Toward Online Robust Zero-Sum Markov Games with Function Approximation

Chenhao Zhou
College of Computer Science and Technology
Zhejiang University

Jiyu Wei*
College of Computer Science and Technology
Zhejiang University

Abstract

Online robust zero-sum Markov games pose a coupled operator-learning and game-solving problem: from nominal interaction data, the learner must recover a worst-case minimax Bellman operator while maintaining strategic coverage and controlling stage-game approximation error. Existing robust reinforcement-learning theory does not yet provide such an operator-learning perspective beyond specialized multi-agent settings. We develop an optimistic dual robust Bellman framework that combines dual robust Bellman fitting, optimistic minimax planning, backward-validated confidence sets, and exploiter-mix data collection. Our analysis is organized around the robust Bellman residuals induced by the candidate value class and yields a regret reduction that separates exploration complexity, robust operator-estimation error, stage-game solution error, and misspecification. Under explicit interpretable assumptions, we obtain conditional theorem interfaces for general and projected function approximation, together with closed results in several structured regimes, including tabular, epochwise/cross-fit linear, self-normalized pure-online linear / finite-rank kernel, and projected RKHS truncation. Small-scale numerical experiments illustrate robustness calibration, strategic coverage, and the return–gap trade-off predicted by the framework.

1 Introduction

Robust reinforcement learning studies deployment-time mismatch between nominal training dynamics and test dynamics [1–3]. In multi-agent settings the mismatch is also strategic: transitions depend on both players, robustness is defined relative to minimax behavior, and a worst-case Bellman operator must be learned from nominal data [4, 5]. This leads to online robust zero-sum Markov games.

The single-agent side of this picture has recently advanced to online robust MDPs with general function approximation [6], but the same extension is substantially harder in zero-sum Markov games. First, the robust Bellman operator now combines minimax optimization with distributional robustness, so both closure and statistical estimation must be carried out for a robust minimax backup rather than a single-agent robust backup. Second, strategic coverage is more delicate: the occupancies relevant for regret are induced by minimax execution, whereas the data used to fit robust operators are generated by adaptive play, so exploration and validation must control a genuinely multi-agent coverage gap. Third, the non-robust large-state Markov-game tools of Huang et al. [7], Jin et al. [8] do not by themselves address robust operator learning, while existing robust Markov-game results remain concentrated in tabular, generative-model, or linear regimes [9–11, 5]. These obstacles explain why a broad robust zero-sum Markov-game theory with function approximation has remained open.

This paper develops an optimistic dual robust Bellman (ODRB) framework for online robust zero-sum Markov games. At a high level, ODRB fits a worst-case Bellman backup from nominal transitions through a dual representation, tests candidate value functions on fresh data using a frozen backup,

*Corresponding author. Email: weijiyu@zju.edu.cn

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

plans optimistically by solving the induced zero-sum stage games, and occasionally executes one-sided response branches to cover deviations that the target pair may not visit. Epochwise freezing and cross-fitting keep the data used to fit the robust backup separate from the data used to validate its residual. The analysis is organized around the robust residual class $(I - \mathcal{T}^\sigma)\mathcal{F}$ and a square-loss robust Markov-game Bellman–Eluder complexity, which together separate exploration, robust operator estimation, stage-game solution error, and misspecification. On the general side, we obtain conditional regret guarantees for general function approximation and projected completeness under interpretable assumptions. On the structured side, we prove tabular, epochwise/cross-fit linear, modified pure-online linear/finite-rank kernel, and projected finite-dimensional or RKHS guarantees. The epochwise/cross-fit linear theorem is the main fully specified scalable result under its stated structural conditions; the appendix also gives a self-normalized pure-online linear/finite-rank specialization.

Our main contributions are:

- We formulate an optimistic dual robust Bellman framework for online robust zero-sum Markov games with function approximation, combining robust Bellman fitting, minimax planning, backward-validated confidence sets, and strategic exploration. This gives a single algorithmic template connecting robust-MDP operator learning with minimax planning in large-state zero-sum Markov games.
- We introduce a robust multi-agent residual perspective based on the Bellman residuals induced by the candidate value class, namely the class $(I - \mathcal{T}^\sigma)\mathcal{F}$. This yields a square-loss robust MG-BE complexity measure, a deterministic regret reduction, and conditional interface theorems for general function approximation and projected completeness. It separates approximation, operator learning, coverage, and stage-game errors in a form that can be instantiated across multiple function classes.
- We prove a structured theorem package including tabular, epochwise/cross-fit linear, self-normalized pure-online linear/finite-rank kernel, and projected finite-dimensional or RKHS-truncation regimes, together with concrete closure and coverage conditions such as TV-linear-mixture closure and primitive BR-based coverage. The epochwise/cross-fit linear theorem is the main fully specified scalable guarantee, and the projected results show how the same framework extends to finite-dimensional lifts and kernel truncations.

We report theorem-aligned experiments illustrating robustness calibration, strategic coverage, and the return–gap trade-off, without claiming uniform empirical superiority across operating points.

1.1 Related Work

Robust RL and robust MDPs. Classical robust dynamic programming and distributionally robust MDPs were developed by [1, 2, 12, 3, 13]; recent surveys include [14]. More recent work studies robust function approximation, online robust learning, and off-dynamics or offline estimation [15–22]. Our starting point is [6], which gives the robust-MDP methodology based on dual Bellman fitting, global confidence sets, and intrinsic robust complexity.

Function approximation in Markov games. For non-robust zero-sum Markov games, recent work develops optimistic elimination, minimax complexity measures, exploiter-assisted exploration, and spectral embeddings for continuous-state games under rich function approximation [7, 8, 23–26]. More generally, the modern RL literature supplies the complexity language behind our analysis, including Eluder-style dimensions, Bellman rank, and coverage-sensitive offline or approximate dynamic programming viewpoints [27–39].

Robust Markov games and robust MARL. Existing robust multi-agent results are mostly tabular, generative-model, or linear [4, 9–11, 40, 5]. The recent model-based online DRMG line, represented by [40], works with explicit TV/KL ambiguity sets and bonus-style optimistic planning in tabular settings, with guarantees expressed in state- and joint-action-dependent terms. The closest large-state comparison is the linear robust Markov-game result of [5], whose algorithm DR-CCE-LSI targets robust coarse correlated equilibrium in general-sum games under linear function approximation, minimum-value style assumptions, and a different equilibrium notion from the zero-sum robust minimax value studied here. Our comparison to these works is therefore assumption- and theorem-level, not only narrative: tabular RONAVI-style methods are model-based and bonus-driven, whereas

Table 1: Positioning relative to representative prior work. "General FA theorem" indicates that a paper states a theorem-level guarantee for general or projected function approximation. "Closed scalable theorem" indicates a fully verified large-state result inside the paper, rather than only a tabular or generative specialization.

Work	Problem class	Robust	Approx. regime	General FA theorem	Closed scalable theorem
Ghosh et al. [6]	robust MDP	✓	GFA	✓	✓
Huang et al. [7], Jin et al. [8]	zero-sum MG	×	GFA	✓	✓
Farhat et al. [40]	robust MG	✓	tabular / model-based	×	✓
Ma et al. [9], Shi et al. [10, 11]	robust MG / MARL	✓	tabular / generative	×	problem-specific
Zheng and Lin [5]	robust MG	✓	linear	×	✓
Ours	**robust zero-sum MG**	✓	**GFA, projected, linear / kernel**	✓	✓

our method learns a robust Bellman operator through dual fitting; DR-CCE-LSI is a linear equilibrium-learning result, whereas our linear theorem is the main scalable specialization of a broader residual-based operator-learning framework. Accordingly, our tabular, linear, and projected guarantees should be read as structured instances of one theorem hierarchy, rather than as isolated regime-specific algorithms, as summarized in Table 1.

2 Preliminaries

We study episodic two-player zero-sum robust Markov games $\mathcal{M} = (\mathcal{S}, \mathcal{A}, \mathcal{B}, H, \{r_h\}, \{P_h^0\}, \{\mathcal{U}_h\})$, where player 1 maximizes reward and player 2 minimizes it. At stage $h \in [H]$, the game is in state $s_h \in \mathcal{S}$, player 1 chooses $a_h \in \mathcal{A}$, player 2 chooses $b_h \in \mathcal{B}$, and player 1 receives reward $r_h(s_h, a_h, b_h) \in [0, 1]$. The nominal next-state law is $P_h^0(\cdot \mid s_h, a_h, b_h)$, but the true transition is allowed to vary inside an uncertainty set $\mathcal{U}_h(s, a, b)$ around $P_h^0(\cdot \mid s, a, b)$, typically defined by TV, KL, or more general ϕ-divergence balls [1, 2, 6]. A policy pair (π, ν) consists of Markov decision rules $\pi_h(\cdot \mid s) \in \Delta(\mathcal{A})$ and $\nu_h(\cdot \mid s) \in \Delta(\mathcal{B})$, where $\Delta(\mathcal{A})$ and $\Delta(\mathcal{B})$ denote the simplices over the two action sets. We regard (a_h, b_h) as a joint action in $\mathcal{A} \times \mathcal{B}$, while keeping the two-player notation explicit because minimax structure is central throughout the paper. Throughout the paper, σ denotes the robustness radius indexing the ambiguity sets; a representative form is

$$\mathcal{U}_h(s, a, b) = \Big\{ P : \; D\big(P, P_h^0(\cdot \mid s, a, b)\big) \leqslant \sigma \Big\},$$

where D may be total variation, KL, or χ^2 divergence. We use σ in the value and Bellman notation to indicate dependence on this robustness level, and $\sigma = 0$ recovers the nominal, non-robust game.

2.1 Robust Value Functions and Bellman Equation

Given a policy pair (π, ν), the robust action-value and value functions are defined backward with terminal condition $V_{H+1}^{\pi,\nu,\sigma} \equiv 0$ by

$$Q_h^{\pi,\nu,\sigma}(s, a, b) = r_h(s, a, b) + \inf_{P \in \mathcal{U}_h(s,a,b)} \mathbb{E}_{s' \sim P}\left[V_{h+1}^{\pi,\nu,\sigma}(s')\right],$$

with $V_h^{\pi,\nu,\sigma}(s) = \mathbb{E}_{a \sim \pi_h(\cdot|s),\, b \sim \nu_h(\cdot|s)}\left[Q_h^{\pi,\nu,\sigma}(s, a, b)\right]$. Thus $Q_h^{\pi,\nu,\sigma}(s, a, b)$ is the robust continuation value after fixing the current action pair, whereas $V_h^{\pi,\nu,\sigma}(s)$ averages that quantity under the two players' mixed actions at state s. Equivalently, the pair $(Q_h^{\pi,\nu,\sigma}, V_h^{\pi,\nu,\sigma})$ satisfies the robust Bellman recursion for the joint Markov policy (π, ν). The optimal robust value at stage h is the minimax quantity $V_h^{\star,\sigma}(s) := \max_\pi \min_\nu V_h^{\pi,\nu,\sigma}(s)$, with the max and min interpreted over Markov policy pairs. This is the zero-sum robust analogue of the usual game value: player 1 optimizes against both player 2 and the worst-case transition selector, while player 2 minimizes the resulting robust return. In the online setting considered here, regret is measured relative to this robust minimax benchmark at the initial state.

The model uses one shared zero-sum payoff $J^\sigma(\pi, \nu) := V_1^{\pi,\nu,\sigma}(s_1)$: player 1 maximizes it, whereas player 2 and Nature minimize it. Thus the apparent asymmetry is operational—the maximizing controller must withstand both a strategic opponent and transition misspecification—rather than a mismatch between two payoff definitions. This covers, for example, a defender facing a strategic attacker under uncertain failures. If each player instead separately robustifies its own subjective payoff, then in general $\inf_P J_P \neq -\inf_P(-J_P) = \sup_P J_P$, so the common zero-sum payoff and the Bellman saddle recursion no longer apply.

2.2 Robust Minimax Bellman Operator

For a candidate Q-sequence $f = \{f_h\}_{h=1}^H$, define the stage-game value

$$V_{f,h}(s) := \max_{p \in \Delta(\mathcal{A})} \min_{q \in \Delta(\mathcal{B})} \sum_{a,b} p(a)q(b) f_h(s,a,b).$$

This is the value of the zero-sum matrix game obtained by freezing the state at s and using $f_h(s,\cdot,\cdot)$ as the payoff matrix. In particular, $V_{f,h}(s)$ compresses a state-dependent action-value table into the corresponding minimax state value. The corresponding robust minimax Bellman operator is

$$(\mathcal{T}_h^\sigma f_{h+1})(s,a,b) = r_h(s,a,b) + \inf_{P \in \mathcal{U}_h(s,a,b)} \mathbb{E}_{s' \sim P}[V_{f,h+1}(s')]\,.$$

This is the multi-agent analogue of the robust Bellman operator in robust MDPs, except that the continuation value is now itself a minimax object. For the optimal robust value, the Bellman equation takes the fixed-point form $Q_h^{\star,\sigma} = \mathcal{T}_h^\sigma Q_{h+1}^{\star,\sigma}$ and $V_h^{\star,\sigma}(s) = V_{Q^\star,h}(s)$. Our algorithmic and theoretical developments are therefore organized around estimating and controlling this robust minimax Bellman operator.

2.3 Online Objective and Occupancies

In episode k, the learner selects a policy pair (π^k, ν^k) and is evaluated against the corresponding worst-case selector inside the uncertainty set. The performance criterion in the main text is robust regret relative to the optimal robust minimax value at the initial state. The analysis therefore uses the executed occupancy induced jointly by the policy pair and the worst-case selector. By contrast, robust Bellman fitting and backward validation are performed from nominal interaction data generated by the data-collection policy. This distinction is important in the multi-agent robust setting: regret is controlled on the worst-case execution trajectory, whereas statistical estimation is carried out on nominal data, and the coverage conditions in the theory are precisely what connect these two occupancy families. More precisely, if s_1 is the fixed initial state, then the cumulative performance criterion studied in the paper is

$$\mathrm{Reg}(K) := \sum_{k=1}^K \Big(V_1^{\star,\sigma}(s_1) - V_1^{\pi^k,\nu^k,\sigma}(s_1) \Big).$$

The comparator is the robust minimax value, but the summand above is a signed, one-sided value-comparator difference for the jointly generated pair. It is not the two-sided saddle exploitability

$$\max_{\pi'} V_1^{\pi',\nu^k,\sigma}(s_1) - \min_{\nu'} V_1^{\pi^k,\nu',\sigma}(s_1).$$

Consequently, a sublinear bound on $\mathrm{Reg}(K)$ controls the stated robust value comparator but does not by itself certify that an output pair is an approximate Nash equilibrium; the latter would require separate control of both unilateral-deviation terms.

2.4 Function Classes and Residual Complexity

Let $\mathcal{F} = \mathcal{F}_1 \times \cdots \times \mathcal{F}_H$ be the candidate Q-function class and let $\mathcal{G}_h$ denote the dual class used to fit the robust backup at stage h. Our analysis is organized around the robust residual class

$$\mathscr{R}_h := \{ f_h - \mathcal{T}_h^\sigma f_{h+1} : f \in \mathcal{F} \}\,.$$

Following the Bellman–Eluder viewpoint of Ghosh et al. [6] and the Markov-game GFA literature [7, 8], we measure exploration difficulty through a square-loss robust MG-BE dimension attached to this residual process. The exact counting definition and its projected variants appear in the appendix; in the main text it suffices to interpret the dimension as the number of statistically distinguishable robust residual directions encountered along learning. When the ambiguity set collapses to the singleton nominal kernel, $\mathcal{T}_h^\sigma$ reduces to the usual Markov-game Bellman operator and the robust MG-BE notion reduces to its non-robust counterpart. The resulting bounds recover the standard $\sqrt{K}$-type episode dependence of non-robust zero-sum Markov games with general or linear function approximation [7, 8], up to the different complexity, coverage, horizon, and logarithmic factors; we

do not claim an improved nominal rate. Under nontrivial ambiguity sets, the residual class may be richer because the backup also contains worst-case transition directions.

The residual class is the right object because it is exactly what appears after Bellman telescoping in the regret decomposition. In other words, the analysis does not need to identify the entire robust Bellman operator uniformly over the ambient function class; it only needs to control those residual directions that are actually induced by candidate functions in $\mathcal{F}$. This viewpoint is what makes it possible to separate exploration complexity from the lower-order errors due to dual approximation, projection bias, and inexact stage-game solution. Two concrete examples are useful. In tabular games, the robust residual class lives in a space with one coordinate per joint state-action triple, so the robust MG-BE complexity is $O(H|\mathcal{S}||\mathcal{A}||\mathcal{B}|)$ up to logarithmic factors. In linear classes, if the residual process admits witness features of dimension d, then the appendix shows that the robust MG-BE dimension scales on the order of $d\log(1+1/\epsilon^2)$, equivalently $O(d\log K)$ when evaluated at $\epsilon = 1/\sqrt{K}$, matching the usual linear Bellman–Eluder behavior.

3 Method

We study a family of optimistic algorithms for online robust zero-sum Markov games. The method has four components: robust Bellman fitting through a dual representation, global confidence sets over the value-function class, minimax policy extraction, and strategically mixed data collection. We refer to this template as *optimistic dual robust Bellman* (ODRB). The main theorem analyzes an epochwise implementation with cross-fit validation; the appendix gives additional structured variants, including a self-normalized pure-online specialization.

3.1 Dual Robust Bellman Fitting

For a candidate action-value function $f = (f_1, \dots, f_H) \in \mathcal{F}$, the robust Bellman target at stage h is

$$(\mathcal{T}_h^\sigma f_{h+1})(s,a,b) = r_h(s,a,b) + \inf_{P \in \mathcal{U}_h(s,a,b)} \mathbb{E}_{s' \sim P}[V_{f,h+1}(s')]\,. \tag{1}$$

Here $V_{f,h+1}(s')$ is the value of the zero-sum stage game with payoff matrix $f_{h+1}(s',\cdot,\cdot)$.

For ϕ-divergence uncertainty sets, the inner robust expectation admits a dual representation. Writing $g = (g_\eta, g_\nu) \in \mathcal{G}_h$ for a stagewise dual variable, and writing $\phi^\star(u) := \sup_t\{ut - \phi(t)\}$ for the convex conjugate of ϕ, the corresponding pointwise dual loss is

$$\ell_\phi(f_{h+1}; s,a,b,s'; g) = g_\eta(s,a,b)\sigma - g_\nu(s,a,b) + g_\eta(s,a,b)\phi^\star\left(\frac{g_\nu(s,a,b) - V_{f,h+1}(s')}{g_\eta(s,a,b)}\right).$$

With this loss convention, the robust continuation value equals $-\inf_g \mathbb{E}_{P_h^0}[\ell_\phi(f_{h+1}; s,a,b,s'; g)]$; the leading minus sign is essential. This converts robust Bellman fitting into empirical risk minimization over the dual class $\mathcal{G}_h$ using nominal interaction data. Concretely, for a stagewise sample buffer $\mathcal{D}_h = \{(s_i, a_i, b_i, s_i')\}_{i=1}^{N_h}$, we define the empirical dual objective

$$\widehat{\mathcal{L}}_h(g; f) := \frac{1}{N_h}\sum_{i=1}^{N_h} \ell_\phi(f_{h+1}; s_i, a_i, b_i, s_i'; g).$$

In the epochwise implementation, $\widehat{\mathcal{L}}_{h,e}^{\mathrm{pre}}(g; f)$ denotes the same empirical average computed on the pre-epoch buffer. The reward term does not appear in $\widehat{\mathcal{L}}_h$ because the dual fit estimates only the robust expectation term in (1); the immediate reward is added back when the plug-in Bellman operator is formed. The resulting plug-in operator is the robust analogue of fitted Bellman evaluation, and the theory isolates its dual approximation error explicitly.

3.2 Confidence Sets and Optimistic Planning

After collecting data up to the current update time, the algorithm refits the dual variable on pooled stage-h data:

$$g_{h,f} \in \operatorname*{argmin}_{g \in \mathcal{G}_h} \widehat{\mathcal{L}}_h(g; f).$$

This induces a plug-in robust Bellman operator, obtained by taking the dual representation of the robust Bellman backup and replacing the population dual optimizer by the empirical refit $g_{h,f}$: $\widehat{\mathcal{T}}^{\sigma}_{h,f}$, defined by

$$\big(\widehat{\mathcal{T}}^{\sigma}_{h,f} f_{h+1}\big)(s,a,b) := r_h(s,a,b) - \mathbb{E}_{s'\sim P^0_h(\cdot|s,a,b)}\left[\ell_\phi(f_{h+1}; s,a,b,s'; g_{h,f})\right].$$

The learner then maintains a global confidence set

$$\mathfrak{F} := \left\{ f \in \mathcal{F} : \widehat{\mathrm{BV}}_h(f) \leqslant \beta_h \text{ for every stage } h \right\}, \tag{2}$$

where $\widehat{\mathrm{BV}}_h(f)$ is built from the plug-in robust Bellman residual $\widehat{\delta}_h(f; s,a,b) := f_h(s,a,b) - (\widehat{\mathcal{T}}^{\sigma}_{h,f} f_{h+1})(s,a,b)$. Concretely, if the past stage-h data are partitioned into batches $\{\mathcal{D}_h^{(j)}\}_{j<k}$ with $n_{h,j} := |\mathcal{D}_h^{(j)}|$, the one-update-per-episode statistic is

$$\widehat{\mathrm{BV}}_{h,k}(f) := \sum_{j=1}^{k-1} \frac{1}{n_{h,j}} \sum_{(s,a,b,r,s')\in\mathcal{D}_h^{(j)}} \left[\widehat{\delta}_{h,k}(f; s,a,b)\right]^2.$$

The epochwise/cross-fit version uses the same score on the fresh within-epoch fold with the frozen operator $\widehat{\mathcal{T}}^{\sigma,(e)}_{h,f}$.

The optimistic value function is chosen from $\mathfrak{F}$ by maximizing the value at the initial state. The role of backward validation is to remove functions whose plug-in robust Bellman residual is inconsistent with the observed data while, with high probability, retaining a Bellman-consistent comparator. This is the key mechanism linking statistical calibration to the regret analysis. Here "backward" means that consistency is checked against the stagewise robust Bellman recursion, starting from the terminal continuation. In the epochwise version, "cross-fit" means that the candidate-dependent dual and plug-in operator are fit and frozen on pre-epoch data, whereas the induced residual witness is evaluated on a fresh within-epoch fold.

3.3 Policy Extraction and Data Collection

Given an optimistic function $f \in \mathfrak{F}$, the policy pair for stage h is obtained by solving the zero-sum matrix game with payoff matrix $f_h(s,\cdot,\cdot)$. This yields a stagewise minimax policy pair (π_h^f, ν_h^f) used both for planning and for defining the robust value $V_{f,h}$.

To guarantee strategic coverage, data are not collected solely from the current target policy. Let $\widehat{V}_1^k(\pi,\nu)$ be the policy-pair value under the current epoch's frozen plug-in robust operator, and define the fitted-game response branches

$$\mathrm{BR}_1^k \in \operatorname*{argmax}_{\pi} \widehat{V}_1^k(\pi, \nu^k), \qquad \mathrm{BR}_2^k \in \operatorname*{argmin}_{\nu} \widehat{V}_1^k(\pi^k, \nu).$$

In each episode, the exploiter-mix rule executes

$$\begin{aligned} &(\pi^k, \nu^k) \quad \text{with probability } 1-p, \\ &(\mathrm{BR}_1^k, \nu^k) \quad \text{with probability } p/2, \qquad (\pi^k, \mathrm{BR}_2^k) \quad \text{with probability } p/2. \end{aligned} \tag{3}$$

Approximate fitted-game responses may be used. Exact response optimality is not an additional premise of the regret conversion, which uses history measurability and bounded rewards; the coverage proof instead invokes the stated branch-induced reachability and excitation conditions. The two response branches expose occupancies reachable under one unilateral strategic deviation that the target pair may not visit.

3.4 Implementation Used in the Main Theory

The main theorem studies an epochwise implementation with cross-fit validation. Here "epochwise" means that the dual operator is refit only at epoch boundaries, while "cross-fit validation" means that the resulting frozen witness family is validated on fresh within-epoch data rather than on the same samples used for refitting. This construction removes the most delicate adaptive-dependence issue

Algorithm 1 Optimistic Dual Robust Bellman (ODRB) with Epochwise Refitting and Cross-Fit Validation

Require: Value class $\mathcal{F}$, dual classes $\{\mathcal{G}_h\}_{h=1}^H$, epoch partition $\{E_e\}_{e\geqslant 1}$, validation radii $\{\beta_{h,e}\}$, exploiter-mix parameter p

Output: deployed sequence $\{(f^e, \pi^e, \nu^e)\}_{e\geqslant 1}$ of optimistic functions and stagewise minimax policy pairs

Initialize pooled stagewise data buffers $\mathcal{D}_h \leftarrow \emptyset$ and confidence set $\mathfrak{F}_1 \leftarrow \mathcal{F}$

for epochs $e = 1, 2, \ldots$ **do**

  Using pre-epoch data $\mathcal{D}_h$, fit frozen dual variables $g_{h,f}^{(e)} \in \arg\min_{g\in\mathcal{G}_h} \widehat{\mathcal{L}}_{h,e}^{\mathrm{pre}}(g; f)$

  Form frozen plug-in operators $\widehat{\mathcal{T}}_{h,f}^{\sigma,(e)}$ and the epoch confidence set

$$\mathfrak{F}_e := \left\{ f \in \mathcal{F} : \widehat{\mathrm{BV}}_{h,e}(f) \leqslant \beta_{h,e},\ \forall h \right\}$$

  Choose $f^e \in \arg\max_{f\in\mathfrak{F}_e} V_{f,1}(s_1)$

  Extract the stagewise minimax policy pair (π^e, ν^e) from $\{f_h^e\}_{h=1}^H$

  for episodes $k \in E_e$ **do**

    Sample and execute one of the three branches in (3); collect trajectory τ^k

    Add τ^k to the within-epoch validation fold and to the pooled buffers $\{\mathcal{D}_h\}_{h=1}^H$

  end for

  Use the fresh within-epoch fold together with the frozen witnesses induced by $\widehat{\mathcal{T}}_{h,f}^{\sigma,(e)}$ to define the next validation statistics $\widehat{\mathrm{BV}}_{h,e+1}(f)$

end for

return $\{(f^e, \pi^e, \nu^e)\}_{e\geqslant 1}$

in the raw online validation rule while preserving the same optimistic planning and exploiter-mix structure.

Algorithm 1 is the template analyzed by the main structured scalable linear theorem. The appendix further studies two structured extensions: a projected version, where exact closure is replaced by stagewise projection and explicit bias terms, and a self-normalized pure-online version, where the raw validation score is replaced by a self-normalized statistic with ridge dual refits.

Epochwise/cross-fit validation is introduced for statistical rather than algorithmic reasons. If the same data are used both to refit the robust plug-in operator and to validate the induced witness, the resulting witness class adapts too tightly to the data stream. Freezing the operator at the start of each epoch and validating it on fresh within-epoch samples breaks this feedback loop while preserving the same optimistic planning rule and exploiter-mix data collection.

The main hyperparameters have transparent roles: increasing p can improve strategic coverage and witness excitation, but incurs an explicit exploration cost and the factor $1/(1-p)$ in the regret bound; the validation radii $\beta_{h,e}$ control how conservatively functions are retained in the confidence set, and the epoch schedule controls the tradeoff between frequent operator refits and statistically separated validation. From a computational perspective, the extra work relative to a matched non-robust backbone comes from stagewise dual refits and repeated zero-sum matrix-game solves; in the linear specialization, the former reduces to ridge problems and the latter to $|\mathcal{A}| \times |\mathcal{B}|$ matrix games.

4 Theory

This section states the main assumptions and the principal regret guarantees. Full proofs are deferred to the appendix. In addition to the results stated here, the appendix establishes a self-normalized pure-online theorem for linear and finite-rank kernel classes. Throughout, $\mathrm{Reg}(K)$ denotes the signed cumulative one-sided comparator difference relative to the robust minimax value defined in the preliminaries.

4.1 Main-Text Assumptions

Assumption 1 (Robust Bellman closure)**.** *For the exact-closure results, the robust Bellman operator preserves the working class: $\mathcal{T}_h^\sigma f_{h+1} \in \mathcal{F}_h$ for every $h \in [H]$ and $f \in \mathcal{F}$. For projected results, this requirement is replaced by stagewise projectors together with explicit projection and operator-remainder terms.*

Assumption 2 (Dual learnability)**.** *For a probability measure ζ on $\mathcal{S} \times \mathcal{A} \times \mathcal{B}$, let*

$$\mathcal{L}_h^\zeta(g; f) := \mathbb{E}_{(s,a,b)\sim\zeta} \mathbb{E}_{s'\sim P_h^0(\cdot|s,a,b)} \left[\ell_\phi(f_{h+1}; s, a, b, s'; g)\right].$$

For every h, $f \in \mathcal{F}$, and relevant ζ, the working dual class satisfies

$$\inf_{g\in\mathcal{G}_h} \mathcal{L}_h^\zeta(g; f) \leqslant \inf_{\bar{g}\in\bar{\mathcal{G}}_h} \mathcal{L}_h^\zeta(\bar{g}; f) + \epsilon_{\mathrm{dual}}, \tag{4}$$

where $\bar{\mathcal{G}}_h$ contains the exact robust dual optimizer. The dual loss is bounded and Lipschitz and the working classes admit the finite covers used for uniform concentration; the corresponding constants are stated in Assumptions 10–12.

Assumption 3 (Support preservation and occupancy transfer)**.** *Measurable worst-case selectors exist, remain inside nominally learnable support, and induce the stage-h executed occupancy $\widetilde{\mu}_h^k$. Let the pooled past data-collection occupancy be $\bar{\widetilde{\eta}}_{h,k} := \sum_{j<k} \lambda_{h,j}^{(k)} \widetilde{\eta}_h^j$, where $\lambda_{h,j}^{(k)}$ is proportional to the stage-h sample count from episode j. For every h and $k \geqslant 2$,*

$$\widetilde{\mu}_h^k \ll \bar{\widetilde{\eta}}_{h,k}, \qquad \frac{d\widetilde{\mu}_h^k}{d\bar{\widetilde{\eta}}_{h,k}} \leqslant C_h^{\mathrm{mix}} \quad \bar{\widetilde{\eta}}_{h,k}\text{-almost surely}. \tag{5}$$

A concrete sufficient condition is useful for interpreting Assumption 3. If each executed policy contains a smoothing distribution $u_h(a, b \mid s)$ with weight at least $\gamma_{\min}$, $\underline{u}_h := \inf_{s,a,b} u_h(a, b \mid s) > 0$, and the current state marginal is dominated by the pooled past marginal with factor $C_{S,h}$, then

$$C_h^{\mathrm{mix}} \leqslant \frac{C_{S,h}}{\gamma_{\min}\underline{u}_h}. \tag{6}$$

The robustness radius σ affects this coefficient through the worst-case state marginal and hence $C_{S,h} = C_{S,h}(\sigma)$. Without an additional lower-mass or mixing condition there is no universal radius-only or monotone law for C_h^{mix}.

Assumption 4 (Inexact stage-game solution)**.** *For every episode k, stage h, and state s, the extracted pair satisfies*

$$V_{f^k,h}(s) \leqslant \mathbb{E}_{a\sim\pi_h^k(\cdot|s),\, b\sim\nu_h^k(\cdot|s)} \left[f_h^k(s, a, b)\right] + \epsilon_{\mathrm{eq}}. \tag{7}$$

The matrix-game value is 1-Lipschitz under the entrywise sup norm. Thus extraction contributes at most $H\epsilon_{\mathrm{eq}}$ per episode.

Assumptions 1–4 separate structural requirements (closure and stage-game solution) from statistical ones (dual learning and occupancy transfer). In spirit, they are robust zero-sum counterparts of the realizability/completeness/regularity assumptions standard in general-function-approximation RL and Markov games [7, 8, 6]. The additional ingredients here are occupancy transfer from nominal interaction to worst-case evaluation and Lipschitz control of approximate stage-game solves. The abstract reduction depends only on residual and operator budgets; concrete classes differ in how they verify them. Theorems 1–2 are conditional GFA interfaces, Theorem 3 closes the budgets under linear coverage/burn-in conditions, and Theorem 4 exposes the RKHS rank–bias trade-off.

4.2 Main Results

We first state the general regret guarantee. Let $d_{\text{rob-MG}} = \dim_{\mathrm{BE},2}^{\text{rob-MG}}(\mathcal{F}, \Pi, 1/\sqrt{K})$ denote the square-loss robust Markov-game Bellman–Eluder dimension, where $\Pi = (\Pi_h)_{h=1}^H$ is the stagewise occupancy family induced by the executed trajectories. As summarized in the preliminaries and verified in the appendix, this reduces to tabular counting in finite games and to a linear witness dimension, up to logarithmic factors, in linear classes. Here $B_\phi(\sigma)$ is the dual-loss envelope (specialized in the appendix for TV, KL, and χ^2). Throughout, ϵ_{dual} denotes dual approximation error, ϵ_{eq} stage-game solution error, and ϵ_{appr} the comparator or model-misspecification term, which vanishes under exact realizability.

Theorem 1 (General regret guarantee)**.** *Suppose Assumptions 1–4 hold, and that the backward-validation and dual-learning arguments provide optimism, a residual budget of order $\widetilde{O}(HB_\phi^2(\sigma)d_{\text{rob-MG}})$, and an operator-learning budget of order $\widetilde{O}(HB_\phi(\sigma)\sqrt{K})$. Then, with probability at least $1-\delta$,*

$$\mathrm{Reg}(K) \lesssim H + \frac{1}{1-p}\left[\widetilde{O}\Big(HB_\phi(\sigma)\sqrt{d_{\text{rob-MG}}K}\Big) + HK\epsilon_{\text{dual}} + HK\epsilon_{\text{eq}} + K\epsilon_{\text{appr}} + pHK\right].$$

Theorem 1 is the compressed main-text form of the appendix statistical-input corollary: residual complexity drives the leading statistical term, while dual approximation, stage-game error, and model misspecification remain additive and scale linearly in K. A sufficient regime for the displayed upper bound to be $o(K)$ is

$$\begin{aligned} p = p_K = o(1), \qquad \epsilon_{\text{dual}}, \epsilon_{\text{eq}}, \epsilon_{\text{appr}} = o(1), \\ B_\phi(\sigma)\sqrt{d_{\text{rob-MG}}K} = o(K). \end{aligned} \tag{8}$$

Under exact realizability, exact dual specification, and exact stage-game solving, the three error terms vanish. If the coverage and excitation constants stay bounded independently of p, then $p_K \asymp K^{-1/2}$ preserves a $\widetilde{O}(\sqrt{K})$-type rate.

For broader classes, exact closure may be replaced by projected closure. We write $R_{\text{BV,proj}}(K)$ and $R_{\text{dual,proj}}(K,\delta)$ for the projected residual-validation and projected dual/transfer contributions, respectively; the latter absorbs the projected operator-learning term together with the dual approximation component that remains explicit in the appendix.

Theorem 2 (Projected function-approximation guarantee)**.** *Suppose Assumptions 2–4 hold, and replace Assumption 1 by stagewise projectors $\{\mathsf{P}_h\}_{h=1}^H$ with projection bias ϵ_{proj} and operator remainder $\epsilon_{\text{op,proj}}$. If the corresponding projected residual and operator budgets are available, then, with probability at least $1-\delta$,*

$$\begin{aligned} \mathrm{Reg}(K) \leqslant H + \frac{1}{1-p}\Big[&R_{\text{BV,proj}}(K) + R_{\text{dual,proj}}(K,\delta) + HK\epsilon_{\text{proj}} + HK\epsilon_{\text{op,proj}} \\ &+ K\epsilon_{\text{appr}} + HK\epsilon_{\text{eq}} + pHK\Big]. \end{aligned}$$

Theorem 2 is the compressed projected counterpart of the appendix projected corollary: $R_{\text{BV,proj}}$ and $R_{\text{dual,proj}}$ abbreviate the explicit projected residual and operator/dual bookkeeping terms expanded there, while departures from exact closure are recorded through ϵ_{proj} and $\epsilon_{\text{op,proj}}$.

Theorem 3 (Epochwise/cross-fit linear theorem)**.** *Assume TV uncertainty, a linear-mixture nominal kernel with augmented features, affine continuation coupling for the dual fit, and a predictable covariance budget after burn-in. Here d_{lin} denotes the resulting linear robust MG-BE dimension, B_F is the uniform bound on the value class, κ_{lin} is the witness feature norm bound, Ψ_{dual} is the dual estimation scale from the ridge/self-normalized analysis, $\underline{\nu}_{\min} := \min_h \underline{\nu}_h/2$ is the post-burn-in covariance lower bound, and c_{ep} is the geometric-epoch constant. If $d_{W,h}$ and $\kappa_{W,h}$ are the stagewise witness dimension and feature-norm bound and M is the number of epochs, define*

$$N_{\text{burn}}(\delta) := \max_{h\in[H]} \left\lceil \frac{64\kappa_{W,h}^4 \log(2d_{W,h}HM/\delta)}{\underline{\nu}_h^2} \right\rceil. \tag{9}$$

Then the epochwise/cross-fit algorithm satisfies, with probability at least $1-\delta$,

$$\begin{aligned} \mathrm{Reg}(K) \leqslant c_{\text{ep}}HN_{\text{burn}}(\delta) + \frac{1}{1-p}\Bigg[&\widetilde{O}\Big(H\Big(B_F + \kappa_{\text{lin}}\Psi_{\text{dual}}/\sqrt{\underline{\nu}_{\min}}\Big)\sqrt{d_{\text{lin}}K}\Big) \\ &+ K\epsilon_{\text{appr}} + HK\epsilon_{\text{eq}} + pHK\Bigg]. \end{aligned}$$

Here N_{burn} is an analysis threshold over the actual run, not a separate warm-up; early episodes cost at most $c_{\text{ep}}HN_{\text{burn}}$. Under BR-only coverage, $\underline{\nu}_h \propto p$ gives costs $\sqrt{K/p}$, p^{-2}, and pK, balanced by $p_K \asymp K^{-1/3}$ for a $\widetilde{O}(K^{2/3})$-type bound. Adapting to unknown coverage or excitation constants is outside the analysis.

Table 2: Main `linear_mixture` results at $d = 8$ and 5000 episodes.

Method	Proxy	Deploy	Proxy gap
SIGMA0	0.230 ± 0.088	0.230 ± 0.088	0.0080 ± 0.0047
ODRB ($\beta = 0.30, \tau = 0.02$)	0.204 ± 0.069	0.239 ± 0.089	0.0072 ± 0.0074
ODRB ($\beta = 0.20, \tau = 0.02$)	0.162 ± 0.057	0.236 ± 0.089	0.0069 ± 0.0066
NOEX ($\beta = 0.30$)	0.254 ± 0.102	0.288 ± 0.115	0.0644 ± 0.0687

Theorem 4 (RKHS spectral-truncation projected theorem)**.** *Let the continuation, dual, and residual classes lie in stagewise RKHSs, and let $\Pi_{h,m}$ denote the rank-m spectral projector at stage h. Applying Theorem 2 to the truncated classes yields a truncation-dependent burn-in threshold*

$$N_{\mathrm{burn}}^{\mathrm{RKHS}}(m, \delta) := \max_{h \in [H]} \left\lceil \frac{64(\kappa_{W,h}^{(m)})^4}{(\gamma_{\min} \bar{\omega}_h^{(m)})^2} \log(2mHM/\delta) \right\rceil,$$

where $\kappa_{W,h}^{(m)}$ and $\bar{\omega}_h^{(m)}$ are the truncated witness-norm and excitation constants. Then

$$\begin{aligned} \mathrm{Reg}(K) \leqslant\ & H + c_{\mathrm{ep}} H N_{\mathrm{burn}}^{\mathrm{RKHS}}(m, \delta) \\ & + \frac{1}{1-p} \Big[\widetilde{O}\Big(H\sqrt{mK} \Big) + \\ & \quad HK\big(\epsilon_{\mathrm{proj}}(m) + \epsilon_{\mathrm{op,proj}}(m) + \epsilon_{\mathrm{dual}} + \epsilon_{\mathrm{eq}} \big) + K\epsilon_{\mathrm{appr}} + pHK \Big], \end{aligned}$$

where the $\widetilde{O}(H\sqrt{mK})$ term hides the truncated RKHS envelopes and self-normalized dual constants, while the projection terms are controlled by the spectral tail.

4.3 Proof Sketch

The confidence set retains a Bellman-consistent comparator, after which optimism and telescoping under the worst-case-selector occupancy $\mu_{h,\mathrm{wc}}^k$ give

$$\mathrm{Reg}(K) \lesssim \sum_{k=1}^{K} \sum_{h=1}^{H} \left\| f_h^k - \mathcal{T}_h^\sigma f_{h+1}^k \right\|_{1, \mu_{h,\mathrm{wc}}^k} + HK\epsilon_{\mathrm{eq}}. \tag{10}$$

Adding and subtracting the plug-in operator separates a residual (backward validation and Bellman–Eluder counting) from an operator gap (dual concentration and occupancy transfer). The game-specific proof fits the dual before each epoch, freezes and cross-validates its witnesses, transfers to worst-case occupancies, and derives covariance growth from mixed response branches.

5 Experiments

We compare ODRB with SIGMA0 and a no-exploiter ablation on linear-mixture, tabular, and support-shift games. The main $d = 8$ comparison uses ten paired seeds; details are in Appendix G.

For ODRB$(\beta = 0.30, \tau = 0.02)$ versus SIGMA0, the deployment-return difference is 0.0094 (paired-t 95% CI $[0.0007, 0.0182]$; exact sign-flip Holm $p = 0.0469$). For $(0.20, 0.02)$, it is 0.0065, but CI $[-0.0017, 0.0148]$ includes zero. This supports a modest primary-setting improvement, not uniform superiority; the higher-return $\tau = 0$ setting has a much larger gap.

6 Conclusion

ODRB gives conditional GFA interfaces and structured linear/RKHS guarantees for online robust zero-sum games under shared-payoff, support, and coverage assumptions. Its signed comparator is not an exploitability certificate, and the experiments are diagnostic. Player-specific ambiguity, adaptive tuning, and equilibrium convergence remain open.

Acknowledgments and Disclosure of Funding

Funding. There was no funding in direct support of this work.
Competing interests. The authors declare no competing interests.

References

[1] Garud N. Iyengar. Robust dynamic programming. *Mathematics of Operations Research*, 30(2):257–280, 2005.

