ICAIF'24
5th ACM International Conference on AI in Finance
Brooklyn, NY, November 14-17, 2024

Navigating the Difficulty of Achieving Global Optimality under Variance-Induced Time Inconsistency

Jingxiang Tang *, Nixie S. Lesmana*†, Chi Seng Pun*

NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE
NUS National University of Singapore

Introduction

Aim

To investigate how TIC permeates into the design and behaviour of mean-variance (MV) reinforcement learning (RL) agents.

Main Contribution

- Study two optimality classes and their corresponding TIC-aware RL methods;
- Study 2 optimality classes under variance-induced TIC (global optimality and SPE) and the corresponding TIC-aware RL methods (EPG and SPERL)
- Characterize the conditions in which equilibrium/SPERL policies attain global optimality.

Time Inconsistency

Bellman's Principle of Optimality (BPO)

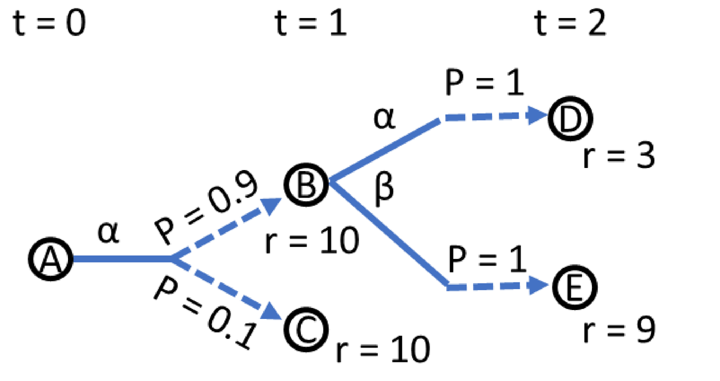
For all $\tau > 0$, $x_\tau \in \mathcal{X}_\tau^{x_0}$,

$$\pi_t^{*\tau,x_\tau}(x) = \pi_t^{*0,x_0}(x), \forall\, t > \tau, x \in \mathcal{X}_t^{x_0}$$

Mean-Variance(MV) Objective

- $V_t^\pi(x) := \mathbb{E}[R_t^\pi(x)] - \lambda\, Var[R_t^\pi(x)]$, **with $\lambda > 0$.**
- $R_t^\pi(x) := \sum_{j=t}^{T-1} r_j\left(X_j, \pi_j(X_j)\right) + r_T(X_T) \mid (t, X_t = x)$

MV objective is a source of TIC



$V_t^\pi(x)$		Policy, π	
		$\{\alpha,\alpha\}$	$\{\alpha,\beta\}$
State (t,x)	(0, A)	11.89	10.81
	(1, B)	3	9

Episodic Policy Gradient

Global Optimality/Precommitment

$$\pi^* := \arg\max_\pi V_0^\pi(x_0)$$

Policy Gradient

$$\Delta\theta \propto \nabla V_0^\theta(x_0)$$

- **TrueEPG**: MDP transitions are known to the agent, $\boldsymbol{\nabla V_0^\theta(x_0)}$ can be explicitly computed.
- **ApproxEPG**: MDP transitions are unknown to the agent, $\boldsymbol{\nabla V_0^\theta(x_0)}$ need to be approximated from trajectories.

Subgame Perfect Equilibrium

Equilibrium

Find $\hat{\pi}$ such that:

$$Q_t^{\hat{\pi}}(x, \hat{\pi}_t(x)) \geq Q_t^{\hat{\pi}}(x, a) \quad \forall t, x, a \in \mathcal{S} \times \mathcal{X} \times \mathcal{U}.$$

Backward Update

For $t \leftarrow T$ to 0:

Compute/Approx $Q_t^\pi(x,a)$ → $\pi_{k+1}(t,x) \leftarrow \arg\max_a Q_t^{\pi_k}(x,a)$

