Objective
The parallel autonomy system learns optimal policies to assist a human operator in regulating a process—from continuous improvements with minimal interventions, to taking over full-control when necessary.

Parallel Autonomy

Human Operator

\[ u(t) \]

\[ u(t) \]

\[ A(t) = AQ(t) + Bu(t), \]

where

\[ x = A(t)u(t), \]

\[ y = C_1(t) \]

System Dynamics

\[ x(t) = Ax(t) + Bu(t), y_{\text{output}} = Cy_{\text{output}}, u_{\text{input}} = \text{input of the plant}, u_{\text{computed}} = \text{autonomy computed control}, A, B, C \in \mathbb{R}^{n \times n}. \]

Reinforcement Learning

- On-Policy:
  \[ V(x(u_t)) = \int_{0}^{\infty} \left( x^T(t)Q x(t) + u^T(t)R u(t) \right) dt + V(x(u_0)), \]
  where \( u_0 = \frac{1}{2} R^{-1} B^T P_0 \).

- Off-Policy:
  \[ V(x(u_0 + \tau)) = \int_{0}^{\infty} \left( x^T(t)Q x(t) + u^T(t)R u(t) \right) dt - \frac{1}{2} \left( x^T(t)Q x(t) + u^T(t)R u(t) \right) \]
  \[ = - \frac{1}{2} \left( x^T(t)Q x(t) + u^T(t)R u(t) \right) \]
  \[ = \frac{1}{2} \left( x^{-T}(t)Q x(t) + u^T(t)R u(t) \right). \]

Human-in-the-Loop Reinforcement Learning

In this case, we have \( \dot{x} = Ax + Bu_0(x) \) where \( A = A + BK_0 C_0 \).

Assumptions
- Human policy is linear and is given by \( u_h(x) = x_0, C_0 \).
- The matrices \( A, K_0, C_0 \) are unknown to the autonomy system.
- Input matrix \( B \) is known to the autonomy system.
- The autonomy system can measure \( u_h, x \).

Optimal Control Formulation

- Problem 1 (Minimum Intervention): Solve the infinite-horizon optimal control problem
  \[ J(x_0, u_h, u_c) = \frac{1}{2} \int_{0}^{\infty} \left( x^T(t)Q x(t) + u^T(t)R u(t) \right) dt + V(x(t)), \]
  where \( u_0 = \frac{1}{2} R^{-1} B^T P_0 \).

- Problem 2 (Take Over): Solve Problem 1 with \( u_h = 0 \).

Application to Car Following

Main Analysis Results
- We avoid learning along a single state-space trajectory which we show leads to data collinearity under certain conditions such as algebraic multiplicity of eigenvalues.
- We show that exploring a minimum number of pairwise distinct state-space trajectories is necessary to avoid collinearity in the learning data.
- We make a clear separation between exploitation of learned policies and exploration of the state-space, and propose an exploration scheme that requires switching to new state-space trajectories rather than injecting noise continuously while evaluating the cost-to-go. This avoidance of continuous noise injection minimizes interference with human action, and avoids bias in the convergence to the stabilizing solution of the underlying algebraic Riccati equation.
- We show conditions under which existence and uniqueness of solutions can be established for off-policy reinforcement learning in continuous-time linear systems; namely a required knowledge of the input matrix \( B \).

More details at: https://arxiv.org/abs/1905.11524