[2] Arnab Nilim and Laurent El Ghaoui. Robust control of markov decision processes with uncertain transition matrices. *Operations Research*, 53(5):780–798, 2005.

[3] Wolfram Wiesemann, Daniel Kuhn, and Berc Rustem. Robust markov decision processes. *Mathematics of Operations Research*, 38(1):153–183, 2013.

[4] Kaiqing Zhang, Tao Sun, Yunzhe Tao, Sahika Genc, Sunil Mallya, and Tamer Basar. Robust multi-agent reinforcement learning with model uncertainty. In *Advances in Neural Information Processing Systems*, volume 33, pages 10571–10583, 2020.

[5] Zewu Zheng and Yuanyuan Lin. Distributionally robust online markov game with linear function approximation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 28892–28900, 2026.

[6] Debamita Ghosh, George K Atia, and Yue Wang. Scaling online distributionally robust reinforcement learning: Sample-efficient guarantees with general function approximation. *arXiv preprint arXiv:2512.18957*, 2025.

[7] Baihe Huang, Jason D. Lee, Zhaoran Wang, and Zhuoran Yang. Towards general function approximation in zero-sum markov games. In *International Conference on Learning Representations*, 2022.

[8] Chi Jin, Qinghua Liu, and Tiancheng Yu. The power of exploiter: Provable multi-agent RL in large state spaces. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 10251–10279. PMLR, 2022.

[9] Shaocong Ma, Ziyi Chen, Shaofeng Zou, and Yi Zhou. Decentralized robust v-learning for solving markov games with model uncertainty. *Journal of Machine Learning Research*, 24(371):1–40, 2023.

[10] Laixi Shi, Eric Mazumdar, Yuejie Chi, and Adam Wierman. Sample-efficient robust multi-agent reinforcement learning in the face of environmental uncertainty. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 44909–44959. PMLR, 2024. URL https://proceedings.mlr.press/v235/shi24d.html.

[11] Laixi Shi, Jingchu Gai, Eric Mazumdar, Yuejie Chi, and Adam Wierman. Breaking the curse of multiagency in robust multi-agent reinforcement learning. In *International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=fFBnsYRsPg.

[12] Huan Xu and Shie Mannor. Distributionally robust markov decision processes. In *Advances in Neural Information Processing Systems*, volume 23, pages 2505–2513, 2010.

[13] Shie Mannor, Ofir Mebel, and Huan Xu. Robust MDPs with k-rectangular uncertainty. *Mathematics of Operations Research*, 41(4):1484–1509, 2016.

[14] Janosch Moos, Kay Hansel, Hany Abdulsamad, Svenja Stark, Debora Clever, and Jan Peters. Robust reinforcement learning: A review of foundations and recent advances. *Machine Learning and Knowledge Extraction*, 4(1):276–315, 2022.

[15] Aviv Tamar, Shie Mannor, and Huan Xu. Scaling up robust MDPs using function approximation. In *Proceedings of the 31st International Conference on Machine Learning*, volume 32 of *Proceedings of Machine Learning Research*, pages 181–189. PMLR, 2014.

[16] Chin Pang Ho, Marek Petrik, and Wolfram Wiesemann. Partial policy iteration for l_1-robust markov decision processes. *Journal of Machine Learning Research*, 22(275):1–46, 2021.

[17] Kishan Panaganti and Dileep M. Kalathil. Sample complexity of robust reinforcement learning with a generative model. In *Proceedings of the 25th International Conference on Artificial Intelligence and Statistics*, volume 151 of *Proceedings of Machine Learning Research*, pages 9582–9602. PMLR, 2022.

[18] Laixi Shi, Gen Li, Yuting Wei, Yuxin Chen, Matthieu Geist, and Yuejie Chi. The curious price of distributional robustness in reinforcement learning with a generative model. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

[19] Yue Wang and Shaofeng Zou. Online robust reinforcement learning with model uncertainty. In *Advances in Neural Information Processing Systems*, volume 34, pages 7193–7206, 2021.

[20] Miao Lu, Han Zhong, Tong Zhang, and Jose H. Blanchet. Distributionally robust reinforcement learning with interactive data collection: Fundamental hardness and near-optimal algorithms. In *Advances in Neural Information Processing Systems*, volume 37, 2024.

[21] Laixi Shi and Yuejie Chi. Distributionally robust model-based offline reinforcement learning with near-optimal sample complexity. *Journal of Machine Learning Research*, 25(200):1–91, 2024.

[22] Zhishuai Liu and Pan Xu. Distributionally robust off-dynamics reinforcement learning: Provable efficiency with linear function approximation. In *Proceedings of the 27th International Conference on Artificial Intelligence and Statistics*, volume 238 of *Proceedings of Machine Learning Research*, pages 2719–2727. PMLR, 2024.

[23] Qiaomin Xie, Yudong Chen, Zhaoran Wang, and Zhuoran Yang. Learning zero-sum simultaneous-move markov games using function approximation and correlated equilibrium. In *Proceedings of the 33rd Conference on Learning Theory*, volume 125 of *Proceedings of Machine Learning Research*, pages 3674–3682. PMLR, 2020.

[24] Pedro Cisneros-Velarde and Sanmi Koyejo. Finite-sample guarantees for nash q-learning with linear function approximation. In *Proceedings of the 39th Conference on Uncertainty in Artificial Intelligence*, volume 216 of *Proceedings of Machine Learning Research*, pages 424–432. PMLR, 2023.

[25] Yuanhao Wang, Qinghua Liu, Yu Bai, and Chi Jin. Breaking the curse of multiagency: Provably efficient decentralized multi-agent RL with function approximation. In *Proceedings of the 36th Conference on Learning Theory*, volume 195 of *Proceedings of Machine Learning Research*, pages 2793–2848. PMLR, 2023.

[26] Chenhao Zhou, Zebang Shen, Chao Zhang, Hanbin Zhao, and Hui Qian. Solving zero-sum markov games with continuous state via spectral dynamic embedding. In *Advances in Neural Information Processing Systems*, volume 37, pages 68622–68657, 2024. doi: 10.52202/079017-2192.

[27] Daniel Russo and Benjamin Van Roy. Eluder dimension and the sample complexity of optimistic exploration. In *Advances in Neural Information Processing Systems*, volume 26, 2013.

[28] Gen Li, Pranav Kamath, Dylan J. Foster, and Nathan Srebro. Understanding the eluder dimension. In *Advances in Neural Information Processing Systems*, volume 35, pages 23737–23750, 2022.

[29] Nan Jiang, Akshay Krishnamurthy, Alekh Agarwal, John Langford, and Robert E. Schapire. Contextual decision processes with low bellman rank are PAC-learnable. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 1704–1713. PMLR, 2017.

[30] Wen Sun, Nan Jiang, Akshay Krishnamurthy, Alekh Agarwal, and John Langford. Model-based RL in contextual decision processes: PAC bounds and exponential improvements over model-free approaches. In *Conference on Learning Theory*, volume 99 of *Proceedings of Machine Learning Research*, pages 2898–2933. PMLR, 2019.

[31] Remi Munos. Error bounds for approximate policy iteration. In *Proceedings of the 20th International Conference on Machine Learning*, pages 560–567, 2003.

[32] Remi Munos and Csaba Szepesvari. Finite-time bounds for fitted value iteration. *Journal of Machine Learning Research*, 9:815–857, 2008.

[33] Jinglin Chen and Nan Jiang. Information-theoretic considerations in batch reinforcement learning. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 1042–1051. PMLR, 2019.

[34] Ying Jin, Zhuoran Yang, and Zhaoran Wang. Is pessimism provably efficient for offline RL? In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 5084–5096. PMLR, 2021.

[35] Tengyang Xie, Ching-An Cheng, Nan Jiang, Paul Mineiro, and Alekh Agarwal. Bellman-consistent pessimism for offline reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 34, pages 6683–6694, 2021.

[36] Chi Jin, Qinghua Liu, and Sobhan Miryoosefi. Bellman eluder dimension: New rich classes of RL problems, and sample-efficient algorithms. In *Advances in Neural Information Processing Systems*, volume 34, pages 13406–13418, 2021.

[37] Andrea Zanette, Alessandro Lazaric, Mykel Kochenderfer, and Emma Brunskill. Learning near optimal policies with low inherent bellman error. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 10978–10989. PMLR, 2020.

[38] Ruosong Wang, Ruslan Salakhutdinov, and Lin Yang. Reinforcement learning with general value function approximation: Provably efficient approach via bounded eluder dimension. In *Advances in Neural Information Processing Systems*, volume 33, pages 6123–6135, 2020.

[39] Tengyang Xie, Dylan J. Foster, Yu Bai, Nan Jiang, and Sham M. Kakade. The role of coverage in online reinforcement learning. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=LQIjzPdDt3q.

[40] Zain Ulabedeen Farhat, Debamita Ghosh, George K. Atia, and Yue Wang. Sample-efficient distributionally robust multi-agent reinforcement learning via online interaction. In *International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=hIBMCDOwMI.

[41] Fangyikang Wang, Hubery Yin, Yuejiang Dong, Huminhao Zhu, Chao Zhang, Hanbin Zhao, Hui Qian, and Chen Li. BELM: Bidirectional explicit linear multi-step sampler for exact inversion in diffusion models. In *Advances in Neural Information Processing Systems*, volume 37, pages 46118–46159, 2024. doi: 10.52202/079017-1466.

[42] Huminhao Zhu, Fangyikang Wang, Tianyu Ding, Qing Qu, and Zhihui Zhu. Analyzing and mitigating model collapse in reflow methods. In *Conference on Parsimony and Learning*, volume 328 of *Proceedings of Machine Learning Research*, pages 314–340. PMLR, 2026.

A Appendix Overview

This appendix expands the proof sketch from the main text into a step-by-step derivation. For clarity, the analysis is separated into two parts:

1. a deterministic regret decomposition, which we prove in full;
2. two statistical verification blocks, which we prove under explicit backward-validation and entropy assumptions.

The deterministic part already makes precise how the robust minimax Bellman operator, the dual plug-in backup, and the stage-game solver interact. The only missing ingredients are the high-probability control needed to instantiate the residual budget and the operator-learning budget for the exact algorithm and target function classes.

Appendix organization. Section B develops the deterministic reduction and its core lemmas. Section C collects the residual-counting, backward-validation, optimism-closure, and dual-concentration ingredients. Section D turns these inputs into closed structured specializations, including tabular, linear, pure-online structured, projected, and RKHS-truncation results. Section E briefly summarizes the main future directions suggested by the present proof package. Finally, Section G records the benchmark definitions, compared methods, evaluation metrics, statistical protocol, and additional result tables used by the theorem-aligned benchmark suite in the main text.

Status of the formal results. The appendix contains two different levels of claims, and we state this distinction explicitly. On the one hand, Theorem 48, Theorem 78, Theorem 81, Theorem 82, and Theorem 83 are structured results whose statistical budgets are verified inside the paper once their stated structural assumptions hold. In particular, Theorem 78 closes the epochwise/cross-fit linear route only after Assumption 20 is imposed; Proposition 72 provides a primitive sufficient condition for that hypothesis. On the other hand, Theorem 53 and Theorem 54 remain conditional broad-function-class interfaces: they record the regret consequences of explicit optimism, residual-budget, operator-learning, and coverage hypotheses, but they do not claim that those hypotheses have already been verified for the original one-update-per-episode backward-validation general-GFA route. That unresolved point is isolated later in Proposition 56. The paper therefore claims a structured theory package with explicit general-function-approximation interfaces, rather than a fully assumption-free broad closure.

Main-text theorem map. For ease of cross-checking, the four main-text theorem statements are compressed from the following appendix results: the General regret guarantee (Theorem 1) from Corollary 12, the Projected function-approximation guarantee (Theorem 2) from Corollary 55, the Epochwise/cross-fit linear theorem (Theorem 3) from Corollary 79, and the RKHS spectral-truncation projected theorem (Theorem 4) from Theorem 83. The appendix-level interface theorems Theorem 53 and Theorem 54 provide the more explicit residual-budget and operator-budget decompositions from which the two main-text interface statements are obtained.

Nominal-game reduction and comparison. When $\sigma = 0$, the ambiguity set is the singleton nominal kernel, the dual robust backup reduces to ordinary conditional expectation, and the robust residual class becomes the nominal Markov-game Bellman-residual class. Table 3 positions this reduction relative to the closest non-robust function-approximation results; the comparison concerns the problem structure and does not claim a better nominal rate.

B Deterministic Reduction Setup

B.1 Setup and Additional Standing Conditions

For clarity, we analyze the on-policy core of the algorithm. That is, in episode k the learner computes $f^{(k)} \in \mathcal{F}$, extracts a policy pair (π^k, ν^k) from $f^{(k)}$, and executes (π^k, ν^k) for one episode. The exploiter-mix variant changes only the occupancy family that appears in the statistical counting lemma; the algebra below is unchanged.

Table 3: Relationship to representative non-robust zero-sum Markov-game results.

Work	Transition model	Complexity / exploration	Distinguishing target
Huang et al. [7]	Nominal	Minimax-Eluder complexity	Value regret under general function approximation
Jin et al. [8]	Nominal	Multi-agent BE dimension and exploiter	Nash-policy learning in large state spaces
This paper, $\sigma = 0$	Nominal	Robust MG-BE reduces to nominal counterpart	Same $\sqrt{K}$-type episode dependence, up to complexity, coverage, horizon, and logs
This paper, $\sigma > 0$	Ambiguity set around nominal kernel	Response-branch coverage plus operator transfer	Candidate-dependent robust dual learned from nominal transitions and evaluated on worst-case occupancies

Standing selector hypothesis. Whenever the analysis refers to a worst-case executed occupancy, we assume that for every Markov policy pair and stage h the inner infimum in the robust Bellman backup admits a measurable selector attaining the robust value. This makes the occupancy process μ_h^k well defined. In finite / tabular settings this is automatic; more generally, one could replace attainment by an ϵ_{sel}-optimal measurable selector at the cost of an additional $O(HK\epsilon_{\mathrm{sel}})$ term in the final regret bounds.

Throughout the deterministic Bellman telescope, μ_h^k denotes the stage-h occupancy of the executed policy pair under a worst-case selector that attains the robust value of that pair. For the statistical lemmas in the present on-policy core, we idealize the stage-h data batch as having state-action marginal μ_h^k and next-state conditional law P_h^0. Section A.11 later separates these two roles: it keeps $\widetilde{\mu}_h^k$ for the worst-case executed occupancy used by the regret telescope, introduces a distinct data-collection occupancy for the empirical-process arguments, and adds an explicit transfer assumption between them.

For each episode k and stage h, define the exact robust residual and the operator error

$$\begin{aligned}
\delta_h^k(s,a,b) &:= f_h^{(k)}(s,a,b) - (\mathcal{T}_h^\sigma f_{h+1}^{(k)})(s,a,b),\\
\widehat{\delta}_h^k(s,a,b) &:= f_h^{(k)}(s,a,b) - \big(\widehat{\mathcal{T}}_{h,f^{(k)}}^{\sigma,(k)} f_{h+1}^{(k)}\big)(s,a,b),\\
e_h^k(s,a,b) &:= \big(\widehat{\mathcal{T}}_{h,f^{(k)}}^{\sigma,(k)} f_{h+1}^{(k)}\big)(s,a,b) - (\mathcal{T}_h^\sigma f_{h+1}^{(k)})(s,a,b),
\end{aligned}$$

so that $\delta_h^k = \widehat{\delta}_h^k + e_h^k$ pointwise.

We will use the following three input conditions.

Assumption 5 (Approximate optimism)**.** *For every episode k,*

$$V_1^{\star,\sigma}(s_1) \leqslant V_{f^{(k)},1}(s_1) + \epsilon_{\mathrm{appr}}.$$

Assumption 6 (Operator-learning budget)**.** *There exists a nonnegative array $\{\alpha_{h,k}\}$ such that for every k and h,*

$$\big\|e_h^k\big\|_{1,\mu_h^k} \leqslant \alpha_{h,k} + \epsilon_{\mathrm{dual}},$$

where μ_h^k is the stage-h occupancy measure induced by the executed policy pair in episode k under a worst-case kernel selector for that pair.

Assumption 7 (Residual budget)**.** *There exists $\Gamma_{\mathrm{res}}(K) \geqslant 0$ such that*

$$\sum_{k=1}^{K}\sum_{h=1}^{H} \mathbb{E}_{\mu_h^k}\Big[\big(\widehat{\delta}_h^k(s,a,b)\big)^2\Big] \leqslant \Gamma_{\mathrm{res}}(K).$$

Assumptions 6 and 7 are exactly the two places where the robust Markov-game extension of the statistical machinery enters. The approximate-optimism assumption is also not meant to be primitive: later in this appendix we show how it follows from retaining a suitable reference function inside the backward-validation confidence set.

B.2 Auxiliary Lemmas

Lemma 5 (Minimax value is 1-Lipschitz)**.** *Let $M, N \in \mathbb{R}^{|\mathcal{A}|\times|\mathcal{B}|}$ be two payoff matrices and define* $\mathrm{val}(M) := \max_{p\in\Delta(\mathcal{A})} \min_{q\in\Delta(\mathcal{B})} p^\top M q$*. Then*

$$\left|\mathrm{val}(M) - \mathrm{val}(N)\right| \leqslant \|M - N\|_\infty.$$

Proof. Fix arbitrary mixed strategies $p \in \Delta(\mathcal{A})$ and $q \in \Delta(\mathcal{B})$. Since p, q are probability vectors,

$$\left|p^\top (M - N) q\right| \leqslant \sum_{a,b} p(a) q(b) \left|M_{ab} - N_{ab}\right| \leqslant \|M - N\|_\infty.$$

Therefore, for every p,

$$\min_q p^\top M q \leqslant \min_q p^\top N q + \|M - N\|_\infty.$$

Taking the maximum over p yields

$$\mathrm{val}(M) \leqslant \mathrm{val}(N) + \|M - N\|_\infty.$$

Interchanging the roles of M and N gives the reverse inequality. □

Lemma 6 (Approximate stage-game extraction)**.** *Suppose that for every (h, s), the extracted policy pair $(\pi_h^k(s), \nu_h^k(s))$ is ϵ_{eq}-optimal for the stage game induced by $f_h^{(k)}$, in the sense that*

$$V_{f^{(k)},h}(s) \leqslant \mathbb{E}_{a\sim\pi_h^k(\cdot|s),\, b\sim\nu_h^k(\cdot|s)} \left[f_h^{(k)}(s, a, b)\right] + \epsilon_{\mathrm{eq}}. \tag{11}$$

Then the stage-game extraction error contributes at most $H\epsilon_{\mathrm{eq}}$ over one episode.

Proof. Equation (11) gives at most ϵ_{eq} loss at each stage. Summing over the H stages of one episode gives the result. □

Lemma 7 (One-step robust Bellman comparison)**.** *Fix an episode k and abbreviate $f = f^{(k)}$, $\pi = \pi^k$, and $\nu = \nu^k$. Let $P_h^\dagger(\cdot \mid s, a, b)$ be a measurable selector that attains the robust value of the executed policy pair (π, ν) at stage h. Define*

$$\Delta_h(s) := V_{f,h}(s) - V_h^{\pi,\nu,\sigma}(s).$$

Then for every state s and stage h,

$$\Delta_h(s) \leqslant \epsilon_{\mathrm{eq}} + \mathbb{E}\left[\delta_h^k(s_h, a_h, b_h) + \Delta_{h+1}(s_{h+1}) \mid s_h = s\right], \tag{12}$$

where the conditional expectation is taken under $a_h \sim \pi_h(\cdot \mid s_h)$, $b_h \sim \nu_h(\cdot \mid s_h)$, and $s_{h+1} \sim P_h^\dagger(\cdot \mid s_h, a_h, b_h)$.

Proof. By Lemma 6,

$$V_{f,h}(s) \leqslant \mathbb{E}_{a,b}\left[f_h(s, a, b)\right] + \epsilon_{\mathrm{eq}}.$$

Subtracting the robust value of the executed policy gives

$$\Delta_h(s) \leqslant \epsilon_{\mathrm{eq}} + \mathbb{E}_{a,b}\left[f_h(s, a, b) - r_h(s, a, b) - \mathbb{E}_{s'\sim P_h^\dagger(\cdot|s,a,b)}\left[V_{h+1}^{\pi,\nu,\sigma}(s')\right]\right].$$

Add and subtract $\mathbb{E}_{s'\sim P_h^\dagger(\cdot|s,a,b)}\left[V_{f,h+1}(s')\right]$ inside the bracket:

$$\begin{aligned}\Delta_h(s) \leqslant \epsilon_{\mathrm{eq}} &+ \mathbb{E}_{a,b}\left[f_h(s, a, b) - r_h(s, a, b) - \mathbb{E}_{s'\sim P_h^\dagger}\left[V_{f,h+1}(s')\right]\right] \\ &+ \mathbb{E}_{a,b,s'}\left[V_{f,h+1}(s') - V_{h+1}^{\pi,\nu,\sigma}(s')\right].\end{aligned}$$

Because $(\mathcal{T}_h^\sigma f_{h+1})(s, a, b) = r_h(s, a, b) + \inf_{P\in\mathcal{U}_h(s,a,b)} \mathbb{E}_{s'\sim P}\left[V_{f,h+1}(s')\right]$, we have

$$r_h(s, a, b) + \inf_{P\in\mathcal{U}_h(s,a,b)} \mathbb{E}_{s'\sim P}\left[V_{f,h+1}(s')\right] \leqslant r_h(s, a, b) + \mathbb{E}_{s'\sim P_h^\dagger}\left[V_{f,h+1}(s')\right].$$

Hence

$$f_h(s, a, b) - r_h(s, a, b) - \mathbb{E}_{s'\sim P_h^\dagger}\left[V_{f,h+1}(s')\right] \leqslant f_h(s, a, b) - (\mathcal{T}_h^\sigma f_{h+1})(s, a, b) = \delta_h^k(s, a, b).$$

Substituting this bound back proves (12). □

Lemma 8 (Bellman telescoping inequality)**.** *Under the setting of Lemma 7,*

$$V_{f,1}(s_1) - V_1^{\pi,\nu,\sigma}(s_1) \leqslant \sum_{h=1}^{H} \mathbb{E}_{\mu_h^k}\left[\delta_h^k(s,a,b)\right] + H\epsilon_{\mathrm{eq}}.$$

Proof. Starting from (12) at $h = 1$ and taking expectations along the trajectory under the executed policy pair and the worst-case selector yields

$$\mathbb{E}\left[\Delta_1(s_1)\right] \leqslant \epsilon_{\mathrm{eq}} + \mathbb{E}_{\mu_1^k}\left[\delta_1^k\right] + \mathbb{E}\left[\Delta_2(s_2)\right].$$

Iterating the same argument for $h = 2, \ldots, H$ and using $\Delta_{H+1} \equiv 0$ gives

$$\mathbb{E}\left[\Delta_1(s_1)\right] \leqslant \sum_{h=1}^{H} \mathbb{E}_{\mu_h^k}\left[\delta_h^k\right] + H\epsilon_{\mathrm{eq}}.$$

Since s_1 is deterministic, $\mathbb{E}[\Delta_1(s_1)] = V_{f,1}(s_1) - V_1^{\pi,\nu,\sigma}(s_1)$. □

Lemma 9 (Residual and operator decomposition)**.** *For every episode k,*

$$\sum_{h=1}^{H} \mathbb{E}_{\mu_h^k}\left[\delta_h^k\right] \leqslant \sum_{h=1}^{H} \mathbb{E}_{\mu_h^k}\left[\widehat{\delta}_h^k\right] + \sum_{h=1}^{H} \mathbb{E}_{\mu_h^k}\left[|e_h^k|\right].$$

Proof. The identity $\delta_h^k = \widehat{\delta}_h^k + e_h^k$ holds pointwise. Taking expectations under μ_h^k and applying the triangle inequality to the operator-error term gives the claim. □

B.3 Main Deterministic Reduction

Theorem 10 (Deterministic regret reduction)**.** *Under Assumptions 5–7,*

$$\mathrm{Reg}(K) \leqslant K\epsilon_{\mathrm{appr}} + HK\epsilon_{\mathrm{eq}} + \sqrt{KH\,\Gamma_{\mathrm{res}}(K)} + \sum_{k=1}^{K}\sum_{h=1}^{H} \alpha_{h,k} + HK\epsilon_{\mathrm{dual}}. \tag{13}$$

Proof. Fix an episode k. By Assumption 5,

$$V_1^{\star,\sigma}(s_1) - V_1^{\pi^k,\nu^k,\sigma}(s_1) \leqslant \epsilon_{\mathrm{appr}} + V_{f^{(k)},1}(s_1) - V_1^{\pi^k,\nu^k,\sigma}(s_1).$$

Applying Lemma 8 and then Lemma 9 yields

$$V_1^{\star,\sigma}(s_1) - V_1^{\pi^k,\nu^k,\sigma}(s_1) \leqslant \epsilon_{\mathrm{appr}} + H\epsilon_{\mathrm{eq}} + \sum_{h=1}^{H} \mathbb{E}_{\mu_h^k}\left[\widehat{\delta}_h^k\right] + \sum_{h=1}^{H} \mathbb{E}_{\mu_h^k}\left[|e_h^k|\right].$$

Summing over $k = 1, \ldots, K$ gives

$$\mathrm{Reg}(K) \leqslant K\epsilon_{\mathrm{appr}} + HK\epsilon_{\mathrm{eq}} + \sum_{k,h} \mathbb{E}_{\mu_h^k}\left[\widehat{\delta}_h^k\right] + \sum_{k,h} \mathbb{E}_{\mu_h^k}\left[|e_h^k|\right]. \tag{14}$$

We now control the two sums in (14). By Cauchy–Schwarz,

$$\sum_{k,h} \mathbb{E}_{\mu_h^k}\left[\widehat{\delta}_h^k\right] \leqslant \sum_{k,h} \sqrt{\mathbb{E}_{\mu_h^k}\left[(\widehat{\delta}_h^k)^2\right]} \leqslant \sqrt{KH \sum_{k,h} \mathbb{E}_{\mu_h^k}\left[(\widehat{\delta}_h^k)^2\right]} \leqslant \sqrt{KH\,\Gamma_{\mathrm{res}}(K)},$$

where the last step uses Assumption 7. For the operator error, Assumption 6 gives

$$\sum_{k,h} \mathbb{E}_{\mu_h^k}\left[|e_h^k|\right] \leqslant \sum_{k,h} \big(\alpha_{h,k} + \epsilon_{\mathrm{dual}}\big) = \sum_{k,h} \alpha_{h,k} + HK\epsilon_{\mathrm{dual}}.$$

Substituting both bounds into (14) proves (13). □

We will also need a projected variant of the same reduction. Fix arbitrary stagewise maps $\mathsf{P}_h : L_\infty(\mathcal{S} \times \mathcal{A} \times \mathcal{B}) \to L_\infty(\mathcal{S} \times \mathcal{A} \times \mathcal{B})$. For each episode k and stage h, define the projected residual, projected plug-in residual, projected operator gap, and projection bias by

$$\begin{aligned}
\delta^k_{h,\mathsf{P}} &:= f^{(k)}_h - \mathsf{P}_h \mathcal{T}^\sigma_h f^{(k)}_{h+1}, \\
\widehat{\delta}^k_{h,\mathsf{P}} &:= f^{(k)}_h - \mathsf{P}_h \widehat{\mathcal{T}}^{\sigma,(k)}_{h,f^{(k)}} f^{(k)}_{h+1}, \\
e^k_{h,\mathsf{P}} &:= \mathsf{P}_h \widehat{\mathcal{T}}^{\sigma,(k)}_{h,f^{(k)}} f^{(k)}_{h+1} - \mathsf{P}_h \mathcal{T}^\sigma_h f^{(k)}_{h+1}, \\
b^k_{h,\mathsf{P}} &:= \mathsf{P}_h \mathcal{T}^\sigma_h f^{(k)}_{h+1} - \mathcal{T}^\sigma_h f^{(k)}_{h+1}.
\end{aligned}$$

Then $\delta^k_h = \delta^k_{h,\mathsf{P}} + b^k_{h,\mathsf{P}}$ and $\delta^k_{h,\mathsf{P}} = \widehat{\delta}^k_{h,\mathsf{P}} + e^k_{h,\mathsf{P}}$ pointwise.

Theorem 11 (Deterministic regret reduction with projected completeness bias)**.** *Fix stagewise maps $\{\mathsf{P}_h\}_{h=1}^H$. Assume approximate optimism, and suppose there exists a residual budget $\Gamma^{\mathsf{P}}_{\mathrm{res}}(K) \geqslant 0$, and a projection-bias constant $\epsilon_{\mathrm{proj}} \geqslant 0$ such that*

$$\sum_{k=1}^K \sum_{h=1}^H \mathbb{E}_{\mu^k_h}\left[\left(\delta^k_{h,\mathsf{P}}\right)^2\right] \leqslant \Gamma^{\mathsf{P}}_{\mathrm{res}}(K), \tag{15}$$

$$\left\|b^k_{h,\mathsf{P}}\right\|_{1,\mu^k_h} \leqslant \epsilon_{\mathrm{proj}} \qquad \textit{for every } (h,k). \tag{16}$$

Then

$$\mathrm{Reg}(K) \leqslant K\epsilon_{\mathrm{appr}} + HK\epsilon_{\mathrm{eq}} + \sqrt{KH\,\Gamma^{\mathsf{P}}_{\mathrm{res}}(K)} + HK\epsilon_{\mathrm{proj}}. \tag{17}$$

Proof. Fix an episode k. By Assumption 5,

$$V^{\star,\sigma}_1(s_1) - V^{\pi^k,\nu^k,\sigma}_1(s_1) \leqslant \epsilon_{\mathrm{appr}} + V_{f^{(k)},1}(s_1) - V^{\pi^k,\nu^k,\sigma}_1(s_1).$$

Applying Lemma 8 gives

$$V^{\star,\sigma}_1(s_1) - V^{\pi^k,\nu^k,\sigma}_1(s_1) \leqslant \epsilon_{\mathrm{appr}} + H\epsilon_{\mathrm{eq}} + \sum_{h=1}^H \mathbb{E}_{\mu^k_h}\left[\delta^k_h\right].$$

Using the pointwise decomposition $\delta^k_h = \delta^k_{h,\mathsf{P}} + b^k_{h,\mathsf{P}}$, we obtain

$$V^{\star,\sigma}_1(s_1) - V^{\pi^k,\nu^k,\sigma}_1(s_1) \leqslant \epsilon_{\mathrm{appr}} + H\epsilon_{\mathrm{eq}} + + \sum_{h=1}^H \mathbb{E}_{\mu^k_h}\left[\delta^k_{h,\mathsf{P}}\right] + \sum_{h=1}^H \mathbb{E}_{\mu^k_h}\left[|b^k_{h,\mathsf{P}}|\right].$$

Summing over $k = 1, \ldots, K$ and applying Cauchy–Schwarz to the projected residual term yields

$$\sum_{k=1}^K \sum_{h=1}^H \mathbb{E}_{\mu^k_h}\left[\delta^k_{h,\mathsf{P}}\right] \leqslant \sqrt{KH \sum_{k=1}^K \sum_{h=1}^H \mathbb{E}_{\mu^k_h}\left[\left(\delta^k_{h,\mathsf{P}}\right)^2\right]} \leqslant \sqrt{KH\,\Gamma^{\mathsf{P}}_{\mathrm{res}}(K)}$$

by (15). Equation (16) yields

$$\sum_{k=1}^K \sum_{h=1}^H \mathbb{E}_{\mu^k_h}\left[|b^k_{h,\mathsf{P}}|\right] \leqslant HK\epsilon_{\mathrm{proj}}.$$

Combining the last two displays proves (17). □

Corollary 12 (Main-text regret bound from statistical inputs)**.** *Suppose in addition that there exist absolute constants $C_1, C_2 > 0$ such that*

$$\begin{aligned}
\Gamma_{\mathrm{res}}(K) &\leqslant C_1 H B^2_\phi(\sigma)\, \dim^{\text{rob-MG}}_{\mathrm{BE},2}(\mathcal{F}, \Pi, 1/\sqrt{K})\, \log(KH/\delta), \\
\sum_{k=1}^K \sum_{h=1}^H \alpha_{h,k} &\leqslant C_2 H B_\phi(\sigma)\sqrt{K}\, \log(KH/\delta).
\end{aligned}$$

Then Theorem 1 follows up to absolute constants and logarithmic factors.

Proof. Plug the two displayed bounds into Theorem 10. The residual term becomes

$$\sqrt{KH\,\Gamma_{\mathrm{res}}(K)} \leqslant \widetilde{O}\left(HB_\phi(\sigma)\sqrt{\dim_{\mathrm{BE},2}^{\text{rob-MG}}(\mathcal{F},\Pi,1/\sqrt{K})\,K}\right).$$

The operator-learning term $\sum_{k,h}\alpha_{h,k}$ is of order $\widetilde{O}(HB_\phi(\sigma)\sqrt{K})$, which is absorbed by the previous term because $\dim_{\mathrm{BE},2}^{\text{rob-MG}}(\mathcal{F},\Pi,1/\sqrt{K}) \geqslant 1$ in any nontrivial problem. The remaining terms match the main-text statement. □

Iterate sample-complexity consequence. Under the assumptions of Corollary 12, there exists at least one iterate $k \in [K]$ such that

$$V_1^{\star,\sigma}(s_1) - V_1^{\pi^k,\nu^k,\sigma}(s_1) \leqslant \frac{\mathrm{Reg}(K)}{K}.$$

Consequently, if $\epsilon_{\mathrm{appr}} + H\epsilon_{\mathrm{eq}} + H\epsilon_{\mathrm{dual}}$ is at most order ϵ, then an iterate whose one-sided comparator difference is at most order ϵ is guaranteed once

$$K = \widetilde{O}\left(\frac{H^2 B_\phi^2(\sigma)\dim_{\mathrm{BE},2}^{\text{rob-MG}}(\mathcal{F},\Pi,1/\sqrt{K})}{\epsilon}^2\right).$$

If every iterate had comparator difference strictly larger than $\mathrm{Reg}(K)/K$, then summing over all K iterates would contradict the definition of regret; the displayed sample-complexity condition is obtained by requiring each term in Theorem 1 divided by K to be at most order ϵ.

C Residual Complexity and Statistical Verification

C.1 Finite-State Fail-State Support-Control Argument

We now prove a concrete finite-state support-preservation statement. The proof only uses the failure-state geometry and does not depend on Bellman–Eluder arguments.

Proposition 13 (Support-preserving optimizer under TV balls)**.** *Assume a finite state space and fix a tuple (h,s,a,b). Let $P^0 \in \Delta(\mathcal{S})$ be the nominal next-state distribution and let $s_F \in \mathrm{supp}(P^0)$ be a failure state satisfying $V(s_F) = \min_{x\in\mathcal{S}} V(x) = 0$. Consider the optimization problem*

$$\inf_{P\in\Delta(\mathcal{S}):\, D_{\mathrm{TV}}(P,P^0)\leqslant\sigma} \mathbb{E}_{x\sim P}\left[V(x)\right].$$

Then there exists an optimal solution $P^\star$ such that $\mathrm{supp}(P^\star) \subseteq \mathrm{supp}(P^0)$.

Proof. Let Q be any feasible distribution. If $\mathrm{supp}(Q) \subseteq \mathrm{supp}(P^0)$, there is nothing to prove. Otherwise define the outside-support mass

$$m := \sum_{x\notin\mathrm{supp}(P^0)} Q(x) > 0.$$

Construct a new distribution $\widetilde{Q}$ by moving all outside-support mass to the failure state:

$$\widetilde{Q}(x) := \begin{cases} 0, & x \notin \mathrm{supp}(P^0), \\ Q(s_F) + m, & x = s_F, \\ Q(x), & \text{otherwise.} \end{cases}$$

Clearly $\widetilde{Q} \in \Delta(\mathcal{S})$ and $\mathrm{supp}(\widetilde{Q}) \subseteq \mathrm{supp}(P^0)$.

We first show that the objective does not increase. Since $V(s_F) = 0$ and $V(x) \geqslant 0$ for all x,

$$\begin{aligned}\mathbb{E}_{\widetilde{Q}}[V] &= \sum_{x\in\mathrm{supp}(P^0)} \widetilde{Q}(x)V(x) = \sum_{x\in\mathrm{supp}(P^0)} Q(x)V(x) + mV(s_F) \\ &= \sum_{x\in\mathrm{supp}(P^0)} Q(x)V(x) \leqslant \sum_{x\in\mathcal{S}} Q(x)V(x) = \mathbb{E}_Q[V].\end{aligned}$$

We next show that the TV distance does not increase. Since $P^0(x) = 0$ for $x \notin \mathrm{supp}(P^0)$,

$$\begin{aligned}
\|\widetilde{Q} - P^0\|_1 &= |\widetilde{Q}(s_F) - P^0(s_F)| + \sum_{\substack{x \in \mathrm{supp}(P^0) \\ x \neq s_F}} |Q(x) - P^0(x)| \\
&= |Q(s_F) + m - P^0(s_F)| + \sum_{\substack{x \in \mathrm{supp}(P^0) \\ x \neq s_F}} |Q(x) - P^0(x)| \\
&\leqslant |Q(s_F) - P^0(s_F)| + m + \sum_{\substack{x \in \mathrm{supp}(P^0) \\ x \neq s_F}} |Q(x) - P^0(x)| \\
&= \sum_{x \in \mathrm{supp}(P^0)} |Q(x) - P^0(x)| + \sum_{x \notin \mathrm{supp}(P^0)} |Q(x) - P^0(x)| \\
&= \|Q - P^0\|_1.
\end{aligned}$$

Therefore $D_{\mathrm{TV}}(\widetilde{Q}, P^0) \leqslant D_{\mathrm{TV}}(Q, P^0) \leqslant \sigma$, so $\widetilde{Q}$ is feasible and has no larger objective value. Starting from an optimal Q and applying this transformation yields an optimal solution supported inside $\mathrm{supp}(P^0)$. □

Corollary 14. *Under the assumptions of Proposition 13, the robust Bellman target for the failure-state normalized value function can be realized without leaving the nominal support.*

Proof. Apply Proposition 13 at each state-action tuple to the continuation value $V = V_{f,h+1}$. The proposition yields an optimal robust kernel whose support is contained in that of the nominal kernel. Hence the robust Bellman target remains an in-support expectation. □

C.2 A Robust MG-BE Counting Lemma

We now formulate a stage-wise counting lemma that captures the exploration term. Because the main text only uses an informal notion of distributional Eluder dimension, we state here a square-loss version tailored to the residual analysis.

Definition 15 (Square-loss distributional independence)**.** *Let $\mathscr{P}$ be a class of real-valued functions on $\mathcal{S} \times \mathcal{A} \times \mathcal{B}$, and let Π_h be a family of probability measures on the same domain. For $\epsilon > 0$, we say that a measure $\mu \in \Pi_h$ is* ϵ-dependent *on $\{\mu^1, \ldots, \mu^m\} \subseteq \Pi_h$ with respect to $\mathscr{P}$ if*

$$\mathbb{E}_{\mu}\left[\psi^2\right] \leqslant \epsilon^2 + \sum_{i=1}^{m} \mathbb{E}_{\mu^i}\left[\psi^2\right] \qquad \text{for all } \psi \in \mathscr{P}.$$

Otherwise, μ is ϵ-independent*. The square-loss distributional Eluder dimension, denoted by $\dim_{\mathrm{DE},2}(\mathscr{P}, \Pi_h, \epsilon)$, is the length of the longest sequence in Π_h whose every element is ϵ-independent of its predecessors.*

Definition 16 (Square-loss robust MG-BE dimension)**.** *For the exact robust residual class $\mathscr{P}_h := (I - \mathcal{T}_h^{\sigma})\mathcal{F}$, define*

$$\dim_{\mathrm{BE},2}^{\text{rob-MG}}(\mathcal{F}, \Pi, \epsilon) := \max_{h \in [H]} \dim_{\mathrm{DE},2}(\mathscr{P}_h, \Pi_h, \epsilon).$$

Lemma 17 (Robust MG-BE counting lemma)**.** *Fix a stage h and let $\mathscr{P}_h = (I - \mathcal{T}_h^{\sigma})\mathcal{F}$. Suppose we are given a sequence $\{(\mu_h^k, \psi_h^k)\}_{k=1}^K$ such that $\mu_h^k \in \Pi_h$, $\psi_h^k \in \mathscr{P}_h$, and for every k,*

$$\sum_{j<k} \mathbb{E}_{\mu_h^j}\left[(\psi_h^k)^2\right] \leqslant \epsilon^2, \tag{18}$$

$$\mathbb{E}_{\mu_h^k}\left[(\psi_h^k)^2\right] > 2\epsilon^2. \tag{19}$$

Then

$$K \leqslant \dim_{\mathrm{DE},2}(\mathscr{P}_h, \Pi_h, \epsilon).$$

Proof. Fix $k \in [K]$. By (18) and (19),

$$\mathbb{E}_{\mu_h^k}\left[(\psi_h^k)^2\right] > \epsilon^2 + \sum_{j<k} \mathbb{E}_{\mu_h^j}\left[(\psi_h^k)^2\right].$$

Hence the witness $\psi_h^k \in \mathscr{P}_h$ certifies that μ_h^k is ϵ-independent of $\{\mu_h^1, \ldots, \mu_h^{k-1}\}$ in the sense of Definition 15. Since this holds for every k, the sequence $\{\mu_h^1, \ldots, \mu_h^K\}$ is an ϵ-independent sequence in Π_h. By the definition of $\dim_{\mathrm{DE},2}(\mathscr{P}_h, \Pi_h, \epsilon)$, its length cannot exceed that dimension. □

Corollary 18 (Residual-budget corollary under backward validation)**.** *Fix a stage h and suppose there exists $B_{\mathrm{res},h} < \infty$ such that $|\psi_h^k(s,a,b)| \leqslant B_{\mathrm{res},h}$ for all (s,a,b), all k, and all $\psi_h^k \in \mathscr{P}_h$. Assume moreover that the sequence $\{(\mu_h^k, \psi_h^k)\}_{k=1}^K$ satisfies the backward-validation condition* (18)*. Then*

$$\sum_{k=1}^K \mathbb{E}_{\mu_h^k}\left[(\psi_h^k)^2\right] \leqslant 2K\epsilon^2 + \dim_{\mathrm{DE},2}(\mathscr{P}_h, \Pi_h, \epsilon)\, B_{\mathrm{res},h}^2.$$

Consequently,

$$\sum_{h=1}^H \sum_{k=1}^K \mathbb{E}_{\mu_h^k}\left[(\psi_h^k)^2\right] \leqslant 2HK\epsilon^2 + \sum_{h=1}^H \dim_{\mathrm{DE},2}(\mathscr{P}_h, \Pi_h, \epsilon)\, B_{\mathrm{res},h}^2.$$

Proof. For a fixed stage h, split the episode indices into

$$\mathcal{L}_h := \left\{k \in [K] : \mathbb{E}_{\mu_h^k}\left[(\psi_h^k)^2\right] > 2\epsilon^2\right\},$$
$$\mathcal{S}_h := [K] \setminus \mathcal{L}_h.$$

By Lemma 17, we have $|\mathcal{L}_h| \leqslant \dim_{\mathrm{DE},2}(\mathscr{P}_h, \Pi_h, \epsilon)$. Therefore,

$$\begin{aligned}\sum_{k=1}^K \mathbb{E}_{\mu_h^k}\left[(\psi_h^k)^2\right] &= \sum_{k\in\mathcal{S}_h} \mathbb{E}_{\mu_h^k}\left[(\psi_h^k)^2\right] + \sum_{k\in\mathcal{L}_h} \mathbb{E}_{\mu_h^k}\left[(\psi_h^k)^2\right] \\ &\leqslant 2K\epsilon^2 + |\mathcal{L}_h| B_{\mathrm{res},h}^2 \\ &\leqslant 2K\epsilon^2 + \dim_{\mathrm{DE},2}(\mathscr{P}_h, \Pi_h, \epsilon)\, B_{\mathrm{res},h}^2.\end{aligned}$$

Summing this inequality over $h \in [H]$ gives the second claim. □

Corollary 18 reduces the residual-budget input Assumption 7 to a backward-validation condition on the exact robust residual witnesses. We now verify the two bridge terms (21) and (27) directly for the current backward-validation update.

