Efficient Risk-aware Decision-making: A Distributional Perspective

Hao Liang

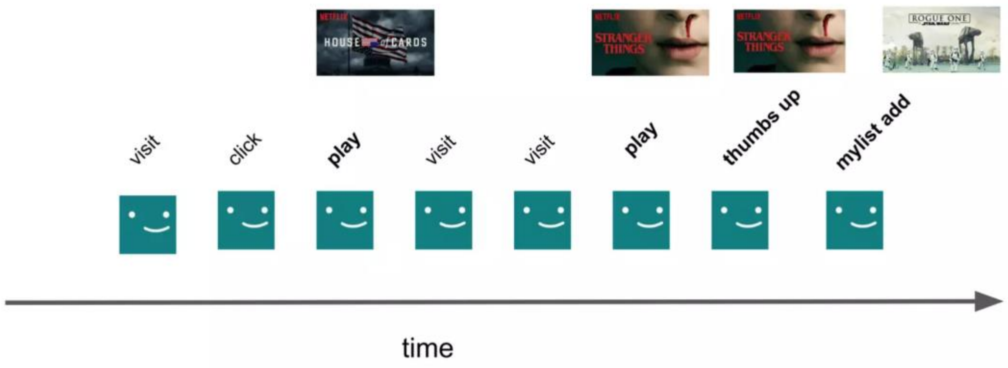
The Chinese University of Hong Kong, Shenzhen

17th, April, 2024

Sequential Decision-making (SDM)

In many applications, decisions are made over time…

online recommendation



maximize average "satisfaction"

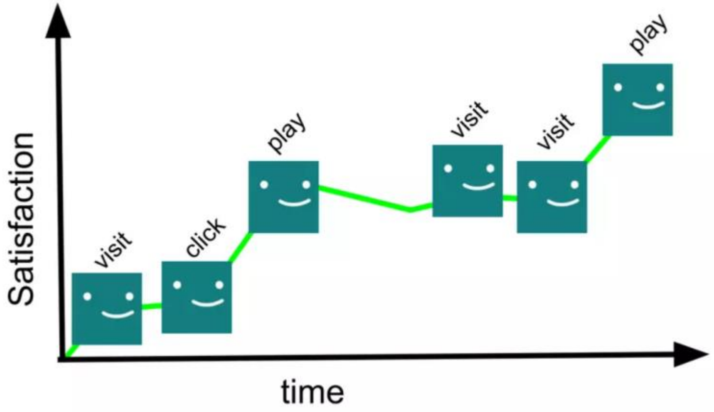
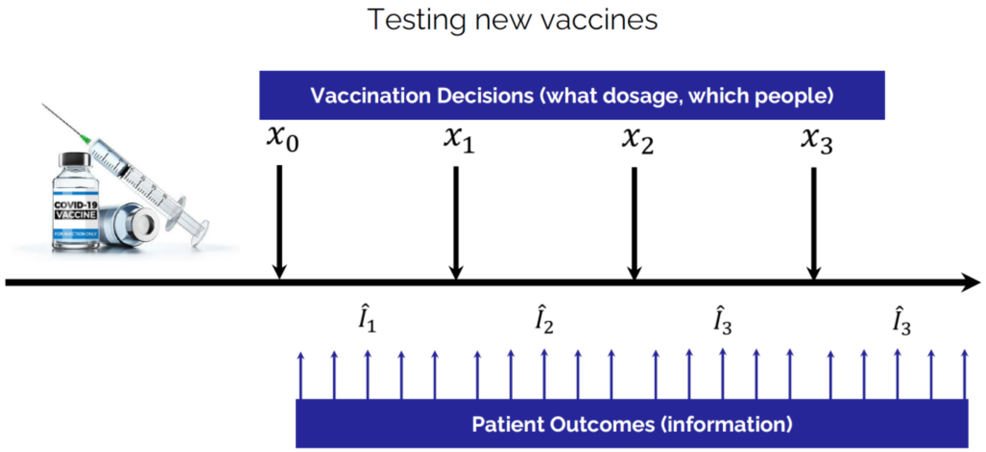


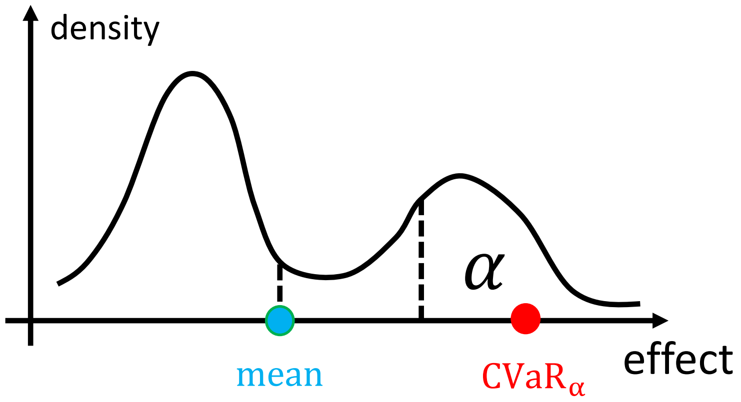
Figure by courtesy of *Warren B Powell "A Unified Framework for Sequential Decisions under Uncertainty"*
And www.slideshare.net/JayaKawale/sequential-decision-making-in-recommendations

