Gradient Flows for Unsupervised Learning

— with Connections to GANs —

Yu-Jui Huang
University of Colorado, Boulder

Joint work with
Yuchong Zhang (*University of Toronto*)

Graduate Institute of Statistics
National Central University

June 6, 2023

UNSUPERVISED LEARNING

- $\mathcal{P}(\mathbb{R}^d)$: the set of density functions on $\mathbb{R}^d$.
- $\rho_{\mathrm{d}} \in \mathcal{P}(\mathbb{R}^d)$: the (unknown) data distribution.
- **Unsupervised Learning:**

$$\inf_{\rho \in \mathcal{P}(\mathbb{R}^d)} d(\rho, \rho_{\mathrm{d}}),$$

where $d(\cdot, \cdot)$ is a metric on $\mathcal{P}(\mathbb{R}^d)$.

LITERATURE

- **Traditional method:** Parametrize ρ_{d}.
  - Parameter fitting by maximum likelihood estimations.
- **Generative adversarial network (GAN):**

  A min-max game between *generator* & *discriminator*.
  - Proposed by Goodfellow et al. (2014), actively studied by

    Dziugaite, Roy, & Ghahramani (2015), Nowozin, Cseke, & Tomioka (2016), Arjovsky, Chintala, & Bottou (2017), Li, Chang, Cheng, Yang, & Poczos (2017), Farnia & Tse (2019), Feizi, Farnia, Ginart, & Tse (2020), ...

    Financial time series, trading strategies:
    Wiese, Knobloch, Korn, & Kretschmer (2020), Koshiyama, Firoozye, & Treleaven (2021), Eckerli & Osterrieder (2021),...

Goodfellow et al. (2014):

$$\min_{G} \max_{D} \left\{ \mathbb{E}_{X\sim\rho_{\mathrm{d}}}[\ln D(X)] + \mathbb{E}_{Z\sim\rho^Z}\left[\ln\left(1 - D\left(G(Z)\right)\right)\right] \right\}. \qquad (1)$$

- This is equivalent to

$$\min_{G} \mathrm{JSD}(\rho^{G(Z)}, \rho_{\mathrm{d}}),$$

  - JSD is the *Jensen-Shannon divergence*

$$\mathrm{JSD}(\rho, \rho_{\mathrm{d}}) := \frac{1}{2} D_{\mathrm{KL}}\left(\rho_{\mathrm{d}} \,\middle\|\, \frac{\rho_{\mathrm{d}} + \rho}{2}\right) + \frac{1}{2} D_{\mathrm{KL}}\left(\rho \,\middle\|\, \frac{\rho_{\mathrm{d}} + \rho}{2}\right),$$

which involves the *Kullback-Leibler divergence*

$$D_{\mathrm{KL}}(\rho \| \bar{\rho}) := \int_{\mathbb{R}^d} \rho(x) \ln\left(\frac{\rho(x)}{\bar{\rho}(x)}\right) dx.$$

Algorithm 1 The GAN Algorithm

1: **for** number of training iterations **do**
2: • Sample m examples $\{z^{(1)}, ..., z^{(m)}\}$ from ρ^Z.
3: • Sample m examples $\{x^{(1)}, ..., x^{(m)}\}$ from ρ_{d}.
4: • Update $D : \mathbb{R}^d \to [0, 1]$ by ascending along

$$\nabla_{\theta_D} \frac{1}{m} \sum_{i=1}^{m} \left[\ln D(x^{(i)}) + \ln \left(1 - D\left(G(z^{(i)}) \right) \right) \right].$$

5: • Sample m examples $\{z^{(1)}, ..., z^{(m)}\}$ from ρ^Z.
6: • Update $G : \mathbb{R}^d \to \mathbb{R}^d$ by descending along

$$-\nabla_{\theta_G} \frac{1}{m} \sum_{i=1}^{m} \ln \left(1 - D(G(z^{(i)})) \right). \tag{2}$$

7: **end for**

LITERATURE

- **Drawback of GANs:** doesn't converge so easily...
  - Salimans et al. (2016): Empirical investigations.
  - Zhu, Jiao, & Tse (2020): The two-player game does *not* have a value:

$$\text{min-max game} \neq \text{max-min game}.$$

  - Cao & Guo (2020): SDE approximations.
  - Guo & Mounjid (2021): stochastic control framework.
  - Cao & Guo (2021): Review of analytical approaches.

Is there alternative (simplified) perspective for GANs?

