Mathematisches Forschungsinstitut Oberwolfach

Report No. 26/2026

DOI: 10.4171/OWR/2026/26

Mathematical Foundations of Digital Twins

Organized by
Omar Ghattas, Austin
Matthias Heinkenschloss, Houston
Sebastian Reich, Potsdam
Claudia Schillings, Berlin

21 June – 26 June 2026

Abstract. A digital twin (DT) is a computational model that evolves over time to persistently represent the structure, behavior, and context of a unique physical system or process. DTs are characterized by a dynamic and continuous two-way flow of information between the computational model and the physical system. Data streams from the physical system are assimilated into the computational model to reduce uncertainties and improve predictions of the model, which in turn is used as a basis for controlling the physical system, optimizing data acquisition, and providing decision support. The DT paradigm brings along challenging mathematical and statistical questions which were at the center of the workshop.

Mathematics Subject Classification (2020): 35Q92, 35R30, 49M25, 60G99, 60H35, 62G05, 62M20, 65C05, 65K99, 65J05.

License: Unless otherwise noted, the content of this report is licensed under CC BY SA 4.0.

Introduction by the Organizers

This workshop focused on the mathematical and statistical foundation of the digital twin paradigm. Presentations were structured according to key digital twin elements: (1) the digital twin virtual representation in the form of evolution equations, generative modeling, operator learning, and reduced order modeling, (2) the information flow from physical to virtual via data assimilation, Bayesian inference, and machine learning, (3) the information flow from virtual to physical leading to optimal control and optimization problems.

The workshop started with talks by Karl Kunisch and Dante Kalise on optimal control. Houman Owhadi presented work on generative modeling approaches

to learned Markov decision processes. Subsequent talks by Georg Stadler and Femke Vossepoel focused on rare event estimates and data assimilation, respectively. Marina Höhne discussed explainable AI in the context of digital twins while Marc Bocquet highlighted connections between AI and data assimilation. Margaret Trautner presented work on operator learning. The talk of Philipp Guth returned to the problem of optimal control for uncertain linear systems. Reduced-order modelling was at the focus of the presentation of Boris Krämer. Contrasting learning and coupling of measure approaches to sampling were presented in talks by Ricardo Baptista and Georg Gottwald. Dirk Hartmann discussed applications of digital twins in engineering. Estimation for partially observed diffusion processes were at the focus of talks by Wilhelm Stannat and Raul Tempone, who also added the control component. The talks by Ion Victor Gosea and Peter Benner centered on operator learning and reduced order modeling. Nisha Chandramoorthy presented simplified model settings to guide our understanding of generative modeling. Prashant Mehta made a connection between transformers and estimation from an optimal control perspective. The topic of rare event estimation was discussed by Tiangang Cui using tensor trains. Doubly intractable Bayesian inference and their Monte Carlo approximation were the focus of the talk by Robert Scheichl. Elaine Spiller talked about the implementation and efficiency of Gaussian process surrogates for geophysical hazards simulations. Sören Christensen discussed offline and online learning settings for drift estimation in an optimal control context. Lars Grüne presented efficient approximations for high-dimensional optimal control problems exploiting sparse coupling structures. Proximal importance sampling for Bayesian inference was the focus of the presentation by George Biros. Roland Herzog spoke about machine learning and optimal experimental design tools for discovering thermodynamic equations of states. Didier Henrion addressed the reduction of complexity in global optimization via low-rank polynomial and tensor-train approximations. Michael Hintermüller presented a combined scale bridging and machine learning framework to model fibrous structures.

The Thursday afternoon session was devoted to a discussion on a formal mathematical formulation of the digital twins paradigm and the need for bench mark test sets. The following continuous-time mathematical description of a digital twin was presented by Sebastian Reich and discussed throughout the session. In order to formalize the problem, it is assumed that the physical reality can be described by the following set of differential equations

$$\dot{Z}_t^\dagger = f^\dagger(Z_t^\dagger, U_t), \tag{1a}$$

$$\dot{Y}_t^\dagger = h^\dagger(Z_t^\dagger) + \dot{V}_t^\dagger. \tag{1b}$$

Here, $V_t^\dagger$ denotes a time-dependent function. Except for the observed $\{Y_s^\dagger\}_{s\leq t}$ the physical reality is inaccessible to the virtual environment. The control $\{U_s\}_{s\leq t}$, on the other hand, is provided by the virtual environment. We assume that the

forward model describing a virtual representation of (1) is of the form

$$\dot{X}_t = f(X_t, U_t), \tag{2a}$$

$$\dot{Y}_t = h(X_t) + R^{1/2}\dot{W}_t. \tag{2b}$$

The functions f and h and the positive definite matrix R characterize the temporal evolution of the virtual model and W_t denotes Brownian motion.

For a given control U_t, data $\dot{Y}_t^\dagger$ collected from the physical object are assimilated into the virtual model using the mean-field differential equations

$$\dot{X}_t = f(X_t, U_t) + K(X_t, \mu_t)\left(\dot{Y}_t^\dagger - \dot{Y}_t\right), \tag{3a}$$

$$\dot{Y}_t = h(X_t) + R^{1/2}\dot{W}_t. \tag{3b}$$

Here, μ_t denotes the law of X_t conditioned on the observed data $\{Y_s^\dagger\}_{s\le t}$. The gain $K(x,\mu)$ satisfies an elliptic partial differential equation, which follows from the requirement that μ_t satisfies the associated Kushner–Stratonovitch equation. We note that the derivation of the Kushner–Stratonovitch equation assumes that the virtual reality is a replica of the physical reality; that is,

$$\dot{Z}_t^\dagger = f(Z_t^\dagger, U_t), \tag{4a}$$

$$\dot{Y}_t^\dagger = h(Z_t^\dagger) + R^{1/2}\dot{W}_t^\dagger \tag{4b}$$

with $W_t^\dagger$ denoting Brownian motion. We call this scenario the ideal digital twin setting. However, important challenges for the digital twin paradigm follow from discrepancies between physical and virtual realities; that is, between (1) and (2).

So far we have assumed that the control U_t is given. Finding the control requires planning for the future $\tau > t$. We formulate the planning problem in the form of a receding horizon stochastic optimal control problem and introduce, for example, the functional

$$J(U_{[t,t+T]}; \mu_t) = \mathbb{E}^\ddagger\left[\int_t^{t+T}\left(\mu_\tau[c] + \frac{1}{2}\|U_\tau\|^2\right)\mathrm{d}\tau + \mu_{t+T}[\psi]\right], \tag{5}$$

where $c(x)$ and $\psi(x)$ denote cost functions,

$$\mu_\tau[c] = \int c(x)\mu_\tau(x)\mathrm{d}x, \quad \mu_{t+T}[\psi] = \int \psi(x)\mu_{t+T}(x)\mathrm{d}x \tag{6}$$

their expectation values, and μ_τ the law of X_τ conditioned on virtual observations $\{\mathrm{Y}_\tau^\ddagger\}_{t\le\tau\le t+T}$. Here, the virtual planning process X_τ, $\tau \ge t$, satisfies

$$\dot{\mathrm{X}}_\tau = f(\mathrm{X}_\tau, U_\tau) + K(\mathrm{X}_\tau, \mu_\tau)\left(\dot{\mathrm{Y}}_\tau^\ddagger - \dot{\mathrm{Y}}_\tau\right), \tag{7a}$$

$$\dot{\mathrm{Y}}_t = h(\mathrm{X}_\tau) + R^{1/2}\dot{\mathrm{W}}_\tau \tag{7b}$$

subject to initial conditions $\mathrm{X}_t \sim \mu_t$ and $\mu_t = \mu_t$. Expectation $\mathbb{E}^\ddagger$ is taken over the virtual observations $\mathrm{Y}_\tau^\ddagger$, $\tau \ge t$.

Invoking again the ideal twin setting, the statistics of virtual observations can be represented via

$$\dot{\mathrm{Y}}^{\ddagger}_{\tau} = \mu_{\tau}[h] + R^{1/2}\dot{\mathrm{W}}^{\ddagger}_{\tau}, \tag{8}$$

where $\mathrm{W}^{\ddagger}_{\tau}$ denotes Brownian motion. Finally, the desired control U_t is obtained by minimizing (5) with respect to

$$U_{[t,t+T]} = \{u_{\tau}(\mu_{\tau})\}_{t\leq\tau<t+T} \tag{9}$$

over suitable families of feedback control laws $u_{\tau}(\mu)$.

The stochastic optimal control problem, defined by (5), (7), and (8), falls under the broad class of stochastic mean-field games with a common noise driver; here $\mathrm{W}^{\ddagger}_{\tau}$. The desired optimal control laws $u^{*}_{\tau}(\mu)$ follow from the Hamilton–Jacobi–Bellman equation for the value function $v_{\tau}(\mu)$ associated with (5), (7), and (8). Finally, $U_t = u^{*}_{t}(\mu_t)$.

The performance of different virtual representations of one and the same physical reality can be assessed by comparing scores; for example, the continuously ranked probability score based on the cumulative distribution functions of $G_t = g(Y_t)$ and $G^{\dagger}_{t} = g(Y^{\dagger}_{t})$ for given scalar-valued functions g.

The discussion on a mathematical formulation of the digital twin paradigm was followed by the presentation of a bench mark test problem by Omar Ghattas.

Workshop: Mathematical Foundations of Digital Twins

Table of Contents

Karl Kunisch (joint with Donato Vasquez-Varas)
A learning based approximation of semi-concave functions with applications to optimal control 1617

Dante Kalise (joint with Michael Herty, Yuyang Huang, Nikolas Kantas, Hicham Kouhkouh)
Global optimization through a control-theoretic lens 1618

Houman Owhadi (joint with Zongren Zou, Théo Bourdais, Ricardo Baptista)
Graphical conditional generative modeling for digital twin modeling 1621

Georg Stadler (joint with Winston Liang, Timo Schorlepp)
Updating failure estimates in digital twins 1622

Femke Vossepoel (joint with Max Frei, Nikolaj Mücke)
On the use of data assimilation in digital twins 1625

Marina Marie-Claire Höhne
Developing trustworthy digital twins by incorporating explainable AI: Detecting learned artefacts for faithful digital twin modeling 1627

Marc Bocquet (joint with Tobias S. Finn, Sibo Cheng, Alban Farchi, Charlotte Durand, Yumeng Chen, Ivo Pasmans, and Alberto Carrassi)
When AI provides the tools DA practitioners have been craving 1629

Margaret Trautner (joint with Andrew M. Stuart)
Enforcing Dirichlet boundary conditions in operator learning 1632

Philipp A. Guth (joint with Karl Kunisch, Sérgio S. Rodrigues, Jesper Schröder)
A digital twin framework for output-feedback control of uncertain linear systems 1634

Boris Krämer (joint with Nick Corbin)
System-theoretic reduced-order digital twin prototypes 1636

Ricardo Baptista (joint with Andrew M. Stuart, Son Tran)
Contrastive learning: A mathematical perspective 1640

Georg A. Gottwald
Localized Schrödinger bridge sampler 1642

Dirk Hartmann
Digital twins for effective thermal management 1643

Wilhelm Stannat (joint with Gregor Pasemann, Jan Szalankiewicz)
State-parameter estimation of partially observed stochastic (partial) differential equations 1645

Raúl Tempone (joint with Christian Bayer, Saifeddine Ben Naamia, Boualem Djehiche, Eliza Rezvanova, Erik von Schwerin)
From data assimilation to active sensing: Belief-space optimal control for digital twins 1648

Ion Victor Gosea (joint with Luisa Peterson, Kai Sundmacher, Peter Benner)
Scientific machine learning and digital twinning through stable operator inference: CO_2 methanation test case 1652

Peter Benner (joint with Leonidas Gkimisis, Ion Victor Gosea, Pawan K. Goyal, Igor Pontes Duff)
Certified stability of learned nonlinear reduced models 1655

Nisha Chandramoorthy (joint with Andrew Dennehy, Ramchandran Muthukumar and Rebecca Willett)
Understanding generative models for scientific applications 1657

Prashant G. Mehta (joint with Heng-Sheng Chang and Jin-Won Kim)
Transformer as a nonlinear predictor 1660

Tiangang Cui (joint with Alex de Beer, Sergey Dolgov, Karina Koval, Robert Scheichl, Georg Stadler)
Rare event estimation using tensor trains and large deviation theory . . . 1663

Robert Scheichl (joint with Josef Martínek, Julian Hofstadler, Alexander M. G. Cox)
Reducing complexity in a class of sequential Monte Carlo samplers 1665

Elaine T. Spiller
Fast uncertainty quantified probabilistic hazard maps for decision support 1667

Sören Christensen (joint with Claudia Strauch, Lukas Trottner)
Some lessons from incorporating data into stochastic control problems . . 1668

Lars Grüne (joint with Dante Kalise, Luca Saluzzi, and Mario Sperl)
Separable approximations of optimal value functions, decaying sensitivity, and neural networks 1670

George Biros (joint with Youguang Chen)
Proximal importance sampling 1671

Roland Herzog (joint with Viktor Martinek)
Machine learning for thermodynamic property modeling 1673

Didier Henrion (joint with Llorenç Balada Gaggioli and Milan Korda)
Global optimization of tensor polynomials 1675

Abstracts

A learning based approximation of semi-concave functions with applications to optimal control

Karl Kunisch
(joint work with Donato Vasquez-Varas)

Our main goal is to find an optimal feedback law for the infinite horizon problem:

$$(1) \qquad (P) \qquad \min_{u\in L^2(0,\infty;\mathbb{R}^m)} J(u,y_0) := \int_0^\infty \left(\ell(y(t)) + \frac{\beta}{2}|u(t)|^2\right) dt$$

where $y \in H^1(0,\infty;\mathbb{R}^d)$ is the unique solution of

$$(2) \qquad y'(t) = f(y(t)) + B(y(t))u(t), \quad t \in (0,\infty), \quad y(0) = y_0$$

with $\beta > 0$, $\ell \in C^1(\overline{\Omega})$, $\ell(0) = 0$, $\ell \geq 0$, $f \in C^1(\overline{\Omega})$, $f(0) = 0$, $B \in C^1(\overline{\Omega};\mathbb{R}^{d\times m})$, and $\Omega \subset \mathbb{R}^d$ open, bounded and convex.

The main tool for investigating this problem involves the value function defined as $V(y_0) = \min_{u\in L^2(0,\infty;\mathbb{R}^m)} J(u,y_0)$. If $V \in C^1(\Omega)$ then it satisfies the Hamilton-Jacobi-Bellman (HJB) equation given by

$$\mathcal{H}(\nabla V(y), y) := \max_{u\in\mathbb{R}^m} \left\{-\nabla V(y)^\top (f(y) + B(y)u) - \frac{\beta}{2}u^2 - \ell(y)\right\} = 0,$$

with boundary conditions $V(0) = 0$ and $\nabla V(0) = 0$. If ∇V is available, then the optimal feedback law is given by:

$$(3) \qquad u(y) = -\frac{1}{\beta}B^\top(y)\nabla V(y),$$

and we may close the loop to obtain:

$$(4) \qquad y'(t) = f(y(t)) - B^\top B\nabla V(y(t)),\ y(0) = y_0.$$

Since no control constraints are incorporated in the HJB equation it simplifies to

$$\nabla V(y)^\top f(y) - \frac{1}{2\beta}|B^\top\nabla V(y)|^2_{\mathbb{R}^d} + \ell(y) = 0.$$

To solve this equation and hence to use (3) in practice one is confronted with at least two difficulties. One results from the curse of dimensionality since (4) is a hyperbolic equation in the dimension given by that of the state space of (2), the other one since V cannot be assumed to be C^1, in general.

Still, in view of the practical relevance of feedback control much effort has been made to develop approaches for solving the HJB equation in high dimensions. In this talk we first survey recent developments, including policy iteration in combination with tensor train techniques, data driven methods, the averaged approximation scheme, Taylor expansions, and max-plus algebra techniques.

Subsequently we focus on our recent work in which we exploit the fact that in many situations the solution to HJB, while not C^1, is at least semi-concave. Semi-concave functions can be expressed as the pointwise infimum of a family

of C^1 functions. In practice we expect that this family is finite. We develop a semi-concavity preserving approximation scheme combining it with a hyperbolic cross framework which naturally arises in with Legrendre polynomials which we utilize in the inf-representation of V by C^1 functions. Approximation results for this scheme for both the value function and the feedback law are derived.

Further, numerical results are presented, with special emphasis on an example for which, in a parameter dependent manner, the value function switches from being C^1 to being semi-concave.

References

[1] K. Kunisch and D. Vásquez-Varas, *Consistent Smooth Approximation of Feedback Laws for Infinite Horizon Control Problems with Non-Smooth Value Functions*, Journal of Differential Equations, 411, pp. 438–477, 2024.

[2] K. Kunisch and D. Vásquez-Varas, *Convergence of Machine Learning Methods for Feedback Control Laws: Averaged Feedback Learning Scheme and Data Driven Methods*, arXiv:2407.18403, 2024.

[3] K. Kunisch and D. Vásquez-Varas, *Structure Preserving Approximation of Semiconcave Functions*, arXiv:2602.07770, 2026.

Global optimization through a control-theoretic lens

Dante Kalise

(joint work with Michael Herty, Yuyang Huang, Nikolas Kantas, Hicham Kouhkouh)

Global optimization problems of the form

$$\min_{\mathbf{x}\in\mathbb{R}^d} f(\mathbf{x}) \tag{1}$$

are ubiquitous in computational science and engineering, and arise naturally within the digital twin paradigm: model calibration, optimal experimental design, and decision support all lead to landscapes which are non-convex, possibly non-smooth, and for which gradients may be expensive or unavailable. In practice, such problems are routinely addressed with metaheuristics, which are effective yet often lack a quantitative convergence theory. This talk presents a research programme that recasts (1) as an optimal stabilization problem for a controlled dynamical system, obtaining a globally informative surrogate of the landscape, exponential convergence guarantees, and a scalable particle-based realization.

The starting point is the well-known correspondence between optimization algorithms and continuous-time dynamics: gradient descent is the explicit Euler discretization of the gradient flow $\dot{\mathbf{x}} = -\nabla f(\mathbf{x})$, while Nesterov's accelerated method corresponds to a damped second-order ODE $\ddot{\mathbf{x}} + \frac{3}{t}\dot{\mathbf{x}} + \nabla f(\mathbf{x}) = 0$ [4], with closed-loop damping strategies studied in [5]. This raises a natural question: instead of prescribing a dissipation mechanism, can one *synthesize* an optimal feedback law for global optimization?

Global optimization as optimal stabilization. We embed (1) into an infinite-horizon, discounted optimal control problem

(2)
$$\inf_{\mathbf{u}(\cdot)\in\mathcal{U}} \mathcal{J}(\mathbf{u}(\cdot),\mathbf{x}) := \int_0^\infty e^{-\mu t}\left(f(\mathbf{y}(t)) + \tfrac{\epsilon}{2}\|\mathbf{u}(t)\|^2\right) dt\,, \quad \dot{\mathbf{y}}(t) = \mathbf{u}(t)\,, \quad \mathbf{y}(0) = \mathbf{x}\,,$$

with $\mu, \epsilon > 0$. The running cost rewards trajectories that reach and remain in low-objective regions, while the quadratic penalty limits the control effort spent navigating the landscape. The value function

$$V_\mu(\mathbf{x}) := \inf_{\mathbf{u}} \mathcal{J}(\mathbf{u}(\cdot),\mathbf{x})$$

satisfies the stationary first-order Hamilton–Jacobi–Bellman (HJB) equation

$$\mu V_\mu(\mathbf{x}) + \tfrac{1}{2\epsilon}\|\nabla V_\mu(\mathbf{x})\|^2 = f(\mathbf{x})\,, \tag{3}$$

and the optimal control is realized in feedback form as $\mathbf{u}^*(\mathbf{x}) = -\frac{1}{\epsilon}\nabla V_\mu(\mathbf{x})$, that is, gradient descent over the surrogate V_μ rather than over f itself. In the limiting case of constrained controls and vanishing discount and penalization, one recovers the Eikonal-type equation with vanishing Lagrangian studied in [6], while parabolic regularizations of (3) connect to the deep relaxation PDEs of [7].

Stability and convergence rates. The first set of results [1] establishes that the closed-loop dynamics indeed solve the global optimization problem. Let $\mathcal{M}$ denote the set of global minimizers of f and $\mathbf{y}^*_{\mathbf{x}}(\cdot)$ an optimal trajectory of (2). For f continuous and bounded, the family of closed-loop systems is practically asymptotically stable with respect to $\mathcal{M}$: for every $\eta > 0$ there exist $\mu > 0$ and $\tau = \tau(\eta,\mu)$ such that $\operatorname{dist}(\mathbf{y}^*_{\mathbf{x}}(t),\mathcal{M}) \le \eta$ for all $t \ge \tau$, in the spirit of stabilization results for discounted optimal control [8].
Quantitative rates require additional structure [2]. Normalizing $f_* := \min f$ via $\widetilde{V}_\mu := V_\mu - f_*/\mu$ and $\tilde{f} := f - f_*$, we introduce

Assumption (A): $\exists\, K > \mu > 0$ such that $K\,\widetilde{V}_\mu(\mathbf{x}) \le \tilde{f}(\mathbf{x}) \quad \forall\, \mathbf{x} \in \mathbb{R}^d$.

Substituting into the shifted HJB equation yields the inequality

$$\tfrac{1}{2\epsilon}\|\nabla\widetilde{V}_\mu\|^2 \ge (K-\mu)\widetilde{V}_\mu,$$

a Polyak–Łojasiewicz-type condition for the *value function* rather than for the objective. As a consequence, $\widetilde{V}_\mu$ is a Lyapunov function for the closed-loop dynamics, decaying as $\widetilde{V}_\mu(\mathbf{y}^*_{\mathbf{x}}(t)) \le e^{-(K-\mu)t}\widetilde{V}_\mu(\mathbf{x})$, and optimal trajectories enjoy three levels of exponential convergence:

(i) *variational*: $\mu V_\mu(\mathbf{y}^*_{\mathbf{x}}(t)) \to \min f$, under Assumption (A) alone;
(ii) *objective*: $f(\mathbf{y}^*_{\mathbf{x}}(t)) \to \min f$, under an additional regularity assumption on f, e.g. Hölder continuity;
(iii) *pathwise*: $\operatorname{dist}(\mathbf{y}^*_{\mathbf{x}}(t),\mathcal{M})^2 \le \frac{C_2}{C_1}e^{-\delta(t-\tau)}\operatorname{dist}(\mathbf{x},\mathcal{M})^2$ for $t \ge \tau$, under a local growth condition near $\mathcal{M}$.

A sufficient condition for the full hierarchy is local quadratic growth,

$$\tfrac{c_1}{2}\operatorname{dist}(\mathbf{x},\mathcal{M})^2 \le \tilde{f}(\mathbf{x}) \le \tfrac{c_2}{2}\operatorname{dist}(\mathbf{x},\mathcal{M})^2$$

near $\mathcal{M}$, which yields Assumption (A) with an explicit Riccati-type constant $K(\mu, c_1, c_2)$. This covers, among others, strongly convex smooth objectives, functions with finitely many non-degenerate global minimizers, and soft k-means or Gaussian-mixture likelihoods; local linear growth or metric regularity of ∂f also suffice. Importantly, the estimates are robust under quasi-optimality: they persist when the exact feedback $-\frac{1}{\epsilon}\nabla V_\mu$ is replaced by a sufficiently accurate approximation, which is the enabling property for the numerical realization below. Numerical validation on Himmelblau's function, whose four global minima form a disconnected $\mathcal{M}$, confirms pathwise stabilization from arbitrary initializations at the predicted exponential rate.

A scalable realization: controlled consensus-based optimization. Direct HJB solvers suffer from the curse of dimensionality. Structured approximations mitigate this: sparse polynomial expansions $V_n = \sum_i c_i \phi_i$ [9], low-rank tensor train decompositions [10], and separable physics-informed architectures. These surrogates are gradient-free and globally informative within the computational domain, but their accuracy in high dimension is limited. We therefore couple a cheap HJB surrogate with consensus-based optimization (CBO) [11, 12], a gradient-free swarm of N particles

$$dX_t^i = -\lambda \left(X_t^i - v_\alpha(\rho_t^N)\right) dt + \beta\, \mathbf{u}_{\mathrm{HJB}}(X_t^i)\, dt + \sigma |X_t^i - v_\alpha(\rho_t^N)|\, dW_t^i\,,$$

where $v_\alpha(\rho_t^N)$ is the Laplace-principle weighted consensus point and $\mathbf{u}_{\mathrm{HJB}} \approx -\nabla V_\mu$ is the approximate feedback, entering as an additional drift with gain β ($\beta = 0$ recovers standard CBO). The two mechanisms are complementary: the feedback provides coarse global guidance from an inexpensive polynomial surrogate of V_μ, steering particles toward low-objective regions and reducing sensitivity to initialization, while the consensus dynamics refine the search through stochastic exploration and remain robust precisely in the quasi-optimality regime where the surrogate is inaccurate. Numerical experiments on the Ackley and Rastrigin benchmarks [3] demonstrate faster and more robust consensus formation than standard CBO, with expected squared Wasserstein errors $\mathbb{E}[W_2^2(\rho_t^N, \delta_{\mathbf{x}^*})]$ of machine precision order in tests up to dimension $d = 30$.

