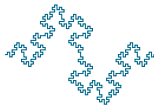


GANs as Gradient Flows that Converge

Yu-Jui Huang
University of Colorado, Boulder

Joint work with
Yuchong Zhang
University of Toronto



HK/SG Joint Seminar in Financial Mathematics & Engineering
July 21, 2022

UNSUPERVISED LEARNING

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

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

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, & Póczos (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):

$$\min_G \max_D \left\{ \mathbb{E}_{X \sim \rho_d} [\ln D(X)] + \mathbb{E}_{Z \sim \rho_Z} [\ln (1 - D(G(Z)))] \right\}. \quad (1)$$

► This is equivalent to

$$\min_G \text{JSD}(\rho^{G(Z)}, \rho_d),$$

► JSD is the *Jensen-Shannon divergence*

$$\text{JSD}(\rho, \rho_d) := \frac{1}{2} D_{\text{KL}} \left(\rho_d \left\| \frac{\rho_d + \rho}{2} \right. \right) + \frac{1}{2} D_{\text{KL}} \left(\rho \left\| \frac{\rho_d + \rho}{2} \right. \right),$$

which involves the *Kullback-Leibler divergence*

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

Algorithm 1 The GAN Algorithm

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

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

- 7: **end for**
-

LITERATURE

- ▶ **Drawback of GANs:** highly unstable.
 - ▶ 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:

$$\text{minimize } J(\rho) := \text{JSD}(\rho, \rho_d) \quad \text{over } \mathcal{P}(\mathbb{R}^d). \quad (2)$$

- ▶ The basics:

- ▶ $0 \leq \text{JSD}(\cdot, \cdot) \leq \ln(2)$.
- ▶ Convergence in JSD \iff in total variation \iff in $L^1(\mathbb{R}^d)$
- ▶ $\rho \mapsto \text{JSD}(\rho, \bar{\rho})$ is strongly convex: $\exists K > 0$ such that

$$\begin{aligned} \text{JSD}(\lambda\rho_1 + (1-\lambda)\rho_2, \bar{\rho}) &\leq \lambda \text{JSD}(\rho_1, \bar{\rho}) + (1-\lambda) \text{JSD}(\rho_2, \bar{\rho}) \\ &\quad - \underbrace{K\lambda(1-\lambda) \frac{\text{TV}(\rho_1, \rho_2)^2}{2}} \end{aligned}$$

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

CONVEX OPTIMIZATION

For a strongly convex $f : \mathbb{R}^d \rightarrow \mathbb{R}$,

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

$$dY_t = -\nabla f(Y_t)dt, \quad Y_0 = y \in \mathbb{R}^d, \quad (3)$$

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

Question: For the strongly convex

$$J(\cdot) = \text{JSD}(\cdot, \rho_d) : \mathcal{P}(\mathbb{R}^d) \rightarrow \mathbb{R},$$

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

DERIVATIVE IN $\rho \in \mathcal{P}_2(\mathbb{R}^d)$

- ▶ $\mathcal{P}_2(\mathbb{R}^d) := \{\rho \in \mathcal{P}(\mathbb{R}^d) : \text{2nd moment of } \rho < \infty\}$.
- ▶ Given $G : \mathcal{P}_2(\mathbb{R}^d) \rightarrow \mathbb{R}$, define $g : L^2(\Omega, \mathcal{F}, \mathbb{P}) \rightarrow \mathbb{R}$ by

$$g(Y) := G(\rho^Y) \quad \forall Y \in L^2(\Omega, \mathcal{F}, \mathbb{P}).$$

Carmona & Delarue (2018):

- 1) Take “derivative of G at $\rho_0 \in \mathcal{P}_2(\mathbb{R}^d)$ ” as

$$\underline{Dg(Y)} \in L^2 \quad (\text{Fréchet derivative of } g \text{ at } Y \in L^2 \text{ s.t. } \rho^Y = \rho_0).$$

- 2) $Dg(Y) = \underline{\xi_{\rho_0}(Y)}$, for some Borel $\xi_{\rho_0} : \mathbb{R}^d \rightarrow \mathbb{R}^d$.

