

Subpolynomial query complexity for well-conditioned log-concave sampling

An expository account of the proof, with consequences for diffusion models

A rewrite, for human readers, of the manuscript

Subpolynomial query complexity for well-conditioned log-concave sampling (OpenAI, September 26, 2026)

Prepared with an AI assistant · October 11, 2026

Abstract

In September 2026 OpenAI posted a manuscript, written by an internal model, proving that the optimal dimension exponent for sampling well-conditioned log-concave densities in the exact first-order oracle model is zero. For every fixed $\varepsilon > 0$ the manuscript’s algorithm uses at most $C_\varepsilon d^\varepsilon$ evaluations of the potential’s gradient on *every* execution, for C^2 potentials on $\mathbb{R}^d$ with a known minimizer and Hessian between I_d and $2I_d$. It also proves an $\Omega(\log d)$ lower bound for arbitrary randomized adaptive algorithms. The manuscript is complete but compressed. The machinery is developed in the order the proof needs it, and the reader is rarely told *why* a step is there. This note has two purposes. The first is expository. It presents the same proof for a reader who knows the elements of log-concave sampling (Langevin discretization, functional inequalities, Wasserstein comparison) but has not read the manuscript. It explains what each step is for, carries out the central calculations in full (the all-split tensor estimates, the material-derivative bookkeeping, the stationary centering flow, the noisy finite evaluation of nested conditional means, and the Fourier-type lower-bound argument), and sketches the parts that adapt known machinery, with pointers to the original. The second purpose is to draw out what the result means for diffusion models, where the score is normally learned from samples. The denoising score and denoiser of the smoothed target are exactly the conditional mean the construction evaluates, every noise level can be accessed with subpolynomially many first-order queries, and the manuscript’s sampler is an ancestral denoiser whose convergence theorem is the upper bound. It is an exposition, not a referee report.

Contents

1	Introduction	2
1.1	The statement	2
1.2	Where things stood, and what was missing	3
1.3	The proof on one page	4
1.4	What is new and what is inherited	6
1.5	How this note is organized	7
2	The near-quadratic sampling problem	8
2.1	Notation and the auxiliary theorem	8
2.2	Convexity, moments, and comparisons	9
2.3	Coupling and smoothing conventions	11

3	Gaussian transport and tensor derivatives	12
3.1	Cumulants from the Poincaré inequality	12
3.2	The transport equation	15
4	Dimension bounds for material derivatives	16
4.1	Material derivatives of the conditional jets	16
4.2	Contractions and the expressions they generate	17
4.3	One final vector dimension factor	21
5	A stationary flow for noisy means	22
5.1	Transport Jacobians and the zero endpoint	23
5.2	A kernel for the centered observable	24
5.3	Stationarity and the path integral	25
5.4	Derivatives on a stationary trajectory	27
6	Finite formulas for the two flows	28
6.1	Parameters and the order of choices	29
6.2	Probability transport by global Picard iteration	29
6.3	The kernel action through harmonic trajectories	30
6.4	The zero endpoint of the kernel integral	32
6.5	A globally Lipschitz approximate centering velocity	33
6.6	The mean-flow circuit and the accuracy ledger	34
6.7	The exact circuit interface and its derivation	36
7	Finite evaluation of nested conditional means	39
7.1	Gaussian center noise and the absorption identity	39
7.2	The two routines and two kinds of depth	40
7.3	Evaluating an expression at a noisy center	41
7.4	Query-preserving shifts and the deterministic cap	42
7.5	Terminal errors and propagation of scales	43
7.6	Stability for arbitrary fixed seeds	45
7.7	Standalone accuracy and the auxiliary theorem	45
8	From noisy conditional samples to the target law	47
8.1	The contracting noised chain	47
8.2	Conditional descent and its ideal histories	49
8.3	The Gaussian base rule	51
8.4	Query counts and all dimensions	52
9	A logarithmic oracle lower bound	53
9.1	Adaptive queries and unrevealed rotations	53
9.2	A statistic invisible on the block span	55
9.3	Separation from the target Gaussian	56
10	Consequences for diffusion-model sampling	58
10.1	Scores of smoothed laws are conditional means	58
10.2	Score and denoiser evaluation in subpolynomial query complexity	59
10.3	The sampler is an ancestral denoiser	60
10.4	Scope, limitations, and open directions	60

1 Introduction

How many evaluations of a potential and its gradient are needed to sample from its Gibbs distribution? Even when the potential is uniformly strongly convex and smooth, the standard discretizations of Langevin and Hamiltonian dynamics produce sampling errors that grow with the dimension. The question is sharpest when the conditioning is fixed and the target accuracy is a constant: is a positive power of the dimension intrinsically necessary, or is it an artefact of the sampling algorithms people happen to use?

The manuscript answers this in the exact first-order oracle model, allowing arbitrary real computation between queries, and shows that the optimal dimension exponent is zero. The construction is aimed at oracle complexity rather than practical running time: its approximation orders may depend on the desired exponent ε , and its constants, while finite, are enormous.

1.1 The statement

For $d \geq 2$, let $\mathcal{V}_d$ consist of the functions $V \in C^2(\mathbb{R}^d)$ satisfying

$$V(0) = 0, \quad \nabla V(0) = 0, \quad I_d \preceq \nabla^2 V(x) \preceq 2I_d \quad \text{for every } x \in \mathbb{R}^d. \quad (1.1)$$

The target law is the Gibbs law

$$\pi_V(dx) = Z_V^{-1} e^{-V(x)} dx, \quad Z_V = \int_{\mathbb{R}^d} e^{-V(x)} dx.$$

The origin is a known minimizer: $\nabla V(0) = 0$ and $\nabla^2 V \succeq I_d$ force it. An *exact first-order query* at a point x returns the pair $(V(x), \nabla V(x))$. Queries may be adaptive, random real variables may be generated, and computation between queries is unrestricted. An algorithm may depend on d , but its only information about V is through the oracle responses. All algorithms are measurable.

Write $\|\mu - \eta\|_{\text{TV}} = \sup_B |\mu(B) - \eta(B)|$ for total variation, and let $Q(d)$ be the smallest integer q such that some algorithm makes at most q queries *on every execution* and returns X with

$$\|\mathcal{L}(X) - \pi_V\|_{\text{TV}} \leq \frac{1}{10} \quad \text{for every } V \in \mathcal{V}_d.$$

The budget is therefore a worst-case cap on every execution, not an expected number of queries. The optimal dimension exponent is

$$\gamma_* = \inf\{\gamma \geq 0 : Q(d) \leq d^{\gamma+o(1)}\} = \limsup_{d \rightarrow \infty} \frac{\log Q(d)}{\log d}.$$

Theorem 1.1 (Subpolynomial query complexity). In the oracle model above, for every $\varepsilon > 0$ there is $C_\varepsilon < \infty$ such that

$$Q(d) \leq C_\varepsilon d^\varepsilon \quad (d \geq 2).$$

There are also absolute constants $c > 0$ and d_0 such that

$$Q(d) \geq c \log d \quad (d \geq d_0).$$

The lower bound applies to arbitrary randomized adaptive first-order algorithms. In particular $\gamma_* = 0$.

The upper assertion is proved in Theorem 8.5 and the lower assertion in Theorem 9.1. Together they determine the exponent but not the finer growth rate of $Q(d)$: the upper bound does *not* assert a polylogarithmic query count, and there is a wide gap between d^ε and $c \log d$. The constants and the algorithmic approximation orders may depend on ε . No arithmetic or bit-complexity bound is imposed or deduced.

Three features of the model. (i) Queries are *exact*: a query returns the true value and gradient, so all error comes from the algorithm’s own approximations. (ii) The budget is a *uniform* cap: the algorithm must fix, in advance, a number of queries that never fails, which is why later sections work with deterministic query caps rather than expected complexity. (iii) Computation between queries is *free*: constants such as the algorithm’s approximation order may depend on ε without affecting the theorem, and no numerical-stability or precision issues need to be discussed.

1.2 Where things stood, and what was missing

Sampling by discretizing Langevin or Hamiltonian dynamics turns functional inequalities and smoothness into algorithmic bounds, but it also introduces a dimension-dependent discretization error. Nonasymptotic analyses of the unadjusted Langevin algorithm were developed by Dalalyan [15] and Durmus–Moulines [16]. Cheng, Chatterji, Bartlett and Jordan [12] analysed the underdamped method under strong convexity and smoothness, and randomized midpoint discretization further improved the quadratic-Wasserstein guarantees of Shen and Lee [22]. Wasserstein and total-variation guarantees must be distinguished when comparing these results.

The best general-purpose bounds before the manuscript were the following, all at fixed conditioning:

- $\tilde{O}(d^{1/3})$ at fixed total-variation accuracy, from the shifted-composition analysis of Altschuler, Chewi and Zhang [3] (a relative-entropy guarantee converted by Pinsker), and independently from the rejection-sampling-on-path-space route of Chen, Chewi, Rakhlin and Zhang [10].
- $\tilde{O}(d^{1/6})$ *expected* gradient queries at fixed relative-entropy accuracy, from smoothed Picard Hamiltonian Monte Carlo of Chen, Chewi, Lu and Zhang [9]. Their theorem applies with the supplied minimizer in our model, and Pinsker’s inequality gives the same dimension exponent at fixed total-variation accuracy. The expected budget can be capped at a constant multiple of its uniform bound, using a fixed fallback on overflow, without changing the exponent.¹
- $d^{1/4}$ in the separate high-accuracy total-variation result of [9, Theorem 1.3], which combines smoothed Picard with the proximal bouncy-particle sampler [6] and has polylogarithmic dependence on inverse accuracy. A follow-up preprint of September 30, 2026 by Chen, Chewi, Lu and Zhang [7] improves this to $\tilde{O}(d^{1/5})$ expected gradient queries at fixed conditioning, while retaining polylogarithmic dependence on inverse total-variation accuracy. Its Gaussian-cloud representation propagates unevaluated Picard iterates using likelihood corrections.

High-order integration and polynomial collocation already reduce the dimensional dependence of Hamiltonian Monte Carlo, but under stronger regularity or structural assumptions than the two-sided Hessian bounds (1.1) (see Lee–Song–Vempala [18] and Mangoubi and Smith [21]). The obstacle for the present problem is exactly this: obtain high-order accuracy under the two-sided

¹Take the relative-entropy parameter $\Delta = 1/20$, so that the total-variation error is at most $1/(20\sqrt{2})$. If $B(d)$ bounds the expected number of queries, capping the run at $\lceil 20B(d) \rceil$ adds at most $1/20$ to the error by Markov’s inequality and a coupling of the two runs. The resulting error is less than $1/10$.

Hessian bounds alone, while implementing every conditional mean with finitely many first-order queries.

On the lower-bound side, Chewi, de Dios Pont, Li, Lu and Narayanan [13] developed a systematic first-order query-complexity treatment of log-concave sampling. Their Gaussian lower bounds use rotational invariance to reduce arbitrary adaptive algorithms to block-Krylov information and already give logarithmic dimension dependence for well-conditioned Gaussian sampling at a suitable fixed total-variation accuracy. The manuscript’s lower bound (our Section 9) follows that strategy, with an explicit spectrum in $[1, 2]$ and a separating statistic. The complete argument for total-variation threshold $1/10$ is given, including unrestricted randomized outputs.

Two methodological precedents deserve emphasis, because the construction reuses their frameworks. First, the Gaussian conditional-sampling framework of Lee, Shen and Tian [19], with the functional-inequality analysis of Chen, Chewi, Salim and Wibisono [11]: its main operation adds a Gaussian quadratic to the potential and samples the resulting restricted conditional law. Second, the smoothed Picard work of Chen, Chewi, Lu and Zhang [9], which develops Gaussian smoothing of C^2 potentials, conditional cumulant estimates for derivatives of the smoothed potential, and high-order Picard approximation (including arbitrary Picard depth when exact smoothed gradients are available). The manuscript keeps the conditional-sampling framework and proves the contraction used here, and it keeps the conditional-cumulant viewpoint but proves the *all-split* tensor estimates needed for its construction. Its genuinely new ingredient on the algorithmic side is the implementation of recursively nested noisy conditional means with finitely many queries to the original gradient, together with the subsequent removal of the auxiliary noise.

1.3 The proof on one page

1. **Reduce to a near-quadratic conditional problem.** Fix $h > 0$, let $X \sim \pi_V$ and let G be an independent standard Gaussian. The conditional law of X given $X + \sqrt{h}G = y$ has density proportional to $\exp(-V(x') - |x' - y|^2/(2h))$. Substituting $x' = y + \sqrt{h}z$ gives the law $\nu_{x,r}(dz) \propto e^{-|z|^2/2 - F(x+rz)}$ with center $x = y$, scale $r = \sqrt{h}$ and $F(z) = V(y + \sqrt{h}z)$, whose gradient has Lipschitz constant at most $2h$. One evaluation of ∇V is one evaluation of ∇F . The problem therefore becomes: sample a *near-quadratic* law $\nu_{x,r}$ for an arbitrary center x and scale $0 < r \leq 1$, when $\text{Lip}(\nabla F) \leq \delta := d^{-e}$ and nothing else about F is known.
2. **Smooth conditional means and bound their tensors.** With $Z \sim \nu$ and an independent standard Gaussian N , form $Y = \rho Z + \sqrt{1 - \rho^2} N$ and let ξ_ρ be its law. The conditional mean $M_\rho(y) = \mathbb{E}[f(Z) \mid Y = y]$, where $f(z) = g(x + rz)$ and $g = \nabla F$, is the velocity of the transport that pushes the standard Gaussian onto ξ_ρ (a stochastic-interpolant identity of Albergo–Boffi–Vanden-Eijnden [2]). Completing the square shows that $M_\rho(y)$ is again a conditional mean of the same type, at center $x + r\rho y$ and scale $r\sqrt{1 - \rho^2}$. Differentiating a Gaussian conditional expectation produces cumulants, and the Poincaré inequality of the conditional law bounds every derivative by a quantity independent of the dimension (Lemma 3.1, Proposition 3.2). This is where the unconditional C^2 hypothesis is exploited: conditioning makes the mean smooth.
3. **Control material derivatives with split norms.** The ODE $\dot{Y}_\rho = -rM_\rho(Y_\rho)$ transports $N(0, I_d)$ to ξ_ρ (Proposition 3.3). To replace path integrals by high-order quadrature one needs derivatives along this flow. A naive estimate loses one power of d per derivative, because a derivative tensor in dimension d has d coordinate values. The construction keeps at least two tensor slots free and measures each tensor by its operator norm under *every* split of its slots into two nonempty groups (“split norm”). A contraction of two tensors then costs nothing

in dimension (Lemma 4.2), and repeated integration-by-parts in the adjoint divergence of ξ_ρ costs only powers of $\log d$: the one coordinate trace of d^b is absorbed by taking the L^p moment $p \asymp (j + \log d)^2$ large (Theorem 4.4). Only the final divergence that produces a vector costs one factor $\sqrt{d}$ (Lemma 4.5). The outcome is that m -th material derivatives of M are bounded by $\text{polylog}(d) L\sqrt{d} (1 - \rho^2)^{-m}$ along the flow (Corollary 4.6), which is what makes quadrature over cells of length $\asymp (1 - \rho)\delta^w$ accurate.

4. **Turn a path integral into a noisy mean.** The centered observable $M_* - u$ can be written as an adjoint divergence $\nabla^* K$ of an explicit matrix kernel K built from the transport Jacobian (Lemma 5.2, which is a Stein-type identity in the spirit of Courtade–Fathi–Pananjady [14]). Coupling the position to a Gaussian momentum through the skew-symmetric drift

$$\dot{Y}_t = -\frac{K(Y_t)^T G_t}{s}, \quad \dot{G}_t = \frac{M_*(Y_t) - u}{s}$$

leaves the product law $\xi_* \otimes N(0, I_d)$ invariant, and along its solution

$$sG_0 + \int_0^1 M_*(Y_t) dt \stackrel{\text{law}}{=} u + sG, \quad G \sim N(0, I_d)$$

(Proposition 5.3). So a *noisy conditional mean* is represented by an exact path integral of conditional means. Time derivatives on the stationary trajectory are bounded in the fixed product law (Lemma 5.4), so high-order quadrature applies to this path integral too.

5. **Discretize both flows with fixed dependency depth.** Both ODEs are discretized by collocation of *global Picard iteration*: interpolate the velocity on cells and iterate the resulting fixed-point map. The number of Picard layers is $O(J + 1)$, independent of the number of cells and of the interpolation order (Propositions 6.1 and 6.5). Two technical points make the centering flow work: the kernel action is differentiated through short harmonic trajectories at time zero rather than by differentiating the kernel itself, which preserves global Lipschitz continuity (Lemma 6.2), and the apparent $1/\rho$ singularity of the kernel integral at zero is removed by the exact identity $B'_0 = 0$, imposed as a Hermite condition on the first cell (Lemma 6.3). The result (Theorem 6.7) is a finite affine expression in exact conditional means, with deterministic coefficients, centers of the form $x + r(\ell_i y + m_i G) + \sum_{j < i} a_{ij} u_j$, scales bounded below, bounded coefficient sums, and dependency depth $3N^2 + 1$ independent of the mesh size.
6. **Evaluate nested means with added noise.** The finite expression is a circuit whose mean nodes sit at centers depending on earlier mean values. Each mean is implemented by a routine that adds a prescribed Gaussian to every center it supplies. A deterministic translation identity, $Q(x + \Delta, z + v \otimes \Delta/r) = Q(x, z)$ for a suitable vector v of seed coefficients, lets a routine absorb the center displacement into its Gaussian seeds by a covariance reduction (Lemma 7.1). Recursing through the routine tree gives a per-execution query cap of $\text{polylog}(d)\delta^{-t}$ and error labels that lose a fixed fraction per level, with no random failure event (Lemma 7.3, Proposition 7.7). This is the step that makes the construction finitely implementable with first-order queries to the *original* gradient.
7. **Return to the target law.** The primitive returns a sample of $\nu_{x,r}$ convolved with a small Gaussian. A proximal (Gaussian-step) chain with step $h = \delta/2$ contracts W_2 at geometric rate $1/(1 + h)$ and approaches $\pi_V * N(0, hI_d)$ in $O(h^{-1} \log d)$ steps (Lemma 8.1, Proposition 8.2). Conditionally on the chain's output y , the algorithm then runs $J \asymp \log d$ variance-halving

steps: if $Z \sim \nu_{x,r}$ and N is standard, then $Z = Y/2 + Z_{\text{new}}/\sqrt{2}$ with $Z_{\text{new}} \mid Y \sim \nu_{x+rY/2, r/\sqrt{2}}$ (identity (8.5)), so noised conditional draws halve the residual scale. At the last level the conditional law is within TV distance $O(\ell\sqrt{d})$ of a mode-centered Gaussian, whose center is found by a simple gradient iteration (Lemma 8.4). Summing the errors gives TV distance $o(1)$ at $O(\delta^{-1} \text{polylog } d)$ calls (Theorem 8.5). The finitely many small dimensions are handled by a rejection fallback with Gaussian proposals and a capped number of trials.

8. **Lower bound.** Restrict to quadratic potentials $V(x) = \frac{1}{2}x^T A x$ with $I \preceq A \preceq 2I$ and $A = O\Lambda O^T$ for a Haar rotation O . A deferred-revelation argument represents the output of any q -query algorithm in a block-Krylov span of $q+1$ auxiliary Gaussian vectors (Lemma 9.2). A cosine statistic of frequency $2q+1$ is at most $40(q+1)^2/\sqrt{d}$ on that span with probability 0.99 (Lemma 9.3), while under the target Gaussian its expectation is $(d/\sqrt{2})(-\varrho)^{2q+1}$ up to an exponentially small remainder, with $\varrho = 3 - \sqrt{8}$ (Lemma 9.4). A suitable threshold event therefore has probability at most 0.01 under the algorithm's joint law but $1 - o(1)$ under the target's, forcing $q = \Omega(\log d)$ (Theorem 9.1).

Three ideas do the work. The first is that *Gaussian conditioning makes everything smooth*: although F has only two derivatives, every derivative of the conditional mean M_ρ exists and is controlled by the Poincaré inequality, and the conditional mean at one scale is again a conditional mean at a smaller scale. The second is that *dimension-free tensor calculus is possible* if one measures tensors by all their matrix unfoldings and saves the single $\sqrt{d}$ for the last vector-valued step. The third is that *the algorithm need not sample conditional means exactly*: a Gaussian added to each center can be absorbed into the routine's seeds exactly, so nested conditional means with noisy centers compose without accumulating failure events.

1.4 What is new and what is inherited

Most of the construction adapts known machinery:

- the Gaussian conditional-sampling framework of Lee–Shen–Tian [19] and its functional-inequality analysis by Chen–Chewi–Salim–Wibisono [11],
- the stochastic-interpolant conditional velocity of Albergo–Boffi–Vanden-Eijnden [2],
- the skew-symmetric stationary-flow recipe of Ma–Chen–Fox [20],
- polynomial collocation of Picard iteration for ODEs and sampling, from Lee–Song–Vempala [18] and from the smoothed-Picard work of Chen–Chewi–Lu–Zhang [9],
- the existence of Stein kernels under a spectral gap (Courtade–Fathi–Pananjady [14]), together with the pointwise bounds that the cited existence statement does not supply and that the manuscript proves,
- the curvature method for Poincaré and comparison inequalities (Bakry–Gentil–Ledoux [5]),
- the block-Krylov lower-bound strategy of Chewi et al. [13].

The new ingredients are:

- **All-split tensor estimates and material-derivative bookkeeping (Sections 3 and 4).** Every derivative tensor of a Gaussian conditional mean is bounded as a matrix under every division of its slots into two nonempty groups, and the resulting expressions retain enough free slots that high-order differentiation and integration by parts cost only $\text{polylog}(d)$, until a single final vector-valued divergence costs $\sqrt{d}$. This is what lets every fixed positive query exponent be reached by choosing the approximation orders.
- **Exact finite formulas for the two flows with fixed dependency depth (Section 6).** The probability transport and the stationary centering flow are turned into finite affine expressions in conditional means whose nesting depth is fixed *before* the interpolation order is chosen. The kernel action is computed through harmonic trajectories, and the endpoint singularity is removed by an exact identity.
- **Recursive noisy evaluation of nested conditional means (Section 7).** The auxiliary Gaussian added to each center is absorbed exactly into the routine’s seeds by a translation identity with a covariance reduction, so nested evaluations are implemented by finitely many queries to the original gradient with a worst-case query cap. This is an exact finite-dimensional identity, not an appeal to an approximate sampling oracle.
- **Removal of the auxiliary noise (Section 8).** A contracting proximal chain and variance-halving conditional descent reduce the problem to an almost-Gaussian law whose mode-centered Gaussian approximation can be computed by gradient iteration, so the final guarantee is for the unnoised target in total variation.

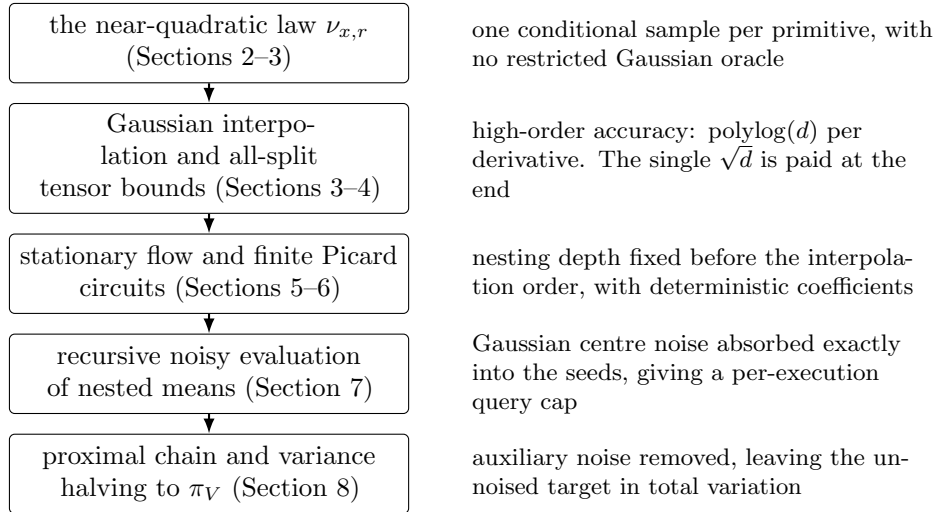


Figure 1: Roadmap. The left column is the construction, and the right column records what each step contributes. All query counts are worst-case and depend only on the fixed parameters and on the dimension.

1.5 How this note is organized

Section 2 states the near-quadratic sampling theorem and the elementary comparison tools used throughout. Sections 3–5 construct and estimate the two flows: Gaussian transport and tensor derivatives (Section 3), material derivatives and their dimension bounds (Section 4), and the

stationary flow that produces noisy means (Section 5). Section 6 discretizes the flows into finite formulas in exact conditional means, and Section 7 implements those formulas with actual gradient queries. Section 8 returns to the original target law and proves the upper bound, and Section 9 proves the independent lower bound. Section 10 records what the result means for diffusion-model sampling.

Throughout, results and equations are numbered by section. Symbols that occur only in one section are defined there.

2 The near-quadratic sampling problem

The conditional density that appeared in item 1 of the one-page proof is a standard Gaussian multiplied by a factor whose Hessian is small: the full potential $|z|^2/2 + F(x + rz)$ has Hessian between $(1 - l)I$ and $(1 + l)I$ with $l = \text{Lip}(g)r^2$. The algorithmic problem is therefore stated for the larger class of all such perturbations, and the center is allowed to be arbitrary: later calls will happen at centers produced by earlier random computations, and there is no way to guarantee that such a center is a minimizer, or even that ∇F is small there. The quantity that controls the difficulty is $D(x) = \sqrt{d} + |g(x)|$, the size of the gradient at the center.

This section states the auxiliary sampling theorem (Theorem 2.1) and collects the elementary facts (moment bounds for strongly log-concave laws, Gaussian comparison and smoothing) that the rest of the construction uses without comment.

2.1 Notation and the auxiliary theorem

Vectors carry the Euclidean norm, and matrices the induced operator norm unless stated otherwise. For a random vector or tensor, $\|X\|_2 = (\mathbb{E}|X|^2)^{1/2}$ with the chosen pointwise norm, and the notation $\|X\|_{F,2}$ specifies the Frobenius norm. The distance W_2 is the quadratic Wasserstein distance on Euclidean space, and $\|\mu - \eta\|_{\text{TV}} = \sup_B |\mu(B) - \eta(B)|$. All probability laws and oracle procedures are measurable.

Fix $0 < e < 1$ and put $\delta = d^{-e}$. Let $F : \mathbb{R}^d \rightarrow \mathbb{R}$ be C^2 with gradient $g = \nabla F$ of Lipschitz constant at most the declared bound $\lambda = \delta$. Values of F need not be available. For $x \in \mathbb{R}^d$ and $0 < r \leq 1$ define

$$\nu_{x,r}(dz) = \frac{\exp(-|z|^2/2 - F(x + rz)) dz}{\int_{\mathbb{R}^d} \exp(-|v|^2/2 - F(x + rv)) dv}, \quad f(z) = g(x + rz), \quad u = \mathbb{E}_{\nu_{x,r}} f. \quad (2.1)$$

We often write $\nu = \nu_{x,r}$, and we use the abbreviations

$$L = \lambda r, \quad l = \lambda r^2 = rL, \quad D(x) = \sqrt{d} + |g(x)|. \quad (2.2)$$

Here x is the actual input center. It need not be a minimizer of F . The positive number λ is a *declared* upper bound on $\text{Lip}(\nabla F)$, so the notation also covers affine or constant F . The full potential in (2.1) has Hessian between $(1 - l)I$ and $(1 + l)I$. Since $l \leq \delta < 1$ for $d \geq 2$, the normalizing integral is finite even when F is not convex. For the quantitative construction we take d large enough that $l \leq 1/2$.