Outlook. Ongoing work includes a structural characterization of objective classes for which the surrogate construction scales, stabilization rates for the controlled CBO dynamics and the interplay between surrogate accuracy and consensus formation, gradient-based schemes derived from the optimal control formulation, extensions to non-smooth and constrained landscapes via metric regularity, and the coupling with mean-field and McKean–Vlasov control for distributional targets.

References

[1] Y. Huang, D. Kalise and H. Kouhkouh, *Non-convex global optimization as an optimal stabilization problem: dynamical properties*, arXiv:2511.10815 (2025).

[2] Y. Huang, D. Kalise and H. Kouhkouh, *Non-convex global optimization as an optimal stabilization problem: convergence rates*, arXiv:2511.11122 (2025).

[3] Y. Huang, M. Herty, D. Kalise and N. Kantas, *Fast and robust consensus-based optimization via optimal feedback control*, SIAM J. Sci. Comput. **48** (2026), A74–A102.

[4] W. Su, S. Boyd and E.J. Candès, *A differential equation for modeling Nesterov's accelerated gradient method: theory and insights*, J. Mach. Learn. Res. **17** (2016), 1–43.
[5] H. Attouch, R.I. Boţ and E.R. Csetnek, *Fast optimization via inertial dynamics with closed-loop damping*, J. Eur. Math. Soc. **25** (2023), 1985–2056.
[6] M. Bardi and H. Kouhkouh, *An Eikonal equation with vanishing Lagrangian arising in global optimization*, Appl. Math. Optim. **87** (2023), 49.
[7] P. Chaudhari, A. Oberman, S. Osher, S. Soatto and G. Carlier, *Deep relaxation: partial differential equations for optimizing deep neural networks*, Res. Math. Sci. **5** (2018), 30.
[8] V. Gaitsgory, L. Grüne and N. Thatcher, *Stabilization with discounted optimal control*, Systems Control Lett. **82** (2015), 91–98.
[9] D. Kalise and K. Kunisch, *Polynomial approximation of high-dimensional Hamilton–Jacobi–Bellman equations and applications to feedback control of semilinear parabolic PDEs*, SIAM J. Sci. Comput. **40** (2018), A629–A652.
[10] S. Dolgov, D. Kalise and K. Kunisch, *Tensor decomposition methods for high-dimensional Hamilton–Jacobi–Bellman equations*, SIAM J. Sci. Comput. **43** (2021), A1625–A1650.
[11] R. Pinnau, C. Totzeck, O. Tse and S. Martin, *A consensus-based model for global optimization and its mean-field limit*, Math. Models Methods Appl. Sci. **27** (2017), 183–204.
[12] J.A. Carrillo, Y.-P. Choi, C. Totzeck and O. Tse, *An analytical framework for consensus-based global optimization method*, Math. Models Methods Appl. Sci. **28** (2018), 1037–1066.

Graphical conditional generative modeling for digital twin modeling

Houman Owhadi

(joint work with Zongren Zou, Théo Bourdais, Ricardo Baptista)

Digital twin modeling, including control and data assimilation under model uncertainty, often faces an open-ended fidelity problem: adding variables, data streams, and time scales can indefinitely increase model complexity, ultimately producing systems that are difficult to maintain, validate, interpret, and use for stress or safety testing. As an alternative, one can seek parsimonious stochastic surrogate models built only on the variables needed to describe the relevant quantities of interest. We introduce a framework for discovering such variables from observational data by identifying which candidate inputs influence the full conditional law of a target quantity, rather than only its conditional mean. This distinction is essential in stochastic, coarse-grained, or partially observed systems, where dependencies may appear through changes in variability, tail behavior, multimodality, or uncertainty rather than through deterministic functional relationships. The framework couples conditional generative modeling, which learns the conditional distribution of the target given candidate inputs, with Gaussian-process-based analysis of variance (through kernel mode decomposition), which enables iterative pruning of non-influential inputs and interpretable structure discovery. In control settings, the resulting surrogate can be interpreted as a learned Markov decision process: the method identifies not only a transition model, but also the state, action, and memory variables needed to make the learned dynamics effectively Markovian. Across examples involving stochastic dynamical systems, missing variables, PDE control, reinforcement learning, and economic data, the discovered structures yield interpretable stochastic surrogates whose downstream performance is comparable to models trained on the full variable set.

References

[1] Z. Zou, T. Bourdais, R. Baptista, and H. Owhadi, *Graphical conditional generative modeling for digital twin modeling*, arXiv preprint `arXiv:2606.16219` [cs.CE], 2026.

Updating failure estimates in digital twins

Georg Stadler

(joint work with Winston Liang, Timo Schorlepp)

When a digital twin (DT) is used to make decisions about its physical twin (PT), one often needs to take into account the probability of rare but important events. Examples of such events include structural failure under stress, the collapse of financial markets, or risks due to extreme weather. Mathematically, we want to quantify the (evolution of the) tail of a quantity of interest (QoI) as data assimilation or Bayesian inference updates the model. Computing posterior samples and then studying the tail of the QoI empirically is expensive, doubly so for rare tail events. Our approach instead uses asymptotics for small probabilities together with measure concentration to replace sampling with optimization. We review the approach first for a fixed prior distribution and then present extensions for the posterior in a data assimilation problem.

Let $q : \boldsymbol{\theta} \in \mathbb{R}^n \to \mathbb{R}$ be a QoI map (typically involving a PDE solve), where $\boldsymbol{\theta}$ is a random, possibly high-dimensional parameter with density π_{pr}. We seek the probability of exceeding a given threshold $z \in \mathbb{R}$, i.e.,

$$P_{\mathrm{pr}}(z) := \mathbb{P}_{\mathrm{pr}}(q(\boldsymbol{\theta}) \geq z).$$

For example, $\boldsymbol{\theta}$ may be estimated material parameters, and q maps to a point stress. Building on ideas from large deviation theory and methods from structural engineering, we show in [4] that

$$P_{\mathrm{pr}}(z) \approx C(z) \exp\big(-I(\boldsymbol{\theta}^\star)\big) \quad \text{as } z \to \infty, \qquad \boldsymbol{\theta}^\star := \operatorname*{argmin}_{q(\boldsymbol{\theta}) \geq z} I(\boldsymbol{\theta}), \tag{1}$$

where $I(\boldsymbol{\theta})$ is the rate function (for Gaussian distributions π_{pr}, $I(\boldsymbol{\theta})$ is the negative log prior and thus quadratic), and the optimizer $\boldsymbol{\theta}^\star$ is called the *instanton* or the *most likely point*. The sub-exponential prefactor $C(z)$ can be approximated using second-order (i.e., Hessian) information at $\boldsymbol{\theta}^\star$. Writing $\hat{\boldsymbol{n}} = \nabla q(\boldsymbol{\theta}^\star)/\|\nabla q(\boldsymbol{\theta}^\star)\|$ for the outward normal to $\{q(\boldsymbol{\theta}) \equiv z\}$ and $\boldsymbol{H}_{\mathrm{LDT}} := (\boldsymbol{I}_n - \hat{\boldsymbol{n}}\hat{\boldsymbol{n}}^\top)\nabla^2 q(\boldsymbol{\theta}^\star)(\boldsymbol{I}_n - \hat{\boldsymbol{n}}\hat{\boldsymbol{n}}^\top)$ for the curvature of the event boundary projected onto the tangent space, a second-order reliability (SORM)-inspired approximation gives

$$C(z) \approx \Big[2\pi\|\boldsymbol{\theta}^\star\|^2 \prod_{i=1}^{n} \big(1 - \lambda\, \lambda_i(\boldsymbol{H}_{\mathrm{LDT}})\big)\Big]^{-1/2}, \qquad \lambda := \|\nabla I(\boldsymbol{\theta}^\star)\|/\|\nabla q(\boldsymbol{\theta}^\star)\|,$$

with $\lambda_i(\cdot)$ the i-th eigenvalue, λ is the Lagrange multiplier for the constraint in the optimization problem in (1), and we assumed for simplicity that $\boldsymbol{\theta} \sim \mathcal{N}(0, I)$. This requires derivatives of q (obtained typically using adjoint PDE solvers) and is asymptotically exact as $z \to \infty$, though it often performs well already at moderate

probabilities. Its main limitation is the assumption of a unique dominating point. Residual bias can be corrected with sampling, e.g., importance sampling from a shifted distribution, cross-entropy proposals, or control variates.

The preceding theory assumes a fixed prior π_{pr}. In a DT, however, the parameter distribution is continually updated by data through a (typically nonlinear) parameter-to-observation map F. Given $\boldsymbol{y}^{\mathrm{d}} = F(\boldsymbol{\theta}) + \boldsymbol{\eta}$ with $\boldsymbol{\theta} \sim \pi_{\mathrm{pr}} = \mathcal{N}(\boldsymbol{\theta}_0, C_0)$ and $\boldsymbol{\eta} \sim \mathcal{N}(\mathbf{0}, \Gamma_{\mathrm{obs}})$, and covariance operators C_0 and Γ_{obs}, one must instead quantify the tail probabilities of the *posterior*. This raises new challenges: the posterior is generally non-Gaussian even for a Gaussian prior, and the rate function, and thus the approximation (1), has no immediate posterior analog.

Since there is no small-noise limit driving concentration, our approach is to replace the rate function with the (unnormalized, negative log-) posterior

$$J^{\boldsymbol{y}^{\mathrm{d}}}(\boldsymbol{\theta}) := \tfrac{1}{2}\|F(\boldsymbol{\theta}) - \boldsymbol{y}^{\mathrm{d}}\|^2_{\Gamma_{\mathrm{obs}}} + \tfrac{1}{2}\|\boldsymbol{\theta} - \boldsymbol{\theta}_0\|^2_{C_0}.$$

The posterior tail probability is defined as

$$P_{\mathrm{post}}(z) := \mathbb{P}_{\mathrm{post}}(q(\boldsymbol{\theta}) \geq z) = \frac{1}{\hat{Z}(\boldsymbol{y}^{\mathrm{d}})} \int_{\{\boldsymbol{\theta}: q(\boldsymbol{\theta}) \geq z\}} \exp\big(-J^{\boldsymbol{y}^{\mathrm{d}}}(\boldsymbol{\theta})\big)\, d\boldsymbol{\theta},$$

where $\hat{Z}(\boldsymbol{y}^{\mathrm{d}})$ is the Bayesian evidence. We use an asymptotic argument based on posterior concentration as a result of rescaling the prior and likelihood, and then interpret the asymptotics as a $z \to \infty$ limit, similar as in [3]. This results in the most likely point under the posterior given by

$$\hat{\boldsymbol{\theta}}^{\star} := \operatorname*{argmin}_{\{\boldsymbol{\theta}: q(\boldsymbol{\theta}) \geq z\}} J^{\boldsymbol{y}^{\mathrm{d}}}(\boldsymbol{\theta}).$$

The corresponding second-order (Laplace-type) expansion has the form

$$P_{\mathrm{post}}(z) \approx \frac{\hat{C}(z)}{\hat{Z}(\boldsymbol{y}^{\mathrm{d}})} \exp\big(-J^{\boldsymbol{y}^{\mathrm{d}}}(\hat{\boldsymbol{\theta}}^{\star})\big), \tag{2}$$

where $\hat{C}(z)$ is a prefactor, formally analogous to the prior case, built from the Hessians of $J^{\boldsymbol{y}^{\mathrm{d}}}$ and q at $\hat{\boldsymbol{\theta}}^{\star}$ projected onto the tangent space as in the prior case. Moreover, $\hat{Z}(\boldsymbol{y}^{\mathrm{d}})$ is the evidence whose computation is difficult in itself, but it is unnecessary if one only needs ratios of failure probabilities. Note that (2) reduces to (1) if the data $\boldsymbol{y}^{\mathrm{d}}$ are non-informative about the parameters.

Next, we illustrate the application of this approach for a simple Darcy flow problem with two parameters. We consider the QoI $q(\theta_1, \theta_2) = p(1/2)$, where p solves $-\frac{d}{dx}(e^{\theta_1}\frac{d}{dx}p(x)) = 1$ on $[0,1]$ with $p(0) = 0$, $p(1) = \theta_2$. The data is $y^{\mathrm{d}} = p(1/4)$ observed with a draw of $\boldsymbol{\theta} \sim \mathcal{N}((0,100)^{\top}, I)$ and $\eta \sim \mathcal{N}(0, 0.1^2)$. Figure 1 shows that the estimate based on the second-order approach agrees closely with both the Monte Carlo reference and the naive integration across a range of thresholds.

The proposed approach can be extended to problems with high parameter dimensions using scalable PDE-constrained optimization solvers and eigenvalue estimation of second-derivative matrices as in [1, 3] for Bayesian inverse problems and rare event probability estimation with prior distributions. Main challenges remain

the estimation of the evidence, as well as verifying the dominant point assumption. Alternatives to estimating small probabilities over posteriors include using tensor train approximations of the posterior, [2], which could possibly be combined with our approach.

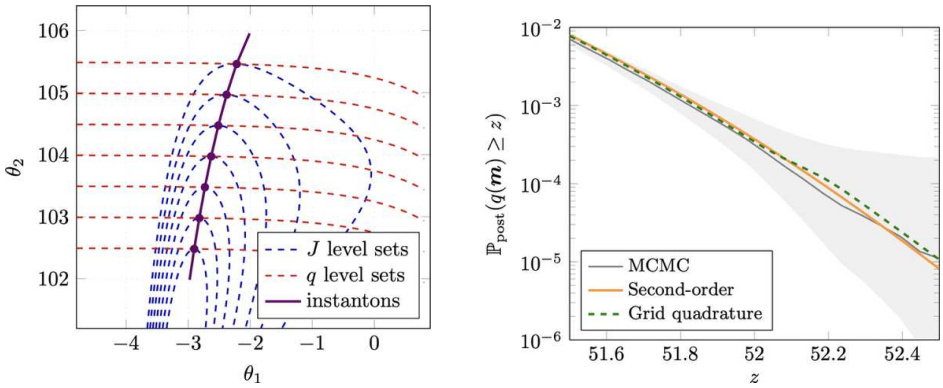

FIGURE 1. Left: Contours of the negative log-posterior J (blue dashed) and of the QoI q (red dashed) in the parameter space (θ_1, θ_2). Shown are several instantons corresponding to $z \in (52, 53)$. Right: Resulting probability estimation using brute-force MCMC sampling of the posterior (in gray we show the uncertainty), the proposed second-order method, and grid quadrature. All these results use a brute-force computation of the evidence.

REFERENCES

[1] T. Bui-Thanh, O. Ghattas, J. Martin, G. Stadler, *A computational framework for infinite-dimensional Bayesian inverse problems Part I: The linearized case, with application to global seismic inversion*, SIAM J. Sci. Comp. **35(6)**, (2013) A2494-A2523.

[2] T. Cui, S. Dolgov, R. Scheichl, *Deep importance sampling using tensor trains with application to a priori and a posteriori rare events*, SIAM J. Sci. Comp. **46(1)** (2024), C1-C34.

[3] T. Schorlepp, S. Tong, T. Grafke, G. Stadler, *Scalable methods for computing sharp extreme event probabilities in infinite-dimensional stochastic systems*, Stats. & Comp., **33** (2023), 137.

[4] S. Tong, E. Vanden-Eijnden, G. Stadler, *Extreme event probability estimation using PDE-constrained optimization and large deviation theory, with application to tsunamis*, Commun. Appl. Math. Comput. Sci. **16** (2021), 181-225.

On the use of data assimilation in digital twins

Femke Vossepoel
(joint work with Max Frei, Nikolaj Mücke)

Ensemble data-assimilation methods are widely used in digital twins to quantify uncertainties and support decision-making. Data assimilation updates the state (and/or the parameters) in such a way that end users can simulate scenarios as realistically as possible. The output of the scenario simulation can then help the users to make well-informed decisions.

The UrbanAIR digital twin (Horizon Europe Grant number 101188131) supports decisions related to the urban atmosphere. Choices in city planning, including green and blue infrastructure, affect the urban temperature and air quality, and simulations of this atmosphere under different boundary and initial conditions can be used by municipalities to make choices in their design of the urban landscape and in the implementation of safety measures. UrbanAIR features several numerical models, ranging from regional weather [1] and city-scale models [2] to models simulating the neighbourhood- and street level atmospheric dynamics [3, 4]. Using benchmark cases that represent the urban landscape [5, 6], we develop and test data assimilation schemes that estimate and update the boundary conditions of the model domain to obtain the most realistic air flow in a simulated urban environment. In this estimation, Bayes' theorem forms the underlying principle to combine the prior samples from the atmospheric models with available observations with data-assimilation methods. The ultimate objective of data assimilation is to estimate the true posterior probability distribution of the state and/or parameters (see, e.g. [7]).

To assess the performance of data-assimilation methods, various metrics are available. For state and parameter estimation, common measures include the root-mean-square error of the ensemble mean or the ensemble members with respect to the available observations. The variance of the state as simulated in the ensemble reflects the spread of the ensemble, which, if small, can indicate ensemble collapse. In forecasting applications, graphical diagnostics such as Taylor diagrams and rank histograms are often used to assess forecast skill and the representation of uncertainty. These diagnostics provide useful information, but they do not always reflect how well the method represents the underlying probability distribution of the state or parameters.

From a Bayesian perspective, the performance of a data-assimilation method should be assessed in terms of its ability to represent the posterior distribution of the state or parameters. This can in principle be measured using probabilistic distances such as the Wasserstein-2 metric or the Kullback–Leibler divergence, but these metrics are difficult to apply in high-dimensional problems. For this reason, [8] suggested the Continuous Ranked Probability Score (CRPS) as an alternative measure, which is a continuous variant of the Brier score: $\mathbb{E}|X-y|-\frac{1}{2}\mathbb{E}|X-X'| = \int\big(F(x)-\mathbf{1}\{x\geq y\}\big)^2 dx$, where x is the model variable of interest, y the observation thereof, μ the mean and σ the variance. Other metrics used in data-assimilation

applications include: the Root Mean Square Error (RMSE), $\sqrt{\mathbb{E}[(y-\mu_F)^2]}$; the Mean Absolute Error MAE, $\mathbb{E}[\,|y-\mu_F|\,]$; the Spread–skill, $\left|\,\mathrm{RMSE}/\sigma_F - 1\,\right|$; the Dawid–Sebastiani Score, DSS; $\mathbb{E}[(y-\mu_F)^2]/\sigma_F^2 + 2\ln\sigma_F$. A comparison of these metrics applied to a simplified geophysical problem illustrates how the CRPS is more informative than the other metrics considered.

Alternative diagnostics inspired by information theory are entropy ($H(x) = -\sum_i p(x_i)\log(p(x_i))$), mutual information $I(z;x)$, written as a function of the entropy and the joint entropy $H(z,x)$ ($I(z;x) = H(x) + H(z) - H(z,x)$) and the derived data assimilation efficiency $\mathcal{E}_{DA}$ that is defined as the ratio of the posterior mutual information $I(z;\hat{x})$ and the total information in model and observations $I(z;x,y)$. An example based on a Kuramoto-Sivashinsky model [9, 10] illustrates how the assessment of data-assimilation methods method can be evaluated with the various metrics and what this implies for hyper-parameter choices and experimental design.

References

[1] L. Bengtsson, U. Andrae, T. Aspelien,Y. Batrak, J. Calvo, W. de Rooy, E. Gleeson, B. Hansen-Sass, M. Homleid, M. Hortal, K.-I. Ivarsson, G. Lenderink, S. Niemelä, K. P. Nielsen, J. Onvlee, L. Rontu, P. Samuelsson, D. S. Muñoz, A. Subias, S. Tijm, V. Toll, X. Yang et M. Ø. Køltzow, *The HARMONIE–AROME Model Configuration in the ALADIN–HIRLAM NWP System*, Monthly Weather Review **145**, (2017) 1919–1935.

[2] A. Zonato, and N. Theeuwes, *Very-high-resolution simulations with Harmonie-AROME of a heatwave case for the city of Paris with different landuse datasets*, EGU General Assembly 2024 (2024) EGU24-18500.

[3] S. O. Owens, D. Majumdar, C. E. Wilson, P. Bartholomew, and M. van Reeuwijk, *A conservative immersed boundary method for the multi-physics urban large-eddy simulation model uDALES v2.0*, Geosci. Model Dev. **17**, (2024).

[4] A. Patil and C. García-Sánchez, *Quantifying the impact of urban geometric detail for urban air mobility risk forecasting*, Sustainable Cities and Society **132**, (2025) 106750.

[5] C. A. Biltoft, *Concentration Fluctuation Measurements in a Plume Dispersing Through a Regular Array of Obstacles*,Boundary-Layer Meteorology **111**, (2004) 363–415.

[6] Z. T. Xie, and I. P. Castro, *Efficient Generation of Inflow Conditions for Large Eddy Simulation of Street-Scale Flows*, Flow Turbulence Combust **81**, (2008) 449–470.

[7] G. Evensen, F. C. Vossepoel, and P. J. van Leeuwen, *Data Assimilation Fundamentals*, Springer International Publishing (2022), XIX, 245.

[8] T. Gneiting, F. Balabdaoui, and A. E. Raftery, *Probabilistic forecasts, calibration and sharpness*, Journal of the Royal Statistical Society: Series B (Statistical Methodology), **69** (2007) 243–268.

[9] Y. Kuramoto, Y.,*Diffusion-induced chaos in reaction systems*, Prog. Theor. Phys. Suppl., **64** (1978) 346–367.

[10] G. I. Sivashinsky, *Nonlinear analysis of hydrodynamic instability in laminar flames–I. Derivation of basic equations*, Acta Astronaut., **4** (1977) 1177–1206.

Developing trustworthy digital twins by incorporating explainable AI: Detecting learned artefacts for faithful digital twin modeling

Marina Marie-Claire Höhne

Digital Twins are increasingly employed to support decision making in complex systems by continuously integrating observational data with computational models. In precision agriculture, for example, Digital Twins have the potential to optimize irrigation strategies, increase crop yield, reduce pesticide use, improve resource efficiency, and support operational planning. As these systems increasingly influence real-world decisions, it becomes essential to understand how they arrive at their predictions and whether these predictions faithfully reflect the underlying physical processes. Since a Digital Twin consists of several interacting components, including sensor data acquisition, data preprocessing, state estimation, uncertainty quantification, machine learning models, decision support, and continuous feedback mechanisms, artificial intelligence is no longer confined to a single prediction model but can be integrated throughout the entire pipeline. Consequently, ensuring the trustworthiness of a Digital Twin requires the understanding and validation of every component rather than only the final prediction model. A particular challenge arises from the fact that machine learning models can learn spurious correlations during training, also known as Clever Hans artifacts or shortcut learning. As a consequence, models may produce correct predictions for the wrong reasons, relying on artifacts rather than the underlying physical mechanisms they are intended to capture. Especially for Digital Twins, whose purpose is to faithfully represent evolving physical systems, decisions based on unintended correlations fundamentally undermine their reliability.

In this presentation I argue that Explainable Artificial Intelligence (XAI) provides an essential foundation for identifying and analysing such failure modes. Local explanation techniques, such as Layer-wise Relevance Propagation [1], reveal which input features contribute to individual predictions. Furthermore, using Bayesian neural networks, we can compute uncertainty-aware explanations incorporating the predictive uncertainty directly into the explanations themselves [6]. Since numerous explanation methods have been proposed, evaluating explanation quality becomes a major challenge. Quantitative evaluation frameworks, such as Quantus [2], allow explanation methods to be assessed with respect to faithfulness and robustness, whereas MetaQuantus [3] extends this perspective by evaluating the evaluation metrics themselves.

Beyond local explanations, understanding the global behavior of learned models is highly important. Feature visualization enables the analysis of internal representations learned by deep neural networks, while our proposed method DORA [4] automatically identifies learned artifacts directly from the network parameters without relying on external datasets, thereby avoiding additional dataset-induced biases. Furthermore, concept-based explanations further reveal semantically meaningful concepts encoded within learned representations.

After introducing XAI, I demonstrated that both local [5] and global [7] explanations can also be manipulated. Consequently, the central challenge is no longer

explainability itself but faithfulness: to what extent does an explanation faithfully represent the features and computations that lead to a certain prediction, and to what extent does a learned model faithfully represent the underlying physical system?

This perspective naturally extends from individual AI models to whole Digital Twin systems. If multiple interacting AI components contribute to state estimation, prediction, uncertainty quantification, and decision support, then the explainability of a single model is no longer sufficient. Instead, the whole Digital Twin itself requires explanation and validation.

I concluded by highlighting several open mathematical questions arising from this point. How should faithfulness be defined mathematically for entire Digital Twins? Must latent state representations themselves satisfy notions of interpretability? What does explainability mean for a Digital Twin comprising many interacting learning components? Finally, how can explanations of Digital Twins be validated quantitatively and independently of human judgment? Addressing these questions will contribute to establish the mathematical foundations required for Digital Twins, such that internal representations, predictions, and decisions are not only accurate but also understandable, faithful, and verifiable.

References

[1] S. Bach, A. Binder, G. Montavon, F. Klauschen, K.-R. Müller, and W. Samek, *On Pixel-Wise Explanations for Non-Linear Classifier Decisions by Layer-Wise Relevance Propagation*, PLoS ONE **10** (2015), e0130140.

