CONVERGENCE OF FEW-STEP IMPORTANCE WEIGHTS: A CHEAP, REFERENCE-FREE CERTIFICATE

Anonymous authors
Paper under double-blind review

ABSTRACT

Diffusion and flow models increasingly use few solver steps as proposals for importance sampling, where the model likelihood enters the final estimator and numerical convergence becomes part of the sampling problem. Few-step likelihoods are only approximate, and confirming their accuracy would require exactly the expensive many-step reference that few-step sampling was meant to avoid. The standard cheap diagnostic, effective sample size, cannot fill this gap: it cancels a constant likelihood shift exactly, and its first-order response to a fluctuating error is governed by a squared-weight alignment that differs from the observable covariance governing estimation bias, so an observable can be biased while ESS is unchanged. We introduce a cheap, reference-free, a-posteriori certificate for the coupled log-weight discretization error. For a fixed source draw the log weight admits a structured expansion in the step size, so extrapolation from a few additional short solves reconstructs the error without invoking the reference; the certificate is provably sound once the step size falls below an explicit threshold, tightens as the budget grows, and uses no hand-tuned constants. The same expansion yields a same-sample truncation diagnostic, and an elementary bridge lemma converts per-sample error bars into a bound on the self-normalized estimate. Across controlled synthetic flows, one-step average-velocity maps, and energy-based molecular samplers, the certificate attains high empirical coverage, substantially reduces importance-sampling error on a real average-velocity map, and reaches near-reference-quality estimates at a fraction of the reference cost on a molecular system, whereas ESS-based budget rules accept severely biased budgets and cannot rank candidates at the degeneracy floor.

1 INTRODUCTION

Diffusion- and flow-based models dominate generative modeling (Lipman et al., 2023; Ho et al., 2020; Song et al., 2021b), and much recent work makes them fast: generating in a few solver steps, or one, through distillation, consistency training, and average-velocity maps (Song et al., 2023; 2021a; Frans et al., 2025; Lu & Song, 2025; Geng et al., 2025; Ding et al., 2024; Karras et al., 2022; Rehman et al., 2026). Beyond producing samples, such models are increasingly used as samplers for unbiased estimation: paired with importance sampling (IS), the model proposes points that are reweighted by the ratio of the target density to the model's own likelihood, so target expectations stay consistent despite an imperfect proposal. The prominent instance is sampling unnormalized (Boltzmann) targets (Noé et al., 2019; Midgley et al., 2023; Akhound-Sadegh et al., 2024), but the same reweighting underlies any use of a generative model for unbiased expectation estimation. Both its samples and its likelihood therefore lie on the critical path.

The recipe hinges on a trustworthy and numerically converged model likelihood, and this is precisely where the field's recent direction creates a problem. Exact likelihoods from continuous normalizing flows require integrating an ODE with many network evaluations, so a wave of recent work instead computes the likelihood in only a few solver steps to make it affordable (Rehman et al., 2026; Geng et al., 2025). The resulting few-step likelihood is only approximate, and the error it carries is left uncharacterized: practitioners have no efficient way to decide whether a chosen step budget is large enough, and confirming correctness would require exactly the expensive many-step reference that few-step sampling was meant to avoid. Worse, the standard inexpensive diagnostic for IS reliability sees only the weight distribution. A constant likelihood error cancels in self-normalization,

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

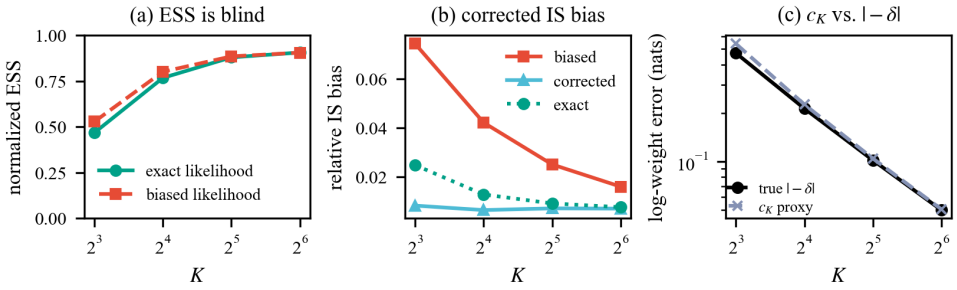


Figure 1: The same-sample diagnostic exposes a failure mode that the standard diagnostic misses. Here c_K is the same-sample diagnostic of Sec. 4.5. Under a low-variance, observable-correlated likelihood error (a truncated log-determinant), (a) the measured effective sample size is essentially unchanged from the exact-likelihood value because this error is nearly orthogonal to the squared-weight direction governing ESS; (b) yet the importance-sampling estimate is biased at first order through the error's covariance with the observable (Prop. 3); (c) c_K follows the magnitude of the same-sample log-weight error in this controlled setting, and subtracting its signed counterpart from the deployed log weight removes the estimation error.

and the first-order ESS response to a fluctuating error is controlled by its alignment with squared normalized weights; the first-order shift of an observable is instead controlled by its covariance with that observable. The latter can be nonzero while the former is negligible, producing a silent failure.

We address the first gap with a cheap, reference-free certificate for coupled log-weight convergence: whether the full sample-and-weight pipeline has converged to its continuous-flow limit on the same source draw. This is distinct from the same-sample deployed-likelihood error caused by a truncated log-determinant or an approximately invertible map, which we analyse separately (Sec. 4). For a fixed source draw, the coupled log-weight gap is a smooth function of solver step size and admits a structured expansion, so it can be reconstructed by extrapolation from a few additional short solver runs at finer budgets $(2K, 4K)$ without invoking the expensive reference. The result is an a-posteriori convergence error bar that is provably valid once the step budget is large enough, becomes tight as the budget grows, and uses no hand-tuned constants. Applying the expansion one level deeper to a same-sample likelihood approximation yields a separate, empirically validated truncation diagnostic; when any method supplies valid per-sample bounds on that error, an elementary bridge lemma converts them into a computable bound on the final self-normalized estimate.

Our contributions are threefold:

- **A reference-free log-weight convergence certificate, with an explicit guarantee and an explicit boundary.** We separate the three error objects a few-step pipeline conflates (discretization drift, deployed-likelihood error, downstream estimation bias), introduce an a-posteriori certificate for the coupled K-step-to-continuous log-weight gap, and prove it soundly upper-bounds that gap above an explicit step threshold $h^\star$, tightening to the exact error as the budget grows. A fourth solve supplies an empirical gate for the threshold, and a free residual screens inconsistent scaling (Remark 1).
- **Exposing and correcting an ESS-invisible failure.** Self-normalized IS cancels the constant part of a likelihood error exactly. For the remaining fluctuation, the first-order change in an observable depends on $\mathrm{Cov}_w(\phi, \delta)$, whereas the first-order ESS change depends on a different, squared-weight alignment. Either can vanish without the other. Our same-sample truncation diagnostic detects the constructed regime where ESS is unchanged but the target observable is biased, and its correction removes the error where the log-determinant expansion converges.
- **Broad validation across few-step generative samplers.** We validate the approach on controlled synthetic flows of growing dimension, under-trained models whose likelihoods are far from ideal, higher-order solvers, a faithful from-scratch reimplementation of FALCON's average-velocity model class, and energy-based molecular samplers, establishing that the certificate transfers from idealized settings to the few-step generative samplers used in practice.

2 Related Work

Few-step samplers and Boltzmann generators. Boltzmann generators pair a flow with IS to obtain consistent equilibrium samples (Noé et al., 2019; Midgley et al., 2023; Köhler et al., 2020; Wu et al., 2020); energy-only diffusion samplers such as iDEM (Akhound-Sadegh et al., 2024) train without data samples, and reducing the sampling budget is an active front (Geng et al., 2025; Ding et al., 2024; Yuan et al., 2026; Schioppa & Salimans, 2026; Nguyen-Cong et al., 2026). Closest to us, FALCON (Rehman et al., 2026) makes the likelihood cheap for IS but explicitly leaves its correctness unguaranteed, and Jutras-Dubé et al. (2026) build a better few-step weight by enforcing volume consistency across step resolutions. This is a constructive remedy, whereas we certify, a-posteriori and reference-free, convergence of the coupled sample-and-log-weight pipeline. We do not build a faster sampler; we diagnose when its few-step log weight has reached the continuous-flow reference.

Reliability diagnostics, and the error they cannot see. ESS is the standard cheap proxy for IS reliability (Owen, 2013; Martino et al., 2017), and Pareto-smoothed IS (Vehtari et al., 2024) adds a tail-index ($\hat{k}$) statistic; both are functions of the weight distribution alone and therefore share its blindness to a near-constant likelihood shift, which we sharpen into an order-separation result (Sec. 4.5). Convergence theory for samplers bounds the sampling-distribution error in KL or Wasserstein (Richter & Berner, 2024; Chen et al., 2023); our complementary object is the accuracy of the computed likelihood q_K that enters the weight. Oracle-based diagnostics for few-step correctness exist for discrete diffusion (Tang et al., 2026), and IS sample efficiency is itself read as the reliability diagnostic in gravitational-wave inference (Dax et al., 2023). The Richardson extrapolation machinery we use is classical for ODE solvers (Hairer et al., 1993); our contribution is to apply it to the few-step likelihood functional, tie it to IS validity through an explicit bias bound (Lemma 4), make it reference-free, and validate it on real samplers. App. L extends this discussion, including the $\widetilde{\Omega}(\sqrt{d})$ score-query lower bound of Xun & Price (2026), which our certificate renders actionable a-posteriori.

3 Preliminaries

Flow likelihood and the discrete map. A flow defines a velocity field $v(x, \tau)$, $\tau \in [0, 1]$. A K-step solver maps a prior draw $z \sim \mathcal{N}(0, I)$ (at τ=1) to a sample $x_K = T_K(z)$ (at τ=0). When each step is a diffeomorphism, T_K is an exact normalizing flow with tractable log-density. As the discrete analogue of the continuous-normalizing-flow likelihood (Chen et al., 2018; Grathwohl et al., 2019), it uses the discrete change of variables

$$\log q_K(x_K) = \log p_{\text{base}}(z) - \textstyle\sum_{k=1}^{K} \log\big|\det\, J_k\big|, \tag{1}$$

where J_k is the Jacobian of the k-th step map. For an Euler step $x \mapsto x - h\, v(x, \tau)$ with $h = 1/K$, $J_k = I - h\, \nabla_x v$.

Self-normalized IS and ESS. For a target p known up to a constant, SNIS uses weights $w_i = p(x_i)/q_K(x_i)$; with $g_K(z) = \log p(x_K) - \log q_K(x_K)$ the log-weight, the normalized ESS is $\text{ESS} = 1/(N \sum_i \tilde{w}_i^2)$, which equals $1/(1 + \chi^2(p \,\|\, q_K))$ in the large-sample limit. ESS depends only on the spread of the weights, the fact we sharpen in Sec. 4.

4 A Reference-Free Log-Weight Convergence Certificate

4.1 Problem: three error objects, and which one corrupts importance sampling

A few-step pipeline involves three distinct error objects that are easily conflated, so we begin by separating them. Fix a source draw z and write the log-weight at budget K as

$$g_K(z) \;=\; \log p(x_K) - \log q_K(x_K), \qquad E_K(z) \;:=\; |g_K(z) - g_\infty(z)|, \tag{2}$$

with g_∞ the $K \to \infty$ value. The three objects are:

(O1) Coupled log-weight discretization drift E_K: the gap between the K-step pipeline's exact log weight and its continuous-flow limit under the same source draw. With the exact change of variables equation 1, q_K is a valid density and SNIS with weights p/q_K is consistent at every K; E_K is therefore a convergence diagnostic for the coupled pipeline, not a same-sample likelihood error or an SNIS-bias bound.

(O2) Deployed-likelihood error $\delta(x) = \log q^{\mathrm{dep}}(x) - \log q_K(x)$: the gap, at the sample the pipeline actually produced, between the number the likelihood code returns and the exact density of the sampler that was deployed. This is the object that corrupts SNIS, and it is nonzero in exactly the situations practitioners face: (a) the log-determinant is computed by a cheap surrogate (a truncated series, a stochastic trace, an approximately-invertible learned map), so q^{dep} is not q_K. This error must be evaluated at the deployed sample x_K; in general it is not $g_K(z) - g_\infty(z)$, whose two terms are evaluated at different coupled endpoints.

(O3) Downstream estimation bias: the shift of the final self-normalized estimate caused by δ. Section 4.6 bounds it computably in terms of per-sample bounds on δ.

Theorem 2 certifies (O1) by extrapolation across budgets without computing g_∞. For the same-sample surrogate error (O2), the truncated log-determinant gives a controlled empirical diagnostic c_K (Sec. 4.5); we do not claim that Theorem 2 proves c_K sound for arbitrary likelihood code. The empirical coverage/tightness studies concern (O1), while the ESS-blindness and correction studies concern the specific truncated-series instance of (O2).

4.2 The log-weight admits an order-p expansion

Assumption 1 (order-p expansion)**.** For $h \in (0, h_0]$ and a one-step solver of (known) order p, $g(h) := g_K(z)$ with $h = 1/K$ satisfies $g(h) = g_0 + a\,h^p + b\,h^{p+1} + r(h)$, with $g_0 = g_\infty$, $b \neq 0$, and $|r(h)| \leq M\,h^{p+2}$.

This is the global-error asymptotic expansion of an order-p one-step method (Hairer et al., 1993) composed with the smooth maps $\log p(\cdot)$ and $\log|\det(\cdot)|$; M collects bounds on the relevant derivatives along the trajectory. The trajectory error and log-determinant quadrature in equation 1 both contribute an expansion in integer powers of h starting at h^p.

4.3 Three short solves recover the leading error

Evaluate g at $h, h/2, h/4$ and form the gaps $\Delta_1 = g(h) - g(h/2)$, $\Delta_2 = g(h/2) - g(h/4)$. Writing $A := a\,h^p$, $B := b\,h^{p+1}$, $s := 2^{-p}$, $t := 2^{-(p+1)} = s/2$, Assumption 1 gives

$$\begin{bmatrix}\Delta_1\\ \Delta_2\end{bmatrix} = \underbrace{\begin{bmatrix}1-s & 1-t\\ s(1-s) & t(1-t)\end{bmatrix}}_{=:\,M_p}\begin{bmatrix}A\\ B\end{bmatrix} + \begin{bmatrix}\rho_1\\ \rho_2\end{bmatrix}, \tag{3}$$

with $\rho_1 = r(h) - r(h/2)$, $\rho_2 = r(h/2) - r(h/4)$, both $O(h^{p+2})$.

Lemma 1 (exact solve)**.** $\det M_p = (1-s)(1-t)(t-s) \neq 0$, so equation 3 is invertible. Solving without ρ gives $\hat{A}, \hat{B}$ with $\hat{A} - A = (s\rho_1 - 2\rho_2)/(s(s-1))$ and $\hat{B} - B = 4(\rho_2 - s\rho_1)/(s(s-2))$; both are $O(h^{p+2})$. (Verified symbolically; see App. B.)

The two-term estimate $\hat{A} + \hat{B}$ thus equals $g(h) - g_0$ up to $O(h^{p+2})$: one order beyond the leading $O(h^p)$ signal, and reference-free. Figure 2 summarizes the construction: three short solves at budgets $K, 2K, 4K$ are extrapolated to $h{=}0$ to recover the (otherwise expensive) limit g_∞, and the resulting estimate is wrapped in the safety margin $|\hat{B}|$ to form a certificate that provably brackets the true error.

4.4 The certificate and its soundness

Define the constant-free certificate

$$C_{\mathrm{sc}}(z) \;=\; |\hat{A} + \hat{B}| \;+\; |\hat{B}|, \tag{4}$$

i.e. the two-term estimate plus the leading correction as a safety margin. With $u := 2^{-(p+2)} = s/4$, define the explicit constants

$$\Gamma_p = \tfrac{|2-3s|}{(1-s)(2-s)}(1{+}u) + \tfrac{2}{(1-s)(2-s)}u(1{+}u) + 1, \quad \Gamma_p' = \tfrac{4}{2-s}(1{+}u) + \tfrac{4}{s(2-s)}u(1{+}u).$$

Theorem 2 (soundness and tightness)**.** Under Assumption 1, for all $h \in (0, h_0]$: (i) $|(\hat{A} + \hat{B}) - (g(h) - g_0)| \leq \Gamma_p\,M\,h^{p+2}$; (ii) $\big|\,|\hat{B}| - |b|h^{p+1}\,\big| \leq \Gamma_p'\,M\,h^{p+2}$; (iii) (soundness) $C_{\mathrm{sc}} \geq E_K$ for

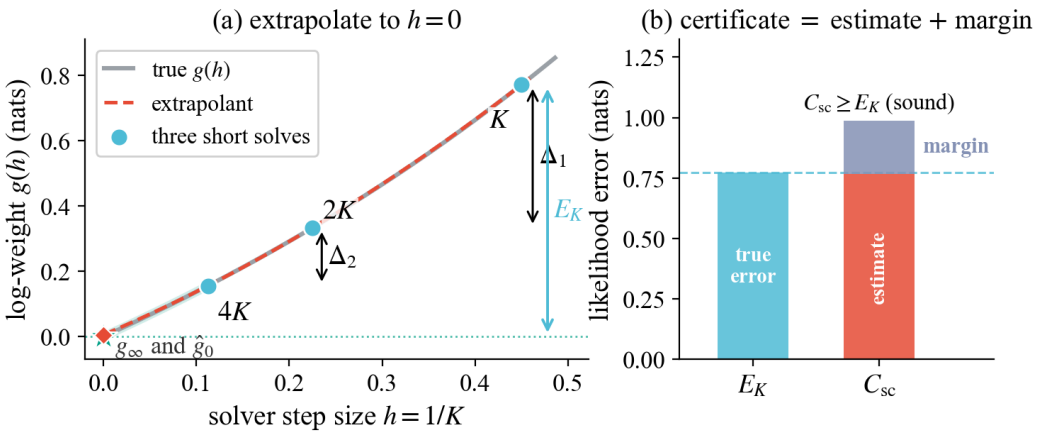


Figure 2: The theorem-backed three-level certificate C_{sc}. *(a)* For a fixed source draw, the log-weight $g(h) = \log p(x) - \log q(x)$ is a smooth function of the step size $h{=}1/K$. Three cheap solves at $K, 2K, 4K$ (blue) determine the leading expansion terms $(\hat{A}, \hat{B})$ via the gaps Δ_1, Δ_2; extrapolating to $h{=}0$ (dashed red) recovers an estimate $\hat{g}_0$ of the expensive many-step value g_∞ (green star) without ever computing it. The true few-step error at budget K is $E_K = |g(h) - g_\infty|$. *(b)* The certificate $C_{sc} = |\hat{A} + \hat{B}| + |\hat{B}|$ stacks the two-term error estimate and the leading-order safety margin; Theorem 2 guarantees it lies above E_K (sound) once $h \leq h^\star$, and that it tightens to E_K as K grows.

every $h \leq h^\star := |b|/((\Gamma_p + \Gamma'_p)M)$; (iv) (tightness) $C_{sc}/E_K \to 1$ as $h \to 0$ (if $a \neq 0$). The constants are exact rationals: $\Gamma_1 = 17/8$, $\Gamma'_1 = 15/4$; $\Gamma_2 = 355/168$, $\Gamma'_2 = 85/28$.

Proof idea. By (i) and the reverse triangle inequality $|\hat{A} + \hat{B}| \geq E_K - \Gamma_p M h^{p+2}$; by (ii) $|\hat{B}| \geq |b|h^{p+1} - \Gamma'_p M h^{p+2}$. Adding, $C_{sc} \geq E_K + |b|h^{p+1} - (\Gamma_p + \Gamma'_p)Mh^{p+2}$, which is $\geq E_K$ exactly when $h \leq h^\star$. App. B gives the full proof and a symbolic and numerical verification. Its adversarial large opposite-sign remainder is sound for all $h \leq h^\star$ and fails just outside. □

Interpretation and guarantee boundary. The safety margin $|\hat{B}|$ is order h^{p+1} while the uncertified remainder is order h^{p+2}. Once the step is fine enough, the certificate is provably sound and becomes tight. The adversarial verification confirms that $h^\star$ is non-vacuous by failing just beyond it. The boundary is explicit: $h^\star$ involves b and M, which are not observable at deployment, so Theorem 2 is conditional and sound only when $h \leq h^\star$. Because neither M nor $h^\star$ is identified by the deployed solves, a fourth-level proxy and the residual ϱ are used only as heuristic gates (Remark 1, Sec. 4.6, App. H.5); gated draws are escalated, not certified.

Remark 1 (estimating the remainder and the threshold). The only non-computable input is M. A fourth solve at $h/8$ lets us solve a 3-term system for $(\hat{A}, \hat{B}, \hat{C})$; since $\hat{C}$ estimates the leading remainder term (not a bound on the full remainder), $\widehat{M} = |\hat{C}|/h^{p+2}$ and the induced $\hat{h}^\star$ are heuristic proxies, not estimates that identify M or certify $h \leq h^\star$. The resulting fourth-level variant $|\hat{A} + \hat{B}| + s_0|\hat{C}|$ uses a computable next term and a data-calibrated factor s_0 as another heuristic gate; its reported 0.05-tolerant empirical coverage is 0.99–1.00 on real flows (Sec. 5).

4.5 From likelihood error to an ESS-invisible bias

A practical few-step likelihood replaces the exact log-determinant in equation 1 by a cheap surrogate, incurring the deployed-likelihood error δ of Sec. 4.1, case (a). A stochastic surrogate (e.g. a Hutchinson trace estimate (Hutchinson, 1989; Grathwohl et al., 2019)) makes δ a mean-zero fluctuation: it inflates weight variance and thus lowers ESS, so ESS already catches it. A systematic surrogate (e.g. a truncated log-determinant, or an approximately-invertible learned map) produces a deterministic $\delta(x)$. Here precision matters, and we state the mechanism exactly: the constant part of δ cancels completely in self-normalization, so it biases nothing and moves nothing; the danger is carried entirely by the fluctuating part, and there the orders separate.

Proposition 3 (different first-order directions for bias and ESS)**.** Let the deployed log-likelihood differ from the exact one by $\delta(x) = \log q_K^{\mathrm{dep}}(x) - \log q_K(x)$, and write $\delta = \bar{\delta} + \tilde{\delta}(x)$ with $\bar{\delta}$ the weight-averaged mean and $\tilde{\delta}$ the fluctuation. Then: (i) the SNIS estimate of every observable and the normalized ESS are both exactly invariant to $\bar{\delta}$; (ii) the SNIS estimate of an observable ϕ shifts by $-\mathrm{Cov}_w(\phi, \tilde{\delta}) + O(\|\tilde{\delta}\|^2)$; (iii), writing the exact normalized weights as $\tilde{w}_i$ and $S = \sum_i \tilde{w}_i^2$, the normalized ESS has relative change

$$\frac{\mathrm{ESS}^{\mathrm{dep}} - \mathrm{ESS}}{\mathrm{ESS}} = 2\,\frac{\sum_i \tilde{w}_i^2 \tilde{\delta}_i}{S} + O(\|\tilde{\delta}\|^2).$$

Thus both quantities may move at first order, but in different directions: an error can be orthogonal to the squared-weight direction while remaining correlated with a target observable. In that regime the observable is biased while ESS is unchanged to first order, as in Fig. 1.

The proof (App. E) is an elementary perturbation of the self-normalized weights. The empirically measured δ of a truncated log-determinant is exactly of this low-variance, observable-correlated kind (Fig. 1). The same series logic that underlies the certificate applies one level up: the leading term is a computable quantity c_K (one extra trace per step). With our convention $\delta = \log q_K^{\mathrm{dep}} - \log q_K$, its signed version estimates the deployed log-weight error $g_K^{\mathrm{dep}} - g_K = -\delta$; it is therefore subtracted from the deployed log weight. The plots use the corresponding nonnegative magnitude proxy. Section 5 validates this relation only where the controlled log-determinant expansion converges.

4.6 From valid same-sample error bars to a bound on estimation bias

Suppose a same-sample likelihood analysis supplies per-sample bounds on the deployed-likelihood error (as the controlled truncation analysis does in our experiments). The practitioner cares about the bias of the final estimate (object (O3) of Sec. 4.1); the following elementary lemma converts any such valid bounds into computable downstream bounds.

Lemma 4 (bridge: likelihood error bars bound the SNIS bias)**.** Let $\widehat{\mathbb{E}}[\phi]$ and $\widehat{\mathbb{E}}^{\mathrm{dep}}[\phi]$ be the SNIS estimates of an observable ϕ under the exact and the deployed likelihoods, and suppose per-sample certified bounds $|\delta(x_i)| \le C_i$ hold on the draws used. Then: (i) (first order) $\left|\widehat{\mathbb{E}}^{\mathrm{dep}}[\phi] - \widehat{\mathbb{E}}[\phi]\right| \le \sqrt{\mathrm{Var}_w(\phi)} \cdot \sqrt{\widehat{\mathbb{E}}_w[C_i^2]} + O(\delta^2)$; (ii) (non-asymptotic) if $C_i \le C$ uniformly, then $\left|\widehat{\mathbb{E}}^{\mathrm{dep}}[\phi] - \widehat{\mathbb{E}}[\phi]\right| \le e^{2C}\left(e^{2C} - 1\right) \widehat{\mathbb{E}}_w\left[|\phi - \widehat{\mathbb{E}}[\phi]|\right]$.

The proof (App. E.2) is a weight-perturbation argument: certified bounds confine every weight distortion factor to $[e^{-2C}, e^{2C}]$, and the observable's weighted dispersion converts that confinement into a bias bound. As written, both right-hand sides are under the exact-likelihood weights and are therefore not directly computable from a deployed-only run; evaluation requires same-sample exact weights or a separate conservative change-of-measure conversion to deployed weights. Two readings follow. First, valid same-sample error bars have end-to-end semantics: "every $|\delta_i| \le C_i$" implies a concrete, checkable error bar on the number the practitioner reports. Second, the lemma quantifies the tail sensitivity: the bound degrades exponentially in the largest certified error among high-weight draws, which is why the deployment recipe below escalates flagged draws rather than averaging over them.

A practical deployment recipe. In deployment: (1) use the theorem-backed three-level certificate equation 4 with the known order p as the default; (2) add a fourth solve when an on-the-spot estimate of $h^\star$ is wanted (a validated heuristic, Rem. 1); (3) screen draws with the free self-consistency residual ϱ and escalate flagged ones rather than certifying them; (4) read the convergence certificate, a same-sample likelihood diagnostic, and ESS as three distinct quantities. Algorithm 1 and App. G give the annotated checklist (Table 3).

5 Experiments

Setup. Experiment-specific fine references are used only for evaluation: $K_{\mathrm{ref}} = 96/128$ in the high-dimensional GMM sweep, 160 for FashionMNIST, 256 for MeanFlow/FALCON, 500 for the iDEM certificate sweeps, 512 for calibration and LJ13 downstream evaluation, and 1024 for the one-gap

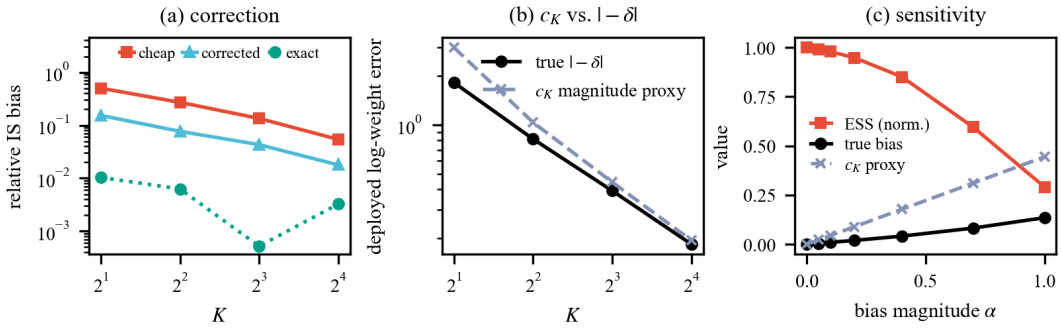


Figure 3: The same-sample diagnostic c_K detects and corrects a real few-step bias. On a MeanFlow map, c_K estimates the magnitude of the deployed log-weight error $-\delta$ in this controlled truncation; subtracting its signed version reduces IS error, while ESS reacts only after the bias is substantial.

Table 1: 0.05-tolerant empirical coverage on real energy-only samplers and on a faithful reimplementation of FALCON's model class. Per-K ranges cover the tested budgets; coverage means $\mathrm{frac}(C + 0.05 \geq E)$, with target ≥ 0.90, and tightness has target ≤ 5. The variant column states which quantity each row archives: GMM-40 and DW4 report the one-gap fixed-factor baseline $C_{\mathrm{gap}} = 5|g_K - g_{2K}|$ of Fig. 5, the only variant archived for those checkpoints; LJ13, MeanFlow, and FALCON report the known-order three-level heuristic C_3 of Fig. 7. Neither is the theorem certificate C_{sc}. GMM-40 exceeds the tightness target at $K \leq 20$. The FALCON row is a from-scratch reimplementation of its average-velocity map and hybrid cycle-consistency objective on a GMM with known density.

sampler	target (d)	variant	0.05-tolerant coverage	tightness $\mathrm{med}(C/E)$
iDEM	GMM-40 (d=2)	C_{gap}	0.91–0.96	3.4–8.1
iDEM	DW4 (d=8)	C_{gap}	0.93–1.00	3.5–5.0
iDEM	LJ13 (d=39)	C_3	0.91–0.98	3.5–5.1
MeanFlow	GMM (d=2, 8)	C_3	0.91–0.99	2.1–3.6
FALCON (reimpl.)	GMM (d=2, 8)	C_3	0.98–1.00	1.9–5.0

baseline (Fig. 5). Unless explicitly marked strict, archived fractions use 0.05-tolerant empirical coverage $\mathrm{frac}(C + 0.05 \geq E_K)$ (target ≥ 0.90), not mathematical soundness $\mathrm{frac}(C \geq E_K)$. We also report median tightness C/E_K (preferred target ≤ 5). Experiments span known-density GMMs, iDEM energy samplers (Akhound-Sadegh et al., 2024), a MeanFlow map (Geng et al., 2025), and molecular systems; implementation details and tail handling are in App. H.1.

Controlled validity and scale. On Euler GMMs C_{sc} has 0.05-tolerant empirical coverage 0.98–1.00 and tightness 1.6–2.3 for $d \leq 16$, remains at 0.98–1.00 through d=512, and has coverage 0.91–1.00 on deliberately under-trained flows. A distinct three-level heuristic, C_3, also covers Heun/RK2 (0.91–1.00, tightness ≤ 4; App. J). With a two-standard-deviation Hutchinson allowance it scales to d=3072 without forming a Jacobian (Fig. 4, top): the Hutchinson variant $C_{\mathrm{rich}} + 2\sigma$ has tolerant coverage 0.91–1.00 and tightness 1.1–2.1 on GMMs. On FashionMNIST the corresponding log-density-only evaluation has tolerant coverage 1.00 and tightness 1.9 at K=32 (Fig. 6); it does not test sample quality or downstream IS. These are empirical scale results, not dimension-free theoretical constants.

Real samplers. Table 1 reports 0.05-tolerant empirical coverage on iDEM, MeanFlow, LJ13, and a FALCON-class average-velocity map.

FALCON-class validation. Because the published FALCON checkpoint is unavailable, we reproduce its average-velocity map and cycle-consistency objective on known-density GMMs (Rehman et al., 2026). The resulting maps have cycle residual 0.02–0.04; C_3 has 0.05-tolerant empirical coverage 0.98–1.00 and tightness 1.9–5.0 (Table 1), while c_K follows the same-sample deployed log-weight error in this controlled truncation and its signed correction removes 68–86% of the IS bias (App. J). This validates the model class rather than a particular released checkpoint.

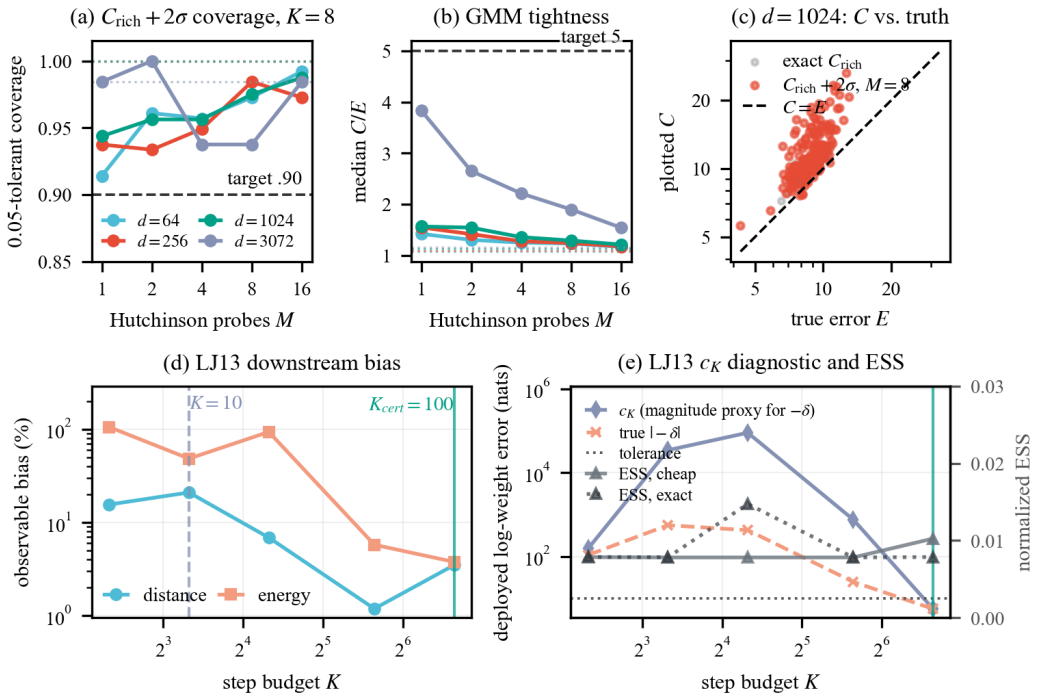


Figure 4: Scale and downstream utility. Top: 0.05-tolerant empirical coverage (a) and tightness (b) of $C_{\text{rich}} + 2\sigma$ versus the number of Hutchinson probes on GMMs up to d=3072 (dotted: exact-Jacobian C_{rich}), and (c) the per-draw scatter of the plotted C against the true error at d=1024. Bottom: the same-sample diagnostic c_K on LJ13 selects K=100, where both physical observables approach their exact-likelihood values (d), while ESS sits near its degeneracy floor and cannot rank the budgets (e). Full sweeps and controls are in App. J.

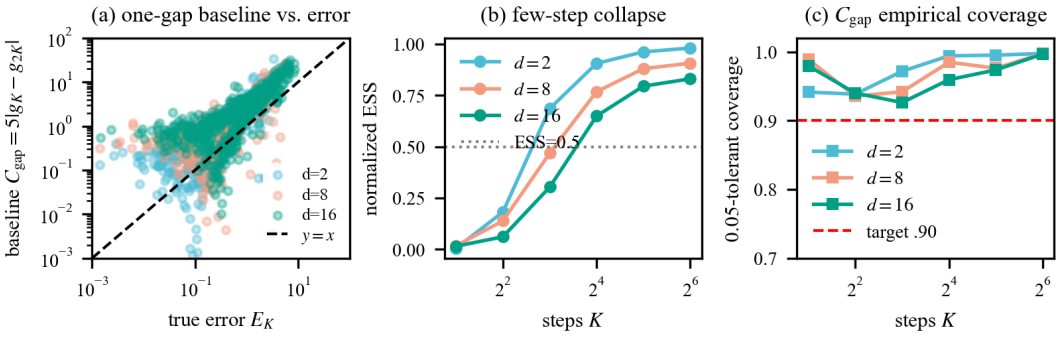


Figure 5: One-gap baseline. This figure plots the fixed-factor baseline $C_{\text{gap}} = 5|g_K - g_{2K}|$; panel (c) reports its 0.05-tolerant empirical coverage $\text{frac}(C_{\text{gap}} + 0.05 \geq E_K)$.

ESS-invisible bias and budget selection. On Euler GMM and a real MeanFlow map (Fig. 3), c_K follows the within-setting magnitude of $-\delta$ while ESS can remain effectively unchanged; signed correction reduces IS bias 3–7×. In this one-off pilot, the c_K rule matches the retrospective grid threshold across tested tolerances. ESS accepts a budget 3–8× smaller (App. J).

Physical payoff. On LJ13, the selected K=100 budget has asymptotic production cost 0.20× the K=512 reference after amortizing a one-off pilot and estimates mean interatomic distance as 1.473 ± 0.013, within 2.9% of the reference 1.517 ± 0.009; the ESS-selected K=5 estimate is 68% off (Fig. 4, bottom). At molecular few-step budgets c_K remains a conservative detector, but its Taylor correction is reliable only once $K \geq 50$.

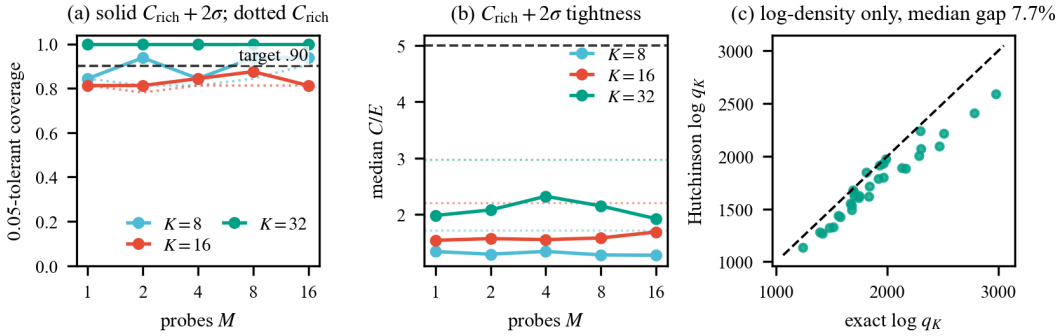


Figure 6: FashionMNIST log-density-only evaluation. Solid curves use $C_{\text{rich}} + 2\sigma$ and dotted curves use C_{rich}; rates are 0.05-tolerant empirical coverage. Panel (c) compares Hutchinson and exact log-density values.

These results rely on the known solver order and the two-solve margin. The theorem controls coupled log-weight convergence, whereas c_K controls the specific same-sample truncation studied here. Detailed ablations and failure modes are in Apps. H.1 and J.

6 Conclusion

Few-step likelihoods enter importance weights without a convergence test, and ESS cannot supply one: its first-order response to a likelihood error can be orthogonal to the direction that biases an observable. We separate coupled discretization drift, deployed-likelihood error, and downstream bias, and give each its own instrument. Theorem 2 certifies the first, reference-free and constant-free, above an explicit threshold; the same-sample diagnostic c_K detects and corrects the second, reducing IS error 3–7$\times$ on a real average-velocity map; and Lemma 4 converts per-sample error bars into a bound on the third.

On LJ13, the resulting budget rule reaches reference-quality physics at $0.20\times$ amortized production cost, whereas the ESS-selected budget is 68% off. The certificate is therefore a convergence and triage tool rather than a replacement for model-specific likelihood checks: its theorem is conditional on the stated solver expansion, and the truncation correction is reliable only in the regime where that expansion converges (App. H.1).

AI use statement

Generative AI tools were used for manuscript polishing and typesetting, organizing code and experimental results, and code standardization. The authors reviewed all AI-assisted work and take responsibility for the final manuscript.

Reproducibility statement

The assumptions, complete proofs, symbolic identities, and numerical verification of the certificate are documented in Apps. B, D, E, and E.2. Fixed seeds, shared source batches, full experimental settings, extended sweeps, and the amortized cost protocol are specified in Apps. K, J, and I. The high-dimensional GMM sweep and an independent LJ13 reproduction were also run on a second machine and agreed on the reported empirical coverage and ESS-invisible-bias conclusions.

References

Tara Akhound-Sadegh, Joey Rector-Brooks, Avishek Joey Bose, Sarthak Mittal, Pablo Lemos, Cheng-Hao Liu, Marcin Sendera, Siamak Ravanbakhsh, Gauthier Gidel, Yoshua Bengio, Nikolay Malkin, and Alexander Tong. Iterated denoising energy matching for sampling from Boltzmann densities. In *International Conference on Machine Learning (ICML)*, 2024.

Ricky TQ Chen, Yulia Rubanova, Jesse Bettencourt, and David K Duvenaud. Neural ordinary differential equations. *NeurIPS*, 2018.

Sitan Chen, Sinho Chewi, Jerry Li, Yuanzhi Li, Adil Salim, and Anru R Zhang. Sampling is as easy as learning the score: theory for diffusion models with minimal data assumptions. In *International Conference on Learning Representations (ICLR)*, 2023.

Maximilian Dax, Stephen R Green, Jonathan Gair, Michael Pürrer, Jonas Wildberger, Jakob H Macke, Alessandra Buonanno, and Bernhard Schölkopf. Neural importance sampling for rapid and reliable gravitational-wave inference. *Physical Review Letters*, 130(17):171403, 2023.

Zhao Ding, Chenguang Duan, Yuling Jiao, Ruoxuan Li, Jerry Zhijian Yang, and Pingwen Zhang. Characteristic learning for provable one step generation. *arXiv preprint arXiv:2405.05512*, 2024.

Kevin Frans, Danijar Hafner, Sergey Levine, and Pieter Abbeel. One step diffusion via shortcut models. *ICLR*, 2025.

Zhengyang Geng, Mingyang Deng, Xingjian Bai, J. Zico Kolter, and Kaiming He. Mean flows for one-step generative modeling. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025. arXiv:2505.13447.

Will Grathwohl, Ricky T. Q. Chen, Jesse Bettencourt, Ilya Sutskever, and David Duvenaud. FFJORD: Free-form continuous dynamics for scalable reversible generative models. In *International Conference on Learning Representations (ICLR)*, 2019.

Ernst Hairer, Syvert P Nørsett, and Gerhard Wanner. *Solving Ordinary Differential Equations I: Nonstiff Problems*. Springer, 1993.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *NeurIPS*, 2020.

Michael F Hutchinson. A stochastic estimator of the trace of the influence matrix for Laplacian smoothing splines. *Communications in Statistics - Simulation and Computation*, 18(3):1059–1076, 1989.

Pascal Jutras-Dubé, Jiaru Zhang, Ziran Wang, and Ruqi Zhang. One-step diffusion samplers via self-distillation and deterministic flow. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2026.

Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the design space of diffusion-based generative models. *NeurIPS*, 2022.

Jonas Köhler, Leon Klein, and Frank Noé. Equivariant flows: Exact likelihood generative learning for symmetric densities. In *International Conference on Machine Learning (ICML)*, 2020.

Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. *ICLR*, 2023.

Cheng Lu and Yang Song. Simplifying, stabilizing and scaling continuous-time consistency models. *ICLR*, 2025.

Luca Martino, Víctor Elvira, and Francisco Louzada. Effective sample size for importance sampling based on discrepancy measures. *Signal Processing*, 131, 2017.

Laurence Illing Midgley, Vincent Stimper, Gregor N C Simm, Bernhard Schölkopf, and José Miguel Hernández-Lobato. Flow annealed importance sampling bootstrap. In *International Conference on Learning Representations (ICLR)*, 2023.

Dat Nguyen-Cong, Tung Kieu, and Hoang Thanh-Tung. FastDiSS: Few-step match many-step diffusion language model on sequence-to-sequence generation. In *Findings of the Association for Computational Linguistics: ACL*, 2026.

Frank Noé, Simon Olsson, Jonas Köhler, and Hao Wu. Boltzmann generators: Sampling equilibrium states of many-body systems with deep learning. *Science*, 365(6457), 2019.

Art B Owen. *Monte Carlo theory, methods and examples*. 2013.

Danyal Rehman, Tara Akhound-Sadegh, Artem Gazizov, Yoshua Bengio, and Alexander Tong. FALCON: Few-step accurate likelihoods for continuous flows. In *International Conference on Learning Representations (ICLR)*, 2026.

Lorenz Richter and Julius Berner. Improved sampling via learned diffusions. In *International Conference on Learning Representations (ICLR)*, 2024.

Andrea Schioppa and Tim Salimans. Covariance-aware sampling for diffusion models. *arXiv preprint arXiv:2605.13910*, 2026.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *ICLR*, 2021a.

Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *ICLR*, 2021b.

Yang Song, Prafulla Dhariwal, Mark Chen, and Ilya Sutskever. Consistency models. In *International Conference on Machine Learning (ICML)*, 2023.

Luhan Tang, Longxuan Yu, Shaorong Zhang, and Greg Ver Steeg. Is your diffusion sampler actually correct? a sampler-centric evaluation of discrete diffusion language models. In *International Conference on Machine Learning (ICML)*, 2026.

Aki Vehtari, Daniel Simpson, Andrew Gelman, Yuling Yao, and Jonah Gabry. Pareto smoothed importance sampling. *Journal of Machine Learning Research (JMLR)*, 25(72), 2024. arXiv:1507.02646.

Hao Wu, Jonas Köhler, and Frank Noé. Stochastic normalizing flows. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

Zhiyang Xun and Eric Price. Query lower bounds for diffusion sampling. *arXiv preprint arXiv:2604.10857*, 2026.

Liangyu Yuan, Ruoyu Wang, Tong Zhao, Dingwen Fu, Mingkun Lei, Beier Zhu, and Chi Zhang. Few-step diffusion sampling through instance-aware discretizations. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2026.

A Notation and standing assumptions

Table 2 collects the symbols used throughout. Two intentional overloads, inherited from the main text, deserve a word of caution: (i) the letter p denotes both the target density $p(x)$ (always written with an argument) and the solver order p (a positive integer, never with an argument); and (ii) M_p is the 2×2 gap matrix of equation 3, while M (no subscript) is the scalar that bounds the remainder of the expansion. The context makes each usage unambiguous.

We restate the standing assumption for self-containedness.

Assumption 2 (order-p expansion; restatement of Assumption 1)**.** For a one-step solver of known order p and a fixed source draw z, there is $h_0 > 0$ such that for all $h \in (0, h_0]$ the log-weight functional $g(h) = g_K(z)$ ($h = 1/K$) admits $g(h) = g_0 + a\,h^p + b\,h^{p+1} + r(h)$ with $g_0 = g_\infty(z)$, $b \neq 0$, and $|r(h)| \leq M\,h^{p+2}$, where a, b, M depend on z but not on h.

The justification of Assumption 2 from standard ODE-solver theory is given in App. D.

B Proof of Theorem 2 and verification

We prove Theorem 2 in full, in the order: (B.1) the gap system and the matrix M_p; (B.2) its determinant and inverse; (B.3) the exact error identities for $\hat{A} - A$ and $\hat{B} - B$; (B.4) the remainder bounds on ρ_1, ρ_2; (B.5) the constants Γ_p, Γ'_p and their exact rational values; (B.6) soundness; (B.7) tightness; (B.8) machine verification. Throughout, Assumption 2 holds and we write $s = 2^{-p}$, $t = 2^{-(p+1)} = s/2$, $u = 2^{-(p+2)} = s/4$, $A = ah^p$, $B = bh^{p+1}$.

Table 2: Notation used in the main text and appendices.

symbol	meaning
$z \sim \mathcal{N}(0, I)$	source (prior) draw at τ=1
$v(x, \tau), \tau \in [0, 1]$	velocity field of the flow
T_K, $x_K = T_K(z)$	K-step solver map and its output sample
K, $h = 1/K$	step budget and step size
p (integer)	known order of the one-step solver (Euler p=1, Heun/RK2 p=2)
$p(x)$	(unnormalized) target density
$q_K(x)$, $q_\infty(x)$	few-step and continuous-flow ($K \to \infty$) model likelihoods
$J_k = \nabla_x$(k-th step)	per-step Jacobian; for Euler $J_k = I - h\,\nabla_x v$
$g(h) = g_K(z)$	log-weight functional $\log p(x_K) - \log q_K(x_K)$ at step h
$g_0 = g_\infty(z)$	its $K \to \infty$ limit (the "reference")
$E_K(z) = \lvert g(h) - g_0 \rvert$	coupled log-weight discretization drift (O1); not deployed-likelihood error (O2)
a, b	leading and next coefficients: $g(h) = g_0 + ah^p + bh^{p+1} + r(h)$
$A = ah^p$, $B = bh^{p+1}$	the leading and next terms (step-scaled)
$r(h)$, M	remainder and its bound: $\lvert r(h) \rvert \le Mh^{p+2}$
$s = 2^{-p}$, $t = s/2$, $u = s/4$	the level ratios $2^{-p}, 2^{-(p+1)}, 2^{-(p+2)}$
Δ_1, Δ_2	log-weight gaps $g(h) - g(h/2)$, $g(h/2) - g(h/4)$
ρ_1, ρ_2	remainder gaps $r(h) - r(h/2)$, $r(h/2) - r(h/4)$
M_p	gap matrix of equation 3: rows $(1{-}s,\ 1{-}t)$ and $(s(1{-}s),\ t(1{-}t))$
$\hat{A}, \hat{B}, \hat{C}$	reference-free estimates of the first three terms
$C_{\mathrm{sc}} = \lvert\hat{A} + \hat{B}\rvert + \lvert\hat{B}\rvert$	the constant-free certificate equation 4
Γ_p, Γ'_p	remainder constants (Thm. 2); $\Gamma_1{=}\frac{17}{8}$, $\Gamma'_1{=}\frac{15}{4}$
$h^\star = \lvert b\rvert/((\Gamma_p + \Gamma'_p)M)$	soundness threshold
s_0	calibrated factor in the fourth-level empirical heuristic $\lvert\hat{A} + \hat{B}\rvert + s_0\lvert\hat{C}\rvert$
$w_i = p(x_i)/q_K(x_i)$	self-normalized importance weight
$\mathrm{ESS} = 1/(N\sum_i \tilde{w}_i^2)$	normalized effective sample size
$\delta(x) = \log q_K^{\mathrm{dep}} - \log q_K$	deployed-likelihood error (O2 of Sec. 4.1); $\bar{\delta}$ mean, $\tilde{\delta}$ fluctuation
c_K	magnitude proxy for deployed log-weight error $-\delta$ (one extra trace/step); not C_{sc}
λ	an eigenvalue of the per-step Jacobian $\nabla_x v$

B.1 The gap system

Evaluating the expansion at the three budgets $K, 2K, 4K$, i.e. at steps $h, h/2, h/4$, and using $(h/2^k)^p = h^p 2^{-pk} = A\,s^k$ and $(h/2^k)^{p+1} = h^{p+1}2^{-(p+1)k} = B\,t^k$,

$$g(h) - g_0 = A + B + r(h), \tag{5}$$

$$g(h/2) - g_0 = A\,s + B\,t + r(h/2), \tag{6}$$

$$g(h/4) - g_0 = A\,s^2 + B\,t^2 + r(h/4). \tag{7}$$

Subtracting consecutive lines gives the gaps

$$\Delta_1 = g(h) - g(h/2) = A(1-s) + B(1-t) + \rho_1, \qquad \rho_1 = r(h) - r(h/2), \tag{8}$$

$$\Delta_2 = g(h/2) - g(h/4) = A\,s(1-s) + B\,t(1-t) + \rho_2, \qquad \rho_2 = r(h/2) - r(h/4), \tag{9}$$

which is exactly equation 3 with $M_p = \left[\begin{smallmatrix} 1-s & 1-t \\ s(1-s) & t(1-t) \end{smallmatrix}\right]$.

B.2 Determinant and inverse

Expanding along the first row,

$$\det M_p = (1-s)\,t(1-t) - (1-t)\,s(1-s) = (1-s)(1-t)(t-s) = -\tfrac{1}{4}\,s(1-s)(2-s), \tag{10}$$

where the last equality uses $t = s/2$. Since $0 < u < t < s \le \frac{1}{2}$ for every integer $p \ge 1$, all three factors $1-s, 1-t, t-s$ are nonzero (the last is negative), so M_p is invertible. Its inverse is

$$M_p^{-1} = \frac{1}{\det M_p}\begin{bmatrix} t(1-t) & -(1-t) \\ -s(1-s) & 1-s \end{bmatrix}. \tag{11}$$

B.3 Exact error identities

Solving equation 3 without the remainder gives $\big[\hat{A}\,;\hat{B}\big] = M_p^{-1}\big[\Delta_1;\Delta_2\big] = \big[A;B\big] + M_p^{-1}\big[\rho_1;\rho_2\big]$, so the estimation errors are linear in (ρ_1,ρ_2):

$$\hat{A} - A = \frac{s\,\rho_1 - 2\,\rho_2}{s(s-1)}, \qquad \hat{B} - B = \frac{4(\rho_2 - s\,\rho_1)}{s(s-2)}, \tag{12}$$

and, adding the two and simplifying ($s^2 - 3s + 2 = (1-s)(2-s)$),

$$(\hat{A} - A) + (\hat{B} - B) = \frac{(2-3s)\,\rho_1 + 2\,\rho_2}{(1-s)(2-s)}. \tag{13}$$

Both equation 12 and equation 13 are obtained by direct substitution of M_p^{-1} and are confirmed symbolically (§B.8). Because each is linear and homogeneous in (ρ_1,ρ_2) and each $\rho_i = O(h^{p+2})$, the estimates $\hat{A},\hat{B}$ recover A,B up to $O(h^{p+2})$; in particular $\hat{A}+\hat{B}$ recovers $g(h)-g_0$ to one order beyond the leading signal, which is Lemma 1.

B.4 Remainder bounds

Since $|r(\cdot)| \le M(\cdot)^{p+2}$ and $u = 2^{-(p+2)}$,

$$|\rho_1| \le |r(h)| + |r(h/2)| \le Mh^{p+2}\big(1 + 2^{-(p+2)}\big) = Mh^{p+2}(1+u), \tag{14}$$

$$|\rho_2| \le |r(h/2)| + |r(h/4)| \le Mh^{p+2}\big(2^{-(p+2)} + 2^{-2(p+2)}\big) = Mh^{p+2}\,u(1+u). \tag{15}$$

B.5 The constants Γ_p, Γ'_p

Claim (i) concerns $(\hat{A}+\hat{B}) - (g(h)-g_0)$. Writing $g(h)-g_0 = A+B+r(h)$ and using equation 13,

$$\big(\hat{A}+\hat{B}\big) - \big(g(h)-g_0\big) = \frac{(2-3s)\rho_1 + 2\rho_2}{(1-s)(2-s)} - r(h). \tag{16}$$

Taking absolute values, applying §B.4 and $|r(h)| \le Mh^{p+2}$,

$$\big|(\hat{A}+\hat{B}) - (g(h)-g_0)\big| \le \Gamma_p\,Mh^{p+2}, \quad \Gamma_p = \frac{|2-3s|}{(1-s)(2-s)}(1+u) + \frac{2}{(1-s)(2-s)}u(1+u) + 1, \tag{17}$$

the last $+1$ being the contribution of $|r(h)|$. Likewise, from equation 12,

$$\big|\,|\hat{B}| - |b|h^{p+1}\,\big| \le |\hat{B} - B| \le \Gamma'_p\,Mh^{p+2}, \quad \Gamma'_p = \frac{4}{2-s}(1+u) + \frac{4}{s(2-s)}u(1+u), \tag{18}$$

using the reverse triangle inequality $\big|\,|\hat{B}| - |B|\,\big| \le |\hat{B} - B|$ and $|B| = |b|h^{p+1}$. These are claims (i)–(ii). Substituting $s = 2^{-p}$, $u = s/4$ gives exact rationals; for $p = 1$ ($s = \frac{1}{2}, u = \frac{1}{8}$),

$$\Gamma_1 = \tfrac{2/3}{1}\cdot\tfrac{9}{8} + \tfrac{8/3}{1}\cdot\tfrac{1}{8}\cdot\tfrac{9}{8} + 1 = \tfrac{3}{4} + \tfrac{3}{8} + 1 = \tfrac{17}{8}, \tag{19}$$

$$\Gamma'_1 = \tfrac{8}{3}\cdot\tfrac{9}{8} + \tfrac{16}{3}\cdot\tfrac{1}{8}\cdot\tfrac{9}{8} = 3 + \tfrac{3}{4} = \tfrac{15}{4}, \tag{20}$$

and similarly $\Gamma_2 = \frac{355}{168}$, $\Gamma'_2 = \frac{85}{28}$, $\Gamma_3 = \frac{577}{280}$, $\Gamma'_3 = \frac{11}{4}$; both sequences decrease in p (the higher the order, the smaller the relative remainder).

B.6 Soundness

By claim (i) and the reverse triangle inequality, $|\hat{A}+\hat{B}| \ge |g(h)-g_0| - \Gamma_p Mh^{p+2} = E_K - \Gamma_p Mh^{p+2}$. By claim (ii), $|\hat{B}| \ge |b|h^{p+1} - \Gamma'_p Mh^{p+2}$. Adding,

$$C_{\mathrm{sc}} = |\hat{A}+\hat{B}| + |\hat{B}| \;\ge\; E_K + |b|h^{p+1} - (\Gamma_p + \Gamma'_p)Mh^{p+2}. \tag{21}$$

The surplus $|b|h^{p+1} - (\Gamma_p+\Gamma'_p)Mh^{p+2}$ is nonnegative exactly when $|b|h^{p+1} \ge (\Gamma_p+\Gamma'_p)Mh^{p+2}$, i.e. $h \le h^\star := |b|/((\Gamma_p+\Gamma'_p)M)$. Hence $C_{\mathrm{sc}} \ge E_K$ for every $h \le h^\star$, which is claim (iii). The mechanism is transparent: the safety margin $|\hat{B}|$ is order h^{p+1}, while everything left uncertified is order h^{p+2}; soundness holds once the step is fine enough that the former dominates the latter.

B.7 TIGHTNESS

Suppose $a \neq 0$. Then $E_K = |g(h) - g_0| \sim |a|h^p$ as $h \to 0$, while $|\hat{A} + \hat{B}| = E_K + O(h^{p+2})$ and $|\hat{B}| = |b|h^{p+1} + O(h^{p+2}) = o(E_K)$. Therefore $C_{\mathrm{sc}}/E_K = \big(E_K + o(E_K)\big)/E_K \to 1$, which is claim (iv). If instead $a = 0$ (the degenerate case of App. H), then $E_K \sim |b|h^{p+1}$, both $|\hat{A} + \hat{B}|$ and $|\hat{B}|$ tend to $|b|h^{p+1}$, and $C_{\mathrm{sc}}/E_K \to 2$: the certificate stays sound but is asymptotically a factor of two loose.

B.8 MACHINE VERIFICATION

Every algebraic step above is checked symbolically with SymPy (the verification script): the determinant and its factorization, the inverse, the error identities equation 12–equation 13, the extraction of the coefficients of (ρ_1, ρ_2), and the exact rational values of Γ_p, Γ'_p for $p \in \{1, 2, 3\}$. Soundness is then checked numerically on synthetic $g(h) = g_0 + ah^p + bh^{p+1} + ch^{p+2}$ (so $r(h) = ch^{p+2}$, $M = |c|$): for both $p = 1$ and $p = 2$, $C_{\mathrm{sc}} \geq E_K$ holds at every tested $h \leq h^\star$ and the bound $|(\hat{A} + \hat{B}) - (g(h) - g_0)| \leq \Gamma_p M h^{p+2}$ is never violated. Crucially, the test includes an adversarial configuration with a large opposite-sign remainder (b and c of opposite sign, $|c|$ large): there $C_{\mathrm{sc}} \geq E_K$ holds for all $h \leq h^\star$ and fails just outside, e.g. a $p = 2$ instance violates soundness at $h = 0.25 > h^\star \approx 0.078$ while satisfying it for every $h \leq h^\star$. This confirms that $h^\star$ is non-vacuous: it is not merely a sufficient condition with slack, but tight enough that soundness can genuinely fail beyond it. □

C GENERAL-ORDER RICHARDSON EXTRAPOLATION AND THE FOURTH-LEVEL HEURISTIC

The three-level construction of Sec. 4 is the first nontrivial member of a family of reference-free extrapolants obtained by reading off more terms of the expansion from more short solves. This appendix gives the general construction and the fourth-level heuristic referenced in Remark 1. It does not make the remainder bound M identifiable or rigorously computable.

C.1 THE $(L+1)$-LEVEL SCALED VANDERMONDE SYSTEM

Fix z and evaluate g at the $L+1$ geometrically refined budgets $K, 2K, \ldots, 2^L K$, i.e. at steps $h/2^j$, $j = 0, \ldots, L$. Extending Assumption 2 to L leading terms,

$$g(h/2^j) - g_0 = \sum_{m=0}^{L-1} A_m \, r_m^j + R_j, \qquad A_m := a_m \, h^{p+m}, \qquad r_m := 2^{-(p+m)}, \qquad |R_j| \leq M \, (h/2^j)^{p+L}, \tag{22}$$

where $a_0 = a$, $a_1 = b$, $a_2 = c$, … are the expansion coefficients and $r_0 = s$, $r_1 = t$, $r_2 = u$. The L consecutive gaps $\Delta_j := g(h/2^{j-1}) - g(h/2^j)$, $j = 1, \ldots, L$, then satisfy a linear system $\Delta = VA + \rho$ with

$$V_{j,m} = r_m^{j-1}(1 - r_m), \qquad j = 1, \ldots, L, \;\; m = 0, \ldots, L-1. \tag{23}$$

Factoring out $1 - r_m$ from each column exposes $V = W \, \mathrm{diag}(1 - r_0, \ldots, 1 - r_{L-1})$ with $W_{j,m} = r_m^{j-1}$ a Vandermonde matrix in the nodes $r_0, \ldots, r_{L-1}$.

Proposition 5 (invertibility of the general system)**.** For every integer $p \geq 1$ and every $L \geq 1$, the nodes $r_m = 2^{-(p+m)}$ are distinct and lie in $(0, 1)$, and

$$\det V = \Big(\prod_{m=0}^{L-1} (1 - r_m) \Big) \prod_{0 \leq m < m' \leq L-1} (r_{m'} - r_m) \neq 0. \tag{24}$$

Hence the gaps determine $(A_0, \ldots, A_{L-1})$ uniquely, and discarding ρ yields estimates $(\hat{A}_0, \ldots, \hat{A}_{L-1})$ with $\hat{A}_m - A_m$ linear in ρ, hence $O(h^{p+L})$.

For $L = 2$ this is $\det V = (1-s)(1-t)(t-s) = \det M_p$, recovering Lemma 1. For $L = 3$,

$$\det V = (1-s)(1-t)(1-u)\,(t-s)(u-s)(u-t) = \tfrac{3}{256}\, s^3(s-1)(s-2)(s-4), \qquad (25)$$

both forms machine-verified symbolically. The two-term and three-term solves are thus the $L \in \{2, 3\}$ cases.

C.2 The fourth level supplies heuristic proxies for M and $h^\star$

The single non-computable input in Theorem 2 is the remainder bound M. A fourth solve at $h/8$ ($L = 3$) recovers the next term $\hat{C} \approx C = c\,h^{p+2}$ from the 3×3 system above. Because the leading behavior of the remainder of the two-term model is precisely $c\,h^{p+2}$, one has $\hat{C} = c\,h^{p+2}(1 + O(h))$, so

$$\widehat{M} := |\hat{C}|/h^{p+2} = |c|\,(1 + O(h)), \qquad \hat{h}^\star := \frac{|\hat{B}|}{(\Gamma_p + \Gamma'_p)\,h\,\widehat{M}} = \frac{|\hat{B}|\,h}{(\Gamma_p + \Gamma'_p)\,|\hat{C}|}, \qquad (26)$$

both computed from the four short solves alone. We are precise about the epistemic status of this step: $|\hat{C}|$ estimates the leading coefficient of the remainder, not a rigorous upper bound on the full remainder (M majorizes all higher terms as well). Thus $\widehat{M}$ and $\hat{h}^\star$ are heuristic next-term proxies, not identified estimates or certified bounds. They provide a practical empirical gate; draws on which the gate or the self-consistency residual ϱ (App. H.5) fires are escalated, not certified.

C.3 The fourth-level empirical variant

Replacing the unknown M-dependent margin by the computable next term gives the empirical fourth-level variant of Remark 1,

$$C_{\text{full}}(z) = |\hat{A} + \hat{B}| + s_0\,|\hat{C}|, \qquad (27)$$

in which s_0 is a small data-calibrated factor multiplying a computable next term. This is an empirically validated heuristic rather than a theorem or a computable bound on M. The role of s_0 is to cover the truncation incurred by stopping the model at three terms. In the strict synthetic run (the self-contained certificate check), s_0=2 gives 1.00 strict coverage at every K with tightness 1.0–1.4, while s_0=1 gives about 0.8 strict coverage. The separate archived real-flow calibration uses 0.05 tolerance and reports coverage 0.99–1.00 for $d \in \{2, 8, 16\}$ and tightness 2.5–5 at small K, the ratio inflating only benignly at large K where the true error itself vanishes.

C.4 Conditioning and the practical choice of L

The nodes $r_m = 2^{-(p+m)}$ form a geometric sequence with ratio $\frac{1}{2}$, so the Vandermonde factor W becomes increasingly ill-conditioned as L grows (its condition number scales roughly like $\prod_{m<m'} |r_{m'} - r_m|^{-1}$). Simultaneously, the gaps Δ_j shrink toward the reference/floating-point floor at fine budgets, so the signal-to-noise ratio of the highest levels degrades. These two effects make $L \geq 4$ unprofitable in practice: the extra term is swamped by amplified noise. This is the same noise wall that makes data-estimating the order p (rather than using its known value) unreliable: at coarse K the next-order term inflates the estimate, at fine K the gaps are floor-limited. We therefore use $L = 2$ (the theorem-backed constant-free certificate) by default and $L = 3$ as a heuristic next-term gate; both are evaluated in App. J.

D Justification of the order-p expansion (Assumption 2)

Assumption 2 is the composition of two classical asymptotic expansions (one for the trajectory, one for the accumulated log-determinant) through smooth maps. We spell out both.

D.1 Trajectory expansion

Let the one-step method have consistent order p and let the velocity field v be C^{p+2} with bounded derivatives on the compact region the trajectories visit. The classical global-error expansion (Gragg's

theorem; see Hairer et al., 1993, II.8, and the asymptotic-expansion-of-the-global-error theorem) states that the numerical endpoint admits

$$x_K = x_\infty + \sum_{m=p}^{N} e_m(z)\, h^m + O(h^{N+1}), \tag{28}$$

where each coefficient function e_m solves a linear variational ODE driven by the elementary differentials of v and is independent of h. The expansion starts at h^p because the method is order p; for a symmetric method the odd coefficients vanish and the expansion proceeds in even powers, which only improves the certificate (fewer terms to track) and is subsumed by the generic integer-power statement we use.

D.2 Composition with the target log-density

The first half of g is $\log p(x_K)$. Since $\log p$ is smooth on the relevant region, a Taylor expansion around x_∞ together with the trajectory expansion yields

$$\log p(x_K) = \log p(x_\infty) + \langle \nabla \log p(x_\infty), e_p(z) \rangle\, h^p + O(h^{p+1}), \tag{29}$$

an integer-power expansion starting at h^p.

D.3 The log-determinant quadrature

The second half is $\log q_K = \log p_{\text{base}}(z) - \sum_k \log|\det J_k|$. For an Euler step $J_k = I - hJ(x_k, \tau_k)$ with $J = \nabla_x v$, the matrix logarithm identity $\log\det(I - X) = \operatorname{tr}\log(I - X) = -\sum_{m\geq 1} \operatorname{tr}(X^m)/m$ (convergent when the spectral radius $\rho(X) < 1$) gives

$$\log|\det(I - hJ)| = -h \operatorname{tr} J - \tfrac{h^2}{2}\operatorname{tr}(J^2) - \tfrac{h^3}{3}\operatorname{tr}(J^3) - \cdots, \qquad \rho(hJ) < 1. \tag{30}$$

Summing over the $K = 1/h$ steps, the leading term $-h\sum_k \operatorname{tr} J(x_k, \tau_k)$ is the left-rectangle quadrature of $-\int_0^1 \operatorname{tr} J(x_\tau, \tau)\, d\tau = -\int_0^1 \nabla\cdot v\, d\tau = \log q_\infty - \log p_{\text{base}}$. For a C^∞ integrand the left-rectangle rule has an Euler–Maclaurin expansion in integer powers of h,

$$h\sum_{k=0}^{K-1} f(\tau_k) = \int_0^1 f\, d\tau + \tfrac{h}{2}\big(f(1) - f(0)\big) + \sum_{\ell\geq 1} \frac{B_{2\ell}}{(2\ell)!} h^{2\ell}\big(f^{(2\ell-1)}(1) - f^{(2\ell-1)}(0)\big), \tag{31}$$

with Bernoulli numbers $B_{2\ell}$. Each higher term $-\frac{h^m}{m}\sum_k \operatorname{tr}(J^m)$ of equation 30 contributes a further $O(h^{m-1})$ to the sum (a sum of K terms each $O(h^m)$). All contributions are integer powers of h.

D.4 Combining

Adding the two halves, $g_K(z) = g_\infty(z) + a\, h^p + b\, h^{p+1} + r(h)$ with $|r(h)| \leq M h^{p+2}$ on $(0, h_0]$, where M collects bounds on the relevant derivatives of v and of $\log p$ along the trajectory, and $b \neq 0$ generically (the next-order coefficient does not vanish for a generic draw). This is Assumption 2. Two caveats are inherited and revisited in App. H: the expansion is asymptotic (useful once h is below the implicit radius), and the log-determinant series equation 30 requires $\rho(hJ) < 1$, i.e. $|h\lambda| < 1$ for every Jacobian eigenvalue λ, a condition that can be violated at molecular few-step budgets.

E Stochastic versus systematic likelihood error

This appendix proves Proposition 3 and Lemma 4, and develops the fuller picture promised in Sec. 4.5: ESS and an observable respond to different first-order projections of a likelihood error. A stochastic error typically inflates weight variance, while a deterministic error can be nearly orthogonal to the ESS direction yet correlated with the observable, which is the regime tested here.

E.1 Proof of Proposition 3

Let $w_i \propto p(x_i)/q_K(x_i) = \exp(g_K(z_i))$ be the exact self-normalized weights and $w_i^{\text{dep}} \propto w_i\, e^{-\delta(x_i)}$ those under the deployed likelihood, where $\delta = \log q_K^{\text{dep}} - \log q_K$. Claim (i). Self-normalization

cancels any additive constant in δ: replacing δ by $\delta - \bar{\delta}$ multiplies every unnormalized weight by the same factor $e^{\bar{\delta}}$, leaving every normalized weight, hence the SNIS estimate of every observable and the normalized ESS $\mathrm{ESS} = 1/(N\sum_i \tilde{w}_i^2)$, exactly unchanged. Both therefore depend on δ only through the fluctuation $\tilde{\delta} = \delta - \bar{\delta}$. Expanding and renormalizing gives $\tilde{w}_i^{\mathrm{dep}} = \tilde{w}_i(1-\tilde{\delta}_i) + O(\|\tilde{\delta}\|^2)$. Therefore, with $S = \sum_i \tilde{w}_i^2$,

$$\sum_i (\tilde{w}_i^{\mathrm{dep}})^2 - S = -2\sum_i \tilde{w}_i^2 \tilde{\delta}_i + O(\|\tilde{\delta}\|^2), \qquad \frac{\mathrm{ESS}^{\mathrm{dep}} - \mathrm{ESS}}{\mathrm{ESS}} = 2\frac{\sum_i \tilde{w}_i^2 \tilde{\delta}_i}{S} + O(\|\tilde{\delta}\|^2),$$

which proves Claim (iii). For Claim (ii), the change in the SNIS estimate of an observable ϕ is

$$\widehat{\mathbb{E}}^{\mathrm{dep}}[\phi] - \widehat{\mathbb{E}}[\phi] = -\operatorname{Cov}_w(\phi, \tilde{\delta}) + O(\tilde{\delta}^2), \tag{32}$$

generically nonzero and first-order in $\tilde{\delta}$. The ESS and observable terms are different linear functionals of $\tilde{\delta}$, so either can vanish while the other does not. □

E.2 Proof of Lemma 4 (the bridge to the estimation bias)

Write the deployed SNIS estimate as $\widehat{\mathbb{E}}^{\mathrm{dep}}[\phi] = \sum_i u_i \phi_i$ with $u_i = \tilde{w}_i e^{-\tilde{\delta}_i} / \sum_j \tilde{w}_j e^{-\tilde{\delta}_j}$, where $\tilde{w}_i$ are the exact normalized weights and $\tilde{\delta}_i = \delta_i - \bar{\delta}$ (the constant part cancels, by Proposition 3(i)). Since $|\delta_i| \le C_i$ implies $|\tilde{\delta}_i| \le 2\max_i C_i =: 2C$ (and, draw-wise, $|\tilde{\delta}_i| \le C_i + \bar{C}$ with $\bar{C} = \widehat{\mathbb{E}}_w[C_i]$), every distortion factor obeys $e^{-\tilde{\delta}_i} \in [e^{-2C}, e^{2C}]$.

Part (ii). Using $\sum_i \tilde{w}_i(\phi_i - \widehat{\mathbb{E}}[\phi]) = 0$,

$$\widehat{\mathbb{E}}^{\mathrm{dep}}[\phi] - \widehat{\mathbb{E}}[\phi] = \frac{\sum_i \tilde{w}_i \left(e^{-\tilde{\delta}_i} - 1\right)\left(\phi_i - \widehat{\mathbb{E}}[\phi]\right)}{\sum_j \tilde{w}_j e^{-\tilde{\delta}_j}}. \tag{33}$$

Bounding $|e^{-\tilde{\delta}_i} - 1| \le e^{2C} - 1$ in the numerator and $\sum_j \tilde{w}_j e^{-\tilde{\delta}_j} \ge e^{-2C}$ in the denominator gives $\left|\widehat{\mathbb{E}}^{\mathrm{dep}}[\phi] - \widehat{\mathbb{E}}[\phi]\right| \le e^{2C}(e^{2C}-1)\,\widehat{\mathbb{E}}_w[|\phi - \widehat{\mathbb{E}}[\phi]|]$, which is claim (ii); note the factor $e^{2C}(e^{2C}-1) = 2C \cdot (1 + O(C))$ for small C, so the non-asymptotic bound degrades gracefully into the first-order one.

Part (i). By Proposition 3(ii), $\widehat{\mathbb{E}}^{\mathrm{dep}}[\phi] - \widehat{\mathbb{E}}[\phi] = -\mathrm{Cov}_w(\phi, \tilde{\delta}) + O(\delta^2)$, and by Cauchy–Schwarz $|\mathrm{Cov}_w(\phi, \tilde{\delta})| \le \sqrt{\mathrm{Var}_w(\phi)}\sqrt{\mathrm{Var}_w(\tilde{\delta})}$. Finally $\mathrm{Var}_w(\tilde{\delta}) = \mathrm{Var}_w(\delta) \le \widehat{\mathbb{E}}_w[\delta^2] \le \widehat{\mathbb{E}}_w[C_i^2]$ (variance is minimized by centering, then $|\delta_i| \le C_i$ pointwise). All quantities on the right-hand sides are under the exact-likelihood weights. They require same-sample exact weights; a deployed-only implementation would need an additional conservative change-of-measure conversion using the C_i envelopes. □

Tail sensitivity. The uniform constant C in part (ii) is a maximum over the draws used; a single high-weight draw with a large certificate inflates it exponentially. This is a feature, not an artifact: importance sampling is dominated by its heaviest weights, so a bias guarantee must be sensitive to exactly those draws. The deployment recipe (Sec. 4.6) therefore escalates flagged draws (large ϱ or large C_i) to a finer budget rather than averaging them away, which shrinks C where it matters before the bound is applied.

E.3 Decomposing the cheap-likelihood error

A practical few-step likelihood replaces the exact per-step log-determinant in equation 1 by a cheap surrogate, incurring $\delta(x) = \delta^{\mathrm{stoch}}(x) + \delta^{\mathrm{sys}}(x)$.

Stochastic component (caught by ESS). A Hutchinson estimator replaces $\mathrm{tr}\, A$ by $\frac{1}{m}\sum_{l=1}^m \xi_l^\top A\, \xi_l$ with ξ_l Rademacher or Gaussian probes, an unbiased estimate whose variance is $\frac{2}{m}\|A\|_F^2$ for Gaussian probes. The induced δ^{stoch} is mean-zero and fluctuating, so it directly inflates $\mathrm{Var}_w(\tilde{\delta})$. By the proposition this lowers ESS (concretely, an independent log-weight perturbation of variance σ^2 multiplies the second weight moment by $\approx e^{\sigma^2}$, so $\mathrm{ESS} \to \mathrm{ESS}/(1+\chi^2)$ shrinks); the standard diagnostic already flags it. This is not the dangerous case.

Systematic component (invisible to ESS). A truncated log-determinant or an approximately-invertible learned map contributes a deterministic $\delta^{\mathrm{sys}}(x)$ with small fluctuation $\tilde{\delta}^{\mathrm{sys}}$ but generally nonzero $\mathrm{Cov}_w(\phi, \delta^{\mathrm{sys}})$. By the proposition ESS is essentially unchanged while $\widehat{\mathbb{E}}[\phi]$ is biased, a silent failure.

E.4 The first-order log-determinant as the controlled model

The cleanest controllable instance truncates equation 30 after its first term, $\log|\det(I - hJ)| \approx -h \operatorname{tr} J$, so the systematic per-step bias is exactly the next term, $+\frac{h^2}{2}\mathrm{tr}(J^2) + O(h^3)$. Summed over steps, the signed leading deployed log-weight error is

$$c_K(x) = \sum_k \tfrac{1}{2}\, h^2 \,\mathrm{tr}\big(J(x_k, \tau_k)^2\big), \tag{34}$$

one extra trace per step and no reference. Under $\delta = \log q_K^{\mathrm{dep}} - \log q_K$, this quantity estimates $g_K^{\mathrm{dep}} - g_K = -\delta$, not δ; the implementation uses its nonnegative magnitude as a diagnostic and subtracts the signed term from the deployed log weight wherever equation 30 converges (the convergence proviso is the subject of App. H).

F The ESS lower bound and its limitation

Theorem 2 converts a per-sample error bar into a bound on the converged effective sample size, but the conversion is exponential and typically vacuous at the small budgets where one would most like it. We give the bound and its limitation.

Lemma 6 (ESS band under a bounded log-weight error)**.** Suppose $|g_K(z) - g_\infty(z)| \le b$ for all z (e.g. $b = \sup_i C_{\mathrm{sc}}(z_i)$, computable). Writing $w_i = w_i^\infty e^{\delta_i}$ with $|\delta_i| \le b$, the normalized ESS satisfies the two-sided band

$$e^{-4b}\,\mathrm{ESS}(K) \;\le\; \mathrm{ESS}(\infty) \;\le\; e^{4b}\,\mathrm{ESS}(K). \tag{35}$$

In particular $\mathrm{ESS}(\infty) \ge e^{-4b}\,\mathrm{ESS}(K)$.

Proof. With $\mathrm{ESS} = (\sum_i w_i)^2/(N\sum_i w_i^2)$ and $w_i = w_i^\infty e^{\delta_i}$, $|\delta_i| \le b$: $\sum_i w_i \in [e^{-b}, e^b]\sum_i w_i^\infty$ and $\sum_i w_i^2 \in [e^{-2b}, e^{2b}]\sum_i (w_i^\infty)^2$. Hence $\mathrm{ESS}(K) \le e^{2b}(\sum_i w_i^\infty)^2/(Ne^{-2b}\sum_i (w_i^\infty)^2) = e^{4b}\mathrm{ESS}(\infty)$, giving the lower bound on $\mathrm{ESS}(\infty)$; the upper bound follows by symmetry. □

The ESS lower bound is weak in the tested regime. The factor e^{-4b} is exponentially sensitive to the pointwise bound b, which is a worst case over draws and is dominated by the few samples with the largest error. Empirically (the value-band check on GMM and Euler): the band is non-vacuous only when $b \lesssim 0.5$, i.e. at large K; for d=8, K=64 it gives $\mathrm{ESS}(\infty) \ge 0.23$ against a true 0.91; but at small/moderate K, $b \approx$ 3–10, so $e^{-4b} \approx 10^{-7}$–10^{-19} and the band collapses to uselessness. A per-sample Richardson extrapolation of the ESS fares no better: ESS is a nonlinear functional of the whole weight vector, and a handful of poorly extrapolated outliers tank the estimate. The converged ESS is therefore not cheaply recoverable at small K, a genuine boundary rather than a slack bound. Accordingly the deliverable headlined in the main text is the per-source log-weight convergence certificate (with high tolerant coverage and low tightness ratios in the tested regime), its go/no-go use, and the separate same-sample bias detection of App. E; the ESS band is reported only for completeness.

F.1 Disentangling a small ESS: proposal gap vs. likelihood bias (LJ13)

A recurring source of confusion is the interpretation of a small self-normalized ESS. The end-to-end error of a certificate-guided estimate, measured against the expensive fine reference, splits cleanly into two independent contributions,

$$\underbrace{\big|\,\widehat{\mathbb{E}}_{q_K^{\mathrm{cheap}}}[\phi] - \widehat{\mathbb{E}}_{q_{K_{\mathrm{ref}}}}[\phi]\,\big|}_{\text{end-to-end (vs. reference)}} \;\le\; \underbrace{\big|\,\widehat{\mathbb{E}}_{q_K^{\mathrm{cheap}}}[\phi] - \widehat{\mathbb{E}}_{q_K}[\phi]\,\big|}_{\text{likelihood bias (same-sample diagnostic)}} \;+\; \underbrace{\big|\,\widehat{\mathbb{E}}_{q_K}[\phi] - \widehat{\mathbb{E}}_{q_{K_{\mathrm{ref}}}}[\phi]\,\big|}_{\text{proposal-convergence gap (not certified)}}\;,$$

where q_K is the exact-log-determinant likelihood at the same budget K. In this controlled truncation experiment c_K diagnoses the first term; the second is a property of the proposal (the distance between

the K-step samples and the converged ones) and no budget rule on a fixed grid controls it. On LJ13 (d=39) the two terms are starkly separated. The deterministic-PF-ODE weights are near-degenerate, so the self-normalized ESS sits at the $1/N$ floor and is nearly identical under the cheap and exact likelihoods at every K; at that floor ESS cannot distinguish those likelihoods or rank the budgets. This small ESS is driven entirely by the proposal-convergence gap, not by the likelihood bias. At the certificate-guided budget the likelihood-bias component is statistically indistinguishable from the exact likelihood (measured bias $z \approx 0.1$ on the mean interatomic distance), while the residual end-to-end error is the uncertified proposal gap.

The practical conclusion, summarized in the recipe (Sec. 4.6, item 4) and visualized in Fig. 14, is that three checks are needed: the certificate tests convergence to the continuous pipeline, c_K diagnoses the controlled same-sample truncation error, and ESS reports weight concentration. A small value of one cannot substitute for the other two.

G THE FULL DEPLOYMENT CHECKLIST

Algorithm 1 Few-step log-weight convergence certificate (per source draw z)

1: **Input:** trained few-step map, budget K, known solver order p, safety factor s_0.
2: Run the sampler at budgets $K, 2K, 4K$ (and optionally $8K$); record $g(h), g(h/2), g(h/4)$ (and $g(h/8)$), where $g(\cdot) = \log p(x) - \log q(x)$ and $h = 1/K$.
3: Form gaps $\Delta_1 = g(h) - g(h/2)$, $\Delta_2 = g(h/2) - g(h/4)$; solve equation 3 for $(\hat{A}, \hat{B})$.
4: **Two-solve certificate:** $C_{\mathrm{sc}} = |\hat{A} + \hat{B}| + |\hat{B}|$ // sound for $h \le h^\star$ (Thm. 2)
5: **(Optional) heuristic gate:** solve the 3-term system from $\{g(h), \dots, g(h/8)\}$ for $(\hat{A}, \hat{B}, \hat{C})$; return $|\hat{A} + \hat{B}| + s_0|\hat{C}|$ (empirical next-term proxy; Rem. 1).
6: **Output:** certified error bar on the coupled log-weight convergence gap E_K; reject the budget if it exceeds tolerance.

We expand the deployment recipe of Sec. 4.6; the branches are mirrored in Algorithm 1.

1. **Default: known order, two extra refined solves.** If the solver order p is known (it usually is, by construction) and you can afford two additional finer solves at $2K, 4K$, use the three-level constant-free certificate $C_{\mathrm{sc}} = |\hat{A} + \hat{B}| + |\hat{B}|$ equation 4. It is theorem-backed (Thm. 2), carries no fitted constant, and is the recommended default.
2. **If you want an on-the-spot heuristic gate.** Add a fourth solve at $8K$ and use the four-level $C_{\mathrm{full}} = |\hat{A} + \hat{B}| + s_0|\hat{C}|$ (Rem. 1, App. C) with the conservative data-calibrated s_0=2 (App. J); this is a validated heuristic layered on the theorem-backed default, not itself a theorem.
3. **Heuristically screen for inconsistent scaling.** From the same solves, compute $\varrho(z) = |\Delta_2 - 2^{-p}\Delta_1|/|\Delta_1|$ (App. H); it is free. A large ϱ is an empirical warning that the observed gaps are inconsistent with leading order-p scaling; it does not identify $h > h^\star$. Reject the draw, escalate it to a larger budget, or inflate its certificate by $1 + \kappa\varrho$. On a clean (well-trained) high-d flow only a small tail is flagged, whereas a hard, stiff (under-trained) flow leaves a sizable pre-asymptotic tail (Fig. 13).
4. **Read convergence, same-sample error, and ESS separately.** A large convergence certificate means the coupled few-step sample-and-weight pipeline has not reached its continuous-flow reference, so enlarge the budget. A small convergence certificate does not by itself certify an arbitrary likelihood surrogate at x_K; use a same-sample diagnostic such as c_K for that question. ESS then reports concentration of the resulting weights, a complementary variance issue. Figure 1 illustrates a systematic same-sample error for which the observable shifts while ESS happens to remain unchanged.

H FAILURE MODES AND BOUNDARY CASES

We document the regimes where the construction degrades and the consequence of each.

Table 3: Deployment reading of the three diagnostics. τ is the user-selected error tolerance.

C_{sc}	c_K	ESS	action
$> \tau$	any	any	increase K: coupled pipeline not converged
$\leq \tau$	$> \tau$	any	refine the same-sample likelihood surrogate
$\leq \tau$	$\leq \tau$	low	likelihood acceptable; proposal/variance limited
$\leq \tau$	$\leq \tau$	healthy	accept the budget

H.1 Scope of each guarantee

(a) Correction is dimension-limited: the bias correction subtracts the next log-determinant series term, valid only while $|h\lambda| < 1$; at molecular few-step budgets this fails, so correction is a low-dimensional / large-K bonus, while detection is evaluated to d=39 and empirical coverage to d=165. (b) The ESS lower bound is weak at small K (App. F), which is why we do not headline it. (c) LJ55 (d=165) is not a usable test: despite extended training, the iDEM sampler plateaued far above usable quality, and its energy explosions violate the certificate's smoothness assumption, so we draw no conclusion there. (d) The threshold is estimated, not observed: $h^\star$ is not observable at deployment; the fourth-solve estimate $\hat{h}^\star$, the calibrated s_0 (≈ 2), and the ϱ-screen are validated heuristics that manage, not remove, this boundary, and only the two-solve certificate itself is theorem-backed. (e) Per-draw coverage vs. heavy tails: IS is dominated by its heaviest weights, so a certificate failure on a high-weight draw matters more than the per-draw fraction suggests. On the archived d=1024, K=8 GMM setting (fresh seeded draw batch, same protocol), re-weighting coverage by the self-normalised IS weights leaves the exact certificate at strict coverage 1.000 and moves the Hutchinson (M=8) certificate from 0.956 unweighted to 0.926 weight-weighted, with the top-weight decile at 0.938 and the single heaviest draw covered in both cases; the recipe's escalation and the tail-sensitive bridge bound (Lemma 4) remain the handles for regimes beyond this check.

H.2 Log-determinant series past its convergence radius (detection vs. correction)

The systematic-bias correction of App. E subtracts $c_K = \sum_k \frac{1}{2} h^2 \mathrm{tr}(J_k^2)$, the next term of equation 30. That series converges only when $\rho(hJ) < 1$, i.e. $|h\lambda| < 1$ for every Jacobian eigenvalue λ. At molecular few-step budgets (LJ13, d=39, K=5–20) the dynamics are stiff, $|h\lambda|$ can exceed 1, and the truncated correction no longer approximates the true (non-summable) tail: subtracting c_K can worsen the estimate. The correction becomes reliable only once K is large enough that $|h\lambda| < 1$ (empirically $K \gtrsim 50$ on LJ13). Crucially, detection is robust where correction is not: c_K remains an empirically conservative magnitude flag ($c_K \geq |-\delta|$ at every tested K on LJ13) because the ordering of magnitudes survives even when the series is not summable for a quantitative correction. We therefore lead with detection (validated to d=39) and scope correction as a low-dimensional / large-K bonus. Two further points: (a) this caveat concerns the log-determinant-order correction c_K, not the step-count certificate C_{sc} of Theorem 2, which concerns the distinct discretization error under its stated assumptions; the LJ13 result is finite-reference 0.05-tolerant empirical coverage, not a proof that $h \leq h^\star$. (b) a stiff-aware correction (exponential integrator on the trace, or a Padé log-determinant) would extend the convergence radius and is a natural next step.

H.3 The degenerate case $a = 0$

If the leading coefficient vanishes ($a = 0$, e.g. accidental cancellation or a sample on which the method behaves a half-order better), then $E_K \sim |b| h^{p+1}$ rather than $|a| h^p$. The two-term estimate still satisfies $|\hat{A} + \hat{B}| \to |b| h^{p+1}$ and the margin $|\hat{B}| \to |b| h^{p+1}$, so $C_{\mathrm{sc}}/E_K \to 2$ (cf. §B.7): the certificate remains sound but is asymptotically a factor of two loose rather than exact. This is the benign boundary of the tightness claim; it does not threaten soundness.

H.4 The degenerate case $b = 0$ (or a, b both near zero)

Theorem 2 assumes the next coefficient $b \neq 0$; this coefficient gives the safety margin $|\hat{B}|$ its h^{p+1} size and makes the threshold $h^\star = |b|/((\Gamma_p + \Gamma'_p)M)$ positive. We record the behavior when b

vanishes (or both a and b are near zero), and, as for the $a = 0$ case, treat it as a flagged-and-rejected boundary rather than a silent one.

The certifiable correction order changes when $b = 0$. If $b = 0$ but $a \neq 0$, the first certifiable correction is no longer $b\,h^{p+1}$ but the next nonzero term $c\,h^{p+2}$: the true error is $E_K \sim |a|h^p$ with leading correction at order h^{p+2}, so the two-solve margin $|\hat{B}| \to 0$ collapses and $h^\star \to 0$. The soundness guarantee of Theorem 2 then becomes vacuous, not because the certificate is wrong but because the h^{p+1} margin it relies on has shrunk into the h^{p+2} remainder it was meant to dominate. If both $a, b \approx 0$ (accidental cancellation, or a draw on which the method is locally super-convergent), the gaps Δ_1, Δ_2 are themselves dominated by the $O(h^{p+2})$ remainder and the two-solve estimate $|\hat{A} + \hat{B}|$ no longer tracks E_K.

Heuristic gating signals. The same three (or four) short solves provide suggestive but non-identifying signals. When $b = 0$ the gaps become near-proportional, $\Delta_2 \approx 2^{-p}\Delta_1$, so the self-consistency residual ϱ of App. H.5 is small, so by itself it cannot distinguish $b = 0$ from a clean fit. The tell-tale is instead a collapsing solved next term: the ratio $|\hat{B}|/(|\hat{A}| + |\hat{B}|) \to 0$ while ϱ stays small can gate a possible $b \approx 0$ case, but finite gaps do not identify the coefficient. When in addition $a \approx 0$, the observed order $\hat{p}(z) = \log_2 |\Delta_1/\Delta_2|$ becomes unstable: it is a ratio of two remainder-dominated, near-zero gaps and fluctuates from budget to budget rather than sitting at the known p; monitoring $\hat{p}$ against the known p across the $\{K, 2K, 4K, 8K\}$ ladder gives another heuristic warning. Neither signal makes M or the degenerate coefficient computable.

Consequence. A flagged draw is rejected or escalated to a larger budget, exactly as for the pre-asymptotic and $a = 0$ cases. When a self-contained margin is wanted on such a draw, the four-level variant (Rem. 1) is an empirical fallback: with $b = 0$ the margin should come from the first nonzero higher term, and $|\hat{C}|$ supplies a next-term proxy rather than a certified margin. Empirically these draws are rare (a measure-zero coincidence, perturbed only by finite precision), and, like $a = 0$, the case concerns tightness/threshold bookkeeping; the signals support conservative gating but add no theorem.

H.5 Near-caustic samples and the self-consistency residual

The expansion can be locally pre-asymptotic on draws where the trajectory's Jacobian varies rapidly within the window $[h/4, h]$ (near a caustic of the flow). Such draws are flagged by the computable self-consistency residual

$$\varrho(z) = \frac{|\Delta_2 - 2^{-p}\Delta_1|}{|\Delta_1|}, \tag{36}$$

which is small under clean leading order-p scaling and can be large otherwise. A gated draw can be excluded, escalated to a larger budget, or assigned the empirical inflation $1 + \kappa\varrho$ (we use $\kappa = 2$). Neither ϱ nor the fourth-level proxy identifies $h > h^\star$ or makes M computable. Figure 13 therefore reports only how often this heuristic gate fires: the clean flow has a smaller ϱ tail than the hard, stiffer flow at the same budget.

I Computational cost of Algorithm 1

Let one network evaluation (NFE) and one per-step log-determinant cost $\mathcal{C}_{\mathrm{NFE}}$ and $\mathcal{C}_{\mathrm{ld}}$. A K-step solve costs $K(\mathcal{C}_{\mathrm{NFE}} + \mathcal{C}_{\mathrm{ld}})$. The certificate runs additional solves only at finer budgets, so its overhead is a small constant multiple of a single few-step solve and is independent of the reference.

Table 4: Per-source-draw cost. The certificate never invokes the expensive reference K_{ref}.

quantity	solves	NFE (network evals)
one few-step likelihood	K	K
three-level certificate C_{sc}	$K, 2K, 4K$	$7K$
four-level self-contained C_{full}	$K, 2K, 4K, 8K$	$15K$
reference (validation only)	experiment-specific K_{ref}	K_{ref}

Three remarks. (i) Reference-free speedup, per draw. For a typical few-step budget $K{=}8$, the three-level certificate costs 56 NFE versus, for example, 512 for the LJ13 reference, a substantial saving that grows with K_{ref}/K. At large selected budgets the per-draw comparison can invert (e.g. $7K = 700 > 512$ at $K{=}100$), which is why the deployment accounting below is amortized, not per-draw. (ii) Per-step cost. An exact log-determinant is $O(d^3)$ per step; the bias-correction trace $\mathrm{tr}(J^2)$ adds $O(d^2)$ per step (or $O(md)$ with m Hutchinson probes when only matrix-vector products are available). The $2 \times 2/3 \times 3$ Richardson solve is $O(1)$ with coefficients that depend only on the known order p. (iii) Memory and batching. Jacobians are formed and detached one step at a time, bounding memory at $O(d^2)$; the entire procedure batches over the N source draws and is embarrassingly parallel, so wall-clock cost scales linearly in N with no per-draw coupling.

Amortized deployment accounting (the protocol behind the "fraction of the reference cost" claims). The certificate is a budget-selection instrument (per-draw costs in Table 4): it is evaluated once, on a pilot batch of n_0 source draws, over a geometric ladder of candidate budgets; production then runs at the selected budget K^* with no further certificate solves. Certifying one candidate budget K costs $7K$ NFE per pilot draw, and a full geometric ladder up to $K_{\max}$ costs at most $\sum_{j\geq 0} 7K_{\max}2^{-j} \leq 14K_{\max}$ NFE per pilot draw. The total cost of a certified production run of N draws is therefore

$$\underbrace{14\,K_{\max}\,n_0}_{\text{one-off selection}} + \underbrace{N\,K^*}_{\text{production}} \qquad \text{versus} \qquad N\,K_{\mathrm{ref}} \ \text{ for the reference,} \tag{37}$$

so the certified run is cheaper whenever $N \geq N^* := 14K_{\max}n_0/(K_{\mathrm{ref}} - K^*)$ and its cost ratio tends to K^*/K_{ref} as N grows. For the LJ13 setting of Sec. 5 ($K_{\max}{=}K^*{=}100$, $K_{\mathrm{ref}}{=}512$) this gives $N^* \approx 3.4\,n_0$ and an asymptotic ratio $100/512 \approx 0.20$, the figure quoted in the main text, which is thus an amortized production-cost ratio, valid once the run size exceeds a few multiples of the pilot batch, not a per-draw claim. For small one-shot batches ($N \lesssim N^*$) the selection overhead dominates and the certificate should be justified by its reliability, not by cost.

J Extended results

We summarize the per-setting numbers underlying Sec. 5. (i) High dimension (known-order certificate): on a full dimension sweep $d \in \{2, 8, 16, 32, 64, 128, 256, 512\}$, Euler has 0.05-tolerant coverage 0.98–1.00 with tight median $C/E \leq 2.2$ at every dimension, and RK2 has coverage 0.89–1.00 with tight median ≤ 4.0. (ii) Fit-independence: on a 300-step under-trained flow, coverage is 0.91–1.00, tightness 1.6–3.8. (iii) Order-generality: Euler ($p{=}1$) and RK2 ($p{=}2$) both have high empirical coverage and low tightness ratios across the sweep; Figure 7 uses C_3, not C_{sc}. (iv) LJ13 detection: per-K ($K{=}5, 10, 20, 50, 100$), ESS under the biased and exact likelihoods coincide; IS bias of mean interatomic distance is $0.16, 0.21, 0.13, 0.01, 0.0003$, while c_K upper-bounds the observed $|-\delta|$ magnitude at every K. The second-order correction reduces the bias only at $K \geq 50$. (v) FALCON-class correction: in the controlled truncation of Sec. 5, the signed c_K correction removes 68–86% of the IS bias of the target observable across the tested budgets.

Ablations (full). The three design choices summarized in Sec. 5 are validated as follows. (i) Known vs. estimated order. Using the known solver order p is essential. Data-estimating it is too noisy because the clean asymptotic window is narrow (coarse budgets inflate the estimate, fine budgets hit the reference/floating-point floor), whereas known p gives high archived tolerant coverage and low tightness. (ii) Safety margin. The two-solve margin $|\hat{B}|$ is the theorem-backed conservative choice; the self-contained variant trades it for a small factor s_0 on the computable next term ($s_0{=}2$ gives high empirical coverage at near-unit tightness; calibrated in Fig. 12). (iii) Constant-free vs. fixed factor. A naive fixed multiplier is marginal on LJ13 at an intermediate budget (a pre-asymptotic effect), whereas the order-aware certificate has high tolerant coverage at every budget there; the failure of the fixed factor, not of the certificate, is the culprit.

constant-free theorem certificate. Figure 8 reports the constant-free two-term certificate equation 4 across dimensions for both solvers. The two-term estimate is asymptotically exact (estimate/true $\to 1$), and archived 0.05-tolerant coverage of C_{sc} is 0.98–1.00 for Euler and 0.96–1.00 for RK2, with tightness in the first-order regime (1.6–2.3); for RK2 it retains high tolerant coverage but can be looser at moderate K where the leading and next terms nearly cancel (the $a \approx -b$ boundary), which the third level (self-contained variant) removes.

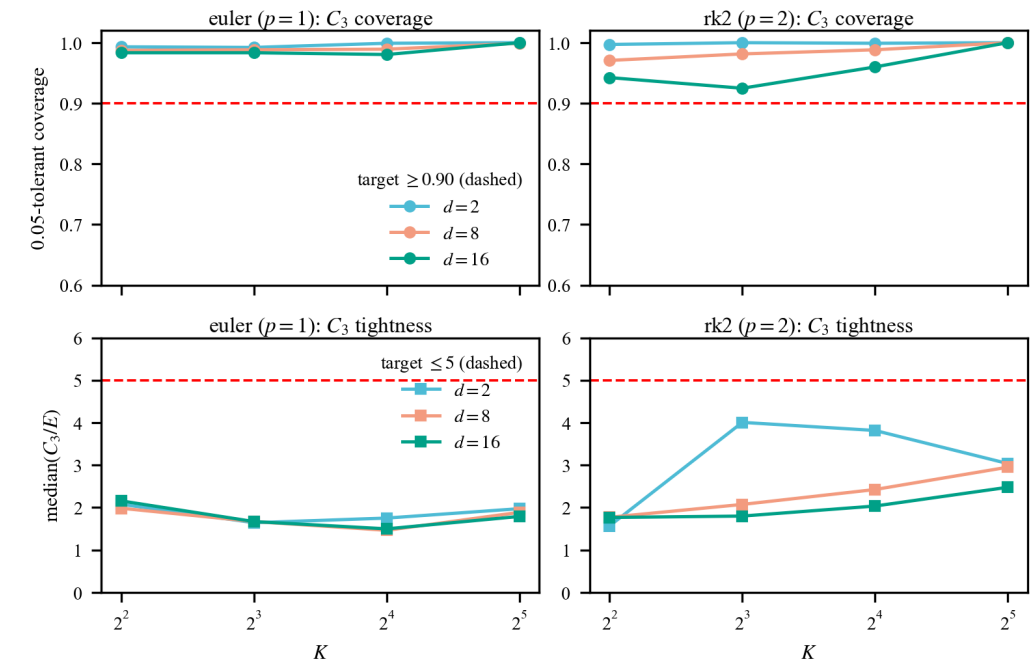


Figure 7: The plotted heuristic is $C_3 = \frac{|g_K - g_{2K}|}{1-2^{-p}}(1 + \kappa\varrho)$, not C_{sc}. The top row is archived 0.05-tolerant empirical coverage and the bottom row is tightness for Euler and Heun/RK2.

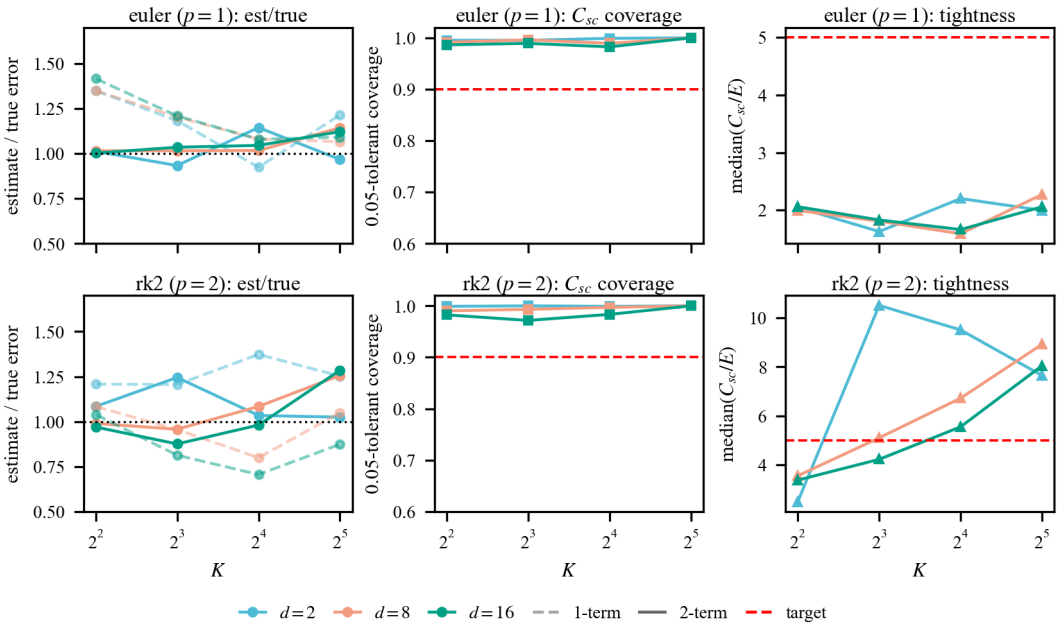


Figure 8: The theorem certificate C_{sc}. For Euler (p=1, top) and RK2 (p=2, bottom): estimate/true ratio (left), archived 0.05-tolerant empirical coverage $\mathrm{frac}(C_{\mathrm{sc}} + 0.05 \geq E)$ (center), and tightness (right). These rates do not replace the conditional mathematical soundness statement in Theorem 2.

Practitioner value: target accuracy at fewer steps. Figure 9 translates the detect-and-correct result on the real MeanFlow map into compute: certificate-corrected SNIS reaches the 5%-bias target at 2–4× fewer steps than the raw few-step likelihood (e.g. the corrected estimator meets the target at K=8, where the raw estimator still needs K=16–32).

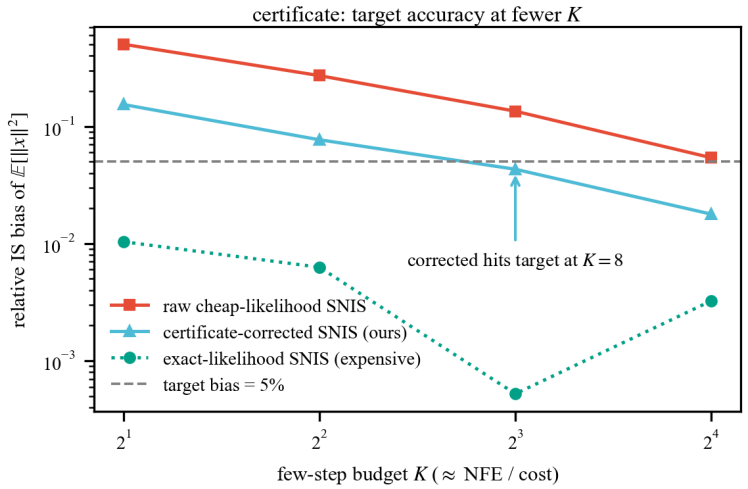


Figure 9: Compute Pareto using the same-sample diagnostic c_K on the real MeanFlow map (GMM d=8). Relative IS bias of $\mathbb{E}_p\|x\|^2$ versus budget K for the raw cheap-likelihood SNIS, the truncation-diagnostic-corrected SNIS, and the expensive exact-likelihood SNIS. Correcting the controlled truncation reaches the 5% target at 2–4× fewer steps.

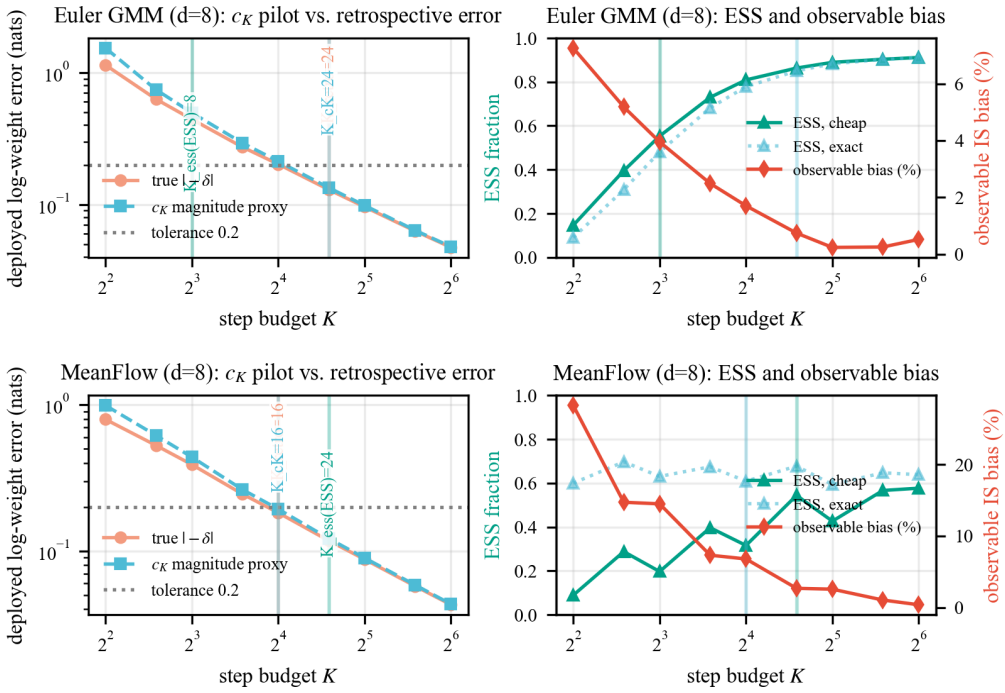


Figure 10: Pilot budget agreement for c_K. On the tested grid, the c_K rule matches the retrospective threshold defined using exact same-sample likelihoods. This empirical agreement is not a theorem-backed correction or an oracle guarantee; the variance-based rule selects a smaller budget with larger observed error.

Budget-acceptance decision quality. Figure 10 expands the pilot decision test of Sec. 5: on the tested grid, the c_K rule matches the retrospective exact-likelihood threshold across tolerances on both an Euler flow and the real MeanFlow map, whereas the variance-based diagnostic accepts a budget 3–8× too small (at which the likelihood error exceeds the tolerance and the estimate is biased), because the systematic error never enters the weight spread it measures.

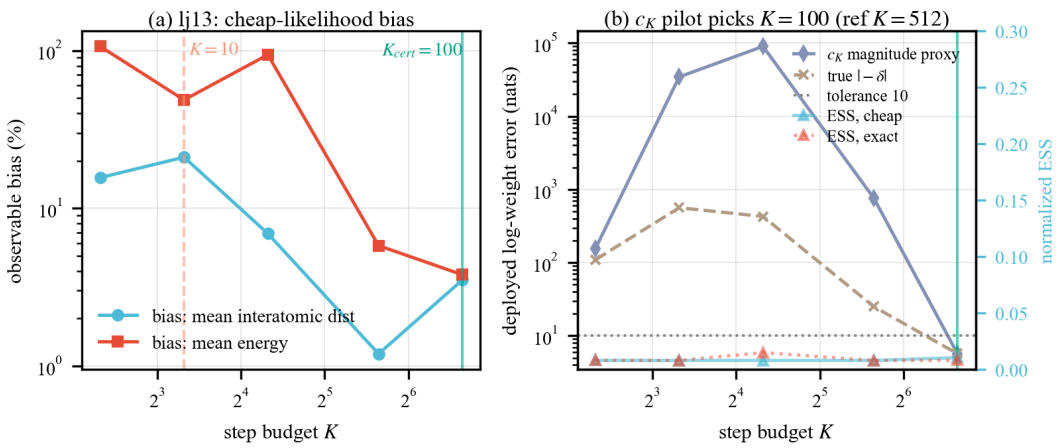


Figure 11: The c_K pilot reaches reference-quality physics at lower asymptotic production cost. Downstream error on LJ13 versus budget. After amortizing one-off budget selection, $K = 100$ tends to $100/512 \approx 0.20$ of reference production cost. Near the degeneracy floor, ESS cannot distinguish or rank the candidate budgets; the ESS-guided estimate is 68% off.

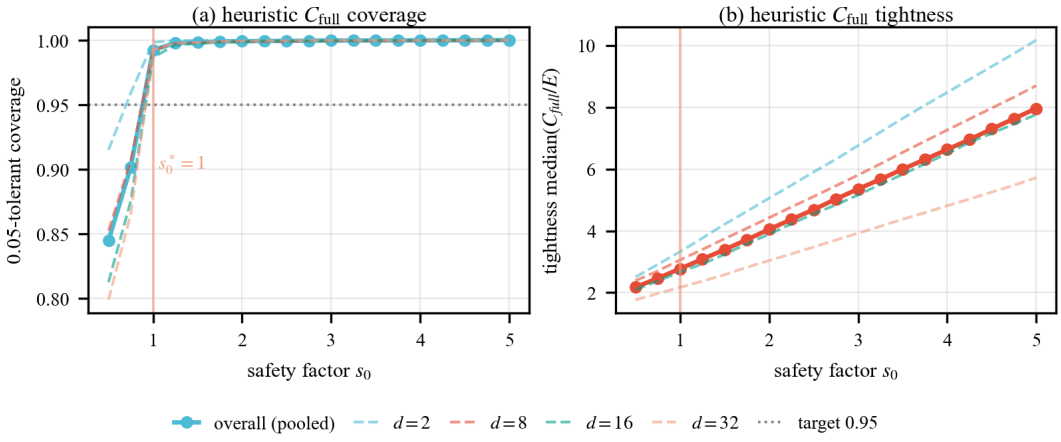


Figure 12: Calibration of the heuristic $C_{\text{full}} = |\hat{A} + \hat{B}| + s_0|\hat{C}|$: archived 0.05-tolerant empirical coverage and tightness versus s_0, pooled across dimensions and budgets. This does not make M computable.

Downstream estimand and cost. On the real LJ13 sampler (d=39), estimating a physical observable (mean interatomic distance and mean energy) with a certificate-guided budget attains a single-digit-percent relative error against the exact many-step reference at roughly one-fifth of its cost, whereas a fixed small budget incurs an order-of-magnitude larger error; throughout, the measured ESS under the biased and exact likelihoods coincides near the degeneracy floor, so it cannot distinguish or rank these budgets (Fig. 11). DW4 is consistent. We stress the caveat that molecular SNIS is variance-noisy at these budgets (small ESS); the coupled convergence error and the separately measured same-sample bias, not the raw estimate variance, are the clean quantities.

Calibration of the fourth-level heuristic factor. Pooling across dimensions and budgets, the smallest safety factor s_0 for the fourth-level empirical heuristic $|\hat{A} + \hat{B}| + s_0|\hat{C}|$ that achieves the target archived 0.05-tolerant coverage is close to one, with tightness well within range (Fig. 12); this calibrates, from data, the conservative value used earlier and confirms that no large constant is needed.

Self-consistency residual and heuristic gate firing rate. Figure 13 reports the distribution of the computable residual $\varrho(z) = |\Delta_2 - 2^{-p}\Delta_1|/|\Delta_1|$ (App. H.5) across source draws at a fixed coarse

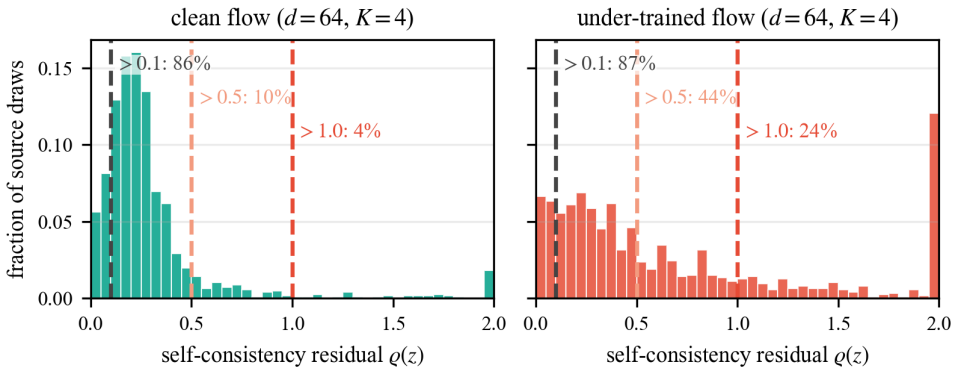


Figure 13: The self-consistency residual ϱ is a heuristic gate. Distribution of $\varrho(z) = |\Delta_2 - 2^{-p}\Delta_1|/|\Delta_1|$ over source draws at a fixed coarse few-step budget (d=64, K=4; dashed lines at the decision thresholds $\varrho = 0.1, 0.5, 1$, with the fraction of draws above each annotated). The clean flow has a smaller threshold-crossing tail than the under-trained flow (≈10%/4% versus $44\%/24\%$ above $0.5/1$). These are gate-firing rates, not identified pre-asymptotic fractions, and ϱ does not make M or $h^\star$ computable.

few-step budget, contrasting two flows of the same high dimension. On the clean setting (a well-trained d=64 GMM Euler flow) the order-p expansion holds and only a small tail crosses the chosen thresholds (≈10% above ϱ=0.5, ≈4% above 1). The deliberately under-trained, stiffer flow has a larger tail (≈44%/24%). This contrast motivates gating, but ϱ does not identify $h > h^\star$ or certify any draw.

The ESS–diagnostic phase plane across targets and budgets. Figure 14 pools every real-sampler setting (GMM, DW4, LJ13, and the MeanFlow map) and budget into one scatter of relative IS bias against measured ESS, coloured by the certificate value. It makes the central message visual: on benign targets (DW4, GMM, MeanFlow) ESS still tracks quality (low bias accompanies higher ESS), whereas on LJ13 the self-normalized ESS collapses to the $1/N$ floor and cannot distinguish or rank the observed bias levels. Colour shows c_K, but its units and calibration are target-specific, so colours across GMM, DW4, MeanFlow, and LJ13 must not be read as a universal bias scale. The figure supports the ESS-floor observation, not a cross-system claim that c_K universally tracks bias.

K Experimental details

Hardware and software. All experiments run on a single RTX 4080 (16 GB) with PyTorch. Per-step Jacobians and log-determinants are computed with PyTorch's functional autodifferentiation and detached each step to bound memory; exact log-determinants use a dense determinant routine.

Synthetic targets. Gaussian mixtures with components on a sphere (known density); a flow-matching velocity field is trained on samples with a time-embedded MLP. Reference budgets are experiment-specific: $K_{\text{ref}} = 512$ for low-dimensional certificate/calibration runs, 1024 for the one-gap baseline, 128 for high-dimensional $d = 64, 256$, and 96 for $d = 1024, 3072$. FashionMNIST uses 160, MeanFlow/FALCON use 256, iDEM certificate sweeps use 500, and the LJ13 downstream study uses 512. Metrics are averaged over multiple seeds and a fixed source batch shared across all budgets (so per-sample differences are meaningful). The under-trained ablation reduces training to a few hundred steps.

Higher-order solver. The second-order solver is the explicit trapezoidal (Heun) method; the per-step map and its exact Jacobian log-determinant are formed by automatic differentiation of the composite step.

Real energy-only samplers. We use publicly available iDEM (Akhound-Sadegh et al., 2024) models on GMM, DW4, and LJ13, and form a deterministic probability-flow likelihood from the trained score; the target log-density is the known energy (LJ13 is mean-free, so the change-of-variables is taken on the mean-free subspace). The MeanFlow average-velocity map is trained with the MeanFlow identity using a forward-mode directional derivative, with the standard adaptive loss weighting and

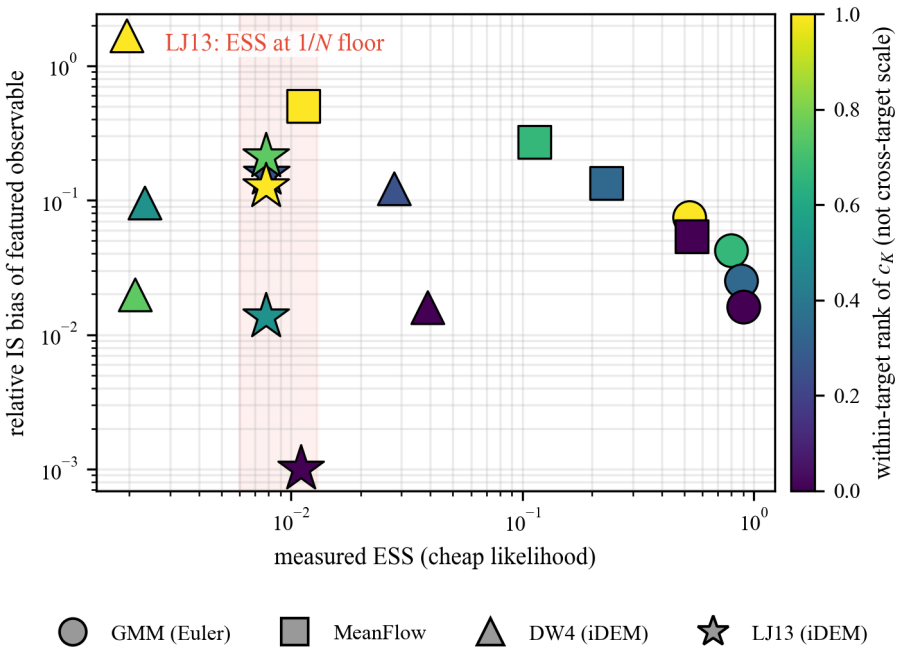


Figure 14: ESS at the LJ13 degeneracy floor cannot rank budgets. Each point is a (target, budget) setting: relative IS bias of the featured observable (vertical) versus measured self-normalized ESS (horizontal, log scale), with marker shape the target and colour the within-target rank of the same-sample diagnostic c_K. On DW4/GMM/MeanFlow, ESS co-varies with bias (moving right as bias falls); on LJ13 the ESS is pinned at the $1/N$ degeneracy floor across all budgets and biases (a vertical stripe), so it cannot distinguish or rank quality. Colour is the target-specific same-sample diagnostic c_K; because the targets and observables have different scales and dimensions, colour magnitudes are not comparable across marker types and do not establish universal bias tracking.

gradient clipping for stability; its few-step map is $x \mapsto x - (t_{\mathrm{hi}} - t_{\mathrm{lo}})\, u(x, t_{\mathrm{lo}}, t_{\mathrm{hi}})$ and converges first-order in $1/K$, so the certificate uses p=1.

Metrics. The stored coverage fields underlying the reported tolerant ranges were computed as $\mathrm{frac}(C + 0.05 \geq E_K)$ and are reported here only as 0.05-tolerant empirical coverage; they are not strict $\mathrm{frac}(C \geq E_K)$. Tightness is the median of C/E_K over draws with non-negligible error. Strict coverage is recomputed on the archived point arrays: the $d = 1024, K = 8$ GMM scatter gives 1.000 for exact C_{rich} and 0.975 for the plotted Hutchinson C, while the FashionMNIST $K = 32, M = 8$ scatter gives 1.000 for both C_{rich} and $C_{\mathrm{rich}} + 2\sigma$. No strict rates are inferred from aggregate coverage fields. The systematic-bias study replaces the exact log-determinant by its first-order (in h) truncation, an exactly-systematic surrogate, and reports the IS bias of a target observable against the exact-likelihood SNIS estimate as the reference.

L Extended related work

This appendix expands the compressed discussion of Sec. 2.

Convergence theory for samplers (a different object). A large body of work bounds the sampling-distribution error of diffusion and flow samplers in KL or Wasserstein, via stochastic control (Richter & Berner, 2024) or score-accuracy analyses (Chen et al., 2023). Our object is different and complementary: it measures the accuracy of the computed likelihood q_K and whether the resulting IS weights can be trusted. This differs from the distributional distance between the sampled distribution and the target; the variance-based reliability proxy does not expose the likelihood error. A sampler can be provably close in distribution while the number its likelihood code returns is still wrong enough to bias every reweighted estimate.

Importance sampling and ESS. ESS remains the field's default reliability readout (Owen, 2013; Martino et al., 2017). Pareto-smoothed importance sampling (Vehtari et al., 2024) adds a tail-index ($\hat{k}$) reliability statistic; being another function of the weight distribution, it is complementary to the variance-based ESS, and for a near-constant likelihood shift it shares that ESS's blindness. Our analysis separates the first-order terms through which a likelihood error enters ESS and a target observable, and turns the gap into a cheap same-sample detector for the regime where the observable moves while ESS stays flat. Richardson extrapolation itself is textbook numerical analysis (Hairer et al., 1993); the new elements here are its application to the few-step likelihood functional, the explicit bias bound tying it to IS validity (Lemma 4), the reference-free construction, and the validation on real samplers.

Diagnosing few-step sampler correctness, and its limits. A parallel line studies the reliability of fast samplers rather than sampling speed. Tang et al. (2026) show, for discrete diffusion language models, that few-step samplers can be distributionally incorrect even when surface metrics improve; their diagnostic is oracle-based and distributional, whereas ours is reference-free, per-sample, and targets the continuous-flow likelihood that enters the importance weight. In gravitational-wave inference, Dax et al. (2023) read a neural proposal's IS sample efficiency as the reliability diagnostic, precisely the variance proxy our certificate complements (Sec. 4.5). Finally, Xun & Price (2026) prove an $\widetilde{\Omega}(\sqrt{d})$ score-query lower bound for diffusion sampling; our certificate makes such limits actionable a-posteriori by detecting when a chosen budget falls below the requirement of the problem.