Throughout the construction, $\text{polylog}(d)$ denotes a factor bounded by $C(1 + \log d)^C$. Its constants may depend on the fixed parameters e , the accuracy orders, and other fixed orders introduced below, but never on F , x , r , or the oracle responses. Different occurrences may have different constants. Every sufficiently-large-dimension threshold has the same uniformity.

Theorem 2.1 (Noisy near-quadratic sampling). Fix $0 < e < 1$, an integer $K_0 \geq 2$, and $t > 0$ satisfying

$$40(K_0 + 2)t < \frac{1}{4}. \quad (2.3)$$

There are constants and a dimension threshold depending only on e, K_0, t with the following property. For every integer d above that threshold, put $\delta = d^{-e}$. One measurable randomized oracle procedure works for every $F \in C^2(\mathbb{R}^d)$ with $\text{Lip}(\nabla F) \leq \delta$, every $x \in \mathbb{R}^d$, and every $0 < r \leq 1$. Given evaluation access to $g = \nabla F$, it uses at most $\text{polylog}(d)\delta^{-t}$ evaluations of g on every execution and returns

$$X = X_{\text{pre}} + \frac{1}{2}G,$$

where $G \sim N(0, I_d)$ is independent of X_{pre} and of all queries, and

$$W_2(\mathcal{L}(X_{\text{pre}}), \nu_{x,r} * N(0, \frac{3}{4}I_d)) \leq \text{polylog}(d)\delta^{K_0}D(x). \quad (2.4)$$

Consequently both

$$W_2(\mathcal{L}(X), \nu_{x,r} * N(0, I_d)) \quad \text{and} \quad \|\mathcal{L}(X) - \nu_{x,r} * N(0, I_d)\|_{\text{TV}}$$

are at most $\text{polylog}(d)\delta^{K_0}D(x)$. All constants in these bounds are uniform in F, x, r . The procedure's scalar parameters, shapes of its computations, and query cap are independent of x and of the primitive. It uses no evaluations of F itself.

The proof occupies Sections 3–7. The independent Gaussian G is reserved until all queries have been made. Adding the same reserve to a coupling preserves its transportation error, and smoothing by that reserve converts the W_2 estimate into the total-variation estimate through Lemma 2.3 below. The target in (2.4) contains the full independent Gaussian convolution $N(0, I_d)$. Section 8 removes that auxiliary noise before returning a sample for the original problem.

Why a whole theorem is needed for one conditional draw. It would be enough for the algorithm to be able to sample $\nu_{x,r}$ exactly, but that is precisely what is not available: the only access to F is through evaluations of g , and there is no restricted Gaussian oracle. The theorem says that a sample from $\nu_{x,r} * N(0, \frac{3}{4}I_d)$ plus an independent reserve $G/2$ can be produced with $\text{polylog}(d)\delta^{-t}$ exact gradient queries, for every center x and every scale r , using a procedure whose *shape* does not depend on the center. The reserve is not an error allowance. It is the noise that the later conditioning steps rely on.

2.2 Convexity, moments, and comparisons

We begin with the elementary diffusion facts. They are familiar forms of the curvature method for functional inequalities [5], and the proof is included because the exact regularity requirements matter later.

Lemma 2.2 (Uniformly convex laws). Let $H \in C^2(\mathbb{R}^n)$ satisfy $mI \preceq \nabla^2 H \preceq MI$ with $0 < m \leq M < \infty$. Write $\mu(dx) \propto e^{-H(x)} dx$, and let z_0 be its unique mode. Then

$$\mathbb{E}_\mu |X - z_0|^2 \leq n/m, \quad \text{Var}_\mu h \leq m^{-1} \mathbb{E}_\mu |\nabla h|^2.$$

The latter applies, in particular, to C^1 functions whose value and gradient have polynomial growth. The diffusion

$$dX_t = -\nabla H(X_t) dt + \sqrt{2} dB_t$$

has invariant law μ , and synchronous solutions contract in distance by e^{-mt} . If $\bar{H}$ is another strongly convex C^2 potential with bounded Hessian and law η , then

$$W_2(\mu, \eta) \leq \frac{1}{m} \left(\mathbb{E}_\eta |\nabla H - \nabla \bar{H}|^2 \right)^{1/2}. \quad (2.5)$$

Proof. Strong convexity implies Gaussian tails and existence of a unique minimizer z_0 . Integration by parts gives $\mathbb{E}_\mu(X - z_0) \cdot \nabla H(X) = n$, and strong monotonicity of ∇H then gives the moment bound $\mathbb{E}_\mu |X - z_0|^2 \leq n/m$. The globally Lipschitz drift $-\nabla H$ makes the diffusion nonexplosive. For synchronous solutions X_t, Y_t started at different points, differentiating the squared distance and using strong monotonicity gives $\frac{d}{dt} |X_t - Y_t|^2 \leq -2m |X_t - Y_t|^2$, hence contraction at rate e^{-mt} .

It remains to justify the Poincaré inequality. First suppose H is smooth and has bounded derivatives of orders at least two. For a smooth compactly supported h , let $P_t h(x) = \mathbb{E}[h(X_t) \mid X_0 = x]$. Differentiating the synchronous flow in the initial state gives bounded first and second derivatives on bounded time intervals, and Itô's formula and the semigroup identity yield

$$\partial_t P_t h = P_t A h = A P_t h, \quad A = \Delta - \nabla H \cdot \nabla.$$

The equality with $A P_t h$ also follows by differentiating $P_{t+s} h = P_s(P_t h)$ at $s = 0$. Its at-most-linear growth is integrable against μ , and integration by parts gives $\int A P_t h d\mu = 0$, hence invariance. Coupling to an invariant starting point and using contraction shows $P_t h \rightarrow \int h d\mu$ in $L^2(\mu)$. Therefore

$$\text{Var}_\mu h = 2 \int_0^\infty \int |\nabla P_t h|^2 d\mu dt.$$

The flow derivative has operator norm at most e^{-mt} . Jensen's inequality and invariance thus bound the inner integral by $\int e^{-2mt} |\nabla h|^2 d\mu$, proving the Poincaré inequality (after a limit as $t \uparrow \infty$).

For a general C^2 potential, convolve H with smooth compactly supported mollifiers of radii tending to zero. Both Hessian bounds are preserved, and all higher derivatives are bounded for each mollification. The normalized laws converge in total variation by the uniform quadratic tail bounds (subtract the potential's value at zero). The gradients converge uniformly, since the original gradient is globally Lipschitz. Synchronous solutions of the mollified and original diffusions consequently converge uniformly in their common starting point on each bounded time interval, by Gronwall's inequality, and testing invariance against bounded Lipschitz functions passes it to the limit. Poincaré first passes on smooth compact tests, then by cutoff and approximation to the stated class.

For (2.5), start the two diffusions with their respective invariant marginals, with any initial coupling, and use a common Brownian motion. Writing $a(t) = (\mathbb{E} |X_t - Y_t|^2)^{1/2}$ and $\beta = (\mathbb{E}_\eta |\nabla H - \nabla \bar{H}|^2)^{1/2}$, strong monotonicity and Cauchy-Schwarz give, in the usual upper-derivative sense,

$$a'(t) \leq -ma(t) + \beta.$$

The marginals remain fixed, so $W_2(\mu, \eta) \leq a(t)$ for every t , and letting $t \rightarrow \infty$ proves the claim. $\square$

These estimates apply to the auxiliary laws without imposing convexity on F . For $\nu_{x,r}$, the mode equation reads $z_0 = -rg(x + rz_0)$. In particular, for $l \leq 1/2$,

$$|z_0| \leq \frac{r|g(x)|}{1-l}, \quad (\mathbb{E}_\nu |Z|^2)^{1/2} \leq C(\sqrt{d} + r|g(x)|), \quad |u| \leq CD(x). \quad (2.6)$$

Indeed, $|g(x + rz_0) - g(x)| \leq L|z_0|$ and $z_0 = -rg(x + rz_0)$ prove the first bound. The lemma with $m = 1-l$ gives $\|Z - z_0\|_2 \leq \sqrt{d/(1-l)}$, and the triangle inequality together with $|u| \leq |g(x)| + L\|Z\|_2$ gives the remaining bounds. Thus $D(x)$ controls the required moments uniformly over all centers, including those with very large gradients.

Why the mode estimate is the right way to see $D(x)$. The center x can be far from a minimizer, so $|g(x)|$ can be large. The first bound above says that the tilted law $\nu_{x,r}$ is centered within $O(r|g(x)|)$ of the origin in the normalized coordinates, and the second and third say that its moments and the mean u are bounded by multiples of $D(x)$. Every error bound in the construction is proportional to $D(x)$, and this is the only place where the size of the gradient at the center enters.

2.3 Coupling and smoothing conventions

Errors of a deterministic map evaluated on a specified input law will be measured in root mean square, and errors of randomized outputs will be measured in W_2 . The following rules keep these two notions separate.

Lemma 2.3 (Conditional comparisons and smoothing). Let (μ_z) and (η_z) be probability kernels on finite-dimensional Euclidean spaces, and let Z have law ζ . Whenever the relevant second moments are finite,

$$W_2^2\left(\int \mu_z d\zeta(z), \int \eta_z d\zeta(z)\right) \leq \int W_2^2(\mu_z, \eta_z) d\zeta(z). \quad (2.7)$$

For finite families, independent sums satisfy

$$W_2\left(\mathcal{L}\left(\sum_i a_i X_i\right), \mathcal{L}\left(\sum_i a_i Y_i\right)\right) \leq \sum_i |a_i| W_2(\mathcal{L}(X_i), \mathcal{L}(Y_i)),$$

where independence is required within each of the two families, not between them. For $s > 0$,

$$\|\mu * N(0, s^2 I) - \eta * N(0, s^2 I)\|_{\text{TV}} \leq s^{-1} W_2(\mu, \eta). \quad (2.8)$$

Moreover, applying a common probability kernel cannot increase total variation, and

$$\left\| \int \mu_z d\zeta(z) - \int \eta_z d\zeta(z) \right\|_{\text{TV}} \leq \int \|\mu_z - \eta_z\|_{\text{TV}} d\zeta(z).$$

Proof. Use conditional near-optimal couplings for (2.7) and their independent product for the sum bound, followed by the triangle inequality in L^2 . Measurability of such conditional comparisons can be seen without an exact optimal-kernel selection: quantize each output law on a finite grid inside a box, sending the complement to zero. A minimizing plan on the finite sets can be chosen measurably by a fixed ordering of the finitely many basic feasible solutions, and then lifted using the conditional measures in the cells as the grids and boxes are refined. Finite unconditional second moments and the triangle inequality give the asserted limit.

For deterministic centers x, y , the relative entropy of $N(x, s^2 I)$ from $N(y, s^2 I)$ is $|x - y|^2 / (2s^2)$. Pinsker's inequality gives total variation at most $|x - y| / (2s)$. Integrate this estimate over any coupling and apply Cauchy–Schwarz to obtain (2.8), with room in its constant. The total-variation rules follow directly by testing measurable functions taking values in $[0, 1]$. $\square$

We will also repeatedly combine deterministic root-mean-square errors with absolute coefficient sums, which uses the ordinary L^2 triangle inequality and does not require independence.

3 Gaussian transport and tensor derivatives

The next step toward Theorem 2.1 is to describe a transport from a Gaussian to the auxiliary law $\nu_{x,r}$. Its velocity will be a conditional gradient mean. Although the original potential has only two derivatives, conditioning by a nondegenerate Gaussian makes that mean smooth, and we can bound each of its derivative tensors as a matrix under *every* division of its slots into two groups. These bounds keep later tensor contractions from reintroducing a dimension factor at each step. The interpolation itself is the Gaussian stochastic interpolant of Albergo, Boffi and Vanden-Eijnden [2, Section 5.1, Equations (5.5)–(5.6)]. We use its conditional velocity and prove the tensor estimates required for the sampling construction.

Retain the primitive notation from Section 2, and assume $l \leq 1/2$, as holds for all sufficiently large d . For independent $Z \sim \nu$ and $N \sim N(0, I_d)$, put

$$a = \sqrt{1 - \rho^2}, \quad Y = \rho Z + aN, \quad \xi_\rho = \mathcal{L}(Y), \quad 0 \leq \rho < 1. \quad (3.1)$$

Write $\xi_\rho(dy) = e^{-H_\rho(y)} dy$, incorporating the normalizing constant in H_ρ , and define

$$M_\rho(y) = \mathbb{E}[f(Z) \mid Y = y].$$

We use the version defined at every y by the conditional density. Completing its quadratic gives

$$\mathcal{L}(Z \mid Y = y)(dz) \propto \exp\left(-\frac{|z - \rho y|^2}{2a} - F(x + rz)\right) dz. \quad (3.2)$$

Under the change of variables $z = \rho y + aw$, $M_\rho(y)$ is exactly the primitive mean with center $x + r\rho y$ and scale $r\sqrt{a}$. In particular $M_0 = u$. Thus the conditional mean is a new instance of the problem already defined in (2.1), at a specified center and a smaller positive scale. This is a density identity, and evaluating that mean from finitely many gradient queries remains part of the construction.

What the interpolation buys. The parameter ρ moves the law of $Y = \rho Z + aN$ continuously from the standard Gaussian $N(0, I_d)$ at $\rho = 0$ to ν as $\rho \rightarrow 1$. The conditional mean $M_\rho(y)$ is the velocity of this motion, and the crucial self-similarity is the observation just made: at scale r and interpolation time ρ , the conditional mean is again a mean of the same type at a smaller scale $r\sqrt{1 - \rho^2}$. Every later step either samples ν or computes u , and both are done by following this transport. The recursion is what allows a finite description.

3.1 Cumulants from the Poincaré inequality

A tensor $T \in (\mathbb{R}^D)^{\otimes n}$ can be viewed as a matrix by partitioning its n slots into two nonempty sets A and B . Contraction on the A slots then defines a linear map $(\mathbb{R}^D)^{\otimes |A|} \rightarrow (\mathbb{R}^D)^{\otimes |B|}$, where both tensor spaces carry their Euclidean, or Frobenius, norm. We call the operator norm of this matrix a *split norm*, and write $|T|_{\text{split}}$ for the bound that holds for every such partition. Equivalently, $|T|_{\text{split}} = \sup |T : (V \otimes W)|$ over coefficient tensors V, W of Frobenius norm at most one with slots in A, B respectively, where the colon denotes contraction over all paired slots. Permuting slots simply permutes these norms. The Frobenius norm of a tensor itself is denoted by $|T|_F$.

For related derivative and cumulant estimates for Gaussian-smoothed log-concave potentials, see [9, Appendix C.2]. Here we need bounds for every nontrivial unfolding of each tensor, and we derive that form directly from Poincaré, using the multivariate Appell polynomials of Anshelevich [4, Section 2.3] and the covariance bounds of Adamczak–Wolff [1, Proposition 3.2 and Theorem 3.3]

and Kothari–Steinhardt [17, Section 4.2]. We give the direct L^2 argument, with its dependence on the polynomial degree, and then use a logarithm with two groups of variables to bound every split of the cumulant tensor.

Lemma 3.1 (All-split cumulant bound). Let S be a random vector in $\mathbb{R}^D$ whose moment-generating function is finite in a neighborhood of zero. Suppose that, for some $\beta \geq 0$,

$$\text{Var}(P(S)) \leq \beta \mathbb{E}|\nabla P(S)|^2$$

for every polynomial P on $\mathbb{R}^D$. For each integer $n \geq 2$, define its cumulant tensor by $\kappa_n(S) = \nabla_\theta^n \log \mathbb{E} e^{\theta \cdot S} \big|_{\theta=0}$. Then, for an absolute constant C ,

$$|\kappa_n(S)|_{\text{split}} \leq C^n n^{Cn} \beta^{n/2}$$

for every nontrivial split. The constant is independent of D and of the mean of S .

Proof. We first prove a dimension-independent covariance bound for centered polynomial tensors, and then recover cumulants from those covariances. If $\beta = 0$, applying the hypothesis to coordinate functions shows that S is almost surely constant, so every cumulant in the statement vanishes. We may therefore suppose $\beta > 0$.

Define the symmetric Appell tensor polynomials by

$$P_j(s) = \nabla_\theta^j \frac{\exp(\theta \cdot s)}{\mathbb{E} \exp(\theta \cdot S)} \bigg|_{\theta=0}.$$

For a coefficient tensor T of order j , let $Q_T(s) = T : P_j(s)$, where the colon denotes contraction on all slots. Differentiating the generating function in s multiplies it by copies of θ . Consequently

$$\mathbb{E} \nabla^h Q_T(S) = 0 \quad (0 \leq h < j), \quad \nabla^j Q_T = j! \text{Sym}(T),$$

where $\text{Sym}(T)$ is the average over all permutations of the slots, so $|\text{Sym}(T)|_F \leq |T|_F$. Apply Poincaré to each component of $\nabla^h Q_T$ and sum: since the expectations vanish, repeated use of $\text{Var}(P) \leq \beta \mathbb{E}|\nabla P|^2$ from $h = 0$ to $h = j - 1$ yields

$$\mathbb{E}|T : P_j(S)|^2 \leq \beta^j (j!)^2 |T|_F^2. \quad (3.3)$$

Only first derivatives in the displayed Poincaré inequality are used at each application. In particular, if T and V have orders $j, k \geq 1$, then Cauchy–Schwarz gives

$$|\mathbb{E}[(T : P_j(S))(V : P_k(S))]| \leq \beta^{(j+k)/2} j!k! |T|_F |V|_F. \quad (3.4)$$

This is the direct covariance estimate. It concerns arbitrary coefficient tensors, rather than products of one-dimensional test directions, which is why it controls a matrix split.

To pass from covariance to cumulants while preserving a specified split, consider the two-variable generating function

$$\mathcal{G}(\theta, \omega) = \frac{\mathbb{E} \exp((\theta + \omega) \cdot S)}{\mathbb{E} \exp(\theta \cdot S) \mathbb{E} \exp(\omega \cdot S)}.$$

Its derivative of orders $j \geq 1$ in θ and $k \geq 1$ in ω at $(0, 0)$ is the cross-moment tensor $\mathbb{E}[P_j(S) \otimes P_k(S)]$. Across the split between the θ and ω slots its operator norm is at most $\beta^{(j+k)/2} j!k!$ by (3.4).

Both $\mathcal{G}(\theta, 0)$ and $\mathcal{G}(0, \omega)$ are identically one. Thus every nonzero term of $\mathcal{G} - 1$ contains variables from both groups. In the formal identity

$$\log \mathcal{G} = \sum_{m \geq 1} \frac{(-1)^{m+1}}{m} (\mathcal{G} - 1)^m,$$

only $m \leq n$ can contribute at total order n . Every differentiated product is a tensor product of the preceding cross-moment operators, up to permutations within the two sides. Its norm is therefore at most the product of their norms. To bound the total coefficient, distribute the n labeled derivative slots among at most n ordered nonempty factors. There are at most n^{n+1} such assignments, and the product of the factorials within any assignment is at most $n!$. Enlarging an absolute constant bounds their total by $C^n n^{Cn}$.

The mixed derivatives of $\log \mathcal{G}$ are precisely the corresponding derivatives of $\log \mathbb{E} \exp((\theta + \omega) \cdot S)$, since the other two logarithms have only one variable group. They are the cumulants, and the two groups can represent any prescribed nontrivial split of the n slots. This proves the claim. $\square$

In the conditional density (3.2), y enters through the linear tilt $\rho y/a$, so each differentiation of M_ρ with respect to y supplies a factor ρ . We divide out these factors to state bounds that remain meaningful at $\rho = 0$.

Proposition 3.2 (Gaussian jet bounds). Under the assumptions above, in particular $l \leq 1/2$, define for each integer $q \geq 1$

$$U_q(\rho, y) = \rho^{-q} \nabla^q M_\rho(y) \quad (\rho > 0).$$

These tensors extend smoothly to $\rho = 0$. There is an absolute constant C such that, for every $0 \leq \rho < 1$, $y \in \mathbb{R}^d$ and every split,

$$|U_q(\rho, y)|_{\text{split}} \leq C^q q^{Cq} L(\sqrt{a})^{-(q-1)}. \quad (3.5)$$

Here and below the vector-output slot is included among the tensor slots. The constant is uniform in the dimension, the primitive, its center and scale, and the derivative order q . Moreover, DM_ρ and U_1 are symmetric, and

$$\nabla H_\rho(y) = y + r\rho M_\rho(y), \quad \|\nabla^2 H_\rho\|_{\text{op}} \leq C. \quad (3.6)$$

For $j \geq 1$, every split of $\nabla^j \nabla^2 H_\rho$ is bounded by

$$C^{j+1} (j+1)^{C(j+1)} (\sqrt{a})^{-j}. \quad (3.7)$$

Proof. The conditional potential in (3.2) has Hessian between $(a^{-1} - l)I$ and $(a^{-1} + l)I$. By Lemma 2.2, its Poincaré constant is at most Ca . Under this conditional law the map

$$z \mapsto S(z) = (z, f(z)/L)$$

has derivative norm at most $\sqrt{2}$, since f is L -Lipschitz. Pulling back an ambient test function shows that the possibly singular law of S satisfies Poincaré with constant Ca . It has exponential moments of every linear functional: f has at most linear growth and the conditional law has Gaussian tails. Lemma 3.1 therefore applies. This use of repeated Poincaré does not differentiate f repeatedly. Each application pulls back a new ambient polynomial and uses only Df .

In the conditional density, the linear tilt parameter for z is $\rho y/a$. Repeated differentiation of an expectation with respect to a linear tilt gives joint cumulants. In coordinates this says

$$(U_q)_{i, k_1, \dots, k_q} = a^{-q} \text{cum}(f_i(Z), Z_{k_1}, \dots, Z_{k_q} \mid Y = y). \quad (3.8)$$

It can also be checked directly by differentiating the logarithm of the conditional moment-generating function with one auxiliary parameter coupled to f_i . Restrict the cumulant of S to one f/L coordinate and q ordinary coordinates. Lemma 3.1 gives a factor $La^{(q+1)/2}$, and the prefactor a^{-q} gives (3.5).

All conditional integrals used here are smooth in (ρ, y) for $\rho < 1$. This follows by differentiating their densities: on compact parameter sets, every derivative inserts a polynomial with an integrable Gaussian majorant. Formula (3.8) gives the smooth extension at zero without division by ρ . For later integrations by parts we record a slightly stronger regularity fact. On each compact interval in $\rho < 1$, conditional moments of any fixed order grow at most polynomially in y . Indeed, the conditional mode satisfies

$$z_0 = \rho y - \arg(x + rz_0),$$

so $|z_0| \leq C(|y| + r|g(x)|)$, and the upper and lower conditional Hessian bounds compare the density and its normalizing integral around this mode to Gaussian integrals. This proves the claimed moment growth, and the same argument applies to time derivatives of the conditional integrals. Dimension-dependent constants are harmless for this regularity justification, while the quantitative estimates above remain dimension-uniform.

Conditional integration by parts gives $\mathbb{E}[Z \mid Y = y] = \rho y - arM_\rho(y)$. Differentiating the marginal density then yields the first identity in (3.6). For $\rho > 0$ it implies symmetry of DM_ρ , hence of U_1 , and continuity handles $\rho = 0$. Its derivative, together with (3.5), gives the bounded Hessian and (3.7). In particular, all these conclusions use only the original C^2 regularity of F . $\square$

3.2 The transport equation

Proposition 3.3 (Probability transport). The ordinary differential equation

$$\frac{dY_\rho}{d\rho} = -rM_\rho(Y_\rho) \tag{3.9}$$

transports $\xi_0 = N(0, I_d)$ to ξ_ρ for every $\rho < 1$. Its velocity is globally Cl -Lipschitz in space. On every compact probability-time interval it has a smooth flow in both directions.

Proof. For the random interpolation used to define ξ_ρ ,

$$\frac{d}{d\rho}(\rho Z + aN) = \frac{Z - \rho Y}{a}.$$

The conditional mean of this derivative is $-rM_\rho(Y)$ by the conditional integration-by-parts identity above. Thus, for compactly supported smooth φ ,

$$\frac{d}{d\rho} \int \varphi d\xi_\rho = -r \int M_\rho \cdot \nabla \varphi d\xi_\rho.$$

On the other hand, (3.5) with $q = 1$ gives $r\|DM_\rho\|_{\text{op}} \leq Cl$. The velocity has at most linear growth, so its smooth ordinary differential equation has a global-in-space flow and inverse on every compact interval in $\rho < 1$. To identify its pushforward, fix a smooth compactly supported terminal test and transport it backward along this flow. The transported test has zero material derivative, and its supports remain in a bounded set on the interval by the linear-growth bound. The last display then shows that its expectation under ξ_ρ is constant, which proves the pushforward assertion. $\square$

The remaining analytic difficulty is not spatial smoothness, which the Gaussian interpolation has now supplied. It is to control repeated derivatives *along* the transport while paying only one factor $\sqrt{d}$ for vector outputs. That is the subject of the next section.

4 Dimension bounds for material derivatives

The Gaussian interpolation has supplied spatial derivatives of every order using only the original C^2 potential. We now need derivatives along its transporting trajectories: these are the derivatives that control the error when a path integral is replaced by polynomial quadrature. Applying a Euclidean norm after each differentiation would lose a new power of d each time. We keep at least two tensor slots free throughout the calculation and convert to a vector norm only at the last step.

Retain the primitive parameters $L = \lambda r$ and $l = rL \leq 1/2$, and the interpolation $\xi_\rho = \mathcal{L}(\rho Z + aN)$, where $a = 1 - \rho^2$ and $0 \leq \rho < 1$. Expectations and L^p norms in this section use ξ_ρ at the displayed probability time. Along a solution of $Y'_\rho = -rM_\rho(Y_\rho)$, the chain rule gives

$$\frac{d}{d\rho}Q(\rho, Y_\rho) = (\mathcal{D}Q)(\rho, Y_\rho), \quad \mathcal{D} = \partial_\rho - rM_\rho \cdot \nabla.$$

Thus the required path derivatives are iterates of $\mathcal{D}$.

The useful form of their identities involves the adjoint of a spatial derivative. If a tensor field V has a distinguished coordinate slot k , define its *adjoint divergence on that slot* by

$$(\nabla^*V)_I = \sum_{k=1}^d \{-\partial_k V_{I,k} + (\partial_k H_\rho) V_{I,k}\}.$$

Here I lists the slots that remain, and the distinguished slot need not have been last before a permutation. For a compactly supported smooth test function ϕ , the defining integration-by-parts relation is

$$\int (\nabla^*V)_I \phi d\xi_\rho = \int \sum_{k=1}^d V_{I,k} \partial_k \phi d\xi_\rho.$$

The next identities express material derivatives using this operation and contractions of the normalized spatial jets $U_q = \rho^{-q} \nabla^q M_\rho$ from Proposition 3.2.

