Measure transport approaches to uncertainty quantification

D. Bigoni A. Spantini Y. Marzouk

Massachusetts Institute of Technology

✉: {dabi,spantini,ymarz}@mit.edu

http://transportmaps.mit.edu

Introduction

The solution of many inference problems requires the evaluation of integrals with respect to a complex ***target*** *distribution ν_π. Choosing a tractable* ***reference*** *distribution ν_ρ, we solve the* ***transportation problem****, i.e., we seek the map T that pushes forward ν_ρ to ν_π, denoted $T_\sharp \nu_\rho = \nu_\pi$. In other words, we seek:*

$$T : \mathbb{R}^d \to \mathbb{R}^d \quad \text{s.t.} \quad \nu_\rho(A) = \nu_\pi(T(A)) \ , \quad \forall A \in \sigma(\mathbb{R}^d) \ .$$

This renders challenging integration problems tractable:

$$I[f] = \int f(\boldsymbol{x}) \nu_\pi(d\boldsymbol{x}) = \int f \circ T(\boldsymbol{x}) \nu_\rho(d\boldsymbol{x}) \ .$$

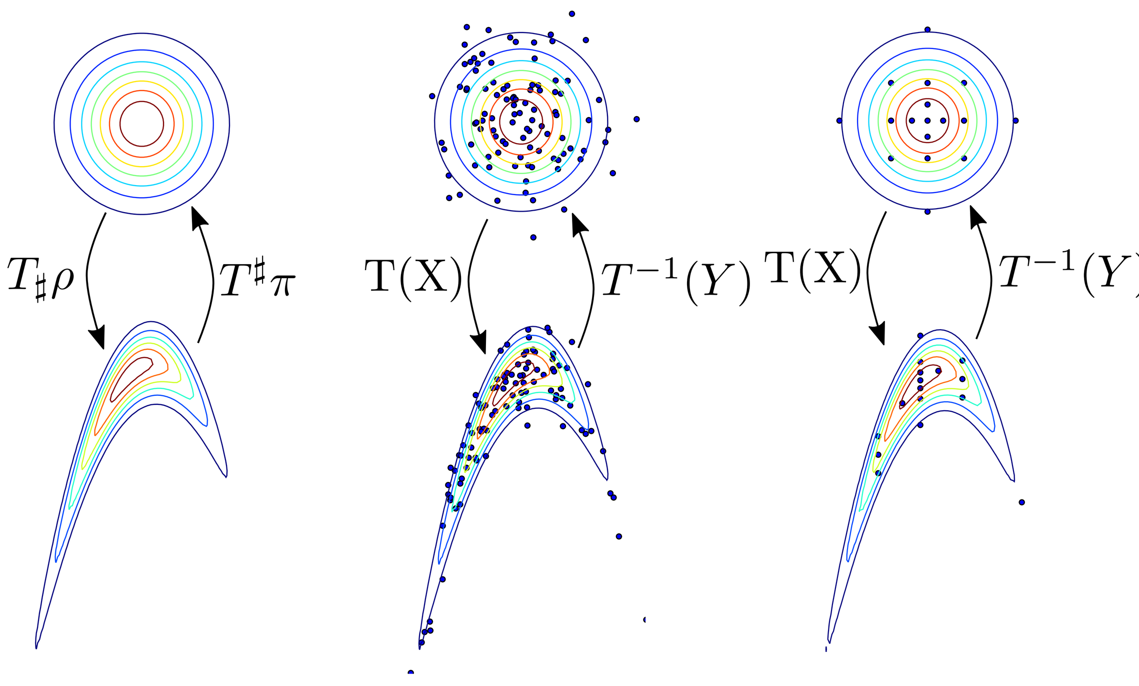
Measure transport

Let ρ be the density of ν_ρ and π be the density of ν_π.
The map T s.t. $T_\sharp \nu_\rho = \nu_\pi$ defines the following identities:

$$\textbf{Pushforward}: \quad T_\sharp \rho(\boldsymbol{x}) = \rho \circ T^{-1}(\boldsymbol{x}) \, |\nabla T^{-1}(\boldsymbol{x})| = \pi \ , \qquad (1)$$

$$\textbf{Pullback}: \quad T^\sharp \pi(\boldsymbol{x}) = \pi \circ T(\boldsymbol{x}) \, |\nabla T(\boldsymbol{x})| = \rho \ , \qquad (2)$$

and for $\mathbf{X} \sim \nu_\rho$, $T(\mathbf{X}) \sim \nu_\pi$.

