Tout ce qui peut mal aller doit être entraîné: une approche antagoniste à la sensibilité au risque en apprentissage par renforcement

Séminaire départemental
14 mars 2025

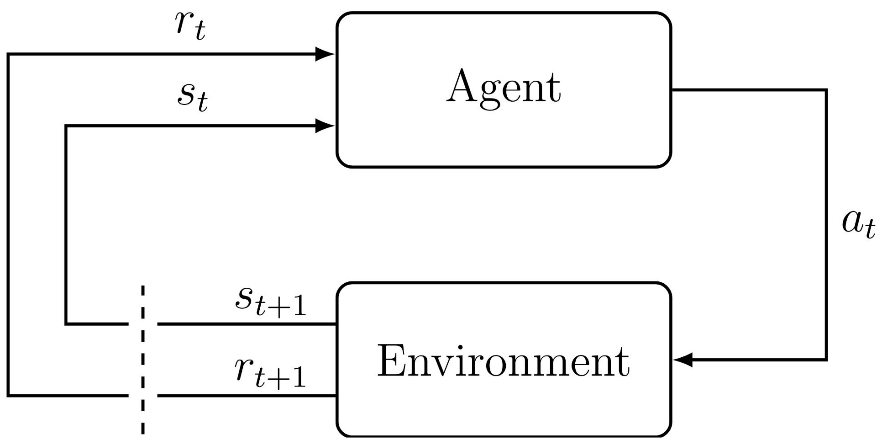
About Me

4th year PhD Student

- Subject: Trustworthy Machine Learning
  - Risk-Sensitive Reinforcement Learning
    - Today's talk!
  - Fairness in Machine Learning
    - Not on today's menu :(!

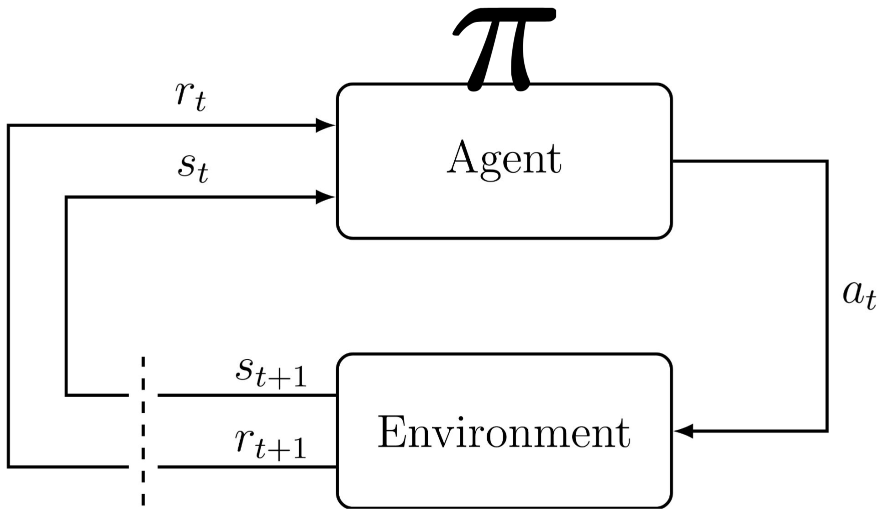
- Advisor: Audrey Durand

RL 101: Markov Decision Process (MDP)



$$\langle \mathcal{S}, \mathcal{A}, P, R, \gamma \rangle$$

RL 101: Markov Decision Process (MDP)



Action $a_t \sim \pi(h_t)$

Next state $s_{t+1} \sim P(\cdot \mid s_t, a_t)$

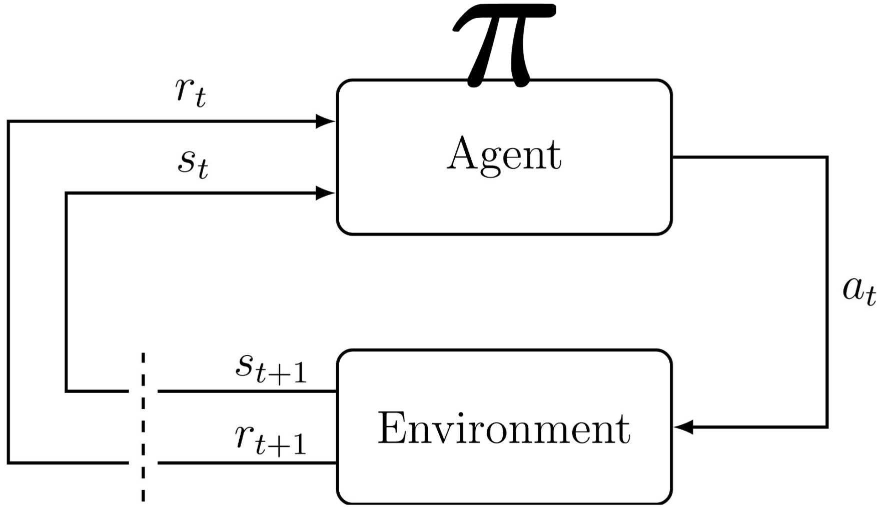
Reward $r_t = R(s_t, a_t)$

History $h_t = (s_0, a_0, r_0, \cdots, s_t)$

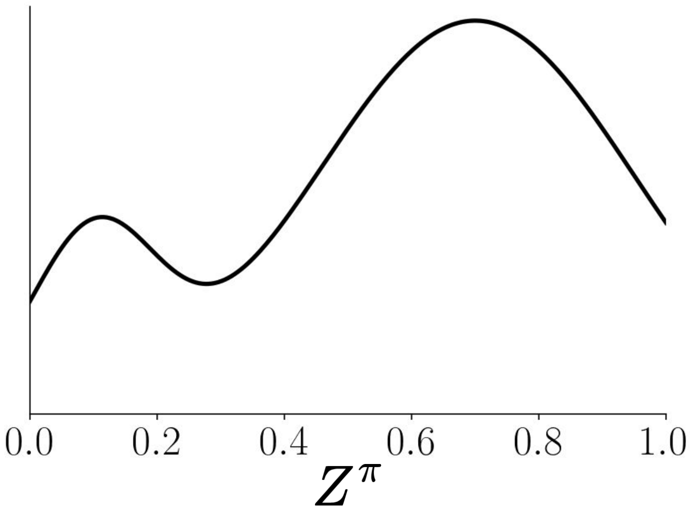
$$\langle \mathcal{S}, \mathcal{A}, P, R, \gamma \rangle$$

Discount Factor

RL 101: Random Total Discounted Return



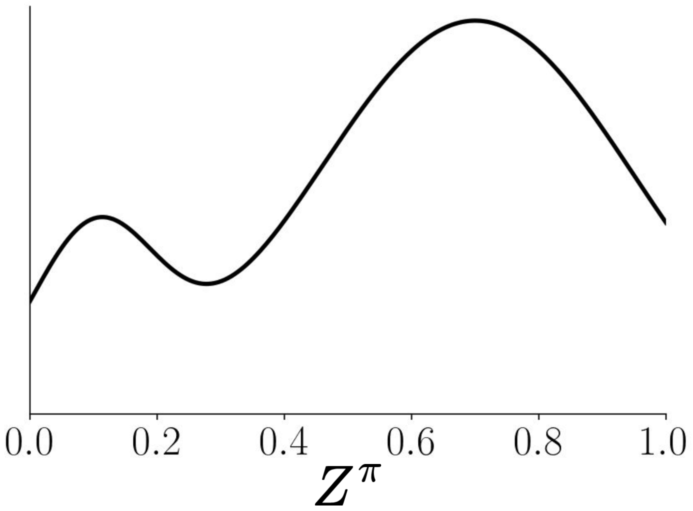
$$Z^{\pi} := \sum_{t=0}^{\infty} \gamma^t r_t$$