4.1 Material derivatives of the conditional jets

Lemma 4.1. The mean and its jets satisfy

$$\mathcal{D}M = \nabla^*U_1, \tag{4.1}$$

$$(\mathcal{D}U_q)_{i,I} = (\nabla^*U_{q+1})_{i,I} + 2r\rho \sum_{\emptyset \neq S \subseteq I} (U_{|S|})_{k,S} (U_{q-|S|+1})_{i,k,I \setminus S}, \quad q \geq 1. \tag{4.2}$$

In the second identity I is a labeled list of q derivative slots. Subsets refer to positions in that list and the index k is summed. For any smooth tensor field V ,

$$\mathcal{D}(\nabla^*V) = \nabla^*(\mathcal{D}V) + r\nabla^*((DM_\rho)V), \tag{4.3}$$

where the matrix acts on the slot removed by the divergence. These identities hold also at $\rho = 0$ by continuous extension.

Proof. Differentiate $\mathbb{E}[f_i(Z)\phi(Y)] = \int (M_\rho)_i \phi d\xi_\rho$. The right side differentiates using the transporting velocity from Proposition 3.3. On the left the random velocity is $(Z - \rho Y)/a$. Subtract its conditional mean from the flux multiplying $\nabla \phi$. The remaining coefficient is $\text{Cov}(f_i(Z), Z_k | Y)/a = (U_1)_{i,k}$ by (3.8). Integration by parts proves (4.1).

For the higher identity, coordinate differentiation gives

$$\mathcal{D}\partial_I M_i = \partial_I \mathcal{D}M_i + r \sum_{\emptyset \neq S \subseteq I} (\partial_S M_k) \partial_k \partial_{I \setminus S} M_i.$$

Insert (4.1) and multiply by ρ^{-q} . When all derivatives pass inside the adjoint divergence, the result is $\nabla^* U_{q+1}$. In the other terms they hit the score $y + r\rho M_\rho$. A derivative hitting its linear part contributes qU_q/ρ , cancelled by differentiating the prefactor ρ^{-q} . The nonlinear score terms and the material/spatial commutator terms are identical after scaling: each is

$$r\rho \sum_{\emptyset \neq S \subseteq I} (U_{|S|})_{k,S} (U_{q-|S|+1})_{i,k,I \setminus S}.$$

This proves (4.2). Each contraction retains at least one free slot in each tensor.

Finally differentiate the adjoint identity against a transported scalar test, componentwise in the other slots. The terms involving $\mathcal{D}\phi$ cancel. The gradient commutator is

$$\mathcal{D}\partial_k \phi - \partial_k \mathcal{D}\phi = r(\partial_k M_j) \partial_j \phi.$$

Moving it to the other side by the adjoint relation gives (4.3). All coefficients in the final formulas are nonsingular at zero, so the extensions follow from Proposition 3.2. $\square$

Reading the identities. Identity (4.1) says that the material derivative of the mean, which is the quantity we actually want to differentiate along trajectories, is the adjoint divergence of the first normalized jet. Identity (4.2) iterates this: the material derivative of a jet is the adjoint divergence of the next jet, plus a sum of contractions of two jets whose orders add to $q + 1$. So material differentiation generates expressions built from the U_q by two operations only (contraction and adjoint divergence), and never contracts two slots of the same tensor. This is exactly the class of expressions that the next subsection estimates.

4.2 Contractions and the expressions they generate

Before iterating the material identities, we isolate the norm estimate that makes their product terms harmless. Each coordinate slot represents a copy of $\mathbb{R}^d$. Contracting one slot of each of two tensors means summing the product over one common coordinate, leaving every other slot free. The restriction to one slot is what lets us interpret the operation as matrix composition.

Lemma 4.2 (A single contraction). Let A and B be tensors, and contract one slot of A with one slot of B , leaving at least one free slot in each. For any nontrivial split of the resulting free slots there are nontrivial splits of A and B such that

$$|A * B|_{\text{split}} \leq |A|_{\text{split}} |B|_{\text{split}}.$$

Proof. Call the two sides of the required output split the input and output sides. Orient the contracted index from one factor to the other. If all free slots of one factor lie on the same side, its contracted slot must be placed on the opposite side of that factor's unfolding. These requirements are compatible: incompatibility would put every free slot of the result on one side, contrary to the hypothesis. If a factor already has free slots on both sides, it imposes no requirement on the orientation. With the resulting orientation, the contraction is composition of the two matrix actions, with identities on the unused input and output coordinates and with coordinate permutations. Identities and these permutations have operator norm one. Submultiplicativity proves the assertion. $\square$

We can now specify exactly which tensor formulas will be estimated. This also records which contractions may occur: contracting two slots of a single tensor is not one of the allowed operations.

Definition 4.3 (Tensor expressions and their weights). A tensor expression is a finite formula built by the following rules.

1. For each integer $q \geq 1$, U_q/L is an expression with $q + 1$ free slots and weight $q - 1$.
2. A permutation of the free slots preserves the weight.
3. Two expressions may be contracted on one chosen slot each, provided each factor retains at least one free slot. The resulting weight is the sum of their weights.
4. An adjoint divergence on a chosen free slot is allowed when at least two free slots remain. It increases the weight by one.

The formula specifies its leaves, their orders, and the slots used at each operation. This finite syntactic data is independent of d , ρ , x , r and of the potential. Repeated occurrences are treated as separate factors, even when their values agree.

For example, $\nabla^*(U_2/L)$ has two free slots and weight two, whereas $\nabla^*(U_1/L)$ is *not* an expression under these rules: it has vector output. The latter is precisely the operation reserved for the final $\sqrt{d}$ estimate. If E is an expression, $\nabla^j E$ has j additional free derivative slots, and a split may place any of these slots on either side. The next theorem bounds all such splits simultaneously. Its dependence on j and p is explicit because its proof must later take a moment order growing with $\log d$.

For a related Gaussian result, [8, Theorem 1.1 and Corollary 1.4] proves sharp divergence inequalities with Hilbert–Schmidt derivative norms. Here we require estimates under the generally non-Gaussian law ξ_ρ that retain every matrix split through the contractions and adjoint divergences of Definition 4.3.

Theorem 4.4 (Tensor divergence bound). Let E be a tensor expression of weight k in Definition 4.3. There is a constant $C_E \geq 1$, depending only on that finite formula, such that for every integer $j \geq 0$, every real $2 \leq p < \infty$, every $0 \leq \rho < 1$, and every nontrivial split,

$$|\nabla^j E|_{\text{split}}^{L^p(\xi_\rho)} \leq C_E^{j+1} (p + j + \log(d + 1))^{C_E(j+1)} (\sqrt{a})^{-(k+j)}. \quad (4.4)$$

The same C_E works for all j, p, d, ρ, x, r . In particular it is not restricted to fixed derivative or moment orders.

Proof. We induct on the finite construction of E , proving all derivative and moment orders simultaneously. Permutations and single contractions follow directly from the preceding lemma. For a divergence, we first integrate by parts in a high trace moment. The resulting coordinate sums can be arranged as a composition of matrix unfoldings with only one closing trace. We then choose the moment large enough to absorb that trace’s dimension factor.

For a normalized jet, $\nabla^j U_q = \rho^j U_{q+j}$, so Proposition 3.2 proves the assertion after enlarging a constant depending on q . Permutations cause no change. For a contraction, distribute the derivatives by Leibniz, use Lemma 4.2 on each summand, and use the inductive estimates at exponent $2p$ on both factors. The number of Leibniz summands is at most 2^j . Their weights and derivative orders add, giving (4.4) with an enlarged fixed constant.

It remains to consider $E = \nabla^* V$, where V has weight k_V and E has at least two slots. Commute the j derivatives past the divergence. Apart from

$$T = \nabla^*(\nabla^j V),$$

the terms are single contractions of $\nabla^{j-h} V$ with an order- h derivative of the score, $1 \leq h \leq j$. Both factors retain slots. The score derivative is bounded pointwise by a constant of the form $C^h h^{C^h} (\sqrt{a})^{-(h-1)}$, by (3.6) and (3.7). Lemma 4.2, the inductive hypothesis, and Leibniz multiplicities therefore give the required bound for all these terms, even with exponent $k_V + j - 1$ in place of the larger allowed exponent $k_V + j + 1$. It remains to bound T .

Removing the adjoints. Fix an unfolding of T , with both sides nonempty, and an integer $s \geq 1$. Pointwise,

$$\|T\|_{\text{op}}^{2s} \leq \text{tr}((TT^*)^s).$$

Expand the trace by coordinate indices. It contains $2s$ labeled primary factors, each an adjoint divergence of $\nabla^j V$. Their free slots are joined in the alternating trace cycle, and the two bundles incident to each primary factor are nonempty.

Remove adjoint divergences one at a time by integration by parts. For a chosen primary, its adjoint differentiates the product of the other factors, never its own inside field. If a derivative hits a remaining adjoint, use

$$\partial_i(\nabla^* W) = \nabla^*(\partial_i W) + (\partial_i \partial_k H_\rho) W_k.$$

The first term leaves that adjoint to be processed later. The second consumes it and creates a Hessian factor joining the two divergence indices. A later derivative may also hit a previously created Hessian factor. After all adjoints have been removed, a typical term has primary factors $\nabla^{j+e_i} V$, $1 \leq i \leq 2s$, and n_H factors $\nabla^{h_v} \nabla^2 H_\rho$. The bookkeeping identity is

$$\sum_{i=1}^{2s} e_i + \sum_{v=1}^{n_H} h_v + 2n_H = 2s. \quad (4.5)$$

Indeed, a removed adjoint either adds one derivative or pairs with another adjoint to create a Hessian. Each step has at most Cs possible target factors and at most two commutator terms, so there are at most $(Cs)^{Cs}$ resulting terms.

We use a graph only to describe these explicit coordinate sums. Its primary vertices are the $2s$ labeled tensor factors, and a Hessian vertex represents one factor $\nabla^{h_v} \nabla^2 H_\rho$. An edge represents a pair of tensor slots summed over a common coordinate, and a bundle represents several such edges. Three properties will control its contraction. All original trace bundles remain. Every new direct edge connects distinct primary factors, because an adjoint never differentiates its own inside field. Finally, each Hessian factor has at least two distinct primary neighbors, the pair that created it, and there are no Hessian-to-Hessian edges. Later derivatives only add further primary-to-Hessian edges. These statements remain true if some derivatives had already been commuted inside an unprocessed adjoint.

Opening the trace. Choose one bundle in the original trace cycle and cut it, leaving its $b \geq 1$ indices unsummed at the two ends. All remaining original bundles form a chain through the $2s$ primary factors. Order these factors along that chain. For each Hessian factor, its two distinct creating primaries have different positions, so we can place it strictly between its earliest and latest primary neighbors. No edge joins two Hessians, so these placements impose no competing condition. Order Hessians occupying the same interval arbitrarily, and orient every uncut edge from its earlier factor to its later factor. The b cut indices enter the first primary and leave the last one.

Every tensor has both an input and an output in this ordering. An interior primary receives one original chain bundle and sends the next. At the two endpoints the cut bundle supplies the missing side. A Hessian receives an edge from an earlier neighbor and sends an edge to a later one. Thus all the matrix unfoldings used below are nontrivial splits, exactly as required by the induction hypothesis.

Here is an explicit matrix-composition interpretation. Immediately before processing a factor, retain one copy of $\mathbb{R}^d$ for each index whose earlier endpoint has been processed but whose later endpoint has not. The initial space is $(\mathbb{R}^d)^{\otimes b}$, belonging to the cut inputs. Permute these tensor coordinates so that the incoming indices of the next factor are adjacent. Apply that factor's matrix unfolding to them and the identity to all other retained coordinates, then permute the outputs into the new retained list. Matrix multiplication performs precisely the sums over the incoming indices. Induction over the factors therefore accounts for every uncut index once and leaves only the b cut outputs at the end. If Q denotes the resulting map from $(\mathbb{R}^d)^{\otimes b}$ to itself, then

$$\|Q\|_{\text{op}} \leq \prod_{i=1}^{2s} |\nabla^{j+e_i} V|_{\text{split}} \prod_{v=1}^{n_H} |\nabla^{h_v} \nabla^2 H_\rho|_{\text{split}}.$$

Indeed coordinate permutations and the intervening identity maps have operator norm one. Closing the original cut is now exactly a single matrix trace, so the absolute value of this completely contracted term is at most

$$|\text{tr } Q| \leq d^b \|Q\|_{\text{op}}. \quad (4.6)$$

This argument explains why other undirected cycles introduce no further dimension factor: their coordinate sums have already been performed by the matrix compositions.

Uniform high-moment bookkeeping. The coordinate expansion and its contraction estimate are now complete. It remains to verify that their constants allow s to grow with d and j . Apply the induction hypothesis to every primary factor with moment exponent $2s$, and use the pointwise Hessian estimates on all other factors. There are exactly $2s$ random primary factors, so Hölder's inequality uses no larger exponent. The power of $1/\sqrt{a}$ before taking the $2s$ -th root is at most

$$2s(k_V + j) + \sum_i e_i + \sum_v h_v \leq 2s(k_V + j + 1),$$

by (4.5). Since $e_i, h_v \leq 2s$, all polynomial bases in the inductive estimates are at most $C(s + j + \log(d + 1))$. The total exponent of those bases is at most

$$C\left(2s(j + 1) + \sum_i e_i + \sum_v (h_v + 1)\right) \leq C's(j + 1),$$

using (4.5). The term count $(Cs)^{Cs}$ is absorbed in the same bound. The cut bundle uses only the original free slots of $T = \nabla^*(\nabla^j V)$, not the new indices introduced by integration by parts. Consequently $b \leq j + C_V$, where C_V depends only on the fixed number of slots of V , and not on s . We have proved

$$\mathbb{E}\|T\|_{\text{op}}^{2s} \leq d^{j+C_V} [C(s + j + \log(d + 1))]^{Cs(j+1)} (\sqrt{a})^{-2s(k_V+j+1)}.$$

Choose s to be the smallest positive integer satisfying $2s \geq p$ and $s \geq (j + C_V) \log(d + 1)$, increasing C_V if necessary. Then $d^{(j+C_V)/(2s)} \leq e^{1/2}$ and

$$s \leq C_V(p + (j + 1) \log(d + 1)).$$

The expression on the right is bounded by a fixed multiple of $(p + j + \log(d + 1))^2$. Taking the $2s$ -th root, using monotonicity of probability-space L^p norms, and increasing C_E now gives (4.4) with $k = k_V + 1$.

All integration-by-parts identities above are legitimate for each finite j, s, d and each $\rho < 1$: the conditional-integral regularity in Proposition 3.2, together with the score formula, gives polynomial growth for every field and derivative appearing in the finite expansion, and the laws ξ_ρ have Gaussian tails, so cutoffs justify the identities with vanishing boundary terms. Uniformity in growing s is supplied by the explicit estimates after the identities, not by an unproved uniform regularity assertion. This completes the induction. $\square$

Where the dimension went. After opening the trace, all coordinate sums except one have been performed by matrix compositions of single-slot contractions, each of which is dimension-free by Lemma 4.2. The one remaining sum is a trace over $b \leq j + C_V$ indices and costs d^b . This cost is then divided among $2s$ primary factors, so it becomes $d^{(j+C_V)/(2s)}$, and choosing the moment order $s \gtrsim (j + \log d)$ makes this factor $O(1)$. The price is a polynomial factor $(p + j + \log d)^{C(j+1)}$ in the moment order, which is why the final estimates contain powers of $\log d$ rather than powers of d .

4.3 One final vector dimension factor

The preceding theorem controls every tensor expression while at least two slots remain. Only the last divergence producing a vector will cost a $\sqrt{d}$ factor. The following elementary identity makes that cost explicit.

Lemma 4.5 (Frobenius divergence estimate). Let $\mu(dy) \propto e^{-H(y)} dy$, where $H \in C^2(\mathbb{R}^d)$ has bounded Hessian and μ has Gaussian tails. Let V be a smooth tensor field with a distinguished divergence slot and with $V, \nabla V \in L^2(\mu)$ in Frobenius norm. Then

$$\|\nabla^* V\|_{F,2} \leq C(\|V\|_{F,2} + \|\nabla V\|_{F,2}), \quad (4.7)$$

where C depends only on $\|\nabla^2 H\|_{\text{op},\infty}$, and is independent of d and the number of remaining tensor slots. For $H = H_\rho$ it is an absolute constant.

Proof. First take a compactly supported smooth vector field, so that the output is scalar. Integrating by parts twice gives

$$\int (\nabla^* V)^2 d\mu = \int \sum_{i,j} [(\partial_i V_j)(\partial_j V_i) + V_i(\partial_{ij} H)V_j] d\mu.$$

The first term is bounded in absolute value by $|\nabla V|_F^2$, and the second by $\|\nabla^2 H\|_{\text{op}}|V|^2$. For tensor output, apply the scalar identity to each fixed collection of remaining indices and sum. Taking a square root proves (4.7) for compactly supported fields. For general V , multiply by smooth cutoffs χ_R equal to one on the ball of radius R , with $|\nabla \chi_R| \leq C/R$. Then $\chi_R V \rightarrow V$ and $\nabla(\chi_R V) \rightarrow \nabla V$ in $L^2(\mu)$, while their adjoint divergences converge pointwise. Fatou's lemma proves the stated estimate. All fields to which we apply it have polynomially growing derivatives and hence satisfy the required integrability. $\square$

Corollary 4.6 (Derivatives along transported paths). For each fixed integer $m \geq 1$,

$$\|\mathcal{D}^m M_\rho\|_{L^2(\xi_\rho)} \leq \text{polylog}(d) L\sqrt{d} a^{-m}. \quad (4.8)$$

For fixed integers $q \geq 1$, $m \geq 0$, and fixed real $2 \leq p < \infty$, every nontrivial split satisfies

$$|\mathcal{D}^m U_q|_{\text{split}}^{L^p(\xi_\rho)} \leq \text{polylog}(d) L(\sqrt{a})^{-(q-1+2m)}. \quad (4.9)$$

The constants hidden in $\text{polylog}(d)$ may depend on m, q, p when these parameters occur, but are uniform in ρ , the primitive center, scale, and oracle.

Proof. Iterate Lemma 4.1, using the product rule on contractions. Replacing U_q by $\nabla^* U_{q+1}$ raises weight by two. Replacing it by one of the contracted products in (4.2) does not raise weight: the two jet weights sum to $q - 1$. The commutator (4.3) inserts $\mathcal{D}M = \rho U_1$ of weight zero on the divergence slot. Every additional jet factor L is accompanied by a factor r , so, after retaining one overall L , the extra factors are powers of $l = rL \leq 1$. Scalar coefficients are fixed polynomials in ρ and remain bounded under the fixed number of differentiations. Starting from a jet U_q , all divergences retain at least two free slots, and all contractions retain a slot of each factor. The resulting terms are therefore expressions of Theorem 4.4, of weight at most $q - 1 + 2m$. Its case $j = 0$ proves (4.9).

Starting instead from M , the first identity is $\mathcal{D}M = \nabla^* U_1$. Its outermost divergence remains a divergence to a vector under further material differentiations. Inside it, the same construction gives finitely many normalized expressions of weight at most $2(m - 1)$, each with one overall factor L . Apply Lemma 4.5 at this outermost operation. For a tensor with at least two slots, its Frobenius norm is at most $\sqrt{d}$ times the operator norm of an unfolding that puts one coordinate slot on one side: that matrix has at most d singular values. Thus Theorem 4.4, for $j = 0, 1$ and $p = 2$, bounds the right side of (4.7) by

$$\text{polylog}(d) L \sqrt{d} (\sqrt{a})^{-(2m-1)} \leq \text{polylog}(d) L \sqrt{d} a^{-m}.$$

This is (4.8). □

To see the quantitative connection with quadrature, start the transport at its correct marginal and put $Q_\rho = M_\rho(Y_\rho)$. Proposition 3.3 ensures that $Y_\rho \sim \xi_\rho$, so (4.8) bounds $\|Q_\rho^{(m)}\|_2$ for every probability time. On a cell $[\rho_0, \rho_1] \subset [0, 1)$ of length h_c , Taylor's integral remainder and the L^2 triangle inequality give, for every $\rho \in [\rho_0, \rho_1]$,

$$\left\| Q_\rho - \sum_{h=0}^{m-1} \frac{(\rho - \rho_0)^h}{h!} Q_{\rho_0}^{(h)} \right\|_2 \leq \frac{h_c^m}{m!} \sup_{t \in [\rho_0, \rho_1]} \|Q_t^{(m)}\|_2 \leq C_m \text{polylog}(d) L \sqrt{d} h_c^m (1 - \rho_1^2)^{-m}.$$

In particular, for a mesh ratio $\theta > 0$, the condition $h_c \leq \theta(1 - \rho_1)$ makes this remainder at most $C_m \text{polylog}(d) L \sqrt{d} \theta^m$. This is a supremum of deterministic RMS bounds. It does not require an expected supremum over a random trajectory. The construction of Proposition 6.1 uses exactly this cancellation between cell length and endpoint singularity in polynomial quadrature. Before constructing those finite formulas, we use the material identity itself to produce a stationary flow whose path integral represents a noisy mean.

5 A stationary flow for noisy means

The probability transport turns conditional means into samples. We now construct the complementary operation: a path integral whose law is $N(u, s^2 I_d)$ for a prescribed $s > 0$. The key is to represent the centered observable $M_* - u$ as a matrix adjoint divergence. That matrix couples a position with law ξ_* to a Gaussian momentum while preserving their product law, and integrating the momentum

equation then produces the noisy mean. This is still an exact mathematical construction: both its velocity and its ideal initial law will need to be implemented later.

Retain the primitive and interpolation notation of Sections 2–4. Thus $L = \lambda r$, $l = \lambda r^2 \leq 1/2$, $M_0 = u$, and the symmetric first jet satisfies $DM_\rho = \rho U_1$. Fix

$$0 < R < 1/2, \quad b_* = (1 - R^2)^{1/2},$$

and use a star for evaluation at $\rho = b_*$. The positive smoothing radius R keeps all coefficients smooth, although F itself is only C^2 . In this section ρ is the Gaussian-interpolation parameter, and the new stationary flow will use a separate time variable t .

The plan of the section. We want to turn the exact conditional mean $u = M_0$ into a *noisy* random variable $u + sG$. The strategy is to write the fluctuation $M_*(y) - u$ of the mean at a nearby interpolation time as a divergence, $\nabla^* K = M_* - u$, for an explicit matrix field K built from the transport, and then to run a Hamiltonian-style flow whose momentum accumulates this divergence. Section 6 will discretize both objects. The present section proves they exist, are smooth, and have the derivative bounds that quadrature needs.

5.1 Transport Jacobians and the zero endpoint

Let $T_\rho = T_{\rho \rightarrow b_*}$ denote the probability transport (3.9) from time ρ to time b_* . Given a terminal position y , define

$$Y_\rho(y) = T_\rho^{-1}(y), \quad B_\rho(y) = DT_\rho(Y_\rho(y)) = (DY_\rho(y))^{-1}.$$

In particular $Y_{b_*}(y) = y$. The backward path satisfies $Y'_\rho = -rM_\rho(Y_\rho)$ with this terminal condition. Its spatial Jacobian $J_\rho = DY_\rho$ solves

$$J'_\rho = -rDM_\rho(Y_\rho)J_\rho, \quad J_{b_*} = I_d.$$

Differentiating its inverse gives the order of multiplication in

$$B'_\rho = rB_\rho DM_\rho(Y_\rho) = r\rho B_\rho U_1(\rho, Y_\rho), \quad B'_0 = 0. \quad (5.1)$$

The last equality is an exact endpoint identity, not an asymptotic estimate: at $\rho = 0$ the right side contains the factor ρ . Also $\|B_\rho\|_{\text{op}} \leq C$, because $r\|DM_\rho\|_{\text{op}} \leq Cl$ and the time interval has length at most one. The following estimates control both spatial differentiation of the kernel and time interpolation of B_ρ .

Lemma 5.1 (Transport derivatives). For each fixed $k \geq 1$, the order- k spatial derivatives of any transport between two times in $[0, b_*]$, after subtracting the identity map, have all nontrivial split norms at most

$$C_k l R^{-(k-1)}.$$

For each fixed $m \geq 1$ and fixed finite $p \geq 1$, uniformly in $0 \leq \rho \leq b_*$,

$$\| \|B_\rho^{(m)}(y)\|_{\text{op}} \|_{L^p(y \sim \xi_*)} \leq \text{polylog}(d) l (1 - \rho^2)^{-m}.$$

The constants and the polylogarithmic factor may depend on the fixed orders and on p , but not on the primitive center or scale.

Proof. Differentiate the transport equation with respect to its starting position. Gronwall's inequality gives a bounded first derivative and an $O(l)$ bound for its difference from I_d . For order $k > 1$, the chain rule has terms consisting of an order- j derivative of the velocity and derivatives of the state of orders $i_1, \dots, i_j \geq 1$, where $i_1 + \dots + i_j = k$. By (3.5), the outer tensor has every split norm at most $C_j l R^{-(j-1)}$. Attach the inner tensors one at a time. Each attachment is a single contraction retaining the outer output and the inner derivative slots, so Lemma 4.2 applies without a dimension factor. Bound a first state derivative by a constant and use the induction hypothesis for all higher state derivatives. Since $l \leq 1$, the total power of R^{-1} is at most

$$(j-1) + \sum_{\nu=1}^j (i_\nu - 1) = k - 1,$$

and each forcing term retains the outer factor l . The term involving the unknown order- k state derivative is linear, with coefficient bounded by C_l , and this derivative starts at zero. Gronwall proves the assertion. The same argument with absolute interval lengths applies to transport in the reverse direction.

For the second assertion, write $C_\rho = rDM_\rho(Y_\rho) = r\rho U_1(\rho, Y_\rho)$, so that $B'_\rho = B_\rho C_\rho$. The product rule expresses $B_\rho^{(m)}$ as a finite sum of ordered products of B_ρ and $k \geq 1$ factors $C_\rho^{(j_1)}, \dots, C_\rho^{(j_k)}$, with $j_1 + \dots + j_k = m - k$. The zero-order jet bound and (4.9) imply, for every fixed moment order,

$$\| \|C_\rho^{(j)}\|_{\text{op}} \|_{L^p} \leq \text{polylog}(d) l (1 - \rho^2)^{-j}.$$

Here the path estimates apply because inverse transport sends $y \sim \xi_*$ to $Y_\rho(y) \sim \xi_\rho$. Apply Hölder with sufficiently high fixed moments to each product. Using $l^k \leq l$ and $(1 - \rho^2)^{-(m-k)} \leq (1 - \rho^2)^{-m}$ gives the claimed conservative exponent. Moments below one, if desired, follow from the L^1 bound. $\square$

5.2 A kernel for the centered observable

For a matrix field K , define its row-wise adjoint divergence by

$$(\nabla_{\xi_*}^* K)_i = \sum_j \{ -\partial_j K_{ij} + (\partial_j H_*) K_{ij} \}.$$

Thus $\nabla_{\xi_*}^* K = h$ means $\mathbb{E}_{\xi_*}[h_i \phi] = \mathbb{E}_{\xi_*} \sum_j K_{ij} \partial_j \phi$ for smooth compactly supported scalar tests ϕ . We seek this identity with $h = M_* - u$.