SDM under Risk

Risk is crucial in some high-stake applications

Clinical trial/Healthcare





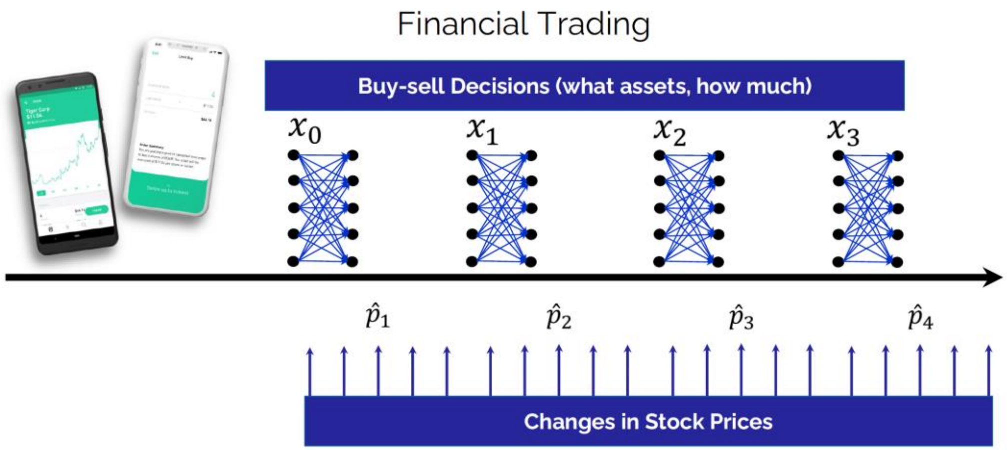
Avoid extreme negative outcomes

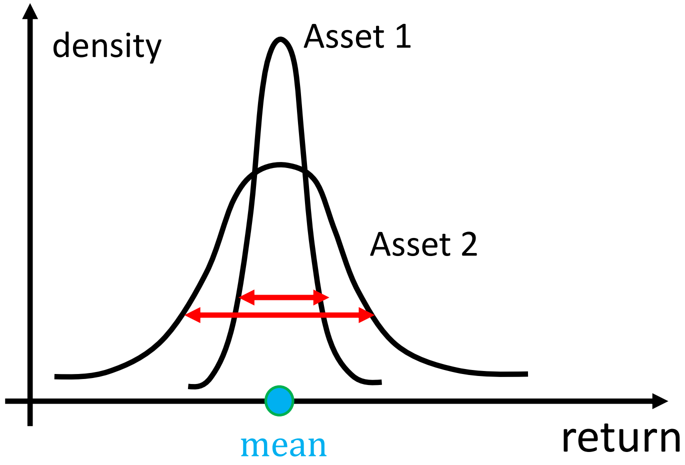
Figure by courtesy of Warren B Powell "A Unified Framework for Sequential Decisions under Uncertainty"

SDM under Risk

Risk is crucial in some high-stake applications

Finance: stock trading





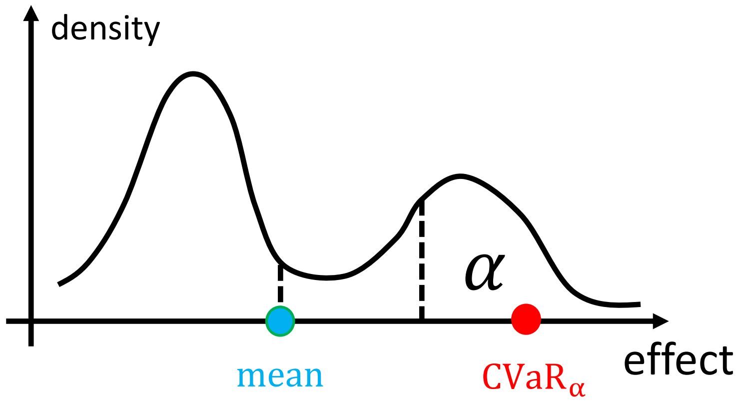
Control volatility

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

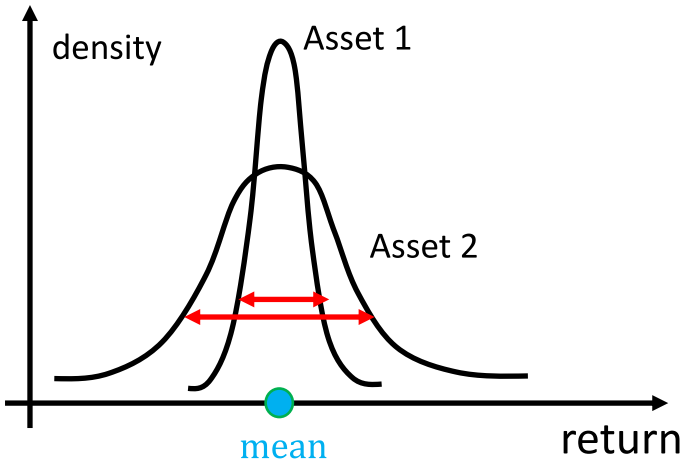
Risk-neutral vs. Risk-aware SDM

Risk neutrality only considers mean

Healthcare/Clinical trial



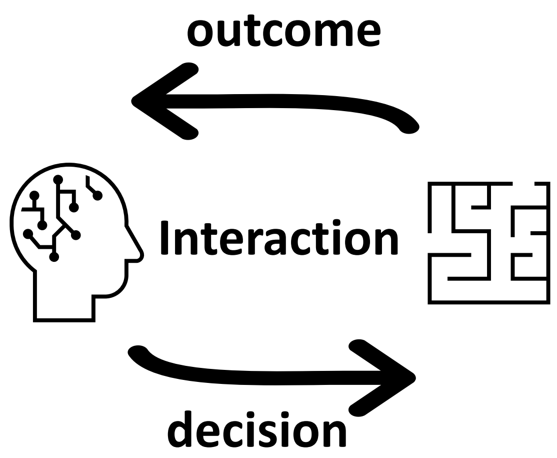
Finance



~~Mean~~ Risk measures captures certain **distributional** characteristics

- Tail mean: extreme negative outcomes
- Higher order moment: violation

Towards **Efficient** Risk-aware SDM



Sample and **computational** efficiency is critical!

- Financial trading
- Healthcare monitoring systems
- Online advertising

Question: How to attain efficient risk-aware SDM?

