RL 6: The Bellman Equation

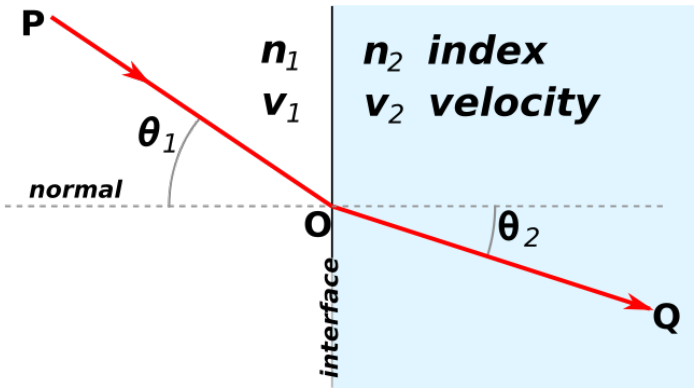
Michael Herrmann

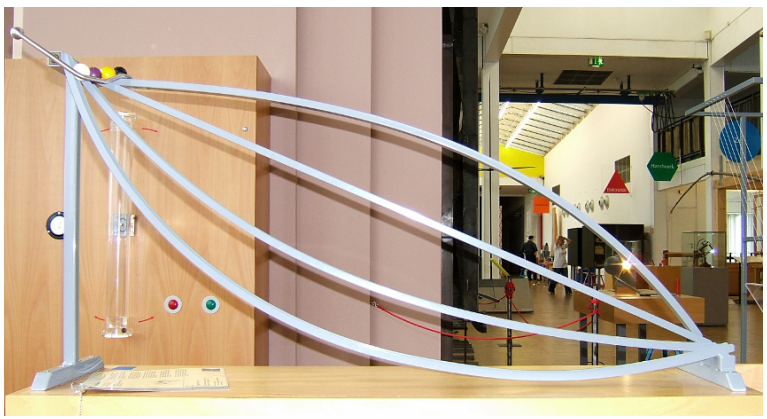
University of Edinburgh, School of Informatics

31/01/2014

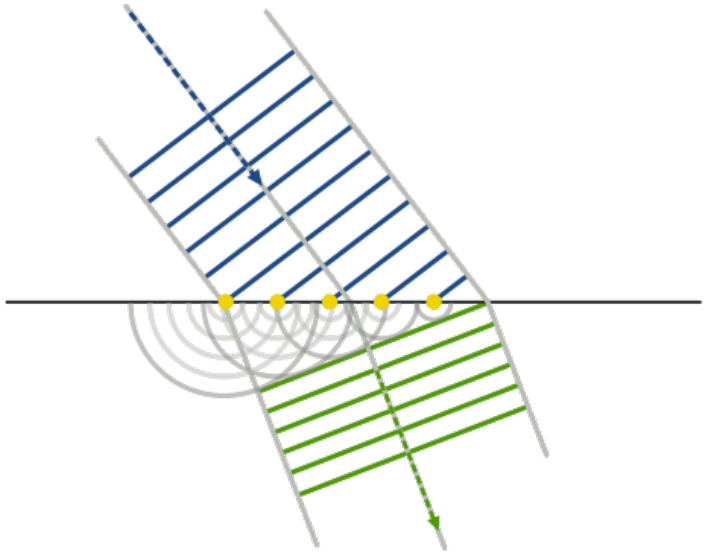
Optimality in Physics

Fermat's principle





Huygens–Fresnel principle



Brachistochrone curve
(Johann Bernoulli, 1696)
$\Rightarrow$ Calculus of variations

Optimal Control

x: state
V: value
C: scalar cost
D: value of final state

Value at starting state

$$V(x(0), 0) = \min_u \left\{ \int_0^T C[x(t), u(t)]\, dt + D[x(T)] \right\}$$

Hamilton-Jacobi-Bellman equation

$$\dot{V}(x, t) + \min_u \{\nabla V(x, t) \cdot F(x, u) + C(x, u)\} = 0$$

with $\dot{x}(t) = F[x(t), u(t)]$ determining the evolution of the state

N.B. This is copied from wikipedia and included here only for comparison. The important part begins on the next slide.

