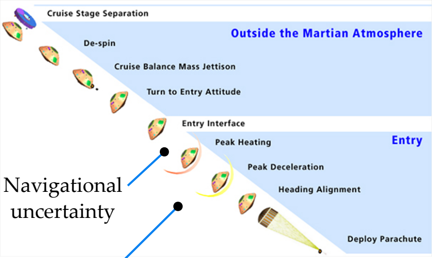
Filtering as Gradient Descent

Abhishek Halder

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

Joint work with Tryphon T. Georgiou

Motivation: Mars Entry-Descent-Landing



Navigational uncertainty

Heating uncertainty

Chute deployment uncertainty

Landing footprint uncertainty

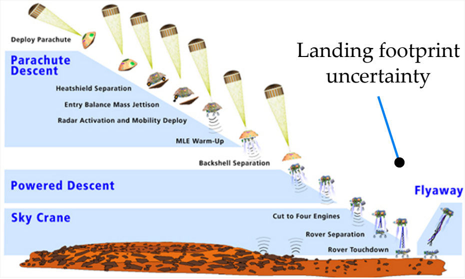
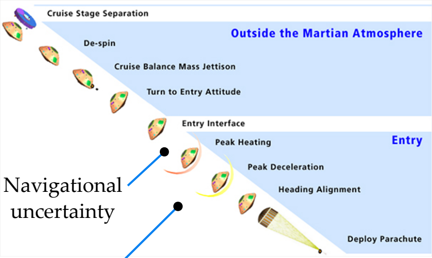


Image credit: NASA JPL

Motivation: Mars Entry-Descent-Landing



Navigational uncertainty

Heating uncertainty

Chute deployment uncertainty

Landing footprint uncertainty

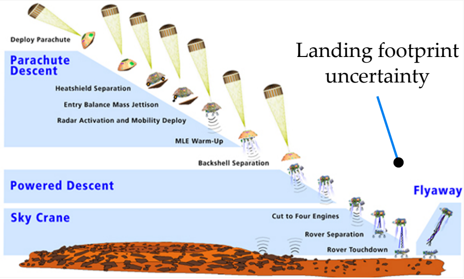
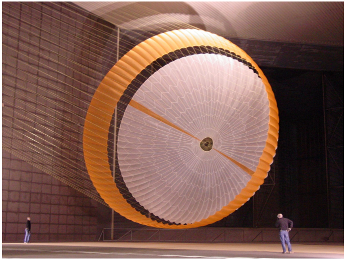


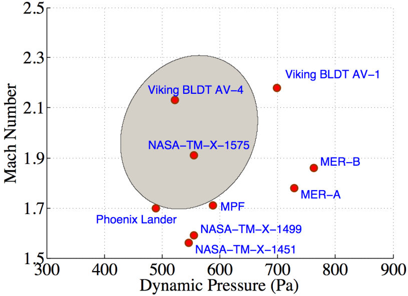
Image credit: NASA JPL

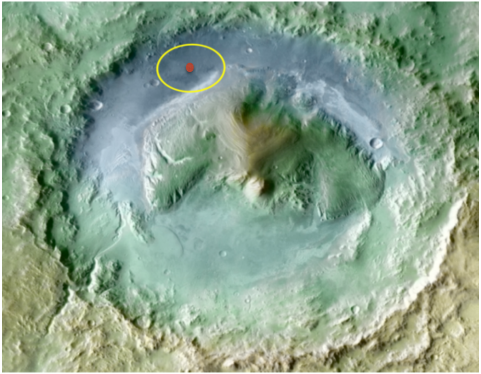
Large number of uncertain scenarios $\rightsquigarrow$ Probability density