Classical RL objective: Expectation Maximization

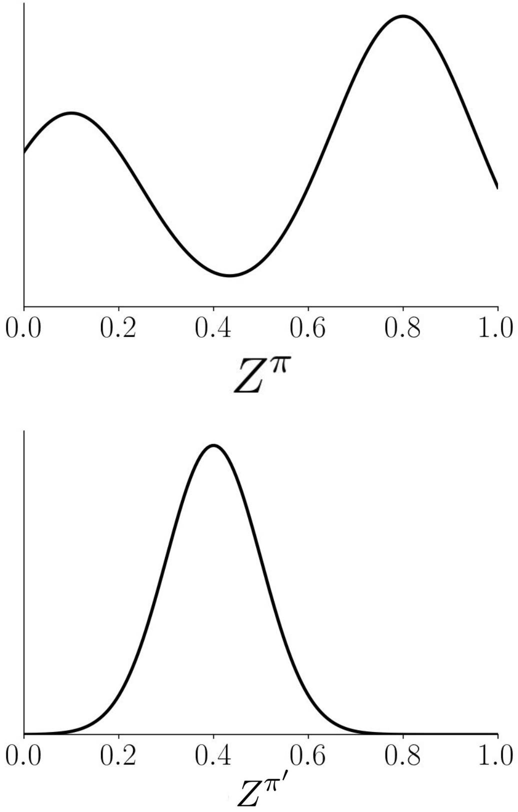
$$Z^{\pi} := \sum_{t=0}^{\infty} \gamma^{t} r_{t}$$

$$\boxed{\pi^{\star} = \arg\max_{\pi} \mathbb{E}[Z^{\pi}]}$$



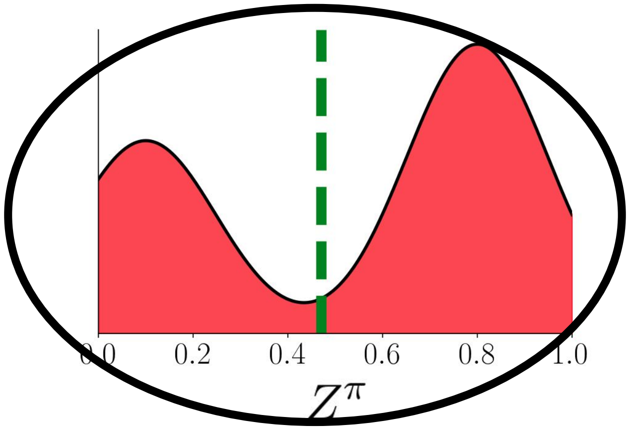
Classical RL objective: Expectation Maximization

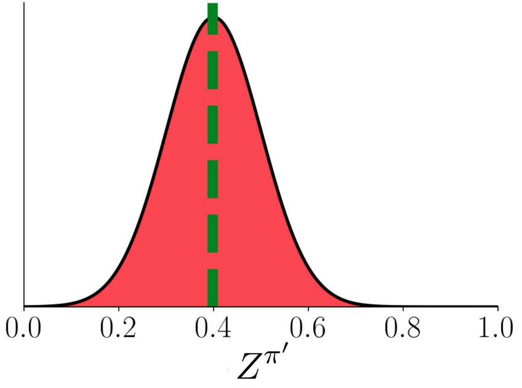
$$\boxed{\pi^\star = \arg\max_{\pi} \mathbb{E}[Z^\pi]}$$



Classical RL objective: Expectation Maximization

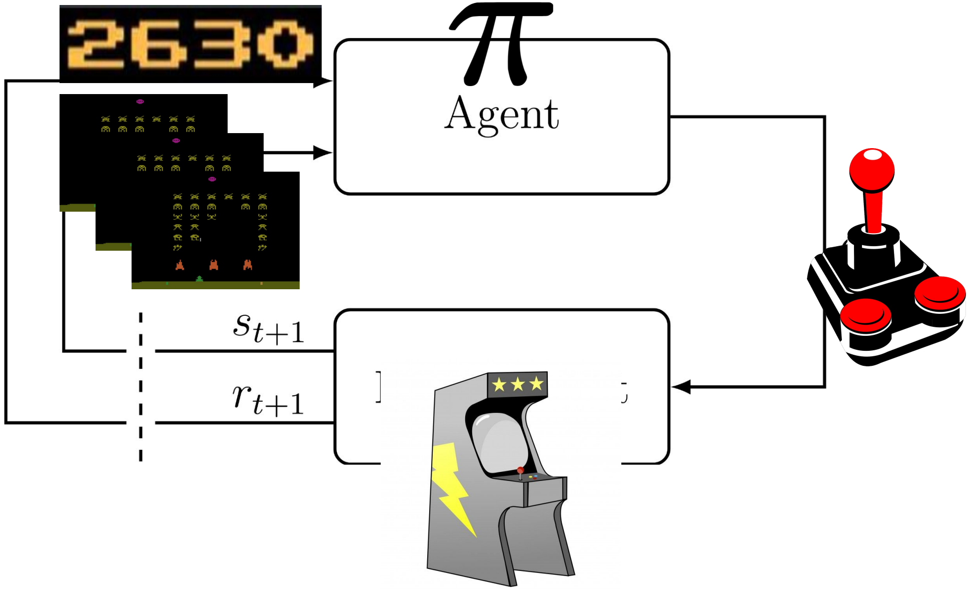
$$\boxed{\pi^\star = \arg\max_\pi \mathbb{E}[Z^\pi]}$$

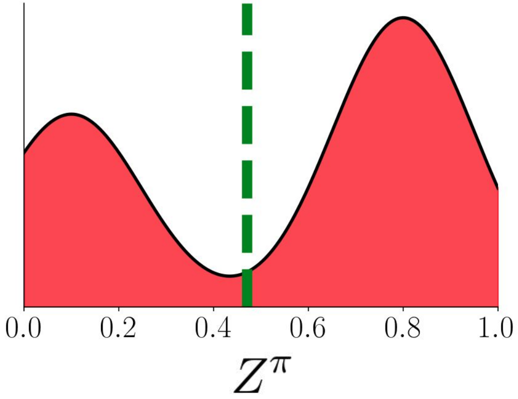


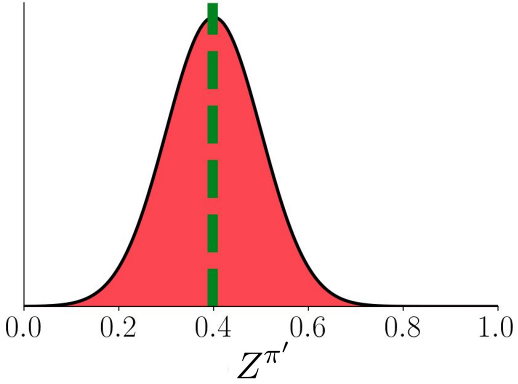


Arcade Learning Environment

[Bellemare et al., 2013]

