Wasserstein Gradient Flow for Stochastic Prediction, Filtering, Learning and Control

Abhishek Halder

Department of Applied Mathematics University of California, Santa Cruz Santa Cruz, CA 95064

Joint work with Kenneth F. Caluya (UC Santa Cruz), and Tryphon T. Georgiou (UC Irvine)

Probabilistics Seminar Series, GE Research, May 27, 2020

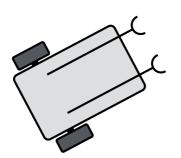


Overarching Theme

Systems-control theory for densities

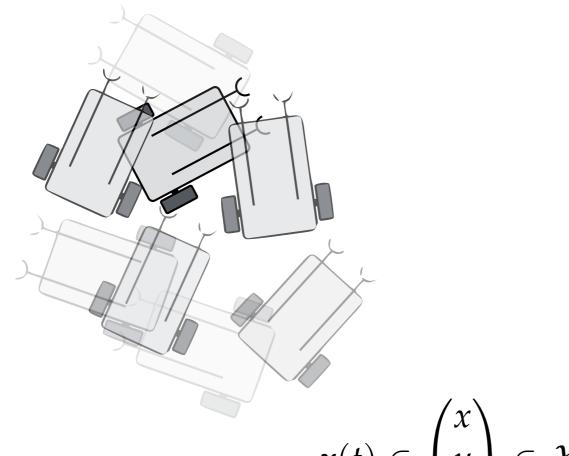
What is density?

Probability Density Fn.



$$\mathbf{x}(t) \in \begin{pmatrix} x \\ y \\ \theta \end{pmatrix} \in \mathcal{X} \equiv \mathbb{R}^2 \times \mathbb{S}^1$$

Probability Density Fn.



$$x(t) \in \begin{pmatrix} x \\ y \\ \theta \end{pmatrix} \in \mathcal{X} \equiv \mathbb{R}^2 \times \mathbb{S}^1$$

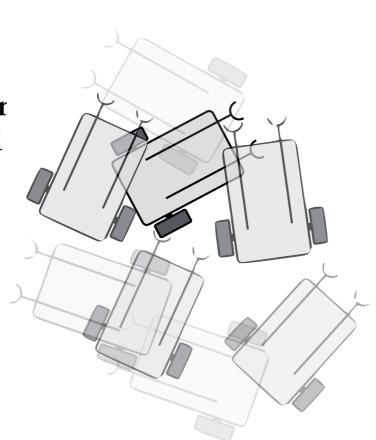
$$\rho\left(\mathbf{x},t\right):\mathcal{X}\times\left[0,\infty\right)\mapsto\mathbb{R}_{\geq0}$$

$$\int_{\mathcal{X}} \rho \, \mathrm{d}x = 1 \quad \text{ for all } t \in [0, \infty)$$

Probability Density Fn.

Population Density Fn.



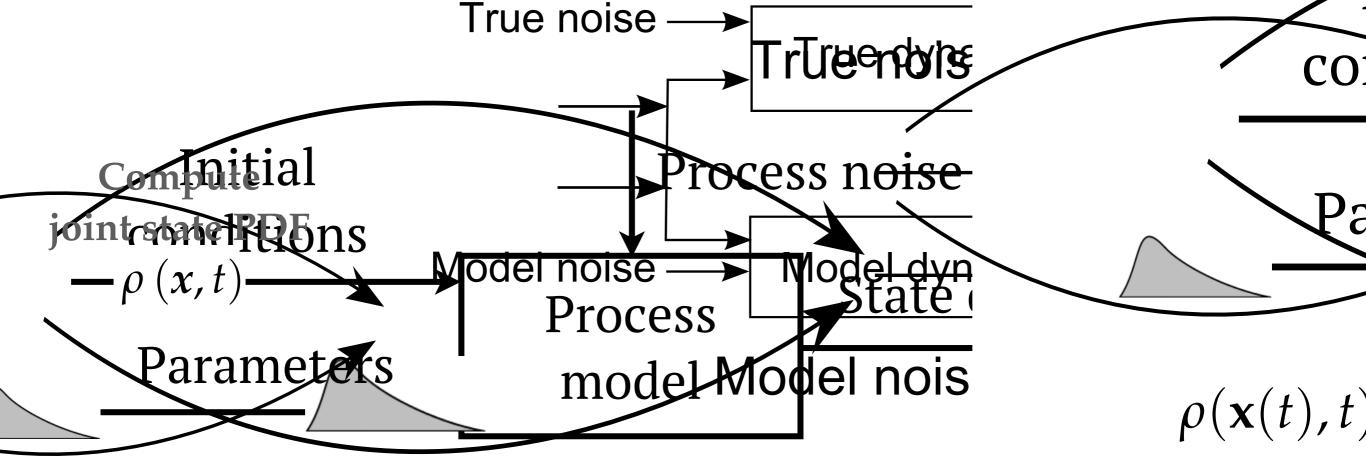


$$x(t) \in \begin{pmatrix} x \\ y \\ \theta \end{pmatrix} \in \mathcal{X} \equiv \mathbb{R}^2 \times \mathbb{S}^1$$

$$\rho\left(\mathbf{x},t\right):\mathcal{X}\times\left[0,\infty\right)\mapsto\mathbb{R}_{>0}$$

$$\int_{\mathcal{X}} \rho \, \mathrm{d}x = 1 \quad \text{ for all } t \in [0, \infty)$$

Why care about densities?

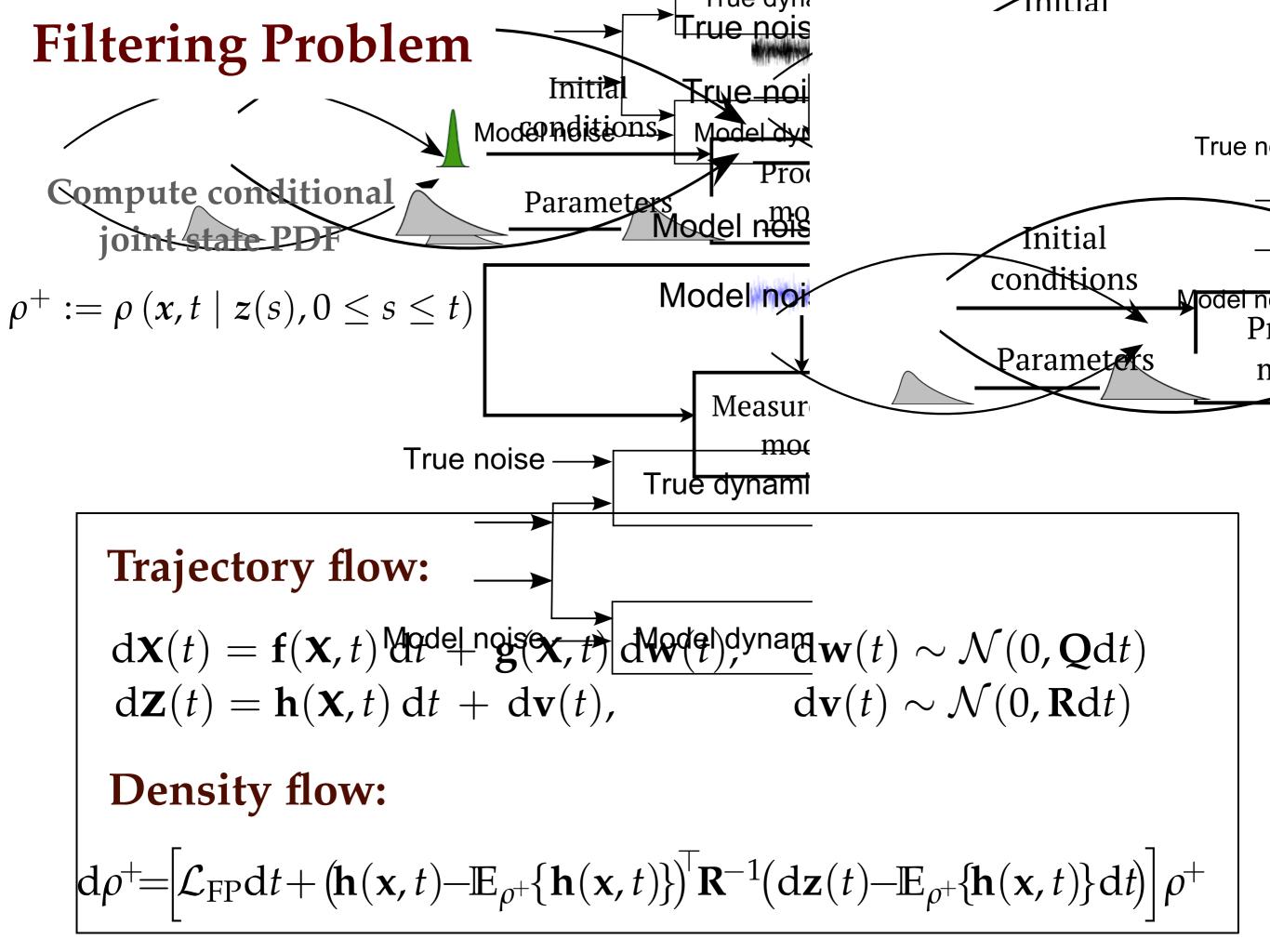


Trajectory flow:

$$d\mathbf{x}(t) = \mathbf{f}(\mathbf{x}, t) dt + \mathbf{g}(\mathbf{x}, t) d\mathbf{w}(t), \quad d\mathbf{w}(t) \sim \mathcal{N}(0, \mathbf{Q}dt)$$

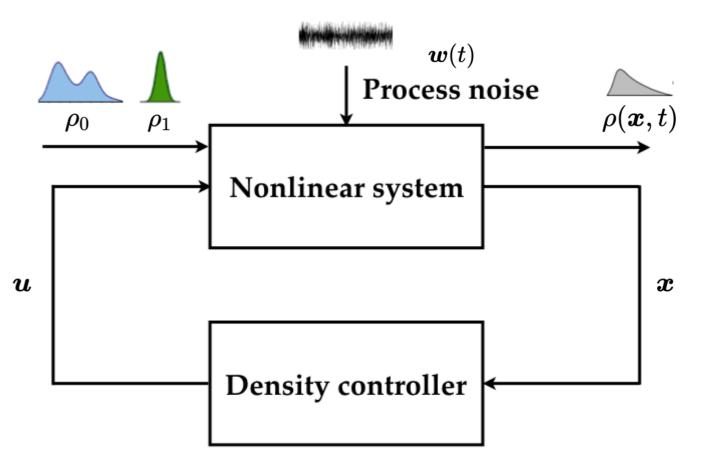
Density flow:

$$\frac{\partial \rho}{\partial t} = \mathcal{L}_{\text{FP}}(\rho) := -\nabla \cdot (\rho \mathbf{f}) + \frac{1}{2} \sum_{i,j=1}^{n} \frac{\partial^2}{\partial x_i \partial x_j} \left(\left(\mathbf{g} \mathbf{Q} \mathbf{g}^{\mathsf{T}} \right)_{ij} \rho \right)$$



Control Problem

Steer joint state PDF via feedback control over finite time horizon

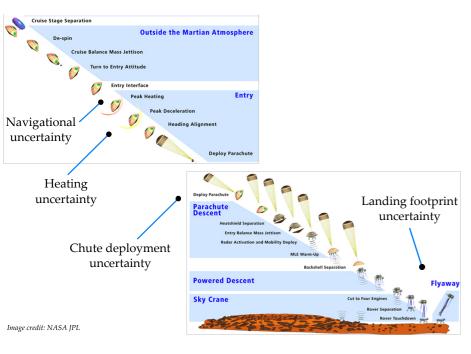


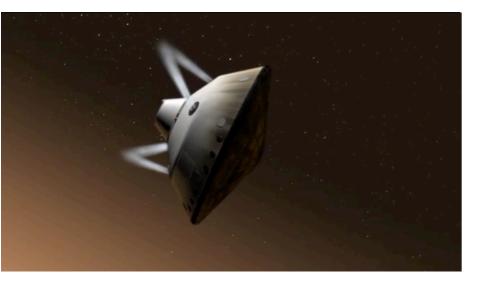
minimize
$$\mathbb{E}\left[\int_0^1 \|u\|_2^2 dt\right]$$
 subject to $dx = f(x, u, t) dt + g(x, t) dw$, $x(t = 0) \sim \rho_0$, $x(t = 1) \sim \rho_1$

PDFs in Mars Entry-Descent-Landing

Prediction Problem Filtering Problem

Control Problem





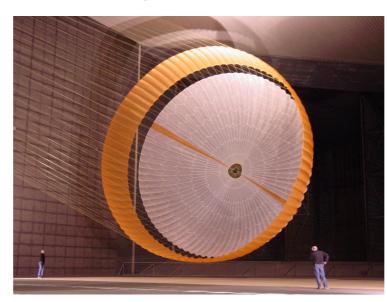
Predict heating rate uncertainty

PDFs in Mars Entry-Descent-Landing

Prediction Problem

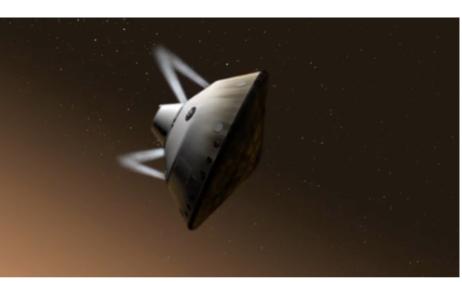
Cruise Balance Mass Jettison Turn to Entry Attitude Entry Interface Peak Heating Uncertainty Deploy Parachute Deploy Parachute Deploy Parachute Descent Uncertainty Chute deployment Uncertainty Flyaway Sky Crane Cut to Four Engines Rover Separation Rover Separation

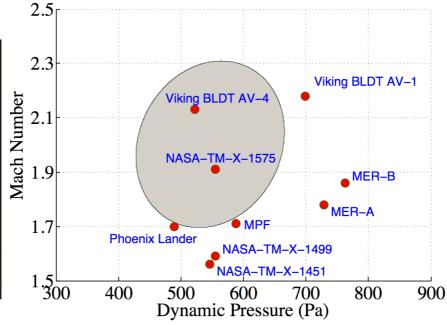
Filtering Problem



Control Problem

Supersonic parachute



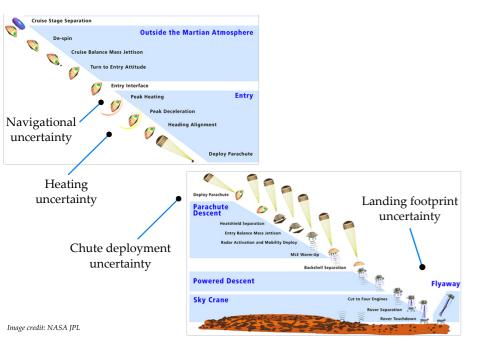


Predict heating rate uncertainty

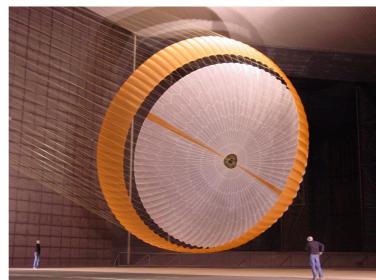
Estimate state to deploy parachute

PDFs in Mars Entry-Descent-Landing

Prediction Problem



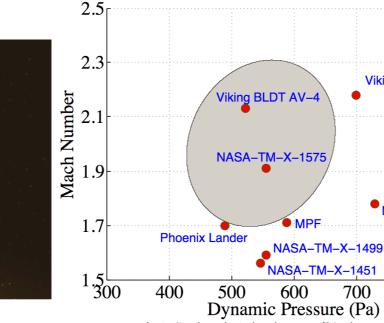
Filtering Problem



Supersonic parachute

Viking BLDT AV-1

800



deployment of a parachute has been a critical event in all Mars EDL deployment (shaqed region).

