Variational Wasserstein Gradient Flow

Presented at
Kantorovich Initiative Retreat, Univesity of Washington, Seattle

Amirhossein Taghvaei

Joint work with J. Fan, Y. Chen

Department of Aeronautics & Astronautics
University of Washington, Seattle

March 18, 2022

Background about myself

September 2021-now:

- Assistant Professor
  Department of Aeronautics & Astronautics

2019-2021

- Postdoctoral Scholar
  University of California, Irvine
  Supervisor: Tryphon Georgiou

UCI media coverage

2013-2019

- Ph.D. in Mechanical Engineering
  University of Illinois at Urbana-Champaign
  Ph.D. advisor: Prashant Mehta

Coordinated Science Laboratory

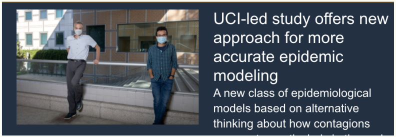
Background about myself

September 2021-now:

- Assistant Professor
  Department of Aeronautics & Astronautics

2019-2021

- Postdoctoral Scholar
  University of California, Irvine
  Supervisor: Tryphon Georgiou



UCI media coverage

2013-2019

- Ph.D. in Mechanical Engineering
  University of Illinois at Urbana-Champaign
  Ph.D. advisor: Prashant Mehta

Coordinated Science Laboratory

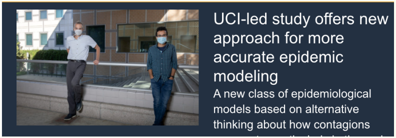
Background about myself

September 2021-now:

- Assistant Professor
  Department of Aeronautics & Astronautics

2019-2021

- Postdoctoral Scholar
  University of California, Irvine
  Supervisor: Tryphon Georgiou



UCI media coverage

2013-2019

- Ph.D. in Mechanical Engineering
  University of Illinois at Urbana-Champaign
  Ph.D. advisor: Prashant Mehta


Coordinated Science Laboratory

Research overveiw

Control & optimization for probability distributions

(I) Optimal filtering & control

- Optimal transportation methods in nonlinear filtering: The feedback particle filter, CSM, 2021
- An optimal transport formulation of the ensemble Kalman filter, TAC, 2021

(III) Stochastic thermodynamics

- Energy harvesting from anisotropic fluctuations, PRE, 2021
- On the relation between information and power in stochastic thermodynamic engines, (L-CSS), 2021
- Maximal power output of a stochastic thermodynamic engine, Automatica, 2021

(II) Machine learning

- OT mapping via input-convex neural networks, ICML, 2020
- Scalable computations of Wasserstein barycenter via input convex neural networks, ICML, 2021
- Variational Wasserstein gradient flow, Submitted to ICML, 2022

Common objectives:

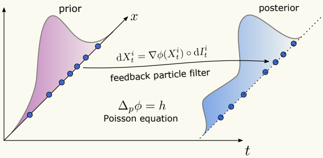
- develop efficient and scalable algorithms
- understand fundamental limitations

Theoretical theme:

- optimal transportation
- (mean-field) optimal control

Research overveiw
Control & optimization for probability distributions

(I) Optimal filtering & control



- Optimal transportation methods in nonlinear filtering: The feedback particle filter, CSM, 2021
- An optimal transport formulation of the ensemble Kalman filter, TAC, 2021

(II) Machine learning

- OT mapping via input-convex neural networks, ICML, 2020
- Scalable computations of Wasserstein barycenter via input convex neural networks, ICML, 2021
- Variational Wasserstein gradient flow, Submitted to ICML, 2022

(III) Stochastic thermodynamics

- Energy harvesting from anisotropic fluctuations, PRE, 2021
- On the relation between information and power in stochastic thermodynamic engines, (L-CSS), 2021
- Maximal power output of a stochastic thermodynamic engine, Automatica, 2021

Common objectives:

- develop efficient and scalable algorithms
- understand fundamental limitations

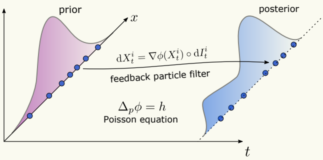
Theoretical theme:

- optimal transportation
- (mean-field) optimal control

Research overveiw

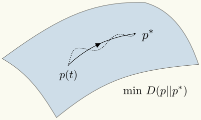
Control & optimization for probability distributions

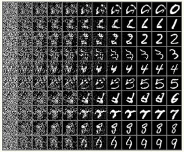
(I) Optimal filtering & control



- Optimal transportation methods in nonlinear filtering: The feedback particle filter, CSM, 2021
- An optimal transport formulation of the ensemble Kalman filter, TAC, 2021

(II) Machine learning





- OT mapping via input-convex neural networks, ICML, 2020
- Scalable computations of Wasserstein barycenter via input convex neural networks, ICML, 2021
- Variational Wasserstein gradient flow, Submitted to ICML, 2022

(III) Stochastic thermodynamics

- Energy harvesting from anisotropic fluctuations, PRE, 2021
- On the relation between information and power in stochastic thermodynamic engines, (L-CSS), 2021
- Maximal power output of a stochastic thermodynamic engine, Automatica, 2021

Common objectives:

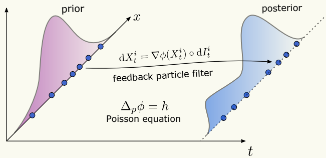
- develop efficient and scalable algorithms
- understand fundamental limitations

Theoretical theme:

- optimal transportation
- (mean-field) optimal control

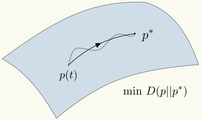
Research overveiw
Control & optimization for probability distributions

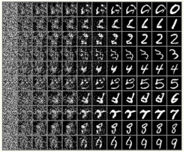
(I) Optimal filtering & control



- Optimal transportation methods in nonlinear filtering: The feedback particle filter, CSM, 2021
- An optimal transport formulation of the ensemble Kalman filter, TAC, 2021

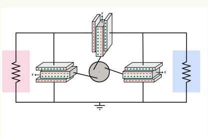
(II) Machine learning

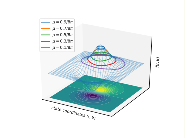




- OT mapping via input-convex neural networks, ICML, 2020
- Scalable computations of Wasserstein barycenter via input convex neural networks, ICML, 2021
- Variational Wasserstein gradient flow, Submitted to ICML, 2022

(III) Stochastic thermodynamics





- Energy harvesting from anisotropic fluctuations, PRE, 2021
- On the relation between information and power in stochastic thermodynamic engines, (L-CSS), 2021
- Maximal power output of a stochastic thermodynamic engine, Automatica, 2021

Common objectives:

- develop efficient and scalable algorithms
- understand fundamental limitations

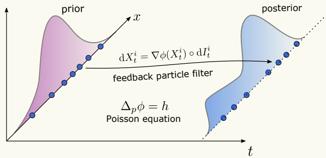
Theoretical theme:

- optimal transportation
- (mean-field) optimal control

Research overveiw

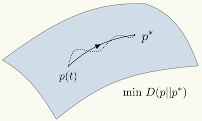
Control & optimization for probability distributions

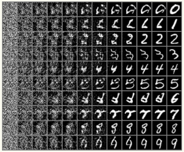
(I) Optimal filtering & control



- Optimal transportation methods in nonlinear filtering: The feedback particle filter, CSM, 2021
- An optimal transport formulation of the ensemble Kalman filter, TAC, 2021

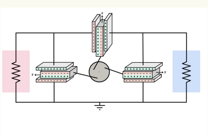
(II) Machine learning

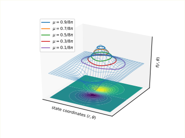




- OT mapping via input-convex neural networks, ICML, 2020
- Scalable computations of Wasserstein barycenter via input convex neural networks, ICML, 2021
- Variational Wasserstein gradient flow, Submitted to ICML, 2022

(III) Stochastic thermodynamics





- Energy harvesting from anisotropic fluctuations, PRE, 2021
- On the relation between information and power in stochastic thermodynamic engines, (L-CSS), 2021
- Maximal power output of a stochastic thermodynamic engine, Automatica, 2021

Common objectives:

- develop efficient and scalable algorithms
- understand fundamental limitations

Theoretical theme:

- optimal transportation
- (mean-field) optimal control

Outline

- Overview of numerical methods to implement Wasserstein gradient flows
- Variational approach

Outline

- Overview of numerical methods to implement Wasserstein gradient flows
- Variational approach

Motivation

- Many machine learning problems are formulated as an optimization problem on the space of probability distributions (e.g. sampling, GAN, policy optimization)

- Optimal transportation theory provides geometrical tools (i.e. Riemannian metric) to employ optimization methods for such problems

- This talk: numerical implementation of Wasserstein gradient flows

Related works:

- pde-based approach (Peyre, 2015; Benamou et al., 2016; Carlier et al., 2017; Li et al., 2020; Carrillo et al., 2021)
- JKO scheme + ICNN (Mokrov et al., 2021; Alvarez-Melis et al., 2021; Yang et al., 2020; Bunne et al., 2021; Bonet et al., 2021)
- kernel methods (Liu & Wang, 2016; Chewi et al., 2020; Korba et al. 2021)
- . . .

