University of New Hampshire

LEARNING TO ACT WITH ROBUSTNESS

Reazul Hasan Russel

University of New Hampshire

Committee members:
Momotaz Begum
Mouhacine Benosman
Ernst Linder
Marek Petrik (Advisor)
Wheeler Ruml

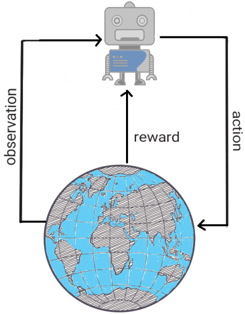
Outline

Reinforcement Learning

- Goal: select actions to maximize total future rewards [29].

Properties:

- No supervisor or labeled data
- Feedback is delayed, not instant
- Subsequent data depends on agent's action



Sequential Decision Making

University of New Hampshire

Markov Decision Process (MDP)

Definition

A Markov Decision Process is a tuple $\langle \mathcal{S}, \mathcal{A}, p, r \rangle$

- A finite set of states $\mathcal{S}$
- A finite set of actions $\mathcal{A}$
- A transition model $p(s'|s, a)$
- A reward function $r(s, a)$

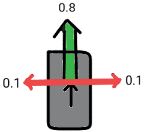
Markov Decision Process (MDP)

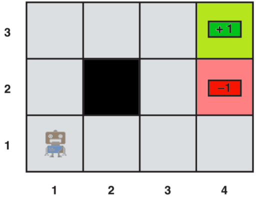
Definition

A Markov Decision Process is a tuple $\langle \mathcal{S}, \mathcal{A}, p, r \rangle$

- A finite set of states $\mathcal{S}$
- A finite set of actions $\mathcal{A}$
- A transition model $p(s'|s,a)$
- A reward function $r(s,a)$

State: Each cell
Action: Up, Down, Left, Right

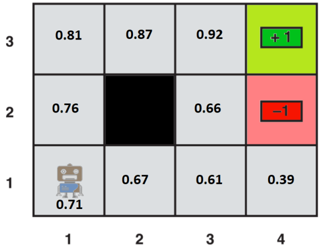




Objective: Maximize γ-discounted return by finding policy $\pi \in \Pi$ [25]:

$$\max_{\pi \in \Pi} \mathbb{E}_s^{\pi} \left[\sum_{t=0}^{\infty} \gamma^t \, r\Big(S_t, \pi(S_t)\Big) \right]$$

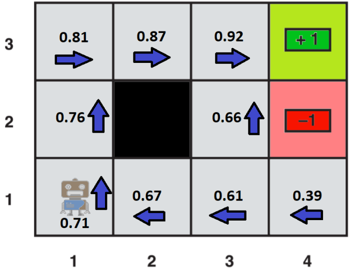
Value Function



Value function: v maps *states* $\rightarrow$ expected return

Return $= p_0^T v$, where p_0 initial state distribution

Optimal Solution



Policy: π maps *states* $\to$ *actions*

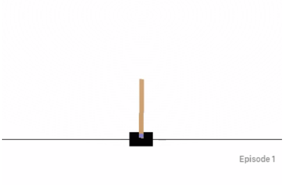
Optimal Solution: $\pi^\star \in \arg\max_\pi \ \text{return}(\pi)$

University of New Hampshire

University of New Hampshire

Applications of RL

Simulated Problems



Cartpole

Atari: Breakout

Cartpole: A classic control problem [5]

- Deterministic dynamics
- Fast and precise simulators
- Failure is cheap and recoverable
- No serious safety constraint

University of New Hampshire

Practical Problems

Agriculture

Precision Medicine

Agriculture: A challenging RL problem

- Stochastic environment, depends on many factors
- No simulator, must learn from historical data
- Delayed reward, one episode = one year
- Crop failure is expensive
- Needs to satisfy safety constraints

My Approach

- *Batch learning* setup because *no* reliable simulator available.

  Logged dataset $\mathcal{D} = (s_0, a_0, r_0, \ldots, s_{t-1}, a_{t-1}, r_{t-1})$

My Approach

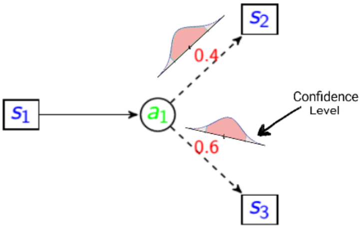
- *Batch learning* setup because *no* reliable simulator available.

$$\text{Logged dataset } \mathcal{D} = (s_0, a_0, r_0, \ldots, s_{t-1}, a_{t-1}, r_{t-1})$$

- How to compute solution and how to evaluate?

1. Learn *plausible models* consistent with $\mathcal{D}$
2. Compute *robust* solution

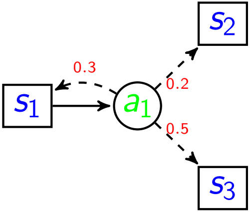
$$\max_{\text{policy}} \min_{\textit{model}} \textit{return}(\textit{policy}, \textit{model})$$



A Toy Example

A small MDP with:

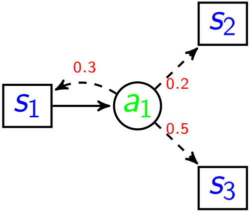
- States $S = \{s_1, s_2, s_3\}$
- Action $A = \{a_1\}$
- Transitions labeled on edges

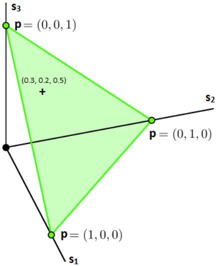


A Toy Example

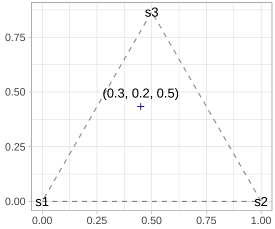
A small MDP with:

- States $S = \{s_1, s_2, s_3\}$
- Action $A = \{a_1\}$
- Transitions labeled on edges





Transition $p(\cdot|s_1, a_1)$



Transition $p(\cdot|s_1, a_1)$ projected onto simplex

Robust MDPs

Definition

A robust Markov Decision Process is a tuple $\langle \mathcal{S}, \mathcal{A}, p, r \rangle$

- A finite set of states $\mathcal{S}$
- A finite set of actions $\mathcal{A}$
- Transition $p(s'|s, a) \sim \mathcal{P}_{s,a}$
- A reward function $r(s, a)$

- **Ambiguity Set:** $\mathcal{P} = \|\bar{p}_{s,a} - p\|_1 \leq \psi_{s,a}$
- **Objective:** Maximize γ-discounted worst-case return [32]:

$$\max_{\pi \in \Pi} \min_{p \in \mathcal{P}} \text{return}(\pi, p)$$

State of The Art in RMDPs

RMDPs:

- *Robust* formulation of discrete dynamic programming.
- Solve RMDPs tractably using VI, PI [Iyengar [18], Nilim et al. [23]].

State of The Art in RMDPs

RMDPs:

- *Robust* formulation of discrete dynamic programming.
- Solve RMDPs tractably using VI, PI [Iyengar [18], Nilim et al. [23]].

