A Multi-Fidelity Control Variate Approach for Policy Gradient Estimation

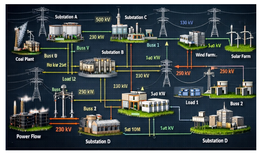
Xinjie Liu*, Cyrus Neary*, Kushagra Gupta, Wesley A. Suttle, Christian Ellis, Ufuk Topcu, and David Fridovich-Keil
***Equal contribution**

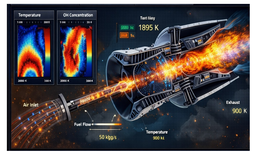
Transactions on Machine Learning Research (TMLR)
J2C Certification (selected to NeurIPS/ICLR/ICML J2C Track)

https://xinjie-liu.github.io/mfpg-rl/

Challenge: Data Scarcity in Reinforcement Learning (RL)

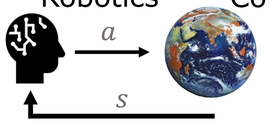
- Online RL algorithms require excessive interaction with **real environments/high-fidelity simulations**





Power systems | Robotics | Combustion simulation

High-Fidelity:

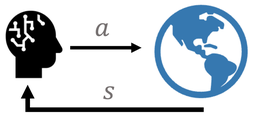


👎 **expensive, slow, even unsafe … but** ✅ **accurate**

≠

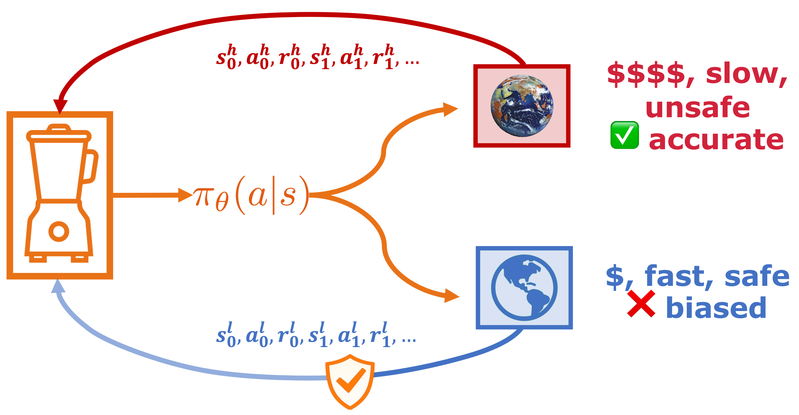
👍 **cheap, fast, safe … but** ❌ **biased**

Low-Fidelity:



- **Low-fidelity simulation** tools provide cheap, **abundant**, but **biased** data: reduced-order models, generative world models, heuristic reward functions, digital twins, etc.

Q: How can we enable sample-efficient RL in the real world by mixing multi-fidelity data, while being robust to low-fidelity data biases?



… How do we build the blender?

Preliminaries: Policy Gradient Methods

- **Objective**: Learn a policy to maximize the **high-fidelity environment** performance

$$\max_\theta J_\theta = \mathbb{E}\left[\sum_{t=0}^{T} \gamma^t r_t^h \;\middle|\; \tau^h \sim \mathcal{M}^h(\pi_\theta)\right]$$

- **Strategy**: Gradient ascent $\theta_{k+1} = \theta_k + \alpha \nabla_{\theta_k} J_{\theta_k}$

$$\nabla_\theta J_\theta \approx \nabla_\theta \frac{1}{N^h} \sum_{i=1}^{N^h} X_{\tau_i^h}^{\pi_\theta}$$

Random variable (R.V.): per-trajectory contribution to policy gradient

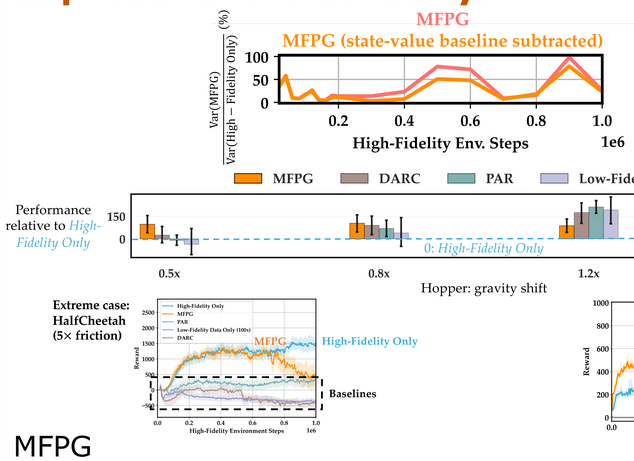
REINFORCE $X_{\tau_i}^{\pi_\theta} = \frac{1}{T}\sum_{t=0}^{T-1}\big(G_t - V_\phi(s_t)\big)\log(\pi_\theta(a_t|s_t))$

PPO $X_{\tau_i}^{\pi_\theta} = \frac{1}{T}\sum_{t=0}^{T-1}\min\{\frac{\pi_\theta(a_t|s_t)}{\pi_{old}(a_t|s_t)}A_\phi(s_t,a_t), clip[\frac{\pi_\theta(a_t|s_t)}{\pi_{old}(a_t|s_t)}, 1-\epsilon, 1+\epsilon]A_\phi(s_t,a_t)\}$