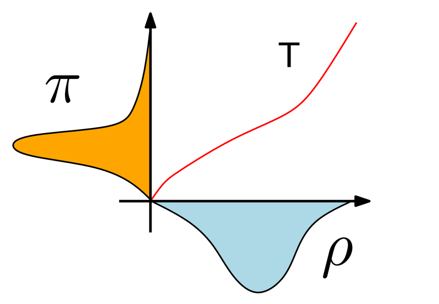


(a) Densities (b) Monte Carlo (c) Sparse grid

Knothe-Rosenblatt rearrangement

For any ν_ρ, ν_π Lebesgue absolutely continuous there exists a **triangular monotone** map $T \in \mathcal{T}_>$ s.t. $T(d\nu_\rho) = d\nu_\pi$

$$T(\boldsymbol{x}) = \begin{bmatrix} T_1(x_1) \\ T_2(x_1, x_2) \\ \vdots \\ T_d(x_1, x_2, \dots, x_d) \end{bmatrix}$$



The transport map framework

The transportation problem is cast as a minimization problem in terms of the **Kullback-Leibler divergence** [1, 2, 3]:

$$T^\star = \underset{T \in \mathcal{T}_>}{\arg\min} \, \mathcal{D}_{\text{KL}}(T_\sharp \nu_\rho \| \nu_\pi) = \underset{T \in \mathcal{T}_>}{\arg\min} \, \mathbb{E}_\rho \left[-\log T^\sharp \tilde{\pi} \right] \ .$$

We approximate $\mathbb{E}_\rho$ using the quadrature $\mathcal{Q}_q$:

$$T^\star \approx T^\star_q = \underset{T \in \mathcal{T}_>}{\arg\min} \, \mathcal{Q}_q \left(-\log T^\sharp \tilde{\pi} \right) \ .$$

The elements in $\mathcal{T}_>$ are defined by their components $T^{(i)}$,

$$T^{(i)}(x_{1:i}) = c^{(i)}(x_{1:i-1}) + \int_0^{x_i} \left(h^{(i)}(x_{1:i-1}, t) \right)^2 + \varepsilon \, dt \ .$$

Let $\mathcal{T}^k_> \subset \mathcal{T}_>$, with $n_k = \dim(\mathcal{T}^k_>) < \dim(\mathcal{T}_>) = \infty$.
This leads to the further approximation:

$$T^\star_q \approx T^\star_{q,k} = \underset{T \in \mathcal{T}^k_>}{\arg\min} \, \mathcal{Q}_q \left(-\log T^\sharp \tilde{\pi} \right) \ .$$

The following **variance diagnostic** is a global convergence criterion for the approximation of $T^\star$ in the space $\mathcal{T}_>$:

$$\mathcal{D}_{\text{KL}}(T_\sharp \nu_\rho \| \nu_\pi) \approx \frac{1}{2} \mathbb{V} \left[\log \frac{\rho}{T^\sharp \tilde{\pi}} \right] \quad \text{as} \quad T \to T^\star$$

[1] T. a. El Moselhy et al. "Bayesian inference with optimal maps". In: *Journal of Computational Physics* 231.23 (Oct. 2012).

[2] Y. Marzouk et al. "Sampling via Measure Transport: An Introduction". In: *Handbook of Uncertainty Quantification*. Ed. by R. G. Ghanem et al. Cham: Springer International Publishing, 2016. arXiv: 1602.05023.

[3] D. Bigoni et al. "On the computation of monotone transports". In: *preprint* (2017).