600

700

Predict heating rate uncertainty

Estimate state to deploy parachute

was minimizedight as 6 Avs. and additional sufetysonicas predsthear Smart & Moure Ligger of that point it be used madigated relied to the primary its. These arinners Decederated By steden RDS to be own in Disagraph is

Control Problem

higher Mach numbers result in increased aerothermal heating of parachute structure, which can reduce material strength; and (3) at Mach numbers above Mach 1.5, DGB parachutes exhibit an instability, known as areal oscillations, which result in multiple partial collapses and violent re-inflations. The chief concern with high Mach number deployments, for parachute deployments in regions where the heating is not a driving factor, is therefore, the increased exposure to area oscillations

The Viking parachute system was qualified to deploy between Mach 1.4 and 2.1, and a dynamic pressure between 250 and 700 Pa [1]. However, Mach 2.1 is not a hard limit for successfully operating DBG parachutes at Mars and there is very little flight test data above Mach 2.1 with which to quantify the amount of increased EDL system risk. Figure 3 shows the relevant flight tests and flight experience in the region of the planned MSh payashure deployte While payashute experts agree that higher Mach numbers result in a higher probability of failure, they have different opinions on where the limit should be braced. For example, Gillis [5] Has proposel xation per bound affe Mach 2 for paraghute aerodynamic decelerators

 $4.49^{o}S$ | $137.42^{o}E$ | This presents a challenge for EDL system designers, who must then weigh the system performance gains and risks associated with a paying the state of the system performance gains and risks associated with a paying the state of the system performance gains and risks associated with a paying the state of the system performance gains and risks associated with a paying the state of the system performance gains and risks associated with a paying the system performance gains and risks associated with a paying the system performance gains and risks associated with a paying the system performance gains and risks associated with a paying the system performance gains and risks associated with a paying the system performance gains and risks associated with a paying the system performance gains and risks associated with a paying the system performance gains and risks associated with a paying the system performance gains and risks associated with a paying the system performance gains and risks associated with a paying the system performance gains and risks associated with a paying the system performance gains and risks associated with a paying the system performance gains and risks associated with a paying the system performance gains and risks as a paying the system performance gains and risks as a paying the system performance gains and risks as a paying the system performance gains and risks as a paying the system performance gains and risks as a paying the system performance gains and risks as a paying the system performance gains and risks as a paying the system performance gains and risks as a paying the system performance gains and risks as a paying the system performance gains and risks as a paying the system performance gains and risks as a paying the system performance gains and risks as a paying the system performance gains and risks as a paying the system performance gains and risks as a paying the system performance gains and risks are paying the system performance gains and risks are paying the system performan

quantified, probability of parachute failure. It is clear that deploying a DGB at Mach 2.5 or 3.0 represents a significant pose of proposing and selecting possible landing sites. While increase in risk ever an initiation at Wach 2.0. However, it is many crites hwere initially incorposed at the first of those workshapsttheauteamearfiles Athlanding Site Workshop in 12008 was usef bis Mafs for Dr. can idea to the extremental largeable entain a sthe 900 four thin all stites with barrowal depending has objected in the state of the s at win45, that 1450/Lin which rise significantly ibe Wealth and inde capability of the system (estimated to be somewhere around

Steer state PDF to achieve desired landing footprint accuracy

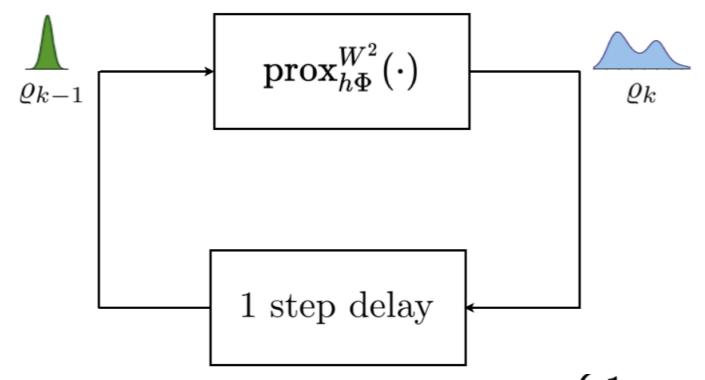
have had to rely on proxy measurements of other states in order to infer whether or not conditions were safe for deploying

Solving prediction problem as Wasserstein gradient flow

What's New?

Main idea: Solve
$$\frac{\partial \rho}{\partial t} = \mathcal{L}_{\mathrm{FP}} \rho, \; \rho(x,t=0) = \rho_0 \; \mathrm{as} \; \mathrm{gradient} \; \mathrm{flow} \; \mathrm{in} \; \mathcal{P}_2(\mathcal{X})$$

Infinite dimensional variational recursion:



 $\text{Proximal operator:} \ \ \varrho_k = \! \operatorname{prox}_{h\Phi}^{W^2}(\varrho_{k-1}) := \! \underset{\varrho \in \mathcal{P}_2(\mathcal{X})}{\operatorname{arg inf}} \bigg\{ \frac{1}{2} W^2(\varrho,\varrho_{k-1}) + h\Phi(\varrho) \bigg\}$

 $\textbf{Optimal transport cost:} \ W^2(\varrho,\varrho_{k-1}) := \inf_{\pi \in \Pi(\varrho,\varrho_{k-1})} \int_{\mathcal{X} \times \mathcal{X}} c(x,y) \ \mathrm{d}\pi(x,y)$

Free energy functional: $\Phi(\varrho) := \int_{\mathcal{X}} \psi \varrho \, \mathrm{d}x + \beta^{-1} \int_{\mathcal{X}} \varrho \log \varrho \, \mathrm{d}x$

Geometric Meaning of Gradient Flow

Gradient Flow in \mathcal{X}

$$\frac{\mathrm{d}\mathbf{x}}{\mathrm{d}t} = -\nabla \varphi(\mathbf{x}), \quad \mathbf{x}(0) = \mathbf{x}_0$$

Recursion:

$$\mathbf{x}_{k} = \mathbf{x}_{k-1} - h\nabla\varphi(\mathbf{x}_{k})$$

$$= \underset{\mathbf{x}\in\mathcal{X}}{\operatorname{arg min}} \left\{ \frac{1}{2} \|\mathbf{x} - \mathbf{x}_{k-1}\|_{2}^{2} + h\varphi(\mathbf{x}) \right\}$$

$$=: \operatorname{prox}_{h\varphi}^{\|\cdot\|_{2}}(\mathbf{x}_{k-1})$$

Convergence:

$$\mathbf{x}_k \to \mathbf{x}(t = kh)$$
 as $h \downarrow 0$

φ as Lyapunov function:

$$rac{\mathrm{d}}{\mathrm{d}t}arphi = -\parallel
abla arphi \parallel_2^2 \ \le \ 0$$

Gradient Flow in $\mathcal{P}_2(\mathcal{X})$

$$\frac{\partial \rho}{\partial t} = -\nabla^W \Phi(\rho), \quad \rho(\mathbf{x}, 0) = \rho_0$$

Recursion:

$$= \mathbf{x}_{k-1} - h\nabla\varphi(\mathbf{x}_{k})$$

$$= \underset{\mathbf{x}\in\mathcal{X}}{\operatorname{arg min}} \left\{ \frac{1}{2} \|\mathbf{x} - \mathbf{x}_{k-1}\|_{2}^{2} + h\varphi(\mathbf{x}) \right\}$$

$$= : \operatorname{prox}_{h\varphi}^{\|\cdot\|_{2}}(\mathbf{x}_{k-1})$$

$$= \operatorname{prox}_{h\varphi}^{W^{2}}(\mathbf{x}_{k-1})$$

$$= \operatorname{prox}_{h\varphi}^{W^{2}}(\rho_{k-1})$$

$$= : \operatorname{prox}_{h\varphi}^{W^{2}}(\rho_{k-1})$$

Convergence:

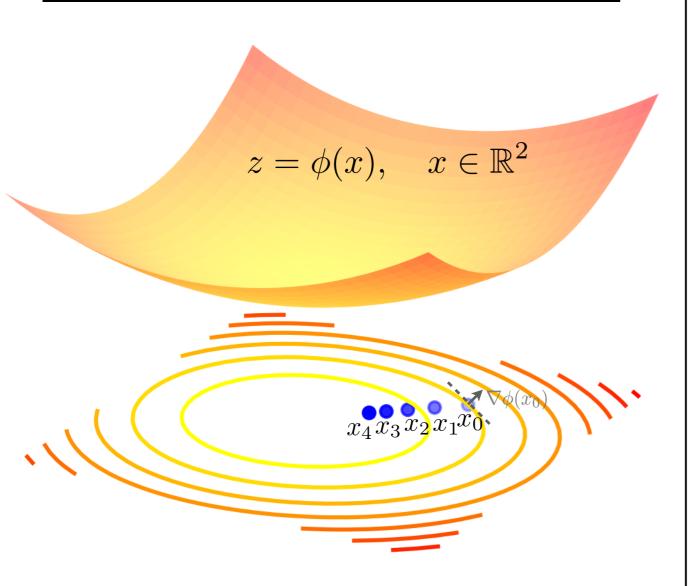
$$\rho_k \to \rho(\cdot, t = kh) \quad \text{as} \quad h \downarrow 0$$

Φ as Lyapunov functional:

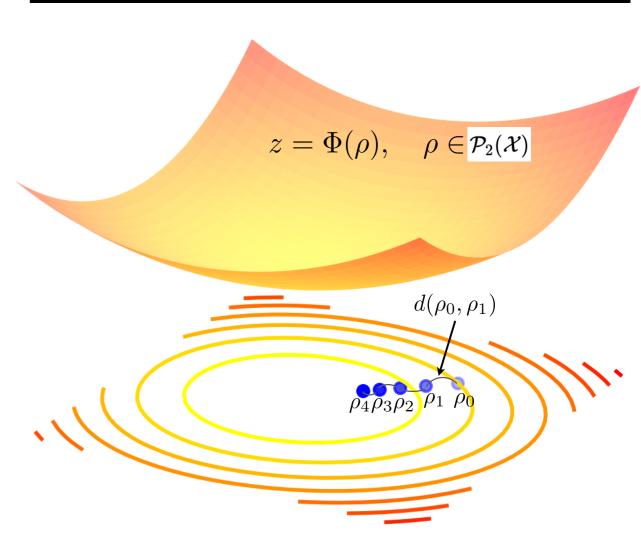
$$rac{\mathrm{d}}{\mathrm{d}t}\Phi = -\mathbb{E}_
hoigg[igg\|
ablarac{\delta\Phi}{\delta
ho}igg\|_2^2igg] ~\leq ~0$$

Geometric Meaning of Gradient Flow

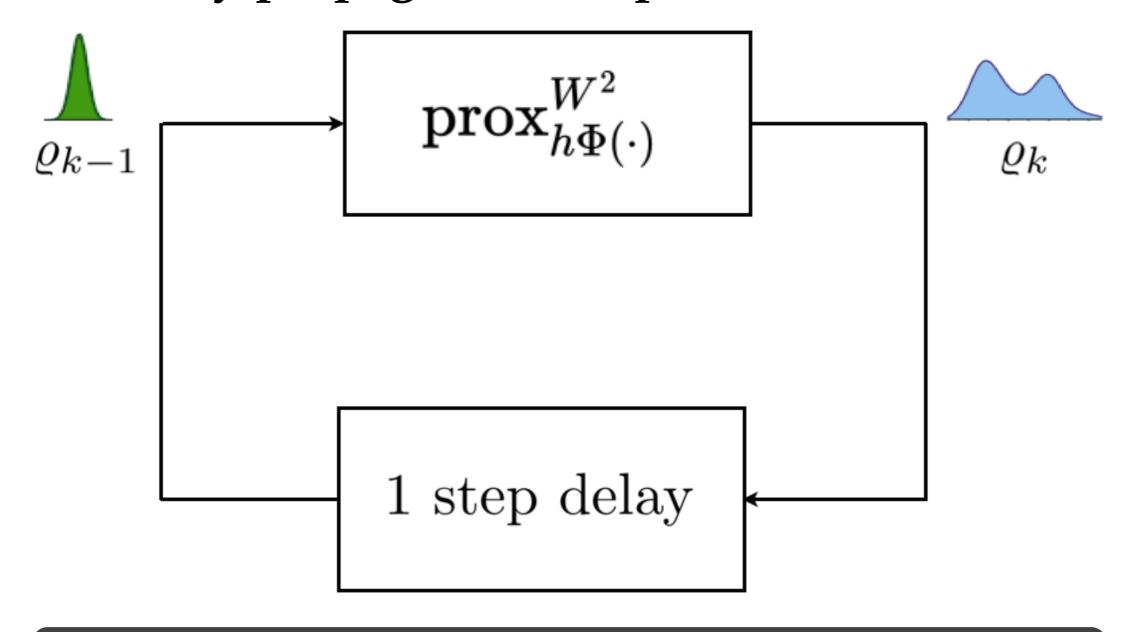
Gradient Flow in \mathcal{X}



Gradient Flow in $\mathcal{P}_2(\mathcal{X})$



Uncertainty propagation via point clouds



No spatial discretization or function approximation

$$\frac{\partial \rho}{\partial t} = \nabla \cdot (\nabla \psi \rho) + \beta^{-1} \Delta \rho$$

$$\updownarrow \quad \text{Proximal Recursion}$$

$$\rho_k = \rho(\mathbf{x}, t = kh) = \operatorname*{arg\ inf}_{\rho \in \mathcal{P}_2(\mathbb{R}^n)} \left\{ \frac{1}{2} W^2(\rho, \rho_{k-1}) + h \, \Phi(\rho) \right\}$$

$$\frac{\partial \rho}{\partial t} = \nabla \cdot (\nabla \psi \rho) + \beta^{-1} \Delta \rho$$

$$\updownarrow \quad \text{Proximal Recursion}$$

$$\rho_k = \rho(\mathbf{x}, t = kh) = \operatorname*{arg\ inf}_{\rho \in \mathcal{P}_2(\mathbb{R}^n)} \left\{ \frac{1}{2} W^2(\rho, \rho_{k-1}) + h \, \Phi(\rho) \right\}$$

$$\boldsymbol{\varrho}_{k} = \arg\min_{\boldsymbol{\varrho}} \left\{ \min_{\boldsymbol{M} \in \Pi(\boldsymbol{\varrho}_{k-1}, \boldsymbol{\varrho})} \frac{1}{2} \langle \boldsymbol{C}_{k}, \boldsymbol{M} \rangle + h \langle \boldsymbol{\psi}_{k-1} + \beta^{-1} \log \boldsymbol{\varrho}, \boldsymbol{\varrho} \rangle \right\}$$

$$\frac{\partial \rho}{\partial t} = \nabla \cdot (\nabla \psi \rho) + \beta^{-1} \Delta \rho$$

$$\updownarrow \quad \text{Proximal Recursion}$$

$$\rho_k = \rho(\mathbf{x}, t = kh) = \operatorname*{arg\ inf}_{\rho \in \mathcal{P}_2(\mathbb{R}^n)} \left\{ \frac{1}{2} W^2(\rho, \rho_{k-1}) + h \, \Phi(\rho) \right\}$$

