Safe RL: Risk and Robustness

Marek Petrik

Department of Computer Science
University of New Hampshire

DLRL Summer School 2023

Safety in RL: Risk and Robustness

Objective: Deploy RL in high-stakes domains

- Health care: automating and improving ER care
- Finance: profitable and safe investments
- Agriculture: profitably grow crops mitigating failure

Safe RL: Compute policies that mitigate return *variability*

1. *Aleatory uncertainty* is inherent to the environment
2. *Epistemic uncertainty* about the model of environment

Markov Decision Process

Model (tabular in this talk)
States $\mathcal{S}$: $s_1, s_2, s_3, \ldots$
Actions $\mathcal{A}$: $a_1, a_2, \ldots$
Transition probabilities p
Rewards r

Solution: Policy $\pi\colon \mathcal{S} \to \mathcal{A}$ (randomized in general)

Return: Discounted random return (random over trajectories):

$$\tilde{\rho}(\pi) = \sum_{t=0}^{\infty} \gamma^t r(\tilde{s}_t^{\pi}, \tilde{a}_t^{\pi})$$

Random variables: $\tilde{\rho}, \tilde{s}, \tilde{a}, \tilde{x}, \ldots$ adorned with tilde

Managing Pest Population with RL

MDP Model

- *States*: Pest population, weather, ...
- *Actions*: How much and which pesticide
- *Transitions*: Pest population dynamics
- *Reward*: Crop yield minus pesticide cost

Challenges

- Stochastic environment, delayed rewards, no reliable simulator
- One episode = one year
- Crop failure can be catastrophic

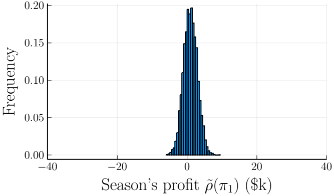
Uncertainty

- *Aleatory uncertainty*: Weather, like temperatures and rain
- *Epistemic uncertainty*: Response of pest to pesticides

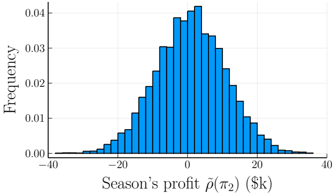
Limitation of Expected Return

Standard RL objective: $\max_\pi \mathbb{E}[\tilde{\rho}(\pi)]$

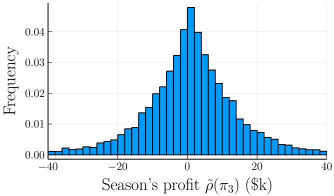
$\mathbb{E}[\tilde{\rho}(\pi_1)] = 1$



$\mathbb{E}[\tilde{\rho}(\pi_2)] = 1$



$\mathbb{E}[\tilde{\rho}(\pi_3)] = 1$



This Talk

Computing policies that mitigate return *variability*

Outline

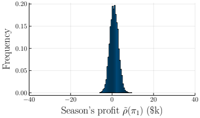
1. *Risk measures*: Measure variability
2. *Risk-averse RL*: Mitigate aleatory uncertainty
3. *Robust RL*: Mitigate epistemic uncertainty

Caution: Mathematical precision matters because ordinary RL intuition fails with risk-aversion

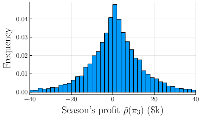
Risk Measures

Measuring Variability of Random Variable

$\mathbb{E}[\tilde{\rho}(\pi_1)] = 1$



$\mathbb{E}[\tilde{\rho}(\pi_3)] = 1$



Variance $\mathbb{V}[\tilde{\rho}(\pi)]$: natural but inflexible and also penalizes upside

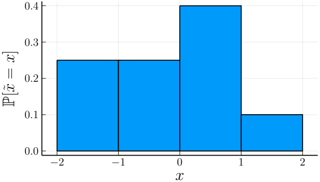
Expected utility $u^{-1}(\mathbb{E}[u(\tilde{\rho}(\pi))])$: powerful but difficult to interpret and optimize

Worst case $\min[\tilde{\rho}(\pi)]$: simple but inflexible and overly conservative

Monetary risk measure $\mathrm{Risk}[\tilde{\rho}(\pi)]$: generalize $\mathbb{E}$ as a maps of random variable to $\mathbb{R}$.

