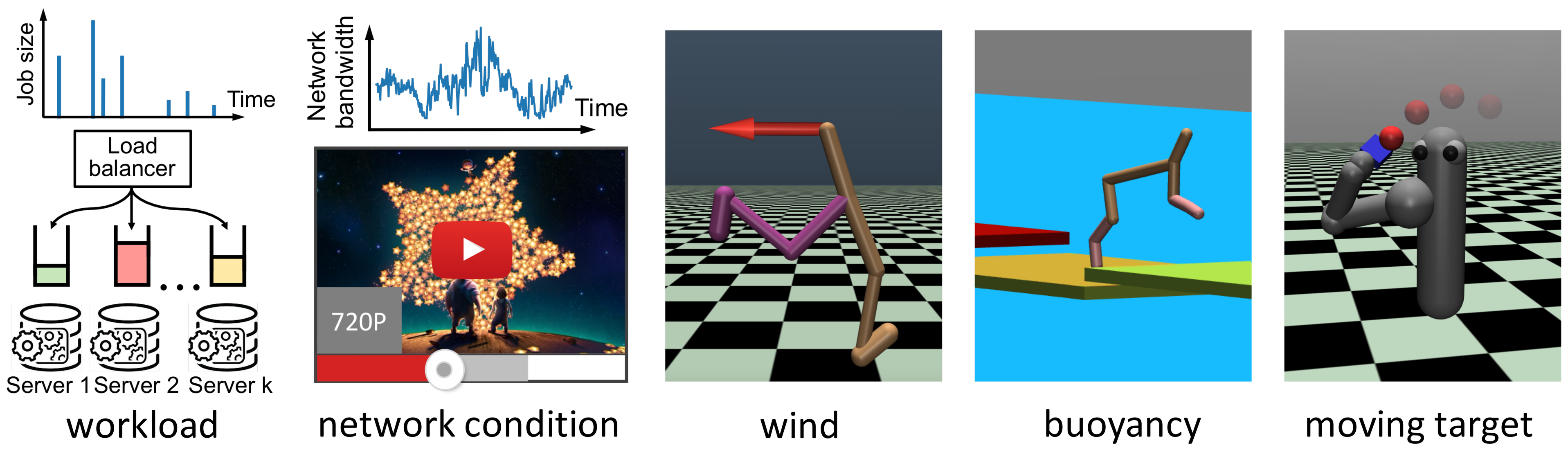
Variance Reduction for Reinforcement Learning in Input-Driven Environments

Hongzi Mao   Shaileshh Bojja Venkatakrishnan   Malte Schwarzkopf   Mohammad Alizadeh
MIT Computer Science and Artificial Intelligence Laboratory

