Structured Critic Adaptation

Soobin Hwang *

Complied on August 14, 2026

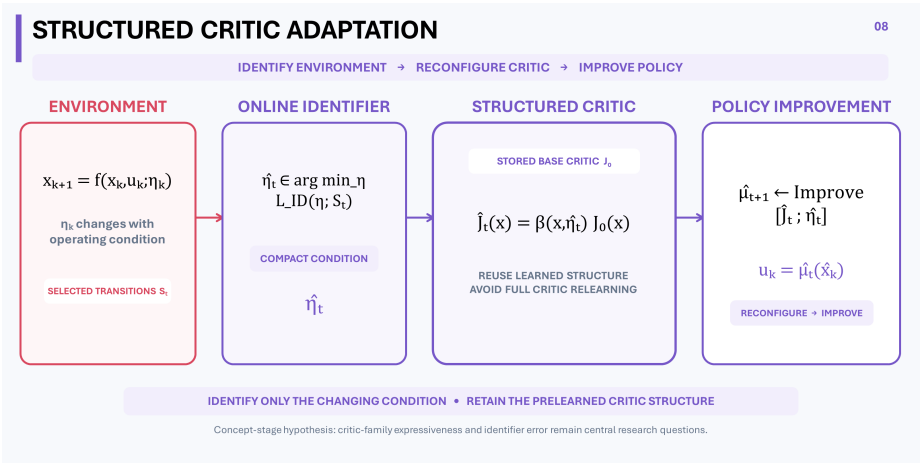

Figure 1: Selected transitions identify the environment, a learned structure reconfigures a stored critic, and the resulting critic improves the policy.

Physical interaction is indexed by k, whereas identification and policy updates are indexed by t. The environment parameter η_k may vary with physical conditions; $\widehat{\eta}_t$ is the estimate available at learner update t, and $k(t)$ is the latest physical sample associated with that update. Identification from a window assumes that η_k is constant or slowly varying over that window, or explicitly models its evolution.

1 Problem definition

Online Learning-Based Optimal Control requires a critic appropriate to the current decision problem. Consider a family of environments

$$x_{k+1} = f(x_k, u_k; \eta_k) + w_k, \qquad y_k = h(x_k) + v_k, \tag{1}$$

where x_k, u_k, and y_k are the state, control input, and measurement; f and h are the state-transition and measurement maps; and w_k and v_k represent

*Soobin Hwang is Integrated M.S./Ph.D. student at KAIST, tnqls2361@kaist.ac.kr

process and measurement uncertainty. The compact parameter η_k represents control-relevant changes in dynamics and, when declared, in the stage cost g or constraints. For notational simplicity, assume the state is available; otherwise, an estimator supplies $\widehat{x}_k$.

Unless the evolution of η_k is included in a Markov decision state, a condition-specific critic J_η below refers to the problem with η fixed over its value horizon. Material variation over that horizon instead requires a nonstationary formulation such as Nonstationary Infinite-Horizon OCP.

Let the transition record and the set selected at learner update t be

$$d_k = (x_k, u_k, g_k, x_{k+1}), \qquad g_k := g(x_k, u_k; \eta_k), \qquad \mathcal{S}_t = \{d_k : k \in \mathcal{I}_t\}, \tag{2}$$

where $\mathcal{I}_t$ is the selected set of physical transition indices. These recent transitions support online identification:

$$\widehat{\eta}_t \in \arg\min_{\eta} \mathcal{L}_{\mathrm{id}}(\eta; \mathcal{S}_t). \tag{3}$$

The loss $\mathcal{L}_{\mathrm{id}}$ must declare which effects identify η: transition residuals when η changes the dynamics, observed costs when it changes the objective, or constraint/activity information when it changes feasibility. A parameter that affects only an unobserved cost or constraint cannot be identified from state transitions alone. The single $\widehat{\eta}_t$ formulation assumes local constancy or sufficiently slow variation over $\mathcal{I}_t$; otherwise, a trajectory-valued condition model is required.

Let J_0 be a stored approximate base critic for a declared reference condition η_0, and let $\beta_\omega(x, \eta)$ be a structural map learned over representative conditions. The current critic is reconfigured as

$$\widehat{J}_t(x) := \beta_\omega(x, \widehat{\eta}_t) J_0(x), \tag{4}$$

and then used for policy improvement:

$$\widehat{\mu}_{t+1} \leftarrow \mathrm{Improve}\left[\widehat{J}_t; \widehat{\eta}_t\right]. \tag{5}$$

The generic Improve operator may denote a declared model-based Bellman improvement or a policy-network update. One-step lookahead is therefore a possible implementation, not part of the Theme definition. Here adaptation means reconfiguration through the identified condition, not unrestricted online retraining of all critic parameters. The multiplicative $\beta_\omega J_0$ form is the current schematic hypothesis; additive corrections or critic-bank interpolation require separate definitions.

2 Why this problem matters

Direct adaptive-DP updates can repeatedly modify the full critic whenever the operating condition changes. Such updates may fit the current condition while overwriting behavior useful under earlier conditions, so a recurring condition can require relearning.

If the environment family is organized by an identifiable parameter η, a prelearned relation from (x, η) to critic shape can reuse structure across conditions. Online computation is then concentrated on estimating η and improving the policy from the reconfigured critic. This is a conditional opportunity, not a guaranteed speedup: it depends on identifiability, offline coverage, and the expressiveness of the critic family.

3 Key challenges

- **Environment representation and identification:**
  η must capture control-relevant variation and be identifiable from the limited, policy-dependent data in $\mathcal{S}_t$.
- **Structured-family expressiveness:** the family $\beta_\omega(x, \eta)J_0(x)$ may be too restrictive, especially near zeros or sign changes of J_0.
- **Error propagation and changing conditions:** identification delay, structural approximation error, stale parameters, and model error perturb both the critic and the improved policy.
- **Bellman and closed-loop consistency:** a plausibly shaped critic is not automatically Bellman-consistent, stabilizing, constraint-satisfying, or reliable outside the representative offline conditions.

