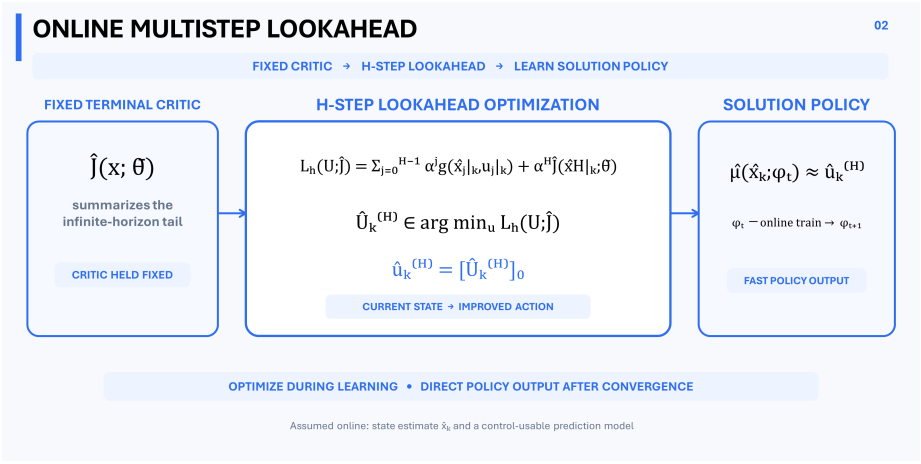
Online Multistep Lookahead

Changeun Park *

Complied on August 14, 2026



Physical interaction is indexed by k, whereas policy-network updates are indexed by t. Here $t(k)$ is the latest policy-network update available at physical step k, and $k(t)$ is the latest physical sample available at learner update t. The critic parameter $\bar{\theta}$ is deliberately fixed in this Theme. A state estimate $\widehat{x}_k$ and a control-usable prediction model f_{use} are assumed available.

1 Problem definition

Online Learning-Based Optimal Control connects an approximate critic to policy improvement through Bellman minimization. Here the critic is already available but imperfect. Rather than using only a one-step greedy action, solve an H-step problem whose terminal cost is the fixed critic:

*Changeun Park is an integrated M.S./Ph.D. student at KAIST, aegis3141@kaist.ac.kr

$$
\begin{aligned}
\widehat{U}_k^{(H)} \in \arg\min_{U_k} \quad & \sum_{j=0}^{H-1} \alpha^j g(\widehat{x}_{j|k}, u_{j|k}) + \alpha^H \widehat{J}_{\bar{\theta}}(\widehat{x}_{H|k}) \\
\text{s.t.} \quad & \widehat{x}_{0|k} = \widehat{x}_k, \\
& \widehat{x}_{j+1|k} = f_{\text{use}}(\widehat{x}_{j|k}, u_{j|k}), \qquad j = 0, \dots, H-1, \\
& u_{j|k} \in \mathcal{U}(\widehat{x}_{j|k}), \qquad j = 0, \dots, H-1, \\
& \widehat{x}_{j|k} \in \mathcal{X}, \qquad j = 0, \dots, H,
\end{aligned}
\tag{1}
$$

$$
\widehat{u}_k^{(H)} = \left[\widehat{U}_k^{(H)}\right]_0 . \tag{2}
$$

H is a positive lookahead horizon, $0 < \alpha < 1$ is the discount factor, and $U_k = (u_{0|k}, \dots, u_{H-1|k})$ is the candidate input sequence. f_{use} is the declared prediction model, $\mathcal{U}(x)$ and $\mathcal{X}$ are the admissible input and state sets, and $\widehat{x}_{j|k}$ is the state predicted j steps from the current estimate. $\widehat{u}_k^{(H)}$ selects the first optimized action. For $H = 1$, the formulation reduces to one-step critic-based policy improvement.

The online-learned policy represents the solution map:

$$
\widehat{\mu}_{\phi_t}(\widehat{x}_k) \approx \widehat{u}_k^{(H)}, \qquad u_k^{\text{NN}} = \widehat{\mu}_{\phi_{t(k)}}(\widehat{x}_k). \tag{3}
$$

The critic is not updated in this Theme. The learned object is the policy induced by the fixed critic and multistep optimization. The model can be known or a declared snapshot of a learned model, but it is held fixed while each lookahead problem is solved. Changes in that snapshot, preview convention, or constraint set between solves move the teacher solution map and must be represented in the training data or policy input.

