Bridging Distributional and Risk-sensitive Reinforcement Learning with Provable Regret Bounds

Hao Liang

Joint work with **Zhi-Quan Luo** (CUHK-Shenzhen)

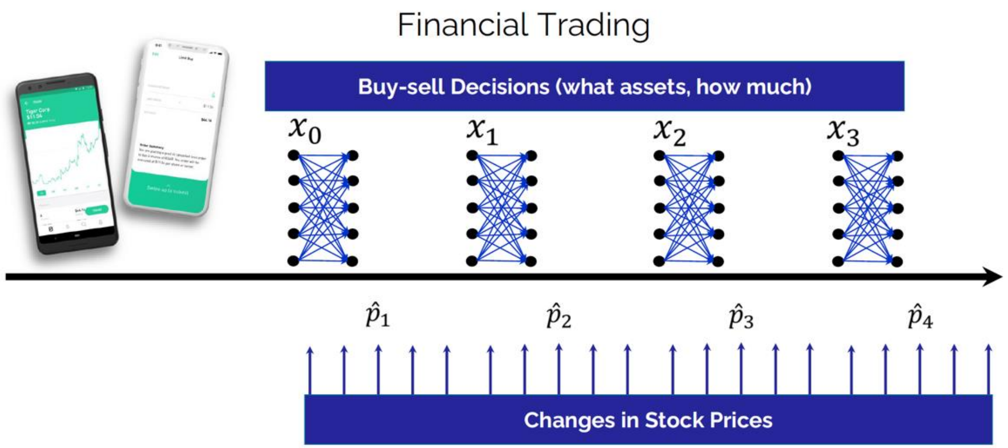
Department of Informatics
King's College London

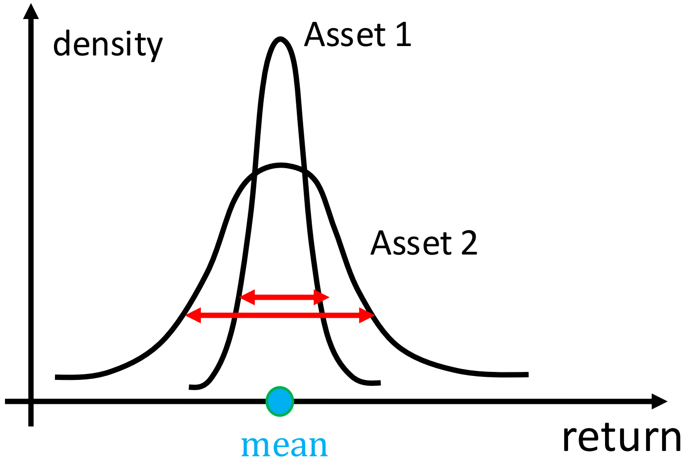
Hao Liang, and Zhi-Quan Luo. "Bridging distributional and risk-sensitive reinforcement learning with provable regret bounds." *Journal of Machine Learning Research* 25.221 (2024): 1-56.

