Risk Aware Reinforcement Learning

Theory and Algorithms

Tosatto Samuele

Intelligent Autonomous Systems
Technische Universität Darmstadt

05 August 2018

Outline

Motivation

What Is our Objective ?

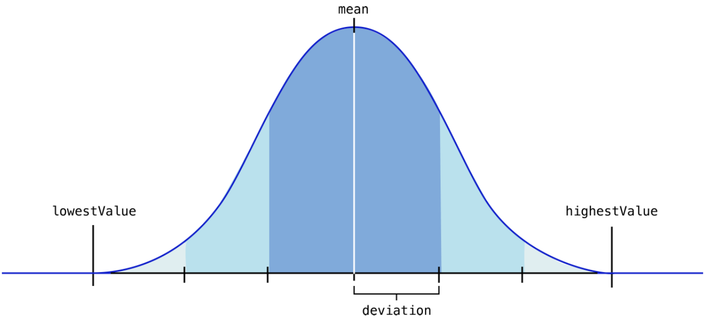


FIGURE – A Gaussian distribution. On the tail rare events might occur..

Rare Events - Douglas Adams

Extremely rare event in Doublas Adams' opinion..

FIGURE – "Oh no, not again..." (Douglas Adams)

Rare Events - Nassim Taleb

Extremely rare event in Nassim Nicholas Taleb opinion..

FIGURE – “Oh no, not again...” (Douglas Adams)

Catastrophic Events

Sometimes rare events might be catastrophic

Catastrophic Events

Sometimes rare events might be catastrophic

Catastrophic Events

Sometimes rare events might be catastrophic

Catastrophic Events

Sometimes rare events might be catastrophic

Avoiding Catastrophes

Of course, we would like to build autonomous systems **sensible to the risk**..
This is important in diverse fields :

Fields

- Finance
- Smart grids
- Health
- Robotics

Not only the average case

Very often people optimize the **average case**.

$$\max_{\theta} \mathbb{E} X(\theta)$$

Not only the average case

Very often people optimize the **average case**.

$$\max_{\theta} \mathbb{E} X(\theta)$$

In a risk aware setting we are interested in the distribution of things.

Risk in Reinforcement Learning

Common Optimization problem in classical RL :

$$\max_{\pi} \mathbb{E} J(\pi)$$

But it is not all just about rewards..

Risk in Reinforcement Learning

Common Optimization problem in classical RL :

$$\max_{\pi} \mathbb{E} J(\pi)$$

But it is not all just about rewards..

- Distribution of the return

Risk in Reinforcement Learning

Common Optimization problem in classical RL :

$$\max_{\pi} \mathbb{E} J(\pi)$$

But it is not all just about rewards..

- Distribution of the return
- Ergodicity

Risk in Reinforcement Learning

Common Optimization problem in classical RL :

$$\max_{\pi} \mathbb{E} J(\pi)$$

But it is not all just about rewards..

- Distribution of the return
- Ergodicity
- Probability of catastrophic states

Risk in Reinforcement Learning

Common Optimization problem in classical RL :

$$\max_{\pi} \mathbb{E} J(\pi)$$

But it is not all just about rewards..

- Distribution of the return
- Ergodicity
- Probability of catastrophic states
- Uncertainty about the model
- ...

Taxonomy

We can divide Risk-Aware RL in two main categories [1] :

Optimization Criterion

- Worst Case
- Risk Sensitive
- Constrained

Exploration Process

- External Knowledge
  - Initial Knowledge
  - Policy from Demonstration
  - Ask for Help
  - Teacher Provide Advices
- Risk Directed

1. **garcia2015comprehensive**.