C.3 Verifying the Backward-Validation Bridge for the Current Update

For a fixed stage h, define the plug-in residual witness class

$$\mathcal{W}_h := \left\{ w_{h,f,g} \;:\; \begin{array}{c} f \in \mathcal{F}, \quad g \in \mathcal{G}_h, \\ w_{h,f,g}(s,a,b) = f_h(s,a,b) - r_h(s,a,b) \\ \qquad + \mathbb{E}_{s' \sim P_h^0(\cdot|s,a,b)}\left[\ell_\phi(f_{h+1}; s, a, b, s'; g)\right] \end{array} \right\}.$$

By construction, for every $k \geqslant 2$ and every $f \in \mathcal{F}$,

$$\widehat{\delta}_{h,k}(f) = w_{h,f,g_{h,f}^{(k)}} \in \mathcal{W}_h.$$

Assumption 8 (Bounded plug-in residual witnesses)**.** *For the fixed stage h, there exists $B_{W,h} < \infty$ such that*

$$\sup_{w \in \mathcal{W}_h} \|w\|_\infty \leqslant B_{W,h}.$$

Assumption 9 (Finite witness covering number)**.** *For the chosen witness scale $\eta_W > 0$,*

$$\mathcal{N}_W(\eta_W) := \mathcal{N}\big(\mathcal{W}_h, \|\cdot\|_\infty, \eta_W\big) < \infty.$$

Proposition 19 (Witness covering from stagewise value and dual covers)**.** *For a fixed stage h, define the stagewise sup-norm covering number*

$$\mathcal{N}_{F,h}^{\infty}(\eta_h) := \mathcal{N}\big(\mathcal{F}_h, \|\cdot\|_{\infty}, \eta_h\big).$$

Assume $\mathcal{N}_{F,h}^{\infty}(\eta_h) < \infty$ and Assumptions 11 and 12 hold. Then

$$\mathcal{N}(\mathcal{W}_h, \|\cdot\|_{\infty}, \eta_h + L_{\ell,h}(\eta_F + \eta_G)) \leqslant \mathcal{N}_{F,h}^{\infty}(\eta_h)\,\mathcal{N}_F(\eta_F)\,\mathcal{N}_G(\eta_G). \tag{20}$$

In particular, Assumption 9 follows once the stagewise value slice $\mathcal{F}_h$, the continuation slice $\mathcal{F}_{h+1}$, and the dual class $\mathcal{G}_h$ all admit finite covers.

Proof. Let $\{u^r\}_{r=1}^{M_h} \subseteq \mathcal{F}_h$ be an η_h-net under $\|\cdot\|_{\infty}$, let $\{f^m\}_{m=1}^{M_F} \subseteq \mathcal{F}_{h+1}$ be an η_F-net under $d_{\mathcal{F},h+1}$, and let $\{g^\ell\}_{\ell=1}^{M_G} \subseteq \mathcal{G}_h$ be an η_G-net under $d_{\mathcal{G},h}$.

Fix any witness $w_{h,f,g} \in \mathcal{W}_h$. Choose $\bar{u} \in \{u^r\}_{r=1}^{M_h}$, $\bar{f} \in \{f^m\}_{m=1}^{M_F}$, and $\bar{g} \in \{g^\ell\}_{\ell=1}^{M_G}$ such that

$$\|f_h - \bar{u}\|_{\infty} \leqslant \eta_h, \qquad d_{\mathcal{F},h+1}(f_{h+1}, \bar{f}) \leqslant \eta_F, \qquad d_{\mathcal{G},h}(g, \bar{g}) \leqslant \eta_G.$$

Define

$$\bar{w}(s,a,b) := \bar{u}(s,a,b) - r_h(s,a,b) + \mathbb{E}_{s' \sim P_h^0(\cdot|s,a,b)}\left[\ell_\phi(\bar{f}; s,a,b,s'; \bar{g})\right].$$

By the stagewise product interpretation $\mathcal{F} = \{\mathcal{F}_t\}_{t=1}^{H}$ from the preliminaries, the pair $(\bar{u}, \bar{f})$ can be completed to a candidate function sequence in $\mathcal{F}$, so $\bar{w} \in \mathcal{W}_h$.

For every (s,a,b),

$$\begin{aligned}
\left|w_{h,f,g}(s,a,b) - \bar{w}(s,a,b)\right| &\leqslant \left|f_h(s,a,b) - \bar{u}(s,a,b)\right| \\
&\quad + \left|\mathbb{E}_{s'}\left[\ell_\phi(f_{h+1}; s,a,b,s'; g) - \ell_\phi(\bar{f}; s,a,b,s'; \bar{g})\right]\right| \\
&\leqslant \eta_h + L_{\ell,h}(\eta_F + \eta_G),
\end{aligned}$$

where the last step uses Assumption 11 and Jensen's inequality. Hence the set of all such $\bar{w}$ forms an $\eta_h + L_{\ell,h}(\eta_F + \eta_G)$-net for $\mathcal{W}_h$ of size at most $M_h M_F M_G$, which proves (20). □

For $k \geqslant 2$, write

$$n_{h,j} := |\mathcal{D}_h^{(j)}|, \qquad S_{h,k} := \sum_{j=1}^{k-1} \frac{1}{n_{h,j}}, \qquad N_{h,<k} := \sum_{j=1}^{k-1} n_{h,j}, \qquad n_{h,\min}^{(k)} := \min_{1 \leqslant j \leqslant k-1} n_{h,j}.$$

We also keep the shorthand

$$\begin{aligned}
\mathrm{BV}_{h,k}^{\mathrm{plug}}(f) &:= \sum_{j=1}^{k-1} \mathbb{E}_{\mu_h^j}\left[\left(\widehat{\delta}_{h,k}(f)\right)^2\right], \\
\mathrm{BV}_{h,k}^{\mathrm{exact}}(f) &:= \sum_{j=1}^{k-1} \mathbb{E}_{\mu_h^j}\left[\left(\delta_h(f)\right)^2\right], \\
\widehat{\mathrm{BV}}_{h,k}(f) &:= \sum_{j=1}^{k-1} \frac{1}{n_{h,j}} \sum_{(s,a,b,r,s') \in \mathcal{D}_h^{(j)}} \left[\widehat{\delta}_{h,k}(f; s,a,b)\right]^2.
\end{aligned}$$

Lemma 20 (Uniform backward-validation deviation)**.** *Under Assumptions 8 and 9, for every $k \geqslant 2$ and every $\delta \in (0,1)$, with probability at least $1-\delta$,*

$$\sup_{f \in \mathcal{F}} \left|\widehat{\mathrm{BV}}_{h,k}(f) - \mathrm{BV}_{h,k}^{\mathrm{plug}}(f)\right| \leqslant \xi_{h,k}(\delta; \eta_W), \tag{21}$$

where

$$\xi_{h,k}(\delta; \eta_W) := B_{W,h}^2 \sqrt{2 S_{h,k} \log\big(2\mathcal{N}_W(\eta_W)/\delta\big)} + 2B_{W,h}(k-1)\eta_W.$$

Proof. Let $\{w^1, \dots, w^M\} \subseteq \mathcal{W}_h$ be an η_W-net under the sup norm, where $M = \mathcal{N}_W(\eta_W)$. Fix one net witness w^m. For every batch $j < k$ and sample index $i \in [n_{h,j}]$, define

$$Y_{j,i}^{(m)} := \frac{\big(w^m(s_{j,i}, a_{j,i}, b_{j,i})\big)^2 - \mathbb{E}_{\mu_h^j}\left[(w^m)^2\right]}{n_{h,j}}.$$

Conditioned on the history before the (j,i)-th stage-h sample is revealed, $Y_{j,i}^{(m)}$ has conditional mean zero, because that sample is drawn from the stage-h occupancy law μ_h^j. By Assumption 8,

$$\big|Y_{j,i}^{(m)}\big| \leqslant \frac{B_{W,h}^2}{n_{h,j}}.$$

Hence Azuma–Hoeffding for bounded martingale differences implies

$$\mathbb{P}\left(\left|\sum_{j=1}^{k-1}\sum_{i=1}^{n_{h,j}} Y_{j,i}^{(m)}\right| > t\right) \leqslant 2\exp\left(-\frac{t^2}{2B_{W,h}^4 S_{h,k}}\right).$$

Applying a union bound over the M net witnesses gives that with probability at least $1-\delta$,

$$\max_{m\in[M]}\left|\sum_{j=1}^{k-1}\frac{1}{n_{h,j}}\sum_{i=1}^{n_{h,j}}\big(w^m(s_{j,i}, a_{j,i}, b_{j,i})\big)^2 - \sum_{j=1}^{k-1}\mathbb{E}_{\mu_h^j}\left[(w^m)^2\right]\right| \leqslant B_{W,h}^2\sqrt{2S_{h,k}\log\big(2M/\delta\big)}. \tag{22}$$

Now fix an arbitrary witness $w \in \mathcal{W}_h$, and choose a net point $\bar{w}$ such that $\|w - \bar{w}\|_\infty \leqslant \eta_W$. Since $|w + \bar{w}| \leqslant 2B_{W,h}$, we have

$$\|w^2 - \bar{w}^2\|_\infty \leqslant \|w - \bar{w}\|_\infty \|w + \bar{w}\|_\infty \leqslant 2B_{W,h}\eta_W.$$

Therefore

$$\left|\sum_{j=1}^{k-1}\frac{1}{n_{h,j}}\sum_{i=1}^{n_{h,j}} w(s_{j,i}, a_{j,i}, b_{j,i})^2 - \sum_{j=1}^{k-1}\frac{1}{n_{h,j}}\sum_{i=1}^{n_{h,j}} \bar{w}(s_{j,i}, a_{j,i}, b_{j,i})^2\right| \leqslant 2B_{W,h}(k-1)\eta_W,$$

$$\left|\sum_{j=1}^{k-1}\mathbb{E}_{\mu_h^j}\left[w^2\right] - \sum_{j=1}^{k-1}\mathbb{E}_{\mu_h^j}\left[\bar{w}^2\right]\right| \leqslant 2B_{W,h}(k-1)\eta_W.$$

Combining these two inequalities with (22) yields

$$\sup_{w\in\mathcal{W}_h}\left|\sum_{j=1}^{k-1}\frac{1}{n_{h,j}}\sum_{i=1}^{n_{h,j}} w(s_{j,i}, a_{j,i}, b_{j,i})^2 - \sum_{j=1}^{k-1}\mathbb{E}_{\mu_h^j}\left[w^2\right]\right| \leqslant \xi_{h,k}(\delta; \eta_W).$$

Since every selected witness $\widehat{\delta}_{h,k}(f)$ belongs to $\mathcal{W}_h$, this proves (21). □

We next introduce the dual setup shared by the pooled backward-validation analysis and the on-policy operator-learning analysis. For a generic probability measure ν on $\mathcal{S}\times\mathcal{A}\times\mathcal{B}$, define

$$\mathcal{L}_h^\nu(g; f) := \mathbb{E}_{(s,a,b)\sim\nu}\mathbb{E}_{s'\sim P_h^0(\cdot|s,a,b)}\left[\ell_\phi(f_{h+1}; s, a, b, s'; g)\right].$$

We also define an ambient dual class $\bar{\mathcal{G}}_h$ large enough that the exact robust Bellman operator admits the pointwise representation

$$(\mathcal{T}_h^\sigma f_{h+1})(s,a,b) = r_h(s,a,b) - \inf_{\bar{g}\in\bar{\mathcal{G}}_h}\mathbb{E}_{s'\sim P_h^0(\cdot|s,a,b)}\left[\ell_\phi(f_{h+1}; s, a, b, s'; \bar{g})\right]. \tag{23}$$

Assumption 10 (Bounded dual loss)**.** *For the fixed stage h,*

$$\big|\ell_\phi(f_{h+1}; s, a, b, s'; g)\big| \leqslant B_\phi(\sigma)$$

for every $(f, g) \in \mathcal{F}_{h+1}\times\mathcal{G}_h$ and every sample (s, a, b, s').

Assumption 11 (Pointwise Lipschitz dual loss)**.** *There exist metrics $d_{\mathcal{F},h+1}$ on $\mathcal{F}_{h+1}$ and $d_{\mathcal{G},h}$ on $\mathcal{G}_h$, together with a constant $L_{\ell,h} \geqslant 0$, such that for every $f, \tilde{f} \in \mathcal{F}_{h+1}$, $g, \tilde{g} \in \mathcal{G}_h$, and every sample (s, a, b, s'),*

$$\left|\ell_\phi(f_{h+1}; s, a, b, s'; g) - \ell_\phi(\tilde{f}_{h+1}; s, a, b, s'; \tilde{g})\right| \leqslant L_{\ell,h}\Big(d_{\mathcal{F},h+1}(f, \tilde{f}) + d_{\mathcal{G},h}(g, \tilde{g})\Big).$$

Assumption 12 (Finite covering numbers at the working scales)**.** *For the chosen scales $\eta_F, \eta_G > 0$,*

$$\mathcal{N}_F(\eta_F) := \mathcal{N}\big(\mathcal{F}_{h+1}, d_{\mathcal{F},h+1}, \eta_F\big) < \infty,$$
$$\mathcal{N}_G(\eta_G) := \mathcal{N}\big(\mathcal{G}_h, d_{\mathcal{G},h}, \eta_G\big) < \infty.$$

Assumption 13 (Dual approximation error)**.** *For every $f \in \mathcal{F}_{h+1}$ and every probability measure ν on $\mathcal{S} \times \mathcal{A} \times \mathcal{B}$,*

$$\inf_{g \in \mathcal{G}_h} \mathcal{L}_h^\nu(g; f) \leqslant \inf_{\bar{g} \in \bar{\mathcal{G}}_h} \mathcal{L}_h^\nu(\bar{g}; f) + \epsilon_{\mathrm{dual}}.$$

We next close the operator-transfer term generated by the same pooled dual refit used by the algorithm. Define the weighted past-occupancy mixture

$$\bar{\mu}_{h,k} := \sum_{j=1}^{k-1} \lambda_{h,j}^{(k)} \mu_h^j, \qquad \lambda_{h,j}^{(k)} := \frac{n_{h,j}}{N_{h,<k}}.$$

For any $(f, g) \in \mathcal{F}_{h+1} \times \mathcal{G}_h$, define the pooled population dual loss

$$\begin{aligned}\bar{\mathcal{L}}_{h,k}(g; f) &:= \sum_{j=1}^{k-1} \lambda_{h,j}^{(k)} \mathcal{L}_h^{\mu_h^j}(g; f) \\ &= \mathbb{E}_{(s,a,b)\sim\bar{\mu}_{h,k}} \mathbb{E}_{s'\sim P_h^0(\cdot|s,a,b)} \left[\ell_\phi(f_{h+1}; s, a, b, s'; g)\right].\end{aligned}$$

The corresponding empirical pooled loss is

$$\widehat{\mathcal{L}}_{h,k}^{\mathrm{pool}}(g; f) := \frac{1}{N_{h,<k}} \sum_{j=1}^{k-1} \sum_{(s,a,b,s') \in \mathcal{D}_h^{(j)}} \ell_\phi(f_{h+1}; s, a, b, s'; g).$$

Let $g_{h,f}^{(k)} \in \operatorname{argmin}_{g \in \mathcal{G}_h} \widehat{\mathcal{L}}_{h,k}^{\mathrm{pool}}(g; f)$ denote the pooled dual fit used by the current algorithm.

Lemma 21 (Pooled dual concentration for the current update)**.** *Under Assumptions 10–12, for every $k \geqslant 2$ and every $\delta \in (0, 1)$, with probability at least $1 - \delta$,*

$$\sup_{f \in \mathcal{F}_{h+1},\, g \in \mathcal{G}_h} \left|\widehat{\mathcal{L}}_{h,k}^{\mathrm{pool}}(g; f) - \bar{\mathcal{L}}_{h,k}(g; f)\right| \leqslant \bar{\Delta}_{h,k}^{\mathrm{cov}}(\delta; \eta_F, \eta_G), \tag{24}$$

where

$$\bar{\Delta}_{h,k}^{\mathrm{cov}}(\delta; \eta_F, \eta_G) := B_\phi(\sigma) \sqrt{\frac{2 \log\big(2\mathcal{N}_F(\eta_F)\mathcal{N}_G(\eta_G)/\delta\big)}{N_{h,<k}}} + 2L_{\ell,h}(\eta_F + \eta_G).$$

Proof. Let $\{f^1, \ldots, f^{M_F}\} \subseteq \mathcal{F}_{h+1}$ be an η_F-net and $\{g^1, \ldots, g^{M_G}\} \subseteq \mathcal{G}_h$ be an η_G-net, where $M_F = \mathcal{N}_F(\eta_F)$ and $M_G = \mathcal{N}_G(\eta_G)$.

Fix a net pair (f^r, g^m). Flatten the pooled stage-h samples $\mathcal{D}_h^{(1)}, \ldots, \mathcal{D}_h^{(k-1)}$ into an arbitrary sequence indexed by $t \in [N_{h,<k}]$. For the t-th sample, define

$$X_t^{(r,m)} := \ell_\phi(f_{h+1}^r; s_t, a_t, b_t, s_t'; g^m) - \mathbb{E}\big[\ell_\phi(f_{h+1}^r; s_t, a_t, b_t, s_t'; g^m) \mid \mathcal{F}_{t-1}\big],$$

where $\mathcal{F}_{t-1}$ is the history before revealing that sample. By construction, $\{X_t^{(r,m)}\}_{t=1}^{N_{h,<k}}$ is a martingale-difference sequence. Assumption 10 gives $|X_t^{(r,m)}| \leqslant 2B_\phi(\sigma)$, and the conditional mean term equals the corresponding summand of $\bar{\mathcal{L}}_{h,k}(g^m; f^r)$. Azuma–Hoeffding therefore yields

$$\mathbb{P}\Big(\left|\widehat{\mathcal{L}}_{h,k}^{\mathrm{pool}}(g^m; f^r) - \bar{\mathcal{L}}_{h,k}(g^m; f^r)\right| > t\Big) \leqslant 2\exp\left(-\frac{N_{h,<k} t^2}{2B_\phi^2(\sigma)}\right).$$

Applying a union bound over all $M_F M_G$ net pairs gives that with probability at least $1-\delta$,

$$\max_{r\in[M_F],\, m\in[M_G]} \left|\widehat{\mathcal{L}}^{\mathrm{pool}}_{h,k}(g^m; f^r) - \bar{\mathcal{L}}_{h,k}(g^m; f^r)\right| \leqslant B_\phi(\sigma)\sqrt{\frac{2\log\!\big(2M_F M_G/\delta\big)}{N_{h,<k}}}. \tag{25}$$

The Lipschitz extension from the finite net to the full class is standard. Specifically, for arbitrary (f, g) and nearest net points $(\bar{f}, \bar{g})$, Assumption 11 implies

$$\begin{aligned}\left|\widehat{\mathcal{L}}^{\mathrm{pool}}_{h,k}(g; f) - \widehat{\mathcal{L}}^{\mathrm{pool}}_{h,k}(\bar{g}; \bar{f})\right| &\leqslant L_{\ell,h}(\eta_F + \eta_G),\\ \left|\bar{\mathcal{L}}_{h,k}(g; f) - \bar{\mathcal{L}}_{h,k}(\bar{g}; \bar{f})\right| &\leqslant L_{\ell,h}(\eta_F + \eta_G).\end{aligned}$$

Combining these two inequalities with (25) proves (24). □

Corollary 22 (Pooled ERM excess risk)**.** *Under Assumptions 10–12, with probability at least* $1-\delta$*, simultaneously for all* $f \in \mathcal{F}_{h+1}$*,*

$$\bar{\mathcal{L}}_{h,k}(g^{(k)}_{h,f}; f) - \inf_{g\in\mathcal{G}_h} \bar{\mathcal{L}}_{h,k}(g; f) \leqslant 2\bar{\Delta}^{\mathrm{cov}}_{h,k}(\delta; \eta_F, \eta_G). \tag{26}$$

Proof. Fix $f \in \mathcal{F}_{h+1}$ and $\tau > 0$. Choose $g^\star_{f,\tau} \in \mathcal{G}_h$ such that

$$\bar{\mathcal{L}}_{h,k}(g^\star_{f,\tau}; f) \leqslant \inf_{g\in\mathcal{G}_h} \bar{\mathcal{L}}_{h,k}(g; f) + \tau.$$

On the event (24),

$$\begin{aligned}\bar{\mathcal{L}}_{h,k}(g^{(k)}_{h,f}; f) &\leqslant \widehat{\mathcal{L}}^{\mathrm{pool}}_{h,k}(g^{(k)}_{h,f}; f) + \bar{\Delta}^{\mathrm{cov}}_{h,k}(\delta; \eta_F, \eta_G)\\ &\leqslant \widehat{\mathcal{L}}^{\mathrm{pool}}_{h,k}(g^\star_{f,\tau}; f) + \bar{\Delta}^{\mathrm{cov}}_{h,k}(\delta; \eta_F, \eta_G)\\ &\leqslant \bar{\mathcal{L}}_{h,k}(g^\star_{f,\tau}; f) + 2\bar{\Delta}^{\mathrm{cov}}_{h,k}(\delta; \eta_F, \eta_G)\\ &\leqslant \inf_{g\in\mathcal{G}_h} \bar{\mathcal{L}}_{h,k}(g; f) + 2\bar{\Delta}^{\mathrm{cov}}_{h,k}(\delta; \eta_F, \eta_G) + \tau.\end{aligned}$$

Letting $\tau \downarrow 0$ proves (26). □

Corollary 23 (Backward-validation operator transfer)**.** *Under Assumptions 10–13, with probability at least* $1-\delta$*, simultaneously for all* $f \in \mathcal{F}_{h+1}$*,*

$$\sum_{j=1}^{k-1} \mathbb{E}_{\mu^j_h}\left[\left(\widehat{\mathcal{T}}^{\sigma,(k)}_{h,f} f_{h+1} - \mathcal{T}^\sigma_h f_{h+1}\right)^2\right] \leqslant \rho_{h,k}(\delta; \eta_F, \eta_G), \tag{27}$$

where

$$\rho_{h,k}(\delta; \eta_F, \eta_G) := \frac{2B_\phi(\sigma)\, N_{h,<k}}{n^{(k)}_{h,\min}}\Big(2\bar{\Delta}^{\mathrm{cov}}_{h,k}(\delta; \eta_F, \eta_G) + \epsilon_{\mathrm{dual}}\Big).$$

Proof. Fix $f \in \mathcal{F}_{h+1}$ and define the pointwise operator gap

$$e_{h,k}(f) := \widehat{\mathcal{T}}^{\sigma,(k)}_{h,f} f_{h+1} - \mathcal{T}^\sigma_h f_{h+1}.$$

By the ambient dual representation (23), the function $-e_{h,k}(f)$ is pointwise nonnegative. Assumption 10 further implies

$$-2B_\phi(\sigma) \leqslant e_{h,k}(f; s, a, b) \leqslant 0 \qquad \text{for all } (s, a, b). \tag{28}$$

Therefore

$$\begin{aligned}\sum_{j=1}^{k-1} \mathbb{E}_{\mu^j_h}\left[e_{h,k}(f)^2\right] &= \sum_{j=1}^{k-1} \frac{N_{h,<k}}{n_{h,j}} \lambda^{(k)}_{h,j} \mathbb{E}_{\mu^j_h}\left[e_{h,k}(f)^2\right] &(29)\\ &\leqslant \frac{N_{h,<k}}{n^{(k)}_{h,\min}} \mathbb{E}_{\bar{\mu}_{h,k}}\left[e_{h,k}(f)^2\right] &(30)\\ &\leqslant \frac{2B_\phi(\sigma)\, N_{h,<k}}{n^{(k)}_{h,\min}} \mathbb{E}_{\bar{\mu}_{h,k}}\left[-e_{h,k}(f)\right], &(31)\end{aligned}$$

where the last step uses (28).

Next,

$$\mathbb{E}_{\bar{\mu}_{h,k}}\left[-e_{h,k}(f)\right] = \bar{\mathcal{L}}_{h,k}(g_{h,f}^{(k)}; f) - \inf_{\bar{g}\in\bar{\mathcal{G}}_h} \bar{\mathcal{L}}_{h,k}(\bar{g}; f).$$

Adding and subtracting $\inf_{g\in\mathcal{G}_h} \bar{\mathcal{L}}_{h,k}(g; f)$ and then applying Corollary 22 together with Assumption 13 yields

$$\mathbb{E}_{\bar{\mu}_{h,k}}\left[-e_{h,k}(f)\right] \leqslant 2\bar{\Delta}_{h,k}^{\mathrm{cov}}(\delta; \eta_F, \eta_G) + \epsilon_{\mathrm{dual}}.$$

Substituting this bound into (31) proves (27). □

Proposition 24 (Backward validation induced by the current confidence set)**.** *Fix a stage h and an episode $k \geqslant 2$. On the intersection of the events from Lemma 20 and Corollary 23, if the confidence set update satisfies $\widehat{\mathrm{BV}}_{h,k}(f) \leqslant \beta_{h,k}$ for every $f \in \mathfrak{F}^{(k-1)}$, then every $f \in \mathfrak{F}^{(k-1)}$ obeys*

$$\mathrm{BV}_{h,k}^{\mathrm{exact}}(f) \leqslant 2\beta_{h,k} + 2\xi_{h,k}(\delta; \eta_W) + 2\rho_{h,k}(\delta; \eta_F, \eta_G). \tag{32}$$

Proof. Fix any $f \in \mathfrak{F}^{(k-1)}$. The identity $\delta_h(f) = \widehat{\delta}_{h,k}(f) + e_{h,k}(f)$ holds pointwise, so

$$\begin{aligned}
\mathrm{BV}_{h,k}^{\mathrm{exact}}(f) &= \sum_{j=1}^{k-1} \mathbb{E}_{\mu_h^j}\left[\left(\widehat{\delta}_{h,k}(f) + e_{h,k}(f)\right)^2\right] \\
&\leqslant 2\sum_{j=1}^{k-1} \mathbb{E}_{\mu_h^j}\left[\left(\widehat{\delta}_{h,k}(f)\right)^2\right] + 2\sum_{j=1}^{k-1} \mathbb{E}_{\mu_h^j}\left[\left(e_{h,k}(f)\right)^2\right] \\
&= 2\,\mathrm{BV}_{h,k}^{\mathrm{plug}}(f) + 2\sum_{j=1}^{k-1} \mathbb{E}_{\mu_h^j}\left[\left(e_{h,k}(f)\right)^2\right].
\end{aligned}$$

On the event of Corollary 23, the second term is at most $2\rho_{h,k}(\delta; \eta_F, \eta_G)$. On the event of Lemma 20,

$$\mathrm{BV}_{h,k}^{\mathrm{plug}}(f) \leqslant \widehat{\mathrm{BV}}_{h,k}(f) + \xi_{h,k}(\delta; \eta_W) \leqslant \beta_{h,k} + \xi_{h,k}(\delta; \eta_W),$$

where the last step uses the defining property of $\mathfrak{F}^{(k-1)}$. Substituting both bounds proves (32). □

In particular, if the selected function $f^{(k)}$ satisfies $2\beta_{h,k} + 2\xi_{h,k}(\delta; \eta_W) + 2\rho_{h,k}(\delta; \eta_F, \eta_G) \leqslant \epsilon^2$, then the exact witness $\psi_h^k := \delta_h(f^{(k)})$ satisfies the backward-validation premise (18) needed by Lemma 17.

C.4 Closing Optimism from the Confidence Set

Assumption 14 (Reference realizability for optimism closure)**.** *There exists a reference function $f^\dagger \in \mathcal{F}$ such that*

$$V_1^{\star,\sigma}(s_1) \leqslant V_{f^\dagger,1}(s_1) + \epsilon_{\mathrm{appr}} \tag{33}$$

and, for every stage h,

$$f_h^\dagger = \mathcal{T}_h^\sigma f_{h+1}^\dagger. \tag{34}$$

Proposition 25 (Exact robust realizability implies an optimism reference)**.** *Suppose there exists $f^\star \in \mathcal{F}$ such that for every stage $h \in [H]$,*

$$f_h^\star = \mathcal{T}_h^\sigma f_{h+1}^\star \qquad \textit{and} \qquad V_{f^\star,1}(s_1) = V_1^{\star,\sigma}(s_1).$$

Then Assumption 14 holds with $f^\dagger = f^\star$ and $\epsilon_{\mathrm{appr}} = 0$.

Proof. Take $f^\dagger := f^\star$. The fixed-point identity is exactly (34), and the value identity implies (33) with $\epsilon_{\mathrm{appr}} = 0$. □

Proposition 26 (Reference retention)**.** *Under Assumption 14, on the intersection of the events from Lemma 20 and Corollary 23, if*

$$\beta_{h,k} \geqslant \xi_{h,k}(\delta; \eta_W) + \rho_{h,k}(\delta; \eta_F, \eta_G) \qquad \textit{for every } h \in [H], \tag{35}$$

then $f^\dagger \in \mathfrak{F}^{(k-1)}$.

Proof. Fix a stage h. By (34), the exact residual of $f^\dagger$ is zero: $\delta_h(f^\dagger) \equiv 0$. Hence $\widehat{\delta}_{h,k}(f^\dagger) = -e_{h,k}(f^\dagger)$ pointwise, and therefore

$$\mathrm{BV}_{h,k}^{\mathrm{plug}}(f^\dagger) = \sum_{j=1}^{k-1} \mathbb{E}_{\mu_h^j}\left[\left(e_{h,k}(f^\dagger)\right)^2\right] \leqslant \rho_{h,k}(\delta; \eta_F, \eta_G)$$

on the event of Corollary 23. Applying Lemma 20 then gives

$$\widehat{\mathrm{BV}}_{h,k}(f^\dagger) \leqslant \mathrm{BV}_{h,k}^{\mathrm{plug}}(f^\dagger) + \xi_{h,k}(\delta; \eta_W) \leqslant \rho_{h,k}(\delta; \eta_F, \eta_G) + \xi_{h,k}(\delta; \eta_W).$$

By (35), the right-hand side is at most $\beta_{h,k}$. Since this holds for every stage h, the definition of $\mathfrak{F}^{(k-1)}$ implies $f^\dagger \in \mathfrak{F}^{(k-1)}$. □

Corollary 27 (Approximate optimism from confidence sets)**.** *Under Assumption 14 and the conditions of Proposition 26, the algorithm's optimistic choice satisfies*

$$V_1^{\star,\sigma}(s_1) \leqslant V_{f^{(k)},1}(s_1) + \epsilon_{\mathrm{appr}}.$$

In particular, Assumption 5 is implied by the backward-validation confidence set construction.

Proof. By Proposition 26, $f^\dagger \in \mathfrak{F}^{(k-1)}$. Since the algorithm chooses $f^{(k)} \in \operatorname{argmax}_{f \in \mathfrak{F}^{(k-1)}} V_{f,1}(s_1)$, we have

$$V_{f^{(k)},1}(s_1) \geqslant V_{f^\dagger,1}(s_1).$$

Combining this inequality with (33) proves the claim. □

Assumption 15 (Projected reference realizability for optimism closure)**.** *Fix stagewise maps $\{\mathsf{P}_h\}_{h=1}^H$. There exists a reference function $f^\dagger \in \mathcal{F}$ such that*

$$V_1^{\star,\sigma}(s_1) \leqslant V_{f^\dagger,1}(s_1) + \epsilon_{\mathrm{appr}} \tag{36}$$

and, for every stage $h \in [H]$,

$$f_h^\dagger = \mathsf{P}_h \mathcal{T}_h^\sigma f_{h+1}^\dagger. \tag{37}$$

Proposition 28 (Projected reference retention)**.** *Fix stagewise maps $\{\mathsf{P}_h\}_{h=1}^H$ and let*

$$\mathfrak{F}_{\mathsf{P}}^{(k-1)} := \left\{ f \in \mathcal{F} : \widehat{\mathrm{BV}}_{h,k}^{\mathsf{P}}(f) \leqslant \beta_{h,k}^{\mathsf{P}} \text{ for every } h \in [H] \right\}. \tag{38}$$

For any $f \in \mathcal{F}$, define the projected plug-in residual

$$\widehat{\delta}_{h,\mathsf{P},k}(f) := f_h - \mathsf{P}_h \widehat{\mathcal{T}}_{h,f}^{\sigma,(k)} f_{h+1},$$

the projected operator error

$$e_{h,\mathsf{P},k}(f) := \mathsf{P}_h \widehat{\mathcal{T}}_{h,f}^{\sigma,(k)} f_{h+1} - \mathsf{P}_h \mathcal{T}_h^\sigma f_{h+1},$$

and the projected plug-in backward-validation score

$$\mathrm{BV}_{h,k}^{\mathsf{P},\mathrm{plug}}(f) := \sum_{j=1}^{k-1} \mathbb{E}_{\widetilde{\mu}_h^j}\left[\left(\widehat{\delta}_{h,\mathsf{P},k}(f)\right)^2\right].$$

Under Assumption 15, if for every stage $h \in [H]$,

$$\left|\widehat{\mathrm{BV}}_{h,k}^{\mathsf{P}}(f^\dagger) - \mathrm{BV}_{h,k}^{\mathsf{P},\mathrm{plug}}(f^\dagger)\right| \leqslant \xi_{h,k}^{\mathsf{P}}, \tag{39}$$

$$\sum_{j=1}^{k-1} \mathbb{E}_{\widetilde{\mu}_h^j}\left[\left(e_{h,\mathsf{P},k}(f^\dagger)\right)^2\right] \leqslant \rho_{h,k}^{\mathsf{P}}, \tag{40}$$

$$\beta_{h,k}^{\mathsf{P}} \geqslant \xi_{h,k}^{\mathsf{P}} + \rho_{h,k}^{\mathsf{P}}, \tag{41}$$

then $f^\dagger \in \mathfrak{F}_{\mathsf{P}}^{(k-1)}$.

Proof. Fix a stage $h \in [H]$. By (37), the projected exact residual of $f^\dagger$ is zero. Therefore

$$\widehat{\delta}_{h,\mathsf{P},k}(f^\dagger) = -e_{h,\mathsf{P},k}(f^\dagger)$$

pointwise, which implies

$$\mathrm{BV}^{\mathsf{P},\mathrm{plug}}_{h,k}(f^\dagger) = \sum_{j=1}^{k-1} \mathbb{E}_{\widetilde{\mu}_h^j}\left[\left(e_{h,\mathsf{P},k}(f^\dagger)\right)^2\right] \leqslant \rho^{\mathsf{P}}_{h,k}$$

by (40). Combining this with (39) gives

$$\widehat{\mathrm{BV}}^{\mathsf{P}}_{h,k}(f^\dagger) \leqslant \mathrm{BV}^{\mathsf{P},\mathrm{plug}}_{h,k}(f^\dagger) + \xi^{\mathsf{P}}_{h,k} \leqslant \rho^{\mathsf{P}}_{h,k} + \xi^{\mathsf{P}}_{h,k} \leqslant \beta^{\mathsf{P}}_{h,k}.$$

Since this holds for every stage h, definition (38) implies $f^\dagger \in \mathfrak{F}^{(k-1)}_{\mathsf{P}}$. □

Corollary 29 (Projected approximate optimism from confidence sets)**.** *Under Assumption 15 and the conditions of Proposition 28, if the algorithm chooses*

$$f^{(k)} \in \operatorname*{argmax}_{f \in \mathfrak{F}^{(k-1)}_{\mathsf{P}}} V_{f,1}(s_1),$$

then

$$V_1^{\star,\sigma}(s_1) \leqslant V_{f^{(k)},1}(s_1) + \epsilon_{\mathrm{appr}}.$$

In particular, Assumption 5 holds for the projected confidence-set route.

Proof. By Proposition 28, $f^\dagger \in \mathfrak{F}^{(k-1)}_{\mathsf{P}}$. Hence the optimistic choice over $\mathfrak{F}^{(k-1)}_{\mathsf{P}}$ satisfies $V_{f^{(k)},1}(s_1) \geqslant V_{f^\dagger,1}(s_1)$. Combining this with (36) proves the claim. □

C.5 A Covering-Number Dual ERM Concentration Lemma

We now return to a fresh-sample, current-occupancy model to verify the on-policy operator-learning budget Assumption 6. Fix a stage h and an occupancy measure μ_h^k. Given i.i.d. samples $\{(s_i, a_i, b_i, s_i')\}_{i=1}^{N_{h,k}}$ from the joint law $\mu_h^k \otimes P_h^0$, define

$$\widehat{\mathcal{L}}_{h,k}(g; f) := \frac{1}{N_{h,k}} \sum_{i=1}^{N_{h,k}} \ell_\phi(f_{h+1}; s_i, a_i, b_i, s_i'; g),$$

and let $\hat{g}^{(k)}_{h,f} \in \operatorname{argmin}_{g \in \mathcal{G}_h} \widehat{\mathcal{L}}_{h,k}(g; f)$. The fitted plug-in operator is

$$\left(\widehat{\mathcal{T}}^{\sigma,(k)}_{h,f} f_{h+1}\right)(s,a,b) := r_h(s,a,b) - \mathbb{E}_{s' \sim P_h^0(\cdot|s,a,b)}\left[\ell_\phi(f_{h+1}; s, a, b, s'; \hat{g}^{(k)}_{h,f})\right].$$

Lemma 30 (Uniform dual concentration via covering numbers)**.** *Under Assumptions 10–12, for every $\delta \in (0,1)$, with probability at least $1-\delta$,*

$$\sup_{f \in \mathcal{F}_{h+1},\, g \in \mathcal{G}_h} \left|\widehat{\mathcal{L}}_{h,k}(g; f) - \mathcal{L}^\mu_h(g; f)\right| \leqslant \Delta^{\mathrm{cov}}_{h,k}(\delta; \eta_F, \eta_G), \tag{42}$$

where

$$\Delta^{\mathrm{cov}}_{h,k}(\delta; \eta_F, \eta_G) := B_\phi(\sigma)\sqrt{\frac{2\log\big(2\mathcal{N}_F(\eta_F)\mathcal{N}_G(\eta_G)/\delta\big)}{N_{h,k}}} + 2L_{\ell,h}(\eta_F + \eta_G).$$

Proof. Let $\{f^1, \dots, f^{M_F}\} \subseteq \mathcal{F}_{h+1}$ be an η_F-net under $d_{\mathcal{F},h+1}$ and $\{g^1, \dots, g^{M_G}\} \subseteq \mathcal{G}_h$ be an η_G-net under $d_{\mathcal{G},h}$, where $M_F = \mathcal{N}_F(\eta_F)$ and $M_G = \mathcal{N}_G(\eta_G)$.

