NEURAL INFORMATION PROCESSING SYSTEMS

Optimizing Functionals in Probability Space with Input Convex Neural Networks

David Alvarez-Melis
Yair Schiff
Youssef Mroueh

Microsoft
IBM Research

Outline

- **Goal**: solve optimization problems over distributions
- **Approach**: follow gradient flow (GF) of optim. objective
- JKO: minimizing movement scheme, discretizes flow
- Brenier Theorem allows to write JKO as an optimization on the space of convex functions
- Our method (JKO-ICNN) implements JKO using Input Convex Neural Networks (ICNN)

Background: Gradient Flows

- GF: steepest descent curve $x(t)$ of functional $F : \mathscr{X} \to \mathbb{R}$
- In Euclidean space $\mathscr{X}$: $\quad x'(t) = -\nabla F(x(t))$
- In Probability space $\mathscr{P}(\mathscr{X})$: $\quad \partial_t \rho_t = \nabla \cdot \left(\rho_t \nabla \frac{\delta F(\rho)}{\delta \rho}\right)$

Class	PDE $\partial_t \rho =$	Flow Functional $F(\rho) =$
Heat Equation	$\Delta\rho$	$\int \rho(x) \log \rho(x)\,\mathrm{d}x$
Advection	$\nabla \cdot (\rho \nabla V)$	$\int V(x)\,\mathrm{d}\rho(x)$
Fokker-Planck	$\Delta\rho + \nabla \cdot (\rho \nabla V)$	$\int \rho(x) \log \rho(x)\,\mathrm{d}x + \int V(x)\,\mathrm{d}\rho(x)$
Porous Media	$\Delta(\rho^m) + \nabla \cdot (\rho \nabla V)$	$\frac{1}{m-1} \int \rho(x)^m\,\mathrm{d}x + \int V(x)\,\mathrm{d}\rho(x)$
Adv.+Diff.+Inter.	$\nabla \cdot \left[\rho(\nabla f'(\rho) + \nabla V + (\nabla W) * \rho)\right]$	$\int V(x)\,\mathrm{d}\rho(x) + \int f(\rho(x))\,\mathrm{d}x + \frac{1}{2} \iint W(x-x')\,\mathrm{d}\rho(x)\,\mathrm{d}\rho(x')$

Equivalences between PDEs and Functional Gradient Flows

Background: JKO Scheme

- A time discretization of gradient flows in prob. space:

$$\rho_{t+1}^{\tau} \in \arg\min_{\rho \in \mathbb{W}_2(\mathscr{X})} F(\rho) + \frac{1}{2\tau} \mathrm{W}_2^2(\rho, \rho_t^{\tau}), \quad \tau > 0$$

step size

Wasserstein distance

- Jordan-Kinderlehrer-Otto (1998)
- Implementation requires tractable modeling of measures

Setting

- We consider problems of the form $\min_{\rho \in \mathscr{P}(\mathscr{X})} F(\rho)$
- We focus on 3 functional families: $\mathscr{V}(\rho) = \int V(x) d\rho$ (potential), $\mathscr{W}(\rho) = \frac{1}{2} \iint W(x - x') d\rho \otimes \rho$ (interaction), $\mathscr{F}(\rho) = \int f(\rho(x)) dx$ (internal energy)
- For these functionals, GF converges to unique minimizer
- Algorithm gist: use JKO iteratively to minimize F

From Measures to Convex Functions

- Under some assumptions, Brenier theorem yields:

$$\mathrm{W}_2^2(\alpha, (\nabla u)_{\sharp}\alpha) = \int_{\mathscr{X}} \|\nabla u(x) - x\|_2^2 d\alpha, \quad u \in \mathrm{CVX}(\mathscr{X})$$

- Therefore, JKO scheme can be written as:

$$\min_{u \in \mathrm{CVX}(\mathscr{X})} F((\nabla u)_{\sharp} \rho_t^{\tau}) + \frac{1}{2\tau} \int_{\mathscr{X}} \|\nabla u(x) - x\|_2^2 d\rho_t^{\tau}$$

- Measures implicitly defined via: $\rho_{t+1}^{\tau} = (\nabla u_{t+1}^{\tau})_{\#}(\rho_t^{\tau})$

From Convex Functions to ICNN

- Parametrize CVX via input-convex neural nets (Amos et al. '17)
- Problem becomes:

$$\min_{u_\theta \in \mathrm{ICNN}(\mathscr{X})} F((\nabla_x u_\theta(x))_{\sharp} \rho_t^{\tau}) + \frac{1}{2\tau} \int_{\mathscr{X}} \|\nabla_x u_\theta(x) - x\|_2^2 d\rho_t^{\tau}$$

- Simple form for potential/interaction functionals:

$$\mathscr{V}\left((\nabla_x u_\theta)_{\sharp} \rho_t^{\tau}\right) = \mathbb{E}_{x \sim \rho_t^{\tau}} V(\nabla_x u_\theta(x))$$

