Information geometry of physics-informed statistical manifolds and its use in data assimilation

F. Boso and D.M. Tartakovsky*

Department of Energy Resources Engineering, Stanford University, Stanford, CA 94305, USA

Abstract

The data-aware method of distributions (DAMD) is a low-dimension data assimilation procedure to forecast the behavior of dynamical systems described by differential equations. The core of DAMD is the minimization of a distance between an observation and a prediction in distributional terms, with prior and posterior distributions constrained on a statistical manifold defined by the method of distributions (MD). We leverage the information-geometric properties of the statistical manifold to reduce predictive uncertainty via data assimilation. Specifically, we exploit the information-geometric structures induced by two discrepancy metrics, the Kullback-Leibler divergence and the Wasserstein distance, which explicitly yield natural gradient descent. The use of a deep neural network as a surrogate model for MD enables automatic differentiation, further accelerating optimization. The manifold's geometry is quantified without sampling, yielding an accurate approximation of the gradient descent direction. Our numerical experiments demonstrate that accounting for the manifold's geometry significantly reduces the computational cost of data assimilation by both facilitating the calculation of gradients and reducing the number of required iterations.

Keywords: method of distributions, Bayesian inference, parameter identification, machine learning

*Corresponding author

Email address: {fboso,tartakovsky}@stanford.com (F. Boso and D.M. Tartakovsky)

1. Introduction

Mathematical models used to represent "reality" are invariably faulty due to a number of mutually reinforcing reasons such as lack of detailed knowledge of the relevant laws of nature, scarcity (in quality and/or quantity) of observations, and inherent spatiotemporal variability of the coefficients used in their parameterizations. Consequently, model predictions must be accompanied by a quantifiable measure of predictive uncertainty (e.g., error bars or confidence intervals); when available, observations should be used to reduce this uncertainty. The probabilistic framework provides a natural means to achieve both goals. For example, a random forcing in Langevin (stochastic ordinary-differential) equations [1] or fluctuating Navier-Stokes (stochastic partial-differential) equations [2] implicitly account for sub-scale variability and processes that are otherwise absent in the underlying model.

Solutions of such stochastic models, and of models with random coefficients, are given in terms of the (joint) probability density function (PDF) or cumulative distribution function (CDF) of the system state(s). They can be computed, with various degrees of accuracy and ranges of applicability, by employing, e.g., Monte Carlo simulations (MCS), polynomial chaos expansions (PCE) and the method of distributions (MD) [3]. MCS are robust, straightforward and trivially parallelizable; yet, they carry (often prohibitively) high computational cost. PCE rely on a finite-dimensional expansion of the solution of a stochastic model; their accuracy and computational efficiency decrease as the correlation length of the random inputs decreases (the so-called curse of dimensionality), making them ill-suited to problems with white noise [4]. MD yields a (generally approximate) partial differential equation (PDE) for the PDF or CDF of a system state (henceforth referred to as a PDF/CDF equation). MD can handle inputs with both long and short correlations, although the correlation length might affect the robustness of the underlying closure approximations when the latter are needed. For Langevin systems driven by white noise, MD yields a Fokker-Planck equation [1] for a system state's PDF. For colored (correlated) noise, PDF/CDF equations become approximate [5], although their computational footprint typically does not change. If a Langevin system is characterized by $N_{\rm st}$ system states, then PDF/CDF equations are defined in an augmented $N_{\rm st}$-dimensional space. Their MD-based derivation requires a closure approximation [3, and references therein] such as the semi-local closure [6, 7, 8] used in our analysis because of its accuracy and manageable computational cost.

The temporal evolution of the PDF of a system state predicted with, e.g., MD provides a measure of the model's predictive uncertainty in the absence of observations of the system state. In Bayesian statistics, this PDF serves as a prior that can be improved (converted into the posterior PDF) via Bayesian update as data become available. When used in combination with ensemble methods like MCS, standard strategies for Bayesian data assimilation, e.g., Markov chain Monte Carlo (MCMC) and its variants, are often prohibitively expensive [9]. The computational expedience is the primary reason for the widespread use of various flavors of Kalman filter; they perform well when the system state's PDF is Gaussian and models are linear, but require adjustments and uncontrollable approximations otherwise [10, 9]. Data-aware MD (DAMD) [11] alleviates this computational bottleneck, rendering Bayesian update feasible even on a laptop. DAMD employs MD to propagate the system state PDF (forecast step) and sequential Bayesian update at measurement locations to assimilate data (analysis step). It offers two major benefits. First, MD replaces repeated model runs, characteristic of both MCMC and ensemble and particle filters, with the solution of a single deterministic equation for the evolving PDF. Second, it dramatically reduces the dimensionality of the PDFs involved in the Bayesian update at each assimilation step because it relies on a single-point PDF rather than a multi-point PDF whose dimensionality is determined by the discretized state being updated. DAMD takes advantage of MD's ability to handle nonlinear models and non-Gaussian distributions [12, 13].

DAMD recasts data assimilation as a minimization problem, whose loss function represents the discrepancy between observed and predicted posterior distributions. The observed posterior PDF is obtained by direct application of Bayes' rule at the measurement point, combining the data model and a prior PDF computed via MD. The predicted PDF is assumed to obey the PDF equation, which acts as a PDE constraint for the loss function. The parameters appearing in MD are the target of minimization and introduce a suitable parameterization for the space of probabilities (a statistical manifold) with quantifiable geometric properties. The computational effort of DAMD is thus determined by the efficiency in the solution of a minimization problem on a manifold. This aspect of DAMD is the central focus of our analysis, in which we exploit information-geometric theory to reformulate the optimization problem by relying on the geometric properties of the MD-defined manifold.

We utilize results from the optimal transport theory and machine learn-

ing. Specifically, we employ both the Kullback-Leibler (KL) divergence and the 2−Wasserstein distance to measure the discrepancy between predicted and observed posterior distributions at each assimilation point. The former underpins much of information theory [14] and variational inference [15],[1] while the latter has its origins in optimal transport and is now increasingly popular in the wider machine learning community [16]. We employ gradient descent (GD) and natural gradient descent (NGD) for optimization [17], with preconditioning matrices expressing the geometry induced on the statistical manifold by the choice of the discrepancy. These formulations are explicit for univariate distributions; thus, they ideally suit our data assimilation procedure.

Finally, we construct a surrogate of the PDF/CDF equation to accelerate sequential minimization of loss functions, taking advantage of the relatively small dimensionality of the statistical manifold. We identify a special architecture of a deep neural network (DNN) that enables the calculation of the terms involved in NGD for both discrepancy choices; its connection with the physical model, enabled by MD, classifies it as a physics-informed neural network (PINN) surrogate. The use of DNNs obviates the need to resort to sampling when assessing the manifold's geometry, a strategy that had a debatable success [18].

The paper is organized as follows. In section 2, we briefly overview the tools and concepts from information geometry and optimal transport that are directly relevant to the subsequent analysis. In section 3, we summarize the DAMD approach (with details in Appendix A) and illustrate how the information-geometric tools and MD can be naturally combined to reduce predictive uncertainty. Section 4 contains results of our numerical experiments conducted on a Langevin equation with either white or colored noise. Main conclusions drawn from this study are summarized in section 5.

2. Preliminaries

Let ${}_p\mathcal{P}(\mathbb{R}^d)$ denote the probability space of PDFs f on $\mathbb{R}^d$ with finite pth moments, where $p \geq 1$. Our key objective is to minimize loss functions

[1] Unlike traditional variational inference, our approach utilizes univariate (single-point) distributions that are characterized by a specific, physics-driven parameterization enabled by MD.

involving PDFs f belonging to ${}_p\mathcal{P}(\mathbb{R}^d)$. In this section, we summarize definitions, tools and theoretical results that will be subsequently used in concert with DAMD.

Measures of discrepancy. Alongside classic measures of discrepancy between generic integrable functions $f_1(\mathbf{X}), f_2(\mathbf{X}) : \mathbb{R}^d \to \mathbb{R}^+$ such as the L_1 and L_2 norms,

$$d_1(f_1, f_2) := \int_{\mathbb{R}^d} |f_1(\mathbf{X}) - f_2(\mathbf{X})| \mathrm{d}\mathbf{X} \tag{1a}$$

and

$$d_2(f_1, f_2) := \left(\int_{\mathbb{R}^2} |f_1(\mathbf{X}) - f_2(\mathbf{X})|^2 \mathrm{d}\mathbf{X} \right)^{1/2}, \tag{1b}$$

we utilize measures of discrepancy that are tailored to the underlying geometry of the probabilistic space ${}_p\mathcal{P}(\mathbb{R}^d)$. The KL divergence,

$$d_{\mathrm{KL}}(f_1, f_2) := \int_{\mathbb{R}^d} f_1(\mathbf{X}) \ln \frac{f_1(\mathbf{X})}{f_2(\mathbf{X})} \mathrm{d}\mathbf{X}, \tag{2}$$

expresses the discrepancy between the PDFs f_1 and f_2 in terms of relative entropy. Used to quantify how well $f_1 : \mathbb{R}^d \to \mathbb{R}^+$ approximates $f_2 : \mathbb{R}^d \to \mathbb{R}^+$, the KL divergence is not a distance since $d_{\mathrm{KL}}(f_1, f_2) \neq d_{\mathrm{KL}}(f_2, f_1)$.

Another discrepancy measure is the p-Wasserstein distance,

$$W_p(f_1, f_2) := \left(\inf_{\gamma \in \Gamma(f_1, f_2)} \int_{\mathbb{R}^d \times \mathbb{R}^d} \|\mathbf{X} - \mathbf{Y}\|^p \gamma(\mathrm{d}\mathbf{X}, \mathrm{d}\mathbf{Y}) \right)^{1/p}, \quad p \geq 1, \tag{3}$$

where Γ is the set of joint probability measures γ on $\mathbb{R}^d \times \mathbb{R}^d$ whose marginals are probability measures corresponding to f_1 and f_2. Originating in the field of optimal transport, (3) quantifies the optimal (infimum) cost of shifting the mass distribution of f_1 to f_2. Such minimum exists and is unique under regularity conditions for the PDFs for $p > 1$, i.e., f must be absolutely continuous with respect to the Lebesgue measure [19]. For $d = 1$, (3) reduces to

$$W_p(f_1, f_2) = \|F_1^{-1}(Y) - F_2^{-1}(Y)\|_p = \left(\int_0^1 |F_1^{-1}(Y) - F_2^{-1}(Y)|^p \mathrm{d}Y \right)^{1/p}, \quad p \geq 1, \tag{4}$$

where $F_i(X) = \int_{-\infty}^{X} f_i(X)\mathrm{d}X$ with $i = 1, 2$ is the CDF corresponding to the PDF $f_i(X)$; and $F_i^{-1}(Y)$ is the inverse of F_i defined as $F_i^{-1}(Y) = \inf\{X : F_i(X) \geq Y\}$, with $Y \in (0, 1)$.

Since DAMD deals with univariate distributions, we are concerned with $d = 1$.

Approximation of distributions. Various fields of science and engineering—e.g., machine learning [20, 21], estimation theory [22], and optimal transport and control theory [19, 23, 24]—deal with a problem of approximating an (empirical) target PDF $\hat{f}(X)$ with a PDF $f(X; \boldsymbol{\varphi}) : \mathbb{R} \to \mathbb{R}^+$ defined on the parameterized probability space $\mathcal{P}_{\boldsymbol{\varphi}}$. The latter consists of PDFs that are uniquely characterized by a set of N_{par} parameters $\boldsymbol{\varphi} \in \Phi \subset \mathbb{R}^{N_{\text{par}}}$ with $N_{\text{par}} \geq 1$. This functional approximation is recast as a problem of finding a parameter set that minimizes a function $\mathcal{C}(\boldsymbol{\varphi})$ depending on a selected measure of discrepancy $D(\boldsymbol{\varphi})$ between the target PDF $\hat{f}(X)$ and its approximation $f(X; \boldsymbol{\varphi})$,

$$\underset{\boldsymbol{\varphi} \in \Phi}{\operatorname{argmin}}\, \mathcal{C}(D(\boldsymbol{\varphi})), \quad \text{with } D(\boldsymbol{\varphi}) = D(f(X; \boldsymbol{\varphi}), \hat{f}(X)), \tag{5}$$

with $f(X; \boldsymbol{\varphi})$ belonging to $\mathcal{P}_{\boldsymbol{\varphi}}$.