Fix a net pair (f^r, g^m). By Assumption 10, the random variables $\ell_\phi(f^r_{h+1}; s_i, a_i, b_i, s_i'; g^m)$ are i.i.d. and take values in $[-B_\phi(\sigma), B_\phi(\sigma)]$. Hoeffding's inequality therefore yields, for every $t > 0$,

$$\mathbb{P}\Big(\left|\widehat{\mathcal{L}}_{h,k}(g^m; f^r) - \mathcal{L}^\mu_h(g^m; f^r)\right| > t\Big) \leqslant 2\exp\left(-\frac{N_{h,k}t^2}{2B_\phi^2(\sigma)}\right).$$

Applying a union bound over all $M_F M_G$ net pairs, we obtain that with probability at least $1-\delta$,

$$\max_{r\in[M_F],\, m\in[M_G]} \left|\widehat{\mathcal{L}}_{h,k}(g^m; f^r) - \mathcal{L}_h^\mu(g^m; f^r)\right| \leqslant B_\phi(\sigma)\sqrt{\frac{2\log\big(2M_F M_G/\delta\big)}{N_{h,k}}}. \tag{43}$$

Now fix arbitrary $(f, g) \in \mathcal{F}_{h+1} \times \mathcal{G}_h$, and choose net points $\bar{f}, \bar{g}$ such that $d_{\mathcal{F},h+1}(f, \bar{f}) \leqslant \eta_F$ and $d_{\mathcal{G},h}(g, \bar{g}) \leqslant \eta_G$. By Assumption 11, for every sample (s, a, b, s'),

$$\left|\ell_\phi(f_{h+1}; s, a, b, s'; g) - \ell_\phi(\bar{f}_{h+1}; s, a, b, s'; \bar{g})\right| \leqslant L_{\ell,h}(\eta_F + \eta_G).$$

Averaging this inequality over the empirical sample and over the population law separately gives

$$\begin{aligned}\left|\widehat{\mathcal{L}}_{h,k}(g; f) - \widehat{\mathcal{L}}_{h,k}(\bar{g}; \bar{f})\right| &\leqslant L_{\ell,h}(\eta_F + \eta_G),\\ \left|\mathcal{L}_h^\mu(g; f) - \mathcal{L}_h^\mu(\bar{g}; \bar{f})\right| &\leqslant L_{\ell,h}(\eta_F + \eta_G).\end{aligned}$$

Combining these two inequalities with (43) yields

$$\begin{aligned}\left|\widehat{\mathcal{L}}_{h,k}(g; f) - \mathcal{L}_h^\mu(g; f)\right| &\leqslant \left|\widehat{\mathcal{L}}_{h,k}(g; f) - \widehat{\mathcal{L}}_{h,k}(\bar{g}; \bar{f})\right| + \left|\widehat{\mathcal{L}}_{h,k}(\bar{g}; \bar{f}) - \mathcal{L}_h^\mu(\bar{g}; \bar{f})\right|\\ &\quad + \left|\mathcal{L}_h^\mu(\bar{g}; \bar{f}) - \mathcal{L}_h^\mu(g; f)\right|\\ &\leqslant B_\phi(\sigma)\sqrt{\frac{2\log\big(2M_F M_G/\delta\big)}{N_{h,k}}} + 2L_{\ell,h}(\eta_F + \eta_G).\end{aligned}$$

Since (f, g) was arbitrary, this proves (42). □

Corollary 31 (ERM excess population risk)**.** *Under Assumptions 10–12, with probability at least $1-\delta$, simultaneously for all $f \in \mathcal{F}_{h+1}$,*

$$\mathcal{L}_h^\mu(\hat{g}_{h,f}^{(k)}; f) - \inf_{g\in\mathcal{G}_h} \mathcal{L}_h^\mu(g; f) \leqslant 2\Delta_{h,k}^{\mathrm{cov}}(\delta; \eta_F, \eta_G). \tag{44}$$

Proof. Fix $f \in \mathcal{F}_{h+1}$ and $\tau > 0$. Choose $g_{f,\tau}^\star \in \mathcal{G}_h$ such that

$$\mathcal{L}_h^\mu(g_{f,\tau}^\star; f) \leqslant \inf_{g\in\mathcal{G}_h} \mathcal{L}_h^\mu(g; f) + \tau.$$

On the event (42),

$$\begin{aligned}\mathcal{L}_h^\mu(\hat{g}_{h,f}^{(k)}; f) &\leqslant \widehat{\mathcal{L}}_{h,k}(\hat{g}_{h,f}^{(k)}; f) + \Delta_{h,k}^{\mathrm{cov}}(\delta; \eta_F, \eta_G)\\ &\leqslant \widehat{\mathcal{L}}_{h,k}(g_{f,\tau}^\star; f) + \Delta_{h,k}^{\mathrm{cov}}(\delta; \eta_F, \eta_G)\\ &\leqslant \mathcal{L}_h^\mu(g_{f,\tau}^\star; f) + 2\Delta_{h,k}^{\mathrm{cov}}(\delta; \eta_F, \eta_G)\\ &\leqslant \inf_{g\in\mathcal{G}_h} \mathcal{L}_h^\mu(g; f) + 2\Delta_{h,k}^{\mathrm{cov}}(\delta; \eta_F, \eta_G) + \tau.\end{aligned}$$

Since $\tau > 0$ is arbitrary, (44) follows. The event (42) already holds uniformly over all f, so the conclusion is simultaneous in f. □

Corollary 32 (Operator-learning budget from dual ERM)**.** *Under Assumptions 10–13, with probability at least $1-\delta$, simultaneously for all $f \in \mathcal{F}_{h+1}$,*

$$\left\|\widehat{\mathcal{T}}_{h,f}^{\sigma,(k)} f_{h+1} - \mathcal{T}_h^\sigma f_{h+1}\right\|_{1,\mu_h^k} \leqslant 2\Delta_{h,k}^{\mathrm{cov}}(\delta; \eta_F, \eta_G) + \epsilon_{\mathrm{dual}}. \tag{45}$$

Therefore Assumption 6 holds with $\alpha_{h,k} = 2\Delta_{h,k}^{\mathrm{cov}}(\delta; \eta_F, \eta_G)$.

Proof. Fix $f \in \mathcal{F}_{h+1}$. By (23), the pointwise operator gap is

$$\begin{aligned}\big(\widehat{\mathcal{T}}_{h,f}^{\sigma,(k)} f_{h+1} - \mathcal{T}_h^\sigma f_{h+1}\big)(s, a, b) = \inf_{\bar{g}\in\bar{\mathcal{G}}_h} &\mathbb{E}_{s'\sim P_h^0(\cdot|s,a,b)}\left[\ell_\phi(f_{h+1}; s, a, b, s'; \bar{g})\right]\\ &- \mathbb{E}_{s'\sim P_h^0(\cdot|s,a,b)}\left[\ell_\phi(f_{h+1}; s, a, b, s'; \hat{g}_{h,f}^{(k)})\right].\end{aligned}$$

Because the first term is a pointwise infimum, the right-hand side is nonpositive. Hence the $L_1(\mu_h^k)$ norm equals the μ_h^k-expectation of its negative:

$$\left\|\widehat{\mathcal{T}}_{h,f}^{\sigma,(k)} f_{h+1} - \mathcal{T}_h^\sigma f_{h+1}\right\|_{1,\mu_h^k} = \mathcal{L}_h^\mu(\hat{g}_{h,f}^{(k)}; f) - \inf_{\bar{g}\in\bar{\mathcal{G}}_h} \mathcal{L}_h^\mu(\bar{g}; f). \tag{46}$$

Add and subtract $\inf_{g\in\mathcal{G}_h} \mathcal{L}_h^\mu(g; f)$ in (46) to get

$$\left\|\widehat{\mathcal{T}}_{h,f}^{\sigma,(k)} f_{h+1} - \mathcal{T}_h^\sigma f_{h+1}\right\|_{1,\mu_h^k} = \underbrace{\mathcal{L}_h^\mu(\hat{g}_{h,f}^{(k)}; f) - \inf_{g\in\mathcal{G}_h} \mathcal{L}_h^\mu(g; f)}_{\text{ERM estimation error}} + \underbrace{\inf_{g\in\mathcal{G}_h} \mathcal{L}_h^\mu(g; f) - \inf_{\bar{g}\in\bar{\mathcal{G}}_h} \mathcal{L}_h^\mu(\bar{g}; f)}_{\text{dual approximation error}}.$$

The first term is bounded by $2\Delta_{h,k}^{\mathrm{cov}}(\delta; \eta_F, \eta_G)$ by Corollary 31; the second term is bounded by ϵ_{dual} by Assumption 13. This proves (45). □

C.6 Concrete Structural Upper Bounds

We now compress the witness and covering assumptions into explicit formulas for common model classes. Throughout this subsection, write $\mathcal{X}_h := \mathcal{S} \times \mathcal{A} \times \mathcal{B}$.

Lemma 33 (Witness-envelope bound)**.** *Assume rewards satisfy $0 \leqslant r_h(s, a, b) \leqslant 1$ and the value class is uniformly bounded at stage h: $\sup_{f\in\mathcal{F}} \|f_h\|_\infty \leqslant B_{F,h}$. Under Assumption 10,*

$$\sup_{w\in\mathcal{W}_h} \|w\|_\infty \leqslant B_{F,h} + 1 + B_\phi(\sigma).$$

Therefore Assumption 8 holds with $B_{W,h} = B_{F,h} + 1 + B_\phi(\sigma)$.

Proof. Fix any $w_{h,f,g} \in \mathcal{W}_h$. For every $(s, a, b) \in \mathcal{X}_h$,

$$\begin{aligned}\left|w_{h,f,g}(s, a, b)\right| &\leqslant |f_h(s, a, b)| + |r_h(s, a, b)| + \left|\mathbb{E}_{s'\sim P_h^0(\cdot|s,a,b)} \left[\ell_\phi(f_{h+1}; s, a, b, s'; g)\right]\right| \\ &\leqslant B_{F,h} + 1 + B_\phi(\sigma),\end{aligned}$$

where the last inequality uses the reward bound and Assumption 10. Taking the supremum over (f, g) proves the claim. □

Lemma 34 (Exact residual envelope under completeness)**.** *Assume robust minimax completeness holds stagewise and that the value class is uniformly bounded: $\sup_{f\in\mathcal{F}} \|f_h\|_\infty \leqslant B_{F,h}$ for every $h \in [H]$. Then every exact robust residual $\psi \in \mathscr{P}_h = (I - \mathcal{T}_h^\sigma)\mathcal{F}$ satisfies*

$$\|\psi\|_\infty \leqslant 2B_{F,h}.$$

Proof. Fix $\psi \in \mathscr{P}_h$. By definition, there exists $f \in \mathcal{F}$ such that $\psi = f_h - \mathcal{T}_h^\sigma f_{h+1}$. By robust minimax completeness, $\mathcal{T}_h^\sigma f_{h+1} \in \mathcal{F}_h$. Hence both f_h and $\mathcal{T}_h^\sigma f_{h+1}$ have sup norm at most $B_{F,h}$, so

$$\|\psi\|_\infty \leqslant \|f_h\|_\infty + \|\mathcal{T}_h^\sigma f_{h+1}\|_\infty \leqslant 2B_{F,h}.$$

□

Proposition 35 (Finite-dimensional linear covering bound)**.** *Let $\mathcal{H} = \{x \mapsto \langle\theta, \psi(x)\rangle : \theta \in \mathbb{R}^d,\ \|\theta\|_2 \leqslant R\}$ be a linear function class on a domain $\mathcal{X}$, and assume $\sup_{x\in\mathcal{X}} \|\psi(x)\|_2 \leqslant \kappa$. Then for every $\eta > 0$,*

$$\log \mathcal{N}\big(\mathcal{H}, \|\cdot\|_\infty, \eta\big) \leqslant d \log\left(1 + \frac{2R\kappa}{\eta}\right). \tag{47}$$

Proof. Let $\Theta := \{\theta \in \mathbb{R}^d : \|\theta\|_2 \leqslant R\}$. The Euclidean ball Θ admits an η/κ-net of cardinality at most $(1 + 2R\kappa/\eta)^d$. For every pair of parameters $\theta, \tilde{\theta}$ and every $x \in \mathcal{X}$,

$$\left|\langle\theta, \psi(x)\rangle - \langle\tilde{\theta}, \psi(x)\rangle\right| \leqslant \|\theta - \tilde{\theta}\|_2 \|\psi(x)\|_2 \leqslant \kappa\|\theta - \tilde{\theta}\|_2.$$

Hence an η/κ-cover in parameter space induces an η-cover of $\mathcal{H}$ under the sup norm. Taking logarithms yields (47). □

Corollary 36 (Linear witness and dual classes)**.** *Suppose that for a fixed stage h,*

$$\mathcal{F}_{h+1} \subseteq \left\{x \mapsto \langle \theta_F, \phi_F(x)\rangle : \theta_F \in \mathbb{R}^{d_{F,h}},\ \|\theta_F\|_2 \leqslant R_{F,h}\right\},$$
$$\mathcal{G}_h \subseteq \left\{x \mapsto \Gamma_h(x)\vartheta_G : \vartheta_G \in \mathbb{R}^{d_{G,h}},\ \|\vartheta_G\|_2 \leqslant R_{G,h}\right\},$$
$$\mathcal{W}_h \subseteq \left\{x \mapsto \langle \theta_W, \phi_W(x)\rangle : \theta_W \in \mathbb{R}^{d_{W,h}},\ \|\theta_W\|_2 \leqslant R_{W,h}\right\},$$

with $\sup_x \|\phi_F(x)\|_2 \leqslant \kappa_{F,h}$, $\sup_x \|\Gamma_h(x)\|_{\mathrm{op}} \leqslant \kappa_{G,h}$, *and* $\sup_x \|\phi_W(x)\|_2 \leqslant \kappa_{W,h}$. *Then*

$$\log \mathcal{N}_F(\eta_F) \leqslant d_{F,h} \log\left(1 + \frac{2R_{F,h}\kappa_{F,h}}{\eta_F}\right),$$
$$\log \mathcal{N}_G(\eta_G) \leqslant d_{G,h} \log\left(1 + \frac{2R_{G,h}\kappa_{G,h}}{\eta_G}\right),$$
$$\log \mathcal{N}_W(\eta_W) \leqslant d_{W,h} \log\left(1 + \frac{2R_{W,h}\kappa_{W,h}}{\eta_W}\right).$$

Consequently, Lemma 20 and the dual concentration lemmas admit explicit dimension-dependent logarithmic terms.

Proof. Apply Proposition 35 to each class separately. For $\mathcal{G}_h$, the operator norm bound implies $\|\Gamma_h(x)(\vartheta - \tilde{\vartheta})\|_2 \leqslant \kappa_{G,h}\|\vartheta - \tilde{\vartheta}\|_2$, so the same covering argument applies. □

Proposition 37 (Linear square-loss DE upper bound)**.** *Let Π_h be any family of probability measures on a domain $\mathcal{X}$. Consider the linear class*

$$\mathcal{H} := \left\{x \mapsto \langle \theta, \phi(x)\rangle : \theta \in \mathbb{R}^d,\ \|\theta\|_2 \leqslant R\right\},$$

and assume $\sup_{x\in\mathcal{X}} \|\phi(x)\|_2 \leqslant \kappa$. *Then for every $\epsilon > 0$,*

$$\dim_{\mathrm{DE},2}(\mathcal{H}, \Pi_h, \epsilon) \leqslant 4d \log_2\left(2 + \frac{\kappa^2 R^2}{\epsilon^2}\right). \tag{48}$$

Proof. Let $\mu^1, \ldots, \mu^n \in \Pi_h$ be an ϵ-independent sequence with respect to $\mathcal{H}$. For each $k \in [n]$, independence yields a witness $\theta_k \in \mathbb{R}^d$ with $\|\theta_k\|_2 \leqslant R$ such that

$$\mathbb{E}_{\mu^k}\left[\langle \theta_k, \phi\rangle^2\right] > \epsilon^2 + \sum_{j<k} \mathbb{E}_{\mu^j}\left[\langle \theta_k, \phi\rangle^2\right]. \tag{49}$$

Define the covariance matrices

$$\Sigma_k := \mathbb{E}_{\mu^k}\left[\phi(x)\phi(x)^\top\right], \qquad V_k := \sum_{j=1}^{k} \Sigma_j, \qquad \lambda := \frac{\epsilon^2}{R^2}, \qquad M_k := \lambda I_d + V_k.$$

Then (49) is exactly

$$\theta_k^\top \Sigma_k \theta_k > \epsilon^2 + \theta_k^\top V_{k-1}\theta_k.$$

Since $\epsilon^2 = \lambda R^2 \geqslant \lambda\|\theta_k\|_2^2$, we obtain

$$\theta_k^\top \Sigma_k \theta_k > \theta_k^\top M_{k-1}\theta_k.$$

Hence the largest eigenvalue of $M_{k-1}^{-1/2}\Sigma_k M_{k-1}^{-1/2}$ is strictly larger than 1, and therefore

$$\begin{aligned}\det(M_k) &= \det(M_{k-1})\,\det\Big(I_d + M_{k-1}^{-1/2}\Sigma_k M_{k-1}^{-1/2}\Big)\\ &\geqslant 2\det(M_{k-1}).\end{aligned}$$

Iterating over $k = 1, \ldots, n$ gives the lower bound

$$\det(M_n) \geqslant 2^n \lambda^d. \tag{50}$$

For the upper bound, note that each Σ_k is positive semidefinite and

$$\mathrm{tr}(\Sigma_k) = \mathbb{E}_{\mu^k}\left[\|\phi(x)\|_2^2\right] \leqslant \kappa^2.$$

Therefore

$$\operatorname{tr}(M_n) = d\lambda + \sum_{k=1}^{n} \operatorname{tr}(\Sigma_k) \leqslant d\lambda + n\kappa^2.$$

By the arithmetic-geometric mean inequality for eigenvalues,

$$\det(M_n) \leqslant \left(\frac{\operatorname{tr}(M_n)}{d}\right)^d \leqslant \left(\lambda + \frac{n\kappa^2}{d}\right)^d. \tag{51}$$

Combining (50) and (51) yields

$$2^{n/d} \leqslant 1 + \frac{n\kappa^2}{d\lambda}. \tag{52}$$

Let $y := n/d$ and $a := \kappa^2/\lambda = \kappa^2 R^2/\epsilon^2$, so (52) becomes

$$2^y \leqslant 1 + ay.$$

We claim that this implies $y \leqslant 4\log_2(2+a)$. Indeed, if $y > 4\log_2(2+a)$, then $y \geqslant 4$ and $y > 2\log_2(1+a)$. Since $y \geqslant 4$, we have $y \leqslant 2^{y/2}$. Thus

$$1 + ay \leqslant 1 + a\, 2^{y/2} \leqslant (1+a)2^{y/2} < 2^y,$$

where the strict inequality uses $2^{y/2} > 1 + a$. This contradicts (52). Therefore

$$n = dy \leqslant 4d\log_2\left(2 + \frac{\kappa^2 R^2}{\epsilon^2}\right),$$

which proves (48). □

Corollary 38 (Linear robust MG-BE upper bound)**.** *Suppose that for every stage $h \in [H]$, the exact robust residual class $\mathscr{P}_h = (I - \mathcal{T}_h^\sigma)\mathcal{F}$ is contained in a linear class of the form*

$$\mathscr{P}_h \subseteq \left\{x \mapsto \langle \theta, \phi_{W,h}(x)\rangle : \theta \in \mathbb{R}^{d_{W,h}},\ \|\theta\|_2 \leqslant R_{W,h}\right\},$$

with $\sup_x \|\phi_{W,h}(x)\|_2 \leqslant \kappa_{W,h}$. Then for every $\epsilon > 0$,

$$\dim_{\mathrm{BE},2}^{\text{rob-MG}}(\mathcal{F}, \Pi, \epsilon) \leqslant 4 \max_{h\in[H]} \left[d_{W,h}\log_2\left(2 + \frac{\kappa_{W,h}^2 R_{W,h}^2}{\epsilon^2}\right)\right].$$

Proof. For each stage h, the square-loss DE dimension is monotone in the function class: if $\mathscr{P}_h \subseteq \bar{\mathscr{P}}_h$, then every ϵ-independent sequence for $\mathscr{P}_h$ is also ϵ-independent for $\bar{\mathscr{P}}_h$. Applying Proposition 37 to the ambient linear class at stage h therefore gives

$$\dim_{\mathrm{DE},2}(\mathscr{P}_h, \Pi_h, \epsilon) \leqslant 4d_{W,h}\log_2\left(2 + \frac{\kappa_{W,h}^2 R_{W,h}^2}{\epsilon^2}\right).$$

Taking the maximum over $h \in [H]$ and using Definition 16 proves the claim. □

Corollary 39 (Tabular classes)**.** *Suppose $\mathcal{S}, \mathcal{A}, \mathcal{B}$ are finite and the value class, dual class, and witness class are tabular with coordinate-wise bounds $\|f_h\|_\infty \leqslant B_{F,h}$, $\|g\|_\infty \leqslant B_{G,h}$, and $\|w\|_\infty \leqslant B_{W,h}$. Let $d_{\mathrm{tab}} := |\mathcal{S}||\mathcal{A}||\mathcal{B}|$. Then*

$$\log \mathcal{N}_F(\eta_F) \leqslant d_{\mathrm{tab}}\log\left(1 + \frac{2B_{F,h}}{\eta_F}\right),$$
$$\log \mathcal{N}_G(\eta_G) \leqslant 2d_{\mathrm{tab}}\log\left(1 + \frac{2B_{G,h}}{\eta_G}\right),$$
$$\log \mathcal{N}_W(\eta_W) \leqslant d_{\mathrm{tab}}\log\left(1 + \frac{2B_{W,h}}{\eta_W}\right).$$

Proof. Use the one-hot feature map on $\mathcal{X}_h$ for $\mathcal{F}_h$ and $\mathcal{W}_h$, and the concatenated one-hot parameterization for the dual pair (g_η, g_ν). Then apply Corollary 36 with $\kappa = 1$. □

Corollary 40 (Tabular robust MG-BE upper bound)**.** *Suppose $\mathcal{S}, \mathcal{A}, \mathcal{B}$ are finite and let $d_{\mathrm{tab}} := |\mathcal{S}||\mathcal{A}||\mathcal{B}|$. Assume that for every stage $h \in [H]$ and every exact robust residual $\psi \in \mathscr{P}_h$,*

$$\|\psi\|_\infty \leqslant B_{\mathrm{res},h}.$$

Then for every $\epsilon > 0$,

$$\dim_{\mathrm{BE},2}^{\text{rob-MG}}(\mathcal{F}, \Pi, \epsilon) \leqslant 4 d_{\mathrm{tab}} \max_{h \in [H]} \log_2\left(2 + \frac{d_{\mathrm{tab}} B_{\mathrm{res},h}^2}{\epsilon^2}\right). \tag{53}$$

Proof. For each stage h, every tabular residual $\psi : \mathcal{X}_h \to \mathbb{R}$ with $\|\psi\|_\infty \leqslant B_{\mathrm{res},h}$ admits the one-hot representation

$$\psi(x) = \langle \theta_\psi, e_x \rangle, \qquad \|\theta_\psi\|_2 \leqslant \sqrt{d_{\mathrm{tab}}}\, B_{\mathrm{res},h}, \qquad \|e_x\|_2 = 1.$$

Thus $\mathscr{P}_h$ is contained in a d_{tab}-dimensional linear class with $R_{W,h} = \sqrt{d_{\mathrm{tab}}}\, B_{\mathrm{res},h}$ and $\kappa_{W,h} = 1$. Applying Corollary 38 gives (53). □

Corollary 41 (Finite-rank kernel classes)**.** *Suppose the value, dual, and witness classes admit explicit finite-dimensional kernel feature maps $\Phi_{F,h} : \mathcal{X}_h \to \mathbb{R}^{m_{F,h}}$, $\Phi_{G,h} : \mathcal{X}_h \to \mathbb{R}^{m_{G,h}}$, and $\Phi_{W,h} : \mathcal{X}_h \to \mathbb{R}^{m_{W,h}}$, with uniform norm bounds $\|\Phi_{F,h}(x)\|_2 \leqslant \kappa_{F,h}$, $\|\Phi_{G,h}(x)\|_2 \leqslant \kappa_{G,h}$, and $\|\Phi_{W,h}(x)\|_2 \leqslant \kappa_{W,h}$. If the corresponding RKHS coefficients are restricted to Euclidean balls of radii $R_{F,h}, R_{G,h}, R_{W,h}$, then the covering numbers satisfy the same bounds as in Corollary 36 with $d_{F,h} = m_{F,h}$, $d_{G,h} = m_{G,h}$, and $d_{W,h} = m_{W,h}$.*

Proof. An explicit finite-rank kernel class is simply a linear class in the corresponding feature coordinates. Apply Corollary 36. □

Corollary 42 (Finite-rank kernel robust MG-BE upper bound)**.** *Suppose that for every stage $h \in [H]$, the exact robust residual class $\mathscr{P}_h$ admits a finite-rank kernel feature map $\Phi_{W,h} : \mathcal{X}_h \to \mathbb{R}^{m_{W,h}}$ with $\|\Phi_{W,h}(x)\|_2 \leqslant \kappa_{W,h}$ and coefficient radius $R_{W,h}$. Then for every $\epsilon > 0$,*

$$\dim_{\mathrm{BE},2}^{\text{rob-MG}}(\mathcal{F}, \Pi, \epsilon) \leqslant 4 \max_{h \in [H]} \left[m_{W,h} \log_2\left(2 + \frac{\kappa_{W,h}^2 R_{W,h}^2}{\epsilon^2}\right)\right].$$

Proof. A finite-rank kernel residual class is a linear residual class in the feature coordinates $\Phi_{W,h}$. Apply Corollary 38 with $d_{W,h} = m_{W,h}$. □

Tabular instantiation. Assume the hypotheses of Corollary 12. Suppose further that $\mathcal{S}, \mathcal{A}, \mathcal{B}$ are finite and the exact robust residuals satisfy the tabular bound in Corollary 40. Let $B_{\mathrm{res}} := \max_{h \in [H]} B_{\mathrm{res},h}$ and $d_{\mathrm{tab}} := |\mathcal{S}||\mathcal{A}||\mathcal{B}|$. Then

$$\mathrm{Reg}(K) \leqslant \widetilde{O}\left(H B_\phi(\sigma) \sqrt{d_{\mathrm{tab}} K \log\bigl(2 + d_{\mathrm{tab}} B_{\mathrm{res}}^2 K\bigr)}\right) + HK\epsilon_{\mathrm{appr}} + HK\epsilon_{\mathrm{eq}} + HK\epsilon_{\mathrm{dual}}. \tag{54}$$

Indeed, Corollary 40 with $\epsilon = 1/\sqrt{K}$ gives

$$\dim_{\mathrm{BE},2}^{\text{rob-MG}}(\mathcal{F}, \Pi, 1/\sqrt{K}) \leqslant 4 d_{\mathrm{tab}} \log_2\bigl(2 + d_{\mathrm{tab}} B_{\mathrm{res}}^2 K\bigr),$$

and substituting this estimate into Corollary 12 yields (54).

Linear instantiation. Assume the hypotheses of Corollary 12. Suppose further that the exact robust residual classes satisfy the linear inclusion in Corollary 38. Let

$$d_W := \max_{h \in [H]} d_{W,h}, \qquad \kappa_W := \max_{h \in [H]} \kappa_{W,h}, \qquad R_W := \max_{h \in [H]} R_{W,h}.$$

Then

$$\mathrm{Reg}(K) \leqslant \widetilde{O}\left(H B_\phi(\sigma) \sqrt{d_W K \log\bigl(2 + \kappa_W^2 R_W^2 K\bigr)}\right) + HK\epsilon_{\mathrm{appr}} + HK\epsilon_{\mathrm{eq}} + HK\epsilon_{\mathrm{dual}}. \tag{55}$$

Indeed, Corollary 38 with $\epsilon = 1/\sqrt{K}$ gives

$$\dim_{\mathrm{BE},2}^{\text{rob-MG}}(\mathcal{F}, \Pi, 1/\sqrt{K}) \leqslant 4 d_W \log_2\bigl(2 + \kappa_W^2 R_W^2 K\bigr),$$

and substituting this bound into Corollary 12 yields (55).

D Structured Specializations and Projected Extensions

D.1 A Fully Closed Tabular Specialization

The generic backward-validation route is intentionally designed for general function approximation. In the tabular case, however, one can simplify the confidence construction substantially and close the route without any residual-class abstraction. The key simplification is that the robust backup can be estimated directly from empirical next-state tables, and optimism can be enforced by pointwise count-based bonuses instead of function-space backward validation.

We restrict this subsection to TV uncertainty sets on a finite state space. Let $\hat{P}_{h,k-1}(\cdot \mid s,a,b)$ be the empirical nominal transition distribution formed from all stage-h samples collected at the triple (s,a,b) before episode k, and let $N_{h,k-1}(s,a,b)$ be the corresponding count. For a bounded continuation value $V : \mathcal{S} \to [0,H]$, define the empirical robust backup by

$$\big(\hat{\mathcal{T}}^{\sigma,\mathrm{tab}}_{h,k} V\big)(s,a,b) := r_h(s,a,b) + \inf_{Q:\, D_{\mathrm{TV}}(Q,\hat{P}_{h,k-1}(\cdot|s,a,b))\leqslant\sigma} \mathbb{E}_Q[V].$$

The tabular specialization uses the confidence set

$$\mathcal{F}^{(k)}_{\mathrm{tab}} := \left\{ f : \begin{array}{l} 0 \leqslant f_h(s,a,b) \leqslant H-h+1 \text{ for all } (h,s,a,b), \\ f_h(s,a,b) \leqslant \big(\hat{\mathcal{T}}^{\sigma,\mathrm{tab}}_{h,k} V_{f,h+1}\big)(s,a,b) + b_{h,k}(s,a,b) \end{array} \right\}, \tag{56}$$

and selects $f^{(k)} \in \operatorname{argmax}_{f\in\mathcal{F}^{(k)}_{\mathrm{tab}}} V_{f,1}(s_1)$.

Lemma 43 (TV-center Lipschitzness of the robust expectation)**.** *Let $V : \mathcal{S} \to [0,H]$ be any bounded function and define*

$$\Psi_\sigma(P;V) := \inf_{Q:\, D_{\mathrm{TV}}(Q,P)\leqslant\sigma} \mathbb{E}_Q[V].$$

Then for any two distributions $P, \hat{P} \in \Delta(\mathcal{S})$,

$$\big|\Psi_\sigma(P;V) - \Psi_\sigma(\hat{P};V)\big| \leqslant H\|P-\hat{P}\|_1.$$

Proof. For TV balls on a finite state space, the worst-case expectation admits the classical form $\Psi_\sigma(P;V) = \mathbb{E}_P[V] - \sigma\,\mathrm{span}(V)$ up to the usual simplex truncation, and the truncation only decreases the magnitude of perturbations. Therefore

$$\big|\Psi_\sigma(P;V) - \Psi_\sigma(\hat{P};V)\big| \leqslant \big|\mathbb{E}_P[V] - \mathbb{E}_{\hat{P}}[V]\big| \leqslant \|V\|_\infty \|P-\hat{P}\|_1 \leqslant H\|P-\hat{P}\|_1.$$

□

Lemma 44 (Uniform tabular transition confidence)**.** *Fix $\delta \in (0,1)$. With probability at least $1-\delta$, simultaneously for all stages $h \in [H]$, episodes $k \in [K]$, and triples $(s,a,b) \in \mathcal{S}\times\mathcal{A}\times\mathcal{B}$,*

$$\|\hat{P}_{h,k-1}(\cdot \mid s,a,b) - P^0_h(\cdot \mid s,a,b)\|_1 \leqslant \beta^{\mathrm{tab}}_{h,k}(s,a,b;\delta), \tag{57}$$

where

$$\beta^{\mathrm{tab}}_{h,k}(s,a,b;\delta) := \sqrt{\frac{2\big(|\mathcal{S}|\log 2 + \log(2HK|\mathcal{S}||\mathcal{A}||\mathcal{B}|/\delta)\big)}{N_{h,k-1}(s,a,b)\vee 1}}.$$

Proof. Apply the Weissman inequality to each empirical next-state table $\hat{P}_{h,k-1}(\cdot \mid s,a,b)$, then union bound over $h \in [H]$, $k \in [K]$, and $(s,a,b) \in \mathcal{S}\times\mathcal{A}\times\mathcal{B}$. □

Corollary 45 (Uniform tabular robust-backup confidence)**.** *On the event of Lemma 44, simultaneously for all k,h,s,a,b and all bounded f with $0 \leqslant V_{f,h+1} \leqslant H$,*

$$\Big|\big(\hat{\mathcal{T}}^{\sigma,\mathrm{tab}}_{h,k} V_{f,h+1}\big)(s,a,b) - \big(\mathcal{T}^\sigma_h f_{h+1}\big)(s,a,b)\Big| \leqslant b^{\mathrm{tab}}_{h,k}(s,a,b;\delta), \tag{58}$$

where

$$b^{\mathrm{tab}}_{h,k}(s,a,b;\delta) := H\,\beta^{\mathrm{tab}}_{h,k}(s,a,b;\delta).$$

Proof. By Lemma 43,

$$\left|\big(\hat{\mathcal{T}}^{\sigma,\mathrm{tab}}_{h,k} V_{f,h+1}\big)(s,a,b) - \big(\mathcal{T}^{\sigma}_{h} f_{h+1}\big)(s,a,b)\right| \leqslant H\,\|\hat{P}_{h,k-1}(\cdot \mid s,a,b) - P^0_h(\cdot \mid s,a,b)\|_1.$$

Substitute (57). □

Proposition 46 (Tabular optimism and residual control)**.** *Assume the exact robust optimal action-value table $f^\star$ satisfies $f^\star_h = \mathcal{T}^\sigma_h f^\star_{h+1}$ and $0 \leqslant f^\star_h \leqslant H - h + 1$ for every stage h. On the event of Corollary 45, the following hold for every episode k:*

(i) $f^\star \in \mathcal{F}^{(k)}_{\mathrm{tab}}$, hence $V^{\star,\sigma}_1(s_1) \leqslant V_{f^{(k)},1}(s_1)$;

(ii) for every stage h and every triple (s,a,b),

$$f^{(k)}_h(s,a,b) - \big(\mathcal{T}^\sigma_h f^{(k)}_{h+1}\big)(s,a,b) \leqslant 2b^{\mathrm{tab}}_{h,k}(s,a,b;\delta). \tag{59}$$

Proof. For the first claim, Bellman optimality and (58) imply

$$f^\star_h(s,a,b) = \big(\mathcal{T}^\sigma_h f^\star_{h+1}\big)(s,a,b) \leqslant \big(\hat{\mathcal{T}}^{\sigma,\mathrm{tab}}_{h,k} V_{f^\star,h+1}\big)(s,a,b) + b^{\mathrm{tab}}_{h,k}(s,a,b;\delta),$$

so $f^\star \in \mathcal{F}^{(k)}_{\mathrm{tab}}$. Since $f^{(k)}$ maximizes $V_{f,1}(s_1)$ over that set, optimism follows.

For the second claim, because $f^{(k)} \in \mathcal{F}^{(k)}_{\mathrm{tab}}$,

$$f^{(k)}_h(s,a,b) \leqslant \big(\hat{\mathcal{T}}^{\sigma,\mathrm{tab}}_{h,k} V_{f^{(k)},h+1}\big)(s,a,b) + b^{\mathrm{tab}}_{h,k}(s,a,b;\delta).$$

Applying (58) once more gives

$$f^{(k)}_h(s,a,b) - \big(\mathcal{T}^\sigma_h f^{(k)}_{h+1}\big)(s,a,b) \leqslant 2b^{\mathrm{tab}}_{h,k}(s,a,b;\delta).$$

□

Lemma 47 (Tabular harmonic bonus sum)**.** *Fix a stage h and let $X_{h,k} = (s^k_h, a^k_h, b^k_h)$ denote the realized stage-h triple in episode k. Then, pathwise,*

$$\sum_{k=1}^{K} b^{\mathrm{tab}}_{h,k}(X_{h,k};\delta) \leqslant 3H\sqrt{2|\mathcal{S}|\log\big(2HK|\mathcal{S}||\mathcal{A}||\mathcal{B}|/\delta\big)}\,\sqrt{|\mathcal{S}||\mathcal{A}||\mathcal{B}|\,K}. \tag{60}$$

Proof. For each $x \in \mathcal{S}\times\mathcal{A}\times\mathcal{B}$, let $N_{h,K}(x)$ be the total number of stage-h visits to x over the K episodes. Every visit to x contributes a term of the form $c/\sqrt{n \vee 1}$ with n equal to the number of previous visits. Hence

$$\sum_{k=1}^{K} \mathbf{1}\{X_{h,k} = x\}\,\frac{1}{\sqrt{N_{h,k-1}(x)\vee 1}} \leqslant 1 + \sum_{n=1}^{N_{h,K}(x)-1} \frac{1}{\sqrt{n}} \leqslant 3\sqrt{N_{h,K}(x)}.$$

Summing over all x and using Cauchy–Schwarz,

$$\sum_x \sqrt{N_{h,K}(x)} \leqslant \sqrt{|\mathcal{S}||\mathcal{A}||\mathcal{B}|\sum_x N_{h,K}(x)} = \sqrt{|\mathcal{S}||\mathcal{A}||\mathcal{B}|\,K}.$$

Multiplying by the common prefactor in $b^{\mathrm{tab}}_{h,k}$ proves (60). □

Theorem 48 (Fully closed tabular specialization)**.** *Consider the tabular specialization* (56) *under TV uncertainty sets and suppose the approximate stage-game solver incurs error at most ϵ_{eq} per stage. Then on the event of Lemma 44,*

$$\mathrm{Reg}(K) \leqslant 2\sum_{k=1}^{K}\sum_{h=1}^{H} \mathbb{E}\big[b^{\mathrm{tab}}_{h,k}(X_{h,k};\delta) \mid \mathcal{H}_{k-1}\big] + HK\epsilon_{\mathrm{eq}}.$$

Consequently,

$$\mathbb{E}[\mathrm{Reg}(K)] \leqslant \widetilde{O}\Big(H^2|\mathcal{S}|\sqrt{|\mathcal{A}||\mathcal{B}|K}\Big) + HK\epsilon_{\mathrm{eq}} + \delta HK. \tag{61}$$

In particular, setting $\delta = 1/K$ yields $\mathbb{E}[\mathrm{Reg}(K)] = \widetilde{O}(H^2|\mathcal{S}|\sqrt{|\mathcal{A}||\mathcal{B}|K}) + HK\epsilon_{\mathrm{eq}}$.

Proof. Fix an episode k. By Proposition 46, the chosen table is optimistic: $V_1^{\star,\sigma}(s_1) \leqslant V_{f^{(k)},1}(s_1)$. Applying Lemma 8 with $\delta_h^k = f_h^{(k)} - \mathcal{T}_h^\sigma f_{h+1}^{(k)}$ and then using (59) gives

$$V_1^{\star,\sigma}(s_1) - V_1^{\pi^k,\nu^k,\sigma}(s_1) \leqslant 2\sum_{h=1}^{H} \mathbb{E}_{\mu_h^k}\left[b_{h,k}^{\mathrm{tab}}(s,a,b;\delta)\right] + H\epsilon_{\mathrm{eq}}.$$

Conditioned on the history $\mathcal{H}_{k-1}$, the stage-h occupancy expectation equals the conditional expectation of the realized stage-h triple:

$$\mathbb{E}_{\mu_h^k}\left[b_{h,k}^{\mathrm{tab}}(s,a,b;\delta)\right] = \mathbb{E}\left[b_{h,k}^{\mathrm{tab}}(X_{h,k};\delta) \mid \mathcal{H}_{k-1}\right].$$

Summing over k proves the first display.

For the expected-regret bound, split according to the good event $\mathcal{G}_\delta$ of Lemma 44:

$$\mathbb{E}[\mathrm{Reg}(K)] \leqslant \mathbb{E}[\mathbf{1}\{\mathcal{G}_\delta\}\mathrm{Reg}(K)] + \mathbb{E}[\mathbf{1}\{\mathcal{G}_\delta^c\}\mathrm{Reg}(K)].$$

The second term is at most δHK because per-episode regret is at most H. On $\mathcal{G}_\delta$, use the first display and then Lemma 47 stage by stage. This yields (61). □

D.2 Adaptive Exploiter-Mix and Reused Data

We now extend the analysis from the on-policy core to the actual adaptive data-collection rule. Let $\mathcal{H}_{k-1}$ denote the history before episode k. The algorithm first computes a target policy pair (π^k, ν^k) from $f^{(k)}$, then draws an executed pair $(\widetilde{\pi}^k, \widetilde{\nu}^k)$ according to the exploiter-mix rule. Let $\widetilde{\mu}_h^k$ be the resulting stage-h occupancy law under the worst-case kernel selector for that executed pair, and let $\widetilde{\eta}_h^k$ be the actual stage-h data-collection occupancy marginal generated by that executed pair under the interaction process.

