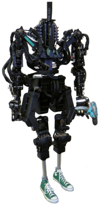
ARes and MaRS - Adversarial and MMD-Minimizing Regression for SDEs

Gabriele Abbati[*1], Philippe Wenk[*23], Michael A Osborne[1], Andreas Krause[2], Bernhard Schölkopf[4], Stefan Bauer[4]

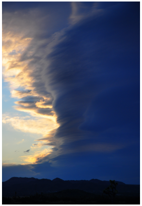
[1]University of Oxford, [2]ETH Zürich, [3]Max Planck ETH Center for Learning Systems, [4]Max Planck Institute for Intelligent Systems

Thirty-sixth International Conference on Machine Learning

Stochastic Differential Equations in the Wild



(a) **Robotics** (source: Athena robot, MPI-IS)



(b) **Atmospheric Modeling** (source: wikipedia)

(c) **Stock Markets** (source: Yahoo Finance)

Gradient Matching

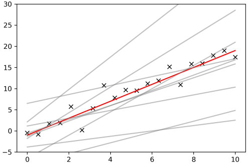
ODE $\begin{cases} \dot{\mathbf{x}} = \mathbf{f}(\mathbf{x}, \theta) \\ \mathbf{y} = \mathbf{x} + \epsilon \quad \text{with } \epsilon \sim \mathcal{N}(\mathbf{0}, \sigma_{\mathbf{y}}) \end{cases}$

Given f and y, infer x and θ

SDE $\begin{cases} \mathrm{d}\mathbf{x} = \mathbf{f}(\mathbf{x}, \theta)\mathrm{d}t + \mathbf{G}\mathrm{d}\mathbf{W} \\ \mathbf{y} = \mathbf{x} + \epsilon \quad \text{with } \epsilon \sim \mathcal{N}(\mathbf{0}, \sigma_{\mathbf{y}}) \end{cases}$

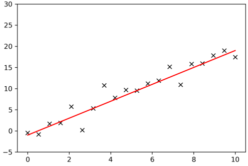
Given f and y, infer x, G and θ

Integration-based methods



parameters $\rightarrow$ trajectory

Integration-free methods

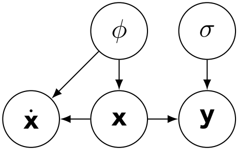


trajectory $\rightarrow$ parameters

Classic Gradient Matching - Model

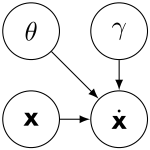
(1) Gaussian Process prior on states

$$p(\mathbf{x} \mid \phi) = \mathcal{N}(\mathbf{x} \mid \boldsymbol{\mu}_{\mathbf{y}}, \mathbf{C}_{\phi})$$
$$p(\dot{\mathbf{x}} \mid \mathbf{x}, \phi) = \mathcal{N}(\dot{\mathbf{x}} \mid \mathbf{D}\mathbf{x}, \mathbf{A})$$



(2) ODE Model

$$p(\dot{\mathbf{x}} \mid \mathbf{x}, \boldsymbol{\theta}, \gamma) = \mathcal{N}(\dot{\mathbf{x}} \mid \mathbf{f}(\mathbf{x}, \boldsymbol{\theta}), \gamma \mathbf{I})$$



Classic Gradient Matching - Inference

- Calderhead, Girolami, and Lawrence (2009) and Dondelinger et al. (2013)

  Product of Experts: $p(\dot{\mathbf{x}}) \propto p_{\text{data}}(\dot{\mathbf{x}}) p_{\text{ODE}}(\dot{\mathbf{x}})$

- Wenk et al. (2018), FGPGM

  Forced equality:
  $$p(\dot{\mathbf{x}}) \propto p_{\text{data}}(\dot{\mathbf{x}}_{\text{data}}) p_{ODE}(\dot{\mathbf{x}}_{\text{ODE}}) \delta(\dot{\mathbf{x}}_{\text{data}} - \dot{\mathbf{x}}) \delta(\dot{\mathbf{x}}_{\text{ODE}} - \dot{\mathbf{x}})$$

- Wenk*, Abbati* et al. (2019), ODIN

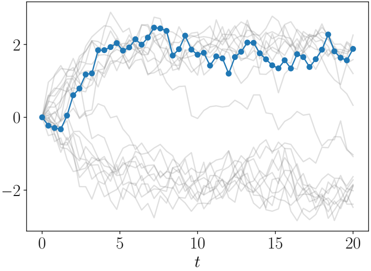
  ODEs as constraints

Stochastic Differential Equations

General SDE Problem

$$\begin{cases} d\mathbf{x} = \mathbf{f}(\mathbf{x}, \theta)dt + \mathbf{G}d\mathbf{W} \\ \mathbf{y} = \mathbf{x} + \epsilon \quad \text{with } \epsilon \sim \mathcal{N}(\mathbf{0}, \sigma_{\mathbf{y}}) \end{cases}$$

