0.28pt

Beyond Cumulative Returns Via Reinforcement Learning Over State-Action Occupancy Measures

Junyu Zhang*, Amrit S. Bedi*, Alec Koppel, and Mengdi Wang

2021 American Control Conference (ACC)
New Orleans, USA, May 26-28, 2021

Acknowledgment

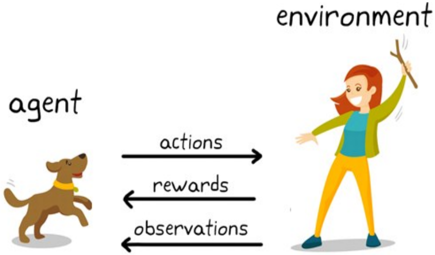
Junyu Zhang (Princeton University)

Amrit Singh Bedi (ARL)

Mengdi Wang (Princeton University)

Reinforcement Learning

- Reinforcement learning: data-driven control



[1]

- Recent successes:
  - ⇒ AlphaGo[2]
  - ⇒ Bipedal walker on terrain[3]
  - ⇒ Personalized web services[4]

---

[1] https://www.kdnuggets.com/2019/10/mathworks-reinforcement-learning.html

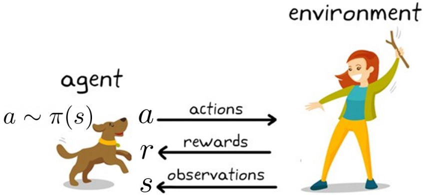
[2] Silver, D. et al., Mastering the game of Go without human knowledge. Nature 550, 354–359 (2017).

[3] Heess, N. et al., Learning continuous control policies by stochastic value gradients. In NeurIPS, 2015.

[4] Theocharous, G., "Ad recommendation systems for life-time value optimization." In ICWWW, pp. 1305-1310. 2015.

Problem Formulation

- Markov decision process (MDP) $(\mathcal{S}, \mathcal{A}, \mathbb{P}, r, \gamma)$
  - $\Rightarrow$ State space $\mathcal{S}$, action space $\mathcal{A}$
  - $\Rightarrow$ Transitions $\mathbb{P}(s' \mid s, a) : \mathcal{S} \times \mathcal{A} \to \mathcal{P}(\mathcal{S})$
  - $\Rightarrow$ Reward $r : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$, discount factor $\gamma \in (0, 1)$



- Infinite-horizon setting value function:

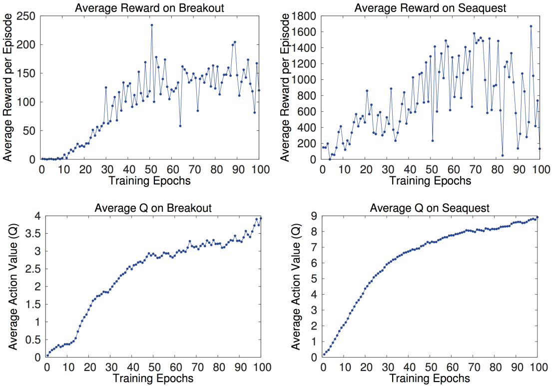
$$V_\pi(s) = \mathbb{E}\left(\sum_{t=0}^{\infty} \gamma^t \cdot r(s_t, a_t) \,\middle|\, s_0 = s, a_t \sim \pi(s_t) \right)$$

- Goal: find $\{a_t = \pi(s_t)\}$ to maximize $V_\pi(s)$

DEVCOM ARMY RESEARCH LABORATORY

Sample Inefficiency and High Variance

Performance on Deep Q Networks

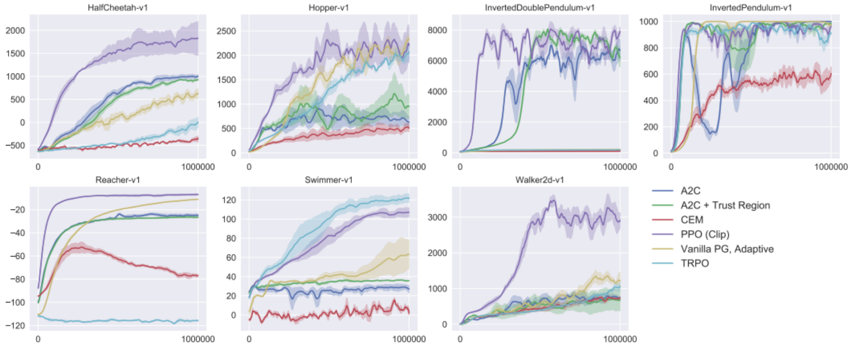


$\Rightarrow$ High variance and millions of samples until convergence[5]