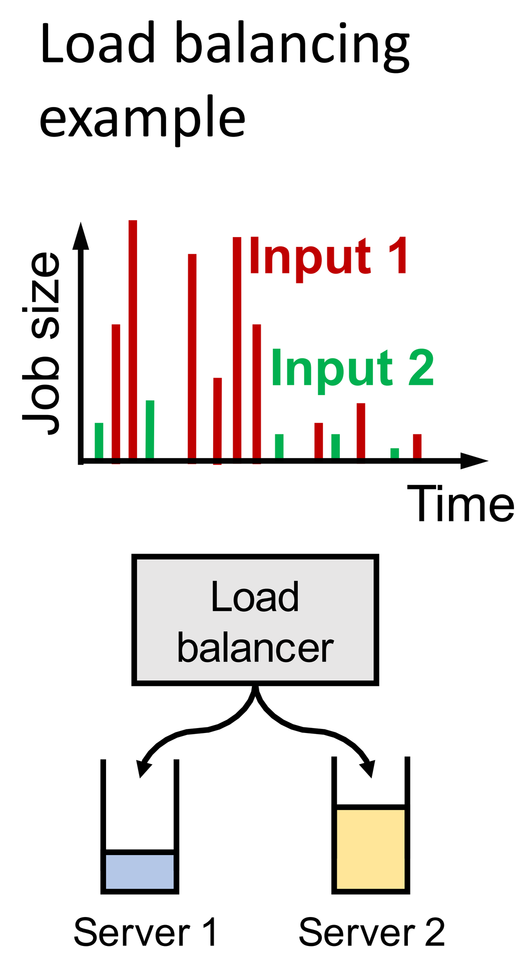
Motivation

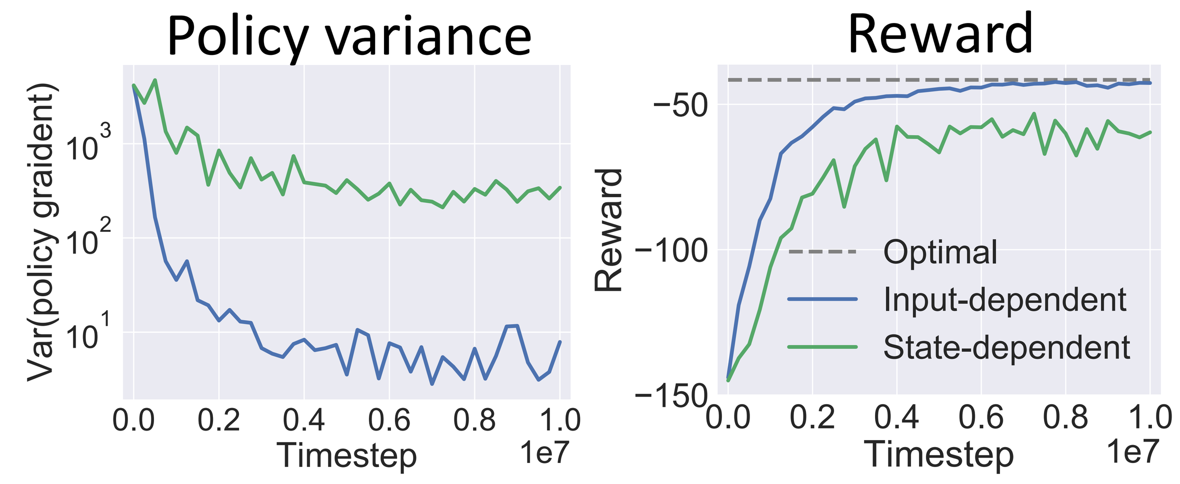


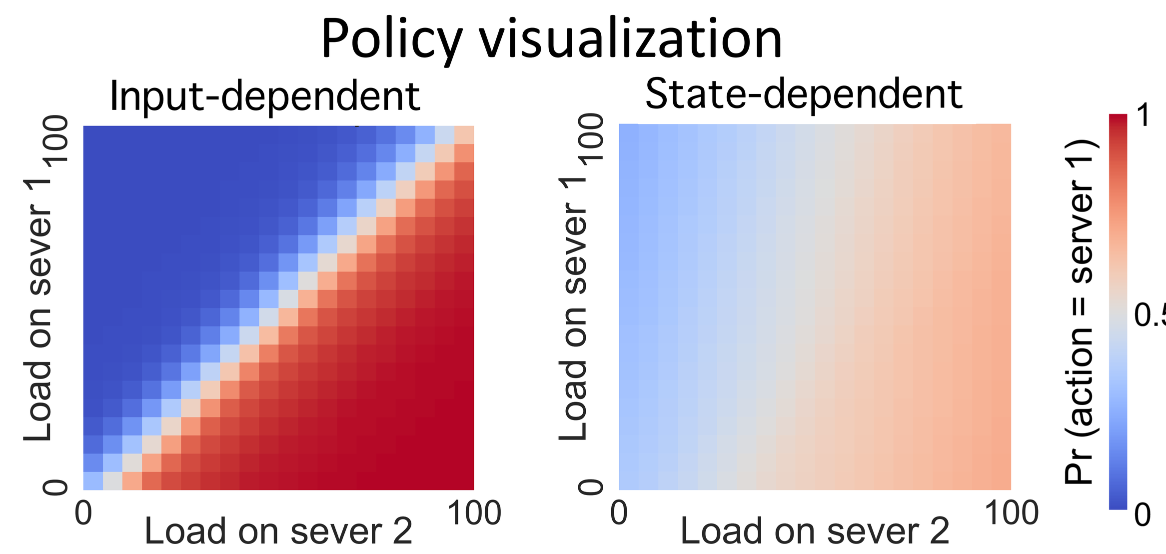
Environments with exogenous, stochastic **input processes** that affect the dynamics

Since the reward is partially dictated by the input process, the state alone only provides limited information to estimate the average return. Thus, policy gradient methods with standard state-dependent baselines suffer from high variance.