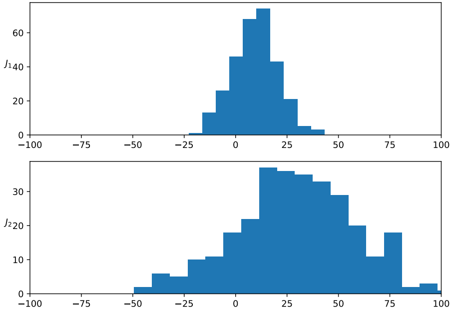
Quick Reinforcement Learning Reminder

Worst Case Criterion

Worst Case Criterion

We want to maximizes the minimum possible expected return :

$$\max_{\pi} \min\{J(\pi)\}$$



$\hat{Q}$-Learning (Heger 1994)

Direct minimization of the worst case :

$$Q(s,a) := \min\{Q(s,a), r + \gamma \max_{a'} Q(s', a')\}$$

The idea is to maintain the memory of the worst sample observed. Note that no learning rate is required.

- Too pessimistic ;
- requests an optimistic initialization of the Q.

β-pessimistic Q-Learning (Gaskett 2003)

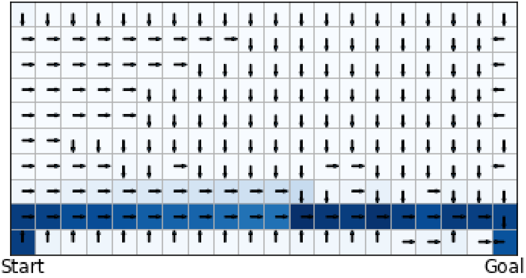
We want to mitigate the strong pessimism of $\hat{Q}$-Learning

$$Q(s,a) = (1-\alpha)Q(s,a) + \alpha\left(r + (1-\beta)\max_{a'} Q(s',a') + \beta \min_{a'} Q(s',a')\right)$$

This method does not truly optimize the worst case criterion, but it works in practice better.

The Cliff Environment

- Gridworld 20x10
- Each step, reward -1
- On the bottom, a cliff. End of episode and reward = -100.
- Hitting the walls : -100.



Exercize 1

Exercize 1 - Risk Aware Q-Learning

1. Open “ex1” with Jupyter
2. Fill out the update rule with a risk update and try it out !
3. Try $\beta = 0.05, 0.1, 1.5$. What happens with $\hat{Q}$-Learning ?

Reminder :

1. $\hat{Q}$-Learning :

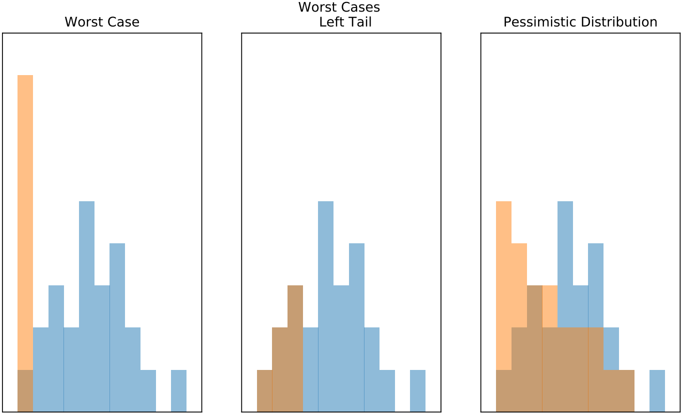
$$Q(s,a) := \min\{Q(s,a), r + \gamma \max_{a'} Q(s', a')\}$$

2. β-pessimistic Q-Learning :

$$Q(s,a) = (1-\alpha)Q(s,a) + \alpha\Big(r + (1-\beta)\max_{a'} Q(s',a') + \beta \min_{a'} Q(s',a')\Big)$$

Risk Sensitivity

Worst Case vs Better Risk Metric



Distribution

- Set of event I, i.e. $\{\mathrm{head}, \mathrm{tail}\}$
- Set of values $X(i) \in \mathbb{R}$
- Set of probability for each event $\mu(i)$

Valuation Function

A valuation function is a mapping between distributions and real values, such that :

1. Is monotonic : $\rho(X, \mu) \leq \rho(Y, \mu)$ whenever $X(i) \leq Y(i) \forall i \in I$;
2. is invariant w.r.t. the translation : $\rho(X + y\mathbf{1}, \mu) = y + \rho(X, \mu)$.

moreover, if the valuation function is concave, e.g.,

$$\rho(\alpha X + (1 - \alpha)Y, \mu) > \alpha\rho(X, \mu) + (1 - \alpha)\rho(Y, \mu)$$

then ρ is risk averse.