Ambiguity Set Construction:

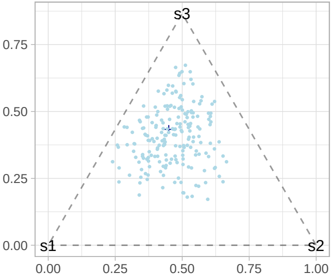
- KL-divergence with MLE or MAP [Nilim and El Ghaoui, 2005 [23]]
  - Disadvantage: No guarantee
- Second order approx. without fixed set [Delage and Mannor, 2010 [9]]
  - Disadvantage: No guarantee
- Confidence region around MLE with prior [Wiesemann et. al. 2013 [32]]
  - Disadvantage: Not optimized, conservative results

University of New Hampshire

Ambiguity Set as Bayesian Credible Region

- Dirichlet prior: $\alpha = (1, 1, 1)$
- Dataset: $\mathcal{D} = s_1 \to a_1 \to [3 \times s_1, 2 \times s_2, 5 \times s_3]$
- Posterior: $\alpha = (4, 3, 6)$

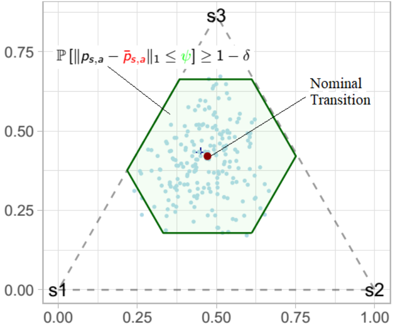
May use MCMC methods for posterior sampling



Samples from posterior

University of New Hampshire

Ambiguity Set as Bayesian Credible Region

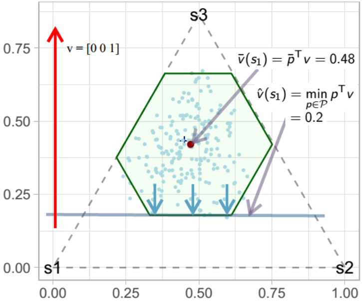


Bayesian Ambiguity set: find minimum ψ to cover $(1-\delta) * N$ samples around nominal point [26].

With $\delta = 0.1$ and $N = 200$, above ambiguity set covers at least $0.9 * 200 = 180$ points around nominal point.

Robust Solution with BCR

With ambiguity set $\mathcal{P}$ and value function being $v = [0, 0, 1]$



<u>Nominal Value</u>	**<u>Robust Value</u>**
$\bar{v}(s_1) = \bar{p}^\mathsf{T} v = 0.48$	$\hat{v}(s_1) = \min_{p \in \mathcal{P}} p^\mathsf{T} v = 0.2$
with NO guarantee	with 90% confidence level

University of New Hampshire

List of Contributions

1. **Weighted Set for RMDPs**: Optimize shape of ambiguity sets with weights *for better high confidence guarantees*.
2. **Near-optimal Set for RMDPs**: Construct near-optimal sets from possible value functions *for better high confidence guarantees*.
3. **Robust Constrained MDPs (RCMDPs)**: Propose robust constrained MDP, optimize *for the worst-case* constraint satisfaction.
4. **Risk-Averse Soft-Robust (RASR) Framework**: Develop risk-averse soft-robust framework to simultaneously *handle model and transition uncertainties*.

Weighted Set

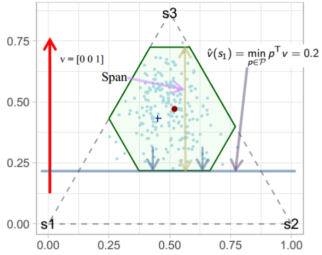
Weighted Ambiguity Sets for RMDPs

Weighted Set: Intuition

Motivation: Reshape by reducing span of the set.

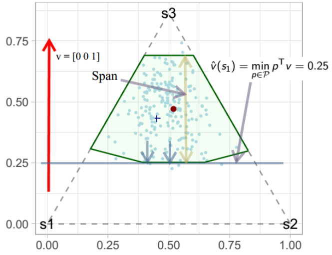
Weighted set: $\mathcal{P}_{s,a} = \left\{ \boldsymbol{p} \in \Delta^{\mathcal{S}} \ : \ \|\boldsymbol{p} - \bar{\boldsymbol{p}}_{s,a}\|_{1,w} \le \psi_{s,a} \right\}$

Unweighted Set



Guaranteed return 0.2

Weighted Set



Guaranteed return 0.25

Weighted Set: Approach

Steps to construct weighted set for $\lambda \in \mathbb{R}$ and $\boldsymbol{z} \in \mathbb{R}^S$:

1. **Maximize lower bound:**

$$\max_{w \in \mathbb{R}^S_{++}} \Big\{ \underbrace{\bar{p}^\mathsf{T} z - \psi \|z - \bar{\lambda}\mathbf{1}\|_{\infty, \frac{1}{w}}}_{\text{lower bound of robust value}} \ : \ \textstyle\sum_{i=1}^S w_i^2 = 1 \Big\}$$

2. **Optimize weights:** $w_i^\star \leftarrow \frac{|z_i - \bar{\lambda}|}{\sqrt{\sum_{j=1}^S |z_j - \bar{\lambda}|}}, \forall i \in \mathcal{S}$

3. **Optimize size:** Minimal ψ with BCR or Hoeffding [27]

Theorem (Weighted Hoeffding bound)

With weights $w \in \mathbb{R}^S_{++}$ sorted in a non-increasing order:

$$\mathbb{P}\left[\|\bar{p}_{s,a} - p^\star_{s,a}\|_{1,w} \geq \psi_{s,a} \right] \leq 2 \sum_{i=1}^{S-1} 2^{S-i} \exp\left(-\frac{\psi^2_{s,a} n_{s,a}}{2 w_i^2} \right)$$

Weighted Set: Evaluation Domains

- **RiverSwim (RS)**: simple and standard benchmark problem with six states and two actions [28].
- **Machine Replacement (MR)**: a small MDP problem modeling progressive deterioration of a mechanical device [9].
- **Population Growth Model (PG)**: an exponential population growth model [19] with 50 states.
- **Inventory Management (IM)**: a classic inventory management problem [34] with discrete inventory levels.
- **Cart-Pole (CP)**: standard RL benchmark problem to balance a pole [6].

University of New Hampshire

Weighted Set: Empirical Evaluation

Normalized Frequentist performance loss

	RS	MR	PG	IM	CP
Standard	0.8	5.83	5.66	1.05	0.78
Optimized	**0.53**	**1.05**	**5.55**	**0.99**	**0.77**

Normalized Bayesian performance loss

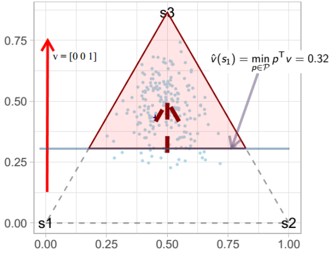
	RS	MR	PG	IM	CP
Standard	0.6	1.56	5.24	0.97	0.77
Optimized	**0.25**	**0.41**	**1.84**	**0.90**	**0.12**

Loss is computed w.r.t. nominal model. confidence level is 95%. *Lower loss is better.*

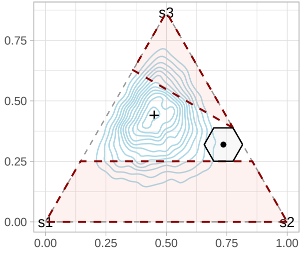
Near-optimal Bayesian Ambiguity Sets for RMDPs

Near-optimal Bayesian Set: Intuition

Motivation: Half space defined by value function good enough.