Motivation

- Many machine learning problems are formulated as an optimization problem on the space of probability distributions (e.g. sampling, GAN, policy optimization)
- Optimal transportation theory provides geometrical tools (i.e. Riemannian metric) to employ optimization methods for such problems
- This talk: numerical implementation of Wasserstein gradient flows

Related works:

- pde-based approach (Peyre, 2015; Benamou et al., 2016; Carlier et al., 2017; Li et al., 2020; Carrillo et al., 2021)
- JKO scheme + ICNN (Mokrov et al., 2021; Alvarez-Melis et al., 2021; Yang et al., 2020; Bunne et al., 2021; Bonet et al., 2021)
- kernel methods (Liu & Wang, 2016; Chewi et al., 2020; Korba et al. 2021)
- . . .

Motivation

- Many machine learning problems are formulated as an optimization problem on the space of probability distributions (e.g. sampling, GAN, policy optimization)
- Optimal transportation theory provides geometrical tools (i.e. Riemannian metric) to employ optimization methods for such problems
- This talk: numerical implementation of Wasserstein gradient flows

Related works:

- pde-based approach (Peyre, 2015; Benamou et al., 2016; Carlier et al., 2017; Li et al., 2020; Carrillo et al., 2021)
- JKO scheme + ICNN (Mokrov et al., 2021; Alvarez-Melis et al., 2021; Yang et al., 2020; Bunne et al., 2021; Bonet et al., 2021)
- kernel methods (Liu & Wang, 2016; Chewi et al., 2020; Korba et al. 2021)
- ...

Motivation

- Many machine learning problems are formulated as an optimization problem on the space of probability distributions (e.g. sampling, GAN, policy optimization)

- Optimal transportation theory provides geometrical tools (i.e. Riemannian metric) to employ optimization methods for such problems

- This talk: numerical implementation of Wasserstein gradient flows

Related works:

- pde-based approach (Peyre, 2015; Benamou et al., 2016; Carlier et al., 2017; Li et al., 2020; Carrillo et al., 2021)

- JKO scheme + ICNN (Mokrov et al., 2021; Alvarez-Melis et al., 2021; Yang et al., 2020; Bunne et al., 2021; Bonet et al., 2021)

- kernel methods (Liu & Wang, 2016; Chewi et al., 2020; Korba et al. 2021)

- ...

Wasserstein gradient flow

- Optimization problem:

$$\min_{p \in \mathcal{P}_2(\mathbb{R}^n)} F(p)$$

- Wasserstein gradient flow:

$$\frac{\partial p}{\partial t} = \nabla \cdot (p \nabla \frac{\delta F}{\delta p})$$

where $\frac{\delta F}{\delta p}$ is the L_2-derivative.

- Example: $F(p) = D(p \| e^{-V})$ (KL divergence)

$$\frac{\partial p}{\partial t} = \nabla \cdot (p \nabla V) + \Delta p, \quad \text{(Fokker-Planck eq.)}$$

- How to numerically implement the Wasserstein gradient flow?
  - pde approach (does not scale with the dimension)
  - probabilistic approach (approximate with an empirical distribution of particles)

Wasserstein gradient flow

- Optimization problem:

$$\min_{p \in \mathcal{P}_2(\mathbb{R}^n)} F(p)$$

- Wasserstein gradient flow:

$$\frac{\partial p}{\partial t} = \nabla \cdot (p \nabla \frac{\delta F}{\delta p})$$

where $\frac{\delta F}{\delta p}$ is the L_2-derivative.

- Example: $F(p) = D(p \| e^{-V})$ (KL divergence)

$$\frac{\partial p}{\partial t} = \nabla \cdot (p \nabla V) + \Delta p, \quad \text{(Fokker-Planck eq.)}$$

- How to numerically implement the Wasserstein gradient flow?
  - pde approach (does not scale with the dimension)
  - probabilistic approach (approximate with an empirical distribution of particles)

Wasserstein gradient flow

- Optimization problem:

$$\min_{p \in \mathcal{P}_2(\mathbb{R}^n)} F(p)$$

- Wasserstein gradient flow:

$$\frac{\partial p}{\partial t} = \nabla \cdot (p \nabla \frac{\delta F}{\delta p})$$

where $\frac{\delta F}{\delta p}$ is the L_2-derivative.

- Example: $F(p) = D(p \| e^{-V})$ (KL divergence)

$$\frac{\partial p}{\partial t} = \nabla \cdot (p \nabla V) + \Delta p, \quad \text{(Fokker-Planck eq.)}$$

- How to numerically implement the Wasserstein gradient flow?
  - pde approach (does not scale with the dimension)
  - probabilistic approach (approximate with an empirical distribution of particles)

Wasserstein gradient flow

- Optimization problem:

$$\min_{p \in \mathcal{P}_2(\mathbb{R}^n)} F(p)$$

- Wasserstein gradient flow:

$$\frac{\partial p}{\partial t} = \nabla \cdot (p \nabla \frac{\delta F}{\delta p})$$

where $\frac{\delta F}{\delta p}$ is the L_2-derivative.

- Example: $F(p) = D(p \| e^{-V})$ (KL divergence)

$$\frac{\partial p}{\partial t} = \nabla \cdot (p \nabla V) + \Delta p, \quad \text{(Fokker-Planck eq.)}$$

- How to numerically implement the Wasserstein gradient flow?
  - pde approach (does not scale with the dimension)
  - probabilistic approach (approximate with an empirical distribution of particles)

Probabilistic approach

Objective: numerically implement the gradient flow $\frac{\partial p}{\partial t} = \nabla \cdot (p \nabla \frac{\delta F}{\delta p})$:

- Step 1: Construct a stochastic process $\{\bar{X}_t\}_{t \geq 0}$ s.t.

$$\text{Law}(\bar{X}_t) = p_t \quad \forall t \geq 0$$

- Step 2: Realize $\bar{X}_t$ with a system of (interacting) particles s.t. $\{X_t^1, \ldots, X_t^N\}$

$$\frac{1}{N} \sum_{i=1}^{N} \delta_{X_t^i} \approx \text{Law}(\bar{X}_t)$$

Questions:

- How to construct $\bar{X}_t$? — uniqueness issue
- How to realize with system of interacting particles? (approximating the mean-field terms that depend on density)
- Error analysis for particle approximation (propagation of chaos)

Probabilistic approach

Objective: numerically implement the gradient flow $\frac{\partial p}{\partial t} = \nabla \cdot (p \nabla \frac{\delta F}{\delta p})$:

- Step 1: Construct a stochastic process $\{\bar{X}_t\}_{t \geq 0}$ s.t.

$$\text{Law}(\bar{X}_t) = p_t \quad \forall t \geq 0$$

- Step 2: Realize $\bar{X}_t$ with a system of (interacting) particles s.t. $\{X_t^1, \ldots, X_t^N\}$

$$\frac{1}{N} \sum_{i=1}^{N} \delta_{X_t^i} \approx \text{Law}(\bar{X}_t)$$

Questions:

- How to construct $\bar{X}_t$? — uniqueness issue
- How to realize with system of interacting particles? (approximating the mean-field terms that depend on density)
- Error analysis for particle approximation (propagation of chaos)

Probabilistic approach

Objective: numerically implement the gradient flow $\frac{\partial p}{\partial t} = \nabla \cdot (p \nabla \frac{\delta F}{\delta p})$:

- Step 1: Construct a stochastic process $\{\bar{X}_t\}_{t \geq 0}$ s.t.

$$\text{Law}(\bar{X}_t) = p_t \quad \forall t \geq 0$$

- Step 2: Realize $\bar{X}_t$ with a system of (interacting) particles s.t. $\{X_t^1, \ldots, X_t^N\}$

$$\frac{1}{N} \sum_{i=1}^{N} \delta_{X_t^i} \approx \text{Law}(\bar{X}_t)$$

Questions:

- How to construct $\bar{X}_t$? — uniqueness issue
- How to realize with system of interacting particles? (approximating the mean-field terms that depend on density)
- Error analysis for particle approximation (propagation of chaos)

Probabilistic approach

Objective: numerically implement the gradient flow $\frac{\partial p}{\partial t} = \nabla \cdot (p \nabla \frac{\delta F}{\delta p})$:

- Step 1: Construct a stochastic process $\{\bar{X}_t\}_{t \geq 0}$ s.t.

$$\text{Law}(\bar{X}_t) = p_t \quad \forall t \geq 0$$

- Step 2: Realize $\bar{X}_t$ with a system of (interacting) particles s.t. $\{X_t^1, \ldots, X_t^N\}$

$$\frac{1}{N} \sum_{i=1}^{N} \delta_{X_t^i} \approx \text{Law}(\bar{X}_t)$$

Questions:

- How to construct $\bar{X}_t$? $\rightarrow$ uniqueness issue
- How to realize with system of interacting particles? (approximating the mean-field terms that depend on density)
- Error analysis for particle approximation (propagation of chaos)

Probabilistic approach

Objective: numerically implement the gradient flow $\frac{\partial p}{\partial t} = \nabla \cdot (p \nabla \frac{\delta F}{\delta p})$:

- Step 1: Construct a stochastic process $\{\bar{X}_t\}_{t\geq 0}$ s.t.

$$\text{Law}(\bar{X}_t) = p_t \quad \forall t \geq 0$$

- Step 2: Realize $\bar{X}_t$ with a system of (interacting) particles s.t. $\{X_t^1, \ldots, X_t^N\}$

$$\frac{1}{N}\sum_{i=1}^{N} \delta_{X_t^i} \approx \text{Law}(\bar{X}_t)$$

Questions:

- How to construct $\bar{X}_t$? $\rightarrow$ uniqueness issue
- How to realize with system of interacting particles? (approximating the mean-field terms that depend on density)
- Error analysis for particle approximation (propagation of chaos)

Uniqueness issue

Step 1: Given a gradient flow $\{p_t\}_{t\geq 0}$, construct a stochastic process $\{\bar{X}_t\}_{t\geq 0}$ s.t.

$$\text{Law}(\bar{X}_t) = p_t, \quad \forall t \geq 0$$

- No unique solution: two-time marginals are not specified $(\text{Law}(\bar{X}_{t_1}, \bar{X}_{t_2}) = ?)$

Example: Fokker-Planck eq. $\dfrac{\partial p}{\partial t} = \nabla \cdot (p \nabla V) + \Delta p,$

- Stochastic:

$$\mathrm{d}\bar{X}_t = -\nabla V(\bar{X}_t)\mathrm{d}t + \sqrt{2}\mathrm{d}B_t, \quad \bar{X}_0 \sim p_0$$

- Deterministic:

$$\dot{\bar{X}}_t = -\nabla V(\bar{X}_t) + \nabla \log \bar{p}_t(\bar{X}_t), \quad \bar{X}_0 \sim p_0$$

  where $\bar{p}_t = \text{Law}(\bar{X}_t)$

- Both systems lead to the same one-time marginal densities
- difference arises with particle approximation

Uniqueness issue

Step 1: Given a gradient flow $\{p_t\}_{t\geq 0}$, construct a stochastic process $\{\bar{X}_t\}_{t\geq 0}$ s.t.

$$\text{Law}(\bar{X}_t) = p_t, \quad \forall t \geq 0$$

- No unique solution: two-time marginals are not specified ($\text{Law}(\bar{X}_{t_1}, \bar{X}_{t_2}) = ?$)

Example: Fokker-Planck eq. $\frac{\partial p}{\partial t} = \nabla \cdot (p \nabla V) + \Delta p,$

- Stochastic:

$$\mathrm{d}\bar{X}_t = -\nabla V(\bar{X}_t)\mathrm{d}t + \sqrt{2}\mathrm{d}B_t, \quad \bar{X}_0 \sim p_0$$

- Deterministic:

$$\dot{\bar{X}}_t = -\nabla V(\bar{X}_t) + \nabla \log \bar{p}_t(\bar{X}_t), \quad \bar{X}_0 \sim p_0$$

where $\bar{p}_t = \text{Law}(\bar{X}_t)$

- Both systems lead to the same one-time marginal densities
- difference arises with particle approximation

Uniqueness issue

Step 1: Given a gradient flow $\{p_t\}_{t\geq 0}$, construct a stochastic process $\{\bar{X}_t\}_{t\geq 0}$ s.t.

$$\text{Law}(\bar{X}_t) = p_t, \quad \forall t \geq 0$$

- No unique solution: two-time marginals are not specified ($\text{Law}(\bar{X}_{t_1}, \bar{X}_{t_2}) = ?$)

Example: Fokker-Planck eq. $\dfrac{\partial p}{\partial t} = \nabla \cdot (p\nabla V) + \Delta p,$

- Stochastic:

$$\mathrm{d}\bar{X}_t = -\nabla V(\bar{X}_t)\mathrm{d}t + \sqrt{2}\mathrm{d}B_t, \quad \bar{X}_0 \sim p_0$$

- Deterministic:

$$\dot{\bar{X}}_t = -\nabla V(\bar{X}_t) + \nabla \log \bar{p}_t(\bar{X}_t), \quad \bar{X}_0 \sim p_0$$

  where $\bar{p}_t = \text{Law}(\bar{X}_t)$

- Both systems lead to the same one-time marginal densities
- difference arises with particle approximation

Uniqueness issue

Step 1: Given a gradient flow $\{p_t\}_{t\geq 0}$, construct a stochastic process $\{\bar{X}_t\}_{t\geq 0}$ s.t.

$$\text{Law}(\bar{X}_t) = p_t, \quad \forall t \geq 0$$

- No unique solution: two-time marginals are not specified ($\text{Law}(\bar{X}_{t_1}, \bar{X}_{t_2}) = ?$)

Example: Fokker-Planck eq. $\frac{\partial p}{\partial t} = \nabla \cdot (p\nabla V) + \Delta p,$

- Stochastic:

$$\mathrm{d}\bar{X}_t = -\nabla V(\bar{X}_t)\mathrm{d}t + \sqrt{2}\mathrm{d}B_t, \quad \bar{X}_0 \sim p_0$$

- Deterministic:

$$\dot{\bar{X}}_t = -\nabla V(\bar{X}_t) + \nabla \log \bar{p}_t(\bar{X}_t), \quad \bar{X}_0 \sim p_0$$

where $\bar{p}_t = \text{Law}(\bar{X}_t)$

- Both systems lead to the same one-time marginal densities
- difference arises with particle approximation

Uniqueness issue

Step 1: Given a gradient flow $\{p_t\}_{t\geq 0}$, construct a stochastic process $\{\bar{X}_t\}_{t\geq 0}$ s.t.

$$\text{Law}(\bar{X}_t) = p_t, \quad \forall t \geq 0$$

- No unique solution: two-time marginals are not specified ($\text{Law}(\bar{X}_{t_1}, \bar{X}_{t_2}) = ?$)

Example: Fokker-Planck eq. $\dfrac{\partial p}{\partial t} = \nabla \cdot (p \nabla V) + \Delta p,$

- Stochastic:

$$\mathrm{d}\bar{X}_t = -\nabla V(\bar{X}_t)\mathrm{d}t + \sqrt{2}\mathrm{d}B_t, \quad \bar{X}_0 \sim p_0$$

- Deterministic:

$$\dot{\bar{X}}_t = -\nabla V(\bar{X}_t) + \nabla \log \bar{p}_t(\bar{X}_t), \quad \bar{X}_0 \sim p_0$$

  where $\bar{p}_t = \text{Law}(\bar{X}_t)$

- Both systems lead to the same one-time marginal densities
- difference arises with particle approximation

Uniqueness issue

Step 1: Given a gradient flow $\{p_t\}_{t\geq 0}$, construct a stochastic process $\{\bar{X}_t\}_{t\geq 0}$ s.t.

$$\text{Law}(\bar{X}_t) = p_t, \quad \forall t \geq 0$$

- No unique solution: two-time marginals are not specified ($\text{Law}(\bar{X}_{t_1}, \bar{X}_{t_2}) = ?$)

Example: Fokker-Planck eq. $\dfrac{\partial p}{\partial t} = \nabla \cdot (p \nabla V) + \Delta p,$

- Stochastic:

$$\mathrm{d}\bar{X}_t = -\nabla V(\bar{X}_t)\mathrm{d}t + \sqrt{2}\mathrm{d}B_t, \quad \bar{X}_0 \sim p_0$$

- Deterministic:

$$\dot{\bar{X}}_t = -\nabla V(\bar{X}_t) + \nabla \log \bar{p}_t(\bar{X}_t), \quad \bar{X}_0 \sim p_0$$

  where $\bar{p}_t = \text{Law}(\bar{X}_t)$

- Both systems lead to the same one-time marginal densities
- difference arises with particle approximation

Uniqueness issue

Step 1: Given a gradient flow $\{p_t\}_{t\geq 0}$, construct a stochastic process $\{\bar{X}_t\}_{t\geq 0}$ s.t.

$$\text{Law}(\bar{X}_t) = p_t, \quad \forall t \geq 0$$

- No unique solution: two-time marginals are not specified ($\text{Law}(\bar{X}_{t_1}, \bar{X}_{t_2}) = ?$)

Example: Fokker-Planck eq. $\frac{\partial p}{\partial t} = \nabla \cdot (p \nabla V) + \Delta p,$

- Stochastic:

$$\mathrm{d}\bar{X}_t = -\nabla V(\bar{X}_t)\mathrm{d}t + \sqrt{2}\mathrm{d}B_t, \quad \bar{X}_0 \sim p_0$$

- Deterministic:

$$\dot{\bar{X}}_t = -\nabla V(\bar{X}_t) + \nabla \log \bar{p}_t(\bar{X}_t), \quad \bar{X}_0 \sim p_0$$

where $\bar{p}_t = \text{Law}(\bar{X}_t)$

- Both systems lead to the same one-time marginal densities
- difference arises with particle approximation

Particle approximation

Step 2: Realize $\bar{X}_t$ with system of (interacting) particles s.t. $\{X_t^1, \dots, X_t^N\}$

$$p_t^{(N)} := \frac{1}{N} \sum_{i=1}^{N} \delta_{X_t^i} \approx \mathsf{Law}(\bar{X}_t)$$

Example: Fokker-Planck eq.

- Stochastic:

$$\mathrm{d}\bar{X}_t = -\nabla V(\bar{X}_t)\mathrm{d}t + \sqrt{2}\mathrm{d}B_t \quad \to \quad \mathrm{d}X_t^i = -\nabla V(X_t^i) + \sqrt{2}\mathrm{d}B_t^i$$

- Deterministic:

$$\dot{\bar{X}}_t = -\nabla V(\bar{X}_t) - \nabla \log \bar{p}_t(\bar{X}_t) \quad \to \quad \dot{X}_t^i = -\nabla V(X_t^i) - I(X_t^i, p_t^{(N)})$$

  where $I(x, p_t^{(N)})$ is approximation of $\nabla \log \bar{p}_t(x)$

- results in interacting particle systems
- How to design the approximation?
- What is the difference between deterministic and stochastic method?

