Reinforcement Learning - Lecture 6

Contents

1 Back to the homework

For the optimization of the quantiles resolving the exact problem does not lead to good results, we could try to find the best beta that gives the best policy for the problem for a certain quantile. Indeed we can define:

$\text{VaR}_\delta = \inf\{t : \mathbb{P}(X \leq t) \geq 1 - \delta\}$

$$\begin{aligned}\mathbb{P}(X \geq u) &= \mathbb{P}(e^{\beta X} \geq e^{\beta u}) \\ &\geq \frac{\mathbb{E}[e^{\beta X}]}{e^{\beta u}} \\ &= \exp\left(\beta\left(\frac{1}{\beta}\ln(\mathbb{E}[e^{\beta X}]) - u\right)\right) \\ &= \exp(\beta(U_\beta(X) - u))\end{aligned}$$

$$\exp(\beta(U_\beta(X) - u^*(\beta))) = \delta \Leftrightarrow u^*(\beta) = U_\beta(X) + \frac{1}{\beta}\ln\left(\frac{1}{\delta}\right)$$

$$\forall \beta > 0, \mathbb{P}(X \geq u^*(\beta)) \leq \delta \Rightarrow \text{VaR}_\delta(X) \leq u^*(\beta)$$

$$\Rightarrow \text{VaR}_\delta(X) \leq \inf_\beta u^*(\beta) = \inf_\beta \left(U_\beta(X) + \frac{1}{\beta}\ln\left(\frac{1}{\delta}\right)\right)$$

This quantity is called the **Entropic Value at Risk (EVaR)**:

$$\text{EVaR}_\delta(X) = \inf_\beta \left(U_\beta(X) + \frac{1}{\beta}\ln\left(\frac{1}{\delta}\right)\right)$$

Measures of Risk are called coherent if they satisfy the following properties:

Note

Definition: Measure of Risk
A risk measure ρ is coherent if it satisfies:

- (P1) **Translation invariance**: $\rho(X + c) = \rho(X) + c$
- (P2) **Subadditivity**: $\rho(X + Y) \leq \rho(X) + \rho(Y)$
- (P3) **Monotonicity**: $X \leq Y \Rightarrow \rho(X) \leq \rho(Y)$
- (P4) **Positive homogeneity**: $\forall \lambda \geq 0, \rho(\lambda X) = \lambda \rho(X)$

Example

Examples of risk measures:

- $\mathbb{E}[X]$ is a coherent risk measure.
- U_β:
  - (P1): $U_\beta(X + c) = \frac{1}{\beta}\ln(\mathbb{E}[e^{\beta(X+c)}]) = \frac{1}{\beta}\ln(e^{\beta c}\mathbb{E}[e^{\beta X}]) = c + U_\beta(X)$
  - (P2): $U_\beta(X + Y)$: OK
  - (P3): OK

  - (P4): $U_\beta(\lambda X) = \frac{1}{\beta}\ln(\mathbb{E}[e^{\beta\lambda X}])$. If $X \sim \mathcal{N}(\mu, \sigma^2)$ then $U_\beta(\lambda X) = \lambda\mu + \frac{\lambda^2\beta\sigma^2}{2} \neq \lambda U_\beta(X)$

- $\mathrm{VaR}_\delta(X)$: Not subadditive
- $\mathbb{E}[X] + \lambda\sqrt{\mathrm{Var}(X)}$: Not monotone

Note

Definition: Conditional Value at Risk

$$\mathrm{CVaR}_\delta(X) = \mathbb{E}[X \mid X \geq \mathrm{VaR}_\delta(X)]$$

is called the Conditional Value at Risk (CVaR). It can also be expressed as:

$$\mathrm{CVaR}_\delta(X) = \inf_{t\in\mathbb{R}}\left(t + \frac{1}{\delta}\mathbb{E}[(X-t)_+]\right) = \int_0^\delta \mathrm{VaR}_u(X)\,du$$

Proposition:

CVaR is a coherent risk measure for all $\delta \in (0, 1)$.

