Control theory and Reinforcement Learning - Lecture 4

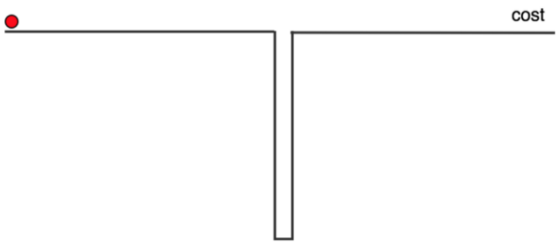
Carlos Esteve Yagüe

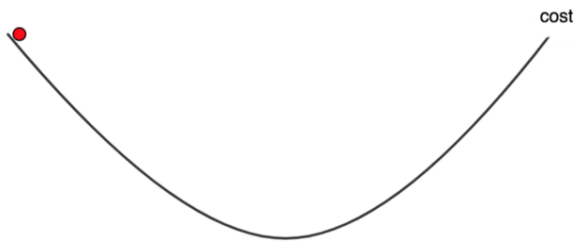
Universidad Autónoma de Madrid - Fundación Deusto

September 2020

We aim to act on a controlled environment in order to achieve a prescribed goal

The choice of a ”good” cost functional is crucial for the performance of the learning algorithm.





Inverse Reinforcement Learning addresses the problem of designing better rewards (or cost) functionals.

One way to design an appropriate cost functional is to use observations from expert demonstrations (supervised learning).

Problem: Consider a dynamical system

$$x_{t+1} = f(x_t, u_t), \qquad x_t \in X,\ u_t \in \mathcal{U},$$

a discount factor $\gamma \in (0, \infty]$ and an optimal feedback control

$$\pi^* : X \longrightarrow \mathcal{U}.$$

Find the set of cost functions $C(x, u)$ such that the control $u_t = \pi^*(x_t)$ minimizes the functional

$$\sum_{t=0}^{\infty} \gamma^t C(x_t, u_t)$$

Main issue: non-uniqueness and degeneracy. Existence of degenerate solutions to the inverse problem (e.g. $C(x, u) \equiv 0$).

Inverse RL history

- 1964, Kalman posed the inverse problem of identifying a quadratic cost for linear dynamics, and solve it in the 1D control case.
- 1973, Jameson and Kreindler studied the same problem for the case of multidimensional control.
- 2000, Ng and Russell: MDP formulation, proposed solutions to the reward degeneracy (based on heuristics).
- 2004, Abbeel and Ng: Inverse RL for apprenticeship learning, reward feature matching

$$C_\theta(x,u) = \sum_{i=1}^{m} \theta_i \varphi_i(x,u)$$

- 2006–. . . : following the formulation of Ng and Russell + new ideas (max-margin formulation, game theoretic formulation, max entropy,...)

Consider the dynamical system

$$\begin{aligned} x'(t) &= A\,x(t) + B\,u(t) \quad t \in (0,\infty) \\ x(0) &= x_0 \end{aligned}$$

and the cost functional

$$J(u) := \frac{1}{2}\int_0^\infty x(t)^T Q x(t) + u(t)^T R u(t)\; dt$$

with (A, B) controllable and (A, Q) observable.

The solution to the direct problem is well known by Riccati Theory:

$$u^*(t) = -R^{-1}B^T P x(t)$$

where P is the unique definite positive solution to the Algebraic Riccati Equation

$$0 = P\,A + A^T P - P\,B\,R^{-1}B^T P + Q \qquad \text{(ARE)}$$

Consider the dynamical system

$$\begin{aligned} x'(t) &= A\,x(t) + B\,u(t) \qquad t \in (0,\infty) \\ x(0) &= x_0, \quad u(t) = Dx(t) \end{aligned}$$

where the matrix D makes the system be asymptotically stable.
Inverse problem: Given the matrices A, B and D, find pairs of matrices (Q, R), with $R > 0$, such that

$$D = -R^{-1}B^T P, \tag{1}$$

where $P > 0$ satisfies

$$0 = PA + A^T P - PBR^{-1}B^T P + Q. \tag{ARE}$$

Goals:

1. Give necessary and sufficient conditions on B and D for the existence of $R > 0$ and $P > 0$ such that (1) is satisfied.
2. Construction of all such (R, P). The matrix Q is then obtained from (ARE).