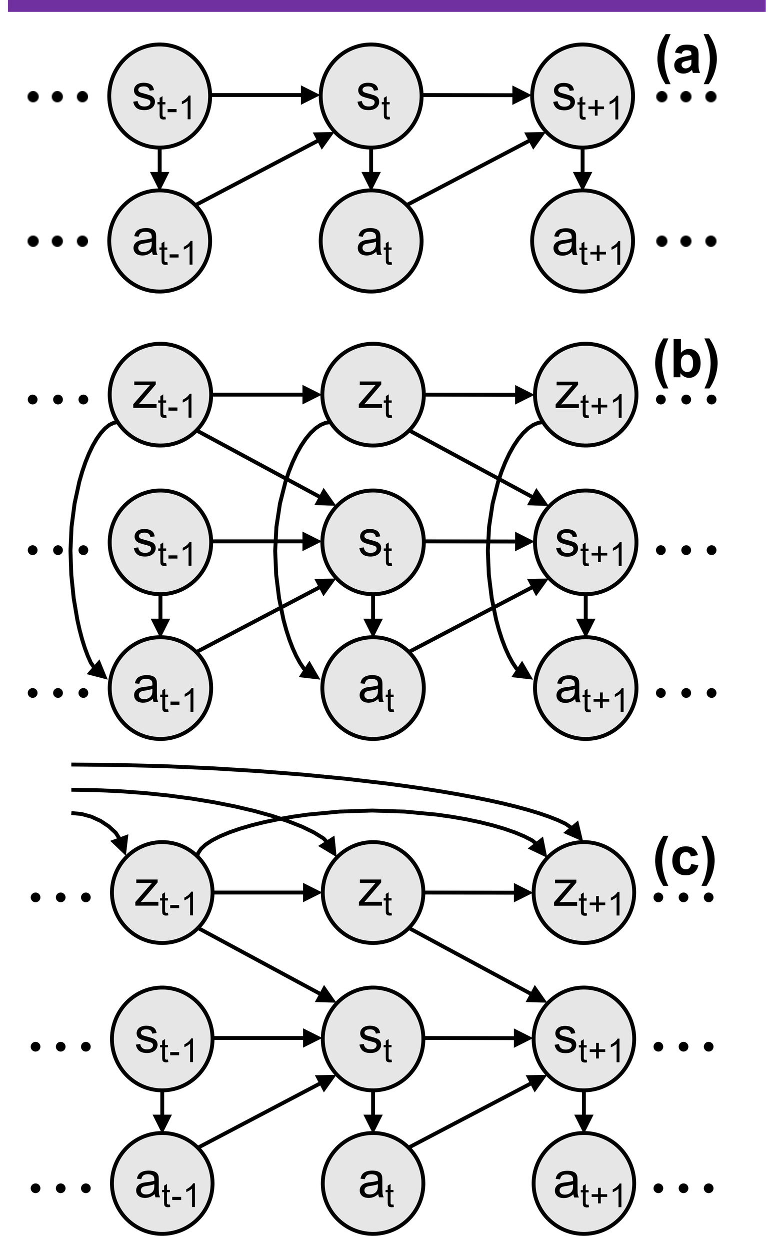
Example







Input-Driven Processes



(a)	Standard MDP
(b)	Input-Driven MDP
(c)	Input-Driven POMDP

Input-Dependent Baselines

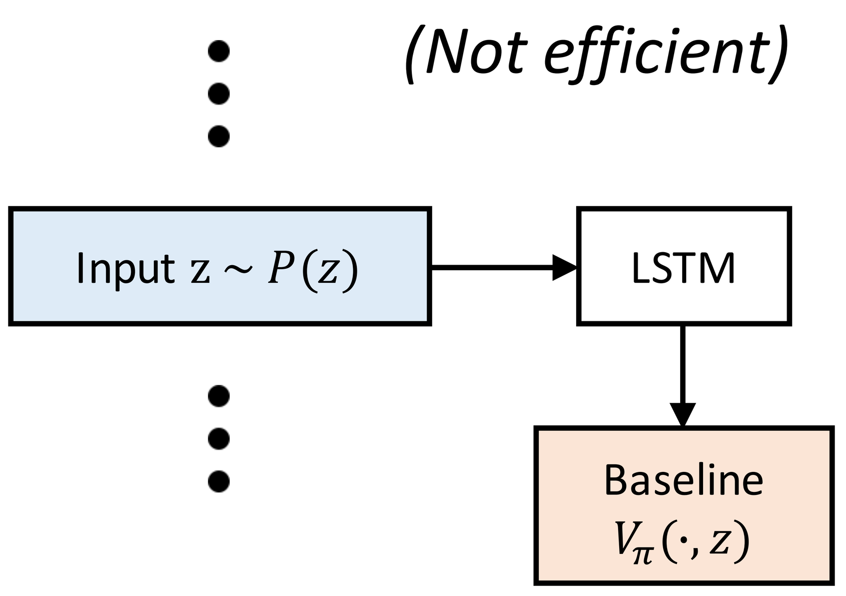
State-dependent baseline: $b(s_t) = V(s_t),\ \forall z_{t:\infty}$

Input-dependent baseline: $b(s_t, z_{t:\infty}) = V(s_t \mid z_{t:\infty})$

