

# Observable-Only AI Safety from Public Data

Robust Bottleneck Diagnosis with Auditable No-Meta Dynamic Programming,  
Anytime Confidence Sequences, and Dynamic IQC

K. Takahashi

ORCID: 0009-0004-4273-3365

February 12, 2026

## Abstract

This paper presents an observable-only AI safety framework for robust bottleneck diagnosis from public data in coupled dynamical systems. Decisions are constrained to replay-visible evidence and authenticated exogenous governance updates (no hidden evaluators, no-meta). The target is not latent point attribution—which is generally non-identifiable under observational aliases—but reproducible interval diagnosis under explicit ambiguity. We integrate robust dynamic programming, partial identification, model-indexed e-processes/confidence sequences, and dynamic IQC analysis. The framework outputs (i) anytime-valid score intervals with explicit optimization, implementation, contamination, dependence, interaction, and rectangularization cushions, (ii) fail-closed declaration rules that emit unique bottleneck labels only under certified interval separation, (iii) time-consistent outer/inner ambiguity recursion, and (iv) deterministic replay contracts with signed governance updates and backup certificates. We prove robust Bellman well-posedness, measurable/constructive selector bridges, identification limits, branchwise guarantees (in-class statistical coverage vs. out-of-class safety behavior), and non-circular lag-one IQC tightening. For deployment, we provide a machine-checkable schema, cross-field replay invariants, and lightweight pseudocode for online operation and third-party full replay verification. The guarantee is accountable best-effort behavior under explicit assumptions, not recovery of latent ground truth.

**Keywords:** observable-only AI safety, robust bottleneck diagnosis, public-data decision-making, no-meta governance, robust MDP, partial identification, anytime-valid confidence sequence, e-processes, dynamic IQC, fail-closed operation, deterministic replay, auditable optimization certificates

## 1 Introduction

AGI/ASI capability scaling is constrained by interacting macro channels: institutions ( $I$ ), markets ( $M$ ), and physical infrastructure ( $P$ ), which together shape capability state ( $C$ ). In high-stakes regimes, delayed feedback, nonlinearity, adversarial contamination, and nonstationary drift are unavoidable. A no-meta governance stance forbids hidden arbiters: all admissibility, diagnosis, and intervention logic must be publicly reconstructible.

This paper addresses the following operational question:

Can a no-meta intelligence *robustly* diagnose macro bottlenecks from public history alone, while preserving third-party auditability?

**Epistemic scope and promise.** The protocol is an accountability machine, not a truth oracle. It guarantees that decisions, abstentions, alarms, and degradations are reproducible from public artifacts

and explicit budgets. It does not guarantee discovery of latent “truth itself” beyond what observable evidence can identify under the declared model class.

**System-generic interpretation.** The channel tuple  $(I, M, P, C)$  is a mnemonic, not a restriction. Any four (or more) coupled subsystems with delayed interactions can be embedded into the same construction by relabeling channels and preserving the public-history interface. Therefore the theory can be used beyond AI policy settings, including critical infrastructure operations, financial-physical supply chains, and federated cyber-physical systems.

The answer is yes—with mathematically explicit limits and explicit abstention regions, using robust MDP, partial identification, e-process, and IQC tools [3, 4, 7, 18, 12]. Exact latent attribution is generally impossible under observational aliases, but robust interval diagnosis is feasible under partial identification and robust control structure [7, 8, 3, 4]. Unique bottleneck declaration is valid only under interval separation; otherwise abstention and probing are required.

### Main technical contributions.

1. Full proofs for robust DP well-posedness, measurable interval endpoints, and dominance correctness.
2. Explicit time-consistent outer/inner ambiguity recursion, with coverage guarantees.
3. Anytime-valid multi-time guarantees using model-indexed e-processes.
4. Dynamic IQC (FIR multipliers) for delayed nonlinearity with a discounted- $\rho$  IQC dissipation proof.
5. Practical decision protocol mapping theory to auditable operations, including deterministic replay and fail-closed handling of empty operational model sets.

## 2 Model

### 2.1 Latent state, public history, and dynamics

Let  $(\Omega, \mathcal{F}, \mathbb{P})$  be complete with filtration  $(\mathcal{F}_t)_{t \geq 0}$ . Latent macro state:

$$X_t = (I_t, M_t, P_t, C_t) \in \mathcal{X} := \mathcal{X}_I \times \mathcal{X}_M \times \mathcal{X}_P \times \mathcal{X}_C.$$

Public observation  $O_t \in \mathcal{O}$ , public receipt  $Q_t \in \mathcal{Q}$ , action  $A_t = (U_t, V_t, W_t, Z_t) \in \mathcal{A}$ . Public history and pre-gate history:

$$H_t := (O_{0:t}, Q_{0:t}, A_{0:t-1}) \in \mathcal{H}_t, \quad \mathcal{G}_t := \sigma(H_t),$$

$$H_t^- := (O_{0:t}, Q_{0:t-1}, A_{0:t-1}), \quad \mathcal{G}_t^- := \sigma(H_t^-).$$

We use the ambient filtration

$$\mathcal{F}_t := \sigma(X_{0:t}, H_t), \quad \mathcal{F}_t^- := \sigma(X_{0:t}, H_t^-),$$

so that  $\mathcal{G}_t^- \subseteq \mathcal{G}_t \subseteq \mathcal{F}_t$  for all  $t$ . The pre-gate filtration  $\mathcal{G}_t^-$  is used when conditioning gate-activation probabilities so that the gate variable is not conditioned on itself.

Dynamics:

$$I_{t+1} = f_I(I_t, U_t, C_{t-\tau_I}, \xi_t^I), \quad (1)$$

$$M_{t+1} = f_M(M_t, V_t, I_t, \xi_t^M), \quad (2)$$

$$P_{t+1} = f_P(P_t, W_t, M_t, \xi_t^P), \quad (3)$$

$$C_{t+1} = (1 - \delta_C)C_t + \chi_t g(I_t, M_t, P_t, Z_t) + \xi_t^C, \quad \delta_C \in (0, 1). \quad (4)$$

Gate variable:

$$\chi_t = \mathbf{1}\{\Gamma(Q_t, H_t^-) \geq \vartheta_t\},$$

where  $\Gamma, \vartheta_t$  are public deterministic objects.

**Public scoring proxies.** To preserve observable-only semantics in optimization and audit, the protocol uses deterministic public maps

$$\widehat{C}_t := \psi_C(H_t), \quad \widehat{\ell}_s := \ell_s(H_t, A_t), \quad \widehat{\ell}_d := \ell_d(H_t, A_t),$$

with  $\psi_C, \ell_s, \ell_d$  published in the rulebook. Latent  $C_t$  remains part of the physical dynamics (4), while all score computations use  $\widehat{C}_t, \widehat{\ell}_s$ , and  $\widehat{\ell}_d$ .

**Initial conditions for delayed states.** Let  $h_{\max} := \max\{\tau_I, L\}$ . A public initialization record provides

$$X_{-h_{\max}:0}, \quad C_{-\tau_I:0}, \quad \zeta_0 = [v_{-1}^\top, \dots, v_{-L}^\top]^\top.$$

If unavailable, a deterministic default (e.g., all zeros) is used and logged. All replay and proofs condition on this published initialization.

## 2.2 Operational semantics of one-step objective

For each stage  $s$ , define a public transition-dependent one-step reward

$$r_s(h, a, h') := u(\widehat{C}'_{s+1}) - u(\widehat{C}_s) - \omega_s \ell_s(h, a) - \omega_d \ell_d(h, a),$$

where  $h' = H_{s+1}$  denotes next public history,  $\widehat{C}'_{s+1} := \psi_C(h')$ ,  $\widehat{C}_s := \psi_C(h)$ , and all rulebook objects  $(u, \psi_C, \ell_s, \ell_d, \omega_s, \omega_d)$  are public. Given kernel  $K$ , the expected stage reward is

$$\bar{R}_s^K(h, a) := \int r_s(h, a, h') \mathbf{P}(dh'|h, a, K).$$

This removes ambiguity about where model uncertainty enters: both transition and expected one-step objective depend on  $K$  through  $\mathbf{P}(\cdot|h, a, K)$ . In implementation, boundedness is enforced by auditable clipping/winsorization of primitive inputs before score computation, making assumptions replay-verifiable.

## 2.3 No-meta policy class

**Definition 2.1** (No-meta admissible policy). A policy  $\pi = (\pi_t)_{t \geq 0}$  belongs to  $\Pi_{\text{NM}}$  if for each  $t$ :

(NM1)  $A_t = \pi_t(H_t)$  is  $\mathcal{G}_t$ -measurable;

(NM2) all admissibility predicates are publicly specified;

- (NM3) no hidden variable can override admissibility;
- (NM4) contestability and challenge paths are public;
- (NM5) fail-closed fallback: verification failure implies  $A_t \in \mathcal{A}_{\text{safe}} \subseteq \mathcal{A}$ .

**Definition 2.2** (System-boundary no-meta semantics). The *no-meta* claim in this paper is internal-algorithmic: online decision rules may depend only on replay-visible observables, public rulebook constants, and authenticated external governance updates. No hidden latent-only constants are allowed inside the algorithmic core.

**Definition 2.3** (Exogenous governance channel). An exogenous governance channel  $g_t^{\text{exo}}$  is a replay-published, versioned, hash-linked stream of externally set parameters (e.g., risk budgets, step sizes, residual caps, and emergency thresholds) adopted by explicit boundary convention. All updates must be authenticated and timestamped; no silent parameter mutation is permitted. Formal channel-security conditions are stated in Assumption 2.4.

**Assumption 2.4** (Threat model and liveness contract for the exogenous governance channel). Governance updates in  $g_t^{\text{exo}}$  must satisfy all replay-verifiable conditions below:

- (g1) (**Threshold authentication**) each accepted update carries a  $k$ -of- $n$  threshold signature by registered signers.
- (g2) (**Anti-rollback**) updates are hash-chained with monotone sequence number and monotone timestamp; stale-sequence acceptance is forbidden.
- (g3) (**Fork choice**) if multiple signed candidates exist at the same height, a deterministic fork-choice rule is published and replayed.
- (g4) (**Multi-path dissemination**) updates are accepted from any authenticated transport among at least  $B_{\text{exo}} \geq 2$  replay-listed paths; path identity is logged.
- (g5) (**Signed contingency ladder with input-bound feasibility floor**) every accepted block carries a pre-signed contingency bundle

$$\mathcal{G}_t^{\text{lad}} := \{g_{t,\tau}^{\text{lad}}\}_{\tau=1}^{H_{\text{grace}}},$$

with monotone-tightening property on declared risk budget

$$\mathcal{R}(g_{t,\tau+1}^{\text{lad}}) \subseteq \mathcal{R}(g_{t,\tau}^{\text{lad}}) \subseteq \mathcal{R}(g_t^{\text{fresh}}).$$

Each ladder element  $g_{t,\tau}^{\text{lad}}$  includes hashes of a replay-public feasibility witness family  $\mathcal{W}_{t,\tau}^{\text{feas}}$  and solver profile  $\mathbf{m}_{t,\tau}^{\text{feas}}$ . The feasibility floor is defined by

$$\Phi_{t,\tau}^{\text{lb}} := \underline{\text{Feas}}(h_t, g_{t,\tau}^{\text{lad}}, \mathfrak{S}_t^{\text{core}}, \mathcal{W}_{t,\tau}^{\text{feas}}, \mathbf{m}_{t,\tau}^{\text{feas}}) \geq \Phi_{\min,t}^{\text{exo}} > 0,$$

where  $\underline{\text{Feas}}$  is a deterministic lower bound on executable safe-action volume computed only from replay-visible inputs.

- (g6) (**Outage classification**) cryptographic invalidity (signature/fork/rollback failure) is distinguished from pure connectivity outage (no new block, last valid block intact).

(g7) (**Bounded stale operation**) under pure connectivity outage of run length  $\ell_t$ , the protocol may execute using  $g_{t,\ell_t}^{\text{lad}}$  only while  $\ell_t \leq H_{\text{grace}}$ . If  $\ell_t > H_{\text{grace}}$ , governance validity is set to zero and certificate-ordered fallback is mandatory (shield-only if core certificate valid, else graceful-degradation if parking certificate valid, else emergency-stop).

(g8) (**Anti-suffocation guard**) if  $\Phi_{t,\ell_t}^{\text{lb}} < \Phi_{\min,t}^{\text{exo}}$  at any stale step, ladder execution is forbidden and the mode must switch to graceful-degradation branch with parking guarantees from Assumption 5.20; hard-stop is used only if the graceful-degradation certificate is also invalid.

All mode transitions  $\{\text{fresh}, \text{ladder}, \text{graceful\_degradation}, \text{shield\_only}, \text{emergency\_stop}\}$ , outage counters, selected ladder indices, and feasibility-floor values are replay-logged.

**Proposition 2.5** (No hidden closure assumption). *If a quantity is not identifiable from observables under the stated model class, it must appear either as (i) an explicit exogenous governance parameter in  $g_t^{\text{exo}}$ , or (ii) an explicit uncertainty term carried through guarantees. Unknown terms may not be hidden inside undocumented constants.*

**Proposition 2.6** (Governance outage resilience with bounded autonomous window). *Assume Assumptions 5.20 and 2.4. If cryptographic validity fails at time  $t$  (signature, rollback, or fork-rule failure), then: (i) shield-only mode is used when a valid nontrivial core certificate exists, (ii) otherwise graceful-degradation mode is used when a valid parking/acceptability certificate exists, (iii) emergency-stop is used only if both certificates are invalid. Alarm ***governance\_invalid*** is mandatory. If only connectivity outage occurs and outage run length satisfies  $\ell_t \leq H_{\text{grace}}$ , execution may continue with the signed contingency profile  $g_{t,\ell_t}^{\text{lad}}$ , preserving all safety guarantees that are monotone under budget tightening (possibly with degraded performance) provided the anti-suffocation floor  $\Phi_{t,\ell_t}^{\text{lb}} \geq \Phi_{\min,t}^{\text{exo}}$  holds. If  $\Phi_{t,\ell_t}^{\text{lb}} < \Phi_{\min,t}^{\text{exo}}$ , the protocol must switch to graceful-degradation mode with the parking certificate  $(\mathcal{X}_t^{\text{park}}, E_{\text{park},t}, S_{\min,t})$ . If  $\ell_t > H_{\text{grace}}$ , bounded autonomous operation is exhausted and the same certificate-ordered fallback applies (shield-only if core certificate valid, else graceful-degradation if parking certificate valid, else emergency-stop), with alarm ***governance\_timeout***. No silent continuation under unbounded staleness is permitted.*

*Proof.* Assumption 2.4 provides replay-verifiable mode predicates, signed contingency ladder, and feasibility-floor guards. Assumption 5.20 provides executable certificate-ordered fallback branches independent of online global optimization. Cryptographic invalidity removes governance authenticity, so fresh-governance execution is forbidden; fallback then follows certificate availability (core-certificate  $\Rightarrow$  shield-only, else parking-certificate  $\Rightarrow$  graceful degradation, else emergency-stop). Pure connectivity outage preserves authenticity of the last valid block, so bounded continuation on pre-signed profiles is admissible only while both stale horizon and feasibility-floor constraints hold. Violation of the feasibility floor triggers graceful degradation rather than continued tightening, preventing self-denial by vanishing feasibility. Beyond  $H_{\text{grace}}$ , certificate-ordered fallback is mandatory and silent continuation is impossible.  $\square$

## 2.4 Objective

For discount  $\gamma \in (0, 1)$ :

$$J_{\theta}^{\pi}(h_t) = \mathbb{E}_{\theta}^{\pi} \left[ \sum_{s=t}^{\infty} \gamma^{s-t} \left( u(\widehat{C}_{s+1}) - u(\widehat{C}_s) - \omega_s \ell_s(H_s, A_s) - \omega_d \ell_d(H_s, A_s) \right) \middle| H_t = h_t \right].$$

This aligns the long-horizon objective with the finite-horizon robust value increment used in diagnostics.

### 3 Uncertainty, contamination, drift, and dependence

**Definition 3.1** (Evidence-consistent ambiguity set). At history  $h_t$ ,  $\Theta_t^{\text{ev}}(h_t) \subseteq \mathcal{T}$  is the set of models not falsified by publicly available evidence and rulebook up to  $t$ .

**Rectangular uncertainty.** For  $\theta \in \Theta_t^{\text{ev}}(h_t)$ , define stagewise uncertainty set  $\mathcal{U}_s^\theta(h_s, a) \subseteq \mathcal{K}$ . Over horizon  $t:T-1$ ,

$$\mathcal{U}_{t:T-1}^\theta(h_t, \pi) := \prod_{s=t}^{T-1} \mathcal{U}_s^\theta(h_s, \pi_s(h_s)).$$

Rectangularity is required for dynamic consistency in robust recursive decision models [3, 4, 5]. Without rectangularity, stagewise minimizers need not compose into a globally worst-case path measure, and Bellman recursion can become time-inconsistent.

**Contamination and drift budgets.**

$$\mathbb{P}_n^{\text{obs}} = (1 - \eta_n) \mathbb{P}_n^* + \eta_n \mathbb{Q}_n, \quad 0 \leq \eta_n \leq \bar{\eta} < \frac{1}{2},$$

$$D_{t,T} := \sum_{s=t}^{T-1} \sup_{x,a} \text{TV}(K_{s+1}^*(\cdot|x, a), K_s^*(\cdot|x, a)).$$

**Dependence.** Allow  $\beta$ -mixing with  $\sum_{\ell \geq 1} \beta(\ell) < \infty$ , and define effective size at time  $t$ :

$$n_{\text{eff},t} := \frac{t+1}{1 + 2 \sum_{\ell=1}^t \beta(\ell)}.$$

This quantity is used in conservative dependence cushions for anytime intervals.

## 4 Interaction-aware bottleneck diagnostics

### 4.1 Relief value

Let  $\mathcal{J} := \{I, M, P\}$ , relief  $r = (r_I, r_M, r_P) \in \mathcal{R} := \prod_{j \in \mathcal{J}} [0, \bar{r}_j]$ , and monotone admissibility

$$r \preceq r' \implies \mathcal{A}_{\text{NM}}(h, r) \subseteq \mathcal{A}_{\text{NM}}(h, r').$$

Define the relief-constrained no-meta policy class

$$\Pi_{\text{NM}}(r) := \{\pi \in \Pi_{\text{NM}} : \pi_s(h) \in \mathcal{A}_{\text{NM}}(h, r), \forall s, \forall h \in \mathcal{H}_s\}.$$

Operationally,  $r$  parametrizes admissibility via publicly specified monotone constraints (e.g., channel caps, budget splits, or actuator-rate limits); it need not be an additive shift in latent dynamics and is interpreted as a governance-level intervention envelope. For horizon  $H \in \mathbb{N}$ ,  $T = t + H$ , define robust value. Unless stated otherwise,  $H$  is fixed ex ante and published. Adaptive horizons  $H_t$  are allowed only when  $H_t$  is  $\mathcal{G}_t$ -predictable and the global risk budget is pre-split across the horizon menu:

$$\sum_{h \in \mathcal{H}} \alpha_{\text{out}}(h) \leq \alpha_{\text{out}}, \quad \sum_{h \in \mathcal{H}} \alpha_{\text{ev}}(h) \leq \alpha_{\text{ev}}, \quad \sum_{h \in \mathcal{H}} \alpha_{\text{dep}}(h) \leq \alpha_{\text{dep}}.$$

$$V_{t,H}^\theta(h_t; r) := \sup_{\pi \in \Pi_{\text{NM}}(r)} \inf_{K \in \mathcal{U}_{t:T-1}^\theta(h_t, \pi)} \mathbb{E}^{\pi, K} \left[ \sum_{s=t}^{T-1} \gamma^{s-t} r_s(H_s, A_s, H_{s+1}) \middle| H_t = h_t \right]. \quad (5)$$

Relief gain:

$$G_{t,H}^\theta(h_t; r) := V_{t,H}^\theta(h_t; r) - V_{t,H}^\theta(h_t; \mathbf{0}).$$

## 4.2 Finite-difference diagnostics

Fix  $\delta > 0$ , unit vectors  $e_j$ .

### First-order pressure

$$\beta_j^\theta := \frac{G_{t,H}^\theta(h_t; \delta e_j) - G_{t,H}^\theta(h_t; \mathbf{0})}{\delta}.$$

### Pair interaction

$$\Delta_{jk}^\theta := G_{t,H}^\theta(h_t; \delta(e_j + e_k)) - G_{t,H}^\theta(h_t; \delta e_j) - G_{t,H}^\theta(h_t; \delta e_k) + G_{t,H}^\theta(h_t; \mathbf{0}).$$

**Third-order interaction** The decomposition follows the Möbius expansion on the Boolean lattice and captures complementarities/supermodularity effects in multi-channel interventions [23, 24].

$$\begin{aligned} \Delta_{IMP}^\theta &:= G_{t,H}^\theta(h_t; \delta(e_I + e_M + e_P)) \\ &\quad - \sum_{\{j,k\} \subset \mathcal{J}} G_{t,H}^\theta(h_t; \delta(e_j + e_k)) + \sum_{j \in \mathcal{J}} G_{t,H}^\theta(h_t; \delta e_j) \\ &\quad - G_{t,H}^\theta(h_t; \mathbf{0}). \end{aligned} \quad (6)$$

### Interaction-aware score

$$S_j^\theta := \beta_j^\theta + \frac{w_2}{2\delta} \sum_{k \neq j} \Delta_{jk}^\theta + \frac{w_3}{3\delta} \Delta_{IMP}^\theta, \quad (7)$$

where  $w_2, w_3 \in [0, 1]$  are published calibration weights (default  $w_2 = w_3 = 1$ ). This keeps units aligned with the first-order term while allowing conservative attenuation of higher-order terms when finite-difference stencils are noisy or sparse. If time-varying weights are used in implementation, replay metadata must publish  $(w_{2,t}, w_{3,t})$  and all contamination cushions must use the corresponding time-indexed values.

## 4.3 Coherent robust intervals and dominance

$$\underline{S}_{j,t} := \inf_{\theta \in \Theta_t^{\text{ev}}(h_t)} S_j^\theta, \quad \overline{S}_{j,t} := \sup_{\theta \in \Theta_t^{\text{ev}}(h_t)} S_j^\theta.$$

(Important: compute by direct optimization over (7), not by independent endpoint aggregation.)

For margin  $\varepsilon \geq 0$ :

$$\mathcal{B}_t^\varepsilon := \left\{ j \in \mathcal{J} : \underline{S}_{j,t} \geq \max_{k \neq j} \overline{S}_{k,t} + \varepsilon \right\}.$$

## 5 Assumptions

**Assumption 5.1** (Spaces and admissibility).  $\mathcal{X}, \mathcal{A}, \mathcal{T}$  are standard Borel.  $\mathcal{A}_{\text{NM}}(h, r)$  is nonempty compact convex, measurable in  $h$ , upper hemicontinuous in  $r$ . If primitive controls are discrete, convexity is interpreted on relaxed (randomized) controls over the simplex; implementation may use measurable purification when available.

**Assumption 5.2** (Bounded objective kernel and weak continuity on Polish history space). For each stage  $s$ , there exists  $B < \infty$  such that for all admissible  $(h, a)$ , all admissible  $K \in \mathcal{U}_s^\theta(h, a)$ , and  $\mathbb{P}(\cdot|h, a, K)$ -a.e.  $h'$ ,

$$|r_s(h, a, h')| \leq B.$$

Hence  $|\bar{R}_s^K(h, a)| \leq B$  for every admissible  $(h, a, K)$ . A public pre-processing rule (e.g., clipping/winsorization thresholds) that guarantees this bound is published and replay-verifiable. For each  $t$ , the public history space  $\mathcal{H}_t$  is Polish (or Borel-isomorphic to a Polish space), and  $(h, a, K) \mapsto \int f(h') \mathbb{P}(dh'|h, a, K)$  is jointly measurable for every bounded Borel  $f$ , with weak continuity in  $(h, a, K)$  under the chosen Polish topology.

**Assumption 5.3** (Rectangular robust uncertainty).  $\mathcal{U}_s^\theta(h, a)$  is nonempty compact convex and measurable in  $(h, a)$ , with product structure across time.

**Proposition 5.4** (Why rectangularity is required for dynamic consistency). *Under robust Bellman recursion, rectangular uncertainty in Assumption 5.3 is sufficient for time consistency of the minimax value. If uncertainty is non-rectangular across time, there exist finite-horizon examples where the ex-ante minimax action is not continuation-optimal after observing interim history, i.e., dynamic inconsistency appears.*

*Proof.* Sufficiency follows from stagewise decomposition used in robust MDP dynamic programming and rectangular recursive ambiguity models [3, 4, 5]. For non-rectangular sets, coupling across stages breaks separability of the inner minimization, so conditional continuation problems need not agree with the ex-ante minimax plan; standard two-stage counterexample constructions apply.  $\square$

**Assumption 5.5** (Random-set measurability and compactness). For each  $t$ ,  $h \mapsto \Theta_t^{\text{ev}}(h)$  has measurable graph and nonempty compact values in a complete separable metric parameter space  $(\Theta_0, d_\Theta)$ .

**Assumption 5.6** (Gate-integrity bounds). There exist public constants  $\varepsilon_{\text{fa}}, \varepsilon_{\text{sv}} \in [0, 1)$  such that

$$\mathbb{P}(\text{false-accept} \mid \mathcal{G}_t^-) \leq \varepsilon_{\text{fa}}, \quad \mathbb{P}(\text{split-view} \mid \mathcal{G}_t^-) \leq \varepsilon_{\text{sv}}.$$

Define  $\varepsilon_{\text{gate}} := \varepsilon_{\text{fa}} + \varepsilon_{\text{sv}}$ .

**Assumption 5.7** (Physical throughput envelope).  $g(I, M, P, z) \leq \bar{g}_0 + \bar{g}_P \psi_P(P) \leq \bar{g}_0 + \bar{g}_P P_{\max}$ .

**Assumption 5.8** (Concentration-fragility relation). Let  $\hat{\kappa}_t := \psi_\kappa(H_t^-) \in [0, 1]$  be a public concentration proxy that is  $\mathcal{G}_t^-$ -measurable (predictable), and let  $\phi : [0, 1] \rightarrow [0, 1]$  be increasing with  $\phi(0) = 0$ . Then

$$\mathbb{E}[\chi_t \mid \mathcal{G}_t^-] \leq \min\{1, 1 - \phi(\hat{\kappa}_t) + \varepsilon_{\text{gate}}\}.$$

**Assumption 5.9** (Auditable optimization certificates). At each time  $t$ , for each channel  $j$ , endpoint optimization returns deterministic certificates

$$L_{j,t}^{\text{inf}}, U_{j,t}^{\text{sup}}, \varepsilon_{\text{opt},j,t}^{(L)}, \varepsilon_{\text{opt},j,t}^{(U)} \geq 0$$

with replay metadata (solver version, tolerance, seed, arithmetic mode), and

$$0 \leq \underline{S}_{j,t} - L_{j,t}^{\text{inf}} \leq \varepsilon_{\text{opt},j,t}^{(L)}, \quad 0 \leq U_{j,t}^{\text{sup}} - \bar{S}_{j,t} \leq \varepsilon_{\text{opt},j,t}^{(U)}.$$

Define channelwise inflation

$$\varepsilon_{\text{opt},j,t} := \max\{\varepsilon_{\text{opt},j,t}^{(L)}, \varepsilon_{\text{opt},j,t}^{(U)}\},$$



and global inflation

$$\varepsilon_{\text{opt},t} := \max_{j \in \mathcal{J}} \varepsilon_{\text{opt},j,t}.$$

For nonconvex endpoint programs, certificates are produced by deterministic global branch-and-bound (or equivalent verified bounding); for convex endpoint programs, primal-dual certificates are acceptable.