Motivation: Mars Entry-Descent-Landing



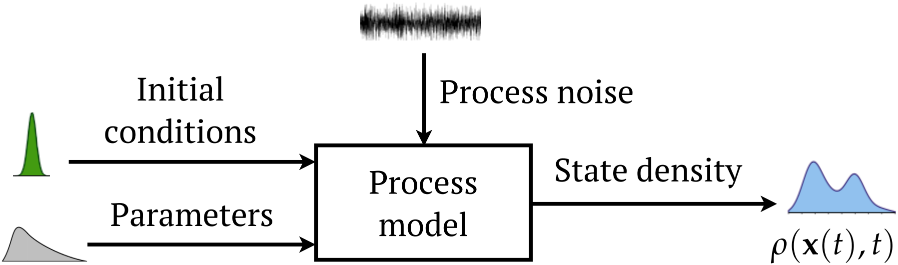
Supersonic parachute



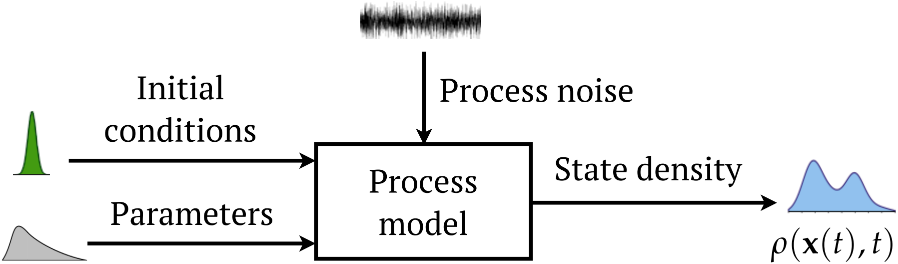


Gale Crater (4.49S, 137.42E)

Problem: Uncertainty Propagation



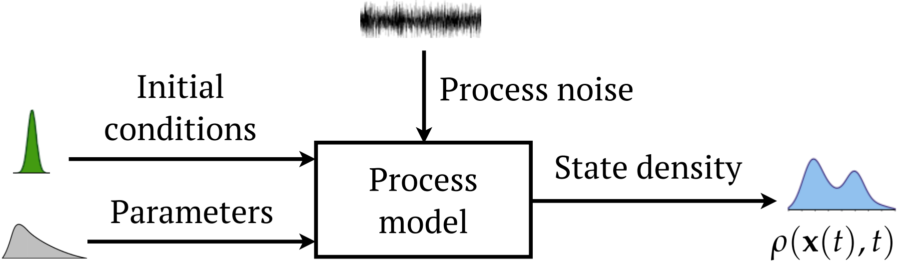
Problem: Uncertainty Propagation



Trajectory flow:

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

Problem: Uncertainty Propagation



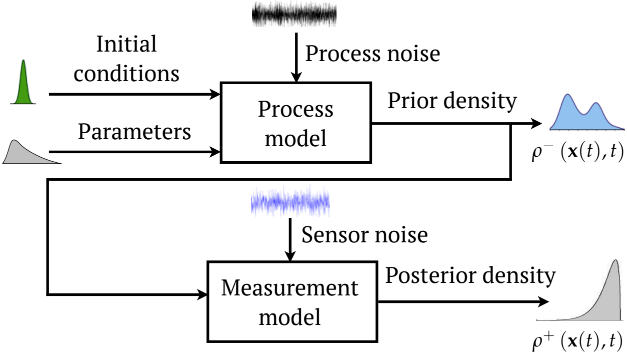
Trajectory flow:

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

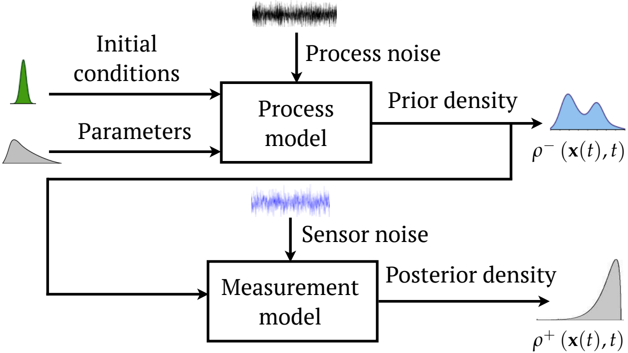
Density flow:

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

Problem: Filtering



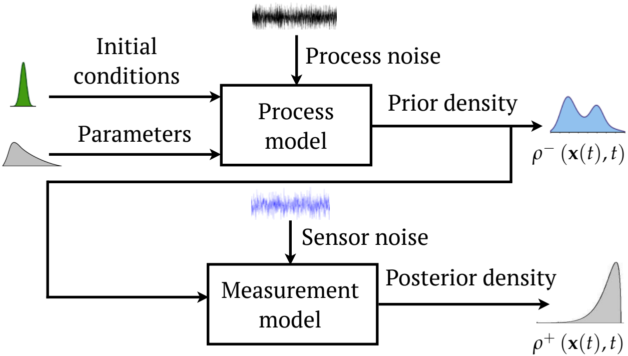
Problem: Filtering



Trajectory flow:

$$
\begin{aligned}
\mathrm{d}\mathbf{X}(t) &= \mathbf{f}(\mathbf{X}, t)\,\mathrm{d}t + \mathbf{g}(\mathbf{X}, t)\,\mathrm{d}\mathbf{w}(t), \qquad && \mathrm{d}\mathbf{w}(t) \sim \mathcal{N}(0, \mathbf{Q}\mathrm{d}t) \\
\mathrm{d}\mathbf{Z}(t) &= \mathbf{h}(\mathbf{X}, t)\,\mathrm{d}t + \mathrm{d}\mathbf{v}(t), && \mathrm{d}\mathbf{v}(t) \sim \mathcal{N}(0, \mathbf{R}\mathrm{d}t)
\end{aligned}
$$

Problem: Filtering



Trajectory flow:

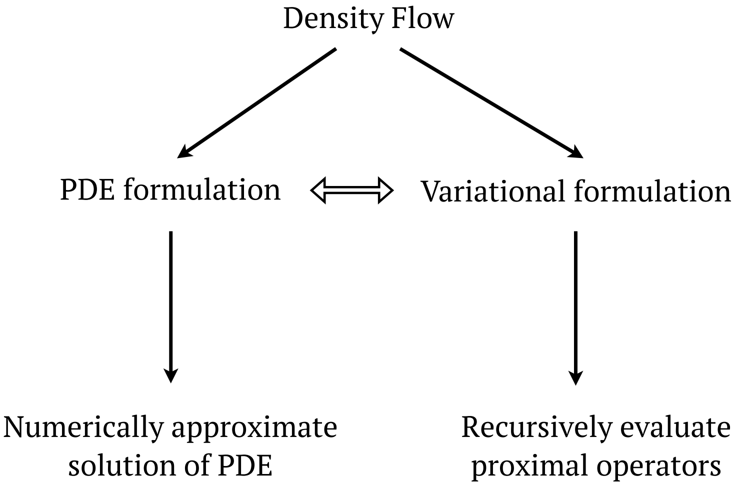
$$\mathrm{d}\mathbf{X}(t) = \mathbf{f}(\mathbf{X},t)\,\mathrm{d}t + \mathbf{g}(\mathbf{X},t)\,\mathrm{d}\mathbf{w}(t), \qquad \mathrm{d}\mathbf{w}(t) \sim \mathcal{N}(0,\mathbf{Q}\mathrm{d}t)$$

$$\mathrm{d}\mathbf{Z}(t) = \mathbf{h}(\mathbf{X},t)\,\mathrm{d}t + \mathrm{d}\mathbf{v}(t), \qquad \mathrm{d}\mathbf{v}(t) \sim \mathcal{N}(0,\mathbf{R}\mathrm{d}t)$$

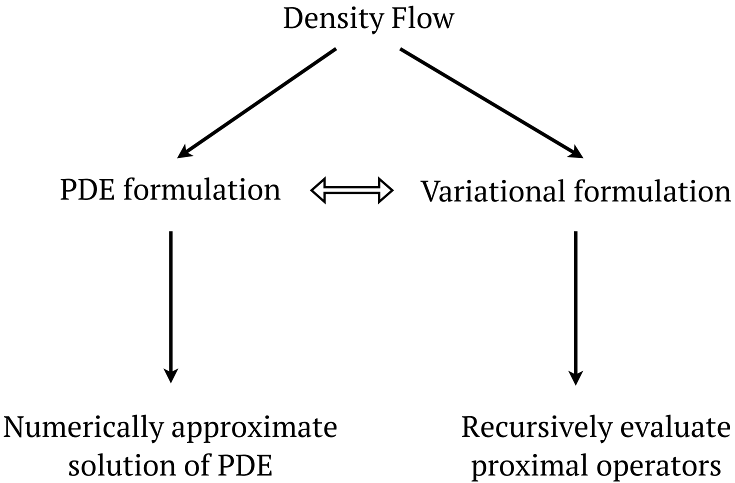
Density flow:

$$\mathrm{d}\rho^{+} = \left[\mathcal{L}_{\mathrm{FP}}\mathrm{d}t + \left(\mathbf{h}(\mathbf{x},t) - \mathbb{E}_{\rho^{+}}\{\mathbf{h}(\mathbf{x},t)\}\right)^{\top}\mathbf{R}^{-1}\left(\mathrm{d}\mathbf{z}(t) - \mathbb{E}_{\rho^{+}}\{\mathbf{h}(\mathbf{x},t)\}\mathrm{d}t\right)\right]\rho^{+}$$