We assume $\mathcal{P}_{\boldsymbol{\varphi}}$ to be a subset of ${}_2\mathcal{P}(\mathbb{R})$. The use of the KL and W_2 metrics in place of D in (5) introduces known geometries to the statistical manifold of parameterized PDFs, facilitating the deployment of efficient optimization algorithms that exploit this geometric structure. Specifically, one of the geometric properties of the KL divergence is its parameterization invariance, i.e., the equivalency between computation of the discrepancy $\mathcal{C}(\boldsymbol{\varphi}) \equiv D(\boldsymbol{\varphi}) \equiv d_{\text{KL}}(f(X; \boldsymbol{\varphi}), \hat{f}(X))$ in the PDF space $\mathcal{P}_{\varphi}$ and in the parameter space Φ; this property facilitates minimization of the loss function via natural gradient descent [25, Sec. 2.1.3]. Moreover, a solution of the minimization problem (5) with $\mathcal{C}(\boldsymbol{\varphi}) \equiv d_{\text{KL}}(f(X; \boldsymbol{\varphi}), \hat{f}(X))$ corresponds to the maximum likelihood estimate of the parameters $\boldsymbol{\varphi}$ [26]. This analogy elucidates the connection between Bayesian inference and information geometry. When $\hat{f}$ is obtained empirically (e.g., from sampling or repeated experiments), the use of the Wasserstein distance, $\mathcal{C}(\boldsymbol{\varphi}) \equiv D^2(\boldsymbol{\varphi})/2 \equiv W_2^2(f(X; \boldsymbol{\varphi}), \hat{f}(X))/2$, is more computationally expedient [19, 20, 21, 23], while possessing geometric properties almost as rigorous as KL [17].

Statistical manifolds. Let the PDF $f(X; \boldsymbol{\varphi})$ be smooth and have a support $\Omega := \{X \in \mathbb{R} | f(X) > 0\}$. We assume this support to be com-

pact, $\Omega = [X_{\min}, X_{\max}] \subset \mathbb{R}$, and the dimensionality of the parameter space $\Phi \subset \mathbb{R}^{N_{\text{par}}}$ to be finite, $N_{\text{par}} < +\infty$. An N_{par}-dimensional manifold is an N_{par}-dimensional topological space that behaves locally like the Euclidean space $\mathbb{R}^{N_{\text{par}}}$. A smooth manifold is equipped with a metric tensor $\mathbf{G}(\boldsymbol{\varphi})$—which facilitates the calculation of distances on the local approximation of the manifold, i.e., the tangent plane—and an affine connection $\nabla_{\boldsymbol{\varphi}}$—which enables differentiation. The second-order tensor $\mathbf{G}$ is positive definite and varies smoothly with $\boldsymbol{\varphi}$. It defines a Riemannian metric on the manifold, and the latter is said to be Riemannian. A statistical manifold $\mathcal{M}$ is a manifold with coordinates $\boldsymbol{\varphi} = (\varphi^1, \dots, \varphi^{N_{\text{par}}}) \in \mathbb{R}^{N_{\text{par}}}$ where each point represents a PDF with assigned support and defined features. A divergence on the statistical manifold $\mathcal{M}$ is a non-negative function $D(f(X;\boldsymbol{\varphi}), f(X;\boldsymbol{\varphi}')) : \mathcal{M} \times \mathcal{M} \to \mathbb{R}^+$, which is equal to zero if and only if $f(X;\boldsymbol{\varphi}) \equiv f(X;\boldsymbol{\varphi}')$ and which can be approximated locally (i.e., when $\boldsymbol{\varphi}$ and $\boldsymbol{\varphi}'$ are close) via the components G_{ij} of the second-order tensor $\mathbf{G}$ as $D(f(\boldsymbol{\varphi}), f(\boldsymbol{\varphi}')) = G_{ij}(\boldsymbol{\varphi})\Delta\varphi^i\Delta\varphi^j/2 + \mathcal{O}(|\Delta\boldsymbol{\varphi}|^3)$, where $\Delta\boldsymbol{\varphi} = \boldsymbol{\varphi} - \boldsymbol{\varphi}'$ and the Einstein summation is implied over the repeated indices $i, j = 1, \dots, N_{\text{par}}$.

Information geometry of statistical manifolds. If the KL divergence is used to quantify the discrepancy between two PDFs on the manifold $\mathcal{M}$, then the tensor metric $\mathbf{G}(\boldsymbol{\varphi})$ (a geometric structure) of the space $\mathcal{P}_{\boldsymbol{\varphi}}$ of parameterized PDFs $f(X;\boldsymbol{\varphi})$ is called Fisher information matrix,

$$\mathbf{G}_F(\boldsymbol{\varphi}) = \int_\Omega \frac{1}{f(X;\boldsymbol{\varphi})} \left(\nabla_{\boldsymbol{\varphi}} f(X;\boldsymbol{\varphi})\right)^\top \nabla_{\boldsymbol{\varphi}} f(X;\boldsymbol{\varphi}) \mathrm{d}X, \tag{6}$$

which is emphasized by the subscript F. The resulting statistical manifold $\mathcal{M}$ is invariant in the sense that, for $\boldsymbol{\varphi}_i \in \Phi$ and $f_i \equiv f(\boldsymbol{\varphi}_i)$ with $i = 1, 2$, the divergence $d_{\text{KL}}(f_1, f_2) = d_{\text{KL}}(f_1(X;\boldsymbol{\varphi}_1), f_2(X;\boldsymbol{\varphi}_2))$ and the metric $\mathbf{G}_F$ describe the same geometry when the random coordinate X is remapped without losing information [27, Ch. 3.1]. This property underpins the Riemannian natural gradient descent (NGD) method (a.k.a. Fisher-Rao gradient descent) for parameter identification [28, and references therein]. The method uses the metric tensor $\mathbf{G}_F$ as a pre-conditioner for gradient descent algorithms to solve (5) with $\mathcal{C} \equiv D \equiv d_{\text{KL}}$,

$$\boldsymbol{\varphi}_{k+1} = \boldsymbol{\varphi}_k - \eta \mathbf{G}_F^{-1}(\boldsymbol{\varphi}_k) \nabla_{\boldsymbol{\varphi}} d_{\text{KL}}(f(X;\boldsymbol{\varphi}), \hat{f})|_{\boldsymbol{\varphi}_k}, \tag{7}$$

where η is the descent step and $\mathbf{G}_F^{-1}$ is the inverse of $\mathbf{G}_F$. The technique presents strong theoretical analogies with classic filtering techniques (namely

Kalman filter and extended Kalman filter) [29, 30]. In the absence of an analytical expression for $\mathbf{G}_F$, the matrix can be approximated empirically, although with debatable accuracy [18].

Geometric structure, including the metric tensor $\mathbf{G}_W(\boldsymbol{\varphi})$, of the finite-dimensional Wasserstein manifolds—hence, the subscript W— of Gaussian PDFs was studied in [31, 32]. These results were subsequently generalized to construct $\mathbf{G}_W(\boldsymbol{\varphi})$ for the manifolds $\mathcal{M}$ of generic discrete [28] and continuous [17] distributions. Specifically, when $d = 1$, the Wasserstein manifold's metric tensor $\mathbf{G}_W$ has an explicit form,

$$\mathbf{G}_W(\boldsymbol{\varphi}) = \int \frac{1}{f(X;\boldsymbol{\varphi})} \left(\nabla_{\boldsymbol{\varphi}} F(X;\boldsymbol{\varphi})\right)^\top \nabla_{\boldsymbol{\varphi}} F(X;\boldsymbol{\varphi}) \mathrm{d}X. \tag{8}$$

Under some mild regularity assumptions, the finite-dimensional Wasserstein manifold $\mathcal{M}$ in the parameter space Φ is Riemannian [17]. It introduces an NGD in the space Φ,

$$\boldsymbol{\varphi}_{k+1} = \boldsymbol{\varphi}_k - \eta \mathbf{G}_W^{-1}(\boldsymbol{\varphi}_k) \nabla_{\boldsymbol{\varphi}} \mathcal{C}(\boldsymbol{\varphi})|_{\boldsymbol{\varphi}_k}, \quad \text{with} \quad \mathcal{C} \equiv \frac{1}{2} D^2 \text{ and } D \equiv W_2. \tag{9}$$

Remark 2.1. *Regardless of whether one chooses the KL divergence or the W_2 distance, NGD orients the optimization problem (9) according to the topology of the statistical manifold $\mathcal{M}$ as expressed by its metric tensor $\mathbf{G}_i$ ($i = F$ or W), thus accelerating the solution. The computational cost of both (7) and (9) depends on the overall number of iterations and on the calculation of $\mathbf{G}_i$ (storage cost $\mathcal{O}(N_{par}^2)$ per iteration) and its inverse $\mathbf{G}_i^{-1}$ (inversion cost $\mathcal{O}(N_{par}^3)$ per iteration) [25]. Thus, the overall cost of optimization is a trade-off between the number of iterations, arguably reduced on information-geometric grounds, and the cost of inverting the metric tensor $\mathbf{G}_i$.*

Remark 2.2. *The finite-dimensional L_2-Wasserstein manifold $\mathcal{M}$ is not exactly geodesic (unless PDFs are Gaussian), and as such the geodesic distance on the manifold is not identical to W_2 [17]. As demonstrated by [17, Th. 1 and Prop. 6], the natural gradient trajectory approximates the geodesic distance up to second order information.*

Remark 2.3. *A unifying framework connecting the KL and W_2 metrics for manifolds of discrete distributions is proposed in [33, and references therein].*

Remark 2.4. *NGD provides an efficient alternative to stochastic gradient descent methods whenever the dimensionality of the parameter space is sufficiently small, thus permitting the calculation of the preconditioning matrices* $\mathbf{G}(\boldsymbol{\varphi})$ *[34].*

3. DAMD with PINN Surrogates

Consider a state variable $x(t) : \mathbb{R}^+ \to \mathbb{R}$, whose dynamics is governed by a stochastic/random ordinary differential equation (SDE)

$$\frac{\mathrm{d}x(t)}{\mathrm{d}t} = s(x(t); w(t), \boldsymbol{\theta}), \qquad t > 0; \tag{10a}$$

subject to a (possibly uncertain, i.e., random) initial condition

$$x(t = 0) = x_0, \qquad x_0 \in \mathbb{R}. \tag{10b}$$

The system is driven by the stationary (statistically homogeneous) random process $w(t)$ characterized by a single-point PDF $f_w(W; t)$ and a two-point auto-correlation function $\rho_w(|t_1 - t_2|)$; these functions involve meta-parameters $\boldsymbol{\varphi}_w$ such as the mean, variance, and correlation length of $w(t)$. The deterministic function $s(x; \cdot)$, parameterized by a set of N_θ (possibly uncertain, i.e., random) coefficients $\boldsymbol{\theta} \in \mathbb{R}^{N_\theta}$, is such that a solution to (10) is smooth almost surely in the probability space of both $w(t)$ and, possibly, $\boldsymbol{\theta}$ and x_0. If $\boldsymbol{\theta}$ and x_0 are random, then they are characterized by PDFs $f_{\boldsymbol{\theta}}(\boldsymbol{\Theta})$ and $f_0(X)$, with meta-parameters $\boldsymbol{\varphi}_\theta$ and $\boldsymbol{\varphi}_0$, respectively. In all, the statistics of $x(t)$ depends on the set of N_{par} meta-parameters $\boldsymbol{\varphi} = (\boldsymbol{\varphi}_w, \boldsymbol{\varphi}_\theta, \boldsymbol{\varphi}_0) \in \Phi \subset \mathbb{R}^{N_{\mathrm{par}}}$ that define statistical properties of the uncertain physical parameters and inputs.

In addition to being described by the model (10), the system state $x(t)$ is sampled at N_{meas} times $t_1, \ldots, t_{N_{\mathrm{meas}}}$. For ease of notation, noisy observations $\hat{\mathbf{x}} = \{\hat{x}_1, \ldots, \hat{x}_{N_{\mathrm{meas}}}\}$ are chronologically ordered although this is not required by DAMD [11]. These measurements satisfy the data model

$$\hat{x}_m = x(t_m) + \varepsilon_m, \qquad m = 1, \ldots, N_{\mathrm{meas}}. \tag{11}$$

The mutually uncorrelated Gaussian measurement errors ε_m have zero mean and variance σ_ε^2.

Data assimilation (DA) improves model predictions by augmenting them with observations. Some DA methods yield the "best" (i.e., unbiased) prediction and quantify its predictive uncertainty in terms of, respectively, ensemble mean, $\langle x(t)\rangle$, and standard deviation, $\sigma_x(t)$, of the state variable $x(t)$. These statistics provide but limited information about $x(t)$, unless it is Gaussian or a known map thereof. Bayesian update and particle filters are examples of DA strategies that overcome this limitation by seeking a solution of (10) in terms of the PDF $f(X;t)$ of $x(t)$—or the corresponding CDF $F(X;t) = \mathbb{P}[x(t) \leq X]$—updated with the data $\hat{\mathbf{x}}$ in (11). Computing such distributions with ensemble methods requires a large number of repeated solves of (10), which can be prohibitively expensive when each forward solve carries significant cost.

