The Score Kalman Filter

Kaito Iwasaki
Department of Mathematics
University of Michigan
kaitoi@umich.edu

Anthony Bloch
Department of Mathematics
University of Michigan
abloch@umich.edu

Taeyoung Lee
Department of Mechanical and
Aerospace Engineering
George Washington University
tylee@gwu.edu

Maani Ghaffari
Department of Naval Architecture
& Marine Engineering and
Department of Robotics
University of Michigan
maanigj@umich.edu

Abstract

A central obstacle in nonlinear Bayesian filtering is representing the belief distribution. Moment-based filters address this by propagating polynomial moments and reconstructing a density from them. Recent work completes the predict-update loop via the maximum-entropy (MaxEnt) principle, but each step requires the partition function and its gradient, both n-dimensional integrals whose cost scales exponentially, restricting the demonstrated MaxEnt moment filtering to $n \leq 4$. We avoid the partition function entirely by combining score matching with Stein's identity. In our setting, score matching reduces the density fit to a single linear solve whose coefficients are assembled directly from the propagated moments. The same parameters then drive Stein's identity to close the moment hierarchy during prediction and to recover posterior moments after each Bayesian update, keeping the full predict-update loop free of partition function evaluation. The resulting **Score Kalman Filter** (SKF) reduces to the classical information-form Kalman filter as a special case and performs every step through linear algebra. On nonlinear coupled-oscillator networks, the SKF runs through $n = 20$ and reports lower RMSE than the EKF, UKF, EnKF, and particle-filter baselines on the tested synthetic benchmarks.

1 Introduction

Bayesian filtering, the recursive estimation of the state of a dynamical system from noisy observations, is fundamental to robotics, control, and signal processing. The Kalman filter [1] solves this problem optimally for linear-Gaussian systems, but many real-world dynamical systems are nonlinear and produce non-Gaussian beliefs: skewed, curved distributions from nonlinear dynamics, and multimodal distributions from systems with multiple stable equilibria. Gaussian filters (the extended Kalman filter (EKF [2, 3]), unscented Kalman filter (UKF [4, 5]), ensemble Kalman filter (EnKF [6, 7]), and their Lie-group variants [8, 9]) are not designed to represent these density shapes directly. Particle filters can, but their convergence rate in the number of particles makes high-accuracy filtering expensive.

A promising middle ground is *moment-based filtering* [10–13], which propagates the moments of the belief density through the system dynamics and reconstructs the density from those moments. Teng et al. [13] recently showed that polynomial MaxEnt distributions $p(x;\lambda) \propto \exp(-\lambda \cdot \phi(x))$ give a flexible, multimodal representation, demonstrated on SE(2) localization. Here $\phi(x)$ collects

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

monomials of total degree at most r in the state $x \in \mathbb{R}^n$ and $\lambda \in \mathbb{R}^M$ with $M = \binom{n+r}{n}$ is the corresponding parameter vector. The main difficulty in this approach is therefore its computational cost. Fitting λ to predicted moments and recovering posterior moments after a Bayesian update require both the *partition function* $Z(\lambda) = \int \exp(-\lambda \cdot \phi(x))\, dx$ and its gradients, n-dimensional quadratures whose cost scales as $O(G^n)$ with G quadrature nodes per dimension and becomes intractable as n grows.

Our key observation. Score matching [14] is a general method for estimating parameters of non-normalized statistical models by minimizing the Fisher divergence between the model score and the data score, without evaluating the normalization constant. We apply this idea to the polynomial exponential family and show that the estimator reduces to a *single linear system* on the non-constant coefficients, with entries assembled directly from the propagated moments (Proposition 1). The resulting fit does not require partition function evaluation, iterative optimization, or numerical integration. The density-fit cost is $O(M^3)$, requiring only linear algebra.

Two additional ingredients are needed to turn score matching into a complete filtering algorithm. First, for nonlinear systems whose moment hierarchy does not close [15], a *moment closure* is required. Classical closures either assume Gaussianity [16] or require the partition function [17, 18]. We instead exploit the *Stein identity* [19], which provides algebraic relations between moments of different orders, parameterized by the score matching parameters λ (Proposition 2). Second, after a Bayesian measurement update, the posterior moments must be recovered from the updated parameters λ^+. For polynomial measurement models, the update is conjugate: $\lambda^+ = \lambda^- + \lambda_{\text{lik}}$, where λ_{lik} is the score parameter induced by the likelihood and $-$ and $+$ denote pre- and post-measurement parameters (analogous to the information-form Kalman update). The same Stein identity yields a linear system for the posterior moments without evaluating Z. The resulting algorithm, the **Score Kalman Filter** (SKF), performs every step of the predict-update loop via linear algebra. The SKF thus belongs to the moment-based family above. It fits λ by score matching where Teng et al. [13] need the partition function and its $O(G^n)$ quadrature, and it propagates continuous-time moments through the generator (Dynkin's formula) with Stein closure. Section 4 and the experiments in Section 7 show how this method extends to higher-dimensional systems.

Contributions. (1) **Score matching reconstruction (Section 3).** We show that score matching for the polynomial exponential family reduces the density fit to a single linear system on the propagated moments, removing the need for partition-function evaluation, iterative optimization, or numerical quadrature. (2) **Stein closure and posterior recovery (Sections 4–5.2).** We use Stein's identity to obtain algebraic moment relations parameterized by the score matching coefficients λ, closing the moment hierarchy during prediction and recovering posterior moments after each Bayesian update, both as linear solves. (3) **Score Kalman Filter and experiments (Section 7).** We introduce the Score Kalman Filter (SKF) and evaluate it on coupled oscillator networks. On this structured benchmark, the SKF scales through n=20 with mean RMSE below the reported EKF/UKF/EnKF/PF baselines at n=4–20 via the active Stein closure, extending beyond the $n \leq 4$ regime demonstrated for partition-function-based MaxEnt moment filtering [13]. (4) **Recovery of the information-form Kalman filter (Appendix D).** We verify that at r=2, the SKF reduces to the information-form Kalman filter exactly.

2 Background

2.1 Stochastic systems and moment propagation

Consider a continuous-time stochastic system with state $x \in \mathbb{R}^n$ evolving via the Itô stochastic differential equation (SDE):

$$dx = X(x)\, dt + h(x)\, dW, \tag{1}$$

where $X : \mathbb{R}^n \to \mathbb{R}^n$ is the drift vector field, $h : \mathbb{R}^n \to \mathbb{R}^{n \times n_w}$ is the diffusion matrix, and W is a standard $\mathbb{R}^{n_w}$-valued Wiener process. The density $p(t, x)$ of the state satisfies the Fokker–Planck equation $\partial_t p = -\nabla \cdot (Xp) + \sum_{ij} \partial_i \partial_j (H_{ij}\, p)$, where $H = \frac{1}{2} h h^\top$ is the diffusion tensor. For constant H, the second term reduces to $\operatorname{Tr}(H \nabla^2 p)$.

Infinitesimal generator. Rather than solving the Fokker–Planck PDE for p directly, we work with the *stochastic Koopman semigroup* $\{\mathcal{K}_t\}_{t \geq 0}$, a family of linear operators acting on observable functions $f \colon \mathbb{R}^n \to \mathbb{R}$ by $\mathcal{K}_t f(x) = \mathbb{E}[f(X_t) \mid X_0 = x]$. The *infinitesimal generator* $\mathcal{A} \colon C^2(\mathbb{R}^n) \to C(\mathbb{R}^n)$

is the derivative of this semigroup at $t = 0$. For $f \in C^2(\mathbb{R}^n)$,

$$\mathcal{A}f(x) := \lim_{t \searrow 0} \frac{\mathcal{K}_t f(x) - f(x)}{t} = \nabla f(x) \cdot X(x) + \mathrm{Tr}\big(H(x)\, \nabla^2 f(x)\big), \tag{2}$$

where the second equality follows from Itô's lemma. The first term captures the deterministic drift and the second captures the stochastic diffusion. The generator gives rise to Dynkin's formula, which describes how expectations of f evolve over time,

$$\frac{d}{dt}\mathbb{E}[f(X_t)] = \mathbb{E}[\mathcal{A}f(X_t)]. \tag{3}$$

Moment propagation. Applying Dynkin's formula to the monomial test functions $\phi_\alpha(x) = x^\alpha$ (multi-index $\alpha \in \mathbb{N}^n$, $|\alpha| = \sum_i \alpha_i$, and $e_i \in \mathbb{N}^n$ denotes the i-th standard basis multi-index) gives a system of ODEs for the moments $m_\alpha(t) := \mathbb{E}[X_t^\alpha]$:

$$\frac{d}{dt} m_\alpha(t) = \mathbb{E}\big[\mathcal{A}\phi_\alpha(X_t)\big]. \tag{4}$$

We now restrict to *polynomial systems*, where each component of the drift $X(x)$ is a polynomial in x of total degree at most d_X, and each entry of the diffusion matrix $h(x)$ is a polynomial of total degree at most d_h. Under this assumption, the generator (2) applied to a monomial ϕ_α of degree $|\alpha|$ produces two polynomial contributions. The drift term $\nabla\phi_\alpha \cdot X$ has degree $|\alpha| - 1 + d_X$, and the diffusion term $\mathrm{Tr}(H\nabla^2\phi_\alpha)$ has degree $|\alpha| - 2 + 2d_h$, since $H = \frac{1}{2}hh^\top$ has entries of degree $2d_h$. Taking expectations, the right-hand side of (4) involves moments up to degree $|\alpha| + \bar{d}$, where

$$\bar{d} = \max(d_X - 1,\ 2d_h - 2) \tag{5}$$

is the *excess degree*, measuring how many additional moment degrees are introduced by the drift nonlinearity and the state-dependence of the noise. If $\bar{d} = 0$, as in systems with linear drift and additive or linear multiplicative noise, the moment ODE closes at any truncation order. If $\bar{d} \geq 1$, as in systems with quadratic drift or quadratic multiplicative noise, propagating degree-K moments requires degree-$(K+\bar{d})$ moments, producing an infinite, hierarchically coupled system.

This is the classical *moment closure problem* [15], arising in SDEs [16, 20], kinetic theory [21, 17], and network dynamics [22–24]. Classical closures either assume Gaussianity [16] or require Z-evaluation [17, 18]. Our Stein closure (Section 4) exploits the exponential family structure to derive algebraic moment relations from λ, without Gaussianity, Z, or independence.

2.2 Maximum-entropy density reconstruction and computational cost

A moment-based filter must also *reconstruct* a density from the propagated moments. The maximum-entropy (MaxEnt) principle [25] solves this by maximizing Shannon entropy subject to moment constraints, yielding the *polynomial exponential family*, or polynomial maximum-entropy distribution (MED),

$$p^*(x;\lambda) = \frac{1}{Z(\lambda)} \exp\big(-\lambda \cdot \phi(x)\big), \quad Z(\lambda) = \int \exp(-\lambda \cdot \phi(x))\, dx, \tag{6}$$

where $\phi(x) = (x^\alpha)_{|\alpha| \leq r}$ and $\lambda \in \mathbb{R}^M$, $M = \binom{n+r}{n}$. Let $\Lambda = \{\lambda : Z(\lambda) < \infty\}$ denote the natural parameter domain. This family includes Gaussians (r=2) and captures multimodality at $r \geq 4$ [13]. The **computational cost** is dominated by evaluating $Z(\lambda)$ and its gradient at every optimization step, an n-dimensional integral with cost $O(G^n)$. This exponential scaling in n is what the score matching formulation in Section 3 replaces with a single linear solve.

2.3 Score matching for unnormalized models

Let p_{data} denote the true (unknown) density of the state at a given time, whose moments $m_\alpha = \mathbb{E}_{p_{\mathrm{data}}}[x^\alpha]$ have been propagated by the moment ODE (4). We wish to fit the polynomial exponential-family model in (6) to p_{data}, without the intractable computation of the normalization $Z(\lambda)$.

Hyvärinen [14] observed that the *score function* $s(x;\lambda) = \nabla_x \log p(x;\lambda) \in \mathbb{R}^n$ does not depend on Z. For the polynomial exponential-family model in (6), $s(x;\lambda) = -\nabla_x(\lambda \cdot \phi) = -J_\phi(x)\lambda$,

where $[J_\phi]_{i\alpha} = \partial_i \phi_\alpha$, so $J_\phi(x) \in \mathbb{R}^{n \times M}$. The Fisher divergence between p_{data} and the model, $\frac{1}{2}\mathbb{E}_{p_{\text{data}}}[\|s_\lambda - s_{\text{data}}\|^2]$ where $s_{\text{data}} = \nabla_x \log p_{\text{data}}$, can be rewritten via integration by parts as

$$J_{\text{SM}}(\lambda) = \mathbb{E}_{p_{\text{data}}}\left[\sum_i \partial_i s_i(x;\lambda) + \frac{1}{2} s_i(x;\lambda)^2\right] + \text{const}, \tag{7}$$

which depends only on the model score and its divergence, not on Z or s_{data}. The expectation in (7) is over p_{data} and reduces to polynomial moments of p_{data} when $s(x;\lambda)$ is polynomial, as we show in Section 3.

3 Score matching for the polynomial exponential family

We now specialize the score matching objective (7) to the polynomial exponential family (6). Substituting the polynomial score $s(x;\lambda) = -J_\phi(x)\lambda$, the expectation over p_{data} reduces to polynomial moments, and the objective becomes a quadratic function of λ, $J_{\text{SM}}(\lambda) = \frac{1}{2}\lambda^\top A\lambda - b^\top\lambda + \text{const}$, where A and b are defined below. Additive constants do not affect the score, so we fix the zero-multi-index coefficient $\lambda_{\mathbf{0}}$ to 0 and write A and b for the system after the constant row and column are removed.

Proposition 1 (Score matching as a linear system)**.** *For the polynomial MED (6), A and b in the right-hand side of the score matching objective have the form*

$$A_{\alpha\beta} = \sum_{i=1}^n \alpha_i\,\beta_i\; m_{\alpha+\beta-2e_i}, \qquad b_\alpha = \sum_{i=1}^n \alpha_i(\alpha_i - 1)\; m_{\alpha-2e_i}, \tag{8}$$

for $\alpha, \beta \in \mathbb{N}^n$ with $|\alpha|, |\beta| \le r$. The constant row and column are then omitted according to the convention above. If A is nonsingular, the score matching estimator is

$$\lambda^* = A^{-1}b, \qquad \lambda^*_{\mathbf{0}} = 0. \tag{9}$$

Every entry of A and b is a polynomial moment. A requires moments up to degree $2r-2$, and b requires moments up to $r-2$.

Proof sketch. $A = \mathbb{E}[J_\phi^\top J_\phi]$ and $b_\alpha = \mathbb{E}[\Delta\phi_\alpha]$ (Laplacian of the monomial ϕ_α). Both reduce to moment lookups via $\partial_i(x^\alpha) = \alpha_i x^{\alpha-e_i}$ with the convention that $\alpha_i x^{\alpha-e_i} = 0$ when $\alpha_i = 0$. The constant row and column vanish because $\nabla\phi_{\mathbf{0}} = 0$. Under this convention, positive definiteness follows from the polynomial-gradient argument in Appendix B.3. See Appendix B. □

Key properties. (i) Solving (9) costs $O(M^3)$, where $M = \binom{n+r}{n}$. (ii) No partition function $Z(\lambda)$ is evaluated. (iii) No iterative optimization (BFGS, Newton, etc.) is needed. (iv) *Sensitivity to moment errors* (Proposition 3, Appendix B.2): if the propagated moments have relative error $\|\delta m\|/\|m\|$, the fitted parameters shift by

$$\frac{\|\delta\lambda\|}{\|\lambda\|} = O\big(\kappa(A)\,\|\delta m\|/\|m\|\big), \tag{10}$$

where $\kappa(A)$ denotes the condition number of A. When propagated moments leave the realizable set, the MaxEnt dual has no finite minimizer, while the linear system still has a solution whenever A is nonsingular (Remark 5).

3.1 Score matching recovers the correct parameters for consistent moments

A natural concern is that score matching (which minimizes the Fisher divergence) does not explicitly enforce moment constraints, and hence the fitted density may not match the propagated moments. The following theorem shows that this concern vanishes when the moments are consistent with a density in the polynomial exponential family.

Theorem 1 (Correct recovery on the model class)**.** *Let $p_0 = p(\cdot;\lambda_{\text{true}})$ be a member of the degree-r polynomial exponential family (6), with $\lambda_{\text{true}} \in \text{int}(\Lambda)$ under the zero-constant convention $\lambda_{\mathbf{0}} = 0$. Form A and b in (8) from the moments of p_0 up to degree $2r-2$. Then the score matching solution recovers the non-constant coefficients,*

$$\lambda^*_\alpha = (\lambda_{\text{true}})_\alpha, \qquad 1 \le |\alpha| \le r. \tag{11}$$

Consequently $p(\cdot;\lambda^*) = p_0$.

On this family the Fisher divergence to p_0 is the quadratic form $\frac{1}{2}(\lambda - \lambda_{\text{true}})^\top A\,(\lambda - \lambda_{\text{true}})$, which vanishes only at λ_{true} because A is positive definite under the convention above. The score matching objective differs from it by a constant independent of λ [14], so the linear solve finds λ_{true} from the moments alone. The proof, with its regularity conditions, is in Appendix B.3.

Remark 1 (Connection to MaxEnt)**.** By a classical result in information theory, the member of the exponential family $\mathcal{P}_r$ with moments m is the unique density that maximizes the Shannon entropy among all densities sharing those moments. Combined with Theorem 1, this implies that *when the moments are consistent with the family*, score matching and MaxEnt produce the same density, even though the two methods optimize different objectives (Fisher divergence vs. KL divergence). When the true density lies outside $\mathcal{P}_r$, the two methods generally give different parameters. In our filtering pipeline, the moments are propagated by Dynkin's formula (exact for polynomial systems with appropriate $\bar{d}$), so the conditions of Theorem 1 are approximately satisfied, explaining the close moment agreement observed empirically (Section 7).

4 Stein closure for the moment hierarchy

4.1 The closure problem

As discussed in Section 2, propagating the moments of degree $|\alpha| \le K$ via (4) requires moments of degree $K + \bar{d}$ on the right-hand side. The case $\bar{d} = 0$, which covers linear drift with at most linearly multiplicative noise, is the easy one because the hierarchy closes at any chosen truncation K and no auxiliary information is needed. The cases of practical interest are $\bar{d} \ge 1$, where the moment ODE for a degree-K state vector depends on moments at degree $K + 1$ or higher, and some rule must supply those higher moments. Classical closures fill this gap by imposing distributional assumptions on the unresolved moments, typically Gaussianity [16] or independence of pairs and triples [22, 23]. We instead derive the higher moments from the score parameters λ that the score matching fit has already produced, using the Stein identity below.

4.2 Score-based closure via Stein's identity

Proposition 2 (Stein's identity for the polynomial MED)**.** *If $p(x;\lambda)$ has score $s = -J_\phi \lambda$ and the displayed moments are finite and all boundary terms vanish, then for any multi-index β and coordinate i,*

$$\sum_{|\alpha| \le r} \lambda_\alpha \, \alpha_i \, m_{\alpha+\beta-e_i} = \beta_i \, m_{\beta-e_i}. \tag{12}$$

Terms with negative multi-indices are interpreted as zero.

Stein's identity $\mathbb{E}[s_i f + \partial_i f] = 0$ is integration by parts against the density. With $f(x) = x^\beta$ and the polynomial score, both terms become moments, so the identity is a linear relation among moments with coefficients λ. The derivation, including the boundary-term conditions, is in Appendix B.

Equation (12) is a **linear relation** between moments, parameterized by λ. To close the hierarchy, we target an unclosed moment m_γ with $|\gamma| = K + 1$ by choosing α_0 with $|\alpha_0| = r$ and setting $\beta = \gamma - \alpha_0 + e_i$, which gives an equation expressing m_γ in terms of moments of degree $\le |\gamma|$ and λ (Appendix G.1 works through a scalar case). Specifically, since $|\beta| = K - r + 2$, we have $|\alpha + \beta - e_i| \le K + 1$ and the left-hand side involves moments up to degree $|\gamma| = K + 1$, while the right-hand side depends only on known moments as $|\beta - e_i| = K - r + 1 \le K$. Collecting all such equations yields a linear system $\Lambda_1(\lambda)\, m^{(K+1)} = c_1(\lambda, m^{(\le K)})$ for the degree-$(K{+}1)$ unknowns. For $\bar{d} \ge 2$, the procedure is applied sequentially, with layer j resolving degree $K{+}j$ from degree $\le K{+}j{-}1$.

The coefficient matrix $\Lambda_1(\lambda)$ has polynomial entries in λ, so its rank-deficient locus is an algebraic variety – either all of $\mathbb{R}^M$ or a proper algebraic subset of Lebesgue measure zero. For $n{=}2$ (any $r \ge 2$), the first-layer system is square and generically nonsingular (Appendix B.5). At $r = 2$, positive precision is a simple sufficient condition. For general n and $r \ge 3$, the simple $n{=}2$ picture gives way to an explicit counting problem. Appendix B.5 carries out this count. At $r{=}3$, the augmented full-basis system is overdetermined through $n{=}15$ and becomes underdetermined at $n{=}16$. At $r{=}4$, the same crossover moves to the gap between $n{=}35$ and $n{=}36$. Our full-basis experiments all fall on the overdetermined side of this count, and Λ_1 had full column rank at every time step in those

tests ($n \leq 10$, $r \leq 8$). For structured systems, we may instead close only the unclosed moments that appear in the moment ODE. This is the active Stein closure used for the coupled oscillator scaling experiment through n=20. The resulting linear system is overdetermined but typically inconsistent, and we take its least-squares solution.

Remark 2 (Factorization caching)**.** The coefficient matrix $\Lambda_1(\lambda)$ depends only on the score parameters λ, not on the moments m. During prediction, λ is fixed while $m(t)$ evolves through N_{sub} ODE substeps. Therefore Λ_1 need be factorized only *once* per prediction window. Each substep rebuilds only the right-hand side $c(\lambda,\, m^{(\leq K)})$ and solves via back-substitution, reducing the closure cost from $O(N_{\text{sub}}\, N_{\text{eq}}\, N_{\text{unk}}^2)$ to $O(N_{\text{eq}}\, N_{\text{unk}}^2 + N_{\text{sub}}\, N_{\text{unk}}^2)$.

Self-consistency loop. At each time step, $\lambda(t)$ from the previous score matching fit provides the coefficients in the Stein closure. After propagation, the new moments yield the next score matching fit, which is used for the following closure.

Remark 3 (Moment budget versus identifiability)**.** Although a member of $\mathcal{P}_r$ is uniquely determined, after fixing the constant parameter, by its moments up to degree r via the bijectivity of the moment map (Appendix A), the score matching estimator in Proposition 1 requires moments up to degree $2r-2$. The gap reflects a distinction between information-theoretic identifiability and algorithmic recoverability, since the quadratic structure of $J_{\text{SM}}(\lambda) = \frac{1}{2}\lambda^\top A\lambda - b^\top\lambda$ makes the Gram matrix A depend on moments of degree up to $2r-2$. Therefore, the resulting Stein closure should be interpreted as a *model-based* reconstruction of higher-order moments (given lower-order moments plus the polynomial-MED assumption), rather than a universal lower-to-higher moment generation procedure.

4.3 Stein-augmented score matching

The Stein identity at $|\beta| = \ell$ with $\ell < r$ reproduces the score matching equations (up to scaling). The first non-redundant constraint is at $\ell = r$, introducing m_{2r-1}. This augmentation tightens the moment consistency of the fitted λ beyond what score matching alone provides. The connection between the generalized score matching and the generalized method of moments is formalized by Kume and Walker [19].

5 Measurement update and moment recovery

5.1 Score-based Bayesian update

Let $z \in \mathbb{R}^{n_z}$ be a measurement of the state, given by a function of the state and noise through the likelihood $p(z \mid x)$. At measurement time, the posterior score adds the likelihood score,

$$s_{\text{post}}(x) = s_{\text{prior}}(x) + \nabla_x \log p(z|x). \tag{13}$$

Equation (13) is Bayes' rule in score form: $\log p(x|z) = \log p(x) + \log p(z|x) - \log p(z)$, and taking ∇_x removes the evidence term because $p(z)$ is independent of x. If the likelihood $p(z|x)$ has a polynomial measurement model $g(x)$ of degree d_g with Gaussian noise, the likelihood score $\nabla_x \log p(z|x) = J_g^\top R^{-1}(z - g(x))$ is polynomial of degree $\leq 2d_g - 1$. When $2d_g - 1 \leq r - 1$ (i.e., $d_g \leq r/2$), the posterior score fits within the basis, giving

$$\lambda^+ = \lambda^- + \lambda_{\text{lik}}(z). \tag{14}$$

Equation (14) is the usual exponential-family conjugate update written in natural parameters. The update changes λ by addition and does not require evaluating the evidence or the partition function.

5.2 Moment recovery from posterior parameters

After the update (14), λ^+ is known but m^+ is not. Although $m^+ = \mu(\lambda^+)$ is identifiable on the non-constant minimal parameterization (Appendix A), evaluating μ would again require Z. We instead impose the componentwise Stein identity at $\lambda = \lambda^+$ and $m_0 = 1$, using directional rows (β, i) with $|\beta| \geq 1$, $\beta_i \geq 1$,

$$\sum_{|\alpha| \leq r} \lambda_\alpha^+\, \alpha_i\, m^+_{\alpha+\beta-e_i} = \beta_i\, m^+_{\beta-e_i}. \tag{15}$$

Algorithm 1 Score Kalman Filter (polynomial MED, degree r)

Require: Initial moments $m(0)$, basis order r, dynamics (X, h), measurement model (g_{obs}, R)

$K \leftarrow 2r - 2$ {moment budget}

$\lambda \leftarrow A(m(0))^{-1} b(m(0))$ {initial score matching}

for each time step **do**

  Predict.

    Propagate moments $m(t) \to m(t+\Delta t)$ via (4), using Stein closure (12) with $\lambda(t)$ if $\bar{d} \geq 1$ (example in Appendix G.1)

    $\lambda^- \leftarrow A(m)^{-1} b(m)$ {score-matching solve}

  Update at measurement z.

    $\lambda^+ \leftarrow \lambda^- + \lambda_{\text{lik}}(z)$ {conjugate score addition}

    Recover m^+ from λ^+ via Stein system (15) {one least-squares solve}

    $\lambda \leftarrow A(m^+)^{-1} b(m^+)$ {re-fit λ}

end for

Rows with $|\beta| = \ell$ couple moments up to degree $r + \ell - 1$. Within the tracked budget $K = 2r - 2$, the exact directional rows are underdetermined (e.g., $R = 12$, $U = 27$ for $n = 2, r = 4$), so we include higher-$|\beta|$ rows and truncate degree $> K$ terms. This gives a least-squares projection onto the Stein consistency manifold, with omitted near-Gaussian centered terms of order $O(\sigma^{K+2})$. Counts, optional $\beta_i = 0$ rows, and conditioning are detailed in Appendix B.5.

Conditioning. Working in centered coordinates $z = x - \mu$ (where μ is the current mean, available from the propagated moments) is essential. The monomial basis in raw coordinates gives $\|\lambda\| \sim 10^5$, while centering reduces this to $O(10)$–$O(100)$. Orthogonal bases (Legendre) provide further improvement [26]. The truncation error decreases as the neglected higher-order centered moments become smaller, which is why centering is important for accuracy.

Iterative consistency refinement. The conjugate update $\lambda^+ = \lambda^- + \lambda_{\text{lik}}$ followed by truncated Stein recovery produces moments m^+ that may not be perfectly consistent with λ^+. We enforce consistency by iterating: (i) re-fit λ via score matching to the recovered m^+, (ii) re-apply the measurement $\lambda \leftarrow \lambda_{\text{refit}} + \lambda_{\text{lik}}$, (iii) re-recover moments. A fixed point of this iteration is a (λ, m) pair that is mutually consistent under both the Stein identity and the measurement update. In experiments where refinement is enabled, a small number of iterations suffices (the residual drops by 2–3 orders of magnitude). The high-dimensional coupled-oscillator sweep uses one refinement step for speed. The construction connects score matching to Stein's method of moments (SMoM) [19, 27], with the canonical decomposition and its convergence implications in Appendix E.

6 The Score Kalman Filter algorithm

Algorithm 1 assembles the preceding pieces into one predict-update loop. The filter carries both moments m, used for Dynkin propagation, and score parameters λ, used for density fitting and Stein relations. Prediction advances m and refits λ. The update adds the likelihood score, recovers posterior moments by the same Stein system, and reports the standard posterior mean $\hat{x}_k = \mathbb{E}[x_k \mid z_{1:k}]$ from the recovered first moments for plots and RMSE.

Complexity. The dense score fit costs $O(M^3)$ for $M = \binom{n+r}{n}$. If $\bar{d} \geq 1$, Stein closure adds a solve with N_{eq} rows and N_{unk} unclosed moments (dense QR: $O(N_{\text{eq}} N_{\text{unk}}^2)$). The factorization or sparse operator is reused over ODE substeps because Λ_1 depends only on λ (Remark 2). Structured generators can restrict closure to active moments, as in the $n = 12$–20 oscillator runs. No partition function is evaluated, and $r = 2$ recovers the information-form Kalman filter (Appendix D).

7 Experiments

We benchmark the SKF across SE(2) rigid-body kinematics, ecosystem dynamics, 3D incompressible fluid advection, and coupled oscillator filtering up to n=20. Full moment panels, density reconstructions, and filter time series for the remaining cases are in Appendices J.2, J.3, and J.5. Note that all runs use Python (NumPy/SciPy) on a single laptop CPU (Intel i9-11900H, no GPU or compiled

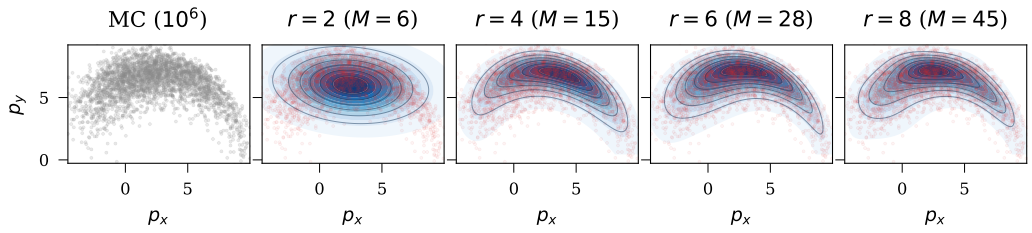


Figure 1: SE(2) density reconstruction. The leftmost panel shows the Monte Carlo reference for the (p_x, p_y) marginal, and the four panels to its right show the score matching reconstruction at basis orders $r = 2, 4, 6, 8$. The Gaussian fit at $r = 2$ misses the banana-shaped curvature entirely, while $r = 4$ already recovers the bulk of the non-Gaussian structure, with $r = 6, 8$ refining the tails.