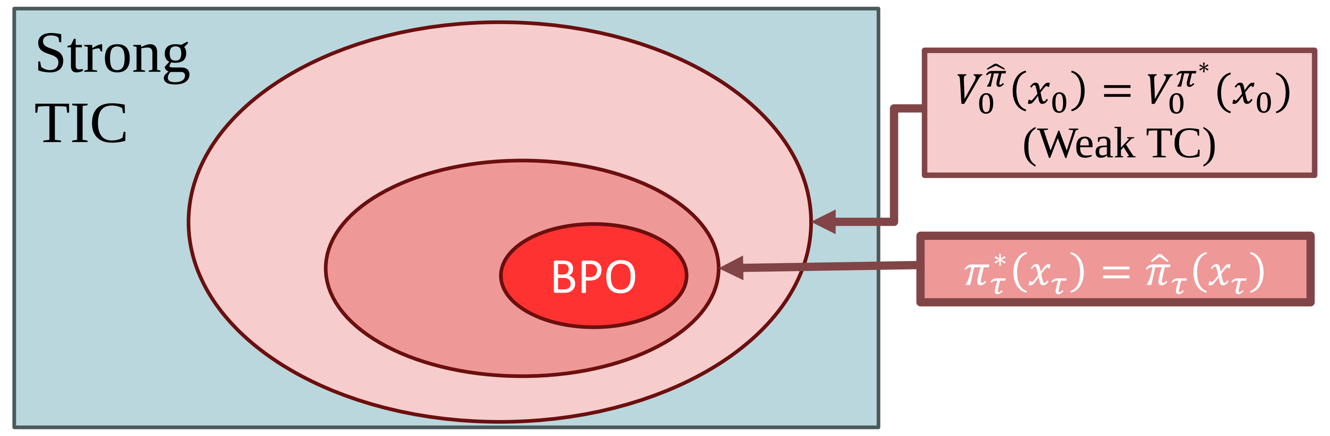
- **TrueSPE**: MDP transitions are known to the agent, Q-values can be explicitly computed.
- **SPERL**: MDP transitions are unknown to the agent, Q-values need to be approximated according to TD algorithm from trajectories.

Hierarchy of TCness

Proposition 1:

- If **BPO** holds, then

$$\pi_\tau^*(x_\tau) = \hat{\pi}_\tau(x_\tau), \qquad \forall\, \tau, x_\tau \in \mathcal{X}_\tau^{x_0}.$$



Experiment Results

Portfolio Optimization (PM)

- Agent's objective: find how much (in % of liquid risk-free asset held) to invest into an illiquid risky asset.
- Rewards: increase in total wealth.

Optimal Control (OE)

- Agent's objective: decide how much (in % of remaining) to sell in to liquidate a huge amount of stock to minimise transaction cost.
- Rewards: negation of transaction cost.

Expected Values and std of $V_0(x_0)$

- Convergence and performance of **EPG** algorithm are sensitive to environment and instance;
- **SPERL** performs better then both **EPG** agents in **OE** setup.

	TrueEPG	ApproxEPG	TrueSPE	SPERL
PM-1	$\mathbf{3.193}_{3.8e-5}$	$3.125_{4.4e-2}$	3.190	$3.188_{4.9e-3}$
PM-2	$\mathbf{2.547}_{2.7e-5}$	$2.469_{4.3e-2}$	**2.547**	$2.533_{1.5e-2}$
OE-1	$-1.85_{0.138}$	$-2.07_{0.179}$	**-1.12**	$-1.61_{0.117}$
OE-2	$-1.75_{0.121}$	$-2.02_{0.251}$	**-1.27**	$-1.74_{0.139}$

Policy Tree-diagrams

- **SPERL** policies are very similar to the corresponding **SPE** policies;
- No clear link is observed between the behaviour of **TrueEPG** policies and **ApproxEPG** policies.