Consider the dynamical system

$$\begin{aligned} x'(t) &= A\,x(t) + B\,u(t) \qquad t \in (0,\infty) \\ x(0) &= x_0, \quad u(t) = Dx(t) \end{aligned}$$

Given B and D (stabilizing the system), we look for matrices P, R satisfying

$$B^T P = -R\,D \tag{2}$$

with $R > 0$ and $P > 0$.
Observe that (2) is equivalent to

$$B^T PB = -RDB, \qquad \text{(this imples } RDB \text{ is symmetric.)}$$

Necessary conditions

Such P and R exist only if

- $\text{rank}(D) = \text{rank}(B)$
- $DB < 0$

Construction of R and P

Let us assume the necessary conditions:

$$\text{rank}(D) = \text{rank}(B) \qquad \text{and} \qquad DB < 0$$

Theorem 1 (Jameson-Kreindler, 1973)

Let $V \in \mathcal{M}_m$ be a matrix whose columns are m linearly independent eigenvectors of $(DB)^T$.
Then, all matrices $R > 0$ such that RDB is symmetric are generated by

$$R = V \Gamma V^T,$$

for any matrix $\Gamma > 0$ such that $\Gamma\Lambda = \Lambda\Gamma$, where Λ is the diagonal matrix with the eigenvalues λ of DB.

Theorem 2 (Jameson-Kreindler, 1973)

Let $m \leq n$ and $\text{rank}(B) = m$. If $DB < 0$ and $\text{rank}(D) = \text{rank}(B)$, then all the solutions (P, R) to (2) are given by R as in the previous Theorem and P by

$$P = -D^T (B^T D^T)^{-1} RD + Y,$$

where Y is any symmetric real matrix such that $B^T Y = 0$.

Summary https://www.researchgate.net/publication/237523986_Inverse_Problem_of_Linear_Optimal_Control
Let $A \in \mathcal{M}_n$, $B \in \mathcal{M}_{n,m}$ and $D \in \mathcal{M}_{m,n}$ be given such that the system

$$x'(t) = A\,x(t) + B\,u(t) \qquad \text{with} \quad u(t) = Dx(t)$$

is asymptotically stable, and assume that $m \leq n$ and $\text{rank}(B) = m$.
Inverse problem: Find all the matrices Q and $R > 0$ such that the feedback $u(x) = Dx$ minimizes the functional

$$J(u) := \frac{1}{2}\int_0^\infty x(t)^T Q\,x(t) + u(t)^T R\,u(t)dt.$$

- The inverse problem has at least one solution if and only if

$$DB < 0 \qquad \text{and} \qquad \text{rank}(D) = \text{rank}(B).$$

- In the case D and B satisfy these conditions, all the solutions (Q, R) are given by R as in Theorem 1 and Q as

$$Q = PBR^{-1}B^T P - PA - A^T P,$$

where P is given as in Theorem 2.

Example

Consider the system

$$x_1'(t) = x_1(t) + x_2(t) + u_1(t)$$
$$x_2'(t) = u_1(t) + u_2(t)$$

and the feedback control

$$u_1(x_1, x_2) = -2x_1 - x_2 \qquad \text{and} \qquad u_2(x_1, x_2) = -x_2.$$

Observe that the matrices

$$A = \begin{pmatrix} 1 & 1 \\ 0 & 0 \end{pmatrix} \qquad B = \begin{pmatrix} 1 & 0 \\ 1 & 1 \end{pmatrix} \qquad D = \begin{pmatrix} -2 & -1 \\ 0 & -1 \end{pmatrix}$$

satisfy $A + BD = \begin{pmatrix} -1 & 0 \\ -2 & -2 \end{pmatrix} < 0$, so the system with the feedback $u(x)$ is asymptotically stable.

In addition, we have $\text{rank}(D) = \text{rank}(B) = 2$ and

$$DB = \begin{pmatrix} -3 & -1 \\ -1 & -1 \end{pmatrix} < 0$$

so the necessary and sufficient conditions for the solvability of the inverse problem are satisfied.