IN THIS TALK...

- We study:

$$\text{minimize} \quad J(\rho) := \mathrm{JSD}(\rho, \rho_{\mathrm{d}}) \quad \text{over } \mathcal{P}(\mathbb{R}^d). \tag{3}$$

- The basics:
  - $0 \leq \mathrm{JSD}(\cdot, \cdot) \leq \ln(2)$.
  - Convergence in JSD $\iff$ in total variation $\iff$ in $L^1(\mathbb{R}^d)$
  - $\rho \mapsto \mathrm{JSD}(\rho, \bar{\rho})$ is *strictly convex*:

$$\mathrm{JSD}(\lambda\rho_1 + (1-\lambda)\rho_2, \bar{\rho}) < \lambda\,\mathrm{JSD}(\rho_1, \bar{\rho}) + (1-\lambda)\,\mathrm{JSD}(\rho_2, \bar{\rho}),$$

for any $\rho_1, \rho_2 \in \mathcal{P}(\mathbb{R}^d)$ and $\lambda \in (0,1)$.

CONVEX OPTIMIZATION

For a *strictly* convex $f : \mathbb{R}^d \to \mathbb{R}$,

- **gradient descent** works efficiently.
- For any $y \in \mathbb{R}^d$, the ODE

$$dY_t = -\nabla f(Y_t)dt, \quad Y_0 = y \in \mathbb{R}^d, \tag{4}$$

converges to global minimizer $y^* \in \mathbb{R}^d$ as $t \to \infty$.

Question: For the *strictly* convex

$$J(\cdot) = \mathrm{JSD}(\cdot, \rho_\mathrm{d}) : \mathcal{P}(\mathbb{R}^d) \to \mathbb{R},$$

can we also do **gradient descent** to find $\rho_\mathrm{d} \in \mathcal{P}(\mathbb{R}^d)$?

Given $G : \mathcal{P}(\mathbb{R}^d) \to \mathbb{R}$,

- **Gradient descent** in $\mathcal{P}(\mathbb{R}^d)$:

$$dY_t = -\partial_\rho G(\rho^{Y_t}, Y_t)dt, \quad \rho^{Y_0} = \rho_0 \in \mathcal{P}(\mathbb{R}^d). \tag{5}$$

  This is a *distribution-dependent* ODE.

  - Y_t is a random variable, with density $\rho^{Y_t} \in \mathcal{P}(\mathbb{R}^d)$.
  - $-\partial_\rho G(\rho^{Y_t}, \cdot) : \mathbb{R}^d \to \mathbb{R}^d$ dictates the direction along which each $y \in \mathbb{R}^d$ moves forward, i.e.,

  $\partial_\rho G(\rho, \cdot) : \mathbb{R}^d \to \mathbb{R}^d$ represents "gradient of G at $\rho \in \mathcal{P}(\mathbb{R}^d)$"

- **Challenge:** How do we define $\partial_\rho G(\rho, \cdot) : \mathbb{R}^d \to \mathbb{R}^d$?

  - $\mathcal{P}(\mathbb{R}^d)$ is not even a vector space
    $\implies$ Fréchet or Gateaux derivatives *not* well-defined.

Linear Functional Derivative

A *linear functional derivative* of $G : \mathcal{P}(\mathbb{R}^d) \to \mathbb{R}$ is a function $\frac{\delta G}{\delta \rho} : \mathcal{P}(\mathbb{R}^d) \times \mathbb{R}^d \to \mathbb{R}$ such that for all $\rho, \bar{\rho} \in \mathcal{P}(\mathbb{R}^d)$,

$$\lim_{\varepsilon \downarrow 0} \frac{G\big(\rho + \varepsilon(\bar{\rho} - \rho)\big) - G(\rho)}{\varepsilon} = \int_{\mathbb{R}^d} \frac{\delta G}{\delta \rho}(\rho, y)(\bar{\rho} - \rho)(y)dy.$$

- Relies only on convexity of $\mathcal{P}(\mathbb{R}^d)$.
- **Challenge remains:**
  Densities in ODE (5) don't evolve *linearly*....