[2] A. Hedström, L. Weber, D. Krakowczyk, D. Bareeva, F. Motzkus, W. Samek, S. Bach, and M. M.-C. Höhne, *Quantus: An Explainable AI Toolkit for Responsible Evaluation of Neural Network Explanations and Beyond*, Journal of Machine Learning Research **24** (2023), Number 34, 1–11.

[3] A. Hedström, P. L. Bommer, K. K. Wickstrøm, W. Samek, S. Lapuschkin, and M. M.-C. Höhne, *The Meta-Evaluation Problem in Explainable AI: Identifying Reliable Estimators with MetaQuantus*, Transactions on Machine Learning Research (TMLR) (2023).

[4] K. Bykov, M. Deb, D. Grinwald, K.-R. Müller, and M. M.-C. Höhne, *DORA: Exploring Outlier Representations in Deep Neural Networks*, Transactions on Machine Learning Research (2023).

[5] A.-K. Dombrowski, M. Alber, C. Anders, M. Ackermann, K.-R. Müller, and P. Kessel, *Explanations Can Be Manipulated and Geometry Is to Blame*, Advances in Neural Information Processing Systems **32** (2019).

[6] K. Bykov, M. M.-C. Höhne, A. Creosteanu, K.-R. Müller, F. Klauschen, S. Nakajima, and M. Kloft, *Explaining Bayesian Neural Networks*, Transactions on Machine Learning Research (2025).

[7] D. Bareeva, M. M.-C. Höhne, A. Warnecke, L. Pirch, K.-R. Müller, K. Rieck, S. Lapuschkin, and K. Bykov, *Manipulating Feature Visualizations with Gradient Slingshots*, Advances in Neural Information Processing Systems **38** (2026).

When AI provides the tools DA practitioners have been craving

Marc Bocquet
(joint work with Tobias S. Finn, Sibo Cheng, Alban Farchi, Charlotte Durand, Yumeng Chen, Ivo Pasmans, and Alberto Carrassi)

Data assimilation (DA) estimates the state of a chaotic system from an imperfect model and noisy, partial observations [1]. The talk surveyed two ways in which artificial intelligence (AI) can supply tools that DA practitioners have long needed: learned probabilistic transition models for general model error, and neural analysis operators that learn how to diagnose uncertainty and non-Gaussian error statistics.

Sequential DA and the IEnKFs

The starting point is the hidden Markov model:

$$\mathbf{x}_{k+1} = \mathcal{M}(\mathbf{x}_k), \tag{1a}$$

$$\mathbf{y}_k = \mathcal{H}_k(\mathbf{x}_k) + \boldsymbol{\xi}_k, \qquad \boldsymbol{\xi}_k \sim \mathcal{N}(\mathbf{0}, \mathbf{R}_k), \tag{1b}$$

where observations arrive sequentially at t_k and forecast errors grow rapidly under chaotic dynamics $\mathcal{M}$. Classical DA exposes a familiar tension: variational methods such as 3D-Var and 4D-Var can handle nonlinear operators and asynchronous observations, but do not directly capture the *errors of the day*; ensemble Kalman filters (EnKFs) represent flow-dependent uncertainty, but are limited by the weak nonlinearity assumption. Ensemble variational methods, especially the iterative ensemble Kalman filter and smoother (IEnKF/IEnKS), combine these strengths by optimising in the ensemble subspace [15, 6].

Focusing on a one time-step DA window $[t_0, t_1]$, the IEnKF minimises

$$\mathcal{L}(\mathbf{x}_0) = \frac{1}{2} \left\| \mathbf{x}_0 - \bar{\mathbf{x}}_0 \right\|^2_{(\mathbf{P}^{\mathrm{a}}_0)^{-1}} + \frac{1}{2} \left\| \mathbf{y}_1 - \mathcal{H}_1 \circ \mathcal{M}(\mathbf{x}_0) \right\|^2_{\mathbf{R}_1^{-1}} . \tag{2}$$

The Lorenz–96 (L96) model with $N_{\mathrm{x}} = 40$ variables provides the low-order testbed used throughout the talk [10]. In L96 experiments, the IEnKF outperforms both a basic 4D–Var and a standard EnKF for moderately long windows, precisely because it combines nonlinear smoothing with ensemble-based errors of the day [5].

Generative dynamics for model error: the IEnKF-G

Model error is unavoidable in realistic geophysical models. A Gaussian additive formulation of model error,

$$\mathbf{x}_{k+1} = \mathcal{M}(\mathbf{x}_k) + \boldsymbol{\eta}_k, \qquad \boldsymbol{\eta}_k \sim \mathcal{N}(\mathbf{0}, \mathbf{Q}_k), \tag{3}$$

leads one to extends the IEnKF into the IEnKF-Q, whose objective over $[t_0, t_1]$ is

$$\mathcal{L}(\mathbf{x}_0) = \frac{1}{2} \left\| \mathbf{x}_0 - \bar{\mathbf{x}}_0 \right\|^2_{(\mathbf{P}^{\mathrm{a}}_0)^{-1}} + \frac{1}{2} \left\| \mathbf{y}_1 - \mathcal{H}_1(\mathbf{x}_1) \right\|^2_{\mathbf{R}_1^{-1}} + \frac{1}{2} \left\| \mathbf{x}_1 - \mathcal{M}(\mathbf{x}_0) \right\|^2_{\mathbf{Q}_1^{-1}} . \tag{4}$$

It improves robustness but retains Gaussian-like assumptions and delicate inflation tuning [13, 14]. We therefore proposed a broader view: learn the Chapman–Kolmogorov transition

$$p(\mathbf{x}_1|\mathbf{x}_0) = \int \mathrm{d}\boldsymbol{\varepsilon}\, \delta\left(\mathbf{x}_1 - \mathcal{M}(\mathbf{x}_0, \boldsymbol{\varepsilon})\right) p_{\boldsymbol{\varepsilon}}(\boldsymbol{\varepsilon}), \qquad \boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}), \tag{5}$$

using flow matching, diffusion, stochastic interpolants, or any related generative models to learn $\mathcal{M}$ conditioned on the Gaussian random variable $\boldsymbol{\varepsilon}$. The point is not to replace DA by a black-box emulator, but to enlarge the mathematical class of sequential DA algorithms while preserving the structure of ensemble and variational methods.

This yields the generative iterative ensemble Kalman filter (IEnKF-G), with the cost function

$$\mathcal{L}\left(\mathbf{x}_0, \boldsymbol{\varepsilon}\right) = \frac{1}{2}\left\|\mathbf{x}_0 - \bar{\mathbf{x}}_0\right\|^2_{\left(\mathbf{P}^{\mathrm{a}}_0\right)^{-1}} + \frac{1}{2}\left\|\mathbf{y}_1 - \mathcal{H}_1 \circ \mathcal{M}(\mathbf{x}_0, \boldsymbol{\varepsilon})\right\|^2_{\mathbf{R}_1^{-1}} + \frac{1}{2}\left\|\boldsymbol{\varepsilon}\right\|^2. \tag{6}$$

The analysis then searches jointly over the initial state and a latent model error variable, while the learned generator supplies a non-Gaussian, state-dependent transition model [3].

The numerical testbed was a two-scale L96-based dynamics, in which $K = 40$ large-scale variables interact with $J = 4$ fine-scale variables per large-scale variable; the large scales alone define a misspecified coarse model [9, 8]. A residual convolutional neural network was trained either deterministically or as a stochastic model using a fair energy score. Inserted into DA, the IEnKF-G with $N_{\mathrm{g}} = 40$ model-error degrees of freedom achieved the best RMSE among the compared methods, outperforming an adaptively inflated EnKF and the more costly IEnKF-Q [12, 3]. Increasing N_{g} from 1 to 40 improved the analysis, indicating that the latent coordinates parametrised by $\boldsymbol{\varepsilon}$ actively resolve missing transition uncertainty.

Learning the analysis itself: DANs

The second part used AI as a discovery tool for DA algorithms. A Data Assimilation Network (DAN) replaces the analysis by a learned incremental operator $\mathbf{a}_{\boldsymbol{\theta}}$ [11, 7, 2, 4]. Given forecast and analysis ensembles $\mathbf{E}^{\mathrm{f}}_k, \mathbf{E}^{\mathrm{a}}_k \in \mathbb{R}^{N_{\mathrm{x}} \times N_{\mathrm{e}}}$, innovations

$$\boldsymbol{\delta}_k \triangleq \mathbf{y}_k - \mathcal{H}_k\left(\bar{\mathbf{x}}^{\mathrm{f}}_k\right), \qquad \bar{\mathbf{x}}^{\mathrm{f}}_k \triangleq \frac{1}{N_{\mathrm{e}}}\sum_{i=1}^{N_{\mathrm{e}}} \mathbf{x}^{\mathrm{f}}_{k,i}, \tag{7}$$

and the linearised observation operator $\mathbf{H}_k$, the update is

$$\mathbf{E}^{\mathrm{a}}_k = \mathbf{E}^{\mathrm{f}}_k + \mathbf{a}_{\boldsymbol{\theta}}\left(\mathbf{E}^{\mathrm{f}}_k, \mathbf{H}_k^{\top}\mathbf{R}_k^{-1}\boldsymbol{\delta}_k\right), \tag{8a}$$

$$\mathbf{E}^{\mathrm{f}}_{k+1} = \mathcal{M}\left(\mathbf{E}^{\mathrm{a}}_k\right). \tag{8b}$$

The parameters $\boldsymbol{\theta}$ are trained over reference trajectories with the cost function

$$\mathcal{L}\left(\boldsymbol{\theta}\right) = \sum_{r=1}^{N_{\mathrm{r}}}\sum_{k=1}^{N_{\mathrm{c}}}\left\|\mathbf{x}^{\mathrm{t}}_{k,r} - \bar{\mathbf{x}}^{\mathrm{a}}_{k,r}\left(\boldsymbol{\theta}\right)\right\|^2, \qquad \bar{\mathbf{x}}^{\mathrm{a}}_{k,r} \triangleq \frac{1}{N_{\mathrm{e}}}\sum_{i=1}^{N_{\mathrm{e}}} \mathbf{x}^{\mathrm{a}}_{k,r,i}. \tag{9}$$

The striking empirical observation is that DAN performance barely depends on N_e, even for $N_e = 1$, while remaining competitive with a well-tuned EnKF. This suggests that $\mathbf{a}_{\boldsymbol{\theta}}$ has internalised part of the analysis error covariance matrix. The evidence is not only scalar RMSE: the leading time-averaged eigenspectra of the covariance diagnosed by DAN and by an EnKF are close, and the learned local patterns scale from $N_x = 40$ to larger L96 grids without retraining [2, 4]. Under stronger nonlinear conditions, obtained by increasing the time step between observations, DAN variants remain competitive with the IEnKF baseline, while being more robust and extending to larger time steps, suggesting that $\mathbf{a}_{\boldsymbol{\theta}}$ learns nonlinear sensitivities compatible with an effectively non-Gaussian prior.

Conclusion

The talk argued for a synthesis of DA and AI. Generative AI gives sequential DA a practical way to represent general, non-Gaussian, state-dependent Chapman–Kolmogorov transitions, leading to IEnKF-G as a principled extension of ensemble variational DA. Conversely, DANs show that neural networks can rediscover and illuminate DA mechanisms: covariance diagnosis, locality, inflation-like behaviour, and nonlinear prior sensitivity. AI thus provides tools DA practitioners have been craving: learned probabilistic structure where models are wrong, and learned analysis maps that reveal where classical schemes can be generalised.

References

[1] M. Asch, M. Bocquet, and M. Nodet. *Data Assimilation: Methods, Algorithms, and Applications*. Fundamentals of Algorithms. SIAM, Philadelphia, 2016.

[2] M. Bocquet, A. Farchi, T. S. Finn, C. Durand, S. Cheng, Y. Chen, I. Pasmans, and A. Carrassi. Accurate deep learning-based filtering for chaotic dynamics by identifying instabilities without an ensemble. *Chaos*, 34:091104, 2024.

[3] M. Bocquet, T. S. Finn, and S. Cheng. Leveraging learned generative dynamics in ensemble variational data assimilation. 2026. in preparation.

[4] M. Bocquet, T. S. Finn, S. Cheng, and A. Farchi. Elucidating the performance of data assimilation neural networks for chaotic dynamics. *EGUsphere*, 2026:1–34, 2026.

[5] M. Bocquet and P. Sakov. Joint state and parameter estimation with an iterative ensemble Kalman smoother. *Nonlin. Processes Geophys.*, 20:803–818, 2013.

[6] M. Bocquet and P. Sakov. An iterative ensemble Kalman smoother. *Q. J. R. Meteorol. Soc.*, 140:1521–1535, 2014.

[7] P. Boudier, A. Fillion, S. Gratton, S. Gürol, and S. Zhang. Data assimilation networks. *J. Adv. Model. Earth Syst.*, 15:e2022MS003353, 2023.

[8] I. Grooms. Analog ensemble data assimilation and a method for constructing analogs with variational autoencoders. *Q. J. R. Meteorol. Soc.*, 147:139–149, 2021.

[9] I. Grooms and Y. Lee. A framework for variational data assimilation with superparameterization. *Nonlin. Processes Geophys.*, 22:601–611, 2015.

[10] E. N. Lorenz and K. A. Emanuel. Optimal sites for supplementary weather observations: simulation with a small model. *J. Atmos. Sci.*, 55:399–414, 1998.

[11] M. McCabe and J. Brown. Learning to assimilate in chaotic dynamical systems. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 12237–12250. Curran Associates, Inc., 2021.

[12] P. N. Raanes, M. Bocquet, and A. Carrassi. Adaptive covariance inflation in the ensemble Kalman filter by Gaussian scale mixtures. *Q. J. R. Meteorol. Soc.*, 145:53–75, 2019.

[13] P. Sakov and M. Bocquet. Asynchronous data assimilation with the EnKF in presence of additive model error. *Tellus A*, 70:1414545, 2018.
[14] P. Sakov, J.-M. Haussaire, and M. Bocquet. An iterative ensemble Kalman filter in presence of additive model error. *Q. J. R. Meteorol. Soc.*, 144:1297–1309, 2018.
[15] P. Sakov, D. S. Oliver, and L. Bertino. An iterative EnKF for strongly nonlinear systems. *Mon. Wea. Rev.*, 140:1988–2004, 2012.

Enforcing Dirichlet boundary conditions in operator learning

Margaret Trautner
(joint work with Andrew M. Stuart)

Interest in, and the use of, neural network-based models for problems in science and engineering has exploded in the past decade. For many of these problems, the underlying phenomena are governed by partial differential equations (PDEs), which map input functions to output functions via an infinite-dimensional operator. *Operator learning* encompasses a branch of machine-learning models that generalize to this infinite-dimensional function space setting; the models themselves are termed *neural operators*.

Operator learning has proved empirically useful in two main cases. First, when the underlying equations of the system are known, operator learning can accelerate scientific computing by serving as a surrogate model for computationally expensive numerical simulations. The second case is where a model is learned from data alone, as the governing equations are unknown. However, in many settings, the reality falls in between these two cases; some information is known about the system, such as a boundary condition or conservation law, but the full descriptive equations remain unknown. In these cases, it is natural to want to enforce the known information without impeding the expressibility of the architecture or the ability to learn an approximator through stochastic optimization. It is this setting which largely motivates the present work: we *enforce boundary conditions* for neural operators that hold *even without training* while retaining the expressibility of the architecture. Furthermore, we develop the method in a manner that can act on *arbitrary mesh data* and thus on *general geometries* in any dimension.

An attractive feature of many operator learning models is that they satisfy the *universal approximation* property: for any desired accuracy $\epsilon > 0$, there exists a parameterization of the model architecture in question that achieves ϵ accuracy for a map of interest over a certain input set. In other words, the model class is *expressive* enough to capture the map of interest. Such results have been established for the architectures DeepONet [1] and the Fourier Neural Operator (FNO) [2], as well as for a broad class of operator learning models that contain both nonlinear and nonlocal operations [3]. In this latter work in particular, it was proved that for a kernel-integral neural operator, taking a single hidden layer that performs an average over the domain, with the remaining layers being pointwise neural network operations, is sufficient for universality. In general, it is a challenge to maintain universal approximation while enforcing known physical laws or constraints explicitly.

We make the following contributions to the state-of-the-art, in relation to enforcing Dirichlet boundary conditions within operator learning, as well as achieving universal approximation in a general setting:

(1) We propose an operator learning architecture that automatically satisfies Dirichlet boundary conditions by using the Dirichlet eigenfunctions of the Laplacian as a basis.
(2) We prove universal approximation for a general class of nonlocal neural operators, subsuming the proposed Dirichlet eigenfunction based architecture as a special case; these approximation results may be of independent interest beyond the application to enforcing Dirichlet boundary conditions. This general class includes neural operators using only a single hidden layer and single basis function as well as neural operators that perform approximation in a particular basis of interest. Sufficient conditions on such basis functions are provided.
(3) We implement the proposed method to approximate two representative operators defined by coefficient-to-solution maps arising in PDE problems: the Darcy flow equation on a square domain and the Helmholtz equation on a circular domain. We demonstrate the advantage of the proposed method over FNO as well as over two possible alternative operator learning approaches that also automatically satisfy the boundary condition.

Specifically, we introduce a new, kernel-integral based, operator learning architecture in which the output space of each layer builds an approximation in the basis of Dirichlet eigenfunctions of the Laplacian in the output domain. These eigenfunctions satisfy a zero boundary condition and thus, coupled with a zero-preserving activation function and an appropriate fixed bias, create an architecture that automatically satisfies the Dirichlet boundary condition of a PDE. In fact, replacing the final hidden layer of any kernel-integral neural operator, including FNO, with this Laplacian eigenfunction layer achieves the desired boundary condition without training. In addition, as a departure from previous work that automatically satisfies the boundary condition, this approach is applicable to complex geometries and domains; we assume only that the output domain is bounded in with Lipschitz boundary. Furthermore, the method requires no restriction on choice of discretization or grid; it may be used on functions defined on arbitrary meshes. Beyond the choice of the eigenfunctions of the Laplacian as the basis elements, we prove approximation theory for a much broader class of potential basis functions. The approach we adopt suggests practical extensions of the method to other classical basis choices, including various finite element bases. Numerical results demonstrate the practicality and competitiveness of the proposed approach.

References

[1] L. Lu, P. Jin, G. Pang, Z. Zhang, and G. Karniadakis, *Learning nonlinear operators via DeepONet based on the universal approximation theorem of operators*, Nature Machine Intelligence **3** (2021) 218–229.

[2] N. Kovachki, S. Lanthaler, and S. Mishra, *On universal approximation and error bounds for Fourier neural operators*, Journal of Machine Learning Research **22** (2021) 1–76.

[3] S. Lanthaler, Z. Li, and A. Stuart, *Nonlocality and nonlinearity implies universality in operator learning*, Constructive Approximation **62** (2025) 1–43.

A digital twin framework for output-feedback control of uncertain linear systems

Philipp A. Guth

(joint work with Karl Kunisch, Sérgio S. Rodrigues, Jesper Schröder)

A digital twin framework for output-feedback stabilization and parameter identification in uncertain dynamical systems is presented. A virtual model evolves in parallel with the physical process, assimilating measurement data in real time. By design, the twin reconstructs the system state and a generates stabilizing feedback, while model parameters are simultaneously inferred from data of the controlled dynamics using a Bayesian approach.

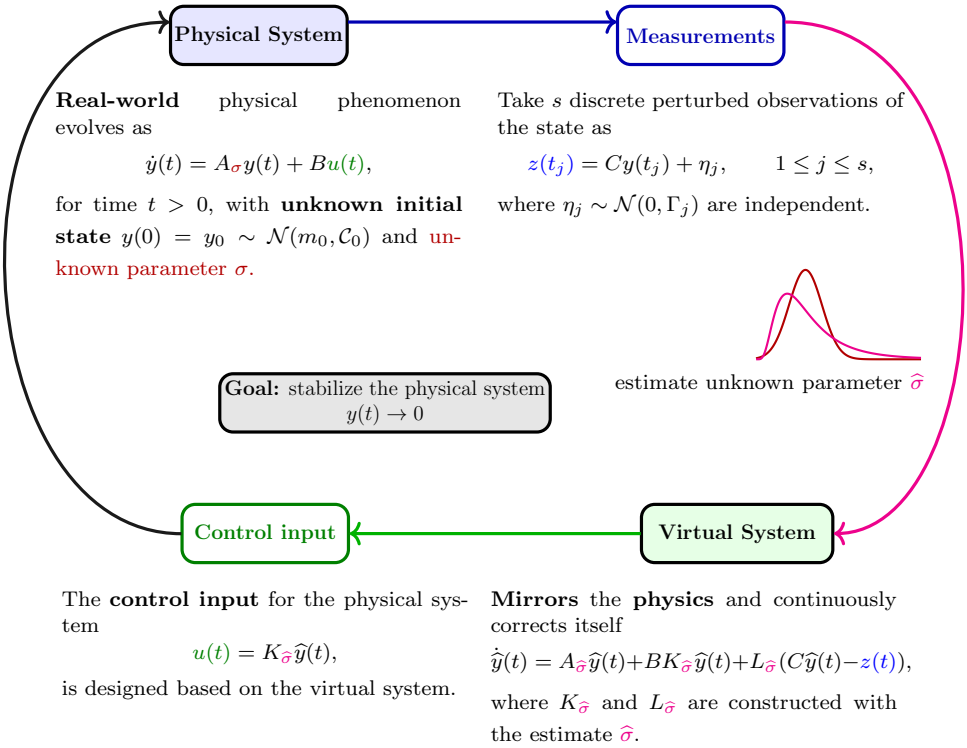

Figure 1. Illustration of the digital twin framework

Designing the virtual twin entails addressing several interrelated challenges: (i) computing a parameter estimate $\widehat{\sigma}$ based on the output measurement data collected from the physical system involves solving a sequence of inverse problems,

which are typically ill-posed and sensitive to noise; (ii) constructing the state-feedback operator $K_{\widehat{\sigma}}$, with the goal of shaping the system's closed-loop behavior in accordance with a performance criterion, which in this work is stabilization. In the presence of uncertainty in the parameter estimate $\widehat{\sigma}$, this becomes a problem of robust or risk-aware control; (iii) designing the observer gain $L_{\widehat{\sigma}}$ to ensure stable and accurate reconstruction of the system state from noisy and possibly partial measurements. This involves balancing responsiveness to new data with robustness to measurement noise and plays a crucial role in ensuring that the virtual state $\widehat{y}$ remains synchronized with the true physical state y.

While each of these tasks has been studied extensively in isolation, a mathematical framework that combines them into a single closed-loop system remains largely unexplored. In particular, parameter estimation modifies the virtual model on which the controller and observer design are based, while the applied control itself influences the data available for subsequent parameter estimation. This creates a strongly coupled interaction between inference and control. The main contribution of this work is the formulation and rigorous analysis of a mathematical framework (illustrated in Figure 1) that jointly addresses the interacting challenges identified above, with the goal of stabilizing the physical system.

Under standard stabilizability and detectability assumptions, we establish stability of the coupled physical-virtual dynamics whenever the parameter estimate belongs to a sufficiently small neighborhood $\mathcal{I}$ of the true parameter. Consequently, the virtual state tracks the evolution of the physical state and the physical system is asymptotically stabilized by the feedback generated by the virtual twin. In this regime, Riccati-based feedback and observer gains associated with the parameter estimate ensure the desired stabilization and tracking properties. In particular, for every $\widehat{\sigma} \in \mathcal{I}$, the distribution of the coupled physical-virtual state converges to a limit distribution, which is characterized explicitly in terms of its moments. Our analysis highlights a fundamental mathematical feature of digital twins: parameter estimation, state reconstruction, and control cannot generally be treated as independent tasks. Instead, they form a dynamically coupled system in which the quality of inference affects control performance, and the chosen control strategy determines the information available for future inference.

Several important directions for future research remain open. A central challenge is to establish conditions under which the evolving parameter estimate enters and remains in the admissible neighborhood $\mathcal{I}$ of the true parameter, thereby connecting the local stability theory with the sequential Bayesian update procedure. Furthermore, adaptive strategies for selecting when model updates should be performed remain to be developed. Our numerical experiments suggest that these issues are critical for the proposed digital twin framework.

System-theoretic reduced-order digital twin prototypes

Boris Krämer
(joint work with Nick Corbin)

1. Digital Twins Prototypes via System-theoretic Model Reduction

Digital twins are collections of models, of varying fidelity, that are continuously synchronized with a physical asset. This positions a digital twin uniquely for control, decision making and scenario planning. There are several stages to developing a digital twin, as there are to developing a physical product. In the prototyping phase, a model class, its parameters, attributes and desirable properties are chosen, and an initial prototype constructed. In the next stage, the twin goes live: as the physical twin is in deployed and produces data, the digital twins continues to get refined, recalibrated, and synchronized to the physical asset, keeping the sim-to-real gap as close as possible.

If a digital twin is to be used for control, it should be efficient to evaluate to allow for control action and decision making at the relevant process timescales. Moreover, we choose a digital twin prototype model class as one that 1) is in state-space form, to enable advanced state- and output feedback control, 2) has suitable system-theoretic properties, such as retaining the most observable and most controllable modes, 3) retains stability properties of the physical twin or a higher-fidelity model thereof, 4) captures nonlinear transient responses of the physical twin, 5) is synchronized with data from the physical twin. To satisfy these requirements, we choose the framework of model reduction via nonlinear balanced truncation, e.g., we start from a high-fidelity model of the physical twin and reduce it by taking into account system-theoretic properties.