Decision Making under Risk

Risk is crucial in high-stake applications

Finance: stock trading

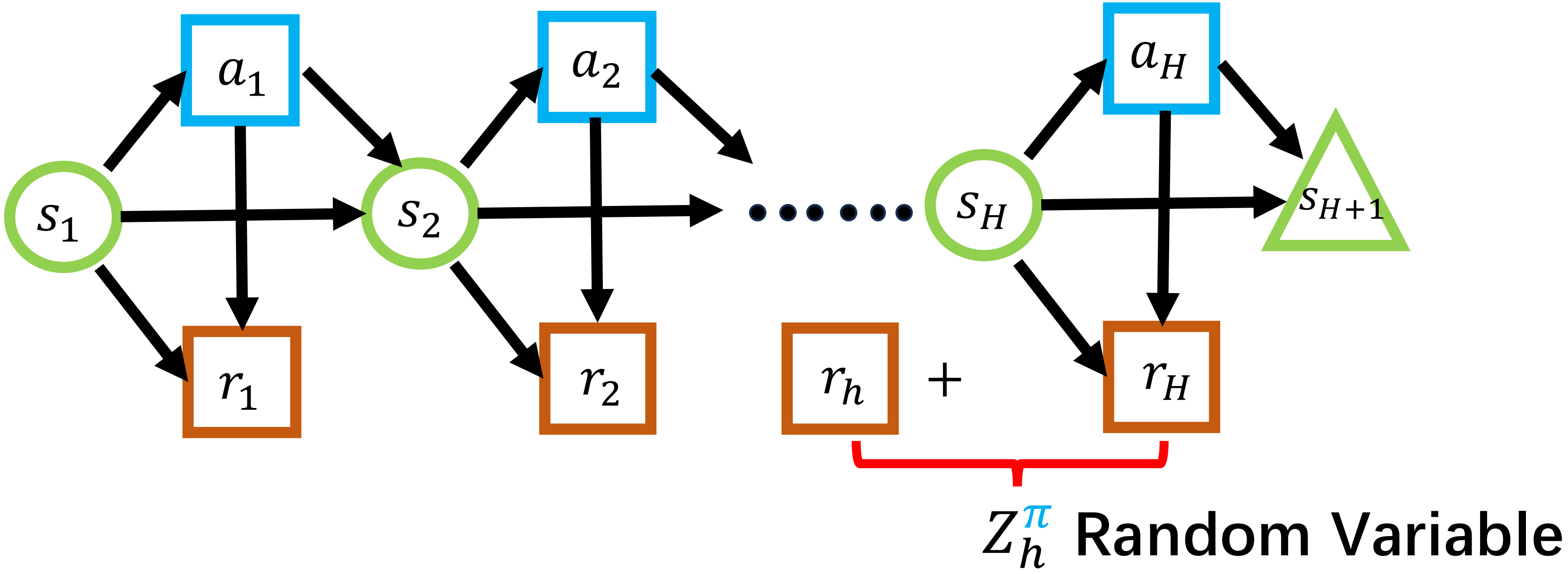




Control volatility

Figure by courtesy of Warren B Powell "A Unified Framework for Sequential Decisions under Uncertainty"

Markov Decision Process (MDP)



- Policy $\pi = (\pi_h)_{h \in [H]}$

  $\Pi \ni \pi_h : S \to A$
- Return = cumulative reward

  $Z_h^\pi = r_h(s_h, a_h) + \cdots + r_H(s_H, a_H)$

  $a_h = \pi_h(s_h), s_{h+1} \sim P_h(s_h, a_h)$

Risk-neutral MDP vs. Risk-aware MDP

Risk-neutral MDP

$$\max \; \mathbf{E}[Z_1^{\pi}]$$

Risk-aware MDP

$$\max \; \rho(Z_1^{\pi})$$

MANAGEMENT SCIENCE
Vol. 18, No. 7, March, 1972
Printed in U.S.A.

RISK-SENSITIVE MARKOV DECISION PROCESSES*

RONALD A. HOWARD† AND JAMES E. MATHESON‡§

Entropic risk measure (ERM) [HM72]

$$\mathbf{U}_{\boldsymbol{\beta}}(X) := \frac{1}{\beta} \log \mathbf{E}[\exp(\beta X)] = \mathbf{E}[X] + \frac{\beta}{2} \mathbf{V}[X] + O(|\beta|^2)$$

β controls risk preference

- Risk-seeking $\beta > 0$
- Risk-averse $\beta < 0$
- Risk-neutral $\beta \to 0$

[HM72] Howard, Ronald A., and James E. Matheson. "Risk-sensitive Markov decision processes." *Management science* 18.7 (1972): 356-369.

Distributional Dynamic Programming: Risk-aware Control

Key property 1: Additivity

$$\boldsymbol{U_\beta}(X + c) = \boldsymbol{U_\beta}(X) + c$$

Key property 2: Independence

$$\boldsymbol{U_\beta}(F) \le \boldsymbol{U_\beta}(G) \Longrightarrow$$
$$\boldsymbol{U_\beta}\big((1-\theta)F + \theta \cdot H\big) \le \boldsymbol{U_\beta}\big((1-\theta)G + \theta \cdot H\big)$$

Distributional Bellman Optimality Equation [LL21]

$$\eta_h^*(s,a) = [\mathbf{T}_{\boldsymbol{d}} \nu_{h+1}^*](s,a)$$
$$\pi_h^*(s) = \operatorname{argmax}_a \boldsymbol{U_\beta}(\eta_h^*(s,a))$$
$$\nu_h^*(s) = \eta_h^*\big(s, \pi_h^*(s)\big)$$

backward recursion

greedy is optimal

Distributional Bellman Operator T_d: $P(R)^S \to P(R)^{S\times A}$

$$\eta_h(s,a) = [\mathrm{T}_d \nu_{h+1}](s,a) \coloneqq \sum P_h(s'|s,a)\nu_{h+1}(s')(\cdot - r_h(s,a))$$

Risk-sensitive Optimistic Distribution Iteration (RODI)