Given f and **y**, **infer x**, **G** and $\boldsymbol{\theta}$



Example

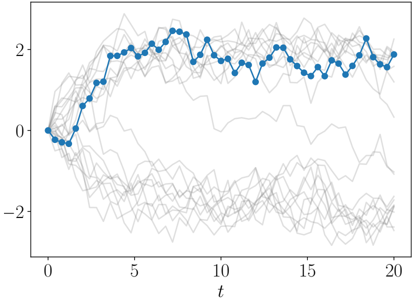
$$\begin{cases} dx = \theta_0 x(\theta_1 - x^2)dt + Gdw \\ \mathbf{y} = \mathbf{x} + \epsilon \quad \text{with } \epsilon \sim \mathcal{N}(\mathbf{0}, \sigma_{\mathbf{y}}) \end{cases}$$

Given f and **y**, **infer x**, G and $\boldsymbol{\theta}$

Stochastic Gradient Matching?

Problems:

- Both **observation** and **process** noise
- **Stochastic** sample paths
- Paths are **not differentiable**



The Doss-Sussmann Transformation

General SDE Problem

$$\begin{cases} d\mathbf{x} = \mathbf{f}(\mathbf{x}, \theta)dt + \mathbf{G}d\mathbf{W} \\ \mathbf{y} = \mathbf{x} + \epsilon \quad \text{with } \epsilon \sim \mathcal{N}(\mathbf{0}, \sigma_{\mathbf{y}}) \end{cases}$$

Given $\mathbf{f}$ and $\mathbf{y}$, infer $\mathbf{x}$, $\mathbf{G}$ and θ

Definition (Ornstein-Uhlenbeck Process) *A stochastic process* $\mathbf{o}$ *defined by the equation:* $d\mathbf{o} = -\mathbf{o}dt + \mathbf{G}d\mathbf{W}$

We introduce the latent variable

$$\mathbf{z} = \mathbf{x} - \mathbf{o}$$

to get the **stochastic gradients**

$$d\mathbf{z}(t) = \{\mathbf{f}(\mathbf{z}(t) + \mathbf{o}(t), \theta) + \mathbf{o}(t)\}\, dt$$

The Doss-Sussmann Transformation

General SDE Problem

$$\begin{cases} \mathrm{d}\mathbf{x} = \mathbf{f}(\mathbf{x}, \theta)\mathrm{d}t + \mathbf{G}\mathrm{d}\mathbf{W} \\ \mathbf{y} = \mathbf{x} + \epsilon \quad \text{with } \epsilon \sim \mathcal{N}(\mathbf{0}, \sigma_{\mathbf{y}}) \end{cases}$$

Given $\mathbf{f}$ and $\mathbf{y}$, infer $\mathbf{x}$, $\mathbf{G}$ and θ

Definition (Ornstein-Uhlenbeck Process) *A stochastic process* $\mathbf{o}$ *defined by the equation:* $\mathrm{d}\mathbf{o} = -\mathbf{o}\mathrm{d}t + \mathbf{G}\mathrm{d}\mathbf{W}$

We introduce the latent variable

$$\mathbf{z} = \mathbf{x} - \mathbf{o}$$

to get the **stochastic gradients**

$$\mathrm{d}\mathbf{z}(t) = \{\mathbf{f}(\mathbf{z}(t) + \mathbf{o}(t), \theta) + \mathbf{o}(t)\}\,\mathrm{d}t$$

The Doss-Sussmann Transformation

General SDE Problem

$$\begin{cases} \mathrm{d}\mathbf{x} = \mathbf{f}(\mathbf{x}, \theta)\mathrm{d}t + \mathbf{G}\mathrm{d}\mathbf{W} \\ \mathbf{y} = \mathbf{x} + \epsilon \quad \text{with } \epsilon \sim \mathcal{N}(\mathbf{0}, \sigma_{\mathbf{y}}) \end{cases}$$

Given f and **y**, **infer x**, **G** and θ

Definition (Ornstein-Uhlenbeck Process) *A stochastic process* $\mathbf{o}$ *defined by the equation:* $\mathrm{d}\mathbf{o} = -\mathbf{o}\mathrm{d}t + \mathbf{G}\mathrm{d}\mathbf{W}$