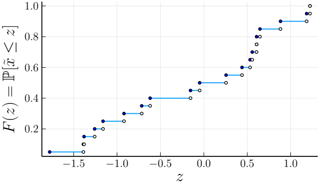
Statistics of Random Variable

Probability $\mathbb{P}[\tilde{x}]$



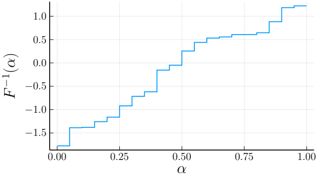
CDF

$$F(z) = \mathbb{P}[\tilde{x} \le z]$$



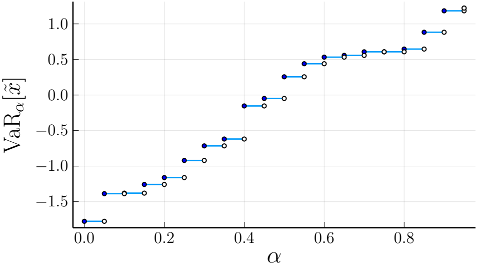
Quantile

$$F^{-1}(\alpha) = \left\{ t \; : \; \begin{array}{l} \mathbb{P}[\tilde{x} \le t] \ge \alpha, \\ \mathbb{P}[\tilde{x} \ge t] \ge 1 - \alpha \end{array} \right\}$$



Basic Risk Measure: Value at Risk (VaR)

$$\mathrm{VaR}_{\alpha}[\tilde{x}] = \sup F^{-1}(\alpha) = \sup\{t \in \mathbb{R} \;:\; \mathbb{P}[\tilde{x} \ge t] \ge 1 - \alpha\}$$



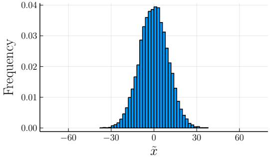
$\mathrm{VaR}_{\alpha}[\tilde{x}]$ = best α-confidence lower bound on $\tilde{x}$

$\mathrm{VaR}_{0.2}[\tilde{x}] = -1.2$ means that 80% of time return is at least -1.2

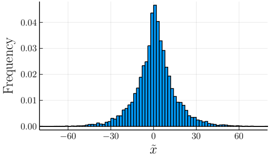
$\mathrm{VaR}_0[\tilde{x}] = \operatorname{ess\,inf}[\tilde{x}]$ $\quad \mathrm{VaR}_{\frac{1}{2}}[\tilde{x}] \approx \mathrm{median}[\tilde{x}]$ $\quad \mathrm{VaR}_1[\tilde{x}] = \infty$

Limitations of VaR

1. **VaR ignores the tail** and catastrophic risk



$\mathrm{VaR}_{0.2}[\tilde{x}] = -8.2$



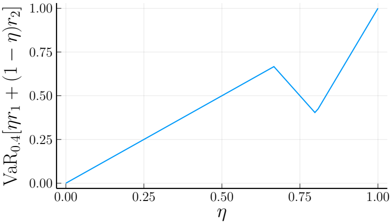
$\mathrm{VaR}_{0.2}[\tilde{x}] = -8.2$

2. **Difficult to optimize**

Stock returns (equal probs.)

	$\tilde{r}_1$	$\tilde{r}_2$
ω_1	1	0
ω_2	1	-2
ω_3	0	2

$\max_{\eta \in [0,1]} \mathrm{VaR}_{0.4}[\eta \tilde{r}_1 + (1-\eta)\tilde{r}_2]$



Concave Risk Measures

Easier to optimize and consider distribution's tail

CVaR: Conditional Value at Risk

$$\begin{aligned}\mathrm{CVaR}_\alpha[\tilde{x}] &= \sup_{z\in\mathbb{R}}\left(z - \frac{1}{\alpha}\mathbb{E}[z-\tilde{x}]_+\right)\\ &\approx \mathbb{E}[\tilde{x} \mid \tilde{x} \le \mathrm{VaR}_\alpha[\tilde{x}]]\end{aligned}$$

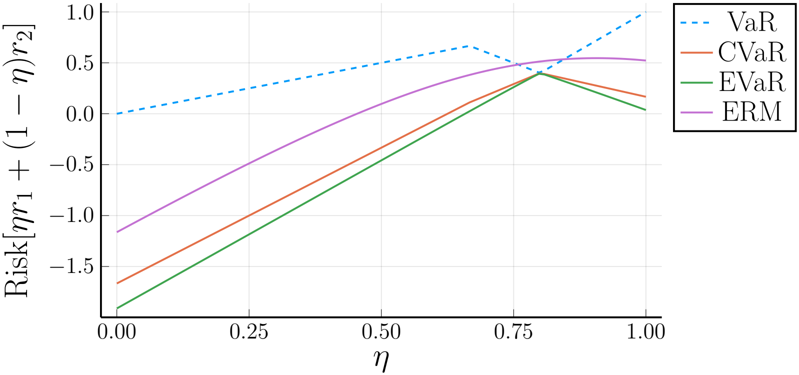
ERM: Entropic risk measure

$$\mathrm{ERM}_\beta[\tilde{x}] = -\beta^{-1}\log\mathbb{E}[\exp(-\beta\tilde{x})], \quad \beta > 0.$$