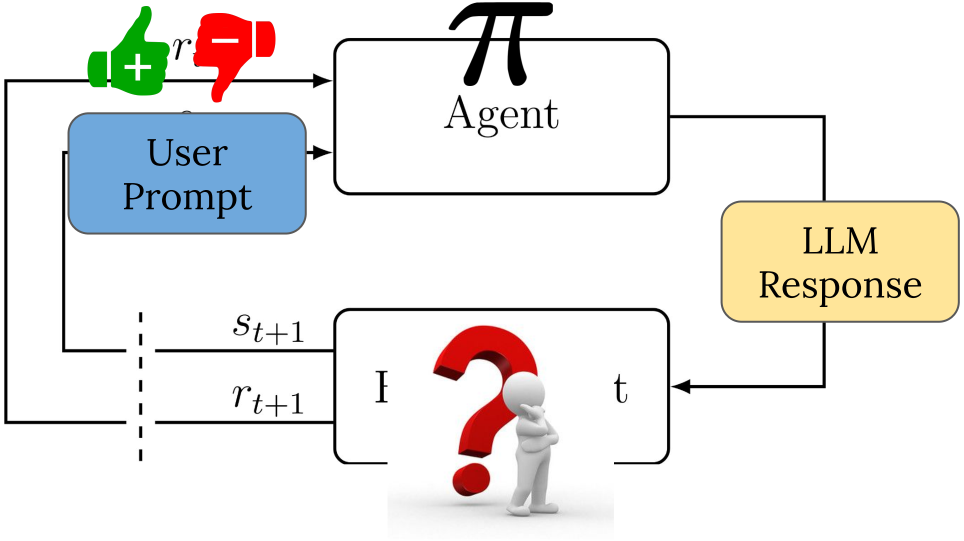


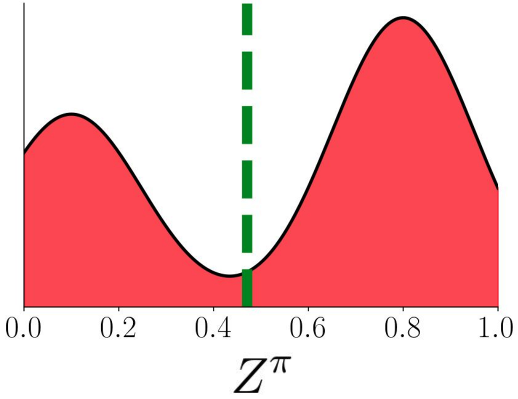


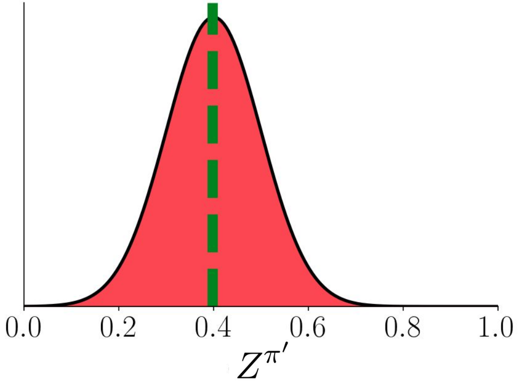


Reinforcement Learning from Human Feedback (RLHF)

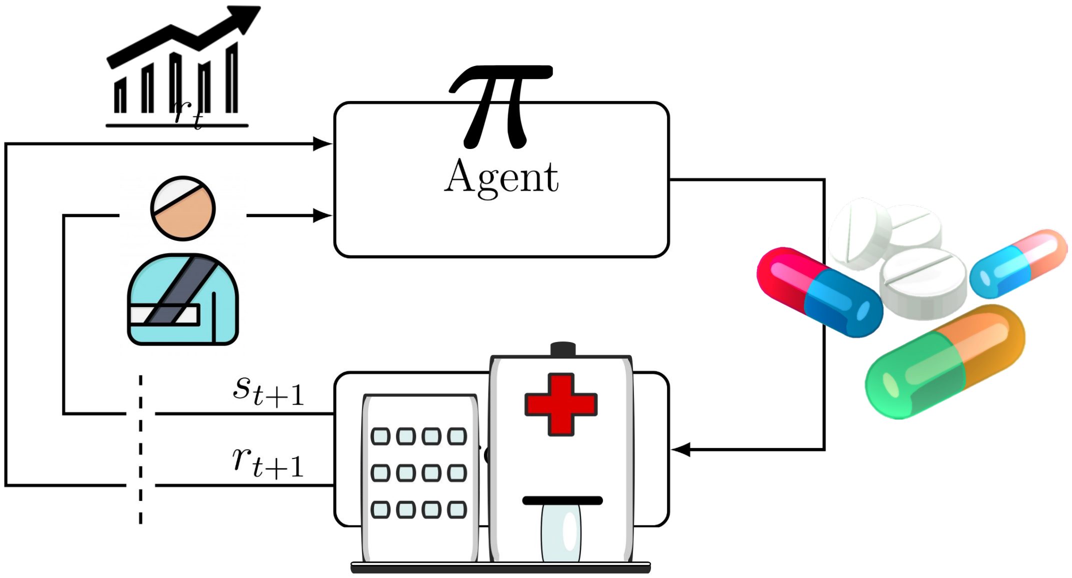
[Ouyang et al., 2022]