We introduce the latent variable

$$\mathbf{z} = \mathbf{x} - \mathbf{o}$$

to get the **stochastic gradients**

$$\mathrm{d}\mathbf{z}(t) = \{\mathbf{f}(\mathbf{z}(t) + \mathbf{o}(t), \theta) + \mathbf{o}(t)\}\, \mathrm{d}t$$

A Novel Generative Model

Previous Generative Model

$$\mathbf{Y} = \mathbf{X} + \mathbf{E}$$

$\longrightarrow$

New Generative Model

$$\mathbf{Y} = \mathbf{Z} + \mathbf{O} + \mathbf{E}$$

Resulting observation marginal distribution:

$$p(\tilde{\mathbf{y}} \mid \phi, \mathbf{G}, \sigma) = \mathcal{N}(\mathbf{0}, \mathbf{C}_\phi + \mathbf{B}\Omega\mathbf{B}^T + \mathbf{T})$$

Gaussian prior **OU process** **Obs. noise**

A Novel Generative Model

Previous Generative Model → **New Generative Model**

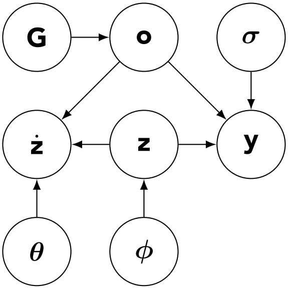
$$\mathbf{Y} = \mathbf{X} + \mathbf{E} \quad \longrightarrow \quad \mathbf{Y} = \mathbf{Z} + \mathbf{O} + \mathbf{E}$$

Resulting observation marginal distribution:

$$p(\tilde{\mathbf{y}} \mid \phi, \mathbf{G}, \sigma) = \mathcal{N}(\mathbf{0}, \underbrace{\mathbf{C}_\phi}_{\text{Gaussian prior}} + \underbrace{\mathbf{B}\Omega\mathbf{B}^T}_{\text{OU process}} + \underbrace{\mathbf{T}}_{\text{Obs. noise}})$$

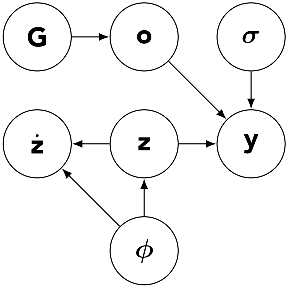
A Tale of Two Graphical Models

SDE-based model



$$p(\dot{\mathbf{z}} \mid \mathbf{o}, \mathbf{z}, \theta) = \delta(\dot{\mathbf{z}} - \mathbf{f}(\mathbf{z} + \mathbf{o}, \theta) - \mathbf{o})$$

Data-based model

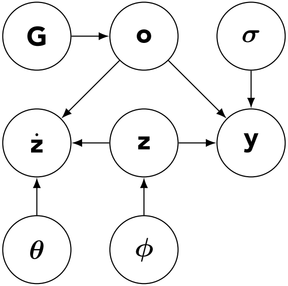


$$p(\dot{\mathbf{z}} \mid \mathbf{z}, \phi) = \mathcal{N}(\dot{\mathbf{z}} \mid \mathbf{D}\mathbf{z}, \mathbf{A})$$

Good θ estimate $\longrightarrow$ $p(\dot{\mathbf{z}} \mid \mathbf{o}, \mathbf{z}, \theta) \sim p(\dot{\mathbf{z}} \mid \mathbf{z}, \phi)$

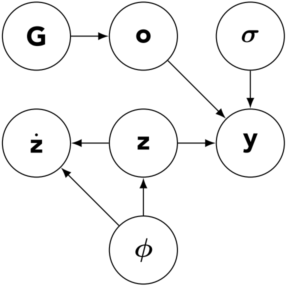
A Tale of Two Graphical Models

SDE-based model



$$p(\dot{\mathbf{z}} \mid \mathbf{o}, \mathbf{z}, \theta) = \delta(\dot{\mathbf{z}} - \mathbf{f}(\mathbf{z} + \mathbf{o}, \theta) - \mathbf{o})$$

Data-based model



$$p(\dot{\mathbf{z}} \mid \mathbf{z}, \phi) = \mathcal{N}(\dot{\mathbf{z}} \mid \mathbf{D}\mathbf{z}, \mathbf{A})$$

Good θ estimate $\longrightarrow$ $p(\dot{\mathbf{z}} \mid \mathbf{o}, \mathbf{z}, \theta) \sim p(\dot{\mathbf{z}} \mid \mathbf{z}, \phi)$

Sample-based Parameter Inference

GP fit
(ZOE noise model)