Optimality and Reward

Bellman's Principle of Optimality: An optimal policy has the property that whatever the initial state and initial decision are, the remaining decisions must constitute an optimal policy with regard to the state resulting from the first decision. (1957)

The reward hypothesis (Sutton): That all of what we mean by goals and purposes can be well thought of as the maximisation of the cumulative sum of a received scalar signal (reward).

Formulate learning problem such that the principle can be applied.

Value functions: Definition

Policy $\pi : \mathcal{S} \to \mathcal{A}$ or $a \sim \pi(\cdot|s)$ (means: $P(a_t = a) = \pi(a|s_t)$)

Value function $V^\pi : \mathcal{S} \to \mathbb{R}$

$$V^\pi(s) = E\left[\sum_{t=0}^{\infty} \gamma^t r_{t+1} | s_0 = s\right]$$

assuming that the initial probability $\pi^0(s_0) > 0$

This is assuming an MDP with fixed π: $\left(\mathcal{S}, P^\pi, \pi^0\right)$ which is extended to $\left(\mathcal{S}, P^\pi, \pi^0, \mathcal{R}\right)$. The latter is a *Markov reward process* which arises naturally by assigning a reward distribution $R(\cdot|s)$ to each state s or to each state-action pair according to

$$R^\pi(r|s) = \sum_{a \in \mathcal{A}} \pi(a|s) R(r|s,a)$$

Value functions for state-action pairs

Policy $\pi : \mathcal{S} \to \mathcal{A}$ or $a \sim \pi(\cdot|s)$

Value function $\mathcal{Q}^\pi : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$

$$\mathcal{Q}^\pi(s,a) = E\left[\sum_{t=0}^{\infty} \gamma^t r_{t+1} | s_0 = s, a_0 = a\right]$$

assuming that the initial probability $\pi^0(s_0) > 0$ and that $\pi(s_0) = a_0$ (deterministic) or $\pi(s_0, a_0) > 0$ (stochastic).

First action a_0 is applied now, later actions are chosen by π.

[Note that the initial distribution π^0 and the policy π are different mathe- matical objects.]

Optimal value functions

For MDPs an optimal policy always exists ($s \in \mathcal{S}$, π fixed)

$$V^*(s) = \sup_\pi V^\pi(s)$$

For state-action pairs we have ($s \in \mathcal{S}$ and $a \in \mathcal{A}$)

$$\begin{aligned} V^*(s) &= \sup_{a \in \mathcal{A}} Q^*(s,a) \\ \mathcal{Q}^*(s,a) &= r(s,a) + \gamma \sum_{u \in \mathcal{S}} P^{a=\pi(s)}(s,u)\, V^*(u) \end{aligned}$$

Optimality of $\mathcal{Q}$

Suppose π satisfies

$$\sum_{a \in \mathcal{A}} \pi(a|s)\, \mathcal{Q}^*(s,a) = V^*(s)$$

for all $s \in \mathcal{S}$. Then π is optimal.

Namely, $\pi(\cdot|s)$ selects the action(s) that maximise(s) $\mathcal{Q}^*(s,\cdot)$.

So, optimality implies greediness and knowing $\mathcal{Q}^*(s,a)$ allows us to act optimally.

Analogously, knowing V^*, r and P suffices to act optimally.

Bellman Equations for deterministic policies in an MDP

How to find the value of a policy?

$$V^{\pi}(s) = r(s, \pi(s)) + \gamma \sum_{u \in \mathcal{S}} P(s, \pi(s), u) V^{\pi}(u)$$

This is the Bellman equation for V^{π}.

Define the Bellman operator for π as $T^{\pi} : \mathbb{R}^{\mathcal{S}} \to \mathbb{R}^{\mathcal{S}}$ (maps value functions to value functions)

$$(T^{\pi} V)(s) = r(s, \pi(s)) + \gamma \sum_{u \in \mathcal{S}} P(s, \pi(s), u) V(u)$$

Then naturally,

$$T^{\pi} V^{\pi} = V^{\pi}$$

which is nothing but a compact formulation of the equation on top of this slide. This is a linear equation in V^{π} and T^{π}.