- ODE (5) takes the form

$$dY_t = \xi(Y_t)dt, \quad \text{for some } \xi : \mathbb{R}^d \to \mathbb{R}^d.$$

- **Discretization:** Given a time step $\varepsilon > 0$, initial points $y \in \mathbb{R}^d$ are transported to

$$\bar{y} := y + \varepsilon\xi(y) = (I + \varepsilon\xi)(y) \in \mathbb{R}^d. \tag{6}$$

  - If $y \in \mathbb{R}^d$ follows $\rho \in \mathcal{P}(\mathbb{R}^d)$, $\bar{y} \in \mathbb{R}^d$ will follow $\rho_\varepsilon^\xi \in \mathcal{P}(\mathbb{R}^d)$ with $\rho_\varepsilon^\xi(y) := \rho((I + \varepsilon\xi)^{-1}(y))/det(\mathrm{Id} + \varepsilon\nabla\xi)$, for $y \in \mathbb{R}^d$.

Proposition

Under suitable regularity of ρ, ξ, and $\frac{\delta G}{\delta\rho}$,

$$\lim_{\varepsilon\downarrow 0} \frac{G(\rho_\varepsilon^\xi) - G(\rho)}{\varepsilon} = \int_{\mathbb{R}^d} \left(\nabla\frac{\delta G}{\delta\rho}(\rho, y) \cdot \xi(y)\right) \rho(y)dy. \tag{7}$$

- We will take

$$\partial_\rho G(\rho,\cdot) = \nabla \frac{\delta G}{\delta \rho}(\rho,\cdot), \quad \rho \in \mathcal{P}(\mathbb{R}^d).$$

- **More general than the literature:**
  - *Lions* and *Wasserstein derivatives* are well-defined on

$$\mathcal{P}_2(\mathbb{R}^d) := \left\{ \rho \in \mathcal{P}(\mathbb{R}^d) : \int_{\mathbb{R}^d} y^2 \rho(y) dy < \infty \right\},$$

    but not on the general space $\mathcal{P}(\mathbb{R}^d)$.
  - All three kinds of derivatives coincide in $\mathcal{P}_2(\mathbb{R}^d)$, under suitable conditions on G. (Carmona & Delarue (2018))

GRADIENT DESCENT IN $\mathcal{P}(\mathbb{R}^d)$

With $G(\rho) = \boxed{J(\rho) := \mathrm{JSD}(\rho, \rho_{\mathrm{d}})}$,

$$dY_t = -\partial_\rho J(\rho^{Y_t})(Y_t)dt = -\nabla \frac{\delta J}{\delta \rho}(\rho^{Y_t}, Y_t)dt, \quad \rho^{Y_0} = \rho_0 \in \mathcal{P}(\mathbb{R}^d).$$

Lemma

$J : \mathcal{P}(\mathbb{R}^d) \to \mathbb{R}$ has a linear functional derivative, given by

$$\frac{\delta J}{\delta \rho}(\rho, y) = \frac{1}{2} \ln \left(\frac{2\rho(y)}{\rho_{\mathrm{d}}(y) + \rho(y)} \right), \ \forall \rho \in \mathcal{P}(\mathbb{R}^d), y \in \mathbb{R}^d. \quad (8)$$

DENSITY-DEPENDENT ODE

Gradient descent in $\mathcal{P}(\mathbb{R}^d)$:

$$dY_t = -\frac{1}{2}\left(\frac{\nabla \rho^{Y_t}(Y_t)}{\rho^{Y_t}(Y_t)} - \frac{\nabla \rho_{\mathrm{d}}(Y_t) + \nabla \rho^{Y_t}(Y_t)}{\rho_{\mathrm{d}}(Y_t) + \rho^{Y_t}(Y_t)}\right) dt,$$
$$\rho^{Y_0} = \rho_0 \in \mathcal{P}(\mathbb{R}^d). \qquad (9)$$

Our Goal

There exists a unique solution Y to ODE (9). Moreover,

$$\rho^{Y_t} \to \rho_{\mathrm{d}} \quad \text{in } L^1(\mathbb{R}^d), \quad \text{as } t \to \infty.$$

- If this holds, can find ρ_{d} by simulating ODE (9)!

How to find a solution Y to ODE (9)?

$$dY_t = -\frac{1}{2}\left(\frac{\nabla\rho^{Y_t}(Y_t)}{\rho^{Y_t}(Y_t)} - \frac{\nabla\rho_{\mathrm{d}}(Y_t) + \nabla\rho^{Y_t}(Y_t)}{\rho_{\mathrm{d}}(Y_t) + \rho^{Y_t}(Y_t)}\right) dt,$$
$$\rho^{Y_0} = \rho_0 \in \mathcal{P}(\mathbb{R}^d).$$

- McKean-Vlasov SDEs typically depend on $\mathcal{L}(Y_t)$.
  - $\mathcal{L}(Y_t)$: the law of Y_t.
  - An *interacting particle system* can approximate $\mathcal{L}(Y_t)$.
- ODE (9) does *not* depend on $\mathcal{L}(Y_t)$, but on
  - Radon-Nikodym derivative of $\mathcal{L}(Y_t)$ (i.e., ρ^{Y_t}),
  - Euclidean derivative of ρ^{Y_t} (i.e., $\nabla\rho^{Y_t}$).