[4] M. Parno et al. "A Multiscale Strategy for Bayesian Inference Using Transport Maps". In: *SIAM/ASA Journal on Uncertainty Quantification* 4.1 (2016).

[5] A. Spantini et al. "Inference via low-dimensional couplings". In: *ArXiv e-prints* (2017). arXiv: 1703.06131 [stat.ME].

[6] T. Cui et al. "Dimension-independent likelihood-informed MCMC". In: *Journal of Computational Physics* 304 (2016), pp. 109–137.

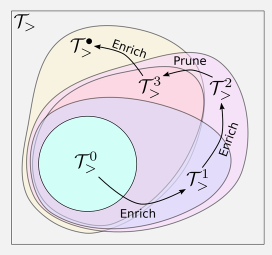
Types of low-dimensional structure

Smoothness and marginal independence [3] (adaptivity)

We seek $\mathcal{T}^\bullet_> \subset \mathcal{T}_>$ ($\dim \mathcal{T}^\bullet_> < \infty$) s.t.

$$\exists \, T \in \mathcal{T}^\bullet_> \quad \text{s.t.} \quad \mathbb{V}\left[\log \rho / T^\sharp \pi \right] < \varepsilon_\bullet$$

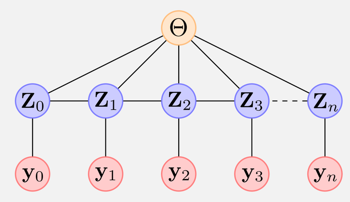
- **First variation** for enrichment
- **Sample average approximation**



Multi-level, multi-fidelity, multi-scale preconditioning [4]

For a sequence of increasingly accurate and expensive distributions $\pi_1, \dots, \pi_\ell$, we use the hierarchy of problems $(T_i)_\sharp \rho = \mathfrak{T}^\sharp_{i-1} \pi_i$, $\mathfrak{T}_k = T_1 \circ \dots \circ T_k$, to obtain $(\mathfrak{T}_\ell)_\sharp \rho = \pi_\ell$.

Conditional independence [5] (filtering and smoothing)



The $d_{\boldsymbol{\Theta}} \times n \cdot d_{\mathbf{Z}}$ dimensional distribution of $\boldsymbol{\Theta}, \mathbf{Z}_{1:n} | \mathbf{y}_{1:n}$ can be represented by

$$\mathfrak{T} = T_1 \circ T_2 \circ \dots \circ T_{n-1}$$

where T_i are $d_{\boldsymbol{\Theta}} \times 2 \cdot d_{\mathbf{Z}}$ maps.

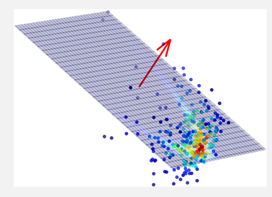
Low-rank transports (likelihood inf./active subspaces) [6]

For $\rho = \mathcal{N}(0, \mathbf{I})$ and a rotation U s.t.

$$U^\sharp \pi(\boldsymbol{x}) = \mu(\boldsymbol{x}_{1:\kappa}) \eta(\boldsymbol{x}_{\kappa+1:d}) \ , \quad \eta = \mathcal{N}(0, \mathbf{I}) \ ,$$

there is a κ-dimensional map T s.t.

$$(U \circ T)_\sharp \rho = \pi$$



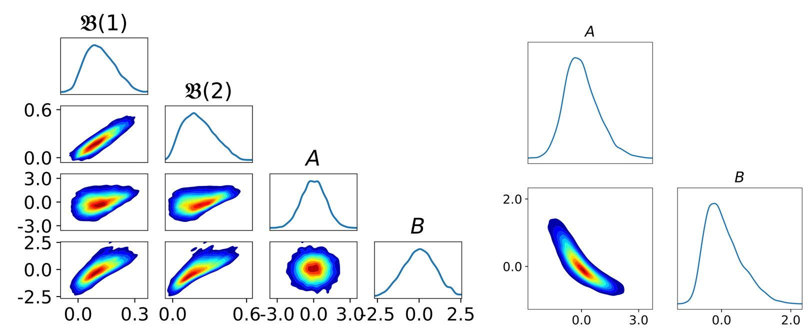
Biochemical oxygen demand model

Inference of the coefficients A and B of the BOD model:

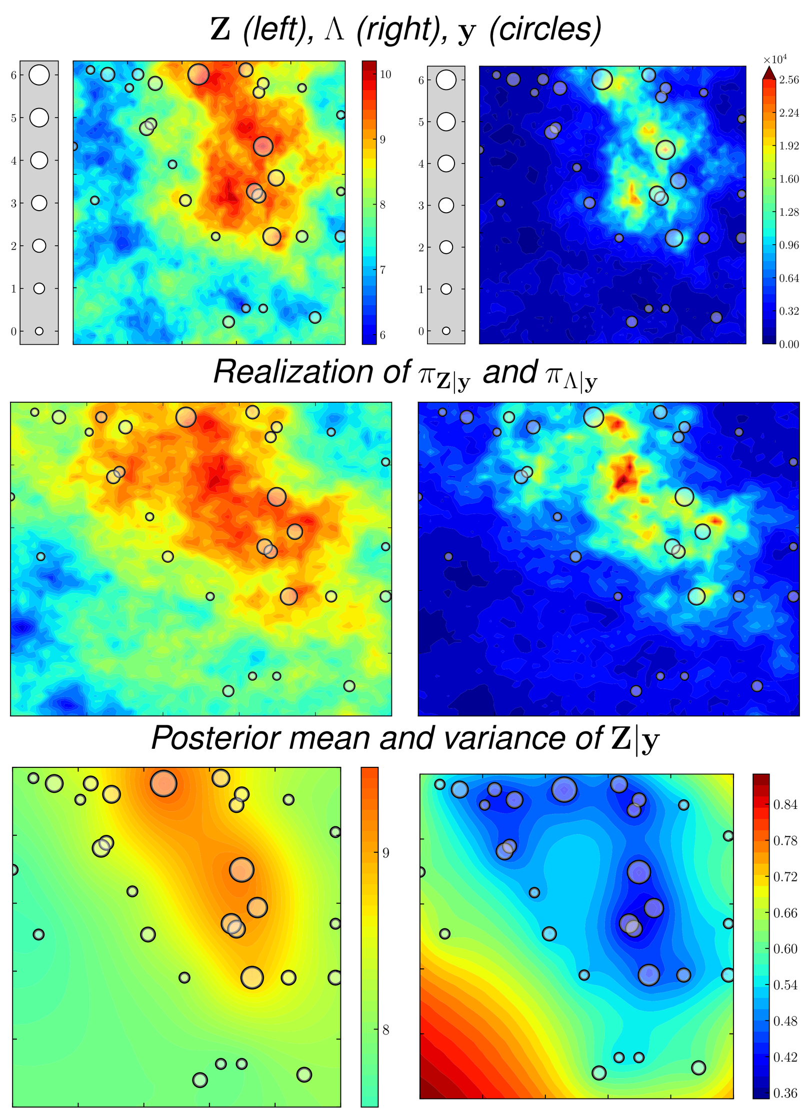
$$\mathfrak{B}(t) = A(1 - \exp(-Bt)) + \varepsilon \ , \quad \varepsilon \sim \mathcal{N}(0, \sigma^2) \ ,$$
$$A \sim \mathcal{U}(0.4, 1.2) \ , \quad B \sim \mathcal{U}(0.01, 0.31) \ ,$$

given n observations $D := [\mathfrak{B}(1), \dots, \mathfrak{B}(n)]$.
We approximate the joint distribution $\pi_{\mathfrak{B}(1), \mathfrak{B}(2), A, B} \approx T_\sharp \rho$.
This map can be used for **fast online inference**, noting

$$x_1, x_2 \sim \rho \quad \text{then} \quad \begin{bmatrix} T_3(b_1, b_2, x_1) \\ T_4(b_1, b_2, x_1, x_2) \end{bmatrix} \sim \pi_{A,B | \mathfrak{B}(1) = b_1, \mathfrak{B}(2) = b_2}$$