Research Scope



Research Scope



Density flow $\rightsquigarrow$ gradient descent in infinite dimensions

Gradient Descent in Finite Dimensions

Problem: $\underset{\mathbf{x}\in\mathbb{R}^n}{\text{minimize}} \quad \phi(\mathbf{x})$

Algorithm: $\mathbf{x}_k = \mathbf{x}_{k-1} - h\nabla\phi(\mathbf{x}_{k-1})$

Gradient Descent in Finite Dimensions

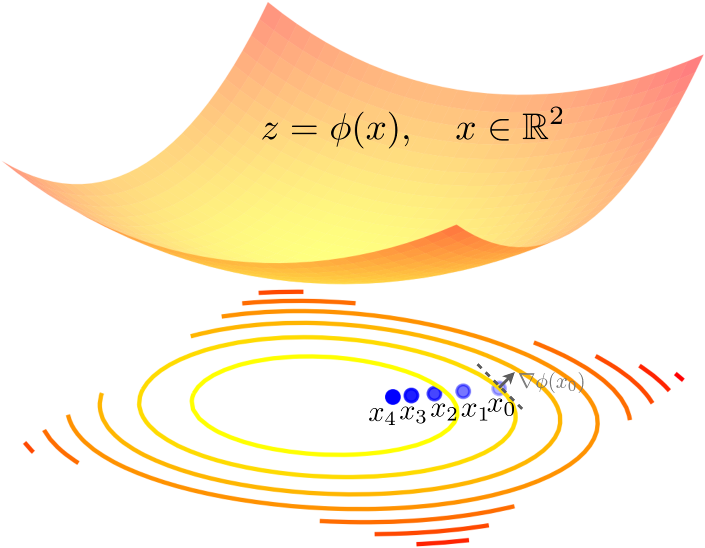
Problem: $\underset{\mathbf{x}\in\mathbb{R}^n}{\text{minimize}} \quad \phi(\mathbf{x})$

Algorithm: $\mathbf{x}_k = \mathbf{x}_{k-1} - h\nabla\phi(\mathbf{x}_{k-1})$

Advantage:

- is a descent method: $\phi(\mathbf{x}_k) \leq \phi(\mathbf{x}_{k-1})$
- convergence under very few assumptions
- simple first order method
- can account constraints (projected gradient descent)

Why does gradient descent work?



$-\nabla\phi(\mathbf{x})$ is the max-rate descending direction (why?)

Rate of Convergence for Gradient Descent

If	**then**
ϕ is $(\frac{1}{h})$-smooth ($\Leftrightarrow \nabla\phi$ is $\frac{1}{h}$ Lipschitz)	$O(\frac{1}{kh})$
ϕ is $(\frac{1}{h})$-smooth AND σ-strongly convex	$O(\frac{1}{h}\exp\left(-\frac{h\sigma}{2}k\right))$

Gradient Descent $\leftrightsquigarrow$ Gradient Flow

- GD is **Euler discretization** of GF

$$\frac{\mathrm{d}\mathbf{x}}{\mathrm{d}t} = -\nabla\phi(\mathbf{x}), \mathbf{x} \in \mathbb{R}^n$$

- **Rate matching:**

GD rate $O(\frac{1}{kh})$ when ϕ is $(\frac{1}{h})$-smooth

GF rate $O(\frac{1}{t})$ when ϕ is convex

Gradient Descent $\leftrightsquigarrow$ Proximal Operator

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

$$\Updownarrow$$

$$\mathbf{x}_k = \mathrm{proximal}_{h\phi}^{\|\cdot\|}(\mathbf{x}_{k-1})$$

$$:= \operatorname*{argmin}_{\mathbf{x}\in\mathbb{R}^n} \left\{\frac{1}{2}\|\mathbf{x} - \mathbf{x}_{k-1}\|^2 + h\phi(\mathbf{x})\right\}$$