$\Longrightarrow$ "interacting particle system" not so promising...

FOKKER-PLANCK EQUATION

If Y is a solution to (9), $u(t,\cdot) := \rho^{Y_t}(\cdot) \in \mathcal{P}(\mathbb{R}^d)$ heuristically satisfies **(nonlinear) Fokker-Planck equation**

$$\begin{aligned} \frac{\partial u}{\partial t}(t,y) &= \frac{1}{2}\,\mathrm{Div}\left(\left(\frac{\nabla u}{u} - \frac{\nabla \rho_{\mathrm{d}} + \nabla u}{\rho_{\mathrm{d}} + u}\right) u\right)(t,y), \\ &= \frac{1}{2}\left(-\,\mathrm{Div}\left(\frac{\nabla \rho_{\mathrm{d}} + \nabla u}{\rho_{\mathrm{d}} + u} u\right) + \Delta u\right)(t,y), \quad u(0,y) = \rho_0(y). \end{aligned} \tag{10}$$

Definition

$u : [0,\infty) \to \mathcal{P}(\mathbb{R}^d)$ is a *weak* solution to FP eqn (10), if $u(t,\cdot)$ is weakly differentiable for a.e. $t \geq 0$ s.t. $\forall \varphi \in C_c^{1,2}((0,\infty) \times \mathbb{R}^d)$,

$$\int_0^\infty \int_{\mathbb{R}^d} \left(\varphi_t + \frac{1}{2}\frac{\nabla \rho_{\mathrm{d}} + \nabla u}{\rho_{\mathrm{d}} + u} \cdot \nabla \varphi + \frac{1}{2}\Delta \varphi\right) u(t,y)dydt = 0.$$

Our Plan

- Motivated by Barbu & Röckner (2020), we will
  1) Find a solution u to Fokker-Planck equation (10).
  2) Use u to construct a solution Y to ODE (9).

Assume: $\rho_{\mathrm{d}} > 0$ on $\mathbb{R}^d$.

- Set $v(t,y) := \frac{u(t,y)}{\rho_{\mathrm{d}}(y)}$ on $[0,\infty) \times \mathbb{R}^d$. Then, (10) becomes

$$\boxed{v_t = \frac{1}{2}\Delta_{\mu_{\mathrm{d}}} \ln(1+v), \quad v(0) = \rho_0/\rho_{\mathrm{d}},} \tag{11}$$

with $\Delta_{\mu_{\mathrm{d}}} := \Delta + \nabla \ln \rho_{\mathrm{d}} \cdot \nabla$.

(μ_{d}: probability measure on $\mathbb{R}^d$ induced by $\rho_{\mathrm{d}} \in \mathcal{P}(\mathbb{R}^d)$)

Definition

$v : [0,\infty) \to L^1(\mathbb{R}^d, \mu_{\mathrm{d}})$ is a *weak solution* to (11) *w.r.t.* μ_{d}, if for any $\varphi \in C_c^{1,2}((0,\infty) \times \mathbb{R}^d)$,

$$\int_0^\infty \int_{\mathbb{R}^d} \left(v\varphi_t + \frac{1}{2}\ln(1+v)\Delta_{\mu_{\mathrm{d}}}\varphi \right)(t,y)\,d\mu_{\mathrm{d}}\,dt = 0.$$

Assume:

$$\rho_0 \leq \beta \rho_{\mathrm{d}} \quad \text{for some} \quad \beta > 0. \tag{12}$$

- Consider the operator

$$Av := -\frac{1}{2}\Delta_{\mu_{\mathrm{d}}} \ln(1+v) \quad \text{for } v \in D(A), \tag{13}$$

where

$$D(A) := \{v \in L^1(\mathbb{R}^d, \mu_{\mathrm{d}}) \cap H_0^1(\mathbb{R}^d, \mu_{\mathrm{d}}) : \\ 0 \leq v \leq \beta,\ Av \in L^1(\mathbb{R}^d, \mu_{\mathrm{d}})\}.$$