Particle approximation

Step 2: Realize $\bar{X}_t$ with system of (interacting) particles s.t. $\{X_t^1, \dots, X_t^N\}$

$$p_t^{(N)} := \frac{1}{N}\sum_{i=1}^{N} \delta_{X_t^i} \approx \text{Law}(\bar{X}_t)$$

Example: Fokker-Planck eq.

- Stochastic:

$$\mathrm{d}\bar{X}_t = -\nabla V(\bar{X}_t)\mathrm{d}t + \sqrt{2}\mathrm{d}B_t \quad \to \quad \mathrm{d}X_t^i = -\nabla V(X_t^i) + \sqrt{2}\mathrm{d}B_t^i$$

- Deterministic:

$$\dot{\bar{X}}_t = -\nabla V(\bar{X}_t) - \nabla \log \bar{p}_t(\bar{X}_t) \quad \to \quad \dot{X}_t^i = -\nabla V(X_t^i) - I(X_t^i, p_t^{(N)})$$

  where $I(x, p_t^{(N)})$ is approximation of $\nabla \log \bar{p}_t(x)$

- results in interacting particle systems
- How to design the approximation?
- What is the difference between deterministic and stochastic method?

Particle approximation

Step 2: Realize $\bar{X}_t$ with system of (interacting) particles s.t. $\{X_t^1, \ldots, X_t^N\}$

$$p_t^{(N)} := \frac{1}{N} \sum_{i=1}^{N} \delta_{X_t^i} \approx \mathsf{Law}(\bar{X}_t)$$

Example: Fokker-Planck eq.

- Stochastic:

$$\mathrm{d}\bar{X}_t = -\nabla V(\bar{X}_t)\mathrm{d}t + \sqrt{2}\mathrm{d}B_t \quad \rightarrow \quad \mathrm{d}X_t^i = -\nabla V(X_t^i) + \sqrt{2}\mathrm{d}B_t^i$$

- Deterministic:

$$\dot{\bar{X}}_t = -\nabla V(\bar{X}_t) - \nabla \log \bar{p}_t(\bar{X}_t) \quad \rightarrow \quad \dot{X}_t^i = -\nabla V(X_t^i) - I(X_t^i, p_t^{(N)})$$

  where $I(x, p_t^{(N)})$ is approximation of $\nabla \log \bar{p}_t(x)$
- results in interacting particle systems
- How to design the approximation?
- What is the difference between deterministic and stochastic method?

Particle approximation

Step 2: Realize $\bar{X}_t$ with system of (interacting) particles s.t. $\{X_t^1, \ldots, X_t^N\}$

$$p_t^{(N)} := \frac{1}{N}\sum_{i=1}^{N} \delta_{X_t^i} \approx \text{Law}(\bar{X}_t)$$

Example: Fokker-Planck eq.

- Stochastic:

$$\mathrm{d}\bar{X}_t = -\nabla V(\bar{X}_t)\mathrm{d}t + \sqrt{2}\mathrm{d}B_t \quad \to \quad \mathrm{d}X_t^i = -\nabla V(X_t^i) + \sqrt{2}\mathrm{d}B_t^i$$

- Deterministic:

$$\dot{\bar{X}}_t = -\nabla V(\bar{X}_t) - \nabla \log \bar{p}_t(\bar{X}_t) \quad \to \quad \dot{X}_t^i = -\nabla V(X_t^i) - I(X_t^i, p_t^{(N)})$$

where $I(x, p_t^{(N)})$ is approximation of $\nabla \log \bar{p}_t(x)$

- results in interacting particle systems
- How to design the approximation?
- What is the difference between deterministic and stochastic method?

Particle approximation

Step 2: Realize $\bar{X}_t$ with system of (interacting) particles s.t. $\{X_t^1, \ldots, X_t^N\}$

$$p_t^{(N)} := \frac{1}{N}\sum_{i=1}^N \delta_{X_t^i} \approx \mathsf{Law}(\bar{X}_t)$$

Example: Fokker-Planck eq.

- Stochastic:

$$\mathrm{d}\bar{X}_t = -\nabla V(\bar{X}_t)\mathrm{d}t + \sqrt{2}\mathrm{d}B_t \quad \rightarrow \quad \mathrm{d}X_t^i = -\nabla V(X_t^i) + \sqrt{2}\mathrm{d}B_t^i$$

- Deterministic:

$$\dot{\bar{X}}_t = -\nabla V(\bar{X}_t) - \nabla \log \bar{p}_t(\bar{X}_t) \quad \rightarrow \quad \dot{X}_t^i = -\nabla V(X_t^i) - I(X_t^i, p_t^{(N)})$$

where $I(x, p_t^{(N)})$ is approximation of $\nabla \log \bar{p}_t(x)$

- results in interacting particle systems
- How to design the approximation?
- What is the difference between deterministic and stochastic method?

Particle approximation

Step 2: Realize $\bar{X}_t$ with system of (interacting) particles s.t. $\{X_t^1, \dots, X_t^N\}$

$$p_t^{(N)} := \frac{1}{N}\sum_{i=1}^N \delta_{X_t^i} \approx \text{Law}(\bar{X}_t)$$

Example: Fokker-Planck eq.

- Stochastic:

$$\mathrm{d}\bar{X}_t = -\nabla V(\bar{X}_t)\mathrm{d}t + \sqrt{2}\mathrm{d}B_t \quad \rightarrow \quad \mathrm{d}X_t^i = -\nabla V(X_t^i) + \sqrt{2}\mathrm{d}B_t^i$$

- Deterministic:

$$\dot{\bar{X}}_t = -\nabla V(\bar{X}_t) - \nabla \log \bar{p}_t(\bar{X}_t) \quad \rightarrow \quad \dot{X}_t^i = -\nabla V(X_t^i) - I(X_t^i, p_t^{(N)})$$

where $I(x, p_t^{(N)})$ is approximation of $\nabla \log \bar{p}_t(x)$

- results in interacting particle systems
- How to design the approximation?
- What is the difference between deterministic and stochastic method?

Particle approximation

Step 2: Realize $\bar{X}_t$ with system of (interacting) particles s.t. $\{X_t^1, \dots, X_t^N\}$

$$p_t^{(N)} := \frac{1}{N} \sum_{i=1}^{N} \delta_{X_t^i} \approx \mathsf{Law}(\bar{X}_t)$$

Example: Fokker-Planck eq.

- Stochastic:

$$\mathrm{d}\bar{X}_t = -\nabla V(\bar{X}_t)\mathrm{d}t + \sqrt{2}\mathrm{d}B_t \quad \rightarrow \quad \mathrm{d}X_t^i = -\nabla V(X_t^i) + \sqrt{2}\mathrm{d}B_t^i$$

- Deterministic:

$$\dot{\bar{X}}_t = -\nabla V(\bar{X}_t) - \nabla \log \bar{p}_t(\bar{X}_t) \quad \rightarrow \quad \dot{X}_t^i = -\nabla V(X_t^i) - I(X_t^i, p_t^{(N)})$$

  where $I(x, p_t^{(N)})$ is approximation of $\nabla \log \bar{p}_t(x)$

- results in interacting particle systems
- How to design the approximation?
- What is the difference between deterministic and stochastic method?

Gaussian approximation

In order to approximate $\nabla \log(\bar{p}_t)$ in terms of particles $\{X_t^1, \ldots, X_t^N\}$:

- Fit a Gaussian distribution $N(m_t^{(N)}, \Sigma_t^{(N)})$ to the particles, where

$$m_t^{(N)} = \frac{1}{N}\sum_{i=1}^N X_t^i, \quad \Sigma_t^{(N)} = \frac{1}{N}\sum_{i=1}^N (X_t^i - m_t^{(N)})(X_t^i - m_t^{(N)})^T$$

- Use this to approximate the interaction term:

$$\nabla \log(\bar{p}_t(x)) \approx -(\Sigma_t^{(N)})^{-1}(x - m_t^{(N)})$$

- Resulting update law for particles

$$\dot{X}_t^i = -\nabla V(X_t^i) + (\Sigma_t^{(N)})^{-1}(X_t^i - m_t^{(N)})$$

- what are the benefits compared to stochastic case

$$\mathrm{d}X_t^i = -\nabla V(X_t^i) + \sqrt{2}\mathrm{d}B_t^i$$

Gaussian approximation

In order to approximate $\nabla \log(\bar{p}_t)$ in terms of particles $\{X_t^1, \ldots, X_t^N\}$:

- Fit a Gaussian distribution $N(m_t^{(N)}, \Sigma_t^{(N)})$ to the particles, where

$$m_t^{(N)} = \frac{1}{N}\sum_{i=1}^{N} X_t^i, \quad \Sigma_t^{(N)} = \frac{1}{N}\sum_{i=1}^{N}(X_t^i - m_t^{(N)})(X_t^i - m_t^{(N)})^T$$

- Use this to approximate the interaction term:

$$\nabla \log(\bar{p}_t(x)) \approx -(\Sigma_t^{(N)})^{-1}(x - m_t^{(N)})$$

- Resulting update law for particles

$$\dot{X}_t^i = -\nabla V(X_t^i) + (\Sigma_t^{(N)})^{-1}(X_t^i - m_t^{(N)})$$

- what are the benefits compared to stochastic case

$$\mathrm{d}X_t^i = -\nabla V(X_t^i) + \sqrt{2}\mathrm{d}B_t^i$$

Gaussian approximation

In order to approximate $\nabla \log(\bar{p}_t)$ in terms of particles $\{X_t^1, \dots, X_t^N\}$:

- Fit a Gaussian distribution $N(m_t^{(N)}, \Sigma_t^{(N)})$ to the particles, where

$$m_t^{(N)} = \frac{1}{N}\sum_{i=1}^N X_t^i, \quad \Sigma_t^{(N)} = \frac{1}{N}\sum_{i=1}^N (X_t^i - m_t^{(N)})(X_t^i - m_t^{(N)})^T$$

- Use this to approximate the interaction term:

$$\nabla \log(\bar{p}_t(x)) \approx -(\Sigma_t^{(N)})^{-1}(x - m_t^{(N)})$$

- Resulting update law for particles

$$\dot{X}_t^i = -\nabla V(X_t^i) + (\Sigma_t^{(N)})^{-1}(X_t^i - m_t^{(N)})$$

- what are the benefits compared to stochastic case

$$\mathrm{d}X_t^i = -\nabla V(X_t^i) + \sqrt{2}\mathrm{d}B_t^i$$

Gaussian approximation

In order to approximate $\nabla \log(\bar{p}_t)$ in terms of particles $\{X_t^1, \ldots, X_t^N\}$:

- Fit a Gaussian distribution $N(m_t^{(N)}, \Sigma_t^{(N)})$ to the particles, where

$$m_t^{(N)} = \frac{1}{N}\sum_{i=1}^N X_t^i, \quad \Sigma_t^{(N)} = \frac{1}{N}\sum_{i=1}^N (X_t^i - m_t^{(N)})(X_t^i - m_t^{(N)})^T$$

- Use this to approximate the interaction term:

$$\nabla \log(\bar{p}_t(x)) \approx -(\Sigma_t^{(N)})^{-1}(x - m_t^{(N)})$$

- Resulting update law for particles

$$\dot{X}_t^i = -\nabla V(X_t^i) + (\Sigma_t^{(N)})^{-1}(X_t^i - m_t^{(N)})$$

- what are the benefits compared to stochastic case

$$\mathrm{d}X_t^i = -\nabla V(X_t^i) + \sqrt{2}\mathrm{d}B_t^i$$

Gaussian setting
comparison between stochastic and deterministic method

- Assume the target distribution is $N(\bar{x}, Q)$, i.e. $V = (x - \bar{x})^T Q^{-1} (x - \bar{x})$
- Compare the error in estimating mean or variance:

$$\text{error} = \mathbb{E}[\|m_t^{(N)} - \bar{x}\|^2]$$

- deterministic:

$$\text{error} \le e^{-\lambda t} \mathbb{E}[\|m_0^{(N)} - \bar{x}\|^2]$$

- stochastic:

$$\text{error} \le e^{-\lambda t} \mathbb{E}[\|m_0^{(N)} - \bar{x}\|^2] + \frac{C}{N}$$

- same result for covariance, but not other moments

Observation:

Gaussian approx. $\Rightarrow$ more accurate estimation of mean and variance

Question: does the observation generalize?

Gaussian setting
comparison between stochastic and deterministic method

- Assume the target distribution is $N(\bar{x}, Q)$, i.e. $V = (x-\bar{x})^T Q^{-1}(x-\bar{x})$
- Compare the error in estimating mean or variance:

$$\text{error} = \mathbb{E}[\|m_t^{(N)} - \bar{x}\|^2]$$

- deterministic:

$$\text{error} \le e^{-\lambda t}\mathbb{E}[\|m_0^{(N)} - \bar{x}\|^2]$$

- stochastic:

$$\text{error} \le e^{-\lambda t}\mathbb{E}[\|m_0^{(N)} - \bar{x}\|^2] + \frac{C}{N}$$

- same result for covariance, but not other moments

Observation:

Gaussian approx. $\Rightarrow$ more accurate estimation of mean and variance

Question: does the observation generalize?

Gaussian setting
comparison between stochastic and deterministic method

- Assume the target distribution is $N(\bar{x}, Q)$, i.e. $V = (x-\bar{x})^T Q^{-1}(x-\bar{x})$
- Compare the error in estimating mean or variance:

$$\text{error} = \mathbb{E}[\|m_t^{(N)} - \bar{x}\|^2]$$

- deterministic:

$$\text{error} \leq e^{-\lambda t}\mathbb{E}[\|m_0^{(N)} - \bar{x}\|^2]$$

- stochastic:

$$\text{error} \leq e^{-\lambda t}\mathbb{E}[\|m_0^{(N)} - \bar{x}\|^2] + \frac{C}{N}$$

- same result for covariance, but not other moments

Observation:

Gaussian approx. $\Rightarrow$ more accurate estimation of mean and variance

Question: does the observation generalize?

Gaussian setting

comparison between stochastic and deterministic method

- Assume the target distribution is $N(\bar{x}, Q)$, i.e. $V = (x-\bar{x})^T Q^{-1}(x-\bar{x})$
- Compare the error in estimating mean or variance:

$$\text{error} = \mathbb{E}[\|m_t^{(N)} - \bar{x}\|^2]$$

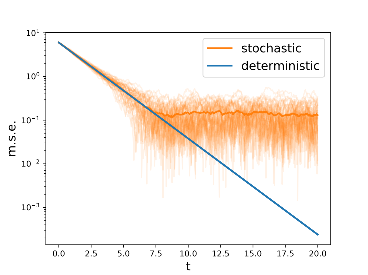
- deterministic:

$$\text{error} \leq e^{-\lambda t}\mathbb{E}[\|m_0^{(N)} - \bar{x}\|^2]$$

- stochastic:

$$\text{error} \leq e^{-\lambda t}\mathbb{E}[\|m_0^{(N)} - \bar{x}\|^2] + \frac{C}{N}$$

- same result for covariance, but not other moments



Observation:

Gaussian approx. $\Rightarrow$ more accurate estimation of mean and variance

Question: does the observation generalize?

Gaussian setting

comparison between stochastic and deterministic method

- Assume the target distribution is $N(\bar{x}, Q)$, i.e. $V = (x-\bar{x})^T Q^{-1}(x-\bar{x})$
- Compare the error in estimating mean or variance:

$$\text{error} = \mathbb{E}[\|m_t^{(N)} - \bar{x}\|^2]$$

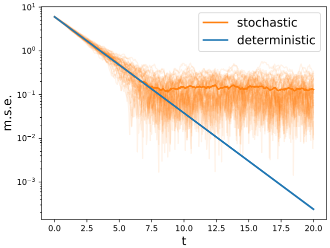
- deterministic:

$$\text{error} \le e^{-\lambda t}\mathbb{E}[\|m_0^{(N)} - \bar{x}\|^2]$$

- stochastic:

$$\text{error} \le e^{-\lambda t}\mathbb{E}[\|m_0^{(N)} - \bar{x}\|^2] + \frac{C}{N}$$

- same result for covariance, but not other moments



Observation:

Gaussian approx. $\Rightarrow$ more accurate estimation of mean and variance

Question: does the observation generalize?

Summary and proposed approach

Objective: numerically implement the gradient flow $\dfrac{\partial p}{\partial t} = \nabla \cdot (p \nabla \dfrac{\delta F}{\delta p})$

- Most existing works (including ours) focus on deterministic approach
- With the hope to trade-off computational effort with improvement in accuracy
- Challenge: approximating the mean-field terms (e.g. $\nabla \log(\bar{p}_t)$)
- SVGD (Liu & Wang, 2016): kernel approximation

$$\nabla \log(p(x)) \approx \int k(x,y) \nabla \log(p(y)) p(y) \mathrm{d}y = \int \nabla_y k(x,y) p(y) \mathrm{d}y$$

- score matching (Maoutsa et al., 2020)

$$\nabla \log(p) = \arg\min_{\phi} \left\{ \int \left(\frac{1}{2} \|\phi(x)\|^2 + \nabla \cdot \phi(x) \right) p(x) \mathrm{d}x \right\}$$

Proposed approach:

- Modify the objective function so that is well defined on empirical distributions
- Directly apply gradient flow on particles
- Achieved with variational characterization of the objective function

Summary and proposed approach

Objective: numerically implement the gradient flow $\frac{\partial p}{\partial t} = \nabla \cdot (p \nabla \frac{\delta F}{\delta p})$

- Most existing works (including ours) focus on deterministic approach
- With the hope to trade-off computational effort with improvement in accuracy
- Challenge: approximating the mean-field terms (e.g. $\nabla \log(\bar{p}_t)$)
- SVGD (Liu & Wang, 2016): kernel approximation

$$\nabla \log(p(x)) \approx \int k(x,y) \nabla \log(p(y)) p(y) \mathrm{d}y = \int \nabla_y k(x,y) p(y) \mathrm{d}y$$

- score matching (Maoutsa et al., 2020)

$$\nabla \log(p) = \arg\min_{\phi} \left\{ \int \left(\frac{1}{2} \|\phi(x)\|^2 + \nabla \cdot \phi(x) \right) p(x) \mathrm{d}x \right\}$$

Proposed approach:

- Modify the objective function so that is well defined on empirical distributions
- Directly apply gradient flow on particles
- Achieved with variational characterization of the objective function