Example

We construct the matrix V whose columns are two eigenventors of the matrix

$$(DB)^T = \begin{pmatrix} -3 & -1 \\ -1 & -1 \end{pmatrix} \quad \text{for instance} \quad V = \begin{pmatrix} -0.9239 & 0.3827 \\ -0.3827 & -0.9239 \end{pmatrix}$$

Now take any matrix $\Gamma > 0$ that commutes with

$$\Lambda = \begin{pmatrix} -3.4142 & 0 \\ 0 & -0.5858 \end{pmatrix} \quad \text{for instance} \quad \Gamma = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}.$$

Then, a solution R to the inverse problem is given by

$$R = V \Gamma V^T = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} > 0.$$

For this R, we compute

$$P = -D^T (B^T D^T)^{-1} R D = \begin{pmatrix} 2 & 0 \\ 0 & 1 \end{pmatrix} > 0.$$

And with this P we compute a solution Q to the inverse problem as

$$Q = PBR^{-1}B^T P - PA - A^T P = \begin{pmatrix} 0 & 0 \\ 0 & 2 \end{pmatrix}$$

Inverse problem for general final costs

For a fixed time-horizon $T > 0$, consider the dynamical system in $\mathbb{R}^n$

$$x'(t) = u(t), \qquad t \in (0, T)$$

and, for any $t \in [0, T)$, let the cost functional

$$J_t(u) = \int_{T-t}^{T} \|u(s)\|^2 ds + C_f(x(T)).$$

We recall that the optimal feedback policy is given by

$$u^*(t) = \text{argmin}_u \left\{ \nabla V(x,t) \cdot u + \|u\|^2 \right\} = -\frac{\nabla_x V(x,t)}{2}$$

where value function is defined as

$$V(x,t) = \inf_{u \in L^2(0,T)} \{ J_t(u) \ : \ \text{s.t. } x(T-t) = x \}.$$

Warning: $V(x,t) \in \text{Lip}_{loc}(\mathbb{R}^n)$ might not be differentiable! We can use the reachable gradient instead

$$\begin{aligned} D_x^* V(x,t) := \quad & \{ p \in \mathbb{R}^n \ : \ \exists \{x_n\}_{n \geq 1} \text{ s.t. } V(\cdot, t) \text{ is diff. at } x_n \\ & x_n \to x, \text{ and } \nabla_x V(x_n, t) \to p \} \end{aligned}$$

Inverse problem for general final costs

Inverse problem: Given a value function $V_T(x) = V(x, T)$ at time T, find all the final costs $C_f(\cdot)$ which are compatible with the given value function.

Motivation: We prescribe a **desired behavior** (feedback control) **at time** T and want to **design a final cost** for which the prescribed feedback control is optimal, and then, will be **learned by the RL algorithm**.

Goals:

1. Give necessary and sufficient conditions on $V_T(\cdot)$ for the existence of at least one compatible cost $C_f(\cdot)$.
2. If the target value function $V_T(\cdot)$ is admissible, construct all the compatible final costs.
3. If target $V_T(\cdot)$ is not admissible, compute a projection $V_T^*(\cdot)$ of $V_T(\cdot)$ on the set of admissible targets.

It is well-known that, given the final cost C_f, the value function $V(x,t)$ as defined above, satisfies (in the viscosity sense) the terminal-value problem

$$\begin{aligned} &\partial_t V(x,t) + \frac{\|\nabla_x V(x,t)\|^2}{4} = 0 \\ &V(x,0) = C_f(x) \end{aligned} \tag{HJ}$$

The inverse problem can then be reformulated as follows:

Given a target function $V_T(\cdot) \in \text{Lip}(\mathbb{R}^n)$, construct all the initial conditions $C_f(\cdot) \in \text{Lip}(\mathbb{R}^n)$ such that