Gradient Descent ⇜⇝ Proximal Operator

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

$$\Updownarrow$$

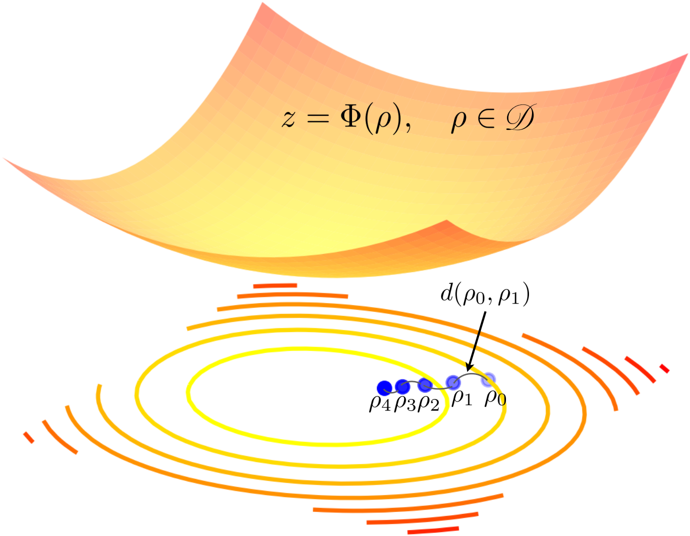
$$\mathbf{x}_k = \text{proximal}_{h\phi}^{\|\cdot\|}(\mathbf{x}_{k-1})$$

$$:= \underset{\mathbf{x}\in\mathbb{R}^n}{\operatorname{argmin}} \left\{ \tfrac{1}{2}\|\mathbf{x} - \mathbf{x}_{k-1}\|^2 + h\phi(\mathbf{x}) \right\}$$

This is nice because

- argmin of $\phi \equiv$ fixed point of prox. operator
- prox. is smooth even when ϕ is not
- reveals metric structure of gradient descent

Gradient Descent in Infinite Dimensions



Proximal recursion: $\rho_k = \underset{\rho \in \mathscr{D}}{\operatorname{arginf}} \left\{ \frac{1}{2} d^2(\rho, \rho_{k-1}) + h\Phi(\rho) \right\}$

Gradient Descent Summary

Finite dimensions

$$\frac{\mathrm{d}\mathbf{x}}{\mathrm{d}t} = -\nabla\phi(\mathbf{x}), \quad \mathbf{x} \in \mathbb{R}^n$$

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

$$= \underset{\mathbf{x}}{\text{argmin}}\{\tfrac{1}{2}\|\mathbf{x} - \mathbf{x}_{k-1}\|^2 + h\phi(\mathbf{x})\}$$

$$= \text{proximal}_{h\phi}^{\|\cdot\|}(\mathbf{x}_{k-1})$$

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

Infinite dimensions

$$\frac{\partial\rho}{\partial t} = \mathcal{L}(\mathbf{x}, \rho), \quad \mathbf{x} \in \mathbb{R}^n,\ \rho \in \mathscr{D}$$

$$\rho_k(\mathbf{x}, h)$$

$$= \underset{\rho}{\text{argmin}}\{\tfrac{1}{2}d(\rho, \rho_{k-1})^2 + h\Phi(\rho)\}$$

$$= \text{proximal}_{h\Phi}^{d(\cdot,\cdot)}(\rho_{k-1})$$

$$\rho_k(\mathbf{x}, h) \rightharpoonup \rho(\mathbf{x}, t = kh), \text{ as } h \downarrow 0$$

Related Work

Transport PDE $\frac{\partial \rho}{\partial t} = \mathcal{L}(\mathbf{x}, \rho)$	Gradient descent scheme	
$\mathcal{L}(\mathbf{x}, \rho)$	$\frac{1}{2} d^2(\rho, \rho_{k-1})$	$\Phi(\rho)$
$\triangle \rho$ Heat equation (1822)	$\frac{1}{2} \parallel \rho - \rho_{k-1} \parallel^2_{L_2(\mathbb{R}^n)}$ Squared L_2 norm of difference	$\frac{1}{2} \int_{\mathbb{R}^n} \parallel \nabla \rho \parallel^2$ Dirichlet energy, CFL (1928)
$\nabla \cdot (\nabla U(\mathbf{x})\rho) + \beta^{-1} \triangle \rho$ Fokker-Planck-Kolmogorov PDE (1914,'17,'31)	$\frac{1}{2} W^2(\rho, \rho_{k-1})$ Optimal transport cost	$\mathbb{E}_\rho \left[U(\mathbf{x}) + \beta^{-1} \log \rho \right]$ Free energy, JKO (1998)
$\left((\mathbf{h} - \mathbb{E}_\rho[\mathbf{h}])^\top \mathbf{R}^{-1} (\mathrm{d}\mathbf{z} - \mathbb{E}_\rho[\mathbf{h}]\mathrm{d}t) \right) \rho$ Kushner-Stratonovich SPDE (1964,'59)	$\mathrm{D}_{KL}(\rho \| \rho_{k-1})$ Kullback-Leibler divergence	$\frac{1}{2} \mathbb{E}_\rho [(\mathbf{y}_k - \mathbf{h})^\top \mathbf{R}^{-1} (\mathbf{y}_k - \mathbf{h})]$ Quadratic surprise, LMMR (2015)