2 Why this problem matters

An approximate critic summarizes the infinite-horizon tail, but local critic error can lead to a poor one-step greedy action. Multistep lookahead evaluates the current action through H explicit stages before applying the critic. Under suitable discounted-DP and rollout assumptions, this can improve the policy even while the critic remains fixed, and a longer horizon can reduce the influence of terminal-critic error. Improvement is possible rather than automatic: model error, incomplete optimization, and policy approximation can offset the benefit.

Repeatedly solving the H-step problem may still be too expensive for an embedded controller. Learning its state-to-solution map amortizes the computation across operating points: lookahead supplies improved action targets during learning, whereas the converged policy network can return an action immediately without a full optimization at every control step.

3 Key challenges

- **Horizon and model trade-off:** increasing H can reduce dependence on the terminal critic but raises computation and accumulated model error.

- **Constrained real-time optimization:** nonlinear dynamics and constraints create feasibility, local-solution, warm-start, and worst-case latency issues.
- **Solution-policy learning:** targets are concentrated on visited states, and a small action-fitting error does not by itself imply small closed-loop performance loss.
- **Safe deployment:** approximating a feasible optimizer does not automatically preserve feasibility or stability; a correction solve, projection, safety filter, or fallback may remain necessary.

4 A conventional approach — repeated online lookahead

A conventional realization solves the problem in Section 1 at every physical control step and applies only the first action:

$$u_k = \widehat{u}_k^{(H)}, \qquad k = 0, 1, \ldots. \tag{4}$$

This receding-horizon implementation uses the current state, model, critic, and constraints directly. Its main limitation is repeated numerical optimization within the control deadline.

This mechanism is MPC-like, but its defining role here is more specific: the fixed approximate infinite-horizon critic closes the finite lookahead problem. Learning a conventional finite-horizon MPC solution without that terminal critic remains a useful amortized-MPC baseline, not this Theme's defining mechanism.

5 Research direction — online learning of the lookahead solution

Let selected online lookahead targets be

$$\mathcal{S}_t^{\mathrm{LA}} \subseteq \left\{ \left(\widehat{x}_i, \widehat{u}_i^{(H)} \right) \right\}_{i=0}^{k(t)}. \tag{5}$$

A conceptual policy-learning problem is

$$\phi_{t+1} \approx \arg\min_{\phi} \sum_{(x,u^{(H)}) \in \mathcal{S}_t^{\mathrm{LA}}} \left\| \widehat{\mu}_\phi(x) - u^{(H)} \right\|_{W_u}^2. \tag{6}$$

Here $W_u \succeq 0$ is the declared action-fitting weight matrix.

Equivalently, the research loop can be summarized as

$$\widehat{J}_{\bar{\theta}} \text{ fixed}, \qquad \phi_t \xrightarrow[\mathcal{S}_t^{\mathrm{LA}}]{\text{online solution learning}} \phi_{t+1}, \qquad u_k^{\mathrm{NN}} = \widehat{\mu}_{\phi_{t(k)}}(\widehat{x}_k). \tag{7}$$

The argmin states the learning objective; a real-time implementation may take only one or a few incremental updates. Training signals may come from a

completed lookahead solve or a corrected warm start. If necessary, deployment can retain a real feasibility mechanism,

$$u_k = \mathcal{C}_{\text{feas}}\left(\widehat{\mu}_{\phi_{t(k)}}(\widehat{x}_k), \widehat{x}_k\right), \tag{8}$$

where $\mathcal{C}_{\text{feas}}$ must denote an implemented projection, filter, or correction procedure rather than an assumed guarantee.

Evidence should compare the network action and objective with the online optimizer, constraint violations before and after correction, closed-loop cost, worst-case latency, model and critic perturbations, and performance on states not represented in the update stream. This page states the research direction and evaluation criteria; it does not claim a validated MIC Lab result.

6 Application domains

- **Electric-drive current or torque control:** combine a learned terminal critic with short online lookahead to make an approximate infinite-horizon MPC realization practical.
- **Coordination of multiple connected and automated vehicles and other complex decision problems:** use multistep optimization to generate a high-quality policy target, amortize that target into a fast policy network, or use the lookahead action as an online policy-improvement target within an RL framework.
- **Vehicle energy and thermal management:** represent long-horizon consequences with the terminal critic while multistep lookahead resolves short-horizon dynamics, constraints, and current operating information.
- **Embedded motion control and MPC-like systems** where a reliable optimizer is available during learning but too costly for permanent execution.

7 References

- D. P. Bertsekas, *Reinforcement Learning and Optimal Control*, Athena Scientific, 2019.
- D. P. Bertsekas, *A Course in Reinforcement Learning*, 2nd ed., Athena Scientific, 2025.

References