$$S_T^+ C_f(x) = V_T(x)$$

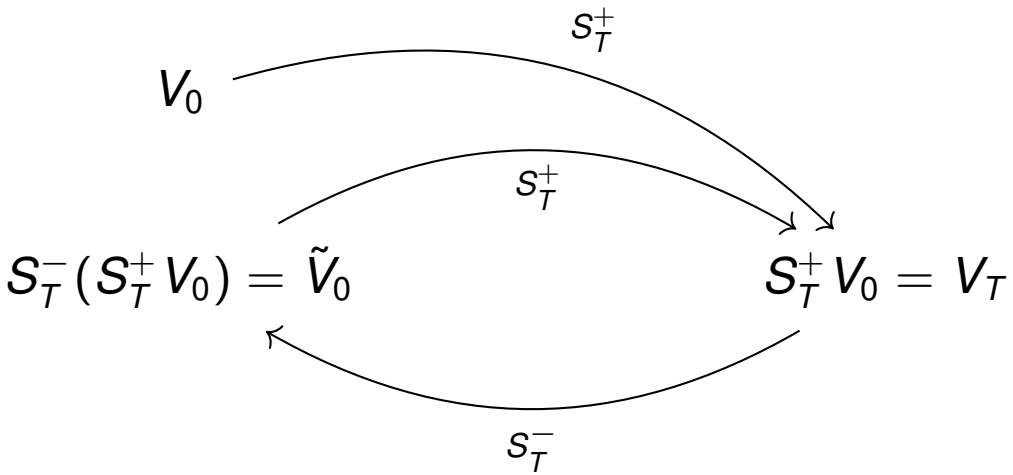
where S_T^+ is the operator that associates to any initial condition C_f, the solution $V(\cdot, T)$ of (HJ) at time T.

Remark: We can as well define the operator S_T^- which associates to any given function V_T the function $W(\cdot,0) = S_T^- V_T(\cdot)$, the *backward viscosity* solution to (HJ) with terminal condition $W(x.T) = V_T(x)$.
See more details in [E, Zuazua, 2020 to appear in SIAM J. on Math. An.]
`https://arxiv.org/abs/2003.06914`

Lemma

For any $V_0 \in \text{Lip}(\mathbb{R}^n$, we have

$$S_T^+ \circ S_T^- \circ S_T^+ (V_0(x)) = S_T^+ V_0(x).$$



Compatibility conditions

Lemma

For any $V_0 \in \text{Lip}(\mathbb{R}^n$, we have

$$S_T^+ \circ S_T^- \circ S_T^+ \left(V_0(x)\right) = S_T^+ V_0(x).$$

Theorem

Given $V_T \in \text{Lip}(\mathbb{R}^n)$, the inverse problem admits at least one solution C_f if and only if

$$S_T^+ \circ S_T^- \left(V_T(x)\right) = V_T(x)$$

which is equivalent to

$$\text{the function } V_T \text{ is semiconcave wtih modulus } \frac{2}{T}.$$

Let $T > 0$ and $V_T(\cdot)$ be an admissible value function at time T.

Then, using the previous Lemma we can construct a compatible terminal cost by means of the Hop-Lax formula:

$$\tilde{C}_f(x) = S_T^- V_T(x) = \max_{y \in \mathbb{R}^n} \left[V_T(y) - \frac{\|y - x\|^2}{T} \right].$$

Theorem [E, Zuazua, 2020]

Let $V_T(\cdot)$ be and admissible value function, then the final cost $C_f \in \text{Lip}(\mathbb{R}^n)$ is compatible with V_T if and only if

$$C_f(x) \geq \tilde{C}_f(x), \ \forall x \in \mathbb{R}^n \quad \text{and} \quad C_f(x) = \tilde{C}_f(x), \ \forall x \in X_T(V_T),$$

where

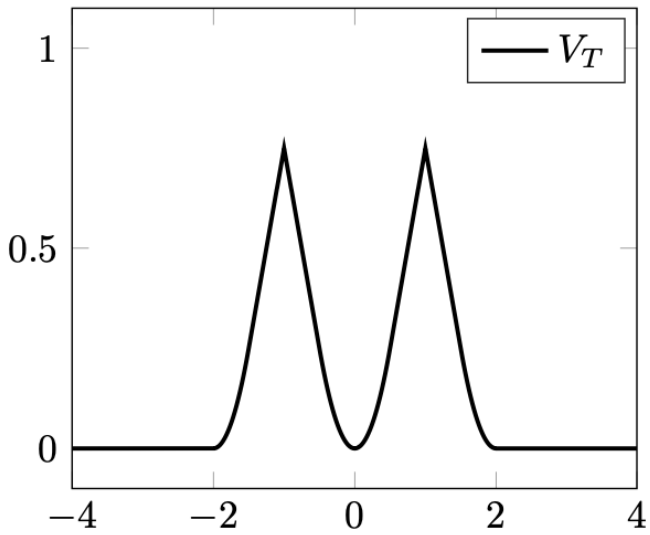
$$X_T(V_T) := \left\{ x - T \frac{\nabla V_T(x)}{2} \ ; \ \forall x \in \mathbb{R}^n \text{ s.t. } V_T \text{ is diff. at } x \right\}.$$

