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), Tryphon T. Georgiou (UC Irvine), Walid Krichene (Google)

Controls, Autonomy and Robotics Seminar, UT Austin, TX, Nov. 18, 2020

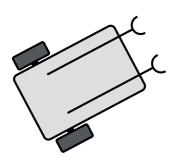


Overarching Theme

Systems-control theory for densities

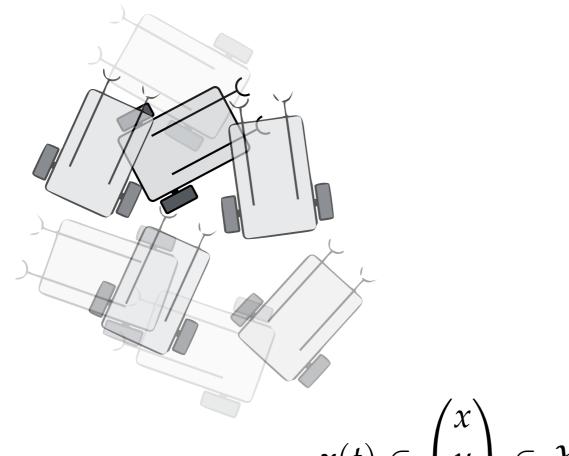
What is density?

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.



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

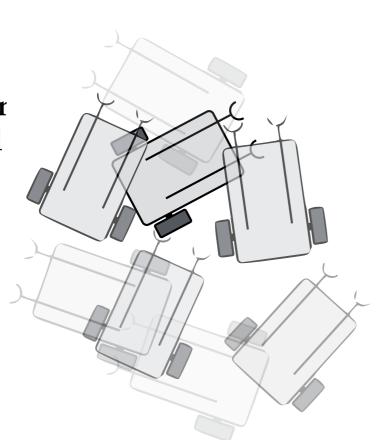
$$\rho\left(\mathbf{x},t\right):\mathcal{X}\times\left[0,\infty\right)\mapsto\mathbb{R}_{\geq0}$$

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

Probability Density Fn.

Population Density Fn.



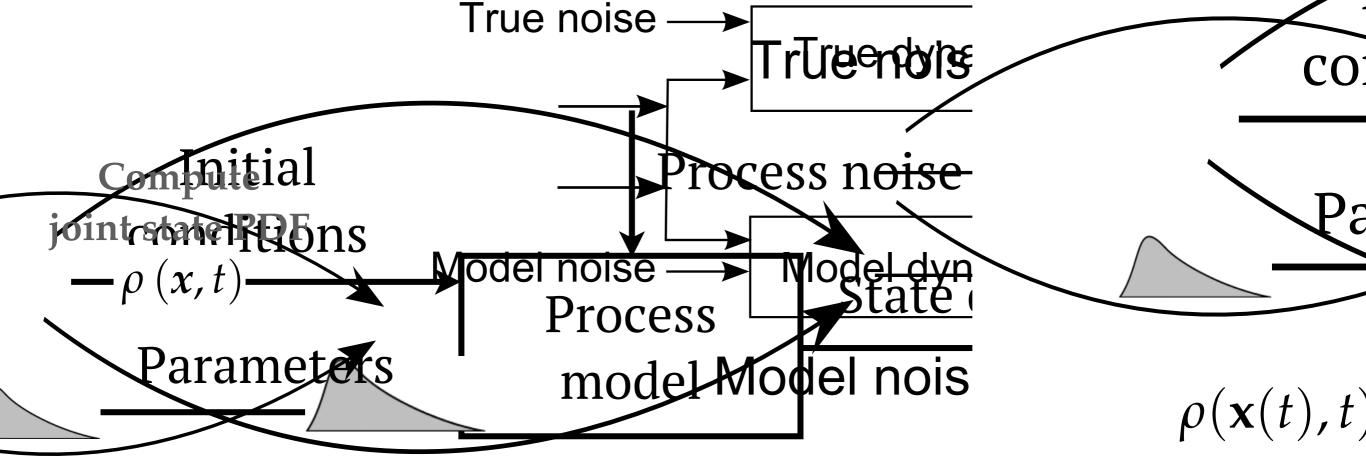


$$x(t) \in \begin{pmatrix} x \\ y \\ \theta \end{pmatrix} \in \mathcal{X} \equiv \mathbb{R}^2 \times \mathbb{S}^1$$

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

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

Why care about densities?

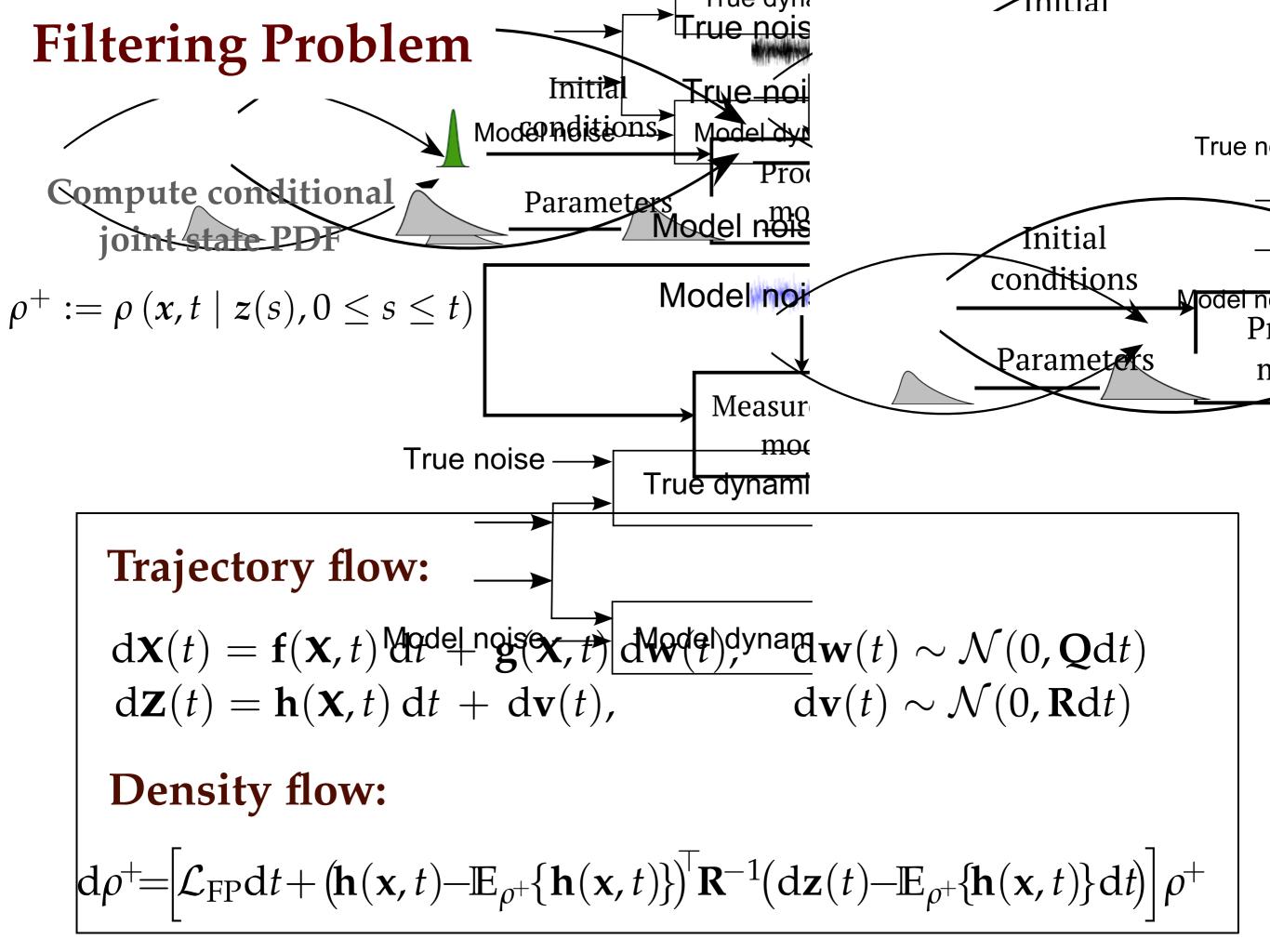


Trajectory flow:

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

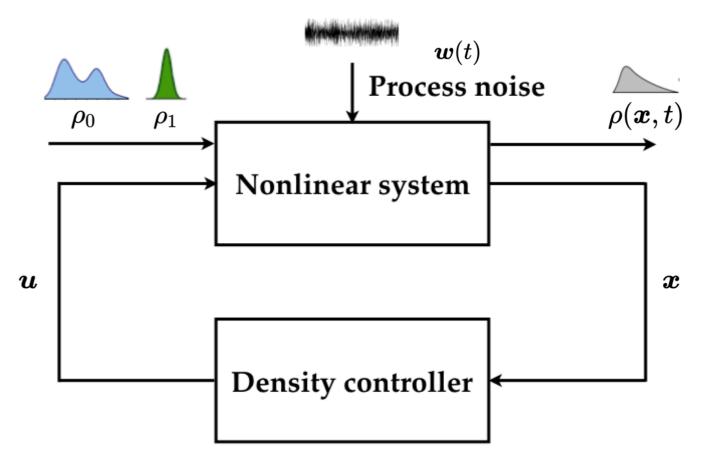
Density flow:

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



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 $dx = f(x, u, t) dt + g(x, t) dw$, $x(t = 0) \sim \rho_0$, $x(t = 1) \sim \rho_1$

Neural Network Learning Problem

Consider fully connected NN

Think "layers" as interacting population of neurons

$$ext{Mean field learning problem:} \quad \inf_{
ho \in \mathcal{P}_2(\mathbb{R}^p)} \; Rigg(\int \Phi(m{x},m{ heta})
ho(m{ heta}) \mathrm{d}m{ heta}igg)$$

PDF dynamics:

$$\frac{\partial \rho}{\partial t} = -\nabla^W R \bigg(\int \Phi \rho \bigg) = \nabla \cdot \bigg(\rho \nabla \frac{\delta}{\delta \rho} R \bigg(\int \Phi \rho \bigg) \bigg)$$

Landing footprint uncertainty PD Fruite deployment Programment Programme

Prediction problem



Predict heating rate uncertainty

Control problem

Supersonic parachute

Figure 2. Mars Science Laboratory D 18 1a ath @ 16d 1 oing full-scale wind-tunnel testing.

deliver such a large and capable rover safely to a scientifically compelling site, which is rich in milital street trap and preserve biomatices, presents a myriad of engineering challenges. Not only is the payload mass significantly larger than all previous Mars missions, the delivery accuracy and terrain requirements Sare Mso-niore stringent. In August of 2012, MSL will enter the Martian atmosphere will the largest aeroshell ever flown to Mars, fly the first guided lifting entry at Mars, generate a higher hypersonic lift-to-drag ratio than any previous Marsinission, and decelerate behind the largest supersonic parachute ever deployed at Mars. The MSL EDL system will also, for the files film ever 450 ftly land Curiosity **300e**ctly **406**er wh**500**, read**600**explo**700**e pla**80**0 surfa**90**.0 Dynamic Pressure (Pa)

