Meta-Reward-Net: Implicitly Differentiable Reward Learning for Preference-based Reinforcement Learning

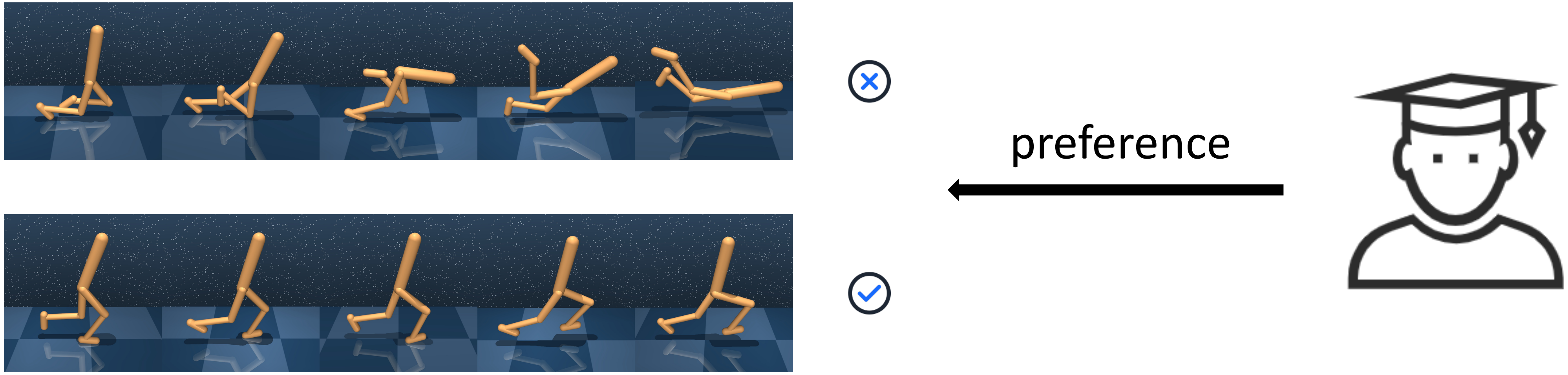
Runze Liu[1,2], Fengshuo Bai[3], Yali Du[4,†], Yaodong Yang[1,5,†]

[1]Institute for AI, Peking University, [2]Shandong University

[3]Institute of Automation, Chinese Academy of Science, [4]King's College London, [5]Beijing Institute for General AI

Preference-based RL

- Traditional RL requires a hand-engineered reward function.
- PbRL constructs a preference predictor, and optimizes the reward function through a classification task.



- Key challenge: feedback-efficiency

[1] Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences. In NeurIPS 2017.
[2] Kimin Lee, Laura M Smith, and Pieter Abbeel. PEBBLE: Feedback-efficient interactive reinforcement learning via relabeling experience and unsupervised pre-training. In ICML 2021.

