

Regularized Parameter Uncertainty for Improving Generalization in Reinforcement Learning

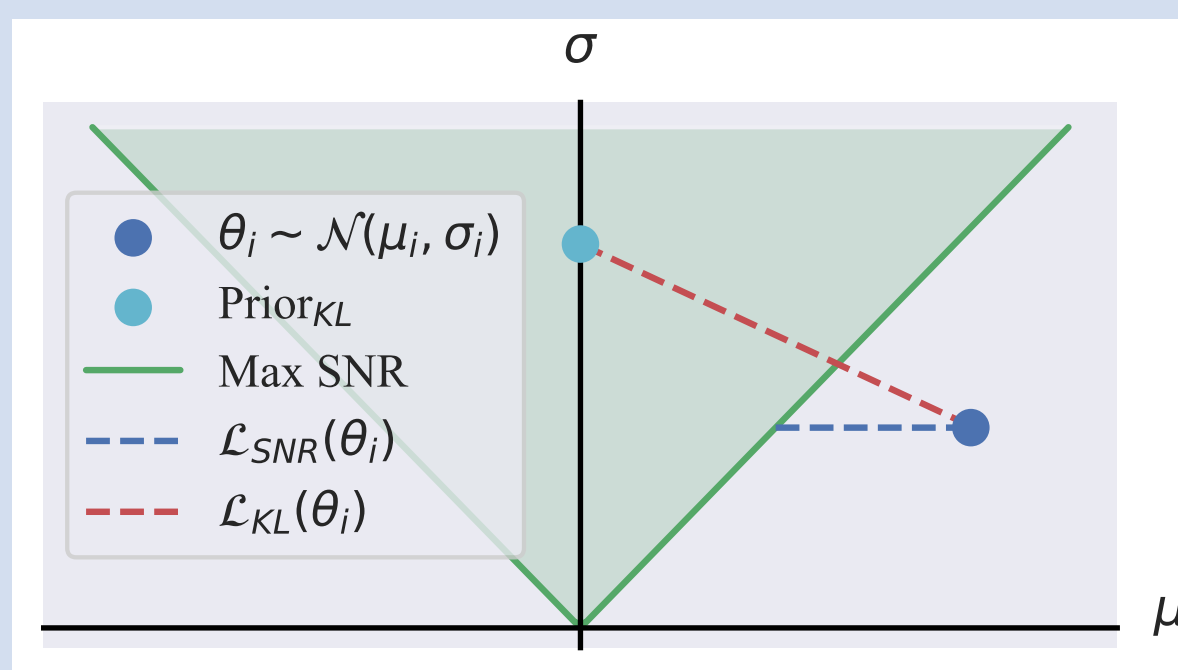
Pehuén Moure¹, Longbiao Cheng¹, Joachim Ott¹, Zuowen Wang¹, Shih-Chii Liu¹

¹Institute of Neuroinformatics, ETH Zurich and University of Zurich

1 Motivation

RL agents struggles with out-of-distribution generalization, often due to over-fitting the particulars of the training environment. [1]

