NEURAL INFORMATION PROCESSING SYSTEMS

Microsoft

IBM Research

Optimizing Functionals on the Space of Probabilities with Input Convex Neural Networks

David Alvarez-Melis (MSR)

Yair Schiff (IBM Watson)

Youssef Mroueh (IBM Research AI)

OTML @ NeurIPS 2021

Motivation:
Distribution fitting

Many problems in ML amount to optimizing over distributions.
E.g., generative modeling:

$$\min_{\rho_\theta \in \mathscr{F}} \mathrm{D}(\rho_\theta, \rho_{\text{data}})$$

parametrized class of distributions ← $\rho_\theta \in \mathscr{F}$

D → Divergence:

- Jensen-Shannon $\Longrightarrow$ GAN
- 1-Wasserstein $\Longrightarrow$ WGAN
- KL $\Longrightarrow$ Maximum Likelihood

More generally:

$$\min_{\rho \in \mathscr{P}(\mathscr{X})} F(\rho)$$

Note we might not have samples of optimal ρ^*, known only implicitly as minimizer of F

F → a functional $F : \mathscr{P}(\mathscr{X}) \to \mathbb{R}$

how to optimize this?

Approach: follow gradient flow of F using JKO scheme [Jordan et al. '98], parametrized via ICNN [Amos et al. '17]

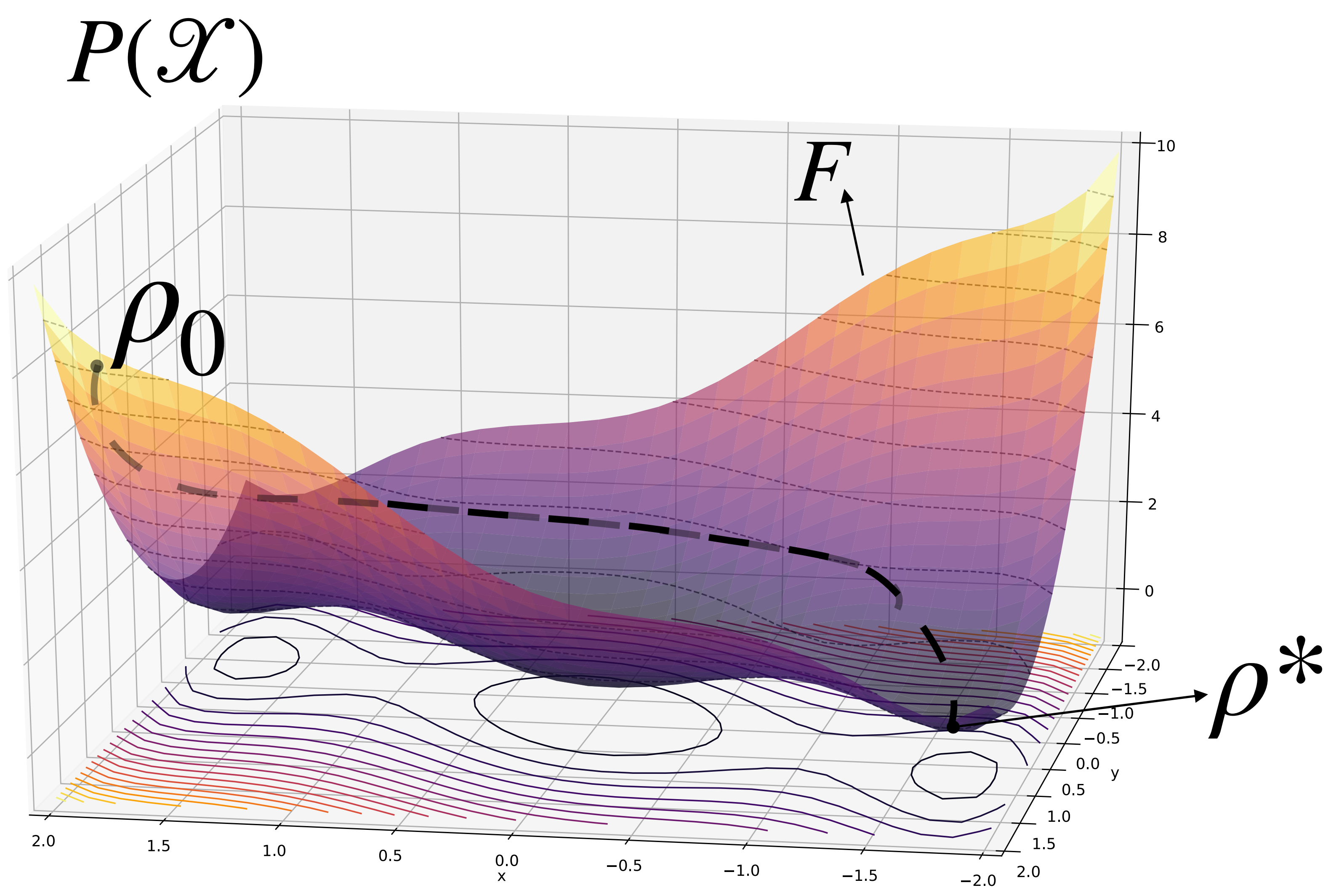
Background:
Gradient Flows in Wasserstein Space