Data assimilation via DAMD [11] removes the need for linearity and Gaussianity approximations, which underpin Kalman filtering, while significantly accelerating the computation. Like many DA strategies, DAMD comprises forecast and analysis. The first of these steps relies on the model (10) and predicts the system state at time t in terms of $f(X;t)$ or $F(X;t)$. Rather than using, e.g., Monte Carlo simulations, MD [3] implements this step by deriving a deterministic equation for $f(X;t)$ or $F(X;t)$. Thus, the single-point CDF $F(X;t)$ of the state variable $x(t)$ in (10) satisfies (sometime approximately) a parabolic PDE (Appendix A)[2]

$$\frac{\partial F}{\partial t} + \mathcal{U}(X,t;\boldsymbol{\varphi})\frac{\partial F}{\partial X} = \frac{\partial}{\partial X}\left(\mathcal{D}(X,t;\boldsymbol{\varphi})\frac{\partial F}{\partial X}\right), \qquad t>0, \quad X\in\Omega=[X_{\min},X_{\max}], \tag{12a}$$

subject to initial and boundary conditions

$$F(X;0) = F_0(X), \qquad F(X_{\min},t) = 0, \qquad F(X_{\max},t) = 1. \tag{12b}$$

The drift velocity, $\mathcal{U}(X,t;\boldsymbol{\varphi}) : \Omega\times\mathbb{R}^+ \to \mathbb{R}$, and the diffusion coefficient, $\mathcal{D}(X;t,\boldsymbol{\varphi}) : \Omega\times\mathbb{R}^+ \to \mathbb{R}^+$, are smooth functions of their arguments, which involve a set of the meta-parameters $\boldsymbol{\varphi}$. The functional forms $\mathcal{U}$ and $\mathcal{D}$ depend on that of $s(x;\cdot)$, on the statistical characterization of the random parameters epitomized by the statistical parameters $\boldsymbol{\varphi}$ of their distributions, and on the degree of approximation introduced by the closure strategy. If

[2]For spatially-dependent physical models, space would appear as a coordinate in a CDF or PDF equation [11]. For systems, MD would yield a PDF equation for the joint PDF of the interacting system states [35, 36].

the initial state of the system, x_0, is known with certainty, then its CDF $F_0(X)$ is the Heaviside step function, $F_0(X) = \mathcal{H}(X - x_0)$. For Itō SDEs with an additive white Gaussian forcing (a.k.a. standard Langevin equation), (12) is derived exactly; it corresponds to the integral of the Fokker-Planck equation [37, Ch. 4]. This setting is explored in Appendix A.1. Solving (12) is usually cheaper than obtaining, with comparable accuracy, an empirical CDF from Monte Carlo realizations of (10) [38].

Remark 3.1. *The CDF equation (12) maps the meta-parameters $\boldsymbol{\varphi}$ onto $F(X; t, \boldsymbol{\varphi})$, the CDF of the system state $x(t)$. In other words, a point $\boldsymbol{\varphi} \in \Phi \subset \mathbb{R}^{N_{par}}$ can be thought of as a coordinate on the statistical manifold $\mathcal{M}$ of the CDF $F(X; t, \boldsymbol{\varphi})$ at time t. At any time t', a solution to (12) provides an estimate of the CDF $F(X; t', \boldsymbol{\varphi})$ dependent on the current characterization of the random inputs expressed by $\boldsymbol{\varphi}$. Equivalently, points $\widetilde{\boldsymbol{\varphi}} = \{t, \boldsymbol{\varphi}\}$ define a dynamic statistical manifold $\mathcal{M}_t$ of the CDF $F(X; \widetilde{\boldsymbol{\varphi}})$.*

The second step of DAMD, analysis via Bayesian update, is performed sequentially for each of the N_{meas} measurements $\hat{x}_m$ in (11). At mth assimilation step, the inference problem is formulated as the minimization (5) of the discrepancy D between the CDF $F(X; t_m, \boldsymbol{\varphi})$ predicted by the model (12) and the observational CDF obtained with Bayes' rule,

$$\hat{F}(X; t_m) = \int_{X_{\min}}^{X} \hat{f}(X; t_m) \mathrm{d}X \tag{13a}$$

with

$$\hat{f}(X; t_m) = \frac{f_L(\hat{x}_m | x(t_m) = X) f(X; t_m, \boldsymbol{\varphi}^{(m-1)})}{\int_\Omega f_L(\hat{x}_m | x(t_m) = X) f(X; t_m, \boldsymbol{\varphi}^{(m-1)}) \mathrm{d}X}. \tag{13b}$$

Here the likelihood function $f_L(\hat{x}_m | x(t_m) = X)$ specifies the choice of a data model; and the PDF $f(X; t_m, \boldsymbol{\varphi}^{(m-1)})$, computed by solving the CDF equation (12) with the parameter set $\boldsymbol{\varphi}^{(m-1)}$ from the previous assimilation step, serves as a prior. Minimization of (5) yields the updated meta-parameters $\boldsymbol{\varphi}^{(m)}$, which are subsequently used in the forecast step in combination with the CDF equation. The calculation of the prior at the first assimilation step ($m = 1$) relies on the initialization of the meta-parameters $\boldsymbol{\varphi}^{(0)}$; the subsequent updates rely on the assumption that observation errors are mutually

uncorrelated,

$$\hat{f}(X;t_m) \propto \prod_{i=1}^{m-1} f_L(\hat{x}_i|x(t_i)=X) f_L(x_m|x(t_m)=X) f(X;\boldsymbol{\varphi}^{(m-1)};t_m)$$
$$\propto f_L(\hat{x}_m|X)\hat{f}(X;t_m,\boldsymbol{\varphi}^{(m-1)}).$$

The formulation of data assimilation in the form of the optimization problem (5) underpins Variational Inference (VI) and provides access to the theoretical results summarized in section 2. While generally faster than MCMC—their performance depends on the properties of the parameterized distributions—and compatible with off-the-self optimization algorithms, e.g., [15], the VI methods do not enjoy asymptotic guarantees of convergence available for MCMC [39].

The use of the L_2 norm in formulation of the discrepancy D incurs significant computational cost of solving the minimization problem (5) [11]. A key innovation of this study is to exploit the geometric structure of the statistical manifolds in the parameter space Φ by using either the KL divergence (2) or the Wasserstein distance (4). Operationally, this means employing either (2) or (4), rather than D in (5), at every assimilation step. This enables us to solve (5) via NGD, which we henceforth refer to as NGD-KL and NGD-W_2 depending on which metric is used. The update of the meta-parameters $\boldsymbol{\varphi}$ is done using NGD-KL (7) or NGD-W_2 (9), taking advantage of the explicit formulations for the manifold's metric tensors $\mathbf{G}_F$ in (6) and $\mathbf{G}_W$ in (8), as detailed in section 2.

Remark 3.2. *The analysis step of DAMD is performed on univariate (one-point) distributions ($d = 1$) regardless of the size of the physical parameter and meta-parameter sets, N_θ and N_{par}. That drastically reduces (to one) the dimensionality of the update effort in classical Bayesian DA. Estimation of the state distributions at different locations is made possible by the physics-based nature of the parameterization: forecast distributions $F(X;t,\boldsymbol{\varphi})$ obey PDEs that depend on the meta-parameters $\boldsymbol{\varphi}$. Updated meta-parameters are used in the CDF equation to forecast at different locations. Availability of a CDF/PDF equation removes the need for Gaussianity and linearity assumptions on the physical model and its random parameters. The CDF/PDF equation is assumed to be valid throughout the assimilation process.*

Remark 3.3. *Parameter update via discrepancy minimization places DAMD in the company of many machine-learning and optimal-transport techniques*

(see the references above). Unlike these methods, DAMD uses CDF or PDF equations and their parameters to define the parameter space for a statistical manifold $\boldsymbol{\varphi}$, such that the discrepancy minimization is constrained by these PDEs. Learning occurs on the statistical manifold defined by $\boldsymbol{\varphi}$ and proceeds by sequential updates of these meta-parameters.

Remark 3.4. *If observational or simulated data abound, a similar approach can be used to discover the coefficients of the CDF equation and/or its differential form. This research direction is explored in [13, 40].*

Loss function minimization. The PDE-constrained minimization problem (5) can be solved with different techniques [41]. We use a surrogate model to accelerate the calculation of the discrepancies $d_{\mathrm{KL}}(f(X;\boldsymbol{\varphi}),\hat{f})$ or $W_2(f(X;\boldsymbol{\varphi}),\hat{f})$, their respective gradients $\nabla_{\boldsymbol{\varphi}} d_{\mathrm{KL}}$ or $\nabla_{\boldsymbol{\varphi}} W_2$, and the preconditioning tensor metrics $\mathbf{G}_F$ or $\mathbf{G}_W$. The workflow of DAMD with a surrogate is detailed in Algorithm 1. We employ a fully-connected deep neural network (DNN) to approximate the solution of the CDF equation given the set of inputs $\mathbf{X} = \{X, t, \boldsymbol{\varphi}\}$. The number of outputs in this DNN equals the number of inputs, $\mathbf{Y} = \{Y_j : j = 1, \ldots, N_{\mathrm{par}} + 2\} = \{Y = F(X;t,\boldsymbol{\varphi}), t, \boldsymbol{\varphi}\}$, such that $\dim(\mathbf{X}) = \dim(\mathbf{Y}) = N_{\mathrm{par}} + 2$. Furthermore, we require the resulting vector function $\mathbf{Y} = \mathbf{F}(\mathbf{X})$ to be continuously differentiable and invertible, i.e., require the determinant of its Jacobian to be non-zero [42].[3] The chosen vector function satisfies these requirements by the very nature of distributions, which are assumed here to have a compact support, hence it fulfills the hypotheses of the Inverse Function theorem [46] for vector functions.

Under these conditions, the inverse of the vector function $\mathbf{Y} = \mathbf{F}(\mathbf{X})$ is differentiable, and the derivative of the inverse is equal to the inverse of the derivative [46]. Automatic differentiation is employed both to verify the inversion theorem hypotheses and to calculate the terms appearing in the minimization algorithms. This is especially useful, since NGD-KL utilizes the derivatives of the forward pass, whereas NGD-W_2 requires the derivatives of the inverse function. A differentiable DNN allows accurate calculation of the metric tensors for both geometries, eliminating potential problems related to their empirical approximation.

[3] The choice of a DNN to approximate the input-output relation $\mathbf{Y} = \mathbf{F}(\mathbf{X})$ is done solely for the sake of concreteness. One can replace it with another approximation technique, such as polynomial representation and symbolic regression [43, 44, 45].

Algorithm 1 DAMD with PINN surrogate workflow.

Identify (or develop) CDF equation (12) and meta-parameters $\boldsymbol{\varphi}$
Train PINN (14) to obtain surrogate model for $F(X;t,\boldsymbol{\varphi})$
Select discrepancy metric in (5), e.g., as a function of (1), (2), (4)
if $\mathcal{C} = d_{\mathrm{KL}}$ (5) **then**
    Select NGD (7) optimization strategy to minimize (5)
else if $\mathcal{C} = 1/2W_2^2$ (5) **then**
    Select NGD (9) optimization strategy to minimize (5)
else
    Select off-the-shelf optimization strategy to minimize (5)
end if
Initialize $\boldsymbol{\varphi}$ to $\boldsymbol{\varphi}^{(0)}$
for Every observation m **do**
    Calculate observational PDF $\hat{f}(X;t_m)$ (13)
    Minimize (5) with the selected optimization strategy
    Update $\boldsymbol{\varphi}$ to $\boldsymbol{\varphi}^{(m)}$
end for