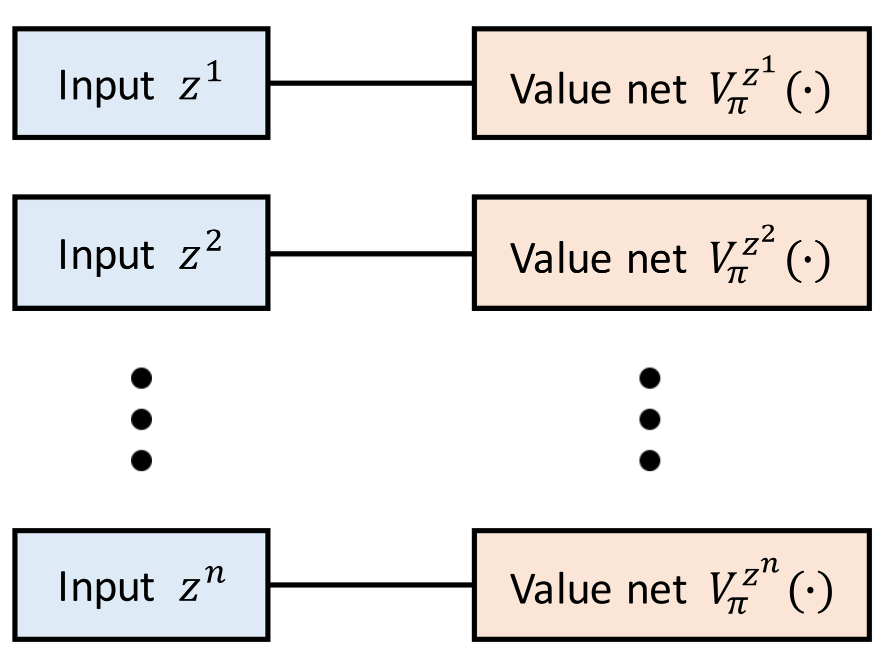
Depend on the entire future input sequence $\{z_t, z_{t+1}, \ldots, z_\infty\}$ during training

Input-dependent baselines are *bias-free* for policy gradients: $\mathbb{E}\left[\nabla_\theta \log \pi_\theta(a_t|s_t) b(s_t, z_{t:\infty})\right] = 0$