Balanced truncation (BT) model reduction is a system-theoretic framework to construct reduced-order models (ROMs) of high-dimensional systems by retaining important structures. It has been very successfully used for open-loop forced systems, both for linear time invariant systems, and more recently for weakly nonlinear systems. The method first balances the system, in that it simultaneously transforms the controllability and observability energies to diagonal form while ordering the transformed states according to their input and output energy. Second, BT then truncates the "hard to reach" and "hard to observe" states to get a reduced-order model that retains the original models' controllability, observability, and stability properties. The method was first proposed for LTI systems by Moore [1], and has been made computationally feasible for large-scale systems by important developments for solving linear Lyapunov matrix equations in the 1990s and 2000s. A comprehensive survey of its variants and extensions for LTI systems are discussed in [2]. For nonlinear systems, the situation is different. The general concept was introduced by Scherpen's seminal work [3] for the open-loop case, and [4] for the closed-loop case. However, to be able to put this theory to use in practice, many computational challenges have to be overcome, yet slow progress

over the past 30 years hindered a broader adoption. Herein, we discuss our recent *computational* developments in balanced truncation for nonlinear systems.

2. Balancing for nonlinear systems

Consider a nonlinear control-affine system starting from initial state $\mathbf{x}(0) = \mathbf{x}_0$:

$$\dot{\mathbf{x}}(t) = \mathbf{f}(\mathbf{x}) + \mathbf{g}(\mathbf{x})\mathbf{u}(t), \qquad \mathbf{y}(t) = \mathbf{h}(\mathbf{x}),$$

The (open-loop) controllability and observability energy for stable systems are defined as

$$\mathcal{E}_c(\mathbf{x}_0) := \min_{\substack{\mathbf{u}\in L_2(-\infty,0] \\ \mathbf{x}(-\infty)=\mathbf{0},\ \mathbf{x}(0)=\mathbf{x}_0}} \frac{1}{2}\int_{-\infty}^{0} ||\mathbf{u}(t)||^2 \mathrm{d}t,$$

$$\mathcal{E}_o(\mathbf{x}_0) := \frac{1}{2}\int_{0}^{\infty} ||\mathbf{y}(t)||^2 \mathrm{d}t, \qquad \mathbf{u}(t) \equiv \mathbf{0}.$$

These energy functions (as well as their counterparts for the unstable case [4]) can be shown to obey high-dimensional Hamilton-Jacobi-Bellman equations [3, 4]. Under certain conditions on a neighborhood $\mathcal{W}$ of zero, there exists a coordinate transformation $\mathbf{x} = \Phi(\mathbf{z})$ such that the energies in the new coordinate $\mathbf{z}$ are

$$\mathcal{E}_c(\mathbf{x}) := \mathcal{E}_c(\Phi(\mathbf{z})) = \frac{1}{2}\sum_{i=1}^{n} \frac{z_i^2}{\sigma_i(z_i)}, \qquad \mathcal{E}_o(\mathbf{x}) := \mathcal{E}_o(\Phi(\mathbf{z})) = \frac{1}{2}\sum_{i=1}^{n} \sigma_i^2(z_i) z_i^2 \tag{1}$$

where $\sigma_1(\mathbf{z}) \geq ... \geq \sigma_m(\mathbf{z})$ are smooth singular value functions. The nonlinear balancing transformation $\Phi(\mathbf{z})$ creates balanced model states that are "hard-to-reach" and "hard-to-observe". Those can be easily removed in the ROM construction.

3. Computational approaches to nonlinear balanced truncation

While the theoretical foundation of nonlinear balancing has been worked out over 30 years ago, its application to moderate to high-dimensional systems has so far been hampered by the need for (i) solving n-dimensional HJB PDEs, (ii) efficiently and uniquely computing a balancing transformation Φ; (iii) constructing an efficient ROM. We present several approaches to achieve computational feasibility for nonlinear balanced truncation into the 1000s of states.

3.1. **Solving the associated HJB equations.** Based on work by Al'brekht in 1961, in [5, 6] we propose—for control-affine dynamical systems with polynomial drift and input functions—a scalable polynomial ansatz for computing energy function approximations of the form

$$\mathcal{E}(\mathbf{x}) = \frac{1}{2}\left(\mathbf{x}^\top \mathbf{V}_2 \mathbf{x} + \mathbf{v}_3^\top \mathbf{x} \otimes \mathbf{x} \otimes \mathbf{x} + \ldots\right) \tag{2}$$

where $\otimes$ denotes the Kronecker product of two vectors and $\mathbf{v}_k \in \mathbb{R}^{n^k}$ are the Taylor series coefficients in vectorized form. We derive algorithms to efficiently solve block-structured equations in n^k unknowns via $\mathcal{L}(\mathbf{M})\mathbf{v}_k = \mathbf{b}$. Here $\mathcal{L}(\cdot)$ denotes the k-way Kronecker sum and $\mathbf{M}$ is a stable (or stabilized) matrix. Computational

examples for semi-discretized PDEs illustrate favorable scalability of our approach up to billions of unknowns in the right-hand side $\mathbf{b}$.

3.2. **Balancing the nonlinear energy functions.** The computation of the nonlinear balancing transformation $\Phi(\mathbf{z})$ in equation (1) is nontrivial due to the difficulty of both achieving certain values (Hankel Singular Value Functions) along the diagonal of the balanced system, as well as the requirement to set off-diagonal entries to zero. While work in the 2000s by Fujimoto [10] and Krener [9] proposed to approximate the nonlinear coordinate transformation as a Taylor series. Similar to how the polynomial coefficients of the energy function can be computed degree-by-degree using Al'brekht's method, they show that the coordinate transformation coefficients can also be computed degree-by-degree. However, the method is merely outlined abstractly; conditions for the existence of the solutions to the coefficients, details regarding the practical computation of the transformation coefficients, and a scalable implementation as needed for model reduction are not provided.

Our work in [7] presents scalable approaches and concrete algorithms to compute the nonlinear transformation

$$\mathbf{x} = \Phi(\mathbf{z}) = \mathbf{T}_1\mathbf{z} + \mathbf{T}_2\mathbf{z} \otimes \mathbf{z} + \dots \tag{3}$$

Therein, we derive two sets of structured linear systems to obtain the transformation's coefficients degree-by-degree, followed by a symmetrization to achieve unique solutions. We also highlight numerical challenges that continue to exist, such as clustering of Hankel singular values which leads to ill-conditioned solutions.

3.3. **Balanced reduced-order model construction.** The third step of nonlinear balanced truncation model reduction is to define a nonlinear ROM

$$\dot{\mathbf{z}}_r(t) = \mathbf{f}_r(\mathbf{z}_r) + \mathbf{g}_r(\mathbf{z}_r)\mathbf{u}(t), \qquad \mathbf{y}_r(t) = \mathbf{h}_r(\mathbf{z}_r),$$

such that the input-to-output map of the ROM approximates the input-to-output map of the FOM well. Moreover, preservation of controllability, observability and stability properties are important features of nonlinear balanced truncation model reduction. In [11] we proposed a nonlinear ROM that achieves input-normal form, but only approximates the output-normal form. In [8] we addressed this issue and derived direct series approximations for the transformations with explicit coefficient computations. This form renders the controllability energy "input-normal" and the output energy approximately "output diagonal", and in a second step guarantees full balancing.

4. Open Problems

While the theory of nonlinear balanced truncation model reduction has been available since the 1990s, scalable methods that can reduce systems of order 100-1000 have only very recently become available, as outlined above. Nevertheless, there remain several open problems for nonlinear balanced truncation. First, it is clear that for complex systems, algorithms are needed that scale to the 1,000,000s of states, so more work needs to be done, mainly on memory efficient and compressed

representations for the energy functions, transformations, and reduced model; this may require considering more numerically advantageous basis functions and approximations than polynomials. Second, there still is not a clear understanding in the community when full nonlinear balanced truncation (for nonlinear models approximated on nonlinear balanced manifolds) significantly outperforms linear projection onto a balanced subspace. A study of regions of stabilizability/stability and accuracy could reveal such recommendations. Third, the algorithms proposed in our work represent a balance-then-truncate approach that, similar to the case in linear balancing, is ill-conditioned in the case of a minimal or nearly minimal realization featuring zero or near-zero Hankel singular values. This presents a major limitation, as zero and near-zero singular values are what indicate reducibility. In our testing, we found that model parameters and, in particular, control and measurement locations had to be artificially chosen in order to ensure that singular values were nonzero and distinct in order to avoid numerical issues. The development of a balance-and-reduce approach akin to square-root balancing in the linear case could enable nonlinear balancing of non-minimal systems in a way that alleviates this numerical ill-conditioning. Similar to the linear case, this could also open the door to a low-rank adaptation to enable scalability to very high dimensions.

As it pertains to actual digital twin development, the nonlinear balanced truncation model reduction is, in our opinion, a good prototype method. However, the model has to be infused with data from the physical twin and continuously optimized. This presents a major hurdle, as no data-driven techniques for nonlinear balancing are available at present. Work in this direction would have tremendous value for certifiable digital twins.

References

[1] B. Moore, *Principal component analysis in linear systems: Controllability, observability, and model reduction*, IEEE Transactions on Automatic Control **26(1)** (1981), 17–32.

[2] S. Gugercin and A. Antoulas, *A survey of model reduction by balanced truncation and some new results*, International Journal of Control **77(8)** (2004), 748–766.

[3] J.M.A. Scherpen, *Balancing for nonlinear systems*, Systems & Control Letters **21(2)** (1993), 143–153.

[4] J.M.A. Scherpen, *$\mathcal{H}_\infty$ balancing for nonlinear systems*, International Journal of Robust and Nonlinear Control **6(7)** (1996), 645–668.

[5] B. Kramer, S. Gugercin, L. Balicki, J. Borggaard, *Scalable computation of energy functions for nonlinear balanced truncation*, Computer Methods for Applied Mechanics and Engineering, 427, (2024), 117011.

[6] N. Corbin and B. Kramer, *Scalable computation of $\mathcal{H}_\infty$ energy functions for polynomial control-affine systems* IEEE Transactions of Automatic Control 70(5), (2025), 3088–3100.

[7] N. Corbin, A. Sarkar, J.M.A. Scherpen, B. Kramer, *Scalable computation of input-normal/output-diagonal balanced realization for control-affine polynomial systems*, Systems & Control Letters, 204 (2025), 106178,.

[8] N. Corbin, B. Kramer, *Nonlinear balanced truncation model reduction through scalable Taylor series* arXiv: 2604.23044

[9] A.J. Krener, *Reduced order modeling of nonlinear control systems*, in: Analysis and Design of Nonlinear Control Systems. Springer Berlin Heidelberg, (2008), 41–62.

[10] K. Fujimoto, D. Tsubakino, D., *On computation of nonlinear balanced realization and model reduction*, in: 2006 American Control Conference, IEEE, 1655399.
[11] B. Kramer, S. Gugercin, J. Borggaard, *Nonlinear Balanced Truncation: Part 2-Model Reduction on Manifolds* arXiv:2302.02036

Contrastive learning: A mathematical perspective

Ricardo Baptista
(joint work with Andrew M. Stuart, Son Tran)

Multimodal contrastive learning is a methodology for linking data across different modalities—for instance, text and images in data science, or Lagrangian and Eulerian observations in data assimilation. The standard formulation identifies a set of modality-specific encoders that map inputs into a shared latent space, where representations are aligned. In this work, we present a probabilistic view that interprets contrastive learning as defining conditional probability distributions for each modality conditioned on the other, in a manner consistent with the joint data distribution.

Let $u \in \mathcal{U}$ and $v \in \mathcal{V}$ represent data from two different modalities with a joint probability distribution $\mu \in \mathcal{P}(\mathcal{U} \times \mathcal{V})$. Bimodal contrastive learning seeks the following encoders that map data to a common d_e-dimensional latent space:

$$g_u \colon \mathcal{U} \times \Theta \to \mathbb{S}^{d_e - 1}$$
$$g_v \colon \mathcal{V} \times \Theta \to \mathbb{S}^{d_e - 1},$$

where Θ denotes a space of parameters. We learn the parameters of the encoders given N i.i.d. pairs $\{(u^{(i)}, v^{(i)})\}_{i=1}^N \sim \mu$. A prototypical instantiation of bimodal contrastive learning is the Contrastive Language-Image Pre-training (CLIP) framework [1] that learns encoders by maximizing the objective:

$$\begin{aligned} \mathsf{L}_{\text{clip}}^N(\theta) = \frac{1}{N} \sum_{i=1}^N \langle g_u(u^{(i)}; \theta_u),\, g_v(v^{(i)}; \theta_v) \rangle / \tau \\ - \frac{1}{2N} \sum_{i=1}^N \log \sum_{j=1}^N e^{\langle g_u(u^{(i)}; \theta_u),\, g_v(v^{(j)}; \theta_v) \rangle / \tau} \\ - \frac{1}{2N} \sum_{j=1}^N \log \sum_{i=1}^N e^{\langle g_u(u^{(i)}; \theta_u),\, g_v(v^{(j)}; \theta_v) \rangle / \tau}, \end{aligned} \tag{1}$$

with respect to the parameters $\theta = (\theta_u, \theta_v, \tau)$, where $\tau > 0$ is a temperature parameter. Maximizing $\mathsf{L}_{\text{clip}}^N$ aligns paired data while penalizing unpaired data.

At the population level, we show that CLIP defines an approximate joint distribution over (u, v). Let μ_u and μ_v be the marginal distributions of μ for u and v, respectively, and let $\mu_{u|v}$ and $\mu_{v|u}$ be the conditional distributions for each modality conditioned on the other. A model for the joint distribution of (u, v) is:

$$\nu(\mathrm{d}u, \mathrm{d}v; \theta) \;\propto\; \exp\Big(\langle g_u(u;\theta_u),\, g_v(v;\theta_v)\rangle/\tau\Big)\, \mu_u(\mathrm{d}u)\, \mu_v(\mathrm{d}v), \tag{2}$$

with corresponding conditionals $\nu_{u|v}$ and $\nu_{v|u}$. Then, one can show that maximizing a rescaled empirical objective in (1) as $N \to \infty$ is equivalent to minimizing the sum of expected Kullback-Leibler (KL) divergences between the true and approximate conditionals:

$$\begin{aligned}\mathsf{J}(\theta) = \tfrac{1}{2}\,\mathbb{E}_{v\sim\mu_v}\big[\mathsf{D}_{\mathrm{KL}}(\mu_{u|v}(\cdot \mid v) \,\|\, \nu_{u|v}(\cdot \mid v;\theta))\big] \\ + \tfrac{1}{2}\,\mathbb{E}_{u\sim\mu_u}\big[\mathsf{D}_{\mathrm{KL}}(\mu_{v|u}(\cdot \mid u) \,\|\, \nu_{v|u}(\cdot \mid u;\theta))\big].\end{aligned}$$

This probabilistic view suggests extensions of classical contrastive learning based on other loss functions. For example, one can instead seek the encoders by minimizing the KL divergence between the joint distributions μ and ν, or by minimizing a one-sided loss for the expected error of only one conditional distribution.

We analyze the various losses in a multivariate Gaussian setting for $u \in \mathbb{R}^{d_u}$ and $v \in \mathbb{R}^{d_v}$. In particular, we derive closed-form solutions for the minimizers of each loss assuming linear encoders of the form $g_u(u) = Gu$ and $g_v(v) = Hv$ for some matrices $G \in \mathbb{R}^{d_e \times d_u}$ and $H \in \mathbb{R}^{d_e \times d_v}$. In this case, the joint distribution (2) depends only on $A = G^\top H \in \mathbb{R}^{d_u \times d_v}$. For the Gaussian data distribution $\mu = \mathcal{N}(0, C)$ with block covariance $C = \left(\begin{smallmatrix} C_{uu} & C_{uv} \\ C_{vu} & C_{vv} \end{smallmatrix}\right)$, we show that the minimizer of J over A when $d_e = \min(d_u, d_v)$ is $A^\star = C_{uu}^{-1} C_{uv} C_{vv}^{-1}$, which recovers the conditional means of $\mu_{u|v}$ and $\mu_{v|u}$ exactly, while overestimating their conditional covariances. On the other hand, minimizing the one-sided conditional loss recovers both the conditional mean *and* covariance of $\mu_{u|v}$ with the same encoders, while minimizing the joint loss yields a closer approximation to the marginal covariances of μ.

Numerically, we validate the results in two settings. First, we consider the MNIST dataset, consisting of images and labels as the two modalities. We show that models trained with the two-sided loss yield greater uncertainty in digit classification and image generation conditioned on a label. Second, we consider the Lagrangian data assimilation problem, first introduced in [2], of identifying a velocity field from time-series trajectories of passive tracer positions, which can model the recovery of ocean currents from the observed locations of sea ice floes. We train a CLIP model on paired velocity and tracer trajectory data to demonstrate how contrastive learning can be used to solve inverse problems without explicit knowledge of a forward model. We then use the model to retrieve the most likely velocity field corresponding to a given trajectory, or to sample velocities from its conditional distribution.

Open directions include: (i) extending the framework to more than two modalities or multiple samples of each modality; (ii) incorporating physical constraints in the probabilistic description; and (iii) further applications in digital twins.

References

[1] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021.

[2] L. Kuznetsov, K. Ide, and C. K. Jones. A method for assimilation of Lagrangian data. *Monthly Weather Review*, 131(10):2247–2260, 2003.
[3] R. Baptista, A. M. Stuart, and S. Tran. A mathematical perspective on contrastive learning. *arXiv:2505.24134*, 2025.

Localized Schrödinger bridge sampler

Georg A. Gottwald

We describe a plug & play Langevin sampler based on Schrödinger bridges. The general problem we address is how to sample from an unknown distribution to which we only have access to via a finite number of training samples. This setting is known as generative modelling. Generative models consist typically of two steps; a noising step in which samples are subjected to random noise and a denoising step, projecting the noised sample back close to the data manifold. In this talk, we propose a generative model combining Schrödinger bridges and Langevin dynamics. Rather than estimating the generator, or the score function, as is routinely done in diffusion models, we instead determine the semi-group mapping directly from the noised sample to the data manifold. In particular, we employ diffusion maps and Schrödinger bridges over an appropriate reversible reference process to approximate the conditional transition probability from the available training samples. The projection onto the manifold turns out to be stable in the sense that the new generated samples are contained in the convex hull of the data. This is contrasted to score-based problems solving stiff stochastic differential equations. Our framework can be naturally extended to generate conditional samples and to Bayesian inference problems with a given likelihood function. A connection with Tweedie's formula establishes the connection with classical diffusion models and denoising methods.

We show that our sampler has a gradient structure and samples from a modified probability density function. Further we highlight a connection to transformers in used in LLMs.

A key bottleneck of Schrödinger bridge samplers is the exponential dependence of the required training samples on the dimension d of the ambient state space. We propose a localization strategy which exploits conditional independence of conditional expectation values. Whereas in the unlocalized setting a single high-dimensional Schrödinger bridge problem is solved for, in the localized setting we solve d low-dimensional Schrödinger bridge problems. We show how detailed balance of an underlying reversible overdamped Langevin process naturally gives rise to our plug & play Langevin sampler architecture. As for the original Schrödinger bridge sampler, the localized sampler is stable and geometric ergodic. We also extend the idea of localization to plug & play Langevin samplers using kernel-based denoising in combination with Tweedie's formula.

This is joint work with Sebastian Reich and is based on [1, 2].

References

[1] G.A. Gottwald, F. Li, Y. Marzouk and S. Reich, *Stable generative modelling using Schrödinger bridges*, Philosophical Transactions of the Royal Society A **383** (2025) 20240332.

[2] G.A. Gottwald and S. Reich, *Localized Schrödinger bridge sampler*, Journal of Computational Physics **548** (2026) 114583.

Digital twins for effective thermal management

Dirk Hartmann

Digital twins act as full computational substitutes for a physical system, used to understand, predict, and optimize its real-world performance through a bidirectional exchange of data and models. The concept is not new: it has evolved steadily since the mid-1980s and is now converging with physical AI and the industrial metaverse [7]. Across industries, five use cases recur: virtual testing and commissioning, for example operator training and automation testing; virtual sensing, such as estimating temperatures inside rotors or pressures in turbines; diagnosis and identification, including predictive maintenance and leakage detection; performance prediction, for example the remaining useful lifetime of motors; and performance optimization, such as model predictive control of reactors. Model order reduction is a recurring enabler across all of these settings [4, 1].

Despite this breadth, the hype around digital twins still outweighs documented real-world success [9]. A survey of practicing engineers ranks the dominant barriers as lack of knowledge (48%), time pressure (45%), cost (35%), insufficient accuracy (27%), computing limits (22%), and lack of automation tools (15%) [3]. Open questions on real-time capability, certification, usability, and model management remain unresolved, and their resolution is as much mathematical as it is technological [7].

Example use case: battery thermal management. Industrial battery models are large, with 1 to 100 million cells and terabytes of associated data. The core difficulty is not the physics itself but building models fast enough for real-time control and virtual sensing. Reduced order modeling (ROM) is the key enabler: proper orthogonal decomposition compresses the high-dimensional thermal state into a handful of dominant modes, turning a terabyte-scale field problem into a low-dimensional system amenable to real-time evaluation [4, 1, 2]. However, despite years of successful research the corresponding methods have not made it to standard industrial research and development workflows. Machine Learning (ML) and Artificial Intelligence are considered to enable a much more effective use of corresponding technologies [10].

The role of ML: operator inference. Given the reduced state, the reduced dynamics must still be identified. Operator inference does this as a machine learning approach: it learns the reduced operators directly from simulation or measurement trajectories by solving a regression problem, rather than projecting the governing equations intrusively [8]. Unlike generic neural surrogates, the learned model retains a physically interpretable polynomial structure, which keeps it explainable

and amenable to analysis. Autoregressive formulations and roll-out training make the procedure robust when data are scarce or noisy [11, 5], and active learning steers data acquisition toward the most informative operating points. For the battery, this yields a real-time capable model built from only a few hours of simulation data. Notably, the model predicts complex drive cycles even when trained solely on simple step functions, indicating genuine generalization rather than interpolation [6].

The role of AI: agentic automation. Building such a model still demands expert effort per product variant. Agentic AI addresses this outer loop as it stands today: an autonomous program built on a large language model, equipped with memory, planning, and tool access, that drives the simulation and model-building tools rather than replacing their numerics. The goal is to automate the full pipeline, from design intent to validated results, in under 60 seconds [10]. Here AI is an orchestrator, not a solver; it composes meshing, simulation, reduction, and operator inference into an end-to-end workflow. Turning this into a dependable capability again hinges on the mathematical questions below, since automation without certified accuracy only accelerates the production of unverified results.

Open mathematical challenges. The battery use case exposes a set of foundational problems that generalize well beyond thermal management.

- *More complex dynamics.* Spray cooling, phase change, and convective flows break the low-dimensional linear-subspace assumption underlying classical POD. Nonlinear and parametric reduction on manifolds, and structure-preserving learned representations, are needed to capture such regimes robustly.
- *Certifying model accuracy.* Deployment for control and virtual sensing requires rigorous, computable error bounds for data-driven reduced models, extending certified reduced-basis and goal-oriented a posteriori error estimation to the operator-inference setting.
- *Model adaptivity and fusion.* Digital twins must combine mechanistic and data-driven components and adapt them online as operating conditions drift. Consistent coupling of hybrid models, with controlled error propagation across components, is largely open.
- *Uncertainty quantification.* Reliable confidence statements for reduced and hybrid models, including distribution-free guarantees under temporal dependence, are prerequisites for engineering sign-off.
- *State and parameter estimation for control.* Assimilating live measurements into reduced models for feedback control raises questions of observability, identifiability, and stability that must be answered at real-time rates.
- *Continual verification and validation.* As the physical asset, its models, and its data evolve, verification, validation, and uncertainty quantification must be performed continually and under extrapolatory conditions rather than once at build time.

Summary. Digital twins are a core technology for engineering across many industries. Making them scalable and trustworthy depends on research and innovation in mathematics and AI in equal measure. Reduced order modeling, operator inference, and uncertainty quantification supply the technical substrate, while agentic automation promises reach, provided that certified accuracy keeps pace. The right balance of usability and caution is decisive for responsible deployment [7, 9].

References

[1] P. Benner, S. Grivet-Talocia, A. Quarteroni, G. Rozza, W. Schilders, and L. M. Silveira, *Model Order Reduction: Volume I–III*, De Gruyter, 2021.

[2] S. L. Brunton and J. N. Kutz, *Data-Driven Science and Engineering: Machine Learning, Dynamical Systems, and Control*, Cambridge University Press, 2022.

[3] Engineering.com, *Research Report: The State of Simulation, Prototyping and Validation*, 2022.

[4] D. Hartmann, M. Herz, and U. Wever, *Model order reduction a key technology for digital twins*, in Reduced-Order Modeling (ROM) for Simulation and Optimization, Springer, Cham, 2018, pp. 167–179.

[5] D. Hartmann and L. Failer, *A differentiable solver approach to operator inference*, arXiv preprint arXiv:2107.02093, 2021.