This is a Stein-type integration-by-parts identity. A classical coordinate Stein kernel for a law ξ represents the observable $y - \mathbb{E}_\xi y$ instead. Courtade, Fathi and Pananjady construct such kernels from spectral-gap assumptions [14, arXiv version 2, Definition 2.1 and Theorem 2.4]. Our observable is $M_* - u$, whose mean is zero by conditional expectation. The transport formula below supplies its particular adjoint-divergence representation together with pointwise derivative bounds. Those bounds are needed for quadrature and do not follow from the cited existence statement. We do not require K to be symmetric or positive.

Define

$$K(y)^T = \int_0^{b_*} B_\rho(y) U_1(\rho, Y_\rho(y)) d\rho = \frac{1}{r} \int_0^{b_*} B'_\rho(y) \frac{d\rho}{\rho}. \quad (5.2)$$

The first integral defines the integrand at zero in the second formula, because $B'_\rho/(r\rho) = B_\rho U_1(\rho, Y_\rho)$ extends smoothly there by (5.1). The transpose is essential: the adjoint divergence will act on the rows of K , while the position velocity will use K^T .

Lemma 5.2 (The centering kernel). The field K is smooth and satisfies

$$\nabla_{\xi_*}^* K = M_* - u, \quad \|K(y)\|_{\text{op}} \leq CL.$$

For each fixed $j \geq 1$, every nontrivial split norm of $\nabla^j K$ is bounded pointwise by $C_j LR^{-j}$.

Proof. Fix an output coordinate i and a smooth compactly supported scalar function ϕ . The transported test $\phi \circ T_\rho$ has zero material derivative: it is constant along each probability-transport path. Differentiate its product with $M_{\rho,i}$ under the transported laws. Lemma 4.1 gives $\mathcal{D}M_\rho = \nabla_{\xi_\rho}^* U_1$, so integration by parts and the chain rule yield

$$\frac{d}{d\rho} \mathbb{E}_{\xi_\rho} [M_{\rho,i} \phi \circ T_\rho] = \mathbb{E}_{\xi_\rho} \sum_{j,k} (U_1)_{ik} (DT_\rho)_{jk} (\partial_j \phi) \circ T_\rho.$$

Smooth transport on the compact time interval and Gaussian tails justify differentiation. Alternatively the test can first be cut off and the resulting identities passed to the limit. Push the expectation forward to $y \sim \xi_*$. Since U_1 is symmetric, its coefficient of $\partial_j \phi(y)$ is

$$\sum_k (U_1)_{ik}(\rho, Y_\rho(y)) (B_\rho(y))_{jk} = (B_\rho(y) U_1(\rho, Y_\rho(y)))_{ji}.$$

The terminal expectation is $\mathbb{E}_{\xi_*} [M_{*,i} \phi]$. At zero, $M_0 = u$ and $(T_0)_\# \xi_0 = \xi_*$, so the initial expectation is $u_i \mathbb{E}_{\xi_*} \phi$. Integrating the displayed derivative in ρ proves

$$\mathbb{E}_{\xi_*} [(M_{*,i} - u_i) \phi] = \mathbb{E}_{\xi_*} \sum_j K_{ij} \partial_j \phi.$$

This proves the adjoint-divergence identity weakly, and smoothness makes it a pointwise identity.

For the bounds, $\|B_\rho\|_{\text{op}} \leq C$ and $\|U_1\|_{\text{op}} \leq CL$ immediately give $\|K\|_{\text{op}} \leq CL$. For a positive-order spatial derivative, use Lemma 5.1 on both DT_ρ and Y_ρ . An order- j derivative of DT_ρ has all split norms at most $C_j LR^{-j}$, and an order- j derivative of U_1 has all split norms at most $C_j LR^{-j}$ by (3.5). When these fields are composed with Y_ρ , attach each chain-rule tensor by a single contraction as in the preceding lemma. First derivatives of Y_ρ are bounded, and a derivative of order $i > 1$ costs $C_i l R^{-(i-1)}$. Every order- j term in the product has total cost at most $C_j LR^{-j}$, since $l \leq 1$. Each factor retains free slots, so the single-contraction bound preserves every nontrivial split. Integrating over an interval of length at most one proves the assertion. All orders are smooth by Gaussian conditional integration and the smooth transport on $\rho \leq b_* < 1$. This uses no higher derivatives of the original unsmoothed gradient. $\square$

5.3 Stationarity and the path integral

The following dynamics specialize the skew-matrix construction of Ma, Chen and Fox [20, Section 2.1, Equation (3), Theorem 1]. The transport-derived K makes the momentum displacement encode $M_* - u$. We prove completeness and invariance directly, since a formal stationary-density calculation alone would not ensure a solution exists for the full integration interval.

Proposition 5.3 (A noisy mean from a path integral). For $s > 0$, let (Y_0, G_0) have law $\xi_* \otimes N(0, I_d)$ and solve

$$\dot{Y}_t = -K(Y_t)^T G_t / s, \quad \dot{G}_t = (M_*(Y_t) - u) / s. \quad (5.3)$$

The solution exists for every finite time, and its law remains $\xi_* \otimes N(0, I_d)$. In particular,

$$sG_0 + \int_0^1 M_*(Y_t) dt \stackrel{\text{law}}{=} u + sG, \quad G \sim N(0, I_d). \quad (5.4)$$

Proof. The vector field v in (5.3) is smooth and hence locally Lipschitz. The bound on K and the global Lipschitz bound on M_* imply $|v(y, g)| \leq C_{x,r,s}(1 + |y| + |g|)$. Gronwall's inequality therefore prevents a solution from escaping to infinity in finite positive or negative time. Local uniqueness gives a complete smooth flow Φ_t , with inverse Φ_{-t} . We only need local Lipschitz continuity here. The factor G does not in general give a globally Lipschitz joint vector field.

Write the product density as

$$p(y, g) = (2\pi)^{-d/2} \exp(-H_*(y) - |g|^2/2).$$

Differentiating in the position and momentum variables separately, with the row convention of the kernel lemma, gives

$$\operatorname{div}_y(pv_Y) = \frac{p}{s} \sum_i g_i (\nabla_{\xi_*}^* K)_i = \frac{p}{s} g \cdot (M_* - u), \quad \operatorname{div}_g(pv_G) = -\frac{p}{s} g \cdot (M_* - u).$$

Thus $\operatorname{div}(pv) = 0$. The Jacobian formula for the complete flow now yields, pointwise,

$$\frac{d}{dt} \{p(\Phi_t(z)) \det D\Phi_t(z)\} = \operatorname{div}(pv)(\Phi_t(z)) \det D\Phi_t(z) = 0.$$

The determinant starts at one and stays positive. Change of variables therefore proves that Φ_t preserves the probability density p .

For later differentiation it is useful to record the same cancellation as an operator identity. Relative to the product law, the Lie derivative is

$$\mathcal{L} = \nabla^* A \nabla, \quad A = \begin{pmatrix} 0 & K^T/s \\ -K/s & 0 \end{pmatrix}. \quad (5.5)$$

Indeed the second-order part vanishes by skew symmetry. The first-order coefficients are $-K^T g/s$ in position and $\nabla_{\xi_*}^* K/s = (M_* - u)/s$ in momentum, exactly as in (5.3).

Finally, integrating the momentum equation over $[0, 1]$ gives the pathwise identity

$$sG_0 + \int_0^1 M_*(Y_t) dt = u + sG_1.$$

Stationarity makes G_1 a standard Gaussian, proving (5.4). $\square$

Stationarity also makes G_1 independent of Y_1 . It does not assert that G_1 is independent of the initial anchors (Y_0, G_0) . The exact mean $u = M_0$ remains in the momentum velocity, and the proposition assumes an ideal sample from ξ_* . Thus the proposition is a comparison construction, not yet an oracle routine. Section 6 will express its approximation by finitely many exact mean evaluations, and Section 7 will replace every such evaluation, including M_0 . The displayed output itself needs no separate addition of u .

Why the skew coupling works. The observable we want to produce, $M_* - u$, is a divergence $\nabla_{\xi_*}^* K$. In the position variable it would simply be the drift of a diffusion, but then the accumulation would happen in the wrong space. The skew coupling instead moves the divergence into the momentum equation, where it integrates directly into a Gaussian displacement: the momentum gains $\frac{1}{s} \int_0^1 (M_*(Y_t) - u) dt$, and the stationarity of the coupled flow guarantees that this displacement plus the initial G_0 has exactly the law $N(0, I_d)$, shifted by u/s . The position drift $-K^T G/s$ has no effect on the law. It exists only to keep the product measure invariant.

5.4 Derivatives on a stationary trajectory

To approximate the path integral by high-order quadrature, we need bounds on time derivatives at each fixed time. Stationarity allows these bounds to be proved in the fixed product measure. The following lemma separates that argument from the particular kernel and avoids requiring a bound on the supremum of a random path.

Lemma 5.4 (Stationary time derivatives). Fix $0 < R \leq 1$. Let a smooth probability density on $\mathbb{R}^n$, with $n \leq Cd$, have Gaussian tails and negative logarithm H . Assume its Hessian is bounded and that every positive-order score derivative $\nabla^j \nabla H$ has every one-slot split norm at most $C^j R^{-(j-1)}$, for $j \geq 1$. Here a one-slot split isolates one tensor slot as the matrix input and groups all remaining slots as its output. Suppose a stationary autonomous flow has Lie derivative $\mathcal{L} = \nabla^* A \nabla$, where the adjoint is for this density, A is skew symmetric, and

$$|\nabla^j A|_{\text{one-slot split}} \leq C^j \alpha R^{-j} \quad (j \geq 0)$$

for every one-slot split. Let ϕ_0 be smooth and scalar- or vector-valued, with

$$|\nabla^j \phi_0|_F \leq C^j a_0 \sqrt{d} R^{-j} \quad (j \geq 1).$$

Assume the fields and all derivatives used have polynomial growth. Then, for each fixed $k \geq 1$, the order- k time derivative of ϕ_0 on the stationary flow has RMS norm at most

$$C_k a_0 \sqrt{d} (\alpha R^{-2})^k. \quad (5.6)$$

Proof. Put $\phi_k = \mathcal{L}^k \phi_0$. We prove, for $j \geq 0$ and $k \geq 1$, the stronger estimate

$$\|\nabla^j \phi_k\|_{F,2} \leq C_{j,k} a_0 \sqrt{d} \alpha^k R^{-(j+2k)}. \quad (5.7)$$

For $k = 0$ the same bound is assumed only for $j \geq 1$, which suffices: the iteration differentiates ϕ_{k-1} before multiplying by A . All L^2 norms below are for the invariant density.

Set $V = A \nabla \phi_{k-1}$. Leibniz's rule and the induction hypothesis give, for each $h \geq 0$,

$$\|\nabla^h V\|_{F,2} \leq C_{h,k} a_0 \sqrt{d} \alpha^k R^{-(h+1+2(k-1))}.$$

To see why no dimension factor appears, in each coefficient tensor $\nabla^v A$ isolate the slot contracted against the gradient of ϕ_{k-1} . Its one-slot operator norm multiplies the Frobenius norm of the other factor. Summing the finitely many Leibniz terms preserves the displayed estimate.

Now commute j derivatives past the adjoint divergence in $\phi_k = \nabla^* V$. The main term is $\nabla^*(\nabla^j V)$. The Frobenius estimate (4.7) bounds its L^2 norm by a constant times $\|\nabla^j V\|_{F,2} + \|\nabla^{j+1} V\|_{F,2}$. These two terms have powers $R^{-(j+2k-1)}$ and $R^{-(j+2k)}$, respectively. Each remaining commutator contracts an order- v derivative of the score, $1 \leq v \leq j$, with $\nabla^{j-v} V$. The coefficient's one-slot norm again suffices, and the total exponent of R^{-1} is

$$(v-1) + (j-v+1+2(k-1)) = j+2k-2.$$

All these terms are bounded by (5.7) because $0 < R \leq 1$. This closes the induction, including $j = 0$ for $k \geq 1$. Gaussian tails and polynomial growth justify the integrations by parts through cutoffs.

Along an autonomous trajectory, the k th time derivative of ϕ_0 is ϕ_k evaluated at the current state. Stationarity identifies its RMS norm at every time with the fixed-measure norm in (5.7) for $j = 0$. $\square$

For (5.3), the ambient dimension is $2d$, and (5.5) satisfies the lemma's coefficient bounds with $\alpha = L/s$. The product potential is $H_*(y) + |g|^2/2$, and its score-derivative bounds follow from (3.5) and the Gaussian factor. For the joint state take $a_0 = 1$: its first derivative has Frobenius norm $\sqrt{2d}$ and all higher derivatives vanish. For M_* take $a_0 = L$. Indeed, isolating any derivative slot in (3.5) gives a matrix with input dimension d , so passage from its operator norm to its Frobenius norm costs at most $\sqrt{d}$. Consequently, at every stationary time,

$$\left\| \frac{d^k}{dt^k}(Y_t, G_t) \right\|_2 \leq C_k \sqrt{d} ((L/s)R^{-2})^k, \quad \left\| \frac{d^k}{dt^k} M_*(Y_t) \right\|_2 \leq C_k L \sqrt{d} ((L/s)R^{-2})^k. \quad (5.8)$$

The strict radius choice $0 < R < 1/2$ made for this construction satisfies the abstract lemma's $0 < R \leq 1$ hypothesis.

We also need an initial displacement estimate for the later iteration. Conditional expectation and the primitive Poincaré inequality give

$$\mathbb{E}_{\xi_*} |M_* - u|^2 \leq \mathbb{E}_\nu |f - u|^2 \leq CL^2 d.$$

At every stationary time, G_t is independent of Y_t , hence

$$\mathbb{E} |K(Y_t)^T G_t|^2 = \mathbb{E} \|K(Y_t)\|_F^2 \leq CL^2 d.$$

Integrate the velocity and apply Minkowski's inequality to obtain

$$\|(Y_t, G_t) - (Y_0, G_0)\|_2 \leq Ct(L/s)\sqrt{d} \leq C(L/s)\sqrt{d} \quad (0 \leq t \leq 1). \quad (5.9)$$

These are fixed-time RMS estimates for the exact stationary flow. They give the derivative and starting-error bounds used in the next section to approximate its path integral by finite formulas.

6 Finite formulas for the two flows

We now turn the probability transport and the stationary centering flow into finite formulas in conditional means. The formulas must meet two requirements at once. Their error on the prescribed input law must be small, and each new mean must be evaluated at a center that is an affine expression in earlier mean values, with scalar coefficients and controlled size. The latter property permits the recursive implementation of Section 7.

The main conclusion is Theorem 6.7. Its proof has three numerical steps. Global Picard iteration discretizes probability transport without making the dependency depth proportional to the number of time cells. Short harmonic trajectories then give a formula for the vector $K(y)^T G$ in the centering velocity. Finally, Picard iteration discretizes that stationary flow. Two cancellations are needed in the middle step: a zero derivative removes the singularity at the lower endpoint of the kernel integral, and retaining only a forward transport correction cancels its apparent inverse power of the primitive scale r .

Polynomial collocation of Picard iteration was developed for sampling by Lee, Song and Vempala [18, Algorithm 2 and Theorem 2.3]. Gaussian smoothing and high-order Picard approximation also appear in [9, Appendix E, Algorithm E.1 and Theorem E.2]. Here we need, in addition, scalar affine formulas whose dependency depth can be fixed before their interpolation order. An exact conditional mean is an analytical node throughout this section, and Section 7 replaces these nodes by finite oracle computations.

6.1 Parameters and the order of choices

Fix an integer $K_0 \geq 2$ and $t > 0$ such that

$$40(K_0 + 2)t < \frac{1}{4}, \quad R = \delta^t, \quad \psi = \delta^{4t}, \quad h_0 = \delta^w, \quad 0 < w < t. \quad (6.1)$$

The iteration count will first fix a depth bound. We then choose w and finally the interpolation orders, and increasing those orders will not change the depth bound. For the centering flow set $s = \sigma/2$, where $\sigma > 0$ is the parent mean-noise parameter. Thus the centering formula itself will approximate noise of standard deviation $\sigma/2$. Throughout its numerical analysis assume

$$\max\{l, L/\sigma\} \leq \text{polylog}(d)\delta^b, \quad b \geq \frac{1}{2}. \quad (6.2)$$

For probability transport alone only the corresponding bound on l is required. All probability-time intervals lie in $[0, (1 - cR^2)^{1/2}]$ for a fixed $c > 0$. Constants may depend on fixed interpolation orders and fixed parameters, but not on the center, dimension, primitive scale r , noise parameter, or oracle. A symbol B_d below denotes a bound $C(1 + \log d)^C$ with precisely this uniformity, where the fixed exponent and constant in a size hypothesis are part of the fixed parameters. All assertions are for sufficiently large d .

The order of the parameter choices. It is worth fixing in advance the order in which the numbers are chosen, because the implementation of Section 7 depends on it. First choose e, K_0, t as in Theorem 2.1. Then, for the circuits, fix the Picard counts: this fixes the dependency depth D_0 of Theorem 6.7 and the routine height H of Lemma 7.2. Then choose the mesh exponent w , and only then the interpolation order m . The reason is that the depth bounds do not involve w or m , so the implementation can choose w after seeing how many routines it needs, while the approximation orders are free to be as large as necessary.

6.2 Probability transport by global Picard iteration

Partition a probability-time interval using equal steps of size at most h_0 in $-\log(1 - \rho)$. For a nonzero interval take the ceiling of its transformed length divided by h_0 as the number of cells. Cell lengths ℓ_c then obey $\ell_c \leq Ch_0(1 - \rho)$ on that cell, and the number of cells is

$$O(1 + h_0^{-1}(1 + |\log R|)).$$

A zero-length interval requires no operations. On each cell interpolate at $m \geq 2$ distinct equally spaced nodes, including both ends, and integrate the resulting degree- $(m - 1)$ polynomial. Include partial integrals to every state node. By scaling the fixed Lagrange basis on $[0, 1]$, its absolute value sum is at most C_m , and its order- k derivative absolute sum is at most $C_m \ell_c^{-k}$. Consequently every partial-integral row has absolute weight sum at most C_m , uniformly in the number of cells.

Let w_{ij} be these partial-integral weights from the starting time ρ_{in} to node ρ_i , with signed weights for backward integration. Duplicate shared endpoints as labeled nodes if convenient. Given a starting position y_{in} , use simultaneous updates

$$y_i^{[0]} = y_{\text{in}}, \quad y_i^{[k+1]} = y_{\text{in}} - r \sum_j w_{ij} M_{\rho_j}(y_j^{[k]}). \quad (6.3)$$

This iteration integrates over the whole interval at every layer. It does not advance successively from cell to cell. That is the point of the “global” in the title: all nodes depend only on the preceding layer, so the dependency depth is the number of layers and not the number of cells.

Proposition 6.1 (Probability-flow circuit). If the input position has marginal law $\xi_{\rho_{\text{in}}}$, then after N_p iterations the maximum, over all state nodes, of the RMS error relative to the exact transport is at most

$$C_m \text{polylog}(d) l \sqrt{d} h_0^m + (C_m l)^{N_p} C r D. \quad (6.4)$$

Every computed state is globally Lipschitz in its initial position with constant C_m . After subtracting the initial position, its Lipschitz constant is at most $C_m l$.

Proof. Write Y_ρ for the exact path from the given random input and $Q_\rho = M_\rho(Y_\rho)$. The marginal identity in Proposition 3.3 and (4.8) give

$$\|Q_\rho^{(m)}\|_2 \leq B_d L \sqrt{d} (1 - \rho^2)^{-m}.$$

On a cell $I = [a_c, b_c]$ of length ℓ_c , let P be the Taylor polynomial of Q of degree $m - 1$ at a_c . Minkowski's inequality applied to the integral remainder gives, at each $v \in I$,

$$\|Q_v - P(v)\|_2 \leq \frac{1}{(m-1)!} \int_{a_c}^v (v-z)^{m-1} \|Q_z^{(m)}\|_2 dz \leq C_m B_d L \sqrt{d} h_0^m. \quad (6.5)$$

The last step uses $\ell_c \leq C h_0 (1 - z)$ on that cell and $1 - z^2 \geq 1 - z$. The same estimate holds at the interpolation nodes. Because interpolation reproduces P and its basis has bounded absolute sum, $\|Q_v - \mathcal{I}_c Q(v)\|_2$ obeys the same bound. Integrate this pointwise estimate over the cells, whose total length is at most one. This proves the required quadrature error, for partial integrals and for either orientation. No supremum of the random path has been taken.

Let $e_k = \max_i \|y_i^{[k]} - Y_{\rho_i}\|_2$. The global bound $\text{Lip}(M_\rho) \leq CL$, together with the absolute row sums, gives

$$e_{k+1} \leq C_m l e_k + C_m B_d L \sqrt{d} h_0^m.$$

On each exact transported marginal, conditional expectation and the primitive moment bound give $\|M_\rho\|_2 \leq CD$. Integrating the velocity shows $e_0 \leq C r D$. Since $C_m l < 1/2$ for large d , summing the geometric recurrence proves (6.4). The norm here is the maximum of the nodewise RMS errors. It is not the RMS of their maximum.

For deterministic inputs the same recurrence, now in maximum node norm, gives a bounded Lipschitz constant for every iterate. In the correction $y_i^{[k+1]} - y_{\text{in}}$, the factor r and the CL -Lipschitz mean give the stronger constant $C_m l$. $\square$

6.3 The kernel action through harmonic trajectories

The position velocity in (5.3) contains $K(y)^T G$. Approximating K and then multiplying the error by an arbitrary momentum would not give the global Lipschitz property needed for Picard iteration. Instead we approximate the vector action directly, using short harmonic trajectories and forward transport corrections.

For the kernel integral use the mesh on $[0, b_*]$ above. Its cell lengths also satisfy $\ell_c \geq c' h_0 (1 - \rho)$ for all sufficiently large d . This follows from the ceiling construction because the transformed interval length tends to infinity. Set $C_\rho(y) = B_\rho(y) - I$. We first describe a circuit for $C_\rho(y)G$ at one probability node ρ .

Backward position. Apply Proposition 6.1 backward from input position y at time b_* to obtain an approximation $\hat{Y}_\rho$ to $Y_\rho(y)$.

Short harmonic trajectories. At fixed probability time ρ , the Hamiltonian equations with potential H_ρ have position integral equation

$$W_z = \cos(z)Y_\rho + \sin(z)G - r\rho \int_0^z \sin(z-v)M_\rho(W_v) dv. \quad (6.6)$$

For every requested angle z , interpolate on the single cell $[0, z]$. Let z_i be its nodes and $\ell_j(v)$ its Lagrange basis, and set

$$w_{ij}^{\sin} = \int_0^{z_i} \sin(z_i - v)\ell_j(v) dv.$$

Use the simultaneous Picard updates

$$W_i^{[0]} = \cos(z_i)\hat{Y}_\rho + \sin(z_i)G, \quad W_i^{[k+1]} = \cos(z_i)\hat{Y}_\rho + \sin(z_i)G - r\rho \sum_j w_{ij}^{\sin} M_\rho(W_j^{[k]}). \quad (6.7)$$

At angle zero there is no integral. The free sine and cosine terms are retained exactly.

Differentiate the transport correction. Let d_j be the weights differentiating at zero the polynomial interpolant on angle nodes $0, \psi, \dots, (m-1)\psi$. Their absolute sum is $C_m\psi^{-1}$. At every such angle node, compute $T_\rho(W_z) - W_z$ by the forward probability-flow circuit, keeping only its integral correction at the output. For exact paths,

$$\left. \frac{d}{dz}(T_\rho - \text{id})(W_z) \right|_{z=0} = (B_\rho - I)G = C_\rho G, \quad (6.8)$$

since $W_0 = Y_\rho$ and $W'_0 = G$. The weighted sum with coefficients d_j therefore approximates $C_\rho G$ using exact mean nodes only.

Here and below the interpolation orders denoted by m may be increased independently. Using a common sufficiently large order merely simplifies the error displays. For the next estimates use the same number $N_p \geq 1$ of Picard layers in the backward, harmonic, and forward computations.

Lemma 6.2 (Errors in the angle computation). Let $y \sim \xi_*$ and $G \sim N(0, I_d)$ independently. The exact-input angle differentiation in (6.8) has RMS truncation error at most

$$C_m l \sqrt{d} \psi^{m-1} R^{-2m} = C_m l \sqrt{d} \delta^{2tm-4t}. \quad (6.9)$$

Before applying the angle differentiation weights, the RMS errors of the computed forward correction values are at most

$$C_m \text{polylog}(d) l \sqrt{d} (h_0^m + \delta^{2tm}) + (C_m l)^{N_p} C r D. \quad (6.10)$$

Every such approximate correction is globally Lipschitz in (y, G) with constant at most $C_m l$.

Proof. For exact inputs $(Y_\rho(y), G)$, harmonic motion $\dot{W} = P, \dot{P} = -\nabla H_\rho(W)$ preserves $\xi_\rho \otimes N(0, I_d)$. The vector field is globally Lipschitz, and its weighted divergence vanishes. Its Lie derivative is $\nabla^* A \nabla$ with constant skew coefficient whose (W, P) block is $-I$. Lemma 5.4 therefore applies with $\alpha = 1$, with $a_0 = l$ for $T_\rho - \text{id}$, and with $a_0 = L$ for M_ρ . Indeed Lemma 5.1 and (3.5) give the required spatial bounds, and changing a one-slot split bound to a Frobenius bound costs only $\sqrt{d}$. To apply the angle stencil, subtract the Taylor polynomial of $z \mapsto (T_\rho - \text{id})(W_z)$ of degree $m-1$ at zero. Its derivative at zero is exact, and its remainder at each stencil node has RMS norm at most $C_m l \sqrt{d} \psi^m R^{-2m}$. Multiplication by the absolute stencil sum C_m/ψ proves (6.9). This uses pointwise RMS remainders on the stationary harmonic path.

For the harmonic state equation, interpolate $M_\rho(W_v)$ in (6.6). Since $z \leq (m-1)\psi$, its Taylor remainder, multiplied by the prefactor r , is bounded by $C_m l \sqrt{d} \psi^m R^{-2m} = C_m l \sqrt{d} \delta^{2tm}$. The

additional integration factors are bounded and can be discarded. The initial free-path error is at most CrD , and Picard contracts by $C_m l$. The computed harmonic state is globally Lipschitz with bounded constant in its two inputs, by the same row-sum argument as before.

Next compare the forward probability correction started at the exact harmonic state. That state has marginal law ξ_ρ , so Proposition 6.1 applies. Errors in its starting state then propagate through the approximate correction with factor $C_m l$, not a factor of order one. This also controls the backward-position error transmitted through the harmonic computation. Combining these estimates proves (6.10) and the asserted global Lipschitz bound. $\square$

6.4 The zero endpoint of the kernel integral

It remains to integrate the derivative of $C_\rho G$ against $1/\rho$. The exact identity $B'_0 = 0$ removes the apparent singularity, and we impose that identity on the interpolant as well.

On all but the first probability cell use ordinary degree- $(m-1)$ interpolation. On the first cell use $m-1$ equally spaced value nodes including both endpoints and the additional Hermite condition that the derivative at zero is zero, for a total of m conditions, with $m \geq 3$. This uniquely determines a polynomial of degree at most $m-1$: a polynomial in the null space would have a double zero at zero and the remaining $m-2$ distinct zeros. Let q_i be the resulting weights on values when the derivative of this piecewise interpolant is integrated against $1/\rho$.