- (11) becomes **(nonlinear) Cauchy problem** in $L^1(\mathbb{R}^d, \mu_{\mathrm{d}})$:

$$\boxed{v_t + Av = 0, \quad v(0) = \rho_0/\rho_{\mathrm{d}}.} \tag{14}$$

Intuition:

- The solution to $v_t + Av = 0$ *should* be "$v(0)e^{-At}$".
- Interpret "$v(0)e^{-At}$" as

$$\lim_{n\to\infty} \left(I + \frac{t}{n}A\right)^{-n} v(0). \tag{15}$$

Questions:

- How to make sense of $(I + \frac{t}{n}A)^{-n}v(0)$?
- The limit (15) well-defined? Solves $v_t + Av = 0$?

Lemma

For any $\lambda > 0$ and $f \in \overline{D(A)}$, there exists a unique weak solution $w \in D(A)$ w.r.t. μ_{d} to

$$(I + \lambda A)w = f. \tag{16}$$

- Denote by $(I + \lambda A)^{-1}f$ the unique weak solution w to (16).

Lemma

The operator $A : D(A) \to L^1(\mathbb{R}^d, \mu_{\mathrm{d}})$ is accretive. That is,

$$\|v_1 - v_2\|_{L^1(\mathbb{R}^d,\mu_{\mathrm{d}})} \le \|(I + \lambda A)v_1 - (I + \lambda A)v_2\|_{L^1(\mathbb{R}^d,\mu_{\mathrm{d}})},$$

for all $v_1, v_2 \in D(A)$ and $\lambda > 0$.

- "Accretive" $\implies$ (15) exists and solves $v_t + Av = 0$ (based on Crandall-Liggett's theory; see Barbu (2010)).

Corollary

Assume $\boxed{\rho_0 \leq \beta \rho_{\mathrm{d}}}$ *for some* $\beta > 0$.

(i) *There is a weak solution* $v \in C([0,\infty); L^1(\mathbb{R}^d, \mu_{\mathrm{d}}))$ *to* (11) *w.r.t.* μ_{d} *s.t.* $0 \leq v \leq \beta$ *and*

$$v(t) = \lim_{n\to\infty} \left(I + \frac{t}{n} A\right)^{-n} v(0) \;\; \textit{in} \;\; L^1(\mathbb{R}^d, \mu_{\mathrm{d}}), \tag{17}$$

uniformly in $t \geq 0$ *on compact intervals.*

(ii) *There is a weak solution* $u : [0,\infty) \to \mathcal{P}(\mathbb{R}^d)$ *to* (10) *given by*

$$u(t) := \rho_{\mathrm{d}} v(t) \quad \forall t \geq 0, \quad \textit{with } v \textit{ from (i)}. \tag{18}$$

Moreover, $u \in C([0,\infty); L^1(\mathbb{R}^d))$.

Theorem

Assume $\boxed{\rho_0 \le \beta\rho_{\mathrm{d}}}$ *for some* $\beta > 0$. *Then,* $u : [0, \infty) \to \mathcal{P}(\mathbb{R}^d)$ *in* (18) *is the unique weak solution to FP eqn* (10) *among*

$$\mathcal{C} := \left\{ \eta \in C([0,\infty); L^1(\mathbb{R}^d)) : \eta/\rho_{\mathrm{d}} \in L^\infty_+([0,\infty) \times \mathbb{R}^d) \right\}. \quad (19)$$

- Relying on (i) uniqueness of solutions to (11), generalized from Brézis & Crandall (1979), and (ii) $\ln \rho_{\mathrm{d}} \in H^1(\mathbb{R}^d, \mu_{\mathrm{d}})$.

To construct a solution Y to ODE (9),

1) Replace $\rho^{Y_t}(Y_t)$ in ODE (9) by $u(t, Y_t)$, with u the unique weak solution to FP eqn. (10):

$$dY_t = -\frac{1}{2}\left(\frac{\nabla u(t, Y_t)}{u(t, Y_t)} - \frac{\nabla \rho_{\mathrm{d}}(Y_t) + \nabla u(t, Y_t)}{\rho_{\mathrm{d}}(Y_t) + u(t, Y_t)}\right) dt,$$
$$\rho^{Y_0} = \rho_0 \in \mathcal{P}(\mathbb{R}^d). \quad (20)$$

2) Find a solution Y to (20) such that

$$\rho^{Y_t} = u(t, \cdot) \in \mathcal{P}(\mathbb{R}^d) \quad \forall t \geq 0.$$

Challenge: What is a "solution" to ODE (20)?