**Assumption 5.10** (Regularity of robust gains and score map). **(A9a) Stencil measurability.** Let

$$\mathcal{R}_\delta := \{\mathbf{0}, \delta e_I, \delta e_M, \delta e_P, \delta(e_I + e_M), \delta(e_I + e_P), \delta(e_M + e_P), \delta(e_I + e_M + e_P)\}.$$

For each  $r \in \mathcal{R}_\delta$ ,  $(h, \theta) \mapsto G_{t,H}^\theta(h; r)$  is jointly measurable.

**(A9b) Score-map regularity.** For each  $j \in \mathcal{J}$ ,  $(h, \theta) \mapsto S_j^\theta(h)$  from (7) is jointly measurable, and  $\theta \mapsto S_j^\theta(h)$  is  $L_S$ -Lipschitz under  $d_\Theta$ :

$$|S_j^\theta(h) - S_j^{\theta'}(h)| \leq L_S d_\Theta(\theta, \theta').$$

**(A9c) Envelope boundedness.** There exists  $\bar{S} < \infty$  such that  $|S_j^\theta(h)| \leq \bar{S}$  for all  $j, \theta, h$  in operational domains.

**Assumption 5.11** (Model-indexed e-process validity and regularity). For each  $\theta \in \Theta_0$ ,  $E_t(\theta) \geq 0$  is adapted with  $E_0(\theta) = 1$  and

$$\mathbb{E}_\theta[E_{t+1}(\theta) \mid \mathcal{F}_t] \leq E_t(\theta) \quad \forall t \geq 0.$$

For each  $t$ ,  $\theta \mapsto E_t(\theta)$  is Borel measurable.

**Definition 5.12** (Operational intersection rule). At each time  $t$ , define

$$\begin{aligned} \Theta_t^{\text{op}, \text{exact}}(\alpha_{\text{out}}) &:= \Theta_t^{\text{ev}}(h_t) \cap \Theta_t^{\text{out}}(\alpha_{\text{out}}), \\ \Theta_t^{\text{op}}(\alpha_{\text{out}}) &:= \Theta_t^{\text{ev}}(h_t) \cap \Theta_t^{\text{out}, \text{num}}(\alpha_{\text{out}}), \end{aligned}$$

where both  $\Theta_t^{\text{ev}}(h_t)$  and  $\Theta_t^{\text{out}, \text{num}}$  are deterministic functions of public logs and published numerical profile. Theorems are stated on  $\Theta_t^{\text{out}}$ , and transferred to  $\Theta_t^{\text{out}, \text{num}}$  via Proposition 8.2.

*Remark 5.13* (Realized-history notation). When  $t$  is fixed and a realized history  $h_t$  is given, we write  $\Theta_t^{\text{ev}}(h_t)$ . Inside probability events under the data-generating process, the random counterpart is written as  $\Theta_t^{\text{ev}}(H_t)$ . This is purely notational:  $\Theta_t^{\text{ev}}(\cdot)$  is the same deterministic set-valued map evaluated at either a fixed argument or the realized random history.

**Assumption 5.14** (Truth-retention of evidence-consistent set). There exists a public budget  $\alpha_{\text{ev}} \in [0, 1)$  such that for the data-generating model  $\theta^*$ ,

$$\mathbb{P}_{\theta^*}(\forall t \geq 0 : \theta^* \in \Theta_t^{\text{ev}}(H_t)) \geq 1 - \alpha_{\text{ev}}.$$

A sufficient auditable construction is to define  $\Theta_t^{\text{ev}}(H_t)$  via model-indexed evidence e-process thresholds  $1/\alpha_{\text{ev}}$  (Proposition 9.1).

**Predictable replay baseline for dependence diagnostics.** For each model  $\theta$ , channel  $j$ , and time  $t \geq 1$ , define a deterministic replay predictor

$$\tilde{S}_{j,t}^\theta := \varphi_{j,t}^\theta(H_{0:t-1}),$$

where  $\varphi_{j,t}^\theta$  is a published map and  $\tilde{S}_{j,t}^\theta$  is  $\mathcal{G}_{t-1}$ -measurable. Set  $\tilde{S}_{j,0}^\theta := 0$ . Residuals are

$$R_{j,t}^\theta := S_j^\theta(H_t) - \tilde{S}_{j,t}^\theta.$$

By convention,  $R_{j,0}^\theta := 0$ .

**Assumption 5.15** (Dependence-cushion primitives). For each  $j \in \mathcal{J}$ , define replay residual

$$R_{j,t} := R_{j,t}^{\theta^*} = S_j^{\theta^*}(H_t) - \tilde{S}_{j,t}^{\theta^*}.$$

Assume:

- (d1) **Predictable centering:**  $\mathbb{E}[R_{j,t} \mid \mathcal{G}_{t-1}] = 0$  for all  $t \geq 1$ .
- (d2) **Conditional MGF bound:** there exist predictable  $\sigma_{j,t}^2$  and deterministic  $c_R$  such that for all  $\lambda \in [0, 1/c_R)$ ,
$$\mathbb{E}\left[e^{\lambda R_{j,t}} \mid \mathcal{G}_{t-1}\right] \leq \exp\left(\frac{\lambda^2 \sigma_{j,t}^2}{2(1 - \lambda c_R)}\right).$$
- (d3) **Dependence-envelope metadata:** replay publishes conservative envelopes  $n_{\text{eff},t}^{\text{lb}} > 0$ ,  $\bar{\sigma}_{S,t}^{\text{pub}}$ , and  $\bar{\beta}_t^{\text{pub}}$  that are deterministic functions of public logs and may only tighten across republications.
- (d4) **Anytime-valid residual envelope target:** for each channel  $j$ , replay specifies a bound sequence  $u_{j,t}(\alpha_{\text{dep}})$  such that

$$\mathbb{P}(\forall t \geq 0 : |R_{j,t}| \leq u_{j,t}(\alpha_{\text{dep}})) \geq 1 - \frac{\alpha_{\text{dep}}}{|\mathcal{J}|}.$$

Define the operational dependence cushion by

$$b_t^{\text{dep}} := \max_{j \in \mathcal{J}} u_{j,t}(\alpha_{\text{dep}}).$$

**Assumption 5.16** (Observable-only calibration of dependence/contamination envelopes). (Deterministic-by-construction *or* anytime-valid calibrated mode.) Either:

- (a) **Deterministic physical mode:** public worst-case engineering limits and metrology certificates imply deterministic bounds  $\bar{\eta}_t^{\text{pub}}, \bar{\beta}_t^{\text{pub}}, \bar{\sigma}_{S,t}^{\text{pub}}, n_{\text{eff},t}^{\text{lb}}$  for all  $t$ , with no stochastic calibration error (set  $\alpha_{\text{env}} := 0$ ); or
- (b) **Calibrated statistical mode:** there exist replay-published observable monitors  $Z_t^{(\eta)}, Z_t^{(\beta)}, Z_t^{(\sigma)}, Z_t^{(n)}, Z_t^{(\text{link})}$  and e-processes  $M_t^{(\eta)}, M_t^{(\beta)}, M_t^{(\sigma)}, M_t^{(n)}, M_t^{(\text{link})}$ , initialized at 1, adapted to  $\mathcal{G}_t^-$ , with each satisfying the e-supermartingale property under the corresponding envelope-validity null. Define channel alarms

$$\mathcal{A}_t^{(q)} := \left\{ \sup_{s \leq t} M_s^{(q)} \geq \frac{1}{\alpha_q} \right\}, \quad q \in \{\eta, \beta, \sigma, n, \text{link}\},$$

with  $\sum_q \alpha_q \leq \alpha_{\text{env}}$ . On  $(\cap_q (\mathcal{A}_\infty^{(q)})^c)$ , published envelopes are valid for all  $t$ .

In both modes, the dependence cushion is replay-defined as

$$b_t^{\text{dep,pub}} := \bar{\sigma}_{S,t}^{\text{pub}} \sqrt{\frac{2 \log\left(\frac{4|\mathcal{J}|(t+1)^2}{\alpha_{\text{dep}}}\right)}{n_{\text{eff},t}^{\text{lb}} \vee 1}} + \frac{3c_R \log\left(\frac{4|\mathcal{J}|(t+1)^2}{\alpha_{\text{dep}}}\right)}{n_{\text{eff},t}^{\text{lb}} \vee 1},$$

and the protocol sets  $b_t^{\text{dep}} := b_t^{\text{dep, pub}}$ , which must satisfy

$$b_t^{\text{dep, pub}} \geq \max_{j \in \mathcal{J}} u_{j,t}(\alpha_{\text{dep}}) \quad \forall t,$$

and a replay-derived interaction remainder bound

$$|\Xi_{j,t}^{\text{int}}| \leq b_t^{\text{int, pub}} \quad \forall j, t.$$

To avoid a single opaque link constant, publish a decomposed monitor-latent cushion

$$b_t^{\text{link, pub}} = \hat{b}_t^{\text{link}} + u_t^{\text{link}}(\alpha_{\text{link}}) + r_t^{\text{ow}},$$

and enforce

$$\left| (\Delta_{j,t}^{\text{cont}} + \Delta_{j,t}^{\text{dep}}) - (\Delta_{j,t}^{\text{cont, proxy}} + \Delta_{j,t}^{\text{dep, proxy}}) \right| \leq b_t^{\text{link, pub}} \quad \forall j, t,$$

where  $\hat{b}_t^{\text{link}}$  is a replay-deterministic proxy residual predictor,  $u_t^{\text{link}}(\alpha_{\text{link}})$  is an anytime-valid uncertainty radius from lower-level calibration, and  $r_t^{\text{ow}}$  is computed by an explicit replay-published map

$$r_t^{\text{ow}} := \left[ \rho_{0,t}^{\text{exo}} + \rho_{1,t}^{\text{exo}} \log^+(M_t^{\text{mis}}) + \rho_{2,t}^{\text{exo}} \text{NVS}_t \right]_0^{r_{\text{max},t}^{\text{exo}}} + u_t^{\text{ow}}(\alpha_{\text{ow}}),$$

where  $\text{NVS}_t$  is a replay-verifiable novelty score and  $(\rho_{0,t}^{\text{exo}}, \rho_{1,t}^{\text{exo}}, \rho_{2,t}^{\text{exo}}, r_{\text{max},t}^{\text{exo}})$  are supplied only via the public exogenous governance channel  $g_t^{\text{exo}}$ . Define envelope-validity event

$$\mathcal{E}_{\text{env}} := \left\{ \forall t, j : |\Delta_{j,t}^{\text{cont}} + \Delta_{j,t}^{\text{dep}}| \leq b_t^{\text{dep, pub}} + b_t^{\text{link, pub}}, |\Xi_{j,t}^{\text{int}}| \leq b_t^{\text{int, pub}} \right\}, \quad \mathbb{P}(\mathcal{E}_{\text{env}}) \geq 1 - \alpha_{\text{env}}.$$

All cushions are computed from  $H_t, Z_t^{(\cdot)}, M_t^{(\cdot)}, g_t^{\text{exo}}$ , public constants, metadata), not latent quantities. Thus monitor calibration is treated as a *necessary but not sufficient* proxy condition, and unresolved proxy-latent mismatch is carried explicitly (never hidden) via  $r_t^{\text{ow}}$ .

**Assumption 5.17** (Continuous-parameter e-process implementation with certified numerical envelope). When  $\Theta_0$  is not finite, implementation uses either:

- (f1) a finite replay-published mesh  $\Theta_0^\Delta \subset \Theta_0$  with covering radius  $\Delta_t$  and Lipschitz envelope  $L_E(t)$ , or
- (f2) a replay-published mixture e-process  $E_t^{\text{mix}}$  with valid Ville-style thresholding.

In either mode, a deterministic replay-verifiable numerical envelope  $\varepsilon_{E,t}^{\text{num}} \geq 0$  is published such that

$$\sup_{\theta \in \Theta_0} \left( \sup_{s \leq t} E_s^{\text{impl}}(\theta) - \sup_{s \leq t} E_s(\theta) \right) \leq \varepsilon_{E,t}^{\text{num}} \quad \forall t \geq 0.$$

(If separate approximation and floating-point envelopes are available,  $\varepsilon_{E,t}^{\text{num}}$  is their sum.) A boundary-stability tolerance  $\tau_{\text{num}} \geq 0$  is published and used in declaration predicates to prevent threshold jitter near decision boundaries.

**Assumption 5.18** (Rectangularization, certified distortion, and model-class alarm). To preserve dynamic-programming time consistency, operational ambiguity set is rectangularized [3, 4, 5]:

$$\Theta_t^{\text{ev}}(h_t) := \mathfrak{R}_t(\Theta_t^{\text{raw}}(h_t)),$$

where  $\Theta_t^{\text{raw}}(h_t)$  is the pre-rectangular evidence-consistent set, and

$$\Theta_t^{\text{raw}}(h_t) \subseteq \Theta_t^{\text{ev}}(h_t) \subseteq \Theta_0$$

holds by protocol construction. Exact global optimization on  $\Theta_t^{\text{raw}}(h_t)$  is *not* required. Instead, replay artifacts must publish a *relaxation ladder*

$$\Theta_t^{\text{in,raw}}(h_t) \subseteq \Theta_t^{\text{raw}}(h_t) \subseteq \Theta_t^{\text{out,raw}}(h_t),$$

where  $\Theta_t^{\text{in,raw}}(h_t)$  is a deterministic feasible archive (inner approximation) and  $\Theta_t^{\text{out,raw}}(h_t)$  is a deterministic convex/SDP-safe outer approximation. For each  $t, j$ , replay-published optimization brackets satisfy

$$\begin{aligned} \underline{U}_{j,t}^{\text{raw}} &\leq \sup_{\theta \in \Theta_t^{\text{raw}}(h_t)} S_{j,t}^\theta \leq \overline{U}_{j,t}^{\text{raw}}, & \underline{U}_{j,t}^{\text{rect}} &\leq \sup_{\theta \in \Theta_t^{\text{ev}}(h_t)} S_{j,t}^\theta \leq \overline{U}_{j,t}^{\text{rect}}, \\ \underline{L}_{j,t}^{\text{raw}} &\leq \inf_{\theta \in \Theta_t^{\text{raw}}(h_t)} S_{j,t}^\theta \leq \overline{L}_{j,t}^{\text{raw}}, & \underline{L}_{j,t}^{\text{rect}} &\leq \inf_{\theta \in \Theta_t^{\text{ev}}(h_t)} S_{j,t}^\theta \leq \overline{L}_{j,t}^{\text{rect}}, \end{aligned}$$

and define the certified distortion

$$\hat{\delta}_{j,t}^{\text{rect}} := \max\left\{0, \overline{U}_{j,t}^{\text{rect}} - \underline{U}_{j,t}^{\text{raw}}, \overline{L}_{j,t}^{\text{raw}} - \underline{L}_{j,t}^{\text{rect}}\right\}, \quad b_t^{\text{rect, pub}} := \max_{j \in \mathcal{J}} \hat{\delta}_{j,t}^{\text{rect}}.$$

The lower/upper brackets may be produced by any deterministic combination of (i) feasible-point evaluation on  $\Theta_t^{\text{in,raw}}(h_t)$ , (ii) convex/SDP dual certificates on  $\Theta_t^{\text{out,raw}}(h_t)$ , and (iii) anytime branch-and-bound residuals. If these certificates are unavailable or numerically invalid, the protocol sets  $b_t^{\text{rect, pub}} := +\infty$  and must output compute-limited soft abstention (fail-closed declaration logic). In addition, a model-class adequacy e-process  $M_t^{\text{mis}}$  is published, with  $M_0^{\text{mis}} = 1$ , adapted to  $\mathcal{G}_t^-$ , and satisfying

$$\mathbb{E}_{\theta^*} \left[ M_t^{\text{mis}} \mid \mathcal{G}_{t-1}^- \right] \leq M_{t-1}^{\text{mis}} \quad \text{whenever } \theta^* \in \Theta_0.$$

Define the replay-verifiable alarm

$$A_t^{\text{mis}} := \mathbf{1} \left\{ \sup_{s \leq t} M_s^{\text{mis}} \geq \frac{1}{\alpha_{\text{mis}}} \right\}.$$

If  $A_t^{\text{mis}} = 1$ , the protocol enters fail-closed safe mode and suppresses unique bottleneck declarations until re-specification.

**Assumption 5.19** (Optional minimax interchange regularity). (Used only for optional saddlepoint statements.) For each  $(s, h, r)$ , the robust stage game has action-independent compact convex kernel set  $\mathcal{U}_s^\theta(h)$ , and payoff  $\Phi(h, a, K; v)$  is concave upper-semicontinuous in  $a$ , convex lower-semicontinuous in  $K$ , and jointly measurable.

**Probe design objects and forced baseline class.** At each time  $t$ , let  $\mathcal{P}_t$  be a compact Borel space of replay-verifiable probe designs, with distinguished null element  $0 \in \mathcal{P}_t$  (no probe). Let  $\Pi_{\text{forced}}$  denote the class of universally measurable policies that always commit to exactly one channel and set  $p_t = 0$  for all  $t$ .

**Assumption 5.20** (Probe optimization feasibility, exogenous-governed hazard dual tracking, and certified backup with graceful degradation). For each  $t, h_t$ , the probe-feasible set  $\mathcal{U}_t^{\text{probe}}(h_t) \subseteq \mathcal{R} \times \mathcal{P}_t$  is nonempty compact with measurable graph and contains a safe fallback element  $(r_t^{\text{safe}}(h_t), 0)$ . The fallback is generated by a replay-published backup stack

$$r_t^{\text{safe}}(h_t) = \Pi_t^{\text{sh}}(\kappa_t^{\text{bk}}(h_t)),$$

where  $\kappa_t^{\text{bk}}$  is a low-order backup controller and  $\Pi_t^{\text{sh}}$  is a safety shield projection.

The shield must be accompanied by one of the following replay-verifiable certificates at each governance epoch:

- (bk1) (**Nontrivial core certificate**) a controlled-invariant core set  $\mathcal{X}_t^{\text{core}} \neq \emptyset$ , witness tuple  $(B_t, \kappa_t^{\text{bk}}, \varepsilon_t^{\text{inv}})$ , and a nontriviality index

$$\nu_t^{\text{core}} := \frac{\text{Vol}(\mathcal{X}_t^{\text{core}})}{\text{Vol}(\mathcal{X}_t^{\text{safe}})} \geq \nu_{\min} > 0.$$

- (bk2) (**Graceful-degradation certificate**) if  $\nu_t^{\text{core}} < \nu_{\min}$ , a nonempty parking set  $\mathcal{X}_t^{\text{park}} \subseteq \mathcal{X}_t^{\text{safe}}$ , horizon  $H_{\text{park}}$ , and bounded-loss contract

$$\sup_{s \in [t, t+H_{\text{park}}]} \text{Excursion}_s \leq E_{\text{park}, t}, \quad \inf_{s \in [t, t+H_{\text{park}}]} \text{Service}_s \geq S_{\min, t}.$$

- (bk3) (**Acceptability envelope contract**) the parking certificate must additionally publish an observable harm vector  $y_s^{\text{harm}} \in \mathbb{R}^{d_h}$ , coordinatewise caps  $h_{i,t}^{\max}$ , and response latency  $L_{\text{acc}}$  such that

$$\sup_{s \in [t, t+H_{\text{park}}]} y_{i,s}^{\text{harm}} \leq h_{i,t}^{\max} \quad (\forall i), \quad \text{and} \quad \mathbb{P}(T_{\text{acc}} > L_{\text{acc}}) \leq \alpha_{\text{acc}},$$

where  $T_{\text{acc}}$  is time-to-mitigation after an acceptability sentinel alarm. If any cap is forecast to be violated, the protocol must escalate to stricter backup mode within  $L_{\text{acc}}$  (graceful  $\rightarrow$  shield-only  $\rightarrow$  emergency-stop if required).

The quoted one-step complexity  $O(d_x^2)$  is valid only for replay-declared shield family  $\mathfrak{F}_{\text{sh}}^{\text{quad}}$  (affine/quadratic projection with sparse Jacobian pattern hash); otherwise a certified bound  $c_{\text{sh}}(d_x)$  must be published and used in runtime budgets. The feasible set contains every forced non-probing action used by  $\Pi_{\text{forced}}$ , i.e., if  $r$  is admissible for forced single-channel commitment then  $(r, 0) \in \mathcal{U}_t^{\text{probe}}(h_t)$ .

The maps  $(r, p) \mapsto \text{Cost}(p)$ ,  $(r, p) \mapsto \text{WidthAfter}(p)$ , and  $(r, p) \mapsto h_t^{\text{alias}}(h_t, p)$  are Borel measurable, bounded below, and lower-semicontinuous on  $\mathcal{U}_t^{\text{probe}}(h_t)$ , with normalization  $\text{Cost}(0) = 0$ ,  $\text{WidthAfter}(0)$  equal to the no-probe posterior width, and  $h_t^{\text{alias}}(h_t, 0)$  equal to the published baseline alias-hazard bound.

An endogenous hazard dual variable  $\lambda_{h,t} \geq 0$  is replay-updated by

$$\lambda_{h,t+1} = \Pi_{[0, \lambda_{\max, t}]} \left( \lambda_{h,t} + \eta_{\lambda, t} \text{sat}_{\delta_h}(\hat{h}_t - \bar{h}_t) \right),$$

with dead-zone  $\text{sat}_{\delta_h}(x) := \text{sign}(x) \max\{|x| - \delta_h, 0\}$ , optional dwell-time update period  $K_\lambda$ , and all hyperparameters  $(\eta_{\lambda, t}, \bar{h}_t, \lambda_{\max, t}, \delta_h, K_\lambda)$  supplied only through the public exogenous governance channel  $g_t^{\text{exo}}$ . No convexity/optimality convergence claim is required: this recursion is used as auditable constraint-tracking, not as proof of global optimality in nonconvex dynamics.

The backup stack must include replay-verifiable safety certificates against a broad core envelope  $\mathfrak{S}_t^{\text{core}}$ , and optionally a narrower performance envelope  $\mathfrak{S}_t^{\text{perf}} \subseteq \mathfrak{S}_t^{\text{core}}$  for  $\kappa_t^{\text{bk}}$ . If performance certificate is unavailable, shield-only execution remains mandatory. If nontrivial core certification is unavailable, graceful-degradation mode with the parking certificate and acceptability envelope is mandatory. Only if both certificates are invalid (or acceptability caps are infeasible) does execution move to explicit emergency-stop with published finite-time shutdown plan and violation budget. All backup certificates are independent of online global optimization over  $\Theta_t^{\text{op}}$ , so fail-closed execution is active stabilization without circular dependence on diagnostic completion.

**Assumption 5.21** (Finite-precision policy compiler and implementation envelope). There exists a deterministic replay-published compiler

$$\mathfrak{C}_t : (h_t, r, \text{policy state, numeric profile}) \mapsto a_t^{\text{fp}} \in \mathcal{A}_{\text{NM}}(h_t, r)$$

with fixed tie-breaking, arithmetic mode, and seed rules, inducing a finite-precision policy class  $\Pi_{\text{NM}}^{\text{fp}} \subseteq \Pi_{\text{NM}}$ . For each  $t$ , a public implementation envelope  $\varepsilon_{\text{impl},t} \geq 0$  is certified such that for every channel  $j$ ,

$$\left| \sup_{\pi \in \Pi_{\text{NM}}} S_{j,t}^\theta(H_t; \pi) - \sup_{\pi \in \Pi_{\text{NM}}^{\text{fp}}} S_{j,t}^\theta(H_t; \pi) \right| \leq \varepsilon_{\text{impl},t},$$

$$\left| \inf_{\pi \in \Pi_{\text{NM}}} S_{j,t}^\theta(H_t; \pi) - \inf_{\pi \in \Pi_{\text{NM}}^{\text{fp}}} S_{j,t}^\theta(H_t; \pi) \right| \leq \varepsilon_{\text{impl},t},$$

uniformly for  $\theta \in \Theta_t^{\text{op}}(\alpha_{\text{out}})$ . The certificate is deterministic and replay-verifiable from published branch-and-bound/mesh artifacts and solver metadata. Define  $b_t^{\text{impl}} := \varepsilon_{\text{impl},t}$ .

## 6 Full proofs: robust dynamic programming and measurability

### 6.1 Bellman operator

For finite horizon  $H$ , set terminal value  $V_{t+H,0}^\theta(\cdot; r) \equiv 0$ . For  $s = t + H - 1, \dots, t$ , define:

$$(\mathcal{T}_s^{\theta,r} v)(h) := \sup_{a \in \mathcal{A}_{\text{NM}}(h,r)} \inf_{K \in \mathcal{U}_s^\theta(h,a)} \int \left( r_s(h, a, h') + \gamma v(h') \right) \mathbb{P}(dh'|h, a, K).$$

**Lemma 6.1** (Stagewise robust payoff well-definedness). *Under Assumptions 5.1–5.3, for bounded measurable  $v$ , define*

$$\Phi(h, a, K; v) := \int \left( r_s(h, a, h') + \gamma v(h') \right) \mathbb{P}(dh'|h, a, K).$$

*Then, for each  $(h, a)$ , the inner value*

$$\Psi(h, a) := \inf_{K \in \mathcal{U}_s^\theta(h,a)} \Phi(h, a, K; v)$$

*is finite and lower semianalytic in  $(h, a)$ . Equivalently,  $-\Psi$  is upper semianalytic. Moreover, for every  $\varepsilon > 0$ , there exists a universally measurable  $\varepsilon$ -inner selector  $K_\varepsilon(h, a) \in \mathcal{U}_s^\theta(h, a)$  such that*

$$\Phi(h, a, K_\varepsilon(h, a); v) \leq \Psi(h, a) + \varepsilon.$$

*If, in addition,  $(h, a, K) \mapsto \Phi(h, a, K; v)$  is continuous in  $K$  and  $\mathcal{U}_s^\theta(h, a)$  is compact-valued with closed graph (Berge conditions), then the inner minimum is attained.*

*Proof.* Bounded integrand and nonempty compact-valued correspondence imply finite infima. For jointly measurable  $\Phi$ , inf-projection over an analytic graph yields lower semianalyticity of  $\Psi$ ; thus  $-\Psi$  is upper semianalytic. Jankov–von Neumann type measurable selection on analytic graphs yields universally measurable  $\varepsilon$ -inner selectors. Under Berge-type continuity and compactness, minimum attainment follows from Weierstrass.  $\square$

**Proposition 6.2** (Optional minimax interchange under action-independent kernels). *Assume*

Assumption 5.19 and the specialization

$$\mathcal{U}_s^\theta(h, a) = \mathcal{U}_s^\theta(h) \quad \forall a \in \mathcal{A}_{\text{NM}}(h, r).$$

Then for each stage  $s$  and history  $h$ ,

$$\sup_{a \in \mathcal{A}_{\text{NM}}(h, r)} \inf_{K \in \mathcal{U}_s^\theta(h)} \Phi(h, a, K; v) = \inf_{K \in \mathcal{U}_s^\theta(h)} \sup_{a \in \mathcal{A}_{\text{NM}}(h, r)} \Phi(h, a, K; v).$$

This equality is optional and is not required for Theorem 6.5; robust Bellman recursion uses the max-inf form directly.