Optimal set



Near-optimal set

Near-optimal set constructed for two possible value functions:
$v_1 = (0, 0, 1)$ and $v_2 = (2, 1, 0)$.

Approach: Find smallest set intersecting all half-spaces corresponding to each value function.

Near-optimal Set: Approach

1. **Optimal set** for a given v and $\zeta = 1 - \delta/(SA)$:
   $\mathcal{K}_{s,a}(v) = \left\{ p \in \Delta^S \;:\; p^\mathsf{T} v \leq \mathrm{V@R}^{\zeta}_{P^\star}\left[(p^\star_{s,a})^\mathsf{T} v \right] \right\}$
2. **Near-optimal set:** with set $\mathcal{V}$
   $\mathcal{L}_{s,a}(\mathcal{V}) = \{ p \in \Delta^S \;:\; \| p - \theta_{s,a}(\mathcal{V}) \|_1 \leq \psi_{s,a}(\mathcal{V}) \}$
   $\psi_{s,a}(\mathcal{V}) = \min_{p \in \Delta^S} f(p), \qquad \theta_{s,a}(\mathcal{V}) \in \arg\min_{p \in \Delta^S} f(p)$
   $$f(p) = \max_{v \in \mathcal{V}} \min_{q \in \mathcal{K}_{s,a}(v)} \| q - p \|_1$$
3. iteratively expand $\mathcal{V}$ and approximate $\mathcal{L}$.

Theorem (Safe return estimates)

Policy $\hat{\pi}_k$ and value function $\hat{v}_k$ computed by near-optimal set in iteration k. The return estimate $\tilde{\rho}(\hat{\pi}) = p_0^\mathsf{T} \hat{v}_k$ is safe:

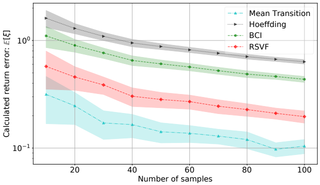
$$\mathbb{P}_{P^\star}\left[p_0^\mathsf{T} \hat{v}_k \leq p_0^\mathsf{T} v^{\hat{\pi}_k}_{P^\star} \;\middle|\; \mathcal{D} \right] \geq 1 - \delta.$$

University of New Hampshire

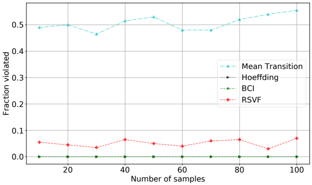
Near-optimal Set: Empirical Evaluation

Single-state Bellman update with uninformative Dirichlet prior.

Regret w.r.t optimal policy



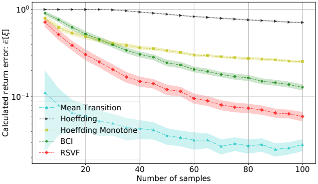
Violation rate



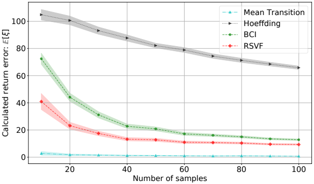
Regret w.r.t optimal policy. Estimates are computed with 95% confidence level. *Lower regret is better.*

Near-optimal Set: Empirical Evaluation

Inventory management



Population model



Regret w.r.t optimal policy. Estimates are computed with 95% confidence level. *Lower regret is better.*

RCMDP

Robust Constrained Markov Decision Processes

Constrained MDPs

Definition

Defined as a tuple $\langle \mathcal{S}, \mathcal{A}, p, \{r_0, r_1, \dots r_n\}, \{\beta_1, \dots, \beta_n\} \rangle$

- Same $\mathcal{S}$, $\mathcal{A}$ and fixed transition kernel P like MDPs
- Contains multiple reward functions $\{r_0, r_1, \dots r_n\}$ and budgets $\{\beta_1, \dots, \beta_n\}$

- **Objective:** Maximize γ-discounted return satisfying constraints [2]:

$$\max_{\pi \in \Pi} \mathbb{E}_s^{\pi} \left[\sum_{t=0}^{\infty} \gamma^t \, r_0 \Big(S_t, A_t \Big) \right]$$

$$\text{s.t. } \mathbb{E}_s^{\pi} \left[\sum_{t=0}^{\infty} \gamma^t \, r_i \Big(S_t, A_t \Big) \right] \geq \beta_i, \text{ for } i = 1, \dots, n$$

State of the Art in CMDPs

Dates back to 1960s, first studied by *Derman and Klein* [11].

CMDP solution methods:

- Linear programming based solutions [11, 2],
- Lagrangian methods [16, 2]
- Surrogate based methods [1, 8],

State of the Art in CMDPs

Dates back to 1960s, first studied by *Derman and Klein* [11].

CMDP solution methods:

- Linear programming based solutions [11, 2],
- Lagrangian methods [16, 2]
- Surrogate based methods [1, 8],

Sensitivity and robustness in CMDPs:

- Sensitivity analysis for LPs with small perturbations (Altman and Shwartz [3]),
- Robustness under small change in constraints (Alex and Shwartz [33]),
- Handling model misspecification in CMDPs (Mankowitz et al. [21])

Robust Constrained MDPs

RMDP

$$\max_{\pi\in\Pi} \min_{p\in\mathcal{P}} \mathbb{E}^{\pi}_{s\sim p}\left[\sum_{t=0}^{\infty} \gamma^t\, r_0(S_t, A_t)\right]$$

CMDP

$$\max_{\pi\in\Pi} \mathbb{E}^{\pi}_{s}\left[\sum_{t=0}^{\infty} \gamma^t\, r_0(S_t, A_t)\right]$$

$$\textbf{s.t.}\;\; \mathbb{E}^{\pi}_{s}\left[\sum_{t=0}^{\infty} \gamma^t\, r_1(S_t, A_t)\right] \geq \beta_1$$

RCMDP

$$\max_{\pi\in\Pi} \min_{p\in\mathcal{P}} \mathbb{E}^{\pi}_{s}\left[\sum_{t=0}^{\infty} \gamma^t\, r_0(S_t, A_t)\right]$$

$$\textbf{s.t.}\;\; \min_{p\in\mathcal{P}} \mathbb{E}^{\pi}_{s}\left[\sum_{t=0}^{\infty} \gamma^t\, r_1(S_t, A_t)\right] \geq \beta_1$$

RCMDP incorporates both constraints and robustness in objective

RCMDP: Approach

- **Lagrange reformulation of RCMDP objective**:
  $\mathfrak{L}(\pi_\theta, \lambda) = \sum_{\xi \in \Xi} p^{\pi_\theta}(\xi)\Big(g(\xi, r) + \lambda g(\xi, d)\Big) - \lambda\beta$
- **Find a saddle point** $(\pi_\theta{}^*, \lambda^*)$ **of** $\mathfrak{L}$ **that satisfies**:
  $\mathfrak{L}(\pi_\theta, \lambda^*) \leq \mathfrak{L}(\pi_\theta{}^*, \lambda^*) \leq \mathfrak{L}(\pi_\theta{}^*, \lambda),\ \forall\theta \in \mathbb{R}^k,\ \forall\lambda \in \mathbb{R}_+$
- Use the gradients of $\mathfrak{L}$ to optimize the RCMDP objective [7]