Discrete Primal Formulation

$$\boldsymbol{\varrho}_{k} = \arg\min_{\boldsymbol{\varrho}} \left\{ \min_{\boldsymbol{M} \in \Pi(\boldsymbol{\varrho}_{k-1}, \boldsymbol{\varrho})} \frac{1}{2} \langle \boldsymbol{C}_{k}, \boldsymbol{M} \rangle + h \left\langle \boldsymbol{\psi}_{k-1} + \beta^{-1} \log \boldsymbol{\varrho}, \boldsymbol{\varrho} \right\rangle \right\}$$

↓ Entropic Regularization

$$\boldsymbol{\varrho}_{k} = \arg\min_{\boldsymbol{\varrho}} \left\{ \min_{\boldsymbol{M} \in \Pi(\boldsymbol{\varrho}_{k-1}, \boldsymbol{\varrho})} \frac{1}{2} \langle \boldsymbol{C}_{k}, \boldsymbol{M} \rangle + \epsilon H(\boldsymbol{M}) + h \langle \boldsymbol{\psi}_{k-1} + \beta^{-1} \log \boldsymbol{\varrho}, \boldsymbol{\varrho} \rangle \right\}$$

Dualization

$$m{\lambda}_0^{ ext{opt}}, m{\lambda}_1^{ ext{opt}} = rg \max_{m{\lambda}_0, m{\lambda}_1 \geq 0} \left\{ \langle m{\lambda}_0, m{arrho}_{k-1}
angle - F^\star(-m{\lambda}_1) - F^\star(-m{\lambda}_1) - rac{\epsilon}{h} \left(\exp(m{\lambda}_0^\top h/\epsilon) \exp(-m{C}_k/2\epsilon) \exp(m{\lambda}_1 h/\epsilon)
ight) \right\}$$

Recursion on the Cone

$$\mathbf{y} = e^{\frac{\lambda_0^*}{\epsilon}h} \qquad \qquad \mathbf{z} = e^{\frac{\lambda_1^*}{\epsilon}h}$$

Coupled Transcendental Equations in y and z

$$\Gamma_{k} = e^{\frac{-C_{k}}{2\epsilon}} \longrightarrow y \odot \Gamma_{k} z = \varrho_{k-1}$$

$$\varrho_{k-1} \longrightarrow \varrho_{k} = z \odot \Gamma_{k}^{\mathsf{T}} y$$

$$\xi_{k-1} = \frac{e^{-\beta \psi_{k-1}}}{e} \longrightarrow z \odot \Gamma_{k}^{\mathsf{T}} y = \xi_{k-1} z^{-\beta \epsilon/2h}$$

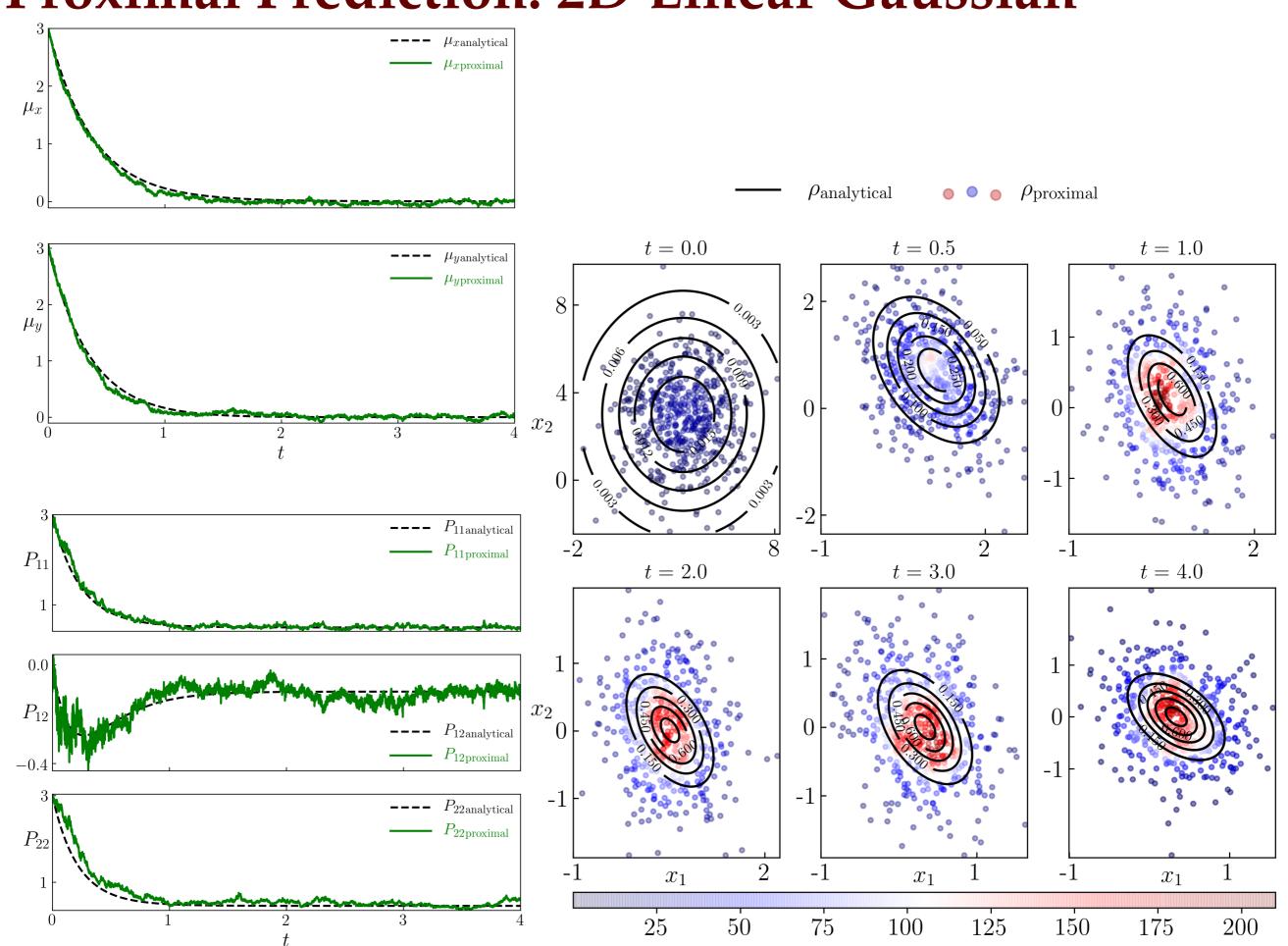
Theorem: Consider the recursion on the cone $\mathbb{R}^n_{\geq 0} \times \mathbb{R}^n_{\geq 0}$

$$oldsymbol{y}\odot(oldsymbol{\Gamma}_koldsymbol{z})=oldsymbol{arrho}_{k-1},\quadoldsymbol{z}\odot\left(oldsymbol{\Gamma}_k^{}^{}^{}oldsymbol{T}oldsymbol{y}
ight)=oldsymbol{\xi}_{k-1}\odotoldsymbol{z}^{-rac{eta\epsilon}{h}},$$

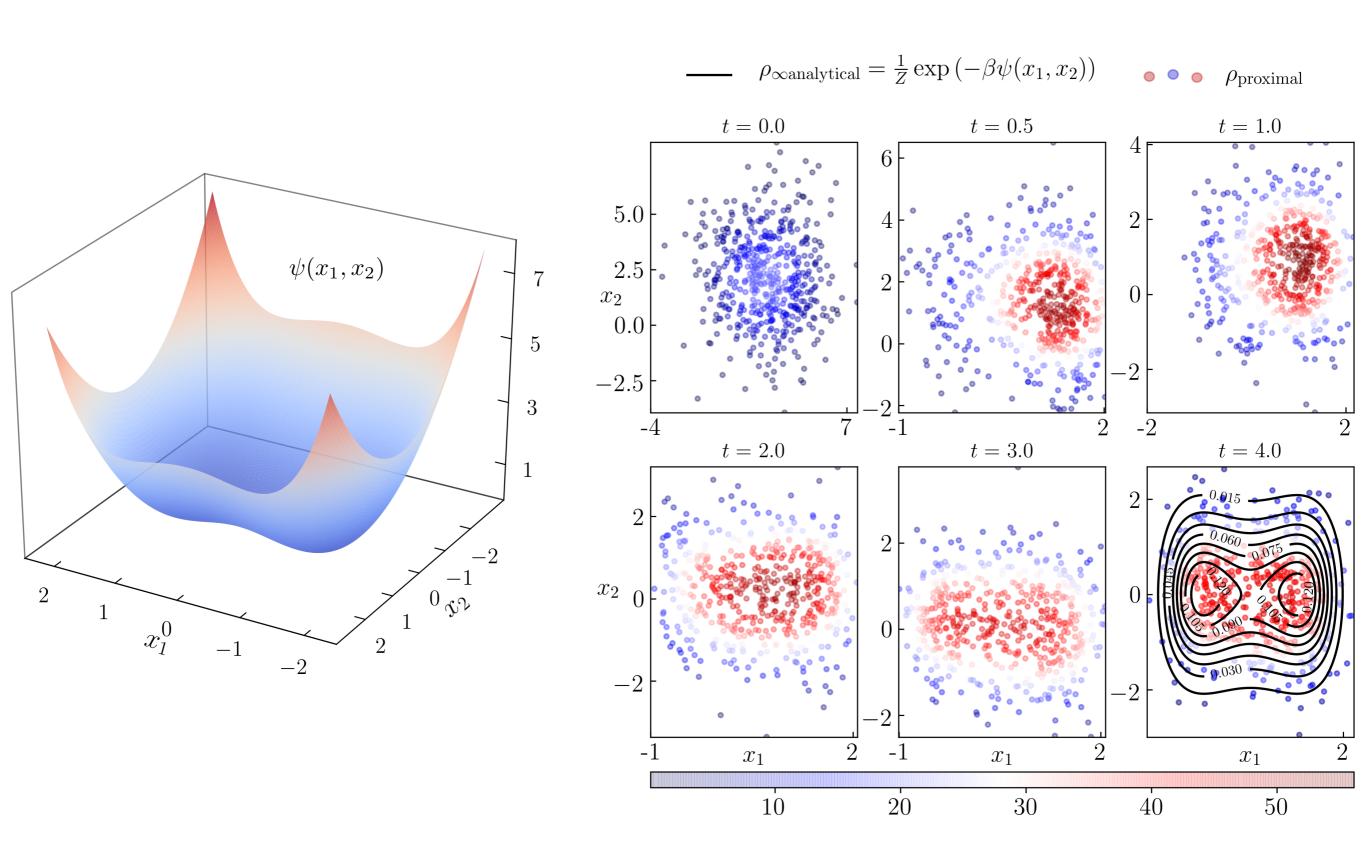
Then the solution $({m y}^*, {m z}^*)$ gives the proximal update ${m arrho}_k = {m z}^* \odot ({m \Gamma}_k^{-1} {m y}^*)$

The Kullback being the state of the state o pulme izapiomi zite naturbe the imizapione line rature. The trappole x_k x_k x_k given by x_k given by x_k x_k x $m_{k} = m_{k} = m_{k$ However, (11) is not all a year iter, sind a in the interior sin the light distript of second of the "proximal presented by and but be referred the of adors in equality the triangle in resplaced by a the proximal results in the proximal results in the proximal results in the 2. The 2-Waskerstein in the Bellines of the Belline $x_k = \max_{\overline{h}\varphi} (x_{k-1}), \quad x_k = \max_{\overline{h}\varphi} (x_{k-1}), \quad x_{k-1}, \quad x_{k-1}), \quad x_k = \max_{\overline{h}\varphi} (x_k) (x_$ Inverges that of the step-size how the sequence (4) with π_1 , the sequence π_1 , the sequence π_1 , the sequence π_1 , the sequence π_2 is the step-size π_2 the step-size π_3 the step-size π_4 in the step-size π_4 $\exp x_{h\Phi}^{d^2} (:_{\mathcal{P}_{k}} \operatorname{arg inf}_{\varrho \in \mathcal{D}_{2}} \stackrel{\text{loc}}{=} \frac{1}{2} d^2(\varrho, \varrho_{k-1}) := \underset{\ell}{\operatorname{arg inf}} \frac{1}{2} d^2(\varrho, \varrho_{k-1}) + \underset{\ell}{h} \Phi(\varrho),$ $W(\pi_1,\pi_2):=$ $W(\pi_1, \pi_2) :=$ as an infinite dimensional proximal operator. As mentioned inite $\int_{\mathbb{R}^{1}} \left(\inf_{\mathbf{z} \in \mathbf{y}} \left(\pi_{1}, \pi_{2} \right) \right) \mathbf{z}^{\frac{1}{2}} - \mathbf{y}$ $\text{where } \Pi\left(\pi_{1}, \pi_{2}\right) \text{ denotes the context}$ By icon (Re) recent to sequence satisfies $\mathcal{C}_{k}(x)$ i.e. the kh as the step-size where if (π_{1}, π_{2}) denotes the sollection of tall probabilities. quatirefresatisfies $h \to 0$ (we t_{also} which asather the privite dimensional case, he **Theorem:** Block co-ordinate iteration of (y, z) recur- $\frac{\mathrm{d}}{\mathrm{d}t}\varphi =$ fotow(sion is contractive on $\mathbb{R}^n_{>0} \times \mathbb{R}^n_{>0}$. nattoWp(2 pilitsip s a me 142 minus con grant the cold append appearing communications were