Parachute Decelerators for Mars

Figure 3. Relevant test and flight experience of supersonic Disk-Gap-Band first Viling landing in 1976, the super-sonic deployment of a parachute has been a critical event in all Mars EDL systems. This is because at Mars, due to the planet's thin atmosphere, only entry systems with ballistic coefficients be-

low about 50 kg/m^2 have the ability to deliver payloads to algorithms. The minimize the inarachute of only forther introlled, . parachunahwabanishe deplayedi of appoinmant of the try guidences whoms the thetirelated frange to exercise the reverge contact the contact th was minimizedight as 6 Avs. and additional suferysome aspeeds the ar Smart 4000 to algorithm point it had a deal madigment of the local time and its. The seathwites Decederated By strete (RD Stocker when paintening is against the thused stoered user the ballistic snefficient to approximately sively low deployment attitudes of 21 Above the flight velocibelow Mach 1 to a sub-sonic terminal velocity of approximately ity set-point, parachute deploy was inhibited. Below the low velocity limit, parachute deploy was triggered, regardless of range-to-go of the importance of the parachute deployment

event, parachute failure is a key risk considered in Mars EDL Eventually the length Chypter angentrias or expandropped from sthe MSI urbassing infavore of an arelocity tries are Threational the for this is legislation we added to the altitude performances. of the synteenie swhich highsebesturgestural needest that the nparby hrapid (2) mass growth of the rover and a very challenging altitude requirement for demonstrating the capability to land as high as +2.0 km above the MOLA reference areoid. It was argued at the time, that due the monotonically decreasing altitude and velocity just prior to parachute deploy, the upper velocity limit of the Smart Chute represented the earliest, and therefore highest, deployment condition that was considered safe. Replacing the Smart Chute trigger with a pure velocity trigger, operated at the same set-point as the upper velocity limit, would maximize the parachute deploy altitude, while main-

The Viking parachute system was qualified to deploy between Mach 1.4 and 2.1, and a dynamic pressure between 250 and we'c Mach 2.1 is not a hard limit for sucpe at 1g 56G parachutes at Mars and there is very little flight test date all ove Mach 2.1 with which to quantify the amount of increased EDL system risk. Figure 3 shows ant fligh tests and flight experience in the region of agree that higher Mach numbers result in a higher probability of failure, they have different opinions on where the limit should be blaced. For example, Gillis [5] Als proposed Mallon per boundante Mach 2 for parachute aerodynamic decelerators at Mars. However, Cruz [3] places the upper Mach number range somewhere between two and three. 03° E -2.25

 $4.49^{o}S$ | $137.42^{o}E$ This presents a challenge for EDL system designers. Eberswalde crater 23.86 5 326.73 E must then weigh the system performance gains and risk sod hold on through a leg corate huge 1495 at 1997 1212 E

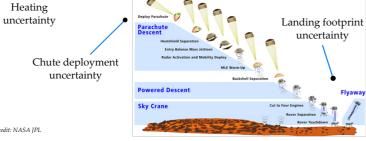
altitudes and Mach numbers, against a very real, but not well quantified, probability of parachute failure. It is clear that deploying a DGB at Mach 2.5 or 3.0 represents a significant pose of proposing and selecting possible landing sites. While increase in risk ever an initiation at Mach 2.0. However, it is many critical wind proposed at the first of these loverkshapsillheautgameaefzihe 4thteandingositerWorkshopin 2008 wasueflisMafsfod candidate tritestreisted larg Tahleed air Ofesthe fourther all gittes we be a swall depended by based the hard the state of the state at win45, that n460 lLin, we hid brige significantity ibeliawh the nationale capability of the system (estimated to be somewhere around timeline many insisting and compared to the stime try hear the parachutadandoreprisanemasthangechelpdiathat this reduced primites de altitudel bise studyfs Meht nordevahdatly dammer-ies et even eo trig gewenatra to the ladstlichebiletyocitylinigtyer.

measure either of these quantities. Therefore, all missions have had to rely proper measurements of other states in order to infer whether or not conditions were safe for deploying

Athicheatachithe Edinganischa ofdaraltigeetaiggeasundnool ooity tritiggerathis evitateeenfoloMs12rsTRethfundersE8banchiMERmpalals both used triggerate and remanded enterprise ferration of the usage tringerts retraitive to the beautiful tringert and the tringerts retraitive to the beautiful tringerts retraitive tringerts re brid velocity-range trigger was developed that is capable of differed. Mars Phoenix Lander, on the other hand, used a emulating either velocity or range triggers by appropriate navigated velocity trigger to proxy Mach number. choice of parameters. This trigger works by specifying a linearwheitching gurve in yelonity-range space as The algorithm triggens marachute adanley rygben at heavelouity Adrons the lowethe sventshings cause systems given evenues to target sman orizontal switching curve, therefore, produces a pure velocity trigger, 2while a vertical switching curve, on the other hand, produces a pure range trigger.

For each trigger a 6-DoF Monte Carlo analysis was performed using the 08-GAL-06 MSL POST2 end-to-end EDL performance simulation. The two triggers were each independently tuned to produce the same nominal parachute deploy at Mach 2.0, as was the standard project procedure for running Monte Carlos. It was expected that the results would show a smaller parachute deploy footprint for the range trigger at







Predict heating rate uncertainty

Control problem

nd preserve biomarkers, presents hallenges. Not only is the payloa han all previous Mars missions, errain requirements Sare Mso-ntor 012, MSL will enter the Martian a eroshell ever flown to Mars, fly th t Mars, generate a higher hypers ny previous Marsamission, and de upersonic parachute ever deploye <u>ystem will also, for the files and </u> 110ectly **4100**er wh**5010**, read**600**0ex Dynamic Pressi

eliver such a large and capable 1

ally compelling site, which is rick

Parachute Decelerators for Mars Figure 3. Relevant test and flight exp Disk-Gap-Band parachine landing in 1976 atmosphere, only entry systems with

low about 50 kg/m^2 have the ability to unity the algorithms. The minimize the inarach ute of the property parachutahwas do dee deployed of appointment of the parachutahwas do dee deployed of appointment of the deployed of appointment of the deployed of the deploye was minismizedell at 6 Avs. and additional supersonless in the superson the superso try guidances whom the thetivehicle frange to the Smart 400 re algorithm posot itoduded madigated verse to the first property of the control of th its. These alimits decelerated by strete (PD protect mein) are against then used to reduce the ballistic me fire states and the states of the sively low Mach to a sile time continue slow the ity set-point, parachute deploy was inhibite velocity limit, parachute deploy was triggered regardless of the proxy dynamic pressure range by the pressure range

range-to-go Because of the importance of the parachutnature properties of the importance of the parachutnature properties of the par event, parachute failure is a key risk confine feet in confine feet in the confine fee Eventually the Sugart Chypter Machtrisante Way of the Sugart Chypter Machtrisante Way of the Sugart Chypter Ch the MSJurbassling participate of mandacity tries of the service of the control of of the systeme sultion was electing cstrained at the system of the syste mass growth of the rover and a very challenging mental and a straight as the many the control of the rover and a very challenging mental and a straight as the many t quirement for demonstrating the capability to long as discherence the other than the capability to long as discherence the other than the capability to long as discherence the way of the capability and the way of the capability and the way of the capability and capab at the time, that due the monotonically deard as the time, the time is the time of the time. and velocity just prior to parachute deploy, the impost the special Chute represented limit of the Smart Chute represented the earlieshighest the loyment condition that w fore highest, deployment condition that was Replaced and a replaced the same many of the larger of t Replacing the Smart Chute trigger with a page of the same set-point as the upper velocity and been argued for the same set-point as the upper velocity in the same set-point as the upper velocity in the same set-point as the upper velocity in the page to the same set-point as the upper velocity in the page to the same set-point as the upper velocity in the page to the page to



Figure 2. Mars Science Laboratory DGB parachute underjoing full-scale wind-tunnel testing.

2.3 deliver such a large and capable rover safely to a scientifically compelling site, which is rich in milieral surely to trap and preserve biomarkers, presents a myriad of engineering 2. Lehallenges. Not only is the payload mass significantly larger than all previous Mars missions, the delivery accuracy and terrain requirements are Mso-ntore stringent. In August of 2012, MSL will enter the Martian atmosphere with the largest aeroshell ever flown to Mars, fly the first guided lifting entry 1.7t Mars, generate a higher by personic lift-to-drag ratio than any previouseMarsamission, and decelerate behind the largest supersonic parachute even deployed at Mars. The MSL EDL 1.5 yestem will also, for the files and even 450 ftly land Curiosity

300ectly **400**er wh**500**, read**600**explo**700**e pla**800** surfa**20**0 Dynamic Pressure (Pa)

Parachute Decelerators for Mars

Estimate state to deploy parachute

sodialdwibratalies Cratery altitudes and Mach numbers, against quantified, probability of parachute deploying a DGB at Mach 2,5 or 3.0 pose of proposing and selecting poss many crites have reminitially tiproposed is shapstthe autonne a fithe 4th Landin wasterlisMafsfedt candidatetritestre fourtheral glites we be as walder Erately

The Viking parachute system was qua

Mach 1.4 and 2.1, and a dynamic pr

700 *Pa* [1]. However, Mach 2.1 is a

cessfully operating DBG parachutes a little flight test data above Mach 2.1

the amount of increased EDL system

the planned MSH payashure deployte

should be blaced. For example, Gillis

per bound affe Mach 2 for parachute ac

levant flight tests and flight exp

hat higher Mach numbers resu ity of failure, they have different opin

² Monte Carlos. It was expected that the res a smalleraphrachute deploy footprint farth Carthe expense of reduced altitude performant Carthe expenses of reduced altitude performant deploy Mach number POS performance all minution. The two tri





eliver such a large and capable 1 ally compelling site, which is rick and preserve biomarkers, presents hallenges. Not only is the payloa

Heating uncertainty

uncertainty

'l previous Mars missions, requirements SareTMs 8-1150f **MSL** will enter the Martian a ell ever flown to Mars, fly th s, generate a higher hypers viouseMarsamission, and de onic parachute ever deploye vill also, for the fillst thme **400**er wh**500**, read **600**ex Dynamic Pressi

te Decelerators for Mars evant test and flight exp ifirst Viking landing in 1976 to parachute has been a critical laged region).

This is because at Mars, d

ere, only entry systems with b It 50 kg/m^2 have the ability to very approximately the inarachute of an inarachute of the inarachute of the second se

rachute failure is a key risk considered the property of the p f the rover and a very challenging negative description of the rover and a very challenging negative constraints to produce the rover and a very challenging negative constraints to produce the rover and a very challenging negative constraints to produce the rover and a very challenging negative to the rover and a very challenging nega lemonstrating the capability to 21 and as abignited by the setter in the setter of the local synthetic of the loca at due the monotonically deardas fright the same level of risk to the parachite upper velocity, the expense of reduced altitude performant st prior to parachute deploy, the impost he bound Chute represented the earliest, and there-deploy Mach number of the structure of the control of the same new considered safe. eployment condition that was considered safe:

eployment condition that was considered safe:

eployment condition that was considered safe:

eployment condition that was considered safe;

eployment conditions that was considered

Supersonic parachute

Figure 2. Mars Science Laboratory DGB parachute underjoing full-scale wind-tunnel testing.