[6] D. Hartmann and A. Ciobanas, *Non-linear real-time battery digital twins – efficient and explainable surrogate modeling*, Proceedings of the NAFEMS World Congress 2025, Salzburg, 2025.

[7] D. Hartmann and H. Van der Auweraer, *Digital twins – a golden age for industrial mathematics*, Journal of Mathematics in Industry **15** (2025), no. 1.

[8] B. Kramer, B. Peherstorfer, and K. E. Willcox, *Learning nonlinear reduced models from data with operator inference*, Annual Review of Fluid Mechanics **56** (2024), no. 1, 521–548.

[9] National Academies of Sciences, Engineering, and Medicine, *Foundational Research Gaps and Future Directions for Digital Twins*, The National Academies Press, Washington, DC, 2024.

[10] J. P. Stein, A. Faure Ragani, J. Walsh, and R. Keene, *On the Brink of a Revolution? Engineering Simulation in the Age of AI*, McKinsey & Company and NAFEMS, 2025.

[11] W. I. T. Uy, D. Hartmann, and B. Peherstorfer, *Operator inference with roll outs for learning reduced models from scarce and low-quality data*, Computers & Mathematics with Applications **145** (2023), 224–239.

State-parameter estimation of partially observed stochastic (partial) differential equations

Wilhelm Stannat

(joint work with Gregor Pasemann, Jan Szalankiewicz)

We present a mathematical framework for filtering and parameter estimation of partially observed stochastic (partial) differential equations (S(P)DE), where observations are modelled as functionals of marked point processes whose intensities depend on the underlying signal. We derive both the unnormalized and normalized filtering equations for these systems. We subsequently analyze maximum-likelihood estimation (MLE) of drift parameters in approximating SDE in the particular case of finite-dimensional linear-Gaussian diffusions and establish large-time identifiability, consistency of the online MLE, and a self-normalized central

limit theorem. The primary application of this work is the statistical inference of data collected from confocal laser scanning microscopy recordings of the space-time evolution of intracellular wave dynamics of biophysical quantities (see [2]).

1. Filtering SPDE with MPP observations

Let us first fix the mathematical framework for our filtering problem. We consider the variational approach to SPDE (see [1]). To this end consider a Gelfand triple $\mathcal{V} \subset \mathcal{H} \cong \mathcal{H}^* \subset \mathcal{V}^*$, consisting of a separable real Hilbert space $\mathcal{H} = L^2(\mathcal{D})$ for some open subset $\mathcal{D} \subset \mathbb{R}^d$ and a reflexive Banach space $\mathcal{V}$, densely and continuously embedded into $\mathcal{H}$.

Given a $\mathcal{U}$-valued Q-Wiener process $(W(t))_{t\geq 0}$ on some underlying probability space $(\Omega, \mathcal{F}, \{\mathcal{F}_t\}_{t\geq 0}, \mathbb{P})$, where $\mathcal{U}$ denotes another separable real Hilbert space, we consider SPDE on $\mathcal{H}$ of the type

$$\text{(S)} \qquad \begin{cases} dX(t) = A(X(t))\,dt + B(X(t))\,dW(t), & t \in (0,T], \\ X(0) = \xi \in \mathcal{H} \end{cases}$$

with $B : \mathcal{U} \to \mathcal{H})$ being Hilbert-Schmidt, and $A : \mathcal{V} \to \mathcal{V}^*$ hemicontinuous. We further impose standard monotonicity and coercivity assumptions on A and B, see (H.1)-(H.4) in Chapter 4 of [1] to ensure the existence of an analytically weak and probabilistically strong solution to (S).

Observations Y of (S) are modelled as a marked point process (MPP) on $[0,T] \times \mathcal{K}$, $\mathcal{K} \subset \mathbb{R}^{d_O}$ compact, directed by X with uniformly strictly positive and bounded conditional intensity $\lambda(s, x \mid X) : [0,T] \times \mathcal{K} \times \mathcal{H} \to \mathbb{R}_+$. The filtering problem now consists of computing the conditional distribution

$$\eta_t(A) := \mathbb{E}[1_A(X(t))|\mathcal{Y}_{0:t}], \quad A \in \mathcal{B}(\mathcal{D}),$$

where $\mathcal{Y}_{0:t} := \sigma\{Y_s \mid s \in [0,t]\}$ is the σ-algebra generated by the observation process up to time t.

1.1. The Filtering Equations.

Under the above assumptions on λ, the stochastic exponential

$$\begin{aligned} Z_t(X,Y) := \exp\Big\{ &\int_0^t \int_{\mathcal{K}} \log\{\lambda(s-, x \mid X(s))\} Y(ds, dx) \\ &- \int_0^t \int_{\mathcal{K}} \big(\lambda(s, x \mid X(s)) - 1\big)\, dx\, ds \Big\}, \quad t \in [0,T], \end{aligned}$$

is well-defined. [3] then proves the following path-space representation of η_t, known as Kallianpur-Striebel formula in nonlinear filtering:

Theorem 1.1. *For $A \in \mathcal{B}(\mathcal{D})$*

$$\eta_t(A) = \frac{\mathbb{E}[1_A(X(t)) Z_t(X,Y)|\mathcal{Y}_{0:t}]}{\mathbb{E}[Z_t(X,Y)|\mathcal{Y}_{0:t}]} \quad \mathbb{P}\text{-}a.s., \quad t \in [0,T]. \tag{1}$$

Based on the Kallianpur-Striebel formula (1) [3] derives the following stochastic evolution equation for $\eta_t(\psi) := \int \psi d\eta_t = \mathbb{E}\left[\psi(X(t)) \mid \mathcal{Y}_{0:t}\right]$, which is the analogue of the Kushner-Stratonovich equation in classical nonlinear filtering:

Theorem 1.2. *For any Itô-function ψ, the following equation holds:*

$$\begin{aligned}(2)\qquad \eta_t(\psi) = \eta_0(\psi) &+ \int_0^t \eta_s(\mathcal{L}\psi)\,ds \\ &+ \int_0^t \int_{\mathcal{K}} \frac{\eta_{s-}(\psi\,\lambda(s-,x\mid\cdot)) - \eta_{s-}(\psi)\eta_{s-}(\lambda(s-,x\mid\cdot))}{\eta_{s-}(\lambda(s-,x\mid\cdot))}\times \\ &\qquad\times (Y(\,ds\times\,dx) - \eta_{s-}(\lambda(s-,x\mid\cdot))\mu_{\mathcal{K}}(\,dx)\,ds),\end{aligned}$$

for any $t\in[0,T]$, where

$$(3)\qquad \mathcal{L}\psi = {}_{\mathcal{V}^*}\langle A(\cdot), D^1\psi\rangle_{\mathcal{V}} + \frac{1}{2}tr\{D^2\psi\; B(\cdot)QB(\cdot)^*\}$$

denotes the (infinitesimal) generator of X.

[3] also discusses filtering of X with finite-dimensional spatially averaged observations and with incomplete observation, provides approximation errors and asymptotic consistency results.

2. Parameter estimation in the partially observed case

The second task in applications is to combine filtering with (drift-) parameter estimation of X. As a prototypical example consider the case of a finite, d-dimensional, linear signal observed with an m-dimensional point process:

$$(4)\qquad \begin{aligned} dX_t &= -\,\theta_0\,A\,X_t\,dt + \Sigma^{1/2}\,dW_t \in \mathbb{R}^d \\ Y_{i,t} &= \mathrm{Poiss}\left(\int_0^t \lambda_i(s,X_s)\,ds\right) \quad i=1,\dots,m\,,\end{aligned}$$

where $-A$ is Hurwitz and the parameter θ_0 unknown and assumed to be contained in some compact interval $\Theta := [\theta_{\min},\theta_{\max}]\subset(0,\infty)$. All subsequent results can be found in [4].

In this setting, the (marginal) likelihood of θ, given point process observations $Y_{0:t}$, admits the representation

$$(5)\qquad L_t(\theta) = \exp\left\{\sum_{i=1}^m \int_0^t \log\big(\hat\lambda^\theta_{i,s-}\big)\,dY_{i,s} - \sum_{i=1}^m \int_0^t \big(\hat\lambda^\theta_{i,s}-1\big)\,ds\right\}$$

where $\hat\lambda^\theta_{i,t} := \eta^\theta_t\big(\lambda_i(t,\cdot)\big)$, $i=1,\dots,m$, is the estimated intensity of the observation w.r.t. the conditional law $\mathbb{P}^\theta_{|\mathcal{Y}_{0:t}}$, assuming the value θ as drift parameter in (4). In the particular case where $\theta=\theta_0$ is the true parameter, the measure $\mathbb{P}^{\theta_0}$ is called the *physical law*. Following the idea of the MLE of θ, (local) maxima of L_t and their asymptotic behaviour as $t\to\infty$ are analyzed. Crucial tools for their analysis are the parameter sensitivity

$$(6)\qquad \mathcal{U}^\theta_T := \partial_\theta \log L_T(\theta) = \sum_{i=1}^m \int_0^T \frac{\partial_\theta \hat\lambda^\theta_{i,s-}}{\hat\lambda^\theta_{i,s-}}\,\Big(dY_{i,t} - \hat\lambda^\theta_{i,t}\,dt\Big)$$

in the likelihood and the observed information $\mathcal{I}^\theta_T := -\partial^2_{\theta\theta}\log L_T(\theta)$. Semimartingale decompositions of both, $\mathcal{U}^\theta_T$ and $\mathcal{I}^\theta_T$, w.r.t. the physical measure $\mathbb{P}^{\theta_0}$ are

derived from (5) in [4]. Imposing suitable assumptions (see Assumptions 4.1 and 4.2 in [4]) on (4) and on the large time asymptotics of the observed information $\mathcal{I}_T^\theta$, the following result is obtained in [4]:

Theorem 2.1. *Let* $\hat{\Theta}_T^{glob} := \arg\max_{\theta\in\Theta} \log L_T(\theta)$. *Then there exists an open neighbourhood* $U_0 \subset \Theta$ *of* θ_0 *such that:*

(i) $\mathbb{P}^{\theta_0}(E_T) \to 1$ *as* $T \to \infty$, *where* $E_T := \{\exists!\, \theta \in U_0 \,:\, \mathcal{U}_T^\theta(Y) = 0\}$.

(ii) $\hat{\theta}_T^{loc} \overset{\mathbb{P}^{\theta_0}}{\to} \theta_0$ *as* $T \to \infty$ *holds for the unique solution* $\hat{\theta}_T^{loc}$ *of* $\mathcal{U}_T^\theta(Y) = 0$.

(iii) *As* $T \to \infty$ *the global MLE is unique and coincides with the local MLE* $\hat{\theta}_T$

$$\mathbb{P}^{\theta_0}\left(\hat{\Theta}_T^{\text{glob}} = \{\hat{\theta}_T^{\text{loc}}\}\right) \to 1.$$

(iv) *The central limit theorem holds for the self-normalized estimator*

$$\sqrt{\mathcal{I}_T^{\theta_0}(Y)}\big(\hat{\theta}_T - \theta_0\big) \Rightarrow \mathcal{N}(0,1),\ T \to \infty.$$

References

[1] W. Liu, and M. Röckner, *Stochastic Partial Differential Equations: An Introduction*. Springer, Cham (2015).

[2] G. Pasemann, C. Beta, and W. Stannat, *Stochastic Reaction-Diffusion Systems in Biophysics: Towards a Toolbox for Quantitative Model Evaluation*, in: Stich, M., Carballido-Landeira, J. (eds) Nonlinear Dynamics for Biological Systems. SEMA SIMAI Springer Series, vol 40. Springer, Cham (2025).

[3] J. Szalankiewicz, C. Martinez-Torres, and W. Stannat, *Filtering SPDEs with Spatio-Temporal Point Process Observations*, Stoch PDE: Anal Comp (2026).

[4] J. Szalankiewicz, *Inference for Partially Observed Stochastic Processes - Filtering SPDEs with marked Point Process Observations and Likelihood Theory under Partial Observations*, PhD-Thesis, TU Berlin (2026).

From data assimilation to active sensing: Belief-space optimal control for digital twins

Raúl Tempone

(joint work with Christian Bayer, Saifeddine Ben Naamia, Boualem Djehiche, Eliza Rezvanova, Erik von Schwerin)

A *digital twin* closes a loop between a physical system and its computational model: data are assimilated to reduce uncertainty, and the model is used both to act on the system and to decide what to measure next. We formulate this loop as *partially observed stochastic optimal control* (POSOC). A hidden state evolves in continuous time under a controlled diffusion,

$$\mathrm{d}X_t = b(t, X_t, \alpha_t)\,\mathrm{d}t + \sigma(t, X_t, \alpha_t)\,\mathrm{d}W_t, \qquad X_0 \sim \mu_0,$$

and is observed only at discrete, noisy, and *controllable* times $0 < t_1 < \cdots < t_N \le T$, through $Y_n \sim \pi_n(\cdot \mid X_{t_n}, \beta_n)$ — for instance $Y_n = H_n X_{t_n} + \beta_n \eta_n$, with $\eta_n \sim \mathcal{N}(0, I)$ and H_n a known, possibly low-dimensional, observation map. Two kinds of decisions are made on the system: the control α acts on the dynamics,

while the *sensing design* β_n sets the measurement quality. Since the controller sees only the data history $\mathbf{y}^{[n]} = (Y_{t_1}, \dots, Y_{t_n})$, the state X_t is not Markov from its viewpoint, and $x \mapsto V(t,x)$ is not the right object.

Belief-space formulation. Lifting to the filtering distribution

$$\mu_t = \mathrm{Law}(X_t \mid \mathcal{F}_t^Y) \in \mathcal{P}(\mathbb{R}^{d_x})$$

restores Markovianity. The belief evolves by a hybrid law: a forward (Fokker–Planck) flow $\partial_t \mu_t = \mathcal{G}_{\alpha_t}^* \mu_t$ between observations, where $\mathcal{G}_\alpha$ is the controlled generator and $\mathcal{G}_\alpha^*$ its adjoint, interrupted by Bayes updates $\mu_{t_n} = K_{\beta_n, Y_n}\, \mu_{t_n^-}$ at the measurement times.

The objective is naturally written as a *state-dependent* cost,

$$J(\alpha,\beta) = \mathbb{E}\Big[\int_0^T f_r(X_r,\alpha_r)\,\mathrm{d}r + \sum_n c_{t_n}(X_{t_n},\beta_n) + g(X_T)\Big].$$

Since the admissible controls are adapted to the observation filtration $\mathcal{F}^Y$ and $\mu_t = \mathrm{Law}(X_t \mid \mathcal{F}_t^Y)$, the tower property of conditional expectation collapses each state-dependent term onto the belief,

$$\mathbb{E}\big[f_r(X_r,\alpha_r)\big] = \mathbb{E}\big[\,\mathbb{E}[f_r(X_r,\alpha_r) \mid \mathcal{F}_r^Y]\,\big] = \mathbb{E}\big[\langle \mu_r,\, f_r(\cdot,\alpha_r)\rangle\big],$$

and likewise for the sensing and terminal terms (against $\mu_{t_n^-}$ and μ_T). The cost therefore depends on the trajectory only through the belief,

$$J(\alpha,\beta) = \mathbb{E}\Big[\int_0^T \langle f_r(\cdot,\alpha_r),\mu_r\rangle\,\mathrm{d}r + \sum_n \langle c_{t_n}(\cdot,\beta_n),\mu_{t_n^-}\rangle + \langle g,\mu_T\rangle\Big],$$

which is precisely what makes μ_t the natural state. Decisions are then posed on the belief through the value function

$$V(t,\mu) = \inf_{\alpha,\beta}\ \mathbb{E}\Big[\int_t^T \langle f_r(\cdot,\alpha_r),\mu_r\rangle\,\mathrm{d}r + \sum_{n:\, t<t_n\le T} \langle c_{t_n}(\cdot,\beta_n),\mu_{t_n^-}\rangle + \langle g,\mu_T\rangle \;\Big|\; \mu_t = \mu\Big],$$

so that $V(0,\mu_0) = \inf_{\alpha,\beta} J(\alpha,\beta)$. Choosing the sensing designs β_n that minimize the expected cost is a sequential *optimal experimental design* embedded in the control problem — *active sensing.*

A shared object. For a realized observation path the relevant quantity is a *pathwise cost-to-go* on the state space,

$$\overline{U}_t(x,\mathbf{y}^{[n]}), \qquad x \in \mathbb{R}^{d_x}, \quad \mathbf{y}^{[n]} \in (\mathbb{R}^{d_y})^{N_t},$$

where N_t is the number of observations up to time t; for each fixed path it is a finite-dimensional function of x. Along the optimal path it coincides with the flat (linear-functional) derivative of the value, $\overline{U}_t = \delta V/\delta\mu(t,\mu_t^{\mathbf{y}^{[n]}},\cdot)$. This object underlies the second and third routes below. We present three routes to the problem, trading the global guarantee of the first for scalability.

(A) Dynamic programming. The value satisfies an *interlaced* Hamilton–Jacobi–Bellman (HJB) system. Between observations it solves a PDE on the space of measures, with terminal datum $V(T,\mu) = \langle g, \mu\rangle$,

$$\partial_t V + \inf_\alpha \Big\langle \mu,\ \mathcal{G}_\alpha \tfrac{\delta V}{\delta \mu}(t,\mu,\cdot) + f_t(\cdot,\alpha)\Big\rangle = 0;$$

and at each observation time t_n it satisfies the Bellman–Bayes pasting condition

$$V(t_n^-,\mu) = \inf_\beta \Big\{ \langle c_{t_n}(\cdot,\beta),\mu\rangle + \mathbb{E}\big[V(t_n, K_{\beta,Y_n}\mu)\big]\Big\},$$

where K_{β,Y_n} is the Bayes update and the expectation is over the next observation Y_n drawn under (μ,β). This route characterizes the *global* optimum and is computable for Gaussian beliefs, via PDE solvers in low dimension and sparse / tensor-train solvers in high dimension [1].

(B) Pontryagin maximum principle. Necessary conditions are expressed through two Hamiltonians. Between observations, the *continuous* Hamiltonian optimizes the control,

$$\widetilde{\mathcal{H}}^c(\mu,\overline{U}) = \min_\alpha \big\langle \mathcal{G}_\alpha \overline{U} + f_t(\cdot,\alpha),\ \mu\big\rangle;$$

at each observation time, the *data* (discrete) Hamiltonian optimizes the sensing,

$$\widetilde{\mathcal{H}}^d_n(\mu^-,\overline{U}^+) = \min_\beta \Big\{ \langle c_{t_n}(\cdot,\beta),\mu^-\rangle + \mathbb{E}_\beta\big[\overline{U}^+(\cdot,\mathbf{y}^{[n]})\big]\Big\},$$

where μ^- is the pre-update belief, $\overline{U}^+$ the post-update costate, and the expectation is over the next datum Y_n drawn under β. Together they generate a forward–backward system for the pair $(\mu,\overline{U})$ — the belief flowing forward, the costate backward, with the jumps generated by $\widetilde{\mathcal{H}}^d$ — whose solution is a *fixed point* on the pathwise cost-to-go $\overline{U}_t(\cdot,\mathbf{y}^{[n]})$. Under convexity the costate coincides, along the optimal path, with A's flat gradient $\delta V/\delta\mu$. Crucially, the forward filter and the backward auxiliary are *deterministic* between observations, so the formulation avoids both the infinite-dimensional HJB equation and backward stochastic differential equations [2].

(C) Potential optimization. We recast control as optimization. Both Hamiltonians are *extended* by a bookkeeping weight γ on the running cost — e.g. $\widetilde{\mathcal{H}}^c(\mu,\overline{U},\gamma) = \min_\alpha\langle \mathcal{G}_\alpha\overline{U} + \gamma f,\mu\rangle$, and $\widetilde{\mathcal{H}}^d$ likewise; this extension is *exact* — at $\gamma = 1$ it reproduces the original problem and preserves the value V. A trainable potential $\widehat{U}^\theta$ replaces $\overline{U}$, and the induced policy, sensing, belief flow, and running cost are read off the extended Hamiltonians, so the total cost becomes an objective $\mathcal{J}(\widehat{U}^\theta)$ to be minimized over the parameters θ. The reduced Hamiltonians are pointwise minima of maps affine in the costate, hence not differentiable in general; we therefore *smooth both of them*, replacing the continuous and the data Hamiltonians $\widetilde{\mathcal{H}}^c, \widetilde{\mathcal{H}}^d$ by smooth surrogates $\widetilde{\mathcal{H}}^c_\delta, \widetilde{\mathcal{H}}^d_\delta$ (entropic / square-root) with

$$\big\|\widetilde{\mathcal{H}}^c_\delta - \widetilde{\mathcal{H}}^c\big\|_\infty \longrightarrow 0, \qquad \big\|\widetilde{\mathcal{H}}^d_\delta - \widetilde{\mathcal{H}}^d\big\|_\infty \longrightarrow 0 \quad \text{as } \delta \downarrow 0.$$

The convergence is *uniform*, and it controls the value function: $V_\delta \to V$ as the smoothing vanishes. Training uses reverse-mode automatic differentiation / discrete adjoints through the belief simulator, and an information-relaxation primal–dual gap serves as an in-development suboptimality diagnostic [3].

A program: costate and potential optimization. The third route is one instance of a broader program that recasts optimal control as *optimization over a costate (or potential)*, rather than over policies. The cost is rewritten through an *extended* Hamiltonian that carries a weight γ on the running cost; the reformulation is *exact*, preserving the value function (*value equivalence*), with the running cost recovered as the γ-derivative of the extended Hamiltonian at $\gamma = 1$. Because the reduced Hamiltonian is generally nonsmooth, a *regularized* (smoothed) Hamiltonian then provides adjoint-computable gradients and, through an HJB sub-solution, a primal–dual suboptimality diagnostic; the smoothing is uniform and controls the value, $V_\delta \to V$. The program was developed first for deterministic control — the costate-optimization reformulation with exact value equivalence and value diagnostics [4], together with a Pontryagin solver that smooths the Hamiltonian via a log-sum-exp / concave piecewise-affine bundle surrogate [6], both building on earlier symplectic-Pontryagin and regularized-Hamiltonian approximations [7] and symplectic-Euler error-density adaptivity [8] — then for fully observed stochastic control, where one optimizes a deterministic state potential $\bar{u}(t, x)$ whose gradient (costate) and Hessian enter the extended Hamiltonian, the policy being executed by the exact selector while trained on a smoothed surrogate [5] — and now for POSOC, where the potential lives on the belief / observation path and the same extended and smoothed Hamiltonians (continuous and data) train $\widehat{U}^\theta$. A structural feature throughout is that none of our stochastic formulations require backward stochastic differential equations.

Numerical evidence. On a linear–quadratic Gaussian (LQG) benchmark the three routes agree on the optimal sensing schedule and the resulting cost, validating the scalable methods (B and C) against A's global optimum. The potential route scales to coupled high-dimensional systems: in a high-dimensional LQG study the trained potential remains within a sub-percent optimality gap, relative to the separation-principle optimum, up to $d = 100$, despite the quadratic growth in the number of trainable parameters.

References

[1] C. Bayer, B. Djehiche, E. Rezvanova, and R. F. Tempone, *Continuous-time stochastic optimal control under discrete-time partial observations*, Preprint, arXiv:2407.18018, 2024.

[2] C. Bayer, S. Ben Naamia, E. von Schwerin, and R. F. Tempone, *A Pontryagin maximum principle on the belief space for continuous-time optimal control with discrete observations*, Preprint, arXiv:2512.24916, 2026.

[3] C. Bayer, S. Ben Naamia, E. von Schwerin, and R. F. Tempone, *Hamiltonian potential optimization for partially observed stochastic control*, in preparation, 2026.

[4] S. Ksous, S. Lalvay Segovia, M. Sandberg, E. von Schwerin, A. Szepessy, and R. F. Tempone, *Deterministic optimal control via costate optimization: exact value equivalence and candidate value diagnostics*, Preprint, 2026.

[5] J. Kiessling, S. Lalvay Segovia, K. Malterling Ruderfors, E. von Schwerin, A. Szepessy, and R. Tempone, *Controlled diffusions via Hamiltonian-partials optimization in a potential*, Preprint, 2026.
[6] S. Ksous, S. Lalvay Segovia, M. Sandberg, E. von Schwerin, A. Szepessy, and R. F. Tempone, *Pontryagin-based solver with smoothed Hamiltonian, adaptive Δt, and PA-bundle refinement*, Preprint, arXiv:2606.25731, 2026.
[7] M. Sandberg and A. Szepessy, *Convergence rates of symplectic Pontryagin approximations in optimal control theory*, ESAIM Math. Model. Numer. Anal. **40** (2006), no. 1, 149–173.
[8] J. Karlsson, S. Larsson, M. Sandberg, A. Szepessy, and R. Tempone, *An error estimate for symplectic Euler approximation of optimal control problems*, SIAM J. Sci. Comput. **37** (2015), no. 2, A946–A969.
[9] J. S. Baras, R. J. Elliott, and M. Kohlmann, *The partially observed stochastic minimum principle*, SIAM J. Control Optim. **27** (1989), no. 6, 1279–1292.
[10] A. Bensoussan, *Stochastic Control of Partially Observable Systems*, Cambridge University Press, 1992.
[11] A. Bain and D. Crisan, *Fundamentals of Stochastic Filtering*, Springer, 2009.

