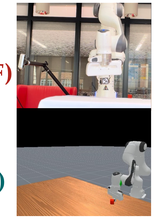
Multi-Fidelity Policy Gradients Stabilize Data-Scarce Reinforcement Learning

TEXAS UBC Xinjie Liu, Ruihan Zhao, Anirban Chaudhuri, Cyrus Neary, Ufuk Topcu, David Fridovich-Keil

Web: https://xinjie-liu.github.io/mfpg-ppo/

TL;DR

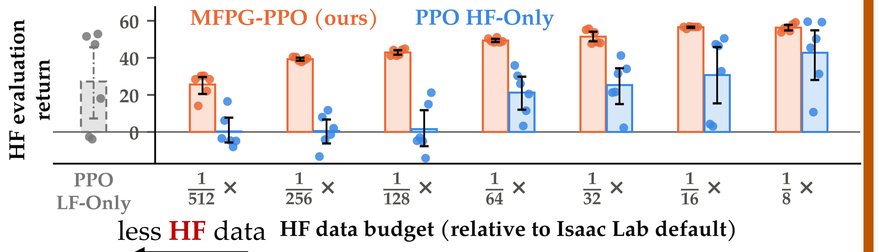


MFPG-PPO stabilizes RL training when **target-domain data** are **scarce** by **mixing real-and-sim data**.

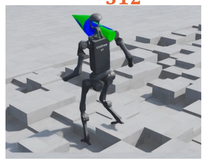
MFPG-PPO forms **multi-fidelity policy gradient (MFPG)** estimates from mixed data. MFPG gives an (approximately) unbiased **HF** policy gradient estimate, so it is **robust** to **LF simulation bias**.

Data Scarcity Challenge

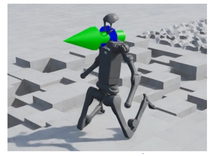
Policy gradient methods such as **PPO** break down when target-domain **HF** data are expensive and scarce due to noisy gradient estimates.



MFPG-PPO, $\frac{1}{512}\times$HF data



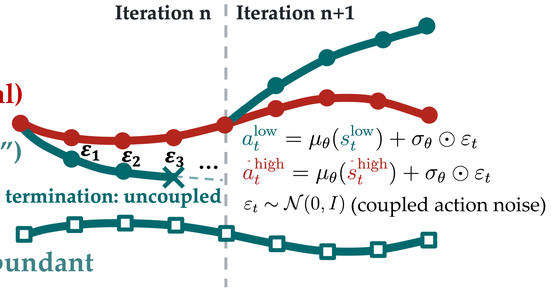
PPO HF-Only, $\frac{1}{128}\times$HF data



Multi-Fidelity Policy Gradients for PPO

MFPG-PPO constructs a **control variate**—a correlated auxiliary random variable with a known mean—for the HF policy gradient estimator.

1. Collect **three batches of transitions**



2. Construct a **control variate estimator** $\nabla_\theta L^{\text{MF}}(\theta)$ for PPO updates

$$L^{\text{MF}}(\theta) = \frac{1}{|\mathcal{P}|}\sum_{(k,t)\in\mathcal{P}}\left[X_{k,t}^{\text{high}}(\theta) + c\left(X_{k,t}^{\text{low}}(\theta) - \hat{\mu}^{\text{low}}(\theta)\right)\right],\quad \hat{\mu}^{\text{low}}(\theta) = \frac{1}{|\mathcal{U}|}\sum_{\substack{(k,t)\in\mathcal{U}\\ |\mathcal{U}|\gg|\mathcal{P}|}} X_{k,t}^{\text{low}}(\theta)$$

HF · Coupled LF · Uncoupled LF

$X_{k,t}(\theta)$: PPO surrogate Variance-minimizing coefficient: $c^* = -\dfrac{\text{Cov}\left(X^{\text{high}}, X^{\text{low}}\right)}{\text{Var}\left(X^{\text{low}}\right)}$

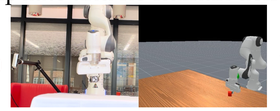
The term $c\left(X_{k,t}^{\text{low}}(\theta) - \hat{\mu}^{\text{low}}(\theta)\right)$ formed from LF simulation data acts as a "**corrector**" that **reduces policy gradient variance without adding bias** (approximately):

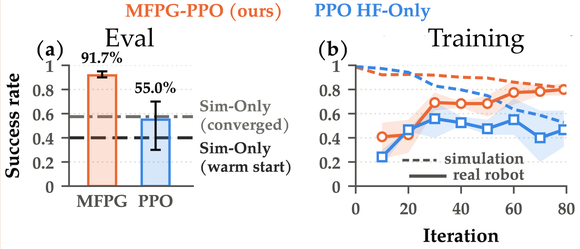
- $\text{Var}\left[X^{\text{high}} + c^*\left(X^{\text{low}} - \mu^{\text{low}}\right)\right] = \left(1-\rho^2\right)\text{Var}\left[X^{\text{high}}\right]$ (**variance reduction**; ρ: correlation between $\left(X^{\text{high}}, X^{\text{low}}\right)$)
- $\mathbb{E}\left[X^{\text{high}} + c\underbrace{\left(X^{\text{low}} - \mu^{\text{low}}\right)}_{\text{zero mean}}\right] = \mathbb{E}\left[X^{\text{high}}\right]$ (**unbiasedness**; holds approximately)

A "perfect" LF simulator ($\rho = 1$) would entirely remove the variance.

Experiments

Real robot: MFPG-PPO enables stable learning on a physical Franka arm using only 4 real-robot episodes per update and no human demonstrations.





Isaac Lab baseline comparison: MFPG-PPO is the only method in our comparison that improves upon PPO HF-Only in nearly every setting, is never distinguishably worse, and scales monotonically with more HF data.