EVaR: Entropic value at risk

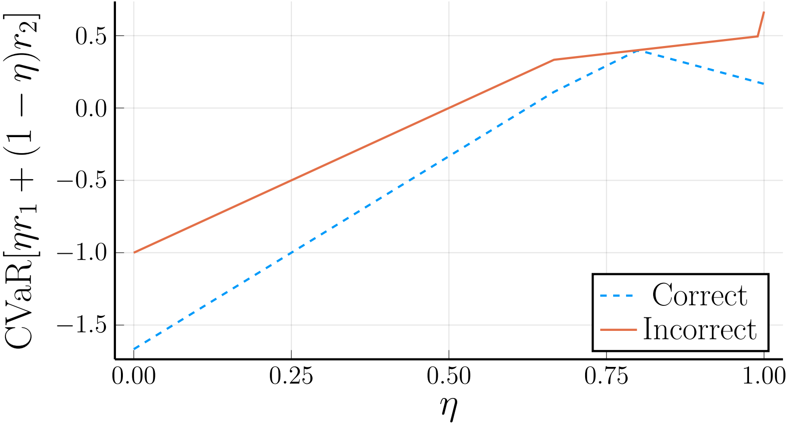
$$\mathrm{EVaR}_\alpha[\tilde{x}] = \sup_{\beta>0}\left(\mathrm{ERM}_\beta[\tilde{x}] + \beta^{-1}\log\alpha\right).$$

Concave Risk Measures: Portfolio Example



Correct CVaR Definition

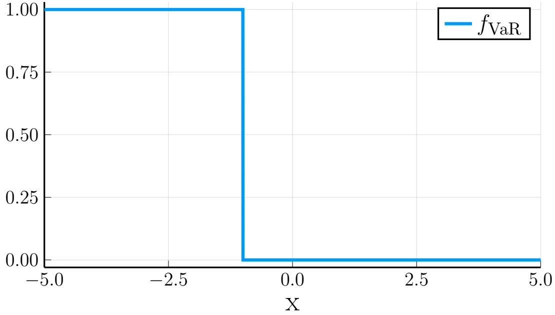
$$\mathrm{CVaR}_\alpha[\tilde{x}] = \sup_{z\in\mathbb{R}} \left(z - \frac{1}{\alpha}\mathbb{E}[z-\tilde{x}]_+ \right)$$
$$\neq \mathbb{E}[\tilde{x} \mid \tilde{x} \le \mathrm{VaR}_\alpha[\tilde{x}]] \text{ for discrete } \tilde{x}$$



EVaR & CVaR: Approximate Value at Risk

$$\begin{aligned}\mathrm{VaR}_\alpha[\tilde{x}] &= \inf\{t \in \mathbb{R} \;:\; \mathbb{P}[\tilde{x} \le t] > \alpha\} \\ &= \inf\{t \in \mathbb{R} \;:\; \mathbb{E}[f_{\mathrm{VaR}}(\tilde{x}; t)] > \alpha\}\end{aligned}$$

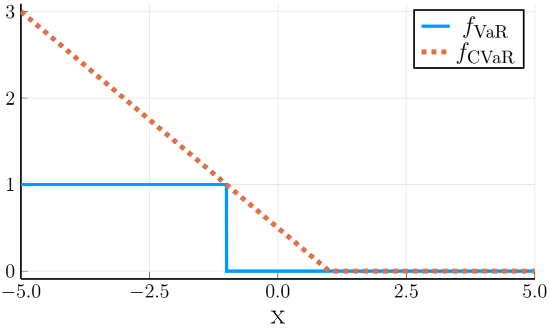
$$f_{\mathrm{VaR}}(x; t) = \mathbb{1}_{x \le t}$$



CVaR Bounds VaR (Markov's Inequality)

$$\mathrm{VaR}_\alpha[\tilde{x}] \geq \sup_{z\in\mathbb{R}} \inf\{t \;:\; \mathbb{E}[f_{\mathrm{CVaR}}(\tilde{x}; t, z)] > \alpha\} = \mathrm{CVaR}_\alpha[\tilde{x}]$$

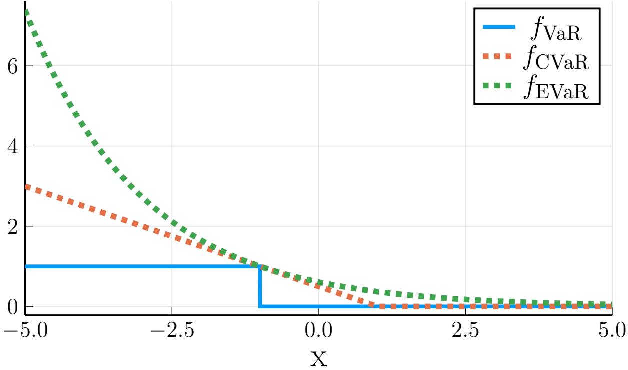
$$f_{\mathrm{CVaR}}(x; t, z) = \frac{[z - x]_+}{[z - t]_+}$$