Theorem (Gradient update formula)

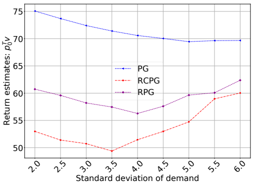
Gradients of $\mathfrak{L}$ with respect to θ and λ are:

$$\nabla_\theta \mathfrak{L}(\pi_\theta, \lambda) = \sum_\xi \hat{p}^{\pi_\theta}(\xi)\Big(g(\xi, r) + \lambda g(\xi, d)\Big) \sum_{t=0}^{T-1} \frac{\nabla_\theta \pi_\theta(a_t|s_t)}{\pi_\theta(a_t|s_t)}$$

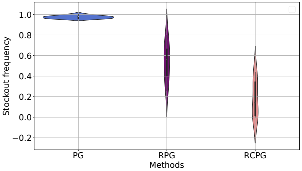
$$\nabla_\lambda \mathfrak{L}(\pi_\theta, \lambda) = \sum_\xi \hat{p}^{\pi_\theta}(\xi) g(\xi, d) - \beta$$

RCMDP: Empirical Evaluation

- Evaluating policy-gradient method on inventory management.



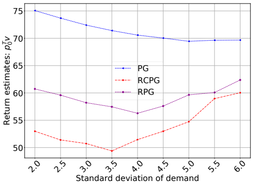
Return estimates with perturbed demand



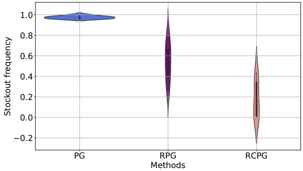
Stock-out frequency

RCMDP: Empirical Evaluation

- Evaluating policy-gradient method on inventory management.



Return estimates with perturbed demand



Stock-out frequency

- Evaluating actor-critic method on cart-pole.

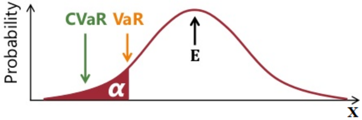
Methods	Expected Return	Constraint Violation
AC	175.45 ± 2.99	2.3%
RAC	118.22 ± 6.07	1.1%
RCAC	123.26 ± 8.64	0.05%

RASR

Risk-Averse Soft-Robust Framework

Risk Measures

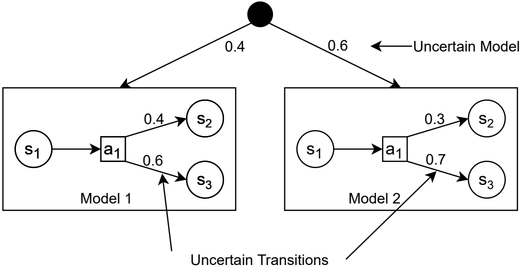
- Risk: a loss, chance of occuring that loss and the significance of that loss to the person concerned.



- **VaR**$^{\alpha}$**(X)**: $\alpha-$percentile of X.
- **CVaR**$^{\alpha}$**(X)**: Expectation of worst $\alpha-$fraction of X.
- **Entropic**$^{\alpha}$**(X)**: $-\frac{1}{\alpha}\log\left(\mathbb{E}\left[\exp(-\alpha X)\right]\right)$

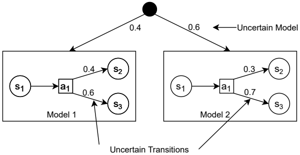
Risk-Averse (RA) and Soft-Robust (SR)

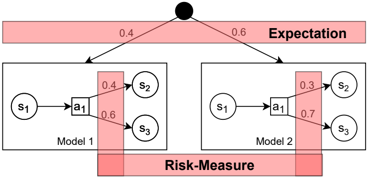
A problem with two possible models



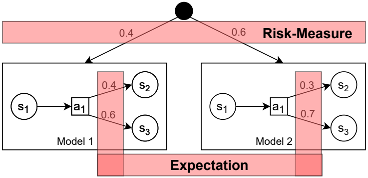
Risk-Averse (RA) and Soft-Robust (SR)

A problem with two possible models



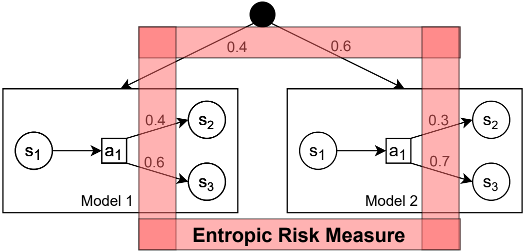


Risk Averse (RA)



Soft Robust (SR)

Risk-Averse Soft-Robust (RASR) Framework



Apply ERM on both model and transition uncertainties

- In RASR, both model parameters $\hat{P}_t$ and transitions to S_{t+1} are dynamically uncertain for each time step t.

$$\psi(\pi,f)=\rho^{\alpha}_{\hat{P},S,A}\left[\sum_{t=0}^{T}\gamma^t\cdot r(S_t,A_t,S_{t+1})\ :\ S_0\sim p_0,S_{t+1}\sim\hat{P}_t(s_t,a_t),A_t\sim\pi(S_t),\hat{P}_t\sim f\right]\ .$$

RASR: Approach

- **Value Iteration:** RASR Bellman equation.
  $\hat{v}(s) \leftarrow \max_{a\in\mathcal{A}} \rho^{\alpha}_{P^\omega \sim \hat{P}, s' \sim P^\omega(\cdot|s,a)} \left[r_{s,a,s'} + \gamma \hat{v}(s') \right]$

- **Actor-Critic:** Parameterize policy and optimize with gradients.
  $J(\pi_\theta) = -\frac{1}{\alpha} \log \left(\mathbb{E}_{\tau \sim p_\theta(\tau)} \left[\exp\left(-\alpha R(\tau) \right) \right] \right)$

Theorem (RASR gradient formula)

Gradient of $J(\pi_\theta)$ with respect to the parameter θ is:

$$\nabla_\theta J(\pi_\theta) = \frac{-\sum_\tau p_\theta(\tau) \sum_{t=0}^{T} \frac{\nabla_\theta \pi_\theta(a_t|s_t)}{\pi_\theta(a_t|s_t)} \cdot \exp\left(-\alpha \sum_{t=0}^{T} r_{s_t,a_t} \right)}{\alpha \sum_\tau p_\theta(\tau) \exp\left(-\alpha R(\tau) \right)}$$

RASR: Empirical Evaluation

Evaluation of RASR-VI policies

	RS	MR	IM
Nominal	16.54	−128.17	60.12
BCR	46.15	−127.53	74.40
RSVF	1.59	−129.03	65.44
RASR-CVaR	43.56	−127.83	69.09
RASR-Entropic	**49.99**	**−120.89**	**83.50**

Evaluation of RASR-AC policies on Cart-Pole problem

General	Soft-Robust	RASR-CVaR	RASR-Entropic
112.11	102.49	127.82	**143.6**

Return estimates under RASR entropic metric

Conclusion

- Introduced basic RL framework and presented concepts regarding robust and risk-averse decision making.
- Presented four novel contributions in robust and risk-averse RL:
  1. Developed methods to construct weighted ambiguity sets for RMDPs.
  2. Developed methods to construct near-optimal Bayesian ambiguity sets for RMDPs.
  3. Developed robust constrained MDP framework and derived methods for policy optimization in RCMDPs
  4. Developed RASR framework and derived methods for policy optimization in RASR setting

Publications

Conferences:

1. *Optimizing Percentile Criterion using Robust MDPs.* Bahram Behzadian, *Reazul Hasan Russel*, Marek Petrik, Chin Pang Ho. Published at AISTATS 2021.
2. *Beyond Confidence Interval: Tight Bayesian Ambiguity Sets for Robust MDPs.* *Reazul Hasan Russel*, Marek Petrik. Published at NeurIPS 2019.
3. *Value Directed Exploration in Multi-Armed Bandits with Structured Priors.* Bence Cserna, Marek Petrik, *Reazul Hasan Russel*, Wheeler Ruml. Published at UAI 2017.
4. Robust Constrained MDP and Stability. Reazul Hasan Russel, Mouhacine Benosman, Jeroen Van Baar, Radu Corcodel. Under review at NeurIPS 2021
5. Risk-Averse Soft-Robust Reinforcement Learning. In preparation.

