Safe Reinforcement Learning with Provable Guarantees

Dr. Mayank S JHA

Associate Professor,
CRAN,
CNRS (UMR 7039) ,
Université de Lorraine, Nancy, France.

Talk: GDR MACS Action – Workshop on Data Driven Control & Analysis

Table of Contents

Acknowledgements

PhD Students :

Former: Dr. S Kanso

Current: Theo Rutschke, Satya Marthi

Collaborators:

Internal

Didier Thielliol (CRAN, Univ of Lorraine)

Hugues Garnier (CRAN, Univ of Lorraine)

External

Bahare Kiumarsi (Univ. of Michigan, USA)

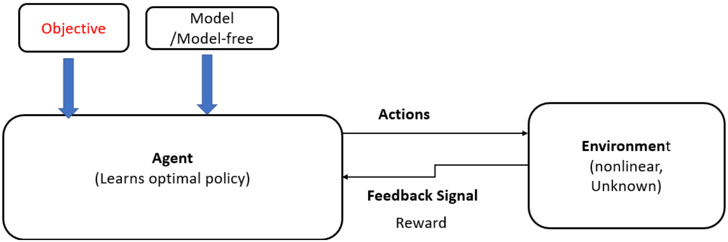
Kyriakos Vamvoudakis (Georgia Tech, USA)

Gautam Biswas (Vanderbilt University, USA)

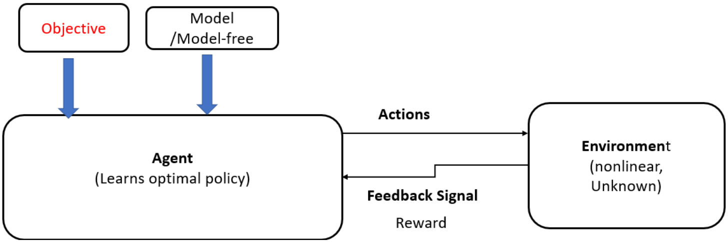
Chetan Kulkarni (NASA Ames Research Center, USA)

Table of Contents

Reinforcement Learning Architecture



Reinforcement Learning Architecture



Reinforcement Learning: Automatic Control

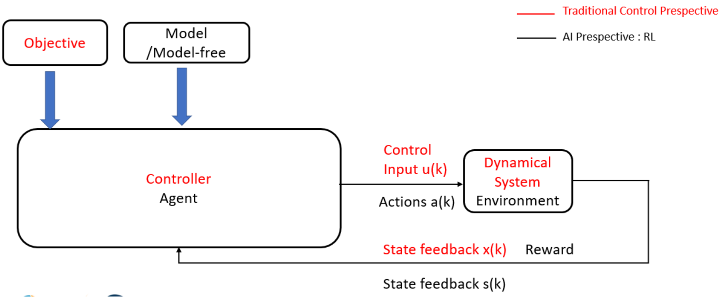


Table of Contents

RL: Discrete time optimal control

System

$$x_{k+1} = f(x_k) + g(x_k)u(x_k) \tag{1}$$

- $x_k \in \Omega \subset \mathbb{R}^n$ is the state variable vector
- Ω being a compact set
- $u(x_k) \in U \subset \mathbb{R}^m$ is the control input vector
- $f(x)$ is C^1 and $x = 0$ is an equilibrium state such that $f(0) = 0$ and $g(0) = 0$.

Note: $u(x_k)$ will be denoted as u_k.

RL: Discrete time optimal control

Control law/ Policy

A control policy is a function from state space to control space $\pi(\cdot) : \mathbb{R}^n \to \mathbb{R}^m$, that defines for every state x_k , a control action:

$$u_k = \pi(x_k) \tag{2}$$

- Such mappings $\to$ feedback controllers.
- Example: linear state-variable feedback $u_k = \pi(x_k) = -Kx_k$

RL: Discrete time optimal control

Goal directed performance

Cost-to-go is a sum of (discounted) future costs from the current time k into the infinite horizon future under a prescribed control law $u_k = \pi(x_k)$:

$$J(x_k, u_k) = \sum_{n=k}^{\infty} r(x_n, u_n) \tag{3}$$

where $r(x_n, u_n)$ is the utility function defined as:
$r(x_n, u_n) = x_n^T Q x_n + u_n^T R\ u_n$

- Q symmetric positive semi-definite matrix $Q = Q^T \geq 0$
- R is a symmetric positive definite matrix $R = R^T > 0$.

RL: Discrete time optimal control

Cost (given a prescribed policy $u_k = \pi(x_k)$)

$$V_\pi(x_k) = \sum_{n=k}^{\infty} r(x_n, u_n), \forall x_k$$

Bellman Eq/ Nonlinear Lyapunov Eq (Recursive):

$$V_\pi(x_k) = r(x_k, u_k) + V_\pi(x_{k+1})$$

Hamiltonian:

$$H(x_k, u_k, V_\pi) = r(x_k, u_k) + V_\pi(x_{k+1}) - V_\pi(x_k)$$

Optimal Cost:

$$V^*(x_k) = \min_{u_k \in U} \left(r(x_k, u_k) + V_\pi(x_{k+1})\right)$$

Bellman principle:

$$V^*(x_k) = \min_{u_k \in U} \left(r(x_k, u_k) + V^*(x_{k+1})\right)$$

Backwards in Time!!

Optimal control (policy):

$$\pi^*(x_k) = \arg\min_{u_k \in U} \left(r(x_k, u_k) + V^*(x_{k+1})\right)$$

RL: Discrete time optimal control

Bellman principle: *(DT Hamilton-Jacobi-Bellman Equation)*

$$
\begin{aligned}
V^*(x_k) &= \min_{u_k \in U} \left(r(x_k, u_k) + V^*(x_{k+1}) \right) \\
&= \min_{u_k \in U} \left(x_k^T Q x_k + u_k^T R\, u_k + V^*(x_{k+1}) \right) \\
&= \min_{u_k \in U} \left(x_k^T Q x_k + u_k^T R\, u_k + V^*(f(x_k) + g(x_k) u_k) \right)
\end{aligned}
$$

Optimal control (policy):

$$
\pi^*(x_k) = \arg\min_{u_k \in U} \left(r(x_k, u_k) + V^*(x_{k+1}) \right)
$$

$$
\pi^*(x_k) = u_k^* = (-1/2) R^{-1} g^T(x_k) \frac{\partial V^*(x_{k+1})}{\partial x_{k+1}}
$$

DT Policy Iteration

Initialization

Select any *stabilizing* /admissible control policy: $\pi_j(x_k)$

Policy Evaluation

Determine the *Value* under the current policy using Bellman Equation/Nonlinear Lyapunov Eq.