[5] Mnih, V., Kavukcuoglu, K., Silver, D., Graves, A., Antonoglou, I., Wierstra, D., & Riedmiller, M. (2013). Playing atari with deep reinforcement learning. arXiv preprint arXiv:1312.5602.

Sample Inefficiency and High Variance

Performance of Proximal Policy Optimization



⇒ High variance and millions of samples until convergence [5]

[5] Schulman, J., Wolski, F., Dhariwal, P., Radford, A., & Klimov, O. (2017). Proximal policy optimization algorithms. arXiv preprint arXiv:1707.06347.

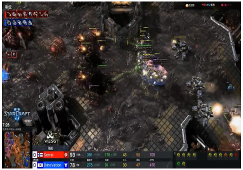
Motivation

- Possible sources:
  - ⇒ higher-order moments of transition dynamics
  - ⇒ reward function is sparse (zero at majority of states)
  - ⇒ “cold start” requires exploration & visiting unrewarding states

- Motivates question of how to **improve reliability of return**
  - ⇒ burgeoning work in policy search: exploration, risk, & imitation

(a) Exploration

(b) Risk sensitivity



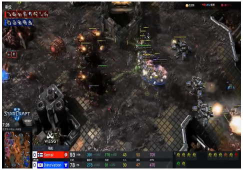
(c) Imitation

Motivation

- Possible sources:
  - ⇒ higher-order moments of transition dynamics
  - ⇒ reward function is sparse (zero at majority of states)
  - ⇒ “cold start” requires exploration & visiting unrewarding states
- Motivates question of how to **improve reliability of return**
  - ⇒ burgeoning work in policy search: exploration, risk, & imitation

(a) Exploration

(b) Risk sensitivity



(c) Imitation

DEVCOM

Context

- Broader decision-making goals incur time-inconsistency
  $\Rightarrow$ lack of additive structure $\Rightarrow$ Bellman's equations to break down
- Existing approaches:
  $\Rightarrow$ Modified Bellman equations[6], multi-stage[7]
  $\Rightarrow$ Do not attain near optimal solution in polynomial time
  $\Rightarrow$ Bayesian[8] and dist. RL [27][9] $\rightarrow$ track full posterior
  $\Rightarrow$ Efficient and convergent dist. models-active research area
- Proposed
  $\Rightarrow$ Given high variance, how to impose risk
  $\Rightarrow$ We develop Cautious RL[10] $\Rightarrow$ builds on LP formulation of RL
  $\Rightarrow$ Can be solved efficiently in polynomial time

---

[6] A. Ruszczynski, "Risk-averse dynamic programming for markov decision processes," ′ Math. Prog., vol. 125, no. 2, pp. 235–261, 2010.

[7] D. R. Jiang et al., "Risk-averse approximate dynamic programming with quantile-based risk measures," Math. Oper. Res., vol. 43, 2018.

[8] M. Ghavamzadeh et al., "Bayesian reinforcement learning: A survey," Found. Trends Mach. Learn., vol. 8, no. 5-6, pp. 359–483, 2015.

[9] M. G. Bellemare et al., "A distributional perspective on reinforcement learning," in 34th Int Conf Mach. Learn. (ICML), 2017, pp. 449–458.

[10] J. Zhang et al., "Cautious RL via Dist. Risk in the Dual Domain" in ACC 2021 (J. sub. to IEEE JSAIT) [*Equal contr.]