Publications

Workshops:

1. *Optimizing Norm-bounded Weighted Ambiguity Sets for Robust MDPs. Reazul Hasan Russel**, Bahram Behzadian*, Marek Petrik. Presented at NeurIPS 2019 workshop on SRDM.
2. *Tight Bayesian Ambiguity Sets for Robust MDPs. Reazul Hasan Russel*, Marek Petrik. Presented at NeurIPS Workshop on Probabilistic Reinforcement Learning and Structured Control, 2018.
3. Robust Exploration with Tight Bayesian Plausibility Sets. Reazul H Russel, Tianyi Gu, Marek Petrik. RLDM 2018.
4. Robust Constrained-MDPs: Soft-Constrained Robust Policy Optimization under Model Uncertainty. Reazul Hasan Russel, Mouhacine Benosman, Jeroen Van Baar. NeurIPS workshop on The Challenges of Real World Reinforcement Learning 2020

Thank you!

Bibliography I

[1] J. Achiam, D. Held, A. Tamar, and P. Abbeel. Constrained Policy Optimization. *International Conference on Machine Learning*, 2017.

[2] E. Altman. Constrained Markov Decision Processes. 2004.

[3] E. Altman and A. Shwartz. Sensitivity of constrained Markov decision processes. *Annals of Operations Research*, 1991.

[4] P. Auer, T. Jaksch, and R. Ortner. Near-optimal regret bounds for reinforcement learning. *Journal of Machine Learning Research*, 2010.

[5] G. Brockman, V. Cheung, L. Pettersson, J. Schneider, J. Schulman, J. Tang, and W. Zaremba. Openai gym, 2016.

[6] G. Brockman, V. Cheung, L. Pettersson, J. Schneider, J. Schulman, J. Tang, and W. Zaremba. OpenAI Gym. Technical report, 2016.

Bibliography II

[7] Y. Chow and M. Ghavamzadeh. Algorithms for CVaR optimization in MDPs. *Advances in Neural Information Processing Systems*, 2014.

[8] G. Dalal, K. Dvijotham, M. Vecerik, T. Hester, C. Paduraru, and Y. Tassa. Safe exploration in continuous action spaces, 2018.

[9] E. Delage and S. Mannor. Percentile Optimization for Markov Decision Processes with Parameter Uncertainty. *Operations Research*, 2010.

[10] E. Derman, D. J. Mankowitz, T. A. Mann, and S. Mannor. Soft-robust actor-critic policy-gradient. *Conference on Uncertainty in Artificial Intelligence (UAI)*, 2018.

[11] M. Derman, Cyrus and Klein. Some Remarks on Finite Horizon Markovian Decision Models. 1965.

University of New Hampshire

Bibliography III

[12] R. Dimitrova, J. Fu, and U. Topcu. Robust optimal policies for Markov decision processes with safety-threshold constraints. *2016 IEEE 55th Conference on Decision and Control, CDC 2016*, 2016.

[13] H. Eriksson and D. Christos. Epistemic Risk-Sensitive Reinforcement Learning. 2019.

[14] H. Eriksson and C. Dimitrakakis. Epistemic risk-sensitive reinforcement learning. *European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning*, 2020.

[15] Y. Fei, Z. Yang, Y. Chen, Z. Wang, and Q. Xie. Risk-sensitive reinforcement learning: Near-optimal risk-sample tradeoff in regret. *arXiv*, 2020.

[16] P. Geibel and F. Wysotzki. Risk-sensitive reinforcement learning applied to control under constraints. *Journal of Artificial Intelligence Research*, 2005.

Bibliography IV

[17] T. Hiraoka, T. Imagawa, T. Mori, T. Onishi, and Y. Tsuruoka. Learning Robust Options by Conditional Value at Risk Optimization. *Neural Information Processing Systems*, 2019.

[18] G. N. Iyengar. Robust dynamic programming. *Mathematics of Operations Research*, 2005.

[19] M. Kery and M. Schaub. *Bayesian Population Analysis Using WinBUGS*. 2012.

[20] E. A. Lobo, M. Ghavamzadeh, and M. Petrik. Soft-Robust Algorithms for Batch Reinforcement Learning. *Arxiv*, 2021.

[21] D. J. Mankowitz, N. Levine, R. Jeong, Y. Shi, J. Kay, A. Abdolmaleki, J. T. Springenberg, T. Mann, T. Hester, and M. Riedmiller. Robust Constrained Rinforcement Learning For Continuous Control With Model Misspecification. 2020.

[22] D. Nass, B. Belousov, and J. Peters. Entropic Risk Measure in Policy Search. *Investment Management and Financial Innovations*, 2020.

University of New Hampshire

Bibliography V

[23] A. Nilim and L. El Ghaoui. Robust control of Markov decision processes with uncertain transition matrices. *Operations Research*, 2005.

[24] L. Prashanth and M. Ghavamzadeh. Variance-constrained Actor-Critic Algorithms for Discounted and Average Reward MDPs. *Machine Learning Journal*, 2016.

[25] M. L. Puterman. *Markov decision processes: Discrete stochastic dynamic programming*. John Wiley & Sons, Inc., 2005.

[26] R. H. Russel and M. Petrik. Beyond confidence regions: Tight Bayesian ambiguity sets for robust MDPs. *Advances in Neural Information Processing Systems*, 2019.

[27] R. H. Russel and M. Petrik. Beyond Confidence Regions: Tight Bayesian Ambiguity Sets for Robust MDPs. *Advances in Neural Information Processing Systems*, 2019.

University of New Hampshire

Bibliography VI

[28] A. L. Strehl and M. L. Littman. An analysis of model-based interval estimation for Markov decision processes. *Journal of Computer and System Sciences*, 74(8):1309–1331, 2008.

[29] R. S. Sutton and A. G. Barto. *Reinforcement learning: An introduction*. MIT press, 2018.

[30] A. Tamar, D. D. Castro, and S. Mannor. Temporal Difference Methods for the Variance of the Reward To Go. *International Conference on Machine Learning*, 2013.

[31] T. Weissman, E. Ordentlich, G. Seroussi, S. Verdu, and M. J. Weinberger. Inequalities for the L_1 deviation of the empirical distribution. 2003.

[32] W. Wiesemann, D. Kuhn, and B. Rustem. Robust Markov decision processes. *Mathematics of Operations Research*, 2013.