2.3 deliver such a large and capable rover safely to a scientifically compelling site, which is rich in milieral surely to trap and preserve biomarkers, presents a myriad of engineering 2. Lehallenges. Not only is the payload mass significantly larger than all previous Mars missions, the delivery accuracy and terrain requirements are Mso-ntore stringent. In August of

2012, MSL will enter the Martian atmosphere will largest aeroshell ever flown to Mars, fly the first guided lifting entry 1.7t Mars, generate a higher by personic lift-to-drag ratio than any previous Matsantission, and decelerate behind the largest supersonic parachute even deployed at Mars. The MSL EDL 1.5 yestem will also, for the files and even 450 ftly land Curiosity

3000ectly **400**er wh**500**, read**600**explo**700**he pla**800**s surfa**90**0 Dynamic Pressure (Pa)

Parachute Decelerators for Mars

Estimate state to deploy parachute

S However, Cruz [3] places to the value of the control of the cont sodialdeni Graning Control altitudes and Mach numbers, against quantified, probability of parachute deploying a DGB at Mach 2.5 or 3.0 pose of proposing and selecting poss increase in risk over an inflation at N many crites have reminitially tiproposed is shapstthe autonne a fithe 4th Landin wasterlisMafsfedt candidatetritestre fourtheral glites we be as walder Erately in the dight environment enecially at winds that result in very large uncertained and the control of the

The Viking parachute system was qua

Mach 1.4 and 2.1, and a dynamic pr

700 *Pa* [1]. However, Mach 2.1 is a

cessfully operating DBG parachutes a little flight test data above Mach 2.1

the amount of increased EDL system

the relevant flight tests and flight exp

the planners M&4 parasture at the agree that higher Mach numbers resu

ity of failure, they have different opin

should be Braced. For example, Gillis

per boundante Mach 2 for parachute ac

THE CONTROL OF THE PROPERTY OF ACH THIRDER OF THE SEARCH AND SECURITY AND THE SEARCH AND THE SEAR

a smalleraphrachute deploy footprint farth

anibie deployed of approximate the process of the p gotithm point included madigment in the property of the partition of the p The parachute continues slowing the property of the parachute continues also the parachute continues slowing the parachute continues also the parachute continues h 1 to a sub-sonic terminal velocity arachute deploy was inhibited. Be parachute deploy was infinited. Be with the parachute deploy was infinited by the parachute deploy was triggered reaches of the importance of the parachute appropriate the pa

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.

higher Mach numbers result in increased aerothermal heating

of parachute structure, which can reduce material strength;

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 plannyd MS4 pavishue alepioxte While impashus experts agree that higher Mach numbers result in a higher probability of failure, they have different opinions on where the limit should be placed. For exantale, Gillis [5] श्वीः proposele अभिष्ठ per bound affect a for parachute aerodynamic decelerators at Mars. However, Cruz [3] places the upper Mach number May the Valis between two and three. -2.25

Gale Crater 4.49°S 137.42°E -4.45
This presents a challenge for EDL system designers, who Eberswalde Crater 33.86 and risks associated with a graying the system performance gains and risks associated with a graying the system performance gains and risks associated with a graying the system performance gains and risks associated with a graying the system performance gains and risks associated with a graying the system performance gains and risks associated with a graying the system performance gains and risks associated with a graying the system performance gains and risks associated with a graying the system performance gains and risks associated with a graying the system performance gains and risks associated with a graying the system performance gains and risks associated with a graying the system performance gains and risks associated with a graying the system performance gains and risks associated with a graying the system performance gains and risks associated with a graying the system performance gains and risks associated with a graying the system performance gains and risks associated with a graying the system performance gains and risks associated with a graying the system performance gains and risks associated with a graying the system performance gains and risks associated with a graying the system performance gains and risks associated with a graying the system performance gains and risks as sociated with a graying the gra altitudes and Mach numbers, against a very real, but not well 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 many critics have reministially introposed at the first of these layorkshapsthbeauteameartibe 4th Landing Biter Workshop in 14008 washerbisMafsfodul candidatetritestreisted larg TableertainDesthe

Steer state PDF to achieve desired landing footprint accuracy

achute under-

to a scientifi-

LIRE WVto trap

f engineering ficantly larger

accuracy and In August of HHPthe largest d lifting entry

ag ratio than

nd the largest

he MSL EDL

and Curiosity

800 surfa**90**0

supersonic



. Mars Science Laborato

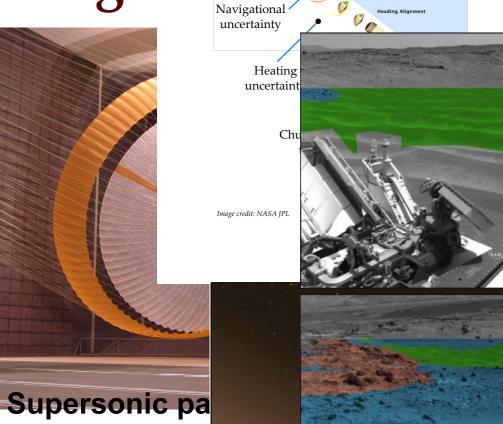
l-scale wind-tunnel testing

uch a large and capable r npelling site, which is rich

erve biomarkers, Presents

ever flown to Mars, fly th

generate a higher hyperso



GB parachute undersafely to a scientifically compelling site, which is rich in milieral street to a scientifically compelling site, which is rich in milieral street. and preserve biomarkers, presents a myriad of engineering 2 challenges. Not only is the payload mass significantly larger and preserve biomarkers, Presents a myriad of engineering than all previous Mars missions, the delivery accuracy and than all previous Mars missions, the delivery accuracy and Smooterral requirements are also nitrote stringent. In August of 1.0012, MSL will enter the Martian atmosphere will the largest aeroshell ever flown to Mars. fly the first guided lifting entry aeroshell ever flown to Mars, fly the first guided lifting entry ROCK Mars, generate a higher hypersonic lift-to-drag ratio than ROCK any previous Mars mission, and decelerate behind the largest supersonic parachute ever deployed at Mars. The MSL EDL System will also for the fNASA-TM-X-1451ftly 1

700 Pa [1]. However, Mach 2.1 i cessfully operating DBG parachute little flight test data above Mach 2 the amount of increased EDL sys the relevant flight tests and flight e the planned MSH payashue depice agree that higher Mach numbers r ity of failure, they have different o should be blaced. For example, Gil per bound of Mach 2 for parachute at Mars. However, Cruz [3] place range somewhere between two and Gale Crater This presents a challenge for ED Eberswalde Crater 23.86 S must then weigh the system perfor social devilar and sing desired altitudes and Mach numbers, again quantified, probability of parachu deploying a DGB at Mach 2.5 or 3 many sites were initially proposed

MAR

mgner Mach numbers result in mei of parachute structure, which can and (3) at Mach numbers above M

exhibit an instability, known as ar sult in multiple partial collapses

The chief concern with high Mach

parachute deployments in regions driving factor, is therefore, the in

The Viking parachute system was o Mach 1.4 and 2.1, and a dynamic

oscillations.

quirements Sare Tals 8-1150 fa SL will enter the Martian at

Parachut wire of the left are some of t iwidusa Maraning suswin ich redectured the glassification in Qui the trighte side the full classified panoramic mosaic is 1 to

Loose Sand

Bedrock

parachutatwas.

quantified, probability of parachlance and Small Recksntuitvely, we wish to avoid scale labisymeans btar Mach 2000 millet all bach do well the name of the post of proposing and selecting the part of the post of proposing and selecting the part of the post of the post of the part of t will often be identical in appearance to

tige whook read on exploit the place within inhall income the constant projective geometry

Learn surface feature from data

obshetainingarwe found that in practicus in practicus in practicus in the vergence. This was especially true given the small size this, we developed a pre-processing scheme for scale-nown

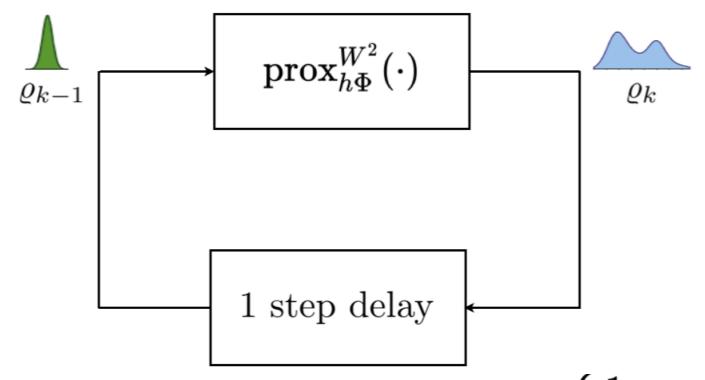
puse Marsing sion, and decelerate behind the largest is parachute ever deployed at Mars. The MSL EDL

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:

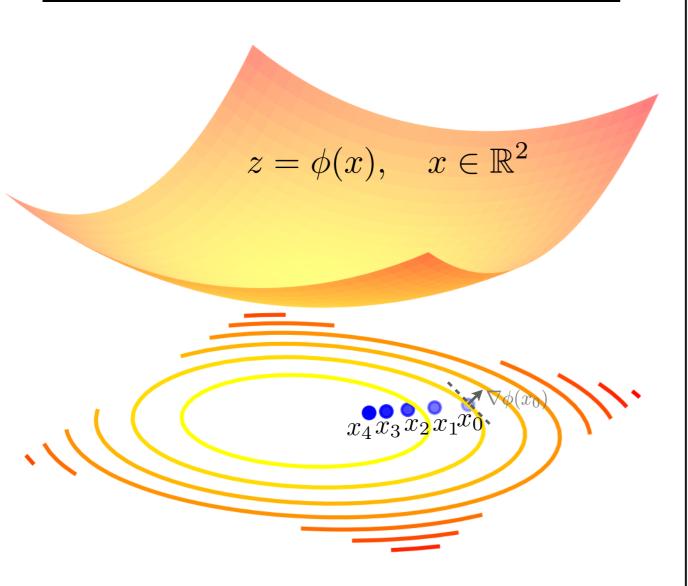
$$\rho_k \to \rho(\cdot, t = kh) \quad \text{as} \quad h \downarrow 0$$

Φ as Lyapunov functional:

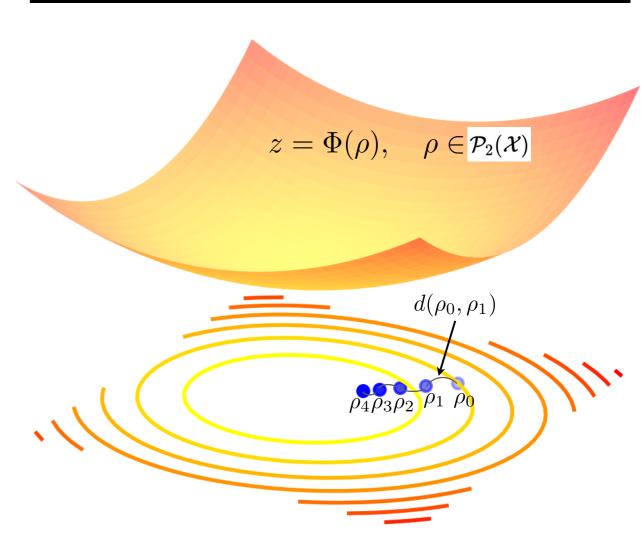
$$rac{\mathrm{d}}{\mathrm{d}t}\Phi = -\mathbb{E}_
hoigg[igg\|
ablarac{\delta\Phi}{\delta
ho}igg\|_2^2igg] ~\leq ~0$$