Lemma 6.3 (Derivative quadrature at zero). Let $y \sim \xi_*$ and $G \sim N(0, I_d)$ independently. For exact values of $C_\rho(y)G$, the RMS error of this quadrature for $\int_0^{b_*} B'_\rho(y)G d\rho/\rho$ is at most

$$C_m \text{polylog}(d) l \sqrt{d} h_0^{m-2}.$$

The value weights satisfy

$$\sum_i |q_i| \leq C_m h_0^{-1} (1 + |\log R| + |\log h_0|). \quad (6.11)$$

These weights are well defined even if the supplied values are approximate: the derivative condition at zero remains exact.

Proof. By Lemma 5.1 and independence of G , the RMS norm of the order- m derivative of $C_\rho(y)G$ is at most $\text{polylog}(d) l \sqrt{d} (1 - \rho^2)^{-m}$. Here multiplication by the independent Gaussian converts the matrix norm to a Frobenius norm and costs at most $\sqrt{d}$.

On an ordinary cell, Taylor expansion and scaling the interpolation basis bound the derivative error by this derivative bound times $C_m \ell_c^{m-1}$. It is therefore at most $C_m \text{polylog}(d) l \sqrt{d} h_0^{m-1} / (1 - \rho)$. After integration against $1/\rho$, all ordinary cells contribute at most this prefactor times $C(1 + |\log R| + |\log h_0|)$, which fits the claimed more conservative h_0^{m-2} bound.

For the first cell, subtract the Taylor polynomial at zero of degree $m-1$. Its derivative at zero agrees with the exact zero derivative. The second derivative of the remainder, and that of the Hermite interpolant of the remainder, are bounded in RMS by $C_m \text{polylog}(d) l \sqrt{d} \ell_c^{m-2}$. Both first derivatives vanish at zero. If E is the difference between the exact function and its Hermite interpolant on this cell, then

$$\|E'(\rho)\|_2 \leq \int_0^\rho \|E''(v)\|_2 dv \leq C_m B_d l \sqrt{d} \rho \ell_c^{m-2}.$$

Consequently the integral of $E'(\rho)/\rho$ over the first cell has RMS norm at most $C_m B_d l \sqrt{d} \ell_c^{m-1}$. This is within the stated, more conservative h_0^{m-2} bound.

On an ordinary cell the absolute coefficient sum of the interpolant derivative is at most C_m/ℓ_c . Integrating and using $\ell_c \geq c'h_0(1-\rho)$ gives a total bounded by

$$C_m h_0^{-1} \int_{\rho_{\text{first}}}^{b_*} \frac{d\rho}{\rho(1-\rho)}.$$

This gives the logarithms in (6.11). On the first cell, scale to the unit interval. The polynomial derivative vanishes at zero and is bounded by distance to zero times the second derivative, so the integral of its absolute value divided by that distance is bounded. Scaling back gives a weight sum $C_m/\ell_c \leq C_m h_0^{-1}$. This proves the claim. $\square$

What just happened, in order. The kernel integral $\int_0^{b_*} B'_\rho d\rho/\rho$ looks divergent at $\rho = 0$, and quadrature would indeed be inaccurate there if we interpolated the integrand naively. The exact endpoint identity $B'_0 = 0$ changes the game: imposing it on the interpolant makes the quadrature error one order better than the general bound, which exactly compensates the extra power of h_0 coming from the weight sum (6.11) and the $1/\rho$ factor.

6.5 A globally Lipschitz approximate centering velocity

Let $\text{corr}_{ij}(y, G)$ denote the forward correction computed at probability node i and angle node j by the three operations above. The approximate position velocity is the explicit expression

$$\hat{v}_Y(y, G) = -\frac{1}{rs} \sum_i q_i \sum_j d_j \text{corr}_{ij}(y, G). \quad (6.12)$$

Each correction retains only the integral part of the forward Picard formula. Writing its weights as w_{ijk} and its mean values as M_{ijk} gives

$$\text{corr}_{ij} = -r \sum_k w_{ijk} M_{ijk}, \quad \text{and hence} \quad \hat{v}_Y = \frac{1}{s} \sum_{i,j,k} q_i d_j w_{ijk} M_{ijk}.$$

The apparent factor $1/r$ has canceled algebraically. There is no additive forward state left in this formula, and all coefficients multiplying mean values are deterministic scalars. Use the exact mean expression $\hat{v}_G(y, G) = (M_*(y) - M_0)/s$ for the momentum velocity. In particular, the momentum occurs inside angle states. It is never multiplied by a separately approximated matrix at the output.

Proposition 6.4 (Kernel-action circuit). On the stationary product input law, the RMS error of $\hat{v} = (\hat{v}_Y, \hat{v}_G)$ relative to (5.3) is bounded by $\text{polylog}(d)$ times

$$\frac{L}{\sigma} \sqrt{d} [h_0^{m-2} + h_0^{-1} \delta^{-4t} (h_0^m + \delta^{2tm})] + \delta^{-4t-w} (C_m l)^{N_p-1} r \frac{L}{\sigma} D. \quad (6.13)$$

It is globally Lipschitz, with constant at most

$$C_m \text{polylog}(d) (L/\sigma) \psi^{-1} h_0^{-1}. \quad (6.14)$$

For every fixed $J \geq 1$, its RMS error can be made at most $\text{polylog}(d) \delta^J D$ with $N_p = O(J+1)$, independently of w and the interpolation orders.

Proof. Combine Lemmas 6.2 and 6.3, and divide by rs . The value errors are amplified only by $C_m\psi^{-1}$ and the weight sum (6.11). The exact stencil truncation has factor $\psi^{m-1}R^{-2m} = \delta^{2tm-4t}$ and the same probability-quadrature weight sum. Crucially,

$$\frac{l}{r\sigma} = \frac{L}{\sigma}, \quad \frac{l^{N_p}rD}{r\sigma} = l^{N_p-1}r \frac{L}{\sigma}D.$$

Thus neither division introduces a negative power of the potentially small scale r . The resulting estimate is (6.13).

Each forward correction has Lipschitz constant C_ml in its inputs, and the backward and harmonic maps have bounded constants. Applying the same absolute weight sums and using $l/(r\sigma) = L/\sigma$ gives (6.14), and the momentum component has the smaller Lipschitz bound CL/σ .

For the last assertion, increase interpolation orders until the first line of (6.13) is of the desired order. Under (6.2), sufficiently many $O(J+1)$ Picard layers make the second line of that order as well. The explicit common choice given below verifies independence from interpolation orders and from w . $\square$

6.6 The mean-flow circuit and the accuracy ledger

We can now discretize the centering flow itself. Interpolate on the single time cell $[0, 1]$, using ordinary degree- $(m-1)$ interpolation and partial-integral rows w_{ij} . With anchor $Z_0 = (Y_0, G_0)$, set

$$Z_i^{[0]} = Z_0, \quad Z_i^{[k+1]} = Z_0 + \sum_j w_{ij} \hat{v}(Z_j^{[k]}).$$

At the end keep the position components y_j and output

$$sG_0 + \sum_j \omega_j M_*(y_j), \quad \sum_j |\omega_j| \leq C_m, \quad (6.15)$$

where ω_j are the full-interval integration weights.

Proposition 6.5 (Mean-flow circuit). For ideal independent anchors $Y_0 \sim \xi_*$ and $G_0 \sim N(0, I_d)$, and every fixed $J \geq 1$, the circuit (6.15) can be chosen to differ in RMS by at most $\text{polylog}(d) \sigma \delta^J D$ from the exact path integral (5.4). The approximate joint states have RMS errors at most $\text{polylog}(d) \delta^J D$. Every integration uses $O(J+1)$ Picard layers, with a bound independent of w and all interpolation orders.

Proof. First compare the exact flow to polynomial integration of its exact velocity. Lemma 5.4, applied to the state with one additional derivative, bounds this quadrature error by

$$C_m \sqrt{d} ((L/s)R^{-2})^{m+1}.$$

The approximation error of $\hat{v}$ at every exact trajectory node is bounded by Proposition 6.4, by stationarity. Previous state errors propagate with its global Lipschitz constant and the absolute row sums. The true centering velocity need not be globally Lipschitz in the joint variables. It is the approximate velocity's bound that is used in this recurrence.

By (6.1), $b \geq 1/2$ and $w < t$ give $b - 4t - w > 1/3$ (indeed $4t + w < 5t < 1/8$). For sufficiently large d , fixed constants and polylogarithms in (6.14) can therefore be absorbed so that the recurrence contracts by at most $\delta^{1/3}$. Its initial error is at most $C(L/s)\sqrt{d}$ by (5.9). Choose the quadrature and velocity error orders sufficiently high, and then $O(J+1)$ Picard layers yield the state error claimed in the statement.

For the output, apply Lemma 5.4 to M_* with $a_0 = L$. Its path quadrature error is at most $C_m L \sqrt{d} ((L/s) R^{-2})^m$, which can be made at most $\text{polylog}(d) \sigma \delta^J D$. The additional error from evaluating on the approximate positions is bounded by their RMS error times $C_m L$, using the CL -Lipschitz bound for M_* . Since L/σ is small, this is within the same budget. $\square$

We first collect the accuracy and depth estimates. We will then expand the actual formulas to obtain the coefficient and scale bounds required by the recursive implementation.

Corollary 6.6 (Accuracy without increasing dependency depth). Fix $J \geq 1$. In all the probability, harmonic, and centering integrations above, one may take at most

$$N = \lceil 10(J + 2) \rceil$$

Picard layers. After any fixed $w \in (0, t)$ is chosen, a common fixed interpolation order satisfying

$$w(m - 3) > J + 10, \quad t(m - 3) > J + 10, \quad m > 4(J + 10) \quad (6.16)$$

suffices for the following conclusions, for all sufficiently large d . Probability transport from its exact starting marginal has endpoint RMS error at most $\text{polylog}(d) \delta^J D$, and the same holds after division by an endpoint parameter bounded away from zero. The mean circuit with ideal anchors has RMS error at most $\text{polylog}(d) \sigma \delta^J D$ relative to (5.4).

The number of layers of exact mean dependencies within either circuit is $O((N + 1)^2)$, independently of m , w , and the number of mesh cells. All mesh nodes, interpolation weights, scalar parameters, and iteration counts are deterministic functions of the prescribed parameters and dimension.

Proof. For large d , probability and harmonic Picard contractions $C_m l \leq \text{polylog}(d) \delta^b$ can be bounded by $\delta^{1/3}$. The only additional inverse powers in kernel evaluation are δ^{-4t-w} , besides the divisions already canceled in Proposition 6.4. The centering iteration also contracts by $\delta^{1/3}$. The probability iteration remainder is at most $C \delta^{N/3} D$, and after the kernel weights the iteration remainder is at most $C \delta^{N/3-4t-w} D$. Both exponents exceed J with a fixed positive margin. For the truncation terms it suffices to check the exponents

$$wm, \quad w(m - 2), \quad w(m - 1) - 4t, \quad 2tm - 4t - w, \quad (m + 1)/3. \quad (6.17)$$

The first four come respectively from probability quadrature, the zero-endpoint rule, its amplified probability-value error, and the harmonic and angle errors. For the last one use $(L/s) R^{-2} \leq \delta^{1/3}$ for large d in both outer quadratures, dividing the mean-output error by σ and using $L/\sigma \leq \delta^{1/3}$. Every displayed exponent exceeds J by (6.16). The outer Picard remainder contracts by the same $\delta^{1/3}$ factor. Thus fixed constants and polylogarithms, including those depending on m , fit within the strict exponent margins as $d \rightarrow \infty$.

Within a probability or harmonic Picard layer, all nodes depend on only the preceding layer, regardless of the number of nodes. A path through one kernel evaluation passes through a backward probability computation, a harmonic computation, and a forward probability computation, each having at most N layers. Such a path traverses at most N layers of the outer centering iteration. This gives depth $O((N + 1)^2)$, and final quadratures add a bounded number of layers. All weights integrate known polynomials, or polynomials against sine or $1/\rho$ with the stated zero-endpoint condition. They require no oracle information and are allowable real constants in the model. $\square$

The exponent ledger. The list (6.17) is the heart of the parameter check, so it is worth reading slowly. Each entry is the power of δ in one error term, and each must exceed the requested accuracy exponent J :

- wm is the probability-quadrature error per iteration,
- $w(m-2)$ is the zero-endpoint quadrature error, weakened by the weight sum,
- $w(m-1) - 4t$ is the same value error after multiplying by the angle stencil weights $\psi^{-1} = \delta^{-4t}$,
- $2tm - 4t - w$ is the harmonic/angle truncation term,
- $(m+1)/3$ is the outer quadrature, using $(L/s)R^{-2} \leq \delta^{1/3}$.

The conditions (6.16) are designed exactly so that all five exponents exceed J , and the extra room 10 absorbs every fixed constant and every power of $\log d$ that appears.

6.7 The exact circuit interface and its derivation

To display a primitive's parameters, for $x' \in \mathbb{R}^d$ and $0 < r' \leq 1$ write

$$u_{x',r'} := \int_{\mathbb{R}^d} g(x' + r'z) \nu_{x',r'}(dz).$$

Thus $u_{x,r} = u$ for the primitive fixed above.

We specify the two formulas to be implemented. In this subsection $\mathcal{C}_M$ and $\mathcal{C}_S$ denote deterministic maps of their listed anchors, with exact primitive means available as analytical nodes. The parameter σ of $\mathcal{C}_M$ is the parent noise parameter, so its construction sets $s = \sigma/2$ exactly once. Define

$$\mathcal{C}_M(x, r, \sigma, y, G) = \frac{\sigma}{2}G + \sum_j \omega_j M_*(y_j), \quad (6.18)$$

$$\mathcal{C}_S(x, r, \eta, y) = \rho_e^{-1} \widehat{T}_{0 \rightarrow \rho_e}(y), \quad \rho_e = (1 + \eta^2/4)^{-1/2}, \quad R \leq \eta \leq 1. \quad (6.19)$$

The positions y_j in the first line are the last centering Picard iterates with anchors (y, G) , using (6.12). The second line uses the probability formula (6.3). No additional Gaussian is included in either map. The next section allocates additional noise and constructs the position anchor needed for $\mathcal{C}_M$.

A *mean dependency path* follows a mean node to a mean value used in its center expression, and then repeats. Direct expression size counts occurrences in one such expression, before expanding the center expressions of its mean nodes.

Theorem 6.7 (Numerical circuit interface). Fix $J \geq 1$, the parameters (6.1), and $N = \lceil 10(J+2) \rceil$. Choose the fixed interpolation order by (6.16). For all sufficiently large d , the maps (6.18)–(6.19) have the following properties.

1. *Deterministic scalar formulas.* Their exact mean nodes can be ordered as $M_1, \dots, M_Q$, where

$$M_i = u_{E_i, r_i}, \quad E_i = b_i + \sum_{j < i} a_{ij} M_j, \quad E_f = b_f + \sum_j a_{fj} M_j. \quad (6.20)$$

All a_{ij} and a_{fj} are deterministic scalars. The whole node list, its scales, and these coefficients can be formed without knowing a mean value or making an oracle query.

2. *Direct terms and scales.* Each node has $r_i = r\tau_i$ and

$$\tau_i = \sqrt{1 - \rho_i^2} \geq R/\sqrt{5}, \quad b_i = x + r(\ell_i y + m_i G). \quad (6.21)$$

For $\mathcal{C}_S$, the G term is absent, $\rho_i \leq \rho_e$, and $(\ell_i, m_i) = (\rho_i, 0)$. For $\mathcal{C}_M$, $\rho_i \leq b_*$ and the pair is either $(\rho_i, 0)$ or $(\rho_i \cos z, \rho_i \sin z)$, with $|z| \leq C_m \psi$. In particular both coefficients have absolute value at most one. A node for $M_0 = u$ has center x and scale r . Every mean occurring directly in the final expression for $\mathcal{C}_M$ is M_* at a computed position and has scale exactly rR . The final direct terms are $b_f = \sigma G/2$ and $b_f = y/\rho_e$, respectively.

3. *Coefficient sums and direct expression size.* There are positive declared bounds A, A_f such that $\sum_{j < i} |a_{ij}| \leq A$ and $\sum_j |a_{fj}| \leq A_f$, namely

$$\begin{aligned} A_M &= C_1(1 + \log d)^{C_1} [(r/\sigma)\delta^{-4t-c_1w} + r^2], & A_S &= C_1 r^2, \\ A_{f,M} &= C_1, & A_{f,S} &= C_1 r, & c_1 &= 10. \end{aligned} \quad (6.22)$$

Here $C_1 \geq 1$ depends only on the fixed parameters and orders. Every direct expression has at most $B_d \delta^{-c_1 w}$ terms. Every mean dependency path has length at most $D_0 = 3N^2 + 1$, and using a larger absolute multiple of $(N + 1)^2$ is harmless. This depth bound is independent of w , m , and the number of mesh cells.

4. *Errors on the ideal anchor laws.* Under (6.2), for independent $Y \sim \xi_*$ and $G \sim N(0, I_d)$,

$$W_2(\mathcal{L}(\mathcal{C}_M(x, r, \sigma, Y, G)), N(u, \sigma^2 I_d/4)) \leq B_d \sigma \delta^J D(x). \quad (6.23)$$

For $\mathcal{C}_S$ only $l \leq B_d \delta^b$, $b \geq 1/2$, is needed. With $Y \sim N(0, I_d)$,

$$W_2(\mathcal{L}(\mathcal{C}_S(x, r, \eta, Y)), \nu_{x,r} * N(0, \eta^2 I_d/4)) \leq B_d \delta^J D(x). \quad (6.24)$$

The coefficient and dependency assertions are deterministic, for every choice of anchors. The distributional estimates require the specified ideal anchor laws.

Proof. We expand only the additive state formulas, stopping whenever an earlier mean value is reached. This convention is what makes the coefficient bounds local: the center expression inside that earlier mean is kept as its own node.

The three transport formulas. A probability Picard state has its input position as direct term and mean coefficients with absolute sum at most $C_m r$. A harmonic state has direct term $\cos z$ times its input position plus $\sin z$ times its input momentum, with mean coefficients of absolute sum at most $C_m r$. The sine-weighted row sums are bounded because their integration intervals have bounded length. These assertions hold at every iteration: its additive direct term is the original input, not the preceding Picard iterate.

For a kernel evaluation, the scalar expansion has absolute coefficient sum at most $B_d \sigma^{-1} \delta^{-4t-w}$, by the derivative-quadrature, angle-stencil, and forward-integral weight bounds. The momentum formula $(M_* - M_0)/s$ has coefficient sum at most C/σ . A centering Picard update has the original anchor as direct term and bounded outer integral rows. Its position and momentum corrections thus have the same respective coefficient bounds.

Centers and their coefficients. Consider a mean evaluation inside any of these formulas. Its physical center is $x + r\rho v$ when its normalized position is v , by (3.2). Substitute its direct starting states: a backward probability state starts at the current centering position, a harmonic state starts at that backward position and the current momentum, and a forward probability state starts at the harmonic position. Stop each substitution at the mean terms just described. Only the free terms pass through this chain. Thus their anchor coefficients are exactly $(\rho, 0)$ or $(\rho \cos z, \rho \sin z)$. For

sampling there is only the first probability computation, so the coefficient is $(\rho, 0)$. Multiplication by $r\rho$, with $\rho \leq 1$, bounds the center coefficient sum in the mean case by

$$B_d(r/\sigma)\delta^{-4t-w} + C_m r^2.$$

The term $C_m r^2$ accounts for the backward, harmonic, and forward integral corrections. In the sampling case the bound is simply $C_m r^2$. Enlarging these positive upper bounds to (6.22) is convenient for the implementation. The mean output is the single quadrature in (6.18), so its coefficient sum is at most C_m and its means all have probability time b_* . For sampling, dividing a probability correction by $\rho_e \geq 2/\sqrt{5}$ gives output coefficient sum at most $C_m r$. These observations also give the stated final direct terms.

The scale at probability time ρ is $r\sqrt{1-\rho^2}$. In the mean computation $\rho \leq b_*$, giving $\tau \geq R$, including equality for the output means. In the sampling computation,

$$\sqrt{1-\rho_e^2} = \frac{\eta}{\sqrt{4+\eta^2}} \geq \frac{R}{\sqrt{5}}.$$

This proves every assertion about the anchors and scales.

Expression size and depth. Let $P \leq B_d \delta^{-w}$ bound the number of nodes in any one probability mesh. Angle stencils, harmonic quadratures, and outer centering quadratures have fixed numbers of nodes once m is fixed. The scalar expansion of $\hat{v}_Y$ has at most $C_m P^2$ mean terms: one factor P comes from probability-time derivative quadrature, and the other from the forward transport integral. A direct centering position has the same bound, and its momentum has only a fixed number of terms. Substituting their free states into a backward, harmonic, or forward center adds at most one backward list, one harmonic list, and one forward list. No substitution continues through a mean value. Hence $B_d \delta^{-2w}$ already bounds the direct expression size, and the exponent $c_1 w = 10w$ is a uniform conservative choice.

All Picard updates at a fixed layer depend only on earlier layers. Order an internal kernel computation as backward transport, harmonic motion, and forward transport, after its input states. A mean path within this computation adds at most $3N$ nodes above a path in those inputs. There are at most N outer Picard layers, and one final mean layer in (6.18). Thus $3N^2 + 1$ bounds the depth. The sampling path has depth at most N . This gives an acyclic order for (6.20). Meshes, weights, angles and partial integral rows depend only on the prescribed scalar parameters. No exact mean is needed to form any of them.

Ideal-law errors. Couple the numerical mean formula and the exact centering path integral by the same ideal anchors. Corollary 6.6 bounds their RMS difference by $B_d \sigma \delta^J D(x)$, while Proposition 5.3 identifies the exact output as $N(u, \sigma^2 I_d/4)$. This coupling proves (6.23). For sampling, the exact probability transport from a standard Gaussian has law ξ_{ρ_e} . After division by ρ_e this is the law of

$$Z + \frac{\sqrt{1-\rho_e^2}}{\rho_e} G = Z + \frac{\eta}{2} G, \quad Z \sim \nu_{x,r}, \quad G \sim N(0, I_d) \text{ independent.}$$

The same-anchor RMS estimate and $\rho_e \geq 2/\sqrt{5}$ prove (6.24). □

For the implementation we use $J = K_0 + 2$. The theorem fixes N and D_0 before w , so the implementation can choose w using its recursion count and then choose m by (6.16). The formulas above introduce no random failure event. Their errors are RMS estimates under the ideal input laws, and their algebraic structure holds for every input.

What the circuit interface is for. Theorem 6.7 is the bridge between analysis and implementation. Its first three parts are purely algebraic: they say that the two maps are finite affine formulas in exact means, with a known list of nodes, known scales, known coefficient bounds, and known depth, all computable from scalar parameters alone, before any oracle query. Its fourth part says that when the anchors are ideal (a sample from ξ_* for the mean circuit, a standard Gaussian for the sampling circuit), the maps have the advertised distributional error. Section 7 takes it from there: the anchors will not be ideal in the actual recursion, and exact means will not be available. The translation identity of the next section repairs both problems at once.

7 Finite evaluation of nested conditional means

Suppose that a deterministic numerical formula calls for a conditional mean at a center that depends on other conditional means. Even a very accurate unbiased estimate is awkward here: its error changes the next center, and the next call must be analysed at that random center. We will instead add a prescribed Gaussian to every computed center. A translation identity of the lower-level routine allows that Gaussian to be absorbed into its random seeds. The identity restores the law of a call at the intended center, which is fixed after conditioning on the parent formula's random input anchors.

We first prove that identity. We then give the full finite recursive construction, including its termination, its query cap on every execution, and its distributional error. The analytical input is a family of finite affine formulas in exact conditional means. Theorem 6.7 supplies these formulas and their complete numerical construction. We now implement them and thereby prove Theorem 2.1.

7.1 Gaussian center noise and the absorption identity

A *seeded mean routine* at scale $r > 0$ is a measurable deterministic map $Q(x, z_1, \dots, z_m) \in \mathbb{R}^d$, whose input is a center $x \in \mathbb{R}^d$ and m seed slots $z_j \in \mathbb{R}^d$. It may query a fixed oracle. A vector $v = (v_1, \dots, v_m) \in \mathbb{R}^m$ gives a *query-preserving translation* if, for every center, displacement and fixed seed list,

$$Q(x + \Delta, z + v \otimes \Delta/r) = Q(x, z), \quad (7.1)$$

and the two computations ask exactly the same physical oracle queries. Here $(v \otimes a)_j = v_j a$. This is a deterministic identity. It does not require independent or Gaussian seeds.

Lemma 7.1 (Absorbing a Gaussian center displacement). Let $D_s > 0$, and let Q satisfy (7.1), with $\|v\|_2 \leq D_s$. Put

$$n = \frac{r}{2D_s}, \quad P = \left(I_m - \frac{vv^T}{4D_s^2} \right)^{1/2}.$$

For a deterministic $x \in \mathbb{R}^d$, let $Z, G_1, \dots, G_m$ be independent standard Gaussian vectors in $\mathbb{R}^d$. Then the output of $Q(x + nZ, (P \otimes I_d)G)$ has exactly the law of $Q(x, G)$.

Proof. The eigenvalues of P lie in $[\sqrt{3}/2, 1]$. The translation identity rewrites the first output as

$$Q\left(x, PG - \frac{v \otimes Z}{2D_s}\right).$$

We suppress the tensor product with I_d on slot-space matrices. The seed array in this expression is centered Gaussian and has slot covariance

$$PP^T + \frac{vv^T}{4D_s^2} = I_m.$$

Its slots are therefore independent standard Gaussian vectors. They may remain correlated with Z . That is harmless because the rewritten call is at the deterministic center x . $\square$

The covariance reduction P is essential. With unmodified fresh seeds, translating the noisy center back would add variance. With P , the center noise restores precisely the variance removed from the seeds. The construction below maintains the deterministic translation identity at every level, so that this argument applies recursively.

Why this identity is the key to the whole construction. A naive implementation would try to estimate each conditional mean accurately and then feed the estimate into the next center. That plan fails for two reasons: the error of the estimate is hard to control after composition, and the algorithm would not have a deterministic query cap. The construction takes the opposite route: it *deliberately* perturbs the center by a Gaussian nZ , and then cancels the perturbation by modifying the seeds according to (7.1). Because the modified seeds remain exactly standard Gaussian and the computation at the ideal noisy center is exactly the standalone law, errors propagate cleanly and the number of queries is a deterministic function of the parameters.

7.2 The two routines and two kinds of depth

We construct a mean routine $M(x, r, \sigma, b, p)$ targeting $u_{x,r} + \sigma G$ and a sampling routine $S(x, r, \eta, b, p)$ targeting $Z + \eta G$, for independent $Z \sim \nu_{x,r}$ and standard Gaussian G . The top call is $S(x, r, 1, 1, K_0)$.

