SEMIPARAMETRICALLY EFFICIENT OFF-POLICY EVALUATION AND LEARNING IN CONSTRAINED MARKOV DECISION PROCESSES

Anonymous authors
Paper under double-blind review

ABSTRACT

We study off-policy evaluation (OPE) in constrained Markov decision processes (CMDPs): estimating the reward value and the constraint cost values of a new policy from data collected by another. Before a policy is deployed as safe, its feasibility must be certified from confidence intervals, not point estimates. Such a certificate concerns a *combination* of the estimates—typically the Lagrangian $L_\lambda = v - \lambda^\top c$—so its validity depends on how the reward and cost errors covary. Existing methods fall short in two ways. Concatenating the step-wise doubly robust estimator across objectives inherits its exponential variance in the horizon. Applying a marginally efficient estimator to each objective separately breaks the curse of horizon, but treating the objectives as independent *for inference* discards the covariance induced by the trajectories they share. We instead treat the reward and cost values as a single vector target, derive its efficient influence function, and propose the *Constrained Doubly Robust* (CDR) estimator. CDR is unbiased whenever either the density ratio or the Q-functions are correctly specified, attains the joint semiparametric efficiency bound under a product-rate condition, and reports the full reward–cost covariance from one cross-fitted pass. The sign of that covariance governs the failure this repairs: when positive, independent treatment is merely conservative; when negative, it is *anti-conservative*—the Lagrangian intervals under-cover and an unsafe policy can be certified. Experiments on tabular and continuous-state CMDPs, and on a constrained Taxi benchmark at horizon 100, confirm the theory, including the negative-covariance regime and a diagnostic in which violating the product-rate condition collapses coverage.

1 INTRODUCTION

We study off-policy evaluation (OPE) in constrained Markov decision processes (CMDPs), the standard formalism of safe RL (Altman, 1999; García & Fernández, 2015; Achiam et al., 2017; Tessler et al., 2019): estimating both the return and the safety cost of a policy from data collected by a different policy (Precup et al., 2000). The constraint is what makes the problem both hard and worth solving. In clinical treatment protocols (Murphy et al., 2001), a regimen is adopted only if it improves outcomes *while* keeping toxicity below a tolerated level; in autonomous driving (Kiran et al., 2022), a controller must improve progress without exceeding a collision-risk budget; in recommendation (Li et al., 2010), engagement is maximized subject to exposure limits. In each case, running the candidate policy to measure its cost is precisely what one cannot afford to do, so the constraint must be checked offline.

This changes what an estimator must deliver. Rather than a point estimate of a single number, a constrained evaluator must *certify feasibility*: declare, with a confidence statement attached, that the cost lies below a threshold, simultaneously for the return and for every cost channel. Certification is a statement about a *combination* of the estimates, so its validity depends on how their errors covary, not only on how accurate each one is on its own.

Two classes of approaches are available, both obtained by applying a single-objective estimator to each objective in turn. The first uses cumulative importance weights, as in the step-wise doubly robust (DR) estimator (Robins & Rotnitzky, 1995; Dudík et al., 2011; Jiang & Li, 2016; Thomas

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

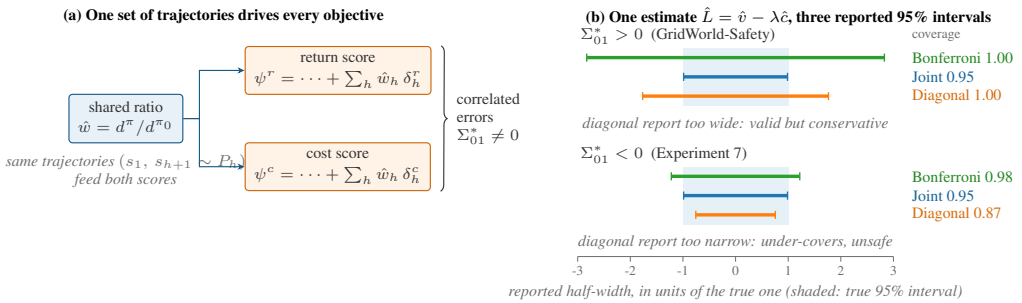


Figure 1: Why the reward–cost covariance matters. **(a)** Every component of the CDR score is computed from the same trajectories and reweighted by the same $\hat{w}$, so the errors of the return and cost estimates are correlated ($\Sigma^*_{01} \neq 0$ in general). **(b)** The Lagrangian estimate $\hat{L} = \hat{v} - \lambda\hat{c}$ and its true 95% interval (shaded) are the same under every report; only the reported width differs. With $\Sigma^*_{01} > 0$ (top, GridWorld-Safety) the diagonal report, which drops $\hat{\Sigma}_{01}$, is too wide: valid but conservative. With $\Sigma^*_{01} < 0$ (bottom, Experiment 7) it is too narrow and covers 0.87 instead of 0.95: an unsafe policy can be certified. Coverage values are asymptotic; Table 1 gives the empirical ones.

& Brunskill, 2016). They are unbiased whenever either the value model or the behavior policy is correct, and need no knowledge of the state-occupancy structure. Their variance, however, grows exponentially in the horizon, and concatenating them across objectives inherits that growth in every component at once. The second class replaces the cumulative weight with a marginal density ratio (Liu et al., 2018; Xie et al., 2019; Nachum et al., 2019; Kallus & Uehara, 2020; Uehara et al., 2020). This breaks the curse of horizon and is efficient for each objective separately. Here the difficulty is not the accuracy of the components but the relationship between them: every objective is estimated from the same logged trajectories (the same initial states and transitions, reweighted by the same density ratio), so the estimation errors of the return and of the costs are correlated in general. A componentwise analysis cannot see it, and combining the components as though they were independent mis-states the uncertainty of exactly the combinations from which a feasibility decision is made. A separate line of work optimizes policies directly under offline constraints (Le et al., 2019; Lee et al., 2022; Xu et al., 2022; Polosky et al., 2022; Liu et al., 2023; Hong et al., 2024; Zheng et al., 2024; Chemingui et al., 2025); it produces a policy, not a joint confidence region for its return and costs.

Whether this correlation is merely wasteful or actively dangerous depends on its sign, and neither case is detectable componentwise (Figure 1). When high-return trajectories also tend to incur high cost, ignoring the correlation *over*-states the uncertainty: the certificate remains valid but is needlessly conservative, and feasible policies are rejected. When return and cost are negatively related, as for a policy that gains reward precisely by avoiding the costly region, ignoring the correlation *under*-states the uncertainty. The certificate then covers less often than it claims, and a policy can be declared safe when it is not. This second regime is the one that matters for safety.

We propose an estimator that combines the non-compounding (polynomial rather than exponential in H) variance of the marginal-density-ratio approach with the robustness of the doubly robust approach, and that reports the joint uncertainty certification requires. Our contributions are threefold.

1. We derive the efficient influence function for the vector-valued target consisting of the return and all cost values, and propose the *Constrained Doubly Robust* (CDR) estimator (Theorems 2, 3), which extends and subsumes the single-objective marginal-efficient estimator. CDR is unbiased whenever either nuisance is correct, and one cross-fitted pass returns both the estimate and its joint covariance at no extra cost.
2. We establish joint asymptotic normality and attainment of the joint efficiency bound (Theorem 6), together with a local asymptotic minimax lower bound that CDR matches (Theorem 17). We then characterize in closed form what ignoring the joint structure costs: the *sign* of the return–cost covariance decides whether the diagonal report, which treats the components as independent, is conservative or anti-conservative (Theorem 5).

3. Empirically, on synthetic tabular and continuous-state CMDPs and on the Taxi benchmark with horizons up to 100, CDR-exact tracks the efficiency bound and CDR-data stays calibrated through its plug-in covariance, both avoid the exponential horizon growth of cumulative-weight estimators, and the diagonal (independence) report under-covers in the negative-covariance regime (Table 1). As an application it yields $O(1/\sqrt{n})$ return regret with high-probability constraint satisfaction (Section 6).

Related work. In the semiparametric tradition (Bickel et al., 1993; Newey, 1990; van der Vaart, 1998), efficiency for *single-objective* OPE was established by Kallus & Uehara (2020) and extended to infinite horizon by Kallus & Uehara (2022). The marginal density ratio $w = d^\pi/d^{\pi_0}$ that makes this possible originates in the marginalized importance-sampling line that first broke the curse of horizon (Liu et al., 2018), developed through the DICE family of distribution-correction estimators (Nachum et al., 2019; Dai et al., 2020) and minimax weight learning (Uehara et al., 2020). We carry w into the constrained, multi-objective setting and characterize the joint covariance structure of the resulting estimates. Offline safe reinforcement learning (RL) methods (Le et al., 2019; Lee et al., 2022; Xu et al., 2022; Polosky et al., 2022; Liu et al., 2023; Hong et al., 2024; Zheng et al., 2024; Chemingui et al., 2025) optimize policies directly; none produces a $\sqrt{n}$-consistent estimator of the joint target with a valid confidence region, which is what a feasibility certificate requires. Closest is Chen et al. (2025), who minimize evaluation variance subject to safety, but over the *behavior* policy; we take the log as given (Appendix A).

2 Setup and Notation

2.1 Constrained Markov Decision Process

A CMDP is $\mathcal{M} = (\mathcal{S}, \mathcal{A}, \{P_h\}_{h=1}^H, \{r_h\}, \{c_h^{(j)}\}_{j\le J}, \mu, \gamma)$ with state space $\mathcal{S}$, action space $\mathcal{A}$, transition kernels P_h, reward r_h, J cost functions $c_h^{(j)} : \mathcal{S}\times\mathcal{A} \to [0,1]$, initial distribution μ, and discount $\gamma \in (0,1]$.

Convention ($\gamma = 1$). The main text sets $\gamma = 1$, standard in finite-horizon OPE (Jiang & Li, 2016; Kallus & Uehara, 2020) and the regime in which the comparison against cumulative-weight estimators is sharpest (Proposition 1). This is presentational: Appendix J establishes the full parallel theory for every $\gamma \in (0,1]$ with the discount-weighted score $\tilde{w}_h = \gamma^{h-1} w_h$, and discounting in fact *strengthens* the horizon guarantee (the bound is uniform in H whenever the marginal ratio is; Corollary 16). The symbol γ is retained below for generality. A policy $\pi = \{\pi_h\}_{h=1}^H$ maps states to action distributions. The state-action occupancy measure under π is $d_h^\pi(s,a) := d_h^\pi(s)\,\pi_h(a|s)$, where $d_h^\pi(s)$ is the marginal state distribution at step h induced by π and $P_1,\dots,P_{h-1}$, with $d_1^\pi = \mu$. Expectations written as $\mathbb{E}_{d_h^\pi}[g(s_h)]$ refer to the state marginal; those involving both (s_h, a_h), such as $\mathbb{E}_{d_h^\pi}[g(s_h,a_h)]$, refer to the joint occupancy measure. Value functions are defined recursively:

$$Q_h^{\pi,\star}(s,a) := f_h^\star(s,a) + \gamma\,\mathbb{E}_{s'\sim P_h}[V_{h+1}^{\pi,\star}(s')], \tag{1}$$

$$V_h^{\pi,\star}(s) := \mathbb{E}_{a\sim\pi_h}[Q_h^{\pi,\star}(s,a)], \tag{2}$$

where $\star \in \{r, c_1, \dots, c_J\}$ with $f_h^r = r_h$ and $f_h^{c_j} = c_h^{(j)}$, and the boundary condition $V_{H+1}^{\pi,\star}(s) \equiv 0$ for all $s \in \mathcal{S}$. The joint target is:

$$\theta^* = (v^\pi,\, c_1^\pi,\, \dots,\, c_J^\pi)^\top \in \mathbb{R}^{J+1}, \tag{3}$$

where $v^\pi = \mathbb{E}_\mu[V_1^{\pi,r}(s_1)]$ and $c_j^\pi = \mathbb{E}_\mu[V_1^{\pi,c_j}(s_1)]$. Throughout, we adopt the 0-indexed convention $\star_0 := r$ and $\star_j := c_j$ for $j = 1,\dots,J$, so that $\theta_0^* = v^\pi$ and $\theta_j^* = c_j^\pi$. More generally we write $v_\star^\pi := \mathbb{E}_\mu[V_1^{\pi,\star}(s_1)]$, so $v_{\star_j}^\pi = \theta_j^*$. When $J = 1$ we write $c^\pi := c_1^\pi$ for brevity.

Notation. Following Jiang & Li (2016), lowercase v^π, c_j^π denote the *scalars* being estimated and uppercase $V_h^{\pi,\star}, Q_h^{\pi,\star}$ the *value functions*; all expectations are with respect to the trajectory distribution induced by (μ, π_0) unless stated otherwise. We write $\mathbb{E}_h[\cdot] := \mathbb{E}[\cdot \mid s_1, a_1, \dots, s_{h-1}, a_{h-1}]$ (and Var_h likewise) in place of explicit filtration notation, so that a quantity "known at step h" is one that passes outside $\mathbb{E}_{h+1}[\cdot]$. Throughout, $\delta_h^\star$ is the temporal-difference error of signal $\star$ (Definition 1), Δ denotes nuisance perturbation directions, $\varrho_{1:h}$ the cumulative importance ratio, and w_h the marginal density ratio.

2.2 Data and Assumptions

We observe n i.i.d. trajectories $\mathcal{D} = \{Z_i\}_{i=1}^n$ from behavior policy π_0.

Assumption 1 (Overlap). $\pi_0(a|s) \geq \eta_{\min} > 0$ *for all* (s, a) *with* $\pi(a|s) > 0$.

Assumption 2 (Bounded signals). $r_h, c_h^{(j)} \in [0, 1]$ *for all* h, j.

2.3 Nuisance Functions

The nuisance vector is $\eta = (w, Q^{\pi,r}, Q^{\pi,c_1}, \ldots, Q^{\pi,c_J})$, where $w_h(s, a) = d_h^\pi(s, a)/d_h^{\pi_0}(s, a)$ is the marginal density ratio. All $J+1$ score components are evaluated on the same trajectories: the same s_1 enters every $V_1^{\pi,\star}(s_1)$ and the same transition s_{h+1} drives every residual $\delta_h^\star$, reweighted by the same w_h. This shared randomness is the source of the off-diagonal covariance; it is present even on-policy ($w \equiv 1$), and w_h^2 only rescales each step's contribution (eq. (8)).

3 Doubly Robust Estimation in CMDPs

3.1 Background: Single-Objective DR and the Curse of Horizon

The step-wise DR estimator of Jiang & Li (2016) (JL) uses cumulative importance-sampling (IS) weights $\varrho_{1:h} = \prod_{t=1}^h \pi(a_t|s_t)/\pi_0(a_t|s_t)$ (we write ϱ for cumulative IS weights, distinct from the correlation ρ of Section 4). Written in the recursive form of Jiang & Li (2016, Eqn. 10), their estimator for signal $\star$ sets $V_{\mathrm{JL}}^{\star,0} \coloneqq 0$ and

$$V_{\mathrm{JL}}^{\star,H+1-h} \coloneqq \hat{V}_h^\star(s_h) + \varrho_h\Big(f_h^\star + \gamma V_{\mathrm{JL}}^{\star,H-h} - \hat{Q}_h^\star(s_h, a_h)\Big). \tag{4}$$

The obvious way to handle $J+1$ signals is to run (4) once per signal and concatenate; we call this baseline JL-Concat.

Two features of (4) drive what follows. First, it is *recursive*: ϱ_h multiplies the entire downstream estimate, so each step's importance ratio compounds into the variance already accumulated by all later steps. This is explicit in the term $\mathbb{E}_h[\gamma^2 \varrho_h^2 \operatorname{Var}_{h+1}[V_{\mathrm{JL}}^{\star,H-h}]]$ of Jiang & Li (2016, Thm. 1); compounding, not accumulation, is what produces exponential growth. Second, it never refers to the occupancy d_h^π, so it treats two trajectories reaching the same state by different routes as unrelated. Kallus & Uehara (2020) replaced $\varrho_{1:h}$ with the marginal ratio w_h for a *single* objective. As Section 3.2 makes precise, this turns the recursion into a flat sum: variance *adds* across steps where (4) makes it *multiply*.

Proposition 1 (Compounding variance of step-wise DR; proof in Appendix H). *Under Assumptions 1–2, even with the optimal control variate* $\hat{Q} = Q^{\pi,\star}$, $\hat{V} = V^{\pi,\star}$, *the step-wise (JL) estimator satisfies*

$$n \operatorname{Var}(\hat{v}_{\star,\mathrm{JL}}^\pi) = \operatorname{Var}(V_1^{\pi,\star}(s_1)) + \sum_{h=1}^H \gamma^{2(h-1)}\, \mathbb{E}\big[\varrho_{1:h-1}^2\, \sigma_h^2(s_h)\big]\,,$$

where $\sigma_h^2(s) \coloneqq \sum_a \frac{\pi(a|s)^2}{\pi_0(a|s)} \gamma^2 \operatorname{Var}_{P_h(\cdot|s,a)}(V_{h+1}^{\pi,\star}(s'))$ *is the weighted Bellman noise variance (which already absorbs the step-h importance weight ϱ_h^2) and all cross-step covariances vanish by the Bellman residual property* $\mathbb{E}[\delta_h^\star \mid s_h, a_h] = 0$. *The per-step second moment is exactly a* χ^2*-divergence,* $\mathbb{E}[\varrho_h^2 \mid s_h = s] = 1 + \chi^2(\pi_h(\cdot|s) \,\|\, \pi_{0,h}(\cdot|s))$. *Let* $1 + \underline{\chi}^2 \coloneqq \inf_{h,s} \mathbb{E}[\varrho_h^2 \mid s_h = s]$ *and* $\underline{\sigma}^2 \coloneqq \inf_{h<H,\, s} \sigma_h^2(s)$, *the infima taken over the support of* $d_h^{\pi_0}$ *(the step $h = H$ is excluded because* $V_{H+1}^{\pi,\star} \equiv 0$ *makes* $\sigma_H^2 \equiv 0$*). Iterating the tower property gives* $\mathbb{E}[\varrho_{1:h}^2] \geq (1 + \underline{\chi}^2)^h$, *hence*

$$n \operatorname{Var}(\hat{v}_{\star,\mathrm{JL}}^\pi) \;\geq\; \underline{\sigma}^2 \sum_{h=1}^{H-1} \big(\gamma^2(1 + \underline{\chi}^2)\big)^{h-1}, \tag{5}$$

Hence, if $\underline{\sigma}^2 > 0$ *and* $\gamma^2(1+\underline{\chi}^2) > 1$, *i.e.* $\underline{\chi}^2 > \gamma^{-2} - 1 \approx 2/H_{\mathrm{eff}}$ *with* $H_{\mathrm{eff}} \coloneqq (1-\gamma)^{-1}$, *the variance grows exponentially in H (a matching upper bound, with suprema in place of the infima, is in Appendix H). At* $\gamma = 1$ *the condition reads* $\underline{\chi}^2 > 0$ *(for finite $\mathcal{S}$:* $\pi_h(\cdot|s) \neq \pi_{0,h}(\cdot|s)$ *at every visited*

(h, s)*): the sharpest end of the comparison. In contrast, the efficiency bound* Σ^*_{00} *of Theorem 3 is a plain sum of* H *within-step terms with no compounding factor: if* $\sup_{h,s,a} \mathrm{Var}_{P_h}(V^{\pi,\star}_{h+1} \mid s, a) \le \bar{\sigma}^2_V$ *and* $\sup_h \|w_h\|_\infty \le M_w$ *with* M_w *not depending on* H*, then* $\Sigma^*_{00} - \mathrm{Var}_\mu(V^{\pi,\star}_1) \le \gamma^2 M_w \bar{\sigma}^2_V H$ *(using* $\mathbb{E}_{d^{\pi_0}_h}[w^2_h g] = \mathbb{E}_{d^{\pi}_h}[w_h g]$*): linear in* H *at* $\gamma = 1$*, at most cubic under Assumption 2 alone (*$|V^{\pi,\star}_{h+1}| \le H - h$*), and bounded uniformly in* H *under discounting (Corollary 16). Without an* H*-free* M_w *the guarantee does not apply: Assumption 1 alone gives only* $\|w_h\|_\infty \le \eta^{-h}_{\min}$*, and on tree MDPs* $w_h = \varrho_{1:h}$ *and* Σ^*_{00} *equals the JL variance (Remark 5).*

This does not conflict with the Cramér–Rao bound of Jiang & Li (2016) for discrete tree MDPs, where $w_h = \varrho_{1:h}$ and CDR coincides with (4); the gap between the two estimators comes from *state merging* of histories with different cumulative ratios (Remark 5, Appendix F).

3.2 The CDR Estimator

The recursion in (4) is not necessary: each step's correction only needs re-weighting from the π_0 to the π occupancy *at that step*, which is what w_h does, and replacing $\varrho_{1:h}$ by w_h turns the recursion into a flat sum in which no step's weight multiplies any other step's correction.

Definition 1 (CDR score)**.** *For* $\star \in \{r, c_1, \dots, c_J\}$*, define:*

$$\psi^\star(Z; v^\pi_\star, Q^{\pi,\star}, w) := V^{\pi,\star}_1(s_1) - v^\pi_\star + \sum_{h=1}^{H} w_h(s_h, a_h)\, \delta^\star_h, \tag{6}$$

where $\delta^\star_h := f^\star_h(s_h, a_h) + \gamma V^{\pi,\star}_{h+1}(s_{h+1}) - Q^{\pi,\star}_h(s_h, a_h)$ *is the Bellman residual at step* h*. The joint CDR score is* $\psi(Z; \theta, \eta) = (\psi^r, \psi^{c_1}, \dots, \psi^{c_J})^\top \in \mathbb{R}^{J+1}$*. Here* $Q^{\pi,\star}$ *and* w *are generic bounded arguments,* $V^{\pi,\star}_h(s) := \sum_a \pi_h(a|s) Q^{\pi,\star}_h(s, a)$ *is always computed from the* Q *argument, and* $V^{\pi,\star}_{H+1} \equiv 0$*.*

Theorem 2 (Double robustness; proof in Appendix I)**.** *Under the finite-horizon convention* $\gamma = 1$ *(see Remark 20 in Appendix I for the case* $\gamma < 1$*):* $\mathbb{E}[\psi^\star] = 0$ *whenever either* $Q^{\pi,\star}_h$ *is the true value function for every* h *(with* $\mathbb{E}_{d^{\pi_0}_h}|w_h| < \infty$*) or* $w_h = d^\pi_h / d^{\pi_0}_h$ *for every* h*.*

The score is also Neyman orthogonal—its derivative with respect to every nuisance direction vanishes at the truth—which is what reduces the requirement on nuisance estimation to the product rate of Assumption 3; Theorem 9 in Appendix C states this precisely.

3.3 Semiparametric Efficiency

By the trajectory factorization $dP(Z) = \mu(s_1) \prod_h \pi_0(a_h|s_h) \prod_h P_h(s_{h+1}|s_h, a_h)$, the tangent space at P decomposes into $L^2(P)$-orthogonal parts, $\dot{\mathcal{P}} = T_\mu \oplus T_{\pi_0} \oplus \bigoplus_{h=1}^H T_{P_h}$, spanned by mean-zero initial-state scores and by conditionally mean-zero policy and transition scores. The efficient influence function (EIF) is the unique element of $\overline{\dot{\mathcal{P}}}$ whose inner product with each score reproduces the pathwise derivative of θ^*.

Theorem 3 (EIF and efficiency bound; proof in Appendix L)**.** *Under the standing convention* $\gamma = 1$ *(see Section 2), the EIF for* $\theta^* \in \mathbb{R}^{J+1}$ *in the nonparametric, time-inhomogeneous CMDP model (*μ*,* π_0 *and each* P_h *vary freely; Remark 8, Appendix F) is* $\psi^*(Z) = \psi(Z; \theta^*, \eta^*)$*. The semiparametric efficiency bound is:*

$$\Sigma^* := \mathbb{E}[\psi^*(Z)\psi^*(Z)^\top] \in \mathbb{R}^{(J+1)\times(J+1)}, \tag{7}$$

with entries (recalling the 0*-indexed convention* $\star_0 = r$*,* $\star_j = c_j$*)*

$$\Sigma^*_{jk} = \mathrm{Cov}_\mu(V^{\pi,\star_j}_1, V^{\pi,\star_k}_1) + \sum_{h=1}^{H} \mathbb{E}_{d^{\pi_0}_h}\big[w^2_h \,\mathrm{Cov}_{P_h}(\delta^{\star_j}_h, \delta^{\star_k}_h \mid s_h, a_h)\big], \tag{8}$$

where $\delta^\star_h = f^\star_h + \gamma V^{\pi,\star}_{h+1}(s_{h+1}) - Q^{\pi,\star}_h(s_h, a_h)$*. Since the signals are deterministic functions of* (s_h, a_h)*, at the true* $Q^{\pi,\star}$ *the within-step covariance simplifies:*

$$\mathrm{Cov}_{P_h}(\delta^{\star_j}_h, \delta^{\star_k}_h \mid s_h, a_h) = \gamma^2 \,\mathrm{Cov}_{P_h}\big(V^{\pi,\star_j}_{h+1}(s'),\, V^{\pi,\star_k}_{h+1}(s') \mid s_h, a_h\big). \tag{9}$$

Reading (8)**: adding instead of multiplying.** No compounding term like the $\mathbb{E}_h[\gamma^2 \varrho_h^2 \operatorname{Var}_{h+1}[\cdot]]$ of Jiang & Li (2016, Thm. 1) appears in (8): w_h^2 multiplies only the *within-step* conditional covariance, and the steps are summed. Estimation error thus has exactly two sources: the randomness of the initial state (irreducible, zero when $\mu = \delta_{s_0}$) and that of each transition, reweighted by w_h^2 because the data come from π_0 while the target is π. (With stochastic signals (8) holds verbatim; Appendix L.)

The lower bound $\Sigma_\infty \succeq \Sigma^*$ for regular estimators is the convolution theorem (Corollary 10); a local asymptotic minimax version for all estimator sequences is Theorem 17 (Appendix T).

4 Joint vs. Independent Estimation: Calibration of Lagrangian Inference

Three ways to report uncertainty. All three attach an interval to the *same* point estimate $\hat{\theta}$ and differ only in the variance they report for the Lagrangian functional $\hat{L}_\lambda = \hat{v} - \lambda \hat{c} = e_\lambda^\top \hat{\theta}$ (single constraint, $\lambda > 0$; for $J > 1$ see Remark 19). *(i) Joint:* report $e_\lambda^\top \hat{\Sigma} e_\lambda$, the correct asymptotic variance of Theorem 6. *(ii) Bonferroni:* componentwise intervals at level $1 - \alpha/2$ combined by the union bound, half-width $z_{1-\alpha/4}\big(\hat{\Sigma}_{00}^{1/2} + \lambda \hat{\Sigma}_{11}^{1/2}\big)/\sqrt{n}$: valid whatever Σ_{01}^* is, at a price in width. *(iii) Diagonal:* report $e_\lambda^\top \operatorname{diag}(\hat{\Sigma}) e_\lambda = \hat{\Sigma}_{00} + \lambda^2 \hat{\Sigma}_{11}$, as though the components were independent. Only (iii) is unsound, and Theorem 5 says exactly when: for $\lambda > 0$ it over-states the variance when $\Sigma_{01}^* > 0$ (merely conservative) and *under*-states it when $\Sigma_{01}^* < 0$ (anti-conservative, hence unsafe). The joint covariance is a calibrated report of the same estimator, not a more efficient one (Remark 4, Appendix F). The substantive comparison is thus (i) vs. (ii). Both are valid, and the joint report is strictly tighter for every value of Σ_{01}^*: Cauchy–Schwarz gives $\sigma_\lambda \le \Sigma_{00}^{*\,1/2} + \lambda \Sigma_{11}^{*\,1/2}$, and $z_{1-\alpha/2} < z_{1-\alpha/4}$. The covariance it exploits is generically non-zero by Proposition 4. Experiments 7, 8 and 12 measure all three reports (Table 1).

Proposition 4 (Non-zero off-diagonal covariance; proof in Appendix N)**.** *Write* $\Sigma_{jk}^* = \operatorname{Cov}_\mu(V_1^{\pi,\star_j}, V_1^{\pi,\star_k}) + \gamma^2 \sum_{h<H} \mathbb{E}_{d_h^\pi}[w_h\, C_h(s_h, a_h)]$, *with* $C_h(s,a) := \operatorname{Cov}_{P_h(\cdot|s,a)}(V_{h+1}^{\pi,\star_j}, V_{h+1}^{\pi,\star_k})$. *(a) If the initial-state covariance and every* C_h *(d_h^π-a.e.) share a common sign, at least one non-zero on a set of positive mass, then* $\Sigma_{jk}^* \neq 0$ *with that sign. (b) In the tabular case (*$|\mathcal{S}| \ge 2$*,* $H \ge 2$*;* μ, π, π_0 *fixed),* $\Sigma_{jk}^* \neq 0$ *for all full-support kernels and signal functions outside a Lebesgue-null set. Without (a), opposite-sign terms—across steps or within one step—can cancel exactly.*

Theorem 5 (Calibration of Lagrangian inference, single constraint; proof in Appendix O)**.** *Consider the single-constraint case (*$J = 1$*) with Lagrangian multiplier* $\lambda > 0$ *and* $e_\lambda = (1, -\lambda)^\top$*. Using the* 0*-indexed notation of* (8) *(*Σ_{00}^* *= reward variance,* Σ_{01}^* *= reward-cost covariance,* Σ_{11}^* *= cost variance):*

(a) CDR Lagrangian variance: $\sigma_\lambda^2 = \Sigma_{00}^* - 2\lambda \Sigma_{01}^* + \lambda^2 \Sigma_{11}^*$.

(b) Diagonal (independence) report: $\sigma_{\lambda,\mathrm{indep}}^2 := e_\lambda^\top \operatorname{diag}(\Sigma^*) e_\lambda = \Sigma_{00}^* + \lambda^2 \Sigma_{11}^*$*, the variance reported for* $\hat{L}_\lambda$ *when* Σ_{01}^* *is dropped (not the variance of any estimator).*

(c) When $\Sigma_{01}^* > 0$ *and* $\lambda > 0$*:*

$$\frac{\sigma_\lambda^2}{\sigma_{\lambda,\mathrm{indep}}^2} = 1 - \frac{2\lambda \Sigma_{01}^*}{\Sigma_{00}^* + \lambda^2 \Sigma_{11}^*} < 1. \tag{10}$$

When $\Sigma_{01}^* < 0$ *the inequality reverses: the independent report then* under-states *the true variance* $e_\lambda^\top \Sigma^* e_\lambda$ *and its inference is anti-conservative, while the joint report is valid for either sign.*

Part (d) of the theorem, giving the variance-minimizing multiplier $\lambda^\dagger = \Sigma_{01}^*/\Sigma_{11}^*$ *(when* $\Sigma_{01}^* > 0$*) and the resulting* $\sigma_{\lambda^\dagger}^2 = \Sigma_{00}^*(1 - \rho^2)$ *with* $\rho := \Sigma_{01}^*/\sqrt{\Sigma_{00}^* \Sigma_{11}^*}$*, is stated in Appendix C.*

The corresponding regret statement (Theorem 11) and the confidence-ellipsoid volume comparison are in Appendix C.

5 Asymptotic Theory and Algorithm

5.1 Cross-Fitted CDR Estimator

We estimate $\hat{Q}$ and the shared $\hat{w}$ on $K-1$ folds and evaluate the score on the held-out fold, cycling over folds; the estimator is the closed-form sample average of the held-out scores (Remark 3), with pseudocode in Algorithm 1 (Appendix C).

Assumption 3 (Consistency and product rates)**.** *Let* $\zeta_w(n) := \max_{h,k} \|\hat{w}_h^{(k)} - w_h\|_{L^2(d_h^{\pi_0})}$ *and* $\zeta_{\star_j}(n) := \max_{h,k} \|\hat{Q}_h^{\star_j,(k)} - Q_h^{\pi,\star_j}\|_{L^2(d_h^{\pi_0})}$*. Assume (i)* $\zeta_w(n) = o_p(1)$ *and* $\max_{0\le j\le J} \zeta_{\star_j}(n) = o_p(1)$*, and (ii)* $\zeta_w(n) \cdot \max_{0\le j\le J} \zeta_{\star_j}(n) = o_p(n^{-1/2})$*, the standard debiased-ML condition (Chernozhukov et al., 2018); sufficient conditions and the equivalent sum form are in Appendix C.*

Assumption 4 (Bounded nuisance estimates)**.** $\|\hat{w}_h^{(k)}\|_\infty \le M$*,* $\|\hat{Q}_h^{\star,(k)}\|_\infty \le M$ *w.p.* $\to 1$*, for a constant* $M \ge \max\{H, \sup_h \|w_h\|_\infty\}$ *(at* $\gamma = 1$*,* $0 \le Q_h^{\pi,\star} \le H - h + 1$*); this strengthens Assumption 1 to marginal overlap and is enforced by clipping* $\hat{w}_h^{(k)}$ *at* M *and* $\hat{Q}_h^{\star,(k)}$ *to* $[0, H-h+1]$*.*

Theorem 6 (Joint asymptotic normality and semiparametric efficiency; proof in Appendix M)**.** *Under Assumptions 1–4 and the continuity condition in Appendix R (Assumption 7):*

$$\sqrt{n}(\hat{\theta} - \theta^*) \xrightarrow{d} \mathcal{N}(0, \Sigma^*) \quad \text{in } \mathbb{R}^{J+1}. \tag{11}$$

Moreover, $\hat{\theta}$ *is* semiparametrically efficient*: no regular estimator of* θ^* *in the time-inhomogeneous CMDP model has asymptotic covariance smaller than* Σ^* *in the Löwner order (Theorem 3).*

Corollary 7 (Inference)**.** *Under the assumptions of Theorem 6,* $\hat{\Sigma} = \frac{1}{n}\sum_i \hat{\psi}_i \hat{\psi}_i^\top \xrightarrow{p} \Sigma^*$ *with* $\hat{\psi}_i := \psi(Z_i; \hat{\theta}, \hat{\eta}^{(k(i))})$ *the held-out scores (the sandwich form simplifies since* $\partial_\theta \psi = -I$*), and if* $\Sigma^* \succ 0$ *(Remark 7) then* $\{\vartheta : n(\hat{\theta} - \vartheta)^\top \hat{\Sigma}^{-1}(\hat{\theta} - \vartheta) \le \chi^2_{J+1,1-\alpha}\}$ *is an asymptotic* $(1-\alpha)$ *confidence ellipsoid for* θ^**.*

6 Application: Constrained Offline Policy Learning

Because CDR returns $\hat{\theta}$ and $\hat{\Sigma}$ from one cross-fitted pass, it can serve as the evaluation subroutine of constrained policy optimization: maximize $\hat{v}^\pi$ subject to $\hat{c}_j^\pi \le b_j - \epsilon_{n,j}$ over a policy class Π, with safety margins $\epsilon_{n,j} = O(n^{-1/2})$ chosen so that the estimated constraints imply the true ones with high probability. Under strict feasibility of the best feasible policy and a Rademacher bound on the CDR score class, this yields $O(1/\sqrt{n})$ reward regret with high-probability satisfaction of every true constraint (Theorem 8). Separately, for the Lagrangian relaxation at a fixed multiplier (Theorem 11), a Bernstein-type bound has variance term $\sup_\pi e_\lambda^\top \Sigma^*(\pi) e_\lambda$, which $\hat{\Sigma}$ estimates consistently and which a diagonal report over-states whenever $\Sigma^*_{01}(\pi) > 0$ for every π. Appendix B states the problem, assumptions and regret theorem, Appendix Q proves them, and Experiment 9 (Appendix G.3) illustrates both conclusions with a normal-approximation margin. This is an application, not the paper's main claim: the joint covariance buys the variance term of the Lagrangian bound, not a different learned policy (Experiment 9).

7 Experiments

Experiments 1–7 use synthetic tabular CMDPs, Experiment 8 a continuous state space with fully learned nuisances, and Experiment 12 the Taxi benchmark. Numbering follows the code; the main text presents Experiments 7, 8 and 12, and Appendix G has the rest. We compare four estimators. **CDR-exact** runs Algorithm 1 with $K=5$ folds and oracle nuisances, isolating the theory's asymptotic predictions from nuisance error. **CDR-data** is the deployable estimator: shrinkage-smoothed tabular fitted Q-evaluation (FQE) for $\hat{Q}$, and an empirical density ratio from an estimated transition kernel and initial distribution; it reads nothing from the environment beyond π_0, which OPE assumes known. **JL-Concat** is step-wise DR with cumulative IS weights (Jiang & Li, 2016), and **IS** is plain importance sampling.

Table 1: Coverage of the nominal-95% Lagrangian interval ($\hat{L}_\lambda = \hat{v} - \lambda\hat{c}$, λ=1) under the three reports of Section 4, and width relative to the joint report. Exp. 7: tabular, oracle nuisances, $\Sigma^*_{01} < 0$, 2000 replications (SE ≤ 0.008); Exp. 8: continuous state, learned nuisances, $\Sigma^*_{01} < 0$, 300 runs (SE ≤ 0.013); Exp. 12: Taxi-Safety, H=100, CDR-data, 300 runs (SE ≤ 0.018). Exp. 12 width ratios reflect CDR-data's plug-in $\hat{\Sigma}$ ($\hat{\rho} \approx -0.1$), not Σ^* ($\rho = -0.84$; CDR-exact's Bonferroni ratio is 1.15). Bold: more than two SEs below nominal.

	Exp. 7 (tabular, oracle)			Exp. 8 (continuous)			Exp. 12 (Taxi, CDR-data)		
n	500	1000	2000	500	2000	8000	500	2000	4000
Joint (full $\hat{\Sigma}$)	0.946	0.951	0.951	0.947	0.950	0.963	**0.910**	0.927	**0.897**
Bonferroni	0.983	0.987	0.985	0.993	0.987	0.980	1.000	0.997	0.997
Diagonal (independent)	**0.863**	**0.868**	**0.874**	**0.857**	**0.840**	**0.840**	**0.890**	**0.910**	**0.880**
Width ratio, Bonferroni	1.23	1.23	1.23	1.25	1.19	1.17	1.50	1.53	1.53
Width ratio, diagonal	0.77	0.77	0.77	0.79	0.74	0.73	0.93	0.95	0.96

Primary environment: GridWorld-Safety. A 4×5 grid ($|\mathcal{S}|$=20, $|\mathcal{A}|$=4, H=10, J=1) with ε-stochastic transitions (ε=0.3), designed so that $\Sigma^*_{01} > 0$: $\theta^* = (6.08, 6.00)^\top$, $\Sigma^*_{00} = 3.80$, $\Sigma^*_{11} = 7.25$, $\Sigma^*_{01} = 3.78$, $\rho = 0.72$.

7.1 Experiment 7: Anti-Conservative Inference When $\Sigma^*_{01} < 0$

This experiment isolates Theorem 5(c) in the cleanest setting: tabular, with oracle nuisances, so any difference is due to the covariance report alone. On a GridWorld variant with $\Sigma^*_{01} = -2.93$ ($\rho = -0.72$) and $\lambda = 1$, the diagonal report under-states the true Lagrangian variance 14.36 as 8.49, a 41% under-statement. Table 1 (left) shows the outcome over 2000 replications (binomial standard error, SE, 0.005): the joint interval is calibrated, Bonferroni over-covers at a constant 23% extra width, and the diagonal report under-covers by about eight points, eleven SEs below nominal, without improving in n: the defect is in the reported variance, not the sample size (Appendix G.2).

7.2 Experiment 8: Continuous State Space with Learned Nuisances

Tabular environments let the nuisances be estimated at a parametric rate, the regime in which Assumption 3 is least demanding. This experiment moves to a linear-Gaussian CMDP on $\mathcal{S} = \mathbb{R}^3$ ($|\mathcal{A}|$=4, H=10), where both nuisances must be learned by nonparametric regression. Rewards and costs are sigmoids of a linear index, anti-aligned so that $\Sigma^*_{01} < 0$; $\theta^* = (3.151, 6.607)^\top$ comes from a 2×10^6-trajectory rollout of π. Since visit counts are unavailable, w is learned through the recursion $w^s_h(s) = \mathbb{E}_{\pi_0}[\, w_{h-1}(s_{h-1}, a_{h-1}) \mid s_h = s\,]$, $w_h = w^s_h\, \pi/\pi_0$: one regression per step on behavior data alone (derived and validated in Appendix D).

Testing the product-rate condition. We run CDR with two nuisance classes, identical in every other respect, that sit on either side of Assumption 3 over the sample sizes studied; both are fixed bases with a fixed ridge, so neither has an approximation error that vanishes as $n \to \infty$. For the *rich* class (random Fourier features plus a low-order polynomial block) the bias induced by the product of the two nuisance errors stays well under the sampling standard deviation for every $n \leq 8000$, so the condition holds in effect over this range; the *coarse* class (small fixed random-feature basis, heavy fixed ridge) has a floor whose bias is comparable to the sampling error already at n=500 and dominates it thereafter. Figure 2 confirms the one-sided prediction in both directions (300 runs per cell, coverage SE 0.013; Table 2 in Appendix D has the full numbers). The rich variant's bias falls to $(-0.0016, -0.0011)$ at n=8000, its mean squared error (MSE) falls 4.9× from n=2000 to 8000 (the $O(1/n)$ rate), and its χ^2_2 coverage holds at 0.92–0.93. The coarse variant's bias is pinned near -0.088 at every n and its coverage collapses to 0: the interval shrinks around a center that does not move. Against the same learned $\hat{Q}$, JL-Concat has 35× CDR's MSE at n=8000 and IS is worse by four orders of magnitude. An external density-ratio estimator (minimax weight learning, Uehara et al. 2020) lowers the ellipsoid coverage, as its less accurate $\hat{w}$ predicts, but leaves the ordering of the three reports unchanged (Experiment 8b, Appendix D). Finally, Experiment 7 reproduces with *estimated* $\hat{\Sigma}$ in a continuous state space (Table 1, middle): the joint interval is at nominal, Bonferroni

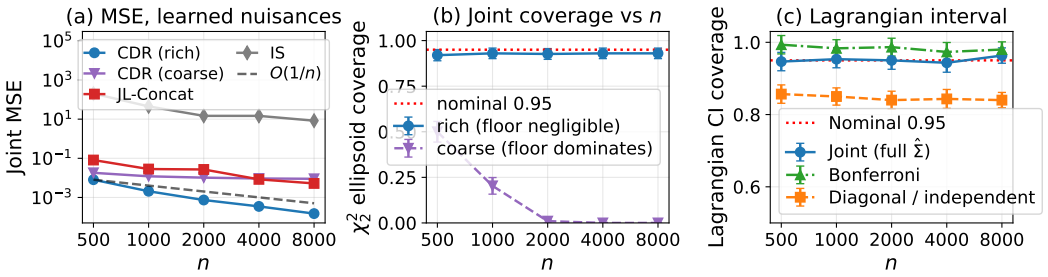


Figure 2: Experiment 8: continuous state space, nuisances learned by random-feature ridge regression (300 runs; ± 1.96 binomial SE). **(a)** CDR attains the $O(1/n)$ rate; JL-CONCAT and IS do not. **(b)** A nuisance class whose approximation floor is negligible at these n keeps coverage near nominal; one whose floor dominates the sampling error drives coverage to zero as the intervals shrink around a biased center. **(c)** Experiment 7 reproduces with estimated $\hat{\Sigma}$: joint at nominal, Bonferroni valid but wider, diagonal anti-conservative.

over-covers at 17–25% extra width, and the diagonal report under-covers at a flat 0.84–0.86: the failure comes from discarding $\hat{\Sigma}_{01}$, not from the tabular setting or oracle nuisances.

7.3 EXPERIMENT 12: TAXI-SAFETY AT HORIZON 100

Taxi is the standard benchmark of the marginalized-IS and DICE literature (Liu et al., 2018; Nachum et al., 2019; Uehara et al., 2020); we use the Gymnasium Taxi-v3 rules (500 states, 6 actions), made continuing and mildly stochastic (10% random moves), with reward $\mathbf{1}\{\text{drop-off}\}$, cost $\mathbf{1}\{\text{illegal pick-up/drop-off}\}$, $H{=}100$, and ϵ-greedy behavior and target policies ($\epsilon_0{=}0.3$, $\epsilon{=}0.1$). The per-step χ^2-divergence is 0.148, so the cumulative weight's second moment grows like $1.148^H \approx 10^6$, while $\max_h \|w_h\|_\infty = 9.5$. Everything is tabular ($\theta^* = (6.40, 2.98)$, $\rho = -0.84$; oracle nuisances exact; CDR-data uses tabular FQE and an estimated stationary kernel; 300 runs per cell; Figure 3, Appendix G.1). CDR-exact attains Σ^* ($n\cdot$Var within 0.84–$1.06\times \mathrm{tr}(\Sigma^*)$, χ^2_2 coverage 0.94–0.96, $\hat{\rho} = -0.836$ vs. -0.836). JL-Concat's $n \cdot$ Var is 10^3–$10^4\times$ larger and its MSE does not improve in n (Figure 3(b): the same linear-versus-exponential growth across $H \in \{10, \ldots, 100\}$). CDR-data has relative bias below 0.4% from $n{=}2000$ and plug-in χ^2_2 coverage of 0.91–0.95. At these n, however, CDR-data's cost covariance is still far from Σ^* ($\hat{\rho} \approx -0.1$ vs. -0.84; χ^2_2 coverage with the true Σ^* is 0.01–0.04); its plug-in intervals stay near nominal (0.90–0.95) because $\hat{\Sigma}$ tracks its larger variance (Appendix G.1). Because the cost is nearly deterministic given the state path ($\Sigma^*_{11} = 0.003$), Theorem 5 predicts only a 7% under-statement by the diagonal report at $\lambda{=}1$ (a coverage gap of ≈ 0.8 points); CDR-exact shows 0–2 points, while CDR-data's 2–5 point gap and $1.5\times$ Bonferroni width (Table 1, right) reflect its own $\hat{\Sigma}$, not Σ^*.

Additional experiments. Appendix G reports Experiments 1–6 and 9–11 (the $O(1/n)$ rate, the curse of horizon to $H{=}100$, five-seed coverage, double robustness, the variance-ratio sweep, $J{=}2$, constrained policy learning with a normal-approximation margin, the orthogonality check, and CartPole).

8 CONCLUSION

We proposed CDR, a jointly efficient OPE estimator for multi-constraint CMDPs. Because reward and cost are estimated from the same trajectories, their estimation errors covary (the shared w only reweights this covariance); reporting it is what calibrates Lagrangian inference, and its *sign* decides whether the diagonal (independence) report is merely conservative or actively unsafe. Limitations: Assumption 1 rules out deterministic behavior policies; all environments are simulated, Taxi being the only standard one; on a CartPole variant with sharp value functions the product-rate condition fails and CDR is biased while step-wise DR is not (Appendix G.11): the escape from the curse of horizon is only as good as the two nuisances; Section 6 is verified only on a finite policy class; infinite horizons and unknown thresholds remain open.

Reproducibility Statement

Every environment is simulated and fully specified in Section 7 and Appendices D and G. The supplementary material contains the code for all experiments (`files_code/`, whose `README` maps each experiment number to its script and result file), the master seeds, and the result files (`results/*.json`) from which every number and figure is produced. Tabular ground truths and efficiency bounds are exact; continuous-state ground truths use 2×10^6 Monte-Carlo trajectories with reported standard errors. The tabular experiments depend only on `numpy` and `scipy`.

AI Use Statement

Large language models (LLMs) were used to aid and polish the writing of this manuscript, and we disclose their role here in accordance with the ICLR 2027 AI policy. An LLM assistant was used to polish sentences, correct grammatical errors and improve wording and concision. The authors have reviewed all LLM-edited text and take full responsibility for the content of this paper.

References

Joshua Achiam, David Held, Aviv Tamar, and Pieter Abbeel. Constrained policy optimization. In *ICML*, pp. 22–31, 2017.

Eitan Altman. *Constrained Markov Decision Processes*. CRC Press, 1999.

Andrew Bennett, Nathan Kallus, Miruna Oprescu, Wen Sun, and Kaiwen Wang. Efficient and sharp off-policy evaluation in robust Markov decision processes. In *NeurIPS*, 2024.

Peter J. Bickel, Chris A. J. Klaassen, Ya'acov Ritov, and Jon A. Wellner. *Efficient and Adaptive Estimation for Semiparametric Models*. Johns Hopkins University Press, 1993.

Olivier Bousquet. A Bennett concentration inequality and its application to suprema of empirical processes. *Comptes Rendus Mathematique*, 334(6):495–500, 2002.

Yassine Chemingui, Aryan Deshwal, Alan Fern, Thanh Nguyen-Tang, and Janardhan Rao Doppa. Online optimization for offline safe reinforcement learning. In *NeurIPS*, 2025.

Claire Chen, Shuze Daniel Liu, and Shangtong Zhang. Efficient off-policy evaluation with safety constraint for reinforcement learning. In *ICLR*, 2025.

Victor Chernozhukov, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins. Double/Debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, 21(1):C1–C68, 2018.

Bo Dai, Ofir Nachum, Yinlam Chow, Lihong Li, Csaba Szepesvári, and Dale Schuurmans. CoinDICE: Off-policy confidence interval estimation. In *NeurIPS*, 2020.

Miroslav Dudík, John Langford, and Lihong Li. Doubly robust policy evaluation and learning. In *ICML*, pp. 1097–1104, 2011.

Yihao Feng, Tongzheng Ren, Ziyang Tang, and Qiang Liu. Accountable off-policy evaluation with kernel Bellman statistics. In *ICML*, 2020.

Javier García and Fernando Fernández. A comprehensive survey on safe reinforcement learning. *Journal of Machine Learning Research*, 16(1):1437–1480, 2015.

Botao Hao, Xiang Ji, Yaqi Duan, Hao Lu, Csaba Szepesvári, and Mengdi Wang. Bootstrapping fitted Q-evaluation for off-policy inference. In *ICML*, 2021.

Kihyuk Hong, Yuhang Li, and Ambuj Tewari. A primal-dual-critic algorithm for offline constrained reinforcement learning. In *AISTATS*, 2024.

Nan Jiang and Jiawei Huang. Minimax value interval for off-policy evaluation and policy optimization. In *NeurIPS*, 2020.

Nan Jiang and Lihong Li. Doubly robust off-policy value evaluation for reinforcement learning. In *ICML*, pp. 652–661, 2016.

Nathan Kallus and Masatoshi Uehara. Double reinforcement learning for efficient off-policy evaluation in Markov decision processes. *Journal of Machine Learning Research*, 21(167):1–63, 2020.

Nathan Kallus and Masatoshi Uehara. Efficiently breaking the curse of horizon in off-policy evaluation with double reinforcement learning. *Operations Research*, 70(6):3282–3302, 2022.

B Ravi Kiran, Ibrahim Sobh, Victor Talpaert, Patrick Mannion, Ahmad A. Al Sallab, Senthil Yogamani, and Patrick Pérez. Deep reinforcement learning for autonomous driving: A survey. *IEEE Transactions on Intelligent Transportation Systems*, 23(6):4909–4926, 2022.

Hoang M Le, Cameron Voloshin, and Yisong Yue. Batch policy learning under constraints. In *ICML*, pp. 3703–3712, 2019.

Jongmin Lee, Cosmin Paduraru, Daniel J Mankowitz, Nicolas Heess, Doina Precup, Kee-Eung Kim, and Arthur Guez. COptiDICE: Offline constrained reinforcement learning via stationary distribution correction estimation. In *ICLR*, 2022.

Lihong Li, Wei Chu, John Langford, and Robert E Schapire. A contextual-bandit approach to personalized news article recommendation. In *WWW*, pp. 661–670, 2010.

Qiang Liu, Lihong Li, Ziyang Tang, and Dengyong Zhou. Breaking the curse of horizon: Infinite-horizon off-policy estimation. In *NeurIPS*, 2018.

Zuxin Liu, Zijian Guo, Yihang Yao, Zhepeng Cen, Wenhao Yu, Tingnan Zhang, and Ding Zhao. Constrained decision transformer for offline safe reinforcement learning. In *ICML*, pp. 21611–21630, 2023.

Susan A Murphy, Mark J van der Laan, and James M Robins. Marginal mean models for dynamic regimes. *Journal of the American Statistical Association*, 96(456):1410–1423, 2001.

Ofir Nachum, Yinlam Chow, Bo Dai, and Lihong Li. DualDICE: Behavior-agnostic estimation of discounted stationary distribution corrections. In *NeurIPS*, 2019.

Whitney K. Newey. Semiparametric efficiency bounds. *Journal of Applied Econometrics*, 5(2): 99–135, 1990.

Nicholas Polosky, Bruno C. Da Silva, Madalina Fiterau, and Jithin Jagannath. Constrained offline policy optimization. In *ICML*, 2022.

Doina Precup, Richard S Sutton, and Satinder P Singh. Eligibility traces for off-policy policy evaluation. In *ICML*, pp. 759–766, 2000.

James M Robins and Andrea Rotnitzky. Semiparametric efficiency in multivariate regression models with missing data. *Journal of the American Statistical Association*, 90(429):122–129, 1995.

Chengchun Shi, Sheng Zhang, Wenbin Lu, and Rui Song. Statistical inference of the value function for reinforcement learning in infinite-horizon settings. *Journal of the Royal Statistical Society: Series B*, 84(3):765–793, 2022.

Chen Tessler, Daniel J. Mankowitz, and Shie Mannor. Reward constrained policy optimization. In *ICLR*, 2019.

Philip S. Thomas and Emma Brunskill. Data-efficient off-policy policy evaluation for reinforcement learning. In *ICML*, pp. 2139–2148, 2016.

Philip S. Thomas, Georgios Theocharous, and Mohammad Ghavamzadeh. High-confidence off-policy evaluation. In *AAAI*, 2015.

Anastasios A Tsiatis. *Semiparametric Theory and Missing Data*. Springer, 2006.

Masatoshi Uehara and Wen Sun. Pessimistic model-based offline reinforcement learning under partial coverage. In *ICLR*, 2022.

Masatoshi Uehara, Jiawei Huang, and Nan Jiang. Minimax weight and Q-function learning for off-policy evaluation. In *ICML*, 2020.

Aad W van der Vaart. *Asymptotic Statistics*. Cambridge University Press, 1998.

Martin J Wainwright. *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*. Cambridge University Press, 2019.

Haoyu Wei. Semiparametric off-policy inference for optimal policy values under possible non-uniqueness. *arXiv preprint arXiv:2505.13809*, 2025.

Tengyang Xie, Yifei Ma, and Yu-Xiang Wang. Towards optimal off-policy evaluation for reinforcement learning with marginalized importance sampling. In *NeurIPS*, 2019.

Haoran Xu, Xianyuan Zhan, and Xiangyu Zhu. Constraints penalized Q-learning for safe offline reinforcement learning. In *AAAI*, 2022.

Yinan Zheng, Jianxiong Li, Dongjie Yu, Yujie Yang, Shengbo Eben Li, Xianyuan Zhan, and Jingjing Liu. Safe offline reinforcement learning with feasibility-guided diffusion model. In *ICLR*, 2024.

A Extended Related Work

Related work. Kallus & Uehara (2020) achieved semiparametric efficiency for single-objective OPE; Kallus & Uehara (2022) extended this to infinite horizon. The marginal density-ratio nuisance $w = d^{\pi}/d^{\pi_0}$ originates in the marginalized importance-sampling line that first broke the curse of horizon (Liu et al., 2018) and was developed through the DICE family (Nachum et al., 2019; Dai et al., 2020) and minimax weight / Q-function learning (Uehara et al., 2020); high-confidence OPE (Thomas et al., 2015) pursues finite-sample safety guarantees that are complementary to our asymptotic confidence regions. We carry w into the constrained, multi-objective setting and characterize the cross-objective covariance structure of the joint EIF. The use of joint influence functions for vector-valued targets has a long history in semiparametric statistics (Tsiatis, 2006; van der Vaart, 1998); our contribution is the specialization to CMDPs, whose objectives are estimated from shared trajectories with a shared density-ratio nuisance, and the resulting characterization of how off-diagonal entries shape Lagrangian inference. Within offline safe RL, Le et al. (2019) reduced batch policy learning under constraints to a Lagrangian game solved with FQI best responses and exponentiated-gradient multiplier updates, with finite-sample FQE-based guarantees but no semiparametric efficiency analysis or asymptotic confidence region, and COptiDICE (Lee et al., 2022), FISOR (Zheng et al., 2024), CDT (Liu et al., 2023), and O3SRL (Chemingui et al., 2025) optimize policies directly; none produces a $\sqrt{n}$-consistent estimator of the joint target with a valid asymptotic confidence region, which is what a feasibility certificate requires. Dai et al. (2020) construct off-policy confidence intervals for single objectives via the DICE framework, as do Jiang & Huang (2020) through minimax value intervals, Feng et al. (2020) through kernel Bellman statistics, Hao et al. (2021) through a bootstrap of fitted Q-evaluation and Shi et al. (2022) through sieve estimation of the value function; Thomas et al. (2015) pursue finite-sample safety guarantees complementary to our asymptotic regions. All are single-objective and therefore silent on the cross-objective covariance. Finally, Wei (2025) target optimal policy values under non-uniqueness rather than a fixed policy, Bennett et al. (2024) consider robust transition perturbations rather than cost constraints, and Uehara & Sun (2022) is single-objective and model-based.

A closely related line, and how it differs. One recent work shares ingredients with ours and is worth separating explicitly. Chen et al. (2025) also pair variance-efficient policy evaluation with a safety constraint, but the constraint acts on the *data-collection* problem: they design an optimal behavior policy that minimizes evaluation variance subject to remaining safe to execute. That is an experimental-design question, and it presumes the ability to choose π_0. We take π_0 as given—the defining feature of the offline setting, where the log already exists—and the constraint in our problem is a property of the *target* policy being certified, not of the policy generating the data. The two are complementary: their result says how to collect, ours how to certify from what was collected.

B Deferred Details of Section 6

Let $\pi^* \in \arg\max_{\pi\in\Pi}\{v^\pi : c_j^\pi \le b_j \ \forall j\}$ denote the best truly feasible policy in Π (assumed to exist, e.g. Π finite or compact). The learning rule is

$$\hat{\pi} = \arg\max_{\pi\in\Pi} \hat{v}^\pi \quad \text{s.t.} \quad \hat{c}_j^\pi \le b_j - \epsilon_{n,j}, \ \ j = 1, \dots, J, \tag{12}$$

with safety margins

$$\epsilon_{n,j} = \frac{2\sqrt{K}\, L\kappa}{\sqrt{n}} + C\sqrt{\frac{\log(4JK/\delta)}{n}}, \tag{13}$$

where κ and L are the constants of Assumption 6, K is the number of cross-fitting folds, and $C = \sqrt{2K}\, B_j + 1$ with $B_j := M + HM(1+2M)$ the boundedness constant of the j-th cost score (Appendix Q); the additive 1 absorbs the cross-fitted nuisance bias.

Assumption 5 (Strict feasibility of the comparator)**.** *$\exists\, \alpha_S > 0$: $c_j^{\pi^*} \le b_j - \alpha_S$ for all j.*

This is stronger than Slater's condition, which asks only for *some* strictly feasible policy in Π: it excludes constraints that are active at π^*. It holds generically for finite Π (as in Experiment 9); for a class closed under occupancy-measure mixtures it forces π^* to be the unconstrained maximizer.

Assumption 6 (CDR score class complexity)**.** *Let $\mathcal{R}_m(\mathcal{F}) := \mathbb{E}\sup_{f\in\mathcal{F}} |m^{-1}\sum_{i\le m}\varepsilon_i f(Z_i)|$. For each fold k and direction $v \in \mathbb{R}^{J+1}$, conditional on $\hat{\eta}^{(k)}$ (on the event where Assumption 4 holds), the class $\mathcal{F}_v^{(k)} = \{Z \mapsto v^\top \psi(Z;\pi,\hat{\eta}^{(k)}) : \pi \in \Pi\}$ satisfies $\mathcal{R}_m(\mathcal{F}_v^{(k)}) \le L\|v\|_1\kappa/\sqrt{m}$ for all m, for constants $\kappa > 0$, $L \ge 1$ (under condition (i) of Remark 15, $L = 1$ and $\kappa = B_0\sqrt{2\log(2|\Pi|)}$; under condition (iii), as given there).*

Remark 1 (Binding constraints)**.** *If instead Π is closed under occupancy-measure mixtures, Assumption 6 holds on this closure, and some $\bar{\pi} \in \Pi$ has $c_j^{\bar{\pi}} \le b_j - \alpha_S$ for all j, then with $t = 2\max_j \epsilon_{n,j}/\alpha_S \le 1$ the mixture $\pi_t = (1-t)\pi^* \oplus t\bar{\pi}$ is feasible for* (12) *on the same event, and $v^{\pi^*} - v^{\hat{\pi}} \le (C_1 L\kappa + C_2\sqrt{\log(4K/\delta)})/\sqrt{n} + 2H\max_j \epsilon_{n,j}/\alpha_S$. The extra $2H\max_j \epsilon_{n,j}/\alpha_S$ is the price of a genuinely binding constraint; Theorem 8 avoids it only through Assumption 5.*

Theorem 8 (Regret and constraint satisfaction)**.** *Under Assumptions 1–6 with safety margins* (13)*, for n sufficiently large that $2\epsilon_{n,j}(\delta) \le \alpha_S$ for all j (where α_S is the feasibility slack of π^* from Assumption 5), with probability at least $1 - \delta - \mathbb{P}(\mathcal{E}_n^c)$:*

$$v^{\pi^*} - v^{\hat{\pi}} \le \frac{C_1\, L\,\kappa + C_2\sqrt{\log(4K/\delta)}}{\sqrt{n}}, \qquad c_j^{\hat{\pi}} \le b_j \ \ \forall\, j, \tag{14}$$

where $\mathcal{E}_n := \{\beta_n \le n^{-1/2}\}$ is the event that the cross-fitted nuisance bias is at most $n^{-1/2}$ uniformly over Π, with $\beta_n := \sup_{\pi\in\Pi}\|\bar{\theta}^\pi - \theta^\pi\|_\infty$ and $\bar{\theta}^\pi := \sum_k \frac{|I_k|}{n}\mathbb{E}_{Z\sim P}[\psi^\circ(Z;\pi,\hat{\eta}^{(k)})]$ the fold-averaged population mean of the score at the fitted nuisances (Z a fresh draw, $\hat{\eta}^{(k)}$ fixed), so that $\mathbb{P}(\mathcal{E}_n^c) \to 0$ under Assumption 3 applied uniformly over Π, i.e. $\sup_\pi \zeta_w = o_p(1)$, $\sup_\pi \max_j \zeta_{\star_j} = o_p(1)$ and $\sup_\pi \zeta_w \cdot \max_j \zeta_{\star_j} = o_p(n^{-1/2})$ (this uniform rate holds, e.g., in the tabular case, since $\sup_\pi \|\hat{Q}^\pi - Q^\pi\|_\infty \le HM\max_h \|\hat{P}_h - P_h\|_{1,\infty}$ and $\sup_\pi \|\hat{w}^\pi - w^\pi\|$ are controlled by π-free model errors); L is the constant of Assumption 6; $C_1 = 4\sqrt{K}$; and $C_2 = 2\sqrt{2K}\, B_0 + 2$, where $B_0 := M + HMB$ with $B := 1 + 2M$ is the boundedness constant for the reward score (see Appendix Q).

Proof. Deferred to Appendix Q. □

Remark 2 (Comparison with variance-based margins)**.** *Let $\bar{\sigma}_j^2 := \sup_{\pi\in\Pi}\Sigma_{jj}^*(\pi)$. A purely variance-based margin $C_0\sqrt{\bar{\sigma}_j^2\log(4J/\delta)/n}$ with no Rademacher term is justified only if its excess over $\sqrt{2\bar{\sigma}_j^2\log(4J/\delta)/n}$ (so necessarily $C_0 > \sqrt{2}$) dominates $2\sqrt{K}L\kappa/\sqrt{n}$; we do not assume this. The margin* (13) *needs neither a variance estimate nor such a condition: under the assumptions of Theorem 8 it bounds $\sup_\pi |\hat{c}_j^\pi - c_j^\pi|$ on $\mathcal{E}_n$, simultaneously for all j, with probability at least $1 - \delta/2 - \mathbb{P}(\mathcal{E}_n^c)$.*

Algorithm 1 Cross-Fitted CDR (CF-CDR)

Require: offline data $\mathcal{D} = \{Z_i\}_{i=1}^n$, target policy π, horizon H, number of folds K
Initialization: partition $\mathcal{D}$ into K disjoint folds $\{I_k\}_{k=1}^K$; write $\mathcal{D}_{I_k} \coloneqq \{Z_i : i \in I_k\}$ and let $k(i)$ denote the fold containing Z_i
// Stage 1: cross-fitted nuisance estimation
for $k = 1$ to K **do**
    On $\mathcal{D} \setminus \mathcal{D}_{I_k}$, fit the Q-functions $\hat{Q}^{\star,(k)} = \{\hat{Q}_h^{\star,(k)}\}_{h=1}^H$ for every $\star \in \{r, c_1, \ldots, c_J\}$
    On $\mathcal{D} \setminus \mathcal{D}_{I_k}$, fit the shared density ratio $\hat{w}^{(k)} = \{\hat{w}_h^{(k)}\}_{h=1}^H$
end for
Set $\hat{V}_h^{\star,(k)}(s) \coloneqq \sum_a \pi_h(a|s)\hat{Q}_h^{\star,(k)}(s,a)$, $\hat{V}_{H+1}^{\star,(k)} \equiv 0$, and $\hat{\delta}_{h,i}^{\star,(k)} \coloneqq f_h^\star(s_{h,i}, a_{h,i}) + \gamma \hat{V}_{h+1}^{\star,(k)}(s_{h+1,i}) - \hat{Q}_h^{\star,(k)}(s_{h,i}, a_{h,i})$
// Stage 2: closed-form estimate (see Remark 3)
for $j = 0, 1, \ldots, J$ **do**
    $\hat{\theta}_j \leftarrow \frac{1}{n}\sum_{i=1}^n \Big[\hat{V}_1^{\pi,\star_j,(k(i))}(s_{1,i}) + \sum_{h=1}^H \hat{w}_h^{(k(i))}(s_{h,i}, a_{h,i})\, \hat{\delta}_{h,i}^{\star_j,(k(i))}\Big]$
end for
// Stage 3: joint covariance (same pass)
for $i = 1$ to n **do**
    $\hat{\psi}_i \leftarrow \big(\hat{V}_1^{\pi,\star_j,(k(i))}(s_{1,i}) + \sum_{h=1}^H \hat{w}_h^{(k(i))}(s_{h,i}, a_{h,i})\, \hat{\delta}_{h,i}^{\star_j,(k(i))} - \hat{\theta}_j\big)_{j=0}^J$
end for
$\hat{\Sigma} \leftarrow \frac{1}{n}\sum_{i=1}^n \hat{\psi}_i \hat{\psi}_i^\top$
return $\hat{\theta} \in \mathbb{R}^{J+1}$ and $\hat{\Sigma} \in \mathbb{R}^{(J+1)\times(J+1)}$

C Secondary Theoretical Results

Theorem 5(d): the variance-minimizing multiplier.

- The *variance-minimizing* multiplier over $\lambda \in \mathbb{R}_+$ is

$$\lambda^\dagger \coloneqq \arg\min_{\lambda \geq 0} \sigma_\lambda^2 = \frac{\Sigma_{01}^*}{\Sigma_{11}^*} \quad (\text{when } \Sigma_{01}^* > 0).$$

  At this choice, $\sigma_{\lambda^\dagger}^2 = \Sigma_{00}^*(1-\rho^2)$, where $\rho = \Sigma_{01}^*/\sqrt{\Sigma_{00}^*\Sigma_{11}^*}$. At the dual-optimal $\lambda^* > 0$, the variance ratio is

$$\frac{\sigma_{\lambda^*}^2}{\sigma_{\lambda^*,\text{indep}}^2} = 1 - \frac{2\lambda^*\Sigma_{01}^*}{\Sigma_{00}^* + (\lambda^*)^2\Sigma_{11}^*} \in [1-\rho,\, 1),$$

  when $\Sigma_{01}^* > 0$ and $\lambda^* > 0$; hence $\in (0,1)$ when $|\rho| < 1$ (i.e. $\Sigma^* \succ 0$).

Theorem 9 (Neyman orthogonality)**.** *Under the convention $\gamma = 1$ (for $\gamma < 1$ see Proposition 14), the joint CDR score satisfies $\partial_\epsilon \mathbb{E}[\psi(Z; \theta^*, \eta^* + \epsilon\,\Delta\eta)]|_{\epsilon=0} = \mathbf{0}$ for all perturbation directions $\Delta\eta = (\Delta Q^r, \Delta Q^{c_1}, \ldots, \Delta Q^{c_J}, \Delta w)$.*

Proof. Deferred to Appendix K. □

Corollary 10 (Convolution lower bound for regular estimators)**.** *At $\gamma = 1$, let $\hat{\theta}_n$ be regular at P relative to the tangent space $\dot{\mathcal{P}}$ of the nonparametric, time-inhomogeneous CMDP model, with $\sqrt{n}(\hat{\theta}_n - \theta^*) \rightsquigarrow L$. Then $L = \mathcal{N}(0, \Sigma^*) * G$ for some distribution G; in particular, if L has covariance Σ_∞, then $\Sigma_\infty \succeq \Sigma^*$ in the Löwner order.*

Proof. Deferred to Appendix T. □

Remark 3 (Closed-form solution)**.** *Since $\psi(Z; \theta, \eta) = \psi^\circ(Z; \eta) - \theta$ is linear in θ, where ψ° collects the θ-free part of the score, the moment equation $\frac{1}{n}\sum_i \psi(Z_i; \hat{\theta}, \hat{\eta}^{(k(i))}) = \mathbf{0}$ has the closed-form solution $\hat{\theta}_j = \frac{1}{n}\sum_i \psi_j^\circ(Z_i; \hat{\eta}^{(k(i))})$: a one-step debiased sample average, with no root-finding.*

Sufficient conditions for the product-rate condition. For fixed J, the max form of Assumption 3 and the sum form $\zeta_w \cdot \sum_j \zeta_{\star_j} = o_p(n^{-1/2})$ are equivalent up to a factor of $(J+1)$; we state the max form to avoid suggesting that all components must simultaneously achieve faster rates as J grows. The product-rate condition is satisfied, for example, when: (i) $\mathcal{S}$ and $\mathcal{A}$ are finite and $\hat{w}$, $\hat{Q}$ are estimated by tabular MLE ($\zeta = O(n^{-1/2})$, so the product is $O(n^{-1})$); (ii) the nuisance functions lie in a reproducing kernel Hilbert space of smoothness s (where $d = \dim(\mathcal{S} \times \mathcal{A})$), giving $\zeta = O(n^{-s/(2s+d)})$ and a product of $o(n^{-1/2})$ as soon as $s > d/2$; or (iii) $\hat{w}$ and $\hat{Q}$ are estimated by neural networks attaining $\zeta = o_p(n^{-1/4})$ (e.g. $O_p(n^{-1/4-\epsilon})$ for some $\epsilon > 0$), so the product is $o_p(n^{-1/2})$; it also suffices that one rate is $o_p(n^{-1/4})$ and the other $O_p(n^{-1/4})$.[1]

These results are referenced from Sections 4 and 6.

Theorem 11 (Variance term of the regret bound)**.** *Let $J = 1$, fix $\lambda \geq 0$ (not data-dependent), let $L_\lambda(\pi) := e_\lambda^\top \theta(\pi) + \lambda b$, $\hat{L}_\lambda(\pi) := e_\lambda^\top \hat{\theta}(\pi) + \lambda b$, $\hat{\pi}_\lambda \in \arg\max_\Pi \hat{L}_\lambda$ and $\pi^*_\lambda \in \arg\max_\Pi L_\lambda$. Under Assumptions 1, 2, 3, 4, 6 and 7, for all sufficiently large n and with probability at least $1 - \delta - o(1)$, the CDR-based Lagrangian optimizer achieves regret*

$$L_\lambda(\pi^*_\lambda) - L_\lambda(\hat{\pi}_\lambda) \leq \frac{C_1 L\kappa + C_2\sqrt{\bar{\sigma}^2_\lambda \log(2/\delta)}}{\sqrt{n}} + \frac{C_3 \log(2/\delta)}{n}, \tag{15}$$

with $C_1 = 8\sqrt{K}(1+\lambda)$ (K the number of folds, cf. Appendix Q), any $C_2 > 2\sqrt{2K}$ for $n \geq n_0$, and $C_3 = c\,K(1+\lambda)(M + HMB)$ for an absolute constant c, where $\bar{\sigma}^2_\lambda := \sup_{\pi\in\Pi} e_\lambda^\top \Sigma^*(\pi) e_\lambda$ is the policy-class-supremum Lagrangian variance under the (correct) joint covariance, and $\bar{\sigma}^2_{\lambda,\mathrm{indep}} := \sup_{\pi\in\Pi} e_\lambda^\top \mathrm{diag}(\Sigma^*(\pi)) e_\lambda$ is the value an independent analysis would instead substitute for this variance term. (The variance that enters the concentration inequality is that of the cross-fitted score $e_\lambda^\top \hat{\psi}(Z;\pi)$, which converges to $e_\lambda^\top \Sigma^*(\pi) e_\lambda$ uniformly over Π under Assumptions 3 and 7; the $o(1)$ in the probability and the "sufficiently large n" absorb this convergence and the nuisance bias, exactly as in Theorem 8.) When $\Sigma^*_{01}(\pi) > 0$ for all $\pi \in \Pi$ and $\lambda > 0$, the diagonal report does not under-state the variance term pointwise, hence $\bar{\sigma}^2_\lambda \leq \bar{\sigma}^2_{\lambda,\mathrm{indep}}$; the inequality is strict if $\inf_{\pi\in\Pi} \Sigma^*_{01}(\pi) > 0$ (then $\bar{\sigma}^2_\lambda \leq \bar{\sigma}^2_{\lambda,\mathrm{indep}} - 2\lambda \inf_\pi \Sigma^*_{01}(\pi)$), or if the supremum defining $\bar{\sigma}^2_\lambda$ is attained (e.g. Π finite, or compact with $\pi \mapsto \Sigma^*(\pi)$ continuous). Pointwise in π, Theorem 5(d) gives $\sigma^2_{\lambda^\dagger(\pi)}(\pi)/\sigma^2_{\lambda^\dagger(\pi),\mathrm{indep}}(\pi) = (1-\rho(\pi)^2)/(1+\rho(\pi)^2)$ at $\lambda^\dagger(\pi) = \Sigma^*_{01}(\pi)/\Sigma^*_{11}(\pi)$, with $\rho(\pi)$ the reward-cost correlation; this is a per-policy statement ($\lambda^\dagger(\pi)$ varies with π, so it is not a fixed λ of (15)*) and does not bound $\bar{\sigma}^2_\lambda/\bar{\sigma}^2_{\lambda,\mathrm{indep}}$, which can be arbitrarily close to 1. This concerns only the variance component of the regret bound* (15) *as* reported*: the learned policy, its true regret, and the Rademacher term $C_1 L\kappa$ are the same under either report. The variance term $\sqrt{\bar{\sigma}^2_\lambda \log(2/\delta)/n}$ is of the same $n^{-1/2}$ order as the Rademacher term; it is not a second-order term.*

Proof. Deferred to Appendix P. □

Corollary 12 (Confidence ellipsoid volume)**.** *Let $\mathcal{E}_{\mathrm{joint}}$ be the ellipsoid of Corollary 7 and $\mathcal{E}_{\mathrm{indep}}$ the axis-aligned ellipsoid with the same $\chi^2_{J+1,1-\alpha}$ radius and $\mathrm{diag}(\hat{\Sigma})$ in place of $\hat{\Sigma}$ (the report of an analysis that ignores correlation; at conventional levels $\alpha \leq 0.1$ it under-covers whenever Σ^* is not diagonal, e.g. 0.931 at $\rho = \pm 0.72$, $\alpha = 0.05$). For $J = 1$ and each π, with $\rho(\pi) := \Sigma^*_{01}(\pi)/\sqrt{\Sigma^*_{00}(\pi)\Sigma^*_{11}(\pi)}$ and $\bar{\rho} := \inf_{\pi\in\Pi} |\rho(\pi)|$, $\mathrm{vol}(\mathcal{E}_{\mathrm{joint}})/\mathrm{vol}(\mathcal{E}_{\mathrm{indep}}) \xrightarrow{p} \sqrt{1-\rho(\pi)^2} \leq \sqrt{1-\bar{\rho}^2}$. For general J and $\Sigma^*(\pi) \succ 0$ the ratio converges to $\sqrt{\det \Sigma^* / \prod_j \Sigma^*_{jj}} \leq 1$ (Hadamard), with equality iff all off-diagonal entries vanish.*

D Continuous-State Experiment: Derivation and Implementation

The density-ratio recursion. Write $w^s_h(s) := d^\pi_h(s)/d^{\pi_0}_h(s)$, so that $w_h(s,a) = w^s_h(s)\,\pi(a|s)/\pi_0(a|s)$ since both occupancies factor the action through the respective policy. The state factor obeys the recursion

$$w^s_h(s) \;=\; \mathbb{E}_{\pi_0}\big[\, w_{h-1}(s_{h-1}, a_{h-1}) \,\big|\, s_h = s \,\big], \qquad w^s_1 \equiv 1, \tag{16}$$

[1] A rate of exactly $O_p(n^{-1/4})$ for both nuisances gives $O_p(n^{-1/2})$, which does not suffice.

Table 2: Experiment 8: continuous state space ($\mathcal{S} = \mathbb{R}^3$), all nuisances learned. 300 runs per cell; coverage standard error ≈ 0.013. "Rich" and "coarse" differ only in the nuisance function class.

	Joint MSE			χ^2_2 coverage		Lagrangian coverage	
n	CDR (rich)	JL-CONCAT	IS	rich	coarse	joint	diagonal
500	0.0081	0.082	202.2	0.920	0.500	0.947	0.857
1000	0.0021	0.028	44.8	0.930	0.203	0.953	0.850
2000	0.00074	0.027	14.7	0.927	0.010	0.950	0.840
4000	0.00035	0.0084	14.7	0.930	0.000	0.943	0.843
8000	0.00015	0.0052	8.3	0.930	0.000	0.963	0.840

which we now derive. For any test function f, one step of the occupancy recursion followed by a change of measure gives

$$\begin{aligned}\mathbb{E}_{d_h^\pi}[f(s_h)] &= \mathbb{E}_{d_{h-1}^\pi}\big[\mathbb{E}_{s'\sim P_{h-1}(\cdot|s,a)}[f(s')]\big] \\ &= \mathbb{E}_{d_{h-1}^{\pi_0}}\big[w_{h-1}(s,a)\,\mathbb{E}_{s'\sim P_{h-1}(\cdot|s,a)}[f(s')]\big] \;=\; \mathbb{E}_{\pi_0}\big[w_{h-1}(s_{h-1},a_{h-1})\,f(s_h)\big],\end{aligned} \tag{17}$$

where the last equality replaces the inner conditional expectation by the *observed* next state, which is legitimate because (s_{h-1}, a_{h-1})-measurable weights pass outside $\mathbb{E}_h[\cdot]$. The left side is also $\mathbb{E}_{d_h^{\pi_0}}[w_h^s(s_h)f(s_h)]$ by definition of w_h^s. Since (17) holds for every f, the two integrands agree $d_h^{\pi_0}$-a.e., which is exactly (16): w_h^s is the $L^2(d_h^{\pi_0})$ projection of the previous step's weight onto $\sigma(s_h)$, i.e. an ordinary regression of w_{h-1} on s_h using behavior data only. Note the contrast with the cumulative ratio $\varrho_{1:h}$, which is the same quantity *without* the conditional expectation: taking $\mathbb{E}[\cdot|s_h]$ at every step is precisely the state-merging operation of Remark 5, and is what keeps the variance from compounding. We verified (16) numerically on random tabular chains, where both sides are computable in closed form; they agree to 2.2×10^{-16}.

Self-normalization. Nothing in the recursion forces $\hat{w}_h$ to have mean one. For the coarse class each ridge fit loses about 3–5% of its target's mean, so the unnormalized training-split mean of $\hat{w}_h$ decays from 1 at h=1 (where $\hat{w}_1 = \pi/\pi_0$ exactly and no regression is involved) to about 0.7 at h=10; for the rich class each fit loses only about 0.1%, but sampling error accumulates and by h=10 the mean drifted to between 0.82 and 1.12 across the training sets we checked. Because $\mathbb{E}_{d_h^{\pi_0}}[w_h] = 1$ is known exactly a priori, we rescale $\hat{w}_h$ at each step to satisfy it on the training split. This uses no additional data and no knowledge of the environment, and it restores the empirical mean to 1.000 at every step.

Nuisance classes. $\hat{Q}$ is fitted by exact backward FQE, one ridge regression per (step, action) cell, and $\hat{w}$ by (16), one ridge regression per step. The *rich* class uses 64 random Fourier features (bandwidth 1.0) concatenated with a constant, linear and quadratic block, and ridge parameter 10^{-3}; the polynomial block matters because the true value functions here are smooth functions of the state (sigmoids of a linear index smoothed by Gaussian transitions) that a low-order polynomial approximates well, and without it the isotropic random features leave a relative approximation bias of 4–5% that dominates the sampling error. The *coarse* class uses 32 random Fourier features at bandwidth 2.5 with ridge 3×10^{-2} and no polynomial block. Weights are clipped at 40; the clip is never active for the rich class at the sample sizes reported. All nuisances are cross-fitted with K=5 folds, and the same fitted nuisances are handed to JL-CONCAT, so the CDR/JL gap is attributable to the weighting scheme alone and not to a different $\hat{Q}$.

Neural nuisances. We also implement every regression above as a two-hidden-layer MLP (64 tanh units per layer, Adam, 150 epochs; `cdr_continuous.py`, `kind='mlp'`; reproduced by `files_code/exp8_mlp_check.py`), a fixed-architecture neural analogue of the ridge class (like it, an illustration of rather than an instance of the sufficient conditions). It costs roughly two orders of magnitude more per replication than the closed-form ridge (97s versus 0.8s), so we run it as a spot check at a single sample size rather than over the full grid. At n=2000 over 30 runs it gives bias $(-0.0007, -0.0059)$, joint MSE 0.00082 and χ^2_2 coverage 0.90 (binomial standard error

0.055), against $(-0.0018,\ -0.0021)$, 0.00074 and 0.927 for the ridge variant at the same n: the two agree to within Monte-Carlo error. This is the expected outcome and a useful negative result—the theory constrains the nuisance *rates*, not the function class, so switching classes should not move the answer, and it does not.

Experiment 8b: an external density-ratio estimator. The recursion above is one way to learn w; the DICE / minimax-weight family (Nachum et al., 2019; Uehara et al., 2020) is the other. We re-ran Experiment 8 (same environment, seeds and learned $\hat{Q}$) with w estimated by finite-horizon linear-class minimax weight learning (MWL): with $w_h(s,a) = \phi_a(s)^\top \beta_{h,a}$ in the span of the same random-feature basis and the discriminator class equal to that span, the MWL saddle point is the linear system $\mathbb{E}_{d_h^{\pi_0}}[w_h(s_h,a_h)\,\phi(s_h,a_h)] = \mathbb{E}_{\pi_0}[w_{h-1}(s_{h-1},a_{h-1})\,\phi_\pi(s_h)]$, with $\phi(s,a) \coloneqq e_a \otimes \phi_a(s)$ and $\phi_\pi(s) = \sum_a \pi(a|s)\phi(s,a)$, solved step by step from $w_1 = \pi/\pi_0$ with ridge 10^{-3} (the same as the recursion; not tuned for MWL), then renormalized and clipped exactly as the recursion is. At $n = 500/1000/2000/4000/8000$ the MWL variant has bias $(-0.11, -0.33)$, $(-0.04, -0.09)$, $(-0.013, -0.026)$, $(-0.005, -0.009)$, $(-0.003, -0.004)$—shrinking faster than $1/n$ (its norm falls from 0.35 to 0.005 as n grows sixteenfold) but 2–30$\times$ the recursion's—joint MSE $38\times$, $10\times$, $3\times$, $1.5\times$, $1.2\times$ the recursion's, and χ^2_2 coverage $0.63/0.56/0.65/0.70/0.69$ against the recursion's $0.92/0.93/0.93/0.93/0.93$: a less accurate $\hat{w}$ means a larger product error, and the ellipsoid is where it shows first. The three Lagrangian reports, however, order identically for both ratios at every n: MWL joint $0.92/0.94/0.94/0.95/0.96$, Bonferroni $0.97/0.98/0.99/0.98/0.99$, diagonal $0.82/0.86/0.88/0.88/0.88$ (recursion: joint $0.95/0.95/0.95/0.94/0.96$, Bonferroni $0.99/0.98/0.99/0.97/0.98$, diagonal $0.86/0.85/0.84/0.84/0.84$). The anti-conservative failure of the diagonal report is a property of *discarding* $\hat{\Sigma}_{01}$, not of the particular estimator that produced $\hat{w}$ (`results/exp8_mwl_results.json`).

E Experimental Protocol and Caveats

Scope of the empirical validation. Every "CDR" curve in the figures below is the *oracle-nuisance* variant (CDR-exact: true $Q^{\pi,\star}$ and w), which most cleanly validates the *asymptotic* predictions of our theory. The coverage experiments (Experiments 2 and 6) report the χ^2 statistic both with the *true* Σ^* and with the plug-in $\hat{\Sigma}$ that Corollary 7 actually prescribes; the two agree closely and both track the nominal levels. The deployable estimator (CDR-data, with data-driven nuisances) is reported alongside throughout, and with the plug-in $\hat{\Sigma}$ it attains nominal 95% coverage at n=2000 (Exp 2).

Note on CDR-data results. Throughout, "CDR-exact" is the CDR estimator with oracle nuisances (true $Q^{\pi,\star}$ and w) and "CDR-data" is the deployable estimator at full settings (K=5 folds, 20 FQE iterations). CDR-data reads *nothing* from the environment except the state, action and horizon dimensions and the behavior policy π_0, which OPE assumes known. Concretely: $\hat{Q}$ is tabular FQE in which sparsely- or never-visited (s,a) cells are handled by count-based shrinkage toward a pooled prior (cell $\rightarrow$ state, pooling over visited actions $\rightarrow$ step, pooling over all visited cells), never by consulting the true reward or cost function; and $\hat{w}$ is built from a transition kernel estimated on the training folds together with the Laplace-smoothed *empirical* initial-state distribution $\hat{\mu}$, not the true μ. The reported CDR-data numbers are therefore attainable by a practitioner with the same data, not an optimistic bound. Across experiments it behaves as the theory predicts (Taxi is the exception: there coverage with the true Σ^* is 0.01–0.04, Section G.1): relative bias $0.06\%/0.08\%$ with both nuisances estimated (Exp 3, Row E, versus 6.5% when both are wrong); MSE converging to CDR-exact's as n grows ($3.2\times$ at n=100 down to $1.14\times$ at n=2000, Exp 1); $n \cdot \mathrm{Var}$ linear rather than exponential in the horizon (Exp 5); and χ^2_2 coverage $0.990/0.950/0.884$ at n=2000 against nominal $0.99/0.95/0.90$ using the plug-in $\hat{\Sigma}$ of Corollary 7 (Exp 2). Coverage computed against the *true* Σ^* instead is lower by $0.8/2.8/2.8$ points ($0.982/0.922/0.856$); this is expected, since the plug-in $\hat{\Sigma}$ is formed from the same estimated scores as $\hat{\theta}$ and widens with the extra nuisance-driven dispersion, while the true Σ^* cannot. Inference in practice uses the plug-in, so it is the plug-in row that characterizes the deployed procedure.

Note on heavy-tailed estimators. For heavy-tailed estimators (JL, IS) at small n, mean MSE is dominated by rare outliers and has SE $\approx$ mean (cf. JSON file `exp1_results.json`); median

MSE or trimmed mean provides more stable summaries. We report mean for compatibility with classical convergence-rate analyses but caution against interpretation at $n \leq 200$.

Note on orthogonality verification. Experiment 3 verifies the zeroth-order moment condition $\mathbb{E}[\psi] = 0$ under mis-specification of one nuisance. The first-order statement of Theorem 9—that the bias is $O(\epsilon^2)$ when *both* nuisances are perturbed by ϵ—is verified directly in Experiment 10 (Appendix G.10), where the exact bias can be computed from the occupancy measures and its ϵ-exponent read off a log-log plot.

F Supplementary Remarks

The following remarks are referenced from the main text and collected here for space.

Remark 4 (Sign of off-diagonal)**.** *The estimator $\hat{L}_\lambda$ and its single true variance $\sigma_\lambda^2 = e_\lambda^\top \Sigma^* e_\lambda$ are the same in both analyses; only the* reported *variance differs. For $\lambda > 0$, when $\Sigma^*_{01} > 0$ the diagonal approximation* ***over-states*** *σ_λ^2 (valid but loose); when $\Sigma^*_{01} < 0$ it* ***under-states*** *it (anti-conservative, potentially unsafe); when $\Sigma^*_{01} = 0$ the two agree; the two agree also at $\lambda = 0$ (inactive constraint). In every case the joint covariance is simply the correctly calibrated report; no estimator is made more efficient.*

Remark 5 (Consistency with the Cramér–Rao bound of Jiang & Li 2016)**.** *Theorem 2 and Observation 1 of Jiang & Li (2016) show that for* discrete tree MDPs*—where the state is the full history, so no two trajectories share a state—the quantity in Proposition 1 is the Cramér–Rao bound and step-wise DR with $\hat{Q} = Q^{\pi,\star}$ attains it. This does not conflict with our efficiency claims, because the model classes differ: on a tree the unique history reaching s_h carries the entire product, so $w_h = \varrho_{1:h}$ exactly, CDR* coincides *with* (4)*, and it inherits the same variance. For discrete layered DAG MDPs, where states can merge, Jiang & Li (2016, App. C, Thm. 3) already show that the bound replaces $\varrho_{1:t-1}$ by the occupancy ratio; in the discrete case this coincides with Σ^*_{00} at $\gamma = 1$ (up to their terminal-reward convention). Precisely, since $\mathbb{E}_{\pi_0}[\varrho_{1:h} \mid s_h, a_h] = w_h$, $n \operatorname{Var}(\hat{v}_{\star,\mathrm{JL}}) - \Sigma^*_{00} = \sum_{h=1}^{H} \gamma^{2h} \mathbb{E}_{\pi_0}\big[\operatorname{Var}(\varrho_{1:h} \mid s_h, a_h) \operatorname{Var}_{P_h}(V^{\pi,\star}_{h+1} \mid s_h, a_h)\big] \geq 0$ (for $\gamma < 1$ with the discounted bound of Appendix J). The gap is positive iff, with positive probability, histories with different cumulative ratios merge into an (s_h, a_h) with non-degenerate transition noise; merging is necessary but not sufficient, and tree MDPs are the case with no merging.*

Remark 6 (Regularity)**.** *Throughout we implicitly assume the following regularity conditions: (i)* Measurability: *$Q^{\pi,\star}_h, V^{\pi,\star}_h, w_h$ are measurable functions on their respective domains; (ii)* Square-integrability: *$w_h \in L^2(d^{\pi_0}_h)$ for each h, ensuring the CDR score has finite second moment; (iii)* Absolute continuity: *$\pi \ll \pi_0$, i.e., $\pi(a|s) > 0 \Rightarrow \pi_0(a|s) > 0$ (already implied by Assumption 1, but stated explicitly here for clarity).*

Remark 7 (Positive definiteness)**.** *$\Sigma^* \succeq 0$ by construction, and $\Sigma^* \succ 0$ iff no non-trivial linear combination $v^\top \psi^*$ vanishes almost surely. A sufficient condition is that* either *the centred initial values $\{V_1^{\pi,\star_j} - \theta_j^*\}_{j=0}^J$ are linearly independent in $L^2(\mu)$,* or *for some h the weighted residuals $\{w_h \delta_h^{\star_j}\}_{j=0}^J$ are linearly independent in $L^2(P)$. These conditions are not necessary: since $\mathbb{E}[\delta_h^{\star_j} \mid s_h, a_h] = 0$, the initial and per-step pieces of ψ^* are mutually orthogonal, so $\Sigma^* = G_0 + \sum_h G_h$ with $G_0 = \big(\operatorname{Cov}_\mu(V_1^{\pi,\star_j}, V_1^{\pi,\star_k})\big)_{jk}$ and $G_h = \big(\mathbb{E}_{d^{\pi_0}_h}[w_h^2 \operatorname{Cov}_{P_h}(\delta_h^{\star_j}, \delta_h^{\star_k} \mid s_h, a_h)]\big)_{jk}$, all positive semidefinite. Hence $\ker \Sigma^* = \ker G_0 \cap \bigcap_h \ker G_h$: Σ^* is singular exactly on directions degenerate for every piece simultaneously (as when one signal is an exact affine combination of the others, cf. Experiment 6).*

Remark 8 (Time-homogeneous dynamics)**.** *If the dynamics are known to be time-homogeneous ($P_h \equiv P$), the tangent space shrinks and the efficiency bound is the (generally smaller) variance of the projection of ψ^* onto it; CDR remains regular and asymptotically normal with covariance Σ^*, so all inference statements are unaffected, but it is efficient only in the time-inhomogeneous model. All experiments use time-homogeneous environments.*

Remark 9 (Deterministic initial state)**.** *When $\mu = \delta_{s_0}$ is a point mass, the first term $\operatorname{Cov}_\mu(V_1^{\pi,\star_j}, V_1^{\pi,\star_k})$ in* (8) *vanishes, and the efficiency bound is determined entirely by the sum over transition-induced covariances.*

Remark 10 (Relationship to single-objective OPE)**.** *For each single component θ_j^*, the individual EIF $\psi^{\star_j,*}$ reduces to the marginal efficient score of Kallus & Uehara (2020); this is by design. The*

substantive content of Theorem 3 is therefore not a new score but the joint object built from these scores: (i) Σ^* generically has non-zero off-diagonal entries (Proposition 4), because all components are driven by the same initial state and transitions; they are present even on-policy ($w \equiv 1$); (ii) these entries are the cross-moments $\mathbb{E}[\psi^{\star_j,*}\psi^{\star_k,*}]$ of the componentwise efficient scores (the EIF of a vector-valued functional is the vector of componentwise EIFs; van der Vaart, 1998, Sec. 25.3), so they follow from the componentwise analysis of Kallus & Uehara (2020), but a report that keeps only the diagonal Σ^*_{jj} discards them; and (iii) their sign decides whether that diagonal report of Lagrangian uncertainty is conservative or anti-conservative (Section 4). The contribution lies in making the joint covariance explicit and exploiting it for feasibility inference, not in new per-component scores.

Remark 11 (Operating point: $\lambda^\dagger$ vs λ^*)**.** *The variance-minimizing $\lambda^\dagger = \Sigma^*_{01}/\Sigma^*_{11}$ used in part (d) is **not** generally the dual-optimal λ^* solving the actual CMDP problem. At λ^*, the ratio of the correctly reported to the diagonally reported Lagrangian variance is $1 - \frac{2\lambda^*\Sigma^*_{01}}{\Sigma^*_{00}+(\lambda^*)^2\Sigma^*_{11}}$, which depends on the problem's Slater geometry. When $\Sigma^*_{01} > 0$, AM–GM ($\Sigma^*_{00} + \lambda^2\Sigma^*_{11} \geq 2\lambda\sqrt{\Sigma^*_{00}\Sigma^*_{11}}$) gives a ratio in $[1-\rho, 1)$ for every $\lambda > 0$ (width ratio in $[\sqrt{1-\rho}, 1)$); the lower end is attained only at $\lambda_r := \sqrt{\Sigma^*_{00}/\Sigma^*_{11}} = \lambda^\dagger/\rho$, not at $\lambda^\dagger$, where the ratio is $(1-\rho^2)/(1+\rho^2) \geq 1-\rho$ (equality only for $\rho \in \{0, 1\}$). At λ^* the discrepancy is problem-dependent: it exceeds the one at $\lambda^\dagger$ exactly when $\lambda^\dagger < \lambda^* < \lambda^\dagger/\rho^2$. (When $\Sigma^*_{01} < 0$ the ratio lies in $(1, 1+|\rho|]$.) In neither case does any estimator change: the discrepancy is between a correct and an incorrect report of the same variance.*

Remark 12 (Bonferroni correction in volume comparison)**.** *$\mathcal{E}_{\text{indep}}$ uses the joint χ^2_{J+1} radius, not marginal quantiles, and is not a valid $(1-\alpha)$ region. A valid axis-aligned alternative is the Bonferroni box with half-widths $z_{1-\alpha/(2(J+1))}(\hat{\Sigma}_{jj}/n)^{1/2}$; for $J = 1$, $\alpha = 0.05$, $\text{vol}(\mathcal{E}_{\text{joint}})/\text{vol}(box) \to 0.937\sqrt{1-\rho^2}$.*

Remark 13 (Practical magnitude)**.** *In Experiment 4 (Section G.8), the construction yields $\rho \approx 0.72$, giving $\sqrt{1-\rho^2} \approx 0.69$; the size of the diagonal report's over-statement depends on λ's operating point. For typical CMDP problems where reward and cost share state dynamics, ρ is often substantial, but precise values are problem-dependent.*

Remark 14 (Nuisance estimation)**.** *For $\hat{Q}^\pi$ we adopt fitted-Q evaluation (FQE): starting from $\hat{V}^{\pi,\star}_{H+1} \equiv 0$ and proceeding backward in h, regress $f^\star_h(s_h, a_h) + \gamma\hat{V}^{\pi,\star}_{h+1}(s_{h+1})$ on (s_h, a_h), where $\hat{V}^{\pi,\star}_{h+1}(s) = \mathbb{E}_{a\sim\pi_{h+1}(\cdot|s)}[\hat{Q}^\star_{h+1}(s,a)]$, separately for each $\star \in \{r, c_1, \ldots, c_J\}$, via, e.g., kernel ridge or neural networks. For $\hat{w}$ (a π-dependent quantity), we use recursive regression of the cumulative ratio (Kallus & Uehara, 2020, Sec. 3.3) or BBR. In policy learning (Section 6), the model $(\hat{\mu}, \hat{P}_h)$ is fit once per cross-fitting fold on π_0 data; for each candidate π, both $\hat{Q}^\pi$ (by backward induction $\hat{Q}^\pi_h = f_h + \gamma\hat{P}_h\hat{V}^\pi_{h+1}$, $\hat{V}^\pi_{H+1} \equiv 0$, equivalently tabular FQE re-run with target π) and $\hat{w}^\pi$ are recomputed analytically from this fixed model. A single π-independent $\hat{Q}$ would not do: Q^π_h depends on $\pi_{h+1:H}$, so a fixed $\hat{Q}$ is inconsistent for all but one π and Assumption 3 then fails uniformly over Π. (Classification or BBR ratio estimators are alternatives for a single fixed π, but must be refit for each π.)*

Remark 15 (Sufficient conditions for Assumption 6)**.** *This assumption is satisfied, for example, when: (i) the policy class Π is finite with $|\Pi|$ elements; or (iii) $\mathcal{S} \times \mathcal{A}$ is finite, $\hat{Q}^\pi$ and $\hat{w}^\pi$ are computed for each π from a fixed estimated model $(\hat{\mu}, \hat{P})$ of probability kernels as in Remark 16, with $\|\hat{Q}^\pi_h\|_\infty \leq M$, the estimated behavior occupancy is bounded below on its support, $\hat{d}^{\pi_0}_h(s,a) \geq \nu > 0$, and Π has sup-norm metric entropy $\log N(\Pi, \epsilon, \|\cdot\|_\infty) \leq D\log(C_\Pi/\epsilon)$ (e.g. softmax policies with bounded parameters); each projected score is then $L\|v\|_1$-Lipschitz in $\|\pi - \pi'\|_\infty$ with $L = HM + H(H+1)(1+2M)/(2\nu) + \gamma H(H-1)M^2$, and Dudley's entropy integral gives $\mathcal{R}_m(\mathcal{F}^{(k)}_v) \leq c\,L\,\|v\|_1\,C_\Pi\sqrt{D/m}$ for every sample size m and an absolute constant c, i.e. Assumption 6 with $\kappa = c\,C_\Pi\sqrt{D}$. (We label the conditions (i) and (iii) to match the code and the proof.) See Appendix S for details.*

Remark 16 (Policy-indexed estimators)**.** *For each candidate $\pi \in \Pi$, the CDR estimates $\hat{v}^\pi$ and $\hat{c}^\pi_j$ are computed via Algorithm 1 with π as the target policy. Following Remark 14, the model $(\hat{\mu}, \hat{P})$ is fit once on π_0 data; for each π, $\hat{Q}^\pi$ is obtained by backward induction on $\hat{P}$, and the marginal density ratio $\hat{w}^\pi_h(s,a) = \frac{\pi_h(a|s)}{\pi_0(a|s)} \cdot \frac{\hat{d}^\pi_h(s)}{\hat{d}^{\pi_0}_h(s)}$ is computed from the state marginals $\hat{d}^{\pi\prime}_1 = \hat{\mu}$,*

$\hat{d}^{\pi'}_{h+1}(s') = \sum_{s,a} \hat{d}^{\pi'}_h(s)\pi'_h(a|s)\hat{P}_h(s'|s,a)$, $\pi' \in \{\pi, \pi_0\}$ *(tabular / model-based case, as in all experiments). This makes the policy optimization in* (12) *computationally tractable.*

Remark 17 (Implementation via Lagrangian)**.** *For the finite class of Experiment 9,* (12) *is solved exactly by enumeration, which is the setting Theorem 8 covers. For large* Π *one may use a primal-dual scheme with a best-response policy oracle and averaged output (Le et al., 2019); its saddle point can differ from* (12) *when* Π *is not mixture-closed, and Theorem 5 at data-dependent iterates additionally requires asymptotic normality uniform over* Π*. The point estimates, and hence the iterates, are unchanged by the choice of report.*

Remark 18 (Finite-sample threshold)**.** *The regret part of Theorem 8 needs* $n \geq n_0(\alpha_S, \kappa, L, M, H, J, K, \delta) := \max_j \left[2(2\sqrt{K}L\kappa + (\sqrt{2K}B_j + 1)\sqrt{\log(4JK/\delta)})/\alpha_S\right]^2$*, which grows like* $KH^2M^4\log(4JK/\delta)/\alpha_S^2$*; this is used only to make* π^* *feasible for the tightened program. The result is asymptotic only through* $\mathbb{P}(\mathcal{E}_n^c) \to 0$*, for which no rate is given. For* $n < n_0$ *(small* n *or small slack* α_S*) the theorem does not apply.*

Remark 19 (Multi-constraint Lagrangian advantage: when it helps and when it does not)**.** *With* $\boldsymbol{\lambda} \in \mathbb{R}^J_+$ *and* $e_{\boldsymbol{\lambda}} = (1, -\lambda_1, \ldots, -\lambda_J)^\top$*, the diagonal report exceeds the correctly-calibrated joint Lagrangian variance by:*

$$\sigma^2_{\boldsymbol{\lambda},\text{indep}} - \sigma^2_{\boldsymbol{\lambda}} = 2\sum_{j=1}^{J} \lambda_j \Sigma^*_{0j} - 2\sum_{j<k} \lambda_j \lambda_k \Sigma^*_{jk}, \tag{18}$$

containing $O(J^2)$ *off-diagonal terms. The diagonal report over-states the Lagrangian variance (so the joint report is smaller) when reward-cost correlations dominate cost-cost correlations, e.g., when* (18) *is positive, and under-states it (the anti-conservative case) when it is negative. The joint approach always yields correctly calibrated inference by using the true* Σ^**.*

G Additional Experimental Results

This appendix reports Experiments 1–6 and 9–11 in full, and gives the details of Experiments 7 and 12 that Section 7 summarizes.

G.1 Experiment 12 in Detail

Figure 3 plots the Taxi-Safety results summarized in Section 7.3, and Table 1 (right) gives the Lagrangian coverage of CDR-data. JL-Concat's joint MSE is 1.50 at n=2000 and 2.46 at n=4000, so it does not improve with n. In the horizon sweep at n=1000, CDR-exact's $n \cdot \mathrm{Var}$ grows from 0.073 at H=10 to 1.46 at H=100 while tracking $\mathrm{tr}(\Sigma^*)$ within 5%; JL-Concat's grows from 0.081 to $1{,}340$. CDR-data's plug-in χ^2_2 ellipsoid coverage is $0.91/0.93/0.95/0.93$ at n=500/1000/2000/4000. Computed with the true Σ^* instead (as in Experiments 2 and 6), its coverage is $0.01/0.03/0.04/0.04$: its plug-in correlation is $\hat{\rho} = -0.15/-0.14/-0.11/-0.09$ against $\rho = -0.84$, and $n\hat{\sigma}^2_\lambda$ falls from 6.97 to 2.72 against the true 1.63, so at these n CDR-data is not yet in the asymptotic regime for the cost, and it is its own larger $\hat{\Sigma}$ that keeps its intervals calibrated. Its joint Lagrangian interval covers 0.90–0.93, and the diagonal report covers 2–5 points less at every n (0.880 vs. 0.897 at n=4000). For CDR-exact the joint interval covers 0.93–0.96, the diagonal report 0–2 points less, and Bonferroni 0.97–0.99 at 15% extra width.

G.2 Experiment 7 in Detail

This experiment isolates the safety-critical case of Theorem 5(c) in the cleanest possible setting: tabular, with *oracle* nuisances, so that any difference is attributable to the covariance report alone and not to estimation error. We use a GridWorld variant whose cost increases with row while the reward decreases with row, giving $\Sigma^*_{01} = -2.93$ ($\rho = -0.72$). For $L(\lambda) = v^\pi - \lambda c^\pi$ at $\lambda = 1$ the true asymptotic variance is $e_\lambda^\top \Sigma^* e_\lambda = 14.36$, whereas the diagonal treatment reports only $\Sigma^*_{00} + \lambda^2\Sigma^*_{11} = 8.49$—an under-statement of 41%. The point estimator $\hat{L}(\lambda)$ is identical under all three reports of Section 4; we measure coverage *and* width of each (2000 replications, binomial standard error 0.005; Figure 4). The joint interval is calibrated ($0.946/0.951/0.951$ at $n \in \{500, 1000, 2000\}$, within one standard error of nominal throughout). The diagonal interval

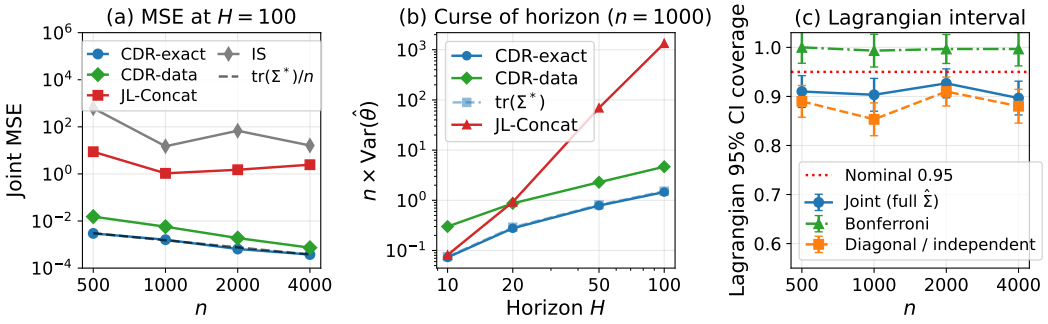


Figure 3: Experiment 12: Taxi-Safety (300 runs). **(a)** Joint MSE at $H{=}100$; CDR-exact tracks $\mathrm{tr}(\Sigma^*)/n$, CDR-data closes on it, JL-Concat and IS are 10^3–$10^5\times$ worse. **(b)** Curse of horizon at $n{=}1000$: $n \cdot \mathrm{Var}$ grows linearly for CDR and exponentially for JL-Concat. **(c)** The three reports of the Lagrangian interval for CDR-data ($\lambda{=}1$): the diagonal report sits 2–5 points below the joint one; this gap and the $1.5\times$ Bonferroni width reflect CDR-data's $\hat{\Sigma}$, which at these n is still far from Σ^* in the cost entry (Section 7.3).

under-covers by roughly eight percentage points ($0.863/0.868/0.874$, some eleven standard errors below nominal) and does not improve with n: the defect is in the reported variance, not the sample size. Bonferroni behaves as its validity guarantee predicts—$0.983/0.987/0.985$, never unsafe and never tight—paying a constant 23.3% in width relative to the joint interval, while the diagonal report buys its 23.1% narrower intervals by being wrong. The comparison that matters is thus joint versus Bonferroni: two valid procedures, of which the joint one is uniformly tighter. This regime is excluded from Experiment 4 (whose $\lambda^\dagger \geq 0$ clipping cannot reach $\Sigma^*_{01} < 0$) (Remark 4).

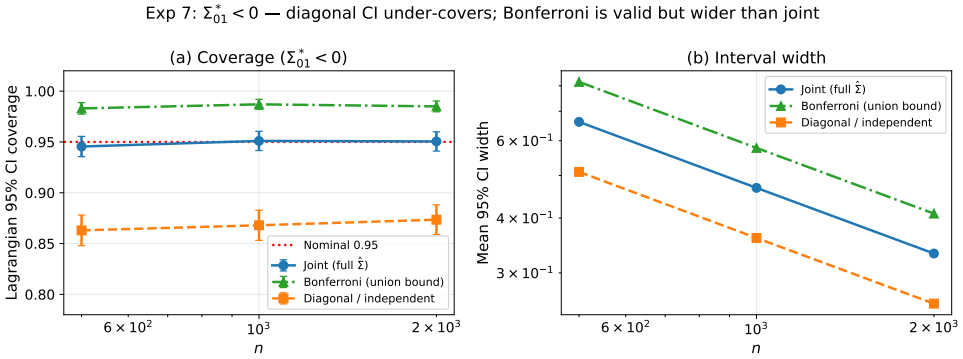


Figure 4: Experiment 7: negative reward-cost covariance ($\Sigma^*_{01} < 0$, $\rho = -0.72$, $\lambda = 1$; 2000 replications, error bars ± 1.96 binomial SE). The point estimator is identical in all three; only the reported variance differs. **(a)** The joint $\hat{\Sigma}$ interval (blue) is calibrated and Bonferroni (green) over-covers, while the diagonal/independent interval (orange) *under-covers* at every n—the anti-conservative failure of Theorem 5(c). **(b)** Bonferroni buys its validity with 23% extra width relative to the joint interval, uniformly in n; the diagonal interval is 23% narrower than the joint one, which is precisely why it fails to cover.

G.3 Experiment 9: Constrained Policy Learning

This experiment illustrates the mechanism of Theorem 8—constraint satisfaction with probability $\geq 1 - \delta$, and reward regret that vanishes—but does not verify the theorem's constants. We use CDR-exact (oracle Q and w) and, for each $\pi \in \Pi$ separately, the one-sided normal-approximation margin $\epsilon^\pi_n = z_{1-\delta}(\hat{\Sigma}^\pi_{11}/n)^{1/2}$, with no union bound over Π and no Rademacher term. The finite-sample margin (13) exceeds $\alpha_S/2$ by orders of magnitude at every n here (with $K{=}5$ and $M \geq \max|Q| \approx 8.4$ it is at least $\approx 1.2 \times 10^4/\sqrt{n}$, about 180 at $n{=}4000$; Remark 18), so it would reject

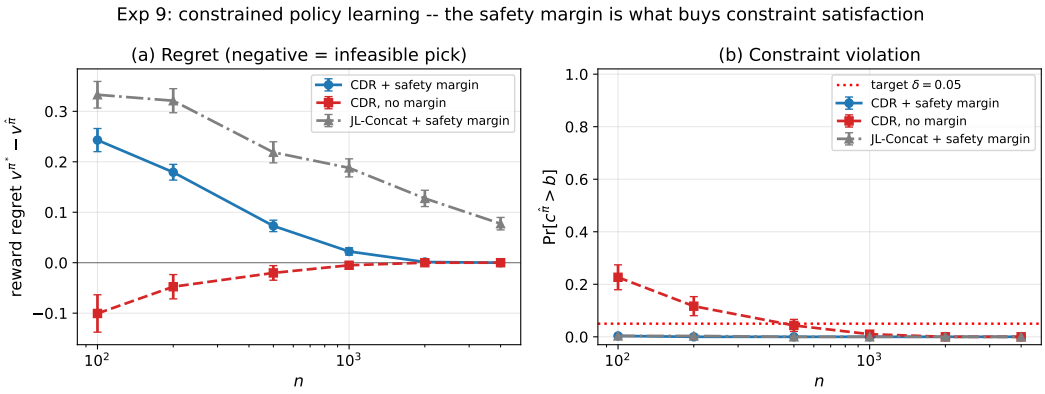


Figure 5: Experiment 9: constrained policy learning (300 runs, error bars ± 1.96 SE). **(a)** Regret; the no-margin rule dips below zero by selecting infeasible policies. **(b)** Constraint violation rate against the δ=0.05 target: the (normal-approximation) safety margin is what keeps violations low, and a noisier evaluator pays for the same safety in regret rather than in violations.

every policy. When no policy clears the test the rule returns π^*; this happens in 8 of 300 runs (2.7%) at n=100 and never for $n \geq 200$. A class of $|\Pi|$=20 softmax policies is evaluated off-policy on a single logged dataset, and a policy is selected by maximizing $\hat{v}^\pi$ subject to an upper-confidence cost clearing the threshold. On GridWorld-Safety with b=6.0 we have π^* at v^{π^*}=5.908, c^{π^*}=5.763, so the feasibility slack of π^* is α_S=0.237; the best *infeasible* policy has v=6.762, which is the trap an unconstrained maximizer falls into. We compare selection rules that differ only in how uncertainty enters the feasibility test (300 runs per cell, δ=0.05; Figure 5 plots regret and violation rate against n).

n	100	200	500	1000	2000	4000
CDR-exact + margin: regret	0.243	0.179	0.073	0.022	0.001	0.000
violation	0.003	0.000	0.000	0.000	0.000	0.000
CDR-exact, no margin: regret	−0.101	−0.048	−0.020	−0.005	0.000	0.000
violation	**0.227**	**0.117**	0.043	0.010	0.000	0.000
JL + margin: regret	0.333	0.321	0.219	0.188	0.127	0.077
violation	0.003	0.003	0.000	0.000	0.000	0.000

Three things follow. *(i) The margin-tightened rule is safe in practice.* CDR-exact with the safety margin violates the constraint in at most 0.3% of runs at every sample size, far inside the δ=0.05 budget, while its regret falls monotonically to zero. *(ii) The margin is what buys it.* Dropping the margin and testing $\hat{c}^\pi \leq b$ directly produces violation rates of 22.7% and 11.7% at n=100, 200—and *negative* regret, because the rule beats v^{π^*} precisely by selecting a policy that is not feasible. Negative regret here is a symptom, not an achievement. *(iii) The evaluator matters, not just the rule.* Given its own variance for the margin, JL-CONCAT is equally safe—its wider margin is doing the work—but pays for that safety in regret, which is never below CDR's and falls far more slowly: the ratio grows from 1.4× at n=100 to 90× at n=2000 as CDR's regret collapses and JL's does not. Its regret at n=4000 (0.077) merely matches CDR's at n=500 (0.073), an eightfold difference in the data needed for the same policy quality. The horizon-driven variance of Section 3 thus propagates directly into the learned policy. We also ran a pessimistic-Lagrangian selection rule with the joint $\hat{\Sigma}$ against the diagonal one: both satisfy the constraint in every run, and their regrets differ by at most 0.002 at every n. This is the expected negative result and matches the claim in Section 6—the covariance report calibrates inference, it does not change which policy is selected. Note also that regret reaches exactly zero once n separates a *finite* Π; the $O(1/\sqrt{n})$ rate of Theorem 8 is an upper bound that a finite well-separated class beats.

G.4 Experiment 5: Curse of Horizon

We fix n=1000 and vary $H \in \{5, 10, 20, 50, 100\}$ on Random-CMDPs ($|\mathcal{S}|$=20, $|\mathcal{A}|$=4, 500 runs per horizon; Figure 6, right). CDR-exact's empirical n·Var matches $\mathrm{tr}(\Sigma^*)$ to within 6% at every horizon (0.196/0.525/0.978/2.47/4.84 against 0.200/0.497/0.994/2.53/4.95) and grows 24.7× as H grows 20×—linear in H, the additive (non-compounding) structure of Theorem 3 at a per-step noise level that does not change with H. JL-Concat instead grows from 1.02 to 4.4×10^4, a 43,000× increase, which is the exponential regime of Proposition 1; at H=100 its mean squared error is 9,000× CDR-exact's (43.8 vs. 0.0048; its median, 0.030, is 6×, the gap between mean and median being the heavy tail of the cumulative weights). The deployable CDR-data grows 23× (0.42 → 9.9)—also linear, confirming that the non-compounding horizon scaling is a property of the estimator and not of the oracle nuisances.

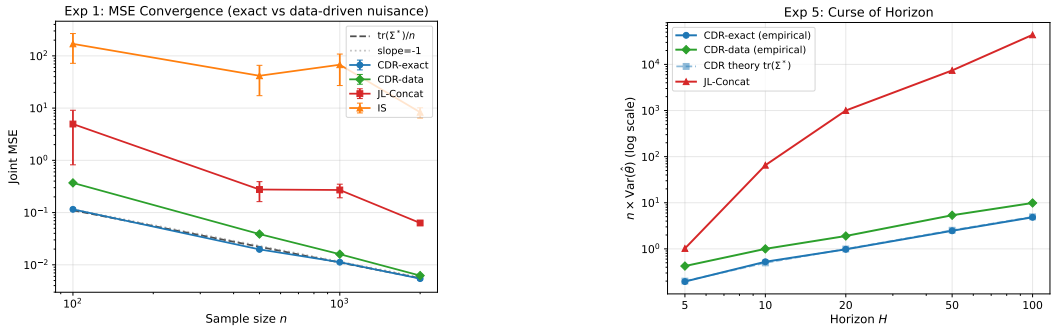


Figure 6: **Left** (Exp. 1): joint MSE vs. sample size (log-log); CDR-exact tracks $\mathrm{tr}(\Sigma^*)/n$ (dashed) while JL-Concat and IS are substantially worse. **Right** (Exp. 5): $n \cdot \mathrm{Var}(\hat{\theta})$ vs. horizon H (log y-axis); CDR-exact grows linearly and tracks $\mathrm{tr}(\Sigma^*)$, JL-Concat grows exponentially.

G.5 Experiment 1: MSE Convergence Rate

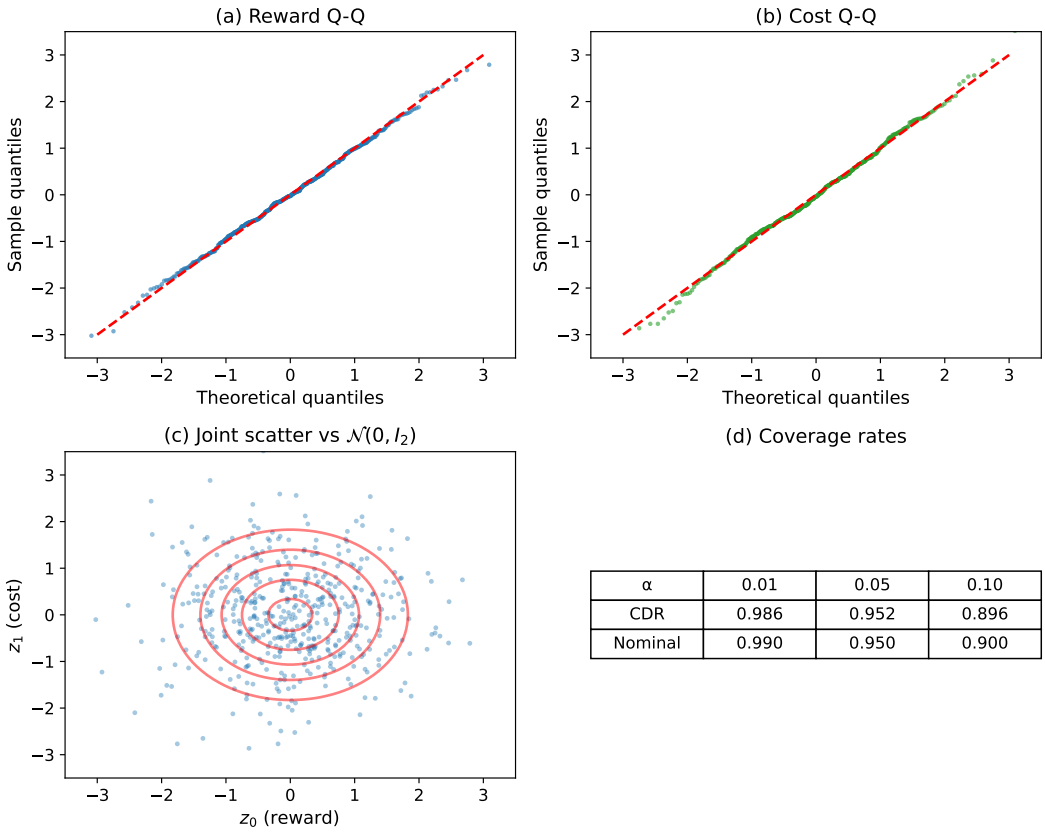
We verify the $O(1/n)$ rate of Theorem 6 over 500 runs at $n \in \{100, 500, 1000, 2000\}$ on GridWorld-Safety (Figure 6, left). CDR-exact decays at a fitted log-log slope of -1.015 and matches $\mathrm{tr}(\Sigma^*)/n$ to within 4% at three of four sample sizes (ratios 1.04, 0.89, 1.02, 0.98); JL-Concat's MSE is 11.6–24.1× larger at n=500–2000. The deployable CDR-data closes on CDR-exact as n grows—from 3.2× its MSE at n=100 to 1.14× at n=2000—as Assumption 3 predicts.

G.6 Experiment 2: Asymptotic Normality

We verify the joint Gaussian limit of Theorem 6 at n=2000 over 500 runs. Both components produce Q-Q plots nearly indistinguishable from $\mathcal{N}(0, 1)$ (Jarque–Bera p-values: reward $p = 0.998$, cost $p = 0.208$; Figure 7a,b). Empirical coverage of the χ^2_2 confidence ellipsoid closely approximates the nominal levels, whether we use the true Σ^* or the plug-in $\hat{\Sigma}$ that Corollary 7 actually prescribes:

Nominal $1-\alpha$	0.99	0.95	0.90
CDR-exact (true Σ^*)	0.986	0.952	0.896
CDR-exact (plug-in $\hat{\Sigma}$)	0.990	0.940	0.888
CDR-data (true Σ^*)	0.982	0.922	0.856
CDR-data (plug-in $\hat{\Sigma}$)	0.990	0.950	0.884

The deployable variant matters most here, and the row that matters is the last one: with the plug-in $\hat{\Sigma}$ that Corollary 7 actually prescribes, CDR-data attains 0.990/0.950/0.884 against nominal 0.99/0.95/0.90—i.e. it is correctly calibrated at n=2000 at the 99% and 95% levels and short by 1.6 points at 90%, about 1.2 binomial standard errors at 500 replications, i.e. not statistically distinguishable from nominal. Using the *true* Σ^* instead leaves CDR-data under-covering nominal by 0.8/2.8/4.4 points at the 99/95/90% levels (0.8/2.8/2.8 points below the plug-in row), which is



α	0.01	0.05	0.10
CDR	0.986	0.952	0.896
Nominal	0.990	0.950	0.900

Figure 7: Experiment 2: Asymptotic normality verification (n=2000, 500 runs). Axes are standardized: $z = \sqrt{n}\,\Sigma^{*-1/2}(\hat{\theta} - \theta^*)$. (a,b) Component Q-Q plots vs. $\mathcal{N}(0,1)$. (c) Joint scatter vs. $\mathcal{N}(0, I_2)$ contours. (d) Coverage table.

the expected direction: the plug-in $\hat{\Sigma}$ is computed from the same estimated scores as $\hat{\theta}$ and so inflates in step with the extra nuisance-driven dispersion, whereas the true Σ^* does not. The Jarque–Bera test does not reject normality for either variant (CDR-data $p = 0.25,\ 0.42$). A coverage-vs-n sweep ($n \in \{2000, 4000, 8000\}$, 300 runs each, binomial standard error ≈ 0.013) shows the residual gap closing with sample size rather than persisting: 95% coverage rises from 0.913 to 0.920 to 0.953 against the true Σ^*, and from 0.943 to 0.933 to 0.960 with the plug-in $\hat{\Sigma}$, so both reach nominal at n=8000; the 90% level behaves the same way (0.870/0.870/0.900 and 0.900/0.877/0.907). This is what Assumption 3 predicts—the nuisance error enters only through a product that is $o_p(n^{-1/2})$—and it is the reason we report CDR-data alongside CDR-exact rather than treating the oracle variant as the only evidence. The Q–Q plots in Fig. 7 use the oracle (CDR-exact) variant.

Robustness to the master seed. All numbers above come from one master seed. Repeating the n=2000 coverage table under five master seeds (300 replications each) gives, with the plug-in $\hat{\Sigma}$, CDR-exact 0.992/0.947/0.900 and CDR-data 0.991/0.957/0.899 against nominal 0.99/0.95/0.90, with across-seed standard deviations of 0.004–0.018—the size of the binomial Monte-Carlo error at 300 replications, so no seed is an outlier. (With the true Σ^*: CDR-exact 0.992/0.952/0.904; CDR-data 0.986/0.931/0.868, a 0.4/1.9/3.2 point under-coverage at the 99/95/90% levels, the same pattern as in the table but smaller.) The coverage-vs-n sweep replicated under three seeds gives CDR-data plug-in 95% coverage 0.960/0.953/0.963 and true-Σ^* coverage 0.939/0.940/0.960 at $n = 2000/4000/8000$ (across-seed SD ≤ 0.023); the per-seed values are in `results/exp2_seeds.json`.

Table 3: Double robustness verification (n=2000, 500 runs).

	$\lvert\text{Bias}\rvert/\theta^*$		RMSE	
Config	Reward	Cost	Reward	Cost
A: Q✓ w✓	0.0000	0.0001	0.042	0.060
B: Q✓ w✗	0.0003	0.0010	0.051	0.069
C: Q✗ w✓	0.0016	0.0019	0.167	0.171
D: Q✗ w✗	0.0659	0.0654	0.433	0.426
E: CDR-data (both estimated)	0.0006	0.0008	0.048	0.064

G.7 Experiment 3: Double Robustness

We verify Theorem 2 under four nuisance configurations (n=2000, 500 runs).

Table 3 confirms that CDR remains unbiased (relative bias $< 0.2\%$) whenever *either* Q or w is correctly specified (Configs A–C), while Config D produces substantial bias ($\approx 6.5\%$). Row E is the deployable variant, in which *neither* nuisance is supplied: both are estimated from the same data. Its relative bias (0.06% / 0.08%) and RMSE (0.048 / 0.064) are close to the both-correct oracle Config A and an order of magnitude below Config D, i.e. two consistently estimated nuisances behave here like one correct one rather than like two wrong ones—which is what the product-rate condition of Assumption 3 asserts, and is not implied by Theorem 2 alone.

Choice of mis-specification. Wrong-Q uses $\hat{Q}_h^{\text{mis}} = r_h$ (immediate reward only), which omits the look-ahead structure but preserves (s, a)-dependence. Wrong-w perturbs $\log \pi_0$ with $\mathcal{N}(0, 0.3)$ noise; the resulting $\hat{w}^w$ corresponds to the exact density ratio under a perturbed behavior policy, providing a controlled but moderate stress test. More aggressive mis-specification (e.g., random function class with no relation to truth) would yield larger RMSE gaps.

G.8 Experiment 4: Joint vs. Diagonal Covariance Report

This experiment directly validates Theorem 5. The point estimate is the same throughout; only the reported variance differs. Figure 8(a) compares the empirical ratio of the two *reported* Lagrangian variances against the theoretical prediction $(1-\rho^2)/(1+\rho^2)$ across a Chain-CMDP sweep that produces realized correlations $\rho_{\text{act}} \in \{0.01, 0.05, 0.09, 0.59, 0.73, 0.88\}$. The empirical ratios track the predicted decay in ρ—from ≈ 1.0 at $\rho \approx 0$ down to 0.113 at $\rho = 0.88$—with deviations of $+2\%$ to $+9\%$ and -8% that are mixed in sign and within the $\approx 8\%$ Monte-Carlo standard error of a variance ratio at 300 replications. For a separate high-correlation Chain-CMDP instance (Figure 8(b); realized $\rho = 0.924$, distinct from the sweep above), the correctly calibrated joint interval for $L(\lambda^\dagger)$ has 95% width 0.023 at n=2000, whereas the diagonal report of the *same* point estimate over-states it as 0.082—an empirical width ratio of 0.2806 against the theoretical $\sqrt{(1-\rho^2)/(1+\rho^2)} = 0.2806$ at $\lambda^\dagger$, and the agreement holds across all four sample sizes (0.278, 0.282, 0.281, 0.281 at $n = 200, 500, 1000, 2000$; see the `exp4b` block of `exp4_results.json`). (The variance ratio at this ρ is $(1-\rho^2)/(1+\rho^2) = 0.0787$; since CI widths scale as the square root of variance, the width ratio is $\sqrt{0.0787} = 0.281$.) On GridWorld-Safety ($\rho = 0.72$), the empirical ellipsoid volume ratio is 0.693, matching $\sqrt{1-\rho^2} = 0.694$ (Corollary 12).

Scope of Experiment 4. The sweep here constructs $\rho \in [-0.04, 0.92]$ via reward-cost coupling and evaluates at the variance-minimizing $\lambda^\dagger \geq 0$, which by construction cannot probe the $\Sigma_{01}^* < 0$ regime (there $\lambda^\dagger$ would be negative, but the constraint multiplier is ≥ 0). The negative-Σ_{01}^* / anti-conservative case—the calibration-advantage case of Remark 4—is validated separately in Experiment 7 (Section 7.1) at a fixed $\lambda > 0$.

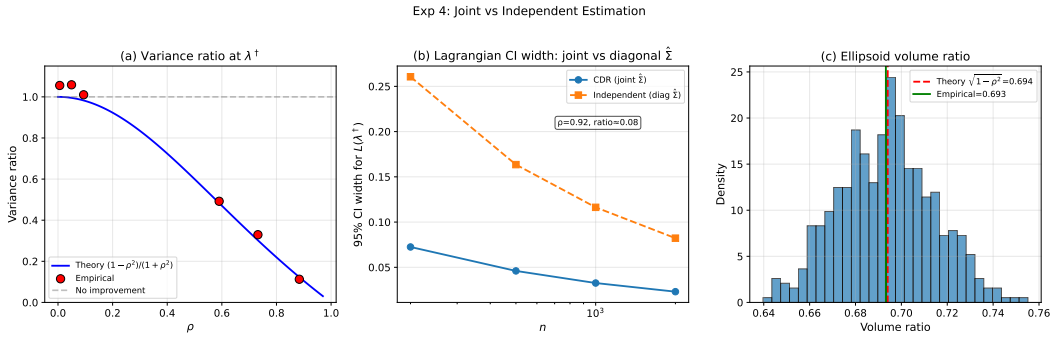


Figure 8: Experiment 4: Joint vs. diagonal covariance report (the point estimate is identical throughout). Here $L(\lambda) = v^\pi - \lambda c^\pi$ is the Lagrangian, and $\lambda^\dagger = \Sigma^*_{01}/\Sigma^*_{11}$ is the variance-minimizing multiplier (Theorem 5d). (a) Ratio of the correctly reported to the diagonally reported Lagrangian variance at $\lambda^\dagger$ vs. ρ: empirical (dots) vs. theory $(1-\rho^2)/(1+\rho^2)$ (curve). (b) Reported 95% CI width for $L(\lambda^\dagger)$: joint $\hat{\Sigma}$ (blue) vs. diagonal report (orange) as a function of n; because $\Sigma^*_{01} > 0$ here, the diagonal report is over-conservative. (c) Confidence ellipsoid volume ratio histogram.

G.9 Experiment 6: Multi-Constraint Setting (J=2)

We verify CDR on GridWorld-Safety-2C with two cost channels. The second cost combines a row-increasing term with a column-bowl term that lies outside $\mathrm{span}\{1, \mathrm{row}, \mathrm{col}\}$; this is deliberate, so that the three value functions are *not* linearly dependent and the efficiency matrix Σ^* is full rank and well-conditioned (a purely linear second cost would make the reward an exact affine combination of the two costs, rendering Σ^* rank-deficient and the χ^2 ellipsoid test ill-posed; cf. Remark 7). True values are $\theta^* = (6.08,\ 6.00,\ 4.44)^\top$, and the 3×3 efficiency matrix is

$$\Sigma^* = \begin{pmatrix} 3.80 & 3.78 & -1.34 \\ 3.78 & 7.25 & 1.26 \\ -1.34 & 1.26 & 3.66 \end{pmatrix} \quad (\lambda_{\min} = 0.37,\ \mathrm{cond}(\Sigma^*) \approx 26),$$

with correlation coefficients $\rho(r, c_1) = 0.72$, $\rho(r, c_2) = -0.36$, and $\rho(c_1, c_2) = 0.24$. CDR-exact tracks $\mathrm{tr}(\Sigma^*)/n$ (Figure 9, left; empirical/theoretical MSE ratio 0.88–1.10 across the four sample sizes) and achieves χ^2_3 ellipsoid coverage of 0.92–0.97 across $n \in \{200, 500, 1000, 2000\}$, essentially identical whether the statistic uses the **true** Σ^* (0.92/0.96/0.96/0.97) or the **plug-in** $\hat{\Sigma}$ of Corollary 7 (0.93/0.96/0.97/0.97). Because the redesigned Σ^* is full rank and well-conditioned, this coverage is meaningful rather than the degenerate 1.00 that a rank-deficient Σ^* produces (its near-singular inverse makes the χ^2 statistic vanish); this is exactly why the second cost is constructed off the row/column span. JL-Concat's MSE is 24–37$\times$ larger than CDR-exact's across n, reflecting its heavy-tailed cumulative importance weights. The deployable CDR-data again closes on CDR-exact with n, from 2.4$\times$ its MSE at n=200 to 1.13$\times$ at n=2000, so the joint construction degrades gracefully when the shared w must serve three components rather than two.

G.10 Experiment 10: Neyman Orthogonality (Bias vs. Nuisance Error)

Theorem 9 says that the CDR score has zero first-order sensitivity to its nuisances: perturbing $Q \to Q + \epsilon\,\Delta Q$ and $w \to w(1 + \epsilon u)$ *simultaneously* changes $\mathbb{E}[\psi]$ by $O(\epsilon^2)$, whereas double robustness (Theorem 2) is the statement that perturbing only one of them changes it not at all. On GridWorld-Safety the bias $\mathbb{E}[\hat{\theta}] - \theta^*$ can be computed *exactly* from the occupancy measures for any pair of nuisances, so the ϵ-exponent can be read off a log-log plot without Monte-Carlo noise. We use structured directions—an optimistic value model $\Delta Q = Q/\|Q\|_\infty$ and a ratio inflated by $u = 0.5 + 0.5\xi$, $\xi \sim U[-1, 1]$—and, for contrast, two non-orthogonal estimators of the same target built from the same perturbed nuisances: the direct method $\hat{\theta}_{\mathrm{DM}} = \mathbb{E}_\mu[\hat{V}_1]$ and marginalized importance sampling $\hat{\theta}_{\mathrm{MIS}} = \sum_h \mathbb{E}[\hat{w}_h f_h]$. Figure 10(a): over $\epsilon \in [0.01, 0.1]$ the fitted slopes are 2.000 for CDR with both nuisances perturbed, and 1.000 for DM and for MIS; with only one nuisance perturbed the CDR bias is at machine precision ($< 3 \times 10^{-15}$) for every ϵ, which is double robustness. Figure 10(b) repeats the comparison empirically at n=2000 (300 runs): at $\epsilon = 0.1$ the CDR

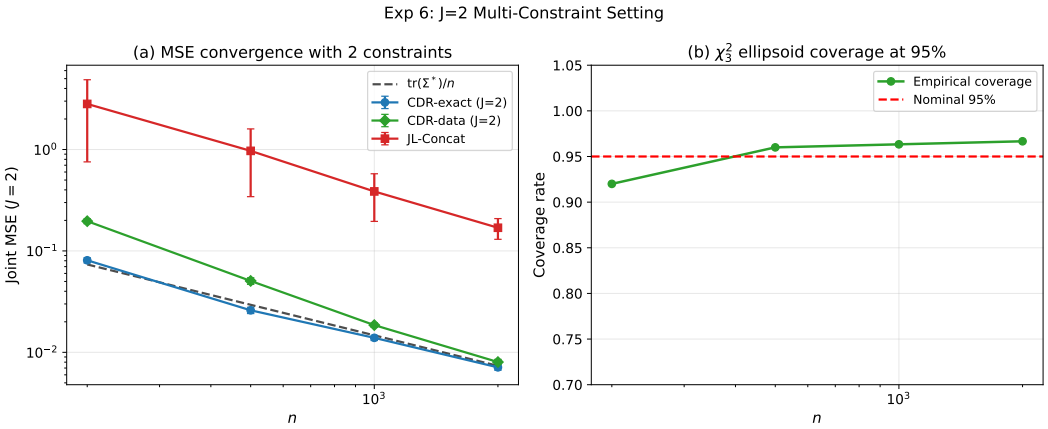


Figure 9: Experiment 6: Multi-constraint setting with $J{=}2$ cost channels (one reward and two costs, so $\theta^* \in \mathbb{R}^3$). (Left) Joint MSE for CDR-exact vs. JL-Concat (log-log); CDR-exact tracks $\mathrm{tr}(\Sigma^*)/n$ (dashed). (Right) χ^2_3 ellipsoid coverage at 95% nominal level.

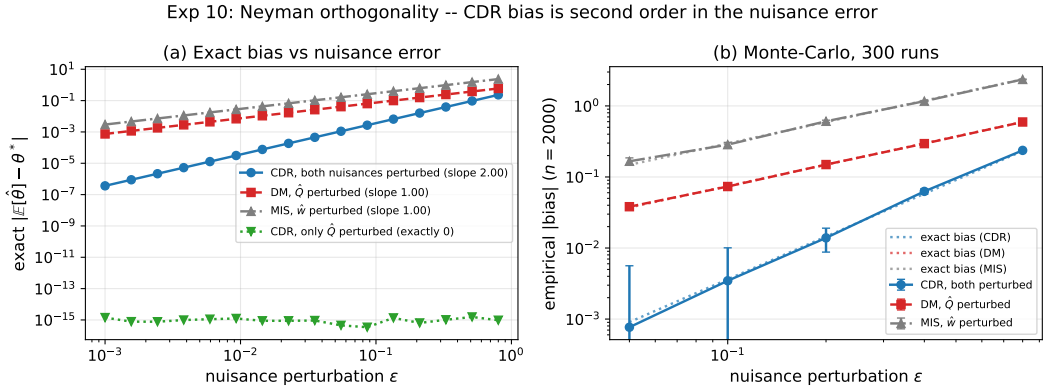


Figure 10: Experiment 10: Neyman orthogonality. **(a)** Exact bias as a function of the nuisance perturbation ϵ (log-log). CDR with both nuisances perturbed has slope 2; the direct method (perturbed $\hat{Q}$) and marginalized IS (perturbed $\hat{w}$) have slope 1; CDR with a single perturbed nuisance is exactly unbiased. **(b)** The same comparison from $n{=}2000$ samples (300 runs, ± 1.96 SE), with the exact curves dotted.

bias is 0.0035 ± 0.0034 (exact 0.0036) against 0.073 for DM and 0.285 for MIS, and at $\epsilon = 0.4$ it is 0.063 (exact 0.058) against 0.295 and 1.17; the empirical biases track the exact curves within Monte-Carlo error at every ϵ. This is the mechanism behind the product-rate condition: an error of size ϵ in each nuisance costs CDR $O(\epsilon^2)$, which is why two estimated nuisances behave like one correct one in Experiment 3 (Row E).

G.11 Experiment 11: CartPole-Safety, a Cautionary Benchmark

We also tried a control benchmark: Gymnasium's CartPole-v1 dynamics with a corridor cost ($c_h = \mathbf{1}\{|x_h| > 0.6\}$, plus 1 per step after a failure), reward $1 - |\theta_h|/\theta_{\max}$, $H{=}40$, and ϵ-floored softmax linear controllers as behavior and target policies. It is worth reporting because CDR does *not* work well there out of the box, for two instructive reasons. *(i) Deterministic dynamics make a tree MDP.* With the standard noiseless dynamics and Gymnasium's ± 0.05 initial box, each state identifies its action history almost exactly, so the marginal ratio w_h collapses to the cumulative importance ratio (Remark 5): there is no state merging for w_h to exploit, and no smooth regressor can represent it. In a pilot ($n{=}2000$, 20 runs) CDR with the random-feature nuisances of Experiment 8 had bias $(-1.7, -1.3)$ on $\theta^* = (33.9,\ 0.13)$, whereas step-wise DR—which needs no w—was un-

biased with joint MSE 0.065. *(ii) With stochastic dynamics the product-rate condition is the binding constraint.* Adding a Gaussian disturbance force (σ=6 N) and a broad initial distribution restores state merging, and the learned $\hat{w}$ then recovers target-policy occupancy moments to within 0.01; but the value functions have sharp features (a kinked reward, an indicator cost, a failure boundary) that the random-feature class approximates poorly. Fitting the nuisances on $50{,}000$ behavior trajectories, the CDR bias on $\theta^* = (32.3,\ 2.77)$ is $(-0.18, -0.20)$ with 128 features and ridge 10^{-3}, $(-0.10, -0.13)$ with 256 features and 10^{-4}, and $(-0.06, -0.08)$ with 512 features and 10^{-5}: it shrinks as the class grows but does not vanish at any budget we could afford, and it is unchanged between $1{,}600$ and $50{,}000$ training trajectories, i.e. it is approximation error, exactly the "coarse class" regime of Experiment 8. Step-wise DR is again nearly unbiased ($(-0.04, -0.03)$) and, because the two controllers are close, its variance is only $1.6\times$ CDR's, so its MSE (0.042) beats CDR's bias-dominated 0.12 at n=2000. The lesson is the one the theory states: the marginal-ratio approach buys its escape from the curse of horizon with two nuisance functions, and when their product error is not $o_p(n^{-1/2})$—here because the value functions are outside the reach of a generic smooth class—the certificate is not valid. Nuisance classes adapted to such value functions (or neural nuisances with enough data) are needed before CDR should be trusted on this kind of task; we did not pursue them here, and report the Taxi benchmark (Experiment 12), where tabular nuisances are consistent (though, at the sample sizes run, CDR-data's cost covariance is still far from Σ^*), as our standard-environment evaluation.

H Proof of Proposition 1: Exponential Variance of Step-wise DR

Proposition (Proposition 1, restated)**.** *Under Assumptions 1–2, even with the optimal control variate $\hat{Q} = Q^{\pi,\star}$, $\hat{V} = V^{\pi,\star}$, the step-wise (JL) estimator satisfies*

$$n \operatorname{Var}(\hat{v}^{\pi}_{\star,\mathrm{JL}}) = \operatorname{Var}(V_1^{\pi,\star}(s_1)) + \sum_{h=1}^{H} \gamma^{2(h-1)}\, \mathbb{E}\big[\varrho^2_{1:h-1}\, \sigma_h^2(s_h)\big]\,,$$

where $\sigma_h^2(s) \coloneqq \gamma^2 \sum_a \frac{\pi(a|s)^2}{\pi_0(a|s)} \operatorname{Var}_{P_h(\cdot|s,a)}(V_{h+1}^{\pi,\star}(s'))$ (absorbing the step-h weight ϱ_h^2) and all cross-step covariances vanish. If $\underline{\sigma}^2 \coloneqq \inf_{h<H,\,s} \sigma_h^2(s) > 0$ and $\gamma^2(1+\underline{\chi}^2) > 1$, with $1+\underline{\chi}^2 \coloneqq \inf_{h,s} \mathbb{E}[\varrho_h^2 \mid s_h = s]$ (infimum over the support of $d_h^{\pi_0}$), the variance grows exponentially in H.

Proof. Unrolling the recursion (4), the step-wise DR estimator of Jiang & Li (2016) is

$$\begin{aligned}\hat{v}^{\pi}_{\star,\mathrm{JL}} = \frac{1}{n}\sum_{i=1}^{n}\Big[&\hat{V}_1^{\star}(s_{1,i}) \\ &+ \sum_{h=1}^{H} \gamma^{h-1} \varrho^{(i)}_{1:h}\big(f_h^{\star}(s_{h,i},a_{h,i}) + \gamma \hat{V}_{h+1}^{\star}(s_{h+1,i}) \\ &\qquad - \hat{Q}_h^{\star}(s_{h,i},a_{h,i})\big)\Big],\end{aligned}$$

where $\varrho_{1:h} = \prod_{t=1}^{h} \pi(a_t|s_t)/\pi_0(a_t|s_t)$. With the optimal control variate $\hat{Q} = Q^{\pi,\star}$, $\hat{V} = V^{\pi,\star}$, each Bellman residual becomes $\gamma\epsilon_h$ with $\epsilon_h \coloneqq V_{h+1}^{\pi,\star}(s_{h+1}) - \mathbb{E}[V_{h+1}^{\pi,\star}(s') \mid s_h, a_h]$, which satisfies $\mathbb{E}[\epsilon_h \mid s_h, a_h] = 0$. The single-trajectory estimation error is $\xi_{\mathrm{JL}} = V_1^{\pi,\star}(s_1) - v_\star^{\pi} + \sum_{h=1}^{H} \gamma^h \varrho_{1:h}\, \epsilon_h$.

Cross-term cancellation. For any $h' > h$,

$$\begin{aligned}&\mathbb{E}[\varrho_{1:h}\epsilon_h \cdot \varrho_{1:h'}\epsilon_{h'}] \\ &\quad = \mathbb{E}\big[\varrho_{1:h}\epsilon_h \cdot \varrho_{1:h'-1} \cdot \underbrace{\mathbb{E}[\varrho_{h'}\epsilon_{h'} \mid s_{h'}]}_{=0}\big] = 0,\end{aligned}$$

since $\mathbb{E}[\varrho_{h'}\epsilon_{h'} \mid s_{h'}] = \sum_a \pi(a|s_{h'})\, \mathbb{E}[\epsilon_{h'} \mid s_{h'}, a] = 0$ by the definition of $Q^{\pi,\star}$. The same argument shows $\mathbb{E}[(V_1^{\pi,\star}(s_1) - v_\star^{\pi})\, \varrho_{1:h}\epsilon_h] = 0$ by conditioning on s_1. Hence

$$n \operatorname{Var}(\hat{v}^{\pi}_{\star,\mathrm{JL}}) = \operatorname{Var}(V_1^{\pi,\star}(s_1)) + \sum_{h=1}^{H} \gamma^{2h}\, \mathbb{E}[\varrho^2_{1:h}\, \epsilon_h^2].$$

Conditional Bellman noise. Define $\sigma_h^2(s) \coloneqq \gamma^2\,\mathbb{E}[\varrho_h^2\,\epsilon_h^2 \mid s_h = s] = \gamma^2 \sum_a \frac{\pi(a|s)^2}{\pi_0(a|s)} \operatorname{Var}_{P_h(\cdot|s,a)}(V_{h+1}^{\pi,\star}(s'))$. Then $\gamma^{2h}\,\mathbb{E}[\varrho_{1:h}^2\,\epsilon_h^2] = \gamma^{2(h-1)}\,\mathbb{E}[\varrho_{1:h-1}^2\,\sigma_h^2(s_h)]$ by iterated conditioning: conditional on (s_h, a_h), the residual ϵ_h depends only on s_{h+1} and is independent of $\varrho_{1:h-1}$ (which is a $\sigma(s_1, a_1, \ldots, s_h, a_h)$-measurable function); using $\gamma^2\,\mathbb{E}_{\pi_0}[\varrho_h^2\,\epsilon_h^2 \mid s_h] = \sigma_h^2(s_h)$ then yields the stated identity.

Exponential growth: the χ^2 threshold. The conditional second moment of the per-step ratio is *exactly* a χ^2-divergence, not merely bounded below by one:

$$\begin{aligned}\mathbb{E}[\varrho_h^2 \mid s_h = s] &= \sum_a \pi_0(a|s) \left(\frac{\pi(a|s)}{\pi_0(a|s)}\right)^2 = \sum_a \frac{\pi(a|s)^2}{\pi_0(a|s)} \\ &= 1 + \sum_a \frac{\big(\pi(a|s) - \pi_0(a|s)\big)^2}{\pi_0(a|s)} = 1 + \chi^2\big(\pi(\cdot|s) \,\|\, \pi_0(\cdot|s)\big),\end{aligned}$$

where the third equality expands the square and uses $\sum_a \pi(a|s) = \sum_a \pi_0(a|s) = 1$. Writing $1+\underline{\chi}^2 \coloneqq \inf_{h,s} \mathbb{E}[\varrho_h^2 \mid s_h = s]$ (infimum over the support of $d_h^{\pi_0}$) and conditioning on s_h, $\mathbb{E}[\varrho_{1:h}^2] = \mathbb{E}\big[\varrho_{1:h-1}^2\,\mathbb{E}[\varrho_h^2 \mid s_h]\big] \geq (1+\underline{\chi}^2)\,\mathbb{E}[\varrho_{1:h-1}^2]$, so by induction

$$\mathbb{E}[\varrho_{1:h}^2] \;\geq\; (1+\underline{\chi}^2)^h,$$

and, with $1+\overline{\chi}^2 \coloneqq \sup_{h,s} \mathbb{E}[\varrho_h^2 \mid s_h = s]$, likewise $\mathbb{E}[\varrho_{1:h}^2] \leq (1+\overline{\chi}^2)^h$. Let $\underline{\sigma}^2 \coloneqq \inf_{h<H,\,s} \sigma_h^2(s)$ and $\overline{\sigma}^2 \coloneqq \sup_{h,s} \sigma_h^2(s)$ (the last step is excluded from the infimum because $V_{H+1}^{\pi,\star} \equiv 0$ forces $\sigma_H^2 \equiv 0$). Combining with the $\gamma^{2(h-1)}$ weight carried by the h-th summand,

$$\underline{\sigma}^2 \sum_{h=1}^{H-1} \big(\gamma^2(1+\underline{\chi}^2)\big)^{h-1} \;\leq\; n \operatorname{Var}(\hat{v}_{\star,\mathrm{JL}}^{\pi}) \;\leq\; \operatorname{Var}(V_1^{\pi,\star}(s_1)) + \overline{\sigma}^2 \sum_{h=1}^{H} \big(\gamma^2(1+\overline{\chi}^2)\big)^{h-1},$$

two geometric series. The lower one diverges exponentially in H when $\underline{\sigma}^2 > 0$ and $\gamma^2(1+\underline{\chi}^2) > 1$, i.e. $\underline{\chi}^2 > \gamma^{-2} - 1$; writing $\gamma = 1 - \varepsilon$ and expanding, $\gamma^{-2} - 1 = 2\varepsilon + O(\varepsilon^2)$, giving the threshold $\underline{\chi}^2 \gtrsim 2/H_{\mathrm{eff}}$ stated in Proposition 1. At $\gamma = 1$ the condition is simply $\underline{\chi}^2 > 0$, i.e. (for finite $\mathcal{S}$) $\pi_h(\cdot|s) \neq \pi_{0,h}(\cdot|s)$ at every visited (h, s), and for general $\mathcal{S}$ a χ^2-divergence bounded away from 0; the upper series shows that the growth is in any case no faster than $(\gamma^2(1+\overline{\chi}^2))^H$, so the two χ^2-extremes pin the exponent. (The condition is sufficient, not necessary: if π and π_0 agree on some states, the variance can still grow exponentially through the states on which they differ, but the clean geometric lower bound no longer applies.) As a special case, if $\pi_h(\cdot|s)$ puts unit mass on $a_h^*(s)$, then $\mathbb{E}[\varrho_h^2 \mid s_h = s] = 1/\pi_{0,h}(a_h^*(s)|s) \in [1, 1/\eta_{\min}]$, so $\mathbb{E}[\varrho_{1:h}^2] \geq \bar{p}^{-h}$ with $\bar{p} \coloneqq \sup_{h,s} \pi_{0,h}(a_h^*(s)|s)$, which grows exponentially whenever $\bar{p} < 1$. Conversely, Assumption 1 caps every per-step factor at $1/\eta_{\min}$, so $\mathbb{E}[\varrho_{1:h}^2] \leq \eta_{\min}^{-h}$. The uniform lower bound $\underline{\sigma}^2 > 0$ holds whenever there is $c > 0$ with $\operatorname{Var}_{P_h(\cdot|s,a)}(V_{h+1}^{\pi,\star}) \geq c$ for every visited (h, s, a), $h < H$, with $\pi_h(a|s) > 0$ (then $\sigma_h^2(s) \geq \gamma^2 c$, since $\sum_a \pi^2/\pi_0 \geq 1$); for finite $\mathcal{S}$ it suffices that each such variance be positive. Stochasticity of P_h alone is not enough (constant rewards make $V_{h+1}^{\pi,\star}$ constant), and the condition is needed: with deterministic P_h, $n \operatorname{Var}(\hat{v}_{\star,\mathrm{JL}}^{\pi}) = \operatorname{Var}(V_1^{\pi,\star}(s_1))$ for every H. Horizon-exponential variance of per-decision importance weighting is well documented (Precup et al., 2000; Jiang & Li, 2016; Liu et al., 2018); the bound above makes the growth rate explicit under the stated conditions. By contrast, the semiparametric efficiency bound Σ_{00}^* is a sum of H within-step terms with no compounding factor (from (8)): each is at most $\gamma^2 M_w \sup_{s,a} \operatorname{Var}_{P_h}(V_{h+1}^{\pi,\star} \mid s, a)$ when $\sup_h \|w_h\|_\infty \leq M_w$ with M_w independent of H, so $\Sigma_{00}^* - \operatorname{Var}_\mu(V_1^{\pi,\star})$ grows linearly in H under a uniform per-step noise bound and at most cubically under Assumption 2 alone—polynomially for such M_w. If M_w itself grows with H (e.g. tree MDPs, Remark 5), this guarantee does not apply. □

I Proof of Theorem 2: Double Robustness (Telescoping)

Theorem (Theorem 2, restated)**.** *Under $\gamma = 1$, $\mathbb{E}[\psi^\star] = 0$ whenever either $Q_h^{\pi,\star}$ equals the true value function for every $h \in [H]$ (with $\mathbb{E}_{d_h^{\pi_0}}|w_h| < \infty$), or $w_h = d_h^\pi / d_h^{\pi_0}$ for every $h \in [H]$; throughout, $V_h^{\pi,\star}$ is the π-average of the Q argument and $V_{H+1}^{\pi,\star} \equiv 0$.*

Proof. **Case (i): $Q^{\pi,\star}$ correct, w arbitrary.** When $Q^{\pi,\star}$ is the true value function: (a) $V_h^{\pi,\star}$ is also correct since $V_h^{\pi,\star}(s) = \mathbb{E}_{a\sim\pi_h}[Q_h^{\pi,\star}(s,a)]$, so the boundary term contributes $\mathbb{E}_\mu[V_1^{\pi,\star}(s_1)] = v_\star^\pi$; (b) each Bellman residual satisfies $\mathbb{E}[\delta_h^\star \mid s_h, a_h] = 0$, hence by the tower property,

$$\mathbb{E}[w_h(s_h,a_h)\,\delta_h^\star] = \mathbb{E}\big[w_h(s_h,a_h)\,\mathbb{E}[\delta_h^\star \mid s_h,a_h]\big] = 0,$$

regardless of whether w is correct. Summing over h and combining with (a): $\mathbb{E}[\psi^\star] = v_\star^\pi + 0 - v_\star^\pi = 0$.

Case (ii): w correct, $Q^{\pi,\star}$ arbitrary. We now prove $\mathbb{E}[\psi^\star] = 0$ when w is correct and Q is misspecified as $\check{Q}^\star$.

Setup. Define the misspecified value function $\check{V}_h^\star(s) := \mathbb{E}_{a\sim\pi_h}[\check{Q}_h^\star(s,a)]$ and the misspecified residual $\check{\delta}_h^\star := f_h^\star(s_h,a_h) + \gamma\check{V}_{h+1}^\star(s_{h+1}) - \check{Q}_h^\star(s_h,a_h)$. The CDR score evaluated at misspecified $\check{Q}^\star$ is:

$$\psi^\star = \check{V}_1^\star(s_1) + \sum_{h=1}^H w_h(s_h,a_h)\,\check{\delta}_h^\star - v_\star^\pi.$$

Key identity. Since w is correct, for any function g_h:

$$\mathbb{E}_{d_h^{\pi_0}}[w_h\,g_h] = \mathbb{E}_{d_h^\pi}[g_h].$$

Telescoping. Taking expectations:

$$\mathbb{E}[\psi^\star] = \mathbb{E}_\mu[\check{V}_1^\star(s_1)] - v_\star^\pi + \sum_{h=1}^H \mathbb{E}_{d_h^\pi}\big[f_h^\star + \gamma\,\mathbb{E}_{P_h}[\check{V}_{h+1}^\star(s')] - \check{V}_h^\star(s_h)\big].$$

Define $A_h := \mathbb{E}_{d_h^\pi}[\check{V}_h^\star(s_h)]$ and $\bar{A}_{h+1} := \mathbb{E}_{d_h^\pi}[\mathbb{E}_{P_h}[\check{V}_{h+1}^\star(s')]] = A_{h+1}$. The identity $\bar{A}_{h+1} = A_{h+1}$ follows from the marginalization

$$\mathbb{E}_{d_h^\pi}[\mathbb{E}_{P_h}[\check{V}_{h+1}^\star(s')]] = \mathbb{E}_{d_{h+1}^\pi}[\check{V}_{h+1}^\star(s_{h+1})],$$

which uses $d_{h+1}^\pi(s') = \sum_{s,a} d_h^\pi(s,a)\,P_h(s'|s,a)$ (definition of the next-step occupancy under π). The Bellman-residual sum becomes a telescoping sum:

$$\sum_{h=1}^H \mathbb{E}_{d_h^\pi}[f_h^\star] + \gamma\sum_{h=1}^H A_{h+1} - \sum_{h=1}^H A_h = \sum_{h=1}^H \mathbb{E}_{d_h^\pi}[f_h^\star] + \sum_{h=1}^H(\gamma A_{h+1} - A_h).$$

With $\gamma = 1$ and $A_{H+1} = 0$ (since $\check{V}_{H+1}^\star = 0$):

$$\sum_{h=1}^H (A_{h+1} - A_h) = A_{H+1} - A_1 = -A_1.$$

Therefore:

$$\mathbb{E}[\psi^\star] = A_1 + \sum_{h=1}^H \mathbb{E}_{d_h^\pi}[f_h^\star] - A_1 - v_\star^\pi = 0,$$

where $v_\star^\pi = \sum_{h=1}^H \mathbb{E}_{d_h^\pi}[f_h^\star]$ by definition. □

Remark 20 (On $\gamma < 1$ in the finite horizon)**.** *If $\gamma < 1$, the telescoping sum yields the residual:*

$$\mathbb{E}[\psi^\star] = (\gamma-1)\textstyle\sum_{h=2}^H A_h + \sum_{h=2}^H(1-\gamma^{h-1})\,F_h,$$

where $F_h := \mathbb{E}_{d_h^\pi}[f_h^\star]$. With w correct, the second term equals $(1-\gamma)\sum_{h=2}^H \mathbb{E}_{d_h^\pi}[V_h^{\pi,\star}(s_h)]$, so $\mathbb{E}[\psi^\star] = (\gamma-1)\sum_{h=2}^H \mathbb{E}_{d_h^\pi}\big[\check{V}_h^\star(s_h) - V_h^{\pi,\star}(s_h)\big]$. Neither term alone is a bias: both are generically non-zero even at the true Q, where they cancel (Case (i) holds for every γ). The sum is non-zero whenever the d^π-averaged value errors at steps $2,\dots,H$ do not cancel, e.g. $\check{Q}_2^\star = Q_2^{\pi,\star} + c$ gives $\mathbb{E}[\psi^\star] = (\gamma-1)c \neq 0$, breaking case (ii) of double robustness. We adopt $\gamma = 1$ throughout; see Appendix U for an extended analysis and Appendix J for the discount-weighted remedy.

J DISCOUNT-WEIGHTED DR: DOUBLE ROBUSTNESS FOR ALL γ

This appendix shows that the entire theory of Sections 3 and 5—double robustness, Neyman orthogonality, the efficient influence function, and the efficiency bound—extends to $\gamma < 1$ once the score carries the discount weight that the target itself carries.

Definition 2 (Discount-weighted CDR score)**.** *Replace w_h by $\tilde{w}_h \coloneqq \gamma^{h-1} w_h$ in Definition 1:*

$$\tilde{\psi}^\star(Z;\, v_\star^\pi, Q^{\pi,\star}, w) \coloneqq V_1^{\pi,\star}(s_1) - v_\star^\pi + \sum_{h=1}^{H} \gamma^{h-1} w_h(s_h, a_h)\, \delta_h^\star. \tag{19}$$

At $\gamma = 1$ this reduces to (6)*.*

The weight γ^{h-1} is not a device. Since $v_\star^\pi = \sum_h \gamma^{h-1} \mathbb{E}_{d_h^\pi}[f_h^\star]$, the signal at step h enters the target discounted by γ^{h-1}; any correction that is to reconstruct $v_\star^\pi$ must therefore discount step h by the same factor. The failure of double robustness at $\gamma < 1$ documented in Remark 20 is exactly the mismatch between an undiscounted correction and a discounted target.

Proposition 13 (Double robustness for all γ)**.** *For every $\gamma \in (0, 1]$ and every component $\star$, $\mathbb{E}[\tilde{\psi}^\star(Z; v_\star^\pi, Q^{\pi,\star}, w)] = 0$ whenever either $Q^{\pi,\star}$ or w is correctly specified.*

Proof. **Case (i): $Q^{\pi,\star}$ correct, w arbitrary.** When $Q^{\pi,\star}$ is correct, $\mathbb{E}[\delta_h^\star \mid s_h, a_h] = 0$, so for any w_h:

$$\mathbb{E}[\gamma^{h-1} w_h \delta_h^\star] = \gamma^{h-1} \mathbb{E}\big[w_h \mathbb{E}[\delta_h^\star \mid s_h, a_h]\big] = 0.$$

Hence $\mathbb{E}[\tilde{\psi}^\star] = 0$.

Case (ii): w correct, $Q^{\pi,\star}$ arbitrary. Let $\check{Q}^\star$ be the misspecified Q, with $\check{V}_h^\star$ as in Appendix I, and $A_h \coloneqq \mathbb{E}_{d_h^\pi}[\check{V}_h^\star(s_h)]$. By the change-of-measure identity:

$$\begin{aligned}\mathbb{E}[\tilde{\psi}^\star] &= A_1 - v_\star^\pi \\ &\quad + \sum_{h=1}^{H} \big(\gamma^h A_{h+1} - \gamma^{h-1} A_h\big) + \sum_{h=1}^{H} \gamma^{h-1} \mathbb{E}_{d_h^\pi}[f_h^\star].\end{aligned}$$

The telescoping sum gives $\sum_{h=1}^{H} (\gamma^h A_{h+1} - \gamma^{h-1} A_h) = \gamma^H A_{H+1} - A_1 = -A_1$ (using $A_{H+1} = 0$). Therefore:

$$\mathbb{E}[\tilde{\psi}^\star] = \sum_{h=1}^{H} \gamma^{h-1} \mathbb{E}_{d_h^\pi}[f_h^\star] - v_\star^\pi = 0,$$

using the definition $v_\star^\pi = \sum_{h=1}^{H} \gamma^{h-1} \mathbb{E}_{d_h^\pi}[f_h^\star]$. For $\gamma = 1$, $\tilde{\psi}^\star$ reduces to the original CDR score. □

NEYMAN ORTHOGONALITY FOR ALL γ

Proposition 14 (Orthogonality of the discount-weighted score)**.** *For every $\gamma \in (0, 1]$, the joint score $\tilde{\psi} = (\tilde{\psi}^r, \ldots, \tilde{\psi}^{c_J})^\top$ satisfies $\partial_\epsilon\, \mathbb{E}[\tilde{\psi}(Z; \theta^*, \eta^* + \epsilon\, \Delta\eta)]\big|_{\epsilon=0} = \mathbf{0}$ in all directions $\Delta\eta$.*

Proof. **Δw direction.** At the true $Q^{\pi,\star}$ we have $\mathbb{E}[\delta_h^\star | s_h, a_h] = 0$, so the derivative equals $\sum_h \gamma^{h-1} \mathbb{E}[\Delta w_h\, \mathbb{E}[\delta_h^\star | s_h, a_h]] = 0$, exactly as in Appendix K; the constants γ^{h-1} play no role.

ΔQ direction. Write $\Delta V_h(s) \coloneqq \mathbb{E}_{a \sim \pi_h}[\Delta Q_h(s, a)]$ and $D_h \coloneqq \mathbb{E}_{d_h^\pi}[\Delta V_h]$. Using the change-of-measure identity $\mathbb{E}_{d_h^{\pi_0}}[w_h^*\, g] = \mathbb{E}_{d_h^\pi}[g]$ together with $\mathbb{E}_{d_h^\pi}\big[\mathbb{E}_{P_h}[\Delta V_{h+1}(s')]\big] = D_{h+1}$,

$$\partial_\epsilon\, \mathbb{E}[\tilde{\psi}^\star]\big|_{\epsilon=0} = D_1 + \sum_{h=1}^{H} \big(\gamma^h D_{h+1} - \gamma^{h-1} D_h\big) = D_1 + \big(\gamma^H D_{H+1} - D_1\big) = 0,$$

using $\Delta V_{H+1} \equiv 0$. Note that the sum telescopes for *every* γ precisely because the score carries γ^{h-1}; with w_h in place of $\gamma^{h-1} w_h$ the sum would be $\sum_h (\gamma D_{h+1} - D_h)$, which does not collapse.

Cross-components. $\tilde{\psi}^{\star_k}$ does not depend on $Q^{\pi,\star_j}$ for $j \neq k$, so all cross-derivatives vanish. □

EFFICIENT INFLUENCE FUNCTION AND EFFICIENCY BOUND FOR ALL γ

Theorem 15 (EIF and bound under discounting)**.** *For every $\gamma \in (0, 1]$, the EIF for θ^* in the nonparametric CMDP model is $\tilde{\psi}^*(Z) = \tilde{\psi}(Z; \theta^*, \eta^*)$, and the efficiency bound $\tilde{\Sigma}^* := \mathbb{E}[\tilde{\psi}^*(Z)\tilde{\psi}^*(Z)^\top]$ has entries*

$$\tilde{\Sigma}^*_{jk} = \mathrm{Cov}_\mu\big(V_1^{\pi,\star_j}, V_1^{\pi,\star_k}\big) + \sum_{h=1}^{H} \gamma^{2h}\, \mathbb{E}_{d_h^{\pi_0}}\Big[w_h^2\, \mathrm{Cov}_{P_h}\big(V_{h+1}^{\pi,\star_j}(s'), V_{h+1}^{\pi,\star_k}(s') \mid s_h, a_h\big)\Big]. \tag{20}$$

Proof. The argument is that of Appendix L with the discount weights carried through; we indicate only the two places where γ enters.

Riesz representation along P_h. For the submodel $P_h^{(\epsilon)} = P_h(1 + \epsilon g_h)$ with $\mathbb{E}_{P_h}[g_h|s, a] = 0$, perturbing P_h shifts $Q_h^{\pi,\star}$ by $\epsilon\gamma\, \mathrm{Cov}_{P_h}(V_{h+1}^{\pi,\star}, g_h)$, and this shift reaches $v_\star^\pi$ discounted by γ^{h-1}, so

$$\partial_\epsilon\, v^\pi_{\star,(\epsilon)}\big|_{\epsilon=0} = \gamma^h\, \mathbb{E}_{d_h^\pi}\big[\mathrm{Cov}_{P_h}(V_{h+1}^{\pi,\star}, g_h \mid s_h, a_h)\big].$$

On the score side only the h-th summand of (19) contributes (the initial term and summands $t < h$ are functions of $(s_1, a_1, \ldots, s_h, a_h)$ and $\mathbb{E}[g_h \mid s_1, \ldots, s_h, a_h] = 0$ by the Markov property; for $t > h$, g_h is $\sigma(s_1, \ldots, s_t, a_t)$-measurable and $\mathbb{E}[\delta_t^\star \mid s_t, a_t] = 0$; within the h-th summand, $f_h^\star - Q_h^{\pi,\star}$ is orthogonal to g_h), giving $\gamma^{h-1}\mathbb{E}_{d_h^{\pi_0}}[w_h \cdot \gamma\, \mathrm{Cov}(V_{h+1}, g_h)] = \gamma^h \mathbb{E}_{d_h^\pi}[\mathrm{Cov}(V_{h+1}, g_h)]$ after change of measure, which matches.

Orthogonality to T_μ and T_{π_0}. The three-case argument of Appendix L uses only $\mathbb{E}[\delta_h^\star|s_h, a_h] = 0$ and the conditional mean-zero property of the π_0-scores; the factors γ^{h-1} are deterministic constants and pass through unchanged.

Covariance. Cross-step terms vanish by the same martingale-difference argument, and the h-th same-step term picks up $(\gamma^{h-1})^2$ from the score and a further γ^2 from $\mathrm{Cov}(\delta_h^{\star_j}, \delta_h^{\star_k} \mid s_h, a_h) = \gamma^2 \mathrm{Cov}_{P_h}(V_{h+1}^{\star_j}, V_{h+1}^{\star_k})$, for a total weight γ^{2h} as stated. □

Corollary 16 (Horizon-free variance under discounting)**.** *Under Assumptions 1–2 and $\sup_h \|w_h\|_\infty \le M_w$ with M_w independent of H, for any $\gamma < 1$ the diagonal entries of* (20) *satisfy*

$$\tilde{\Sigma}^*_{jj} \;\le\; \mathrm{Var}_\mu\big(V_1^{\pi,\star_j}\big) + \frac{\gamma^2 M_w}{(1-\gamma)^2\,(1-\gamma^2)},$$

which is bounded uniformly in *H (at fixed M_w; on tree MDPs M_w can be $\eta_{\min}^{-H}$ and $\tilde{\Sigma}^*_{00}$ coincides with the JL variance). By contrast, the $\gamma = 1$ bound of Theorem 3 grows polynomially in H (for H-uniform M_w; linearly under a uniform per-step Bellman-noise bound, Proposition 1), and* (5) *grows exponentially whenever $\gamma^2(1 + \underline{\chi}^2) > 1$.*

Proof. Each summand of (20) equals $\gamma^{2h}\mathbb{E}_{d_h^\pi}[w_h \mathrm{Var}_{P_h}(V_{h+1}^{\pi,\star_j} \mid s_h, a_h)]$ after the change of measure $\mathbb{E}_{d_h^{\pi_0}}[w_h^2 g] = \mathbb{E}_{d_h^\pi}[w_h g]$, hence is at most $\gamma^{2h} M_w \sup_s |V_{h+1}^{\pi,\star_j}(s)|^2 \le \gamma^{2h} M_w/(1-\gamma)^2$ by Assumption 2, and $\sum_{h\ge1} \gamma^{2h} = \gamma^2/(1-\gamma^2)$. □

Remark 21 (Asymptotic normality)**.** *Theorem 6 carries over, with $\tilde{\Sigma}^*$ of Theorem 15 in place of Σ^*, for the discount-weighted estimator, i.e. Algorithm 1 with $\hat{w}_h^{(k(i))}$ replaced by $\gamma^{h-1}\hat{w}_h^{(k(i))}$. The only step of Appendix M written for $\gamma = 1$, the telescoping of the first-order $\hat{Q} - Q$ term, holds with these weights by Proposition 14; the other steps are unchanged. The boundedness constants of Assumption 4 are, if anything, improved, since $\gamma^{h-1} \le 1$. The calibration analysis of Section 4 is likewise unaffected: Theorem 5 depends only on the sign and magnitude of the off-diagonal entries, not on how they are weighted across steps.*

K Proof of Theorem 9: Neyman Orthogonality

Theorem (Theorem 9, restated)**.** *Under $\gamma = 1$, the joint CDR score satisfies Neyman orthogonality: $\partial_\epsilon \mathbb{E}[\psi(Z; \theta^*, \eta^* + \epsilon\,\Delta\eta)]|_{\epsilon=0} = \mathbf{0}$ for all perturbation directions $\Delta\eta$.*

Proof. **Note.** Neyman orthogonality is verified at the truth (w^*, Q^*): we show that the directional derivative of $\mathbb{E}[\psi]$ at this point is zero in all L^2 directions $(\Delta w, \Delta Q)$.

Direction $\Delta Q^{\star_j}$ (component j). Perturbing $Q^{\pi,\star_j} \to Q^{\pi,\star_j} + \epsilon\,\Delta Q^{\star_j}$ induces $V_h^{\pi,\star_j} \to V_h^{\pi,\star_j} + \epsilon\,\Delta V_h^{\star_j}$ where $\Delta V_h^{\star_j}(s) = \mathbb{E}_{a\sim\pi_h}[\Delta Q_h^{\star_j}(s,a)]$. The perturbed CDR score for component j is:

$$\begin{aligned}\psi_\epsilon^{\star_j} &= (V_1^{\pi,\star_j} + \epsilon\,\Delta V_1^{\star_j})(s_1) - v_{\star_j}^{\pi} \\ &\quad + \sum_{h=1}^{H} w_h^* \big[\delta_h^{\star_j} + \epsilon(\gamma\,\Delta V_{h+1}^{\star_j}(s_{h+1}) \\ &\qquad - \Delta Q_h^{\star_j}(s_h, a_h))\big].\end{aligned}$$

Taking $\partial_\epsilon|_{\epsilon=0}$ and applying $\mathbb{E}[\cdot]$:

$$\begin{aligned}&\partial_\epsilon \mathbb{E}[\psi_\epsilon^{\star_j}]\big|_{\epsilon=0} \\ &= \mathbb{E}_\mu[\Delta V_1^{\star_j}] + \sum_{h=1}^{H} \mathbb{E}_{d_h^{\pi_0}}\big[w_h^*\big(\gamma\,\Delta V_{h+1}^{\star_j}(s_{h+1}) \\ &\qquad - \Delta Q_h^{\star_j}(s_h, a_h)\big)\big].\end{aligned}$$

Applying the density-ratio identity $\mathbb{E}_{d_h^{\pi_0}}[w_h^* f] = \mathbb{E}_{d_h^\pi}[f]$, its next-state form $\mathbb{E}_{\pi_0}[w_h^*\, g(s_{h+1})] = \mathbb{E}_{d_h^\pi}[(P_h g)(s_h, a_h)] = \mathbb{E}_{d_{h+1}^\pi}[g]$, and $\mathbb{E}_{d_h^\pi}[\Delta Q_h^{\star_j}] = \mathbb{E}_{d_h^\pi}[\Delta V_h^{\star_j}]$:

$$= \mathbb{E}_\mu[\Delta V_1^{\star_j}] + \sum_{h=1}^{H} \big(\gamma\,\mathbb{E}_{d_{h+1}^\pi}[\Delta V_{h+1}^{\star_j}] - \mathbb{E}_{d_h^\pi}[\Delta V_h^{\star_j}]\big).$$

With $\gamma = 1$, the sum telescopes:

$$\begin{aligned}&\sum_{h=1}^{H} \big(\mathbb{E}_{d_{h+1}^\pi}[\Delta V_{h+1}^{\star_j}] - \mathbb{E}_{d_h^\pi}[\Delta V_h^{\star_j}]\big) \\ &\quad = \mathbb{E}_{d_{H+1}^\pi}[\Delta V_{H+1}^{\star_j}] - \mathbb{E}_{d_1^\pi}[\Delta V_1^{\star_j}] = -\mathbb{E}_\mu[\Delta V_1^{\star_j}],\end{aligned}$$

where $\Delta V_{H+1}^{\star_j} = 0$ and $d_1^\pi = \mu$. This cancels the boundary term, giving $\partial_\epsilon \mathbb{E}[\psi_\epsilon^{\star_j}]|_{\epsilon=0} = 0$.

Direction Δw. Perturbing $w_h^* \to w_h^* + \epsilon\,\Delta w_h$:

$$\begin{aligned}\partial_\epsilon \mathbb{E}[\psi_\epsilon^{\star_j}]\big|_{\epsilon=0} &= \sum_{h=1}^{H} \mathbb{E}_{d_h^{\pi_0}}[\Delta w_h \cdot \delta_h^{\star_j}] \\ &= \sum_{h=1}^{H} \mathbb{E}_{d_h^{\pi_0}}\big[\Delta w_h \cdot \mathbb{E}[\delta_h^{\star_j} \mid s_h, a_h]\big] = 0,\end{aligned}$$

since at the true $Q^{\pi,\star_j}$ we have $\mathbb{E}[\delta_h^{\star_j} | s_h, a_h] = 0$.

Cross-component derivatives. $\partial_{\Delta Q^{\star_j}} \psi^{\star_k} = 0$ for $j \neq k$ because $\psi^{\star_k}$ depends on $Q^{\pi,\star_k}$ and w but not on $Q^{\pi,\star_j}$. □

L Proof of Theorem 3: EIF Derivation

Theorem (Theorem 3, restated)**.** *Under the standing convention $\gamma = 1$, the EIF for $\theta^* \in \mathbb{R}^{J+1}$ in the nonparametric, time-inhomogeneous CMDP model is $\psi^*(Z) = \psi(Z; \theta^*, \eta^*)$, with efficiency bound $\Sigma^* = \mathbb{E}[\psi^*(Z)\psi^*(Z)^\top]$ given by* (8)*.*

Proof. **Pathwise derivative along P_h.** Consider the submodel $P_h^{(\epsilon)}(s'|s,a) = P_h(s'|s,a)(1+\epsilon g_h(s,a,s'))$ with $\mathbb{E}_{P_h}[g_h|s,a]=0$. Perturbing P_h shifts $Q_h^{\pi,\star}(s,a)$ by $\epsilon\gamma\,\mathrm{Cov}_{P_h}(V_{h+1}^{\pi,\star}, g_h \mid s,a)$ (the value functions $V_t^{\pi,\star}$, $t>h$, do not depend on P_h), and this shift propagates to $v_\star^\pi = \mathbb{E}_\mu[V_1^{\pi,\star}]$ through the $h-1$ earlier Bellman steps, each contributing a factor γ and one step of the occupancy recursion, so the pathwise derivative of the target is

$$\begin{aligned}\partial_\epsilon v^\pi_{\star,(\epsilon)}|_{\epsilon=0} &= \gamma^h\,\mathbb{E}_{d_h^\pi}\big[\mathrm{Cov}_{P_h}\big(V_{h+1}^{\pi,\star}(s'), g_h(s_h,a_h,s') \mid s_h,a_h\big)\big]\\ &= \mathbb{E}_{d_h^\pi}\big[\mathrm{Cov}_{P_h}\big(V_{h+1}^{\pi,\star}(s'), g_h(s_h,a_h,s') \mid s_h,a_h\big)\big] \quad \text{at the standing convention } \gamma=1.\end{aligned}$$

The Riesz representation is verified: in the sum defining $\psi^{\star,*}$, only the h-th term contributes. Indeed, the initial term and the summands $t<h$ are functions of $(s_1,a_1,\dots,s_h,a_h)$ and $\mathbb{E}[g_h \mid s_1,\dots,s_h,a_h]=0$ by the Markov property; for $t>h$, g_h is $\sigma(s_1,\dots,s_t,a_t)$-measurable and $\mathbb{E}[\delta_t^\star \mid s_t,a_t]=0$; and within the h-th summand the (s_h,a_h)-measurable part $f_h^\star - Q_h^{\pi,\star}$ is orthogonal to g_h (which has zero conditional mean given (s_h,a_h)). The remaining term equals:

$$\begin{aligned}&\gamma\,\mathbb{E}_{d_h^{\pi_0}}[w_h\,V_{h+1}^{\pi,\star}(s_{h+1})\,g_h(s_h,a_h,s_{h+1})]\\ &\quad=\gamma\,\mathbb{E}_{d_h^\pi}[V_{h+1}^{\pi,\star}(s_{h+1})\,g_h],\end{aligned}$$

which matches the pathwise derivative at $\gamma=1$. (For $\gamma<1$ the two sides differ by the factor γ^{h-1}, which is exactly why the discount-weighted score of Theorem 15 carries $\gamma^{h-1}w_h$ in place of w_h.)

Along μ (initial distribution). Consider the submodel $\mu^{(\epsilon)}(s)=\mu(s)(1+\epsilon g_0(s))$ with $\mathbb{E}_\mu[g_0]=0$. The pathwise derivative of the component target $v_\star^\pi = \mathbb{E}_\mu[V_1^{\pi,\star}]$ is $\partial_\epsilon \mathbb{E}_{\mu^{(\epsilon)}}[V_1^{\pi,\star}]\big|_{\epsilon=0} = \mathbb{E}_\mu[g_0\,V_1^{\pi,\star}] = \mathbb{E}_\mu[g_0(V_1^{\pi,\star}-v_\star^\pi)]$ (the last step uses $\mathbb{E}_\mu[g_0]=0$). On the score side, $\langle\psi^{\star,*},g_0\rangle = \mathbb{E}[(V_1^{\pi,\star}(s_1)-v_\star^\pi)\,g_0(s_1)] + \sum_{h\ge1}\mathbb{E}[w_h\delta_h^\star\,g_0(s_1)]$, and each summand vanishes because $g_0(s_1)$ is $\sigma(s_1,\dots,s_h,a_h)$-measurable and $\mathbb{E}[\delta_h^\star \mid s_h,a_h]=0$. Hence $\langle\psi^{\star,*},g_0\rangle = \mathbb{E}_\mu[g_0(V_1^{\pi,\star}-v_\star^\pi)]$, matching the T_μ pathwise derivative. This verifies the Riesz matching for the T_μ and $\bigoplus_h T_{P_h}$ directions; the T_{π_0} direction is treated next.

Along π_0. $V^{\pi,\star}$ does not depend on π_0, so the pathwise derivative of θ^* along any $g_h^\pi \in T_{\pi_0}$ is 0. This shows the π_0-component of the pathwise derivative vanishes; Riesz matching along T_{π_0} therefore requires $\langle\psi^*, g_{h'}^\pi\rangle_{L^2(P)}=0$ for every score $g_{h'}^\pi$ of π_0 (i.e., with $\mathbb{E}_{\pi_0(\cdot|s_{h'})}[g_{h'}^\pi(s_{h'},\cdot)]=0$ for all $s_{h'}$). Split $\langle w_h\delta_h^\star, g_{h'}^\pi\rangle$ by the relative order of h,h':

(i) $h'<h$ (cross-step, $g_{h'}^\pi$ earlier): by the tower property, condition on (s_h,a_h) in the inner expectation; since $\mathbb{E}[\delta_h^\star \mid s_h,a_h]=0$, the product vanishes: $\mathbb{E}[w_h\delta_h^\star\cdot g_{h'}^\pi(s_{h'},a_{h'})] = \mathbb{E}[w_h\,g_{h'}^\pi(s_{h'},a_{h'})\,\underbrace{\mathbb{E}[\delta_h^\star \mid s_h,a_h]}_{=0}]=0$.

(ii) $h'=h$ (same-step): $\mathbb{E}[w_h\delta_h^\star\cdot g_h^\pi(s_h,a_h)] = \mathbb{E}[w_h\,g_h^\pi(s_h,a_h)\,\underbrace{\mathbb{E}[\delta_h^\star \mid s_h,a_h]}_{=0}]=0$.

(iii) $h'>h$ (cross-step, $g_{h'}^\pi$ later): condition on $s_{h'}$ in the inner expectation; since $\mathbb{E}_{\pi_0(\cdot|s_{h'})}[g_{h'}^\pi(s_{h'},\cdot)]=0$, the product vanishes: $\mathbb{E}[w_h\delta_h^\star\cdot g_{h'}^\pi(s_{h'},a_{h'})] = \mathbb{E}[w_h\delta_h^\star\,\underbrace{\mathbb{E}_{\pi_0(\cdot|s_{h'})}[g_{h'}^\pi(s_{h'},\cdot)]}_{=0}]=0$.

The initial-state term $V_1^{\pi,\star}(s_1)-v_\star^\pi$ depends only on s_1 and is likewise orthogonal to $g_{h'}^\pi$ for any $h'\ge1$ by the conditional mean-zero property of $g_{h'}^\pi$ given $s_{h'}$. Hence $\psi^*\perp T_{\pi_0}$, which is the Riesz matching along T_{π_0}.

Membership. $V_1^{\pi,\star}(s_1)-v_\star^\pi$ is a mean-zero function of s_1, hence in T_μ; each $w_h\delta_h^\star$ is a square-integrable function of (s_h,a_h,s_{h+1}) with $\mathbb{E}[w_h\delta_h^\star \mid s_h,a_h]=0$, hence in T_{P_h}. So $\psi^{\star,*}\in T_\mu\oplus\bigoplus_h T_{P_h}\subseteq\overline{\dot{\mathcal{P}}}$, and ψ^* is the EIF.

Covariance formula. Expanding $\mathbb{E}[\psi^{\star_j,*}\psi^{\star_k,*}]$: cross-step terms ($h\neq h'$) vanish by the tower property and $\mathbb{E}[\delta_h^{\star_j}|s_h,a_h]=0$.

For $h < h'$:

$$\mathbb{E}\big[w_h^* \delta_h^{\star j} \cdot w_{h'}^* \delta_{h'}^{\star k}\big] = \mathbb{E}\big[w_h^* \delta_h^{\star j} \cdot w_{h'}^* \cdot \underbrace{\mathbb{E}[\delta_{h'}^{\star k} \mid s_{h'}, a_{h'}]}_{=0}\big] = 0.$$

For $h > h'$:

$$\mathbb{E}\big[w_h^* \delta_h^{\star j} \cdot w_{h'}^* \delta_{h'}^{\star k}\big] = \mathbb{E}\big[w_{h'}^* \delta_{h'}^{\star k} \cdot w_h^* \cdot \underbrace{\mathbb{E}[\delta_h^{\star j} \mid s_h, a_h]}_{=0}\big] = 0.$$

For the same-step term ($h = h'$): since $\mathbb{E}[\delta_h^{\star j} | s_h, a_h] = 0$,

$$\mathbb{E}[w_h^2 \delta_h^{\star j} \delta_h^{\star k}] = \mathbb{E}_{d_h^{\pi_0}}[w_h^2 \mathrm{Cov}_{P_h}(\delta_h^{\star j}, \delta_h^{\star k} \mid s_h, a_h)].$$

The initial-state term contributes $\mathrm{Cov}_\mu(V_1^{\pi,\star_j}, V_1^{\pi,\star_k})$. Combining yields (8). With stochastic signals, (8) holds verbatim with $\delta_h^\star$ built from the realized signal $R_h^\star$, while (9) acquires the terms $\mathrm{Cov}(R_h^{\star j}, R_h^{\star k} \mid s_h, a_h) + \gamma \mathrm{Cov}(R_h^{\star j}, V_{h+1}^{\pi,\star k} \mid s_h, a_h) + \gamma \mathrm{Cov}(V_{h+1}^{\pi,\star j}, R_h^{\star k} \mid s_h, a_h)$, and the $h = H$ term no longer vanishes. □

M Proof of Theorem 6: Asymptotic Normality

Theorem (Theorem 6, restated)**.** *Under Assumptions 1–4 and the regularity conditions of Appendix R,* $\sqrt{n}(\hat{\theta} - \theta^*) \xrightarrow{d} \mathcal{N}(0, \Sigma^*)$.

Proof. **Step 1 (Decomposition).** $\sqrt{n}(\hat{\theta} - \theta^*) = \frac{1}{\sqrt{n}} \sum_i \psi^*(Z_i) + \frac{1}{\sqrt{n}} \sum_i [\psi(Z_i; \theta^*, \hat{\eta}^{(k(i))}) - \psi(Z_i; \theta^*, \eta^*)]$.

Step 2 (Second-order bias). Let $R_n := \frac{1}{\sqrt{n}} \sum_i [\psi(Z_i; \theta^*, \hat{\eta}^{(k(i))}) - \psi(Z_i; \theta^*, \eta^*)]$. Expanding:

$$\begin{aligned}
&\psi^{\star_j}(Z_i; \theta^*, \hat{\eta}) - \psi^{\star_j}(Z_i; \theta^*, \eta^*) \\
&= \underbrace{\big[\hat{V}_1^{\pi,\star_j}(s_{1,i}) - V_1^{\pi,\star_j}(s_{1,i})\big]}_{\text{boundary, first-order in } \hat{Q}-Q} \\
&\quad + \underbrace{\sum_{h=1}^{H} w_h(\hat{\delta}_h^{\star j} - \delta_h^{\star j})}_{\text{first-order in } \hat{Q}-Q} \\
&\quad + \underbrace{\sum_{h=1}^{H} (\hat{w}_h - w_h)\delta_h^{\star j}}_{\text{first-order in } \hat{w}-w} \\
&\quad + \underbrace{\sum_{h=1}^{H} (\hat{w}_h - w_h)(\hat{\delta}_h^{\star j} - \delta_h^{\star j})}_{\text{second-order}}.
\end{aligned}$$

First-order in $\hat{Q} - Q$: Writing $\epsilon_h^V := \hat{V}_h - V_h$ and $\epsilon_h^Q := \hat{Q}_h - Q_h$, the conditional expectation (with $\hat{\eta}$ fixed via cross-fitting) telescopes:

$$\begin{aligned}
&\mathbb{E}\big[\epsilon_1^V(s_1) + \textstyle\sum_{h=1}^H w_h^*\big(\epsilon_{h+1}^V(s_{h+1}) - \epsilon_h^Q(s_h, a_h)\big) \,\big|\, \hat{\eta}\big] \\
&= \mathbb{E}_\mu[\epsilon_1^V] + \sum_{h=1}^{H} \big(\mathbb{E}_{d_{h+1}^\pi}[\epsilon_{h+1}^V] - \mathbb{E}_{d_h^\pi}[\epsilon_h^V]\big) \\
&= \mathbb{E}_\mu[\epsilon_1^V] - \mathbb{E}_{d_1^\pi}[\epsilon_1^V] = 0,
\end{aligned}$$

using the density-ratio change-of-measure, Q-to-V averaging ($\mathbb{E}_{d_h^\pi}[\epsilon_h^Q] = \mathbb{E}_{d_h^\pi}[\epsilon_h^V]$), telescoping, and boundary conditions ($\epsilon_{H+1}^V = 0$, $d_1^\pi = \mu$).

First-order in $\hat{w} - w$: $\mathbb{E}[(\hat{w}_h - w_h)\delta_h^{\star_j} \mid \hat{\eta}] = 0$ since $\mathbb{E}[\delta_h^{\star_j} | s_h, a_h] = 0$; cross-fitting gives unconditional mean zero.

Second-order remainder: Let $\zeta_{Q,h}$ denote the $L^2(d_h^{\pi_0})$ error of $\hat{Q}_h^{\star_j} - Q_h^{\star_j}$ and $\zeta_{\star_j} := \max_h \zeta_{Q,h}$ the worst-step error of the j-th value head (as in Assumption 3). By Cauchy–Schwarz step by step, with $\hat{\delta}_h^{\star_j} - \delta_h^{\star_j} = \gamma\epsilon_{h+1}^V(s_{h+1}) - \epsilon_h^Q(s_h, a_h)$ and, by Jensen and $\pi/\pi_0 \le \eta_{\min}^{-1}$ (Assumption 1), $\|\epsilon_{h+1}^V\|_{L^2(d_{h+1}^{\pi_0})} \le \eta_{\min}^{-1/2}\zeta_{Q,h+1}$ (with $\zeta_{Q,H+1} := 0$):

$$\mathbb{E}\Big[\Big|\sum_h (\hat{w}_h - w_h)(\hat{\delta}_h^{\star_j} - \delta_h^{\star_j})\Big| \,\Big|\, \hat{\eta}^{(k)}\Big] \le (1 + \eta_{\min}^{-1/2})\, H\, \zeta_w\, \zeta_{\star_j} = o_p(n^{-1/2}).$$

Since the two first-order terms have zero conditional mean, this gives $\sqrt{n}\,\mathbb{E}[\psi(Z; \theta^*, \hat{\eta}^{(k)}) - \psi^* \mid \hat{\eta}^{(k)}] = o_p(1)$ for each k.[2]

Step 2c (Empirical process term). Define $\nu_n^{(k)} := \mathbb{G}_n^{(k)}[\psi(\cdot; \theta^*, \hat{\eta}^{(k)}) - \psi(\cdot; \theta^*, \eta^*)]$, where $\mathbb{G}_n^{(k)} = \sqrt{|I_k|}(\mathbb{P}_n^{(k)} - P)$ acts on fold k. Conditional on $\hat{\eta}^{(k)}$ (estimated on $\mathcal{D} \setminus \mathcal{D}_{I_k}$, hence independent of $\mathcal{D}_{I_k}$), $\nu_n^{(k)}$ has mean zero and conditional variance bounded by $\|\psi(\cdot; \theta^*, \hat{\eta}^{(k)}) - \psi^*\|_{L^2(P)}^2$. By Assumption 4 and the L^2-consistency of $\hat{\eta}$ (part (i) of Assumption 3) together with continuity of $\eta \mapsto \psi(\cdot; \theta^*, \eta)$ in L^2 (Assumption 7), this L^2-norm $\to 0$ in probability. Chebyshev gives $\nu_n^{(k)} = o_p(1)$. Aggregating over K folds yields that the empirical-process contribution is $o_p(1)$. Combining with Step 2, $R_n = o_p(1)$.

Step 3 (CLT). $\mathbb{E}[\psi^*] = 0$ and $\mathbb{E}[\|\psi^*\|^2] < \infty$ imply $\frac{1}{\sqrt{n}}\sum_i \psi^*(Z_i) \xrightarrow{d} \mathcal{N}(0, \Sigma^*)$.

Step 4 (Slutsky). Since $R_n = o_p(1)$, $\sqrt{n}(\hat{\theta} - \theta^*) \xrightarrow{d} \mathcal{N}(0, \Sigma^*)$. Efficiency: Steps 1–2c show $\sqrt{n}(\hat{\theta} - \theta^*) = n^{-1/2}\sum_i \psi^*(Z_i) + o_p(1)$, i.e. $\hat{\theta}$ is asymptotically linear with influence function equal to the EIF of Theorem 3. By van der Vaart (1998, Lemma 25.23), $\hat{\theta}$ is regular and attains the semiparametric efficiency bound Σ^* in the nonparametric, time-inhomogeneous CMDP model. □

N Proof of Proposition 4: Non-vanishing Off-diagonal Covariance

Proposition (Proposition 4, restated)**.** *By* (8)–(9), $\Sigma_{jk}^* = \mathrm{Cov}_\mu(V_1^{\pi,\star_j}, V_1^{\pi,\star_k}) + \gamma^2 \sum_{h=1}^{H-1} \mathbb{E}_{d_h^\pi}\big[w_h\, C_h(s_h, a_h)\big]$ *with* $C_h(s,a) := \mathrm{Cov}_{P_h(\cdot|s,a)}(V_{h+1}^{\pi,\star_j}, V_{h+1}^{\pi,\star_k})$*. (a) If the initial-state covariance and every C_h (d_h^π-a.e.) share a common sign, and at least one of them is non-zero on a set of positive μ- resp. d_h^π-mass, then $\Sigma_{jk}^* \neq 0$ with that sign. (b) In the tabular case ($|\mathcal{S}|, |\mathcal{A}| < \infty$, $|\mathcal{S}| \ge 2$, $H \ge 2$, with μ, π, π_0 fixed), $\Sigma_{jk}^* \neq 0$ for all full-support kernels and signal functions outside a Lebesgue-null set. Without (a), terms of opposite sign—across steps, or across (s,a) within one step—can cancel exactly.*

Remark 22 (Sense of "generic")**.** *The condition above is generic in the following precise sense. Work on the connected open domain of full-support kernels (the interior of the product of transition simplices, where $d_h^{\pi_0}(s) > 0$ for all s and $h \ge 2$, so each w_h is finite on the support of d_h^π). At the true Q, Σ_{jk}^* is there a fixed real-analytic function of the transition kernels $\{P_h\}$ and signal functions $\{f_h^\star\}$—a finite sum of terms rational in the entries of P_h (through w_h and the occupancies) and polynomial in the induced value functions. This function is not identically zero: take $f_H^{\star_k} = f_H^{\star_j}$ with values in $(0,1)$ and $V_H^{\pi,\star_j}$ non-constant, set $f_{H-1}^{\star_k}(s,a) = 1 - \gamma(P_{H-1}V_H^{\pi,\star_k})(s,a)$ and $f_h^{\star_k} \equiv \frac{1}{2}$ for $h \le H-2$. Then $V_h^{\pi,\star_k}$ is constant for every $h \le H-1$, so the initial-state term and all steps $h \le H-2$ vanish, and $\Sigma_{jk}^* = \gamma^2 \mathbb{E}_{d_{H-1}^\pi}[w_{H-1}\mathrm{Var}_{P_{H-1}}(V_H^{\pi,\star_j})] > 0$ for full-support P_{H-1}, an interior point of the domain. A real-analytic function on a connected domain that is not*

[2] Throughout, H and J are fixed as $n \to \infty$; a growing-horizon version would need a self-normalized triangular-array CLT and horizon-dependent versions of Assumptions 3–4, which we do not pursue.

identically zero vanishes only on a set of Lebesgue measure zero; hence the exact-cancellation set $\{(\{P_h\},\{f_h^\star\}) : \Sigma^*_{jk} = 0\}$ *has measure zero under any reference measure absolutely continuous with respect to Lebesgue. Consequently,* $\Sigma^*_{jk} \neq 0$ *for Lebesgue-almost every* $(\{P_h\},\{f_h^\star\})$*, which is part (b) of the proposition. (The statement is void when* $|\mathcal{S}| = 1$ *or* $H = 1$*: with a point-mass* μ *and a single step,* $\Sigma^*_{jk} \equiv 0$*.)*

Proof. At true Q, by (8) and (9):

$$\Sigma^*_{jk} = \mathrm{Cov}_\mu(V_1^{\pi,\star_j}, V_1^{\pi,\star_k}) + \gamma^2 \sum_{h=1}^{H} \mathbb{E}_{d_h^{\pi_0}}\big[w_h^2 \,\mathrm{Cov}_{P_h}\big(V_{h+1}^{\pi,\star_j}(s'), V_{h+1}^{\pi,\star_k}(s') \mid s_h, a_h\big)\big].$$

The last step $h = H$ contributes nothing since $V_{H+1}^{\pi,\star} \equiv 0$. By Assumption 1, $w_h \in (0,\infty)$ d_h^π-a.e., and $\mathbb{E}_{d_h^{\pi_0}}[w_h^2 C_h] = \mathbb{E}_{d_h^\pi}[w_h C_h]$, which gives the displayed form in the proposition. (a) Every term of the displayed sum has the common sign and at least one is non-zero, so the sum is non-zero with that sign. (b) See Remark 22. Unlike the diagonal case ($j = k$), where each summand is a non-negative variance, the off-diagonal terms are covariances of either sign, and without (a) they can cancel exactly, both across steps and across (s,a) within a single step. □

O Proof of Theorem 5: Calibration of Lagrangian Inference

Theorem (Theorem 5, restated)**.** *Consider the single-constraint case (*$J = 1$*) with Lagrangian multiplier* $\lambda > 0$ *and* $e_\lambda = (1, -\lambda)^\top$*. The true and diagonally reported Lagrangian variances are as stated in parts (a)–(b), the variance ratio satisfies* (10)*, and the variance-minimizing multiplier is* $\lambda^\dagger = \Sigma^*_{01}/\Sigma^*_{11}$ *(when* $\Sigma^*_{01} > 0$*, over* $\lambda \geq 0$*).*

Proof. **Parts (a)–(b).** Direct expansion: $\sigma^2_\lambda = e_\lambda^\top \Sigma^* e_\lambda = \Sigma^*_{00} - 2\lambda\Sigma^*_{01} + \lambda^2\Sigma^*_{11}$ and $\sigma^2_{\lambda,\mathrm{indep}} = e_\lambda^\top \mathrm{diag}(\Sigma^*_{00}, \Sigma^*_{11}) e_\lambda = \Sigma^*_{00} + \lambda^2\Sigma^*_{11}$. Their difference is $\sigma^2_{\lambda,\mathrm{indep}} - \sigma^2_\lambda = 2\lambda\Sigma^*_{01}$, which (since $\lambda > 0$) is positive iff $\Sigma^*_{01} > 0$.

Part (c). Dividing:

$$\frac{\sigma^2_\lambda}{\sigma^2_{\lambda,\mathrm{indep}}} = 1 - \frac{2\lambda\Sigma^*_{01}}{\Sigma^*_{00} + \lambda^2\Sigma^*_{11}} < 1 \quad \text{when } \Sigma^*_{01} > 0,\ \lambda > 0.$$

When $\Sigma^*_{01} < 0$ and $\lambda > 0$, this ratio exceeds 1, meaning the diagonal report underestimates the true Lagrangian variance, leading to anti-conservative confidence regions.

Part (d). Setting $\partial_\lambda \sigma^2_\lambda = -2\Sigma^*_{01} + 2\lambda\Sigma^*_{11} = 0$ gives $\lambda^\dagger = \Sigma^*_{01}/\Sigma^*_{11}$ (assuming $\Sigma^*_{01} > 0$ so $\lambda^\dagger > 0$). Substituting:

$$\sigma^2_{\lambda^\dagger} = \Sigma^*_{00} - 2\frac{(\Sigma^*_{01})^2}{\Sigma^*_{11}} + \frac{(\Sigma^*_{01})^2}{\Sigma^*_{11}} = \Sigma^*_{00} - \frac{(\Sigma^*_{01})^2}{\Sigma^*_{11}} = \Sigma^*_{00}(1 - \rho^2),$$

where $\rho^2 = (\Sigma^*_{01})^2/(\Sigma^*_{00}\Sigma^*_{11})$. The formula at the dual-optimal λ^* follows directly from part (c) evaluated at $\lambda = \lambda^*$; the lower bound $1 - \rho$ follows from $\Sigma^*_{00} + \lambda^2\Sigma^*_{11} \geq 2\lambda\sqrt{\Sigma^*_{00}\Sigma^*_{11}}$. □

P Proof of Theorem 11: Variance Term of the Regret Bound

Theorem (Theorem 11, restated)**.** *Let* $J = 1$ *and fix* $\lambda \geq 0$*; let* $\hat{\pi}_\lambda$ *and* π^*_λ *maximize* $\hat{L}_\lambda$ *and* L_λ *over* Π*. Under Assumptions 1, 2, 3, 4, 6 and 7, for all sufficiently large* n *and with probability at least* $1 - \delta - o(1)$*, the CDR-based Lagrangian optimizer achieves the regret bound* (15)*; when* $\Sigma^*_{01}(\pi) > 0$ *for all* $\pi \in \Pi$ *the diagonal report does not under-state its variance (Bernstein) term,* $\bar{\sigma}^2_\lambda \leq \bar{\sigma}^2_{\lambda,\mathrm{indep}}$*, strictly if* $\inf_\pi \Sigma^*_{01}(\pi) > 0$ *or the supremum is attained. The learned policy and the Rademacher term* $C_1 L\kappa$ *are the same under either report.*

Proof. The Lagrangian $L_\lambda(\pi) = e_\lambda^\top \theta(\pi) + \lambda b$ has EIF $e_\lambda^\top \psi^*(Z)$ and variance $\sigma_\lambda^2(\pi) = e_\lambda^\top \Sigma^*(\pi) e_\lambda$.

Rademacher complexity. For each fold k, the projected score class $\mathcal{F}_\lambda^{(k)} = \{Z \mapsto e_\lambda^\top \psi(Z; \pi, \hat{\eta}^{(k)}) : \pi \in \Pi\}$ has $\mathcal{R}_m(\mathcal{F}_\lambda^{(k)}) \le (1+\lambda) L\kappa/\sqrt{m}$ by Assumption 6.

Bernstein concentration. Under Assumption 4, $|e_\lambda^\top \hat{\psi}(Z;\pi)| \le (1+\lambda)(M + HMB) := U$ uniformly in π. Conditional on $\hat{\eta}^{(k)}$, the $n_k = n/K$ held-out scores of fold k are i.i.d., so Bousquet's version of Bernstein's inequality for suprema of empirical processes (Bousquet, 2002) gives, with probability $\ge 1 - \delta/K$ for each fold, and hence after averaging the K fold bounds and a union bound over k (which is where the factors $\sqrt{K}$ and $\log(2K/\delta)$ below come from), with probability $\ge 1 - \delta$,

$$\sup_\pi \big|\hat{L}_\lambda(\pi) - \bar{L}_\lambda(\pi)\big| \le \frac{4\sqrt{K}(1+\lambda)L\kappa}{\sqrt{n}} + \sqrt{2K\hat{\sigma}_{\lambda,n}^2 \log(2K/\delta)/n} + C_3\, K\, U \log(2K/\delta)/n,$$

where $\bar{L}_\lambda(\pi)$ is the fold-averaged population mean of the score at the fitted nuisances (as $\bar{\theta}^\pi$ in Appendix Q), $\hat{\sigma}_{\lambda,n}^2 := \max_k \sup_\pi \mathrm{Var}(e_\lambda^\top \psi(Z;\pi,\hat{\eta}^{(k)}) \mid \hat{\eta}^{(k)})$ and C_3 is an absolute constant (the Rademacher term is $4\mathcal{R}$ rather than $2\mathcal{R}$ because Bousquet's inequality is applied to the two-sided supremum, i.e. to $\mathcal{F} \cup -\mathcal{F}$). Under Assumptions 3 and 7, $\hat{\sigma}_{\lambda,n}^2 \to \bar{\sigma}_\lambda^2$ in probability, and the nuisance bias $\sup_\pi |\bar{L}_\lambda(\pi) - L_\lambda(\pi)|$ is $o_p(n^{-1/2})$ (Appendix Q); on the event of probability $1 - o(1)$ where both are within the stated tolerances, the regret argument of Appendix Q (optimality of $\hat{\pi}_\lambda$ for $\hat{L}_\lambda$, so that the regret is at most twice the uniform deviation) gives the stated bound with $C_1 = 8\sqrt{K}(1+\lambda)$, any $C_2 > 2\sqrt{2K}$ for $n \ge n_0$, and C_3 replaced by $2C_3K(1+\lambda)(M + HMB)$.

Comparison of the variance term with its diagonal report. When $\Sigma_{01}^*(\pi) > 0$ for all $\pi \in \Pi$ and $\lambda > 0$, Theorem 5(c) gives $e_\lambda^\top \Sigma^*(\pi) e_\lambda = e_\lambda^\top \mathrm{diag}(\Sigma^*(\pi)) e_\lambda - 2\lambda \Sigma_{01}^*(\pi) < e_\lambda^\top \mathrm{diag}(\Sigma^*(\pi)) e_\lambda$ pointwise in π, so $\bar{\sigma}_\lambda^2 \le \bar{\sigma}_{\lambda,\mathrm{indep}}^2$. A pointwise strict inequality does not by itself make the suprema strictly ordered (take $\Sigma_{00}^* = \Sigma_{11}^* = 1$, $\Sigma_{01}^*(\pi_k) = 1/k$, $\lambda = 1$: both suprema equal 2); the inequality is strict if $\inf_\pi \Sigma_{01}^*(\pi) > 0$, since then $\bar{\sigma}_\lambda^2 \le \bar{\sigma}_{\lambda,\mathrm{indep}}^2 - 2\lambda \inf_\pi \Sigma_{01}^*(\pi)$, or if the supremum defining $\bar{\sigma}_\lambda^2$ is attained at some π_0, since then $\bar{\sigma}_\lambda^2 = \sigma_\lambda^2(\pi_0) < \sigma_{\lambda,\mathrm{indep}}^2(\pi_0) \le \bar{\sigma}_{\lambda,\mathrm{indep}}^2$. The per-policy ratio $(1-\rho(\pi)^2)/(1+\rho(\pi)^2)$ at $\lambda^\dagger(\pi)$ (Theorem 5(d)) is not a consequence of this bound: $\lambda^\dagger(\pi)$ varies with π and the ratio of the two suprema at a fixed λ can be arbitrarily close to 1. Since the point estimates, and hence $\hat{\pi}_\lambda$ and its true regret, do not depend on which variance is reported, the only difference between the two analyses is that the diagonal one would substitute the larger $\bar{\sigma}_{\lambda,\mathrm{indep}}^2$ for $\bar{\sigma}_\lambda^2$ in (15); the Rademacher term $C_1 L\kappa$ is unaffected. □

Q Proof of Theorem 8: Regret Bound

Theorem (Theorem 8, restated)**.** *Under Assumptions 1–6 with safety margins* (13) *and n large enough that $2\epsilon_{n,j} \le \alpha_S$ for all j, with probability at least $1 - \delta - \mathbb{P}(\mathcal{E}_n^c)$: $v^{\pi^*} - v^{\hat{\pi}} \le (4\sqrt{K}L\kappa + (2\sqrt{2K}B_0 + 2)\sqrt{\log(4K/\delta)})/\sqrt{n}$ and $c_j^{\hat{\pi}} \le b_j$ for all j, where L is the constant of Assumption 6, K the number of folds, and $\mathcal{E}_n$ the nuisance-bias event defined below.*

Proof. **Boundedness of CDR scores.** Under Assumption 4, each Bellman residual satisfies $|\hat{\delta}_h^{\star_j}| \le 1 + 2M := B$, so $|\hat{\psi}^{\star_j}(Z_i;\pi)| \le M + HMB := B_j$ uniformly over all $\pi \in \Pi$ (the same constant for every j; the index only records which score is meant).

Nuisance bias. For each fold k, $\hat{\eta}^{(k)}$ depends only on $\mathcal{D} \setminus \mathcal{D}_{I_k}$. Let $\bar{\theta}^\pi := \sum_k \frac{|I_k|}{n} \mathbb{E}_{Z\sim P}\big[\psi^\circ(Z;\pi,\hat{\eta}^{(k)})\big]$ (Z a fresh draw, $\hat{\eta}^{(k)}$ fixed) and $\beta_n := \sup_{\pi\in\Pi} \|\bar{\theta}^\pi - \theta^\pi\|_\infty$. By the second-order bound of Appendix M, applied fold by fold and under Assumption 3 uniformly over Π, $\beta_n = o_p(n^{-1/2})$, so $\mathcal{E}_n := \{\beta_n \le n^{-1/2}\}$ has $\mathbb{P}(\mathcal{E}_n^c) \to 0$. We work on $\mathcal{E}_n$ and center each fold's empirical process at its own population mean, paying $n^{-1/2}$ for the recentering. (Conditioning on

all K fitted nuisances at once would destroy the independence of the held-out samples, since every observation enters the training set of the other folds; the argument is therefore made fold by fold.)

Uniform concentration. Define $\mathcal{F}_j^{(k)} := \mathcal{F}_{e_j}^{(k)}$, where e_j is the j-th standard basis vector. Conditional on $\hat{\eta}^{(k)}$, the $n_k = |I_k|$ held-out scores of fold k are i.i.d. and $[-B_j, B_j]$-valued, so Wainwright (2019, Theorem 4.10) (symmetrization followed by McDiarmid's inequality) together with Assumption 6 gives, with probability at least $1 - \delta/(2JK)$, $\sup_\pi |\mathbb{P}_{n_k}^{(k)} \psi_j^\circ - P\psi_j^\circ| \le 2L\kappa/\sqrt{n_k} + B_j\sqrt{2\log(2JK/\delta)/n_k}$. Averaging over folds ($n_k = n/K$; use $\min_k |I_k|$ otherwise), adding $\beta_n \le n^{-1/2}$, using $\log(4JK/\delta) \ge \log 4 > 1$ and $\log(4JK/\delta) \ge \log(2JK/\delta)$, and taking a union bound over (j,k),

$$\sup_{\pi\in\Pi} |\hat{c}_j^\pi - c_j^\pi| \le \frac{2\sqrt{K}\,L\kappa}{\sqrt{n}} + (\sqrt{2K}\,B_j + 1)\sqrt{\frac{\log(4JK/\delta)}{n}} = \epsilon_{n,j},$$

which fixes the constant $C = \sqrt{2K}B_j + 1$ in (13). The union bound over (j,k) is already included, so all J constraint events hold simultaneously with probability $\ge 1 - \delta/2$.

Constraint satisfaction of $\hat{\pi}$. On the constraint events, $c_j^{\hat{\pi}} \le \hat{c}_j^{\hat{\pi}} + \epsilon_{n,j} \le (b_j - \epsilon_{n,j}) + \epsilon_{n,j} = b_j$ for every j.

Feasibility of π^*. By Assumption 5, $c_j^{\pi^*} \le b_j - \alpha_S$, so on the constraint events $\hat{c}_j^{\pi^*} \le c_j^{\pi^*} + \epsilon_{n,j} \le b_j - \alpha_S + \epsilon_{n,j} \le b_j - \epsilon_{n,j}$ as soon as $2\epsilon_{n,j} \le \alpha_S$; that is, π^* is feasible for the margin-tightened problem (12). (It is the tightened constraint $\hat{c}_j^{\pi^*} \le b_j - \epsilon_{n,j}$, not $\hat{c}_j^{\pi^*} \le b_j$, that must be checked, which is why the threshold is $\alpha_S/2$ rather than α_S.)

Regret. The same fold-by-fold argument applied to the reward score classes $\mathcal{F}_0^{(k)} := \mathcal{F}_{e_0}^{(k)}$ gives, with probability $\ge 1 - \delta/2$, $\sup_{\pi\in\Pi} |\hat{v}^\pi - v^\pi| \le 2\sqrt{K}L\kappa/\sqrt{n} + (\sqrt{2K}B_0 + 1)\sqrt{\log(4K/\delta)/n}$. By optimality of $\hat{\pi}$ and feasibility of π^*, $\hat{v}^{\hat{\pi}} \ge \hat{v}^{\pi^*}$, hence $v^{\pi^*} - v^{\hat{\pi}} \le 2\sup_{\pi\in\Pi} |\hat{v}^\pi - v^\pi| \le (4\sqrt{K}L\kappa + (2\sqrt{2K}B_0 + 2)\sqrt{\log(4K/\delta)})/\sqrt{n}$, i.e. $C_1 = 4\sqrt{K}$ and $C_2 = 2\sqrt{2K}B_0 + 2$. Combining the reward event ($\delta/2$), the constraint union event ($\delta/2$) and $\mathcal{E}_n$ gives overall probability at least $1 - \delta - \mathbb{P}(\mathcal{E}_n^c)$. □

R Regularity Conditions

Remark 23 (No Donsker condition needed)**.** *Following the cross-fitting framework of Chernozhukov et al. (2018), no Donsker condition is required for our asymptotic analysis: the proof of Theorem 6 uses only consistency and the product rate (Assumption 3), overlap and bounded signals (Assumptions 1–2), bounded nuisances (Assumption 4) and Assumption 7. Theorems 8 and 11 additionally use Assumption 6 and Assumption 3 uniformly over* Π*. Cross-fitting decouples the empirical process from the nuisance estimator, so pointwise* L^2 *convergence rates suffice in place of a Donsker envelope (Chernozhukov et al., 2018).*

Assumption 7 (Continuity)**.** $\|\psi(\cdot;\theta^*,\eta) - \psi(\cdot;\theta^*,\eta^*)\|_{L^2(P)} \to 0$ *as* $\max_h \|w_h - w_h^*\|_{L^2(d_h^{\pi_0})} + \max_{j,h} \|Q_h^{\star_j} - Q_h^{\pi,\star_j}\|_{L^2(d_h^{\pi_0})} \to 0$ *within the* M*-bounded class. (This is implied by Assumptions 1, 2 and 4: write* $\hat{w}\hat{\delta} - w\delta = (\hat{w} - w)\hat{\delta} + w(\hat{\delta} - \delta)$ *and use* $\|\hat{V}_h - V_h\|_{L^2(d_h^{\pi_0})} \le \eta_{\min}^{-1/2}\|\hat{Q}_h - Q_h\|_{L^2(d_h^{\pi_0})}$*.)*

Justification of $\hat{\Sigma} \to \Sigma^*$. Write $\psi_i(\theta^*) := \psi(Z_i;\theta^*,\hat{\eta}^{(k(i))})$. (1) Since ψ is bounded and $\|\psi(\cdot;\theta^*,\hat{\eta}^{(k)}) - \psi^*\|_{L^2} = o_p(1)$ (Step 2c), $\|\mathbb{E}[\psi_i(\theta^*)\psi_i(\theta^*)^\top \mid \hat{\eta}^{(k)}] - \Sigma^*\| = o_p(1)$. (2) Conditional on $\hat{\eta}^{(k)}$ the fold-k scores are i.i.d. and bounded, so conditional Chebyshev gives $\frac{1}{|I_k|}\sum_{i\in I_k} \psi_i(\theta^*)\psi_i(\theta^*)^\top - \mathbb{E}[\cdot \mid \hat{\eta}^{(k)}] = o_p(1)$. (3) ψ is linear in θ and $\frac{1}{n}\sum_i \hat{\psi}_i = 0$ (Remark 3), so $\hat{\Sigma} = \frac{1}{n}\sum_i \psi_i(\theta^*)\psi_i(\theta^*)^\top - (\hat{\theta} - \theta^*)(\hat{\theta} - \theta^*)^\top$, whose last term is $O_p(n^{-1})$. Hence $\hat{\Sigma} \xrightarrow{p} \Sigma^*$.

S Proof of Sufficient Conditions for Assumption 6

Remark 24 (Standing assumption for this proof)**.** *The estimated model $(\hat{\mu}, \hat{P})$ (probability kernels) is held fixed across Π, and for each π both $\hat{Q}^\pi$ (backward induction on $\hat{P}$) and $\hat{w}^\pi$ are computed from it (Remark 16). A fixed, π-independent $\hat{Q}$ is not used: it would make the nuisance bias first order in $\hat{w} - w$ and break Assumption 3 uniformly over Π. The Lipschitz dependence of $\hat{Q}^\pi$ on π established below relies on backward induction through a sup-norm non-expansive (tabular or model-based) operator; it is not claimed for neural FQE.*

Proof. We prove condition (iii) of Remark 15; condition (i) is standard (Massart's lemma applied to $\mathcal{F}_v \cup -\mathcal{F}_v$ gives $\kappa = B_0\sqrt{2\log(2|\Pi|)}$ with $L = 1$).

Fix $v \in \mathbb{R}^{J+1}$ and write $\epsilon := \|\pi - \pi'\|_\infty$, where $\|\pi - \pi'\|_\infty := \sup_{s,h} \|\pi_h(\cdot|s) - \pi'_h(\cdot|s)\|_1$. Assume $\hat{\mu}$ and each $\hat{P}_h(\cdot|s,a)$ are probability vectors. Since $\hat{d}_1^\pi = \hat{d}_1^{\pi'} = \hat{\mu}$ and $\|\hat{d}_{h+1}^\pi - \hat{d}_{h+1}^{\pi'}\|_1 \le \sum_{s,a} |\hat{d}_h^\pi(s)\pi_h(a|s) - \hat{d}_h^{\pi'}(s)\pi'_h(a|s)| \le \|\hat{d}_h^\pi - \hat{d}_h^{\pi'}\|_1 + \|\pi - \pi'\|_\infty$, induction gives $\|\hat{d}_h^\pi - \hat{d}_h^{\pi'}\|_1 \le (h-1)\epsilon$. The projected CDR score $g_\pi(Z) = v^\top \hat{\psi}(Z;\pi)$ depends on π through three channels: the boundary term $\hat{V}_1^\pi(s_1)$, the value-function terms $\hat{V}_{h+1}^\pi(s_{h+1})$ and $\hat{Q}_h^\pi(s_h, a_h)$ inside each Bellman residual, and the density ratio $\hat{w}_h^\pi$.

Lipschitz bound. *Q channel.* Since $\hat{Q}_h^\pi - \hat{Q}_h^{\pi'} = \gamma \hat{P}_h(\hat{V}_{h+1}^\pi - \hat{V}_{h+1}^{\pi'})$ and $\hat{V}_h^\pi - \hat{V}_h^{\pi'} = (\pi_h - \pi'_h)\hat{Q}_h^\pi + \pi'_h(\hat{Q}_h^\pi - \hat{Q}_h^{\pi'})$, backward induction from $\hat{V}_{H+1} \equiv 0$ gives $\|\hat{Q}_h^\pi - \hat{Q}_h^{\pi'}\|_\infty \le \gamma(H-h)M\epsilon$ and $\|\hat{V}_h^\pi - \hat{V}_h^{\pi'}\|_\infty \le (H-h+1)M\epsilon$ (for each component $\star_j$; $\hat{P}_h$ is sup-norm non-expansive).

Boundary term: $|\hat{V}_1^{\pi,\star_j}(s_1) - \hat{V}_1^{\pi',\star_j}(s_1)| \le HM\epsilon$.

Bellman residual term: The residual $\hat{\delta}_h^{\pi,\star_j} = f_h^{\star_j} + \gamma \hat{V}_{h+1}^{\pi,\star_j}(s_{h+1}) - \hat{Q}_h^{\pi,\star_j}(s_h, a_h)$ satisfies $|\hat{\delta}_h^{\pi,\star_j} - \hat{\delta}_h^{\pi',\star_j}| \le \gamma\|\hat{V}_{h+1}^\pi - \hat{V}_{h+1}^{\pi'}\|_\infty + \|\hat{Q}_h^\pi - \hat{Q}_h^{\pi'}\|_\infty \le 2\gamma(H-h)M\epsilon$.

Density ratio term: Following Remark 16, $\hat{w}_h^\pi(s,a) = \hat{d}_h^\pi(s,a)/\hat{d}_h^{\pi_0}(s,a)$ is recomputed from the fixed estimated model. By the occupancy bound above, $\hat{d}_h^{\pi'}(s) \le 1$, and $\hat{d}_h^\pi(s,a) = \hat{d}_h^\pi(s)\,\pi_h(a|s)$, $|\hat{d}_h^\pi(s,a) - \hat{d}_h^{\pi'}(s,a)| \le |\hat{d}_h^\pi(s) - \hat{d}_h^{\pi'}(s)| + \hat{d}_h^{\pi'}(s)\,|\pi_h(a|s) - \pi'_h(a|s)| \le h\,\epsilon$. Dividing by $\hat{d}_h^{\pi_0}(s,a) \ge \nu$—the lower bound assumed in condition (iii); note that the bound $\|\hat{w}\|_\infty \le M$ of Assumption 4 does *not* imply it—gives $|\hat{w}_h^\pi(s,a) - \hat{w}_h^{\pi'}(s,a)| \le (h/\nu)\,\epsilon$. Since $\sum_{s,a} \hat{d}_h^{\pi_0}(s,a) = 1$, necessarily $\nu \le 1/|\mathrm{supp}\,\hat{d}_h^{\pi_0}|$, so L grows with the support size.

Combining all three channels: decomposing $\hat{w}_h^\pi\hat{\delta}_h^\pi - \hat{w}_h^{\pi'}\hat{\delta}_h^{\pi'} = (\hat{w}_h^\pi - \hat{w}_h^{\pi'})\hat{\delta}_h^\pi + \hat{w}_h^{\pi'}(\hat{\delta}_h^\pi - \hat{\delta}_h^{\pi'})$, and using $|\hat{\delta}_h^\pi| \le B := 1 + 2M$, $|\hat{w}_h^{\pi'}| \le M$:

$$
\begin{aligned}
|g_\pi(Z) - g_{\pi'}(Z)| &\le \|v\|_1 \Big[HM + \sum_{h=1}^{H} \big(hB/\nu + 2\gamma(H-h)M^2 \big) \Big]\, \epsilon \\
&= \|v\|_1 \big(HM + H(H+1)B/(2\nu) + \gamma H(H-1)M^2 \big)\, \epsilon \;:=\; L\,\|v\|_1\,\|\pi - \pi'\|_\infty .
\end{aligned}
$$

From Lipschitz continuity to Rademacher complexity. A Lipschitz dependence on π in the sup norm over the *whole state space* is not a coordinatewise contraction of a finite vector of sample evaluations, so Talagrand's contraction principle does not apply directly; the entropy route does. Since g_π is $L\|v\|_1$-Lipschitz in $\|\pi - \pi'\|_\infty$ uniformly in Z (the argument is conditional on the fold-k model, so it applies to each class $\mathcal{F}_v^{(k)}$ and every sample size m), every $\epsilon/(L\|v\|_1)$-cover of Π in $\|\cdot\|_\infty$ induces an ϵ-cover of $\mathcal{F}_v^{(k)}$ in the sup norm (hence in every empirical L^2 norm), so

$$
\log N(\mathcal{F}_v^{(k)}, \epsilon, \|\cdot\|_\infty) \le \log N\big(\Pi, \epsilon/(L\|v\|_1), \|\cdot\|_\infty\big) \le D \log\big(C_\Pi L \|v\|_1/\epsilon\big).
$$

Dudley's entropy integral (Wainwright, 2019, Theorem 5.22) then gives, for an absolute constant c_0,

$$
\mathcal{R}_m(\mathcal{F}_v^{(k)}) \le \frac{c_0}{\sqrt{m}} \int_0^{\|v\|_1 B_0} \sqrt{D \log\big(C_\Pi L \|v\|_1/\epsilon\big)}\, d\epsilon \le \frac{c_0\sqrt{\pi D}}{2} \cdot \frac{C_\Pi L \|v\|_1}{\sqrt{m}},
$$

using the substitution $\epsilon = C_\Pi L\|v\|_1 t$ and $\int_0^u \sqrt{\log(1/t)}\,dt \le \sqrt{\pi}/2$ for $u \le 1$ (enlarge C_Π if necessary so that $C_\Pi L \ge B_0$). This is Assumption 6 with $\kappa = c\,C_\Pi\sqrt{D}$, $c = c_0\sqrt{\pi}/2$, and the Lipschitz constant $L = HM + H(H+1)(1+2M)/(2\nu) + \gamma H(H-1)M^2$. □

T SEMIPARAMETRIC LOWER BOUND AND LOCAL ASYMPTOTIC NORMALITY

We establish that Σ^* lower-bounds the asymptotic covariance of every regular estimator (convolution theorem) and the local asymptotic minimax risk of every estimator.

Full proof of Corollary 10. Fix any $v \in \mathbb{R}^{J+1}$ with $\|v\|_2 = 1$. The scalar functional $\tau_v = v^\top\theta^*$ is pathwise differentiable with EIF $\phi_v(Z) = v^\top\psi^*(Z)$, having mean zero and variance $\sigma_v^2 = v^\top\Sigma^* v$. Since $\hat\tau_n = v^\top\hat\theta_n$ is a regular estimator of τ_v, the convolution theorem (van der Vaart, 1998, Theorem 25.20) gives:

$$\sqrt{n}(\hat\tau_n - \tau_v) \xrightarrow{d} \mathcal{N}(0,\sigma_v^2) * G_v$$

for some distribution G_v on $\mathbb{R}$, so the limit law of $\sqrt{n}(\hat\tau_n - \tau_v)$ has variance $\sigma_v^2 + \mathrm{Var}(G_v) \ge \sigma_v^2 = v^\top\Sigma^* v$ (possibly $+\infty$). Since v is arbitrary, $\Sigma_\infty \succeq \Sigma^*$. CDR is regular by asymptotic linearity in ψ^* (van der Vaart, 1998, Lemma 25.23) and, under the assumptions of Theorem 6, achieves equality with $G_v = \delta_0$ for every v. □

Theorem 17 (Local asymptotic minimax lower bound)**.** *Under Assumptions 1–2 and $\gamma = 1$, the CMDP model satisfies local asymptotic normality (LAN) at every $P \in \mathcal{P}$ with limiting Fisher information yielding Σ^*. Consequently, for every estimator sequence $\hat\theta_n$ and every subconvex (bowl-shaped), lower-semicontinuous loss $\ell : \mathbb{R}^{J+1} \to \mathbb{R}_+$,*

$$\sup_I \liminf_{n\to\infty} \max_{g\in I} \mathbb{E}_{P_{1/\sqrt{n},\,g}}\Big[\ell\big(\sqrt{n}\,(\hat\theta_n - \theta(P_{1/\sqrt{n},\,g}))\big)\Big] \ge \mathbb{E}_{Z\sim\mathcal{N}(0,\Sigma^*)}[\ell(Z)], \tag{21}$$

where the supremum is over finite subsets I of the tangent set of bounded scores $g = (g_0, \{g_h^\pi\}, \{g_{h'}\})$ and $t \mapsto P_{t,g}$ is the joint submodel constructed in the proof. (A fixed-radius version with $\sup_{\|h\|_2\le 1}$ is false: shrinking toward the known centre beats it.) For squared loss $\ell(u) = \|u\|^2$ the right-hand side equals $\mathrm{tr}(\Sigma^*)$. Under the assumptions of Theorem 6, CDR is asymptotically linear with influence function ψ^*, hence regular (van der Vaart, 1998, Lemma 25.23), and attains the bound for every bounded, a.e.-continuous subconvex ℓ; for squared loss it attains $\mathrm{tr}(\Sigma^*)$ provided $n\|\hat\theta_n - \theta(P_{1/\sqrt{n},g})\|^2$ is uniformly integrable.

Proof. **Joint local submodel.** The data generating distribution depends on three components: the initial distribution μ, the behavior policy π_0, and the transition kernels $\{P_h\}_{h=1}^H$. To obtain a sharp LAN expansion that matches every entry of Σ^*, we perturb all three simultaneously. Let g_0, g_h^π ($h = 1,\dots,H$) and $g_{h'}$ ($h' = 1,\dots,H$) be *bounded* score functions (bounded scores are dense in the respective L^2 tangent spaces, e.g. by truncation and conditional re-centering, so $\dot{\overline{\mathcal{P}}}$ and the EIF are unchanged) satisfying the orthogonality constraints $\mathbb{E}_\mu[g_0] = 0$, $\mathbb{E}_{\pi_0(\cdot|s_h)}[g_h^\pi(s_h,\cdot)] = 0$ for all s_h, and $\mathbb{E}_{P_{h'}}[g_{h'} \mid s,a] = 0$ for all (s,a). Define the local submodel at $t = u/\sqrt{n}$: $\mu^{(t)}(s) = \mu(s)(1 + t\,g_0(s))$, $\pi_0^{(t)}(a|s_h) = \pi_0(a|s_h)(1 + t\,g_h^\pi(s_h,a))$, and $P_{h'}^{(t)}(s'|s,a) = P_{h'}(s'|s,a)(1 + t\,g_{h'}(s,a,s'))$, for $|t| \le 1/(2\max\|g\|_\infty)$, so that all three remain probability measures; write $P_{t,g}$ for the resulting trajectory law.

Log-likelihood expansion. By the trajectory factorization $dP(Z) = \mu(s_1)\prod_h \pi_0(a_h|s_h)\prod_h P_h(s_{h+1}|s_h,a_h)$, the total log-likelihood ratio decomposes additively over the three components:

$$\log\frac{dP_{u/\sqrt{n}}^{(n)}}{dP^{(n)}}(Z_{1:n}) = \frac{u}{\sqrt{n}}\sum_{i=1}^n S_g(Z_i) - \frac{u^2}{2}\,I_g + o_p(1),$$

with combined score $S_g(Z) = g_0(s_1) + \sum_{h=1}^{H} g_h^\pi(s_h, a_h) + \sum_{h'=1}^{H} g_{h'}(s_{h'}, a_{h'}, s_{h'+1})$. Because the three component scores are mutually orthogonal under P (their zero-mean conditioning sets are nested σ-algebras), the Fisher information is additive: $I_g = \mathrm{Var}_\mu(g_0) + \sum_h \mathbb{E}_{d_h^{\pi_0}}[\mathrm{Var}_{\pi_0(\cdot|s_h)}(g_h^\pi(s_h, \cdot))] + \sum_{h'} \mathbb{E}_{d_{h'}^{\pi_0}}[\mathbb{E}_{P_{h'}}[g_{h'}^2 \mid s_{h'}, a_{h'}]]$. The CLT for S_g together with a second-order Taylor expansion of the log-likelihood ratio yields the LAN expansion (van der Vaart, 1998, Chapter 7).

Matching Σ^* via the three tangent directions. By the EIF derivation (Appendix L): (i) the g_0-direction generates the initial-state term $\mathrm{Cov}_\mu(V_1^{\pi,\star_j}, V_1^{\pi,\star_k})$ in (8); (ii) the $\{g_{h'}\}$-directions generate the within-step transition covariance $\sum_h \mathbb{E}_{d_h^{\pi_0}}[w_h^2 \mathrm{Cov}_{P_h}(\delta_h^{\star_j}, \delta_h^{\star_k} \mid s_h, a_h)]$; (iii) by the orthogonality established above (i.e., $\psi^* \perp T_{\pi_0}$ in $L^2(P)$, see Theorem 3 proof in Appendix L), the Riesz representation of the parameter functional restricted to T_{π_0} vanishes; hence the π_0-direction does not contribute to the efficiency bound. Concretely, for the scalar functional $\tau_v = v^\top \theta^*$ the canonical gradient is $v^\top \psi^*$, which lies in $T_\mu \oplus \bigoplus_h T_{P_h}$ (since $V_1 - \theta \in T_\mu$ and $w_h \delta_h \in T_{P_h}$; its T_{π_0}-projection is zero by (iii)). The local-asymptotic-minimax efficiency bound for τ_v is the squared $L^2(P)$-norm of this gradient, $\mathbb{E}[(v^\top \psi^*)^2] = v^\top \Sigma^* v$, i.e. the inverse Fisher information of the least-favourable submodel in the S_g-expansion above (the supremum of $(\partial_t \tau_v)^2 / I_g$ over score directions g is attained by the gradient and equals $v^\top \Sigma^* v$). Since $v \in \mathbb{R}^{J+1}$ is arbitrary, the Gaussian limit experiment has covariance exactly Σ^*, with the π_0-tangent slice contributing no information.

Connecting LAN to the efficiency bound. Theorem 25.21 of van der Vaart (1998), applied with the linear-space tangent set of bounded scores and efficient influence function ψ^*, yields (21) for the local minimax risk over finite subsets of the tangent set. Attainment by CDR follows from its asymptotic linearity (Steps 1–4 of Appendix M) and van der Vaart (1998, Lemma 25.23). □

Remark 25 (Statistical limits of CMDP offline evaluation)**.** *The LAN bound and Corollary 10 imply that no* regular *estimator of θ^* can have $\sqrt{n}(\hat{\theta}_n - \theta^*)$ with asymptotic covariance smaller than Σ^*, and that no estimator can have local asymptotic minimax squared-error risk (in the sense of* (21)*) below* $\mathrm{tr}(\Sigma^*)$*. Superefficiency at isolated points (Hodges-type estimators) is not excluded. CDR is therefore both statistically optimal and computationally tractable.*

U Discussion: Scope of the $\gamma = 1$ Assumption

U.1 Settings Where $\gamma = 1$ Is Natural

The $\gamma = 1$ (undiscounted, finite-horizon) formulation covers a wide class of practically important problems, including healthcare treatment protocols (Murphy et al., 2001), robotics with fixed episode lengths, and recommender systems evaluated over a session. The average-reward setting ($\gamma = 1$, $H \to \infty$, target normalized by $1/H$) is not covered by our finite-horizon analysis and is left open (see the end of this section).

U.2 Why the Undiscounted Score Fails at $\gamma < 1$

The barrier is specific to the *undiscounted* score (6), not to discounting itself. For $\gamma < 1$, applying (6) unchanged with w correct makes the telescoping sum yield a non-zero residual:

$$\mathbb{E}[\psi^\star] = (\gamma - 1) \sum_{h=2}^{H} \mathbb{E}_{d_h^\pi}\big[\check{V}_h^\star(s_h)\big] + \sum_{h=2}^{H} (1 - \gamma^{h-1})\, \mathbb{E}_{d_h^\pi}[f_h^\star], \tag{22}$$

where $\check{V}_h^\star$ is the value function under the misspecified Q. The first term arises from the failed telescoping of A-terms; the second from the mismatch between the undiscounted reward sum $\sum_h F_h$ and the discounted target $v_\star^\pi = \sum_h \gamma^{h-1} F_h$. The two terms cancel at the true Q but not in general: the residual equals $(\gamma - 1) \sum_{h \geq 2} \mathbb{E}_{d_h^\pi}[\check{V}_h^\star - V_h^{\pi,\star}]$, which is non-zero for generic misspecification and is uncontrolled by w alone, breaking double robustness. Both terms trace to the same mismatch: the target discounts step h by γ^{h-1} while the correction does not.

U.3 Resolution: Discount-Weighted Density Ratios

Carrying that same weight into the score—replacing w_h by $\tilde{w}_h = \gamma^{h-1} w_h$—removes the mismatch and restores the *entire* theory for every $\gamma \in (0, 1]$. Appendix J establishes, for the discount-weighted score (19): double robustness (Proposition 13), Neyman orthogonality (Proposition 14), the efficient influence function and the joint efficiency bound $\tilde{\Sigma}^*$ including its off-diagonal entries (Theorem 15), and joint asymptotic normality (Remark 21). The cross-objective covariance structure survives intact: all components are still evaluated on the same trajectories with the same $\tilde{w}_h$, and the off-diagonal entries of $\tilde{\Sigma}^*$ have the same form up to the known factor $\gamma^{2(h-1)}$, so the calibration analysis of Section 4—which depends only on the sign and magnitude of the off-diagonal entries—applies verbatim. Discounting in fact *strengthens* the horizon guarantee: by Corollary 16, $\tilde{\Sigma}^*_{jj}$ is bounded uniformly in H for any $\gamma < 1$ (for an H-independent bound on $\|w_h\|_\infty$), whereas the $\gamma = 1$ bound grows polynomially in H (linearly under a uniform per-step noise bound). What remains open is the infinite-horizon ($H \to \infty$) stationary formulation, not the discounted finite-horizon one.