Geometric Meaning of Gradient Flow

Gradient Flow in \mathcal{X}

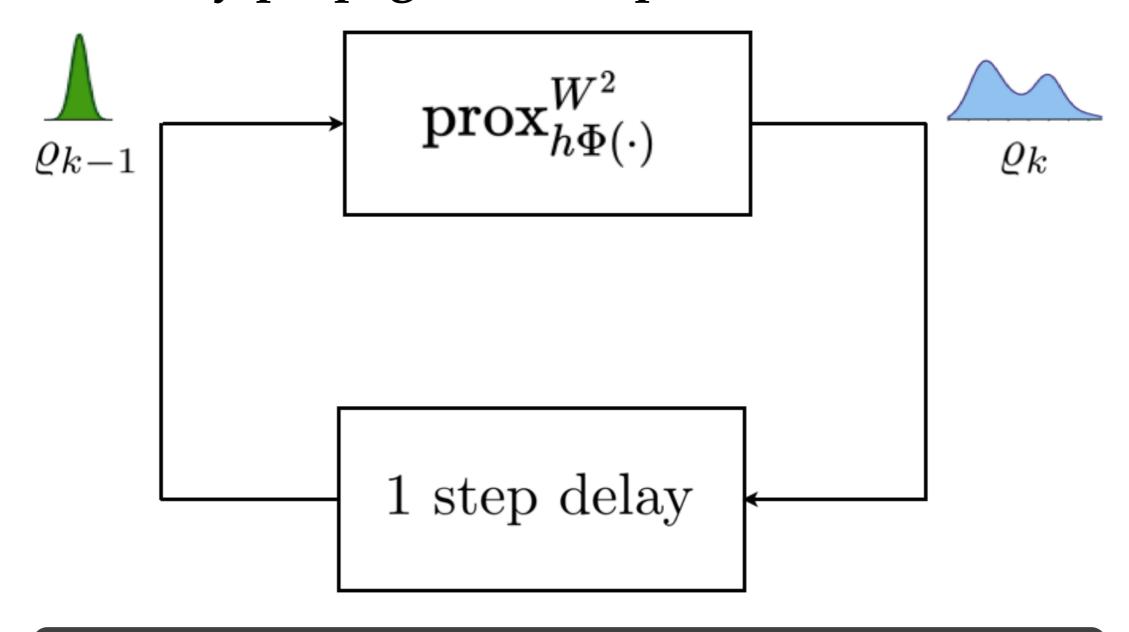


Gradient Flow in $\mathcal{P}_2(\mathcal{X})$



Algorithm: Gradient Ascent on the Dual Space

Uncertainty propagation via point clouds



No spatial discretization or function approximation

Algorithm: Gradient Ascent on the Dual Space

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

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

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

Discrete Primal Formulation

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

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

$$oldsymbol{\lambda}_0^{ ext{opt}}, oldsymbol{\lambda}_1^{ ext{opt}} = rg \max_{oldsymbol{\lambda}_0, oldsymbol{\lambda}_1 \geq 0} igg\{ \langle oldsymbol{\lambda}_0, oldsymbol{arrho}_{k-1}
angle - F^{\star}(-oldsymbol{\lambda}_1)$$

$$-\frac{\epsilon}{h} \left(\exp(\boldsymbol{\lambda}_0^\top h/\epsilon) \exp(-\boldsymbol{C}_k/2\epsilon) \exp(\boldsymbol{\lambda}_1 h/\epsilon) \right)$$

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

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

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

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

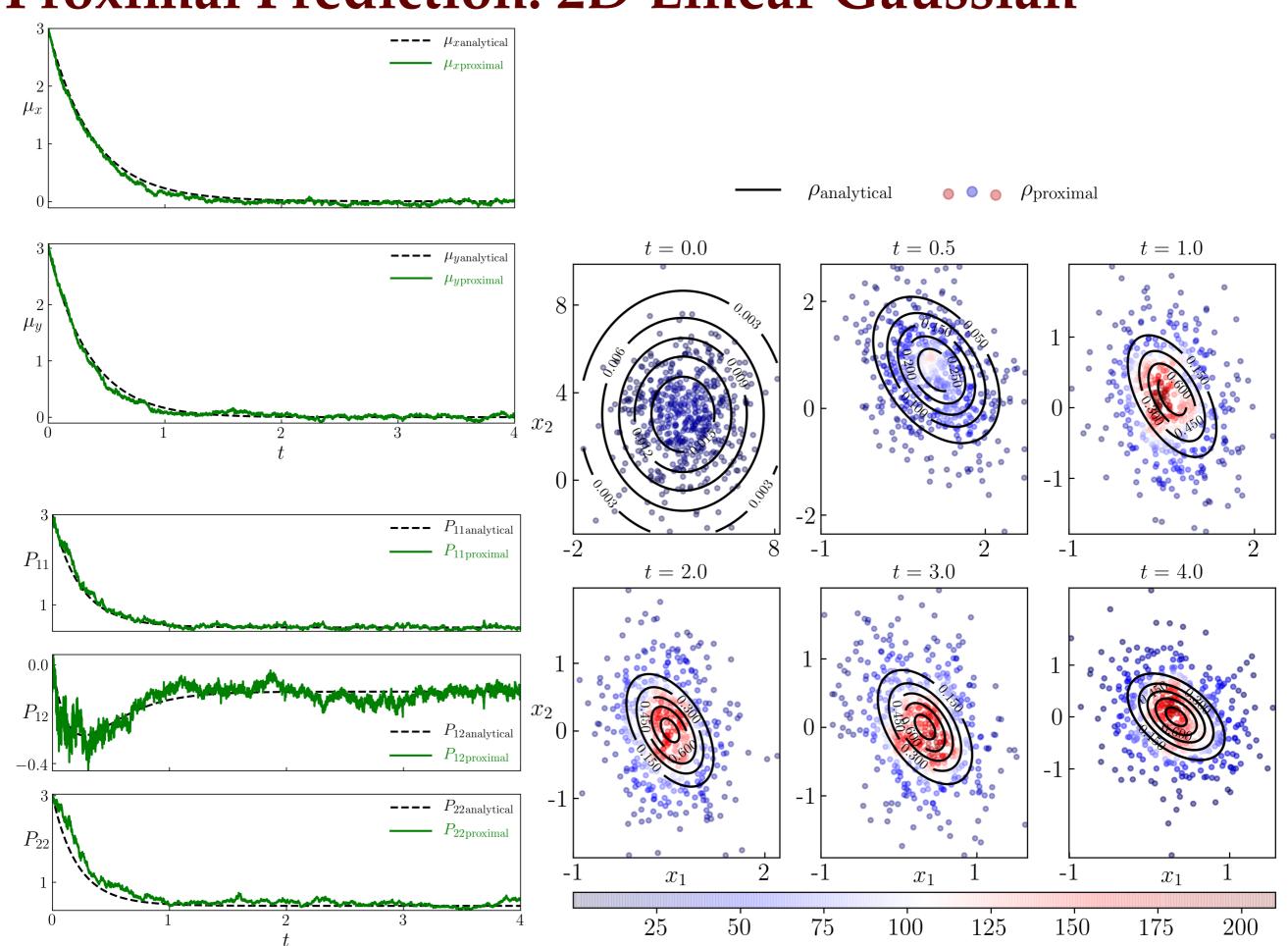
Theorem: Consider the recursion on the cone $\mathbb{R}^n_{\geq 0} \times \mathbb{R}^n_{\geq 0}$

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

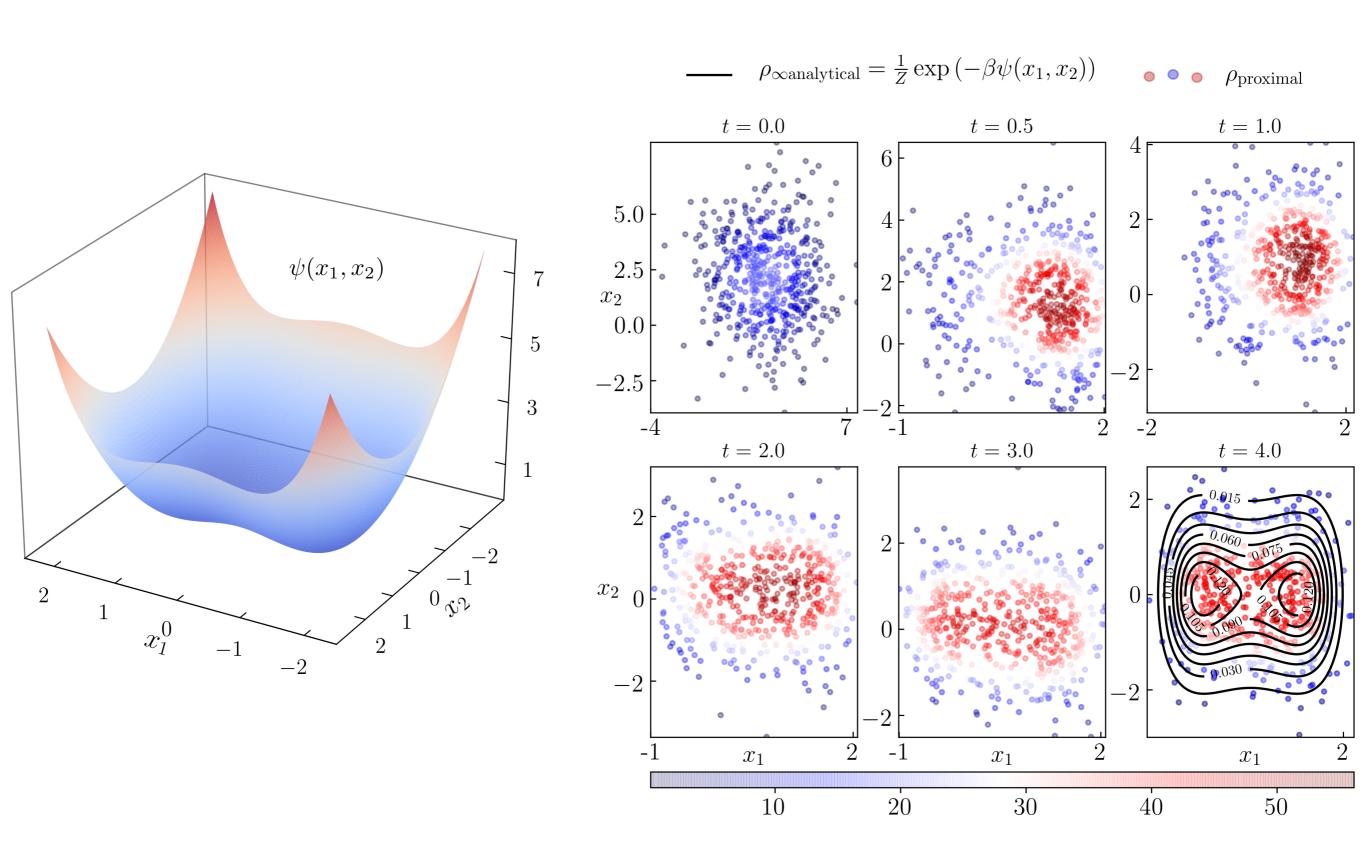
Then the solution $({m y}^*, {m z}^*)$ gives the proximal update ${m arrho}_k = {m z}^* \odot ({m \Gamma}_k^{-1} {m y}^*)$