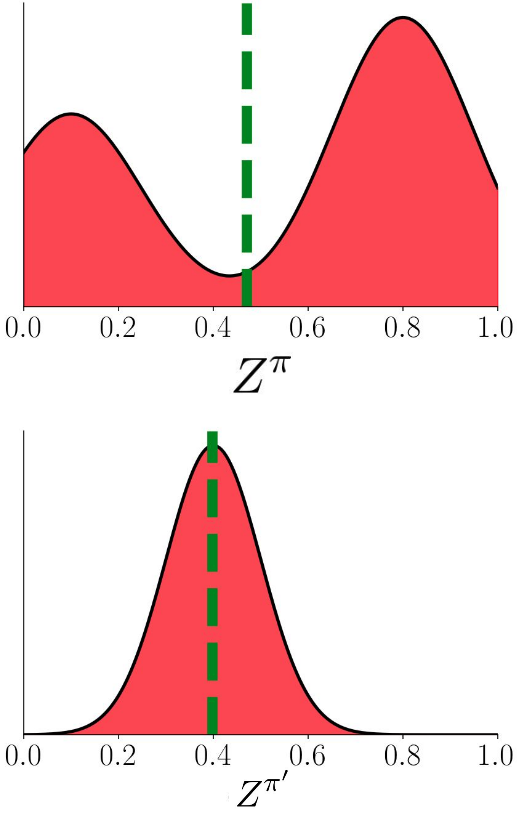




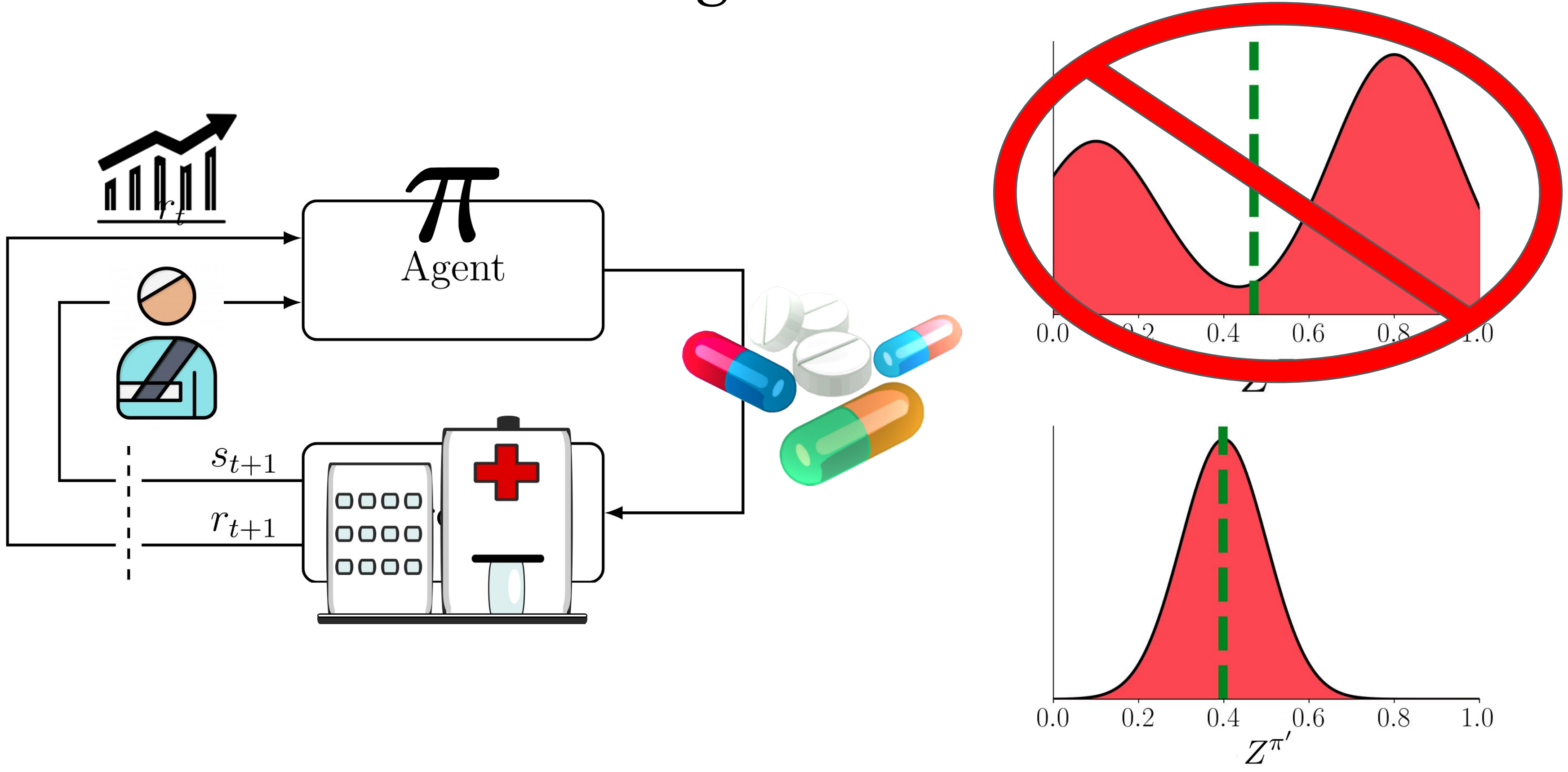


Clinical Treatment Design



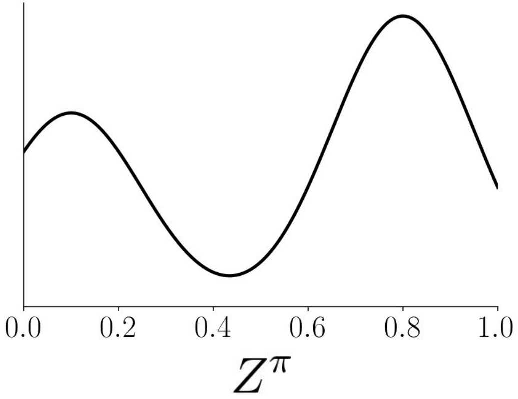


Clinical Treatment Design



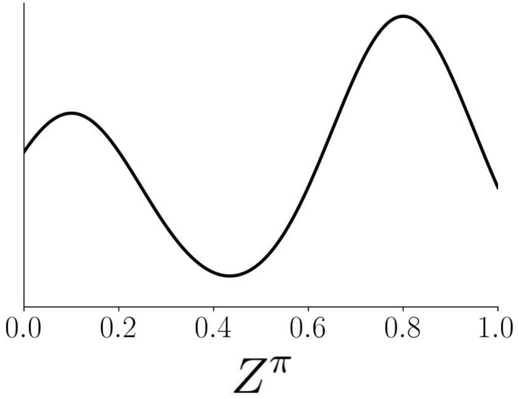
Measuring Risk: Conditional-Value-at-Risk (CVaR)

$$\mathbf{CVaR}_\alpha\left(Z\right) = \mathbb{E}\left[z \;\mid z \leq F_\alpha^{-1}\left(Z\right)\right]$$

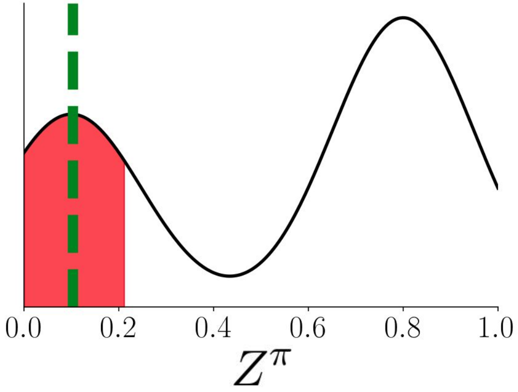


Measuring Risk: Conditional-Value-at-Risk (CVaR)

$$\mathrm{CVaR}_{\alpha}(Z) = \mathbb{E}\left[z \mid z \leq F_{\alpha}^{-1}(Z)\right]$$

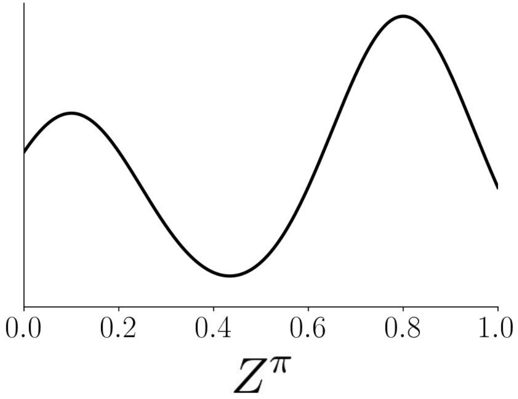


$\mathrm{CVaR}_{0.25}$

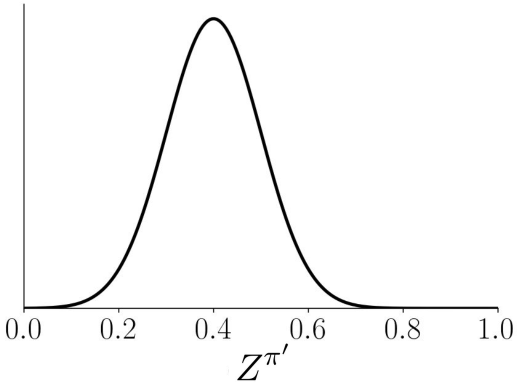


CVaR RL: Risk-Sensitive Objective

$$\mathrm{CVaR}_\alpha(Z) = \mathbb{E}\left[z \mid z \leq F_\alpha^{-1}(Z)\right]$$



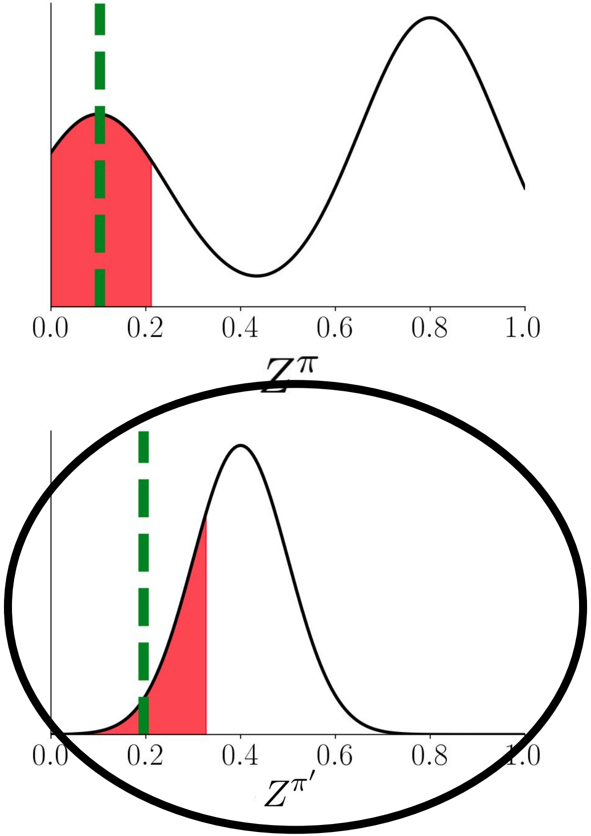
$$\boxed{\pi^\star = \arg\max_\pi \mathrm{CVaR}_\alpha[Z^\pi]}$$



CVaR RL: Risk-Sensitive Objective

$$\mathrm{CVaR}_\alpha(Z) = \mathbb{E}\left[z \mid z \leq F_\alpha^{-1}(Z)\right]$$

$$\boxed{\pi^\star = \arg\max_\pi \mathrm{CVaR}_\alpha[Z^\pi]}$$



Why Use CVaR RL for Safe RL?