- 3) **L-derivative** of G at $\rho_0 \in \mathcal{P}_2(\mathbb{R}^d)$: $\partial_\rho G(\rho_0)(\cdot) := \xi_{\rho_0}(\cdot)$

- ▶ **Gradient descent** in $\mathcal{P}_2(\mathbb{R}^d)$:

$$dY_t = -\partial_\rho G(\rho^{Y_t})(Y_t)dt, \quad \rho^{Y_0} = \rho_0 \in \mathcal{P}_2(\mathbb{R}^d).$$

Linear Functional Derivative

A *linear functional derivative* of $G : \mathcal{P}_2(\mathbb{R}^d) \rightarrow \mathbb{R}$ is a function $\frac{\delta G}{\delta \rho} : \mathcal{P}_2(\mathbb{R}^d) \times \mathbb{R}^d \rightarrow \mathbb{R}$ that satisfies

- (a) “continuity”, “growth” conditions...
- (b) for any $\rho, \bar{\rho} \in \mathcal{P}_2(\mathbb{R}^d)$,

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

► Carmona & Delarue (2018):

$$\partial_{\rho} G(\rho_0)(\cdot) = \nabla \frac{\delta G}{\delta \rho}(\rho_0, \cdot), \quad \rho_0 \in \mathcal{P}_2(\mathbb{R}^d).$$

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

With $G(\rho) = \boxed{J(\rho) := \text{JSD}(\rho, \rho_d)}$,

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

Lemma

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

$$\frac{\delta J}{\delta \rho}(\rho, y) = \frac{1}{2} \ln \left(\frac{2\rho(y)}{\rho_d(y) + \rho(y)} \right), \quad \forall \rho \in \mathcal{P}(\mathbb{R}^d), y \in \mathbb{R}^d. \quad (4)$$

- ▶ $\frac{\delta J}{\delta \rho}$ exists on $\mathcal{P}(\mathbb{R}^d)$, not the smaller $\mathcal{P}_2(\mathbb{R}^d)$!
- ▶ No a priori continuity, growth conditions on $\frac{\delta J}{\delta \rho}$.

DENSITY-DEPENDENT ODE

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

$$dY_t = -\frac{1}{2} \left(\frac{\nabla \rho^{Y_t}(Y_t)}{\rho^{Y_t}(Y_t)} - \frac{\nabla \rho_d(Y_t) + \nabla \rho^{Y_t}(Y_t)}{\rho_d(Y_t) + \rho^{Y_t}(Y_t)} \right) dt, \\ \rho^{Y_0} = \rho_0 \in \mathcal{P}(\mathbb{R}^d). \quad (5)$$

- ▶ 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 (5) 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., $\nabla \rho^{Y_t}$).
- ⇒ “interacting particle system” not so promising...

FOKKER-PLANCK EQUATION

If Y is a solution to (5), $u(t, \cdot) := \rho^{Y_t}(\cdot) \in \mathcal{P}(\mathbb{R}^d)$ heuristically satisfies **(nonlinear) Fokker-Planck equation**

$$\begin{aligned} \frac{\partial u}{\partial t}(t, y) &= \frac{1}{2} \operatorname{Div} \left(\left(\frac{\nabla u}{u} - \frac{\nabla \rho_d + \nabla u}{\rho_d + u} \right) u \right) (t, y), \\ &= \frac{1}{2} \left(-\operatorname{Div} \left(\frac{\nabla \rho_d + \nabla u}{\rho_d + u} u \right) + \Delta u \right) (t, y), \quad u(0, y) = \rho_0(y). \end{aligned} \tag{6}$$

Definition

$u : [0, \infty) \rightarrow \mathcal{P}(\mathbb{R}^d)$ is a weak solution to FP eqn (6), if $u(t, \cdot)$ is weakly differentiable for a.e. $t \geq 0$ s.t. $\forall \varphi \in C_c^{1,2}((0, \infty) \times \mathbb{R}^d)$,

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

OUR PLAN

- ▶ Motivated by Barbu & Röckner (2020), we will
 - 1) Find a solution u to Fokker-Planck equation (6).
 - 2) Use u to construct a solution Y to (5).