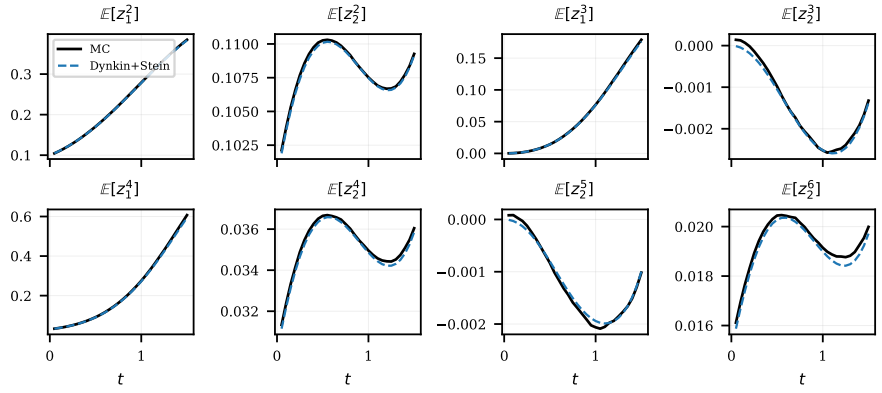


Figure 2: Representative LV centered moments (n=2, r=6, $\bar{d}$=1): Dynkin + Stein (blue dashed) vs. MC (black), degrees 2–6 (up to the score matching truncation order r). Error grows monotonically with degree.

extensions), so timings are conservative. Generator calculations, moment ODE derivations, and centered-coordinate formulas are in Appendices G–H.

7.1 SE(2) prediction with Fourier $\times$ position moments ($\bar{d} = 0$)

SE(2) rigid body kinematics in embedded coordinates (c, s, p_x, p_y) with ω=1 rad/s, v=2 m/s, heading noise σ=0.3. Drift and diffusion are linear, so $\bar{d} = 0$ and $m(T) = e^{LT} m(0)$ in the geometry-adapted Fourier $\times$ position basis $\mathbb{E}[\cos(k\theta)\, p_x^a p_y^b]$, $\mathbb{E}[\sin(k\theta)\, p_x^a p_y^b]$. All moments match $5{\times}10^5$-particle Monte Carlo (MC) simulation to within 0.5% (Appendix J.2, Figure A4). Score matching on the (p_x, p_y) marginal captures the non-Gaussian curved shape from $r = 4$ ($M = 15$, 0.4 ms, timing in Table A1) (Figure 1).

7.2 Stochastic Lotka-Volterra with Stein closure ($\bar{d} = 1$)

Stochastic predator-prey dynamics $dx_1 = (\alpha x_1 - \beta x_1 x_2)\, dt + \sigma_1\, dW_1$, $dx_2 = (-\gamma x_2 + \delta x_1 x_2)\, dt + \sigma_2\, dW_2$ (α=1, β=0.5, γ=0.8, δ=0.3, σ_1=0.3, σ_2=0.2). The bilinear interaction gives d_X=2, $\bar{d}$=1, the textbook moment closure benchmark [16]. With r=6, T=1.5 s, and centered coordinates, Dynkin+Stein moments match 10^6-particle MC to 0.065% (variance) and 2–7% (third-order), error growing monotonically with degree (Figure 2, full grid in Appendix J.2). The Dynkin+Stein run takes 0.7 s on the CPU used for Table A1. The MC reference used for validation takes 69 s. Density reconstruction in Appendix J.3 (Figure A8).

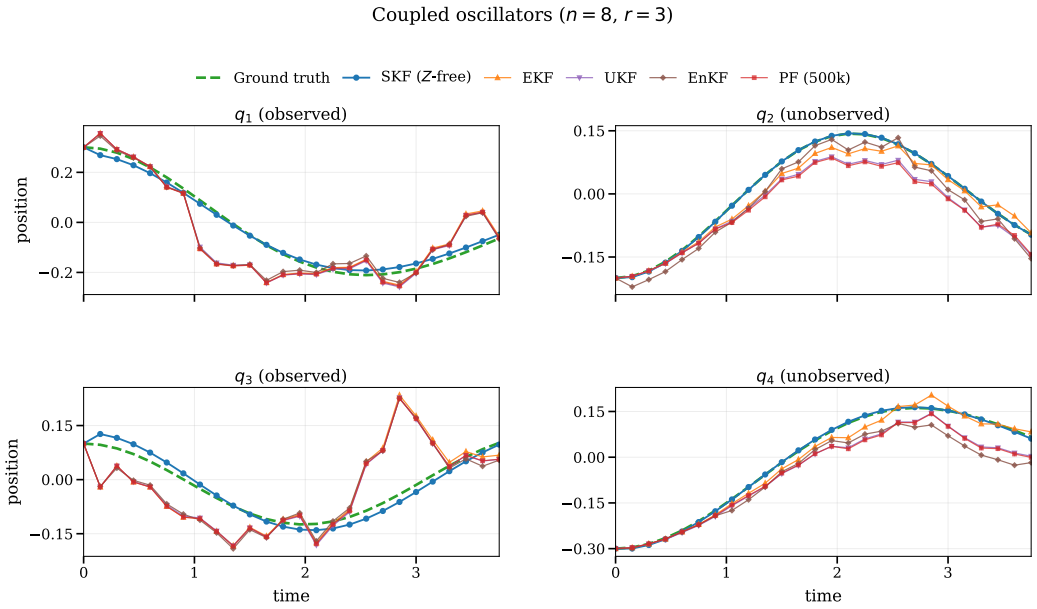


Figure 3: Coupled oscillators (n=8, r=3). The SKF (blue) tracks all four positions through 25 steps, including unobserved oscillators (q_2, q_4) inferred from coupling alone. EKF (orange), UKF (purple), EnKF (brown), and PF (red) have larger RMSE. Multi-seed statistics are reported in Table A1.

7.3 Filtering on coupled oscillator networks

We demonstrate the full predict-update SKF loop on N coupled Duffing oscillators with nearest-neighbor coupling ($\dot{q}_i = p_i$, $\dot{p}_i = -\gamma p_i - \alpha q_i - \beta q_i^2 + \kappa(q_{i+1} - 2q_i + q_{i-1}) + \sigma\, dW_i$, $n = 2N$, $d_X = 2$, $\bar{d} = 1$, centered coordinates, Appendix H.4, constants in Appendix J.5). The single-oscillator n=2 Duffing filter is in Appendix J.4. At N=3, 4, 5 (n=6, 8, 10, r=3) observing the odd-indexed positions, the SKF has lower mean RMSE than EKF/UKF/EnKF/PF across 10 independent seeds in our implementation (Figure 3, Table A1).

We further scale to $n = 12, \ldots, 20$ with $r = 3$ using an active Stein closure that keeps only the degree-five moments that the moment ODE actually requests, rather than the full augmented Stein system that the structural count of Appendix B.5 predicts to become underdetermined at $n = 16$ (Appendix J.5). Even at these dimensions, the SKF stays more accurate than every baseline in Table A1. The fact that EKF, UKF, and EnKF all converge to a similar error band suggests that the Gaussian belief approximation, not the choice of update law, is what is limiting them on this benchmark. Factorization caching also keeps the SKF runtime below the $5{\times}10^5$-particle bootstrap filter at $n = 10$, where the particle count is the only setting that gave the PF stable tracking. From n=16 the PF is faster, taking 646 s against 2079 s for the SKF at n=20, while its RMSE stays higher. All numbers in Table A1 are averaged over 10 random seeds, with the SKF reported as mean $\pm$ std.

8 Related work

Gaussian filtering. The KF [1], EKF [2, 3], UKF [4, 5], EnKF [6, 7], and geometric variants (InEKF [8, 28], EqF [29], UKF-M [9]) form a mature lineage, but their standard forms do not provide an explicit multimodal density representation. Key et al. [30] scale Gaussian inference to weather data assimilation by message passing on a Gaussian Markov random field. The SKF generalizes the information-form KF (Appendix D) to polynomial exponential family beliefs.

Moment-based filtering. Moment-based filtering for polynomial systems has been developed primarily for robotics state estimation, including closed-form moment recursions for discrete-time trigonometric-polynomial systems [10], the moment-based Kalman filter [11], the generalized mo-

ment Kalman filter [12], and the MEM-KF [13] which couples moment propagation with MaxEnt density reconstruction. These filters operate on discrete-time prediction/update maps. A nonlinear system needs moments above the tracked order in discrete and continuous time alike, so discretizing does not remove the closure problem (Appendix I). The generator raises a degree-K moment additively to degree $K + \bar{d}$, whereas a discrete transition of polynomial degree d_F raises it to $d_F K$. The SKF extends moment-based filtering to continuous-time polynomial SDEs by closing the hierarchy with Stein's identity.

Score matching and Stein's method of moments. Score matching [14] fits unnormalized models without Z, with extensions to generative modeling [31] and S^2 filtering [32]. Stein's method of moments (SMoM) derives Z-free estimators from Stein characterizations of a distribution [33, 27]. Kume and Walker [19] showed that a special case of the SMoM estimator coincides with generalized score matching, and that extending it to the generalized method of moments yields generalized score matching estimators with optimal properties. Nagai and Yano [27] decompose SMoM estimators around the score matching estimator and use the decomposition to construct one with smaller asymptotic variance. This connection motivates our iterative consistency refinement (Section 5.2). In filtering, Cao et al. [34] use Stein's lemma, the Gaussian (r=2) case of (12), to derive the optimality conditions of a natural-gradient Gaussian filter.

Particle filters and data-driven methods. Particle filters [35] represent non-Gaussian beliefs by Monte Carlo samples but suffer weight degeneracy under partial observation, with accuracy controlled by the particle count. KalmanNet [36] replaces the Kalman gain by a learned neural network at the cost of supervised training data.

9 Conclusion

The SKF replaces partition-function evaluation by linear solves for density fitting, moment closure, and posterior recovery. It agrees with MaxEnt on the polynomial exponential family (Theorem 1) and recovers the information-form Kalman filter at $r = 2$, giving a partition-function-free route to high-order moment filtering. The same algebra also opens a research program around automatic basis selection, adaptive truncation, active Stein closures for sparse generators, and diagnostics that decide when higher-order moments are worth propagating. We discuss these directions in Appendix L.

Acknowledgements

A.M.B. was supported in part by NSF grants 2103026 and DMS-2605296 and AFOSR grants FA9550-23-1-021. Authors supported by FA9550-23-1-0400 (MURI). M. Ghaffari was supported by AFOSR YIP FA9550-25-1-0224. The authors declare no competing interests.

References

[1] R. E. Kalman. A new approach to linear filtering and prediction problems. *Journal of Basic Engineering (ASME Transactions, Series D)*, 82(1):35–45, 1960.

[2] Stanley F. Schmidt. Application of state-space methods to navigation problems. In *Advances in Control Systems*, volume 3, pages 293–340. Academic Press, 1966.

[3] Andrew H. Jazwinski. *Stochastic Processes and Filtering Theory*. Academic Press, New York, 1970.

[4] Simon J. Julier and Jeffrey K. Uhlmann. New extension of the Kalman filter to nonlinear systems. In *Signal Processing, Sensor Fusion, and Target Recognition VI*, volume 3068 of *Proceedings of SPIE*, pages 182–193, 1997.

[5] Eric A. Wan and Rudolph van der Merwe. The unscented Kalman filter for nonlinear estimation. In *Proceedings of the IEEE Adaptive Systems for Signal Processing, Communications, and Control Symposium (AS-SPCC)*, pages 153–158, 2000.

[6] Geir Evensen. Sequential data assimilation with a nonlinear quasi-geostrophic model using Monte Carlo methods to forecast error statistics. *Journal of Geophysical Research: Oceans*, 99 (C5):10143–10162, 1994.

[7] Gerrit Burgers, Peter Jan van Leeuwen, and Geir Evensen. Analysis scheme in the ensemble Kalman filter. *Monthly Weather Review*, 126(6):1719–1724, 1998.

[8] Axel Barrau and Silvère Bonnabel. The invariant extended Kalman filter as a stable observer. *IEEE Transactions on Automatic Control*, 62(4):1797–1812, 2017.

[9] Martin Brossard, Axel Barrau, and Silvère Bonnabel. A code for unscented Kalman filtering on manifolds (UKF-M). In *IEEE International Conference on Robotics and Automation (ICRA)*, pages 5701–5708, 2020.

[10] Ashkan Jasour, Allen Wang, and Brian C. Williams. Moment-based exact uncertainty propagation through nonlinear stochastic autonomous systems. *arXiv preprint arXiv:2101.12490*, 2021.

[11] Yutaka Shimizu, Ashkan Jasour, Maani Ghaffari, and Shinpei Kato. Moment-based Kalman filter: Nonlinear Kalman filtering with exact moment propagation. In *IEEE International Conference on Robotics and Automation (ICRA)*, pages 3948–3954. IEEE, 2023.

[12] Sangli Teng, Harry Zhang, David Jin, Ashkan Jasour, Maani Ghaffari, and Luca Carlone. GMKF: Generalized moment Kalman filter for polynomial systems with arbitrary noise. *arXiv preprint arXiv:2403.04712*, 2024.

[13] Sangli Teng, Harry Zhang, David Jin, Ashkan Jasour, Ram Vasudevan, Maani Ghaffari, and Luca Carlone. Max entropy moment Kalman filter for polynomial systems with arbitrary noise. In *NeurIPS*, 2025.

[14] Aapo Hyvärinen. Estimation of non-normalized statistical models by score matching. *Journal of Machine Learning Research*, 6:695–709, 2005.

[15] Christian Kuehn. Moment closure—a brief review. In *Control of Self-Organizing Nonlinear Systems*, Understanding Complex Systems, pages 253–271. Springer, Cham, 2016.

[16] Peter Whittle. On the use of the normal approximation in the treatment of stochastic processes. *Journal of the Royal Statistical Society, Series B*, 19(2):268–281, 1957.

[17] C. David Levermore. Moment closure hierarchies for kinetic theories. *Journal of Statistical Physics*, 83(5–6):1021–1065, 1996.

[18] Abhyudai Singh and João P. Hespanha. Approximate moment dynamics for chemically reacting systems. *IEEE Transactions on Automatic Control*, 56(2):414–418, 2011.

[19] Alfred Kume and Stephen G. Walker. On Stein's method of moments and generalized score matching. *arXiv preprint arXiv:2602.06482*, 2026.

[20] Colin S. Gillespie. Moment-closure approximations for mass-action models. *IET Systems Biology*, 3(1):52–58, 2009.

[21] Harold Grad. On the kinetic theory of rarefied gases. *Communications on Pure and Applied Mathematics*, 2(4):331–407, 1949.

[22] John G. Kirkwood. Statistical mechanics of fluid mixtures. *The Journal of Chemical Physics*, 3 (5):300–313, 1935.

[23] Matt J. Keeling. The effects of local spatial structure on epidemiological invasions. *Proceedings of the Royal Society of London B*, 266(1421):859–867, 1999.

[24] István Z. Kiss, Joel C. Miller, and Péter L. Simon. *Mathematics of Epidemics on Networks: From Exact to Approximate Models*, volume 46 of *Interdisciplinary Applied Mathematics*. Springer Cham, 2017.

[25] Lawrence R. Mead and N. Papanicolaou. Maximum entropy in the problem of moments. *Journal of Mathematical Physics*, 25(8):2404–2417, 1984.

[26] Hang Liu, Sangli Teng, and Maani Ghaffari. MePoly: Max entropy polynomial policy optimization. *arXiv preprint arXiv:2602.17832*, 2026.

[27] Mitsuki Nagai and Keisuke Yano. The geometry of Stein's method of moments: A canonical decomposition via score matching. *arXiv preprint arXiv:2603.12843*, 2026.

[28] Ross Hartley, Maani Ghaffari, Ryan M. Eustice, and Jessy W. Grizzle. Contact-aided invariant extended Kalman filtering for robot state estimation. *The International Journal of Robotics Research*, 39(4):402–430, 2020.

[29] Pieter van Goor, Tarek Hamel, and Robert Mahony. Equivariant filter (EqF). *IEEE Transactions on Automatic Control*, 68(6):3501–3512, 2023.

[30] Oscar Key, So Takao, Daniel Giles, and Marc Peter Deisenroth. Scalable data assimilation with message passing. *Environmental Data Science*, 4:e1, 2025. doi: 10.1017/eds.2024.47.

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

[32] Mario Bukal, Ivan Marković, and Ivan Petrović. Score matching based assumed density filtering with the von Mises-Fisher distribution. In *20th International Conference on Information Fusion (FUSION)*, 2017.

[33] Bruno Ebner, Adrian Fischer, Robert E. Gaunt, Babette Picker, and Yvik Swan. Stein's method of moments. *Scandinavian Journal of Statistics*, 52(4):1594–1624, 2025.

[34] Wenhan Cao, Tianyi Zhang, Zeju Sun, Chang Liu, Stephen S.-T. Yau, and Shengbo Eben Li. Nonlinear Bayesian filtering with natural gradient Gaussian approximation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 48(7):8581–8597, 2026. doi: 10.1109/TPAMI.2026.3674120.

[35] Renato Zanetti, Andrey A. Popov, Kristen Michaelson, Felipe Giraldo-Grueso, Dalton Durant, Simone Servadio, and Uwe D. Hanebeck. A survey of nonlinear estimation filters. *Journal of Advances in Information Fusion*, 2025. Accepted for publication.

[36] Guy Revach, Nir Shlezinger, Xiaoyong Ni, Adrià López Escoriza, Ruud J. G. van Sloun, and Yonina C. Eldar. KalmanNet: Neural network aided Kalman filtering for partially known dynamics. *IEEE Transactions on Signal Processing*, 70:1532–1547, 2022.

[37] Martin J. Wainwright and Michael I. Jordan. Graphical models, exponential families, and variational inference. *Foundations and Trends in Machine Learning*, 1(1–2):1–305, 2008.

[38] Edwin T. Jaynes. Information theory and statistical mechanics. *Physical Review*, 106(4):620–630, 1957.

[39] Charles Stein. A bound for the error in the normal approximation to the distribution of a sum of dependent random variables. *Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability*, 2:583–602, 1972.

[40] Bernt Øksendal. *Stochastic Differential Equations: An Introduction with Applications*. Springer, Berlin, 6th edition, 2003.

[41] Daniel Choukroun, Itzhack Y. Bar-Itzhack, and Yaakov Oshman. Novel quaternion Kalman filter. *IEEE Transactions on Aerospace and Electronic Systems*, 42(1):174–190, 2006.

[42] Brian C. Hall. *Lie Groups, Lie Algebras, and Representations: An Elementary Introduction*, volume 222 of *Graduate Texts in Mathematics*. Springer, 2nd edition, 2015.

[43] Gregory S. Chirikjian. *Stochastic Models, Information Theory, and Lie Groups, Volume 2: Analytic Methods and Modern Applications*. Birkhäuser, Boston, 2012.

[44] Jean B. Lasserre. Global optimization with polynomials and the problem of moments. *SIAM Journal on Optimization*, 11(3):796–817, 2001. doi: 10.1137/S1052623400366802.

[45] Pablo A. Parrilo. Semidefinite programming relaxations for semialgebraic problems. *Mathematical Programming*, 96(2):293–320, 2003. doi: 10.1007/s10107-003-0387-5.

[46] Jackson Gorham and Lester Mackey. Measuring sample quality with Stein's method. In *NeurIPS*, 2015.

[47] Alessandro Barp, François-Xavier Briol, Andrew Duncan, Mark Girolami, and Lester Mackey. Minimum Stein discrepancy estimators. In *NeurIPS*, 2019.

[48] Andreas Anastasiou, Alessandro Barp, François-Xavier Briol, Bruno Ebner, Robert E. Gaunt, Fatemeh Ghaderinezhad, Jackson Gorham, Arthur Gretton, Christophe Ley, Qiang Liu, Lester Mackey, Chris J. Oates, Gesine Reinert, and Yvik Swan. Stein's method meets computational statistics: A review of some recent developments. *Statistical Science*, 38(1):120–139, 2023.

[49] Valentin De Bortoli, Emile Mathieu, Michael Hutchinson, James Thornton, Yee Whye Teh, and Arnaud Doucet. Riemannian score-based generative modelling. In *NeurIPS*, 2022.

[50] Weixin Wang and Taeyoung Lee. Uncertainty propagation for general stochastic hybrid systems on compact Lie groups. *SIAM Journal on Applied Dynamical Systems*, 21(3):2215–2240, 2022.

[51] René Thom. *Structural Stability and Morphogenesis*. W. A. Benjamin, 1975.

[52] R. E. Kalman. Contributions to the theory of optimal control. *Boletín de la Sociedad Matemática Mexicana*, 5(2):102–119, 1960.

[53] D. Q. Mayne, J. B. Rawlings, C. V. Rao, and P. O. M. Scokaert. Constrained model predictive control: Stability and optimality. *Automatica*, 36(6):789–814, 2000.

[54] Aaron D. Ames, Xiangru Xu, Jessy W. Grizzle, and Paulo Tabuada. Control barrier function based quadratic programs for safety critical systems. *IEEE Transactions on Automatic Control*, 62(8):3861–3876, 2017.

A Background on exponential families and maximum entropy

This section collects standard results on exponential families [37, 38] that are used in our proofs. We state them for the polynomial exponential family $\mathcal{P}_r = \{p(\cdot;\lambda) = \exp(-\lambda \cdot \phi)/Z(\lambda) : \lambda \in \Lambda\}$ with natural parameter domain $\Lambda = \{\lambda \in \mathbb{R}^M : Z(\lambda) < \infty\}$, but with the usual identifiability convention that the constant statistic is removed, or equivalently that $\lambda_{\mathbf{0}} = 0$ is fixed. Without this convention, shifting the constant parameter only rescales $Z(\lambda)$ and leaves the density unchanged, so strict convexity and injectivity cannot hold on the full redundant parameter space. The results below hold for this minimal representation, and more generally for any exponential family with affinely independent sufficient statistics.

Log-partition function. The log-partition function $\Psi(\lambda) := \log Z(\lambda) = \log \int \exp(-\lambda \cdot \phi(x))\, dx$ is convex on Λ and strictly convex on $\mathrm{int}(\Lambda)$ after the constant direction has been removed. Its gradient and Hessian are

$$\nabla_\lambda \Psi = -\mathbb{E}_{p_\lambda}[\phi(x)], \qquad \nabla^2_\lambda \Psi = \mathrm{Cov}_{p_\lambda}[\phi(x)], \tag{A.1}$$

where $\mathrm{Cov}_{p_\lambda}[\phi]$ is the covariance matrix of the sufficient statistics under $p(\cdot;\lambda)$. Since the covariance matrix is positive semidefinite (and positive definite when the sufficient statistics in the minimal representation are not a.s. affinely dependent), Ψ is strictly convex on that representation.

Moment map. The *moment space* $\mathcal{M} \subset \mathbb{R}^{M-1}$ is the set of all non-constant moment vectors $m = \mathbb{E}_q[\phi(x)]$ realizable by some probability distribution q on $\mathbb{R}^n$. Define the moment map $\mu\colon \mathrm{int}(\Lambda) \to \mathcal{M}$ by $\mu(\lambda) = \mathbb{E}_{p_\lambda}[\phi(x)]$. Since $\mu = -\nabla\Psi$ and Ψ is strictly convex on the minimal parameterization, μ is injective: distinct parameters give distinct moments. Moreover, μ is a diffeomorphism from $\mathrm{int}(\Lambda)$ onto $\mathrm{int}(\mathcal{M})$ in this minimal representation: bijectivity follows from strict convexity of Ψ, smoothness of μ from analyticity of Ψ, and smoothness of μ^{-1} from the inverse function theorem since $\nabla^2\Psi \succ 0$ on $\mathrm{int}(\Lambda)$ implies $\nabla\mu = -\nabla^2\Psi$ is invertible. This means that for any moment vector $m \in \mathrm{int}(\mathcal{M})$, there is a *unique* $\lambda \in \mathrm{int}(\Lambda)$ such that $\mathbb{E}_{p_\lambda}[\phi] = m$.

Maximum entropy characterization. Among all densities q satisfying $\mathbb{E}_q[\phi_\alpha(x)] = m_\alpha$ for $1 \le |\alpha| \le r$ (and $\int q = 1$, $q \ge 0$), the one that maximizes the Shannon entropy $\mathcal{H}(q) = -\int q \log q\, dx$ is $p(\cdot;\lambda(m))$, the unique exponential family member with those moments. This follows from the Gibbs inequality: for any q with $\mathbb{E}_q[\phi] = m$,

$$\mathcal{H}(q) = -\mathbb{E}_q[\log q] \le -\mathbb{E}_q[\log p_\lambda] = \mathbb{E}_q[\lambda \cdot \phi] + \log Z(\lambda) = \lambda \cdot m + \Psi(\lambda), \tag{A.2}$$

with equality if and only if $q = p_\lambda$ a.e. The second inequality uses $\mathrm{KL}(q\|p_\lambda) \ge 0$, which expands to $-\mathcal{H}(q) - \mathbb{E}_q[\log p_\lambda] \ge 0$.

Implication for this paper. Score matching recovers λ from moments without evaluating Ψ or solving the entropy optimization. Theorem 1 shows that the recovered parameter agrees with the MaxEnt/moment-matching parameter when the moments are consistent with the family. The bijectivity of μ guarantees that this parameter is unique.

B Proofs

B.1 Proof of Proposition 1 (score matching linear system)

For the polynomial MED with energy $E_\lambda(x) = \lambda \cdot \phi(x)$, the model score is $s_\lambda(x) = -\nabla_x E_\lambda = -J_\phi(x)\lambda$, where $[J_\phi]_{i\alpha} = \partial_i \phi_\alpha = \alpha_i x^{\alpha - e_i}$. The Hyvärinen score matching objective (7) becomes

$$\begin{aligned} J_{\mathrm{SM}}(\lambda) &= \mathbb{E}\left[\sum_i \partial_i s_i(x;\lambda) + \frac{1}{2} s_i(x;\lambda)^2\right] \\ &= \mathbb{E}\left[-\sum_i \partial_i (J_\phi(x)\lambda)_i + \frac{1}{2}\|J_\phi(x)\lambda\|^2\right] \\ &= -\mathbb{E}[\Delta\phi(x)]^\top \lambda + \frac{1}{2}\lambda^\top \mathbb{E}[J_\phi(x)^\top J_\phi(x)]\lambda \\ &= \frac{1}{2}\lambda^\top A \lambda - b^\top \lambda, \end{aligned} \tag{A.3}$$

where $A = \mathbb{E}[J_\phi^\top J_\phi]$ and $b = \mathbb{E}[\Delta\phi]$. We now compute the entries explicitly.

Matrix A**.** The (i,α)-entry of J_ϕ is $[J_\phi]_{i\alpha} = \partial_i(x^\alpha) = \alpha_i x^{\alpha-e_i}$. Therefore

$$\begin{aligned}[J_\phi^\top J_\phi]_{\alpha\beta} &= \sum_{i=1}^{n} [J_\phi]_{i\alpha}\,[J_\phi]_{i\beta} \\ &= \sum_{i=1}^{n} \alpha_i x^{\alpha-e_i} \cdot \beta_i x^{\beta-e_i} \\ &= \sum_{i=1}^{n} \alpha_i\,\beta_i\,x^{\alpha+\beta-2e_i}. \end{aligned} \tag{A.4}$$

Taking the expectation under p_{data} and writing $m_\gamma = \mathbb{E}_{p_{\text{data}}}[x^\gamma]$ gives

$$A_{\alpha\beta} = \mathbb{E}_{p_{\text{data}}}[J_\phi^\top J_\phi]_{\alpha\beta} = \sum_{i=1}^{n} \alpha_i\,\beta_i\,m_{\alpha+\beta-2e_i}. \tag{A.5}$$

The multi-index $\alpha+\beta-2e_i$ has degree $|\alpha|+|\beta|-2 \le 2r-2$, so every entry of A is a known propagated moment.

Vector b**.** The Laplacian of ϕ_α is $\Delta\phi_\alpha = \sum_i \partial_i^2(x^\alpha) = \sum_i \alpha_i(\alpha_i-1)\,x^{\alpha-2e_i}$. Taking the expectation gives $b_\alpha = \sum_i \alpha_i(\alpha_i-1)\,m_{\alpha-2e_i}$, which requires moments up to degree $r-2$.

The minimizer satisfies $\nabla_\lambda J_{\text{SM}} = A\lambda - b = 0$, solved on the non-constant subspace. The row and column corresponding to $|\alpha|=0$ (the constant monomial) are identically zero because $\partial_i(x^0)=0$, so the constant component $\lambda_{\mathbf{0}}$ (corresponding to the zero multi-index) is unidentifiable and is dropped.

After this deletion, A is the Gram matrix of the gradients of the remaining monomials. Indeed, for any coefficient vector v and $q_v(x) = \sum_{1\le|\alpha|\le r} v_\alpha x^\alpha$,

$$v^\top A v = \mathbb{E}_{p_{\text{data}}}\left[\|J_\phi(x)v\|^2\right] = \mathbb{E}_{p_{\text{data}}}\left[\|\nabla q_v(x)\|^2\right]. \tag{A.6}$$

Thus A is positive semidefinite. It is positive definite precisely when no nonzero polynomial in the chosen non-constant basis has zero gradient almost surely under p_{data}. For the smooth densities used in this paper this holds: if the last display is zero, then ∇q_v vanishes on an open set, hence everywhere by polynomiality. Therefore q_v is constant, and since q_v has no constant term, $v=0$. □

Remark 4 (A pathological example for positive definiteness)**.** The positive definiteness statement is not meant to cover laws that live on a lower-dimensional set. For example, if a two-dimensional state is supported on the line $x_2=0$, then the nonzero quadratic $q(x)=x_2^2$ has $\nabla q = 0$ on the support. The associated coefficient vector is therefore in the nullspace of A.

B.2 Perturbation bound

Proposition 3 (Moment perturbation bound)**.** *Let $A(m)$ and $b(m)$ be the score matching matrix and vector assembled from moments m via (8), and let $\lambda^* = A^{-1}b$ on the non-constant subspace, where $A(m)$ is nonsingular and $\lambda^*, b(m)$ are nonzero. Suppose the propagated moments are perturbed: $\tilde{m} = m + \delta m$, with δm small enough that $A(\tilde{m})$ is also nonsingular. Then the perturbed parameters $\tilde{\lambda} = A(\tilde{m})^{-1}b(\tilde{m})$ satisfy*

$$\frac{\|\tilde{\lambda}-\lambda^*\|}{\|\lambda^*\|} \le \kappa(A)\left(\frac{\|\delta A\|}{\|A\|} + \frac{\|\delta b\|}{\|b\|}\right) + O(\|\delta m\|^2), \tag{A.7}$$

where $\delta A = A(\tilde{m}) - A(m)$, $\delta b = b(\tilde{m}) - b(m)$, and $\kappa(A) = \|A\|\,\|A^{-1}\|$.

Proof. Since each entry of A and b is a linear combination of moments (equations (A.68)–(A.69)), both $A(m)$ and $b(m)$ are *linear* in m. Therefore $\delta A = A(\delta m)$ and $\delta b = b(\delta m)$ (using linearity), and $\|\delta A\| \le C_A\|\delta m\|$, $\|\delta b\| \le C_b\|\delta m\|$ for constants C_A, C_b depending only on the basis.

The perturbed system is $(A+\delta A)(\lambda^*+\delta\lambda) = b + \delta b$. Expanding and dropping the $O(\|\delta m\|^2)$ term $\delta A\cdot\delta\lambda$:

$$\delta\lambda = A^{-1}(\delta b - \delta A\,\lambda^*). \tag{A.8}$$

Taking norms: $\|\delta\lambda\| \le \|A^{-1}\|(\|\delta b\| + \|\delta A\|\,\|\lambda^*\|)$. Dividing by $\|\lambda^*\|$ and using $\|A^{-1}\| = \kappa(A)/\|A\|$:

$$\frac{\|\delta\lambda\|}{\|\lambda^*\|} \le \kappa(A)\,\frac{\|\delta A\|}{\|A\|} + \kappa(A)\,\frac{\|\delta b\|}{\|A\|\,\|\lambda^*\|} \le \kappa(A)\left(\frac{\|\delta A\|}{\|A\|} + \frac{\|\delta b\|}{\|b\|}\right), \tag{A.9}$$

where the last step uses $\|b\| = \|A\lambda^*\| \le \|A\|\,\|\lambda^*\|$. □

Remark 5 (Contrast with MaxEnt realizability)**.** The MaxEnt dual is $\min_\lambda\{\lambda \cdot m + \log Z(\lambda)\}$, a convex program on $\Lambda = \{\lambda : Z(\lambda) < \infty\}$. On the non-constant minimal parameterization, $\log Z$ is strictly convex on $\mathrm{int}(\Lambda)$, so the moment map $\lambda \mapsto m(\lambda) = -\nabla_\lambda \log Z(\lambda)$ is a diffeomorphism from $\mathrm{int}(\Lambda)$ onto an open convex set $\mathcal{M} \subset \mathbb{R}^M$, the *mean parameter space* (Wainwright and Jordan [37], Theorem 3.3). When the target moments m lie outside $\mathcal{M}$ (which can happen when moments are propagated numerically and accumulate error), the MaxEnt dual has no finite minimizer: $\|\lambda_k\| \to \infty$ along the optimization iterates. In practice, this manifests as the optimizer failing to converge, returning a degenerate density (mass concentrating on domain vertices), or producing numerical overflow. In this case the optimization problem has no finite minimizer.

