Risk-averse Distributional Reinforcement Learning

A CVaR optimization approach

Silvestr Stanko[1]

[1]Department of Computer Science
Czech Technical University

Thursday 24th May, 2018

Outline

Motivation

Figure: Robotics

Figure: Finance

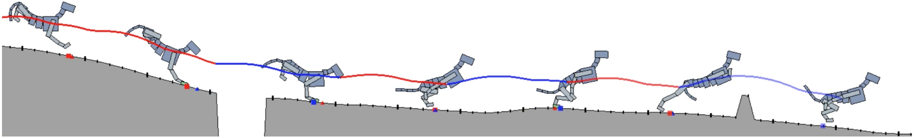


Figure: AI safety

Ultimate goals of AI

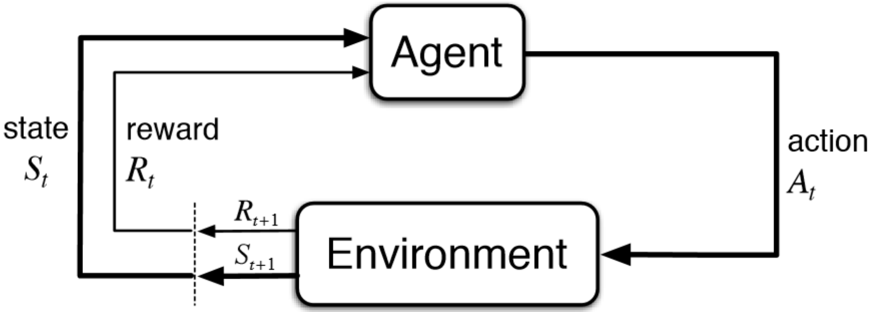
General AI

- Learning from experience
- Learning *tabula rasa*
- Beyond purpose-specific AI
- Beyond human-level performance

Safe AI

- Avoiding catastrophic events
- Robust to environment changes or adversaries

Reinforcement Learning



Recent successes

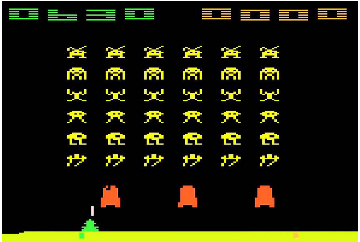


Figure: Atari games

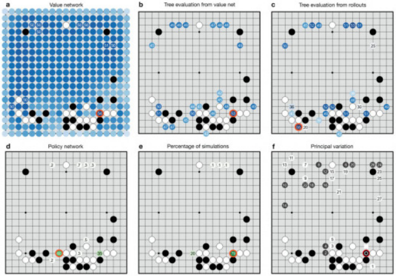


Figure: Go

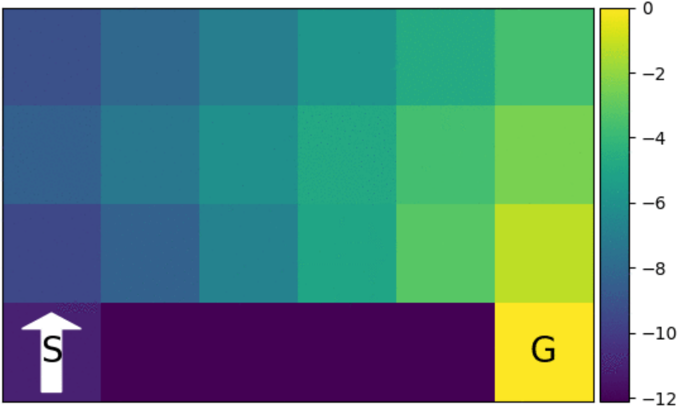
Markov Decision Processes

Definition

An MDP is a 5-tuple $\mathcal{M} = (\mathcal{X}, \mathcal{A}, R, P, \gamma)$, where

- $\mathcal{X}$ is the state space
- $\mathcal{A}$ is the action space
- $R(x, a)$ is a random variable representing the reward generated by being in state x and selecting action a
- $P(\cdot|x, a)$ is the transition probability distribution
- $\gamma \in [0, 1)$ is a discount factor