$V_{j+1}(x_k) = r(x_k, \pi_j(x_k)) + V_{j+1}(x_{k+1})$; $V_{j+1}(0) = 0$

Policy Improvement

Determine an improved policy

$$\pi_{j+1}(x_k) = \arg\min_{u_k \in U} \left(r(x_k, u_k) + V_{j+1}(x_{k+1})\right)$$

DT Policy Iteration

Initialization

Select any *stabilizing* /admissible control policy: $\pi_j(x_k)$

Policy Evaluation

Determine the *Value* under the current policy using Bellman Equation/Nonlinear Lyapunov Eq.

$$V_{j+1}(x_k) = r(x_k, \pi_j(x_k)) + V_{j+1}(x_{k+1}) \; ; \; V_{j+1}(0) = 0$$

Policy Improvement

Determine an improved policy

$$\pi_{j+1}(x_k) = \arg\min_{u_k \in U} \left(r(x_k, u_k) + V_{j+1}(x_{k+1}) \right)$$

DT Policy Iteration

Initialization

Select any *stabilizing* /admissible control policy: $\pi_j(x_k)$

Policy Evaluation

Determine the *Value* under the current policy using Bellman Equation/Nonlinear Lyapunov Eq.

$$V_{j+1}(x_k) = r(x_k, \pi_j(x_k)) + V_{j+1}(x_{k+1}) \; ; \; V_{j+1}(0) = 0$$

Policy Improvement

Determine an improved policy

$$\pi_{j+1}(x_k) = \arg\min_{u_k \in U} \left(r(x_k, u_k) + V_{j+1}(x_{k+1}) \right)$$

DT Policy Iteration: Observations

- Initial policy must be stabilizing.
- Policy Iteration (Howard, 1960; Leake and Liu, 1967) $\Rightarrow$
  - $V_{j+2}(x_k) \leq V_{j+1}(x_k)$
- As $j \rightarrow \infty$:
  - $V_j(x_k) \rightarrow V^*(x_k)$
  - $\pi_j \rightarrow \pi^*$
- Convergence to optimal cost and thus, optimal control policy.

Forward-in-time Learning

Temporal Difference Error (TD error):
$e_k = r(x_k, \pi_{x_k}) + V_\pi(x_{k+1}) - V_\pi(x_k)$

- RHS is DT Hamiltonian
- If Bellman Eq holds, e_k is zero.
- Linear in x.
- Thus, given a policy $\pi(x)$, Least Square based solution at each time k for $e_k = 0$.

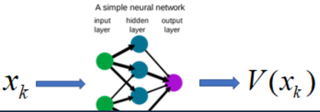
NN based approximation

Value Function approximation (VFA): Neural Networks

- Value function is sufficiently smooth over compact space
- Consider dense basis set $\{\phi_i(x)\}$ with basis vector (Weierstrass Theorem):
  $\phi(x) = [\varphi_1(x)\varphi_2(x)...\varphi_L(x)] : \mathbb{R}^n \to \mathbb{R}^L$

$$V_\pi(x) = \sum_{i=1}^{L} w_i \varphi_i(x) = W^T \phi(x)$$

Substituting in Bellman TD equation:
$$e_k = r(x_k, \pi_{x_k}) + W^T \phi(x_{k+1}) - W^T \phi(x_k)$$



Online Policy Iteration

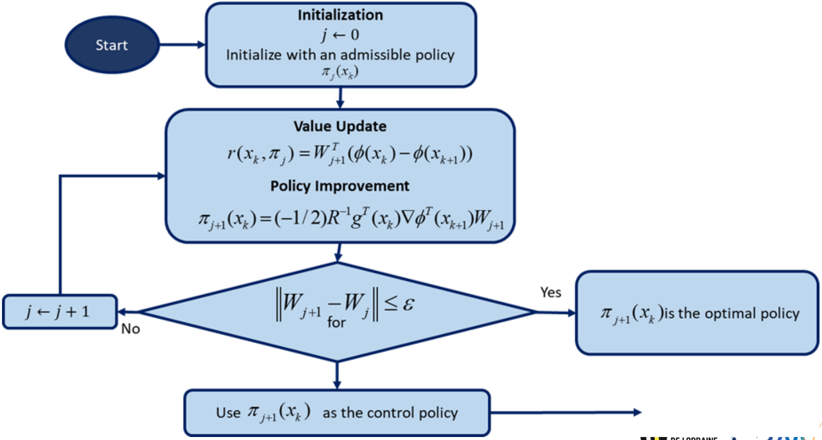


Figure: Online PI

Safe RL Motivations

Conventional RL:

- Stability
- Optimality: Performance, energy consumption etc.

Does NOT:

- ensure SAFETY.

Poses "Threat"

- during Exploration: **data collection** phase.
- during Exploitation: **learning phase.**

Treatment remains different from SATURATION

- nearness to safety frontier also important
- action at time k may leads to violation at $k+l$
- may vary with environment
- unmodelled effects, stochastic etc.

Safe RL Motivations

Conventional RL:

- Stability
- Optimality: Performance, energy consumption etc.

Does NOT:

- ensure SAFETY.

Poses "Threat"

- during Exploration: **data collection** phase.
- during Exploitation: **learning phase.**

Treatment remains different from SATURATION

- nearness to safety frontier also important
- action at time k may leads to violation at $k+l$
- may vary with environment
- unmodelled effects, stochastic etc.

Safe RL Motivations

Conventional RL:

- Stability
- Optimality: Performance, energy consumption etc.

Does NOT:

- ensure SAFETY.

Poses "Threat"

- during Exploration: **data collection** phase.
- during Exploitation: **learning phase.**

Treatment remains different from SATURATION

- nearness to safety frontier also important
- action at time k may leads to violation at $k + l$
- may vary with environment
- unmodelled effects, stochastic etc.

Safe RL Motivations

Conventional RL:

- Stability
- Optimality: Performance, energy consumption etc.

Does NOT:

- ensure SAFETY.

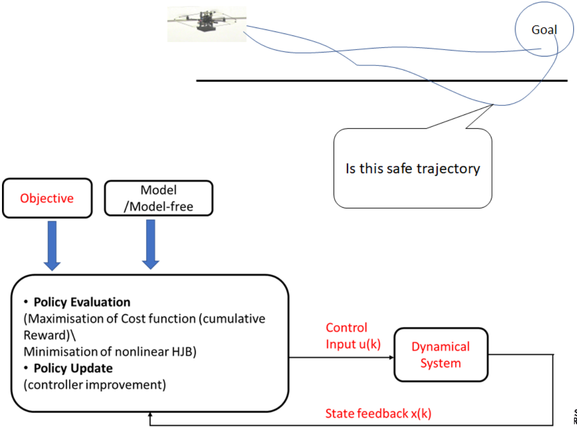
Poses "Threat"

- during Exploration: **data collection** phase.
- during Exploitation: **learning phase.**

Treatment remains different from SATURATION

- nearness to safety frontier also important
- action at time k may leads to violation at $k + l$
- may vary with environment
- unmodelled effects, stochastic etc.