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

► Set $v(t, y) := \frac{u(t, y)}{\rho_d(y)}$ on $[0, \infty) \times \mathbb{R}^d$. Then, (6) becomes

$$\boxed{v_t = \frac{1}{2} \Delta_{\mu_d} \ln(1 + v), \quad v(0) = \rho_0 / \rho_d,} \quad (7)$$

with $\Delta_{\mu_d} := \Delta + \nabla \ln \rho_d \cdot \nabla$.

(μ_d : probability measure on \mathbb{R}^d induced by $\rho_d \in \mathcal{P}(\mathbb{R}^d)$)

Definition

$v : [0, \infty) \rightarrow L^1(\mathbb{R}^d, \mu_d)$ is a weak solution to (7) *w.r.t.* μ_d , if for any $\varphi \in C_c^{1,2}((0, \infty) \times \mathbb{R}^d)$,

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

Assume:

$$\rho_0 \leq \beta \rho_d \quad \text{for some } \beta > 0. \quad (8)$$

- Consider the operator

$$Av := -\frac{1}{2} \Delta_{\mu_d} \ln(1+v) \quad \text{for } v \in D(A), \quad (9)$$

where

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

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

$$\boxed{v_t + Av = 0, \quad v(0) = \rho_0 / \rho_d.} \quad (10)$$

Intuition:

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

$$\lim_{n \rightarrow \infty} \left(I + \frac{t}{n} A \right)^{-n} v(0). \quad (11)$$

Questions:

- ▶ How to make sense of $(I + \frac{t}{n} A)^{-n} v(0)$?
- ▶ The limit (11) 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

$$(I + \lambda A)w = f. \quad (12)$$

- Denote by $\underbrace{(I + \lambda A)^{-1}f}$ the unique weak solution \underbrace{w} to (12).

Lemma

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

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

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

- “Accretive” \implies (11) exists and solves $v_t + Av = 0$ (based on Crandall-Liggett’s theory; see Barbu (2010)).

Corollary

Assume $\rho_0 \leq \beta \rho_d$ for some $\beta > 0$.

- (i) There is a weak solution $v \in C([0, \infty); L^1(\mathbb{R}^d, \mu_d))$ to (7) w.r.t. μ_d s.t. $0 \leq v \leq \beta$ and

$$v(t) = \lim_{n \rightarrow \infty} \left(I + \frac{t}{n} A \right)^{-n} v(0) \text{ in } L^1(\mathbb{R}^d, \mu_d), \quad (13)$$

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

- (ii) There is a weak solution $u : [0, \infty) \rightarrow \mathcal{P}(\mathbb{R}^d)$ to (6) given by

$$u(t) := \rho_d v(t) \quad \forall t \geq 0, \quad \text{with } v \text{ from (i)}. \quad (14)$$

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

Theorem

Assume $\rho_0 \leq \beta \rho_d$ for some $\beta > 0$. Then, $u : [0, \infty) \rightarrow \mathcal{P}(\mathbb{R}^d)$ in (14) is the unique weak solution to FP eqn (6) among

$$\mathcal{C} := \left\{ \eta \in C([0, \infty); L^1(\mathbb{R}^d)) : \eta/\rho_d \in L^{\infty}_+([0, \infty) \times \mathbb{R}^d) \right\}. \quad (15)$$

- Relying on (i) uniqueness of solutions to (7), generalized from Brézis & Crandall (1979), and (ii) $\ln \rho_d \in H^1(\mathbb{R}^d, \mu_d)$.