Two levels of randomness at time 0:

- *Randomness* of initial point $y \in \mathbb{R}^d$ (through $\rho_0 \in \mathcal{P}(\mathbb{R}^d)$).
- Once initial point $y \in \mathbb{R}^d$ is sampled, there can be multiple solutions $t \mapsto Y_t$ to ODE (20) (with $Y_0 = y$ fixed).

$\implies$ *Randomness* of which continuous path $t \mapsto Y_t$ to pick, among those that solve ODE (20) (with $Y_0 = y$ fixed).

Idea: Use a probability measure $\mathbb{P}$ on the path space

$$(\Omega, \mathcal{F}) := (C([0,\infty);\mathbb{R}^d), \mathcal{B}(C([0,\infty);\mathbb{R}^d)))$$

to express the joint randomness!

Definition

A process $Y : [0,\infty) \times \Omega \to \mathbb{R}^d$ is a solution to ODE (9) if

$$Y_t(\omega) := \omega(t) \quad \forall (t,\omega) \in [0,\infty) \times \Omega,$$

and there is a probability measure $\mathbb{P}$ on $(\Omega, \mathcal{F})$ under which

(i) the density $\eta_t \in \mathcal{P}(\mathbb{R}^d)$ of $Y_t : \Omega \to \mathbb{R}^d$ exists and is weakly differentiable $\forall t \geq 0$, and $\eta_0 = \rho_0$ Leb-a.e.;

(ii) $\mathbb{P}(\Gamma) = 1$, with $\Gamma \subseteq \Omega$ defined as

$$\left\{ \omega \in \Omega : \omega(t) = \omega(0) - \frac{1}{2} \int_0^t \left(\frac{\nabla \eta_s}{\eta_s} - \frac{\nabla \rho_{\mathrm{d}} + \nabla \eta_s}{\rho_{\mathrm{d}} + \eta_s} \right)(\omega(s))ds, \ t \geq 0 \right\}.$$

- $\mathbb{P}$ samples continuous paths $\omega : [0,\infty) \to \mathbb{R}^d$ from Γ, in a way that $\omega(0)$ has density ρ_0.

Consider an SDE

$$dX_t = b(t, X_t)dt + \sigma(t, X_t)dW_t. \tag{21}$$

Superposition Principle [Trevisan (2016)]:

- If $\{\nu_t\}_{t\geq 0}$ is a weak solution to Fokker-Planck eqn. associated with (21) s.t.

$$\int_0^T \int_{\mathbb{R}^d} (|b(t,x)| + |\sigma\sigma^T(t,x)|)d\nu_t dt < \infty \quad T > 0, \tag{22}$$

  there exists $\mathbb{P}$ on $(\Omega, \mathcal{F}) = (C([0,\infty);\mathbb{R}^d), \mathcal{B}(C([0,\infty);\mathbb{R}^d)))$ such that

  (i) $\mathbb{P}$ is a solution to local martingale problem for (21) (i.e., $X_t(\omega) := \omega(t)$, $t \geq 0$, satisfies (21) under $\mathbb{P}$);

  (ii) $\mathbb{P} \circ (X_t)^{-1} = \nu_t$ for all $t \geq 0$.

In our case, condition (22) becomes

$$\int_0^T \int_{\mathbb{R}^d} \left| \frac{\nabla u}{u} - \frac{\nabla \rho_{\mathrm{d}} + \nabla u}{\rho_{\mathrm{d}} + u} \right| u \, dy dt = \int_0^T \int_{\mathbb{R}^d} \left| \frac{\nabla v}{1+v} \right| d\mu_{\mathrm{d}} dt$$
$$\leq \int_0^T \int_{\mathbb{R}^d} |\nabla v| \, d\mu_{\mathrm{d}} dt \leq \int_0^T \|\nabla v(t)\|^2_{L^2(\mathbb{R}^d, \mu_{\mathrm{d}})} dt.$$

Lemma

Assume $\boxed{\rho_0 \leq \beta \rho_{\mathrm{d}}}$ *for some* $\beta > 0$*. Then, the unique weak solution* $v \in C([0,\infty); L^1(\mathbb{R}^d, \mu_{\mathrm{d}}))$ *to* (11) *w.r.t.* μ_{d} *satisfies*

$$\int_0^\infty \|\nabla v\|^2_{L^2(\mathbb{R}^d, \mu_{\mathrm{d}})} dt \leq (1+\beta)\beta^2. \tag{23}$$

- Relying on approximation of v in (17).