Entropic Mapping

A notable valuation function is the entropic mapping :

$$\rho_\eta(X, \mu) = \frac{1}{\eta} \log \sum_i \mu(i) e^{\eta X(i)}$$

Entropic Mapping

A notable valuation function is the entropic mapping :

$$\rho_\eta(X, \mu) = \frac{1}{\eta} \log \sum_i \mu(i) e^{\eta X(i)}$$

It is very interesting to note that the entropic mapping is the solution to the problem :

$$\rho_\eta(X, \mu) = \min_q \sum_i X(i) q(i) + \frac{1}{\eta} KL(q || \mu)$$

Entropic Mapping

A notable valuation function is the entropic mapping :

$$\rho_\eta(X,\mu) = \frac{1}{\eta}\log\sum_i \mu(i)e^{\eta X(i)}$$

It is very interesting to note that the entropic mapping is the solution to the problem :

$$\rho_\eta(X,\mu) = \min_q \sum_i X(i)q(i) + \frac{1}{\eta}KL(q||\mu)$$

1. $\eta \to -\infty$ we have min operator

Entropic Mapping

A notable valuation function is the entropic mapping :

$$\rho_\eta(X, \mu) = \frac{1}{\eta} \log \sum_i \mu(i) e^{\eta X(i)}$$

It is very interesting to note that the entropic mapping is the solution to the problem :

$$\rho_\eta(X, \mu) = \min_q \sum_i X(i)q(i) + \frac{1}{\eta} KL(q||\mu)$$

1. $\eta \to -\infty$ we have min operator
2. $\eta \to 0$ we have the average $\mathbb{E}$

Entropic Mapping

A notable valuation function is the entropic mapping :

$$\rho_\eta(X, \mu) = \frac{1}{\eta} \log \sum_i \mu(i) e^{\eta X(i)}$$

It is very interesting to note that the entropic mapping is the solution to the problem :

$$\rho_\eta(X, \mu) = \min_q \sum_i X(i) q(i) + \frac{1}{\eta} KL(q || \mu)$$

1. $\eta \to -\infty$ we have $\min$ operator
2. $\eta \to 0$ we have the average $\mathbb{E}$
3. $\eta \to +\infty$ we have $\max$ operator

Exercize 2

Exercize 2 - Entropic Map

1. Open “ex2” with Jupyter
2. Fill out the entropic map function (you can use either `np.exp` and `np.log` or from `scipy.special` the `logsumexp`, where b is the parameter weighting the summation)
3. Run the script, and observe how the distribution changes. We can notice that the entropic map is equivalent to the definition of the optimization problem defined.

Reminder :

$$\rho_\eta(X, \mu) = \frac{1}{\eta} \log \sum_i \mu(i) e^{\eta X(i)}$$

Utility Shortfall

Let's assume $u : \mathbb{R} \to \mathbb{R}$ a continuous and strictly increasing function. The

$$\rho^u_{x_0}(X, \mu) := \sup\Big\{m \in \mathbb{R} \Big| \sum_i u(X(i) - m) \geq x_0\Big\}$$

is a shortfall induced by u with **acceptance level** x_0. It is possible to show that

- ρ is a proper valuation function (cite Föllmer and Schied 2004) ;
- if $u(x) = x$ and $x_0 = 0$ we have the expected value ;
- $u(x)$ being concave determines risk-adversity, or risk-seeking in the opposite case ;
- $u(x) = e^{\eta x}$ and $x_0 = 1$ determines the entropic map.