The score matching linear system $A\lambda = b$, by contrast, has a solution whenever A is nonsingular, regardless of whether m lies in $\mathcal{M}$. This is because the score matching objective $J_{\mathrm{SM}} = \frac{1}{2}\lambda^\top A\lambda - b^\top\lambda$ is a quadratic in λ with no domain constraint. The fitted λ^* may not correspond to a normalizable density when $m \notin \mathcal{M}$, but the score function $s(x) = -J_\phi(x)\lambda^*$ remains well-defined and can still be used for Langevin sampling or downstream computations.

B.3 Proof of Theorem 1 (score matching and MaxEnt equivalence)

We require the following setup and regularity conditions.

Setup. Let $\mathcal{P}_r = \{p(\cdot;\lambda) = \exp(-\lambda\cdot\phi(x))/Z(\lambda) : \lambda \in \Lambda\}$ be the degree-r polynomial exponential family, where $\phi(x) = (x^\alpha)_{|\alpha|\le r}$ is the monomial sufficient statistic and $\Lambda = \{\lambda \in \mathbb{R}^M : Z(\lambda) < \infty\}$ is the natural parameter space. Let $\lambda_{\mathrm{true}} \in \mathrm{int}(\Lambda)$ and denote $p_0 = p(\cdot;\lambda_{\mathrm{true}})$. Suppose the propagated moments $m_\alpha = \mathbb{E}_{p_0}[x^\alpha]$ for $|\alpha| \le 2r-2$ are the exact moments of p_0.

Regularity conditions.

- (R1) $\lambda_{\mathrm{true}} \in \mathrm{int}(\Lambda)$ with $\lambda_{\mathbf{0}} = 0$, and the support of p_0 has nonempty interior. For the polynomial family on $\mathbb{R}^n$, $p_0(x) > 0$ on all of $\mathbb{R}^n$.
- (R2) All moments $\mathbb{E}_{p_0}[|x^\alpha|]$ for $|\alpha| \le 2r-2$ are finite (guaranteed by $\lambda_{\mathrm{true}} \in \mathrm{int}(\Lambda)$ for exponential families).
- (R3) The boundary terms in the integration-by-parts step of score matching vanish, i.e., $\lim_{\|x\|\to\infty} p_0(x)\,s_\lambda(x) = 0$ for all $\lambda \in \Lambda$. This holds when the leading-degree term of E_λ grows sufficiently fast (e.g., even degree with positive leading coefficient).

Proof. We proceed in three steps.

Step 1. The Fisher divergence vanishes at λ_{true}.

The Fisher divergence between p_0 and $p(\cdot;\lambda)$ is

$$D_F(p_0\|p_\lambda) = \frac{1}{2}\mathbb{E}_{p_0}\left[\|s_0(x) - s_\lambda(x)\|^2\right], \tag{A.10}$$

where $s_0 = \nabla_x \log p_0 = -J_\phi\lambda_{\mathrm{true}}$ and $s_\lambda = -J_\phi\lambda$. Substituting,

$$D_F(p_0\|p_\lambda) = \frac{1}{2}\mathbb{E}_{p_0}\left[\|J_\phi(\lambda - \lambda_{\mathrm{true}})\|^2\right] = \frac{1}{2}(\lambda - \lambda_{\mathrm{true}})^\top A_0(\lambda - \lambda_{\mathrm{true}}), \tag{A.11}$$

where $A_0 = \mathbb{E}_{p_0}[J_\phi^\top J_\phi]$ is the Gram matrix evaluated under p_0. At $\lambda = \lambda_{\mathrm{true}}$, we have $D_F = 0$.

Step 2. A_0 is positive definite (on the non-constant subspace), so λ_{true} is the unique minimizer.

We must show that $A_0 v = 0$ implies $v = 0$ (restricted to $|\alpha| \ge 1$). $A_0 v = 0$ means $\mathbb{E}_{p_0}[\|J_\phi v\|^2] = 0$, which implies $J_\phi(x)v = 0$ for p_0-a.e. x. By condition (R1), this polynomial vector field vanishes on a set with nonempty interior, and hence vanishes identically. Writing $q_v(x) = \sum_{1\le|\alpha|\le r} v_\alpha x^\alpha$, this says $\nabla q_v \equiv 0$, so q_v is constant. Since q_v has no constant term, $q_v \equiv 0$, and therefore $v = 0$.

This shows A_0 is positive definite on $\{v : v_{\mathbf{0}} = 0\}$, so $D_F(p_0\|p_\lambda) = 0$ if and only if $\lambda_\alpha = (\lambda_{\text{true}})_\alpha$ for all $|\alpha| \geq 1$.

Step 3. The score matching objective has the same minimizer.

The Hyvärinen score matching objective, evaluated under p_0, is

$$J_{\text{SM}}(\lambda) = \mathbb{E}_{p_0}\left[\sum_i \partial_i(s_\lambda)_i + \frac{1}{2}(s_\lambda)_i^2\right] = D_F(p_0\|p_\lambda) + \mathbb{E}_{p_0}\left[\sum_i \partial_i(s_0)_i + \frac{1}{2}(s_0)_i^2\right], \quad \text{(A.12)}$$

where the second term is independent of λ (this is the standard identity from Hyvärinen [14], valid under (R3)). Therefore J_{SM} and D_F differ by a constant, and they share the same minimizer.

From Steps 1–2, the unique minimizer of D_F (restricted to $\lambda_{\mathbf{0}} = 0$) is λ_{true}. Therefore $\lambda^*_\alpha = (\lambda_{\text{true}})_\alpha$ for all $|\alpha| \geq 1$.

Moment matching follows. Since λ^* and λ_{true} agree on all non-constant components, $p(\cdot;\lambda^*) = p(\cdot;\lambda_{\text{true}}) = p_0$, and the fitted density has exactly the moments m. The uniqueness of λ_{true} within $\mathcal{P}_r$ follows from the bijectivity of the moment map (Appendix A). □

Remark 6 (When the conditions fail)**.** If the propagated moments m do *not* correspond to any $p_0 \in \mathcal{P}_r$ (the generic case when the true filtering density is not in the polynomial exponential family), then λ^* minimizes D_F over $\mathcal{P}_r$ but does not match m exactly. The moment mismatch $\|m_{\lambda^*} - m\|$ (where m_{λ^*} are the moments of $p(\cdot;\lambda^*)$) is then controlled by the approximation quality of $\mathcal{P}_r$ for the true density. Bounding this mismatch in terms of r and the smoothness of the true density is an interesting direction for future work.

B.4 Proof of Proposition 2 (Stein identity)

Stein's identity [39] states that for a smooth density p with score $s_i = \partial_i \log p$ and any smooth test function f with $\mathbb{E}[|s_i f|] < \infty$,

$$\mathbb{E}[s_i(x)f(x) + \partial_i f(x)] = 0, \quad \text{(A.13)}$$

provided $\lim_{\|x\|\to\infty} p(x)f(x) = 0$ (the boundary term from integration by parts vanishes). This follows from $\int (s_i f + \partial_i f)p\,dx = \int (\partial_i p \cdot f/p + \partial_i f)p\,dx = \int \partial_i(pf)\,dx = [pf]_{\text{boundary}} = 0$.

Substituting $s_i = -\sum_\alpha \lambda_\alpha \alpha_i x^{\alpha - e_i}$ (the polynomial MED score) and $f(x) = x^\beta$,

$$\mathbb{E}\Big[-\sum_\alpha \lambda_\alpha \alpha_i x^{\alpha+\beta-e_i} + \beta_i x^{\beta-e_i}\Big] = 0. \quad \text{(A.14)}$$

Rearranging gives (12). □

B.5 Well-posedness of the Stein closure system

The one-layer Stein closure assembles a linear system $\Lambda_1(\lambda)\, m^{(K+1)} = c(\lambda, m^{(\leq K)})$ for the degree-$(K{+}1) = (2r{-}1)$ unknowns. We count the rows (equations) and columns (unknowns) for general n and r.

Unknowns. The number of monomials of degree exactly $2r{-}1$ in n variables is $U(n,r) = \binom{2r+n-2}{n-1}$.

Equations (first layer, $|\beta| = r$ only). The implemented directional subsystem uses each (β, i) pair with $|\beta| = r$ and $\beta_i \geq 1$. For a given i, the number of such β is $\binom{r+n-2}{n-1}$ (substitute $\beta'_i = \beta_i - 1$). Summing over i gives $R_1(n,r) = n\binom{r+n-2}{n-1}$. For $n = 2$, $R_1 = 2r = U$, so the system is square. For $n \geq 3$, $R_1 < U$ and the first layer alone is underdetermined. The component-wise Stein identity is also valid when $\beta_i = 0$. Those rows have zero right-hand side and may be appended as additional least-squares constraints. The counts in this appendix are therefore conservative counts for the directional subsystem used in the reported implementation.

Proposition 4 (Well-posedness of the directional Stein closure for $n{=}2$)**.** *For $n{=}2$ and any $r \geq 2$, the directional Stein closure subsystem is square ($2r$ equations for $2r$ unknowns). Its determinant is a nonzero polynomial in the degree-r score coefficients. Hence the closure is well-posed for all λ outside a proper algebraic variety of Lebesgue measure zero. For $r = 2$, nonsingularity holds whenever the precision matrix $\Omega \succ 0$.*

Proof. Each entry of Λ_1 is a monomial in λ (specifically $\lambda_\alpha \cdot \alpha_i$), so every $k \times k$ minor is a polynomial in λ. The rank-deficient locus is the zero set of all maximal minors, which is an algebraic variety. It remains to show that this variety is proper. Set the only nonzero degree-r coefficients to $\lambda_{(r,0)}$ and $\lambda_{(0,r)}$. Then the equations from the x_1-derivative solve the degree-$(2r-1)$ moments with x_1-exponent at least r, while the equations from the x_2-derivative solve the remaining moments with x_2-exponent at least r. After ordering the unknowns this matrix is diagonal, with nonzero diagonal entries $r\lambda_{(r,0)}$ and $r\lambda_{(0,r)}$. Thus $\det \Lambda_1$ is not the zero polynomial, and its zero set has Lebesgue measure zero. At r=2, the four Stein equations at $|\beta|$=2 for the four degree-3 moments reduce to the Gaussian moment–cumulant relation $\kappa_3 = 0$, uniquely solvable when $\Omega \succ 0$. □

Augmented directional equations ($|\beta| \in [r, K]$). Stein equations at $|\beta| = r+j$ ($j = 1, \ldots, r-2$) also contribute to the degree-$(2r-1)$ unknowns via $|\alpha| = r-j$ terms. Terms with $|\alpha| < r-j$ involve already tracked moments, while terms with $|\alpha| > r-j$ involve degrees above $K+1$ and are omitted in the truncated closure solve. The augmented row count is

$$R(n, r) = \sum_{j=0}^{r-2} n \binom{r+j+n-2}{n-1}.$$

Asymptotically, $R/U \sim (2r-1)(2r-2)/n$, so the crossover occurs at a dimension of order $(2r-1)(2r-2)$. For the regimes used in the experiments, the exact counts are concrete: at r=3, $R/U > 1$ through n=15 and $R/U < 1$ starting at n=16 ($R/U = 1.37$ at n=10, 1.09 at n=14, and 0.45 at n=40). At r=4, the crossover is between n=35 and n=36.

Extended directional equations ($|\beta| \in [r, K+1]$). The Stein identity at $|\beta| = K+1$ also involves degree-$(K+1)$ unknowns (through the $|\alpha| = 1$ terms), while the right-hand side $\beta_i\, m_{\beta - e_i}$ involves only degree-K moments (still tracked). Including these equations adds $n\binom{K+n-1}{n-1}$ rows to the system. Higher-degree moments (degree $\geq K+2$) that appear in these equations through $|\alpha| \geq 2$ terms are truncated (set to zero). For near-Gaussian densities this truncation error is $O(\sigma^{K+2})$.

The extended row count is

$$R_{\text{ext}}(n, r) = \sum_{j=0}^{r-1} n \binom{r+j+n-2}{n-1}.$$

The additional $j = r-1$ term (from $|\beta| = K+1 = 2r-1$) contributes the dominant number of equations at large n. At n=40, r=3: R_{ext} = 5,428,400 for U = 1,086,008 unknowns (5.00× overdetermined), compared to R = 492,000 (0.45×, underdetermined) without the extension. This restores overdetermination at large n at the cost of truncating higher-degree moments in the extended equations.

Active closure for structured generators. The counts above are for closing *all* degree-$(K+1)$ moments. In a concrete moment ODE, only moments that actually appear in the generator must be closed. For the coupled oscillator network at r=3 (K=4), the only unclosed terms have the form $m_{\alpha - e_{p_j} + 2e_{q_j}}$ from the quadratic force $-\beta q_j^2$ in dp_j. At n=20, the full degree-five closure would have $\binom{24}{5} = 42{,}504$ unknowns and the standard augmented system is underdetermined. Restricting to the dynamically active targets leaves 14,500 unknown moments and 32,800 Stein equations, a 2.26× overdetermined system. This is the active closure used in the high-dimensional coupled oscillator experiments. This also explains why the n=16, 18, 20 runs use r=3 rather than r=4: the full-basis heuristic counts moments that the coupled-oscillator generator never requests, while r=4 would increase the score basis at n=20 from $\binom{23}{3} = 1771$ to $\binom{24}{4} = 10{,}626$ functions and the propagated moment budget from degree 4 to degree 6.

Remark 7 (Empirical well-posedness)**.** In all full-basis experiments with $n \leq 10$ and $r \leq 8$ (using the standard augmented system, $|\beta| \in [r, K]$), Λ_1 had full column rank at every time step and we have not encountered a rank-deficient instance. For the coupled oscillator scaling sweep through n=20, the active closure systems were overdetermined after restricting to the moments requested by the generator. For larger dense closures where the standard system becomes underdetermined, the extended system ($|\beta|$ up to $K+1$) can restore overdetermination. A rigorous truncation error bound for the extended equations as a function of the departure from Gaussianity remains open.

C The Fokker–Planck equation and the score evolution PDE

The SKF propagates finitely many moments rather than the full density. This section derives the underlying PDEs (Fokker–Planck equation and the score evolution equation) governing the density and its score, providing the continuous-time foundation for the moment propagation via Dynkin's formula used in Section 4.

C.1 The Fokker–Planck equation

The probability density $p(t, x)$ of the state X_t satisfying the SDE (1) evolves according to the *Fokker–Planck / forward Kolmogorov* equation [40]:

$$\frac{\partial p}{\partial t} = -\sum_i \frac{\partial}{\partial x_i}\big(X_i\, p\big) + \sum_{i,j} \frac{\partial^2}{\partial x_i\, \partial x_j}\big(H_{ij}\, p\big), \tag{A.15}$$

or in compact notation, $\partial_t p = -\nabla \cdot (Xp) + \mathrm{Tr}(H\nabla^2 p)$ (for constant $H = \frac{1}{2}hh^\top$). This is a *linear* PDE in p, but it lives in an infinite-dimensional function space, meaning $p(t, \cdot)$ must be tracked as a full field over $\mathbb{R}^n$. Direct numerical computation by finite-difference or finite-element methods scales as $O(G^n)$ for grid resolution G, the same exponential cost as the partition function (6).

C.2 Derivation of the moment ODE (Dynkin's formula)

The moment ODE (4) arises [40, Ch. 7] by testing the Fokker–Planck equation against monomials $\phi_\alpha(x) = x^\alpha$. Multiplying (A.15) by x^α and integrating over $\mathbb{R}^n$,

$$\frac{d}{dt} m_\alpha = \frac{d}{dt} \int x^\alpha\, p(t, x)\, dx = \int x^\alpha\, \partial_t p(t, x)\, dx. \tag{A.16}$$

Substituting the Fokker–Planck equation and integrating by parts (assuming p and its derivatives decay sufficiently fast at infinity),

$$\int x^\alpha \big[-\nabla \cdot (Xp)\big] dx = \int \nabla(x^\alpha) \cdot X(x)\, p(x)\, dx = \mathbb{E}\big[\nabla\phi_\alpha \cdot X\big], \tag{A.17}$$

$$\int x^\alpha \sum_{i,j} \partial_i\partial_j (H_{ij} p)\, dx = \int \sum_{i,j} H_{ij}(x)\, \partial_i\partial_j (x^\alpha)\, p(x)\, dx = \mathbb{E}\big[\mathrm{Tr}(H(X_t)\nabla^2\phi_\alpha)\big]. \tag{A.18}$$

The first equality in each line uses integration by parts (once for the drift term, twice for the diffusion term), transferring the derivatives from p onto the test function x^α. Combining both terms gives

$$\frac{d}{dt} m_\alpha = \mathbb{E}\big[\nabla\phi_\alpha \cdot X + \mathrm{Tr}(H\nabla^2\phi_\alpha)\big] = \mathbb{E}\big[\mathcal{A}\phi_\alpha(X_t)\big], \tag{A.19}$$

which is the moment ODE (4). This reduces the infinite-dimensional Fokker–Planck PDE to a system of ODEs for the moments, at the cost of the closure problem discussed in Section 2.

C.3 The score evolution PDE

An alternative to tracking p is to track the *score* $s(t, x) = \nabla_x \log p(t, x)$. We derive its evolution equation from (A.15) and show that it is equally or more intractable than the Fokker–Planck equation.

Step 1: Time derivative of $\log p$. The Fokker–Planck equation $\partial_t p = -\nabla \cdot (Xp) + \mathrm{Tr}(H\nabla^2 p)$ gives, after dividing by p and using $\nabla p/p = s$:

$$\partial_t \log p = -(\nabla \cdot X) - X \cdot s + s^\top H s + \mathrm{Tr}(HS), \tag{A.20}$$

where $S = \nabla s = \nabla^2 \log p$ is the *score Hessian*. To obtain (A.20), we used $\partial_i\partial_j p/p = s_i s_j + S_{ij}$ (from differentiating $\partial_i p = p\, s_i$).

Step 2: Gradient to get $\partial_t s$. Since $s = \nabla \log p$, we have $\partial_t s = \nabla(\partial_t \log p)$. Taking ∇ of each term in (A.20):

$$\nabla(-X \cdot s) = -J_X^\top s - SX, \tag{A.21}$$

$$\nabla(s^\top H s) = 2\, SHs, \tag{A.22}$$

$$[\nabla\, \mathrm{Tr}(HS)]_k = \textstyle\sum_{ij} H_{ij}\, T_{ijk}, \quad T_{ijk} := \partial_k S_{ij}, \tag{A.23}$$

where J_X is the Jacobian of X ($[J_X]_{ij} = \partial_j X_i$) and T is the *third-order score tensor*.

Result: score evolution PDE.

$$\boxed{\partial_t s = -\nabla(\nabla \cdot X) - J_X^\top s - SX + 2SHs + \nabla[\mathrm{Tr}(HS)].} \tag{A.24}$$

Why this is intractable. The PDE (A.24) for the score s (a vector field) is not self-contained. The term $2SHs$ involves the score Hessian $S_{ij} = \partial_i s_j$ (a matrix field), and the term $\nabla[\mathrm{Tr}(HS)]$ involves the third-order tensor $T_{ijk} = \partial_k S_{ij}$ (the spatial derivative of the Hessian). To evolve s, one must know S. To evolve S, one must know T. To evolve T, one must know the fourth derivative of $\log p$, and so on. This is an *infinite hierarchy of coupled PDEs* in which each level requires the next, structurally analogous to the moment hierarchy (4) where each moment order couples to the next. Truncating the score hierarchy at any finite order introduces an uncontrolled approximation, just as truncating the moment hierarchy does without a proper closure.

Directly propagating the score therefore does not avoid the closure problem. It merely reformulates it in score space. Our approach works with finite-dimensional moment equations rather than solving either the Fokker–Planck PDE or the score PDE directly. It does so by propagating *moments* via the ODE (4) and using score matching (9) only for density reconstruction from those moments.

D Linear-Gaussian specialization and recovery of the information-form Kalman filter

We show that at $r = 2$, Algorithm 1 reduces exactly to the information-form Kalman filter, thus establishing the SKF as a strict generalization of the classical linear filter.

D.1 The information form of the Kalman filter

The standard Kalman filter maintains the mean μ and covariance P of a Gaussian belief. The *information form* is reparameterized in terms of the *precision* (information) matrix $\Omega := P^{-1}$ and the *information vector* $\eta := \Omega\mu = P^{-1}\mu$. The Gaussian density is:

$$p(x) = \frac{|\Omega|^{1/2}}{(2\pi)^{n/2}} \exp\Big(-\frac{1}{2}(x-\mu)^\top \Omega\,(x-\mu)\Big) \propto \exp\Big(-\frac{1}{2}x^\top \Omega\, x + \eta^\top x\Big). \tag{A.25}$$

The information form is preferred in multi-sensor fusion because the update is additive (no matrix inversion), while the prediction requires inverting Ω, the opposite of the covariance form.

D.2 Correspondence between SKF and information-form parameters

At $r = 2$, the polynomial MED is $p(x;\lambda) \propto \exp(-\lambda_{\mathbf{0}} - \lambda_1^\top x - x^\top \Lambda_2 x)$, where $\lambda_{\mathbf{0}} \in \mathbb{R}$ is the constant component, $\lambda_1 \in \mathbb{R}^n$ collects the degree-1 parameters, and $\Lambda_2 \in \mathbb{R}^{n\times n}_{\mathrm{sym}}$ collects the degree-2 parameters. Comparing with (A.25):

$$\Omega = 2\Lambda_2, \qquad \eta = -\lambda_1, \qquad \mu = \Omega^{-1}\eta = -\frac{1}{2}\Lambda_2^{-1}\lambda_1. \tag{A.26}$$

The score is $s(x) = -\nabla(\lambda\cdot\phi) = -\lambda_1 - 2\Lambda_2 x = \eta - \Omega x = -\Omega(x-\mu)$, confirming that the score of a Gaussian is affine in x with slope $-\Omega$ (negative precision). The score Hessian is $S = \nabla s = -\Omega$ (constant), and all third-order derivatives vanish ($T_{ijk} = 0$).

D.3 Score matching recovers the information parameters

At $r = 2$ in 1D (for clarity), the score matching system $A\lambda = b$ (Proposition 1) with basis (x, x^2) and moments $m_0 = 1$, $m_1 = \mu$, $m_2 = \mu^2 + P$ gives:

$$\begin{pmatrix} 1 & 2m_1 \\ 2m_1 & 4m_2 \end{pmatrix} \begin{pmatrix} \lambda_1 \\ \lambda_2 \end{pmatrix} = \begin{pmatrix} 0 \\ 2 \end{pmatrix}.$$

Solving this yields

$$\lambda_1 = -\frac{m_1}{m_2 - m_1^2} = -\frac{\mu}{P} = -\eta, \qquad \lambda_2 = \frac{1}{2(m_2 - m_1^2)} = \frac{1}{2P} = \frac{\Omega}{2}. \tag{A.27}$$

Hence, score matching recovers the information vector η and the precision matrix Ω exactly from the first two moments.

D.4 Prediction step reduces to Riccati

For affine dynamics $dx = (Ax + b)\,dt + h\,dW$ with $H = \frac{1}{2}hh^\top$ constant, the score PDE (A.24) with $S = -\Omega$, $T = 0$, $J_X = A$, $\nabla(\nabla \cdot X) = 0$ reduces to $\partial_t s = -A^\top s + \Omega(Ax + b) - 2\Omega H s$. Substituting $s = -\Omega(x - \mu)$ and matching the coefficient of $(x - \mu)$ and the constant term:

$$\dot{\Omega} = -A^\top \Omega - \Omega A - 2\Omega H \Omega, \tag{A.28}$$

$$\dot{\mu} = A\mu + b. \tag{A.29}$$

Equation (A.28) is the **continuous-time information-form Riccati equation**. Using $\Omega = P^{-1}$ and $\dot{\Omega} = -P^{-1}\dot{P}P^{-1}$, one recovers the standard covariance Riccati $\dot{P} = AP + PA^\top + 2H$.

In terms of the information vector $\eta = \Omega\mu$:

$$\dot{\eta} = \dot{\Omega}\mu + \Omega\dot{\mu} = (-A^\top\Omega - \Omega A - 2\Omega H\Omega)\mu + \Omega(A\mu + b) = -(A^\top + 2\Omega H)\eta + \Omega b. \tag{A.30}$$

D.5 Update step recovers information-form Kalman

At measurement time with $z = Cx + v$, $v \sim \mathcal{N}(0, R)$, the likelihood score is $\nabla_x \log p(z|x) = C^\top R^{-1}(z - Cx) = C^\top R^{-1} z - C^\top R^{-1} C x$. The score addition $s^+ = s^- + \nabla_x \log p(z|x)$ gives:

$$\underbrace{-\Omega^+ x + \eta^+}_{s^+} = \underbrace{-\Omega^- x + \eta^-}_{s^-} + \underbrace{-C^\top R^{-1} C\, x + C^\top R^{-1} z}_{\nabla_x \log p(z|x)}.$$

Matching coefficients, we obtain

$$\boxed{\Omega^+ = \Omega^- + C^\top R^{-1} C, \qquad \eta^+ = \eta^- + C^\top R^{-1} z.} \tag{A.31}$$

These are **exactly the information-form Kalman filter update equations**. The posterior mean is $\mu^+ = (\Omega^+)^{-1}\eta^+$.

In the SKF framework, the same update is written as $\lambda^+ = \lambda^- + \lambda_{\text{lik}}$ (eq. (14)), where $\lambda_{\text{lik}} = (-C^\top R^{-1} z,\ \frac{1}{2} C^\top R^{-1} C)$ in the (λ_1, Λ_2) parameterization. The correspondence is exact, and the SKF conjugate update *is* the information-form Kalman update, expressed in natural parameters. We summarize the preceding calculations as follows.

Proposition 5 (Kalman filter recovery)**.** *Consider an affine-Gaussian system with dynamics $dx = (Ax + b)\,dt + h\,dW$, where $A \in \mathbb{R}^{n\times n}$, $b \in \mathbb{R}^n$, and $h \in \mathbb{R}^{n\times n_w}$ are constant, together with an affine measurement model $z = Cx + v$, $v \sim \mathcal{N}(0, R)$. Then Algorithm 1 at $r = 2$ produces the same precision matrix $\Omega(t)$ and information vector $\eta(t)$ as the continuous-discrete information-form Kalman filter.*

Proof. Each step is verified in the preceding subsections. Score matching recovers (η, Ω) from the first two moments (Section D.3), the score PDE reduces to the information-form Riccati equation (Section D.4), and the conjugate score update reproduces the information-form Kalman measurement update (Section D.5). □

The comparison between the SKF and the information-form Kalman filter is summarized in the table below.

SKF quantity	**Information-form Kalman**
Score $s(x) = -J_\phi \lambda$	$-\Omega(x - \mu) = -\Omega x + \eta$
Score Hessian $S = -\Omega$	Negative precision
Score matching solve	(Ω, η) from (m_1, m_2)
$\lambda^+ = \lambda^- + \lambda_{\text{lik}}$ (conjugate)	$\Omega^+ = \Omega^- + C^\top R^{-1} C$, $\eta^+ = \eta^- + C^\top R^{-1} z$
Score PDE (A.24)	Riccati ODE for Ω
Stein identity (trivial at $r = 2$)	Moment-covariance relation

E Iterative consistency refinement via Stein's method of moments

The conjugate measurement update $\lambda^+ = \lambda^- + \lambda_{\text{lik}}$ (Section 5) produces posterior natural parameters λ^+, but the subsequent Stein moment recovery (Section 5.2) yields only approximate moments $\hat{m}^+$ due to truncation. In general, $\hat{m}^+$ and λ^+ are not mutually consistent: re-fitting score matching to $\hat{m}^+$ produces $\hat{\lambda} \neq \lambda^+$. This section formalizes the iterative refinement that enforces consistency, drawing on recent work connecting score matching to Stein's method of moments (SMoM).

E.1 Background: Stein's method of moments

Stein's method [39] characterizes a distribution $p(x;\theta)$ through an operator $\mathcal{A}_\theta$ – called a *Stein operator* – that satisfies $\mathbb{E}_{p(\cdot;\theta)}[\mathcal{A}_\theta g(X)] = 0$ for a sufficiently rich class of test functions g. Ebner et al. [33] developed Stein's method of moments (SMoM), which derives parameter estimators by solving the empirical Stein equations for selected test functions. For the polynomial MED $p(x;\lambda) \propto \exp(-\lambda \cdot \phi(x))$, the divergence-based Stein operator is $\mathcal{A}_\lambda g(x) = \nabla \cdot g(x) - (\nabla E(x;\lambda)) \cdot g(x)$, where $E(x;\lambda) = \lambda \cdot \phi(x)$ is the energy. The condition $\mathbb{E}[\mathcal{A}_\lambda g] = 0$ reduces to the Stein identity (Proposition 2) via integration by parts.

An SMoM estimator solves the empirical Stein equations $\frac{1}{n}\sum_{i=1}^n \mathcal{A}_\theta g_k(X_i) = 0$ for selected test functions $\{g_k\}$. Different choices of g_k yield different estimators with different asymptotic variances.

E.2 Score matching as the center of SMoM

Kume and Walker [19] proved that the score matching estimator $\hat{\lambda}_{\text{SM}}$ is a special case of SMoM, corresponding to the choice $g_k(x) = \nabla\phi_k(x)$. Nagai and Yano [27] showed that SMoM estimators admit a canonical decomposition centered at score matching. In their setting,

$$\hat{\lambda}_{\text{SMoM}} = \hat{\lambda}_{\text{SM}} + \Delta(\hat{\lambda}_{\text{SM}}), \tag{A.32}$$

where Δ is an augmentation term that depends on additional Stein equations beyond the score matching ones. They use this decomposition to construct SMoM augmentations that can improve the asymptotic variance of score matching.

E.3 Application to the SKF measurement update

In the SKF, we do not estimate λ from samples. Instead, after the conjugate update $\lambda^+ = \lambda^- + \lambda_{\text{lik}}$, we need to find moments m^+ consistent with λ^+. The Stein identity with $\lambda = \lambda^+$ provides a linear system $\Lambda_1(\lambda^+)m^+ = c_1(\lambda^+)$ (equation (15)), but the truncated system is approximate.

The iterative refinement enforces a fixed-point condition on the (λ, m) pair. Define the map $\Phi\colon \lambda \mapsto m \mapsto \lambda'$:

(i) $m = \text{SteinRecover}(\lambda)$: solve the truncated Stein system (15) for the moments.
(ii) $\lambda_{\text{refit}} = \text{SM}(m)$: re-fit the score matching parameters to the recovered moments.
(iii) $\lambda' = \lambda_{\text{refit}} + \lambda_{\text{lik}}$: re-apply the measurement update.