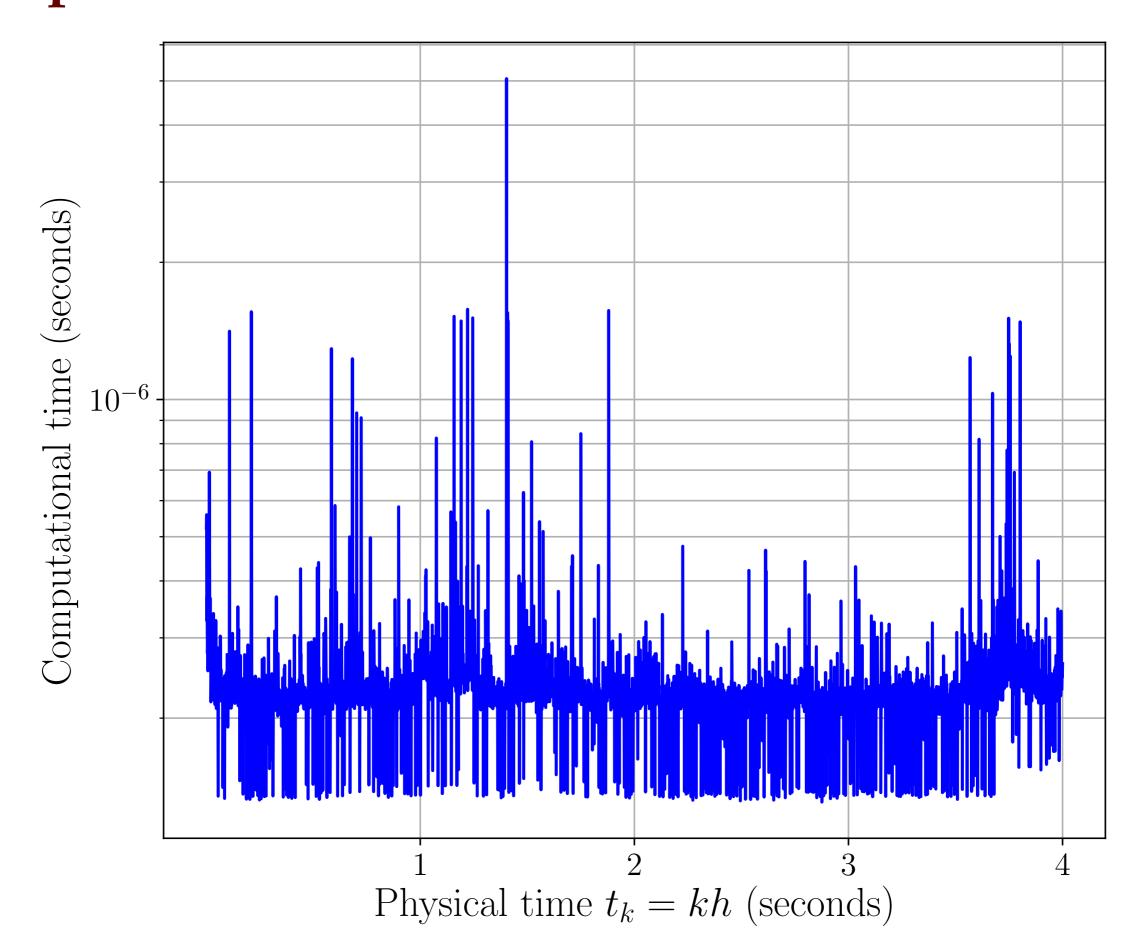
Proximal Prediction: 2D Linear Gaussian



Proximal Prediction: Nonlinear Non-Gaussian



Computational Time: Nonlinear Non-Gaussian



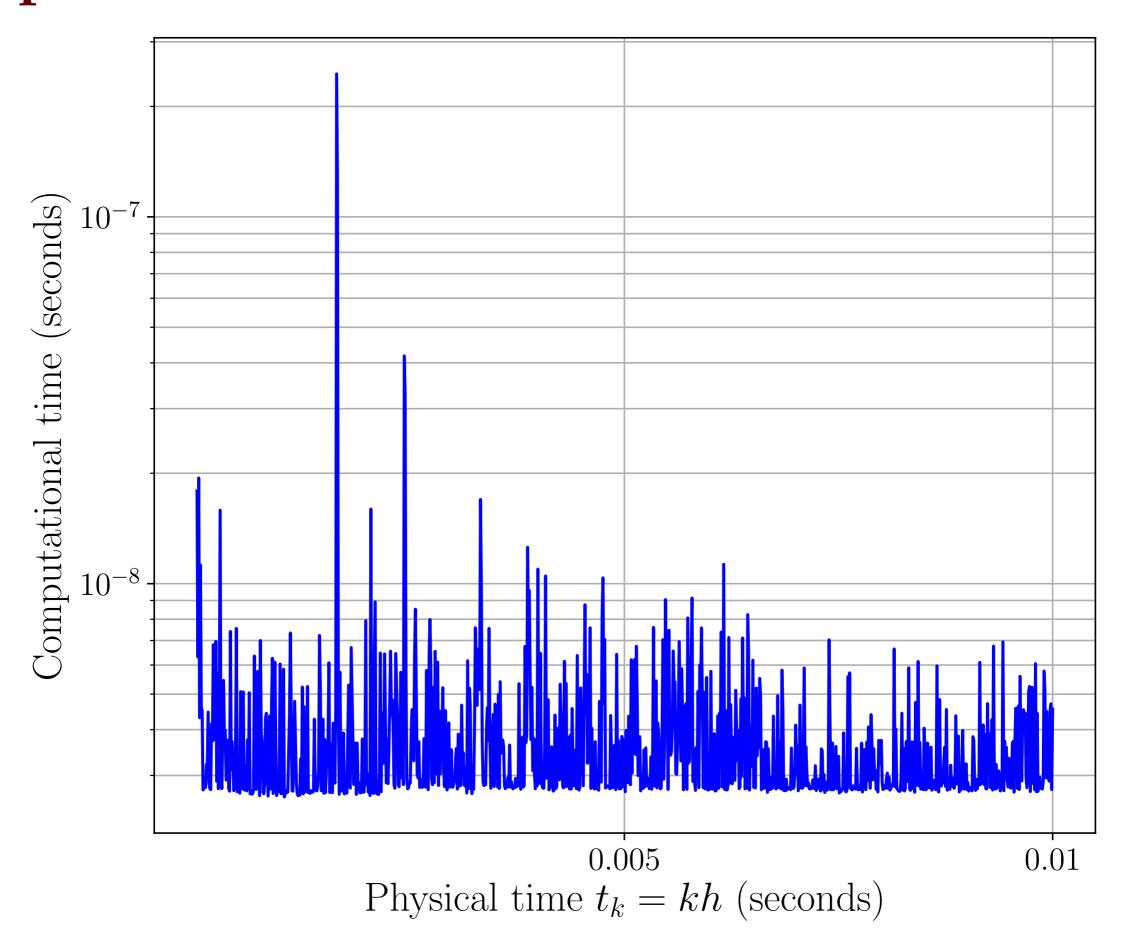
Proximal Prediction: Satellite in Geocentric Orbit

Here, $\mathcal{X} \equiv \mathbb{R}^6$

$$\begin{pmatrix} \mathrm{d}x \\ \mathrm{d}y \\ \mathrm{d}z \\ \mathrm{d}v_x \\ \mathrm{d}v_y \\ \mathrm{d}v_z \end{pmatrix} = \begin{pmatrix} v_x \\ v_y \\ -\frac{\mu x}{r^3} + (f_x)_{\mathsf{pert}} - \gamma v_x \\ -\frac{\mu y}{r^3} + (f_y)_{\mathsf{pert}} - \gamma v_y \\ -\frac{\mu z}{r^3} + (f_z)_{\mathsf{pert}} - \gamma v_z \end{pmatrix} dt + \sqrt{2\beta^{-1}\gamma} \begin{pmatrix} 0 \\ 0 \\ 0 \\ \mathrm{d}w_1 \\ \mathrm{d}w_2 \\ \mathrm{d}w_3 \end{pmatrix},$$

$$\begin{pmatrix} f_{\mathsf{x}} \\ f_{\mathsf{y}} \\ f_{\mathsf{z}} \end{pmatrix}_{\mathsf{pert}} = \begin{pmatrix} s\theta \ c\phi & c\theta \ c\phi & -s\phi \\ s\theta \ s\phi & c\theta \ s\phi & c\phi \\ c\theta & -s\theta & 0 \end{pmatrix} \begin{pmatrix} \frac{k}{2r^4} \left(3(s\theta)^2 - 1\right) \\ -\frac{k}{r^5}s\theta \ c\theta \\ 0 \end{pmatrix}, k := 3J_2R_{\mathrm{E}}^2, \mu = \mathsf{constant}$$

Computational Time: Satellite in Geocentric Orbit



Extensions: Nonlocal Interactions

PDF dependent sample path dynamics:

$$d\mathbf{x} = -\left(\nabla U\left(\mathbf{x}\right) + \nabla \rho * V\right) dt + \sqrt{2\beta^{-1}} d\mathbf{w}$$

Mckean-Vlasov-Fokker-Planck-Kolmogorov integro PDE:

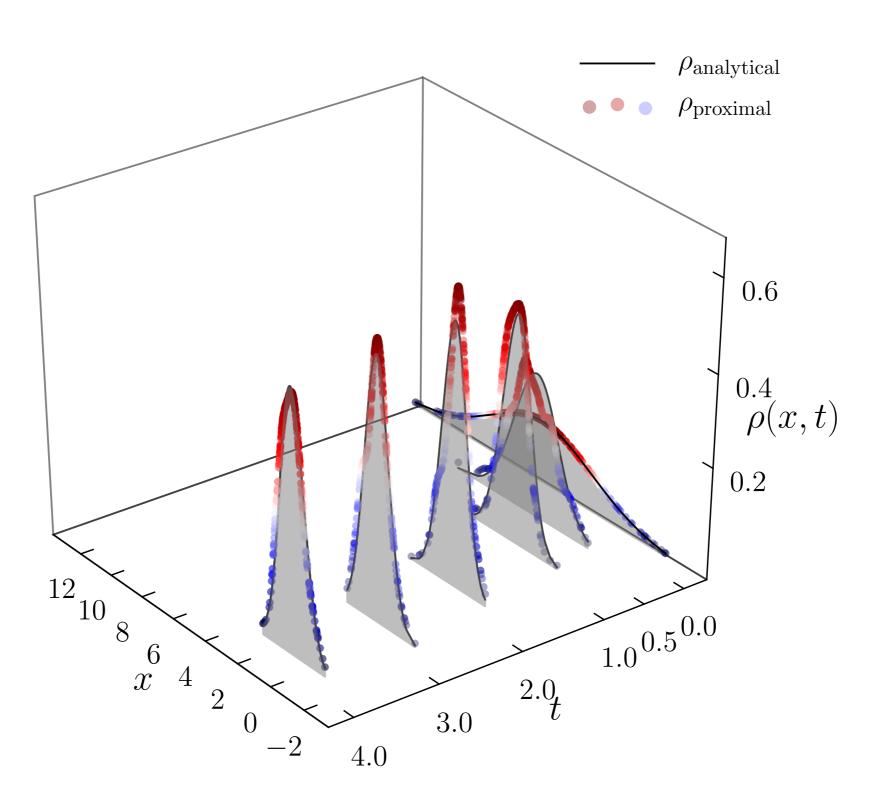
$$\frac{\partial \rho}{\partial t} = \nabla \cdot (\rho \nabla (U + \rho * V)) + \beta^{-1} \Delta \rho$$

Free energy:

$$F(\rho) := \mathbb{E}_{\rho} \left[U + \beta^{-1} \rho \log \rho + \rho * V \right]$$