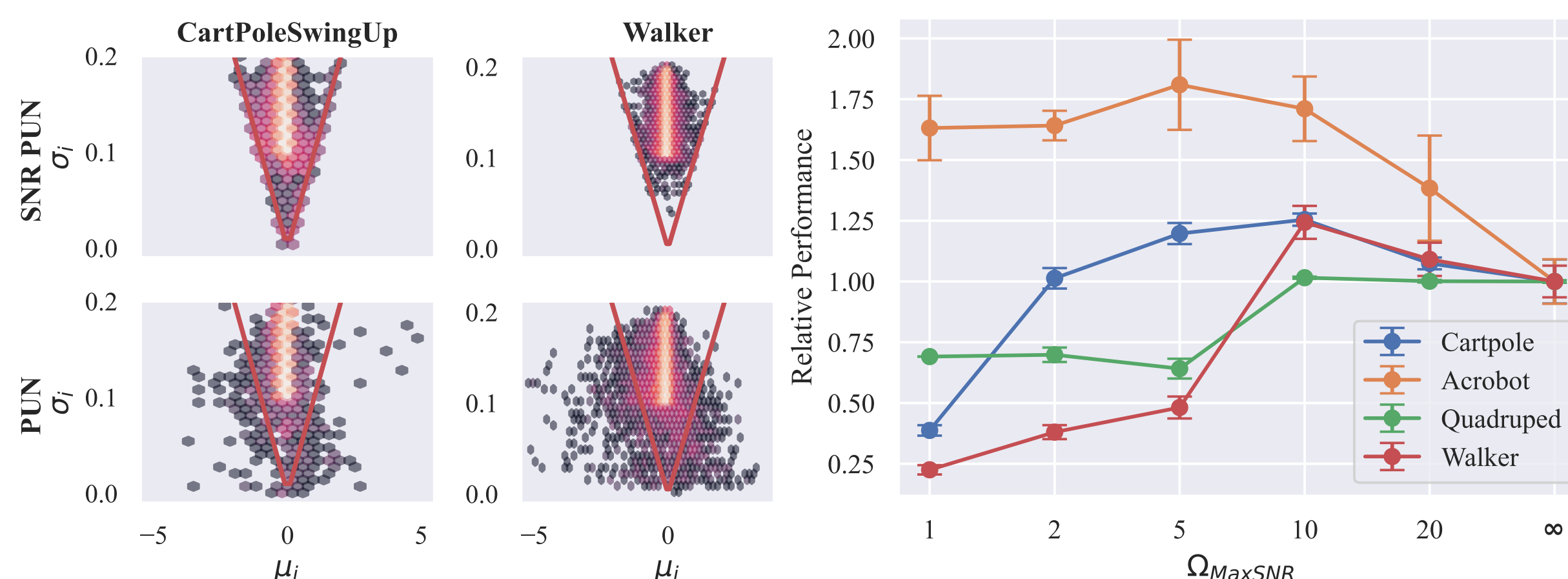
We propose a Signal-to-Noise Ratio regulated Parameter Uncertainty Network (SNR PUN) for improving generalization.



2 SNR Loss Term Effect

When training a PUN the KL divergence term is often dropped (or set to a value that is effectively 0) [2]. As a consequence, the learned distributions are under-regulated, often collapsing each parameter's standard deviation to zero [3].

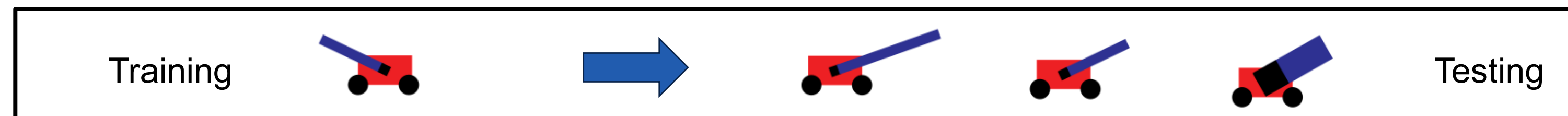
SNR regularization allows a wide range of parameter distributions and can be more heavily penalized during training, this combination lead to overall improved generalization.