A fixed point λ^* of the iteration $\lambda^{(k+1)} = \Phi(\lambda^{(k)})$ satisfies $\text{SM}(\text{SteinRecover}(\lambda^*)) + \lambda_{\text{lik}} = \lambda^*$, i.e., score matching applied to the moments recovered from λ^* returns $\lambda^* - \lambda_{\text{lik}}$.

In the language of [27], step (ii) projects back to the score matching solution (the "center" of the SMoM family), and step (iii) re-applies the measurement. The net effect is that each iteration reduces the inconsistency introduced by truncation in step (i).

Empirical convergence. We do not yet have a convergence proof for this iteration. Empirically, in the full-basis experiments ($n \leq 10$, $r \leq 8$), 2–3 iterations reduce the Stein recovery residual by 2–3 orders of magnitude (from $\sim 10^{-5}$ to $\sim 10^{-8}$ on the Duffing oscillator benchmark). In the active-closure coupled oscillator sweep through n=20, one refinement step was used for speed, and we have not observed divergence. The per-iteration cost is one reassembled score-matching solve ($O(M^3)$ dense, lower if a structured or regularized solve is used) and one Stein recovery solve. In the reported runs this refinement cost is smaller than the prediction step, but it is not included in the stand-alone score-matching $O(M^3)$ density-fit cost.

F Degree accounting and moment budget

This appendix records the degree bookkeeping used throughout the experiments. Suppose the drift has polynomial degree d_X and the diffusion coefficient has polynomial degree d_h. Applying the generator to a monomial of total degree k gives two possible sources of degree growth. The drift term differentiates once and then multiplies by X, so it reaches degree $k + d_X - 1$. The diffusion term differentiates twice and multiplies by $hh^\top$, so it reaches degree $k + 2d_h - 2$. Thus the moment equation for degree k can involve moments up to degree $k + \bar{d}$, where

$$\bar{d} = \max(d_X - 1,\, 2d_h - 2).$$

This number is the amount by which the moment hierarchy reaches beyond the moments currently being propagated. It therefore gives the number of Stein closure layers needed. The table gives the common cases used in the examples.

System type	d_X	d_h	$\bar{d}$	**Closure**
Linear + additive noise	1	0	0	Exact
Linear + linear mult. noise	1	1	0	Exact
Quadratic + additive	2	0	1	1 layer
Bilinear	2	0–1	1	1 layer
Cubic + additive	3	0	2	2 layers

Moment budget. Score matching (Proposition 1) requires moments up to degree $K = 2r-2$. This fixes the natural propagated budget for an order-r polynomial score model. During prediction, however, the ODE for the degree-K moments may request moments beyond this budget, up to degree $K + \bar{d}$. The Stein closure is applied one degree at a time: the first layer estimates degree $K + 1$ moments from moments of degree $\le K$ and the current score parameters, the next layer estimates degree $K + 2$ from the enlarged collection, and so on. When $\bar{d} = 0$, no higher moments are requested and the moment hierarchy closes exactly at the tracked degree. When $\bar{d} > 0$, the prediction step needs exactly $\bar{d}$ sequential closure layers.

G Generator calculations for the experimental systems

G.1 Worked example of the prediction step

We trace line 5 of Algorithm 1 for $n = 1$ and $r = 2$, so $K = 2r - 2 = 2$, $M = \binom{3}{1} = 3$, and the basis is $\{1, x, x^2\}$. Consider $dX = (X - X^2)\,dt + \sigma\,dW$. Its generator is $\mathcal{A}f = (x - x^2)f'(x) + \frac{1}{2}\sigma^2 f''(x)$, so with $f = x^2$ the moment ODE (4) gives

$$\dot{m}_2 = 2\,m_2 - 2\,m_3 + \sigma^2.$$

The quadratic drift gives $\bar{d} = 1$, and the right-hand side reaches m_3, one degree above $K = 2$. The Stein identity (12) with $\beta = 2$ reads $\lambda_1 m_2 + 2\lambda_2 m_3 = 2m_1$, so

$$m_3 = \frac{2m_1 - \lambda_1 m_2}{2\lambda_2},$$

with $\lambda(t)$ from the preceding score matching fit. At $r = 2$ that fit is the Gaussian with $\lambda_2 = 1/(2v)$ and $\lambda_1 = -\mu/v$, where $\mu = m_1$ and $v = m_2 - m_1^2$, and the expression returns $m_3 = \mu^3 + 3\mu v$, the Gaussian third moment. Substituting m_3 closes the ODE in (m_1, m_2), which line 5 integrates over Δt.

G.2 Duffing oscillator

SDE. State $x = (x_1, x_2)$.

$$\begin{aligned} dx_1 &= x_2\,dt, \\ dx_2 &= (-\delta x_2 - \alpha x_1 - \beta x_1^2)\,dt + \sigma\,dW, \end{aligned} \tag{A.33}$$

with parameters δ (damping), α (linear stiffness), β (quadratic stiffness), and σ (noise intensity).

Drift and diffusion classification.

$$X(x) = (x_2,\ -\delta x_2 - \alpha x_1 - \beta x_1^2)^\top,$$
$$h = (0,\ \sigma)^\top. \tag{A.34}$$

Since X has degree $d_X = 2$ (from the βx_1^2 term) and h is constant ($d_h = 0$), we have $\bar{d} = \max(1, -2) = 1$.

Generator. Applied to $\phi_\alpha = x_1^a x_2^b$:

$$\begin{aligned}\mathcal{A}\phi_\alpha &= \nabla\phi_\alpha \cdot X + \mathrm{Tr}(H\nabla^2\phi_\alpha) \\ &= a\, x_1^{a-1}x_2^b \cdot x_2 + b\, x_1^a x_2^{b-1} \cdot (-\delta x_2 - \alpha x_1 - \beta x_1^2) + \frac{\sigma^2}{2}\, b(b{-}1)\, x_1^a x_2^{b-2} \\ &= a\, x_1^{a-1}x_2^{b+1} - b\delta\, x_1^a x_2^b - b\alpha\, x_1^{a+1}x_2^{b-1} \\ &\quad - b\beta\, x_1^{a+2}x_2^{b-1} + \frac{\sigma^2}{2}\, b(b{-}1)\, x_1^a x_2^{b-2}.\end{aligned} \tag{A.35}$$

All terms have degree $|\alpha| = a + b$ except $x_1^{a+2}x_2^{b-1}$, which has degree $|\alpha| + 1$. This is the unclosed term that the Stein closure (Section 4) resolves.

G.3 Lotka-Volterra (stochastic predator-prey)

SDE. State $x = (x_1, x_2)$ (prey, predator).

$$\begin{aligned}dx_1 &= (\alpha x_1 - \beta x_1 x_2)\, dt + \sigma_1\, dW_1, \\ dx_2 &= (-\gamma x_2 + \delta x_1 x_2)\, dt + \sigma_2\, dW_2,\end{aligned} \tag{A.36}$$

Drift and diffusion classification. Bilinear drift $X(x) = (\alpha x_1 - \beta x_1 x_2,\ -\gamma x_2 + \delta x_1 x_2)^\top$, $d_X = 2$ (from $x_1 x_2$), $d_h = 0$, $\bar{d} = 1$.

Generator. Applied to $\phi_\alpha = x_1^a x_2^b$:

$$\begin{aligned}\mathcal{A}\phi_\alpha = {} & a\alpha\, x_1^a x_2^b - a\beta\, x_1^a x_2^{b+1} - b\gamma\, x_1^a x_2^b + b\delta\, x_1^{a+1}x_2^b \\ & + \frac{\sigma_1^2}{2}\, a(a{-}1)\, x_1^{a-2}x_2^b + \frac{\sigma_2^2}{2}\, b(b{-}1)\, x_1^a x_2^{b-2}.\end{aligned} \tag{A.37}$$

Taking expectations: $\frac{d}{dt} m_{(a,b)} = \mathbb{E}[\mathcal{A}\phi_\alpha]$. All terms yield moments of degree $|\alpha|$ except $x_1^a x_2^{b+1}$ and $x_1^{a+1}x_2^b$, which have degree $|\alpha| + 1$. These are the unclosed terms resolved by Stein closure.

Centered coordinates. In strictly mean-centered coordinates $z = x - \mu(t)$, the reference evolves as $\dot{\mu} = \mathbb{E}[X(x)]$, not generally as $X(\mu)$ for nonlinear drift. Thus the centered drift is $X(z{+}\mu) - \mathbb{E}[X(x)]$ and has zero expectation. For Lotka–Volterra this gives

$$\begin{aligned}dz_1 &= \big[(\alpha - \beta\mu_2)z_1 - \beta\mu_1 z_2 - \beta(z_1 z_2 - m^z_{(1,1)})\big]\, dt + \sigma_1\, dW_1, \\ dz_2 &= \big[\delta\mu_2 z_1 + (-\gamma + \delta\mu_1)z_2 + \delta(z_1 z_2 - m^z_{(1,1)})\big]\, dt + \sigma_2\, dW_2.\end{aligned} \tag{A.38}$$

The linear coefficients $(\alpha - \beta\mu_2)$ etc. depend on the current mean $\mu(t)$ and change at each time step. The purely quadratic terms $\beta z_1 z_2$ and $\delta z_1 z_2$ remain unclosed at $|\alpha| = K$ and are resolved by Stein closure.

G.4 Coupled Duffing oscillators

SDE. N oscillators with positions q_i and momenta p_i, $i = 1, \ldots, N$. State dimension $n = 2N$.

$$\begin{aligned}dq_i &= p_i\, dt, \\ dp_i &= \big[-\gamma p_i - \alpha q_i - \beta q_i^2 + \kappa(q_{i+1} {-} 2q_i {+} q_{i-1})\big]\, dt + \sigma\, dW_i,\end{aligned} \tag{A.39}$$

with free boundary conditions ($q_0 = q_1$, $q_{N+1} = q_N$).

Drift and diffusion classification.

$$X_i(x) = p_i, \quad i = 1, \dots, N \quad \text{(kinematics)},$$
$$X_{N+j}(x) = -\gamma p_j - \alpha q_j - \beta q_j^2 + \kappa(q_{j+1} - 2q_j + q_{j-1}),$$
$$h = \begin{bmatrix} 0_{N\times N} \\ \sigma I_N \end{bmatrix} \quad \text{(independent noise on momenta only)}. \tag{A.40}$$

Since X has degree $d_X = 2$ (from βq_j^2) and h is constant ($d_h = 0$), we have $\bar{d} = \max(1, -2) = 1$.

Diffusion tensor. $H = \frac{\sigma^2}{2}\,\mathrm{diag}(0, \dots, 0, 1, \dots, 1)$ (nonzero only on the momentum block).

Generator. Applied to $\phi_\alpha = \prod_{i=1}^N q_i^{a_i} p_i^{b_i}$, with contributions from each coordinate. For the kinematic dimensions ($dq_i = p_i\,dt$):

$$a_i \cdot q_i^{a_i - 1} p_i^{b_i + 1} \prod_{j \neq i} q_j^{a_j} p_j^{b_j} \quad \Longrightarrow \quad a_i\, m_{\alpha - e_{q_i} + e_{p_i}}. \tag{A.41}$$

For the momentum dimensions (dp_j), the drift terms give:

$$-\gamma\, b_j\, m_\alpha \;-\; \alpha_{\text{stiff}}\, b_j\, m_{\alpha - e_{p_j} + e_{q_j}} \;-\; \beta\, b_j\, m_{\alpha - e_{p_j} + 2e_{q_j}}$$
$$+\; \kappa\, b_j \big(m_{\alpha - e_{p_j} + e_{q_{j-1}}} + m_{\alpha - e_{p_j} + e_{q_{j+1}}} - 2\, m_{\alpha - e_{p_j} + e_{q_j}} \big), \tag{A.42}$$

and the diffusion gives $\frac{\sigma^2}{2} b_j (b_j - 1)\, m_{\alpha - 2e_{p_j}}$. The term $m_{\alpha - e_{p_j} + 2e_{q_j}}$ has degree $|\alpha| + 1$ and is unclosed at $|\alpha| = K$.

Centered coordinates. In strictly mean-centered coordinates $z = x - \mu(t)$ with $\dot{\mu} = \mathbb{E}[X(x)]$, the momentum equation for oscillator j becomes

$$dz_{N+j} = \big[-\gamma z_{N+j} - (\alpha + 2\beta\mu_j)\, z_j - \beta(z_j^2 - m^z_{2e_{q_j}})$$
$$+ \kappa(z_{j-1} + z_{j+1} - 2z_j) \big]\, dt + \sigma\, dW_j. \tag{A.43}$$

The centered linear coefficient for z_j is $-(\alpha + 2\beta\mu_j)$, not $-\alpha$: this linearization around μ is essential for numerical stability.

G.5 Double-well Langevin

SDE. State $x = (x_1, x_2)$. The gradient Langevin dynamics is $dX = -\nabla V(X)\, dt + \sigma\, dW$ with the separable quartic potential

$$V(x) = \frac{1}{4}x_1^4 + \frac{1}{4}x_2^4 - \frac{1}{2}x_1^2 - \frac{1}{2}x_2^2. \tag{A.44}$$

Drift and diffusion classification. The drift is $X(x) = -\nabla V = (x_1 - x_1^3,\ x_2 - x_2^3)^\top$, $h = \sigma I_2$. Since X has degree $d_X = 3$ (from x_i^3) and h is constant ($d_h = 0$), we have $\bar{d} = \max(d_X - 1,\ 2d_h - 2) = 2$. This is the highest excess degree among all experimental systems and requires *two* Stein closure layers.

Generator. Applied to $\phi_\alpha = x_1^a x_2^b$:

$$\mathcal{A}\phi_\alpha = a(x_1 - x_1^3)\, x_1^{a-1} x_2^b + b(x_2 - x_2^3)\, x_1^a x_2^{b-1}$$
$$+ \frac{\sigma^2}{2} \big[a(a-1)\, x_1^{a-2} x_2^b + b(b-1)\, x_1^a x_2^{b-2} \big]. \tag{A.45}$$

Taking expectations:

$$\dot{m}_\alpha = a\, m_\alpha - a\, m_{\alpha + 2e_1} + b\, m_\alpha - b\, m_{\alpha + 2e_2}$$
$$+ \frac{\sigma^2}{2} \big[a(a-1)\, m_{\alpha - 2e_1} + b(b-1)\, m_{\alpha - 2e_2} \big]. \tag{A.46}$$

The terms $m_{\alpha+2e_1}$ and $m_{\alpha+2e_2}$ have degree $|\alpha| + 2$, confirming $\bar{d} = 2$. At $|\alpha| = K = 2r - 2$, the first Stein closure layer resolves degree-$(K+1)$ moments, and a second layer resolves degree-$(K+2)$.

Because the potential is separable ($V = V_1(x_1) + V_2(x_2)$), the drift couples coordinates only through the shared moment vector. The independent noise ($H = \frac{\sigma^2}{2} I_2$) preserves this separability: if the initial density is a product $p_0(x) = p_1(x_1)\, p_2(x_2)$, then $p(t, x)$ remains a product for all t. In the experiments (Section 7), we use a near-product initial condition and propagate to $T = 3$ s, by which time the density has split into 4 modes at $(\pm 1, \pm 1)$.

G.6 3D tracer advection

SDE. State $X = (x_1, x_2, x_3) \in \mathbb{R}^3$, velocity field

$$u(X) = \begin{pmatrix} \varepsilon x_1 + \omega x_2 \\ -\omega x_1 + \varepsilon x_2 \\ -2\varepsilon x_3 + \alpha(x_1^2 + x_2^2) \end{pmatrix}, \qquad dX = u(X)\,dt + \sqrt{2\kappa}\,dW. \tag{A.47}$$

Drift classification. u_1 and u_2 are linear in X, while u_3 is quadratic via $\alpha(x_1^2 + x_2^2)$. Drift degree $d_X = 2$, noise additive ($d_h = 0$), $\bar{d} = d_X - 1 = 1$. One Stein closure layer. Incompressibility: $\nabla \cdot u = \varepsilon + \varepsilon - 2\varepsilon = 0$.

Generator. For $\phi(x) = x_1^a x_2^b x_3^c$:

$$\begin{aligned} \mathcal{A}\phi = {} & a(\varepsilon x_1 + \omega x_2) x_1^{a-1} x_2^b x_3^c + b(-\omega x_1 + \varepsilon x_2) x_1^a x_2^{b-1} x_3^c \\ & + c\big[-2\varepsilon x_3 + \alpha(x_1^2 + x_2^2)\big] x_1^a x_2^b x_3^{c-1} \\ & + \kappa\big[a(a-1) x_1^{a-2} x_2^b x_3^c + b(b-1) x_1^a x_2^{b-2} x_3^c + c(c-1) x_1^a x_2^b x_3^{c-2}\big]. \end{aligned} \tag{A.48}$$

Centered moment ODE. In $z = x - \mu(t)$ with $\dot{\mu} = u(\mu)$:

$$\begin{aligned} f_1^{\text{cen}} &= \varepsilon z_1 + \omega z_2, \\ f_2^{\text{cen}} &= -\omega z_1 + \varepsilon z_2, \\ f_3^{\text{cen}} &= -2\varepsilon z_3 + \alpha(z_1^2 + z_2^2 + 2\mu_1 z_1 + 2\mu_2 z_2). \end{aligned} \tag{A.49}$$

The z_1, z_2 equations are linear and close without Stein closure. Only the z_3 equation has quadratic terms (αz_1^2, αz_2^2) that couple to the closure at degree K.

G.7 SE(2) kinematics

SDE. State $x = (c, s, p_x, p_y)$ with $c = \cos\theta$, $s = \sin\theta$, $c^2 + s^2 = 1$. The SDE in embedded coordinates is

$$\begin{aligned} dc &= \Big(-\omega s - \frac{\sigma^2}{2} c\Big) dt - \sigma s\, dW, \\ ds &= \Big(\omega c - \frac{\sigma^2}{2} s\Big) dt + \sigma c\, dW, \\ dp_x &= v c\, dt, \quad dp_y = v s\, dt, \end{aligned} \tag{A.50}$$

where ω is the angular velocity, v is the forward speed, and σ is the heading noise intensity.

Drift and diffusion classification. Both are linear in x: $d_X = 1$, $d_h = 1$, $\bar{d} = 0$.

Diffusion tensor.

$$H = \frac{\sigma^2}{2} \begin{pmatrix} s^2 & -cs & 0 & 0 \\ -cs & c^2 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}. \tag{A.51}$$

Generator. For $\phi_\alpha = c^{a_1} s^{a_2} p_x^{a_3} p_y^{a_4}$, the generator $\mathcal{L}\phi_\alpha = \sum_i X_i \partial_i \phi_\alpha + \sum_{ij} H_{ij} \partial_i \partial_j \phi_\alpha$ gives the moment ODE $\dot{m}_\alpha = \mathbb{E}[\mathcal{L}\phi_\alpha]$:

$$\begin{aligned} \dot{m}_\alpha = {} & \omega\big(a_2\, m_{\alpha + e_c - e_s} - a_1\, m_{\alpha - e_c + e_s}\big) \\ & - \frac{\sigma^2}{2}(a_1 + a_2)\, m_\alpha + v\big(a_3\, m_{\alpha + e_c - e_{p_x}} + a_4\, m_{\alpha + e_s - e_{p_y}}\big) \\ & + \frac{\sigma^2}{2}\big[a_1(a_1 - 1)\, m_{\alpha - 2e_c + 2e_s} + a_2(a_2 - 1)\, m_{\alpha + 2e_c - 2e_s}\big] \\ & - \sigma^2 a_1 a_2\, m_\alpha. \end{aligned} \tag{A.52}$$

The first line is the rotation coupling (drift of c and s), the second is the Itô correction plus the translation coupling, and the last two lines are the state-dependent diffusion. Every moment on the right-hand side has degree $|\alpha|$, confirming $\bar{d} = 0$.

On the constraint manifold $c^2+s^2=1$, the diffusion terms simplify via $m_{\alpha-2e_c+2e_s} = m_{\alpha-2e_c} - m_\alpha$ and $m_{\alpha+2e_c-2e_s} = m_{\alpha-2e_s} - m_\alpha$ (replacing $s^2 \to 1-c^2$ and $c^2 \to 1-s^2$ respectively), reducing to lower-degree moments plus diagonal corrections to m_α. The resulting ODE is linear: $\dot{m}(t) = L\, m(t)$, so $m(T) = e^{LT} m(0)$.

Fourier–Legendre basis. A natural choice of orthogonal basis for $S^1 \times \mathbb{R}^2$ replaces the embedded monomials with: Fourier modes $\cos(m\theta)$, $\sin(m\theta)$ for the heading ($m = 0, \dots, K_\theta$), and centered Legendre polynomials $\widetilde{P}_j(z_{p_x})\, \widetilde{P}_l(z_{p_y})$ for the position. For the propagation formula below, write the raw Fourier–position moments $\mathrm{mc}_{m,j,l} := \mathbb{E}[\cos(m\theta)\, p_x^j\, p_y^l]$ and $\mathrm{ms}_{m,j,l}$ similarly with sin. The SE(2) kinematics give:

$$\begin{aligned}\frac{d}{dt}\mathrm{mc}_{m,j,l} = &-m\omega\, \mathrm{ms}_{m,j,l} - \frac{m^2\sigma^2}{2}\, \mathrm{mc}_{m,j,l} \\ &+ \frac{vj}{2}\big(\mathrm{mc}_{m-1,j-1,l} + \mathrm{mc}_{m+1,j-1,l}\big) \\ &+ \frac{vl}{2}\big(\mathrm{ms}_{m+1,j,l-1} - \mathrm{ms}_{m-1,j,l-1}\big),\end{aligned} \tag{A.53}$$

using $\cos\theta\cos(m\theta) = \frac{1}{2}(\cos((m-1)\theta) + \cos((m+1)\theta))$. The coupling matrix L is *tridiagonal in* m: the position drift $v\cos\theta$ couples Fourier mode m to $m \pm 1$. With $K_\theta = 6$ and position degree 4, the matrix is $<1\%$ nonzero, and the solution is $m(T) = e^{LT} m(0)$. In the Fourier–Legendre implementation, the same operator is conjugated into the Legendre position basis using the change-of-basis formulas in Appendix H.3. Position derivatives are then evaluated with the Legendre recurrence below.

G.7.1 Stein identity on $S^1 \times \mathbb{R}^2$

For the measurement update on SE(2), we need to recover posterior moments from posterior score parameters in the Fourier–Legendre basis. The Stein identity on $S^1 \times \mathbb{R}^2$ takes the form: for a test function $f(\theta, p_x, p_y)$ and a density $p(\theta, p_x, p_y) \propto \exp(-E(\theta, p_x, p_y))$,

$$\mathbb{E}\Big[\frac{\partial^2 f}{\partial\theta^2} + \frac{\partial^2 f}{\partial p_x^2} + \frac{\partial^2 f}{\partial p_y^2} - \frac{\partial E}{\partial\theta}\frac{\partial f}{\partial\theta} - \frac{\partial E}{\partial p_x}\frac{\partial f}{\partial p_x} - \frac{\partial E}{\partial p_y}\frac{\partial f}{\partial p_y}\Big] = 0, \tag{A.54}$$

where the angular part uses the standard Laplace–Beltrami operator $\Delta_{S^1} = \partial^2/\partial\theta^2$ on the circle.

Fourier–Legendre specialization. Let the energy be $E(\theta, p_x, p_y) = \sum_\alpha \lambda_\alpha\, \psi_\alpha(\theta, p_x, p_y)$, where ψ_α are Fourier–Legendre basis functions $\cos(m\theta)\, P_j(z_{p_x})\, P_l(z_{p_y})$ or $\sin(m\theta)\, P_j(z_{p_x})\, P_l(z_{p_y})$. The derivatives of basis functions are:

$$\begin{aligned}&\frac{\partial}{\partial\theta}\cos(m\theta) = -m\sin(m\theta), && \frac{\partial}{\partial\theta}\sin(m\theta) = m\cos(m\theta), \\ &\frac{\partial^2}{\partial\theta^2}\cos(m\theta) = -m^2\cos(m\theta), && \frac{\partial^2}{\partial\theta^2}\sin(m\theta) = -m^2\sin(m\theta), \\ &\frac{\partial}{\partial p_x}P_j(z_{p_x}) = \frac{1}{\sigma_{p_x}}P_j'(z_{p_x}), && \frac{\partial^2}{\partial p_x^2}P_j(z_{p_x}) = \frac{1}{\sigma_{p_x}^2}P_j''(z_{p_x}),\end{aligned} \tag{A.55}$$

where $z_{p_x} = (p_x - \mu_{p_x})/\sigma_{p_x}$ and P_j', P_j'' are Legendre derivative coefficients expressible in the Legendre basis via the recurrence $P_n'(x) = \sum_{k<n,\, n-k \text{ odd}} (2k+1)\, P_k(x)$.

Substituting a Fourier–Legendre test function ψ_β into (A.54) and using the product linearization $\cos(m\theta)\cos(m'\theta) = \frac{1}{2}[\cos((m-m')\theta) + \cos((m+m')\theta)]$ and $P_j P_{j'} = \sum_k g_{jj'k}\, P_k$ (Clebsch–Gordan), each Stein equation becomes a linear constraint on the Fourier–position moments $\mathrm{mc}_{m,j,l}$, $\mathrm{ms}_{m,j,l}$. The full system has the structure $\Lambda_1(\lambda)\,\mu = c_1(\lambda)$ where μ is the vector of unknown posterior moments, exactly as in the Euclidean case (Section 5.2).

Key advantage over the embedded basis. the Fourier basis diagonalizes the S^1 Laplacian ($\Delta_{S^1}\cos(m\theta) = -m^2\cos(m\theta)$), so no constraint $c^2 + s^2 = 1$ appears in the system. The Stein matrix Λ_1 is well-conditioned in this basis.

Observation likelihood. For a landmark $L \in \mathbb{R}^2$ observed as $y = R(\theta)^\top(L + v - p)$ with non-Gaussian noise v (e.g., the 4-modal distribution of [13]), the log-likelihood $\log p(y \mid \theta, p)$ is not

polynomial in (θ, p), but it can be projected onto the Fourier–Legendre basis by regression: evaluate $\log p(y \mid \theta_i, p_i)$ at random samples (θ_i, p_i) and solve for λ_{lik} via least squares. The conjugate update $\lambda^+ = \lambda^- + \lambda_{\text{lik}}$ then proceeds as in Section 5, followed by Stein moment recovery and iterative consistency refinement (Appendix E).

G.8 SO(3) quaternion kinematics

State $q = (q_0, q_1, q_2, q_3) \in S^3 \subset \mathbb{R}^4$ with $|q|^2 = 1$. The unit quaternion represents a rotation $R \in \mathrm{SO}(3)$ via the double cover $S^3 \to \mathrm{SO}(3)$ (both q and $-q$ map to the same rotation R).

SDE. The rotation kinematics SDE in Itô form [41] is

$$dq = \underbrace{\Big(\frac{1}{2}Q(q)\omega_0 - \frac{3\sigma^2}{8}q\Big)}_{X(q)}\,dt + \underbrace{\frac{\sigma}{2}Q(q)}_{h(q)}\,dW, \tag{A.56}$$

where $\omega_0 \in \mathbb{R}^3$ is the nominal angular velocity, $\sigma > 0$ is the noise intensity, $W \in \mathbb{R}^3$ is a standard Wiener process, and

$$Q(q) = \begin{pmatrix} -q_1 & -q_2 & -q_3 \\ q_0 & -q_3 & q_2 \\ q_3 & q_0 & -q_1 \\ -q_2 & q_1 & q_0 \end{pmatrix} \in \mathbb{R}^{4\times 3} \tag{A.57}$$

is the quaternion multiplication matrix satisfying $Q(q)^\top Q(q) = |q|^2 I_3$ and $Q(q)Q(q)^\top = |q|^2 I_4 - qq^\top$. The Itô correction $-\frac{3\sigma^2}{8}q$ ensures $\frac{d}{dt}\mathbb{E}[|q|^2] = 0$, preserving the unit norm constraint.

Drift and diffusion classification. Both $X(q)$ and $h(q)$ are *linear* in q, giving $d_X = 1$ and $d_h = 1$. The excess degree is $\bar{d} = \max(d_X - 1,\ 2d_h - 2) = \max(0, 0) = 0$: the moment hierarchy closes exactly.

Diffusion tensor.

$$H(q) = \frac{1}{2}h(q)\,h(q)^\top = \frac{\sigma^2}{8}\big(|q|^2 I_4 - qq^\top\big). \tag{A.58}$$

On S^3 ($|q|^2 = 1$), $H = \frac{\sigma^2}{8}(I_4 - qq^\top)$ is the projection onto the tangent space of S^3 at q, scaled by $\sigma^2/8$.

Wigner D-matrix basis. The monomial basis in $(q_0, \ldots, q_3)$ fails on S^3 ($\kappa(A) \sim 10^{14}$) because the constraint $|q|^2 = 1$ introduces $\binom{4+r}{4} - (r+1)^2$ algebraic dependencies among the monomials at each degree. The Wigner D-matrices $\{D^l_{mm'}\}_{l=0}^{l_{\max}}$, $m, m' \in \{-l, \ldots, l\}$, form a complete orthonormal basis for $L^2(\mathrm{SO}(3))$ by the Peter–Weyl theorem [42, 43]. At $l_{\max} = 4$: $M = \sum_{l=0}^{4}(2l+1)^2 = 165$ real basis functions. Each $D^l_{mm'}$ is a polynomial of degree $2l$ in the quaternion components. Functions on SO(3) correspond to *even* polynomials in q (invariant under $q \mapsto -q$).

Generator in the Wigner basis. Define the Wigner moments $\hat{P}^l_{mm'}(t) := \mathbb{E}[D^l_{mm'}(R_t)]$. The infinitesimal generator of the SDE (A.56), when applied to $D^l_{mm'}$, yields:

- **Deterministic rotation.** $\mathcal{A}_{\text{rot}} D^l_{mm'} = -i[\omega_0 \cdot \mathbf{J}^{(l)}]_{mn}\, D^l_{nm'}$, where $\mathbf{J}^{(l)} = (\mathbf{J}^{(l)}_x, \mathbf{J}^{(l)}_y, \mathbf{J}^{(l)}_z)$ are the $(2l{+}1) \times (2l{+}1)$ angular momentum matrices in the spin-l representation.
- **Isotropic diffusion.** $\mathcal{A}_{\text{diff}} D^l_{mm'} = -\frac{\sigma^2}{2} l(l{+}1)\, D^l_{mm'}$, using the Laplace–Beltrami eigenvalue $\Delta_{\mathrm{SO}(3)} D^l_{mm'} = -l(l{+}1) D^l_{mm'}$.

The moment ODE is therefore *block-diagonal in* l:

$$\frac{d}{dt}\hat{P}^l = -\Big(\frac{\sigma^2}{2} l(l{+}1)\, I_{(2l+1)^2} + i\,\omega_0 \cdot \mathbf{J}^{(l)}\Big)\hat{P}^l, \tag{A.59}$$

where $\hat{P}^l \in \mathbb{C}^{(2l+1)\times(2l+1)}$ and the angular momentum acts by left multiplication on the row index m. The solution is $\hat{P}^l(T) = e^{L_l T}\hat{P}^l(0)$, with each block L_l of size $(2l{+}1)^2 \times (2l{+}1)^2$. No coupling between different l-blocks occurs.

