Real-World RL

Hyochan Lee [*] and Kyunghwan Choi [†]

Compiled on August 14, 2026

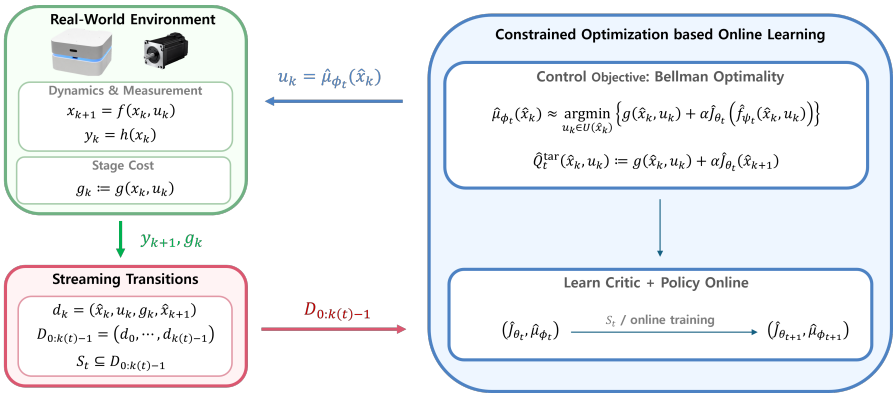

Figure 1: Real interactions generate streaming transitions that train both the critic and policy online .

The environment returns measured transition records; an assumed state estimator converts the measurements into $\widehat{x}_k$ before the state-based transition enters $\mathcal{S}_t$. Without an estimator, $\widehat{x}_k$ is replaced by an information state o_k. Physical interaction is indexed by k, whereas network updates are indexed by t; the two clocks need not advance at the same rate. Here $k(t)$ is the latest physical-state sample available at learner update t, and $t(k)$ is the latest learner update available at physical step k.

1 Problem Definition

Online Learning-Based Optimal Control defines the desired policy through Bellman optimality. The broader Real-World RL Theme addresses the case in which its critic and, when explicitly parameterized, policy are learned from real closed-loop data. A representative approximate policy is

$$\widehat{\mu}_{\phi_t}(\widehat{x}) \approx \arg\min_{u \in \mathcal{U}(\widehat{x})} \left\{ g(\widehat{x}, u) + \alpha \widehat{J}_{\theta_t}\left(\widehat{f}_{\psi_t}(\widehat{x}, u)\right) \right\}. \tag{1}$$

[*]Hyochan Lee is Ph.D. student at KAIST, hyochanlee@kaist.ac.kr
[†]Kyunghwan Choi is Assistant Professor at KAIST

Here $\widehat{J}_{\theta_t}$ is the critic, $\widehat{\mu}_{\phi_t}$ is the policy network, $\widehat{f}_{\psi_t}$ is a successor model with parameter ψ_t, and $0 < \alpha < 1$ is the discount factor. The expression above is model-based. Without a successor model, an observed transition provides a sample-based Q-factor target directly:

$$\widehat{Q}_t^{\mathrm{tar}}(\widehat{x}_k, u_k) := g(\widehat{x}_k, u_k) + \alpha \widehat{J}_{\theta_t}(\widehat{x}_{k+1}). \tag{2}$$

A learned Q-critic fits such targets over state-action pairs; model-free policy improvement then minimizes the learned Q-factor over u.

For the primary notation, a state estimator is assumed to supply the controller-facing state estimate. The real system then supplies a stream

$$d_k = (\widehat{x}_k, u_k, g_k, \widehat{x}_{k+1}), \qquad \mathcal{D}_{0:k(t)-1} := (d_0, \ldots, d_{k(t)-1}), \qquad \mathcal{S}_t \subseteq \mathcal{D}_{0:k(t)-1}. \tag{3}$$

Here $g_k := g(\widehat{x}_k, u_k)$ is the realized stage cost. Because a complete transition ending at the latest state sample $k(t)$ is indexed by $k(t)-1$, $\mathcal{D}_{0:k(t)-1}$ is the ordered collection of all completed transition records available to the learner. The update set $\mathcal{S}_t$ may contain the latest transition alone or selected prior transitions. Replay or minibatch training is optional rather than a defining assumption.