Samples from p_{SDE}

$\dot{\mathbf{z}}_{\text{SDE}} \sim p(\dot{\mathbf{z}} \mid \mathbf{o}, \mathbf{z}, \theta)$

Samples from p_{data}

$\dot{\mathbf{z}}_{\text{data}} \sim p(\dot{\mathbf{z}} \mid \mathbf{z}, \phi)$

Iterative
Gradient-based
optimization
$\rightarrow \dot{\mathbf{z}}_{\text{SDE}} \sim \dot{\mathbf{z}}_{\text{data}}$

(1) AReS (WGAN) $\theta \leftarrow -\nabla_\theta \left[\frac{1}{M} \sum_{i=1}^{M} f_\omega(\dot{\mathbf{z}}_{\text{SDE}}^{(i)})\right]$

(2) MaRS (MMD) $\theta \leftarrow \nabla_\theta MMD_u^2 \left[\dot{\mathbf{z}}_{\text{SDE}}, \dot{\mathbf{z}}_{\text{data}}\right]$

Sample-based Parameter Inference

GP fit
(ZOE noise model)

Samples from p_{SDE}

$\dot{\mathbf{z}}_{\text{SDE}} \sim p(\dot{\mathbf{z}} \mid \mathbf{o}, \mathbf{z}, \boldsymbol{\theta})$

Samples from p_{data}

$\dot{\mathbf{z}}_{\text{data}} \sim p(\dot{\mathbf{z}} \mid \mathbf{z}, \phi)$

Iterative
Gradient-based
optimization
$\rightarrow \dot{\mathbf{z}}_{\text{SDE}} \sim \dot{\mathbf{z}}_{\text{data}}$

(1) AReS (WGAN) $\theta \leftarrow -\nabla_{\theta} \left[\frac{1}{M} \sum_{i=1}^{M} f_{\omega}(\dot{\mathbf{z}}_{\text{SDE}}^{(i)}) \right]$

(2) MaRS (MMD) $\theta \leftarrow \nabla_{\theta} MMD_u^2 \left[\dot{\mathbf{z}}_{\text{SDE}}, \dot{\mathbf{z}}_{\text{data}} \right]$

Sample-based Parameter Inference

GP fit
(ZOE noise model)

Samples from p_{SDE}

$\dot{\mathbf{z}}_{\text{SDE}} \sim p(\dot{\mathbf{z}} \mid \mathbf{o}, \mathbf{z}, \boldsymbol{\theta})$

Samples from p_{data}

$\dot{\mathbf{z}}_{\text{data}} \sim p(\dot{\mathbf{z}} \mid \mathbf{z}, \phi)$

Iterative
Gradient-based
optimization
$\rightarrow \dot{\mathbf{z}}_{\text{SDE}} \sim \dot{\mathbf{z}}_{\text{data}}$

(1) AReS (WGAN) $\theta \leftarrow -\nabla_{\theta} \left[\frac{1}{M} \sum_{i=1}^{M} f_{\omega}(\dot{\mathbf{z}}^{(i)}_{\text{SDE}}) \right]$

(2) MaRS (MMD) $\theta \leftarrow \nabla_{\theta} MMD_u^2 \left[\dot{\mathbf{z}}_{\text{SDE}}, \dot{\mathbf{z}}_{\text{data}} \right]$

Sample-based Parameter Inference

GP fit
(ZOE noise model)

Samples from p_{SDE}

$\dot{\mathbf{z}}_{\text{SDE}} \sim p(\dot{\mathbf{z}} \mid \mathbf{o}, \mathbf{z}, \boldsymbol{\theta})$

Samples from p_{data}

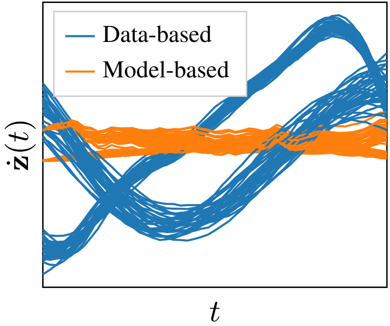
$\dot{\mathbf{z}}_{\text{data}} \sim p(\dot{\mathbf{z}} \mid \mathbf{z}, \phi)$

Iterative
Gradient-based
optimization
$\rightarrow \dot{\mathbf{z}}_{\text{SDE}} \sim \dot{\mathbf{z}}_{\text{data}}$