Riemannian score matching. The polynomial MED on SO(3) takes the form $p(R;\lambda) \propto \exp\big(-\sum_{l,m,m'} \lambda^l_{mm'} D^l_{mm'}(R)\big)$. Using the Hyvärinen objective with the Laplace–Beltrami operator on SO(3) [43], we assemble:

$$b_k = -l_k(l_k{+}1)\,\hat{p}^{l_k}_{m_k m'_k} \quad \text{(diagonal in the Wigner basis)}, \tag{A.60}$$

$$A_{kj} = \mathbb{E}[\nabla_{\mathrm{SO}(3)}\phi_k \cdot \nabla_{\mathrm{SO}(3)}\phi_j]. \tag{A.61}$$

The b-vector requires only the propagated Wigner moments because $\Delta_{\mathrm{SO}(3)} D^l_{mm'} = -l(l{+}1) D^l_{mm'}$. The A-matrix is computed from products of differentiated Wigner functions. Equivalently, the left-invariant vector fields act through angular-momentum matrices, and products such as $D^{l_1}_{m_1 m'_1} D^{l_2}_{m_2 m'_2}$ decompose via Clebsch–Gordan coefficients [42] into $D^L_{MM'}$ with $|l_1{-}l_2| \le L \le l_1{+}l_2$, giving A a banded structure in l (bandwidth $2l_{\max}$). We do not use the simplified integration-by-parts identity $\mathbb{E}_p[\nabla\phi_k \cdot \nabla\phi_j] = -\mathbb{E}_p[\phi_j \Delta\phi_k]$, which would only hold without the extra $\nabla \log p$ term under the uniform Haar measure.

G.9 SE(3) rigid body kinematics

SDE. State $x = (q, p) \in S^3 \times \mathbb{R}^3$ with $q = (q_0, q_1, q_2, q_3)$ a unit quaternion ($|q|^2 = 1$) and $p = (p_1, p_2, p_3)$ the world-frame position. The manifold dimension is $n = 6$ (embedded in $\mathbb{R}^7$ via the quaternion constraint $|q|^2 = 1$). The Itô SDE [41] for the rotation is the same as the SO(3) system (A.56):

$$dq = \Big(\frac{1}{2} Q(q)\omega_0 - \frac{3\sigma_\omega^2}{8} q\Big)\,dt + \frac{\sigma_\omega}{2} Q(q)\,dW_\omega, \tag{A.62}$$

where $\omega_0 \in \mathbb{R}^3$ is the nominal angular velocity and σ_ω is the rotation noise intensity. The position evolves by the body-frame velocity $v_0 \in \mathbb{R}^3$ rotated into the world frame:

$$dp = R(q)\,v_0\,dt + \sigma_v\,dW_v, \tag{A.63}$$

where $R(q) \in \mathrm{SO}(3)$ is the rotation matrix corresponding to q and $\sigma_v\,dW_v$ is additive translational noise.

Drift and diffusion classification. The rotation drift is linear in q ($d_X = 1$ for the q-block). Each entry of $R(q)$ is *quadratic* in (q_0, q_1, q_2, q_3), so the position drift $R(q)v_0$ has $d_X = 2$. The rotation diffusion $Q(q)$ is linear ($d_h = 1$) and the position diffusion is constant ($d_h = 0$). Overall: $\bar{d} = \max(d_X - 1,\ 2d_h - 2) = \max(1, 0) = 1$. One Stein closure layer is needed.

Generator. For $\phi_\alpha = q_0^{a_0} q_1^{a_1} q_2^{a_2} q_3^{a_3} p_1^{b_1} p_2^{b_2} p_3^{b_3}$ with $|\alpha| = \sum a_i + \sum b_j$, the generator $\mathcal{A}\phi_\alpha = \nabla\phi \cdot X + \mathrm{Tr}(H\nabla^2\phi)$ has three groups of terms:

(i) Rotation block (same as SO(3)): the drift $\frac{1}{2}Q(q)\omega_0 - \frac{3\sigma_\omega^2}{8}q$ and diffusion $H_q = \frac{\sigma_\omega^2}{8}(I_4 - qq^\top)$ contribute terms involving only q-moments. These preserve the degree of q^a (i.e., $\bar{d}_q = 0$) by the same mechanism as in Appendix G.8.

(ii) Position drift ($R(q)v_0$):

$$\sum_{j=1}^{3} b_j \Big(\sum_{k=1}^{3} R_{jk}(q)\,v_{0,k}\Big) q^a\,p^{b-e_j}. \tag{A.64}$$

Since $R_{jk}(q)$ is degree 2 in q, this produces moments of degree $|a| + 2 + |b| - 1 = |\alpha| + 1$. These are the **unclosed terms** resolved by Stein closure.

(iii) Position diffusion ($\sigma_v^2/2 \cdot \nabla_p^2\phi$): $\frac{\sigma_v^2}{2}\sum_j b_j(b_j{-}1)\,m_{\alpha - 2e_{p_j}}$, which has degree $|\alpha| - 2$ (closed).

Rotation matrix entries. On $|q|^2 = 1$:

$$\begin{aligned} R_{11} &= 2(q_0^2 + q_1^2) - 1, & R_{12} &= 2(q_1 q_2 - q_0 q_3), \\ R_{21} &= 2(q_1 q_2 + q_0 q_3), & R_{22} &= 2(q_0^2 + q_2^2) - 1, \\ R_{31} &= 2(q_1 q_3 - q_0 q_2), & R_{33} &= 2(q_0^2 + q_3^2) - 1, \end{aligned} \tag{A.65}$$

with $R_{13} = 2(q_1 q_3 + q_0 q_2)$, $R_{23} = 2(q_2 q_3 - q_0 q_1)$, $R_{32} = 2(q_2 q_3 + q_0 q_1)$. Each R_{jk} contributes a degree-2 monomial in q to the position drift, confirming $\bar{d} = 1$.

Wigner–Legendre product basis. A natural choice of orthogonal basis for $\mathrm{SO}(3) \times \mathbb{R}^3$ is

$$\phi^{l,\alpha}_{mm'}(R,p) = D^l_{mm'}(R)\ \widetilde{P}_\alpha(p-\bar{p}), \tag{A.66}$$

where $D^l_{mm'}$ are Wigner D-matrices ($l \le l_{\max}$) and $\widetilde{P}_\alpha$ are centered Legendre polynomials on $\mathbb{R}^3$ ($|\alpha| \le r_p$). At $l_{\max}=2$, $r_p=4$: $M = 35\times 35 = 1225$. Define the product moments $\hat{m}^l_{mm',\alpha}(t) := \mathbb{E}[D^l_{mm'}(R_t)\,\widetilde{P}_\alpha(p_t-\bar{p})]$.

The generator $\mathcal{A}\phi^{l,\alpha}_{mm'}$ decomposes into three terms:

(i) Rotation block: $\mathcal{A}_{\mathrm{rot}}\phi^{l,\alpha}_{mm'} = (-\frac{\sigma_\omega^2}{2} l(l{+}1)\,\delta_{mn} - i[\omega_0{\cdot}\mathbf{J}^{(l)}]_{mn})\,D^l_{nm'}\,\widetilde{P}_\alpha$. Maps (l,α) to (l,α). Always closed.

(ii) Position diffusion: $\frac{\sigma_v^2}{2}\Delta_p\phi^{l,\alpha}_{mm'} = \frac{\sigma_v^2}{2} D^l_{mm'}\sum_j \alpha_j(\alpha_j{-}1)\widetilde{P}_{\alpha-2e_j}$. Maps $(l,|\alpha|)$ to $(l,|\alpha|{-}2)$. Always closed.

(iii) Position drift (rotation–translation coupling): $(Rv_0)\cdot\nabla_p\,\phi^{l,\alpha}_{mm'} = \sum_j \alpha_j(\sum_k R_{jk}v_{0,k})\,D^l_{mm'}\,\widetilde{P}_{\alpha-e_j}$. Each R_{jk} is a real combination of D^1 functions, so $R_{jk}D^l_{mm'}$ decomposes via the Clebsch–Gordan series:

$$D^1_{m_1m_1'}D^l_{mm'} = \sum_{L=|l-1|}^{l+1} \langle 1\,m_1; l\,m|L\,M\rangle\,\langle 1\,m_1'; l\,m'|L\,M'\rangle\,D^L_{MM'}. \tag{A.67}$$

This maps $(l,|\alpha|) \to (l{-}1,|\alpha|{-}1)\oplus(l,|\alpha|{-}1)\oplus(l{+}1,|\alpha|{-}1)$.

Closure. Only term (iii) is unclosed, producing Wigner level $l_{\max}{+}1$ when $l = l_{\max}$. One Stein closure layer resolves these, with the CG coefficients playing the role of the monomial product rule. The moment ODE matrix L is *tridiagonal in* l (each Wigner level couples only to $l \pm 1$), the direct SO(3) analog of the tridiagonal-in-m Fourier coupling (A.53) for SE(2).

H Concrete matrix forms

H.1 Score matching matrix A and vector b

For the polynomial MED $p(x;\lambda) \propto \exp(-\lambda\cdot\phi(x))$ with monomial basis $\phi_\alpha(x) = x^\alpha$, $|\alpha| \le r$, the score matching normal equations $A\lambda = b$ have entries:

$$A_{\alpha\beta} = \sum_{i=1}^n \alpha_i\beta_i\ m_{\alpha+\beta-2e_i}, \tag{A.68}$$

$$b_\alpha = \sum_{i=1}^n \alpha_i(\alpha_i-1)\ m_{\alpha-2e_i}, \tag{A.69}$$

where α,β index non-constant basis functions with $1 \le |\alpha|,|\beta| \le r$, and $m_\gamma = \mathbb{E}[x^\gamma]$ are the propagated moments. After eliminating the constant term ($|\alpha|=0$), A is $(M{-}1)\times(M{-}1)$ with $M = \binom{n+r}{n}$. All entries are moments of degree $\le 2r-2$, which is exactly the degree propagated by Dynkin's formula.

H.2 Stein closure matrix Λ_1

To resolve degree-$(K{+}1)$ moments, the Stein identity (12) at $|\beta| = r$ gives the system $\Lambda_1(\lambda)\,m^{(K+1)} = c(\lambda, m^{(\le K)})$ where:

$$[\Lambda_1]_{(\beta,i),\gamma} = \lambda_{\gamma-\beta+e_i}\ (\gamma-\beta+e_i)_i \quad \text{if } |\gamma-\beta+e_i| \le r, \tag{A.70}$$

with rows indexed by (β,i) with $|\beta|=r$, $\beta_i \ge 1$, and columns indexed by degree-$(K{+}1)$ multi-indices γ. For $n=2$, the first-layer system is $2r\times 2r$ (square). For general n and $r \ge 3$, the augmented system using $|\beta| \in [r,K]$ has a dimension-dependent crossover. For $r{=}3$, it is overdetermined through $n{=}15$ and underdetermined starting at $n{=}16$. Appendix B.5 gives the full row/column count, and Proposition 4 gives the $n{=}2$ well-posedness result. For structured generators, the same matrix construction may be restricted to the active columns requested by the moment ODE.

H.3 Change of basis

General framework. Let $\{\phi_\alpha\}$ (e.g., monomials) and $\{\psi_\alpha\}$ (e.g., product Legendre) be two polynomial bases of degree $\leq r$, related by a nonsingular matrix C:

$$\psi_\alpha(x) = \sum_\beta C_{\alpha\beta}\,\phi_\beta(x), \qquad \phi_\alpha(x) = \sum_\beta C^{-1}_{\alpha\beta}\,\psi_\beta(x). \tag{A.71}$$

For product bases, C has Kronecker structure: $C_{\alpha\beta} = \prod_{i=1}^n C^{(i)}_{\alpha_i,\beta_i}$, where $C^{(i)}_{k,j}$ is the j-th monomial coefficient of the k-th basis function in dimension i.

Natural parameters. The energy function $E(x) = \sum_\alpha \lambda^\phi_\alpha \phi_\alpha(x) = \sum_\alpha \lambda^\psi_\alpha \psi_\alpha(x)$ transforms as

$$\lambda^\phi = C^\top \lambda^\psi, \qquad \lambda^\psi = C^{-\top}\lambda^\phi. \tag{A.72}$$

Moments. The generalized moments $\mu^\psi_\alpha = \mathbb{E}[\psi_\alpha(x)]$ and $\mu^\phi_\alpha = \mathbb{E}[\phi_\alpha(x)]$ transform as

$$\mu^\psi = C\,\mu^\phi, \qquad \mu^\phi = C^{-1}\,\mu^\psi. \tag{A.73}$$

Score matching matrices. The score $s(x) = -\sum_\alpha \lambda_\alpha \nabla\phi_\alpha(x)$ is basis-invariant. Since $\nabla\psi = C\nabla\phi$, equivalently $\nabla\phi = C^{-1}\nabla\psi$, the normal equations $A^\psi\lambda^\psi = b^\psi$ in the new basis have

$$A^\psi = D^{-1}A^\phi D^{-\top}, \qquad b^\psi = D^{-1}b^\phi, \tag{A.74}$$

where D is the matrix such that $\nabla\phi = D\nabla\psi$. For a linear change of polynomial basis, $D = C^{-1}$. Derivative recurrences are still useful for assembling Legendre-basis matrices directly, but they do not change the parameter-space transformation above.

Conditioning. The key property of orthogonal bases (e.g., Legendre on $[-1,1]$) is that A^ψ is better conditioned than A^ϕ when the data distribution is supported near $[-1,1]^n$. For Legendre, $\mathbb{E}[\psi_\alpha\psi_\beta] \approx \delta_{\alpha\beta}\prod_i(2\alpha_i+1)^{-1}$ (orthogonality under a nearly uniform product measure), so A^ψ is approximately diagonal, giving $\kappa(A^\psi) \sim O(1)$–$O(10^2)$ compared to $\kappa(A^\phi) \sim O(10^4)$–$O(10^6)$ for monomials. This conditioning improvement propagates to the Stein closure and moment recovery systems, since their coefficients are products of λ entries and moment entries.

Stein identity under basis change. The Stein identity $\mathbb{E}[s_i f + \partial_i f] = 0$ is basis-independent. When expressed in basis ψ:

$$\sum_\alpha \lambda^\psi_\alpha\,\mathbb{E}[\partial_i\psi_\alpha\cdot\psi_\beta] = \mathbb{E}[\partial_i\psi_\beta]. \tag{A.75}$$

The products $\partial_i\psi_\alpha\cdot\psi_\beta$ and $\partial_i\psi_\beta$ are polynomials that can be re-expanded in the ψ basis using the linearization coefficients $g^\psi_{\alpha\beta\gamma}$ satisfying $\psi_\alpha\psi_\beta = \sum_\gamma g^\psi_{\alpha\beta\gamma}\,\psi_\gamma$. For Legendre, these are the Adams–Neumann coefficients, computable via Gauss–Legendre quadrature.

Practical workflow. In our implementation:

1. Moment propagation via Dynkin operates on monomial moments (the generator acts naturally on monomials).
2. Before score matching: convert monomial moments to the chosen basis using (A.73), or equivalently assemble A, b directly via (A.68)–(A.69) in centered/scaled monomial coordinates (which approximate orthogonality).
3. After score matching: λ is in the working basis. For Stein closure during propagation, convert to monomial λ via (A.72) if needed.
4. For measurement update: the conjugate addition $\lambda^+ = \lambda^- + \lambda_{\text{lik}}$ is performed in whichever basis both are expressed in.

H.4 Centered coordinate transformation

Given raw moments $m_\alpha = \mathbb{E}[x^\alpha]$ and mean $\mu_i = m_{e_i}$, the centered moments $m^z_\alpha = \mathbb{E}[(x-\mu)^\alpha]$ are obtained via the multinomial expansion:

$$m^z_\alpha = \sum_{\beta\leq\alpha}\binom{\alpha}{\beta}(-\mu)^{\alpha-\beta}\,m_\beta, \tag{A.76}$$

where $\binom{\alpha}{\beta} = \prod_i \binom{\alpha_i}{\beta_i}$ and $\mu^{\alpha-\beta} = \prod_i \mu_i^{\alpha_i-\beta_i}$. The inverse (uncentering) replaces $(-\mu)$ with $(+\mu)$.

Working in centered coordinates is essential for conditioning: after centering, $m^z_{e_i} = 0$ and m^z_α is a central moment of order $|\alpha|$, which is $O(\sigma^{|\alpha|})$ for a distribution with standard deviation σ. This prevents the score matching matrix A from being dominated by mean-related terms.

Scaling. Further conditioning improvement is obtained by normalizing to $w = z/s$ where $s_i = c \cdot \sigma_i$ (typically $c = 3.5$ maps $\pm 3.5\sigma$ to $[-1, 1]$). The scaled moments $m^w_\alpha = m^z_\alpha / \prod_i s_i^{\alpha_i}$ are $O(1)$ for all degrees, giving $\|\lambda\| = O(10)$–$O(100)$ instead of $O(10^4)$–$O(10^6)$ with raw monomials.

I Moment closure: continuous-time vs. discrete-time

The MEM-KF system class. MEM-KF [13] operates on discrete-time polynomial systems $X_{k+1} = f(X_k, U_k, W_k)$ where f is polynomial and W_k is *state-independent* noise. Their "extended dynamical system" (Eq. 6 in [13]) lifts the dynamics into the monomial basis $\phi_r(X_{k+1}) = A(W_k, U_k)\, \phi_r(X_k)$, and the moment propagation factors as $\mathbb{E}[\phi_r(X_{k+1})] = \mathbb{E}[A(W_k, U_k)]\, \mathbb{E}[\phi_r(X_k)]$ *only because* W_k *is independent of* X_k. This is a restricted class from the perspective of stochastic systems theory:

- **No multiplicative/state-dependent noise.** Continuous-time SDEs generically have state-dependent diffusion $dX = g(X)\, dt + h(X)\, dW$ where $h(X) \neq$ const. This arises in rigid body kinematics ($h(q) = \frac{\sigma}{2} Q(q)$ for SE(2), SO(3), SE(3)), geometric Brownian motion ($h = \sigma X$), and aerodynamic systems with quadratic drag under turbulence. State-dependent noise contributes to the excess degree via $\bar{d} = \max(d_X - 1,\ 2d_h - 2)$ and produces an Itô correction in the drift. These effects have no analog in their additive discrete-time framework $X_{k+1} = f(X_k) + W_k$.
- **No continuous-time structure.** The infinitesimal generator $\mathcal{A}$, Dynkin's formula [40], and the Fokker–Planck PDE provide exact moment evolution equations for continuous-time systems. Discrete-time polynomial maps do not expose this structure, replacing it with a single-step algebraic relation.
- **Discretization error for non-group systems.** Systems with native Lie group structure (e.g., SE(2), SO(3)) admit exact discrete composition maps $X_{k+1} = X_k \cdot U_k \cdot W_k$ that preserve the group and introduce no time-stepping error. For general continuous-time SDEs that lack this structure, however, converting to the discrete polynomial form via Euler–Maruyama introduces $O(\Delta t)$ weak error per step, and higher-order schemes (Milstein, stochastic Runge–Kutta) produce more complex polynomial structure in the noise coupling.

The SE(2) system in [13] has the special structure $X_{k+1} = X_k \cdot U_k \cdot W_k$ (Lie group composition), which is bilinear in the embedded coordinates (c, s, p_x, p_y). Combined with the quotient ring constraint $c^2 + s^2 = 1$, the moment propagation closes automatically within a fixed basis. This closure is a consequence of the bilinear group structure, not a general property of the method.

Closure obstruction for nonlinear systems. For systems with polynomial drift of degree $d_X \geq 2$ (Duffing: $d_X = 2$, Lotka-Volterra: $d_X = 2$, coupled oscillators: $d_X = 2$, double-well Langevin: $d_X = 3$), the moment hierarchy does not close, under *either* the continuous generator or a discrete polynomial map.

Under Euler discretization with step Δt:

$$X_{k+1} = X_k + \Delta t\, g(X_k) + \sigma\sqrt{\Delta t}\, \xi_k, \tag{A.77}$$

the map $f(x) = x + \Delta t\, g(x)$ has degree d_X. Propagating degree-K moments requires: $\mathbb{E}[\phi(f(X_k))]$ for $|\phi| \leq K$, which involves monomials of X_k up to degree $d_X \cdot K$. The hierarchy is unclosed: degree-K moments depend on degree-$d_X K$ moments.

Truncating at degree K (setting higher moments to zero) introduces $O(\Delta t)$ error per step that compounds. This is the discrete analog of the continuous-time excess degree $\bar{d} = d_X - 1 \geq 1$: both formulations face the *same* fundamental obstruction.

Stein closure as the resolution. The SKF resolves this via the Stein identity (Section 4), which provides algebraic relations between moments of different orders, parameterized by λ. Unlike truncation, the Stein closure is *exact* for the polynomial exponential family: it exploits the parametric

structure of the density to express unclosed moments as functions of closed ones. The empirical evidence is in Figures A2–A3 (Appendix J.2): on both the Duffing oscillator ($\bar{d} = 1$, $d_X = 2$) and Lotka-Volterra ($\bar{d} = 1$, $d_X = 2$), all 14 centered moments up to degree 4 match 5×10^5-particle MC to within 1–7% over the full propagation interval. This capability (handling continuous-time SDEs with $\bar{d} \geq 1$ without moment truncation) is absent in the discrete-time MEM-KF framework and is a core contribution of this work.

SE(2) landmark benchmark. To verify that the SKF can match MEM-KF on its *own* benchmark, we reproduce the SE(2) localization experiment of [13]: a robot traverses a unit circle with heading noise $\sigma = 0.05\,\text{rad}/\sqrt{\text{s}}$, observing four symmetric landmarks with unknown data association (4-modal likelihood). We use an intrinsic Fourier $\times$ Legendre basis on $S^1 \times \mathbb{R}^2$ with $K_\theta = 8$ angular modes and position degree $r = 4$, avoiding the $c^2 + s^2 = 1$ quotient ring constraint of the embedded (c, s, p_x, p_y) coordinates used in [13]. The result is shown in Figure A1: the SKF mean trajectory tracks the ground truth comparably to a 5×10^5-particle bootstrap filter.

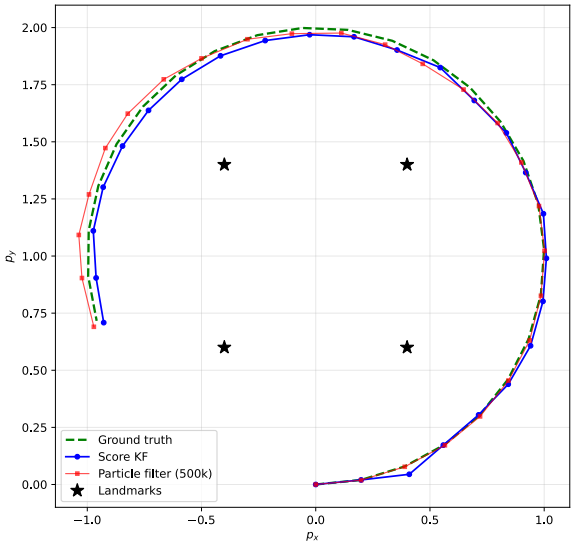


Figure A1: SE(2) landmark filter reproducing the benchmark of [13]. The SKF (Fourier $\times$ Legendre, K_θ=8, r=4) tracks the circular ground truth comparably to the particle filter. Stars mark the four landmarks with unknown data association.

Observe that a 5×10^5-particle bootstrap filter already tracks this benchmark well: the effective state dimension after the $c^2 + s^2 = 1$ constraint is $n = 3$, which is low enough for the particle filter to operate without weight degeneracy. We refer the reader to [13] for a more detailed performance comparison on this system with EKF, InEKF, UKF, UKF-M, and particle filters. Their results confirm that MEM-KF tracks well here, but so does the particle filter, and neither the partition function reconstruction nor the score matching machinery provides a clear advantage over simple importance resampling at this dimension. The scalability contrast with MEM-KF and the accuracy gap relative to the reported bootstrap PF become apparent at higher dimensions: as shown in Table A1, the SKF maintains practical 25-step filtering cost through $n = 20$ while the partition function integral in MEM-KF would require $O(G^n)$ evaluations, which in practice hits the scalability limit at $n \leq 8$ even with the GPU parallelization or compiled extensions.

J Additional experimental details

J.1 Discussion of Table A1

Prediction timing. Across the prediction benchmarks, the SKF avoids sampling during moment propagation and uses less wall-clock time than the 500k-particle Monte Carlo reference in Table A1. The largest timing gaps occur on manifold-valued systems. For SO(3) with Wigner D-matrices, the propagation reduces to independent small matrix exponentials on each l-block of size $(2l+1)^2$, while

MC must perform a quaternion multiplication for every particle at every time step. The $\bar{d} = 1$ systems (LV, 3D tracer) have a narrower gap because Stein closure adds per-step linear algebra that MC does not incur.

Filter accuracy. On the coupled Duffing oscillators, the SKF has lower mean RMSE than the reported baselines (EKF, UKF, EnKF, and PF) over 10 independent seeds. The EKF, UKF, and EnKF cluster at similar RMSE. In these runs, changing the Gaussian propagation rule does not remove the error associated with a non-Gaussian posterior. The 500k-particle PF also performs comparably to the Gaussian filters, which is consistent with particle degeneracy under partial observation. This PF baseline is a standard bootstrap filter using the transition prior as the proposal. Proposal-adapted PF variants (for example EKF- or UKF-informed proposals) are a useful stronger baseline for future study. The SKF behavior is consistent with the polynomial exponential-family belief retaining higher-order moment information about the nonlinear coupling.

Filter timing and scalability. In the multi-seed PF comparison, the SKF filter runtime grows from 4 s (n=6) to 41 s (n=10), scaling polynomially. At n=10, the SKF run uses less wall-clock time than the 500k-particle PF run on the same CPU (41 s vs. 312 s). The larger active-closure sweep reaches n=20 in 2079 s for a 25-step trajectory on the same CPU. The bootstrap PF is faster than the SKF in the largest rows ($n \geq 16$), although it remains much less accurate on this benchmark. The factorization-caching structure (Remark 2) matters at this scale: the Stein closure matrix $\Lambda_1(\lambda)$ is factorized once per prediction window and reused across all ODE substeps.

Comparison with MEM-KF. The MEM-KF [13] requires $O(G^n)$ quadrature for the partition function $Z(\lambda)$ at each step. At n=6 this is borderline feasible ($G^6 \sim 10^7$). At n=10 it exceeds 10^{16} operations per step, and at n=20 even a coarse $G = 15$ grid exceeds 10^{23} evaluations. The SKF avoids evaluating Z via score matching (Proposition 1), replacing the exponential-cost quadrature with a polynomial-cost linear solve. Additionally, MEM-KF is restricted to discrete-time bilinear maps whose moment propagation closes automatically. Continuous-time SDEs with $\bar{d} \geq 1$ require the Stein closure (Section 4), which is a contribution of this work. Once the posterior parameter λ^+ is available, another option is to extract a MAP state by minimizing the polynomial energy $\lambda^+ \cdot \phi(x)$. This is commonly done with a polynomial optimization problem and an SDP relaxation in the moment-SOS hierarchy [44, 45], as in MEM-KF-style state extraction [13]. This POP/SDP state-extraction step is compatible with the posterior parameter produced by the SKF, but the experiments here do not use it because it would add a separate SDP layer. We instead recover posterior moments, which the recursive filter already needs for the next prediction step, and report the posterior mean rather than a MAP point.

Multi-seed variability for all filters. Table A1 reports the SKF column as mean $\pm$ std and the baseline columns as mean only, to keep the summary table compact. Table A2 below repeats the coupled-oscillator filter rows with mean $\pm$ std for every method. Two patterns are visible. First, the Gaussian baselines have $\sigma/\mu \approx 10$–30% across all dimensions, i.e., they are seed-robust at the cost of being non-Gaussian-blind. Second, the SKF has comparatively wider σ/μ: about 40–80% for $n \leq 10$, about 60% at n=12, and above 100% in several larger rows because the mean RMSE is very small. Thus the absolute standard deviation remains small, but the relative ratio does not monotonically tighten. The SKF mean stays below every baseline in every row and at every dimension.

J.2 Moment trajectory accuracy

To verify that the Stein closure propagates moments correctly, we compare all centered moments up to degree 4 against large-sample Monte Carlo estimates over the full propagation interval. The Duffing moment-trajectory test initializes from $\mathcal{N}((0.5, 0)^\top, \mathrm{diag}(0.04, 0.04))$. The compact Lotka–Volterra moment panel initializes at the equilibrium $(\gamma/\delta, \alpha/\beta)$ with covariance $\mathrm{diag}(0.1, 0.1)$.

Duffing oscillator (n=2, r=4, $\bar{d}$=1): Figure A2 shows all 14 centered moments up to degree 4 over $t \in [0, 2]$. All moments (including variances, third-order skewness, and fourth-order kurtosis) agree with MC to within 1% relative error, confirming that a single Stein closure layer resolves the unclosed hierarchy with high accuracy.

Lotka-Volterra (n=2, r=6, $\bar{d}$=1): Figure A3 shows the same comparison over $t \in [0, 1.5]$, starting from the predator-prey equilibrium. Variances match MC within 1%. Third- and fourth-order moments stay within 2–7%, reflecting the larger dynamic range of the neutrally stable oscillatory dynamics.

SE(2) kinematics (n=3, $\bar{d}$=0): Figure A4 shows Fourier $\times$ position moments ($\mathbb{E}[\cos(k\theta)\, p_x^a p_y^b]$ and $\mathbb{E}[\sin(k\theta)\, p_x^a p_y^b]$) propagated via Dynkin's formula ($m(T) = e^{LT} m(0)$, no Stein closure needed). All 16 moments match MC within 0.5%.

SO(3) attitude (n=3, $\bar{d}$=0): Figure A5 shows 16 representative Wigner D-matrix moments $\mathbb{E}[D^l_{mm'}(R_t)]$ for $l = 1, 2, 3, 4$, propagated via block-diagonal matrix exponential ($m(T) = e^{L_l T} m(0)$ for each l-block). All significant moments match 2×10^5-particle MC to within 0.3%. The Wigner basis avoids the constraint-induced ill-conditioning of the monomial basis ($\kappa(A)$ from 10^{14} to $O(10^2)$).

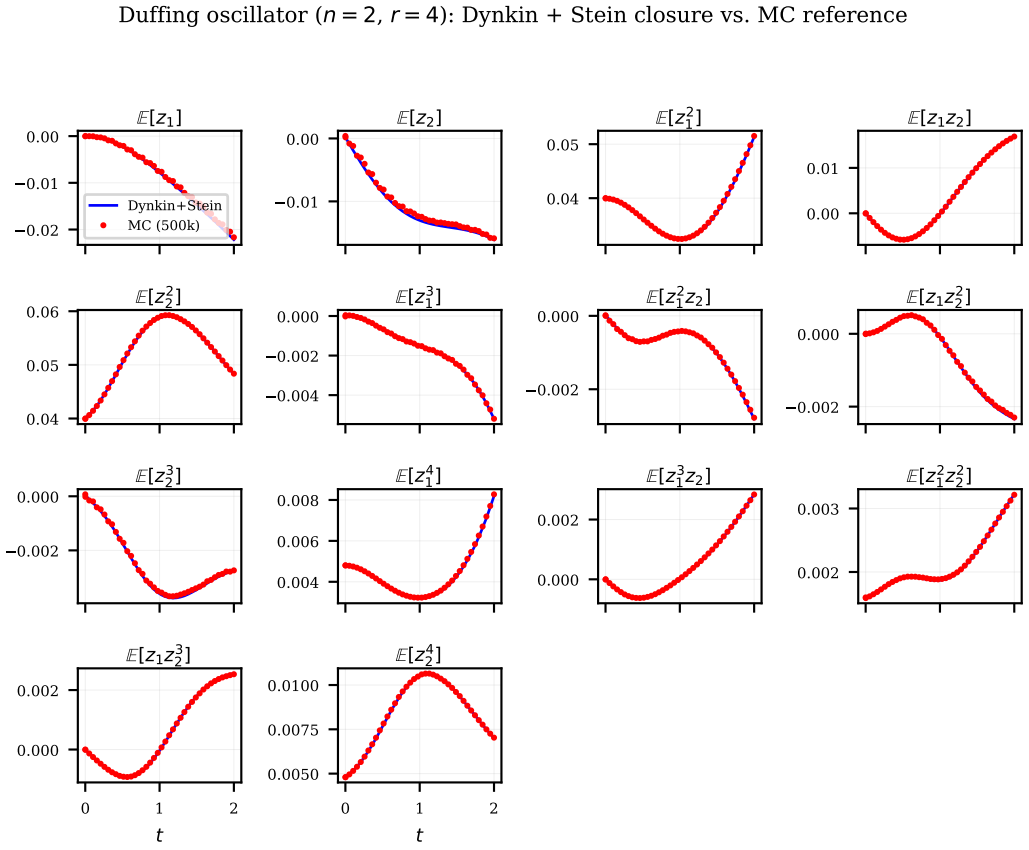


Figure A2: Duffing (n=2, r=4): Dynkin + Stein (blue) vs. MC (red) for all centered moments up to degree 4. Relative errors are below 1% across all displayed moments.