Without a state estimator, the problem becomes partially observed. Then $\widehat{x}_k$ should be replaced by an observation or information state o_k: $o_k = y_k$ only when the current measurement is sufficient for control; otherwise o_k should summarize an observation history, belief state, or learned latent state. The corresponding record is $d_k = (o_k, u_k, g_k, o_{k+1})$.

In an actor-critic realization, the selected update data train both learned objects:

$$\left(\widehat{J}_{\theta_t}, \widehat{\mu}_{\phi_t}\right) \xrightarrow[\mathcal{S}_t]{\text{online training}} \left(\widehat{J}_{\theta_{t+1}}, \widehat{\mu}_{\phi_{t+1}}\right), \qquad u_k = \widehat{\mu}_{\phi_{t(k)}}(\widehat{x}_k). \tag{4}$$

2 Why This Problem Matters

Many control-oriented RL methods are trained and evaluated primarily in simulation. Their deployed performance can then be limited by the sim-to-real gap: unmodeled dynamics, sensing imperfections, disturbances, and operating conditions absent from the simulator.

Learning directly from the real system can capture this missing information and continue improving when deployment reveals new operating conditions. Here, adaptability means accumulating and reusing experience to learn a critic and policy that remain useful over a declared task domain. It does not mean only tracking a local parameter change along the currently excited trajectory, as in an adaptive-control-like update.

3 Key Challenges

- **Limited and policy-dependent data:** real transitions arrive sequentially and are correlated. Unlike simulator rollouts, they cannot be generated freely over the entire state-action domain.

- **Local fit versus task-level learning:** reducing loss on the latest trajectory can produce a local fit, forget earlier regimes, or repeatedly relearn the same task. The learned functions must retain broader task-relevant structure.
- **Learning inside the closed loop:** critic and policy change while the system is operating. Update time, action latency, stability, constraints, fallback behavior, and safety evidence must therefore be part of the learning problem.

4 A Conventional Approach: TD-Error-Based Approximate Policy Iteration

One conventional approach is TD-error-based approximate policy iteration: fix the current policy, reduce its one-step temporal-difference error, and then perform policy improvement. For notational simplicity, this section writes x_k as the state available to the controller; when an estimator is used, it is replaced by $\widehat{x}_k$. Here π_i is the policy at policy-iteration step i.

$$\begin{aligned} &\delta_k^{\pi_i} = g(x_k, \pi_i(x_k)) + \alpha \widehat{J}^{\pi_i}(x_{k+1}) - \widehat{J}^{\pi_i}(x_k), \\ &\widehat{J}^{\pi_i} \xleftarrow{\text{TD error reduction}} \mathrm{Fit}(\delta_k^{\pi_i}), \qquad \pi_{i+1} \leftarrow \mathrm{Greedy}\left(\widehat{J}^{\pi_i}\right). \end{aligned} \tag{5}$$

This is a valid approximate-DP mechanism, but streaming data may make it adaptation-like: a small TD error on visited samples establishes sampled policy consistency, not task-wide Bellman optimality. Limited coverage can therefore yield episode- or trajectory-dependent critic parameters.

5 Current Precursor: BOCO Critic and Constrained Control Learning

BOCO (Bellman Optimality via Constrained Optimization) first parameterizes the unknown optimal critic,

$$J^*(x) \approx \widehat{J}_\theta(x), \tag{6}$$

and, at each selected update state, expresses the local Bellman minimization and equality as a finite-dimensional constrained optimization problem. This makes the local condition amenable to numerical KKT or primal-dual implementation:

$$\begin{aligned} &F_k(u, \theta) := g(x_k, u) + \alpha \widehat{J}_\theta(x_{k+1}(u)), \\ &\delta_k(u, \theta) := F_k(u, \theta) - \widehat{J}_\theta(x_k). \end{aligned} \tag{7}$$

$$\begin{aligned} \min_{u,\theta} \quad & F_k(u, \theta) \\ \text{s.t.} \quad & \delta_k(u, \theta) = 0, \\ & c_i(x_k, u) \le 0, \qquad i = 1, \ldots, n_c. \end{aligned} \tag{8}$$