Safe Learning



Safe Learning

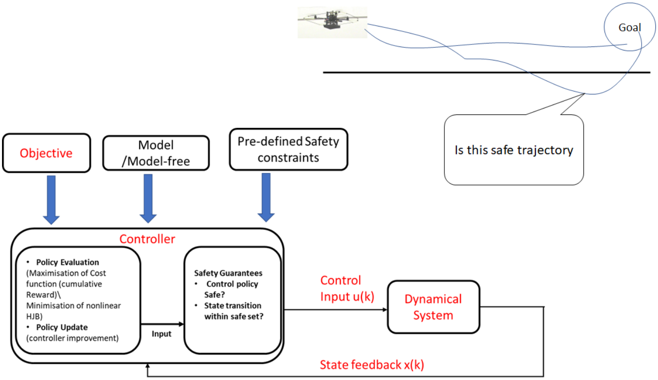


Table of Contents

System

$$x_{k+1} = f(x_k) + g(x_k)u(x_k) \tag{1}$$

where:

- $x_k \in \Omega \subset \mathbb{R}^n$ states of the system
- $u(x_k) \in U \subset \mathbb{R}^m$ are the and the control input
- U denotes the set of all admissible control inputs
- $f(x_k) \in \mathbb{R}^n$represents the drift dynamics
- $g(x_k) \in \mathbb{R}^{n \times m}$ is the input dynamics.
- $f(x_k)$ is C^1 and $x = 0$ is an equilibrium state such that $f(0) = 0$ and $g(0) = 0$.
  It is assumed that system (1) is stabilizable on a prescribed set $\Omega \in \mathbb{R}^n$.

Safe Set

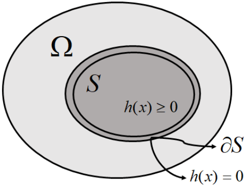
Definition

The safe set $\mathcal{S}$ and its boundary $\partial\mathcal{S}$ can be mathematically defined as:

$\mathcal{S} = \{x \in \Omega | h(x) \geqslant 0\}$

$\partial\mathcal{S} = \{x \in \Omega | h(x) = 0\}$

where $h(x) : \mathbb{R}^n \to \mathbb{R}$ belongs to C^1 and $h(x) > 0$ represents the admissible state space that respects the safety requirements.



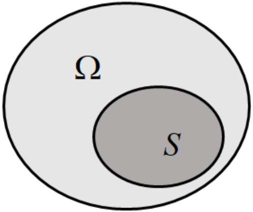
Strategy

Definition. *A set $\mathcal{S} \in \Omega$ is control invariant set if*
$x_k \in \mathcal{S} \Rightarrow \exists u_k \in U \quad | \quad x_{k+1} \in \mathcal{S} \quad \forall k \in \mathbb{Z}^+$
where $x_{k+1} = f(x_k) + g(x_k)u_k$
with $x_k \in \Omega \subset \mathbb{R}^n$ *and* $u_k \in U \subset \mathbb{R}^m$

Strategy:

Learning control law (sequence of control actions)

- *that ensures positive invariant property of safe set S,*
- *Optimality : performance + energy consumption etc.*

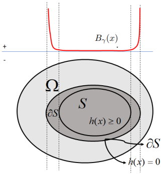


Barrier Function

Definition

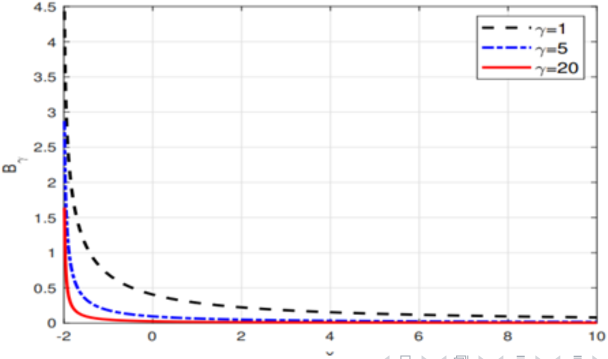
BF candidate (Ames et al., 2016; Brunke et al., 2022; Wabersich et al., 2023) $B_\gamma(x) : \mathcal{S} \to \mathbb{R}$ satisfies the following properties:

1. $B_\gamma(x) > 0 \;\; \forall x \in \mathcal{S}$
2. $B_\gamma(x) \to \infty \;\; \forall x \in \partial\mathcal{S}$
3. $B_\gamma(x)$ is monotonically decreasing $\forall x \in \mathcal{S}$



Barrier Function Candidate

$$B_\gamma(x_k) = -\log\left(\frac{\gamma h(x_k)}{\gamma h(x_k)+1}\right)$$



Control Barrier Function CBF

Definition

Control Barrier functions for DT systems Agrawal and Sreenath, 2017: A function $B_\gamma(x) : \mathcal{S} \to \mathbb{R}$ is a CBF on the safe set $\mathcal{S}$ and for the nonlinear DT control system (1) if there exists:

1. locally Lipschitz class $\mathcal{K}$ functions α_1 and α_2 such that

$$\frac{1}{\alpha_1(h(x_k))} \leqslant B_\gamma(x_k) \leqslant \frac{1}{\alpha_2(h(x_k))}, \quad \forall x \in int\mathcal{S} \tag{4}$$

2. a safe control input $u_k \in \mathcal{U}^s$, $\forall x \in intS$ such that

$$\Delta B_\gamma(x_{k+1}, x_k) := B_\gamma(f(x_k) + g(x_k)\,u_k) - B_\gamma(x_k) \leqslant \alpha_3(h(x_k)) \tag{5}$$

Control Barrier Function CBF

These conditions imply:

- u_k maintains the barrier function $B_\gamma(x_k) \geqslant 0,\ \forall k \in \mathbb{Z}^+$ given $B_\gamma(x_0) \geqslant 0$
- safe input maintains the trajectory of system within the safe set $\mathcal{S}$ if the initial state x_0 is within $\mathcal{S}$.

Safety Aware Control design

Modified Cost

Classical cost-to-go modified and augmented with a CBF candidate as:

$$\min_{u \in U} \quad J_s(x_k, u) = \sum_{n=k}^{\infty} r_s(x_n, u_n) = \sum_{n=k}^{\infty} x_n^T Q x_n + u_n^T R\, u_n + B_\gamma(x_n)$$

$B_\gamma(x) : \mathcal{S} \to \mathbb{R}$ is augmented utility function $r_s(x_k, u_k)$ as:

$$r_s(x_k, u_k) = x_k^T Q x_k + u_k^T R\, u_k + B_\gamma(x_k) \tag{6}$$

The candidate CBF $B_\gamma(x)$ is sensitive to a coefficient γ that models the relative importance of the CBF to the utility function.

Safe Admissible policy and strict interiority

Definition

Safe admissible control policy: $\mathcal{U}^a = U \cap \mathcal{U}^s$

Definition

Strict interiority of initial condition:
The initial condition of system (1) remains strictly in the interior of the safe set $\mathcal{S}$, i.e. $x_0 \in int\mathcal{S}$.

Assumption

$\mathcal{U}^a = U \cap \mathcal{U}^s \neq \emptyset$

Safety Analysis

Lemma

Given an arbitrary admissible control policy $u^{(1)}(x_k) \in \mathcal{U}^a$ *(denoted as* u_k^1*), if there exists a positive definite value function* $W(x) \in \mathcal{C}^1$ *on* Ω *such that*

$$\frac{1}{2}\left(f(x_k) + g(x_k)u_k^{(1)} - x_k\right)^T \nabla^2 W_k \left(f(x_k) + g(x_k)u_k^{(1)} - x_k\right)$$
$$+ \nabla W_k^T \left(f(x_k) + g(x_k)u_k^{(1)} - x_k)\right)$$
$$+ \left(x_k^T Q x_k + (u_k^{(1)})^T R u_k^{(1)} + B_\gamma(x_k)\right) = 0$$

and $W(x_0, u_0^{(1)}) = J_s(x_0, u_0^{(1)})$.
Then, $W(x_k, u_k^{(1)})$ *is the value function of the system for all* $k = 0, ..., \infty$ *applying the feedback control input* $u_k^{(1)}$ *and* $W(x_k, u(x_k)) = J_s(x_k, u(x_k))$.