3D tracer advection (n=3, r=6, $\bar{d}$=1): Figure A6 shows 12 selected centered moment trajectories at degrees 2–5. Second moments match MC within 1.3%, third moments within 3.2%, and fourth moments within 5.7% at T=3. The largest errors are concentrated on pure z_3 powers, the direction of the nonlinear quadratic lift. The z_1 and z_2 moments, which are governed by linear equations, match MC closely. Figure A7 summarizes the per-degree RMS scaled error $|m_{\mathrm{SKF}} - m_{\mathrm{MC}}| / \max(|m_{\mathrm{MC}}|, \tau_d)$ where τ_d is the RMS moment magnitude at degree d. The hierarchy is clean: error grows monotonically with degree, from $<1\%$ at degree 2 to $\sim 8\%$ at degree 5.

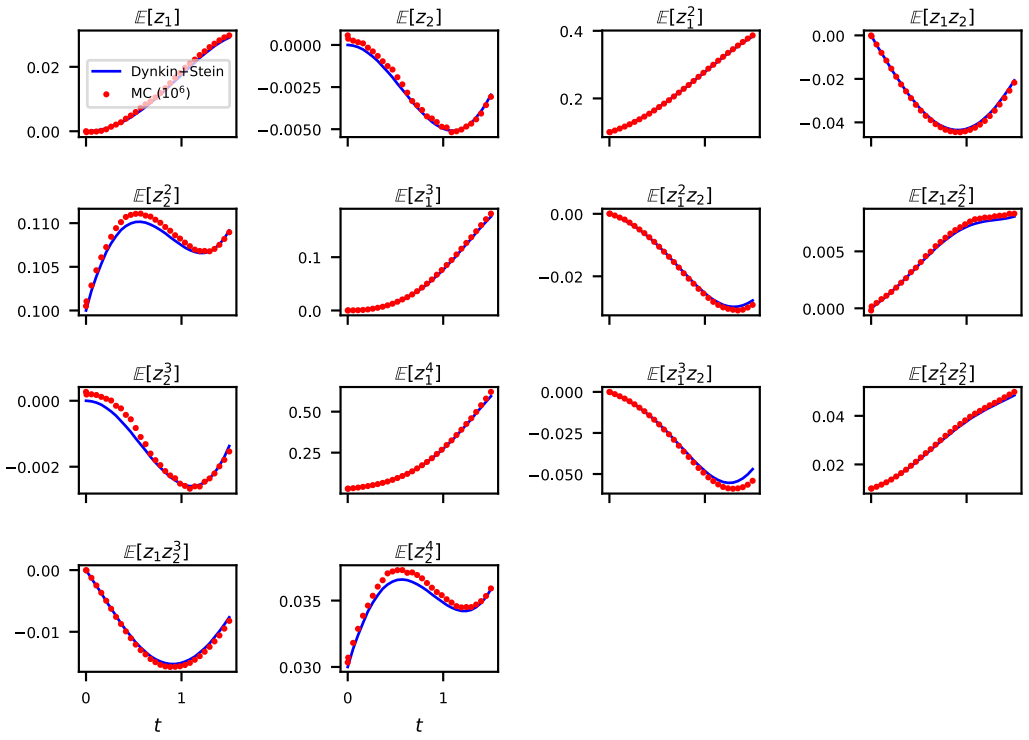


Figure A3: Lotka-Volterra (n=2, r=6): Dynkin + Stein (blue) vs. MC (red). Variance errors stay below 1%, while third- and fourth-order moments are within 2–7%.

Table A1: Comparison across methods. The SKF score-matching fit costs $O(M^3)$ with $M = \binom{n+r}{n}$. When Stein closure is needed, the closure solve adds a factor depending on the active unclosed moment system. MEM-KF [13] cost is dominated by $O(G^n)$ quadrature for $Z(\lambda)$, exponential in n. The coupled-oscillator rows report RMSE for the SKF and baselines. The high-dimensional coupled oscillator rows (n=12–20) use active Stein closure.

Experiment	**SKF (ours)**	**EKF**	**UKF**	**EnKF**	**PF (500k)**	**MEM-KF**
Prediction (timing):						
SE(2) (n=4, $\bar{d}$=0)	0.6 s	<0.1 s	–	–	209 s	est. ∼35 s[§]
SO(3) ($l_{\max}$=4, $\bar{d}$=0)	<0.1 s	<0.1 s	–	–	396 s	est. ∼100 s[§]
Lotka-Volterra (n=2, $\bar{d}$=1)	0.7 s	<0.1 s	–	–	69 s	n/a[‡]
3D tracer (n=3, r=6, $\bar{d}$=1)	8 s	<0.1 s	–	–	122 s	n/a[‡]
Filter (mean RMSE / 25 steps):						
Duffing (n=2, r=4, $\bar{d}$=1)	tracks	tracks	tracks	tracks	tracks	n/a[‡]
Coupled osc. (n=4, r=3)	**.0103±.0040**	.0676	.0725	.0762	.0727	n/a[‡†]
Coupled osc. (n=6, r=3)	**.0200±.0154**	.0746	.0746	.0762	.0754	n/a[‡†]
Coupled osc. (n=8, r=3)	**.0181±.0112**	.0713	.0775	.0806	.0747	n/a[‡†]
Coupled osc. (n=10, r=3)	**.0197±.0123**	.0728	.0743	.0752	.0744	n/a[‡†]
Coupled osc. (n=12, r=3)	**.0029±.0017**	.0708	.0731	.0768	.0741	n/a[‡†]
Coupled osc. (n=14, r=3)	**.0039±.0042**	.0801	.0810	.0842	.0818	n/a[‡†]
Coupled osc. (n=16, r=3)	**.0055±.0041**	.0752	.0773	.0820	.0781	n/a[‡†]
Coupled osc. (n=18, r=3)	**.0039±.0044**	.0772	.0784	.0816	.0791	n/a[‡†]
Coupled osc. (n=20, r=3)	**.0039±.0050**	.0734	.0752	.0806	.0761	n/a[‡†]
Filter runtime (25 steps, single CPU):						
Duffing (n=2)	0.3 s	<0.1 s	<0.1 s	1.8 s	46 s	–
Coupled osc. (n=4)	1 s	<0.1 s	<0.1 s	1.2 s	110 s	–
Coupled osc. (n=6)	4 s	<0.1 s	0.1 s	1.7 s	184 s	–
Coupled osc. (n=8)	14 s	<0.1 s	0.1 s	2.9 s	291 s	–
Coupled osc. (n=10)	41 s	<0.1 s	0.2 s	3.1 s	312 s	–
Coupled osc. (n=12)	195 s	0.02 s	0.21 s	4.0 s	385 s	–
Coupled osc. (n=14)	344 s	0.02 s	0.34 s	4.8 s	441 s	–
Coupled osc. (n=16)	810 s	0.03 s	0.44 s	5.1 s	540 s	–
Coupled osc. (n=18)	1182 s	0.03 s	0.47 s	5.8 s	603 s	–
Coupled osc. (n=20)	2079 s	0.03 s	0.57 s	6.5 s	646 s	–
Z evaluations per step	**0**	0	0	0	0	≥ 1
Iterative optimization	**no**	no	no	no	no	BFGS
Cost per step	$O(M^3)$ + closure	$O(n^3)$	$O(n^3)$	$O(N_e n)$	$O(N_p n)$	$O(G^n M)$

PF uses 5×10^5 bootstrap particles. EnKF uses 500 ensemble members.

[‡]MEM-KF [13] propagates moments via discrete-time polynomial maps $X_{k+1} = f(X_k, U_k, W_k)$ with state-independent noise, which close automatically for bilinear maps. Continuous-time SDEs with $\bar{d} \geq 1$ (e.g., Duffing, LV, coupled oscillators, etc.) require moment closure (Appendix I), which MEM-KF does not address.

[†][13] demonstrated MEM-KF on SE(2) (3D integration after constraint reduction) using adaptive Gauss–Kronrod quadrature (order 7, 15 Kronrod points/dim per subregion). The partition function $Z(\lambda)$ requires n-dimensional numerical integration at each BFGS iteration. At $n\geq6$ the cost per Z evaluation grows as $O(G^n)$ and has not been demonstrated on a single laptop CPU. For example, at n=8: $15^8 \approx 2.6 \times 10^9$, and at n=10: $15^{10} \approx 5.8 \times 10^{11}$ evaluations per subregion.

[§]*MEM-KF timing estimates.* SE(2): 3D quadrature ($c^2+s^2=1$ reduces n=4 to 3 effective dimensions) with M=45 basis functions. At G=30 points/dim: G^3=27,000 evaluations per Z call, ∼20 BFGS iterations with function + gradient ≈40 × 27,000 × 45 ≈49M operations per reconstruction. Estimated ∼35 s in Python. SO(3): same 3D integration but M=165 Wigner D-matrix basis functions ($l_{\max}$=4: $\sum_{l=0}^{4}(2l+1)^2$=165). Each $D^l_{mm'}(R)$ evaluation requires $O(l^2)$ operations via the Wigner d-matrix recurrence, ∼3× costlier per basis function than monomial evaluation. With more parameters and costlier evaluations: estimated ∼100 s.

Table A2: Coupled oscillator filter RMSE: full mean $\pm$ std for every method, averaged over 10 independent seeds at every dimension.

System	SKF (ours)	EKF	UKF	EnKF	PF (500k)
Coupled osc. (n=4, r=3)	**.0103±.0040**	.0676±.0143	.0725±.0136	.0762±.0150	.0727±.0136
Coupled osc. (n=6, r=3)	**.0200±.0154**	.0746±.0232	.0746±.0255	.0762±.0265	.0754±.0232
Coupled osc. (n=8, r=3)	**.0181±.0112**	.0713±.0160	.0775±.0129	.0806±.0121	.0747±.0155
Coupled osc. (n=10, r=3)	**.0197±.0123**	.0728±.0134	.0743±.0222	.0752±.0221	.0744±.0140
Coupled osc. (n=12, r=3)	**.0029±.0017**	.0708±.0089	.0731±.0086	.0768±.0085	.0741±.0087
Coupled osc. (n=14, r=3)	**.0039±.0042**	.0801±.0084	.0810±.0079	.0842±.0087	.0818±.0082
Coupled osc. (n=16, r=3)	**.0055±.0041**	.0752±.0077	.0773±.0072	.0820±.0080	.0781±.0072
Coupled osc. (n=18, r=3)	**.0039±.0044**	.0772±.0074	.0784±.0069	.0816±.0073	.0791±.0068
Coupled osc. (n=20, r=3)	**.0039±.0050**	.0734±.0069	.0752±.0063	.0806±.0077	.0761±.0062

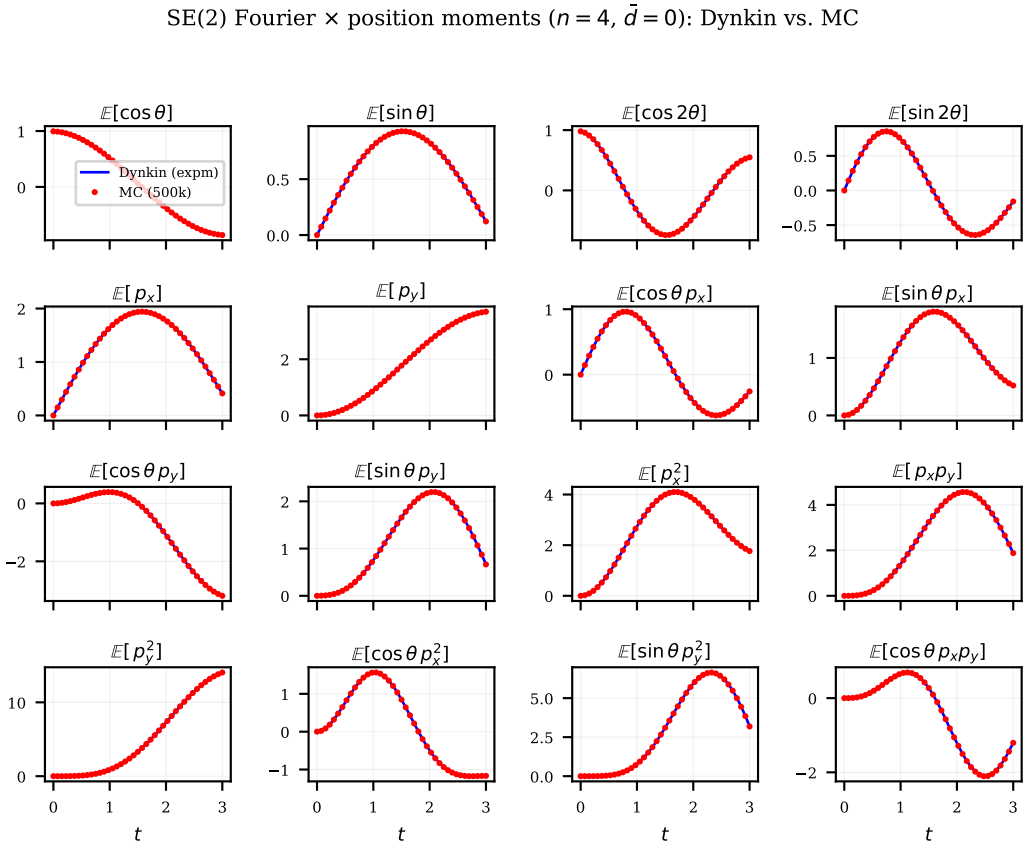


Figure A4: SE(2) Fourier $\times$ position moments (n=4, $\bar{d}$=0): Dynkin (blue) vs. MC (red) for 16 moments, including the Fourier harmonics ($\cos\theta$, $\sin\theta$, $\cos 2\theta$, $\sin 2\theta$), position means, second moments, and cross-moments ($\cos\theta \cdot p_x$, $\cos\theta \cdot p_x^2$, etc.). Relative errors stay below 0.5% on all of them.

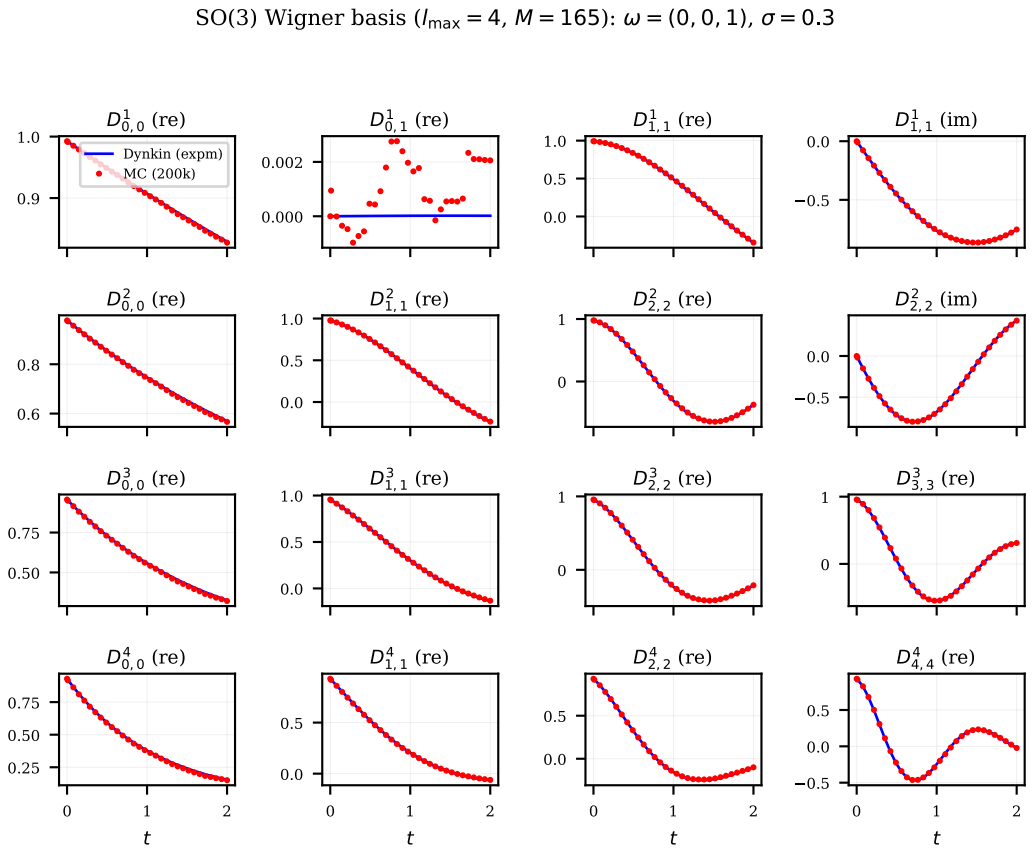


Figure A5: SO(3) Wigner basis ($l_{\max}$=4, M=165): Dynkin (blue) vs. MC (red) for representative Wigner moments. Moments with appreciable magnitude ($D^1_{0,0}$, $D^1_{1,1}$) agree to within $\sim$0.3%, and moments that are near zero in the MC reference remain near zero in the Dynkin trajectory.

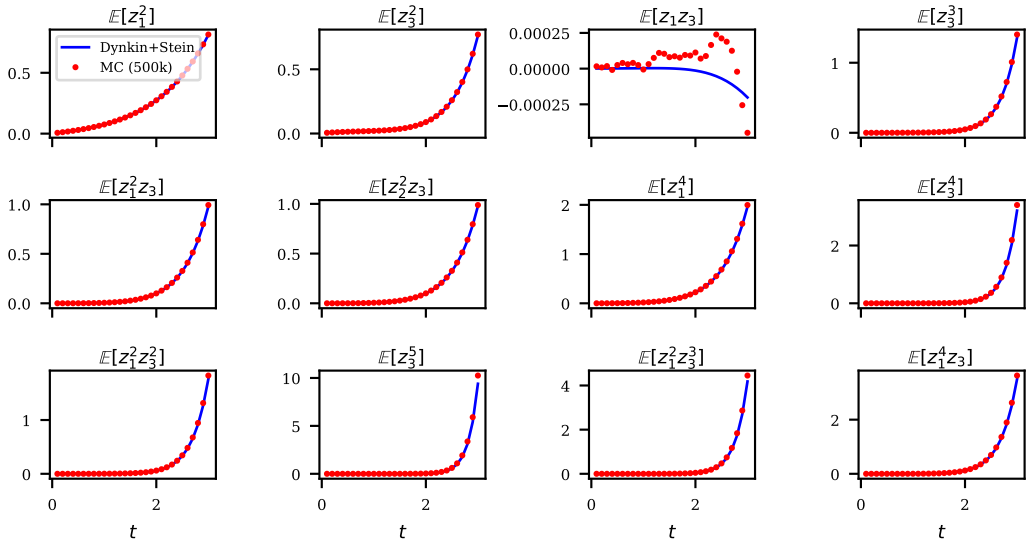


Figure A6: Twelve selected centered moment trajectories for the 3D tracer (n=3, r=6): Dynkin + Stein (blue) vs. MC (red). Degree 2–3 moments (rows 1–2) show excellent agreement through T=3. Degree 4–5 moments (rows 3–4) track well through T=2.5, with a small gap appearing at T=3.

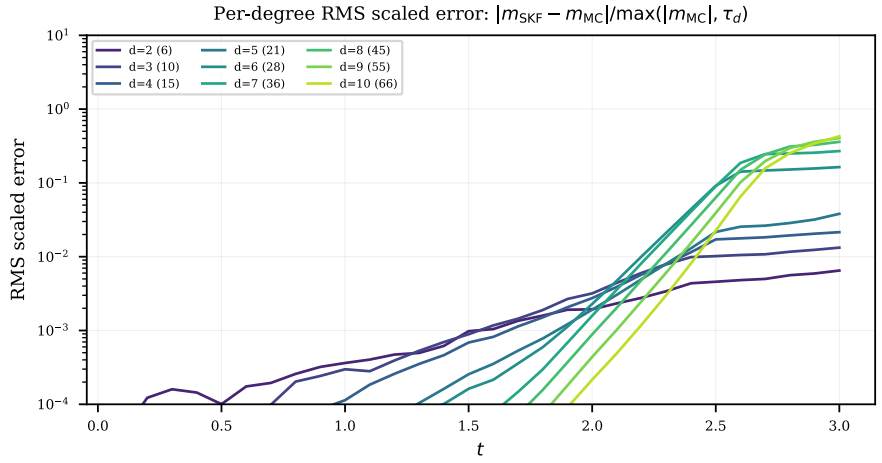


Figure A7: Per-degree RMS scaled error for the 3D tracer. Each curve aggregates all $\binom{d+2}{2}$ moments of degree d. The propagated state includes degrees up to K=2r−2=10 (not just the basis order r=6). Degrees 2–5 remain below 8% through T=3. Higher degrees degrade faster, consistent with the exponential amplification of the nonlinear z_3 coupling.

J.3 Density reconstruction from propagated moments

The polynomial MED $p(x;\lambda) \propto \exp(-\lambda \cdot \phi(x))$ provides a closed-form density from the fitted parameters λ. We visualize this density on several systems, ordered as in Appendix G.

Lotka-Volterra (n=2, r=6, $\bar{d}$=1): Figure A8 shows the (x_1, x_2) density of the stochastic predator-prey system at $T = 1.5$ s. The bilinear interaction $x_1 x_2$ produces an unclosed moment hierarchy that Stein closure resolves (Section 4). The polynomial MED at $r = 6$ reconstructs the anisotropic, non-Gaussian shape, matching the 10^6-particle MC reference. This density reconstruction uses initial law $\mathcal{N}((\gamma/\delta, \alpha/\beta)^\top, \mathrm{diag}(0.3^2, 0.2^2))$.

Double-well Langevin (n=2, r=6, $\bar{d}$=2): Figure A9 shows a 2D double-well Langevin system $dX = (-\nabla V(X))\,dt + \sigma\,dW$ with $V(x) = \frac{1}{4}x_1^4 + \frac{1}{4}x_2^4 - \frac{1}{2}x_1^2 - \frac{1}{2}x_2^2$ ($\sigma = 0.5$, $T = 3$ s). Starting near the hilltop, the distribution splits into 4 modes at $(\pm 1, \pm 1)$. Score matching (r=6, M=28, $\kappa(A) = 42$) reconstructs all 4 modes, while the EKF Gaussian does not capture the multimodal structure (kurtosis 1.4 vs. 3 for Gaussian).

3D tracer advection (n=3, r=6, $\bar{d}$=1): Starting from $X_0 \sim \mathcal{N}([0, 0, 0.5]^\top, 0.05^2 I)$, the nonlinear coupling creates an asymmetric plume: Figure A10 shows the (x_1, x_3) marginal of the tracer density at four times. The quadratic lift $\alpha(x_1^2 + x_2^2)$ produces a parabolic plume that the EKF (Gaussian, symmetric ellipse) cannot represent. SKF propagates 286 centered moments (K=10) and reconstructs the non-Gaussian marginal, matching the 500k-particle MC reference. Moment accuracy: $\mathbb{E}[z_3^2]$ within 1.3% and $\mathbb{E}[z_3^3]$ within 3.2% of MC at T=3 (Appendix J.2). Timing: SKF 8 s, MC (500k) 122 s, EKF <0.1 s.

SE(2) localization (n=4, $\bar{d}$=0): The reconstructed (p_x, p_y) marginal is Figure 1 in the main text. The non-Gaussian distribution arises from heading uncertainty coupled to translational motion, and the polynomial MED captures the curved, asymmetric shape starting at $r = 4$ ($M = 15$ parameters, one linear solve in 0.4 ms). The MC reference uses 10^6 particles.

SO(3) attitude ($l_{\max}$=4, $\bar{d}$=0): Figure A11 shows the marginal density of each body-axis direction on S^2 after rotation kinematics with angular velocity noise. The MED fitted from Wigner D-matrix moments captures the concentration near the north pole, matching the MC reference. The Gaussian (EKF) approximation produces a similar shape for this axially symmetric system. The two reconstructions are expected to differ visibly under asymmetric dynamics or multimodal posteriors from measurement ambiguity, which we defer to a companion paper on Lie group filtering.

SE(3) position uncertainty (n=6, $\bar{d}$=1): Figure A12 shows the (x, y) marginal of the position density for an SE(3) rigid body following a curved trajectory with rotation noise ($T_{k+1} = T_k \exp(U_k + w_k)$, $w_k \sim \mathcal{N}(0, Q)$ in $\mathfrak{se}(3)$). Moments are computed from 5×10^5 MC samples. Dynkin propagation on SE(3) requires a Wigner $\times$ Legendre basis and is left for the Lie-group filter extension. The rotation-translation coupling produces a curved, asymmetric shape that the Gaussian (r=2) does not capture. At r=5 (M=56, $\kappa(A) = 86$ after Jacobi preconditioning), the reconstructed density matches the MC reference.

J.4 Duffing filter

The single Duffing oscillator (n=2, r=4, $\bar{d}$=1) with σ=0.15 serves as a basic validation of the full predict-update loop. The measurement model is $y = x_1 + v$ with $v \sim \mathcal{N}(0, 0.04)$, which is linear ($d_g$=1) and therefore conjugate for any $r \geq 2$. We run 25 predict-update cycles at $\Delta t_{\mathrm{pred}} = 0.2$ s, with prediction via Dynkin + Stein closure (RK4) and update via the conjugate score addition $\lambda^+ = \lambda^- + \lambda_{\mathrm{lik}}$ followed by Stein moment recovery (42 equations, 27 unknowns, full column rank). The initial law is $\mathcal{N}((0.5, 0)^\top, 0.1^2 I)$. All computations are performed in centered coordinates $z = x - \mu$, so each step reduces to a single linear system solve.

On this mildly nonlinear system, all methods track the ground truth (Figure A13). The EKF, UKF, and EnKF perform comparably, indicating that for weak nonlinearity the Gaussian approximation itself is adequate. A clearer separation between the SKF and the Gaussian baselines emerges at higher dimensions with stronger nonlinearity (Section 7.3).

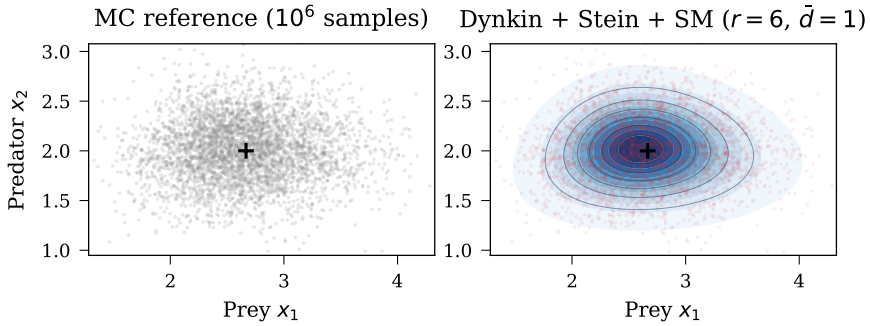


Figure A8: Stochastic Lotka-Volterra ($\bar{d} = 1$): the bilinear interaction $x_1 x_2$ produces an unclosed moment hierarchy that is resolved algebraically by the Stein closure. Left: 10^6-particle MC reference. Right: reconstructed density from Dynkin + Stein closure + score matching ($r = 6$), with MC samples overlaid in red. The reconstruction tracks the anisotropic, non-Gaussian shape of the reference.

Double-well Langevin ($n = 2$, $r = 6$): 4-modal density at $T = 3.0$s
$V(x) = \frac{1}{4}x^4 - \frac{1}{2}x^2$, $\sigma = 0.5$. Black +: potential minima.

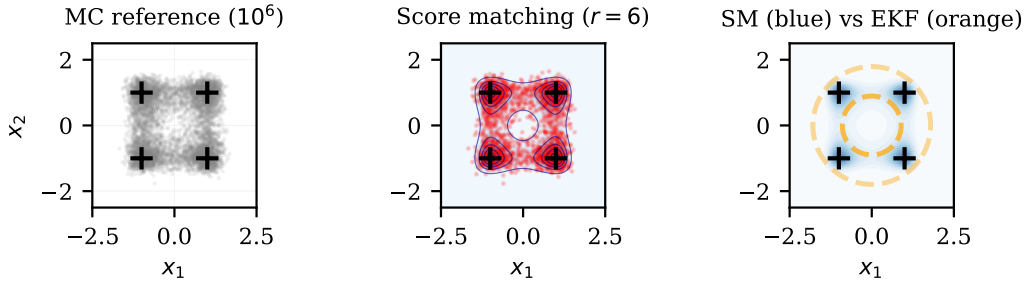


Figure A9: Separable double-well Langevin (n=2, r=6): $V(x) = \frac{1}{4}x_1^4 + \frac{1}{4}x_2^4 - \frac{1}{2}x_1^2 - \frac{1}{2}x_2^2$, 4-modal density at $T = 3$ s. Left: 10^6-particle MC reference. Center: score matching density reconstruction (r=6, $\kappa(A)$=42) with MC samples (red). Right: score matching (blue) vs. EKF Gaussian (orange, $1\sigma/2\sigma$). Black crosses: potential minima at $(\pm 1, \pm 1)$. The score-matching reconstruction reproduces the four-mode structure of the reference. The Gaussian fit, restricted to a unimodal density, does not.