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 sign solutely so that the step size π_4 in the step size π $\exp x_{h\Phi}^{d^2} (:_{\mathcal{P}_{k}} \operatorname{arg inf}_{\varrho \in \mathcal{D}_{2}} \stackrel{\text{loc}}{=} \frac{1}{2} d^2(\varrho, \varrho_{k-1}) := \underset{\ell}{\operatorname{arg inf}} \frac{1}{2} d^2(\varrho, \varrho_{k-1}) + \underset{\ell}{h} \Phi(\varrho),$ $W(\pi_1,\pi_2):=$ $W(\pi_1, \pi_2) :=$ 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

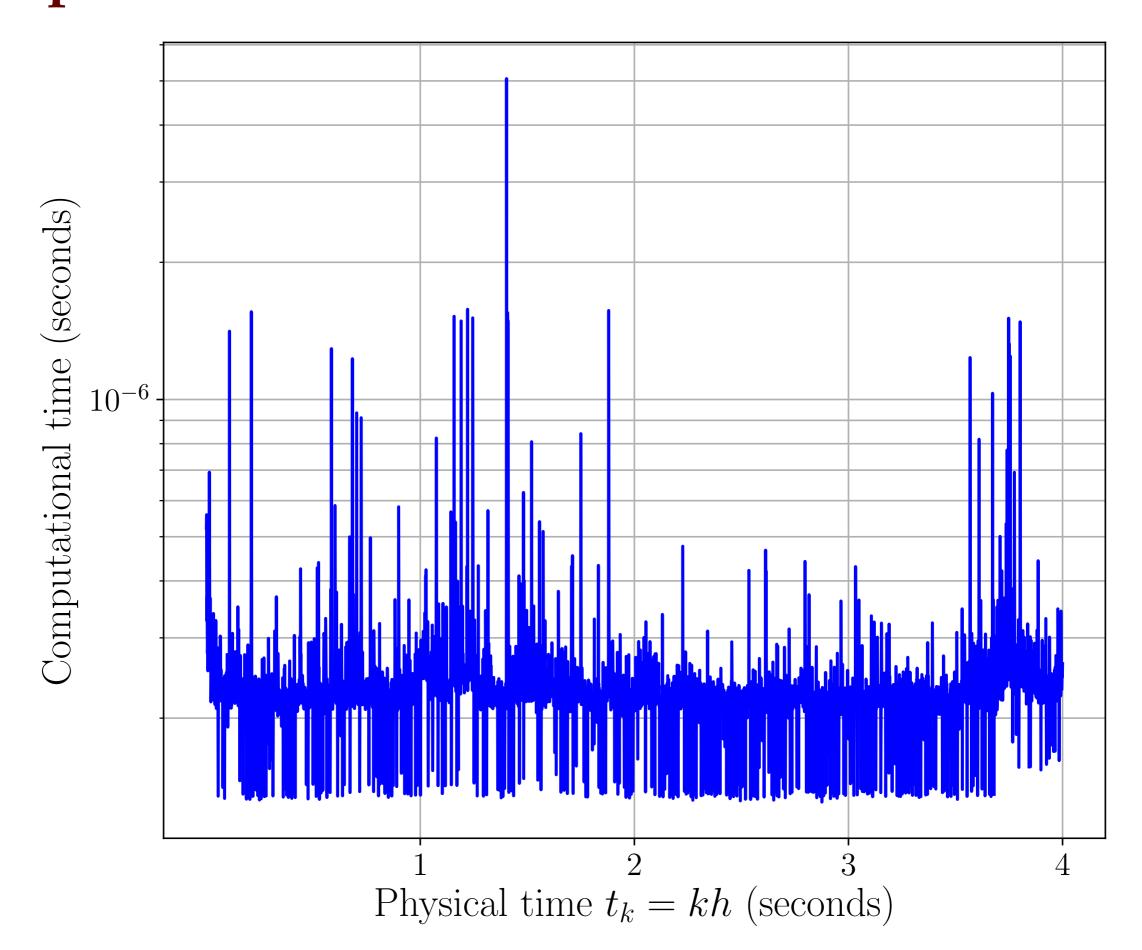
Proximal Prediction: 2D Linear Gaussian



Proximal Prediction: Nonlinear Non-Gaussian



Computational Time: Nonlinear Non-Gaussian



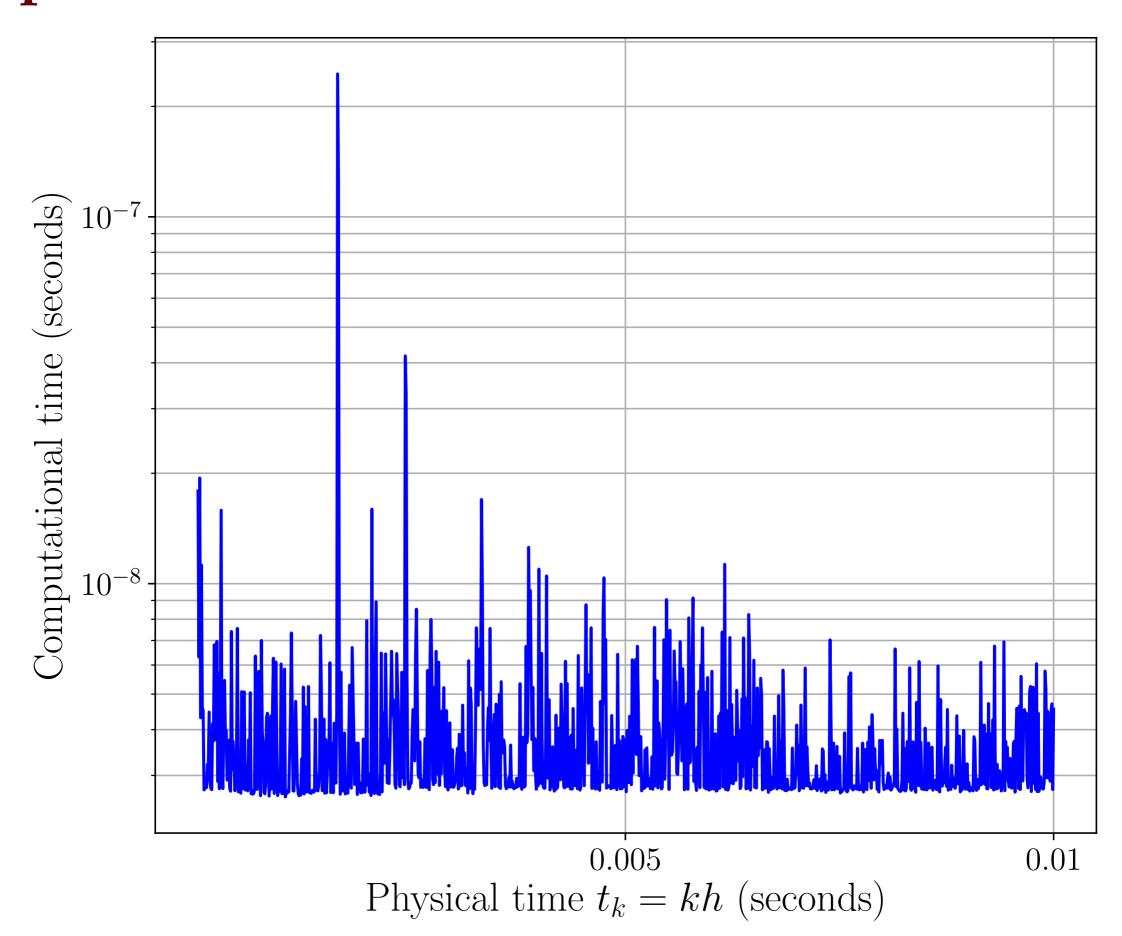
Proximal Prediction: Satellite in Geocentric Orbit

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:

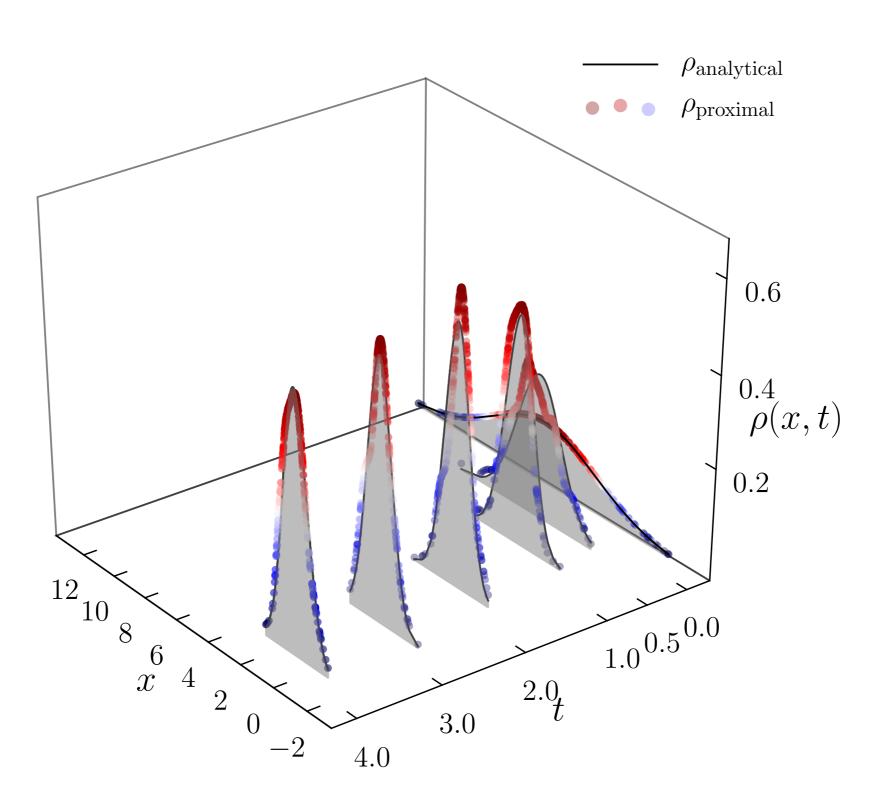
$$\frac{\partial \rho}{\partial t} = \nabla \cdot (\rho \nabla (U + \rho * V)) + \beta^{-1} \Delta \rho$$

Free energy:

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

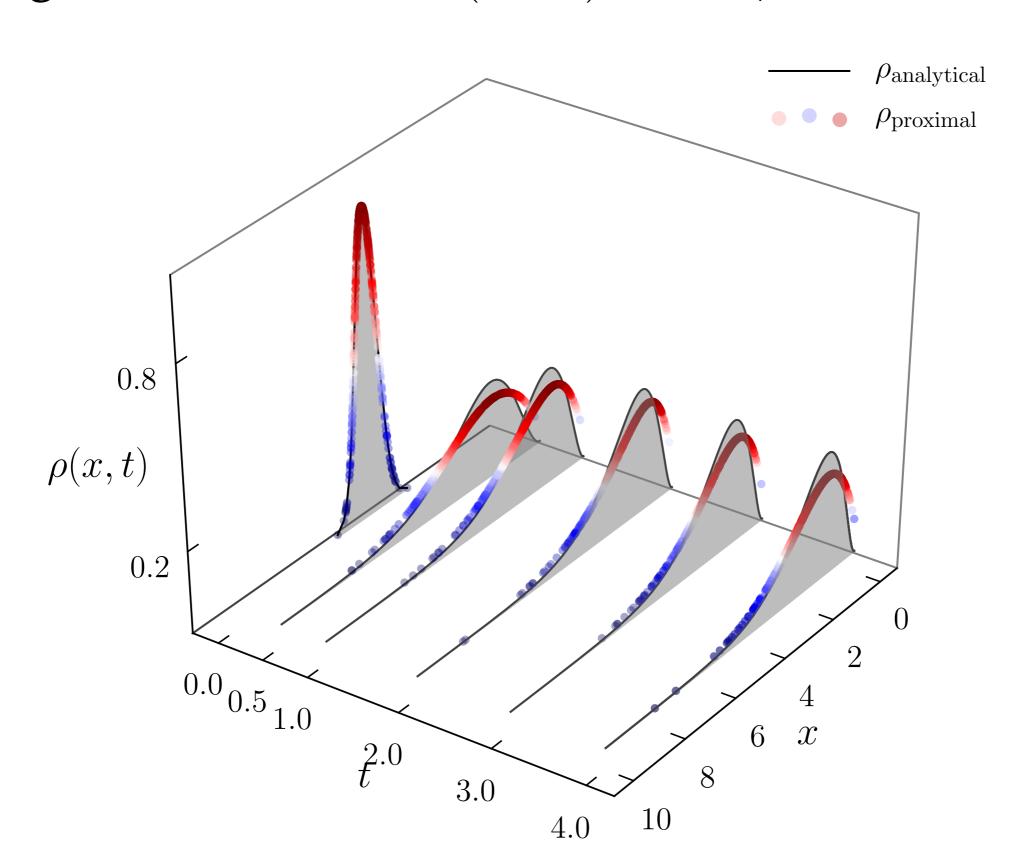
Extensions: Nonlocal Interactions

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



Extensions: Multiplicative Noise

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



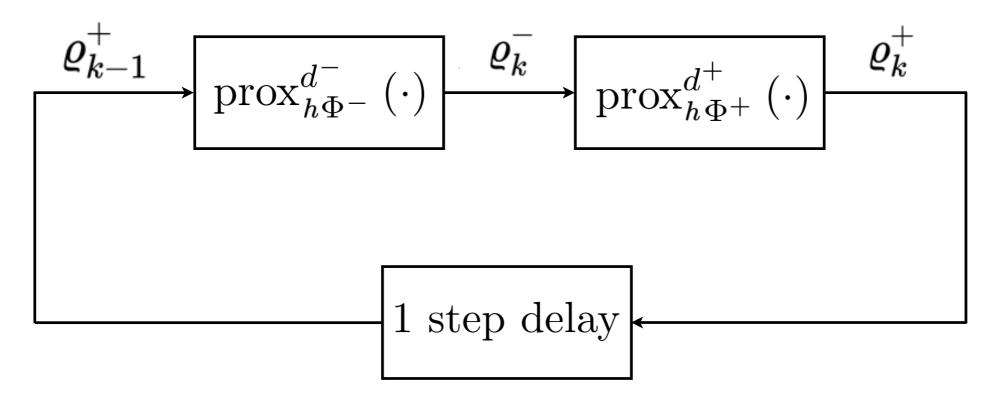
Solving filtering as Wasserstein gradient flow

What's New?

Main idea: Solve the Kushner-Stratonovich SPDE

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

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:

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

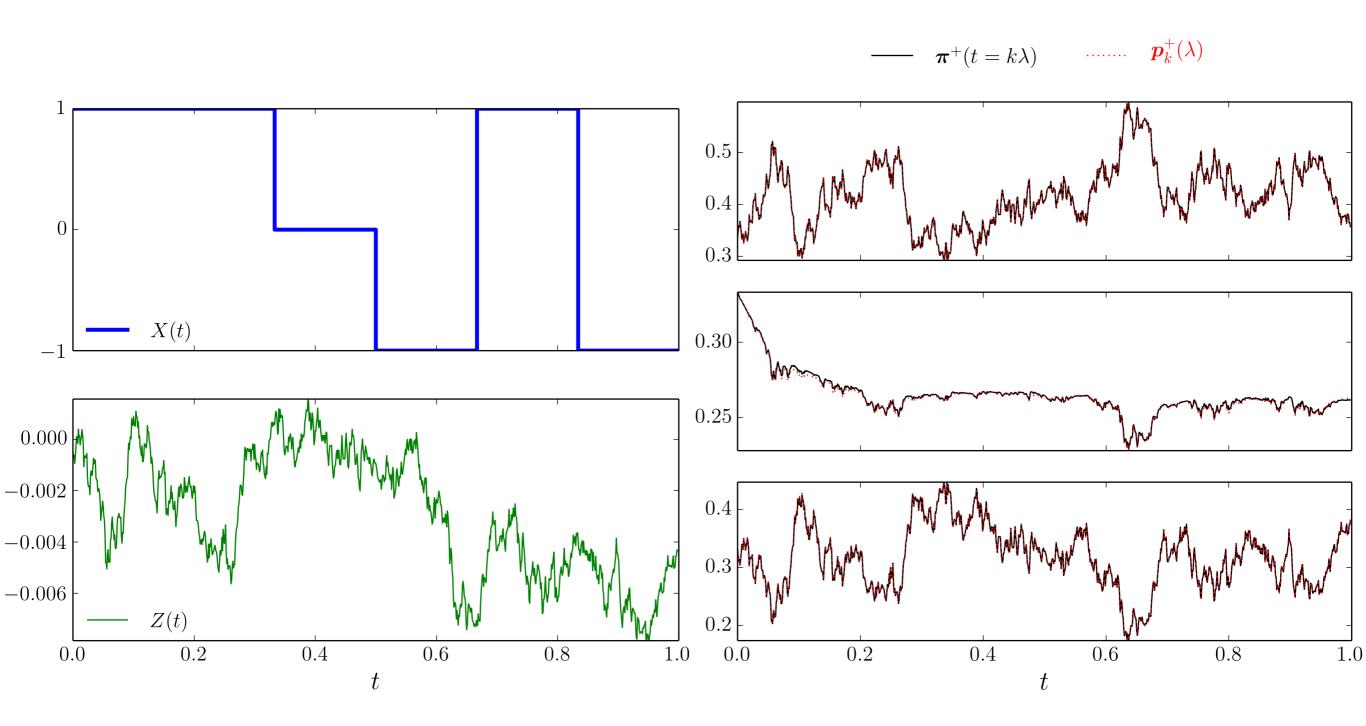
where
$$H:=\operatorname{diag}(h(a_1),\ldots,h(a_m)), \quad \widehat{h}(t):=\sum_{i=1}^m h(a_i)\pi_i^+(t),$$

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.

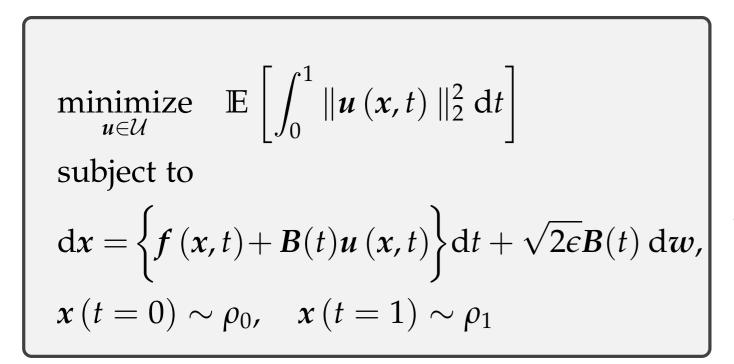
Numerical Results for the Wonham Filter

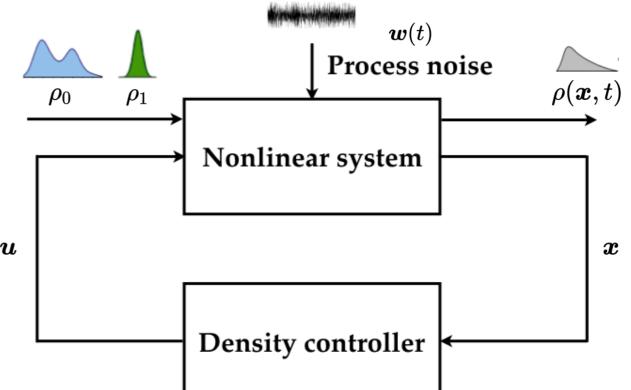


— A.H. and T.T. Georgiou, Proximal Recursion for the Wonham Filter, CDC 2019.

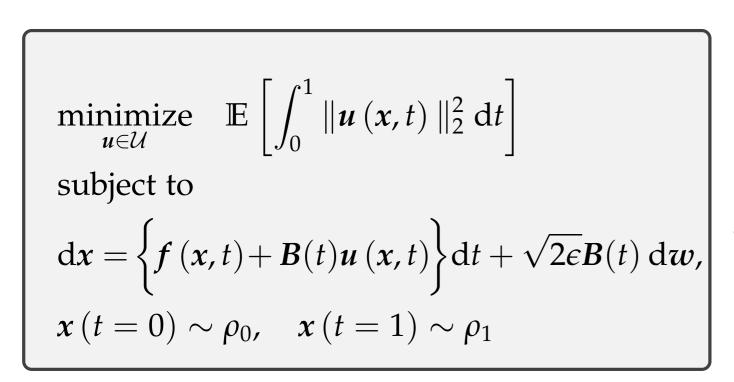
Solving density control as Wasserstein gradient flow

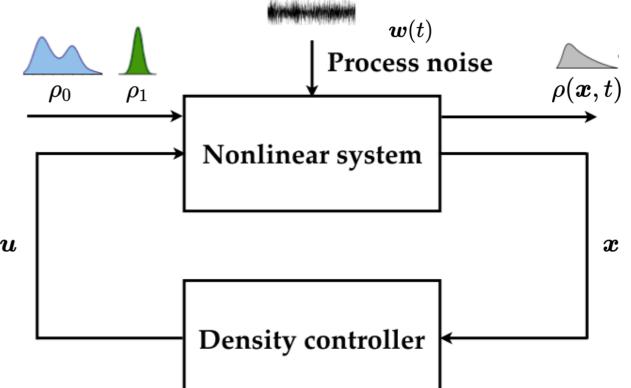
Finite Horizon Feedback Density Control





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:

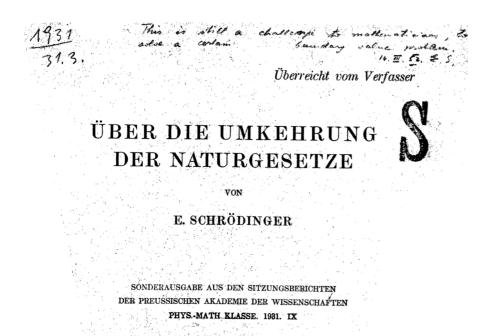
$$\rho^{\text{opt}}(x,0) = \rho_0(x), \quad \rho^{\text{opt}}(x,1) = \rho_1(x)$$

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

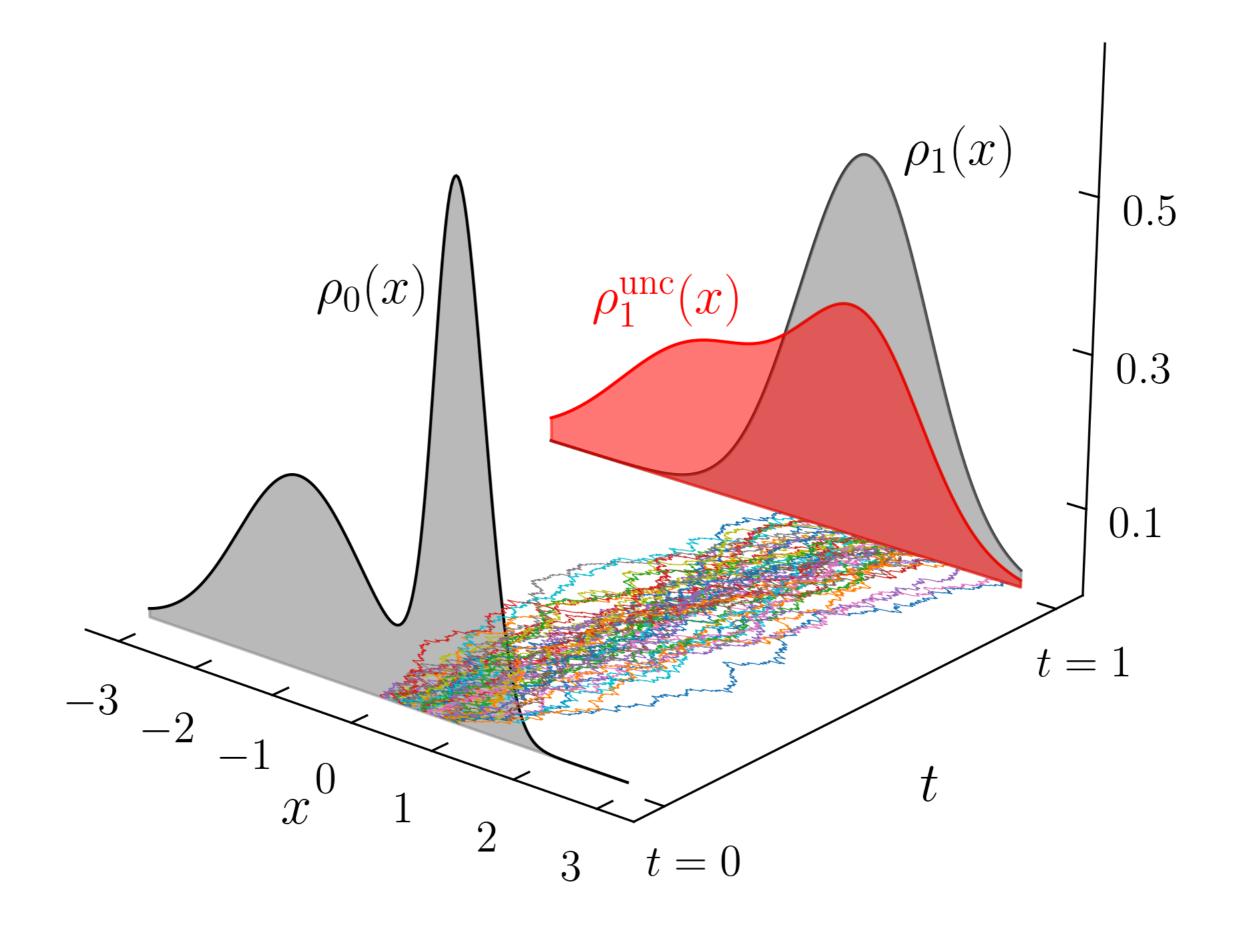
Wasserstein proximal algorithm \longrightarrow fixed point recursion over $(\hat{\varphi}_0, \varphi_1)$

(Contractive in Hilbert metric)

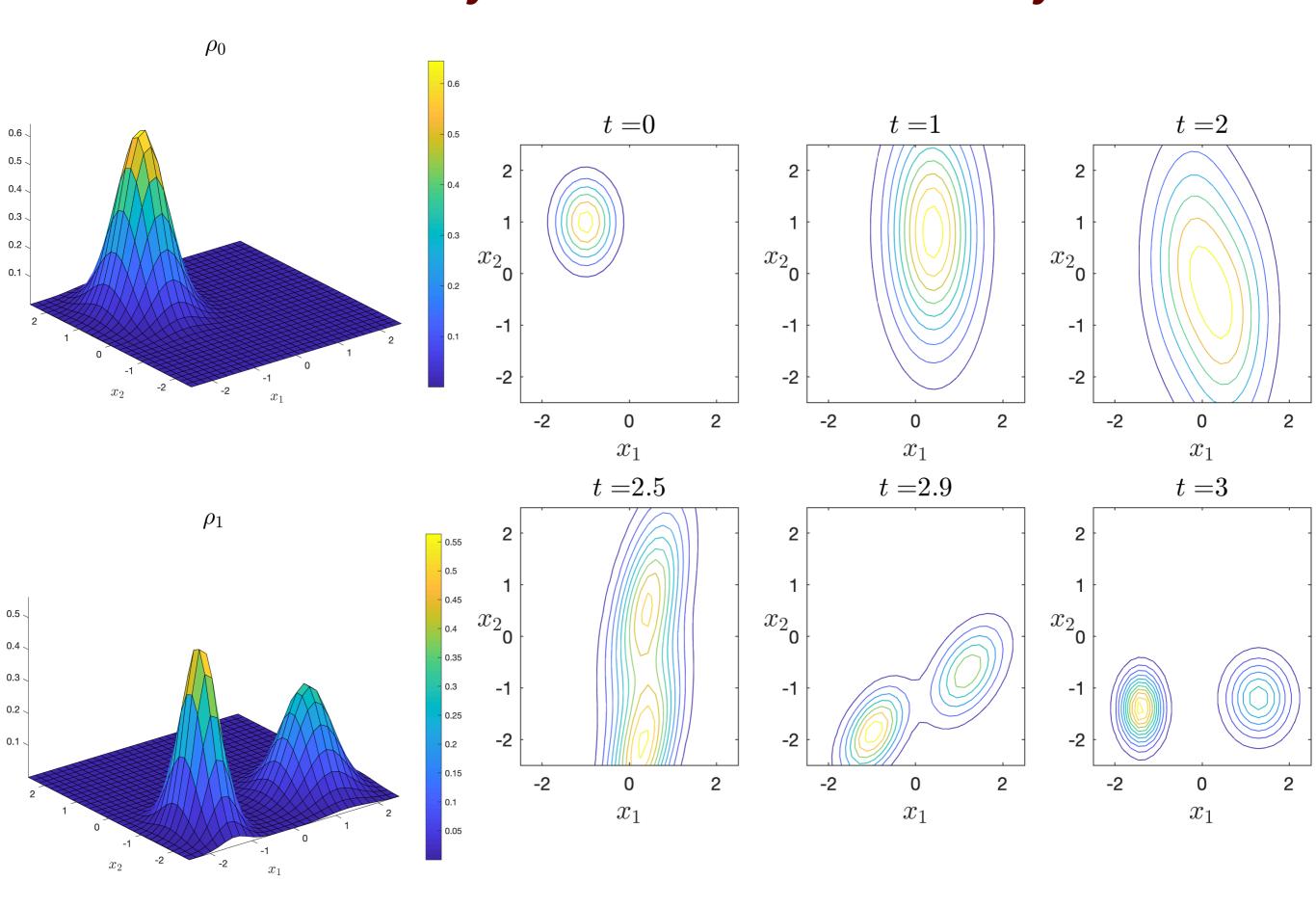
Fixed Point Recursion over $(\hat{\varphi}_0, \varphi_1)$