Gradient Flow: curve of steepest descent of some functional F

In probability space: [Ambrosio et al. '05; Santambrogio '17; Figalli; Villani; etc]

$$\partial_t \rho(t) = -\nabla_{\mathbb{W}_2} F(\rho(t)) = -\nabla \cdot \left(\rho(t) \nabla \frac{\delta F}{\delta \rho}(\rho(t)) \right)$$

$$\rho(0) = \rho_0$$



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

Equivalence between PDE's and Gradient Flows

Our Approach:
JKO-ICNN

Setting: $\min_{\rho \in \mathcal{P}(\mathcal{X})} F(\rho)$, one of: $\mathcal{V}(\rho) = \int V(x) d\rho$ (potential), $\mathcal{W}(\rho) = \frac{1}{2} \iint W(x - x') d\rho \otimes \rho$ (interaction), $\mathcal{F}(\rho) = \int f(\rho(x)) dx$ (internal energy)

Base: JKO scheme to discretize gradient flow in probability space: $\rho_{t+1}^{\tau} \in \arg\min_{\rho \in \mathbb{W}_2(\mathcal{X})} F(\rho) + \frac{1}{2\tau} \mathrm{W}_2^2(\rho, \rho_t^{\tau})$

From Measures to Convex Functions

Under some assumptions, Brenier theorem yields:

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

So JKO scheme can be written as [Benamou et al. '14]:

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

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

From Convex Functions to ICNN

Parametrize CVX w/ input-convex neural nets [Amos et al. '17]:

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

Simple form for potential/interaction functionals:

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

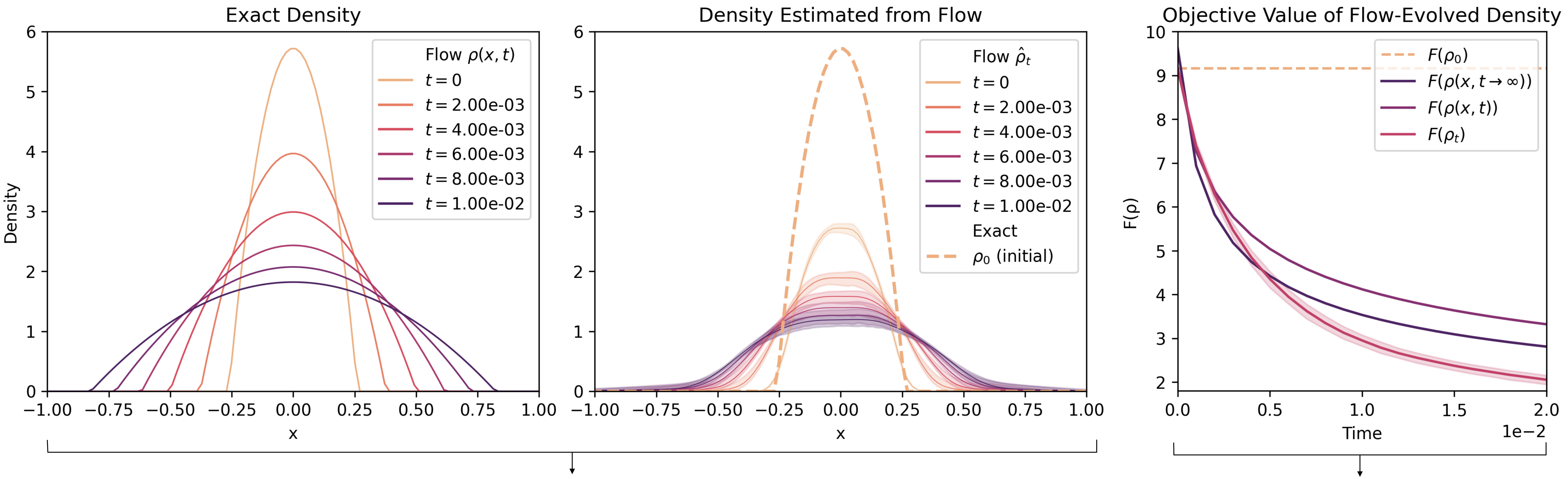
$$\mathcal{W}\left((\nabla_x u_\theta)_\sharp \rho_t^{\tau}\right) = \frac{1}{2} \mathbb{E}_{x,y \sim \rho_t^{\tau}} W(\nabla_x u_\theta(x) - \nabla_x u_\theta(y))$$

Surrogate objectives for certain internal energies

Evaluation:
Evolving PDEs with known solutions

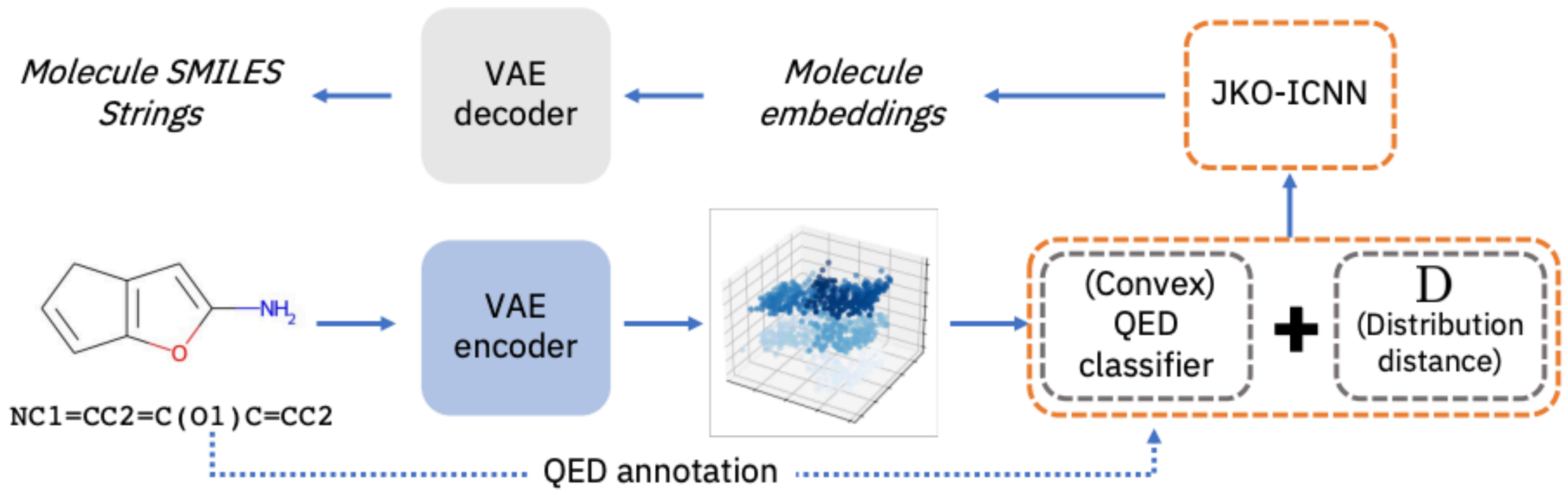
Porous medium equation: $\partial_t \rho = \Delta \rho^m, m > 1$, corresponds to gradient flow of $\mathscr{F}(\rho) = \frac{1}{m-1}\int \rho^m(x)dx$