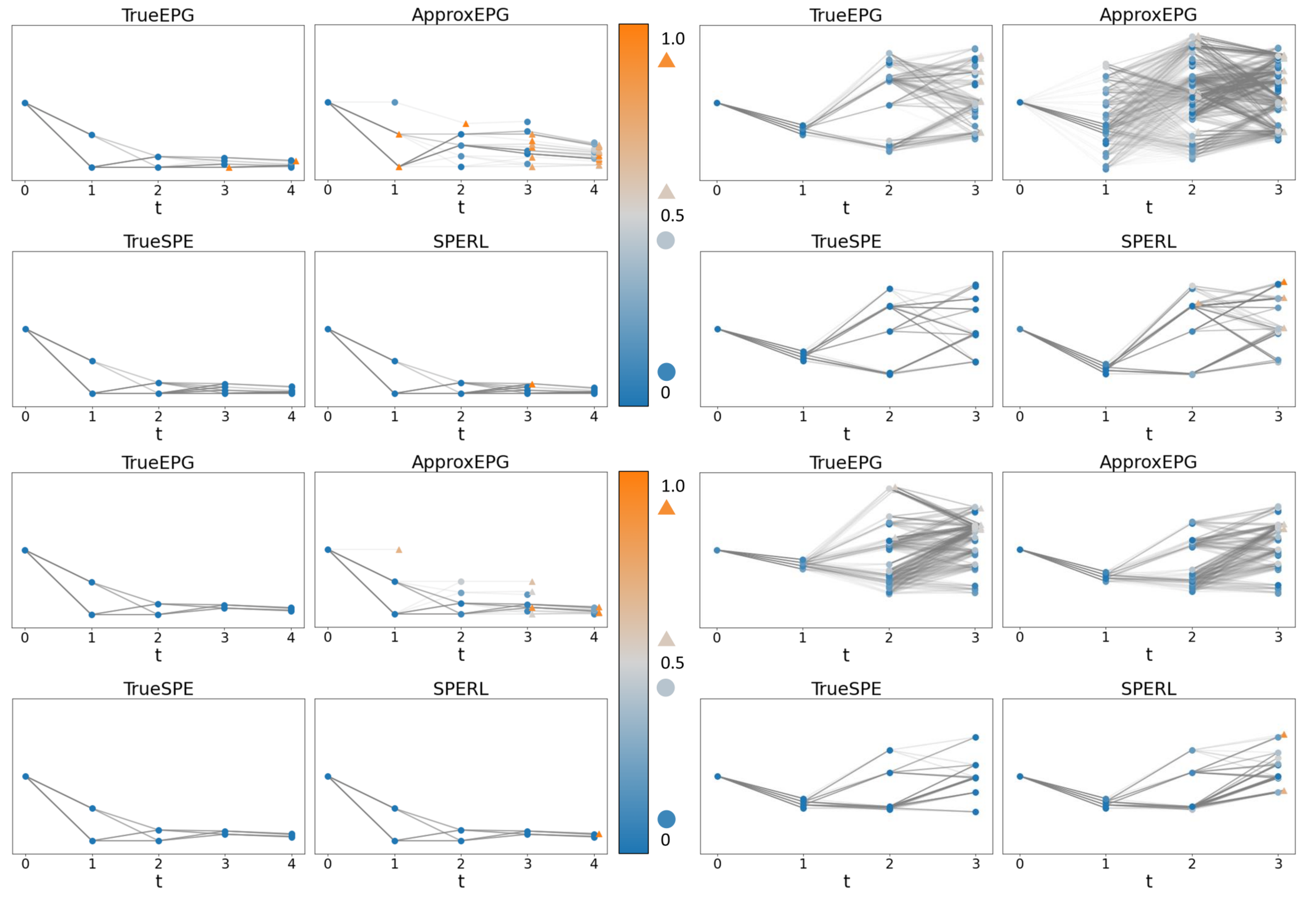


Figure 1. Agents' Policy in **PM** (left); Agents' Policy in **OE** (right).