Log-Gaussian Cox process



In the $n = 64 \times 64$ grid on $\mathcal{D} = [0,1]^2$, with $\mathbf{s}_i$ being the center of cell i, we consider the process $(\mathbf{Y}_i)_{i=1}^n$ representing the number of points in a cell:

$$\mathbf{Y}_i \sim \text{Poisson}(\boldsymbol{\Lambda}_i) \ , \quad \mathbf{Y}_i \perp\!\!\!\perp \mathbf{Y}_j | \boldsymbol{\Lambda} \ , \quad \boldsymbol{\Lambda} = \exp \mathbf{Z}$$

$$\mathbf{Z} \sim \mathcal{N}(2 \log 64, \boldsymbol{\Sigma}) \ , \quad \boldsymbol{\Sigma}_{ij} = \exp\left(-\frac{\|\mathbf{s}_i - \mathbf{s}_j\|_2}{7/10} \right)$$

Given observations $\boldsymbol{y} = (\boldsymbol{y}_1, \dots, \boldsymbol{y}_d)$ we characterize

$$\pi_{\mathbf{Z}|\boldsymbol{y}}(\boldsymbol{z}) \propto \pi_{\boldsymbol{y}|\mathbf{Z}}(\boldsymbol{z}) \pi_{\mathbf{Z}}(\boldsymbol{z})$$

The likelihood $\pi_{\mathbf{y}|\mathbf{Z}}(\mathbf{z}) = g(\mathbf{z}_1, \dots, \mathbf{z}_d)$ is local, then the map

$$T_{\text{pr}}(\boldsymbol{x}) = \begin{Bmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{Bmatrix} + \begin{bmatrix} \mathbf{L}_{11} & \mathbf{0} \\ \mathbf{L}_{12} & \mathbf{L}_{22} \end{bmatrix} \boldsymbol{x} \ , \quad \boldsymbol{\Sigma} = \mathbf{L}\mathbf{L}^\top \ ,$$

is such that $T^\sharp_{\text{pr}} \pi_{\mathbf{Z}|\boldsymbol{y}}(\boldsymbol{z}) = g(\boldsymbol{z}_{1:d}) \eta(\boldsymbol{z})$, where $\eta = \mathcal{N}(0, \mathbf{I})$.
In other words, $T^\sharp_{\text{pr}} \pi_{\mathbf{Z}|\boldsymbol{y}}$ **departs from** $\mathcal{N}(0, \mathbf{I})$ **in a** $d \ll n$ **dimensional subspace**.
Hence a d dimensional map T is sufficient $(T_{\text{pr}} \circ T)_\sharp \rho = \pi_{\mathbf{Z}|\boldsymbol{y}}$.

Stochastic volatility

AR(1) process to model the log-volatility X_t of an asset:

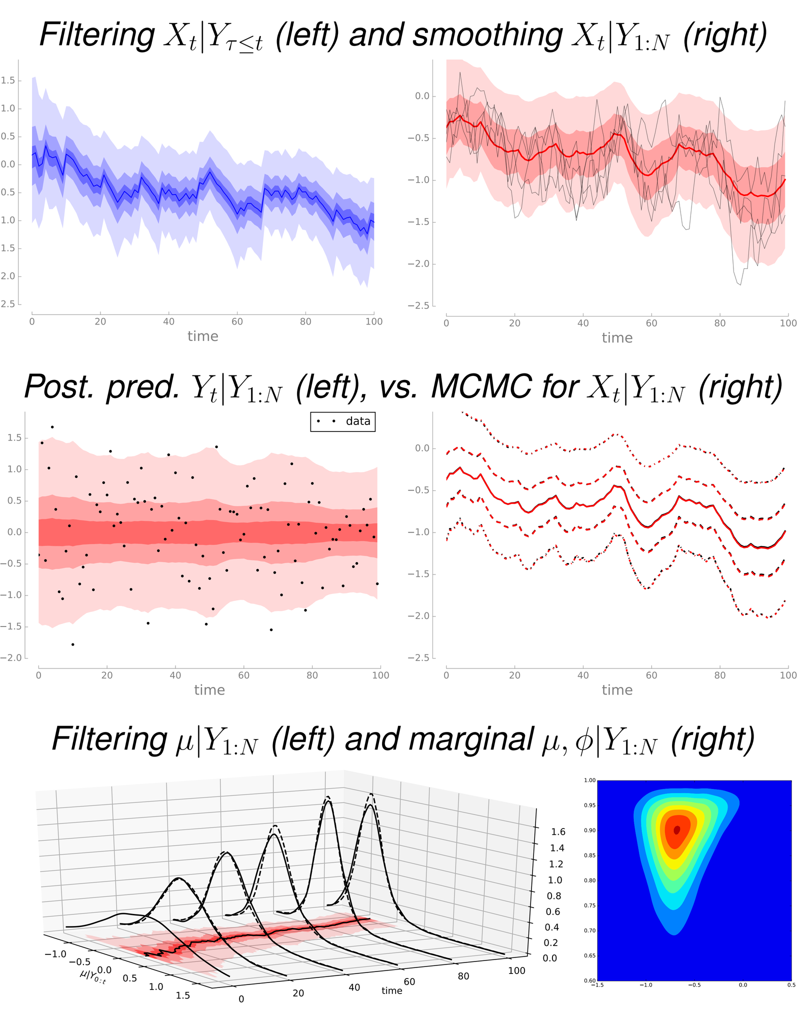
$$X_{t+1} = \mu + \phi(X_t - \mu) + \eta_t \ ,$$
$$X_1 \sim \mathcal{N}\left(0, 1/\left(1 - \phi^2\right)\right) \ , \quad \eta_t \sim \mathcal{N}(0,1) \ ,$$
$$\mu \sim \mathcal{N}(0,1) \ , \quad \phi = 2 \frac{\exp(\phi^\star)}{1 + \exp(\phi^\star)} - 1 \ , \quad \phi^\star \sim \mathcal{N}(3,1) \ .$$

The observed return Y_t follows the price evolution model

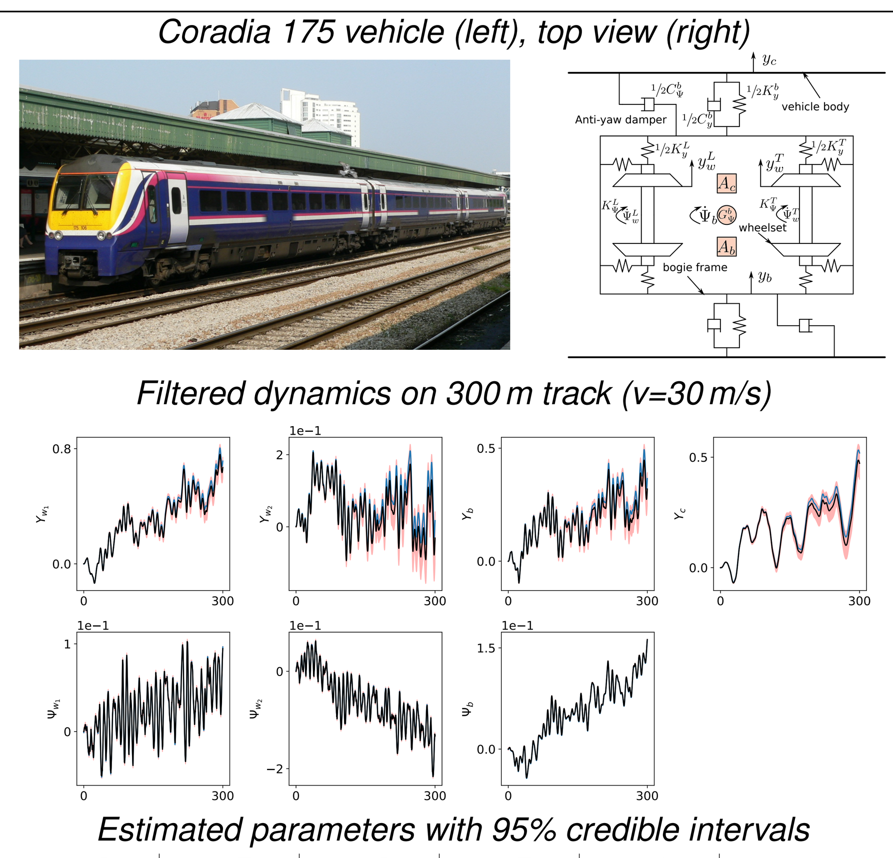
$$Y_t = \varepsilon_t \exp(X_t / 2) \ , \quad \varepsilon_t \sim \mathcal{N}(0,1) \ .$$

