WHERE TO QUERY: DATASET AGGREGATION AS QUADRATURE IN INTERACTIVE IMITATION LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Interactive imitation learning often collects data under mixtures of expert and learner behaviour to mitigate covariate shift, but mixture points and weights are typically specified as a training schedule without an explicit approximation target. We show that, along a fixed expert–learner path, weighted dataset aggregation is a numerical quadrature rule for the path-average occupancy. This target assigns positive mass to every state reachable under some mixture on the discounted path, whereas aggregates supported only on the endpoint policies can miss states that occur under interior mixtures. For finite-state Markov Decision Processes (MDPs), classical quadrature theory yields schedule-dependent rates, including an $O(n^{-1})$ endpoint-Riemann bound and geometric convergence for Gauss–Legendre, with constants that grow with the effective horizon or mixing time. Exact evaluation reproduces the rate hierarchy and, on Garnets, the predicted endpoint limit. Under independent finite sampling, Gauss–Legendre and midpoint reach an error no more than 20% above that of direct path-average sampling with the same budget in at most three rounds at every tested budget for the median instance, whereas endpoint Riemann, DAgger, and random-uniform schedules require many more rounds or fail to reach this level as the budget grows. In continuous control, with retraining held fixed, path-covering node sets lower reference-path imitation loss relative to DAgger-P's endpoint nodes on three of four environments but improve return significantly on at most two.

1 INTRODUCTION

Imitation learning by supervised prediction from expert demonstrations (Osa et al., 2018; Argall et al., 2009) is vulnerable to covariate shift, because small prediction errors can move the learner into states that are rare in the expert data, where further errors may compound (Ross & Bagnell, 2010). Interactive methods such as Dataset Aggregation (DAgger) instead collect expert labels on states encountered during learning. DAgger rolls out the current learner, optionally mixed with the expert, aggregates the labelled states, and retrains (Ross et al., 2011). The expert–learner mixture used for collection is specified by a per-round schedule, but this schedule is usually treated as part of the training procedure rather than as an approximation design with an explicit occupancy target. We ask what occupancy such a schedule represents and how the placement and weighting of its rounds affect that representation. Our focus is therefore the global placement of collection points along an expert–learner mixture path, rather than state-conditional or uncertainty-gated decisions about when to query the expert.

We study this question along a fixed path between an initial learner and an expert. Each mixture policy induces a state occupancy, so a finite collection schedule samples this occupancy path at selected mixture points and combines the resulting occupancies with prescribed weights. At the population level, this is a numerical quadrature rule: the mixture coefficients are its nodes, the aggregation coefficients are its weights, and the quantity being approximated is the occupancy averaged along the path. We use the uniform average as the coverage target. This weighting is a design choice, motivated by α being the per-step probability of selecting the expert action. For the discounted path, the resulting average assigns positive mass to every state reached by some interior expert–learner mixture, whereas an aggregate supported only on the two endpoint policies can miss states that arise only under interior mixtures.

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.

The usefulness of the quadrature view depends on the regularity of the occupancy path. For finite-state Markov decision processes (MDPs), discounted occupancies are analytic along the mixture path, stationary occupancies are analytic under a uniqueness condition, and finite-horizon occupancies are polynomial. Classical quadrature theory (Davis & Rabinowitz, 2007) then gives schedule-dependent approximation rates: first order for endpoint Riemann, second order for midpoint and trapezoidal rules, fourth order for Simpson's rule, and geometric convergence for Gauss–Legendre. The associated constants grow with the effective horizon or mixing time. The limiting aggregate is determined by the schedule measure: schedules whose measures converge to the uniform measure recover the path average, whereas schedules concentrating on an endpoint recover that endpoint's occupancy.

Exact evaluation on Garnet MDPs, together with an independent linear quadratic regulator (LQR) check, reproduces the predicted approximation hierarchy, and on Garnets the endpoint-concentrating schedules approach the predicted limit. Under independent finite sampling, Gauss–Legendre and midpoint reach an error no more than 20% above that of direct path-average sampling with the same budget in at most three collection rounds at every tested budget for the median held-out instance. Endpoint Riemann, the tested DAgger schedules, and random-uniform schedules require substantially more rounds or fail to reach this level within the tested grid as the budget grows. On four MuJoCo environments (Todorov et al., 2012), with retraining held fixed, frozen-path schedules containing interior mixtures lower reference-path imitation loss relative to DAgger-P's endpoint nodes on three environments, but improve return significantly on at most two (Section 6.4). These results characterise coverage along a fixed reference path rather than the performance of a learned policy. Retraining moves collection away from that path, and the learned policy's own occupancy need not lie on it. Consequently, fixed-path coverage error does not by itself provide a bound on the learned policy's loss or return.

Contributions. We formulate weighted dataset aggregation along a fixed expert–learner path as numerical quadrature and characterise the support coverage of its target. We then establish the regularity required by classical quadrature and use it to derive schedule-dependent convergence rates with horizon- and mixing-dependent constants. We also characterise the occupancy approached by each schedule in the limit and derive a finite-sample bound separating quadrature bias, weight mismatch, and sampling error. We test these predictions under exact and finite-sample evaluation and in continuous control.

2 Related Work

Interactive imitation learning. Compounding error under covariate shift motivated interactive methods based on stochastic policy mixing (Ross & Bagnell, 2010) and DAgger's no-regret dataset aggregation (Ross et al., 2011). Later work changes the learning signal, as AggreVaTe and AggreVaTeD do using the expert's cost-to-go (Ross & Bagnell, 2014; Sun et al., 2017), or changes when expert supervision is requested. HG-DAgger relies on human intervention, SafeDAgger uses a learned safety criterion, EnsembleDAgger uses model uncertainty, ThriftyDAgger combines novelty and risk under an intervention budget, and AIM learns an adaptive intervention rule (Kelly et al., 2019; Zhang & Cho, 2017; Menda et al., 2019; Hoque et al., 2021; Cai et al., 2025). These methods adapt supervision or control at the state level. We instead study how a prescribed set of mixture coefficients and aggregation weights represents a fixed expert–learner occupancy path. Related theoretical work studies convergence of value aggregation (Cheng & Boots, 2018), the minimax limits of active expert queries in tabular MDPs (Rajaraman et al., 2020), and the horizon dependence of on- and off-policy moment matching (Swamy et al., 2021). Our focus is the occupancy represented by a weighted collection schedule.

Policy mixtures and occupancy perturbation. Conservative policy iteration updates towards per-state policy mixtures and bounds the resulting change in state distribution through the probability of departing from the base policy (Kakade & Langford, 2002). We instead treat the full occupancy curve as the object to approximate and bound its derivatives through the resolvent of the mixed transition matrix. Perturbation theory for Markov chains gives derivatives of stationary distributions with respect to transition probabilities (Schweitzer, 1968) and condition numbers that quantify their sensitivity (Cho & Meyer, 2001). The quantity M_∞ in Corollary 6 plays the corresponding sensitivity role in our stationary-path bounds.

Numerical integration and path sampling. Expected Policy Gradients integrates the policy gradient over actions, analytically where possible and numerically otherwise, and notes that designed abscissae such as Gauss–Legendre nodes can replace sampled actions (Ciosek & Whiteson, 2020). Our integration variable is instead the expert–learner mixing coefficient, so each quadrature node is a data-collection policy rather than an action. Thermodynamic integration and path sampling integrate along scalar paths of distributions and study discretisation error, point placement, and adaptive schedules (Gelman & Meng, 1998; Calderhead & Girolami, 2009; Friel et al., 2014; Hug et al., 2016). The corresponding distinction here is between quadrature bias and finite-sample error. Our Gauss–Legendre bound uses the classical Bernstein-ellipse argument (Trefethen, 2008).

3 Coverage Along a Reference Policy Path

Throughout the theory, $\mathcal{S}$ is finite, occupancies are probability vectors in $\Delta(\mathcal{S})$, $d_{\mathrm{TV}}(\mu,\nu) = \frac{1}{2}\|\mu - \nu\|_1$, and μ_0 is the initial-state distribution. We study the mixture path

$$\pi_\alpha = (1-\alpha)\hat{\pi}_0 + \alpha\pi^\star, \qquad \alpha \in [0,1], \tag{1}$$

with induced occupancy $\rho_\alpha := \rho_{\pi_\alpha}$, so that ρ_0 and ρ_1 are the learner and expert endpoint occupancies. We write ρ_α^γ, ρ_α^∞, and ρ_α^H for the normalised discounted, stationary, and H-step truncated occupancies, respectively.

Per-step mixing (PM). At every step, the executed action is drawn from $\pi_\alpha(\cdot \mid s)$ independently of the past, as in DAgger's β-mixtures and in our implementation. Under PM the transition matrix is affine in α (Section 5.1). Under episode-level mixing the occupancy itself is affine in α, and Proposition 7 gives what that implies for the rates.

For $\eta \in \mathcal{P}([0,1])$, let $\bar{\rho}_\eta := \int_{[0,1]} \rho_\alpha \, \eta(d\alpha)$ be the occupancy averaged along the path according to η (Appendix A.1 notes that reparameterising α leaves it unchanged when the measure is transported accordingly). We take the uniform measure λ on α, a design choice motivated by α being the per-step probability of an expert action, and write $\bar{\rho} := Q_\lambda[\rho] := \int_0^1 \rho_\alpha \, \mathrm{d}\alpha$.

To quantify how well a single distribution represents the path, we define, for $\mu \in \Delta(\mathcal{S})$,

$$r(\mu) := \sup_{\alpha\in[0,1]} d_{\mathrm{TV}}(\mu, \rho_\alpha), \qquad r_{\mathrm{Cheb}} := \inf_{\mu\in\Delta(\mathcal{S})} r(\mu). \tag{2}$$

Theorem 1 (Coverage properties of the path average)**.** *If $\alpha \mapsto \rho_\alpha$ is L-Lipschitz in total variation, then $r(\bar{\rho}) \le \min\{1, L/2\}$. Independently of this assumption,*

$$r(\bar{\rho}) \le 2r_{\mathrm{Cheb}}, \qquad \bar{\rho} = \arg\min_{\mu\in\Delta(\mathcal{S})} \int_0^1 \|\mu - \rho_\alpha\|_2^2 \, \mathrm{d}\alpha, \tag{3}$$

and for every $g : \mathcal{S} \to [0,1]$,

$$\sup_{\alpha\in[0,1]} |\mathbb{E}_{\bar{\rho}}[g] - \mathbb{E}_{\rho_\alpha}[g]| \le r(\bar{\rho}). \tag{4}$$

Theorem 1 gives two interpretations of $\bar{\rho}$. It is the squared-ℓ_2 barycentre of the occupancy path, and its total-variation radius bounds how far the expected value of any $[0,1]$-valued loss under $\bar{\rho}$ can be from its value under any path occupancy. The radius bounds alone do not single out the path average, because the same argument gives $r(\rho_\beta) \le 2r_{\mathrm{Cheb}}$ for every β, including ρ_0, and $r(\rho_{1/2}) \le \min\{1, L/2\}$. The Lipschitz bound is non-trivial only when $L < 2$, and for the discounted path, Corollary 6 bounds it by $L \le \gamma\varepsilon_\pi/(1-\gamma)$, with $\varepsilon_\pi = \max_s d_{\mathrm{TV}}(\hat{\pi}_0(\cdot \mid s), \pi^\star(\cdot \mid s))$, a bound that grows with the effective horizon. Neither bound survives severe shift: the triangle inequality gives $2r_{\mathrm{Cheb}} \ge d_{\mathrm{TV}}(\rho_0, \rho_1)$, so both are trivial when learner and expert visit almost disjoint states.

The radius view does not capture whether an aggregate assigns positive mass to every state encountered along the path. To express this form of coverage, for $\eta \in \mathcal{P}([0,1])$, we define the concentrability coefficient

$$C(\eta) := \sup_{\alpha\in[0,1]} \max_{s:\, \rho_\alpha(s)>0} \frac{\rho_\alpha(s)}{\bar{\rho}_\eta(s)} \in [1,\infty]. \tag{5}$$

For every α and every $g : \mathcal{S} \to [0,\infty)$, $\mathbb{E}_{\rho_\alpha}[g] \le C(\eta)\, \mathbb{E}_{\bar{\rho}_\eta}[g]$, so when $C(\eta) < \infty$ the expected loss under the aggregate controls the expected loss at every point of the reference path.

Table 1: **Collection schedules as node–weight designs on the fixed reference path.** The DAgger rows read those schedules open-loop, without retraining (Section 5.3). The last column is the weak limit of ν_n, which determines the limit of the aggregate (Proposition 8). Here λ is uniform on $[0,1]$, and (x_i, ω_i) are the Gauss–Legendre nodes and weights on $[-1,1]$.

Schedule	Nodes α_i, $i = 1, \dots, n$	Weights w_i	Quadrature view	Limit
DAgger-P	$1, 0, \dots, 0$	$1/n$	endpoint-concentrating	δ_0
DAgger-G	$p^{\,i-1},\ p \in (0,1)$	$1/n$	geometric nodes	δ_0
DAgger-L	$1 - (i-1)/n$	$1/n$	right Riemann	λ
Random-uniform	$\alpha_i \overset{\text{iid}}{\sim} \mathrm{Unif}[0,1]$	$1/n$	Monte Carlo	λ a.s.
Endpoint Riemann	$(i-1)/n$	$1/n$	left Riemann	λ
Midpoint	$(i-1/2)/n$	$1/n$	composite midpoint	λ
Trapezoid ($n \geq 2$)	$(i-1)/(n-1)$	$1/(n-1)$, ends $1/(2n-2)$	composite trapezoid	λ
Simpson ($n \geq 3$ odd)	$(i-1)/(n-1)$	$(1,4,2,\dots,4,1)/(3n-3)$	composite Simpson	λ
Gauss–Legendre	$(1+x_i)/2$	$\omega_i/2$	n-point Gaussian	λ

Proposition 2 (Support coverage of the discounted path)**.** *Assume PM, let $\rho_\alpha = \rho_\alpha^\gamma$ with $\gamma < 1$, and let $\mathcal{S}^\circ := \operatorname{supp} \rho_{1/2}$. Then $\operatorname{supp} \rho_\alpha = \mathcal{S}^\circ$ for every $\alpha \in (0,1)$ and $\operatorname{supp} \rho_0 \cup \operatorname{supp} \rho_1 \subseteq \mathcal{S}^\circ$. Consequently $C(\eta) < \infty$ whenever $\eta((0,1)) > 0$, in particular $C(\lambda) < \infty$, whereas every η supported on $\{0,1\}$ has $C(\eta) = \infty$ whenever $\mathcal{S}^\circ \neq \operatorname{supp} \rho_0 \cup \operatorname{supp} \rho_1$.*

Proposition 2 shows that $C(\eta)$ is finite for schedules with interior weight, but gives no bound on its size. In a chain whose state k is reached only by k consecutive expert actions, $\rho_\alpha(k) = (\gamma\alpha)^k \rho_\alpha(0)$, so density ratios between nearby mixtures grow geometrically in k (Appendix A.3).

4 Dataset Aggregation as Numerical Quadrature

A collection schedule is a *node–weight design* $\mathcal{D}_n = \{(\alpha_i, w_i)\}_{i=1}^n$, with $\alpha_i \in [0,1]$, $w_i \geq 0$, and $\sum_i w_i = 1$. The index n counts nodes with multiplicity, and for DAgger schedules it is the number of collection rounds. Its schedule measure and the population aggregate it induces are

$$\nu_n := \sum_{i=1}^n w_i \delta_{\alpha_i}, \qquad Q_\nu[\rho] := \int_{[0,1]} \rho_\alpha \, \nu(\mathrm{d}\alpha), \tag{6}$$

Proposition 3 (Weighted dataset aggregation induces a quadrature rule)**.** *For any schedule $\mathcal{D}_n$, $Q_{\nu_n}[\rho] = \sum_{i=1}^n w_i \rho_{\alpha_i}$, and $d_{\mathrm{TV}}(Q_{\nu_n}[\rho], \bar{\rho})$ is the error of the node–weight rule $\mathcal{D}_n$ for the integral $Q_\lambda[\rho] = \bar{\rho}$.*

Proposition 3 identifies schedule design with numerical integration along the fixed path, in which the collection mixtures α_i are quadrature nodes, the aggregation weights w_i are quadrature weights, and the collection policies are not refitted between rounds. Schedules can therefore be compared by their approximation error (Section 5.2) and by the limit of their schedule measures (Section 5.3). This is a population-level statement: finite datasets estimate the node occupancies, as analysed in Section 5.4. The intended weights need not match those a pooled dataset realises: if B_i samples are collected at α_i and pooled with equal per-sample weight, node i has realised weight $\widetilde{w}_i := B_i / \sum_j B_j$. Equal-size rounds therefore implement equal weights, and rules with non-uniform weights—including the trapezoid, Simpson, and Gauss–Legendre rules with more than two nodes—require proportional allocation, $B_i \propto w_i$, or explicit reweighting.

Table 1 lists the DAgger schedules, read as node–weight designs on the fixed path, alongside random and classical quadrature designs. Because α is the expert weight, endpoint Riemann, with nodes $(i-1)/n$, includes the learner endpoint, and DAgger-L, with nodes $1-(i-1)/n$, the expert endpoint. As reflections of one another, they share the bound of Theorem 5(ii) and the leading term of Proposition 7.

5 What Schedule Design Predicts

We first derive schedule-dependent approximation rates from occupancy-path regularity, then characterise schedule limits and finite-sample error.

5.1 Occupancy-path regularity

Let $P_0 := P_{\hat{\pi}_0}$, $P_1 := P_{\pi^\star}$, and $\Delta P := P_1 - P_0$. Under PM the transition matrix is affine,

$$P_\alpha = P_0 + \alpha \Delta P, \tag{7}$$

and rollouts that reset to μ_0 on termination retain this form once the reset is absorbed into P. For the stationary case we assume (SU), the *stationary uniqueness* condition that every P_α, $\alpha \in [0,1]$, has a unique stationary distribution.

Theorem 4 (Regularity of occupancy paths)**.** *For every $\gamma < 1$,*

$$\rho_\alpha^\gamma = (1-\gamma)(I - \gamma P_\alpha^\top)^{-1}\mu_0, \tag{8}$$

and $\alpha \mapsto \rho_\alpha^\gamma$ is real analytic on $[0,1]$ with a holomorphic extension to a complex neighbourhood of the interval. Under (SU)*, $\rho_\alpha^\infty = (I - P_\alpha^\top + \mathbf{1}\mathbf{1}^\top)^{-1}\mathbf{1}$ has the same property. For every horizon H, the truncated occupancy $\rho_\alpha^H := H^{-1}\sum_{t<H}(P_\alpha^\top)^t\mu_0$ is a polynomial of degree at most $H-1$ in α.*

Theorem 4 provides the regularity that high-order quadrature needs. It does not use aperiodicity, though estimating stationary occupancies from finite trajectories may require mixing assumptions, and (SU) constrains the induced chains in a way endpoint action support alone does not imply.

5.2 Approximation hierarchy

For a schedule measure ν_n, define the exact coverage error, the population-level error before any sampling, $\mathcal{E}(\nu_n) := d_{\mathrm{TV}}(Q_{\nu_n}[\rho], \bar{\rho})$. For the canonical rules, n is the number of distinct exact path evaluations. Midpoint uses n cells, trapezoid $n-1$ subintervals, Simpson $n-1$ even subintervals, and Gauss–Legendre n nodes (Appendix A.7). Write $M_k := \sup_\alpha \|\partial_\alpha^k \rho_\alpha\|_1$.

Theorem 5 (Schedule-dependent approximation rates)**.** *Let $d := |\mathcal{S}|$ and consider exact occupancy evaluations.* (i) *For $\alpha_i \overset{\text{iid}}{\sim} \mathrm{Unif}[0,1]$ and any measurable path, with probability at least $1-\delta$, $\mathcal{E}(\nu_n^{\mathrm{RU}}) \le \frac{1}{2}\sqrt{d/n} + \sqrt{\log(1/\delta)/(2n)}$.* (ii) *If the path is L-Lipschitz in total variation, $\mathcal{E}(\nu_n^{\mathrm{ER}}) \le L/(2n)$.* (iii) *If it is C^2, $\mathcal{E}(\nu_n^{\mathrm{mid}}) \le M_2/(48n^2)$ and $\mathcal{E}(\nu_n^{\mathrm{trap}}) \le M_2/(24(n-1)^2)$.* (iv) *If it is C^4 and n is odd, $\mathcal{E}(\nu_n^{\mathrm{Simp}}) \le M_4/(360(n-1)^4)$.* (v) *If it extends holomorphically to a neighbourhood of the region bounded by a shifted Bernstein ellipse $\mathcal{B}_\varrho$, $\varrho > 1$, with $M_\varrho := \sup_{z\in\mathcal{B}_\varrho}\|\rho(z)\|_1$, then $\mathcal{E}(\nu_n^{\mathrm{GL}}) \le C_\varrho \varrho^{-2n}$ with $C_\varrho := 2M_\varrho\varrho/(\varrho-1)$.*

The hierarchy reflects how much regularity each design exploits. Random-uniform placement needs no smoothness, endpoint Riemann uses Lipschitz continuity, the composite rules use higher derivatives, and Gauss–Legendre uses analyticity. Theorem 4 supplies this regularity for (ii)–(v), whose bounds carry no explicit d, unlike the Monte Carlo term in (i), though the constants can worsen with the effective horizon or mixing time. With ε_π as in Section 3, set $\kappa := \gamma/(1-\gamma)$ for the discounted path, $\kappa := M_\infty := \sup_\alpha \|(I - P_\alpha^\top + \mathbf{1}\mathbf{1}^\top)^{-1}\|_{1\to 1}$ for the stationary path under (SU), a path-wide condition number (Cho & Meyer, 2001) that can exceed its endpoint values, and $\kappa := H-1$ for the truncated path.

Corollary 6 (Horizon-explicit constants)**.** *In Theorem 5, $L \le \kappa\varepsilon_\pi$ and $M_k \le k!\,(2\kappa\varepsilon_\pi)^k$. Hence $\mathcal{E}(\nu_n^{\mathrm{ER}}) \le \kappa\varepsilon_\pi/(2n)$, $\mathcal{E}(\nu_n^{\mathrm{mid}}) \le (\kappa\varepsilon_\pi)^2/(6n^2)$, $\mathcal{E}(\nu_n^{\mathrm{trap}}) \le (\kappa\varepsilon_\pi)^2/(3(n-1)^2)$, and $\mathcal{E}(\nu_n^{\mathrm{GL}}) \le 4(1+2\kappa\varepsilon_\pi)(1+\frac{1}{2\kappa\varepsilon_\pi})^{-2n}$, which is at most ϵ once $n \ge (\frac{1}{2}+\kappa\varepsilon_\pi)\log\big(4(1+2\kappa\varepsilon_\pi)/\epsilon\big)$.*

For fixed $\kappa\varepsilon_\pi$, the Gauss–Legendre bound thus guarantees accuracy ϵ with a number of exact path evaluations logarithmic in $1/\epsilon$, while the horizon or mixing dependence enters through κ, both in the rate and in the count needed before the bound falls below one; these are path evaluations rather than expert labels, which enter in Section 5.4. The upper bounds need not determine the finite-n ordering of the rules, and for endpoint Riemann, midpoint, and trapezoid, Proposition 7 gives the leading terms.

Proposition 7 (Leading error terms). *If $\alpha \mapsto \rho_\alpha$ is C^2, then $\mathcal{E}(\nu_n^{\mathrm{ER}}) = d_{\mathrm{TV}}(\rho_0, \rho_1)/(2n) + O(n^{-2})$. If it is C^4, then with $\rho_a' := \partial_\alpha \rho_\alpha|_{\alpha=a}$,*

$$\mathcal{E}(\nu_n^{\mathrm{mid}}) = \frac{\|\rho_1' - \rho_0'\|_1}{48n^2} + O(n^{-4}), \qquad \mathcal{E}(\nu_n^{\mathrm{trap}}) = \frac{\|\rho_1' - \rho_0'\|_1}{24(n-1)^2} + O(n^{-4}).$$

The endpoint rule is exactly first order iff $\rho_0 \neq \rho_1$; midpoint and trapezoid are exactly second order iff $\rho_0' \neq \rho_1'$, and then their error ratio tends to 2. If the path is affine in α, every rule exact for linear functions has zero error and only endpoint rules retain a bias.

The orders of the upper bounds are thus attained under the stated non-degeneracy conditions, while affine paths are integrated exactly by every rule exact for linear functions. Appendix G checks Corollary 6 and Proposition 7 numerically.

5.3 Consistency and endpoint concentration

The rates in Theorem 5 describe schedules whose measures converge to the uniform measure λ; for a general schedule, the limit of the aggregate is determined by the weak limit of its schedule measure.

Proposition 8 (Schedule-limit principle). *Suppose $\alpha \mapsto \rho_\alpha$ is continuous in total variation and $\nu_n \Rightarrow \nu$ weakly in $\mathcal{P}([0,1])$. Then $d_{\mathrm{TV}}(Q_{\nu_n}[\rho], Q_\nu[\rho]) \to 0$. In particular, if $\nu_n \Rightarrow \lambda$, then $d_{\mathrm{TV}}(Q_{\nu_n}[\rho], \bar{\rho}) \to 0$. If instead $\nu_n \Rightarrow \delta_a$, then*

$$d_{\mathrm{TV}}(Q_{\nu_n}[\rho], \rho_a) \to 0, \qquad \mathcal{E}(\nu_n) \to d_{\mathrm{TV}}(\rho_a, \bar{\rho}). \tag{9}$$

Weak convergence to λ is sufficient, but not necessary, for consistency on a particular path. In Table 1, DAgger-P and DAgger-G converge weakly to δ_0, random-uniform to λ almost surely, and the rest to λ. For DAgger-P with n equal-weight rounds on the fixed path,

$$\nu_n^{\mathrm{P}} = \frac{n-1}{n}\delta_0 + \frac{1}{n}\delta_1 \Longrightarrow \delta_0, \qquad Q_{\nu_n^{\mathrm{P}}}[\rho] \to \rho_0, \tag{10}$$

so its coverage error tends to $d_{\mathrm{TV}}(\rho_0, \bar{\rho})$ on the fixed path. If the learner is retrained between rounds, later rounds sample the retrained learners' occupancies instead, so (10) does not describe the resulting DAgger procedure.

5.4 Finite sampling

At node α_i, collect $B_i \geq 0$ samples $S_{i,j} \sim \rho_{\alpha_i}$, let $B := \sum_i B_i > 0$, and pool them with equal weight:

$$\widetilde{w}_i := \frac{B_i}{B}, \qquad \widehat{Q}_{n,B} := \frac{1}{B}\sum_{i=1}^{n}\sum_{j=1}^{B_i} e_{S_{i,j}}, \tag{11}$$

where e_s is the one-hot vector for state s.

Theorem 9 (Finite-sample coverage bound). *Let $d := |\mathcal{S}|$ and assume all samples in (11) are mutually independent. For every $\delta \in (0,1)$, with probability at least $1-\delta$,*

$$d_{\mathrm{TV}}(\widehat{Q}_{n,B}, \bar{\rho}) \leq \mathcal{E}(\nu_n) + \tfrac{1}{2}\|\widetilde{w} - w\|_1 + \tfrac{1}{2}\sqrt{d/B} + \sqrt{\log(1/\delta)/(2B)}. \tag{12}$$

Theorem 9 separates the empirical coverage error into the quadrature bias $\mathcal{E}(\nu_n)$ of the node–weight design, the mismatch between the intended weights and those realised by the collected data, and sampling noise. Under equal pooling the sampling term depends only on the total budget B, and proportional allocation keeps the realised weights aligned with the intended ones. If the node estimates are instead explicitly re-weighted by the intended weights, the mismatch term vanishes, but the sampling term depends on the allocation through $\sum_i w_i^2/B_i \geq 1/B$, with equality under proportional allocation (Lemma 6).

Theorem 9 assumes independent samples from a finite state space. Samples collected along trajectories are temporally correlated, so their analysis would require a mixing or effective-sample-size argument. Continuous state spaces would also require a different discrepancy, such as one defined over a class of loss functions, rather than total variation. The theorem therefore does not apply directly to our continuous-control experiments.

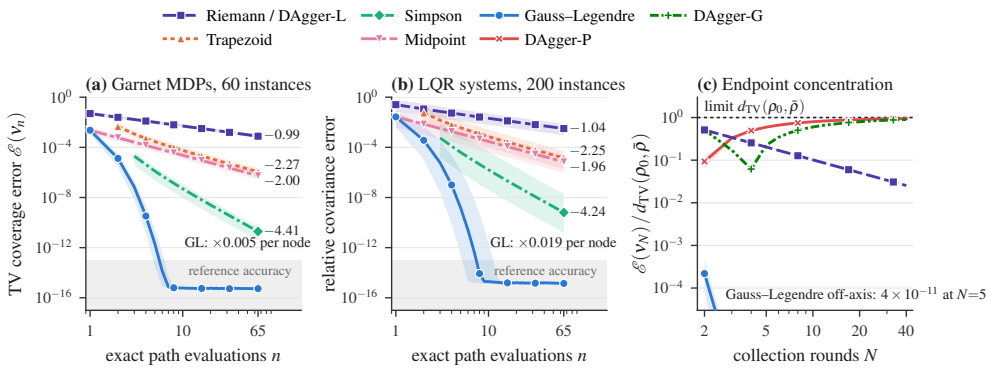

Figure 1: **Exact coverage behaviour** on frozen paths. Curves show per-n medians with interquartile ranges. (a) TV error on Garnets and (b) relative covariance error on LQR; labels give mean log–log slopes, and Gauss–Legendre contracts by median factors 0.005 and 0.019 per node, respectively, until reference accuracy. (c) Garnet errors normalised by $d_{\mathrm{TV}}(\rho_0, \bar{\rho})$; the dashed line is the endpoint-concentrating limit of Proposition 8.

6 Experiments

We test the main predictions of Sections 5.2–5.4 at three levels of realism. Exact path evaluation gives the rate hierarchy and the limiting behaviour of the schedule measures. Under finite sampling we count the collection nodes a schedule needs, at a fixed budget, to reach an error no more than 20% above that of direct path-average sampling. Continuous control then asks which differences remain attributable to the schedule once policy fitting and retraining enter.

6.1 Experimental setup

Exact and finite-sampling experiments. We use Garnet MDPs (Archibald et al., 1995; Bhatnagar et al., 2009) for the finite-state experiments. Their state occupancies can be computed to high accuracy, allowing us to measure the coverage error directly in total variation. As a continuous-state check we also use LQR systems. The LQR experiment is not an instance of the finite-state occupancy theory: we interpolate the feedback gain between a perturbed controller and the optimal one, integrate the resulting state covariance along that path, and measure approximation error in relative Frobenius norm. For each instance, we compute the corresponding reference path integral in at least two independent ways and require agreement to 10^{-10}. Instances that fail this check are excluded and reported.

For the finite-sampling experiments, each of the n nodes receives $\lfloor B/n \rfloor$ states drawn independently from its exact stationary occupancy, and the node estimates are combined with the rule's weights. We define the *rounds to reference* as the smallest tested n whose coverage error is at most 1.2 times the error obtained from B states drawn directly from $\bar{\rho}$, a *direct-sampling reference* with the same total budget but no collection schedule (Appendix F gives the full settings).

Continuous control. We evaluate Hopper, HalfCheetah, Walker2d, and Ant with 40 seeds over $N \in \{3, 5, 10\}$ rounds, label budgets $B \in \{500, 2000, 10000\}$, two initialisations, and six methods: BC, DAgger-P, DAgger-L, DAgger-G, Gauss–Legendre, and trapezoid. The pre-specified centre cell is $N = 5$, $B = 2000$, with a behavioural-cloning initialisation $\hat{\pi}_0$. The quadrature methods roll out the fixed mixtures of $\hat{\pi}_0$ and the expert at their nodes, allocate rollouts in proportion to the weights, and fit the learner once. The DAgger methods begin with an expert node and run $N-1$ rounds under mixtures of the *current* learner and the expert, refitting after every round, so their later collection policies leave the fixed path. The two families also differ in policy fits and environment interaction (Table 9), so between-family contrasts compare procedures rather than node sets. All methods are scored by

$$J(\hat{\pi}) := \frac{1}{20}\sum_{k=0}^{19} \mathbb{E}_{s\sim\hat{\rho}_{\alpha_k}} \left\|\hat{\pi}(s) - \pi^\star(s)\right\|_2^2, \qquad \alpha_k = \frac{2k+1}{41}, \tag{13}$$

Table 2: **Rounds to the direct-sampling reference** on 60 held-out Garnet instances. Entries are the median smallest n with error $\leq 1.2\,\varepsilon_B$, where ε_B is the error from B direct samples from $\bar{\rho}$ (medians over 10 draws). Parentheses give the percentage reaching the reference, and >33 means the median instance never does within the grid. Sampling is i.i.d. from exact frozen-path occupancies.

	Total sample budget B				
Schedule	10^3	10^4	10^5	10^6	10^7
ε_B	5.3×10^{-2}	1.8×10^{-2}	5.3×10^{-3}	1.7×10^{-3}	5.5×10^{-4}
Gauss–Legendre	1	1	1 (98)	2 (98)	2
Midpoint	1	1	1	2 (98)	3
Simpson	5	5	5 (98)	5 (95)	5
Trapezoid	2	2	3	4 (98)	5
Random-uniform	1	2 (98)	5 (83)	29 (58)	>33 (25)
Endpoint Riemann	2	5 (98)	20 (80)	>33 (35)	>33 (0)
DAgger-L	2	5	20 (77)	>33 (38)	>33 (0)
DAgger-P	2 (88)	2 (75)	>33 (43)	>33 (12)	>33 (0)

where $\hat{\rho}_{\alpha_k}$ holds states from episodes of the fixed mixture π_{α_k}, shared across methods. J thus estimates imitation loss along the reference path, and is neither return nor imitation loss under the learned policy's own state distribution. Contrasts are normalised by J_{bc}. Runs whose rollouts end too early to supply their label budget are recorded as refusals; there are none at the centre cell (Appendix D). Appendices E and G report the preregistration, amendments, failed criteria, and follow-up protocols.

6.2 Exact coverage and endpoint concentration

Figures 1a and 1b show the predicted rate hierarchy. Endpoint Riemann is first order, midpoint and trapezoid second order, Simpson fourth order, and Gauss–Legendre geometric until it reaches the accuracy of the reference integral. On Garnets the mean fitted midpoint slope is -1.996, and at $n = 33$ the median ratio of trapezoid to midpoint error is 2.127, the value $2(33/32)^2$ predicted by Proposition 7. The same pattern appears for the discounted targets with $\gamma \in \{0.9, 0.95, 0.99\}$ (Appendix B). Figure 1c tests the limiting behaviour predicted by Proposition 8: the coverage errors of DAgger-P and DAgger-G approach $d_{\mathrm{TV}}(\rho_0, \bar{\rho})$ rather than zero, while those of DAgger-L and Gauss–Legendre tend to zero.

6.3 Finite sampling and rounds to reference

On 60 held-out Garnet instances disjoint from those used to formulate the hypotheses, the one-point Gauss–Legendre and midpoint rules, both using $\alpha = \frac{1}{2}$, reach the reference through $B = 10^5$ (Table 2). Gauss–Legendre needs two nodes at $B \in \{10^6, 10^7\}$ and midpoint three at $B = 10^7$. Random-uniform placement needs 29 nodes at $B = 10^6$ and does not reach the reference within the tested grid at $B = 10^7$; endpoint Riemann and DAgger-L do not reach it within the grid from $B = 10^6$, and DAgger-P from $B = 10^5$.

Simpson needs five nodes at every budget, which its fourth-order quadrature error alone does not explain. Each node receives about B/n samples and the node estimates are reweighted, so the sampling term depends on $\sum_i w_i^2/B_i$ (Lemma 6), and with nearly equal B_i the variance relative to proportional allocation is scaled by about $n\sum_i w_i^2$. For three-node Simpson that factor puts the error scale just above the 1.2 tolerance, and at the looser tolerances 1.5 and 2.0 Simpson needs three nodes (Appendix C). This is consistent with the separation of quadrature bias from sampling error in Section 5.4, though it does not test its constants.

6.4 Continuous control

The continuous-control experiment lies outside the fixed-path coverage theory: the state space is continuous, DAgger retrains and leaves the reference path, and J measures reference-path imitation loss rather than occupancy error or return.

Table 3: **Frozen-path contrasts at the centre cell** ($N = 5$, $B = 2000$, behavioural-cloning initialisation, 40 seeds). Rows 1–4 vary the node set with retraining fixed; row 5 varies retraining at DAgger-P's nodes. Entries are mean paired differences: a negative J/J_{bc} or a positive return favours the first method, with return normalised per seed by the expert-minus-$\hat{\pi}_0$ return. Bold is resolved after Holm correction over the 44 tests per outcome; intervals are in Tables 14 and 15. Hop., HC, Wlk.: Hopper, HalfCheetah, Walker2d.

	J/J_{bc}				Normalised return			
Contrast	Hop.	HC	Wlk.	Ant	Hop.	HC	Wlk.	Ant
$\mathrm{GL^f} - \mathrm{P^f}$	$\mathbf{-0.320}$	$\mathbf{-0.110}$	$\mathbf{-0.265}$	$+0.042$	$\mathbf{+0.209}$	-0.005	$+0.023$	$+0.142$
$\mathrm{Trap^f} - \mathrm{P^f}$	$\mathbf{-0.315}$	$\mathbf{-0.098}$	$\mathbf{-0.266}$	$+0.013$	$\mathbf{+0.130}$	$+0.002$	$+0.020$	$\mathbf{+0.288}$
$\mathrm{L^f} - \mathrm{P^f}$	$\mathbf{-0.303}$	$\mathbf{-0.110}$	$\mathbf{-0.247}$	-0.028	$\mathbf{+0.232}$	$+0.010$	$+0.021$	$+0.105$
$\mathrm{G^f} - \mathrm{P^f}$	$\mathbf{-0.313}$	$\mathbf{-0.108}$	$\mathbf{-0.264}$	$\mathbf{-0.095}$	$\mathbf{+0.123}$	$+0.022$	$+0.022$	$+0.179$
$\mathrm{P} - \mathrm{P^f}$	$\mathbf{-0.103}$	$\mathbf{-0.074}$	$\mathbf{-0.063}$	$\mathbf{-0.278}$	$\mathbf{+0.117}$	$\mathbf{+0.073}$	$\mathbf{+0.034}$	$+0.201$

The frozen-path methods isolate node and weight choice by sharing one episode collector and one fit. Relative to DAgger-P's endpoint nodes, Gauss–Legendre, trapezoid, DAgger-L, and DAgger-G all have lower J on Hopper, HalfCheetah, and Walker2d after Holm correction, and DAgger-G also on Ant (Table 3). Return favours these alternatives only on Hopper and, for trapezoid, on Ant. Holding DAgger-P's nodes fixed, retraining lowers own-state action loss on all four environments and raises return on three, while also changing the number of fits and environment interactions.

In the original six-method comparison, Gauss–Legendre has lower J than DAgger-P on Hopper, HalfCheetah, and Walker2d, but the two procedures also differ in retraining, fits, and environment steps. Retrained DAgger-G has the lowest median J of the six on all four environments and, after Holm correction, lower J than both quadrature methods on all four (Appendix G). The main comparison therefore does not isolate node and weight choice. The pre-registered prediction that DAgger-P's median J/J_{bc} would increase monotonically by at least 20% from $N = 3$ to $N = 10$ on three environments held on none; it extrapolated the fixed-path limit in eqrefeq:dagger-p-endpoint to retrained DAgger.

7 Conclusion and Limitations

This paper shows that weighted dataset aggregation along a fixed expert–learner path can be interpreted as numerical quadrature for the path-average occupancy. The placement and weighting of collection rounds determine both the finite-n coverage error and the limiting occupancy as the number of rounds grows. For sufficiently smooth occupancy paths, higher-order quadrature rules can approximate the path average with few exact evaluations, while under finite sampling, simpler designs such as midpoint retain much of this benefit. Our experiments recover the predicted approximation hierarchy and endpoint limits under exact evaluation, and illustrate how schedule design interacts with sampling and retraining.

These results concern a specific coverage target and collection model. We define the path average using the uniform measure over the mixing coefficient, and the finite-sample analysis assumes independent samples from a finite state space. Extending the theory to trajectory data would require mixing or effective-sample-size arguments, while continuous state spaces would require a discrepancy defined over an appropriate loss class. A further limitation is the fixed-path assumption. Retraining changes the learner and therefore the subsequent collection path, so the fixed-path theory does not directly characterise retrained DAgger. Consistent with this distinction, extrapolating the fixed-path endpoint limit to retrained DAgger did not yield the pre-specified prediction on any of the four continuous-control environments. Nevertheless, the continuous-control experiments help disentangle schedule design from retraining. With retraining held fixed, frozen-path schedules containing interior mixtures reduced reference-path imitation loss on more environments than they improved return or own-state loss. Retraining further altered these outcomes, while the compared procedures also differ in expert participation, policy fitting, and environment interactions.

Reproducibility Statement

Appendix A contains complete proofs, and Appendix F lists every experimental setting, software version, seed, and result artifact. Every data-bearing figure and table is generated by scripts that read recorded run outputs, and the protocols, instance registries, seeds, and scripts (afterwards) are provided as supplementary material. Appendix E records each pre-registered criterion, every amendment, and every defect found. The re-runs of Appendix G reproduced 241,753 recorded values bit for bit.

Ethics Statement

This work uses simulated control environments and synthetic MDPs only and involves no human subjects or personal data.

LLM Usage Disclosure

We used Anthropic Claude (Opus 5) and OpenAI ChatGPT to assist with language polishing, mathematical checking and code debugging. All final claims, proofs, experiments, and interpretations were reviewed and approved by the authors.

References

T. W. Archibald, K. I. M. McKinnon, and L. C. Thomas. On the generation of Markov decision processes. *Journal of the Operational Research Society*, 46(3):354–361, 1995. doi: 10.1038/sj/jors/0460308.

Brenna D Argall, Sonia Chernova, Manuela Veloso, and Brett Browning. A survey of robot learning from demonstration. *Robotics and autonomous systems*, 57(5):469–483, 2009.

Shalabh Bhatnagar, Richard S. Sutton, Mohammad Ghavamzadeh, and Mark Lee. Natural actor–critic algorithms. Technical Report TR09-10, Department of Computing Science, University of Alberta, 2009.

Haoyuan Cai, Zhenghao Peng, and Bolei Zhou. Robot-gated interactive imitation learning with adaptive intervention mechanism. In *Proceedings of the 42nd International Conference on Machine Learning*, 2025.

Ben Calderhead and Mark Girolami. Estimating bayes factors via thermodynamic integration and population mcmc. *Computational Statistics & Data Analysis*, 53(12):4028–4045, 2009.

Ching-An Cheng and Byron Boots. Convergence of value aggregation for imitation learning. In *Proceedings of the 21st International Conference on Artificial Intelligence and Statistics*, volume 84 of *PMLR*, pp. 1801–1809, 2018.

Grace E. Cho and Carl D. Meyer. Comparison of perturbation bounds for the stationary distribution of a Markov chain. *Linear Algebra and its Applications*, 335:137–150, 2001. doi: 10.1016/S0024-3795(01)00320-2.

Kamil Ciosek and Shimon Whiteson. Expected policy gradients for reinforcement learning. *Journal of Machine Learning Research*, 21(52):1–51, 2020.

Philip J Davis and Philip Rabinowitz. *Methods of numerical integration*. Dover Publications, second edition, 2007.

Nial Friel, Merrilee Hurn, and Jason Wyse. Improving power posterior estimation of statistical evidence. *Statistics and Computing*, 24(5):709–723, 2014.

Andrew Gelman and Xiao-Li Meng. Simulating normalizing constants: From importance sampling to bridge sampling to path sampling. *Statistical Science*, 13(2):163–185, 1998. doi: 10.1214/ss/1028905934.

Ryan Hoque, Ashwin Balakrishna, Ellen Novoseller, Albert Wilcox, Daniel S. Brown, and Ken Goldberg. ThriftyDAgger: Budget-aware novelty and risk gating for interactive imitation learning. In *Proceedings of the 5th Conference on Robot Learning*, volume 164 of *PMLR*, pp. 598–608, 2021.

Sabine Hug, Michael Schwarzfischer, Jan Hasenauer, Carsten Marr, and Fabian J Theis. An adaptive scheduling scheme for calculating bayes factors with thermodynamic integration using simpson's rule. *Statistics and Computing*, 26(3):663–677, 2016.

Sham Kakade and John Langford. Approximately optimal approximate reinforcement learning. In *Proceedings of the 19th International Conference on Machine Learning*, pp. 267–274, 2002.

Michael Kelly, Chelsea Sidrane, Katherine Driggs-Campbell, and Mykel J. Kochenderfer. HG-DAgger: Interactive imitation learning with human experts. In *IEEE International Conference on Robotics and Automation*, pp. 8077–8083, 2019. doi: 10.1109/ICRA.2019.8793698.

Kunal Menda, Katherine Driggs-Campbell, and Mykel J. Kochenderfer. EnsembleDAgger: A Bayesian approach to safe imitation learning. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2019. doi: 10.1109/IROS40897.2019.8968287.

Takayuki Osa, Joni Pajarinen, Gerhard Neumann, J Andrew Bagnell, Pieter Abbeel, and Jan Peters. An algorithmic perspective on imitation learning. *Foundations and Trends® in Robotics*, 7(1-2): 1–179, 2018.

Nived Rajaraman, Lin F. Yang, Jiantao Jiao, and Kannan Ramchandran. Toward the fundamental limits of imitation learning. In *Advances in Neural Information Processing Systems*, volume 33, pp. 2914–2924, 2020.

Stéphane Ross and Drew Bagnell. Efficient reductions for imitation learning. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pp. 661–668. JMLR Workshop and Conference Proceedings, 2010.

Stéphane Ross and J. Andrew Bagnell. Reinforcement and imitation learning via interactive no-regret learning. *arXiv preprint arXiv:1406.5979*, 2014.

Stéphane Ross, Geoffrey Gordon, and Drew Bagnell. A reduction of imitation learning and structured prediction to no-regret online learning. In *Proceedings of the fourteenth international conference on artificial intelligence and statistics*, pp. 627–635. JMLR Workshop and Conference Proceedings, 2011.

Paul J. Schweitzer. Perturbation theory and finite Markov chains. *Journal of Applied Probability*, 5 (2):401–413, 1968. doi: 10.2307/3212261.

Wen Sun, Arun Venkatraman, Geoffrey J. Gordon, Byron Boots, and J. Andrew Bagnell. Deeply AggreVaTeD: Differentiable imitation learning for sequential prediction. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *PMLR*, pp. 3309–3318, 2017.

Gokul Swamy, Sanjiban Choudhury, J. Andrew Bagnell, and Zhiwei Steven Wu. Of moments and matching: A game-theoretic framework for closing the imitation gap. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *PMLR*, pp. 10022–10032, 2021.

Emanuel Todorov, Tom Erez, and Yuval Tassa. MuJoCo: A physics engine for model-based control. In *2012 IEEE/RSJ international conference on intelligent robots and systems*, pp. 5026–5033. IEEE, 2012.

Lloyd N. Trefethen. Is Gauss quadrature better than Clenshaw–Curtis? *SIAM Review*, 50(1):67–87, 2008. doi: 10.1137/060659831.

Jiakai Zhang and Kyunghyun Cho. Query-efficient imitation learning for end-to-end simulated driving. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*, 2017.

A Proofs and Technical Details

This appendix supplies the proofs for Sections 3–5. The order follows the logical dependencies: basic measure conventions, the coverage result, the schedule identity, occupancy-path regularity, the common quadrature machinery, approximation rates, schedule limits, and finally finite-sample estimation.

A.1 Measure, norm, and integration conventions

Let $d := |\mathcal{S}| < \infty$ and identify each state distribution with a vector in $\Delta(\mathcal{S}) \subset \mathbb{R}^d$. Our ambient Banach space is

$$X := (\mathbb{R}^d, \|\cdot\|_1). \tag{14}$$

For probability vectors $\mu, \nu \in \Delta(\mathcal{S})$,

$$d_{\mathrm{TV}}(\mu, \nu) := \frac{1}{2}\|\mu - \nu\|_1. \tag{15}$$

For signed vectors we work directly with $\|\cdot\|_1$. When a complex extension is used below, the same notation denotes $\|z\|_1 := \sum_{j=1}^d |z_j|$ on $\mathbb{C}^d$.

Because X is finite dimensional, every measurable bounded map $\alpha \mapsto \rho_\alpha \in \Delta(\mathcal{S})$ is Bochner integrable, and its Bochner integral agrees with coordinatewise integration. Hence for $\eta \in \mathcal{P}([0,1])$,

$$Q_\eta[\rho] := \int_{[0,1]} \rho_\alpha \, \eta \mathrm{d}\alpha \tag{16}$$

is again a probability vector. In particular, $\bar{\rho} = Q_\lambda[\rho]$ for $\lambda(\mathrm{d}\alpha) = \mathrm{d}\alpha$.

Lemma 1 (Barycentre inequality)**.** *For any $\eta \in \mathcal{P}([0,1])$, measurable occupancy path $\alpha \mapsto \rho_\alpha$, and $\mu \in \Delta(\mathcal{S})$,*

$$d_{\mathrm{TV}}(\mu, Q_\eta[\rho]) \le \int_{[0,1]} d_{\mathrm{TV}}(\mu, \rho_\alpha)\, \eta(d\alpha). \tag{17}$$

Proof. Linearity of the integral gives

$$\mu - Q_\eta[\rho] = \int_{[0,1]} (\mu - \rho_\alpha)\, \eta(\mathrm{d}\alpha).$$

Applying the integral triangle inequality in X and then 15 yields

$$d_{\mathrm{TV}}(\mu, Q_\eta[\rho]) \le \int_{[0,1]} d_{\mathrm{TV}}(\mu, \rho_\alpha)\, \eta(\mathrm{d}\alpha).$$

□

A.2 Coverage properties of the path average

Proof of Theorem 1. Fix $\alpha \in [0,1]$. Since $\bar{\rho} = \int_0^1 \rho_\beta \, \mathrm{d}\beta$, the TV-Lipschitz assumption gives

$$\begin{aligned} d_{\mathrm{TV}}(\bar{\rho}, \rho_\alpha) &\le \int_0^1 d_{\mathrm{TV}}(\rho_\beta, \rho_\alpha)\, \mathrm{d}\beta \\ &\le L \int_0^1 |\beta - \alpha|\, \mathrm{d}\beta \\ &= \frac{L}{2}(\alpha^2 + (1-\alpha)^2) \\ &\le \frac{L}{2}. \end{aligned} \tag{18}$$

Taking the supremum over α and using $d_{\mathrm{TV}} \le 1$ on probability distributions proves

$$r(\bar{\rho}) \le \min\left\{1, \frac{L}{2}\right\}.$$

For the comparison with the best TV centre, take any $\mu \in \Delta(\mathcal{S})$. Lemma 1 with $\eta = \lambda$ gives

$$d_{\mathrm{TV}}(\bar{\rho}, \mu) \le \int_0^1 d_{\mathrm{TV}}(\rho_\beta, \mu)\, \mathrm{d}\beta \le r(\mu). \tag{19}$$

Thus, for every α,

$$d_{\mathrm{TV}}(\bar{\rho}, \rho_\alpha) \le d_{\mathrm{TV}}(\bar{\rho}, \mu) + d_{\mathrm{TV}}(\mu, \rho_\alpha) \le 2r(\mu). \tag{20}$$

Taking first the supremum over α and then the infimum over μ yields $r(\bar{\rho}) \le 2r_{\mathrm{Cheb}}$.

It remains to prove the squared-ℓ_2 characterisation. Expanding around $\bar{\rho}$ gives, for any $\mu \in \Delta(\mathcal{S})$,

$$\int_0^1 \|\mu - \rho_\alpha\|_2^2\, \mathrm{d}\alpha = \|\mu - \bar{\rho}\|_2^2 + \int_0^1 \|\bar{\rho} - \rho_\alpha\|_2^2\, \mathrm{d}\alpha, \tag{21}$$

where the cross term vanishes because $\int_0^1 (\bar{\rho} - \rho_\alpha)\, \mathrm{d}\alpha = 0$. The second term in (21) does not depend on μ, and the first vanishes only when $\mu = \bar{\rho}$. Hence $\bar{\rho}$ is the unique squared-ℓ_2 Fréchet mean of the path. □

Bounded-loss certificate. For probability vectors μ, ν and any $g : \mathcal{S} \to [0, 1]$,

$$\begin{aligned} |\mathbb{E}_\mu[g] - \mathbb{E}_\nu[g]| &= |\sum_{s \in \mathcal{S}} (g(s) - \tfrac{1}{2})(\mu(s) - \nu(s))| \\ &\le \frac{1}{2} \|\mu - \nu\|_1 = d_{\mathrm{TV}}(\mu, \nu). \end{aligned} \tag{22}$$

Substituting $\mu = \bar{\rho}$ and $\nu = \rho_\alpha$, then taking the supremum over α, proves (4). Nothing in this argument concerns optimisation over policies. It is only an expectation-shift bound for a fixed bounded function.

A.3 Support coverage of the discounted path

Proof of Proposition 2. By (26), $\rho_\alpha(s) = (1 - \gamma) \sum_{t \ge 0} \gamma^t \Pr_\alpha(s_t = s)$, and $\Pr_\alpha(s_t = s)$ is the sum, over state sequences $(s_0, \dots, s_t)$ with $s_t = s$, of $\mu_0(s_0) \prod_{k<t} P_\alpha(s_k, s_{k+1})$. By (24), each factor $(1 - \alpha)P_0(s_k, s_{k+1}) + \alpha P_1(s_k, s_{k+1})$ is either zero on $[0, 1]$, when both entries vanish, or strictly positive on $(0, 1)$. Products of such factors and non-negative sums of such products inherit this dichotomy, so each map $\alpha \mapsto \rho_\alpha(s)$ is either identically zero on $(0, 1)$ or strictly positive there. Hence $\operatorname{supp} \rho_\alpha = \mathcal{S}^\circ$ for all $\alpha \in (0, 1)$. If $\rho_0(s) > 0$, some sequence has every factor positive at $\alpha = 0$, so every factor has $P_0(s_k, s_{k+1}) > 0$ and is positive on $(0, 1)$, so $s \in \mathcal{S}^\circ$. The same argument applies at $\alpha = 1$.

If $\eta((0, 1)) > 0$, then for $s \in \mathcal{S}^\circ$, $\bar{\rho}_\eta(s) \ge \int_{(0,1)} \rho_\alpha(s)\, \eta(\mathrm{d}\alpha) > 0$ because the integrand is positive on $(0, 1)$. Every ρ_α is supported in the finite set $\mathcal{S}^\circ$ and bounded by one, so $C(\eta) \le 1/\min_{s \in \mathcal{S}^\circ} \bar{\rho}_\eta(s) < \infty$. If instead η is supported on $\{0, 1\}$ and some $s \in \mathcal{S}^\circ \setminus (\operatorname{supp} \rho_0 \cup \operatorname{supp} \rho_1)$ exists, then $\bar{\rho}_\eta(s) = 0 < \rho_{1/2}(s)$ and $C(\eta) = \infty$. □

Concentrability can be exponential in the horizon. Take states $\{0, \dots, K - 1\}$, $\mu_0 = \delta_0$, an expert that moves from k to $\min(k + 1, K - 1)$, and a learner that moves to 0. For $1 \le k \le K - 2$, the event $s_t = k$ occurs exactly when $s_{t-k} = 0$ and the next k actions are expert actions, so $\Pr_\alpha(s_t = k) = \alpha^k \Pr_\alpha(s_{t-k} = 0)$ and $\rho_\alpha(k) = (\gamma\alpha)^k \rho_\alpha(0)$. Since $1 - \gamma \le \rho_\alpha(0) \le 1$, for $\alpha > \beta$, $\rho_\alpha(k)/\rho_\beta(k) \ge (1 - \gamma)(\alpha/\beta)^k$, which grows geometrically in k.

A.4 Schedule measures and weighted aggregation

Proof of Proposition 3. Let $\nu_n = \sum_{i=1}^n w_i \delta_{\alpha_i}$. For every state $s \in \mathcal{S}$,

$$\begin{aligned} Q_{\nu_n}[\rho](s) &= \int_{[0,1]} \rho_\alpha(s)\, \nu_n(\mathrm{d}\alpha) \\ &= \sum_{i=1}^n w_i \rho_{\alpha_i}(s). \end{aligned}$$

Therefore

$$Q_{\nu_n}[\rho] = \sum_{i=1}^{n} w_i \rho_{\alpha_i}. \tag{23}$$

Since $Q_\lambda[\rho] = \bar{\rho}$, the quantity $d_{\mathrm{TV}}(Q_{\nu_n}[\rho], \bar{\rho})$ is precisely the quadrature error of the node–weight design applied to the occupancy-valued map $\alpha \mapsto \rho_\alpha$. □

Equation 23 is a population-level identity. Finite-data estimation and any mismatch between intended and realised aggregation weights are handled separately in Section A.11.

A.5 Discounted occupancy: resolvent regularity

Set

$$P_0 := P_{\hat{\pi}_0}, \qquad P_1 := P_{\pi^\star}, \qquad \Delta P := P_1 - P_0.$$

Linearity of the induced kernel in the policy gives

$$P_\alpha = P_0 + \alpha \Delta P. \tag{24}$$

For $\gamma < 1$, define

$$R_\gamma(\alpha) := (I - \gamma P_\alpha^\top)^{-1}. \tag{25}$$

Because P_α is stochastic, its spectral radius satisfies $\mathrm{spr}(\gamma P_\alpha^\top) = \gamma < 1$, so the inverse exists and

$$R_\gamma(\alpha) = \sum_{t=0}^{\infty} \gamma^t (P_\alpha^\top)^t, \qquad \rho_\alpha^\gamma = (1-\gamma) R_\gamma(\alpha) \mu_0. \tag{26}$$

This is 8.

Lemma 2 (Discounted derivative formula)**.** *For every integer $k \geq 1$,*

$$R_\gamma^{(k)}(\alpha) = k! \gamma^k \big(R_\gamma(\alpha) \Delta P^\top\big)^k R_\gamma(\alpha), \tag{27}$$

$$\frac{d^k}{d\alpha^k} \rho_\alpha^\gamma = k! \gamma^k \big(R_\gamma(\alpha) \Delta P^\top\big)^k \rho_\alpha^\gamma. \tag{28}$$

Moreover,

$$\sup_{\alpha \in [0,1]} \left\| \frac{d^k}{d\alpha^k} \rho_\alpha^\gamma \right\|_1 \leq k! \left(\frac{\gamma}{1-\gamma} \|\Delta P\|_\infty \right)^k. \tag{29}$$

Proof. Differentiate $R_\gamma(\alpha)(I - \gamma P_\alpha^\top) = I$. Since $P_\alpha' = \Delta P$,

$$R_\gamma'(\alpha) = \gamma R_\gamma(\alpha) \Delta P^\top R_\gamma(\alpha).$$

Repeated differentiation preserves the order of the factors, and induction gives 27. Multiplication by $(1-\gamma)\mu_0$ gives 28.

The Neumann series also gives $\|R_\gamma(\alpha)\|_{1\to 1} \leq (1-\gamma)^{-1}$. Together with $\|\Delta P^\top\|_{1\to 1} = \|\Delta P\|_\infty$ and $\|\rho_\alpha^\gamma\|_1 = 1$, submultiplicativity yields 29. □

Holomorphic extension. For $z \in \mathbb{C}$, let $P_z := P_0 + z\Delta P$ and

$$D_\gamma(z) := \det(I - \gamma P_z^\top).$$

This determinant is a polynomial in z. It is nonzero on $[0,1]$, because $I - \gamma P_\alpha^\top$ is invertible there, so it is not the zero polynomial. Its zero set is therefore finite and has positive distance from the compact interval $[0,1]$. On a complex neighbourhood avoiding those zeros,

$$\rho^\gamma(z) := (1-\gamma)(I - \gamma P_z^\top)^{-1} \mu_0 \tag{30}$$

is holomorphic and agrees with ρ_α^γ on the real interval. This proves the discounted clause of Theorem 4.

A.6 STATIONARY OCCUPANCY: AUGMENTED-SYSTEM REGULARITY

Assume (SU) from Section 5.1, that each P_α, $\alpha \in [0,1]$, has a unique stationary distribution. For a finite stochastic matrix, this is equivalent to $\dim\ker(I - P_\alpha^\top) = 1$. Define

$$A_\alpha := I - P_\alpha^\top + \mathbf{1}\mathbf{1}^\top. \tag{31}$$

Lemma 3 (Invertibility of the augmented stationary system)**.** *Under* (SU)*, A_α is invertible for every $\alpha \in [0,1]$ and*

$$\rho_\alpha^\infty = A_\alpha^{-1}\mathbf{1}. \tag{32}$$

Proof. Stationarity and normalisation give $P_\alpha^\top \rho_\alpha^\infty = \rho_\alpha^\infty$ and $\mathbf{1}^\top \rho_\alpha^\infty = 1$, hence $A_\alpha \rho_\alpha^\infty = \mathbf{1}$.

To show uniqueness of this solution, suppose $A_\alpha x = 0$. Left multiplication by $\mathbf{1}^\top$, using $\mathbf{1}^\top(I - P_\alpha^\top) = 0$, gives $d\,\mathbf{1}^\top x = 0$. Thus $\mathbf{1}^\top x = 0$ and $(I - P_\alpha^\top)x = 0$. By (SU), this nullspace is spanned by ρ_α^∞, so $x = c\rho_\alpha^\infty$. The condition $\mathbf{1}^\top x = 0$ then forces $c = 0$. Hence A_α is nonsingular and 32 follows. □

Write $G(\alpha) := A_\alpha^{-1}$. Since $A_\alpha = A_0 - \alpha\Delta P^\top$,

$$G'(\alpha) = G(\alpha)\Delta P^\top G(\alpha). \tag{33}$$

The same product-rule induction as above gives

$$G^{(k)}(\alpha) = k!\big(G(\alpha)\Delta P^\top\big)^k G(\alpha), \tag{34}$$

$$\frac{d^k}{d\alpha^k}\rho_\alpha^\infty = k!\big(G(\alpha)\Delta P^\top\big)^k \rho_\alpha^\infty. \tag{35}$$

Matrix inversion is continuous on the set of nonsingular matrices. Since $\alpha \mapsto A_\alpha$ is continuous and $[0,1]$ is compact,

$$M_\infty := \sup_{\alpha\in[0,1]} \|A_\alpha^{-1}\|_{1\to 1} < \infty. \tag{36}$$

Consequently,

$$\sup_{\alpha\in[0,1]} \left\| \frac{d^k}{d\alpha^k}\rho_\alpha^\infty \right\|_1 \le k!\big(M_\infty \|\Delta P\|_\infty\big)^k. \tag{37}$$

Holomorphic extension. For $z \in \mathbb{C}$, define

$$A(z) := I - (P_0 + z\Delta P)^\top + \mathbf{1}\mathbf{1}^\top.$$

The polynomial $\det A(z)$ is nonzero on $[0,1]$ by Lemma 3. Hence its zero set is finite and separated from that compact interval. On the resulting complex neighbourhood,

$$\rho^\infty(z) := A(z)^{-1}\mathbf{1} \tag{38}$$

is holomorphic and restricts to ρ_α^∞ on $[0,1]$. This proves the stationary clause of Theorem 4.

Truncated occupancies. By 7, $(P_\alpha^\top)^t$ is a product of t matrices affine in α, so $\rho_\alpha^H = H^{-1}\sum_{t<H}(P_\alpha^\top)^t\mu_0$ is a polynomial of degree at most $H-1$, which is the last clause of Theorem 4. Its k-th derivative replaces k of the t factors by $\Delta P^\top$ in $t!/(t-k)!$ ordered ways, and the remaining factors have $\|\cdot\|_{1\to1} = 1$, so $\|\partial_\alpha^k \rho_\alpha^H\|_1 \le (H-1)^k\|\Delta P\|_\infty^k$. Resetting to μ_0 on termination is a transition kernel that is linear in the policy, so it preserves 7.

Regularity is not mixing. Aperiodicity is not used in this argument. The result concerns how the stationary solution varies with α. Estimating that solution from a finite trajectory is a different question and requires an appropriate mixing or effective-sample-size assumption. Likewise, action support at the endpoint policies does not by itself imply (SU), and the induced chains must also have the requisite environmental connectivity.

A.7 ℓ_1-VALUED QUADRATURE AND NODE ACCOUNTING

The rate proofs all use the same scalar-to-vector reduction.

Lemma 4 (ℓ_1-valued quadrature lifting)**.** *Let $Q_n[f] = \sum_{i=1}^n w_i f(\alpha_i)$ be a scalar quadrature functional. For continuous $\rho : [0,1] \to \mathbb{R}^d$,*

$$\left\| Q_n[\rho] - \int_0^1 \rho_\alpha \, d\alpha \right\|_1 = \sup_{\|u\|_\infty \le 1} \left| Q_n[u^\top \rho] - \int_0^1 u^\top \rho_\alpha \, d\alpha \right|. \tag{39}$$

If $\rho \in C^k$, then

$$\sup_{\|u\|_\infty \le 1} \sup_{\alpha \in [0,1]} \left| \frac{d^k}{d\alpha^k} (u^\top \rho_\alpha) \right| \le \sup_{\alpha \in [0,1]} \|\rho_\alpha^{(k)}\|_1. \tag{40}$$

Proof. The first identity is the usual ℓ_1/ℓ_∞ duality formula applied to the quadrature-error vector. The derivative bound follows from $|u^\top \rho_\alpha^{(k)}| \le \|u\|_\infty \|\rho_\alpha^{(k)}\|_1$. □

All deterministic rules below use nonnegative weights summing to one. Their aggregates and $\bar{\rho}$ are therefore probability vectors, so an ℓ_1 error bound converts to TV by a factor $1/2$.

Node conventions. Throughout this appendix, n is the number of distinct exact occupancy evaluations. The schedule measures are

$$\nu_n^{\mathrm{ER}} := \frac{1}{n} \sum_{i=0}^{n-1} \delta_{i/n}, \tag{41}$$

$$\nu_n^{\mathrm{mid}} := \frac{1}{n} \sum_{i=1}^{n} \delta_{(i-1/2)/n}, \tag{42}$$

and, for $n \ge 2$,

$$\nu_n^{\mathrm{trap}} := \frac{1}{2(n-1)} \delta_0 + \frac{1}{n-1} \sum_{i=2}^{n-1} \delta_{(i-1)/(n-1)} + \frac{1}{2(n-1)} \delta_1. \tag{43}$$

For composite Simpson, $n \ge 3$ is odd, $m := n-1$ is even, $\alpha_i = (i-1)/m$, and

$$w_i^{\mathrm{Simp}} := \frac{1}{3m} \begin{cases} 1, & i \in \{1, n\}, \\ 4, & i \text{ even}, \\ 2, & i \text{ odd and } 1 < i < n. \end{cases} \tag{44}$$

Thus $\nu_n^{\mathrm{Simp}} := \sum_{i=1}^n w_i^{\mathrm{Simp}} \delta_{\alpha_i}$. Finally, if $(x_i^{\mathrm{GL}}, \omega_i^{\mathrm{GL}})_{i=1}^n$ are the standard Gauss–Legendre nodes and weights on $[-1,1]$, then

$$\alpha_i^{\mathrm{GL}} := \frac{1 + x_i^{\mathrm{GL}}}{2}, \qquad w_i^{\mathrm{GL}} := \frac{\omega_i^{\mathrm{GL}}}{2}, \qquad \nu_n^{\mathrm{GL}} := \sum_{i=1}^n w_i^{\mathrm{GL}} \delta_{\alpha_i^{\mathrm{GL}}}. \tag{45}$$

The endpoint rule may be reversed without changing its order. The trapezoid rule uses $n-1$ subintervals, Simpson uses $n-1$ even subintervals, and Gauss–Legendre uses exactly n nodes.

A.8 SCHEDULE-DEPENDENT APPROXIMATION RATES

We now prove Theorem 5. Whenever the derivative exists, write

$$M_k := \sup_{\alpha \in [0,1]} \|\rho_\alpha^{(k)}\|_1. \tag{46}$$

Random-uniform nodes. Let $\alpha_1, \dots, \alpha_n \overset{\text{iid}}{\sim} \text{Unif}[0,1]$ and $X_i := \rho_{\alpha_i}$. Then $\mathbb{E}[X_i] = \bar{\rho}$. For each state s, $0 \le X_i(s) \le 1$, hence $\text{Var}(X_i(s)) \le \mathbb{E}[X_i(s)] = \bar{\rho}(s)$. Jensen's inequality, followed by Cauchy–Schwarz over the d coordinates, gives

$$\mathbb{E}\left\| \frac{1}{n}\sum_{i=1}^{n} X_i - \bar{\rho} \right\|_1 \le \frac{1}{\sqrt{n}} \sum_{s \in \mathcal{S}} \sqrt{\bar{\rho}(s)} \le \sqrt{\frac{d}{n}}. \tag{47}$$

Therefore

$$\mathbb{E}\, \mathcal{E}(\nu_n^{\text{RU}}) \le \frac{1}{2}\sqrt{\frac{d}{n}}. \tag{48}$$

If one X_i is replaced, the empirical mean changes by TV distance at most $1/n$. McDiarmid's inequality therefore yields, with probability at least $1 - \delta$,

$$\mathcal{E}(\nu_n^{\text{RU}}) \le \frac{1}{2}\sqrt{\frac{d}{n}} + \sqrt{\frac{\log(1/\delta)}{2n}}. \tag{49}$$

This proves the $O_{\mathbb{P}}(n^{-1/2})$ row. Notice that no path smoothness is used.

Endpoint Riemann rule. Assume $d_{\text{TV}}(\rho_\alpha, \rho_\beta) \le L|\alpha - \beta|$. With $I_i = [i/n, (i+1)/n]$,

$$\begin{aligned} \mathcal{E}(\nu_n^{\text{ER}}) &\le \sum_{i=0}^{n-1} \int_{I_i} d_{\text{TV}}(\rho_\alpha, \rho_{i/n})\, d\alpha \\ &\le L \sum_{i=0}^{n-1} \int_{i/n}^{(i+1)/n} (\alpha - i/n)\, d\alpha \\ &= \frac{L}{2n}. \end{aligned} \tag{50}$$

Thus the endpoint rule is first order.

Midpoint, trapezoid, and Simpson rules. For scalar f on $[0,1]$, the standard composite error bounds are

$$\left| Q_n^{\text{mid}}[f] - \int_0^1 f \right| \le \frac{\|f''\|_\infty}{24n^2}, \tag{51}$$

$$\left| Q_n^{\text{trap}}[f] - \int_0^1 f \right| \le \frac{\|f''\|_\infty}{12(n-1)^2}, \tag{52}$$

for $f \in C^2$, and, for $f \in C^4$ with odd n,

$$\left| Q_n^{\text{Simp}}[f] - \int_0^1 f \right| \le \frac{\|f^{(4)}\|_\infty}{180(n-1)^4}. \tag{53}$$

Lemma 4, followed by the factor $1/2$ converting ℓ_1 error to TV, gives

$$\mathcal{E}(\nu_n^{\text{mid}}) \le \frac{M_2}{48n^2}, \tag{54}$$

$$\mathcal{E}(\nu_n^{\text{trap}}) \le \frac{M_2}{24(n-1)^2}, \tag{55}$$

$$\mathcal{E}(\nu_n^{\text{Simp}}) \le \frac{M_4}{360(n-1)^4}. \tag{56}$$

These are second, second, and fourth order under the common evaluation budget.

Gauss–Legendre rule. Let $\mathcal{B}_\varrho$ be a shifted Bernstein ellipse for $[0,1]$ with $\varrho > 1$, contained in the domain of a holomorphic extension $z \mapsto \rho(z)$, and set

$$M_\varrho := \sup_{z \in \mathcal{B}_\varrho} \|\rho(z)\|_1 < \infty. \tag{57}$$

For fixed real u with $\|u\|_\infty \le 1$, the scalar function $f_u(z) := u^\top \rho(z)$ is holomorphic on the same ellipse and $|f_u(z)| \le M_\varrho$. The standard Chebyshev truncation bound therefore provides, for each $m \ge 0$, a polynomial p_m of degree at most m with

$$\|f_u - p_m\|_{\infty,[0,1]} \le \frac{2M_\varrho}{\varrho - 1}\, \varrho^{-m}. \tag{58}$$

The n-point Gauss–Legendre rule is exact for polynomials of degree at most $2n-1$ and its mapped weights are positive and sum to one. Taking $m = 2n-1$ gives

$$\begin{aligned}\left| Q_n^{\mathrm{GL}}[f_u] - \int_0^1 f_u(\alpha)\, d\alpha \right| &\le 2\|f_u - p_m\|_\infty \\ &\le \frac{4M_\varrho \varrho}{\varrho - 1}\, \varrho^{-2n}.\end{aligned} \tag{59}$$

Taking the supremum over $\|u\|_\infty \le 1$ in Lemma 4, then converting to TV, yields

$$\mathcal{E}(\nu_n^{\mathrm{GL}}) \le C_\varrho \varrho^{-2n}, \qquad C_\varrho := \frac{2M_\varrho \varrho}{\varrho - 1}. \tag{60}$$

The constant worsens as the admissible ellipse approaches a complex singularity. Corollary 6 gives an explicit admissible ellipse.

Together, 49, 50, 54–56, and 60 prove Theorem 5. Theorem 4 supplies the required smoothness for the discounted path and, under (SU), the stationary path.

A.9 Horizon-explicit constants and leading error terms

Proof of Corollary 6. Let $K_\alpha := \gamma R_\gamma(\alpha)\Delta P^\top$ in the discounted case and $K_\alpha := G(\alpha)\Delta P^\top$ in the stationary case. The truncated case is treated at the end of each step. Then 28 and 35 read $\partial_\alpha^k \rho_\alpha = k!\, K_\alpha^k \rho_\alpha$, with $\|K_\alpha\|_{1\to 1} \le \kappa \|\Delta P\|_\infty$. Row s of ΔP is the mixture of $P(\cdot \mid s,a)$ under the signed measure $\pi^\star(\cdot \mid s) - \hat\pi_0(\cdot \mid s)$, so its ℓ_1 norm is at most $2d_{\mathrm{TV}}(\pi^\star(\cdot \mid s), \hat\pi_0(\cdot \mid s))$ and $\|\Delta P\|_\infty \le 2\varepsilon_\pi$. This gives $M_k \le k!(2\kappa\varepsilon_\pi)^k$ and $L \le \frac{1}{2}M_1 \le \kappa\varepsilon_\pi$, and the truncated bound above gives the same with $\kappa = H-1$. Substituting into 50 and 54 gives the endpoint and midpoint bounds.

For Gauss–Legendre, set $r := 1/(2\kappa\varepsilon_\pi)$. For real $\alpha \in [0,1]$ and complex h with $|h| \le r/2$, the factorizations $I - \gamma P_{\alpha+h}^\top = (I - \gamma P_\alpha^\top)(I - hK_\alpha)$ and $A(\alpha+h) = A_\alpha(I - hK_\alpha)$, together with $\|hK_\alpha\|_{1\to 1} \le \frac{1}{2}$, show that the matrix in 30, respectively 38, is invertible at $\alpha+h$ and that $\rho(\alpha+h) = \sum_{k\ge 0}(hK_\alpha)^k \rho_\alpha$, so $\|\rho(\alpha+h)\|_1 \le 2$. For the truncated occupancy, which is a polynomial, Taylor's formula and the derivative bound give $\|\rho^H(\alpha+h)\|_1 \le \exp\big(2(H-1)\varepsilon_\pi|h|\big) \le e^{1/2} < 2$ for the same h. Let $\mathcal{B}_\varrho$ be the shifted Bernstein ellipse for $[0,1]$ with semi-minor axis $b = r/2$. Its semi-major axis is $a = \frac{1}{2}\sqrt{1+r^2} \le \frac{1}{2} + b$, its parameter is $\varrho = 2(a+b) = r + \sqrt{1+r^2} \ge 1 + r$, and a short computation shows that every point of $\mathcal{B}_\varrho$ lies within distance b of $[0,1]$. Hence $M_\varrho \le 2$, and 60 gives $\mathcal{E}(\nu_n^{\mathrm{GL}}) \le 4\varrho^{1-2n}/(\varrho-1)$, which decreases in ϱ for $n \ge 1$. With $\varrho \ge 1 + r$, $\mathcal{E}(\nu_n^{\mathrm{GL}}) \le 4\frac{1+r}{r}(1+r)^{-2n} = 4(1+2\kappa\varepsilon_\pi)(1+r)^{-2n}$. Finally, $\log(1+r) \ge r/(1+r)$ shows that $n \ge \frac{1+r}{2r}\log\frac{4(1+r)}{r\epsilon}$ suffices for error at most ϵ, which is the stated condition. □

Proof of Proposition 7. Applying the Euler–Maclaurin expansions (Davis & Rabinowitz, 2007) coordinatewise, which is legitimate because d is finite. With $h = 1/n$, the endpoint rule equals the trapezoid rule on the grid $\{0, h, \dots, 1\}$ minus $\frac{h}{2}(\rho_1 - \rho_0)$, and that trapezoid rule has error $O(h^2)$ for C^2 integrands, so $Q_{\nu_n^{\mathrm{ER}}}[\rho] - \bar\rho = -\frac{1}{2n}(\rho_1 - \rho_0) + O(n^{-2})$. For C^4 integrands, $Q_{\nu_n^{\mathrm{mid}}}[\rho] - \bar\rho = -\frac{h^2}{24}(\rho_1' - \rho_0') + O(h^4)$ with $h = 1/n$, and $Q_{\nu_n^{\mathrm{trap}}}[\rho] - \bar\rho = \frac{h^2}{12}(\rho_1' - \rho_0') + O(h^4)$ with $h = 1/(n-1)$. By the triangle inequality, $\|v + w\|_1 = \|v\|_1 + O(\|w\|_1)$, and taking ℓ_1 norms and halving gives the three expansions. The remaining statements follow from whether the leading coefficient vanishes. □

A.10 Schedule limits, consistency, and endpoint concentration

Lemma 5 (Continuity of occupancy barycentres)**.** *If $\alpha \mapsto \rho_\alpha$ is TV-continuous and $\nu_n \Rightarrow \nu$ weakly in $\mathcal{P}([0,1])$, then*

$$\|Q_{\nu_n}[\rho] - Q_\nu[\rho]\|_1 \to 0. \tag{61}$$

Proof. For each $s \in \mathcal{S}$, the coordinate map $f_s(\alpha) := \rho_\alpha(s)$ is bounded and continuous, since

$$|f_s(\alpha) - f_s(\beta)| \le \|\rho_\alpha - \rho_\beta\|_1 = 2d_{\mathrm{TV}}(\rho_\alpha, \rho_\beta).$$

Weak convergence therefore gives $\int f_s \, d\nu_n \to \int f_s \, d\nu$ for every s. There are only finitely many coordinates, so the convergence also holds in ℓ_1; this is the only place finiteness of $\mathcal{S}$ enters the argument. □

Proof of Proposition 8. By Lemma 5,

$$d_{\mathrm{TV}}(Q_{\nu_n}[\rho], Q_\nu[\rho]) = \frac{1}{2}\|Q_{\nu_n}[\rho] - Q_\nu[\rho]\|_1 \to 0.$$

Taking $\nu = \lambda$ gives the uniform-target consistency statement. If instead $\nu = \delta_a$, then $Q_{\nu_n}[\rho] \to Q_{\delta_a}[\rho] = \rho_a$. The reverse triangle inequality gives

$$|\mathcal{E}(\nu_n) - d_{\mathrm{TV}}(\rho_a, \bar{\rho})| \le d_{\mathrm{TV}}(Q_{\nu_n}[\rho], \rho_a) \to 0,$$

which proves (9) and the strictly positive limit when $\rho_a \neq \bar{\rho}$. □

DAgger-P. For the equal-round endpoint schedule used in the main text,

$$\nu_n^{\mathrm{P}} = \frac{n-1}{n}\delta_0 + \frac{1}{n}\delta_1.$$

For every bounded continuous f,

$$\int f \, d\nu_n^{\mathrm{P}} = \frac{n-1}{n} f(0) + \frac{1}{n} f(1) \to f(0),$$

so $\nu_n^{\mathrm{P}} \Rightarrow \delta_0$. Proposition 8 then gives (10). This is a statement about limiting aggregate coverage, not about monotonic policy-return behaviour as n increases.

Consistency of the canonical designs. For endpoint Riemann, midpoint, trapezoid, and Simpson, weak convergence to λ follows from convergence of the corresponding quadrature sums on continuous test functions. For Gauss–Legendre, take any $f \in C([0,1])$ and choose a polynomial p with $\|f - p\|_\infty < \varepsilon$. For all sufficiently large n, the n-point rule integrates p exactly. Since the Gauss–Legendre weights are positive and normalised,

$$\left| \int f \, d\nu_n^{\mathrm{GL}} - \int f \, d\lambda \right| \le 2\varepsilon$$

for all such n. Hence $\nu_n^{\mathrm{GL}} \Rightarrow \lambda$.

For random-uniform scheduling, the empirical-measure strong law gives

$$\nu_n^{\mathrm{RU}} \Rightarrow \lambda \qquad \text{almost surely.} \tag{62}$$

Thus the canonical designs in Table 1 share the same limiting reference measure but approach it at different rates. A schedule with a different weak limit is handled by the same proposition, and interior node placement alone does not determine that limit.

A.11 Finite-sampling bounds

We first prove a weighted concentration result. It contains the equal-pooling estimator of Theorem 9 as a special case. If $B_i = 0$, we take $\widetilde{w}_i = 0$, and otherwise assume $B_i > 0$ whenever $\widetilde{w}_i > 0$.

Lemma 6 (Weighted empirical aggregate concentration)**.** *At node α_i, let $S_{i,1}, \dots, S_{i,B_i} \overset{\mathrm{iid}}{\sim} \rho_{\alpha_i}$, with mutual independence across all active nodes and samples, and let $\widehat{\rho}_i$ be the empirical distribution. For realised weights $\widetilde{w}_i \ge 0$, $\sum_i \widetilde{w}_i = 1$, define*

$$\widehat{Q} := \sum_i \widetilde{w}_i \widehat{\rho}_i, \quad Q_{\widetilde{\nu}_n}[\rho] := \sum_i \widetilde{w}_i \rho_{\alpha_i}, \quad \kappa_{\mathrm{alloc}} := \sum_{i: B_i > 0} \frac{\widetilde{w}_i^2}{B_i}. \tag{63}$$

Then

$$\mathbb{E}\, d_{\mathrm{TV}}\Big(\widehat{Q}, Q_{\widetilde{\nu}_n}[\rho]\Big) \le \frac{1}{2}\sqrt{d\kappa_{\mathrm{alloc}}}, \tag{64}$$

and, for every $\delta \in (0,1)$, with probability at least $1 - \delta$,

$$d_{\mathrm{TV}}\Big(\widehat{Q}, Q_{\widetilde{\nu}_n}[\rho]\Big) \le \frac{1}{2}\sqrt{d\kappa_{\mathrm{alloc}}} + \sqrt{\frac{\kappa_{\mathrm{alloc}}}{2}\log\frac{1}{\delta}}. \tag{65}$$

Proof. Write $X_{i,j} := e_{S_{i,j}}$. For each coordinate s, independence gives

$$\begin{aligned}\mathrm{Var}(\widehat{Q}(s)) &= \sum_i \frac{\widetilde{w}_i^2}{B_i}\,\mathrm{Var}(X_{i,1}(s)) \\ &\le \sum_i \frac{\widetilde{w}_i^2}{B_i}\rho_{\alpha_i}(s).\end{aligned}$$

Hence, by Jensen's inequality coordinatewise and then Cauchy–Schwarz,

$$\begin{aligned}\mathbb{E}\|\widehat{Q} - Q_{\widetilde{\nu}_n}[\rho]\|_1 &\le \sum_s \sqrt{\sum_i \frac{\widetilde{w}_i^2}{B_i}\rho_{\alpha_i}(s)} \\ &\le \sqrt{d\sum_s\sum_i \frac{\widetilde{w}_i^2}{B_i}\rho_{\alpha_i}(s)} \\ &= \sqrt{d\kappa_{\mathrm{alloc}}}.\end{aligned}$$

This proves (64) after dividing by two.

Changing one observation at node i changes $\widehat{Q}$ by TV distance at most $\widetilde{w}_i/B_i$. Thus the sum of squared bounded-difference constants is $\sum_i B_i(\widetilde{w}_i/B_i)^2 = \kappa_{\mathrm{alloc}}$. McDiarmid's inequality then gives (65). □

Lemma 7 (Weight mismatch)**.** *For intended weights w and realised weights $\widetilde{w}$ on the same nodes,*

$$d_{\mathrm{TV}}(Q_{\widetilde{\nu}_n}[\rho], Q_{\nu_n}[\rho]) \le \frac{1}{2}\|\widetilde{w} - w\|_1. \tag{66}$$

Proof. By the definitions of the two aggregates,

$$Q_{\widetilde{\nu}_n}[\rho] - Q_{\nu_n}[\rho] = \sum_i (\widetilde{w}_i - w_i)\rho_{\alpha_i}.$$

Applying the triangle inequality,

$$\|Q_{\widetilde{\nu}_n}[\rho] - Q_{\nu_n}[\rho]\|_1 \le \sum_i |\widetilde{w}_i - w_i|\|\rho_{\alpha_i}\|_1.$$

Each ρ_{α_i} is a probability distribution, so $\|\rho_{\alpha_i}\|_1 = 1$. Hence

$$\|Q_{\widetilde{\nu}_n}[\rho] - Q_{\nu_n}[\rho]\|_1 \le \|\widetilde{w} - w\|_1.$$

Since both aggregates are probability distributions, $d_{\mathrm{TV}}(\mu,\nu) = \frac{1}{2}\|\mu - \nu\|_1$, which proves (66). □

Proof of Theorem 9. Under equal pooling, $\widetilde{w}_i = B_i/B$, and therefore

$$\kappa_{\mathrm{alloc}} = \sum_{i:B_i>0} \frac{(B_i/B)^2}{B_i} = \frac{1}{B}. \tag{67}$$

Inserting the exact aggregates corresponding to the realised and intended schedule measures:

$$\begin{aligned}d_{\mathrm{TV}}(\widehat{Q}_{n,B}, \bar{\rho}) \le\ & d_{\mathrm{TV}}(\widehat{Q}_{n,B}, Q_{\widetilde{\nu}_n}[\rho]) \\ & + d_{\mathrm{TV}}(Q_{\widetilde{\nu}_n}[\rho], Q_{\nu_n}[\rho]) + \mathcal{E}(\nu_n).\end{aligned} \tag{68}$$

Lemma 6 with $\kappa_{\mathrm{alloc}} = 1/B$ controls the first term, and Lemma 7 the second. With probability at least $1-\delta$,

$$d_{\mathrm{TV}}(\widehat{Q}_{n,B}, \bar{\rho}) \le \mathcal{E}(\nu_n) + \frac{1}{2}\|\widetilde{w} - w\|_1 + \frac{1}{2}\sqrt{\frac{d}{B}} + \sqrt{\frac{\log(1/\delta)}{2B}},$$

which is (12). □

Deterministic fallback. If arbitrary node estimates $\widehat{\rho}_i$ are aggregated with the intended weights, no independence assumption is needed for

$$d_{\mathrm{TV}}\left(\sum_i w_i \widehat{\rho}_i, \bar{\rho}\right) \le \sum_i w_i d_{\mathrm{TV}}(\widehat{\rho}_i, \rho_{\alpha_i}) + \mathcal{E}(\nu_n). \tag{69}$$

This is just the triangle inequality after inserting $Q_{\nu_n}[\rho]$. It is the appropriate statement when the independent-sampling model is not available.

Allocation. In the general weighted model, $\kappa_{\mathrm{alloc}} = \sum_i \widetilde{w}_i^2 / B_i$. If the intended weights are realised exactly, $\widetilde{w}_i = w_i$, then for continuous allocations with $\sum_i B_i = B$, Cauchy–Schwarz gives

$$1 = (\textstyle\sum_i w_i)^2 \le \left(\sum_i \frac{w_i^2}{B_i}\right)(\textstyle\sum_i B_i) = B\kappa_{\mathrm{alloc}}.$$

Hence $\kappa_{\mathrm{alloc}} \ge 1/B$, with equality when $B_i = B w_i$. Under equal pooling, $\kappa_{\mathrm{alloc}} = 1/B$ automatically, while proportional allocation is what keeps $\widetilde{w}_i = B_i/B$ close to the intended quadrature weights.

Trajectory-based estimates. The concentration lemma assumes independent draws from each occupancy. For temporally correlated trajectories, the deterministic decomposition (69) remains valid, but the sampling term must be replaced by a concentration result based on mixing or an effective sample size. Analyticity of the stationary path does not by itself provide such a statistical guarantee.

B Additional Exact-Evaluation Results

Figure 2 repeats Figures 1a and 1b for the discounted targets, from the FU-2 sweep (Appendix G). The registered slopes of the endpoint and Simpson rules do not move with γ. Instances whose two reference constructions disagree by more than 10^{-10} are excluded before any rule is evaluated, which removes 1, 1, and 7 of the 60 Garnet instances at $\gamma = 0.9$, 0.95, and 0.99, and no LQR instance.

C Additional Finite-Sampling Results

The variance inflation $n\sum_i w_i^2$ of Section 6.3 is 1.5 for three-node Simpson, an error scale of about $\sqrt{1.5} \approx 1.22$ against the 1.2 tolerance, and 1.32 (error scale 1.15) for five nodes. Table 4 and Figure 3 report rounds to reference at the two looser pre-registered tolerances, with the same conventions as Table 2, from the FU-4 run (Appendix G). The pre-registered criteria R1–R3 pass at all three tolerances, both on the registered grid and on the FU-4 grid.

Table 4: **Rounds to reference at tolerances $1.5\,\varepsilon_B$ and $2.0\,\varepsilon_B$.** Conventions as in Table 2.

	tolerance $1.5\,\varepsilon_B$					tolerance $2.0\,\varepsilon_B$				
Schedule	10^3	10^4	10^5	10^6	10^7	10^3	10^4	10^5	10^6	10^7
Gauss–Legendre	1	1	1	2	2	1	1	1	1	2
Midpoint	1	1	1	2	2.5	1	1	1	1	2
Simpson	3	3	3	3	3	3	3	3	3	3
Trapezoid	2	2	2	3	4	2	2	2	3	4
Random-uniform	1	2	3 (90)	12 (73)	>33 (43)	1	1	3 (98)	6 (80)	33 (53)
Endpoint Riemann	1	3	9.5 (87)	33 (55)	>33 (8)	1	2	6 (97)	20 (75)	>33 (33)
DAgger-L	1	3	9 (87)	33 (55)	>33 (8)	1	2	7 (98)	20 (73)	>33 (33)
DAgger-P	1 (98)	2 (78)	2 (60)	>33 (32)	>33 (0)	1	2 (85)	2 (73)	>33 (37)	>33 (2)

D Additional Continuous-Control Results

Table 6 reports the median evaluation loss of every method at the pre-specified cell. Table 7 reports the pre-registered F1 statistic under the primary complete-case reading and under a penalised reading

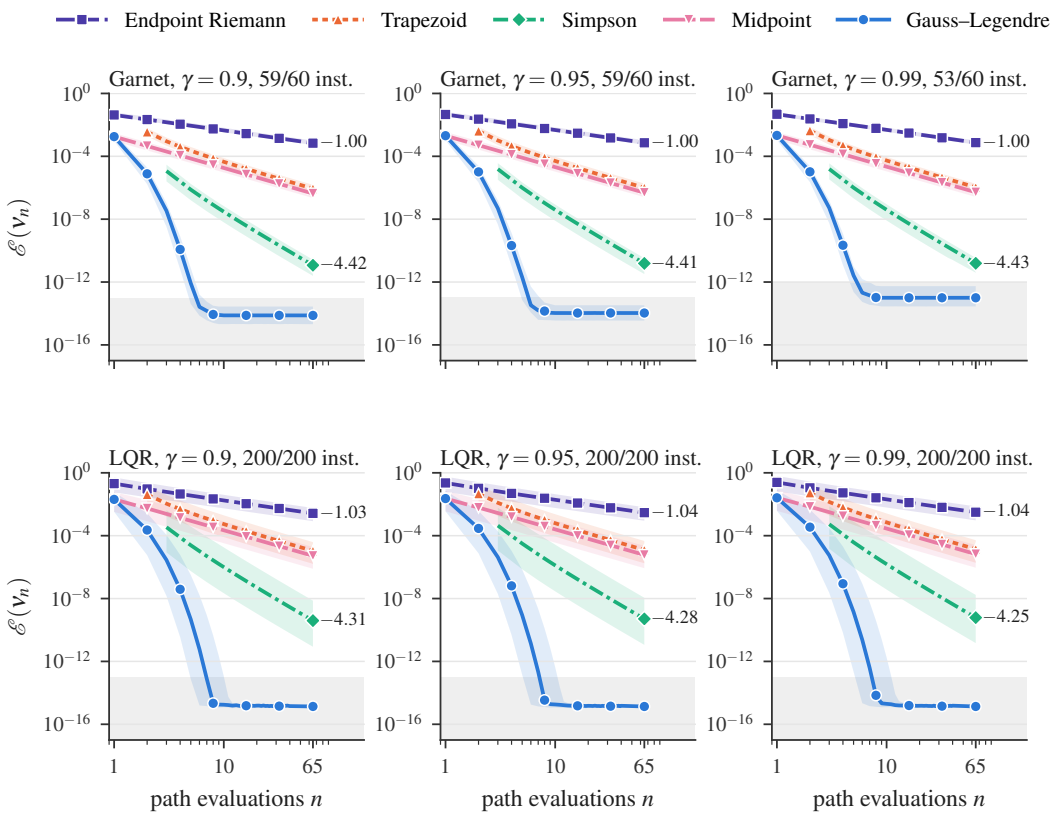

Figure 2: **Exact coverage error on discounted targets.** Same statistic and style as Figure 1a,b, with panel titles giving accepted over drawn instances. Labels are mean fitted log–log slopes.

that imputes each refusal at the worst J observed in its own cell, since dropping refused seeds flatters the refusing method. Across the full grid, 125 of 17,280 runs were refused (122 DAgger-P and 3 DAgger-G, all on Hopper and Walker2d, 107 of them from the random initialisation). The two readings agree.

The two quadrature methods of Table 5 share the fixed path and fitting procedure, so their contrast isolates the rule. The difference $J_{\mathrm{GL}} - J_{\mathrm{trap}}$ is unresolved on every environment at $B = 500$ and favours Gauss–Legendre on Hopper and Walker2d at $B = 10^4$. These are eight unadjusted comparisons, and the finite-sampling theorem does not apply to this continuous-state setting, so the change from $0/4$ to $2/4$ is suggestive only.

Lower reference-path loss also need not mean higher return. On Ant, frozen DAgger-P has the lowest median J of the frozen methods (Table 13) and the lowest median return, and the trapezoid nodes are resolved above it on return while their J difference is not resolved at all (Tables 14 and 15). Off-path behaviour, which J does not see, separates the methods there.

E Pre-Registration Record

Table 8 lists every pre-registered criterion the paper relies on. Protocols, instance registries, and graders were committed before the data they judge were generated.

Multiplicity. H2, H3 (with H3′), H4, and H7 are four members of a confirmatory family of seven. H2 and H7 concern results reported separately, and the remaining three belong to experiments not yet run. Holm correction over the family is applied once all seven exist. The verdicts here are declared at $0.05/7$, and any p-value below that level is rejected by Holm whatever the other six are.

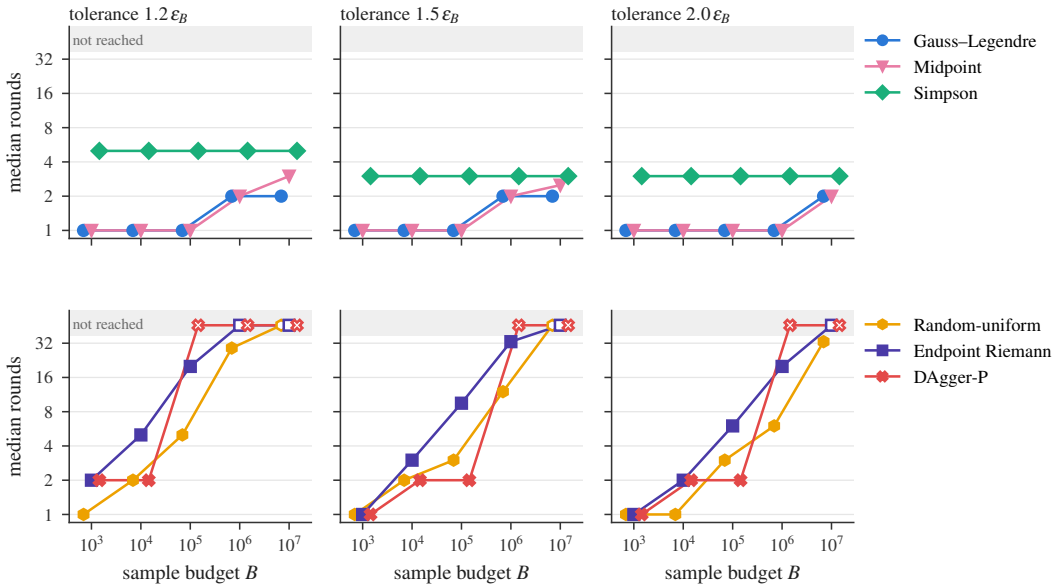

Figure 3: **Rounds to reference against the sample budget** at three tolerances. Designed rules are shown on top and baselines below. Points are medians over 60 held-out instances, horizontally offset for legibility, and open markers in the grey band mean the median instance never reaches the reference.

Table 5: **Factorial continuous-control contrasts** at $N = 5$ with behavioural-cloning initialisation (40 seeds/cell). $\Delta_{\mathrm{GL,P}}$ is the mean $(J_{\mathrm{GL}} - J_{\mathrm{P}})/J_{\mathrm{bc}}$ at $B = 2000$. The other columns report 10^2 times the mean paired $J_{\mathrm{GL}} - J_{\mathrm{trap}}$. Brackets are 95% percentile-bootstrap intervals (10^4 resamples), unadjusted for multiplicity, and bold intervals exclude zero. Negative values favour Gauss–Legendre. The two families also differ in retraining, in the number of fits, and in environment steps (Table 9), so the first column compares procedures; Table 3 isolates the node set.

	$\Delta_{\mathrm{GL,P}}$	$10^2\,(J_{\mathrm{GL}} - J_{\mathrm{trap}})$	
Environment	$B = 2000$	$B = 500$	$B = 10^4$
Hopper	$\mathbf{-0.133}$ [−0.178, −0.091]	−0.96 [−2.18, +0.25]	$\mathbf{-1.25}$ [−1.93, −0.59]
HalfCheetah	$\mathbf{-0.060}$ [−0.075, −0.048]	+1.54 [−7.18, +10.74]	−0.55 [−1.22, +0.13]
Walker2d	$\mathbf{-0.140}$ [−0.165, −0.115]	−0.31 [−1.90, +1.30]	$\mathbf{-2.25}$ [−3.03, −1.50]
Ant	+0.020 [−0.006, +0.048]	+0.07 [−2.92, +3.25]	+0.08 [−0.84, +0.96]

The binomial p-value of the count-based criterion H3′ on the Garnet $\gamma = 0.95$ cell, 7.3×10^{-3}, is just above that level, and its pass is decided by the registered counts rather than by the p-value.

Amendments and failures. C4a was registered in an $R^2 \geq 0.95$ form, which passed on 29–67% of instances in the primary cells. After it failed we replaced it by a residual measured in nodes, which respects the integer lattice of node counts; both readings are reported, and the direct log-linear form passes on every instance. P1's registered slope band tested flatness, whereas the claim is non-convergence, and it passed on no instance. We replaced it by $\mathcal{E}(40) \geq \frac{1}{2}\mathcal{E}(4)$ read together with P3, and the superseded reading is retained in the released artifact. P7 compared Gauss–Legendre at 5 nodes with endpoint Riemann at 20 nodes at matched labels; it passed at 2 of 5 budgets and is withdrawn, because at matched labels the sampling error barely depends on n. The rounds-to-reference statistic was found after P1 and P7 failed, so it was registered as a new test on 60 freshly drawn instances, with disjointness from the hypothesis-forming instances verified by hashing. A defect in its criterion, a median that was undefined when the hypothesis held strongly, was found in a smoke test on three held-out instances. Those instances and their seed were discarded before any graded instance existed. F1 failed under both readings (Table 7), F2 passed, and the secondary directional criterion F3 held on 2 of the required 3 environments.

Table 6: **Median evaluation loss J of every method** at $N = 5$, $B = 2000$, behavioural-cloning initialisation (40 seeds, lower is better). Bold marks the lowest median among the interactive methods. Medians are not paired contrasts and carry no interval.

		DAgger			Quadrature	
Environment	BC	P	G	L	Trapezoid	Gauss–Legendre
Hopper	0.881	0.580	**0.411**	0.469	0.485	0.484
HalfCheetah	5.967	1.600	**1.255**	1.339	1.315	1.289
Walker2d	0.959	0.935	**0.743**	0.755	0.789	0.800
Ant	0.766	0.458	**0.456**	0.492	0.487	0.462

Table 7: **F1 under complete-case and penalised readings**: median $J_{\mathrm{P}}/J_{\mathrm{bc}}$ at $B = 2000$ with the behavioural-cloning initialisation, and its growth from $N = 3$ to $N = 10$. Parentheses give the number of refused seeds. F1 required monotone growth of at least 20% on three environments.

	complete case			penalised			growth $N{=}3 \to 10$	
Environment	$N{=}3$	5	10	$N{=}3$	5	10	complete	penalised
Hopper	0.721 (2)	0.692	0.597	0.724	0.692	0.597	0.829	0.824
HalfCheetah	0.280	0.268	0.266	0.280	0.268	0.266	0.949	0.949
Walker2d	0.896 (1)	0.946	1.015	0.900	0.946	1.015	1.133	1.128
Ant	0.587	0.583	0.580	0.587	0.583	0.580	0.987	0.987

Defects and mismatches found after grading. Four defects were found by reading code while preparing the follow-up protocol, before any follow-up value existed. None changes a registered verdict, and each is corrected in the presentation.

- The quadrature library returns the single midpoint node for the trapezoid rule at $n = 1$. The E1 sweep therefore records a midpoint value as trapezoid at $n = 1$, and the registered H3 trapezoid slope includes it (FU-2 reports the slope without it).
- The rounds-to-reference runner applied no admissibility rule, so the registered Simpson row used the two-node trapezoid rule at $n = 2$ and a mixed 3/8–1/3 rule at even n. Table 2 uses the FU-4 re-run, which evaluates trapezoid only for $n \geq 2$ and Simpson only for odd $n \geq 3$.
- An earlier draft described the LQR experiment as a policy-mixture path with total-variation error. It is a gain-interpolation path with a relative covariance error.
- An earlier draft described the rounds-to-reference sampling as proportional allocation. Each node receives the same number of samples, and the node estimates are reweighted.

F Experimental Details and Reproducibility

Every number in this paper is read from a recorded run artifact by a script, and none is entered by hand. Each artifact records the code commit that produced it, the interpreter and library versions, the thread settings, and, for continuous control, the SHA-256 digest of every expert checkpoint. Table 10 lists the artifacts behind each result.

Software and hardware. Python 3.11.15, NumPy 2.4.4, SciPy 1.17.1, PyTorch 2.11.0 (CUDA build, with no device visible), Gymnasium 1.3.0 with MuJoCo 3.8.1, and Stable-Baselines3 2.9.0. Every process runs with `OMP_NUM_THREADS`, `MKL_NUM_THREADS`, and `OPENBLAS_NUM_THREADS` set to 1 and `PYTHONHASHSEED`=0, on an Intel Core i7-14700 (28 hardware threads) with no GPU in the computation. Randomness is drawn from named substreams keyed by a run seed and a string, so the stream an operation uses does not depend on execution order. Re-running completed cells reproduced their recorded values bit for bit (below).

Finite-state instances. Garnet MDPs have $S \in \{8, 20, 50\}$ states, $A \in \{3, 5\}$ actions drawn uniformly, and branching $b = \lceil S/4 \rceil$, so each state–action pair has b successors drawn without

Table 8: **Pre-registered criteria and outcomes.** Primary cells are the stationary and $\gamma = 0.95$ discounted targets on Garnet and LQR.

	Criterion	Outcome	Status
H3	slope bands (endpoint Riemann, trapezoid, Simpson)	CIs inside bands on all 4 primary cells; TOST $p \leq 1.0\times10^{-4}$	as registered
H3′	ordering GL < Simpson < trapezoid < Riemann	0 violations; binomial $p \in [1.6\times10^{-10}, 7.3\times10^{-3}]$	as registered
H4	GL geometric decay, per-node factor < 0.5	direct form 100% of instances; binomial $p \leq 2.2\times10^{-3}$	as registered
C4a	nodes-to-accuracy linear in $\log(1/\epsilon)$	pre-registered $R^2 \geq 0.95$ form 29%–67%; node-residual form passes	amended after failing
P1	DAgger-P does not converge to $\bar{\rho}$	slope band 0%; $\mathcal{E}(40) \geq \frac{1}{2}\mathcal{E}(4)$ on 100%	amended after failing
P2	DAgger-L slope in the Riemann band	100%; median slope -1.0015	as registered
P3	DAgger-P limit is $d_{\mathrm{TV}}(\rho_0, \bar{\rho})$ within 10%	100%; median gap 5.05%	as registered
P7	GL (n=5) beats Riemann (n=20) at matched labels	passed at 2 of 5 budgets	failed; withdrawn
R1–R3	rounds to reference on held-out instances	pass at tolerances 1.2, 1.5, 2.0	as registered
F1	DAgger-P loss grows $\geq 20\%$ from N=3 to 10	0 of 4 environments; penalised reading agrees	failed
F2	GL−trapezoid resolved on more environments at B=10^4	0 of 4 at $B = 500$, 2 of 4 at $B = 10^4$	passed
FU-2 M1	midpoint slope in second-order band	mean slopes -1.997 to -1.956 on 4 primary cells	passed
FU-2 M3	trapezoid/midpoint error ratio at $n = 33$ in $[1.8, 2.3]$	medians 2.127, 2.127	passed
FU-3 T-C6	Corollary 6 bounds hold	0 violations in 18,216 checks	passed
FU-3 T-P7	measured error over Prop. 7 leading term at $n = 256$ in $[0.9, 1.1]$	ratios 0.9988–1.0002	passed
FU-3 T-P2	support constancy; chain identity	all checks hold	passed
FU-2 repro	re-run equals committed values	0 of 209,293 differ	passed
FU-4 repro	re-run equals committed values	0 of 31,500 differ	passed
FU-5 repro	re-run J equals factorial v3	0 of 960 differ	passed
FU-5 P-J1	$\mathrm{GL}^{\mathrm{f}} < \mathrm{P}^{\mathrm{f}}$ on J, ≥ 3 of 4	3 of 4 after Holm	passed
FU-5 P-J2	$\mathrm{L}^{\mathrm{f}} < \mathrm{P}^{\mathrm{f}}$ on J, ≥ 3 of 4	3 of 4 after Holm	passed

replacement with Dirichlet($\mathbf{1}$) probabilities, and $\mu_0 \sim$ Dirichlet($\mathbf{1}$). The expert is Dirichlet(0.5) per state and $\hat{\pi}_0$ is uniform. An instance is accepted if the mixture chain is ergodic on an acceptance grid of α, and 20 are accepted per size (60 total). LQR systems have state dimension $n \in \{2, 4, 8, 16, 32\}$ and input dimension $m \in \{1, \max(2, \lfloor n/2 \rfloor)\}$, 20 per (n, m) (200 total), with open-loop spectral radius 0.85. The path interpolates the gain, $K_\alpha = (1-\alpha)K_0 + \alpha K^\star$, between a perturbed gain K_0 and the optimal gain $K^\star$. The integrand is the stationary or discounted state covariance Σ_α, and the error is $\|\widehat{\Sigma} - \bar{\Sigma}\|_F / \|\bar{\Sigma}\|_F$. Instances are generated from seed 20260815, and the registry digest begins `fe619d2f`. Targets are stationary and discounted with $\gamma \in \{0.9, 0.95, 0.99\}$, and the stationary and $\gamma = 0.95$ targets are primary.

Exact evaluation. The reference integral $\bar{\rho}$ is computed by tanh–sinh quadrature (level 8) and adaptive vector Gauss–Kronrod quadrature (absolute tolerance 10^{-14}, relative 10^{-13}), plus a closed form for discounted Garnet targets. An instance whose constructions disagree by more than 10^{-10} is excluded before any rule is evaluated, and errors below $\max(10^{-13}, 10\delta)$, with δ the disagreement, are treated as at the reference accuracy. Node counts are $n \in \{1, \ldots, 40, 65\}$. Slopes are least-squares fits of $\log_{10}$ error on $\log_{10} n$ (for Gauss–Legendre, on n) over the contiguous prefix of

points above that accuracy, with at least four points, and Simpson is fitted on odd n. Intervals are stratified percentile bootstraps over instances (strata S or n) with 10^4 resamples.

Schedule degeneracy and rounds to reference. Figure 1c uses the 60 Garnet instances, the stationary target, $N \in \{2, \dots, 10, 12, 17, 25, 33, 40\}$, and $p = 0.5$ for DAgger-G. Rounds to reference uses 60 Garnet instances drawn afresh from seed 20260904 with the same specification, and their parameters were hashed and found disjoint from the 60 above. Aggregates are formed by multinomial sampling of $\lfloor B/n \rfloor$ states per node from the exact stationary occupancies, 10 draws per (instance, method, B, n), budgets $B \in \{10^3, \dots, 10^7\}$ and tolerances $\{1.2, 1.5, 2.0\}$. A median instance that never reaches the reference is ordered at 34.

Continuous control. Environments are Hopper-v5, HalfCheetah-v5, Walker2d-v5, and Ant-v5 with the Gymnasium default time limit of 1000 steps. Experts are Stable-Baselines3 SAC checkpoints, accepted before use against fixed return bars (HalfCheetah 8581.9, Hopper 1744.2, Walker2d 2897.4, Ant 3461.8). Each bar is a fraction of the published return for the same algorithm and environment: 0.9 where the reward function is unchanged, 0.75 where it is not. Acceptance was decided over 200 evaluation episodes, which also had to resolve the expert-minus-random gap to within 5% and keep the expert's episode spread below half that gap. The SHA-256 digest of every accepted checkpoint is recorded in each artifact.

The learner is an MLP with two hidden layers of 256 tanh units, trained with Adam (learning rate 10^{-3}, batch 256) on squared action error, with 15% of the data held out for early stopping (minimum improvement 10^{-5}). $\hat{\pi}_0$ is fit by behavioural cloning on 10 expert episodes (at most 30 epochs, patience 5). All other fits use at most 60 epochs with patience 8, and every expert label is the deterministic expert action. Reporting seeds are 0–39, and seeds 100–104 were used only for pilots and code checks.

- **BC:** expert episodes until at least B states and 20 episodes are pooled, then B states drawn uniformly and one fit.
- **DAgger-P/L/G:** an expert node of $\max(5, \lceil B/(25N) \rceil)$ episodes, then $N-1$ rounds of the same number of episodes under the per-step mixture of the current learner and the expert with expert weight β_i from Table 1. Labels are drawn uniformly with an equal share per round, and the learner is refit after every round.
- **Quadrature (Gauss–Legendre, trapezoid):** $N\lfloor B/N \rfloor$ particles allocated across the N nodes in proportion to the weights. Each particle rolls out the fixed mixture of $\hat{\pi}_0$ and the expert for 200 steps, resetting on termination. B states are drawn uniformly from the pool, followed by one fit.
- **Frozen-path methods** (Appendix G): whole episodes of the fixed mixture at each node until the node's label count $B\,w_i$ (largest-remainder rounding) is pooled, $B\,w_i$ states drawn uniformly per node, and one fit.

At the centre cell the quadrature methods take $200\,N\lfloor B/N \rfloor = 400{,}000$ environment steps and the DAgger methods at most $80{,}000$; Table 9 gives the recorded medians for every method.

J uses 20 mixture levels $\alpha_k = (2k+1)/41$, three episodes of the fixed mixture each, and at most 400 states per level, shared by all methods of a seed. The denominator is 41 rather than 40 so that no evaluation level coincides with a collection node of any schedule: 41 is coprime to the node denominators in use, whereas $(2k+1)/40$ hits DAgger-G's node at 0.125. The runner recomputes this check and aborts on a collision. Return and on-policy loss use 10 episodes of the deterministic policy on reset seeds shared by all methods of a seed. Intervals are percentile bootstraps over seeds with 10^4 resamples, and Holm corrections are applied within each reported family.

G Follow-Up Experiments

A second protocol, committed before any of its cells ran, specified the five follow-up experiments FU-1 to FU-5 below. They re-run committed code, add opt-in flags whose defaults leave every earlier artifact reproducible, or add new runners. All criteria, tolerances, and contrast families were fixed in that protocol. Two tolerances were revised before any FU-3 value existed, because the originals would have tested round-off rather than the claim.

Table 9: **Median environment steps per run at the centre cell** ($N = 5$, $B = 2000$, behavioural-cloning initialisation, 40 seeds). Every method is given the same label budget, so these counts are what that budget costs in environment interaction. The frozen-path methods of Appendix G are below the rule; they collect whole episodes and fit once, which is why their counts are both small and close to one another.

Method	Hopper	HalfCheetah	Walker2d	Ant
BC	20,000	20,000	19,418	18,856
DAgger-P	25,440	80,000	22,849	69,067
DAgger-L	56,510	80,000	46,900	70,631
DAgger-G	39,262	80,000	31,274	69,139
Gauss–Legendre	400,000	400,000	400,000	400,000
Trapezoid	400,000	400,000	400,000	400,000
P^{f}	2,943	5,000	2,860	5,000
L^{f}	3,786	5,000	4,084	5,000
G^{f}	3,212	5,000	3,249	5,000
GL^{f}	3,682	5,000	3,836	5,000
$\mathrm{Trap}^{\mathrm{f}}$	3,695	5,000	3,696	5,000

Table 10: **Result artifacts and the code commits recorded in them.** Commits are abbreviated to seven characters.

Used in	Content	Commit
Fig. 1a,b; App. B (exact hierarchy)	E1 sweep with midpoint (FU-2)	`8807a27`
H3, H3′, H4 verdicts	registered E1 grading	`c94bc20`
Fig. 1c	frozen-path DAgger schedules	`c54f89a`
P7	placement vs. count	`7f4e03f`
R1–R3 (registered)	rounds to reference	`4d44e9f`
Tab. 2; App. C	rounds, $n \geq 1$ (FU-4)	`8807a27`
Tab. 3; App. D; F1–F5	continuous-control factorial and grader	`91cfd6b, 4f24678`
App. G, FU-1	contrasts (derived)	`8807a27`
App. G, FU-2/FU-4 grading	reproduction, M1–M3	`8807a27`
App. G, FU-3	theory checks, hard instances	`8807a27`
App. G, FU-5	centre cell re-run	`8807a27`

Reproduction. Re-running existing cells reproduced every recorded value bit for bit: 209,293 exact-evaluation errors of the E1 sweep (FU-2), 31,500 per-instance sampled errors of the rounds-to-reference run (FU-4), and 960 values of J from the continuous-control factorial (FU-5).

FU-1: contrasts among the factorial's DAgger methods. Table 11 gives the registered contrasts on the factorial's centre cell.

Table 11: **Paired contrasts of J/J_{bc} among factorial methods** at $N = 5$, $B = 2000$ (40 seeds). Means with 95% percentile-bootstrap intervals. Bold entries are resolved after Holm correction over all 20 tests. G, L, and P denote DAgger-G, DAgger-L, and DAgger-P, and GL and Trap the quadrature methods.

Contrast	Hopper	HalfCheetah	Walker2d	Ant
GL − G	$\mathbf{+0.084}\,[+0.071,\ +0.098]$	$\mathbf{+0.008}\,[+0.004,\ +0.013]$	$\mathbf{+0.051}\,[+0.027,\ +0.074]$	$\mathbf{+0.037}\,[+0.017,\ +0.057]$
Trap − G	$\mathbf{+0.089}\,[+0.076,\ +0.104]$	$\mathbf{+0.011}\,[+0.006,\ +0.017]$	$\mathbf{+0.047}\,[+0.025,\ +0.068]$	$\mathbf{+0.056}\,[+0.030,\ +0.084]$
GL − L	$+0.011\,[-0.007,\ +0.030]$	$\mathbf{-0.010}\,[-0.020,\ -0.002]$	$\mathbf{+0.036}\,[+0.010,\ +0.060]$	$-0.027\,[-0.064,\ +0.009]$
G − P	$\mathbf{-0.218}\,[-0.259,\ -0.177]$	$\mathbf{-0.069}\,[-0.083,\ -0.057]$	$\mathbf{-0.191}\,[-0.217,\ -0.166]$	$-0.017\,[-0.037,\ +0.002]$
L − P	$\mathbf{-0.145}\,[-0.182,\ -0.109]$	$\mathbf{-0.051}\,[-0.060,\ -0.042]$	$\mathbf{-0.176}\,[-0.203,\ -0.150]$	$\mathbf{+0.046}\,[+0.013,\ +0.087]$

FU-2: midpoint in exact evaluation. The E1 sweep was re-run with the midpoint rule added.

- **Midpoint slopes** (95% intervals): Garnet stationary -1.996 $[-1.998, -1.995]$; Garnet $\gamma = 0.95$ -1.997 $[-1.999, -1.995]$; LQR stationary -1.956 $[-1.965, -1.947]$; LQR $\gamma = 0.95$

-1.967 $[-1.973, -1.961]$. Each interval lies inside the second-order band $[-2.4, -1.7]$, as registered.

- **Trapezoid slopes.** Excluding the $n = 1$ point, which is the midpoint node (Appendix E), gives Garnet stationary -2.273, Garnet $\gamma = 0.95$ -2.273, LQR stationary -2.247, and LQR $\gamma = 0.95$ -2.254, against registered values that include it of -2.124, -2.125, -2.083, and -2.094 for the same four cells, in the same order.
- **Trapezoid/midpoint error ratio at** $n = 33$ (median over Garnet instances): 2.127 (stationary) and 2.127 ($\gamma = 0.95$), inside the registered interval $[1.8, 2.3]$.

FU-3: numerical checks of the new results. The checks use 133 exact occupancy paths:

- the 60 Garnet instances, each with its stationary and $\gamma = 0.95$ targets;
- chain MDPs with $K \in \{10, 30, 100\}$ states and $\gamma \in \{0.9, 0.99, 0.999\}$, where the expert advances and the learner resets;
- two-state near-decomposable chains with $\varepsilon \in \{10^{-1}, 10^{-2}, 10^{-3}\}$;
- a three-state example.

Node counts run to $n = 256$. Results:

- **Corollary 6:** its three bounds hold in all 18,216 comparisons.
- **Proposition 7:** on the Garnet targets, the ratio of measured error to leading term at $n = 256$ lies in $[0.9988,\ 1.0002]$ for the endpoint, midpoint, and trapezoid rules (119 targets each, with 1 target excluded by the reference check).
- **Proposition 2:** on the three-state example, $\operatorname{supp}\rho_\alpha = \{0, 1, 2\}$ at all 99 interior grid points, $\operatorname{supp}\rho_0 = \{0, 1\}$ and $\operatorname{supp}\rho_1 = \{0\}$, so every measure on $\{0, 1\}$ leaves state 2 uncovered, while $C(\lambda) = 1.80$.
- **Chain identity:** $\rho_\alpha(k) = (\gamma\alpha)^k \rho_\alpha(0)$ holds on every chain to the registered tolerance.

Table 12 reports how many nodes each rule needs on the hard families.

Table 12: **Nodes needed for total variation at most** 10^{-4} on hard instances (exact evaluation, $n \le 256$). L is the TV-Lipschitz constant on a 2001-point grid. Simpson is evaluated at odd n and trapezoid for $n \ge 2$.

Instance	L	$d_{\mathrm{TV}}(\rho_0, \bar{\rho})$	GL	Simpson	Midpoint	Trapezoid	Riemann
chain, $K = 10$, $\gamma = 0.9$	3.5	0.45	4	13	39	64	>256
chain, $K = 10$, $\gamma = 0.99$	8.2	0.49	4	15	64	96	>256
chain, $K = 10$, $\gamma = 0.999$	8.9	0.50	5	15	64	96	>256
chain, $K = 30$, $\gamma = 0.9$	3.5	0.45	6	17	39	64	>256
chain, $K = 30$, $\gamma = 0.99$	21.5	0.49	8	33	96	192	>256
chain, $K = 30$, $\gamma = 0.999$	28.0	0.50	8	35	128	192	>256
chain, $K = 100$, $\gamma = 0.9$	3.5	0.45	6	17	39	64	>256
chain, $K = 100$, $\gamma = 0.99$	35.7	0.49	12	>256	128	192	>256
chain, $K = 100$, $\gamma = 0.999$	87.5	0.50	13	>256	256	>256	>256
two-state, $\varepsilon = 10^{-1}$	10.0	0.76	7	25	96	96	>256
two-state, $\varepsilon = 10^{-2}$	95.2	0.95	16	>256	256	>256	>256
two-state, $\varepsilon = 10^{-3}$	666.7	0.99	32	>256	>256	>256	>256
three-state, $\gamma = 0.9$	1.7	0.21	4	9	27	39	>256

FU-4: rounds to reference with $n = 1$**.** Table 2 and Appendix C use this run. The fraction of instances whose rounds to reference is exactly one at tolerance 1.2 is 97, 85, 63, 32, 0% for Gauss–Legendre and 95, 78, 62, 32, 0% for midpoint at $B = 10^3, \ldots, 10^7$. The registered criteria R1–R3, recomputed on this grid, still pass at all three tolerances.

FU-5: frozen-path methods, return, and on-policy loss. The six factorial methods were re-run at the centre cell ($N = 5$, $B = 2000$, behavioural-cloning initialisation, 40 seeds). Their J equals the factorial's recorded value in 960 of 960 rows. Five frozen-path methods (superscript f) were added. Each collects whole episodes of the fixed mixture of $\hat{\pi}_0$ and the expert at the nodes of DAgger-P, DAgger-L, DAgger-G, Gauss–Legendre, or trapezoid, with the rule's label allocation, and fits once. Every trained policy, the expert, and $\hat{\pi}_0$ are scored by mean return and by squared action loss over 10 evaluation episodes on shared reset seeds. Return differences are divided by the seed's expert-minus-$\hat{\pi}_0$ return, and no seed has a non-positive denominator. No run was refused and none failed.

Table 13 gives medians, and Tables 14–16 give the registered contrasts. Bold entries are resolved after Holm correction over the 44 tests of each outcome. The registered predictions P-J1 and P-J2 (frozen Gauss–Legendre and frozen DAgger-L below frozen DAgger-P on J on at least three environments) both pass, each on three.

Table 13: **Median J and median return of every policy** at the centre cell (40 seeds). Superscript f marks frozen-path methods.

	Hopper		HalfCheetah		Walker2d		Ant	
Policy	J	return	J	return	J	return	J	return
BC	0.881	468	5.967	897	0.959	278	0.766	2737
P	0.580	626	1.600	1611	0.935	358	0.458	3431
L	0.469	1039	1.339	1165	0.755	368	0.492	3374
G	0.411	692	1.255	1706	0.743	439	0.456	3698
Trap	0.485	562	1.315	881	0.789	387	0.487	3404
GL	0.484	671	1.289	1063	0.800	375	0.462	3608
P^{f}	0.680	243	1.987	724	0.968	221	0.622	2775
L^{f}	0.427	877	1.392	853	0.746	309	0.567	3187
G^{f}	0.416	583	1.399	889	0.750	290	0.588	3364
$\mathrm{Trap}^{\mathrm{f}}$	0.414	530	1.410	821	0.728	312	0.701	3543
GL^{f}	0.416	671	1.385	638	0.735	303	0.745	3404
$\hat{\pi}_0$	–	493	–	858	–	276	–	2804
expert	–	3435	–	11101	–	4442	–	4779

Table 14: **FU-5 contrasts of J/J_{bc}** (negative favours the first method). Rows 1–5 vary the schedule with retraining fixed, rows 6–8 vary retraining with the nodes fixed, and rows 9–11 compare the factorial methods.

Contrast	Hopper	HalfCheetah	Walker2d	Ant
$\mathrm{GL}^{\mathrm{f}} - \mathrm{P}^{\mathrm{f}}$	$\mathbf{-0.320}\,[-0.363,\ -0.276]$	$\mathbf{-0.110}\,[-0.155,\ -0.077]$	$\mathbf{-0.265}\,[-0.292,\ -0.238]$	$+0.042\,[-0.030,\ +0.125]$
$\mathrm{Trap}^{\mathrm{f}} - \mathrm{P}^{\mathrm{f}}$	$\mathbf{-0.315}\,[-0.360,\ -0.271]$	$\mathbf{-0.098}\,[-0.128,\ -0.071]$	$\mathbf{-0.266}\,[-0.287,\ -0.244]$	$+0.013\,[-0.039,\ +0.069]$
$\mathrm{L}^{\mathrm{f}} - \mathrm{P}^{\mathrm{f}}$	$\mathbf{-0.303}\,[-0.349,\ -0.259]$	$\mathbf{-0.110}\,[-0.154,\ -0.073]$	$\mathbf{-0.247}\,[-0.273,\ -0.220]$	$-0.028\,[-0.090,\ +0.030]$
$\mathrm{G}^{\mathrm{f}} - \mathrm{P}^{\mathrm{f}}$	$\mathbf{-0.313}\,[-0.359,\ -0.266]$	$\mathbf{-0.108}\,[-0.151,\ -0.074]$	$\mathbf{-0.264}\,[-0.289,\ -0.238]$	$\mathbf{-0.095}\,[-0.158,\ -0.033]$
$\mathrm{GL}^{\mathrm{f}} - \mathrm{Trap}^{\mathrm{f}}$	$-0.005\,[-0.014,\ +0.005]$	$-0.012\,[-0.042,\ +0.015]$	$+0.000\,[-0.014,\ +0.015]$	$+0.029\,[-0.039,\ +0.103]$
$\mathrm{P} - \mathrm{P}^{\mathrm{f}}$	$\mathbf{-0.103}\,[-0.133,\ -0.076]$	$\mathbf{-0.074}\,[-0.113,\ -0.042]$	$\mathbf{-0.063}\,[-0.090,\ -0.036]$	$\mathbf{-0.278}\,[-0.339,\ -0.214]$
$\mathrm{L} - \mathrm{L}^{\mathrm{f}}$	$\mathbf{+0.055}\,[+0.040,\ +0.071]$	$-0.015\,[-0.035,\ +0.000]$	$+0.007\,[-0.012,\ +0.027]$	$\mathbf{-0.204}\,[-0.271,\ -0.134]$
$\mathrm{G} - \mathrm{G}^{\mathrm{f}}$	$-0.008\,[-0.026,\ +0.007]$	$\mathbf{-0.035}\,[-0.048,\ -0.023]$	$+0.009\,[-0.011,\ +0.030]$	$\mathbf{-0.200}\,[-0.264,\ -0.141]$
GL − P	$\mathbf{-0.133}\,[-0.177,\ -0.091]$	$\mathbf{-0.060}\,[-0.075,\ -0.048]$	$\mathbf{-0.140}\,[-0.164,\ -0.115]$	$+0.020\,[-0.006,\ +0.048]$
GL − G	$\mathbf{+0.084}\,[+0.071,\ +0.098]$	$\mathbf{+0.008}\,[+0.004,\ +0.013]$	$\mathbf{+0.051}\,[+0.027,\ +0.074]$	$\mathbf{+0.037}\,[+0.018,\ +0.057]$
L − P	$\mathbf{-0.145}\,[-0.182,\ -0.109]$	$\mathbf{-0.051}\,[-0.060,\ -0.042]$	$\mathbf{-0.176}\,[-0.203,\ -0.150]$	$+0.046\,[+0.012,\ +0.088]$

Table 15: **FU-5 contrasts of normalised return** (positive favours the first method). Row groups as in Table 14.

Contrast	Hopper	HalfCheetah	Walker2d	Ant
$GL^f - P^f$	$+\mathbf{0.209}\,[+0.154,\ +0.273]$	$-0.005\,[-0.025,\ +0.015]$	$+0.023\,[+0.004,\ +0.042]$	$+0.142\,[-0.168,\ +0.400]$
$Trap^f - P^f$	$+\mathbf{0.130}\,[+0.092,\ +0.168]$	$+0.002\,[-0.018,\ +0.022]$	$+0.020\,[+0.001,\ +0.039]$	$+\mathbf{0.288}\,[+0.127,\ +0.455]$
$L^f - P^f$	$+\mathbf{0.232}\,[+0.173,\ +0.296]$	$+0.010\,[-0.011,\ +0.031]$	$+0.021\,[+0.002,\ +0.039]$	$+0.105\,[-0.183,\ +0.362]$
$G^f - P^f$	$+\mathbf{0.123}\,[+0.089,\ +0.159]$	$+0.022\,[+0.000,\ +0.044]$	$+0.022\,[+0.003,\ +0.043]$	$+0.179\,[-0.165,\ +0.474]$
$GL^f - Trap^f$	$+0.079\,[+0.017,\ +0.144]$	$-0.007\,[-0.029,\ +0.016]$	$+0.002\,[-0.017,\ +0.021]$	$-0.145\,[-0.435,\ +0.100]$
$P - P^f$	$+\mathbf{0.117}\,[+0.092,\ +0.141]$	$+\mathbf{0.073}\,[+0.055,\ +0.090]$	$+\mathbf{0.034}\,[+0.020,\ +0.047]$	$+0.201\,[-0.084,\ +0.448]$
$L - L^f$	$+0.034\,[-0.058,\ +0.125]$	$+0.027\,[+0.007,\ +0.048]$	$+0.014\,[-0.008,\ +0.036]$	$+0.128\,[-0.071,\ +0.361]$
$G - G^f$	$+0.041\,[-0.002,\ +0.084]$	$+\mathbf{0.065}\,[+0.042,\ +0.087]$	$+\mathbf{0.050}\,[+0.019,\ +0.080]$	$+0.183\,[-0.066,\ +0.464]$
GL − P	$+0.035\,[-0.006,\ +0.076]$	$-\mathbf{0.049}\,[-0.067,\ -0.032]$	$+0.013\,[-0.005,\ +0.032]$	$+0.181\,[-0.025,\ +0.414]$
GL − G	$-0.013\,[-0.058,\ +0.032]$	$-\mathbf{0.064}\,[-0.087,\ -0.040]$	$-0.025\,[-0.053,\ +0.002]$	$+0.020\,[-0.146,\ +0.187]$
L − P	$+\mathbf{0.149}\,[+0.090,\ +0.211]$	$-\mathbf{0.035}\,[-0.053,\ -0.016]$	$+0.001\,[-0.015,\ +0.016]$	$+0.032\,[-0.129,\ +0.204]$

Table 16: **FU-5 contrasts of on-policy loss divided by BC's** (negative favours the first method). Row groups as in Table 14.

Contrast	Hopper	HalfCheetah	Walker2d	Ant
$GL^f - P^f$	$-\mathbf{0.183}\,[-0.265,\ -0.113]$	$-0.011\,[-0.080,\ +0.064]$	$-0.209\,[-0.364,\ -0.046]$	$-0.270\,[-0.707,\ +0.138]$
$Trap^f - P^f$	$-\mathbf{0.140}\,[-0.221,\ -0.074]$	$-0.058\,[-0.121,\ +0.002]$	$-0.154\,[-0.271,\ -0.031]$	$-0.431\,[-0.935,\ +0.007]$
$L^f - P^f$	$-\mathbf{0.164}\,[-0.254,\ -0.094]$	$-0.033\,[-0.090,\ +0.027]$	$-0.107\,[-0.266,\ +0.048]$	$-0.329\,[-0.975,\ +0.307]$
$G^f - P^f$	$-\mathbf{0.166}\,[-0.255,\ -0.090]$	$-\mathbf{0.086}\,[-0.135,\ -0.036]$	$-0.145\,[-0.285,\ -0.006]$	$-0.522\,[-1.024,\ -0.055]$
$GL^f - Trap^f$	$-0.042\,[-0.086,\ +0.003]$	$+0.047\,[-0.017,\ +0.118]$	$-0.055\,[-0.209,\ +0.102]$	$+0.161\,[-0.259,\ +0.575]$
$P - P^f$	$-\mathbf{0.175}\,[-0.266,\ -0.100]$	$-\mathbf{0.201}\,[-0.258,\ -0.152]$	$-\mathbf{0.520}\,[-0.621,\ -0.421]$	$-\mathbf{1.052}\,[-1.537,\ -0.616]$
$L - L^f$	$-0.052\,[-0.112,\ +0.012]$	$-0.049\,[-0.111,\ +0.011]$	$-\mathbf{0.239}\,[-0.395,\ -0.098]$	$-\mathbf{0.576}\,[-1.064,\ -0.186]$
$G - G^f$	$-\mathbf{0.188}\,[-0.225,\ -0.152]$	$-\mathbf{0.137}\,[-0.179,\ -0.101]$	$-\mathbf{0.410}\,[-0.540,\ -0.289]$	$-\mathbf{0.483}\,[-0.857,\ -0.189]$
GL − P	$-0.017\,[-0.055,\ +0.019]$	$+\mathbf{0.083}\,[+0.050,\ +0.119]$	$+0.021\,[-0.042,\ +0.086]$	$+0.136\,[-0.003,\ +0.321]$
GL − G	$+\mathbf{0.161}\,[+0.128,\ +0.193]$	$+\mathbf{0.104}\,[+0.078,\ +0.135]$	$+0.056\,[-0.004,\ +0.115]$	$+0.089\,[-0.084,\ +0.297]$
L − P	$-0.041\,[-0.101,\ +0.020]$	$+\mathbf{0.119}\,[+0.077,\ +0.164]$	$+\mathbf{0.173}\,[+0.102,\ +0.251]$	$+0.147\,[+0.031,\ +0.280]$