Methodology: A Distributional Perspective

Universality

Improved sample efficiency

Computational efficiency

Outline of Research

Risk Awareness

+

SDM

Risk Estimation

↓

Risk-aware Bandits

↓

Risk-aware Reinforcement Learning

Confidence interval of risk [**L**L23a]

Generic risk measures [**L**L23a]

Entropic risk measure [**L**L21][**L**L22]

Dynamic risk measure [**L**L23b][**L**L24]

Risk Estimation

Risk Estimation

Confidence interval of risk
[**L**L23a]

Risk-aware Bandits

Generic risk measures
[LL23a]

Risk-aware
Reinforcement Learning

Entropic risk measure
[LL21][LL22]

Dynamic risk measure
[LL23b][LL24]

Estimation of Risk Measures (RM)

RM reflects the risk preference towards uncertainty

Cannot evaluate the RM exactly

- Unknown distribution
- Finite samples

Confidence interval (CI) of RM

- Provides a **reliable range** in risk-aware context
- Allows **better decision-making**

Towards Tight Confidence Intervals

Problem setting

- A risk measure ρ assigns risk value $\rho(F)$ to a **bounded** distribution $F \in D(a, b)$
- Given n iid samples $X_1, X_2, \cdots, X_n \sim F$
- Goal: Derive **CI** of $\rho(F)$ given $X_1, X_2, \cdots, X_n$

$$\boxed{l(\rho)} \leq \rho(F) \leq \boxed{u(\rho)} \text{ w.h.p.}$$

Classical concentration bounds on mean

$$|\mu - \hat{\mu}_n| \leq c(\mu) \Rightarrow$$
$$\boxed{\hat{\mu}_n - c(\mu)} \leq \mu \leq \boxed{\hat{\mu}_n + c(\mu)}$$

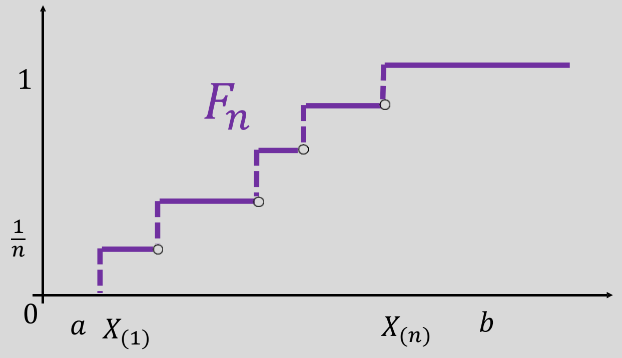
$\longrightarrow$ **Generalization to risk measures?**

- Nonlinearity
- Diversity

Global Lipschitz Constant-based Methods [LB22,LHLA22]

Step 1: Empirical distribution

$$F_n := \frac{1}{n}\sum_{i=1} \delta_{X_i}$$



Step 2: Concentration bound on F_n

$$\|F - F_n\|_p \le c_p \, (1)$$

DKW, Wasserstein bound

Step 3: Global Lipschitz constant (GLC) of ρ

$$\mathrm{GLC} = \sup_{G,G' \in D(a,b)} \frac{\rho(G)-\rho(G')}{\|G-G'\|_p} \, (2)$$

Step 4: Global linearization

$$|\rho(F) - \rho(F_n)| \underbrace{\le}_{(2)} \mathrm{GLC} \cdot \|F - F_n\|_p \underbrace{\le}_{(1)} \mathrm{GLC} \cdot c_p$$

$$\boxed{\rho(F_n) - \mathrm{GLC} \cdot c_p} \le \rho(F) \le \boxed{\rho(F_n) + \mathrm{GLC} \cdot c_p}$$

[LB22] LA, P. and Bhat, S. P. "A wasserstein distance approach for concentration of empirical risk estimates." *J. Machine Learn. Res*, 23(238):1–61, 2022.
[LHLA22] Leqi, Liu, et al. "Supervised learning with general risk functionals." *International Conference on Machine Learning*. PMLR, 2022.

A Distribution Optimization Framework

Limitations of GLC-based method

- Not easy to obtain
- **Loose** due to global linearization of **nonlinear** risk measure
- Specific for each RM

Question 1: How to obtain tight CI of generic risks?

Answer: A Distribution Optimization Framework

Universality

Tightness

Computational efficiency

A Distribution Optimization Framework [LL23a]

risk value	$\longrightarrow$	objective
distribution	$\longrightarrow$	opt variable
concentration	$\longrightarrow$	constraint

$$\max \ \rho(G)$$
$$\text{s.t.} \ \|G - F_n\|_p \leq c_p$$

Optimal solutions

- Maximizer $\overline{F_n^p}$
- Minimizer $\underline{F_n^p}$

Plug-in $\longrightarrow$

Tight CI

$$\boxed{\rho\left(\underline{F_n^p}\right)} \leq \rho(F) \leq \boxed{\rho\left(\overline{F_n^p}\right)}$$

[LL23a] Hao Liang, and Zhi-Quan Luo. "A distribution optimization framework for confidence bounds of risk measures." *International Conference on Machine Learning*. PMLR, 2023.

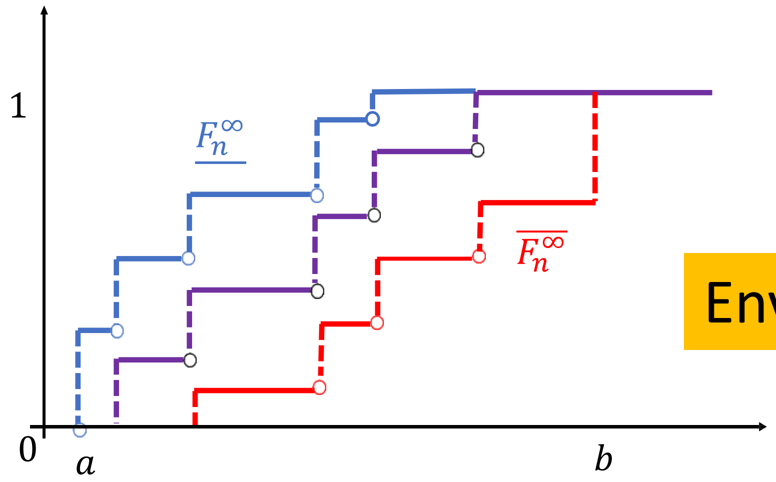
Optimal Solution

Computational challenges
- **Infinite-dimensional** CDF
- **Diverse** and **nonlinear** risk measures