G-SHJB

Definition

G-SHJB Generalised Safety-aware Hamiltonian Jacobi Bellman (G-SHJB) for DT systems

$$
\begin{aligned}
&(1/2)\Delta x^T \nabla^2 W(x) \Delta x + \nabla W(x)^T \Delta x \\
&+ x^T Q x + u(x)^T R u(x) + B_\gamma(x) = 0 \\
&W(0) = 0 \\
&\Delta x = f(x) + g(x)u(x) - x
\end{aligned}
\tag{7}
$$

G-SHJB

- The G-SHJB with boundary condition can be used to solve infinite-time problems.
- Given an admissible control input, solve G-SHJB to obtain the value function $W(x)$
- Then, $W(x_0)$ to calculate the cost of the admissible control in J_s .

However, the objective is to improve the performance of the system and guarantee safety over time by updating the control law.

G-SHJB

Definition

G-SHJB Hamiltonian

$$\begin{aligned} &H(x, W(x), u(x), B_\gamma(x)) = \\ &(1/2)\Delta x^T \nabla^2 W(x) \Delta x + \nabla W(x)^T \Delta x \\ &+ x^T Q x + u(x)^T R u(x) + B_\gamma(x) \end{aligned} \tag{8}$$

Policy Improvement

$$\begin{aligned} \frac{\partial H^i\left(x, W^{(i)}(x), u^{(i+1)}, B_\gamma(x)\right)}{\partial u^{(i+1)}} &= 0 \\ u^{(i+1)} = \frac{-g^T(x)\left[\nabla W^{(i)} + \nabla^2 W^{(i)}(f(x) - x)\right]}{\left[g^T(x)\nabla^2 W^{(i)} g(x) + 2R\right]} \end{aligned} \tag{9}$$

Bounded CBF at each step

Lemma

Consider the policy improvement step (9) with corresponding control policy sequence $\{u_k^{(i)}\}_{i=1}^{i+1} = \{u_k^1, u_k^2 ... u_k^{(i+1)}\}$ and corresponding sequence of value functions due to sequential minimization $\{W_k^{(i)}(x_k, u_k^{(i)})\}_{i=1}^{i=i+1} = \{W_k^{(1)}, W_k^{(2)} ... W_k^{(i+1)}\}$. Then, the CBF is bounded at each sequential step i.

Invariance of Safe Set

Theorem

Consider $B_\gamma(x)$, Safety aware cost, and the control policy obtained through sequential steps (9) , then the safe set $\mathcal{S}$ is invariant along the system trajectories.
That is, if the initial state lies within the interior of safe set $\mathcal{S}$, i.e. $x_0 \in int\mathcal{S}$, then $x_k \in int\mathcal{S}\ \forall k \in \mathbb{Z}^+$.

Stability analysis

Theorem

Assuming $x = 0$ is the equilibrium, within the safe region $\mathcal{D} \subset \mathrm{R}$, the CBF candidate $B_\gamma(x)$, cost to go and consider the policy improvement step (9) with corresponding control policy sequence $\{u_k^{(i)}\}_{i=1}^{i+1}$ $=\{u_k^1, u_k^2 ... u_k^{(i+1)}\}$ along with corresponding sequence of positive definite value functions due to sequential minimization $\{W_k^{(i)}(x_k, u_k^i)\}_{i=1}^{i=i+1} = \{W_k^{(1)}, W_k^{(2)} ... W_k^{(i+1)}\}$, then the control inputs obtained from policy sequence asymptotically stabilizes the system within the safe region $\mathcal{D}$.

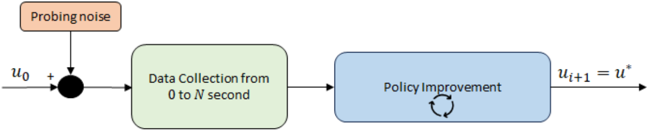
$$\Delta W_k^{(i)} \leqslant -x_k^T Q x_k \leqslant -\lambda_{\min}(Q) \|x_k\|^2$$

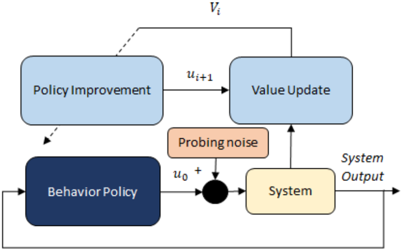
Optimality Analysis

Theorem

Given an initial admissible control $u_k^0 \in \mathcal{U}^a$, solving G-SHJB in an iterative manner and improving the control law using (9), the sequence of solutions i.e. sequence of value functions $W_k^{(i)}$ and sequence of control laws $u_k^{(i)}$ converge, respectively, to the optimal value function W_k^* and corresponding optimal safe control law u_k^* i.e. $W_k^{(i)} \rightarrow W_k^*$ and $u_k^{(i)} \rightarrow u_k^*$.

On-policy vs Off-policy





Off-policy Approach

Off-policy Equation

$$x_{k+1} = f_k + g_k u_k^{(i)} + g_k(u_k - u_k^{(i)}) \tag{10}$$

- **Behaviour policy** is a safe policy that is applied to the system to execute data collection under various scenarios including those that remain close to boundary of safe set.
- **Target policy** is the policy that is improved towards the optimal policy using the data collected.

Off-policy S-GHJB

Theorem

The successive differences of value function W^i along an off-policy based system trajectory $(f, g, u^{(i)}, u)$ can be derived as:

$$\begin{aligned} W_{k+1}^{(i)} - W_k^{(i)} = & - x_k^T Q x_k - B_\gamma(x_k) - u_k^{(i)T} R u_k^{(i)} \\ & - 2u_k^{(i+1)T} R(u_k - u_k^{(i+1)}) \end{aligned}$$

NN based approximation

NN approximation

$$\hat{W}_k^{(i)} := \hat{W}^{(i)}(x_k) = \hat{\Omega}_c^{(i)T}\Phi(x) = \sum_{j=1}^{Lc} \omega_j^{\Omega_c^{(i)}} \phi_j(x) \tag{11}$$

$$\hat{u}^{(i)}(x_k) := \hat{u}_k^{(i)} = \hat{\Omega}_a^{(i)T}\Psi(x) = \sum_{j=1}^{La} \omega_j^{\Omega_a^{(i)}} \sigma_j(x) \tag{12}$$

Off-policy temporal difference

NN based expression Off-policy G-SHJB

$$e_k^{(i)} = \hat{\Omega}_c^{(i)T}\Phi(x_{k+1}) - \hat{\Omega}_c^{(i)T}\Phi(x_k) + \left(x_k^T Q x_k + u_k^{(i)^T} R u_k^{(i)} + B_\gamma(x_k)\right) + 2\sum_{j=1}^{m} \rho_j \hat{\Omega}_{a,j}^{(i)T}\Psi(x_k) v_j^{(i)} \quad (13)$$