Linear Programming for RL

- Goal: find $\{a_t = \pi(s_t)\}$ to maximize $V_\pi(s) := \mathbb{E}[V(s) \,|\, a \sim \pi(s)]$
- Bellman's optimality principle[11] $[r(s,a) = \mathbb{E}_{s' \sim \mathcal{P}(\cdot|a,s)}[\hat{r}_{ss'a}]]$

$$v^*(s) = \max_{a \in \mathcal{A}} \left\{ r(s,a) + \gamma \sum_{s' \in \mathcal{S}} P_a(s,s') v^*(s') \right\}$$

- Linear prog. reformulation[12] $[(v, \xi, r_a) \in \mathbb{R}^{|\mathcal{S}|}, P_a \in \mathbb{R}^{|\mathcal{S}| \times |\mathcal{S}|}]$

$$\min_{v \geq 0} \langle \xi, v \rangle \;\text{ s.t. }\; (I - \gamma P_a) v - r_a \geq 0, \;\text{ for all } a \in \mathcal{A}$$

- Dual LP $[\lambda_a \in \mathbb{R}^{|\mathcal{S}|}]$

$$\max_{\lambda \geq 0} \sum_{a \in \mathcal{A}} \langle \lambda_a, r_a \rangle \;\text{ s.t. }\; \sum_{a \in \mathcal{A}} (I - \gamma P_a^\top) \lambda_a = \xi, \;\text{ for all } a \in \mathcal{A}$$

- λ denotes the occupancy measure across state-action space

$$\lambda_{sa} = \sum_{t=0}^{\infty} \gamma^t \cdot \mathbb{P}\left(s_t = s, a_t = a \,\middle|\, s_0 \sim \xi, a_t \sim \pi(\cdot|s_t) \right) \text{ and } \pi(a|s) = \frac{\lambda_{sa}}{\sum_{a'} \lambda_{sa'}}$$

---

[11] Bertsekas, D. P. et al., Stochastic optimal control: the discrete-time case. 2004.

[12] De Farias, D. P., Van Roy, B.., The linear programming approach to approximate dynamic programming. Operations research, 2003.

Linear Programming for RL

- Goal: find $\{a_t = \pi(s_t)\}$ to maximize $V_\pi(s) := \mathbb{E}[V(s) \,|\, a \sim \pi(s)]$
- Bellman's optimality principle[11] $[r(s,a) = \mathbb{E}_{s' \sim \mathcal{P}(\cdot|a,s)}[\hat{r}_{ss'a}]]$

$$v^*(s) = \max_{a \in \mathcal{A}} \left\{ r(s,a) + \gamma \sum_{s' \in \mathcal{S}} P_a(s,s') v^*(s') \right\}$$

- Linear prog. reformulation[12] $[(v, \xi, r_a) \in \mathbb{R}^{|\mathcal{S}|}, P_a \in \mathbb{R}^{|\mathcal{S}| \times |\mathcal{S}|}]$

$$\min_{v \geq 0} \langle \xi, v \rangle \text{ s.t. } (I - \gamma P_a)v - r_a \geq 0, \quad \text{for all } a \in \mathcal{A}$$

- Dual LP $[\lambda_a \in \mathbb{R}^{|\mathcal{S}|}]$

$$\max_{\lambda \geq 0} \sum_{a \in \mathcal{A}} \langle \lambda_a, r_a \rangle \text{ s.t. } \sum_{a \in \mathcal{A}} (I - \gamma P_a^\top)\lambda_a = \xi, \quad \text{for all } a \in \mathcal{A}$$

- λ denotes the occupancy measure across state-action space

$$\lambda_{sa} = \sum_{t=0}^{\infty} \gamma^t \cdot \mathbb{P}\left(s_t = s, a_t = a \,\middle|\, s_0 \sim \xi, a_t \sim \pi(\cdot|s_t) \right) \text{ and } \pi(a|s) = \frac{\lambda_{sa}}{\sum_{a'} \lambda_{sa'}}$$

[11] Bertsekas, D. P. et al., Stochastic optimal control: the discrete-time case. 2004.

[12] De Farias, D. P., Van Roy, B.., The linear programming approach to approximate dynamic programming. Operations research, 2003.