Other alternatives exist:

- Reward shaping

- Adversarial training

- Constrained RL

Why Use CVaR RL for Safe RL?

Other alternatives exist:

- Reward shaping
- Adversarial training
- Constrained RL

But CVaR RL is:

- Meaning preserving
- More interpretable
- Assumption-free

CVaR: Actually Used in Practice!

[Basel Committee on Banking Supervision, 2019]

3. Quantitative standards

181. Banks will have flexibility in devising the precise nature of their models, but the following minimum standards will apply for the purpose of calculating their capital charge. Individual banks or their supervisory authorities will have discretion to apply stricter standards.

(a) "*Expected shortfall*" must be computed on a daily basis for the bank-wide internal model for regulatory capital purposes. Expected shortfall must also be computed on a daily basis for each trading desk that a bank wishes to include within the scope for the internal model for regulatory capital purposes.

(b) In calculating the expected shortfall, a 97.5th percentile, one-tailed confidence level is to be used.

The Million Dollar Question:
What difference does it make to change the RL objective to CVaR?

Classical RL: The Fundamental Theorem

The classical RL optimization problem

$$\pi^\star = \arg\max_\pi \mathbb{E}[Z^\pi]$$

can be solved by repeatedly applying the *Bellman Optimality Operator*

$$V_{k+1}^\star(s) = T\left[V_k^\star\right](s) = \max_a \left[R(s,a) + \gamma \sum_{s'} P\left(s' \mid s,a\right) V_k^\star\left(s'\right)\right]$$

which can be cast as a Dynamic Programming (DP).

Classical RL: The Fundamental Theorem

Proof components:

1. A Markovian policy is optimal.

$$\exists\, \pi_m : \boxed{S} \to \Delta(A) \text{ with } \mathbb{E}\left[Z^{\pi_m}\right] = \max_{\pi} \mathbb{E}\left[Z^{\pi}\right]$$

Classical RL: The Fundamental Theorem

Proof components:

1. A Markovian policy is optimal.

$$\exists\, \pi_m : \boxed{S} \to \Delta(A) \text{ with } \mathbb{E}\left[Z^{\pi_m}\right] = \max_{\pi} \mathbb{E}\left[Z^{\pi}\right]$$

2. A deterministic policy is optimal.

$$\exists\, \pi_d : S \to \boxed{A} \text{ with } \mathbb{E}\left[Z^{\pi_d}\right] = \max_{\pi} \mathbb{E}\left[Z^{\pi}\right]$$

Classical RL: The Fundamental Theorem

3. *Fixed Policy Bellman Operator* performs **Policy Evaluation**.

   Define the average discounted returned at state s:

$$V^{\pi}(s) := \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t r_t \mid s_0 = s,\ \pi\right]$$

Classical RL: The Fundamental Theorem

3. *Fixed Policy Bellman Operator* performs **Policy Evaluation**.

   Define the average discounted returned at state s:

$$V^{\pi}(s) := \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t r_t \mid s_0 = s,\ \pi\right]$$

   Unrolling the first term, we have

$$V^{\pi}(s) := \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t r_t \mid s_0 = s,\ \pi\right] = R(s, \pi(s)) + \gamma \sum_{s'} P(s')\, \mathbb{E}\left[\sum_{t=1}^{\infty} \gamma^{t-1} r_t \mid\ s_1 = s',\ \pi\right]$$

Classical RL: The Fundamental Theorem

3. *Fixed Policy Bellman Operator* performs **Policy Evaluation**.

   Define the average discounted returned at state s:

$$V^{\pi}(s) := \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^{t} r_t \mid s_0 = s,\, \pi\right]$$

   Unrolling the first term, we have

$$\boxed{V^{\pi}(s)} := \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^{t} r_t \mid s_0 = s,\, \pi\right] = R(s, \pi(s)) + \gamma \sum_{s'} P(s')\, \underbrace{\mathbb{E}\left[\sum_{t=1}^{\infty} \gamma^{t-1} r_t \mid s_1 = s',\, \pi\right]}_{\boxed{V^{\pi}(s')}}$$

Classical RL: The Fundamental Theorem

3. *Fixed Policy Bellman Operator* performs **Policy Evaluation**.

Because we have

$$V^{\pi}(s) := \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t r_t \mid s_0 = s,\ \pi\right] = R(s, \pi(s)) + \gamma \sum_{s'} P(s')\, \underbrace{\mathbb{E}\left[\sum_{t=1}^{\infty} \gamma^{t-1} r_t \mid s_1 = s',\ \pi\right]}_{V^{\pi}(s')}$$

We can define

$$\boxed{V_{k+1}^{\pi}(s)} = T^{\pi}\left[V_k^{\pi}\right](s) = R(s, \pi(s)) + \gamma \sum_{s'} P(s' \mid s, \pi(s)) \boxed{V_k^{\pi}(s')}$$

Classical RL: The Fundamental Theorem

3. *Fixed Policy Bellman Operator* performs **Policy Evaluation**.

Because we have

$$V^{\pi}(s) := \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^{t} r_t \mid s_0 = s,\ \pi\right] = R(s, \pi(s)) + \gamma \sum_{s'} P(s')\, \underbrace{\mathbb{E}\left[\sum_{t=1}^{\infty} \gamma^{t-1} r_t \mid\ s_1 = s',\ \pi\right]}_{V^{\pi}(s')}$$

We can define

$$V_{k+1}^{\pi}(s) = T^{\pi}\left[V_k^{\pi}\right](s) = R(s, \pi(s)) + \gamma \sum_{s'} P\big(s' \mid s, \pi(s)\big) V_k^{\pi}(s')$$

and it follows that

$$\|V^{\pi} - V_k^{\pi}\|_{\infty} \le \boxed{\gamma^{k} \|V^{\pi} - V_0^{\pi}\|_{\infty}} \qquad \boxed{\lim_{k \to \infty} V_k^{\pi}(s) = V^{\pi}(s)}$$

Classical RL: The Fundamental Theorem

4. *Optimality Bellman Operator* finds **optimal policy**.

   Define the average discounted returned at state s *of the* ***optimal policy***:

$$V^{\star}(s) := \boxed{\max_{\pi}} \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t r_t \mid s_0 = s,\, \pi\right]$$

Classical RL: The Fundamental Theorem

4. *Optimality Bellman Operator* finds **optimal policy**.

   Define the average discounted returned at state s *of the* ***optimal policy***:

$$V^{\star}(s) := \max_{\pi} \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^{t} r_t \mid s_0 = s,\ \pi\right]$$

   Unrolling the first term, we have

$$V^{\star}(s) := \max_{\pi} \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^{t} r_t \mid s_0 = s,\ \pi\right] = \boxed{\max_{a}}\left[R(s,a) + \gamma \sum_{s'} P(s') \boxed{\max_{\pi}} \mathbb{E}\left[\sum_{t=1}^{\infty} \gamma^{t-1} r_t \mid s_1 = s', \pi\right]\right]$$

Classical RL: The Fundamental Theorem

4. *Optimality Bellman Operator* finds **optimal policy**.