Proposition

Assume $\boxed{\rho_0 \leq \beta\rho_{\mathrm{d}}}$ for some $\beta > 0$. Then, there exists a solution Y to ODE (9).

Theorem

Let Y be a solution to ODE (9) s.t. $\eta(t, y) := \rho^{Y_t}(y)$ satisfies

$$\eta \in \mathcal{C} \quad \textit{and} \quad \nabla\eta \in L^1_{\mathrm{loc}}([0, \infty) \times \mathbb{R}^d).$$

Then, $\eta(t, \cdot) = u(t, \cdot) \in \mathcal{P}(\mathbb{R}^d)$ for all $t \geq 0$, where $u(t, x)$ is the unique weak solution to FP eqn. (10).

- Weaker than standard "weak uniqueness" of SDEs.

- **Want:** minimize $J(\rho) := \mathrm{JSD}(\rho, \rho_{\mathrm{d}})$.
- **Hope:** *Gradient flow* in $\mathcal{P}(\mathbb{R}^d)$ converges to ρ_{d}, i.e.,

$$\rho^{Y_t} \to \rho_{\mathrm{d}} \quad \text{as } t \to \infty,$$

where Y is the solution to ODE (9).

Proposition

Let Y be the unique solution to ODE (9). For any $0 \le t_1 < t_2$,

$$J(\rho^{Y_{t_2}}) - J(\rho^{Y_{t_1}}) \le -\int_{t_1}^{t_2} \int_{\mathbb{R}^d} \left| \nabla \frac{\delta J}{\delta \rho}(\rho^{Y_t}, y) \right|^2 \rho^{Y_t}(y) dy dt \le 0.$$

- Relying on approximation of v in (17).
- **Gradient descent works!**

Lemma

Let Y be the unique solution to ODE (9). There exist $\{t_n\}_{n\in\mathbb{N}}$ in $[0,\infty)$ with $t_n \uparrow \infty$ s.t. $\rho^{Y_{t_n}} \to \rho_{\mathrm{d}}$ in $L^1(\mathbb{R}^d)$.

- **Proof ideas:**
  - By (23), $\|\nabla v(t_n)\|_{L^2(\mathbb{R}^d,\mu_{\mathrm{d}})} \to 0$.
  - So, $\nabla v(t_n) \to 0$ and thus $v(t_n) \to v_\infty \equiv$ constant.
  - As $\int_{\mathbb{R}^d} v(t_n) d\mu_{\mathrm{d}} = \int_{\mathbb{R}^d} u(t_n) dy = 1$, we have $\int_{\mathbb{R}^d} v_\infty d\mu_{\mathrm{d}} = 1$.
  - So, $v_\infty \equiv 1$. Then, $\rho^{Y_{t_n}} = u(t_n) = \rho_{\mathrm{d}} v(t_n) \to \rho_{\mathrm{d}}$.

Theorem

Let Y be the unique solution to ODE (9). Then,

$$\|\rho^{Y_t} - \rho_{\mathrm{d}}\|_{L^1(\mathbb{R}^d)} \downarrow 0, \quad \text{as } t \to \infty.$$

- Established under $\rho_{\mathrm{d}} > 0$ and $\ln \rho_{\mathrm{d}} \in H^1(\mathbb{R}^d, \mu_{\mathrm{d}})$.

SIMULATION OF ODE (9)

Now, we set out to simulate

$$dY_t = -\frac{1}{2}\left(\frac{\nabla\rho^{Y_t}(Y_t)}{\rho^{Y_t}(Y_t)} - \frac{\nabla\rho_{\mathrm{d}}(Y_t) + \nabla\rho^{Y_t}(Y_t)}{\rho_{\mathrm{d}}(Y_t) + \rho^{Y_t}(Y_t)}\right) dt,$$
$$\rho^{Y_0} = \rho_0 \in \mathcal{P}(\mathbb{R}^d).$$

Challenge: The dynamics involves *unknown* ρ_{d}!