3 Simulated Domain Generalization

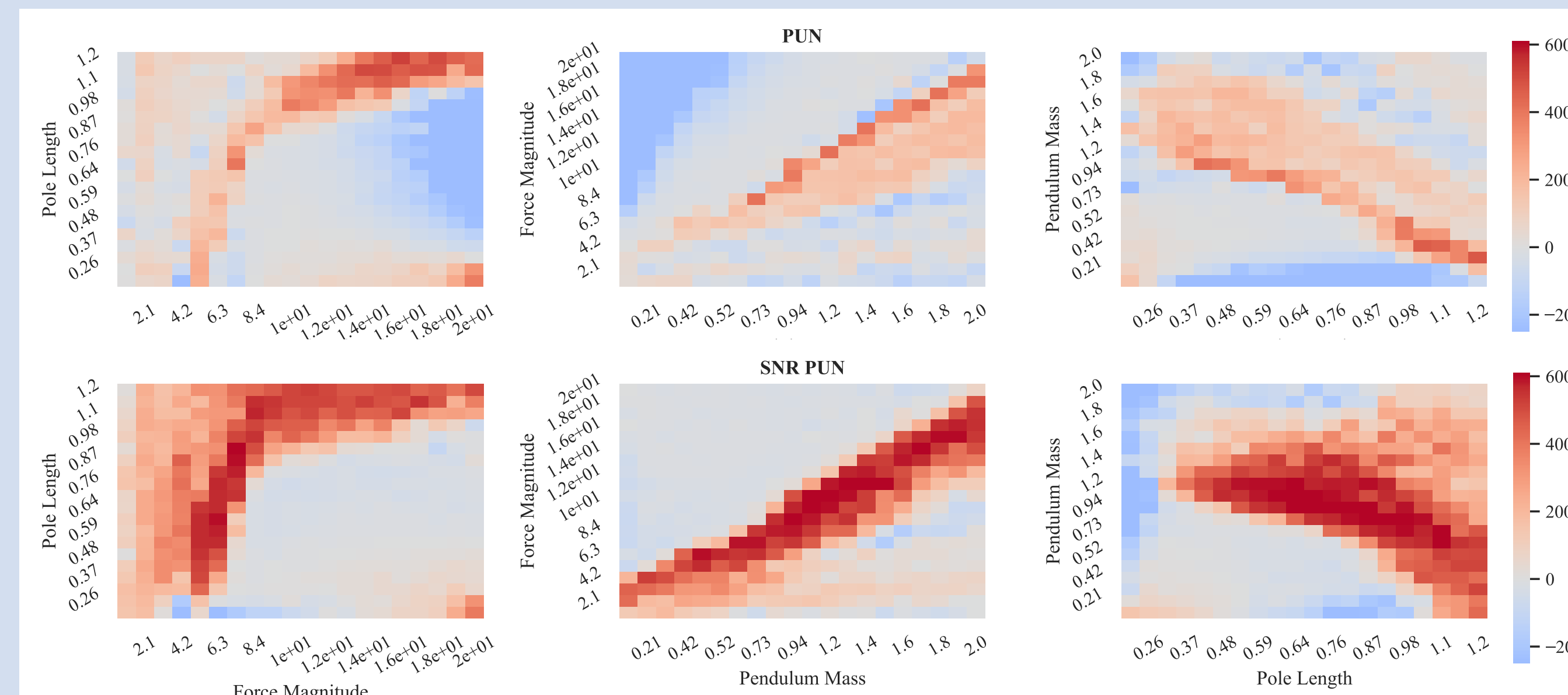
We train on a single environment and then stop model updates. We evaluate our approach by performing sweep over the range of key values for each environment parameter. SNR PUN achieves SOTA performance.

Model	Experiment	Cartpole	Acrobot	Quadruped	Walker
MLP	–	529.64 ± 10.19	160.58 ± 19.07	670.57 ± 28.82	212.96 ± 14.89
	+Dropout	597.61 ± 9.67	148.52 ± 23.28	713.87 ± 2.81	180.84 ± 15.84
	+gSDE	571.46 ± 19.51	–	688.08 ± 8.21	166.03 ± 10.38
	+Entropy Maximization +Weight Decay	662.03 ± 17.77	145.47 ± 27.96	540.25 ± 45.91	301.19 ± 23.48
PUN	–	534.43 ± 48.40	246.57 ± 22.55	738.95 ± 3.91	250.11 ± 16.28
	+Dropout	500.09 ± 64.89	414.34 ± 31.91	686.79 ± 1.71	179.18 ± 13.86
VIB	–	622.72 ± 23.06	302.95 ± 34.92	716.11 ± 3.77	255.27 ± 18.41
	Use Mean	623.06 ± 17.84	303.51 ± 34.93	739.84 ± 16.93	259.18 ± 17.81
SNR PUN	–	670.48 ± 13.42	421.77 ± 32.74	750.47 ± 3.23	310.96 ± 16.96
	Sample Once	664.88 ± 24.35	423.43 ± 50.98	751.21 ± 3.44	309.56 ± 17.33
	Use Mean	693.86 ± 12.21	417.32 ± 31.97	749.13 ± 3.41	310.70 ± 17.01