ΔV across environment parameters

- $\Delta V := V_0^{SPE}(x_0) - V_0^{EPG}(x_0)$

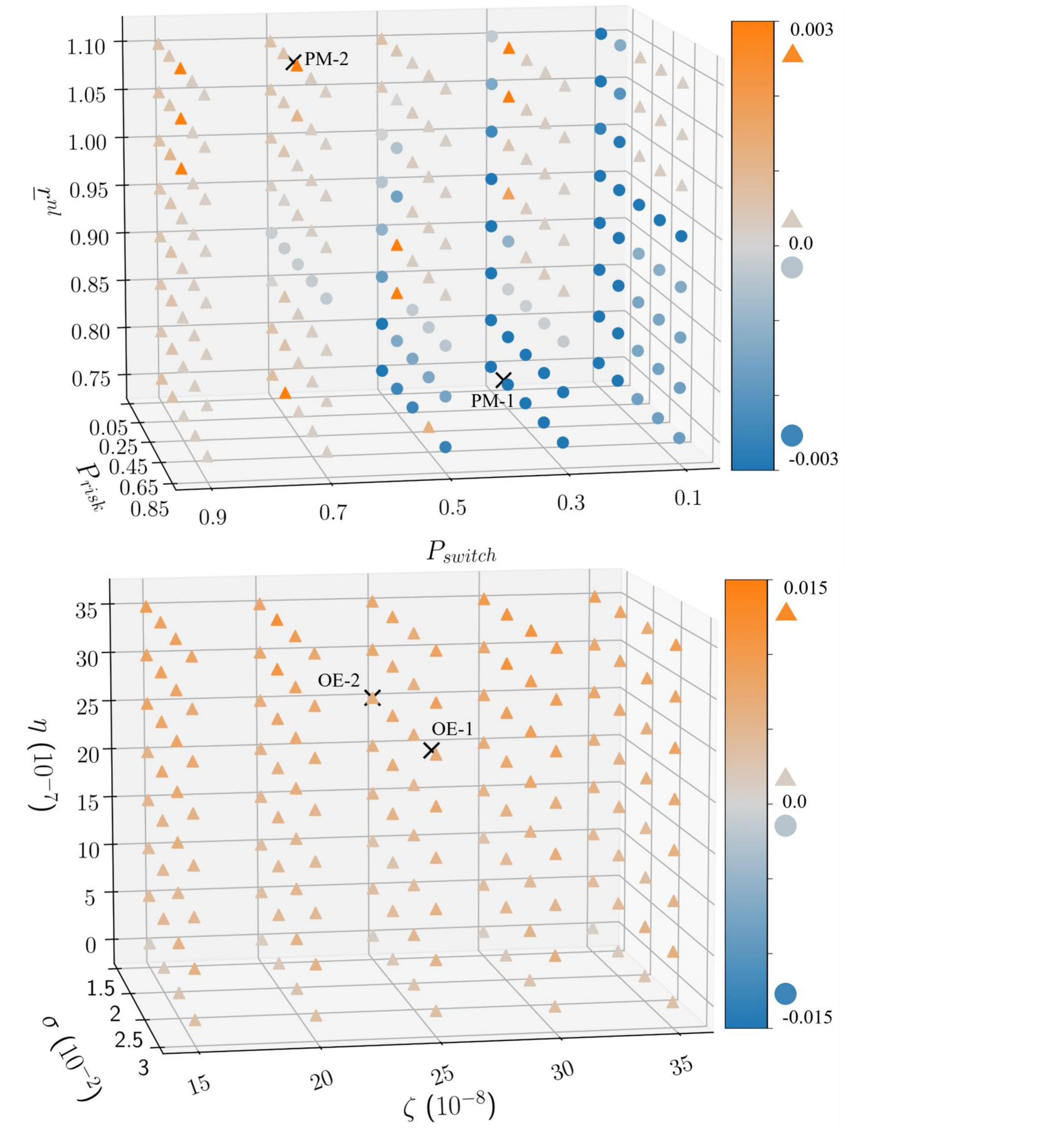


Figure 2. TrueEPG and **TrueSPE** initial value differences in **PM** (above) and **OE** (below).

Conclusions

- **TrueEPG** dominates **SPE** only when *both* (i) the environment is **strong-TIC** and (ii) **TrueEPG** is inclined to attain the global optimum (more likely in simpler environments like PM than those like OE).
- Whether or not an environment is amenable for **TrueEPG** can be gleaned from **TrueEPG**'s behaviour across multiple seeds.
- **SPERL** algorithm can learn **SPE** policy.

Acknowledgments

Chi Seng Pun gratefully acknowledges Ministry of Education (MOE) Singapore, AcRF Tier 2 grant (Reference No.: MOE-T2EP20220-0013) for the funding of this research. Nixie S Lesmana acknowledges the financial support from MoE AcRF grant R-144-000-457-733 (A-0004550-00-00).

Link to paper: https://doi.org/10.1145/3677052.3698657

*Division of Mathematical Sciences, Nanyang Technological University, Singapore 637371
†Department of Physics, Faculty of Science, National University of Singapore, Singapore 117546