→ **Can we obtain optimal solutions?**

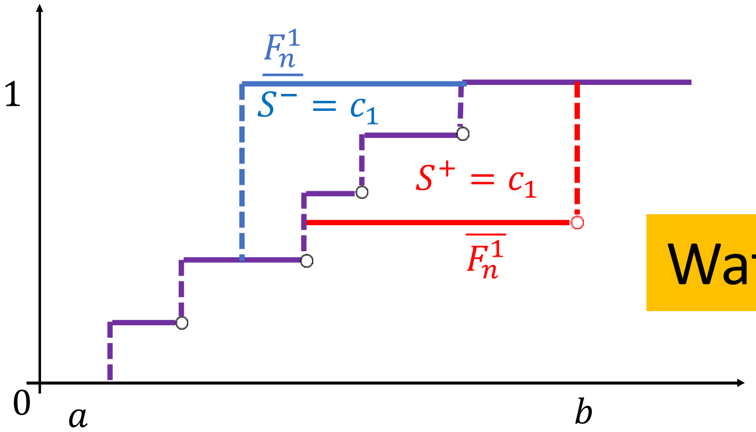
Closed form as **transformation** of F_n, computationally efficient

$$\|G - F_n\|_\infty = \sup_x |G(x) - F_n(x)| \le c_\infty$$



Envelope

$$\|G - F_n\|_1 = \int_a^b |G(x) - F_n(x)| dx \le c_1$$



Water-filling

Intrinsic Tightness

Our new baseline [LL3a]

$$\mathrm{LLC} = \sup_{G,G' \in B(F,c)} \frac{\rho(G)-\rho(G')}{\|G-G'\|_p}$$

Our bounds improve the **tightest Local Lipschitz Constant**!

$$\rho\left(\overline{F_n^p}\right) < \underbrace{\rho(F_n) + \mathrm{LLC} \cdot c_n^p}_{\text{LLC bound}} < \underbrace{\rho(F_n) + \mathrm{GLC} \cdot c_n^p}_{\text{GLC bound}}$$

LLC vs. GLC

RM	LLC ($p = \infty$)	GLC ($p = \infty$)	Improvement
CVaR	$\frac{b-F_n^{-1}((1-\alpha-c)^+)}{\alpha}$	$\frac{b-a}{\alpha}$	✓
SRM	$\left\|\phi(\underline{F_n^\infty})\right\|_1$	$(b-a)\phi(1)$	✓
DRM	$\left\|g'(1-\underline{F_n^\infty})\right\|_1$	$(b-a)\,\|g'\|_\infty$	✓
ERM	$\frac{\exp(\beta b)-\exp(\beta a)}{\beta \int_a^b \exp(\beta x) d\overline{F_n^\infty}(x))}$	$\frac{\exp(\beta(b-a))-1}{\beta}$	✓
RDEU	$\left\|w'(\underline{F_n^\infty})v'\right\|_1$	$\|w'\|_\infty \|v'\|_1$	✓

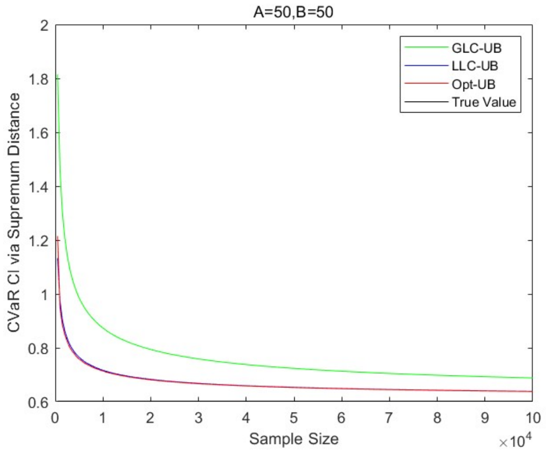
Ours vs. LLC

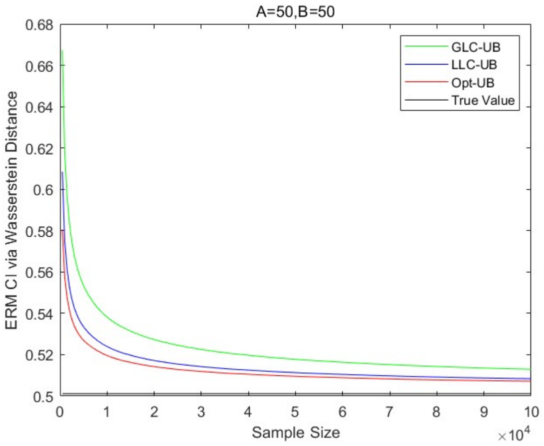
RM	CVaR	SRM	DRM	ERM	RDEU
LLC	$\frac{b-F^{-1}(1-\alpha-c)}{\alpha}$	$\|\phi(\underline{F^\infty})\|_1$	$\|g'(1-\underline{F^\infty})\|_1$	$\frac{\exp(\beta b)-\exp(\beta a)}{\int_a^b \exp(\beta x) d\underline{F^\infty}(x)}$	$\|w'(\underline{F^\infty})v'\|_1$
$\frac{\mathbf{T}(\overline{F^\infty})-\mathbf{T}(F)}{c}$	$\frac{b-F^{-1}(1-\alpha)}{\alpha}$	$\|\phi(F)\|_1$	$\|g'(1-F)\|_1$	$\frac{\exp(\beta b)-\exp(\beta a)}{\beta \int_a^b \exp(\beta x) dF(x)}$	$\|w'(F)v'\|_1$
Improvement	✓	✓	✓	✓	✓