Summary and proposed approach

Objective: numerically implement the gradient flow $\frac{\partial p}{\partial t} = \nabla \cdot (p \nabla \frac{\delta F}{\delta p})$

- Most existing works (including ours) focus on deterministic approach
- With the hope to trade-off computational effort with improvement in accuracy
- Challenge: approximating the mean-field terms (e.g. $\nabla \log(\bar{p}_t)$)
- SVGD (Liu & Wang, 2016): kernel approximation

$$\nabla \log(p(x)) \approx \int k(x,y) \nabla \log(p(y)) p(y) \mathrm{d}y = \int \nabla_y k(x,y) p(y) \mathrm{d}y$$

- score matching (Maoutsa et al., 2020)

$$\nabla \log(p) = \arg\min_{\phi} \left\{ \int \left(\frac{1}{2} \|\phi(x)\|^2 + \nabla \cdot \phi(x) \right) p(x) \mathrm{d}x \right\}$$

Proposed approach:

- Modify the objective function so that is well defined on empirical distributions
- Directly apply gradient flow on particles
- Achieved with variational characterization of the objective function

Summary and proposed approach

Objective: numerically implement the gradient flow $\frac{\partial p}{\partial t} = \nabla \cdot (p \nabla \frac{\delta F}{\delta p})$

- Most existing works (including ours) focus on deterministic approach
- With the hope to trade-off computational effort with improvement in accuracy
- Challenge: approximating the mean-field terms (e.g. $\nabla \log(\bar{p}_t)$)
- SVGD (Liu & Wang, 2016): kernel approximation

$$\nabla \log(p(x)) \approx \int k(x,y) \nabla \log(p(y)) p(y) \mathrm{d}y = \int \nabla_y k(x,y) p(y) \mathrm{d}y$$

- score matching (Maoutsa et al., 2020)

$$\nabla \log(p) = \arg\min_{\phi} \left\{ \int \left(\frac{1}{2} \|\phi(x)\|^2 + \nabla \cdot \phi(x) \right) p(x) \mathrm{d}x \right\}$$

Proposed approach:

- Modify the objective function so that is well defined on empirical distributions
- Directly apply gradient flow on particles
- Achieved with variational characterization of the objective function

Summary and proposed approach

Objective: numerically implement the gradient flow $\frac{\partial p}{\partial t} = \nabla \cdot (p \nabla \frac{\delta F}{\delta p})$

- Most existing works (including ours) focus on deterministic approach
- With the hope to trade-off computational effort with improvement in accuracy
- Challenge: approximating the mean-field terms (e.g. $\nabla \log(\bar{p}_t)$)
- SVGD (Liu & Wang, 2016): kernel approximation

$$\nabla \log(p(x)) \approx \int k(x,y) \nabla \log(p(y)) p(y) \mathrm{d}y = \int \nabla_y k(x,y) p(y) \mathrm{d}y$$

- score matching (Maoutsa et al., 2020)

$$\nabla \log(p) = \arg\min_{\phi} \left\{ \int \left(\frac{1}{2} \|\phi(x)\|^2 + \nabla \cdot \phi(x) \right) p(x) \mathrm{d}x \right\}$$

Proposed approach:

- Modify the objective function so that is well defined on empirical distributions
- Directly apply gradient flow on particles
- Achieved with variational characterization of the objective function

Outline

- Overview of numerical methods to implement Wasserstein gradient flows
- Variational approach

Outline

- Overview of numerical methods to implement Wasserstein gradient flows
- Variational approach

Variational f-divergences

- Consider f-divergence objective functionals

$$F(p) = D_f(p\|q) := \int f(\frac{p(x)}{q(x)})q(x)\mathrm{d}x$$

where $f:[0,\infty]\to\mathbb{R}$ is convex and $f(1)=0$ (e.g. $f(x)=x\log(x)\to$ KL)

- It admits variational representation

$$D_f(p\|q) = \sup_{h\in\mathcal{C}}\left\{\int h(x)p(x)\mathrm{d}x - \int f^*(h(x))q(x)\mathrm{d}x\right\}$$

- Approximate f-divergence

$$D_f^{\mathcal{H}}(p\|q) = \sup_{h\in\mathcal{H}}\left\{\int h(x)p(x)\mathrm{d}x - \int f^*(h(x))q(x)\mathrm{d}x\right\}$$

- It is well-defined for empirical distributions

$$D_f^{\mathcal{H}}(p^{(N)}\|q) = \sup_{h\in\mathcal{H}}\left\{\frac{1}{N}\sum_{i=1}^{N}h(X^i) - \int f^*(h(x))q(x)\mathrm{d}x\right\}$$

- Is the new objective function meaningful?

Variational f-divergences

- Consider f-divergence objective functionals

$$F(p) = D_f(p\|q) := \int f(\frac{p(x)}{q(x)})q(x)\mathrm{d}x$$

where $f : [0,\infty] \to \mathbb{R}$ is convex and $f(1) = 0$ (e.g. $f(x) = x\log(x) \to$ KL)

- It admits variational representation

$$D_f(p\|q) = \sup_{h\in\mathcal{C}} \left\{ \int h(x)p(x)\mathrm{d}x - \int f^*(h(x))q(x)\mathrm{d}x \right\}$$

- Approximate f-divergence

$$D_f^{\mathcal{H}}(p\|q) = \sup_{h\in\mathcal{H}} \left\{ \int h(x)p(x)\mathrm{d}x - \int f^*(h(x))q(x)\mathrm{d}x \right\}$$

- It is well-defined for empirical distributions

$$D_f^{\mathcal{H}}(p^{(N)}\|q) = \sup_{h\in\mathcal{H}} \left\{ \frac{1}{N}\sum_{i=1}^{N} h(X^i) - \int f^*(h(x))q(x)\mathrm{d}x \right\}$$

- Is the new objective function meaningful?

Variational f-divergences

- Consider f-divergence objective functionals

$$F(p) = D_f(p\|q) := \int f(\frac{p(x)}{q(x)})q(x)\mathrm{d}x$$

where $f : [0,\infty] \to \mathbb{R}$ is convex and $f(1) = 0$ (e.g. $f(x) = x\log(x) \to$ KL)

- It admits variational representation

$$D_f(p\|q) = \sup_{h\in\mathcal{C}} \left\{ \int h(x)p(x)\mathrm{d}x - \int f^*(h(x))q(x)\mathrm{d}x \right\}$$

- Approximate f-divergence

$$D_f^{\mathcal{H}}(p\|q) = \sup_{h\in\mathcal{H}} \left\{ \int h(x)p(x)\mathrm{d}x - \int f^*(h(x))q(x)\mathrm{d}x \right\}$$

- It is well-defined for empirical distributions

$$D_f^{\mathcal{H}}(p^{(N)}\|q) = \sup_{h\in\mathcal{H}} \left\{ \frac{1}{N}\sum_{i=1}^{N} h(X^i) - \int f^*(h(x))q(x)\mathrm{d}x \right\}$$

- Is the new objective function meaningful?

Variational f-divergences

- Consider f-divergence objective functionals

$$F(p) = D_f(p\|q) := \int f(\frac{p(x)}{q(x)})q(x)\mathrm{d}x$$

where $f : [0,\infty] \to \mathbb{R}$ is convex and $f(1) = 0$ (e.g. $f(x) = x\log(x) \to$ KL)

- It admits variational representation

$$D_f(p\|q) = \sup_{h\in\mathcal{C}} \left\{\int h(x)p(x)\mathrm{d}x - \int f^*(h(x))q(x)\mathrm{d}x\right\}$$

- Approximate f-divergence

$$D_f^{\mathcal{H}}(p\|q) = \sup_{h\in\mathcal{H}} \left\{\int h(x)p(x)\mathrm{d}x - \int f^*(h(x))q(x)\mathrm{d}x\right\}$$

- It is well-defined for empirical distributions

$$D_f^{\mathcal{H}}(p^{(N)}\|q) = \sup_{h\in\mathcal{H}} \left\{\frac{1}{N}\sum_{i=1}^N h(X^i) - \int f^*(h(x))q(x)\mathrm{d}x\right\}$$

- Is the new objective function meaningful?

Variational f-divergences

- Consider f-divergence objective functionals

$$F(p) = D_f(p\|q) := \int f(\frac{p(x)}{q(x)})q(x)\mathrm{d}x$$

where $f : [0,\infty] \to \mathbb{R}$ is convex and $f(1) = 0$ (e.g. $f(x) = x\log(x) \to$ KL)

- It admits variational representation

$$D_f(p\|q) = \sup_{h\in\mathcal{C}} \left\{ \int h(x)p(x)\mathrm{d}x - \int f^*(h(x))q(x)\mathrm{d}x \right\}$$

- Approximate f-divergence

$$D_f^{\mathcal{H}}(p\|q) = \sup_{h\in\mathcal{H}} \left\{ \int h(x)p(x)\mathrm{d}x - \int f^*(h(x))q(x)\mathrm{d}x \right\}$$

- It is well-defined for empirical distributions

$$D_f^{\mathcal{H}}(p^{(N)}\|q) = \sup_{h\in\mathcal{H}} \left\{ \frac{1}{N}\sum_{i=1}^{N} h(X^i) - \int f^*(h(x))q(x)\mathrm{d}x \right\}$$

- Is the new objective function meaningful?