[33] A. Zadorojniy and A. Shwartz. Robustness of policies in constrained Markov decision processes. *IEEE Transactions on Automatic Control*, 2006.

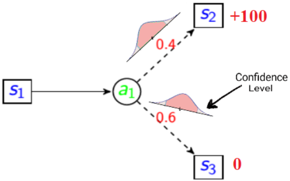
[34] P. H. Zipkin. *Foundations of Inventory Management*. 2000.

Supplementary Materials

Summary of the work to be done

- Soft-robust with entropic risk:
  - Theoretical understanding: time consistency of entropic risk measure for our formulation. ✓
  - More empirical evaluation: run more experiments on bigger and complex domain. ✓
- Robust constrained MDP:
  - Exploring and understanding new ideas for further contribution ✓
  - Theoretical understanding and empirical evaluation. ✓

Robustness: Policy Evaluation



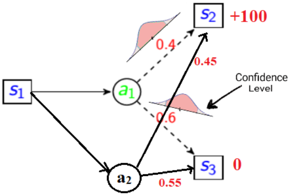
- True expected return: $0.4 * 100 + 0.6 * 0 = 40$
- $\mathcal{D} = s_1 \to a_1 \to [5 \times s_2, 5 \times s_3]$
- Nominal transition: $[0.5, 0.5]$.

- Non-robust return: $0.5 * 100 + 0.5 * 0 = 50$
- Ambiguity budget: $\psi = 0.4$
- Worst-case transition: $0.3, 0.7$.
- Robust return: $0.3 * 100 + 0.7 * 0 = 30$.

Non-robust evaluation: promises $50, but delivers $40.
Robust evaluation: promises at least $30, and delivers $40.

Robustness: Policy Evaluation

- True expected return: $a_1 = 40$, $a_2 = 45$
- $\mathcal{D} = \{s_1 \rightarrow a_1 \rightarrow [5 \times s_2, 5 \times s_3], s_1 \rightarrow a_2 \rightarrow [45 \times s_2, 55 \times s_3]\}$



a_1	a_2	
Nominal: $[0.5, 0.5]$	Nominal: $[0.45, 0.55]$	
Return: 50	Return: 45	Decision: a_1
Robust Return: 40	Robust Return: 45	Decision: a_2

Robustness makes it possible to pick the best action a_2

RASR: State of the Art in Risk and RL

References	Uncertainty Types		Risk Measures		
	RA	SR	Variance	CVaR	Entropic
Lobo et al. [20]	✗	✓	✗	✓	✗
Nass et al. [22]	✓	✗	✗	✗	✓
Fei et al. [15]	✓	✗	✗	✗	✓
Eriksson et al. [14]	✗	✓	✗	✓	✓
Hiraoka et al.[17]	✗	✓	✗	✓	✗
Prashanth et al. [24]	✓	✗	✓	✗	✗
Chow et al. [7]	✓	✗	✗	✓	✗
Tamar et al.[30]	✓	✗	✗	✓	✗
RASR	✓	✓	✗	✗	✓

RASR: Empirical Evaluation

Methods		RS	MR	IM
Nominal	Mean	221.90	−12.46	226.47
	RASR	16.54	−128.17	60.12
BCR	Mean	107.77	−15.68	208.73
	RASR	46.15	−127.53	74.40
RSVF	Mean	220.81	−14.14	216.54
	RASR	1.59	−129.03	65.44
RASR-CVaR	Mean	132.92	−14.08	216.52
	RASR	43.56	−127.83	69.09
RASR-Entropic	Mean	49.99	−24.11	118.54
	RASR	49.99	−120.89	83.50

Pest Control as MDP

States: Pest population: $[0, 50]$

Actions:

- 0 No pesticide
- 1-4 Pesticides P1, P2, P3, P4 with increasing effectiveness

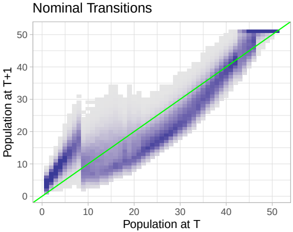
Transition probabilities: Pest population dynamics

Reward:

1. Crop yield minus pest damage
2. Spraying cost: P4 more expensive than P1

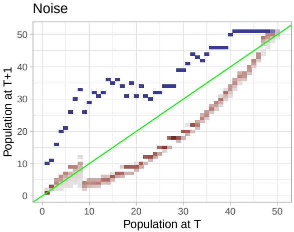
Non-robust Solution

Return: $\$8,820$



$+$

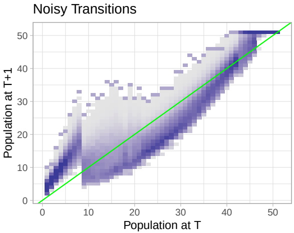
$L_1 \leq 0.05$



$=$

$=$

Return: $-\$6,725$

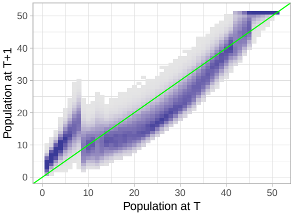


University of New Hampshire

Robust Solution

Return: $\$7,125$

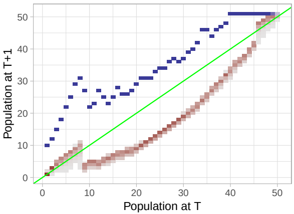
Nominal Transitions



$+$

$L_1 \leq 0.05$

Noise

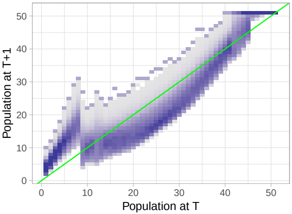


$=$

$=$

Return: $-\$27$

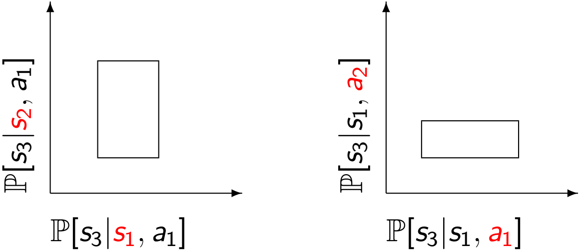
Noisy Transitions



SA-Rectangular Ambiguity

Nature is constrained for each state and action separately e.g. [23]

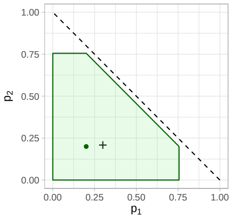
Sets are rectangles over s and a:



Frequentist Ambiguity Set

For $\bar{p}_{s,a} = \mathbb{E}_{P^\star}[p^\star_{s,a} \mid \mathcal{D}]$ with prob. $1-\delta$ (using Hoeffding's Ineq. see e.g. [31, 4, 26]):

$$\mathcal{P}^H_{s,a} = \left\{ p \in \Delta^S : \|p - \bar{p}_{s,a}\|_1 \le \sqrt{\frac{2}{n_{s,a}} \log \frac{SA2^S}{\delta}} \right\}$$

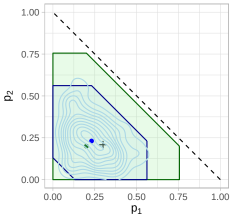


Few samples ⟶ large ambiguity set ⟶ Very conservative

Bayesian Ambiguity Set

Use posterior distribution to optimize for the *smallest* ambiguity set.

$$\mathcal{P}^B_{s,a} = \left\{ p \in \Delta^S \;:\; \|p - \bar{p}_{s,a}\|_1 \leq \psi^B_{s,a} \right\}, \quad \bar{p}_{s,a} = \mathbb{E}_{P^\star}[p^\star_{s,a} \mid \mathcal{D}] \,.$$



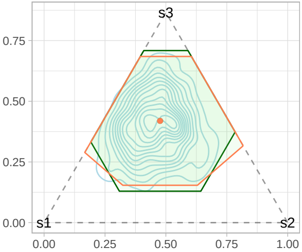
Hoeffding (green) vs Bayesian(blue), uniform Dirichlet Prior, 3 states

Tighter than frequentist but require prior and omputationally demanding

Idea 1: Weighted Ambiguity Sets

Optimize ambiguity sets with problem specific weights.