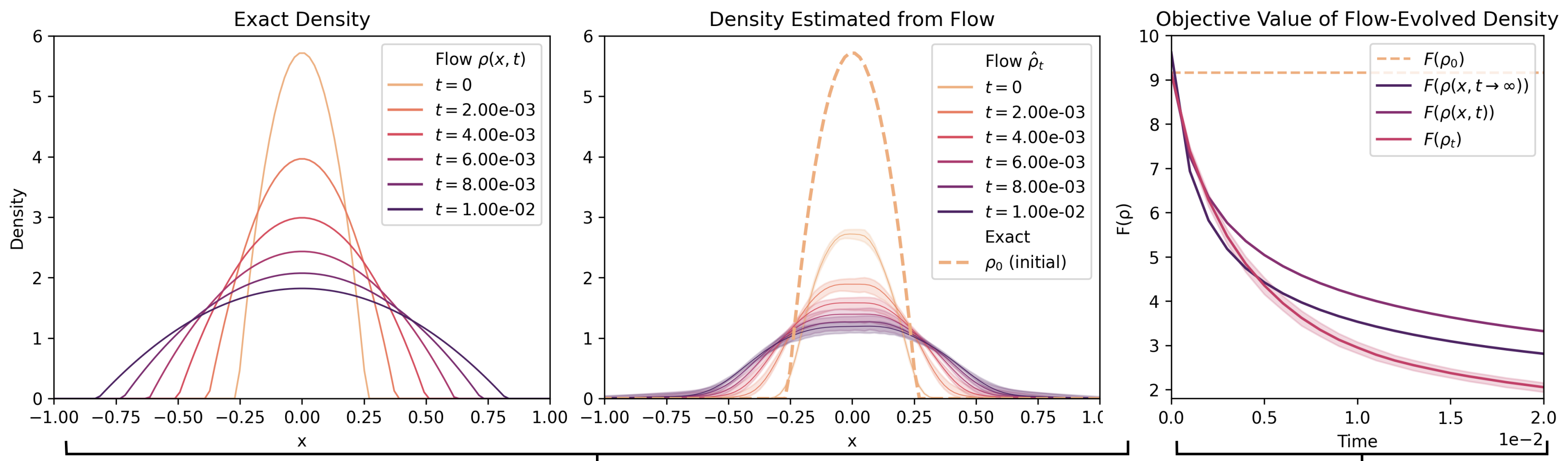
$$\mathscr{W}\left((\nabla_x u_\theta)_{\sharp} \rho_t^{\tau}\right) = \frac{1}{2} \mathbb{E}_{x,y \sim \rho_t^{\tau}} W(\nabla_x u_\theta(x) - \nabla_x u_\theta(y))$$

approximate with finite samples

- Surrogate objectives for certain internal energies

Evolving PDEs with Known Solutions

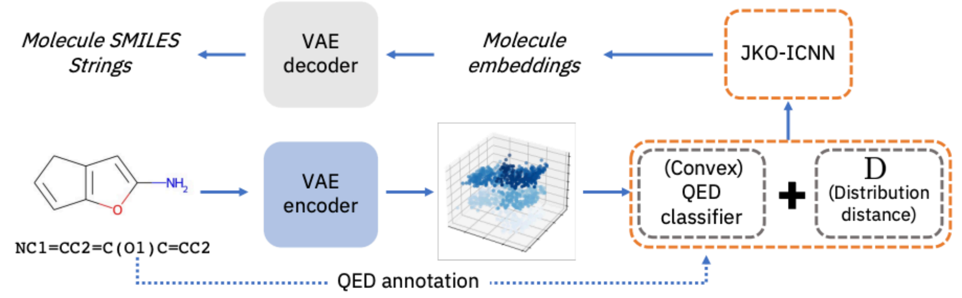
- Low-dim PDEs w/ analytic (full or asymptotic) solution
- Porous medium equation: $\partial_t \rho = \Delta \rho^m, m > 1$



JKO-ICNN flow tracks true solution, distributionally… …and in objective value!

- Additional results: Fokker-Planck and Aggregation Eqs.

Application: Molecule Discovery



- Goal: Transport molecular embeddings to areas with desirable properties (convex potential V), while staying near original set, as measured by divergence D:

$$\min_{\rho \in \mathscr{P}(\mathscr{X})} F(\rho) := \lambda_1 \mathbb{E}_\rho V(x) + \lambda_2 \mathrm{D}(\rho, \rho_0)$$

λ_2	LR	Validity	Uniqueness	QED Median	Final SD
ρ_0					
N/A	N/A	100.000 ± 0.000	99.980 ± 0.045	0.630 ± 0.001	N/A
JKO-ICNN					
$1e^4$	$1e^{-4}$	93.940 ± 0.336	100.000 ± 0.000	0.750 ± 0.001	0.620 ± 0.010
Baseline - SGD					
0	$5e^{-1}$	43.440 ± 1.092	100.000 ± 0.000	0.772 ± 0.004	9792.93 ± 76.913
1	$5e^{-1}$	49.440 ± 1.128	100.000 ± 0.000	0.768 ± 0.006	8881.38 ± 69.736
$1e^3$	$5e^{-1}$	87.240 ± 0.777	100.000 ± 0.000	0.767 ± 0.002	2515.08 ± 49.870
Baseline - ADAM					
0	$1e^{-1}$	92.080 ± 0.973	100.000 ± 0.000	0.793 ± 0.005	18.261 ± 0.134
0	$1e^{-2}$	93.900 ± 0.781	99.979 ± 0.048	0.758 ± 0.006	1.650 ± 0.006
1	$1e^{-1}$	91.200 ± 0.539	99.978 ± 0.049	0.792 ± 0.005	17.170 ± 0.097
$1e^3$	$1e^{-1}$	99.980 ± 0.045	99.980 ± 0.045	0.630 ± 0.001	0.077 ± 0.003
$1e^4$	$1e^{-1}$	99.900 ± 0.122	99.980 ± 0.045	0.630 ± 0.001	0.240 ± 0.019