# Policy Learning and Evaluation with Randomized Quasi-Monte Carlo

Séb Arnold, Pierre L'Ecuyer, Liyu Chen, Yi-fan Chen, Fei Sha

*March 10, 2022*

USC University of Southern California
Université de Montréal
Google AI

# Overview

USC University of Southern California

**Summary**

- Replacing MC with RQMC accelerates learning and improves value estimation in RL.

**Main Contributions**

- We propose to combine policy gradients with randomized QMC.
  - Retains flexibility of policy gradients (eg, continuous actions, non-linear policies, etc).
  - Readily compatible with different policy gradient formulations (eg, actor-critic).

- Empirically, we show:
  - RQMC improves policy learning and evaluation, even for SOTA algorithms.
  - RQMC reduces variance in gradients and policy values.
  - RQMC complements other variance reduction techniques.

# Background



**Policy Gradients**

- Iterate: $\pi \leftarrow \pi - \eta \nabla_\pi \mathbb{E}_{s,a}[Q^\pi(s,a)]$

**Randomized Quasi-Monte Carlo (RQMC)**

Monte Carlo:

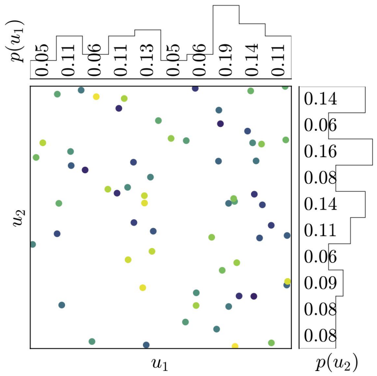
- Sample points $u \sim U(0;1)$ uniformly at random.

Quasi-Monte Carlo:

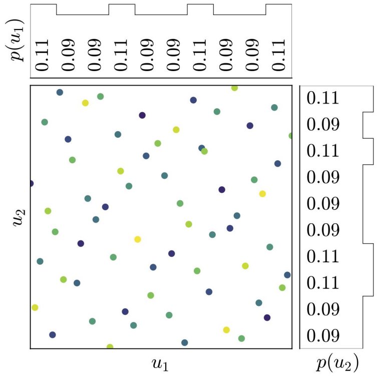
- Deterministically generate a low-discrepancy point set.

Randomized Quasi-Monte Carlo:

- Scramble & randomly shift a QMC point set to retain low-discrepancy.




Monte Carlo




Randomized QMC

# Policy Evaluation with RQMC

USC University of Southern California

**Goal**

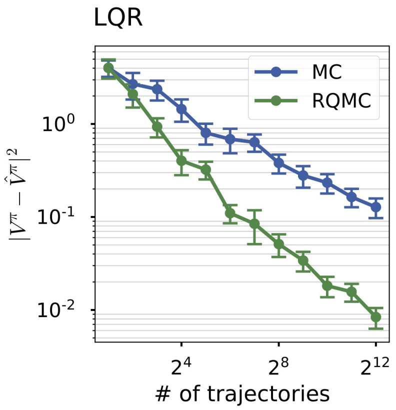
- Efficiently estimate: $V^\pi = \mathbb{E}_{s,a}[Q^\pi(s,a)]$

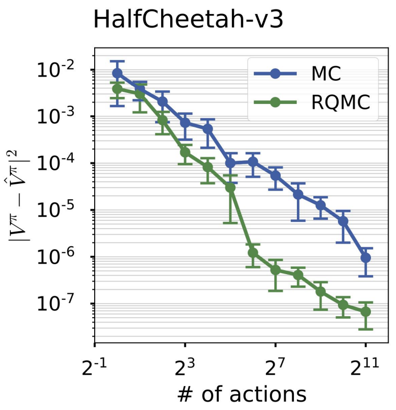
**Method**

- Let: $a = \pi(s,u) = \mu(s) + \sigma(s) \odot F^{-1}(u)$, where $u$ is an RQMC point.

**Policy evaluation** when approximating $V^\pi$ with:

- Expected Returns: $V^\pi \approx \frac{1}{N}\sum_{i=0}^{N}\left[\sum_{t=0}^{T} R(s_t^{(i)}, a_t^{(i)})\right]$
  - Sample trajectories, average sum of rewards.
- Learned Critic: $V^\pi \approx \mathbb{E}_{s_k}\left[\frac{1}{N}\sum_{i=0}^{N} \hat{Q}^\pi(s_k, \pi(s_k, u_k^{(i)}))\right]$
  - Sample states from buffer replay, average Q-values.



# Policy Learning with RQMC

USC University of Southern California

**Goal**

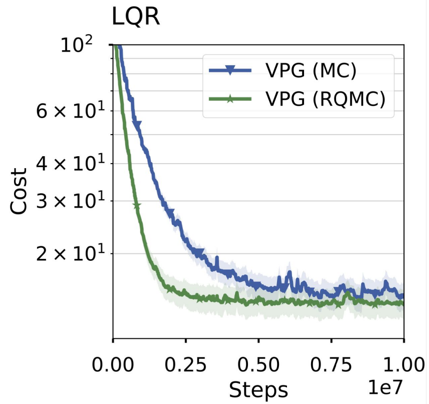
- Efficiently learn a policy: $\arg\max_\pi \mathbb{E}_{s,a}[Q^\pi(s,a)]$

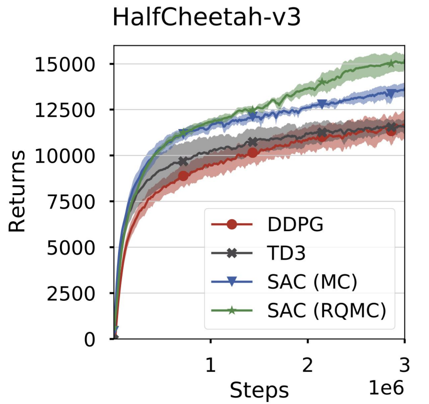
**Method**

- Let: $a = \pi(s,u) = \mu(s) + \sigma(s) \odot F^{-1}(u)$, where $u$ is an RQMC point.
- Learn with
  - Expected Returns $\rightarrow$ Vanilla Policy Gradient (VPG)
  - Learned Critic $\rightarrow$ Soft Actor-Critic (SAC)

**Experimental results**

- RQMC outperforms MC on all scenarios.
  - Significantly improves learning with VPG.
  - Combines with and improves upon SOTA algorithms.



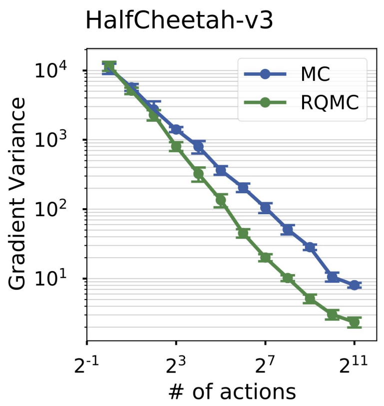



# Analyses and Ablations

USC University of Southern California

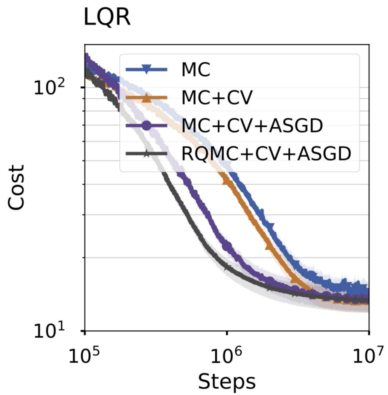
**RQMC improves gradient estimation**

- Why does RQMC improve upon MC?
  - Hypothesis: variance reduction.
- Experiment:
  - Collect trajectories mid-training.
  - Measure gradient variance and alignment.
  - Results: 5x lower gradient variance.




**RQMC combines with other variance reduction techniques**

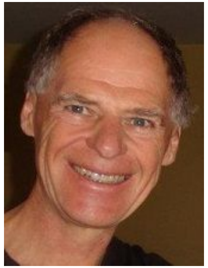
- Can RQMC complement other variance reduction techniques (VRTs)?
- Experiment:
  - Compare MC with different VRT combinations.
  - Results: RQMC further improves upon
    - Control variates (CV)
    - Accelerated SGD (ASGD)



# Thank You!

| | |
|---|---|
| **Poster # 3166** | Mon 28 Mar 10:15 a.m. PDT — 11:45 a.m. PDT |
| **Code** | github.com/seba-1511/qrl |
| **Website** | sebarnold.net/projects/qrl |
| **Contact** | seb.arnold@usc.edu |

Séb Arnold



Pierre L'Ecuyer

Liyu Chen

Yi-fan Chen

Fei Sha