Proposition:

- EVaR is a coherent risk measure.
- $\mathrm{CVaR}_\delta(X) \leq \mathrm{EVaR}_\delta(X)$

Example

For $X \sim \mathcal{N}(0, 1)$:

$$\mathbb{P}(X \geq u) \approx_{u\to+\infty} \frac{1}{u\sqrt{2\pi}} e^{-\frac{u^2}{2}} = \delta$$

$$u_\delta \approx \sqrt{-2\ln\left(\frac{1}{\delta}\right)} - c\sqrt{2\ln\left(\frac{1}{\delta}\right)} = \sqrt{2\ln\left(\frac{1}{\delta}\right)} + o\left(\sqrt{\ln\left(\frac{1}{\delta}\right)}\right)$$

Note

Algorithm in practice:

- Take a grid of β values
- For each β compute the optimal policy for U_β with the distributed planning algorithm
- Choose the best policy among the different β for the quantile of interest

2 Infinite Horizon MDPs

MDP: $(\mathcal{S}, \mathcal{A}, k, r)$

We introduce a discount factor $\gamma \in (0,1)$ which measures the preference of present rewards over future rewards.

$$W = \sum_{t=1}^{+\infty} \gamma^t R_t$$

- $t \ll \frac{1}{1-\gamma} \Rightarrow \gamma^t \approx 1$
- $t \gg \frac{1}{1-\gamma} \Rightarrow \gamma^t \approx 0$

Alternative interpretation: $W = \mathbb{E}_\tau \left[\sum_{t=1}^{\tau} R_t\right]$ with $\tau \sim \text{Geometric}(1-\gamma)$

2.1 Policy Evaluation

Fix a policy π and a state $s \in \mathcal{S}$:

$$\begin{aligned}
V^\pi(s) &= \mathbb{E}\left[\sum_{t=1}^{+\infty} \gamma^t R_t \mid S_1 = s\right] \\
&= \mathbb{E}\left[\mathbb{E}\left[\gamma R_1 + \sum_{t=2}^{+\infty} \gamma^t R_t \mid S_1 = s, S_2\right]\right] \\
&= \mathbb{E}[\gamma R_1 \mid S_1 = s] + \sum_{s'} k(s, \pi(s), s') \mathbb{E}\left[\sum_{t=2}^{+\infty} \gamma^t R_t \mid S_2 = s'\right] \\
&= \mathbb{E}[\gamma R_1 \mid S_1 = s] + \gamma \sum_{s'} k(s, \pi(s), s') V^\pi(s')
\end{aligned}$$

The last part is equal to $\gamma \mathbb{E}\left[\sum_{t'=1}^{+\infty} \gamma^{t'} R_{t'} \mid S_1 = s'\right]$.
If $R^\pi(s) = \mathbb{E}[R_1 \mid S_1 = s] = \sum_{s'} k(s, \pi(s), s') r(s, \pi(s), s')$ then:

$$V^\pi(s) = R^\pi(s) + \gamma \sum_{s'} k(s, \pi(s), s') V^\pi(s')$$

2.1.1 Matrix form

$$V^\pi = \begin{pmatrix} V^\pi(s_1) \\ V^\pi(s_2) \\ \vdots \\ V^\pi(s_k) \end{pmatrix} \in \mathbb{R}^k$$

where $\mathcal{S} = \{1, 2, \ldots, k\}$ and $k = |\mathcal{S}|$.

$$R^\pi = \begin{pmatrix} R^\pi(1) \\ R^\pi(2) \\ \vdots \\ R^\pi(k) \end{pmatrix} \in \mathbb{R}^k$$

$$K^\pi = (k(s, \pi(s), s'))_{1 \le s, s' \le k} \in \mathbb{R}^{k \times k}$$

is a stochastic matrix.
We have then:

$$\begin{aligned} V^\pi &= R^\pi + \gamma K^\pi V^\pi \\ &\Leftrightarrow (I - \gamma K^\pi)V^\pi = R^\pi \\ &\Leftrightarrow V^\pi = (I - \gamma K^\pi)^{-1} R^\pi \\ &= (I + \gamma K^\pi + \gamma^2 (K^\pi)^2 + \cdots + \gamma^k (K^\pi)^k) R^\pi \end{aligned}$$

This matrix is invertible because K^π is stochastic: $\forall s, \sum_j K^\pi(s,j) = 1$. Thus for all λ the eigenvalues of K^π satisfy: $|\lambda| \leq 1$ and $K^\pi u = \lambda u$.