We characterize the **full posterior** $\mu, \phi, \mathbf{X}_{1:N} | \mathbf{Y}_{1:N} \sim \pi$.
For $N = 100$, this is a 102-dimensional problem.
We use the **conditional independence** structure of π and solve instead a sequence of $(N-1)$ 4-dimensional problems.
The composition of these low-dimensional maps is s.t.:

$$(T_1 \circ \dots \circ T_{N-1})_\sharp \rho \approx \pi$$



Railway vehicle dynamics



Estimated parameters with 95% credible intervals

K^L_y	K^T_y	K^L_Ψ	K^T_Ψ	K^b_y	C^b_y
4.0×10^6	4.0×10^6	4.0×10^6	4.1×10^6	1.6×10^5	1.6×10^4
7.5%	7.9%	9.0%	8.2%	1.5%	0.8%
C^b_Ψ	f_{11}	f_{22}	I_w	I_b	λ
5.0×10^5	7.4×10^6	6.1×10^6	5.9×10^2	2.2×10^3	1.5×10^{-1}
7.0%	9.4%	8.9%	9.6%	8.1%	7.1%

Dynamics of a vehicle running on perturbed tracks:

$$\mathbf{X}_{k+1} = \Phi(\boldsymbol{\theta}) \mathbf{X}_k + \varepsilon_k \ , \quad \varepsilon_k \sim \mathcal{N}(0, \Sigma(\boldsymbol{\theta})) \ .$$

We observe $\mathbf{Y}_k = [\ddot{y}_b, \ddot{y}_c, \dot{\Psi}_b]_k$ through accelerometers:

$$\mathbf{Y}_k = A(\boldsymbol{\theta}) \mathbf{X}_k + w_k \ , \quad w_k \sim \mathcal{N}(0, \Sigma_w) .$$

In this setting $\Phi(\boldsymbol{\theta})$, $\Sigma(\boldsymbol{\theta})$, and $A(\boldsymbol{\theta})$ are nonlinear in $\boldsymbol{\theta}$.
We perform Bayesian inference for the parameters:

$$\pi(\boldsymbol{\theta} | \mathbf{Y}_{1:N}) \propto \underbrace{\int \pi(\mathbf{Y}_{1:N} | \mathbf{X}_{1:N}, \boldsymbol{\theta}) \pi(\mathbf{X}_{1:N} | \boldsymbol{\theta}) d\mathbf{X}_{1:N}}_{\text{Likelihood}} \, \pi(\boldsymbol{\theta}) \ .$$

For a fixed $\boldsymbol{\theta}$ the state-space model is linear; thus the likelihood can be evaluated using Kalman recursions.
Therefore, the **online update** of $T_\sharp \rho \approx \pi(\boldsymbol{\theta} | \mathbf{Y}_{1:N})$ can be quickly performed exploiting the conditional independence of the model.

This work was supported by the US Department of Energy, Office of Advanced Scientific Computing (ASCR), under grant numbers DE-SC0010518 and DE-SC000929.7