*Proof.* Under Assumption 5.19,  $\mathcal{A}_{\text{NM}}(h, r)$  and  $\mathcal{U}_s^\theta(h)$  are compact convex sets,  $a \mapsto \Phi(h, a, K; v)$  is upper-semicontinuous and quasi-concave for each fixed  $K$ , and  $K \mapsto \Phi(h, a, K; v)$  is affine (hence both convex and concave, and continuous) for each fixed  $a$ . Therefore Sion's minimax theorem applies and yields the stated equality [10].  $\square$

**Lemma 6.3** (Measurable  $\varepsilon$ -selector for outer maximization). *Under Assumptions 5.1, 5.2, for bounded measurable  $v$  and each  $\varepsilon > 0$ , there exists a universally measurable selector  $a_\varepsilon(h) \in \mathcal{A}_{\text{NM}}(h, r)$  such that*

$$\Psi(h, a_\varepsilon(h)) \geq \sup_{a \in \mathcal{A}_{\text{NM}}(h, r)} \Psi(h, a) - \varepsilon,$$

where  $\Psi$  is the inner-value map from Lemma 6.1. If additional upper-semicontinuity/compactness conditions for exact maximization hold, an exact maximizer exists.

*Proof.* The feasible correspondence  $h \mapsto \mathcal{A}_{\text{NM}}(h, r)$  is measurable with nonempty compact values. By Lemma 6.1,  $\Psi(h, a)$  is lower semianalytic; equivalently  $-\Psi(h, a)$  is upper semianalytic. Applying approximate measurable minimization to  $-\Psi$  on analytic graphs yields a universally measurable  $\varepsilon$ -maximizer for  $\Psi$ . Exact attainment requires extra regularity, stated separately.  $\square$

**Lemma 6.4** (Backward recursion preserves bounded measurability). *If  $v$  is bounded measurable, then  $\mathcal{T}_s^{\theta, r} v$  is bounded measurable.*

*Proof.* Boundedness follows from  $|r_s| \leq B$  and  $\gamma < 1$ . Measurability follows from measurable selectors and measurable inf-projection.  $\square$

**Theorem 6.5** (Robust value well-posedness). *Under Assumptions 5.1–5.3, for finite horizon  $H$ :*

- (i)  $V_{t, H}^\theta(h_t; r)$  in (5) is finite for all  $(h_t, r, \theta)$ ;
- (ii) Bellman recursion is dynamically consistent (rectangular uncertainty);
- (iii) measurable  $\varepsilon$ -maximizing selectors exist at each stage for every  $\varepsilon > 0$ ; this is an existential statement, while finite-precision replayability is handled by Assumption 5.21 and Theorem 6.6.

*Proof.* By Lemma 6.4, if terminal value is bounded measurable (zero), then one-step backward application preserves bounded measurability. Induct backward  $H$  steps to obtain bounded measurable value functions. Lemma 6.1 gives stagewise  $\varepsilon$ -inner selectors for each action; with the optional extra regularity in Lemma 6.1, exact attainment is recovered. Rectangularity ensures that local worst-case choices compose over time (time consistency) in robust MDPs [3, 4, 6]. Lemma 6.3 yields measurable  $\varepsilon$ -maximizing selectors at each stage (arbitrarily tight). Therefore value is well-defined, finite, and operationally approximable to arbitrary precision via measurable policies.  $\square$

**Theorem 6.6** (Constructive bridge from measurable selectors to replayable implementation). *Assume 5.1–5.3, 5.9, 5.10, and 5.21. For each  $t, j$ , let  $S_{j,t}^{\text{opt}}$  denote the ideal model-class endpoint under  $\Pi_{\text{NM}}$ , and  $S_{j,t}^{\text{fp}}$  the implemented endpoint under  $\Pi_{\text{NM}}^{\text{fp}}$  with deterministic certificates. Then*

$$|S_{j,t}^{\text{opt}} - S_{j,t}^{\text{fp}}| \leq \varepsilon_{\text{impl},t} + \varepsilon_{\text{opt},j,t}.$$

*Hence interval inflation remains replay-sound once  $b_t^{\text{impl}} = \varepsilon_{\text{impl},t}$  is included.*

*Proof.* Assumption 5.21 bounds approximation loss from  $\Pi_{\text{NM}}$  to  $\Pi_{\text{NM}}^{\text{fp}}$ . Assumption 5.9 bounds optimization suboptimality within  $\Pi_{\text{NM}}^{\text{fp}}$ . Triangle inequality gives the claim.  $\square$

**Theorem 6.7** (Universal measurability of score interval endpoints). *Under Assumptions 5.2, 5.5, 5.10, the maps*

$$h \mapsto \underline{S}_{j,t}(h), \quad h \mapsto \overline{S}_{j,t}(h)$$

*are universally measurable for each  $j \in \mathcal{J}$ . If, in addition, the corresponding argmin/argmax correspondences are Borel measurable, then the endpoints are Borel measurable.*

*Proof.* By Assumption 5.10, the finite-difference stencil values of  $G_{t,H}^\theta(h; \cdot)$  are jointly measurable in  $(h, \theta)$ ; therefore  $S_j^\theta(h)$  in (7) is jointly measurable. Because  $h \mapsto \Theta_t^{\text{ev}}(h)$  has measurable graph and closed nonempty values (Assumption 5.5), measurable inf/sup projection theorems apply [15, 17]:

$$\underline{S}_{j,t}(h) = \inf_{\theta \in \Theta_t^{\text{ev}}(h)} S_j^\theta(h), \quad \overline{S}_{j,t}(h) = \sup_{\theta \in \Theta_t^{\text{ev}}(h)} S_j^\theta(h).$$

Hence both are universally measurable; the Borel claim follows under the additional selector regularity condition stated above.  $\square$

**Proposition 6.8** (Constructive replayable  $\varepsilon$ -policy extraction). *Assume that for each  $t, h$ , the admissible action set  $\mathcal{A}_t(h) \subset \mathbb{R}^{d_a}$  is compact and a deterministic  $\delta_{a,t}$ -net  $\mathcal{A}_t^{(\delta_{a,t})}(h)$  is published in replay metadata. Suppose  $Q_t^\theta(h, a)$  is jointly Lipschitz on  $\Theta_t^{\text{ev}}(h) \times \mathcal{A}_t(h)$ , with action modulus  $L_{a,t}$ , and endpoint optimization over  $\Theta_t^{\text{ev}}(h)$  is solved with certificate  $\varepsilon_{\text{opt},t}$ . Let  $\varepsilon_{\text{fp},t}$  bound floating-point/profile replay discrepancy. Define*

$$\pi_t^{\text{rep}}(h) \in \arg \max_{a \in \mathcal{A}_t^{(\delta_{a,t})}(h)} \inf_{\theta \in \Theta_t^{\text{ev}}(h)} Q_t^\theta(h, a).$$

*Then*

$$\sup_h \left( \sup_{a \in \mathcal{A}_t(h)} \inf_{\theta \in \Theta_t^{\text{ev}}(h)} Q_t^\theta(h, a) - \inf_{\theta \in \Theta_t^{\text{ev}}(h)} Q_t^\theta(h, \pi_t^{\text{rep}}(h)) \right) \leq L_{a,t} \delta_{a,t} + \varepsilon_{\text{opt},t} + \varepsilon_{\text{fp},t}.$$

*Hence measurable-existence results can be operationalized by finite deterministic selectors with explicit auditable suboptimality.*

*Proof.* For any  $a^* \in \mathcal{A}_t(h)$ , choose a net point  $\tilde{a} \in \mathcal{A}_t^{(\delta_{a,t})}(h)$  with  $\|a^* - \tilde{a}\| \leq \delta_{a,t}$ . Lipschitz continuity yields  $\inf_{\theta} Q_t^\theta(h, \tilde{a}) \geq \inf_{\theta} Q_t^\theta(h, a^*) - L_{a,t} \delta_{a,t}$ . Maximizing over net points and adding deterministic optimization and replay envelopes gives the bound.  $\square$

## 7 Full proofs: identification limits and interaction structure

**Definition 7.1** (No-meta observational equivalence).  $\theta \sim_{\Pi_{\text{NM}}} \theta'$  iff for every  $t$ , every measurable  $B \subseteq \mathcal{H}_t$ , and every  $\pi \in \Pi_{\text{NM}}$ ,

$$\mathbb{P}_\theta^\pi(H_t \in B) = \mathbb{P}_{\theta'}^\pi(H_t \in B).$$



**Theorem 7.2** (Non-identifiability of latent point attribution). *If  $\theta \sim_{\Pi_{\text{NM}}} \theta'$ , then for any measurable estimator  $\hat{Q}_t : \mathcal{H}_t \rightarrow \mathbb{R}^m$ ,*

$$\mathcal{L}_\theta^\pi(\hat{Q}_t(H_t)) = \mathcal{L}_{\theta'}^\pi(\hat{Q}_t(H_t)) \quad \forall \pi \in \Pi_{\text{NM}}.$$

*Hence latent point attribution cannot be identified from public history alone [7, 8].*

*Proof.* Fix  $t, \pi$ , and measurable  $A \subseteq \mathbb{R}^m$ . Then

$$\mathbb{P}_\theta^\pi(\hat{Q}_t(H_t) \in A) = \mathbb{P}_\theta^\pi(H_t \in \hat{Q}_t^{-1}(A)).$$

Since  $\hat{Q}_t^{-1}(A) \in \mathcal{B}(\mathcal{H}_t)$  and  $\theta \sim_{\Pi_{\text{NM}}} \theta'$ ,

$$\mathbb{P}_\theta^\pi(H_t \in \hat{Q}_t^{-1}(A)) = \mathbb{P}_{\theta'}^\pi(H_t \in \hat{Q}_t^{-1}(A)) = \mathbb{P}_{\theta'}^\pi(\hat{Q}_t(H_t) \in A).$$

Thus induced laws coincide for all measurable  $A$ , proving equality in distribution.  $\square$

**Proposition 7.3** (Sufficient condition for nontrivial observational aliases). *Let admissible policies depend only on public history. Suppose  $\theta \neq \theta'$ , and:*

- (a) *initial public-history law agrees:  $\mathcal{L}_\theta(H_0) = \mathcal{L}_{\theta'}(H_0)$ ;*
- (b) *for every  $t \geq 0$ , every public history  $h \in \mathcal{H}_t$ , every admissible action  $a \in \mathcal{A}_{\text{NM}}(h, \bar{r})$ , and every measurable next-history event  $B \subseteq \mathcal{H}_{t+1}$ ,*

$$\mathbb{P}_\theta(H_{t+1} \in B \mid H_t = h, A_t = a) = \mathbb{P}_{\theta'}(H_{t+1} \in B \mid H_t = h, A_t = a).$$

*Then  $\theta \sim_{\Pi_{\text{NM}}} \theta'$ , hence point attribution is impossible on that pair.*

*Proof.* Condition (b) gives equality of one-step public kernels conditioned on the full public history/action pair. Together with (a), induction over  $t$  and Ionescu–Tulcea construction imply equality of the full public-history law under every  $\pi \in \Pi_{\text{NM}}$ , hence  $\theta \sim_{\Pi_{\text{NM}}} \theta'$  [2]. Apply Theorem 7.2.  $\square$

**Proposition 7.4** (Möbius decomposition on  $2^{\{I, M, P\}}$ ). *Define  $F^\theta(S) := G_{t, H}^\theta(h_t; \delta \sum_{j \in S} e_j)$ ,  $S \subseteq \mathcal{J}$ . Then*

$$F^\theta(\mathcal{J}) - F^\theta(\emptyset) = \sum_{j \in \mathcal{J}} \Delta_j^\theta + \sum_{\{j, k\} \subset \mathcal{J}} \Delta_{jk}^\theta + \Delta_{IMP}^\theta,$$

*where  $\Delta_j^\theta := F^\theta(\{j\}) - F^\theta(\emptyset)$ .*

*Proof.* Use inclusion–exclusion (equivalently, Möbius inversion on  $2^{\mathcal{J}}$ ):

$$F^\theta(\mathcal{J}) - F^\theta(\emptyset) = \sum_{\emptyset \neq T \subseteq \mathcal{J}} \delta_T^\theta, \quad \delta_T^\theta := \sum_{U \subseteq T} (-1)^{|T| - |U|} F^\theta(U).$$

For  $|\mathcal{J}| = 3$ ,  $\delta_{\{j\}}^\theta = \Delta_j^\theta$ ,  $\delta_{\{j, k\}}^\theta = \Delta_{jk}^\theta$ , and  $\delta_{\{I, M, P\}}^\theta = \Delta_{IMP}^\theta$ , yielding the displayed decomposition.  $\square$

**Theorem 7.5** (Coherent dominance certificate). *If for some  $j^* \in \mathcal{J}$ ,*

$$\underline{S}_{j^*, t} \geq \max_{k \neq j^*} \bar{S}_{k, t} + \varepsilon,$$

*then  $j^*$  is uniquely  $\varepsilon$ -dominant for all  $\theta \in \Theta_t^{\text{ev}}(h_t)$ .*

*Proof.* Take arbitrary  $\theta \in \Theta_t^{\text{ev}}(h_t)$ . By definition of endpoints:

$$S_{j^*}^\theta \geq \underline{S}_{j^*,t}, \quad S_k^\theta \leq \overline{S}_{k,t} \quad \forall k \neq j^*.$$

Hence

$$S_{j^*}^\theta - S_k^\theta \geq \underline{S}_{j^*,t} - \overline{S}_{k,t} \geq \varepsilon, \quad \forall k \neq j^*.$$

So  $j^*$  strictly dominates all competitors at margin  $\varepsilon$  for every model in the ambiguity set. Uniqueness follows.  $\square$

**Corollary 7.6** (Abstention correctness). *If  $|\mathcal{B}_t^\varepsilon| \neq 1$ , unique bottleneck declaration is not certifiable under model ambiguity.*

*Proof.* Direct contrapositive of Theorem 7.5.  $\square$

## 8 Time-consistent outer/inner set recursion

### 8.1 Model-indexed e-processes

By Assumption 5.11, for each  $\theta \in \Theta_0$ ,  $E_t(\theta)$  is a nonnegative supermartingale under  $P_\theta$  with  $E_0(\theta) = 1$ .

### 8.2 Outer recursion

For confidence level  $1 - \alpha_{\text{out}}$ , define the exact set

$$\Theta_t^{\text{out}}(\alpha_{\text{out}}) := \left\{ \theta \in \Theta_0 : \sup_{s \leq t} E_s(\theta) < \frac{1}{\alpha_{\text{out}}} \right\}.$$

Operationally (for replay with floating-point arithmetic), use

$$\Theta_t^{\text{out,num}}(\alpha_{\text{out}}) := \left\{ \theta \in \Theta_0 : \sup_{s \leq t} E_s^{\text{impl}}(\theta) \leq \frac{1}{\alpha_{\text{out}}} + \varepsilon_{E,t}^{\text{num}} + \tau_{\text{num}} \right\},$$

where  $(\varepsilon_{E,t}^{\text{num}}, \tau_{\text{num}})$  are published and  $E_s^{\text{impl}}$  follows Assumption 5.17. Equivalent exact recursion:

$$\Theta_{t+1}^{\text{out}}(\alpha_{\text{out}}) = \Theta_t^{\text{out}}(\alpha_{\text{out}}) \cap \left\{ \theta : E_{t+1}(\theta) < \frac{1}{\alpha_{\text{out}}} \right\}.$$

**Theorem 8.1** (Anytime-valid outer coverage). *Under Assumption 5.11,*

$$\mathbb{P}_{\theta^*}(\forall t \geq 0 : \theta^* \in \Theta_t^{\text{out}}(\alpha_{\text{out}})) \geq 1 - \alpha_{\text{out}}.$$

*Proof.* For fixed  $\theta^*$ ,  $E_t(\theta^*)$  is a nonnegative supermartingale with  $E_0 = 1$ ; the theorem applies to the exact set  $\Theta_t^{\text{out}}$ , while implementation uses  $\Theta_t^{\text{out,num}}$  with published tolerance. By Ville's inequality [18, 21],

$$\mathbb{P}_{\theta^*} \left( \sup_{t \geq 0} E_t(\theta^*) \geq \frac{1}{\alpha_{\text{out}}} \right) \leq \alpha_{\text{out}}.$$

Complementing gives the claim.  $\square$

**Proposition 8.2** (Numerical outer-set containment). *Under Assumption 5.17, for every  $t \geq 0$ ,*

$$\Theta_t^{\text{out}}(\alpha_{\text{out}}) \subseteq \Theta_t^{\text{out,num}}(\alpha_{\text{out}}).$$

Hence

$$\mathbb{P}_{\theta^*}(\forall t \geq 0 : \theta^* \in \Theta_t^{\text{out,num}}(\alpha_{\text{out}})) \geq 1 - \alpha_{\text{out}}.$$

*Proof.* Take  $\theta \in \Theta_t^{\text{out}}(\alpha_{\text{out}})$ . Then  $\sup_{s \leq t} E_s(\theta) < 1/\alpha_{\text{out}}$ . By Assumption 5.17,

$$\sup_{s \leq t} E_s^{\text{impl}}(\theta) \leq \sup_{s \leq t} E_s(\theta) + \varepsilon_{E,t}^{\text{num}} < \frac{1}{\alpha_{\text{out}}} + \varepsilon_{E,t}^{\text{num}} \leq \frac{1}{\alpha_{\text{out}}} + \varepsilon_{E,t}^{\text{num}} + \tau_{\text{num}},$$

so  $\theta \in \Theta_t^{\text{out,num}}(\alpha_{\text{out}})$ . The probability claim follows from Theorem 8.1.  $\square$

**Remark 8.3** (Operational set used in guarantees). By Definition 5.12, intervals and decisions are computed on  $\Theta_t^{\text{op}}(\alpha_{\text{out}}) = \Theta_t^{\text{ev}}(h_t) \cap \Theta_t^{\text{out,num}}(\alpha_{\text{out}})$ . This ties evidence-consistency and anytime-valid filtering in a single set-valued object.

**Proposition 8.4** (Relaxation-ladder certification for raw-set endpoints). *Fix  $t, j$ . Assume deterministic sets  $\Theta_t^{\text{in,raw}}(h_t) \subseteq \Theta_t^{\text{raw}}(h_t) \subseteq \Theta_t^{\text{out,raw}}(h_t)$ . Define*

$$\begin{aligned} \underline{U}_{j,t}^{\text{raw}} &:= \sup_{\theta \in \Theta_t^{\text{in,raw}}(h_t)} S_j^\theta, & \overline{U}_{j,t}^{\text{raw}} &:= \sup_{\theta \in \Theta_t^{\text{out,raw}}(h_t)} S_j^\theta, \\ \underline{L}_{j,t}^{\text{raw}} &:= \inf_{\theta \in \Theta_t^{\text{in,raw}}(h_t)} S_j^\theta, & \overline{L}_{j,t}^{\text{raw}} &:= \inf_{\theta \in \Theta_t^{\text{out,raw}}(h_t)} S_j^\theta. \end{aligned}$$

*Then the bracket relations in Assumption 5.18 hold without solving global optimization on  $\Theta_t^{\text{raw}}(h_t)$ . If  $\Theta_t^{\text{out,raw}}(h_t)$  is represented by convex/SDP constraints,  $\overline{U}_{j,t}^{\text{raw}}$  and  $\underline{L}_{j,t}^{\text{raw}}$  admit deterministic dual certificates.*

*Proof.* Set inclusion immediately yields

$$\begin{aligned} \sup_{\theta \in \Theta_t^{\text{in,raw}}(h_t)} S_j &\leq \sup_{\theta \in \Theta_t^{\text{raw}}(h_t)} S_j^\theta \leq \sup_{\theta \in \Theta_t^{\text{out,raw}}(h_t)} S_j, \\ \inf_{\theta \in \Theta_t^{\text{out,raw}}(h_t)} S_j &\leq \inf_{\theta \in \Theta_t^{\text{raw}}(h_t)} S_j^\theta \leq \inf_{\theta \in \Theta_t^{\text{in,raw}}(h_t)} S_j. \end{aligned}$$

Thus the displayed definitions give valid raw-set endpoint brackets. When outer relaxation is convex, dual feasibility/optimality residuals provide replay-verifiable upper/lower endpoint certificates.  $\square$

**Proposition 8.5** (Operational rectangularization preserves dynamic consistency with endpoint-certified distortion). *Assume Assumptions 5.3 and 5.18. Then robust Bellman recursion is dynamically consistent on the operational set*

$$\Theta_t^{\text{ev}}(h_t) = \mathfrak{R}(\Theta_t^{\text{raw}}(h_t)).$$

*For each channel  $j$ , true endpoint distortions satisfy*

$$\begin{aligned} 0 &\leq \sup_{\theta \in \Theta_t^{\text{ev}}(h_t)} S_j^\theta(h_t) - \sup_{\theta \in \Theta_t^{\text{raw}}(h_t)} S_j^\theta(h_t) \leq \hat{\delta}_{j,t}^{\text{rect}}, \\ 0 &\leq \inf_{\theta \in \Theta_t^{\text{raw}}(h_t)} S_j^\theta(h_t) - \inf_{\theta \in \Theta_t^{\text{ev}}(h_t)} S_j^\theta(h_t) \leq \hat{\delta}_{j,t}^{\text{rect}}. \end{aligned}$$

*Hence interval inflation by  $b_t^{\text{rect,pub}} = \max_j \hat{\delta}_{j,t}^{\text{rect}}$  transfers raw-set score validity to the rectangularized operational recursion.*

*Proof.* Time consistency follows because recursion is executed on the rectangular set  $\Theta_t^{\text{ev}}(h_t)$ . Set

inclusion  $\Theta_t^{\text{raw}}(h_t) \subseteq \Theta_t^{\text{ev}}(h_t)$  implies

$$\sup_{\Theta_t^{\text{ev}}(h_t)} S_j \geq \sup_{\Theta_t^{\text{raw}}(h_t)} S_j, \quad \inf_{\Theta_t^{\text{ev}}(h_t)} S_j \leq \inf_{\Theta_t^{\text{raw}}(h_t)} S_j.$$

Using certified endpoint brackets in Assumption 5.18,

$$\begin{aligned} \sup_{\Theta_t^{\text{ev}}(h_t)} S_j - \sup_{\Theta_t^{\text{raw}}(h_t)} S_j &\leq \overline{U}_{j,t}^{\text{rect}} - \underline{U}_{j,t}^{\text{raw}} \leq \hat{\delta}_{j,t}^{\text{rect}}, \\ \inf_{\Theta_t^{\text{raw}}(h_t)} S_j - \inf_{\Theta_t^{\text{ev}}(h_t)} S_j &\leq \overline{L}_{j,t}^{\text{raw}} - \underline{L}_{j,t}^{\text{rect}} \leq \hat{\delta}_{j,t}^{\text{rect}}. \end{aligned}$$

Taking channelwise maxima gives the published bound  $b_t^{\text{rect, pub}}$ . □

### 8.3 Inner recursion (time-consistent definition)

Choose radius schedule  $r_t \downarrow 0$ , metric  $d_\Theta$ , and define naive erosion

$$\tilde{\Theta}_t^{\text{in}}(\alpha_{\text{out}}, r_t) := \left\{ \theta \in \Theta_t^{\text{out}}(\alpha_{\text{out}}) : \mathbb{B}_{d_\Theta}(\theta, r_t) \subseteq \Theta_t^{\text{out}}(\alpha_{\text{out}}) \right\}.$$

Naive erosion alone need not be temporally monotone. We therefore define the operational inner core recursively:

$$\begin{aligned} \Theta_0^{\text{in}}(\alpha_{\text{out}}, r_0) &:= \tilde{\Theta}_0^{\text{in}}(\alpha_{\text{out}}, r_0), \\ \Theta_{t+1}^{\text{in}}(\alpha_{\text{out}}, r_{t+1}) &:= \left\{ \theta \in \Theta_t^{\text{in}}(\alpha_{\text{out}}, r_t) \cap \Theta_{t+1}^{\text{out}}(\alpha_{\text{out}}) : \mathbb{B}_{d_\Theta}(\theta, r_{t+1}) \subseteq \Theta_{t+1}^{\text{out}}(\alpha_{\text{out}}) \right\}. \end{aligned}$$

**Theorem 8.6** (Inner/outer consistency and monotonicity). *For all  $t$ ,*

$$\Theta_t^{\text{in}}(\alpha_{\text{out}}, r_t) \subseteq \Theta_t^{\text{out}}(\alpha_{\text{out}}), \quad \Theta_{t+1}^{\text{in}}(\alpha_{\text{out}}, r_{t+1}) \subseteq \Theta_t^{\text{in}}(\alpha_{\text{out}}, r_t).$$

Hence  $(\Theta_t^{\text{in}})_t$  is nested and time-consistent.

*Proof.* Outer inclusion is immediate by construction at  $t = 0$ , and preserved at  $t + 1$  by explicit intersection with  $\Theta_{t+1}^{\text{out}}(\alpha_{\text{out}})$ . Monotonicity is immediate from recursive intersection with  $\Theta_t^{\text{in}}(\alpha_{\text{out}}, r_t)$ . No additional geometric argument is required. □

*Remark 8.7.* The recursive definition is the one used for guarantees.  $\tilde{\Theta}_t^{\text{in}}$  is diagnostic only.

## 9 Anytime-valid multi-time score guarantees

Define auditable intervals from operational sets  $\Theta_t^{\text{op}}(\alpha_{\text{out}})$  with contamination/dependence cushions:

$$\begin{aligned} c_{S, \text{cont}} &:= 8 \left( 1 + 2w_2 + \frac{4}{3}w_3 \right), \quad b_{t,H}^{\text{cont}} := \frac{c_{S, \text{cont}} B}{(1 - \gamma) \delta} \bar{\eta}_t^{\text{pub}}, \\ b_t^{\text{dep}} &:= b_t^{\text{dep, pub}}, \end{aligned}$$

where  $\bar{\eta}_t^{\text{pub}} := \max_{0 \leq s \leq t} \bar{\eta}_s^{\text{pub}}$ , and all envelope quantities are public and replay-verifiable (Assumption 5.16). Define additionally

$$b_t^{\text{int}} := b_t^{\text{int, pub}}, \quad b_t^{\text{link}} := b_t^{\text{link, pub}}, \quad b_t^{\text{rect}} := b_t^{\text{rect, pub}}, \quad b_t^{\text{impl}} := \varepsilon_{\text{impl}, t},$$

and the aggregate auditable inflation

$$b_t^\Sigma := b_t^{\text{impl}} + b_{t,H}^{\text{cont}} + b_t^{\text{dep}} + b_t^{\text{int}} + b_t^{\text{link}} + b_t^{\text{rect}}.$$

The contamination coefficient  $c_{S,\text{cont}}$  is a conservative score-level constant induced by (7) (Lemma A.2); with weights  $(w_2, w_3)$ , the certified coefficient is

$$c_{S,\text{cont}}(w_2, w_3) := 8 \left( 1 + 2w_2 + \frac{4}{3}w_3 \right),$$

and the default  $w_2 = w_3 = 1$  gives  $c_{S,\text{cont}} = 104/3$ . For  $j \in \mathcal{J}$ , define

$$(\hat{L}_{j,t}, \hat{U}_{j,t}) = \begin{cases} \left( \inf_{\theta \in \Theta_t^{\text{op}}(\alpha_{\text{out}})} S_j^\theta - \varepsilon_{\text{opt},j,t} - b_t^\Sigma, \sup_{\theta \in \Theta_t^{\text{op}}(\alpha_{\text{out}})} S_j^\theta + \varepsilon_{\text{opt},j,t} + b_t^\Sigma \right), & \Theta_t^{\text{op}}(\alpha_{\text{out}}) \neq \emptyset, \\ (-\infty, +\infty), & \Theta_t^{\text{op}}(\alpha_{\text{out}}) = \emptyset. \end{cases}$$

$$\mathcal{I}_{j,t} := [\hat{L}_{j,t}, \hat{U}_{j,t}].$$

When  $\Theta_t^{\text{op}}(\alpha_{\text{out}}) = \emptyset$ , the protocol is fail-closed: declare  $\hat{\mathcal{B}}_t^\varepsilon = \emptyset$ , abstain from unique attribution, and enforce  $A_t \in \mathcal{A}_{\text{safe}}$ .

