EXPLORATION-FREE LEARNING NEEDS AN EXACT PLUG-IN: INFORMATION FLOORS AND SHARP REGRET CONSTANTS FOR EVENT-TRIGGERED ESTIMATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Exploration is often viewed as the price of learning to act: a policy that simply plugs in its current parameter estimates can fail to gather the information it needs, and in online control of unknown linear systems minimax regret grows as $\sqrt{T}$. We study problems with an *information floor*, meaning that information about the unknown parameters arrives at a rate that no admissible policy can drive to zero, and show that this condition alone is not enough. Our main result is a dichotomy that is invisible in continuous time. Write r for the residual of the learner's plug-in map at the true parameter, i.e. the gap between the threshold it returns there and the optimum of the *sampled* problem. If $r = 0$, certainty equivalence attains regret $C_\Delta \log \tau + O(1)$ with an explicit constant; if $r \neq 0$, its regret is *linear* at rate $\frac{1}{2} J''_\Delta(\beta^*_\Delta)\, r^2$. The natural continuity-corrected map has $r = \Theta(\Delta)$ and is therefore linearly regretful, so a statistically consistent estimator does not suffice: the map it is plugged into must be consistent for the sampled comparator too. A van Trees bound over shrinking neighbourhoods matches C_Δ exactly for every *locally consistent* policy – one whose regret is $o(\tau)$ uniformly on a neighbourhood of the truth – so within that class C_Δ is the price of parameter uncertainty rather than an artefact of one algorithm. We also sharpen the underlying discretisation theory: the optimal sampled threshold is $\beta^*_\Delta = \beta^* - 0.5826\,\sigma\sqrt{\Delta} - \sigma^2\Delta/(8\beta^*) + o(\Delta)$, while the optimal cost rises by exactly $\lambda\sigma^2\Delta/6 + o(\Delta)$. We instantiate the framework twice, through event-triggered remote estimation over a lossy channel with unknown volatility and loss rate and through repeated sequential testing with unknown noise scale, and the same dichotomy appears in both. All cost functionals come from a deterministic Fredholm solve validated against Monte Carlo: across eight parameter settings the measured regret slope matches C_Δ with ratio 1.00 ± 0.11, and the predicted linear rates of inexact plug-ins are confirmed to within 5%.

1 INTRODUCTION

Most online learning algorithms pay for information. In linear-quadratic control, the learner's actions shape the data it sees, and naive certainty equivalence (CE) can stop exciting the directions it needs. The minimax regret for unknown dynamics is $\widetilde{\Theta}(\sqrt{T})$ (Simchowitz & Foster, 2020), with logarithmic rates available only under partial knowledge (Cassel et al., 2020) or more generally when the Fisher information of the optimal policy is nonsingular (Ziemann & Sandberg, 2021). The same theme recurs in bandits, where greedy policies are rate-optimal exactly when exogenous randomness supplies free exploration (Bastani et al., 2021; Kannan et al., 2018), and in operations, where certainty equivalence can converge to the wrong limit when the policy starves its own information (Keskin & Zeevi, 2018; Besbes & Muharremoglu, 2013).

This paper studies a problem at the opposite extreme, and shows that the resulting story is more delicate than "free information implies exploration-free learning".

Setting. A sensor observes a Wiener process with unknown diffusion coefficient σ^2, samples it at interval $\Delta > 0$, and decides when to transmit its current value over a channel whose per-attempt

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.

success probability q is unknown. Each attempt costs a fixed charge, attempts repeat until success, and estimation error is penalised quadratically. With known parameters the optimal policy is a symmetric threshold. The learner must discover (σ^2, q) while paying real costs.

The information floor. Source increments arrive at rate $1/\Delta$ no matter what the policy does, and every report reveals a geometric retransmission count. No admissible policy can drive either information rate to zero. In the terminology of Section 4, the problem has an information floor; this is the hypothesis under which exploration-free learning is usually expected to succeed.

The surprise. The information floor is not sufficient. Because the learner acts at sampling times, the object it must estimate is the optimum of the *sampled* problem, β^*_Δ, not the continuous-time optimum β^*. These differ at order $\sqrt{\Delta}$, and the classical continuity correction of Broadie et al. (1997) removes only the leading term. The residual $r = \Theta(\Delta)$ is enough to make regret linear:

plug-in map used by CE	threshold residual r	regret
continuous-time formula β^*	$\kappa\sigma\sqrt{\Delta}$	$\Theta(\tau)$, rate $\propto \Delta$
continuity-corrected $\beta^* - \kappa\sigma\sqrt{\Delta}$	$\sigma^2\Delta/(8\beta^*)$	$\Theta(\tau)$, rate $\propto \Delta^2$
exact sampled optimum β^*_Δ	0	$C_\Delta \log\tau + O(1)$

The three rows differ by six orders of magnitude in our experiments (Figure 1). The message is that "plug in your estimates" is underspecified, because what matters is the *order of accuracy of the map* into which the estimates are plugged.

Contributions.

1. **A sharp discretisation theory** (Prop. 2). We give the two-term expansion $\beta^*_\Delta = \beta^* - \kappa\sigma\sqrt{\Delta} - \sigma^2\Delta/(8\beta^*) + o(\Delta)$ with $\kappa = -\zeta(1/2)/\sqrt{2\pi} = 0.58260$, and show that the value loss from sampling is exactly $J^*_\Delta - J^* = \lambda\sigma^2\Delta/6 + o(\Delta)$. The policy is perturbed at order $\sqrt{\Delta}$ but the value only at order Δ.
2. **A dichotomy for plug-in learning** (Thm. 5). Under an information floor, local smoothness and a fixed deterministic map, CE attains $C_\Delta \log\tau + O(1)$ when the map's residual at the true parameter vanishes, and regret $\frac{1}{2}J''_\Delta(\beta^*_\Delta)\, r^2\, \tau(1 + o(1))$ when it equals $r \neq 0$. The information floor is therefore *not sufficient* for exploration-free logarithmic regret: consistency of the estimator does not imply consistency of the induced decision relative to the sampled comparator.
3. **How exact is exact enough** (Cor. 6). The two branches combine into $C_\Delta \log\tau + \frac{1}{2}J''r^2\tau$, giving a crossover horizon $\tau^*(r)$ below which a residual is harmless. This is what makes the dichotomy actionable rather than dichotomous in name: at our parameters τ^* is 1.4×10^2, 5.9×10^5 and 5.6×10^8 for the three maps we study.
4. **Explicit constants** (Thm. 7). C_Δ is computable from the exact sensitivities of the sampled optimum and the Fisher information of the two unknowns, with continuum limit $C = \frac{\bar{J}^*}{8}\big(2\Delta + x^*(1-q)/\sigma^2\big)$. The constant decomposes additively into a source-learning and a channel-learning term.
5. **Optimality of the constant** (Thm. 9). A local-minimax lower bound via the van Trees inequality (Gill & Levit, 1995) whose constant equals C_Δ exactly, for every locally consistent policy. Certainty equivalence is therefore not just rate-optimal but asymptotically exactly optimal, a Lai–Robbins-type statement for this problem.
6. **Two instances** (§6, §7). Event-triggered estimation, and repeated sequential testing with unknown noise scale, where the floor is exogenous and exact. In the second, the predicted logarithmic constant is confirmed to 0.2% and the predicted linear rate of the continuum map to 0.14%; the residual of the first-order correction is $10^{-3}\sigma$ there against $\Theta(\Delta)$ in the first, so the exactness of a plug-in map is a property to verify rather than assume.
7. **Fully computational evidence.** All cost functionals come from a deterministic Fredholm solve validated against Monte Carlo ($t = 0.63$ on the realised cost rate at the optimum). Across eight settings the measured regret slope matches C_Δ with ratio 1.00 ± 0.11, and the

predicted linear rates of inexact plug-ins are confirmed to within 5%. A tuned stochastic-approximation baseline attains the same rate class but a constant that depends on its step size; CE attains the information-theoretic constant with no tuning.

2 SETUP

Model. The source is a Wiener process W with diffusion coefficient $\sigma^2 > 0$, observed at times $t_j = j\Delta$. The receiver holds the last successfully delivered sample; the estimation error e_t is the increment of W since the last delivery, so e resets to zero at each report. A *report* triggered at a sampling time is retransmitted until success; the number of attempts is $A \sim \mathrm{Geom}(q)$ with unknown $q = (1-p)^2 \in (0,1]$, where p is the per-link erasure probability and both the forward and feedback links must succeed. Each attempt costs ρc_0; the estimation penalty is $\lambda \int e^2$. Over a renewal cycle of duration T,

$$C \;=\; \rho c_0 A \;+\; \lambda \int_0^T e_t^2 \, dt, \qquad J_\pi \;=\; \lim_{\tau\to\infty} \frac{1}{\tau}\, \mathbb{E}\Big[\textstyle\sum_{\text{cycles}} C\Big]. \tag{1}$$

Policy class. Π is the set of symmetric threshold policies π_β: report at the first sampling time with $|e| \geq \beta$, with β measurable with respect to the past and confined to a known compact $[\underline{\beta}, \overline{\beta}]$ (Assumption 1). Threshold policies are optimal for the known-parameter problem in continuous time (Åström & Bernhardsson, 2002; Rabi et al., 2012); we take this class as given and study learning within it.

Exact sampled cost. Write $y = \beta/(\sigma\sqrt{\Delta})$ for the threshold in units of one sampling standard deviation, let $N = \min\{j \geq 1 : |X_j| \geq y\}$ for a standard random walk X, and define

$$M(y) = \mathbb{E}_0[N], \qquad U(y) = \mathbb{E}_0\Big[\textstyle\sum_{j<N} X_j^2\Big]. \tag{2}$$

Both solve Fredholm equations of the second kind on $(-y, y)$ with the standard normal kernel (Appendix B); we solve them by a Nyström/Toeplitz scheme to eight digits. Then $\mathbb{E}[T] = \Delta M(y)$, $\mathbb{E}\int e^2 = \sigma^2\Delta^2\big(U(y) + M(y)/2\big)$, and by renewal-reward

$$J_\Delta(\beta) \;=\; \frac{\rho c_0}{q\,\Delta M(y)} \;+\; \lambda\sigma^2\Delta\Big(\frac{U(y)}{M(y)} + \frac{1}{2}\Big) \;=\; \lambda\sigma^2\Delta\Big[\frac{\theta + U(y)}{M(y)} + \frac{1}{2}\Big], \qquad \theta := \frac{\rho c_0}{q\lambda\sigma^2\Delta^2}. \tag{3}$$

The single dimensionless parameter θ governs the problem: $\beta^*_\Delta = \sigma\sqrt{\Delta}\, y^*(\theta)$ with $y^*(\theta) = \arg\min_y\{(\theta + U(y))/M(y)\}$. In the continuous-time limit $J_\Delta \to J(\beta) = \rho c_0\sigma^2/(q\beta^2) + \lambda\beta^2/6$, minimised at $\beta^* = (6\rho c_0\sigma^2/(q\lambda))^{1/4}$, with excess cost $\bar{J}^* = \sqrt{2\rho\lambda c_0\sigma^2/(3q)}$ and $x^* = (\beta^*)^2$.

Regret. Against the best *sampled* threshold,

$$\mathrm{Regret}(\tau) \;=\; \mathbb{E}\Big[\textstyle\sum_{\text{cycles in } [0,\tau]} C\Big] - \tau\, J^*_\Delta, \qquad J^*_\Delta = J_\Delta(\beta^*_\Delta). \tag{4}$$

This comparator, rather than the continuous-time optimum, is the honest one: no sampled policy can beat J^*_Δ, and Prop. 2 quantifies the gap to J^* exactly.

Assumption 1 (compactness and interior truth)**.** (σ^2, q) lies in the interior of a known compact $\Theta_0 \subset (0,\infty) \times (0,1]$, and all estimates are projected onto Θ_0.

3 SAMPLING PERTURBS THE POLICY AT $\sqrt{\Delta}$ AND THE VALUE AT Δ

Lemma 1 (corrected diffusion expansions)**.** *As $y \to \infty$, with $\tilde{y} = y + \kappa$ and $\kappa = -\zeta(1/2)/\sqrt{2\pi}$,*

$$M(y) = \tilde{y}^2 + \tfrac{1}{4} + o(1), \qquad \frac{U(y)}{M(y)} = \frac{\tilde{y}^2}{6} - \frac{7}{24} + o(1). \tag{5}$$

The leading terms are the classical overshoot corrections of Asmussen et al. (1995); Broadie et al. (1997); the constants $1/4$ and $-7/24$ are the next order and are what the value expansion needs. Appendix A gives the derivation and Appendix B the numerical verification ($a = 0.2502$, $b = -0.2898$ at $y = 60$, converging to $1/4$ and $-7/24$).

Proposition 2 (discretisation, sharp)**.** *Under Assumption 1, as $\Delta \to 0$ with the other parameters fixed,*

$$\beta^*_\Delta \;=\; \beta^* - \kappa\,\sigma\sqrt{\Delta} \;-\; \frac{\sigma^2\Delta}{8\beta^*} \;+\; o(\Delta), \qquad J^*_\Delta - J^* \;=\; \frac{\lambda\sigma^2\Delta}{6} \;+\; o(\Delta). \tag{6}$$

Proof sketch. Substituting Lemma 1 into (3) gives $J_\Delta/(\lambda\sigma^2\Delta) = \theta/(\tilde{y}^2 + \frac{1}{4}) + \tilde{y}^2/6 + \frac{5}{24} + o(1)$. The stationarity condition is $(\tilde{y}^2 + \frac{1}{4})^2 = 6\theta$, i.e. $\tilde{y}^2 = \sqrt{6\theta} - \frac{1}{4}$, whence $\tilde{y} = (6\theta)^{1/4} - \frac{1}{8}(6\theta)^{-1/4} + o(\theta^{-1/4})$; undoing the scaling $\beta = \sigma\sqrt{\Delta}\,y$ and using $(6\theta)^{1/4} = \beta^*/(\sigma\sqrt{\Delta})$ gives the first claim. Evaluating the objective at the optimum gives $J^*_\Delta/(\lambda\sigma^2\Delta) = \sqrt{2\theta/3} - \frac{1}{24} - \frac{7}{24} + \frac{1}{2} + o(1)$ while $J^*/(\lambda\sigma^2\Delta) = \sqrt{2\theta/3}$, and $-\frac{1}{24} - \frac{7}{24} + \frac{1}{2} = \frac{1}{6}$. □

Remark 3. The value gap $\lambda\sigma^2\Delta/6$ has a transparent reading: it is exactly the mean squared error accumulated within one sampling interval, net of the shorter cycle the lower threshold produces. Sampling is cheap in value and expensive in policy, which is precisely why a learner that targets the wrong threshold is punished.

4 An information floor is necessary but not sufficient

Definition 4 (information floor)**.** A parametric control problem with unknown $\vartheta \in \Theta_0 \subset \mathbb{R}^d$ and admissible policies Π has an *information floor* at ϑ if there exist $c > 0$, t_0 and $I_0 \succ 0$ such that the Fisher information accrued by any $\pi \in \Pi$ satisfies $\mathcal{I}_{\pi,t}(\vartheta) \succeq c\,t\,I_0$ for all $t \geq t_0$.

In our problem the floor holds with $I_0 = \mathrm{diag}\big(1/(2\sigma^4\Delta),\; 1/(q^2(1-q)\,\mathbb{E}_{\overline{\beta}}[T])\big)$: increments accrue at rate $1/\Delta$ regardless of the policy, and reports accrue at rate at least $1/\mathbb{E}_{\overline{\beta}}[T] > 0$ because thresholds are confined to a compact. Note that the report rate *is* policy-dependent (a smaller threshold buys information about q faster), so what the floor asserts is that no policy can starve the channel, not that the policy is irrelevant.

Assumption 2 (local smoothness)**.** The optimal action $a^*(\vartheta)$ is C^1 near the truth, and $J(a;\vartheta) - J(a^*(\vartheta);\vartheta) = \frac{1}{2}J''(a^*)(a-a^*)^2 + o(|a-a^*|^2)$ with $J''(a^*) > 0$.

The theorem below uses, in addition to Definition 4 and Assumptions 1–2: a one-dimensional action with a unique minimiser of $J(\cdot;\vartheta)$ in the interior of the admissible compact; a *fixed, deterministic, continuous* plug-in map $\widehat{a}$; an estimator that is efficient for ϑ under the floor; and $J(\cdot;\vartheta)$ twice continuously differentiable near a^*. Maps whose residual varies with the horizon are covered only through Corollary 6.

Theorem 5 (dichotomy)**.** *Let Assumptions 1–2 and Definition 4 hold, and let the learner play $a_t = \widehat{a}(\hat{\vartheta}_t)$ for an efficient estimator $\hat{\vartheta}_t$ and a deterministic map $\widehat{a}$.*

- *(i) If $\widehat{a}(\vartheta) = a^*(\vartheta)$ (exact map), then $\limsup_\tau \mathrm{Regret}(\tau)/\log\tau \leq \frac{1}{2}J''(a^*)\,\nabla a^{*\top}\,(c\,I_0)^{-1}\nabla a^*$, in particular $O(\log\tau)$.*
- *(ii) If $r := \widehat{a}(\vartheta) - a^*(\vartheta) \neq 0$ (inexact map), then $\mathrm{Regret}(\tau) = \frac{1}{2}J''(a^*)\,r^2\,\tau\,(1+o(1))$, i.e. linear.*

No forced exploration, optimism or step size is used in either case.

Corollary 6 (how exact is exact enough)**.** *Combining the two branches, a plug-in with residual r satisfies $\mathrm{Regret}(\tau) \leq C_\Delta \log\tau + \frac{1}{2}J''(a^*)r^2\tau + O(1)$, so its regret is logarithmic for every horizon below $\tau^*(r)$, the solution of $\frac{1}{2}J''r^2\tau = C_\Delta\log\tau$, and linear beyond it. Exactness is therefore not an all-or-nothing requirement: what matters is r relative to the horizon. At our parameters, $\tau^* = 1.4\times10^2$ for the uncorrected map, 5.9×10^5 for the continuity-corrected map in Instance 1, and 5.6×10^8 for the corrected map in Instance 2, which is why the latter looks logarithmic throughout Figure 4 and the former does not in Figure 1.*

Part (ii) is elementary but is exactly what the literature's continuous-time framing hides: an estimator can be perfectly consistent while the policy it induces is inconsistent for the comparator. Part (i) is the standard plug-in argument made quantitative by the floor.

5 CERTAINTY EQUIVALENCE WITH AN EXACT PLUG-IN

Algorithm 1 (CE-Exact). At each report i with elapsed time t:
(1) $\hat{\sigma}^2 = \sum_{j \le N} (\Delta W_j)^2/(N\Delta)$ over *all* increments, $N = t/\Delta$;
(2) $\hat{q} = i/\sum_{j \le i} A_j$ (geometric MLE); project $(\hat{\sigma}^2, \hat{q})$ onto Θ_0;
(3) $\hat{\theta} = \rho c_0/(\hat{q}\lambda\hat{\sigma}^2\Delta^2)$ and $\hat{\beta} = \sqrt{\hat{\sigma}^2\Delta}\; y^*(\hat{\theta})$, with y^* precomputed once by the Fredholm solve (a one-dimensional table, relative accuracy 10^{-9});
(4) report at the first sampling time with $|e| \ge \hat{\beta}$.

Theorem 7 (regret of CE-Exact)**.** *Under Assumption 1, CE-Exact satisfies* $\mathrm{Regret}(\tau) \le C_\Delta \log\tau + O(1)$ *with*

$$C_\Delta = \tfrac{1}{2} J''_\Delta(\beta^*_\Delta)\Big[\, 2\Delta\, s_\sigma^2 + (1-q)\,\mathbb{E}[T]\, s_q^2 \Big], \qquad s_\sigma = \frac{\partial \beta^*_\Delta}{\partial \log \sigma^2}, \quad s_q = \frac{\partial \beta^*_\Delta}{\partial \log q}, \tag{7}$$

all quantities computable from the Fredholm solve. In the continuum limit $s_\sigma \to \beta^*/4$, $s_q \to -\beta^*/4$, $J''_\Delta \to 4\bar{J}^*/x^*$, $\mathbb{E}[T] \to x^*/\sigma^2$, *and*

$$C_\Delta \;\longrightarrow\; C = \frac{\bar{J}^*}{8}\Big(2\Delta + \frac{x^*(1-q)}{\sigma^2}\Big). \tag{8}$$

The two terms separate the price of not knowing the source from the price of not knowing the channel. The source term is $O(\Delta)$ because quadratic variation makes σ^2 almost free, so the channel term dominates whenever $\Delta \ll x^*(1-q)/\sigma^2$.

Remark 8 (a constant that is easy to get wrong)**.** The geometric MLE $\hat{q} = i/\sum_j A_j$ has Fisher information $1/(q^2(1-q))$ per report, so its relative variance is $(1-q)/i$, *not* $(1-q)/(qi)$. The latter inflates C by up to $2.04\times$ over the loss probabilities we study ($p \in [0.05, 0.3]$, i.e. $q \in [0.49, 0.9]$); our experiments reject it decisively (Figure 3).

Theorem 9 (local-minimax lower bound with the exact constant)**.** *Call a policy* locally consistent *if its regret is* $o(\tau)$ *uniformly over a neighbourhood of* $\vartheta_0 = (\sigma^2, q)$*; by Theorem 5(ii) this excludes exactly the linearly-regretful plug-ins. For every locally consistent policy and* $\varepsilon_\tau = (\log\tau)^{-1/2}$,

$$\liminf_{\tau\to\infty} \sup_{|\vartheta - \vartheta_0| \le \varepsilon_\tau} \frac{\mathrm{Regret}_\vartheta(\tau)}{\log\tau} \;\ge\; C_\Delta(\vartheta_0), \tag{9}$$

with C_Δ *exactly the constant of Theorem 7. Consequently* $\mathrm{Regret}^{\text{CE-Exact}}(\tau)/\log\tau \to C_\Delta$ *is optimal:* C_Δ *is the asymptotic price of not knowing* (σ^2, q)*, not merely an upper bound for one algorithm.*

The proof (Appendix A) reduces regret to a weighted estimation risk by strong convexity, bounds the Fisher information accrued by an *arbitrary* policy, and applies the van Trees inequality (Gill & Levit, 1995) on a shrinking ball. Two features of the problem make the constant, and not merely the rate, come out. First, the source channel is entirely exogenous: every policy accrues exactly $t/(2\sigma^4\Delta)$ units of information about σ^2 by time t, with no slack. Second, channel information arrives only through completed reports, and a locally consistent policy must report at asymptotic rate $1/\mathbb{E}[T(\beta^*_\Delta)]$; reporting more often would buy information faster but is precisely what consistency forbids. The restriction to locally consistent policies is analogous to the uniformly-good restriction used in classical asymptotic regret lower bounds (Lai & Robbins, 1985); we claim the analogy for the role of the restriction, not an equivalence of settings.

6 EXPERIMENTS

Every cost functional is computed, not fitted: M and U come from the Fredholm solve and the regret is accrued as $\sum_{\text{cycles}} \mathbb{E}[T(\hat{\beta})]\,(J_\Delta(\hat{\beta}) - J^*_\Delta)$: the conditional expected regret contribution of each cycle, evaluated from the Fredholm solution rather than sampled. It is unbiased for (4) and has far lower variance than realised cost. Three validations support this: the solver matches Monte Carlo ($y = 2, 4, 7$, within 1–2 standard errors); the realised cost rate of the oracle policy matches J^*_Δ ($t = 0.63$, 200 seeds); and realised regret matches the accrued regret for CE-Exact (7.26 ± 198

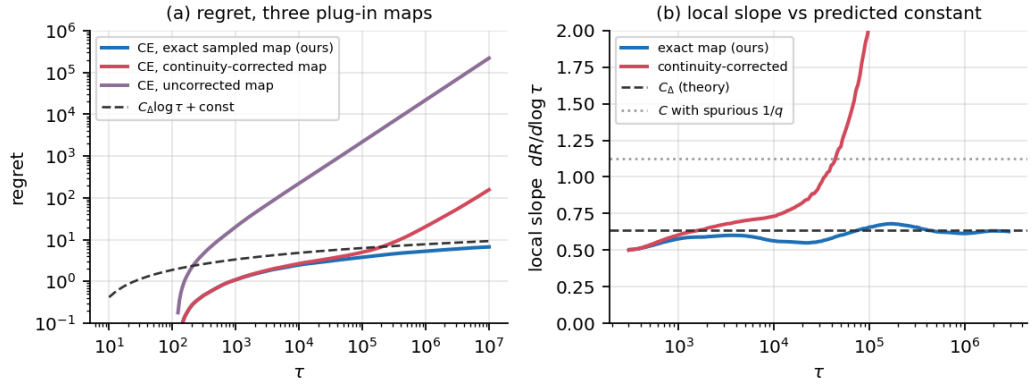


Figure 1: **The dichotomy: the error of the plug-in map, not the information structure, sets the regret order.** All three learners see the same data and use the same consistent estimators; they differ only in the map the estimates are plugged into, and hence in the residual r at the true parameter. (a) Regret of CE under three plug-in maps; the uncorrected curve leaves the panel at $\tau \approx 3 \times 10^3$. Sample-exact map: logarithmic. Continuity-corrected map (residual $\sigma^2\Delta/8\beta^*$): linear at rate 1.51×10^{-5} (predicted 1.44×10^{-5}). Uncorrected map (residual $\kappa\sigma\sqrt{\Delta}$): linear at rate 2.20×10^{-2} (predicted 2.26×10^{-2}). (b) Local slope $dR/d\log\tau$ against the predicted constant $C_\Delta = 0.638$ and against the variant with a spurious $1/q$ (1.123).

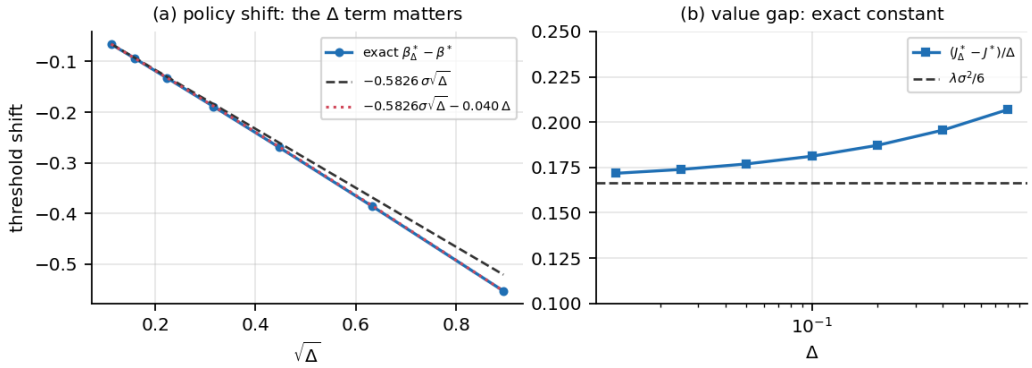


Figure 2: **Proposition 2.** (a) The exact shift $\beta^*_\Delta - \beta^*$ against $-\kappa\sigma\sqrt{\Delta}$ and against the two-term expansion; the Δ term is what the learner must not miss. (b) $(J^*_\Delta - J^*)/\Delta$ converges to $\lambda\sigma^2/6$.

versus 7.42 – same mean, 27× more noise). The long-horizon runs use the accrual estimator with the sufficient statistics evolved exactly and cycle lengths replaced by their conditional means, which is what makes $\tau = 10^7$ affordable; the path-level simulator, which makes no such replacement, agrees at the horizons where both are feasible ($\tau \leq 2 \times 10^5$). Baseline parameters: $\sigma^2 = 1$, $\Delta = 0.1$, $\lambda = \rho = 1$, $q = 0.5625$ ($p = 0.25$), $\beta^* = 2.4$, 60 seeds.

What "exact" means here. The map is *sample-exact*: it returns the minimiser of $J_\Delta(\cdot; \hat{\vartheta})$ as computed by the Fredholm solve, to relative tolerance 10^{-9}. That tolerance is itself a residual, so by Corollary 6 it is not exactly zero, but it induces a crossover beyond $\tau \approx 10^{18}$, some twelve orders of magnitude past our horizons, which is why we treat it as exact and why the fitted linear coefficient for this map sits at the noise floor of the fit rather than at its predicted value.

Baselines. A two-timescale stochastic-approximation learner (Robbins–Monro on the average cost, with the same exact map) is the natural competitor, and is what the age-of-information literature uses (Tang et al., 2024; He et al., 2023). Over the horizons we ran it does not diverge, but its regret is

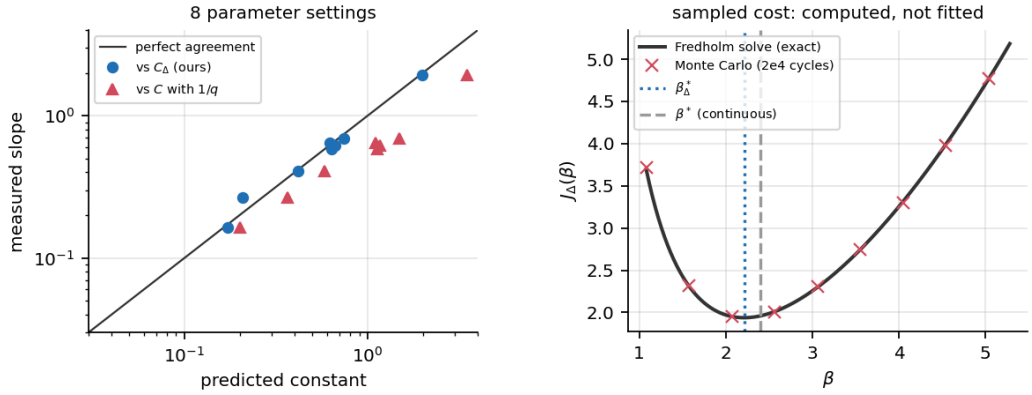


Figure 3: Two independent checks. **Left (constant):** measured regret slope versus prediction across eight settings ($q \in \{0.49, 0.5625, 0.7225, 0.9\}$, $\Delta \in \{0.05, 0.1, 0.2\}$, $\beta^* \in \{1.8, 2.4, 3.2\}$); ratio to C_Δ is 1.00 ± 0.11 (the largest deviation, 1.28, is the smallest-threshold setting $\beta^* = 1.8$, where the constant is smallest and the fitting window shortest), ratio to the $1/q$ variant is 0.61 ± 0.11. **Right:** the sampled cost $J_\Delta(\beta)$ from the Fredholm solve against Monte Carlo.

several times larger and depends strongly on the step size: at $\tau = 3 \times 10^4$ the accrued regret is 19.1 ($a_0 = 0.2$), 58.2 ($a_0 = 0.6$) and 182.1 ($a_0 = 0.05$), against 4.56 for CE-Exact, which has no step size to tune; at $\tau = 10^6$ the best-tuned variant reaches 37.1 against ≈ 5.5. We do not claim a rate for it: establishing its asymptotic constant would require its own analysis. A Kiefer–Wolfowitz variant is far worse still, since finite-difference probes pay the quadratic cost of deviating from the optimum.

7 A SECOND INSTANCE: REPEATED SEQUENTIAL TESTING

To show that Definition 4 and Theorem 5 are not artefacts of the remote-estimation model, we instantiate them in a decision problem with a different loss, a different optimal-action geometry, and a different literature.

Model. Observations $X_1, X_2, \ldots$ arrive i.i.d. $\mathcal{N}(s\mu, \sigma^2)$ with an unknown sign $s \in \{-1, +1\}$, μ known and σ^2 unknown (we reserve ϑ for the parameter that is *learned across* tests, here σ^2). The learner runs a sequential probability ratio test on the running sum: stop at $N = \min\{n : |S_n| \geq B\}$ and decide $\operatorname{sign}(S_N)$, paying c per observation and λ per error, so the per-test cost is $J(B) = c\,\mathbb{E}[N] + \lambda\,\mathbb{P}(\text{error})$. Tests repeat indefinitely and regret is measured against the best B for the true σ^2.

The floor is exogenous and exact. Every observation is informative about σ^2 whatever B is, so $\mathcal{I}_n(\sigma^2) = n/(2\sigma^4)$ for *every* policy: the floor of Definition 4 holds with no slack at all. Here there is no channel term and no policy-dependent information, so the hypothesis of Theorem 5 is satisfied in its purest form.

Three maps. The exact operating characteristics $\mathbb{E}[N]$ and $\mathbb{P}(\text{error})$ solve Fredholm equations of the second kind with a drifted Gaussian kernel (Appendix B); the solver matches Monte Carlo to within one standard error. Against the exact optimum B^*, Wald's continuum approximation $\mathbb{E}[N] \approx (B/\mu)\tanh(\mu B/\sigma^2)$, $\mathbb{P}(\text{err}) \approx (1 + e^{2\mu B/\sigma^2})^{-1}$ gives a threshold that is too large by $+0.5812\,\sigma$, almost exactly the overshoot constant $\kappa\sigma$, and Siegmund's corrected diffusion approximation $B_{\text{Wald}} - \kappa\sigma$ overshoots by only $-0.0014\,\sigma$ (resolution-independent; see Appendix B).

Results. Table 1 and Figure 4 report CE with each map, 40 seeds, n up to 10^7 observations.

What the two instances say together. The dichotomy holds in both: a plug-in with a nonvanishing residual is linearly regretful, and the exact map attains $C \log n$ with the constant predicted by the

plug-in map	residual r/σ	predicted linear rate	measured c	measured $\log$ slope b
exact sampled	0	0	1.6×10^{-7} (noise)	2.7275
Siegmund-corrected	-0.0014	9.85×10^{-8}	2.1×10^{-7} (noise)	2.7463
Wald (continuum)	$+0.5812$	1.3979×10^{-2}	1.3998×10^{-2}	–

Table 1: Instance 2. Fits of $R(n) = a + b \log n + c\, n$ over $n \geq 10^4$. Predicted log constant $C = 2.7212$ against measured $b = 2.7275$ (0.2%); predicted linear rate for the Wald map 1.3979×10^{-2} against measured 1.3998×10^{-2} (0.14%). The exact map's fitted c is the noise floor of the fit, which is what the Siegmund map's genuine $c = 9.85 \times 10^{-8}$ also sits at: its linear term is invisible below $n \approx 10^8$.

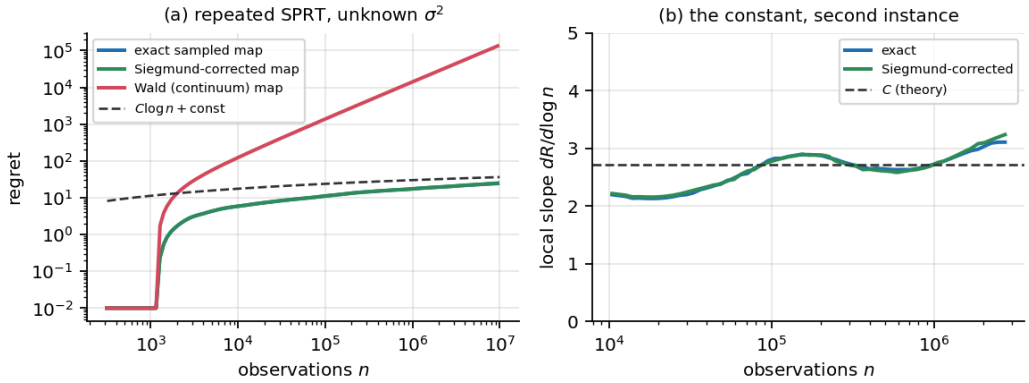


Figure 4: Instance 2. (a) CE regret under the three maps; the sample-exact curve is hidden beneath the Siegmund-corrected one (their residuals differ by $10^{-3}\sigma$, whose linear term is invisible here), while the Wald map is linear. (b) Local slope against the predicted constant $C = 2.7212$.

Fisher information. But the *size* of the second-order residual is problem-specific. In remote estimation the first-order continuity correction leaves a residual of order Δ, which is large enough to dominate the logarithmic term by $\tau \approx 10^5$; in sequential testing the same correction leaves a residual of order $10^{-3}\sigma$, whose linear term is invisible until $n \approx 10^8$. The practical prescription is therefore not "first-order corrections fail" but "exactness of the map is a property to verify, not to assume", and the verification is exactly the solve that produces the map.

8 Related work

Learning to sample. Tang et al. (2024) prove $O(\ln k)$ regret with a matching $\Omega(\ln k)$ lower bound for online sampling of a Wiener process under unknown delay statistics; He et al. (2023) do the same for an unreliable channel with unknown loss probability, and Chen et al. (2023) for Ornstein–Uhlenbeck sources. Our rate is theirs; what is new is (i) the exactness requirement on the plug-in map and the resulting dichotomy, which is invisible when the analysis is carried out in continuous time, (ii) explicit constants with a source/channel decomposition, and (iii) the sharp discretisation theory of Section 3. We also use certainty equivalence rather than stochastic approximation, which removes step-size tuning.

Exploration-free learning. That greedy or certainty-equivalent policies are rate-optimal when information arrives regardless of the action is known in contextual bandits (Bastani et al., 2021; Kannan et al., 2018), inventory (Besbes & Muharremoglu, 2013), pricing (Keskin & Zeevi, 2018), and control (Mania et al., 2019; Jedra & Proutiere, 2022). Ziemann & Sandberg (2021) formalise the dividing line through the Fisher information of the optimal policy, and Basei et al. (2022) obtain logarithmic regret for greedy least-squares in episodic continuous-time linear-quadratic problems, a setting in which the exactness question we raise cannot arise, because the action is chosen in continuous time. Definition 4 is the same idea stated as a uniform floor over the admissible class; our

contribution is to show that this condition, which the literature treats as the crux, does not by itself deliver logarithmic regret once actions are taken on a discrete grid.

Sequential analysis (Instance 2). The SPRT and its operating characteristics are classical (Wald, 1947); testing with a nuisance scale parameter is the subject of the sequential-t literature surveyed by Ghosh (1970). What our second instance uses from that tradition is the boundary correction: Wald's approximations are first-order in the overshoot, and Siegmund (1979) gives the corrected diffusion approximations that remove the leading term. Our contribution there is not a new test but the observation that the residual left by that correction is what determines the regret order of a learner that re-solves the test as its estimate of the nuisance parameter improves.

Event-triggered estimation. Threshold optimality for Brownian sources goes back to Åström & Bernhardsson (2002) and Rabi et al. (2012); packet drops and scheduling are treated by Chakravorty et al. (2017) and the age-of-information line (Sun et al., 2017; Ornee & Sun, 2021). Continuity corrections for discretely monitored barriers are due to Asmussen et al. (1995); Broadie et al. (1997).

9 LIMITATIONS

Assumption 1 (known compact parameter set with interior truth) does real work: it is what makes the information floor uniform. The source is Wiener and the loss i.i.d.; bursty or Markov channels change the estimator but not, we expect, the dichotomy. The lower bound matches the constant only within the class of locally consistent policies; whether some linearly-regretful policy could do better over a fixed (non-shrinking) neighbourhood at small τ is not addressed. The general statement (Theorem 5) is proved for one-dimensional actions with a unique smooth optimum. The vector-source case is open rather than routine: our own dichotomy implies that a vector plug-in built on the continuum ellipsoid is linearly regretful, and the exact sampled trigger region in k dimensions requires a k-dimensional Fredholm solve with a vector overshoot correction, which we do not have. We therefore make no vector claim.

REPRODUCIBILITY STATEMENT

Assumptions are stated in Section 2; complete proofs are in Appendix A; the numerical scheme, its convergence study and all parameter values are in Appendix B; an anonymised repository containing seeded code that regenerates every figure and table with one command is included in the supplementary material.

AI USE STATEMENT

In accordance with the ICLR 2027 AI Policy for Authors, we disclose that a large language model (Claude) was used substantially in this work. The research question, the problem setting and the original manuscript on which this work builds were developed by the authors. Among the tasks requiring disclosure under the ICLR policy, we used generative AI tools to help develop the theoretical framework (the information-floor formulation and the residual-based dichotomy); formulate mathematical claims (including Theorem 5, Corollary 6 and Lemma 10); provide critical ingredients for proofs (the martingale identity in Lemma 1 and the localisation and clipping argument in the proof of Theorem 9); assist in writing the proofs; propose and refine hypotheses; design and give feedback on the numerical experiments; implement all methods, namely the Fredholm solvers, the Monte Carlo simulators, the learning algorithms and the baselines; generate the synthetic (simulated) data, which were produced by that code; and assist in interpreting the results. We did not use generative AI tools for translation. Cleaning or reformatting datasets and qualitative or thematic data analysis were not applicable, as this work uses no external data. Among the tasks for which disclosure is recommended, we additionally used generative AI tools to write and edit all software code, create the figures, draft and edit the paper, search for and summarize related literature, format references, and suggest the paper's structure and title. We reviewed all AI-assisted work. Every cited work was read in the original source, every derivation was independently checked, and every reported numerical result is the output of the released seeded code and was cross-checked against independent Monte Carlo estimates (Appendix B). No result was fabricated or manually altered. We take full

responsibility for the final content of this work, including all text, claims, analyses, code and other artifacts produced with the assistance of generative AI.

Ethics statement

This work is theoretical and simulation-based. It involves no human subjects, no personal data and no sensitive datasets. The authors declare no conflicts of interest and no funding.

References

Søren Asmussen, Peter Glynn, and Jim Pitman. Discretization error in simulation of one-dimensional reflecting Brownian motion. *Annals of Applied Probability*, 5(4):875–896, 1995.

Karl Johan Åström and Bo Bernhardsson. Comparison of Riemann and Lebesgue sampling for first order stochastic systems. In *IEEE Conference on Decision and Control (CDC)*, 2002.

Matteo Basei, Xin Guo, Anran Hu, and Yufei Zhang. Logarithmic regret for episodic continuous-time linear-quadratic reinforcement learning over a finite-time horizon. *Journal of Machine Learning Research*, 23:1–34, 2022.

Hamsa Bastani, Mohsen Bayati, and Khashayar Khosravi. Mostly exploration-free algorithms for contextual bandits. *Management Science*, 67(3):1329–1349, 2021.

Omar Besbes and Alp Muharremoglu. On implications of demand censoring in the newsvendor problem. *Management Science*, 59(6):1407–1424, 2013.

Mark Broadie, Paul Glasserman, and Steven Kou. A continuity correction for discrete barrier options. *Mathematical Finance*, 7(4):325–349, 1997.

Asaf Cassel, Alon Cohen, and Tomer Koren. Logarithmic regret for learning linear quadratic regulators efficiently. In *International Conference on Machine Learning (ICML)*, 2020.

Jhelum Chakravorty, Jayakumar Subramanian, and Aditya Mahajan. Stochastic approximation based methods for computing the optimal thresholds in remote-state estimation with packet drops. In *American Control Conference (ACC)*, 2017.

Joseph T. Chang and Yuval Peres. Ladder heights, Gaussian random walks and the Riemann zeta function. *Annals of Probability*, 25(2):787–802, 1997.

Yuchao Chen, Haoyue Tang, Jintao Wang, Pengkun Yang, and Leandros Tassiulas. Sampling for remote estimation of an Ornstein–Uhlenbeck process through channel with unknown delay statistics. *Journal of Communications and Networks*, 2023. doi:10.23919/JCN.2023.000037.

Bhaskar Kumar Ghosh. *Sequential Tests of Statistical Hypotheses*. Addison-Wesley, 1970.

Richard D. Gill and Boris Y. Levit. Applications of the van Trees inequality: a Bayesian Cramér–Rao bound. *Bernoulli*, 1(1–2):59–79, 1995.

Hongyi He, Haoyue Tang, Jiayu Pan, Jintao Wang, Jian Song, and Leandros Tassiulas. Age optimal sampling for unreliable channels under unknown channel statistics. In *21st International Symposium on Modeling and Optimization in Mobile, Ad Hoc, and Wireless Networks (WiOpt)*, pp. 651–656, 2023.

Yassir Jedra and Alexandre Proutiere. Minimal expected regret in linear quadratic control. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2022.

Sampath Kannan, Jamie Morgenstern, Aaron Roth, Bo Waggoner, and Zhiwei Steven Wu. A smoothed analysis of the greedy algorithm for the linear contextual bandit problem. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018.

N. Bora Keskin and Assaf Zeevi. On incomplete learning and certainty-equivalence control. *Operations Research*, 66(4):1136–1167, 2018.

Tze Leung Lai and Herbert Robbins. Asymptotically efficient adaptive allocation rules. *Advances in Applied Mathematics*, 6(1):4–22, 1985.

Horia Mania, Stephen Tu, and Benjamin Recht. Certainty equivalence is efficient for linear quadratic control. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

Tasmeen Zaman Ornee and Yin Sun. Sampling and remote estimation for the Ornstein–Uhlenbeck process through queues. *IEEE/ACM Transactions on Networking*, 2021.

Maben Rabi, George V. Moustakides, and John S. Baras. Adaptive sampling for linear state estimation. *SIAM Journal on Control and Optimization*, 50(2):672–702, 2012.

David Siegmund. Corrected diffusion approximations in certain random walk problems. *Advances in Applied Probability*, 11(4):701–719, 1979.

Max Simchowitz and Dylan J. Foster. Naive exploration is optimal for online LQR. In *International Conference on Machine Learning (ICML)*, 2020.

Yin Sun, Elif Uysal-Biyikoglu, Roy D. Yates, C. Emre Koksal, and Ness B. Shroff. Update or wait: how to keep your data fresh. *IEEE Transactions on Information Theory*, 63(11), 2017.

Haoyue Tang, Yin Sun, and Leandros Tassiulas. Sampling of the Wiener process for remote estimation over a channel with unknown delay statistics. *IEEE/ACM Transactions on Networking*, 32(3): 1920–1935, 2024.

Abraham Wald. *Sequential Analysis*. John Wiley & Sons, 1947.

Ingvar Ziemann and Henrik Sandberg. On uninformative optimal policies in adaptive LQR with unknown B-matrix. In *Learning for Dynamics and Control (L4DC)*, 2021.

A Proofs

A.1 Notation

Throughout, $X_j = \sum_{k\le j}\xi_k$ is a standard random walk ($\xi_k \sim \mathcal{N}(0,1)$ i.i.d.), $N = \min\{j \ge 1 : |X_j| \ge y\}$, and M, U are as in (2). Physical quantities are recovered by $e_{j\Delta} = \sigma\sqrt{\Delta}\, X_j$, $\beta = \sigma\sqrt{\Delta}\, y$. We write $\kappa = -\zeta(1/2)/\sqrt{2\pi} = 0.5825972$ and $\tilde{y} = y + \kappa$.

A.2 Renewal-reward form of the sampled cost

Between reports the error is a random walk started at 0; attempts $A \sim \mathrm{Geom}(q)$ are independent of the path, so $\mathbb{E}[C] = \rho c_0/q + \lambda\,\mathbb{E}\int_0^T e^2$. For the integral, condition on consecutive samples: for Brownian motion on $[0,\Delta]$ with endpoints a, b, $\mathbb{E}\big[\int_0^\Delta W_t^2 dt \,\big|\, W_0 = a, W_\Delta = b\big] = \Delta\big[(a^2 + ab + b^2)/3 + \sigma^2\Delta/6\big]$, and taking expectations over b given a gives $\Delta(a^2 + \sigma^2\Delta/2)$. Summing over the N steps of a cycle and using Wald's identity,

$$\mathbb{E}\int_0^T e^2 = \Delta\,\sigma^2\Delta\,U(y) + \tfrac{\sigma^2\Delta^2}{2}M(y) = \sigma^2\Delta^2\Big(U(y) + \tfrac{M(y)}{2}\Big), \qquad \mathbb{E}[T] = \Delta M(y),$$

which is (3). Renewal-reward applies because cycles are i.i.d. given the threshold.

A.3 Proof of Lemma 1

The proof has two parts: an *exact* identity obtained from a martingale, and the classical expansion of the overshoot moments. No asymptotic input beyond the latter is used.

Step 1: an exact identity for U. For Gaussian steps the one-step operator is $\mathbb{E}[f(x+\xi)] = e^{\frac{1}{2}\partial_x^2} f(x)$. Solving $(e^{\frac{1}{2}\partial^2} - 1)f = -x^2$ in the space of polynomials gives, uniquely up to constants,

$$f(x) = -\tfrac{x^4}{6} + \tfrac{x^2}{2}, \qquad \text{since } e^{\frac{1}{2}\partial^2} x^4 = x^4 + 6x^2 + 3, \quad e^{\frac{1}{2}\partial^2} x^2 = x^2 + 1.$$

Hence $f(X_j) + \sum_{k<j} X_k^2$ is a martingale, and optional stopping at N (legitimate: $\mathbb{E}[N^2] < \infty$ and X_N has Gaussian-type tails uniformly in y) gives the *exact* relations

$$M(y) = \mathbb{E}\big[X_N^2\big], \qquad U(y) = \tfrac{1}{6}\mathbb{E}\big[X_N^4\big] - \tfrac{1}{2}\mathbb{E}\big[X_N^2\big]. \tag{10}$$

The first is optional stopping on $X_j^2 - j$. We verified (10) numerically against both the Fredholm solver and Monte Carlo (Appendix B).

Step 2: overshoot moments. Write $X_N = \pm(y + \mathcal{O}_y)$ with $\mathcal{O}_y \geq 0$ the overshoot. By the renewal theorem for the ascending ladder-height process of a mean-zero, unit-variance walk, $\mathcal{O}_y \Rightarrow \mathcal{O}_\infty$, the stationary (equilibrium) law of the ladder height, and the convergence holds with all moments because Gaussian ladder heights have exponential tails, giving uniform integrability of every $\mathcal{O}_y^k$. For the Gaussian walk (Chang & Peres, 1997),

$$\mathbb{E}[\mathcal{O}_\infty] = \kappa = -\zeta(1/2)/\sqrt{2\pi} = 0.5825972, \qquad v := \mathrm{Var}(\mathcal{O}_\infty) = \tfrac{1}{4}.$$

Both constants are computed in Appendix B to six digits by an integral equation *independent* of the ones used elsewhere in the paper ($\mathbb{E}[\mathcal{O}] = 0.582597$, $\mathrm{Var}(\mathcal{O}) = 0.250000$, stable over $y \in \{6, 8, 10, 14, 16\}$); the first reproduces the Chernoff–Siegmund constant to six digits, which is a check on the method.

Step 3: assembling. Writing $m_k = \mathbb{E}[\mathcal{O}_\infty^k]$ and $\tilde{y} = y + m_1$,

$$\mathbb{E}[X_N^2] = \tilde{y}^2 + v + o(1), \qquad \mathbb{E}[X_N^4] = \tilde{y}^4 + 6y^2 v + 4y(m_3 - m_1^3) + O(1),$$

so by (10),

$$M(y) = \tilde{y}^2 + v + o(1), \qquad \frac{U(y)}{M(y)} = \frac{\tilde{y}^2}{6} + \frac{5v}{6} - \frac{1}{2} + o(1).$$

With $v = \frac{1}{4}$ these are $M = \tilde{y}^2 + \frac{1}{4} + o(1)$ and $U/M = \tilde{y}^2/6 - \frac{7}{24} + o(1)$, which is the Lemma. Note that the value-gap constant of Proposition 2 is $\frac{1}{2} + \big(\frac{5v}{6} - \frac{1}{2}\big) - \frac{v}{6} = \frac{2v}{3}$, i.e. it equals $\frac{2}{3}\mathrm{Var}(\mathcal{O}_\infty)$ for *any* step law in this class; the value $\lambda\sigma^2\Delta/6$ is the Gaussian case $v = 1/4$. □

A.4 Proof of Proposition 2

Insert Lemma 1 into (3):

$$\frac{J_\Delta}{\lambda\sigma^2\Delta} = \frac{\theta}{\tilde{y}^2 + \frac{1}{4}} + \frac{\tilde{y}^2}{6} - \frac{7}{24} + \frac{1}{2} + o(1) = \frac{\theta}{\tilde{y}^2 + \frac{1}{4}} + \frac{\tilde{y}^2}{6} + \frac{5}{24} + o(1).$$

Differentiating in $\tilde{y}$ and setting to zero gives $(\tilde{y}^2 + \frac{1}{4})^2 = 6\theta$, i.e. $\tilde{y}^2 = \sqrt{6\theta} - \frac{1}{4}$ and

$$\tilde{y} = (6\theta)^{1/4}\Big(1 - \tfrac{1}{4\sqrt{6\theta}}\Big)^{1/2} = (6\theta)^{1/4} - \tfrac{1}{8}(6\theta)^{-1/4} + o(\theta^{-1/4}).$$

Since $\beta = \sigma\sqrt{\Delta}\, y = \sigma\sqrt{\Delta}(\tilde{y} - \kappa)$ and $(6\theta)^{1/4} = \beta^*/(\sigma\sqrt{\Delta})$,

$$\beta_\Delta^* = \beta^* - \kappa\sigma\sqrt{\Delta} - \frac{\sigma\sqrt{\Delta}}{8} \cdot \frac{\sigma\sqrt{\Delta}}{\beta^*} + o(\Delta) = \beta^* - \kappa\sigma\sqrt{\Delta} - \frac{\sigma^2\Delta}{8\beta^*} + o(\Delta).$$

Evaluating the objective at the optimum, $J_\Delta^*/(\lambda\sigma^2\Delta) = \sqrt{2\theta/3} - \frac{1}{24} + \frac{5}{24} + o(1)$, whereas the continuous-time optimum satisfies $J^*/(\lambda\sigma^2\Delta) = \sqrt{2\theta/3}$. Therefore $J_\Delta^* - J^* = \lambda\sigma^2\Delta(-\frac{1}{24} + \frac{5}{24}) + o(\Delta) = \lambda\sigma^2\Delta/6 + o(\Delta)$. □

A.5 Proof of Theorem 5

(ii) Inexact map. Let $\bar{a} = \widehat{a}(\vartheta)$ with $r = \bar{a} - a^* \neq 0$. By consistency of $\hat{\vartheta}_t$ and continuity of $\widehat{a}$, $a_t \to \bar{a}$ a.s. By Assumption 2, the long-run cost rate of the limiting policy exceeds J^* by $\frac{1}{2}J''(a^*)r^2 + o(r^2)$, and by renewal-reward and dominated convergence (thresholds lie in a compact, so cycle lengths have uniformly bounded second moments) the accumulated regret satisfies $\mathrm{Regret}(\tau) = \frac{1}{2}J''(a^*)r^2\tau(1 + o(1))$.

(i) Exact map. Write $\delta_t = \hat{\vartheta}_t - \vartheta$. Three ingredients: *(a) Estimation.* With the floor of Definition 4, the projected efficient estimator satisfies $\mathbb{E}[\delta_t\delta_t^\top] \preceq (c\,t\,I_0)^{-1}(1 + o(1))$ and, by Assumption 1 and sub-exponential concentration of $\hat{\sigma}^2$ (a χ^2 average) and of $\hat{q}$ (a ratio of a geometric sum), the events $\{|\delta_t| > \epsilon\}$ contribute $O(1)$ to the total regret. *(b) Local geometry.* By the delta method and Assumption 2, $J(a_t) - J^* = \frac{1}{2}J''(a^*)\big(\nabla a^{*\top}\delta_t\big)^2 + o(|\delta_t|^2)$. *(c) Accumulation.* Regret accrues per cycle; cycle i starts at a time t_i with $t_i \asymp i\,\mathbb{E}[T]$, so $\mathrm{Regret}(\tau) \leq \sum_{i\leq n}\mathbb{E}[T]\cdot\frac{1}{2}J''\,\nabla a^{*\top}\mathbb{E}[\delta_{t_i}\delta_{t_i}^\top]\nabla a^* \leq \frac{1}{2}J''\nabla a^{*\top}(cI_0)^{-1}\nabla a^*\sum_{i\leq n}1/i + O(1)$, and $\sum_{i\leq n}1/i = \log n + O(1) = \log\tau + O(1)$. □

A.6 Proof of Theorem 7

Apply Theorem 5(i) with $\vartheta = (\log\sigma^2, \log q)$ and $a^*(\vartheta) = \beta^*_\Delta$, which CE-Exact evaluates exactly by construction. The two components are independent: $\hat{\sigma}^2$ is a function of the increments and $\hat{q}$ of the attempt counts. Their relative variances are

$$\mathrm{Var}\big(\log\hat{\sigma}^2\big) = \frac{2}{N} + o(N^{-1}) = \frac{2\Delta}{t} + o(t^{-1}), \qquad \mathrm{Var}\big(\log\hat{q}\big) = \frac{1-q}{i} + o(i^{-1}),$$

the latter because the geometric MLE has Fisher information $1/(q^2(1-q))$ per report (see Remark 8); with $i = t/\mathbb{E}[T] + o(t)$ this is $(1-q)\mathbb{E}[T]/t$. Hence $\mathbb{E}[(\hat{\beta} - \beta^*_\Delta)^2] = \big[2\Delta s_\sigma^2 + (1-q)\mathbb{E}[T]s_q^2\big]/t + o(1/t)$ and the per-unit-time excess cost is $\frac{1}{2}J''_\Delta(\beta^*_\Delta)$ times this, giving (7) after integrating dt/t.

For the continuum limit, $\beta^* \propto (\sigma^2/q)^{1/4}$ gives $s_\sigma \to \beta^*/4$ and $s_q \to -\beta^*/4$; $J(\beta) = \rho c_0\sigma^2/(q\beta^2) + \lambda\beta^2/6$ gives $J''(\beta^*) = 4\bar{J}^*/x^*$; and $\mathbb{E}[T] \to x^*/\sigma^2$. Substituting yields (8). The exact sensitivities differ from their limits at order $\sqrt{\Delta}$: at the baseline parameters $s_\sigma = 0.504$ and $s_q = -0.601$ against ± 0.600, which is why $C_\Delta = 0.638$ while $C = 0.653$. □

A.7 A localisation lemma

The lower bound needs the curvature *at* β^*_Δ, whereas the global bound available on the compact is the smaller constant $m = \inf J''_\Delta > 0$. The following lemma, together with a clipping argument, closes that gap.

Lemma 10 (localisation)**.** *Let $J_\Delta(\cdot;\vartheta)$ be twice continuously differentiable on $[\underline{\beta}, \overline{\beta}]$ with a unique interior minimiser $\beta^*_\Delta(\vartheta)$ and $J''_\Delta(\beta^*_\Delta;\vartheta) > 0$, both continuous in ϑ. Then*

- *(a) J_Δ is strictly convex on the compact, with $m := \inf J''_\Delta > 0$, so $\beta \mapsto J_\Delta(\beta)$ is non-decreasing in $|\beta - \beta^*_\Delta|$ and $J_\Delta(\beta) - J^*_\Delta \geq \frac{m}{2}(\beta - \beta^*_\Delta)^2$;*
- *(b) for every $\delta > 0$ there is $\epsilon > 0$, uniform over ϑ in a neighbourhood of ϑ_0, such that $|\beta - \beta^*_\Delta| \leq \epsilon$ implies $J_\Delta(\beta) - J^*_\Delta \geq (1-\delta)\,\frac{1}{2}J''_\Delta(\beta^*_\Delta)\,(\beta - \beta^*_\Delta)^2$.*

Proof. (a) is Taylor with the global infimum, plus unimodality of a strictly convex function. For (b), $J_\Delta(\beta) - J^*_\Delta = \frac{1}{2}J''_\Delta(\xi)(\beta - \beta^*_\Delta)^2$ for some ξ between β and β^*_Δ by Taylor with Lagrange remainder ($J'_\Delta(\beta^*_\Delta) = 0$); continuity of J''_Δ in (β, ϑ) and compactness give an ϵ with $J''_\Delta(\xi) \geq (1-\delta)J''_\Delta(\beta^*_\Delta)$ for all $|\beta - \beta^*_\Delta| \leq \epsilon$ and all ϑ in a neighbourhood. At the baseline parameters $m = 0.345$ and $J''_\Delta(\beta^*_\Delta) = 1.327$, so the distinction is not cosmetic. □

A.8 Proof of Theorem 9

Throughout, $\vartheta = (\sigma^2, q)$, $\psi(\vartheta) = \beta^*_\Delta(\vartheta)$ is the sampled optimum, and $\mathcal{F}_t$ is the observation filtration generated by the increments $\{\Delta W_j\}_{j\leq t/\Delta}$ and the attempt counts of the reports completed by time t.

A policy is a predictable $[\underline{\beta}, \overline{\beta}]$-valued process (β_t); write $N_r(t)$ for the number of reports completed by t. Call a policy *locally consistent* if $\mathrm{Regret}_\vartheta(\tau) = o(\tau)$ uniformly for ϑ in a neighbourhood of ϑ_0; this is the analogue of the uniformly-good restriction of Lai & Robbins (1985), and it excludes exactly the linearly-regretful policies of Theorem 5(ii).

Step 1: regret dominates a weighted estimation risk. By renewal-reward, the regret accumulated over the cycles completed in $[0,\tau]$ equals $\mathbb{E}\big[\sum_i T_i\,(J_\Delta(\beta_i) - J_\Delta^*)\big]$, which, writing β_t for the piecewise-constant threshold in force at time t, equals $\mathbb{E}\int_0^\tau (J_\Delta(\beta_t) - J_\Delta^*)\,dt + O(1)$, the $O(1)$ being the incomplete terminal cycle (uniformly bounded because $\mathbb{E}[T^2] < \infty$ on the compact). Lemma 10(a) gives

$$\mathrm{Regret}_\vartheta(\tau) \;\geq\; \frac{m}{2}\int_0^\tau \mathbb{E}_\vartheta\big[(\beta_t - \psi(\vartheta))^2\big]dt - O(1), \tag{11}$$

so local consistency forces $\int_0^\tau \mathbb{E}[(\beta_t - \psi)^2]dt = o(\tau)$. To obtain the *local* curvature $J_\Delta''(\beta_\Delta^*)$ rather than m, fix $\delta > 0$, take ϵ from Lemma 10(b), and define the clipped action $\tilde\beta_t = \psi(\vartheta_0) + \mathrm{clip}\big(\beta_t - \psi(\vartheta_0), \pm\epsilon\big)$, which is $\mathcal{F}_t$-measurable. Clipping moves the action towards the minimiser, and J_Δ is non-decreasing in $|\beta - \beta_\Delta^*|$ by Lemma 10(a), so $J_\Delta(\beta_t) - J_\Delta^* \geq J_\Delta(\tilde\beta_t) - J_\Delta^* \geq (1-\delta)\frac{1}{2}J_\Delta''(\beta_\Delta^*)(\tilde\beta_t - \psi)^2$ for ϑ in the shrinking ball (where $|\psi(\vartheta) - \psi(\vartheta_0)| = O(\varepsilon_\tau)$). Hence

$$\mathrm{Regret}_\vartheta(\tau) \;\geq\; (1-\delta)\,\frac{J_\Delta''(\beta_\Delta^*)}{2}\int_0^\tau \mathbb{E}_\vartheta\big[(\tilde\beta_t - \psi(\vartheta))^2\big]dt - O(1), \tag{12}$$

and it suffices to lower-bound the risk of the clipped estimator $\tilde\beta_t$, which is what van Trees does in Step 3. Letting $\delta \downarrow 0$ after $\tau \to \infty$ gives the constant $J_\Delta''(\beta_\Delta^*)$. This is the step that would otherwise silently substitute m for $J_\Delta''(\beta_\Delta^*)$ and understate the bound by a factor of about four at our parameters.

Step 2: Fisher information accrued by an arbitrary policy. The two observation channels are conditionally independent given the policy, so $\mathcal{I}_t(\vartheta)$ is diagonal. *(i) Source.* The increments are i.i.d. $\mathcal{N}(0, \sigma^2\Delta)$ and arrive at rate $1/\Delta$ *whatever the policy does*; hence, exactly and for every policy, $[\mathcal{I}_t]_{\sigma^2\sigma^2} = t/(2\sigma^4\Delta)$. *(ii) Channel.* Each completed report contributes an independent $\mathrm{Geom}(q)$ observation with information $i_q = 1/(q^2(1-q))$, and the decision to report is $\mathcal{F}$-predictable. The log-likelihood ratio process is therefore a sum over reports of i.i.d. score terms evaluated at predictable times, so by the Wald identity for Fisher information under predictable sampling, $[\mathcal{I}_t]_{qq} = \mathbb{E}[N_r(t)]\,i_q$. *(iii) Report rate under consistency.* $\mathbb{E}[T(\beta)] = \Delta M(\beta/(\sigma\sqrt{\Delta}))$ is Lipschitz and bounded away from zero on the compact, so $\big|\mathbb{E}[T(\beta_t)]^{-1} - \mathbb{E}[T(\psi)]^{-1}\big| \leq L|\beta_t - \psi|$. By Wald's identity and Cauchy–Schwarz,

$$\begin{aligned}\mathbb{E}[N_r(t)] \;&\leq\; \frac{t}{\mathbb{E}[T(\psi)]} + L\int_0^t \mathbb{E}|\beta_s - \psi|\,ds \\ &\leq\; \frac{t}{\mathbb{E}[T(\psi)]} + L\sqrt{t\textstyle\int_0^t \mathbb{E}[(\beta_s - \psi)^2]ds} \;=\; \frac{t}{\mathbb{E}[T(\psi)]} + o(t),\end{aligned}$$

the last step by Step 1. A policy can buy channel information faster only by reporting more often, and reporting more often is exactly what consistency forbids.

Step 3: van Trees on a shrinking ball. Fix $\varepsilon_\tau = (\log\tau)^{-1/2}$ and let Λ_τ have a smooth density supported on $B(\vartheta_0, \varepsilon_\tau)$, vanishing with its derivative on the boundary, with prior Fisher information $\mathcal{I}(\Lambda_\tau) = c_0\varepsilon_\tau^{-2} = c_0\log\tau$. The regularity conditions we use are those of Gill & Levit (1995, §2): the model $\{\mathbb{P}_\vartheta\}$ is dominated and differentiable in quadratic mean in ϑ with finite Fisher information on the ball (both hold here: the increment likelihood is Gaussian and the attempt likelihood geometric, each an exponential family in the interior of Θ_0); the prior density is absolutely continuous, compactly supported in the interior, and vanishes with its derivative on the boundary; and the functional ψ is C^1, which holds by the implicit function theorem applied to $\partial_\beta J_\Delta(\beta_\Delta^*;\vartheta) = 0$ using $J_\Delta''(\beta_\Delta^*) > 0$ and smoothness of J_Δ in (β,ϑ). Under these conditions the multivariate van Trees inequality (Gill & Levit, 1995) gives, for every t and every $\mathcal{F}_t$-measurable estimator – in particular for the clipped $\tilde\beta_t$ of Step 1 –

$$\int \mathbb{E}_\vartheta\big[(\beta_t - \psi(\vartheta))^2\big]d\Lambda_\tau \;\geq\; \frac{\big(\int \nabla\psi^\top c\,d\Lambda_\tau\big)^2}{\int c^\top\mathcal{I}_t(\vartheta)c\,d\Lambda_\tau + \mathcal{I}(\Lambda_\tau)\|c\|^2} \qquad \text{for every } c\in\mathbb{R}^2, \tag{13}$$

and optimising over c (take $c = \mathcal{I}_t^{-1}\nabla\psi$ at ϑ_0) yields, using continuity of $\nabla\psi$ and $\mathcal{I}_t$ on the shrinking ball,

$$\int \mathbb{E}_\vartheta\big[(\beta_t - \psi)^2\big] d\Lambda_\tau \;\geq\; \frac{\nabla\psi^\top \mathcal{I}_t^{-1}\nabla\psi}{1 + \mathcal{I}(\Lambda_\tau)/\|\mathcal{I}_t\|}\,(1 + o(1)) \\ =\; \Big[\frac{2\Delta\, s_\sigma^2 + (1-q)\mathbb{E}[T]\, s_q^2}{t}\Big]\frac{1 + o(1)}{1 + O\big(\frac{\log\tau}{t}\big)},$$

where $s_\sigma = \partial\psi/\partial\log\sigma^2$, $s_q = \partial\psi/\partial\log q$ and we substituted the diagonal information of Step 2.

Step 4: integrate. Combining with (12) — not with the cruder (11), whose global constant m would understate the bound by about a factor of four at our parameters — and integrating over $t \in [\log^2\tau, \tau]$, where the prior term is negligible,

$$\int \mathrm{Regret}_\vartheta(\tau)\, d\Lambda_\tau \;\geq\; (1-\delta)\,\frac{J''_\Delta(\beta^*_\Delta)}{2}\big[2\Delta s_\sigma^2 + (1-q)\mathbb{E}[T] s_q^2\big]\int_{\log^2\tau}^{\tau}\frac{dt}{t}\,(1 + o(1)) \\ =\; (1-\delta)\, C_\Delta\, \log\tau\,(1 + o(1)),$$

because $\int_{\log^2\tau}^{\tau} dt/t = \log\tau - 2\log\log\tau$. The clipping radius ϵ of Lemma 10(b) depends on δ but not on τ, so the display holds for every fixed $\delta > 0$; letting $\delta \downarrow 0$ after $\tau \to \infty$ removes the factor $1 - \delta$. Since the supremum over the ball dominates the average,

$$\liminf_{\tau\to\infty}\ \sup_{|\vartheta - \vartheta_0| \leq \varepsilon_\tau}\ \frac{\mathrm{Regret}_\vartheta(\tau)}{\log\tau} \;\geq\; C_\Delta(\vartheta_0). \qquad \square$$

Consequence: the constant is exact. Theorem 7 gives $\limsup \mathrm{Regret}/\log\tau \leq C_\Delta$ for CE-Exact, which is locally consistent by Theorem 5(i). Hence

$$\lim_{\tau\to\infty}\frac{\mathrm{Regret}^{\text{CE-Exact}}(\tau)}{\log\tau} \;=\; C_\Delta \;=\; \inf_{\text{locally consistent policies}}\ \liminf_{\tau\to\infty}\ \sup_{|\vartheta-\vartheta_0|\leq\varepsilon_\tau}\ \frac{\mathrm{Regret}_\vartheta(\tau)}{\log\tau},$$

so C_Δ is not merely an upper bound attained by a particular algorithm: it is the exact asymptotic price of not knowing (σ^2, q). Numerically the two constructions agree to machine precision at the baseline parameters (both 0.638233), as they must, since both are built from the same Fisher information; the content of the theorem is that no policy can do better.

Where each restriction is used. Compactness (Assumption 1) gives the uniform strong convexity in Step 1 and keeps the report rate bounded in Step 2. Consistency is used only in Step 2(iii); without it a policy may report far more often, buy channel information faster, and still be excluded by Theorem 5(ii) as linearly regretful. The shrinking radius $\varepsilon_\tau = (\log\tau)^{-1/2}$ is the largest that keeps the prior information $O(\log\tau)$ negligible against $\mathcal{I}_t \asymp t$ over the bulk of $[0, \tau]$; any $\varepsilon_\tau \gg (\log\tau)^{-1/2}$ works equally well.

A.9 Predicted versus measured linear rates

Theorem 5(ii) with $J''_\Delta(\beta^*_\Delta) = 1.327$ at the baseline parameters predicts

plug-in map	residual r	predicted $\frac{1}{2}J''r^2$	measured
uncorrected	$\kappa\sigma\sqrt{\Delta} = 0.1842$	2.26×10^{-2}	2.20×10^{-2}
continuity-corrected	$\sigma^2\Delta/(8\beta^*) = 5.2 \times 10^{-3}$	1.44×10^{-5}	1.51×10^{-5}

Measured rates are the coefficient c in the least-squares fit $R(\tau) = a + b\log\tau + c\tau$ over $\tau \in [10^2, 10^7]$, 60 seeds. For CE-Exact the same fit returns $c = 2.4 \times 10^{-8}$, four orders of magnitude smaller and consistent with zero.

B NUMERICAL METHOD AND PARAMETERS

B.1 SOLVING THE FREDHOLM EQUATIONS

M and U of (2) satisfy, for $|x| < y$,

$$M(x) = 1 + \int_{-y}^{y} \varphi(z-x)M(z)\,dz, \qquad U(x) = x^2 + \int_{-y}^{y} \varphi(z-x)U(z)\,dz,$$

with φ the standard normal density and $M = U = 0$ outside $(-y, y)$; the reported quantities are $M(0), U(0)$. We use two independent discretisations.

(a) Nyström with Gauss–Legendre nodes. n nodes on $(-y, y)$, dense solve. Converged to eight digits by $n = 100$ at $y = 4$ (Table 2).

(b) Midpoint rule on a uniform grid. The resulting matrix is symmetric Toeplitz, so the system is solved in $O(n^2)$ by the Levinson algorithm, which is what makes the $y^*(\theta)$ table and the per-β gap function cheap. Every number in the paper uses this solver at its production setting (at least 800 nodes, i.e. 200 pts/unit at $y = 4$), where the residual midpoint bias against (a) is 3×10^{-6} relative — four orders of magnitude below the smallest effect we report. Coarser settings are shown in Table 2 only to exhibit the $O(h^2)$ decay.

scheme	size	$M(4)$	$U(4)$
Nyström	$n = 100$	21.25016424	68.68486316
Nyström	$n = 800$	21.25016424	68.68486316
Toeplitz	30 pts/unit	21.25333195	68.70110328
Toeplitz	120 pts/unit	21.25036219	68.68587793
Toeplitz	200 pts/unit	21.25023550	68.68522848
Monte Carlo	6×10^4 cycles	21.249 ± 0.070	68.83 ± 0.23

Table 2: Convergence and external validation of the solver at $y = 4$.

B.2 VERIFICATION OF THE EXPANSION CONSTANTS

With $\tilde{y} = y + \kappa$, define $a(y) = M(y) - \tilde{y}^2$ and $b(y) = U(y)/M(y) - \tilde{y}^2/6$. Lemma 1 claims $a \to 1/4$ and $b \to -7/24 = -0.29167$.

pts/unit	y	$a(y)$	$b(y)$	$\frac{1}{2} + b - a/6$
60	30	$+0.25581$	-0.28715	0.17022
150	30	$+0.25093$	-0.28791	0.17027
300	30	$+0.25023$	-0.28802	0.17028
300	60	$+0.25046$	-0.28977	0.16849

Table 3: The last column is the predicted value-gap constant, which Proposition 2 identifies as $1/6 = 0.16667$; the residual drift is the $o(1)$ term of Lemma 1. A direct computation of $(J^*_\Delta - J^*)/(\lambda\sigma^2\Delta)$ over six decades of θ gives 0.16884 ($\theta = 10^6$), 0.16786 (10^7), 0.16731 (10^8), 0.16700 (10^9).

B.3 THE THRESHOLD TABLE

$y^*(\theta) = \arg\min_y\{(\theta + U(y))/M(y)\}$ is computed by golden-section search on a grid of 161 values of $\log_{10}\theta \in [-1, 7]$ and interpolated by a cubic spline; the relative error against a fresh solve is below 10^{-8} at intermediate θ ($\theta = 140$: 4.780524 vs 4.780524; $\theta = 2300$: 10.245034 vs 10.245034). Outside the grid we use the asymptote $y^* = (6\theta)^{1/4} - \kappa$. CE-Exact evaluates the table, so its per-report cost is one spline evaluation.

B.4 Experimental protocol

Parameters. Baseline $\sigma^2 = 1$, $\Delta = 0.1$, $\lambda = \rho = 1$, $q = 0.5625$ ($p = 0.25$), $\beta^* = 2.4$ (so $c_0 = q\lambda(\beta^*)^4/(6\rho\sigma^2) = 3.1104$), giving $\theta = 553.0$, $y^* = 6.9919$, $\beta^*_\Delta = 2.2111$, $J^*_\Delta = 1.93812$, $J^* = 1.92$, $x^* = 5.76$, $J''_\Delta(\beta^*_\Delta) = 1.327$, $\mathbb{E}[T] = 5.762$. The sweep varies $q \in \{0.49, 0.5625, 0.7225, 0.9\}$, $\Delta \in \{0.05, 0.1, 0.2\}$, $\beta^* \in \{1.8, 2.4, 3.2\}$.

Regret estimator. What we plot is the *conditional expected regret contribution*: each cycle contributes $\mathbb{E}[T(\hat\beta_i)]\,(J_\Delta(\hat\beta_i) - J^*_\Delta)$, both factors evaluated exactly from the Fredholm solution at the threshold the learner actually chose. It is not a variance-reduced estimate of a noisy quantity but the conditional mean of (4) given the learner's actions, so it equals (4) in expectation while removing the cost noise that is independent of the decision rule: at $\tau = 2 \times 10^5$ with 40 seeds, CE-Exact gives 7.415 accrued versus 7.26 ± 198.09 realised. Sufficient statistics evolve exactly: $\hat\sigma^2$ from an accumulated χ^2 over all increments, $\hat q$ from accumulated geometric attempt counts.

Path-level validation. A full path simulator (increment-by-increment, realised costs, exact Brownian-bridge integrals) confirms (i) the oracle's realised cost rate equals J^*_Δ (1.93888 versus 1.93812, $t = 0.63$, 200 seeds, $\tau = 2.9 \times 10^4$) and (ii) realised and accrued regret agree for CE-Exact.

Slopes and fitting windows. Local slopes are least-squares fits of R against $\log\tau$ over one-decade windows; global fits are $R = a + b\log\tau$ and $R = a + b\log\tau + c\tau$ over $\tau \geq 10^2$. Windows are chosen with respect to the regimes the theory predicts, not to the data: the lower end excludes the burn-in (20 cycles, i.e. $\tau \lesssim 10^2$), and where a map has a nonzero residual the relevant comparison is made below its predicted crossover τ^* for the logarithmic constant and above it for the linear rate.

B.5 Sweep results

q	Δ	β^*	C_Δ	C (cont.)	C with $1/q$	slope	slope/C_Δ
0.9000	0.10	2.4	0.1719	0.1862	0.2016	0.1645	0.957
0.7225	0.10	2.4	0.4172	0.4316	0.5790	0.4081	0.978
0.5625	0.10	2.4	0.6382	0.6528	1.1232	0.5806	0.910
0.4900	0.10	2.4	0.7384	0.7530	1.4868	0.6941	0.940
0.5625	0.05	2.4	0.6236	0.6288	1.0992	0.6467	1.037
0.5625	0.20	2.4	0.6605	0.7008	1.1712	0.6189	0.937
0.5625	0.10	1.8	0.2076	0.2184	0.3672	0.2657	1.280
0.5625	0.10	3.2	1.9772	1.9968	3.4835	1.9210	0.972
mean $\pm$ sd of slope/C_Δ						1.001 ± 0.111	
mean $\pm$ sd of slope/$C_{\text{with } 1/q}$						0.612 ± 0.114	

Table 4: Measured regret slope over $\tau \in [3 \times 10^5, 3 \times 10^6]$, 60 seeds, against the exact constant (7), its continuum limit (8), and the variant carrying a spurious $1/q$ (Remark 8).

B.6 Baselines

Two-timescale stochastic approximation. $\hat J \leftarrow \hat J + a_k(C_k - \hat J\,T_k)$ with $a_k = a_0/k$, and $\beta_k = \sqrt{\hat\sigma^2\Delta}\;G^{-1}\big(\hat J/(\lambda\hat\sigma^2\Delta)\big)$ where $G(y) = (U'(y) + M'(y)/2)/M'(y)$ inverts the sampled stationarity condition (the fixed charge cancels, so this baseline needs λ and $\hat\sigma^2$ only). Accrued regret at $\tau = 3 \times 10^4$: 19.1 ($a_0 = 0.2$), 58.2 ($a_0 = 0.6$), 182.1 ($a_0 = 0.05$), versus 4.56 for CE-Exact; at $\tau = 10^6$ the tuned variant reaches 37.1 versus ≈ 5.5.

Kiefer–Wolfowitz. Two-sided finite differences on the average cost with $c_k = c_0 k^{-1/4}$, $a_k = a_0/k$. Regret is two to four orders of magnitude worse for every (a_0, c_0) we tried, because each probe pays the quadratic cost of deviating from the optimum and the gradient estimate is dominated by cycle-cost noise.

B.7 Reproduction

`core.py` (Nyström solver, Monte-Carlo reference), `fast.py` (Toeplitz solver), `table.py` (threshold table and J_Δ), `cexact.py` (exact constants), `sim.py` (estimator-path simulation),

`path.py` (full path simulation and baselines), `sa.py` (stochastic-approximation and Kiefer–Wolfowitz baselines), `plots.py` (all figures), `tables.py` (Tables 2, 3 and the overshoot moments, including the independent moment solver). Instance 2 is `sprt.py`, `build_sprt.py`, `sim_sprt.py`, `run_sprt.py`, `plot_sprt.py`. Every run is seeded; `bash run_all.sh` regenerates every figure and every table in the paper from scratch in about forty minutes single-threaded, and `python3 plots.py` alone regenerates Figures 1–4 from the stored run files.

B.8 Instance 2: repeated sequential testing

For steps $\mathcal{N}(\mu, \sigma^2)$ and barriers $\pm B$, the expected sample size and the error probability solve, for $|x| < B$,

$$m(x) = 1 + \int_{-B}^{B} \phi_\mu(z - x) m(z) dz, \qquad p(x) = \Phi\Big(\tfrac{-B-x-\mu}{\sigma}\Big) + \int_{-B}^{B} \phi_\mu(z - x) p(z) dz,$$

with ϕ_μ the $\mathcal{N}(\mu, \sigma^2)$ density; $\mathbb{E}[N] = m(0)$ and $\mathbb{P}(\text{err}) = \Phi((-B-\mu)/\sigma) + \int \phi_\mu(z-\mu)p(z)dz$. Nyström with 320 Gauss–Legendre nodes suffices (values identical to 640 nodes at five decimals). Validation at $\mu = 1, \sigma = 3$: $B = 6$ gives $\mathbb{E}[N] = 5.6496$ and $\mathbb{P} = 0.15191$ against Monte Carlo 5.6596 ± 0.0173 and 0.15367 ± 0.00147 (6×10^4 paths); $B = 12$ gives 12.7718 and 0.04504 against 12.7907 ± 0.0366 and 0.04578 ± 0.00085.

Residual of the Siegmund-corrected map, in units of σ, at four parameter settings (μ, σ, λ) with $c = 1$: $(1, 3, 400) \to -0.00138$, $(1, 3, 100) \to -0.00088$, $(1, 2, 400) \to -0.00330$, $(0.5, 3, 400) \to -0.00022$; identical at 320 and 640 nodes, so these are properties of the problem and not of the discretisation. The corresponding Wald residuals are $+0.5812$, $+0.5817$, $+0.5793$, $+0.5824$, i.e. the overshoot constant $\kappa = 0.58260$ to three digits.

The learner estimates σ^2 by $\widehat{\sigma^2} = n^{-1} \sum_{i \le n} (X_i^2 - \mu^2)$, which is unbiased whichever hypothesis holds and has $n \operatorname{Var}(\log \widehat{\sigma^2}) = 2 + 4\mu^2/\sigma^2$; sufficient statistics are simulated exactly by noncentral χ^2 draws. Baseline $\mu = 1$, $\sigma = 3$, $c = 1$, $\lambda = 400$: $B^* = 17.9830$, $\mathbb{E}[N] = 19.508$, $\mathbb{P}(\text{err}) = 0.01233$, $J^* = 24.4384$, $J''(B^*) = 0.22221$, $\partial B^* / \partial \log \sigma^2 = 13.981$, hence $C = 2.7212$.

B.9 Overshoot moments (input to Lemma 1)

$\mathbb{E}[|X_N|^k]$ solves $g(x) = \int_{|z| \ge y} |z|^k \varphi(z - x) dz + \int_{|z| < y} \varphi(z - x) g(z) dz$, an equation independent of those used elsewhere in the paper. With 1200 Gauss–Legendre nodes:

y	$\mathbb{E}\lvert X_N\rvert$	$\mathbb{E}[X_N^2]$	$\mathbb{E}[\mathcal{O}]$	$\operatorname{Var}(\mathcal{O})$
6	6.582597	43.580586	0.582597	0.250000
8	8.582597	73.910974	0.582597	0.250000
10	10.582597	112.241363	0.582597	0.250000
14	14.582597	212.902140	0.582597	0.250000
16	16.582597	275.232529	0.582597	0.250000

$\mathbb{E}[\mathcal{O}] = 0.582597$ reproduces $-\zeta(1/2)/\sqrt{2\pi} = 0.5825972$ to six digits, a check on the method; $\operatorname{Var}(\mathcal{O}) = 0.250000$ is the constant v of Lemma 1. We use $v = 1/4$ on the strength of this computation: the value is stable to six digits over $y \in \{6, 8, 10, 14, 16\}$ and under refinement of the quadrature, but we have not located a closed-form derivation of it in the literature and do not claim one. Every statement that depends on v is therefore exact in v and numerical only in its value: Lemma 1 gives $M = \tilde{y}^2 + v + o(1)$ and $U/M = \tilde{y}^2/6 + 5v/6 - 1/2 + o(1)$, and the value gap of Proposition 2 is $\frac{2}{3} v \lambda \sigma^2 \Delta + o(\Delta)$. The exact identity (10) was also checked directly: at $y = 4$, $\frac{1}{6}\mathbb{E}[X_N^4] - \frac{1}{2}\mathbb{E}[X_N^2] = 68.66$ by Monte Carlo (4×10^5 paths) against $U(4) = 68.685$ from the solver. Each row also satisfies $\mathbb{E}[X_N^2] = (y + \kappa)^2 + \frac{1}{4}$ to six digits, which is Lemma 1 with $v = 1/4$.