Numerical Experiments

Comparisons of CIs for **CVaR** and **ERM** with varying sample sizes

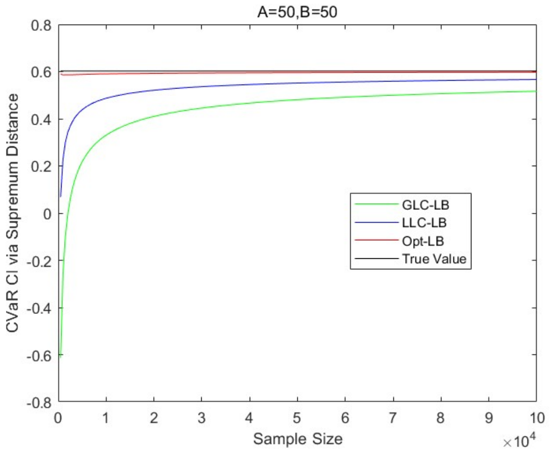
CVaR UCB w/ $\|\cdot\|_\infty$

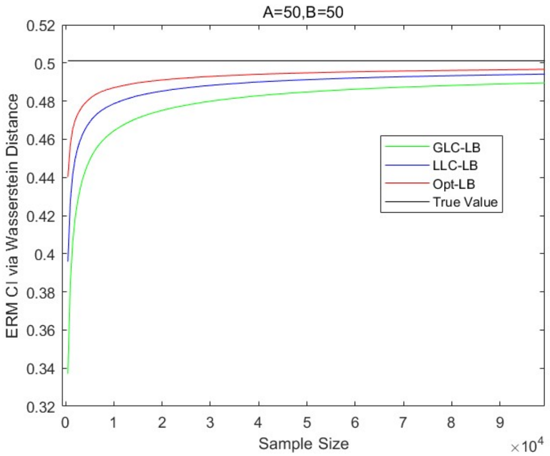




ERM UCB w/ $\|\cdot\|_1$

CVaR LCB w/ $\|\cdot\|_\infty$





ERM LCB w/ $\|\cdot\|_1$

Risk-aware SDM： Bandits

Risk Estimation

Confidence interval of risk
[LL23a]

↓

Risk-aware Bandits

Generic risk measures
[**L**L23a]

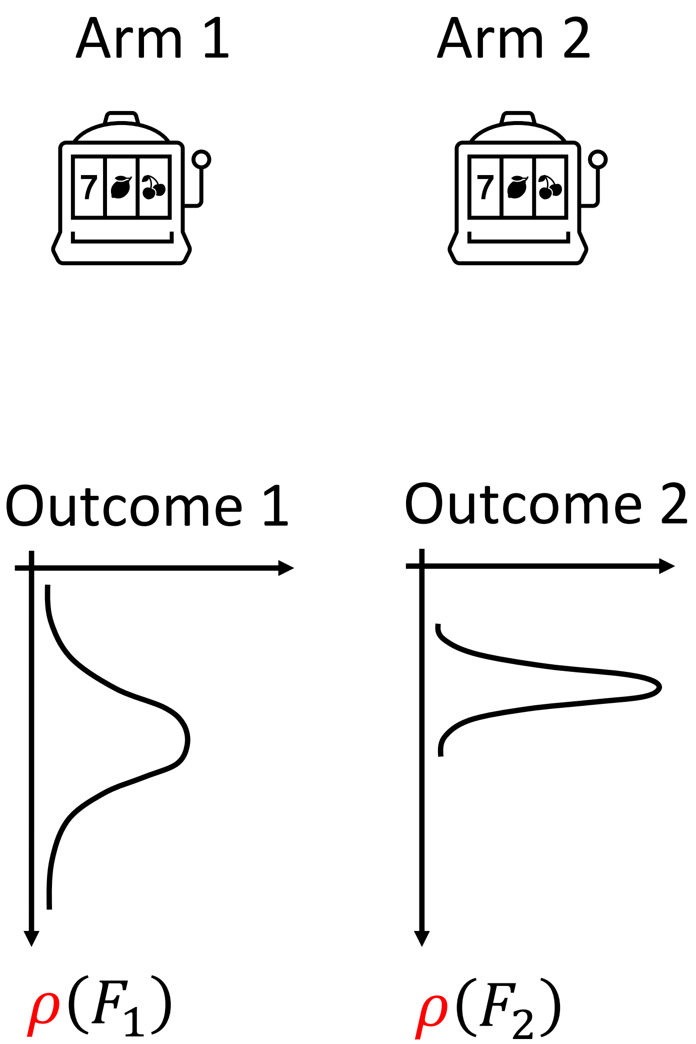
↓

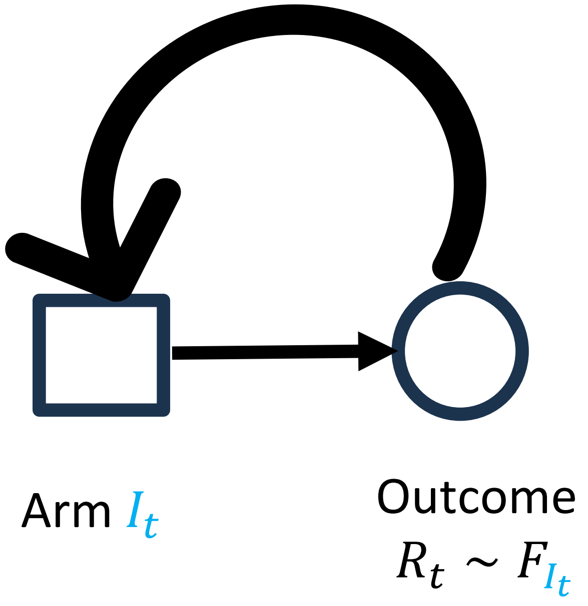
Risk-aware
Reinforcement Learning

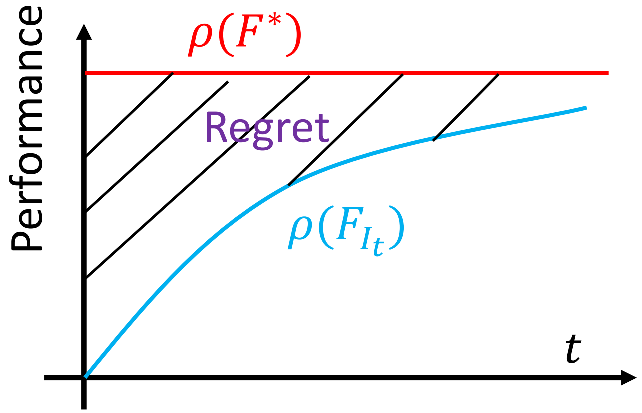
Entropic risk measure
[LL21][LL22]

Dynamic risk measure
[LL23][LL24]