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

Feedback Density Control: Zero Prior Dynamics

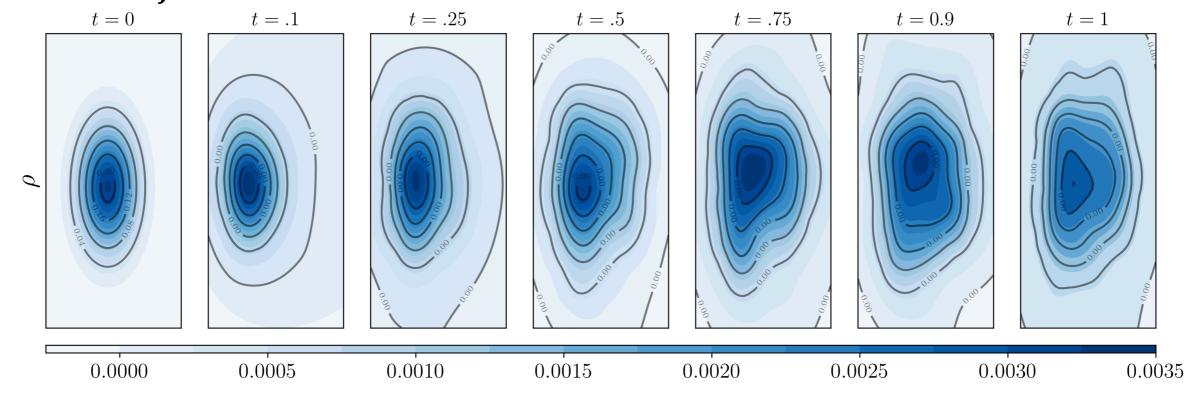


Feedback Density Control: LTI Prior Dynamics

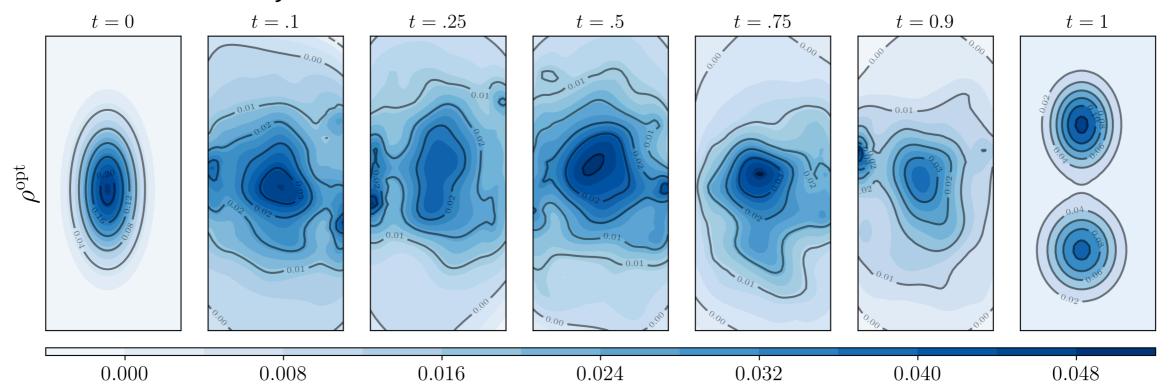


Feedback Density Control: Nonlinear Grad. Drift

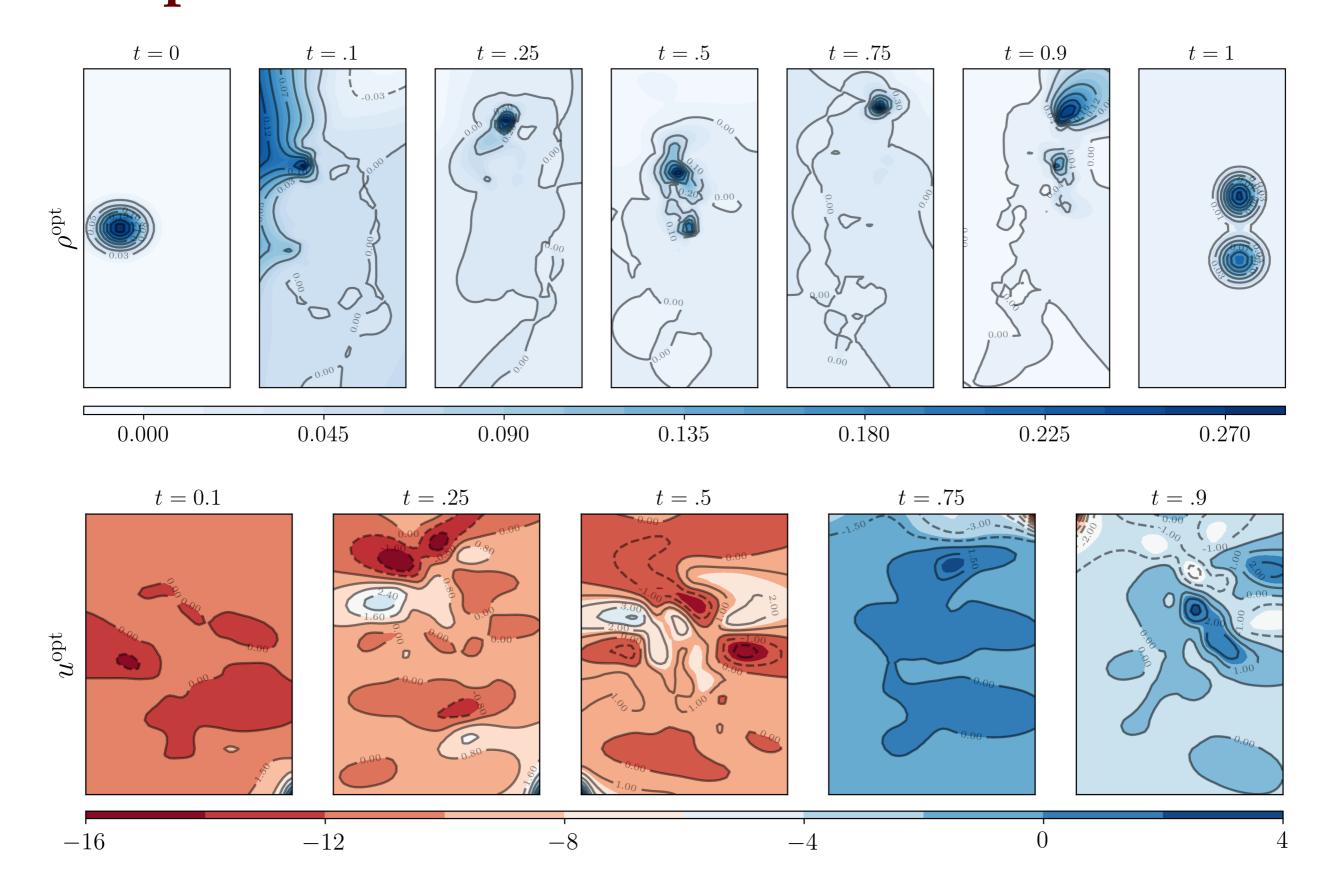
Uncontrolled joint PDF evolution:



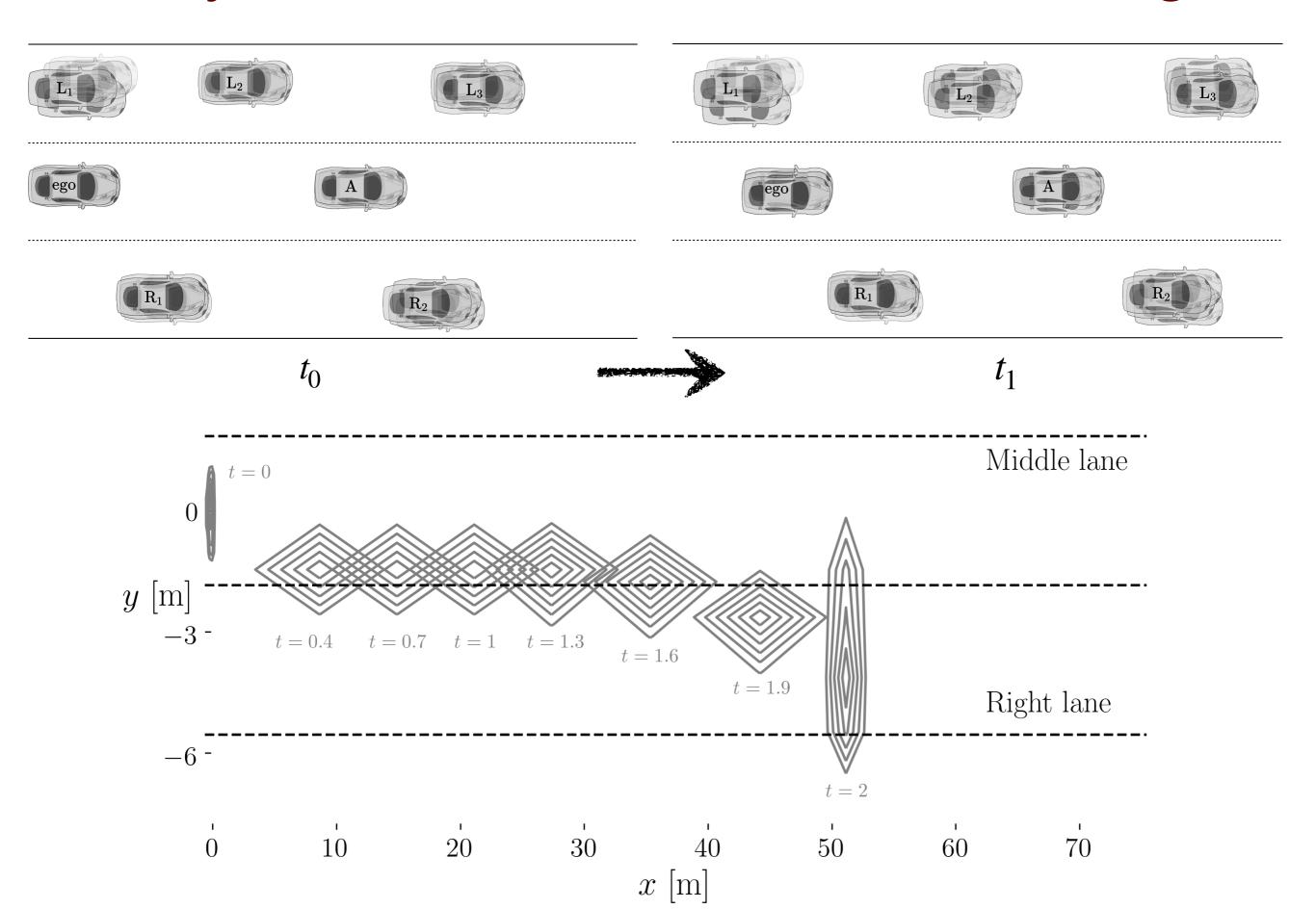
Optimal controlled joint PDF evolution:



Feedback Density Control: Mixed Conservative-Dissipative Drift



Density Control for Safe Automated Driving



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:

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

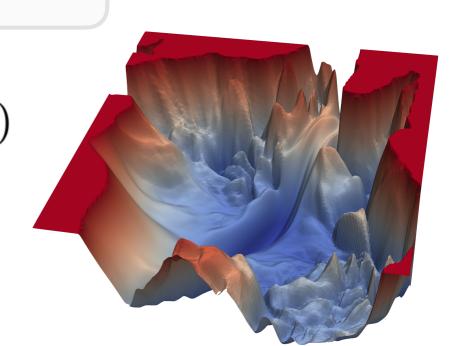
 $\text{Learning problem: } \min_{\boldsymbol{\theta} \in \mathbb{R}^{pn_{\text{H}}}} R(\hat{f}\,)$

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!!

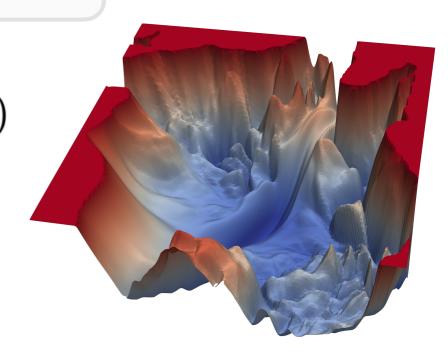


Learning Neural Network from Data

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

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:

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

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

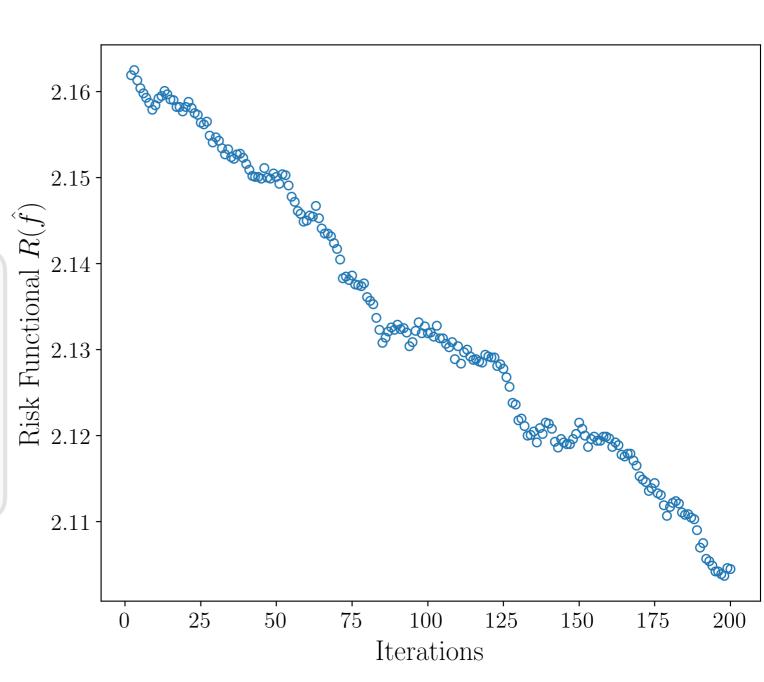
Joint law of $(\boldsymbol{x},y) \in \mathbb{R}^d \times \mathbb{R}$:

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

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

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

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



Take Home Message



Thank You