…

- **Challenge**: High-fidelity data **scarcity** (small N^h) causing **high estimation variance** for $\mathbb{E}\big[X_{\tau^h}^{\pi_\theta}\big]$ and slow convergence
- **Idea**: **Ground** learning in **high-fidelity samples** (**unbiased**); use abundant **low-fidelity samples** to **reduce variance** for policy gradient estimation

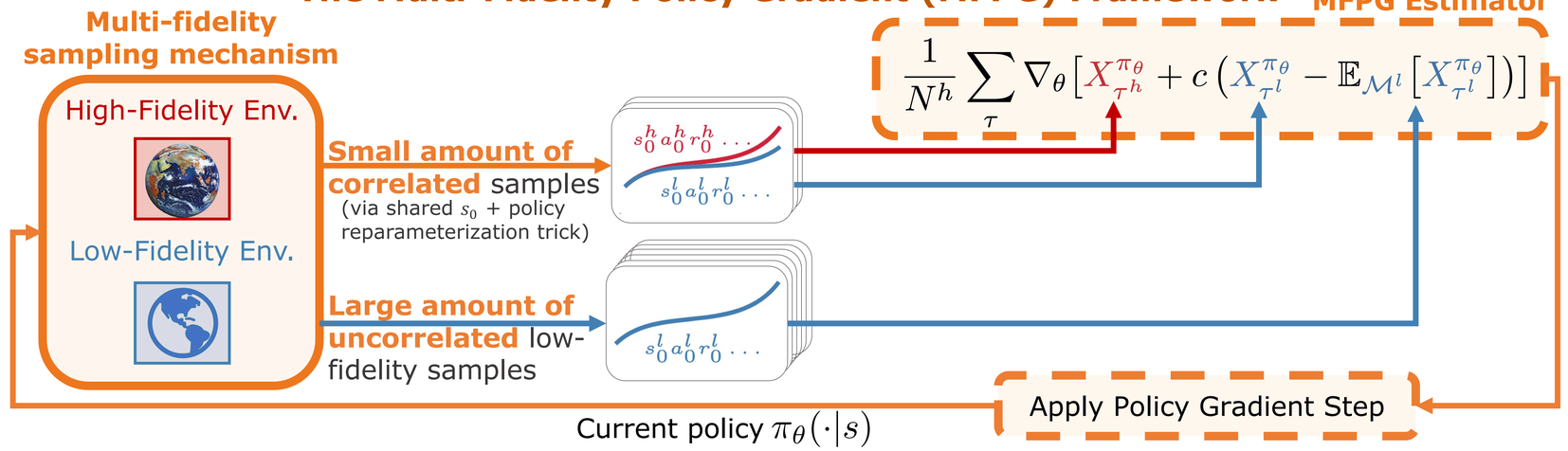
Experimental Takeaways



MFPG

- Can achieve **substantially greater variance reduction** than single-fidelity baseline subtraction
- Handles both **dynamics** and **reward** gaps, even when the low-fidelity environment is **anti-correlated**
- Demonstrates the **strongest consistency** and **robustness** compared to off-dynamics RL baselines, thanks to MFPG's **unbiasedness**

The Multi-Fidelity Policy Gradient (MFPG) Framework



- The **MFPG estimator** mixes **scarce high-fidelity samples** with **abundant low-fidelity samples**
  - Uses abundant low-fidelity samples to form a **control variate** that **corrects** high-fidelity estimates
- Instantiate MFPG with an existing policy gradient method, e.g., REINFORCE $X_{\tau_i}^{\pi_\theta} = \frac{1}{T}\sum_t \big(G_t - V_\phi(s_t)\big)\log(\pi_\theta(a_t|s_t))$

- Properties of the multi-fidelity estimator $Z^{\pi_\theta}(c) := X_{\tau^h}^{\pi_\theta} + c\big(X_{\tau^l}^{\pi_\theta} - \mathbb{E}_{\mathcal{M}^l}\big[X_{\tau^l}^{\pi_\theta}\big]\big)$:
  - **Unbiasedness**: $\mathbb{E}_{\mathcal{M}^h}[Z^{\pi_\theta}(c)] = \mathbb{E}_{\mathcal{M}^h}\big[X_{\tau^h}^{\pi_\theta}\big]$
  - **Reduced variance**: $\mathrm{Var}(Z^{\pi_\theta}(c^\star)) = \big(1-\rho^2\big(X_{\tau^h}^{\pi_\theta}, X_{\tau^l}^{\pi_\theta}\big)\big)\,\mathrm{Var}\big(X_{\tau^h}^{\pi_\theta}\big)$ (driven by **correlation** $\rho^2\big(X_{\tau^h}^{\pi_\theta}, X_{\tau^l}^{\pi_\theta}\big)$)
  - **Faster finite-sample convergence** of MFPG-REINFORCE than the standard, single-fidelity baseline

$$\rho^2\big(X_{\tau^h}^{\pi_\theta}, X_{\tau^l}^{\pi_\theta}\big) \uparrow \implies \mathrm{Var}(Z^{\pi_\theta}(c^\star)) \quad \text{faster MFPG convergence}$$