Linear Programming for RL

- Goal: find $\{a_t = \pi(s_t)\}$ to maximize $V_\pi(s) := \mathbb{E}[V(s) \mid a \sim \pi(s)]$
- Bellman's optimality principle[11] $[r(s,a) = \mathbb{E}_{s' \sim \mathcal{P}(\cdot|a,s)}[\hat{r}_{ss'a}]]$

$$v^*(s) = \max_{a \in \mathcal{A}} \left\{ r(s,a) + \gamma \sum_{s' \in \mathcal{S}} P_a(s,s') v^*(s') \right\}$$

- Linear prog. reformulation[12] $[(v, \xi, r_a) \in \mathbb{R}^{|\mathcal{S}|}, P_a \in \mathbb{R}^{|\mathcal{S}| \times |\mathcal{S}|}]$

$$\min_{v \geq 0} \langle \xi, v \rangle \text{ s.t. } (I - \gamma P_a) v - r_a \geq 0, \text{ for all } a \in \mathcal{A}$$

- Dual LP $[\lambda_a \in \mathbb{R}^{|\mathcal{S}|}]$

$$\max_{\lambda \geq 0} \sum_{a \in \mathcal{A}} \langle \lambda_a, r_a \rangle \text{ s.t. } \sum_{a \in \mathcal{A}} (I - \gamma P_a^\top) \lambda_a = \xi, \quad \text{for all } a \in \mathcal{A}$$

- λ denotes the occupancy measure across state-action space

$$\lambda_{sa} = \sum_{t=0}^{\infty} \gamma^t \cdot \mathbb{P}\left(s_t = s, a_t = a \;\middle|\; s_0 \sim \xi, a_t \sim \pi(\cdot|s_t) \right) \text{ and } \pi(a|s) = \frac{\lambda_{sa}}{\sum_{a'} \lambda_{sa'}}$$

[11] Bertsekas, D. P. et al., Stochastic optimal control: the discrete-time case. 2004.

[12] De Farias, D. P., Van Roy, B.., The linear programming approach to approximate dynamic programming. Operations research, 2003.

Linear Programming for RL

- Goal: find $\{a_t = \pi(s_t)\}$ to maximize $V_\pi(s) := \mathbb{E}[V(s) \,|\, a \sim \pi(s)]$
- Bellman's optimality principle[11] $[r(s,a) = \mathbb{E}_{s' \sim \mathcal{P}(\cdot|a,s)}[\hat{r}_{ss'a}]]$

$$v^*(s) = \max_{a \in \mathcal{A}} \left\{ r(s,a) + \gamma \sum_{s' \in \mathcal{S}} P_a(s,s') v^*(s') \right\}$$

- Linear prog. reformulation[12] $[(v, \xi, r_a) \in \mathbb{R}^{|\mathcal{S}|}, P_a \in \mathbb{R}^{|\mathcal{S}| \times |\mathcal{S}|}]$

$$\min_{v \geq 0} \langle \xi, v \rangle \text{ s.t. } (I - \gamma P_a)v - r_a \geq 0, \quad \text{for all } a \in \mathcal{A}$$

- Dual LP $[\lambda_a \in \mathbb{R}^{|\mathcal{S}|}]$

$$\max_{\lambda \geq 0} \sum_{a \in \mathcal{A}} \langle \lambda_a, r_a \rangle \text{ s.t. } \sum_{a \in \mathcal{A}} (I - \gamma P_a^\top)\lambda_a = \xi, \quad \text{for all } a \in \mathcal{A}$$

- λ denotes the occupancy measure across state-action space

$$\lambda_{sa} = \sum_{t=0}^{\infty} \gamma^t \cdot \mathbb{P}\left(s_t = s, a_t = a \,\middle|\, s_0 \sim \xi, a_t \sim \pi(\cdot|s_t) \right) \text{ and } \pi(a|s) = \frac{\lambda_{sa}}{\sum_{a'} \lambda_{sa'}}$$

[11] Bertsekas, D. P. et al., Stochastic optimal control: the discrete-time case. 2004.

[12] De Farias, D. P., Van Roy, B.., The linear programming approach to approximate dynamic programming. Operations research, 2003.