Scientific machine learning and digital twinning through stable operator inference: CO_2 methanation test case

Ion Victor Gosea
(joint work with Luisa Peterson, Kai Sundmacher, Peter Benner)

In systems coupled to renewable energy grids, the intermittent nature of the energy supply presents a fundamental design choice. Theoretically, this volatility could be buffered by large-scale intermediate storage (such as hydrogen tanks or battery parks), allowing the downstream chemical plant to remain in steady-state operation. However, to minimize capital expenditures and improve overall system efficiency, modern plant concepts increasingly aim to reduce or eliminate these costly storage units. Under this economic premise, the full dynamic footprint of the renewable energy supply propagates directly into the process. The reactor systems must therefore be capable of withstanding and efficiently navigating highly dynamic load fluctuations, which establishes the demanding control scenario that motivates the methodology developed in this work. A prime candidate requiring such operational flexibility is the catalytic CO2 methanation reactor within a Power-to-X (PtX) process, which we consider as a case study in this work. Specifically, we focus on a single fixed-bed tube of industrial dimensions, acting as a representative unit operation for large-scale production plants scaled via numbering-up. While previous contributions have successfully addressed the physical multi-period design of such load-flexible reactors, the present work focuses on an operational challenge: deriving fast, predictive models to dynamically control these tailored systems. As a core mechanism for renewable energy storage, this process converts hydrogen (H_2) and carbon dioxide (CO_2) into methane (CH_4) and water (H_2O) via an exothermic methanation reaction (the so-called "Sabatier reaction", discovered more than 125 years ago).

Since the hydrogen feedstock is produced via electrolysis using intermittent renewable electricity, the reactor inlet flow fluctuates significantly. The need to

operate dynamically under volatile loads creates a demanding control benchmark, further complicated by strongly nonlinear couplings between mass and heat transport. To capture these phenomena, mechanistic models remain the foundation of chemical engineering. They describe transport, reaction, and thermodynamics from first principles. However, while modern solvers can integrate these models efficiently for single simulations, they often become computationally prohibitive for outer-loop applications such as dynamic optimization, where thousands of model evaluations and gradient computations are executed sequentially within a nonlinear programming solver. Furthermore, they are labor-intensive to set up and require expert knowledge to calibrate them against real plant data. These structural constraints are further tightened in the context of Digital Twins [3]: acting as bi-directional links to a physical asset, they demand models that are not only predictive but can be rapidly updated from live data and embedded into repeated optimal control problems. Conversely, chemical plants offer a complementary resource in the form of dense streams of sensor data, including temperatures, conversions, and flow rates measured in real time. This availability allows for the inference of dynamic models directly from operation [5]. However, purely data-driven approaches often fall short in closed-loop settings: due to a lack of physical structure, they are prone to violations of conservation laws and instability when extrapolating beyond the training regime. The remaining challenge is therefore to exploit operational data while retaining analytical tractability, stability, and computational efficiency tailored for optimization. To address this gap between data-driven flexibility and physical structure, Operator Inference (OpInf) offers a viable alternative [7, 6]. Instead of learning a black-box model, it identifies the explicit matrices of a reduced-order model directly from data. By prescribing a polynomial structure (typically quadratic) motivated by the governing equations, it yields reliable reduced-order dynamic models. This structural data-driven approach provides a significant advantage for process systems engineering: the resulting models possess a low system order, are easily integrated into standard Nonlinear Programming solvers for repeated optimization, and are amenable to system-theoretic analysis and stability certification. Leveraging these structural advantages, this work investigates OpInf for dynamically operated methanation systems. The main contributions of this talk, included in the recent contribution [4], are summarized below:

- We apply OpInf to experimental data from a pilot-scale reactor, demonstrating its applicability to an industrially relevant system beyond standard academic benchmarks.
- We incorporate various existing approaches to account for partially observed states and stability constraints, and extend the framework with (i) a neural decoder for nonlinear state reconstruction and (ii) a hybrid physics-informed formulation for robust coefficient identification from noisy/perturbed measurements.
- We demonstrate the multi-query capability of the resulting reduced-order models by solving a nonlinear open-loop optimal control problem. This is

conducted on a synthetic PtX test case, serving as a proof-of-concept for navigating active constraints under model mismatch.

As mentioned before, the main test case for examining the improvements to OpInf proposed in this contribution (summarized above) stems from process engineering applications. Digital twinning for complex chemical processes requires models that balance predictive accuracy with the ability of rapid execution of simulations [3]. The case study is the catalytic CO_2 methanation reaction, which is achieved by a physical twin represented by a single fixed-bed reactor tube with industrial-scale dimensions. Structured low-dimensional reduced-order models are learned directly from reactor data using specifically tailored OpInf extensions that impose stability guarantees [1] (described in detail in the talk by Peter Benner). An extensive study comparing various surrogate models for this problem was presented in [2]. We follow up on this work and discuss recent results and challenges, belonging to real-world scenarios in which experimental data is used (collected in the laboratory from the actual reactor, i.e., the physical twin).

This work establishes OpInf as a viable framework for data-driven modeling of dynamically operated reactors within PtX processes. Using experimental sensor data, we demonstrated that structured and stable reduced-order models can be identified directly from operation while retaining an interpretable quadratic format. This structure enables a favorable balance between computational efficiency, analytical tractability, and stability, extending the applicability of data-driven models beyond purely predictive tasks. Validation on pilot-scale experimental data confirms that the inferred models remain robust under realistic noise and simultaneous thermal and fluid dynamic load fluctuations.

References

[1] P. Goyal, I. Pontes Duff, and P. Benner, *Guaranteed stable quadratic models and their applications in SINDy and operator inference*, Physica D: Nonlinear Phenomena, page 134893, 2025, `https://doi.org/10.1016/j.physd.2025.134893`.

[2] L. Peterson, A. Forootani, E. I. Sanchez Medina, I. V. Gosea, K. Sundmacher, and P. Benner, *Toward digital twins for Power-to-X processes: Comparing surrogate models for a catalytic CO_2 methanation reactor*, IEEE Transactions on Automation Science and Engineering, 23:3469–3483, 2025, `https://doi.org/10.1109/TASE.2025.3564637`

[3] L. Peterson, I. V. Gosea, P. Benner, and K. Sundmacher, *Digital twins in process engineering: An overview on computational and numerical methods*, Computers & Chemical Engineering, 193:108917, 2025, `https://doi.org/10.1016/j.compchemeng.2024.108917`.

[4] L. Peterson, A. Geschke, P. Goyal, I. V. Gosea, J. Bremer, R. T. Zimmermann, P. Benner, and K. Sundmacher, *Data-driven reduced-order model for optimal control of dynamically operated reactors in Power-to-X processes*, submitted to Chemical Engineering Journal, 2026, `https://doi.org/10.2139/ssrn.6669954`.

[5] P. Benner, W. Schilders, S. Grivet-Talocia, A. Quarteroni, G. Rozza, L. M. Silveira (editors), *Model Order Reduction: Volume 3 Applications*, de Gruyter, 2020. `https://doi:10.1515/9783110499001`.

[6] B. Kramer, Benjamin Peherstorfer, and Karen E Willcox, *Learning nonlinear reduced models from data with operator inference*, Annual Review of Fluid Mechanics, 56(1):521–548, 2024. `https://doi.org/10.1146/annurev-fluid-121021-025220`.

[7] B. Peherstorfer and K. E. Willcox, *Data-driven operator inference for nonintrusive projection- based model reduction*, Computer Methods in Applied Mechanics and Engineering, 306:196–215, 2016, https://doi.org/10.1137/15M1046472

Certified stability of learned nonlinear reduced models

Peter Benner
(joint work with Leonidas Gkimisis, Ion Victor Gosea, Pawan K. Goyal, Igor Pontes Duff)

The integration of prior knowledge into learning frameworks has become increasingly relevant for creating accurate and interpretable surrogate models in various engineering applications, including, in particular also Digital Twins. This is of particular importance when prior knowledge exists, such as model hypotheses or a form of governing equations, but information about system parameters and discretization schemes is lacking (and one would have access solely to time-domain measurements or simulated data).

Adjacent to the realm of black-box modeling that relies on neural networks and machine learning methods, scientific machine learning (SciML) has established itself as a complementary field that bridges physics-based modeling and data-driven learning, combining the theoretical foundations of the former with the representational flexibility of the latter. Following this philosophy, Operator Inference (OpInf) [7, 6] has emerged as a SciML technique that learns the (linear and quadratic) operators of a reduced-order model directly from snapshot data, without requiring access to the full-order model operators. The reduced-order quadratic model is described by the following equations:

$$\frac{d}{dt}\hat{\mathbf{x}}(t) = \hat{\mathbf{A}}\hat{\mathbf{x}}(t) + \hat{\mathbf{H}}\big[\hat{\mathbf{x}}(t) \otimes \hat{\mathbf{x}}(t)\big] + \hat{\mathbf{B}}\mathbf{u}(t). \tag{1}$$

Here, $\hat{\mathbf{A}}, \hat{\mathbf{H}}, \hat{\mathbf{B}}$ denote the reduced linear, quadratic, and input operators, respectively. In contrast to intrusive projection-based techniques, which compute these operators by projecting the discretized governing equations onto a low-dimensional subspace, OpInf estimates them directly from data without requiring access to the original equations.

One of the fundamental challenges in data-driven modeling and SciML is ensuring the stability of learned models. This is also true for OpInf. In recent years, due to the increased popularity of the OpInf method, several solutions have been proposed, including those in [1, 3, 5].

Concerning the linear component of the learned models (matrix $\hat{\mathbf{A}}$), asymptotic stability is achieved provided that all eigenvalues of this matrix possess strictly negative real components. We follow here the explicit analytic solution proposed in [2, 5], where a parameterization of the matrix $\hat{\mathbf{A}}$ is enforced as $\hat{\mathbf{A}} = (\hat{\mathbf{J}} - \hat{\mathbf{R}})\hat{\mathbf{P}}$, which bounds all eigenvalues of it into the left-half complex plane. This is due to the structure of the matrices involved, i.e., $\hat{\mathbf{J}}$ is a skew-symmetric matrix ($\hat{\mathbf{J}}^\top = -\hat{\mathbf{J}}$), $\hat{\mathbf{R}}$ is a symmetric positive definite matrix ($\hat{\mathbf{R}} \succ 0$), and $\hat{\mathbf{P}}$ is a symmetric positive definite matrix ($\hat{\mathbf{P}} \succ 0$) that defines a Lyapunov function $V(\hat{\mathbf{q}}) = \frac{1}{2}\hat{\mathbf{q}}^\top \hat{\mathbf{P}}\hat{\mathbf{q}}$, while $\hat{\mathbf{J}}$

represents energy conservation and $\hat{\mathbf{R}}$ energy dissipation. Hence, local asymptotic stability (LAS) is enforced this way.

However, one cannot guarantee global stability properties of the learned quadratic model if one enforces only LAS, as before. Hence, the problem of enforcing global asymptotic stability (GAS) is studied for the class of quadratic systems. To this aim, the concept of energy-preserving nonlinearities is defined; this allows using a direct Lyapunov method to establish sufficient and necessary conditions for a quadratic system with generalized energy-preserving nonlinearities to be GAS. Then, a parameterization of quadratic systems that inherently has the global stability property is proposed [4], which is then leveraged to infer GAS quadratic models. This was all achieved without using any matrix inequality-based constraint or relying on eigenvalues of the system's matrices, but by using dedicated parameterizations of operators. Moreover, to avoid using derivative information while inferring the operators, a numerical integration scheme is proposed (which is both robust to noise and to scarce data).

Moreover, the problem of inferring quadratic systems with attractor dynamics can be addressed following the explicit parametrization approach. The necessary shift of the dynamics rendering such systems GAS can be learned directly from the data by incorporating it into the OpInf learning problem.

This strong form of learning explicit parameterizations of the learned systems provides a certificate of global stability of the dynamics of the reduced-order systems, in contrast to approaches based on weak stability enforcement via the loss function. A disadvantage of this approach is that the OpInf learning problem, which in its original formulation is a regularized linear least-squares problems becomes a nonlinear, nonconvex optimization problem. As for now, this is solved by (stochastic) gradient descent, but methods with faster convergence should be adapted to this problem in the future.

References

[1] P. den Boef, D. Manvelyan-Stroot, J. Maubach, W. Schilders, and N. van de Wouw, *Stable sparse operator inference for nonlinear structural dynamics*, Computers & Mathematics with Applications, 200:228–242, 2025, `https://doi.org/10.1016/j.camwa.2025.09.017`.

[2] I. Pontes Duff, P. Goyal, and P. Benner, *Stability-certified learning of control systems with quadratic nonlinearities*, IFAC-PapersOnLine, 58(17):151–156, 2024, `https://doi.org/10.1016/j.ifacol.2024.10.129`.

[3] L. Gkimisis, N. Aretz, M. Tezzele, T. Richter, P. Benner, and K. E. Willcox, *Non-intrusive reduced-order modeling for dynamical systems with spatially localized features*, Computer Methods in Applied Mechanics and Engineering, 444:118115, 2025, `https://doi.org/10.1016/j.cma.2025.118115`.

[4] L. Gkimisis, I. Pontes Duff, P. Goyal, and P. Benner, *On the representation of energy-preserving quadratic operators with application to operator inference*, Applied Mathematics Letters, 173:109761, 2026, `https://doi.org/10.1016/j.aml.2025.109761`.

[5] P. Goyal, I. Pontes Duff, and P. Benner, *Guaranteed stable quadratic models and their applications in SINDy and operator inference*, Physica D: Nonlinear Phenomena, page 134893, 2025, `https://doi.org/10.1016/j.physd.2025.134893`.

[6] B. Kramer, B. Peherstorfer, and K. E Willcox, *Learning nonlinear reduced models from data with operator inference*, Annual Review of Fluid Mechanics, 56(1):521–548, 2024. https://doi.org/10.1146/annurev-fluid-121021-025220.

[7] B. Peherstorfer and K. E. Willcox, *Data-driven operator inference for nonintrusive projection- based model reduction*, Computer Methods in Applied Mechanics and Engineering, 306:196–215, 2016, https://doi.org/10.1137/15M1046472.

Understanding generative models for scientific applications

Nisha Chandramoorthy

(joint work with Andrew Dennehy, Ramchandran Muthukumar and Rebecca Willett)

The problem of *generation* is to sample from a target distribution, henceforth denoted p_{data}, given a finite number of data points from the same distribution. No other information about p_{data} is known a priori. A popular algorithm, known as diffusion model or score-based generative model [1, 2] works on the principle of iteratively adding noise to the given samples and learning (again iteratively) a denoiser. The denoiser is approximated as a neural network that is learned by optimizing a variant of a *denoising score-matching loss*. As a result of learning errors, the samples generated after denoising are not distributed according to p_{data} but a close approximation of it. Several theoretical works have studied convergence of diffusion models, which guarantees that Wasserstein-1 or TV metric between the generated and target distributions tends to zero in the limit of zero learning errors. Unfortunately, these integral probability metrics are difficult to approximate in high dimensions and moreover are not informative about how well desired features of p_{data} are represented in the generated distribution. In this work, we pose this problem as an inverse problem for the recovery of a *parameter* associated with p_{data}, which is now considered a member of a parameterized family of probability distributions.

The noising process is typically a variance-preserving Ornstein-Uhlenbeck process of the form: $dX_t = -(\beta_t/2)\, X_t\, dt + \sqrt{\beta_t}\, dW_t$, with a scalar β_t typically decreasing with time. We denote by p_t the marginals of the noising process initialized with p_{data}, i.e., $X_t \sim p_t$ if $X_0 \sim p_{\text{data}}$. The idea of denoising in a diffusion model comes from the observation by Anderson 1982 that the following Ito process shares the same marginals, $p_t : dX_t = -(\beta_t/2)(X_t + 2\,\nabla \log p_t(X_t)) + \sqrt{\beta_t}\, dW_t$. Now, since only the samples from p_{data} are given, we do not know p_t exactly, but these marginals can be approximated by running the OU process above. Then, the idea is that the denoising process can be run backward in time starting with a noisy sample obtained by stopping the forward process at any time, say T. As a result of the shared marginals, the sample returned from the denoising process run backward in time is a sample of $p_0 \equiv p_{\text{data}}$. The marginals p_t are not approximated directly but rather their scores, defined as $\nabla \log p_t$, are solved for to run the denoising. The objective function that is typically used to solve for a neural network that approximates $s(x,t) \approx \nabla \log p_t(x)$ is a form of denoising score matching loss,

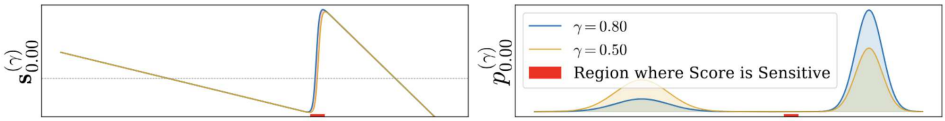

FIGURE 1. Left: Scores of bimodal Gaussians in 1D (shown on the right) at two different mode weights.

which is defined as

$$\ell_{\mathrm{DSM}}(p_0, q_0) := \int_0^T \lambda_t \mathbb{E}_{x\sim q_t} \|\nabla \log p_t(x) - \nabla \log q_t(x)\|^2 \, dt. \tag{1}$$

Here, p_t, q_t are respectively the marginal distributions along the noising process initialized with p_0, q_0. Now let p_{data} be a member of a parameterized family of distributions, with scalar parameter $\gamma = \gamma^*$. Our goal is to analyze if γ^* can be recovered accurately from the generated distribution. Our first contribution is to define a sensitivity index that is predictive of parameter recovery in this context. Specifically, we define a diffusion sensitivity index,

$$L(\gamma^*) := \inf_{\gamma\in\Gamma} \frac{\ell_{\mathrm{DSM}}(p^{(\gamma)}, p^{(\gamma^*)})}{|\gamma - \gamma^*|^2}, \tag{2}$$

which directly compares the value of the loss function and the squared deviation from the true parameter value. First, we note that, when the parameter is the weight of a bimodal distribution, i.e., $p^{(\gamma)}(x) = \gamma\, a(x) + (1-\gamma) b(x)$, where a and b are the mixture components, significant results are known about parameter recovery in the context of score matching. For instance, a number of different statistical learning results [Li, van Breugel, and der Schaar 2024; Koehler Hecket Risteski 2023; Chen Chewi Niles-Weed 2021] point to vanilla score matching (minimizing the integrand in ℓ_{DSM} only at time 0) being insufficient for weight recovery. An intuitive picture of their results is shown in Figure 1. Here we see that the scores of a bimodal Gaussian with well-separated components are not very sensitive to changing the mixture weight. Hence, when we learn the score function, $s_0^{(\gamma^*)} := \nabla \log p^{(\gamma^*)}$, approximately, and sample using the approximate score using Langevin dynamics, we may generate samples according to $p^{(\gamma)}$, with γ not close to γ^*. On the other hand, using the denoising score matching loss above to estimate the scores, $s_t^{(\gamma^*)}$, and then using the denoising process that uses the sequence of approximate scores, in practice, generates samples from $p^{(\gamma)}$ with γ close to γ^*. We hypothesize that this is due to $L(\gamma^*)$ becoming larger as a result of the time-integration in (1). At least for Gaussian mixtures, this is because the sensitivity of the score to the weight increases as the separation between the modes decreases (see Figure 2), and the separation decreases along the OU process, as any distribution mixes exponentially fast to a unimodal Gaussian. By an explicit calculation involving the noising process, we show a lower bound for the sensitivity index, $L(\gamma^*)$ for any $\gamma^* \in (0,1)$ for a bimodal mixture in any dimension.

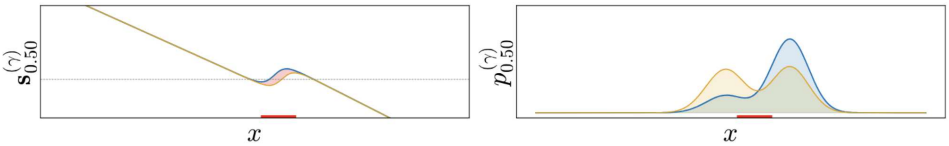

FIGURE 2. Left: Scores at an intermediate time along the noising process starting with the bimodal Gaussians in 1D (shown on the right of Figure 1) at two different mode weights. The corresponding marginals, $p_t^{(\gamma)}$, are shown on the right.

Theorem. [Diffusion sensitivity in bimodal Gaussian mixtures] Consider the family $p^{(\gamma)}(x) = (1-\gamma)\,\mathcal{N}(x;\mu_0, I) + \gamma\,\mathcal{N}(x;\mu_1, I)$ and a linear noise schedule $\beta_t = \beta_1 t$ for $\beta_1 > 0$. Let $\alpha_t = \exp\left(\int_0^t \beta_s \mathrm{d}s\right)$. Then for any $a \in (0, 1/2)$

$$\min_{\hat{\gamma}\in[a,1-a]} \frac{\ell_{\mathrm{DSM}}(p^{(\hat{\gamma})}, p^{(\gamma^*)})}{(\hat{\gamma}-\gamma^*)^2} > \frac{(1-\Phi(\|\mu_1-\mu_0\|\sqrt{\alpha_1}))(\sqrt{k^2+16}-k)}{7\beta_1\gamma^*(1-\gamma^*)}\exp(-(k^2+k)/4) \tag{3}$$

with the minimum value of $C \approx 0.37$ achieved at $\gamma^* = 0.5$. Equivalently, $L(\gamma^*) \geq C$.

While the above result only holds for bimodal mixtures with Gaussian components, we observe empirically that on benchmark datasets, such as MNIST and CIFAR10, $L(\gamma^*)$ is higher than the score sensitivity only at time 0 (the target score sensivity). We also observe that the value of $L(\gamma^*)$ predicts how accurately a well-trained diffusion model can recover γ^*. Our next result explains the underlying mechanism for why sensitivity of scores at intermediate times along the noising process affects parameter recovery. The proof of the result comes from an application of classical results (e.g., Yang and Barron 1999) from minimax estimation theory. The key idea is that, in order for the generated distribution to have an accurate γ, the value of γ must be recoverable from the "score data" along the noising process. It does not matter that the noised samples themselves, which are distributed according to $p_t^{(\gamma^*)}$, may not be sufficiently informative to recover γ^*. Our result below therefore considers a class of "score" estimators, $\Gamma^{(m)}_{\mathrm{score},t}$ that use *score observations* at time t as data to recover the parameter: $\mathcal{S}_m := \{s_t^{(\gamma^*)}(x_t^{(1)}), \cdots, s^t_{(\gamma^*)}(x_t^{(m)})\}$, where $x_t^{(i)} \sim p_t^{(\gamma^*)}$ are noised samples at time t. We denote by $q_t^{(\gamma^*)}$ the distribution of the scores, i.e., $q_t^{(\gamma^*)} = (s_t^{(\gamma^*)})_\sharp p_t^{(\gamma^*)}$. Further, we use $\mathbb{E}_{\gamma^*,t}$ as a shorthand for expectations with respect to the score data distribution, $(q_t^{(\gamma^*)})^{\otimes m}$.

Theorem. Let $\epsilon_{m,t} > 0$ be a covering radius that solves $\epsilon_{m,t} := \sqrt{V(\epsilon_{m,t})/m}$, where $V(\epsilon_{m,t})$ is the $\epsilon_{m,t}$-covering entropy with respect to square root KL divergence of the class of densities $\{q_t^{(\gamma)}\}$. Suppose the bimodal mixture family satisfies $(\gamma-\gamma')^2 \leq A_t^2\,\mathrm{KL}(q_t^{(\gamma)}\|q_t^{(\gamma')})$ for some $A_t > 0$ and all $\gamma, \gamma' \in (0,1)$.

Further, suppose the mixture family exhibits an (ϵ, δ)-sensitivity at time t with $(\epsilon, \delta) = (\epsilon_{m,t}, \epsilon_{m,t}^2/4)$ (see the paper [3] for details). Then, we have the following minimax upper bound on the score restricted estimators using m samples of score data from $\{q_t^{(\gamma)}\}$ at time t:

$$\min_{\hat{\gamma} \in \Gamma^{(m)}_{\text{score},t}} \max_{\gamma^*} \mathbb{E}_{\gamma^*,t} |\gamma^* - \hat{\gamma}|^2 \leq 2\, A_t^2\, \epsilon_{m,t}^2 \,. \tag{4}$$

Simultaneously, if the packing set cardinality $N(\gamma^*, \epsilon_{m,t}) \geq e^{V(\epsilon_{m,t})}$, for all γ^*, it holds that the minimax lower bound achievable for target score-restricted estimators is the following:

$$\min_{\hat{\gamma} \in \Gamma^{(m)}_{\text{score},0}} \max_{\gamma^*} \mathbb{E}_{\gamma^*,0} |\gamma^* - \hat{\gamma}|^2 \geq 4\, A_t^2\, \epsilon_{m,t}^2 \tag{5}$$

□

The above analysis supports the unexpected conclusion that a diffusion model may accurately model the weights of a multimodal distribution even when the target score is insensitive to the weights. Such an analysis therefore has practical implications for rare event modeling (γ^* close to 0 or 1) and improves our understanding of diffusion models and their applicability for scientific data generation.

References

[1] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. *Deep unsupervised learning using nonequilibrium thermodynamics*, In International conference on machine learning, pp. 2256-2265. pmlr, 2015.