Define declared set

$$\hat{\mathcal{B}}_t^\varepsilon := \left\{ j \in \mathcal{J} : \Theta_t^{\text{op}}(\alpha_{\text{out}}) \neq \emptyset, \hat{L}_{j,t} \geq \max_{k \neq j} \hat{U}_{k,t} + \varepsilon + \tau_{\text{num}} \right\}.$$

**Proposition 9.1** (Auditable sufficient protocol for Assumption 5.14). *Suppose the evidence-consistent set is implemented as*

$$\Theta_t^{\text{ev}}(H_t) = \left\{ \theta \in \Theta_0 : \sup_{s \leq t} E_s^{\text{ev}}(\theta) < \frac{1}{\alpha_{\text{ev}}} \right\},$$

where for each fixed  $\theta$ ,  $E_s^{\text{ev}}(\theta)$  is a nonnegative  $(\mathcal{F}_s)$ -supermartingale under  $P_\theta$  with  $E_0^{\text{ev}}(\theta) = 1$ , and  $\theta \mapsto E_s^{\text{ev}}(\theta)$  is Borel measurable. Then Assumption 5.14 holds.

*Proof.* Apply Ville's inequality modelwise under  $P_{\theta^*}$ :

$$\mathbb{P}_{\theta^*} \left( \sup_{s \geq 0} E_s^{\text{ev}}(\theta^*) \geq \frac{1}{\alpha_{\text{ev}}} \right) \leq \alpha_{\text{ev}}.$$

Therefore  $\theta^*$  is retained for all  $t$  with probability at least  $1 - \alpha_{\text{ev}}$ .  $\square$

**Proposition 9.2** (Auditable sufficient protocol for Assumption 5.15 under published blocking certificate). *Suppose Assumption 5.15 holds and, in addition, the public replay bundle provides for each  $(j, t)$ :*

(d1) *cumulative residual process*

$$C_{j,t} := \sum_{s=1}^t R_{j,s}, \quad C_{j,0} := 0,$$

*and a blocked martingale surrogate  $M_{j,t}^{\text{blk}}$ ;*

(d2) *a deterministic coupling remainder bound  $c_{j,t}^{\text{mix}} \geq 0$  such that*

$$|C_{j,t} - M_{j,t}^{\text{blk}}| \leq c_{j,t}^{\text{mix}};$$

(d3) a predictable variance process  $V_{j,t}$  and range constant  $c_j$  for  $M_{j,t}^{\text{blk}}$ , together with a certified cumulative radius  $U_{j,t}(\alpha_{\text{dep}})$  satisfying

$$\mathbb{P}\left(\forall t \geq 0 : |M_{j,t}^{\text{blk}}| \leq U_{j,t}(\alpha_{\text{dep}})\right) \geq 1 - \frac{\alpha_{\text{dep}}}{|\mathcal{J}|}.$$

Define per-time residual radius

$$u_{j,0}(\alpha_{\text{dep}}) := 0, \quad u_{j,t}(\alpha_{\text{dep}}) := U_{j,t}(\alpha_{\text{dep}}) + U_{j,t-1}(\alpha_{\text{dep}}) + c_{j,t}^{\text{mix}} + c_{j,t-1}^{\text{mix}} \quad (t \geq 1).$$

If the published dependence cushion obeys

$$b_t^{\text{dep}} \geq \max_{j \in \mathcal{J}} u_{j,t}(\alpha_{\text{dep}}) \quad \forall t,$$

then

$$\mathbb{P}\left(\forall t \geq 0, \forall j \in \mathcal{J} : |S_j^{\theta^*}(H_t) - \tilde{S}_{j,t}^{\theta^*}| \leq b_t^{\text{dep}}\right) \geq 1 - \alpha_{\text{dep}}.$$

*Proof.* Fix  $j$ , and define event

$$\mathcal{E}_j := \{\forall t \geq 0 : |M_{j,t}^{\text{blk}}| \leq U_{j,t}(\alpha_{\text{dep}})\}.$$

By premise,  $\mathbb{P}(\mathcal{E}_j) \geq 1 - \alpha_{\text{dep}}/|\mathcal{J}|$ . On  $\mathcal{E}_j$ , for  $t \geq 1$ ,

$$\begin{aligned} |R_{j,t}| &= |C_{j,t} - C_{j,t-1}| \\ &\leq |M_{j,t}^{\text{blk}} - M_{j,t-1}^{\text{blk}}| + |C_{j,t} - M_{j,t}^{\text{blk}}| + |C_{j,t-1} - M_{j,t-1}^{\text{blk}}| \\ &\leq U_{j,t}(\alpha_{\text{dep}}) + U_{j,t-1}(\alpha_{\text{dep}}) + c_{j,t}^{\text{mix}} + c_{j,t-1}^{\text{mix}} = u_{j,t}(\alpha_{\text{dep}}). \end{aligned}$$

Hence  $|R_{j,t}| \leq b_t^{\text{dep}}$  for all  $t$  by the displayed domination condition. Union bound over  $j \in \mathcal{J}$  gives the simultaneous probability bound  $1 - \alpha_{\text{dep}}$ .  $\square$

**Proposition 9.3** (Observable-only calibration protocol implies auditable envelope validity). *Assume Assumption 5.16(b). If each monitor  $e$ -process  $M_t^{(q)}$  is valid under its channel null and  $\sum_q \alpha_q \leq \alpha_{\text{env}}$ , then with probability at least  $1 - \alpha_{\text{env}}$ , all replay-published monitor-side envelopes are simultaneously valid for all  $t$ . Moreover, if the link decomposition*

$$b_t^{\text{link, pub}} = \hat{b}_t^{\text{link}} + u_t^{\text{link}}(\alpha_{\text{link}}) + r_t^{\text{ow}},$$

uses the explicit open-world map

$$r_t^{\text{ow}} = \left[ \rho_{0,t}^{\text{exo}} + \rho_{1,t}^{\text{exo}} \log^+(M_t^{\text{mis}}) + \rho_{2,t}^{\text{exo}} \text{NVS}_t \right]_0^{r_{\text{max},t}^{\text{exo}}} + u_t^{\text{ow}}(\alpha_{\text{ow}}),$$

with exogenous tuple from public governance channel  $g_t^{\text{exo}}$ , then latent-side envelope validity is obtained with explicit residual accounting and no hidden constants. This proposition is intentionally one-way: monitor validity is a necessary evidence condition, not a proof of latent normality.

*Proof.* For each channel  $q$ , Ville's inequality gives  $\mathbb{P}(\sup_t M_t^{(q)} \geq 1/\alpha_q) \leq \alpha_q$ . Hence

$$\mathbb{P}\left(\bigcup_q \mathcal{A}_{\infty}^{(q)}\right) \leq \sum_q \alpha_q \leq \alpha_{\text{env}}.$$

On the complement event, all monitor-side channel envelopes hold for all  $t$ . The link decomposition then contributes an additive certified mismatch budget  $\hat{b}_t^{\text{link}} + u_t^{\text{link}} + r_t^{\text{ow}}$ . Because  $r_t^{\text{ow}}$  is computed from

an explicit monotone map with hash-linked exogenous tuple and replay-visible inputs ( $M_t^{\text{mis}}, \text{NVS}_t$ ), envelope closure is auditable rather than tautological.  $\square$

**Proposition 9.4** (Composability of implementation/optimization/contamination/dependence errors with interaction and proxy-link residuals). *Suppose for each  $j, t$ ,*

$$S_j^{\theta^*}(H_t) - \tilde{S}_{j,t}^{\theta^*} = \Delta_{j,t}^{\text{cont}} + \Delta_{j,t}^{\text{dep}} + \Xi_{j,t}^{\text{int}} + \Xi_{j,t}^{\text{link}},$$

*with  $|\Delta_{j,t}^{\text{cont}}| \leq b_{t,H}^{\text{cont}}$ ,  $|\Delta_{j,t}^{\text{dep}}| \leq b_t^{\text{dep}}$ ,  $|\Xi_{j,t}^{\text{int}}| \leq b_t^{\text{int}}$ , and  $|\Xi_{j,t}^{\text{link}}| \leq b_t^{\text{link}}$ , and optimization certificates satisfy Assumption 5.9, with implementation envelope  $b_t^{\text{impl}}$  from Assumption 5.21. Then interval inflation by  $b_t^{\text{impl}} + \varepsilon_{\text{opt},j,t} + b_{t,H}^{\text{cont}} + b_t^{\text{dep}} + b_t^{\text{int}} + b_t^{\text{link}}$  is sufficient for score containment.*

*Proof.* Apply triangle inequality:

$$|S_j^{\theta^*} - \tilde{S}_{j,t}^{\theta^*}| \leq |\Delta_{j,t}^{\text{cont}}| + |\Delta_{j,t}^{\text{dep}}| + |\Xi_{j,t}^{\text{int}}| + |\Xi_{j,t}^{\text{link}}| \leq b_{t,H}^{\text{cont}} + b_t^{\text{dep}} + b_t^{\text{int}} + b_t^{\text{link}}.$$

Combining with implementation inflation  $b_t^{\text{impl}}$  (Assumption 5.21) and endpoint optimization bracket inflation in Assumption 5.9 gives the stated certificate.  $\square$

**Proposition 9.5** (Per-time practical certificate). *Fix  $t$  and assume  $\Theta_t^{\text{op}}(\alpha_{\text{out}}) \neq \emptyset$ . If for some  $j^* \in \mathcal{J}$ ,*

$$\hat{L}_{j^*,t} \geq \max_{k \neq j^*} \hat{U}_{k,t} + \varepsilon + \tau_{\text{num}},$$

*then on the event  $\bigcap_{j \in \mathcal{J}} \{S_j^{\theta^*}(H_t) \in [\hat{L}_{j,t}, \hat{U}_{j,t}]\}$ ,  $j^*$  is the unique true  $\varepsilon$ -dominant channel at time  $t$ .*

*Proof.* The proof is immediate from interval containment and the displayed strict separation inequality.  $\square$

**Theorem 9.6** (Anytime-valid simultaneous score coverage with explicit error budget). *Let*

$$\alpha_{\text{tot}} := \alpha_{\text{out}} + \alpha_{\text{ev}} + \alpha_{\text{dep}} + \alpha_{\text{env}} + \alpha_{\text{mis}} < 1,$$

*where  $\alpha_{\text{env}} := 0$  in deterministic-by-construction envelope mode (Assumption 5.16). Under Assumptions 5.9, 5.11, 5.14–5.18, 5.21, with Propositions 9.2, 9.3, and 9.4, with probability at least  $1 - \alpha_{\text{tot}}$ ,*

$$\forall t \geq 0, \forall j \in \mathcal{J} : S_j^{\theta^*}(H_t) \in \mathcal{I}_{j,t}.$$

*Proof.* Define events

$$\mathcal{E}_{\text{out}} := \{\forall t \geq 0 : \theta^* \in \Theta_t^{\text{out}}(\alpha_{\text{out}})\},$$

$$\mathcal{E}_{\text{ev}} := \{\forall t \geq 0 : \theta^* \in \Theta_t^{\text{ev}}(H_t)\},$$

$$\mathcal{E}_{\text{dep}} := \{\text{dependence-cushion inequality in Assumption 5.15 holds for all } t, j\},$$

$$\mathcal{E}_{\text{env}} := \text{envelope-validity event in Assumption 5.16},$$

$$\mathcal{E}_{\text{mis}} := \{\forall t \geq 0 : A_t^{\text{mis}} = 0\}.$$

Theorem 8.1 gives  $\mathbb{P}(\mathcal{E}_{\text{out}}) \geq 1 - \alpha_{\text{out}}$ ; Assumption 5.14 (or Proposition 9.1) gives  $\mathbb{P}(\mathcal{E}_{\text{ev}}) \geq 1 - \alpha_{\text{ev}}$ , Proposition 9.2 gives  $\mathbb{P}(\mathcal{E}_{\text{dep}}) \geq 1 - \alpha_{\text{dep}}$ , Assumption 5.16 gives  $\mathbb{P}(\mathcal{E}_{\text{env}}) \geq 1 - \alpha_{\text{env}}$ , and Proposition 9.8 gives  $\mathbb{P}(\mathcal{E}_{\text{mis}}) \geq 1 - \alpha_{\text{mis}}$ . Hence, by union bound,

$$\mathbb{P}(\mathcal{E}_{\text{out}} \cap \mathcal{E}_{\text{ev}} \cap \mathcal{E}_{\text{dep}} \cap \mathcal{E}_{\text{env}} \cap \mathcal{E}_{\text{mis}}) \geq 1 - \alpha_{\text{tot}}.$$

On  $\mathcal{E}_{\text{out}}$ , Proposition 8.2 yields  $\theta^* \in \Theta_t^{\text{out,num}}(\alpha_{\text{out}})$  for all  $t$ . Together with  $\mathcal{E}_{\text{ev}}$ , this implies  $\theta^* \in \Theta_t^{\text{op}}(\alpha_{\text{out}})$  for all  $t$ , so

$$\inf_{\theta \in \Theta_t^{\text{op}}(\alpha_{\text{out}})} S_j^\theta \leq S_j^{\theta^*} \leq \sup_{\theta \in \Theta_t^{\text{op}}(\alpha_{\text{out}})} S_j^\theta.$$

Adding auditable optimization margin  $\varepsilon_{\text{opt},j,t}$  and aggregate cushion  $b_t^\Sigma$  (implementation, contamination, dependence, interaction, proxy-link, and rectangularization components) preserves containment; if  $\Theta_t^{\text{op}} = \emptyset$ ,  $\mathcal{I}_{j,t} = (-\infty, \infty)$  by definition. Therefore  $S_j^{\theta^*}(H_t) \in \mathcal{I}_{j,t}$  for all  $t, j$ .  $\square$

**Corollary 9.7** (Anytime false declaration control). *Let*

$$\mathcal{B}_t^{*,\varepsilon} = \left\{ j : S_j^{\theta^*}(H_t) \geq \max_{k \neq j} S_k^{\theta^*}(H_t) + \varepsilon \right\},$$

and let  $\hat{j}_t$  denote the unique element of  $\widehat{\mathcal{B}}_t^\varepsilon$  when  $|\widehat{\mathcal{B}}_t^\varepsilon| = 1$ . Define

$$\mathcal{M}_1 := \left\{ \exists t \geq 0 : |\widehat{\mathcal{B}}_t^\varepsilon| = 1, |\mathcal{B}_t^{*,\varepsilon}| = 1, \hat{j}_t \neq j_t^* \right\},$$

$$\mathcal{M}_2 := \left\{ \exists t \geq 0 : |\widehat{\mathcal{B}}_t^\varepsilon| = 1, |\mathcal{B}_t^{*,\varepsilon}| \neq 1 \right\},$$

and  $\mathcal{M} := \mathcal{M}_1 \cup \mathcal{M}_2$ , where  $j_t^*$  is the unique element of  $\mathcal{B}_t^{*,\varepsilon}$  on  $\{|\mathcal{B}_t^{*,\varepsilon}| = 1\}$ . Then  $\mathbb{P}(\mathcal{M}) \leq \alpha_{\text{tot}}$ .

*Proof.* Let  $\mathcal{E}_{\text{all}}$  be the event in Theorem 9.6. On  $\mathcal{E}_{\text{all}}$ , for all  $t, j$ ,  $S_j^{\theta^*}(H_t) \in [\widehat{L}_{j,t}, \widehat{U}_{j,t}]$ . Fix  $t$  with  $|\widehat{\mathcal{B}}_t^\varepsilon| = 1$ , and denote its element by  $\hat{j}_t$ . By definition of  $\widehat{\mathcal{B}}_t^\varepsilon$ ,

$$\widehat{L}_{\hat{j}_t,t} \geq \max_{k \neq \hat{j}_t} \widehat{U}_{k,t} + \varepsilon + \tau_{\text{num}}.$$

Hence for each  $k \neq \hat{j}_t$ ,

$$S_{\hat{j}_t}^{\theta^*}(H_t) \geq \widehat{L}_{\hat{j}_t,t} \geq \widehat{U}_{k,t} + \varepsilon + \tau_{\text{num}} \geq S_k^{\theta^*}(H_t) + \varepsilon.$$

Therefore  $|\mathcal{B}_t^{*,\varepsilon}| = 1$  and its unique element equals  $\hat{j}_t$ . So neither  $\mathcal{M}_1$  nor  $\mathcal{M}_2$  can occur on  $\mathcal{E}_{\text{all}}$ , i.e.  $\mathcal{M} \subseteq \mathcal{E}_{\text{all}}^c$ . Thus  $\mathbb{P}(\mathcal{M}) \leq \alpha_{\text{tot}}$ .  $\square$

**Proposition 9.8** (Misspecification alarm control under in-class truth). *Assume 5.18 and  $\theta^* \in \Theta_0$ . Then*

$$\mathbb{P}_{\theta^*}(\exists t \geq 0 : A_t^{\text{mis}} = 1) \leq \alpha_{\text{mis}}.$$

*If  $\theta^* \notin \Theta_0$ , no coverage claim is made for outer/inner sets; the protocol remains conservative by switching to fail-closed safe mode whenever  $A_t^{\text{mis}} = 1$ .*

*Proof.* Under  $\theta^* \in \Theta_0$ ,  $M_t^{\text{mis}}$  is a nonnegative supermartingale with  $M_0^{\text{mis}} = 1$ . Ville's inequality yields  $\mathbb{P}(\sup_t M_t^{\text{mis}} \geq 1/\alpha_{\text{mis}}) \leq \alpha_{\text{mis}}$ , equivalent to the alarm bound. The second statement is by protocol definition in Assumption 5.18.  $\square$

**Theorem 9.9** (Branchwise guarantees with explicit in-class/out-of-class split). *Define the coverage event*

$$\mathcal{E}_{\text{cov}} := \{\forall t \geq 0 : \theta^* \in \Theta_t^{\text{op}}(\alpha_{\text{out}})\},$$

and no-alarm event

$$\mathcal{E}_{\text{no-mis}} := \left\{ \sup_{t \geq 0} M_t^{\text{mis}} < \frac{1}{\alpha_{\text{mis}}} \right\}.$$



Under Assumptions 5.15–5.18, let  $\alpha_{\text{cov}}$  be any certified bound such that  $\mathbb{P}(\mathcal{E}_{\text{cov}}) \geq 1 - \alpha_{\text{cov}}$  (baseline:  $\alpha_{\text{cov}} = \alpha_{\text{out}} + \alpha_{\text{ev}}$ ). Define envelope/dependence validity events

$$\mathcal{E}_{\text{env}} := \left\{ \forall t \geq 0 : \eta_t \leq \bar{\eta}_t^{\text{pub}}, n_{\text{eff},t} \geq n_{\text{eff},t}^{\text{lb}}, \sigma_{S,t} \leq \bar{\sigma}_{S,t}^{\text{pub}}, b_t^{\text{int}} \text{ is certified} \right\},$$

$$\mathcal{E}_{\text{dep}} := \left\{ \forall t \geq 0, \forall j \in \mathcal{J} : \left| S_j^{\theta^*}(H_t) - \tilde{S}_{j,t}^{\theta^*} \right| \leq b_t^{\text{dep}} \right\}.$$

Then

(a) (**In-class branch, statistical validity**) If  $\theta^* \in \Theta_0$ , then

$$\mathbb{P}(\mathcal{E}_{\text{cov}} \cap \mathcal{E}_{\text{env}} \cap \mathcal{E}_{\text{dep}} \cap \mathcal{E}_{\text{no-mis}}) \geq 1 - \alpha_{\text{cov}} - \alpha_{\text{env}} - \alpha_{\text{dep}} - \alpha_{\text{mis}}.$$

On this event, all interval/declaration guarantees in Theorem 9.6 apply with the certified numerical/implementation cushions.

(b) (**Alarm branch, operational safety**) If  $A_t^{\text{mis}} = 1$  for some  $t$ , the protocol is fail-closed by design: unique bottleneck declarations are suppressed and only actions in  $\mathcal{A}_{\text{safe}}$  are permitted until re-specification.

(c) (**Out-of-class branch**) If  $\theta^* \notin \Theta_0$ , no statistical coverage claim is asserted. The certified guarantee is purely operational and given by (b).

*Proof.* Part (a): by construction,  $\mathbb{P}(\mathcal{E}_{\text{cov}}) \geq 1 - \alpha_{\text{cov}}$ ,  $\mathbb{P}(\mathcal{E}_{\text{env}}) \geq 1 - \alpha_{\text{env}}$ , and Proposition 9.2 gives  $\mathbb{P}(\mathcal{E}_{\text{dep}}) \geq 1 - \alpha_{\text{dep}}$ . Under  $\theta^* \in \Theta_0$ , Proposition 9.8 gives  $\mathbb{P}(\mathcal{E}_{\text{no-mis}}) \geq 1 - \alpha_{\text{mis}}$ . A union bound yields the displayed probability lower bound. On the intersection event, all premises of Theorem 9.6 are satisfied, hence interval/declaration guarantees hold. Part (b) is Assumption 5.18. Part (c) is definitional: without in-class truth, no coverage interpretation is claimed.  $\square$

**Proposition 9.10** (Out-of-class alarm-delay bound under certified drift). *Let  $\tau$  denote the first time the system enters an out-of-class regime. Define log-evidence increments for the misspecification monitor*

$$\ell_t := \log M_t^{\text{mis}} - \log M_{t-1}^{\text{mis}}, \quad t \geq 1.$$

Assume replay-certified constants  $\kappa_{\text{ow}} > 0$ ,  $\sigma_{\text{ow}} > 0$  such that for all  $u \geq 1$ ,

$$\mathbb{P}\left(\sum_{s=\tau}^{\tau+u-1} (\ell_s - \kappa_{\text{ow}}) \leq -x\right) \leq \exp\left(-\frac{x^2}{2u\sigma_{\text{ow}}^2}\right) \quad (\forall x \geq 0).$$

Let

$$T_{\text{alarm}} := \inf \left\{ t \geq \tau : M_t^{\text{mis}} \geq 1/\alpha_{\text{mis}} \right\}.$$

Then for every  $u$  with  $u\kappa_{\text{ow}} > \log(1/\alpha_{\text{mis}})$ ,

$$\mathbb{P}(T_{\text{alarm}} - \tau > u) \leq \exp\left(-\frac{(u\kappa_{\text{ow}} - \log(1/\alpha_{\text{mis}}))^2}{2u\sigma_{\text{ow}}^2}\right).$$

Hence detection delay is explicitly upper-bounded once minimal drift power is certified.

*Proof.* If  $T_{\text{alarm}} - \tau > u$ , then  $\sum_{s=\tau}^{\tau+u-1} \ell_s < \log(1/\alpha_{\text{mis}})$ . Rearranging gives  $\sum_{s=\tau}^{\tau+u-1} (\ell_s - \kappa_{\text{ow}}) < \log(1/\alpha_{\text{mis}}) - u\kappa_{\text{ow}}$ , and the stated tail bound follows from the concentration assumption.  $\square$

## 10 Dynamic IQC with FIR multipliers: full proof

### 10.1 Augmented delayed system

Consider local deviation dynamics:

$$x_{t+1} = A_0 x_t + \sum_{q=1}^h A_q x_{t-q} + B_w w_t + B_\varphi \varphi_t.$$

Define delay-augmented state

$$\xi_t = [x_t^\top, x_{t-1}^\top, \dots, x_{t-h}^\top]^\top.$$

Then

$$\xi_{t+1} = \bar{A} \xi_t + \bar{B}_w w_t + \bar{B}_\varphi \varphi_t.$$

Let

$$v_t = \begin{bmatrix} z_t \\ \varphi_t \end{bmatrix} = \begin{bmatrix} C_z \xi_t + D_{z\varphi} \varphi_t \\ \varphi_t \end{bmatrix}.$$

Choose FIR multiplier filter  $\Psi(z) = \sum_{\ell=0}^L \Psi_\ell z^{-\ell}$ , define

$$\psi_t := \sum_{\ell=0}^L \Psi_\ell v_{t-\ell}.$$

Assume discounted ( $\rho$ -hard) dynamic IQC [12, 13, 14]:

$$\sum_{t=0}^{N-1} \rho^{-2t} \psi_t^\top M \psi_t \geq 0, \quad \forall N \geq 1, \quad M = M^\top.$$

Introduce filter memory  $\zeta_t = [v_{t-1}^\top, \dots, v_{t-L}^\top]^\top$ , augmented state

$$s_t := [\xi_t^\top, \zeta_t^\top]^\top.$$

There exist matrices  $A_\chi, B_{\chi w}, B_{\chi\varphi}, C_\chi, D_{\chi\varphi}$  such that

$$s_{t+1} = A_\chi s_t + B_{\chi w} w_t + B_{\chi\varphi} \varphi_t, \quad \psi_t = C_\chi s_t + D_{\chi\varphi} \varphi_t.$$

**Theorem 10.1** (Dynamic IQC LMI certificate). *Fix  $0 < \rho < 1$ ,  $\gamma_{\text{iqc}} > 0$ . If there exist  $P \succ 0$ ,  $\lambda_{\text{iqc}} \geq 0$  with*

$$\underbrace{\begin{bmatrix} A_\chi^\top P A_\chi - \rho^2 P & A_\chi^\top P B_{\chi w} & A_\chi^\top P B_{\chi\varphi} \\ * & B_{\chi w}^\top P B_{\chi w} - \gamma_{\text{iqc}}^2 I & B_{\chi w}^\top P B_{\chi\varphi} \\ * & * & B_{\chi\varphi}^\top P B_{\chi\varphi} \end{bmatrix}}_{\Xi_0} + \lambda_{\text{iqc}} \underbrace{\begin{bmatrix} C_\chi^\top \\ 0 \\ D_{\chi\varphi}^\top \end{bmatrix} M \begin{bmatrix} C_\chi & 0 & D_{\chi\varphi} \end{bmatrix}}_{\Xi_{\text{IQC}}} \prec 0,$$

then there exists  $\eta > 0$  such that:

(i) *weighted finite-energy bound holds:*

$$\sum_{t=0}^{\infty} \rho^{-2(t+1)} \|s_t\|^2 \leq \frac{V_0}{\eta} + \frac{\gamma_{\text{iqc}}^2}{\eta} \sum_{t=0}^{\infty} \rho^{-2(t+1)} \|w_t\|^2;$$

(ii) *if additionally  $w \equiv 0$  and  $\psi_t^\top M \psi_t \geq 0$  for all  $t$ , then  $V_{t+1} \leq \rho^2 V_t$ , hence  $s_t$  is exponentially stable*

with rate  $\rho$ .