Example

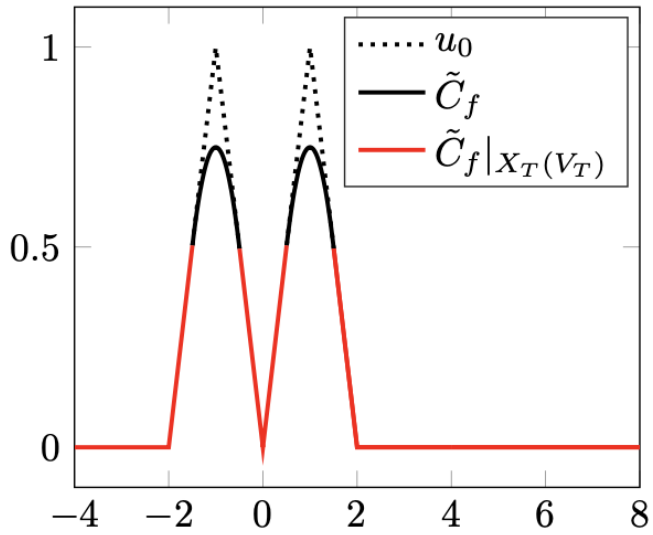
Consider $T = 0.5$ and the target value function

$$V_T(x) := S_T^+ u_0(x), \quad \text{where} \quad u_0(x) := \begin{cases} 1 - |x+1| & \text{if } -2 < x \leq 0 \\ 1 - |x-1| & \text{if } 0 < x < 2 \\ 0 & \text{else.} \end{cases}$$

Clearly, the target $_T$ is admissible. The construction of all C_f compatible with V_T can be done using the following scheme.



(A) The target V_T.



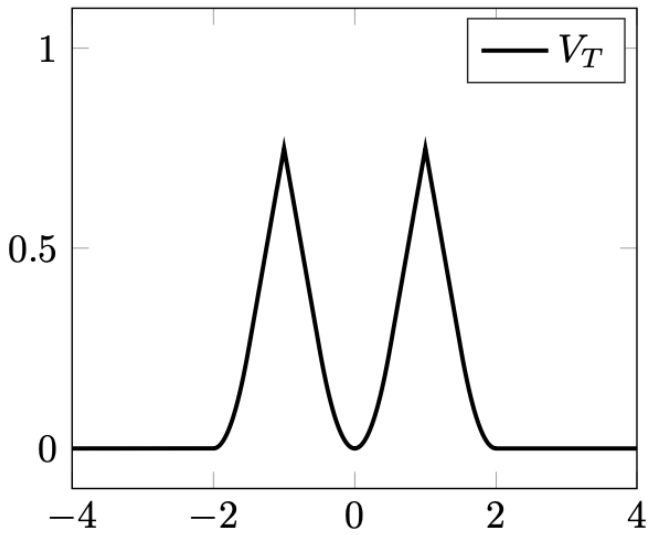
(B) The initial data $\tilde{C}_f$ and u_0 satisfy $S_T^+ \tilde{C}_f = V_T$.

Example

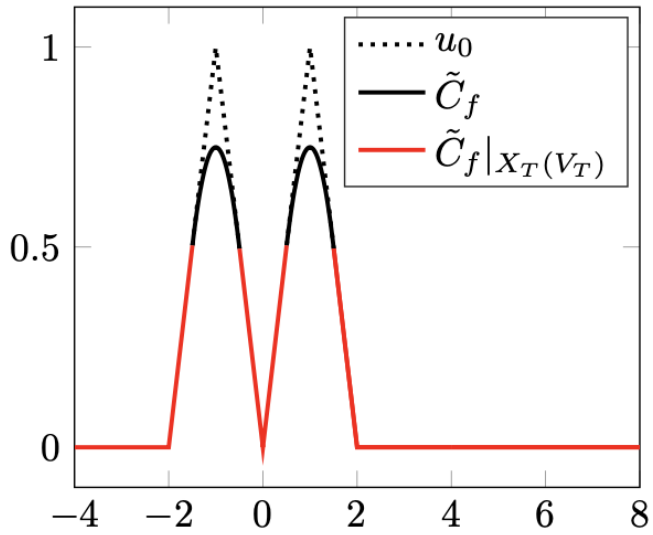
Consider $T = 0.5$ and the target value function