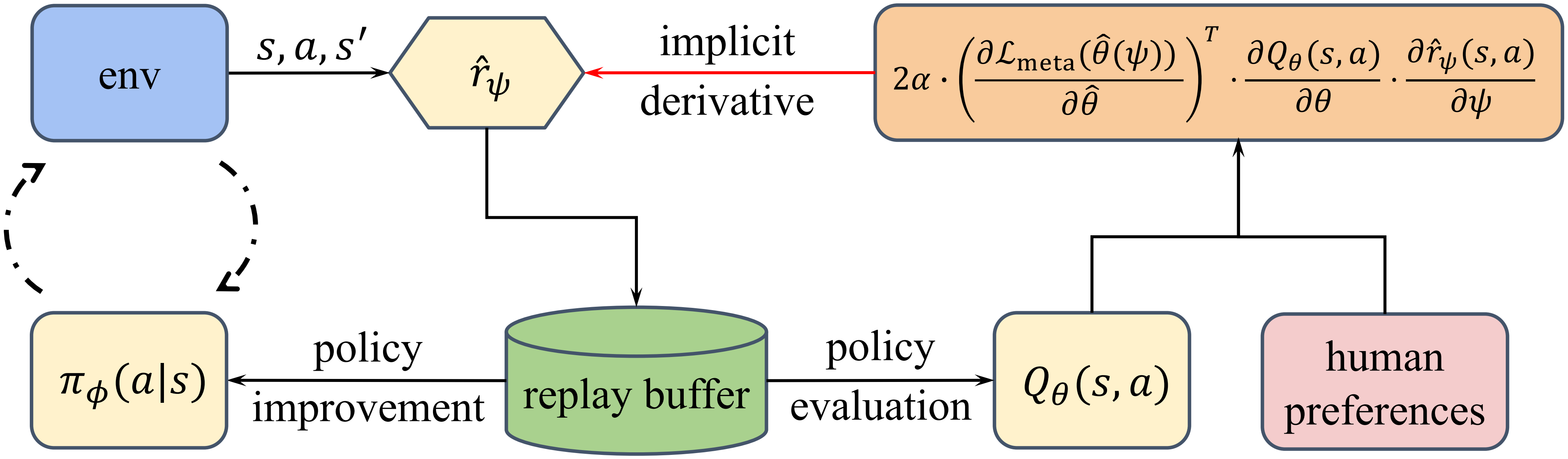
Motivation

- Confirmation bias: a network overfits to inaccurate targets predicted by another network.
- When there are few preference labels, PbRL methods will likely learn an inaccurate reward function, therefore the Q-function may overfit to the inaccurate outputs of the reward function.

[1] Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In NeurIPS 2017.
[2] Hieu Pham, Zihang Dai, Qizhe Xie, and Quoc V. Le. Meta pseudo labels. In CVPR 2021.

Meta-Reward-Net

- Main idea: consider the performance of the Q-function in the reward learning



Theoretical Results

Theorem 1. *Assume the outer loss $\mathcal{L}_{\text{meta}}$ is Lipschitz smooth with constant L, and the gradient of $\mathcal{L}_{\text{meta}}$ and J_Q is bounded by ρ. Let $\widehat{r}_\psi$ be twice differential, with its gradient and Hessian respectively bounded by δ and $\mathcal{B}$. For some $c_1 > 0$, suppose the learning rate of the inner updating $\alpha_k = \min\{1, \frac{c_1}{T}\}$, where $c_1 < T$. For some $c_2 > 0$, suppose the learning rate of the outer updating $\beta_k = \min\{\frac{1}{L}, \frac{c_2}{\sqrt{T}}\}$, where $\frac{\sqrt{T}}{c_2} \geq L$, $\sum_{k=1}^{\infty} \beta_k \leq \infty$ and $\sum_{k=1}^{\infty} \beta_k^2 \leq \infty$. Meta-Reward-Net can achieve:*