Family of exact solutions: Barenblatt profiles $\rho(x,t) = t^{-\alpha}\left(C - k\|x\|^2 t^{-2\beta}\right)_+^{\frac{1}{m-1}}, \quad x \in \mathbb{R}^d, t > 0$



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

Application:
Molecule Discovery



Goal: Transport molecular embeddings to areas with desirable properties (encoded via convex potential V) **while** staying close to original (feasible) distribution

Functional:

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

encodes 'drug-likeness' (QED) — enforce proximity to original molecules

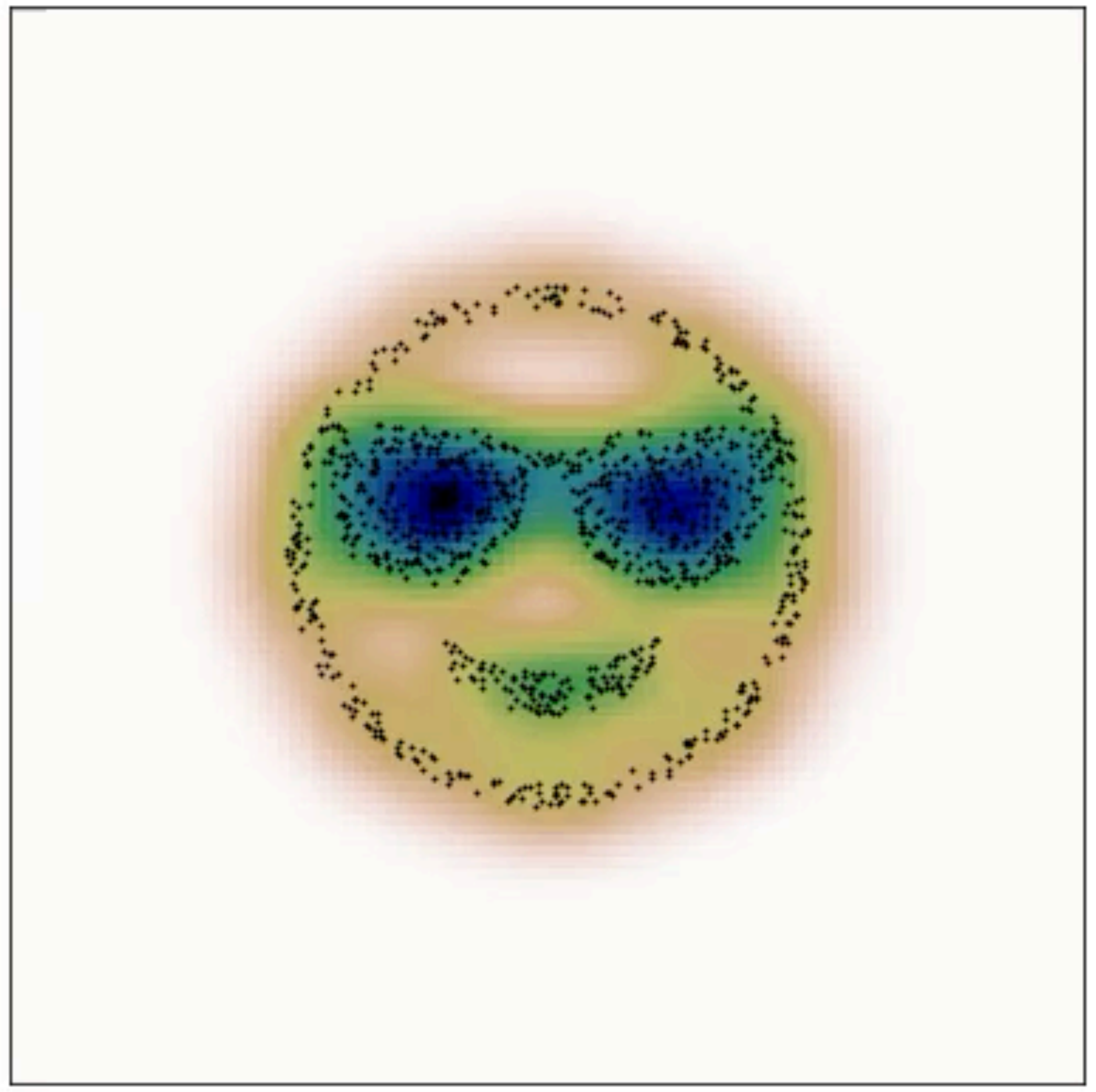
λ_2	LR	Validity	Uniqueness	QED Median	Final SD
ρ_0					
N/A	N/A	100.000 ± 0.000	99.980 ± 0.045	0.630 ± 0.001	N/A
JKO-ICNN					
$1e^4$	$1e^{-4}$	93.940 ± 0.336	100.000 ± 0.000	0.750 ± 0.001	0.620 ± 0.010
Baseline - SGD					
0	$5e^{-1}$	43.440 ± 1.092	100.000 ± 0.000	0.772 ± 0.004	9792.93 ± 76.913
1	$5e^{-1}$	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