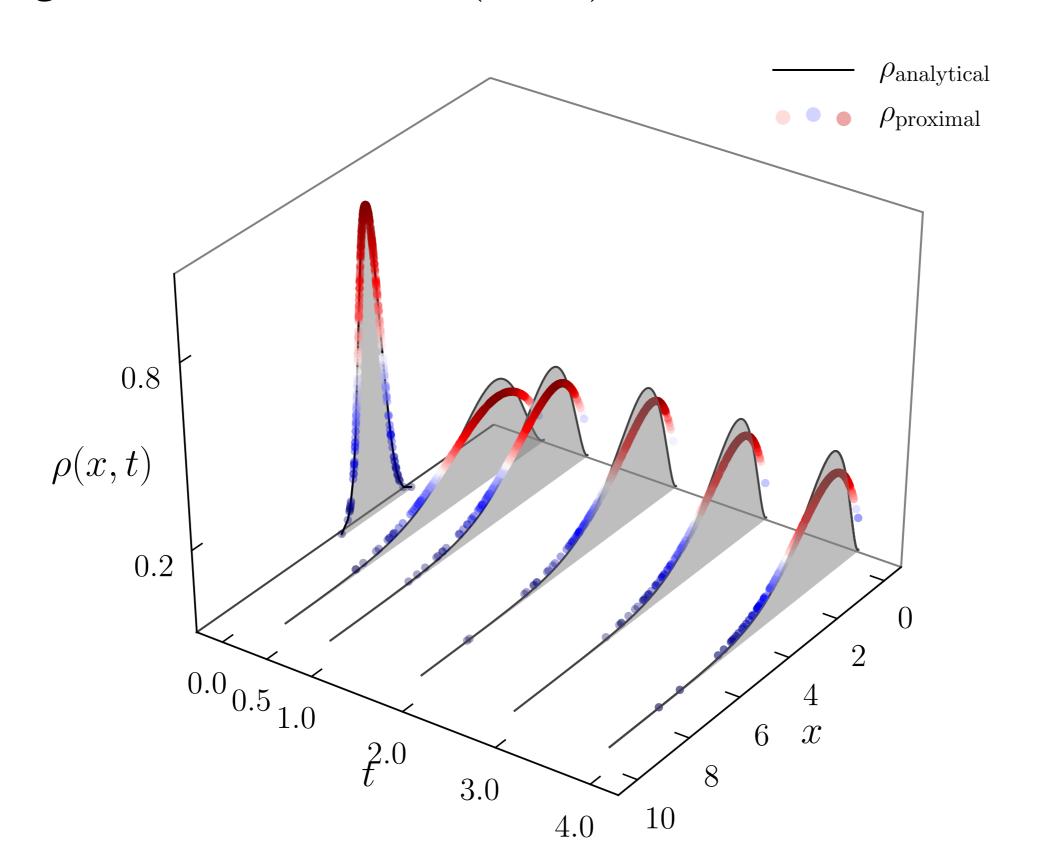
Extensions: Nonlocal Interactions

$$U(\cdot) = V(\cdot) = \|\cdot\|_2^2$$

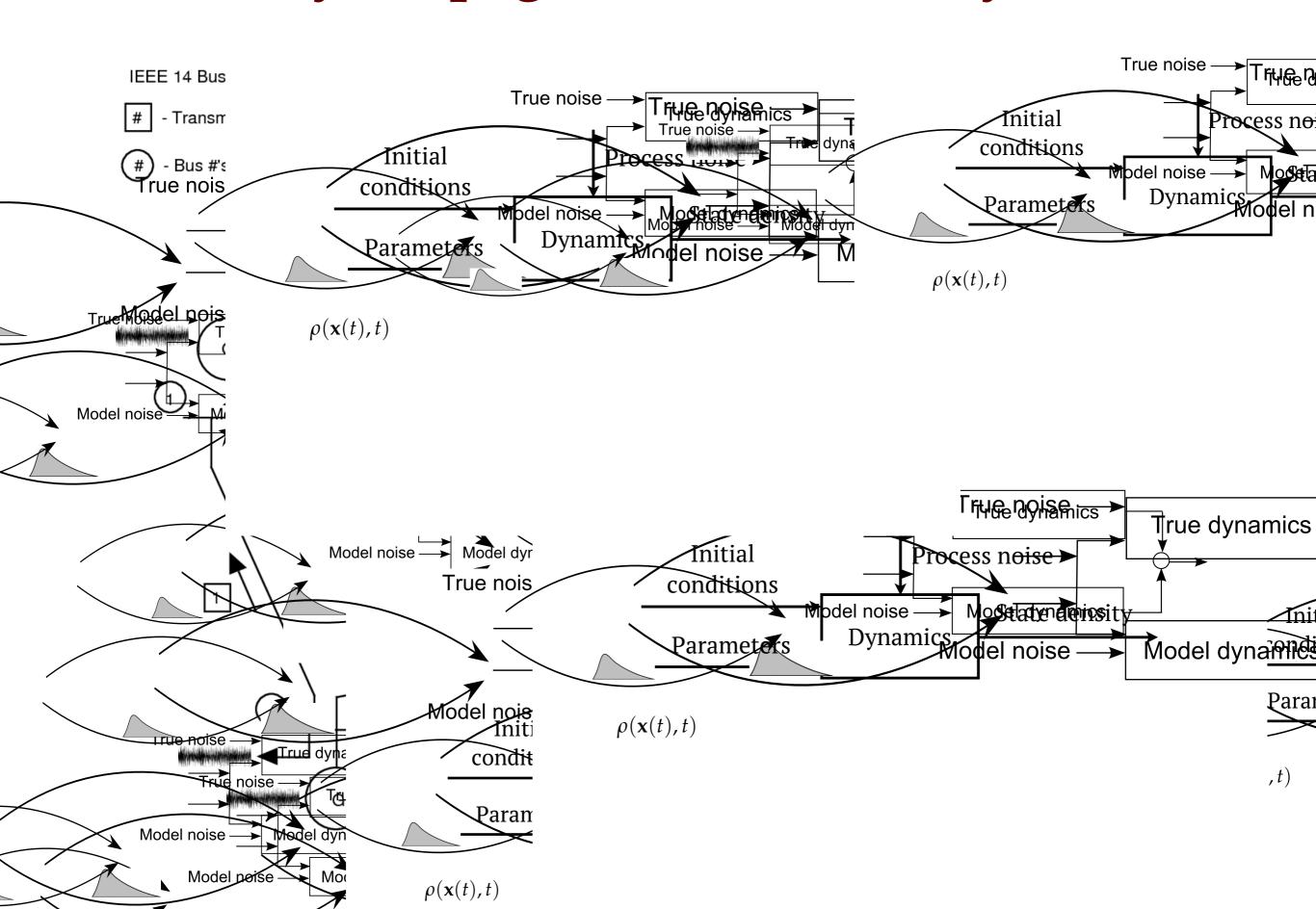


Extensions: Multiplicative Noise

Cox-Ingersoll-Ross: $dx = a(\theta - x) dt + b\sqrt{x} dw$, $2a > b^2$, $\theta > 0$



Uncertainty Propagation in Power Systems



Network Reduced Power System Model

Structure preserving power network model with noise

$$m_i\ddot{ heta}_i + \gamma_i\dot{ heta}_i = P_i^{ ext{mech}} - \sum_{j=1}^n k_{ij}\sin(heta_i - heta_j) + \sigma_i imes ext{stochastic forcing}, \; i = 1, \dots, n$$

Mixed Conservative-Dissipative SDE over state variables $(\theta, \omega) \in \mathbb{T}^n \times \mathbb{R}^n$

$$d\boldsymbol{\theta} = \boldsymbol{\omega} dt$$

 $d\boldsymbol{\omega} = (-(\boldsymbol{\gamma} \oslash \boldsymbol{m}) \odot \boldsymbol{\omega} - \nabla_{\boldsymbol{\theta}} V(\boldsymbol{\theta})) dt + (\boldsymbol{\sigma} \oslash \boldsymbol{m}) \odot d\boldsymbol{w}$

Potential function $V: \mathbb{T}^n \mapsto \mathbb{R}_{\geq 0}$

$$V(oldsymbol{ heta}) := \sum_{i=1}^n rac{1}{m_i} P_i^{\mathrm{mech}} heta_i + \sum_{(i,j) \in \mathcal{E}} rac{1}{m_i} k_{ij} (1 - \cos(heta_i - heta_j))$$

Proximal Recursion for Power System Model

Consider simple case: homogeneous generators with $\,\sigma^2=2eta^{-1}\gamma$

Lyapunov functional:

$$\Phi(
ho) = \int_{\mathbb{T}^n imes \mathbb{R}^n} igg(rac{1}{2} \parallel oldsymbol{\omega} \parallel_2^2 + V(oldsymbol{ heta})igg)
ho \, \mathrm{d}oldsymbol{ heta} \mathrm{d}oldsymbol{\omega} + eta^{-1} \int_{\mathbb{T}^n imes \mathbb{R}^n}
ho \log
ho \, \mathrm{d}oldsymbol{ heta} \mathrm{d}oldsymbol{\omega}$$

However, the FPK PDE is NOT a gradient descent of Φ w.r.t. W

Instead, do:
$$arrho_k = \mathrm{prox}_{h\gamma\widetilde{\Phi}}^{\widetilde{W}}(arrho_{k-1}), \quad k \in \mathbb{N},$$

$$\widetilde{\Phi}(
ho) = \int_{\mathbb{T}^n imes \mathbb{R}^n} rac{1}{2} \parallel oldsymbol{\omega} \parallel_2^2
ho \, \mathrm{d}oldsymbol{ heta} \mathrm{d}oldsymbol{\omega} + eta^{-1} \int_{\mathbb{T}^n imes \mathbb{R}^n}
ho \log
ho \, \mathrm{d}oldsymbol{ heta} \mathrm{d}oldsymbol{\omega}$$

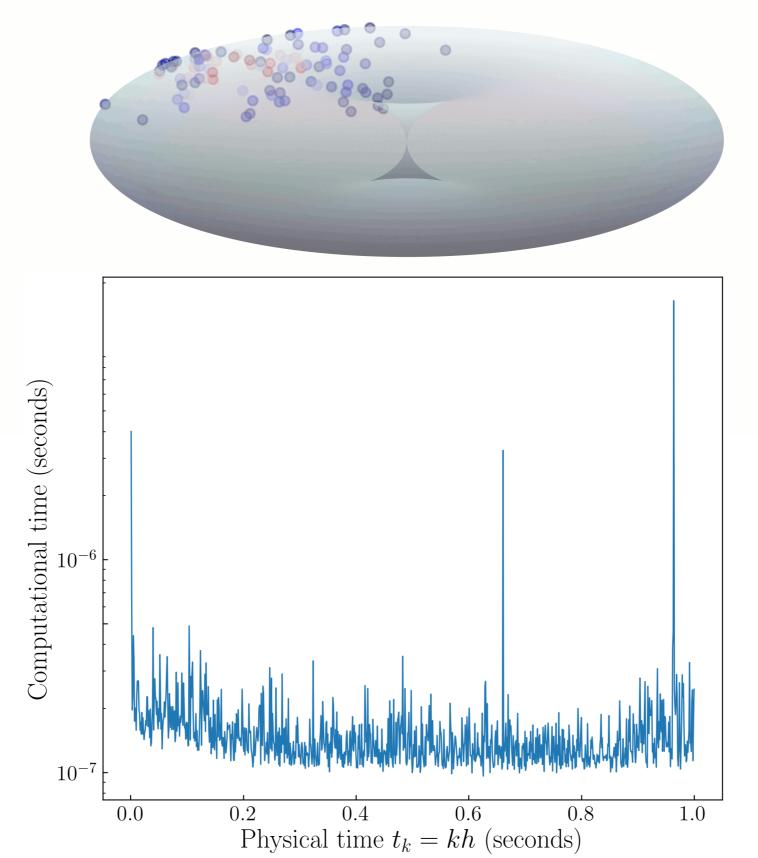
$$\widetilde{W}^2(arrho,arrho_{k-1}) = \inf_{\pi\in\Pi(arrho,arrho_{k-1})} \int_{\mathbb{T}^{2n} imes\mathbb{R}^{2n}} s_h(oldsymbol{ heta},oldsymbol{\omega},\overline{oldsymbol{ heta},\overline{oldsymbol{\omega}}}) \,\mathrm{d}\pi(oldsymbol{ heta},oldsymbol{\omega},\overline{oldsymbol{ heta},\overline{oldsymbol{\omega}}})$$

$$\mathbf{where} \ \ s_h(\boldsymbol{\theta}, \boldsymbol{\omega}, \overline{\boldsymbol{\theta}}, \overline{\boldsymbol{\omega}}) := \left\| \overline{\boldsymbol{\omega}} - \boldsymbol{\omega} + h \nabla V(\boldsymbol{\theta}) \right\|_2^2 \ + \ 12 \left\| \frac{\boldsymbol{\theta} - \boldsymbol{\theta}}{h} - \frac{\boldsymbol{\omega} + \boldsymbol{\omega}}{2} \right\|_2^2$$

Proximal Prediction: Power System with n = 2

Projection of the joint PDF on \mathbb{T}^2

$$t = 0.0000 s$$



Details on Proximal Prediction

Publications:

- K.F. Caluya, and A.H., Proximal Recursion for Solving the Fokker-Planck Equation, ACC 2019.
- K.F. Caluya, and A.H., Gradient Flow Algorithms for Density Propagation in Stochastic Systems, *IEEE Trans. Automatic Control* 2020, doi: 10.1109/TAC.2019.2951348.
- A.H., K.F. Caluya, B. Travacca, and S.J. Moura, Hopfield Neural Network Flow: A Geometric Viewpoint, *IEEE Trans. Neural Networks and Learning Systems* 2020, doi: 10.1109/TNNLS.2019.2958556.

Git repo: github.com/kcaluya/UncertaintyPropagation

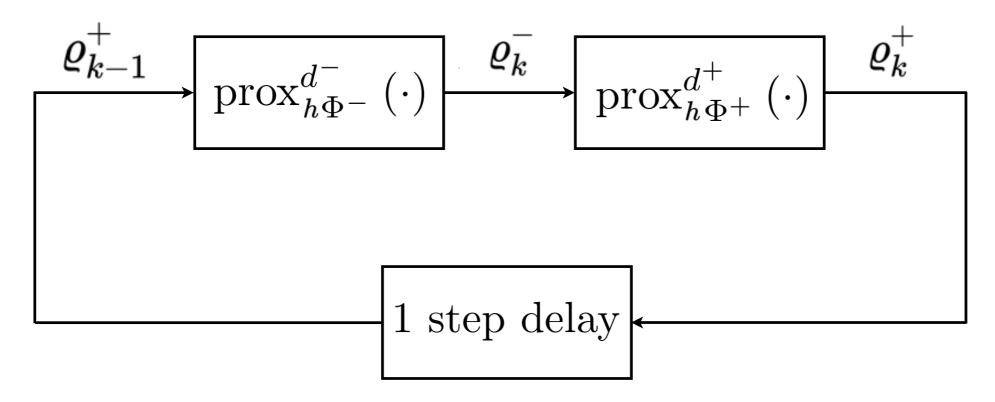
Solving filtering as Wasserstein gradient flow

What's New?

Main idea: Solve the Kushner-Stratonovich SPDE

$$\mathrm{d}
ho^+ = ig[\mathcal{L}_{\mathrm{FP}}\mathrm{d}t + \mathcal{L}ig(\mathrm{d}z,\mathrm{d}t,
ho^+ig)ig]
ho^+,\;
ho(x,t=0) =
ho_0 ext{ as gradient flow in } \mathcal{P}_2(\mathcal{X})$$

Recursion of {deterministic o stochastic} proximal operators:



Convergence: $\varrho_k^+(h) o
ho^+(x,t=kh)$ as $h\downarrow 0$

For prior, as before: $d^- \equiv W^2, \quad \Phi^- \equiv \ \mathbb{E}_{arrho} ig[\psi + eta^{-1} \log arrho ig]$

For posterior: $d^+ \equiv d_{ ext{FR}}^2 ext{ or } D_{ ext{KL}}, \quad \Phi^+ \equiv \; rac{1}{2} \mathbb{E}_{arrho^+} \Big[(y_k - h(x))^ op R^{-1} (y_k - h(x)) \Big]$

Explicit Recovery of the Kalman-Bucy Filter

Model:

$$d\mathbf{x}(t) = \mathbf{A}\mathbf{x}(t)dt + \mathbf{B}d\mathbf{w}(t), \quad d\mathbf{w}(t) \sim \mathcal{N}(0, \mathbf{Q}dt)$$
$$d\mathbf{z}(t) = \mathbf{C}\mathbf{x}(t)dt + d\mathbf{v}(t), \quad d\mathbf{v}(t) \sim \mathcal{N}(0, \mathbf{R}dt)$$

Given $\mathbf{x}(0) \sim \mathcal{N}(\mu_0, \mathbf{P}_0)$, want to recover:

$$\mathbf{P}^{+}\mathbf{C}\mathbf{R}^{-1}$$

$$\mathbf{d}\mu^{+}(t) = \mathbf{A}\mu^{+}(t)\mathbf{d}t + \mathbf{K}(t) \quad (\mathbf{d}\mathbf{z}(t) - \mathbf{C}\mu^{+}(t)\mathbf{d}t),$$

$$\dot{\mathbf{P}}^{+}(t) = \mathbf{A}\mathbf{P}^{+}(t) + \mathbf{P}^{+}(t)\mathbf{A}^{\top} + \mathbf{B}\mathbf{Q}\mathbf{B}^{\top} - \mathbf{K}(t)\mathbf{R}\mathbf{K}(t)^{\top}.$$

- A.H. and T.T. Georgiou, Gradient Flows in Uncertainty Propagation and Filtering of Linear Gaussian Systems, CDC 2017.
- A.H. and T.T. Georgiou, Gradient Flows in Filtering and Fisher-Rao Geometry, ACC 2018.