Related Work

Transport PDE $\frac{\partial \rho}{\partial t} = \mathcal{L}(\mathbf{x}, \rho)$	Gradient descent scheme	
$\mathcal{L}(\mathbf{x}, \rho)$	$\frac{1}{2} d^2(\rho, \rho_{k-1})$	$\Phi(\rho)$
$\triangle \rho$ Heat equation (1822)	$\frac{1}{2} \parallel \rho - \rho_{k-1} \parallel^2_{L_2(\mathbb{R}^n)}$ Squared L_2 norm of difference	$\frac{1}{2} \int_{\mathbb{R}^n} \parallel \nabla \rho \parallel^2$ Dirichlet energy, CFL (1928)
$\nabla \cdot (\nabla U(\mathbf{x})\rho) + \beta^{-1} \triangle \rho$ Fokker-Planck-Kolmogorov PDE (1914,'17,'31)	$\frac{1}{2} W^2(\rho, \rho_{k-1})$ Optimal transport cost	$\mathbb{E}_\rho \left[U(\mathbf{x}) + \beta^{-1} \log \rho \right]$ Free energy, JKO (1998)
$\Big(\big(\mathbf{h} - \mathbb{E}_\rho[\mathbf{h}]\big)^\top \mathbf{R}^{-1} \big(\mathrm{d}\mathbf{z} - \mathbb{E}_\rho[\mathbf{h}]\mathrm{d}t\big) \Big) \rho$ Kushner-Stratonovich SPDE (1964,'59)	$\mathrm{D}_{KL}(\rho \vert\vert \rho_{k-1})$ Kullback-Leibler divergence	$\frac{1}{2} \mathbb{E}_\rho \big[(\mathbf{y}_k - \mathbf{h})^\top \mathbf{R}^{-1} (\mathbf{y}_k - \mathbf{h}) \big]$ Quadratic surprise, LMMR (2015)

Process dynamics is stochastic gradient flow:

Gibbs density

$$\mathrm{d}\mathbf{x}(t) = -\nabla U(\mathbf{x})\,\mathrm{d}t + \sqrt{2\beta^{-1}}\,\mathrm{d}\mathbf{w}(t), \qquad \rho_\infty(\mathbf{x}) \propto e^{-\beta U(\mathbf{x})}$$

Related Work

Transport PDE $\frac{\partial\rho}{\partial t} = \mathcal{L}(\mathbf{x},\rho)$	Gradient descent scheme	
$\mathcal{L}(\mathbf{x},\rho)$	$\frac{1}{2}d^2(\rho,\rho_{k-1})$	$\Phi(\rho)$
$\triangle\rho$ Heat equation (1822)	$\frac{1}{2}\parallel \rho-\rho_{k-1} \parallel^2_{L_2(\mathbb{R}^n)}$ Squared L_2 norm of difference	$\frac{1}{2}\int_{\mathbb{R}^n} \parallel \nabla\rho \parallel^2$ Dirichlet energy, CFL (1928)
$\nabla\cdot(\nabla U(\mathbf{x})\rho)+\beta^{-1}\triangle\rho$ Fokker-Planck-Kolmogorov PDE (1914,'17,'31)	$\frac{1}{2}W^2(\rho,\rho_{k-1})$ Optimal transport cost	$\mathbb{E}_\rho\left[U(\mathbf{x})+\beta^{-1}\log\rho\right]$ Free energy, JKO (1998)
$\left(\left(\mathbf{h}-\mathbb{E}_\rho[\mathbf{h}]\right)^\top\mathbf{R}^{-1}\left(\mathrm{d}\mathbf{z}-\mathbb{E}_\rho[\mathbf{h}]\mathrm{d}t\right)\right)\rho$ Kushner-Stratonovich SPDE (1964,'59)	$\mathrm{D}_{KL}(\rho\|\|\rho_{k-1})$ Kullback-Leibler divergence	$\frac{1}{2}\mathbb{E}_\rho[(\mathbf{y}_k-\mathbf{h})^\top\mathbf{R}^{-1}(\mathbf{y}_k-\mathbf{h})]$ Quadratic surprise, LMMR (2015)