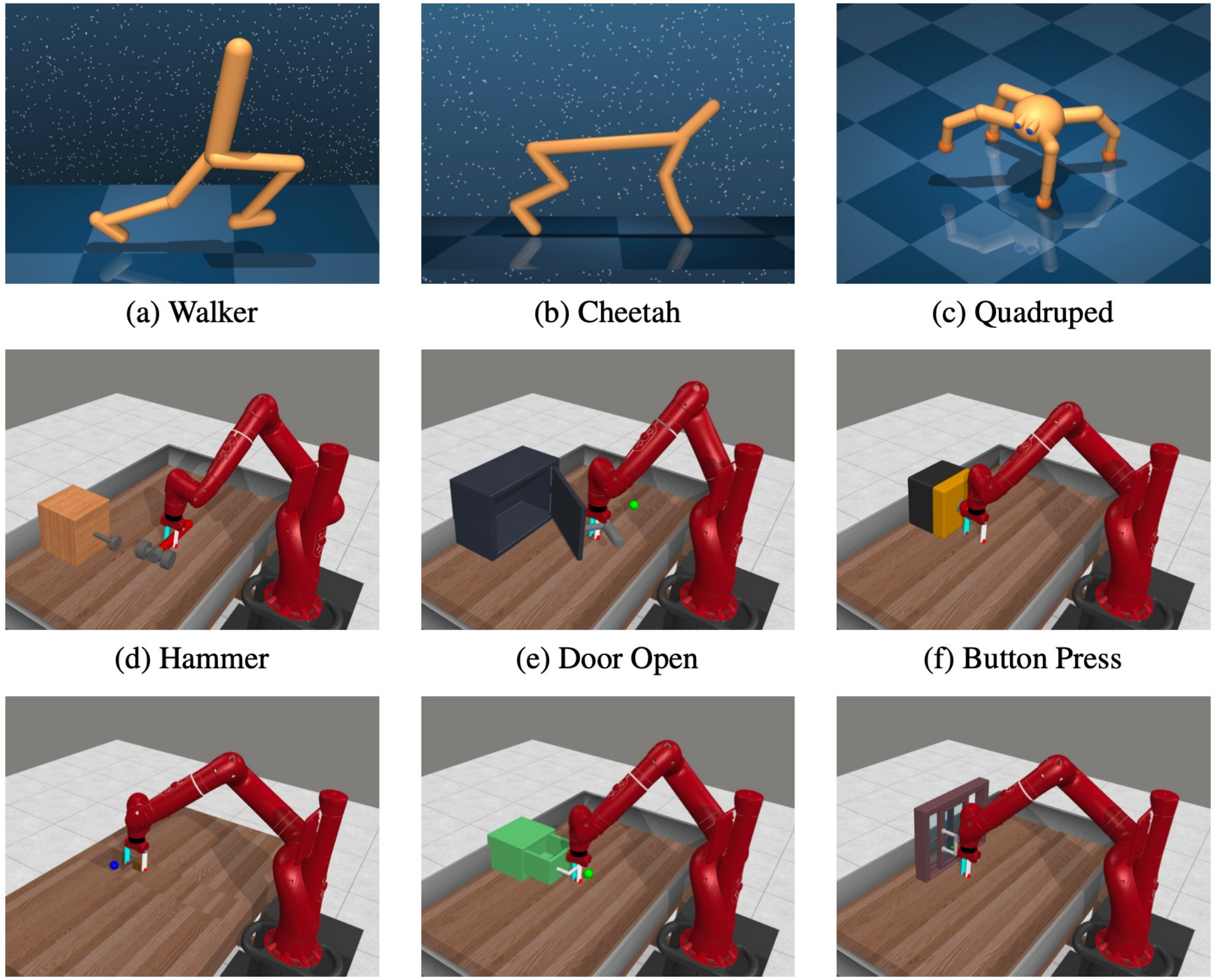
$$\min_{1 \leq k \leq T} \mathbb{E}\left[\left\|\nabla_\psi \mathcal{L}_{\text{meta}}(\hat{\theta}^{(k)}(\psi^{(k)}))\right\|^2\right] \leq \mathcal{O}\left(\frac{1}{\sqrt{T}}\right).$$

Theoretical Results

Theorem 2. *Assume the outer loss $\mathcal{L}_{\text{meta}}$ is Lipschitz smooth with constant L, and the gradient of $\mathcal{L}_{\text{meta}}$ and J_Q is bounded by ρ. Let $\widehat{r}_\psi$ be twice differential, with its gradient and Hessian respectively bounded by δ and $\mathcal{B}$. For some $c_1 > 0$, suppose the learning rate of the inner updating $\alpha_k = \min\{1, \frac{c_1}{T}\}$, where $c_1 < T$. For some $c_2 > 0$, suppose the learning rate of the outer updating $\beta_k = \min\{\frac{1}{L}, \frac{c_2}{\sqrt{T}}\}$, where $\frac{\sqrt{T}}{c_2} \geq L$, $\sum_{k=1}^{\infty} \beta_k \leq \infty$ and $\sum_{k=1}^{\infty} \beta_k^2 \leq \infty$. Meta-Reward-Net can achieve:*

$$\lim_{k \to \infty} \mathbb{E}\left[\left\|\nabla_\theta J_Q(\theta^{(k)}; \psi^{(k+1)})\right\|^2\right] = 0.$$