*Proof.* Define storage  $V_t = s_t^\top P s_t$ . Using system equation,

$$\begin{bmatrix} s_t \\ w_t \\ \varphi_t \end{bmatrix}^\top \Xi_0 \begin{bmatrix} s_t \\ w_t \\ \varphi_t \end{bmatrix} = V_{t+1} - \rho^2 V_t - \gamma_{\text{iqc}}^2 \|w_t\|^2.$$

Similarly,

$$\begin{bmatrix} s_t \\ w_t \\ \varphi_t \end{bmatrix}^\top \Xi_{\text{IQC}} \begin{bmatrix} s_t \\ w_t \\ \varphi_t \end{bmatrix} = \psi_t^\top M \psi_t.$$

Strict negativity of the LMI implies existence of  $\eta > 0$  with

$$V_{t+1} - \rho^2 V_t - \gamma_{\text{iqc}}^2 \|w_t\|^2 + \lambda_{\text{iqc}} \psi_t^\top M \psi_t \leq -\eta \|s_t\|^2.$$

Multiply by  $\rho^{-2(t+1)}$  and sum  $t = 0, \dots, N-1$ :

$$\rho^{-2N} V_N - V_0 - \gamma_{\text{iqc}}^2 \sum_{t=0}^{N-1} \rho^{-2(t+1)} \|w_t\|^2 + \lambda_{\text{iqc}} \sum_{t=0}^{N-1} \rho^{-2(t+1)} \psi_t^\top M \psi_t \leq -\eta \sum_{t=0}^{N-1} \rho^{-2(t+1)} \|s_t\|^2.$$

By discounted IQC, the weighted IQC sum is nonnegative. Dropping it and using  $V_N \geq 0$  gives

$$\eta \sum_{t=0}^{N-1} \rho^{-2(t+1)} \|s_t\|^2 \leq V_0 + \gamma_{\text{iqc}}^2 \sum_{t=0}^{N-1} \rho^{-2(t+1)} \|w_t\|^2.$$

Let  $N \rightarrow \infty$  to obtain (i). For (ii), under  $w \equiv 0$  and pointwise nonnegativity  $\psi_t^\top M \psi_t \geq 0$ , the one-step inequality yields  $V_{t+1} \leq \rho^2 V_t - \eta \|s_t\|^2 \leq \rho^2 V_t$ , hence exponential decay.  $\square$

*Remark 10.2* (From energy to peaks: two auditable bridges). Theorem 10.1 yields a weighted energy certificate. Peak-level claims for  $x_t, z_t, g_t$ , and gate probabilities require an explicit bridge. This manuscript supports two replay-verifiable options: (i) a reachable-tube bridge (deterministic but potentially expensive), and (ii) an energy-to-peak bridge that maps weighted energy certificates directly to conservative pointwise envelopes. If neither bridge is certified, the protocol reverts to baseline (fail-closed).

**Proposition 10.3** (Conservatism ordering). *Static IQC is recovered by  $L = 0, \Psi_0 = I$ , and scalar delay-margin tests embed as restricted multiplier/Lyapunov subclasses. Hence dynamic FIR-IQC feasible set weakly contains those subclasses and is weakly less conservative.*

*Proof.* Set  $L = 0$  to obtain static multiplier. Restrict  $P$  and  $M$  to diagonal/sector forms corresponding to scalar margin tests. Every such restricted feasible point remains feasible in the unrestricted dynamic-IQC program. Therefore feasible-set inclusion holds.  $\square$

**Theorem 10.4** (Bridge from IQC feasibility to diagnostic constants). *Assume the LMI in Theorem 10.1 is feasible at time  $t$ , and a deterministic calibration map*

$$\mathcal{C}_{\text{iqc}} : (P, \lambda_{\text{iqc}}, \Psi_\ell, M, \text{noise envelope}) \mapsto (\bar{g}_t^{\text{pk}}, \phi_t^{\text{pk}}, \varepsilon_{\phi,t}^{\text{pk}})$$

*is published and replay-verifiable, with*

$$\bar{g}_t^{\text{pk}} \leq \bar{g}, \quad \phi_t^{\text{pk}}(\kappa) \geq \phi(\kappa) \quad \forall \kappa \in [0, 1], \quad \varepsilon_{\phi,t}^{\text{pk}} \leq \varepsilon_{\text{gate}}.$$

Then Assumptions 5.7–5.8 remain valid with these tightened constants. If score endpoints are Lipschitz in these envelopes with modulus  $L_{\text{env}}$ , then interval half-width contracts by at least

$$\Delta W_t \geq L_{\text{env}} \left( (\bar{g} - \bar{g}_t^{\text{pk}}) + (\varepsilon_{\text{gate}} - \varepsilon_{\phi,t}^{\text{pk}}) + \inf_{\kappa \in [0,1]} (\phi_t^{\text{pk}}(\kappa) - \phi(\kappa)) \right)_+.$$

Consequently, any unique-declaration time under baseline envelopes is weakly improved (no later) under tightened IQC-derived envelopes.

*Proof.* By construction of  $\mathcal{C}_{\text{iqc}}$ , the tightened constants satisfy the same inequality templates as Assumptions 5.7–5.8 with smaller (or equal) adverse terms. Thus all downstream guarantees that depend monotonically on these constants remain valid and weakly sharpened. The width statement follows from envelope-to-width Lipschitz continuity and monotonicity of the interval construction.  $\square$

**Proposition 10.5** (Reachability-free energy-to-peak bridge). *Fix time  $t$ , horizon  $H$ , and state channel  $j$ . Assume IQC pipeline produces certified nonnegative bounds  $\bar{e}_{t,u}^{\text{IQC}}$  on per-step state-energy increment over  $u = 0, \dots, H-1$ , and suppose output map satisfies  $\|z_{t+u}\| \leq c_{z,t,u} \|x_{t+u}\|$ , with state recursion local Lipschitz bound*

$$\|x_{t+u+1}\| - \|x_{t+u}\| \leq L_{g,t,u} \|x_{t+u}\| + L_{\chi,t,u} \chi_{t+u} \|x_{t+u}\| + \bar{w}_{t,u},$$

where  $0 \leq \chi_{t+u} \leq 1$  and  $\bar{w}_{t,u}$  is certified. The structural coefficients  $L_{g,t,u}, L_{\chi,t,u}, c_{z,t,u}$  need not be static constants: it suffices to publish deterministic upper envelopes

$$\bar{L}_{g,t,u} \geq L_{g,t,u}, \quad \bar{L}_{\chi,t,u} \geq L_{\chi,t,u}, \quad \bar{c}_{z,t,u} \geq c_{z,t,u},$$

obtained from replay-verifiable observable secant-slope monitors (input/output increments only), open-world residual cushions, and misspecification alarms; no latent-state Jacobian observation is assumed. Then a replay-verifiable peak envelope

$$\bar{g}_{t,u}^{\text{eng}} := \bar{c}_{z,t,u} \left( \|x_t\| + \sum_{\nu=0}^{u-1} \sqrt{\bar{e}_{t,\nu}^{\text{IQC}}} + \sum_{\nu=0}^{u-1} \bar{w}_{t,\nu} \right) \exp \left( \sum_{\nu=0}^{u-1} (\bar{L}_{g,t,\nu} + \bar{L}_{\chi,t,\nu}) \right)$$

satisfies  $\|z_{t+u}\| \leq \bar{g}_{t,u}^{\text{eng}}$  for all  $u \leq H$ . Hence one may set

$$\delta_t^{\text{peak}} := \sup_{0 \leq u \leq H} \bar{g}_{t,u}^{\text{eng}}$$

without reachable-set computation.

*Proof.* From the recursion inequality, discrete Grönwall with time-varying coefficients gives

$$\|x_{t+u}\| \leq \left( \|x_t\| + \sum_{\nu=0}^{u-1} \Delta_{\nu} \right) \exp \left( \sum_{\nu=0}^{u-1} (L_{g,t,\nu} + L_{\chi,t,\nu}) \right),$$

with  $\Delta_{\nu} := \|x_{t+\nu+1}\| - \|x_{t+\nu}\|$  upper-bounded by  $\sqrt{\bar{e}_{t,\nu}^{\text{IQC}}} + \bar{w}_{t,\nu}$  from energy certificate plus disturbance envelope. Replacing unknown local coefficients by published upper envelopes preserves the inequality. Multiplying by  $\bar{c}_{z,t,u}$  yields  $\|z_{t+u}\| \leq \bar{g}_{t,u}^{\text{eng}}$ . Taking supremum over  $u \in \{0, \dots, H\}$  gives  $\delta_t^{\text{peak}}$ . All terms are deterministic functions of replay artifacts, so certificate is auditable.  $\square$

**Proposition 10.6** (Lag-one non-circular IQC tightening). *Let  $\Theta_t^{\text{op}}$  and score intervals at time  $t$  be computed with baseline envelopes only. Any IQC-based tightening computed from  $\Theta_t^{\text{op}}$  (using either*

Proposition 10.7 or 10.5) is activated only for time  $t + 1$  and later. Then no fixed-point circularity occurs between set construction and IQC tightening at the same time index.

*Proof.* By protocol definition,  $\Theta_t^{\text{op}}$  is measurable with respect to data and artifacts available before IQC tightening is applied at  $t + 1$ . Hence the map  $\Theta_t^{\text{op}} \mapsto$  tightened constants is feed-forward in time, not self-referential at time  $t$ .  $\square$

**Proposition 10.7** (Reachability-certified replay calibration map). *Assume the LMI of Theorem 10.1 is feasible and a deterministic reachable-tube certificate is published for horizon  $H_r$ :*

$$\|x_{t+\tau}\|_\infty \leq r_{t,\tau}^x, \quad \|z_{t+\tau}\|_\infty \leq r_{t,\tau}^z, \quad \tau = 0, \dots, H_r,$$

*valid for the declared initial/disturbance envelope up to replay error  $\varepsilon_{\text{reach},t}$ . Assume published structural bounds*

$$g_{t+\tau} \leq g_0 + L_g \|x_{t+\tau}\|_\infty,$$

*and a replay-verifiable shock-probability upper bound*

$$\mathbb{P}(\chi_{t+\tau} = 1 \mid \mathcal{G}_{t+\tau}^-) \leq p_{t+\tau}^{\text{ub}}(\hat{\kappa}_{t+\tau}; \|z_{t+\tau}\|_\infty),$$

*where  $p_s^{\text{ub}}(\kappa; \cdot)$  is nondecreasing. Define*

$$\bar{g}_t^{\text{pk}} := \min \left\{ \bar{g}, g_0 + L_g \max_{0 \leq \tau \leq H_r} (r_{t,\tau}^x + \varepsilon_{\text{reach},t}) \right\},$$

$$p_t^{\text{max}}(\kappa) := \max_{0 \leq \tau \leq H_r} p_{t+\tau}^{\text{ub}}(\kappa; r_{t,\tau}^z + \varepsilon_{\text{reach},t}),$$

$$\varepsilon_{\phi,t}^{\text{pk}} := \varepsilon_{\text{gate}}, \quad \phi_t^{\text{pk}}(\kappa) := 1 + \varepsilon_{\text{gate}} - p_t^{\text{max}}(\kappa).$$

*If  $p_t^{\text{max}}(\kappa) \leq 1 - \phi(\kappa) + \varepsilon_{\text{gate}}$  for all  $\kappa \in [0, 1]$ , then*

$$\bar{g}_t^{\text{pk}} \leq \bar{g}, \quad \phi_t^{\text{pk}}(\kappa) \geq \phi(\kappa) \quad \forall \kappa, \quad \varepsilon_{\phi,t}^{\text{pk}} \leq \varepsilon_{\text{gate}}.$$

*Hence the calibration map is deterministic and replay-verifiable from public tube artifacts, and Theorem 10.4 applies.*

*Proof.* The  $\bar{g}_t^{\text{pk}}$  inequality follows from the structural bound on  $g$ , monotonicity of  $\max$ , and truncation by  $\bar{g}$ . By definition of  $p_t^{\text{max}}$ , conditional shock probabilities are bounded by  $p_t^{\text{max}}(\kappa)$ , so

$$\mathbb{P}(\chi = 1 \mid \mathcal{G}^-) \leq p_t^{\text{max}}(\kappa) = 1 - \phi_t^{\text{pk}}(\kappa) + \varepsilon_{\phi,t}^{\text{pk}}.$$

The dominance condition  $p_t^{\text{max}}(\kappa) \leq 1 - \phi(\kappa) + \varepsilon_{\text{gate}}$  implies  $\phi_t^{\text{pk}}(\kappa) \geq \phi(\kappa)$ . All quantities are deterministic functions of published certificates and tolerances, so replay verification is direct.  $\square$

**Corollary 10.8** (Fail-closed rule for IQC-based tightening). *If the replay bundle at time  $t$  includes neither (a) a valid reachable-tube certificate satisfying Proposition 10.7, nor (b) a valid energy envelope satisfying Proposition 10.5, then no peak-level tightening is allowed:*

$$\bar{g}_t^{\text{pk}} := \bar{g}, \quad \phi_t^{\text{pk}} := \phi, \quad \varepsilon_{\phi,t}^{\text{pk}} := \varepsilon_{\text{gate}}.$$

*Hence IQC feasibility never by itself justifies peak-level claims without an auditable bridge artifact.*

*Proof.* Without (a) or (b), assumptions required to map IQC energy certificates into peak envelopes are missing. The protocol therefore reverts to baseline by definition, preserving soundness.  $\square$

## 11 Practical structural bounds and decision rules

**Theorem 11.1** (Physical throughput ceiling). *Let  $\bar{g} := \bar{g}_0 + \bar{g}_P P_{\max}$ . Under Assumptions 5.7–5.8 and  $\sup_t \mathbb{E}|\xi_t^C| \leq \mu_C$ ,*

$$\mathbb{E}[C_{t+1}] \leq (1 - \delta_C)\mathbb{E}[C_t] + \bar{g} \cdot \mathbb{E}[\min\{1, 1 - \phi(\hat{\kappa}_t) + \varepsilon_{\text{gate}}\}] + \mu_C \leq (1 - \delta_C)\mathbb{E}[C_t] + \bar{g} + \mu_C.$$

Therefore

$$\limsup_{t \rightarrow \infty} \mathbb{E}[C_t] \leq \frac{\bar{g} + \mu_C}{\delta_C}.$$

*Proof.* From (4),

$$\mathbb{E}[C_{t+1}] = (1 - \delta_C)\mathbb{E}[C_t] + \mathbb{E}[\chi_t g_t] + \mathbb{E}[\xi_t^C].$$

Using  $g_t \leq \bar{g}$  and tower property,

$$\mathbb{E}[\chi_t g_t] \leq \bar{g} \mathbb{E}[\chi_t] = \bar{g} \mathbb{E}[\mathbb{E}[\chi_t \mid \mathcal{G}_t^-]] \leq \bar{g} \mathbb{E}[\min\{1, 1 - \phi(\hat{\kappa}_t) + \varepsilon_{\text{gate}}\}].$$

Also  $\mathbb{E}[\xi_t^C] \leq \mathbb{E}|\xi_t^C| \leq \mu_C$ . Substitute and iterate the scalar recursion.  $\square$

**Theorem 11.2** (Concentration-induced drift erosion). *If  $g_t \leq \bar{g}$  almost surely and Assumption 5.8 holds:*

$$\mathbb{E}[C_{t+1} - C_t \mid \mathcal{G}_t^-] \leq -\delta_C \mathbb{E}[C_t \mid \mathcal{G}_t^-] + \bar{g} \min\{1, 1 - \phi(\hat{\kappa}_t) + \varepsilon_{\text{gate}}\} + \mathbb{E}[\xi_t^C \mid \mathcal{G}_t^-].$$

*Proof.* Condition on  $\mathcal{G}_t^-$ :

$$\mathbb{E}[C_{t+1} - C_t \mid \mathcal{G}_t^-] = -\delta_C \mathbb{E}[C_t \mid \mathcal{G}_t^-] + \mathbb{E}[\chi_t g_t \mid \mathcal{G}_t^-] + \mathbb{E}[\xi_t^C \mid \mathcal{G}_t^-].$$

Because  $g_t \leq \bar{g}$ ,

$$\mathbb{E}[\chi_t g_t \mid \mathcal{G}_t^-] \leq \bar{g} \mathbb{E}[\chi_t \mid \mathcal{G}_t^-] \leq \bar{g} \min\{1, 1 - \phi(\hat{\kappa}_t) + \varepsilon_{\text{gate}}\}.$$

Substitute and conclude.  $\square$

### 11.1 Identification-aware no-meta decision protocol

At each time  $t$ :

(D1) Update  $E_t(\theta)$ ,  $M_t^{\text{mis}}$ ,  $\Theta_t^{\text{out}}(\alpha_{\text{out}})$ ,  $\Theta_t^{\text{out,num}}(\alpha_{\text{out}})$ , and  $\Theta_t^{\text{in}}(\alpha_{\text{out}}, r_t)$ .

(D2) Compute auditable score intervals  $\mathcal{I}_{j,t}$  and the top-gap

$$\text{Gap}_t^{\text{top}} := \hat{L}_{j_t^L, t} - \max_{k \neq j_t^L} \hat{U}_{k, t}, \quad j_t^L \in \arg \max_j \hat{L}_{j, t}.$$

(D3) (**Alarm gate**) If  $A_t^{\text{mis}} = 1$ , suppress unique attribution, enforce  $A_t \in \mathcal{A}_{\text{safe}}$ , and record fail-closed alarm branch.

(D4) (**Compute-certification gate**) Else, if

$$2 \left( \max_j \varepsilon_{\text{opt}, j, t} + b_t^{\text{impl}} \right) > (\varepsilon + \tau_{\text{num}} + \text{Gap}_t^{\text{top}})_+,$$

output **compute-limited soft abstention**: no unique bottleneck declaration, but emit machine-only computational-slack indices  $\pi_{j,t}^{\text{soft}}$  together with operator-facing ordinal bands  $b_{j,t}^{\text{soft}} \in \{R0, R1, R2, R3, R4\}$ , then execute the robust fallback/backup action.

(D5) Else, if  $|\hat{\mathcal{B}}_t^\varepsilon| = 1$ , allocate primary relief to the declared channel.

(D6) Else soft-abstain from unique attribution. If  $\Theta_t^{\text{op}}(\alpha_{\text{out}}) \neq \emptyset$ , solve the robust probe-diversification problem over  $\mathcal{U}_t^{\text{probe}}(h_t)$ :

$$\max_{(r,p) \in \mathcal{U}_t^{\text{probe}}(h_t)} \inf_{\theta \in \Theta_t^{\text{op}}(\alpha_{\text{out}})} \left( G_{t,H}^\theta(h_t; r, p) - \lambda_p \text{Cost}(p) - \lambda_u \text{WidthAfter}(p) - \lambda_{h,t} h_t^{\text{alias}}(h_t, p) \right).$$

Here  $G_{t,H}^\theta(h_t; r, p)$  allows probe-dependent immediate utility; when probe only affects penalties and information state, set  $G_{t,H}^\theta(h_t; r, p) := G_{t,H}^\theta(h_t; r)$ . The term  $h_t^{\text{alias}}(h_t, p)$  is the replay-published worst-case hazard envelope over current observational-alias classes after applying probe design  $p$ ; it penalizes probe choices that reduce interval width while increasing catastrophic alias risk. If  $\Theta_t^{\text{op}}(\alpha_{\text{out}}) = \emptyset$ , execute fail-closed fallback  $(r^{\text{safe}}, 0)$ .

(D7) (**Persistent identifiability guard**) If ambiguity persists and a deterministic counter shows probe density below  $\rho_{\text{probe}}$ , force exploration from a replay-published finite catalog  $\mathcal{P}_t^{\text{grid}} \subset \mathcal{P}_t$  that satisfies the Assumption 11.6 lower-information certificate.

**Proposition 11.3** (Existence of the D6 optimizer and fail-closed determinism). *Fix  $t, h_t$ , and define*

$$\mathcal{J}_t(r, p; h_t) := \inf_{\theta \in \Theta_t^{\text{op}}(\alpha_{\text{out}})} G_{t,H}^\theta(h_t; r, p) - \lambda_p \text{Cost}(p) - \lambda_u \text{WidthAfter}(p) - \lambda_{h,t} h_t^{\text{alias}}(h_t, p).$$

*Assume  $\Theta_t^{\text{op}}(\alpha_{\text{out}})$  is nonempty compact and  $(\theta, r, p) \mapsto G_{t,H}^\theta(h_t; r, p)$  is jointly continuous on  $\Theta_t^{\text{op}}(\alpha_{\text{out}}) \times \mathcal{U}_t^{\text{probe}}(h_t)$ . Under Assumption 5.20,  $\mathcal{J}_t(\cdot, \cdot; h_t)$  is upper-semicontinuous on the compact feasible set  $\mathcal{U}_t^{\text{probe}}(h_t)$ , and therefore*

$$\arg \max_{(r,p) \in \mathcal{U}_t^{\text{probe}}(h_t)} \mathcal{J}_t(r, p; h_t) \neq \emptyset$$

*is compact. If  $\Theta_t^{\text{op}}(\alpha_{\text{out}}) = \emptyset$ , the fallback action  $(r^{\text{safe}}, 0) \in \mathcal{U}_t^{\text{probe}}(h_t)$  is feasible and deterministic by Assumption 5.20.*

*Proof.* For each  $(r, p)$ , continuity in  $\theta$  and compactness of  $\Theta_t^{\text{op}}$  imply that  $\inf_{\theta \in \Theta_t^{\text{op}}} G_{t,H}^\theta(h_t; r, p)$  is attained and continuous in  $(r, p)$  by Berge's maximum theorem [16]. Assumption 5.20 gives lower-semicontinuity of Cost, WidthAfter, and  $h_t^{\text{alias}}$ , hence  $-\lambda_p \text{Cost} - \lambda_u \text{WidthAfter} - \lambda_{h,t} h_t^{\text{alias}}$  is upper-semicontinuous. Therefore  $\mathcal{J}_t$  is upper-semicontinuous on compact  $\mathcal{U}_t^{\text{probe}}(h_t)$ , so Weierstrass yields a nonempty compact argmax. When  $\Theta_t^{\text{op}} = \emptyset$ , D6 is defined to execute the published fail-closed fallback; feasibility is guaranteed by Assumption 5.20.  $\square$

**Theorem 11.4** (Abstention+probe weak dominance under ambiguity). *Define, for policy  $\pi$ , model  $\theta$ , and decision time  $t$ ,*

$$W_{t,H}^\theta(h_t; \pi) := \mathbb{E}_\theta^\pi \left[ \sum_{s=t}^{t+H-1} \gamma^{s-t} \left( r_s(H_s, A_s, H_{s+1}) - \lambda_p \text{Cost}(P_s) - \lambda_u \text{WidthAfter}(P_s) - \lambda_{h,s} h_s^{\text{alias}}(H_s, P_s) \right) \mid H_t = h_t \right].$$

where  $A_s = (R_s, P_s)$  splits relief and probe components. Let  $\Pi_{\text{forced}}$  be the nonempty class of universally measurable policies that always commit to a single channel and never probe. Let  $\Pi_{\text{AP}}$  be the nonempty class of universally measurable abstention+probe policies in (D6), with  $p_t = 0$  admissible and safe fallback always feasible. Assume  $\Theta_t^{\text{op}}(\alpha_{\text{out}}) \neq \emptyset$ . Then

$$\sup_{\pi \in \Pi_{\text{AP}}} \inf_{\theta \in \Theta_t^{\text{op}}(\alpha_{\text{out}})} W_{t,H}^\theta(h_t; \pi) \geq \sup_{\pi \in \Pi_{\text{forced}}} \inf_{\theta \in \Theta_t^{\text{op}}(\alpha_{\text{out}})} W_{t,H}^\theta(h_t; \pi).$$

Under Assumption 11.7, the same dominance statement holds for the executable finite-catalog policy class, up to additive  $\varepsilon_{\text{pol},t}$ . Moreover, fix  $\varepsilon_{\text{cmp}} > 0$ , and choose an  $\varepsilon_{\text{cmp}}$ -optimal forced comparator

$$\pi^{\text{for},\varepsilon} \in \Pi_{\text{forced}} \quad \text{such that} \quad \inf_{\theta \in \Theta_t^{\text{op}}(\alpha_{\text{out}})} W_{t,H}^\theta(h_t; \pi^{\text{for},\varepsilon}) \geq \sup_{\pi \in \Pi_{\text{forced}}} \inf_{\theta \in \Theta_t^{\text{op}}(\alpha_{\text{out}})} W_{t,H}^\theta(h_t; \pi) - \varepsilon_{\text{cmp}}.$$

If there exist  $\pi^\dagger \in \Pi_{\text{AP}}$  and  $\delta_0 > 0$  such that

$$\inf_{\theta \in \Theta_t^{\text{op}}(\alpha_{\text{out}})} \left( W_{t,H}^\theta(h_t; \pi^\dagger) - W_{t,H}^\theta(h_t; \pi^{\text{for},\varepsilon}) \right) \geq \delta_0,$$

then

$$\sup_{\pi \in \Pi_{\text{AP}}} \inf_{\theta \in \Theta_t^{\text{op}}(\alpha_{\text{out}})} W_{t,H}^\theta(h_t; \pi) \geq \sup_{\pi \in \Pi_{\text{forced}}} \inf_{\theta \in \Theta_t^{\text{op}}(\alpha_{\text{out}})} W_{t,H}^\theta(h_t; \pi) + \delta_0 - \varepsilon_{\text{cmp}}.$$

In particular, since  $\varepsilon_{\text{cmp}} > 0$  is arbitrary, AP attains a strict robust gain margin arbitrarily close to  $\delta_0$  whenever the displayed model-uniform advantage certificate holds.

*Proof.*  $\Pi_{\text{forced}} \subseteq \Pi_{\text{AP}}$  because Assumption 5.20 embeds all forced non-probing actions as  $(r, 0)$ -choices in  $\mathcal{U}_t^{\text{probe}}(h_t)$ , and AP can always select deterministic one-channel relief with  $p = 0$ . Taking sup inf over a superset of policies gives weak dominance.

For the strict part, by definition of supremum,

$$\sup_{\pi \in \Pi_{\text{AP}}} \inf_{\theta \in \Theta_t^{\text{op}}} W_{t,H}^\theta(h_t; \pi) \geq \inf_{\theta \in \Theta_t^{\text{op}}} W_{t,H}^\theta(h_t; \pi^\dagger) \geq \inf_{\theta \in \Theta_t^{\text{op}}} W_{t,H}^\theta(h_t; \pi^{\text{for},\varepsilon}) + \delta_0.$$

Using  $\varepsilon_{\text{cmp}}$ -optimality of  $\pi^{\text{for},\varepsilon}$ ,

$$\inf_{\theta \in \Theta_t^{\text{op}}} W_{t,H}^\theta(h_t; \pi^{\text{for},\varepsilon}) \geq \sup_{\pi \in \Pi_{\text{forced}}} \inf_{\theta \in \Theta_t^{\text{op}}} W_{t,H}^\theta(h_t; \pi) - \varepsilon_{\text{cmp}}.$$

Combine both inequalities. □

**Definition 11.5** (Probe information lower envelope). Fix  $\delta_\Theta > 0$ , time  $t$ , and realized history  $h_t$ . For any admissible probe  $p \in \mathcal{P}_t$ , define

$$\text{IGLower}_t(p; h_t) := \inf_{\substack{\theta, \theta' \in \Theta_t^{\text{op}}(\alpha_{\text{out}}) \\ d_\Theta(\theta, \theta') \geq \delta_\Theta}} \min \left\{ \mathbb{E}_\theta \left[ \log \frac{dP_{\theta,p}}{dP_{\theta',p}} \middle| \mathcal{G}_t^- \right], \mathbb{E}_{\theta'} \left[ \log \frac{dP_{\theta',p}}{dP_{\theta,p}} \middle| \mathcal{G}_t^- \right] \right\},$$

with the convention  $\inf \emptyset := +\infty$ .

**Assumption 11.6** (Probe-identifiability under no-meta constraints). There exist constants  $\delta_\Theta > 0$ ,  $\kappa_{\text{id}} > 0$ ,  $c_{\text{id}} > 0$ ,  $\nu_{\text{id}} > 0$ , and replay-published finite probe grids  $\mathcal{P}_t^{\text{grid}} \subset \mathcal{P}_t$  such that:

- (p1) (**Separation over observable inequivalence classes**) for any  $\theta, \theta' \in \Theta_t^{\text{op}}(\alpha_{\text{out}})$  with  $d_\Theta(\theta, \theta') \geq \delta_\Theta$  and  $\theta \not\sim_t^{\text{obs}} \theta'$ , there exists  $p \in \mathcal{P}_t^{\text{grid}}$  with one-step public-history divergence

$$D_{\text{KL}}(\mathbb{P}_\theta(\cdot \mid h_t, p) \parallel \mathbb{P}_{\theta'}(\cdot \mid h_t, p)) \geq \kappa_{\text{id}}.$$



- (p2) (**Probe realization frequency**) when protocol enforces exploration (D7), the realized effective probe density is at least  $\rho_{\text{probe}} > 0$  along ambiguity epochs.
- (p3) (**Information accumulation**) there exists an adapted information score  $\mathcal{I}_t$  such that on ambiguity epochs

$$\mathcal{I}_{t+1} - \mathcal{I}_t \geq \nu_{\text{id}} \mathbf{1}\{\text{effective probe at } t\}$$

up to bounded martingale noise  $m_t$  with sub-Gaussian parameter  $c_{\text{id}}$ .