   Define the average discounted returned at state s *of the **optimal policy***:

$$V^\star(s) := \max_\pi \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t r_t \mid s_0 = s,\ \pi\right]$$

   Unrolling the first term, we have

$$\boxed{V^\star(s)} := \max_\pi \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t r_t \mid s_0 = s,\ \pi\right] = \max_a \left[R(s,a) + \gamma \sum_{s'} P(s') \underbrace{\max_\pi \mathbb{E}\left[\sum_{t=1}^{\infty} \gamma^{t-1} r_t \mid s_1 = s', \pi\right]}_{\boxed{V^\star(s')}}\right]$$

Classical RL: The Fundamental Theorem

4. *Optimality Bellman Operator* finds **optimal policy**.

   Because we have

$$V^\star(s) := \max_\pi \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t r_t \mid s_0 = s,\ \pi\right] = \max_a \left[R(s,a) + \gamma \sum_{s'} P(s') \underbrace{\max_\pi \mathbb{E}\left[\sum_{t=1}^{\infty} \gamma^{t-1} r_t \mid s_1 = s', \pi\right]}_{V^\star(s')}\right]$$

   We can define

$$\boxed{V_{k+1}^\star(s)} = T\left[V_k^\star\right](s) = \max_a \left[R(s,a) + \gamma \sum_{s'} P(s' \mid s, a) \boxed{V_k^\star(s')}\right]$$

Classical RL: The Fundamental Theorem

4. *Optimality Bellman Operator* finds **optimal policy**.

   Because we have

$$V^\star(s) := \max_\pi \mathbb{E}\left[\sum_{t=0}^{\infty}\gamma^t r_t \mid s_0 = s,\, \pi\right] = \max_a\left[R(s,a) + \gamma \sum_{s'} P(s') \underbrace{\max_\pi \mathbb{E}\left[\sum_{t=1}^{\infty}\gamma^{t-1} r_t \mid s_1 = s', \pi\right]}_{V^\star(s')}\right]$$

   We can define

$$V_{k+1}^\star(s) = T\left[V_k^\star\right](s) = \max_a\left[R(s,a) + \gamma \sum_{s'} P(s' \mid s,a) V_k^\star(s')\right]$$

   and it follows that

$$\|V^\star - V_k^\star\|_\infty \leq \boxed{\gamma^k \|V^\star - V_0^\star\|_\infty} \qquad \boxed{\lim_{k\to\infty} V_k^\star(s) = V^\star(s)}$$

Classical RL: The Fundamental Theorem

Recap:

1. A deterministic Markovian policy is optimal.

2. *Fixed Policy Bellman Operator* performs **Policy Evaluation**.

3. *Optimality Bellman Operator* finds **optimal policy**.

We can compute optimal policy *efficiently* using DP!

A Fundamental Theorem for CVaR RL?

Can we cast the CVaR RL problem

$$\pi^\star = \arg\max_\pi \mathrm{CVaR}_\alpha[Z^\pi]$$

as a Dynamic Program that can be solved efficiently?

A Fundamental Theorem for CVaR RL?

Proof components:

1. A deterministic Markovian policy is optimal.

Classical RL: $$\exists\, \pi_m : S \rightarrow A \text{ with } \mathbb{E}\left[Z^{\pi_m}\right] = \max_{\pi} \mathbb{E}\left[Z^{\pi}\right]$$

CVaR RL: This does not hold in general! [Artzner et al., 1999]

- Intuition: Are you betting yesterday's profit or today's lunch money?

A Fundamental Theorem for CVaR RL?

Can augment the MDP with enough info to have Markovian optimality!

Two options (based on CVaR reformulations):

A Fundamental Theorem for CVaR RL?

Can augment the MDP with enough info to have Markovian optimality!

Two options (based on CVaR reformulations):

- [Bäuerle and Ott, 2011]: Leverage *primal* formulation of CVaR
    - Add reward *floor* “c” to state.
    - “c” is unbounded real
    - “c0” is optimized

A Fundamental Theorem for CVaR RL?

Can augment the MDP with enough info to have Markovian optimality!

Two options (based on CVaR reformulations):

- [Bäuerle and Ott, 2011]: Leverage *primal* formulation of CVaR
  - Add reward *floor* “c” to state.
  - “c” is unbounded real
  - “c0” is optimized
- [Chow et al., 2015]: Leverage *dual* formulation of CVaR
  - Add confidence level “y” to state.
  - “y” is bounded between 0 and 1!
  - Requires optimizing dual variables.

Dual Time Decomposition of CVaR

[Pflug and Pichler, 2016]

CVaR at time t can be decomposed as a function of CVaR at time $t + 1$:

$$\boxed{\mathrm{CVaR}_\alpha\left[\sum_{t=0}^{\infty}\gamma^t r_t \mid s_0 = s,\, \pi\right]} = R(s,\, a) + \gamma \min_{\xi \in \Xi_\alpha(P,\, s,\, a)} \sum_{s'} P(s')\xi(s') \boxed{\mathrm{CVaR}_{\alpha\xi(s')}\left[\sum_{t=1}^{\infty}\gamma^{t-1} r_t \mid s_1 = s',\, \pi\right]}$$

$$\Xi_\alpha(P,\, s,\, a) := \left\{\xi : S \to \left[0,\, \frac{1}{\alpha}\right] \,:\, \sum_{s'} P(s' \mid s,\, a)\xi(s') = 1\right\}$$

Dual Time Decomposition of CVaR

[Pflug and Pichler, 2016]

CVaR at time t can be decomposed as a function of CVaR at time $t+1$:

$$\mathrm{CVaR}_\alpha\left[\sum_{t=0}^{\infty}\gamma^t r_t \mid s_0 = s,\, \pi\right] = R(s,\, a) + \gamma \boxed{\min_{\xi \in \Xi_\alpha(P,\, s,\, a)}} \sum_{s'} P(s')\xi(s')\mathrm{CVaR}_{\alpha\xi(s')}\left[\sum_{t=1}^{\infty}\gamma^{t-1} r_t \mid s_1 = s',\, \pi\right]$$

$$\boxed{\Xi_\alpha(P,\, s,\, a) := \left\{\xi : S \to \left[0,\, \frac{1}{\alpha}\right] \,:\, \sum_{s'} P(s' \mid s,\, a)\xi(s') = 1\right\}}$$

- Involves an inner minimization problem on worst next states

Dual Time Decomposition of CVaR

[Pflug and Pichler, 2016]

CVaR at time t can be decomposed as a function of CVaR at time $t + 1$:

$$\mathrm{CVaR}_\alpha\left[\sum_{t=0}^{\infty}\gamma^t r_t \mid s_0 = s,\, \pi\right] = R(s,\, a) + \gamma \boxed{\min_{\xi\,\in\,\Xi_\alpha(P,\, s,\, a)}} \sum_{s'} \boxed{P(s')\,\xi(s')}\,\mathrm{CVaR}_{\boxed{\alpha\xi(s')}}\left[\sum_{t=1}^{\infty}\gamma^{t-1} r_t \mid s_1 = s',\, \pi\right]$$

$$\Xi_\alpha(P,\, s,\, a) := \left\{\xi : S \to \left[0,\, \frac{1}{\alpha}\right] \;:\; \sum_{s'} P(s' \mid s,\, a)\,\xi(s') = 1\right\}$$

- Involves an inner minimization problem on worst next states
- Dual variables change next state and next confidence level

Dual Time Decomposition of CVaR

[Pflug and Pichler, 2016]

CVaR at time *t* can be decomposed as a function of CVaR at time *t* + 1:

$$\mathrm{CVaR}_\alpha\left[\sum_{t=0}^{\infty}\gamma^t r_t \mid s_0 = s, \pi\right] = R(s, a) + \gamma \boxed{\min_{\xi \in \Xi_\alpha(P, s, a)}} \sum_{s'} \boxed{P(s')\xi(s')} \mathrm{CVaR}_{\boxed{\alpha\xi(s')}}\left[\sum_{t=1}^{\infty}\gamma^{t-1} r_t \mid s_1 = s', \pi\right]$$

$$\Xi_\alpha(P, s, a) := \left\{\xi : S \to \left[0, \frac{1}{\alpha}\right] : \sum_{s'} P(s' \mid s, a)\xi(s') = 1\right\}$$

- Involves an inner minimization problem on worst next states
- Dual variables change next state and next confidence level
- Our interpretation: dual variables are *calibrated* Murphy's Law!
  - Increases likelihood of adverse events

A Fundamental Theorem for CVaR RL?