Risk-aware Multi-armed Bandits







Maximize cumulative value

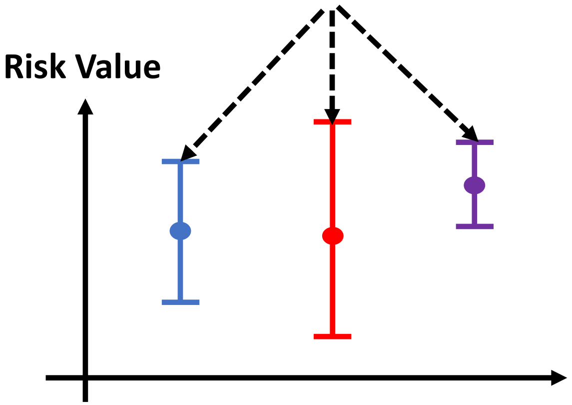
$$\sum_{t=1}^{N} \rho(F_{I_t})$$

Low regret

↓

High sample efficiency

Optimism in Face of Uncertainty (OFU) in SDM



- Act greedily w.r.t. Upper Confidence Bound of **Risk Value**
- Explore actions with the **best possible** outcomes

Tighter UCB ⟶ Less Optimism ⟶ Higher efficiency

Improved risk estimation ⟶ Better decision-making

Meta Bandit Algorithm for Generic Risk Measures

Upper Confidence Band [LL23a]

For $t = 1:N$

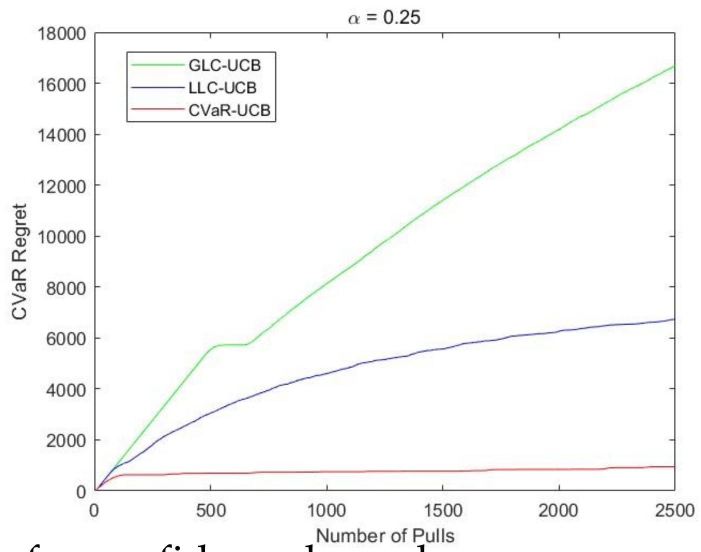
- Maintain EDF for each arm $\widehat{F_{i,t}}$
- Choose action

$$I_t = \operatorname{argmax}_{i\in[K]} \rho\left(\overline{F_{i,t}}\right)$$

Regret gain

$$\frac{\sum_{i>1} \mathrm{GLC}^2(\rho)/\Delta_i}{\sum_{i>1} \mathrm{LLC}^2(\rho; F_i, 2c_i^*)/\Delta_i}$$

Numerical experiments



[LL23a] Hao Liang, and Zhi-Quan Luo. "A distribution optimization framework for confidence bounds of risk measures." *International Conference on Machine Learning*. PMLR, 2023.

Risk-aware RL with Entropic Risk Measure

Risk Estimation

Confidence interval of risk
[LL23a]

Risk-aware Bandits

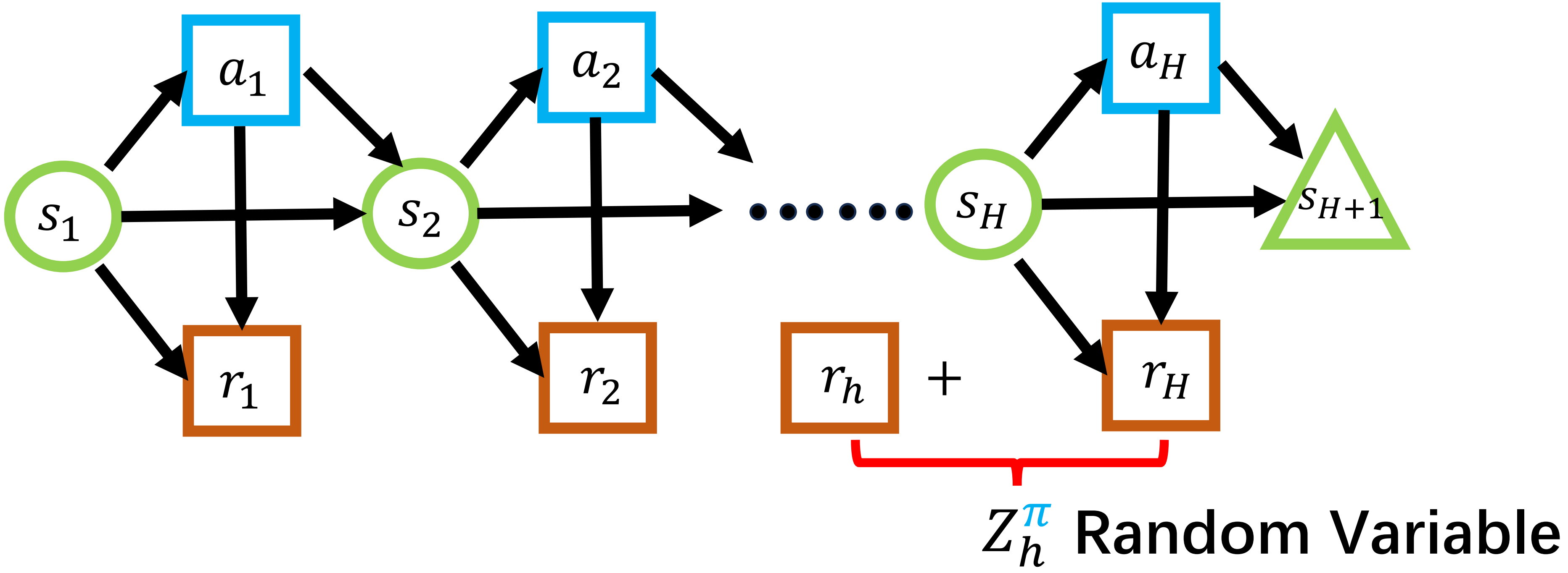
Generic risk measures
[LL23a]

Risk-aware
Reinforcement Learning

Entropic risk measure
[**L**L21][**L**L22]

Dynamic risk measure
[LL23b][LL24]

Markov Decision Process (MDP)



Tabular MDP $\boldsymbol{M} = (S, A, P, r, H)$

- Finite state space S, action space A
- Transition kernel $P_h(s,a)$

  $s' \sim P_h(s,a)$
- Reward function $r_h(s,a)$
- Horizon H

- 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

$\max \ \mathrm{E}[Z_1^\pi]$

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.