Markov Decision Process - example



Reinforcement Learning - Goal

Definition

$Z^\pi(x_t)$ Is a random variable representing the discounted reward along a trajectory generated by the MDP by following the policy π, starting at state x_t.

$$Z^\pi(x_t) = \sum_{t=0}^{\infty} \gamma^t R(x_t, \pi(x_t))$$

Reinforcement Learning goals

Our goal is to find a globally optimal policy π^*

$$\pi^* = \arg\max_\pi \exp Z^\pi(x_0)$$

Potential problems

- Solutions must avoid catastrophic events and be **safe**
- RL is sample inefficient $\rightarrow$ expensive training
- Solutions must be **robust** to small model changes

Solution

Instead of maximizing the expected reward, focus on other criteria that take into account the **risk** of the potential reward.

Risk

Definition

Risk is the potential of gaining or losing something of value.

Risk-averse: disinclined or reluctant to take risks

Risk-neutral: indifferent to or balanced with respect to risk.

Risk-seeking: inclined or eager to take risks

Example

Choose between recieving:

1. \$100 in 100% cases
2. \$200 in 50% cases and \$0 in 50% cases
3. \$10,000 in 1% cases and \$0 in 99% cases

Measuring Risk

Value-at-Risk (VaR)

- Easy to understand
- Historically the most used risk-measure
- Undesirable computational properties
- Does not differentiate between large and catastrophic losses

Definition

Let Z be a random variable representing reward, with cumulative distribution function $F(z) = \mathbb{P}(Z \leq z)$. The Value-at-Risk at confidence level $\alpha \in (0, 1)$ is the α-quantile of Z, i.e.

$$\mathrm{VaR}_\alpha(Z) = F^{-1}(\alpha) = \inf\left\{z \middle| \alpha \leq F(z)\right\}$$

Measuring Risk

Conditional Value-at-Risk (CVaR)

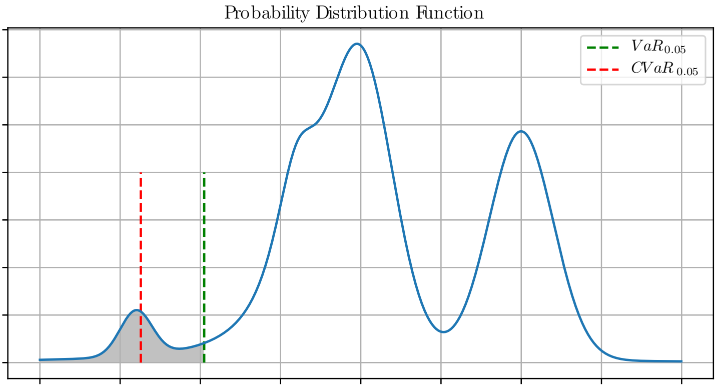
- Good computational properties
- Basel Committee on Banking Supervision: VaR $\rightarrow$ CVaR
- Equivalent to robustness

Definition

The Conditional Value-at-Risk (CVaR) at confidence level $\alpha \in (0,1)$ is defined as the expected reward of of outcomes worse than the α-quantile (VaR_α):

$$\mathrm{CVaR}_\alpha(Z) = \frac{1}{\alpha}\int_0^\alpha F_Z^{-1}(\beta)\mathrm{d}\beta = \frac{1}{\alpha}\int_0^\alpha \mathrm{VaR}_\beta(Z)\mathrm{d}\beta$$