$v = (0, 0, 1)$



Green: L_1-norm bounded set:

$$\mathcal{P}_{s,a} = \left\{ \boldsymbol{p} \in \Delta^{\mathcal{S}} \; : \; \|\boldsymbol{p} - \bar{\boldsymbol{p}}_{s,a}\|_1 \leq \psi_{s,a} \right\}$$

Orange: Weighted L_1-norm bounded:

$$\mathcal{P}_{s,a} = \left\{ \boldsymbol{p} \in \Delta^{\mathcal{S}} \; : \; \|\boldsymbol{p} - \bar{\boldsymbol{p}}_{s,a}\|_{1,w} \leq \psi_{s,a} \right\}$$

Idea 1: Weighted Ambiguity Sets

Optimizing weights:

- **Step 1:** Estimate a value function $\hat{v}$
- **Step 2:** Lower bound the robust value:

$$\min_{\boldsymbol{p}\in\Delta^{\mathcal{S}}} \left\{ r_{s,a} + \gamma\, \boldsymbol{p}^{\mathsf{T}} \hat{\boldsymbol{v}} : \|\boldsymbol{p} - \bar{\boldsymbol{p}}\|_{1,\boldsymbol{w}} \le \psi \right\}$$

- **Step 3:** Compute weights w maximizing the lower bound:

$$\max_{\boldsymbol{w}\in\mathbb{R}^S_{++}} \left\{ \bar{\boldsymbol{p}}^{\mathsf{T}} \boldsymbol{z} - \psi \|\boldsymbol{z} - \bar{\lambda}\mathbf{1}\|_{\infty,\frac{1}{\boldsymbol{w}}} \; : \; \sum_{i=1}^{S} {w_i}^2 = 1 \right\}$$

- **Step 4:** Use w to compute weighted sets.

Idea 1: Optimizing Weights

- Define $\boldsymbol{z} = r_{s,a}\mathbf{1} + \gamma\,\hat{\boldsymbol{v}}$ and $q(\boldsymbol{z})$ with L_∞ norm for some $\boldsymbol{w} > 0$ as: $q(\boldsymbol{z}) = \min_{\boldsymbol{p}\in\Delta^S}\left\{\boldsymbol{p}^\mathsf{T}\boldsymbol{z} : \|\boldsymbol{p} - \bar{\boldsymbol{p}}\|_{1,\boldsymbol{w}} \le \psi\right\}$.

Theorem

$q(\boldsymbol{z})$ can be lower-bounded as follows:

$$q(\boldsymbol{z}) \ge \bar{\boldsymbol{p}}^\mathsf{T}\boldsymbol{z} - \psi\|\boldsymbol{z} - \lambda\mathbf{1}\|_{\infty,\frac{1}{\boldsymbol{w}}}$$

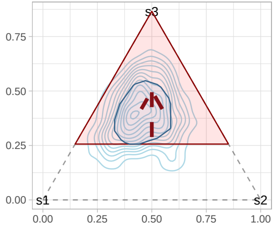
for any $\lambda \in \mathbb{R}$. Moreover, when $\boldsymbol{w} = \mathbf{1}$, the bound is tightest when $\lambda = (\max_i z_i + \min_i z_i)/2$ and the bound turns to $q(\boldsymbol{z}) \ge \bar{\boldsymbol{p}}^\mathsf{T}\boldsymbol{z} - \frac{\psi}{2}\|\boldsymbol{z}\|_s$ with $\|\cdot\|_s$ representing the *span semi-norm*.

- We choose $\boldsymbol{w}$ that will maximize the lower bound on $q(\boldsymbol{z})$:

$$\max_{\boldsymbol{w}\in\mathbb{R}^S_{>0}} \left\{\bar{\boldsymbol{p}}^\mathsf{T}\boldsymbol{z} - \psi\|\boldsymbol{z} - \bar{\lambda}\mathbf{1}\|_{\infty,\frac{1}{\boldsymbol{w}}} \;:\; \sum_{i=1}^{S} w_i^2 = 1\right\}$$

Idea 2: Near-optimal Bayesian Ambiguity Sets

Value-function driven near-optimal ambiguity sets

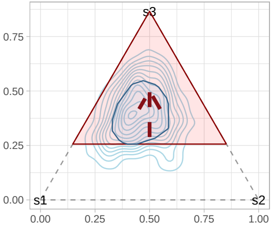


Red: Optimal set for for a known value function $v = [0, 0, 1]$

Blue: Optimal set for all possible value functions.

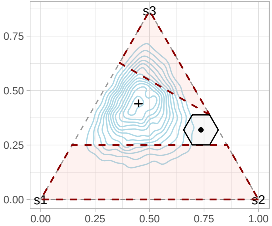
Idea 2: Near-optimal Bayesian Ambiguity Sets

Value-function driven near-optimal ambiguity sets



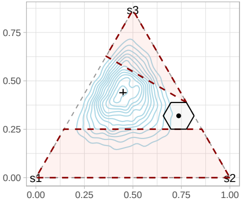
Red: Optimal set for for a known value function $v = [0, 0, 1]$

Blue: Optimal set for all possible value functions.



Near-optimal ambiguity sets constructed for two possible value functions: $v_1 = (0, 0, 1)$ and $v_2 = (2, 1, 0)$

Idea 2: Near-optimal Bayesian Ambiguity Sets



Near-optimal sets:

- **Step 1:** Find the half-space for each value function:

$$\mathcal{K}_{s,a}(v) = \left\{ p \in \Delta^S \ : \ p^\mathsf{T} v \le \text{V@R}^{\zeta}_{P^\star} \left[(p^\star_{s,a})^\mathsf{T} v \right] \right\}$$

- **Step 2:** Find minimal set intersecting each half-space.
- **Step 3:** Compute robust solution and iterate.