[2] 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, 2021.

[3] Andrew Dennehy, Ramchandran Muthukumar, Rebecca Willett, and Nisha Chandramoorthy. *Diffusion models recover correct mixture weights despite score function insensitivity*, Submitted, 2026

Transformer as a nonlinear predictor

Prashant G. Mehta

(joint work with Heng-Sheng Chang and Jin-Won Kim)

A decoder-only transformer computes the conditional probability of the next token through a sequence of L identical layer operations:

$$\text{(Layer for transformer)} \qquad [\sigma_1, \sigma_2, \ldots, \sigma_T]_{d\times T} \mapsto [\sigma_1^+, \sigma_2^+, \ldots, \sigma_T^+]_{d\times T}, \tag{1a}$$

where each layer maps a $(d \times T)$-sequence of (embedding) vectors to a $(d \times T)$-sequence of the same dimension, with σ_t^+ depending only on $\{\sigma_1, \ldots, \sigma_t\}$ (causal structure), for $1 \leq t \leq T$ [1]. The output of the final layer is decoded to give the next-token prediction. This layer structure is the defining architectural feature of the transformer — yet its mathematical justification remains unclear. Why should the optimal prediction architecture take this particular form?

This talk, closely adapted from our paper [2], offers one answer using optimal control theory. We adopt a model-based approach where the observed tokens are generated from an underlying hidden stochastic process, and the prediction problem is cast as causal inference in a partially observed system:

(hidden state process) $X := [X_0, X_1, \ldots, X_T]$ taking values in state-space $\mathbb{S}$,

(observation process) $Z := [Z_1, Z_2 \ldots, Z_T, Z_{T+1}]$ taking values in obs-space $\mathbb{O}$,

where the joint distribution of (X, Z) is according to a probabilistic graphical model. We consider two model classes, based on definition of $\mathbb{S}$ and $\mathbb{O}$, as follows:

Models	**State-space $\mathbb{S}$**	**Observation-space $\mathbb{O}$**
Nonlinear	$\{1, 2, \ldots, d\}$	$\{0, 1, \ldots, m\}$
Linear	$\mathbb{R}^d$	$\mathbb{R}^m$

The nonlinear model setting is directly motivated by the transformer: $\mathbb{O}$ is the vocabulary, and $|\mathbb{S}| = d$ is the embedding dimension. The prediction of interest is the conditional probability $\pi_T(\cdot) := \mathsf{P}(\mathsf{X_T} = \cdot \mid \mathsf{Z_{1:T}})$, a probability vector of dimension d, from which the next-token prediction $\mathsf{P}(\mathsf{Z_{T+1}} = \cdot \mid \mathsf{Z_{1:T}})$ is computed. The linear Gaussian model serves as a tractable baseline in which exact, closed-form results are available. The same notation (X, Z) is used for both model classes; the appropriate class will be clear from context.

Models	**Prediction**	**Motivation**
Nonlinear	$\pi_T := \mathsf{P}(\mathsf{X_T} \mid \mathsf{Z_{1:T}})$	transformer-like
Linear	$\hat{X}_T := \mathsf{E}(\mathsf{X_T} \mid \mathsf{Z_{1:T}})$	tractable baseline

Our objective is to derive transformer-inspired inference architectures where the prediction — conditional probability for the nonlinear model and conditional expectation for the linear model — is computed as a causal transformation of $Z_{1:T}$. The representations are formally introduced in Table 1.

Table 1. Inference architectures for the two model classes: The prediction is computed as a weighted sum of past observations. In the nonlinear predictor, weights are causally adapted to the observation data (i.e., U_t is allowed to depend upon $Z_{1:t}$). This data-dependence is what makes the predictor nonlinear and is the direct analogue of attention weights in a transformer.

Representation	**Weights (or control)**
$\pi_T(f) = (\text{const.}) - \sum_{t=1}^{T} U_{t-1}^\top e(Z_t)$, for $f : \mathbb{S} \to \mathbb{R}$.	$U = \{U_0, \ldots, U_{T-1}\}$ is an $\mathbb{R}^m$-valued adapted process
$f^\top \hat{X}_T = (\text{const.}) - \sum_{t=1}^{T} u_{t-1}^\top Z_t$, for $f \in \mathbb{R}^d$.	$u = \{u_0, u_1, \ldots, u_{T-1}\}$ is an $\mathbb{R}^m$-valued deterministic process

A key finding is that for both model classes, the optimal predictor is computed by iterating a $d \times T$ sequence-to-sequence transformation — structurally analogous to a transformer layer, but derived here from optimality conditions of the prediction problem rather than specified by architecture.

While the representation for the linear predictor is standard for the Gaussian processes, the representation for the nonlinear predictor is an original contribution of our work. The other original contributions are as follows:

Optimal control framework for computing weights. For both model classes, the synthesis of optimal weights — u for the linear predictor and U for the nonlinear predictor — is formulated as an optimal control problem (OCP).

For the linear model, the OCP is solved for the general non-Markovian X, yielding a non-recursive algorithm that is provably more efficient than the recursive Kalman filter architecture ($O(T^2d^2)$ vs $O(T^3d^2)$). For the nonlinear model, the OCP and explicit formula for U are obtained only for the Markovian X, i.e., when (X, Z) is a hidden Markov model (HMM), with complexity $O(T^2d^2)$ — same as a transformer layer — independent of vocabulary size m.

Transformer-like layer operation. The OCP solution defines a layer transformation on a $d \times T$-sequence — directly analogous to the transformer. For the linear model, the $(d \times T)$-sequence is the momentum process, and the layer transformation is defined using Pontryagin's maximum principle:

(1b) (Layer for linear predictor)

$$[\eta_0, \dots, \eta_{T-1}]_{d\times T} \xrightarrow{\text{(Hamilton's equation)}} [\eta_0^+, \dots, \eta_{T-1}^+]_{d\times T}.$$

At convergence, the layer transformation yields the optimal momentum sequence, from which the prediction weights u are computed via an explicit formula.

For the nonlinear model (analysis is limited to HMM), the layer maps the probability sequence:

(1c) (Layer for nonlinear predictor)

$$[\rho_1, \dots, \rho_T]_{d\times T} \xrightarrow{\text{(optimality equation)}} [\rho_1^+, \dots, \rho_T^+]_{d\times T}.$$

At convergence, the layer transformation yields the conditional probability of the hidden state at each time step, from which the weights U are computed via an explicit formula. For both models, the layer transformation emerges from the optimality conditions rather than being specified by architecture.

Numerical experiments provide a direct side-by-side comparison of optimal control weights and attention weights from a trained nanoGPT. When the embedding dimension is insufficient, the dual filter degrades while nanoGPT retains near-optimal performance — revealing that the transformer implicitly exploits non-Markovian structure.

References

[1] Mary Phuong and Marcus Hutter. *Formal algorithms for transformers.* arXiv preprint arXiv:2207.09238, 2022.

[2] Aditya Kudre, Heng-Sheng Chang, and Prashant G. Mehta. *Transformer-like Inference from Optimal Control.* arXiv preprint arXiv:2605.15608, 2026.

Rare event estimation using tensor trains and large deviation theory

Tiangang Cui

(joint work with Alex de Beer, Sergey Dolgov, Karina Koval, Robert Scheichl, Georg Stadler)

Rare-event probabilities arise in applications where low-probability events can have severe consequences. Their accurate estimation plays a vital role in safeguarding digital twins used for critical decision-making. Let θ denote a vector of uncertain parameters taking value in $\mathbb{R}^d$, equipped with a prior density π_0. Let $F : \mathbb{R}^d \to \mathbb{R}$ be a parameter-to-event map. For a threshold z, we define the failure event set $\Omega(z) = \{\theta \in \Omega : F(\theta) \geq z\}$. A typical rare event simulation paradiagm is to estimate the *a priori* failure probability

$$P(z) = \int_\Omega \mathbf{1}_{\Omega(z)}(\theta)\pi_0(\theta)\, d\theta,$$

where θ is distributed according to the prior density π_0. In practical applications where observed data y are available, the prior density is updated to the posterior density

$$\pi(\theta \mid y) = \frac{1}{Z}\ell(y \mid \theta)\pi_0(\theta),$$

where $\ell(y \mid \theta)$ denotes the likelihood function, and Z denotes the unknown normalizing constant. Thus, a more practically relevant paradigm is to estimate the *a posteriori* failure probability

$$P^y(z) = \int_\Omega \mathbf{1}_{\Omega(z)}(\theta)\pi(\theta \mid y)\, d\theta.$$

There are several difficulties in estimating rare events. Firstly, it is well-known that the relative error of a direct Monte Carlo estimator is of order $(NP(z))^{-1/2}$, making direct Monte Carlo computationally prohibitive when $P(z)$ is very small. A standard remedy is importance sampling. However, numerical approximation to the optimal (zero-variance) importance density $\pi^*(\theta) \propto \mathbf{1}_{\Omega(z)}(\theta)\pi_0(\theta)$ using methods such as cross-entropy and Gaussian mixture approximation can suffer from the dimensionality and may fail to resolve accurately the highly concentrated geometry of the failure set. Secondly, the *a posteriori* setting introduces an additional difficulty. In this case one must not only estimate an integral over a small, concentrated tail region, but also account for the unknown normalizing constant of the posterior distribution.

We propose to overcome these challenges by constructing accurate measure transport maps that approximate the optimal biasing densities for rare events,

while also estimating the normalising constant in the *a posteriori* setting. We realise these transport maps numerically using deep compositions of functional tensor trains [1, 2]. To enable adaptive function approximation and reduce the cost of cross interpolation [3], we also consider extended FTT approximations. These first compress the target functions into Tucker form using mode-wise reduced bases and interpolation points, and then apply a tensor train approximation to the reduced core [4].

Another acceleration strategy is to use large-deviation theory (LDT) [5] to obtain a good initial approximation. We also apply a Taylor expansion about the large-deviation point, defined as the point in the failure set $\Omega(z)$ that minimises the rate function, to characterise the parameter sensitivities associated with the failure event. This naturally yields a dimension-reduced approximation, for which we derive error estimators. We further accelerate the construction of the measure transport map using preintegration, or line sampling. LDT provides a natural candidate for defining the coordinate transformation required for preintegration. The resulting higher-order smoothness is expected to make the approximation problem substantially easier.

The combined measure-transport strategy can be naturally extended to the *a posteriori* setting. We first construct a deep tensor measure transport for approximating the posterior, i.e., $\mathcal{T}_\sharp \pi_0(\cdot) \approx \pi(\cdot \mid y)$. Using $\mathcal{T}$ as a preconditioner, the *a posteriori* failure probability can be written as

$$P^y(z) = \frac{1}{Z} \int_\Omega \mathbf{1}_{\Omega(z)}\big(\mathcal{T}(\theta)\big) w(\theta) \pi_0(\theta) \, \mathrm{d}\theta,$$

where

$$w(\theta) = \frac{\ell\big(y \mid \mathcal{T}(\theta)\big) \pi_0\big(\mathcal{T}(\theta)\big) \, |\det \nabla \mathcal{T}(\theta)|}{\pi_0(\theta)}.$$

The numerator integral can again be estimated using the LDT-based acceleration, while the normalising constant is estimated separately using the same measure transport.

We demonstrate the numerical accuracy and dimension scalability of our methods on a range of examples, including a networked susceptible-infectious-recovered model with uncertain infection and recovery rates, and a contaminant transport problem with an uncertain diffusivity field. Our proposed methods achieve substantial acceleration over existing cross-entropy-based methods.

References

[1] T. Cui and S. Dolgov, *Deep composition of tensor-trains using squared inverse Rosenblatt transports*, Found. Comput. Math. **22** (2022), 1863–1922.

[2] T. Cui, S. Dolgov and R. Scheichl, *Deep importance sampling using tensor trains with application to a priori and a posteriori rare events*, SIAM J. Sci. Comput. **46** (2024), C1–C29.

[3] I. Oseledets and E. Tyrtyshnikov, *TT-cross approximation for multidimensional arrays*, Linear Algebra Appl. **432** (2010), 70–88.

[4] C. Strössner, B. Sun, D. Kressner, *Approximation in the extended functional tensor train format*, Adv. Comput. Math. **50** (2024), 54.

[5] S. Tong and G. Stadler, *Large deviation theory-based adaptive importance sampling for rare events in high dimensions*, SIAM/ASA J. Uncertain. Quantif. **11** (2023), 788–813.

Reducing complexity in a class of sequential Monte Carlo samplers

Robert Scheichl

(joint work with Josef Martínek, Julian Hofstadler, Alexander M. G. Cox)

Among the computationally most challenging Bayesian inference problems are *doubly intractable* problems, where constructing even an unbiased likelihood estimator is infeasible [5]. We consider a class of such problems in which a biased likelihood estimator arises from a stochastic approximation of the forward map (a *stochastic solver*), and develop a density-tempered marginalised sequential Monte Carlo (SMC) sampler that exploits a multilevel hierarchy of stochastic solvers of increasing accuracy. The accuracy is increased within both the importance sampling and mutation stages of the SMC sampler, thus leading to significant efficiency gains proportional to the number of tempering and mutation steps.

Setting. Let $\mathcal{U} \subseteq \mathbb{R}^d$ be the parameter space, $\mathcal{Y} \subseteq \mathbb{R}^m$ the observation space, $G\colon \mathcal{U} \to \mathcal{Y}$ the forward map, and let $y = G(u) + E$, where the observational noise $E \sim \mathcal{N}(0, \Gamma)$ is independent of the prior μ_0. Applying Bayes' formula we get $\mathrm{d}\mu/\mathrm{d}\mu_0 \propto \exp(-\Phi)$ with potential $\Phi(u) = \frac{1}{2}\|y - G(u)\|^2_{\Gamma^{-1}}$.

In practice, exact evaluation of G is often impossible and random estimators of G are used instead [2, 1]. To quantify the induced randomness an abstract *noise model* is introduced, i.e., a measurable space $(\mathcal{E}, \mathcal{B}(\mathcal{E}))$ and a Markov kernel $\kappa_S\colon \mathcal{U} \times \mathcal{B}(\mathcal{E}) \to [0,1]$, indexed by an accuracy (and cost) parameter $S > 0$; a *stochastic solver* is then a $(\mu_0 \otimes \kappa_S)$-measurable map $\widetilde{G}\colon \mathcal{U} \times \mathcal{E} \to \mathcal{Y}$. As a representative example, the parameter S can represent the number of samples in a Monte Carlo estimator of G. Writing $\widetilde{\Phi}(u,e) = \frac{1}{2}\|y - \widetilde{G}(u,e)\|^2_{\Gamma^{-1}}$ the *marginal approximation* μ_S^m of μ is defined by

$$\frac{\mathrm{d}\mu_S^m}{\mathrm{d}\mu_0}(u) \;\propto\; \int_{\mathcal{E}} \exp\bigl(-\widetilde{\Phi}(u,e)\bigr)\, \kappa_S(u, \mathrm{d}e),$$

generalising the marginal approximation of [3]. Crucially, κ_S is only required for analysis, not in the implementation. The estimator is biased, i.e., $\mu_S^m \neq \mu$, and depends on S.

The sequence of tempered measures targeted by the Sequential Monte Carlo sampler is defined as follows: First, the number of tempering steps $T \in \mathbb{N}$, tempering parameters $0 = \beta_0 \leq \cdots \leq \beta_T = 1$ and solver accuracies $S_0 \leq \cdots \leq S_T$ are fixed. On the extended space $\mathcal{U} \times \mathcal{E}$ we define

$$\frac{\mathrm{d}\widetilde{\mu}_t^{S_t}}{\mathrm{d}(\mu_0 \otimes \kappa_{S_t})}(u,e) \;\propto\; \exp\bigl(-\beta_t \widetilde{\Phi}(u,e)\bigr), \qquad t = 0, \ldots, T, \tag{1}$$

so that $\widetilde{\mu}_t^{S_t}(\cdot \times \mathcal{E}) = \mu_{t,S_t}^m(\cdot)$ and $\widetilde{\mu}_T^{S_T}$ marginalises to $\mu_{S_T}^m$. The sampler reweights between consecutive measures in (1), then resamples and mutates by applying ℓ transitions with a pseudo-marginal MCMC kernel. We adapt the solver accuracy

S_t across the T tempering steps, and we further optimally split the ℓ transitions within one mutation stage between a cheap and an expensive mutation kernel.

Increasing accuracy in the reweight step. A tempered posterior tolerates a larger amount of solver noise, so there is no need to use the target accuracy S_T at every step. Following [4, Sec. 5], we measure cost by the total number ℓ of mutation kernel transitions, one transition at level t costing S_t.

We fix a growth factor $\gamma \geq 1$ for the solver accuracy and define a sequence of solver accuracies $S_t = \gamma^{t-T} S_T$. Taking $\gamma = 1$ gives the *standard SMC* algorithm, in which $S_t \equiv S_T$; for $\gamma > 1$ we refer to it as the *adaptive SMC* algorithm. The analysis for the adaptive algorithm rests on three assumptions on the sequence of measures in (1), all of them uniform in t and S:

(A1) Bounded weights. The weights to move from the measure at step t with accuracy S_t to the measure at $t+1$ with accuracy γS_t are uniformly bounded.
(A2) Bounded normalising constants. The corresponding ratios of normalising constants are uniformly bounded.
(A3) Bias rate. For S large enough and for all test functions $f : \mathcal{U} \to \mathbb{R}$ in the class $\mathcal{F}_b$ of measurable functions with $|f| \leq 1$, the marginal approximation converges to the posterior at a rate $|\mu_S^m(f) - \mu(f)| \leq C_{\mathrm{B}}\, S^{-1}$.
(A4) Mixing. The mutation kernels are reversible and, for S_t large, geometrically ergodic with L^2 spectral gaps $\rho_t(S_t) \in (0, 1]$.

Given an error tolerance $\varepsilon > 0$, the bias rate is used to determine the target accuracy $S_T = 2C_{\mathrm{B}}/\varepsilon$. From the finite-sample theory in [4] one obtains for both the standard and the adaptive algorithm for any $\delta \in (0, 1)$ that there exists an N and ℓ such that

$$\mathbb{P}\Big(\big|\tfrac{1}{N}\textstyle\sum_n f(u_{T+1}^n) - \mu(f)\big| \leq \varepsilon\Big) \;\geq\; 1 - \delta \qquad \text{for every } f \in \mathcal{F}_b.$$

The adaptive algorithm achieves this accuracy at a much lower cost. The cost gain is characterised by the following theorem.

Theorem (Complexity gain)**.** *Let $\gamma_a > 1$, assume* (A1)–(A4) *for both γ_a and $\gamma = 1$, and let $\liminf_{S\to\infty} \rho_t(S) \geq \rho^* > 0$ uniformly in t. Then the total costs $\mathcal{C}^{\mathrm{St}}$ and $\mathcal{C}^{\mathrm{Ad}}$ of the standard ($S_t \equiv S_T$) and the adaptive ($S_t = \gamma_a^{t-T} S_T$) SMC algorithm satisfy*

$$\liminf_{\varepsilon \searrow 0} \frac{\mathcal{C}^{\mathrm{St}}}{\mathcal{C}^{\mathrm{Ad}}} \;\geq\; T\,\rho^*\,(1 - \gamma_a^{-1}),$$

with equality, and ρ^* omitted, if for all t the limits $\lim_{S\to\infty} \rho_t(S) > 0$ exist.

Thus the gains are proportional to the number of tempering steps T.

Increasing accuracy in mutation. The theorem above takes the cost $\ell \times S_t$ of ℓ transitions with target μ_{t,S_t}^m in the mutation stage as given. The second contribution further reduces that cost by replacing most transitions to the more expensive target chain by cheap approximate ones: apply k transitions of a cheap kernel P_c with invariant measure μ_c and cost $\mathcal{C}^c$ per transition, to decorrelate the resampled particles, and only then $\ell - k$ transitions of the expensive kernel P_e with

invariant measure $\mu_e = \mu^m_{t,S_t}$ and cost $\mathcal{C}^e$, to correct. A coupling argument shows that geometric ergodicity of a kernel is inherited by its N-fold product acting on the N SMC particles, at the same rate, which leads to the following error estimate.

Proposition (Accelerated mutation error)**.** *The error of one mutation stage consisting of k transitions of a cheap kernel P_c followed by $\ell - k$ transitions of an expensive kernel P_e, measured in a weighted Wasserstein distance $\mathcal{W}$ on the product space of N particles and starting from an initial measure η, is bounded by*

$$\mathcal{W}\big(\eta P_c^k P_e^{\ell-k},\, \mu_e\big) \;\le\; K_e \tau_e^{\ell-k}\Big(K_c \tau_c^k\, \mathcal{W}(\eta, \mu_c) + \mathcal{W}(\mu_c, \mu_e)\Big), \tag{2}$$

where (K_e, τ_e) and (K_c, τ_c) are the ergodicity constants of the two kernels.

Now, one can minimise (2) subject to a cost budget of $\mathcal{C}^c k + \mathcal{C}^e(\ell - k) = \mathcal{C}_{\text{budg}}$, which has a unique solution that can be characterised in closed form. Informally, it falls into two regimes. In the *edge cases* one uses only P_c (if the budget is too tight, or P_c is highly accurate) or only P_e (if P_c is too inaccurate, or too expensive to be worth its bias). Otherwise a mix of both kernels is used and there is an optimal number $k^* > 0$ of cheap transitions that is independent of the budget C_{budg}; all remaining resources should then be spent on the expensive chain.

Under appropriate assumptions this two level analysis can again be extended to a multilevel analysis leading to gains proportional to the number of mutation steps ℓ. In practical applications in high dimensions and with highly concentrating posteriors, T and ℓ can both go into the hundreds or even thousands, thus offering substantial gains for our *adaptive SMC* algorithm.

References

[1] P. R. Conrad, M. Girolami, S. Särkkä, A. Stuart and K. Zygalakis, *Statistical analysis of differential equations: introducing probability measures on numerical solutions*, Stat. Comput. **27** (2017), 1065–1082.

[2] M. B. Giles, *Multilevel Monte Carlo methods*, Acta Numer. **24** (2015), 259–328.

[3] H. C. Lie, T. J. Sullivan and A. L. Teckentrup, *Random forward models and log-likelihoods in Bayesian inverse problems*, SIAM/ASA J. Uncertain. Quantif. **6** (2018), 1600–1629.

[4] J. Marion, J. Mathews and S. C. Schmidler, *Finite-sample complexity of sequential Monte Carlo estimators*, Ann. Statist. **51** (2023), 1357–1375.

[5] J. Park and M. Haran, *Bayesian inference in the presence of intractable normalizing functions*, J. Amer. Statist. Assoc. **113** (2018), 1372–1390.

Fast uncertainty quantified probabilistic hazard maps for decision support

Elaine T. Spiller

Large-scale natural geophysical hazards – debris flows, storm surge from hurricanes, granular volcanic flows – pose a constant threat to exposed communities due to the possibility of rare catastrophic events. Further, there is a paucity of data for these large-scale, rare events. As such we rely on physics-based, computational models of the hazardous phenomena (debris flows and volcanic flows) solved over terrain represented by digital elevation models. We further model the aleatory

variability as a probability density function (pdf) on possible event scenarios. This aleatory pdf and the computational model of the mass flows are both imperfect. Thus we also seek to model, where possible, the associated epistemic uncertainties through Monte Carlo (MC) or importance-sampled Monte Carlo. Evaluating one scenario of the flow model is computationally intensive (1-10 hours) and as such is a bottleneck for MC simulations. We exploit various Gaussian process (GP) surrogates of the computational models that are well-suited to geophysical flows – zero-censored GPs appropriate for non-negative data (flow inundation is positive or zero) [1] and the parallel partial emulator (PPE) for large-dimensional output data [2] (flow inundation at millions of pixels). Once trained, we can compute probabilistic hazard maps with MC utilizing GP surrogates in a few minutes on a laptop. This allows for the generation of many probabilistic hazard maps thus enabling aleatory scenario model exploration as MC simulations no longer require expensive simulations. We show examples of uncertainty quantified probabilistic hazard maps and hazard analyses for volcanic flow hazards at several volcanoes including Aluto (Ethiopia) [3] and Soufrière Hills (Montserrat) [4] as well as post-fire debris flows in Southern California [5].

References

[1] E.T. Spiller, R.W. Wolpert, P. Tierz, and T.G. Asher. *The zero problem: Gaussian process emulators for range constrained computer models.* SIAM/ASA J. Uncert. Quant, **11** (2023) 540–566.

[2] M. Gu, and J.O. Berger, *Parallel partial Gaussian process emulation for computer models with massive output.* Ann. Appl. Stat. **10** (2016) 1317–1347.

[3] P. Tierz, E.T. Spiller, B.A. Clarke, F. Dessalegn, Y. Bekele, E. Lewi, G. Yirgu, R.L. Wolpert, S.C. Loughlin, and E.S. Calder. *Topographic controls on pyroclastic density current hazard at Pluto volcano (Ethiopia) as seen with a novel zero-censored Gaussian Process emulator.* Journal of Geophysical Research, **129** (2024) e2023JB028645.

[4] E.T. Spiller, R.L. Wolpert, S.E. Ogburn, E.S. Calder, J.O. Berger, A.K. Patra, and E.B. Pitman. *Volcanic hazard assessment during an eruption hiatus, or post-eruption unrest context: Modeling continued dome collapse hazards for Soufrière Hills Volcano.* Frontiers in Earth Science Geohazards and Georisks, **8** (2020) 396.

