AIMS Mathematics, 11(9): 30581–30618.
DOI: 10.3934/math.20261212
Received: 09 August 2026
Revised: 03 September 2026
Accepted: 11 September 2026
Published: 20 September 2026

http://www.aimspress.com/journal/Math

Research article

Proximal policy learning with latent confounding and limited overlap: identification and oracle-calibrated regret bounds

Siyang Bai[1,*], **Zheng Fang**[2] **and Jie Chen**[2,*]

[1] School of Statistics and Data Science, Nankai University, Tianjin 300071, China
[2] College of Electronic Engineering, National University of Defense Technology, Changsha 410073, China

* **Correspondence:** Email: 2413855@mail.nankai.edu.cn; jchen202209@163.com.

Abstract: Data-driven sequential policy optimization from observational time series is difficult when latent states affect both treatment assignment and rewards and when overlap deteriorates over time. We develop temporally adaptive proximal learning under overlap extremes (Taple), a proximal policy learner that addresses this optimization problem through nested intervention-indexed bridge operators and temporally adaptive cumulative-weight capping. Its identifying functional avoids an invalid stagewise Bellman substitution and is stagewise doubly robust when the bridge choices are recursively compatible. For the proposed behavior-law implementation, learning a later policy-indexed outcome bridge from transported moments still requires valid preceding treatment bridges, unless another valid estimator of the relevant intervention-law moments is available. Under explicit conditions on bridge estimation, score entropy, policy indexing, numerical optimization, and cap calibration, we establish an oracle-calibrated fixed-horizon regret upper bound together with a minimax lower bound under polynomial correction-weight tails. When score-class and policy-class complexities are comparable and the nuisance and optimization remainders are of oracle order, the two bounds match in their overlap and policy-complexity exponents up to logarithmic and fixed-horizon factors. The executable factorized analogue, Taple-F, is deliberately restricted to an action-invariant factorized submodel and is not empirical validation of the full cumulative-score algorithm or its global optimization guarantee. Here "data-driven" means that bridge functions, caps, and policies are learned from observational data within an explicitly stated causal and statistical model; it does not mean model-free reinforcement learning. In controlled simulations, the locally capped factorized benchmark has lower mean regret than the corresponding proximal baseline without the adaptive local cap. A source-grouped PhysioNet semi-synthetic stress test shows only a modest local-capping improvement; with only nine training source records, observable-history doubly robust methods perform better in that experiment.

Keywords: data-driven optimization; sequential policy learning; proximal causal learning; limited overlap; minimax regret

Mathematics Subject Classification: 62C20, 62D20, 62M10, 68T05

Notation. Throughout the paper, n denotes the number of independent trajectories and T the number of decision stages. H_t is the observed pre-action history available to the target policy, excluding the current designated proxies Z_t and W_t; U_t is the latent time-varying confounder; A_t is the logged action; and Y_t is the stage reward. Z_t and W_t denote the action-side and outcome-side proxies, respectively. P_t^π is the intervention law through stage $t-1$, $\mathcal{D}_t^\pi$ is the one-step proximal operator, h_t^π is the policy-indexed outcome bridge, and q_t is the policy-stable treatment bridge. η_t^π and ω_t^π are the cumulative matched and correction bridge weights; Λ_t is the stagewise cumulative cap; and α_t is the weight-tail exponent. v_t^{pol} and v_t^{sc} are the stage policy and capped-score complexity indices, $d_{n,t}$ is the local entropy quantity, and d_t^{br} is the effective bridge-domain dimension in the illustrative sieve rate. $r_{h,t}$ and $r_{q,t}$ denote outcome- and treatment-bridge errors, $R_{t,n}^{\mathrm{idx}}$ is the stage-specific nested transport/policy-indexing remainder, and $\rho_{t,n}$ is the oracle overlap-rate benchmark. Finally, γ is the discount factor, $\mathbf{\Pi}$ is the candidate policy class, and $\mathrm{Reg}(\pi)$ denotes regret relative to $\pi^\star$.

1. Introduction

Dynamic treatment regimes assign actions repeatedly as information accumulates. Their statistical formulation builds on optimal sequential decision rules and the associated inferential challenges [1, 2]; outcome-weighted learning provides a complementary classification view of individualized treatment choice [3]. Modern longitudinal policy research formalizes this objective through stage-dependent decision rules and counterfactual value functions [4], while observational policy learning supplies regret analyses for data-dependent rules [5]. Recent stage-aware learning highlights the sample-efficiency loss caused by requiring entire observed trajectories to match a target regime [6], and advantage-based doubly robust methods establish welfare-regret guarantees for when-to-treat decisions [7]. The longitudinal setting is nevertheless harder than a single-stage decision problem: early actions may alter later histories, rewards, and support. A policy can therefore look favorable under a fitted model while relying on histories that are scarcely represented under the logging mechanism. This is simultaneously an identification problem and a data-driven optimization problem, because the learned decision rule must maximize policy value under unstable observational support.

Offline reinforcement learning and dynamic treatment research increasingly use routinely collected trajectories. Hidden clinical severity, operator workload, or system health may affect both actions and outcomes. Proximal reinforcement learning identifies policy values through proxy bridge functions under partial observability [8], whereas analyses of confounded offline optimization show that unrestricted hidden-state effects can make policy evaluation impossible [9]. Complementary algorithmic work in [10] analyzes value and policy iteration for Markov decision processes, while structured two-way latent confounding has been studied for off-policy evaluation [11]; robust Markov decision process (MDP) sensitivity analysis instead targets sharp policy-value bounds under transition perturbations [12]. A finite-time Q-learning analysis in [13] provides another optimization perspective under explicit convergence constraints, but these optimization results do not, by themselves, remove the identification difficulty created by latent confounding. These observations motivate guarantees that reveal, rather than conceal, the statistical price of latent state dynamics.

The reinforcement-learning (RL) motivation in this paper is specifically offline and causal: the decision maker receives a sequence of rewards and optimizes a policy, but no online exploration is available to repair confounding or weak support. In contrast, Q-learning, policy iteration, actor–critic learning, graph RL, and metaheuristic optimization address important computational aspects of sequential decision making when the state–action objective is directly learnable or simulatable [14–16]. Related population-based and unsupervised optimization studies illustrate how structural information can improve difficult search and representation problems [17, 18]; deep RL combined with alternative or metaheuristic optimization has also been used for constrained communication-resource allocation [19]. These works motivate the optimization layer of our framework, while the present contribution addresses the distinct preceding question of how an observational sequential objective can be identified when latent confounding remains.

Proxy-variable identification offers a principled alternative to sequential ignorability. Nonparametric negative-control arguments use two measurements that carry information about an unobserved confounder while satisfying exclusion restrictions [20]. Recent individualized-treatment work extends this idea to policy search [21]. Proximal regime and conditional-effect methods distinguish outcome and treatment bridge representations [22, 23]. Our use of cross-fitting and orthogonal scores is also connected to double/debiased machine learning [24] and double reinforcement learning for off-policy evaluation [25]. Existing results, however, generally analyze fixed policies or assume bridge moments with sufficiently light tails. They do not characterize longitudinal policy regret when proxy inversion and support deterioration vary across stages.

Limited overlap creates a second obstacle. Classical trimming stabilizes inverse weighting by changing the target population [26], while high-dimensional analyses show why strict positivity becomes implausible as histories grow richer [27]. In long-horizon off-policy evaluation, overlap degradation is one mechanism behind the curse of horizon [28]; adaptive data collection can similarly produce diminishing action propensities even when minimax policy learning remains possible under appropriate weighting [29]. Positivity-free stochastic interventions provide a different estimand [30], localized intervention policies deliberately remain near observational support [31], and recent tests quantify whether overlap failure can be detected from observed data [32]. In a trajectory, deleting an entire record because one stage is weakly supported discards information at every other stage and may change the longitudinal target in an opaque way.

Safety, selective observation, and inverse problems further complicate policy construction. Safe learning under partial identifiability compares policies without pretending that all causal quantities are point identified [33]. Off-policy evaluation under nonignorable missingness shows that observation mechanisms must enter the identification argument [34]. Longitudinal proximal causal theory shows that time-varying hidden confounding requires recursively compatible bridge operators rather than isolated one-stage moments [35]. Longitudinal clinical analyses increasingly evaluate realistic strategies with doubly robust procedures [36]. Together, these developments call for a stage-aware method whose causal assumptions, tail requirements, and numerical diagnostics are explicit.

Temporal dependence and nonstationarity also require attention. The authors in [37, 38] developed stochastic optimization under Markov sampling and studied generalization for Markov-ensemble learning, illustrating how dependent sampling changes optimization and learning analyses. EpiCare shows that standard estimators can fail even in carefully designed dynamic-treatment benchmarks [39]. A credible learner should therefore report stagewise support and weight-concentration diagnostics, not

merely a final policy score.

This paper develops TAPLE, short for temporally adaptive proximal learning under overlap extremes, as a data-driven framework for sequential policy optimization under latent confounding and weak support. The term data-driven is used in the statistical sense: the nuisance bridges, overlap caps, and policy are estimated from observed trajectories, but their estimability is governed by explicit proxy, transport, support, and bridge-class assumptions. The method is therefore not model free.

The revised contribution can be summarized in four points. First, we introduce intervention-indexed nested proximal operators whose outcome bridges depend on the downstream target policy; this avoids replacing a latent continuation value by an observed-history Bellman regression. Second, repeated expansion reveals cumulative direct and correction bridge weights, which are the longitudinal quantities that must be regularized when overlap deteriorates. Third, we give a fixed-horizon regret analysis that separates bridge-product error, policy-indexing error, capping–nuisance interaction, empirical fluctuation, cap calibration, and numerical optimization error, together with a matching-exponent minimax lower bound under the stated rate-matching conditions. Fourth, we provide an executable factorized analogue, TAPLE-F, to isolate proxy correction and local adaptive capping; this experiment does not implement the full carryover algorithm, and the revised empirical claims are restricted accordingly.

Our theoretical analysis pairs a fixed-horizon minimax lower bound under polynomial cumulative bridge-weight tails with an oracle-calibrated upper bound under additional bridge-learning, entropy, policy-indexing, cap-oracle, and optimization conditions. A polynomial tail condition is imposed on the policy-specific cumulative correction weight, which is the quantity that actually drives longitudinal variance. For the oracle-calibrated upper bound, a tail exponent $\alpha < 1$ yields an overlap component of order $(d_n/n)^{\alpha/(1+\alpha)}$; for $\alpha > 1$, the component has square-root complexity order $(d_n/n)^{1/2}$, with a logarithmic correction at $\alpha = 1$. The lower bound has the same overlap exponents with policy-class complexity in place of score-class complexity. A one-stage lower bound is lifted to the longitudinal class by reduction. The upper bound includes bridge-product, capping-interaction, policy-indexing, and optimization errors explicitly; rate matching requires comparable policy and score-class complexity in addition to the displayed nuisance conditions. Controlled experiments use a factorized design with a known oracle, while a source-grouped PhysioNet protocol retains real intensive-care-unit (ICU) covariate paths but generates actions and rewards transparently. The latter is a semi-synthetic stress test, not evidence of a clinical treatment effect.

Overall, the framework links proximal identification, statistical regularization, and policy optimization while keeping their assumptions and error contributions explicit.

2. Problem formulation

2.1. Observed trajectories and policy value

For trajectory i, let

$$O_i = (H_{i1}, Z_{i1}, W_{i1}, A_{i1}, Y_{i1}, \ldots, H_{iT}, Z_{iT}, W_{iT}, A_{iT}, Y_{iT}), \tag{2.1}$$

where H_t is the observed pre-action history available to the target policy, while Z_t and W_t are the designated action-side and outcome-side proxies used by the proximal identification argument. They are kept separate from the current policy input to prevent the policy class from exploiting variables whose role is reserved for bridge identification. The trajectories $O_1, \ldots, O_n$ are independent, while dependence

within a trajectory is unrestricted. Throughout the main theory, the horizon $T < \infty$ is fixed and known, actions are binary, and the target policy is deterministic. The observed history H_t may contain baseline covariates, past observed states, past actions, and past rewards, but excludes the current designated proxies unless they have already entered through earlier history. Potential outcomes are defined under no interference between sampling units: the potential trajectory of unit i depends on its own action history and not on the actions assigned to another unit. Continuous or combinatorial actions are outside the present theorem and are listed as an extension. A deterministic policy $\pi = (\pi_1, \ldots, \pi_T)$ satisfies

$$E_t^\pi = \pi_t(H_t) \in \{0, 1\}. \tag{2.2}$$

Let $Y_t(\bar{a}_t)$ denote the potential stage reward under action history $\bar{a}_t$. The discounted value is

$$V(\pi) = \mathbb{E}\left[\sum_{t=1}^{T} \gamma^{t-1} Y_t(\bar{A}_t^\pi)\right], \qquad 0 < \gamma \leq 1. \tag{2.3}$$

For a class $\mathbf{\Pi} = \Pi_1 \times \cdots \times \Pi_T$, define

$$\pi^\star \in \arg\max_{\pi \in \mathbf{\Pi}} V(\pi) \tag{2.4}$$

and

$$\mathrm{Reg}(\widehat{\pi}) = V(\pi^\star) - V(\widehat{\pi}). \tag{2.5}$$

The logging law may depend on an unobserved state U_t:

$$e_t(a \mid H_t, U_t) = \mathbb{P}(A_t = a \mid H_t, U_t). \tag{2.6}$$

Figure 1 displays the local proxy roles. The figure is a schematic proxy graph rather than a fully saturated causal diagram. Solid arrows show the core latent-confounding and history-update structure. Dashed arrows $Z_t \to A_t$ and $W_t \to Y_t$ indicate direct treatment- and outcome-inducing links that are permitted, but not required, by Assumption 2.2. Conditioning on A_t in the Z_t negative-control restriction is essential when Z_t may directly influence treatment assignment; this is consistent with the treatment- and outcome-inducing proxy interpretation used in proximal causal models [20, 35]. In contrast, shortcuts such as $Z_t \to Y_t$, $A_t \to W_t$, and $Z_t \to W_t$ are excluded by the negative-control restrictions. Other history-to-outcome, history-to-proxy, latent-state-transition, and history-update arrows are suppressed for readability.

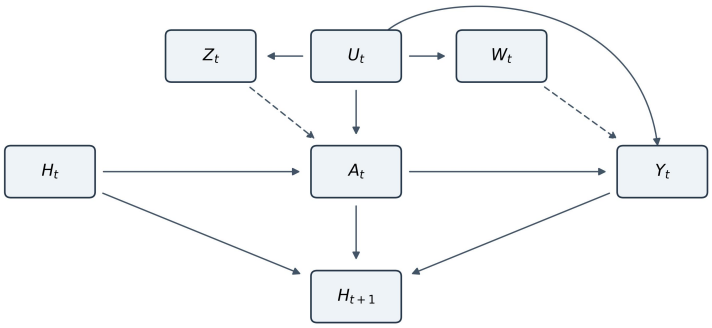

Figure 1. Schematic longitudinal proxy structure. Dashed $Z_t \to A_t$ and $W_t \to Y_t$ arrows are allowed but not required by the proxy assumptions; prohibited negative-control shortcuts are absent, and nonessential transition arrows are suppressed.

2.2. Intervention laws and admissible proxies

For a policy π, let P_t^π denote the law under which stages $1, \dots, t-1$ follow π and stages $t, \dots, T$ follow the logging mechanism. Thus,

$$P_1^\pi = P_b, \qquad P_{T+1}^\pi = P_\pi, \tag{2.7}$$

where P_b is the observed law and P_π is the full policy law. The sequence P_t^π is essential: a one-step proximal identity is valid under P_t^π, not automatically under the observed law after arbitrary future interventions.

Assumption 2.1 (Consistency and latent sequential exchangeability)**.** Whenever factual and counterfactual action histories agree,

$$Y_t = Y_t(\bar{A}_t), \qquad H_{t+1} = H_{t+1}(\bar{A}_t). \tag{2.8}$$

For every t, a, and future measurable outcome $G_t(a)$,

$$A_t \perp G_t(a) \mid H_t, U_t \quad \text{under } P_t^\pi. \tag{2.9}$$

Assumption 2.2 (Sequential negative controls)**.** For every admissible π, stage t, action a, and integrable potential future pseudo-outcome $G_t(a)$ measurable with respect to variables temporally downstream of A_t,

$$Z_t \perp G_t(a) \mid H_t, U_t, A_t, \tag{2.10}$$

$$W_t \perp (A_t, Z_t) \mid H_t, U_t \tag{2.11}$$

under a regular conditional version of P_t^π. The target action $E_t^\pi = \pi_t(H_t)$ is measurable with respect to H_t and does not use Z_t or W_t beyond variables already included in H_t.

For clarity, let $\mathcal{G}_t$ denote the admissible class of downstream pseudo-outcomes generated by the recursion in Section 3 from policies in $\mathbf{\Pi}$ and bridge functions in the declared bridge classes $\mathcal{H}_t$ and $\mathcal{Q}_t$. We require joint measurability in (O, π) and integrability under the relevant intervention law. The bridge-existence and completeness statements below are asserted only on these declared classes, not for every arbitrary measurable function. Uniform statements over $\mathbf{\Pi}$ refer to these jointly measurable classes and to the envelope/entropy conditions stated later.

Assumption 2.3 (Bridge existence and treatment-bridge stability)**.** For each t, π, and integrable future pseudo-outcome G_t for which the displayed conditional expectations exist, there are functions q_t and $h_t^{\pi,G}$ such that, for $a \in \{0, 1\}$,

$$\mathbb{E}_{P_t^\pi}\{q_t(Z_t, a, H_t) \mid U_t, H_t, A_t = a\} = e_t(a \mid H_t, U_t)^{-1}, \tag{2.12}$$

$$\mathbb{E}_{P_t^\pi}\{h_t^{\pi,G}(W_t, a, H_t) \mid U_t, H_t, A_t = a\} = \mathbb{E}_{P_t^\pi}(G_t \mid U_t, H_t, A_t = a). \tag{2.13}$$

The treatment bridge q_t is assumed to solve the latent equation under every P_t^π. This policy stability is justified when the logging and proxy mechanisms conditional on (U_t, H_t) are invariant to earlier interventions. The outcome bridge remains policy indexed because its pseudo-outcome changes with the target tail policy. Without treatment-bridge stability, q_t must also be indexed by π.

Assumption 2.4 (Observed identification, transport support, and boundedness)**.** The behavior-law treatment-bridge operator and the transported outcome-bridge operator under every P_t^π are complete on the declared bridge classes. Hence the unique solutions of

$$\mathbb{E}_{P_b}\{q_t(Z_t, a, H_t) \mid W_t, H_t, A_t = a\} = \mathbb{P}_{P_b}(A_t = a \mid W_t, H_t)^{-1}, \tag{2.14}$$

$$\mathbb{E}_{P_t^\pi}\{h_t^{\pi,G}(W_t, a, H_t) \mid Z_t, H_t, A_t = a\} = \mathbb{E}_{P_t^\pi}(G_t \mid Z_t, H_t, A_t = a) \tag{2.15}$$

coincide with valid latent bridges. Completeness in the second equation concerns the intervention-law conditional operator; it is not inferred from behavior-law regression fit.

For every action that can be selected by an admissible policy, $e_t(a \mid H_t, U_t) > 0$ almost surely, and the corresponding observed conditional probability $\mathbb{P}_{P_b}(A_t = a \mid W_t, H_t)$ is positive almost surely; no uniform lower bound away from zero is imposed. The support of (H_t, W_t, Z_t) under each P_t^π is contained in its behavior-law support. Together with treatment-bridge stability, the first equation identifies one q_t from behavior data that remains valid under all admissible intervention laws. For the recursively induced bridge solutions used in the value representation and regret analysis, assume fixed constants $B_Y, B_h < \infty$ such that

$$|Y_t| \le B_Y, \qquad |h_t^{\pi,G}(W_t, a, H_t)| \le B_h \tag{2.16}$$

almost surely. This boundedness is imposed on rewards and bridge evaluations, not on the weighted nested pseudo-outcomes themselves: the latter can remain heavy tailed because the treatment bridges may be unbounded. Their cumulative tails are handled explicitly by Assumption 4.1 and by capping.

These proximal assumptions have concrete application-level implications. Action-side proxies should predict treatment after conditioning on recorded history while having no direct pathway to the downstream counterfactual outcome beyond the latent state; outcome-side proxies should carry information about the latent state without being caused by the current action. Completeness is not directly testable, but weak identification can be diagnosed by ill-conditioned conditional-moment operators, unstable bridge coefficients, and large out-of-fold moment residuals. In applications we therefore recommend reporting residual moment norms, the smallest retained singular value or condition number of the bridge system, proxy perturbation analyses, and support diagnostics. Such diagnostics cannot prove the negative-control assumptions, but they can reveal weak or numerically unstable proxy designs.

Remark 2.1 (A finite-state model satisfying the proximal identification assumptions). The assumptions are strong but mutually compatible in nontrivial sequential models. For example, condition on a finite observed history $H_t = h$ and let the latent state $U_t \in \{0, 1\}$ evolve through a finite-state Markov kernel. Let $Z_t, W_t \in \{0, 1\}$ be generated before action selection with conditionally independent proxy noises given (U_t, H_t), with $\mathbb{P}(Z_t = 1 \mid U_t = 1, h) \neq \mathbb{P}(Z_t = 1 \mid U_t = 0, h)$ and $\mathbb{P}(W_t = 1 \mid U_t = 1, h) \neq \mathbb{P}(W_t = 1 \mid U_t = 0, h)$. Let the binary logging action satisfy $\mathbb{P}(A_t = 1 \mid U_t = u, h) \in [\epsilon, 1 - \epsilon]$ for some $\epsilon > 0$, and let the next-history kernel $K_t(h' \mid h, u, a)$ have positive mass on every policy-reachable state and depend on a, so genuine treatment carryover is allowed. Finally, let each bounded downstream pseudo-outcome have the form $G_t(a) = m_t(U_t, h, a) + \xi_t$, where ξ_t is bounded and conditionally independent of (Z_t, W_t, A_t) given (U_t, H_t). Then the stated negative-control relations hold by construction. If, for each (h, a), the finite conditional-probability matrices $[\mathbb{P}(W_t = w \mid Z_t = z, h, A_t = a)]_{z,w}$ and $[\mathbb{P}(Z_t = z \mid W_t = w, h, A_t = a)]_{w,z}$ are nonsingular, the outcome- and treatment-bridge equations are finite linear systems with unique bounded solutions; nonsingularity also gives the corresponding finite-state completeness property.

Positivity and the full-support transition condition give intervention-law transport support, while the local proxy/action kernels remain unchanged under earlier policy interventions, yielding treatment-bridge stability across P_t^π. Thus, the principal identification assumptions can hold simultaneously even with action-dependent longitudinal state evolution.

Remark 2.2. The intervention-law formulation avoids a subtle conditioning error that can arise in a naive Bellman proof. In general, conditioning an identified future value on (U_t, H_t, A_t) cannot be exchanged with conditioning on H_{t+1}. The nested construction below avoids that interchange by solving a new proximal problem under each P_t^π. Bridge existence remains a strong assumption and must be justified by proxy design and diagnostics.

3. Nested proximal identification

3.1. One-step proximal operator

For a future pseudo-outcome G_t, define

$$\begin{aligned}\mathcal{D}_t^\pi(G_t; h_t, q_t) = \; & h_t(W_t, E_t^\pi, H_t) \\ & + \mathbf{1}(A_t = E_t^\pi) q_t(Z_t, A_t, H_t)\{G_t - h_t(W_t, A_t, H_t)\}.\end{aligned} \tag{3.1}$$

The direct bridge is evaluated at the target action. The augmentation is also target-action specific; summing over all actions would identify a different functional. For fixed (π, G_t), let h_t^0 and q_t^0 denote valid outcome and treatment bridges.

Lemma 3.1 (One-step proximal transport)**.** *Under Assumptions 2.1–2.4, let h_t and q_t be candidate bridge functions for which the displayed score is integrable. If either $h_t = h_t^0$ or $q_t = q_t^0$, then*

$$\mathbb{E}_{P_t^\pi}\{\mathcal{D}_t^\pi(G_t; h_t, q_t)\} = \mathbb{E}_{P_{t+1}^\pi}(G_t). \tag{3.2}$$

In particular, the equality holds when both bridges are valid.

Proof. If $h_t = h_t^0$, (2.15) implies

$$\mathbb{E}_{P_t^\pi}[G_t - h_t^0(W_t, A_t, H_t) \mid Z_t, H_t, A_t] = 0. \tag{3.3}$$

Therefore, the augmentation has mean zero for every integrable q_t, and

$$\mathbb{E}_{P_t^\pi}\{\mathcal{D}_t^\pi(G_t; h_t^0, q_t)\} = \mathbb{E}_{P_t^\pi}\{h_t^0(W_t, E_t^\pi, H_t)\}. \tag{3.4}$$

By (2.13), (2.11), latent exchangeability (2.9), and iterated expectation, the right-hand side equals the mean of G_t after replacing the logged stage-t action by E_t^π, which is $\mathbb{E}_{P_{t+1}^\pi}(G_t)$.

If $q_t = q_t^0$, the treatment bridge and (2.11) imply, for every integrable $f(W_t, H_t, a)$,

$$\mathbb{E}_{P_t^\pi}[\mathbf{1}(A_t = E_t^\pi) q_t^0(Z_t, A_t, H_t) f(W_t, H_t, A_t)] = \mathbb{E}_{P_t^\pi}\{f(W_t, H_t, E_t^\pi)\}. \tag{3.5}$$

Taking $f = h_t$ cancels the direct bridge and the bridge part of the augmentation. Next, consistency, latent exchangeability, and (2.10) give $\mathbb{E}\{G_t(a) \mid Z_t, A_t, H_t, U_t\} = \mathbb{E}\{G_t(a) \mid H_t, U_t\}$. Combining this relation with (2.12) gives

$$\mathbb{E}_{P_t^\pi}[\mathbf{1}(A_t = E_t^\pi) q_t^0(Z_t, A_t, H_t) G_t \mid U_t, H_t] = \mathbb{E}_{P_t^\pi}\{G_t(E_t^\pi) \mid U_t, H_t\}. \tag{3.6}$$

Integrating (3.6) yields $\mathbb{E}_{P_{t+1}^\pi}(G_t)$ and proves both cases. □

3.2. Nested value representation

Set

$$G_T^\pi = Y_T. \tag{3.7}$$

For $t = T, T-1, \ldots, 2$, recursively define

$$G_{t-1}^\pi = Y_{t-1} + \gamma \mathcal{D}_t^\pi(G_t^\pi; h_t^\pi, q_t). \tag{3.8}$$

At each step, h_t^π abbreviates h_t^{π, G_t^π} for the current pseudo-outcome. Finally, let

$$\Psi_\pi(O; \vartheta) = \mathcal{D}_1^\pi(G_1^\pi; h_1^\pi, q_1), \tag{3.9}$$

where $\vartheta = \{(h_t, q_t) : 1 \le t \le T\}$ collects the policy-indexed nuisance bridges. This symbol is distinct from the cumulative weight η_t^π introduced below.

Theorem 3.1 (Nested proximal identification)**.** *Under Assumptions 2.1–2.4,*

$$V(\pi) = \mathbb{E}_{P_b}\{\Psi_\pi(O; \vartheta^0)\}. \tag{3.10}$$

At every stage, it is sufficient for the population identifying functional that either the treatment bridge is valid or the outcome bridge is valid for the recursively induced downstream pseudo-outcome generated by the selected later-stage bridges. This representation-level double robustness does not by itself guarantee that an arbitrary valid policy-indexed outcome bridge can be learned from behavior data when preceding treatment bridges are misspecified; the transported-moment implementation in Lemma 3.3 requires valid preceding treatment bridges (or another valid route to the relevant intervention-law moments).

Proof. For $1 \le t \le T$, define the policy tail value

$$\mathcal{V}_t^\pi = \sum\nolimits_{s=t}^{T} \gamma^{s-t} \mathbb{E}_{P_{s+1}^\pi}(Y_s). \tag{3.11}$$

At $t = T$, Lemma 3.1 gives

$$\mathbb{E}_{P_T^\pi}\{\mathcal{D}_T^\pi(Y_T)\} = \mathbb{E}_{P_{T+1}^\pi}(Y_T) = \mathcal{V}_T^\pi. \tag{3.12}$$

Suppose $\mathbb{E}_{P_{t+1}^\pi}(G_t^\pi) = \mathcal{V}_t^\pi$. By Lemma 3.1 and (3.8),

$$\mathbb{E}_{P_t^\pi}(G_{t-1}^\pi) = \mathbb{E}_{P_t^\pi}(Y_{t-1}) + \gamma \mathbb{E}_{P_{t+1}^\pi}(G_t^\pi) = \mathbb{E}_{P_t^\pi}(Y_{t-1}) + \gamma \mathcal{V}_t^\pi = \mathcal{V}_{t-1}^\pi. \tag{3.13}$$

The last equality uses the fact that actions after stage $t-1$ cannot affect Y_{t-1}, so $\mathbb{E}_{P_t^\pi}(Y_{t-1}) = \mathbb{E}_{P_{T+1}^\pi}(Y_{t-1})$. Backward induction and one final use of Lemma 3.1 yield

$$\mathbb{E}_{P_1^\pi}\{\Psi_\pi(O; \vartheta^0)\} = \sum\nolimits_{t=1}^{T} \gamma^{t-1} \mathbb{E}_{P_{t+1}^\pi}(Y_t) = V(\pi). \tag{3.14}$$

Because $P_1^\pi = P_b$, (3.10) follows. Selecting the valid part of Lemma 3.1 at each induction step proves stagewise double robustness, provided an outcome bridge declared valid at stage t is valid for the recursively induced pseudo-outcome entering that stage. □

A two-stage example makes the policy indexing explicit. When $T = 2$,

$$\Psi_\pi = \mathcal{D}_1^\pi(Y_1 + \gamma \mathcal{D}_2^\pi(Y_2; h_2^\pi, q_2); h_1^\pi, q_1)\,. \tag{3.15}$$

The stage-1 outcome bridge is therefore a bridge for a pseudo-outcome that already contains the target stage-2 action $E_2^\pi = \pi_2(H_2)$. Replacing it by a bridge for a behavior-law continuation generally targets a different conditional mean whenever the stage-1 intervention changes the distribution of H_2. This is why h_1^π is indexed by the future intervention and why the recursion proceeds through $P_1^\pi \to P_2^\pi \to P_3^\pi$ instead of using a single observed-law Bellman substitution. Joint measurability of $(O, \pi) \mapsto h_t^\pi$ and the uniform entropy/indexing conditions are explicitly included in the declared bridge and score classes.

Remark 3.1 (Scope of stagewise double robustness). "Stagewise doubly robust" refers first to the population representation: at each application of Lemma 3.1, either the current treatment bridge or the outcome bridge for the current recursively induced pseudo-outcome may be valid. This does not mean that every such mixed representation is directly estimable from behavior data. Under the transported-moment implementation, a later outcome bridge generally requires valid preceding treatment bridges to construct its intervention-law moments. Thus, population identification, estimability of a policy-indexed bridge, and finite-sample robustness of the complete learned policy are distinct claims.

3.3. Expanded score and sequential orthogonality

Define

$$\eta_1^\pi = 1, \qquad \eta_{t+1}^\pi = \eta_t^\pi \mathbf{1}(A_t = E_t^\pi) q_t(Z_t, A_t, H_t), \tag{3.16}$$

and

$$\omega_t^\pi = \eta_t^\pi q_t(Z_t, A_t, H_t). \tag{3.17}$$

Repeated substitution in (3.9) gives

$$\begin{aligned}\Psi_\pi(O; \vartheta) = \sum\nolimits_{t=1}^{T} \gamma^{t-1} [&\eta_t^\pi h_t^\pi(W_t, E_t^\pi, H_t) \\ &+ \omega_t^\pi \mathbf{1}(A_t = E_t^\pi)\{Y_t - h_t^\pi(W_t, A_t, H_t)\}].\end{aligned} \tag{3.18}$$

Thus, longitudinal instability is driven by cumulative matched products and cumulative correction weights, not by a single propensity or bridge factor.

Lemma 3.2 (Local drift identity)**.** *Fix G_t and let h_t^0, q_t^0 be valid bridges. For arbitrary h_t, q_t for which the displayed expectations are finite,*

$$\begin{aligned}&\mathbb{E}_{P_t^\pi}\{\mathcal{D}_t^\pi(G_t; h_t, q_t) - \mathcal{D}_t^\pi(G_t; h_t^0, q_t^0)\} \\ &= -\mathbb{E}_{P_t^\pi}[\mathbf{1}(A_t = E_t^\pi)\{q_t - q_t^0\}\{h_t - h_t^0\}(W_t, A_t, H_t)].\end{aligned} \tag{3.19}$$

Proof. Add and subtract $\mathcal{D}_t^\pi(G_t; h_t, q_t^0)$. The difference produced by $h_t - h_t^0$ is

$$\mathbb{E}_{P_t^\pi}[(h_t - h_t^0)(W_t, E_t^\pi, H_t)] - \mathbb{E}_{P_t^\pi}[\mathbf{1}(A_t = E_t^\pi) q_t^0 (h_t - h_t^0)(W_t, A_t, H_t)] = 0 \tag{3.20}$$

by (3.5). In the remaining difference, the component involving $G_t - h_t^0$ has conditional mean zero by (3.3). The only surviving term is the product in (3.19). □

Lemma 3.3 (Behavior-law representation of transported moments)**.** *Let $t \geq 2$ and suppose $q_1^0, \ldots, q_{t-1}^0$ satisfy the policy-stable treatment-bridge equations. For any stage-t or downstream functional f_t for which f_t and the successively weighted products below are integrable, and for which the sequential exchangeability and negative-control transport conditions in Assumptions 2.1 and 2.2 apply at stages $1, \ldots, t-1$,*

$$\mathbb{E}_{P_b}\{\eta_t^{\pi,0} f_t(O)\} = \mathbb{E}_{P_t^\pi}\{f_t(O)\}. \tag{3.21}$$

In particular, if $h_t^{\pi,G}$ satisfies (2.15)*, then for every bounded test function $\phi_t(Z_t, H_t)$ and $a \in \{0, 1\}$,*

$$\mathbb{E}_{P_b}[\eta_t^{\pi,0} \mathbf{1}(A_t = a)\phi_t(Z_t, H_t)\{h_t^{\pi,G}(W_t, a, H_t) - G_t\}] = 0. \tag{3.22}$$

Thus, the intervention-law outcome-bridge equation can be fitted from behavior data through transported unconditional moments, provided the preceding treatment bridges are valid. The identity is only asserted for the declared admissible moment class; it does not make $\eta_t^{\pi,0}$ a probability density for arbitrary functionals.

Proof. For $t = 2$, repeat the treatment-bridge argument used to obtain (3.6), replacing its downstream pseudo-outcome by the admissible functional f_2. This gives

$$\mathbb{E}_{P_b}[\mathbf{1}(A_1 = E_1^\pi) q_1^0(Z_1, A_1, H_1) f_2] = \mathbb{E}_{P_2^\pi}(f_2). \tag{3.23}$$

For general t, repeat the same argument successively under the laws P_s^π, $s = 1, \ldots, t-1$. The product accumulated by these steps is exactly $\eta_t^{\pi,0}$ in (3.16), proving (3.21). Taking $f_t = \mathbf{1}(A_t = a)\phi_t(Z_t, H_t)$ $\{h_t^{\pi,G}(W_t, a, H_t) - G_t\}$ and using (2.15) under P_t^π proves (3.22). □

Assumption 3.1 (Nested transport and indexing stability)**.** For fixed T, the recursively induced pseudo-outcomes and all products appearing in the one-step transport identities are integrable under the relevant P_t^π laws, uniformly over the declared policy class. When an estimated downstream pseudo-outcome replaces its recursively compatible population target, any approximation, transported-functional, or policy-indexing discrepancy in the associated outcome bridge is recorded in a nonnegative remainder $R_{t,n}^{\mathrm{idx}}$ rather than absorbed into the orthogonal product term. Apart from this declared discrepancy, the true treatment bridge transports the signed mean of an admissible downstream perturbation from P_t^π to P_{t+1}^π as in Lemma 3.3. Here, $R_{t,n}^{\mathrm{idx}}$ is uniform over folds and the declared policy class. Define

$$R_n^{\mathrm{idx}} = \sum\nolimits_{t=1}^{T} \gamma^{t-1} R_{t,n}^{\mathrm{idx}}. \tag{3.24}$$

Theorem 3.2 (Sequential orthogonality)**.** *Under Assumptions 2.1–2.4, at the true bridges the Gâteaux derivative of $\mathbb{E}_{P_b}\Psi_\pi$ is zero in every admissible bridge direction for which the perturbed score and the displayed products are integrable:*

$$\partial_\vartheta \mathbb{E}_{P_b}\{\Psi_\pi(O; \vartheta^0)\}[b] = 0. \tag{3.25}$$

If, in addition, Assumption 3.1 holds and the nuisance estimators are cross-fitted over independent sampling units, let $\widehat{\Psi}_\pi^{(-k)}$ denote the uncapped nested score constructed with nuisances trained on I_k^c. Conditional on each training fold and uniformly over the declared policy class,

$$\max_{1 \leq k \leq K} \sup_{\pi \in \mathbf{\Pi}} \left| P_b\{\widehat{\Psi}_\pi^{(-k)} - \Psi_\pi^0\} \right| \leq C_T \sum\nolimits_{t=1}^{T} \gamma^{t-1} r_{h,t} r_{q,t} + R_n^{\mathrm{idx}}, \tag{3.26}$$

where R_n^{idx} is the aggregate uniform remainder in (3.24)*.*

Proof. First perturb one stage while holding every other bridge at truth. Lemma 3.2 shows that the local mean drift is bilinear in the two stage-t perturbations. Hence its derivative in either an h_t direction or a q_t direction is zero at (h_t^0, q_t^0). A perturbation of a later bridge changes the input pseudo-outcome G_t^π, but the true stage-t operator transports its mean from P_t^π to P_{t+1}^π by Lemma 3.1. Backward induction from T to one therefore propagates a zero derivative through all preceding true operators. Summing directional perturbations over the finite collection of stages proves (3.25).

For the finite-error statement, fix a fold k, condition on I_k^c, and let Δ_t be the behavior-law mean drift of the estimated nested tail beginning at stage t. At stage t, add and subtract a recursively compatible population outcome bridge for the downstream pseudo-outcome currently being transported. Lemma 3.2 and Cauchy–Schwarz bound the same-stage bridge perturbation by $Cr_{h,t}r_{q,t}$. The true treatment bridge then transports the signed mean of the admissible downstream perturbation to P_{t+1}^π, while any mismatch between the recursively compatible bridge and the policy-indexed bridge family actually fitted is assigned to $R_{t,n}^{\mathrm{idx}}$ under Assumption 3.1. Consequently,

$$|\Delta_t| \le Cr_{h,t}r_{q,t} + \gamma|\Delta_{t+1}| + R_{t,n}^{\mathrm{idx}}, \qquad \Delta_{T+1} = 0. \tag{3.27}$$

The final remainder therefore records errors in the recursively estimated input G_t^π, the transported empirical functional, and uniform policy indexing; these terms are not assumed orthogonal. The policy-uniform L_2 norms defining $r_{h,t}$ and $r_{q,t}$ are given in (4.16)–(4.17). Iterating the recursion yields

$$|\Delta_1| \le C_T \sum_{t=1}^{T} \gamma^{t-1} r_{h,t}r_{q,t} + \sum_{t=1}^{T} \gamma^{t-1} R_{t,n}^{\mathrm{idx}}. \tag{3.28}$$

By definition (3.24), the last sum equals R_n^{idx}. Taking the policy supremum and then the maximum over folds yields (3.26). Cross-fitting ensures that each evaluation trajectory is scored with nuisance functions trained on different sampling units. □

Remark 3.2. The nested score is not obtained by replacing a latent continuation value with an observed-history regression. It repeatedly transports expectations from P_t^π to P_{t+1}^π. The direct bridge must be evaluated at the target action, and the augmentation must be restricted to observations matching that action. These details determine both double robustness and the cumulative weights that govern longitudinal instability.

4. Temporally adaptive cumulative capping

4.1. Capped global score

Define the odd, 1-Lipschitz winsorization map

$$C_\Lambda(x) = \mathrm{sign}(x)\min\{|x|, \Lambda\}, \qquad \Lambda \ge 1. \tag{4.1}$$

The stage-t cap is applied to both cumulative terms in the expanded score:

$$\begin{aligned}\Psi_\pi^{\Lambda}(O;\vartheta) = \sum_{t=1}^{T} \gamma^{t-1}\Big[& C_{\Lambda_t}(\eta_t^\pi) h_t^\pi(W_t, E_t^\pi, H_t) \\ & + C_{\Lambda_t}(\omega_t^\pi)\mathbf{1}(A_t = E_t^\pi)\{Y_t - h_t^\pi(W_t, A_t, H_t)\}\Big].\end{aligned} \tag{4.2}$$

No trajectory is deleted. A large cumulative correction is capped only in its stage contribution, while the trajectory remains available elsewhere. We use "capping" and "winsorization" synonymously for the map in (4.1). "Clipping" refers only to fixed numerical safeguards imposed on fitted propensities, bridge predictions, or final scores in the executable experiments. "Truncation" is reserved for deleting or zeroing observations/terms beyond a threshold; the main TAPLE score does not truncate whole trajectories.

Assumption 4.1 (Polynomial cumulative-weight tails)**.** Under the behavior law P_b, for constants $C_t < \infty$ and $\alpha_t > 0$, uniformly over $\pi \in \mathbf{\Pi}$ and $u \geq 1$,

$$\mathbb{P}(|\eta_t^{\pi,0}| > u) \leq C_t u^{-(1+\alpha_t)}, \tag{4.3}$$

$$\mathbb{P}(|\omega_t^{\pi,0}| > u) \leq C_t u^{-(1+\alpha_t)}. \tag{4.4}$$

Because $\alpha_t > 0$, these tail bounds guarantee finite first moments of the cumulative weights. Assumption 3.1 separately requires the weighted products used by the transport identities to be integrable. The tail condition does not require uniformly bounded second moments; for $\alpha_t \leq 1$, second moments may grow or diverge, which is precisely the regime in which capping changes the fluctuation rate. Let

$$\chi_\alpha(\Lambda) = \begin{cases} \Lambda^{1-\alpha}, & 0 < \alpha < 1, \\ 1 + \log \Lambda, & \alpha = 1, \\ 1, & \alpha > 1. \end{cases} \tag{4.5}$$

Proposition 4.1 (Global capping bias and second moments)**.** *Under Assumptions 2.4 and 4.1,*

$$\sup_{\pi \in \mathbf{\Pi}} |\mathbb{E}\{\Psi_\pi^{\mathbf{\Lambda}}(\vartheta^0) - \Psi_\pi(\vartheta^0)\}| \leq C_T \sum\nolimits_{t=1}^{T} \gamma^{t-1} \Lambda_t^{-\alpha_t}. \tag{4.6}$$

Moreover,

$$\mathbb{E}\{C_{\Lambda_t}^2(\eta_t^{\pi,0})\} \leq C \chi_{\alpha_t}(\Lambda_t), \tag{4.7}$$

$$\mathbb{E}\{C_{\Lambda_t}^2(\omega_t^{\pi,0})\} \leq C \chi_{\alpha_t}(\Lambda_t). \tag{4.8}$$

Proof. For any X satisfying $\mathbb{P}(|X| > u) \leq C u^{-(1+\alpha)}$,

$$\mathbb{E}(|X| - \Lambda)_+ = \int_\Lambda^\infty \mathbb{P}(|X| > u)\, du \leq C \alpha^{-1} \Lambda^{-\alpha}. \tag{4.9}$$

The stage reward and both bridge evaluations are bounded by Assumption 2.4. Applying (4.9) to $\eta_t^{\pi,0}$ and $\omega_t^{\pi,0}$ in (4.2), followed by the triangle inequality, proves (4.6). For the second moments,

$$\mathbb{E}\{C_\Lambda^2(X)\} = 2 \int_0^\Lambda u \mathbb{P}(|X| > u)\, du. \tag{4.10}$$

Splitting the integral at one yields the three cases in (4.5). □

4.2. Uniform policy fluctuation

Assumption 4.2 (Score entropy and fitted-score increments)**.** For each stage, the true-bridge capped score class is measurable, has an envelope of order $C_T\Lambda_t$, and satisfies a uniform Vapnik–Chervonenkis (VC)-subgraph or polynomial covering-number bound with complexity index v_t^{sc}. Conditional on each complementary training sample, the fitted-minus-true capped score increment class obeys the same order of covering complexity. The number of folds K is fixed, the folds are balanced over independent sampling units, and v_t^{sc} dominates the joint complexity of the policy components entering the stage-t cumulative weights and policy-indexed bridge, together with the bridge class. For fixed T, this has the same order as stagewise policy complexity when stage dimensions are comparable.

Let $\mathcal{G}_t$ be a finite, complementary-training-fold-measurable geometric cap grid with $M_t = |\mathcal{G}_t|$. Set

$$d_{n,t}(\delta) = v_t^{\mathrm{sc}} \log\left(\frac{en}{v_t^{\mathrm{sc}} \vee 1}\right) + \log\left(\frac{8KTM_t}{\delta}\right). \tag{4.11}$$

We suppress δ and write $d_{n,t}$ when unambiguous. For the rate comparison, the grid is assumed to contain a cap within a fixed multiplicative factor of the continuous oracle in (4.14); this changes only constants in the displayed rates. For the true bridges, let $\Psi_{\pi,t}^{\Lambda_t,0}$ denote the stage-t summand in (4.2) and define

$$\mathcal{F}_t^{\Lambda_t,0} = \{O \mapsto \Psi_{\pi,t}^{\Lambda_t}(O;\vartheta^0) : \pi \in \mathbf{\Pi}\}. \tag{4.12}$$

Lemma 4.1 (Uniform true-score deviation on an evaluation fold)**.** *For any evaluation fold I_k, conditional on a complementary-training-fold-measurable choice of cap grid, with probability at least $1-\delta$ simultaneously over all $\Lambda_t \in \mathcal{G}_t$,*

$$\begin{aligned}\sup_{\pi\in\mathbf{\Pi}} |(P_{n,k} - P)\Psi_\pi^{\mathbf{\Lambda}}(\vartheta^0)| \le C_T \sum_{t=1}^{T} \gamma^{t-1}\Bigg[&\sqrt{\frac{\chi_{\alpha_t}(\Lambda_t) d_{n,t}(\delta)}{n_k}} + \frac{\Lambda_t d_{n,t}(\delta)}{n_k} \\ &+ \sqrt{\frac{d_{n,t}(\delta)}{n_k}}\Bigg].\end{aligned} \tag{4.13}$$

Proof. The true-bridge capped direct and residual terms have envelopes of order $C(B_Y + B_h)\Lambda_t$. Proposition 4.1 controls their second moments under the behavior law. Conditional on the cap grid, symmetrization and Assumption 4.2 yield the square-root terms, while Bernstein's inequality contributes the linear envelope term. A union bound over stages and candidate caps proves the foldwise statement; the theorem below then unions over the fixed number of folds. The fitted-score increment is not covered by Proposition 4.1 and is controlled separately through (4.21). □

Balancing (4.6) and (4.13) gives

$$\Lambda_{t,n}^{\circ} \asymp \begin{cases} \{n/d_{n,t}\}^{1/(1+\alpha_t)}, & 0 < \alpha_t < 1, \\ \{n/[d_{n,t}\log n]\}^{1/2}, & \alpha_t = 1, \\ \{n/d_{n,t}\}^{1/(2\alpha_t)}, & \alpha_t > 1, \end{cases} \tag{4.14}$$

and

$$\rho_{t,n} = \begin{cases} \{d_{n,t}/n\}^{\alpha_t/(1+\alpha_t)}, & 0 < \alpha_t < 1, \\ \{d_{n,t}\log n/n\}^{1/2}, & \alpha_t = 1, \\ \{d_{n,t}/n\}^{1/2}, & \alpha_t > 1. \end{cases} \tag{4.15}$$

4.3. Capping and nuisance interaction

For a measurable perturbation g, define policy-uniform intervention-law norms

$$\begin{aligned} \|g\|_{h,t,2}^2 &= \sup_{\pi\in\mathbf{\Pi}} \mathbb{E}_{P_t^\pi}\Big[g^2(W_t, E_t^\pi, H_t) + \mathbf{1}(A_t = E_t^\pi)g^2(W_t, A_t, H_t)\Big], \\ \|g\|_{q,t,2}^2 &= \sup_{\pi\in\mathbf{\Pi}} \mathbb{E}_{P_t^\pi}\Big[\mathbf{1}(A_t = E_t^\pi)g^2(Z_t, A_t, H_t)\Big]. \end{aligned} \tag{4.16}$$

The cross-fitted nuisance rates are the maxima over folds,

$$r_{h,t} = \max_k \sup_{\pi\in\mathbf{\Pi}} \|\widehat{h}_t^{\pi,(-k)} - h_t^{\pi,0}\|_{h,t,2}, \qquad r_{q,t} = \max_k \|\widehat{q}_t^{(-k)} - q_t^0\|_{q,t,2}. \tag{4.17}$$

Capping removes exact global orthogonality. We therefore define the interaction remainder directly by

$$\zeta_{t,n}^{(-k)}(\Lambda_t) = \sup_{\pi\in\mathbf{\Pi}} \Big|\mathbb{E}[\{\widehat{\Psi}_{\pi,t}^{\Lambda_t,(-k)} - \Psi_{\pi,t}^{\Lambda_t,0}\} - \{\widehat{\Psi}_{\pi,t}^{(-k)} - \Psi_{\pi,t}^0\}]\Big|. \tag{4.18}$$

Where the subscript t denotes the stage-t summand and the expectation is over a fresh behavior-law trajectory, conditional on the complementary training sample I_k^c. The upper theorem treats $\max_k \zeta_{t,n}^{(-k)}$ as a separately auditable remainder.

Assumption 4.3 (Capping–nuisance interaction control)**.** Bridge perturbations are uniformly bounded on the declared sieve classes, the cumulative-product map is locally Lipschitz in the relevant L_2 neighborhoods, and, uniformly over $\Lambda_t \in \mathcal{G}_t$ and the declared policy class,

$$\max_k \zeta_{t,n}^{(-k)}(\Lambda_t) \le C_T r_{h,t} r_{q,t} + C_T(r_{h,t} + r_{q,t})\Lambda_t^{-\alpha_t/2} + R_{t,n}^{\mathrm{idx}}. \tag{4.19}$$

This is a high-level regularity condition to be verified for the chosen bridge learner; it is not implied by cross-fitting.

Cross-fitting also leaves a centered stochastic nuisance term. On evaluation fold k, write

$$\Delta_{\pi,t}^{\Lambda_t,(-k)} = \widehat{\Psi}_{\pi,t}^{\Lambda_t,(-k)} - \Psi_{\pi,t}^{\Lambda_t,0} \tag{4.20}$$

and define the conditional empirical-process radius

$$\xi_{t,n}^{(-k)}(\Lambda_t) = \sup_{\pi\in\mathbf{\Pi}} \|\Delta_{\pi,t}^{\Lambda_t,(-k)}\|_{P,2} \sqrt{\frac{d_{n,t}(\delta)}{n_k}} + \sup_{\pi\in\mathbf{\Pi}} \|\Delta_{\pi,t}^{\Lambda_t,(-k)}\|_{P,\infty} \frac{d_{n,t}(\delta)}{n_k}. \tag{4.21}$$

Both P-norms are taken under the behavior law for a fresh trajectory and are conditional on the complementary training sample I_k^c. The second, Bernstein-envelope term is required for a high-probability bound unless a sharper localized empirical-process condition is imposed. In cross-fitted statements we take the maximum over k. This radius is distinct from the mean interaction $\zeta_{t,n}^{(-k)}$: cross-fitting removes own-observation bias but does not make the centered nuisance fluctuation vanish.

In a canonical mildly ill-posed sieve model with singular values $\sigma_j \asymp j^{-\nu_t}$, smoothness β_t, effective bridge-domain dimension d_t^{br}, and an L_2 loss matched to the operator, one representative bridge rate is

$$r_{h,t} \vee r_{q,t} = O_p\left(n^{-\beta_t/(2\beta_t+2\nu_t+d_t^{\mathrm{br}})} + a_{J,t}\right), \tag{4.22}$$

where $a_{J,t}$ is sieve approximation error. Equation (4.22) is illustrative and model specific; neither it nor Assumption 4.3 follows automatically from cross-fitting. When $\alpha_t \leq 1$, the true treatment bridge or cumulative correction can fail to have a finite second moment. In that regime, the full-L_2 nuisance rate in (4.17) is an additional bridge-learnability restriction (or is trivially satisfied for a known/exact bridge), not a consequence of Assumption 4.1. A theory based entirely on localized or capped nuisance norms would be needed to cover arbitrary infinite-variance bridge estimation.

For a concrete finite-dimensional connection to the ridge bridge fits used in the executable benchmark, suppose a stage-t bridge is represented by d_t bounded features, the population conditional-moment matrix has smallest singular value at least $\kappa_t > 0$, and ridge regularization λ_t is applied on a complementary fold. Under standard sub-Gaussian feature/noise concentration and correct finite-dimensional specification, the conditional-moment solution obeys the representative bound

$$r_{h,t} \vee r_{q,t} = O_p\left[\kappa_t^{-1}\sqrt{\frac{d_t + \log(1/\delta)}{n}} + \frac{\lambda_t}{\kappa_t}\right]. \tag{4.23}$$

Hence fixed d_t, $\inf_t \kappa_t > 0$, and $\lambda_t = O(n^{-1/2})$ give the parametric $n^{-1/2}$ order. When κ_t is small, the bound deteriorates, making singular-value and residual diagnostics directly relevant to Assumptions 2.3–4.3. Equation (4.23) is a sufficient benchmark, not a claim that an ill-posed nonparametric bridge automatically attains a parametric rate.

Remark 4.1. Capping cumulative weights is more conservative than capping isolated bridge factors, but it targets the actual longitudinal envelope. The price is explicit: truncation bias, loss of exact orthogonality, and a policy-indexing remainder. A minimax claim is therefore conditional on uniform bridge learning and on numerical optimization being accurate enough. Reporting these terms is preferable to hiding them inside an unspecified constant.

5. The Taple procedure

Figure 2 separates identification, regularization, and optimization in the proposed data-driven decision pipeline. Independent sampling units are first assigned to folds. Treatment bridges are estimated from behavior-law moments, while policy-indexed outcome bridges use recursively transported training functionals. The resulting cumulative correction weights determine stage-specific caps and diagnostics. Candidate policies are then compared through the global cross-fitted value, with any unresolved search gap retained as ϵ_{opt}. The supplied experiment code executes the factorized Taple-F analogue, not the full cumulative-score or general carryover solver represented by the pipeline.

Figure 2. Cross-fitted proximal learning and policy optimization pipeline.

5.1. Cross-fitting over independent sampling units

Split independent sampling units into folds $I_1, \dots, I_K$. For independent trajectories, a unit is one trajectory. When multiple generated trajectories share a source template, all descendants of that source must remain in the same fold. For $i \in I_k$, every nuisance bridge and cap entering the score of O_i is fitted on I_k^c. Candidate policies are common across folds and may be selected using the aggregate cross-fitted objective in (5.5); their data dependence is handled by the uniform policy-class bounds rather than by fitting a separate policy on each complementary fold. Let

$$\widehat{q}_t^{(-k)} = \mathcal{A}_{q,t}(I_k^c) \tag{5.1}$$

and, for a current policy π and nested outcome $\widehat{G}_t^{\pi,(-k)}$,

$$\widehat{h}_t^{\pi,(-k)} = \mathcal{A}_{h,t}(I_k^c, \widehat{G}_t^{\pi,(-k)}). \tag{5.2}$$

The policy-indexed outcome bridge at stage t is not fitted as if observations were sampled directly from P_t^π. Instead, earlier valid treatment bridges induce the transported training functional

$$\widehat{P}_{t,-k}^{\pi} f = \frac{1}{|I_k^c|} \sum\nolimits_{i \in I_k^c} \widehat{\eta}_{it}^{\pi,(-k)} f(O_i). \tag{5.3}$$

For bounded stage-t test functions, or more generally for functions whose successively weighted products are integrable within the declared bridge moment class, Lemma 3.3 gives the corresponding population identity when the preceding treatment bridges are valid. Consequently, this functional estimates the relevant P_t^π moments used in (3.22). It is signed and generally unnormalized, so it is not a probability measure and must not be used as a generic estimator of arbitrary intervention-law expectations. Regularized conditional-moment fitting for h_t^π uses only these admissible transported moments. The policy-stable treatment bridge is estimated once per fold from the behavior-law equation (2.14). Assumption 2.4 supplies the required support inclusion and invariance for reusing that behavior-law solution under P_t^π. This recursive weighting step is necessary in general carryover settings.

Let $n_k = |I_k|$ and $P_{n,k}$ be the empirical measure on fold I_k. The fold score is

$$\widehat{\Psi}_{i,\pi}^{(-k)} = \Psi_\pi^{\widehat{\mathbf{\Lambda}}^{(-k)}}(O_i; \widehat{\vartheta}^{(-k)}), \qquad i \in I_k. \tag{5.4}$$

The actual cross-fitted empirical value, allowing the selected cap vector to differ across folds, is

$$\widehat{V}_{\mathrm{cf}}(\pi) = \sum\nolimits_{k=1}^{K} \frac{n_k}{n} P_{n,k} \Psi_\pi^{\widehat{\mathbf{\Lambda}}^{(-k)}}(\widehat{\vartheta}^{(-k)}). \tag{5.5}$$

When a common deterministic cap vector is analyzed, the shorter notation $\widehat{V}_{\mathbf{\Lambda}}$ is used.

5.2. Temporally calibrated caps

On each training fold, let $\widehat{\eta}_{it}^\pi$ and $\widehat{\omega}_{it}^\pi$ denote the estimated cumulative direct and correction weights. Use the finite geometric grid $\mathcal{G}_t$ introduced above and define the fitted stage contribution

$$\widehat{S}_{it}^{\pi}(\Lambda) = C_\Lambda(\widehat{\eta}_{it}^\pi)\widehat{h}_t^\pi(W_{it}, E_{it}^\pi, H_{it}) + C_\Lambda(\widehat{\omega}_{it}^\pi)\mathbf{1}(A_{it} = E_{it}^\pi)\widehat{\varepsilon}_{it}^\pi, \tag{5.6}$$

where $\widehat{\varepsilon}_{it}^{\pi} = \widehat{G}_{it}^{\pi} - \widehat{h}_t^{\pi}(W_{it}, A_{it}, H_{it})$. Its uncapped version is $\widehat{S}_{it}^{\pi}(\infty)$. For a computational pilot class $\mathbf{\Pi}_0$, set

$$\widehat{B}_t(\Lambda) = \sup_{\pi \in \mathbf{\Pi}_0} \left| P_{-k}\{\widehat{S}_t^{\pi}(\infty) - \widehat{S}_t^{\pi}(\Lambda)\} \right| \tag{5.7}$$

and

$$\widehat{\mathcal{V}}_t(\Lambda) = \sup_{\pi \in \mathbf{\Pi}_0} \mathrm{Var}_{-k}\{\widehat{S}_t^{\pi}(\Lambda)\}. \tag{5.8}$$

Select the fold-specific cap

$$\widehat{\Lambda}_t^{(-k)} = \arg\min_{\Lambda \in \mathcal{G}_t} \left[\widehat{B}_t(\Lambda) + c_{\mathrm{cap}} \sqrt{\frac{\widehat{\mathcal{V}}_t(\Lambda) \log n}{|I_k^c|}} \right], \tag{5.9}$$

where P_{-k} and Var_{-k} denote the empirical mean and variance on I_k^c, and $c_{\mathrm{cap}} > 0$ is a fixed tuning constant chosen before evaluating the held-out fold. This criterion calibrates both capped terms in (4.2); calibrating only the residual correction would leave the cumulative direct term uncontrolled.

Remark 5.1 (Scope of the adaptive guarantee). The grid criterion (5.9) is implementable, but its finite-sample oracle behavior is not implied by cross-fitting alone. The regret result below is therefore conditional on the explicit calibration inequality (6.13). Establishing that inequality for a particular bridge learner, pilot policy class, and grid is a separate statistical task; without it, the method remains a data-adaptive stabilized estimator but is not claimed to be unconditionally minimax adaptive.

Proposition 5.1 (A finite-grid sufficient condition for cap calibration)**.** *Fix a stage and a training fold. Let $B_t(\Lambda)$ and $\mathcal{V}_t(\Lambda)$ denote the population analogues of* (5.7) *and* (5.8) *on the pilot class. Suppose that, with probability at least $1 - \delta$, simultaneously over $\Lambda \in \mathcal{G}_t$,*

$$|\widehat{B}_t(\Lambda) - B_t(\Lambda)| \le b_{t,n}, \qquad |\sqrt{\widehat{\mathcal{V}}_t(\Lambda)} - \sqrt{\mathcal{V}_t(\Lambda)}| \le v_{t,n}, \tag{5.10}$$

and the worst capping bias over $\mathbf{\Pi}$ exceeds that over $\mathbf{\Pi}_0$ by at most $a_{t,n}$. Then the selected cap satisfies an oracle inequality for the bias-plus-standard-error criterion with additive remainder

$$2b_{t,n} + 2c_{\mathrm{cap}} v_{t,n} \sqrt{\frac{\log n}{|I_k^c|}} + a_{t,n}. \tag{5.11}$$

Consequently, the calibration condition used by the regret theorem follows whenever the finite-grid deviation terms and pilot-class approximation error are of the same or smaller order than $\rho_{t,n}$.

Proof. Apply the defining optimality of $\widehat{\Lambda}_t^{(-k)}$ to the empirical criterion, add and subtract the corresponding population criterion at the selected cap and at its grid oracle, and use (5.10) twice. The pilot-class approximation contributes $a_{t,n}$. □

Let $k(i)$ denote the fold containing trajectory i, and evaluate the diagnostic at the fitted policy. Define the matched capped correction weight

$$\widetilde{\omega}_{it} = \mathbf{1}(A_{it} = E_{it}^{\widehat{\pi}}) C_{\widehat{\Lambda}_t^{(-k(i))}}\left(\widehat{\omega}_{it}^{\widehat{\pi},(-k(i))}\right). \tag{5.12}$$

For score-concentration auditing, define the normalized policy-matched absolute-weight effective-sample-size (ESS) diagnostic

$$N_{t,\mathrm{eff}} = \frac{(\sum_i |\widetilde{\omega}_{it}|)^2}{n \sum_i \widetilde{\omega}_{it}^2}, \tag{5.13}$$

with $N_{t,\mathrm{eff}} = 0$ when all matched capped correction weights are zero. Thus $N_{t,\mathrm{eff}} \in [0, 1]$. Because proximal bridge corrections may be signed, $N_{t,\mathrm{eff}}$ is a concentration diagnostic based on weight magnitudes, not a design-based effective sample size or an estimate of the number of independent trajectories. The executable factorized benchmark reports a different, explicitly local analogue based on the observed-action bridge corrections q_{A_i} after local capping. That TAPLE-F quantity is useful for score-concentration auditing but is not an estimate of the general cumulative policy-matched $N_{t,\mathrm{eff}}$.

5.3. Policy optimization

The statistical target is the supremum of the global cross-fitted empirical value over $\mathbf{\Pi}$, but the returned policy $\widehat{\pi} \in \mathbf{\Pi}$ need not attain that supremum. Define its empirical optimization error by

$$\epsilon_{\mathrm{opt}} = \sup_{\pi \in \mathbf{\Pi}} \widehat{V}_{\mathrm{cf}}(\pi) - \widehat{V}_{\mathrm{cf}}(\widehat{\pi}) \geq 0. \tag{5.14}$$

Thus, an exact empirical maximizer has $\epsilon_{\mathrm{opt}} = 0$, while an approximate search is covered without redefining the estimator. A general implementation may use backward coordinate optimization: update π_T, then π_{T-1}, and continue to π_1 while recomputing the nested outcomes and score. Accept only nondecreasing coordinate sweeps and repeat until, for a prespecified tolerance $\tau_{\mathrm{opt}} \geq 0$,

$$0 \leq \widehat{V}_{\mathrm{cf}}(\widehat{\pi}^{(m+1)}) - \widehat{V}_{\mathrm{cf}}(\widehat{\pi}^{(m)}) \leq \tau_{\mathrm{opt}}. \tag{5.15}$$

If a proposed sweep decreases the score, it is rejected or damped before applying this rule; a negative improvement is therefore not a convergence certificate. The theorem below retains ϵ_{opt} rather than assuming the nonconvex search is exact. In the supplied experiments, actions do not affect later state evolution; the value objective therefore factorizes. The implemented local capped score is valid for each factor but is not algebraically identical to capping the cumulative products in (4.2).

The constraints in the optimization problem are therefore explicit rather than implicit: candidate policies must remain in $\mathbf{\Pi}$; actions satisfy $\pi_t(H_t) \in \{0, 1\}$; cap variables are restricted to the prespecified finite grids $\mathcal{G}_t \subset [1, \Lambda_{t,\max}]$; and any parameterized policy update is projected or rejected if it leaves its declared feasible set. The cap lower bound prevents a degenerate zero score, while the finite upper bound supplies a finite envelope for optimization and concentration. In a general constrained application, feasibility constraints belong in the definition of $\mathbf{\Pi}$ and the policy-update subproblem, not in the proximal identification equations.

Proposition 5.2 (Factorized experimental submodel)**.** *Under Assumptions 2.1–2.4, suppose in addition that, for every stage, the joint law of the policy-relevant pre-action state H_t and latent state U_t is invariant to past policy actions, the conditional law of future states and proxies is invariant to the current action, and the conditional mean reward has the additive form*

$$\mathbb{E}\{Y_t(a) \mid H_t, U_t\} = m_t(H_t, U_t) + a\Delta_t(H_t), \tag{5.16}$$

and the policy class is a Cartesian product of stagewise classes. Then

$$V(\pi) - V(\widetilde{\pi}) = \sum_{t=1}^{T} \gamma^{t-1} \mathbb{E}[\{\pi_t(H_t) - \widetilde{\pi}_t(H_t)\}\Delta_t(H_t)], \tag{5.17}$$

so policy optimization separates by stage. Each stage value is identified by the one-step proximal operator in Lemma 3.1. The executable Taple-F *method caps the local one-stage correction in each separated factor. It is therefore a valid robust estimator for this submodel, but it is not a pointwise implementation of the full cumulative capped score in* (4.2)*.*

Proof. The stated policy-state invariance makes the distribution of (H_t, U_t) common to all admissible policies; this extra condition is needed because a generic history containing past actions would otherwise change distribution when the policy changes. Subtracting the two additive reward means and summing the discounted differences gives (5.17). Lemma 3.1 identifies each stage mean, and Cartesian-product optimization separates the maximization across t. The final distinction follows because local winsorization does not generally commute with cumulative multiplication. □

Algorithm 1 Temporally adaptive proximal policy learning.

Split independent sampling units into K folds, grouping shared-source descendants.
Estimate the policy-stable treatment bridges on each complementary fold.
Initialize a policy with an observable-history doubly robust learner.
repeat
  for $t = T, T-1, \ldots, 1$ **do**
    Construct nested outcomes and the transported functional in (5.3).
    Estimate outcome bridges and cumulative correction weights.
    Select $\widehat{\Lambda}_t^{(-k)}$ using (5.9).
    Update π_t to increase the global cross-fitted score.
  end for
until an accepted sweep satisfies (5.15)
Return $\widehat{\pi}$, caps, bridge residuals, and the declared policy-matched or local ESS diagnostics.

With sieve dimension J_t, K folds, and M policy sweeps, the bridge-fitting component of a direct linear implementation costs

$$O\left(MK \sum_{t=1}^{T} (nJ_t^2 + J_t^3)\right). \tag{5.18}$$

Let C_{pol} denote the additional cost of the chosen policy-update/search routine; it may depend on M, n, and $\mathbf{\Pi}$. The total computational cost is the bridge-fitting order in (5.18) plus C_{pol}, and generic policy search can dominate. Excluding optimizer-specific state, the bridge-fitting memory requirement is

$$O\left(n \sum_{t=1}^{T} J_t + \sum_{t=1}^{T} J_t^2\right). \tag{5.19}$$

Remark 5.2. The theorem concerns the global capped empirical value and retains the optimization gap ϵ_{opt}. This distinction is material. The executable benchmark does not implement the general backward coordinate optimizer or compute a certificate for ϵ_{opt}: after the factorized score is constructed, it fits a stagewise ridge least-squares regression to the cross-fitted pseudo-contrast and thresholds the fitted

contrast at zero. Thus, factorization makes the population value separable, but the supplied policy fit remains a surrogate rather than an exact empirical maximizer over a general policy class. The experiments therefore test the proximal score and local capping mechanism, not the global optimization guarantee. For carryover dynamics, both statistical and optimization errors must be assessed.

6. Minimax regret theory

6.1. Model class and oracle rate

Let $\mathcal{P}(\boldsymbol{\alpha}, \mathbf{v}^{\mathrm{pol}})$ be a fixed-horizon class satisfying Assumptions 2.1–2.4 and 4.1, with stagewise policy VC dimensions v_t^{pol}. Define

$$\mathfrak{M}_n = \inf_{\widehat{\pi}} \sup_{P \in \mathcal{P}} \mathbb{E}_P \mathrm{Reg}_P(\widehat{\pi}). \tag{6.1}$$

For stage t, set

$$\underline{\rho}_{t,n} = \begin{cases} \{v_t^{\mathrm{pol}}/n\}^{\alpha_t/(1+\alpha_t)}, & 0 < \alpha_t < 1, \\ \{v_t^{\mathrm{pol}}/n\}^{1/2}, & \alpha_t \geq 1. \end{cases} \tag{6.2}$$

For the one-stage construction, $\underline{\rho}_n(\alpha, v^{\mathrm{pol}})$ denotes the same expression with the stage index suppressed, and $\mathcal{P}(\alpha, v^{\mathrm{pol}})$ denotes the corresponding one-stage subclass. Constants below are uniform for $1 \leq v^{\mathrm{pol}} \leq c_0 n$ with c_0 sufficiently small.

6.2. One-stage lower bound

Let $m = \lfloor v^{\mathrm{pol}} \rfloor$ and choose m points shattered by the policy class. In the heavy-correction-weight submodel, $X \in \{0, 1, \ldots, m\}$, the union of rare cells has probability

$$\mathbb{P}(X \in [m]) = p = \varepsilon^{\alpha}, \qquad \mathbb{P}(X = j) = p/m, \quad j \in [m]. \tag{6.3}$$

On every rare cell,

$$\mathbb{P}(A = 1 \mid X = j) = \varepsilon, \qquad q^0(1, j) = \varepsilon^{-1}. \tag{6.4}$$

Set $\mathbb{P}(A = 1 \mid X = 0) = 1/2$ outside the rare cells and restrict the construction to $\varepsilon \leq 1/4$. Thus, all nonrare correction weights are at most two, whereas the action-one correction on a rare cell is ε^{-1}. For the target action one, the realized correction weight therefore obeys

$$\mathbb{P}\{|\omega_1^0| > (2\varepsilon)^{-1}\} = p\varepsilon = \varepsilon^{1+\alpha}. \tag{6.5}$$

For $2 < u < \varepsilon^{-1}$, this probability remains $p\varepsilon = \varepsilon^{1+\alpha} \leq u^{-(1+\alpha)}$; for $1 \leq u \leq 2$ it is bounded by one, and for $u \geq \varepsilon^{-1}$ it is zero. Hence Assumption 4.1 holds uniformly in ε after adjusting the tail constant. Take $H = X$, let U be degenerate, and let W and Z be independent ancillary variables. Restrict the bridge classes in this submodel to functions of (a, H). Then $q^0(a, H) = \mathbb{P}(A = a \mid H)^{-1}$, the negative-control restrictions hold, and completeness is immediate on the restricted observable bridge class. Hence the construction is contained in the larger proximal class. The logging law, and therefore q^0, may even be revealed to the learner in this submodel; doing so can only make the policy-learning problem easier, so the lower bound below is not driven by treatment-bridge estimation error. Let the reward under $A = 0$ have Bernoulli mean $1/2$. On rare cell j, let the reward under $A = 1$ have Bernoulli mean $1/2 + \theta_j \delta/2$, where $\theta_j \in \{-1, 1\}$, and make the reward laws identical across alternatives outside the rare cells.

Theorem 6.1 (Minimax lower bound under polynomial correction-weight tails)**.** *For a policy class with VC dimension v^{pol} satisfying $1 \le v^{\mathrm{pol}} \le c_0 n$, there is $c > 0$ such that*

$$\inf_{\widehat{\pi}} \sup_{P \in \mathcal{P}(\alpha, v^{\mathrm{pol}})} \mathbb{E}_P \mathrm{Reg}_P(\widehat{\pi}) \ge c\, \underline{\rho}_n(\alpha, v^{\mathrm{pol}}). \tag{6.6}$$

Proof. First suppose $0 < \alpha < 1$ and use the rare-cell hypercube above. Two vertices differing only in coordinate j differ in their observed reward law only when $X = j$ and $A = 1$. Since Bernoulli means remain bounded away from zero and one,

$$\mathrm{KL}(P_\theta^{\otimes n}, P_{\theta^{(j)}}^{\otimes n}) \le Cn \frac{p}{m} \varepsilon \delta^2 = Cn \frac{\varepsilon^{1+\alpha}}{m} \delta^2. \tag{6.7}$$

Choose

$$\varepsilon = c_1 \left(\frac{m}{n}\right)^{1/(1+\alpha)}, \qquad \delta = c_2, \tag{6.8}$$

with fixed sufficiently small constants. Then every adjacent divergence is bounded by an absolute constant. A wrong sign on cell j incurs regret $p\delta/(2m)$ under the Bernoulli means specified above. Assouad's lemma gives a constant expected fraction of sign errors, and hence

$$\inf_{\widehat{\pi}} \sup_{\theta} \mathbb{E}\mathrm{Reg}(\widehat{\pi}) \ge c_3 p \delta \asymp \left(\frac{m}{n}\right)^{\alpha/(1+\alpha)}. \tag{6.9}$$

Because m is comparable to v^{pol}, this proves the first case of (6.6).

For $\alpha \ge 1$, use a bounded-overlap hypercube on the same m shattered points, assign probability $1/m$ to each point, and let the logging action be Bernoulli$(1/2)$. On cell j, set the Bernoulli reward mean under action zero to $1/2$ and under action one to $1/2 + \theta_j \delta$, with $\delta \le 1/4$. The treatment contrast is therefore $\theta_j \delta$. Adjacent n-sample divergences are at most $Cn\delta^2/m$, while a wrong sign costs δ/m. Taking $\delta = c_4 \sqrt{m/n}$ and applying Assouad's lemma yields

$$\inf_{\widehat{\pi}} \sup_{\theta} \mathbb{E}\mathrm{Reg}(\widehat{\pi}) \ge c_5 \sqrt{m/n}. \tag{6.10}$$

The correction weights in this submodel are bounded, so Assumption 4.1 holds for every $\alpha > 0$. This proves the second case. □

6.3. Longitudinal reduction

Corollary 6.1 (Fixed-horizon lower bound)**.** *For fixed T,*

$$\mathfrak{M}_n \ge c \max_{1 \le t \le T} \gamma^{t-1} \underline{\rho}_{t,n}. \tag{6.11}$$

Proof. Embed the one-stage hard experiment at stage t, make all other rewards deterministic, and make all transitions independent of its sign. Any longitudinal estimator then induces a one-stage estimator. The value loss is multiplied by γ^{t-1}. Apply Theorem 6.1 and maximize over t. □

6.4. Oracle-calibrated adaptive upper bound

For the upper-bound statement, define the theoretical envelopes

$$\begin{aligned} \mathrm{Bias}_t(\Lambda) &= C_T \Lambda^{-\alpha_t}, \\ \mathrm{Dev}_t(\Lambda) &= C_T \left[\sqrt{\frac{\chi_{\alpha_t}(\Lambda) d_{n,t}(\delta)}{n}} + \frac{\Lambda d_{n,t}(\delta)}{n} + \sqrt{\frac{d_{n,t}(\delta)}{n}} \right]. \end{aligned} \tag{6.12}$$

Assume the following high-level calibration inequality holds almost surely, simultaneously for every fold k and stage t,

$$\begin{aligned} &\mathrm{Bias}_t(\widehat{\Lambda}_t^{(-k)}) + \mathrm{Dev}_t(\widehat{\Lambda}_t^{(-k)}) \\ &\quad \le C \inf_{\Lambda \in \mathcal{G}_t} \{\mathrm{Bias}_t(\Lambda) + \mathrm{Dev}_t(\Lambda)\} + C \sqrt{\frac{\log(KM_tT/\delta)}{n}}. \end{aligned} \tag{6.13}$$

Balanced fixed-K folds make replacing fold sizes by n immaterial up to constants. If (6.13) fails with probability δ_{cap}, add δ_{cap} to the failure probability in Theorem 6.2. Under Assumption 4.3, define the fold-robust nuisance remainder

$$\begin{aligned} \mathcal{R}_{t,n}^{\mathrm{cf}} &= r_{h,t} r_{q,t} + R_{t,n}^{\mathrm{idx}} \\ &\quad + \max_{1 \le k \le K} \left\{ (r_{h,t} + r_{q,t}) (\widehat{\Lambda}_t^{(-k)})^{-\alpha_t/2} + \xi_{t,n}^{(-k)} (\widehat{\Lambda}_t^{(-k)}) \right\}, \end{aligned} \tag{6.14}$$

where $\xi_{t,n}^{(-k)}$ is the conditional centered nuisance fluctuation on evaluation fold k; the maxima in (4.17) already define $r_{h,t}$ and $r_{q,t}$ uniformly over folds.

Theorem 6.2 (Oracle-calibrated longitudinal regret upper bound)**.** *Under Assumptions 2.1–2.4, 3.1, and 4.1–4.3, balanced fixed-K cross-fitting over independent sampling units, and* (6.13)*, with probability at least* $1 - \delta$*,*

$$\mathrm{Reg}(\widehat{\pi}_{\mathrm{TAPLE}}) \le C_T \sum_{t=1}^{T} \gamma^{t-1} \left[\rho_{t,n} + \mathcal{R}_{t,n}^{\mathrm{cf}} + \sqrt{\frac{\log(KM_tT/\delta)}{n}} \right] + \epsilon_{\mathrm{opt}}. \tag{6.15}$$

If $\mathcal{R}_{t,n}^{\mathrm{cf}} = O(\rho_{t,n})$*,* $\epsilon_{\mathrm{opt}} = O(\max_t \rho_{t,n})$*, and the score-class complexity scale* v_t^{sc} *is comparable to the corresponding policy-class complexity scale, the upper bound matches* (6.11) *in overlap exponent and policy-complexity order up to logarithmic and fixed-horizon factors.*

Proof. For fold-specific fitted caps, define the corresponding fold-averaged population functional

$$\overline{V}_{\mathrm{cf}}(\pi) = \sum_{k=1}^{K} \frac{n_k}{n} P \Psi_\pi^{\widehat{\mathbf{\Lambda}}^{(-k)}}(\vartheta^0), \qquad \widehat{V}_{\mathrm{cf}}(\pi) = \sum_{k=1}^{K} \frac{n_k}{n} P_{n,k} \Psi_\pi^{\widehat{\mathbf{\Lambda}}^{(-k)}}(\widehat{\vartheta}^{(-k)}). \tag{6.16}$$

The population expectation P in the kth term is evaluated at a fresh behavior-law trajectory while holding the training-fold fit fixed. Apply the empirical-process argument separately to each fold k, conditioning only on I_k^c. The fitted bridges and selected caps are then fixed and independent of the trajectories in I_k. Uniformity over $\mathbf{\Pi}$ permits the final policy to be data dependent. The empirical near-maximization property and a telescoping decomposition give

$$\begin{aligned} V(\pi^\star) - V(\widehat{\pi}) \le\ & 2 \sup_{\pi \in \mathbf{\Pi}} |\widehat{V}_{\mathrm{cf}}(\pi) - \overline{V}_{\mathrm{cf}}(\pi)| + \epsilon_{\mathrm{opt}} \\ & + |V(\pi^\star) - \overline{V}_{\mathrm{cf}}(\pi^\star)| + |V(\widehat{\pi}) - \overline{V}_{\mathrm{cf}}(\widehat{\pi})|. \end{aligned} \tag{6.17}$$

For each fold and policy, decompose its contribution to the first supremum into

$$\begin{aligned}(P_{n,k}-P)\Psi_{\pi}^{\widehat{\mathbf{\Lambda}}^{(-k)}}(\vartheta^0)+(P_{n,k}-P)\{\Psi_{\pi}^{\widehat{\mathbf{\Lambda}}^{(-k)}}(\widehat{\vartheta}^{(-k)})-\Psi_{\pi}^{\widehat{\mathbf{\Lambda}}^{(-k)}}(\vartheta^0)\}\\ +P\{\Psi_{\pi}^{\widehat{\mathbf{\Lambda}}^{(-k)}}(\widehat{\vartheta}^{(-k)})-\Psi_{\pi}^{\widehat{\mathbf{\Lambda}}^{(-k)}}(\vartheta^0)\}.\end{aligned} \tag{6.18}$$

Lemma 4.1 controls the first term for every candidate cap. Conditional VC/Bernstein control under Assumption 4.2 bounds the second term by $C_T\sum_t\gamma^{t-1}\max_k\xi_{t,n}^{(-k)}(\widehat{\Lambda}_t^{(-k)})$; the $L_\infty d/n_k$ component in (4.21) prevents the fitted-score increment from being treated as if an L_2 radius alone implied a high-probability bound. Sequential orthogonality, (3.26), and the capping interaction definition control the third term uniformly over folds by $C_T\sum_t\gamma^{t-1}\{r_{h,t}r_{q,t}+\max_k\zeta_{t,n}^{(-k)}(\widehat{\Lambda}_t^{(-k)})+R_{t,n}^{\mathrm{idx}}\}$. Assumption 4.3 converts this expression, up to a fixed constant, into the fold-robust remainder in (6.14). A union bound over the K folds, T stages, and M_t candidate caps is legitimate because each cap is selected from its training fold before evaluation.

Proposition 4.1 and the oracle inequality (6.13) control the two capping-bias terms in (6.17). Combining these bounds with the oracle balance in (4.14) yields

$$C_T\sum\nolimits_t\gamma^{t-1}\left[\rho_{t,n}+\mathcal{R}_{t,n}^{\mathrm{cf}}+\sqrt{\frac{\log(KM_tT/\delta)}{n}}\right]. \tag{6.19}$$

Substitution into (6.17) proves (6.15). For fixed T, a stagewise sum is at most T times its maximum; therefore the upper and lower rates agree up to logarithmic factors and the displayed nuisance and optimization remainders. □

Corollary 6.2 (Regular-overlap limit)**.** *If cumulative weights are uniformly bounded and* $\mathcal{R}_{t,n}^{\mathrm{cf}}=O\{(v_t^{\mathrm{sc}}\log n/n)^{1/2}\}$*, then*

$$\mathrm{Reg}(\widehat{\pi}_{\mathrm{TAPLE}})=O_p\left(\sum\nolimits_{t=1}^{T}\gamma^{t-1}\sqrt{\frac{v_t^{\mathrm{sc}}\log n}{n}}\right)+\epsilon_{\mathrm{opt}}. \tag{6.20}$$

Corollary 6.3 (Sufficient inverse-problem condition)**.** *If* (4.22) *holds uniformly over* $\mathbf{\Pi}$*,* $a_{J,t}=0$*, and* (4.19) *is valid, a sufficient condition for overlap-rate adaptation is*

$$\begin{aligned}n^{-2\beta_t/(2\beta_t+2\nu_t+d_t^{\mathrm{br}})}+\max_k n^{-\beta_t/(2\beta_t+2\nu_t+d_t^{\mathrm{br}})}(\widehat{\Lambda}_t^{(-k)})^{-\alpha_t/2}\\ +\max_k\xi_{t,n}^{(-k)}(\widehat{\Lambda}_t^{(-k)})+R_{t,n}^{\mathrm{idx}}=O(\rho_{t,n}).\end{aligned} \tag{6.21}$$

The centered nuisance term in (6.21) is essential: bridge L_2 rates alone do not control the capped-score fluctuation. If $\widehat{\Lambda}_t^{(-k)}\asymp\Lambda_{t,n}^{\circ}$ uniformly and $\xi_{t,n}^{(-k)}$ is stable on that scale, the selected-cap terms may instead be checked at $\Lambda_{t,n}^{\circ}$.

Remark 6.1. The lower bound is deliberately a maximum, not an unsupported sum. The upper bound is a sum because errors can accumulate across a fixed horizon; the two are equivalent only up to a T-dependent factor. The theorem also separates statistical and computational difficulty. A method that estimates bridges accurately but fails to optimize the global nested score does not inherit the minimax guarantee.

7. Numerical experiments

7.1. Controlled longitudinal benchmark

The controlled benchmark has $T = 6$ stages and action-invariant state transitions. The executable policy state is $H_t = (X_{1t}, X_{2t}, r_t/(T-1))$, with the current proxies Z_t and W_t excluded from the policy input. This choice makes the observed-history optimal action known exactly, so regret is not estimated by the same nuisance model used for training. The manuscript indexes stages by $t = 1, \ldots, T$, whereas the executable script uses the zero-based index $r_t = t - 1 \in \{0, \ldots, T-1\}$. To make the data-generating process identical in the paper and code, all periodic and drift terms below use r_t. Initially, $U_1 \sim N(0, 1)$, $X_{1,1} = 0.65U_1 + \varepsilon_1^{x0}$ with $\varepsilon_1^{x0} \sim N(0, 0.9^2)$, and $X_{2,1} \sim N(0, 1)$. The latent state evolves as

$$U_{t+1} = 0.72U_t + \varepsilon_{t+1}^U, \tag{7.1}$$

while observed states satisfy

$$X_{1,t+1} = 0.62X_{1t} + 0.30U_t + \varepsilon_{t+1}^1, \tag{7.2}$$
$$X_{2,t+1} = 0.57X_{2t} + 0.12U_t + \varepsilon_{t+1}^2. \tag{7.3}$$

The proxies are

$$Z_t = U_t + \varepsilon_t^Z, \tag{7.4}$$
$$W_t = 0.80U_t + 0.20X_{1t} + \varepsilon_t^W. \tag{7.5}$$

The logging law is

$$\mathbb{P}(A_t = 1 \mid H_t, U_t) = \operatorname{expit}\{0.60X_{1t} - 0.40X_{2t} + gU_t + s(r_t - 2.5)\}, \tag{7.6}$$

where g controls the strength of latent confounding in treatment assignment and s controls stagewise logging drift. The reward is

$$Y_t = 0.35X_{1t} - 0.20X_{2t} + 0.75U_t + 0.15\cos(r_t) + A_t\Delta_t + \varepsilon_t^Y, \tag{7.7}$$

where

$$\Delta_t = 0.90X_{1t} - 0.50X_{2t} + 0.28\sin(r_t + 1). \tag{7.8}$$

The initial variables U_1, $X_{2,1}$, and ε_1^{x0} are mutually independent. All listed innovation sequences are mutually independent of these initial variables, independent across stages, and independent across trajectories. Specifically, $\varepsilon_{t+1}^U \sim N(0, 0.68^2)$, $\varepsilon_{t+1}^1 \sim N(0, 0.58^2)$, $\varepsilon_{t+1}^2 \sim N(0, 0.62^2)$, $\varepsilon_t^Z, \varepsilon_t^W \sim N(0, 0.75^2)$, and $\varepsilon_t^Y \sim N(0, 1)$. Because actions do not affect future states in this benchmark, the oracle rule is

$$\pi_t^{\mathrm{or}}(H_t) = \mathbf{1}(\Delta_t \geq 0). \tag{7.9}$$

Discounted test regret is computed as

$$\widehat{\mathrm{Reg}}_{\mathrm{test}}(\pi) = \sum_{t=1}^{T} \gamma^{t-1} \frac{1}{n_{\mathrm{test}}} \sum_i \left[\max(\Delta_{it}, 0) - \pi_t(H_{it})\Delta_{it}\right], \tag{7.10}$$

with $\gamma = 0.95$. We compare Naive outcome regression, inverse-probability weighting (IPW), doubly robust learning (DR), fixed propensity-trimmed DR (Trim-DR), Proximal-DR without the data-adaptive local cap, and Taple-F. For every method, the executable policy learner regresses the cross-fitted pseudo-contrast on the declared stagewise policy features by ridge least squares and takes the fitted sign; this is a common surrogate policy-fitting rule, not an exact empirical-value optimizer. All controlled-benchmark scores use trajectory-level two-fold cross-fitting; the PhysioNet analysis instead groups folds by source record. In Taple-F, each training fold forms a grid from empirical quantiles of the absolute local treatment-bridge correction and chooses the cap by minimizing the absolute change from the locally un-winsorized residual correction plus a standard-error penalty. Because the factorized direct contrast is unweighted, this executable local rule caps only the residual correction; it is not the cumulative two-term selector in (5.9).

The implementation also uses fixed numerical safeguards that are distinct from the adaptive cap and are applied before policy fitting: estimated observable propensities are restricted to $[0.01, 0.99]$, the inverse target used in the treatment-bridge fit uses a $[0.02, 0.98]$ propensity restriction, treatment-bridge predictions are bounded to $[-80, 80]$, and inverse-weighted contrast scores are finally restricted to $[-100, 100]$. Therefore, "without adaptive local cap" is the precise baseline description; it is not literally unbounded. The run manifest records these constants. In the twelve-replication main benchmark the final $[-100, 100]$ safeguard is activated 15 times for IPW and zero times for DR, Trim-DR, Proximal-DR, and Taple-F; thus the reported main Proximal-DR versus Taple-F comparison is unaffected by the outer score safeguard.

All tuning quantities used to score a held-out fold are determined from its complementary training fold. In particular, the adaptive local-cap candidates are constructed from training-fold empirical quantiles and are selected by the training-fold bias–standard-error criterion; held-out outcomes are not used for cap selection. By contrast, the propensity restrictions, treatment-bridge prediction bound, and final score safeguard are fixed numerical protections applied consistently to the relevant competing methods rather than data-adaptive performance-tuning devices. We therefore distinguish sensitivity of the scientific data-generating parameters (g, s) from sensitivity to algorithmic and numerical settings. The former is displayed in Figure 4. A complete algorithmic-sensitivity audit would vary the cap-grid endpoints and density, the cap penalty constant, ridge regularization, the pilot policy class, propensity restrictions, bridge-prediction bounds, and the final score safeguard one factor at a time while reporting regret, action agreement, selected caps, and safeguard activation counts. The current replication archive is too small to support that high-dimensional analysis, so no claim of hyperparameter insensitivity is made and no parameter was re-tuned to improve a reported test result.

The controlled generator uses Gaussian state and reward disturbances, and its ridge outcome-bridge predictions are not constrained to a common almost-sure bound. Consequently these simulations do not literally instantiate the bounded-outcome/bridge condition in Assumption 2.4. They are finite-sample stress tests of the factorized score and local capping mechanism, not a numerical verification that every condition of Theorem 6.2 holds. Extending the theorem to sub-Gaussian outcomes and fitted bridges would require replacing the bounded-envelope empirical-process argument and is left open.

Table 1 reports twelve independent replications with $n = 1100$, $g = 0.8$, and $s = 0.4$. In this controlled benchmark, proxy-aware estimation provides the largest improvement: Proximal-DR reduces mean regret by about 72% relative to DR, and adaptive local capping lowers it further to 0.0370. IPW has the largest dispersion because rare logged actions amplify reward noise. The difference

between DR and fixed Trim-DR remains small relative to the proxy-aware gains, indicating that a single propensity floor does not resolve latent confounding. Action agreement confirms the same ordering and is evaluated against the known oracle rather than a fitted surrogate. With twelve replications, the Monte Carlo standard errors for mean regret are 0.0105 (Naive), 0.0269 (IPW), 0.0111 (DR), 0.0096 (Trim-DR), 0.0044 (Proximal-DR), and 0.0041 (TAPLE-F). The corresponding marginal 95% t intervals are approximately $[0.188, 0.234]$, $[0.259, 0.378]$, $[0.182, 0.230]$, $[0.177, 0.220]$, $[0.047, 0.067]$, and $[0.028, 0.046]$. These intervals quantify Monte Carlo uncertainty in each mean but are not paired intervals for method differences; the latter require seed-level covariance and should be reported in a larger rerun.

Table 1. Controlled policy performance under confounding and support drift.

Method	Mean regret	Regret SD	Action agreement
Naive	0.2109	0.0365	0.8296
IPW	0.3185	0.0933	0.8204
DR	0.2059	0.0384	0.8337
Trim-DR	0.1984	0.0333	0.8351
Proximal-DR	0.0571	0.0153	0.9323
TAPLE-F	**0.0370**	**0.0141**	**0.9449**

Figure 3 displays log–log regret curves for the two proxy-aware methods. Both improve as the trajectory count rises, which is expected because the bridge equations become more stable. The adaptive method is already better at $n = 300$, shows a larger gap at $n = 600$, and retains an advantage at $n = 2200$. The figure is not used to estimate an asymptotic exponent: four sample sizes and eight replications per sample size are still insufficient for that claim. Its purpose is to show that no visible bias floor appears over the examined sample-size range and that the empirical trend is compatible with the theoretical rate decomposition.

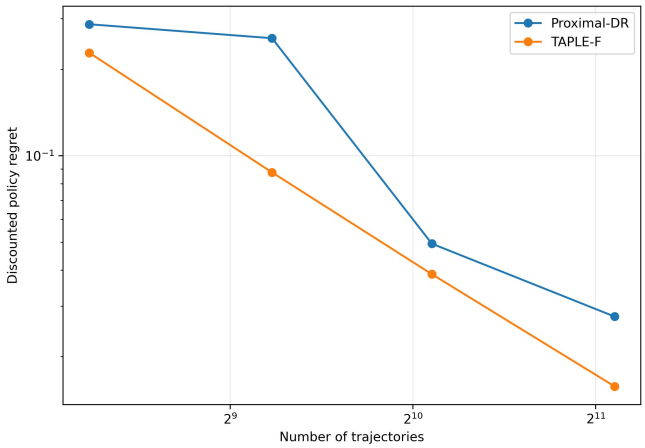

Figure 3. Regret scaling of the two proximal learners.

Table 2 provides the numerical values behind Figure 3. At $n = 300$, TAPLE-F has both lower mean regret and slightly lower variability. By $n = 600$, its mean regret is about one third that of Proximal-DR without the adaptive local cap. At $n = 1100$, the gap narrows because both estimators are accurate,

whereas the $n = 2200$ result again favors adaptive capping with smaller dispersion. The scaling experiment uses its own eight fixed-seed replications at each sample size, so its $n = 1100$ mean need not equal the twelve-replication main-table mean. These nonmonotone finite-sample standard deviations are retained rather than smoothed, and every entry is computed from executed replications in the supplied result file.

Table 2. Regret scaling of the proxy-aware learners.

n	Proximal-DR mean	Proximal-DR sd	TAPLE-F mean	TAPLE-F sd
300	0.2869	0.1247	0.2282	0.1171
600	0.2565	0.1120	0.0875	0.0338
1100	0.0494	0.0201	0.0387	0.0176
2200	0.0276	0.0106	0.0158	0.0058

Figure 4 reports the regret reduction Reg(DR) − Reg(TAPLE-F) over nine combinations of hidden-confounding strength and support drift. Every displayed value is positive, but the gain is not uniform. Stronger latent confounding generally produces a larger separation because observable-only DR does not adjust for the latent reward component in this generator. Drift changes the correction-weight distribution and makes adaptive capping more consequential. The heat map therefore supports an interaction interpretation: proximal identification addresses structural bias, while stagewise capping controls the additional variance that appears when the logging distribution becomes more polarized over time.

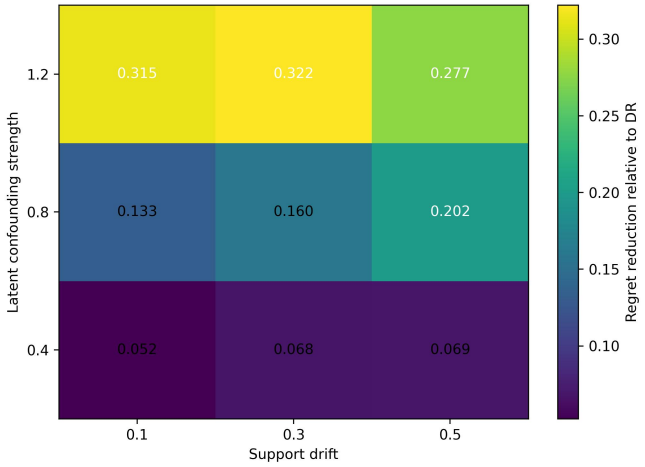

Figure 4. Regret reduction across confounding and support regimes.

Figure 5 shows the average selected correction cap and the observed-action local correction concentration/ESS diagnostic across decision stages in the executable TAPLE-F benchmark. Caps vary rather than following a monotone pattern because the empirical bridge distribution reflects both logging drift and estimation noise. The reported stagewise diagnostic lies between approximately 0.88 and 0.93, so local capping reduces tail leverage without concentrating the factorized correction on only a few observations. This quantity is not the cumulative policy-matched ESS in (5.13), and neither diagnostic validates the proxy assumptions; both only quantify weight concentration.

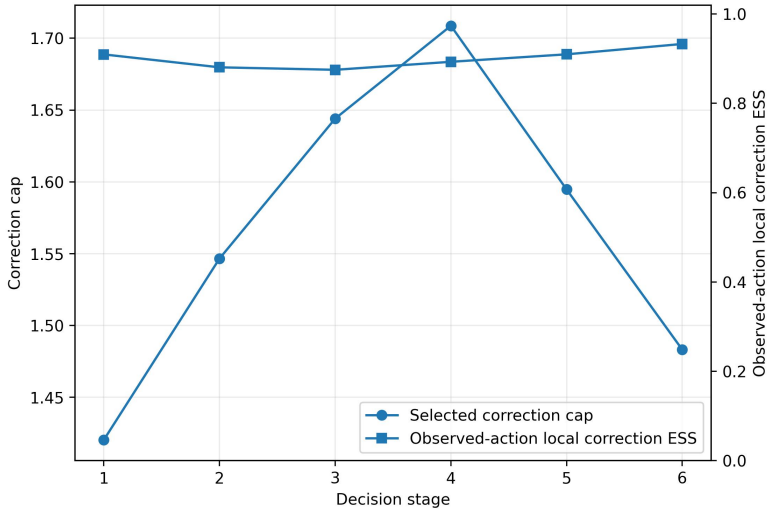

Figure 5. Stagewise correction caps and observed-action local correction concentration diagnostic.

7.2. Executed ablations

The ablations rerun the estimator after removing or replacing one component. No result is created by adding a constant or random perturbation to the full method. "No adaptive capping" uses Proximal-DR without the data-adaptive local cap while retaining the fixed numerical safeguards described above, "Fixed correction cap" uses the same cap at every stage, "Hard stage deletion" zeros the residual correction when its bridge magnitude exceeds the selected cap, "No cross-fitting" trains and scores on the same fold, and "No proxy bridges" uses observable-history DR.

Table 3 shows that removing proxy bridges is by far the most damaging change, increasing mean regret from 0.0363 to 0.3073. Fixed and absent adaptive capping also increase both mean regret and variability, while hard stage deletion is moderately worse than the full cross-fitted learner. The no-cross-fitting diagnostic has a slightly lower mean regret (0.0349 versus 0.0363) and nearly identical action agreement in these twelve fixed-seed runs. Because it trains and scores on the same observations, however, it is deliberately leakage-prone and is not an out-of-fold estimator; this finite-sample comparison therefore cannot be interpreted as evidence that cross-fitting is unnecessary for the theoretical construction. All six variants were independently executed for twelve seeds.

Table 3. Executed ablation study for the factorized learner.

Variant	Mean regret	Regret standard deviation	Action agreement
Full TAPLE-F	0.0363	0.0118	0.9462
No cross-fitting	0.0349	0.0123	0.9463
Hard stage deletion	0.0413	0.0122	0.9420
Fixed correction cap	0.0624	0.0333	0.9368
No adaptive capping	0.0786	0.0556	0.9316
No proxy bridges	0.3073	0.1491	0.8016

Figure 6 displays the seed-level regret distributions used in Table 3. Full TAPLE-F, no cross-fitting, and hard stage deletion overlap substantially across seeds; in fact, the leakage-prone no-cross-fitting diagnostic has the lowest median in this finite run. Fixed and absent adaptive capping show wider upper tails, whereas the no-proxy distribution is substantially shifted upward relative to the other variants. Showing all twelve replications prevents a mean-only summary from hiding these finite-sample reversals and capping failures. The empirical ablation therefore supports the importance of proxies and local robustification, while cross-fitting is justified here by out-of-fold construction and the theory rather than by a uniformly better finite-sample mean.

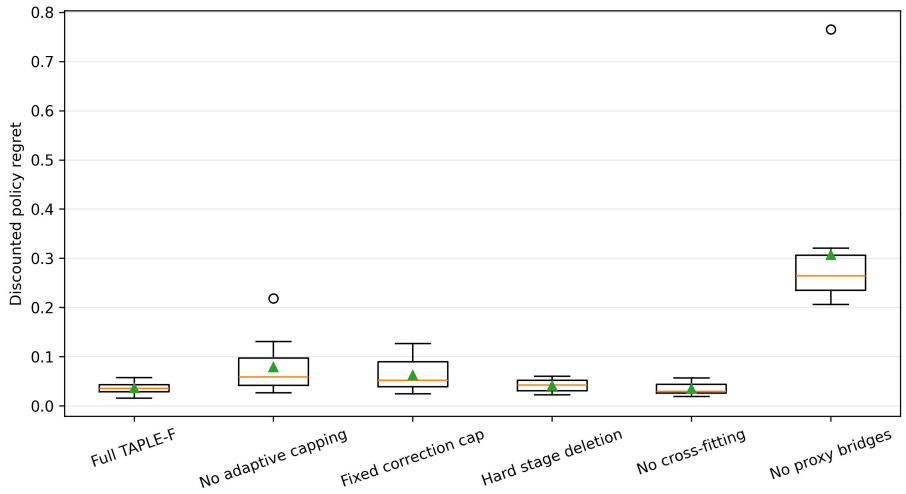

Figure 6. Seed-level regret distributions for executed ablations.

7.3. Source-grouped PhysioNet stress test

We use 13 public records from the PhysioNet/Computing in Cardiology Challenge 2012 [40], accessed through the PhysioNet platform [41, 42]. Measurements from the first 48 hours are aggregated into eight six-hour intervals. A fixed source-level split is selected before resampling: the subset's sole death source is retained in training, eight of the twelve survivor sources are sampled without replacement with seed 2701, and the four remaining survivor sources form the test set. The nine training sources generate 320 trajectories and the four disjoint test sources generate 180 trajectories. Logged actions, proxies, and rewards are then generated by a documented semi-synthetic mechanism. Within training, the two nuisance folds are also formed by source identity, so descendants of one template cannot appear on both sides of a cross-fitting split. No source template leaks across train/test or nuisance folds. However, because all held-out sources are survivors, this experiment cannot assess mortality transport, does not identify an actual treatment effect, and is not covered by the independent-trajectory asymptotic theorem.

Table 4 is a deliberately small-source stress test. Trim-DR has the smallest regret, closely followed by DR. TAPLE-F improves only modestly on Proximal-DR without the adaptive local cap and remains worse than both observable-history doubly robust estimators. IPW is worst under the fixed split, while Naive also trails the proximal learners. The result does not support a claim of universal proxy-method superiority and exposes the instability of bridge inversion with only nine independent training sources. Naive has the highest unweighted action agreement in Table 4 but worse regret than DR and Trim-DR;

this is not contradictory because action agreement weights all decisions equally, whereas regret weights mistakes by treatment-contrast magnitude and discounting.

Table 4. Source-grouped PhysioNet semi-synthetic stress test.

Method	Discounted regret	Action agreement
Naive	1.2752	**0.6667**
IPW	1.7220	0.5167
DR	1.0011	0.6597
Trim-DR	**0.9814**	0.6292
Proximal-DR	1.1508	0.6215
TAPLE-F	1.1161	0.6424

Figure 7 visualizes the clinical stress-test regrets. Trim-DR and DR are best under the fixed source split. Local adaptive capping reduces Proximal-DR regret from 1.1508 to 1.1161, an improvement of about 3%, but TAPLE-F still trails both observable-history doubly robust estimators. The common final $[-100, 100]$ safeguard is active for 114 of the 2560 cross-fitted Proximal-DR training contrast scores and for none of the TAPLE-F scores in this stress test. Consequently this is a comparison between adaptive local capping and a Proximal-DR baseline that already retains an emergency numerical bound, not a comparison with a mathematically unbounded estimator. This single-split comparison supports only a modest numerical improvement from local capping; it is not evidence of clinical efficacy, general superiority, or variance reduction. No confidence interval is shown because the experiment uses one prespecified split with nine training and four test source records.

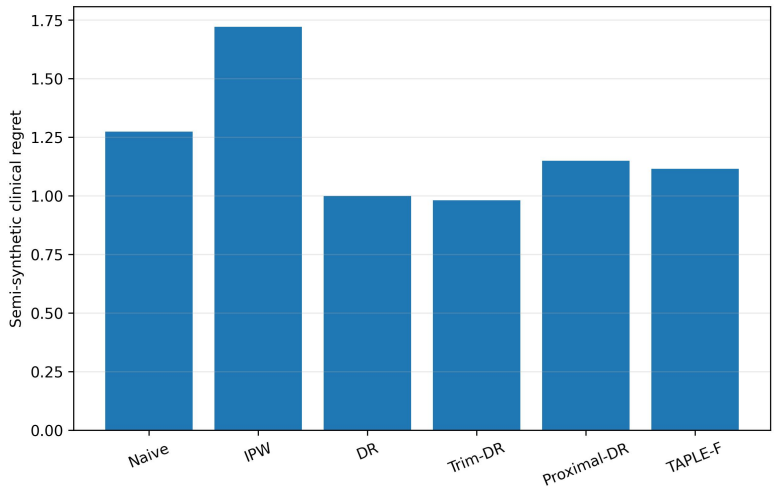

Figure 7. Method comparison in the source-grouped clinical stress test.

Figure 8 plots cross-fitted, observed-action proximal correction magnitudes at three controlled stages before the data-adaptive local cap (but after the fixed $[-80, 80]$ bridge-prediction safeguard). The distributions are right-skewed and retain nonnegligible mass far from their centers. Later stages do not simply shift uniformly; their shapes change because the logging law, state distribution, and stage-specific bridge fit interact. This observation motivates a cap chosen from each stage's own training distribution. The plot is descriptive rather than a tail-index estimate. Reliable estimation

of α_t would require substantially more trajectories and a separate extreme-value analysis, which the proposed selector intentionally avoids. For a full TAPLE implementation, the corresponding audit should additionally report cumulative-weight quantiles, out-of-fold conditional-moment residual norms, smallest retained singular values/condition numbers for each bridge system, the selected cap-grid location, and an empirical or optimization-based bound on ϵ_{opt}. The present TAPLE-F files report local corrections and clipping counts only; they do not contain the full cumulative quantities, and we do not relabel those local diagnostics as evidence for the general theorem.

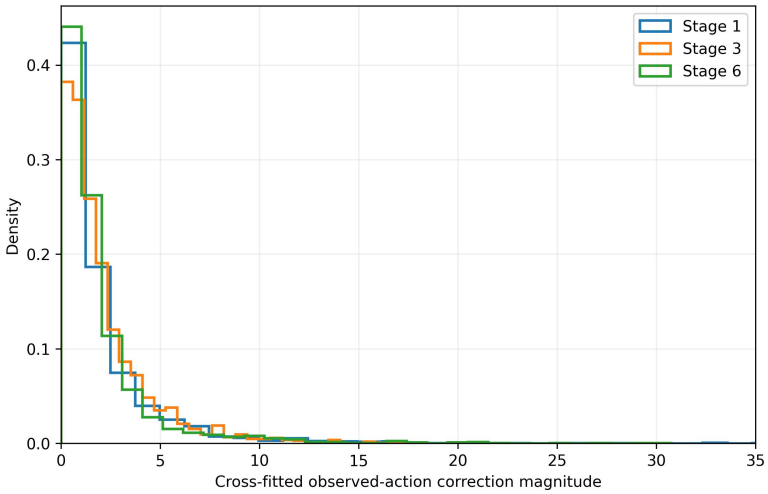

Figure 8. Estimated proximal correction magnitudes at selected stages.

7.4. Reproducibility checks

The experiment script generates every processed comma-separated-value (CSV) file, table, and figure from fixed seeds. The main benchmark has twelve replications; scaling experiments have eight per sample size; the sensitivity grid has four per cell; and ablations have twelve. Executable assertions verify that source-record train/test sets and every source-grouped nuisance split are disjoint. The package writes raw seed-level results, a source-group audit, score-clipping counts, and summaries. The machine-readable run manifest records

$$\mathcal{M} = \{\text{seeds, folds, source split, cap rule, methods, file paths}\}. \tag{7.11}$$

The current replication counts are adequate only for an illustrative finite-sample comparison, not for strong significance claims. A definitive simulation study should use at least 30–50 common-seed replications per configuration, report paired confidence intervals or a justified paired nonparametric test, and vary proxy noise, bridge misspecification, horizon length, treatment carryover, cap-grid endpoints, ridge regularization, and fixed safeguard bounds one factor at a time. Those additional experiments require new simulation runs and are not inferred from the existing twelve/eight/four-replication summaries.

Numerical consistency is checked by recomputing table summaries from the raw files:

$$\max_j \left| \widehat{T}_j^{\mathrm{paper}} - \widehat{T}_j^{\mathrm{raw}} \right| \le 5 \times 10^{-5}, \tag{7.12}$$

where the tolerance reflects four-decimal display rounding.

Remark 7.1. The controlled results support the local capping mechanism in the factorized regime, but the clinical stress test sharply limits the empirical claim. Proxy correction can be unstable when only a few independent source trajectories identify the bridge operators. Under the fixed source split, local adaptive capping yields only a modest improvement over proximal learning without the adaptive cap; larger source cohorts with greater outcome diversity are required for a credible comparison of proximal and conventional doubly robust learners.

8. Discussion

The central conceptual point is that longitudinal identification cannot be reduced to isolated stagewise contrasts. In particular, a stagewise doubly robust contrast is not, by itself, a valid longitudinal value representation when hidden states persist. The correct object is the nested proximal score in (3.9), whose proof moves through the intervention laws P_t^{π}. Lemma 3.3 additionally shows how the intervention-law bridge moments used by the general procedure can be represented by behavior-law weighted moments when preceding treatment bridges are valid. The expanded form (3.18) also shows that longitudinal overlap is governed by cumulative bridge products. This explains why capping only a local propensity or bridge factor can leave later-stage variance uncontrolled.

The model class remains demanding. Sequential negative controls must hold for the policy class being searched, outcome bridges are policy indexed, and completeness is an operator property rather than an empirical goodness-of-fit statistic. The upper theorem therefore carries the stage-specific $R_{t,n}^{\mathrm{idx}}$ remainders, aggregated as the uniform quantity R_n^{idx} in the uncapped nested drift, instead of pretending that one bridge fit works uniformly for every candidate policy. Sensitivity analysis is still required when proxy roles are uncertain. A practical analysis should deliberately weaken candidate proxies, perturb exclusion assumptions through bias functions, and track bridge residuals, condition numbers, cumulative-weight quantiles, subgroup regret, and support failures. These checks do not turn an untestable exclusion restriction into a testable one; instead, they make the consequences of weak or partially invalid proxies visible.

The controlled benchmark is intentionally factorized. In that design, the population policy value separates and the oracle action is known without using the fitted nuisance models. The executable learner nevertheless uses a ridge-sign surrogate fit to the cross-fitted pseudo-contrast; it does not certify exact empirical maximization or estimate ϵ_{opt}. This isolates latent confounding and heavy-correction-weight stabilization while deliberately leaving the general numerical optimization problem unvalidated. A separate carryover benchmark with a verified global optimizer and an auditable optimization gap is an important next step. This separates our problem from standard RL optimization: Q-learning, policy iteration, actor–critic methods, graph RL, and metaheuristics primarily specify how to search a sequential objective [14, 16, 19]; recent general-purpose population-based optimizers such as the glider snake and gray langurs methods further illustrate the breadth of numerical search mechanisms available for difficult nonconvex objectives [43, 44]. TAPLE first constructs an observationally identified objective under latent confounding and then leaves the feasible optimizer as a modular layer. Fair future comparisons should therefore use a common policy class, common feasibility constraints, and the same identified objective whenever possible, while separately reporting computational cost and optimization gap.

The PhysioNet protocol preserves real measurement patterns, separates source records before

resampling, and groups all descendants of a source record in the same nuisance-training fold. Actions and rewards remain semi-synthetic. Its result is informative because it reveals a regime in which local bridge inversion remains noisy despite leakage-resistant grouping. It cannot establish clinical benefit, safety, or transportability. Any substantive deployment would require prespecified proxies, external validation, sensitivity regions, and prospective governance. Because there are only 13 independent sources and all four held-out sources are survivors, the present split is now described strictly as an illustrative source-grouped stress test. A deployment-oriented study would additionally require subgroup auditing, uncertainty-based abstention under poor support, explicit harm constraints, clinician or operator oversight, external cohorts, prospective evaluation, distribution-shift monitoring, and post-deployment surveillance.

Remark 8.1. The practical recommendation is to audit the full identification and optimization chain. Reports should state intervention-law assumptions, admissible policy inputs, bridge residuals, cap values, signed-weight concentration diagnostics, source separation, and optimization gaps. Predictive accuracy cannot validate negative-control assumptions. A favorable off-policy score also cannot establish safety. The method controls one statistical failure mode but does not replace domain knowledge or prospective evidence.

9. Conclusions and future work

This paper introduced TAPLE as a data-driven sequential policy optimization framework under latent confounding and limited longitudinal overlap. Its principal contribution is an intervention-indexed nested proximal value representation that avoids an invalid observed-history Bellman substitution; expanding that representation identifies cumulative proximal weights as the relevant longitudinal instability mechanism. Temporally adaptive capping then regularizes those cumulative terms, while the regret decomposition retains bridge, indexing, cap-calibration, empirical-process, and optimization errors rather than hiding them in a generic remainder. The minimax lower bound and the oracle-calibrated upper bound have matching overlap exponents under the stated fixed-horizon rate-matching conditions. Controlled experiments support local robustification in the factorized submodel, while the source-grouped PhysioNet stress test shows only a modest capping benefit and favors observable-history doubly robust methods when only nine source records are available for training. In the twelve-replication controlled benchmark, TAPLE-F attains mean regret 0.0370 and action agreement 0.9449, compared with regret 0.0571 and agreement 0.9323 for Proximal-DR without the adaptive local cap. In the prespecified PhysioNet split, local capping reduces Proximal-DR regret only from 1.1508 to 1.1161, while DR and Trim-DR achieve lower regrets of 1.0011 and 0.9814. Accordingly, the numerical evidence supports the proxy/local-capping mechanism in TAPLE-F under the controlled factorized design but does not validate the full cumulative-score TAPLE algorithm, establish clinical superiority, or justify a general significance claim from the present number of independent replications and source records.

Three directions deserve priority. First, growing horizons and dependent trajectories require mixing conditions and nonasymptotic control of cumulative bridge products. Second, continuous and combinatorial actions require vector-valued bridges and scalable constrained policy search. Third, partially invalid proxies should be handled through sensitivity regions and lower-value optimization, producing policies that remain defensible when point identification is doubtful.

Author contributions

Siyang Bai: Conceptualization, methodology, software, formal analysis, investigation, data curation, visualization, and writing–original draft. Zheng Fang: Methodology, validation, formal analysis, and writing–review and editing. Jie Chen: Conceptualization, supervision, project administration, validation, and writing–review and editing. All authors have read and agreed to the final version of the manuscript.

Use of Generative-AI tools declaration

The authors used ChatGPT (OpenAI) only for limited language proofreading and editorial assistance during the production-proof stage. It was not used to generate data, analyses, results, or scientific conclusions. All AI-assisted wording was reviewed and approved by the authors, who take full responsibility for the final content.

Data availability statement

The supplementary package contains executable code, requirements, the raw PhysioNet subset, deterministic processing and source-split code, the run manifest, replication results, audits, summary tables, and all figure files in JPG format. The source dataset is distributed by PhysioNet under the Open Data Commons Attribution License v1.0; the manuscript includes both the challenge citation and PhysioNet platform citations, and users of the packaged subset should comply with the current source terms and attribution requirements. The controlled implementation evaluates the stated factorized Taple-F analogue; it is not presented as the full cumulative-score implementation or as a generic solver for arbitrary carryover dynamics. The clinical protocol is explicitly semi-synthetic and must not be used to infer a treatment effect in the original cohort.

Conflict of interest

All authors declare no conflict of interest in this paper.

References

1. S. A. Murphy, Optimal dynamic treatment regimes, *J. R. Stat. Soc. B*, **65** (2003), 331–355. https://doi.org/10.1111/1467-9868.00389
2. E. B. Laber, D. J. Lizotte, M. Qian, W. E. Pelham, S. A. Murphy, Dynamic treatment regimes: Technical challenges and applications, *Electron. J. Statist.*, **8** (2014), 1225–1272. https://doi.org/10.1214/14-EJS920
3. Y. Zhao, D. Zeng, A. J. Rush, M. R. Kosorok, Estimating individualized treatment rules using outcome weighted learning, *J. Am. Stat. Assoc.*, **107** (2012), 1106–1118. https://doi.org/10.1080/01621459.2012.695674
4. N. T. Williams, K. L. Hoffman, I. Díaz, K. E. Rudolph, Learning optimal dynamic treatment regimes from longitudinal data, *Am. J. Epidemiol.*, **193** (2024), 1768–1775. https://doi.org/10.1093/aje/kwae122

5. S. Athey, S. Wager, Policy learning with observational data, *Econometrica*, **89** (2021), 133–161. https://doi.org/10.3982/ECTA15732
6. H. Ye, W. Zhou, R. Zhu, A. Qu, Stage-aware learning for dynamic treatments, *J. Mach. Learn. Res.*, **25** (2024), 1–51.
7. X. Nie, E. Brunskill, S. Wager, Learning when-to-treat policies, *J. Am. Stat. Assoc.*, **116** (2021), 392–409. https://doi.org/10.1080/01621459.2020.1831925
8. A. Bennett, N. Kallus, Proximal reinforcement learning: Efficient off-policy evaluation in partially observed Markov decision processes, *Oper. Res.*, **72** (2024), 1071–1086. https://doi.org/10.1287/opre.2021.0781
9. C. Kausik, Y. Lu, K. Tan, M. Makar, Y. Wang, A. Tewari, Offline policy evaluation and optimization under confounding, In: *Proceedings of The 27th International Conference on Artificial Intelligence and Statistics*, **238** (2024), 1459–1467.
10. H. Zheng, D. Wang, A study of value iteration and policy iteration for Markov decision processes in deterministic systems, *AIMS Mathematics*, **9** (2024), 33818–33842. https://doi.org/10.3934/math.20241613
11. S. Yu, S. Fang, R. Peng, Z. Qi, F. Zhou, C. Shi, Two-way deconfounder for off-policy evaluation in causal reinforcement learning, In: *Advances in Neural Information Processing Systems 37*, 2024, 78169–78200. https://doi.org/10.52202/079017-2485
12. A. Bennett, N. Kallus, M. Oprescu, W. Sun, K. Wang, Efficient and sharp off-policy evaluation in robust Markov decision processes, In: *NIPS '24: Proceedings of the 38th International Conference on Neural Information Processing Systems*, 2024, 112962–113000.
13. P. Miao, A finite-time Q-Learning algorithm with finite-time constraints, *AIMS Mathematics*, **10** (2025), 23380–23393. https://doi.org/10.3934/math.20251038
14. I. A. Zamfirache, R. E. Precup, E. M. Petriu, Q-learning, policy iteration and actor-critic reinforcement learning combined with metaheuristic algorithms in servo system control, *Facta Univ. Ser. Mech. Eng.*, **21** (2023), 615–630. https://doi.org/10.22190/FUME231011044Z
15. H. Wang, B. R. Sarker, J. Li, J. Li, Adaptive scheduling for assembly job shop with uncertain assembly times based on dual Q-learning, *Int. J. Prod. Res.*, **59** (2021), 5867–5883. https://doi.org/10.1080/00207543.2020.1794075
16. F. Zhao, Z. Fu, L. Wang, H. Sang, A heterogeneous graph reinforcement learning framework with question-aware neighborhood aggregation and interoption prompt attention for dynamic flexible job shop scheduling problem, *IEEE T. Ind. Inform.*, **22** (2026), 2863–2874. https://doi.org/10.1109/TII.2025.3646962
17. Z. Pan, D. Lei, L. Wang, A knowledge-based two-population optimization algorithm for distributed energy-efficient parallel machines scheduling, *IEEE T. Cybernetics*, **52** (2022), 5051–5063. https://doi.org/10.1109/TCYB.2020.3026571
18. Z.-L. Lu, U. H. Lok, Dimension-reduced modeling for local volatility surface via unsupervised learning, *Rom. J. Inf. Sci. Technol.*, **27** (2024), 255–266. https://doi.org/10.59277/ROMJIST.2024.3-4.01

19. S. Meng, K. Chhea, S. Muy, J. R. Lee, Deep reinforcement learning and metaheuristic approaches to maximize downlink sum-rate for Internet of Things systems in non-orthogonal multiple access-based space-air-ground integrated networks, *Rom. J. Inf. Sci. Technol.*, **28** (2025), 327–340. https://doi.org/10.59277/ROMJIST.2025.4.02

20. W. Miao, Z. Geng, E. J. Tchetgen Tchetgen, Identifying causal effects with proxy variables of an unmeasured confounder, *Biometrika*, **105** (2018), 987–993. https://doi.org/10.1093/biomet/asy038

21. Z. Qi, R. Miao, X. Zhang, Proximal learning for individualized treatment regimes under unmeasured confounding, *J. Am. Stat. Assoc.*, **119** (2024), 915–928. https://doi.org/10.1080/01621459.2022.2147841

22. T. Shen, Y. Cui, Optimal treatment regimes for proximal causal learning, *Advances in Neural Information Processing Systems 36*, 2023, 47735–47748. https://doi.org/10.52202/075280-2068

23. E. Sverdrup, Y. Cui, Proximal causal learning of conditional average treatment effects, In: *Proceedings of the 40th International Conference on Machine Learning*, **202** (2023), 33285–33298.

24. V. Chernozhukov, D. Chetverikov, M. Demirer, E. Duflo, C. Hansen, W. Newey, J. Robins, Double/debiased machine learning for treatment and structural parameters, *Econ. J.*, **21** (2018), C1–C68. https://doi.org/10.1111/ectj.12097

25. N. Kallus, M. Uehara, Double reinforcement learning for efficient off-policy evaluation in Markov decision processes, *J. Mach. Learn. Res.*, **21**, (2020), 1–63.

26. R. K. Crump, V. J. Hotz, G. W. Imbens, O. A. Mitnik, Dealing with limited overlap in estimation of average treatment effects, *Biometrika*, **96** (2009), 187–199. https://doi.org/10.1093/biomet/asn055

27. A. D'Amour, P. Ding, A. Feller, L. Lei, J. Sekhon, Overlap in observational studies with high-dimensional covariates, *J. Econ.*, **221** (2021), 644–654. https://doi.org/10.1016/j.jeconom.2019.10.014

28. N. Kallus, M. Uehara, Efficiently breaking the curse of horizon in off-policy evaluation with double reinforcement learning, *Oper. Res.*, **70** (2022), 3282–3302. https://doi.org/10.1287/opre.2021.2249

29. R. Zhan, Z. Ren, S. Athey, Z. Zhou, Policy learning with adaptively collected data, *Manage. Sci.*, **70** (2024), 5270–5297. https://doi.org/10.1287/mnsc.2023.4921

30. P. Zhao, A. Chambaz, J. Josse, S. Yang, Positivity-free policy learning with observational data, In: *Proceedings of the 27th International Conference on Artificial Intelligence and Statistics*, **238** (2024), 1918–1926.

31. M. G. Marmarelis, F. Morstatter, A. Galstyan, G. Ver Steeg, Policy learning for localized interventions from observational data, In: *Proceedings of the 27th International Conference on Artificial Intelligence and Statistics*, **238** (2024), 4456–4464.

32. X. Ma, Y. Sasaki, Y. Wang, Testing limited overlap, *Economet. Theor.*, **41** (2025), 1129–1162.

33. S. Joshi, J. Zhang, E. Bareinboim, Towards safe policy learning under partial identifiability: A causal approach, In: *Proceedings of the AAAI Conference on Artificial Intelligence*, **38** (2024), 13004–13012. https://doi.org/10.1609/aaai.v38i12.29198

34. H. Wang, Y. Xu, W. Lu, R. Song, Off-policy evaluation under nonignorable missing data, In: *Proceedings of the 42nd International Conference on Machine Learning (ICML)*, **267** (2025), 65020–65058.

35. A. Ying, W. Miao, X. Shi, E. J. Tchetgen Tchetgen, Proximal causal inference for complex longitudinal studies, *J. R. Stat. Soc. B*, **85** (2023), 684–704. https://doi.org/10.1093/jrsssb/qkad020

36. Y. Zhang, A. Bennett, A. Manca, M. Mittelman, M. Hoeks, A. Smith, et al., Estimating the causal effect of realistic treatment strategies using longitudinal observational data, *Med. Decis. Making*, **46** (2026), 144–157. https://doi.org/10.1177/0272989X251379819

37. H. Zan, D. Chen, Distributed stochastic optimization algorithm based on Markov sampling, *AIMS Mathematics*, **11** (2026), 4123–4146. https://doi.org/10.3934/math.2026166

38. H. Jiang, Y. Yang, B. Zou, J. Xu, Generalization analysis of tuning-free, Markov-ensemble SVM with distributed applications, *AIMS Mathematics*, **11** (2026), 13683–13709. https://doi.org/10.3934/math.2026564

39. M. Hargrave, A. Spaeth, L. Grosenick, EpiCare: A reinforcement learning benchmark for dynamic treatment regimes, In: *Advances in Neural Information Processing Systems 37*, 2024, 130536–130568.

40. I. Silva, G. Moody, D. J. Scott, L. A. Celi, R. G. Mark, Predicting in-hospital mortality of ICU patients: The PhysioNet/Computing in Cardiology Challenge 2012, In: *2012 Computing in Cardiology*, **39** (2012), 245–248.

41. T. Pollard, B. E. Moody, L. H. Lehman, B. J. Gow, C. Fernandes, C. Xie, et al., PhysioNet as a global platform for biomedical research, *Nat. Health*, **1** (2026), 792–795. https://doi.org/10.1038/s44360-026-00096-z

42. A. L. Goldberger, L. A. N. Amaral, L. Glass, J. M. Hausdorff, P. Ch. Ivanov, R. G. Mark, et al., PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals, *Circulation*, **101** (2000), e215–e220. https://doi.org/10.1161/01.CIR.101.23.e215

43. E. M. El-Kenawy, N. Khodadadi, S. Mirjalili, A. M. Zaki, A. Ibrahim, A. A. Alhussan, et al., Glider snake optimizer (GSO): A nature-inspired metaheuristic algorithm for global and engineering optimization problems, *Artif. Intell. Rev.*, **59** (2026), 91. https://doi.org/10.1007/s10462-026-11504-x

44. S. Barshandeh, N. Khodadadi, B. Abdollahzadeh, A. Mohammadzadeh, E. M. El-Kenawy, M. M. Eid, et al., Gray langurs optimizer: A multi-group bio-inspired optimization algorithm, *Artif. Intell. Rev.*, **59** (2026), 166. https://doi.org/10.1007/s10462-026-11529-2

AIMS Press