The DNN training is accomplished by solving an optimization problem [47],

$$\underset{\mathbf{w},\mathbf{b}}{\operatorname{argmin}}(\mathrm{MSE}_{\mathrm{ts}} + \mathrm{MSE}_{\mathrm{R}} + \mathrm{MSE}_{\mathrm{aux}} + \mathrm{SC}), \tag{14a}$$

with respect to the weights and biases of the DNN, $\mathbf{w}$ and $\mathbf{b}$, respectively. Here,

$$\mathrm{MSE}_{\mathrm{ts}} = \sum_{j=1}^{n+2} \lambda_j \frac{1}{N_{\mathrm{ts}}} \sum_{i=1}^{N_{\mathrm{ts}}} |Y_j(\mathbf{X}_{\mathrm{ts}}^i) - Y_{j,\mathrm{ts}}^i|^2, \quad \lambda_j = (\max Y_{j,\mathrm{ts}})^{-1} \tag{14b}$$

$$\mathrm{MSE}_{\mathrm{R}} = \frac{1}{N_{\mathrm{R}}} \sum_{i=1}^{N_{\mathrm{R}}} |R(\mathbf{X}_{\mathrm{R}}^i)|^2 \tag{14c}$$

$$\mathrm{MSE}_{\mathrm{aux}} = \frac{1}{N_{\mathrm{aux}}} \sum_{i=1}^{N_{\mathrm{aux}}} |Y(\mathbf{X}_{\mathrm{aux}}^i) - Y_{\mathrm{aux}}^i|^2 \tag{14d}$$

$$\mathrm{SC} = \left(\max \left| \frac{\partial Y}{\partial X}(\mathbf{X}_{\mathrm{SC}}) \right| \right)^{-1} \sum_{i=1}^{N_{\mathrm{SC}}} \max \left(0, -\frac{\partial Y}{\partial X}(\mathbf{X}_{\mathrm{SC}}^i) \right), \tag{14e}$$

and $\mathbf{Y}(\mathbf{X}^i)$ represents the $N_{\text{par}}+2$ outputs of the DNN with inputs $\mathbf{X}^i$. The DNN is trained on a data set consisting of N_{ts} pairs $(\mathbf{X}^i_{\text{ts}}, \mathbf{Y}^i_{\text{ts}})$, for $i = 1, \ldots, N_{\text{ts}}$ in terms of Mean Square Error, MSE_{ts}. The training set is generated by solving the CDF equation (12) for N_{ts} combinations of meta-parameters $\boldsymbol{\varphi}$, i.e., at points $\boldsymbol{\varphi}_i \in \Phi$ with $i = 1, \ldots, N_{\text{ts}}$.[4] Moreover, the mean square errors MSE_{R} and MSE_{aux} enforce the fulfillment of the CDF equation and its initial/boundary conditions at collocation points $\{\mathbf{X}^i_{\text{R}}\}_{i=1}^{N_{\text{R}}}$ and $\{\mathbf{X}^i_{\text{aux}}\}_{i=1}^{N_{\text{aux}}}$, respectively.[5] Our DNN is physics-informed, as the CDF equation is rigorously derived from the physical model, albeit with closure approximations. Consequently, we hereafter refer to it as PINN. The residual is defined as

$$R(\mathbf{X}^i_{\text{R}}) = \frac{\partial Y(\mathbf{X}^i_{\text{R}})}{\partial t} + \left(\mathcal{U}(\mathbf{X}^i_{\text{R}}) - \frac{\partial \mathcal{D}}{\partial X}(\mathbf{X}^i_{\text{R}})\right)\frac{\partial Y}{\partial X}(\mathbf{X}^i_{\text{R}}) - \mathcal{D}(\mathbf{X}^i_{\text{R}})\frac{\partial^2 Y}{\partial X^2}(\mathbf{X}^i_{\text{R}}), \tag{15}$$

and Y^i_{aux} represent the auxiliary conditions for the CDF equation at points $\mathbf{X}^i_{\text{aux}}$, which represents initial or boundary conditions (12b). The term SC is a Soft Constraint [48] that regularizes the DNN by enforcing monotonicity of the output $Y = F(X; t, \boldsymbol{\varphi})$ along the X direction at points $\{\mathbf{X}^i_{\text{SC}}\}_{i=1}^{N_{\text{SC}}} = \{\{\mathbf{X}^i_{\text{ts}}\}_{i=1}^{N_{\text{R}}}, \{\mathbf{X}^i_{\text{R}}\}_{i=1}^{N_{\text{R}}}, \{\mathbf{X}^i_{\text{aux}}\}_{i=1}^{N_{\text{aux}}}\}$. The physics-aware component of (14), $\text{MSE}_{\text{R}}+\text{MSE}_{\text{aux}}$, makes training less data-intensive and increases confidence in the PINN predictions outside the training range (but within the residual points range).

4. Numerical Experiments

In this section, we apply the information-theoretic DA strategy introduced above to two problems described by (10). Section 4.1 (Example 1) deals with a Langevin equation with white noise $w(t)$, a problem for which the CDF equation (12) is exact. In other words, the forecast component of DAMD is exact, whereas the analysis step introduces an approximation. This setting allows us to ascertain the impact of the PINN surrogate of the

[4] For each i, the data pairs $(\mathbf{X}^i_{\text{ts}}, \mathbf{Y}^i_{\text{ts}})$ are extracted from these solutions at regularly-spaced time intervals and at spatial locations (in the X direction) refined with a cosine mapping around a solution of (10) with mean parameters.

[5] We select a regularly spaced set of points for the enforcement of (14c) in all but the X direction, wherein points are refined around the solution of (10) with mean parameters; N_{aux} points are regularly spaced in all directions.

CDF solution on the accuracy of DAMD. In section 4.2 (Example 2), we consider a Langevin equation with colored noise $w(t)$ that is modeled as an Ornstein-Uhlenbeck process [1, Ch. 3.2]; the derivation of the CDF equation (12) requires a closure approximation. In this case, the performance of DAMD depends also on the accuracy and robustness of the CDF equation as forecasting tool.

In both cases, one realization $\boldsymbol{\theta}^\star$ of the relevant unknown parameters $\boldsymbol{\theta}$ represents ground truth. Statistical models for these parameters are chosen such that the state variable $x(t)$ has a compact support $\Omega \subset \mathbb{R}^+$, or can be approximated as such with probability guarantees. This ensures that the information geometry induced by the W_2 divergence is rigorously defined. The N_{meas} observations $\hat{\mathbf{x}}$ are taken at regular time intervals, with the time step $\Delta t = t_{N_{\text{meas}}}/(N_{\text{meas}} + 1)$. They are generated by adding zero-mean Gaussian noise with standard deviation σ_ε to the solution of (10) with $\boldsymbol{\theta}^\star$ (i.e., the synthetic truth). This procedure results in the Gaussian likelihood function f_L, although other choices are possible [6].

We use the `JITCSDE` Python module [49] to solve the stochastic differential equation (10) with $\boldsymbol{\theta} = \boldsymbol{\theta}^\star$. The module implements the adaptive integration method [50] for both Itō (considered here) and Stratonovich SDEs. The CDF equations (12) are solved with a finite volumes (FV) scheme, implemented using the `Fipy` library [51], to provide a training set for the surrogate model. PINN is trained by employing `Tensorflow`; optimization in (14) is performed using L-BFGS-B method [52], with a random initialization of $\mathbf{w}$ and $\mathbf{b}$. Automatic differentiation is used to compute both the derivatives in the residual R in (15) and the PDF from CDF. The CDF equation for the problem in section 4.1 has an analytical solution (Appendix A.1), which is used to evaluate the impact of the PINN approximation on the assimilation procedure. Minimization of the KL and W_2 discrepancies is performed via both standard GD and NGD. In the case of NGD, convergence is accelerated by the use of the pre-conditioners $\mathbf{G}_F$ and $\mathbf{G}_W$ in (7) and (9). For each direction established by the gradient of the loss function (adjusted by the pre-conditioners when NGD is used) we employ the `Scipy` library's implementation of step calculation [53, Sec. 5.2]. A conver-

[6] While not investigated here, data models constructed from repeated observations of the same phenomenon may be more suitable for processes that are inherently random, like those described by Langevin equations. Crucially, DAMD can seamlessly incorporate nonlinear one-to-one observation maps, emerging when $x(t)$ cannot be observed directly.

gence criterion for NGD in (7) and (9) is defined by $|\nabla_{\varphi} D| \leq \epsilon$. Because of the different order of magnitude of the KL and W_2 discrepancies D, the convergence threshold ϵ is discrepancy-specific; we select a KL-based minimization threshold, ϵ_{KL}, and assign the threshold for W_2, ϵ_{W_2}, such that $\epsilon_{\mathrm{W}_2}/\mathcal{C}(\mathrm{W}_2(f(X;t_1,\boldsymbol{\varphi}^{(0)});\hat{f}(X;t_1))) = \epsilon_{\mathrm{KL}}/\mathcal{C}(d_{\mathrm{KL}}(f(X;t_1,\boldsymbol{\varphi}^{(0)});\hat{f}(X;t_1)))$.

4.1. Example 1: Langevin equation with white noise

The dynamics of state variable $x(t)$ is described by a Langevin equation,

$$\frac{\mathrm{d}x}{\mathrm{d}t} = -a(t)x(t), \qquad x(0) = x_0^\star, \tag{16}$$

where the statistically homogeneous (stationary) random process $a(t) = \mu_a + \sigma_a w(t)$ has mean $\mu_a \in \mathbb{R}^+$ and standard deviation $\sigma_a \in \mathbb{R}^+$, and $w(t)$ is standard Gaussian white noise. The initial state $x_0^\star \in \mathbb{R}^+$ is deterministic; to be specific, we set $x_0^\star = 1$. We impose $2\sigma_a < \mu_a$, such that $\mathbb{P}[a(t) > 0] > 0.97$ at each time step, and the support of $x(t)$ is approximately compact, $\Omega = [0, x_0^\star] \subset \mathbb{R}^+$.

The single-point CDF $F(X;t)$ of $x(t)$ in (16) satisfies exactly a CDF equation (Appendix A)

$$\frac{\partial F}{\partial t} - \mu_a X \frac{\partial F}{\partial X} = \frac{1}{2}\frac{\partial}{\partial X}\left(\sigma_a^2 X^2 \frac{\partial F}{\partial X}\right), \tag{17a}$$

subject to initial and boundary conditions

$$F(x;0) = \mathcal{H}(X - x_0^\star), \qquad F(X_{\min};t) = 0, \qquad F(X_{\max};t) = 1. \tag{17b}$$

In this example, CDF F is parameterized by $\boldsymbol{\varphi} = \{\mu_a, \sigma_a\}$, which we make explicit by writing $F(X;t,\boldsymbol{\varphi})$. The values of $\boldsymbol{\varphi}$ are refined by assimilating observations $\hat{\mathbf{x}}$.

A PINN serves as a surrogate model that approximates the solution of the CDF equation (17). The training set consists of the finite-volumes solutions [51] of (17) at selected points (X,t), computed for a number of different combinations of meta-parameters $\boldsymbol{\varphi}$. The details of this and other computations are provided in the opening of section 4. In this experiment, PINN function approximation is considered satisfactory upon reaching a value of the loss function of $4 \cdot 10^{-4}$. This high accuracy enables the deployment of the PINN surrogate for both the analysis and forecast steps, further accelerating the information-geometric optimization of (5) with the `Scipy` conjugate gradient routine. We derive an analytical solution for (17) in Appendix B, and then use it to quantify the impact of the PINN approximation on DAMD.

Remark 4.1. *A surrogate model is introduced to accelerate loss function minimization in cases where an analytical solution is not available. For complex problems, it might be advantageous to use the surrogate model only for the approximation of the gradients, while retaining the finite-volume solution of the CDF equation for prediction. Alternatively, it might be necessary to construct a surrogate model for the local CDF at each assimilation time t_m, hence reducing the dimensionality of the surrogate model.*

For this problem, the parameters $\boldsymbol{\varphi} = \{\mu_a, \sigma_a\}$ can be estimated via the linear Kalman filter (LKF) [54] (Appendix B). LKF forecast is performed analytically, whereas LKF analysis relies on the (approximate) linearization of the observation map [10]. We use LKF to evaluate the accuracy of DAMD in the identification of the meta-parameters $\boldsymbol{\varphi}$ via GD in (5) with $\mathcal{C} \equiv d_{\mathrm{KL}}$ or $W_2^2/2$, NGD-KL (7), and NGD-W$_2$ (9), performed using either the analytical CDF F or its DNN approximation. Unlike LKF, DAMD requires neither a linearizing approximation nor the Gaussianity assumption. The availability of analytical formulations for $F(X;t)$ and $f(X;t)$ (either the exact solution or its DNN approximation) facilitates the calculation of the forecast PDFs at each measurement time (13), as well as the (semi-)analytical computation of both the metric tensors $\mathbf{G}_F$ in (6) and $\mathbf{G}_W$ in (8), and the gradient of the discrepancy, $\nabla_\varphi \mathcal{D}$, for the KL and W_2 measures[7]. The integrals in the metric tensors, the discrepancy gradient, and the normalization constant in (13), are computed via numerical quadrature from the Fortran library `QUADPACK`.