Simulation of ODE (9)

1. **Approximate** Y_t **by** $G(Z)$
   - Z is a simple r.v. (e.g., Gaussian), *fixed* over time.
   - $G : \mathbb{R}^d \to \mathbb{R}^d$ is complicated and *updated* over time.
2. **Substituting** $\rho^{G(Z)}$ **for** ρ^{Y_t} **in ODE** (9) $\implies$

$$dY_t = \frac{\nabla D(Y_t)}{2(1 - D(Y_t))} dt, \quad \text{with} \quad D(\cdot) := \frac{\rho_{\mathrm{d}}(\cdot)}{\rho_{\mathrm{d}}(\cdot) + \rho^{G(Z)}(\cdot)}$$

- Goodfellow et al. (2014):
  - With G given, $D : \mathbb{R}^d \to [0, 1]$ is the unique maximizer of

$$\max_{D:\ \mathbb{R}^d \to [0,1]} \left\{ \mathbb{E}_{y \sim \rho_{\mathrm{d}}} [\ln D(y)] + \mathbb{E}_{z \sim \rho_Z} [\ln (1 - D(G(z)))] \right\}.$$

  - **Note:** 1st half of **GAN algorithm** (i.e, Algorithm 1) estimates D *without* knowledge ρ_{d} or $\rho^{G(Z)}$!

Algorithm 2 Simulating ODE (9)

for number of training iterations **do**

- Sample m examples $\{z^{(1)}, ..., z^{(m)}\}$ from ρ^Z.
- Sample m examples $\{x^{(1)}, ..., x^{(m)}\}$ from ρ_d.
- Update $D : \mathbb{R}^d \to [0,1]$ by ascending along

$$\nabla_{\theta_D} \frac{1}{m} \sum_{i=1,\dots,m} \left[\ln D(x^{(i)}) + \ln \left(1 - D\left(G(z^{(i)})\right)\right)\right].$$

- Sample m examples $\{z^{(1)}, ..., z^{(m)}\}$ from ρ^Z.
- Set $Y = \{y^{(1)}, ..., y^{(m)}\}$ by

$$y^{(i)} := G(z^{(i)}) + \frac{\nabla D(G(z^{(i)}))}{2(1 - D(G(z^{(i)})))} \varepsilon, \quad \forall i = 1, 2, ..., m. \tag{24}$$

- Update $G : \mathbb{R}^d \to \mathbb{R}^d$ by descending along

$$-\nabla_{\theta_G} \frac{1}{m} \sum_{i=1,\dots,m} |G(z^{(i)}) - y^{(i)}|^2. \tag{25}$$

end for

Proposition

Algorithm 2 is equivalent to **GAN algorithm** (i.e, Algorithm 1), up to adjustment of learning rates.

- This is because

$$\begin{aligned}\nabla_{\theta_G}\frac{1}{m}\sum_{i=1}^{m}|G(z^{(i)})-y^{(i)}|^2 &= \frac{2}{m}\sum_{i=1}^{m}(G(z^{(i)})-y^{(i)})\cdot\nabla_{\theta_G}G(z^{(i)})\\ &= \frac{2}{m}\sum_{i=1}^{m}\left(\frac{-\nabla D(G(z^{(i)}))}{2(1-D(G(z^{(i)})))}\varepsilon\right)\cdot\nabla_{\theta_G}G(z^{(i)})\\ &= \varepsilon\nabla_{\theta_G}\frac{1}{m}\sum_{i=1}^{m}\ln\left(1-D(G(z^{(i)}))\right).\end{aligned}$$

GAN algorithm performs simulation of ODE (9)!

IMPLICATIONS

A New Theoretic Framework for GANs:

Non-cooperative game
(*generator* v.s. *discriminator*)

min max

$\downarrow$

Cooperative game
(*navigator* with *calibrator*)

min

- Theoretic convergence to ρ_{d} established rigorously (complements Section 4.2 of Goodfellow et al. (2014)).

IMPLICATIONS

A New Cause for GANs to Diverge:

- We've shown the equivalence between
  - updating G in GAN algorithm (i.e., (2))
  - moving along ODE + MSE fitting (i.e., (24)-(25))
- *MSE fitting is too strong a criterion!*
  - MSE demands **point-wise similarity** :

$$G(z^{(i)}) \text{ close to } y^{(i)} \text{ for all } i = 1, 2, ..., m.$$

  - What's needed is only **set-wise similarity** :

$$\text{distribution of } \{G(z^{(i)})\}_{i=1}^{m} \text{ close to that of } \{y^{(i)}\}_{i=1}^{m}.$$

- Work in progress: algorithms based on a set-wise criterion.

THANK YOU!!

Q & A

Preprint available @ arXiv: 2205.02910

"GANs as Gradient Flows that Converge"