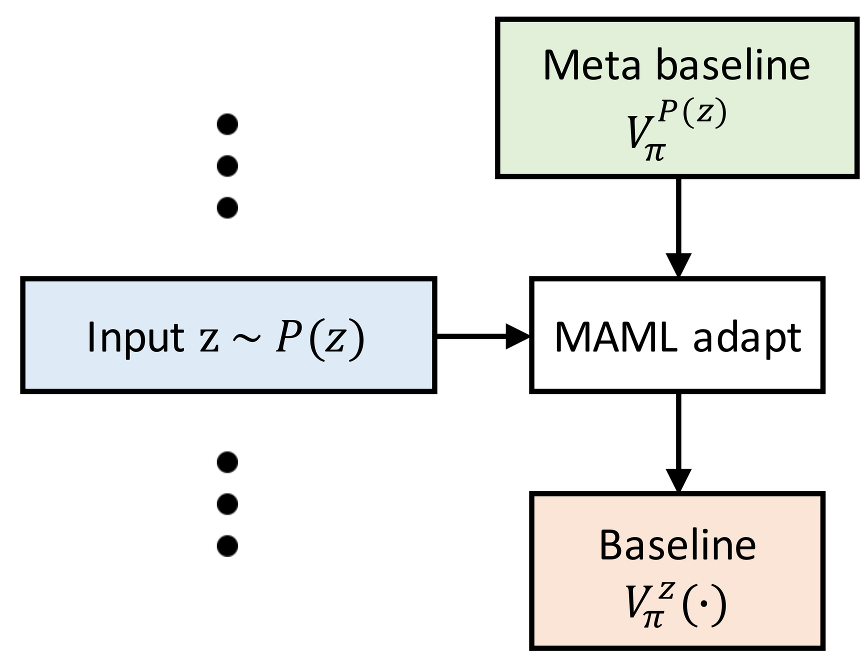
Implementations of input-dependent baselines:



Sample $z \sim P(z)$, use LSTM to compute $V_\pi(\cdot, z)$ for policy gradient: $\mathbb{E}[\nabla_\theta \log \pi_\theta(a|s)(R(s) - V_\pi(s,z))]$



Sample $z^i \sim \{z^1, z^2, \ldots, z^n\}$, use the corresponding value net for policy gradient: $\mathbb{E}[\nabla_\theta \log \pi_\theta(a|s)(R(s) - V_\pi^{z^i}(s))]$

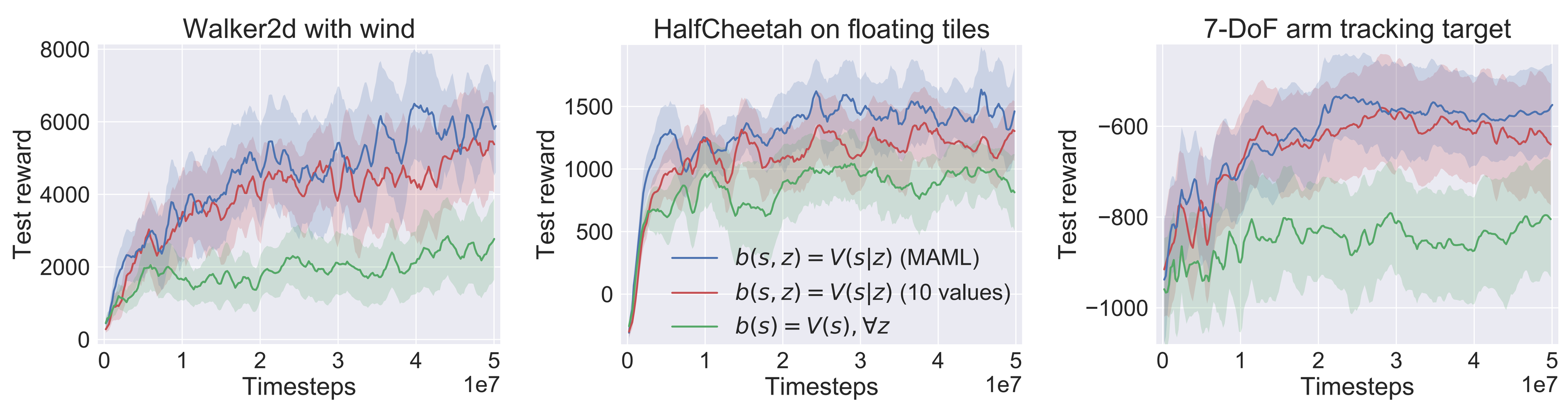


Sample $z \sim P(z)$ and trajectories, then use MAML to adapt meta baseline $V_\pi^{P(z)}$ for policy gradient: $\mathbb{E}[\nabla_\theta \log \pi_\theta(a|s)(R(s) - V_\pi^{z}(s))]$