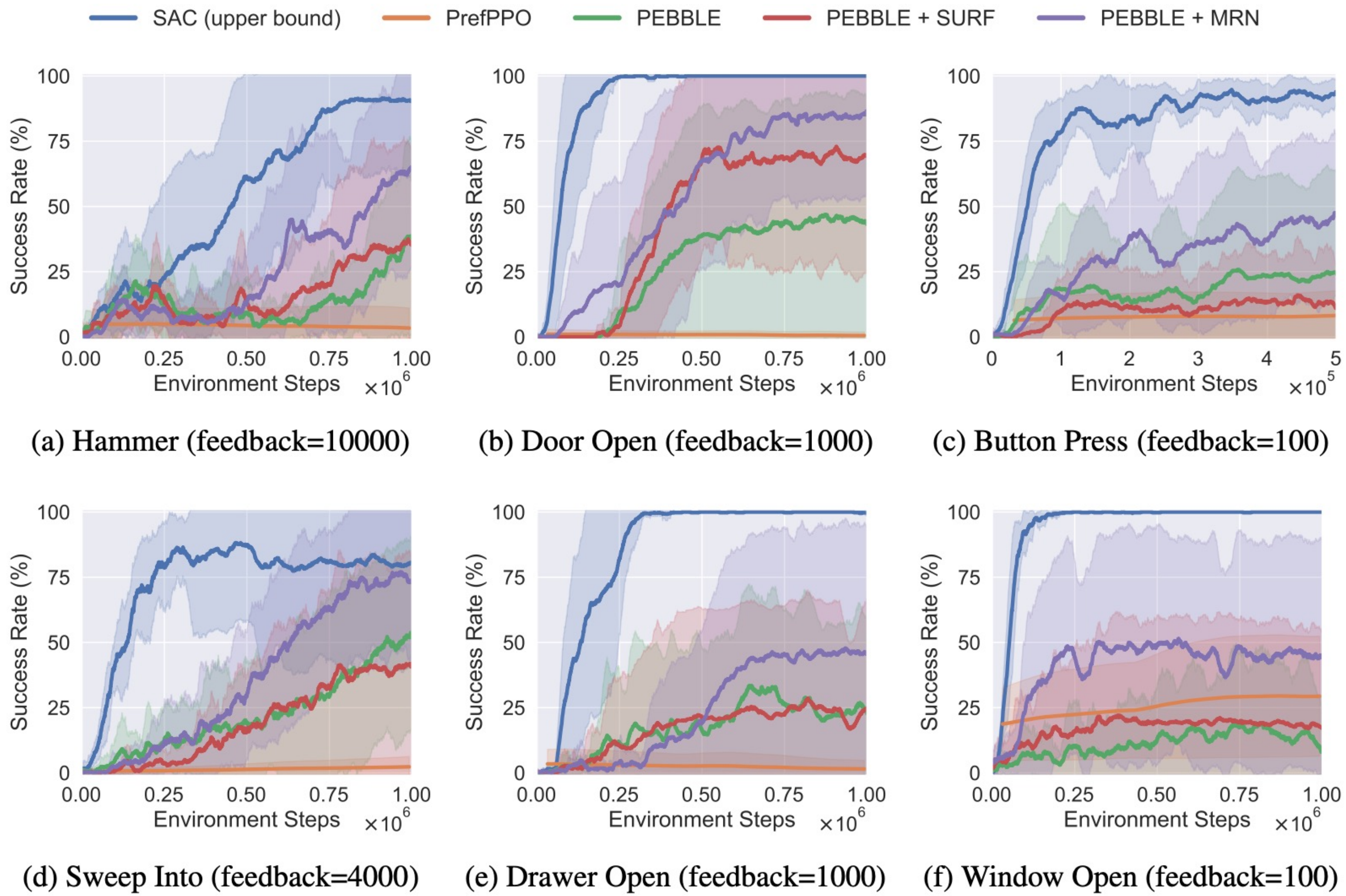
Experiments



(a) Walker (b) Cheetah (c) Quadruped
(d) Hammer (e) Door Open (f) Button Press
(g) Sweep Into (h) Drawer Open (i) Window Open

[1] Tianhe Yu, Deirdre Quillen, Zhanpeng He, Ryan Julian, Karol Hausman, Chelsea Finn, and Sergey Levine. Meta-world: A benchmark and evaluation for multi-task and meta reinforcement learning. In CoRL 2020.
[2] Yuval Tassa, Yotam Doron, Alistair Muldal, Tom Erez, Yazhe Li, Diego de Las Casas, David Budden, Abbas Abdolmaleki, Josh Merel, Andrew Lefrancq, et al. Deepmind control suite. arXiv preprint arXiv:1801.00690, 2018.
[3] Saran Tunyasuvunakool, Alistair Muldal, Yotam Doron, Siqi Liu, Steven Bohez, Josh Merel, Tom Erez, Timothy Lillicrap, Nicolas Heess, and Yuval Tassa. dm_control: Software and tasks for continuous control. Software Impacts, 6:100022, 2020.