If $0 < \gamma < 1$ then T^* is a contraction w.r.t. the maximum norm.

Bellman optimality equations

How to characterise the optimal policy? Use the Bellman optimality principle.

$$V^*(s) = \sup_{a \in \mathcal{A}} \left(r(s,a) + \gamma \sum_{u \in \mathcal{S}} P(s,a,u) V^*(u) \right)$$

Bellman optimality operator $T^* : \mathbb{R}^{\mathcal{S}} \to \mathbb{R}^{\mathcal{S}}$

$$(T^* V)(s) = \sup_{a \in \mathcal{A}} \left(r(s,a) + \gamma \sum_{u \in \mathcal{S}} P(s,a,u) V(u) \right)$$

Then naturally,

$$T^* V^* = V^*$$

which is nothing but a compact formulation of the equation on top of this slide.

If $0 < \gamma < 1$ then T^* is a contraction w.r.t. the maximum norm.

Bellman Operators for state-action value functions

$$T^{\pi} : \mathbb{R}^{\mathcal{S}\times\mathcal{A}} \to \mathbb{R}^{\mathcal{S}\times\mathcal{A}}, \; T^{*} : \mathbb{R}^{\mathcal{S}\times\mathcal{A}} \to \mathbb{R}^{\mathcal{S}\times\mathcal{A}}$$

$$\begin{aligned} T^{\pi}\mathcal{Q}(s,a) &= r(s,a) + \gamma \sum_{u\in\mathcal{S}} P(s,a,u)\, \mathcal{Q}(u,\pi(s)) \\ T^{*}\mathcal{Q}(s,a) &= r(s,a) + \gamma \sum_{u\in\mathcal{S}} P(s,a,u) \sup_{b\in\mathcal{A}} \mathcal{Q}(u,b) \end{aligned}$$

T^{π} is a linear operator, but T^{*} is not. Both, T^{π} and T^{*} are contractions w.r.t. the maximum norm.

Defining $\mathcal{Q}(s,\pi(s)) = \mathcal{Q}^{\pi}$ we have $T^{\pi}\mathcal{Q}^{\pi} = \mathcal{Q}^{\pi}$ and $\mathcal{Q}^{\pi}$ is the unique solution of this equation. Similarly, we have $T^{*}\mathcal{Q}^{*} = \mathcal{Q}^{*}$ and $\mathcal{Q}^{*}$ is the unique solution of this equation.

Starting from arbitrary V_0

$$V_{t+1} = T^* V_t$$

global convergence?

For state-action values functions

$$\mathcal{Q}_{t+1} = T^* \mathcal{Q}_t$$

Once V_t (or $\mathcal{Q}_t$) is close to V^* (or $\mathcal{Q}^*$) then the greedy policy is close to optimal.

Suppose $\mathcal{Q}$ is fixed and π is the greedy policy w.r.t. $\mathcal{Q}$. Then

$$V^\pi(s) \geq V^*(s) - \frac{2}{1-\gamma} \|\mathcal{Q} - \mathcal{Q}^*\|_\infty$$

Singh and Yee, 1994

Fix an arbitrary initial policy π_0.

Policy evaluation: At iteration $t > 0$ compute the action-value function underlying π_t

Policy improvement: Given $\mathcal{Q}^{\pi_t}$ define π_{t+1} as the policy that is greedy w.r.t. $\mathcal{Q}^{\pi_t}$.

Works similar to value iteration, but policy evaluation is computationally more costly.