Risk-aware MDP: Optimality

Risk-neutral optimality equation → **Optimal substructure**

$$Q_h^*(s,a) = r_h(s,a) + \sum P_h(s'|s,a) V_{h+1}^*(s')$$
$$V_h^*(s) = \max_a Q_h^*(s,a)\,, V_{H+1}^*(s) = 0$$

- Break into **multiple single-stage** problems
- Recursion of value functions

Question 2: Optimal substructure for risk-aware MDP?

Answer: Yes. Distributional dynamic programing

Distributional Dynamic Programming: Policy Evaluation

Return = reward + future return

Recursion of **R.V.s**

$$Z_h(s,a) = r_h(s,a) + Y_{h+1}(S')$$
$$S' \sim P_h(\cdot \mid s,a)$$
$$Y_h(s) = Z_h(s,\pi_h(s))$$

Recursion of **distributions**

$$\eta_h(s,a) = \sum P_h(s' \mid s,a)\nu_{h+1}(s')(\cdot - r_h(s,a))$$
$$\nu_h(s) = \eta_h(s,\pi_h(s))$$

mixture **shift**

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

$$\eta_h(s,a) = [\mathrm{T}_d \nu_{h+1}](s,a)$$

Distributional Dynamic Programming: Risk-aware Control

$$\max_{\pi} \quad \boldsymbol{U_\beta}\,(Z_1^\pi)$$

Key property 1: Additivity

Key property 2: Independence

Distributional Bellman Optimality Equation [LL21]

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

backward recursion

greedy is optimal

[LL21] Hao Liang, and Zhi-Quan Luo. " Model-based Distributional Reinforcement Learning for Risk-sensitive Control." NeurIPS 2021 Workshop on Ecological Theory of RL.

Risk-aware Optimistic Distribution Iteration (RODI)

Approximate Bellman recursion

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

Distributional Optimism Operator

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

Policy Execution

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

RODI [LL22]

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

Optimism

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

[LL22] Hao Liang, and Zhi-Quan Luo. "Bridging distributional and risk-sensitive reinforcement learning with provable regret bounds." arXiv preprint arXiv:2210.14051v3 (2022). Under review at *Journal of Machine Learning Research*.

Regret Lower Bound: Fundamental Hardness

$T := KH$ **total time steps**

[FWCWX20]

$$\mathrm{E}[\mathrm{Regret}(K)] \geq \Omega\left(\frac{\exp(|\beta|H/2)-1}{|\beta|}\sqrt{K\log K}\right)$$

Missing S, A

Loose dependency on H

- Reduction to **2-armed bandit**

[LL22]

$$\mathrm{E}[\mathrm{Regret}(K)] \geq \Omega\left(\frac{\exp(\beta H/6)-1}{\beta}\sqrt{SAT}\right)$$

Fundamental trade-off between risk awareness and sample complexity

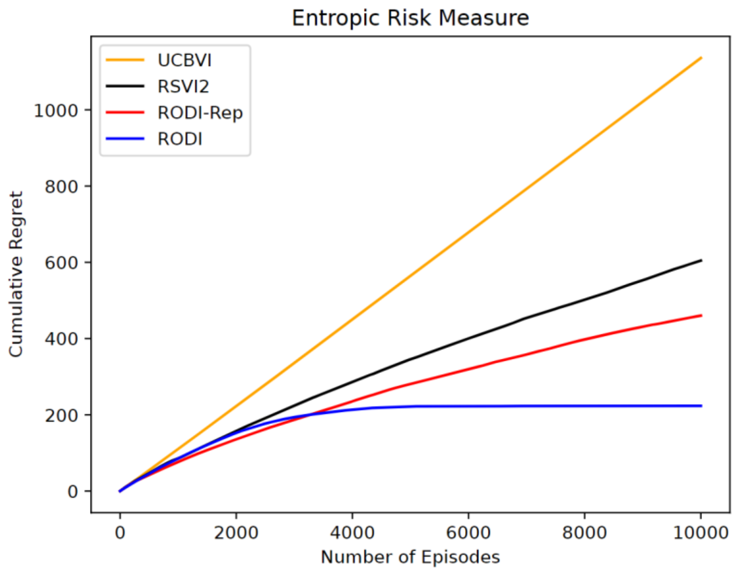
- Fix and tighten the previous result
- Recover tight risk-neutral result
- Hold for $\beta > 0$

[FWCWX20] Fei, Yingjie, et al. "Risk-sensitive reinforcement learning: Near-optimal risk-sample tradeoff in regret." *Advances in Neural Information Processing Systems* 33 (2020): 22384-22395.

Regret Upper Bound: Performance Guarantee

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

- First regret analysis of DRL
- Matching the **best known result** in [FYCW21]
- Computational efficiency
- Outperform **RSVI2** [FYCW21] empirically



[FYCW21] 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.

Risk-aware RL with Dynamic Risk Measure

Risk Estimation

Confidence interval of risk
[LL23a]

Risk-aware Bandits

Generic risk measures
[LL23a]

Risk-aware
Reinforcement Learning

Entropic risk measure
[LL21][LL22]

Dynamic risk measure
[**L**L23b][**L**L24]

Risk-aware RL with Dynamic Risk Measure (DRM)

General static risk measure may NOT support Bellman equation

$$\max_{\pi} \boldsymbol{\rho}(Z_1^{\pi}) = \boldsymbol{\rho}(r_1 + \cdots + r_H) \neq \max_{\pi_1} \boldsymbol{\rho}(r_1) + \max_{\pi_2 \cdots \pi_H} \boldsymbol{\rho}(r_2 + \cdots + r_H)$$

Dynamic risk measure assigns values via a recursive application of $\boldsymbol{\rho}$

$$Q_h^*(s,a) = r_h(s,a) + \boldsymbol{\rho_h}\left(V_{h+1}^{\pi}(S')\right)$$
$$V_h^*(s) = \max Q_h^*(s, \pi_h(s)), V_{H+1}^*(s) = 0$$