Can augment the MDP with enough info to have Markovian optimality!

- [Chow et al., 2015]: Leverage *dual* formulation of CVaR
  - Add confidence level "y" to state.
  - "y" is bounded between 0 and 1!
  - Requires optimizing dual variables.

$$a_t = \pi\left(s_t, \boxed{y_t}\right)$$

$$s_{t+1} \sim P\left(\cdot \mid s_t, a_t\right) \boxed{\xi\left(\cdot \mid s_t, a_t, y_t\right)}$$

$$\boxed{y_{t+1} = y_t \xi\left(s_{t+1} \mid s_t, a_t, y_t\right)}$$

A Fundamental Theorem for CVaR RL?

Can augment the MDP with enough info to have Markovian optimality!

- [Chow et al., 2015]: Leverage *dual* formulation of CVaR
  - Add confidence level "y" to state.
  - "y" is bounded between 0 and 1!
  - Requires optimizing dual variables.

$$a_t = \pi(s_t, \boxed{y_t})$$
$$s_{t+1} \sim P(\cdot \mid s_t, a_t)\boxed{\xi(\cdot \mid s_t, a_t, y_t)}$$
$$\boxed{y_{t+1} = y_t \xi(s_{t+1} \mid s_t, a_t, y_t)}$$

1. A deterministic Markovian policy is optimal. ✅

A Fundamental Theorem for CVaR RL?

2. *Fixed Policy Bellman Operator* performs **Policy Evaluation**.

   Because we have

$$\underbrace{\mathrm{CVaR}_{\alpha}\left[\sum_{t=0}^{\infty}\gamma^{t}r_{t} \mid s_0 = s,\, \pi\right]}_{\boxed{V^{\pi}(s,\alpha)}} = R(s,\, a) + \gamma \min_{\xi \in \Xi_{\alpha}(P,\, s,\, a)} \sum_{s'} P(s')\xi(s') \underbrace{\mathrm{CVaR}_{\alpha\xi(s')}\left[\sum_{t=1}^{\infty}\gamma^{t-1}r_{t} \mid s_1 = s',\, \pi\right]}_{\boxed{V^{\pi}(s',\alpha\xi(s'))}}$$

A Fundamental Theorem for CVaR RL?

2. *Fixed Policy Bellman Operator* performs **Policy Evaluation**.

Because we have

$$\underbrace{\mathrm{CVaR}_\alpha\left[\sum_{t=0}^{\infty}\gamma^t r_t \mid s_0 = s,\, \pi\right]}_{V^\pi(s,\alpha)} = R(s,\, a) + \gamma \min_{\xi \in \Xi_\alpha(P,\, s,\, a)} \sum_{s'} P(s')\xi(s')\underbrace{\mathrm{CVaR}_{\alpha\xi(s')}\left[\sum_{t=1}^{\infty}\gamma^{t-1} r_t \mid s_1 = s',\, \pi\right]}_{V^\pi(s',\alpha\xi(s'))}$$

We can define

$$\boxed{V_{k+1}^\pi(s, y)} = T^\pi\left[V_k^\pi\right](s, y) = R(s, a) + \gamma \min_{\xi\in\Xi_\alpha(P,\, s,\, a)} \sum_{s'} P(s')\xi(s')\boxed{V_k^\pi\left(s', y\xi(s')\right)}$$

and it follows that

$$\|V^\pi - V_k^\pi\|_\infty \le \gamma^k \|V^\pi - V_0^\pi\|_\infty \qquad \lim_{k\to\infty} V_k^\pi(s) = V^\pi(s)$$

A Fundamental Theorem for CVaR RL?

2. *Fixed Policy Bellman Operator* performs **Policy Evaluation**. ✅

Because we have

$$\underbrace{\mathrm{CVaR}_\alpha\left[\sum_{t=0}^{\infty}\gamma^t r_t \mid s_0 = s,\ \pi\right]}_{V^\pi(s,\alpha)} = R(s,\ a) + \gamma \min_{\xi \in \Xi_\alpha(P,\ s,\ a)} \sum_{s'} P(s')\xi(s')\underbrace{\mathrm{CVaR}_{\alpha\xi(s')}\left[\sum_{t=1}^{\infty}\gamma^{t-1} r_t \mid s_1 = s',\ \pi\right]}_{V^\pi(s',\alpha\xi(s'))}$$

We can define

$$\boxed{V_{k+1}^\pi(s,y)} = T^\pi\left[V_k^\pi\right](s,y) = R(s,a) + \gamma \min_{\xi\in\Xi_\alpha(P,\ s,\ a)} \sum_{s'} P(s')\xi(s')\boxed{V_k^\pi(s',y\xi(s'))}$$

and it follows that

$$\|V^\pi - V_k^\pi\|_\infty \le \gamma^k \|V^\pi - V_0^\pi\|_\infty \qquad \lim_{k\to\infty} V_k^\pi(s) = V^\pi(s)$$

A Fundamental Theorem for CVaR RL?

3. *Optimality Bellman Operator* finds **optimal policy**.

   Taking the maximum on both sides of CVaR Decomposition gives:

$$\max_{\pi} \mathrm{CVaR}_{\alpha}\left[\sum_{t=0}^{\infty} \gamma^{t} r_{t} \mid s_{0}=s, \pi\right]=\max_{\pi}\left[R(s, \pi(s, y))+\gamma \min_{\xi \in \Xi_{\alpha}(P, s, \pi(s, y))} \sum_{s^{\prime}} P\left(s^{\prime}\right) \xi\left(s^{\prime}\right) \mathrm{CVaR}_{\alpha \xi\left(s^{\prime}\right)}\left[\sum_{t=1}^{\infty} \gamma^{t-1} r_{t} \mid s_{1}=s^{\prime}, \pi\right]\right]$$

   We want to show that we have

$$\max_{\pi} \mathrm{CVaR}_{\alpha}\left[\sum_{t=0}^{\infty} \gamma^{t} r_{t} \mid s_{0}=s, \pi\right]=\boxed{\max_{a}}\left[R(s, a)+\gamma \min_{\xi \in \Xi_{\alpha}(P, s, a)} \sum_{s^{\prime}} P\left(s^{\prime}\right) \xi\left(s^{\prime}\right) \boxed{\max_{\pi}} \mathrm{CVaR}_{\alpha \xi\left(s^{\prime}\right)}\left[\sum_{t=1}^{\infty} \gamma^{t-1} r_{t} \mid s_{1}=s^{\prime}, \pi\right]\right]$$

A Fundamental Theorem for CVaR RL?