Off-policy temporal difference

Least Square Problem: $\widehat{\mathbf{W}}^{(i)T} H^{(i)} = Y^{(i)}$

- $\widehat{\mathbf{W}}^{(i)T} \in \mathrm{R}^{1\times(Lc+mLa)}$ as $\widehat{\mathbf{W}}^{(i)T} = \left[\hat{\Omega}_c^{(i)}, \hat{\Omega}_{a,1}^{(i)}, \hat{\Omega}_{a,2}^{(i)}, ..., \hat{\Omega}_{a,m}^{(i)}\right]$,
- independent data vector $H^{(i)} \in \mathrm{R}^{(Lc+mLa)\times N}$ as $H^{(i)} = \left[h_1^{(i)} h_2^{(i)} ... h_N^{(i)}\right]$ wherein $j \in (1, ...N)$ $h_j^{(i)} = \left[\overline{\theta}, 2\rho_1\Psi(x_k)v_1^{(i)}, ..., 2\rho_m\Psi(x_k)v_m^{(i)}\right] \in \mathrm{R}^{(Lc+mLa)}$
- dependant data vector $Y^{(i)} \in \mathrm{R}^{1\times N}$ as $Y^{(i)} = \left[y_1^{(i)}, y_2^{(i)}, ..., y_N^{(i)}\right]$ wherein the data collected $\forall k \in (1, ..., N)$ is given by the observed reward (augmented utility) $y_k^{(i)} = -r_{s,k}^{(i)}$.

Off-policy temporal difference

Least Square Solution

$$\widehat{\mathbf{W}}^{(i)T} = \left(H^{(i)} H^{(i)T}\right)^{-1} H^{(i)} Y^{(i)} \tag{14}$$

The unique solution exists if the number of points of data collection is greater or equal to the order of approximation or $N > (Lc + mLa)$.

Algorithm

Algorithm 1: Off-policy safe policy iteration

1: **procedure** DATA COLLECTION
2: Employ an initial noisy stabilizing control policy $\mathcal{U}^a = U \cap \mathcal{U}^s$ until number of points of data collection is greater or equal to the order of approximation or $N > (Lc + mLa)$.
3: **end procedure**
4: **procedure** OFF-POLICY POLICY EVALUATION AND IMPROVEMENT
5: **Policy Iteration** Solve for $\hat{\mathbf{W}}$ and terminate the process when the following approximation error is within a prefixed convergence threshold ϵ , chosen sufficiently small. $\sum_{j=1}^{m} \left\| \hat{W}_{i,j} - \hat{W}_{i-1,j} \right\| \leq \epsilon$
6: **Update** If not, let $i \leftarrow i + 1$ and go to step 5.
7: **Application** Update the controller using learned weights and apply safe optimal policy to the system.
8: **end procedure**

Simulations

Car model

$$\begin{bmatrix} y_{k+1} \\ v_{k+1} \\ \phi_{k+1} \\ \psi_{k+1} \end{bmatrix} = \begin{bmatrix} 1 & Ts & v_{l0}.Ts & 0 \\ 0 & 1+(-\frac{C_f+C_r}{Mv_{l0}})Ts & 0 & (\frac{bC_r-aC_f}{Mv_{l0}} - v_{l0})Ts \\ 0 & 0 & 1 & Ts \\ 0 & (\frac{bC_r-aC_f}{I_z v_{l0}})Ts & 0 & 1 \end{bmatrix} \begin{bmatrix} y_k \\ v_k \\ \phi_k \\ \psi_k \end{bmatrix} + \begin{bmatrix} 0 \\ \frac{C_f}{M} \\ 0 \\ a\frac{C_f}{I_z} \end{bmatrix} .Ts.u_k + \begin{bmatrix} 0 \\ 0 \\ -1 \\ 0 \end{bmatrix} .Ts.d_k \tag{15}$$

Simulations

Safety aware Reward/Utility function

$$r_s(x_k, u_k) = x_k^T Q x_k + u_k^T R u_k - m(log(\frac{\gamma(x_{1,k} + y_{max})}{\gamma(x_{1,k} + y_{max}) + 1}) + log(\frac{\gamma(-x_{1,k} + y_{max})}{\gamma(-x_{1,k} + y_{max}) + 1}))$$

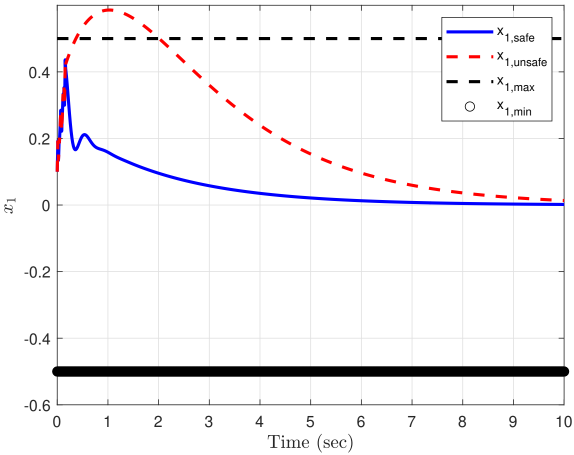
- y_k and v_k are lateral displacement and its velocity
- y_{max} expresses the absolute value of maximum safe displacement from the center of the road.
- ϕ_k is error yaw angel and ψ_k is its derivative,
- u_k is the steering angle,
- d_k is the desired yaw rate obtained from the curvature of the

Simulations

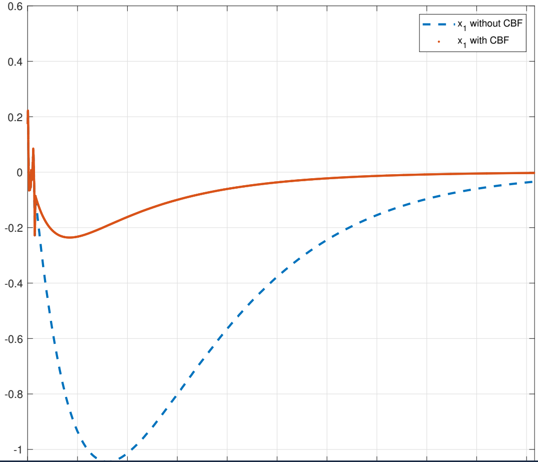
Actor and Critic NNs

$$\Phi(x) = [x_1^2 \ x_2^2 \ x_3^2 \ x_4^2 \ x_1x_2 \ x_1x_3, \ x_1x_4 \ x_2x_3$$
$$x_2x_4 \ x_3x_4 \ (x_1 - y_{max})^2 \ {x_1}^4, {x_2}^4]$$
$$\Psi(x) = [x_1 \ x_2 \ x_3 \ x_4]^T$$

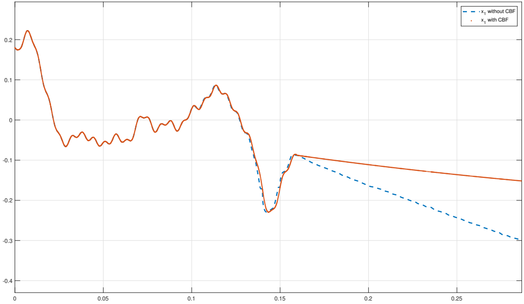
Lateral displacement zoomed



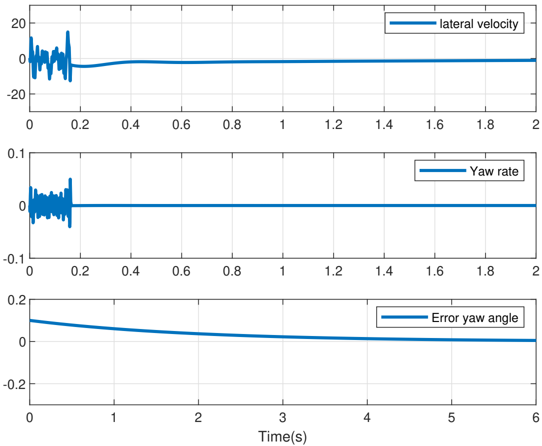
Lateral displacement



Lateral displacement zoomed



Other states



Conclusions

- Model free approach (data based)
- Optimality
- Stability
- Safety during operation—OK!
- Safety during EXPLORATION ???
- Initial admissible policy ?????

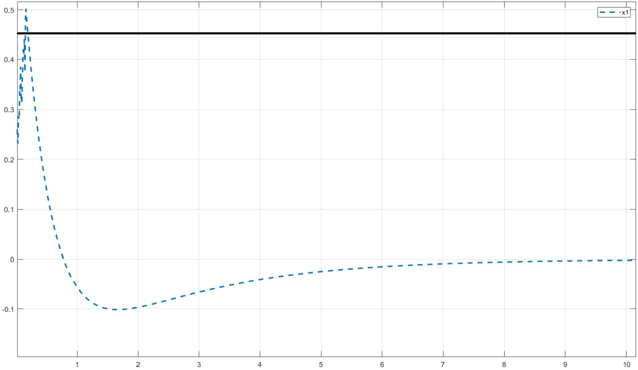
Mayank Shekhar Jha, Bahare Kiumarsi, Off-policy safe reinforcement learning for nonlinear discrete-time systems, Neurocomputing, Elsevier, Volume 611, 2025, 128677, ISSN 0925-2312, https://doi.org/10.1016/j.neucom.2024.128677.

Jha, M. S., Kiumarsi, B., Theilliol, D. (2024, July). Safe Reinforcement Learning Based on Off-Policy Approach for Nonlinear Discrete-Time Systems. In 2024 American Control Conference (ACC) (pp. 1574-1579). IEEE.

SHhhhhhh.....!

BEHIND SCENES!!!

Safety FAILURE during Exploration!!!!



Safety FAILURE during Exploration!!!!

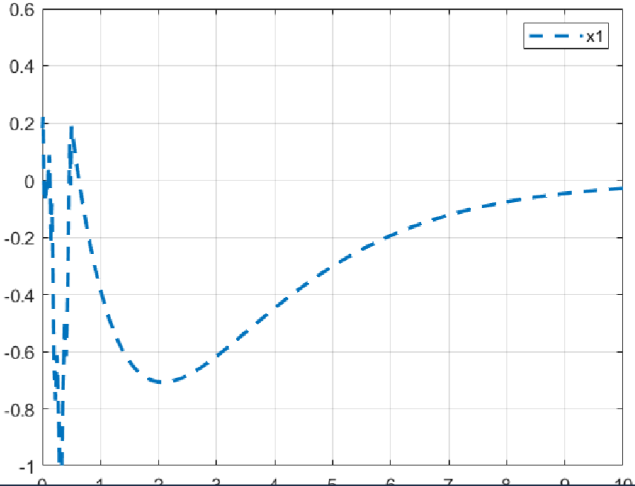


Table of Contents

System under exploration

Under exploration noise

$$\dot{x} = f(x) + g(x)(u + e) \tag{16}$$

$$\dot{x} = f(x) + g(x)u + p(x)w \tag{17}$$

Key Idea: The system (16) is input-to-state stabilizable if and only if there exists an ISS-CLF.

Safe Exploration

Robust QP Problem

Find the control u_{safe} and the relaxation variable δ that satisfy

$$\begin{aligned} \min_{u_{safe},\delta} \quad & \frac{1}{2}(u_{safe}^T u_{safe} + \ell \delta^T \delta) \\ s.t. \quad & F_1 = a_1 + b_1(u + u_{safe}) + \delta \leq 0 \\ & F_2 = a_2 + b_2(u + u_{safe}) \leq 0 \end{aligned} \tag{18}$$

with

$$\begin{aligned} a_1 &= L_f V(x) + L_g V(x) \eta^{-1}(x) + \alpha(x) \\ a_2 &= L_f B_\gamma(x) + L_g B_\gamma(x) e(t) - \alpha_B(h(x)) \\ b_1 &= L_g V(x) \\ b_2 &= L_g B_\gamma(x) \end{aligned}$$

The gradients of the R-CBF B_γ and ISS-CLF V, α, α_B, η^{-1} are

Safe off-policy

$$\dot{x} = f(x) + g(x)[u_0 + e + u_{safe}] \tag{19}$$

The initial policy $u_{0,random}$ is randomly generated then by adding the solution of the Robust-QP problem u_{safe}, $u_{0,random}$ is modified to ensure that the resulting control policy u_0 is both safe and admissible. Then, above can be rewritten as

$$\dot{x} = f(x) + g(x)u_i + g(x)\nu_i \tag{20}$$

where $\nu_i = u_0 + e + u_{safe} - u_i = u_s - u_i$ and $u_{noisy} = u_0 + e$.

Lemma

The weights $\hat{C}_i$ and $\hat{U}_i$ can be obtained by solving the following least-squares (LS) equation:

$$\tilde{\Theta}_i^N \begin{bmatrix} vec(\hat{C}_i) \\ vec(\hat{U}_i^T) \end{bmatrix} = \tilde{E}_i^N \tag{21}$$

for $N > N_1 + mN_2$ and

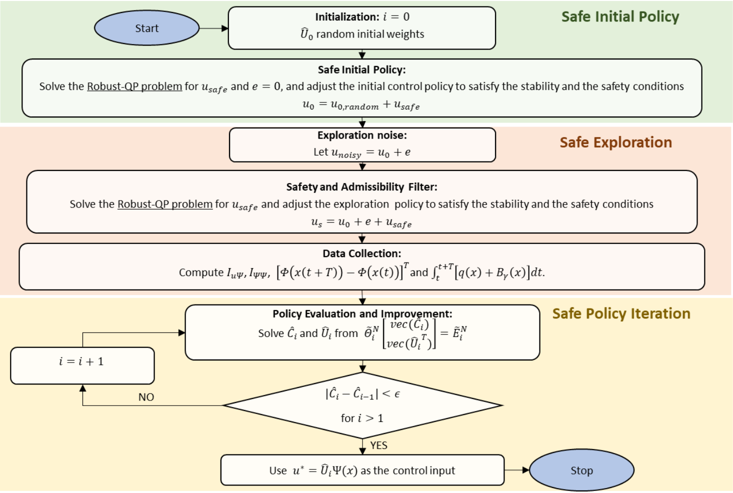
$$\begin{aligned} \tilde{\Theta}_i^N &= [\tilde{\Theta}_i(t_1), \ldots, \tilde{\Theta}_i(t_N)]^T \\ \tilde{E}_i^N &= [\tilde{E}_i(t_1), \ldots, \tilde{E}_i(t_N)]^T \end{aligned} \tag{22}$$

where

$$\tilde{\Theta}_i(t) = \begin{bmatrix} [\Phi(x(t+T)) - \Phi(x(t))]^T \\ 2[I_{u\Psi}(R \otimes I_{N_2}) - I_{\Psi\Psi}(\hat{U}_{i-1}^T R \otimes I_{N_2})] \end{bmatrix}^T \tag{23}$$

$$\tilde{E}_i(t) = -I_{\psi\psi}[\hat{U}_{i-1}^T \otimes \hat{U}_{i-1}^T] vec(R) - \int_t^{t+T} [q(x) + B_\gamma(x)] dt \tag{24}$$

End to End Safe Learning-CT



Safe Initialization, Exploration, and Exploitation (operation)

Jet engine surge and stall dynamics

Consider the following jet engine surge and stall dynamics

$$\dot{x}_1 = -\sigma x_1^2 - \sigma x_1 \left(2x_2 + x_2^2\right)$$
$$\dot{x}_2 = -a x_2^2 - b x_2^3 - \left(u + 3x_1 x_2 + 3x_1\right)$$

$$\boxed{\begin{aligned}\sigma &= 0.35\\ a &= 1.4\\ b &= 0.5\end{aligned}}$$

- x_1 is the normalized rotating stall amplitude
- x_2 is the deviation of the scaled annulus-averaged flow with $-1.1 < x_2 < 0.45$
- u is the deviation of the plenum pressure rise and is considered as the control input

- Initial states : $x_0 = [1 \quad -1]^T$
- Initial Actor Weights : $\widehat{U}_0 = [-3 \quad 0 \quad 0 \quad 0 \quad 0 \quad 0 \quad 0 \quad 0 \quad 0 \quad 0 \quad 0]$
- Probing noise : $e(t) = 2\sum \omega \times \sin([1\ 3\ 7\ 11\ 13\ 15\ 17\ 19\ 21\ 23\ 25\ 27\ 29] \times t)$
  ω random Gaussian noise

Example

Exploration Phase

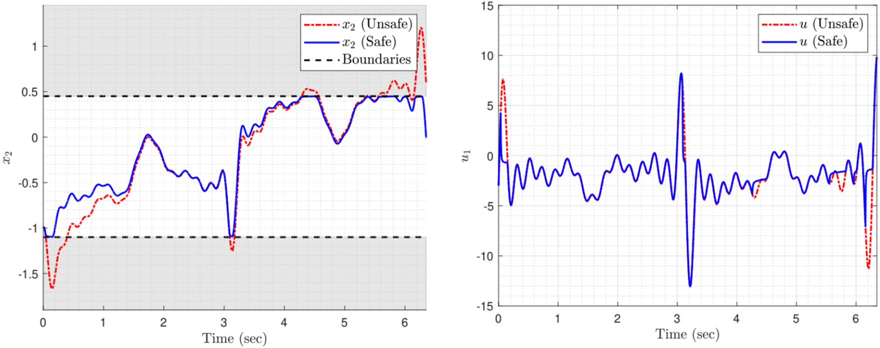
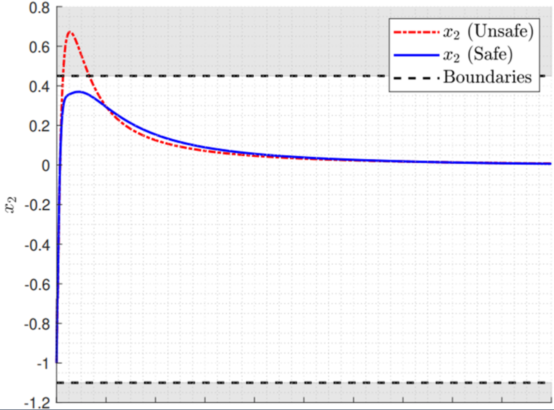


Fig.1 Trajectory of x_2 during exploration

Fig.2 Exploration policy under probing noise

Example

Exploitation of Learned Policy



Conclusions

- Optimality
- Stability
- Safety during operation—OK!
- Safety during EXPLORATION –OK!
- Initial admissible policy –OK!
- BUT,
  - Tracking?
  - Exploration Quality ?
  - Input saturation ?
  - Model Based

Kanso, S, Jha, MS, Theilliol, D. Off-policy model-based end-to-end safe reinforcement learning. *Int J Robust Nonlinear Control*. 2023; 1-26. doi: 10.1002/rnc.7109

Table of Contents

Exploration as Robust QP

Consider system under probing noise $e_u(t)$ during the exploration phase $\forall t \geq 0$ as:

$$\dot{x} = f(x) + g(x)(u + e_u) \tag{25}$$

where $e_u : \mathbb{R}_{\geq 0} \to \mathbb{R}^m$ is a time-varying probing noise, $\|e_u(t)\|_\infty = \sup_{t \geq 0} \|e_u(t)\| < \infty$.

Tunable input to state safe exploration

- probing noise $e_u(t)$ as a matched disturbance,
- a larger safe set $\mathcal{C}_{\xi,\mathrm{T}} \subset \mathbb{R}^n$ is considered parameterized by $\xi \geq 0$ such that $\mathcal{C} \subseteq \mathcal{C}_{\xi,\mathrm{T}}$.
- This larger set $\mathcal{C}_{\xi,\mathrm{T}}$ should remain forward invariant for all $\|e_u(t)\|$ satisfying $\|e_u(t)\|_\infty \leq \xi$ to ensure safety during data collection phase. To that end, consider a function $h_{\xi,\mathrm{T}} : \mathbb{R}^n \times \mathbb{R}_{\geq 0} \to \mathbb{R}$ as:

$$h_{\xi,\mathrm{T}}(x,\xi) = h(x) + \gamma_{\mathrm{T}}(h(x),\xi) \tag{26}$$

$\gamma_{\mathrm{T}}(a,\cdot) \in \mathcal{K}_\infty$ for all $a \in \mathbb{R}$. Then, a larger set $\mathcal{C}_{\xi,\mathrm{T}}$ becomes:

$$\mathcal{C}_{\xi,\mathrm{T}} \triangleq \{x \in \mathbb{R}^n : h(x) + \gamma_{\mathrm{T}}(h(x),\xi) \geq 0\} \tag{27}$$

$$\partial\mathcal{C}_{\xi,\mathrm{T}} \triangleq \{x \in \mathbb{R}^n : h(x) + \gamma_{\mathrm{T}}(h(x),\xi) = 0\} \tag{28}$$

$$\mathrm{Int}\,(\mathcal{C}_{\xi,\mathrm{T}}) \triangleq \{x \in \mathbb{R}^n : h(x) + \gamma_{\mathrm{T}}(h(x),\xi) > 0\}. \tag{29}$$

Input to State Safety

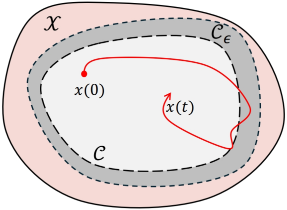
Input-to-State Safety

Adding probing noise ϵ to the control input leading to the following dynamics:

$$\dot{x} = f(x) + g(x)(\underbrace{u_0 + \epsilon}_{\text{matched disturbance}})$$

The probing noise is assumed to not destabilize the system, and:

$$|\epsilon|_\infty = \text{ess} \sup_{t \in \mathbb{R}_{\geq 0}} |\epsilon(t)|$$



Input-to-State Safe (ISSf) [Romdlony and Jayawardhana, 2016], [Kolathaya et al., 2018]

Given $\mathcal{C} \subset \mathcal{X}$ the 0-superlevel set of a continuously differentiable function $h : \mathcal{X} \to \mathbb{R}$, the system is **ISSf** with respect to $\mathcal{C}$ if there exist $\epsilon \in \mathbb{R}_{>0}$ and $\mu \in \kappa$ such that for all $\epsilon \in [0, \bar{\epsilon}]$, the set $\mathcal{C}_\epsilon \subset \mathcal{X}$ defined by:

$$\mathcal{C}_\epsilon = \{x \in \mathcal{X} \mid h(x) + \mu(|\epsilon|_\infty) \geq 0\}$$

is forward invariant.

Figure:

Exploration near safety boundries

Tunable Input-to-State Safety Control Barrier Function

Tunable Input-to-State Safe Control Barrier Function (TISSf-CBF) [Alan et al.,2021]

The function h is an **TISSf-CBF** on $\mathcal{C}$ if there exist an extended κ_∞ function α and λ: $\mathbb{R} \to \mathbb{R}_{>0}$ that is continuously differentiable on $\mathbb{R}$ such that:

$$\sup_{u \in \mathcal{U}} \left[\frac{\partial h(x)}{dx} f(x) + \frac{\partial h(x)}{dx} g(x) u - \frac{1}{\boldsymbol{\lambda(h(x))}} \left\| \frac{\partial h(x)}{dx} g(x) \right\|^2 \right] > -\alpha(h(x))$$

for all $x \in \mathcal{X}$,

$$\frac{\partial \lambda}{\partial r}(r) \geq 0$$

for all $r \in \mathcal{X}$.

Safety during Exploration

Solve the **QP** problem for u_{QP} to marginally adjust the exploration input:

$$\min_{u_{QP}} \frac{1}{2} u_{QP}^T M u_{QP}$$

s.t.

Condition of **TISSf-CBF** is satisfied

Figure:

Exploration near safety boundries

TISSf-CBF vs ISSf-CBF

$$\frac{\partial h(x)}{dx} f(x) + \frac{\partial h(x)}{dx} g(x)u > -\alpha\big(h(x)\big) + \frac{1}{\lambda\big(h(x)\big)} \left\| \frac{\partial h(x)}{dx} g(x) \right\|^2$$

Why are we using TISSf-CBF instead of ISSf-CBF?

In **ISSf-CBF**, λ is a constant

- Small λ values → More conservative → Compromised data quality
- Large λ values → Less conservative → Violation of Safety

In **TISSf-CBF**, λ is a function of $h(x)$

- Near safety boundaries → $h(x)$ is small → λ smaller → More conservative
- Within the safe set → $h(x)$ is large → λ increases → Less conservative

Exploration near safety boundries

Simulation and Results

λ Tunable

Blue curve: λ is a function of $h(.)$
$h_1(\rho X) = X_2 + 2$ and $h_2(\rho X) = 2 - X_2$

λ Fix

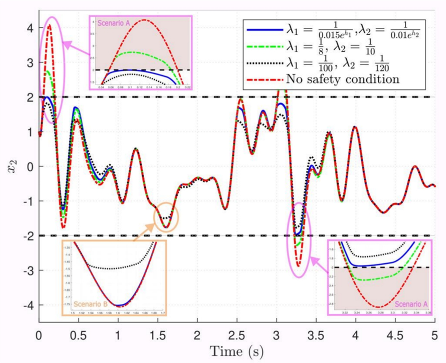
Green curve: λ is a large value
Black curve: λ is a small value

No safety

Red curve: Exploration input is not adjusted

Scenario A: Exploration input is unsafe

Scenario B: Exploration input is safe



Trajectory of x_2 during exploration for different values of λ

Figure:

Exploration near safety boundries

Simulation and Results

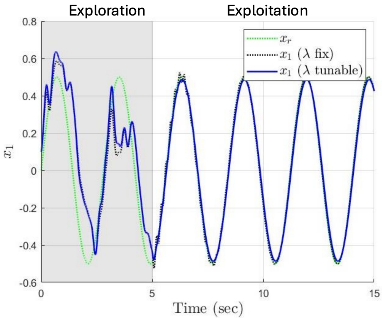


Fig 1: Trajectory of x_1 during exploration and exploitation

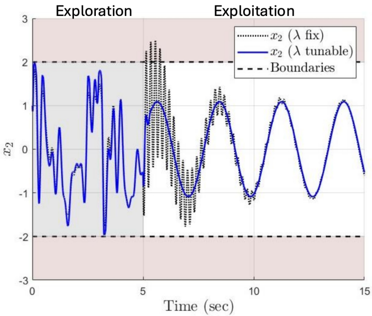


Fig 2: Trajectory of x_2 during exploration and exploitation

Figure:

Table of Contents

- Under saturation
- Learning CBFs, CLFs
  - Gaussian process,
  - Neural ODEs
- Abruptly/slowly varying environments
- Varying dynamics
- Stochastic dynamics
- Stochastic noise : Excitation noise with probability distribution.

Fin. ?

References I

Agrawal, A., & Sreenath, K. (2017).Discrete Control Barrier Functions for Safety-Critical Control of Discrete Systems with Application to Bipedal Robot Navigation.. *Robotics: Science and Systems, 13*.

Ames, A. D., Xu, X., Grizzle, J. W., & Tabuada, P. (2016).Control barrier function based quadratic programs for safety critical systems. *IEEE Transactions on Automatic Control, 62*(8), 3861–3876.

References II

Brunke, L., Greeff, M., Hall, A. W., Yuan, Z., Zhou, S., Panerati, J., & Schoellig, A. P. (2022).Safe learning in robotics: From learning-based control to safe reinforcement learning. *Annual Review of Control, Robotics, and Autonomous Systems*, *5*, 411–444.

Howard, R. A. (1960).Dynamic programming and markov processes..

Leake, R., & Liu, R.-W. (1967).Construction of suboptimal control sequences. *SIAM Journal on Control*, *5*(1), 54–63.

References III

Wabersich, K. P., Taylor, A. J., Choi, J. J., Sreenath, K., Tomlin, C. J., Ames, A. D., & Zeilinger, M. N. (2023).Data-driven safety filters: Hamilton-jacobi reachability, control barrier functions, and predictive methods for uncertain systems. *IEEE Control Systems Magazine*, *43*(5), 137–177.