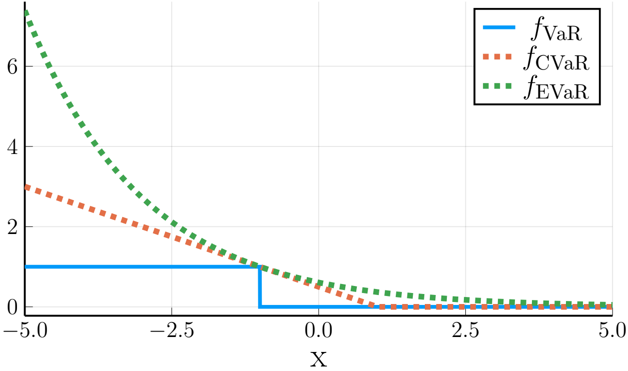
EVaR Bounds VaR (Chernoff Bound)

$$\mathrm{VaR}_\alpha[\tilde{x}] \geq \sup_{\beta \in \mathbb{R}} \inf \{t \in \mathbb{R} \ : \ \mathbb{E}[f_{\mathrm{EVaR}}(\tilde{x}; t, \beta)] > \alpha\} = \mathrm{EVaR}_\alpha[\tilde{x}]$$

$$f_{\mathrm{EVaR}}(x; t, \beta) = e^{\beta t} \cdot e^{-\beta x}$$



Hierarchy of Risk Measures



For any r.v. $\tilde{x}$ and $\alpha \in [0, 1]$

$$\mathrm{VaR}_{\alpha}[\tilde{x}] \quad \geq \quad \mathrm{CVaR}_{\alpha}[\tilde{x}] \quad \geq \quad \mathrm{EVaR}_{\alpha}[\tilde{x}]$$

Common Risk Measures

Property	$\mathbb{E}$, min	VaR	CVaR	ERM	EVaR
Translation invariance	✓	✓	✓	✓	✓
Monotonicity	✓	✓	✓	✓	✓
Positive homogeneity	✓	✓	✓	✗	✓
Concavity	✓	✗	✓	✓	✓
Coherence	✓	✗	✓	✗	✓
Tower property	✓	✗	✗	✓	✗

Risk-averse RL

Risk-averse Reinforcement Learning

Return: Discounted random return (random variable):

$$\tilde{\rho}(\pi) = \sum_{t=0}^{\infty} \gamma^t r(\tilde{s}_t^\pi, \tilde{a}_t^\pi)$$

Risk neutral RL: Maximize *expected* return

$$\max_{\pi} \; \mathbb{E}\left[\tilde{\rho}(\pi)\right]$$

Risk-averse RL: Maximize high-confidence *guarantee* on the return

$$\max_{\pi} \; \mathrm{VaR}_\alpha \left[\tilde{\rho}(\pi)\right]$$

Risk-averse RL

$$\max_{\pi} \; \mathrm{VaR}_{\alpha}\left[\tilde{\rho}(\pi)\right]$$

Difference from ordinary RL:

1. Optimal policy is history-dependent
2. No optimal stationary policy
3. No notion of value function
4. No Bellman optimality equation
5. NP hard to compute optimal policy

Risk-Neutral RL: Dynamic Programming

Optimal value function

$$v_t^\star(s) \;=\; \max_{\pi} \mathbb{E}\left[\sum_{t'=t}^{T} \gamma^{t'-t} r(\tilde{s}_{t'}, \pi_t(\tilde{s}_{t'})) \mid \tilde{s}_t = s\right]$$

Dynamic program: Compute optimal $v^\star$ efficiently

$$v_t^\star(s) \;=\; \max_{a\in\mathcal{A}} \left(r(s,a) + \gamma \sum_{s'\in\mathcal{S}} p(s' \mid s,a)\cdot v_{t+1}^\star(s')\right)$$

RL use of dynamic programs

1. (Fitted) value and policy iteration, TD, Q-learning
2. Actor-critic policy gradient methods, LP formulations

Why is Dynamic Programming Possible?

$$v_t^\star(s) \;=\; \max_{a\in\mathcal{A}} r(s,a) + \gamma \sum_{s'\in\mathcal{S}} p(s' \mid s,a)\cdot v_{t+1}^\star(s')$$

Dynamic program: Compute v_t from v_{t+1} (fixed a)

$$v_0(s) = \mathbb{E}\left[r(s,a) + \gamma\cdot r(\tilde{s}_1,a) + \gamma^2\cdot r(\tilde{s}_2,a) \mid \tilde{s}_0 = s\right]$$

Use positive homogeneity and translation invariance

$$= r(s,a) + \gamma\cdot\mathbb{E}\left[r(\tilde{s}_1,a) + \gamma\cdot r(\tilde{s}_2,a) \mid \tilde{s}_0 = s\right]$$

Use tower property and translation invariance

$$= r(s,a) + \gamma\cdot\mathbb{E}\left[r(\tilde{s}_1,a) + \mathbb{E}\left[\gamma\cdot r(\tilde{s}_2,a) \mid \tilde{s}_1\right] \mid \tilde{s}_0 = s\right]$$