Figure 1 shows the updated $\boldsymbol{\varphi}^{(m)}$ as function of the assimilation step m for both the KL and W_2 metrics of discrepancy, either taking (NGD) or not taking (GD) advantage of the information-geometric structure of the statistical manifold of F. Also shown in this figure are estimates of the meta-parameters $\boldsymbol{\varphi}^{(m)}$ obtained with both LKF and GD for the L_2 norm of discrepancy, $D(\boldsymbol{\varphi}) = d_2(f(X;t,\boldsymbol{\varphi}), \hat{f}(X;t))$ in (1). To facilitate comparison between the various discrepancy metrics, we assign the minimization convergence threshold for d_2, ϵ_{d_2}, such that

$$\frac{\epsilon_{d_2}}{\mathcal{C}\{d_2(f(X;t_1,\boldsymbol{\varphi}^{(0)}), \hat{f}(X;t_1))\}} = \frac{\epsilon_{\mathrm{KL}}}{\mathcal{C}(d_{\mathrm{KL}}(f(X;t_1,\boldsymbol{\varphi}^{(0)}); \hat{f}(X;t_1)))}.$$

All assimilation algorithms improve estimates of both the exact mean μ_a and standard deviation σ_a, but DAMD with the surrogate yields a more

[7] The code is available at https://github.com/DDMS-ERE-Stanford/DAMD-NGD

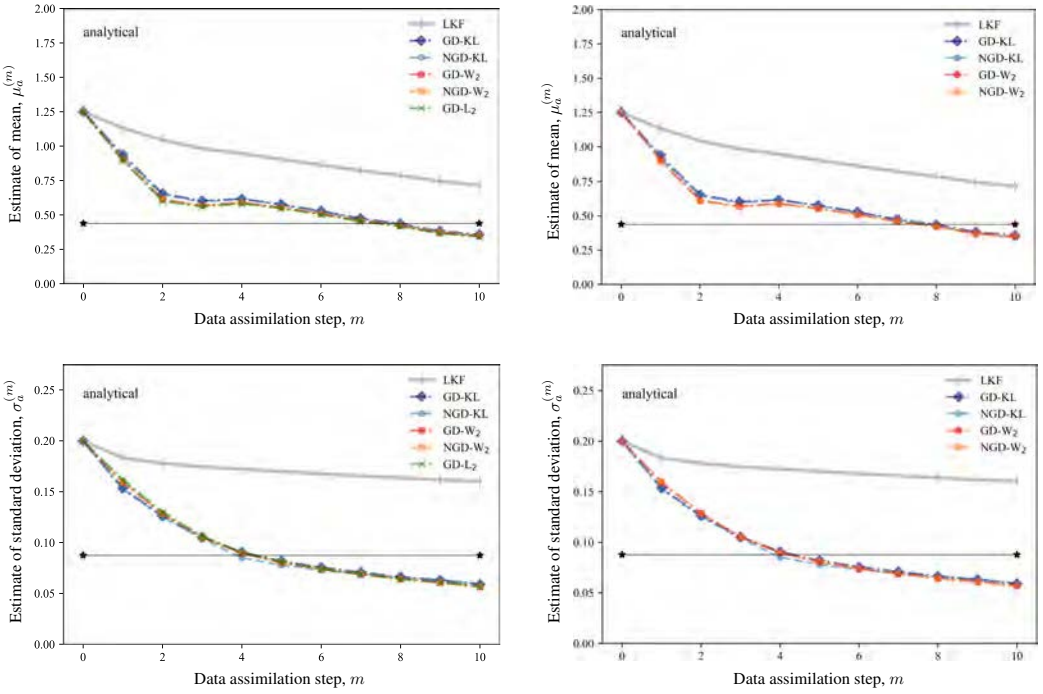

Figure 1: Example 1: Estimation of the meta-parameters $\boldsymbol{\varphi} = \{\mu_a, \sigma_a\}$ for the Langevin equation with white noise via DAMD using either the analytical CDF solution (left column) or its DNN surrogate (right column). The parameters $\boldsymbol{\varphi}^{(m)}$ are plotted as function of the assimilation step m for the four information-geometric optimization strategies: GD and NGD, for the KL and W_2 discrepancies. Also shown are the corresponding estimates obtained with LKF and GD for the L_2 measure of discrepancy. The starred values correspond to the statistical parameters used to generate the observations. The simulation parameter values are set to $x_0^\star = 1$, $\boldsymbol{\varphi}^\star = \{0.44, 0.088\}$, $\boldsymbol{\varphi}^{(0)} = \{1.25, 0.2\}$, $\theta_1^{(0)} = \mu_a^{(0)}/\sigma_a^{(0)}$, $\theta_2^{(0)} = \log \sigma_a^{(0)}$, $\sigma_\varepsilon = 0.1$, $N_{\text{meas}} = 10$, $t_{N_{\text{meas}}} = 2$, $\epsilon_{\text{KL}} = 10^{-2}$, $N_T = 18850$, $N_B = 5632$, $N_I = 1280$, and $N_R = 47872$.

erratic estimate of σ_a. After N_{meas} are assimilated, LKF's estimates of μ_a and σ_a have errors equal to 32% and 64% of their respective initial errors. The DAMD errors (averaged over the alternative optimization strategies) for μ_a and σ_a are respectively 11% and 26% of their initial values when the analytical CDF solution is used; and 22% and 44% when the DNN surrogate is deployed. Thus, DAMD yields appreciably more accurate results than LKF does, regardless of whether the exact CDF solution or its DNN approximation is used. As expected, the reliance on the DNN surrogate of the CDF solution affects the DAMD accuracy; however, its performance is advantageous when only an approximate CDF equation is available, the setting explored in the next section.

Remark 4.2. *The performance of LKF depends on the initialization of the covariance matrix. The results reported above are for LKF, whose augmented state's covariance is initialized as*

$$diag[0, (|\mu_a^{(0)}/\sigma_a^{(0)} - \mu_a^\star/\sigma_a^\star|/2)^2, (|\log\sigma_a^{(0)} - \log\sigma_a^\star|/2)^2].$$

Remark 4.3. *Given the small dimensionality of the augmented state (one state variable and two parameters), each LKF assimilation step is computationally inexpensive. Each DAMD iterative step involves the calculation of one-dimensional numerical integrals to evaluate the loss function and compute the* $\mathbf{G}$ *components; once the DNN surrogate is trained, there is no significant difference in the computational effort required using the analytical* F *or its DNN surrogate.*

The number of iterations over the assimilation time window is smaller for NGD than for GD for both choices of the loss function (fig. 2). The difference is apparent when the CDF solution is analytical, and less pronounced when its DNN approximation is used. In the absence of error in the approximation of the distributions, the updated estimate of the meta-parameters obtained with $D = d_2$ is as good as those for $D = d_{\text{KL}}$ and $D = W_2$ (fig. 1), although with a higher number of iterations per assimilation step (fig. 2).

The physics-driven parameterization of the statistical manifold yields an isotropic geometry of the loss function in the search area, which reduces the benefits of preconditioning. This is shown in fig. 3, where the KL and W_2 loss functions are plotted, at the first and last assimilation steps, as function of the meta-parameters $\boldsymbol{\varphi}$, also highlighting the true solution and the prior

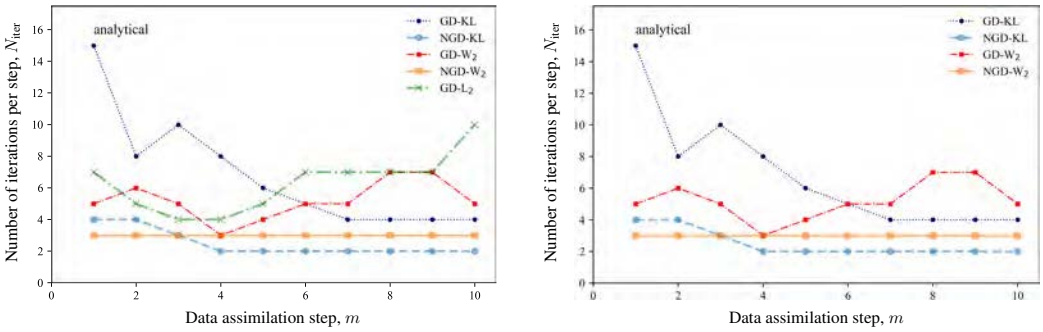

Figure 2: Example 1: Number of iterations per assimilation step m for the four information-geometric optimization strategies: GD and NGD, for the KL and W_2 discrepancies. DAMD employs either the analytical CDF solution (left) or its DNN surrogate (right). The GD with $D = d_2$ for the analytical distributions is also included. The simulation parameter values are the same as in fig. 1.

location.[8] The loss functions depicted in fig. 3 are obtained for the exact CDF solution; a similar behavior is exhibited by the corresponding functions computed using the DNN surrogate (not shown here). Although superficially similar throughout the assimilation process, the minor differences in the topology of the KL and W_2 loss functions are enough to prevent convergence for the KL loss function for a slightly worst choice of the prior (P_2 in the Figure), which results in diverging iterates of the DAMD procedure for both GD and NGD. This is because the KL divergence is more sensitive to numerical errors in the calculation of the integrals, especially for sharp or non-overlapping distributions, which mislead the direction of the search.

The computational cost of the different optimization strategies within the DAMD framework depends on the number of iterations (fig. 2); on the computational cost per iteration; and, in case of information-geometric optimization, on the cost of computing the tensor metrics. Since the function- and gradient-evaluations for this example are not expensive, the computa-

[8]The loss functions at assimilation step $m = 1$ are obtained using the initial $\boldsymbol{\varphi}^{(0)}$ for the calculation of the observational PDF/CDF, whereas the loss functions at assimilation step $N_{\text{meas}} = 10$ are computed using $\boldsymbol{\varphi}^{(N_{\text{meas}}-1)}$ for the prior obtained using either NGD-KL or NKD-W_2. The initial guess of the prior $\boldsymbol{\varphi}^{(0)}$ is the same for both KL and W_2 metrics (P_1 in the Figure), and yields similar outcomes in terms of identification of the meta-parameters, as illustrated above.

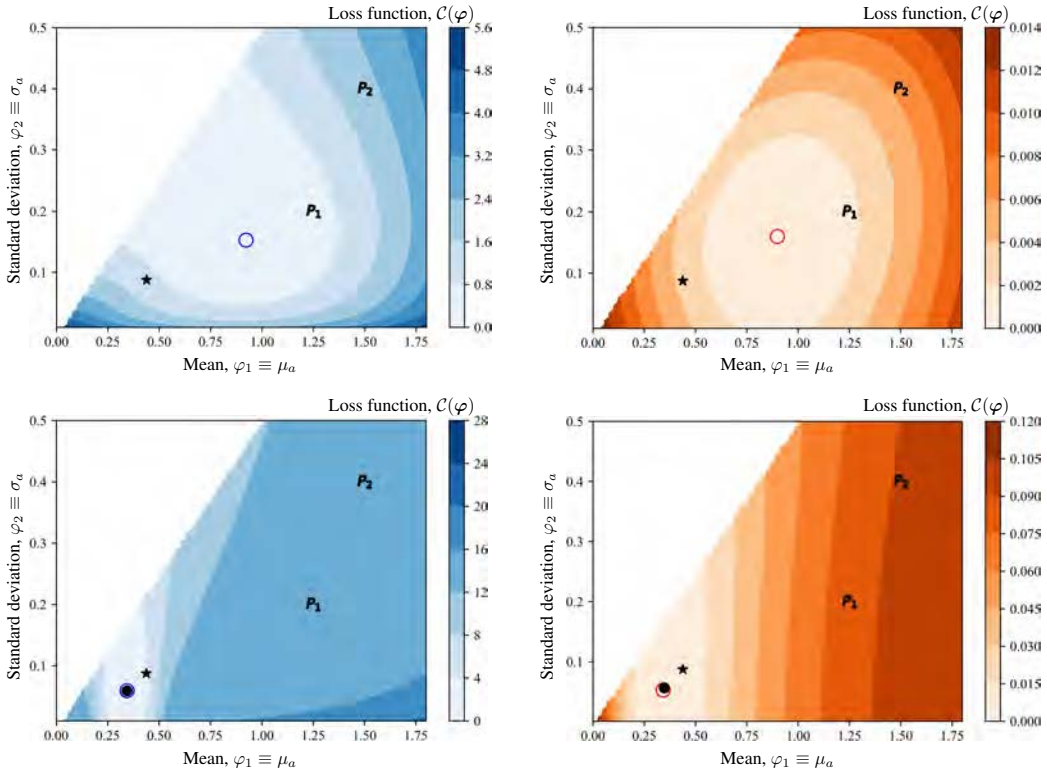

Figure 3: Example 1: The KL (left column) and W_2 (right column) loss functions at the first ($m = 1$, top row) and last ($m = N_{\text{meas}}$, bottom row) steps of DAMD. All calculations are performed using the analytical CDF solution. The star indicates the true values of the meta-parameters (used to generate the synthetic reality). The points P_1 and P_2 indicates the priors $\boldsymbol{\varphi}^{(0)} = (\varphi_1^{(0)}, \varphi_2^{(0)})$ for which the optimization of the KL loss function converges and fails to converge, respectively. The larger (empty) circles indicate the posterior parameters at the mth assimilation step, $\boldsymbol{\varphi}^{(m+1)}$, and the smaller (full) circles in the bottom row indicate $\boldsymbol{\varphi}^{(N_{\text{meas}}-1)}$. The blank region reflects the assumption $2\sigma_a < \mu_a$. The simulation parameter values are the same as in fig. 1

tional gain of having a smaller number of evaluations is not significant, and it is compensated by the additional cost of the calculation of the preconditioning matrices.