No process dynamics, only measurement update:

$$\mathrm{d}\mathbf{x}(t)=0,\quad \mathrm{d}\mathbf{z}(t)=\mathbf{h}(\mathbf{x},t)\,\mathrm{d}t+\mathrm{d}\mathbf{v}(t),\quad \mathrm{d}\mathbf{v}(t)\sim\mathcal{N}(0,\mathbf{R}\mathrm{d}t)$$

Our Contribution

Transport description	Gradient descent scheme	
PDE/SDE/ODE	$\frac{1}{2}d^2(\rho, \rho_{k-1})$	$\Phi(\rho)$
Mean ODE, Lyapunov ODE Linear Gaussian uncertainty propagation	$\frac{1}{2}W^2(\rho, \rho_{k-1})$ Optimal transport cost	$\mathbb{E}_\rho\left[U(\mathbf{x}, t) + \frac{\text{tr}(\mathbf{P}_\infty)}{n}\log\rho\right]$ Generalized free energy
Conditional mean SDE, Riccati ODE Kalman-Bucy filter	$D_{KL}(\rho \vert\vert \rho_{k-1})$ Kullback-Leibler divergence	$\frac{1}{2}\mathbb{E}_\rho[(\mathbf{y}_k - \mathbf{h})^\top \mathbf{R}^{-1}(\mathbf{y}_k - \mathbf{h})]$ Quadratic surprise
ditto	$\frac{1}{2}d^2_{\text{FR}}(\rho, \rho_{k-1})$ Fisher-Rao metric	ditto
Kushner-Stratonovich SPDE Nonlinear filter	ditto Fisher-Rao metric	ditto

The Case for Linear Gaussian Systems

Model:

$$\mathrm{d}\mathbf{x}(t) = \mathbf{A}\mathbf{x}(t)\mathrm{d}t + \mathbf{B}\mathrm{d}\mathbf{w}(t), \qquad \mathrm{d}\mathbf{w}(t) \sim \mathcal{N}(0, \mathbf{Q}\mathrm{d}t)$$

$$\mathrm{d}\mathbf{z}(t) = \mathbf{C}\mathbf{x}(t)\mathrm{d}t + \mathrm{d}\mathbf{v}(t), \qquad \mathrm{d}\mathbf{v}(t) \sim \mathcal{N}(0, \mathbf{R}\mathrm{d}t)$$

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

For uncertainty propagation:

$$\dot{\mu} = \mathbf{A}\mu, \ \mu(0) = \mu_0; \qquad \dot{\mathbf{P}} = \mathbf{A}\mathbf{P} + \mathbf{P}\mathbf{A}^\top + \mathbf{B}\mathbf{Q}\mathbf{B}^\top, \ \mathbf{P}(0) = \mathbf{P}_0.$$

For filtering:

$$\mathrm{d}\mu^+(t) = \mathbf{A}\mu^+(t)\mathrm{d}t + \underbrace{\mathbf{K}(t)}_{\mathbf{P}^+\mathbf{C}\mathbf{R}^{-1}} (\mathrm{d}\mathbf{z}(t) - \mathbf{C}\mu^+(t)\mathrm{d}t),$$

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

The Case for Linear Gaussian Systems

Challenge 1:

How to actually perform the infinite dimensional optimization over $\mathscr{D}_2$?

Challenge 2:

If and how one can apply the variational schemes for generic linear system with Hurwitz $\mathbf{A}$ and controllable $(\mathbf{A}, \mathbf{B})$?