Value-at-Risk, Conditional Value-at-Risk



Conditional Value-at-Risk as an optimal point

Definition

$$\mathrm{CVaR}_\alpha(Z) = \max_s \left\{ \frac{1}{\alpha} \mathbb{E}\left[(Z - s)^-\right] + s \right\}$$

where $(x)^- = \min(x, 0)$ and in the optimal point it holds that $s^* = VaR_\alpha(Z)$

$$\mathrm{CVaR}_\alpha(Z) = \frac{1}{\alpha} \mathbb{E}\left[(Z - VaR_\alpha(Z))^-\right] + VaR_\alpha(Z)$$

it's dual is

$$\mathrm{CVaR}_\alpha(Z) = \min_{\xi \in \mathcal{U}_{\mathrm{CVaR}}(\alpha, p(\cdot))} \mathbb{E}_\xi[Z]$$

$$\mathcal{U}_{\mathrm{CVaR}}(\alpha, p(\cdot)) = \left\{ \xi : \xi(z) \in \left[0, \frac{1}{\alpha}\right], \int \xi(z) p(z) \mathrm{d}z = 1 \right\}$$

Risk-averse Reinforcement Learning - goals

Definition

$Z^\pi(x_t)$ Is a random variable representing the discounted reward along a trajectory generated by the MDP by following the policy π, starting at state x_t.

$$Z^\pi(x_t) = \sum_{t=0}^{\infty} \gamma^t R(x_t, \pi(x_t))$$

Reinforcement Learning with CVaR

For a given α, our goal is to find a globally optimal policy π^*

$$\pi^* = \arg\max_\pi CVaR_\alpha^\pi(Z^\pi(x_0))$$

Risk-averse Reinforcement Learning - example

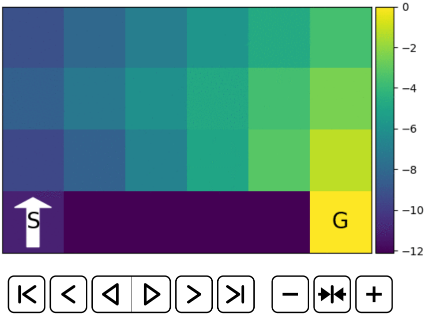


Figure: Greedy agent

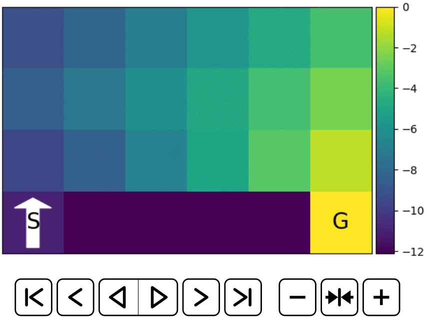


Figure: Risk-averse agent

Value Iteration

Definition

Value function $V(x)$ represents the expected return when starting in state x and following the optimal policy π^* thereafter.

Value Iteration

Initialize $V_0(x)$ for each state (arbitrary value, e.g. 0).
Update each state:

$$V_{k+1}(x) = \max_a \left[R(x,a) + \gamma \sum_{x'} p(x'|x,a) V_k(x') \right]$$

Repeat.

The algorithm converges to the optimal policy π^*: $\lim_{k\to\infty} V_k(x) = V(x)$

Value Iteration with CVaR

Theorem (CVaR decomposition)

The conditional CVaR under policy π obeys the following decomposition:

$$CVaR_\alpha\left(Z^\pi(x,a)\right) = \min_{\xi \in \mathcal{U}_{CVaR}(\alpha, p(\cdot|x,a))} \sum_{x'} p(x'|x,a)\xi(x') CVaR_{\xi(x')\alpha}\left(Z^\pi(x')\right)$$

TODO: pic of mdp with cvars

CVaR Value Iteration

Theorem (CVaR Value Iteration)

The following Bellman operator is a contraction:

$$\mathbf{T}C(x,y) = \max_a \left[R(x,a) + \gamma \min_\xi \sum_{x'} p(x'|x,a)\xi(x')C\left(x', y\xi(x')\right) \right]$$

The operator $\mathbf{T}$ describes the following relationship:

$$\mathbf{T}CVaR_y(Z(x)) = \max_a \left[R(x,a) + \gamma CVaR_y(Z(x')) \right]$$
$$x' \sim p(\cdot|x,a)$$

Linear interpolation

Computing operator $\mathbf{T}$ s intractable, as the state-space is continuous. A solution would be to approximate the operator with linear interpolation.

Theorem

The function $\alpha CVaR_\alpha$ is convex. The operator $\mathbf{T}_\mathcal{I}$ is a contraction.

$$\mathcal{I}_x[C](y) = y_i C(x, y_i) + \frac{y_{i+1} C(x, y_{i+1}) - y_i C(x, y_i)}{y_{i+1} - y_i}(y - y_i)$$

$$\mathbf{T}_\mathcal{I} C(x, y) = \max_a \left[R(x, a) + \gamma \min_\xi \sum_{x'} p(x'|x, a) \frac{\mathcal{I}_{x'}[C](y\xi(x'))}{y} \right]$$

This iteration can be formulated and solved as a linear program.

TODO: pic of cvar alpha

Original Contributions

1. **Faster CVaR Value Iteration**
   - Polynomial $\rightarrow$ linear time.
   - Formally proved for increasing, unbounded distributions.
   - Experimentally verified for general distributions.
2. **CVaR Q-learning**
   - Sampling version of CVaR Value Iteration.
   - Based on the distributional approach.
   - Experimentally verified.
3. **Distributional Policy improvement**
   - Proved monotonic improvement for distributional RL.
   - Used as a heuristic for extracting π^* from CVaR Q-learning.
4. **Deep CVaR Q-learning**
   - TD update $\rightarrow$ loss function.
   - Experimentally verified in a deep learning context.

αCVaR$_\alpha$ describes a quantile function

Lemma

Any discrete distribution has a piece-wise linear $\alpha CVaR_\alpha$ function. Similarly, any a piece-wise linear $\alpha CVaR_\alpha$ function can be seen as representing a certain discrete distribution.

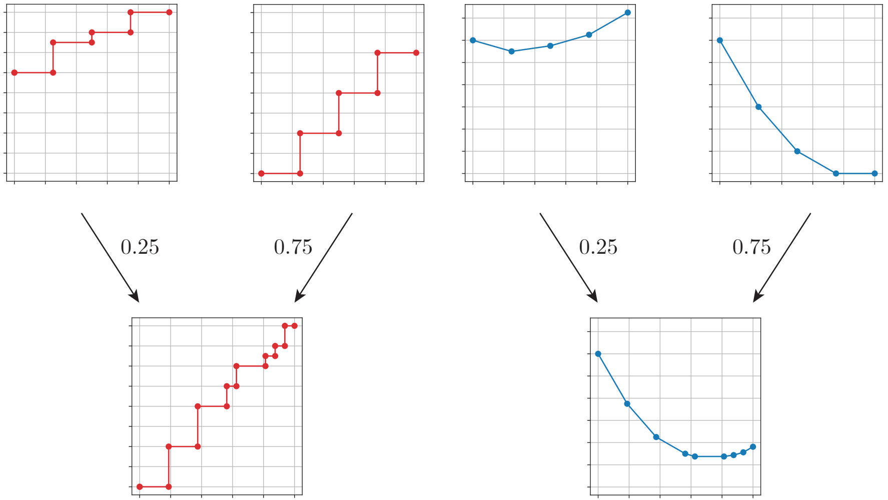
αCVaR$_\alpha$ $\Leftarrow$ VaR

$$\frac{\partial}{\partial\alpha}\alpha\text{CVaR}_\alpha(Z) = \frac{\partial}{\partial\alpha}\int_0^\alpha VaR_\beta(Z)d\beta = VaR_\alpha(Z)$$