The posterior NGD parameters $\boldsymbol{\varphi}^{(N_{\text{meas}})}$ are used to compute the posterior CDF and PDF of $x(t)$ in fig. 4. NGD yields accurate posteriors, with the W_2 optimization (9) performing better than the KL optimization (7) when the approximate CDF solution is used. In order to highlight the accuracy of the DNN surrogate model, we show the finite-volume solution of the CDF equation (17) with $\boldsymbol{\varphi} = \boldsymbol{\varphi}^{(N_{\text{meas}})}$ and its corresponding PDF computed via numerical differentiation, and their DNN-based counterparts. We conclude that the W_2 optimization is more robust to inaccuracies of the CDF solution, thus yielding better overall predictions. Moreover, in agreement within [17], we found the W_2 minimization to be more robust to the choice of the prior.

Remark 4.4. *An additional advantage of the W_2 loss function stems from its reliance on a CDF rather than a PDF that enters the KL loss function. CDFs are smoother and easier to compute as a solution of the CDF equation than PDFs, which are obtained by solving the PDF equation. This facilitates the generation of a training set and the training of a surrogate model. On the other hand, approximation of the solution to a CDF equation with a DNN surrogate possesses a potential challenge for the W_2 optimization, since (5) calls for invertible surrogate models. We overcome this difficulty by selecting a special structure for the DNN that guarantees automatic inversion, as detailed in section 3.*

4.2. Example 2: Langevin equation with colored noise

The dynamics of state variable $x(t)$ is described by (10) with $s \equiv -a(t)x(t)$. Here, $a(t) = \mu_a + w(t)$ with $\mu_a \in \mathbb{R}^+$, and $w(t)$ is the derivative of an Ornstein–Uhlenbeck process characterized by the exponential auto-covariance function $C_w(t,\tau) = \sigma_a^2/(2\theta_a)[\exp(-\theta_a|t-\tau|) + \exp(-\theta_a(t+\tau))]$ with parameters σ_a and $\theta_a \in \mathbb{R}^+$. By construction, the latter is also the auto-covariance function of $a(t)$, $C_w(t,\tau) = C_a(t,\tau)$. Taking the initial state x_0 to be deterministic, the stochastic solution of this problem depends on three meta-parameters $\boldsymbol{\varphi} = \{\mu_a, \sigma_a, \theta_a\}$. We impose $\mu_a - 2\sqrt{\sigma_a^2/\theta_a} > 0$, such that the support of $x(t)$ is approximately compact, $\Omega \subset \mathbb{R}^+$. One realization of this solution, drawn from the distribution with the "true" meta-parameters $\boldsymbol{\varphi}^\star$, serves as ground truth for which observations $\hat{\mathbf{x}}$ are constructed in accordance with (11).

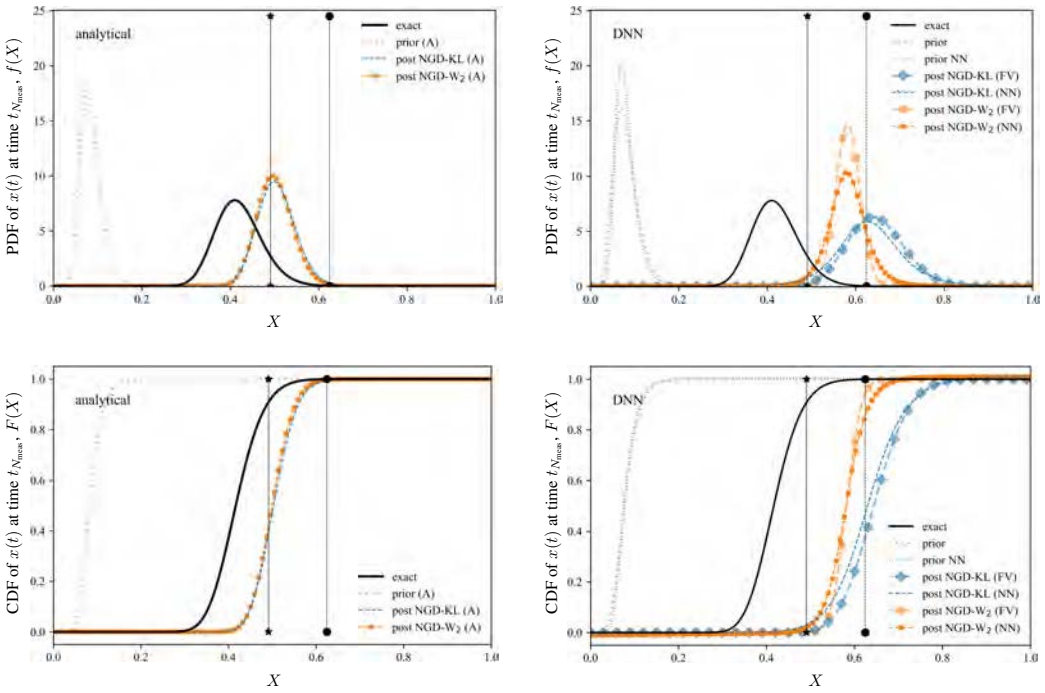

Figure 4: Example 1: Prior and posterior PDFs (top row) and corresponding CDFs (bottom row) at time $t_{N_{\text{meas}}}$ obtained via NGD minimization for the KL and W_2 loss functions, with either the analytical solution of the CDF equation (17) (left column) or its NN surrogate (right column). The FV CDF solution used to train the DNN is also shown in the right column. Black stars and circles mark the exact value $x(t_{N_{\text{meas}}})$ and its noisy observation $x_{N_{\text{meas}}}$, respectively. The black solid line represents the analytical CDF solution at $t_{N_{\text{meas}}}$ with $\boldsymbol{\varphi} = \{\mu_a^\star, \sigma_a^\star\}$, describing the synthetic truth. The simulation parameter values are the same as in fig. 1.

We show in Appendix A.2 that the CDF $F(X;t)$ of $x(t)$ satisfies the CDF equation (12) with

$$\mathcal{U}(X,t;\boldsymbol{\varphi}) = -\mu_a X + X\int_0^t C_w(t,\tau)\mathrm{d}\tau \quad \text{and} \quad \mathcal{D}(X,t;\boldsymbol{\varphi}) = X^2\int_0^t C_w(t,\tau)\mathrm{d}\tau. \tag{18}$$

In the absence of an analytical solution of the CDF equation in this case, we rely on a surrogate model to accelerate optimization. The FV solution of (12) with (18) and its DNN surrogate are used to assimilate observations $\hat{\mathbf{x}}$ via our information-geometric DAMD framework. Similar to the case of white noise (section 4.1), we found the KL-based implementation of DAMD to be less robust to the choice of the prior. Hence, only the W_2-based results are displayed below.

Figure 5 exhibits the identification of the meta-parameters $\boldsymbol{\varphi}$ as function of the data assimilation step m. Since the W_2 loss function is relatively insensitive to the third meta-parameter θ_a, we present the convergence results for $\tilde{\sigma} = \sqrt{\sigma^2/(2\theta_a)}$ instead.[9] Both GD-W_2 and NGD-W_2 converge after assimilation of about 20 observations, which are generated every $\Delta t = 0.055$. NGD converges, for the given combination of observations and the prior, in fewer iterations over the assimilation window than GD (fig. 5d) .

In fig. 6, we present the posterior PDF and CDF of the state $x(t)$ at the final assimilation time $t_{N_{\text{meas}}}$. The CDF is computed as a FV solution of the CDF equation with meta-parameters $\boldsymbol{\varphi}^{(N_{\text{meas}})}$, and the PDF as its derivative. Observations $\hat{\mathbf{x}}$ are assimilated, alternatively, via the GD-W_2 and NGD-W_2 optimization strategies. Both approaches yield posterior distributions that are close to the true state, with negligible differences between NGD-W_2 and GD-W_2. The use of the FV solution of the CDF equation leads to a slightly wider posterior than the reliance on its DNN surrogate does, possibly because of numerical diffusion.

Although not shown here, we found the KL- and W_2-based loss functions at the first and later assimilation steps to be smooth and not significantly different from each other. Yet, similar to the example in section 4.1, the differences are sufficient to prevent convergence in the KL case for poor choices of the prior.

[9] This lack of sensitivity reflects the challenge of inferring the correlation length, $1/\theta_a$, from observations over a time window spanning only two true correlation lengths, $1/\theta_a^\star$.

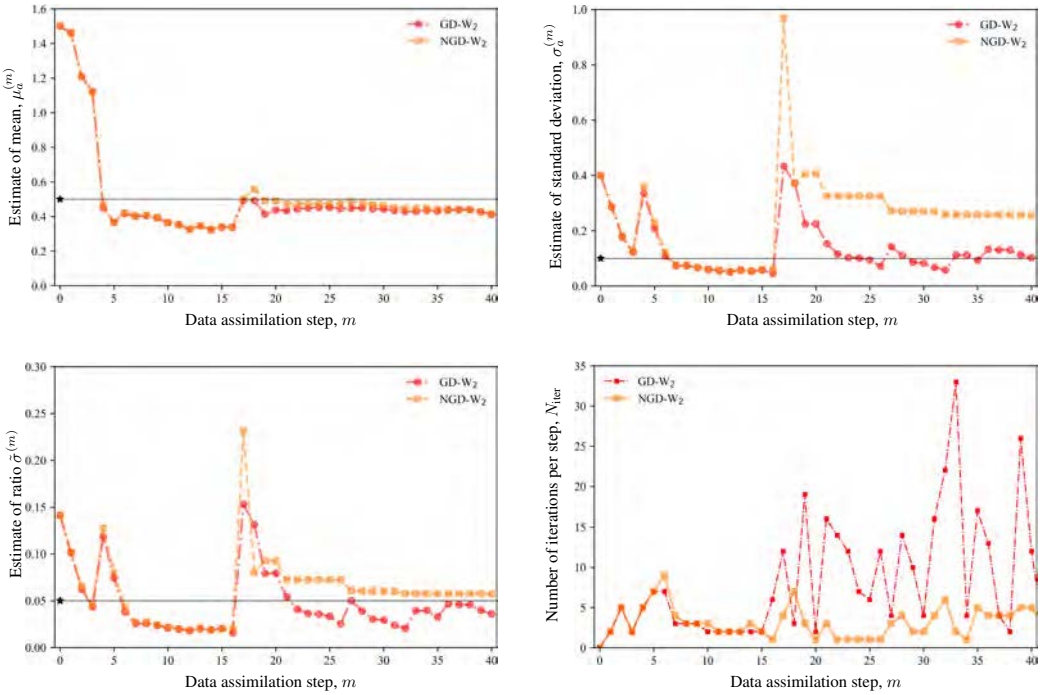

Figure 5: Example 2: Estimation of meta-parameters $\boldsymbol{\varphi} = \{\mu_a, \sigma_a, \tilde{\sigma} = \sqrt{\sigma_a^2/(2\theta_a)}\}$, as function of the assimilation step m, with GD and NGD for the W_2 loss functions. The bottom right panel shows the number of iterations per assimilation step for GD and NGD. The simulation parameter values are set to $x_0^\star = 1$, $\boldsymbol{\varphi}^* = \{0.5, 0.1, 0.05\}$, $\boldsymbol{\varphi}^{(0)} = \{1.5, 0.4, 0.14\}$, $\sigma_\varepsilon = 0.05$, $N_{\text{meas}} = 41$, $t_{N_{\text{meas}}} = 2.2$, $\epsilon_{\text{KL}} = 10^{-2}$, $N_T = 13550$, $N_B = 29282$, $N_I = 6655$, $N_R = 248897$.

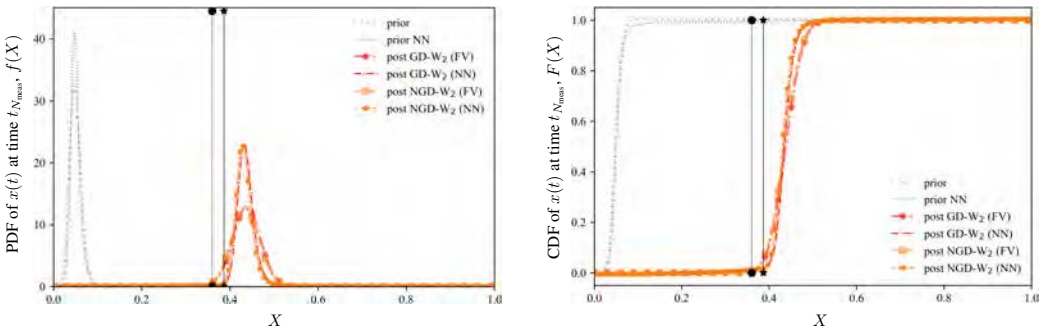

Figure 6: Example 2: Prior and posterior distributions at time $t_{N_{\text{meas}}}$ (both PDFs, on the left, and CDFs, on the right) obtained via GD-W$_2$ and NGD-W$_2$ minimization for the Langevin equation with colored noise. For each distribution, both the FV solution and NN approximation are shown. Black stars and circles mark the exact value $x(t_{N_{\text{meas}}})$ and its noisy observation $x_{N_{\text{meas}}}$, respectively. The simulation parameter values are the same as in fig. 5.