Risk Aware Q-Learning (Shen et al, 2014)

We want to solve the risk aware bellman equation :

$$Q^*(s,a) = \mathcal{U}\left(R(s,a) + \gamma \sum_{s'} P(s'|s,a) \max_{a'} Q^*(s',a')\right)$$

where $\mathcal{U}$ is a valuation function (i.e. entropic map).

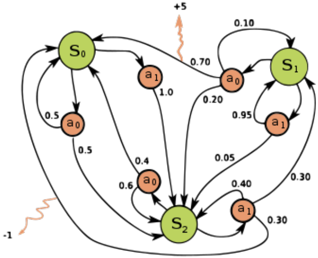
If $\mathcal{U}$ is generated by the utility-based short-fall with utility u and acceptance level x_0, then the correspondend Q-Learning will have update formula

$$Q(s,a) := Q(s,a) + \alpha\left[u\left(r + \gamma \sum_{s'} P(s'|s,a) \max_{a'} Q^*(s',a') - Q(s,a)\right) - x_0\right]$$

Robust Markov Decision Process

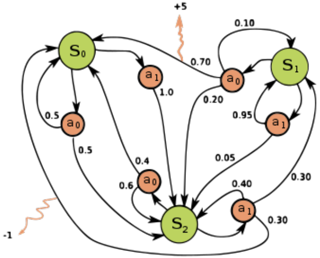
Markov Decision Process

- Set of states $\mathcal{S}$
- Set of actions $\mathcal{A}$
- Transition probability $\mathcal{P}$
- Reward function $\mathcal{R}$
- Initial distribution μ



Robust Markov Decision Process

- Set of states $\mathcal{S}$
- Set of actions $\mathcal{A}$
- Set of transition probabilities $\mathcal{P}$
- Reward function $\mathcal{R}$
- Initial distribution μ



Robust Value Iteration

Let's define the Robust Bellman Equation

$$V^*(s) = \max_a r(s,a) + \gamma \inf_{P \in \mathcal{P}} \mathbb{E}_{s' \sim P(s,a)} V^*(s')$$

For convenience, let's use a vector notation,

$$\sigma_{\mathcal{P}(s,a)} V := \inf\{P^t V : P \in \mathcal{P}(s,a)\}$$

and the Bellman Operator T^* as

$$T^* V := \max_\pi r^\pi + \gamma \sigma_{\mathcal{P},\pi} V.$$

It is possible to show that T^* is a contraction, and V^* is its unique fixed point.

Robust Least Square Policy Iteration

- Let D be a positive diagonal matrix.
- $V(s)$ is encoded as $\phi(s)\omega$

Classic LSPI :

$$\omega_{k+1} = (\phi^T D\phi)^{-1}\left(\phi^T Dr + \gamma D\phi\omega_k\right)$$

Robust LSPI (under some assumptions) :

$$\omega_{k+1} = (\phi^T D\phi)^{-1}\left(\phi^T Dr + \gamma D\sigma_\pi\phi\omega_k\right)$$

But how to solve σ_π ?

The Inner Problem

How do we solve

$$\inf_{p \in \mathcal{P}(s,a)} \sum_{s'} p(s')\phi(s')\omega_k?$$

It much depends about how we define the set $\mathcal{P}(s, a)$.

- $\mathcal{P}(s, a = \{p : \mathrm{Dist}(p, \hat{p}) \leq \epsilon, p^T\mathbf{1} = 1, p \geq 0\})$
  - L_1 Distance : (Strehl and Littman 2005)
  - KL Distance : (Iyengar 2005) and (Nilim and El Ghaoui 2005)
- Interval or ellipsoidal models
- For parametric p_θ we can use policy gradient $\nabla_{\theta \mid \theta}[\phi(s)^T\omega_k] = \mathbb{E}[\nabla_\theta \log p_\theta(s)\phi(x)^T\omega_k]$