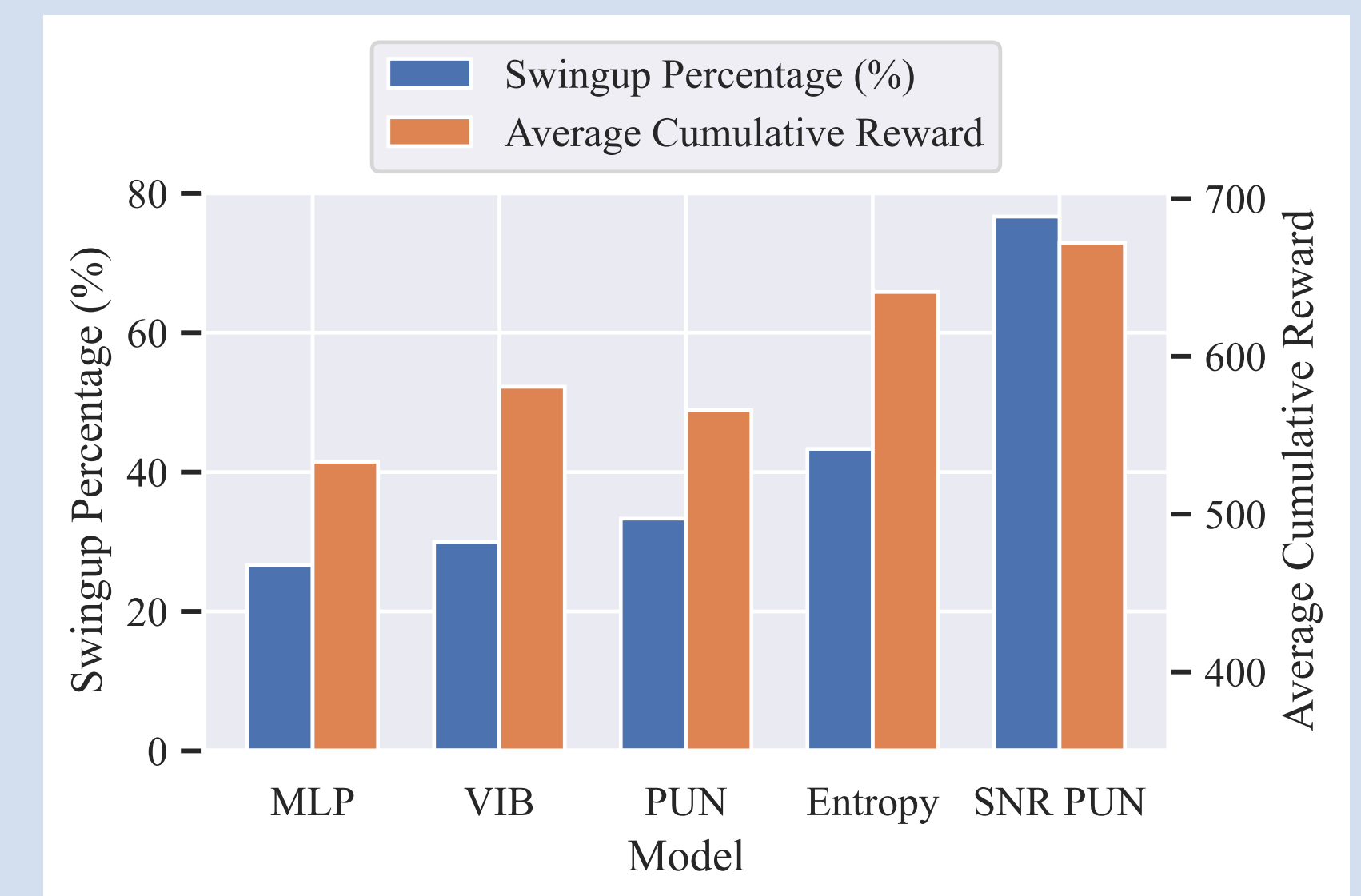
4 Cartpole Case Study Results

We find that in general an unconstrained PUN performs better than a standard MLP but in certain cases does not. By including the SNR loss term many of these cases disappear.



5 Transfer to Real-world Robotic Platform

We find that the improvements in simulations still hold when transferring to a physical CartPole. Apart from maintaining a higher average cumulative reward, we also found that the SNR PUN had the highest successful swingup percentage, suggesting a more stable overall solution.



References

1. Igl, Maximilian, Kamil Ciosek, Yingzhen Li, Sebastian Tschitschek, Cheng Zhang, Sam Devlin, and Katja Hofmann. "Generalization in reinforcement learning with selective noise injection and information bottleneck." *Advances in neural information processing systems* 32 (2019).
2. Plappert, M., Houthoofd, R., Dhariwal, P., Sidor, S., Chen, R. Y., Chen, X., ... & Andrychowicz, M. (2018, February). Parameter Space Noise for Exploration. In *International Conference on Learning Representations*.
3. Gawlikowski, Jakob, Cedric Rovile Njietcheu Tassi, Mohsin Ali, Jongseok Lee, Matthias Humt, Jianxiang Feng, Anna Kruspe et al. "A survey of uncertainty in deep neural networks." *Artificial Intelligence Review* 56, no. Suppl 1 (2023): 1513-1589.

Partner/Sponsor:

