Adaptive optimal control of continuous-time nonlinear affine systems via hybrid iteration - NOTES

Devesh

November 25, 2024

1 Problem Formulation

Value Iteration is slow to converge.

Policy Iteration is prone to numerical instability.

Consider the nonlinear affine system given by:

$$\dot{x} = f(x) + g(x)u, \quad x(0) = x_0$$

The cost functional is defined as:

$$J(x, u) = \int_0^\infty \left[Q(x) + u^T R u\right] dt,$$

where $Q(x)$ is a positive semi-definite state cost and R is a positive definite control cost matrix.

The optimal value function is defined as:

$$V^*(x_0) = \inf_u J(x_0; u), \quad u^* = \arg\inf_u J(x_0; u)$$

The optimal value function V^* solves the Hamilton–Jacobi–Bellman (HJB) equation:

$$\inf_u H(x, \nabla_x V, u) = 0, \quad V(0) = 0, \quad \forall x \in \mathbb{R}^n,$$

The Hamiltonian is defined as:

$$H(x, \nabla_x V, u) := (\nabla_x V)^T (f(x) + g(x)u) + Q(x) + u^T R u,$$

The term $(\nabla_x V)^T (f(x) + g(x)u)$ represents the rate of change of V along the system's trajectory. It indicates how the value function $V(x)$ changes in the direction of the vector $f(x) + g(x)u$. Since $\nabla_x V(x)$ points in the direction of the steepest ascent of $V(x)$, the entire expression is a scalar that quantifies this rate of change.

Q(x) is the control cost and $u^T Ru$ is the state cost.

The HJB equation can be written as:

$$0 = (\nabla_x V^*)^T f(x) + Q(x) - \frac{1}{4}(\nabla_x V^*)^T g(x) R^{-1} g(x)^T \nabla_x V^*,$$

The optimal control law is given by:

$$u^*(x) = -\frac{1}{2} R^{-1} g(x)^T \nabla_x V^*, \quad \forall x \in \mathbb{R}^n.$$

2 Model Based Policy Iteration

Policy Evaluation: Using $u_i \in \mathcal{D}$, solve V_i from

$$H(x, \nabla_x V_i, u_i) = 0, \quad V_i(0) = 0.$$

$\mathcal{D}$ is the set of admissible control inputs.

Policy Improvement: Update the control policy by

$$u_{i+1} = -\frac{1}{2} R^{-1} g(x)^T \nabla_x V_i.$$

3 Model Based Value Iteration

Value Update: Solve $V_i(x)$ from

$$\inf_u \{H(x, \nabla_x V_i, u_{i+1})\} = 0,$$

where u_{i+1} is defined in.

In practice, can be solved using different numerical methods, such as stochastic approximation and forward Euler method.

$$V_{i+1}(x) \leftarrow V_i(x) + \epsilon_i \left((\nabla_x V_i(x))^T f(x) + Q(x) - (u_{i+1}(x))^T R u_{i+1}(x) \right)$$

where the sequence $\{\epsilon_i\}_{i=0}^{\infty}$ is a deterministic sequence. .

.

.

.

.

.

4 Model Based Hybrid Iteration

Algorithm 1 Model-based Hybrid Iteration

Choose a proper $V_0 \in \mathcal{P}$, $V_0(0) = 0$, and $\hat{Q}(x) \succ Q(x)$.
$i \leftarrow 0$
repeat
  Compute $u_{i+1} = -\frac{1}{2}R^{-1}(g(x))^T \nabla_x V_i(x)$.
  Update the value function using

$$V_{i+1}(x) \leftarrow V_i(x) + \epsilon_i \left((\nabla_x V_i(x))^T f(x) + \hat{Q}(x) - (u_{i+1}(x))^T R u_{i+1}(x) \right).$$

  $i \leftarrow i + 1$
until $V_i(x) - V_{i-1}(x) \preceq \epsilon_{i-1}[\hat{Q}(x) - Q(x)]$ and $V_{i-1} \in \mathcal{P}^+$.
loop
  Solve V_i from

$$H(x, \nabla_x V_i, u_i) = 0.$$

  Update the control policy by

$$u_{i+1} = -\frac{1}{2}R^{-1}(g(x))^T \nabla_x V_i(x).$$

  $i \leftarrow i + 1$
end loop

Phase 1: This warm-up phase learns a baseline control policy ensuring system stability and cost-effectiveness. The stopping condition prevents excessive refinement of the value function. The criterion $V_i(x) - V_{i-1}(x) \preceq \epsilon_{i-1}[\hat{Q}(x) - Q(x)]$ ensures diminishing updates and penalizes suboptimal states via $\hat{Q}(x)$.

It ensures that the control policy $u_i(x)$ becomes admissible (stabilizing and finite-cost). It establishes a proper value function $V(x)$, which corresponds to the learned admissible policy.

Phase 2: This phase refines the control policy by solving the HJB equation, focusing on precision whereas in Phase 1 $V(x)$ was calculated using relaxed constraints $\hat{Q}$. The value function is updated to reflect the new control policy. The process iterates until convergence.

5 Data Driven Hybrid Iteration

Phase 1: Learning towards an admissible control policy

To avoid the prior knowledge of an initial admissible control policy, Phase 1 is initiated, wherein no admissible control policy is required. Instead, an arbitrary initial value function, i.e., $V_0 \in \mathcal{P}, V_0(0) = 0$ is used. By iteratively updating the value function, and after the satisfaction of a certain condition, one can ensure that an admissible control policy to system (1) is found, which is then used to accelerate the convergence in Phase 2.

To begin with, we start by taking the time derivative of V along the trajectory of system (1). Under

Assumption 1, the following is obtained.

$$\dot{V}(x,s) = \nabla_x V(x,s)\dot{x}(t) = \bar{H}(x, \nabla_x V(x,s), u) - \hat{Q}(x) - u^T R u,$$

where $\bar{H}(x, \nabla_x V, u) = H(x, \nabla_x V, u) - Q(x) + \hat{Q}(x)$.

By function approximation, three sets of continuously differentiable and linearly independent basis functions $\{\phi_j\}_{j=1}^{\infty}$, $\{\psi_j^{(0)}\}_{j=1}^{\infty}$, and $\{\psi_j^{(1)}\}_{j=1}^{\infty}$ are used to approximate V and $\bar{H}$ at each time instance s for all $x \in \mathcal{D}$. The approximated functions are

$$\hat{V}_N(x) = \sum_{j=1}^{N_w} \hat{w}_j \phi_j(x),$$

$$\hat{H}_N(x, u, \hat{c}) = \sum_{j=1}^{N_0} \hat{c}_j^{(0)} \psi_j^{(0)}(x) + \sum_{j=1}^{N_1} \hat{c}_j^{(1)} \psi_j^{(1)}(x) u + u^T R u.$$

Given the sets of basis functions $\{\phi_j\}_{j=1}^{N_w}$, $\{\psi_j^{(0)}\}_{j=1}^{N_0}$, and $\{\psi_j^{(1)}\}_{j=1}^{N_1}$, with $N_w, N_0, N_1 \in \mathbb{Z}^+$, the following are denoted for any $t_f > 0$.

$$K_\phi(t_f) = \int_0^{t_f} \Phi(x)\Phi^T(x)\, dt,$$

$$K_\psi(t_f) = \int_0^{t_f} \Psi(x,u)\Psi^T(x,u)\, dt,$$

where

$$\Phi(x) = \begin{bmatrix} \phi_1(x) & \phi_2(x) & \cdots & \phi_{N_w}(x) \end{bmatrix}^T,$$

$$\Psi(x,u) = \begin{bmatrix} \Psi^{(0)}(x) & \Psi^{(1)}(x,u) & u^T R u \end{bmatrix}^T,$$

$$\Psi^{(0)}(x) = \begin{bmatrix} \psi_1^{(0)}(x) & \psi_2^{(0)}(x) & \cdots & \psi_{N_0}^{(0)}(x) \end{bmatrix},$$

$$\Psi^{(1)}(x,u) = \begin{bmatrix} (\psi_1^{(1)}(x)u)^T & (\psi_2^{(1)}(x)u)^T & \cdots & (\psi_{N_1}^{(1)}(x)u)^T \end{bmatrix}.$$

From Eqs. (17) and (18) along with (19)–(22), the following differential equation is obtained.

$$\frac{d\hat{w}}{ds} = K_\phi^{-1}(t_f) \int_0^{t_f} \Phi(x) \hat{H}_N(x, \hat{\mu}_N(x, \hat{c}))\, dt,$$

where

$$\hat{c} = K_\psi^{-1}(t_f) \int_0^{t_f} \Psi(x,u) \left(\frac{d\hat{V}_N(x,\hat{w})}{dt} + \hat{Q}(x) + u^T R u \right) dt,$$

$$\hat{\mu}_N = -\frac{1}{2} R^{-1} \left(\sum_{j=1}^{N_1} \hat{c}_j^{(1)} \psi_j^{(1)}(x) \right)^T.$$

The vector $\hat{c}$ is comprised of $\{\hat{c}_j^{(0)}\}_{j=1}^{N_0}$ and $\{\hat{c}_j^{(1)}\}_{j=1}^{N_1}$ that are the weights corresponding to the sets $\{\psi_j^{(0)}\}_{j=1}^{N_0}$ and $\{\psi_j^{(1)}\}_{j=1}^{N_1}$, respectively. During this phase, the weights $\hat{w}$ and $\hat{c}$ are updated by applying an essentially bounded input (exploration noise) over a time interval $[0, t_f]$, $t_f > 0$ and selecting $\hat{w}(0) \in \{w : \hat{V}_N(\cdot, w) \in$

$\mathcal{P}$ is proper$\}$ and $\hat{c}(0) = 0$. The collected state/input information is used to solve Eq. (23). This process is repeated until an admissible control policy is obtained.