5. Discussion and Conclusion

We presented an information-geometric implementation of DAMD, which yields computationally efficient data assimilation and parameter estimation for nonlinear problems with non-Gaussian system states. The forecast step is performed by employing MD, an uncertainty propagation technique that yields a deterministic evolution equation for the CDF (or, equivalently, the PDF) of the state. This equation maps a set of meta-parameters (statistical properties of the random inputs) onto the system-state's distribution, and defines a parameter space for a dynamic manifold of distributions. Such probabilistic forecasts are physics-based but generally not exact, as they often require closure approximations; their accuracy can and should be ascertained a priori or drawn from the MD literature [55, 3, 12, 56, 13].

The analysis step is performed on this statistical manifold; it is formulated as sequential minimization of the discrepancy between an observational distribution and a predictive posterior distribution obeying the CDF equation with unknown (posterior) parameters. The observational PDF is the Bayesian posterior obtained as the product of the data model (i.e., the likelihood function), and the prior PDF obeys the CDF equation with the parameters from the previous assimilation step. DAMD can be classified as a Variational Inference method. Unlike classical VI methods, DAMD minimizes the discrepancy between univariate distributions, as observations are collected locally and the CDF equation acts as a physics-informed generative

model that allows one to compute univariate distributions of the state at an observation location. Like many VI methods, DAMD is typically faster than MCMC but does not enjoy asymptotic guarantees of convergence available for the latter.

Reliance on statistical discrepancy measures—the Kullback-Leibler divergence and the L_2 Wasserstein distance—confers exploitable geometric properties to the manifold of distributions. Specifically, it enables the use of NGD, an efficient optimization technique. Our numerical experiments revealed the W_2-based DAMD to be more robust to the choice of a prior than its KL-based counterpart.

For one-dimensional (univariate) distributions, W_2 is defined in terms of system-state CDFs, and KL in terms of corresponding PDFs. This argues in favor of the W_2-based DAMD, since CDFs are smoother and numerical solution of CDF equations is easier. This facilitates the use of invertible DNNs as a surrogate model in the probabilistic space to facilitate and accelerate optimization and calculation of the geometric metric tensors. In our numerical experiments, the W_2 optimization with the DNN surrogate yields more accurate results than the KL optimization.

Future work will focus on the identification of ambiguity sets and their dynamics on statistical manifolds [24], their evolution and their update with observations. We also plan to explore the use of different data models, the impact of alternative parameterizations of a statistical manifold on DAMD performance, and the latter's implications for sensitivity analysis.

Appendix A. CDF equation for the stochastic ODE

We summarize MD for the two test problems from section 4. The original derivations can be found in [37] and [7], respectively. The first result is exact, whereas the second is approximate and has been verified against Monte Carlo simulations in [6, 7].

Appendix A.1. MD for the Langevin equation with white noise

Consider a Langevin equation, (10) with $s(x;w) \equiv s_d(x,t) + s_w(x,t)w(t)$ where $w(t)$ is a white standard Gaussian process (with zero mean and unit variance). The deterministic functions s_d and s_w are such that $s(x;w)$ is integrable with respect to t in the mean square sense [37, Sec. 4.1]. The derivation of a PDF equation for $x(t)$ is relatively straightforward, and leads

to the Fokker-Planck equation (a.k.a. Kolmogorov's forward equation) [37, Sec. 4.9]

$$\frac{\partial f}{\partial t} + \frac{\partial\, s_d(X,t) f}{\partial X} = \frac{1}{2}\frac{\partial^2 s_w^2(X,t) f}{\partial X^2}, \tag{A.1}$$

It is formally valid if $f(X;t)$ is well-behaved at infinity, and is subject to initial and boundary conditions condition $f(X;0) = f_0(x)$ and $f(\pm\infty;t) = 0$. An equivalent CDF version of the Fokker-Planck equation (A.1) can be obtained via integration of (A.1) over $X \in \Omega$

$$\frac{\partial F}{\partial t} + s_d(X,t)\frac{\partial F}{\partial X} = \frac{1}{2}\frac{\partial}{\partial X}\left(s_w^2(X,t)\frac{F}{\partial X}\right), \tag{A.2}$$

subject to $F(x;0) = F_0(X)$, $F(X_{\min},t) = 0$, and $F(X_{\max},t) = 1$.

In (16), $s(x;w) = -a(t)x(t)$ where the random process $a(t)$ has the constant mean μ_a and standard deviation σ_a. This translates into $s_d(x,t) = -\mu_a x$ and $s_w = -\sigma_a x$, so that the coefficients $\mathcal{U}$ and $\mathcal{D}$ in (12) become $\mathcal{U} = -\mu_a X$ and $\mathcal{D} = (\sigma_a^2/2)X^2$, with $\boldsymbol{\varphi} = \{\mu_a, \sigma_a\}$.

Appendix A.2. MD for the Langevin equation with colored noise

Consider (10) with $s(x;w) \equiv -a(t)x(t)$, where $a(t) = \mu_a + w(t)$ and $w(t)$ is a correlated stationary Gaussian process (colored noise) [37, sec. 4.8]. MD for stochastic/random (Langevin) ODEs with temporally correlated forcings requires closure approximations. These include the semi-local approximation [6, 7], which compares favorably with Monte Carlo simulations and a local closure approximation in terms of both accuracy and computational efficiency. For the sake of completeness, we summarize the derivation of the PDF equation and its semi-local closure approximation for the specific form of the Langevin equation described above. We start by deriving an equation for the raw PDF $\pi(X,t) = \delta(X - x(t))$, whose ensemble mean is the PDF, $f(X;t) = \langle\pi\rangle$. Multiplying our ODE by $-\partial\pi/\partial X$ and using the properties of the Dirac delta function $\delta(\cdot)$, we obtain

$$\frac{\partial \pi}{\partial t} + a(t)\frac{\partial \pi}{\partial X} = 0. \qquad s(X,t) = \langle s(X,t)\rangle + s'(X,t;w); \quad \langle s\rangle = -\mu_a X, s' = -w(t)X \tag{A.3}$$

We use the Reynolds decomposition $\mathcal{A} = \langle\mathcal{A}\rangle + \mathcal{A}'$ to represent relevant random processes $\mathcal{A}$ as the sums of their ensemble means $\langle\mathcal{A}\rangle$ and zero-mean

fluctuations, $\mathcal{A}'$. Since $\pi = f + \pi'$, taking the ensemble mean of this equation yields an unclosed equation for the PDF $f(X;t)$,

$$\frac{\partial f}{\partial t} + \mu_a \frac{\partial f}{\partial X} + \frac{\partial \langle w'(t)\pi'(X,t)\rangle}{\partial X} = 0, \qquad \text{subject to} \quad f(X;0) = f_0. \quad \text{(A.4)}$$

A closure approximation is needed to render the cross-correlation term $\langle w'(t)\pi'(X,t)\rangle$ computable. Subtracting (A.4) from (A.3), we obtain an equation for random fluctuations $\pi'(X,t)$,

$$\frac{\partial \pi'}{\partial t} + \mu_a \frac{\partial \pi'}{\partial X} = \frac{\partial(\langle s'(X,t)\pi'(X,t)\rangle - s'\pi)}{\partial X}, \qquad \text{subject to} \quad \pi'(X, t=0) = 0. \quad \text{(A.5)}$$

The deterministic Green's function for (A.5), $G(X,t;\Xi,\tau)$, is a solution of

$$\frac{\partial G}{\partial \tau} + \mu_a \frac{\partial G}{\partial \Xi} = -\delta(X-\Xi)\delta(t-\tau) \quad \text{(A.6)}$$

with homogeneous initial (at $\tau = t$) and boundary conditions at infinity. Its analytical solution, obtained, e.g., via the method of characteristics, is $G(X,t;\Xi,\tau) = \mathcal{H}(t-\tau)\delta(X - \Xi\exp(-\mu_a(t-\tau)))$. Hence, the path-wise solution of (A.5) is

$$\pi'(X,t) = \int_0^t \int_{-\infty}^{\infty} G(X,t;\Xi,\tau) \frac{\partial}{\partial \Xi} \left[\langle w'(\Xi,\tau)\pi'(\Xi,\tau)\rangle - w'(\Xi,\tau)\pi(\Xi,\tau)\right] \mathrm{d}\tau\mathrm{d}\Xi. \quad \text{(A.7)}$$

A closure approximation for $\langle w'(t)\pi'(X,t)\rangle$ is constructed by multiplying (A.7) with $w'(t)$, taking the ensemble mean, and neglecting the third-order correlation term,

$$\langle w'(X,t)\pi'(X,t)\rangle = -\int_0^t \int_{-\infty}^{\infty} G(X,t;\Xi,\tau) \frac{\partial}{\partial \Xi} \left(C_w(X,t;\Xi,\tau) f(\Xi,\tau)\right) \mathrm{d}\Xi\mathrm{d}\tau, \quad \text{(A.8)}$$

where $C_w(X,t;\Xi,\tau) = \langle w'(X,t)w'(\Xi,\tau)\rangle$ is the auto-covariance of the random noise $w(t)$. Substituting this expression into (A.4) yields a nonlocal (integro-differential) PDF equation. Accounting for the analytical expression for G, (A.8) is approximated semi-locally as

$$\langle w'(X,t)\pi'(X,t)\rangle = -Xf(X,t)\int_0^t C_w(t,\tau)\mathrm{d}\tau - X^2 \frac{\partial f(X,t)}{\partial X} \int_0^t C_w(t,\tau)\mathrm{d}\tau. \quad \text{(A.9)}$$

This yields the closed CDF equation (12) with (18). If $w(t)$ were white noise, i.e., if $C_w(t,\tau) = \delta(t-\tau)$, then the resulting PDF equation would reduce to the Fokker-Planck equation.

Appendix B. Linear Kalman filter for Langevin equation with white noise

The nonlinear assimilation problem for the scenario explored in section 4 4.1 with $x_0 = 1$ is transformed into a linear problem with nonlinear observational map by introducing the parameters $\theta_1 = \mu_a/\sigma_a$ and $\theta_2 = \ln \sigma_a$ and the transformed variable $\xi(t) = -\mathrm{e}^{-\theta_2} \ln x(t)$ obeying

$$\frac{\mathrm{d}\xi}{\mathrm{d}t} = -\theta_1 + w(t), \quad \text{subject to} \quad \xi(t=0) = \xi_0 = 0. \tag{B.1}$$

The remapping of the physical parameters yields a linear parameter estimation problem, and enforces the positivity of σ_a. Observations of $x(t)$ are mapped onto $\xi(t)$ via a nonlinear observation map $\hat{x}_m = h(\xi(t_m)) + \varepsilon_m = \exp(-\mathrm{e}^{\theta_2}\xi(t_m)) + \varepsilon_m$, which replaces (11). An exact equation for the PDF $f_\xi(\Xi;t)$ of $\xi(t)$ is obtained following the steps outlined in Appendix A (a):

$$\frac{\partial f_\xi}{\partial t} + \frac{\partial \theta_1 f_\xi}{\partial \Xi} = \frac{1}{2}\frac{\partial^2 f_\xi}{\partial \Xi^2}, \quad \text{subject to} \quad f_\xi(\Xi; t=0) = \delta(\Xi - \xi_0). \tag{B.2}$$

The analytical solution of (B.2) is easily obtained, and mapped onto the PDF of $x(t)$, $f(X;t,\mu_a,\sigma_a)$. This is exploited in 4.1.