University of New Hampshire

Near-optimal Bayesian Ambiguity Sets

- *Optimal* ambiguity set for a known value function v:

$$\mathcal{K}_{s,a}(v) = \left\{ p \in \Delta^S \ : \ p^T v \leq \text{V@R}^{\zeta}_{P^\star} \left[(p^\star_{s,a})^T v \right] \right\} \ ,$$

- Approximation of optimal ambiguity set for a set of possible value functions:

$$\mathcal{L}_{s,a}(\mathcal{V}) = \{ p \in \Delta^S \ : \ \|p - \theta_{s,a}(\mathcal{V})\|_1 \leq \psi_{s,a}(\mathcal{V}) \},$$

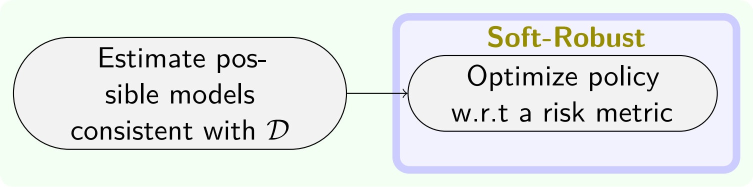
$$\psi_{s,a}(\mathcal{V}) = \min_{p \in \Delta^S} f(p), \quad \theta_{s,a}(\mathcal{V}) \in \arg\min_{p \in \Delta^S} f(p),$$

$$f(p) = \max_{v \in \mathcal{V}} \min_{q \in \mathcal{K}_{s,a}(v)} \|q - p\|_1$$

Theorem

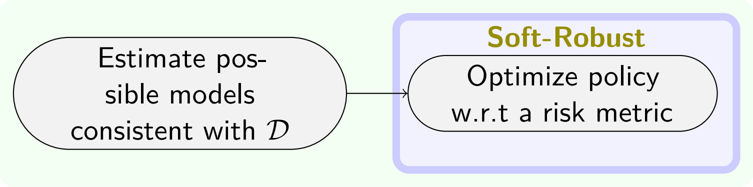
Suppose that the algorithm terminates with a policy $\hat{\pi}_k$ and a value function $\hat{v}_k$ in the iteration k. Then, the return estimate $\tilde{\rho}(\hat{\pi}) = p_0^T \hat{v}_k$ is safe: $\mathbb{P}_{P^\star} \left[p_0^T \hat{v}_k \leq p_0^T v^{\hat{\pi}_k}_{P^\star} \;\middle|\; \mathcal{D} \right] \geq 1 - \delta$.

Soft-Robust Methods



University of New Hampshire

Soft-Robust Methods



Related Works:

- [10] proposed soft-robust policy-gradient and actor-critic methods constrained by a fixed ambiguity set.
- [13] propose entropic and CV@R risk constrained policy gradient in Bayesian setting.

University of New Hampshire

Idea 3: Soft-Robustness with Entropic Risk

- **Objective:**

$$\max_{\theta} \mathbb{E}_{\mathcal{M}}\left[\mathbb{E}_{\xi}\left[g^{\theta}(\xi)\right]\right]$$
$$\text{s.t.} \quad -\frac{1}{\alpha}\log\left(\mathbb{E}_{\mathcal{M}}\left[e^{-\alpha \mathbb{E}_{\xi}[g^{\theta}(\xi)]}\right]\right) \geq \beta$$

University of New Hampshire

Idea 3: Soft-Robustness with Entropic Risk

- **Objective:**

$$\max_{\theta} \mathbb{E}_{\mathcal{M}}\left[\mathbb{E}_{\xi}\left[g^{\theta}(\xi)\right]\right]$$

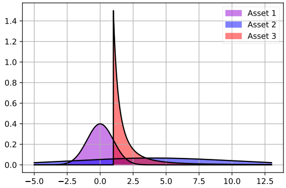
$$\text{s.t.} \quad -\frac{1}{\alpha}\log\left(\mathbb{E}_{\mathcal{M}}\left[e^{-\alpha \mathbb{E}_{\xi}[g^{\theta}(\xi)]}\right]\right) \geq \beta$$

- Derive gradient update rule:

$$\nabla_{\theta} L(\theta,\lambda) = \sum_{\mathcal{M}} P(\mathcal{M}) \sum_{\xi: P_{\theta,\mathcal{M}}(\xi)\neq 0} g(\xi) P_{\theta,\mathcal{M}}(\xi)\Big(1 - \alpha\lambda e^{-\alpha \sum_{\xi: P_{\theta,\mathcal{M}}(\xi)\neq 0} P_{\theta,\mathcal{M}}(\xi) g(\xi)}\Big) \sum_{k=0}^{T-1} \frac{\nabla_{\theta}\pi_{\theta}(a_k|s_k)}{\pi_{\theta}(a_k|s_k)}$$

$$\nabla_{\lambda} L(\theta,\lambda) = \sum_{\mathcal{M}} P(\mathcal{M}) e^{-\alpha \sum_{\xi: P_{\theta}(\xi)\neq 0} P_{\theta,\mathcal{M}}(\xi) g(\xi)} - e^{-\alpha\beta}$$

Idea 3: Empirical Evaluation



- **Asset 1:** Standard normal.
- **Asset 2:** Normal with $\mu = 4$ and $\sigma = 6$.
- **Asset 3:** Pareto distribution with shape $a = 1.5$, scale $m = 1$ and pdf $p(x) = \frac{am^a}{x^{a+1}}$.

Convergence Analysis

Corollary

When $S_\phi(X) = \rho^\alpha_{ent}(X)$ for some $\alpha \in (0, 1]$, then we have:

$$P(|\hat{\rho}^\alpha_{ent}(X_1, \ldots, X_N) - \rho^\alpha_{ent}(X)| \geq \varepsilon) \leq 2e^{-2\alpha^2\varepsilon^2 N}$$

Theorem

Under assumptions **(A1)** - **(A7)** stated in Appendix of the paper, the sequence of parameter updates of the policy gradient algorithm converges almost surely to a locally optimal policy $\theta^\star$ as $k \to \infty$.

Theorem

Under assumptions **(A1)** - **(A7)** stated in Appendix of the paper, the sequence of parameter updates of actor-critic Algorithm converges almost surely to a locally optimal policy

Robust Constrained MDP

Constrained MDPs: MDPs with multiple reward functions [2].
Robust CMDPs: Incorporate robustness into CMDPs.

Related Works:

- [12] Proposes methods to find robust optimal policies with safety-threshold constraints.
- [21] Proposes methods to optimize policy robust to constrained model misspecification.

Idea 4: Robust Constrained Policy Optimization

- Objective:

$$\max_{\pi\in\Pi}\min_{p\in\mathcal{P}}\mathbb{E}_p\left[\sum_{t=0}^{\infty}\gamma^t c(s_t,a_t)\right]$$

$$\text{s.t.}\quad \min_{p\in\mathcal{P}}\mathbb{E}_p\left[\sum_{t=0}^{\infty}\gamma^t d(s_t,a_t)\right]\leq d_0$$

- Formulate Lagrange:

$$\max_{\lambda\geq 0}\min_{\theta}\left(L(\theta,\lambda)=\hat{v}^{\pi}_{\mathcal{P}}(s)+\lambda\Big(\hat{u}^{\pi}_{\mathcal{P}}(s)-d_0\Big)\right)$$

- Derive gradient update rule:

$$\nabla_\theta L(\theta,\lambda)=\sum_{\xi}\hat{p}^{\theta}(\xi)\Big(g(\xi)+\lambda h(\xi)\Big)\sum_{t=0}^{T-1}\frac{\nabla_\theta\pi_\theta(a_t|s_t)}{\pi_\theta(a_t|s_t)}$$

$$\nabla_\lambda L(\theta,\lambda)=\sum_{\xi}\hat{p}^{\theta}_{u}(\xi)h(\xi)-d_0$$