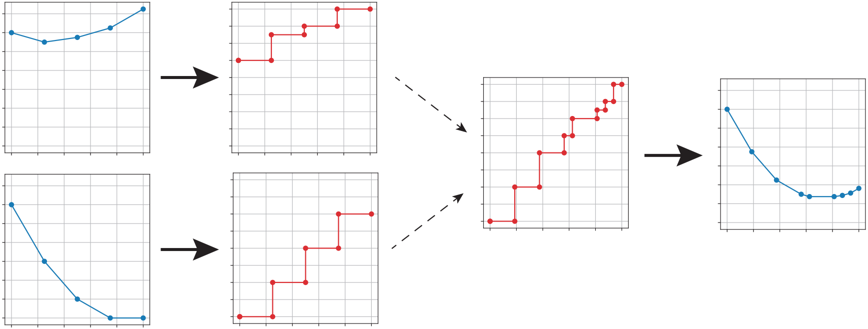
αCVaR$_\alpha$ $\Rightarrow$ VaR

$$\alpha\text{CVaR}_\alpha(Z) = \int_0^\alpha VaR_\beta(Z)\text{d}\beta$$

Next state CVaR computation



Next state CVaR computation



Linear-time Computation

Theorem

Solution to minimization problem

$$\min_{\xi \in \mathcal{U}_{CVaR}(\alpha, p(\cdot|x,a))} \sum_{x'} p(x'|x,a)\xi(x') CVaR_{\xi(x')\alpha}\left(Z^{\pi}(x')\right)$$

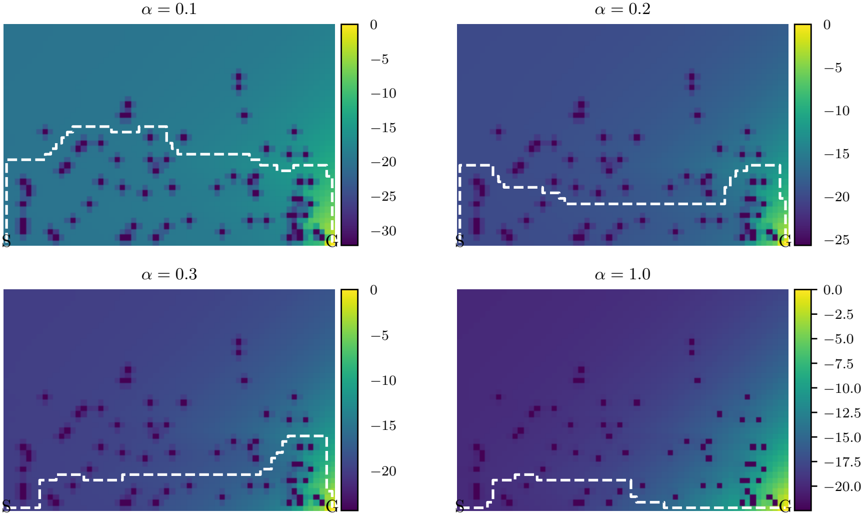
can be computed by setting

$$\xi(x') = \frac{F_{Z(x')}(F^{-1}_{Z(x,a)}(\alpha))}{\alpha}$$

The computational complexity is $O(n \cdot m)$ where n is the number of transition states and m is the number of atoms.

- Proved for increasing unbounded distributions

CVaR Value Iteration - Experiments



Optimal policy extraction

TODO: policy extraction is problematic

VaR-based Policy Improvement

TODO:visual

Theorem

Let π be a fixed policy, $\alpha \in (0, 1]$. By following policy π' from the following algorithm, we will improve $CVaR_\alpha(Z)$ in expectation:

$$CVaR_\alpha(Z^\pi) \leq CVaR_\alpha(Z^{\pi'})$$

input α, x_0, γ
$a = \arg\max_a CVaR_\alpha(Z(x_0, a))$
$s = VaR_\alpha(Z(x_0, a))$
$x_t, r_t = \text{envTransition}(x_0, a)$
while x_t is not terminal **do**
$\quad s = \frac{s - r_t}{\gamma}$
$\quad a = \arg\max_a \mathbb{E}\left[(Z(x_t, a) - s)^-\right]$
$\quad x_t, r_t = \text{envTransition}(x_t, a)$

TODO

- CVaR Q-learning
  - (?) Use Wasserstein distance with quantile improvement
  - (?) Extend the VaR-based algorithm
  - (?) Combine with quantile regression
- Experiments
  - Value Iteration + Q-learning
  - Deep Q-learning