Recursive definition

$$= r(s,a) + \gamma\cdot\mathbb{E}\left[v_1(\tilde{s}_1) \mid \tilde{s}_0 = s\right]$$

Dynamic Programming for MDPs

1. **Tower property**

$$\mathbb{E}[\tilde{x}_1] \;=\; \mathbb{E}\left[\mathbb{E}[\tilde{x}_1 \mid \tilde{x}_2]\right]$$

2. **Positive homogeneity** for $\gamma \geq 0$

$$\mathbb{E}[\gamma \cdot \tilde{x}] \;=\; \gamma \cdot \mathbb{E}[\tilde{x}]$$

3. **Translation invariance**

$$\mathbb{E}[c + \tilde{x}] \;=\; c + \mathbb{E}[\tilde{x}]$$

Dynamic Programming for Risk-Averse RL

$$\max_{\pi} \text{ Risk}\left[\sum_{t=0}^{T} \gamma^t \, r(\tilde{s}_t, \pi_t(\tilde{s}_t))\right]$$

Properties needed for a dynamic program

Property	$\mathbb{E}$, min	VaR	CVaR	ERM	EVaR
Tower property	✓	✗	✗	✓	✗
Positive homogeneity	✓	✓	✓	✗	✓
Translation invariance	✓	✓	✓	✓	✓

Building Risk-averse Dynamic Programs

1. Use a nested risk measure
2. Use entropic risk measure (ERM)
3. Reduce to simpler risk measure
4. Dual decomposition

1. Nested Risk Measures: Pros

Nested risk measures (or Markov risk measure) for CVaR

$$\mathrm{nCVaR}_\alpha[\tilde{\rho}(\pi)] = \mathrm{CVaR}_\alpha\left[\tilde{r}_0^\pi + \mathrm{CVaR}_\alpha\left[\gamma\,\tilde{r}_1^\pi + \mathrm{CVaR}_\alpha\left[\gamma^2\,\tilde{r}_2^\pi + \dots\right]\right]\right]$$

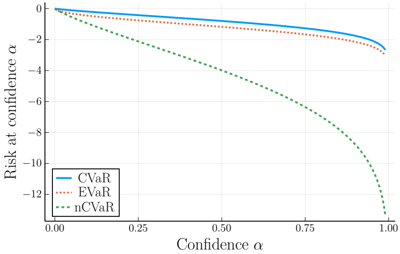
Dynamic program and value function

$$v_t^\star(s) \;=\; \max_{a\in\mathcal{A}}\left(r(s,a) + \gamma\,\mathrm{CVaR}_\alpha\left[p(\tilde{s}' \mid s,a)\cdot v_{t+1}^\star(\tilde{s}')\right]\right)$$

Ruszczynski, Andrzej. "Risk-Averse Dynamic Programming for Markov Decision Processes." Mathematical Programming B, 2010

1. Nested Risk Measures: Cons

Poor approximation of static risk



NOT law invariant

$$\tilde{\rho}(\pi_1) = \tilde{\rho}(\pi_2) \qquad \text{but} \qquad \mathrm{nCVaR}_\alpha[\tilde{\rho}(\pi_1)] \neq \mathrm{nCVaR}_\alpha[\tilde{\rho}(\pi_2)]$$

Difficult to interpret

2. ERM is Special in RL

Properties needed for dynamic programming

Property	VaR	CVaR	ERM	EVaR	Nested
Tower property	✗	✗	✓	✗	✓
Translation invariance	✓	✓	✓	✓	✓
Law invariance	✓	✓	✓	✓	✗

ERM is unique: No other risk measure checks all boxes

Note that $\mathbb{E}[\tilde{x}] = \mathrm{ERM}_0\left[\tilde{x}\right]$, $\min[\tilde{x}] = \mathrm{ERM}_\infty\left[\tilde{x}\right]$

2. Formulating ERM DP

Challenge: ERM is NOT positively homogeneous

$$\mathrm{ERM}_{\beta}\left[\gamma \cdot \tilde{x}\right] \;\neq\; \gamma \cdot \mathrm{ERM}_{\beta}\left[\tilde{x}\right]$$

Solution: ERM is positive quasi-homogeneous

$$\mathrm{ERM}_{\beta}\left[\gamma \cdot \tilde{x}\right] \;=\; \gamma \cdot \mathrm{ERM}_{\gamma \cdot \beta}\left[\tilde{x}\right]$$

2. Dynamic Program for ERM-MDPs

ERM-MDP objective

$$\max_{\pi} \ \mathrm{ERM}_{\beta}\left[\sum_{t=0}^{T} \gamma^{t}\, r(\tilde{s}_t, \pi_t(\tilde{s}_t))\right]$$

ERM Dynamic Program: Time-dependent risk level

$$v_t^{\star}(s) = \max_{a \in \mathcal{A}} \ \mathrm{ERM}_{\beta \cdot \gamma^t}\left[r(s,a) + \gamma \cdot v_{t+1}^{\star}(\tilde{s}')\right]$$

Hau, Jia Lin, Marek Petrik, and Mohammad Ghavamzadeh. "Entropic Risk Optimization in Discounted MDPs." In Artificial Intelligence and Statistics (AISTATS), 2023.

