QPRL: Learning Optimal Policies with Quasi-Potential Functions for Asymmetric Traversal

Jumman Hossain and Nirmalya Roy

Department of Information Systems, University of Maryland, Baltimore County, USA.

UMBC · Mobile Pervasive & Sensor Computing · CYPRESS · ArtIAMAS · DEVCOM Army Research Laboratory · ICML International Conference On Machine Learning 2025

Introduction

- Real-world robotic navigation often involves **asymmetric** and **irreversible** traversal costs (e.g., uphill vs. downhill paths, one-way transitions).
- Traditional RL and potential-based reward shaping implicitly assume symmetric costs, limiting their effectiveness in such scenarios.
- Recent **quasimetric RL** methods relax symmetry constraints but:
  - Do not explicitly model **path-dependent** traversal costs.
  - Lack rigorous **safety guarantees**.
- Our approach addresses these limitations through explicit **quasi-potential decomposition** and a **Lyapunov-based safety mechanism**.

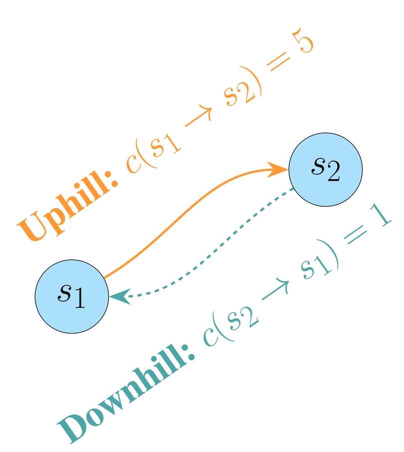


Figure 1: Illustration of asymmetric traversal costs: uphill ($s_1 \to s_2$, cost=5) vs. downhill ($s_2 \to s_1$, cost=1). QPRL explicitly addresses this direction-dependent asymmetry.

Quasi-Potential Reinforcement Learning (QPRL)

Novel Decomposition:

$$d(s,g) = \underbrace{\Phi(g) - \Phi(s)}_{\text{Path-Independent}} + \underbrace{\Psi(s \to g)}_{\text{Path-Dependent}}$$

- **Path-Independent Potential (Φ)**: Reusable costs, analogous to gravitational potentials.
- **Path-Dependent Residual (Ψ)**: Irreversible or dissipative costs, like friction.

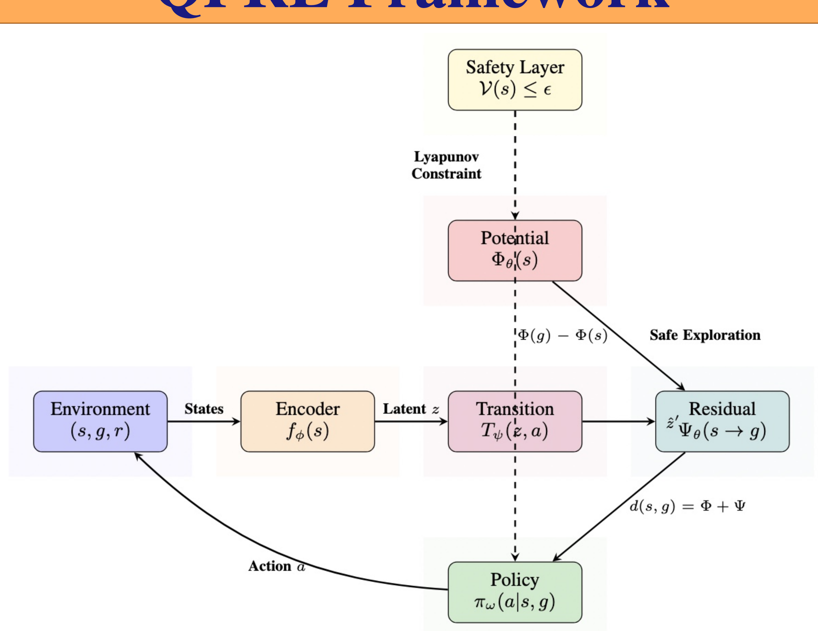
Key Benefits:

- Explicit and interpretable modeling of **directional asymmetry**.
- Improved exploration efficiency and accelerated policy learning.

Theoretical Contributions:

- Faster convergence rate: $\tilde{\mathcal{O}}(\sqrt{T})$, improving upon prior $\tilde{\mathcal{O}}(T)$.
- Provable Lyapunov safety guarantees significantly reduce constraint violations.

QPRL Framework



- Models traversal costs explicitly with two distinct functions:
  - **Potential function Φ:** represents reversible, global state costs.
  - **Residual function Ψ:** captures irreversible, direction-specific costs.
- Uses a Lyapunov-inspired constraint (Φ-based safety) to bound state transitions and guide exploration.
- Updates policy and value functions based on decomposed costs for targeted optimization.

Algorithm

Algorithm: Quasi-Potential Reinforcement Learning (QPRL)

1: **Input:** Replay buffer $\mathcal{D}$, learning rates $\alpha_\phi, \alpha_\psi, \alpha_\theta, \alpha_\omega$, threshold ϵ
2: **for** iteration $= 1$ to N **do**
3: Sample batch $\{(s_i, a_i, s_i', c_i, g_i)\}_{i=1}^B \sim \mathcal{D}$
4: **Update Encoder & Transition Model:**
5: $z_i = f_\phi(s_i),\ \hat{z}_i' = T_\psi(z_i, a_i)$
6: $\mathcal{L}_T = \frac{1}{B}\sum_i \|\hat{z}_i' - f_\phi(s_i')\|^2$
7: Update ϕ, ψ using $\nabla_{\phi,\psi}\mathcal{L}_T$
8: **Update Quasi-Potential Function:**
9: $\mathcal{L}_U = \frac{1}{B}\sum_i \left(\Phi_\theta(g_i) - \Phi_\theta(s_i) + \Psi_\theta(s_i \to g_i) - c_i\right)^2$
10: $\mathcal{L}_{\text{constraint}} = \frac{1}{B}\sum_i \left(\max\left(0,\ \Psi_\theta(s_i \to s_i') - (c_i - \Phi_\theta(s_i') + \Phi_\theta(s_i))\right)\right)^2$
11: Update θ using $\nabla_\theta(\mathcal{L}_U + \lambda\mathcal{L}_{\text{constraint}})$
12: **Update Policy with Safety Layer:**
13: $z_i = f_\phi(s_i),\ a_i = \pi_\omega(s_i, g_i)$
14: $\hat{z}_i' = T_\psi(z_i, a_i)$
15: $\hat{d}_i = \Phi_\theta(g_i) - \Phi_\theta(s_i) + \Psi_\theta(s_i \to g_i)$
16: $\mathcal{L}_\pi = \frac{1}{B}\sum_i \hat{d}_i + \lambda \cdot \max\left(0, \Phi_\theta(\hat{z}_i') - \Phi_\theta(s_i) - \epsilon\right)$
17: Update ω using $\nabla_\omega \mathcal{L}_\pi$
18: **end for**

- **State Encoder (f_ϕ) and Transition Model (T_ψ):**
  - Learn compact latent state representations.
  - Efficiently predict next latent states for planning.
- **Quasi-Potential Components (Φ, Ψ):**
  - Decompose asymmetric traversal costs explicitly.
  - Maintain quasimetric properties (triangle inequality, non-negativity).
- **Lyapunov-Based Safety Constraint:**

$$\mathbb{E}_{s' \sim P(\cdot|s,a)}\left[\Phi_\theta(s')\right] \leq \Phi_\theta(s) + \epsilon$$

- **Safety-Aware Policy Loss:**

$$\mathcal{L}_\pi = \frac{1}{B}\sum_{i=1}^{B}\left[\hat{d}_i + \lambda \cdot \mathrm{ReLU}\left(\Phi_\theta(\hat{z}_i') - \Phi_\theta(s_i) - \epsilon\right)\right]$$

- **Dynamic Lagrange Multiplier (λ):**
  - Adaptively enforces safety constraints during training.

Theoretical Analysis

Theorem (Convergence) Assuming Lipschitz continuity of Φ and Ψ, **QPRL** attains $\tilde{\mathcal{O}}(\sqrt{T})$ regret, improving on the $\tilde{\mathcal{O}}(T)$ bound of monolithic quasimetric RL.

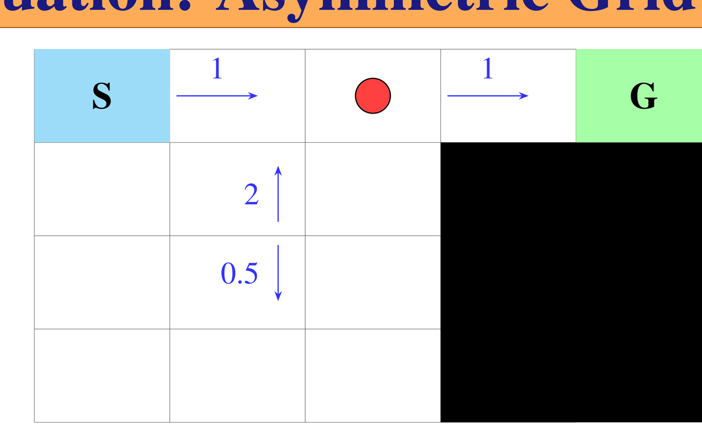
Lemma (Lyapunov Safety) Under the policy π_{safe},

$$\mathbb{E}_{s' \sim P(\cdot|s,a)}\left[\Phi(s')\right] \leq \Phi(s) + \epsilon, \quad \forall t,$$

guaranteeing recoverability from ϵ-bounded unsafe states.

See main paper for complete proofs.

Evaluation: Asymmetric GridWorld



- Agent must navigate from start (**S**) to goal (**G**).
- Costs: horizontal = 1, up = 2, down = 0.5.
- Walls are impassable, illustrating direction-dependent navigation.
- Evaluates QPRL's handling of asymmetric costs and safety constraints.

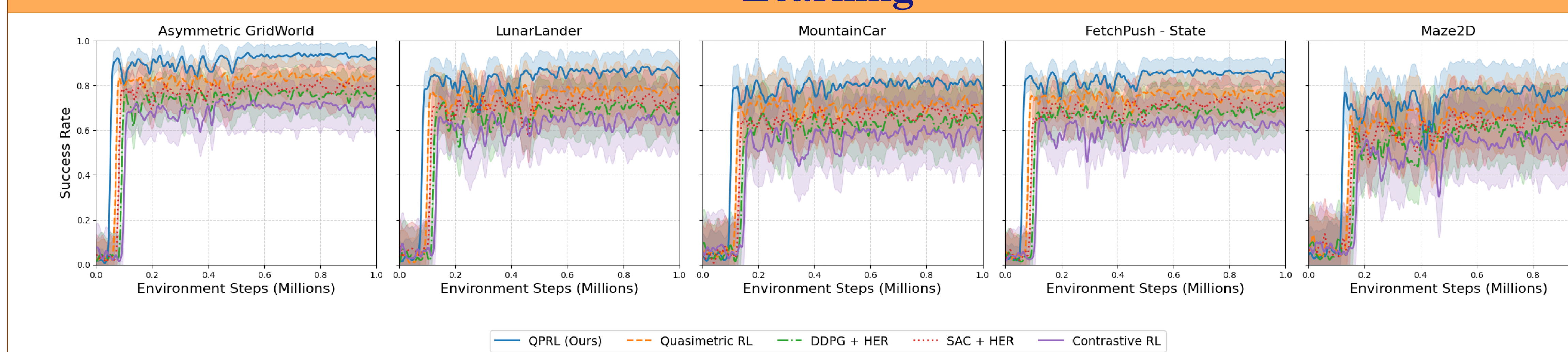
Performance Comparison

Environment	Metric	QPRL (Ours)	QRL	Contrastive RL	DDPG+HER	SAC+HER
Asymmetric GridWorld	Success Rate (%)	**92.5** $\pm$ **2.2**	87.3 $\pm$ 3.0	82.4 $\pm$ 3.5	78.9 $\pm$ 4.2	80.3 $\pm$ 4.0
MountainCar	Normalized Return	**-95.6** $\pm$ **4.1**	-108.4 $\pm$ 6.7	-118.3 $\pm$ 8.1	-125.5 $\pm$ 7.6	-121.2 $\pm$ 7.0
FetchPush	Success Rate (%)	**91.2** $\pm$ **3.0**	85.5 $\pm$ 3.6	79.3 $\pm$ 4.1	73.8 $\pm$ 4.5	77.0 $\pm$ 4.3
LunarLander	Success Rate (%)	**88.9** $\pm$ **3.4**	81.4 $\pm$ 4.0	76.7 $\pm$ 4.5	72.5 $\pm$ 5.0	74.2 $\pm$ 4.8
Maze2D	Success Rate (%)	**85.3** $\pm$ **3.7**	78.1 $\pm$ 4.3	72.6 $\pm$ 4.7	68.9 $\pm$ 5.2	70.1 $\pm$ 4.9

Table 1: Mean $\pm$ std performance over 5 random seeds on asymmetric-cost benchmarks. QPRL attains the highest success rate (or least-negative return).

- QPRL consistently achieves **highest success rates** and **best returns**.
- Notably reduces variance across multiple random seeds.
- Demonstrates clear empirical advantage in asymmetric environments.

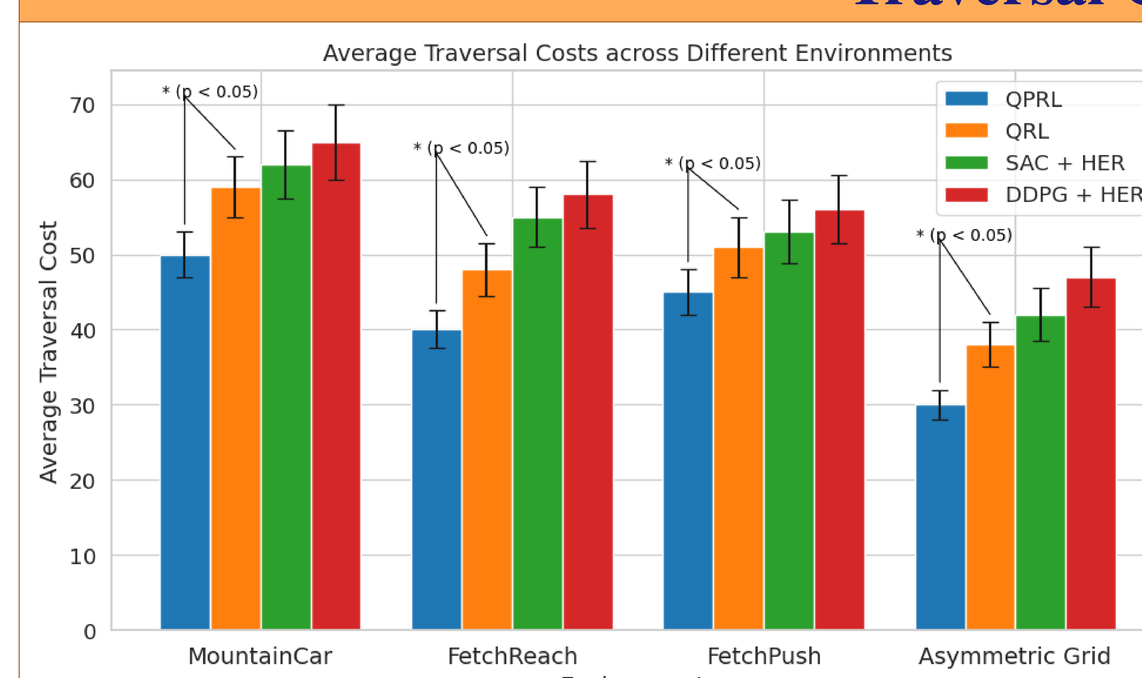
Learning



Sample-efficiency and stability across tasks. Success-rate learning curves for all five asymmetric environments. The x-axis shows *environment interactions*; the y-axis shows mean *success rate*. QPRL (blue) reaches high performance earliest and maintains the highest asymptotic success with visibly lower variance.

- **Fastest convergence:** QPRL reaches $\geq 90\%$ success in $\sim 2-3\times$ fewer steps than the best baseline.
- **Highest final performance:** achieves the top asymptotic success rate on every environment.
- **Stable learning:** narrow confidence band indicates significantly lower variance across seeds.

Traversal Cost Comparison



Average traversal cost comparison. QPRL demonstrates the **lowest cost**, showing its advantage in exploiting asymmetric dynamics. Results statistically significant ($p < 0.01$, paired t-test).

Project Website

Environment	Method	Symmetric (%)	Asymmetric (%)	Gap (%)
Asymmetric GridWorld				
	QPRL	94.1 ± 1.8	88.7 ± 2.5	5.4
	QRL	92.3 ± 2.0	83.5 ± 2.8	8.8
	SAC + HER	90.2 ± 2.3	81.0 ± 3.2	9.2
	DDPG + HER	89.8 ± 2.5	80.5 ± 3.5	9.3
MountainCar				
	QPRL	-90.5 ± 4.3	-98.2 ± 5.0	7.7
	QRL	-88.2 ± 4.1	-96.5 ± 5.2	8.3
	SAC + HER	-87.0 ± 4.0	-95.8 ± 5.3	8.8
	DDPG + HER	-86.5 ± 4.2	-94.5 ± 5.1	8.0
FetchPush				
	QPRL	92.0 ± 2.2	85.3 ± 3.1	6.7
	QRL	90.5 ± 2.3	81.0 ± 3.2	9.5
	SAC + HER	89.8 ± 2.5	79.8 ± 3.5	10.0
	DDPG + HER	88.5 ± 2.4	78.5 ± 3.4	10.0
LunarLander				
	QPRL	88.6 ± 3.4	82.4 ± 3.7	6.2
	QRL	87.0 ± 3.5	80.0 ± 4.0	7.0
	SAC + HER	85.5 ± 3.8	77.5 ± 4.2	8.0
	DDPG + HER	84.0 ± 3.6	76.0 ± 4.1	8.0

Table 2: Performance on symmetric vs. asymmetric variants of each environment (mean $\pm$ 1 s.d. over 5 seeds). **Gap (%)** is the absolute difference between the two settings—lower is better, indicating robustness to asymmetric traversal costs.

Acknowledgments: This work has been partially supported by ONR Grant #N00014-23-1-2119, U.S. Army Grant #W911NF2120076, U.S. Army Grant #W911NF2410367, NSF REU Site Grant #2050999, NSF CNS EAGER Grant #2233879, and NSF CAREER Award #1750936