Linear Programming for RL

- Goal: find $\{a_t = \pi(s_t)\}$ to maximize $V_\pi(s) := \mathbb{E}[V(s) \,|\, a \sim \pi(s)]$
- Bellman's optimality principle[11] $[r(s,a) = \mathbb{E}_{s' \sim \mathcal{P}(\cdot|a,s)}[\hat{r}_{ss'a}]]$

$$v^*(s) = \max_{a \in \mathcal{A}} \left\{ r(s,a) + \gamma \sum_{s' \in \mathcal{S}} P_a(s,s') v^*(s') \right\}$$

- Linear prog. reformulation[12] $[(v, \xi, r_a) \in \mathbb{R}^{|\mathcal{S}|}, P_a \in \mathbb{R}^{|\mathcal{S}| \times |\mathcal{S}|}]$

$$\min_{v \geq 0} \langle \xi, v \rangle \ \text{ s.t. } \ (I - \gamma P_a) v - r_a \geq 0, \ \text{ for all } a \in \mathcal{A}$$

- Dual LP $[\lambda_a \in \mathbb{R}^{|\mathcal{S}|}]$

$$\max_{\lambda \geq 0} \sum_{a \in \mathcal{A}} \langle \lambda_a, r_a \rangle \ \text{ s.t. } \ \sum_{a \in \mathcal{A}} (I - \gamma P_a^\top) \lambda_a = \xi, \ \text{ for all } a \in \mathcal{A}$$

- λ denotes the occupancy measure across state-action space

$$\lambda_{sa} = \sum_{t=0}^{\infty} \gamma^t \cdot \mathbb{P}\left(s_t = s, a_t = a \,\middle|\, s_0 \sim \xi, a_t \sim \pi(\cdot|s_t) \right) \ \text{ and } \ \pi(a|s) = \frac{\lambda_{sa}}{\sum_{a'} \lambda_{sa'}}$$

---

[11] Bertsekas, D. P. et al., Stochastic optimal control: the discrete-time case. 2004.

[12] De Farias, D. P., Van Roy, B.., The linear programming approach to approximate dynamic programming. Operations research, 2003.

Cautious RL

- Proposed Formulation[13] $\Rightarrow$ non-standard notion of risk "Caution"
  $\Rightarrow$ introduce convex caution function $\rho(\lambda)$ into the dual objective

$$\max_{\lambda \geq 0} \langle \lambda, r \rangle - c\rho(\lambda)$$

$$\text{s.t.} \sum_{a \in \mathcal{A}} (I - \gamma P_a^\top)\lambda_a = \xi$$

- Examples:
  $\Rightarrow$ Barrier risk: $\rho(\lambda) = -\log\left(\lambda(\bar{S}) - (1-\delta)\right) \Rightarrow$ staying in $\bar{S}$
  $\Rightarrow$ Incorporating priors: $\rho(\lambda) = \text{KL}\left((1-\gamma)\lambda || p\right)$
  $\Rightarrow$ Variance risk: $\rho(\lambda) = \langle (1-\gamma)\lambda, R \rangle - \langle (1-\gamma)\lambda, r \rangle^2$
  $\Rightarrow R(s,a) = \mathbb{E}_{s' \sim \mathcal{P}(\cdot|a,s)}[\hat{r}^2_{ss'a}]$
- Solution: Stochastic Primal-Dual Algorithm

$$\max_{\lambda \in \mathcal{L}} \min_{v \in \mathcal{V}} L(\lambda, v) = \langle \lambda, r \rangle - c\rho(\lambda) + \langle \xi, v \rangle + \sum_{a \in \mathcal{A}} \lambda_a^\top (\gamma P_a - I) v,$$

[13] Zhang*, Bedi*, Koppel, and Wang, "Cautious Reinforcement Learning via Distributional Risk in the Dual Domain" in ACC 2020 (Journal submitted to IEEE JSAIT)

Cautious RL

- Proposed Formulation[13] $\Rightarrow$ non-standard notion of risk "Caution"
  $\Rightarrow$ introduce convex caution function $\rho(\lambda)$ into the dual objective

$$\max_{\lambda \geq 0} \langle \lambda, r \rangle - c\rho(\lambda)$$
$$\text{s.t.} \sum_{a \in \mathcal{A}} (I - \gamma P_a^\top)\lambda_a = \xi$$

- Examples:
  $\Rightarrow$ Barrier risk: $\rho(\lambda) = -\log\left(\lambda(\bar{S}) - (1-\delta)\right)$ $\Rightarrow$ staying in $\bar{S}$
  $\Rightarrow$ Incorporating priors: $\rho(\lambda) = \text{KL}\left((1-\gamma)\lambda || p\right)$
  $\Rightarrow$ Variance risk: $\rho(\lambda) = \langle (1-\gamma)\lambda, R \rangle - \langle (1-\gamma)\lambda, r \rangle^2$
  $\Rightarrow R(s,a) = \mathbb{E}_{s' \sim \mathcal{P}(\cdot|a,s)}[\hat{r}^2_{ss'a}]$
- Solution: Stochastic Primal-Dual Algorithm

$$\max_{\lambda \in \mathcal{L}} \min_{v \in \mathcal{V}} L(\lambda, v) = \langle \lambda, r \rangle - c\rho(\lambda) + \langle \xi, v \rangle + \sum_{a \in \mathcal{A}} \lambda_a^\top (\gamma P_a - I)v,$$

