Multi-Source Policy Aggregation in Heterogeneous and Private Environmental Dynamics

Mohammadamin Barekatain[1,2] Ryo Yonetani[1] Masashi Hamaya[1]

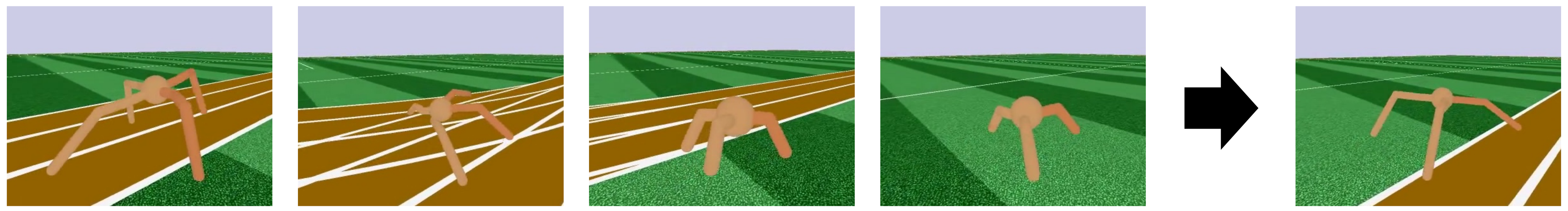
1 OMRON SINIC X 2 Technical University of Munich

Problem: Transfer RL between heterogeneous and private dynamics

- Without getting access to source environmental dynamics
- Learning a target policy efficiently from the policies acquired in the source envs
- No prior knowledge about source policies (differentiable or not, how optimal they are)

Example: Robotic ants with different leg designs

- Just reusing source policies won't work for a new robot with different legs

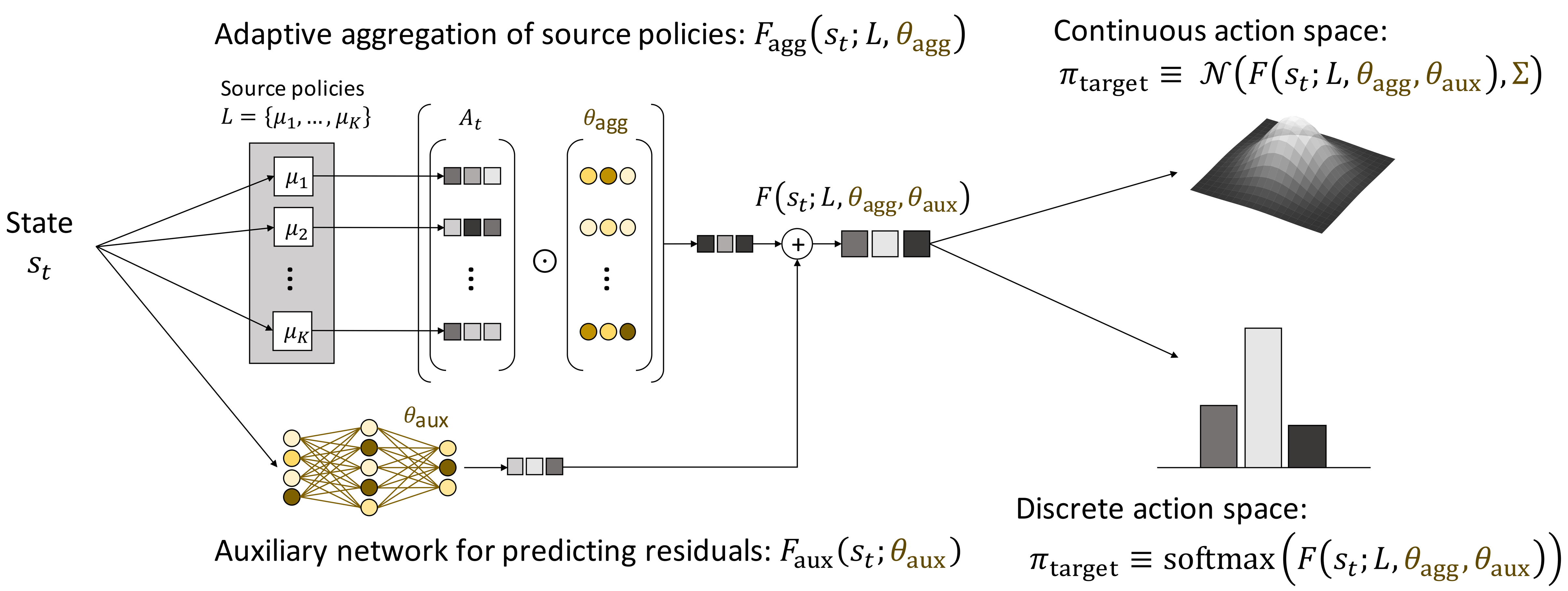


Our Solution: Learning to aggregate source policies ``adaptively''