Question 3: Can we design RaRL algorithms for general DRM

Answer: Yes. Lipschitz continuous risk measure

$$|\boldsymbol{\rho}(F) - \boldsymbol{\rho}(G)| \leq L_{p,M} \cdot \|F - G\|_p, \forall F, G \in D(0, M)$$

Optimistic Value Iteration with DRM (OVI-DRM)

Optimistic Model

$$\tilde{P}_h^k \leftarrow \mathbf{OM}(\hat{P}_h^k, V_{h+1}^k, c_h^k)$$

Bellman Recursion

$$Q_h^k(s,a) \leftarrow r_h(s,a) + \boldsymbol{\rho_h}(V_{h+1}^k, \tilde{P}_h^k(s,a))$$

$$V_h^k(s) = \max Q_h^k(s,a)$$

Policy Execution

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

Optimism

$$Q_h^k(s,a) \geq Q_h^*(s,a)$$

[LL24] Hao Liang, and Zhi-Quan Luo, Regret Bounds for Risk-sensitive Reinforcement Learning with Lipschitz Dynamic Risk Measures, *AISTATS 2024*.

Regret Analysis

Worst-case regret bound of OVI-DRM

$$\text{Regret}(K) \leq O\left(\sum_{h=1}^{H-1} L_{\infty,h}\, \widetilde{L}_{1,h-1} \sqrt{S^2 A K}\right)$$

$$\widetilde{L}_{1,h-1} := \prod_{i=1}^{h-1} L_{1,i}$$

Minimax Lower Bound

$$\mathrm{E}[\text{Regret}(K)] \geq \Omega(c_\rho H \sqrt{SAT})$$

OVI-DRM

$$\text{Regret}(K) \leq O\left(\frac{S^2 A H \left(\sum_{h=1}^{H-1} L_{\infty,h}\, \widetilde{L}_{1,h-1}\right)^2}{\Delta_{min}} \log(SAT)\right)$$

Lower Bound

$$\lim_{K\to\infty} \frac{\text{Regret}}{\log K} \geq \Omega\left(\sum_{s,a:\Delta_1(s,a)>0} \frac{(c_\rho H)^2}{\Delta_1(s,a)}\right)$$

Summary

Risk Estimation

↓

Risk-aware Bandits

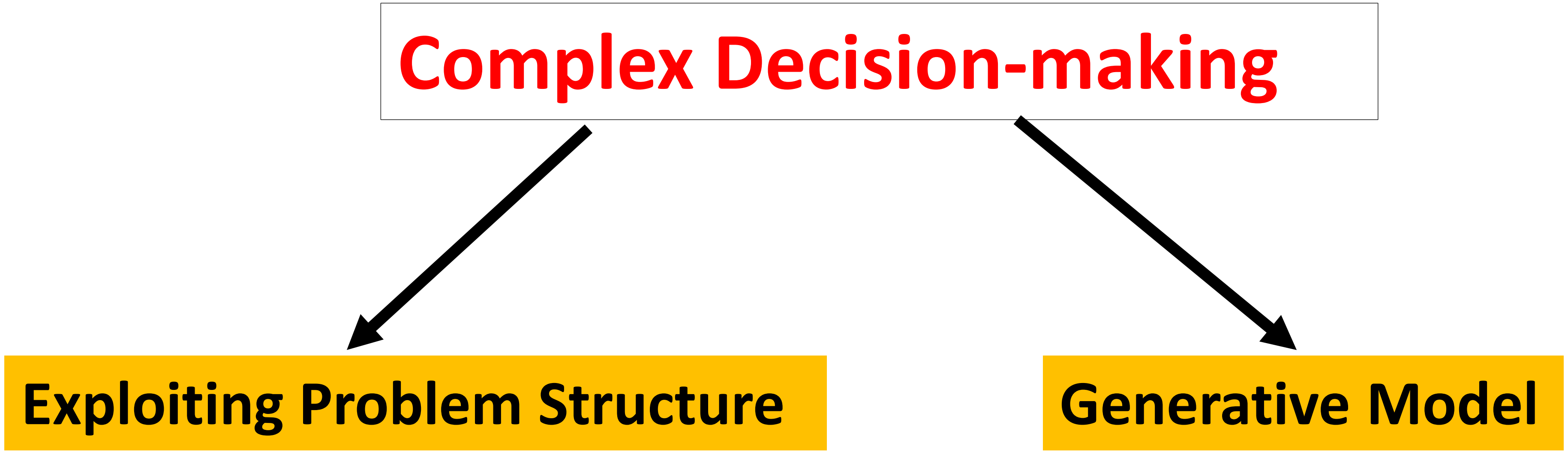
↓

Risk-aware Reinforcement Learning

Distributional perspective facilities the design of algorithms

- Universality
- Finer risk estimation
- Improved sample efficiency
- Computational efficiency

Research Plan



Exploiting Problem Structure

- Inherent structure improves efficiency
  - Optimal value structure: **Lipschitz continuity** [SBY22], **monotonicity** [JP15], **convexity** [P19],…
  - System model: **deterministic** [TP21], **exo-mdp** [S23],…
  - Optimal policy: **monotonicity** [AP22],…
- Common in real world applications
  - Operations Research: optimal replacement [FR74], batch servicing of customers [PP02]
  - Energy: energy storage and allocation [SP12]
  - Healthcare: optimal dosing of glycemic control [H10], managing patient service [G06]
  - Finance, Economics…

Combining Generative Model with Decision-making

Generative AI has led to significant advances in NLP, vision, audio, and video

- Generative Models for Decision Making
  - LLMs: planning, reward generation, simulation
  - Diffusion Models: planning, RL, and robotic control
  - Sample Efficiency, Exploration: long-horizon, high-dimensional and sparse reward
- LLMs and Human/Social behavior
  - LLMs for Human/Social behavior: voting, opinion dynamics, …
  - Human/Social behavior for LLM: behavioral economics, social choice theory
- Planning and Risk in LLM agent, LLMs in multi-agent environments…

Thank You