Linear Kalman filter (LKF) [54] is applied to the transformed linear problem. Our focus is on forecast and analysis of the augmented state $\boldsymbol{\xi}_A = [\xi, \theta_1, \theta_2]$ obeying

$$\frac{\mathrm{d}\boldsymbol{\xi}_A}{\mathrm{d}t} = \mathbf{A}\boldsymbol{\xi}_A, \quad \text{with} \quad \mathbf{A} = \begin{pmatrix} 010 \\ 000 \end{pmatrix}. \tag{B.3}$$

LKF relies on the Gaussianity of the involved distributions, such that it is sufficient to propagate (forecast) and update (analysis) only the mean $\langle \boldsymbol{\xi}_A \rangle$ and covariance $\mathbf{P}_A = \langle (\boldsymbol{\xi}_A - \langle \boldsymbol{\xi}_A \rangle)(\boldsymbol{\xi}_A - \langle \boldsymbol{\xi}_A \rangle)^\top \rangle$ of the distribution describing the augmented state. Propagation of $\langle \boldsymbol{\xi}_A \rangle$ and $\mathbf{P}_A$ on the time interval $[t_{m-1}, t_m]$ between the previous and the current assimilation times,

t_{m-1} and t_m, respectively, is given by

$$\begin{aligned}
\frac{\mathrm{d}\langle \boldsymbol{\xi}_A \rangle}{\mathrm{d}t} &= \mathbf{A}\,\langle \boldsymbol{\xi}_A \rangle, \qquad\qquad \text{subject to} \quad \langle \boldsymbol{\xi}_A(t_{m-1}) \rangle = \langle \boldsymbol{\xi}_A \rangle_{m-1|m-1}, \\
\frac{\mathrm{d}\mathbf{P}_A}{\mathrm{d}t} &= \mathbf{A}\mathbf{P}_A + \mathbf{P}\mathbf{A}^\top + \mathbf{Q}, \qquad \text{subject to} \quad \mathbf{P}_A(t_{m-1}) = \mathbf{P}_{A,m-1|m-1}. \quad \text{(B.4)}
\end{aligned}$$

Here, the initial conditions are the mean and the covariance of the previous assimilation step, respectively, and $\mathbf{Q} = \mathrm{diag}\,[1, 0, 0]$ is the covariance of the model error. The mean and covariance at $m = 1$ are initialized as $\langle \boldsymbol{\xi}_A \rangle_{0|0} = [\xi_0, \theta_1^{(0)}, \theta_2^{(0)}]$ and $\mathbf{P}_{A,0|0} = \mathrm{diag}[0, (\sigma_{\theta_1}^{(0)})^2, (\sigma_{\theta_2}^{(0)})^2]$, respectively. A solution of (B.4) at $t = t_m$ (analytical in this case) yields $\langle \boldsymbol{\xi}_A(t_m) \rangle = \langle \boldsymbol{\xi}_A \rangle_{m|m-1}$ and $\mathbf{P}_A(t_m) = \mathbf{P}_{A,m|m-1}$, i.e., the forecast mean and covariance conditional on the observations up to t_{m-1}. The mean and covariance are then updated at the assimilation time t_m according to

$$\langle \boldsymbol{\xi}_A \rangle_{m|m} = \langle \boldsymbol{\xi}_A \rangle_{m|m-1} + \mathbf{K}\left(\hat{x}_m - h(\langle \boldsymbol{\xi}_A \rangle_{m|m-1})\right), \quad \mathbf{P}_{A,m|m} = (\mathbf{I} - \mathbf{K}\mathbf{H})\,\mathbf{P}_{A,m|m-1}, \tag{B.5}$$

where the Kalman gain matrix is defined as $\mathbf{K} = \mathbf{P}_{A,m|m-1}\mathbf{H}^\top(\mathbf{H}\mathbf{P}_{A,m|m-1}\mathbf{H}^\top + \mathbf{R})^{-1}$, $\mathbf{I}$ is the identity matrix, $\mathbf{H} = \frac{\mathrm{d}h}{\mathrm{d}\boldsymbol{\xi}_A}(\boldsymbol{\xi}_A = \langle \boldsymbol{\xi}_A \rangle_{m|m-1})$ is the linearized observation map, and $\mathbf{R} = [\sigma_\varepsilon^2]$ is the variance of the observations. Kalman filtering for this problem introduces a linearization approximation due to the nonlinearity of the observation map.

Appendix C. Acknowledgements

This work was supported in part by Air Force Office of Scientific Research under award number FA9550-21-1-0381, and National Science Foundation under award number 2100927.

References

[1] H. Risken, The Fokker-Planck Equation, Springer, 1996.

[2] L. D. Landau, E. M. Lifshitz, Statistical Physics, Part 1, Elsevier, Amsterdam, 1980.

[3] D. M. Tartakovsky, P. A. Gremaud, Method of distributions for uncertainty quantification, in: R. Ghanem, D. Higdon, H. Owhadi (Eds.), Handbook of Uncertainty Quantification, Springer, 2015, pp. 763–783. doi:10.1007/978-3-319-12385-1-27.

[4] M. Branicki, A. J. Majda, Fundamental limitations of polynomial chaos for uncertainty quantification in systems with intermittent instabilities, Commun. Math. Sci. 11 (1) (2013) 55–103.

[5] P. Wang, A. M. Tartakovsky, D. M. Tartakovsky, Probability density function method for Langevin equations with colored noise, Phys. Rev. Lett. 110 (14) (2013) 140602. doi:10.1103/PhysRevLett.110.140602.

[6] D. A. Barajas-Solano, A. M. Tartakovsky, Probabilistic density function method for nonlinear dynamical systems driven by colored noise, Phys. Rev. E 93 (5) (2016) 052121.

[7] T. Maltba, P. A. Gremaud, D. M. Tartakovsky, Nonlocal PDF methods for Langevin equations with colored noise, J. Comput. Phys. 367 (2018) 87–101. doi:10.1016/j.jcp.2018.04.023.

[8] T. E. Maltba, H. Zhao, D. M. Tartakovsky, Autonomous learning of nonlocal stochastic neuron dynamics, J. Cogn. Neurodyn. 16 (2022) 683–705. doi:10.1007/s11571-021-09731-9.

[9] C. K. Wikle, L. M. Berliner, A Bayesian tutorial for data assimilation, Physica D 230 (1-2) (2007) 1–16.

[10] G. Evensen, Data assimilation: the ensemble Kalman filter, Springer, 2009.

[11] F. Boso, D. M. Tartakovsky, Learning on dynamic statistical manifolds, Proc. Roy. Soc. A 476 (2239) (2020) 20200213. doi:10.1098/rspa.2020.0213.

[12] F. Boso, S. V. Broyda, D. M. Tartakovsky, Cumulative distribution function solutions of advection-reaction equations with uncertain parameters, Proc. R. Soc. A 470 (2166) (2014) 20140189. doi:10.1098/rspa.2014.0189.

[13] F. Boso, D. M. Tartakovsky, Data-informed method of distributions for hyperbolic conservation laws, SIAM J. Sci. Comput. 42 (1) (2020) A559–A583. doi:10.1137/19M1260773.

[14] S. Kullback, Information theory and statistics, Courier Corporation, 1997.

[15] D. M. Blei, A. Kucukelbir, J. D. McAuliffe, Variational inference: A review for statisticians, J. Am. Stat. Assoc. 112 (518) (2017) 859–877.

[16] G. Peyré, M. Cuturi, Computational optimal transport: With applications to data science, Found. Trends Mach. Learn. 11 (5-6) (2019) 355–607.

[17] Y. Chen, W. Li, Wasserstein natural gradient in statistical manifolds with continuous sample space, arXiv:1805.08380 (2018).

[18] F. Kunstner, P. Hennig, L. Balles, Limitations of the empirical Fisher approximation for natural gradient descent, in: Adv. Neural Inf. Process. Sys., 2019, pp. 4156–4167.

[19] C. Villani, Topics in optimal transportation, Vol. 58, American Mathematical Society, 2003.

[20] C. Frogner, C. Zhang, H. Mobahi, M. Araya, T. A. Poggio, Learning with a Wasserstein loss, in: Adv. Neural Inf. Process. Sys., 2015, pp. 2053–2061.

[21] M. Arjovsky, S. Chintala, L. Bottou, Wasserstein GAN, arXiv:1701.07875 (2017).

[22] J. Neyman, E. L. Scott, Consistent estimates based on partially consistent observations, Econometrica (1948) 1–32.

[23] P. M. Esfahani, D. Kuhn, Data-driven distributionally robust optimization using the Wasserstein metric: Performance guarantees and tractable reformulations, Math. Progr. 171 (1-2) (2018) 115–166.

[24] F. Boso, D. Boskos, J. Cortés, S. Martínez, D. M. Tartakovsky, Dynamics of data-driven ambiguity sets for hyperbolic conservation laws with uncertain inputs, SIAM J. Sci. Comput 43 (2021) A2102–A2129.

[25] Y. Ollivier, L. Arnold, A. Auger, N. Hansen, Information-geometric optimization algorithms: A unifying picture via invariance principles, J. Mach. Learn. Res. 18 (1) (2017) 564–628.

[26] Y. Li, Y. Cheng, X. Li, H. Wang, X. Hua, Y. Qin, Bayesian nonlinear filtering via information geometric optimization, Entropy 19 (12) (2017) 655.

[27] S.-I. Amari, Information geometry and its applications, Vol. 194, Springer, 2016.

[28] W. Li, G. Montúfar, Natural gradient via optimal transport, Inform. Geom. 1 (2) (2018) 181–214.

[29] Y. Ollivier, Online natural gradient as a Kalman filter, El. J. Stat. 12 (2) (2018) 2930–2961.

[30] Y. Ollivier, The Extended Kalman Filter is a natural gradient descent in trajectory space, arXiv:1901.00696 (2019).

[31] A. Takatsu, Wasserstein geometry of Gaussian measures, Osaka J. Math. 48 (4) (2011) 1005–1026.

[32] L. Malagò, L. Montrucchio, G. Pistone, Wasserstein Riemannian geometry of positive definite matrices, arXiv:1801.09269 (2018).

[33] S.-I. Amari, R. Karakida, M. Oizumi, Information geometry connecting Wasserstein distance and Kullback–Leibler divergence via the entropy-relaxed transportation problem, Information Geom. 1 (1) (2018) 13–37.

[34] J. Martens, New insights and perspectives on the natural gradient method, Journal of Machine Learning Research 21 (2020) 1–76.

[35] F. Boso, A. Marzadri, D. M. Tartakovsky, Probabilistic forecasting of nitrogen dynamics in hyporheic zone, Water Resour. Res. 54 (7) (2018) 4417–4431. doi:10.1029/2018WR022525.

[36] A. A. Alawadhi, F. Boso, D. M. Tartakovsky, Method of distributions for water-hammer equations with uncertain parameters, Water Resour. Res. 54 (11) (2018) 9398–9411.

[37] A. H. Jazwinski, Stochastic Processes and Filtering, Dover Publications, 1970.

[38] H. J. Yang, F. Boso, H. A. Tchelepi, D. M. Tartakovsky, Probabilistic forecast of flow in porous media with uncertain properties, Water Resour. Res. (2019).

[39] C. P. Robert, G. Casella, G. Casella, Monte Carlo statistical methods, Vol. 2, Springer, 2004.

[40] J. Bakarji, D. M. Tartakovsky, Data-driven discovery of coarse-grained equations, J. Comput. Phys. 434 (2021) 110219.

[41] R. Herzog, K. Kunisch, Algorithms for PDE-constrained optimization, GAMM-Mitteilungen 33 (2) (2010) 163–176.

[42] A. Guzman, Derivatives and integrals of multivariable functions, Springer, 2012.

[43] H. Voss, M. J. Bünner, M. Abel, Identification of continuous, spatiotemporal systems, Phys. Rev. E 57 (1998) 2820–2823.

[44] M. D. Schmidt, H. Lipson, Co-evolving fitness predictors for accelerating and reducing evaluations, in: R. L. Riolo, T. Soule, B. Worzel (Eds.), Genetic Programming Theory and Practice IV, Vol. 5 of Genetic and Evolutionary Computation, Springer, Ann Arbor, 2006, Ch. 17.

[45] C. M. Bishop, Pattern Recognition and Machine Learning, Springer-Verlag, New York, 2006.

[46] M. Spivak, Calculus on manifolds: A modern approach to classical theorems of advanced calculus, Harper Collins, 1965.

[47] M. Raissi, P. Perdikaris, G. E. Karniadakis, Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations, J. Comput. Phys. 378 (2019) 686–707.

[48] A. Gupta, N. Shukla, L. Marla, A. Kolbeinsson, K. Yellepeddi, How to incorporate monotonicity in deep networks while preserving flexibility?, arXiv:1909.10662 (2019).

[49] G. Ansmann, Efficiently and easily integrating differential equations with JiTCODE, JiTCDDE, and JiTCSDE, Chaos 28 (4) (2018) 043116. doi:10.1063/1.5019320.

[50] C. Rackauckas, Q. Nie, Adaptive methods for stochastic differential equations via natural embeddings and rejection sampling with memory, Discrete Contin. Dyn. Syst. Ser. B 22 (7) (2017) 2731.

[51] J. E. Guyer, D. Wheeler, J. A. Warren, FiPy: Partial differential equations with Python, Comput. Sci. Engrg. 11 (3) (2009) 6–15.

[52] D. C. Liu, J. Nocedal, On the limited memory BFGS method for large scale optimization, Math. Program. 45 (1-3) (1989) 503–528.

[53] J. Nocedal, S. Wright, Numerical Optimization, 2nd Edition, Springer, 2006.

[54] R. E. Kalman, A new approach to linear filtering and prediction problems, 1960.

[55] M. Dentz, D. M. Tartakovsky, Probability density functions for passive scalars dispersed in random velocity fields, Geophys. Res. Lett. 37 (2010) L24406. doi:10.1029/2010GL045748.

[56] F. Boso, D. M. Tartakovsky, The method of distributions for dispersive transport in porous media with uncertain hydraulic properties, Water Resources Research 52 (6) (2016) 4700–4712.