- Agnostic to source environmental dynamics and performances of source policies
- Enabling sample-efficient training for a variety of environments

MULTI-source POLicy AggRegation (MULTIPOLAR)

- Aggregating source policies adaptively to maximize expected returns
- Auxiliary network predicting residuals around aggregated actions
- Working on both of continuous and discrete action spaces



Full paper available at arXiv

Call for interns: Robotics (2019/12 ~) CV/ML (2020/04 ~)

Contact: Ryo Yonetani ryo.yonetani@sinicx.com

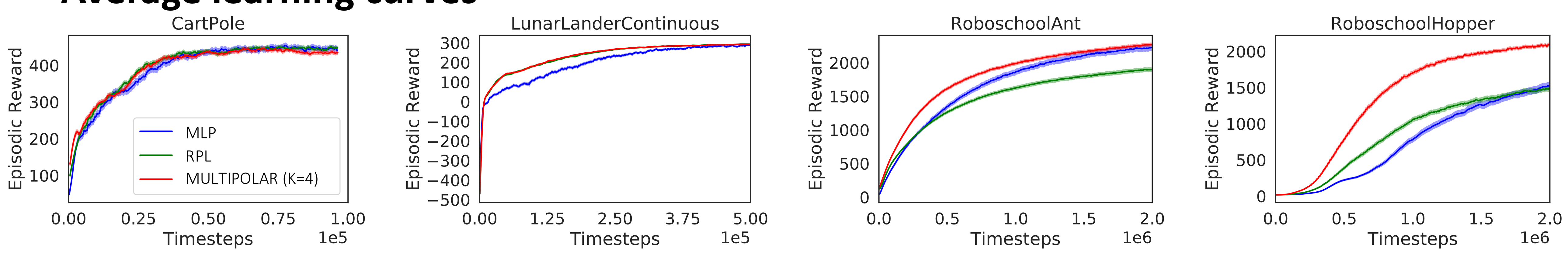
Related Work (see our full paper for more details):

- **Transfer RL between different dynamics:** typically requires full access to source environment MDPs or state sequences sampled from the source environments
- **Meta-RL:** requires a target policy to be trained over a distribution of environments
- **Leveraging multiple policies:** requires the source policies to be trained in a single env (e.g., option frameworks) or in the same environmental dynamics (e.g., policy reuse)
- **No prior work can transfer knowledge from the only policies obtained in diverse and unknown environmental dynamics**

Experimental Evaluations

- A variety of environments provided in OpenAI Gym, from classic control problems to challenging robotic simulations
- Randomly modifying kinematics and dynamics to create diverse environment instances
  - E.g. link lengths, link mass, damping factor, friction
- Baselines: MLP trained from scratch | RPL (residual policy learning that learns residuals around actions from a single source policy)

Average learning curves



Average episodic rewards over various training samples

Methods	CartPole 25K	50K	75K	100K
MLP	171 (164,179)	229 (220,237)	266 (258,275)	291 (282,300)
RPL	185 (179,192)	238 (231,245)	269 (262,276)	289 (282,296)
MULTIPOLAR (K=4)	**202 (195,209)**	**252 (245,260)**	**283 (276,290)**	**299 (292,306)**

Methods	LunarLander 125K	250K	375K	500K
MLP	10 (2,18)	112 (104,121)	178 (171,185)	216 (210,221)
RPL	92 (87,96)	178 (174,182)	223 (220,226)	246 (243,248)
MULTIPOLAR (K=4)	**95 (90,99)**	**181 (177,185)**	**224 (221,228)**	**246 (244,249)**

Methods	Roboschool Ant 0.5M	1M	1.5M	2M
MLP	714 (674,756)	1088 (1030,1146)	1332 (1267,1399)	1500 (1430,1572)
RPL	807 (785,830)	1120 (1088,1152)	1307 (1269,1344)	1432 (1391,1473)
MULTIPOLAR (K=4)	**1025 (995,1056)**	**1397 (1361,1432)**	**1606 (1568,1644)**	**1744 (1705,1783)**

Methods	Roboschool Hopper 0.5M	1M	1.5M	2M
MLP	26 (25,27)	43 (42,45)	67 (64,70)	92 (88,96)
RPL	37 (36,39)	75 (70,79)	114 (107,121)	152 (142,160)
MULTIPOLAR (K=4)	**61 (59,64)**	**138 (132,143)**	**213 (206,221)**	**283 (273,292)**