- (p4) (**Alias is non-rejection, not equivalence**) if  $\theta \sim_t^{\text{obs}} \theta'$ , protocol may not collapse them into a single trusted point. Each alias class must carry replay-published hazard envelopes, and the protocol must provide an aggregate penalty map  $p \mapsto h_t^{\text{alias}}(h_t, p)$  computed from stress-probe outcomes and worst-case consequence bounds; D6 optimization is required to remain robust over this hazard envelope.

**Assumption 11.7** (Constructive policy catalog and approximation envelope). For each  $(t, h_t)$ , there is a finite replay-published catalog  $\mathcal{C}_t(h_t) \subseteq \mathcal{A}(h_t)$ , a deterministic tie-break rule, and a public approximation radius  $\varepsilon_{\text{pol},t} \geq 0$  such that:

- (c1) (**Comparator retention**) all forced one-channel comparators used by  $\Pi_{\text{forced}}$  are contained in  $\mathcal{C}_t(h_t)$ .
- (c2) (**Uniform approximation envelope**) for the horizon- $H$  robust objective kernel

$$G_{t,H}^\theta(h_t; r, p) := \mathbb{E}_\theta \left[ \sum_{s=t}^{t+H-1} \gamma^{s-t} \left( r_s(H_s, A_s, H_{s+1}) - \lambda_p \text{Cost}(P_s) - \lambda_u \text{WidthAfter}(P_s) - \lambda_{h,s} h_s^{\text{alias}}(H_s, P_s) \right) \mid \begin{array}{l} H_t = h_t, \\ (R_t, P_t) = (r, p) \end{array} \right].$$

the catalog satisfies

$$\sup_{(r,p) \in \mathcal{A}(h_t)} G_{t,H}^\theta(h_t; r, p) - \max_{(r,p) \in \mathcal{C}_t(h_t)} G_{t,H}^\theta(h_t; r, p) \leq \varepsilon_{\text{pol},t} \quad \forall \theta \in \Theta_t^{\text{op}}(\alpha_{\text{out}}).$$

Score endpoint optimization over  $\theta$  still uses certified  $\varepsilon_{\text{opt},j,t}$  (Assumption 5.9);  $\varepsilon_{\text{pol},t}$  covers only policy-space discretization/approximation.

- (c3) (**Replay determinism**) catalog hash, tie-break rule, and argmax convention are published so that the selected action is deterministic under replay.

A replay-published nested refinement family

$$\mathcal{C}_t^{(0)}(h_t) \subseteq \mathcal{C}_t^{(1)}(h_t) \subseteq \dots, \quad \varepsilon_{\text{pol},t}^{(m+1)} \leq \varepsilon_{\text{pol},t}^{(m)}$$

must be available, with executed budget level  $m_t$  satisfying

$$\mathcal{C}_t(h_t) = \mathcal{C}_t^{(m_t)}(h_t), \quad \varepsilon_{\text{pol},t} = \varepsilon_{\text{pol},t}^{(m_t)}.$$

All catalog hashes, level indices, and  $\varepsilon_{\text{pol},t}^{(m)}$  values are part of replay metadata.

**Proposition 11.8** (Constructive replacement of measurable selection). *Assume 5.20 and 11.7. Then the D6 decision can be implemented by deterministic maximization over  $\mathcal{C}_t(h_t)$ , and its robust value is*

within  $\varepsilon_{\text{pol},t}$  of the measurable-policy robust optimum. In particular, no non-constructive selector is needed for executable protocol semantics.

*Proof.* By Assumption 11.7(c2), the catalog optimum upper-approximates the measurable-policy robust value up to  $\varepsilon_{\text{pol},t}$ . Assumption 11.7(c1) ensures forced one-channel comparators are retained in the executable set. Assumption 11.7(c3) makes the argmax and tie-breaking replay-deterministic. Hence D6 is both constructive and auditable, with explicit approximation error  $\varepsilon_{\text{pol},t}$ .  $\square$

**Theorem 11.9** (Alias-aware liveness decomposition). *Assume 11.6. Let  $\mathcal{A}_t^{\text{obs}}$  be the partition of  $\Theta_t^{\text{op}}(\alpha_{\text{out}})$  induced by  $\sim_t^{\text{obs}}$ , and define the set of  $\delta_\Theta$ -separable unresolved pairs not yet in the same alias class by*

$$\mathfrak{P}_t^{\text{sep}} := \{(\theta, \theta') : d_\Theta(\theta, \theta') \geq \delta_\Theta, \theta \not\sim_t^{\text{obs}} \theta'\}.$$

Then:

- (a) *On episodes where  $\mathfrak{P}_t^{\text{sep}} \neq \emptyset$  infinitely often and probe budgets continue, the ambiguity mass outside alias classes shrinks almost surely; perpetual abstention on separable alternatives has probability 0.*
- (b) *If perpetual abstention occurs, it is confined to (i) true  $\varepsilon$ -ties, or (ii) certified observable-alias classes  $\mathcal{A}_t^{\text{obs}}$ .*

Thus liveness is guaranteed on identifiable structure, while non-identifiability is isolated and explicitly reported.

*Proof.* Under (p1)–(p3), each effective informative probe produces positive expected information gain  $\kappa_{\text{id}}$  with bounded increments. Freedman/Bernstein-type concentration implies cumulative gain over separable unresolved pairs diverges when such probes occur infinitely often [19, 20]. By (p2), informative probes occur with positive frequency whenever separable unresolved pairs remain. Hence posterior/plausibility mass on incorrect separable alternatives vanishes almost surely, eliminating perpetual abstention on separable structure. Residual abstention can only persist when dominance is not uniquely identifiable in observables: true  $\varepsilon$ -ties or certified alias classes.  $\square$

**Corollary 11.10** (Eventual unique declaration under persistent true margin). *Assume the premises of Theorem 9.6,  $\theta^* \in \Theta_0$ , and suppose there exist  $m_\star > 0$ ,  $T_m < \infty$ , and a measurable index process  $j_t^\star$  such that on the no-alarm branch:*

$$\forall t \geq T_m : S_{j_t^\star}^{\theta^*}(H_t) - \max_{k \neq j_t^\star} S_k^{\theta^*}(H_t) - \varepsilon \geq m_\star,$$

and  $j_t^\star$  is unique. Let

$$b_t^\Sigma := b_{t,H}^{\text{cont}} + b_t^{\text{dep}} + b_t^{\text{int}} + b_t^{\text{rect,pub}} + b_t^{\text{impl}} + \max_j \varepsilon_{\text{opt},j,t}.$$

If

$$\forall t \geq T_m : 2b_t^\Sigma + \tau_{\text{num}} \leq m_\star,$$

then, with probability at least  $1 - \alpha_{\text{tot}}$ ,

$$\forall t \geq T_m : |\widehat{\mathcal{B}}_t^\varepsilon| = 1 \quad \text{and} \quad \hat{j}_t = j_t^\star.$$

In particular,  $T^\star := T_m$  is a valid finite certification time.

*Proof.* On the event in Theorem 9.6, for all  $t, j$ ,

$$S_j^{\theta^*}(H_t) \in [\hat{L}_{j,t}, \hat{U}_{j,t}],$$

hence for  $t \geq T_m$ ,

$$\hat{L}_{j_t^*,t} \geq S_{j_t^*}^{\theta^*}(H_t) - b_t^\Sigma, \quad \hat{U}_{k,t} \leq S_k^{\theta^*}(H_t) + b_t^\Sigma \quad (k \neq j_t^*).$$

Therefore

$$\hat{L}_{j_t^*,t} - \max_{k \neq j_t^*} \hat{U}_{k,t} \geq \left( S_{j_t^*}^{\theta^*} - \max_{k \neq j_t^*} S_k^{\theta^*} \right) - 2b_t^\Sigma \geq \varepsilon + m_\star - 2b_t^\Sigma \geq \varepsilon + \tau_{\text{num}}.$$

By definition of  $\hat{\mathcal{B}}_t^\varepsilon$ , channel  $j_t^*$  is uniquely declared. The probability is at least  $1 - \alpha_{\text{tot}}$  by Theorem 9.6.  $\square$

**Corollary 11.11** (Eventual positive computational-slack declaration under bounded compute). *Assume the premises of Corollary 11.10 except that hard uniqueness margin may be insufficient for all large  $t$ . If there exists  $m_\star^{\text{soft}} > 0$  such that*

$$\forall t \geq T_m : \quad \hat{L}_{j_t^*,t} - \max_{k \neq j_t^*} \hat{U}_{k,t} \geq -\varepsilon - \tau_{\text{num}} + m_\star^{\text{soft}},$$

*and endpoint certificate inflation is uniformly bounded, then with probability at least  $1 - \alpha_{\text{tot}}$ ,  $\pi_{j_t^*,t}^{\text{soft}}$  from Definition 12.2 is eventually bounded away from 0 and nondecreasing under refinement levels  $m_t$ . Thus liveness can be achieved as graded computational slack even when hard uniqueness is unavailable.*

*Proof.* The claim follows from monotonicity of  $\pi_{j,t}^{\text{soft}}$  in endpoint separation and from the monotone refinement property of certificates (Assumption 11.7). Whenever the lower-separation floor  $m_\star^{\text{soft}}$  holds, the numerator in Definition 12.2 remains positive eventually, yielding a strictly positive lower bound.  $\square$

## 12 Computational profile and complexity

**Complexity sketch (certificate-oriented).** For each time  $t$ , the dominant cost is robust endpoint certification over  $\Theta_t^{\text{op}}$ . Let  $N_{\text{node},t}$  be explored branch-and-bound nodes,  $N_{a,t}$  pruned action candidates,  $d_\theta$  model dimension, and  $C_{\text{sim}}(H)$  one rollout/backward-pass cost. A deterministic certificate run scales as

$$O(|\mathcal{J}| \cdot N_{\text{node},t} \cdot N_{a,t} \cdot C_{\text{sim}}(H))$$

plus LMI/reachability solves for IQC updates. Worst-case global search remains exponential in  $d_\theta$  for nonconvex models; therefore the protocol is explicitly *compute-aware*: it publishes  $(\varepsilon_{\text{opt},j,t}, b_t^{\text{impl}}, \varepsilon_{\text{pol},t})$ , declares compute-limited soft abstention when certificates are too loose, and treats branch-and-bound as an anytime certificate generator rather than an oracle. The policy-level approximation  $\varepsilon_{\text{pol},t}$  is separated from score-endpoint certification and is reported independently.

**Proposition 12.1** (Compute-limited abstention is mandatory when certificates are loose). *Let  $j_t^L \in \arg \max_j \hat{L}_{j,t}$ , and define*

$$\text{Gap}_t^{\text{top}} := \hat{L}_{j_t^L,t} - \max_{k \neq j_t^L} \hat{U}_{k,t}.$$

*If*

$$2 \left( \max_j \varepsilon_{\text{opt},j,t} + b_t^{\text{impl}} \right) > (\varepsilon + \tau_{\text{num}} + \text{Gap}_t^{\text{top}})_+,$$

the protocol must suppress unique declaration and enter *compute-limited soft abstention*. This rule is replay-verifiable, prevents false precision from prematurely terminated nonconvex optimization, and preserves graded liveness through  $\pi_t^{\text{soft}}$ .

*Proof.* A certified unique declaration at margin  $\varepsilon + \tau_{\text{num}}$  requires existence of some  $j^*$  with

$$\widehat{L}_{j^*,t} \geq \max_{k \neq j^*} \widehat{U}_{k,t} + \varepsilon + \tau_{\text{num}}.$$

The declared winner lower endpoint can shift down by at most  $\varepsilon_{\text{opt},j^*,t} + b_t^{\text{impl}}$ , and competitor upper endpoints can shift up by at most  $\max_k (\varepsilon_{\text{opt},k,t} + b_t^{\text{impl}})$ . Hence a  $2(\max_j \varepsilon_{\text{opt},j,t} + b_t^{\text{impl}})$  overturn budget is sufficient. If it exceeds the available certified margin  $(\varepsilon + \tau_{\text{num}} + \text{Gap}_t^{\text{top}})_+$ , uniqueness is not certifiable and abstention is mandatory.  $\square$

**Definition 12.2** (Soft-abstention computational-slack outputs). When unique declaration is uncertified, the protocol outputs a pair

$$(\pi_t^{\text{soft}}, b_t^{\text{soft}}), \quad \pi_t^{\text{soft}} = (\pi_{j,t}^{\text{soft}})_{j \in \mathcal{J}} \in [0, 1]^{|\mathcal{J}|}, \quad b_t^{\text{soft}} = (b_{j,t}^{\text{soft}})_{j \in \mathcal{J}} \in \{R0, \dots, R4\}^{|\mathcal{J}|},$$

with

$$\pi_{j,t}^{\text{soft}} := \left[ \frac{\widehat{L}_{j,t} - \max_{k \neq j} \widehat{U}_{k,t} + \varepsilon + \tau_{\text{num}}}{2 \max\{\epsilon_{\min}, \max_i \varepsilon_{\text{opt},i,t} + b_t^{\text{impl}}\} + \varepsilon + \tau_{\text{num}}} \right]_0^1,$$

where  $[x]_0^1 := \min\{1, \max\{0, x\}\}$ , with replay-published  $\epsilon_{\min} > 0$ . Bands are computed by a replay-published quantizer  $Q_{\text{soft}}$ :

$$b_{j,t}^{\text{soft}} := Q_{\text{soft}}(\pi_{j,t}^{\text{soft}}), \quad Q_{\text{soft}} : [0, 1] \rightarrow \{R0, \dots, R4\},$$

with explicit cutpoints in the rulebook.

$\pi_{j,t}^{\text{soft}}$  is a *machine-only computational-slack index*: it quantifies certified endpoint separation relative to the numerical-overturn budget. It is **not** a posterior probability and **not** a statistical confidence level. Operator dashboards must expose only  $b_t^{\text{soft}}$  and rank order, not raw  $\pi_t^{\text{soft}}$  by default.

**Proposition 12.3** (Semantic firebreak for soft outputs). *Soft outputs are admissible only for (i) ranking/triage, (ii) probe-allocation priority, and (iii) operator diagnostics. They are prohibited as direct threshold variables for hard actuation, hard risk-limit switching, or uniqueness declaration. The control-plane schema must exclude  $\pi_t^{\text{soft}}$  and  $b_t^{\text{soft}}$  from gate predicates; hard transitions must depend only on certified intervals  $(\widehat{L}_{j,t}, \widehat{U}_{j,t})$ , declared margins, and explicit gate predicates. If manual override is enabled, it requires an explicit dual-signature `manual_override` event with reason code and replay trace.*

*Proof.* By construction,  $\pi_t^{\text{soft}}$  and  $b_t^{\text{soft}}$  are uncalibrated geometry summaries. Type-level separation between diagnostics and control payloads yields non-interference for automatic actuation. Therefore semantic misinterpretation in the diagnostic channel cannot silently alter hard control logic.  $\square$

**Proposition 12.4** (Human-factors firewall for soft outputs). *Operational dashboards must render only ordinal bands  $\{R0, \dots, R4\}$ , rank order, and textual cautions. Real-time numeric display of  $\pi_t^{\text{soft}}$  is disabled by default and can be unlocked only in forensic mode with dual authorization and delayed reveal. Manual override decisions must log an explicit declaration that soft outputs are non-probabilistic diagnostics.*

*Proof.* Removing real-time numeric exposure suppresses default probability-style anchoring. Dual-authorization and delayed reveal prevent reflexive thresholding on  $\pi_t^{\text{soft}}$  in fast loops. Mandatory declaration and replay logging convert semantic misuse into a detectable policy violation rather than a silent drift of meaning.  $\square$

**Proposition 12.5** (Soft abstention preserves soundness while exposing computational slack). *Suppose the compute-certification gate triggers or  $|\hat{\mathcal{B}}_t^\varepsilon| \neq 1$ . Then emitting  $(\pi_t^{\text{soft}}, b_t^{\text{soft}})$  with no unique claim cannot increase misdeclaration risk relative to hard abstention, and yields an auditable ranking signal whenever interval geometry is informative. Moreover, if a unique hard declaration is certified, then  $\pi_{\hat{j}_t, t}^{\text{soft}} = 1$ ,  $b_{\hat{j}_t, t}^{\text{soft}} = \text{R4}$ , and  $\pi_{k, t}^{\text{soft}} < 1$  for all  $k \neq \hat{j}_t$ .*

*Proof.* No unique claim is made in soft-abstention mode, so false-unique declarations are impossible by construction. The map in Definition 12.2 is deterministic and monotone in certified endpoint separation, and  $Q_{\text{soft}}$  is replay-published. Under hard-certifiable uniqueness, the certified gap exceeds the overturn budget, making the winner score saturate at 1 and competitors strictly below 1, hence winner band R4.  $\square$

**Proposition 12.6** (No-paralysis under compute limitation). *Under Assumption 5.20, for every history  $h_t$ , a replay-computable fallback action exists and can be executed within a fixed one-step budget. Therefore the protocol does not require idle waiting for global optimization completion.*

*Proof.* Assumption 5.20 guarantees  $(r_t^{\text{safe}}(h_t), 0) \in \mathcal{U}_t^{\text{probe}}(h_t)$  for every  $t, h_t$ , and the shielded action  $r_t^{\text{safe}}(h_t) = \Pi_t^{\text{sh}}(\kappa_t^{\text{bk}}(h_t))$  has replay-published one-step complexity bounds. Hence actuation never waits for completion of heavy bottleneck inference.  $\square$

**Proposition 12.7** (Non-circular active backup safety). *Let  $\mathfrak{S}_t^{\text{core}}$  be the replay-published shield-core envelope in Assumption 5.20 and let  $r_t^{\text{safe}}(h_t) = \Pi_t^{\text{sh}}(\kappa_t^{\text{bk}}(h_t))$ . If the shield-core certificate is valid, then executing  $r_t^{\text{safe}}(h_t)$  preserves the controlled-invariant core set  $\mathcal{X}_t^{\text{core}}$  without requiring completion of online global optimization over  $\Theta_t^{\text{op}}$ . Hence fallback safety is non-circular with respect to diagnostic inference.*

*Proof.* Assumption 5.20 publishes  $\Pi_t^{\text{sh}}$ ,  $\kappa_t^{\text{bk}}$ , and shield-core certificates independently of the online global solve. By definition of the shield projection and controlled invariance of  $\mathcal{X}_t^{\text{core}}$ , executed backup actions remain admissible and safe whenever the shield-core certificate holds. Therefore backup safety does not depend on successful completion of the heavy diagnostic optimization.  $\square$

**Proposition 12.8** (Constructive shield existence or certified graceful degradation). *Under Assumption 5.20, at every time  $t$  exactly one replay-verifiable backup mode is active:*

- (m1) *nontrivial-core mode with  $\nu_t^{\text{core}} \geq \nu_{\min}$  and invariant-set preservation in  $\mathcal{X}_t^{\text{core}}$ , or*
- (m2) *graceful-degradation mode with certified parking set  $\mathcal{X}_t^{\text{park}}$ , horizon  $H_{\text{park}}$ , excursion cap  $E_{\text{park}, t}$ , and service floor  $S_{\min, t}$ .*

*Hence fallback control is never undefined; absence of a nontrivial core does not imply implicit crash, but a bounded and auditable degradation contract.*

*Proof.* Assumption 5.20 requires publication of either the nontrivial-core certificate or the graceful-degradation certificate. Both certificates include executable controllers and admissibility checks independent of online global optimization. Therefore the fallback branch is always well-defined and replay-auditable, with bounded behavior in graceful-degradation mode.  $\square$

**Proposition 12.9** (Acceptability-closed graceful degradation). *Assume Assumption 5.20 with items (bk2)–(bk3). During graceful-degradation mode, either*

- (a1) *the bounded-loss contract and all acceptability caps remain satisfied on  $[t, t + H_{\text{park}}]$ , or*
- (a2) *an acceptability sentinel is triggered and escalation to a stricter backup mode occurs within  $L_{\text{acc}}$ , with explicit alarm and replay trace.*

*Thus “acceptable loss” is not an ungrounded scalar promise: it is tied to observable harm caps and bounded mitigation latency.*

*Proof.* Item (bk3) provides replay-verifiable caps and response-latency requirement for each degradation window. If caps are forecast to hold, case (a1) is immediate. If not, the sentinel clause enforces bounded-time escalation, giving case (a2). No third silent branch exists under Assumption 5.20.  $\square$

**Proposition 12.10** (Two-tier backup prevents freeze and controls coverage-control tension). *Assume Assumption 5.20 and that the shield core set  $\mathcal{X}_t^{\text{core}}$  is nonempty and controlled-invariant under  $\Pi_t^{\text{sh}}$ . Then:*

- (i) *if performance-backup certification on  $\mathfrak{S}_t^{\text{perf}}$  is available, execution uses  $\Pi_t^{\text{sh}}(\kappa_t^{\text{bk}}(h_t))$  and inherits both performance and shield guarantees;*
- (ii) *if performance-backup certification is unavailable, shield-only execution still preserves core invariance;*
- (iii) *if shield feasibility fails, the protocol emits explicit emergency-degradation alarm with violation-budget certificate (no silent execution).*

*Thus fail-closed behavior is active degraded control, not static freezing.*

*Proof.* (i) follows from composition of certified performance backup with shield projection. (ii) follows from shield invariance alone. (iii) is mandated by Assumption 5.20, so infeasibility is surfaced as an explicit alarm and budgeted degradation.  $\square$

**Proposition 12.11** (Anytime branch-and-bound with monotone certificates). *For each channel  $j$ , let branch-and-bound return monotone certified bounds after  $m$  node expansions:*

$$L_{j,t}(m) \uparrow, \quad U_{j,t}(m) \downarrow, \quad r_{j,t}(m) := U_{j,t}(m) - L_{j,t}(m).$$

*Define*

$$\text{Gap}_t^{\text{top}}(m) := \max_j L_{j,t}(m) - \max_{k \neq j^*(m)} U_{k,t}(m), \quad j^*(m) \in \arg \max_j L_{j,t}(m).$$

*If for some  $m$ ,*

$$\text{Gap}_t^{\text{top}}(m) \geq \varepsilon + \tau_{\text{num}} + 2 \left( \max_j r_{j,t}(m) + b_t^{\text{impl}} \right),$$

*then a unique declaration is already certifiable at budget  $m$  without solving the global problem to completion.*

*Proof.* The unresolved numerical uncertainty at budget  $m$  is bounded by  $\max_j r_{j,t}(m)$  per endpoint plus  $b_t^{\text{impl}}$ . If the certified top-gap exceeds this overturn budget and decision margin, uniqueness cannot be reversed by any refinement.  $\square$

**Proposition 12.12** (Adaptive catalog refinement for practical liveness). *Assume Assumption 11.7 and a replay-published nested catalog*

$$\mathcal{C}_t^{(0)} \subseteq \mathcal{C}_t^{(1)} \subseteq \dots, \quad \varepsilon_{\text{pol},t}^{(m+1)} \leq \varepsilon_{\text{pol},t}^{(m)}.$$

*If true robust margin at time  $t$  satisfies*

$$\text{Gap}_t^\star > \varepsilon + \tau_{\text{num}} + 2 \left( \varepsilon_{\text{pol},t}^{(m)} + \max_j \varepsilon_{\text{opt},j,t}^{(m)} + l_t^{\text{impl}} \right)$$

*for some finite level  $m$ , then the protocol certifies a unique declaration at that level and need not reach asymptotic refinement.*

*Proof.* At level  $m$ , policy approximation and optimization errors are explicitly bounded by published envelopes. If the true margin dominates the total envelope as displayed, certified intervals are strictly separated and uniqueness follows.  $\square$

**Proposition 12.13** (Replay-estimable lower bound on hard-certification rate). *Fix a window length  $W \geq 1$  and define*

$$Y_s := \mathbf{1}\{\text{hard singleton declaration emitted at step } s\}, \quad s = t - W + 1, \dots, t.$$

*Let  $\mathcal{F}_s^{\text{pub}} := \sigma(H_s, Q_{0:s}, A_{0:s})$ , and define the predictable window-average hard-certification rate*

$$\bar{p}_{t,W}^{\text{hard}} := \frac{1}{W} \sum_{s=t-W+1}^t \mathbb{E}[Y_s \mid \mathcal{F}_{s-1}^{\text{pub}}].$$

*For any  $\beta_{\text{hard}} \in (0, 1)$ , let*

$$\underline{p}_{t,W}^{\text{hard}} := \text{LCB}_{\text{CS}}(Y_{t-W+1:t}, \beta_{\text{hard}}; \mathbf{m}_t^{\text{CS}})$$

*be the replay-deterministic one-sided lower endpoint returned by a published nonnegative-supermartingale confidence-sequence routine (predictable-mixture or stitching profile in metadata  $\mathbf{m}_t^{\text{CS}}$ ). Then*

$$\mathbb{P}(\bar{p}_{t,W}^{\text{hard}} \geq \underline{p}_{t,W}^{\text{hard}}) \geq 1 - \beta_{\text{hard}}.$$

*If the replay bundle additionally certifies exchangeability/i.i.d. within the same window, an optional exact Clopper–Pearson lower bound may be reported as a supplementary metric [22]. Therefore liveness claims must be stated in terms of  $\underline{p}_{t,W}^{\text{hard}}$ , not point estimates.*

*Proof.* For bounded  $Y_s \in [0, 1]$ , the published  $\text{LCB}_{\text{CS}}$  routine is constructed from a nonnegative supermartingale family that is valid under optional stopping; by Ville’s inequality, its confidence-sequence coverage is at least  $1 - \beta_{\text{hard}}$  [20, 21]. Evaluating that routine on the fixed window  $Y_{t-W+1:t}$  yields the displayed one-sided lower bound for  $\bar{p}_{t,W}^{\text{hard}}$ . The Clopper–Pearson statement is conditional on the stronger exchangeability/i.i.d. certificate.  $\square$

## Operational hardening addenda: proxy–latent link and bridge sentinels

**Assumption 12.14** (Two-layer monitor for open-world residual linkage). The open-world residual term  $r_t^{\text{ow}}$  is governed by a two-layer monitor: a proxy misspecification monitor  $M_t^{\text{mis}}$  and an independent linkage monitor  $M_t^{\text{link}}$  that tests whether proxy evidence remains informative for latent-side risk. If

$M_t^{\text{link}}$  crosses threshold  $1/\alpha_{\text{link}}$ , the protocol must apply deterministic residual inflation

$$r_t^{\text{ow}} \leftarrow \min\left\{r_{\text{max},t}^{\text{exo}}, r_t^{\text{ow}} + u_t^{\text{link}}\right\},$$

publish alarm label `link_broken`, and tighten to conservative mode.

**Assumption 12.15** (Structure-constant violation sentinel for IQC energy-to-peak bridge). When the energy-to-peak bridge is active, replay logs must publish a structure-residual statistic  $R_t^{\text{str}}$  that checks consistency of the deployed envelopes  $(\bar{L}_{g,t,u}, \bar{L}_{\chi,t,u}, \bar{c}_{z,t,u})$  against observed increments. If  $R_t^{\text{str}} > \tau_{\text{str}}$ , the bridge is invalidated for that window:

$$\text{iqc\_bridge\_mode} \leftarrow \text{none},$$

and only direct certified envelopes are allowed for hard guarantees.