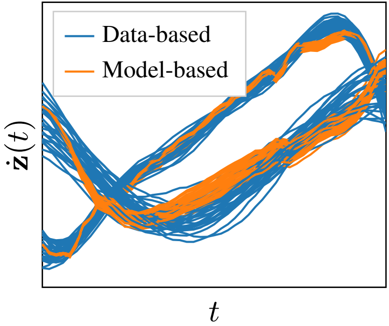
(1) AReS (WGAN) $\quad \boldsymbol{\theta} \leftarrow -\nabla_{\boldsymbol{\theta}} \left[\frac{1}{M} \sum_{i=1}^{M} f_{\boldsymbol{\omega}}(\dot{\mathbf{z}}_{\text{SDE}}^{(i)}) \right]$

(2) MaRS (MMD) $\quad \boldsymbol{\theta} \leftarrow \nabla_{\boldsymbol{\theta}} MMD_u^2 \left[\dot{\mathbf{z}}_{\text{SDE}}, \dot{\mathbf{z}}_{\text{data}} \right]$

Samples during Training



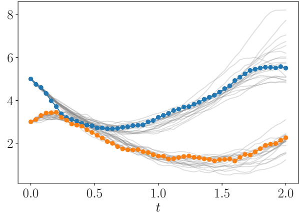
(a) Samples before training



(b) Samples after training

Experimental Results - Lotka Volterra

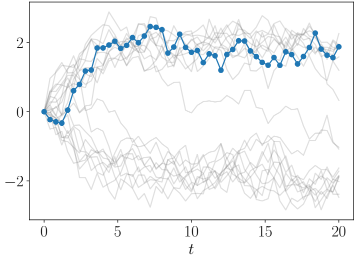
LV, GT	NPSDE	ESGF	AReS	MaRS
$\theta_0 = 2$	1.58 ± 0.71	2.04 ± 0.09	2.36 ± 0.18	$\mathbf{2.00 \pm 0.09}$
$\theta_1 = 1$	0.74 ± 0.31	1.02 ± 0.05	1.18 ± 0.9	$\mathbf{1.00 \pm 0.04}$
$\theta_2 = 4$	2.26 ± 1.51	3.87 ± 0.59	$\mathbf{3.97 \pm 0.63}$	3.70 ± 0.51
$\theta_3 = 1$	0.49 ± 0.35	0.96 ± 0.14	$\mathbf{0.98 \pm 0.18}$	0.91 ± 0.14
$\mathbf{H}_{1,1} = 0.05$	/	0.01 ± 0.03	$\mathbf{0.03 \pm 0.004}$	
$\mathbf{H}_{1,2} = 0.03$	/	0.01 ± 0.01	$\mathbf{0.02 \pm 0.01}$	
$\mathbf{H}_{2,1} = 0.03$	/	0.01 ± 0.01	$\mathbf{0.02 \pm 0.01}$	
$\mathbf{H}_{2,2} = 0.09$	/	0.03 ± 0.02	$\mathbf{0.09 \pm 0.03}$	



$$
\begin{aligned}
dx_1(t) &= [\theta_1 x_1(t) - \theta_2 x_1(t) x_2(t)]dt \\
&\quad + G_{11} dw_1(t) \\
dx_2(t) &= [-\theta_3 x_2(t) + \theta_4 x_1(t) x_2(t)]dt \\
&\quad + G_{21} dw_1(t) + G_{22} dw_2(t)
\end{aligned}
$$

Experimental Results - Double Well Potential

DW, GT	NPSDE	VGPA	ESGF	AReS	MaRS
$\theta_0 = 0.1$	0.09 ± 7.00	0.05 ± 0.04	0.01 ± 0.03	0.09 ± 0.04	$\mathbf{0.10 \pm 0.05}$
$\theta_1 = 4$	3.36 ± 248.82	1.11 ± 0.66	0.11 ± 0.16	3.68 ± 1.34	$\mathbf{3.85 \pm 1.10}$
$H = 0.25$	0.00 ± 0.02	/	0.20 ± 0.05	$\mathbf{0.21 \pm 0.09}$	



$$\mathrm{d}x(t) = \theta_0 x(\theta_1 - x^2)\mathrm{d}t + G\mathrm{d}w(t)$$

Contributions

- We extend classical gradient matching to SDEs
- We introduce a novel statistical framework combining the Doss-Sussmann transformation and GPs
- We introduce a novel parameter inference scheme that leverages adversarial and moment matching loss functions
- We improve parameter inference accuracy in systems of SDEs

Thank you

Come and catch us → poster #216

Bonus Round: check out our paper on classic gradient matching!
Wenk*, P., Abbati*, G., Bauer, S., Osborne, M. A., Krause, A., Schölkopf, B. (2019). **ODIN: ODE-Informed Regression for Parameter and State Inference in Time-Continuous Dynamical Systems**. ArXiv Preprint ArXiv:1902.06278.