$$V_T(x) := S_T^+ u_0(x), \quad \text{where} \quad u_0(x) := \begin{cases} 1 - |x+1| & \text{if } -2 < x \leq 0 \\ 1 - |x-1| & \text{if } 0 < x < 2 \\ 0 & \text{else.} \end{cases}$$

Clearly, the target $_T$ is admissible. The construction of all C_f compatible with V_T can be done using the following scheme.



(A) The target V_T.



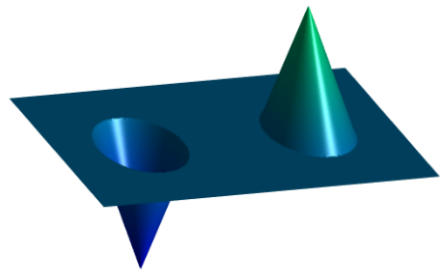
(B) The initial data $\tilde{C}_f$ and u_0 satisfy $S_T^+ \tilde{C}_f = V_T$.

Example

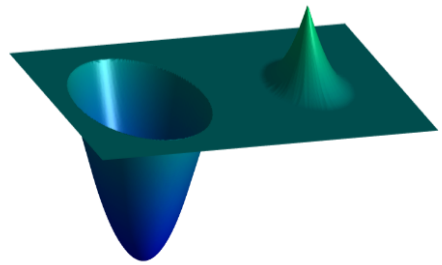
Consider $T = 0.5$ and the target value function

$$V_T := S_T^+ u_0, \quad \text{where} \quad u_0(x, y) := \begin{cases} \|x - z_1\| - 1, & \text{if } \|x - z_1\| < 1 \\ 1 - \|x - z_2\|, & \text{if } \|x - z_2\| < 1 \\ 0, & \text{else.} \end{cases}$$

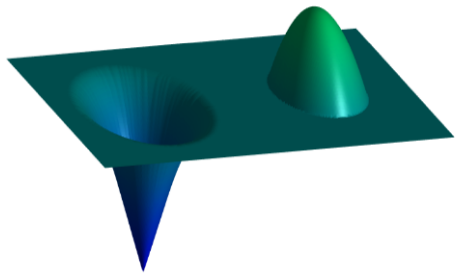
Clearly, the target $_T$ is admissible. The construction of all C_f compatible with V_T can be done using the following scheme.



(A) The function u_0.



(B) The target V_T defined as $V_T := S_T^+ u_0$.



(C) The compatible final cost $\tilde{C}_f$.

(D) The set $X_T(V_T)$ in dark green.

FIGURE 2. The compatible costs C_f are those functions which coincide with $\tilde{C}_f$ on the blue region while on its complement they are bigger or equal than it.

Non-admissible target functions

If the given target value function $V_T(x)$ is not admissible, we can make it admissible by applying the backward-forward HJ operator

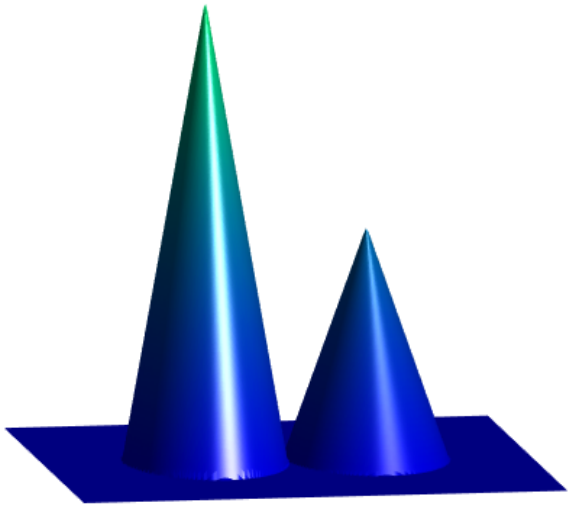
$$V_T^*(x) := S_T^+ \circ S_T^-(V_T(x)).$$