**Assumption 12.16** (Detection-lag buffer and preemptive reserve guard with shock inflation). Let  $\sigma_t$  denote a replay-computable safety reserve (distance-to-boundary surrogate for the active certified set), and let  $g_t^{\text{max}}$  be a replay-published baseline upper bound on one-step reserve consumption under current actuation limits. Let  $L_{\text{det}}$  be the certified worst-case confirmation lag of linkage/structure alarms, and let  $\varphi_t^{\text{lag}}(u)$  be a replay-published upper bound on alarm delay tails:

$$\mathbb{P}(T_{\text{alarm}} - t > u \mid \mathcal{G}_t) \leq \varphi_t^{\text{lag}}(u), \quad u \in \mathbb{N}.$$

A finite-time power floor must also be published: there exists  $u_{\text{pow},t}$  such that

$$\varphi_t^{\text{lag}}(u_{\text{pow},t}) \leq \alpha_{\text{pow},t}^{\text{exo}},$$

with  $(u_{\text{pow},t}, \alpha_{\text{pow},t}^{\text{exo}})$  included in the signed governance block. Define a replay-computable jump sentinel  $J_t \in \{0, 1\}$  from high-frequency residuals and a shock multiplier  $\Gamma_t \geq 1$ :

$$\tilde{g}_t^{\text{max}} := \Gamma_t g_t^{\text{max}}, \quad \Gamma_t = \begin{cases} \Gamma_{\text{shock},t}^{\text{exo}} & \text{if } J_t = 1, \\ 1 & \text{otherwise.} \end{cases}$$

The controller must trigger a preemptive safety mode whenever

$$\sigma_t \leq (L_{\text{det}} + 1) \tilde{g}_t^{\text{max}} + \zeta_t,$$

where  $\zeta_t \geq 0$  is a replay-published implementation cushion. When  $J_t = 1$ , hard-certification is temporarily disabled until recertification clears the shock flag.

**Proposition 12.17** (Latency-compensated conservatism under monitor lag). *Under Assumptions 12.14–12.16, linkage failure or structure-constant violation cannot silently produce over-optimistic hard declarations. Before confirmed alarms, the reserve guard in Assumption 12.16 forces preemptive tightening with shock-inflated consumption  $\tilde{g}_t^{\text{max}}$ . After alarm confirmation, Assumptions 12.14–12.15 further tighten by residual inflation or bridge deactivation. Consequently, hard-certification predicates are monotone-tightened through the latency window, including jump-sentinel phases.*

*Proof.* Assumption 12.16 enforces an early guard based on worst-case reserve consumption, certified alarm lag, and shock multiplier  $\Gamma_t$ . Assumption 12.14 enforces monotone non-decreasing residual cushions after linkage alarm. Assumption 12.15 removes potentially invalid bridge reductions after structure violation. When  $J_t = 1$ , temporary hard-certification disablement prevents exploiting



unmodeled phase-transition windows. Each transition only tightens (never loosens) hard-certification conditions, so over-optimistic declarations are excluded throughout pre- and post-alarm phases.  $\square$

**Proposition 12.18** (Budget-level lower bound for hard certification). *Fix a replay budget level  $m$ , and define total certified uncertainty envelope*

$$E_t(m) := \varepsilon_{\text{pol},t}^{(m)} + \max_j \varepsilon_{\text{opt},j,t}^{(m)} + b_t^{\text{impl}}.$$

*Let  $\Gamma_t^*$  be the true top-margin at time  $t$ . If*

$$\mathbb{P}(\Gamma_t^* > \varepsilon + \tau_{\text{num}} + 2E_t(m)) \geq 1 - \beta_m,$$

*then*

$$\mathbb{P}(\text{hard certification is achieved by budget } m) \geq 1 - \beta_m.$$

*Hence empirical lower bounds on margin exceedance directly translate into lower bounds on hard-certification frequency.*

*Proof.* Whenever  $\Gamma_t^* > \varepsilon + \tau_{\text{num}} + 2E_t(m)$ , certified intervals are strictly separated at level  $m$ , so hard uniqueness is certifiable by the same envelope argument as Proposition 12.12. Taking probability on both sides yields the claim.  $\square$

**Deterministic replay contract.** Each published decision must include a hash-linked replay bundle

$$\mathbf{r}_t := \left\{ \begin{array}{l} H_t, \text{ rulebook hash, solver hash,} \\ \text{numeric profile, seed, threading mode,} \\ \text{certificate payload} \end{array} \right\}.$$

and a deterministic verifier that recomputes  $\Theta_t^{\text{op}}$ ,  $\mathcal{I}_{j,t}$ , and  $\widehat{\mathcal{B}}_t^\varepsilon$  up to declared tolerances. This makes disagreement falsifiable at the artifact level, not at narrative level.

## 13 Auditability guarantees

**Theorem 13.1** (Public recomputability). *If logs publish  $H_t$ , rulebook, gate parameters, e-process definitions, budget tuple  $(\alpha_{\text{out}}, \alpha_{\text{ev}}, \alpha_{\text{dep}}, \alpha_{\text{env}}, \alpha_{\text{mis}})$ , optimization certificates  $(L_{j,t}^{\text{inf}}, U_{j,t}^{\text{sup}}, \varepsilon_{\text{opt},j,t})_{j \in \mathcal{J}}$ , contamination/dependence/interaction envelopes  $(\bar{\eta}_t^{\text{pub}}, n_{\text{eff},t}^{\text{lb}}, \bar{\sigma}_{S,t}^{\text{pub}}, b_t^{\text{int,pub}})$ , numerical envelope  $\varepsilon_{E,t}^{\text{num}}$ , dynamic-IQC design  $(\Psi_\ell, M)$ , and deterministic replay metadata*

$$\mathbf{m}_t := \left\{ \begin{array}{l} \text{compiler/solver versions, BLAS/LAPACK backend,} \\ \text{floating-point mode, rounding mode, seeds, stopping rules,} \\ \text{thread count, canonical serialization profile, hash algorithm/domain tags} \end{array} \right\}.$$

*then all outputs*

$$\Theta_t^{\text{out}}, \Theta_t^{\text{in}}, \widehat{L}_{j,t}, \widehat{U}_{j,t}, \widehat{\mathcal{B}}_t^\varepsilon$$

*are externally reproducible up to declared numerical tolerances.*

*Proof.* Each output is a deterministic functional of public inputs and metadata  $\mathbf{m}_t$ . Fixing  $\mathbf{m}_t$  removes implementation degrees of freedom in linear algebra kernels and stopping logic. Hence recomputation

reproduces the same optimization brackets  $(L_{j,t}^{\inf}, U_{j,t}^{\sup})_{j \in \mathcal{J}}$  and therefore identical declarations up to certified tolerances.  $\square$

**Theorem 13.2** (Tiered verification with finite-lag detectability). *Let the audit plan include three replay-public tiers:*

- (v1) micro-check each step: schema, signatures, hash-chain, gate inequalities;
- (v2) meso-check randomized spot recomputation with per-step inclusion probability  $q_t \in (0, 1]$  drawn from a public beacon;
- (v3) macro-check full deterministic replay every  $K_{\text{full}}$  steps.

*If at least  $r$  manipulated steps occur between two macro checks and each manipulated step can be detected by meso recomputation when sampled, then*

$$\mathbb{P}(\text{no detection before next macro check}) \leq \prod_{i=1}^r (1 - q_{t_i}) \leq (1 - q_{\min})^r.$$

*In addition, each tier has explicit semantic obligations: micro-check validates syntax/integrity invariants only; meso-check validates step-local semantic invariants on sampled steps; macro-check validates full trajectory semantics. Hence detection delay is almost surely finite, bounded deterministically by  $K_{\text{full}}$  and probabilistically sharpened by meso sampling.*

*Proof.* Micro checks guarantee low-cost syntactic validity each step but not full semantic replay. For each manipulated step  $t_i$ , meso sampling catches manipulation with probability at least  $q_{t_i}$ , independent of attacker choice conditional on beacon unpredictability. Failure of all  $r$  spot checks yields product bound above. Even if meso checks miss, the next macro full replay detects any replay-inconsistent manipulation, giving deterministic finite-lag detection.  $\square$

## 14 Counterexamples and resolved failure modes

**Counterexample 14.1** (Aliasing despite rich dynamics). Different latent decompositions over  $(I, M, P)$  can generate identical public histories under all  $\Pi_{\text{NM}}$  policies. Resolved by interval diagnosis and abstention logic, not by forced point estimates.

**Counterexample 14.2** (Naive endpoint aggregation inconsistency). If endpoints are built by separately summing per-term bounds, dependence can produce impossible intervals. Resolved by direct optimization of full score  $S_j^\theta$ .

**Counterexample 14.3** (Naive inner-set update non-monotonicity). A one-shot erosion each time does not ensure temporal monotonicity. Resolved by recursive inner update in Theorem 8.6.

**Counterexample 14.4** (Repeated fixed-time testing inflation). Applying fixed-time confidence repeatedly inflates false alarms. Resolved by anytime-valid e-process construction and Corollary 9.7.

## 15 Falsifiable predictions

1. **Dynamic vs static IQC feasibility gap.** Define  $Y_n = \mathbf{1}\{F_n^{\text{dyn}} = 1, F_n^{\text{stat}} = 0\}$  on benchmark instance  $n$ . Primary target:  $\Delta_{\text{feas}} := \mathbb{E}[Y_n] > 0$ . Rejective criterion: one-sided lower confidence bound for  $\Delta_{\text{feas}}$  exceeds 0.

2. **Anytime misdeclaration control.** Track  $Z_t = \mathbf{1}\{\text{misdeclaration at } t\}$  with  $\mathcal{M}$  from Corollary 9.7. Primary target:  $\mathbb{P}(\mathcal{M}) \leq \alpha_{\text{tot}}$ . Operational test: nonasymptotic upper confidence bound for cumulative violation indicator remains below  $\alpha_{\text{tot}} + \varepsilon_{\text{audit}}$ .
3. **Interaction lead signal.** Define  $\Delta_{IMP,t}$  and first time of certified unique dominance  $T_{\text{dom}}$ . Prediction: in strongly coupled regimes, significantly positive  $\Delta_{IMP,t}$  appears before  $T_{\text{dom}}$  with positive lead  $\ell$ , tested by lagged association statistics with pre-registered threshold.
4. **Inner/outer geometry.** Let  $V_t^{\text{in}}$  be inner-core volume surrogate and  $C_t^{\text{out}}$  be outer coverage indicator. Prediction:  $V_t^{\text{in}}$  exhibits negative trend while empirical coverage remains compatible with  $1 - \alpha_{\text{out}}$ .
5. **Abstention+probe value gain.** Define per-episode robust regret difference  $\Delta_{\text{AP}} := W^{\text{AP}} - W^{\text{forced}}$ . Prediction: under observational aliasing episodes ( $|\hat{\mathcal{B}}_t^\varepsilon| \neq 1$ ), lower confidence bound of  $\mathbb{E}[\Delta_{\text{AP}}]$  is strictly positive.

**Pre-registered measurement design (minimum).** For item 1, estimate  $\Delta_{\text{feas}}$  with one-sided time-uniform confidence sequences (nonnegative-supermartingale construction) and report the smallest  $n$  such that the lower endpoint exceeds 0 at level  $1 - \beta_{\text{feas}}$ ; optional exact binomial bounds may be added only with an explicit i.i.d./exchangeability certificate. For item 2, use the anytime event  $\mathcal{M}$  and report an e-value or confidence sequence that is valid under optional stopping. For item 3, pre-register lag window  $\ell$ , statistic family, and permutation/bootstrap calibration under the null of no lead. For item 4, report both  $V_t^{\text{in}}$  trend statistic and empirical non-coverage counts against budget  $\alpha_{\text{out}}$ . For item 5, pre-register ambiguity episodes and evaluate  $\mathbb{E}[\Delta_{\text{AP}}]$  with one-sided confidence sequences.

## 16 Limitations

- Dynamic IQC remains sufficient, not necessary; conservative certificates can trade liveness for safety.
- E-process quality depends on test-martingale design and calibration diagnostics.
- Inner erosion can be conservative in poorly scaled model metrics.
- Raw-set optimization is generally intractable; convex/SDP/branch-and-bound relaxations introduce certified but potentially large gaps. When `raw_relax_gap` is large, compute-limited abstention can dominate.
- Observable monitors are only proxies for latent properties. The decomposed link cushion  $b_t^{\text{link, pub}} = \hat{b}_t^{\text{link}} + u_t^{\text{link}} + r_t^{\text{ow}}$  makes this mismatch explicit, but cannot eliminate delayed or strategic monitor deception.
- The energy-to-peak IQC bridge depends on replay-published observable slope envelopes. If these envelopes drift faster than recertification, tightening may become unavailable and the protocol must revert to baseline/fallback.
- Observational alias is handled as non-rejection, not equivalence. Hazard-envelope penalties reduce risk but cannot eliminate sleeper-trigger behavior that is fundamentally unidentifiable from current observables.

- Fail-closed branches use an active low-order backup controller instead of static freezing, but this still prioritizes safety over performance and may be conservative in fast transients.
- Soft abstention improves liveness but does not prove latent correctness. The semantic firebreak ( $\pi_t^{\text{soft}}$  machine-only,  $b_t^{\text{soft}}$  operator-facing) reduces probability-style misuse, but cannot remove risks from explicit human manual overrides.
- Endogenous  $\lambda_{h,t}$  is a nonconvex-safe constraint-tracking mechanism, not an optimality oracle. Hyperparameters ( $\bar{h}_t, \lambda_{\max,t}, \eta_{\lambda,t}, \delta_h, K_\lambda$ ) must be injected via the public exogenous governance channel and stress-tested with anti-chattering diagnostics.
- Constructive executability relies on the quality of the finite-catalog approximation envelope  $\varepsilon_{\text{pol},t}$ ; if too loose, policy-level claims become conservative.
- All guarantees are model-class-conditional and do not claim access to “true” latent reality; unknown unknowns are handled via misspecification alarms, recertification, and fallback.
- Strategic adversaries reacting to probes require explicit game-theoretic extensions.
- Grounded diversity checks reduce descriptor blind spots, but challenge suites can still miss rare shared failure modes; periodic adversarial test-suite refresh is required.
- Signed contingency ladders reduce governance outages but can still reduce feasible action volume in prolonged crises; anti-suffocation guards and graceful-degradation escalation are required to avoid self-denial of service.
- Tiered verification improves scalability, yet full replay remains expensive; practical assurance depends on maintaining adequate macro-check cadence and unbiased meso sampling.
- Quantities not identifiable from observables are treated as explicit exogenous governance inputs (hash-linked, versioned, replay-visible). The framework excludes hidden internal meta-constants but does not claim self-identification of external normative budgets.

## 17 Future robust extension: multi-agent no-meta consensus

A natural extension is multi-agent, multi-view certification without hidden overrides. Let  $\{\mathbf{A}_m\}_{m=1}^M$  be no-meta agents running the same public rulebook with diversified sensing/probe families, model classes, compilers, and randomness beacons.

Each agent emits

$$\left(\widehat{\mathcal{B}}_{t,m}^\varepsilon, \text{hard-certificate hash}, b_{t,m}^{\text{soft}}, \mathbf{d}_{m,t}^{\text{str}}, \mathbf{p}_{m,t}^{\text{prov}}, \mathbf{z}_{m,t}^{\text{beh}}\right),$$

where  $\mathbf{d}_{m,t}^{\text{str}}$  is a structural descriptor,  $\mathbf{p}_{m,t}^{\text{prov}}$  is a provenance descriptor (data lineage / training-source commitments), and  $\mathbf{z}_{m,t}^{\text{beh}}$  is a behavioral fingerprint on blind challenge suites.

**Assumption 17.1** (Grounded diversity certificates with anti-gaming challenge design). At each consensus window, the rulebook publishes a two-part challenge protocol:

$$\mathcal{Q}_t^{\text{pub}} \cup \mathcal{Q}_t^{\text{sec}},$$

where  $\mathcal{Q}_t^{\text{pub}}$  is public and  $\mathcal{Q}_t^{\text{sec}}$  is commit-then-reveal (hash committed before decision, revealed after decision commit). Minimum effective coverage  $N_{\min}$  applies to the union. Each agent must provide:

- (d1) structural descriptor commitment  $\mathbf{d}_{m,t}^{\text{str}}$ ,
- (d2) provenance commitment  $\mathbf{p}_{m,t}^{\text{prov}}$ ,
- (d3) behavioral error signatures on both suites

$$\mathbf{z}_{m,t}^{\text{pub}} := (\mathbf{1}\{\text{error on } q\})_{q \in \mathcal{Q}_t^{\text{pub}}}, \quad \mathbf{z}_{m,t}^{\text{sec}} := (\mathbf{1}\{\text{error on } q\})_{q \in \mathcal{Q}_t^{\text{sec}}}.$$

Define per-agent gaming gap

$$\Delta_{m,t}^{\text{game}} := |\bar{e}_{m,t}^{\text{pub}} - \bar{e}_{m,t}^{\text{sec}}|,$$

where  $\bar{e}$  is mean error rate on the respective suite. If challenge coverage  $< N_{\min}$  or  $\max_m \Delta_{m,t}^{\text{game}} > \gamma_{\max}$ , hard consensus is disabled for that window.

Define

$$\begin{aligned} \text{BehCorr}_t &:= \max \left\{ \max_{m \neq m'} \text{Corr}(\mathbf{z}_{m,t}^{\text{pub}}, \mathbf{z}_{m',t}^{\text{pub}}), \max_{m \neq m'} \text{Corr}(\mathbf{z}_{m,t}^{\text{sec}}, \mathbf{z}_{m',t}^{\text{sec}}) \right\}, \\ \text{GameGap}_t &:= \max_m \Delta_{m,t}^{\text{game}}, \\ \text{ProvOverlap}_t &:= \text{Overlap}(\mathbf{p}_{1,t}^{\text{prov}}, \dots, \mathbf{p}_{M,t}^{\text{prov}}), \end{aligned}$$

and structural overlap

$$\text{StrOverlap}_t := \text{Overlap}(\mathbf{d}_{1,t}^{\text{str}}, \dots, \mathbf{d}_{M,t}^{\text{str}}).$$

Hard consensus on label  $\hat{j}$  is permitted only if: (i) at least  $q$  agents output singleton hard declarations equal to  $\hat{j}$ , (ii) certificate-distance checks pass, and (iii) common-mode sentinel

$$C_t^{\text{cm}} := \mathbf{1} \left\{ \begin{array}{l} \lambda_{\max}(\hat{R}_t^{\text{res}}) > \rho_{\max} \vee \text{StrOverlap}_t > \omega_{\max} \\ \vee \text{BehCorr}_t > \kappa_{\max} \vee \text{ProvOverlap}_t > \xi_{\max} \\ \vee \text{GameGap}_t > \gamma_{\max} \end{array} \right\} = 0.$$

If  $C_t^{\text{cm}} = 1$  or agreement fails, the protocol blocks hard consensus, retains soft abstention, and executes active backup.

A further extension is polycentric governance for exogenous parameters: multiple rulebook signers (institutions, auditors, cryptoeconomic committees, or formal councils) publish  $g_t^{\text{exo}}$  updates with threshold signatures, multi-path dissemination, and signed contingency ladders. This keeps meta decisions explicit, contestable, and replay-auditable while reducing governance single-point failures.

**Proposition 17.2** (False-consensus bounds under grounded diversity caps). *Suppose hard consensus on label  $\hat{j}$  requires at least  $q$  agreeing agents among  $M$ . Let  $I_{m,t}^{\text{fh}} \in \{0, 1\}$  indicate “agent  $m$  outputs a false hard declaration for  $\hat{j}$ ” and*

$$S_t^{\text{fh}} := \sum_{m=1}^M I_{m,t}^{\text{fh}}.$$

*Assume the certified common-mode sentinel satisfies  $C_t^{\text{cm}} = 0$ , and define the replay-certified marginal envelope*

$$\bar{p}_t^{\text{fh}} := \min\{1, p_t^{\text{fh}} + \rho_{\max} + \omega_{\max} + \kappa_{\max} + \xi_{\max} + \gamma_{\max}\},$$

*where  $p_t^{\text{fh}}$  is an upper bound on the per-agent marginal false-hard probability under current certificates. Then:*

(i) (**Dependence-robust universal bound**)

$$\mathbb{P}(\text{false consensus at } t) = \mathbb{P}(S_t^{\text{fh}} \geq q) \leq \min\left\{1, \frac{M \bar{p}_t^{\text{fh}}}{q}\right\}.$$

This requires no independence assumption.

(ii) (**Sharper optional bound under replay-certified exchangeable overdispersion**) if, in addition,  $S_t^{\text{fh}}$  is replay-certified to follow a beta-binomial envelope  $\text{BB}(M, \alpha_t, \beta_t)$  with  $\alpha_t, \beta_t > 0$ , then

$$\mathbb{P}(\text{false consensus at } t) \leq \sum_{r=q}^M \binom{M}{r} \frac{B(r + \alpha_t, M - r + \beta_t)}{B(\alpha_t, \beta_t)}.$$

Hence dependence reduction and overlap controls tighten either bound through  $\bar{p}_t^{\text{fh}}$  and, when available, through smaller overdispersion.

*Proof.* For (i),  $S_t^{\text{fh}} \geq 0$  and  $\mathbb{E}[S_t^{\text{fh}}] = \sum_{m=1}^M \mathbb{P}(I_{m,t}^{\text{fh}} = 1) \leq M \bar{p}_t^{\text{fh}}$ . Applying Markov to  $S_t^{\text{fh}}$  yields  $\mathbb{P}(S_t^{\text{fh}} \geq q) \leq \mathbb{E}[S_t^{\text{fh}}]/q$ , then cap at 1. For (ii), the beta-binomial tail is the exact mixture tail under the certified envelope; using it as an upper envelope gives the displayed bound.  $\square$

## 18 Conclusion

Under observable-only no-meta constraints, exact latent bottleneck attribution is generally impossible. The correct target is robust, auditable, interaction-aware interval diagnosis with principled abstention. The framework combines: robust DP foundations, coherent interval dominance logic, time-consistent outer/inner model-set recursion, anytime-valid multi-time guarantees, dynamic IQC delay robustness, signed contingency governance with anti-suffocation guards, acceptability-closed graceful degradation, anti-gaming grounded diversity, and tiered public verification. Its formal guarantees are guarantees of *accountability and best effort under explicit assumptions*: reproducible decisions, explicit uncertainty, bounded-latency alarms, and certified fallback behavior. They are not guarantees that latent ground truth has been discovered. In short, the system is engineered to surface limits honestly and act conservatively under uncertainty, not to claim epistemic omniscience. It guarantees accountable rule-following under explicit assumptions, not truth itself.

## A Appendix A: Additional technical lemmas

**Lemma A.1** (Contamination bias bound). *Let one-step reward kernel satisfy  $|r_s(h, a, h')| \leq B$ . Under contamination model  $\mathbb{P}^{\text{obs}} = (1 - \eta)\mathbb{P}^{\star} + \eta\mathbb{Q}$ , finite-horizon return bias satisfies*

$$|V_{t,H}^{\text{obs}} - V_{t,H}^{\star}| \leq 2B \sum_{s=0}^{H-1} \gamma^s \eta \leq \frac{2B}{1 - \gamma} \eta.$$

*Proof.* For any bounded measurable  $f$  with  $\|f\|_{\infty} \leq B$ , mixture decomposition gives

$$|\mathbb{E}_{\mathbb{P}^{\text{obs}}}[f] - \mathbb{E}_{\mathbb{P}^{\star}}[f]| = \eta |\mathbb{E}_{\mathbb{Q}}[f] - \mathbb{E}_{\mathbb{P}^{\star}}[f]| \leq 2B\eta.$$

Apply this one-step bound to discounted reward increments and sum geometrically.  $\square$

**Lemma A.2** (Score-level contamination cushion induced by (7)). Assume  $|r_s(h, a, h')| \leq B$  and contamination level  $\eta \leq \bar{\eta}_t^{\text{pub}}$ . Then each robust gain term in the finite-difference stencil satisfies

$$\left| G_{t,H}^{\text{obs}}(h_t; r) - G_{t,H}^*(h_t; r) \right| \leq \frac{4B}{1-\gamma} \eta.$$

For  $|\mathcal{J}| = 3$  and score weights  $(w_2, w_3) \in [0, 1]^2$ , the score error obeys

$$\left| S_j^{\text{obs}} - S_j^* \right| \leq \left( 1 + 2w_2 + \frac{4}{3}w_3 \right) \frac{8B}{(1-\gamma)\delta} \eta \leq \left( 1 + 2w_2 + \frac{4}{3}w_3 \right) \frac{8B}{(1-\gamma)\delta} \bar{\eta}_t^{\text{pub}}.$$

Hence

$$c_{S,\text{cont}}(w_2, w_3) := 8 \left( 1 + 2w_2 + \frac{4}{3}w_3 \right)$$

is a conservative valid constant (default  $w_2 = w_3 = 1 \Rightarrow c_{S,\text{cont}} = 104/3$ ). If weights are time-varying, use

$$c_{S,\text{cont},t} := c_{S,\text{cont}}(w_{2,t}, w_{3,t}),$$

and publish  $b_{t,H}^{\text{cont}} := c_{S,\text{cont},t} \frac{B}{(1-\gamma)\delta} \bar{\eta}_t^{\text{pub}}$  with the same replay index  $t$ .