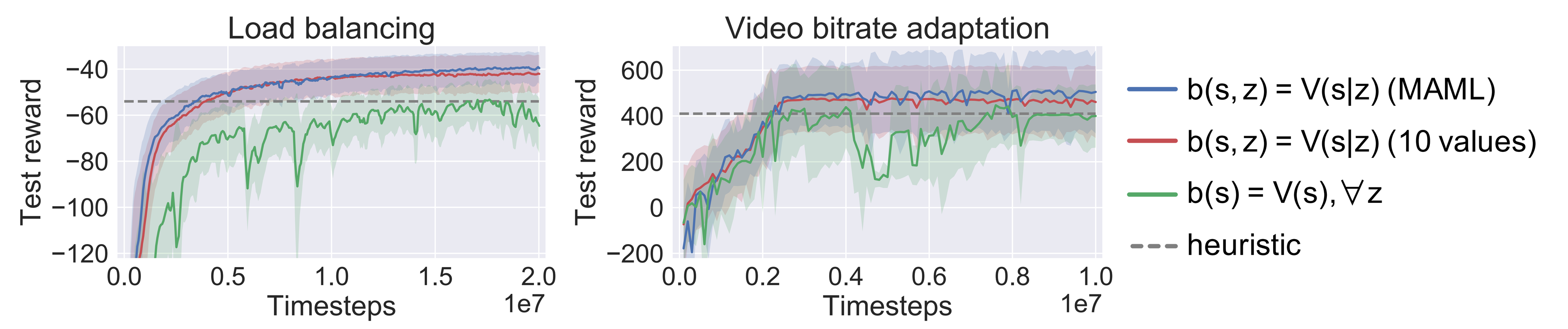
Experiments

Input-dependent baselines are applicable to many policy gradient methods, such as **A2C**, **TRPO**, **PPO**, and they are complementary and orthogonal to robust adversarial RL methods such as **RARL** (Pinto et al., 2017) and meta-policy optimization such as **MPO** (Clavera et al., 2018).

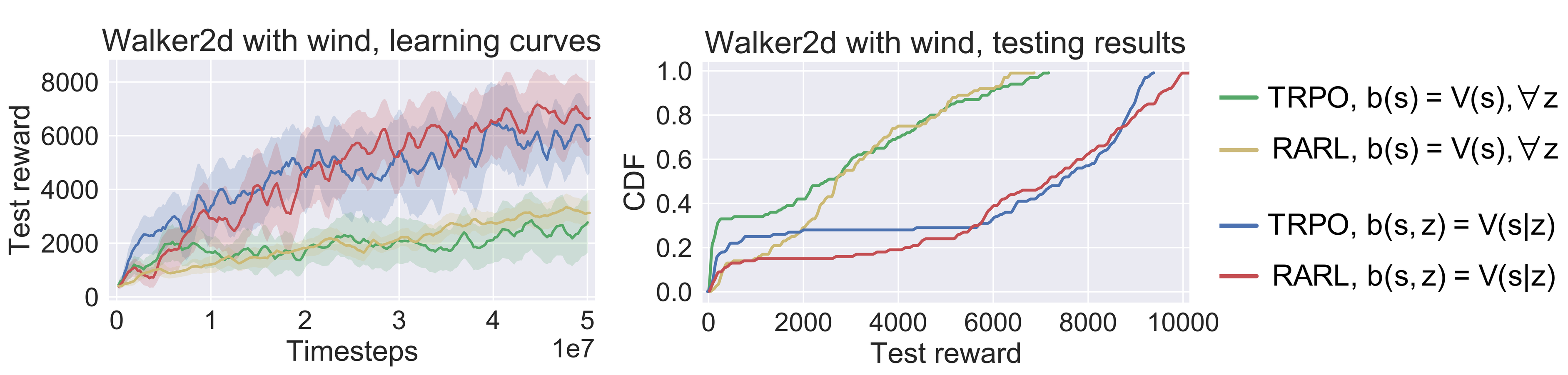
TRPO



A2C



Robust Adversarial RL



Meta-Policy Optimization