Properties of the approximate f divergence

- upper-bound:

$$D_f^{\mathcal{H}}(p\|q) \leq D_f(p\|q) \quad \text{with equality if} \quad f'(\frac{p}{q}) \in \mathcal{H}$$

- positivity: If $\mathcal{H}$ contains all constant functions, then

$$D_f^{\mathcal{H}}(p\|q) \geq 0, \quad \forall p, q$$

- moment-matching: If for all $h \in \mathcal{H}$, $a + bh \in \mathcal{H}$ for $a, b \in \mathbb{R}$

$$D_f^{\mathcal{H}}(p\|q) = 0 \quad \Longleftrightarrow \quad \int hp\mathrm{d}x = \int hq\mathrm{d}x, \quad \forall h \in \mathcal{H}$$

- embedding: Additionally, if f is α-strongly convex and L-smooth, then

$$\frac{\alpha}{2} d_{\mathcal{H}}(p,q)^2 \leq D_f^{\mathcal{H}}(p\|q) \leq \frac{L}{2} d_{\mathcal{H}}(p,q)^2$$

where $d_{\mathcal{H}}(p,q)$ is a type of integral probability metric

$$d_{\mathcal{H}}(p,q) = \sup_{h \in \mathcal{H}} \frac{1}{\|h\|_{2,q}} \left\{ \int hp\mathrm{d}x - \int hq\mathrm{d}x \right\}$$

Properties of the approximate f divergence

- upper-bound:

$$D_f^{\mathcal{H}}(p\|q) \leq D_f(p\|q) \quad \text{with equality if} \quad f'(\frac{p}{q}) \in \mathcal{H}$$

- positivity: If $\mathcal{H}$ contains all constant functions, then

$$D_f^{\mathcal{H}}(p\|q) \geq 0, \quad \forall p, q$$

- moment-matching: If for all $h \in \mathcal{H}$, $a + bh \in \mathcal{H}$ for $a, b \in \mathbb{R}$

$$D_f^{\mathcal{H}}(p\|q) = 0 \quad \Longleftrightarrow \quad \int hp\mathrm{d}x = \int hq\mathrm{d}x, \quad \forall h \in \mathcal{H}$$

- embedding: Additionally, if f is α-strongly convex and L-smooth, then

$$\frac{\alpha}{2} d_{\mathcal{H}}(p,q)^2 \leq D_f^{\mathcal{H}}(p\|q) \leq \frac{L}{2} d_{\mathcal{H}}(p,q)^2$$

where $d_{\mathcal{H}}(p,q)$ is a type of integral probability metric

$$d_{\mathcal{H}}(p,q) = \sup_{h \in \mathcal{H}} \frac{1}{\|h\|_{2,q}} \left\{ \int hp\mathrm{d}x - \int hq\mathrm{d}x \right\}$$

Properties of the approximate f divergence

- upper-bound:

$$D_f^{\mathcal{H}}(p\|q) \leq D_f(p\|q) \quad \text{with equality if} \quad f'(\frac{p}{q}) \in \mathcal{H}$$

- positivity: If $\mathcal{H}$ contains all constant functions, then

$$D_f^{\mathcal{H}}(p\|q) \geq 0, \quad \forall p, q$$

- moment-matching: If for all $h \in \mathcal{H}$, $a + bh \in \mathcal{H}$ for $a, b \in \mathbb{R}$

$$D_f^{\mathcal{H}}(p\|q) = 0 \quad \Longleftrightarrow \quad \int hp\mathrm{d}x = \int hq\mathrm{d}x, \quad \forall h \in \mathcal{H}$$

- embedding: Additionally, if f is α-strongly convex and L-smooth, then

$$\frac{\alpha}{2} d_{\mathcal{H}}(p,q)^2 \leq D_f^{\mathcal{H}}(p\|q) \leq \frac{L}{2} d_{\mathcal{H}}(p,q)^2$$

where $d_{\mathcal{H}}(p,q)$ is a type of integral probability metric

$$d_{\mathcal{H}}(p,q) = \sup_{h \in \mathcal{H}} \frac{1}{\|h\|_{2,q}} \left\{ \int hp\mathrm{d}x - \int hq\mathrm{d}x \right\}$$

Properties of the approximate f divergence

- upper-bound:

$$D_f^{\mathcal{H}}(p\|q) \leq D_f(p\|q) \quad \text{with equality if} \quad f'(\frac{p}{q}) \in \mathcal{H}$$

- positivity: If $\mathcal{H}$ contains all constant functions, then

$$D_f^{\mathcal{H}}(p\|q) \geq 0, \quad \forall p, q$$

- moment-matching: If for all $h \in \mathcal{H}$, $a + bh \in \mathcal{H}$ for $a, b \in \mathbb{R}$

$$D_f^{\mathcal{H}}(p\|q) = 0 \quad \Longleftrightarrow \quad \int hp\mathrm{d}x = \int hq\mathrm{d}x, \quad \forall h \in \mathcal{H}$$

- embedding: Additionally, if f is α-strongly convex and L-smooth, then

$$\frac{\alpha}{2} d_{\mathcal{H}}(p,q)^2 \leq D_f^{\mathcal{H}}(p\|q) \leq \frac{L}{2} d_{\mathcal{H}}(p,q)^2$$

where $d_{\mathcal{H}}(p,q)$ is a type of integral probability metric

$$d_{\mathcal{H}}(p,q) = \sup_{h \in \mathcal{H}} \frac{1}{\|h\|_{2,q}} \left\{ \int hp\mathrm{d}x - \int hq\mathrm{d}x \right\}$$

Variational Wasserstein gradient flow

- New optimization problem:

$$\min_p D_f^{\mathcal{H}}(p\|q) = \min_p \max_{h\in\mathcal{H}} \underbrace{\left\{\int hp\mathrm{d}x - \int f^*(h)q\mathrm{d}x\right\}}_{\mathcal{V}(p,h)}$$

- Gradient flow:

$$\frac{\partial p_t}{\partial t} = \nabla \cdot (p_t \nabla h_t)$$

where h_t is the maximizer for $p = p_t$

- Representation in terms of $\bar{X}_t$:

$$\dot{\bar{X}}_t = -\nabla h_t(\bar{X}_t)$$

- Particle approximation

$$\dot{X}_t^i = -\nabla h_t^{(N)}(X_t^i)$$

where $h_t^{(N)}$ is the maximizer for $p = p_t^{(N)} = \frac{1}{N}\sum_{i=1}^{N} X_t^i$

- How about the sampling problem where we do not have access to q?

Variational Wasserstein gradient flow

- New optimization problem:

$$\min_p D_f^{\mathcal{H}}(p\|q) = \min_p \max_{h\in\mathcal{H}} \underbrace{\left\{\int hp\mathrm{d}x - \int f^*(h)q\mathrm{d}x\right\}}_{\mathcal{V}(p,h)}$$

- Gradient flow:

$$\frac{\partial p_t}{\partial t} = \nabla\cdot(p_t\nabla h_t)$$

where h_t is the maximizer for $p = p_t$

- Representation in terms of $\bar{X}_t$:

$$\dot{\bar{X}}_t = -\nabla h_t(\bar{X}_t)$$

- Particle approximation

$$\dot{X}_t^i = -\nabla h_t^{(N)}(X_t^i)$$

where $h_t^{(N)}$ is the maximizer for $p = p_t^{(N)} = \frac{1}{N}\sum_{i=1}^N X_t^i$

- How about the sampling problem where we do not have access to q?

Variational Wasserstein gradient flow

- New optimization problem:

$$\min_p D_f^{\mathcal{H}}(p\|q) = \min_p \max_{h\in\mathcal{H}} \underbrace{\left\{\int hp\mathrm{d}x - \int f^*(h)q\mathrm{d}x\right\}}_{\mathcal{V}(p,h)}$$

- Gradient flow:

$$\frac{\partial p_t}{\partial t} = \nabla\cdot(p_t \nabla h_t)$$

where h_t is the maximizer for $p = p_t$

- Representation in terms of $\bar{X}_t$:

$$\dot{\bar{X}}_t = -\nabla h_t(\bar{X}_t)$$

- Particle approximation

$$\dot{X}_t^i = -\nabla h_t^{(N)}(X_t^i)$$

where $h_t^{(N)}$ is the maximizer for $p = p_t^{(N)} = \frac{1}{N}\sum_{i=1}^N X_t^i$

- How about the sampling problem where we do not have access to q?

Variational Wasserstein gradient flow

- New optimization problem:

$$\min_p \; D_f^{\mathcal{H}}(p\|q) = \min_p \max_{h\in\mathcal{H}} \underbrace{\left\{ \int hp\mathrm{d}x - \int f^*(h)q\mathrm{d}x \right\}}_{\mathcal{V}(p,h)}$$

- Gradient flow:

$$\frac{\partial p_t}{\partial t} = \nabla \cdot (p_t \nabla h_t)$$

where h_t is the maximizer for $p = p_t$

- Representation in terms of $\bar{X}_t$:

$$\dot{\bar{X}}_t = -\nabla h_t(\bar{X}_t)$$

- Particle approximation

$$\dot{X}_t^i = -\nabla h_t^{(N)}(X_t^i)$$

where $h_t^{(N)}$ is the maximizer for $p = p_t^{(N)} = \dfrac{1}{N}\sum_{i=1}^N X_t^i$

- How about the sampling problem where we do not have access to q?