Our method provides a strictly better tradeoff between the two objectives

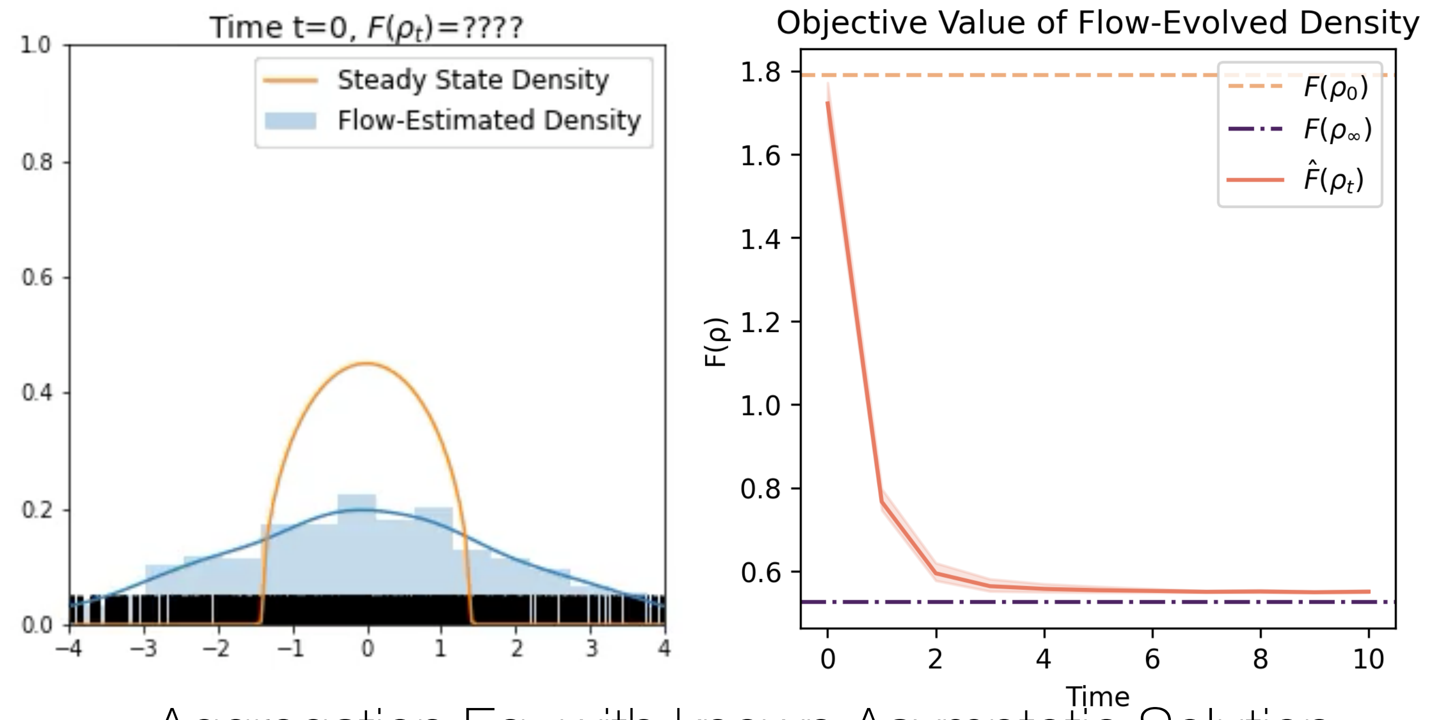
"Direct" (particle) optimization of F

Wrap-Up
and pointers

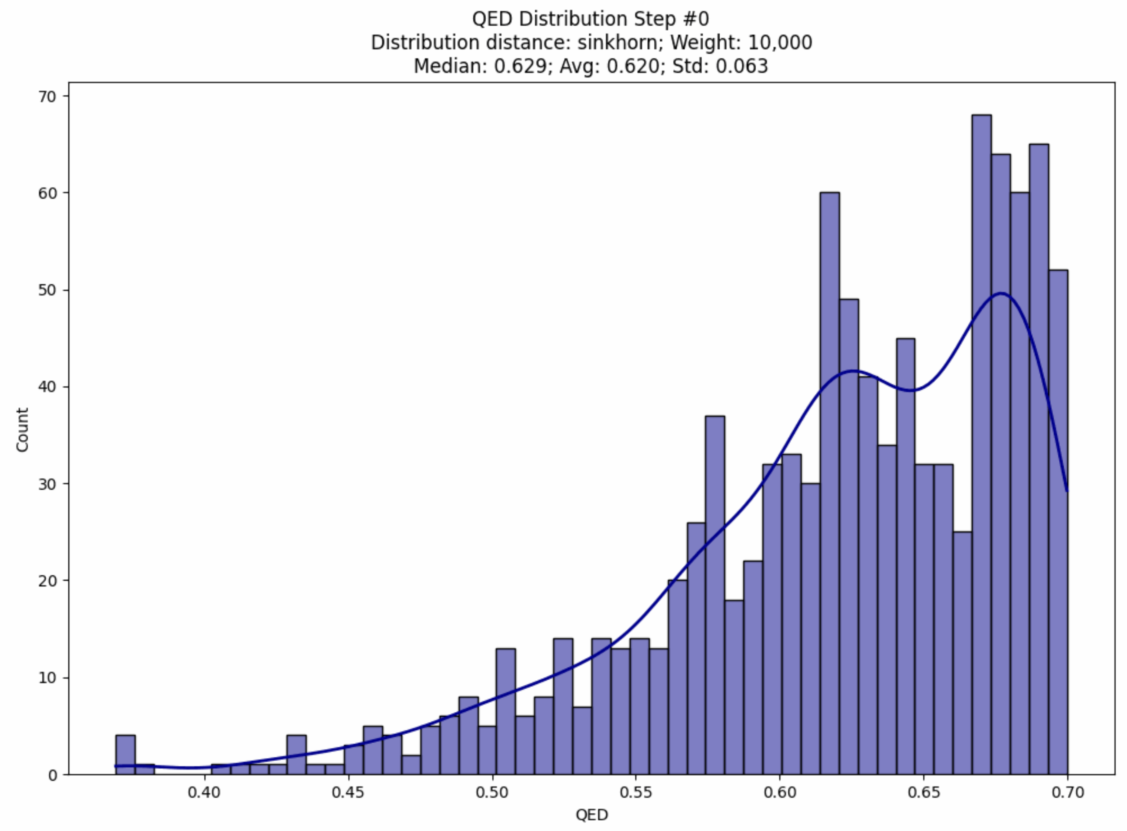
See the paper (arXiv:2106.00774) for more experiments … … and implementation details



Aggregation/Fokker-Planck/Heat EQ. In 2D



Aggregation Eq. with known Asymptotic Solution



Lots more experiments with molecule generation

- surrogate objectives
- density estimation
- out-of-sample mapping
- efficient computation