3. *Optimality Bellman Operator* finds **optimal policy**. ✕

Taking the maximum on both sides of CVaR Decomposition gives:

$$\max_{\pi} \mathrm{CVaR}_{\alpha}\left[\sum_{t=0}^{\infty} \gamma^{t} r_{t} \mid s_{0}=s,\ \pi\right]=\max_{\pi}\left[R(s, \pi(s, y))+\gamma \min_{\xi \in \Xi_{\alpha}(P, s, \pi(s, y))} \sum_{s^{\prime}} P\left(s^{\prime}\right) \xi\left(s^{\prime}\right) \mathrm{CVaR}_{\alpha \xi\left(s^{\prime}\right)}\left[\sum_{t=1}^{\infty} \gamma^{t-1} r_{t} \mid s_{1}=s^{\prime}, \pi\right]\right]$$

We want to show that we have

$$\max_{\pi} \mathrm{CVaR}_{\alpha}\left[\sum_{t=0}^{\infty} \gamma^{t} r_{t} \mid s_{0}=s,\ \pi\right] \boxed{\leq} \max_{a}\left[R(s, a)+\gamma \min_{\xi \in \Xi_{\alpha}(P, s, a)} \sum_{s^{\prime}} P\left(s^{\prime}\right) \xi\left(s^{\prime}\right) \max_{\pi} \mathrm{CVaR}_{\alpha \xi\left(s^{\prime}\right)}\left[\sum_{t=1}^{\infty} \gamma^{t-1} r_{t} \mid s_{1}=s^{\prime}, \pi\right]\right]$$

Only lower bound in general! [Hau et al., 2024]

Impact of Hau et al.'s result

$$\max_{\pi} \mathrm{CVaR}_{\alpha}\left[\sum_{t=0}^{\infty} \gamma^{t} r_{t} \mid s_{0}=s, \pi\right] \boxed{\leq} \max_{a}\left[R(s, a)+\gamma \min_{\xi \in \Xi_{\alpha}(P, s, a)} \sum_{s'} P\left(s'\right) \xi\left(s'\right) \max_{\pi} \mathrm{CVaR}_{\alpha \xi\left(s'\right)}\left[\sum_{t=1}^{\infty} \gamma^{t-1} r_{t} \mid s_{1}=s', \pi\right]\right]$$

- The Optimality bellman operator does not hold for dual CVaR:
  - Chow et al. and numerous works building on it are refuted.
  - Convergence does occur, but to an overestimation of the CVaR of the policy found.
  - Suboptimality gap can be made arbitrarily large: no quick fix!

Impact of Hau et al.'s result

$$\max_{\pi} \mathrm{CVaR}_{\alpha}\left[\sum_{t=0}^{\infty} \gamma^{t} r_{t} \mid s_{0}=s,\, \pi\right] \boxed{\leq} \max_{a}\left[R(s, a)+\gamma \min_{\xi \in \Xi_{\alpha}(P, s, a)} \sum_{s'} P\left(s'\right) \xi\left(s'\right) \max_{\pi} \mathrm{CVaR}_{\alpha \xi\left(s'\right)}\left[\sum_{t=1}^{\infty} \gamma^{t-1} r_{t} \mid s_{1}=s', \pi\right]\right]$$

- The Optimality bellman operator does not hold for dual CVaR:
  - Chow et al. and numerous works building on it are refuted.
  - Convergence does occur, but to an overestimation of the CVaR of the policy found.
  - Suboptimality gap can be made arbitrarily large: no quick fix!

- All other properties remain:
  - Optimal policy is deterministic and Markovian on augmented MDP
  - CVaR evaluation can be cast as a Dynamic Program

Our Proposal: CVaR RL as a Game

Consider dual variables to yield an **adversarial** MDP:

- Motivated by antagonist objective (max-min structure)
- When the adversary is optimal for the policy, expected return = CVaR
- Notation abuse: adversary is set of dual variables for all (s, a, y)

Our Proposal: CVaR RL as a Game

Define the expected value of a (policy, adversary) pair:

$$V_{\xi}^{\pi}(\boldsymbol{s}, y) := \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t \boldsymbol{r}_t \mid \boldsymbol{s}_0 = \boldsymbol{s},\ \pi,\ \xi\right]$$

Our Proposal: CVaR RL as a Game

Define the expected value of a (policy, adversary) pair:

$$V_{\xi}^{\pi}(s, y) := \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^{t} r_{t} \mid s_{0}=s,\, \pi,\, \xi\right]$$

We propose to alternate between policy and adversary optimization:

Improve policy

$$\pi_{k+1}(s, y)=\arg\max_{a}\left[R(s,\, a)+\gamma \sum_{s'} P(s' \mid s,\, a) \xi(s' \mid s,\, a,\, y) V_{\xi_{k}}^{\pi_{k}}(s',\, y \xi_{k}(s' \mid s,\, a,\, y))\right]$$

Propagate new policy

$$V_{\xi_{k}}^{\pi_{k+1}}(s, y)=R(s, \pi_{k+1}(s, y))+\gamma \sum_{s'} P(s' \mid s,\, \pi_{k+1}(s, y)) \xi_{k}(s' \mid s,\, \pi_{k+1}(s, y),\, y) V_{\xi_{k}}^{\pi_{k+1}}(s',\, y \xi_{k}(s'))$$

Compute CVaR

$$V_{\xi_{k+1}}^{\pi_{k+1}}(s, y)=R(s, \pi_{k+1}(s, y))+\gamma \min_{\xi \in \Xi_{y}(P,\, s,\, \pi_{k+1})} \sum_{s'} P(s' \mid s,\, \pi_{k+1}(s, y)) \xi(s') V_{\xi_{k+1}}^{\pi_{k+1}}(s',\, y \xi(s'))$$

Compute adversary

$$\xi_{k+1}(s' \mid s, a, y)=\arg\min_{\xi \in \Xi_{y}(P,\, s,\, a)} \sum_{s'} P(s' \mid s,\, \pi_{k+1}(s, y)) \xi(s') V_{\xi_{k+1}}^{\pi_{k+1}}(s',\, y \xi(s'))$$

Convergence Analysis: Work In Progress

Intuition: next policy selects best greedy action wrt previous CVaR

- *What can go wrong should be trained on*
- Early empirical results are encouraging

Not being able to rely on simultaneous optimization complicates analysis.

- Hard to quantify if next policy has higher CVaR
- Current idea: try to prove (policy, adversary) pairs converge to equilibrium

Conclusion

The CVaR RL formulation offers an **attractive** pathway to safe RL

Finding a **CVaR** equivalent of the **Fundamental Theorem** still an open question

Proposal: **separately optimizing the policy and adversary as a game** could be a good foundation for CVaR Fundamental Theorem

References

Bellemare, Marc G., et al. "The arcade learning environment: An evaluation platform for general agents." *Journal of artificial intelligence research* 47 (2013): 253-279.

Ouyang, Long, et al. "Training language models to follow instructions with human feedback." Advances in neural information processing systems 35 (2022): 27730-27744.

Basel Committee on Banking Supervision. Minimum capital requirements for market risk. In Basel III: international regulatory framework for banks, (2019).

Bäuerle, Nicole, and Jonathan Ott. "Markov decision processes with average-value-at-risk criteria." Mathematical Methods of Operations Research 74 (2011): 361-379.

References

Chow, Yinlam, et al. "Risk-sensitive and robust decision-making: a cvar optimization approach." Advances in neural information processing systems 28 (2015).

Pflug, Georg Ch, and Alois Pichler. "Time-consistent decisions and temporal decomposition of coherent risk functionals." Mathematics of Operations Research 41.2 (2016): 682-699.

Hau, Jia Lin, et al. "On dynamic programming decompositions of static risk measures in Markov decision processes." Advances in Neural Information Processing Systems 36 (2023): 51734-51757.