[5] E.T. Spiller, L.A. McGuire, P. Patel, A. Patra, and E.B. Pitman. *Quantifying fire effects on debris flow runout using a morphodynamic model and stochastic surrogates.* Natural Hazards and Earth System Sciences **26** (2026) 1705–1725.

Some lessons from incorporating data into stochastic control problems

Sören Christensen

(joint work with Claudia Strauch, Lukas Trottner)

A central feature of digital twins is the interaction between data and decision making. On the one hand, data are used to estimate unknown system characteristics that are relevant for decision making. On the other hand, the implemented control policy influences the future evolution of the system and therefore the data available for subsequent learning. The talk discussed this interplay in a class of classical stochastic control problems with explicit solutions. Instead of assuming

that the diffusion dynamics are fully known, the talk considered the statistically more realistic situation where the drift function is unknown and has to be learned nonparametrically from observations.

As a first step, an *offline learning* setting was studied, where a continuous trajectory of the uncontrolled diffusion is available for statistical inference. For two classical long-run average control problems – impulse control and singular control – the explicit structure of the optimal policy makes it possible to reduce the learning problem to estimating suitable functionals of the invariant density. This leads to plug-in policies based on nonparametric estimators of the invariant density rather than the drift itself, avoiding a statistically much more difficult task. The resulting data-driven control policies satisfy non-asymptotic regret guarantees. For singular control, the regret converges at the rate $\mathcal{O}(\sqrt{\log(T)/T})$, whereas for impulse control the rate is $\mathcal{O}(T^{-1/2})$, which is minimax optimal up to a logarithmic factor. A main observation in the analysis is that the control objective itself determines both the statistical quantities that need to be estimated (stationary density instead of drift) and the appropriate error metric: while impulse control only requires L^1 accuracy of the density estimator, singular control depends on uniform (L^∞) control.

The second part of the talk considered the more realistic *online learning* setting, where observations are generated by the controlled system itself. In this case, learning and control become intrinsically coupled, since the implemented policy determines which regions of the state space are explored and hence which information becomes available for future estimation. To address this exploration-exploitation trade-off, an episodic learning strategy was proposed that alternates between dedicated exploration phases and exploitation phases based on the current statistical estimate. This approach yields non-asymptotic regret bounds of order $T^{2/3}$ (up to logarithmic factors). For several extensions, including parametric models and Lévy-driven dynamics, faster rates can be achieved.

Finally, recent work was briefly discussed that suggests that explicit exploration is not always necessary. By using the nested structure of threshold policies, the thresholds can be learned more directly, leading to regret of essentially order $\mathcal{O}(\sqrt{T})$, matching the statistical lower bound up to logarithmic factors.

Although the models considered are rather simple, they already exhibit three characteristic features that might also be relevant for digital twins. First, the decision problem determines which quantities should be estimated. Second, it determines which statistical error measures are relevant for guaranteeing good decisions. Third, data generation and decision making form a feedback loop: control actions influence the observations available for future learning.

References

[1] S. Christensen and C. Strauch, *Nonparametric learning for impulse control problems–Exploration vs. exploitation*, Annals of Applied Probability **33** (2023), no. 2, 1369–1387.

[2] S. Christensen, C. Strauch and L. Trottner, *Learning to reflect: A unifying approach for data-driven stochastic control strategies*, Bernoulli **30** (2024), no. 3, 2074–2101.

[3] S. Christensen, A. H. Thomsen and L. Trottner, *Data-Driven Rules for Multidimensional Reflection Problems*, SIAM/ASA Journal on Uncertainty Quantification **12** (2024), no. 4, 1240–1272.

Separable approximations of optimal value functions, decaying sensitivity, and neural networks

Lars Grüne

(joint work with Dante Kalise, Luca Saluzzi, and Mario Sperl)

In this talk I have presented the continuation of our studies to find out when deep neural networks are able to represent high-dimensional functions appearing in optimal control problems, such as the optimal value function $V : \mathbb{R}^d \to \mathbb{R}$ or the Q-function $Q : \mathbb{R}^d \times U \to \mathbb{R}$. Here the dimension d should be thought of in the order of magnitude of up to several hundred.

Such functions typically do not enjoy a high degree of smoothness, at least not for deterministic optimal control problems, which is why approaches requiring smoothness are not generally applicable. However, large-scale optimal control problems often have dynamics consisting of many low-dimensional subsystems, which are only coupled to a set of neighboring systems but not to all other subsystems. Formally, this means that we can split the large scale system into q subsystems

$$z_i^+ = f_i(z_i, z_{-i}, u_i), \quad i = 1, \dots, q, \qquad z_{-i} = (z_j)_{j \in \mathcal{N}(i)},$$

where $z_i \in \mathbb{R}^{d_i}$ is the state of the i-th subsystem and $\mathcal{N}(i)$ collects the indices of the neighbors of the i-th subsystem. This distributed structure can be formalized by an interconnection graph, where each node represents a subsystem and an edge between two nodes indicates that one system is a neighbor of the other.

It is known that this distributed structure carries over to the costs of the problem over longer time horizons [1], but the distributed structure becomes increasingly more complex as time increases. Much better results can be obtained if one specifically looks at the *optimal* cost. Here, it was shown in [2] for linear quadratic infinite horizon problems that under suitable uniform controllability and detectability conditions, the influence that one system has on another decreases exponentially with the distance in the interconnection graph, a property known as (exponentially) decaying sensitivity. This property induces a corresponding property of the optimal value function V, which allows to approximate this function by separable functions of the form

$$V(x) \approx \sum_{k=1}^{s} \Psi_k^l(z_{i_{k,1}}, z_{i_{k,2}}, \dots, z_{i_{k,l}}),$$

where the approximation error can be estimated depending on the number l, see [3] for details. The right hand side of this expression can then, in turn, be approximated by specially structured separable neural networks, whose number of neurons

grows only polynomially with the dimension d and the degree of the polynomial depends on l.

While numerical results suggest that the decisive decaying sensitivity property also holds for nonlinear problems, so far it has only be shown rigorously for linear quadratic problems. The reason is that the proof in [2] crucially relies on sophisticated matrix analysis, which is not easily carried over to nonlinear problems. In order to fix this, in [4] we developed a proof technique that uses graph-theoretic methods, employing the interconnection graph described above, instead of matrix analysis. Using this technique, a first decaying sensitivity result was obtained for nonlinear dynamics.

References

[1] Wei Kang and Qi Gong, *Feedforward neural networks and compositional functions with applications to dynamical systems*, SIAM Journal on Control and Optimization **60** (2022), 786–813

[2] Sungho Shin, Mihai Anitescu, and Victor M. Zavala, *Exponential decay of sensitivity in graph-structured nonlinear programs*, SIAM Journal on Optimization **32**, (2023), 1156–1183

[3] Mario Sperl, Luca Saluzzi, Dante Kalise, and Lars Grüne, *Separable approximations of optimal value functions and their representation by neural networks*, SIAM Journal on Control and Optimization **64** (2026), 1099–1126.

[4] Lars Grüne and Mario Sperl, *Decaying sensitivity of the zero solution for a class of nonlinear optimal control problems*, arXiv preprint 2602.05020. To appear in the Proceeding of the International Symposium on Mathematical Theory of Networks and Systems, MTNS 2026

Proximal importance sampling

George Biros

(joint work with Youguang Chen)

We are considering the problem of sampling from a posterior distribution related to Bayesian inverse problems arising in science, engineering, and digital twin technologies. Our method belongs to the family of independence Metropolis–Hastings (IMH) sampling algorithms. These are quite common in Bayesian inference. Relying on the existence of an approximate posterior distribution that is cheaper to sample from but can have significant bias, we introduce Proximal-IMH, a scheme that removes this bias: it corrects samples from the approximate posterior solving an auxiliary optimization problem, yielding a local adjustment that trades off adherence to the exact model against stability around the approximate reference point. For idealized settings, we prove that the proximal correction tightens the match between approximate and exact posteriors, and thereby improves acceptance rates and mixing. The new method works with both linear and nonlinear input-output operators and is especially suitable for inverse problems where exact posterior sampling is too expensive. We perform several numerical experiments that include multimodal and data-driven priors and nonlinear input-output operators. The results show that Proximal-IMH reliably outperforms existing IMH variants. The results appeared in [1].

Problem statement

The problem setting is given

$$y = \mathbf{A}x + e, \qquad e \sim \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I}), \qquad x \sim p(x): \text{ prior},$$

where $\mathbf{A}$ is the forward operator indicating the physics of the problem, x are the unknown parameters, e is the observation noise with zero mean and specified variance σ^2, and a $p(x)$ a given prior for the target vector parameter x.

Given observations y, our goal is to develop efficient algorithms for sampling from the posterior distribution

$$\pi(x \mid y) \propto \exp\left(-\|y - \mathbf{A}x\|^2/2\sigma^2\right) p(x).$$

The most reliable method is using a Metropolis Hastings (MH) method. This requires a proposal distribution. Existing approaches are very slow to converge. In this work our main assumption is that we have access on an approximate operator $\widetilde{\mathbf{A}}$ that is spectrally equivalent to $\mathbf{A}$ and faster to compute. Then using any method we assume we can sample from

$$\widetilde{\pi}(x \mid y) \propto \exp\left(-\|y - \widetilde{\mathbf{A}}x\|^2/2\sigma^2\right) p(x).$$

Main results

We present and analyze an **importance sampling MH** method, which we term **proximal IMH**. The method first samples from $\widetilde{\pi}(x \mid y)$ and the corrects the sample by solving an optimization problem:

$$x = \arg\min_x \|\mathbf{A}(x) - \widetilde{\mathbf{A}}(\widetilde{x})\|^2 + \beta\|x - \widetilde{x}\|^2,$$

and then filtering x using a standard MH acceptance.

We show the following:

- We prove mixing times for proximal IMH and compare with MALA mixing times. Proximal IMH is insensitive to the condition number for $\mathbf{A}$ as opposed to MALA.
- We empirically extend the method to nonlinear forward problems.
- We test the method on problems in which $p(x)$ is generated by variational autoencoders.
- We apply the method to medium frequency inverse scattering.

References

[1] Youguang Chen and George Biros, *Proximal-IMH: Proximal Posterior Proposals for Independent Metropolis-Hastings with Approximate Operators*, arXiv preprint arXiv:2602.21426 (2026)

Machine learning for thermodynamic property modeling

Roland Herzog
(joint work with Viktor Martinek)

Accurate knowledge of thermophysical properties underpins the design, simulation, and optimization of processes throughout the energy and process industries. These properties are commonly encoded in thermodynamic equations of state (EOS), which relate quantities such as pressure, temperature, density, and composition. Such EOS are precisely the physics-based component on which a digital twin of a chemical or energy process must ultimately rely: the predictive fidelity of the twin, and hence its usefulness for control, optimization, and decision support, is bounded by the accuracy of the property models embedded in it. The development of an EOS is expensive and time-consuming. On the experimental side it rests on extensive measurement campaigns, and on the modeling side it relies on empirical functional forms that are assembled by trial and error and refined manually using expert knowledge. The resulting models may therefore be structurally suboptimal, and their creation generally scales poorly to the growing number of substances and mixtures of practical interest.

This talk describes an effort to accelerate and partly automate this process by combining machine learning for model discovery with optimal experimental design for data acquisition. Its centerpiece is Thermodynamics-informed Symbolic Regression (TiSR), an open-source symbolic regression tool tailored to the specific needs of thermodynamic EOS development [4]. Symbolic regression searches simultaneously for the functional form of a model and for its parameters, and thereby produces closed-form, interpretable expressions rather than black-box surrogates. TiSR is built on genetic programming, in which a population of candidate expressions, is iteratively selected, mutated, and recombined, while the numerical parameters of each candidate are estimated by a Levenberg–Marquardt method. This methodology is generic, but its value for thermodynamics rests on a set of domain-specific extensions that form the core of our work.

A first extension concerns the incorporation of prior knowledge in the form of shape constraints. A shape constraint prescribes a descriptive attribute that the model is required to satisfy — for instance a sign, a direction of monotonicity, a curvature, an asymptotic limit, or the value of an integral. Such constraints are indispensable in thermodynamics, where data are often sparse and noisy, and where inter- and extrapolation must respect physical laws. We enforce them during parameter estimation by minimizing constraint violations through a penalized least-squares formulation solved with gradient-based optimization [6]. This step is standard in classical thermophysical property modeling, but had not previously been integrated into the parameter estimation of symbolic regression [3].

A second extension addresses shared expressions. The idea is to reuse one and the same expression across a whole set of substances, e.g., fluids, re-fitting only part of its parameters per substance, so that abundant measurements on some fluids help constrain the description of others – a form of transfer learning. We

push this further through partial parameter sharing and the simultaneous use of several categorical variables, which broadens the class of problems that symbolic regression can address in this setting [5].

A third direction targets fundamental equations of state formulated in the Helmholtz energy, which occupy a special place because they capture the complete thermodynamic behavior of a fluid within a single potential. The challenge is that this potential is not itself observable; it enters the measurable quantities only through its derivatives. A model must therefore be calibrated against a collection of heterogeneous property measurements at once, rather than against samples of the target function, and it is this setting that we approach with symbolic regression [1, 2].

Beyond these methodological ingredients, a large part of the design effort in TiSR is devoted to usability: the intended audience is the domain scientist with no background in evolutionary computation, and the tool accordingly provides protective defaults, curated presets that bundle related settings, and extensive documentation. The tool is under active development and is already being used by other groups; it is available at `https://github.com/scoop-group/TiSR`.

Upcoming additions to TiSR include robustness to outliers, model-selection support, automated development of fundamental EOS, and a tighter coupling of model discovery with optimal experimental design in order to reduce measurement effort. This coupling mirrors the two-way information flow that characterizes a digital twin: data acquired from the physical system are assimilated to refine the property model, while the model in turn guides where new measurements are most informative. Symbolic regression contributes interpretable, physically constrained models to this loop, which is a prerequisite for the trust that safety- and cost-critical digital-twin applications demand.

References

[1] O. Frotscher, V. Martinek, R. Fingerhut, X. Yang, J. Vrabec, R. Herzog, and M. Richter, *Proof of concept for fast equation of state development using an integrated experimental-computational approach*, International Journal of Thermophysics **44** (2023), no. 7, doi:10.1007/s10765-023-03197-z.

[2] O. Frotscher, V. Martinek, X. Xiao, P. Falloon, E. May, R. Herzog, and M. Richter, *Comparison of strategies for the development of a Helmholtz equation of state for solid benzene I*, Poster presented at the Thermodynamik-Kolloquium, Hannover/Garbsen, Germany (2023), doi:10.5281/zenodo.11387692.

[3] G. Kronberger, F. Olivetti de França, B. Burlacu, C. Haider, and M. Kommenda, *Shape-constrained symbolic regression—improving extrapolation with prior knowledge*, Evolutionary Computation **30** (2022), no. 1, 75–98, doi:10.1162/evco_a_00294.

[4] V. Martinek, O. Frotscher, M. Richter, and R. Herzog, *Introducing thermodynamics-informed symbolic regression—a tool for thermodynamic equations of state development*, arXiv:2309.02805 (2023).

[5] V. Martinek and R. Herzog, *Symbolic regression for shared expressions: introducing partial parameter sharing*, arXiv:2601.04051 (2026).

[6] V. Martinek, J. Reuter, O. Frotscher, S. Mostaghim, M. Richter, and R. Herzog, *Shape constraints in symbolic regression using penalized least squares*, Machine Learning and Principles and Practice of Knowledge Discovery in Databases (M. Cerrato, D. Kalinauskaitė,

M. Lukoševičius, M. Pechenizkiy, and K. Šutienė, eds.), Springer Nature Switzerland, 2026, pp. 224–239, doi:10.1007/978-3-032-25305-7_16.

Global optimization of tensor polynomials

Didier Henrion
(joint work with Llorenç Balada Gaggioli and Milan Korda)

We introduce a new moment-SOS (sum of squares a.k.a. Lasserre) hierarchy of semidefinite programming (SDP) relaxations for minimizing high-dimensional polynomials with tensor structure, generating certified lower bounds with convergence guarantees to the global minimum. The standard moment-SOS hierarchy suffers from the curse of dimensionality, where the size of the SDP blocks grows exponentially with the number of variables. Sparse variants (exploiting correlative or term sparsity) mitigate this but still struggle if the sparsity graph of the polynomial is dense.

First, we introduce LRPOP (Low-Rank Polynomial Optimization) by exploiting the Canonical Polyadic (CP) tensor decomposition of the objective polynomial. A polynomial $p(x)$ is viewed as a tensor and decomposed into a sum of r separable products:

$$p(x) = \sum_{k=1}^{r} \prod_{i=1}^{n} p_{k,i}(x_i)$$

where each $p_{k,i}(x_i)$ is a univariate polynomial of the single variable x_i. By introducing lifting variables for the intermediate products of this decomposition, we construct a specific sparsity graph that admits a clique tree decomposition with small maximal clique size. Crucially, the size of the resulting SDP blocks depends linearly on the CP rank r rather than the ambient dimension n. This allows LRPOP to solve global optimization problems with thousands of variables and high degrees, provided the polynomial has a low CP rank. The method is shown to generate certified lower bounds converging to the global minimum. It uses Bernstein basis polynomials to maintain numerical stability in high dimensions.

Then, we introduce TTPOP (Tensor-Train Polynomial Optimization) by replacing the CP decomposition with the Tensor Train (TT) decomposition. The high-dimensional polynomial $p(x)$ is represented as a sequence of 3-way core tensors contracted together:

$$p(x) = \prod_{i=1}^{n} P_i(x_i)$$

where each core $P_i(x_i)$ is a matrix-valued polynomial of the single variable x_i. Unlike CP, the TT decomposition relies on stable singular value decomposition algorithms, avoiding the divergence issues often seen when approximating high-dimensional tensors with CP. Instead of a single rank r, the complexity is controlled by a vector of TT ranks, corresponding to the dimensions of the matrices $P_i(x_i)$. This allows for more flexible modelling of polynomials that might require a high CP rank but have low TT ranks. The storage and optimization complexity in TTPOP

scales linearly with the number of variables and polynomially with the TT ranks, making it effective for very high-dimensional polynomials for which LRPOP might struggle. This extension effectively adapts the lifting and graph-sparsity techniques to the chain-like structure of the TT, resulting in an optimization hierarchy that is robust and scalable for a broader class of high-dimensional functions.

Reporter: Philip Schär

Participants

Dr. Sandra Babyale
Institut für Mathematik
Universität Potsdam
14415 Potsdam
GERMANY

Dr. Ricardo Baptista
Department of Statistical Sciences
University of Toronto
700 University Avenue
Toronto ON M5G 1X6
CANADA

Prof. Dr. Peter Benner
Max-Planck-Institut für
Dynamik komplexer techn. Systeme
Sandtorstr. 1
39106 Magdeburg
GERMANY

Prof. Dr. George Biros
Oden Institute for Computational
Engineering and Sciences
The University of Texas at Austin
1 University Station C1200
Austin, TX 78712-1082
UNITED STATES

Prof. Dr. Marc Bocquet
CEREA, ENPC, EDF R&D
Institut Polytechnique de Paris
6-8 Avenue Blaise Pascal
77455 Marne-la-Vallée
FRANCE

Dr. Nisha Chandramoorthy
Department of Statistics
Committee on Computational and
Applied Mathematics
The University of Chicago
5747 S Ellis Avenue
Chicago, IL 60637
UNITED STATES

Dr. Liangying Chen
Fachbereich Mathematik
Technische Universität Berlin
Straße des 17. Juni 136
10623 Berlin
GERMANY

Prof. Dr. Sören Christensen
Mathematisches Seminar
Christian-Albrechts-Universität Kiel
Heinrich-Hecht-Platz 6
24118 Kiel
GERMANY

Dr. Tiangang Cui
School of Mathematics & Statistics
University of Sydney
Sydney NSW 2006
AUSTRALIA

Prof. Dr. Omar Ghattas
Oden Institute for Computational
Engineering and Sciences,
The University of Texas at Austin
1 University Station C 0200
Austin TX 78712-1085
UNITED STATES

Dr. Ion Victor Gosea
Max-Planck-Institut für
Dynamik komplexer techn. Systeme
Sandtorstr. 1
39106 Magdeburg
GERMANY

Prof. Dr. Georg A. Gottwald
School of Mathematics and Statistics
The University of Sydney
Sydney 2206
AUSTRALIA

Prof. Dr. Lars Grüne
Mathematisches Institut
Lehrstuhl für Angewandte Mathematik
Universität Bayreuth
95440 Bayreuth
GERMANY

Dr. Philipp Guth
Johann Radon Institute for
Computational and Applied
Mathematics
Austrian Academy of Sciences
4040 Linz
AUSTRIA

Dirk Hartmann
Siemens Digital Industries Software
Otto-Hahn-Ring 6
81739 Munich
GERMANY

Prof. Dr. Matthias Heinkenschloss
Department of Computational Applied
Mathematics and Operations Research
Rice University
6100 S. Main Street
Houston, TX 77005-1892
UNITED STATES

Prof. Dr. Didier Henrion
LAAS – CNRS
Université de Toulouse
7, avenue du Colonel Roche
31400 Toulouse
FRANCE

Prof. Dr. Roland Herzog
Heidelberg University
Interdisciplinary Center for Scientific
Computing
Im Neuenheimer Feld 205
69120 Heidelberg
GERMANY

Prof. Dr. Michael Hintermüller
Weierstraß-Institut für
Angewandte Analysis und Stochastik
Anton-Wilhelm-Amo-Straße 39
10117 Berlin
GERMANY

Prof. Dr. Marina Höhne
Leibniz-Institut für Agrartechnik
und Bioökonomie
Max-Eyth-Allee 100
14469 Potsdam
GERMANY

Dr. Dante Kalise
Department of Mathematics
Imperial College London
South Kensington Campus
London SW7 2AZ
UNITED KINGDOM

Prof. Dr. Boris Krämer
Department of Mechanical and
Aerospace Engineering
University of California San Diego
9500 Gilman Drive
La Jolla CA 92093-0411
UNITED STATES

Prof. Dr. Karl Kunisch
Institut für Mathematik und
wissenschaftliches Rechnen
Karl-Franzens-Universität Graz
Heinrichstrasse 36
8010 Graz
AUSTRIA

Charlotte Lämbgen
Institut für Mathematik
FU Berlin
Arnimalle 7
14195 Berlin
GERMANY

Prof. Dr. Prashant Mehta
Coordinated Science Laboratory
Dept. of Mechanical Science and Engineering
University of Illinois at Urbana Champaign
1308 W. Main Street
Urbana, IL 61801
UNITED STATES

Marco Morik
BIFOLD
TU Berlin
Franklinstr. 28/29
Postfach FR6-1
10587 Berlin
GERMANY

Prof. Dr. Houman Owhadi
Department of Computing and Mathematical Sciences
California Institute of Technology
1200 E California blvd
Pasadena CA 91125
UNITED STATES

Prof. Dr. Sebastian Reich
Institut für Mathematik
Universität Potsdam
Karl-Liebknecht-Straße 24-25
14476 Potsdam
GERMANY

Hans Reimann
Institut für Mathematik
Universität Heidelberg
Im Neuenheimer Feld 205
69120 Heidelberg
GERMANY

Dr. Philip Schär
Institut für Mathematik
Universität Potsdam
Campus Golm, Haus 9
Karl-Liebknecht-Str. 24-25
14476 Potsdam
GERMANY

Prof. Dr. Robert Scheichl
Institut für Mathematik
Universität Heidelberg
Im Neuenheimer Feld 205
69120 Heidelberg
GERMANY

Zakhar Shumaylov
Centre for Mathematical Sciences
University of Cambridge
Wilberforce Road
Cambridge CB3 0WB
UNITED KINGDOM

Elaine Spiller
Dept. of Mathematical and Statistical Sciences
Marquette University
Milwaukee, WI 53233
UNITED STATES

Prof. Dr. Georg Stadler
Courant Institute of Mathematical Sciences
New York University
251, Mercer Street
New York NY 10012-1110
UNITED STATES

Prof. Dr. Wilhelm Stannat
Institut für Mathematik
Technische Universität Berlin
Sekr. MA 7-2
Straße des 17. Juni 136
10623 Berlin
GERMANY

Prof. Dr. Ing. Raul F. Tempone
Mathematics
King Abdullah University of Science and Technology (KAUST)
Thuwal 23955
SAUDI ARABIA

Maia Tienstra
Fachbereich Mathematik
Universität Potsdam
Karl-Liebknecht Str. 24-25
14476 Potsdam
GERMANY

Margaret Trautner
Departement Mathematik
ETH Zürich
Rämistrasse 101
8092 Zürich
SWITZERLAND

Paolo Villani
Institut für Mathematik
FU Berlin
Arnimalle 6
14195 Berlin
GERMANY

Prof. Dr. Ir. Femke Vossepoel
Faculty of Civil Engineering and
Geosciences
Delft University of Technology
Stevinweg 1
Delft, 2628 CN
NETHERLANDS

Dr. André-Alexander Zepernick
Institut für Mathematik
Freie Universität Berlin
Arnimallee 6
14195 Berlin
GERMANY