Approximate Bellman recursion

$$\eta_h^k \leftarrow \widehat{\mathbf{T}_d}^k \nu_{h+1}^k$$

Distributional Optimism Operator

$$\eta_h^k \leftarrow \mathbf{O}_{c^k} \eta_h^k$$

Policy Execution

$$\pi_h^k(s) \leftarrow \mathrm{argmax}_a \boldsymbol{U_\beta}(\eta_h^k(s,a))$$

RODI in one line

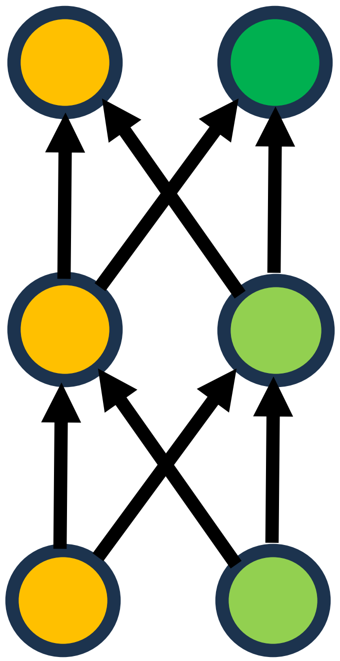
$$\eta_h^k \leftarrow \mathbf{O}_{c^k} \widehat{\mathbf{T}_d}^k \nu_{h+1}^k$$

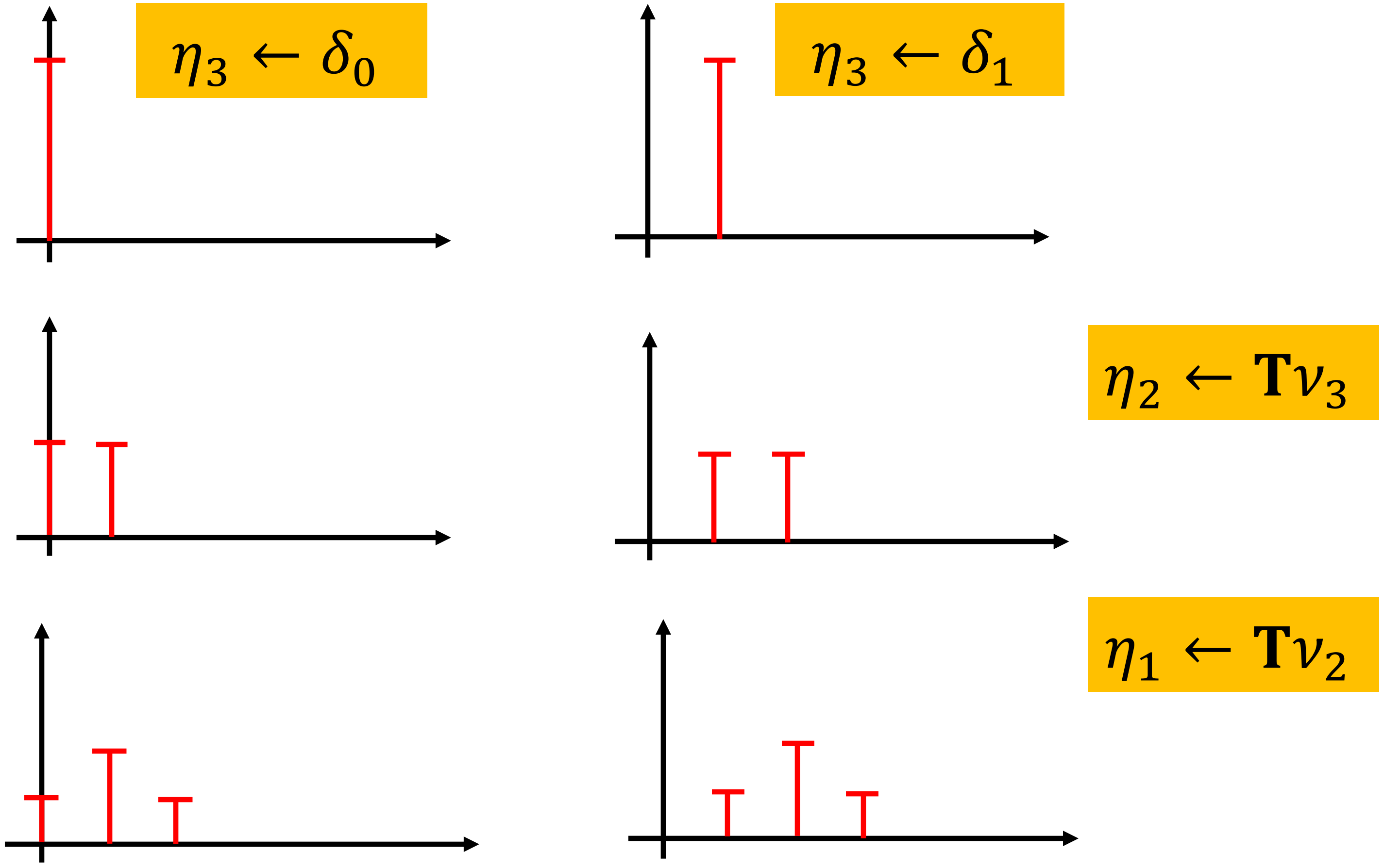
Distributional Optimism

$$\boldsymbol{U_\beta}(\eta_h^k(s,a)) \geq \boldsymbol{U_\beta}(\eta_h^*(s,a))$$
$$\forall\ (s,a,k,h)$$

Computational Inefficiency of RODI

State space = { , }
Uniform transition
$r(\ ) = 0, r(\ ) = 1$





Operator $\mathbf{T}$ expands support!