Here $x_{k+1}(u)$ denotes the successor associated with candidate input u. Evaluating this quantity for arbitrary candidate inputs requires a declared control-usable model, simulator, or other counterfactual mechanism; one realized transition supplies the successor only for the input that was actually applied. The set $\mathcal{U}(x_k)$ contains admissible inputs, n_c is the number of explicit inequality constraints, and c_i is the ith such constraint. It may represent an actuator limit, a state-dependent feasibility condition, or a safety-related operating bound. These constraints can be added explicitly instead of being handled only by a penalty.

A global critic still requires accumulated streaming constraints, a declared collocation or update set, or another function-approximation condition; one local equality does not identify the complete value function.

The Lagrangian is

$$\mathcal{L}_k(u, \theta, \lambda) = F_k(u, \theta) + \lambda_\delta \delta_k(u, \theta) + \sum_{i=1}^{n_c} \lambda_i c_i(x_k, u), \qquad \lambda_i \geq 0. \tag{9}$$

At online update t, the primal control, critic parameter, and multipliers can be updated conceptually as

$$\begin{aligned} u_k^{\mathrm{BOCO}} &\in \arg\min_{u \in \mathcal{U}(x_k)} \mathcal{L}_k(u, \theta_t, \lambda_t), \\ \theta_{t+1} &= \theta_t - \eta_{\theta,t} \nabla_\theta \mathcal{L}_k\left(u_k^{\mathrm{BOCO}}, \theta_t, \lambda_t\right), \\ \lambda_{\delta,t+1} &= \lambda_{\delta,t} + \eta_{\delta,t}\, \delta_k\left(u_k^{\mathrm{BOCO}}, \theta_t\right), \\ \lambda_{i,t+1} &= \left[\lambda_{i,t} + \eta_{\lambda_i,t}\, c_i\left(x_k, u_k^{\mathrm{BOCO}}\right)\right]_+ . \end{aligned} \tag{10}$$

Here $\lambda_t = (\lambda_{\delta,t}, \lambda_{1,t}, \ldots, \lambda_{n_c,t})$ collects the current multipliers, the η terms are positive update step sizes, and $[\cdot]_+$ projects each inequality multiplier onto the nonnegative orthant. The equality multiplier λ_δ is unrestricted in sign. The physical state x_k is embedded in F_k, δ_k, c_i, and therefore $\mathcal{L}_k$.

Unlike TD-error reduction alone, BOCO keeps control minimization, critic consistency, and explicit constraints together. In the current formulation, u_k^{BOCO} is the primal actor decision and a separate policy-network parameter ϕ is not required. A future amortized policy network could be trained to reproduce this solution, but that is a distinct extension.

The current BOCO paper provides this formulation and constrained nonlinear simulation evidence. Formal stability, safety, broader task-domain learning, separate policy-network training, and real-system validation require additional assumptions and evidence.

6 Application Domains

- **Automatic tuning and calibration of control systems for electric drives and vehicle or mobility systems**, using operational data to improve the deployed critic and controller.
- **Controller learning in complex, unstructured real-world environments** where a faithful simulator and exhaustive offline training data are difficult to obtain.

7 References

- H. Lee and K. Choi, "Constrained Optimization Formulation of Bellman Optimality Equation for Online Reinforcement Learning," International Conference on Robot Intelligence Technology and Applications (RiTA), accepted, in press, 2025.
- H. Lee and K. Choi, "Online Constrained Reinforcement Learning for Optimal Tracking," International Federation of Automatic Control (IFAC) World Congress, accepted, in press, 2026.
- H. Lee and K. Choi, "Online Actor Critic Learning for Optimal Tracking in Servo Positioning Systems," Annual Conference of the IEEE Industrial Electronics Society (IECON), pp. 1–7, 2025.

References