J.5 Coupled oscillator filter time series

Figures A14–A22 show the full predict-update filter trajectories for the coupled Duffing oscillator network at n=4, 6, . . . , 20. At all dimensions, the SKF (blue) tracks the ground truth (green dashed) more closely than the reported baselines: EKF (orange), UKF (purple), EnKF (brown), and the 500k-particle bootstrap PF (red). The EKF, UKF, and EnKF cluster together at similar RMSE, suggesting that the limiting factor is the Gaussian approximation, not the specific approximation strategy (linearization vs. sigma points vs. ensemble). The PF RMSE is also close to these Gaussian baselines under partial observation of the odd-indexed positions. For n>10, the r=3 filter uses active Stein closure for the degree-five moments requested by the moment ODE. This keeps the closure overdetermined without introducing a chordal or banded sparsification. Because this algorithmic change begins at n=12, the drop in RMSE from n=10 to n=12 should be read as the effect of the active closure on this structured generator, not as a pure dimension-scaling phenomenon. The SKF infers the unobserved oscillators from coupling alone, exploiting the full moment-based belief representation. Quantitative RMSE comparisons are in Table A1. For reproducibility, the coupled oscillator filter uses $\gamma = 0.3$, $\alpha = 1.0$, $\beta = 0.6$, $\kappa = 0.3$, $\sigma = 0.4$, $\Delta t_{\text{pred}} = 0.15$, RK4 step $\Delta t_{\text{ode}} = 0.005$, and 25 predict-update cycles. The observed coordinates are $q_1, q_3, \ldots$ with measurement covariance

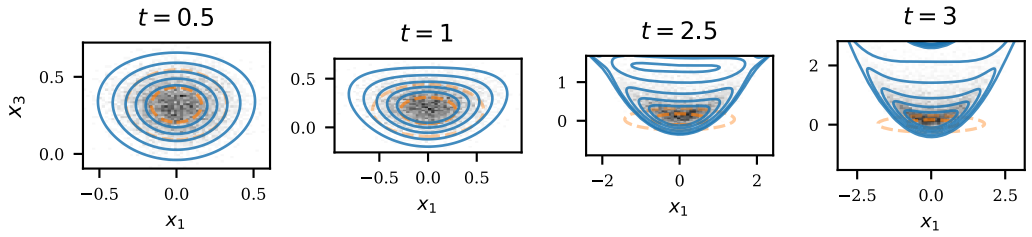


Figure A10: 3D tracer advection (n=3, r=6, $\bar{d}$=1): (x_1, x_3) marginal at four times. The SKF reconstruction (blue) tracks the parabolic plume induced by the quadratic lift, while the EKF reconstruction (orange) remains symmetric and does not represent this curvature. Reference: 500k MC particles (gray).

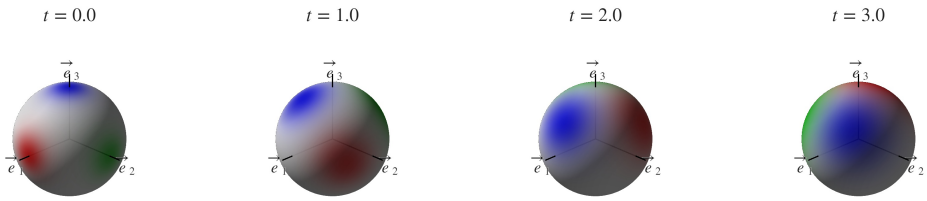


Figure A11: SO(3) attitude density evolution: marginal distribution of each body axis on S^2 (Red=$\vec{b}_1$, Green=$\vec{b}_2$, Blue=$\vec{b}_3$). $\omega_0 = (0.5, 0.3, 1.0)$ rad/s, $\sigma = 0.15$. At t=0 the distribution is concentrated near the identity rotation. As time evolves, the tilted rotation creates asymmetric spreading.

$R = 0.3^2 I$. The initial law is Gaussian with covariance $0.15^2 I$ and mean $(q_1, \ldots, q_N, p_1, \ldots, p_N)$ whose position block cycles through $(0.3, -0.2, 0.1, -0.3, 0.15, 0.25, -0.1, 0.2)$ and whose momentum block is zero.

K Stein perspective on the duality of parameter and moment recovery

This section provides a unified, high-level interpretation of both score parameter estimation (Proposition 1) and moment recovery (Proposition 2) through the Stein identity. Rather than viewing these procedures as separate constructions, we show that they arise from the same underlying weak formulation.

Stein identity as a weak constraint. For a smooth density p with score $s(x) = \nabla \log p(x)$, the Stein identity [39] states that

$$\mathbb{E}[s(x) \cdot \nabla f(x) + \Delta f(x)] = 0, \tag{A.78}$$

for all sufficiently regular test functions f. This is naturally interpreted as a *weak formulation*: instead of imposing a pointwise condition on p, the constraint is enforced in an integrated sense against arbitrary test functions. This weak-form perspective underlies a range of modern computational-statistics methods built on Stein operators, including kernel Stein discrepancies [46] and minimum Stein discrepancy estimation [47]. Anastasiou et al. [48] review these developments.

Score matching as a projected Stein system. The score matching formulation in Proposition 1 can be interpreted as a projection of the weak Stein identity onto a finite-dimensional test space with a parametric score model, paralleling the broader Stein-discrepancy framing of score-matching

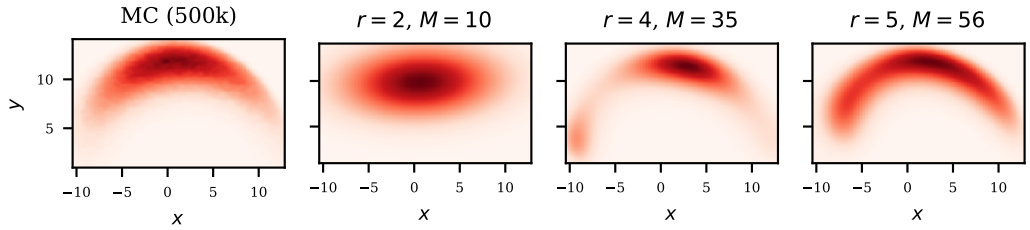


Figure A12: SE(3) position density at the final time step, shown as the (x, y) marginal. The leftmost panel is the 500k-particle MC reference. The remaining panels show the score-matching reconstruction at $r = 2$, 4, and 5. The non-Gaussian curvature induced by rotation-translation coupling is captured from $r = 4$ onward.

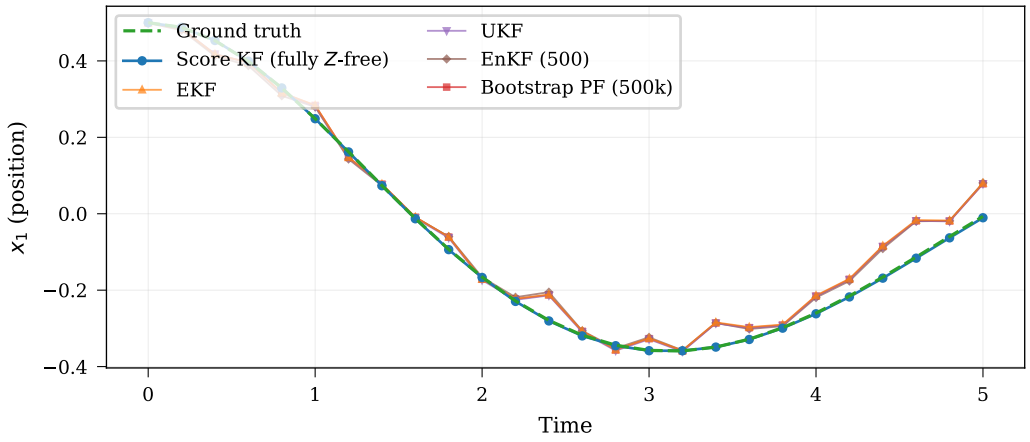


Figure A13: Score Kalman Filter on the Duffing oscillator (n=2, r=4). Ground truth (green dashed), SKF (blue), EKF (orange), UKF (purple), EnKF (brown), and bootstrap PF (5×10^5 particles, red). On this mildly nonlinear benchmark all methods track the ground truth.

estimators in [47, 19]. Restrict the test functions to

$$\mathcal{V}_r = \mathrm{span}\{\phi_\alpha(x) = x^\alpha \mid |\alpha| \le r\}.$$

Substituting the score model $s_\lambda(x) = -J_\phi(x)\lambda$ into (A.78) and taking $f = \phi_\beta$ yields

$$\mathbb{E}_p[-(J_\phi\lambda)\cdot\nabla\phi_\beta + \Delta\phi_\beta] = 0,$$

for all $|\beta| \le r$. Collecting these equations for all basis functions gives

$$\mathbb{E}_p[J_\phi^\top J_\phi]\,\lambda = \mathbb{E}_p[\Delta\phi],$$

which recovers (9).

Moment recovery as the dual problem. Conversely, if the score parameters λ are given, the same projected Stein identity becomes a system of linear relations among the moments. The unknowns are now the higher-order moments, and the Stein equations provide a mechanism for reconstructing them from lower-order statistics. This is precisely the role of the Stein closure in Proposition 2.

In the ideal (infinite-dimensional) setting, the Stein identity characterizes the distribution through an infinite family of constraints. In practice, both the test space and the moment representation are

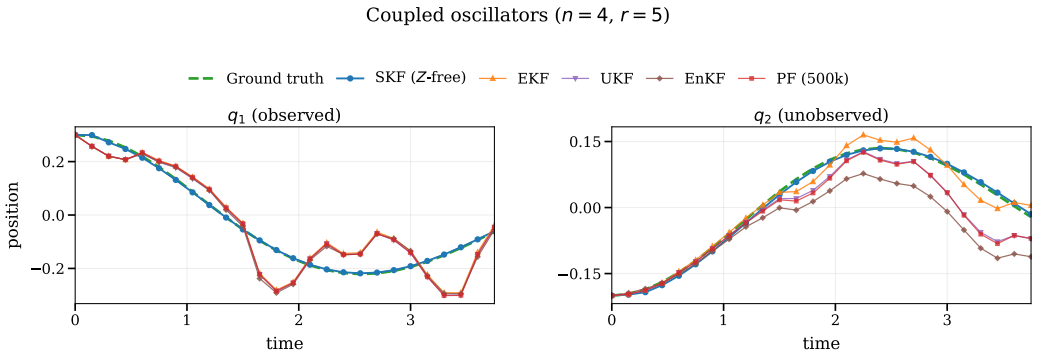


Figure A14: Coupled oscillators (n=4, r=3). RMSE averaged over 25 steps for the displayed seed: SKF 0.005, EKF 0.057, UKF 0.061, EnKF 0.067, PF 0.062.

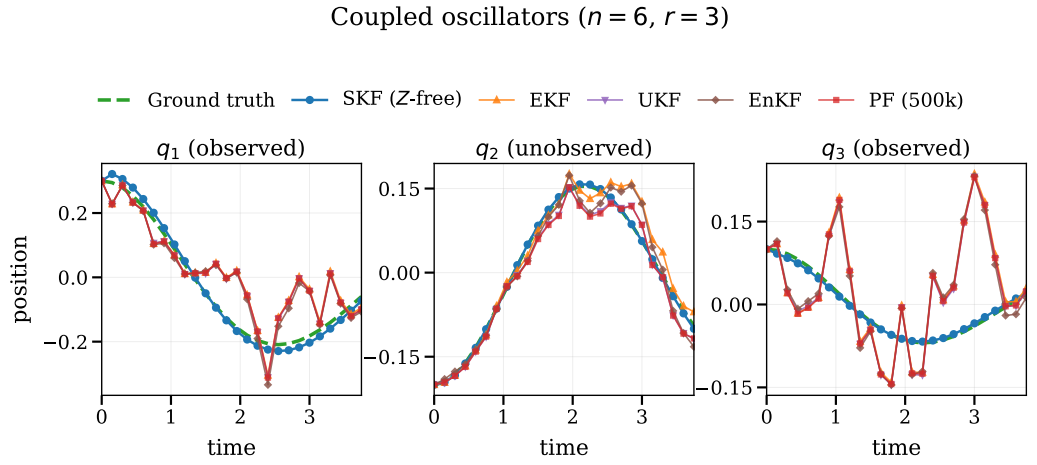


Figure A15: Coupled oscillators (n=6, r=3). RMSE averaged over 25 steps for the displayed seed: SKF 0.011, EKF 0.078, UKF 0.077, EnKF 0.078, PF 0.079.

truncated. For moment recovery we use the component-wise identities underlying Proposition 2, not only the scalar collapsed form (A.78). With x^β restricted to $\mathcal{V}_K = \mathrm{span}\{x^\beta : |\beta| \le K\}$, given λ and moments $m^{(\le K)}$, the moments of degree $K{+}1$ are obtained by minimizing the residual of the projected component-wise Stein operator:

$$m^{(K+1)} = \arg\min_m \sum_{|\beta|\le K} \sum_{i=1}^{n} \left(\mathbb{E}\left[s_{\lambda,i}(x)x^\beta + \partial_i x^\beta\right]\right)^2 .$$

Summary. Score matching and Stein-based moment recovery arise from the same fundamental principle: enforcing the Stein identity over a finite set of test functions. The difference lies only in the choice of unknowns. Score matching solves for the score parameters given moments (Proposition 1). Stein closure solves for moments given the score parameters (Proposition 2). In both cases, an infinite collection of constraints is approximated by a finite one, typically solved in a least-squares sense due to truncation. This unifies parameter estimation and moment closure as dual instances of the same projection principle.

This connects to weak formulations of stochastic dynamics more broadly. Moment propagation follows from Dynkin's formula, which describes how expectations evolve over time, while the Stein identity enforces consistency between the score representation and the moments at each time.

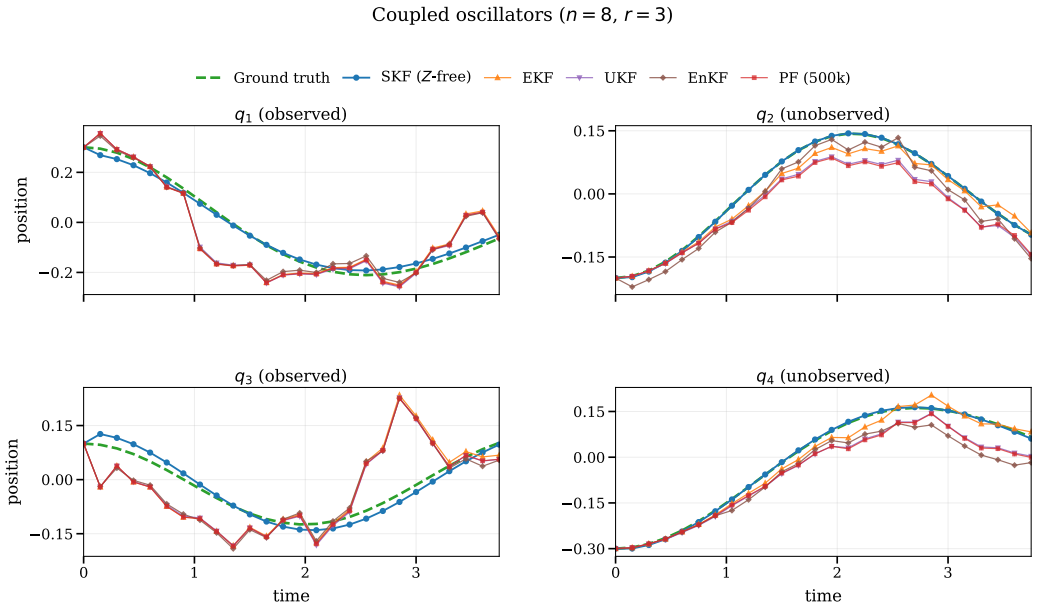


Figure A16: Coupled oscillators (n=8, r=3). RMSE averaged over 25 steps for the displayed seed: SKF 0.013, EKF 0.068, UKF 0.070, EnKF 0.073, PF 0.072.

The proposed framework can therefore be viewed as combining two complementary projections: one governing temporal evolution (via Dynkin) and one enforcing distributional consistency (via Stein). Together, they provide a finite-dimensional approximation of the underlying density evolution problem.

L Future research directions and open questions

(i) *Off-model moment consistency.* The reconstruction step minimizes Fisher divergence rather than KL divergence. When the true density lies in the polynomial exponential family, Theorem 1 shows that the distinction disappears: score matching recovers the same parameters as maximum entropy. Away from the model class, however, the fitted density need not reproduce every propagated moment exactly. The Stein augmentation in Section 4 reduces this mismatch, but the remaining error should depend on both the basis order r and the regularity of the target density. Making that dependence explicit is still open.

(ii) *Automatic bases and conditioning.* The theory is basis-invariant, but the linear algebra is not. Raw monomials give $\kappa(A) \sim 10^4$–10^6 in practice. Centering coordinates and switching to orthogonal bases such as Legendre polynomials bring this down to $O(10)$–$O(100)$ in the examples here. For that reason, centered orthogonal coordinates are the default implementation rather than a cosmetic change (Appendix H.4 gives the centering formulas and Appendix H.3 gives the change-of-basis theory). A practical open problem is that this basis choice is still hand tuned. Automatic basis selection or adaptive preconditioning along the filter trajectory remains future work.

(iii) *Lie group score matching beyond compact groups.* On SO(3), the Wigner D-matrix basis is already intrinsic: the block-diagonal moment propagation follows from the representation-theoretic structure of the generator (Casimir element + Lie algebra drift), and the embedding-coordinate derivation in Appendix G.8 is merely a convenient route to the same result. Score matching on general Riemannian manifolds is well-studied [49], and on compact groups (e.g., tori, SO(N)), harmonic analysis provides a finite spectral basis with closed-form moment propagation [50]. The gap is in noncompact groups such as SE(3) = SO(3) $\ltimes \mathbb{R}^3$: the irreducible unitary representations are infinite-dimensional (induced representations from the semidirect structure). This is exactly where the compact-group story stops being automatic: we do not know a finite spectral basis with

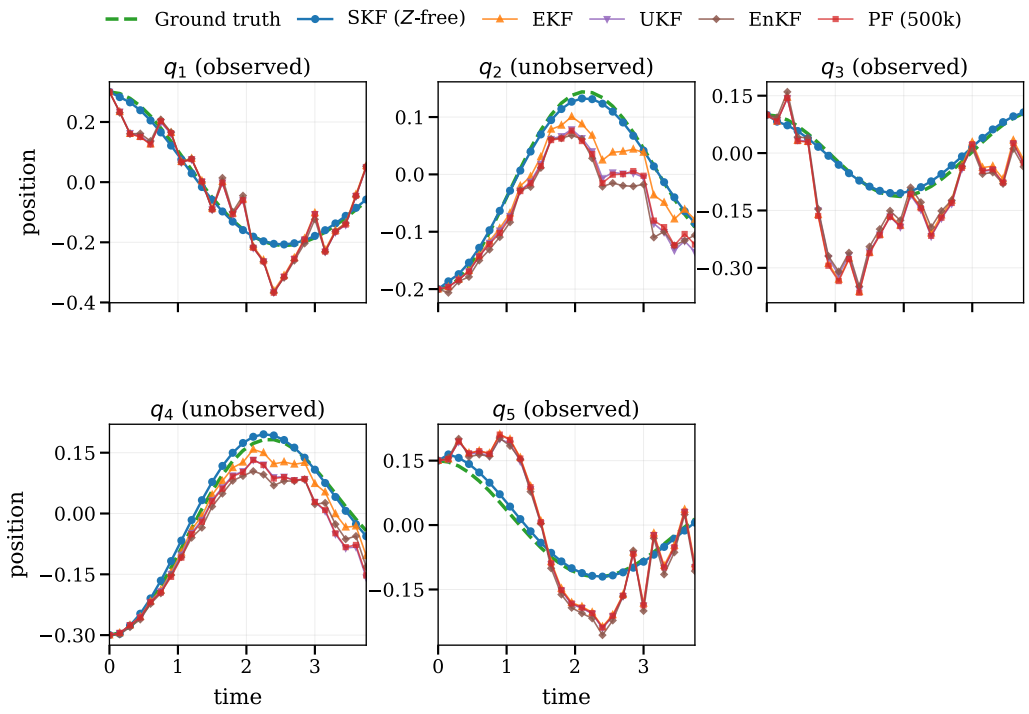


Figure A17: Coupled oscillators (n=10, r=3). RMSE averaged over 25 steps for the displayed seed: SKF 0.012, EKF 0.080, UKF 0.081, EnKF 0.083, PF 0.083.

comparable moment-closure properties. Our SE(2) experiments use a product basis (Fourier for the compact SO(2) factor, Legendre for $\mathbb{R}^2$) adapted to the semidirect structure, but this is not the full SE(2) harmonic analysis. For SE(3), the natural product basis is Wigner $\times$ Legendre (Wigner for the rotation subgroup, Legendre for the translational coordinates). The SE(3) density reconstruction in Appendix J.3 uses MC-sampled moments as a proof of concept. Dynkin + Stein closure on SE(3) with the Wigner $\times$ Legendre basis, and the corresponding predict-update filter on Lie groups, are the next step rather than part of the present paper.

(iv) *Stein closure counts and conditioning.* There are two questions hiding under "well-posedness." The first is combinatorial: does the finite Stein system provide enough equations for the degree-$(K{+}1)$ moments? The second is numerical: once there are enough equations, is the resulting matrix full column rank and well conditioned along the trajectory? Appendix B.5 answers the counting part for the full-basis system and gives a sufficient condition for when the Stein closure is overdetermined. For r=3, the standard augmented closure is overdetermined through n=15 (including the even-dimensional oscillator cases $n \leq 14$) and becomes underdetermined at n=16. The oscillator sweep at n=12–20 uses the generator itself: the active closure solves only for the degree-five moments requested by the moment ODE, a restriction that becomes essential at n=16, 18, 20. At n=20, this leaves 14,500 targets and 32,800 Stein equations. The remaining gap is not the equation count but a structural rank and conditioning theory for these counted and active systems. Even with full column rank, $\Lambda_1(\lambda)$ can become poorly conditioned as the fitted parameters $\lambda(t)$ approach the bifurcation surface $\mathcal{B}$ discussed below, making the closure numerically fragile near modality transitions.

(v) *Stein recovery is a projection after truncation.* The posterior moment recovery (Section 5.2) inverts the parameter-to-moment map via a truncated Stein system. The self-contained system (moments $\leq K$) is underdetermined. The system we use instead adds higher-$|\beta|$ equations and truncates the moments above the tracked degree. This is a projection onto the consistency manifold of the (λ, m)

map, not an exact inversion. After propagation and measurement update, the moments being projected are no longer exactly on that manifold, and the projection can move the recovered moments away from the physically propagated ones—even when the fitted density looks good. The approximation is accurate in centered coordinates on $\mathbb{R}^n$ (Duffing: residual $\sim 10^{-5}$–10^{-9}), but on manifold state spaces with truncated product algebras (e.g., the trigonometric ring for SE(2)), the product truncation introduces additional error. The iterative SM-Stein consistency refinement (Appendix E) reduces this truncation error empirically by 2–3 orders of magnitude, but a formal convergence proof for the fixed-point iteration remains open.

(vi) *Generator-aware active closure.* The n=12–20 coupled oscillator experiments exploit the generator: only the degree-five moments produced by the local quadratic force need to be closed. This active closure is structured in a useful way, because it is tied directly to the moment ODE, but it does not amount to a general chordal or banded sparse moment method. A system whose generator requests a dense set of high-degree moments may still require larger r, extended Stein equations, or a genuinely sparse basis/closure construction. This is where a true sparse moment method would enter beyond the present n=20 demonstration. We view this as a promising direction for future work building on the SKF framework developed here.

(vii) *Model-based filtering under misspecification.* Like MEM-KF and other model-based nonlinear filters, the SKF assumes that the dynamics, measurement model, and noise statistics are accurate enough for moment propagation to be useful. The contribution here is to remove the partition-function bottleneck in MaxEnt reconstruction. It does not by itself solve model misspecification or partially known state-space models.

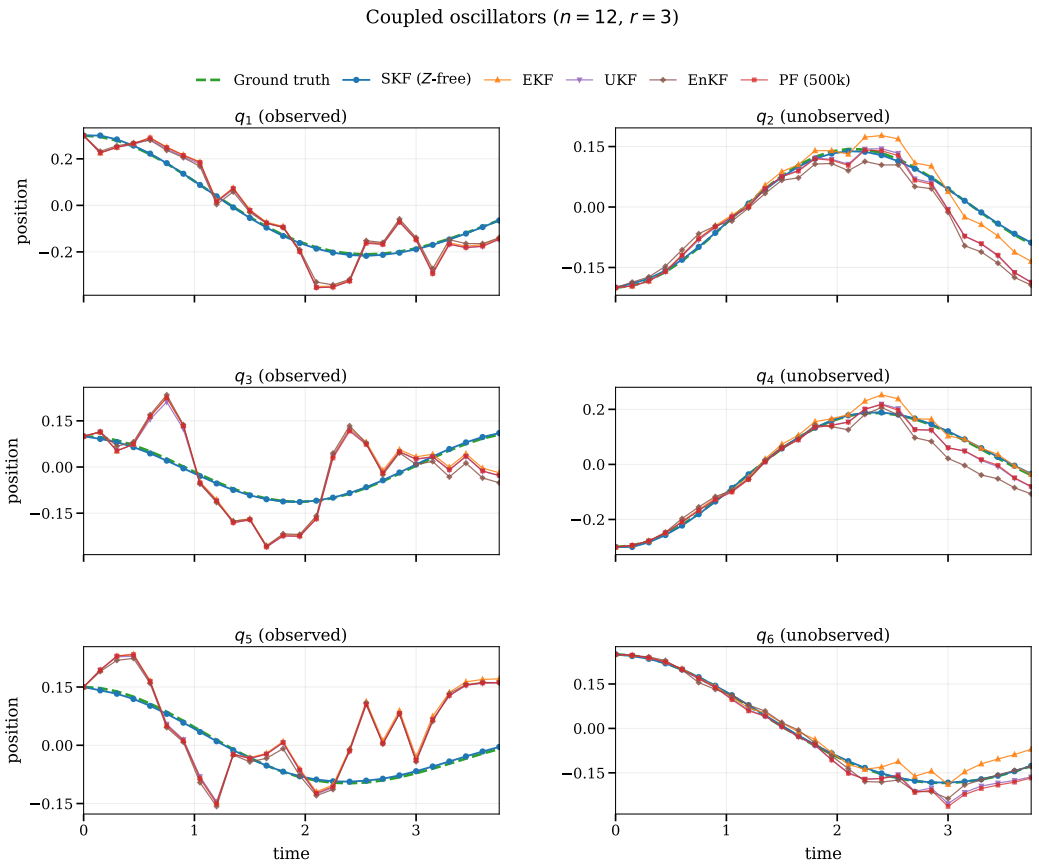


Figure A18: Coupled oscillators (n=12, r=3, active Stein closure). RMSE averaged over 25 steps for the displayed seed: SKF 0.0040, EKF 0.0696, UKF 0.0718, EnKF 0.0756, PF 0.0731.

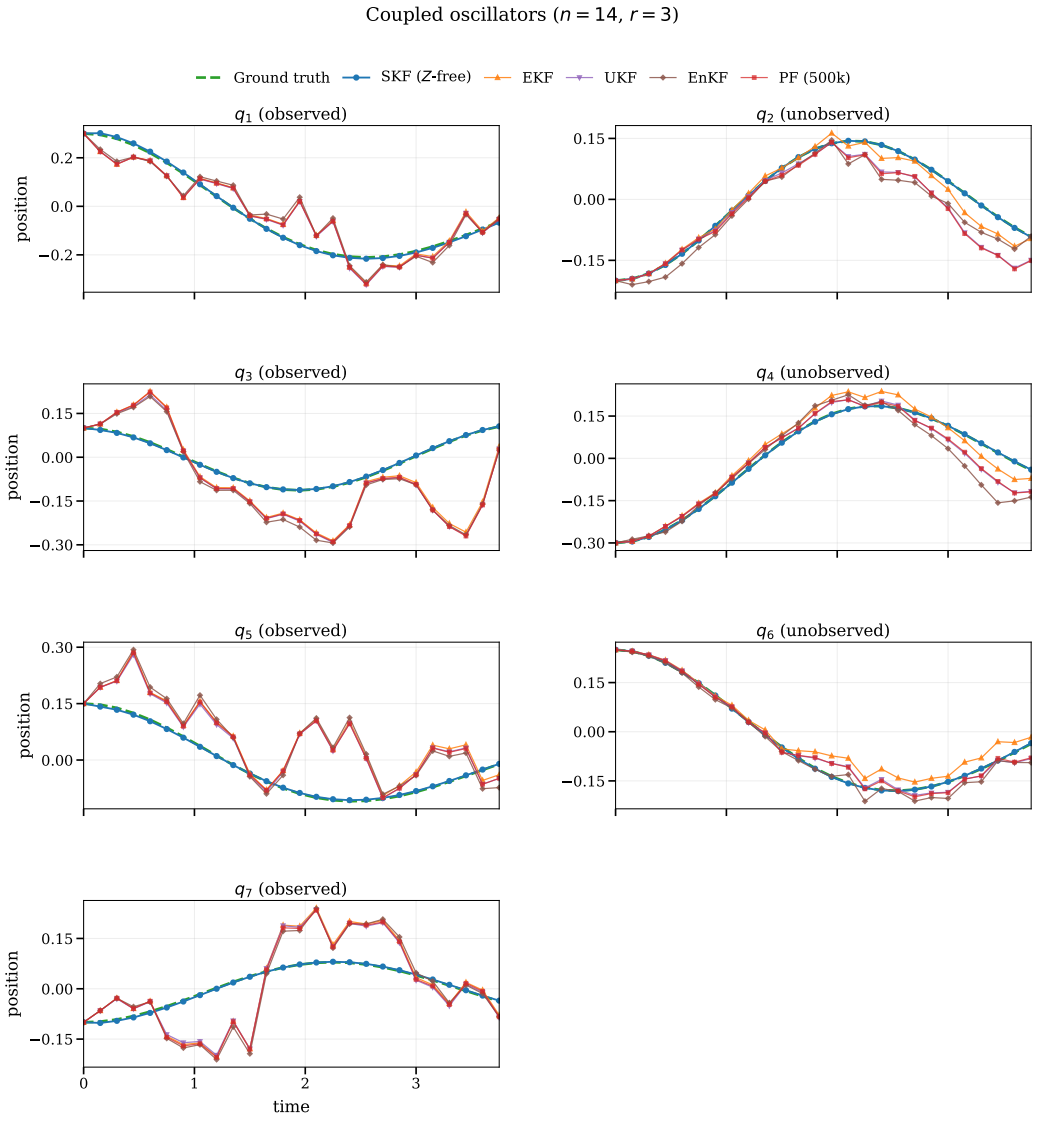


Figure A19: Coupled oscillators (n=14, r=3, active Stein closure). RMSE averaged over 25 steps for the displayed seed: SKF 0.0024, EKF 0.0814, UKF 0.0836, EnKF 0.0880, PF 0.0843.