RODI with Distribution Representation (RODI-Rep)

$$\eta_h \leftarrow \mathbf{T}_{\boldsymbol{d}} \nu_{h+1} \longrightarrow |\eta_h| = S \cdot |\nu_{h+1}|$$

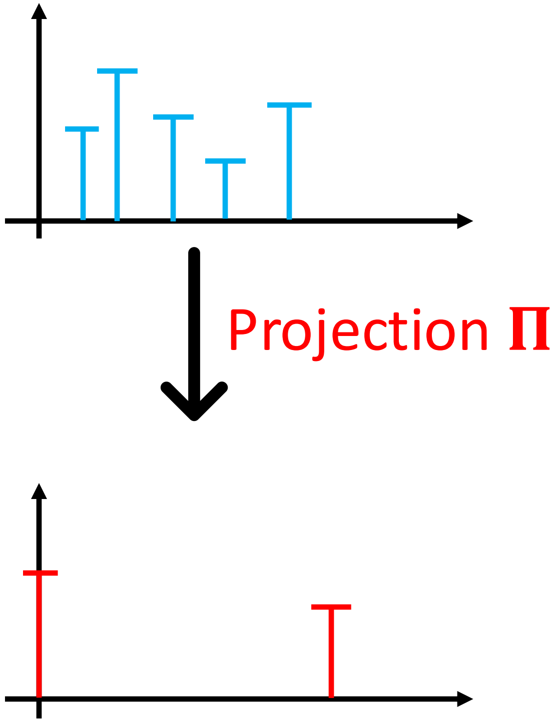
Represent distribution with fixed support via projection

$$\eta_h \leftarrow \boldsymbol{\Pi} \mathbf{T}_{\boldsymbol{d}} \nu_{h+1} \longrightarrow |\eta_h| = |\nu_{h+1}| = |\eta_{h+1}|$$

RODI-Rep

$$\eta_h \leftarrow \boldsymbol{\Pi} \mathbf{O}_{\boldsymbol{c}} \widehat{\mathbf{T}_d} \nu_{h+1} \longrightarrow |\eta_h| = |\nu_{h+1}|$$

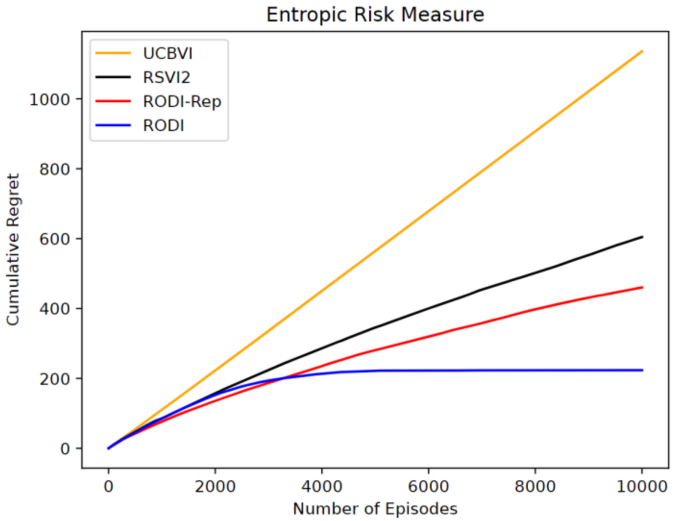
Ensure optimism while maintaining computational efficiency



Bernoulli representation

Regret Bounds and Numerical Experiments

Algorithm	Regret bound	Time	Space
RSVI	$\tilde{\mathcal{O}}\left(\exp(\lvert\beta\rvert H^2)\frac{\exp(\lvert\beta\rvert H)-1}{\lvert\beta\rvert}\sqrt{HS^2AT}\right)$	$\mathcal{O}\left(TS^2A\right)$	$\mathcal{O}\left(HSA+T\right)$
RSVI2	$\tilde{\mathcal{O}}\left(\frac{\exp(\lvert\beta\rvert H)-1}{\lvert\beta\rvert}\sqrt{HS^2AT}\right)$		
RODI-Rep			
RODI		$\mathcal{O}(KS^H)$	$\mathcal{O}(S^H)$
lower bound	$\Omega\left(\frac{\exp(\beta H/6)-1}{\beta}\sqrt{SAT}\right)$	-	-



[FYC+21] Fei, Yingjie, et al. "Exponential bellman equation and improved regret bounds for risk-sensitive reinforcement learning." *Advances in Neural Information Processing Systems* 34 (2021): 20436-20446.

Thank You!