Variational Wasserstein gradient flow

- New optimization problem:

$$\min_p D_f^{\mathcal{H}}(p\|q) = \min_p \max_{h\in\mathcal{H}} \underbrace{\left\{\int hp\mathrm{d}x - \int f^*(h)q\mathrm{d}x\right\}}_{\mathcal{V}(p,h)}$$

- Gradient flow:

$$\frac{\partial p_t}{\partial t} = \nabla\cdot(p_t\nabla h_t)$$

where h_t is the maximizer for $p = p_t$

- Representation in terms of $\bar{X}_t$:

$$\dot{\bar{X}}_t = -\nabla h_t(\bar{X}_t)$$

- Particle approximation

$$\dot{X}_t^i = -\nabla h_t^{(N)}(X_t^i)$$

where $h_t^{(N)}$ is the maximizer for $p = p_t^{(N)} = \dfrac{1}{N}\sum_{i=1}^N X_t^i$

- How about the sampling problem where we do not have access to q?

- Objective function for sampling: $(f_s(x) = x\log(x))$

$$D_{f_s}^{\mathcal{H}}(p\|q) = \max_{h\in\mathcal{H}}\left\{\int hp\mathrm{d}x - \int e^{h-1}q\mathrm{d}x\right\}$$

- With change of variable $h \to h + 1 + \log(\frac{\eta}{q})$

$$D_{f_s}^{\mathcal{H}}(p\|q) = 1 + \int \log(\frac{\eta}{q})p\mathrm{d}x + \max_{h\in\mathcal{H}}\left\{\int hp\mathrm{d}x - \int e^{h}\eta\mathrm{d}x\right\}$$

  where η is a distribution easy to sample (e.g. $N(m_t, \Sigma_t)$)

- Resulting gradient flow $(q = e^{-V})$

$$\dot{\bar{X}}_t = -\nabla V(\bar{X}_t) + \Sigma_t^{-1}(\bar{X}_t - m_t) - \nabla h_t(\bar{X}_t)$$

- It simplifies to the algorithm with Gaussian approx. when $\mathcal{H} = \{0\}$

- Objective function for sampling: $(f_s(x) = x\log(x))$

$$D_{f_s}^{\mathcal{H}}(p\|q) = \max_{h\in\mathcal{H}} \left\{ \int hp\mathrm{d}x - \int e^{h-1} q\mathrm{d}x \right\}$$

- With change of variable $h \to h + 1 + \log(\frac{\eta}{q})$

$$D_{f_s}^{\mathcal{H}}(p\|q) = 1 + \int \log(\frac{\eta}{q}) p\mathrm{d}x + \max_{h\in\mathcal{H}} \left\{ \int hp\mathrm{d}x - \int e^{h} \eta\mathrm{d}x \right\}$$

where η is a distribution easy to sample (e.g. $N(m_t, \Sigma_t)$)

- Resulting gradient flow $(q = e^{-V})$

$$\dot{\bar{X}}_t = -\nabla V(\bar{X}_t) + \Sigma_t^{-1}(\bar{X}_t - m_t) - \nabla h_t(\bar{X}_t)$$

- It simplifies to the algorithm with Gaussian approx. when $\mathcal{H} = \{0\}$

- Objective function for sampling: $(f_s(x) = x\log(x))$

$$D_{f_s}^{\mathcal{H}}(p\|q) = \max_{h\in\mathcal{H}}\left\{\int hp\mathrm{d}x - \int e^{h-1}q\mathrm{d}x\right\}$$

- With change of variable $h \to h + 1 + \log(\frac{\eta}{q})$

$$D_{f_s}^{\mathcal{H}}(p\|q) = 1 + \int \log(\frac{\eta}{q})p\mathrm{d}x + \max_{h\in\mathcal{H}}\left\{\int hp\mathrm{d}x - \int e^{h}\eta\mathrm{d}x\right\}$$

where η is a distribution easy to sample (e.g. $N(m_t, \Sigma_t)$)

- Resulting gradient flow $(q = e^{-V})$

$$\dot{\bar{X}}_t = -\nabla V(\bar{X}_t) + \Sigma_t^{-1}(\bar{X}_t - m_t) - \nabla h_t(\bar{X}_t)$$

- It simplifies to the algorithm with Gaussian approx. when $\mathcal{H} = \{0\}$

- Objective function for sampling: $(f_s(x) = x\log(x))$

$$D_{f_s}^{\mathcal{H}}(p\|q) = \max_{h\in\mathcal{H}} \left\{ \int hp\mathrm{d}x - \int e^{h-1} q\mathrm{d}x \right\}$$

- With change of variable $h \to h + 1 + \log(\frac{\eta}{q})$

$$D_{f_s}^{\mathcal{H}}(p\|q) = 1 + \int \log(\frac{\eta}{q})p\mathrm{d}x + \max_{h\in\mathcal{H}} \left\{ \int hp\mathrm{d}x - \int e^{h} \eta\mathrm{d}x \right\}$$

where η is a distribution easy to sample (e.g. $N(m_t, \Sigma_t)$)

- Resulting gradient flow $(q = e^{-V})$

$$\dot{\bar{X}}_t = -\nabla V(\bar{X}_t) + \Sigma_t^{-1}(\bar{X}_t - m_t) - \nabla h_t(\bar{X}_t)$$

- It simplifies to the algorithm with Gaussian approx. when $\mathcal{H} = \{0\}$

Computational algorithms

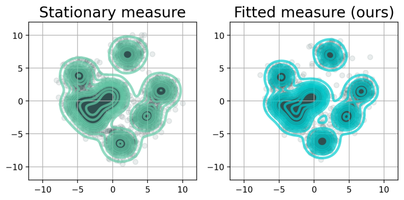
- time discretization with JKO scheme

$$\bar{X}_{k+1} = \nabla\phi_k(\bar{X}_k),$$
$$\phi_k = \underset{\phi\in\mathsf{ICNN}}{\arg\min}\,\max_{h\in\mathcal{H}}\{\frac{1}{2\Delta t}W_2^2(\bar{p}_k,\nabla\phi\#\bar{p}_k) + \mathcal{V}(h,\nabla\phi\#\bar{p}_k)\}$$

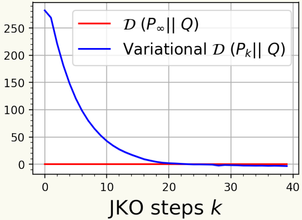
- results in min-max optimization at each time-step
- solve using stochastic optimization algorithms
- represent ϕ with input convex neural networks (ICNN) (Amos et al., 2017)
- represent h with feed-forward neural networks

Setup:

- objective function is $D(p\|q)$
- target is Gaussian mixture with 10 components

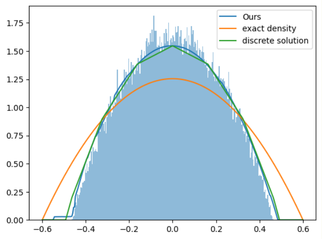


dimension $= 128$

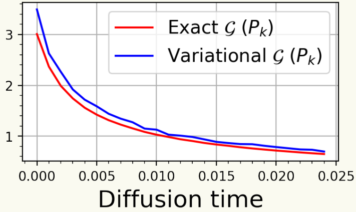


Setup:

- objective function is generalized entropy $\mathcal{G}(p) = \frac{1}{m-1}\int p^m(x)\mathrm{d}x$
- gradient flow is $\frac{\partial p}{\partial t} = \Delta p^m$



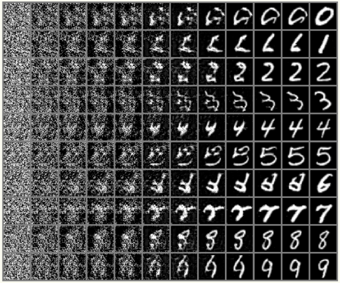
comparison with exact solution



convergence of the objective function

Setup:

- objective function is JS distance $\text{JSD}(p\|q) = D(p\|\frac{p+q}{2}) + D(q\|\frac{p+q}{2})$
- assuming access to samples from q (GAN setup)



MNIST dataset

CIFAR dataset

Concluding remarks

Summary:

- Variational approach to construct gradient flows

$$\min_p F(p) \quad \to \quad \min_p \max_{h\in\mathcal{H}} \mathcal{V}(p,h)$$

- established elementary results about the variational divergence
- numerical results illustrating scalability with dimension

Open questions:

- Does the gradient flow converge

$$D_f^{\mathcal{H}}(p_t\|q) \to 0, \quad \text{as} \quad t\to\infty$$

- Under what conditions we have log-Sobolev type inequality

$$\frac{\mathrm{d}}{\mathrm{d}t} D_f^{\mathcal{H}}(p_t\|q) \le -\lambda D_f^{\mathcal{H}}(p_t\|q)$$

- For sampling, what is the benefit compared to simulating Langevin eq.?

Concluding remarks

Summary:

- Variational approach to construct gradient flows

$$\min_p F(p) \quad \rightarrow \quad \min_p \max_{h\in\mathcal{H}} \mathcal{V}(p,h)$$

- established elementary results about the variational divergence
- numerical results illustrating scalability with dimension

Open questions:

- Does the gradient flow converge

$$D_f^{\mathcal{H}}(p_t\|q) \rightarrow 0, \quad \text{as} \quad t\rightarrow\infty$$

- Under what conditions we have log-Sobolev type inequality

$$\frac{\mathrm{d}}{\mathrm{d}t} D_f^{\mathcal{H}}(p_t\|q) \leq -\lambda D_f^{\mathcal{H}}(p_t\|q)$$

- For sampling, what is the benefit compared to simulating Langevin eq.?