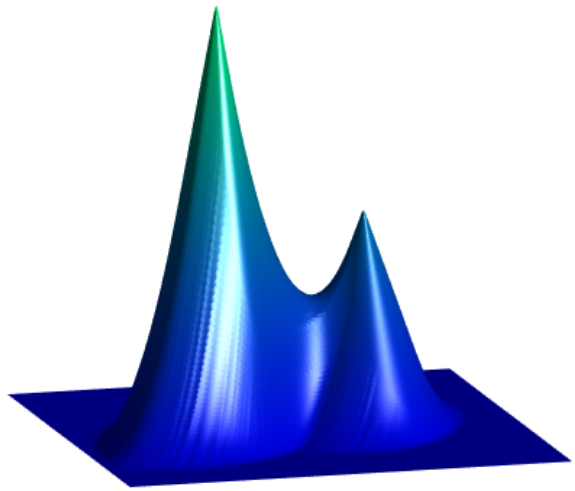
Theorem [E, Zuazua, 2020]

For a $V_T \in \text{Lip}(\mathbb{R}^n)$, the function V_T^* is the unique viscosity solution to the obstacle problem

$$\min\left\{V - V_T,\ -\lambda_n\left[D^2 V - \frac{2}{T} I_n\right]\right\} = 0,$$

The function V_T^* is the smallest admissible value function which is above V_T.





Summary

For $T \in (0, \infty]$ and $\gamma \in (0, 1]$, consider the following optimal control problem:

$$\underset{u_0, u_1, u_3 \dots}{\text{minimize}} \quad \sum_{t=0}^{T} \gamma^t C(x_t, u_t) \quad \text{subject to} \quad \begin{cases} x_{t+1} = f(x_t, u_t) & t = 0, 1, \dots, T \\ x_0 = x \in \mathbb{R}^n. \end{cases}$$

Goal: Find (learn) an optimal feedback control

$$\pi^* : \mathbb{R}^n \to \mathcal{U}, \quad \text{such that} \quad u_t^* = \pi^*(x_t) \text{ is optimal.}$$

The **value function** for $T = \infty$:

$$V(x) = \min_{\pi(\cdot)} \sum_{t=0}^{\infty} \gamma^t C(x_t, u_t)$$

Bellman equation:

$$V(x) = \min_{u} \{C(x, u) + \gamma V(f(x, u))\}$$

Value iteration: Set an initial guess $V_0(x)$, and then improve the approximation

$$V_{k+1}(x) := \min_{u} \{C(x, u) + \gamma V_k(f(x, u))\}$$

For $T \in (0, \infty]$ and $\gamma \in (0, 1]$, consider the following stochastic optimal control problem:

$$\underset{u_0, u_1, u_3 \dots}{\text{minimize}} \quad \mathbb{E}_w \left[\sum_{t=0}^{T} \gamma^t C(x_t, u_t) \right] \quad \text{subject to} \quad \begin{cases} x_{t+1} = f(x_t, u_t) + w_t & t = 0, 1, \dots, \\ x_0 = x \in \mathbb{R}^n. \end{cases}$$

where $w_0, w_1, w_2, \dots$ is an stochastic process.
Goal: Find (learn) an optimal feedback control

$$\pi^* : \mathbb{R}^n \to \mathcal{U}, \quad \text{such that} \quad u_t^* = \pi^*(x_t) \text{ is optimal.}$$

The **value function** for $T = \infty$:

$$V(x) = \min_{\pi(\cdot)} \mathbb{E}_w \left[\sum_{t=0}^{\infty} \gamma^t C(x_t, u_t) \right]$$

Bellman equation:

$$V(x) = \min_{u} \mathbb{E}_w \left\{ C(x, u) + \gamma V(f(x, u) + w) \right\}$$

Value iteration: Set an initial guess $V_0(x)$, and then improve the approximation

$$V_{k+1}(x) := \min_{u} \mathbb{E}_w \left\{ C(x, u) + \gamma V_k(f(x, u) + w) \right\}$$