Addressing Challenge 1: How to Compute

Two Step Optimization Strategy

– Notice that the objective is a *sum*:

first functional

second functional

$$\operatorname*{arginf}_{\rho \in \mathscr{D}_2} \left\{ \frac{1}{2} d(\rho, \rho_{k-1})^2 + h\Phi(\rho) \right\}$$

– Choose a parametrized subspace of $\mathscr{D}_2$ such that the individual minimizers over that subspace match

– Then optimize over parameters

– $\mathscr{D}_{\mu,\mathbf{P}} \subset \mathscr{D}_2$ works!

Addressing Challenge 2: Generic $(\mathbf{A}, \sqrt{2}\mathbf{B})$

Two Successive Coordinate Transformations

#1. Equipartition of energy:

- Define *thermodynamic temperature* $\theta := \frac{1}{n}\mathrm{tr}(\mathbf{P}_\infty)$, and *inverse temperature* $\beta := \theta^{-1}$
- State vector: $\mathbf{x} \mapsto \mathbf{x}_{\mathrm{ep}} := \sqrt{\theta}\mathbf{P}_\infty^{-\frac{1}{2}}\mathbf{x}$
- System matrices:

$$\mathbf{A}, \sqrt{2}\mathbf{B} \mapsto \underbrace{\mathbf{P}_\infty^{-\frac{1}{2}}\mathbf{A}\mathbf{P}_\infty^{\frac{1}{2}}}_{\mathbf{A}_{\mathrm{ep}}}, \sqrt{2\theta}\;\underbrace{\mathbf{P}_\infty^{-\frac{1}{2}}\mathbf{B}}_{\mathbf{B}_{\mathrm{ep}}}$$

- Stationary covariance: $\mathbf{P}_\infty \mapsto \theta\mathbf{I}$

Addressing Challenge 2: Generic $(\mathbf{A}, \sqrt{2}\mathbf{B})$

Two Successive Coordinate Transformations

#2. Symmetrization:

- State vector: $\mathbf{x}_{\text{ep}} \mapsto \mathbf{x}_{\text{sym}} := e^{-\mathbf{A}_{\text{ep}}^{\text{skew}} t}\mathbf{x}_{\text{ep}}$
- System matrices:

$$\mathbf{A}_{\text{ep}}, \sqrt{2\theta}\mathbf{B}_{\text{ep}} \mapsto \underbrace{e^{-\mathbf{A}_{\text{ep}}^{\text{skew}} t}\mathbf{A}_{\text{ep}}^{\text{sym}} e^{\mathbf{A}_{\text{ep}}^{\text{skew}} t}}_{\mathbf{F}(t)}, \sqrt{2\theta}\, \underbrace{e^{-\mathbf{A}_{\text{ep}}^{\text{skew}} t}\mathbf{B}_{\text{ep}}}_{\mathbf{G}(t)}$$

- Stationary covariance:
  $\theta\mathbf{I} \mapsto \theta\mathbf{I}$
- Potential: $U(\mathbf{x}_{\text{sym}}, t) := -\frac{1}{2}\mathbf{x}_{\text{sym}}^{\top}\mathbf{F}(t)\mathbf{x}_{\text{sym}} \geq 0$

Summary

- Two successive coordinate transformations bring generic linear system to JKO canonical form
- Can apply two step optimization strategy in $\mathbf{x}_{\text{sym}}$ coordinate
- Recovers mean-covariance propagation, and Kalman-Bucy filter in $h \downarrow 0$ limit
- Changing the distance in LMMR from D_{KL} to $\frac{1}{2}W_2^2$ gives Luenberger-type observers
- **Future work:** computation for nonlinear filtering

Thank You

Backup Slides

Gradient Descent with Constraints

$$\underset{\mathbf{x}\in\mathcal{C}}{\text{minimize}} \quad \phi(\mathbf{x})$$

$$\Updownarrow$$

$$\mathbf{x}_k = \text{proj}_{\mathcal{C}}\left(\mathbf{x}_{k-1} - h\nabla\phi(\mathbf{x}_{k-1})\right)$$

$$\Updownarrow$$

$$\mathbf{x}_k = \text{proximal}_{h\phi}^{\|\cdot\|}\left(\mathbf{x}_{k-1}\right)$$

$$:= \underset{\mathbf{x}\in\mathcal{C}}{\text{argmin}}\left\{\tfrac{1}{2}\|\mathbf{x} - \mathbf{x}_{k-1}\|^2 + h\phi(\mathbf{x})\right\}$$