4 A conventional approach — direct critic adaptation

A conventional adaptive-DP realization updates critic parameters directly from recent data and then improves the policy:

$$\begin{aligned} \theta_{t+1} &\approx \arg\min_{\theta} \mathcal{L}_{\text{critic}}(\theta; \mathcal{S}_t), \\ \widehat{\mu}_{t+1} &\leftarrow \text{Improve}\Big[\widehat{J}_{\theta_{t+1}}\Big]. \end{aligned} \tag{6}$$

One representative choice for $\mathcal{L}_{\text{critic}}$ is a sampled one-step TD loss. Holding the target critic $\widehat{J}_{\bar{\theta}_t}$ fixed during learner update t and taking $0 < \alpha < 1$, define

$$\begin{aligned} y_{k,t}^{\text{TD}} &= g_k + \alpha \widehat{J}_{\bar{\theta}_t}(x_{k+1}), \\ \mathcal{L}_{\text{critic}}(\theta; \mathcal{S}_t) &= \frac{1}{|\mathcal{S}_t|} \sum_{k \in \mathcal{I}_t} \Big(\widehat{J}_\theta(x_k) - y_{k,t}^{\text{TD}}\Big)^2. \end{aligned} \tag{7}$$

This form evaluates the policy that generated the transitions, subject to the usual on-policy or correction assumptions. If $\mathcal{S}_t$ mixes behavior policies or materially different environment conditions, the window must be restricted or an explicit off-policy/condition correction must be supplied. If a model and action minimization are available, an empirical Bellman-optimality target can be used instead. The selected critic loss and its policy-evaluation or policy-improvement role must be declared.

This can track the currently visited condition, but it does not explicitly preserve a reusable relation between the environment parameter and critic shape. Under limited data, it may repeatedly overwrite and relearn critic behavior as conditions change. This is a baseline mechanism, not a claim that every direct critic update necessarily forgets.

5 Candidate direction — identifier-conditioned critic reconfiguration

Offline, suppose approximate critics $\{\widehat{J}^{(m)}\}$ are available for representative conditions $\{\eta^{(m)}\}$ under a common state representation, objective, discount, and solution concept. In the optimal-control case, $\widehat{J}^{(m)} \approx J^{\star}_{\eta^{(m)}}$, where $J^{\star}_{\eta}$ denotes the optimal critic for the fixed-condition problem. Hold the reference critic J_0 fixed and fit the structural map through

$$\omega^{\dagger} \in \arg\min_{\omega} \sum_{m} \left\| \widehat{J}^{(m)}(\cdot) - \beta_{\omega}(\cdot, \eta^{(m)}) J_0(\cdot) \right\|^2_{\nu_m}, \tag{8}$$

where ν_m is a declared weighting distribution or measure over control-relevant states and $\|e\|^2_{\nu_m} := \int |e(x)|^2 \, d\nu_m(x)$. If J_0 is to be learned jointly instead, it must appear as an optimization variable and a normalization is needed to remove the scaling ambiguity between J_0 and β_{ω}. Online operation then follows

$$\begin{aligned} \mathcal{S}_t &\xrightarrow{\text{online identification}} \widehat{\eta}_t, \\ \widehat{J}_t(x) &= \beta_{\omega^{\dagger}}(x, \widehat{\eta}_t) J_0(x), \\ \widehat{\mu}_{t+1} &\leftarrow \text{Improve}\left[\widehat{J}_t; \widehat{\eta}_t\right]. \end{aligned} \tag{9}$$

A useful diagnostic, rather than a guarantee, separates two critic errors. For compactness, let $\eta_t := \eta_{k(t)}$ denote the physical environment condition associated with learner update t. The norm below must use the same declared control-relevant domain or measure for all terms:

$$\begin{aligned} \left\| \widehat{J}_t - J^{\star}_{\eta_t} \right\| \leq\ & \left\| \beta_{\omega^{\dagger}}(\cdot, \widehat{\eta}_t) J_0 - \beta_{\omega^{\dagger}}(\cdot, \eta_t) J_0 \right\| \\ & + \left\| \beta_{\omega^{\dagger}}(\cdot, \eta_t) J_0 - J^{\star}_{\eta_t} \right\|. \end{aligned} \tag{10}$$

The first term audits sensitivity to identification error; the second audits the structured critic family. Policy-improvement and model errors remain additional. Research questions include how to select η, how to learn J_0 and β_{ω}, when an additive or critic-bank structure is more appropriate, how to represent uncertainty in $\widehat{\eta}_t$, and when to reject reconfiguration in favor of a safe fallback.

Future evidence should compare a fixed J_0, direct critic adaptation, the structured method with oracle η, and the method with estimated η. Relevant metrics include identification error and delay, critic and Bellman error, closed-loop performance, violations, update latency, memory, and return-to-prior-condition performance under abrupt, gradual, recurring, and held-out conditions.

6 Application domains

- **Mobility controllers under changing mass, load, road–tire condition, temperature, or component aging:** estimate a compact condition parameter and reconfigure a stored critic so recurring operating regimes can reuse previously learned long-run control structure.

- **Electric-drive and mechatronic control** with compact identifiable plant parameters, such as resistance, inertia, friction, or load, that can condition a reusable critic family and its subsequent policy improvement.
- **Systems that revisit a finite or smoothly parameterized environment family:** recover or interpolate an appropriate critic when a known condition returns, rather than relearning the full value function from the latest trajectory. The benefit depends on identification quality and offline coverage of that environment family.