[13] Zhang*, Bedi*, Koppel, and Wang, "Cautious Reinforcement Learning via Distributional Risk in the Dual Domain" in ACC 2020 (Journal submitted to IEEE JSAIT)

Cautious RL

- Proposed Formulation[13] $\Rightarrow$ non-standard notion of risk "Caution"
  $\Rightarrow$ introduce convex caution function $\rho(\lambda)$ into the dual objective

$$\max_{\lambda \geq 0} \langle \lambda, r \rangle - c\rho(\lambda)$$
$$\text{s.t.} \sum_{a \in \mathcal{A}} (I - \gamma P_a^\top)\lambda_a = \xi$$

- Examples:
  $\Rightarrow$ Barrier risk: $\rho(\lambda) = -\log\left(\lambda(\bar{S}) - (1-\delta)\right)$ $\Rightarrow$ staying in $\bar{S}$
  $\Rightarrow$ Incorporating priors: $\rho(\lambda) = \text{KL}\left((1-\gamma)\lambda || p\right)$
  $\Rightarrow$ Variance risk: $\rho(\lambda) = \langle (1-\gamma)\lambda, R\rangle - \langle (1-\gamma)\lambda, r\rangle^2$
  $\Rightarrow$ $R(s,a) = \mathbb{E}_{s' \sim \mathcal{P}(\cdot|a,s)}[\hat{r}^2_{ss'a}]$
- Solution: Stochastic Primal-Dual Algorithm

$$\max_{\lambda \in \mathcal{L}} \min_{v \in \mathcal{V}} L(\lambda, v) = \langle \lambda, r \rangle - c\rho(\lambda) + \langle \xi, v \rangle + \sum_{a \in \mathcal{A}} \lambda_a^\top (\gamma P_a - I) v,$$

[13] Zhang*, Bedi*, Koppel, and Wang, "Cautious Reinforcement Learning via Distributional Risk in the Dual Domain" in ACC 2020 (Journal submitted to IEEE JSAIT)

Cautious RL Algorithm

- **Input:** Sample size T. Parameter $\xi = \frac{1}{|\mathcal{S}|} \cdot \mathbf{1}$. Stepsizes $\alpha, \beta > 0$. Discount $\gamma \in (0, 1)$
- **Initialize:** Arbitrary $v^1 \in \mathcal{V}$ and $\lambda^1 := \frac{1}{|\mathcal{S}||\mathcal{A}|(1-\gamma)} \cdot \mathbf{1} \in \mathcal{L}$.
- **For** $t = 1, 2, \cdots, T$
  - $\Rightarrow$ Sample (s_t, a_t) uniformly and $\bar{s}_t \sim \xi$.
  - $\Rightarrow$ Generate $s'_t \sim \mathcal{P}(\cdot|a_t, s_t)$ & $\hat{r}_{s_t s'_t a_t}$ from generative model.
  - $\Rightarrow$ Update v and λ as

$$v^{t+1} = \Pi_{\mathcal{V}}(v^t - \alpha \hat{\nabla}_v L(v^t, \lambda^t)) \quad (1)$$

$$\lambda' = \underset{\lambda}{\text{argmax}} \, \langle \hat{\partial}_\lambda L(v^t, \lambda^t), \lambda - \lambda^t \rangle - (1/(1-\gamma)\beta) KL\big((1-\gamma)\lambda \,||(1-\gamma)\lambda^t\big).$$

$$\lambda^{t+1} = \lambda' / (1-\gamma) \|\lambda'\|_1 \quad (2)$$

- **Output:** $\bar{\lambda} := \frac{1}{T} \sum_{t=1}^{T} \lambda^t$ and $\bar{v} := \frac{1}{T} \sum_{t=1}^{T} v^t$.