2. ERM-MDP Optimal Policies

$$\max_{\pi} \; \mathrm{ERM}_\beta \left[\sum_{t=0}^{T} \gamma^t \, r(\tilde{s}_t, \pi_t(\tilde{s}_t)) \right]$$

Theorem

Exist optimal policy that is

1. **Markov** *(history independent)*
2. **Deterministic** *(no hedging)*
3. **More risk-neutral over time**

ERM is often impractical because

1. Risk aversion depends on rewards scale (currency)
2. Hard to interpret

3. Reduce EVaR-MDP to ERM-MDP

Objective

$$\max_{\pi} \ \mathrm{EVaR}_{\alpha}\left[\sum_{t=0}^{T} \gamma^{t}\, r(\tilde{s}_t, \pi_t(\tilde{s}_t))\right]$$

Reformulate from EVaR definition

$$\sup_{\beta>0} \underbrace{\max_{\pi} \left(\mathrm{ERM}_{\beta}\left[\sum_{t=0}^{T} \gamma^{t}\, r(\tilde{s}_t, \pi_t(\tilde{s}_t))\right] + \frac{\log(1-\alpha)}{\beta}\right)}_{=h(\beta)}$$

Theorem

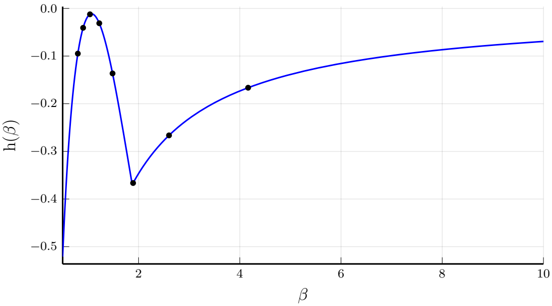
There exists EVaR-MDP optimal policy also optimal in ERM-MDP

Hau, Jia Lin, Marek Petrik, and Mohammad Ghavamzadeh. "Entropic Risk Optimization in Discounted MDPs." In Artificial Intelligence and Statistics (AISTATS), 2023.

3. EVaR-MDP Algorithm

Discretize the non-concave objective function:

$$h(\beta) = \max_{\pi} \left(\mathrm{ERM}_\beta \left[\sum_{t=0}^{T} \gamma^t \, r(\tilde{s}_t, \pi_t(\tilde{s}_t)) \right] + \frac{\log(1-\alpha)}{\beta} \right)$$



FPTAS algorithm when discretized properly

3. Numerical Results: EVaR-MDP

Method	MR	GR	INV1	INV2	RS
EVaR-MDP	**-6.73**	**5.34**	**67.4**	**189**	**303**
Risk neutral	**-6.53**	2.29	40.6	**186**	**300**
Nested CVaR	-10.00	-0.02	-0.0	132	217
Nested EVaR	-10.00	4.61	-0.0	164	217
ERM	**-6.72**	5.19	50.7	178	217
Nested ERM	-10.00	4.76	24.9	150	217
CVaR	-7.06	3.64	49.0	82	93

Similar reductions for VaR and CVaR

Bäuerle, Nicole, and Jonathan Ott. Markov Decision Processes with Average-Value-at-Risk Criteria. Mathematical Methods of Operations Research 74, no. 3 (2011): 361–79.

4. Dual Decomposition

Augment states with risk level, using

$$\max_{\pi \in \Pi} \mathrm{VaR}_\alpha[r(\tilde{s}, \tilde{a}, \tilde{s}')] =$$

$$= \sup_{\zeta \in \Delta_S} \left\{ \min_{s \in \mathcal{S}} \max_{d \in \Delta_A} \mathrm{VaR}_{\alpha \zeta_s \hat{p}_s^{-1}} \left[r(s, \tilde{a}, \tilde{s}') \mid \tilde{s} = s \right] \; : \; \alpha \cdot \zeta \leq \hat{p} \right\}.$$

Properties

+ A single DP for all risk levels α
– Only optimal and practical for VaR
– Conceptually complex

Hau, Jia Lin, Erick Delage, Mohammad Ghavamzadeh, and Marek Petrik. On Dynamic Programming Decompositions of Static Risk Measures in Markov Decision Processes. arXiv, 2023.