Throughout, a *reachable* call means a descendant of this call, including the call itself. The first label b records the size conditions

$$\max\{l, L/\sigma\} \leq \text{polylog}(d)\delta^b \quad (\text{M}), \quad l/\eta \leq \text{polylog}(d)\delta^b \quad (\text{S}), \quad (7.2)$$

where $R \leq \eta \leq 1$. The second label p records the requested error power. The intended errors are $\text{polylog}(d)\sigma\delta^p D(x)$ and $\text{polylog}(d)\delta^p D(x)$, respectively. A call is *terminal* precisely when $p \leq b$, and it is *nonterminal* when $p > b$. Every sampling routine reserves a final noise $\eta G/2$, unused in its own queries. Before adding that reserve, its target law is $\nu_{x,r} * N(0, 3\eta^2 I_d/4)$.

For a nonterminal sampling call, use $\mathcal{C}_S$ with a standard Gaussian position anchor, supply additional final noise $n = \eta/\sqrt{2}$, and then add the reserve. For a nonterminal mean call, pass the parent parameter σ to $\mathcal{C}_M$ and supply final noise $n = \sqrt{3}\sigma/2$. Obtain its position anchor from an S call at the same center and scale with $\eta = R/b_*$, multiplying that output by b_* . An ideal such anchor has exactly the required law $b_*Z + RN$. Its momentum anchor is an independent standard Gaussian. For all sufficiently large d , $R/b_* \in [R, 1]$, so the anchor call has an admissible noise parameter.

An expression dependency and a routine call play different roles. For example, suppose a parent circuit contains $E_f = au_{x_2, r_2}$ and $x_2 = x + cu_{x_1, r_1}$. Evaluating E_f invokes two new mean routines: one to help supply x_2 , and one at that supplied center. Both are children of the *same* parent routine. Their own internal circuits begin the next routine level. The two nested affine expressions do not create two successive losses in the parent's accuracy label.

To evaluate an expression, we expand its mean dependencies as an uncached tree, duplicating every mean occurrence together with its entire center-supplying subtree. All mean calls at any expression tier are direct children of the routine whose circuit is being evaluated.

The children of a parent labeled (b, p) receive labels as follows:

child's role	child labels	
term of E_f in an M parent	$(b + t, p)$	(7.3)
term of E_f in an S parent	(b, p)	
any other term, or the M position anchor	$(b - 10t, p - \frac{1}{4})$	

In the last row the labels are always those of the parent routine, not successively decreased while descending its expression tree.

Lemma 7.2 (Bounded routine height). Starting at $(b, p) = (1, K_0)$, the rules (7.3) give a finite routine tree of height at most a constant $H = H(K_0, t)$. Every reachable label satisfies $b > 1/2$.

Proof. Before $4K_0 + 1$ decreasing transitions, the total loss in b is less than $10t(4K_0 + 1) < 1/4$, so $b > 3/4$. At that many transitions, $p \leq K_0 - (4K_0 + 1)/4 < 0$, forcing termination. Therefore no branch can have more such transitions. Positive transitions increase b by t , while a nonterminal satisfies $b < p \leq K_0$, so their total number is bounded in terms of K_0, t and the already bounded total decrease in b . A neutral transition is an S-to-M transition. Another neutral transition requires first reaching another S, which can occur only through a decreasing transition. This bounds the neutral transitions as well and proves both claims. $\square$

A routine is a deterministic function of its center, scalar parameters, oracle, and a finite ordered list of d -dimensional seed slots. Standalone use fills these slots by independent standard Gaussians. In addition, each routine will have a deterministic vector v of scalar coefficients, one per slot, with

$$\|v\|_2 \leq D_s, \quad D_s := \delta^{-t}. \quad (7.4)$$

For *every* fixed seed list, changing x to $x + \Delta$ and adding $v \otimes \Delta/r$ to the seeds will preserve every physical query point. The output of M will be unchanged, while the normalized output of S will decrease by Δ/r . These are pathwise properties, not statements restricted to Gaussian seeds.

Within each call write $u = u_{x,r}$ and $\nu = \nu_{x,r}$. At a terminal call define

$$\begin{aligned} \mathbf{M} &= g(x + rZ) + \sigma G, \\ \mathbf{S} &= Z - rg(x + rZ) + (\sqrt{3}\eta/2)G_1 + (\eta/2)G_2, \end{aligned}$$

with separate seed slots. The last term in S is its reserved noise. Taking $v = -1$ on the Z slot and zero elsewhere proves the shift property and $\|v\|_2 = 1$.

Two kinds of depth, and why they must be separated. There are two different notions of nesting in this construction, and conflating them would break the proof. The *expression depth* measures how many layers of affine expressions separate a mean node from the anchors of the circuit. It grows like $O((N + 1)^2)$ but is fixed before the interpolation order. The *routine height* measures how many times the algorithm calls itself in the recursive evaluation, and it is bounded by the constant H of Lemma 7.2. Both are bounded independently of the mesh size, the dimension, and the accuracy exponent. That is what makes the query cap possible.

7.3 Evaluating an expression at a noisy center

Fix a parent circuit and its anchors, and write $M_i = u_{x_i, r_i}$ for its exact mean nodes. For each expression $E = b_0 + \sum_i a_i M_i$, the operation $\text{Eval}(E, n)$ will approximate its exact value plus an independent $N(0, n^2 I_d)$ noise. Assign disjoint seed blocks to every term occurrence, its center-supplying computation, and each residual noise, and keep all these blocks separate from the anchors. Thus the term blocks are independent conditional on the anchors in a standalone call.

Here is the recursive evaluation rule, assuming the child routines and their shift vectors have already been constructed. Take $A' = A$ for a center expression and $A' = A_f$ at the output, and put $\sigma_i = n/(2A')$. For the term of scale $r_i = r\tau_i$, first evaluate its center expression with noise

$$n_i = \frac{r_i}{2D_s}. \quad (7.5)$$

At this supplied center run the child $\mathbf{M}$ with scale r_i , noise σ_i , and labels from (7.3). If its shift vector is v_i , apply to its fresh seed block G_i the deterministic slot-space transformation

$$P_i = \left(I - \frac{v_i v_i^T}{4D_s^2} \right)^{1/2}, \quad \text{and use seeds } (P_i \otimes I_d)G_i. \quad (7.6)$$

Denote its output by Q_i . Return

$$b_0 + \sum_i a_i Q_i + \left(n^2 - \sum_i a_i^2 \sigma_i^2 \right)^{1/2} G_E, \quad (7.7)$$

where G_E is a separate slot. The square root is defined, since

$$\sum_i a_i^2 \sigma_i^2 \leq \left(\sum_i |a_i| \right)^2 \frac{n^2}{4(A')^2} \leq n^2/4.$$

If there are no terms, return $b_0 + nG_E$. The mean-node ordering makes the expression recursion finite. At the parent's output use the value of n specified above, and add the reserved noise for $\mathbf{S}$. A routine supplied with correlated seeds performs these same deterministic operations, without any change of rule.

At an ideal supplied center $x_i + n_i Z$, Lemma 7.1 now identifies this child with its standalone law at x_i . We still owe the construction of the shift vectors and a proof that the noisy supplied centers are accurate. We prove these in that order.

7.4 Query-preserving shifts and the deterministic cap

We next show that the shift vectors required in Lemma 7.1 exist with the universal bound (7.4). This also verifies that the definition is noncircular and has a deterministic query cap.

For a nonterminal $\mathbf{S}$, shift its Gaussian position anchor by $-\rho_e \Delta/r$. For a nonterminal $\mathbf{M}$, we require its position anchor y_* to shift by $-\Delta/(b_* r)$ and leave its momentum anchor unchanged. To obtain this, apply the shift of its anchor $\mathbf{S}$ call and additionally shift that call's reserved seed by the coefficient $-2R/b_*$ times Δ/r . The ordinary shift changes its output by $-\Delta/r$. Since the reserved standard deviation is $R/(2b_*)$, the additional change is $-R^2 \Delta/(b_*^2 r)$. Multiplication by b_* makes the total change

$$-\left(b_* + \frac{R^2}{b_*} \right) \frac{\Delta}{r} = -\frac{\Delta}{b_* r}.$$

The extra slot is used in no query, so query invariance is preserved.

For a term of scale $r\tau_i$, the direct part of its supplied center then changes by $k_i\Delta$, where

$$k_i = 1 - \ell_i \rho_e \quad (\text{S}), \quad k_i = 1 - \ell_i/b_* \quad (\text{M}).$$

These coefficients satisfy $|k_i|/\tau_i \leq C$. For S, $\rho_i \leq \rho_e$ gives $1 - \rho_i \rho_e \leq 2(1 - \rho_i) \leq 2\tau_i$. For M, its two possible coefficients obey

$$0 \leq 1 - \rho_i \cos z/b_* \leq C(1 - \rho_i) + Cz^2,$$

with $z = 0$ when there is no cosine. Use $1 - \rho_i \leq \tau_i$, $\tau_i \geq cR$, and $z^2/R \leq C\delta^{7t} \rightarrow 0$.

Shift that child's untransformed seed block by

$$\frac{k_i}{\tau_i} P_i^{-1} v_i \otimes \frac{\Delta}{r}. \quad (7.8)$$

If (7.4) holds for the child, the smallest eigenvalue of P_i is at least $\sqrt{3}/2$, so $\|P_i^{-1}\|_{\text{op}} \leq 2$. Formula (7.8) induces exactly the child's shift for center displacement $k_i\Delta$. Leave all residual noise slots fixed. Induction from leaf expressions upward shows that every mean output is unchanged, and therefore that the actual supplied-center change is precisely its direct-term change. All queries are preserved. The parent's final direct term is unchanged for M and changes by $-\Delta/r$ for S, giving the required output behaviour.

The center-supplying subtree and the invoked child have disjoint slots. In particular, the factor k_i/τ_i in (7.8) multiplies only the shift of that child, not the shifts of routines used to supply its center. Those are other direct children of the same parent and are charged once in its total shift vector.

Lemma 7.3 (Well-defined construction and query cap). One can choose $w > 0$, depending only on K_0 , t and the fixed Picard counts, and then choose the interpolation orders, so that the entire construction is well defined for all sufficiently large d . Every routine reached from a top call satisfies (7.4). Its total number of seed slots and oracle queries is at most $\text{polylog}(d)\delta^{-t}$ on every execution.

Proof. Let D_0 be the dependency-depth bound in Theorem 6.7, and let F_0 be its direct fan-out bound. Expanding a circuit into an expression tree costs at most $(D_0 + 1)(1 + F_0)^{D_0}$ nodes. Its size is thus $\text{polylog}(d)\delta^{-C_2 w}$, where C_2 depends on the Picard counts but not on w or the interpolation orders. This bounds its direct child invocations and its own anchor and residual slots. Expanding through the height H from Lemma 7.2 gives $\text{polylog}(d)\delta^{-C_3 w}$ slots and terminal queries. Interpolation orders change only fixed multiplicative constants in these bounds, not C_2 or C_3 .

The terminal shift has norm 1. Assuming the shift bound for the children, the bounded ratios and inverse norms above show that a parent's shift norm is at most a fixed constant times one plus the sum of its direct children's shift norms, including the anchor call. The extra reserved-slot shift has bounded coefficient. The triangle inequality is sufficient even if ordinary and extra anchor shifts affect the same slot. Iterating through the fixed routine height gives the same bound $\text{polylog}(d)\delta^{-C_3 w}$, after increasing C_3 if necessary.

Choose $w > 0$ with

$$C_3 w < t/2, \quad w < t, \quad c_1 w < t,$$

and only then choose the fixed interpolation orders needed in Theorem 6.7. For sufficiently large d , the last shift bound is smaller than $D_s = \delta^{-t}$. This is a valid bottom-up induction: it proves each child's bound before using its inverse transformation in the parent. The count bound is also at most $\text{polylog}(d)\delta^{-t}$.

To make the parameter order explicit: first fix the Picard counts, hence D_0, H, C_3 , then w , then the interpolation orders. The universal scalar D_s fixes every center-noise budget. The circuit shapes,

labels, requested noises, and slot counts are determined by these scalar parameters before any shifts are computed. Compute child shifts before their matrices and the parent shift. No center value, oracle response, numerical-error estimate, or assumption (7.2) enters this finite construction or the count. Only terminal routines query g , once each, so the bound is a cap on every execution, not an expected query count. $\square$

7.5 Terminal errors and propagation of scales

The finite construction and its cap are now established independently of accuracy. We next verify its terminal errors and the size conditions needed by the numerical estimates.

Lemma 7.4 (Terminal accuracy). In standalone use, a terminal call satisfying (7.2) obeys

$$W_2(\mathcal{L}(\mathbf{M}), N(u, \sigma^2 I_d)) \leq \text{polylog}(d) \sigma \delta^p D(x),$$

$$W_2(\mathcal{L}(\mathbf{S}_{\text{pre}}), \nu_{x,r} * N(0, 3\eta^2 I_d/4)) \leq \text{polylog}(d) \delta^p D(x),$$

where $\mathbf{S}_{\text{pre}}$ is the output before adding its reserved noise $\eta G_2/2$. The sampling bound also holds after adding that same independent reserve to both laws. The fixed-seed center Lipschitz constant is at most λ for $\mathbf{M}$ and $r\lambda$ for $\mathbf{S}$.

Proof. For the standard Gaussian seed Z , the primitive moment bounds give

$$\|g(x + rZ) - u\|_{L^2} \leq CL(\sqrt{d} + r|g(x)|) \leq CLD(x).$$

Since $L/\sigma \leq \text{polylog}(d)\delta^b$ and $p \leq b$, this is within the mean error budget. Couple the added Gaussian identically.

Let z_0 be the mode of $\nu_{x,r}$, so $z_0 = -rg(x + rz_0)$. Then

$$\|Z - rg(x + rZ) - (Z + z_0)\|_{L^2} \leq l\|Z - z_0\|_{L^2} \leq ClD(x).$$

Also $W_2(N(z_0, I_d), \nu_{x,r}) \leq Cl\sqrt{d}$. Apply the synchronous comparison in Lemma 2.2 with Gaussian potential $H_0(z) = |z - z_0|^2/2$ (contraction constant 1) and tilted potential $H(z) = |z|^2/2 + F(x + rz)$. Their gradient difference is $r[g(x + rz) - g(x + rz_0)]$, whose $L^2(\nu_{x,r})$ norm is at most $l\|Z - z_0\|_{L^2(\nu_{x,r})} \leq Cl\sqrt{d}$. This proves the displayed Wasserstein bound under exactly the C^2 and bounded-Hessian hypotheses of that lemma. Adding the same independent comparison noises proves the sampling error $ClD(x) \leq \text{polylog}(d)\eta\delta^b D(x)$, which is at most the requested bound because $\eta \leq 1$ and $p \leq b$. Omitting G_2 proves the pre-reserved assertion. The Lipschitz bounds follow directly from the two displayed terminal formulas. $\square$

Lemma 7.5 (Size conditions descend through the routine tree). Starting with a top call satisfying (7.2), every descendant satisfies that condition with its assigned labels, up to a uniform polylogarithmic factor. At each nonterminal circuit,

$$\lambda A \leq \text{polylog}(d) \delta^{1/3}. \quad (7.9)$$

For nonfinal expression terms within an $\mathbf{M}$ parent, $\sigma_i \leq C\sigma$, while within an $\mathbf{S}$ parent, $r\sigma_i \leq C$.

Proof. For a final term of an $\mathbf{M}$ parent, the output uses only means of scale rR , so $r_i = rR$, while σ_i is a fixed positive multiple of σ . Thus L_i/σ_i gains a factor R and l_i gains R^2 , allowing labels $(b+t, p)$. For a final term of an $\mathbf{S}$ parent, σ_i is a fixed positive multiple of η/r . Therefore $L_i/\sigma_i \leq Cl/\eta$ and $l_i \leq l$, allowing the unchanged labels.

For a nonfinal term, let $r\tau_{\text{prev}}$ be the scale of the node whose center is currently being supplied. Its expression is evaluated with budget $r\tau_{\text{prev}}/(2D_s)$, irrespective of the tier at which it occurs. Consequently

$$\sigma_i = \frac{r\tau_{\text{prev}}}{4D_s A}, \quad \frac{L_i}{\sigma_i} \leq \frac{CD_s \lambda A}{R}.$$

The coefficient estimates give $\lambda A \leq \text{polylog}(d)\delta^{b-4t-c_1w}$ for an M parent. For an S parent, $\lambda A = C_1 l \leq \text{polylog}(d)\delta^b$, since $\eta \leq 1$. Using $D_s/R = \delta^{-2t}$ and $c_1w < t$ costs less than the allowed $10t$ decrease in b . Also $l_i \leq l$. For the position-anchor S call of M, $l/\eta \leq l/R$, which has the same allowance. This proves the size conditions by the bounded routine height. All accumulated fixed and polylogarithmic factors are still one polylogarithmic factor.

As $b > 1/2$ and $4t + c_1w < 5t < 1/6$, the coefficient bounds imply (7.9). Finally, the positive declared bound for A includes a fixed multiple of r/σ in an M parent, so $\sigma_i \leq r/(4D_s A) \leq C\sigma$. For S, $A = C_1 r^2$ gives $r\sigma_i \leq 1/(4C_1 D_s) \leq C$. $\square$

7.6 Stability for arbitrary fixed seeds

Lemma 7.6 (Fixed-seed stability). For every seed array, including correlated or transformed arrays, the compiled M output is a function of x with Lipschitz constant at most $\text{polylog}(d)\lambda$, and the normalized S output has constant at most $\text{polylog}(d)r\lambda$. If the anchors of a nonterminal M are treated as separate inputs, its Lipschitz constant in the position anchor is at most $\text{polylog}(d)L$.

Proof. Induct on routine height, with the terminal case supplied by Lemma 7.4. First hold the anchors fixed. A direct center term has Lipschitz constant 1 in x and at most Cr in a position anchor. The center-supplying expression combines child outputs with total absolute coefficient at most A . Using the child M bound, the contribution of earlier supplied-center sensitivities is at most $\text{polylog}(d)\lambda A$ times their maximum. By (7.9), this is small. Induction through the bounded expression depth therefore bounds center sensitivities by a constant times their direct sensitivities, with an allowed polylogarithmic factor.

At the final expression the coefficient sum is A_f . Its direct term has no center dependence when anchors are held fixed, and for M it also has no position-anchor dependence. Hence its center sensitivity is at most $\text{polylog}(d)A_f\lambda$, and its M position-anchor sensitivity at most $\text{polylog}(d)A_f\lambda r$. These are the asserted bounds with anchors treated as separate inputs. The S Gaussian anchor is indeed center independent. The M position anchor is supplied by S at the same center, whose center Lipschitz constant by induction is $\text{polylog}(d)r\lambda$. Its additional contribution is at most $\text{polylog}(d)Lr\lambda = \text{polylog}(d)L^2$, and $L^2 \leq \lambda$ for $r \leq 1$ and $\lambda \leq 1$. All seed transformations depend only on scalar parameters, not on x . The proof uses only fixed-seed identities and Lipschitz bounds, and therefore requires no seed independence. $\square$

7.7 Standalone accuracy and the auxiliary theorem

We prove the law guarantees by induction on routine height. This is the only stage at which a routine's seeds are assumed independent standard Gaussians. Such an assumption will be used for a child only after the absorption identity has restored its full standalone seed law.

Proposition 7.7 (Accuracy of the compiled routines). For every call reached from the top labels, the standalone routines satisfy

$$W_2(\mathcal{L}(\text{M}), N(u, \sigma^2 I_d)) \leq \text{polylog}(d)\sigma\delta^p D(x),$$

$$W_2(\mathcal{L}(\text{S}_{\text{pre}}), \nu_{x,r} * N(0, 3\eta^2 I_d/4)) \leq \text{polylog}(d)\delta^p D(x),$$

where S_{pre} denotes the output before its reserved noise. The latter bound also holds after adding the reserved noise to both laws, and

$$\|\mathcal{L}(S) - \nu_{x,r} * N(0, \eta^2 I_d)\|_{\text{TV}} \leq \text{polylog}(d) \frac{\delta^p D(x)}{\eta}. \quad (7.10)$$

Proof. The terminal assertion is Lemma 7.4. All random variables below have finite second moments: the actual algorithms are finite compositions of affine maps and globally Lipschitz oracle evaluations, while ideal tilted variables have the primitive moment bounds. Thus all W_2 comparisons are defined.

For a nonterminal M , first couple its approximate S position anchor to an ideal anchor, independently of all remaining seed blocks. The child labels are $(b - 10t, p - 1/4)$. By induction and Lemma 7.6, replacing this anchor costs at most

$$\text{polylog}(d) L \delta^{p-1/4} D(x).$$

After division by the mean error unit $\sigma \delta^p D(x)$, the cost is at most $\text{polylog}(d) \delta^{b-1/4}$, hence is bounded for large d . We may now assume that the position anchor has law $b_* Z + RN$, where $Z \sim \nu_{x,r}$ and N is standard Gaussian, and that the momentum anchor is an independent standard Gaussian. These are exactly the input laws in Theorem 6.7.

Let x_i be the centers of the deterministic exact-mean circuit under these anchors, and put $D_i = \sqrt{d} + |g(x_i)|$. Their RMS values over anchors obey

$$\max_i \|D_i\|_{L^2} \leq \text{polylog}(d) D(x). \quad (7.11)$$

Indeed, both position and momentum anchors have RMS at most $CD(x)$. Use $|g(x_i)| \leq |g(x)| + \lambda|x_i - x|$ and the circuit's direct terms. An exact primitive mean at a preceding center has magnitude at most CD_j . Their contribution to this maximum-of-RMS recurrence is at most $C\lambda A$ times its previous maximum. This is small by (7.9), while the direct terms cost at most $CD(x)$. Induction through the node ordering proves (7.11) without a maximum inside the expectation.

Condition on the anchors. For an expression E , write ϵ_E for the W_2 error of $\text{Eval}(E, n)$ relative to its deterministic exact value plus $N(0, n^2 I_d)$. Couple the supplied center of its i th term to $x_i + n_i Z$ with its conditional W_2 error ϵ_{E_i} . Use the same transformed child seeds on these two centers, drawn independently of this coupling. The fixed-seed Lipschitz bound costs at most $\text{polylog}(d) \lambda \epsilon_{E_i}$. At the ideal noisy center, Lemma 7.1 identifies the child's law with its standalone law at x_i . The induction hypothesis therefore supplies the additional error $\text{polylog}(d) \sigma_i \delta^{p_i} D_i$. Different term occurrences have disjoint blocks, so their outputs are independent conditional on the anchors. Product couplings compare them to independent noised exact means. Together with the independent residual in (7.7), the ideal sum has exactly covariance $n^2 I_d$. We conclude that

$$\epsilon_E \leq \sum_i |a_i| [\text{polylog}(d) \lambda \epsilon_{E_i} + \text{polylog}(d) \sigma_i \delta^{p_i} D_i]. \quad (7.12)$$

Notice the order of this argument: inaccurate supplied centers use only pathwise stability, and standalone accuracy is invoked only after Gaussian absorption. No assertion is made that the actual algorithm's intermediate states have the ideal input laws.

Take RMS in the anchors and expand (7.12) through the finite expression tree. At its root, direct-child labels have $p_i = p$ and

$$\sum_i |a_i| \sigma_i \leq A_f \frac{n}{2A_f} = n/2.$$

Thus their intrinsic errors fit the requested budget. At tier $j \geq 2$, absolute coefficient sums and Lipschitz factors contribute at most

$$A_f(\text{polylog}(d)\lambda A)^{j-1}.$$

All such children have the single accuracy-label loss $p_i = p - 1/4$, not one loss per expression tier. Their noise units are at most $C\sigma$ in an **M** parent and C/r in an **S** parent by Lemma 7.5. These units combine with $A_f = C_1$ or C_1r , respectively. Already the first factor $\text{polylog}(d)\lambda A \leq \text{polylog}(d)\delta^{1/3}$ pays for the loss $\delta^{-1/4}$, leaving a positive power $\delta^{1/12}$. The bounded expression depth allows all further tiers and polylogarithmic factors to be summed. Leaf affine expressions have zero error. This proves the required final evaluation error.

Conditionally on the anchors, the ideal final evaluation is its deterministic exact-mean circuit value plus a Gaussian of the specified fixed covariance. Its noise may therefore be taken independent of the anchors. By Lemma 7.5, Theorem 6.7 applies. It compares the ideal mean circuit to $N(u, \sigma^2 I_d/4)$ and the ideal sampling circuit to $\nu_{x,r} * N(0, \eta^2 I_d/4)$, with the respective errors $\text{polylog}(d)\sigma\delta^J D(x)$ and $\text{polylog}(d)\delta^J D(x)$. Since $J = K_0 + 2$ and $p \leq K_0$, both errors fit the requested bounds. Couple the independent evaluation noise identically in each comparison. For **M**, its variance $3\sigma^2/4$ raises the Gaussian target variance from $\sigma^2/4$ to σ^2 . For **S**, its variance $\eta^2/2$ raises the Gaussian convolution variance from $\eta^2/4$ to $3\eta^2/4$. This closes the standalone induction. Adding the common reserved noise preserves the W_2 bound, while Gaussian smoothing by its standard deviation $\eta/2$ gives the stated total variation bound by (2.8). $\square$

Proof of Theorem 2.1. Use the top **S** call with $(b, p, \eta) = (1, K_0, 1)$. Because $l = \delta r^2 \leq \delta$, its size condition holds with constant 1. Proposition 7.7 gives both W_2 and total variation error at most $\text{polylog}(d)\delta^{K_0} D(x)$ from $\nu_{x,r} * N(0, I_d)$, uniformly in x and $0 < r \leq 1$. Lemma 7.3 gives at most $\text{polylog}(d)\delta^{-t}$ evaluations of g on every execution. The output has the form $X_{\text{pre}} + G/2$, where the reserved standard Gaussian G is independent of X_{pre} and all queries, and the pre-reserved conclusion of Proposition 7.7 gives $W_2(\mathcal{L}(X_{\text{pre}}), \nu_{x,r} * N(0, 3I_d/4)) \leq \text{polylog}(d)\delta^{K_0} D(x)$. The exact means and numerical comparison laws serve only to analyse this finite algorithm. Its only oracle operations are the terminal evaluations. $\square$

8 From noisy conditional samples to the target law

The auxiliary algorithm returns a conditional sample with an added Gaussian. We now use it to sample the original, unnoised law π_V . The construction has two stages. A contracting proximal chain first approaches $\tilde{\pi} = \pi_V * N(0, hI_d)$. Conditionally on its output, a sequence of variance-halving steps reduces the remaining sampling problem to a law close to a mode-centred Gaussian. The numbers of calls in both stages are prescribed in advance.

Throughout this section, $V \in C^2(\mathbb{R}^d)$ satisfies $V(0) = 0$, $\nabla V(0) = 0$, and $I_d \preceq \nabla^2 V \preceq 2I_d$ everywhere, so the origin is the supplied minimizer. Each original query returns both V and ∇V at the queried point. The main construction uses only the gradient, and the bounded-dimensional fallback also uses the value. We fix $0 < e < 1$, put $\delta = d^{-e}$ and $h = \delta/2$, and take an integer $K_0 > 2 + 1/e$. Choose $t > 0$ as in (2.3), and fix the approximation orders required by Theorem 2.1. All sufficiently-large-dimension thresholds below are uniform in V .

8.1 The contracting noised chain

The Gaussian proximal sampler alternates a Gaussian step with a restricted Gaussian conditional draw [19, Definition 1 and Lemmas 2 and 4]. See also [11, Section 3 and Theorem 1]. We use the

chain on the noised variable and prove its contraction directly.