2.2 Bellman Operator

We define $T^\pi : \mathbb{R}^k \to \mathbb{R}^k$ as: $T^\pi(V) = R^\pi + \gamma K^\pi V$, the Bellman operator for policy π.

Proposition:

T^π is affine, isotonic and contractant:

- **Isotonic**: if $V \leq V'$ (in the sense that $\forall s, V(s) \leq V'(s)$) then $T^\pi(V) \leq T^\pi(V')$
  Indeed:
  $$T^\pi(V_1)(i) = R^\pi(i) + \gamma \sum_j K^\pi(i,j) V_1(j) \leq R^\pi(i) + \gamma \sum_j K^\pi(i,j) V_2(j) = T^\pi(V_2)(i)$$
- **Contractant**: $\forall V_1, V_2 \in \mathbb{R}^k, ||T^\pi(V_1) - T^\pi(V_2)||_\infty \leq \gamma ||V_1 - V_2||_\infty$
  Indeed:
  $$\begin{aligned} |T^\pi(V_1)(i) - T^\pi(V_2)(i)| &= \gamma \left| \sum_j K^\pi(i,j)(V_1(j) - V_2(j)) \right| \\ &\leq \gamma \sum_j K^\pi(i,j) |V_1(j) - V_2(j)| \\ &\leq \gamma ||V_1 - V_2||_\infty \end{aligned}$$

Consequences: T^π has a unique fixed point V^π such that $T^\pi(V^\pi) = V^\pi$ and for any $V \in \mathbb{R}^k$, $(T^\pi)^n(V) \to V^\pi$ when $n \to +\infty$.
Let $V_n = (T^\pi)^n V$, V^π satisfies: $T^\pi(V^\pi) = V^\pi$ then:

$$||V_n - V^\pi||_\infty = ||T^\pi(V_{n-1}) - T^\pi(V^\pi)||_\infty \leq \gamma ||V_{n-1} - V^\pi||_\infty \leq \gamma^n ||V_0 - V^\pi||_\infty \to 0$$

when $n \to +\infty$.

2.3 Evaluating a policy from samples

Two main approaches:

- Monte Carlo evaluation
- Temporal Difference (TD) learning

Algorithm: Temporal Difference Learning

Input: V_0 the initial guess $\in \mathbb{R}^k$, T the number of iterations

1. $V \leftarrow V_0$
2. $S = S_1$
3. for $t = 1$ to T do
   - $R \leftarrow r(S, \pi(S), S')$ where
   - $S' \leftarrow$ next state from using $\pi(S)$
   - $V(S) \leftarrow (1 - \alpha_t)V(S) + \alpha_t(R + \gamma V(S'))$
   - $S \leftarrow S'$
4. return V

Evolution of V:

$$V : \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{pmatrix} \rightarrow \begin{pmatrix} \alpha_1 R_1 \\ 0 \\ 0 \\ 0 \\ 0 \end{pmatrix} \rightarrow \begin{pmatrix} \alpha_1 R_1 \\ 0 \\ \vdots \\ \alpha_2 R_2 \\ 0 \end{pmatrix} \rightarrow \cdots \rightarrow \begin{pmatrix} V^\pi(1) \\ V^\pi(2) \\ V^\pi(3) \\ V^\pi(4) \\ V^\pi(5) \end{pmatrix}$$

Note

One can prove that if $\alpha_t \rightarrow 0$, $\sum_t \alpha_t = +\infty$ and $\sum_t \alpha_t^2 < +\infty$ then $V_t \rightarrow V^\pi$ when $t \rightarrow +\infty$ with probability 1.