Explicit Recovery of the Wonham Filter

Model:

$$egin{aligned} x(t) \sim \operatorname{Markov}(Q), \ \operatorname{d}\!z(t) = h(x(t)) \operatorname{d}\!t \, + \, \sigma_v(t) \mathrm{d}v(t) \end{aligned}$$

State space: $\Omega := \{a_1, \ldots, a_m\}$

J.SIAM CONTROL Ser. A, Vol. 2, No. 3 Printed in U.S.A., 1965

SOME APPLICATIONS OF STOCHASTIC DIFFERENTIAL EQUATIONS TO OPTIMAL NONLINEAR FILTERING*

W. M. WONHAM†

Posterior $\pi^+(t) := \{\pi_1^+(t), \dots, \pi_m^+(t)\}$ solves the nonlinear SDE:

$$\mathrm{d}\pi^+(t) = \pi^+(t)Q\,\mathrm{d}t \;+\; rac{1}{\left(\sigma_v(t)
ight)^2}\pi^+(t)\Big(H-\widehat{h}(t)I\Big)\Big(\mathrm{d}z(t)-\widehat{h}(t)\mathrm{d}t\Big),$$

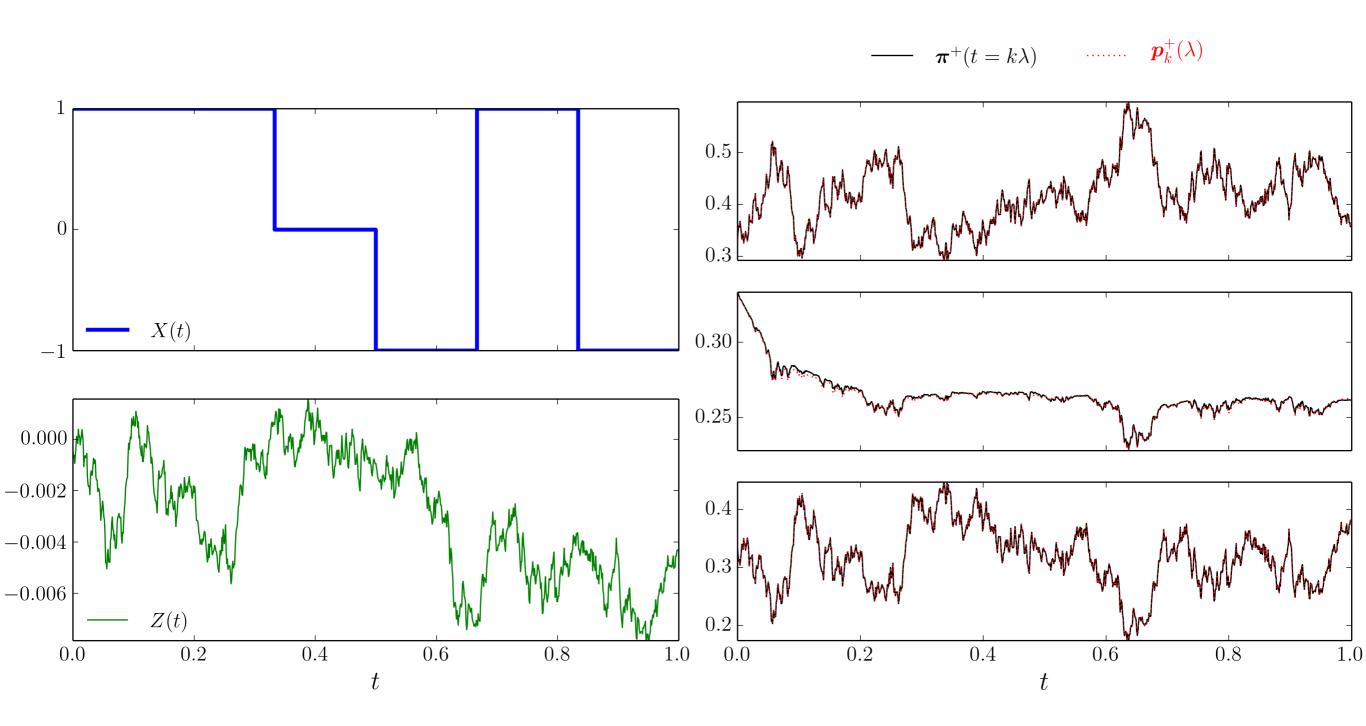
where
$$H:=\operatorname{diag}(h(a_1),\ldots,h(a_m)), \quad \widehat{h}(t):=\sum_{i=1}^m h(a_i)\pi_i^+(t),$$

Initial condition: $\pi^+(t=0)=\pi_0,$

By defn.
$$\pi^+(t)=\mathbb{P}(x(t)=a_i\mid z(s), 0\leq s\leq t)$$

— A.H. and T.T. Georgiou, Proximal Recursion for the Wonham Filter, CDC 2019.

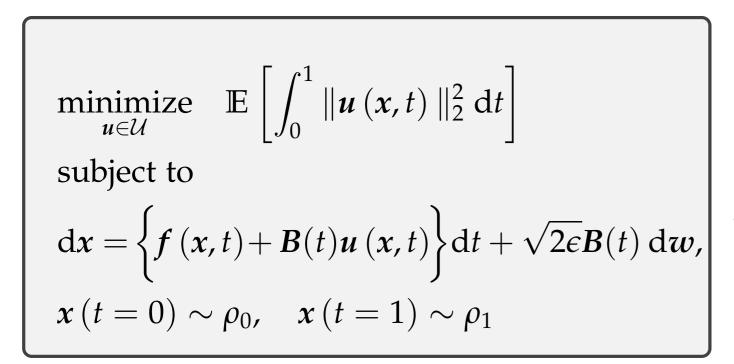
Numerical Results for the Wonham Filter

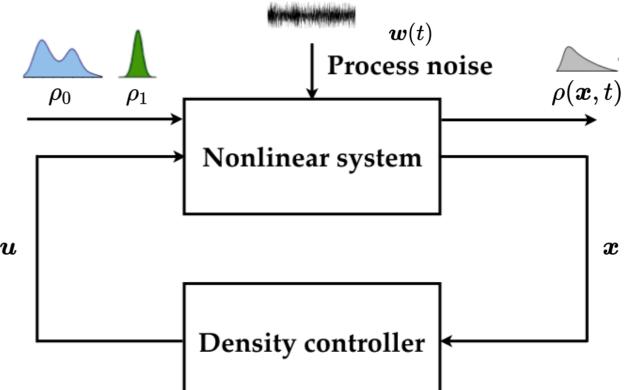


— A.H. and T.T. Georgiou, Proximal Recursion for the Wonham Filter, CDC 2019.

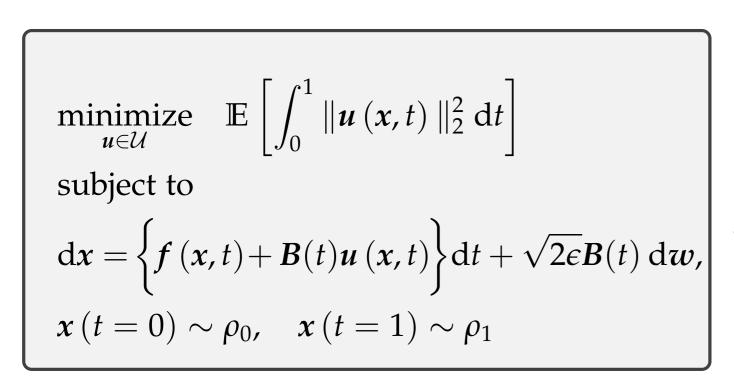
Solving density control as Wasserstein gradient flow

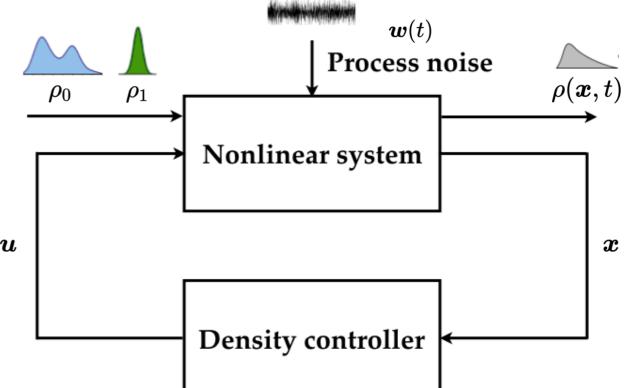
Finite Horizon Feedback Density Control





Finite Horizon Feedback Density Control





Necessary conditions for optimality: coupled nonlinear PDEs (FPK + HJB)

$$\frac{\partial \rho^{\text{opt}}}{\partial t} + \nabla \cdot \left(\rho^{\text{opt}} \left(f + \mathbf{B}(t)^{\mathsf{T}} \nabla \psi \right) \right) = \epsilon \mathbf{1}^{\mathsf{T}} \left(\mathbf{D}(t) \odot \text{Hess} \left(\rho^{\text{opt}} \right) \right) \mathbf{1},$$

$$\frac{\partial \psi}{\partial t} + \frac{1}{2} \| \boldsymbol{B}(t)^{\top} \nabla \psi \|_{2}^{2} + \langle \nabla \psi, \boldsymbol{f} \rangle = -\epsilon \langle \boldsymbol{D}(t), \operatorname{Hess}(\psi) \rangle$$

Boundary conditions:

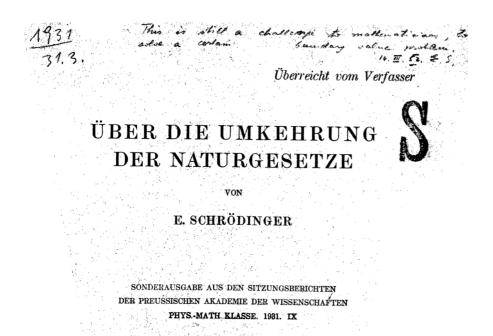
$$\rho^{\text{opt}}(x,0) = \rho_0(x), \quad \rho^{\text{opt}}(x,1) = \rho_1(x)$$

Optimal control:

$$u^{\mathrm{opt}}(x,t) = B(t)^{\mathsf{T}} \nabla \psi$$

Feedback Synthesis via the Schrödinger System

Schrödinger's (until recently) forgotten papers:



Sur la théorie relativiste de l'électron et l'interprétation de la mécanique quantique

E. SCHRÖDINGER

I. — Introduction

J'ai l'intention d'exposer dans ces conférences diverses idées concernant la mécanique quantique et l'interprétation qu'on en donne généralement à l'heure actuelle ; je parlerai principalement de la théorie quantique relativiste du mouvement de l'électron. Autant que nous pouvons nous en rendre compte aujourd'hui, il semble à peu près sûr que la mécanique quantique de l'électron, sous sa forme idéale, que nous ne possédons pas encore, doit former un jour la base de toute la physique. A cet intérêt tout à fait général, s'ajoute, ici à Paris, un intérêt particulier : vous savez tous que les bases de la théorie moderne de l'électron ont été posées à Paris par votre célèbre compatriote Louis de BROGLIE.



Hopf-Cole transform: $(\rho^{\text{opt}}, \psi) \mapsto (\varphi, \hat{\varphi})$

$$\varphi(x,t) = \exp\left(\frac{\psi(x,t)}{2\epsilon}\right),$$

$$\hat{\varphi}(x,t) = \rho^{\text{opt}}(x,t) \exp\left(-\frac{\psi(x,t)}{2\epsilon}\right),$$

Optimal controlled joint state PDF: $\rho^{\text{opt}}(x,t) = \hat{\varphi}(x,t)\varphi(x,t)$

Optimal control: $u^{\text{opt}}(x,t) = 2\epsilon B(t)^{\top} \nabla \log \varphi(x,t)$

Feedback Synthesis via the Schrödinger System

2 coupled nonlinear PDEs → boundary-coupled linear PDEs!!

$$\frac{\partial \hat{\varphi}}{\partial t} = -\nabla \cdot (\hat{\varphi}f) + \epsilon \mathbf{1}^{\top} (\mathbf{D}(t) \odot \operatorname{Hess}(\hat{\varphi})) \mathbf{1}, \ \varphi_0 \hat{\varphi}_0 = \rho_0,$$
 forward Kolmogorov PDE
$$\frac{\partial \varphi}{\partial t} = -\langle \nabla \varphi, f \rangle - \epsilon \langle \mathbf{D}(t), \operatorname{Hess}(\varphi) \rangle, \qquad \varphi_1 \hat{\varphi}_1 = \rho_1.$$
 backward Kolmogorov PDE

Wasserstein proximal algorithm \longrightarrow fixed point recursion over $(\hat{\varphi}_0, \varphi_1)$