For $y \in \mathbb{R}^d$, denote by κ_y the probability law with density proportional to

$$\exp\left(-V(x') - \frac{|x' - y|^2}{2h}\right). \quad (8.1)$$

The ideal transition P_h draws $x' \sim \kappa_y$ and returns $x' + \sqrt{h}N$, where N is an independent standard Gaussian. Thus $P_h(y, \cdot) = \kappa_y * N(0, hI_d)$.

Lemma 8.1 (Proximal contraction). The law $\tilde{\pi}$ is invariant for P_h , and

$$W_2(\kappa_{y_1}, \kappa_{y_2}) \leq \frac{|y_1 - y_2|}{1 + h}.$$

Consequently, for input laws μ, η of finite second moment,

$$W_2(\mu P_h, \eta P_h) \leq \frac{W_2(\mu, \eta)}{1 + h}.$$

Proof. Under the joint law $\pi_V(dx)N(x, hI_d)(dy)$, the conditional law of x given y is κ_y . Drawing this conditional and then redrawing y given x preserves the joint law, whose y -marginal is $\tilde{\pi}$.

The potential in (8.1) has curvature at least $1 + 1/h$. Replacing y_1 by y_2 changes its gradient by the constant vector $(y_1 - y_2)/h$. The synchronous comparison (2.5) therefore bounds the conditional W_2 distance by $|y_1 - y_2|/(1 + h)$. Couple the subsequent Gaussian increments identically. Integrating these conditional couplings against an input coupling, as in Lemma 2.3, proves the kernel contraction. $\square$

To implement this transition, substitute $x' = y + \sqrt{h}z$ in (8.1). The normalized conditional potential becomes $|z|^2/2 + F_y(z)$, where

$$F_y(z) = V(y + \sqrt{h}z), \quad g_y(z) = \nabla F_y(z) = \sqrt{h} \nabla V(y + \sqrt{h}z). \quad (8.2)$$

The Hessian bound and $\nabla V(0) = 0$ give

$$\text{Lip}(g_y) \leq 2h = \delta, \quad |g_y(0)| \leq 2\sqrt{h}|y|.$$

An evaluation of g_y is one original first-order query. Apply Theorem 2.1 to F_y at center zero and scale one, with fresh standalone seeds, and map its output S to $y + \sqrt{h}S$. This defines the actual transition $\hat{P}_h$. The primitive already approximates $Z + N$, so the ideal comparison law of this affine image is exactly $P_h(y, \cdot)$. In particular, the primitive's independent final reserve $G/2$ is retained. Its pre-reserve W_2 guarantee and Gaussian smoothing are what supply the following total variation guarantee:

$$\begin{aligned} W_2(\hat{P}_h(y, \cdot), P_h(y, \cdot)) &\leq \sqrt{h} B_d \delta^{K_0} (\sqrt{d} + 2\sqrt{h}|y|), \\ \|\hat{P}_h(y, \cdot) - P_h(y, \cdot)\|_{\text{TV}} &\leq B_d \delta^{K_0} (\sqrt{d} + 2\sqrt{h}|y|). \end{aligned} \quad (8.3)$$

Here and below $B_d = C(1 + \log d)^C$ may be increased by a fixed factor or power. Its constants depend only on the fixed parameters, not on the center, oracle, or random history.

Why a proximal chain first. The primitive of Section 7 samples $\nu_{x,r}$ convolved with a Gaussian. Applied to F_y at centre zero and scale one, it returns a sample of κ_y convolved with $N(0, hI_d)$. Adding the h -Gaussian in (8.1) to the chain makes the target the smoothed law $\tilde{\pi}$. The unsmoothed target is recovered by the conditional descent of

Section 8.2, which uses the exact chain structure rather than a crude deconvolution. The chain is useful because its transition is contracting at the explicit rate $1/(1+h)$: the error committed by the primitive at each step is damped geometrically, so $O(h^{-1} \log d)$ steps turn a local error $\sqrt{h} \text{polylog}(d) \delta^{K_0}$ into a total-variation error $o(1)$.

Proposition 8.2 (Mixing of the approximate noised chain). With $K_0 > 2 + 1/e$ and t chosen above, start the $\widehat{P}_h$ -chain at zero and run $\lceil Ch^{-1} \log d \rceil + 1$ steps. For a sufficiently large fixed C , its output has total-variation distance $o(1)$ from $\tilde{\pi}$, uniformly over all admissible V .

Proof. Lemma 2.2, applied at the mode zero, gives $\mathbb{E}_{\pi_V} |X|^2 \leq d$. Hence the root mean square under $\tilde{\pi}$ is at most $C\sqrt{d}$. Let μ_k be the actual law after k steps, and set $a_k = W_2(\mu_k, \tilde{\pi})$. Then

$$\left(\int |y|^2 \mu_k(dy) \right)^{1/2} \leq a_k + C\sqrt{d}.$$

All these moments are finite. Indeed they are finite at $k = 0$, and the local W_2 bound in (8.3), together with the conditional moment bounds for P_h , propagates finite second moments to the next step.

Take the root mean square of the first local error in (8.3). Conditional coupling and Lemma 8.1 give

$$a_{k+1} \leq \left(\frac{1}{1+h} + 2hB_d\delta^{K_0} \right) a_k + C\sqrt{h} B_d\delta^{K_0}\sqrt{d}.$$

For all large d , the coefficient of a_k is at most $1 - ch$ for an absolute $c > 0$. Since $a_0 \leq C\sqrt{d}$, summing this geometric recurrence yields

$$a_k \leq C\sqrt{d} e^{-chk} + Ch^{-1/2} B_d\delta^{K_0}\sqrt{d}. \quad (8.4)$$

The second term is $o(\sqrt{h})$: after division by $\sqrt{h}$, it is bounded by a constant times $B_d\delta^{K_0-1}\sqrt{d}$, which tends to zero because $e(K_0 - 1) > 1/2$. At $k = \lceil Ch^{-1} \log d \rceil$, a sufficiently large fixed C makes the first term $o(\sqrt{h})$ as well.

Use one more transition to obtain total-variation control. The mixture bound in Lemma 2.3 gives

$$\|\mu_k \widehat{P}_h - \mu_k P_h\|_{\text{TV}} \leq B_d\delta^{K_0} [\sqrt{d} + 2\sqrt{h} (a_k + C\sqrt{d})] = o(1).$$

For the remaining comparison, the conditional x' -laws from μ_k and $\tilde{\pi}$ can be coupled with root mean square at most $a_k/(1+h)$. Their common independent $\sqrt{h}$ -Gaussian then gives, by (2.8),

$$\|\mu_k P_h - \tilde{\pi} P_h\|_{\text{TV}} \leq \frac{a_k}{(1+h)\sqrt{h}} = o(1).$$

Invariance of $\tilde{\pi}$ completes the proof. $\square$

8.2 Conditional descent and its ideal histories

An exact draw from κ_y with $y \sim \tilde{\pi}$ has marginal π_V . We next implement this unnoised conditional draw. For its analysis, first replace the incoming approximate y by $y \sim \tilde{\pi}$. Applying the same conditional algorithm to both inputs costs at most the total-variation error of Proposition 8.2, by data processing.

For a related recursive restricted-Gaussian-oracle framework, see [9, Algorithm 3.3 and Section 6.3]. The construction here uses fixed variance halving, mode-centred Gaussian termination, and an every-execution query cap.

Fix y temporarily. Keep F_y and its oracle g_y fixed throughout the conditional algorithm, and only the primitive centre and scale change. At centre x and scale r , the desired normalized law is

$$\nu_{x,r}(dz) \propto \exp(-|z|^2/2 - F_y(x + rz))dz.$$

Suppose first that $Z \sim \nu_{x,r}$ and that N is an independent standard Gaussian. Given $Y = Z + N$, the conditional density of Z is proportional to

$$\exp(-|z - Y/2|^2 - F_y(x + rz)).$$

Substitute $z = Y/2 + z_{\text{new}}/\sqrt{2}$ to obtain the exact identity

$$Z = \frac{Y}{2} + \frac{Z_{\text{new}}}{\sqrt{2}}, \quad Z_{\text{new}} \mid Y \sim \nu_{x+rY/2, r/\sqrt{2}}. \quad (8.5)$$

The actual algorithm starts at $x_0 = 0$, $r_0 = 1$ and uses $J = \lceil C_1 \log d \rceil$ levels. At each $0 \leq j < J$, it calls Theorem 2.1 with fresh seeds at (x_j, r_j) to produce a noised observation Y_j , and sets

$$x_{j+1} = x_j + r_j Y_j / 2, \quad r_{j+1} = r_j / \sqrt{2}.$$

At level J , it produces a base sample Z_J by the Gaussian rule given below and returns

$$y + \sqrt{h}(x_J + r_J Z_J).$$

To identify this output, define the backwards variables $Z_j = Y_j/2 + Z_{j+1}/\sqrt{2}$ for the analysis. The center update gives

$$x_j + r_j Z_j = x_j + r_j Y_j / 2 + (r_j / \sqrt{2}) Z_{j+1} = x_{j+1} + r_{j+1} Z_{j+1}.$$

Since $x_0 = 0$ and $r_0 = 1$, the returned point is therefore $y + \sqrt{h} Z_0$. By (8.5), ideal observations and an exact base conditional sample would make this an exact draw from κ_y .

We must control the errors at random centers. The useful moment bound concerns the history in which all preceding observations have their ideal laws.

Lemma 8.3 (Moments along the ideal descent). If $y \sim \tilde{\pi}$ and every conditioning observation preceding level j is ideal, then

$$b_j := (\mathbb{E}|g_y(x_j)|^2)^{1/2} \leq C\sqrt{d} \quad (0 \leq j \leq J),$$

with an absolute C for all sufficiently large d .

Proof. The initial estimate follows from $|g_y(0)| \leq 2\sqrt{h}|y|$ and the moment bound for $\tilde{\pi}$. Conditional on an ideal history, the next observation has the law of $Z + N$ at center x_j and scale $r_j = 2^{-j/2}$. Equation (2.6) therefore gives

$$(\mathbb{E}[|Y_j|^2 \mid x_j, y])^{1/2} \leq C(\sqrt{d} + r_j |g_y(x_j)|).$$

The center update and $\text{Lip}(g_y) \leq \delta$ imply, by the conditional and then unconditional L^2 triangle inequalities,

$$b_{j+1} \leq (1 + C\delta r_j^2)b_j + C\delta r_j \sqrt{d}.$$

Both $\sum_{j \geq 0} r_j$ and $\sum_{j \geq 0} r_j^2$ are bounded by absolute constants. In particular the products of $1 + C\delta r_j^2$ are bounded by $\exp(C\delta \sum_{j \geq 0} r_j^2)$. Expanding the recurrence proves the stated uniform bound. $\square$

Here is the precise replacement order used to sum the total-variation errors. For $0 \leq j \leq J$, let H_j be the output law when the first j observations are ideal, all later observations use the actual primitive, and the base rule is the actual Gaussian rule. For $0 \leq j < J$, the laws H_j and H_{j+1} have the same ideal prefix of length j . Conditional on that prefix, only the next observation kernel changes, and all subsequent computation is the same probability kernel. Data processing and the mixture total-variation bound therefore give

$$\|H_j - H_{j+1}\|_{\text{TV}} \leq B_d \delta^{K_0} \mathbb{E}_{\text{ideal}}[\sqrt{d} + |g_y(x_j)|].$$

The expectation includes $y \sim \tilde{\pi}$ and the common ideal prefix. Lemma 8.3 now yields

$$\|H_0 - H_J\|_{\text{TV}} \leq C J B_d \delta^{K_0} \sqrt{d} = o(1). \quad (8.6)$$

Thus errors are summed using only ideal-history moments. It remains to compare the actual and exact base laws under that same ideal history.

8.3 The Gaussian base rule

At depth J , write $l = \delta r_J^2$ and start at $z^{(0)} = 0$. Make exactly $m_0 = \lceil C_2 \log d \rceil$ gradient iterations

$$z^{(k+1)} = -r_J g_y(x_J + r_J z^{(k)}), \quad 0 \leq k < m_0. \quad (8.7)$$

The base output is $Z_J = z^{(m_0)} + G$, for a fresh standard Gaussian G . The iteration map has Lipschitz constant at most l . Its unique fixed point is the mode z_0 of ν_{x_J, r_J} , and (2.6) gives

$$|z^{(m_0)} - z_0| \leq l^{m_0} |z_0| \leq l^{m_0} \frac{r_J |g_y(x_J)|}{1 - l}. \quad (8.8)$$

The following comparison explains why a Gaussian centred there completes the construction.

Lemma 8.4 (Mode-centred Gaussian approximation). For any $F \in C^2(\mathbb{R}^d)$ with $\text{Lip}(\nabla F) \leq \lambda$, center x , and scale $0 < r \leq 1$ such that $l = \lambda r^2 \leq 1/2$, let z_0 be the mode of $\nu_{x, r}$. Then

$$\|\nu_{x, r} - N(z_0, I_d)\|_{\text{TV}} \leq C l \sqrt{d}.$$

Proof. Translate by z_0 . Up to an additive constant, the potential of $w = z - z_0$ is $|w|^2/2 + J(w)$, where

$$J(w) = z_0 \cdot w + F(x + r(z_0 + w)) - F(x + r z_0).$$

The mode equation gives $\nabla J(0) = 0$, and $\text{Lip}(\nabla J) \leq l$. In particular $J(0) = 0$ and $|J(w)| \leq l|w|^2/2$. For $0 \leq s \leq 1$, the potential $|w|^2/2 + sJ(w)$ has mode zero and Hessian between $(1 - l)I_d$ and $(1 + l)I_d$. Its law μ_s satisfies, by Lemma 2.2,

$$\text{Var}_{\mu_s} J \leq \frac{1}{1 - l} \mathbb{E}_{\mu_s} |\nabla J|^2 \leq \frac{l^2}{1 - l} \mathbb{E}_{\mu_s} |w|^2 \leq \frac{l^2 d}{(1 - l)^2} \leq 4l^2 d.$$

The same quadratic bounds justify differentiating

$$\Phi(s) = \log \int_{\mathbb{R}^d} e^{-|w|^2/2 - sJ(w)} dw$$

twice under the integral, and give $\Phi''(s) = \text{Var}_{\mu_s} J$. Since $\mu_0 = N(0, I_d)$,

$$\text{KL}(\mu_0 \|\mu_1) = \Phi(1) - \Phi(0) - \Phi'(0) = \int_0^1 (1 - s) \Phi''(s) ds \leq 2l^2 d.$$

Pinsker's inequality, followed by translation back by z_0 , proves the assertion. $\square$

Under an ideal history, the Gaussian shift bound and (8.8) bound the expected total-variation error of the base rule by

$$C\delta 2^{-J}\sqrt{d} + C(\delta 2^{-J})^{m_0} \frac{2^{-J/2}\sqrt{d}}{1 - \delta 2^{-J}}. \quad (8.9)$$

Here the second term uses Lemma 8.3. Choose the fixed C_1 large enough that $\delta 2^{-J}\sqrt{d} \rightarrow 0$. A sufficiently large fixed C_2 makes the second term tend to zero as well. Replacing the base law under the ideal history therefore changes total variation by $o(1)$. By (8.5), the resulting completely ideal conditional algorithm returns κ_y . Averaging over $y \sim \tilde{\pi}$ gives π_V .

Combining this base comparison, (8.6), and the initial replacement of the incoming y proves

$$\|\mathcal{L}(X_{\text{out}}) - \pi_V\|_{\text{TV}} = o(1), \quad (8.10)$$

uniformly over admissible V . The final Gaussian is used to approximate the last conditional law, and the law in (8.10) is the original target without an additional convolution.

8.4 Query counts and all dimensions

Theorem 8.5 (Subpolynomial upper bound). For every fixed $\varepsilon > 0$, there is $C_\varepsilon < \infty$ such that, for every integer $d \geq 2$, a measurable algorithm using the exact value-and-gradient oracle makes at most $C_\varepsilon d^\varepsilon$ queries on every execution and returns a sample within total-variation distance $1/10$ of π_V for every V satisfying (1.1). Thus

$$Q(d) \leq C_\varepsilon d^\varepsilon \quad (d \geq 2).$$

Proof. Given $\varepsilon > 0$, first choose $e \in (0, 1)$ with $2e < \varepsilon$, then an integer $K_0 > 2 + 1/e$, and then $t > 0$ satisfying (2.3). Fix the auxiliary construction's orders in its prescribed order, and finally the constants in the chain, descent depth, and base iteration count. These choices may depend on ε , but they are fixed before observing the oracle or the random seeds.

Every invocation of Theorem 2.1 uses at most $\text{polylog}(d)\delta^{-t}$ evaluations of the gradient oracle g_y . By (8.2), each evaluation is one original query, including calls at random descent centers. The cap is uniform in those centers. There are $O(\delta^{-1} \log d)$ proximal calls, $O(\log d)$ descent calls, and $O(\log d)$ additional gradient queries in the base rule. Summing gives the deterministic cap

$$\text{polylog}(d) \delta^{-(1+t)} = d^{e(1+t)+o(1)}. \quad (8.11)$$

The query counts and recursion shapes of this construction are fixed by the scalar parameters, and they do not depend on oracle magnitudes or approximation events. All maps are measurable by the primitive theorem and the explicit finite operations above. Since $t < 1$, we have $e(1+t) < 2e < \varepsilon$. The fixed polylogarithmic factors fit $C_\varepsilon d^\varepsilon$ for all sufficiently large d . Equation (8.10) gives total-variation error at most $1/10$ above a uniform dimension threshold.

For each of the finitely many smaller dimensions, use independent standard Gaussian proposals. Accept a proposal X with probability

$$\exp(-V(X) + |X|^2/2).$$

Integrating the Hessian bounds from the known minimizer gives $|x|^2/2 \leq V(x) \leq |x|^2$. The acceptance probability is therefore at most one, and its average is at least $\mathbb{E}e^{-|X|^2/2} = 2^{-d/2}$. The proposal density times its acceptance probability is proportional to $e^{-V(x)}$, so an accepted proposal has law π_V .

Cap the number of trials at $\lceil 4 \cdot 2^{d/2} \rceil$. Return the first accepted proposal, or zero if none is accepted. The failure probability is at most $e^{-4} < 1/10$. Conditional on success, the returned proposal has law π_V . The fallback thus has total-variation error at most e^{-4} and makes at most the prescribed number of original first-order queries on every execution, one per trial. Increasing C_ε covers these finitely many caps. This proves the assertion for all $d \geq 2$. $\square$

Where each hypothesis is actually used. It is worth recording that the strong convexity and smoothness enter at three separate places, and nowhere else. They give (i) the moment and Poincaré estimates of Lemma 2.2 and hence the moment bounds (2.6), (ii) the curvature lower bound in the proximal contraction Lemma 8.1, and (iii) the quadratic sandwich $|x|^2/2 \leq V(x) \leq |x|^2$ used by the rejection fallback. Everything else in the construction uses only that $\text{Lip}(\nabla F) \leq \delta$ for the perturbed potential and that the full potential $|z|^2/2 + F(x + rz)$ has Hessian bounded between $(1 - l)I$ and $(1 + l)I$ with $l \leq 1/2$.

9 A logarithmic oracle lower bound

The lower bound already holds for centred Gaussian targets, whose minimizer is known. The obstruction is therefore present in the same oracle model as the upper bound. We adapt the Gaussian block-Krylov strategy of Chewi, de Dios Pont, Li, Lu and Narayanan [13, Section 5.3.3, Lemma 39 and Theorem 45], which already yields logarithmic dependence on dimension at fixed condition number. We give a complete argument for the spectral interval $[1, 2]$ and total-variation accuracy $1/10$ used here, covering arbitrary measurable adaptive algorithms and unrestricted real computation.

Theorem 9.1. There are absolute constants $c > 0$ and d_0 such that $Q(d) \geq c \log d$ for every $d \geq d_0$. This bound holds even when the inputs are restricted to potentials $V(x) = x^T A x / 2$ with $I \preceq A \preceq 2I$. The query cap is imposed on every execution, and the algorithm may use arbitrary private randomness independent of its input.

For these potentials the gradient reply is Ax , and the value $x^T A x / 2$ is already determined by the query and its gradient reply. Thus an exact value-and-gradient query is equivalent to one query to $x \mapsto Ax$. Each potential belongs to $\mathcal{V}_d$, has minimizer 0, and has target $N(0, A^{-1})$.

The proof has three steps. First, deferred revelation of a random rotation represents any output in a Gaussian block-Krylov span, including outputs outside the span of the actual queries and replies. Second, a cosine statistic is small uniformly on this enlarged random span. Finally, the same statistic has a large bias under the target Gaussian, which separates the two joint experiments in total variation. These steps use no result from the sampling upper bound.

9.1 Adaptive queries and unrevealed rotations

We first record a representation of an arbitrary algorithm's output. The auxiliary Gaussian vectors in this representation are independent of one another, but need not be independent of the input matrix.

Lemma 9.2 (Deferred rotation). Let Λ be a fixed real diagonal $d \times d$ matrix, let O be Haar distributed on the orthogonal group, and set $A = O\Lambda O^T$. Consider any measurable randomized algorithm, with private randomness independent of O , making at most q adaptive queries to the

map $x \mapsto Ax$, where $2q + 1 < d$, and producing an output $X \in \mathbb{R}^d$. There is an extension of this experiment carrying independent standard Gaussian vectors $g_1, \dots, g_{q+1} \in \mathbb{R}^d$ such that

$$O^T X \in \text{span}\{\Lambda^j g_i : 1 \leq i \leq q + 1, 0 \leq j \leq q\} \quad \text{almost surely.} \quad (9.1)$$

The joint law of the original variables, including the algorithm's randomness and O , is unchanged. The vectors g_i need not be independent of O or of A .

Proof. Include all the algorithm's randomness, including any randomness used at output, in a private seed independent of O . Fix this seed for the moment. Pad executions with zero queries so that there are exactly q query slots.

We use the following conditional property of Haar measure. For a fixed restriction $T : E \rightarrow W$ of an orthogonal map, the conditional law of its remaining restriction is uniform among isometries from $E^\perp$ to $W^\perp$. For a specified unit vector in either complement, its partner is uniform on the unit sphere of the other complement. After revealing that pair, the map between the smaller complements is again uniform. This follows from invariance under the orthogonal transformations fixing E and W pointwise. Applying this rule successively also permits each specified direction to be a measurable function of the already revealed data.

All conditional laws below are laws of finite-dimensional data after the seed has been fixed. To represent the subspaces and partial isometries measurably, apply Gram–Schmidt to the revealed vectors in their order of appearance, discard zero residuals, and complete a basis by scanning the standard coordinate vectors in order. These operations are Borel on each rank stratum. Thus regular conditional laws and the successive Gaussian draws can be chosen as measurable kernels. No continuity of the algorithm's query maps is required.

Initially $E = W = \{0\}$. After each reply, reveal the restriction $T = O|_E$, where E is the span of the eigenframe coordinates of all queries and replies so far, and $W = OE$ is their span in the original coordinates. Our conditional invariant is that the unrevealed map is Haar on the two complements, given the full latent history: the seed, the transcript, this restriction, and all auxiliary variables constructed so far.

Let x be the next query, which is fixed given that history. Write $x = x_W + x_\perp$ according to $W \oplus W^\perp$. The preimage of x_W is known. If $x_\perp \neq 0$, put

$$w = \frac{x_\perp}{|x_\perp|}, \quad v = O^T w \in E^\perp.$$

Conditionally on the history, v is uniform on the complementary unit sphere. Let h be a standard Gaussian vector in E , and let a have the distribution of the Euclidean norm of a standard Gaussian in $\mathbb{R}^{d-\dim E}$. Draw h and a independently of one another and of the unrevealed map, conditional on the history, and define

$$g_i = h + av.$$

The Gaussian polar decomposition shows that g_i is a standard Gaussian in $\mathbb{R}^d$ conditional on the entire previous history. Moreover, $a > 0$ almost surely, so $v = (g_i - h)/a \in E + \text{span}\{g_i\}$. Reveal $Ov = w$ and extend T to $E + \text{span}\{v\}$. Conditioning on the full vector g_i does not reveal any further part of O : its parallel component h and complementary radius a are independent auxiliary variables, and its normalized complementary component is precisely v . Thus the conditional Haar invariant persists. If $x_\perp = 0$, instead draw a fresh independent standard Gaussian g_i and make no extension at this step.

In either case the query preimage $u = O^T x$ is now known and lies in the old space $E + \text{span}\{g_i\}$. Its reply has preimage Λu . Project Λu onto the enlarged eigenframe space. The image of this

projection is known. If the residual is nonzero, reveal the image under O of its unit direction and extend the isometry once more. This is the other form of the same Haar-revelation rule: the direction in the eigenframe is fixed by the current history, and its image is uniform in the unrevealed complementary sphere. The resulting reply is exactly Ax , and the conditional invariant is preserved.

Every g_i has conditional law $N(0, I_d)$ given all earlier Gaussian vectors and the earlier history. Consequently these vectors are independent. To check the polynomial degrees, let E_i be the revealed eigenframe space after reply i . Induction gives

$$E_i \subseteq \text{span}\{\Lambda^j g_a : 1 \leq a \leq i, 0 \leq j \leq i\}.$$

Indeed, the new query preimage is in $E_{i-1} + \text{span}\{g_i\}$, and adjoining its reply applies Λ once. All old directions remain in the indicated larger span.

Finally treat the algorithm's output as one more query direction, with no reply: repeat the preimage construction using g_{q+1} . This gives $O^T X \in E_q + \text{span}\{g_{q+1}\}$, proving (9.1). At most two dimensions were exposed per oracle query and one for the output, so the assumed dimension inequality ensures that every required complement is available. The construction started with the genuine Haar input and never altered it. Its added variables are obtained by measurable kernels from the original variables and independent auxiliary draws. The construction therefore applies jointly measurably for the fixed-seed algorithms, and integrating over the original independent seed proves the assertion for randomized algorithms. $\square$

What the lemma says, informally. An algorithm that queries Ax only ever sees the image of its queries under A , and each reply reveals one more direction of the rotation O . The lemma says that one can postpone all of these revelations to the end: the algorithm's entire transcript, and even its output, is a polynomial of degree at most q in $q + 1$ Gaussian vectors, in eigenframe coordinates. The vectors are independent, but the algorithm's coefficients may depend on everything else, including the rotation. That uniformity is what makes the next statistic so powerful.

9.2 A statistic invisible on the block span

For the rest of the section define

$$\theta_k = \frac{(k - 1/2)\pi}{d}, \quad \lambda_k = \frac{3 + \cos \theta_k}{2}, \quad \Lambda = \text{diag}(\lambda_1, \dots, \lambda_d), \quad 1 \leq k \leq d. \quad (9.2)$$

All eigenvalues lie strictly between 1 and 2. For an integer $q \geq 0$ put

$$n = 2q + 1, \quad p_k = \cos(n\theta_k), \quad D_p = \text{diag}(p_1, \dots, p_d).$$

The frequency n exceeds twice the polynomial degree in (9.1). This makes the quadratic form $u^T D_p u$ small simultaneously for all vectors in that random span.