Summary on Bellman equations

- *The* Bellman equation is the Bellman optimality equation. It characterises the optimal strategy based on the Bellman optimality principle
- It uses the transition probabilities
- Outlook: Use the actual process to estimate the transition probabilities or to directly sample the value function (or the state-action value function)
- Next: value prediction

Value prediction

- The Bellman (optimality) equation characterises an optimal value function
- In general this equation is not solvable
- Solution is possible by iterative schemes
- Need to take into account the embeddedness of the agent

Temporal Difference (TD) Learning for Value Prediction

Ideal value function

$$\begin{aligned} V_t &= \sum_{\tau=t}^{\infty} \gamma^{\tau-t} r_\tau = r_t + \gamma r_{t+1} + \gamma^2 r_{t+2} + \cdots \\ &= r_t + \gamma \left(r_{t+1} + \gamma r_{t+2} + \cdots \right) \\ &= r_t + \gamma \sum_{\tau=t+1}^{\infty} \gamma^{\tau-(t+1)} r_\tau \\ &= r_t + \gamma V_{t+1} \end{aligned}$$

Real value function is based on estimates of V_t and V_{t+1}, which may not obey this relation. Even if the estimates $\hat{V}_t$ and $\hat{V}_{t+1}$ are far from their true values we can at least require consistency, i.e. minimise the absolute value of the δ error (δ for $\delta\iota\alpha\varphi o\rho\acute{\alpha}$)

$$\delta_{t+1} = r_t + \gamma \hat{V}_{t+1} - \hat{V}_t$$

The simplest TD algorithm

Let $\hat{V}_t$ be the t-th iterate of a learning rule for estimating the value function V.

Let s_t the state of the system at time step t.

$$\delta_{t+1} = r_t + \gamma \hat{V}_t(s_{t+1}) - \hat{V}_t(s_t)$$

$$\hat{V}_{t+1} = \begin{cases} \hat{V}_t(s) + \eta\delta_{t+1} & \text{if } s = s_t \\ \hat{V}_t(s) & \text{otherwise} \end{cases}$$

$$\begin{aligned} \hat{V}_{t+1}(s_t) = \hat{V}_t(s_t) + \eta\delta_{t+1} &= \hat{V}_t(s_t) + \eta\left(r_t + \gamma \hat{V}_t(s_{t+1}) - \hat{V}_t(s_t)\right) \\ &= (1-\eta)\hat{V}_t(s_t) + \eta\left(r_t + \gamma \hat{V}_t(s_{t+1})\right) \end{aligned}$$

The update of the estimate $\hat{V}$ is an exponential average over the cumulative expected reward.

TD(0) Algorithm

Initialise η and γ and execute after each state transition

function TD0(s,r,$s1$,V) {

$\delta := r + \gamma * V[s1] - V[s];$

$V[s] := V[s] + \eta * \delta;$

return V;

}

Remarks on the TD(0) Algorithm

- If the algorithm converges it must converge to a value function where the expected temporal differences are zero for all states. This state satisfies the Bellman optimality equation.
- The continuous version of the algorithm can be shown to be globally asymptotically stable
- TD(0) is a stochastic approximation algorithm. If the system is ergodic and the learning rate is appropriately decreased, it behaves like the continuous version.

Robbins-Monro conditions

How to choose learning rates? If

$$\sum_{t=0}^{\infty} \eta_t = \infty, \qquad \sum_{t=0}^{\infty} \eta_t^2 < \infty,$$

then $V_t(\cdot)$ will behave as the temporally continuous variant

$$\frac{dV(\cdot)}{dt} = r + (\gamma P - I) V(\cdot)$$

Choosing e.g. $\eta_t = c\, t^{-\alpha}$, the conditions hold for $\alpha \in \left(\frac{1}{2}, 1\right]$:

- $\alpha > 1$: goal is reachable even after been temporally trapped
- $\alpha = 1$: smallest step sizes, but still possible
- $\alpha \leq \frac{1}{2}$: large fluctuations can happen even after long time

Iterate-averaging (Polyak & Juditsky, 1992) gives best possible asymptotic rate of convergence

Practically: fixed step sizes or finite-time reduction (see earlier slide)

Acknowledgements

Many slides are adapted from web resources associated with Sutton and Barto's Reinforcement Learning book

. . . before being used by Dr. Subramanian Ramamoorthy in this course in the last three years.

... today's lecture followed mostly the book on Algorithms for Reinforcement learning by C. Szepesvari, Ch. 1.