Proposition 49 (Adaptive-data validity of the backward-validation lemmas)**.** *Suppose that for every episode k, the executed pair $(\widetilde{\pi}^k, \widetilde{\nu}^k)$ is $\mathcal{H}_{k-1}$-measurable, and stage-h samples in episode k are drawn from the conditional law $\widetilde{\eta}_h^k(ds, da, db)\, P_h^0(ds' \mid s, a, b)$. Then the conclusions of Lemma 20, Lemma 21, and Corollaries 22–23 remain valid with μ_h^j replaced everywhere by $\widetilde{\eta}_h^j$. Consequently, rerunning Proposition 26 and Corollary 27 with $\widetilde{\eta}_h^j$ in place of μ_h^j yields the same reference-retention and optimism-closure conclusions for the adaptive data sequence.*

Proof. Each of those proofs already proceeds through bounded martingale-difference sequences whose conditional means are taken with respect to the stage-wise occupancy law of the currently executed episode. Replacing the deterministic on-policy sequence by an arbitrary history-measurable executed sequence therefore changes only the notation of the conditional law, not the argument. □

Assumption 16 (Coverage dominance from exploiter mixing)**.** *For every stage h and episode $k \geqslant 2$, the executed occupancy $\widetilde{\mu}_h^k$ is absolutely continuous with respect to the pooled past data-collection mixture*

$$\bar{\widetilde{\eta}}_{h,k} := \sum_{j=1}^{k-1} \lambda_{h,j}^{(k)} \widetilde{\eta}_h^j, \qquad \lambda_{h,j}^{(k)} := \frac{n_{h,j}}{N_{h,<k}},$$

and there exists a constant $C_h^{\mathrm{mix}} \geqslant 1$ such that

$$\frac{d\widetilde{\mu}_h^k}{d\bar{\widetilde{\eta}}_{h,k}} \leqslant C_h^{\mathrm{mix}} \qquad \bar{\widetilde{\eta}}_{h,k}\text{-almost surely.}$$

Corollary 50 (Current-occupancy operator budget from pooled adaptive ERM)**.** *Under Assumptions 10–13 and 16, with probability at least $1-\delta$, simultaneously for all $f \in \mathcal{F}_{h+1}$,*

$$\left\| \widehat{\mathcal{T}}_{h,f}^{\sigma,(k)} f_{h+1} - \mathcal{T}_h^\sigma f_{h+1} \right\|_{1,\widetilde{\mu}_h^k} \leqslant C_h^{\mathrm{mix}} \Big(2\bar{\Delta}_{h,k}^{\mathrm{cov}}(\delta; \eta_F, \eta_G) + \epsilon_{\mathrm{dual}} \Big). \tag{62}$$

Therefore Assumption 6 holds with

$$\alpha_{h,k} := 2C_h^{\mathrm{mix}} \bar{\Delta}_{h,k}^{\mathrm{cov}}(\delta; \eta_F, \eta_G) + \big(C_h^{\mathrm{mix}} - 1\big)\epsilon_{\mathrm{dual}}.$$

Proof. Fix $f \in \mathcal{F}_{h+1}$ and define $e_{h,k}(f) := \widehat{\mathcal{T}}^{\sigma,(k)}_{h,f} f_{h+1} - \mathcal{T}^{\sigma}_{h} f_{h+1}$. By the ambient dual representation, $e_{h,k}(f) \geqslant 0$ pointwise. Hence

$$\begin{aligned}\|e_{h,k}(f)\|_{1,\widetilde{\mu}^k_h} &= \mathbb{E}_{\widetilde{\eta}_{h,k}}\left[\frac{d\widetilde{\mu}^k_h}{d\widetilde{\eta}_{h,k}} e_{h,k}(f)\right] \\ &\leqslant C^{\mathrm{mix}}_h \mathbb{E}_{\widetilde{\eta}_{h,k}}\left[e_{h,k}(f)\right].\end{aligned}$$

By Proposition 49, the pooled dual-ERM argument continues to hold with $\widetilde{\eta}^j_h$ in place of μ^j_h, so the same proof as Corollary 23 yields

$$\mathbb{E}_{\widetilde{\eta}_{h,k}}\left[e_{h,k}(f)\right] \leqslant 2\bar{\Delta}^{\mathrm{cov}}_{h,k}(\delta; \eta_F, \eta_G) + \epsilon_{\mathrm{dual}}.$$

Multiplying by C^{mix}_h proves (62). The final display is just a rearrangement into the form required by Assumption 6. □

Proposition 51 (Target regret versus executed regret under exploiter mix)**.** *Assume rewards lie in* $[0,1]$*, so every robust value is in* $[0,H]$*. Let*

$$\widetilde{\mathrm{Reg}}(K) := \sum_{k=1}^{K}\left(V_1^{\star,\sigma}(s_1) - \mathbb{E}\Big[V_1^{\widetilde{\pi}^k,\widetilde{\nu}^k,\sigma}(s_1) \mid \mathcal{H}_{k-1}\Big]\right).$$

Then the regret of the target planned-pair sequence satisfies

$$\sum_{k=1}^{K}\Big(V_1^{\star,\sigma}(s_1) - V_1^{\pi^k,\nu^k,\sigma}(s_1)\Big) \leqslant \frac{\widetilde{\mathrm{Reg}}(K) + pHK}{1-p}. \tag{63}$$

Proof. Condition on $\mathcal{H}_{k-1}$. By definition of the exploiter-mix rule,

$$\begin{aligned}\mathbb{E}\Big[V_1^{\widetilde{\pi}^k,\widetilde{\nu}^k,\sigma}(s_1) \mid \mathcal{H}_{k-1}\Big] &= (1-p)V_1^{\pi^k,\nu^k,\sigma}(s_1) \\ &\quad + \frac{p}{2}V_1^{\pi^k,\mathrm{BR}^k_2,\sigma}(s_1) + \frac{p}{2}V_1^{\mathrm{BR}^k_1,\nu^k,\sigma}(s_1).\end{aligned}$$

Each value term is at most H, so

$$\mathbb{E}\Big[V_1^{\widetilde{\pi}^k,\widetilde{\nu}^k,\sigma}(s_1) \mid \mathcal{H}_{k-1}\Big] \leqslant (1-p)V_1^{\pi^k,\nu^k,\sigma}(s_1) + pH.$$

Rearranging gives

$$V_1^{\pi^k,\nu^k,\sigma}(s_1) \geqslant \frac{\mathbb{E}\Big[V_1^{\widetilde{\pi}^k,\widetilde{\nu}^k,\sigma}(s_1) \mid \mathcal{H}_{k-1}\Big] - pH}{1-p}.$$

Therefore

$$\begin{aligned}(1-p)\Big(V_1^{\star,\sigma}(s_1) - V_1^{\pi^k,\nu^k,\sigma}(s_1)\Big) &\leqslant V_1^{\star,\sigma}(s_1) - \mathbb{E}\Big[V_1^{\widetilde{\pi}^k,\widetilde{\nu}^k,\sigma}(s_1) \mid \mathcal{H}_{k-1}\Big] + p\Big(H - V_1^{\star,\sigma}(s_1)\Big) \\ &\leqslant V_1^{\star,\sigma}(s_1) - \mathbb{E}\Big[V_1^{\widetilde{\pi}^k,\widetilde{\nu}^k,\sigma}(s_1) \mid \mathcal{H}_{k-1}\Big] + pH,\end{aligned}$$

because $0 \leqslant V_1^{\star,\sigma}(s_1) \leqslant H$. Summing over k and dividing by $1-p$ proves (63). □

Corollary 52 (Adaptive exploiter-mix reduction)**.** *Replace the on-policy occupancies* μ^k_h *in the deterministic reduction by the executed occupancies* $\widetilde{\mu}^k_h$*. Under the assumptions of Proposition 49 and Corollary 50, the same deterministic Bellman telescope and operator-learning decomposition hold for* $\widetilde{\mathrm{Reg}}(K)$*. Combining it with Proposition 51 yields a regret bound for the target policy sequence* $\{(\pi^k,\nu^k)\}_{k=1}^K$ *with an additional multiplicative factor* $1/(1-p)$ *and an additive exploration cost* $pHK/(1-p)$*, provided the residual-budget premise is verified directly on the executed occupancy sequence.*

Proof. The deterministic Bellman telescoping proof only depends on the executed policy pair and its occupancy measure. Thus it applies verbatim with $(\widetilde{\pi}^k, \widetilde{\nu}^k, \widetilde{\mu}^k_h)$ in place of the on-policy sequence. The statistical ingredients are provided by Proposition 49 and Corollary 50; the former already re-establishes the reference-retention and optimism-closure argument on the adaptive data sequence, while the latter transfers the operator term to the executed occupancies. Any residual-budget premise on $\widetilde{\mu}^k_h$ must still be supplied separately by the downstream theorem. Finally, Proposition 51 transfers the executed-pair bound to the target planned-pair sequence. □

D.3 General GFA Theorem Under Interpretable Assumptions

We now package the abstract regret reduction into a single theorem whose assumptions are aligned with the template used by Ghosh et al. [6]. The robust Markov-game counterparts are:

1. robust minimax completeness together with a bounded value envelope;
2. a Bellman-consistent reference function for optimism closure;
3. bounded dual losses, finite covers of the dual and plug-in witness classes, and dual approximation error;
4. support control and stable stage-game extraction;
5. an adaptive coverage-dominance condition for exploiter-mix data reuse.

The theorem below is written as a conditional interface: optimism and operator transfer are discharged by labeled results in this appendix, while the executed-occupancy residual premise is stated explicitly rather than claimed from the original adaptive one-update-per-episode route.

Theorem 53 (General GFA theorem under interpretable assumptions)**.** *Fix $\delta \in (0,1)$ and set $\delta_0 := \delta/(4HK)$. Assume:*

(i) rewards lie in $[0,1]$, support control holds, and the stage-game solver incurs value error at most ϵ_{eq} per stage;

(ii) robust minimax completeness holds stagewise and $\sup_{f\in\mathcal{F}} \|f_h\|_\infty \leqslant B_{F,h}$ for every $h \in [H]$;

(iii) Assumption 14 holds;

(iv) Assumptions 8, 9, 10, 11, 12, and 13 hold stagewise; in particular, Proposition 19 gives a sufficient condition for 9;

(v) Proposition 49 applies to the executed policy sequence and Assumption 16 holds;

(vi) the algorithm uses the backward-validation confidence-set update from the main method section, and the corresponding radii are calibrated so that for every stage $h \in [H]$ and every episode $k \geqslant 2$,

$$\beta_{h,k} \geqslant \xi_{h,k}(\delta_0; \eta_W) + \rho_{h,k}(\delta_0; \eta_F, \eta_G), \tag{64}$$

(vii) on an event of probability at least $1-\delta/4$, the selected exact residual witnesses $\psi_h^k := \delta_h(f^{(k)})$ satisfy

$$\sum_{j<k} \mathbb{E}_{\widetilde{\mu}_h^j}\left[(\psi_h^k)^2\right] \leqslant \epsilon_{\mathrm{BV}}^2 \tag{65}$$

for every $h \in [H]$ and every $k \geqslant 2$, for some $\epsilon_{\mathrm{BV}} > 0$.

Define

$$d_{\mathrm{GFA}}(\epsilon_{\mathrm{BV}}) := \dim^{\text{rob-MG}}_{\mathrm{BE},2}(\mathcal{F}, \Pi, \epsilon_{\mathrm{BV}}),$$

where Π_h is any family of measures containing all executed stage-h occupancies, and let

$$B_F := \max_{h\in[H]} B_{F,h}, \qquad C_{\mathrm{mix}} := \max_{h\in[H]} C_h^{\mathrm{mix}},$$

and

$$\Upsilon_{\mathrm{dual}}(K,\delta) := \sum_{k=2}^{K}\sum_{h=1}^{H} \bar{\Delta}^{\mathrm{cov}}_{h,k}(\delta_0; \eta_F, \eta_G).$$

Then with probability at least $1-\delta$, the target-policy regret satisfies

$$\begin{aligned}\mathrm{Reg}(K) \leqslant H + \frac{1}{1-p}\Big[& K\epsilon_{\mathrm{appr}} + HK\epsilon_{\mathrm{eq}} \\ & + \sqrt{KH}\,\sqrt{2HK\epsilon_{\mathrm{BV}}^2 + 4HB_F^2 d_{\mathrm{GFA}}(\epsilon_{\mathrm{BV}})} \\ & + 2C_{\mathrm{mix}}\Upsilon_{\mathrm{dual}}(K,\delta) + HKC_{\mathrm{mix}}\epsilon_{\mathrm{dual}} + pHK\Big].\end{aligned} \tag{66}$$

Proof. Let $\mathcal{G}$ be the intersection over all $h \in [H]$ and $k \in \{2, \dots, K\}$ of the good events from Lemma 20, Corollary 23, and Corollary 50, each invoked with confidence parameter δ_0, together with the high-probability event in (65). By a union bound, $\mathbb{P}(\mathcal{G}) \geqslant 1 - \delta$.

Fix the event $\mathcal{G}$. On $\mathcal{G}$, (64) and Corollary 27 imply that the optimistic choice is valid for every episode $k \geqslant 2$. By assumption (65), the selected exact residual witness $\psi_h^k := \delta_h(f^{(k)})$ satisfies $\sum_{j<k} \mathbb{E}_{\widetilde{\mu}_h^j}\left[(\psi_h^k)^2\right] \leqslant \epsilon_{\mathrm{BV}}^2$ for every h and every $k \geqslant 2$.

By Lemma 34, every exact residual in $\mathscr{P}_h = (I - \mathcal{T}_h^\sigma)\mathcal{F}$ has sup norm at most $2B_{F,h}$. Applying Corollary 18 stage by stage to the executed occupancy sequence $\{\widetilde{\mu}_h^k\}_{k=2}^K$ therefore gives

$$\Gamma_{\mathrm{res}}(K) \leqslant 2HK\epsilon_{\mathrm{BV}}^2 + 4\sum_{h=1}^{H} B_{F,h}^2 \dim_{\mathrm{DE},2}(\mathscr{P}_h, \Pi_h, \epsilon_{\mathrm{BV}}) \leqslant 2HK\epsilon_{\mathrm{BV}}^2 + 4HB_F^2 d_{\mathrm{GFA}}(\epsilon_{\mathrm{BV}}). \tag{67}$$

Next, Corollary 50 yields $\alpha_{h,k} = 2C_h^{\mathrm{mix}} \bar{\Delta}_{h,k}^{\mathrm{cov}}(\delta_0; \eta_F, \eta_G) + (C_h^{\mathrm{mix}} - 1)\epsilon_{\mathrm{dual}}$ for every h and $k \geqslant 2$. Hence

$$\sum_{k=2}^{K}\sum_{h=1}^{H} \alpha_{h,k} \leqslant 2C_{\mathrm{mix}} \Upsilon_{\mathrm{dual}}(K, \delta) + HK(C_{\mathrm{mix}} - 1)\epsilon_{\mathrm{dual}}. \tag{68}$$

Apply the same proof as Theorem 10 to the executed sequence over episodes $k = 2, \dots, K$, using (67), (68), and the optimism conclusion above. This bounds the executed regret from episodes $2, \dots, K$ by the bracketed term in (66) without the final pHK. Proposition 51 then transfers the executed-sequence bound to the target policy sequence, introducing the factor $1/(1-p)$ and the additive exploiter cost $pHK/(1-p)$. Finally, the first episode contributes at most H regret because all robust values lie in $[0, H]$. Combining these bounds proves (66). □

Theorem 54 (Projected GFA theorem under interpretable assumptions)**.** *Fix stagewise maps $\{\mathsf{P}_h\}_{h=1}^H$ and define the projected residual class*

$$\mathscr{P}_h^{\mathsf{P}} := \{f_h - \mathsf{P}_h \mathcal{T}_h^\sigma f_{h+1} : f \in \mathcal{F}\}.$$

Let

$$d_{\mathrm{GFA}}^{\mathsf{P}}(\epsilon_{\mathrm{BV}}) := \max_{h \in [H]} \dim_{\mathrm{DE},2}\big(\mathscr{P}_h^{\mathsf{P}}, \Pi_h, \epsilon_{\mathrm{BV}}\big).$$

Assume items (i)*,* (iv)*, and* (v) *from Theorem 53. Replace exact completeness by the following projected conditions:*

(a) Assumption 15 holds, and the algorithm uses the projected confidence-set update $\mathfrak{F}_{\mathsf{P}}^{(k-1)}$ from (38) *with radii $\beta_{h,k}^{\mathsf{P}}$ such that the conditions of Corollary 29 hold for every $k \geqslant 2$;*

(b) there are constants $B_{W,h}^{\mathsf{P}}$ such that $\sup_{\psi \in \mathscr{P}_h^{\mathsf{P}}} \|\psi\|_\infty \leqslant B_{W,h}^{\mathsf{P}}$ for every stage h;

(c) the stagewise projection bias obeys

$$\|\mathsf{P}_h \mathcal{T}_h^\sigma f_{h+1} - \mathcal{T}_h^\sigma f_{h+1}\|_{1,\nu} \leqslant \epsilon_{\mathrm{proj}} \tag{69}$$

for every $f \in \mathcal{F}$, every stage h, and every executed occupancy measure $\nu \in \Pi_h$;

(d) on an event of probability at least $1 - \delta$, the selected projected exact residual witnesses $\psi_{h,\mathsf{P}}^k := f_h^{(k)} - \mathsf{P}_h \mathcal{T}_h^\sigma f_{h+1}^{(k)}$ satisfy

$$\sum_{j<k} \mathbb{E}_{\widetilde{\mu}_h^j}\left[\left(\psi_{h,\mathsf{P}}^k\right)^2\right] \leqslant \epsilon_{\mathrm{BV}}^2 \tag{70}$$

for every $h \in [H]$ and $k \geqslant 2$;

(e) on the same event, the projected operator-learning terms satisfy

$$\sum_{k=2}^{K}\sum_{h=1}^{H} \alpha_{h,k}^{\mathsf{P}} \leqslant R_{\mathrm{dual}}^{\mathsf{P}}(K, \delta),$$

where $\alpha_{h,k}^{\mathsf{P}}$ is any nonnegative bookkeeping array used to record separate projected operator-learning costs.

Let

$$B_W^{\mathsf{P}} := \max_{h\in[H]} B_{W,h}^{\mathsf{P}}.$$

Then with probability at least $1-\delta$*, the target-policy regret satisfies*

$$\begin{aligned}\mathrm{Reg}(K) \leqslant H + \frac{1}{1-p}\Big[&K\epsilon_{\mathrm{appr}} + HK\epsilon_{\mathrm{eq}} + HK\epsilon_{\mathrm{proj}} \\ &+ \sqrt{KH}\,\sqrt{2HK\epsilon_{\mathrm{BV}}^2 + 4H\big(B_W^{\mathsf{P}}\big)^2 d_{\mathrm{GFA}}^{\mathsf{P}}(\epsilon_{\mathrm{BV}})} \\ &+ R_{\mathrm{dual}}^{\mathsf{P}}(K,\delta) + HK\epsilon_{\mathrm{dual}} + pHK\Big].\end{aligned} \tag{71}$$

Proof. Fix the high-probability event from assumptions *(d)* and *(e)*. Assumption *(a)* states that Assumption 15 holds and that the projected confidence radii satisfy the hypotheses of Corollary 29 for every $k \geqslant 2$. Hence approximate optimism holds for the selected sequence on this event. By (70), the projected residual witnesses satisfy the premise of Corollary 18. Applying that corollary stage by stage to the executed occupancy sequence $\{\widetilde{\mu}_h^k\}_{k=2}^K$ with witness class $\mathscr{P}_h^{\mathsf{P}}$ and envelope $B_{W,h}^{\mathsf{P}}$ gives

$$\Gamma_{\mathrm{res}}^{\mathsf{P}}(K) \leqslant 2HK\epsilon_{\mathrm{BV}}^2 + 4H\big(B_W^{\mathsf{P}}\big)^2 d_{\mathrm{GFA}}^{\mathsf{P}}(\epsilon_{\mathrm{BV}}). \tag{72}$$

Assumption (69) is exactly the projection-bias budget (16). Therefore Theorem 11 applies to the executed sequence over episodes $2,\dots,K$, with $\Gamma_{\mathrm{res}}^{\mathsf{P}}(K)$ from (72). This bounds the executed regret by the bracketed term in (71) without the final pHK. Proposition 51 then transfers this executed-sequence bound to the target policy sequence, introducing the factor $1/(1-p)$ and the additive exploiter cost $pHK/(1-p)$. Finally, the first episode contributes at most H regret. Since the operator bookkeeping term $R_{\mathrm{dual}}^{\mathsf{P}}(K,\delta) + HK\epsilon_{\mathrm{dual}}$ is nonnegative, it can be added back without invalidating the bound. Combining these bounds proves (71). □

Corollary 55 (Main-text projected regret bound from projected statistical inputs)**.** *Suppose the setting of Theorem 54 holds. Define*

$$R_{\mathrm{BV,proj}}(K) := \sqrt{KH}\,\sqrt{2HK\epsilon_{\mathrm{BV}}^2 + 4H\big(B_W^{\mathsf{P}}\big)^2 d_{\mathrm{GFA}}^{\mathsf{P}}(\epsilon_{\mathrm{BV}})},$$

and let $R_{\mathrm{dual,proj}}(K,\delta)$ *be any upper bound on* $R_{\mathrm{dual}}^{\mathsf{P}}(K,\delta) + HK\epsilon_{\mathrm{dual}}$*. If one further separates a projected operator-transfer remainder bounded by* $HK\epsilon_{\mathrm{op,proj}}$*, then*

$$\begin{aligned}\mathrm{Reg}(K) \leqslant H + \frac{1}{1-p}\Big[&R_{\mathrm{BV,proj}}(K) + R_{\mathrm{dual,proj}}(K,\delta) + HK\epsilon_{\mathrm{proj}} + HK\epsilon_{\mathrm{op,proj}} \\ &+ K\epsilon_{\mathrm{appr}} + HK\epsilon_{\mathrm{eq}} + pHK\Big].\end{aligned}$$

In particular, Theorem 2 is this compressed projected form.

Proof. Theorem 54 already yields the same inequality with the projected residual term written explicitly and with $R_{\mathrm{dual}}^{\mathsf{P}}(K,\delta) + HK\epsilon_{\mathrm{dual}}$ displayed instead of $R_{\mathrm{dual,proj}}(K,\delta)$. Thus the corollary is only a bookkeeping rewrite. A separately tracked projected operator-transfer remainder may either be left explicit as $HK\epsilon_{\mathrm{op,proj}}$ or absorbed into $R_{\mathrm{dual,proj}}(K,\delta)$. □

Proposition 56 (Current scope of the pure-online BV bridge)**.** *Suppose each episode contributes exactly one stage-*h *sample, so* $n_{h,j} = 1$ *for every* $j < k$*. Then the currently proved bridge terms satisfy*

$$S_{h,k} = k-1, \qquad N_{h,<k} = k-1, \qquad n_{h,\min}^{(k)} = 1.$$

Consequently, Lemma 20 and Corollary 23 only yield

$$\xi_{h,k}(\delta;\eta_W) = B_{W,h}^2\sqrt{2(k-1)\log\big(2\mathcal{N}_W(\eta_W)/\delta\big)} + 2B_{W,h}(k-1)\eta_W, \tag{73}$$

$$\rho_{h,k}(\delta;\eta_F,\eta_G) = 2B_\phi(\sigma)(k-1)\Big(2\bar{\Delta}_{h,k}^{\mathrm{cov}}(\delta;\eta_F,\eta_G) + \epsilon_{\mathrm{dual}}\Big). \tag{74}$$

In particular, the present proof does not verify the uniform calibration condition (65) *in the pure online regime unless one can prove substantially stronger operator-transfer or backward-validation rates than the current* $1/\sqrt{N_{h,<k}}$*-type concentration bounds.*

Proof. The identities for $S_{h,k}$, $N_{h,<k}$, and $n^{(k)}_{h,\min}$ follow immediately from $n_{h,j} = 1$ for all $j < k$. Substituting these values into the definitions of $\xi_{h,k}(\delta;\eta_W)$ in Lemma 20 and $\rho_{h,k}(\delta;\eta_F,\eta_G)$ in Corollary 23 yields (73) and (74). Since both bounds grow with k under the currently available concentration rates, they do not imply the existence of a uniform-in-k constant ϵ^2_{BV} that certifies the executed-occupancy residual premise (65). This identifies the present scope of the bridge analysis, not a statement of impossibility. □

Proposition 57 (Pure-online self-normalized partial closure for linear and finite-rank kernel witnesses)**.** *Fix a stage $h \in [H]$ and suppose each episode contributes one realized stage-h tuple $X^k_h = (S^k_h, A^k_h, B^k_h)$. Assume the exact residual witness selected in episode k admits the finite-dimensional representation*

$$\psi^k_h(x) = \langle \theta^k_h, \phi_{W,h}(x)\rangle, \qquad \|\theta^k_h\|_2 \leqslant R_{W,h}, \qquad \sup_x \|\phi_{W,h}(x)\|_2 \leqslant \kappa_{W,h}, \tag{75}$$

where $\phi_{W,h} : \mathcal{X}_h \to \mathbb{R}^{d_{W,h}}$. Define the online witness Gram matrices

$$\Lambda_{W,h,1} := \lambda I_{d_{W,h}}, \qquad \Lambda_{W,h,k} := \lambda I_{d_{W,h}} + \sum_{j=1}^{k} \phi_{W,h}(X^j_h)\phi_{W,h}(X^j_h)^\top \qquad (k \geqslant 1),$$

and suppose the algorithm enforces the self-normalized validation rule

$$\|\theta^k_h\|^2_{\Lambda_{W,h,k-1}} \leqslant \beta^{\mathrm{SN}}_{h,k} \qquad \text{for every } k \geqslant 2.$$

Let $\beta^{\mathrm{SN}}_{W,h} := \sup_{2\leqslant k\leqslant K} \beta^{\mathrm{SN}}_{h,k}$, and for any $\epsilon > 0$ define the large-residual index set

$$\mathcal{L}^{\mathrm{SN}}_h(\epsilon) := \left\{k \in \{2,\dots,K\} : |\psi^k_h(X^k_h)| > \epsilon\right\}.$$

Then

$$|\mathcal{L}^{\mathrm{SN}}_h(\epsilon)| \min\left\{1, \frac{\epsilon^2}{\beta^{\mathrm{SN}}_{W,h}}\right\} \leqslant 2\log\frac{\det(\Lambda_{W,h,K})}{\det(\lambda I_{d_{W,h}})} \leqslant 2d_{W,h}\log\left(1 + \frac{\kappa^2_{W,h}K}{\lambda}\right). \tag{76}$$

In particular, whenever $\epsilon^2 \leqslant \beta^{\mathrm{SN}}_{W,h}$,

$$|\mathcal{L}^{\mathrm{SN}}_h(\epsilon)| \leqslant \frac{2\beta^{\mathrm{SN}}_{W,h}}{\epsilon^2} d_{W,h}\log\left(1 + \frac{\kappa^2_{W,h}K}{\lambda}\right). \tag{77}$$

The same statement applies to finite-rank kernel classes after lifting them into their finite feature coordinates. This is only a partial closure result: it controls the number of large realized residuals under a self-normalized rule, but it does not by itself verify the fixed occupancy-space residual budget required by Theorem 53.

Proof. Fix $k \in \mathcal{L}^{\mathrm{SN}}_h(\epsilon)$ and abbreviate $\phi_k := \phi_{W,h}(X^k_h)$. By (75) and Cauchy–Schwarz in the $\Lambda_{W,h,k-1}$ inner product,

$$\epsilon^2 < |\psi^k_h(X^k_h)|^2 = |\langle \theta^k_h, \phi_k\rangle|^2 \leqslant \|\theta^k_h\|^2_{\Lambda_{W,h,k-1}} \|\phi_k\|^2_{\Lambda^{-1}_{W,h,k-1}} \leqslant \beta^{\mathrm{SN}}_{W,h}\|\phi_k\|^2_{\Lambda^{-1}_{W,h,k-1}}.$$

Therefore

$$\min\left\{1, \frac{\epsilon^2}{\beta^{\mathrm{SN}}_{W,h}}\right\} \leqslant \min\left\{1, \|\phi_k\|^2_{\Lambda^{-1}_{W,h,k-1}}\right\}.$$

Summing over $k \in \mathcal{L}^{\mathrm{SN}}_h(\epsilon)$ gives

$$|\mathcal{L}^{\mathrm{SN}}_h(\epsilon)| \min\left\{1, \frac{\epsilon^2}{\beta^{\mathrm{SN}}_{W,h}}\right\} \leqslant \sum_{k=2}^{K} \min\left\{1, \|\phi_k\|^2_{\Lambda^{-1}_{W,h,k-1}}\right\}.$$

For every $k \geqslant 2$, the matrix-determinant identity yields

$$\det(\Lambda_{W,h,k}) = \det(\Lambda_{W,h,k-1})\left(1 + \|\phi_k\|^2_{\Lambda^{-1}_{W,h,k-1}}\right),$$

so

$$\log \frac{\det(\Lambda_{W,h,K})}{\det(\lambda I_{d_{W,h}})} = \sum_{k=2}^{K} \log\Big(1 + \|\phi_k\|^2_{\Lambda^{-1}_{W,h,k-1}}\Big).$$

Since $\log(1+u) \geqslant \frac{1}{2}\min\{1,u\}$ for every $u \geqslant 0$,

$$\sum_{k=2}^{K} \min\Big\{1, \|\phi_k\|^2_{\Lambda^{-1}_{W,h,k-1}}\Big\} \leqslant 2\log \frac{\det(\Lambda_{W,h,K})}{\det(\lambda I_{d_{W,h}})}.$$

This proves the first inequality in (76). For the second, use $\Lambda_{W,h,K} \preceq (\lambda + \kappa^2_{W,h}K) I_{d_{W,h}}$. Finally, (77) follows from (76) when $\epsilon^2 \leqslant \beta^{\mathrm{SN}}_{W,h}$. Finite-rank kernel classes are finite-dimensional linear classes in feature coordinates, so the same argument applies verbatim after lifting into that coordinate system. □

D.4 An Epochwise / Cross-Fit Linear Theorem Under Explicit Dual Specialization

We now show how to bypass Proposition 56 by changing only the validation schedule. Partition the interaction into geometrically growing epochs $\mathcal{E}_1, \dots, \mathcal{E}_M$, with $M \leqslant \lceil \log_2(2K) \rceil$ and epoch lengths $m_e := |\mathcal{E}_e|$ satisfying $m_e \asymp N_e^-$, where $N_e^- := \sum_{r<e} m_r$ is the number of pre-epoch episodes. At the beginning of epoch e, the learner refits the dual operator on pre-epoch data only and freezes the corresponding plug-in witness family throughout that epoch. The confidence update for epoch e is then computed solely from fresh within-epoch data against that frozen witness class. This is the epochwise / cross-fit construction: operator fitting uses old data, witness validation uses new data, and the adaptive witness loop responsible for Proposition 56 is cut.

We package the linear instantiation into a theorem whose inputs are no longer the abstract robust MG-BE dimension and abstract operator budget, but explicit linear representation and coverage assumptions.

Assumption 17 (TV-linear-mixture nominal model)**.** *For each stage $h \in [H]$, let $x = (s,a,b)$ denote the stagewise state-action tuple. Assume:*

(i) the nominal kernel is a linear mixture,

$$P^0_h(\cdot \mid x) = \sum_{m=1}^{d_h} \phi_{h,m}(x)\, \mu_{h,m}(\cdot),$$

for some feature map $\phi_h : \mathcal{X}_h \to \mathbb{R}^{d_h}$ and base measures $\{\mu_{h,m}\}_{m=1}^{d_h}$;

(ii) the stage reward is linear in the same feature map,

$$r_h(x) = \langle \beta^r_h, \phi_h(x) \rangle;$$

(iii) the robust uncertainty set is a stagewise TV ball of radius σ_h, and under the fail-state support-control condition the robust continuation admits the exact representation

$$\inf_{P:\, D_{\mathrm{TV}}(P, P^0_h(\cdot|x)) \leqslant \sigma_h} \mathbb{E}_{s' \sim P}\left[V(s')\right] = \mathbb{E}_{s' \sim P^0_h(\cdot|x)}\left[V(s')\right] - \sigma_h \operatorname{span}(V)$$

for every bounded continuation value $V : \mathcal{S} \to \mathbb{R}$.

Proposition 58 (TV-linear-mixture robust minimax completeness)**.** *Assume 17. Define the augmented feature map*

$$\widetilde{\phi}_h(x) := \big(1, \phi_h(x)\big) \in \mathbb{R}^{d_h+1}.$$

Then for every bounded f_{h+1}, the robust minimax Bellman backup belongs to the augmented linear class generated by $\widetilde{\phi}_h$. More precisely,

$$\big(\mathcal{T}^\sigma_h f_{h+1}\big)(x) = \Big\langle \widetilde{\theta}_h(f_{h+1}), \widetilde{\phi}_h(x) \Big\rangle, \tag{78}$$

where

$$\widetilde{\theta}_h(f_{h+1}) := \Big(-\sigma_h \operatorname{span}(V_f),\ \beta^r_h + u_h(f_{h+1})\Big)$$

and

$$u_h(f_{h+1}) := \left(\int V_f(s')\, \mu_{h,1}(ds'), \ldots, \int V_f(s')\, \mu_{h,d_h}(ds') \right).$$

If $0 \leqslant V_f \leqslant H$*, then*

$$\|\widetilde{\theta}_h(f_{h+1})\|_2 \leqslant \sigma_h H + \|\beta_h^r\|_2 + \sqrt{d_h}\, H. \tag{79}$$

Proof. Fix $x = (s, a, b)$ and abbreviate $V_f(\cdot) := V_{f,h+1}(\cdot)$. By Assumption 17,

$$\begin{aligned}\big(\mathcal{T}_h^\sigma f_{h+1}\big)(x) &= r_h(x) + \inf_{P:\, D_{\mathrm{TV}}(P, P_h^0(\cdot|x)) \leqslant \sigma_h} \mathbb{E}_{s'\sim P}\left[V_f(s')\right] \\ &= \langle \beta_h^r, \phi_h(x)\rangle + \mathbb{E}_{s'\sim P_h^0(\cdot|x)}\left[V_f(s')\right] - \sigma_h \operatorname{span}(V_f).\end{aligned}$$

Using the linear-mixture representation of P_h^0,

$$\mathbb{E}_{s'\sim P_h^0(\cdot|x)}\left[V_f(s')\right] = \sum_{m=1}^{d_h} \phi_{h,m}(x) \int V_f(s')\, \mu_{h,m}(ds') = \langle u_h(f_{h+1}), \phi_h(x)\rangle.$$

Substituting this into the previous display proves (78). Finally, if $0 \leqslant V_f \leqslant H$, then $\operatorname{span}(V_f) \leqslant H$ and each coordinate of $u_h(f_{h+1})$ has magnitude at most H, which yields (79). □

Corollary 59 (TV-linear-mixture residual inclusion)**.** *Under Assumption 17, suppose $\mathcal{F}_h$ is the augmented linear class generated by $\widetilde{\phi}_h$ with parameter radius $R_{F,h}$. Then the exact robust residual class $\mathscr{P}_h = (I - \mathcal{T}_h^\sigma)\mathcal{F}$ is contained in the same augmented linear witness class with parameter radius at most*

$$R_{W,h}^{\text{TV-mix}} := R_{F,h} + \sigma_h H + \|\beta_h^r\|_2 + \sqrt{d_h}\, H. \tag{80}$$

Proof. Every $f_h \in \mathcal{F}_h$ already has the form $f_h(x) = \langle \theta_h, \widetilde{\phi}_h(x)\rangle$ with $\|\theta_h\|_2 \leqslant R_{F,h}$. By Proposition 58, $\mathcal{T}_h^\sigma f_{h+1}$ also belongs to the same augmented linear class with parameter norm bounded by $\sigma_h H + \|\beta_h^r\|_2 + \sqrt{d_h}H$. Subtracting the two linear representations proves the inclusion and the radius bound (80). □

Assumption 18 (Linear residual closure)**.** *For every stage $h \in [H]$, the exact robust residual class $\mathscr{P}_h = (I - \mathcal{T}_h^\sigma)\mathcal{F}$ is contained in a linear witness class*

$$\left\{ x \mapsto \langle \theta, \phi_{W,h}(x)\rangle : \theta \in \mathbb{R}^{d_{W,h}},\ \|\theta\|_2 \leqslant R_{W,h} \right\},$$

with $\sup_x \|\phi_{W,h}(x)\|_2 \leqslant \kappa_{W,h}$. Moreover, the value class and dual class satisfy the linear parametrization assumptions of Corollary 36.

Proposition 58 and Corollary 59 show that this assumption is concrete for TV uncertainty sets with nominal linear-mixture kernels and augmented linear features: in that regime, both robust minimax completeness and residual-class inclusion hold with an explicit witness radius.

Assumption 19 (Linear coverage-excitation under exploiter mix)**.** *For every stage $h \in [H]$, the adaptive executed occupancies satisfy Assumption 16 with a constant C_h^{mix}. We interpret this as a linear coverage-excitation condition: the current occupancy is controlled by the pooled past mixture on which the linear dual refit is trained.*

Proposition 60 (Explicit occupancy-ratio constant from smoothing)**.** *Fix a stage h and an epoch e. Suppose every executed policy in the pooled past and current epoch uses the same smoothing design $u_h(a, b \mid s)$ with mixing weight at least $\gamma_{\min} > 0$, that is,*

$$\widetilde{\pi}_h^k(a, b \mid s) \geqslant \gamma_{\min}\, u_h(a, b \mid s) \qquad \text{for all executed episodes } k, \tag{81}$$

and let

$$\underline{u}_h := \inf_{s\in\mathcal{S},\ a\in\mathcal{A},\ b\in\mathcal{B}} u_h(a, b \mid s) > 0.$$

Let d_h^j denote the stage-h state marginal of executed episode j, and define the pooled past state marginal $\bar{d}_{h,e}^-$ by

$$\bar{d}_{h,e}^- := \sum_{r<e} \sum_{j\in\mathcal{E}_r} \lambda_{r,j}^{(e)} d_h^j, \qquad \lambda_{r,j}^{(e)} \geqslant 0, \quad \sum_{r<e} \sum_{j\in\mathcal{E}_r} \lambda_{r,j}^{(e)} = 1.$$

Assume moreover that every current stage-h state marginal d_h^k with $k \in \mathcal{E}_e$ is dominated by the pooled past state marginal $\bar{d}_{h,e}^-$ according to

$$d_h^k(s) \leqslant C_{S,h}\, \bar{d}_{h,e}^-(s) \qquad \textit{for all } s \in \mathcal{S},\ k \in \mathcal{E}_e. \tag{82}$$

Then Assumption 16 holds on epoch e with the explicit constant

$$C_h^{\mathrm{mix}} \leqslant \frac{C_{S,h}}{\gamma_{\min}\underline{u}_h}. \tag{83}$$

Proof. Let $\bar{\widetilde{\mu}}_{h,e}^-$ denote the pooled past stage-h occupancy measure averaged over all episodes before epoch e, and let $\widetilde{\mu}_h^k$ be the current occupancy for an episode $k \in \mathcal{E}_e$. For any (s,a,b),

$$\begin{aligned}
\bar{\widetilde{\mu}}_{h,e}^-(s,a,b) &= \sum_{r<e}\sum_{j\in\mathcal{E}_r} \lambda_{r,j}^{(e)} d_h^j(s)\widetilde{\pi}_h^j(a,b \mid s) \\
&\geqslant \sum_{r<e}\sum_{j\in\mathcal{E}_r} \lambda_{r,j}^{(e)} d_h^j(s)\gamma_{\min} u_h(a,b \mid s) \\
&= \gamma_{\min} u_h(a,b \mid s)\, \bar{d}_{h,e}^-(s) \\
&\geqslant \gamma_{\min}\underline{u}_h\, \bar{d}_{h,e}^-(s).
\end{aligned}$$

On the other hand,

$$\widetilde{\mu}_h^k(s,a,b) = d_h^k(s)\widetilde{\pi}_h^k(a,b \mid s) \leqslant d_h^k(s) \leqslant C_{S,h}\, \bar{d}_{h,e}^-(s)$$

by (82). Dividing the two displays yields

$$\frac{d\widetilde{\mu}_h^k}{d\bar{\widetilde{\mu}}_{h,e}^-}(s,a,b) \leqslant \frac{C_{S,h}}{\gamma_{\min}\underline{u}_h}$$

whenever $\bar{\widetilde{\mu}}_{h,e}^-(s,a,b) > 0$. This is exactly Assumption 16 with the constant in (83). □

Proposition 61 (Predictive Gram domination from feature excitation)**.** *Assume Assumption 18. Fix a stage h and an epoch e, and define the pre-epoch witness Gram matrix*

$$\Lambda_{W,h,e}^- := \lambda I_{d_{W,h}} + \sum_{r<e}\sum_{j\in\mathcal{E}_r} \phi_{W,h}(X_h^j)\phi_{W,h}(X_h^j)^\top, \qquad \lambda > 0.$$

Suppose there exists a smoothing design $u_h(a,b \mid s)$ and constants $\gamma_{\min}, \omega_h > 0$ such that every executed policy in the pooled past and current epoch satisfies (81) *and*

$$M_h(s) := \mathbb{E}_{(a,b)\sim u_h(\cdot|s)}\left[\phi_{W,h}(s,a,b)\phi_{W,h}(s,a,b)^\top\right] \succeq \omega_h I_{d_{W,h}} \qquad \textit{for all } s \in \mathcal{S}. \tag{84}$$

Then:

(i) conditioned on the stage-h state sequence of the pre-epoch episodes,

$$\mathbb{E}\Big[\Lambda_{W,h,e}^- \;\Big|\; \{s_h^j\}_{j\in\cup_{r<e}\mathcal{E}_r}\Big] \succeq \big(\lambda + \gamma_{\min}\omega_h N_e^-\big) I_{d_{W,h}}; \tag{85}$$

(ii) on any event $\Omega_{h,e}^{\mathrm{Gram}}(\nu_h) := \Big\{\Lambda_{W,h,e}^- \succeq \big(\lambda + \nu_h N_e^-\big) I_{d_{W,h}}\Big\}$ with $\nu_h > 0$, every linear witness $w_\theta(x) = \langle\theta, \phi_{W,h}(x)\rangle$ satisfies

$$\mathbb{E}_\mu\left[w_\theta^2\right] \leqslant \frac{\kappa_{W,h}^2}{\lambda + \nu_h N_e^-}\|\theta\|^2_{\Lambda_{W,h,e}^-} \tag{86}$$

for every current stage-h occupancy measure μ generated during epoch e.

Proof. Fix a pre-epoch episode $j \in \cup_{r<e}\mathcal{E}_r$. Conditioned on the stage-h state s_h^j, the executed action pair is sampled from $\widetilde{\pi}_h^j(\cdot \mid s_h^j)$, so by (81),

$$\begin{aligned}\mathbb{E}\Big[\phi_{W,h}(X_h^j)\phi_{W,h}(X_h^j)^\top \;\Big|\; s_h^j\Big] &= \mathbb{E}_{(a,b)\sim\widetilde{\pi}_h^j(\cdot|s_h^j)}\left[\phi_{W,h}(s_h^j,a,b)\phi_{W,h}(s_h^j,a,b)^\top\right] \\ &\succeq \gamma_{\min}\mathbb{E}_{(a,b)\sim u_h(\cdot|s_h^j)}\left[\phi_{W,h}(s_h^j,a,b)\phi_{W,h}(s_h^j,a,b)^\top\right] \\ &= \gamma_{\min}M_h(s_h^j) \succeq \gamma_{\min}\omega_h I_{d_{W,h}}.\end{aligned}$$

Summing over all N_e^- pre-epoch episodes and adding $\lambda I_{d_{W,h}}$ proves (85).

For part (ii), fix a current occupancy measure μ on epoch e. Because $\sup_x \|\phi_{W,h}(x)\|_2 \leqslant \kappa_{W,h}$,

$$\mathbb{E}_\mu\left[w_\theta^2\right] = \mathbb{E}_\mu\left[\langle\theta,\phi_{W,h}\rangle^2\right] \leqslant \kappa_{W,h}^2\|\theta\|_2^2.$$

On the event $\Omega_{h,e}^{\mathrm{Gram}}(\nu_h)$, $\Lambda_{W,h,e}^- \succeq \big(\lambda+\nu_h N_e^-\big)I_{d_{W,h}}$, so

$$\|\theta\|_2^2 \leqslant \frac{1}{\lambda+\nu_h N_e^-}\|\theta\|_{\Lambda_{W,h,e}^-}^2.$$

Combining the last two displays yields (86). □

Corollary 62 (Operator transfer from predictive Gram control)**.** *Assume the event $\Omega_{h,e}^{\mathrm{Gram}}(\nu_h)$ from Proposition 61. Suppose further that for a candidate continuation function $f \in \mathcal{F}_{h+1}$, the epochwise operator gap admits the witness-feature representation*

$$e_{h,e}(f;x) := \big(\widehat{\mathcal{T}}_{h,f}^{\sigma,(e)}f_{h+1} - \mathcal{T}_h^\sigma f_{h+1}\big)(x) = \langle\vartheta_{h,e}(f),\phi_{W,h}(x)\rangle$$

for some parameter $\vartheta_{h,e}(f) \in \mathbb{R}^{d_{W,h}}$. Then for every executed episode $k \in \mathcal{E}_e$,

$$\|e_{h,e}(f)\|_{1,\widetilde{\mu}_h^k} \leqslant \frac{\kappa_{W,h}}{\sqrt{\lambda+\nu_h N_e^-}}\|\vartheta_{h,e}(f)\|_{\Lambda_{W,h,e}^-}. \tag{87}$$

Proof. By Cauchy–Schwarz,

$$\|e_{h,e}(f)\|_{1,\widetilde{\mu}_h^k} \leqslant \sqrt{\mathbb{E}_{\widetilde{\mu}_h^k}\left[e_{h,e}(f)^2\right]}.$$

Applying Proposition 61(ii) to the linear witness $w_{\vartheta_{h,e}(f)} = e_{h,e}(f)$ gives

$$\mathbb{E}_{\widetilde{\mu}_h^k}\left[e_{h,e}(f)^2\right] \leqslant \frac{\kappa_{W,h}^2}{\lambda+\nu_h N_e^-}\|\vartheta_{h,e}(f)\|_{\Lambda_{W,h,e}^-}^2.$$

Taking square roots proves (87). □

The corollary isolates the quantity that must be controlled in order to remove the current state-domination-based coverage proof. The next three results do exactly this for the structured linear specialization used by the main theorem: they prove a high-probability Gram lower bound, introduce a ridge/self-normalized dual refit over the whole linear value class, and then feed that bound back through (87) so that the feature-excitation route can replace (82).