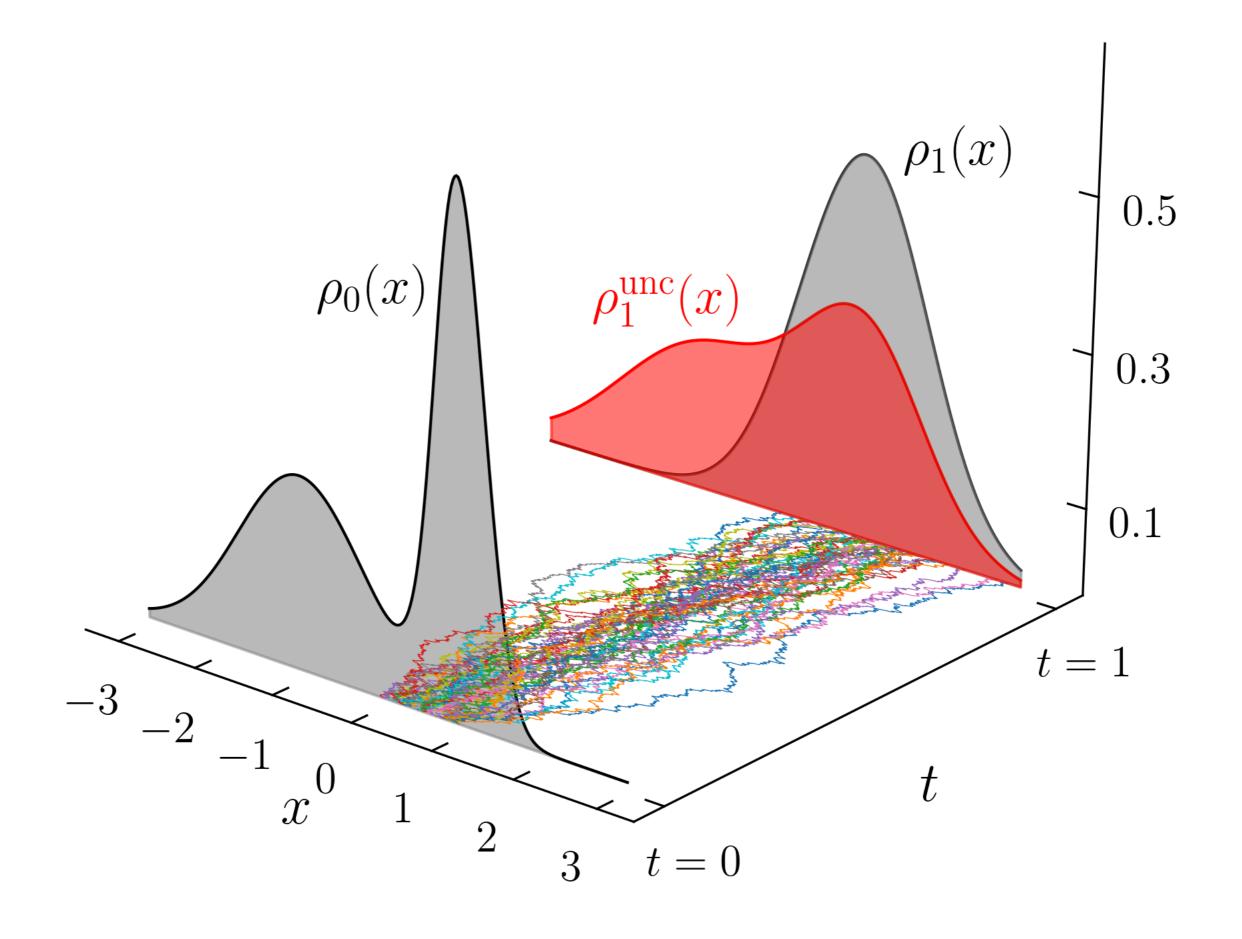
(Contractive in Hilbert metric)

— Y. Chen, T.T. Georgiou, and M. Pavon, Entropic and displacement interpolation: a computational approach using the Hilbert metric, *SIAM J. Applied Mathematics*, 2016.

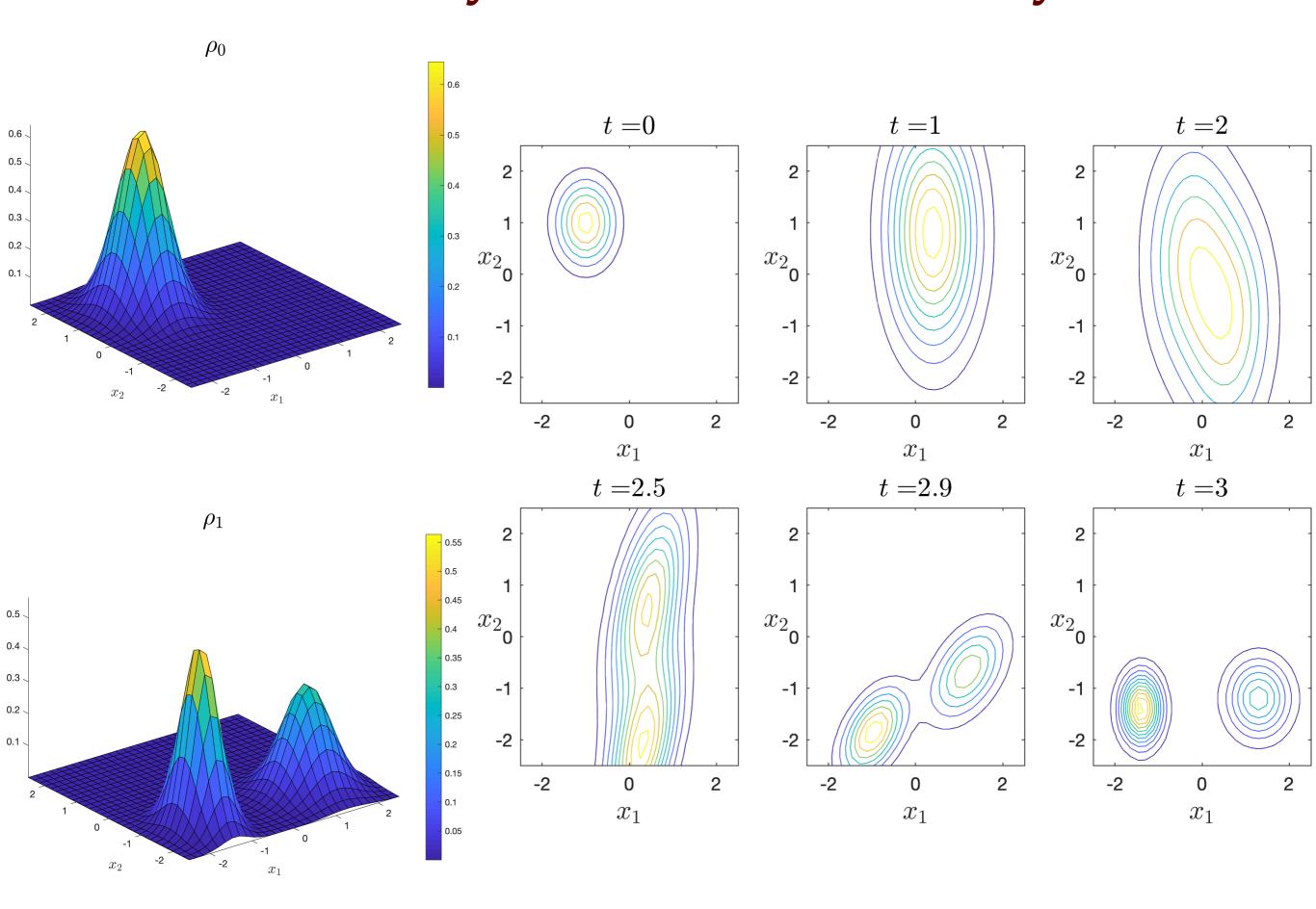
Fixed Point Recursion over $(\hat{\varphi}_0, \varphi_1)$

$$\begin{array}{cccc}
& & & \int & \\
\hat{\varphi}_0(x) & \longrightarrow & \hat{\varphi}_1(x) \\
& & \downarrow & \\
\rho_0(x)/\hat{\varphi}_1(x) & & \downarrow \\
\varphi_0(x) & \longleftarrow & \varphi_1(x)
\end{array}$$

Feedback Density Control: Zero Prior Dynamics

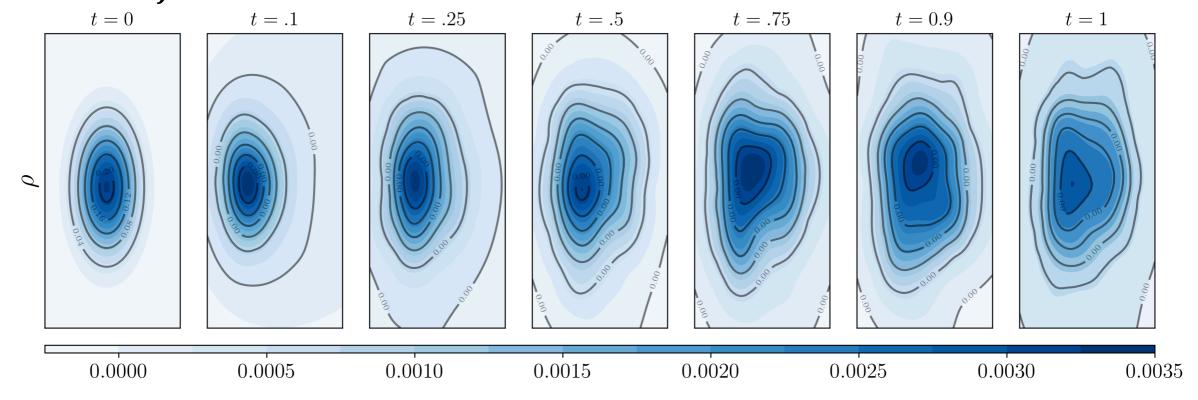


Feedback Density Control: LTI Prior Dynamics

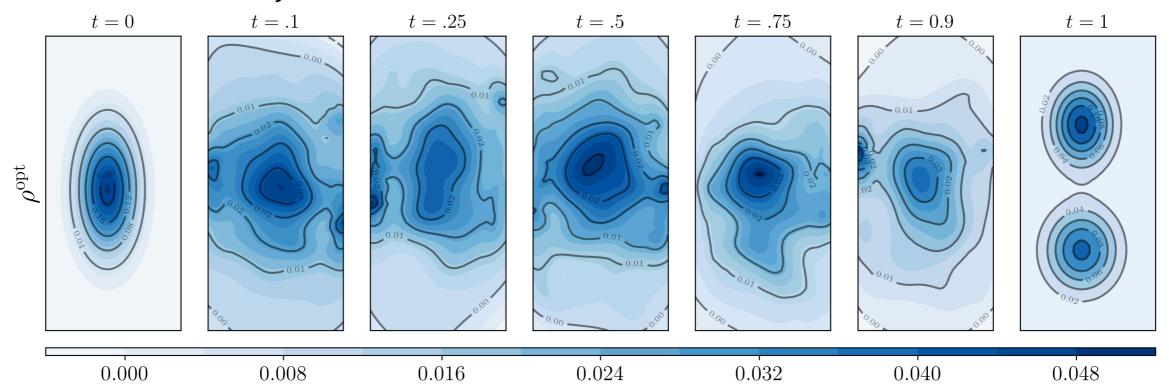


Feedback Density Control: Nonlinear Grad. Drift

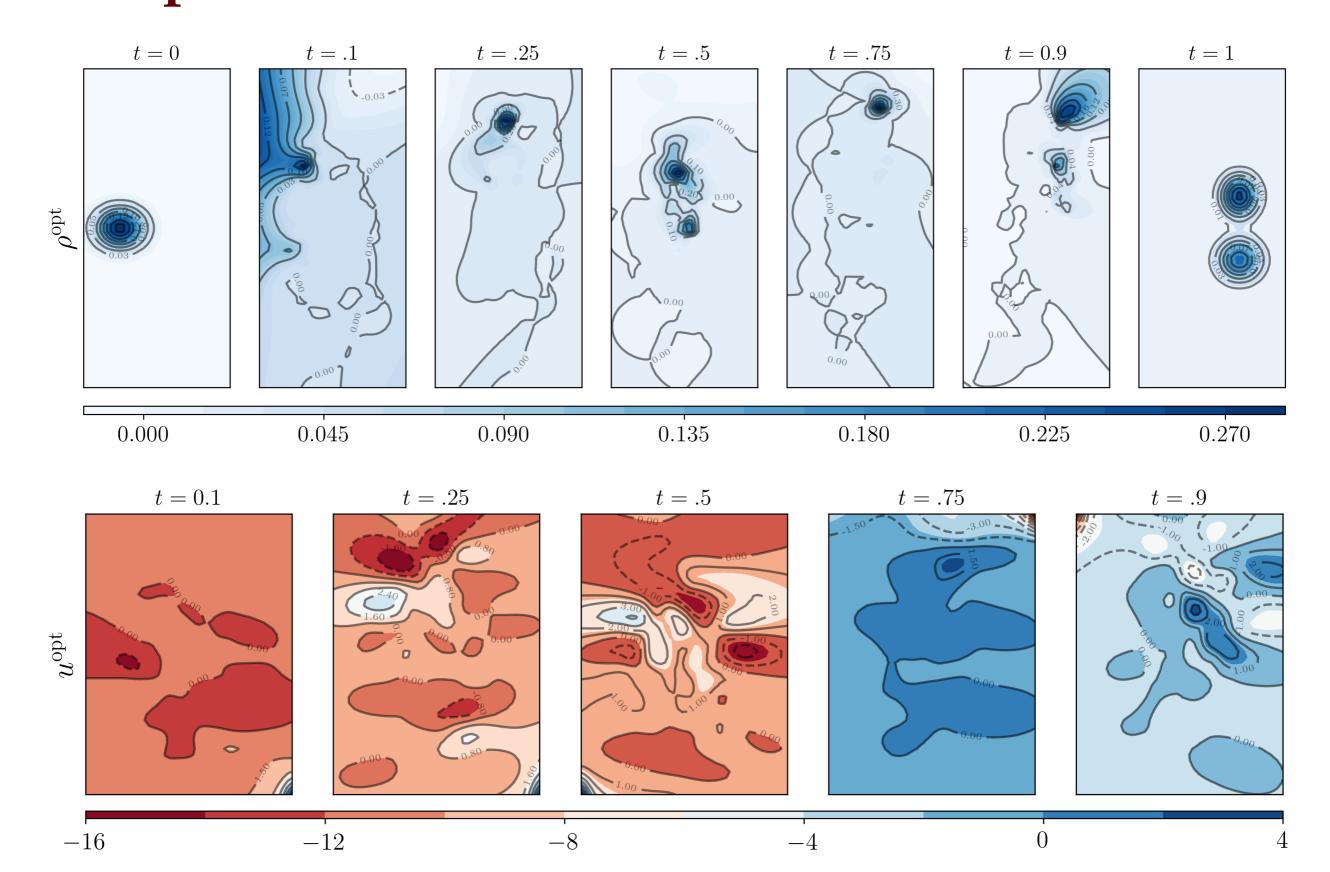
Uncontrolled joint PDF evolution:



Optimal controlled joint PDF evolution:



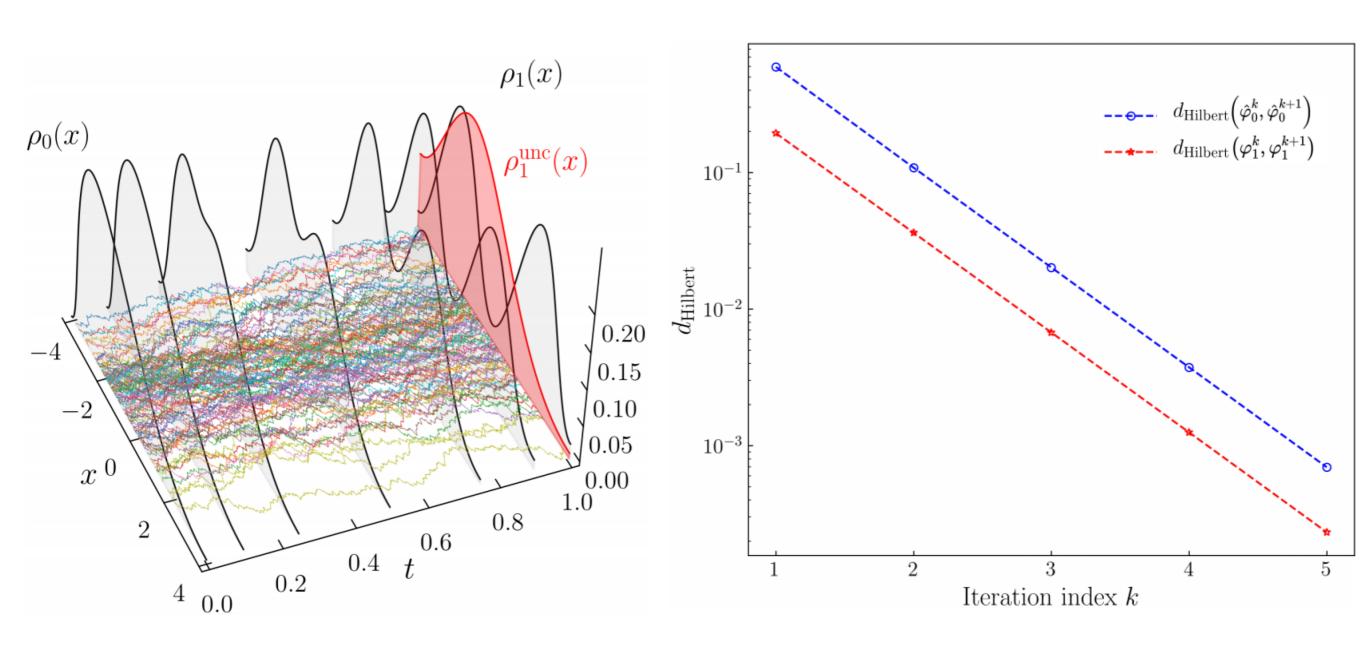
Feedback Density Control: Mixed Conservative-Dissipative Drift



Density Control with Det. Path Constraints

Reflecting Schrödinger Bridge

Contraction in the Hilbert metric



Details on Density Control

Publications:

- A.H., and E.D.B. Wendel, Finite horizon linear quadratic Gaussian density regulator with Wasserstein terminal cost, *ACC* 2016.
- K.F. Caluya, and A.H., Wasserstein proximal algorithms for the Schrödinger Bridge Problem: density control with nonlinear drift, *IEEE Trans. Automatic Control*, under review, 2019.
- K.F. Caluya, and A.H., Finite horizon density control for static state feedback linearizable systems, *IEEE Trans. Automatic Control*, in revision, 2020.
- K.F. Caluya, and A.H., Finite Horizon Density Steering for Multi-input State Feedback Linearizable Systems, *ACC* 2020.
- K.F. Caluya, and A.H., Reflected Schrödinger bridge: density control with path constraints, CDC 2020.

Learning a neural network as Wasserstein gradient flow

Learning Neural Network from Data

$$(ext{feature vector, label}) = (oldsymbol{x}_i, y_i) \in \mathbb{R}^d imes \mathbb{R}, \quad i = 1, \dots, n$$

Consider shallow NN: 1 hidden layer with $n_{\rm H}$ neurons

NN parameter vector
$$\boldsymbol{\theta} := (\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \dots, \boldsymbol{\theta}_{n_{\mathrm{H}}})^{\top} \in \mathbb{R}^{pn_{\mathrm{H}}}$$

Approximating function:

$$\hat{f}\left(oldsymbol{x},oldsymbol{ heta}
ight) = rac{1}{n_{ ext{H}}} \sum_{i=1}^{n_{ ext{H}}} \Phi(oldsymbol{x},oldsymbol{ heta}_i), \; ext{ example: } \Phi(oldsymbol{x},oldsymbol{ heta}_i) = a_i \sigmaig(oldsymbol{w}_i^{ op}oldsymbol{x} + b_iig)$$

Population risk functional:

$$R(\hat{f}) = \mathbb{E}_{(oldsymbol{x},y)}igg[\Big(y - \hat{f}\left(oldsymbol{x},oldsymbol{ heta}
ight) \Big)^2 igg] pprox rac{1}{n} \sum_{i=1}^n \Big(y_i - \hat{f}\left(oldsymbol{x}_i,oldsymbol{ heta}
ight) \Big)^2$$

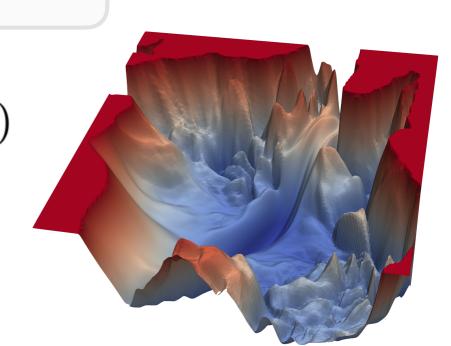
 $\text{Learning problem: } \min_{\boldsymbol{\theta} \in \mathbb{R}^{pn_{\text{H}}}} R(\hat{f}\,)$

Learning Neural Network from Data

 $\text{Learning problem: } \min_{\boldsymbol{\theta} \in \mathbb{R}^{pn_{\mathrm{H}}}} \mathrm{ER}(\hat{f})$

Challenge: highly non-convex (many local minima)

Surprise: SGD and its variants work in practice!!

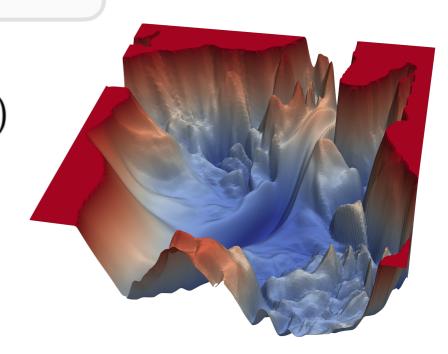


Learning Neural Network from Data

 $\text{Learning problem: } \min_{\boldsymbol{\theta} \in \mathbb{R}^{pn_{\mathrm{H}}}} \mathrm{R}(\hat{f}\,)$

Challenge: highly non-convex (many local minima)

Surprise: SGD and its variants work in practice!!



Good news: emerging theory (starting in 2018!!)

Chizat and Bach (NIPS 2018), Mei, Montanari and Nguyen (PNAS 2018), Rotskoff and Vanden-Eijnden (arXiv:1805.00915, 2018), Williams et al (arXiv:1906.07842, 2019)

Idea: Think of the mean field, i.e., infinite width $(n_{
m H} o \infty)$ limit

$$\hat{f} \equiv \hat{f}\left(oldsymbol{x},
ho
ight) = \int_{\mathbb{R}^p} \Phi(oldsymbol{x},oldsymbol{ heta})
ho(oldsymbol{ heta}) \, \mathrm{d}oldsymbol{ heta}$$

Then, learning problem: $\min_{
ho \in \mathcal{P}_2(\mathbb{R}^p)} R(\hat{f})$

Mean Field Density Dynamics of SGD

Free energy functional: $F(\rho) := R(\hat{f}(\boldsymbol{x}, \rho))$

For quadratic loss:

$$F(\rho) = F_0 + \underbrace{\int_{\mathbb{R}^p} V(\boldsymbol{\theta}) \, \rho(\boldsymbol{\theta}) d\boldsymbol{\theta}}_{\text{independent of } \rho} + \underbrace{\int_{\mathbb{R}^p} \int_{\mathbb{R}^p} U(\boldsymbol{\theta}, \tilde{\boldsymbol{\theta}}) \, \rho(\boldsymbol{\theta}) \rho(\tilde{\boldsymbol{\theta}}) d\boldsymbol{\theta} d\tilde{\boldsymbol{\theta}}}_{\text{interaction potential energy, linear in } \rho} + \underbrace{\int_{\mathbb{R}^p} \int_{\mathbb{R}^p} U(\boldsymbol{\theta}, \tilde{\boldsymbol{\theta}}) \, \rho(\boldsymbol{\theta}) \rho(\tilde{\boldsymbol{\theta}}) d\boldsymbol{\theta} d\tilde{\boldsymbol{\theta}}}_{\text{interaction potential energy, nonlinear in } \rho},$$

where

$$F_0 := \mathbb{E}_{(\boldsymbol{x},y)} [y^2], \qquad V(\boldsymbol{\theta}) := \mathbb{E}_{(\boldsymbol{x},y)} [-2y\Phi(\boldsymbol{x},\boldsymbol{\theta})], \qquad U(\boldsymbol{\theta},\tilde{\boldsymbol{\theta}}) := \mathbb{E}_{(\boldsymbol{x},y)} [\Phi(\boldsymbol{x},\boldsymbol{\theta})\Phi(\boldsymbol{x},\tilde{\boldsymbol{\theta}})]$$

PDF dynamics for SGD:

$$\frac{\partial \rho}{\partial t} = \nabla \cdot (\rho \nabla (\underbrace{V + U \circledast \rho})), \text{ where } (U \circledast \rho)(\boldsymbol{\theta}) := \int_{\mathbb{R}^p} U(\boldsymbol{\theta}, \tilde{\boldsymbol{\theta}}) \rho(\tilde{\boldsymbol{\theta}}) d\tilde{\boldsymbol{\theta}}$$

$$\frac{\delta F}{\delta \rho}$$

This PDE is the gradient flow of functional F w.r.t. the Wasserstein metric W

Wasserstein Proximal Recursion for Training NN

$$\varrho_{k}(\tau, \boldsymbol{\theta}) = \underset{\varrho \in \mathcal{P}(\mathbb{R}^{p})}{\operatorname{arg \, min}} \frac{1}{2} \left(W \left(\varrho(\boldsymbol{\theta}), \varrho_{k-1}(\tau, \boldsymbol{\theta}) \right) \right)^{2} + \tau F(\varrho(\boldsymbol{\theta}))$$
$$= \operatorname{prox}_{\tau F}^{W} \left(\varrho_{k-1} \right)$$

Classifying two Gaussians:

$$d = 40, n = 100,$$

$$a=1, b=0, \sigma(\cdot)=\tanh(\cdot),$$

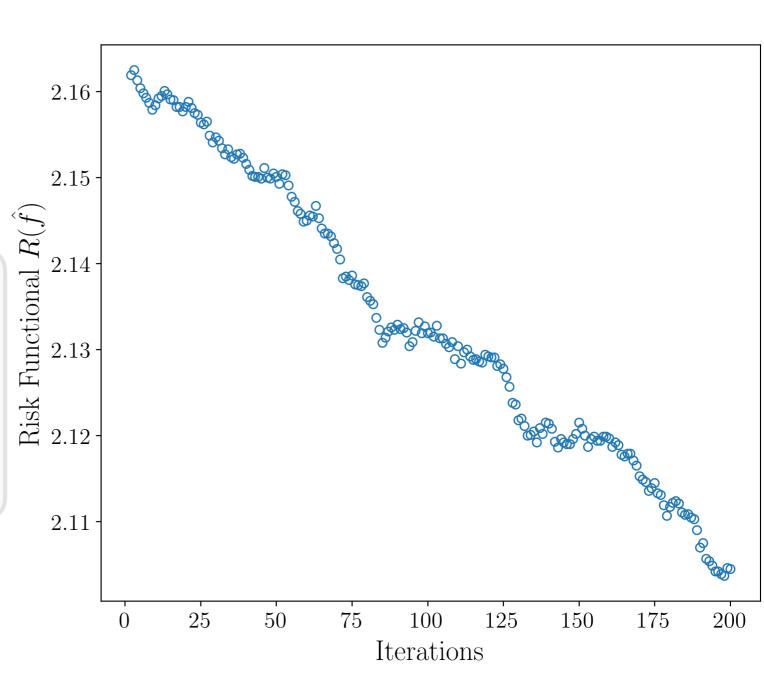
Joint law of $(\boldsymbol{x},y) \in \mathbb{R}^d \times \mathbb{R}$:

$$ext{Prob}ig(y=+1,oldsymbol{x}\sim\mathcal{N}ig(oldsymbol{0},(1+\Delta)^2oldsymbol{I}_dig)ig)=rac{1}{2},$$

$$ext{Prob}ig(y=-1,oldsymbol{x}\sim\mathcal{N}ig(oldsymbol{0},(1-\Delta)^2oldsymbol{I}_dig)ig)=rac{1}{2},$$

$$au = 10^{-3}, n_{
m sample} = 100, \Delta = 0.2,$$

Noisy SGD with
$$\beta = \frac{1}{3}$$



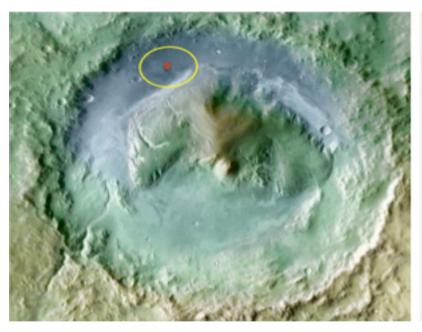
Take Home Message

Emerging system-control theory for densities

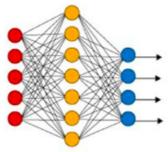
Wasserstein gradient flow: one unifying framework for the prediction, estimation, learning, and feedback control

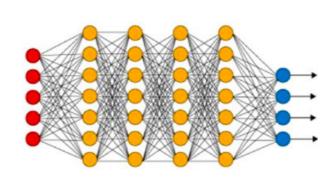
Feedback density control theory: many recent progress, much remains to be done

Several applications: controlling biological and robotic swarm, process control









Thank You









Pictorial Summary