Phase 2: Exploring the optimal control policy

Since this phase is PI-based, an admissible control policy is required to initiate this phase. We define a virtual input v and rewrite the system described in (1) to be in the following form for all $t \geq 0$.

$$\dot{x} = f(x) + g(x)u_i(x) + g(x)v_i,$$

where $v_i = u - u_i$. By definition, for each $i \in \mathbb{Z}^+$, considering the time derivative of $V_i(x)$ along the solutions of (24) we have

$$\dot{V}_i(x) = \nabla_x V_i(x)^T[f(x) + g(x)u_i(x) + g(x)v_i] = -Q(x) - u_i(x)^T R u_i(x) - 2u_{i+1}(x)^T R v_i.$$

Integrating both sides of (25) over any time interval $[t_{k-1}, t_k]$, it follows that

$$V_i(x(t_k)) - V_i(x(t_{k-1})) = \int_{t_{k-1}}^{t_k} \left(-Q(x) - u_i(x)^T R u_i(x) - 2u_{i+1}(x)^T R v_i\right) dt,$$

such that $\{t_k\}_{k=1}^{l}$ is a strictly increasing sequence with a sufficiently large $l \in \mathbb{Z}^+$, and $t_l = t_{l-1} + \Delta t$, $\Delta t > 0$.

Replacing the value function $V_i(x)$ in (26) with the approximation (17), and the control policy $u_{i+1}(x)$ with its following approximation,

$$\hat{u}_{i+1} = -\frac{1}{2} R^{-1} \left(\sum_{j=1}^{N_1} \hat{c}_{i,j}^{(1)} \psi_j^{(1)}(x) \right)^T,$$

we obtain

$$\sum_{j=1}^{N_w} \hat{w}_{i,j} \left[\phi_j(x(t_k)) - \phi_j(x(t_{k-1}))\right] = -\int_{t_{k-1}}^{t_k} \left[Q(x) + u_i(x)^T R u_i(x)\right] dt - \int_{t_{k-1}}^{t_k} 2 \left(\sum_{j=1}^{N_1} \hat{c}_{i,j}^{(1)} \psi_j^{(1)}(x) \right)^T R v_i \, dt + e_{i,k},$$

where $\hat{u}_0 = u_0$ and $\hat{v}_i = u - \hat{u}_i$. The weight vectors $\hat{c}_i^{(1)}$ and $\hat{w}_i$ can be solved in a least squares sense, i.e., by minimizing $e_i := \sum_{k=1}^{l} e_{i,k}^2$.

Algorithm 2: Data-Driven Hybrid Iteration

Algorithm 2 Data-Driven Hybrid Iteration

Choose t_f, N_0, N_1, N_w, and sets of basis functions, $\{\psi_j^{(0)}(x)\}_{j=1}^{N_0}$, $\{\psi_j^{(1)}(x)\}_{j=1}^{N_1}$, and $\{\phi_j(x)\}_{j=1}^{N_w}$.
Select $\hat{w}_0 \in \{w : \hat{V}_N(\cdot, w) \in \mathcal{P} \text{ is proper}\}$, $\hat{c}_0 = 0$, a small constant $\epsilon > 0$, and $\hat{Q}(x) \succ Q(x)$.
$i \leftarrow 0$
Apply a measurable locally essentially bounded input u to system (1).
Collect the state/input data over the interval $[0, t_f]$.
repeat
    Solve $\hat{w}_{i+1}$ and $\hat{c}_i$ by

$$\hat{c}_i \leftarrow K_\psi^{-1}(t_f) \int_0^{t_f} \Psi(x,u) \left(\frac{d\hat{V}_N(x,\hat{w}_i)}{dt} + \hat{Q}(x) + u^T R u \right) dt,$$

$$\hat{w}_{i+1} \leftarrow \hat{w}_i + \epsilon_i K_\phi^{-1}(t_f) \int_0^{t_f} \Phi(x) \hat{H}_N(x, \hat{\mu}_N(x, \hat{c}_i), \hat{c}_i)\, dt.$$

    $i \leftarrow i + 1$
until $(\hat{w}_i - \hat{w}_{i-1})^T \Phi(x) \preceq \epsilon_{i-1}[\hat{Q}(x) - Q(x)]$ and $\hat{w}_{i-1}^T \Phi(x) \in \mathcal{P}^+$
$u_i(x) \leftarrow \hat{\mu}_N(x, \hat{c}_{i-1})$
repeat
    Solve $\hat{w}_i$ and $\hat{c}_i^{(1)}$ from (29).
    Update the control input from (27).
    $i \leftarrow i + 1$
until $\|\hat{w}_i - \hat{w}_{i-1}\| < \epsilon$
Set $\hat{u}^* = \hat{u}_i$ as an approximation to the optimal control policy.