sometimes perform better than the negative entropy benchmark even if they were learned on a different environment.