Experiments

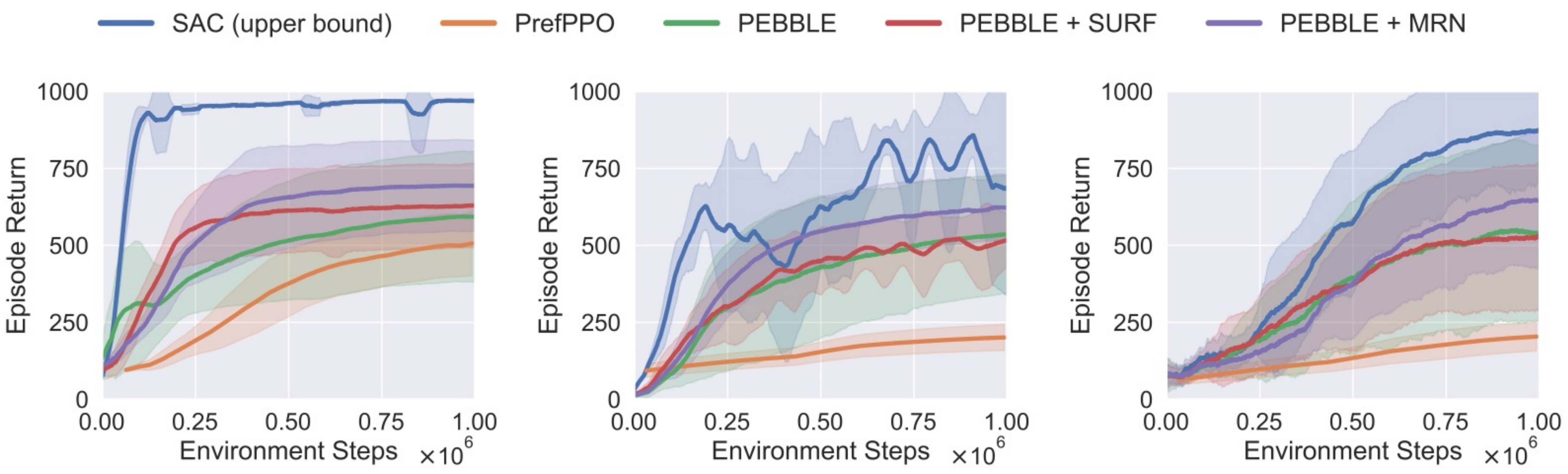


(a) Hammer (feedback=10000) (b) Door Open (feedback=1000) (c) Button Press (feedback=100)

(d) Sweep Into (feedback=4000) (e) Drawer Open (feedback=1000) (f) Window Open (feedback=100)

[1] Tuomas Haarnoja, Aurick Zhou, Pieter Abbeel, and Sergey Levine. Soft actor-critic: Off- policy maximum entropy deep reinforcement learning with a stochastic actor. In ICML 2018.
[2] Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences. In NeurIPS 2017.
[3] Kimin Lee, Laura M Smith, and Pieter Abbeel. PEBBLE: Feedback-efficient interactive reinforcement learning via relabeling experience and unsupervised pre-training. In ICML 2021.
[4] Jongjin Park, Younggyo Seo, Jinwoo Shin, Honglak Lee, Pieter Abbeel, and Kimin Lee. SURF: Semi-supervised reward learning with data augmentation for feedback-efficient preference-based reinforcement learning. In ICLR 2022.

Experiments



(a) Walker (feedback=100) (b) Cheetah (feedback=100) (c) Quadruped (feedback=700)

[1] Tuomas Haarnoja, Aurick Zhou, Pieter Abbeel, and Sergey Levine. Soft actor-critic: Off- policy maximum entropy deep reinforcement learning with a stochastic actor. In ICML 2018.
[2] Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences. In NeurIPS 2017.
[3] Kimin Lee, Laura M Smith, and Pieter Abbeel. PEBBLE: Feedback-efficient interactive reinforcement learning via relabeling experience and unsupervised pre-training. In ICML 2021.
[4] Jongjin Park, Younggyo Seo, Jinwoo Shin, Honglak Lee, Pieter Abbeel, and Kimin Lee. SURF: Semi-supervised reward learning with data augmentation for feedback-efficient preference-based reinforcement learning. In ICLR 2022.

Conclusion

- We propose a novel preference-based RL algorithm, Meta-Reward-Net (MRN), which considers the performance of the Q-function in reward learning with convergence guarantee.
- We demonstrate that MRN outperforms preference-based RL baselines on several complex control tasks and improves the feedback efficiency.

Thank you!