*Proof.* Each  $G$  is a difference of two values, and each value has contamination bias at most  $2B(1-\gamma)^{-1}\eta$ , giving  $4B(1-\gamma)^{-1}\eta$  for each  $G$ -term. In (7), the first-order part contributes at most  $2 \cdot \frac{4B}{(1-\gamma)\delta}\eta$ , the pair-interaction part contributes at most  $4w_2 \cdot \frac{4B}{(1-\gamma)\delta}\eta$ , and the third-order part contributes at most  $\frac{8}{3}w_3 \cdot \frac{4B}{(1-\gamma)\delta}\eta$ . Summing yields the displayed bound.  $\square$

**Lemma A.3** (Drift perturbation bound). If Bellman operator is  $L_K$ -Lipschitz in one-step kernel TV distance,

$$|V_{t,H}^K - V_{t,H}^{\tilde{K}}| \leq L_K \sum_{s=t}^{t+H-1} \sup_{x,a} \text{TV}(K_s(\cdot|x, a), \tilde{K}_s(\cdot|x, a)).$$

*Proof.* Apply one-step perturbation recursively and telescope across horizon.  $\square$

**Lemma A.4** (Score perturbation). Under Assumption 5.10, for any two nonempty compact model sets  $\Xi_t, \Xi'_t \subseteq \Theta_0$ ,

$$|\underline{S}_{j,t} - \underline{S}'_{j,t}| \leq L_S d_H(\Xi_t, \Xi'_t), \quad |\bar{S}_{j,t} - \bar{S}'_{j,t}| \leq L_S d_H(\Xi_t, \Xi'_t).$$

*Proof.* For any  $\theta \in \Xi_t$ , choose  $\theta' \in \Xi'_t$  within Hausdorff distance. Use Lipschitz continuity of  $S_j$ , then take inf/sup.  $\square$

## B Appendix B: Operational checklist

Item	Public artifact	If violated
No-meta admissibility	rulebook + fallback policy	hidden override risk
Outer recursion	e-process definition + threshold $1/\alpha_{\text{out}}$	no uniform validity
Inner recursion	recursive erosion logs + radius schedule $r_t$	non-monotone robust core
Score intervals	direct optimization certificate $\varepsilon_{\text{opt},j,t}$	incoherent dominance claims
Envelope origin	observable calibration e-process logs, monitor-latent link cushion, or deterministic physical bound source tags	hidden meta-knowledge constants or proxy drift
Rectangularization distortion	raw/rect endpoint brackets + $b_t^{\text{rect},\text{pub}}$ certificate	phantom-model inflation unquantified
Constructive D6	finite catalog $\mathcal{C}_t(h_t)$ , $\varepsilon_{\text{pol},t}$ , tie-break rule, alias-hazard map $h_t^{\text{alias}}(h_t, p)$ , dual $\lambda_{h,t}$ logs	non-executable measurable selector risk or hidden alias risk
Soft abstention output	machine-only $\pi_t^{\text{soft}}$ , operator bands $b_t^{\text{soft}}$ , compute-gate evidence, semantics tag= <code>computational_slack</code> , control-plane soft-field denial	semantic leakage into hard control
Fail-closed backup control	shielded backup stack $(\Pi_t^{\text{sh}}, \kappa_t^{\text{bk}})$ , <code>backup_cert_hash</code> , <code>shield_hash</code> , <code>core_envelope_id</code> , <code>perf_envelope_id</code>	static freeze fallback or uncertified backup execution
Dynamic IQC	FIR multipliers $\Psi_\ell$ , $M$ , LMI tolerance + bridge artifact (reachable tube or energy-to-peak)	peak claims without certificate



## C Appendix C: Notation

Symbol	Meaning
$X_t = (I_t, M_t, P_t, C_t)$	latent macro state
$H_t$	public history
$\Pi_{\text{NM}}$	no-meta admissible policy class
$\Theta_t^{\text{ev}}(h_t)$	evidence-consistent model set
$G_{t,H}^\theta$	robust relief gain (possibly probe-dependent in D6 objective before penalties)
$\beta_j, \Delta_{jk}, \Delta_{\text{IMP}}$	first-/second-/third-order diagnostic quantities
$S_j^\theta$	interaction-aware score
$\Theta_t^{\text{out}}, \Theta_t^{\text{in}}$	outer/inner ambiguity recursions
$E_t(\theta)$	model-indexed e-process
$\hat{L}_{j,t}, \hat{U}_{j,t}$	anytime-valid score interval endpoints
$b_t^\Sigma$	aggregate auditable interval inflation (impl + cont + dep + int + link + rect)
$b_t^{\text{rect, pub}}$	certified endpoint distortion from raw-to-rectangular operational filtering
$h_t^{\text{alias}}(h_t, p)$	probe-dependent worst-case alias hazard penalty
$\lambda_{h,t}$	endogenous hazard dual weight (public risk-budget update)
$\pi_t^{\text{soft}}, b_t^{\text{soft}}$	machine-only slack index and operator-facing ordinal soft bands (both non-probabilistic)
$\hat{\mathcal{B}}_t^\varepsilon$	declared dominant set at margin $\varepsilon$
$\chi_t$	gating variable
$s_t$	IQC augmented state $[\xi_t, \zeta_t]$
$\Psi_\ell, M$	dynamic IQC FIR multiplier parameters
$\mathcal{C}_t(h_t), \varepsilon_{\text{pol},t}$	replay-published constructive action catalog and policy approximation envelope

## D Appendix D: Minimal public certificate schema (informative)

A minimal JSON-like payload for one decision time  $t$  is:

```
{
  "time": t,
  "history_hash": "...",
  "rulebook_hash": "...",
  "alpha": {"out":..., "ev":..., "dep":..., "env":..., "mis":...},
  "theta_op_nonempty": true/false,

  "numerics": {
    "tau_num":...,
    "eps_E_num":...,
    "theta_mesh_radius":...,
    "impl_mode": "grid|mixture",
    "bnb_nodes_used":...,
    "eps_pol_t":...,
    "catalog_level":...,
    "raw_relax_mode": "none|convex|sdp|hybrid",
    "raw_relax_gap":...,
    "floating_point_mode": "..."
  }
}
```

```

},

"envelopes": {
  "eta_pub": ...,
  "n_eff_lb": ...,
  "sigma_pub": ...,
  "b_int_pub": ...,
  "b_link_pub": ...,
  "b_link_components":{"b_link_hat":...,"u_link":...,"r_open_world":...},
  "link_alarm_evalue": ...,
  "mis_alarm_evalue": ...,
  "env_source_mode":"deterministic|calibrated",
  "env_evalues":{"eta":...,"sigma":...,"ninv":...,"int":...},
  "rect_endpoint_bounds":{"
    "U_raw_lb":...,"U_raw_ub":...,"U_rect_lb":...,"U_rect_ub":...,
    "L_raw_lb":...,"L_raw_ub":...,"L_rect_lb":...,"L_rect_ub":...
  },
  "b_rect_pub": ...,
  "iqc_bridge_mode":"none|reachable_tube|energy_peak",
  "structure_residual": ...,
  "tau_str": ...,
  "delta_peak_iqc": ...,
  "delta_impl": ...
},

"identifiability": {
  "delta_theta": ...,
  "covering_number": ...,
  "kappa_id": ...,
  "rho_probe": ...,
  "nu_id": ...,
  "v_id_deprecated_alias": ...,
  "c_id": ...,
  "iglower_t": ...,
  "ambiguity_counter": ...,
  "effective_probe_counter": ...,
  "alias_class_count": ...,
  "alias_class_hash": "...",
  "alias_hazard_t": ...,
  "lambda_h_t": ...,
  "hazard_budget_t": ...
},

"soft_abstention": {
  "enabled":true/false,
  "semantics":"computational_slack",
  "pi_soft_internal":{"I":...,"M":...,"P":...},
  "soft_band":{"I":"R0|R1|R2|R3|R4","M":"R0|R1|R2|R3|R4","P":"R0|R1|R2|R3|R4"},
  "ui_numeric_exposed":false,
  "control_plane_soft_fields_allowed":false,
  "pi_soft_deprecated_alias":{"I":...,"M":...,"P":...}
},

"governance_exogenous": {

```

```

    "g_exo_hash": "...",
    "g_exo_version": "...",
    "block_seq": ...,
    "block_time_unix": ...,
    "signer_set_hash": "...",
    "k_required": ...,
    "n_total": ...,
    "fork_choice_id": "...",
    "transport_path_id": "...",
    "outage_type": "none|connectivity|crypto_invalid",
    "dwell_steps": ...,
    "mode": "fresh|ladder|graceful_degradation|shield_only|emergency_stop",
    "outage_len": ...,
    "H_grace": ...,
    "ladder_hash": "...",
    "ladder_index": ...,
    "phi_feas_lb": ...,
    "phi_min": ...,
    "disturbance_jump_sentinel": 0/1,
    "risk_budget": {"h_bar": ..., "eta_lambda": ..., "lambda_max": ..., "deadzone": ...},
    "open_world": {"rho0": ..., "rho1": ..., "rho2": ..., "rmax": ...}
},
"governance_mode": "fresh|ladder|graceful_degradation|shield_only|emergency_stop",
"backup_mode": "fresh|ladder|graceful_degradation|shield_only|emergency_stop",

"backup_control": {
    "mode": "fresh|ladder|graceful_degradation|shield_only|emergency_stop",
    "shield_hash": "...",
    "core_envelope_id": "...",
    "perf_envelope_id": "...",
    "nu_core": ...,
    "nu_min": ...,
    "park_set_hash": "...",
    "H_park": ...,
    "E_park": ...,
    "S_min": ...,
    "harm_caps_hash": "...",
    "L_acc": ...,
    "alpha_acc": ...,
    "controller_hash_deprecated": "...",
    "backup_cert_hash_deprecated": "...",
    "backup_envelope_hash_deprecated": "..."
},

"diversity_challenge": {
    "Q_pub_hash": "...",
    "Q_sec_commit_hash": "...",
    "Q_sec_reveal_hash": "...",
    "N_min": ...,
    "game_gap_max": ...
},

"audit_plan": {
    "micro_enabled": true/false,

```

```

    "meso_sample_prob":...,
    "macro_full_replay_period":...,
    "audit_beacon_hash":"..."
  },

  "intervals": {
    "I":{"L":..., "U":..., "eps_opt":...},
    "M":{"L":..., "U":..., "eps_opt":...},
    "P":{"L":..., "U":..., "eps_opt":...}
  },

  "epsilon_margin": ...,
  "declared_set": [...],
  "alarm": {"mis":0/1, "link_broken":0/1, "structure_violation":0/1,
    ↪ "branch":"normal|fail_closed"},
  "solver_meta": {
    "solver_version":"...",
    "tolerance":...,
    "seed":...,
    "threads":...,
    "catalog_hash":"...",
    "tie_break_rule":"lexicographic|score_then_id",
    "cs_profile_hash":"...",
    "iqc_activation_lag":"1-step"
  },
  "certificate_hash":"..."
}

```

**Path-level required-key rule.**

**Required keys and path ownership.** Required keys are enforced at their owning object path. In particular, `shield_hash`, `core_envelope_id`, `perf_envelope_id`, `nu_core`, `nu_min`, ... are required inside `backup_control.*` (not at root), and governance-liveness keys are required inside `governance_exogenous.*`.

Canonical mode alphabet is

`{fresh, ladder, graceful_degradation, shield_only, emergency_stop}`,

and root `backup_mode` must equal `backup_control.mode`.

**Theorem-relevant required fields.** The following fields are required for replaying Assumptions 5.9, 5.15, 5.16, 5.18, 5.20, 5.21, 11.6, 11.7, 2.4, 12.16, and 17.1:

- **Numeric / envelope / optimization** `eps_E_num`, `eta_pub`, `n_eff_lb`, `sigma_pub`, `b_int_pub`, `b_link_pub`, `env_source_mode`, `env_evalues`, `rect_endpoint_bounds`, `b_rect_pub`, `iqc_bridge_mode`, `delta_peak_iqc`, `delta_impl`, `eps_pol_t`, `catalog_level`, `bnb_nodes_used`, `raw_relax_mode`, `raw_relax_gap`, channelwise `eps_opt`, full alpha tuple (including `env`).
- **Backup-control** `backup_control.mode`, `backup_control.shield_hash`, `backup_control.core_envelope_id`, `backup_control.perf_envelope_id`, `backup_control.nu_core`, `backup_control.nu_min`, `backup_control.park_set_hash`, `backup_control.H_park`,

backup\_control.E\_park, backup\_control.S\_min, backup\_control.harm\_caps\_hash, backup\_control.L\_acc, backup\_control.alpha\_acc.

- **Identifiability / alias-hazard** delta\_theta, covering\_number, kappa\_id, rho\_probe, nu\_id, c\_id, iglower\_t, alias\_class\_count, alias\_class\_hash, alias\_hazard\_t, lambda\_h\_t.
- **Soft-abstention contract** soft\_abstention.semantics, soft\_abstention.soft\_band, soft\_abstention.ui\_numeric\_exposed, soft\_abstention.control\_plane\_soft\_fields\_allowed.
- **Governance exogenous** governance\_exogenous.g\_exo\_hash, governance\_exogenous.mode, governance\_exogenous.outage\_len, governance\_exogenous.H\_grace, governance\_exogenous.ladder\_hash, governance\_exogenous.ladder\_index, governance\_exogenous.phi\_feas\_lb, governance\_exogenous.phi\_min, governance\_exogenous.disturbance\_jump\_sentinel, governance\_exogenous.risk\_budget.\*, governance\_exogenous.open\_world.\*.
- **Diversity challenge** diversity\_challenge.Q\_pub\_hash, diversity\_challenge.Q\_sec\_commit\_hash, diversity\_challenge.Q\_sec\_reveal\_hash, diversity\_challenge.N\_min, diversity\_challenge.game\_gap\_max.
- **Audit plan** audit\_plan.micro\_enabled, audit\_plan.meso\_sample\_prob, audit\_plan.macro\_full\_replay\_period, audit\_plan.audit\_beacon\_hash.
- **Link decomposition (fully qualified path)** envelopes.b\_link\_components.b\_link\_hat, envelopes.b\_link\_components.u\_link, envelopes.b\_link\_components.r\_open\_world.

For Assumptions 2.4–12.15, replay bundles must additionally require: governance\_exogenous.b\_lock\_seq, governance\_exogenous.block\_time\_unix, governance\_exogenous.signer\_set\_hash, governance\_exogenous.k\_required, governance\_exogenous.n\_total, governance\_exogenous.fork\_choice\_id, governance\_exogenous.transport\_path\_id, governance\_exogenous.outage\_type, envelope-side envelopes.link\_alarm\_evaluate, envelopes.mis\_alarm\_evaluate, envelopes.structure\_residual, envelopes.tau\_str, and alarm bits alarm.link\_broken, alarm.structure\_violation.

Legacy aliases v\_id, pi\_soft, controller\_hash, backup\_cert\_hash, and backup\_envelope\_hash may appear in transitional logs but are deprecated and must not be used as canonical required keys. Implementations may extend this schema, but theorem-relevant fields must remain public and deterministic.

## Machine-checkable JSON Schema (normative subset).

```
{
  "$schema": "https://json-schema.org/draft/2020-12/schema",
  "type": "object",
  "required": [
    "time", "history_hash", "rulebook_hash", "alpha", "theta_op_nonempty",
    "numerics", "envelopes", "identifiability", "governance_exogenous",
    "governance_mode", "backup_mode", "backup_control", "intervals",
    "alarm", "solver_meta", "certificate_hash"
  ],
  "properties": {
    "time": { "type": "integer", "minimum": 0 },
    "history_hash": { "type": "string", "minLength": 1 },
```

```

"rulebook_hash": { "type": "string", "minLength": 1 },
"theta_op_nonempty": { "type": "boolean" },
"alpha": {
  "type": "object",
  "required": ["out", "ev", "dep", "env", "mis"],
  "properties": {
    "out": { "type": "number", "minimum": 0, "maximum": 1 },
    "ev": { "type": "number", "minimum": 0, "maximum": 1 },
    "dep": { "type": "number", "minimum": 0, "maximum": 1 },
    "env": { "type": "number", "minimum": 0, "maximum": 1 },
    "mis": { "type": "number", "minimum": 0, "maximum": 1 }
  },
  "additionalProperties": false
},
"governance_mode": {
  "type": "string",
  "enum": ["fresh", "ladder", "graceful_degradation", "shield_only", "emergency_stop"]
},
"backup_mode": {
  "type": "string",
  "enum": ["fresh", "ladder", "graceful_degradation", "shield_only", "emergency_stop"]
},
"governance_exogenous": {
  "type": "object",
  "required": [
    "g_exo_hash", "g_exo_version", "block_seq", "block_time_unix",
    "signer_set_hash", "k_required", "n_total", "fork_choice_id",
    "transport_path_id", "outage_type", "mode", "outage_len", "H_grace",
    "ladder_hash", "ladder_index", "phi_feas_lb", "phi_min",
    "disturbance_jump_sentinel", "risk_budget", "open_world"
  ],
  "properties": {
    "outage_type": {
      "type": "string",
      "enum": ["none", "connectivity", "crypto_invalid"]
    },
    "mode": {
      "type": "string",
      "enum":
        ↪ ["fresh", "ladder", "graceful_degradation", "shield_only", "emergency_stop"]
    }
  }
},
"alarm": {
  "type": "object",
  "required": ["mis", "link_broken", "structure_violation", "branch"],
  "properties": {
    "mis": { "type": "integer", "enum": [0,1] },
    "link_broken": { "type": "integer", "enum": [0,1] },
    "structure_violation": { "type": "integer", "enum": [0,1] },
    "branch": { "type": "string", "enum": ["normal", "fail_closed"] }
  },
  "additionalProperties": false
}

```

```

    },
    "additionalProperties": true
}

```

Cross-field checks (enforced by replay verifier code, not plain JSON Schema):

- (cv1) `backup_mode == backup_control.mode`.
- (cv2) `governance_mode == governance_exogenous.mode`.
- (cv3) If `governance_exogenous.outage_type == crypto_invalid`, then `governance_mode != fresh`.
- (cv4) If `alarm.mis==1` or `alarm.link_broken==1` or `alarm.structure_violation==1`, then `declared_set` must be empty.

## E Appendix E: Reference implementation pseudocode

**Design goal.** This appendix gives a lightweight deterministic implementation blueprint. It is intentionally explicit about data contracts, branch logic, and replay determinism.

**State container (minimal).**

```

State:
  t: int
  model_sets:
    theta_ev_raw
    theta_ev_rect
    theta_out_exact
    theta_out_num
    theta_op
    theta_in
  monitors:
    E_out(theta), E_ev(theta), M_mis, M_link, M_env[q]
  envelopes:
    b_cont, b_dep, b_int, b_link, b_rect, b_impl
  scores:
    Lhat[j], Uhat[j], eps_opt[j]
  control:
    governance_mode
    backup_mode
    lambda_h

```

**Algorithm 1: one-step online update and decision.**

```

function STEP(bundle_t, state_{t-1}):
  VERIFY_BUNDLE(bundle_t, state_{t-1})
  INGEST_NUMERICS_AND_ENV(bundle_t)
  eps      <- bundle_t.epsilon_margin
  tau_num  <- bundle_t.numerics.tau_num
  (b_impl, b_cont, b_dep, b_int, b_link, b_rect)
    <- READ_ENVELOPES(bundle_t)

  # 1) Update e-processes and model sets

```

```

UPDATE_EVIDENCE_PROCESS(E_ev, bundle_t)
UPDATE_OUTER_PROCESS(E_out, bundle_t)
UPDATE_MISSPEC_PROCESS(M_mis, bundle_t)
theta_ev_raw <- BUILD_EVIDENCE_SET(bundle_t)
theta_ev_rect <- RECTANGULARIZE(theta_ev_raw, bundle_t.rect_cert)
theta_out_num <- BUILD_OUTER_NUM_SET(
  E_out, bundle_t.alpha.out, bundle_t.numerics
)
theta_op <- INTERSECT(theta_ev_rect, theta_out_num)
theta_in <- UPDATE_INNER_RECURSION(
  state_{t-1}.theta_in, theta_out_num
)

# 2) Endpoint certification and interval construction
if theta_op is empty:
  intervals <- WIDE_INTERVALS({I,M,P})
  declared_set <- empty
  action <- FAIL_CLOSED_BRANCH(bundle_t, state_{t-1})
  iqc_tightening_for_t_plus_1 <- "skipped_fail_closed"
  PUBLISH_DECISION_AND_CERTIFICATE(
    action, declared_set, intervals,
    iqc_tightening_for_t_plus_1
  )
  return ADVANCE_STATE(
    state_{t-1}, bundle_t, action, intervals, declared_set
  )

for j in {I,M,P}:
  (L_raw, U_raw, eps_opt_j) <- CERTIFY_SCORE_ENDPOINT(
    theta_op, j, bundle_t.solver_meta
  )
  inflate <- b_impl + b_cont + b_dep + b_int + b_link
  inflate <- inflate + b_rect + eps_opt_j
  Lhat[j] <- L_raw - inflate
  Uhat[j] <- U_raw + inflate
  eps_opt[j] <- eps_opt_j
  intervals[j] <- (Lhat[j], Uhat[j], eps_opt_j)

declared_set <- {
  j : Lhat[j] >= max_{k!=j} Uhat[k] + eps + tau_num
}

# 3) Alarm / compute gates
if MIS_OR_LINK_OR_STRUCTURE_ALARM(bundle_t):
  declared_set <- empty
  action <- FAIL_CLOSED_BRANCH(bundle_t, state_{t-1})
  iqc_tightening_for_t_plus_1 <- "skipped_alarm"
  PUBLISH_DECISION_AND_CERTIFICATE(
    action, declared_set, intervals,
    iqc_tightening_for_t_plus_1
  )
  return ADVANCE_STATE(
    state_{t-1}, bundle_t, action, intervals, declared_set
  )

```



```

if COMPUTE_GATE_FAILS(
  Lhat, Uhat, eps_opt, b_impl, eps, tau_num
):
  EMIT_SOFT_ABSTENTION(Lhat, Uhat, eps_opt, b_impl, tau_num)
  declared_set <- empty
  action <- COMPUTE_LIMITED_SAFE_OR_PROBE(
    theta_op, bundle_t, state_{t-1}
  )
  iqc_tightening_for_t_plus_1 <- "skipped_compute_gate"
  PUBLISH_DECISION_AND_CERTIFICATE(
    action, declared_set, intervals,
    iqc_tightening_for_t_plus_1
  )
  return ADVANCE_STATE(
    state_{t-1}, bundle_t, action, intervals, declared_set
  )

# 4) Normal action branch
if |declared_set| == 1:
  action <- PRIMARY_RELIEF_POLICY(
    singleton(declared_set), bundle_t, state_{t-1}
  )
else:
  action <- SOLVE_D6_PROBE_POLICY(theta_op, bundle_t, state_{t-1})

# 5) Lag-one IQC tightening (non-circular)
iqc_tightening_for_t_plus_1 <- COMPUTE_IQC_TIGHTENING(
  theta_op, bundle_t
)
PUBLISH_DECISION_AND_CERTIFICATE(
  action, declared_set, intervals,
  iqc_tightening_for_t_plus_1
)
return ADVANCE_STATE(
  state_{t-1}, bundle_t, action, intervals, declared_set
)

```

#### Algorithm 2: deterministic endpoint certification API.

```

function CERTIFY_SCORE_ENDPOINT(theta_op, channel_j, solver_meta):
  # deterministic branch-and-bound / convex dual / hybrid
  # mandatory deterministic controls:
  #   fixed seed, fixed tie-break, fixed rounding mode, fixed stopping rules
  lower_bound <- GLOBAL_LOWER_BOUND(theta_op, channel_j, solver_meta)
  upper_bound <- GLOBAL_UPPER_BOUND(theta_op, channel_j, solver_meta)
  eps_opt_j <- CERTIFIED_GAP(upper_bound, lower_bound, solver_meta)
  return (lower_bound, upper_bound, eps_opt_j)

```

#### Algorithm 3: fail-closed backup controller.

```

function FAIL_CLOSED_BRANCH(bundle_t, state):
  if VALID_CORE_CERT(bundle_t.backup_control):
    mode <- "shield_only"

```

```

    action <- SHIELDED_BACKUP(bundle_t, state)
else if VALID_PARKING_CERT(bundle_t.backup_control):
    mode <- "graceful_degradation"
    action <- PARKING_POLICY(bundle_t, state)
else:
    mode <- "emergency_stop"
    action <- EMERGENCY_STOP(bundle_t, state)
PUBLISH_MODE_AND_ALARM(mode)
return action

```

**Algorithm 4: full replay verifier (third-party audit).**

```

function REPLAY_VERIFY(log_stream):
    state <- INIT_FROM_GENESIS(log_stream.init)
    for each bundle_t in log_stream in chronological order:
        CHECK_HASH_CHAIN(bundle_t)
        CHECK_SIGNATURE_THRESHOLD(bundle_t.governance_exogenous)
        CHECK_FORK_CHOICE(bundle_t.governance_exogenous)
        CHECK_SCHEMA_AND_CROSS_FIELD_INVARIANTS(bundle_t)
        state <- STEP(bundle_t, state)
        ASSERT(bundle_t.certificate_hash == HASH(CANONICAL_SERIALIZE(state.outputs)))
    return "replay_ok"

```

**Implementation notes (lightweight profile).**

- (im1) Use one deterministic solver profile per endpoint family; avoid runtime solver auto-tuning.
- (im2) Keep  $\Theta_t^{\text{op}}$  sparse by early pruning with e-process thresholds and rectangle certificates.
- (im3) Run endpoint certification in anytime mode with hard wall-clock budget and emit compute-limited soft abstention when the compute gate fails.
- (im4) Store only hashes for large artifacts (tube proofs, catalog snapshots); keep full artifacts in content-addressed storage.
- (im5) Use fixed-width decimal serialization for all published floating-point values before hashing.

**Reference map (reader guidance).** Foundational robust MDP and stochastic-control machinery: [1, 2, 3, 4, 6, 5]. Partial identification and observational limits: [7, 8]. Anytime/e-process validity and confidence-sequence construction: [18, 19, 20, 21]. IQC and delay-robust analysis: [12, 13, 14, 11]. Set-valued and measurable-analysis tools: [15, 17, 9, 16].

## References

- [1] M. L. Puterman. *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. Wiley, 1994.
- [2] D. P. Bertsekas and S. E. Shreve. *Stochastic Optimal Control: The Discrete-Time Case*. Athena Scientific, 1996.
- [3] G. Iyengar. Robust dynamic programming. *Mathematics of Operations Research*, 30(2):257–280, 2005. DOI: <https://doi.org/10.1287/moor.1040.0129>.

- [4] A. Nilim and L. El Ghaoui. Robust control of Markov decision processes with uncertain transition matrices. *Operations Research*, 53(5):780–798, 2005. DOI: <https://doi.org/10.1287/opre.1050.0216>.
- [5] L. G. Epstein and M. Schneider. Recursive multiple-priors. *Journal of Economic Theory*, 113(1):1–31, 2003. DOI: [https://doi.org/10.1016/S0022-0531\(03\)00097-8](https://doi.org/10.1016/S0022-0531(03)00097-8).
- [6] L. P. Hansen and T. J. Sargent. *Robustness*. Princeton University Press, 2008.
- [7] C. F. Manski. *Partial Identification of Probability Distributions*. Springer, 2003.
- [8] C. F. Manski. *Identification for Prediction and Decision*. Harvard University Press, 2007.
- [9] R. T. Rockafellar and R. J.-B. Wets. *Variational Analysis*. Springer, 1998.
- [10] M. Sion. On general minimax theorems. *Pacific Journal of Mathematics*, 8(1):171–176, 1958. DOI: <https://doi.org/10.2140/pjm.1958.8.171>.
- [11] S. Boyd, L. El Ghaoui, E. Feron, and V. Balakrishnan. *Linear Matrix Inequalities in System and Control Theory*. SIAM, 1994.
- [12] A. Megretski and A. Rantzer. System analysis via integral quadratic constraints. *IEEE Transactions on Automatic Control*, 42(6):819–830, 1997. DOI: <https://doi.org/10.1109/9.587335>.
- [13] L. Lessard, B. Recht, and A. Packard. Analysis and design of optimization algorithms via integral quadratic constraints. *SIAM Journal on Optimization*, 26(1):57–95, 2016. DOI: <https://doi.org/10.1137/15M1009597>.
- [14] E. Fridman. *Introduction to Time-Delay Systems: Analysis and Control*. Birkhäuser, 2014.
- [15] J.-P. Aubin and H. Frankowska. *Set-Valued Analysis*. Birkhäuser, 2009.
- [16] C. D. Aliprantis and K. C. Border. *Infinite Dimensional Analysis: A Hitchhiker’s Guide*. 3rd edition, Springer, 2006.
- [17] I. Molchanov. *Theory of Random Sets*. Springer, 2005.
- [18] J. Ville. *Étude Critique de la Notion de Collectif*. Gauthier-Villars, 1939.
- [19] D. A. Freedman. On tail probabilities for martingales. *The Annals of Probability*, 3(1):100–118, 1975. DOI: <https://doi.org/10.1214/aop/1176996452>.
- [20] S. R. Howard, A. Ramdas, J. McAuliffe, and J. Sekhon. Time-uniform Chernoff bounds via nonnegative supermartingales. *Probability Surveys*, 17:257–317, 2020. DOI: <https://doi.org/10.1214/18-PS321>.
- [21] S. R. Howard, A. Ramdas, J. McAuliffe, and J. Sekhon. Time-uniform, nonparametric, nonasymptotic confidence sequences. *The Annals of Statistics*, 49(2):1055–1080, 2021. DOI: <https://doi.org/10.1214/20-AOS1991>.
- [22] C. J. Clopper and E. S. Pearson. The use of confidence or fiducial limits illustrated in the case of the binomial. *Biometrika*, 26(4):404–413, 1934. DOI: <https://doi.org/10.1093/biomet/26.4.404>.
- [23] D. M. Topkis. *Supermodularity and Complementarity*. Princeton University Press, 1998.
- [24] J. Tirole. *The Theory of Industrial Organization*. MIT Press, 1988.