Building Risk-averse Dynamic Programs

1. Use a nested risk measure
2. Use entropic risk measure (ERM)
3. Reduce to simpler risk measure
4. Dual decomposition

Robust RL

MDP with Epistemic Uncertainty

Epistemic (model) uncertainty in RL: limited data, missing observations, violated Markov assumption, . . .

Random return: uncertain transitions $\tilde{p}$ and $\tilde{r}$

$$\tilde{\rho}(\pi, \tilde{p}, \tilde{r}) = \sum_{t=0}^{\infty} \gamma^t \tilde{r}(\tilde{s}_t^\pi, \tilde{a}_t^\pi) \qquad \tilde{s}_{t+1}^\pi \sim \tilde{p}(\tilde{s}_t^\pi, \tilde{a}_t^\pi)$$

Expected return: uncertain transition probabilities

$$\rho(\pi, \tilde{p}, \tilde{r}) = \mathbb{E}\left[\tilde{\rho}(\pi) \mid \tilde{p}, \tilde{r}\right]$$

Robust RL

Soft-robust RL: epistemic risk aversion

$$\max_{\pi} \operatorname{Risk}\left[\rho(\pi, \tilde{p}, \tilde{r})\right] \;=\; \operatorname{Risk}\left[\mathbb{E}\left[\tilde{\rho}(\pi) \mid \tilde{p}, \tilde{r}\right]\right]$$

Robust RL: use $\min$ as the risk measure with some $\mathcal{P}$ and $\mathcal{R}$

$$\max_{\pi} \min_{p \in \mathcal{P},\, r \in \mathcal{R}} \rho(\pi, p, r)$$

Difference from aleatory uncertainty

- Distribution over $\tilde{p}$ and $\tilde{r}$ is often unknown
- Model is unknown but does not change

Adversarial Robustness for RL

Robust optimization: Best π with respect to the inputs with *all* possible *small errors*:

$$\max_{\pi} \min_{p,r} \left\{ \rho(\pi, p, r) \;:\; \begin{matrix} \|\bar{p} - p\| \le \text{small} \\ \|\bar{r} - r\| \le \text{small} \end{matrix} \right\}$$

Game in which adversarial nature chooses p, r

Robust Representation

Nominal values: $\bar{p}$, $\bar{r}$

Robustness to rewards

$$\max_{\pi} \min_{r} \{\rho(\pi, \bar{p}, r) \ : \ \|r - \bar{r}\| \le \psi\}$$

Robustness to transitions

$$\max_{\pi} \min_{p} \{\rho(\pi, p, \bar{r}) \ : \ \|p - \bar{p}\| \le \psi\}$$

Robustness to Reward Errors

Objective:

$$\max_{\pi} \min_{r} \{\rho(\pi, \bar{p}, r) \;:\; \|r - \bar{r}\| \le \psi\}$$

Linear program reformulation ($\|\cdot\|_\star$ is dual norm):

$$\begin{aligned} \max_{u \in \mathbb{R}^{SA}} \quad & \bar{r}^\top u - \psi \|u\|_\star \\ \text{s. t.} \quad & \sum_a (I - \gamma P_a^\top) u_a = p_0 \\ & u \ge 0 \end{aligned}$$

Robustness to Transition Errors

Objective:

$$\max_{\pi} \min_{p} \{\rho(\pi, p, \bar{r}) \ : \ \|p - \bar{p}\| \le \psi\}$$

Ambiguity set (aka uncertainty set):

$$\mathcal{P} \ = \ \{p \ : \ \|p - \bar{p}\| \le \psi\}$$

- **NP-hard** to solve
- No value function, or dynamic program

Dynamic Program for Rectangular Robust RL

S-rectangular: $\mathcal{P}$ constrained for each state separately

$$\max_{\pi} \min_{p} \{\rho(\pi, p, \bar{r}) \ : \ \|p_s - \bar{p}_s\| \le \psi_s, \ \forall s\}$$

Nature sees last state

SA-rectangular: $\mathcal{P}$ constrained for each state and action separately

$$\max_{\pi} \min_{p} \{\rho(\pi, p, \bar{r}) \ : \ \|p_{s,a} - \bar{p}_{s,a}\| \le \psi_{s,a}, \ \forall s, a\}$$

Nature sees last state and action

Optimal Robust Value Function

Bellman operator in MDPs:

$$v(s) = \max_a \left(r_{s,a} + \gamma \cdot \bar{p}_{s,a}^\top v \right)$$

Robust Bellman operator: SA-rectangular ambiguity set

$$v(s) \; = \; \max_a \min_{p \in \Delta_S} \left\{ r_{s,a} + \gamma \cdot p^\top v \; : \; \|p - \bar{p}_{s,a}\| \leq \psi_{s,a} \right\}$$

Robust Bellman operator: S-rectangular ambiguity set

$$v(s) = \max_{d \in \Delta_A} \min_{p_a \in \Delta_S} \left\{ \sum_a d(s,a)(r_{s,a} + \gamma \cdot p_a{}^\top v) \; : \; \sum_a \|p_a - \bar{p}_{s,a}\| \leq \psi_s \right\}$$

Solving Robust MDPs

Robust Bellman operator is:

1. A contraction in L_∞ norm
2. Monotone elementwise

Algorithms

1. Value iteration works but slow
2. Naive policy iteration may loop forever
3. Approximate convex optimization formulation

Grand-Clément, Julien, and Marek Petrik. Towards Convex Optimization Formulations for Robust MDPs, 2022.