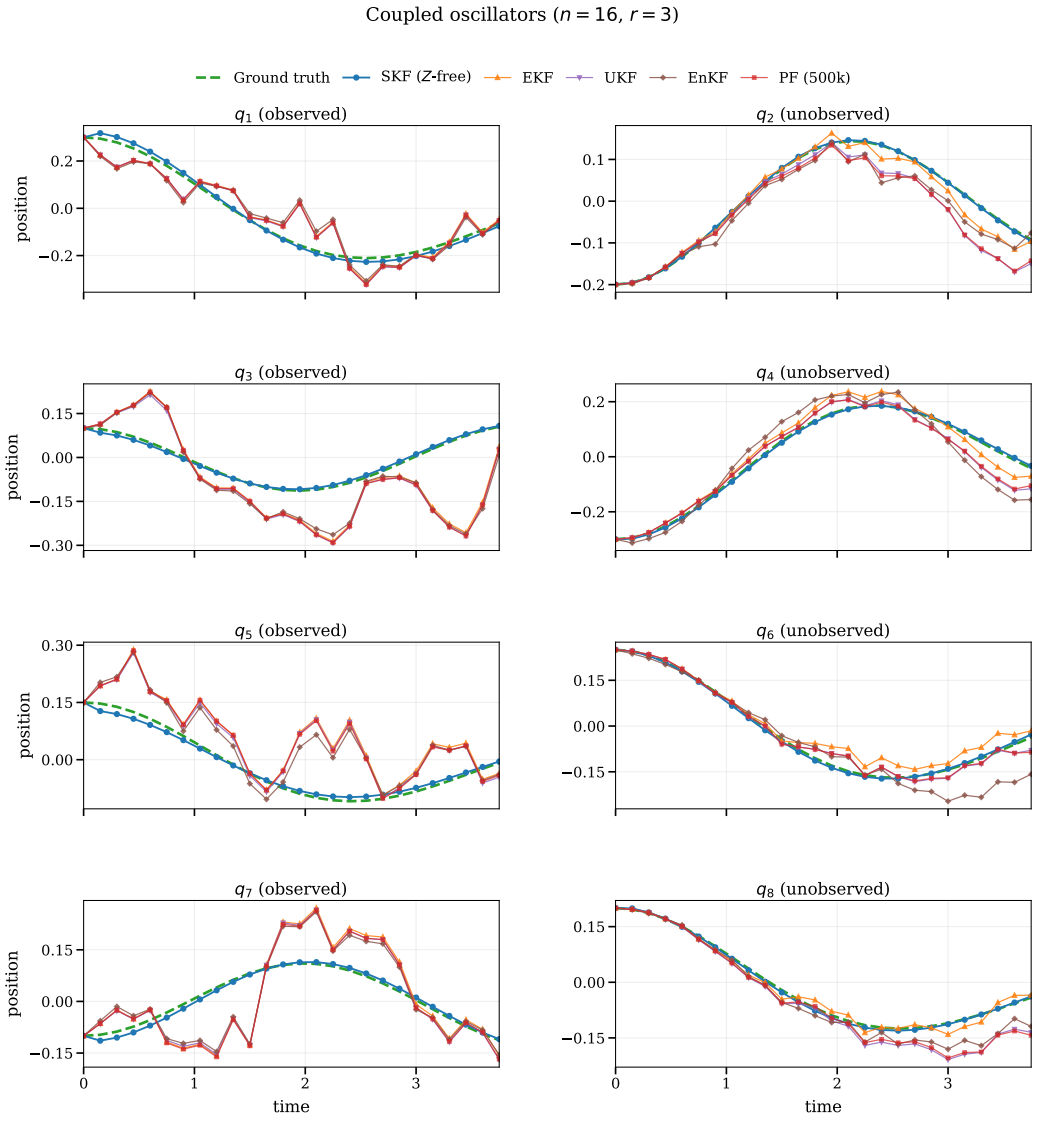


Figure A20: Coupled oscillators (n=16, r=3, active Stein closure). RMSE averaged over 25 steps for the displayed seed: SKF 0.0085, EKF 0.0763, UKF 0.0795, EnKF 0.0845, PF 0.0796.

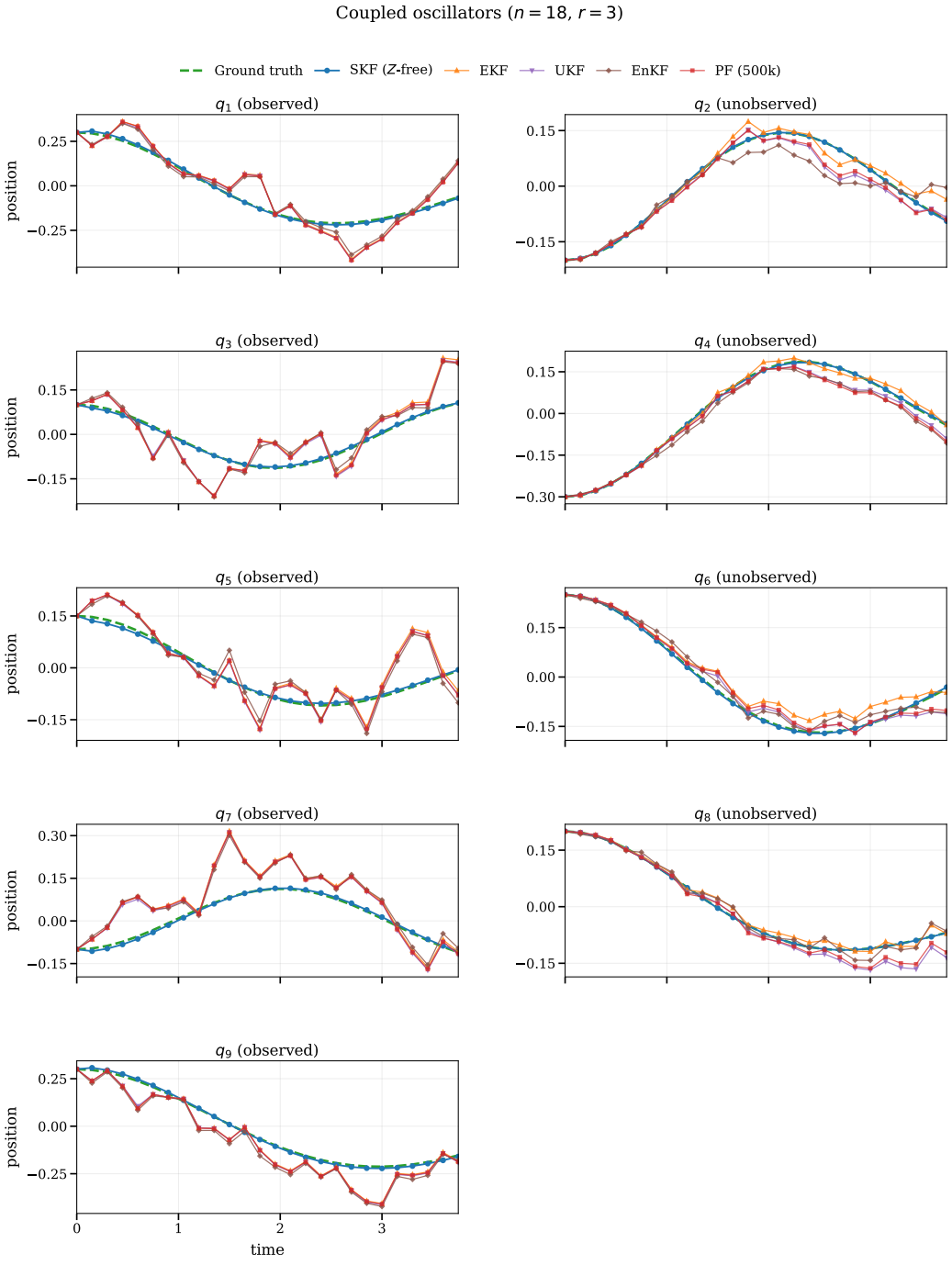


Figure A21: Coupled oscillators (n=18, r=3, active Stein closure). RMSE averaged over 25 steps for the displayed seed: SKF 0.0052, EKF 0.0646, UKF 0.0660, EnKF 0.0673, PF 0.0668.

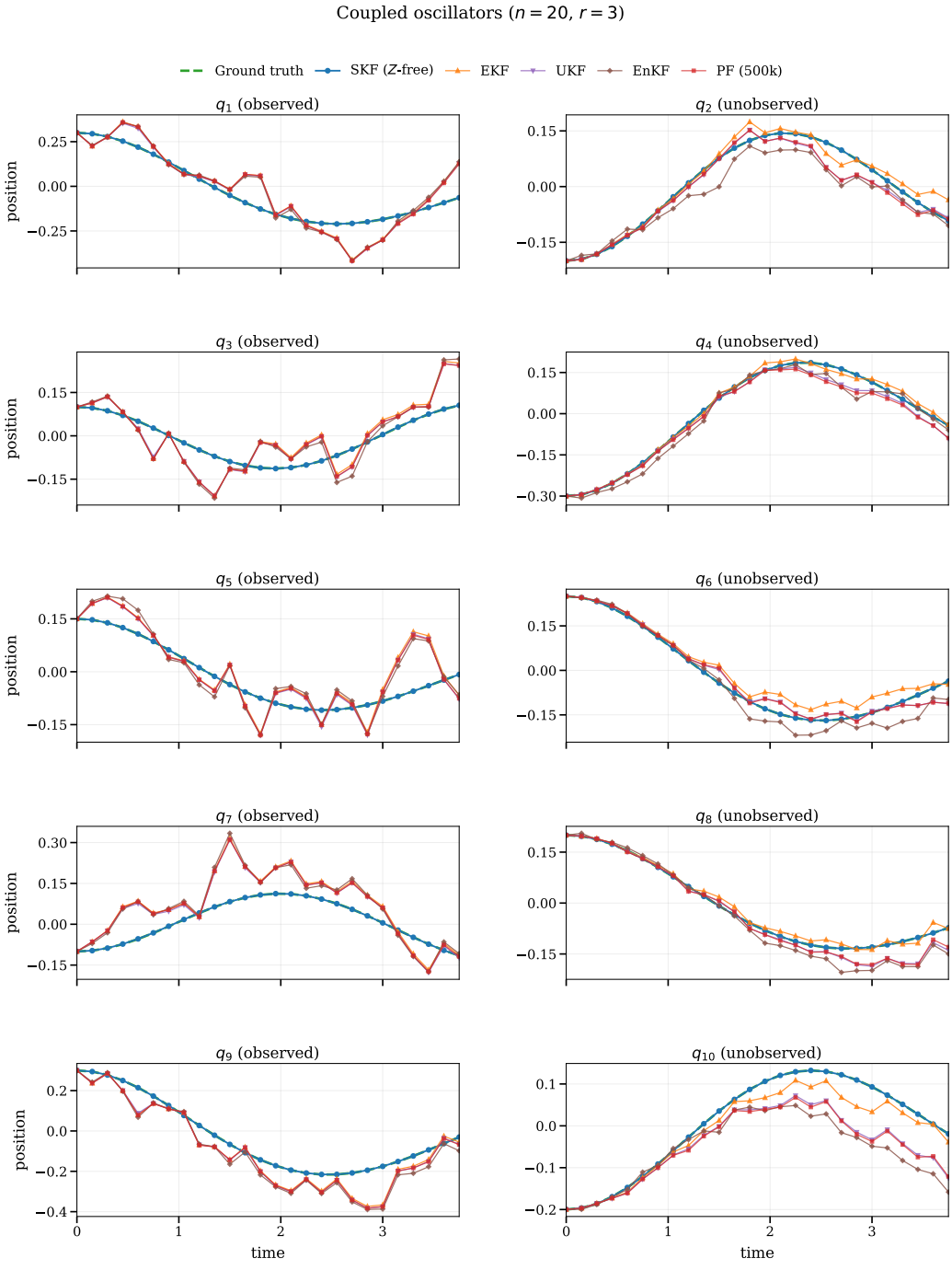


Figure A22: Coupled oscillators (n=20, r=3, active Stein closure). RMSE averaged over 25 steps for the displayed seed: SKF 0.0002, EKF 0.0617, UKF 0.0644, EnKF 0.0680, PF 0.0650.

(viii) *Polynomial families favor moderate-order structure.* The polynomial exponential family is most effective when the filtering density is reasonably captured by finitely many moderate-order moments in the chosen basis. For broad heavy-tailed distributions, or beliefs requiring very high-order moments to represent accurately, richer bases or different density families may be more appropriate. In the settings considered here (nonlinear oscillators, predator-prey dynamics, fluid tracer advection, rigid body kinematics), the dominant challenge is multimodality and skewness rather than extreme tail behavior, which is why this family is a useful approximation. Broadening the scope to heavy-tailed or multi-scale distributions would require either different sufficient statistics or a different density family.

(ix) *Boundaries and hybrid jumps reintroduce density values.* The present work addresses SDEs on open domains without boundaries. More generally, moment propagation via Dynkin's formula can involve additional integral terms (e.g., boundary flux in guard-triggered hybrid systems, codimension-1 surface integrals for reflecting boundaries) where the probability current across the boundary requires pointwise density evaluation (not just the score) to ensure mass conservation. This reintroduces the partition function Z, and with it the obstacle that SKF was designed to avoid. Extending the method to stochastic hybrid systems or domains with nontrivial boundary conditions therefore requires new boundary-aware closure machinery.

(x) *Modality transitions for single-component closures when $\bar{d} \geq 2$.*

A key open case is filtering through changes in modality. The score-matching step can still be well behaved, but the propagated moments must first survive the closure. For filtering, this issue is most relevant when prediction must carry a multimodal belief across a long, noisy, or partially observed interval. Frequent informative measurements can collapse the posterior before the closure error dominates. If those moments are already consistent with the polynomial exponential family, Theorem 1 recovers the corresponding parameters. The trouble begins earlier, inside the closure used to propagate those moments. When $\bar{d} \geq 2$, Stein closure has to resolve more than one layer of missing moments. A single component that moves from one mode to two modes also has to pass through a degenerate geometry where the parameter-to-moment map is badly conditioned. In that regime the multi-layer Stein backsolve can amplify small errors until the closed moment ODE becomes unusable. The double-well example separates representation from propagation. The polynomial MED can still *represent* the multimodal density (Figure A9: r=6, $\kappa(A)$=42, all four modes recovered), but the single-component closure does not propagate cleanly through the modal split.

Bifurcation surface. As an illustrating example, we use one-dimensional systems to show this obstruction. Consider the polynomial exponential family $p(x;\lambda) \propto \exp(-E(x;\lambda))$ with $E = \sum_{k=1}^{2r} \lambda_k x^k$ and $\lambda_{2r} > 0$ (normalizability). The derivative $E'(x) = \sum_{k=1}^{2r} k\lambda_k x^{k-1}$ is a polynomial of degree $2r-1$ in x whose coefficients depend on λ. The density p is unimodal when E has a single local minimum, i.e., E' has exactly one real root, and bimodal when E has two local minima separated by a local maximum, i.e., E' has at least three real roots. The number of distinct real roots of E' is locally constant on each connected component of the complement of the *bifurcation surface* $\mathcal{B} := \{\lambda \in \mathbb{R}^{2r} : \lambda_{2r} > 0,\ \mathrm{disc}(E') = 0\}$, where $\mathrm{disc}(E')$ is the discriminant of E'. It vanishes exactly when E' has a repeated root. Since $\mathrm{disc}(E')$ is a polynomial in λ, $\mathcal{B}$ is a codimension-1 algebraic variety. Thus any continuous path in λ-space from a unimodal to a bimodal configuration has to cross $\mathcal{B}$.

Two-layer Stein closure and its sensitivity. For the sensitivity calculation, apply the Stein identity (Proposition 2) to the test function x^k. For each $k \geq 1$,

$$\textstyle\sum_{j=1}^{2r} j\lambda_j\, m_{k+j-1} = k\, m_{k-1}. \tag{A.79}$$

If moments are truncated at degree $K = 2r-2$, the unclosed moments m_{K+1} and m_{K+2} are resolved in two layers. Layer 1 sets $k = K-2r+2$ in (A.79) and isolates the leading term $2r\lambda_{2r}\, m_{K+1}$:

$$m_{K+1} = \frac{(K-2r+2)\, m_{K-2r+1} - \sum_{j=1}^{2r-1} j\lambda_j\, m_{K-2r+j+1}}{2r\, \lambda_{2r}}. \tag{A.80}$$

Layer 2 sets $k = K-2r+3$ and isolates m_{K+2}. The resulting equation contains the layer 1 quantity m_{K+1} with coefficient $(2r-1)\lambda_{2r-1}$. Substituting (A.80) yields the composed map $\widetilde{m}_{K+2}(\lambda, m_{\leq K})$ with sensitivity

$$\frac{\partial \widetilde{m}_{K+2}}{\partial m_l} = O\Big(\frac{1}{\lambda_{2r}}\Big) + O\Big(\frac{\lambda_{2r-1}}{\lambda_{2r}^2}\Big), \qquad l \leq K. \tag{A.81}$$

The second term is the fragile one: layer 2 inherits the $1/\lambda_{2r}$ amplification from layer 1 and then divides by λ_{2r} again. For a single-layer closure ($\bar{d} = 1$), only the first term is present and remains bounded as long as $\lambda_{2r} > 0$. For two layers ($\bar{d} = 2$), the cross-term $\lambda_{2r-1}/\lambda_{2r}^2$ can grow without bound as $\lambda(t)$ evolves toward $\mathcal{B}$.

Stiffness and spurious finite-time blowup. Specializing the separable double-well moment equation (A.46) to one coordinate gives the cubic double-well $dX = (X - X^3)\,dt + \sigma\,dW$ ($\bar{d} = 2$), for which Dynkin's formula gives

$$\dot{m}_K = K m_K - K\,\widetilde{m}_{K+2}(\lambda, m_{\le K}) + \frac{\sigma^2}{2} K(K-1)\,m_{K-2}. \tag{A.82}$$

The Jacobian $J = \partial\dot{m}/\partial m$ inherits the $O(\lambda_{2r-1}/\lambda_{2r}^2)$ entries from (A.81) through the $-K\,\widetilde{m}_{K+2}$ coupling. As the density evolves from unimodal toward bimodal, the score-matching fit tracks a path $\lambda(t)$ moving toward $\mathcal{B}$, and the spectral radius $\rho(J)$ grows rapidly. Once $\rho(J)$ exceeds the stability boundary of the integrator, the numerical solution diverges. The observed divergence time t^* is mainly determined by the path of $\lambda(t)$ in parameter space, not by the step size Δt. It is important to note that this is the signature of a singularity in the closed moment ODE rather than an ordinary time-discretization artifact. The true density does not have this singularity. The additive noise makes the Fokker–Planck equation uniformly parabolic, and the cubic drift is dissipative at infinity, so standard Lyapunov estimates give a non-explosive solution with finite polynomial moments for all finite times. The blowup is therefore an artifact of the closure, not a physical singularity.

Numerical evidence. On the 2D double-well (Appendix G.5, Figure A9), the closed moment ODE diverges at $t^* \approx 1.1$ s. Refining Δt by $16\times$ shifts t^* by less than 5%, which rules out a plain integration error as the main explanation. Increasing the truncation order accelerates the blowup (t^*=2.9 for r=4, t^*=0.54 for r=12), because higher K exposes more eigendirections of J to the $1/\lambda_{2r}^2$ amplification. The score matching Gram matrix remains well-conditioned ($\kappa(A) \approx 1$) throughout, confirming the instability is in the closure.

Universality. This bifurcation picture is not special to the linear parameterization $E = \lambda^\top \phi$. In a smooth finite-dimensional family $p(x;\theta)$, unimodal and bimodal regions are generically separated by a codimension-1 boundary in parameter space, which is the usual singularity-theory picture [51]. The specific amplification mechanism ($1/\lambda_{2r}^2$ from two-layer Stein closure) is particular to the polynomial exponential family. The broader geometric obstruction is that a single component has to pass through a root-merging event in order to change modality.

Possible resolutions. The same example also points to a useful path forward. The issue is not that moments stop carrying information in highly nonlinear systems. Rather, the exact Stein backsolve can be too rigid as a closure rule. This makes it natural to treat closure as a small optimal control problem in moment space. We only borrow standard control ingredients here: linear-quadratic optimal control for local feedback design [52], constrained MPC for receding-horizon optimization [53], and CBF-QPs for enforcing forward-invariance constraints through input inequalities [54]. Note that in our setting the input is not a physical actuator but is the unresolved moment, chosen to keep the retained moment trajectory realizable and stable.

The resulting closure architecture has three parts. First, choose a retained moment vector x and treat the first unresolved moment block as an input u, so the truncated Dynkin hierarchy has the control-affine form

$$\dot{x} = F(x) + G(x)u. \tag{A.83}$$

Second, restrict u by moment-realizability constraints, such as Hankel positive semidefiniteness. Third, use barrier inequalities to keep the retained trajectory inside the feasible moment cone. If these constraints leave a small amount of freedom, that freedom must be fixed by system-specific structure such as the invariant law. This is not a free lunch: the control architecture must know something about the system, just as the higher-order moments in the moment ODE carry physical information about that system. The point is that such structure can be analytic or physics-informed, and need not come from particles. In the example below, the anchor is the known invariant density of the one-dimensional double well. Specializing the separable generator (A.45) to one coordinate, consider the factor

$$dX = (X - X^3)\,dt + \sigma\,dW, \qquad \sigma = 0.5,$$

let $s_p = \mathbb{E}[X^{2p}]$ denote the even moments, set the retained state to $x = (s_1, s_2, s_3)$, and use the unclosed eighth moment $u = s_4 = m_8$ as the control input. The retained hierarchy is affine in u:

$$\dot{s}_p = 2p\,s_p - 2p\,s_{p+1} + \sigma^2 p(2p-1)s_{p-1}, \qquad p = 1, 2, 3. \tag{A.84}$$

Equivalently, with $s_0 = 1$ absorbed into the constant term, this is the standard affine control system

$$\dot{x} = Ax + Bu + c,$$
$$A = \begin{bmatrix} 2 & -2 & 0 \\ 6\sigma^2 & 4 & -4 \\ 0 & 15\sigma^2 & 6 \end{bmatrix}, \quad B = \begin{bmatrix} 0 \\ 0 \\ -6 \end{bmatrix}, \quad c = \begin{bmatrix} \sigma^2 \\ 0 \\ 0 \end{bmatrix}. \tag{A.85}$$

The pair (A, B) is controllable, since $\operatorname{rank}[B, AB, A^2B] = 3$. Thus, before imposing realizability constraints, the unresolved moment can move the retained moment dynamics. The only unusual part is what the input means: $u = m_8$ is supplied by the closure, not by a physical actuator. Instead of backsolving a second Stein layer, we require the next moment to keep the truncated Stieltjes sequence feasible. With $Y = X^2$, the even moments of X become moments of a nonnegative variable: $(1, s_1, s_2, s_3, u) = (\mathbb{E}[Y^0], \ldots, \mathbb{E}[Y^4])$. We therefore impose the Hankel condition

$$\begin{bmatrix} 1 & s_1 & s_2 \\ s_1 & s_2 & s_3 \\ s_2 & s_3 & u \end{bmatrix} \succeq 0 \tag{A.86}$$

which gives the Schur-complement Hankel lower bound $u \geq L_H(x)$. For the lower-order minor, we use the moment-cone barrier $g(x) = s_1 s_3 - s_2^2$ and impose the control-barrier inequality

$$\dot{g}(x, u) + \kappa_{\text{cbf}} g(x) \geq 0, \qquad \kappa_{\text{cbf}} > 0, \tag{A.87}$$

where κ_{cbf} is the barrier rate and $\dot{g}(x, u) = \nabla g(x) \cdot (Ax + Bu + c)$. This gives an upper bound $u \leq U_{\text{CBF}}(x)$, the CBF upper bound. We then use the one-step constrained closure

$$u(x) = L_H(x) + \theta_\infty(\kappa_{\text{cbf}})\big(U_{\text{CBF}}(x) - L_H(x)\big). \tag{A.88}$$

This last scalar is the system-specific part of the construction: $\theta_\infty(\kappa_{\text{cbf}}) \in [0, 1]$, the stationary calibration fraction, is calibrated from the known invariant density, not from rollout samples. Thus the closure is physics-informed rather than sample-informed. The one-dimensional double well has $p_\infty(x) \propto \exp((x^2 - x^4/2)/\sigma^2)$, so we compute its stationary moments by deterministic quadrature and choose θ_∞ so that (A.88) is exact at that stationary point. For $\kappa_{\text{cbf}} = 6$, this gives $\theta_\infty = 0.3496$.

Figure A23 shows the closure tracking the MC reference moments through $T = 3$ without using MC moments in the closure. Figure A24 reconstructs the 2D density at $t = 0, 1, 2, 3$ from the same particle-free moments and recovers the transition to the four-modal structure.

This is not a general solution to the two-layer closure problem yet, but it is concrete evidence that the modality-transition failure is not fatal to a moment and score-matching filter. Score matching can still supply the density representation. The closure step just needs a stabilizing layer that respects the deterministic moment cone. The lesson is to treat Stein consistency as a soft preference or residual, while realizability and barrier inequalities define the admissible closure.

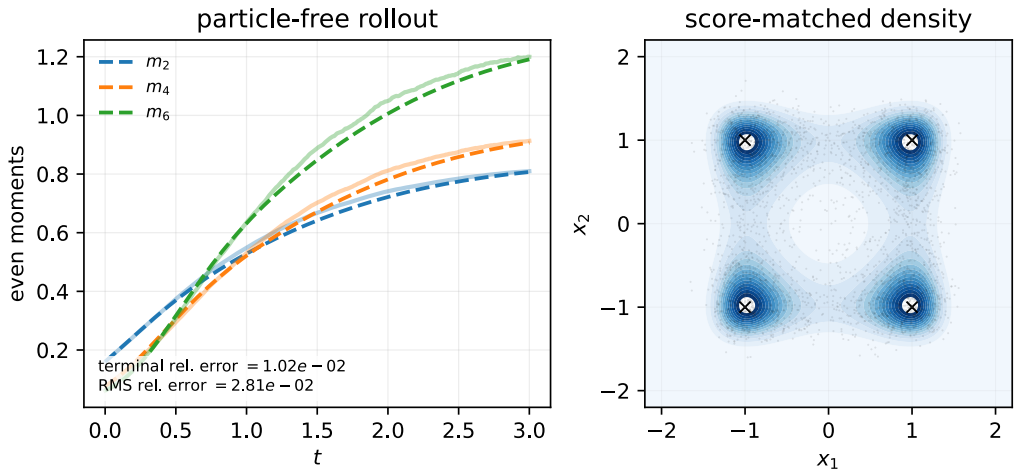


Figure A23: Particle-free constrained closure on the double-well test case. Left: even moments of the one-dimensional factor through $T = 3$. Solid faint curves are MC evaluation moments. Dashed curves are the cone-CBF closure from (A.88), which does not use MC moments. Right: score-matched density for the separable 2D double well, obtained from the product of the two particle-free marginal reconstructions. MC samples are shown only as a reference. The rollout has terminal relative error 1.0×10^{-2} and RMS relative trajectory error 2.8×10^{-2} for (m_2, m_4, m_6).

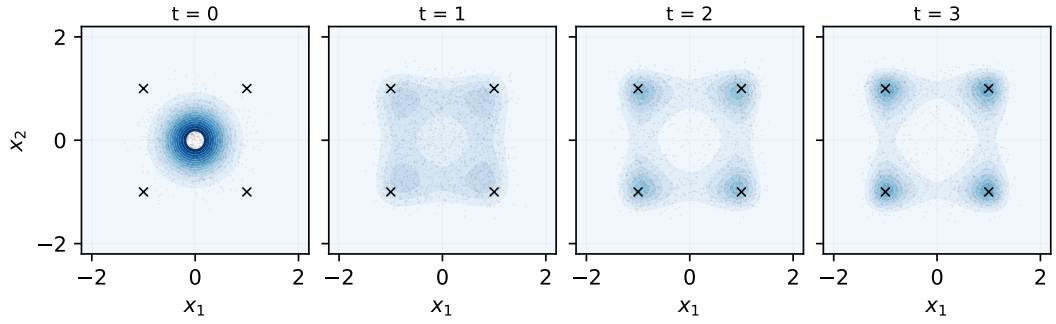


Figure A24: Particle-free density evolution on the separable double-well test case. Each panel shows the 2D score-matched density reconstructed from the propagated moment sequence at the indicated time. MC samples are shown only as a visual reference. The closure captures the transition from the initial central mass to the four-well structure through $T = 3$ without using particle moments.

M Broader impacts

The Score Kalman Filter is a general-purpose nonlinear estimation algorithm for stochastic dynamical systems. Its primary positive impact is computational. By replacing the partition-function evaluation of maximum-entropy density reconstruction with a linear solve, the SKF makes high-order moment-based filtering tractable on commodity hardware, which broadens access to uncertainty quantification for applications that have until now relied on Gaussian-only or particle-based filters. Likely beneficiaries are robotics and autonomous systems, scientific data assimilation, and downstream tasks in estimation-aware control where calibrated higher moments matter.

As with any improvement to nonlinear state estimation, the same machinery could be deployed in surveillance or tracking pipelines whose societal use is contested. We do not view these dual-use concerns as unique to our method, since they apply broadly to the nonlinear filtering literature, but we acknowledge them as a negative impact. The paper releases no pretrained models, datasets, or other artifacts with elevated risk of misuse, and all experiments use synthetic dynamical systems.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction frame the claims as algorithmic contributions for polynomial moment-based filtering: score-matching reconstruction, Stein closure and recovery, the SKF algorithm and experiments, and the $r{=}2$ Kalman-filter special case. These claims are tied to the stated propositions/theorem and to the synthetic benchmark results in Section 7 and Table A1, with assumptions and limitations discussed in the conclusion and Appendix L.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The conclusion points to Appendix L, which gives the detailed limitations. These include off-model score matching, hand-tuned basis/preconditioning choices, incomplete Lie-group extensions beyond compact groups, finite Stein-system rank and conditioning, truncation in posterior moment recovery, the generator-specific nature of active closure, model misspecification, heavy-tailed or high-order densities, boundary/hybrid systems, and modality transitions for $\bar{d} \geq 2$.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.

- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: Proposition 1 (score matching linear system) is proved in Appendix B.1. Theorem 1 (parameter recovery) states all assumptions (moment consistency, positive definiteness of A) and is proved in Appendix B.3. Proposition 2 (Stein identity) is derived in Appendix B. Proposition 3 (moment perturbation bound) is proved in Appendix B.2. Proposition 4 (generic well-posedness of the directional Stein closure) gives the polynomial determinant argument in Appendix B.5.

   Guidelines:

   - The answer [N/A] means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: All system parameters, basis orders, time steps, and observation models are specified in Section 7 and Appendices G and J. The generator calculations (Appendix G) and matrix assembly formulas (Appendix H) provide the complete mathematical specification needed to reimplement the method. Algorithm 1 gives the full pseudocode.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.

- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: An anonymized code release is provided as supplementary material. It includes a README, requirements file, license, a `reproduce.sh` driver, and experiment scripts for the synthetic benchmarks and figures. The scripts generate their data from specified synthetic systems and fixed seeds; no external datasets, pretrained models, or private assets are required.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: All system parameters, basis orders, time steps, observation models, and initial conditions are given in Section 7 and Appendix J, with generator details in Appendix G. No training or hyperparameter tuning is involved. The method is purely model-based with the basis order r as the only user-specified parameter.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: For the coupled oscillator filter experiments, Table A1 reports mean RMSE $\pm$ standard deviation over 10 independent random seeds (different process noise and measurement noise realizations). The prediction experiments use deterministic moment propagation validated against 10^6-particle MC references. The standard deviation captures variability due to the random trajectory realization.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
   - If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Section 7 states all experiments are run in Python (NumPy/SciPy) on a single laptop CPU (Intel i9-11900H, no GPU). Per-experiment timings are reported in Section 7.1 (0.6 s), Section 7.2 (0.7 s), and Table A1 includes cost per step for all filter experiments.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics https://neurips.cc/public/EthicsGuidelines?

   Answer: [Yes]

   Justification: This is a theoretical and algorithmic contribution to nonlinear filtering. It does not involve human subjects or personal data, and the general dual-use concerns for state-estimation methods are acknowledged in the broader-impact discussion.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: Broader impacts are discussed in Appendix M. The SKF is a general-purpose estimation algorithm for stochastic dynamical systems whose primary impact is computational, broadening access to high-order moment-based filtering for robotics, control, and scientific data assimilation. We acknowledge that any improvement in state estimation can in principle support surveillance applications, but these dual-use concerns are not unique to our method.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

Justification: The paper does not release pre-trained models, datasets, or any artifacts with risk for misuse. It presents an algorithmic framework for nonlinear filtering.

Guidelines:

- The answer [N/A] means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [N/A]

Justification: The paper does not use external code, datasets, or pre-trained models. All experiments are on synthetic dynamical systems implemented by the authors.

Guidelines:

- The answer [N/A] means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [Yes]

Justification: The new asset is the anonymized code package submitted with the paper. Its README documents installation, commands for reproducing results, generated synthetic data, and expected outputs; the package also includes a license and dependency specification.

Guidelines:

- The answer [N/A] means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing or research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: The paper does not involve research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: LLMs were not used as an important, original, or non-standard component of the core methodology. Any assistance was limited to manuscript preparation and routine coding support, and did not determine the algorithm, experiments, results, or claims.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.