For $T \in (0, \infty]$ and $\tau \in (0, 1]$, consider the following optimal control problem:

$$\underset{u \in L^2(0,T)}{\text{minimize}} \quad \int_0^T e^{-\tau t} C(x(t), u(t)) dt \quad \text{subject to} \quad \begin{cases} x'(t) = f(x(t), u(t)) & t \in (0, T) \\ x_0 = x \in \mathbb{R}^n. \end{cases}$$

Goal: Find (learn) an optimal feedback control

$$\pi^* : \mathbb{R}^n \to \mathcal{U}, \quad \text{such that} \quad u^*(t) = \pi^*(x(t)) \text{ is optimal.}$$

The **value function** for $T = \infty$:

$$V(x) = \min_{\pi(\cdot)} \int_0^\infty e^{-\tau t} C(x(t), u(t)) dt$$

Bellman equation:

$$\tau V(x) = H(x, \nabla V(x))$$

with $H(x, p) := \min_u \{p \cdot f(x, u) + C(x, u)\}$.
Value iteration: Set an initial guess $V_0(x)$, and then improve the approximation by taking $t >> 1$ in

$$\begin{aligned} \partial_t V(x, t) + \tau V(x, t) &= H(x, \nabla_x V(x, t)) \\ V(x, 0) &= V_0(x) \end{aligned}$$

Summary

For $T \in (0, \infty]$ and $\tau \in (0, 1]$, consider the following optimal control problem:

$$\underset{u \in L^2(0,T)}{\text{minimize}}\ \mathbb{E}\left[\int_0^T e^{-\tau t} C(x(t), u(t)) dt\right] \quad \text{subject to} \quad \begin{cases} x'(t) = f(x(t), u(t)) + \sigma \dot{w}(t) \\ x_0 = x \in \mathbb{R}^n. \end{cases}$$

where $w(t)$ is a Gaussian brownian motion.
Goal: Find (learn) an optimal feedback control

$$\pi^* : \mathbb{R}^n \to \mathcal{U}, \quad \text{such that} \quad u^*(t) = \pi^*(x(t)) \text{ is optimal.}$$

The **value function** for $T = \infty$:

$$V(x) = \min_{\pi(\cdot)} \mathbb{E}\left[\int_0^\infty e^{-\tau t} C(x(t), u(t)) dt\right]$$

$$\tau V(x) - \sigma \Delta V(x) = H(x, \nabla V(x))$$

with $H(x, p) := \min_u \{p \cdot f(x, u) + C(x, u)\}$.
Value iteration: Set an initial guess $V_0(x)$, and then improve the approximation by taking $t >> 1$ in

$$\begin{aligned} &\partial_t V(x, t) - \sigma \Delta V(x, t) + \tau V(x, t) = H(x, \nabla_x V(x, t)) \\ &V(x, 0) = V_0(x) \end{aligned}$$

The Q-function

For each state x and control $u \in \mathcal{U}$. We define the Q-function as

$$Q(x,u) := C(x,u) + \gamma V(f(x,u))$$

If we known the Q-function, we can construct the optimal policy without using the dynamics.

Optimal feedback control

$$u^*(x) = \text{argmin}_u Q(x,u)$$

Dynamic Programming for Q

For all $t = 0, 1, 2, 3, \ldots$

$$Q(x,u) = C(x,u) + \gamma \min_v Q(f(x,u), v).$$

Q-learning, Watkins 1989

- Initialize an arbitrary Q-function, $\widehat{Q}_0(x, u)$.
- Run a number N of experiments (episodes) of T steps each one.
- In total there will be $N\,T$ number of steps.
- **ε-greedy policy:**

$$u_t = min_u \widehat{Q}_k(x_t,\, u) \quad \text{with probability } 1 - \varepsilon$$

and u_k is chosen randomly with probability ε (exploration vs exploitation).

- **Improvement of $\widehat{Q}(x_t, u_t)$:**

$$\widehat{Q}_{k+1}(x_t, u_t) = (1 - \alpha)\widehat{Q}_k(x_t, u_t) + \alpha\big(C_k + \gamma \min_u \widehat{Q}_k(x_{t+1}, u)\big)$$