Performance Guarantees

- Convex, non-smooth $\rho(\lambda)$, bounded subgradients, for Dulity Gap $\leq \epsilon$

$$T \geq \mathcal{O}\left(\frac{|\mathcal{S}||\mathcal{A}|\log(|\mathcal{S}||\mathcal{A}|)}{\epsilon^2} \cdot \frac{(1+2c\sigma)^2}{(1-\gamma)^4}\right)$$

- After T iterations, the constraint violation is ($\bar{\lambda}$ is the output)

$$\left\|\sum_{a\in\mathcal{A}}(I-\gamma P_a^\top)\bar{\lambda}_a - \xi\right\|_1 \leq \frac{(1-\gamma)\epsilon}{1+c\sigma} \leq (1-\gamma)\epsilon$$

- After T iterations, the sub-optimality is given by ($\bar{v}$ is the output)

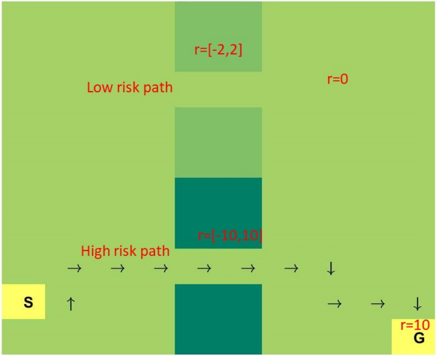
$$\mathbb{E}[(\langle\lambda^*, r\rangle - c\rho(\lambda^*)) - (\langle\bar{\lambda}, r\rangle - c\rho(\bar{\lambda}))] \leq \epsilon$$

A Simple Motivating Example

Maze World:

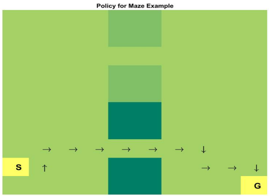
- Consider the problem of reaching the goal

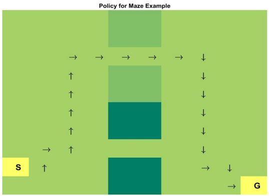
Policy for Maze Example

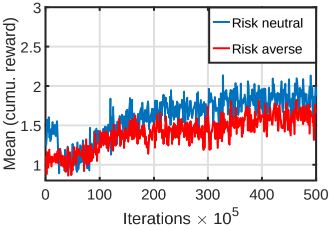


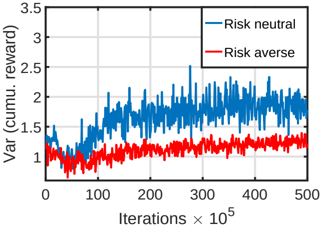
Proof of Concept Experiments

- Variance sensitive policy optimization





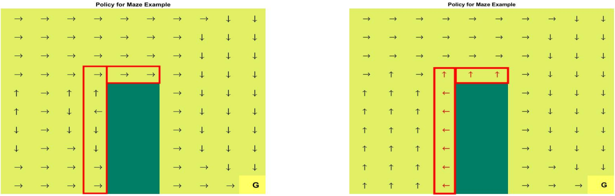


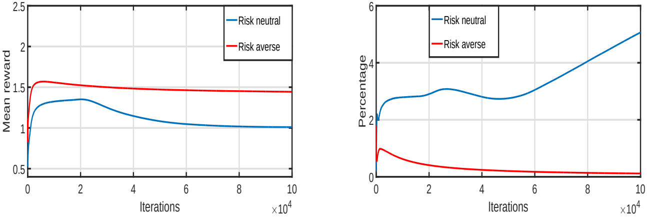


Proof of Concept Experiments

- Caution as Proximity to Prior

  ⇒ Left: risk-neutral policy; Right: risk sensitive policy





⇒ Left: cumulative return of risk-sensitive/neutral policies

⇒ Right: comparison of percentage of time visiting costly states

Conclusion and Future Directions

- Proposed a new definition of risk named "Caution"
- Solved the resulting risk aware RL problem in model free manner
- Derived sample complexities for the proposed primal-dual algorithm
- Verified the approach via experiments
- Future Directions:
  $\Rightarrow$ Deriving Bellman equations associated with Cautious RL
  $\Rightarrow$ Generalizations to continuous spaces

Thank You