Proposition 63 (High-probability Gram lower bound under feature excitation)**.** *Under the conditions of Proposition 61, with probability at least $1-\delta$, simultaneously for all stages $h \in [H]$ and epochs $e \in [M]$,*

$$\Lambda_{W,h,e}^- \succeq \Big(\lambda + \gamma_{\min}\omega_h N_e^- - 4\kappa_{W,h}^2\sqrt{N_e^-\log\big(2d_{W,h}HM/\delta\big)}\Big)I_{d_{W,h}}. \tag{88}$$

In particular, if

$$N_e^- \geqslant \frac{64\kappa_{W,h}^4\log\big(2d_{W,h}HM/\delta\big)}{\gamma_{\min}^2\omega_h^2}, \tag{89}$$

then the event $\Omega_{h,e}^{\mathrm{Gram}}(\nu_h)$ holds with $\nu_h = \gamma_{\min}\omega_h/2$.

Proof. Fix a stage h and an epoch e. For each pre-epoch episode $j \in \cup_{r<e}\mathcal{E}_r$, define

$$Y_j := \phi_{W,h}(X_h^j)\phi_{W,h}(X_h^j)^\top, \qquad \bar{Y}_j := \mathbb{E}\Big[Y_j \mid \mathcal{H}_{j-1}, s_h^j\Big].$$

By Proposition 61, $\bar{Y}_j \succeq \gamma_{\min}\omega_h I_{d_{W,h}}$. Also, $\|Y_j\|_{\mathrm{op}} \leqslant \kappa_{W,h}^2$ and $\|\bar{Y}_j\|_{\mathrm{op}} \leqslant \kappa_{W,h}^2$, so the martingale-difference increments $Z_j := Y_j - \bar{Y}_j$ satisfy $\|Z_j\|_{\mathrm{op}} \leqslant 2\kappa_{W,h}^2$.

Matrix Azuma's inequality therefore gives

$$\Pr\left(\left\|\sum_{j\in\cup_{r<e}\mathcal{E}_r} Z_j\right\|_{\mathrm{op}} \geqslant t\right) \leqslant 2d_{W,h}\exp\left(-\frac{t^2}{8\kappa_{W,h}^4 N_e^-}\right).$$

Choosing $t = 4\kappa_{W,h}^2\sqrt{N_e^-\log(2d_{W,h}HM/\delta)}$ and union bounding over all $h \in [H]$ and $e \in [M]$ proves that, on an event of probability at least $1-\delta$,

$$\sum_{j\in\cup_{r<e}\mathcal{E}_r} Y_j \succeq \sum_{j\in\cup_{r<e}\mathcal{E}_r} \bar{Y}_j - 4\kappa_{W,h}^2\sqrt{N_e^-\log\big(2d_{W,h}HM/\delta\big)}\, I_{d_{W,h}}.$$

Since each $\bar{Y}_j \succeq \gamma_{\min}\omega_h I_{d_{W,h}}$, adding $\lambda I_{d_{W,h}}$ proves (88). If (89) holds, then the deviation term is at most $\gamma_{\min}\omega_h N_e^-/2$, which proves $\Omega_{h,e}^{\mathrm{Gram}}(\gamma_{\min}\omega_h/2)$. □

Proposition 64 (High-probability Gram lower bound under a predictable excitation budget)**.** *Fix a stage $h \in [H]$ and an epoch $e \in [M]$. For each pre-epoch episode $j \in \cup_{r<e}\mathcal{E}_r$, let $Y_j := \phi_{W,h}(X_h^j)\phi_{W,h}(X_h^j)^\top$ and suppose there exists an $\mathcal{H}_{j-1}$-measurable positive semidefinite matrix $\underline{Y}_j$ such that*

$$\mathbb{E}[Y_j \mid \mathcal{H}_{j-1}] \succeq \underline{Y}_j. \tag{90}$$

Define the predictable excitation budget

$$\Omega_{h,e}^{\mathrm{pred}} := \lambda_{\min}\left(\sum_{j\in\cup_{r<e}\mathcal{E}_r} \underline{Y}_j\right). \tag{91}$$

Then with probability at least $1-\delta$, simultaneously for all stages $h \in [H]$ and epochs $e \in [M]$,

$$\Lambda_{W,h,e}^- \succeq \Big(\lambda + \Omega_{h,e}^{\mathrm{pred}} - 4\kappa_{W,h}^2\sqrt{N_e^-\log\big(2d_{W,h}HM/\delta\big)}\Big) I_{d_{W,h}}. \tag{92}$$

Proof. Fix a stage h and an epoch e. Define $Z_j := Y_j - \mathbb{E}[Y_j \mid \mathcal{H}_{j-1}]$. Because $\|Y_j\|_{\mathrm{op}} \leqslant \kappa_{W,h}^2$ and $\|\mathbb{E}[Y_j \mid \mathcal{H}_{j-1}]\|_{\mathrm{op}} \leqslant \kappa_{W,h}^2$, we have $\|Z_j\|_{\mathrm{op}} \leqslant 2\kappa_{W,h}^2$. The same matrix Azuma argument as in Proposition 63 therefore yields, with probability at least $1-\delta$, simultaneously over all stages and epochs,

$$\left\|\sum_{j\in\cup_{r<e}\mathcal{E}_r} Z_j\right\|_{\mathrm{op}} \leqslant 4\kappa_{W,h}^2\sqrt{N_e^-\log\big(2d_{W,h}HM/\delta\big)}. \tag{93}$$

On this event,

$$\begin{aligned}\Lambda_{W,h,e}^- &= \lambda I_{d_{W,h}} + \sum_{j\in\cup_{r<e}\mathcal{E}_r} \mathbb{E}[Y_j \mid \mathcal{H}_{j-1}] + \sum_{j\in\cup_{r<e}\mathcal{E}_r} Z_j \\ &\succeq \lambda I_{d_{W,h}} + \sum_{j\in\cup_{r<e}\mathcal{E}_r} \underline{Y}_j - 4\kappa_{W,h}^2\sqrt{N_e^-\log\big(2d_{W,h}HM/\delta\big)}\, I_{d_{W,h}}\end{aligned}$$

by (90) and (93). Taking minimum eigenvalues proves (92). □

Proposition 65 (Scalar design value implies occupancy-averaged excitation)**.** *Assume the explicit mixing rule* (81)*. Fix a stage $h \in [H]$ and an epoch $e \in [M]$. For every state s, define the smoothing-design covariance*

$$M_h(s) := \mathbb{E}_{(a,b)\sim u_h(\cdot|s)}\left[\phi_{W,h}(s,a,b)\phi_{W,h}(s,a,b)^\top\right]$$

and its scalar design value

$$\omega_h(s) := \lambda_{\min}\big(M_h(s)\big). \tag{94}$$

Let s_h^j denote the stage-h state of pre-epoch episode $j \in \cup_{r<e}\mathcal{E}_r$. If there exists $\bar{\omega}_h > 0$ such that

$$\mathbb{E}\Big[\omega_h(s_h^j) \mid \mathcal{H}_{j-1}\Big] \geqslant \bar{\omega}_h \qquad \textit{for every } j \in \cup_{r<e}\mathcal{E}_r, \tag{95}$$

then the matrix excitation condition (96) *holds. Consequently, Proposition 66 applies with the same constant $\bar{\omega}_h$.*

Proof. For every state s, the minimum-eigenvalue definition gives

$$M_h(s) \succeq \omega_h(s) I_{d_{W,h}}.$$

Applying this with $s = s_h^j$ and taking conditional expectations yields

$$\mathbb{E}\Big[M_h(s_h^j) \mid \mathcal{H}_{j-1}\Big] \succeq \mathbb{E}\Big[\omega_h(s_h^j) \mid \mathcal{H}_{j-1}\Big]\, I_{d_{W,h}} \succeq \bar{\omega}_h I_{d_{W,h}},$$

which is exactly (96). The final claim is then immediate from Proposition 66. □

Proposition 66 (Occupancy-averaged excitation implies a predictable budget)**.** *Assume the explicit mixing rule* (81)*. Fix a stage $h \in [H]$ and an epoch $e \in [M]$, and let s_h^j denote the stage-h state of pre-epoch episode $j \in \cup_{r<e}\mathcal{E}_r$. Suppose the smoothing design $u_h(\cdot \mid s)$ satisfies*

$$M_h(s) := \mathbb{E}_{(a,b)\sim u_h(\cdot|s)}\left[\phi_{W,h}(s,a,b)\phi_{W,h}(s,a,b)^\top\right],$$

and that there exists $\bar{\omega}_h > 0$ such that

$$\mathbb{E}\Big[M_h(s_h^j) \mid \mathcal{H}_{j-1}\Big] \succeq \bar{\omega}_h I_{d_{W,h}} \qquad \textit{for every } j \in \cup_{r<e}\mathcal{E}_r. \tag{96}$$

Then the premise of Proposition 64 holds with $\underline{Y}_j = \gamma_{\min}\bar{\omega}_h I_{d_{W,h}}$. Consequently,

$$\Omega_{h,e}^{\mathrm{pred}} \geqslant \gamma_{\min}\bar{\omega}_h N_e^-. \tag{97}$$

Proof. Fix a pre-epoch episode $j \in \cup_{r<e}\mathcal{E}_r$. Conditioned on s_h^j, the explicit mixing rule (81) gives

$$\begin{aligned}
\mathbb{E}\Big[Y_j \mid \mathcal{H}_{j-1}, s_h^j\Big] &= \mathbb{E}_{(a,b)\sim\widetilde{\pi}_h^j(\cdot|s_h^j)}\left[\phi_{W,h}(s_h^j,a,b)\phi_{W,h}(s_h^j,a,b)^\top\right] \\
&\succeq \gamma_{\min}\mathbb{E}_{(a,b)\sim u_h(\cdot|s_h^j)}\left[\phi_{W,h}(s_h^j,a,b)\phi_{W,h}(s_h^j,a,b)^\top\right] \\
&= \gamma_{\min} M_h(s_h^j).
\end{aligned}$$

Taking conditional expectations given $\mathcal{H}_{j-1}$ and using (96) yields

$$\mathbb{E}[Y_j \mid \mathcal{H}_{j-1}] \succeq \gamma_{\min}\mathbb{E}\Big[M_h(s_h^j) \mid \mathcal{H}_{j-1}\Big] \succeq \gamma_{\min}\bar{\omega}_h I_{d_{W,h}}.$$

Thus Proposition 64 applies with $\underline{Y}_j = \gamma_{\min}\bar{\omega}_h I_{d_{W,h}}$, and summing over the N_e^- pre-epoch episodes proves (97). □

Corollary 67 (Uniform action-pair smoothing gives a concrete design score)**.** *Assume finite action spaces and let the smoothing design be uniform over action pairs,*

$$u_h^{\mathrm{unif}}(a,b \mid s) := \frac{1}{|\mathcal{A}||\mathcal{B}|}.$$

For each state s, form the witness-feature matrix

$$\Phi_h(s) := \begin{bmatrix}\phi_{W,h}(s,a_1,b_1) & \cdots & \phi_{W,h}(s,a_{|\mathcal{A}|},b_{|\mathcal{B}|})\end{bmatrix} \in \mathbb{R}^{d_{W,h}\times|\mathcal{A}||\mathcal{B}|},$$

whose columns enumerate all action pairs. Then

$$M_h(s) = \frac{1}{|\mathcal{A}||\mathcal{B}|}\Phi_h(s)\Phi_h(s)^\top, \qquad \omega_h(s) = \frac{1}{|\mathcal{A}||\mathcal{B}|}\lambda_{\min}\big(\Phi_h(s)\Phi_h(s)^\top\big). \tag{98}$$

Hence, if there exists $\bar{\omega}_h > 0$ such that

$$\mathbb{E}\Big[\lambda_{\min}\big(\Phi_h(s_h^j)\Phi_h(s_h^j)^\top\big) \mid \mathcal{H}_{j-1}\Big] \geqslant |\mathcal{A}||\mathcal{B}|\,\bar{\omega}_h \tag{99}$$

for every pre-epoch episode j, then the scalar design condition (95) *holds and therefore Proposition 66 applies.*

Proof. The identity (98) follows directly from the definition of the uniform distribution over action pairs. Substituting (98) into (95) shows that (99) implies the premise of Proposition 65. The conclusion then follows immediately. □

Proposition 68 (Good-state reachability implies a design-value lower bound)**.** *Fix a stage $h \in [H]$ and an epoch $e \in [M]$. Let $\mathcal{G}_h \subseteq \mathcal{S}$ be a measurable good-state set and suppose the scalar design value from* (94) *satisfies*

$$\omega_h(s) \geqslant \omega_{h,\mathrm{good}} \qquad \textit{for every } s \in \mathcal{G}_h \tag{100}$$

for some constant $\omega_{h,\mathrm{good}} > 0$. If the stage-$h$ state of every pre-epoch episode obeys

$$\Pr\Big(s_h^j \in \mathcal{G}_h \mid \mathcal{H}_{j-1}\Big) \geqslant \rho_h \qquad \textit{for every } j \in \cup_{r<e}\mathcal{E}_r \tag{101}$$

for some $\rho_h > 0$, then the scalar design condition (95) *holds with*

$$\bar{\omega}_h = \rho_h\, \omega_{h,\mathrm{good}}. \tag{102}$$

Consequently, Proposition 65 and Proposition 66 apply with that value of $\bar{\omega}_h$.

Proof. By (100),

$$\omega_h(s_h^j) \geqslant \omega_{h,\mathrm{good}}\, \mathbf{1}\{s_h^j \in \mathcal{G}_h\}.$$

Taking conditional expectations given $\mathcal{H}_{j-1}$ and using (101) yields

$$\mathbb{E}\Big[\omega_h(s_h^j) \mid \mathcal{H}_{j-1}\Big] \geqslant \omega_{h,\mathrm{good}} \Pr\Big(s_h^j \in \mathcal{G}_h \mid \mathcal{H}_{j-1}\Big) \geqslant \rho_h\, \omega_{h,\mathrm{good}},$$

which is exactly (95) with $\bar{\omega}_h$ given by (102). The remaining claim follows from Proposition 65 and then Proposition 66. □

Corollary 69 (Episode-level explorer branch implies reachability excitation)**.** *Fix a stage $h \in [H]$ and an epoch $e \in [M]$. Assume the setting of Proposition 68, and suppose the executed policy-selection rule is augmented by an episode-level explorer branch: before every pre-epoch episode $j \in \cup_{r<e}\mathcal{E}_r$, the algorithm draws an $\mathcal{H}_{j-1}$-measurable Bernoulli variable E_j with*

$$\Pr(E_j = 1 \mid \mathcal{H}_{j-1}) \geqslant q_{\mathrm{exp},h}, \qquad q_{\mathrm{exp},h} > 0, \tag{103}$$

and whenever $E_j = 1$ the episode is rolled in with an explorer pair whose stage-h state marginal $d_{h,\mathrm{exp}}^j$ satisfies

$$d_{h,\mathrm{exp}}^j(\mathcal{G}_h) \geqslant \rho_h. \tag{104}$$

Then (101) *holds with ρ_h replaced by $q_{\mathrm{exp},h}\rho_h$, and therefore*

$$\bar{\omega}_h = q_{\mathrm{exp},h}\rho_h\, \omega_{h,\mathrm{good}} \tag{105}$$

is a valid scalar design-value lower bound.

Proof. Condition on $\mathcal{H}_{j-1}$. By the law of total expectation and the nonnegativity of conditional probabilities,

$$\begin{aligned}\Pr(s_h^j \in \mathcal{G}_h \mid \mathcal{H}_{j-1}) &= \Pr(E_j = 1 \mid \mathcal{H}_{j-1}) \Pr(s_h^j \in \mathcal{G}_h \mid \mathcal{H}_{j-1}, E_j = 1) \\ &\quad + \Pr(E_j = 0 \mid \mathcal{H}_{j-1}) \Pr(s_h^j \in \mathcal{G}_h \mid \mathcal{H}_{j-1}, E_j = 0) \\ &\geqslant \Pr(E_j = 1 \mid \mathcal{H}_{j-1}) \Pr(s_h^j \in \mathcal{G}_h \mid \mathcal{H}_{j-1}, E_j = 1).\end{aligned}$$

Under the explorer branch, $\Pr(s_h^j \in \mathcal{G}_h \mid \mathcal{H}_{j-1}, E_j = 1) = d_{h,\mathrm{exp}}^j(\mathcal{G}_h) \geqslant \rho_h$ by (104), while (103) gives $\Pr(E_j = 1 \mid \mathcal{H}_{j-1}) \geqslant q_{\mathrm{exp},h}$. Hence

$$\Pr(s_h^j \in \mathcal{G}_h \mid \mathcal{H}_{j-1}) \geqslant q_{\mathrm{exp},h}\rho_h.$$

This is exactly (101) with ρ_h replaced by $q_{\mathrm{exp},h}\rho_h$. Applying Proposition 68 proves (105). □

Proposition 70 (Existing exploiter-mix implies reachability from BR component dynamics)**.** *Fix a stage $h \in [H]$ and an epoch $e \in [M]$. Assume the executed pair in every pre-epoch episode $j \in \cup_{r<e}\mathcal{E}_r$ is drawn according to the existing episode-level exploiter-mix rule used in Proposition 51: with conditional probabilities $1-p$, $p/2$, and $p/2$, the episode executes one of (π^j, ν^j), (BR_1^j, ν^j), and (π^j, BR_2^j), respectively. Let $d^j_{h,\mathrm{tar}}, d^j_{h,\mathrm{br1}}, d^j_{h,\mathrm{br2}}$ denote the corresponding stage-h state marginals under the interaction process. Suppose there exists a measurable good-state set $\mathcal{G}_h \subseteq \mathcal{S}$ such that, for every pre-epoch episode j,*

$$d^j_{h,\mathrm{tar}}(\mathcal{G}_h) \geqslant \rho_{h,\mathrm{tar}}, \qquad d^j_{h,\mathrm{br1}}(\mathcal{G}_h) \geqslant \rho_{h,\mathrm{br1}}, \qquad d^j_{h,\mathrm{br2}}(\mathcal{G}_h) \geqslant \rho_{h,\mathrm{br2}}, \tag{106}$$

with deterministic nonnegative constants $\rho_{h,\mathrm{tar}}, \rho_{h,\mathrm{br1}}, \rho_{h,\mathrm{br2}}$. Then the stage-$h$ state of the actually executed episode satisfies

$$\Pr\Big(s_h^j \in \mathcal{G}_h \mid \mathcal{H}_{j-1}\Big) \geqslant (1-p)\rho_{h,\mathrm{tar}} + \frac{p}{2}\rho_{h,\mathrm{br1}} + \frac{p}{2}\rho_{h,\mathrm{br2}}. \tag{107}$$

If in addition $\omega_h(s) \geqslant \omega_{h,\mathrm{good}}$ for all $s \in \mathcal{G}_h$, then Proposition 68 applies with

$$\bar{\omega}_h = \Big((1-p)\rho_{h,\mathrm{tar}} + \frac{p}{2}\rho_{h,\mathrm{br1}} + \frac{p}{2}\rho_{h,\mathrm{br2}}\Big)\omega_{h,\mathrm{good}}.$$

Proof. Condition on $\mathcal{H}_{j-1}$. Let $B_j \in \{\mathrm{tar}, \mathrm{br1}, \mathrm{br2}\}$ denote the branch selected by the exploiter-mix rule in episode j. By the law of total probability,

$$\begin{aligned}\Pr(s_h^j \in \mathcal{G}_h \mid \mathcal{H}_{j-1}) &= \sum_{b\in\{\mathrm{tar},\mathrm{br1},\mathrm{br2}\}} \Pr(B_j = b \mid \mathcal{H}_{j-1}) \Pr(s_h^j \in \mathcal{G}_h \mid \mathcal{H}_{j-1}, B_j = b) \\ &= (1-p)\, d^j_{h,\mathrm{tar}}(\mathcal{G}_h) + \frac{p}{2}\, d^j_{h,\mathrm{br1}}(\mathcal{G}_h) + \frac{p}{2}\, d^j_{h,\mathrm{br2}}(\mathcal{G}_h).\end{aligned}$$

Using (106) yields (107). The design-value claim then follows from Proposition 68. □

Corollary 71 (BR-only reachability lower bound)**.** *Under the assumptions of Proposition 70, if no lower bound is assumed for the target branch but both BR branches satisfy $\rho_{h,\mathrm{br1}}, \rho_{h,\mathrm{br2}} \geqslant \rho_{h,\mathrm{BR}}$ for some $\rho_{h,\mathrm{BR}} > 0$, then*

$$\Pr\Big(s_h^j \in \mathcal{G}_h \mid \mathcal{H}_{j-1}\Big) \geqslant p\,\rho_{h,\mathrm{BR}}$$

and therefore

$$\bar{\omega}_h \geqslant p\,\rho_{h,\mathrm{BR}}\,\omega_{h,\mathrm{good}}.$$

Proof. Set $\rho_{h,\mathrm{tar}} = 0$ in Proposition 70 and use $\rho_{h,\mathrm{br1}}, \rho_{h,\mathrm{br2}} \geqslant \rho_{h,\mathrm{BR}}$. □

Primitive BR-mix coverage from reachability and local nondegeneracy. Assume finite action spaces and the explicit smoothing rule (81) with the uniform action-pair design u_h^{unif} from Corollary 67. Fix a stage $h \in [H]$ and an epoch $e \in [M]$. Suppose there exists a measurable good-state set $\mathcal{G}_h \subseteq \mathcal{S}$, a stagewise local nondegeneracy constant $s_h > 0$, and branch-mass constants $\rho_{h,\mathrm{tar}}, \rho_{h,\mathrm{br1}}, \rho_{h,\mathrm{br2}} \geqslant 0$ such that:

(i) for every $s \in \mathcal{G}_h$,

$$\lambda_{\min}\big(\Phi_h(s)\Phi_h(s)^\top\big) \geqslant s_h^2, \tag{108}$$

where $\Phi_h(s)$ is the witness-feature matrix from (98);

(ii) the current BR-based exploiter-mix obeys the branch good-state lower bounds (106).

Define

$$\alpha_h := (1-p)\rho_{h,\mathrm{tar}} + \frac{p}{2}\rho_{h,\mathrm{br1}} + \frac{p}{2}\rho_{h,\mathrm{br2}}. \tag{109}$$

Then the predictable excitation budget satisfies

$$\Omega^{\mathrm{pred}}_{h,e} \geqslant \gamma_{\min}\,\frac{\alpha_h s_h^2}{|\mathcal{A}||\mathcal{B}|}\,N_e^-. \tag{110}$$

Consequently, the predictable-growth premise (132) of Corollary 79 holds with

$$\underline{\nu}_h = \gamma_{\min}\,\frac{\alpha_h s_h^2}{|\mathcal{A}||\mathcal{B}|}. \tag{111}$$

In the BR-only case of Corollary 71, α_h can be replaced by $p\rho_{h,\mathrm{BR}}$. This follows because (108) and (98) imply $\omega_h(s) \geqslant s_h^2/(|\mathcal{A}||\mathcal{B}|)$ for every $s \in \mathcal{G}_h$. Proposition 70 and Proposition 68 then yield $\bar{\omega}_h = \alpha_h s_h^2/(|\mathcal{A}||\mathcal{B}|)$, and Proposition 66 converts this scalar design-value lower bound into (110) and (111).

Assumption 20 (Uniform ridge-linear dual specialization over the linear value class)**.** *For each stage $h \in [H]$, consider the full linear value class $\mathcal{F}_{h+1}$ equipped with the metric $d_{\mathcal{F},h+1}$ from Assumption 11. For every epoch e, stage h, and continuation function $f_{h+1} \in \mathcal{F}_{h+1}$, there exists a parameter $\vartheta_h^\star(f_{h+1}) \in \mathbb{R}^{d_{W,h}}$ with $\|\vartheta_h^\star(f_{h+1})\|_2 \leqslant S_{G,h}$ and scalar pseudo-labels $\{Y_{h,j}(f_{h+1})\}_{j\in\cup_{u<e}\mathcal{E}_u}$ such that*

$$Y_{h,j}(f_{h+1}) = \left\langle \vartheta_h^\star(f_{h+1}), \phi_{W,h}(X_h^j) \right\rangle + \eta_{h,j}(f_{h+1}), \tag{112}$$

where $\eta_{h,j}(f_{h+1})$ is conditionally mean zero and $R_{G,h}$-sub-Gaussian given $\mathcal{H}_{j-1}$. Assume moreover that the pseudo-label and target-parameter maps are uniformly Lipschitz:

$$\left|Y_{h,j}(f_{h+1}) - Y_{h,j}(\tilde{f}_{h+1})\right| \leqslant L_{Y,h}\, d_{\mathcal{F},h+1}(f_{h+1}, \tilde{f}_{h+1}), \tag{113}$$

$$\|\vartheta_h^\star(f_{h+1}) - \vartheta_h^\star(\tilde{f}_{h+1})\|_2 \leqslant L_{\star,h}\, d_{\mathcal{F},h+1}(f_{h+1}, \tilde{f}_{h+1}) \tag{114}$$

for all $f_{h+1}, \tilde{f}_{h+1} \in \mathcal{F}_{h+1}$. The epochwise dual refit is the ridge estimator

$$\widehat{\vartheta}_{h,e}(f_{h+1}) \in \operatorname*{argmin}_{\vartheta\in\mathbb{R}^{d_{W,h}}} \left\{ \sum_{j\in\cup_{u<e}\mathcal{E}_u} \Big(Y_{h,j}(f_{h+1}) - \langle \vartheta, \phi_{W,h}(X_h^j)\rangle\Big)^2 + \lambda\|\vartheta\|_2^2 \right\}, \tag{115}$$

and the induced operator gap satisfies

$$\big(\widehat{\mathcal{T}}_{h,f_{h+1}}^{\sigma,(e)} f_{h+1} - \mathcal{T}_h^\sigma f_{h+1}\big)(x) = \left\langle \widehat{\vartheta}_{h,e}(f_{h+1}) - \vartheta_h^\star(f_{h+1}), \phi_{W,h}(x) \right\rangle. \tag{116}$$

Proposition 72 (Affine continuation coupling removes separate Lipschitz inputs)**.** *Fix a stage $h \in [H]$ and suppose the continuation class admits the parameterization*

$$\mathcal{F}_{h+1} = \left\{ f_{h+1,\alpha} : x \mapsto \langle \alpha, \phi_{F,h+1}(x)\rangle : \alpha \in \mathbb{R}^{d_{F,h}},\ \|\alpha\|_2 \leqslant R_{F,h} \right\}.$$

Assume that for every pre-epoch sample index $j \in \cup_{u<e}\mathcal{E}_u$, there exist a scalar $y_{h,j}^0$, a vector $a_{h,j} \in \mathbb{R}^{d_{F,h}}$, a parameter $\vartheta_{h,0}^\star \in \mathbb{R}^{d_{W,h}}$, and a matrix $B_h \in \mathbb{R}^{d_{W,h}\times d_{F,h}}$ such that

$$Y_{h,j}(f_{h+1,\alpha}) = y_{h,j}^0 + \langle a_{h,j}, \alpha\rangle + \eta_{h,j}(\alpha), \tag{117}$$

$$\vartheta_h^\star(f_{h+1,\alpha}) = \vartheta_{h,0}^\star + B_h\alpha, \tag{118}$$

where $\eta_{h,j}(\alpha)$ is conditionally mean zero and $R_{G,h}$-sub-Gaussian given $\mathcal{H}_{j-1}$ for every α. If

$$\sup_{j\in\cup_{u<e}\mathcal{E}_u} \|a_{h,j}\|_2 \leqslant A_{Y,h}, \qquad \|B_h\|_{\mathrm{op}} \leqslant A_{\star,h},$$

then Assumption 20 holds with the parameter metric $d_{\mathcal{F},h+1}(f_{h+1,\alpha}, f_{h+1,\tilde{\alpha}}) := \|\alpha - \tilde{\alpha}\|_2$, the radius choice $S_{G,h} = \|\vartheta_{h,0}^\star\|_2 + A_{\star,h}R_{F,h}$, and the derived continuity constants $L_{Y,h} = A_{Y,h}$ and $L_{\star,h} = A_{\star,h}$.

Proof. Fix any $\alpha, \tilde{\alpha} \in \mathbb{R}^{d_{F,h}}$ with $\|\alpha\|_2, \|\tilde{\alpha}\|_2 \leqslant R_{F,h}$. By (118),

$$\|\vartheta_h^\star(f_{h+1,\alpha})\|_2 \leqslant \|\vartheta_{h,0}^\star\|_2 + \|B_h\|_{\mathrm{op}}\|\alpha\|_2 \leqslant \|\vartheta_{h,0}^\star\|_2 + A_{\star,h}R_{F,h} = S_{G,h}.$$

Also, (117) gives

$$\left|Y_{h,j}(f_{h+1,\alpha}) - Y_{h,j}(f_{h+1,\tilde{\alpha}})\right| = \left|\langle a_{h,j}, \alpha - \tilde{\alpha}\rangle\right| \leqslant A_{Y,h}\|\alpha - \tilde{\alpha}\|_2,$$

while (118) yields

$$\|\vartheta_h^\star(f_{h+1,\alpha}) - \vartheta_h^\star(f_{h+1,\tilde{\alpha}})\|_2 = \|B_h(\alpha - \tilde{\alpha})\|_2 \leqslant A_{\star,h}\|\alpha - \tilde{\alpha}\|_2.$$

These are exactly (113) and (114) under the parameter metric. The remaining parts of Assumption 20 are already built into (117)–(118). □

Proposition 73 (Uniform self-normalized ridge control over the value class)**.** *Assume Assumption 20. Then with probability at least $1-\delta$, simultaneously for all stages $h \in [H]$, epochs $e \in [M]$, and continuation functions $f_{h+1} \in \mathcal{F}_{h+1}$,*

$$\left\|\widehat{\vartheta}_{h,e}(f_{h+1}) - \vartheta_h^\star(f_{h+1})\right\|_{\Lambda^-_{W,h,e}} \leqslant \beta^{\mathrm{ridge}}_{h,e}(\delta), \tag{119}$$

where

$$\begin{aligned}\beta^{\mathrm{ridge}}_{h,e}(\delta) \;&:= R_{G,h}\sqrt{2d_{W,h}\log\left(1+\frac{\kappa^2_{W,h}N^-_e}{\lambda}\right)+2\log\left(\frac{2HM\mathcal{N}_F(\eta^{\mathrm{ridge}}_{F,e})}{\delta}\right)} \\ &\quad + \sqrt{\lambda}\,S_{G,h} + L_{Y,h} + (\sqrt{\lambda}+\kappa_{W,h})L_{\star,h},\end{aligned} \tag{120}$$

and $\eta^{\mathrm{ridge}}_{F,e} := (N^-_e \vee 1)^{-1/2}$.

Proof. Fix a stage h, an epoch e, and let $\mathcal{F}^{\sharp,e}_{h+1} \subseteq \mathcal{F}_{h+1}$ be an $\eta^{\mathrm{ridge}}_{F,e}$-net under $d_{\mathcal{F},h+1}$ with cardinality at most $\mathcal{N}_F(\eta^{\mathrm{ridge}}_{F,e})$. Fix a net point $\bar{f}_{h+1} \in \mathcal{F}^{\sharp,e}_{h+1}$. By Assumption 20, the pre-epoch samples satisfy the standard linear regression model (112) with design vectors $\phi_{W,h}(X^j_h)$, parameter $\vartheta^\star_h(\bar{f}_{h+1})$, and conditionally $R_{G,h}$-sub-Gaussian noise. Applying the standard self-normalized bound for ridge regression to the estimator (115) yields, with probability at least $1-\delta'$,

$$\left\|\widehat{\vartheta}_{h,e}(\bar{f}_{h+1}) - \vartheta^\star_h(\bar{f}_{h+1})\right\|_{\Lambda^-_{W,h,e}} \leqslant R_{G,h}\sqrt{2\log\left(\frac{\det(\Lambda^-_{W,h,e})^{1/2}}{\delta'\det(\lambda I_{d_{W,h}})^{1/2}}\right)} + \sqrt{\lambda}\,S_{G,h}.$$

Because $\Lambda^-_{W,h,e} \preceq (\lambda + \kappa^2_{W,h}N^-_e)I_{d_{W,h}}$, we have

$$\log\frac{\det(\Lambda^-_{W,h,e})^{1/2}}{\det(\lambda I_{d_{W,h}})^{1/2}} \leqslant \frac{d_{W,h}}{2}\log\left(1+\frac{\kappa^2_{W,h}N^-_e}{\lambda}\right).$$

Choosing $\delta' = \delta/(HM\mathcal{N}_F(\eta^{\mathrm{ridge}}_{F,e}))$ and union bounding over all stages, epochs, and net points gives the stated bound simultaneously for every $\bar{f}_{h+1} \in \mathcal{F}^{\sharp,e}_{h+1}$.

Now fix an arbitrary $f_{h+1} \in \mathcal{F}_{h+1}$ and choose $\bar{f}_{h+1} \in \mathcal{F}^{\sharp,e}_{h+1}$ such that $d_{\mathcal{F},h+1}(f_{h+1}, \bar{f}_{h+1}) \leqslant \eta^{\mathrm{ridge}}_{F,e}$. By the ridge normal equations,

$$\widehat{\vartheta}_{h,e}(f_{h+1}) = \left(\Lambda^-_{W,h,e}\right)^{-1}\sum_{j\in\cup_{u<e}\mathcal{E}_u}\phi_{W,h}(X^j_h)\,Y_{h,j}(f_{h+1}),$$

so

$$\widehat{\vartheta}_{h,e}(f_{h+1}) - \widehat{\vartheta}_{h,e}(\bar{f}_{h+1}) = \left(\Lambda^-_{W,h,e}\right)^{-1}\sum_{j\in\cup_{u<e}\mathcal{E}_u}\phi_{W,h}(X^j_h)\Big(Y_{h,j}(f_{h+1}) - Y_{h,j}(\bar{f}_{h+1})\Big).$$

Let $\mathbf{\Phi}^-_{W,h,e}$ be the design matrix whose rows are $\phi_{W,h}(X^j_h)^\top$. Then

$$\left\|\widehat{\vartheta}_{h,e}(f_{h+1}) - \widehat{\vartheta}_{h,e}(\bar{f}_{h+1})\right\|^2_{\Lambda^-_{W,h,e}} = \Delta y^\top \mathbf{\Phi}^-_{W,h,e}\left(\Lambda^-_{W,h,e}\right)^{-1}\left(\mathbf{\Phi}^-_{W,h,e}\right)^\top \Delta y \leqslant \|\Delta y\|_2^2,$$

where $\Delta y_j := Y_{h,j}(f_{h+1}) - Y_{h,j}(\bar{f}_{h+1})$, and the last step uses $\mathbf{\Phi}^-_{W,h,e}\left(\Lambda^-_{W,h,e}\right)^{-1}\left(\mathbf{\Phi}^-_{W,h,e}\right)^\top \preceq I$. By (113),

$$\|\Delta y\|_2 \leqslant L_{Y,h}\sqrt{N^-_e}\,\eta^{\mathrm{ridge}}_{F,e} \leqslant L_{Y,h}.$$

Also, by (114),

$$\begin{aligned}\left\|\vartheta^\star_h(f_{h+1}) - \vartheta^\star_h(\bar{f}_{h+1})\right\|_{\Lambda^-_{W,h,e}} &\leqslant \left\|\vartheta^\star_h(f_{h+1}) - \vartheta^\star_h(\bar{f}_{h+1})\right\|_2\sqrt{\|\Lambda^-_{W,h,e}\|_{\mathrm{op}}} \\ &\leqslant L_{\star,h}\eta^{\mathrm{ridge}}_{F,e}\sqrt{\lambda+\kappa^2_{W,h}N^-_e} \\ &\leqslant (\sqrt{\lambda}+\kappa_{W,h})L_{\star,h}.\end{aligned}$$

Combining these two displays with the net-point self-normalized bound and the triangle inequality proves (119) with (120). □

Corollary 74 (Ridge operator transfer under feature excitation)**.** *Assume Assumption 20 and the warm-start condition* (89)*. Then with probability at least* $1-\delta$*, simultaneously for all stages* $h \in [H]$*, epochs* $e \in [M]$*, continuation functions* $f_{h+1} \in \mathcal{F}_{h+1}$*, and executed episodes* $k \in \mathcal{E}_e$*,*

$$\left\| \widehat{\mathcal{T}}^{\sigma,(e)}_{h,f_{h+1}} f_{h+1} - \mathcal{T}^{\sigma}_{h} f_{h+1} \right\|_{1,\widetilde{\mu}^k_h} \leqslant \bar{\rho}^{\mathrm{ridge}}_{h,e}(\delta), \tag{121}$$

where

$$\bar{\rho}^{\mathrm{ridge}}_{h,e}(\delta) := \frac{\kappa_{W,h}}{\sqrt{\lambda + \nu_h N_e^-}}\, \beta^{\mathrm{ridge}}_{h,e}(\delta), \qquad \nu_h := \gamma_{\min}\omega_h/2.$$

Proof. Under (89), Proposition 63 gives the Gram event $\Omega^{\mathrm{Gram}}_{h,e}(\nu_h)$ with $\nu_h = \gamma_{\min}\omega_h/2$. Assumption 20 identifies the operator-gap parameter in (116), and Proposition 73 controls that parameter in the pre-epoch Gram norm. Applying Corollary 62 with $\vartheta_{h,e}(f_{h+1}) = \widehat{\vartheta}_{h,e}(f_{h+1}) - \vartheta^\star_h(f_{h+1})$ therefore yields (121). □

Corollary 75 (Warm-start-free ridge operator transfer under predictable excitation)**.** *Assume Assumption 20 and the conditions of Proposition 64. Then with probability at least* $1-\delta$*, simultaneously for all stages* $h \in [H]$*, epochs* $e \in [M]$*, continuation functions* $f_{h+1} \in \mathcal{F}_{h+1}$*, and executed episodes* $k \in \mathcal{E}_e$*,*

$$\left\| \widehat{\mathcal{T}}^{\sigma,(e)}_{h,f_{h+1}} f_{h+1} - \mathcal{T}^{\sigma}_{h} f_{h+1} \right\|_{1,\widetilde{\mu}^k_h} \leqslant \bar{\rho}^{\mathrm{pred}}_{h,e}(\delta), \tag{122}$$

where

$$\bar{\rho}^{\mathrm{pred}}_{h,e}(\delta) := \frac{\kappa_{W,h}}{\sqrt{\underline{\Lambda}^{\mathrm{pred}}_{h,e}(\delta)}}\, \beta^{\mathrm{ridge}}_{h,e}(\delta),$$
$$\underline{\Lambda}^{\mathrm{pred}}_{h,e}(\delta) := \lambda + \left[\Omega^{\mathrm{pred}}_{h,e} - 4\kappa^2_{W,h}\sqrt{N_e^- \log\!\big(2d_{W,h}HM/\delta\big)} \right]_+ .$$

Proof. Proposition 64 implies $\Lambda^-_{W,h,e} \succeq \underline{\Lambda}^{\mathrm{pred}}_{h,e}(\delta) I_{d_{W,h}}$ on the stated high-probability event. Combining this with Proposition 73 and Corollary 62 gives (122). □

Proposition 76 (Projected operator transfer with an approximation remainder)**.** *Fix a stage* $h \in [H]$*, an epoch* $e \in [M]$*, and a continuation function* $f_{h+1} \in \mathcal{F}_{h+1}$*. Let* $\mathsf{P}_h : L_\infty(\mathcal{S}\times\mathcal{A}\times\mathcal{B}) \to L_\infty(\mathcal{S}\times\mathcal{A}\times\mathcal{B})$ *be such that, for every executed episode* $k \in \mathcal{E}_e$*,*

$$\|\mathsf{P}_h u - \mathsf{P}_h v\|_{1,\widetilde{\mu}^k_h} \leqslant \|u - v\|_{1,\widetilde{\mu}^k_h} \qquad \textit{for all } u, v. \tag{123}$$

Suppose the stagewise operator gap admits the decomposition

$$\widehat{\mathcal{T}}^{\sigma,(e)}_{h,f_{h+1}} f_{h+1} - \mathcal{T}^{\sigma}_{h} f_{h+1} = \zeta_{h,e}(f_{h+1}) + r_{h,e}(f_{h+1}), \tag{124}$$

and that for every executed episode $k \in \mathcal{E}_e$*,*

$$\|\zeta_{h,e}(f_{h+1})\|_{1,\widetilde{\mu}^k_h} \leqslant \bar{\rho}_{h,e}, \qquad \|r_{h,e}(f_{h+1})\|_{1,\widetilde{\mu}^k_h} \leqslant \epsilon_{\mathrm{op,proj}}. \tag{125}$$

Then

$$\left\| \mathsf{P}_h \widehat{\mathcal{T}}^{\sigma,(e)}_{h,f_{h+1}} f_{h+1} - \mathsf{P}_h \mathcal{T}^{\sigma}_{h} f_{h+1} \right\|_{1,\widetilde{\mu}^k_h} \leqslant \bar{\rho}_{h,e} + \epsilon_{\mathrm{op,proj}} \qquad \textit{for every } k \in \mathcal{E}_e. \tag{126}$$

Proof. Fix an executed episode $k \in \mathcal{E}_e$. By (123),

$$\begin{aligned}
\left\| \mathsf{P}_h \widehat{\mathcal{T}}^{\sigma,(e)}_{h,f_{h+1}} f_{h+1} - \mathsf{P}_h \mathcal{T}^{\sigma}_{h} f_{h+1} \right\|_{1,\widetilde{\mu}^k_h} &\leqslant \left\| \widehat{\mathcal{T}}^{\sigma,(e)}_{h,f_{h+1}} f_{h+1} - \mathcal{T}^{\sigma}_{h} f_{h+1} \right\|_{1,\widetilde{\mu}^k_h} \\
&\leqslant \|\zeta_{h,e}(f_{h+1})\|_{1,\widetilde{\mu}^k_h} + \|r_{h,e}(f_{h+1})\|_{1,\widetilde{\mu}^k_h} \\
&\leqslant \bar{\rho}_{h,e} + \epsilon_{\mathrm{op,proj}},
\end{aligned}$$

where the second step uses (124) and the triangle inequality, and the last step uses (125). This proves (126). □

Proposition 77 (Epochwise / cross-fit BV calibration for linear classes)**.** *Assume support control, Assumption 14, Assumption 18, Assumption 20, and the warm-start condition* (89)*. Consider the epochwise / cross-fit variant described above, and let $\mathfrak{F}_e$ denote the confidence set certified for deployment on epoch e. Define*

$$\Psi_{\mathrm{dual}} := \max_{h\in[H]}\Big[R_{G,h}+\sqrt{\lambda}S_{G,h}+L_{Y,h}+(\sqrt{\lambda}+\kappa_{W,h})L_{\star,h}\Big], \qquad \nu_{\min} := \min_{h\in[H]}\frac{\gamma_{\min}\omega_h}{2},$$

and

$$B_{\mathrm{lin}}^{\mathrm{ridge}} := B_F + \frac{\kappa_{\mathrm{lin}}\Psi_{\mathrm{dual}}}{\sqrt{\nu_{\min}}}.$$

Then there exist stagewise radii $\beta_{h,e}^{\mathrm{lin}}$ depending only on the pre-epoch sample size N_e^-, the linear dimensions, and logarithmic confidence terms such that, with probability at least $1-\delta$, simultaneously for all epochs $e\in[M]$ and all stages $h\in[H]$,

1. *the optimism reference function is retained in $\mathfrak{F}_e$;*
2. *every function deployed on epoch e satisfies an exact residual budget of order*

$$\sum_{k\in\mathcal{E}_e}\mathbb{E}_{\mu_h^k}\left[\left(\delta_h^k\right)^2\right] \leqslant \widetilde{O}\Big(\big(B_{\mathrm{lin}}^{\mathrm{ridge}}\big)^2 d_{\mathrm{lin}}\log\big(2+\kappa_{\mathrm{lin}}^2R_{\mathrm{lin}}^2K\big)\Big); \tag{127}$$

3. *consequently Assumptions 5 and 7 hold for the deployed epochwise sequence with*

$$\Gamma_{\mathrm{res}}^{\text{ep-lin}}(K) \leqslant \widetilde{O}\Big(H\big(B_{\mathrm{lin}}^{\mathrm{ridge}}\big)^2 d_{\mathrm{lin}}\log\big(2+\kappa_{\mathrm{lin}}^2R_{\mathrm{lin}}^2K\big)\log K\Big). \tag{128}$$

Proof. Fix a stage h and an epoch e. Let $\mathcal{H}_e^-$ denote the sigma-field generated by all samples collected before epoch e. By construction of the epochwise / cross-fit scheme, the pre-epoch dual refit and the induced plug-in witness family are $\mathcal{H}_e^-$-measurable because they are computed only from pre-epoch data. For every $k\in\mathcal{E}_e$, let $\widehat{\mathrm{BV}}_{h,e,k}, \mathrm{BV}_{h,e,k}^{\mathrm{plug}}, \mathrm{BV}_{h,e,k}^{\mathrm{exact}}$ denote the epoch-local backward-validation statistics obtained by restricting the sums in (21)–(32) to the fresh prefix $\{j\in\mathcal{E}_e : j<k\}$ and by using the frozen epoch-e plug-in operator.

Conditional on $\mathcal{H}_e^-$, the only randomness in these epoch-local statistics comes from fresh within-epoch samples. The proof of Lemma 20 can therefore be replayed after restricting the sample index set to the fresh epoch block. Concretely, fix $k\in\mathcal{E}_e$ with nonempty fresh prefix $\mathcal{J}_{e,k} := \{j\in\mathcal{E}_e : j<k\}$, and let $\{w^1,\dots,w^M\}$ be an $\eta_{W,e}$-net of the frozen epoch-e witness class under $\|\cdot\|_\infty$. For a fixed net witness w^m and each $j\in\mathcal{J}_{e,k}$, define

$$Y_j^{(m,k)} := \frac{\big(w^m(X_h^j)\big)^2-\mathbb{E}_{\mu_h^j}\left[(w^m)^2\right]}{|\mathcal{J}_{e,k}|},$$

where $X_h^j=(S_h^j,A_h^j,B_h^j)$ is the realized stage-h tuple in episode j. Because the frozen witness family is $\mathcal{H}_e^-$-measurable and the stage-h sample of episode j is drawn from μ_h^j, we have $\mathbb{E}[Y_j^{(m,k)}\mid \mathcal{H}_{j-1}]=0$ and $|Y_j^{(m,k)}|\leqslant B_F^2/|\mathcal{J}_{e,k}|$. Azuma–Hoeffding therefore gives

$$\mathbb{P}\left(\left|\sum_{j\in\mathcal{J}_{e,k}}Y_j^{(m,k)}\right|>t\;\middle|\;\mathcal{H}_e^-\right)\leqslant 2\exp\left(-\frac{|\mathcal{J}_{e,k}|\,t^2}{2B_F^4}\right).$$

Union bounding over the net witnesses, and then over the at most m_e choices of $k\in\mathcal{E}_e$, shows that with the stated confidence allocation,

$$\max_{k\in\mathcal{E}_e}\max_{m\in[M]}\left|\frac{1}{|\mathcal{J}_{e,k}|}\sum_{j\in\mathcal{J}_{e,k}}\big(w^m(X_h^j)\big)^2-\frac{1}{|\mathcal{J}_{e,k}|}\sum_{j\in\mathcal{J}_{e,k}}\mathbb{E}_{\mu_h^j}\left[(w^m)^2\right]\right|\lesssim B_F^2\sqrt{\frac{\log(Mm_e/\delta)}{m_e}}.$$

For an arbitrary frozen witness w, choose a net point $\bar{w}$ with $\|w-\bar{w}\|_\infty\leqslant\eta_{W,e}$. Exactly as in Lemma 20, $\|w^2-\bar{w}^2\|_\infty\leqslant 2B_F\eta_{W,e}$, so both the empirical averages and the population averages

differ by at most $2B_F\eta_{W,e}$. Taking $\eta_{W,e} \asymp m_e^{-1/2}$ and using Corollary 36 to bound $\log M$ by $\widetilde{O}\big(d_{\mathrm{lin}} \log(2+\kappa_{\mathrm{lin}}^2 R_{\mathrm{lin}}^2 K)\big)$ yields an event $\Omega_{h,e}^{\mathrm{BV}}$ on which, simultaneously for all $k \in \mathcal{E}_e$ and $f \in \mathcal{F}$,

$$\left|\widehat{\mathrm{BV}}_{h,e,k}(f) - \mathrm{BV}_{h,e,k}^{\mathrm{plug}}(f)\right| \leqslant \bar{\xi}_{h,e}^{\mathrm{lin}}, \qquad \bar{\xi}_{h,e}^{\mathrm{lin}} = \widetilde{O}\left(B_F^2 \sqrt{\frac{d_{\mathrm{lin}} \log\big(2+\kappa_{\mathrm{lin}}^2 R_{\mathrm{lin}}^2 K\big)}{m_e}}\right).$$

Next, the pre-epoch dual fit is based on N_e^- episodes. Proposition 73 and Corollary 74 give an event $\Omega_{h,e}^{\mathrm{op}}$ on which, simultaneously for all $k \in \mathcal{E}_e$ and $f_{h+1} \in \mathcal{F}_{h+1}$,

$$\left\|\widehat{\mathcal{T}}_{h,f_{h+1}}^{\sigma,(e)} f_{h+1} - \mathcal{T}_h^{\sigma} f_{h+1}\right\|_{1,\mu_h^k} \leqslant \bar{\rho}_{h,e}^{\mathrm{lin}}, \qquad \bar{\rho}_{h,e}^{\mathrm{lin}} = \widetilde{O}\left(\frac{\kappa_{\mathrm{lin}} \Psi_{\mathrm{dual}}}{\sqrt{\nu_{\min}}} \sqrt{\frac{d_{\mathrm{lin}} \log\big(2+\kappa_{\mathrm{lin}}^2 R_{\mathrm{lin}}^2 K\big)}{N_e^-}}\right).$$

Because the epoch schedule is geometric, $m_e \asymp N_e^-$, so these two radii are of the same order up to constants and logarithms. Allocate confidence $\delta/(2HM)$ to each of the events $\Omega_{h,e}^{\mathrm{BV}}$ and $\Omega_{h,e}^{\mathrm{op}}$; the logarithmic factors in the two radii absorb this choice, and a union bound over all $h \in [H]$ and $e \in [M]$ produces a global event of probability at least $1-\delta$ on which all displays below hold simultaneously.

Choose $\beta_{h,e}^{\mathrm{lin}}$ to satisfy $\beta_{h,e}^{\mathrm{lin}} \geqslant \bar{\xi}_{h,e}^{\mathrm{lin}} + \bar{\rho}_{h,e}^{\mathrm{lin}}$. On $\Omega_{h,e}^{\mathrm{BV}} \cap \Omega_{h,e}^{\mathrm{op}}$, every deployed $f \in \mathfrak{F}_e$ satisfies the epoch-local confidence inequality

$$\widehat{\mathrm{BV}}_{h,e,k}(f) \leqslant \beta_{h,e}^{\mathrm{lin}} \qquad \text{for every } k \in \mathcal{E}_e,$$

because $\mathfrak{F}_e$ is exactly the confidence set certified for deployment on epoch e using the fresh block $\mathcal{E}_e$. Proposition 24 applied to the epoch-local statistics implies that every candidate admitted by the epoch-e confidence update, and in particular every deployed function on epoch e, satisfies

$$\mathrm{BV}_{h,e,k}^{\mathrm{exact}}(f) \leqslant 2\beta_{h,e}^{\mathrm{lin}} + 2\bar{\xi}_{h,e}^{\mathrm{lin}} + 2\bar{\rho}_{h,e}^{\mathrm{lin}} \lesssim \beta_{h,e}^{\mathrm{lin}} \qquad \text{for all } k \in \mathcal{E}_e.$$

The same epoch-local application of Proposition 26 retains the Bellman-consistent reference function, so $f^\dagger \in \mathfrak{F}_e$, which proves item 1.

The previous display is exactly the backward-validation premise (18) on the epoch-restricted deployed sequence, with a threshold of order $\beta_{h,e}^{\mathrm{lin}}$. Equivalently, for a constant $C > 0$, we may set $\epsilon_{h,e}^2 := C\beta_{h,e}^{\mathrm{lin}}$ and obtain $\sum_{j \in \mathcal{E}_e : j<k} \mathbb{E}_{\mu_h^j}\left[(\delta_h^k)^2\right] \leqslant \epsilon_{h,e}^2$ for every deployed $k \in \mathcal{E}_e$. Corollary 38 bounds the square-loss robust MG-BE dimension by $\widetilde{O}(d_{\mathrm{lin}})$, and the linear envelope bound is $B_{\mathrm{lin}}^{\mathrm{ridge}}$. Therefore Corollary 18 applied on the restricted epoch block gives

$$\sum_{k \in \mathcal{E}_e} \mathbb{E}_{\mu_h^k}\left[\left(\delta_h^k\right)^2\right] \leqslant \widetilde{O}\Big(\big(B_{\mathrm{lin}}^{\mathrm{ridge}}\big)^2 d_{\mathrm{lin}} \log\big(2+\kappa_{\mathrm{lin}}^2 R_{\mathrm{lin}}^2 K\big)\Big),$$

which is exactly (127) and proves item 2.

Summing the last display over H stages and over the $M = O(\log K)$ epochs of the geometric schedule proves (128). Finally, on $\Omega_{h,e}^{\mathrm{op}}$ we have $\widehat{\delta}_h^k = \delta_h^k - e_h^k$ with $\|e_h^k\|_{1,\mu_h^k} \leqslant \bar{\rho}_{h,e}^{\mathrm{lin}}$ for all $k \in \mathcal{E}_e$, so the plug-in residual budget has the same order by the same $(a+b)^2 \leqslant 2a^2 + 2b^2$ algebra used earlier. This is exactly Assumption 7 for the deployed epochwise sequence, proving item 3. □

Theorem 78 (Formal epochwise / cross-fit linear theorem)**.** *Assume support control, stable stage-game extraction, and the epochwise / cross-fit variant just described. Suppose further that Assumptions 14, 18, and 20 hold, and that the explicit excitation schedule satisfies* (81)*,* (84)*, and the warm-start condition* (89)*. In particular, by Proposition 58 and Corollary 59, robust minimax completeness and residual inclusion are concretely verified for TV-linear-mixture models with augmented linear features, while Proposition 63 and Corollary 74 replace the previous occupancy-ratio constant*

by a feature-excitation / self-normalized operator budget. Let

$$
\begin{aligned}
d_{\mathrm{lin}} &:= \max_{h\in[H]} \{d_{F,h}, d_{G,h}, d_{W,h}\}, \\
\kappa_{\mathrm{lin}} &:= \max_{h\in[H]} \{\kappa_{F,h}, \kappa_{G,h}, \kappa_{W,h}\}, \\
R_{\mathrm{lin}} &:= \max_{h\in[H]} \{R_{F,h}, R_{G,h}, R_{W,h}\}, \\
B_F &:= \max_{h\in[H]} B_{F,h}, \\
\Psi_{\mathrm{dual}} &:= \max_{h\in[H]} \left\{R_{G,h} + \sqrt{\lambda} S_{G,h} + L_{Y,h} + (\sqrt{\lambda} + \kappa_{W,h}) L_{\star,h}\right\}, \\
\nu_{\min} &:= \min_{h\in[H]} \frac{\gamma_{\min}\omega_h}{2}.
\end{aligned}
$$

Also define

$$
B_{\mathrm{lin}}^{\mathrm{ridge}} := B_F + \frac{\kappa_{\mathrm{lin}}\Psi_{\mathrm{dual}}}{\sqrt{\nu_{\min}}}.
$$

Then the target-policy regret sequence generated by the epochwise / cross-fit adaptive exploiter-mix algorithm satisfies

$$
\begin{aligned}
\mathrm{Reg}(K) \leqslant \frac{1}{1-p}\Bigg[& \widetilde{O}\left(HB_F\sqrt{d_{\mathrm{lin}}K\log\bigl(2+\kappa_{\mathrm{lin}}^2 R_{\mathrm{lin}}^2 K\bigr)}\right) \\
& + \widetilde{O}\left(H\frac{\kappa_{\mathrm{lin}}\Psi_{\mathrm{dual}}}{\sqrt{\nu_{\min}}}\sqrt{d_{\mathrm{lin}}K\log\bigl(2+\kappa_{\mathrm{lin}}^2 R_{\mathrm{lin}}^2 K\bigr)}\right) \\
& + K\epsilon_{\mathrm{appr}} + HK\epsilon_{\mathrm{eq}} + pHK\Bigg].
\end{aligned} \tag{129}
$$

Theorem 78 is therefore closed relative to Assumption 20; Proposition 72 supplies a primitive sufficient condition via affine continuation coupling, but TV-linear-mixture closure alone does not imply that operator-learning hypothesis.

Proof. Proposition 77 verifies both approximate optimism and the residual budget for the epochwise deployed sequence. In particular, Assumption 7 holds with $\Gamma_{\mathrm{res}}(K) = \Gamma_{\mathrm{res}}^{\text{ep-lin}}(K)$, so Theorem 10 contributes the residual term

$$
\sqrt{KH\,\Gamma_{\mathrm{res}}^{\text{ep-lin}}(K)} \leqslant \widetilde{O}\left(HB_{\mathrm{lin}}^{\mathrm{ridge}}\sqrt{d_{\mathrm{lin}}K\log\bigl(2+\kappa_{\mathrm{lin}}^2 R_{\mathrm{lin}}^2 K\bigr)}\right), \tag{130}
$$

where the extra epoching factor is hidden in $\widetilde{O}(\cdot)$.

For the operator-learning term, Corollary 74 gives a self-normalized epochwise radius of order $\kappa_{\mathrm{lin}}\Psi_{\mathrm{dual}}\sqrt{d_{\mathrm{lin}}/(\nu_{\min}N_e^-)}$ up to logarithmic factors. Summing the resulting epochwise operator budgets over a geometric schedule gives

$$
\sum_{k=1}^{K}\sum_{h=1}^{H}\alpha_{h,k} \leqslant \widetilde{O}\left(H\frac{\kappa_{\mathrm{lin}}\Psi_{\mathrm{dual}}}{\sqrt{\nu_{\min}}}\sqrt{d_{\mathrm{lin}}K\log\bigl(2+\kappa_{\mathrm{lin}}^2 R_{\mathrm{lin}}^2 K\bigr)}\right) \tag{131}
$$

because geometric epochs still satisfy the same $\sum_e m_e/\sqrt{N_e^-} \lesssim \sqrt{K}$ summability as the pure online sequence.

Applying Proposition 77, Theorem 10, and then Proposition 51 proves the executed-pair bound and finally transfers it to the target policy sequence with the $1/(1-p)$ multiplicative factor and $pHK/(1-p)$ additive exploiter cost. Substituting (130) and (131) yields (129). □

Corollary 79 (Warm-start-free epochwise linear theorem with burn-in)**.** *Assume the setting of Theorem 78, except that the explicit warm-start condition* (89) *is replaced by the following predictable excitation budget: for every stage $h \in [H]$ and epoch $e \in [M]$,*

$$
\Omega_{h,e}^{\mathrm{pred}} \geqslant \underline{\nu}_h N_e^-, \tag{132}
$$

where $\Omega_{h,e}^{\mathrm{pred}}$ is defined in (91) *and $\underline{\nu}_h > 0$ is stage dependent. Define*

$$N_{\mathrm{burn}}(\delta) := \max_{h\in[H]} \left\lceil \frac{64\kappa_{W,h}^4 \log\bigl(2d_{W,h}HM/\delta\bigr)}{\underline{\nu}_h^2} \right\rceil, \qquad \underline{\nu}_{\min} := \min_{h\in[H]} \frac{\underline{\nu}_h}{2}.$$

Let c_{ep} be the epoch-schedule constant such that the cumulative number of deployed episodes before the first epoch with $N_e^- \geqslant n$ is at most $c_{\mathrm{ep}} n$ for every $n \geqslant 1$. Then the target-policy regret satisfies the same bound as Theorem 78, with $\nu_{\min}$ replaced by $\underline{\nu}_{\min}$, plus the additive burn-in term

$$c_{\mathrm{ep}} H N_{\mathrm{burn}}(\delta). \tag{133}$$

Proof. By (132) and Proposition 64, every epoch with $N_e^- \geqslant N_{\mathrm{burn}}(\delta)$ satisfies

$$\Lambda_{W,h,e}^- \succeq \left(\lambda + \frac{\underline{\nu}_h N_e^-}{2}\right) I_{d_{W,h}}$$

simultaneously for all stages $h \in [H]$, because the deviation term in (92) is then at most $\underline{\nu}_h N_e^-/2$. Hence Corollary 75 recovers the same operator-transfer rate as Corollary 74, but with $\nu_{\min}$ replaced by $\underline{\nu}_{\min}$. From that point onward, the proof of Theorem 78 is unchanged.

Before the first warm epoch, the total number of deployed episodes is at most $c_{\mathrm{ep}} N_{\mathrm{burn}}(\delta)$ by definition of c_{ep}. Each episode incurs regret at most H because the horizon is H and rewards lie in $[0,1]$. Therefore the pre-warm contribution is bounded by (133), which yields the claimed warm-start-free result. □

Proposition 80 (Pure-online self-normalized linear calibration)**.** *Assume support control, Assumptions 14, 18, and 20. Consider the singleton-epoch schedule $\mathcal{E}_k = \{k\}$ for $k \in [K]$, so $M = K$ and $N_k^- = k-1$. For every stage $h \in [H]$ and episode $k \geqslant 2$, define*

$$\Lambda_{W,h,k-1} := \lambda I_{d_{W,h}} + \sum_{j=1}^{k-1} \phi_{W,h}(X_h^j)\phi_{W,h}(X_h^j)^\top$$

and

$$\beta_{h,k}^{\mathrm{on}}(\delta) := R_{G,h}\sqrt{2d_{W,h}\log\left(1+\frac{\kappa_{W,h}^2(k-1)}{\lambda}\right) + 2\log\left(\frac{4HK\,\mathcal{N}_F(\eta_{F,k}^{\mathrm{on}})}{\delta}\right)} + \sqrt{\lambda}\,S_{G,h} + L_{Y,h} + (\sqrt{\lambda}+\kappa_{W,h})L_{\star,h},$$

where $\eta_{F,k}^{\mathrm{on}} := ((k-1)\vee 1)^{-1/2}$. Assume the pure-online algorithm refits the ridge dual operator after every episode, defines the plug-in residual coordinates by

$$\widehat{\delta}_{h,k}(f;x) = \langle \widehat{\theta}_{h,k}(f), \phi_{W,h}(x)\rangle,$$

and deploys only functions in the self-normalized confidence set

$$\mathfrak{F}_{\mathrm{on}}^{(k-1)} := \left\{ f \in \mathcal{F} : \|\widehat{\theta}_{h,k}(f)\|_{\Lambda_{W,h,k-1}} \leqslant \beta_{h,k}^{\mathrm{on}}(\delta) \text{ for all } h \in [H] \right\}. \tag{134}$$

Suppose furthermore that for every stage $h \in [H]$ and episode $k \geqslant 2$,

$$\Omega_{h,k}^{\mathrm{pred}} := \lambda_{\min}\left(\sum_{j=1}^{k-1} \underline{Y}_j\right) \geqslant \underline{\nu}_h(k-1), \tag{135}$$

where $\underline{Y}_j$ is the predictable lower bound from Proposition 64 and $\underline{\nu}_h > 0$. Define

$$N_{\mathrm{burn}}^{\mathrm{on}}(\delta) := \max_{h\in[H]} \left\lceil \frac{64\kappa_{W,h}^4 \log\bigl(4d_{W,h}HK/\delta\bigr)}{\underline{\nu}_h^2} \right\rceil, \qquad \underline{\nu}_{\min} := \min_{h\in[H]} \frac{\underline{\nu}_h}{2}. \tag{136}$$

Then with probability at least $1-\delta$, simultaneously for all episodes $k \geqslant N_{\mathrm{burn}}^{\mathrm{on}}(\delta) + 2$, the following hold.

(i) $f^\dagger \in \mathfrak{F}_{\mathrm{on}}^{(k-1)}$, hence the optimistic choice over $\mathfrak{F}_{\mathrm{on}}^{(k-1)}$ satisfies Assumption 5.

(ii) For every stage $h \in [H]$,

$$\left\| \widehat{\mathcal{T}}^{\sigma,(k)}_{h,f^{(k)}} f^{(k)}_{h+1} - \mathcal{T}^{\sigma}_{h} f^{(k)}_{h+1} \right\|_{1,\widetilde{\mu}^{k}_{h}} \leqslant \alpha^{\mathrm{on}}_{h,k}(\delta) := \frac{\kappa_{W,h}\beta^{\mathrm{on}}_{h,k}(\delta)}{\sqrt{\lambda + \underline{\nu}_h(k-1)/2}}. \tag{137}$$

(iii) For every stage $h \in [H]$,

$$\mathbb{E}_{\widetilde{\mu}^{k}_{h}}\left[(\delta^{k}_{h})^2\right] \leqslant \frac{8\kappa^2_{W,h}\big(\beta^{\mathrm{on}}_{h,k}(\delta)\big)^2}{\underline{\nu}_h(k-1)}. \tag{138}$$

Consequently,

$$\Gamma^{\mathrm{on}}_{\mathrm{res}}(K,\delta) := \sum_{h=1}^{H} \sum_{k=N^{\mathrm{on}}_{\mathrm{burn}}(\delta)+2}^{K} \frac{8\kappa^2_{W,h}\big(\beta^{\mathrm{on}}_{h,k}(\delta)\big)^2}{\underline{\nu}_h(k-1)} \tag{139}$$

is a valid exact residual budget for the post-burn pure-online deployed sequence.

Proof. Apply Proposition 73 with the singleton schedule $\mathcal{E}_k = \{k\}$, so $M = K$ and $N^{-}_{k} = k-1$, using confidence level $\delta/2$. This yields, with probability at least $1-\delta/2$, simultaneously for all stages $h \in [H]$, episodes $k \geqslant 2$, and continuation functions $f_{h+1} \in \mathcal{F}_{h+1}$,

$$\left\| \widehat{\vartheta}_{h,k-1}(f_{h+1}) - \vartheta^{\star}_{h}(f_{h+1}) \right\|_{\Lambda_{W,h,k-1}} \leqslant \beta^{\mathrm{on}}_{h,k}(\delta), \tag{140}$$

where $\widehat{\vartheta}_{h,k-1}$ is the ridge refit on the prefix $\{1,\dots,k-1\}$. Likewise, Proposition 64 with the same singleton schedule and confidence level $\delta/2$ gives, with probability at least $1-\delta/2$,

$$\Lambda_{W,h,k-1} \succeq \Big(\lambda + \Omega^{\mathrm{pred}}_{h,k} - 4\kappa^2_{W,h}\sqrt{(k-1)\log\big(4d_{W,h}HK/\delta\big)}\Big) I_{d_{W,h}}.$$

On the intersection of these two events, if $k \geqslant N^{\mathrm{on}}_{\mathrm{burn}}(\delta)+2$, then (135) and (136) imply

$$\Lambda_{W,h,k-1} \succeq \left(\lambda + \frac{\underline{\nu}_h(k-1)}{2}\right) I_{d_{W,h}}. \tag{141}$$

We first prove (i). By Assumption 14, the reference function satisfies $\delta_h(f^{\dagger}) \equiv 0$ for every stage h. Hence, by the operator-gap representation (116),

$$\widehat{\delta}_{h,k}(f^{\dagger};x) = -\Big(\widehat{\mathcal{T}}^{\sigma,(k)}_{h,f^{\dagger}} f^{\dagger}_{h+1} - \mathcal{T}^{\sigma}_{h} f^{\dagger}_{h+1}\Big)(x) = -\left\langle \widehat{\vartheta}_{h,k-1}(f^{\dagger}_{h+1}) - \vartheta^{\star}_{h}(f^{\dagger}_{h+1}), \phi_{W,h}(x)\right\rangle.$$

Therefore the plug-in residual coordinate of $f^{\dagger}$ is exactly

$$\widehat{\theta}_{h,k}(f^{\dagger}) = -\Big(\widehat{\vartheta}_{h,k-1}(f^{\dagger}_{h+1}) - \vartheta^{\star}_{h}(f^{\dagger}_{h+1})\Big).$$

Taking $\Lambda_{W,h,k-1}$ -norms and using (140) yields

$$\|\widehat{\theta}_{h,k}(f^{\dagger})\|_{\Lambda_{W,h,k-1}} \leqslant \beta^{\mathrm{on}}_{h,k}(\delta),$$

so $f^{\dagger} \in \mathfrak{F}^{(k-1)}_{\mathrm{on}}$ by (134). Since the algorithm chooses $f^{(k)}$ optimistically over $\mathfrak{F}^{(k-1)}_{\mathrm{on}}$, the same argument as in Corollary 27 implies Assumption 5.

For (ii), Corollary 75 specialized to the singleton schedule gives

$$\left\| \widehat{\mathcal{T}}^{\sigma,(k)}_{h,f^{(k)}} f^{(k)}_{h+1} - \mathcal{T}^{\sigma}_{h} f^{(k)}_{h+1} \right\|_{1,\widetilde{\mu}^{k}_{h}} \leqslant \frac{\kappa_{W,h}\beta^{\mathrm{on}}_{h,k}(\delta)}{\sqrt{\lambda + \underline{\nu}_h(k-1)/2}},$$

which is exactly (137).

For (iii), let θ^{k}_{h} denote the exact residual parameter of the deployed function $f^{(k)}$, so by Assumption 18,

$$\delta^{k}_{h}(x) = \langle \theta^{k}_{h}, \phi_{W,h}(x)\rangle.$$

Pointwise, $\delta_h^k = \widehat{\delta}_{h,k}(f^{(k)}) + e_h^k$, and both terms are represented in the same witness feature map. Hence

$$\theta_h^k = \widehat{\theta}_{h,k}(f^{(k)}) + \widehat{\vartheta}_{h,k-1}(f_{h+1}^{(k)}) - \vartheta_h^\star(f_{h+1}^{(k)}).$$

Because $f^{(k)} \in \mathfrak{F}_{\mathrm{on}}^{(k-1)}$ and (140) holds,

$$\|\theta_h^k\|_{\Lambda_{W,h,k-1}} \leqslant \|\widehat{\theta}_{h,k}(f^{(k)})\|_{\Lambda_{W,h,k-1}} + \left\|\widehat{\vartheta}_{h,k-1}(f_{h+1}^{(k)}) - \vartheta_h^\star(f_{h+1}^{(k)})\right\|_{\Lambda_{W,h,k-1}} \leqslant 2\beta_{h,k}^{\mathrm{on}}(\delta).$$

Using $\sup_x \|\phi_{W,h}(x)\|_2 \leqslant \kappa_{W,h}$ and (141),

$$\begin{aligned}
\mathbb{E}_{\widetilde{\mu}_h^k}\left[(\delta_h^k)^2\right] = \mathbb{E}_{\widetilde{\mu}_h^k}\left[\langle\theta_h^k, \phi_{W,h}\rangle^2\right] &\leqslant \kappa_{W,h}^2 \|\theta_h^k\|_2^2 \\
&\leqslant \frac{\kappa_{W,h}^2 \|\theta_h^k\|_{\Lambda_{W,h,k-1}}^2}{\lambda + \underline{\nu}_h(k-1)/2} \leqslant \frac{4\kappa_{W,h}^2 \left(\beta_{h,k}^{\mathrm{on}}(\delta)\right)^2}{\lambda + \underline{\nu}_h(k-1)/2} \\
&\leqslant \frac{8\kappa_{W,h}^2 \left(\beta_{h,k}^{\mathrm{on}}(\delta)\right)^2}{\underline{\nu}_h(k-1)},
\end{aligned}$$

which proves (138). Summing this bound over all h and $k \geqslant N_{\mathrm{burn}}^{\mathrm{on}}(\delta) + 2$ gives (139). □

Theorem 81 (Formal pure-online self-normalized linear / finite-rank kernel theorem)**.** *Assume support control, stable stage-game extraction, and the pure-online self-normalized variant of Proposition 80. Suppose further that Assumptions 14, 18, and 20 hold, and that the predictable growth condition* (135) *is satisfied. Define*

$$\begin{aligned}
d_{\mathrm{lin}} &:= \max_{h\in[H]} \{d_{F,h}, d_{G,h}, d_{W,h}\}, \\
\kappa_{\mathrm{lin}} &:= \max_{h\in[H]} \{\kappa_{F,h}, \kappa_{G,h}, \kappa_{W,h}\}, \\
R_{\mathrm{lin}} &:= \max_{h\in[H]} \{R_{F,h}, R_{G,h}, R_{W,h}\}, \\
\Psi_{\mathrm{dual}} &:= \max_{h\in[H]} \left\{R_{G,h} + \sqrt{\lambda} S_{G,h} + L_{Y,h} + (\sqrt{\lambda} + \kappa_{W,h}) L_{\star,h}\right\}.
\end{aligned}$$

Then the target-policy regret sequence generated by the pure-online self-normalized algorithm satisfies

$$\begin{aligned}
\mathrm{Reg}(K) \leqslant\ & H\left(N_{\mathrm{burn}}^{\mathrm{on}}(\delta) + 1\right) \\
&+ \frac{1}{1-p}\left[\widetilde{O}\left(H\kappa_{\mathrm{lin}}\Psi_{\mathrm{dual}}/\sqrt{\underline{\nu}_{\min}}\sqrt{d_{\mathrm{lin}} K \log\left(2 + \kappa_{\mathrm{lin}}^2 R_{\mathrm{lin}}^2 K\right)}\right)\right. \\
&\left. + K\epsilon_{\mathrm{appr}} + HK\epsilon_{\mathrm{eq}} + pHK\right].
\end{aligned} \tag{142}$$

The same statement applies to finite-rank kernel classes after lifting them into their explicit finite feature coordinates and interpreting $d_{\mathrm{lin}}, \kappa_{\mathrm{lin}}, R_{\mathrm{lin}}$ in those coordinates.

Proof. Let $K_0 := N_{\mathrm{burn}}^{\mathrm{on}}(\delta) + 1$. The first K_0 episodes contribute at most HK_0 to regret because each episode has horizon H and rewards lie in $[0, 1]$.

For the suffix $\{K_0 + 1, \ldots, K\}$, Proposition 80 verifies approximate optimism and gives the exact residual budget $\Gamma_{\mathrm{res}}^{\mathrm{on}}(K, \delta)$ together with the operator radii $\alpha_{h,k}^{\mathrm{on}}(\delta)$. Since $\beta_{h,k}^{\mathrm{on}}(\delta) = \widetilde{O}\left(\Psi_{\mathrm{dual}}\sqrt{d_{\mathrm{lin}} \log\left(2 + \kappa_{\mathrm{lin}}^2 R_{\mathrm{lin}}^2 k\right)}\right)$, (139) implies

$$\Gamma_{\mathrm{res}}^{\mathrm{on}}(K, \delta) \leqslant \widetilde{O}\left(H\frac{\kappa_{\mathrm{lin}}^2 \Psi_{\mathrm{dual}}^2}{\underline{\nu}_{\min}} d_{\mathrm{lin}} \log\left(2 + \kappa_{\mathrm{lin}}^2 R_{\mathrm{lin}}^2 K\right)\right),$$

where the extra logarithmic factor coming from $\sum_{k>K_0} 1/(k-1)$ is absorbed by $\widetilde{O}(\cdot)$. Therefore the residual contribution over the suffix is

$$\sqrt{KH\,\Gamma_{\mathrm{res}}^{\mathrm{on}}(K, \delta)} \leqslant \widetilde{O}\left(H\frac{\kappa_{\mathrm{lin}}\Psi_{\mathrm{dual}}}{\sqrt{\underline{\nu}_{\min}}}\sqrt{d_{\mathrm{lin}} K \log\left(2 + \kappa_{\mathrm{lin}}^2 R_{\mathrm{lin}}^2 K\right)}\right). \tag{143}$$

Likewise, using (137) and the same bound on $\beta^{\mathrm{on}}_{h,k}(\delta)$,

$$\sum_{k=K_0+1}^{K}\sum_{h=1}^{H}\alpha^{\mathrm{on}}_{h,k}(\delta) \leqslant \widetilde{O}\left(H\frac{\kappa_{\mathrm{lin}}\Psi_{\mathrm{dual}}}{\sqrt{\underline{\nu}_{\min}}}\sqrt{d_{\mathrm{lin}}}\sum_{k=K_0+1}^{K}\frac{1}{\sqrt{k-1}}\right) \tag{144}$$

$$\leqslant \widetilde{O}\left(H\frac{\kappa_{\mathrm{lin}}\Psi_{\mathrm{dual}}}{\sqrt{\underline{\nu}_{\min}}}\sqrt{d_{\mathrm{lin}}K\log\bigl(2+\kappa^2_{\mathrm{lin}}R^2_{\mathrm{lin}}K\bigr)}\right). \tag{145}$$

Applying the proof of Theorem 10 to the post-burn executed sequence, and then Proposition 51 to transfer back to the target-policy sequence, gives (142) after combining (143) and (145). The finite-rank kernel claim is immediate because an explicit finite-rank kernel lift is a finite-dimensional linear class in its feature coordinates, so the same argument applies verbatim after replacing the linear features by the lifted kernel features. □

Theorem 82 (Formal structured projected theorem under finite-dimensional lifts)**.** *Assume the setting of Theorem 54. Suppose in addition that for every stage $h \in [H]$, the continuation class, the dual class, and the projected residual class admit finite-dimensional feature lifts:*

$$\begin{aligned}
\mathcal{F}_{h+1} &\subseteq \bigl\{x \mapsto \langle \alpha, \varphi_{F,h+1}(x)\rangle : \alpha \in \mathbb{R}^{d_{F,h}},\ \|\alpha\|_2 \leqslant R_{F,h}\bigr\},\\
\mathcal{G}_h &\subseteq \bigl\{x \mapsto \Gamma_h(x)\vartheta : \vartheta \in \mathbb{R}^{d_{G,h}},\ \|\vartheta\|_2 \leqslant R_{G,h}\bigr\},\\
\mathscr{P}^{\mathsf{P}}_h &\subseteq \bigl\{x \mapsto \langle \theta, \varphi_{W,h}(x)\rangle : \theta \in \mathbb{R}^{d_{W,h}},\ \|\theta\|_2 \leqslant R^{\mathsf{P}}_{W,h}\bigr\},
\end{aligned}$$

with feature-norm bounds $\sup_x\|\varphi_{F,h+1}(x)\|_2 \leqslant \kappa_{F,h}, \sup_x\|\Gamma_h(x)\|_{\mathrm{op}} \leqslant \kappa_{G,h}, \sup_x\|\varphi_{W,h}(x)\|_2 \leqslant \kappa_{W,h}$, and projected residual envelopes $\sup_{\psi\in\mathscr{P}^{\mathsf{P}}_h}\|\psi\|_\infty \leqslant B^{\mathsf{P}}_{W,h}$. Assume moreover that:

- *(i) the coefficient-space continuation model satisfies the affine coupling of Proposition 72, with derived constants $A^{\mathsf{P}}_{Y,h}$ and $A^{\mathsf{P}}_{\star,h}$;*
- *(ii) the explicit mixing rule* (81) *holds and the witness feature map satisfies the occupancy-averaged excitation condition* (96) *with constant $\bar{\omega}_h$;*
- *(iii) each P_h is $L_1(\widetilde{\mu}^k_h)$-nonexpansive in the sense of* (123)*;*
- *(iv) the stagewise projection bias obeys* (69)*;*
- *(v) the projected operator gap satisfies the decomposition* (124)–(125) *with remainder level $\epsilon_{\mathrm{op,proj}}$;*
- *(vi) the projected backward-validation calibration assumption* (70) *holds for the epochwise / cross-fit sequence.*

Define

$$\begin{aligned}
d_{\mathrm{proj}} &:= \max_{h\in[H]}\{d_{F,h}, d_{G,h}, d_{W,h}\},\\
\kappa_{\mathrm{proj}} &:= \max_{h\in[H]}\{\kappa_{F,h}, \kappa_{G,h}, \kappa_{W,h}\},\\
R_{\mathrm{proj}} &:= \max_{h\in[H]}\bigl\{R_{F,h}, R_{G,h}, R^{\mathsf{P}}_{W,h}\bigr\},\\
B^{\mathsf{P}}_{W,\mathrm{proj}} &:= \max_{h\in[H]} B^{\mathsf{P}}_{W,h},\\
\Psi^{\mathsf{P}}_{\mathrm{dual}} &:= \max_{h\in[H]}\Bigl\{R_{G,h} + \sqrt{\lambda}\bigl(\|\vartheta^\star_{h,0}\|_2 + A^{\mathsf{P}}_{\star,h}R_{F,h}\bigr) + A^{\mathsf{P}}_{Y,h} + (\sqrt{\lambda}+\kappa_{W,h})A^{\mathsf{P}}_{\star,h}\Bigr\},\\
\underline{\nu}_{\mathrm{proj}} &:= \min_{h\in[H]}\frac{\gamma_{\min}\bar{\omega}_h}{2}.
\end{aligned}$$

Let

$$N^{\mathsf{P}}_{\mathrm{burn}}(\delta) := \max_{h\in[H]}\left\lceil\frac{64\kappa^4_{W,h}\log\bigl(2d_{W,h}HM/\delta\bigr)}{(\gamma_{\min}\bar{\omega}_h)^2}\right\rceil.$$

Then the target-policy regret satisfies

$$\begin{aligned}\mathrm{Reg}(K) \leqslant H + c_{\mathrm{ep}} H N^{\mathsf{P}}_{\mathrm{burn}}(\delta) \\ + \frac{1}{1-p}\widetilde{O}\left(H\left(B^{\mathsf{P}}_{W,\mathrm{proj}} + \frac{\kappa_{\mathrm{proj}}\Psi^{\mathsf{P}}_{\mathrm{dual}}}{\sqrt{\underline{\nu}_{\mathrm{proj}}}}\right)\sqrt{d_{\mathrm{proj}} K \log\bigl(2+\kappa^2_{\mathrm{proj}} R^2_{\mathrm{proj}} K\bigr)}\right) \\ + \frac{K\epsilon_{\mathrm{appr}} + HK\epsilon_{\mathrm{eq}} + HK\epsilon_{\mathrm{proj}} + HK\epsilon_{\mathrm{op,proj}} + HK\epsilon_{\mathrm{dual}} + pHK}{1-p}.\end{aligned} \tag{146}$$

Proof. Proposition 37 applied to the projected witness feature lift gives

$$d^{\mathsf{P}}_{\mathrm{GFA}}(\epsilon_{\mathrm{BV}}) \leqslant 4 \max_{h\in[H]}\left[d_{W,h}\log_2\left(2 + \frac{\kappa^2_{W,h}(R^{\mathsf{P}}_{W,h})^2}{\epsilon^2_{\mathrm{BV}}}\right)\right]. \tag{147}$$

Proposition 72 turns $L_{Y,h}, L_{\star,h}$ into the coefficient-space operator norms $A^{\mathsf{P}}_{Y,h}, A^{\mathsf{P}}_{\star,h}$. Proposition 66 yields the predictable budget $\Omega^{\mathrm{pred}}_{h,e} \geqslant \gamma_{\min}\bar{\omega}_h N^-_e$, so after the burn-in threshold $N^{\mathsf{P}}_{\mathrm{burn}}(\delta)$ Corollary 75 gives a stagewise self-normalized operator radius of order

$$\widetilde{O}\left(\frac{\kappa_{\mathrm{proj}}\Psi^{\mathsf{P}}_{\mathrm{dual}}}{\sqrt{\underline{\nu}_{\mathrm{proj}}}}\sqrt{\frac{d_{\mathrm{proj}}\log(2+\kappa^2_{\mathrm{proj}}R^2_{\mathrm{proj}}K)}{N^-_e}}\right).$$

Proposition 76 then adds the approximation remainder $\epsilon_{\mathrm{op,proj}}$, while (69) contributes the projection-bias term ϵ_{proj}. Applying Theorem 54 with the dimension bound (147), the operator budget above, and the burn-in argument from Corollary 79 proves (146). □

Theorem 83 (Formal RKHS spectral-truncation projected theorem)**.** *Assume the setting of Theorem 54. For every stage $h \in [H]$, let $\mathcal{H}_{F,h+1}, \mathcal{H}_{G,h}, \mathcal{H}_{W,h}$ be separable RKHSs on the corresponding stagewise domains, with Mercer decompositions*

$$\begin{aligned}k_{F,h+1}(x,x') &= \sum_{\ell\geqslant 1}\tau_{F,h+1,\ell}\, e_{F,h+1,\ell}(x)e_{F,h+1,\ell}(x'),\\ k_{G,h}(x,x') &= \sum_{\ell\geqslant 1}\tau_{G,h,\ell}\, e_{G,h,\ell}(x)e_{G,h,\ell}(x'),\\ k_{W,h}(x,x') &= \sum_{\ell\geqslant 1}\tau_{W,h,\ell}\, e_{W,h,\ell}(x)e_{W,h,\ell}(x'),\end{aligned}$$

relative to reference measures $\rho_{F,h+1}, \rho_{G,h}, \rho_{W,h}$, with nonincreasing eigenvalues. For an integer truncation level $m \geqslant 1$, define the rank-m Mercer feature lifts

$$\begin{aligned}\varphi^{(m)}_{F,h+1}(x) &:= \bigl(\sqrt{\tau_{F,h+1,1}}\, e_{F,h+1,1}(x), \dots, \sqrt{\tau_{F,h+1,m}}\, e_{F,h+1,m}(x)\bigr),\\ \varphi^{(m)}_{G,h}(x) &:= \bigl(\sqrt{\tau_{G,h,1}}\, e_{G,h,1}(x), \dots, \sqrt{\tau_{G,h,m}}\, e_{G,h,m}(x)\bigr),\\ \varphi^{(m)}_{W,h}(x) &:= \bigl(\sqrt{\tau_{W,h,1}}\, e_{W,h,1}(x), \dots, \sqrt{\tau_{W,h,m}}\, e_{W,h,m}(x)\bigr),\end{aligned}$$

and let $\Pi^{(m)}_{F,h+1}, \Pi^{(m)}_{G,h}, \Pi^{(m)}_{W,h}$ be the corresponding orthogonal truncation operators onto the spans of the first m coordinates. Assume:

(i) $\mathcal{F}_{h+1} \subseteq \{f \in \mathcal{H}_{F,h+1} : \|f\|_{\mathcal{H}_{F,h+1}} \leqslant R_{F,h}\}$, $\mathcal{G}_h \subseteq \{g \in \mathcal{H}_{G,h} : \|g\|_{\mathcal{H}_{G,h}} \leqslant R_{G,h}\}$, and $\mathscr{P}^{\mathsf{P}}_h \subseteq \{\psi \in \mathcal{H}_{W,h} : \|\psi\|_{\mathcal{H}_{W,h}} \leqslant R^{\mathsf{P}}_{W,h}\}$;

(ii) the projected optimistic dual robust Bellman algorithm is run on the truncated classes

$$\mathcal{F}^{(m)}_{h+1} := \Pi^{(m)}_{F,h+1}\mathcal{F}_{h+1}, \qquad \mathcal{G}^{(m)}_h := \Pi^{(m)}_{G,h}\mathcal{G}_h, \qquad \mathscr{P}^{\mathsf{P},(m)}_h := \Pi^{(m)}_{W,h}\mathscr{P}^{\mathsf{P}}_h;$$

(iii) after this truncation, the six structured hypotheses of Theorem 82 hold with truncation-dependent quantities $A^{\mathsf{P},(m)}_{Y,h}, A^{\mathsf{P},(m)}_{\star,h}, B^{\mathsf{P},(m)}_{W,h}, \bar{\omega}^{(m)}_h, \epsilon_{\mathrm{proj}}(m), \epsilon_{\mathrm{op,proj}}(m)$, and with feature envelopes

$$\kappa^{(m)}_{F,h} := \sup_x \|\varphi^{(m)}_{F,h+1}(x)\|_2, \qquad \kappa^{(m)}_{G,h} := \sup_x \|\varphi^{(m)}_{G,h}(x)\|_2, \qquad \kappa^{(m)}_{W,h} := \sup_x \|\varphi^{(m)}_{W,h}(x)\|_2.$$

Define

$$
\begin{aligned}
\kappa_{\mathrm{RKHS}}^{(m)} &:= \max_{h\in[H]} \left\{\kappa_{F,h}^{(m)}, \kappa_{G,h}^{(m)}, \kappa_{W,h}^{(m)}\right\},\\
R_{\mathrm{RKHS}}^{(m)} &:= \max_{h\in[H]} \left\{R_{F,h}, R_{G,h}, R_{W,h}^{\mathsf{P}}\right\},\\
B_{W,\mathrm{RKHS}}^{\mathsf{P},(m)} &:= \max_{h\in[H]} B_{W,h}^{\mathsf{P},(m)},\\
\Psi_{\mathrm{dual}}^{\mathsf{P},(m)} &:= \max_{h\in[H]} \left\{R_{G,h} + \sqrt{\lambda}\big(\|\vartheta_{h,0}^{\star}\|_2 + A_{\star,h}^{\mathsf{P},(m)} R_{F,h}\big) + A_{Y,h}^{\mathsf{P},(m)} + (\sqrt{\lambda} + \kappa_{W,h}^{(m)}) A_{\star,h}^{\mathsf{P},(m)}\right\},\\
\underline{\nu}_{\mathrm{RKHS}}^{(m)} &:= \min_{h\in[H]} \frac{\gamma_{\min}\bar{\omega}_h^{(m)}}{2},\\
N_{\mathrm{burn}}^{\mathrm{RKHS}} &:= \max_{h\in[H]} \\
&\qquad \left\lceil \frac{64(\kappa_{W,h}^{(m)})^4}{(\gamma_{\min}\bar{\omega}_h^{(m)})^2} \log\big(2mHM/\delta\big)\right\rceil.
\end{aligned}
$$

Then the target-policy regret of the rank-m projected sequence satisfies

$$
\begin{aligned}
\mathrm{Reg}(K) \leqslant\ & H + c_{\mathrm{ep}} H N_{\mathrm{burn}}^{\mathrm{RKHS}}\\
&+ \frac{1}{1-p}\widetilde{O}\Big(H\Big(B_{W,\mathrm{RKHS}}^{\mathsf{P},(m)} + \kappa_{\mathrm{RKHS}}^{(m)} \Psi_{\mathrm{dual}}^{\mathsf{P},(m)} / \sqrt{\underline{\nu}_{\mathrm{RKHS}}^{(m)}}\Big)\\
&\quad \cdot \sqrt{mK \log\big(2 + (\kappa_{\mathrm{RKHS}}^{(m)})^2 (R_{\mathrm{RKHS}}^{(m)})^2 K\big)}\Big)\\
&+ \frac{K\epsilon_{\mathrm{appr}} + HK\epsilon_{\mathrm{eq}} + HK\epsilon_{\mathrm{proj}}(m) + HK\epsilon_{\mathrm{op,proj}}(m) + HK\epsilon_{\mathrm{dual}} + pHK}{1-p}.
\end{aligned}
\tag{148}
$$

Proof. For every stage h, the truncated classes $\mathcal{F}_{h+1}^{(m)}, \mathcal{G}_h^{(m)}, \mathscr{P}_h^{\mathsf{P},(m)}$ are finite-dimensional linear lifts of dimension at most m, with feature maps $\varphi_{F,h+1}^{(m)}, \varphi_{G,h}^{(m)}, \varphi_{W,h}^{(m)}$. Hence Theorem 82 applies with $d_{F,h} = d_{G,h} = d_{W,h} = m$, feature envelopes $\kappa_{F,h}^{(m)}, \kappa_{G,h}^{(m)}, \kappa_{W,h}^{(m)}$, and the truncation-dependent structural constants from the statement. Substituting these identifications into (146) yields (148). □

Eigenvalue-decay RKHS projected consequence. Assume the setting of Theorem 83. Suppose further that there exist constants $C_{F,h}, C_{G,h}, C_{W,h} > 0$ and exponents $r_{F,h}, r_{G,h}, r_{W,h} > 0$ such that

$$
\begin{aligned}
\tau_{F,h,\ell} &\leqslant C_{F,h}\ell^{-1-2r_{F,h}},\\
\tau_{G,h,\ell} &\leqslant C_{G,h}\ell^{-1-2r_{G,h}},\\
\tau_{W,h,\ell} &\leqslant C_{W,h}\ell^{-1-2r_{W,h}} \qquad \text{for all } \ell \geqslant 1.
\end{aligned}
\tag{149}
$$

Let

$$
r_{\min} := \min_{h\in[H]} \{r_{F,h}, r_{G,h}, r_{W,h}\}.
$$

Assume in addition that there exist constants $\bar{\kappa}_{\mathrm{RKHS}}, \bar{\Psi}_{\mathrm{dual}}, \underline{\nu}_{\mathrm{RKHS}}, \bar{B}^{\mathsf{P}}_{W,\mathrm{RKHS}}, C_{\mathrm{tail}} > 0$ such that for every truncation level $m \geqslant 1$,

$$\kappa^{(m)}_{\mathrm{RKHS}} \leqslant \bar{\kappa}_{\mathrm{RKHS}}, \qquad \Psi^{\mathsf{P},(m)}_{\mathrm{dual}} \leqslant \bar{\Psi}_{\mathrm{dual}}, \tag{150}$$

$$\underline{\nu}^{(m)}_{\mathrm{RKHS}} \geqslant \underline{\nu}_{\mathrm{RKHS}}, \qquad B^{\mathsf{P},(m)}_{W,\mathrm{RKHS}} \leqslant \bar{B}^{\mathsf{P}}_{W,\mathrm{RKHS}}, \tag{151}$$

$$\epsilon_{\mathrm{proj}}(m) + \epsilon_{\mathrm{op,proj}}(m) \leqslant C_{\mathrm{tail}} \max_{h\in[H]} \tag{152}$$

$$\left(R_{F,h}\sqrt{\sum_{\ell>m} \tau_{F,h,\ell}} + R_{G,h}\sqrt{\sum_{\ell>m} \tau_{G,h,\ell}} \right. \tag{153}$$

$$\left. + R^{\mathsf{P}}_{W,h}\sqrt{\sum_{\ell>m} \tau_{W,h,\ell}} \right). \tag{154}$$

Then for every $m \geqslant 1$,

$$\begin{aligned} \mathrm{Reg}(K) \leqslant\ & H + c_{\mathrm{ep}} H N^{\mathrm{RKHS}}_{\mathrm{burn}} \\ & + \frac{1}{1-p}\widetilde{O}\left(H\left(\bar{B}^{\mathsf{P}}_{W,\mathrm{RKHS}} + \frac{\bar{\kappa}_{\mathrm{RKHS}}\bar{\Psi}_{\mathrm{dual}}}{\sqrt{\underline{\nu}_{\mathrm{RKHS}}}} \right)\sqrt{mK} \right. \\ & \qquad \left. \times \sqrt{\log\!\big(2 + \bar{\kappa}^2_{\mathrm{RKHS}}(R^{(m)}_{\mathrm{RKHS}})^2 K\big)} \right) \\ & + \frac{K\epsilon_{\mathrm{appr}} + HK\epsilon_{\mathrm{eq}} + HKC_{\mathrm{spec}}\, m^{-r_{\min}} + HK\epsilon_{\mathrm{dual}} + pHK}{1-p}, \end{aligned} \tag{155}$$

where

$$\begin{aligned} C_{\mathrm{spec}} := C_{\mathrm{tail}} \max_{h\in[H]} \Bigg(& R_{F,h}\sqrt{\frac{C_{F,h}}{2r_{F,h}}} + R_{G,h}\sqrt{\frac{C_{G,h}}{2r_{G,h}}} \\ & + R^{\mathsf{P}}_{W,h}\sqrt{\frac{C_{W,h}}{2r_{W,h}}} \Bigg). \end{aligned}$$

By the integral test, for every $r > 0$ and $m \geqslant 1$,

$$\sum_{\ell>m} \ell^{-1-2r} \leqslant \int_m^\infty x^{-1-2r}\, dx = \frac{m^{-2r}}{2r}.$$

Applying this to (149) gives

$$\sqrt{\sum_{\ell>m} \tau_{F,h,\ell}} \leqslant \sqrt{\frac{C_{F,h}}{2r_{F,h}}}\, m^{-r_{F,h}},$$

$$\sqrt{\sum_{\ell>m} \tau_{G,h,\ell}} \leqslant \sqrt{\frac{C_{G,h}}{2r_{G,h}}}\, m^{-r_{G,h}},$$

$$\sqrt{\sum_{\ell>m} \tau_{W,h,\ell}} \leqslant \sqrt{\frac{C_{W,h}}{2r_{W,h}}}\, m^{-r_{W,h}}.$$

Since each exponent is at least $r_{\min}$, (154) implies

$$\epsilon_{\mathrm{proj}}(m) + \epsilon_{\mathrm{op,proj}}(m) \leqslant C_{\mathrm{spec}}\, m^{-r_{\min}}.$$

Substituting this and the uniform bounds (150) into (148) proves (155).

Finite-rank kernel projected extension with burn-in. Assume the setting of Theorem 82, and suppose the finite-dimensional lifts there are induced by explicit finite-rank kernel feature maps of

ranks $m_{F,h}, m_{G,h}, m_{W,h}$. Let

$$\begin{aligned}
m_{\mathrm{ker}} &:= \max_{h\in[H]} \left\{m_{F,h}, m_{G,h}, m_{W,h}\right\},\\
\kappa_{\mathrm{ker}} &:= \max_{h\in[H]} \left\{\kappa_{F,h}, \kappa_{G,h}, \kappa_{W,h}\right\},\\
R_{\mathrm{ker}} &:= \max_{h\in[H]} \left\{R_{F,h}, R_{G,h}, R^{\mathsf{P}}_{W,h}\right\},\\
B^{\mathsf{P}}_{W,\mathrm{ker}} &:= \max_{h\in[H]} B^{\mathsf{P}}_{W,h},\\
\Psi^{\mathrm{ker}}_{\mathrm{dual}} &:= \max_{h\in[H]} \left\{R_{G,h} + \sqrt{\lambda}\big(\|\vartheta^{\star}_{h,0}\|_2 + A^{\mathsf{P}}_{\star,h}R_{F,h}\big) + A^{\mathsf{P}}_{Y,h} + (\sqrt{\lambda}+\kappa_{W,h})A^{\mathsf{P}}_{\star,h}\right\},\\
\underline{\nu}_{\mathrm{ker}} &:= \min_{h\in[H]} \frac{\gamma_{\min}\bar{\omega}_h}{2}.
\end{aligned}$$

Then

$$\begin{aligned}
\mathrm{Reg}(K) \leqslant\ & H + c_{\mathrm{ep}} H N^{\mathsf{P}}_{\mathrm{burn}}(\delta)\\
&+ \frac{1}{1-p}\widetilde{O}\left(H\left(B^{\mathsf{P}}_{W,\mathrm{ker}} + \frac{\kappa_{\mathrm{ker}}\Psi^{\mathrm{ker}}_{\mathrm{dual}}}{\sqrt{\underline{\nu}_{\mathrm{ker}}}}\right)\sqrt{m_{\mathrm{ker}}K\log\big(2+\kappa^2_{\mathrm{ker}}R^2_{\mathrm{ker}}K\big)}\right)\\
&+ \frac{K\epsilon_{\mathrm{appr}} + HK\epsilon_{\mathrm{eq}} + HK\epsilon_{\mathrm{proj}} + HK\epsilon_{\mathrm{op,proj}} + HK\epsilon_{\mathrm{dual}} + pHK}{1-p}.
\end{aligned} \tag{156}$$

Finite-rank kernel classes are finite-dimensional linear classes in feature coordinates, so Theorem 82 applies with $d_{F,h} = m_{F,h}$, $d_{G,h} = m_{G,h}$, and $d_{W,h} = m_{W,h}$. Substituting these identifications into (146) proves (156).

Proposition 84 (Compactly clipped ϕ-dual losses are bounded and Lipschitz)**.** *Fix a stage h and consider the scalar dual integrand*

$$\mathfrak{l}_{\phi,h}(V,\eta,\nu) := \eta\sigma - \nu + \eta\phi^{\star}\left(\frac{\nu - V}{\eta}\right), \qquad (V,\eta,\nu)\in[0,H]\times[\underline{\eta}_h,\overline{\eta}_h]\times[-\overline{\nu}_h,\overline{\nu}_h],$$

with $0 < \underline{\eta}_h \leqslant \overline{\eta}_h$. Define

$$M_h := H + \overline{\nu}_h, \qquad I_h := \left[-\frac{M_h}{\underline{\eta}_h}, \frac{M_h}{\underline{\eta}_h}\right].$$

Assume $\phi^{\star}$ is continuously differentiable on I_h, and let

$$B_{\phi^{\star},h} := \sup_{u\in I_h}|\phi^{\star}(u)|, \quad L_{\phi^{\star},h} := \sup_{u\in I_h}|(\phi^{\star})'(u)|, \quad M_{\phi^{\star},h} := \sup_{u\in I_h}|u(\phi^{\star})'(u)|.$$

Then:

(i) for every $(V,\eta,\nu)\in[0,H]\times[\underline{\eta}_h,\overline{\eta}_h]\times[-\overline{\nu}_h,\overline{\nu}_h]$,

$$\big|\mathfrak{l}_{\phi,h}(V,\eta,\nu)\big| \leqslant \overline{\eta}_h\sigma + \overline{\nu}_h + \overline{\eta}_h B_{\phi^{\star},h} =: B^{\mathrm{clip}}_{\phi,h}; \tag{157}$$

(ii) for any two triples (V,η,ν) and $(\widetilde{V},\widetilde{\eta},\widetilde{\nu})$ in the same domain,

$$\begin{aligned}
\big|\mathfrak{l}_{\phi,h}(V,\eta,\nu) - \mathfrak{l}_{\phi,h}(\widetilde{V},\widetilde{\eta},\widetilde{\nu})\big| \leqslant\ & L_{\phi^{\star},h}|V-\widetilde{V}| + \big(\sigma + B_{\phi^{\star},h} + M_{\phi^{\star},h}\big)|\eta-\widetilde{\eta}|\\
&+ \big(1 + L_{\phi^{\star},h}\big)|\nu-\widetilde{\nu}|.
\end{aligned} \tag{158}$$

Consequently, any stagewise dual class whose (η,ν) outputs are clipped into those intervals satisfies the boundedness and pointwise Lipschitz requirements of Assumptions 10 and 11 with explicit geometry-dependent constants. In particular, the proposition applies to KL and χ^2 ambiguity sets because their conjugates are continuously differentiable on every compact interval.

Proof. For every admissible (V,η,ν), the quantity $z := (\nu - V)/\eta$ satisfies $|z| \leqslant M_h/\underline{\eta}_h$, so $z\in I_h$. This implies

$$\big|\mathfrak{l}_{\phi,h}(V,\eta,\nu)\big| \leqslant \eta\sigma + |\nu| + \eta|\phi^{\star}(z)| \leqslant \overline{\eta}_h\sigma + \overline{\nu}_h + \overline{\eta}_h B_{\phi^{\star},h},$$

which proves (157).

To prove (158), define $F(V,\eta,\nu) := \mathfrak{l}_{\phi,h}(V,\eta,\nu)$ on the compact convex set $[0,H]\times[\underline{\eta}_h,\overline{\eta}_h]\times[-\overline{\nu}_h,\overline{\nu}_h]$. By the chain rule, for every point in this set,

$$\begin{aligned}
\partial_V F(V,\eta,\nu) &= -(\phi^\star)'\left(\frac{\nu-V}{\eta}\right),\\
\partial_\nu F(V,\eta,\nu) &= -1+(\phi^\star)'\left(\frac{\nu-V}{\eta}\right),\\
\partial_\eta F(V,\eta,\nu) &= \sigma+\phi^\star\left(\frac{\nu-V}{\eta}\right)-\frac{\nu-V}{\eta}(\phi^\star)'\left(\frac{\nu-V}{\eta}\right).
\end{aligned}$$

Hence

$$|\partial_V F|\leqslant L_{\phi^\star,h},\qquad |\partial_\nu F|\leqslant 1+L_{\phi^\star,h},\qquad |\partial_\eta F|\leqslant \sigma+B_{\phi^\star,h}+M_{\phi^\star,h}$$

uniformly on the whole domain. Applying the mean-value theorem coordinatewise along the segment joining (V,η,ν) and $(\widetilde{V},\widetilde{\eta},\widetilde{\nu})$ gives (158). The final claim follows immediately from the definitions of Assumptions 10 and 11. For KL, $\phi^\star(u)=e^u-1$; for χ^2, the conjugate is continuously differentiable and piecewise quadratic/affine; in both cases the required suprema over I_h are finite. □

KL/χ^2 projected finite-rank kernel consequence. Assume the setting of Theorem 82. Suppose further that each stagewise ambiguity set is either a KL ball or a χ^2 ball, and that the continuation, dual, and projected residual classes are represented by finite-rank kernel feature maps

$$\begin{aligned}
\mathcal{F}_{h+1} &\subseteq \{x\mapsto\langle\alpha,\varphi_{F,h+1}(x)\rangle : \alpha\in\mathbb{R}^{m_{F,h}},\ \|\alpha\|_2\leqslant R_{F,h}\},\\
\mathcal{G}_h &\subseteq \left\{x\mapsto\Big(\langle\vartheta_\eta,\varphi_{\eta,h}(x)\rangle,\langle\vartheta_\nu,\varphi_{\nu,h}(x)\rangle\Big) : \vartheta_\eta\in\mathbb{R}^{m_{\eta,h}},\vartheta_\nu\in\mathbb{R}^{m_{\nu,h}}\right\},\\
\mathscr{P}_h^{\mathsf{P}} &\subseteq \{x\mapsto\langle\theta,\varphi_{W,h}(x)\rangle : \theta\in\mathbb{R}^{m_{W,h}},\ \|\theta\|_2\leqslant R_{W,h}^{\mathsf{P}}\},
\end{aligned}$$

with pointwise clipping

$$\langle\vartheta_\eta,\varphi_{\eta,h}(x)\rangle\in[\underline{\eta}_h,\overline{\eta}_h],\qquad \langle\vartheta_\nu,\varphi_{\nu,h}(x)\rangle\in[-\overline{\nu}_h,\overline{\nu}_h] \tag{159}$$

for every admissible x, and with all remaining assumptions of Theorem 82 unchanged. Define

$$m_{\phi,\mathrm{ker}} := \max_{h\in[H]}\{m_{F,h},m_{\eta,h}+m_{\nu,h},m_{W,h}\}.$$

Then the regret bound (146) holds with d_{proj} replaced by $m_{\phi,\mathrm{ker}}$, and the explicit projected-bias and projected-operator terms ϵ_{proj} and $\epsilon_{\mathrm{op,proj}}$ unchanged. The clipping condition (159) and Proposition 84 show that the KL / χ^2 dual losses satisfy the boundedness and pointwise Lipschitz requirements used by the projected general theorem. The stacked dual feature map

$$\Gamma_h(x) := \begin{bmatrix}\varphi_{\eta,h}(x)^\top & 0\\ 0 & \varphi_{\nu,h}(x)^\top\end{bmatrix}$$

has feature dimension $m_{\eta,h}+m_{\nu,h}$, so Theorem 82 applies with $d_{G,h}=m_{\eta,h}+m_{\nu,h}$, $d_{F,h}=m_{F,h}$, and $d_{W,h}=m_{W,h}$. Substituting these identifications into (146) proves the claim.

E Future Directions

The appendix establishes a structured theory package with internally verified statistical budgets under the stated assumptions, including tabular, epochwise / cross-fit linear, modified pure-online self-normalized linear / finite-rank kernel, and projected structured results. The remaining directions are therefore comparatively localized. The most natural ones are:

1. extending the current structured verifications beyond TV-linear-mixture and finite-dimensional projected settings, especially to richer uncertainty geometries and broader function classes;
2. closing the original one-update-per-episode backward-validation broad-function route, for which Proposition 56 isolates the present obstacle;
3. weakening the remaining coverage / excitation conditions and sharpening the current linear and kernel rates.

Learned simulators and error propagation. The operator-error terms also suggest studying robust planning with learned transition simulators. Such an extension would need to quantify both numerical sampling error and the effects of training on model-generated transitions. Relevant analyses in generative modeling include the local truncation error and stability analysis of BELM diffusion samplers [41], and the study of recursive-training collapse and mitigation through real-data augmentation in Reflow [42]. Applying these insights to transition modeling would require new bounds connecting simulator errors to the robust Bellman operator while accounting for strategic coverage.

F Broader Impacts

This work is foundational: it studies robust zero-sum Markov games as a formal model for learning under adversarial or distributional deployment shift. A potential positive impact is improved robustness calibration in safety-critical competitive settings, where nominally strong policies can fail under modest misspecification. A potential negative impact is that the same tools could be used to design stronger adaptive strategies in adversarial domains. We therefore view the present results primarily as analytical guidance for robust evaluation and controlled experimentation rather than as a stand-alone argument for unrestricted deployment.

G Additional Experimental Details

This appendix gives the experimental details for the theorem-aligned benchmark suite reported in the main text. The current submission focuses on this suite because it most directly matches the paper's theoretical hierarchy and permits exact or near-exact deployment evaluation.

G.1 Benchmarks

Linear-mixture robust Markov game. The main large-state benchmark is a zero-sum robust Markov game with linear-mixture nominal dynamics, linear rewards, and moderate action sets. We report results for feature dimension $d \in \{8, 16, 32\}$. The $d = 8$ setting is used for the main comparison and for a dedicated $\beta \times \tau$ resweep, while the $d \in \{16, 32\}$ settings are used to test whether the same qualitative tradeoff persists as the approximation problem becomes harder.

Exact tabular games. The tabular suite contains two small stochastic zero-sum games, `tabular_attack_defense` and `tabular_queue_guard`. These environments are small enough that the relevant robust backups can be computed exactly, so they serve as controlled validation benchmarks for the qualitative behavior of the practical method.

Support-shift benchmarks. We report two support-shift benchmarks. The original `support_shift` benchmark acts mainly as a calibration test, since its shift does not strongly separate the learned policies. The redesigned `support_shift` benchmark perturbs the aggressive deployment branch and is used to test whether ODRB changes the deployed strategy in a robustness-favorable direction.

G.2 Methods

Compared methods. The primary comparisons are the full ODRB practical variant, the matched non-robust baseline SIGMA0, and a no-exploiter control. In the main text and tables, SIGMA0 corresponds to the same practical implementation with the robust term disabled, while the no-exploiter control removes the exploiter-assisted data-collection component.

Model selection in the linear suite. For `linear_mixture`, we distinguish between proxy-tuned and deployment-selected ODRB operating points. The proxy-tuned point is selected by the backup-based worst-case proxy, while the deployment-selected point is an *oracle diagnostic* selected after evaluating exact deployment worst-case return within the low-gap regime. The latter is not a training-time selection rule and is reported only to expose the proxy–deployment mismatch. To study this distinction, we run an additional $d = 8$ $\beta \times \tau$ resweep over $\beta \in \{0.20, 0.30, 0.40, 0.50\}$, $\tau \in \{0, 0.01, 0.02\}$.

Support-shift ablations. For the redesigned `support_shift` benchmark, we sweep $\beta \in \{0.10, 0.20, 0.30\}$ and record the induced stage-0 action masses. This allows us to test not only whether ODRB improves deployment robustness, but also whether it changes the learned strategy from policies concentrated on the aggressive action toward a more conservative deployment policy.

G.3 Metrics

Primary metrics. The main reported metrics are nominal return, backup-based proxy worst-case return, exact deployment worst-case return, and approximate Nash gap. The proxy metric matches the internal robust Bellman evaluation used by the practical method, whereas the deployment metric evaluates the learned policy directly in the perturbed environment family.

Auxiliary diagnostics. In the linear suite we additionally report occupancy entropy as a strategic-coverage diagnostic. In the redesigned support-shift benchmark we report stage-0 safe versus aggressive action mass in order to diagnose whether robustness gains arise from a meaningful policy change rather than only from a change in value estimation.

G.4 Statistical Protocol

Seeds and aggregation. Unless otherwise noted, the main $d = 8$ linear comparison and the $d = 8$ $\beta \times \tau$ resweep use 10 seeds. The $d \in \{16, 32\}$ linear sweep and redesigned support-shift ablations use 5 seeds. All tables report means with 95% confidence intervals when the metric is aggregated across seeds. The primary $d = 8$ comparison is paired by environment seed. We report the mean paired deployment-return difference, its paired-t 95% confidence interval, and a two-sided exact sign-flip test; the two ODRB-versus-SIGMA0 tests are corrected by Holm's procedure. These paired tests were added after reviewer feedback and should be read as small-sample evidence rather than a broad performance claim.

Scope of reported results. The main text emphasizes the strongest and most decision-relevant findings. The appendix records the full theorem-aligned result tables, including the dimension sweep, the deployment-driven linear resweep, the tabular sanity suite, and the redesigned support-shift ablation.

G.5 Compute Resources

Hardware and implementation. All reported experiments are lightweight synthetic experiments. They were run on an internal lab workstation/server with a multi-core CPU and four NVIDIA RTX 2080 Ti GPUs with 11GB GPU memory each. The implementation is NumPy/SciPy based and CPU-bound; the GPUs were available on the machine but were not used by the reported runs. Each job is a single Python process. The reported raw seed-level JSON files occupy about 7.9MB, and the full collection of exploratory and reported seed-level outputs occupies about 11.1MB.

Per-run and total compute. Table 4 summarizes the compute needed to reproduce the reported experimental tables. The full reproduction manifest contains 389 single-seed jobs. Individual tabular and support-shift runs typically finish in under one minute, while the 5000-episode linear-mixture runs typically finish in under two minutes on the CPU workstation described above. Running the full manifest sequentially therefore requires only a few CPU-hours; in our server runs, the main reported queues completed within a few hours of wall-clock time.

Additional exploratory compute. The full research project required more compute than the final reported experiment manifest. In addition to the 389 seed-level runs used to regenerate the reported tables, we ran preliminary debugging, calibration, and failed-setting diagnostics while correcting the zero-sum backup and tuning the support-shift benchmarks. These exploratory runs were of the same small synthetic scale as the reported runs; the complete local archive contains 576 seed-level JSON outputs. The supplementary material includes the exact reproduction manifest and the raw JSON outputs used for the reported tables.

Table 4: Compute resources for the reported experiment groups. Runtime ranges are approximate single-job wall-clock times on the internal CPU workstation used for the paper.

Experiment group	Jobs	Episodes/run	Seeds	Time/run
Tabular sanity suite	54	1500	3	< 1 min
Original support-shift calibration	30	2000	3	< 1 min
Linear-mixture dimension sweep	45	5000	5	< 2 min
Linear-mixture $d = 8$ seed extension	15	5000	5	< 2 min
Redesigned support-shift sweep	125	2000	5	< 1 min
Deployment-driven $\beta \times \tau$ resweep	120	5000	10	< 2 min
Total reported reproduction manifest	389	–	–	few CPU-hours

Table 5: Linear-mixture main comparison at $d = 8$ with 10 seeds. ODRB rows are labeled by (β, τ), where β is the robustness weight and τ is the validation threshold.

Method	Seeds	Proxy	Deploy	Proxy gap
SIGMA0	10	0.230 ± 0.088	0.230 ± 0.088	0.0080 ± 0.0047
ODRB $(\beta = 0.30, \tau = 0.02)$	10	0.204 ± 0.069	0.239 ± 0.089	0.0072 ± 0.0074
ODRB $(\beta = 0.20, \tau = 0.02)$	10	0.162 ± 0.057	0.236 ± 0.089	0.0069 ± 0.0066
NOEX $(\beta = 0.30)$	10	0.254 ± 0.102	0.288 ± 0.115	0.0644 ± 0.0687

G.6 Additional Results for Theorem-Aligned Benchmarks

We report final results for the theorem-aligned benchmark suite described in the main text. Unless otherwise noted, the main $d = 8$ linear-mixture comparison and the $d = 8$ $\beta \times \tau$ resweep use 10 seeds, while the $d \in \{16, 32\}$ sweep and the redesigned support-shift ablation use 5 seeds.

Linear-mixture main comparison. At $d = 8$, the ordering depends on the evaluation metric. Under the backup-based proxy, SIGMA0 is the strongest low-gap baseline. Under exact deployment worst-case evaluation, the paired comparison of ODRB$(0.30, 0.02)$ against SIGMA0 gives mean difference 0.0094, paired-t 95% confidence interval $[0.0007, 0.0182]$, and exact sign-flip Holm-adjusted $p = 0.0469$. For ODRB$(0.20, 0.02)$, the corresponding values are 0.0065, $[-0.0017, 0.0148]$, and $p = 0.0977$; this second comparison is therefore inconclusive. Both ODRB settings preserve a similarly small approximate gap. The no-exploiter control attains the highest return, but only by moving to a substantially higher-gap regime.

Linear-mixture dimension sweep. Across dimensions, ODRB remains competitive but not uniformly dominant. The mean return is close to SIGMA0 at $d = 16$, and is higher in the five-seed $d = 32$ sample; both settings have a small reported gap. These small dimension sweeps are descriptive rather than confirmatory. The sharpest proxy–deployment mismatch appears at $d = 8$, where the choice of operating point is most consequential.

Deployment-driven $\beta \times \tau$ resweep at $d = 8$. The resweep exposes a clean tradeoff. Setting $\tau = 0$ raises deployment return but pushes ODRB into the same high-gap regime as the no-exploiter control and sharply lowers occupancy entropy. Positive τ restores the low-gap regime and substantially increases entropy, suggesting broader strategic coverage. Within this calibrated region, $(0.20, 0.02)$ is an oracle-selected low-gap diagnostic point, whereas $(0.30, 0.02)$ has the larger sample mean deployment return at the cost of a slightly larger gap. Exact deployment return was not used as a training-time selection criterion.

Tabular sanity suite. The tabular benchmarks provide controlled diagnostics in a setting where worst-case backups can be computed exactly. On `tabular_attack_defense`, the tested ODRB setting reaches worst-case return 0.660 at $\beta = 0.20$, below 0.822 for SIGMA0; hence this benchmark does not show an ODRB advantage. On `tabular_queue_guard`, SIGMA0 attains nominal return 4.077 but falls to worst-case return 2.008, whereas ODRB at $\beta = 0.20$ changes only from 1.154 to 1.080. These exact small-game results are consistent with the view that robustness calibration can trade away nominal value, but do not establish uniform improvement over SIGMA0.

Table 6: Paired seed-level deployment-return differences, ODRB minus SIGMA0, for the main $d = 8$ comparison.

Seed	ODRB $(0.30, 0.02)$	ODRB $(0.20, 0.02)$
0	0.002070	0.002015
1	0.024937	0.008538
2	0.031841	0.032091
3	0.020964	0.016377
4	-0.001782	-0.001730
5	-0.000085	-0.001470
6	0.005018	-0.005283
7	0.010971	0.014524
8	0.000000	0.000000
9	0.000299	0.000102

Table 7: Linear-mixture dimension sweep across $d \in \{8, 16, 32\}$.

Dim	Method	Proxy	Deploy	Gap
8	SIGMA0	0.230 ± 0.088	0.230 ± 0.088	0.0080 ± 0.0047
8	ODRB $(\beta = 0.30, \tau = 0.02)$	0.204 ± 0.069	0.239 ± 0.089	0.0072 ± 0.0074
8	NOEX $(\beta = 0.30)$	0.254 ± 0.102	0.288 ± 0.115	0.0644 ± 0.0687
16	SIGMA0	0.0556 ± 0.0313	0.0556 ± 0.0313	0.0166 ± 0.0115
16	ODRB $(\beta = 0.30, \tau = 0.02)$	0.0532 ± 0.0336	0.0532 ± 0.0336	0.0093 ± 0.0074
16	NOEX $(\beta = 0.30)$	0.105 ± 0.074	0.105 ± 0.074	0.0604 ± 0.0638
32	SIGMA0	0.0686 ± 0.0426	0.0686 ± 0.0426	0.0102 ± 0.0111
32	ODRB $(\beta = 0.30, \tau = 0.02)$	0.0757 ± 0.0488	0.0757 ± 0.0488	0.0088 ± 0.0088
32	NOEX $(\beta = 0.30)$	0.0720 ± 0.0558	0.0720 ± 0.0558	0.0160 ± 0.0128

Support-shift benchmarks. The original `support_shift` benchmark is best interpreted as a calibration test rather than a separator: aggregated over all overlap levels, ODRB attains worst-case return 0.4561 ± 0.0038 and SIGMA0 attains 0.4562 ± 0.0038. By contrast, the redesigned deployment-oriented benchmark is genuinely discriminative: ODRB with $\beta \in \{0.10, 0.20\}$ consistently exceeds SIGMA0 on exact deployment worst-case return throughout the tested overlap band. For example, at $m = 0.10$, ODRB attains 0.5742 while SIGMA0 attains 0.4983.

Policy diagnostics. The redesigned support-shift benchmark also clarifies why the methods separate. At $m = 0.05$, 0.10, and 0.15, SIGMA0 places most of its stage-0 mass on the aggressive action in all $5/5$ seeds. By contrast, ODRB with $\beta = 0.10$ places most of its stage-0 mass on the safe action in $5/5$ seeds at $m = 0.05$, 0.08, 0.12, and 0.15, and in $4/5$ seeds at $m = 0.10$. The $\beta = 0.20$ setting shows the same safe-biased pattern with slightly more mixing. This diagnostic supports the intended mechanism story: the redesigned benchmark does not merely reward a different value estimator, but actually induces a different deployed strategy.

Interpretation. Taken together, the theorem-aligned suite supports a qualified but positive empirical conclusion. The strongest evidence comes from the redesigned deployment-oriented support-shift benchmark, where ODRB consistently improves deployment robustness and changes the learned policy in the intended safe direction. The linear-mixture results are more nuanced: ODRB is competitive across dimensions and admits deployment-favorable low-gap operating points, but it does not uniformly dominate the matched SIGMA0 baseline under the backup proxy. Overall, the experiments illustrate robustness calibration, strategic coverage, and the return–gap trade-off; they do not establish uniform improvement over non-robust self-play.

Table 8: Deployment-driven $d = 8$ ODRB resweep over (β, τ).

Setting	Proxy	Deploy	Gap	Entropy
ODRB $(\beta = 0.20, \tau = 0.01)$	0.168 ± 0.063	0.245 ± 0.092	0.0120 ± 0.0164	4.585 ± 0.229
ODRB $(\beta = 0.20, \tau = 0.02)$	0.162 ± 0.057	0.236 ± 0.089	0.0068 ± 0.0065	5.067 ± 0.197
ODRB $(\beta = 0.30, \tau = 0.02)$	0.204 ± 0.069	0.239 ± 0.089	0.0073 ± 0.0074	5.076 ± 0.192
ODRB $(\beta = 0.30, \tau = 0)$	0.254 ± 0.102	0.288 ± 0.115	0.0624 ± 0.0669	3.781 ± 0.209
ODRB $(\beta = 0.50, \tau = 0.02)$	0.232 ± 0.091	0.236 ± 0.094	0.0126 ± 0.0148	5.022 ± 0.191

Table 9: Redesigned support-shift benchmark: deployment worst-case return across overlap levels.

Method	$m = 0.05$	$m = 0.08$	$m = 0.10$	$m = 0.12$	$m = 0.15$
SIGMA0	0.5365	0.5618	0.4983	0.5049	0.4600
ODRB $\beta = 0.10$	0.5952	0.5940	0.5742	0.5924	0.5912
ODRB $\beta = 0.20$	0.5835	0.5940	0.5742	0.5924	0.5912
ODRB $\beta = 0.30$	0.5600	0.5779	0.5552	0.5486	0.5912
NOEX $\beta = 0.20$	0.5739	0.6000	0.6000	0.6000	0.6000

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and Introduction state the theorem hierarchy explicitly, distinguishing conditional broad-function interfaces from the structured closed results and noting that the unchanged raw backward-validation pure-online route is not claimed as closed.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The Introduction and Appendix E delimit the current scope of the theory package, state the remaining structural assumptions, and identify the unresolved raw backward-validation pure-online route and broader coverage extensions.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes]

Justification: The main-text Theory section states the assumptions used by the compressed results, and the appendix provides the formal theorem map, the full structured assumptions, and complete proofs.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: The Experiments section and Appendix G specify the benchmark families, compared methods, metrics, seed counts, and the main hyperparameter sweeps used to generate the reported results.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).

(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: We release code of our experiment and it can reproduce experimental results.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (https://neurips.cc/public/guides/CodeSubmissionPolicy) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (https://neurips.cc/public/guides/CodeSubmissionPolicy) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: The main text summarizes the theorem-aligned benchmark suite, and Appendix G details the benchmark definitions, compared variants, selection sweeps, metrics, and statistical protocol used in evaluation.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: The experimental appendix reports means with 95% confidence intervals across seeds and states the number of seeds used for each benchmark family and sweep in Appendix G.

   Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Appendix G.5 reports the compute hardware, implementation type, per-group job counts, approximate per-run wall-clock time, total reproduction compute, storage footprint, and additional exploratory compute.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics https://neurips.cc/public/EthicsGuidelines?

   Answer: [Yes]

   Justification: The work is a theoretical and synthetic-benchmark study; it does not involve human subjects, sensitive personal data, or deployment of high-risk models, and the submission preserves anonymity.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [Yes]

Justification: Appendix F discusses both the positive role of robustness calibration in safety-critical competitive settings and the possibility of misuse in adversarial domains.

Guidelines:

- The answer [N/A] means that there is no societal impact of the work performed.
- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

Answer: [N/A]

Justification: The paper does not release high-risk models, scraped datasets, or other assets that require a controlled release mechanism.

Guidelines:

- The answer [N/A] means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [N/A]

Justification: The reported experiments use author-constructed theorem-aligned synthetic benchmarks and do not rely on external datasets, released models, or third-party benchmark assets as part of the reported results.

Guidelines:

- The answer [N/A] means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, paperswithcode.com/datasets has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

    Justification: The submission does not release new datasets, models, or code assets alongside the paper.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing or human-subject research.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [N/A]

Justification: The paper does not involve human-subject research or data collection.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [N/A]

Justification: The core method development in this research does not involve LLMs as any important, original, or non-standard components.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.