Lemma 9.3. Suppose $4q + 1 < 2d$, and let $g_1, \dots, g_{q+1}$ be independent standard Gaussian vectors. Set

$$\mathcal{K}_q = \text{span}\{\Lambda^j g_i : 1 \leq i \leq q + 1, 0 \leq j \leq q\}, \quad \tau_d = \frac{40(q + 1)^2}{\sqrt{d}}.$$

If $\tau_d \leq 1/2$, then, with probability at least 0.99,

$$|u^T D_p u| \leq 2\tau_d |u|^2 \quad \text{for every } u \in \mathcal{K}_q. \quad (9.3)$$

Proof. Let $m = (q + 1)^2$ and form a $d \times m$ matrix B , whose columns are indexed by (i, j) with $1 \leq i \leq q + 1$ and $0 \leq j \leq q$, by

$$B_{k,ij} = d^{-1/2}(g_i)_k c_j \cos(j\theta_k), \quad c_0 = 1, \quad c_j = \sqrt{2} \quad (j \geq 1).$$

The functions $\cos(j\theta)$ are degree- j polynomials in $\cos \theta$ with nonzero leading coefficients. Since $\cos \theta_k = 2\lambda_k - 3$, the range of B is precisely $\mathcal{K}_q$.

Discrete cosine orthogonality gives

$$\mathbb{E} B^T B = I_m, \quad \mathbb{E} B^T D_p B = 0. \quad (9.4)$$

For completeness, the equally spaced half-grid satisfies

$$\sum_{k=1}^d \cos(\ell\theta_k) = \frac{\sin(\ell\pi)}{2\sin(\ell\pi/(2d))} = 0, \quad 1 \leq \ell < 2d, \quad \ell \in \mathbb{Z}. \quad (9.5)$$

Here the finite cosine sum follows by summing a geometric progression, and its denominator is nonzero in the indicated range. Expand products of two cosines for the first identity in (9.4). For the second, expand the product of three cosines with frequencies n, j, j' . Its frequencies are nonzero because $n > j + j'$, and have absolute value at most $4q + 1 < 2d$. Different Gaussian blocks have zero cross means in both cases.

An entry of either Gram matrix is a sum of independent row terms. The deterministic coefficient in each term has absolute value at most $2/d$, and the variance of a product of two standard Gaussian entries is at most 2. Thus every entry has variance at most $8/d$, and

$$\mathbb{E}[\|B^T B - I_m\|_F^2 + \|B^T D_p B\|_F^2] \leq \frac{16m^2}{d}.$$

By Markov's inequality,

$$\mathbb{P}(\|B^T B - I_m\|_F^2 + \|B^T D_p B\|_F^2 > \tau_d^2) \leq \frac{16m^2/d}{1600m^2/d} = 0.01.$$

Since $\|C\|_{\text{op}} \leq \|C\|_F$, outside this event both operator norms are at most $40m/\sqrt{d} = \tau_d$. Whenever these bounds hold, for every $a \in \mathbb{R}^m$,

$$|Ba|^2 \geq (1 - \tau_d)|a|^2 \geq \frac{1}{2}|a|^2, \quad |(Ba)^T D_p (Ba)| \leq \tau_d |a|^2.$$

These inequalities imply (9.3) on the entire range of B . $\square$

The uniformity in this lemma is essential: the output's coefficients in the block span can depend arbitrarily on the whole transcript. No independence between these coefficients, the input rotation, and the Gaussian vectors is required.

9.3 Separation from the target Gaussian

An ideal target has a detectable bias at the same frequency. Write

$$\varrho = 3 - \sqrt{8} \in (0, 1).$$

The geometric-series formula gives the absolutely convergent identity

$$\frac{2}{3 + \cos \theta} = \frac{1}{\sqrt{2}} \left(1 + 2 \sum_{j \geq 1} (-\varrho)^j \cos(j\theta) \right). \quad (9.6)$$

Indeed, the series in parentheses sums to $(1 - \varrho^2)/(1 + 2\varrho \cos \theta + \varrho^2)$, and $1 + \varrho^2 = 6\varrho$ and $1 - \varrho^2 = 4\sqrt{2}\varrho$.

Lemma 9.4. Let $U \sim N(0, \Lambda^{-1})$, with Λ from (9.2), and suppose $n = 2q + 1 < d$. Then

$$\mu_{d,q} := \mathbb{E}[U^T D_p U] = \frac{d}{\sqrt{2}}(-\varrho)^n + R_{d,q}, \quad |R_{d,q}| \leq \frac{\sqrt{2} d \varrho^{2d-n}}{1 - \varrho}, \quad (9.7)$$

$$\text{Var}(U^T D_p U) \leq 2d, \quad \mathbb{E}|U|^2 \leq d, \quad \text{Var}(|U|^2) \leq 2d. \quad (9.8)$$

Proof. Independence of the coordinates gives $\mu_{d,q} = \sum_k p_k / \lambda_k$. Substitute (9.6). The constant term vanishes by (9.5). For $1 \leq j < 2d - n$, use

$$\cos(n\theta_k) \cos(j\theta_k) = \frac{1}{2} \cos((n-j)\theta_k) + \frac{1}{2} \cos((n+j)\theta_k).$$

If $j \neq n$, both frequencies have nonzero absolute value below $2d$, so their sums vanish. If $j = n$, the constant term contributes $d/2$ and the frequency $2n < 2d$ contributes zero. Thus only $j = n$ contributes in this range. Each remaining cosine product has absolute sum at most d , so summing the geometric tail proves (9.7). Finally, a centred normal variable with variance v satisfies $\text{Var}(U_k^2) = 2v^2$. Since $|p_k| \leq 1$ and $\lambda_k \geq 1$, summing these coordinate variances proves (9.8). $\square$

Proof of Theorem 9.1. Take $A = O\Lambda O^T$ with O Haar and Λ as in (9.2). The reduction at the start of the section identifies an exact value-and-gradient algorithm on $V_A(x) = x^T A x / 2$ with an algorithm querying $x \mapsto Ax$, with the same query cap and output law.

Fix any algorithm using at most q queries on every execution, where

$$0 \leq q \leq c_0 \log d, \quad c_0 = \frac{1}{16 \log(1/\varrho)}.$$

All dimension restrictions in Lemmas 9.2, 9.3 and 9.4 hold for sufficiently large d , uniformly over these q . Let $\mathbb{P}_{\text{alg}}$ be the joint law of the Haar rotation and the algorithm's output on this input. Define the event on pairs (O, x)

$$\mathcal{E}_{d,q} = \left\{ (O, x) : |O^T x|^2 \leq 2d, \quad (O^T x)^T D_p (O^T x) > 160(q+1)^2 \sqrt{d} \right\}. \quad (9.9)$$

This event is measurable. It is also determined by (A, x) : if T_n is the polynomial defined by $T_n(\cos \theta) = \cos(n\theta)$, then $OD_p O^T = T_n(2A - 3I)$. We retain O in the notation to use eigenframe coordinates throughout the comparison.

Apply Lemma 9.2 to the algorithm's output and then Lemma 9.3 to its auxiliary Gaussian vectors. On the latter lemma's event, the norm condition in (9.9) forces the quadratic form to be at most $2\tau_d(2d) = 160(q+1)^2 \sqrt{d}$. Hence

$$\mathbb{P}_{\text{alg}}(\mathcal{E}_{d,q}) \leq 0.01. \quad (9.10)$$

For the target joint law $\mathbb{P}_{\text{target}}$, sample O from the same prior and then X from $N(0, O\Lambda^{-1}O^T)$. Conditional on O , the vector $U = O^T X$ has law $N(0, \Lambda^{-1})$. Our choice of c_0 gives

$$\varrho^{2q+1} \geq \varrho d^{-1/8}, \quad \frac{d\varrho^{2q+1}}{(q+1)^2 \sqrt{d}} \geq \frac{\varrho d^{3/8}}{(1 + c_0 \log d)^2} \longrightarrow \infty.$$

The remainder in (9.7) satisfies

$$\frac{|R_{d,q}|}{d\varrho^n / \sqrt{2}} \leq \frac{2\varrho^{2d-2n}}{1 - \varrho} \longrightarrow 0$$

uniformly, since $n \leq 1 + 2c_0 \log d$. Thus it is negligible relative to the leading term. In particular, with $T_{d,q} = 160(q+1)^2\sqrt{d}$, we have $|\mu_{d,q}| > T_{d,q}$ for large d , and

$$\mathbb{P}_{\text{target}}(\mathcal{E}_{d,q}) \leq \frac{2d}{(|\mu_{d,q}| - T_{d,q})^2} = o(1)$$

uniformly, by Lemma 9.4 and Chebyshev's inequality. Thus the probabilities of $\mathcal{E}_{d,q}$ under the two joint experiments differ by at least $0.99 - o(1)$.

To compare these joint laws with the required guarantee on each fixed input, let K_O be the algorithm's output law on $A = O\Lambda O^T$ and let $G_O = N(0, A^{-1})$. For a measurable joint event E , write $E_O = \{x : (O, x) \in E\}$. Both laws use the same Haar prior, so

$$|\mathbb{P}_{\text{alg}}(E) - \mathbb{P}_{\text{target}}(E)| = \left| \int [K_O(E_O) - G_O(E_O)] d\text{Haar}(O) \right| \leq \int \|K_O - G_O\|_{\text{TV}} d\text{Haar}(O).$$

The pointwise sampling guarantee would bound the last integral by $1/10$. The event used to disprove that guarantee is a test of the joint laws, and the algorithm need not know or compute it. This is impossible for (9.9). Therefore no such algorithm with $q \leq c_0 \log d$ can meet the required accuracy for all inputs. Decreasing the constant if necessary proves the theorem. $\square$

Why the lower bound has the right shape. The statistic $u^T D_p u$ is chosen so that its frequency $2q+1$ beats the polynomial degree q of the block span: every vector in the span is a degree- $\leq q$ combination of the Gaussian vectors, and the cosine kernel at frequency $2q+1$ is orthogonal to those, which is exactly what the discrete orthogonality (9.5) reflects. The same frequency is not orthogonal to the target $N(0, \Lambda^{-1})$ because the target has a spectral mass at frequency $2q+1$. This is the geometric identity (9.6). The gap is the factor q^{2q+1} in (9.7), and requiring it to beat the noise $160(q+1)^2\sqrt{d}$ inside the event forces $q = O(\log d)$ but no further. This is why the logarithmic lower bound is the natural one for this family of statistics.

10 Consequences for diffusion-model sampling

The proof was carried out in the language of conditional means, proximal chains and repeated conditioning. This section records that the main objects also have a second reading, as the scores and denoisers of a diffusion model, so that the results can be stated as theorems about diffusion-model sampling in which the score is *computed* from first-order queries rather than *estimated* from samples. Nothing here is used in the proof. It is included because the translation is exact and requires no new estimates. We stay in the oracle model of Section 1. No training data is used, and all error bounds are uniform over the potential class $\mathcal{V}_d$.

Which parameterization is used. Diffusion models are usually written in a time-indexed form (variance preserving or variance exploding). We use the variance-exploding, or denoising, parameterization, in which the law at noise level r is $p_r = \pi_V * N(0, r^2 I_d)$. This is the parameterization that matches the auxiliary problem (2.1) without rescaling the potential, and a statement in one parameterization converts to the other by a deterministic time change.

10.1 Scores of smoothed laws are conditional means

Fix $V \in \mathcal{V}_d$ and let $X_0 \sim \pi_V$. For $r > 0$ put $X_r = X_0 + rN$, with N standard Gaussian, and write $p_r = \mathcal{L}(X_r) = \pi_V * N(0, r^2 I_d)$, with density again denoted by $p_r(x)$.

Lemma 10.1. For every $r > 0$ and every $x \in \mathbb{R}^d$,

$$\nabla \log p_r(x) = -u_{x,r}, \quad \mathbb{E}[X_0 \mid X_r = x] = x - r^2 u_{x,r}, \quad \mathbb{E}_{\nu_{x,r}}[Z] = -r u_{x,r}, \quad (10.1)$$

where $u_{x,r} = \mathbb{E}_{\nu_{x,r}}[\nabla V(x + rZ)]$ is the primitive mean of (2.1).

Proof. Substituting $y = x + rz$ in the posterior of X_0 given $X_r = x$ shows that $Z = (X_0 - x)/r$ has law $\nu_{x,r}$. This is the density identity (3.2) of Section 3. For the score, differentiate p_r under the integral sign. Since $\nabla_x e^{-|x-y|^2/(2r^2)} = (y-x)r^{-2}e^{-|x-y|^2/(2r^2)}$,

$$\nabla p_r(x) = \int \frac{y-x}{r^2} e^{-V(y)} \frac{e^{-|x-y|^2/(2r^2)}}{(2\pi r^2)^{d/2}} \frac{dy}{Z_V}.$$

The quadratic growth of V makes $\int \nabla_y [e^{-V(y)} e^{-|x-y|^2/(2r^2)}] dy = 0$, so the same integral also equals $-\int \nabla V(y) e^{-V(y)} e^{-|x-y|^2/(2r^2)} (2\pi r^2)^{-d/2} dy / Z_V$. Dividing by $p_r(x) > 0$ gives the first identity. In the variable $z = (y-x)/r$ the same integration by parts reads $\int [z_i + r \partial_i V(x + rz)] \nu_{x,r}(dz) = 0$, which is the third identity. It gives the second one, because $\mathbb{E}[X_0 \mid X_r = x] = x + r \mathbb{E}[Z \mid X_r = x]$. $\square$

For a quadratic potential $V(x) = \frac{1}{2}x^T A x$ one has $u_{x,r} = A(I + r^2 A)^{-1}x$ and $\nabla \log p_r(x) = -(A^{-1} + r^2 I)^{-1}x$, the score of $N(0, A^{-1} + r^2 I)$, as a check.

The score is not a separate object. Lemma 10.1 says that the score and the denoiser of the smoothed law are the conditional mean around which the auxiliary problem (2.1) was built. This is why the construction never needs a score network: the quantity the algorithm computes and the quantity a diffusion model needs are the same one.

10.2 Score and denoiser evaluation in subpolynomial query complexity

Corollary 10.2. For every fixed $\theta \in (0, 1/4)$ there are constants and a dimension threshold, depending only on e, K_0, t, θ , such that for every $V \in \mathcal{V}_d$, every $x \in \mathbb{R}^d$ and every $0 < r \leq 1$, some measurable randomized procedure using at most $\text{polylog}(d)\delta^{-t}$ evaluations of ∇V returns $\hat{s} \in \mathbb{R}^d$ with

$$(\mathbb{E}|\hat{s} - \nabla \log p_r(x)|^2)^{1/2} \leq \text{polylog}(d) \delta^\theta r (\sqrt{d} + |\nabla V(x)|), \quad (10.2)$$

and then $x - r^2 \hat{s}$ estimates the denoiser $\mathbb{E}[X_0 \mid X_r = x]$ in RMS with error at most $\text{polylog}(d)\delta^\theta r^3(\sqrt{d} + |\nabla V(x)|)$. The constants depend on θ , while the shape of the procedure and its query cap do not depend on x, r , or the oracle.

Proof. Use the mean routine of Section 7 with labels (b_0, K_0) , where $b_0 = 1 - \theta$, and with noise parameter $\sigma = (1 + \log d)^C L \delta^{-b_0}$, taking C fixed large enough. The size condition (7.2) at the top reads $\max\{l, L/\sigma\} \leq \text{polylog}(d)\delta^{b_0}$, and it holds because $l \leq \delta < \delta^{b_0}$ while $L/\sigma = (1 + \log d)^{-C} \delta^{b_0}$. The label descent (7.3) loses at most $10t(4K_0 + 1) < 1/4$ in b along any routine path, so every reachable label has $b > b_0 - 1/4 > 1/2$. The height bound of Lemma 7.2 and the propagation of the size conditions in Lemma 7.5 use only this, hence apply with the top label (b_0, K_0) , with constants enlarged by a factor depending on θ , and Lemma 7.3 again gives a cap of $\text{polylog}(d)\delta^{-t}$ queries. By

Proposition 7.7, the routine's output M satisfies $W_2(\mathcal{L}(M), N(u_{x,r}, \sigma^2 I_d)) \leq \text{polylog}(d) \sigma \delta^{K_0} D(x)$. Coupling M to its Gaussian target gives

$$(\mathbb{E}|M - u_{x,r}|^2)^{1/2} \leq \sigma \sqrt{d} + \text{polylog}(d) \sigma \delta^{K_0} D(x) \leq \text{polylog}(d) \sigma (\sqrt{d} + D(x)),$$

and $\sigma \leq \text{polylog}(d) \delta^\theta r$. Returning $\hat{s} = -M$, and using $\nabla \log p_r(x) = -u_{x,r}$ together with $D(x) = \sqrt{d} + |\nabla V(x)|$, proves (10.2). The denoiser bound follows by multiplying by r^2 , since $\mathbb{E}[X_0 | X_r = x] = x - r^2 u_{x,r}$. $\square$

Where the two error terms come from. The bound (10.2) has two ingredients: the deliberate Gaussian σG of the mean routine, which contributes $\sigma \sqrt{d}$, and the numerical error of the routine, which contributes the much smaller second term. The exponent $1/4$ is the margin available in the label condition $b > 1/2$ after the descent loses at most $1/4$. The Gaussian noise is the price of implementing a conditional *mean*, rather than a conditional draw, with finitely many queries.

Computed scores versus learned scores. Estimating a score from samples and estimating it from queries are different problems. The statistical theory of diffusion models bounds the number of samples needed to learn a score that is good enough for sampling [25, 26]. Here no sample is drawn from π_V and nothing is learned, so (10.2) is a worst-case bound over the class with no generalization gap. Neither set of results implies the other.

10.3 The sampler is an ancestral denoiser

Recall the variance-halving identity (8.5) of Section 8: if $Z \sim \nu_{x,r}$ and $Y = Z + N$ with N standard Gaussian, then

$$Z = \frac{Y}{2} + \frac{Z_{\text{new}}}{\sqrt{2}}, \quad Z_{\text{new}} | Y \sim \nu_{x+rY/2, r/\sqrt{2}}.$$

This is the posterior of the denoising step that halves the noise level in a variance-exploding diffusion model [23, 24]: given a noisy observation Y of a draw from $\nu_{x,r}$, the conditional law of the underlying draw is $\nu_{x+rY/2, r/\sqrt{2}}$, and the state moves by $x \mapsto x + rY/2$ with the scale replaced by $r/\sqrt{2}$. Theorem 2.1 produces exactly such observations: its output is within $\text{polylog}(d) \delta^{K_0} D(x)$ in W_2 of $\nu_{x,r} * N(0, I_d)$, so it plays the role of the noisy observation Y , and the independent Gaussian it carries is the observation noise.

Corollary 10.3 (Diffusion-model reading of the upper bound). For every fixed $\varepsilon > 0$ there is C_ε such that the following ancestral sampler has total-variation distance at most $1/10$ from π_V for every $V \in \mathcal{V}_d$, using at most $C_\varepsilon d^\varepsilon$ first-order queries on every execution and $O(\log d)$ reverse steps:

1. *Warm start.* Obtain a sample y within total-variation distance $o(1)$ of $\tilde{\pi} = \pi_V * N(0, hI_d)$, $h = \delta/2$, from the proximal chain of Section 8, whose steps are Gaussian conditional draws.
2. *Reverse steps.* Set $x_0 = 0$ and $r_0 = 1$. For $0 \leq j < J$, with J the depth $\lceil C_1 \log d \rceil$ of Section 8, call the primitive of Theorem 2.1 at center x_j and scale r_j with fresh seeds to obtain the noisy observation Y_j , then set $x_{j+1} = x_j + r_j Y_j/2$ and $r_{j+1} = r_j/\sqrt{2}$.
3. *Return.* Output $y + \sqrt{h}(x_J + r_J Z_J)$, where Z_J is produced by the mode-centred Gaussian base rule (8.7).

Proof. This is Theorem 8.5 read in the diffusion vocabulary. The identification of the reverse step is the identity (8.5) together with the backward-variable computation preceding it. The warm start is the approximate noised chain of Section 8, the base rule is the mode-centred Gaussian approximation of Lemma 8.4, and the query count is (8.11). $\square$

Why the sampler never computes a score field. The reverse steps consume each score through a conditional draw rather than a value. The finite evaluation of Section 7 adds a Gaussian to every center and absorbs it exactly into the seeds, so the deliberate noise becomes part of the next conditioning instead of an error to be summed. This is what makes the uniform query cap and the noise-halving composition possible, and it is the structural difference from samplers that evaluate an approximate score along a trajectory.

10.4 Scope, limitations, and open directions

- *Oracle versus statistical model.* The scores here are computed from exact first-order queries of V . No samples of π_V are used and no score network is trained. The statistical theory of diffusion models [25, 26] asks how many samples suffice to learn a score that is good enough for sampling. That question is different, and its bounds neither transfer to this model nor follow from it.
- *Pointwise scores do not compose naively.* Standard convergence theory for score-based generative models charges the total squared score error along the trajectory, for instance through Girsanov’s theorem: the relative entropy between the approximate and exact reverse processes is bounded by a constant times $\int \mathbb{E}|s_t - \nabla \log p_t|^2 dt$, plus initialization and discretization terms. Inserting the per-slice bound (10.2) into such an estimate gives an error of order $\delta^{2\theta}d = d^{1-2e\theta}$, which is not $o(1)$ for fixed $\theta < 1/4$. So Corollary 10.2 by itself does not give a constant-accuracy reverse-SDE sampler, and Corollary 10.3 does not follow this route.
- *Accuracy versus dimension.* The accuracy of the construction is tied to a fixed power δ^{K_0} of the perturbation size. To reach a target error η one takes $K_0 \approx \log(1/\eta)/\log(1/\delta)$, which shrinks the admissible t in (2.3) and raises the query exponent, whereas the high-accuracy results cited in Section 1.2 have polylogarithmic dependence on inverse accuracy at the price of a fixed positive dimension exponent ($d^{1/4}$ or $d^{1/5}$). The two regimes are complementary, and the present result does not claim the optimal accuracy dependence.

The diffusion reading leaves several questions open. Can the per-evaluation score bound (10.2) be improved below the exponent $1/4$, or below the factor $\sqrt{d}$? Can the warm start be replaced by a purely reverse-diffusion procedure? Does score or denoiser evaluation admit a lower bound analogous to the $\Omega(\log d)$ bound of Section 9? The upper bound itself gives no answer to any of these.

References

- [1] R. Adamczak and P. Wolff. Concentration inequalities for non-Lipschitz functions with bounded derivatives of higher order. *Probability Theory and Related Fields*, 162:531–586, 2015.
- [2] M. S. Albergo, N. M. Boffi, and E. Vanden-Eijnden. Stochastic interpolants: a unifying framework for flows and diffusions. *Journal of Machine Learning Research*, 26(209):1–80, 2025.

- [3] J. M. Altschuler, S. Chewi, and M. S. Zhang. Shifted composition IV: toward ballistic acceleration for log-concave sampling. In *Proceedings of the 58th Annual ACM Symposium on Theory of Computing*, pages 1739–1750, 2026.
- [4] M. Anshelevich. Appell polynomials and their relatives. *International Mathematics Research Notices*, 2004(65):3469–3531, 2004.
- [5] D. Bakry, I. Gentil, and M. Ledoux. *Analysis and Geometry of Markov Diffusion Operators*, volume 348 of Grundlehren der mathematischen Wissenschaften. Springer, Cham, 2014.
- [6] F. Chen, S. Chewi, J. Lu, and M. S. Zhang. Accelerated high-accuracy sampling from a warm start via the proximal bouncy particle sampler, 2026. arXiv:2609.06905.
- [7] F. Chen, S. Chewi, J. Lu, and M. S. Zhang. High-accuracy simulation of Picard HMC, part I: Gaussian cloud correction, 2026. arXiv:2609.38710.
- [8] F. Chen, S. Chewi, J. Lu, and M. S. Zhang. Sharp Gaussian divergence and Bernstein–Markov inequalities, 2026. arXiv:2609.32851.
- [9] F. Chen, S. Chewi, J. Lu, and M. S. Zhang. Smoothed Picard Hamiltonian Monte Carlo, 2026. arXiv:2609.06906.
- [10] F. Chen, S. Chewi, A. Rakhlin, and M. S. Zhang. Exact simulation of diffusions and improved algorithms for log-concave sampling, 2026. arXiv:2608.05022.
- [11] Y. Chen, S. Chewi, A. Salim, and A. Wibisono. Improved analysis for a proximal algorithm for sampling. In *Proceedings of the 35th Conference on Learning Theory*, volume 178 of PMLR, pages 2984–3014, 2022.
- [12] X. Cheng, N. S. Chatterji, P. L. Bartlett, and M. I. Jordan. Underdamped Langevin MCMC: a non-asymptotic analysis. In *Proceedings of the 31st Conference on Learning Theory*, volume 75 of PMLR, pages 300–323, 2018.
- [13] S. Chewi, J. de Dios Pont, J. Li, C. Lu, and S. Narayanan. Query lower bounds for log-concave sampling. *Journal of the ACM*, 71(4):29:1–29:42, 2024.
- [14] T. A. Courtade, M. Fathi, and A. Pananjady. Existence of Stein kernels under a spectral gap, and discrepancy bounds. *Annales de l’Institut Henri Poincaré, Probabilités et Statistiques*, 55(2):777–790, 2019.
- [15] A. S. Dalalyan. Theoretical guarantees for approximate sampling from smooth and log-concave densities. *Journal of the Royal Statistical Society: Series B*, 79(3):651–676, 2017.
- [16] A. Durmus and E. Moulines. Nonasymptotic convergence analysis for the unadjusted Langevin algorithm. *The Annals of Applied Probability*, 27(3):1551–1587, 2017.
- [17] P. K. Kothari and J. Steinhardt. Better agnostic clustering via relaxed tensor norms, 2017. arXiv:1711.07465.
- [18] Y. T. Lee, Z. Song, and S. S. Vempala. Algorithmic theory of ODEs and sampling from well-conditioned logconcave densities, 2018. arXiv:1812.06243.

- [19] Y. T. Lee, R. Shen, and K. Tian. Structured logconcave sampling with a restricted Gaussian oracle. In *Proceedings of the 34th Conference on Learning Theory*, volume 134 of PMLR, pages 2993–3050, 2021.
- [20] Y.-A. Ma, T. Chen, and E. B. Fox. A complete recipe for stochastic gradient MCMC. In *Advances in Neural Information Processing Systems*, volume 28, 2015.
- [21] O. Mangoubi and A. Smith. Rapid mixing of Hamiltonian Monte Carlo on strongly log-concave distributions, 2017. arXiv:1708.07114.
- [22] R. Shen and Y. T. Lee. The randomized midpoint method for log-concave sampling. In *Advances in Neural Information Processing Systems*, volume 32, 2019.
- [23] Y. Song, J. Sohl-Dickstein, D. P. Kingma, A. Kumar, S. Ermon, and B. Poole. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations*, 2021.
- [24] J. Ho, A. Jain, and P. Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, volume 33, 2020.
- [25] S. Chen, S. Chewi, J. Li, Y. Li, A. Salim, and A. R. Zhang. Sampling is as easy as learning the score: theory for diffusion models with minimal data assumptions. In *International Conference on Learning Representations*, 2023.
- [26] V. De Bortoli. Convergence of denoising diffusion models under the manifold hypothesis. *Transactions on Machine Learning Research*, 2022.