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.

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: \( b(s_t, z_{t,\infty}) = V(s_t | z_{t,\infty}) \)

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:

\( z \sim P(z) \)
- Sample \( z \sim P(z) \), use LSTM to compute \( V_{\phi}(z, s) \) for policy gradient: \( \mathbb{E} \left[ \nabla_{\phi} \log \pi_{\phi}(a | s) (R(s_t) - V_{\phi}(s, z)) \right] \)
- Sample \( z \sim P(z) \), trajectory, then use MAML to adapt meta baseline \( V_{\phi}^{P(z)}(z) \) for policy gradient: \( \mathbb{E} \left[ \nabla_{\phi} \log \pi_{\phi}(a | s) (R(s) - V_{\phi}^{z}(s)) \right] \)

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
- Walker2d with wind
- HalfCheetah on sliding tiles
- 7.09F arm tracking target
- Load balancing
- Video bitrate adaptation

A2C
- Walker2d with wind, learning curves
- Walker2d with wind, testing results

Robust Adversarial RL
- Walker2d with wind, learning curves
- Walker2d with wind, testing results

Meta-Policy Optimization
- Walker2d with wind, learning curves
- Walker2d with wind, policy adaptation