Solving Robust MDPs

Robust Bellman Optimality: SA-rectangular ambiguity set

$$v(s) = \max_{a} \min_{p \in \Delta_S} \left\{ r_{s,a} + p^\top v \ : \ \|\bar{p} - p\|_1 \leq \psi \right\}$$

How to solve for p?

Linear programming is **polynomial time** for polyhedral sets

Is it really **tractable**?

Benchmarking Robust Bellman Update

Bellman update: Inventory optimization, 200 states and actions

$$r_{s,a} + p^\top v$$

Time: 0.04s

Robust Bellman update: Gurobi LP

$$\min_{p \in \Delta_S} \left\{ r_{s,a} + p^\top v \;:\; \|\bar{p} - p\|_1 \le \psi \right\}$$

Rectangularity	**Time**
SA-	1.1 min
S	16.7 min

Fast Robust RL Algorithms

Homotopy algorithm + PPI:

Rectangularity	Time
SA-	1.1 min / 0.6s
S-	16.7 min / 0.7s

- Ho, Chin Pang, Marek Petrik, and Wolfram Wiesemann. Robust Phi-Divergence MDPs, Neurips, 2023
- Derman, Esther, Matthieu Geist, and Shie Mannor. Twice Regularized MDPs and the Equivalence between Robustness and Regularization, Neurips, 2021
- Grand-Clément, Julien. From Convex Optimization to MDPs: A Review of First-Order, Second-Order and Quasi-Newton Methods for MDPs, 2021

Other Robust RL Results

- Other notions of rectangularity

  Goyal, V. and Grand-Clement, J., Robust Markov decision process: Beyond rectangularity, Mathematics of Operations Research, 2022.

- Model free algorithms

  Panaganti, K. et al., Robust reinforcement learning using offline data, NIPS, 2022).

- Robust policy gradient

  Qiuhao Wang, Chin Pang Ho, Marek Petrik, Policy Gradient in Robust MDPs with Global Convergence Guarantee, ICML, 2023.

- Average reward criteria

  Wang, Y. et al., Robust Average-Reward Markov Decision Processes, AAAI, 2023.

- ...

Summary

Risk and Robustness in RL

- Monetary risk measures: VaR, CVaR, EVaR, ERM

- Risk-aversion
  1. Aleatory: risk-averse RL
  2. Epistemic: (soft-)robust RL

- Formulating a dynamic program
  1. Make assumptions on the risk: nested risk measures, ERM, rectangular uncertainty
  2. Reduce to a simpler risk measure: EVaR to ERM
  3. Augment state space: VaR, CVaR

Research Questions

1. Scalable risk-averse RL with guarantees
2. Distributional RL for risk-aversion
3. Relaxing rectangularity in robust RL
4. Unifying risk-averse and robust RL

Thank You

Thanks to my collaborators: Bahram Behzadian, Erick Delage, Mohammad Ghavamzadeh, Julien Grand-Clement, Jia Lin Hau, Chin Pang Ho, Reazul Russel, Xihong Su, Wiuhao Wang, Wolfram Wiesemann

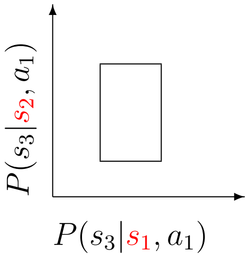
Appendix

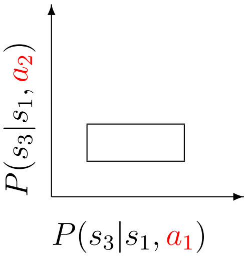
SA-Rectangular Ambiguity

Example: For each state s and action a:

$$\left\{ p_{s,a} \; : \; \|p_{s,a} - \bar{p}_{s,a}\|_1 \le \psi_{s,a} \right\} = \left\{ p_{s,a} \; : \; \sum_{s'} |p_{s,a,s'} - \bar{p}_{s,a,s'}| \le \psi_{s,a} \right\}$$

Sets are rectangles over s and a:





S-Rectangular Ambiguity

Example: For each state s:

$$\left\{ p_{s,a} \;:\; \sum_{a} \|p_{s,a} - \bar{p}_{s,a}\|_1 \le \psi_s \right\} = \left\{ p_{s,a} \;:\; \sum_{a,s'} |p_{s,a,s'} - \bar{p}_{s,a,s'}| \le \psi_s \right\}$$

Sets are rectangles over s only: