GRADIENT-FREE SAMPLING FROM GENERATIVE MODELS VIA STOCHASTIC BOUNDED EXTREMUM SEEKING

Anonymous authors
Paper under double-blind review

ABSTRACT

We introduce a sampling approach for energy-based and score-based generative models that requires no gradient evaluations of the model. Replacing the drift term that would normally contain the score $\nabla_{\mathbf{x}} \log p_\theta(\mathbf{x})$ with a high-frequency dithered cosine of the model's *value*, $\sqrt{\alpha\omega}\,\cos(\omega t + k \log p_\theta(\mathbf{x}))$, produces, in the high-frequency averaging limit, Langevin Markov chain Monte Carlo for energy-based models and the reverse-time SDE of score-based diffusion. We prove that trajectories of the dithered Itô SDE converge to those of the target SDE, driven by the same Brownian motion, uniformly on compact time intervals in probability, by an averaging argument that extends bounded extremum seeking (ES) to Itô processes, with an explicit $O(\omega^{-1/2})$ mean-square rate under global bounds. The algorithm does not require derivatives: only forward evaluations are used, one per time step regardless of dimension. The approach is not confined to smooth targets: it extends to $C^{1,1}$ energies with discontinuous curvature (without ellipticity requirement) and to Sobolev energies whose Hessians exist only off measure zero sets; for Lipschitz energies with gradient kinks the averaged limit remains well posed, and the convergence proof uses an ω-uniform occupation estimate for the dithered process, which we prove by a near-identity change of variables that removes the dither pathwise before Girsanov's theorem is applied; the Krylov–Röckner integrability class is the boundary of provability. The approach provides a hard *a priori* bound on the per-step update rate and applies to explicitly time-varying targets on finite horizons. Gradient-free pixel-space sampling is not competitive with well-tuned backpropagation-based samplers at practical evaluation budgets; the regime where the approach offers an advantage is latent-space sampling when the model is a black box and the target drifts in time. We demonstrate latent-space tracking for time-varying images on CelebA-HQ (256×256) from limited 1D projection measurements and latent-space EBM sampling on CIFAR-10.

1 INTRODUCTION

Modern generative models are trained, sampled, and conditioned primarily through gradients of large neural networks. Score-based diffusion (Song et al., 2021; Ho et al., 2020; Song & Ermon, 2019), latent diffusion (Rombach et al., 2022), flow matching (Lipman et al., 2023; Liu et al., 2023; Albergo & Vanden-Eijnden, 2023), and energy-based models (EBMs) (Du & Mordatch, 2019; Grathwohl et al., 2020; Du et al., 2021) all rely on gradients of a trained network: guidance or likelihood signals for diffusion and flow models, and the energy gradient itself for EBMs. When the model is available inside an automatic-differentiation framework this is efficient: a backward pass costs a small constant multiple of a forward pass. Our interest is in settings where even this is unavailable or insufficient.

Three such regimes recur in practice. *(i)*: The quantity of interest is computed by a simulator (Cranmer et al., 2020), a black-box scoring function, a quantized or compiled inference engine, or a model served behind an API, so that only function *values* can be obtained (Nesterov & Spokoiny, 2017). *(ii)*: The target distribution drifts continuously, which is the operating condition of feedback-based generative reconstruction in scientific instruments, such as evolving microstructures in dynamic coherent diffraction imaging (Scheinker & Pokharel, 2024), time-varying particle accelerator beam diagnostics (Scheinker et al., 2026), and more broadly of streaming inverse problems in which each new measurement redefines the posterior. Restarting an MCMC chain at every measurement is wasteful, and running a single gradient-based chain against a drifting target has no tracking guarantee.

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.

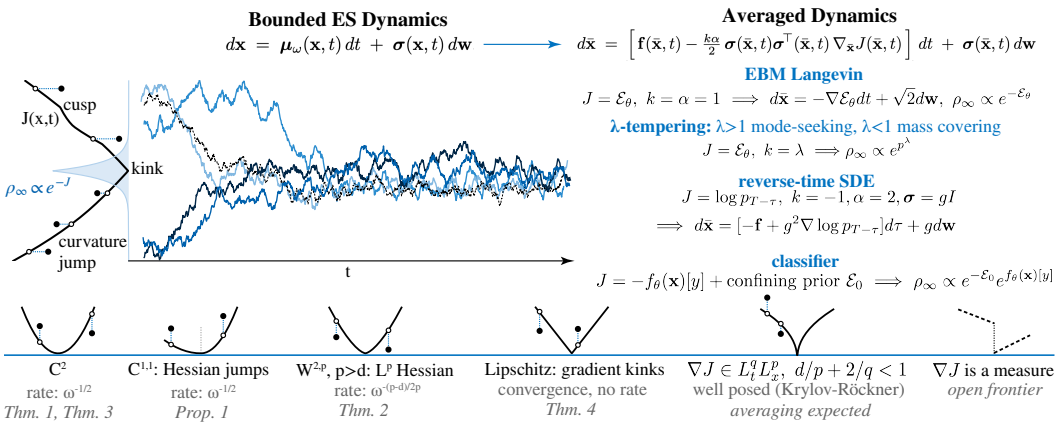


Figure 1: Bounded ES dynamics results in various samplers depending on the choice of $(\mathbf{f}, \boldsymbol{\sigma}, k, \alpha, J)$, where $\boldsymbol{\sigma}(\mathbf{x},t) = \operatorname{diag}\big(g_1(\mathbf{x},t),\dots,g_d(\mathbf{x},t)\big)$, $\mu_{\omega,i}(\mathbf{x},t) = f_i(\mathbf{x},t) + g_i(\mathbf{x},t)\sqrt{\alpha\omega_i}\cos\big(\omega_i t + k\,J(\mathbf{x},t)\big)$. Averaged dynamics not changed by rough energies, but the provable results are. Above C^2 means the a.e.-equivalence class of J contains a C^2 function. J can have measure zero jumps.

(iii): Sampling in the latent space of a pretrained encoder-decoder whose internals are frozen or otherwise outside the user's autodiff graph, as in generative-prior inverse problems (Bora et al., 2017) and latent-space molecular design (Gómez-Bombarelli et al., 2018).

This paper develops a single sampling procedure for all three regimes by replacing model gradients with a *high-frequency dithered cosine* of the model's value. Given a function $J(\mathbf{x},t)$ that can be evaluated but not differentiated by the user, consider the stochastic feedback law

$$dx_i \;=\; \sqrt{2}\,\sqrt{\alpha\omega_i}\,\cos\big(\omega_i t + k\,J(\mathbf{x},t)\big)\,dt \;+\; \sqrt{2}\,dw_i, \qquad i=1,\dots,d, \tag{1}$$

with distinct dither frequencies $\omega_i = \omega r_i$. As $\omega\to\infty$ the trajectories of Eq. (1) converge (in the precise sense of Theorem 1) to those of $d\bar{\mathbf{x}} = -k\alpha\,\nabla_{\bar{\mathbf{x}}}J(\bar{\mathbf{x}},t)\,dt + \sqrt{2}\,d\mathbf{w}$. With $J=\mathcal{E}_\theta$ and $k\alpha=1$ this is Langevin MCMC for the EBM $p\propto e^{-\mathcal{E}_\theta}$; embedded in the appropriate drift with $J=\log p_\theta$ it reproduces the reverse-time SDE of score-based diffusion (Anderson, 1982; Song et al., 2021). The gradient $\nabla_{\mathbf{x}}J$ is never computed, only evaluated once per time step, shared across all d coordinates. Flow-matching velocity networks are score networks up to an affine map (Lipman et al., 2023; Albergo & Vanden-Eijnden, 2023), so they are covered exactly as score-only diffusion models are in Sec. 3.3, (II): through the line-integral surrogate, not directly. The dithered feedback law is *bounded extremum seeking*, which originated in adaptive control of analytically unknown time-varying systems (Scheinker, 2013) and was generalized to discontinuous dithers in (Scheinker & Scheinker, 2016). For generative-model sampling we extend the averaging analysis to stochastic differential equations, which is the subject of Theorem 1 and its extensions, as summarized in Figure 1. The main contributions of this paper are:

Theory: We prove that multivariate stochastic bounded ES SDE converges to Langevin / reverse-time diffusion SDE, driven by the same Brownian motion, uniformly on compact time intervals in probability as the dither frequencies tend to infinity and prove convergence in distribution on path space. This is, to our knowledge, the first averaging analysis of bounded ES in the Itô setting. Under global bounds the same proof yields the quantitative mean-square rate $\mathbb{E}\sup_{t\le T}\|\mathbf{x}_\omega - \bar{\mathbf{x}}\|^2 \le C/\omega$, and we show that the same sampler converges on $C^{1,1}$ energies with discontinuous curvature and on Sobolev energies whose Hessian exists only off a null set; for Lipschitz energies with gradient kinks the averaged limit remains well posed. The boundary of provability is the Krylov-Röckner class.

Algorithm 1: The finite-step sampler is provided and guidance for choosing ω, $\{r_i\}$, Δt, k, and α.

Instantiations and tempering, as corollaries: Choices of $(\mathbf{f}, \boldsymbol{\sigma}, k, \alpha, J)$ recover gradient-free Langevin for EBMs, gradient-free reverse-SDE sampling for models with an explicit log-density, gradient-free sampling from a classifier via the EBM–classifier correspondence (Grathwohl et al., 2020), and a single-knob λ-tempering whose stationary law is p^λ.

Experiments mapping the operating regime: We evaluate against gradient-based (SGLD, MALA) and zeroth-order (Roy et al., 2022) baselines at matched evaluation budgets, on a synthetic testbed, on latent-space EBM sampling (CIFAR-10), and on a time-varying latent tracking problem (CelebA-HQ).

Scope of claims. We do not claim that gradient-free ES sampling outperforms well-tuned gradient-based Langevin samplers in pixel space on standard benchmarks; the need to resolve the dither makes the per-step / per-quality tradeoff unfavorable there (Sec. 3.2). The claim is a *regime expansion*: the same algorithm, with guarantees, operates in settings where backpropagation is unavailable.

2 Background

2.1 Langevin MCMC and score-based diffusion

Sampling from a target $p(\mathbf{x}) \propto \exp(-\mathcal{E}(\mathbf{x}))$ on $\mathbb{R}^d$ via overdamped Langevin dynamics (Roberts & Tweedie, 1996; Welling & Teh, 2011) simulates

$$d\bar{\mathbf{x}} \;=\; -\nabla_{\bar{\mathbf{x}}}\mathcal{E}(\bar{\mathbf{x}})\,dt \;+\; \sqrt{2}\,d\mathbf{w} \;=\; \nabla_{\bar{\mathbf{x}}}\log p(\bar{\mathbf{x}})\,dt + \sqrt{2}\,d\mathbf{w}, \tag{2}$$

whose unique stationary law is p. (The $\sqrt{2}$ matters: with unit diffusion the stationary law of $d\mathbf{x} = \nabla \log p\,dt + d\mathbf{w}$ is proportional to p^2, not p.) For modern EBMs $\mathcal{E}_\theta$ trained via contrastive-divergence-style objectives (Du & Mordatch, 2019; Du et al., 2021; Grathwohl et al., 2020; Nijkamp et al., 2020), $\nabla_{\mathbf{x}}\mathcal{E}_\theta(\mathbf{x})$ is obtained by automatic differentiation. Score-based generative models (Song & Ermon, 2019; Song et al., 2021; Ho et al., 2020) extend this to a continuum of noise levels; sampling integrates the reverse-time SDE (Anderson, 1982)

$$d\bar{\mathbf{x}} \;=\; \big[\,\mathbf{f}(\bar{\mathbf{x}},t) - g^2(t)\,\nabla_{\bar{\mathbf{x}}}\log p(\bar{\mathbf{x}},t)\,\big]\,dt + g(t)\,d\mathbf{w}, \tag{3}$$

where the score $\nabla_{\mathbf{x}}\log p(\mathbf{x},t)$ is approximated by a network $\mathbf{s}_\theta(\mathbf{x},t)$ trained by denoising score matching (Vincent, 2011; Song & Ermon, 2019). In Eq. (3) time runs backward from T to 0 ($dt<0$, $\mathbf{w}$ a reverse-time Brownian motion); in the forward computational time $\tau = T - t$ used by every integrator, including Algorithm 1 below, the same dynamics read $d\bar{\mathbf{x}} = [-\mathbf{f}(\bar{\mathbf{x}}, T-\tau) + g^2(T-\tau)\nabla_{\bar{\mathbf{x}}}\log p(\bar{\mathbf{x}}, T-\tau)]\,d\tau + g(T-\tau)\,d\mathbf{w}$, with the score drift *contracting* mass onto the data modes. Both Eq. (2) and Eq. (3) are ordinarily implemented by differentiating through θ.

2.2 Bounded extremum seeking

Bounded ES (Scheinker, 2013; Scheinker & Scheinker, 2016) is a model-independent feedback law that steers a parameter vector toward the (possibly time-varying) minimizer of an unknown cost $J(\mathbf{x},t)$ using only evaluations of J. Its defining feature is the dithered cosine update of Eq. (1), with two properties that distinguish it from generic finite-difference, simultaneous-perturbation, or smoothing-based gradient estimates (Spall, 1992; Nesterov & Spokoiny, 2017; Roy et al., 2022): (i) the update rate obeys the hard *a priori* bound $|dx_i/dt - \text{noise}| \le \sqrt{\alpha\omega_i}\,|g_i|$ *regardless of the magnitude of* J, and (ii) a single scalar evaluation of J per time instant drives all d coordinates simultaneously, rather than $d{+}1$ or 2 evaluations per estimated gradient. Existing averaging analyses of bounded ES concern deterministic ODEs; we extend the analysis to Itô SDEs.

3 Gradient-Free Sampling via Stochastic ES

3.1 Setup and main result

Throughout, $\mathbf{w}$ is a standard d-dimensional Brownian motion on a filtered probability space, and we write the dithered system in standard Itô drift–diffusion form. For $\mathbf{x} \in \mathbb{R}^d$ consider

$$d\mathbf{x} \;=\; \boldsymbol{\mu}_\omega(\mathbf{x},t)\,dt \;+\; \boldsymbol{\sigma}(\mathbf{x},t)\,d\mathbf{w}, \qquad \boldsymbol{\sigma}(\mathbf{x},t) \;=\; \mathrm{diag}\big(g_1(\mathbf{x},t),\dots,g_d(\mathbf{x},t)\big), \tag{4}$$

$$\mu_{\omega,i}(\mathbf{x},t) \;=\; f_i(\mathbf{x},t) \;+\; g_i(\mathbf{x},t)\,\sqrt{\alpha\omega_i}\,\cos\big(\omega_i t + k\,J(\mathbf{x},t)\big), \qquad \omega_i = \omega\,r_i, \tag{5}$$

with $\mathbf{f} = (f_1,\dots,f_d)$ the drift, $\boldsymbol{\sigma}$ a diagonal diffusion, k,α gains, and J a black-box target function.

Assumption 1 (Regularity)**.** $\mathbf{f}(\cdot,t)$ is continuous and locally Lipschitz in $\mathbf{x}$, uniformly for t in compacts; each g_i and J is C^2 in $\mathbf{x}$ and C^1 in t, with continuous mixed derivatives $\partial_t\nabla_{\mathbf{x}} g_i$ and $\partial_t\nabla_{\mathbf{x}} J$, and with $\mathbf{f}$, g_i, J, all first- and second-order $\mathbf{x}$-derivatives, the first-order t-derivatives, and the mixed derivatives of g_i, J bounded on $K\times[0,T]$ for every compact $K\subset\mathbb{R}^d$ and $T>0$.

Table 1: Parameter dictionary: choices of $(\mathbf{f}, \boldsymbol{\sigma}, k, \alpha, J)$ in Eq. (4)–Eq. (5) and the resulting averaged dynamics Eq. (6). $\mathcal{E}_\theta$ denotes an energy, p a density, $\lambda > 0$ a tempering exponent.

Instantiation	$\mathbf{f}$	$\boldsymbol{\sigma}$	k	α	**averaged SDE / stationary law**
(I) EBM Langevin	$\mathbf{0}$	$\sqrt{2}\,I$	1	1	$d\bar{\mathbf{x}} = -\nabla\mathcal{E}_\theta\,dt + \sqrt{2}\,d\mathbf{w}$; $\rho_\infty \propto e^{-\mathcal{E}_\theta}$
(I_λ) tempered	$\mathbf{0}$	$\sqrt{2}\,I$	λ	1	$d\bar{\mathbf{x}} = -\lambda\nabla\mathcal{E}_\theta\,dt + \sqrt{2}\,d\mathbf{w}$; $\rho_\infty \propto e^{-\lambda\mathcal{E}_\theta}$
(II) reverse SDE ($\tau{=}T{-}t$)	$-\mathbf{f}(\mathbf{x}, T{-}\tau)$	$g(T{-}\tau)I$	-1	2	$d\bar{\mathbf{x}} = [-\mathbf{f} + g^2\nabla \log p_{T-\tau}]d\tau + g\,d\mathbf{w}$
(III) classifier	$\mathbf{0}$	$\sqrt{2}\,I$	1	1	(I) with $\mathcal{E}_\theta(\mathbf{x}\,|\,y) = -f_\theta(\mathbf{x})[y]$

Assumption 2 (Dither frequencies)**.** $\omega_i = \omega r_i$ with $r_i \neq r_j\ \forall\, i \neq j$ $r_i > 0$, $i = 1, \ldots, d$.

Assumption 3 (Well-posedness)**.** $\mathbf{f}$, $\boldsymbol{\sigma}$, and $\boldsymbol{\sigma}\boldsymbol{\sigma}^\top\nabla_{\mathbf{x}} J$ satisfy a linear-growth bound $\|\cdot\| \leq C(1{+}\|\mathbf{x}\|)$ uniformly for $t \in [0, T]$, so that Eq. (4) and Eq. (6) below admit unique global strong solutions.

Theorem 1 (Stochastic ES averaging)**.** *Under Assumptions 1–3, let $\mathbf{x}_\omega$ solve Eq. (4) and let $\bar{\mathbf{x}}$ solve*

$$d\bar{\mathbf{x}} \;=\; \underbrace{\Big[\mathbf{f}(\bar{\mathbf{x}}, t) - \tfrac{k\alpha}{2}\,\boldsymbol{\sigma}(\bar{\mathbf{x}}, t)\boldsymbol{\sigma}^\top(\bar{\mathbf{x}}, t)\,\nabla_{\bar{\mathbf{x}}} J(\bar{\mathbf{x}}, t)\Big]}_{\bar{\boldsymbol{\mu}}(\bar{\mathbf{x}}, t)}\,dt \;+\; \boldsymbol{\sigma}(\bar{\mathbf{x}}, t)\,d\mathbf{w}, \tag{6}$$

driven by the same *Brownian motion and same initial condition $\mathbf{x}_\omega(0) = \bar{\mathbf{x}}(0) = \mathbf{x}_0$. For any $T > 0$,*

$$\sup_{t\in[0,T]} \big\|\mathbf{x}_\omega(t) - \bar{\mathbf{x}}(t)\big\| \;\longrightarrow\; 0 \quad \textit{in probability as } \omega \to \infty; \tag{7}$$

for every $\delta > 0$, $\varepsilon > 0$ $\exists\, \omega^* = \omega^*(\delta, \varepsilon, T)$ such that $\mathbb{P}\big(\sup_{t\leq T} \|\mathbf{x}_\omega(t) - \bar{\mathbf{x}}(t)\| > \delta\big) < \varepsilon$, $\forall \omega \geq \omega^*$.

The proof is in Appendix A with main steps: (i) localization by exit times from a large ball; (ii) Itô integration by parts; (iii) an oscillatory-integral lemma (Lemma 1) showing that integrals of the form $\int_0^t \Phi(\mathbf{x}_\omega(s), s)\, e^{\mathrm{i}c\omega s}\, ds$ with $c \neq 0$ vanish in probability; (iv) extraction of the resonant terms, which produce the drift correction $-\frac{k\alpha}{2}\boldsymbol{\sigma}\boldsymbol{\sigma}^\top\nabla J$; and (v) an L^2 Grönwall comparison with $\bar{\mathbf{x}}$. Intuitively, the state has a small ripple of amplitude $O(\omega^{-1/2})$ so that $x_i(t) \approx \bar{x}_i(t) + g_i\sqrt{\alpha/\omega_i}\,\sin(\omega_i t + kJ)$, where g_i is the coefficient of the dither ($g_i = \sqrt{2}$ in Eq. (1)). Plugging this ripple into the cosine, the dither "feels" the change in J that its own motion induces; the product of the dither and its induced phase shift has a nonvanishing time average, and the average is $-\frac{k\alpha}{2}\,g_i^2\,\partial J/\partial x_i$.

3.2 The finite-step algorithm and practical guidance

Algorithm 1 is the procedure used in all experiments. The methods for choosing ES gain k, amplitude α, and frequencies ω_i are studied in more detail in Appendix C and D. Any pairwise-distinct positive constants satisfy Assumption 2; we use r_i evenly spaced (or i.i.d. uniform) in $[1, N]$ with modest $N \in [1.75, 2]$. Euler-Maruyama must resolve the fastest dither: we use $\omega r_{\max}\,\Delta t \leq 2\pi/10$ so that at least 10 steps are required per fastest dither period. The product $k\alpha$ sets the strength of the emergent drift (Eq. 6) and the temperature of the averaged law ($\rho_\infty \propto e^{-k\alpha\mathcal{E}_\theta}$ in (I_λ)). Furthermore, if J is changing slowly, we see from Eq. 5 that components will oscillate with amplitude proportional to $\sqrt{\alpha/\omega_i}$, therefore in practice once ω_i is fixed, α is slowly increased until the dither is big enough to produce a good signal in the cost function and then k is slowly increased until a reasonable rate of convergence is achieved. Excessive gain $k\alpha$ at any finite ω can destabilize the ES dynamics.

3.3 Instantiations

(I) Gradient-free Langevin MCMC for EBMs. With $\mathbf{f} \equiv \mathbf{0}$, $g_i \equiv \sqrt{2}$, $k, \alpha = 1$, & $J = \mathcal{E}_\theta$,

$$dx_i \;=\; \sqrt{2\omega_i}\,\cos\big(\omega_i t + \mathcal{E}_\theta(\mathbf{x}, t)\big)\,dt \;+\; \sqrt{2}\,dw_i, \tag{8}$$

with average Langevin diffusion dynamics Eq. (2) with potential $\mathcal{E}_\theta$ and stationary law $\propto e^{-\mathcal{E}_\theta}$.

Algorithm 1 ES sampling (Euler–Maruyama form). One evaluation of J per step, for any d.

Require: evaluator $J(\cdot,\cdot)$; base drift $\mathbf{f}$, diffusion $\boldsymbol{\sigma} = \mathrm{diag}(g_i)$; gains k, α; base frequency ω; distinct ratios $\{r_i\}_{i=1}^{d} \subset [1, r_{\max}]$; step Δt; steps M; initial $\mathbf{x}^{(0)}$

1: **for** $m = 0, \dots, M-1$ **do**
2: $\quad t_m \leftarrow m\,\Delta t$; $\quad J_m \leftarrow J(\mathbf{x}^{(m)}, t_m)$ ▷ single forward evaluation
3: $\quad$ draw $\boldsymbol{\xi} \sim \mathcal{N}(0, I_d)$
4: $\quad$ **for** $i = 1, \dots, d$ **do** ▷ vectorized in practice
5: $\quad\quad x_i^{(m+1)} \leftarrow x_i^{(m)} + \big[f_i(\mathbf{x}^{(m)}, t_m) + g_i(\mathbf{x}^{(m)}, t_m)\sqrt{\alpha\omega r_i}\,\cos(\omega r_i t_m + k J_m)\big]\Delta t + g_i(\mathbf{x}^{(m)}, t_m)\sqrt{\Delta t}\,\xi_i$
6: $\quad$ **end for**
7: **end for**
8: **return** $\{\mathbf{x}^{(m)}\}$ (samples after burn-in, or the full trajectory for tracking)

($\mathbf{I}_\lambda$) Single-knob λ-tempering. Replacing the unit gain in Eq. (8) by $k = \lambda > 0$ yields averaged dynamics $d\bar{\mathbf{x}} = -\lambda \nabla \mathcal{E}_\theta\, dt + \sqrt{2}\, d\mathbf{w}$ with stationary law

$$\rho_\infty(\mathbf{x}) \;\propto\; e^{-\lambda \mathcal{E}_\theta(\mathbf{x})} \;\propto\; p^\lambda(\mathbf{x}) \tag{9}$$

(proof via the stationary Fokker–Planck equation, Appendix A.5). $\lambda = 1$ recovers the target; $\lambda > 1$ is mode-seeking; $\lambda < 1$ is mass-covering. Unlike parallel tempering (Earl & Deem, 2005) or annealed Langevin (Song & Ermon, 2019), no schedule, replica exchange, or noise-conditional training is required—the tempering exponent *is* the ES gain. Figure 7 illustrates the effect.

(II) Gradient-free reverse SDE for models with explicit log-density. Write the reverse-time SDE Eq. (3) in the forward computational time $\tau = T - t$ (Sec. 2), so that it is integrated with positive increments. With base drift $-\mathbf{f}(\mathbf{x}, T-\tau)$, $\boldsymbol{\sigma} = g(T-\tau)I$ from the forward diffusion, $k = -1$, $\alpha = 2$, and $J(\mathbf{x}, \tau) = \log p(\mathbf{x}, T-\tau)$, Eq. (4) becomes

$$dx_i \;=\; \big[\, -f_i(\mathbf{x}, T-\tau) + g(T-\tau)\sqrt{2\omega_i}\,\cos\big(\omega_i \tau - \log p(\mathbf{x}, T-\tau)\big)\,\big]\, d\tau + g(T-\tau)\, dw_i, \tag{10}$$

whose averaged dynamics, $d\bar{\mathbf{x}} = [-\mathbf{f} + g^2 \nabla \log p_{T-\tau}]\, d\tau + g\, d\mathbf{w}$, are exactly Eq. (3) run in forward time: the sign $k = -1$ (equivalently, $J = -\log p$ with $k = 1$) is what makes the emergent drift $+g^2 \nabla \log p$ contract onto the data modes when Algorithm 1 steps forward. This instantiation requires a *scalar* log-density evaluator: normalizing flows (Rezende & Mohamed, 2015), energy-parameterized diffusion models (Salimans & Ho, 2021), or any model exposing $\log p_t$ up to an additive constant (constants shift the cosine phase globally and drop out of the averaged drift). Score-only networks of the DDPM/EDM type do not expose $\log p_t$ and fall outside this instantiation, unless one forms the line-integral surrogate $\hat{J}(\mathbf{x}) = \int_0^1 \langle \mathbf{s}_\theta(\mathbf{x}_0 + u(\mathbf{x} - \mathbf{x}_0), t),\, \mathbf{x} - \mathbf{x}_0 \rangle\, du$, exact when $\mathbf{s}_\theta$ is conservative, at the cost of extra evaluations per step and a bias when it is not.

(III) Gradient-free sampling from a classifier. By the EBM-classifier correspondence (Grathwohl et al., 2020), a K-class classifier $f_\theta : \mathbb{R}^d \to \mathbb{R}^K$ defines class-conditional energies $\mathcal{E}_\theta(\mathbf{x} \mid y) = -f_\theta(\mathbf{x})[y]$. Substituting into Eq. (8) samples from any pretrained classifier using only its forward pass, which need not sit in an autodiff graph. Since a logit $f_\theta(\mathbf{x})[y]$ trained by cross-entropy is in general unbounded above and $e^{f_\theta(\mathbf{x})[y]}$ need not be integrable on $\mathbb{R}^d$, this instantiation presupposes a confining prior: either an intrinsically dissipative classifier (Grathwohl et al., 2020) or a non-zero base drift $\mathbf{f} = -\nabla \mathcal{E}_0$ implementing a prior density $\propto e^{-\mathcal{E}_0}$ (e.g., a Gaussian in latent space), so that the averaged dynamics target the well-posed density $\propto e^{-\mathcal{E}_0}\, e^{f_\theta(\mathbf{x})[y]}$.

3.4 Latent-space sampling

Applied in the latent space of a pretrained encoder–decoder (E, D), ES sampling becomes practical. Given an energy or log-density on latent codes $\mathbf{z} \in \mathbb{R}^{d_z}$, $d_z \ll d$, run Eq. (8) or Eq. (10) on $\mathbf{z}$ and decode $\hat{\mathbf{x}} = D(\mathbf{z})$. Three points: (i) d_z is small (3–512 in our experiments), so the frequency-gap constraint of Sec. 3.2 is mild and moderate ω suffices; (ii) the decoder is used strictly as a black box—relevant when D is frozen, exported, or quantized; (iii) latent diffusion (Rombach et al., 2022) and latent EBMs are the dominant deployment pattern, so this is the natural setting. The geometry of the latent energy landscape governs whether ES (or any local sampler) succeeds there.



Figure 2: Latent-space tracking from 1D RGB projections of time-varying images.

3.5 CelebA-HQ tracking statistics and derivative-free benchmark

For tracking time-varying CELEB-HQ images, the setup is shown in Figure 2. Time-varying targets are generated by spherical interpolation in latent space between pairs of CelebA-HQ images over 8192 steps, for 500 images each from the training and the held-out data, with 3 seeds per image. The tuned latent has shape $8{\times}8{\times}4$ ($d = 256$ parameters) and is decoded to an RGB image $\hat{I}[n] = D(z[n])$ of size $256{\times}256{\times}3$. The bounded-ES dynamics used in this experiment are the quadrature-pair variant

$$dx_i = \cos\big(\omega_i t + k\, J(\mathbf{x}, t)\big)\sqrt{\alpha\,\omega_i}\,dt + \sigma\, dW_i \quad (\text{even } i;\ \sin(\cdot) \text{ for odd } i), \tag{11}$$

with $x = \mathrm{vec}(z)$, integrated by Euler–Maruyama with step h:

$$x_i[n{+}1] = x_i[n] + \cos\big(\omega_i n h + k J[n]\big)\sqrt{\alpha\,\omega_i}\,h + \sigma\sqrt{h}\,\xi_i[n], \qquad \xi_i[n] \sim \mathcal{N}(0,1) \text{ i.i.d.}, \tag{12}$$

($\sin$ for odd i), with $k = 1$, $\alpha = 0.03$, $\sigma = 10^{-3}$, the ω_i linearly spaced over $[1.0, 3.5]$ with $8{\times}8{\times}2 = 128$ distinct values (each $\cos$ / $\sin$ pair shares one frequency), and $h = 2\pi/17.5$. With $\sigma = 10^{-3}$ this benchmark runs in the near-deterministic mode-seeking regime of Appendix D.6: it tests the tracking behavior of bounded ES as a derivative-free optimizer of a kinked objective, and the non-degenerate-noise theory of the Lipschitz rung (Proposition 3, Theorem 4), whose ellipticity constant would here be $\lambda_0 = 10^{-6}$, does not formally cover it. Hyperparameters were selected by a sweep over repeated optimization of a training image (3 seeds, first 2048 steps).

Feedback cost. $J[n]$ has four components; images are in $[0, 255]$ for the projection terms and in the decoder's native $[-1, 1]$ range (with standard VGG16 preprocessing) for the perceptual term. (i)–(ii) Projection terms: for an image I with pixels $I_{r,c,k}$ (r, c spatial, k channel), the projections are

$$P_{x/y}I(c,k) = \sum_{r/c=1}^{256} I_{r,c,k} \in \mathbb{R}^{256\times 3},\ c_{x/y}[n] = \tfrac{1}{255}\,\mathrm{MAE}\big(P_{x/y}(\hat{I}[n]),\ P_{x/y}(I^{\mathrm{tgt}}[n])\big), \tag{13}$$

with MAE the mean absolute difference over all entries. (iii) A perceptual term from an ImageNet-pretrained frozen VGG16: with ϕ_ℓ the feature maps of layers `block1_conv2`, `block2_conv2`, `block3_conv3`, `block4_conv3`, `block5_conv3`, $c_p[n] = \sum_{\ell=1}^{5} \mathrm{mean}\big[(\phi_\ell(\hat{I}[n]) - \phi_\ell(I^{\mathrm{tgt}}[n]))^2\big]$. (iv) A latent-norm regularizer $c_z[n] = \big(\|z[0]\|_2 - \|z[n]\|_2\big)^2$. The total is

$$J[n] = c_x[n] + c_y[n] + c_p[n]/10^8 + c_z[n], \tag{14}$$

the 10^{-8} weight compensating the scale of the raw summed VGG feature MSE. J drives ES; the performance metric is the NMSE computed directly between full images, as summarized in Table 2.

Table 2: Tracking study: 500 held-out pairs $\times$ 3 seeds, mean$\pm$SEM over the 1,500 runs.

NMSE	t=1170	2340	3510	4680	5850	7020	8191
Val, drift (vs. t=0)	.025$\pm$.0006	.078$\pm$.0022	.146$\pm$.0040	.221$\pm$.0061	.305$\pm$.0086	.400$\pm$.0119	.510$\pm$.0167
Val, bounded ES	.010$\pm$.0002	.019$\pm$.0004	.029$\pm$.0005	.040$\pm$.0007	.052$\pm$.0009	.067$\pm$.0013	.083$\pm$.0017

Table 3: Tracked vs. target image quality (left), and its evolution over the tracking horizon (right).

Metric	Validation (500 pairs)
PSNR (dB) ↑	21.08 ± 3.30
SSIM ↑	0.855 ± 0.095
FID ↓	8.68 ± 0.11
KID $\times 10^3$ ↓	5.23 ± 0.47

	t=0	1170	2340	3510	4680	5850	7020	8191
SSIM ↑	1.000	0.949	0.907	0.868	0.832	0.797	0.761	0.729
PSNR (dB) ↑	∞	26.50	23.62	21.84	20.48	19.36	18.33	17.41

FID and KID are computed per seed against the corresponding 4,000 target images (500 pairs $\times$ 8 checkpoints along the trajectory), mean$\pm$std across 3 seeds; PSNR/SSIM are paired and pooled over pairs$\times$seeds$\times$checkpoints (PSNR over the seven checkpoints $t > 0$, where it is finite; SSIM over all eight, where t=0 contributes a trivial value of 1). Summary in Table 3. We also compared bounded ES against, all at an identical budget of 8192 forward measurements per run: SPSA (Spall, 1992), Spall's two-measurement simultaneous perturbation with constant gains for tracking; GS-ES, the Gaussian-smoothing evolution-strategies estimator with antithetic sampling (Nesterov & Spokoiny, 2017; Salimans et al., 2017); central coordinate-wise finite differences (FD); and an automatic-differentiation oracle (exact gradients backpropagated through the trained decoder, with Adam). The benchmark covers 219 image pairs $\times$ 3 seeds per method (657 runs each). SPSA, GS-ES, and FD each received *two rounds* of grid search over their gains and estimator hyperparameters (a coarse round followed by a refined round around the best cell), and are reported at their best-tuned settings; bounded ES runs at the fixed setting of the tracking protocol above. Results summarized in Table 4.

Table 4: Final image-NMSE at t=8191. $\pm$ is the std over all 219 $\times$ 3 runs; median (IQR) across the 219 pairs after seed averaging (the per-pair distributions are right-skewed). "Updates" is the number of parameter updates each method uses, m is the tuned number of GS-ES antithetic pairs per update.

Method	Meas./update	Updates	Final NMSE (mean$\pm$std)	median (IQR)
Bounded ES (ours)	1	8192	$\mathbf{0.080 \pm 0.035}$	**0.072** (0.055–0.101)
SPSA	2	4096	0.142 ± 0.059	0.131 (0.104–0.181)
FD (central)	512	16	0.276 ± 0.312	0.219 (0.169–0.307)
GS-ES (antithetic)	$2m$	$8192/2m$	1.168 ± 0.627	1.028 (0.773–1.348)
AutoDiff oracle	—	—	0.048 ± 0.032	0.041 (0.024–0.063)

3.6 Latent representation and diffusion energy-based prior

Images $x \in [-1, 1]^{32\times32\times3}$ from CIFAR-10 (Krizhevsky, 2009) are represented in the latent space of a frozen convolutional variational autoencoder with self-attention (Kingma & Welling, 2014; Higgins et al., 2017), whose encoder $q_\phi(z \mid x) = \mathcal{N}\big(\mu_\phi(x), \operatorname{diag} \sigma^2_\phi(x)\big)$ and decoder D_θ act on a latent $z \in \mathbb{R}^{8\times8\times4}$ ($d = 256$, a 12$\times$ compression of the pixel space) and are trained with

$$\mathcal{L}_{\text{VAE}} = \mathbb{E}_{q_\phi(z|x)}\big[\|x - D_\theta(z)\|_2^2\big] + \beta\, \mathrm{KL}(q_\phi(z \mid x) \,\|\, \mathcal{N}(0, I))\,, \qquad \beta = 10^{-4}. \tag{15}$$

On the standardized latents (see Appendix E), we train a diffusion energy-based model, a latent diffusion model (Rombach et al., 2022; Ho et al., 2020) whose score is the gradient of an explicit energy Salimans & Ho (2021); Gao et al. (2021); Du et al. (2023). With forward process $z_t = \sqrt{\bar\alpha_t}\, z_0 + \sigma_t\, \epsilon$, $\sigma_t = \sqrt{1 - \bar\alpha_t}$, $\epsilon \sim \mathcal{N}(0, I)$, energy is parameterized through a denoising U-Net f_θ

$$E_\theta(z, t) = \|z - f_\theta(z, t)\|_2^2 / 2\,\sigma_t, \qquad \hat z_0(z, t) = f_\theta(z, t)/\sqrt{\bar\alpha_t}, \tag{16}$$

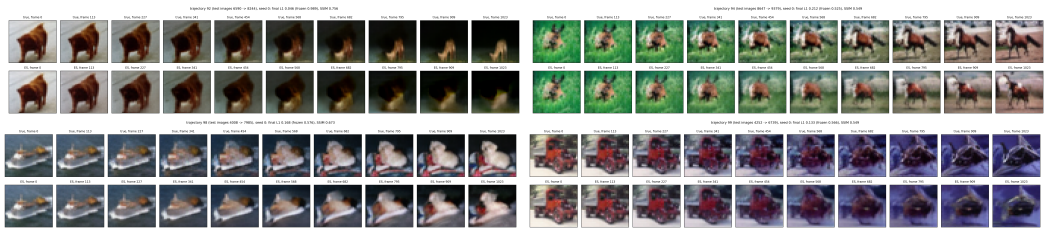


Figure 3: ES dynamics results for tracking time-varying CIFAR-10 images based on 1D projections.

Table 5: Final errors ($\mu \pm \sigma$) between tracked and true images, and FID between final tracked and final true images at the sample size N of the run, with the same-N floor (FID between two independent samples of the true distribution). L1 on $[-1, 1]$ pixel scale, RMSE/SSIM on $[0, 1]$ images.

Tracker	N_s	N	L1 final ↓	RMSE ↓	SSIM ↑	FID (floor) ↓
Frozen	–	5000	0.568 ±0.151	0.342 ±0.082	0.046 ±0.068	9.95 (9.95)
Bounded ES	200	5000	0.160 ±0.070	0.110 ±0.044	0.556 ±0.130	34.5 (9.95)

so $-\nabla_z E_\theta(z, t) \approx \sigma_t \nabla_z \log p_t(z)$ is a conservative score estimate and $\hat{z}_0$ the corresponding denoised estimate. The network is trained across noise levels with a denoising objective and its EMA weights are used at test time. Decoding 50,000 samples of the latent EBM through D_θ gives a Fréchet Inception distance Heusel et al. (2017) of 14.3 against the CIFAR-10 training set and 16.4 against the test set under the legacy TensorFlow protocol of the original FID implementation. For tracking, the prior is evaluated at a single low noise level $t_p = 0.05$ and standardised by its mean μ_p and standard deviation s_p over 2,000 held-out real latents,

$$J_p(z) = \sigma_{t_p} E_\theta(z, t_p) = \tfrac{1}{2} \|z - f_\theta(z, t_p)\|_2^2. \tag{17}$$

The unknown target is a moving image $x_n^* = D_\theta(z_n^*)$, $n = 0, \ldots, 1023$, whose latent follows a spherical interpolation between the latents of two random test images, a constant-speed path with $\|z_{n+1}^* - z_n^*\|_2 \approx 0.0125$ per frame. The tracker never sees x_n^*; it observes only four one-dimensional projections, $p_n = A\, x_n^*$, the sums along rows, columns, diagonals and anti-diagonals ($32 + 32 + 63 + 63 = 190$ bins per colour channel, 570 numbers per frame), the measurement model of few-view tomography. The data-fit cost of a candidate latent is the ℓ_1 projection mismatch, scaled by C_{scale}, the median mismatch between projections of unrelated real images. Tracking cost is

$$C(z; p_n) = \|A\, D_\theta(z) - p_n\|_1, \qquad J_n(z) = \frac{J_p(z) - \mu_p}{s_p} + \lambda \frac{C(z; p_n)}{C_{\text{scale}}}, \qquad \lambda = 100. \tag{18}$$

Bounded ES minimizes J_n with every latent coordinate $j = 1, \ldots, 256$ dithered at its own frequency $\omega_j \in [1,\ 1.75]$, $\Delta t = 2\pi/10\omega_{\max}$, $\alpha = 4 \times 10^{-3}$, and $k = 0.03$. Between frames a filtered velocity feed-forward carries the estimate along, $v \leftarrow v + (\Delta\bar{z} - v)/30$ and $z \leftarrow z + g\, v$ with $g = 1$ after the first 30 frames. Latents are clipped to $[-4, 4]$, the tracker is warm-started at z_0^*, and spends $N_s = 100$ cost evaluations on each frame. The results are summarized in Table 5 for tracking 5,000 trajectories of 1,024 frames. Errors are between the reported estimate $D_\theta(\bar{z})$ and the true image of the same frame: mean absolute error (L1) on the $[-1, 1]$ pixel scale, RMSE and SSIM Wang et al. (2004) on $[0, 1]$ images, and the Fréchet Inception distance (FID) Heusel et al. (2017) between the set of final tracked images and the set of true final images. FID uses Inception-v3 pool features Szegedy et al. (2016) with Keras ImageNet weights on bilinearly upsampled 299×299 images. Some tracking results are shown in Fig. 3 and many more in Fig. 5.

3.7 Beyond smooth energies: a regularity ladder

The C^2 hypothesis of Theorem 1 is a convenience of the proof, not a property the method needs. Algorithm 1 only ever evaluates $\cos(\omega_i t + kJ)$, which is well defined for any measurable J; what the *guarantee* needs is that the averaged drift's gradient ∇J be an admissible drift for the sampler's own noise. Here the noise is not an obstacle but the enabling mechanism: when the diffusion is uniformly

Table 6: Provability degrades gracefully as J roughens. Statements and proofs in Appendix B.

Regularity of $J(\cdot,t)$	**Permits**	**Drift** ∇J	**Status**
C^2 (Assumption 5)	classical smoothness	continuous	Thm 1; rate $\omega^{-1/2}$ (Thm 3)
$C^{1,1} = W^{2,\infty}$	bounded, non-C^1 Hessian	Lipschitz	Prop. 1 (proved)
$W^{2,p}_{\mathrm{uloc}}$, $p > d$, ∇J bounded	unbounded L^p Hessian	continuous	Thm 2 (proved)
Lipschitz, $W^{1,\infty}$	gradient kinks	a.e., $\in L^\infty$	Prop. 3; convergence, no rate
$\nabla J \in L^q_t L^p_{\mathbf{x}}$, $\frac{d}{p}+\frac{2}{q} < 1$	unbounded gradient; non-Lipschitz cusps	$\in L^p$	limit well posed (Krylov–Röckner); averaging expected
∇J distributional, $C^{-\beta}$	jumps in J	surface measure	open frontier

non-degenerate, the law of the process is absolutely continuous (Lemma 5), so the trajectory spends zero time on any Lebesgue-null set, and derivatives defined only *almost everywhere* are as good as continuous ones. Here the unmodified J is dithered directly, no proximal oracle or convexity is assumed. Throughout this subsection we adopt the global form of the regularity hypotheses on $\mathbf{f}$ and $\boldsymbol{\sigma}$ (Assumption 5, Appendix B), under which the localization in the proof of Theorem 1 is unnecessary and the argument sharpens to an explicit mean-square rate, $\mathbb{E}\sup_{t\le T}\|\mathbf{x}_\omega - \bar{\mathbf{x}}\|^2 \le C/\omega$ (Theorem 3), without the C^3 smoothness of Remark 1; the hypothesis on J is then relaxed rung by rung, with all statements and proofs in Appendix B and the full ladder summarized in Table 6.

Proposition 1 ($C^{1,1}$ energies)**.** *In Assumption 5, replace the hypothesis on J by: $J(\cdot,t) \in C^{1,1}(\mathbb{R}^d)$ uniformly in t, i.e., ∇J bounded and globally Lipschitz (so $\nabla^2 J$ exists a.e. and is bounded, but may be discontinuous; J itself may be unbounded). Then, with no non-degeneracy assumption on $\boldsymbol{\sigma}$, the conclusion of Theorem 3 holds unchanged, including the rate: $\mathbb{E}\sup_{t\le T}\|\mathbf{x}_\omega - \bar{\mathbf{x}}\|_2^2 \le C/\omega$.*

Theorem 2 (Sobolev energies)**.** *Assume the diffusion is uniformly non-degenerate, $\boldsymbol{\sigma}\boldsymbol{\sigma}^\top \succeq \lambda_0 I$ for some $\lambda_0 > 0$ (i.e., $\min_i \inf g_i^2 \ge \lambda_0$), with $\mathbf{f}, \boldsymbol{\sigma}$ bounded and Lipschitz, and that $J(\cdot,t) \in W^{2,p}_{\mathrm{uloc}}(\mathbb{R}^d)$ uniformly in t for some $p > \max(d,2)$ (so ∇J is continuous and $\nabla^2 J \in L^p_{\mathrm{uloc}}$), with ∇J and $\partial_t J$ bounded on $\mathbb{R}^d \times [0,T]$. Then $\mathbf{x}_\omega \to \bar{\mathbf{x}}$ in probability, uniformly on $[0,T]$, with*

$$\mathbb{E}\Big[\sup_{t\le T}\big\|\mathbf{x}_\omega(t) - \bar{\mathbf{x}}(t)\big\|_2^2\Big] \;\le\; C\,\omega^{-(p-d)/p},$$

where $\bar{\mathbf{x}}$ is the unique strong solution of the averaged SDE Eq. (6), whose drift is continuous because ∇J is (Morrey's embedding, $p > d$); the a.e.-defined Hessian $\nabla^2 J$ enters only the proof, along the trajectories, where the non-degenerate noise makes it unambiguous (Appendix B.3).

The proof is given in Appendix B.4, it utilizes the machinery of Theorem 3 on a mollified corrector. Theorem 2 covers energies that are C^2 off a null set as long as the gradient stays continuous: piecewise-quadratic energies glued in C^1 and, more generally, any energy with a merely L^p Hessian. When the gradient kinks, $\|\cdot\|_1$ or ReLU-type energies, or the absolute-value errors that drive tracking energy of Eq. (14), J leaves $W^{2,p}$: $\nabla^2 J$ acquires a singular surface part, and neither result above applies. The averaged limit remains well posed by Veretennikov's theorem for bounded measurable drifts of non-degenerate diffusions (Proposition 3). Appendix B.5 gives the two-scale scheme for this, and Appendix B.6 proves that bound (Theorem 4). Convergence holds without a rate whenever J is Lipschitz in time as well as in space; no rate is claimed, and none is expected from the argument, since $\|\nabla J - \nabla J_\varepsilon\|$ carries no rate for merely Lipschitz J.

How rough can J be? By Krylov-Röckner (Krylov & Röckner, 2005), the averaged SDE Eq. (6) is strongly well posed when the weak gradient satisfies the Ladyzhenskaya-Prodi-Serrin-type condition

$$\nabla J \;\in\; L^q\big([0,T];\, L^p(\mathbb{R}^d)\big), \qquad d/p + 2/q \;<\; 1, \tag{19}$$

which permits ∇J to be unbounded, defined only almost everywhere, and merely p-integrable in space with q-integrability in time; this is the most general regularity under which the limit is provably well posed by classical means, and we expect the averaging to extend throughout it (Remark 7). Requiring $J(\cdot,t)$ to be merely *locally Lipschitz*, the weakest classical condition under which ∇J exists a.e. at all, by Rademacher's theorem (Evans & Gariepy, 1992), still forces ∇J to be locally bounded, hence locally in L^p for every $p \le \infty$, so in local regularity Eq. (19) is strictly more general than any Lipschitz-type hypothesis: it additionally admits gradients with genuine local blow-ups of merely L^p type (cusps $|\nabla J| \asymp |\mathbf{x} - \mathbf{x}_0|^{-\gamma}$, $0 < \gamma < d/p$), which no Lipschitz condition tolerates.

AI USE STATEMENT

In this work, generative AI tools were used to proof-read the paper, to find typos or mathematical errors, for polishing the writing, and for discussion, suggestions, and collaboration on some of the mathematical results. We have reviewed all AI-assisted work. We checked LLM-suggested references through a manual literature survey. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

For all of the theoretical results, clear explanations of any assumptions and complete proofs of the claims are included in the appendix. For the proposed algorithm a clear explanation of exactly how it is applied has been provided. A complete description of data processing steps has also been provided.

REFERENCES

Michael S. Albergo and Eric Vanden-Eijnden. Building normalizing flows with stochastic interpolants. In *International Conference on Learning Representations (ICLR)*, 2023.

Brian D. O. Anderson. Reverse-time diffusion equation models. *Stochastic Processes and their Applications*, 12(3):313–326, 1982.

Dirk V. Arnold and Hans-Georg Beyer. Local performance of the (1+1)-ES in a noisy environment. *IEEE Transactions on Evolutionary Computation*, 6(1):30–41, 2002.

Patrick Billingsley. *Convergence of Probability Measures*. Wiley, 2nd edition, 1999.

Nikolai N. Bogoliubov and Yurii A. Mitropolsky. *Asymptotic Methods in the Theory of Non-Linear Oscillations*. Gordon and Breach, New York, 1961.

Ashish Bora, Ajil Jalal, Eric Price, and Alexandros G. Dimakis. Compressed sensing using generative models. In *International Conference on Machine Learning (ICML)*, 2017.

Frank H. Clarke. *Optimization and Nonsmooth Analysis*. Wiley, 1983.

Kyle Cranmer, Johann Brehmer, and Gilles Louppe. The frontier of simulation-based inference. *Proceedings of the National Academy of Sciences*, 117(48):30055–30062, 2020.

Yilun Du and Igor Mordatch. Implicit generation and modeling with energy based models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

Yilun Du, Shuang Li, Joshua Tenenbaum, and Igor Mordatch. Improved contrastive divergence training of energy-based models. In *International Conference on Machine Learning (ICML)*, 2021.

Yilun Du, Conor Durkan, Robin Strudel, Joshua B. Tenenbaum, Sander Dieleman, Rob Fergus, Jascha Sohl-Dickstein, Arnaud Doucet, and Will Sussman Grathwohl. Reduce, reuse, recycle: Compositional generation with energy-based diffusion models and MCMC. In *Proceedings of the 40th International Conference on Machine Learning (ICML)*, volume 202 of *Proceedings of Machine Learning Research*, pp. 8489–8510, 2023.

John C. Duchi, Michael I. Jordan, Martin J. Wainwright, and Andre Wibisono. Optimal rates for zero-order convex optimization: The power of two function evaluations. *IEEE Transactions on Information Theory*, 61(5):2788–2806, 2015.

David J. Earl and Michael W. Deem. Parallel tempering: Theory, applications, and new perspectives. *Physical Chemistry Chemical Physics*, 7(23):3910–3916, 2005.

Lawrence C. Evans and Ronald F. Gariepy. *Measure Theory and Fine Properties of Functions*. CRC Press, 1992.

Aleksei F. Filippov. *Differential Equations with Discontinuous Righthand Sides*. Kluwer Academic Publishers, 1988.

Franco Flandoli, Massimiliano Gubinelli, and Enrico Priola. Well-posedness of the transport equation by stochastic perturbation. *Inventiones Mathematicae*, 180(1):1–53, 2010.

Mark I. Freidlin and Alexander D. Wentzell. *Random Perturbations of Dynamical Systems*. Springer, 3rd edition, 2012.

Avner Friedman. *Partial Differential Equations of Parabolic Type*. Prentice-Hall, 1964.

Ruiqi Gao, Yang Song, Ben Poole, Ying Nian Wu, and Diederik P. Kingma. Learning energy-based models by diffusion recovery likelihood. In *International Conference on Learning Representations (ICLR)*, 2021.

Saeed Ghadimi and Guanghui Lan. Stochastic first- and zeroth-order methods for nonconvex stochastic programming. *SIAM Journal on Optimization*, 23(4):2341–2368, 2013.

Rafael Gómez-Bombarelli, Jennifer N. Wei, David Duvenaud, José Miguel Hernández-Lobato, Benjamín Sánchez-Lengeling, Dennis Sheberla, Jorge Aguilera-Iparraguirre, Timothy D. Hirzel, Ryan P. Adams, and Alán Aspuru-Guzik. Automatic chemical design using a data-driven continuous representation of molecules. *ACS Central Science*, 4(2):268–276, 2018.

Will Grathwohl, Kuan-Chieh Wang, Jörn-Henrik Jacobsen, David Duvenaud, Mohammad Norouzi, and Kevin Swersky. Your classifier is secretly an energy based model and you should treat it like one. In *International Conference on Learning Representations (ICLR)*, 2020.

Nikolaus Hansen, Anne Auger, Raymond Ros, Olaf Mersmann, Tea Tušar, and Dimo Brockhoff. COCO: A platform for comparing continuous optimizers in a black-box setting. *Optimization Methods and Software*, 36(1):114–144, 2021.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. GANs trained by a two time-scale update rule converge to a local Nash equilibrium. In *Advances in Neural Information Processing Systems*, volume 30, pp. 6626–6637, 2017.

Irina Higgins, Loic Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. beta-VAE: Learning basic visual concepts with a constrained variational framework. In *International Conference on Learning Representations (ICLR)*, 2017.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, volume 33, pp. 6840–6851, 2020.

Pavel Izmailov, Dmitrii Podoprikhin, Timur Garipov, Dmitry Vetrov, and Andrew Gordon Wilson. Averaging weights leads to wider optima and better generalization. In *Conference on Uncertainty in Artificial Intelligence (UAI)*, 2018.

Kevin G. Jamieson, Robert D. Nowak, and Benjamin Recht. Query complexity of derivative-free optimization. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2012.

Ioannis Karatzas and Steven E. Shreve. *Brownian Motion and Stochastic Calculus*. Springer, 2nd edition, 1991.

Jack Kiefer and Jacob Wolfowitz. Stochastic estimation of the maximum of a regression function. *The Annals of Mathematical Statistics*, 23(3):462–466, 1952.

Diederik P. Kingma and Max Welling. Auto-encoding variational Bayes. In *International Conference on Learning Representations (ICLR)*, 2014.

Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009.

Nicolai V. Krylov. *Controlled Diffusion Processes*. Springer, 1980.

Nicolai V. Krylov. On stochastic equations with drift in L_d. *The Annals of Probability*, 49(5): 2371–2398, 2021.

Nicolai V. Krylov and Michael Röckner. Strong solutions of stochastic equations with singular time dependent drift. *Probability Theory and Related Fields*, 131(2):154–196, 2005.

Harold J. Kushner. *Approximation and Weak Convergence Methods for Random Processes, with Applications to Stochastic Systems Theory*. MIT Press, 1984.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. In *International Conference on Learning Representations (ICLR)*, 2023.

Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. In *International Conference on Learning Representations (ICLR)*, 2023.

Yurii Nesterov and Vladimir Spokoiny. Random gradient-free minimization of convex functions. *Foundations of Computational Mathematics*, 17(2):527–566, 2017.

Erik Nijkamp, Mitch Hill, Tian Han, Song-Chun Zhu, and Ying Nian Wu. On the anatomy of MCMC-based maximum likelihood learning of energy-based models. In *AAAI Conference on Artificial Intelligence*, 2020.

Grigorios A. Pavliotis and Andrew M. Stuart. *Multiscale Methods: Averaging and Homogenization*. Springer, 2008.

Boris T. Polyak and Anatoli B. Juditsky. Acceleration of stochastic approximation by averaging. *SIAM Journal on Control and Optimization*, 30(4):838–855, 1992.

Ingo Rechenberg. *Evolutionsstrategie: Optimierung technischer Systeme nach Prinzipien der biologischen Evolution*. Frommann–Holzboog, Stuttgart, 1973.

Danilo Jimenez Rezende and Shakir Mohamed. Variational inference with normalizing flows. In *International Conference on Machine Learning (ICML)*, 2015.

Gareth O. Roberts and Richard L. Tweedie. Exponential convergence of Langevin distributions and their discrete approximations. *Bernoulli*, 2(4):341–363, 1996.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10684–10695, 2022.

Abhishek Roy, Lingqing Shen, Krishnakumar Balasubramanian, and Saeed Ghadimi. Stochastic zeroth-order discretizations of Langevin diffusions for Bayesian inference. *Bernoulli*, 28(3): 1810–1834, 2022.

David Ruppert. Efficient estimations from a slowly convergent Robbins–Monro process. Technical Report 781, School of Operations Research and Industrial Engineering, Cornell University, 1988.

Tim Salimans and Jonathan Ho. Should EBMs model the energy or the score? In *ICLR 2021 Workshop on Energy-Based Models*, 2021.

Tim Salimans, Jonathan Ho, Xi Chen, Szymon Sidor, and Ilya Sutskever. Evolution strategies as a scalable alternative to reinforcement learning. *arXiv preprint arXiv:1703.03864*, 2017.

Jan A. Sanders, Ferdinand Verhulst, and James Murdock. *Averaging Methods in Nonlinear Dynamical Systems*, volume 59 of *Applied Mathematical Sciences*. Springer, 2nd edition, 2007.

Alexander Scheinker. Model independent beam tuning. In *International Particle Accelerator Conference (IPAC'13)*, pp. 1862–1864, 2013.

Alexander Scheinker and Reeju Pokharel. Enabling dynamic 3D coherent diffraction imaging via adaptive latent space tuning of generative autoencoders. *npj Computational Materials*, 10(1):283, 2024.

Alexander Scheinker and David Scheinker. Bounded extremum seeking with discontinuous dithers. *Automatica*, 69:250–257, 2016.

Alexander Scheinker, Alexander Plastun, and Peter Ostroumov. PhaseFlow4D: Physically constrained 4D beam reconstruction via feedback-guided latent diffusion. *arXiv preprint arXiv:2604.03885*, 2026.

John H. Scofield. Frequency-domain description of a lock-in amplifier. *American Journal of Physics*, 62(2):129–133, 1994.

Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

Yang Song, Jascha Sohl-Dickstein, Diederik P. Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations (ICLR)*, 2021.

James C. Spall. Multivariate stochastic approximation using a simultaneous perturbation gradient approximation. *IEEE Transactions on Automatic Control*, 37(3):332–341, 1992.

Daniel W. Stroock and S. R. Srinivasa Varadhan. *Multidimensional Diffusion Processes*, volume 233 of *Grundlehren der mathematischen Wissenschaften*. Springer, Berlin, 1979.

Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the Inception architecture for computer vision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2818–2826, 2016.

Alexander Yu. Veretennikov. On strong solutions and explicit formulas for solutions of stochastic integral equations. *Mathematics of the USSR-Sbornik*, 39(3):387–403, 1981.

Cédric Villani. *Optimal Transport: Old and New*. Springer, 2009.

Pascal Vincent. A connection between score matching and denoising autoencoders. *Neural Computation*, 23(7):1661–1674, 2011.

Zhou Wang, Alan C. Bovik, Hamid R. Sheikh, and Eero P. Simoncelli. Image quality assessment: From error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4): 600–612, 2004.

Max Welling and Yee Whye Teh. Bayesian learning via stochastic gradient Langevin dynamics. In *International Conference on Machine Learning (ICML)*, 2011.

Alexander K. Zvonkin. A transformation of the phase space of a diffusion process that removes the drift. *Mathematics of the USSR-Sbornik*, 22(1):129–149, 1974.

The appendix includes all detailed proofs and a large amount of numerical experiments testing various aspects of this approach such as dithering frequency spread versus state dimension, as well as additional details about the experiments which could not fit in the main text.

A Proof of Theorem 1 and related results

A.1 Notation and standing assumptions

For $\mathbf{x} \in \mathbb{R}^N$ (we write N for the state dimension; $N = d$ in the notation of the main text) consider the coupled nonlinear system of Itô stochastic differential equations

$$d\mathbf{x} \;=\; \boldsymbol{\mu}_\omega(\mathbf{x},t)\,dt \;+\; \boldsymbol{\sigma}(\mathbf{x},t)\,d\mathbf{w}, \qquad \boldsymbol{\sigma}(\mathbf{x},t) \;=\; \operatorname{diag}\big(g_1(\mathbf{x},t),\dots,g_N(\mathbf{x},t)\big), \tag{20}$$

on a filtered probability space $(\Omega,\mathcal{F},(\mathcal{F}_t)_{t\ge0},\mathbb{P})$ carrying a standard N-dimensional Brownian motion $\mathbf{w}$, with drift components

$$\mu_{\omega,n}(\mathbf{x},t) \;=\; f_n(\mathbf{x},t) \;+\; g_n(\mathbf{x},t)\,\sqrt{\alpha\omega_n}\,\cos\big(\omega_n t + k\,J(\mathbf{x},t)\big), \qquad \omega_n = \omega r_n, \tag{21}$$

and the associated averaged system

$$d\bar{\mathbf{x}} \;=\; \bar{\boldsymbol{\mu}}(\bar{\mathbf{x}},t)\,dt + \boldsymbol{\sigma}(\bar{\mathbf{x}},t)\,d\mathbf{w}, \qquad \bar{\boldsymbol{\mu}} \;=\; \mathbf{f} - \tfrac{k\alpha}{2}\,\boldsymbol{\sigma}\boldsymbol{\sigma}^{\top}\nabla_{\mathbf{x}}J, \tag{22}$$

driven by the *same* $\mathbf{w}$ and started from the same $\mathcal{F}_0$-measurable initial condition $\mathbf{x}_\omega(0) = \bar{\mathbf{x}}(0) = \mathbf{x}_0$ with $\mathbb{E}\|\mathbf{x}_0\|^2 < \infty$, under Assumptions 1–3 of the main text. Under Assumption 3 both Eq. (20) (for each fixed ω) and Eq. (22) have unique global strong solutions with $\mathbb{E}\sup_{t\le T}\|\cdot\|^2 < \infty$. Because the dither term is bounded by $\sqrt{\alpha\omega r_{\max}}\,|g_n|$, $\boldsymbol{\mu}_\omega$ satisfies a linear-growth bound with an ω-dependent constant, which guarantees global existence at each fixed ω (Karatzas & Shreve, 1991, Ch. 5).

Throughout the analysis that follows we fix $T > 0$, C will be used to denote a constant that may change from line to line, and C_R denotes a constant depending only on $(R, T, N, \alpha, k, r_1, \dots, r_N)$ and on the bounds from Assumption 1 of $\mathbf{f}$, g_n, J and their derivatives on $\bar{B}_R \times [0,T]$, where $\bar{B}_R = \{\mathbf{x} : \|\mathbf{x}\| \le R\}$.

A.2 Warm-up: the deterministic one-dimensional mechanism

The following is meant to give intuition about how this algorithm is able to extract gradients, without yet presenting any stochastic technicalities. We consider $N = 1$, $g \equiv 1$, $\mathbf{f} \equiv 0$, and the dynamics $\dot{x} = \sqrt{\alpha\omega}\cos\theta$ with phase $\theta(t) = \omega t + kJ(x(t),t)$, where $J \in C^2$ has bounded derivatives. Consider the near-identity variable

$$y = x - \sqrt{\alpha/\omega}\,\sin\theta, \tag{23}$$

which differs from x by at most $\sqrt{\alpha/\omega}$. Integrating $\dot{x} = \sqrt{\alpha\omega}\cos\theta$ at leading order ($\theta \approx \omega t$) gives $x \approx \text{const} + \sqrt{\alpha/\omega}\sin\theta$, so the subtraction in Eq. (23) removes the fast oscillation. Differentiating Eq. (23) and plugging in $\dot{\theta} = \omega + k(\partial_x J\,\dot{x} + \partial_t J)$, we get

$$\begin{aligned}
\dot{y} &= \dot{x} - \sqrt{\alpha/\omega}\,\cos\theta\,\dot{\theta} \\
&= \dot{x} - \sqrt{\alpha/\omega}\cos\theta\,\big[\omega + k(\partial_x J\,\dot{x} + \partial_t J)\big] \\
&= \underbrace{\dot{x} - \sqrt{\alpha\omega}\cos\theta}_{=0} \;-\; \sqrt{\alpha/\omega}\,k\,\partial_x J\,\big(\sqrt{\alpha\omega}\cos\theta\big)\cos\theta \;-\; \sqrt{\alpha/\omega}\,k\,\partial_t J\cos\theta \\
&= -\underbrace{\sqrt{\alpha/\omega}\,\sqrt{\alpha\omega}}_{=\alpha}\,k\,\partial_x J\cos^2\theta \;-\; \sqrt{\alpha/\omega}\,k\,\partial_t J\cos\theta \\
&= -\tfrac{k\alpha}{2}\,\partial_x J\,\big(1+\cos 2\theta\big) \;-\; \sqrt{\alpha/\omega}\,k\,\partial_t J\cos\theta \\
&= -\tfrac{k\alpha}{2}\,\partial_x J(x,t) \;-\; \tfrac{k\alpha}{2}\,\partial_x J(x,t)\cos 2\theta \;-\; \sqrt{\alpha/\omega}\,k\,\partial_t J(x,t)\cos\theta \\
&= -\tfrac{k\alpha}{2}\,\partial_x J(x,t) \;-\; \tfrac{k\alpha}{2}\,\partial_x J(x,t)\cos 2\theta \;+\; O(\omega^{-1/2}).
\end{aligned} \tag{24}$$

The final line follows because, for fixed k and α,

$$\big|\sqrt{\alpha/\omega}\,k\,\partial_t J(x,t)\cos\theta\big| \;\le\; k\sqrt{\alpha}\,\sup_{x,t}|\partial_t J|\;\omega^{-1/2} \;=\; O(\omega^{-1/2}) \qquad \text{as } \omega\to\infty. \tag{25}$$

The first term of Eq. (24) is the desired gradient flow. The $\cos 2\theta$ term oscillates at frequency 2ω with slowly varying coefficient; one further integration by parts in time shows its running integral is $O(\omega^{-1/2})$ on $[0,T]$ when $\partial_x J$ is Lipschitz along the trajectory (the boundary term is $O(\omega^{-1})$, but the coefficient $\partial_x J(x(t),t)$ varies at the dither speed $\dot{x} = O(\sqrt{\omega})$, which costs $\sqrt{\omega}/\omega$). Hence y, and therefore $x = y + O(\omega^{-1/2})$, converges uniformly on $[0,T]$ to the solution of $\dot{\bar{x}} = -\frac{k\alpha}{2}\partial_x J(\bar{x},t)$ as $\omega \to \infty$. The mechanism is as described in the main text: the state carries an $O(\omega^{-1/2})$ ripple phase-locked to the dither; re-inserting the ripple into the cosine produces a product whose time average is exactly $-\frac{k\alpha}{2}\,\partial_x J$. The remainder of this appendix carries this argument out for the coupled Itô system Eq. (20), where three complications arise: the ripple interacts with the Brownian forcing, the coordinates are coupled through the shared phase kJ, and the oscillatory remainders must be controlled *along the ω-dependent random trajectory*.

A.3 An oscillatory-integral lemma along ES trajectories

We start with a helper Lemma that will be used in the proof of Theorem 1.

For $R > 0$ let

$$\begin{aligned} \tilde{\tau}_R &= \inf\{t \geq 0 : \ \|\mathbf{x}_\omega(t)\| \geq R \text{ or } \|\bar{\mathbf{x}}(t)\| \geq R\} \in [0,\infty], \\ \tau_R &= \tilde{\tau}_R \wedge T, \quad \text{where} \quad a \wedge b := \min\{a,b\}. \end{aligned} \tag{26}$$

Since the paths are continuous, $\|\mathbf{x}_\omega(t)\| \leq R$ and $\|\bar{\mathbf{x}}(t)\| \leq R$ for all $t \leq \tau_R$, and $\|\mathbf{x}_\omega(\tilde{\tau}_R)\| = R$ or $\|\bar{\mathbf{x}}(\tilde{\tau}_R)\| = R$ whenever $\tilde{\tau}_R < \infty$.

Lemma 1 (Oscillatory integrals along the dithered trajectory)**.** *Let Assumptions 1–3 hold, fix $R, T > 0$, and let $\Phi : \mathbb{R}^N \times [0,T] \to \mathbb{R}$ be bounded by C_Φ and jointly Lipschitz with constant L_Φ on $\bar{B}_R \times [0,T]$. Then for every real $c \neq 0$ and all $\omega \geq 1$,*

$$\mathbb{E}\left[\sup_{t \leq T}\left|\int_0^{t\wedge\tau_R} \Phi\big(\mathbf{x}_\omega(s), s\big)\, e^{\mathrm{i}c\omega s}\, ds\right|^2\right] \ \leq \ C(R,T,N,\Phi,|c|,\alpha,r_{\max})\,\omega^{-1/2}. \tag{27}$$

The same bound holds with $e^{\mathrm{i}c\omega s}$ replaced by $\cos(c\omega s)$ or $\sin(c\omega s)$.

Proof. Write $\mathbf{x} = \mathbf{x}_\omega$ and abbreviate $t \wedge \tau_R$ by t^τ. Partition $[0,T]$ into $M = \lceil T/\Delta \rceil$ blocks $[t_j, t_{j+1}]$ of length $\leq \Delta \in (0,1]$, $t_j = j\Delta$, and write

$$\begin{aligned} \int_0^{t^\tau} \Phi(\mathbf{x}_s, s) e^{\mathrm{i}c\omega s} ds \ &= \underbrace{\sum_j \Phi(\mathbf{x}_{t_j^\tau}, t_j) \int_{t_j \wedge t^\tau}^{t_{j+1}\wedge t^\tau} e^{\mathrm{i}c\omega s} ds}_{A(t)} \\ &\quad + \underbrace{\sum_j \int_{t_j \wedge t^\tau}^{t_{j+1}\wedge t^\tau} \big[\Phi(\mathbf{x}_s, s) - \Phi(\mathbf{x}_{t_j^\tau}, t_j)\big] e^{\mathrm{i}c\omega s} ds}_{B(t)}. \end{aligned} \tag{28}$$

In $A(t)$, each trigonometric integral over a (possibly truncated) block is at most $2/(|c|\omega)$ in modulus, and there are at most $M \leq T/\Delta + 1$ blocks, so path-wise $\sup_{t\leq T}|A(t)| \leq 2C_\Phi(T/\Delta + 1)/(|c|\omega)$.

For $B(t)$, for s in block j with $s \leq \tau_R$,

$$\big|\Phi(\mathbf{x}_s, s) - \Phi(\mathbf{x}_{t_j}, t_j)\big| \ \leq \ L_\Phi\big(\|\mathbf{x}_s - \mathbf{x}_{t_j}\| + \Delta\big), \tag{29}$$

and from Eq. (20), $\mathbf{x}_s - \mathbf{x}_{t_j} = \int_{t_j}^s \boldsymbol{\mu}_\omega\, du + \mathbf{N}_j(s)$, $\mathbf{N}_j(s) := \int_{t_j}^s \boldsymbol{\sigma}\, d\mathbf{w}$, so that on $\{s \leq \tau_R\}$,

$$\|\mathbf{x}_s - \mathbf{x}_{t_j}\| \ \leq \ C_R\big(1 + \sqrt{\alpha\omega r_{\max}}\big)\Delta \ + \ \|\mathbf{N}_j(s)\|, \qquad \mathbb{E}\|\mathbf{N}_j(s\wedge\tau_R)\|^2 \leq C_R^2 N\Delta. \tag{30}$$

Because the integrand of B is dominated by a nonnegative process, $\sup_{t\leq T}|B(t)|$ is at most $L_\Phi \int_0^{T\wedge\tau_R}(\|\mathbf{x}_s - \mathbf{x}_{t_{j(s)}}\| + \Delta)\, ds$, and by Cauchy–Schwarz in s,

$$\mathbb{E}\sup_{t\leq T}|B|^2 \ \leq \ 3L_\Phi^2 T^2\Big[C_R^2\big(1 + \alpha\omega r_{\max}\big)\Delta^2 + \Delta^2\Big] \ + \ 3L_\Phi^2\, T \int_0^T \mathbb{E}\|\mathbf{N}_{j(s)}(s\wedge\tau_R)\|^2 ds, \tag{31}$$

the last term being at most $3L_\Phi^2 T^2 C_R^2 N\Delta$. Combining,

$$\mathbb{E}\sup_{t\le T}\Big|\int_0^{t^\tau}\Phi e^{\mathrm{i}c\omega s}ds\Big|^2 \;\le\; C\Big[\frac{1}{\Delta^2\omega^2}+\omega\Delta^2+\Delta^2+\Delta\Big]. \tag{32}$$

Choosing $\Delta=\omega^{-3/4}$ makes the first two brackets $O(\omega^{-1/2})$ and the last two $O(\omega^{-3/4})$, which proves Eq. (27); real and imaginary parts give the trigonometric versions. □

Remark 1 (Sharper rate under one more derivative). If $g_n, J\in C^3$ in $\mathbf{x}$ (so that Φ below is C^1 with Lipschitz derivative on $\bar B_R$), the partition argument can be replaced by a genuine Itô integration by parts against $e^{\mathrm{i}c\omega s}/(\mathrm{i}c\omega)$: the boundary terms are $O(\omega^{-1})$, the drift part of $d\Phi(\mathbf{x}_s,s)$ is $O(\sqrt{\omega})$ on the localized set, contributing $O(\omega^{-1}\cdot\sqrt{\omega})=O(\omega^{-1/2})$, and the martingale part contributes $O_{L^2}(\omega^{-1})$ by Doob's inequality. This sharpens Eq. (27) to $C\omega^{-1}$, i.e., $O_{L^2}(\omega^{-1/2})$ for the integral itself, and correspondingly sharpens the overall rate in Theorem 1 from $O_{\mathbb{P}}(\omega^{-1/4})$ to $O_{\mathbb{P}}(\omega^{-1/2})$.

A.4 Proof of Theorem 1

Step 0: localization. Fix $R\ge 2$ and let τ_R be as in Eq. (26). All estimates in Steps 1–3 are on $[0,\tau_R]$, where both processes lie in $\bar B_R$ and all bounds C_R of Assumption 1 apply.

Step 1: Itô integration by parts on the dither term. Fix a coordinate n and write the mild form of Eq. (20):

$$x_{\omega,n}(t)=x_{0,n}+\int_0^t f_n\,ds+I_{2,n}(t)+\int_0^t g_n\,dw_n,\qquad I_{2,n}(t)=\int_0^t g_n\sqrt{\alpha\omega_n}\,\cos\big(\omega_n s+kJ\big)\,ds, \tag{33}$$

all integrands evaluated along $(\mathbf{x}_\omega(s),s)$. Expand the cosine and define the $C^{2,1}$ compositions

$$u^c \;=\; g_n\cos(kJ),\qquad u^s \;=\; g_n\sin(kJ), \tag{34}$$

so that $I_{2,n}=\int u^c\,dv_c-\int u^s\,dv_s$ with the deterministic, finite-variation dither antiderivatives

$$v_c(s)=\sqrt{\alpha/\omega_n}\,\sin(\omega_n s),\qquad v_s(s)=-\sqrt{\alpha/\omega_n}\,\cos(\omega_n s),\qquad |v_c|,|v_s|\le\sqrt{\alpha/\omega_n}. \tag{35}$$

Since v_c, v_s are C^1 and deterministic, the Itô product rule has vanishing covariation, and

$$I_{2,n}(t) \;=\; \big[u^c v_c-u^s v_s\big]_0^t \;-\;\int_0^t v_c\,du^c \;+\;\int_0^t v_s\,du^s. \tag{36}$$

By Itô's formula applied to $u\in\{u^c,u^s\}$ along Eq. (20),

$$du \;=\; \Big[\partial_t u+(\nabla u)^\top\mathbf{f}+\tfrac12\,\mathrm{tr}\,\big(\boldsymbol{\sigma}\boldsymbol{\sigma}^\top\nabla^2 u\big)\Big]ds \;+\; (\nabla u)^\top\boldsymbol{\sigma}\,d\mathbf{w} \;+\; (\nabla u)^\top\mathbf{D}\,ds, \tag{37}$$

where $\mathbf{D}$ is the dither drift, $D_i = g_i\sqrt{\alpha\omega_i}\cos(\omega_i s+kJ)$. Insert this into Eq. (36) and collect all terms whose prefactor carries $\sqrt{\alpha/\omega_n}$ *without* a compensating $\sqrt{\omega_i}$: the boundary terms (bounded by $4C_R\sqrt{\alpha/\omega_n}$, noting $v_s(0)\neq 0$), the $\partial_t u$, $(\nabla u)^\top\mathbf{f}$, and Hessian terms (each bounded by $C_R\sqrt{\alpha/\omega_n}\,T$), and the stochastic integrals $-\int v_c(\nabla u^c)^\top\boldsymbol{\sigma}\,d\mathbf{w}+\int v_s(\nabla u^s)^\top\boldsymbol{\sigma}\,d\mathbf{w}$, whose supremum has second moment at most $16(\alpha/\omega_n)C_R^2T$ by Doob's maximal inequality and the Itô isometry (a factor 4 from Doob and a factor 2 from the sum of the two martingales). Denoting their sum by $E_n^{(1)}(t)$,

$$\mathbb{E}\Big[\sup_{t\le T}\big|E_n^{(1)}(t\wedge\tau_R)\big|^2\Big] \;\le\; C_R\,\omega^{-1}. \tag{38}$$

Step 2: resonant and non-resonant dither products. The remaining terms of Eq. (36) are those where du contributes the dither $\mathbf{D}$:

$$\begin{aligned}&-\sum_{i=1}^N\alpha\sqrt{\tfrac{r_i}{r_n}}\int_0^t\sin(\omega_n s)\,\partial_i u^c\,g_i\big[\cos(\omega_i s)\cos kJ-\sin(\omega_i s)\sin kJ\big]ds\\ &-\sum_{i=1}^N\alpha\sqrt{\tfrac{r_i}{r_n}}\int_0^t\cos(\omega_n s)\,\partial_i u^s\,g_i\big[\cos(\omega_i s)\cos kJ-\sin(\omega_i s)\sin kJ\big]ds,\end{aligned} \tag{39}$$

where the second group, which carries the same overall sign as the first, arises from the term $+\int v_s(\nabla u^s)^\top \mathbf{D}\, ds$ with $v_s = -\sqrt{\alpha/\omega_n}\cos(\omega_n s)$. Every product of two trigonometric factors in Eq. (39) reduces by the product formulas to $\frac{1}{2}[\mathrm{trig}((r_n - r_i)\omega s) \pm \mathrm{trig}((r_n + r_i)\omega s)]$; for $i \neq n$ all resulting frequencies $(r_n \pm r_i)\omega$ are nonzero by Assumption 2, and for $i = n$ the frequencies are 0 and $2r_n\omega$. Each nonzero-frequency term is an oscillatory integral of the form covered by Lemma 1 with $\Phi \in \{\partial_i u^{c/s}\, g_i \cos kJ,\ \partial_i u^{c/s}\, g_i \sin kJ\}$, which are bounded and jointly Lipschitz on $\bar{B}_R \times [0,T]$ because $g_n, J \in C^{2,1}$ with continuous mixed derivatives $\partial_t \nabla_{\mathbf{x}} g_n, \partial_t \nabla_{\mathbf{x}} J$ (Assumption 1), so that $\partial_i u^{c/s} \in C^1$ jointly in $(\mathbf{x}, t)$, hence Lipschitz on the compact set. Their total, $E_n^{(2)}(t)$, therefore satisfies

$$\mathbb{E}\Big[\sup_{t\le T}\big|E_n^{(2)}(t\wedge\tau_R)\big|^2\Big] \;\le\; C_R\,\omega^{-1/2}. \tag{40}$$

The zero-frequency (resonant) halves survive averaging. From the first group of Eq. (39) with $i = n$, using $\sin(\omega_n s)\sin(\omega_n s) = \frac{1}{2} - \frac{1}{2}\cos(2\omega_n s)$, the surviving term is $+\frac{\alpha}{2}\int_0^t g_n\,\partial_n u^c\,\sin(kJ)\,ds$; from the second group, using $\cos^2(\omega_n s) = \frac{1}{2} + \frac{1}{2}\cos(2\omega_n s)$, it is $-\frac{\alpha}{2}\int_0^t g_n\,\partial_n u^s\,\cos(kJ)\,ds$. Substituting $\partial_n u^c = \partial_n g_n \cos kJ - k g_n\,\partial_n J\,\sin kJ$ and $\partial_n u^s = \partial_n g_n \sin kJ + k g_n\,\partial_n J\,\cos kJ$,

$$g_n\big[\partial_n u^c \sin kJ - \partial_n u^s \cos kJ\big] \;=\; -k\,g_n^2\,\partial_n J, \tag{41}$$

the $\partial_n g_n$ terms cancelling identically. The resonant contribution to $I_{2,n}$ is therefore

$$-\frac{k\alpha}{2}\int_0^t g_n^2\big(\mathbf{x}_\omega(s), s\big)\,\partial_n J\big(\mathbf{x}_\omega(s), s\big)\,ds. \tag{42}$$

Collecting Eq. (38)–Eq. (42) and writing $\mathbf{E} = (E_n^{(1)} + E_n^{(2)})_{n=1}^N$, we obtain on $[0, \tau_R]$ the exact representation

$$\begin{aligned}\mathbf{x}_\omega(t) \;&=\; \mathbf{x}_0 + \int_0^t \bar{\boldsymbol{\mu}}\big(\mathbf{x}_\omega(s), s\big)ds + \int_0^t \boldsymbol{\sigma}\big(\mathbf{x}_\omega(s), s\big)d\mathbf{w} + \mathbf{E}(t),\\ \mathbb{E}\sup_{t\le T}\big\|\mathbf{E}(t\wedge\tau_R)\big\|^2 \;&\le\; C_R\,\omega^{-1/2}.\end{aligned} \tag{43}$$

Thus, up to the explicitly controlled remainder $\mathbf{E}$, the dithered process satisfies the *averaged* SDE along its own trajectory.

Step 3: Grönwall comparison under the shared Brownian motion. Let $\mathbf{z} = \mathbf{x}_\omega - \bar{\mathbf{x}}$. Subtracting the mild form of Eq. (22) from Eq. (43),

$$\mathbf{z}(t) = \int_0^t \big[\bar{\boldsymbol{\mu}}(\mathbf{x}_\omega) - \bar{\boldsymbol{\mu}}(\bar{\mathbf{x}})\big]ds + \int_0^t \big[\boldsymbol{\sigma}(\mathbf{x}_\omega) - \boldsymbol{\sigma}(\bar{\mathbf{x}})\big]d\mathbf{w} + \mathbf{E}(t). \tag{44}$$

On $\bar{B}_R \times [0,T]$, $\bar{\boldsymbol{\mu}}$ and $\boldsymbol{\sigma}$ are Lipschitz with a constant L_R: for $\boldsymbol{\sigma}$ this is Assumption 1 directly, and for $\bar{\boldsymbol{\mu}} = \mathbf{f} - \frac{k\alpha}{2}\boldsymbol{\sigma}\boldsymbol{\sigma}^\top \nabla J$ it holds because products of C^1 functions of $(g_n, \partial_n J)$ are Lipschitz on compacts. Set $\varphi(t) = \mathbb{E}\big[\sup_{s\le t}\|\mathbf{z}(s\wedge\tau_R)\|^2\big]$. By Cauchy–Schwarz on the drift difference and Doob plus the Itô isometry on the (diagonal) diffusion difference,

$$\varphi(t) \;\le\; 3(T+4)L_R^2\int_0^t \varphi(s)\,ds \;+\; 3\,\mathbb{E}\sup_{s\le T}\|\mathbf{E}(s\wedge\tau_R)\|^2, \tag{45}$$

and Grönwall's inequality with Eq. (43) gives

$$\mathbb{E}\Big[\sup_{t\le T}\|\mathbf{z}(t\wedge\tau_R)\|^2\Big] \;\le\; C_R'\,\omega^{-1/2}, \qquad C_R' = 3\,C_R\,e^{3(T+4)L_R^2 T}. \tag{46}$$

Step 4: removing the localization. Let $\delta \in (0,1]$ and $\varepsilon \in (0,1)$ be given. By the standard moment bound for Eq. (22) under Assumption 3, $\mathbb{E}\sup_{t\le T}\|\bar{\mathbf{x}}\|^2 < \infty$, so we may fix R with $\mathbb{P}\big(\sup_{t\le T}\|\bar{\mathbf{x}}(t)\| \ge R-2\big) \le \varepsilon/2$. By Eq. (46) and Chebyshev's inequality, choose ω^* so that for all $\omega \ge \omega^*$, $\mathbb{P}\big(\sup_{t\le T}\|\mathbf{z}(t\wedge\tau_R)\| > \delta\big) \le C_R'\,\omega^{-1/2}\delta^{-2} \le \varepsilon/2$. On the event $A = \{\sup_{t\le T}\|\bar{\mathbf{x}}\| \le R-2\} \cap \{\sup_{t\le T}\|\mathbf{z}(t\wedge\tau_R)\| \le \delta\}$, suppose that the uncapped exit time satisfies $\tilde{\tau}_R \le T$; then $\tau_R = \tilde{\tau}_R$ and by continuity one of the processes has norm exactly R at τ_R, yet $\|\bar{\mathbf{x}}(\tau_R)\| \le R-2$ and

$\|\mathbf{x}_\omega(\tau_R)\| \le \|\bar{\mathbf{x}}(\tau_R)\| + \delta \le R - 1 < R$, a contradiction. Hence $\tilde{\tau}_R > T$ on A, so $\tau_R = T$ there, the stopped and unstopped suprema coincide, and

$$\mathbb{P}\Big(\sup_{t\le T} \|\mathbf{x}_\omega(t) - \bar{\mathbf{x}}(t)\| > \delta\Big) \;\le\; \mathbb{P}(A^c) \;\le\; \varepsilon \qquad \text{for all } \omega \ge \omega^*. \tag{47}$$

This proves Theorem 1; Eq. (46) also yields the localized rate $O_{\mathbb{P}}(\omega^{-1/4})$, sharpened to $O_{\mathbb{P}}(\omega^{-1/2})$ under the extra smoothness of Remark 1. □

A.5 Corollaries

Corollary 1 (Convergence in distribution on path space)**.** *Under the assumptions of Theorem 1,* $\mathrm{Law}(\mathbf{x}_\omega) \Rightarrow \mathrm{Law}(\bar{\mathbf{x}})$ *on* $C([0,T];\mathbb{R}^N)$ *as* $\omega \to \infty$.

Proof. Theorem 1 states convergence in probability of $\mathbf{x}_\omega \to \bar{\mathbf{x}}$ in the Polish space $C([0,T];\mathbb{R}^N)$ with the uniform norm; convergence in probability implies convergence in distribution (Billingsley, 1999). □

Corollary 2 (Sampling in the iterated limit)**.** *Take instantiation* (I_λ)*:* $\mathbf{f} \equiv \mathbf{0}$, $\boldsymbol{\sigma} = \sqrt{2}\, I$, $k\alpha = \lambda > 0$, $J = \mathcal{E} \in C^2$ *independent of* t *with* $\nabla\mathcal{E}$ *of linear growth (Assumption 3; e.g.,* $\nabla\mathcal{E}$ *globally Lipschitz), and assume additionally the dissipativity condition* $\langle \nabla\mathcal{E}(\mathbf{x}), \mathbf{x}\rangle \ge c_1\|\mathbf{x}\|^2 - c_2$ *for constants* $c_1 > 0, c_2 \ge 0$ *and all* $\|\mathbf{x}\|$ *large. (Superlinear potentials such as quartic double wells violate the linear-growth condition and would require a Lyapunov-based localization of Theorem 1.) Then the averaged process is the Langevin diffusion with unique stationary law* $\pi_\lambda \propto e^{-\lambda\mathcal{E}}$*, it is exponentially ergodic (Roberts & Tweedie, 1996), and for every bounded Lipschitz* $h : \mathbb{R}^N \to \mathbb{R}$*,*

$$\lim_{T\to\infty} \lim_{\omega\to\infty} \mathbb{E}\big[h\big(\mathbf{x}_\omega(T)\big)\big] \;=\; \int h\, d\pi_\lambda. \tag{48}$$

The order of the limits cannot be interchanged on the basis of Theorem 1, whose threshold ω^* *depends on* T.

Proof. For fixed T, split $|\mathbb{E}h(\mathbf{x}_\omega(T)) - \pi_\lambda(h)| \le |\mathbb{E}h(\mathbf{x}_\omega(T)) - \mathbb{E}h(\bar{\mathbf{x}}(T))| + |\mathbb{E}h(\bar{\mathbf{x}}(T)) - \pi_\lambda(h)|$. For any $\delta > 0$ the first term is at most $\mathrm{Lip}(h)\,\delta + 2\|h\|_\infty \mathbb{P}(\|\mathbf{x}_\omega(T) - \bar{\mathbf{x}}(T)\| > \delta) \to \mathrm{Lip}(h)\,\delta$ as $\omega \to \infty$ by Theorem 1, and δ is arbitrary. The second term decays as $C_h e^{-\rho T}$ by exponential ergodicity of the Langevin diffusion under dissipativity (Roberts & Tweedie, 1996). Let $\omega \to \infty$, then $T \to \infty$. □

Under the additional strong log-concavity hypothesis $\nabla^2\mathcal{E} \succeq mI$, Appendix A.6 upgrades this iterated limit to $\lim_{\omega\to\infty} \limsup_{T\to\infty} \mathcal{W}_2(\mathrm{Law}(\mathbf{x}_\omega(T)), \pi_\lambda) = 0$, i.e., the frequency limit may be taken after (or jointly with) the time limit—conditionally on the start-uniform window estimate of Assumption 4 (unconditionally at the level of the transfer principle, Proposition 2).

Corollary 3 (λ-tempering)**.** *In the setting of Corollary 2, the stationary density of the averaged dynamics* $d\bar{\mathbf{x}} = -\lambda\nabla\mathcal{E}\, dt + \sqrt{2}\, d\mathbf{w}$ *is* $\rho_\infty \propto e^{-\lambda\mathcal{E}} \propto p^\lambda$.

Proof. The stationary Fokker–Planck equation is $0 = \nabla\cdot\big(\lambda\nabla\mathcal{E}\,\rho + \nabla\rho\big)$, solved by any ρ with $\nabla\rho = -\lambda\rho\nabla\mathcal{E}$, i.e., $\rho \propto e^{-\lambda\mathcal{E}}$; uniqueness under dissipativity is classical (Roberts & Tweedie, 1996). □

A.6 Uniform-in-time accuracy under strong log-concavity

Throughout this subsection $\mathbf{f}, \boldsymbol{\sigma}, J$ are independent of t, and Assumption 3 is in force, so that for each fixed ω both Eq. (20) and Eq. (22) have unique global strong solutions with finite second moments at every time. We write $\mathcal{P}_2(\mathbb{R}^N)$ for the probability measures with finite second moment, $\mathcal{W}_2$ for the Wasserstein-2 distance, under which $\mathcal{P}_2(\mathbb{R}^N)$ is a complete metric space (Villani, 2009), and $(\bar{P}_t)_{t\ge0}$ for the Markov semigroup of the averaged SDE acting on laws, $\mu\bar{P}_t := \mathrm{Law}(\bar{\mathbf{x}}(t))$ for $\bar{\mathbf{x}}(0) \sim \mu$; under Assumption 3 with locally Lipschitz coefficients, $\bar{P}_t$ maps $\mathcal{P}_2$ into itself (Karatzas & Shreve, 1991, Ch. 5).

Lemma 2 (Start invariance of the window estimate). *Let Assumption 5 hold with $\mathbf{f}, \boldsymbol{\sigma}, J$ independent of t, fix $T_0 > 0$ and $t_0 \geq 0$, and let $\mathbf{x}^*$ be $\mathcal{F}_{t_0}$-measurable with $\mathbb{E}\|\mathbf{x}^*\|^2 < \infty$. Let $\mathbf{x}_\omega$ solve the ES SDE Eq. (20) and $\bar{\mathbf{x}}$ the averaged SDE Eq. (22) on $[t_0, t_0 + T_0]$, driven by the same Brownian motion, from the common initial condition $\mathbf{x}_\omega(t_0) = \bar{\mathbf{x}}(t_0) = \mathbf{x}^*$. Then*

$$\mathbb{E}\Big[\sup_{t_0 \leq t \leq t_0+T_0} \big\|\mathbf{x}_\omega(t) - \bar{\mathbf{x}}(t)\big\|_2^2\Big] \;\leq\; \frac{C(T_0)}{\omega},$$

with $C(T_0) = C_1 e^{C_2 T_0}$ the constant of Theorem 3, independent of t_0 and of the law of $\mathbf{x}^*$.

Proof. In the shifted time $u = t - t_0$ the pair solves the same two equations, driven by the Brownian motion $\tilde{\mathbf{w}}_u := \mathbf{w}_{t_0+u} - \mathbf{w}_{t_0}$ from the (time-0-measurable, for the shifted filtration) start $\mathbf{x}^*$, except that the dither phases carry the constant offsets $\theta_n := \omega_n t_0$: the dither drift reads $g_n\sqrt{\alpha\omega_n}\cos(\omega_n u + \theta_n + kJ)$. Every step of the proof of Theorem 3 is invariant under such offsets: the dither antiderivatives become $v_c(u) = \sqrt{\alpha/\omega_n}\,\sin(\omega_n u + \theta_n)$ and $v_s(u) = -\sqrt{\alpha/\omega_n}\,\cos(\omega_n u + \theta_n)$ and obey the same bounds; the product-to-sum identities of Step 2 produce zero-mean trigonometric factors at the same frequencies $\omega|r_n \pm r_i|$ and $2\omega r_n$, now with constant phases, to which Lemma 3 applies verbatim; and the resonant extraction $\sin^2(\omega_n u + \theta_n) = \frac{1}{2} - \frac{1}{2}\cos(2\omega_n u + 2\theta_n)$ is unchanged. No constant in that proof depends on the initial condition (beyond $\mathbb{E}\|\mathbf{x}^*\|^2 < \infty$, used only for a priori moment finiteness) or on t_0. □

Assumption 4 ((W): start-uniform window averaging). There exist $T_0 > 0$ and $\varepsilon : [1, \infty) \to (0, \infty)$ with $\varepsilon(\omega) \to 0$ as $\omega \to \infty$ such that the coupled pair of Lemma 2 satisfies, for every $t_0 \geq 0$ and every $\mathcal{F}_{t_0}$-measurable $\mathbf{x}^*$ with $\mathbb{E}\|\mathbf{x}^*\|^2 < \infty$,

$$\mathbb{E}\Big[\sup_{t_0 \leq t \leq t_0+T_0} \big\|\mathbf{x}_\omega(t) - \bar{\mathbf{x}}(t)\big\|_2^2\Big] \;\leq\; \varepsilon(\omega)^2\,\big(1 + \mathbb{E}\|\mathbf{x}^*\|^2\big). \tag{49}$$

By Lemma 2, (W) holds with $\varepsilon(\omega) = (C(T_0)/\omega)^{1/2}$ (for any fixed T_0, say $T_0 = 1$), and even without the factor $(1 + \mathbb{E}\|\mathbf{x}^*\|^2)$, whenever Assumption 5 does—in particular whenever ∇J is bounded. The factor is what makes (W) a meaningful hypothesis for coercive J: there the averaged drift grows linearly while the dither drift of Eq. (20) is bounded by $C\sqrt{\omega}$, so from a start with $\|\mathbf{x}^*\| \gg \sqrt{\omega}$ the two processes separate at a rate proportional to $\|\mathbf{x}^*\|$, and no bound uniform in $\mathbf{x}^*$ can hold. For coercive J, (W) is precisely the quantitative, start-uniform content that the localized Theorem 1 does not provide; see Remark 2.

Proposition 2 (Uniform-in-time sampling accuracy). *Let $\mathbf{f}, \boldsymbol{\sigma}, J$ be independent of t, let Assumption 3 hold with locally Lipschitz coefficients, and assume: (i) Assumption 4; and (ii) there is $\rho > 0$ such that $\mathcal{W}_2(\mu\bar{P}_t, \nu\bar{P}_t) \leq e^{-\rho t}\,\mathcal{W}_2(\mu, \nu)$ for all $\mu, \nu \in \mathcal{P}_2(\mathbb{R}^N)$ and all $t \geq 0$. Then $\bar{P}$ has a unique invariant law $\bar{\pi} \in \mathcal{P}_2(\mathbb{R}^N)$, with second moment $\bar{m} := (\int \|\mathbf{x}\|^2\, d\bar{\pi})^{1/2}$, and, writing $\eta := 1 - e^{-\rho T_0} \in (0, 1)$ and $\rho' := \eta/(2T_0)$, for every ω with $\varepsilon(\omega) \leq \eta/2$ and every $t \geq 0$,*

$$\mathcal{W}_2\big(\mathrm{Law}(\mathbf{x}_\omega(t)),\, \bar{\pi}\big) \;\leq\; 2e^{\eta/2}\, e^{-\rho' t}\, \mathcal{W}_2\big(\mathrm{Law}(\mathbf{x}_0),\, \bar{\pi}\big) \;+\; \Big(1 + \frac{4}{\eta}\Big)(1 + \bar{m})\,\varepsilon(\omega). \tag{50}$$

If (W) holds without the factor $(1 + \mathbb{E}\|\mathbf{x}^*\|^2)$, as it does under Assumption 5, the same argument gives the sharper bound $e^{-\rho t}\mathcal{W}_2(\mathrm{Law}(\mathbf{x}_0), \bar{\pi}) + (1 + 1/\eta)\,\varepsilon(\omega)$ for all $\omega \geq 1$. In either case $\lim_{\omega\to\infty} \limsup_{t\to\infty} \mathcal{W}_2(\mathrm{Law}(\mathbf{x}_\omega(t)), \bar{\pi}) = 0$, and after a burn-in of order $\rho'^{-1} \log\big(1/\varepsilon(\omega)\big)$ the law of the ES iterate remains within $C\varepsilon(\omega)$ of $\bar{\pi}$ in $\mathcal{W}_2$ for all *subsequent times simultaneously; the time limit is a* lim sup *because at fixed ω the law of the dithered process is asymptotically periodic in t, with oscillations of amplitude $O(\varepsilon(\omega))$, rather than convergent. Moreover, hypothesis (ii) holds whenever $\boldsymbol{\sigma} \equiv \boldsymbol{\sigma}_0$ is constant and the averaged drift is ρ-strongly monotone, $\langle \bar{\boldsymbol{\mu}}(\mathbf{x}) - \bar{\boldsymbol{\mu}}(\mathbf{y}),\, \mathbf{x} - \mathbf{y}\rangle \leq -\rho\|\mathbf{x} - \mathbf{y}\|^2$ for all $\mathbf{x}, \mathbf{y}$.*

Proof. Sufficiency of strong monotonicity for (ii). Let $\bar{\mathbf{x}}, \bar{\mathbf{y}}$ solve the averaged SDE from initial conditions $(\bar{\mathbf{x}}_0, \bar{\mathbf{y}}_0)$ distributed according to an optimal $\mathcal{W}_2$-coupling of (μ, ν), driven by the *same* Brownian motion. With $\boldsymbol{\sigma}$ constant the noise cancels in the difference, and $\frac{d}{dt}\|\bar{\mathbf{x}} - \bar{\mathbf{y}}\|^2 = 2\langle \bar{\boldsymbol{\mu}}(\bar{\mathbf{x}}) - \bar{\boldsymbol{\mu}}(\bar{\mathbf{y}}), \bar{\mathbf{x}} - \bar{\mathbf{y}}\rangle \leq -2\rho\|\bar{\mathbf{x}} - \bar{\mathbf{y}}\|^2$, so $\|\bar{\mathbf{x}}(t) - \bar{\mathbf{y}}(t)\| \leq e^{-\rho t}\|\bar{\mathbf{x}}_0 - \bar{\mathbf{y}}_0\|$ pathwise; taking $L^2(\mathbb{P})$-norms yields $\mathcal{W}_2(\mu\bar{P}_t, \nu\bar{P}_t) \leq e^{-\rho t}\mathcal{W}_2(\mu, \nu)$.

Invariant law. By (ii), $\mu \mapsto \mu\bar{P}_{T_0}$ is a strict contraction on the complete metric space $(\mathcal{P}_2, \mathcal{W}_2)$ (Villani, 2009), hence has a unique fixed point $\bar{\pi}$. For $s > 0$, $(\bar{\pi}\bar{P}_s)\bar{P}_{T_0} = (\bar{\pi}\bar{P}_{T_0})\bar{P}_s = \bar{\pi}\bar{P}_s$, so $\bar{\pi}\bar{P}_s$ is also a fixed point of $\bar{P}_{T_0}$ and therefore equals $\bar{\pi}$: $\bar{\pi}$ is invariant for all times, and unique in $\mathcal{P}_2$ by (ii).

Window recursion. Fix $\omega \geq 1$, set $t_j := jT_0$, write $\mu_t := \mathrm{Law}(\mathbf{x}_\omega(t)) \in \mathcal{P}_2$ (second moments are finite at every time for fixed ω by Assumption 3), and $a_j := \mathcal{W}_2(\mu_{t_j}, \bar{\pi}) < \infty$. On $[t_j, t_{j+1}]$ let $\bar{\mathbf{x}}^{(j)}$ solve the averaged SDE from $\bar{\mathbf{x}}^{(j)}(t_j) = \mathbf{x}_\omega(t_j)$, driven by the same Brownian motion as $\mathbf{x}_\omega$. Since the averaged coefficients are time-homogeneous and the increments $(\mathbf{w}_{t_j+s} - \mathbf{w}_{t_j})_{s\geq 0}$ are independent of $\mathcal{F}_{t_j}$, $\mathrm{Law}(\bar{\mathbf{x}}^{(j)}(t_j + s)) = \mu_{t_j}\bar{P}_s$ for $s \in [0, T_0]$; and $\big(\mathbf{x}_\omega(t_j + s), \bar{\mathbf{x}}^{(j)}(t_j + s)\big)$ is a coupling of $(\mu_{t_j+s},\ \mu_{t_j}\bar{P}_s)$, so by Eq. (49) with $\mathbf{x}^* = \mathbf{x}_\omega(t_j)$,

$$\begin{aligned}\mathcal{W}_2\big(\mu_{t_j+s},\ \mu_{t_j}\bar{P}_s\big) \ &\leq\ \Big(\mathbb{E}\big\|\mathbf{x}_\omega(t_j + s) - \bar{\mathbf{x}}^{(j)}(t_j + s)\big\|_2^2\Big)^{1/2} \\ &\leq\ \varepsilon(\omega)\,\big(1 + m_2(\mu_{t_j})\big)\ \leq\ \varepsilon(\omega)\,\big(1 + \bar{m} + a_j\big), \qquad s \in [0, T_0],\end{aligned}$$

where $m_2(\mu) := (\int \|\mathbf{x}\|^2 d\mu)^{1/2} = \mathcal{W}_2(\mu, \delta_{\mathbf{0}}) \leq \mathcal{W}_2(\mu, \bar{\pi}) + \bar{m}$. Hence, by the triangle inequality, invariance of $\bar{\pi}$, and (ii),

$$\begin{aligned}a_{j+1} \ &\leq\ \mathcal{W}_2\big(\mu_{t_{j+1}}, \mu_{t_j}\bar{P}_{T_0}\big) + \mathcal{W}_2\big(\mu_{t_j}\bar{P}_{T_0}, \bar{\pi}\bar{P}_{T_0}\big) \\ &\leq\ \varepsilon(\omega)(1 + \bar{m}) + \big(e^{-\rho T_0} + \varepsilon(\omega)\big)a_j\ \leq\ \varepsilon(\omega)(1 + \bar{m}) + \big(1 - \tfrac{\eta}{2}\big)a_j\end{aligned}$$

for $\varepsilon(\omega) \leq \eta/2$, so that $a_j \leq (1 - \frac{\eta}{2})^j a_0 + \frac{2}{\eta}(1 + \bar{m})\varepsilon(\omega)$: the second moment of μ_{t_j}, which the weighted form of (W) feeds back into the recursion, is controlled by a_j itself, so no separate moment bound is needed. For arbitrary $t = t_j + s$ with $s \in [0, T_0)$,

$$\begin{aligned}\mathcal{W}_2(\mu_t, \bar{\pi}) \ &\leq\ \mathcal{W}_2\big(\mu_t, \mu_{t_j}\bar{P}_s\big) + \mathcal{W}_2\big(\mu_{t_j}\bar{P}_s, \bar{\pi}\bar{P}_s\big)\ \leq\ \varepsilon(\omega)(1 + \bar{m} + a_j) + e^{-\rho s}a_j \\ &\leq\ \varepsilon(\omega)(1 + \bar{m}) + 2a_j\ \leq\ 2e^{\eta/2}e^{-\rho' t}a_0 + \Big(1 + \tfrac{4}{\eta}\Big)(1 + \bar{m})\,\varepsilon(\omega),\end{aligned}$$

using $(1 - \eta/2)^j \leq e^{-\eta j/2} = e^{-\eta t_j/(2T_0)} \leq e^{\eta/2}e^{-\rho' t}$, which is Eq. (50). Without the factor $(1 + \mathbb{E}\|\mathbf{x}^*\|^2)$ in (W) the first display gives $\mathcal{W}_2(\mu_{t_j+s}, \mu_{t_j}\bar{P}_s) \leq \varepsilon(\omega)$, the recursion reads $a_{j+1} \leq \varepsilon(\omega) + e^{-\rho T_0}a_j$, and the same steps give the sharper bound quoted after Eq. (50). □

Corollary 4 (Strongly log-concave targets)**.** *Take instantiation (I_λ): $\mathbf{f} \equiv \mathbf{0}$, $\boldsymbol{\sigma} = \sqrt{2}\,I$, $k\alpha = \lambda > 0$, and $J = \mathcal{E} \in C^2$ independent of t with $\nabla\mathcal{E}$ globally Lipschitz and $\nabla^2\mathcal{E}(\mathbf{x}) \succeq mI$ for all $\mathbf{x}$ and some $m > 0$ (a strongly log-concave target). Then hypothesis (ii) of Proposition 2 holds with $\rho = \lambda m$, the unique invariant law is $\pi_\lambda \propto e^{-\lambda\mathcal{E}} \in \mathcal{P}_2$, and consequently, under Assumption 4, with $\eta = 1 - e^{-\lambda m T_0}$, $\rho' = \eta/(2T_0)$ and $\bar{m}$ the second moment of π_λ,*

$$\mathcal{W}_2\big(\mathrm{Law}(\mathbf{x}_\omega(t)),\ \pi_\lambda\big)\ \leq\ 2e^{\eta/2}e^{-\rho' t}\,\mathcal{W}_2\big(\mathrm{Law}(\mathbf{x}_0),\ \pi_\lambda\big) + \Big(1 + \tfrac{4}{\eta}\Big)(1 + \bar{m})\,\varepsilon(\omega)$$

for all $t \geq 0$ and all ω with $\varepsilon(\omega) \leq \eta/2$, so the iterated limit of Corollary 2—whose dissipativity hypothesis is implied by strong convexity—upgrades to $\lim_{\omega\to\infty} \limsup_{t\to\infty} \mathcal{W}_2(\mathrm{Law}(\mathbf{x}_\omega(t)), \pi_\lambda) = 0$, with the frequency limit taken after, or jointly with, the time limit.

Proof. The averaged drift is $\bar{\boldsymbol{\mu}} = -\lambda\nabla\mathcal{E}$, and strong convexity gives $\langle\bar{\boldsymbol{\mu}}(\mathbf{x}) - \bar{\boldsymbol{\mu}}(\mathbf{y}), \mathbf{x} - \mathbf{y}\rangle = -\lambda\langle\nabla\mathcal{E}(\mathbf{x}) - \nabla\mathcal{E}(\mathbf{y}), \mathbf{x} - \mathbf{y}\rangle \leq -\lambda m\|\mathbf{x} - \mathbf{y}\|^2$, so the sufficient condition of Proposition 2 applies with $\rho = \lambda m$ and constant $\boldsymbol{\sigma}$. The law π_λ is invariant by Corollary 3, lies in $\mathcal{P}_2$ because $\mathcal{E}(\mathbf{x}) \geq \mathcal{E}(\mathbf{x}_*) + \frac{m}{2}\|\mathbf{x} - \mathbf{x}_*\|^2$ at the minimizer $\mathbf{x}_*$ makes $e^{-\lambda\mathcal{E}}$ sub-Gaussian, and is the unique invariant law in $\mathcal{P}_2$ by the contraction. The well-posedness and moment bounds required by Assumption 3 hold because $\nabla\mathcal{E}$ is Lipschitz, hence of linear growth. □

Remark 2 (Status of the uniform-in-time program). The two ingredients of Proposition 2 are individually proved but, at present, never simultaneously. Under Assumption 5, (W) is a theorem (Lemma 2), but no bounded drift satisfies (ii): if $\|\bar{\boldsymbol{\mu}}\|_\infty = B < \infty$, the mean functions $\mathbf{m}_\mathbf{x}(t) := \mathbb{E}[\bar{\mathbf{x}}(t) \mid \bar{\mathbf{x}}(0) = \mathbf{x}]$ obey $\|\mathbf{m}_\mathbf{x}(t) - \mathbf{m}_\mathbf{y}(t)\| \geq \|\mathbf{x} - \mathbf{y}\| - 2Bt$, and since every coupling satisfies $\mathbb{E}\|X - Y\| \geq \|\mathbb{E}X - \mathbb{E}Y\|$, also $\mathcal{W}_2(\delta_\mathbf{x}\bar{P}_t, \delta_\mathbf{y}\bar{P}_t) \geq \|\mathbf{m}_\mathbf{x}(t) - \mathbf{m}_\mathbf{y}(t)\|$; the contraction $e^{-\rho t}\|\mathbf{x} - \mathbf{y}\| \geq \|\mathbf{x} - \mathbf{y}\| - 2Bt$ then fails whenever $\|\mathbf{x} - \mathbf{y}\| > 2Bt/(1 - e^{-\rho t})$. Conversely, under strong log-concavity (ii) holds, but $\nabla^2\mathcal{E} \succeq mI$ forces $\|\nabla\mathcal{E}(\mathbf{x})\| \geq m\|\mathbf{x}\| - \|\nabla\mathcal{E}(\mathbf{0})\|$, so $\nabla\mathcal{E}$ is

unbounded and Assumption 5—hence the proof of (W)—is unavailable; the localization behind Theorem 1 covers such coercive energies, but yields an in-probability window statement without the start-uniform mean-square form that the recursion consumes. (The growth weight $(1+\mathbb{E}\|\mathbf{x}^*\|^2)$ in Eq. (49) is not a convenience but a necessity here: the dither drift is bounded by $C\sqrt{\omega}$ while $\|\bar{\boldsymbol{\mu}}(\mathbf{x})\| \geq \lambda m\|\mathbf{x}\| - C$, so the window error from a start $\mathbf{x}^*$ is at least of order $(\lambda m\|\mathbf{x}^*\| - C\sqrt{\omega})_+ T_0$, which no $\mathbf{x}^*$-free $\varepsilon(\omega)$ can bound.) What a proof of (W) for m-strongly convex $\mathcal{E}$ with Lipschitz gradient requires is a linear-growth-weighted extension of the quantitative machinery: versions of Lemmas 3–4 for slow factors of linear growth along $\mathbf{x}_\omega$, together with ω-uniform second- and fourth-moment bounds for the dithered process, in the spirit of the perturbed-Lyapunov-function method of stochastic averaging (Kushner, 1984). We expect (W) to hold there with $\varepsilon(\omega) = C\omega^{-1/2}$ and record it as an assumption rather than assert it. In summary: the transfer principle and the contraction are proved (Proposition 2, Corollary 4), and the start-uniform window estimate is the single isolated remaining ingredient in the coercive case—in the same spirit as Conjecture 1 on the Lipschitz rung, which is proved in Theorem 4 under Lipschitz time dependence (Theorem 4); the pathwise near-identity transformation of Appendix B.6, which is a perturbed-Lyapunov device in all but name, is the natural starting point for the moment bounds that (W) requires.

B Relaxed regularity: sampling non-differentiable energies

(Throughout Appendices B–B.7, as in Appendix A.4, $N = d$ denotes the state dimension.) Theorem 1 assumes $J(\cdot, t) \in C^2$, but this is a convenience of the proof, not a requirement of the method: the algorithm Eq. (20)–Eq. (21) only ever evaluates $\cos(\omega_n t + kJ)$, which is well defined for any measurable J, and the averaged drift only needs ∇J to exist in a suitably weak sense. Two mechanisms make weak derivatives admissible. The first is purely analytic and needs no noise at all: if ∇J is Lipschitz, mollification costs only ε-powers that the dither frequency can absorb (Appendix B.2). The second is the noise: for a uniformly non-degenerate diffusion, the law of the process is absolutely continuous, so quantities defined only almost everywhere are seen unambiguously along the trajectory (Appendix B.3). We use the first to prove the $C^{1,1}$ case with a rate, and both together for the Sobolev case.

Throughout this appendix, mollification is in $\mathbf{x}$ only: $J_\varepsilon := J *_{\mathbf{x}} \rho_\varepsilon$ for a standard mollifier ρ_ε, and Assumption 5 below is retained for $\mathbf{f}, \boldsymbol{\sigma}$ and for the t-regularity of J (bounded variation, uniformly in $\mathbf{x}$; Theorem 2 strengthens this to $\partial_t J$ bounded, which its Itô–Krylov step consumes). "Uniformly locally" refers to the norms $\|h\|_{L^p_{\mathrm{uloc}}} := \sup_{\mathbf{x}_0} \|h\|_{L^p(B_1(\mathbf{x}_0))}$. On every rung below, ∇J remains (essentially) bounded, so the dither drift $\mathbf{x} \mapsto g_n(\mathbf{x}, t)\sqrt{\alpha\omega_n}\cos(\omega_n t + kJ(\mathbf{x}, t))$ stays globally Lipschitz in $\mathbf{x}$ and Eq. (20) retains unique global strong solutions for each fixed ω; the well-posedness questions below concern only the averaged limit.

B.1 Global hypotheses and the quantitative baseline

The rough-energy results are cleanest under the following global form of the hypotheses, which removes the localization of Steps 0 and 4 of Appendix A.4 and upgrades the convergence to a mean-square estimate with an explicit rate.

Assumption 5 (Global regularity)**.** There exist finite constants such that, uniformly on $\mathbb{R}^N \times [0, T]$: (i) each f_n, g_n is twice continuously differentiable in $\mathbf{x}$, with the function and its first two $\mathbf{x}$-derivatives bounded and globally Lipschitz in $\mathbf{x}$; J is twice continuously differentiable in $\mathbf{x}$ with $\nabla_{\mathbf{x}} J$ and $\nabla^2_{\mathbf{x}} J$ bounded and globally Lipschitz in $\mathbf{x}$ (J itself may be unbounded: it enters Eq. (21) only through the 2π-periodic phase of the cosine, and no bound on $|J|$ is used anywhere below, so linear-growth energies with bounded gradient—Laplace-type tails—are covered as stated); (ii) f_n, g_n, J and their first two $\mathbf{x}$-derivatives are of bounded variation in t, uniformly in $\mathbf{x}$ (continuous differentiability in t with bounded derivative, as in Assumption 1, is a special case); (iii) the dither ratios $r_1, \ldots, r_N > 0$ are pairwise distinct (Assumption 2) and normalized so that $\min_n r_n = 1$ (a common scale is absorbed into ω; without this normalization the ratio-only dependence of the constants below would not be scale-invariant).

We write C_0 for a common bound on $|f_n|, |g_n|$, on their first two $\mathbf{x}$-derivatives, and on $|\nabla_{\mathbf{x}} J|, |\nabla^2_{\mathbf{x}} J|$; L for a common Lipschitz constant (in $\mathbf{x}$) of these quantities; and V for a common bound on their total variation in t, all uniform on $\mathbb{R}^N \times [0, T]$. We abbreviate $r_{\min} = \min_n r_n = 1$, $r_{\max} = \max_n r_n$,

and

$$c_r \;:=\; \max\Big(\max_{i\neq j}|r_i-r_j|^{-1},\;(2r_{\min})^{-1}\Big), \tag{51}$$

the reciprocal frequency spacing that carries the dimension dependence of the constants (cf. Sec. 3.2 and Appendix F). All constants C in this appendix depend only on $(T, N, \alpha, k, C_0, L, V, r_{\max}, c_r)$ and may change from line to line; we take $\omega \geq 1$ so that $\omega^{-1} \leq \omega^{-1/2}$.

Lemma 3 (Dither antiderivative). *Let $\zeta_\omega(s) = \sum_m c_m \,\mathrm{trig}(\omega\nu_m s + \phi_m)$ be a finite sum with $\mathrm{trig} \in \{\sin, \cos\}$, arbitrary fixed phases $\phi_m \in \mathbb{R}$, $\nu_m \neq 0$, and $\sum_m |c_m| \leq 1$. Then $A_\omega(t) := \int_0^t \zeta_\omega(s)\,ds$ satisfies $\|A_\omega\|_{L^\infty[0,T]} \leq 2c_\zeta/\omega$ with $c_\zeta := \max_m |\nu_m|^{-1}$. For the zero-mean trigonometric factors produced by the dither products Eq. (39) of Step 2, $\nu_m \in \{r_n \pm r_i,\, 2r_n\}$, and hence $c_\zeta \leq c_r$ of Eq. (51).*

Proof. Integrate each summand explicitly: $\big|\int_0^t \mathrm{trig}(\omega\nu s + \phi)\,ds\big| = \big|\mathrm{trig}^*(\omega\nu t + \phi) - \mathrm{trig}^*(\phi)\big|/(\omega|\nu|) \leq 2(\omega|\nu|)^{-1}$, where trig^* denotes the corresponding antiderivative ($\sin$ for $\cos$, and $-\cos$ for $\sin$). □

Lemma 4 (Oscillatory integrals along the ES process, quantitative form). *Let Assumption 5 hold for $\mathbf{f}, \boldsymbol{\sigma}$ (for J, only boundedness and Lipschitz continuity of ∇J enter, through the drift bound), let $\mathbf{x}_\omega$ solve Eq. (20), and let ζ_ω, c_ζ be as in Lemma 3. Let $\Phi : \mathbb{R}^N \times [0,T] \to \mathbb{R}$ be bounded, with $\nabla_{\mathbf{x}}\Phi$ bounded and Lipschitz in $\mathbf{x}$ uniformly in t, and with $\Phi(\mathbf{x}, \cdot)$ of bounded variation uniformly in $\mathbf{x}$. Then, for $\omega \geq 1$,*

$$\mathbb{E}\Big[\sup_{t\leq T}\Big|\int_0^t \Phi\big(\mathbf{x}_\omega(s), s\big)\,\zeta_\omega(s)\,ds\Big|^2\Big] \;\leq\; \frac{C_\Phi\,(1+c_\zeta)^2}{\omega}, \tag{52}$$

with C_Φ depending only on $T, N, \alpha, r_{\max}$, the coefficient bounds, and $\|\Phi\|_\infty, \|\nabla_{\mathbf{x}}\Phi\|_\infty, \mathrm{Lip}(\nabla_{\mathbf{x}}\Phi), \mathrm{TV}_t\Phi$.

Proof. Write $A = A_\omega$ as in Lemma 3, so $\|A\|_\infty \leq 2c_\zeta/\omega$; for readability the bounds below are written with c_ζ/ω, the factor 2 being absorbed into the generic constants. Mollify Φ in $\mathbf{x}$ only: $\Phi_\delta := \Phi *_{\mathbf{x}} \rho_\delta$ with a standard mollifier, $\delta := \omega^{-1/2}$. Then $\|\Phi - \Phi_\delta\|_\infty \leq \|\nabla_{\mathbf{x}}\Phi\|_\infty\delta$, $\|\nabla_{\mathbf{x}}\Phi_\delta\|_\infty \leq \|\nabla_{\mathbf{x}}\Phi\|_\infty$, $\|\nabla^2_{\mathbf{x}}\Phi_\delta\|_\infty \leq \mathrm{Lip}(\nabla_{\mathbf{x}}\Phi)$ (mollification of a Lipschitz gradient does not inflate the bound), and $\mathrm{TV}_t\Phi_\delta \leq \mathrm{TV}_t\Phi$. Split

$$\int_0^t \Phi\,\zeta_\omega\,ds \;=\; \int_0^t (\Phi-\Phi_\delta)\,\zeta_\omega\,ds \;+\; \int_0^t \Phi_\delta\,\zeta_\omega\,ds;$$

the first term is bounded pathwise by $\|\nabla_{\mathbf{x}}\Phi\|_\infty T\,\omega^{-1/2}$. For the second, integrate by parts against the finite-variation function A (the covariation $[\Phi_\delta(\mathbf{x}_\cdot, \cdot), A]$ vanishes):

$$\int_0^t \Phi_\delta\big(\mathbf{x}_\omega(s), s\big)\,dA(s) \;=\; \Phi_\delta\big(\mathbf{x}_\omega(t), t\big)A(t) - \int_0^t A(s)\,d\big[\Phi_\delta\big(\mathbf{x}_\omega(s), s\big)\big],$$

and expand $d[\Phi_\delta(\mathbf{x}_\omega(s), s)]$ by Itô's formula ($\Phi_\delta \in C^2$ in $\mathbf{x}$; the t-dependence contributes a Stieltjes term):

$$d\Phi_\delta \;=\; \Phi_\delta(\mathbf{x}, ds) + \nabla_{\mathbf{x}}\Phi_\delta\cdot\boldsymbol{\mu}_\omega\,ds + \tfrac{1}{2}\,\mathrm{tr}\big(\boldsymbol{\sigma}\boldsymbol{\sigma}^\top\nabla^2_{\mathbf{x}}\Phi_\delta\big)\,ds + (\nabla_{\mathbf{x}}\Phi_\delta)^\top\boldsymbol{\sigma}\,d\mathbf{w}.$$

The resulting terms are bounded as follows, using $\|A\|_\infty \leq c_\zeta/\omega$: the boundary term by $\|\Phi\|_\infty c_\zeta/\omega$; the Stieltjes term by $(c_\zeta/\omega)\,\mathrm{TV}_t\Phi$; the drift term—the dominant one, since $\|\boldsymbol{\mu}_\omega\| \leq C_0\sqrt{N}\,(1+\sqrt{\alpha r_{\max}\omega})$—by $(c_\zeta/\omega)\,\|\nabla_{\mathbf{x}}\Phi\|_\infty\,C_0\sqrt{N}\,(1+\sqrt{\alpha r_{\max}\omega})\,T \leq C\,c_\zeta\,\omega^{-1/2}$; the Itô-correction term by $(c_\zeta/\omega)\frac{N}{2}C_0^2\,\mathrm{Lip}(\nabla_{\mathbf{x}}\Phi)\,T$; and the martingale term $M_t = \int_0^t A\,(\nabla_{\mathbf{x}}\Phi_\delta)^\top\boldsymbol{\sigma}\,d\mathbf{w}$ by Doob's maximal inequality and the Itô isometry, $\mathbb{E}\sup_{t\leq T}|M_t|^2 \leq 4(c_\zeta/\omega)^2\|\nabla_{\mathbf{x}}\Phi\|^2_\infty NC_0^2T$. Collecting: the pathwise-bounded parts total at most $C(1+c_\zeta)\,\omega^{-1/2}$, and the martingale part has mean-square supremum $O(c_\zeta^2\omega^{-2})$; squaring and using $(x+y)^2 \leq 2x^2+2y^2$ gives Eq. (52). □

The mollification step is why Lemma 4 asks only that $\nabla_{\mathbf{x}}\Phi$ be Lipschitz rather than that $\Phi \in C^2$ with Lipschitz second derivatives (compare the genuine integration by parts of Remark 1, which asked for one more derivative): in the application, Φ is built from first derivatives of the correctors u^c, u^s of Step 1, so $\nabla_{\mathbf{x}}\Phi$ involves second derivatives of J and g_n, whose Lipschitz continuity is exactly Assumption 5(i); no third derivatives are ever required.

Theorem 3 (Quantitative averaging under global bounds)**.** *Under Assumption 5 there is a constant $C = C_1 e^{C_2 T}$, with C_1, C_2 depending polynomially on $N, \alpha, |k|, C_0, L, V, r_{\max}$ and c_r, such that for all $\omega \geq 1$,*

$$\mathbb{E}\Big[\sup_{0\leq t\leq T} \big\|\mathbf{x}_\omega(t) - \bar{\mathbf{x}}(t)\big\|_2^2\Big] \;\leq\; \frac{C}{\omega}. \tag{53}$$

In particular $\mathbb{P}\big(\sup_{t\leq T} \|\mathbf{x}_\omega - \bar{\mathbf{x}}\|_2 > \delta\big) \leq C/(\delta^2\omega)$ for every $\delta > 0$, i.e., $\sup_{t\leq T} \|\mathbf{x}_\omega - \bar{\mathbf{x}}\| = O_{\mathbb{P}}(\omega^{-1/2})$: the localized $\omega^{-1/4}$ is sharpened without the C^3 smoothness of Remark 1.

Proof. Repeat Steps 1–3 of Appendix A.4 with $\tau_R \equiv T$: every bound of Assumption 5 is global, so no localization is needed and every constant C_R becomes a global constant. (Under the bounded-variation form of Assumption 5(ii), the $\partial_t u$ contribution of Step 1 is read as a Stieltjes differential and bounded by the total variation V, exactly as the Stieltjes term in the proof of Lemma 4.) The one upgrade is in Step 2: the non-resonant family $E_n^{(2)}$ is estimated by Lemma 4 in place of Lemma 1. Each nonzero-frequency term of Eq. (39) has the form $\alpha\sqrt{r_i/r_n}\int_0^t \Phi(\mathbf{x}_\omega(s), s)\,\zeta_\omega(s)\,ds$ with prefactor $\alpha\sqrt{r_i/r_n} \leq \alpha\sqrt{r_{\max}}$, slow factor $\Phi \in \{\partial_i u^{c/s} g_i \cos kJ,\ \partial_i u^{c/s} g_i \sin kJ\}$, and ζ_ω a zero-mean trigonometric factor at frequency $\omega|r_n \pm r_i|$ or $2\omega r_n$, so that $c_\zeta \leq c_r$ by Lemma 3. Under Assumption 5(i)–(ii) each such Φ is bounded by $C_0^2(1 + |k|C_0)$, with $\nabla_{\mathbf{x}}\Phi$ bounded and Lipschitz and $\mathrm{TV}_t\Phi \leq CV$; Lemma 4 therefore bounds each term's mean-square supremum by $C(1 + c_r)^2/\omega$, and summing the at most $8N$ terms per coordinate, with $(\sum_{j\leq m} y_j)^2 \leq m\sum_j y_j^2$, gives

$$\mathbb{E}\sup_{t\leq T}\big\|\mathbf{E}(t)\big\|^2 \;\leq\; \frac{C}{\omega}$$

in place of Eq. (43), the family $E^{(1)}$ being $O(\omega^{-1})$ in mean-square supremum with global constants exactly as in Step 1. The Grönwall comparison of Step 3 closes the estimate: the averaged drift $\bar{\boldsymbol{\mu}} = \mathbf{f} - \frac{k\alpha}{2}\boldsymbol{\sigma}\boldsymbol{\sigma}^\top \nabla_{\mathbf{x}} J$ is bounded and globally Lipschitz (with constant $L_b \leq L(1 + |k|\alpha C_0^2)$), as is $\boldsymbol{\sigma}$, so Eq. (46) holds with $\tau_R \equiv T$, a global constant, and ω^{-1} in place of $\omega^{-1/2}$. (Both suprema have finite second moments at each fixed ω: the dither drift is bounded by $C_0\sqrt{\alpha r_{\max}\omega}$, so the standard moment bounds for bounded-coefficient Itô SDEs apply.) □

Remark 3 (Where the dither ratios enter). The dependence of C in Eq. (53) on the dither ratios enters in exactly two places: the prefactors $\alpha\sqrt{r_i/r_n} \leq \alpha\sqrt{r_{\max}}$ of the dither products, and the antiderivative bound $c_\zeta \leq c_r$ of Lemma 3, which is dominated by the reciprocal spacing $\max_{i\neq j}|r_i - r_j|^{-1}$. Packing N pairwise-distinct ratios into a fixed band therefore costs a factor $O(N)$ in c_r and hence $O(N^2)$ in Eq. (53); spreading the ratios out instead inflates $\omega r_{\max}$ and the evaluation budget at fixed step resolution. This is the quantitative counterpart of the frequency-gap discussion of Sec. 3.2, whose sharp finite-ω version is the subject of Appendix F.

B.2 $C^{1,1}$ ENERGIES: ROUGH CURVATURE WITHOUT ELLIPTICITY

Proposition 1 (**$C^{1,1}$ energies; restated**)**.** *In Assumption 5, replace the hypothesis on J by: ∇J bounded and globally Lipschitz with constant M (so $\nabla^2 J$ exists a.e. with $\|\nabla^2 J\|_{L^\infty} \leq M$ but may be discontinuous; J itself may be unbounded). Then, with no non-degeneracy assumption on $\boldsymbol{\sigma}$, the conclusion of Theorem 3 holds unchanged: $\mathbb{E}\sup_{t\leq T}\|\mathbf{x}_\omega - \bar{\mathbf{x}}\|_2^2 \leq C/\omega$.*

Proof. Set $\varepsilon = \omega^{-1/2}$ and let $J_\varepsilon = J * \rho_\varepsilon$. Because ∇J is M-Lipschitz,

$$\|J - J_\varepsilon\|_\infty \leq \|\nabla J\|_\infty \varepsilon, \quad \|\nabla J - \nabla J_\varepsilon\|_\infty \leq M\varepsilon, \quad \|\nabla^2 J_\varepsilon\|_\infty \leq M, \quad \|\nabla^3 J_\varepsilon\|_\infty \leq CM/\varepsilon. \tag{54}$$

Split each dither integral $I_{2,n}$ of Step 1 as $I_{2,n}^\varepsilon + M_{2,n}$, where $I_{2,n}^\varepsilon$ carries $\cos(\omega_n s + kJ_\varepsilon)$ and

$$\begin{aligned} M_{2,n} &= \int_0^t g_n\sqrt{\alpha\omega_n}\,\big[\cos(\omega_n s + kJ) - \cos(\omega_n s + kJ_\varepsilon)\big]\,ds \\ &= -2\int_0^t g_n\sqrt{\alpha\omega_n}\,\sin\big(\omega_n s + k\Theta_\varepsilon\big)\,\psi_\varepsilon\,ds, \end{aligned} \tag{55}$$

with $\Theta_\varepsilon := \frac{J+J_\varepsilon}{2}$ and $\psi_\varepsilon := \sin\big(\frac{k}{2}(J - J_\varepsilon)\big)$, so that $\|\psi_\varepsilon\|_\infty \le \frac{|k|}{2}\|\nabla J\|_\infty \varepsilon$ and, crucially, $\|\nabla\psi_\varepsilon\|_\infty \le \frac{|k|}{2} M\varepsilon$ by Eq. (54). (Throughout this proof, $O(\cdot)$ refers to the pathwise supremum on $[0,T]$ for finite-variation contributions and to the mean-square supremum, via Doob, for martingale contributions, as in Appendix A.4.)

The ε-corrector stream. Run Steps 1–2 of Appendix A.4 on $I^\varepsilon_{2,n}$ with correctors $u^c_\varepsilon = g_n\cos(kJ_\varepsilon)$, $u^s_\varepsilon = g_n\sin(kJ_\varepsilon)$, estimating the non-resonant dither products by Lemma 4 as in the proof of Theorem 3. Every contribution to the family $E^{(1)}_n$ is bounded exactly as in Step 1, with constants now involving $\|\nabla^2 J_\varepsilon\|_\infty \le M$ (uniform in ε), so the family remains $O(\omega^{-1})$ in mean-square supremum. In the non-resonant family the slow factors are mixed, e.g. $\Phi = \partial_i u^c_\varepsilon\, g_i\cos(kJ)$: decompose $\cos(kJ) = \cos(kJ_\varepsilon) + [\cos kJ - \cos kJ_\varepsilon]$. The all-$\varepsilon$ part is handled by Lemma 4, whose Itô-correction contribution carries $\mathrm{Lip}(\nabla\Phi^\varepsilon) \le C(M)(1 + \|\nabla^3 J_\varepsilon\|_\infty) \le CM/\varepsilon$ and is therefore $O\big(c_r/(\varepsilon\omega)\big) = O(\omega^{-1/2})$, all other contributions being $O(\omega^{-1/2})$ with ε-uniform constants; the difference part has amplitude $\le |k|\,\|\nabla J\|_\infty\varepsilon$ against $O(1)$ integrands, contributing $O(\varepsilon) = O(\omega^{-1/2})$ pathwise. The surviving resonant means are computed as in Step 2 but with mixed phases: the same trigonometric algebra gives

$$-\tfrac{k\alpha}{2}\, g_n^2\,\partial_n J_\varepsilon(\mathbf{x}_\omega)\,\cos\big(k(J-J_\varepsilon)\big) + \tfrac{\alpha}{2}\, g_n\,\partial_n g_n\,\sin\big(k(J-J_\varepsilon)\big) \;=\; -\tfrac{k\alpha}{2}\, g_n^2\,\partial_n J(\mathbf{x}_\omega) + O(\varepsilon) \quad (56)$$

pathwise, using $|\partial_n J_\varepsilon - \partial_n J| \le M\varepsilon$, $\cos(k(J-J_\varepsilon)) = 1 + O(k^2\varepsilon^2)$, and $\sin(k(J-J_\varepsilon)) = O(|k|\varepsilon)$; the $\partial_n g_n$ cross terms, which cancelled identically in Step 2, are now merely $O(\varepsilon)$.

The mismatch stream. Expand $\sin(\omega_n s + k\Theta_\varepsilon) = \sin(\omega_n s)\cos(k\Theta_\varepsilon) + \cos(\omega_n s)\sin(k\Theta_\varepsilon)$ in Eq. (55) and integrate each branch by parts against the antiderivative of the fast trigonometric factor, exactly as in the proof of Lemma 4, with slow factor $\tilde u = g_n\psi_\varepsilon\,\mathrm{trig}(k\Theta_\varepsilon)$. The key point of $C^{1,1}$ regularity is that $\nabla\tilde u$ is a sum of products of *bounded Lipschitz* functions (each of $\psi_\varepsilon, \nabla\psi_\varepsilon, \nabla\Theta_\varepsilon, g_n, \nabla g_n$ and the trig factors is bounded with a Lipschitz constant $\le C(M, |k|, C_0, L)$ *independent of* ε), so $\mathrm{Lip}(\nabla\tilde u) \le C$ and the mollification device inside Lemma 4 applies with ε-uniform constants. The resulting terms are: boundary $\le C\varepsilon\,\omega^{-1/2}$; Stieltjes and Itô-correction $\le C\omega^{-1/2}$; the drift term, which is the dominant one,

$$\le\; \sqrt{\alpha\omega_n}\cdot\frac{C}{\omega_n}\cdot\|\nabla\tilde u\|_\infty\cdot C_0\sqrt{N}\big(1+\sqrt{\alpha r_{\max}\omega}\big)\,T \;\le\; C\,\|\nabla\tilde u\|_\infty \;\le\; C\varepsilon,$$

using $\|\nabla\tilde u\|_\infty \le C\big(\|\nabla\psi_\varepsilon\|_\infty + |k|\,\|\psi_\varepsilon\|_\infty\|\nabla\Theta_\varepsilon\|_\infty + \|\psi_\varepsilon\|_\infty\big) \le C\varepsilon$; and a martingale with mean-square supremum $\le C\varepsilon^2\omega^{-1}$.

Collecting both streams, the remainder $\mathbf{E}$ in the exact representation Eq. (43) (now derived with $\partial_n J$ via Eq. (56), and with $\tau_R \equiv T$) obeys $\mathbb{E}\sup_{t\le T}\|\mathbf{E}\|^2 \le C(\omega^{-1} + \varepsilon^2) = C\omega^{-1}$. The averaged drift $\mathbf{f} - \frac{k\alpha}{2}\boldsymbol{\sigma}\boldsymbol{\sigma}^\top\nabla J$ is Lipschitz (∇J is), so the Grönwall comparison of Step 3 applies verbatim. No step used non-degeneracy of $\boldsymbol{\sigma}$. □

This rung already covers the marquee examples of "twice differentiable off a null set": piecewise-quadratic energies glued in C^1, max/soft-max compositions with matched gradients, and any energy whose curvature jumps across hypersurfaces while staying bounded—with the full rate of Theorem 3, and for arbitrary (possibly vanishing) diffusion amplitude $\boldsymbol{\sigma}$.

B.3 The regularizing role of the noise

When the Hessian is not merely discontinuous but unbounded ($\nabla^2 J \in L^p$ only), pathwise bounds on the mollified quantities are no longer enough and terms must be integrated along the trajectory.

Lemma 5 (Occupation density and the Krylov estimate)**.** *Let $\mathbf{x}$ solve $d\mathbf{x} = \mathbf{b}(\mathbf{x},t)\,dt + \boldsymbol{\sigma}(\mathbf{x},t)\,d\mathbf{w}$ on $[0,T]$ with $\mathbf{b}, \boldsymbol{\sigma}$ bounded and $\boldsymbol{\sigma}\boldsymbol{\sigma}^\top \succeq \lambda_0 I$ uniformly ($\lambda_0 > 0$). Then there is a constant $N_* = N_*(\lambda_0, N, p, q, T, \|\mathbf{b}\|_\infty)$ such that for every $h \in L^q\big([0,T]; L^p(\mathbb{R}^N)\big)$ with $\frac{N}{p} + \frac{2}{q} < 2$,*

$$\mathbb{E}\int_0^T \big|h(t,\mathbf{x}(t))\big|\,dt \;\le\; N_*\,\|h\|_{L^q_t L^p_{\mathbf{x}}}. \quad (57)$$

In particular the expected occupation measure $\mu(A) = \mathbb{E}\int_0^T \mathbf{1}_A(t,\mathbf{x}(t))\,dt$ is absolutely continuous with respect to Lebesgue measure on $[0,T]\times\mathbb{R}^N$. Consequently $\mathbf{x}$ charges no Lebesgue-null set, and two drifts that agree Lebesgue-a.e. produce indistinguishable dynamics.

This is the Krylov estimate for uniformly non-degenerate diffusions (Krylov, 1980; Krylov & Röckner, 2005). Along the averaged process, a drift or Hessian defined only almost everywhere is as good as a continuous one, because the elliptic noise smooths the law of $\mathbf{x}$ into an absolutely continuous measure that never sees the exceptional set. This is exactly why Langevin and SGLD samplers operate on rough energies.

Lemma 5 cannot, however, be applied as is along the ES process $\mathbf{x}_\omega$: its constant depends on the drift bound, and the dither drift grows like $\sqrt{\omega}$. That gap is closed by the following estimate, whose proof exploits the parabolic scaling of the dither: over a window of length $1/\omega$ the dither displaces the process by only $O(\omega^{-1/2})$, which is exactly the diffusive spread, so Girsanov applies on each window.

Lemma 6 (Microinterval occupation estimate)**.** *Let $\mathbf{x}_\omega$ solve Eq. (20) with $\mathbf{f}, \boldsymbol{\sigma}$ bounded and Lipschitz, ∇J bounded, and, additionally, $\boldsymbol{\sigma}\boldsymbol{\sigma}^\top \succeq \lambda_0 I$ uniformly. Let $p > \max(N, 2)$ and let $h \geq 0$ be measurable with $M_h := \sup_{t \leq T} \|h(t, \cdot)\|_{L^p_{\mathrm{uloc}}} < \infty$. Then for all $\omega \geq 1$,*

$$\mathbb{E} \int_0^T h\big(t, \mathbf{x}_\omega(t)\big)\, dt \;\leq\; C\big(1 + \omega^{N/(2p)}\big)\, M_h, \tag{58}$$

with $C = C(N, p, \lambda_0, T, C_0, L, \alpha, r_{\max})$ independent of ω. Note $N/(2p) < \frac{1}{2}$ if and only if $p > N$: the dither prefactor $\omega^{-1/2}$ of the averaging estimates beats the occupation growth exactly at the Morrey threshold.

Proof. Partition $[0, T]$ into $\lceil T\omega \rceil$ intervals $I_j = [t_j, t_{j+1}]$ of length $h_0 \leq 1/\omega$ and condition on $\mathcal{F}_{t_j}$. Write $\boldsymbol{\theta} := \boldsymbol{\sigma}^{-1}\boldsymbol{\mu}_\omega$, so $\|\boldsymbol{\theta}\|^2 \leq \lambda_0^{-1} N C_0^2 \big(1 + \sqrt{\alpha r_{\max}\omega}\big)^2$ and hence

$$\tfrac{1}{2} \int_{I_j} \|\boldsymbol{\theta}\|^2\, ds \;\leq\; \tfrac{1}{2}\lambda_0^{-1} N C_0^2 \big(1 + \sqrt{\alpha r_{\max}\omega}\big)^2 h_0 \;\leq\; \Lambda_0 := C\lambda_0^{-1} N C_0^2 \alpha r_{\max} \qquad \text{for } \omega \geq 1,$$

uniformly in ω: this is the parabolic scaling $K^2 h_0 = O(1)$ with $K \asymp \sqrt{\omega}$, $h_0 = 1/\omega$. Novikov's condition therefore holds conditionally on each I_j, and Girsanov's theorem defines $\mathbb{Q}_j \sim \mathbb{P}$ on $\mathcal{F}_{t_{j+1}}$ under which $\mathbf{x}$ solves the driftless non-degenerate SDE $d\mathbf{x} = \boldsymbol{\sigma}(\mathbf{x}, s)\, d\tilde{\mathbf{w}}$ on I_j from $\mathbf{x}_{t_j}$. With $Z_j := d\mathbb{P}/d\mathbb{Q}_j$ one has $\mathbb{E}^{\mathbb{Q}_j}[Z_j^2 \,|\, \mathcal{F}_{t_j}] \leq e^{2\Lambda_0}$ (bounded $\boldsymbol{\theta}$ makes the Doléans exponential a true martingale, and the standard second-moment bound applies). By conditional Cauchy–Schwarz and then Cauchy–Schwarz in s,

$$\begin{aligned}
\mathbb{E}\Big[\int_{I_j} h(s, \mathbf{x}_s)\, ds \,\Big|\, \mathcal{F}_{t_j}\Big] \;&\leq\; e^{\Lambda_0}\Big(\mathbb{E}^{\mathbb{Q}_j}\Big[\Big(\int_{I_j} h\, ds\Big)^2 \Big| \mathcal{F}_{t_j}\Big]\Big)^{1/2} \\
&\leq\; e^{\Lambda_0} h_0^{1/2}\Big(\mathbb{E}^{\mathbb{Q}_j}\Big[\int_{I_j} h^2\, ds \Big| \mathcal{F}_{t_j}\Big]\Big)^{1/2}.
\end{aligned}$$

Under $\mathbb{Q}_j$ the process is a uniformly elliptic diffusion with bounded Lipschitz coefficients, whose transition density over the elapsed time $s - t_j$ from $\mathbf{x}_{t_j}$ obeys the classical Gaussian upper bound $p_{s-t_j}(\mathbf{x}_{t_j}, \mathbf{y}) \leq C_A (s - t_j)^{-N/2} e^{-c\|\mathbf{y} - \mathbf{x}_{t_j}\|^2/(s-t_j)}$ (Friedman, 1964). Hence, with $r = p/2 > 1$ and r' its conjugate, $\|p_{s-t_j}(\mathbf{x}_{t_j}, \cdot)\|_{L^{r'}(B_1(\mathbf{y}_0))} \leq C(s - t_j)^{-N/p} e^{-c\,\mathrm{dist}(\mathbf{x}_{t_j}, B_1(\mathbf{y}_0))^2/(s-t_j)}$, and summing over a cover of $\mathbb{R}^N$ by unit balls (the Gaussian tails dominate the polynomially many relevant balls),

$$\mathbb{E}^{\mathbb{Q}_j}\big[h^2(s, \mathbf{x}_s) \,\big|\, \mathcal{F}_{t_j}\big] \leq C\,(s - t_j)^{-N/p} M_h^2, \qquad \mathbb{E}^{\mathbb{Q}_j}\Big[\int_{I_j} h^2\, ds \,\Big|\, \mathcal{F}_{t_j}\Big] \leq C\, M_h^2\, h_0^{1 - N/p},$$

the time integral converging precisely because $N/p < 1$. Combining, each interval contributes at most $C M_h\, h_0^{1 - N/(2p)}$, and summing over $\lceil T\omega \rceil$ intervals gives $\mathbb{E}\int_0^T h(s, \mathbf{x}_\omega(s))\, ds \leq C M_h T\, h_0^{-N/(2p)} \leq C(1 + \omega^{N/(2p)}) M_h$. □

B.4 Sobolev energies: proof of Theorem 2

Theorem 2 (Sobolev energies; restated). *Assume $\boldsymbol{\sigma}\boldsymbol{\sigma}^\top \succeq \lambda_0 I$ uniformly with $\mathbf{f}, \boldsymbol{\sigma}$ bounded and Lipschitz, and that $J(\cdot, t) \in W^{2,p}_{\mathrm{uloc}}(\mathbb{R}^N)$ uniformly in t for some $p > \max(N, 2)$, with ∇J and $\partial_t J$*

bounded; write $M_p := \sup_t \|\nabla^2 J(\cdot,t)\|_{L^p_{\mathrm{uloc}}}$. *Then* $\mathbf{x}_\omega \to \bar{\mathbf{x}}$ *in probability, uniformly on* $[0,T]$, *with* $\mathbb{E}\sup_{t\le T}\|\mathbf{x}_\omega - \bar{\mathbf{x}}\|_2^2 \le C\,\omega^{-(p-N)/p}$, *where* $\bar{\mathbf{x}}$ *is the unique strong solution of the averaged SDE Eq. (22), whose drift is continuous* ($\nabla J \in W^{1,p}_{\mathrm{uloc}} \hookrightarrow C^{0,1-N/p}$ *by Morrey); the a.e.-defined Hessian* $\nabla^2 J$ *enters only the proof, along the trajectories.*

Proof. (0). The averaged drift $\mathbf{f} - \frac{k\alpha}{2}\boldsymbol{\sigma}\boldsymbol{\sigma}^\top \nabla J$ is bounded and measurable (∇J is bounded by hypothesis and continuous by Morrey's embedding, $p > N$), so strong existence and uniqueness for Eq. (22) follow from Veretennikov's theorem (Veretennikov, 1981); by Lemma 5, the a.e.-defined $\nabla^2 J$ never enters ambiguously along $\bar{\mathbf{x}}$.

(i). Uniformly-local Morrey gives the global modulus

$$\|\nabla J - \nabla J_\varepsilon\|_{L^\infty} \le C(N,p)\,M_p\,\varepsilon^{1-N/p}, \qquad \|\nabla^2 J_\varepsilon\|_{L^\infty} \le C M_p\,\varepsilon^{-N/p},$$
$$\|\nabla^3 J_\varepsilon\|_{L^\infty} \le C M_p\,\varepsilon^{-1-N/p}, \tag{59}$$

together with $\|J - J_\varepsilon\|_\infty \le \|\nabla J\|_\infty \varepsilon$ and $\|\nabla^2 J_\varepsilon\|_{L^p_{\mathrm{uloc}}} \le C M_p$. (The first bound follows from $|\nabla J(\mathbf{x}) - \nabla J(\mathbf{y})| \le C\|\nabla^2 J\|_{L^p(B(\mathbf{x},2\|\mathbf{x}-\mathbf{y}\|))}\|\mathbf{x}-\mathbf{y}\|^{1-N/p}$ for $\|\mathbf{x}-\mathbf{y}\| \le 1$.)

(ii). As in Proposition 1, split $I_{2,n} = I^\varepsilon_{2,n} + M_{2,n}$ and run Steps 1–2 of Appendix A.4 on $I^\varepsilon_{2,n}$ with correctors built from J_ε, which is smooth with the bounds Eq. (59). All estimates are now pathwise in the mollified quantities: the boundary, Stieltjes ($\partial_t u$), $\mathbf{f}$-drift, and stochastic-integral constituents of $E^{(1)}_n$ are $O(\omega^{-1/2})$ with ε-uniform constants; the Hessian (Itô-correction) constituent is $O(\omega^{-1/2}\varepsilon^{-N/p})$; the non-resonant family $E^{(2)}_n$, after the decomposition $\cos kJ = \cos kJ_\varepsilon + [\cos kJ - \cos kJ_\varepsilon]$ of its mixed slow factors, contributes $O\big(\omega^{-1/2}\varepsilon^{-N/p} + c_r\,\omega^{-1}\varepsilon^{-1-N/p}\big)$ from the all-ε parts via Lemma 4 (whose Itô-correction term now carries $\mathrm{Lip}(\nabla\Phi^\varepsilon) \lesssim \varepsilon^{-1-N/p}$) and $O(\varepsilon)$ from the small difference parts. The surviving resonant means give, exactly as in Eq. (56),

$$-\tfrac{k\alpha}{2}\,g_n^2\,\partial_n J_\varepsilon(\mathbf{x}_\omega)\cos\big(k(J-J_\varepsilon)\big) + O(\varepsilon) \;=\; -\tfrac{k\alpha}{2}\,g_n^2\,\partial_n J(\mathbf{x}_\omega) + O\big(\varepsilon + M_p\varepsilon^{1-N/p}\big)$$

pathwise, the replacement $\partial J_\varepsilon \to \partial J$ costing $M_p\varepsilon^{1-N/p}$ globally by Eq. (59)—no occupation estimate is needed for this step.

(iii). Write $M_{2,n}$ as in Eq. (55). Now $\|\psi_\varepsilon\|_\infty \le C|k|\varepsilon$ and $\|\nabla\psi_\varepsilon\|_\infty \le C|k|M_p\varepsilon^{1-N/p}$ by Eq. (59), but $\nabla\tilde{u}$ is no longer Lipschitz: its weak gradient contains $\nabla^2\Theta_\varepsilon \ni \nabla^2 J \in L^p_{\mathrm{uloc}}$ only. Integrate by parts against the fast antiderivative as before; the Itô expansion of $\tilde{u}$ along $\mathbf{x}_\omega$ is justified, for each fixed ω, by the Itô–Krylov formula (Krylov, 1980, Ch. 2) ($\tilde{u}(\cdot,t) \in W^{2,p}_{\mathrm{uloc}}$ by the chain rule, $\partial_t\tilde{u}$ bounded, diffusion non-degenerate, drift bounded for fixed ω). The resulting terms: boundary $O(\varepsilon\omega^{-1/2})$; Stieltjes $O(\omega^{-1/2})$; the drift term $O(\|\nabla\tilde{u}\|_\infty) = O\big(\varepsilon + M_p\varepsilon^{1-N/p}\big)$; the martingale term $O(\varepsilon\omega^{-1/2})$ in mean-square supremum; and the second-order term, the only one requiring the occupation estimate,

$$\sqrt{\alpha\omega_n}\cdot\frac{C}{\omega_n}\,\mathbb{E}\int_0^T \big|\operatorname{tr}\big(\boldsymbol{\sigma}\boldsymbol{\sigma}^\top\nabla^2\tilde{u}\big)\big|\big(\mathbf{x}_\omega(s),s\big)\,ds$$
$$\le\; C\,\omega^{-1/2}\,\mathbb{E}\int_0^T \big(1 + |\nabla^2 J| + |\nabla^2 J_\varepsilon|\big)\big(\mathbf{x}_\omega(s),s\big)\,ds \;\le\; C\,M_p\,\omega^{-1/2+N/(2p)} \tag{60}$$

by Lemma 6 (mollification does not increase $\|\cdot\|_{L^p_{\mathrm{uloc}}}$), vanishing precisely because $p > N$.

(iv). Choosing $\varepsilon = \omega^{-1/2}$, every pathwise contribution above is $O\big(\omega^{-(p-N)/(2p)}\big)$ (the slowest terms being $\omega^{-1/2}\varepsilon^{-N/p}$, $\omega^{-1}\varepsilon^{-1-N/p}$, $M_p\varepsilon^{1-N/p}$, and $\omega^{-1/2+N/(2p)}$, all equal to $\omega^{-(p-N)/(2p)}$ at this choice), and every martingale contribution is at least as small in mean-square supremum. Hence the representation Eq. (43) holds, with $\tau_R \equiv T$, with $\bar{\boldsymbol{\mu}}$ containing the a.e.-defined ∇J, and with $\mathbb{E}\sup_{t\le T}\|\mathbf{E}\|^2 \le C\omega^{-(p-N)/p}$. The averaged drift is bounded and, by uniformly-local Morrey ($W^{1,p}_{\mathrm{uloc}} \hookrightarrow C^{0,1-N/p}$, $p > N$, applied to ∇J), globally $(1-N/p)$-Hölder in $\mathbf{x}$ uniformly in t—but not necessarily Lipschitz—so the Grönwall step of Step 3 is replaced by a Zvonkin-transformed comparison (Zvonkin, 1974; Veretennikov, 1981). The Hölder modulus is what makes the transformation quantitatively usable here. Because the coefficients depend on t, the transformation is built from the *backward parabolic* Cauchy problem rather than an elliptic resolvent (Zvonkin,

1974; Veretennikov, 1981): for ϱ large, let $\mathbf{u}$ solve $\partial_t \mathbf{u} + \bar{\mathcal{L}}_t \mathbf{u} - \varrho\,\mathbf{u} = -\bar{\boldsymbol{\mu}}$ on $[0,T]\times\mathbb{R}^N$ with $\mathbf{u}(\cdot,T)=\mathbf{0}$, where $\bar{\mathcal{L}}_t$ is the averaged generator at time t. Its drift $\bar{\boldsymbol{\mu}}$ is jointly Hölder: $(1-N/p)$-Hölder in $\mathbf{x}$ by Morrey, and Hölder in t with exponent at least $(1-N/p)/2$ by interpolation between the boundedness of $\partial_t J$ and the spatial modulus (mollify J at scale ε and optimize: $|\nabla J(\mathbf{x},t) - \nabla J(\mathbf{x},s)| \le C\varepsilon^{1-N/p} + C|t-s|/\varepsilon$). Parabolic Schauder theory (Friedman, 1964) therefore gives $\mathbf{u} \in C^{2+\beta,1+\beta/2}$ with $\|\nabla\mathbf{u}\|_\infty \le \frac{1}{2}$ once ϱ is large, so $\Phi(\cdot,t) := \mathrm{id} + \mathbf{u}(\cdot,t)$ is, for each t, a global bi-Lipschitz diffeomorphism with $\nabla\Phi$, $\nabla^2\Phi$ bounded—a conclusion that would fail for merely bounded measurable drift, where Calderón–Zygmund theory yields only $\mathbf{u} \in W^{2,r}_{\mathrm{uloc}}$—and by Itô's formula the time derivative $\partial_t\mathbf{u}$ is exactly compensated: $d\Phi(\bar{\mathbf{x}},t) = \varrho\,\mathbf{u}(\bar{\mathbf{x}},t)\,dt + (I+\nabla\mathbf{u})\boldsymbol{\sigma}(\bar{\mathbf{x}},t)\,d\mathbf{w}$, so that in the coordinates $\mathbf{y} = \Phi(\mathbf{x},t)$ the averaged equation has Lipschitz drift $\varrho\,\mathbf{u}\circ\Phi^{-1}$ and Lipschitz diffusion $(I+\nabla\mathbf{u})\boldsymbol{\sigma}\circ\Phi^{-1}$. (An elliptic resolvent $\varrho\mathbf{u} - \bar{\mathcal{L}}_t\mathbf{u} = \bar{\boldsymbol{\mu}}$ solved separately for each t would leave an uncompensated $\partial_t\mathbf{u}\,dt$ in this expansion, controlled only if $\nabla\partial_t J$ were bounded, which is not assumed; for autonomous J the two formulations coincide.) Both processes are continuous semimartingales, $\mathbf{x}_\omega$ by Eq. (20) and, equivalently through the representation Eq. (43), as the averaged dynamics along its own trajectory plus $\mathbf{E}$, where $\mathbf{E}$ collects the finite-variation families and the stochastic-integral family with integrands of size $O(\omega^{-1/2})$. Itô's formula applied to $\Phi(\mathbf{x}_\omega)$ and $\Phi(\bar{\mathbf{x}})$ therefore gives, for the difference, the transformed Lipschitz comparison plus remainders of two kinds: $\nabla\Phi$ acting on $d\mathbf{E}$, controlled by $\|\nabla\Phi\|_\infty^2\,\mathbb{E}\sup\|\mathbf{E}\|^2$; and second-order cross-variation corrections, in which only the martingale part of $\mathbf{E}$ participates: its quadratic variation has integrands of size $O(\omega^{-1})$, and its cross-variation with $\boldsymbol{\sigma}\,d\mathbf{w}$ has integrands of size $O(\omega^{-1/2})$ which oscillate at the dither frequencies, so that their time integrals are $O(\omega^{-1})$ by Lemma 4; even without exploiting the oscillation these corrections are bounded by $\|\nabla^2\Phi\|_\infty\,O(\omega^{-1/2})$, which is already smaller than the target rate $\omega^{-(p-N)/(2p)}$. Grönwall–Burkholder–Davis–Gundy in the $\mathbf{y}$-coordinates, followed by the bi-Lipschitz transfer back through Φ^{-1}, yields $\mathbb{E}\sup_{t\le T}\|\mathbf{x}_\omega - \bar{\mathbf{x}}\|^2 \le C\,\omega^{-(p-N)/p}$. Convergence in probability follows by Chebyshev. □

Theorem 2 covers energies whose Hessian is unbounded but p-integrable, $p > N$: energies with isolated curvature blow-ups of order $\|\mathbf{x}-\mathbf{x}_0\|^{-\gamma}$, $\gamma < N/p$, and, more generally, any J whose only failure of C^2 occurs on a null set across which ∇J stays continuous. The rate degrades continuously to zero as $p \downarrow N$, the Morrey threshold at which ∇J can first fail to be continuous.

B.5 Lipschitz energies and gradient kinks

When the gradient itself jumps—kinks, as in $\|\cdot\|_1$ or ReLU-type energies, and the mean-absolute-difference measurement terms of the tracking energy Eq. (14)—J is no longer in $W^{2,p}$ ($\nabla^2 J$ acquires a singular surface part), and neither proof above applies: $\|\nabla J - \nabla J_\varepsilon\|_\infty$ no longer vanishes, and $\|\nabla^2 J_\varepsilon\|_{L^p_{\mathrm{uloc}}} \asymp \varepsilon^{1/p-1}$ blows up. Well-posedness of the limit nevertheless survives in full, and we isolate exactly what the convergence proof still needs—an occupation estimate for the dithered process that is uniform in ω—which Appendix B.6 then proves.

Proposition 3 (Lipschitz energies). *Assume $\boldsymbol{\sigma}\boldsymbol{\sigma}^\top \succeq \lambda_0 I$ uniformly with $\mathbf{f}, \boldsymbol{\sigma}$ bounded and Lipschitz, and that $J(\cdot,t)$ is globally Lipschitz uniformly in t (J need not be bounded). By Rademacher's theorem (Evans & Gariepy, 1992), ∇J exists Lebesgue-a.e. with $\|\nabla J\|_{L^\infty} \le L_J$. Then: (a) the averaged SDE Eq. (22) has a unique strong solution (Veretennikov, 1981; Zvonkin, 1974), and by Lemma 5 it does not depend on the values assigned to ∇J on its null exceptional set; (b) if, in addition, Conjecture 1 below holds—by Theorem 4 this is the case whenever J is also Lipschitz in t uniformly in $\mathbf{x}$ and $\mathbf{f}, \boldsymbol{\sigma}$ satisfy hypotheses (H1)–(H2) of Appendix B.6—then $\mathbf{x}_\omega \Rightarrow \bar{\mathbf{x}}$ as $\omega \to \infty$.*

Conjecture 1 (ω-uniform occupation estimate). *Under the hypotheses of Proposition 3, the estimate Eq. (58) holds with exponent zero: $\mathbb{E}\int_0^T h(t,\mathbf{x}_\omega(t))\,dt \le C\,M_h$ uniformly in $\omega \ge 1$, for $h \in L^\infty_t L^p_{\mathrm{uloc}}$, $p > \max(N,2)$.*

Conjecture 1 asserts that the occupation measure of the dithered process is no rougher, uniformly in ω, than that of a bounded-drift diffusion. It is what the averaging limit itself predicts (the occupation measure of $\mathbf{x}_\omega$ converges to that of $\bar{\mathbf{x}}$, which satisfies Lemma 5), and it holds with the harmless $\omega^{N/(2p)}$ correction by Lemma 6. Removing the correction without presupposing the averaging one is trying to prove requires taking the dither out of the drift *before* Girsanov's theorem is applied, which Appendix B.6 does by a pathwise near-identity change of variables: Theorem 4 proves Conjecture 1

whenever J is Lipschitz in t (the time regularity that Theorem 2 also carries) and $\partial_t g_n$ is bounded, and Corollary 5 supplies the ω-uniform tightness that the localization below uses. Only the case of merely bounded-variation time dependence remains open, and only because of the time regularity, not because of the kinks (Remark 6(i)); we keep the statement here, in the form in which the two-scale argument consumes it.

Proof sketch of Proposition 3(b). Mollify: $J_\varepsilon = J * \rho_\varepsilon$ with $\|\nabla J_\varepsilon\|_\infty \le L_J$, $\nabla J_\varepsilon \to \nabla J$ Lebesgue-a.e., $\|\nabla^2 J_\varepsilon\|_\infty \le C L_J/\varepsilon$, and run the two-stream architecture of Proposition 1 / Theorem 2. At no point is the crude bound $\sqrt{\omega}\,\|J - J_\varepsilon\|_\infty$ used: the dither mismatch Eq. (55) is itself a zero-mean oscillation and is integrated by parts, so the mismatch enters only through (i) remainders carrying ε-powers of the mollified norms, all of the form $\omega^{-1/2}\varepsilon^{-\gamma}$ ($\gamma \le 2$) and vanishing along, e.g., $\varepsilon = \omega^{-1/8}$; and (ii) drift-replacement and slow-factor terms proportional to $|\nabla J - \nabla J_\varepsilon|(\mathbf{x}_\omega(s), s)$ integrated in time. For (ii), $h_\varepsilon := |\nabla J - \nabla J_\varepsilon|$ satisfies $h_\varepsilon \le 2L_J$ and $h_\varepsilon \to 0$ a.e. in $(\mathbf{x}, t)$; by Theorem 4, whose constant does not depend on ω, $\mathbb{E}\int_0^T h_\varepsilon(\mathbf{x}_\omega)\,ds \to 0$ as $\varepsilon \to 0$ uniformly in ω (localize to a ball around $\mathbf{x}_0$, where dominated convergence applies to the mixed norm $\|h_\varepsilon \mathbf{1}_{B_R}\|_{p,q}$ with $q < \infty$, and control the complement by the ω-uniform tightness of Corollary 5; the details are in Remark 5), so both limits $\varepsilon \to 0, \omega \to \infty$ may be taken along the single sequence $\varepsilon = \omega^{-1/8}$. The identified limit drift is $-\frac{k\alpha}{2}\boldsymbol{\sigma}\boldsymbol{\sigma}^\top \nabla J$ with the a.e. gradient. The argument then closes at the level of laws rather than paths, because the limit equation has only bounded measurable drift, so that no pathwise Lipschitz comparison of the kind used in step (iv) of Theorem 2 is available (for such drift the Zvonkin map satisfies only $\mathbf{u} \in W^{2,r}_{\mathrm{uloc}}$ for every $r < \infty$ (Veretennikov, 1981; Krylov & Röckner, 2005), so $\nabla\Phi$ is merely Hölder and a Grönwall–BDG comparison under a shared Brownian motion does not close in $N \ge 2$ dimensions). Instead: (1) *tightness*: under the hypotheses (H1)–(H4) of Appendix B.6, which are those under which Conjecture 1 is established, $\mathbf{y} = \Psi(\mathbf{x}_\omega, t)$ of Eq. (70) is an Itô process whose drift and diffusion are bounded uniformly in ω, so its laws are tight on $C([0,T];\mathbb{R}^N)$ by the Kolmogorov–Chentsov moment criterion (Billingsley, 1999), and $\sup_t \|\mathbf{x}_\omega - \mathbf{y}\| \le C\omega^{-1/2}$ transfers tightness to $\{\mathrm{Law}(\mathbf{x}_\omega)\}_{\omega \ge \omega_0}$ (Corollary 5 is the quantitative form); (2) *identification*: for $\varphi \in C_c^\infty(\mathbb{R}^N)$, the two-stream expansion applied to $\varphi(\mathbf{x}_\omega)$ in place of $\mathbf{x}_\omega$—with slow factors $\partial_i\varphi\, g_i \cos kJ_\varepsilon$ and $\partial_i\varphi\, g_i \sin kJ_\varepsilon$—shows that $\varphi(\mathbf{x}_\omega(t)) - \int_0^t \bar{\mathcal{L}}_s\varphi(\mathbf{x}_\omega(s))\,ds$ is a martingale up to an error vanishing in L^1 as $\omega \to \infty$, the oscillatory terms by Lemma 4 and (i), the drift-replacement terms by (ii); hence every subsequential weak limit solves the martingale problem for the averaged generator $\bar{\mathcal{L}}$ with drift $-\frac{k\alpha}{2}\boldsymbol{\sigma}\boldsymbol{\sigma}^\top\nabla J$; (3) *uniqueness*: the martingale problem for a uniformly elliptic diffusion with bounded measurable drift and Lipschitz diffusion coefficient is well posed (Stroock & Varadhan, 1979; Veretennikov, 1981), so every subsequential limit is $\mathrm{Law}(\bar{\mathbf{x}})$, which proves $\mathbf{x}_\omega \Rightarrow \bar{\mathbf{x}}$. The occupation estimate thus enters the Lipschitz rung through step (2) alone, in the drift-replacement terms and in the mollified-corrector remainders; Theorem 4 supplies it with any $p > \max(N, 2)$ and any finite $q > 2p/(p-N)$. (With Lemma 6 in place of the conjecture, the argument delivers everything except the uniformity of step (ii), whose bound degrades to $C\omega^{N/(2p)}\|h_\varepsilon\|$; since $\|h_\varepsilon\|$ carries no rate for merely Lipschitz J, the two limits cannot then be coupled.) Equivalently, one may read $\bar{\mathbf{x}}$ as a Filippov solution of the differential inclusion $d\bar{\mathbf{x}} \in \big[\mathbf{f} - \frac{k\alpha}{2}\boldsymbol{\sigma}\boldsymbol{\sigma}^\top \partial_C J\big]\,dt + \boldsymbol{\sigma}\,d\mathbf{w}$ with $\partial_C J$ the Clarke gradient (Clarke, 1983; Filippov, 1988); the non-degenerate noise collapses the inclusion to the a.e.-gradient SDE. □

Remark 4 (Why Lemma 6 does not suffice, even for structured kinks). One might hope to bypass Conjecture 1 for the practically relevant case in which the kink set of J is a fixed $(N-1)$-rectifiable hypersurface of finite surface measure: there $h_\varepsilon = |\nabla J - \nabla J_\varepsilon|$ is supported in an ε-collar of the surface, so $\|h_\varepsilon\|_{L^\infty_t L^p_{\mathrm{uloc}}} \le C L_J \varepsilon^{1/p}$ carries a genuine rate. Feeding this into Lemma 6 bounds the step-(ii) terms by $C\omega^{N/(2p)}\varepsilon^{1/p}$, which vanishes only if $\varepsilon \ll \omega^{-N/2}$. But the mollified-corrector remainders of step (i) contain $\omega^{-1/2}\varepsilon^{-1}$ and $\omega^{-1}\varepsilon^{-2}$ and hence require $\varepsilon \gg \omega^{-1/2}$. The two windows $\varepsilon \ll \omega^{-N/2}$ and $\varepsilon \gg \omega^{-1/2}$ are disjoint for every $N \ge 1$. The $\omega^{N/(2p)}$ loss in Lemma 6 is therefore not a slack constant one can trade against geometric structure in J; removing it—Conjecture 1—is the genuine obstruction. Appendix B.6 removes it by changing the order of operations rather than the scale: the dither is taken out of the drift by a pathwise change of variables first, at its own mollification scale $\varepsilon = \omega^{-1/2}$, and Girsanov's theorem is applied to what remains, whose drift is bounded uniformly in ω.

B.6 THE ω-UNIFORM OCCUPATION ESTIMATE: PROOF OF CONJECTURE 1

Lemma 6 loses the factor $\omega^{N/(2p)}$ because it applies Girsanov's theorem to the dither drift itself: that drift has size $\sqrt{\omega}$, so Novikov's exponent stays bounded only on windows of length $1/\omega$, and $\lceil T\omega \rceil$ windows have to be summed. This subsection removes the loss by removing the dither first. The device is the near-identity change of variables of the deterministic warm-up, Eq. (23)—the classical averaging transformation (Bogoliubov & Mitropolsky, 1961; Sanders et al., 2007), in its stochastic form (Freidlin & Wentzell, 2012; Pavliotis & Stuart, 2008)—applied pathwise to the Itô process: subtracting the phase-locked ripple $g_n\sqrt{\alpha/\omega_n}\,\sin(\omega_n t + kJ)$ from each coordinate produces a process whose drift is bounded uniformly in ω and whose diffusion coefficient is an $O(\omega^{-1/2})$ perturbation of $\boldsymbol{\sigma}$. The classical occupation estimate then applies to the transformed process with ω-uniform constants, and the transformation is bi-Lipschitz with ω-uniform constants, so the estimate transfers back to $\mathbf{x}_\omega$. What makes this work where Remark 4 found an obstruction is a difference of purpose. An occupation estimate needs the transformed coefficients to be *bounded*, not small, so the mollification scale inside the transformation can be tied to the dither scale, $\varepsilon = \omega^{-1/2}$, exactly where the dither mismatch and the Itô-correction term are both $O(1)$. A convergence proof needs those same terms to *vanish*, which forces $\varepsilon \gg \omega^{-1/2}$. The two requirements are incompatible for a single scale, which is the disjoint-windows obstruction of Remark 4; the resolution is that the occupation estimate carries its own scale.

Throughout this subsection we assume:

- (H1) $\mathbf{f}$ is bounded and Lipschitz in $\mathbf{x}$; each g_n is twice continuously differentiable in $\mathbf{x}$, with g_n and its first two $\mathbf{x}$-derivatives bounded by C_0 and Lipschitz in $\mathbf{x}$ with constant L, and continuously differentiable in t with $|\partial_t g_n| \le C_0$ (this is Assumption 5(i)–(ii) for $\mathbf{f}, \boldsymbol{\sigma}$, with (ii) in its continuously-differentiable form);
- (H2) $\boldsymbol{\sigma}\boldsymbol{\sigma}^\top \succeq \lambda_0 I$ for some $\lambda_0 > 0$;
- (H3) J is globally Lipschitz in $\mathbf{x}$ uniformly in t, and in t uniformly in $\mathbf{x}$: $\|\nabla J\|_{L^\infty} \le L_J$ and $\|\partial_t J\|_{L^\infty} \le L_t$ (the derivatives exist a.e. by Rademacher's theorem and are the weak derivatives);
- (H4) Assumption 2, normalized so that $\min_n r_n = 1$ (Assumption 5(iii)).

(H3) is the hypothesis of Proposition 3 together with the Lipschitz-in-time regularity that Theorem 2 also carries; Remark 6 explains why merely bounded variation in t does not suffice for this argument. Constants C below depend only on $(N, p, q, \lambda_0, T, C_0, L, \alpha, |k|, r_{\max}, L_J, L_t)$ and never on ω or on h.

Theorem 4 (ω-uniform occupation estimate)**.** *Let (H1)–(H4) hold and let $\mathbf{x}_\omega$ solve Eq. (20) from an $\mathcal{F}_0$-measurable $\mathbf{x}_0$. Let $p > \max(N, 2)$ and $q \in \left(\frac{2p}{p-N}, \infty\right]$, and for measurable $h \ge 0$ on $[0,T]\times\mathbb{R}^N$ write*

$$\|h\|_{p,q} \;:=\; \Big\| \, \|h(t,\cdot)\|_{L^p_{\mathrm{uloc}}} \Big\|_{L^q(0,T)}, \qquad \textit{so that } \|h\|_{p,\infty} = M_h \textit{ in the notation of Lemma 6.}$$

Then there is a constant C, independent of ω and of h, such that for all $\omega \ge 1$,

$$\mathbb{E}\int_0^T h\big(t, \mathbf{x}_\omega(t)\big)\, dt \;\le\; C\,\|h\|_{p,q}. \tag{61}$$

In particular, Conjecture 1 holds under the hypotheses of Proposition 3 strengthened by (H1) and (H3).

Proof. Step 1: mollification at the dither scale. Fix $\omega \ge 1$ and set $\varepsilon := \omega^{-1/2} \le 1$. Let $\rho \in C_c^\infty(B_1)$ be a nonnegative mollifier with $\int \rho = 1$, $\rho_\varepsilon(\mathbf{z}) := \varepsilon^{-N}\rho(\mathbf{z}/\varepsilon)$, $c_\rho := \max_j \|\partial_j \rho\|_{L^1}$, and $J_\varepsilon := J *_{\mathbf{x}} \rho_\varepsilon$. Since $\nabla J_\varepsilon = (\nabla J) *_{\mathbf{x}} \rho_\varepsilon$, $\partial_{ij} J_\varepsilon = (\partial_i J) *_{\mathbf{x}} \partial_j \rho_\varepsilon$, $\partial_t J_\varepsilon = (\partial_t J) *_{\mathbf{x}} \rho_\varepsilon$, and $J_\varepsilon(\mathbf{x},t) - J(\mathbf{x},t) = \int \big(J(\mathbf{x}-\mathbf{z},t) - J(\mathbf{x},t)\big)\rho_\varepsilon(\mathbf{z})\, d\mathbf{z}$,

$$\|J - J_\varepsilon\|_\infty \le L_J\varepsilon, \qquad \|\nabla J_\varepsilon\|_\infty \le L_J, \qquad \|\nabla^2 J_\varepsilon\|_\infty \le c_\rho L_J\, \varepsilon^{-1}, \qquad \|\partial_t J_\varepsilon\|_\infty \le L_t. \tag{62}$$

Note that J_ε is C^∞ in $\mathbf{x}$, Lipschitz in t, and that $\nabla^2 J_\varepsilon = J *_{\mathbf{x}} \nabla^2 \rho_\varepsilon$ is continuous in $(\mathbf{x}, t)$.

Step 2: the transformation and its bounds. Write $\theta_n(\mathbf{x},t) := \omega_n t + kJ(\mathbf{x},t)$ for the dither phase actually driving Eq. (21), and $\theta_n^\varepsilon := \omega_n t + kJ_\varepsilon(\mathbf{x},t)$ for its mollified version. Define

$$v_n(\mathbf{x},t) := g_n(\mathbf{x},t)\sqrt{\alpha/\omega_n}\,\sin\theta_n^\varepsilon(\mathbf{x},t), \qquad \Psi(\mathbf{x},t) := \mathbf{x} - \mathbf{v}(\mathbf{x},t), \qquad n = 1,\dots,N, \tag{63}$$

the stochastic counterpart of Eq. (23), with the phase mollified. By (H1) and Eq. (62), $\mathbf{v}$ is C^2 in $\mathbf{x}$ with $\nabla_{\mathbf{x}}^2\mathbf{v}$ continuous, and Lipschitz in t. Since $\partial_i v_n = \sqrt{\alpha/\omega_n}\,\big[\partial_i g_n \sin\theta_n^\varepsilon + kg_n\cos\theta_n^\varepsilon\,\partial_i J_\varepsilon\big]$, using $\omega_n \ge \omega \ge 1$ and $\sqrt{\alpha/\omega}\,\varepsilon^{-1} = \sqrt{\alpha}$ one obtains, uniformly on $\mathbb{R}^N \times [0,T]$,

$$|v_n| \le C_0\sqrt{\alpha/\omega}, \qquad |\partial_i v_n| \le \sqrt{\alpha/\omega}\,C_0(1+|k|L_J), \tag{64}$$

$$\begin{aligned}|\partial_{ij} v_n| &\le \sqrt{\alpha/\omega_n}\Big[|\partial_{ij}g_n| + |k|\big(|\partial_i g_n\partial_j J_\varepsilon| + |\partial_j g_n\partial_i J_\varepsilon|\big) + |k|\,|g_n\partial_{ij}J_\varepsilon| + k^2|g_n\partial_i J_\varepsilon\partial_j J_\varepsilon|\Big]\\ &\le \sqrt{\alpha}\,C_0\big[1+2|k|L_J + k^2L_J^2 + |k|c_\rho L_J\big] =: c_2,\end{aligned} \tag{65}$$

$$\partial_t v_n = g_n\sqrt{\alpha\omega_n}\,\cos\theta_n^\varepsilon + \varrho_n,$$

$$\varrho_n := \sqrt{\alpha/\omega_n}\big[\partial_t g_n\sin\theta_n^\varepsilon + kg_n\,\partial_t J_\varepsilon\cos\theta_n^\varepsilon\big], \qquad |\varrho_n| \le \sqrt{\alpha}\,C_0(1+|k|L_t), \tag{66}$$

where in Eq. (65) the Itô-critical term $|k|\,|g_n\partial_{ij}J_\varepsilon| \le |k|C_0c_\rho L_J\varepsilon^{-1}$ has been multiplied by $\sqrt{\alpha/\omega_n} \le \sqrt{\alpha/\omega} = \sqrt{\alpha}\,\varepsilon$: this is the only place where the choice $\varepsilon = \omega^{-1/2}$ is used, and it makes the second derivatives of the transformation bounded rather than small. In Eq. (66), the leading term arises from $\partial_t\theta_n^\varepsilon = \omega_n + k\partial_t J_\varepsilon$ and is the dither drift of Eq. (21) with mollified phase. Let $\nabla\mathbf{v} := (\partial_i v_n)_{n,i}$. By Eq. (64)–Eq. (65),

$$\begin{aligned}&\|\nabla\mathbf{v}(\mathbf{x},t)\|_{\mathrm{op}} \le \|\nabla\mathbf{v}(\mathbf{x},t)\|_F \le c_1\,\omega^{-1/2}, && c_1 := NC_0\sqrt{\alpha}\,(1+|k|L_J);\\ &\|\nabla\mathbf{v}(\mathbf{x},t) - \nabla\mathbf{v}(\mathbf{x}',t)\|_{\mathrm{op}} \le c_3\|\mathbf{x}-\mathbf{x}'\|, && c_3 := N^2c_2.\end{aligned} \tag{67}$$

Step 3: $\Psi(\cdot,t)$ is a bi-Lipschitz diffeomorphism, uniformly in ω. Set $\omega_0 := \max\{1, 4c_1^2\}$ and let $\omega \ge \omega_0$, so that $\|\nabla\mathbf{v}\|_{\mathrm{op}} \le \frac{1}{2}$ and $\mathbf{v}(\cdot,t)$ is $\frac{1}{2}$-Lipschitz. Then for all $\mathbf{x},\mathbf{x}'$,

$$\tfrac{1}{2}\|\mathbf{x}-\mathbf{x}'\| \;\le\; \|\Psi(\mathbf{x},t) - \Psi(\mathbf{x}',t)\| \;\le\; \tfrac{3}{2}\|\mathbf{x}-\mathbf{x}'\|, \tag{68}$$

so $\Psi(\cdot,t)$ is injective; it is also surjective, because for every $\mathbf{y}\in\mathbb{R}^N$ the map $\mathbf{x}\mapsto \mathbf{y}+\mathbf{v}(\mathbf{x},t)$ is a contraction of $\mathbb{R}^N$ whose unique fixed point $\mathbf{x}$ satisfies $\Psi(\mathbf{x},t)=\mathbf{y}$. Its inverse $\Xi_t := \Psi(\cdot,t)^{-1}$ is therefore globally defined, 2-Lipschitz, and C^1 by the inverse function theorem ($\nabla_{\mathbf{x}}\Psi = I - \nabla\mathbf{v}$ is invertible). The singular values of $I-\nabla\mathbf{v}$ lie in $[\frac{1}{2},\frac{3}{2}]$, whence

$$2^{-N} \le |\det\nabla_{\mathbf{x}}\Psi(\mathbf{x},t)| \le (\tfrac{3}{2})^N, \qquad |\det\nabla_{\mathbf{y}}\Xi_t(\mathbf{y})| \le 2^N. \tag{69}$$

Step 4: Itô's formula—the dither cancels. Let $\mathbf{y}(t) := \Psi(\mathbf{x}_\omega(t),t)$. Since Ψ is C^2 in $\mathbf{x}$ with continuous second derivatives and Lipschitz in t, and since for fixed ω the process $\mathbf{x}_\omega$ has bounded drift and non-degenerate diffusion, Itô's formula applies with the t-derivative entering as a Lebesgue integral (the Itô–Krylov formula, Krylov, 1980, Ch. 2, exactly as invoked in the proof of Theorem 2). With $\boldsymbol{\sigma}$ diagonal,

$$dy_n = dx_{\omega,n} - \partial_t v_n\,dt - \sum_i \partial_i v_n\,dx_{\omega,i} - \tfrac{1}{2}\sum_i g_i^2\,\partial_{ii}v_n\,dt.$$

Insert $dx_{\omega,i} = \big[f_i + g_i\sqrt{\alpha\omega_i}\cos\theta_i\big]dt + g_i\,dw_i$ and Eq. (66): the dither drift $g_n\sqrt{\alpha\omega_n}\cos\theta_n$ of $dx_{\omega,n}$ and the leading term of $\partial_t v_n$ cancel up to the mismatch between θ_n and θ_n^ε, and

$$\begin{aligned}dy_n &= b_n(t)\,dt + \sum_i \big(\delta_{ni} - \partial_i v_n\big)g_i\,dw_i,\\ b_n &:= f_n + g_n\sqrt{\alpha\omega_n}\big[\cos\theta_n - \cos\theta_n^\varepsilon\big] - \varrho_n - \sum_i \partial_i v_n\big[f_i + g_i\sqrt{\alpha\omega_i}\cos\theta_i\big] - \tfrac{1}{2}\sum_i g_i^2\,\partial_{ii}v_n,\end{aligned} \tag{70}$$

all coefficients evaluated at $(\mathbf{x}_\omega(t),t)$. Each term of b_n is bounded uniformly in $\omega \ge 1$: by Eq. (62), $|\cos\theta_n - \cos\theta_n^\varepsilon| \le |k|\,|J-J_\varepsilon| \le |k|L_J\varepsilon$, so the mismatch term is at most $C_0\sqrt{\alpha r_{\max}\omega}\,|k|L_J\varepsilon = C_0\sqrt{\alpha r_{\max}}\,|k|L_J$ (here $\sqrt{\omega}\,\varepsilon = 1$ is used a second time); by Eq. (64) the cross terms with the base drift and with the dither are at most $N\sqrt{\alpha}\,C_0^2(1+|k|L_J)$ and $N\sqrt{\alpha/\omega}\,C_0(1+|k|L_J)\cdot C_0\sqrt{\alpha r_{\max}\omega} =$

$N\alpha\sqrt{r_{\max}}\,C_0^2(1+|k|L_J)$ respectively; and by Eq. (65) the Itô-correction term is at most $\frac{1}{2}NC_0^2c_2$. Hence

$$|b_n(t)| \le B := C_0\big[1+\sqrt{\alpha r_{\max}}\,|k|L_J+\sqrt{\alpha}(1+|k|L_t)\big] + NC_0^2\big[\sqrt{\alpha}\,(1+\sqrt{\alpha r_{\max}})(1+|k|L_J)+\tfrac{1}{2}c_2\big]. \tag{71}$$

The diffusion matrix of Eq. (70) is $\Sigma_\omega(t) := (I-\nabla\mathbf{v})\boldsymbol{\sigma}$ at $(\mathbf{x}_\omega(t),t)$; for $\omega \ge \omega_0$, by (H2) and Eq. (68),

$$\Sigma_\omega\Sigma_\omega^\top = (I-\nabla\mathbf{v})\boldsymbol{\sigma}\boldsymbol{\sigma}^\top(I-\nabla\mathbf{v})^\top \succeq \lambda_0(I-\nabla\mathbf{v})(I-\nabla\mathbf{v})^\top \succeq \tfrac{\lambda_0}{4}I, \qquad \|\Sigma_\omega\|_{\mathrm{op}} \le \tfrac{3}{2}C_0. \tag{72}$$

The transformed process thus has a drift bounded by B and a diffusion coefficient that is bounded and uniformly elliptic, all independently of $\omega \ge \omega_0$. This is the whole point: the $O(\sqrt{\omega})$ dither has been traded for $O(1)$ terms, at the price of an $O(\omega^{-1/2})$ deformation of the diffusion.

Step 5: Markov form and Girsanov. Since $\mathbf{x}_\omega(t) = \Xi_t(\mathbf{y}(t))$, the coefficients of Eq. (70) are functions of $(\mathbf{y}(t),t)$: $b_n(t) = \tilde{b}_n(\mathbf{y}(t),t)$ and $\Sigma_\omega(t) = \tilde{\Sigma}(\mathbf{y}(t),t)$, where $\tilde{\mathbf{b}}(\mathbf{y},t)$ is the expression in Eq. (70) evaluated at $(\Xi_t(\mathbf{y}),t)$ and $\tilde{\Sigma}(\mathbf{y},t) := \big[(I-\nabla\mathbf{v})\boldsymbol{\sigma}\big](\Xi_t(\mathbf{y}),t)$. Then $\tilde{\mathbf{b}}$ is measurable and bounded by B, and $\tilde{\Sigma}$ is bounded by $\frac{3}{2}C_0$, uniformly elliptic with constant $\lambda_0/4$, and Lipschitz in $\mathbf{y}$ uniformly in t and in $\omega \ge \omega_0$, with $\mathrm{Lip}_{\mathbf{y}}(\tilde{\Sigma}) \le 2\big[\frac{3}{2}L + c_3C_0\big] =: \ell$ by Eq. (67) and Eq. (68). Let $\boldsymbol{\theta}(t) := \Sigma_\omega(t)^{-1}\mathbf{b}(t)$, so that $\|\boldsymbol{\theta}\| \le 2\sqrt{N}\,B/\sqrt{\lambda_0}$ by Eq. (72). Novikov's condition holds, and Girsanov's theorem (Karatzas & Shreve, 1991, Ch. 3) provides $\mathbb{Q} \sim \mathbb{P}$ on $\mathcal{F}_T$, with $d\mathbb{Q}/d\mathbb{P} = \exp\big(-\int_0^T \boldsymbol{\theta}\cdot d\mathbf{w} - \frac{1}{2}\int_0^T\|\boldsymbol{\theta}\|^2 ds\big)$ and $\mathbb{Q} = \mathbb{P}$ on $\mathcal{F}_0$, under which $\tilde{\mathbf{w}} := \mathbf{w} + \int_0^{\cdot}\boldsymbol{\theta}\,ds$ is an $(\mathcal{F}_t)$-Brownian motion and

$$d\mathbf{y} = \tilde{\Sigma}(\mathbf{y},t)\,d\tilde{\mathbf{w}}, \qquad \mathbf{y}(0) = \mathbf{y}_0 := \Psi(\mathbf{x}_0,0).$$

Moreover $Z := d\mathbb{P}/d\mathbb{Q} = \exp\big(\int_0^T \boldsymbol{\theta}\cdot d\tilde{\mathbf{w}} - \frac{1}{2}\int_0^T\|\boldsymbol{\theta}\|^2 ds\big)$ satisfies $\mathbb{E}^{\mathbb{Q}}[Z^2\,|\,\mathcal{F}_0] \le e^{2\Lambda}$ with $\Lambda := 2NB^2T/\lambda_0$, because $Z^2 = \exp\big(\int 2\boldsymbol{\theta}\cdot d\tilde{\mathbf{w}} - \frac{1}{2}\int\|2\boldsymbol{\theta}\|^2 ds\big)\exp\big(\int\|\boldsymbol{\theta}\|^2 ds\big)$ and the first factor, a Doléans exponential with bounded integrand, is a $\mathbb{Q}$-martingale. The constant Λ is independent of ω: this is where the argument differs from Lemma 6, whose Novikov exponent grows like ωT over the full horizon.

Step 6: occupation estimate for the transformed process, and transfer. Let $\tilde{h}(t,\mathbf{y}) := h(t,\Xi_t(\mathbf{y}))$. (a) *The transformation preserves uniformly local integrability.* For any unit ball $B_1(\mathbf{y}_0)$, $\Xi_t(B_1(\mathbf{y}_0)) \subset B_2(\Xi_t(\mathbf{y}_0))$, which is covered by K_N unit balls, so by the change of variables $\mathbf{x} = \Xi_t(\mathbf{y})$ and Eq. (69),

$$\int_{B_1(\mathbf{y}_0)} h\big(t,\Xi_t(\mathbf{y})\big)^p\,d\mathbf{y} = \int_{\Xi_t(B_1(\mathbf{y}_0))} h(t,\mathbf{x})^p\,|\det\nabla_{\mathbf{x}}\Psi(\mathbf{x},t)|\,d\mathbf{x} \le (\tfrac{3}{2})^N K_N\,\|h(t,\cdot)\|^p_{L^p_{\mathrm{uloc}}},$$

i.e. $\|\tilde{h}(t,\cdot)\|_{L^p_{\mathrm{uloc}}} \le c_\Psi\|h(t,\cdot)\|_{L^p_{\mathrm{uloc}}}$ with $c_\Psi := \big((\frac{3}{2})^N K_N\big)^{1/p}$.

(b) *Gaussian bound.* Under $\mathbb{Q}$, conditionally on $\mathcal{F}_0$, $\mathbf{y}$ is the (pathwise unique) diffusion $d\mathbf{y} = \tilde{\Sigma}(\mathbf{y},t)\,d\tilde{\mathbf{w}}$ from $\mathbf{y}_0$, whose diffusion coefficient is bounded, uniformly elliptic and Lipschitz in space with constants independent of ω. Its transition density therefore obeys the Gaussian upper bound $p_s(\mathbf{y}_0,\mathbf{y}) \le C_G s^{-N/2}e^{-c_G\|\mathbf{y}-\mathbf{y}_0\|^2/s}$, $0 < s \le T$, with C_G, c_G depending only on (N,λ_0,C_0,ℓ,T) (Friedman, 1964).[1] Let $\{B_1(\mathbf{z}_j)\}_j$ be a cover of $\mathbb{R}^N$ by unit balls with centres on a lattice, so that at most $c_N(1+m)^{N-1}$ of them satisfy $\mathrm{dist}(\mathbf{y}_0,B_1(\mathbf{z}_j)) \in [m,m+1)$. With $r := p/2 > 1$ and

[1] Only the ellipticity, the sup norm and the spatial Lipschitz (indeed Hölder) seminorm of $\tilde{a} := \tilde{\Sigma}\tilde{\Sigma}^\top$ enter C_G, c_G: the parametrix may be built by freezing the coefficient in space only, i.e. from the Gaussian with covariance $\int_\tau^s \tilde{a}(\boldsymbol{\xi},u)\,du$, whose residual $\frac{1}{2}(\tilde{a}(\boldsymbol{\xi},s)-\tilde{a}(\mathbf{y},s)):\nabla^2$ is controlled by the spatial modulus of $\tilde{a}$ alone, so that no regularity in s is consumed. This matters here because the modulus of $\tilde{\Sigma}$ in t is of order $\sqrt{\omega}$. For $p \ge N+1$ one can avoid transition densities altogether: Krylov's estimate for Itô processes with bounded coefficients and non-degenerate diffusion, $\mathbb{E}\int_0^T h(s,\mathbf{y}_s)\,ds \le C\|h\|_{L^{N+1}([0,T]\times\mathbb{R}^N)}$ (Krylov, 1980, Ch. 2), requires no continuity of $\tilde{\Sigma}$ at all, and the uniformly-local norm is recovered from it by localizing to unit cubes and using the tail bound of Corollary 5.

$r' := r/(r-1)$, Hölder's inequality on each ball gives

$$\mathbb{E}^{\mathbb{Q}}\big[\tilde{h}^2(s,\mathbf{y}_s)\,\big|\,\mathcal{F}_0\big] \le \sum_j \|\tilde{h}(s,\cdot)\|^2_{L^p(B_1(\mathbf{z}_j))}\,\|p_s(\mathbf{y}_0,\cdot)\|_{L^{r'}(B_1(\mathbf{z}_j))} \le c_\Psi^2\,\|h(s,\cdot)\|^2_{L^p_{\mathrm{uloc}}} \sum_j \|p_s(\mathbf{y}_0,\cdot)\|_{L^{r'}(B_1(\mathbf{z}_j))}.$$

For the boundedly many (at most c'_N) balls with $\operatorname{dist}(\mathbf{y}_0, B_1(\mathbf{z}_j)) < 2$ we use the global bound

$$\|p_s(\mathbf{y}_0,\cdot)\|_{L^{r'}(\mathbb{R}^N)} \le C_G s^{-N/2}\Big(\int e^{-r' c_G\|\mathbf{y}\|^2/s}\,d\mathbf{y}\Big)^{1/r'} = C s^{-N/2+N/(2r')} = C s^{-N/p};$$

for the others, on which $\|\mathbf{y}-\mathbf{y}_0\| \ge m := \lfloor \operatorname{dist}(\mathbf{y}_0, B_1(\mathbf{z}_j))\rfloor \ge 2$, we use

$$\|p_s(\mathbf{y}_0,\cdot)\|_{L^{r'}(B_1(\mathbf{z}_j))} \le C_G|B_1|^{1/r'} s^{-N/2} e^{-c_G m^2/s} \le C\, e^{-c_G m^2/(2T)}\, m^{-N} \sup_{u>0} u^{N/2} e^{-c_G u/2},$$

and $\sum_{m\ge 2} c_N (1+m)^{N-1} m^{-N} e^{-c_G m^2/(2T)} < \infty$. Hence, for $0 < s \le T$,

$$\mathbb{E}^{\mathbb{Q}}\big[\tilde{h}^2(s,\mathbf{y}_s)\,\big|\,\mathcal{F}_0\big] \le C\big(s^{-N/p}+1\big)\,\|h(s,\cdot)\|^2_{L^p_{\mathrm{uloc}}}. \tag{73}$$

(c) *Integration in time.* Let $a := q/2 \in \big(\frac{p}{p-N}, \infty\big]$ and a' its conjugate exponent, so that $a'N/p < 1$ precisely when $q > 2p/(p-N)$ (for $q=\infty$, $a'=1$ and the condition is $N/p<1$). By Hölder's inequality in s applied to Eq. (73),

$$\mathbb{E}^{\mathbb{Q}}\Big[\int_0^T \tilde{h}^2(s,\mathbf{y}_s)\,ds\;\Big|\;\mathcal{F}_0\Big] \le C\,\big\|s^{-N/p}+1\big\|_{L^{a'}(0,T)}\,\Big\|\,\|h(s,\cdot)\|^2_{L^p_{\mathrm{uloc}}}\Big\|_{L^a(0,T)} = C'\,\|h\|^2_{p,q}, \tag{74}$$

the first factor being finite because $a'N/p < 1$: this is the only place where the lower bound on q enters.

(d) *Transfer.* By the Cauchy–Schwarz inequality under $\mathbb{Q}$, conditionally on $\mathcal{F}_0$,

$$\mathbb{E}\Big[\int_0^T h\big(s,\mathbf{x}_\omega(s)\big)ds\;\Big|\;\mathcal{F}_0\Big] = \mathbb{E}^{\mathbb{Q}}\Big[Z\int_0^T \tilde{h}(s,\mathbf{y}_s)\,ds\;\Big|\;\mathcal{F}_0\Big] \le \big(\mathbb{E}^{\mathbb{Q}}[Z^2\,|\,\mathcal{F}_0]\big)^{1/2}\,T^{1/2}\Big(\mathbb{E}^{\mathbb{Q}}\Big[\int_0^T \tilde{h}^2(s,\mathbf{y}_s)\,ds\;\Big|\;\mathcal{F}_0\Big]\Big)^{1/2} \le e^{\Lambda}\,T^{1/2}\,(C')^{1/2}\,\|h\|_{p,q},$$

and taking expectations proves Eq. (61) for $\omega \ge \omega_0$. For $1 \le \omega < \omega_0$, Steps 5–6 apply verbatim to $\mathbf{x}_\omega$ itself, with $\Psi = \mathrm{id}$ and no transformation: the drift of Eq. (20) is then bounded by $C_0\big(1+\sqrt{\alpha r_{\max}\omega_0}\big)$, and the resulting constant depends on ω_0, which depends only on the admissible parameters. Taking the larger of the two constants gives Eq. (61) for all $\omega \ge 1$. □

The transformation yields a second ω-uniform statement for free, which the two-scale argument of Appendix B.5 also needs (Remark 5).

Corollary 5 (ω-uniform exponential tightness of the dithered process)**.** *Under (H1)–(H4), with B as in Eq. (71) and ω_0 as in Step 3, for all $\omega \ge \omega_0$ and all $R > 0$,*

$$\mathbb{P}\Big(\sup_{t\le T}\|\mathbf{x}_\omega(t)-\mathbf{x}_0\| \ge R + \sqrt{N}\,BT + 2\sqrt{N}\,C_0\sqrt{\alpha}\Big) \;\le\; 2N\exp\Big(-\frac{2R^2}{9NC_0^2T}\Big). \tag{75}$$

Proof. From Eq. (70), $\mathbf{y}(t)-\mathbf{y}_0 = \int_0^t \mathbf{b}\,ds + \mathbf{M}_t$ with $\|\int_0^t \mathbf{b}\,ds\| \le \sqrt{N}\,BT$ and $\mathbf{M}_t = \int_0^t \Sigma_\omega\,d\mathbf{w}$. The nth coordinate M^n is a continuous martingale with $\langle M^n\rangle_t \le \|\Sigma_\omega\|^2_{\mathrm{op}} t \le \frac{9}{4}C_0^2 t$ by Eq. (72), so the exponential martingale inequality (Karatzas & Shreve, 1991, Ch. 3) gives $\mathbb{P}(\sup_{t\le T}|M^n_t| \ge R/\sqrt{N}) \le 2\exp\big(-2R^2/(9NC_0^2T)\big)$. Since $\mathbf{x}_\omega(t)-\mathbf{x}_0 = \mathbf{y}(t)-\mathbf{y}_0 + \mathbf{v}(\mathbf{x}_\omega(t),t) - \mathbf{v}(\mathbf{x}_0,0)$ and $\|\mathbf{v}\| \le \sqrt{N}\,C_0\sqrt{\alpha/\omega} \le \sqrt{N}\,C_0\sqrt{\alpha}$ by Eq. (64), we have $\sup_{t\le T}\|\mathbf{x}_\omega(t)-\mathbf{x}_0\| \le \sqrt{N}\,BT + 2\sqrt{N}\,C_0\sqrt{\alpha} + \sqrt{N}\max_n \sup_{t\le T}|M^n_t|$, and a union bound over n gives Eq. (75). □

Remark 5 (Uniformity in the two-scale argument). Step (ii) of the proof sketch of Proposition 3(b) needs $\mathbb{E}\int_0^T h_\varepsilon(s,\mathbf{x}_\omega(s))\,ds \to 0$ along $\varepsilon = \varepsilon(\omega) \to 0$, where $h_\varepsilon := |\nabla J - \nabla J_\varepsilon|$ satisfies $0 \le h_\varepsilon \le 2L_J$ and $h_\varepsilon \to 0$ a.e. in $(\mathbf{x},t)$ (at every Lebesgue point of $\nabla J(\cdot,t)$, hence a.e. by Fubini). Pointwise convergence alone does not make $\|h_\varepsilon\|_{L_t^\infty L^p_{\mathrm{uloc}}}$ vanish—the supremum over all unit balls of $\mathbb{R}^N$ and over t need not commute with the limit if the kinks of J accumulate at infinity or move erratically in time—and this is where Theorem 4 and Corollary 5 are used together. Fix $R > 0$ and $q \in (\frac{2p}{p-N}, \infty)$, and split $h_\varepsilon = h_\varepsilon \mathbf{1}_{B_R(\mathbf{x}_0)} + h_\varepsilon \mathbf{1}_{B_R(\mathbf{x}_0)^c}$ (Theorem 4 was proved conditionally on $\mathcal{F}_0$, so an $\mathcal{F}_0$-measurable localizing ball is admissible). By Eq. (61),

$$\mathbb{E}\int_0^T \big(h_\varepsilon \mathbf{1}_{B_R(\mathbf{x}_0)}\big)\big(s,\mathbf{x}_\omega(s)\big)\,ds \;\le\; C\,\mathbb{E}\Big(\int_0^T \|h_\varepsilon(s,\cdot)\|^q_{L^p(B_{R+1}(\mathbf{x}_0))}\,ds\Big)^{1/q},$$

which tends to zero as $\varepsilon \to 0$ by dominated convergence (the integrand is bounded by $(2L_J)^q|B_{R+1}|^{q/p}$ and vanishes a.e.), *uniformly in* $\omega \ge 1$, because C does not depend on ω. By Eq. (75), the complementary term is at most $2L_J T\,\mathbb{P}(\sup_{t\le T}\|\mathbf{x}_\omega(t) - \mathbf{x}_0\| \ge R)$, which tends to zero as $R \to \infty$ uniformly in $\omega \ge \omega_0$. Hence $\lim_{\omega\to\infty}\mathbb{E}\int_0^T h_{\varepsilon(\omega)}(s,\mathbf{x}_\omega(s))\,ds = 0$ along any $\varepsilon(\omega) \to 0$, in particular along $\varepsilon = \omega^{-1/8}$ as used in Appendix B.5. The same argument applies to every drift-replacement term of the identification step (2) of that sketch, whose integrands are bounded and vanish a.e. as $\varepsilon \to 0$. No rate in ε is available for merely Lipschitz J, and none is claimed: this is why the Lipschitz rung of Table 6 comes with convergence but without a rate.

Remark 6 (What the transformation does and does not give). (i) *Time regularity.* The Lipschitz-in-time hypothesis of (H3) is used exactly once, to bound ϱ_n in Eq. (66). If J is only of bounded variation in t, as in the standing hypotheses of this appendix, then $\varrho_n\,dt$ becomes a signed measure of total variation $O(\omega^{-1/2}(1+|k|V))$, which is small but singular, and Girsanov's theorem cannot absorb it; mollifying J in t as well does not help, because the mismatch bound $\sqrt{\omega}\,\|J - J_\varepsilon\|_\infty = O(1)$ then requires a modulus of continuity of J in t that bounded variation does not provide. Conjecture 1 therefore remains open for merely bounded-variation time dependence, and only for that reason: the kinks of J in $\mathbf{x}$ play no role in the obstruction. In every instantiation of Sec. 3.3, and in the tracking energies, the time dependence of J and of $\boldsymbol{\sigma}$ is smooth. (ii) *Bounded is not small.* At $\varepsilon = \omega^{-1/2}$ the mismatch term and the Itô-correction term of $\mathbf{b}$ are of order one, not $o(1)$; the transformed process is therefore *not* close to the averaged one, and Theorem 4 does not by itself prove Proposition 3(b). Making those terms small forces $\varepsilon \gg \omega^{-1/2}$, and the mismatch then has to be integrated by parts as a zero-mean oscillation, which is the two-stream architecture of Appendices B.2–B.5. The transformation supplies the occupation estimate that architecture consumes, on a scale of its own, and nothing more. (iii) *The Krylov–Röckner class.* The same computation shows why the transformation does not reach Eq. (19). If ∇J is merely in L^p_{uloc}, $p > N$, then $\|\nabla J_\varepsilon\|_\infty \lesssim \varepsilon^{-N/p}$, $\|\nabla^2 J_\varepsilon\|_\infty \lesssim \varepsilon^{-1-N/p}$, and by Morrey $\|J - J_\varepsilon\|_\infty \lesssim \varepsilon^{1-N/p}$. The Itô-correction term of $\mathbf{b}$ is then bounded only if $\varepsilon \gtrsim \omega^{-p/(2(p+N))}$, while the mismatch term $\sqrt{\omega}\,\|J - J_\varepsilon\|_\infty$ is bounded only if $\varepsilon \lesssim \omega^{-p/(2(p-N))}$; since $p/(2(p-N)) > p/(2(p+N))$, these windows are disjoint for every $N \ge 1$. Two further obstacles are structural rather than technical: Eq. (19) constrains ∇J only, whereas the algorithm evaluates J itself, whose time dependence—an additive $c(t)$, invisible to ∇J—modulates the dither phase directly; and the averaged drift is extracted by differentiating the corrector twice, which requires $\nabla^2 J$ in some function class, whereas Eq. (19) is a condition on the limit drift alone. This is the precise sense in which Eq. (19) marks the boundary of provability for the *limit* (Remark 7) and not, at present, for the averaging.

B.7 How rough can J be?

The hierarchy above is not bottomless, and its floor has a clean description.

Remark 7 (The boundary of provability). The differentiability of J is not the binding constraint; the constraint is that the limiting drift $-\frac{k\alpha}{2}\boldsymbol{\sigma}\boldsymbol{\sigma}^\top\nabla J$ be an admissible drift for the non-degenerate diffusion. By Krylov–Röckner (Krylov & Röckner, 2005) the averaged SDE is strongly well posed whenever the weak gradient satisfies the Ladyzhenskaya–Prodi–Serrin-type condition Eq. (19), i.e. $\nabla J \in L^q([0,T];L^p(\mathbb{R}^N))$ with $\frac{N}{p} + \frac{2}{q} < 1$. Condition Eq. (19) permits ∇J to be unbounded, defined only a.e., and merely p-integrable in space with q-integrability in time, so long as ∇J remains an L^p function; it is the most general regularity under which the limit is provably well posed by classical means. (Note that Eq. (19) forces $p > N$, so Morrey's embedding keeps $J(\cdot,t)$ itself

continuous in $\mathbf{x}$ at a.e. fixed time—what Eq. (19) adds over the Lipschitz rung is genuine local blow-up of the gradient, together with irregularity in time; spatial jumps in J belong to the frontier below.) We expect the averaging convergence to extend throughout Eq. (19): the mollified-corrector proof of Theorem 2 extends once the corrector lies in the matching parabolic Sobolev class and the occupation estimates of Appendix B.3 are available at the corresponding integrability, and the critical case $\frac{N}{p} + \frac{2}{q} = 1$ is treated in (Krylov, 2021). We record in Remark 6(iii) why the pathwise transformation of Appendix B.6, which settles the Lipschitz rung, does not by itself reach Eq. (19): with ∇J merely in L^p, the scale at which the transformation is a diffeomorphism with bounded second derivatives is incompatible with the scale at which the dither mismatch is bounded, and Eq. (19) constrains only the limit drift, not the phase J that the algorithm actually dithers. The genuine outer boundary is reached when ∇J leaves the function-valued world entirely—a true jump in J, whose gradient is a surface measure, i.e. $\nabla J \in C^{-\beta}$ in a Besov sense. There the limiting SDE has distributional drift, well posed only in low dimension or under Besov restrictions and resting on regularization-by-noise arguments (Flandoli et al., 2010). In one line: we can prove convergence with rates up to $W^{2,p}$ energies and without a rate for Lipschitz energies (through the occupation estimate of Theorem 4, under Lipschitz time dependence), expect it up to the Krylov–Röckner class Eq. (19), and reach the open frontier at distributional gradients—and in every case it is the non-degenerate sampling noise, not any smoothness of J, that carries the argument.

C Choosing Bounded ES Parameters

Frequency ratios $\{r_i\}$. Any pairwise-distinct positive constants satisfy Assumption 2; we use r_i evenly spaced (or i.i.d. uniform) in $[1, N]$ with modest $N \in [1.75, 2]$, fixed once. At finite ω the cross-coordinate error terms decay like $(\omega\,|r_i - r_j|)^{-1/2}$ (Appendix A), so the minimal gap $\min_{i\neq j} |r_i - r_j| \approx \frac{3}{4(d-1)}$ ties the required base frequency to dimension: *ω^* grows at least linearly in d at fixed accuracy*. Cross-coordinate beat terms average out once the minimum gap frequency $(N-1)\omega/(d-1)$ (band $[\omega, N\omega]$) exceeds a constant of order ten, giving $\omega^*(d) \approx O(10)(d-1)/(N-1)$ with the constant empirically 5–15; a fixed sweep window returns the $\omega^{-1/2}$ law while its top satisfies this criterion, degrades toward slope 0 once it does not, and recovers when the window is raised — verified up to $d = 1024$ (Appendix D.3, Table 11). Appendix F proves both directions of this law: the threshold is at least linear in d (Proposition 6) and at most polynomial (Proposition 4), and the number of evaluations needed to track the averaged dynamics grows at least linearly in d for every band and horizon (Theorem 5).

Step size Δt and evaluation count. Euler–Maruyama must resolve the fastest dither: we use $\omega r_{\max}\,\Delta t \leq 2\pi/10$, i.e., at least 10 steps per fastest dither period — a threshold validated directly: measured rates are invariant from 10 to 100 steps per period (a $10\times$ range in Δt) and deviate only at 5 (Appendix D.2). Since one step costs exactly one evaluation of J (line 2), the number of function evaluations (NFE) per unit diffusion time is $1/\Delta t \approx 10\,\omega r_{\max}/2\pi$: *NFE scales linearly with ω, while the averaging error decays like $\omega^{-1/2}$ at best*. By contrast, a two-point Gaussian-smoothing estimate (Nesterov & Spokoiny, 2017; Roy et al., 2022) costs 2 evaluations per Langevin step and a coordinate-wise finite difference costs $d + 1$; ES costs 1 but takes more, smaller steps, and measurement noise on J enters only through the dither phase, where it is averaged over the demodulation window, rather than being amplified by a $1/\nu$ divided difference (this claim is quantified — including a validated noise-floor law and its breakdown point — in Appendix D.6). The two measured laws compose: the physical settling time is independent of ω and equals the averaged system's own (Appendix D.4), so evaluations-to-converge grow $\propto \omega$ while accuracy improves as $\omega^{-1/2}$ — accuracy $\sim$ NFE$^{-1/2}$, measured.

Gains. $k\alpha$ sets the strength of the emergent drift (Eq. 6); the instantiations of Sec. 3.3 fix it to match the target dynamics. Because $k\alpha$ thereby also sets the temperature of the averaged law ($\rho_\infty \propto e^{-k\alpha\mathcal{E}_\theta}$ in (I_λ)), any dimension-dependent reduction of k adopted for stability at finite ω—as in the mode-seeking benchmark of Appendix D.6, where the temperature is immaterial—must be compensated by α, or by rescaling J, whenever the goal is sampling from a fixed target. Excessive k at finite ω destabilizes the fast dynamics—Fig. 4(bottom) shows exactly this failure mode—so we sweep k around the nominal value in every experiment.

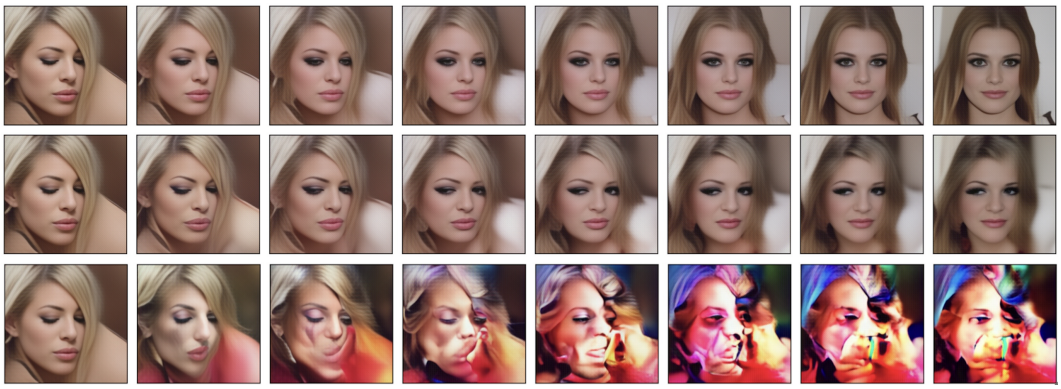


Figure 4: Latent-space tracking from 1D RGB projections of time-varying images. The top row shows the true images, the second row shows decent tracking, and the bottom row shows catastrophic instability and failure when the ES gain $k\alpha > 0$ is extremely large.

D Detailed numerical studies on the synthetic testbed

This appendix documents, in full, the quantitative studies of the direct verification of Theorem 1 and its $\omega^{-1/2}$ rate across dimension, noise level, and time discretization (D.2); the finite-ω transition, its dimension scaling, and the robustness checks behind it (D.3); the convergence-speed and cost measurements (D.4); the matched-budget comparison against gradient-based and derivative-free samplers (D.5); a fixed-target benchmark of the mode-seeking configuration against derivative-free *optimizers* under measurement noise, with a validated noise-floor law and a readout (iterate-averaging) study (D.6).

D.1 Testbed, conventions, and the coupled protocol

Target. All synthetic studies use a four-mode Gaussian mixture in $\mathbb{R}^d$ with modes placed symmetrically in the first two coordinates and isotropic unit covariance:

$$p(\mathbf{x}) \;=\; \frac{1}{4}\sum_{m=1}^{4} \mathcal{N}\big(\mathbf{x};\, \boldsymbol{\mu}_m,\, I_d\big), \qquad \boldsymbol{\mu}_m \;=\; \big(\pm s,\, \pm s,\, 0,\, \ldots,\, 0\big) \in \mathbb{R}^d, \qquad s = 3, \tag{76}$$

with energy $\mathcal{E}(\mathbf{x}) = -\log p(\mathbf{x})$. The four modes are well separated ($\|\boldsymbol{\mu}_m - \boldsymbol{\mu}_{m'}\| \in \{6, 6\sqrt{2}\}$ against unit component width) in the first two coordinates, while the remaining $d-2$ coordinates are a product of standard Gaussians; $\log p$ and $\nabla \log p$ are available in closed form (via a log-sum-exp and its responsibilities), which supplies exact-gradient oracles and exact target samples for the metrics below. The sampler, of course, only ever evaluates the scalar $\mathcal{E}$.

ES configuration. The studies run the discretization of Eq. (4) in the convention

$$dx_i \;=\; g\sqrt{\alpha\omega_i}\,\cos\big(\omega_i t + k\,J(\mathbf{x})\big)\,dt \;+\; g\,\sigma\,dw_i, \qquad g = 1,\; k = \alpha = 1,\; \sigma \in \{0,\, 0.1,\, 1\}, \tag{77}$$

with $\omega_i = \omega r_i$, ratios r_i evenly spaced in $[1, N]$ (fixed once), Euler step $\Delta t = 2\pi/(cN\omega)$ (c steps per fastest dither period), dispersed initialization $\mathbf{x}_0 \sim \mathcal{N}(0,\, 2^2 I_d)$, 256 chains, and a fixed physical horizon $T = 4$ (so the step count $\lceil T/\Delta t \rceil$ grows linearly in ω). In this convention the averaged dynamics are $d\bar{\mathbf{x}} = \frac{k\alpha}{2} g^2 \nabla \log p\, dt + g\sigma\, d\mathbf{w}$, whose stationary law is $\propto p^{\,k\alpha/\sigma^2}$ (the zero-flux condition of the stationary Fokker–Planck equation gives $\nabla \log \rho_\infty = 2\bar{\boldsymbol{\mu}}/(g\sigma)^2 = (k\alpha/\sigma^2)\nabla \log p$, in which g^2 cancels): the $\sigma = 1$ rows therefore sample p exactly and constitute the sampling setting, while $\sigma \in \{0, 0.1\}$ probe the deterministic-to-stochastic range of the trajectory statement of Theorem 1, which holds for any $\sigma \geq 0$.

Coupled protocol. Following the coupling in Theorem 1, each ES run is paired with an Euler simulation of its averaged SDE using the *same* initial condition and the *same* per-step Gaussian

increments $\boldsymbol{\xi}_j$:

$$x_{j+1,i} = x_{j,i} + g\sqrt{\alpha\,\omega_i}\,\cos\big(\omega_i t_j + k\,J(\mathbf{x}_j)\big)\,\Delta t + g\sigma\sqrt{\Delta t}\,\xi_{j,i}, \tag{78}$$

$$\bar{\mathbf{x}}_{j+1} = \bar{\mathbf{x}}_j + \tfrac{k\alpha}{2} g^2\,\nabla \log p(\bar{\mathbf{x}}_j)\,\Delta t + g\sigma\sqrt{\Delta t}\,\boldsymbol{\xi}_j. \tag{79}$$

The primary metric is the per-chain pathwise supremum $\sup_{j\le T/\Delta t}\|\mathbf{x}_j - \bar{\mathbf{x}}_j\|_2$, averaged over the 256 chains; for each seed we fit the slope of $\log(\text{metric})$ against $\log\omega$ by least squares over the sweep window, and report the mean $\pm$ std of the per-seed slopes. (Time-averaged normalized errors — NMSE and NMAE — were also recorded; the supremum is the theorem-faithful quantity, and since NMSE is a squared error its fitted slopes are approximately twice those quoted here.)

D.2 Rate in ω: dimension, noise level, and discretization

Tables 7–9 report the fitted slopes over the window $\omega \in \{10, 31.6, 100, 316, 1000\}$ with band $N = 1.75$, for $d \in \{2, \dots, 32\}$, noise $\sigma \in \{0, 0.1, 1\}$, and steps-per-period $c \in \{5, 10, 20, 100\}$.

Table 7: Fitted log–log slope of the coupled sup-error vs. ω; $\sigma = 0$ (deterministic ES vs. gradient flow), window $\omega \in [10, 10^3]$, $N = 1.75$. Mean$\pm$std over 5 seeds.

d	$c = 5$	$c = 10$	$c = 20$	$c = 100$
2	-0.410 ± 0.026	-0.489 ± 0.048	-0.487 ± 0.057	-0.489 ± 0.056
4	-0.494 ± 0.035	-0.587 ± 0.042	-0.594 ± 0.043	-0.596 ± 0.038
8	-0.574 ± 0.020	-0.655 ± 0.014	-0.650 ± 0.014	-0.655 ± 0.015
16	-0.652 ± 0.005	-0.705 ± 0.013	-0.703 ± 0.018	-0.701 ± 0.018
32	-0.641 ± 0.006	-0.645 ± 0.005	-0.649 ± 0.006	-0.652 ± 0.006

Table 8: As Table 7, for $\sigma = 0.1$.

d	$c = 5$	$c = 10$	$c = 20$	$c = 100$
2	-0.395 ± 0.037	-0.499 ± 0.050	-0.504 ± 0.055	-0.448 ± 0.058
4	-0.465 ± 0.023	-0.550 ± 0.050	-0.552 ± 0.041	-0.571 ± 0.023
8	-0.578 ± 0.025	-0.653 ± 0.016	-0.661 ± 0.021	-0.657 ± 0.024
16	-0.646 ± 0.014	-0.705 ± 0.019	-0.709 ± 0.014	-0.720 ± 0.017
32	-0.635 ± 0.008	-0.642 ± 0.007	-0.645 ± 0.007	-0.646 ± 0.006

Table 9: As Table 7, for $\sigma = 1$ (the sampling setting: the averaged dynamics target p exactly).

d	$c = 5$	$c = 10$	$c = 20$	$c = 100$
2	-0.387 ± 0.031	-0.459 ± 0.021	-0.466 ± 0.009	-0.452 ± 0.011
4	-0.448 ± 0.018	-0.500 ± 0.013	-0.493 ± 0.010	-0.474 ± 0.016
8	-0.522 ± 0.011	-0.526 ± 0.012	-0.514 ± 0.014	-0.527 ± 0.009
16	-0.558 ± 0.006	-0.550 ± 0.007	-0.532 ± 0.010	-0.539 ± 0.008
32	-0.544 ± 0.003	-0.530 ± 0.004	-0.509 ± 0.009	-0.511 ± 0.007

Three conclusions. **(a) Discretization.** The slopes are invariant, at every (d, σ), across $c \in \{10, 20, 100\}$ — a $10\times$ range in Δt — while $c = 5$ deviates systematically toward shallower values: the resolved-discretization regime is $c \ge 10$, and the deviation of $c = 5$ identifies the threshold rather than any property of the continuous system. This validates the default $c = 20$ as conservative and $c = 10$ as the cost-optimal resolved choice (halving evaluations per unit time). **(b) The sampling setting.** At $\sigma = 1$ the fitted rate is -0.50 ± 0.04 across the entire range $d \in [2, 32]$ with no systematic dimension dependence: $\omega^{-1/2}$, dimension-independent, as Theorem 1 requires. **(c) Low noise.** At $\sigma \in \{0, 0.1\}$, mid-dimensional fits steepen to ≈ -0.7, i.e., the error decays *faster* than the one-sided $\omega^{-1/2}$ bound over the measured window. This excess is invariant under the $10\times$ step refinement above, under randomization of the dither-frequency grid (D.3), and across the noise levels probed; we attribute it to higher-order error terms of the expansion (the $O(\omega^{-1})$-type contributions in the proof) whose constants grow with d, and note that decay faster than the bound is fully consistent with Theorem 1, which is one-sided. The local-slope analysis below resolves this excess into a transition band that moves with d.

D.3 THE FINITE-ω TRANSITION AND ITS DIMENSION SCALING

Two clocks. The iteration Eq. (77) with $\Delta t = 2\pi/(cN\omega)$ looks self-similar in ω: the fastest dither always takes c steps per oscillation and the slowest cN. The substitution $\tau = \omega t$ (with the Brownian scaling $dw(t) = \omega^{-1/2} dW(\tau)$) makes this precise and shows what does change:

$$dx_i \;=\; g\sqrt{\tfrac{\alpha r_i}{\omega}}\,\cos\big(r_i\tau + k\mathcal{E}\big)\,d\tau \;+\; g\,\sigma\,\omega^{-1/2}\,dW_i(\tau). \tag{80}$$

On this fast clock the dither frequencies are the fixed ratios r_i and the Euler grid $\Delta\tau = 2\pi/(cN)$ is ω-independent, but every term that moves the state carries $\omega^{-1/2}$: per dither period the state moves only $O(\omega^{-1/2})$, so the phase $k\mathcal{E}(\mathbf{x})$ is frozen to that order over each oscillation (the working condition of averaging, and the origin of the $\omega^{-1/2}$ rate), while each step covers only $2\pi/(cN\omega)$ of slow time (the origin of the linear-in-ω cost). ω is thus the exchange rate between the clocks: the number of dither oscillations per unit of the averaged dynamics' own timescale.

Beats and the gap criterion. Expanding the coordinate ripples (cf. the warm-up of A.2) into the phase, the drift of coordinate n contains products of its dither with every coordinate's ripple; by the product formulas these oscillate at the sum frequencies $(r_n{+}r_i)\omega$ (always fast) and the *beat* frequencies $\Delta_{ni} = \omega|r_n - r_i|$, of which the smallest, for evenly spaced ratios, is

$$\Delta_{\min} \;=\; \frac{(N-1)\,\omega}{d-1}. \tag{81}$$

The resonant term $i = n$ (zero beat frequency) is the emergent gradient; every other term is an oscillatory integral $\int_0^T A(t)\cos(\Delta t)\,dt$ with slowly varying envelope, bounded after one integration by parts by $(2\|A\|_\infty + T\|\dot{A}\|_\infty)/\Delta$: it is negligible precisely when the beat completes many cycles per unit of the slow timescale. Empirically "many" is a factor of order ten, giving the finite-ω threshold

$$\Delta_{\min} \;\gtrsim\; O(10) \qquad\Longleftrightarrow\qquad \omega^*(d) \;\approx\; \frac{O(10)\,(d-1)}{N-1}, \tag{82}$$

with the constant measured below to be 5–15. This converts the existence statement of Theorem 1 into a design formula: linear in d, controlled by the band spread.

Local slopes locate the transition. Table 10 reports segment-by-segment (local) slopes between consecutive ω's of the sweep, per dimension and noise level. At $\sigma \in \{0, 0.1\}$ a steep band (local slopes -0.8 to -1.0) is visible whose location moves to higher ω with d: the steepest segments sit at geometric segment centers $\approx 56,\ 178,\ 562$ for $d = 8, 16, 32$ — in order-of-magnitude agreement with Eq. (82) ($13.3\,(d-1) = 93,\ 200,\ 413$ at $N = 1.75$), growing slightly faster than linearly, unsurprising since the *number* of near-minimal gaps also grows with d. On either side of the band the local rate is near $-1/2$; at $\sigma = 1$ the band is smeared by the diffusion and only a weak bump remains. Fitted whole-window slopes steeper than $-1/2$ (Tables 7–8) are chords through this band.

High-dimensional falloff and verification by window shifting. Extending the sweep to $d = 1024$ on a fixed window makes the threshold's motion visible from the other side: once even the window top is sub-threshold, the fitted chord returns the pre-asymptotic plateau and slides toward 0. Table 11 reports the fitted slopes for two windows — A: $\omega \in [10, 10^3]$, $N = 1.75$; B: $\omega \in [10, 10^4]$, $N = 2$, so that the top-of-window gap frequency $\Delta_{\min}(\omega_{\text{hi}})$ grows by a factor 13.3 — at $c = 10$ throughout. If Eq. (82) is right, raising the window top must restore the slope at fixed d; it does, monotonically, at every d.

The slope tracks $\Delta_{\min}$, not d per se: wherever the window top satisfies the criterion Eq. (82) the fitted rate sits near $-1/2$ (or in the overshoot band at low σ); the largest dimensions recover monotonically as the window climbs past $\omega^*(d)$, their chords still averaging over the sub-threshold lower window (hence the partial values at $d \geq 512$). At $d = 1024$ the top of window A gives $\Delta_{\min} = 0.73 < 1$: adjacent dithers complete less than one relative beat per unit slow time, no averaging of the closest pairs is possible, and the measured slope is correspondingly near zero. Full recovery at $d \geq 512$ is predicted at $\omega \gtrsim 10\,(d-1)/(N-1)$, i.e., a $[10^3, 10^5]$ window.

Table 10: Local (segment) slopes of the coupled sup-error between consecutive sweep frequencies $\omega \in \{10, 31.6, 100, 316, 1000\}$ (segment centers $\approx 17.8,\ 56,\ 178,\ 562$), and the fit over the top three points. Mean over 5 seeds; $N = 1.75$.

	local slope per segment				
d	10→31.6	31.6→100	100→316	316→1000	top-3 fit
$\sigma = 0$					
2	−0.47	−0.51	−0.50	−0.40	−0.45
4	−0.63	−0.64	−0.51	−0.59	−0.55
8	−0.53	−0.77	−0.68	−0.54	−0.61
16	−0.43	−0.67	−0.89	−0.74	−0.82
32	−0.27	−0.53	−0.80	−0.99	−0.90
$\sigma = 0.1$					
2	−0.45	−0.47	−0.51	−0.56	−0.53
4	−0.65	−0.56	−0.57	−0.39	−0.48
8	−0.54	−0.77	−0.69	−0.57	−0.63
16	−0.43	−0.66	−0.89	−0.79	−0.84
32	−0.26	−0.53	−0.79	−0.98	−0.88
$\sigma = 1$					
2	−0.43	−0.52	−0.47	−0.42	−0.44
4	−0.46	−0.54	−0.48	−0.47	−0.48
8	−0.46	−0.56	−0.51	−0.49	−0.50
16	−0.36	−0.63	−0.57	−0.50	−0.53
32	−0.16	−0.57	−0.64	−0.57	−0.60

Table 11: Whole-window slopes vs. dimension for the two sweep windows, with the top-of-window minimum gap frequency $\Delta_{\min}(\omega_{\text{hi}}) = (N-1)\,\omega_{\text{hi}}/(d-1)$. Window A: $\omega_{\text{hi}} = 10^3$, $N = 1.75$ ($\sigma = 1$ shown; the $d \le 32$ block coincides with the $c = 10$ column of Table 9). Window B: $\omega_{\text{hi}} = 10^4$, $N = 2$, all three noise levels. Mean±std over 5 seeds.

	window A (σ=1)		window B			
d	$\Delta_{\min}$	slope	$\Delta_{\min}$	slope, σ=0	slope, σ=0.1	slope, σ=1
2	750	-0.459 ± 0.021	10000	-0.469 ± 0.023	-0.440 ± 0.031	-0.431 ± 0.005
4	250	-0.500 ± 0.013	3333	-0.529 ± 0.010	-0.497 ± 0.016	-0.467 ± 0.019
8	107	-0.526 ± 0.012	1429	-0.558 ± 0.008	-0.557 ± 0.026	-0.498 ± 0.009
16	50	-0.550 ± 0.007	667	-0.615 ± 0.022	-0.642 ± 0.004	-0.520 ± 0.012
32	24	-0.530 ± 0.004	323	-0.673 ± 0.013	-0.662 ± 0.016	-0.523 ± 0.006
64	11.9	-0.465 ± 0.003	159	-0.671 ± 0.009	-0.673 ± 0.005	-0.508 ± 0.005
128	5.9	-0.365 ± 0.002	79	-0.594 ± 0.003	-0.598 ± 0.002	-0.464 ± 0.004
256	2.9	-0.241 ± 0.001	39	-0.495 ± 0.003	-0.500 ± 0.002	-0.408 ± 0.001
512	1.5	-0.122 ± 0.001	19.6	-0.392 ± 0.001	-0.391 ± 0.001	-0.337 ± 0.001
1024	0.73	-0.038 ± 0.000	9.8	-0.285 ± 0.000	-0.289 ± 0.001	-0.254 ± 0.000

Why widening the band does not help. Resolving the fastest dither forces $\Delta t = 2\pi/(cN\omega)$, so the evaluation cost per unit slow time is $cN\omega/2\pi = c\,\omega_{\max}/2\pi$ — set by the *top* of the band $\omega_{\max} = N\omega$ regardless of how it is split. At fixed $\omega_{\max}$, widening N buys relative minimum gap $(N-1)/N$, which saturates immediately (already 0.67 at $N = 3$), while the slowest coordinates' own averaging error remains at $(\omega_{\max}/N)^{-1/2}$ and the cross-term prefactors $\sqrt{\omega_i/\omega_n} \le \sqrt{N}$ grow. We therefore fix a modest band ($N \in [1.75, 2]$) and scale the base frequency; the dimension cost is then intrinsic: reaching the asymptotic regime at dimension d requires $\omega_{\max} = \Omega(d)$ and hence $\Omega(d)$ evaluations per unit slow time — the quantitative basis for the latent-space scoping of Sec. 3.2.

Robustness checks. *(i) Frequency-grid randomization.* The evenly spaced grid is an arithmetic progression and therefore carries exact higher-order sum resonances ($r_i + r_k = 2r_j$ for every consecutive triple). Jittering each ratio by up to 20% of the grid spacing (destroying the exact resonances while keeping the gap scale; minimum gap 0.035 vs. 0.050) reproduces the linspace errors to within 2% at every ω and the local slopes to ± 0.02 (checked at $d = 16$, $\sigma = 0$, $c = 20$;

$\omega \in \{31.6, 100, 316, 1000\}$: errors $3.62/1.68/0.616/0.245$ vs. $3.60/1.71/0.621/0.251$, local slopes $-0.67/-0.87/-0.80$ vs. $-0.65/-0.88/-0.79$; 2 seeds, 64 chains). Grid design is thus not the origin of the low-noise steepening, and the practical recommendation "evenly spaced or i.i.d. ratios, fixed once" carries a data point. *(ii) Floating point.* The largest dither phase in these sweeps is $\omega_{\max} T \leq 2 \times 10^5$ radians, resolved by double precision to $\sim 10^{-11}$ radians, and the measured differences $\mathbf{x} - \bar{\mathbf{x}}$ are $O(10^{-1})$ against states $O(1)$: no cancellation is involved. Together with the $10\times$ step-refinement invariance of D.2, the transition structure documented here is a property of the continuous-time coupled system, not of its implementation.

D.4 Convergence speed and the cost law

Settling-time detector. Convergence speed is measured as the *last-exit settling time* of the chain-mean energy $m(t) = \frac{1}{256} \sum_{\text{chains}} J(\mathbf{x}(t))$, the standard settling convention of control engineering: m is smoothed over $\min(4c,\ \text{series length}/10)$ steps ($\approx$ two slowest dither periods) to remove the ripple; the plateau m_∞ and its noise s_∞ are taken from the final 25% of the run; the band is $\max\big(0.05\,|m(0) - m_\infty|,\ 4\,s_\infty\big)$; and t_{conv} is the time after which m never again leaves the band. Two guard flags accompany every measurement: *settled* is false when the last exit lands in the final 10% of the horizon (plateau not reached), and *transient* is false when $|m(0) - m_\infty| < 3s_\infty$, i.e., there was no transient to time — the signature of the sub-ω^* regime, where the effective drift is too weak to move the ensemble; such runs are excluded from slope fits. The same detector is applied to the coupled averaged twin, whose settling time is the reference. Horizon $T = 4$, $c = 20$, $N = 1.75$, all (σ, ω, d) of the main sweep; two clocks are reported: physical time t_{conv} and steps-to-settle (equal to the per-chain evaluation count).

Table 12: Convergence-speed summary. Fitted log–log slopes of t_{conv} and of the evaluations-to-settle vs. ω over $\omega \in [10, 10^3]$, and the ES-to-averaged settling-time ratio at $\omega = 10^3$. Entries $\sigma = 0\,/\,0.1\,/\,1$; 5 seeds; all 75 configurations settled $5/5$.

d	slope of t_{conv}	slope of evals-to-settle	ES/averaged at ω=10^3
2	$+0.02\,/\,+0.02\,/\,+0.02$	$+1.02\,/\,+1.02\,/\,+1.02$	$1.00\,/\,1.00\,/\,0.99$
4	$+0.02\,/\,+0.02\,/\,+0.07$	$+1.02\,/\,+1.02\,/\,+1.07$	$1.00\,/\,1.00\,/\,1.01$
8	$+0.01\,/\,+0.01\,/\,+0.02$	$+1.01\,/\,+1.01\,/\,+1.02$	$1.00\,/\,1.00\,/\,1.00$
16	$+0.05\,/\,+0.05\,/\,+0.08$	$+1.05\,/\,+1.05\,/\,+1.08$	$1.00\,/\,1.00\,/\,1.00$
32	$+0.04\,/\,+0.04\,/\,+0.04$	$+1.04\,/\,+1.04\,/\,+1.04$	$0.99\,/\,0.99\,/\,1.00$

Results. Table 12: the physical settling time is independent of ω (all fitted slopes $\leq +0.08$ over two decades) and coincides with the averaged system's own settling time (ratio 0.99–1.01) for every (d, σ) — convergence speed is governed by the averaged dynamics, an independent verification of Theorem 1 in a different observable than the trajectory error. Correspondingly the evaluation count to settle grows linearly in ω (slopes $+1.01$ to $+1.08$). Combining the two measured laws with the rate of D.2 gives the efficiency scaling quoted in Sec. 3.2: accuracy $\sim$ (evaluations)$^{-1/2}$. Tables 13–14 list the full measurements at $\sigma = 0$ and $\sigma = 1$; $\sigma = 0.1$ matches $\sigma = 0$ within the quoted uncertainties cell by cell and is omitted. The small common upward drift of t_{conv} at low ω (e.g., $2.38 \to 2.60$ for $d = 2$) is finite detector resolution — the smoothing window is fixed in steps, hence O(one dither period) wide in time at small ω — and affects ES and twin identically, which is why the per-ω ratio is the trustworthy statistic. The isolated low-ω/large-d cells where ES settles *early* to a degraded plateau ($\sigma = 0$: $d \geq 16$, $\omega \leq 31.6$; $\sigma = 1$: $d = 16$, $\omega = 31.6$, and the flagged $d = 8$, $\omega = 10$ cell with transient $2/5$) lie below the transition threshold of D.3 — the two studies locate the same band independently.

D.5 Matched-budget comparison with gradient-based and derivative-free samplers

Protocol. The comparison is designed around two symmetries identified during pilot experiments. *Equal effective sample size:* metrics are computed on the *final states* of 1024 chains for every method — no trajectory pooling, whose effective sample size depends on each method's decorrelation time. The finite-sample floor of the metric (sliced-W_2 between an independent 1024-sample draw of the target and the reference set; 0.245 at d=2, 0.110 at d=10) is reported, and

Table 13: Full settling measurements, $\sigma = 0$ ($c = 20$; 5 seeds; all cells settled $5/5$, transient $5/5$). Per cell: t_{conv} and evaluations-to-settle.

d	ω=10	ω=31.6	ω=100	ω=316	ω=10^3
2	2.38±0.02; 132	2.40±0.01; 423	2.54±0.01; 1412	2.58±0.02; 4551	2.60±0.02; 14478
4	2.33±0.01; 130	2.33±0.00; 411	2.48±0.00; 1384	2.55±0.01; 4485	2.56±0.01; 14261
8	2.58±0.01; 144	2.26±0.02; 399	2.43±0.00; 1355	2.51±0.00; 4425	2.54±0.00; 14143
16	2.09±0.05; 117	2.07±0.07; 364	2.36±0.00; 1312	2.48±0.00; 4368	2.52±0.00; 14049
32	2.11±0.01; 118	2.11±0.03; 372	2.39±0.01; 1334	2.42±0.00; 4258	2.50±0.00; 13943

Table 14: As Table 13, for $\sigma = 1$ (the sampling setting). All cells settled $5/5$; transient $5/5$ except (d=8, ω=10) with $2/5$ — the below-threshold cell excluded from the slope fits of Table 12.

d	ω=10	ω=31.6	ω=100	ω=316	ω=10^3
2	2.15±0.28; 120	2.13±0.17; 375	2.22±0.29; 1237	2.05±0.08; 3604	2.43±0.19; 13513
4	1.68±0.31; 94	2.30±0.17; 405	2.41±0.19; 1342	2.32±0.13; 4095	2.47±0.19; 13734
8	0.46±0.92; 26	2.32±0.13; 409	2.29±0.12; 1277	2.46±0.14; 4326	2.46±0.07; 13709
16	2.15±0.04; 120	1.31±0.11; 232	2.39±0.07; 1333	2.49±0.12; 4380	2.48±0.06; 13835
32	2.13±0.05; 118	2.21±0.15; 389	2.17±0.05; 1209	2.49±0.12; 4378	2.51±0.08; 14001

entries within $\sim 15\%$ of it are marked unresolved. *Symmetric tuning:* every method's hyperparameters are chosen by an identical pilot on a held-out seed (seed 0; evaluation uses seeds 1–5): step size $h \in \{3\times10^{-3}, 10^{-2}, 3\times10^{-2}, 10^{-1}\}$ for the Langevin samplers, smoothing radius $\nu \in \{10^{-3}, 10^{-2}, 10^{-1}\}$ for the zeroth-order and finite-difference baselines, and the time/frequency split $T \in \{2, 4, 8, 16, 32\}$ (with $\omega = 2\pi\,\text{NFE}/(c \cdot 1.75\,T)$ implied by the budget, $c = 10$, and ω constrained above the measured threshold) for ES. All methods target p of Eq. (76); the ES rows use $g = \sqrt{2}$, $\sigma = 1$, so that the averaged dynamics are the standard-speed Langevin diffusion.

Methods and per-step evaluation accounting. Oracle ULA, $\mathbf{x} \leftarrow \mathbf{x} + h\nabla \log p + \sqrt{2h}\,\boldsymbol{\xi}$ (1 gradient); oracle MALA, the same proposal with Metropolis correction (1 gradient); ZO-ULA with the two-point Gaussian-smoothing estimate $\hat{\mathbf{g}} = -\mathbf{u}\,[J(\mathbf{x}+\nu\mathbf{u}) - J(\mathbf{x}-\nu\mathbf{u})]/(2\nu)$, $\mathbf{u} \sim \mathcal{N}(0, I)$ (Nesterov & Spokoiny, 2017; Roy et al., 2022) (2 evaluations); FD-ULA with forward coordinate differences (d+1 evaluations); and ES (1 evaluation, all coordinates). Each method receives the same per-chain evaluation budget; its step count is the budget divided by its per-step cost.

Table 15: Sliced-W_2 to ground truth on final states of 1024 chains (mean±std, seeds 1–5). Metric floors: 0.245 (d=2), 0.110 (d=10); entries within 15% of the floor are at metric resolution. The ES pilot selected $(T, \omega) = (8, 90), (4, 1795), (4, 180), (16, 449)$ for the four columns, respectively. All methods recovered $4/4$ modes in every run, and unbiased RBF-MMD^2 values were zero within noise throughout.

	$d = 2$		$d = 10$	
Method	NFE 2×10^3	NFE 2×10^4	NFE 2×10^3	NFE 2×10^4
oracle ULA (1 grad)	0.191 ± 0.038	0.215 ± 0.069	0.114 ± 0.011	0.110 ± 0.018
oracle MALA (1 grad)	0.213 ± 0.056	0.239 ± 0.053	0.113 ± 0.025	0.102 ± 0.008
ZO-ULA (2 ev)	0.295 ± 0.052	0.218 ± 0.041	0.123 ± 0.010	0.135 ± 0.030
FD-ULA (d+1 ev)	0.308 ± 0.054	0.259 ± 0.095	0.307 ± 0.007	0.118 ± 0.012
ES, ours (1 ev)	0.211 ± 0.026	0.232 ± 0.042	0.126 ± 0.007	0.118 ± 0.022

Findings. ES is statistically indistinguishable from the tuned exact-gradient oracles in all four settings, at one evaluation per step; it resolves ahead of both derivative-free baselines at d=2, NFE=2×10^3; and the one clearly resolved deficit in the table is FD-ULA at d=10, 2×10^3, whose d+1 evaluations per step leave only NFE/11 steps and show as incomplete burn-in (mean final energy 20.4 vs. ≈ 15.7 for all other methods). We state the scope plainly: a smooth, low-dimensional testbed with every sampler tuned — the table supports parity at matched budget, not superiority.

D.6 THE MODE-SEEKING CONFIGURATION: A FIXED-TARGET BENCHMARK AGAINST DERIVATIVE-FREE OPTIMIZERS UNDER MEASUREMENT NOISE

The comparison of D.5 evaluates the *sampling* configuration ($\sigma = 1$: the averaged dynamics are the Langevin diffusion and the metric is distributional). This subsection evaluates the complementary *mode-seeking* configuration: the diffusion is taken to (numerically) zero, so that by the tempering corollary (Corollary 3, stationary law $\propto p^{k\alpha/\sigma^2}$) the averaged dynamics degenerate to the gradient flow $\dot{\bar{\mathbf{x}}} = -\frac{k\alpha}{2}\nabla\mathcal{E}(\bar{\mathbf{x}})$ and the algorithm is being used as a derivative-free *optimizer* of the energy $\mathcal{E} = -\log p$. This is the operating mode relevant to the tracking experiments, where the objective is to sit on the moving minimizer rather than to cover the posterior, and it is the mode in which the noise-robustness claim of Sec. 3.2 ("measurement noise on J enters only through the dither phase") can be tested quantitatively. The benchmark below therefore answers three questions on the testbed Eq. (76): (i) at matched total evaluation budgets, how does ES compare with the standard derivative-free optimization baselines — simultaneous-perturbation stochastic approximation (SPSA) (Spall, 1992), Gaussian-smoothing (GS) zeroth-order (ZO) gradient estimates (Nesterov & Spokoiny, 2017; Ghadimi & Lan, 2013; Duchi et al., 2015; Salimans et al., 2017), coordinate-wise finite differences (FD) (Kiefer & Wolfowitz, 1952), and adaptive random search (Rechenberg, 1973) — when every evaluation of $\mathcal{E}$ is corrupted by i.i.d. (independent and identically distributed) measurement noise; (ii) *why* ES behaves as it does under noise, via a derived and validated error-floor law; and (iii) what fraction of each method's residual error is recoverable by cost-free iterate averaging, which isolates the coherent (dither-ripple) character of the ES residual.

Why a fixed-target protocol. Distributional metrics cannot score this regime: for the temperature-one target, the stationary energy gap is χ^2-distributed to excellent accuracy (the modes are ≥ 6 component widths apart), so a perfect *sampler* has $\mathbb{E}\,\Delta\mathcal{E} = d/2$ by equipartition and essentially never enters a fixed energy ball — e.g. $\mathbb{P}(\Delta\mathcal{E} < \frac{1}{2}) = \mathbb{P}(\chi^2_{16} < 1) \approx 6 \times 10^{-8}$ at $d = 16$. Optimizers and samplers must therefore be benchmarked on different axes, and here we adopt the fixed-target (hitting-time) methodology standard in black-box optimization (Hansen et al., 2021): fix an accuracy target, record when each chain first reaches it, and report quantiles of the hitting-time distribution. The Langevin samplers of D.5 are correspondingly excluded from this table by construction, not by defeat.

Success criterion, dwell, and hitting times. With $\mathcal{E}^\star = \min_m \mathcal{E}(\boldsymbol{\mu}_m)$ (all four modes are equivalent), define the energy gap

$$\Delta\mathcal{E}(\mathbf{x}) \;=\; \mathcal{E}(\mathbf{x}) - \mathcal{E}^\star, \qquad \Delta\mathcal{E}(\mathbf{x}) \;\approx\; \tfrac{1}{2}\,\|\mathbf{x} - \boldsymbol{\mu}_{m(\mathbf{x})}\|^2 \;\text{ near a mode,} \tag{83}$$

so the target $\Delta\mathcal{E} < \varepsilon$ with $\varepsilon = \frac{1}{2}$ is a ball of radius $\approx \sqrt{2\varepsilon} = 1$ (one component width) around some mode, in every dimension. Let e count the noisy oracle evaluations a method has consumed for a given chain (its only clock), and let $\mathbf{x}(e)$ be that chain's state, piecewise constant between the method's updates. The per-chain hitting time is the first entry into a visit that *persists*,

$$\tau \;=\; \min\Big\{\, e \;:\; \Delta\mathcal{E}\big(\mathbf{x}(e')\big) < \varepsilon \;\text{ for all } e' \in [e,\, e + D]\,\Big\}, \qquad D = 20 \text{ evaluations}, \tag{84}$$

where the dwell requirement D disqualifies transits in which a large dither excursion merely slices through the ball without settling (without it, hitting times can be gamed by amplitude). We report the success rate $\mathbb{P}(\tau \leq B)$ over chains at the total per-chain budget B, and the median hitting time E_{50} (the empirical 50% quantile of τ, in evaluations), quoting E_{50} only when at least half the chains succeed. Scoring uses the *exact* $\mathcal{E}$ — available because the testbed is synthetic — and is not charged to any budget; the methods themselves never see it.

Noisy oracle. Every method interacts with the energy only through the corrupted oracle

$$\widehat{\mathcal{E}}(\mathbf{x}) \;=\; \mathcal{E}(\mathbf{x}) + \sigma_J\, n, \qquad n \sim \mathcal{N}(0,1) \text{ i.i.d. per evaluation and per chain,} \tag{85}$$

at noise levels $\sigma_J \in \{0.03,\, 0.1,\, 0.3,\, 1.0\}$; for reference, the initial gaps are $\Delta\mathcal{E} \approx 6d$ and the target is $\varepsilon = \frac{1}{2}$, so a single read at $\sigma_J = 0.1$ already carries noise equal to 20% of the target resolution. This is the synthetic stand-in for a mini-batch or measurement-limited evaluation of a model's energy.

Readout: last iterate vs. iterate averaging. Under a noisy oracle, scoring the final iterate of a stochastic trajectory is a high-variance convention; the standard remedy in stochastic approximation is Polyak–Ruppert iterate averaging (Ruppert, 1988; Polyak & Juditsky, 1992), whose deep-learning analogue is weight averaging (Izmailov et al., 2018). We therefore maintain, purely as a *readout*, an exponential moving average (EMA) in evaluation time,

$$\bar{\mathbf{x}} \leftarrow \bar{\mathbf{x}} + \lambda\,(\mathbf{x} - \bar{\mathbf{x}}), \qquad \lambda = \min(1,\, \Delta e/\tau_{\mathrm{f}}), \qquad \tau_{\mathrm{f}} = 300 \text{ evaluations}, \tag{86}$$

updated whenever the method updates (Δe = evaluations since the last update), and score $\Delta\mathcal{E}(\bar{\mathbf{x}})$. Three properties keep this fair: it is applied *identically* to every method; it costs zero evaluations (it reuses the trajectory already paid for); and it is feedback-free ($\bar{\mathbf{x}}$ never enters any method's dynamics). Its one side effect is a warm-up lag: the average retains weight $\approx e^{-(B-e_{\mathrm{conv}})/\tau_{\mathrm{f}}}$ on pre-convergence states, so all EMA hitting times carry a common offset of order $\tau_{\mathrm{f}}\ln(\Delta\mathcal{E}_0/\varepsilon)$, and a method that converges only near the end of its budget can show a *larger* EMA error than raw error. The clean-gradient oracle row below, which converges in ≤ 40 updates, measures this lag directly ($E_{50} = 514$–1211 across d) and serves as the ruler; entries dominated by warm-up rather than by an error floor are marked †.

Baselines: updates, per-update cost, and tuning. All baselines consume the oracle Eq. (85) and pay their true evaluation price per update; their step counts are the budget divided by that price.

- *Adaptive random search* (the (1+1) evolution strategy, i.e. one-parent/one-offspring; equivalently zero-temperature random-walk Metropolis with step adaptation). Proposal $\mathbf{x}' = \mathbf{x} + s\,\boldsymbol{\xi}$, $\boldsymbol{\xi} \sim \mathcal{N}(0, I_d)$; accept iff $\widehat{\mathcal{E}}(\mathbf{x}') < \widehat{\mathcal{E}}_{\mathrm{inc}}$; per-chain step adaptation by the 1/5th-success rule $s \leftarrow s\,e^{\eta(\mathbf{1}_{\mathrm{acc}} - 1/5)}$ (Rechenberg, 1973), with $s_0 = 0.5$, $\eta = 0.1$. *Cached-incumbent* variant: the incumbent value is measured once and stored (1 evaluation per proposal). *Re-evaluating* variant: the incumbent is re-measured at every comparison (2 evaluations per proposal), the standard remedy for the elitist noise pathology (Arnold & Beyer, 2002) in which a single lucky noise draw makes the cached incumbent unbeatable.
- *FD-GD* (central-difference gradient descent) (Kiefer & Wolfowitz, 1952):

$$\hat{g}_i = \frac{\widehat{\mathcal{E}}(\mathbf{x} + h\mathbf{e}_i) - \widehat{\mathcal{E}}(\mathbf{x} - h\mathbf{e}_i)}{2h}, \qquad \mathbf{x} \leftarrow \mathbf{x} - \eta_{\mathrm{g}}\,\hat{\mathbf{g}}, \tag{87}$$

  $2d$ evaluations per update, $\eta_{\mathrm{g}} = 0.1$. The divided difference amplifies oracle noise to $\sigma_J/(\sqrt{2}\,h)$ per gradient component, so h must scale with the noise: we use $h = 0.5$ (the bias is $O(h^2)$ and negligible on the near-quadratic modes); $h = 10^{-4}$, appropriate for a clean oracle, diverges immediately at any $\sigma_J > 0$.
- *ZO-GD* (GS gradient estimate with m antithetic pairs) (Nesterov & Spokoiny, 2017; Salimans et al., 2017):

$$\hat{\mathbf{g}} = \frac{1}{m}\sum_{j=1}^{m} \mathbf{u}_j\,\frac{\widehat{\mathcal{E}}(\mathbf{x} + \nu\mathbf{u}_j) - \widehat{\mathcal{E}}(\mathbf{x} - \nu\mathbf{u}_j)}{2\nu}, \qquad \mathbf{u}_j \sim \mathcal{N}(0, I_d), \tag{88}$$

  then $\mathbf{x} \leftarrow \mathbf{x} - \eta_{\mathrm{g}}\hat{\mathbf{g}}$; $2m$ evaluations per update, $m = 10$, $\eta_{\mathrm{g}} = 0.05$, $\nu = \max(0.1, \sigma_J)$ (probe scaled to the noise, the standard bias–variance remedy).
- *Two-point Gaussian-smoothing estimator (SPSA-style)*: Eq. (88) with $m = 1$, the one-random-direction scheme in the spirit of Spall (1992) (classical SPSA uses Rademacher directions, which satisfy its bounded-inverse-moment condition, and decaying gains; our variant uses Gaussian directions and constant gains, so it is a two-point Gaussian-smoothing gradient estimator rather than SPSA proper). The single-probe estimator is unbiased but heavy-tailed: on a unit-curvature quadratic, one step contracts the squared error by $1 - 2\eta_{\mathrm{g}} + \eta_{\mathrm{g}}^2(d+2)$, so stability requires $\eta_{\mathrm{g}} < 2/(d+2)$ with one-step optimum $\eta_{\mathrm{g}} = 1/(d+2)$; we use $\eta_{\mathrm{g}} = 1/d$ throughout, which is essential — a d-independent step ($\eta_{\mathrm{g}} = 0.05$) diverges for $d \geq 64$. With this tuning SPSA is the strongest derivative-free baseline in the tables below, consistent with the information-theoretic optimality of two-point estimators for this query model (Duchi et al., 2015; Jamieson et al., 2012).
- *Clean-gradient descent* (reference only): $\mathbf{x} \leftarrow \mathbf{x} + \eta_{\mathrm{g}}\nabla \log p(\mathbf{x})$, $\eta_{\mathrm{g}} = 0.1$, using the *exact, noiseless* gradient. It is not purchasable in the zeroth-order query model at any price and is reported as an upper bound and as the EMA-lag ruler.

ES configuration. Algorithm 1 in the convention Eq. (77) with $g = 1$, dither amplitude $\alpha = 1.5$, base frequency $\omega = 100$, band $N = 1.75$ (ratios r_i evenly spaced in $[1, 1.75]$), $c = 5$ steps per fastest dither period, hence $\Delta t = 2\pi/(cN\omega) \approx 7.18 \times 10^{-3}$ and one oracle evaluation per step; diffusion $\sigma = 10^{-3}$ (the numerical mode-seeking limit); and the dimension-dependent feedback gain

$$k(d) \;=\; \frac{4}{\sqrt{d}}. \tag{89}$$

The scaling Eq. (89) is forced by the same mechanism as the threshold Eq. (82): the phase carries $k\mathcal{E}$, and $\mathcal{E}$ is extensive — both its fluctuations under the target (std $\approx \sqrt{d/2}$) and its per-step increments far from the modes grow with d — so at fixed k the per-step phase modulation grows with dimension until it scrambles the demodulation. Keeping $k\sqrt{d}$ constant keeps it bounded. Furthermore: (i) $c = 5$ sits below the $c \geq 10$ recommendation of Sec. 3.2, which concerns the trajectory-error *rate measurement*; here the criterion is the end-task hitting time and the setting is flagged rather than re-derived; (ii) the initialization is $\mathbf{x}_0 \sim \mathrm{Unif}([-6, 6]^d)$ (dispersed across all four basins; initial gap $\Delta\mathcal{E}_0 \approx 6d$), not the Gaussian initialization of D.1. We run both the base variant (d distinct frequencies) and the *quadrature-pair* variant of Eq. (11) ($\lceil d/2 \rceil$ carriers, each driving a $(\cos, \sin)$ coordinate pair at phases $0, \pi/2$).

The averaged dynamics predict the ES constants. On the near-quadratic basins, the averaged flow gives $\frac{d}{dt}\Delta\mathcal{E} = -k\alpha\,\Delta\mathcal{E}$, hence the parameter-free prediction

$$E_{50} \;\approx\; \frac{\ln(\Delta\mathcal{E}_0/\varepsilon)}{k\alpha\,\Delta t} \;=\; \frac{\ln(12d)\,\sqrt{d}}{4 \cdot 1.5 \cdot \Delta t}. \tag{90}$$

Against the clean-oracle, last-iterate measurement (single replicate, 2000 chains): predicted vs. observed $E_{50} = 180/155, 488/502, 781/856, 1234/1493$ at $d = 4, 16, 32, 64$ — the discretized dither runs at the speed of its own averaged flow to within $\pm 15\%$ for $d \leq 32$, with a 21% finite-gain drag appearing at $d = 64$. Under noise the *rate* is nearly unchanged, as the phase-averaging claim of Sec. 3.2 predicts: from $\sigma_J = 0$ to 0.1 the raw E_{50} moves $155 \to 175$, $502 \to 536$, $856 \to 925$ at $d = 4, 16, 32$ ($+13/+7/+8\%$; single matched-stream replicates), while — for contrast — the same noise destroys both random-search variants at $d \geq 32$ (Table 17).

Budgets, replication, and evaluation accounting. The per-chain budget is identical for every method: $B = 2000$ evaluations for $d \leq 64$, 4000 at $d = 128$, 8000 at $d = 256$. Headline tables are mean$\pm$sd (standard deviation) over 5 replicates of 500 chains; a replicate re-draws the initialization, every method's internal randomness, and the oracle noise stream, while *within* a replicate all methods share the same initialization and the same noise stream (common random numbers, so method-to-method comparisons are paired). Total evaluations is the fair currency for ES specifically because its evaluations are inherently *sequential* (each step's evaluation feeds the next state), whereas the $2m$ probes of ZO-GD, the $2d$ of FD, and the 2 of SPSA are embarrassingly parallel across evaluators; wall-clock comparisons on parallel hardware would favor the batched estimators, and we say so rather than hide it. The converse structural advantage of the one-evaluation-per-step design — every coordinate updates after every single measurement, with no staleness window — is exercised by the time-varying benchmarks of D.7, not by this static one.

Findings at low and moderate noise (Tables 16–17). Tuned SPSA is the fastest derivative-free method at every dimension, as its query-model optimality leads one to expect; we state that plainly. The ES result is of a different kind: it reaches the target at *every* $d \leq 128$ at both noise levels (quadrature variant: success $\geq 97.4\%$ everywhere), with E_{50} within 1.1–$1.5\times$ of SPSA's (1.10–$1.43\times$ for the quadrature variant across $d = 4$–128), while operating under the strictly sequential one-evaluation-per-step constraint that is the point of the design. Every other baseline breaks somewhere in this range: cached-incumbent random search noise-locks (its stored incumbent value is a lucky order statistic that fresh proposals cannot beat (Arnold & Beyer, 2002)) already at $(d, \sigma_J) = (32, 0.03)$; the re-evaluating repair pays $2\times$ per step and still fails for $d \geq 64$, because accepting on noisy comparisons turns the elitist walk into a diffusion whose drift vanishes once the per-step improvement falls below the $\sqrt{2}\,\sigma_J$ comparison noise; FD-GD is budget-starved by its $2d$-evaluation tax (≤ 15 updates at $d \geq 64$) on top of its noise amplification; and ZO-GD with $m = 10$ is squeezed by budget quantization ($\lfloor B/20 \rfloor$ updates — visible as the 74–79% success at $d = 64$, where it affords only 100 updates). The quadrature variant is never worse than the base variant and is consistently

Table 16: Fixed-target benchmark at $\sigma_J = 0.03$, EMA readout Eq. (86): success rate in % (mean±sd over 5 replicates of 500 chains; sd 0.0 omitted) with the median hitting time E_{50} in evaluations in parentheses (seed-to-seed sd of $E_{50} \leq 18$ evaluations in every entry shown); "—" where fewer than half the chains succeed. Budget $B = 2000$ ($d \leq 64$), 4000 (d=128), 8000 (d=256). All EMA hitting times include the common filter warm-up measured by the clean-gradient row (which converges in ≤ 40 updates). †: warm-up artifact of the filtered readout only — under the last-iterate readout FD-GD succeeds at d=32 (99.9%, final gap 0.267; Table 18) but its $\lfloor B/2d \rfloor = 31$ updates arrive too late in the budget for the $\tau_f = 300$ filter to settle; at $d \geq 64$ FD-GD fails under both readouts.

Method (evals/update)	d=4	d=16	d=32	d=64	d=128	d=256
rand. search, cached (1)	100 (526)	99.1±0.3 (884)	39.3±2.6 (—)	0	0	0
rand. search, re-eval (2)	100 (530)	100 (908)	100 (1346)	0	0	0
FD-GD ($2d$)	100 (589)	100 (1248)	0.2±0.1 (—)†	0	0	0
ZO-GD, m=10 (20)	100 (1060)	100 (1460)	100 (1660)	79.0±0.8 (1920)	100 (2440)	100 (4784)
SPSA, $\eta_g = 1/d$ (2)	100 (526)	100 (811)	100 (967)	100 (1246)	100 (2111)	100 (4303)
bounded ES (1)	100 (608)	100 (1036)	99.8±0.1 (1322)	95.9±0.7 (1797)	99.4±0.4 (3084)	47.7±1.0 (—)
bounded ES, quadrature (1)	100 (580)	100 (1002)	100 (1278)	99.1±0.7 (1723)	99.9±0.1 (2927)	78.0±0.9 (7506)
clean-grad. descent (oracle)	100 (514)	100 (782)	100 (893)	100 (1000)	100 (1106)	100 (1211)

Table 17: As Table 16, at $\sigma_J = 0.1$ (seed-to-seed sd of $E_{50} \leq 16$ evaluations except re-evaluating random search at d=32, ±48).

Method (evals/update)	d=4	d=16	d=32	d=64	d=128	d=256
rand. search, cached (1)	99.8±0.2 (540)	30.5±1.7 (—)	0	0	0	0
rand. search, re-eval (2)	100 (530)	100 (948)	56.4±3.5 (1878)	0	0	0
FD-GD ($2d$)	100 (586)	100 (1248)	0.1±0.1 (—)†	0	0	0
ZO-GD, m=10 (20)	100 (1060)	100 (1464)	100 (1672)	74.0±0.9 (1940)	100 (2480)	100 (5228)
SPSA, $\eta_g = 1/d$ (2)	100 (532)	100 (818)	100 (980)	100 (1269)	100 (2173)	100 (4485)
bounded ES (1)	100 (620)	100 (1054)	99.8±0.2 (1349)	90.7±1.2 (1850)	97.4±1.2 (3328)	2.5±1.0 (—)
bounded ES, quadrature (1)	100 (590)	100 (1016)	100 (1300)	98.0±0.7 (1768)	99.7±0.2 (3104)	12.5±1.3 (—)
clean-grad. descent (oracle)	100 (514)	100 (782)	100 (893)	100 (1000)	100 (1106)	100 (1211)

better in the tails (e.g. 99.1 vs. 95.9% at $(64, 0.03)$; 78.0 vs. 47.7% at $(256, 0.03)$), consistent with its elimination of the nearest-beat cross terms at fixed band. At $d = 256$ the current configuration reaches its measured boundary: the dither-envelope floor derived below ($\approx 0.0028\, d = 0.72$ at $d = 256$) exceeds ε under the raw readout, the gain schedule Eq. (89) crosses the finite-gain ceiling discussed below, and $E_{50} = 7506$ presses the $B = 8000$ budget — three compounding effects, all of which the frequency-scaling prescription of D.3 (raise ω with d) is designed to relieve; we report the boundary rather than re-tune past it.

High-noise regime and the limits of the comparison. At $\sigma_J = 0.3$ the ES success degrades (98.3/94.7% at $d = 16$ for the quadrature/base variants, 48.1±4.4/31.3±2.6% at $d = 32$, zero beyond), and at $\sigma_J = 1$ ES fails at all d while SPSA and ZO-GD retain 100% success with essentially σ-independent hitting times. Two mechanisms, one per side. On the ES side, the per-evaluation phase jitter is $k\sigma_J$ radians: once it approaches $O(1)$ the demodulation itself decoheres — with Eq. (89), $k\sigma_J \approx 0.71$ at (d=32, σ_J=1) — and, below that, the injected noise floor derived next crosses ε. On the baseline side, the smoothing estimators were granted the standard noise adaptation $\nu = \max(0.1, \sigma_J)$, which on this testbed is free of charge: the modes are near-quadratic, so enlarging the probe adds no smoothing bias — a testbed-specific gift that would expire on landscapes with curvature variation at the probe scale. ES, by contrast, ran a single gain schedule Eq. (89) across all σ_J; the symmetric remedy — reducing k with σ_J to hold $k\sigma_J$ bounded, the ES analogue of SPSA's noise-scaled probe — is the natural extension and is left to future work. The honest summary of the regime map: at $k\sigma_J \lesssim 0.3$ the lock-in structure delivers near-noiseless rates; well beyond it, noise-adapted stochastic smoothing dominates.

The ES residual is coherent (Table 18). The readout study answers a question the headline tables cannot: *what kind* of error does each method leave behind? Averaging the trajectory can only remove zero-mean scatter about the right place; it does nothing for a chain that never arrived. The table shows exactly this taxonomy. The noise-locked random search is untouched by the filter (0.575 → 0.572): its error is dynamical, and the experiment thereby also certifies that the noise is *not* being filtered

Table 18: Identical-trajectory readout study at $d = 32$, $B = 2000$: final median gap $\Delta\mathcal{E}$ under the last-iterate (raw) readout $\rightarrow$ under the EMA readout Eq. (86), for the *same* runs (the two scorers observe one trajectory per method per replicate). Mean over 5 replicates of 500 chains; seed-to-seed sd ≤ 0.01 for entries below 1 and $\leq 3\%$ of the mean for entries above 1. The raw$\rightarrow$EMA ratio measures each method's *recoverable* error — zero-mean scatter about the optimum — as opposed to dynamical failure, which no readout repairs. †: EMA exceeds raw due to residual filter warm-up weight $e^{-(B-e_{\text{conv}})/\tau_f}$ on pre-convergence states (FD-GD and ZO-GD converge late in their update-sparse budgets), not an error floor.

Method	σ_J=0.03	σ_J=0.1	σ_J=0.3	σ_J=1.0
rand. search, cached	0.575 $\rightarrow$ 0.572	1.91 $\rightarrow$ 1.90	5.66 $\rightarrow$ 5.64	19.0 $\rightarrow$ 19.0
rand. search, re-eval	0.532 $\rightarrow$ 0.177	1.78 $\rightarrow$ 0.572	5.37 $\rightarrow$ 1.73	18.0 $\rightarrow$ 5.82
FD-GD	0.267 $\rightarrow$ 0.756†	0.282 $\rightarrow$ 0.766†	0.413 $\rightarrow$ 0.856†	1.91 $\rightarrow$ 1.89
ZO-GD, m=10	0.017 $\rightarrow$ 0.106†	0.037 $\rightarrow$ 0.117†	0.037 $\rightarrow$ 0.117†	0.039 $\rightarrow$ 0.123†
SPSA	0.023 $\rightarrow$ 0.005	0.260 $\rightarrow$ 0.045	0.260 $\rightarrow$ 0.045	0.264 $\rightarrow$ 0.047
bounded ES	0.113 $\rightarrow$ 0.015	0.224 $\rightarrow$ 0.070	1.25 $\rightarrow$ 0.609	25.8 $\rightarrow$ 21.6
bounded ES, quadrature	0.108 $\rightarrow$ 0.009	0.218 $\rightarrow$ 0.060	1.25 $\rightarrow$ 0.563	25.3 $\rightarrow$ 21.1

before the optimizers — if it were, every method would revert to its clean numbers. SPSA improves $4.6\times$ at $\sigma_J = 0.03$ (its residual is a weakly correlated stochastic wander). ES improves $7.5\times$ (base) and $12.0\times$ (quadrature, $0.108 \rightarrow 0.009$): its residual is dominated by the deterministic dither ripple — a zero-mean, phase-locked oscillation of known period (5–9 steps here) that any window much longer than a beat cancels coherently. This is the same structure exploited by a lock-in amplifier's output low-pass filter (Scofield, 1994): the dithered phase demodulation of Sec. 3.2 is the mixer, and Eq. (86) is the missing output stage. The practical consequence is that the correct deployment of the mode-seeking configuration reports the averaged iterate, at zero additional evaluations — which is also what converts the $d = 128$ column of Table 17 from a raw-readout floor sitting near ε into a 97–100% success rate.

A quantitative error-floor model. The stationary gap of the ES iterates (raw readout) decomposes into three terms:

$$F(d, \sigma_J) \;\approx\; \underbrace{\frac{\alpha}{4}\sum_{i=1}^{d}\frac{1}{\omega_i}}_{F_{\text{osc}} \approx 0.0028\, d} \;+\; F_{\text{dec}}(k, d) \;+\; \underbrace{C_n\, k\, \sigma_J^2\, d}_{F_{\text{noise}}}, \tag{91}$$

in which each term has a mechanism. F_{osc} is the energy stored in the dither ripple itself: to leading order $x_i \approx \bar{x}_i + \sqrt{\alpha/\omega_i}\sin(\omega_i t + k\mathcal{E})$ (cf. A.2), so each coordinate carries variance $\alpha/(2\omega_i)$ and, on the unit-curvature basin Eq. (83), gap $\alpha/(4\omega_i)$; for our band this is $0.00280\, d$, independent of k. F_{noise} is noise-injected heating: each evaluation kicks the phase by $k\sigma_J n$, hence the velocity by $\approx -\sqrt{\alpha\omega_i}\,\sin(\omega_i t + k\mathcal{E})\, k\sigma_J n$, an injected position variance of $\frac{1}{2}\alpha\omega_i(k\sigma_J)^2\Delta t^2$ per step per coordinate; balancing against the averaged relaxation (rate $k\alpha/2$ in position) gives $F_{\text{noise}} \approx \frac{1}{4}\bar{\omega}\Delta t\, k\sigma_J^2 d$ with $\bar{\omega}\Delta t \approx 0.99$ here, i.e. the stated form with $C_n \approx 0.25$ predicted; the constant fitted once, globally, is $C_n = 0.55$ — the derivation captures the functional form (linear in k and d, quadratic in σ_J) and the order of the constant. F_{dec} is the finite-gain decoherence residual, the term the schedule Eq. (89) is chosen to suppress; empirically it is the small remainder of the clean floors above F_{osc} (below) and scales $\propto k^2$ across the two gain settings we tested at $d = 128$. Table 19 confronts Eq. (91) with measurement: the noise-induced excess matches the derived law within 3% for $32 \leq d \leq 128$ at $\sigma_J = 0.1$, and within its error bars across $\sigma_J \in \{0.03, 0.1, 0.3\}$ at $d = 32$; at $\sigma_J = 1$ the law underpredicts by $\approx 2\times$, precisely where the small-jitter expansion behind it fails ($k\sigma_J \approx 0.7$ rad) — the same boundary as the success collapse above, giving the validity condition $k\sigma_J \lesssim 0.3$ a second, independent measurement.

Finite-gain ceiling. The counterpart of the frequency threshold Eq. (82) on the gain side: sweeping k at fixed (α, c, ω) (single replicate, 500 chains, clean oracle, dither diffusion $\sigma = 0.02$, $B = 2000$, raw readout), convergence is reliable up to a ceiling that halves as d doubles — at $d = 32$, $k = 1$ gives $E_{50} = 713$ (97.0% success; quadrature 696, 99.2%) while $k = 2$ collapses to 2.8%; at $d = 64$, $k = 0.5$ gives $E_{50} = 1521$ (99.4%) while $k = 1$ collapses to 15.0% (20.4% quadrature). That is,

Table 19: Validation of the floor model Eq. (91) under the raw (last-iterate) readout. Panel (a): dimension scan at $\sigma_J = 0.1$ (single matched-stream replicate, 2000 chains: the clean and noisy runs share initialization; the gain schedule Eq. (89) is in force, so the prediction is $0.55\,k(d)\,\sigma_J^2\,d = 2.2\,\sigma_J^2\sqrt{d}$). Panel (b): noise scan at $d = 32$ (5 replicates of 500 chains, mean±sd; excess relative to the clean floor 0.102). The σ_J=1 row lies outside the small-jitter validity of the law ($k\sigma_J \approx 0.71$) and is included to locate the breakdown.

(a) dimension scan, $\sigma_J = 0.1$

d	clean floor	F_{osc} pred.	noisy floor	excess (obs.)	excess (pred.)
4	0.013	0.011	0.048	0.035	0.044
16	0.042	0.045	0.124	0.082	0.088
32	0.102	0.090	0.223	0.121	0.124
64	0.240	0.179	0.416	0.176	0.176
128	0.456	0.358	0.712	0.256	0.249

(b) noise scan, $d = 32$ (5 seeds)

σ_J	noisy floor	excess (obs.)	excess (pred.)
0.03	0.113 ± 0.001	0.011	0.011
0.1	0.224 ± 0.002	0.122	0.124
0.3	1.252 ± 0.013	1.150	1.120
1.0	25.8 ± 0.26	25.7	12.4

$k_{\max}\,d \approx 30$–60 for this configuration, consistent with the per-step phase-slew mechanism above ($k\,|\Delta\mathcal{E}$ per step$|$ growing $\propto k\,d$ in the far field). The schedule Eq. (89) has $k(d)\,d = 4\sqrt{d}$, which sits below this ceiling for $d \lesssim 100$, grazes it near $d \approx 128$ (the observed 21% rate drag), and exceeds it at $d = 256$ — one of the three compounding boundary effects noted above. Below the ceiling, the rate is linear in k as Eq. (90) predicts ($E_{50} = 1182 \to 713$ for $k = 0.5 \to 1$ at $d = 32$, i.e. 83% of exact $1/k$ scaling).

Scope. We state the limits of this benchmark plainly, as in D.5. The testbed is smooth with near-quadratic, unit-curvature basins, which grants the smoothing baselines their free noise-scaled probes and makes tuned SPSA close to its query-model optimum; none of the results here claim ES superiority on clean, smooth objectives, and the parallelism asymmetry noted in the accounting paragraph favors the batched estimators on parallel hardware. What the benchmark does establish, with five-seed error bars: (i) a strictly sequential, one-evaluation-per-step method reaches the fixed target at every $d \leq 128$ under evaluation noise, within 1.1–1.5× the evaluation cost of the strongest (noise-tuned, parallelizable) baseline, where every other sequential-friendly baseline fails; (ii) the noise-robustness mechanism claimed in Sec. 3.2 is quantitative — rates within 8–13% of noiseless up to $k\sigma_J \approx 0.2$, with the residual cost captured by the derived floor law Eq. (91) to a few percent over an order of magnitude in d and in σ_J^2; and (iii) the ES residual is coherent rather than stochastic, so a zero-cost averaged readout recovers up to 12× accuracy — the lock-in signature that distinguishes the dither mechanism from stochastic smoothing.

D.7 CelebA-HQ: steady-state lag and paired comparison

This appendix supplements the CelebA-HQ tracking study of Sec. 3.5: the first paragraph reads the checkpoint-by-checkpoint NMSE of Table 2 (target drift vs. bounded ES) as a quantitative measurement of the tracking mechanism, and the second returns to the matched-budget benchmark of Table 4 to explain the ranking it reports.

Steady-state-lag signature. The checkpoint increments expose the tracking mechanism quantitatively. Between consecutive checkpoints the target's NMSE advances by $\Delta_{\text{drift}} = .053, .068, .075, .084, .095, .110$ (the drift accelerates), while the tracker's residual advances by $\Delta_{\text{ES}} = .009, .010, .011, .012, .015, .016$; the ratio $\Delta_{\text{ES}}/\Delta_{\text{drift}} = .170, .147, .147, .143, .158, .145$ is constant at 0.152 ± 0.010 across all six intervals, after a standing-start first checkpoint at level ratio 0.40. A constant incremental ratio is the signature of steady-state lag tracking: the residual is proportional to the target's local speed with a fixed effective response time, the stochastic analogue

of the deterministic ES tracking bounds of (Scheinker & Scheinker, 2016), and it implies the $6\times$ headline reduction is not a plateau but improves whenever the target moves more slowly.

Tuning improves the baselines considerably, SPSA most of all, yet bounded ES remains the strongest forward-only method: $1.8\times$ lower final NMSE than best-tuned SPSA, $3.5\times$ lower than FD, $15\times$ lower than GS-ES, and within $1.7\times$ of the exact-gradient oracle. Because all methods are run on the same 219 pairs and seeds, the comparison is paired: *bounded ES achieves the lowest final NMSE among the four forward-only methods on* 100% *of the* 219 *pairs* — $219/219$ head-to-head wins against each of SPSA, FD, and GS-ES — while outperforming even the exact-gradient oracle on $\approx 9\%$ of pairs (exact instantaneous gradients are not optimal for a moving target: Adam's momentum can overshoot or lag the target's turns, and on those pairs the ES trajectory ends closer). The ranking is explained by the update rate against a moving target rather than by per-estimate variance. Bounded ES commits an update after every single measurement (8192 updates); SPSA after every 2 (4096); GS-ES only after $2m$ consecutive measurements — its antithetic pairs are evaluated at different times against a moving target, so temporal drift contaminates the spatial differences, and the tuned m trades estimator variance against staleness with no good operating point at this drift rate: its final NMSE (1.17) exceeds the targets' own final drift (≈ 0.51, Table 2), i.e., it ends farther from the target than the untouched initial image. Beyond the raw evaluation-count tax (cf. Table 15), coordinate FD suffers the structural staleness penalty in its purest form: with $d = 256$ parameters, central differences must hold the state essentially fixed for 512 measurements to assemble one gradient estimate, so its entire budget purchases 16 steps, each applying the gradient of a target that no longer exists — the estimate answers a question the system is no longer asking — whereas ES updates every coordinate after every single measurement.

E CIFAR-10 Additional Details

Latents are standardized per channel to zero mean and standard deviation 0.5 with the statistics of $45{,}000$ training latents ($5{,}000$ are held out for validation). On the $10{,}000$ test images the VAE reconstructs with PSNR $= 28.2$ dB, SSIM $= 0.90$ Wang et al. (2004) and a mean absolute error of 0.060 on the $[-1, 1]$ pixel scale, so the decoder's own reconstruction error is about $2.6\times$ smaller than the tracker's final-frame error against the decoded targets reported below. Encoder and decoder use exponential-moving-average weights, and every image in the tracking study is a decoder output $D_\theta(z)$, so the targets lie on the decoder's image manifold by construction.

The latent diffusion U-Net has 17.0 M parameters; base width 128, two residual blocks per level, multi-head self-attention at every level and at bottleneck.

F Dimension dependence of the finite-frequency threshold and the evaluation cost of ES

Theorem 1 is dimension-free *in form*: for every (δ, ε, T) there is a finite ω^*, whatever the state dimension. All dependence on the dimension is compressed into $\omega^*(\delta, \varepsilon, T)$ through the constant C'_R of Eq. (46), which silently sums the $O(d^2)$ cross-coordinate oscillatory integrals of Step 2, and through the frequency gaps $|r_n - r_i|$ entering the constants of Lemma 1. Read naively, the theory appears to work equally well in any dimension; in experiments the finite-ω threshold grows linearly with d (Eq. (82), Table 11). This appendix closes the gap between those two statements. We prove: (i) an explicit finite-ω error bound whose constants carry the dimension polynomially, so that a *sufficient* threshold $\omega^*(d)$ is polynomial in d (Proposition 4); (ii) a matching-in-d *lower* bound: for the evenly spaced frequency design there are C^∞ targets for which no $\omega \lesssim d/T$ suffices, so $\omega^*(d) = \Omega\big(d/((r_{\max} - 1)T)\big)$ (Proposition 6), the rigorous counterpart of the measured law Eq. (82); and (iii) the consequence the practitioner feels: combined with the step-resolution requirement of Algorithm 1, the number of model evaluations needed to track the averaged dynamics grows at least linearly in d, for *every* choice of the frequency band (Theorem 5). The averaged limit itself never sees the dimension; the entire cost is paid through the exchange rate ω between the fast and slow clocks of Eq. (80).

Throughout this appendix d denotes the state dimension (the N of Appendix A.1), and the frequency ratios are $\{r_i\}_{i=1}^d \subset [1, r_{\max}]$; to avoid a notation collision we write $r_{\max}$ for the top of the

band, denoted N in Appendix D.3. For the evenly spaced design, adjacent ratios differ by $\delta_r = (r_{\max} - 1)/(d-1)$, so the smallest beat frequency is $\Delta_{\min} = \omega\delta_r$ as in Eq. (81).

F.1 Dimension-normalized setting

Comparing constants across dimensions is meaningful only if the regularity of the target is normalized per coordinate. We therefore work, in this appendix, under the following strengthening of Assumptions 1–3, chosen to cover the sampling instantiations (I), (I$_\lambda$), (III) of Sec. 3.3 and the synthetic testbed of Appendix D.1.

Assumption 6 (Dimension-normalized regularity; constant diffusion)**.** $\mathbf{f} \equiv \mathbf{0}$ and $g_n \equiv g > 0$ for all n (constant scalar diffusion, $\boldsymbol{\sigma} = gI$). $J \in C^{2,1}(\mathbb{R}^d \times [0,T])$ with the *global* bounds

$$\sup_{\mathbf{x},t} |\partial_i J| \le G_1 \ \ \forall i, \qquad \sup_{\mathbf{x},t} \|\nabla_{\mathbf{x}} J\|_2 \le \bar{G}_1, \qquad \sup_{\mathbf{x},t} \|\nabla^2_{\mathbf{x}} J\|_{\mathrm{op}} \le H_2, \qquad \sup_{\mathbf{x},t} |\partial_t J| \le G_t. \tag{92}$$

The four primitives are deliberately kept separate: G_1, H_2, and G_t are per-coordinate (dimension-robust) quantities, while $\bar{G}_1 \le \sqrt{d}\, G_1$ may grow with d in the worst case but is $O(1)$ for targets whose gradient is carried by $O(1)$ coordinates. On the Gaussian-mixture testbed Eq. (76), restricted to any fixed ball, all four are bounded uniformly in d. Global (rather than compact-set) bounds let us drop the localization τ_R and keep the radius R out of the dimension accounting; on compacts, every statement below holds verbatim with the primitives evaluated on $\bar{B}_R \times [0,T]$ and the stopping time of Eq. (26) reinstated.

F.2 A frequency-resolved oscillatory bound

The first step is to expose the two hidden dependences of Lemma 1: on the oscillation frequency $|c|\omega$ (through the block averages) and on the dimension (through the speed $\|\boldsymbol{\mu}_\omega\|_2 = O(\sqrt{\alpha\omega r_{\max} d}\,)$ of the dithered state and the size $O(\sqrt{d\Delta})$ of its Brownian increments).

Lemma 7 (Oscillatory integrals, explicit in $|c|$ and d)**.** *Let Assumption 6 hold, let $\Phi : \mathbb{R}^d \times [0,T] \to \mathbb{R}$ be bounded by C_Φ and jointly Lipschitz with constant L_Φ (Euclidean norm in $\mathbf{x}$), and let $c \neq 0$. Then for all $\omega \ge 1$,*

$$\mathbb{E}\Big[\sup_{t\le T}\Big|\int_0^t \Phi\big(\mathbf{x}_\omega(s), s\big)\, e^{\mathrm{i}c\omega s}\, ds\Big|^2\Big] \ \le\ \kappa\,(1+d)\Big(1+\tfrac{1}{|c|}\Big)\,\omega^{-1/2}, \tag{93}$$

$$\kappa := c_0(1+T)^2\big[C_\Phi^2 + L_\Phi^2 g^2(1+\alpha r_{\max})\big], \tag{94}$$

with c_0 a universal constant. If Φ depends on $\mathbf{x}$ only through a fixed subset S of coordinates, then $(1+d)$ may be replaced by $(1+|S|)$.

Proof. Repeat the proof of Lemma 1 with the block length chosen per frequency, $\Delta = |c|^{-1/2}\omega^{-3/4}$ (for $\Delta > T$ or ω so small that the right side of Eq. (94) exceeds $(C_\Phi T)^2$, the claim is trivial after enlarging c_0). The frozen part obeys $\sup_t |A|^2 \le 8C_\Phi^2 T^2/(\Delta^2 c^2 \omega^2) = 8C_\Phi^2 T^2/(|c|\,\omega^{1/2})$. For the Lipschitz part, with $\mathbf{f} \equiv 0$ the drift is the dither alone, so on each block $\|\mathbf{x}_s - \mathbf{x}_{t_j}\| \le g\sqrt{\alpha\omega r_{\max} d}\,\Delta + \|\mathbf{N}_j(s)\|$ with $\mathbb{E}\|\mathbf{N}_j\|^2 \le g^2 d\,\Delta$; if Φ depends only on coordinates S, the same holds with d replaced by $|S|$ because only the increments of those coordinates enter. Exactly as in Lemma 1,

$$\begin{aligned}\mathbb{E}\sup_{t\le T}|B|^2 &\le 3L_\Phi^2 T^2\big[g^2\alpha r_{\max}\, d\,\omega\Delta^2 + \Delta^2\big] + 3L_\Phi^2 T^2 g^2 d\,\Delta \\ &\le 3L_\Phi^2 T^2 g^2(1+\alpha r_{\max})\, d\Big[\tfrac{1}{|c|\,\omega^{1/2}} + \tfrac{1}{|c|^{1/2}\omega^{3/4}}\Big] + 3L_\Phi^2 T^2\Delta^2.\end{aligned} \tag{95}$$

Since $\omega \ge 1$, $|c|^{-1/2}\omega^{-3/4} \le \frac{1}{2}(|c|^{-1}\omega^{-1/2} + \omega^{-1}) \le (1+|c|^{-1})\omega^{-1/2}$, and $\Delta^2 \le \omega^{-1/2}$; collecting terms gives Eq. (94). □

F.3 Upper bound: a sufficient threshold polynomial in d

Define the *spectral crowding* of a ratio set,

$$\mathfrak{S}(r) \;:=\; \max_{1\le n\le d} \sum_{i\neq n} \frac{1}{\sqrt{|r_n - r_i|}}, \qquad \mathfrak{S}(r) \;\le\; \frac{4d}{\sqrt{r_{\max}-1}} \quad \text{(evenly spaced)}, \tag{96}$$

the bound following from $\sum_{i\neq n}|r_n-r_i|^{-1/2} \le 2\delta_r^{-1/2}\sum_{g=1}^{d-1} g^{-1/2} \le 4\sqrt{(d-1)/\delta_r}$. $\mathfrak{S}$ is the quantity through which the frequency design enters the error: it is small when the ratios are few or well separated, and grows when d frequencies crowd a fixed band.

Proposition 4 (Explicit finite-ω error bound)**.** *Let Assumptions 2 and 6 hold, let $\mathbf{x}_\omega$ and $\bar{\mathbf{x}}$ be as in Theorem 1, and suppose the moment bound $\mathbb{E}\sup_{t\le T}\|\bar{\mathbf{x}}\|^2 < \infty$ of Assumption 3. There is a constant $Q = Q(k,\alpha,g,r_{\max},T,G_1,\bar{G}_1,H_2,G_t)$, independent of d and ω, such that for all $\omega \ge 1$,*

$$\mathbb{E}\Big[\sup_{t\le T}\big\|\mathbf{x}_\omega(t)-\bar{\mathbf{x}}(t)\big\|^2\Big] \;\le\; Q\,e^{\Gamma T}\Big[\underbrace{d\,\big(1+d^2H_2^2+\bar{G}_1^4\big)\,\omega^{-1}}_{\text{ripple, boundary, martingale}} + \underbrace{d\,(1+d)\,\big(\mathfrak{S}(r)+d\big)^2\,\omega^{-1/2}}_{\text{beat pairs}}\Big], \tag{97}$$

with $\Gamma := 3(T+4)\big(k\alpha g^2 H_2/2\big)^2$. Consequently, for the evenly spaced design, a sufficient condition for $\mathbb{P}\big(\sup_{t\le T}\|\mathbf{x}_\omega-\bar{\mathbf{x}}\|>\delta\big)\le\varepsilon$ is

$$\omega \;\ge\; \omega^*_{\text{suff}}(d) \;=\; C_{\text{suff}}\,\frac{d^8}{(r_{\max}-1)^2\,\delta^4\varepsilon^2}, \tag{98}$$

with $C_{\text{suff}} = C_{\text{suff}}(k,\alpha,g,r_{\max},T,G_1,\bar{G}_1,H_2,G_t)$. The exponent 8 treats the four primitives as d-free, which holds for coordinate-sparse gradients ($\bar{G}_1 = O(1)$); for dense isotropic energies with $\bar{G}_1 = \Theta(\sqrt{d})$ the Lipschitz constant $L_\Phi \lesssim H_2 + |k|G_1\bar{G}_1$ of the slow factors grows like $\sqrt{d}$, the beat-pair term of Eq. (97) acquires an extra factor d through κ, and the sufficient exponent becomes 10 (and 5 rather than 4 under Remark 8).

Proof. We retrace Steps 1–3 of Appendix A.4, tracking the dimension. *Step 1 terms.* With g constant, $u^c = g\cos kJ$, $u^s = g\sin kJ$, so $\|\nabla u\|_2 \le kg\bar{G}_1$ and $|\operatorname{tr}(\boldsymbol{\sigma}\boldsymbol{\sigma}^\top\nabla^2 u)| \le g^3k\big(dH_2 + k\bar{G}_1^2\big)$ (the trace of a symmetric matrix with operator norm H_2 is at most dH_2). Following the estimates leading to Eq. (38), each coordinate obeys $\mathbb{E}\sup_t|E_n^{(1)}|^2 \le C\big(1+d^2H_2^2+\bar{G}_1^4+G_t^2\big)\,\omega^{-1}$ with $C = C(k,\alpha,g,r_{\max},T)$, and summing over n multiplies by d; this is the first bracket of Eq. (97). *Step 2 terms.* For each ordered pair (n,i) the nonresonant integrals of Eq. (39) have envelopes $\Phi \in \{\partial_i u^{c/s}\, g\cos kJ,\ \partial_i u^{c/s}\, g\sin kJ\}$ with $C_\Phi \le kg^2G_1$ and $L_\Phi \le Ckg^2(H_2 + kG_1\bar{G}_1)$, oscillating at the frequencies $(r_n \mp r_i)\omega$ (and $2r_n\omega$ for $i=n$), with prefactor $\alpha\sqrt{r_i/r_n} \le \alpha\sqrt{r_{\max}}$. Lemma 7 and the triangle inequality in $L^2(\sup_t)$ give

$$\begin{aligned}\Big\|\sup_t|E_n^{(2)}|\Big\|_{L^2} \;&\le\; C\sqrt{\kappa}\,(1+d)^{1/2}\,\omega^{-1/4}\Big[\sum_{i\neq n}\Big(1+\tfrac{1}{|r_n-r_i|}\Big)^{1/2} + \sum_{i\neq n}\Big(1+\tfrac{1}{r_n+r_i}\Big)^{1/2} + 2\Big]\\ &\le\; C'\sqrt{\kappa}\,(1+d)^{1/2}\big(\mathfrak{S}(r)+d\big)\,\omega^{-1/4},\end{aligned} \tag{99}$$

and summing squares over n contributes the second bracket. *Step 3.* With $\mathbf{f}\equiv 0$ and $\boldsymbol{\sigma} = gI$ the averaged drift is $\bar{\boldsymbol{\mu}} = -\frac{k\alpha g^2}{2}\nabla J$, whose Euclidean Lipschitz constant is exactly $\frac{k\alpha g^2}{2}H_2$ — crucially free of d under the operator-norm bound of Eq. (92) — and the diffusion difference vanishes. The Grönwall step of Eq. (46) then multiplies the remainder by $3e^{\Gamma T}$ with Γ as stated, giving Eq. (97). For Eq. (98), insert Eq. (96), note that the second bracket dominates for $\omega \ge 1$ up to a constant, apply Chebyshev's inequality, and absorb numerical factors into C_{suff}. □

Remark 8 (Sharper ω-power under one more derivative). If in addition $J \in C^3$ with $\sup_{\mathbf{x},t}\|\nabla_{\mathbf{x}}\nabla^2_{\mathbf{x}}J\|_{\text{op}}$ bounded, the per-pair estimate can be run through the Itô integration by parts of Remark 1 instead of the partition argument, replacing $(1+|c|^{-1})\omega^{-1/2}$ in Eq. (94) by $(1+|c|^{-2})\omega^{-1}$ up to constants involving $\|\nabla J\|_1 \le dG_1$. Since $\sum_{i\neq n}|r_n-r_i|^{-1} \le 2(d-1)(1+\ln(d-1))/(r_{\max}-1)$ for the evenly spaced design, this trades powers between d and ω and yields the alternative sufficient threshold $\omega^*_{\text{suff}} \le C\,d^4(1+\ln d)^2/\big((r_{\max}-1)^2\delta^2\varepsilon\big)$. We do not optimize further: the role of Proposition 4 is to certify that the hidden constants are *polynomial* in d, and Remark 11 explains why the true threshold is smaller still.

F.4 Lower bound: the beat obstruction is real

Distinct frequencies (Assumption 2) are not a technicality. We first record what happens when they fail, then turn the same mechanism, quantitatively, into a lower bound on ω^*.

Proposition 5 (A coincident pair changes the limit)**.** *Let Assumption 6 hold and suppose $r_n = r_m$ for one pair $n \neq m$, all other ratios pairwise distinct and distinct from r_n. Then the conclusion of Theorem 1 holds with the averaged SDE Eq. (22) replaced by*

$$d\bar{\mathbf{x}} = -\tfrac{k\alpha g^2}{2}\, M\, \nabla_{\bar{\mathbf{x}}} J(\bar{\mathbf{x}}, t)\, dt + g\, d\mathbf{w}, \qquad M := I + \mathbf{e}_n\mathbf{e}_m^\top + \mathbf{e}_m\mathbf{e}_n^\top. \tag{100}$$

M is singular, with eigenvalue 0 along $(\mathbf{e}_n - \mathbf{e}_m)/\sqrt{2}$. In particular, for instantiation (I) with the quadratic energy $\mathcal{E}(\mathbf{x}) = \frac{1}{2}\|\mathbf{x}\|^2$ the modified dynamics have no invariant probability law*: along $\mathbf{e}_n - \mathbf{e}_m$ the drift vanishes identically and the state performs free Brownian motion.*

Proof. Rerun Step 2 of Appendix A.4 for coordinate n. The term $i = m$ of Eq. (39) now oscillates at the frequencies 0 and $2r_n\omega$, exactly as the term $i = n$ does; retaining its zero-frequency half and applying the same algebra that produced Eq. (42) (with ∂_m in place of ∂_n; the ∂g terms are absent since g is constant) yields, in addition to Eq. (42), the resonant contribution $-\frac{k\alpha g^2}{2}\int_0^t \partial_m J(\mathbf{x}_\omega(s), s)\, ds$, and symmetrically for coordinate m. All remaining estimates are unchanged, giving Eq. (100). The spectral claim is immediate; for the quadratic energy, $M\nabla\mathcal{E}(\mathbf{x}) = M\mathbf{x}$ vanishes on $\ker M$, so the $(\mathbf{e}_n - \mathbf{e}_m)$-component of the state solves $d\zeta = \sqrt{2}\, g\, d\beta$ for a scalar Brownian motion β, which is null recurrent. □

At finite ω with a small but nonzero gap, the extra term of Proposition 5 does not vanish; it merely oscillates at the beat frequency $\Delta = \omega|r_n - r_m|$, and until the beat has completed a fraction of a cycle the trajectory cannot distinguish the correct limit from Eq. (100). Quantifying this yields the lower bound. The construction is deliberately *two-body*: a linear target supported on a single coordinate makes every cross term except the designated pair vanish identically, so the obstruction can be isolated with dimension-free constants.

Proposition 6 ($\Omega(d)$ frequency threshold for the evenly spaced design)**.** *Fix $k, \alpha, g, T > 0$, $\gamma > 0$, and $r_{\max} > 1$, and let the ratios be evenly spaced on $[1, r_{\max}]$. Let $J(\mathbf{x}, t) = \gamma x_1$, which satisfies Assumption 6 with $G_1 = \bar{G}_1 = \gamma$, $H_2 = G_t = 0$. Set*

$$\delta_0 := \frac{k\alpha g^2 \gamma\, T}{2\pi\sqrt{r_{\max}}}\,. \tag{101}$$

There exist constants $\omega_1, V < \infty$ and $\varepsilon_0 := \delta_0^2/(4V) > 0$, all independent of d and ω, such that for every $d \geq d_0$ and every

$$\omega_1 \;\leq\; \omega \;\leq\; \frac{\pi\,(d-1)}{2T\,(r_{\max} - 1)} \tag{102}$$

one has $\mathbb{P}\big(\sup_{t\leq T} \|\mathbf{x}_\omega(t) - \bar{\mathbf{x}}(t)\| > \delta_0/2\big) \geq \varepsilon_0$. Consequently

$$\omega^*(\delta, \varepsilon, T) \;\geq\; \frac{\pi\,(d-1)}{2T\,(r_{\max} - 1)} \qquad \text{for all } \delta < \delta_0/2,\ \varepsilon < \varepsilon_0,\ d \geq d_0, \tag{103}$$

where $d_0 := 1 + \lceil 2T(r_{\max} - 1)\,\omega_1/\pi \rceil$. The threshold of Theorem 1 therefore grows at least linearly in the dimension at fixed band, matching the measured law Eq. (82) up to the constant.

Proof. Take the adjacent pair $(1, 2)$, so $\Delta := \omega(r_2 - r_1) = \omega(r_{\max} - 1)/(d-1)$, and let $\mathbf{x}_\omega, \bar{\mathbf{x}}$ share the Brownian motion and start from $\mathbf{x}_0 = \mathbf{0}$. Since $\partial_i J = 0$ for $i \geq 2$, the averaged dynamics of coordinate 2 are pure noise, $d\bar{x}_2 = g\, dw_2$, and $u^c = g\cos(k\gamma x_1)$, $u^s = g\sin(k\gamma x_1)$ depend on $\mathbf{x}$ only through x_1, with $\partial_i u^{c/s} = 0$ for $i \geq 2$.

Apply Steps 1–2 of Appendix A.4 to coordinate $n = 2$. Because ∇u has a single nonzero component, every constant in the Step-1 estimate Eq. (38) is independent of d, and in Step 2 the sum over i in Eq. (39) reduces to the single term $i = 1$. Expanding its trigonometric products, the sum-frequency parts oscillate at $(r_1 + r_2)\omega \geq 2\omega$ and the self term of coordinate 2 at $2r_2\omega$; all are covered by Lemma 7 with envelopes depending on $\mathbf{x}$ only through x_1 (so with $(1 + d)$ replaced by 2) and with

$C_\Phi \le k\gamma g^2$, $L_\Phi \le k\gamma g^2(1+k\gamma)$. The difference-frequency part is special: writing $S = \sin k\gamma x_1$, $C = \cos k\gamma x_1$, the two groups of Eq. (39) combine to

$$k\alpha g^2\gamma\sqrt{\tfrac{r_1}{r_2}}\int_0^t \Big[SC\,\sin\big((\omega_1+\omega_2)s\big) - \tfrac{1}{2}\big(S^2+C^2\big)\cos\big(\Delta s\big) + \tfrac{1}{2}\big(S^2-C^2\big)\cos\big((\omega_1+\omega_2)s\big)\Big]ds, \tag{104}$$

in which the identity $S^2+C^2=1$ removes the state from the beat term entirely: the slow contribution is the *deterministic* integral $-\frac{k\alpha g^2\gamma}{2}\sqrt{r_1/r_2}\int_0^t \cos(\Delta s)\,ds$. Collecting, there is a d-free constant $C_4 = C_4(k,\alpha,\gamma,g,r_{\max},T)$ with

$$x_{\omega,2}(t) - \bar{x}_2(t) \;=\; -\frac{k\alpha g^2\gamma}{2}\sqrt{\frac{r_1}{r_2}}\int_0^t \cos(\Delta s)\,ds \;+\; \mathcal{R}(t), \qquad \Big\|\sup_{t\le T}|\mathcal{R}(t)|\Big\|_{L^2} \le C_4\,\omega^{-1/4}. \tag{105}$$

Under Eq. (102), $\Delta T \le \pi/2$, so $\int_0^T \cos(\Delta s)\,ds = \sin(\Delta T)/\Delta \ge (2/\pi)T$, and $\sqrt{r_1/r_2} \ge r_{\max}^{-1/2}$; hence $|\mathbb{E}[\,x_{\omega,2}(T) - \bar{x}_2(T)\,]| \ge 2\delta_0 - C_4\omega^{-1/4} \ge \delta_0$ once $\omega \ge \omega_1 := (C_4/\delta_0)^4$. Moreover the deterministic term in Eq. (105) is bounded by $\bar{m} := \frac{k\alpha g^2\gamma}{2}T$ pathwise, so $\mathbb{E}\big[(x_{\omega,2}(T) - \bar{x}_2(T))^2\big] \le V := 2\bar{m}^2 + 2C_4^2$. Writing $X = |x_{\omega,2}(T) - \bar{x}_2(T)|$, the Paley–Zygmund inequality gives $\mathbb{P}(X \ge \frac{1}{2}\mathbb{E}X) \ge (\mathbb{E}X)^2/(4\,\mathbb{E}X^2) \ge \delta_0^2/(4V) = \varepsilon_0$, and since $\sup_{t\le T}\|\mathbf{x}_\omega - \bar{\mathbf{x}}\| \ge X$ pathwise, the displayed probability bound follows. Eq. (103) is its contrapositive against the definition of $\omega^*(\delta,\varepsilon,T)$ in Theorem 1. □

Remark 9 (Phase averaging of the Brownian ripple). The proof extends verbatim to any diffusion level because the beat term in Eq. (105) is deterministic; the Brownian motion influences only $\mathcal{R}$. Had the beat phase retained the state (as the sum-frequency terms do), the noise in x_1 would multiply the mean by the Gaussian phase-averaging factor $e^{-k^2\gamma^2 g^2 s/2}$ — the same mechanism by which measurement noise entering the ES phase is demodulated rather than amplified (Appendix D.6). The cancellation $S^2 + C^2 = 1$ that removes the state from the beat phase is thus what makes the obstruction robust at every noise level.

F.5 Evaluation complexity: convergence speed versus dimension

Propositions 4 and 6 bound the frequency; the practitioner pays in *model evaluations*. Algorithm 1 performs one evaluation of J per step and must resolve the fastest dither: with $c \ge 1$ steps per fastest period, $\Delta t \le 2\pi/(c\,r_{\max}\omega)$, so the number of evaluations on the horizon $[0,T]$ is $\mathrm{NFE} = T/\Delta t \ge (c/2\pi)\,r_{\max}\,\omega\,T$ (the validated choice throughout this paper is $c = 10$; Appendix D.2).

Theorem 5 (Linear-in-dimension evaluation cost)**.** *In the setting of Proposition 6, consider a run of Algorithm 1 with evenly spaced ratios on $[1, r_{\max}]$, any $r_{\max} > 1$, any horizon $T > 0$, and $c \ge 1$ steps per fastest dither period, simulating the continuous process Eq. (20) at a base frequency $\omega \ge \omega_1$ in the averaging regime of Proposition 6. If that continuous process achieves*

$$\mathbb{P}\Big(\sup_{t\le T}\|\mathbf{x}_\omega(t) - \bar{\mathbf{x}}(t)\| > \delta_0/2\Big) < \varepsilon_0 \tag{106}$$

for the target of Proposition 6 with $d \ge d_0$ (equivalently, $T \le \pi(d-1)/(2(r_{\max}-1)\omega_1)$, so that the window Eq. (102) is nonempty), then the run must use

$$\mathrm{NFE} \;\ge\; \frac{c}{2\pi}\,r_{\max}\,\omega\,T \;\ge\; \frac{c}{4}\,(d-1)\,\frac{r_{\max}}{r_{\max}-1} \;\ge\; \frac{c}{4}\,(d-1)\,. \tag{107}$$

The bound is algebraically independent of the horizon and of the band: widening $[1, r_{\max}]$ relaxes the frequency threshold Eq. (103) and tightens the step size in exact proportion, so no choice of the band escapes the linear cost; the horizon enters only through the applicability condition $d \ge d_0(T)$. The restriction $\omega \ge \omega_1$ excludes the sub-threshold regime $\omega < \omega_1$, which Proposition 6 does not analyze (there the remainder bound in Eq. (105) exceeds δ_0, and the dither is too slow to average at all); the accuracy condition is placed on the continuous process, and the step count is what Algorithm 1 needs to resolve it at c steps per period. Conversely, by Proposition 4 (or Remark 8), a polynomial budget $\mathrm{NFE} = O\big(\mathrm{poly}(d)\,T/(\delta^p\varepsilon^q)\big)$ always suffices under Assumption 6.

Proof. By Proposition 6, since $\omega \ge \omega_1$ and the window Eq. (102) is nonempty for $d \ge d_0$, the accuracy requirement forces $\omega > \pi(d-1)/(2T(r_{\max}-1))$; insert this into $\mathrm{NFE} \ge (c/2\pi)r_{\max}\omega T$ and use $r_{\max}/(r_{\max}-1) > 1$. □

Theorem 5 is the precise sense in which "convergence speed depends on dimension" for stochastic bounded ES, and it is invisible at the level of Theorem 1: the *averaged* dynamics, and hence the physical settling time, are dimension-free (measured directly in Appendix D.4); what grows with d is the number of scalar evaluations needed per unit of averaged time, because ω — the exchange rate between the fast and slow clocks of Eq. (80) — must itself grow linearly with d to keep the beat spectrum resolvable. Equivalently, in the language of lock-in detection (Scofield, 1994): the band $[\omega, r_{\max}\omega]$ must host d carrier channels whose pairwise beats each complete $\Omega(1)$ cycles within the horizon, a bandwidth budget $(r_{\max} - 1)\,\omega \geq \pi(d-1)/(2T)$ that is exactly Eq. (103).

Remark 10 (Consistency with zeroth-order information bounds). A linear-in-d evaluation cost is not an artifact of the dither design: any procedure that extracts a d-dimensional descent direction from one scalar evaluation per step is subject to $\Omega(d)$-type oracle lower bounds (Jamieson et al., 2012; Duchi et al., 2015), the same accounting that gives simultaneous-perturbation and smoothing estimators their dimension factors (Spall, 1992; Nesterov & Spokoiny, 2017). What Proposition 6 and Theorem 5 add is *where* the universal cost surfaces in ES — not in the number of evaluations per gradient estimate, which remains one, but in the frequency budget of the beat spectrum — and that the constants of the mechanism match the measured design law Eq. (82) and Table 11.

Remark 11 (The gap between the bounds, and the measured law). Between the sufficient threshold $\omega^*_{\text{suff}} = O\big(d^4\text{polylog}(d)\big)$ of Remark 8 and the necessary threshold $\Omega(d)$ of Proposition 6, the measurements of Appendix D.3 place the truth at the lower end: $\omega^*(d) \approx O(10)\,(d-1)/(r_{\max}-1)$, linear up to the mildly superlinear drift of the transition centers in Table 10. The lossy step in Proposition 4 is identifiable: the triangle inequality over the $O(d^2)$ beat pairs treats their oscillations as if they could align, whereas distinct beat frequencies are mutually incoherent; replacing the triangle inequality by the orthogonality heuristic replaces $\mathfrak{S}(r)^2$ by $\sum_{i\neq n} |r_n - r_i|^{-1} = O\big(d \log d/(r_{\max} - 1)\big)$ and predicts $\omega^* = O(d\,\text{polylog}(d))$, consistent with the measured law — and the growing number of near-minimal gaps supplies the observed superlinear correction. Converting the heuristic into a proof requires a joint (rather than pairwise) treatment of the beat family; we record it as an open problem. None of this affects the conclusions used in the main text, which rely only on the two proved directions: the threshold is finite and polynomial (Proposition 4) and at least linear (Proposition 6), and the evaluation cost of ES grows at least linearly with dimension (Theorem 5).

G Additional CelebA-HQ tracking results

Figure 6 shows additional time-varying tracking runs on CelebA-HQ 256×256 under the 1-D projection objective.

H λ-tempering example

We numerically demonstrate the λ-tempering result of Eq. 9 with a numerical example as shown in Figure 7.

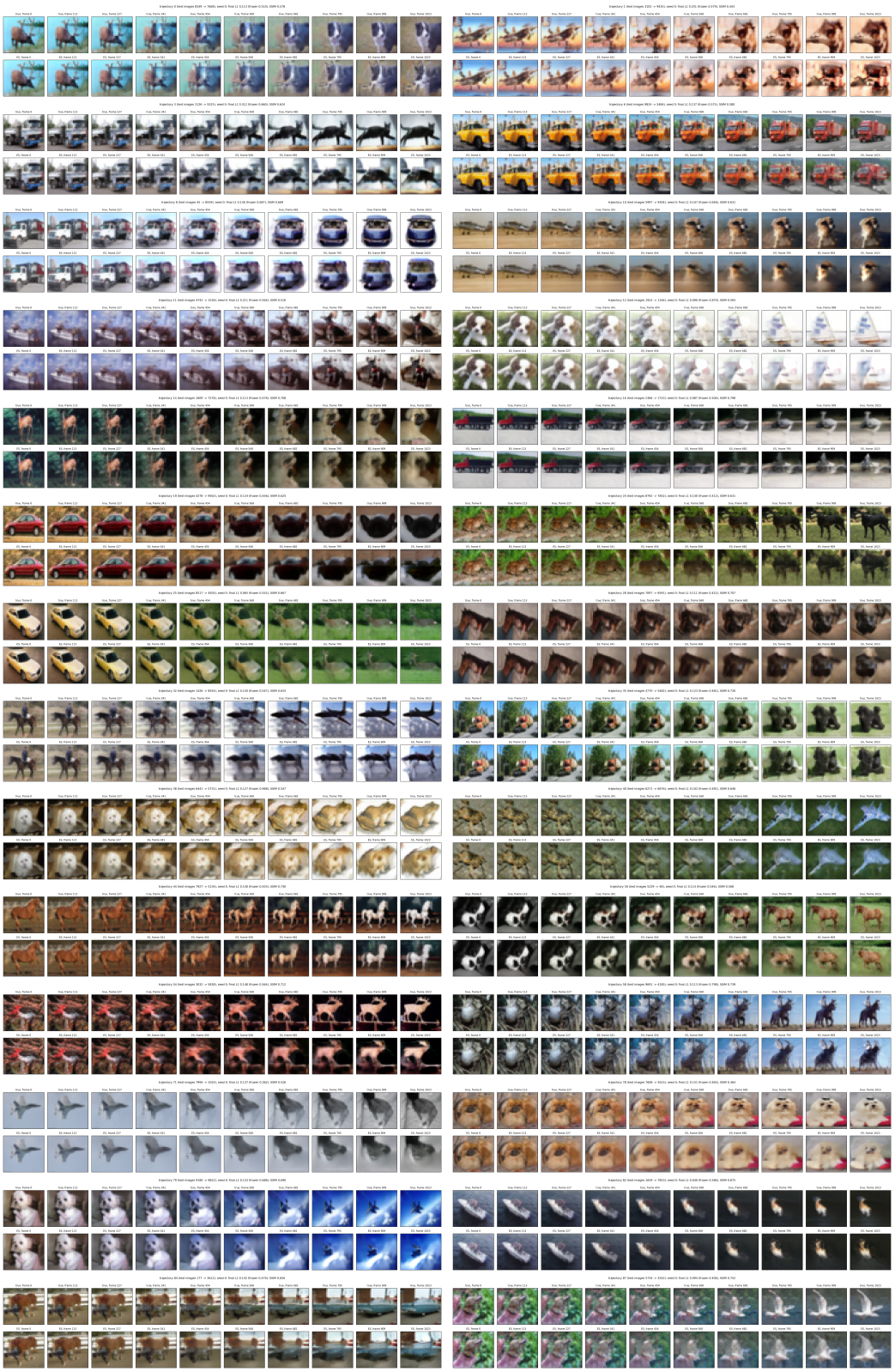


Figure 5: ES dynamics results for tracking time-varying CIFAR-10 images based on 1D projections.

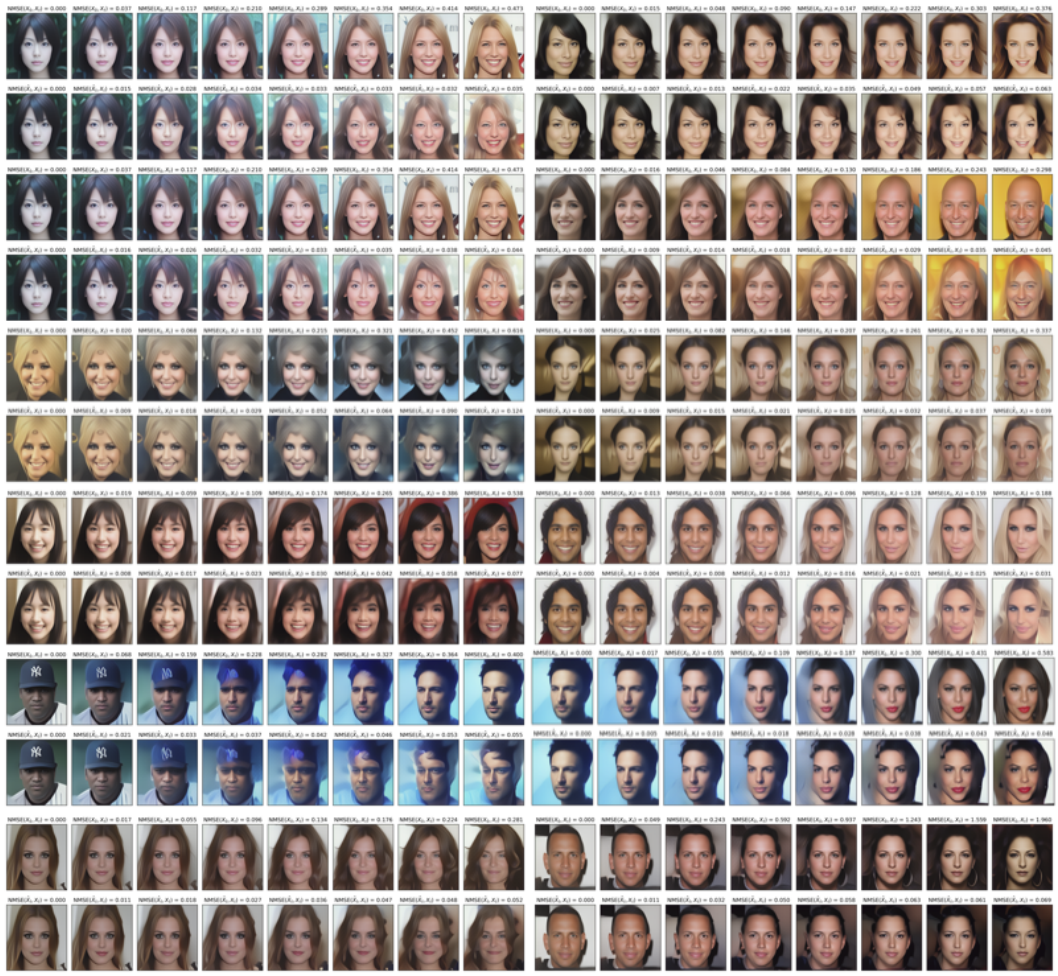


Figure 6: Additional CELEBA-HQ tracking examples.

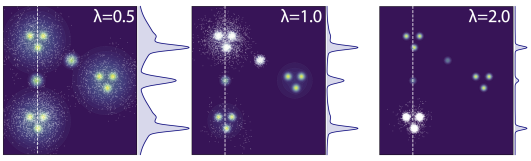


Figure 7: Effect on ρ^{λ} of λ-tempering ES sampling of 2D multimodal target (white dots: visited locations). As λ decreases, local peaks and valleys are escaped, and the sampler becomes mass-covering; large λ concentrates on modes.