To construct a solution Y to ODE (5),

- 1) Replace $\rho^{Y_t}(Y_t)$ in (5) by $u(t, Y_t)$, with u the unique weak solution to FP eqn. (6):

$$dY_t = -\frac{1}{2} \left(\frac{\nabla u(t, Y_t)}{u(t, Y_t)} - \frac{\nabla \rho_d(Y_t) + \nabla u(t, Y_t)}{\rho_d(Y_t) + u(t, Y_t)} \right) dt,$$
$$\rho^{Y_0} = \rho_0 \in \mathcal{P}(\mathbb{R}^d). \quad (16)$$

- 2) Find a solution Y to (16) 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 (16)?

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 (16) (with $Y_0 = y$ fixed).
 \implies *Randomness* of which continuous path $t \mapsto Y_t$ to pick, among those that solve ODE (16) (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 $Y : [0, \infty) \times \Omega \rightarrow \mathbb{R}^d$ is a solution to ODE (5) if

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

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

- (i) the density $\eta_t \in \mathcal{P}(\mathbb{R}^d)$ of $Y_t : \Omega \rightarrow \mathbb{R}^d$ 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

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

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

Consider an SDE

$$dX_t = b(t, X_t)dt + \sigma(t, X_t)dW_t. \quad (17)$$

Superposition Principle [Trevisan (2016)]:

- If $\{\nu_t\}_{t \geq 0}$ is a weak solution to **Fokker-Planck eqn.** associated with (17) s.t.

$$\int_0^T \int_{\mathbb{R}^d} (|b(t, x)| + |\sigma\sigma^T(t, x)|) d\nu_t dt < \infty \quad T > 0, \quad (18)$$

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 (17) (i.e., $Y_t(\omega) := \omega(t)$, $t \geq 0$, satisfies (17));
- (ii) $\mathbb{P} \circ (Y_t)^{-1} = \nu_t$ for all $t \geq 0$.

In our case, condition (18) becomes

$$\begin{aligned} \int_0^T \int_{\mathbb{R}^d} \left| \frac{\nabla u}{u} - \frac{\nabla \rho_d + \nabla u}{\rho_d + u} \right| u \, dy dt &= \int_0^T \int_{\mathbb{R}^d} \left| \frac{\nabla v}{1+v} \right| d\mu_d dt \\ &\leq \int_0^T \int_{\mathbb{R}^d} |\nabla v| \, d\mu_d dt \leq \int_0^T \|\nabla v(t)\|_{L^2(\mathbb{R}^d, \mu_d)}^2 dt. \end{aligned}$$

Lemma

Assume $\rho_0 \leq \beta \rho_d$ for some $\beta > 0$. Then, the unique weak solution $v \in C([0, \infty); L^1(\mathbb{R}^d, \mu_d))$ to (7) w.r.t. μ_d satisfies

$$\int_0^\infty \|\nabla v\|_{L^2(\mathbb{R}^d, \mu_d)}^2 dt \leq (1 + \beta)\beta^2. \quad (19)$$

- Relying on approximation of v in (13).

Proposition

Assume $\rho_0 \leq \beta \rho_d$ for some $\beta > 0$. Then, there exists a solution Y to ODE (5).

Theorem

Let Y be a solution to ODE (5) s.t. $\eta(t, y) := \rho^{Y_t}(y)$ satisfies

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

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. (6).

- ▶ Weaker than standard “weak uniqueness” of SDEs.

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

$$\rho^{Y_t} \rightarrow \rho_d \quad \text{as } t \rightarrow \infty,$$

where Y is the solution to ODE (5).

Proposition

Let Y be the unique solution to ODE (5). For any $0 \leq t_1 < t_2$,

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

- ▶ Relying on approximation of v in (13).
- ▶ **Gradient descent works!**

Lemma

Let Y be the unique solution to ODE (5). There exist $\{t_n\}_{n \in \mathbb{N}}$ in $[0, \infty)$ with $t_n \uparrow \infty$ s.t. $\rho^{Y_{t_n}} \rightarrow \rho_d$ in $L^1(\mathbb{R}^d)$.

► Proof ideas:

- By (19), $\|\nabla v(t_n)\|_{L^2(\mathbb{R}^d, \mu_d)} \rightarrow 0$.
- So, $\nabla v(t_n) \rightarrow 0$ and thus $v(t_n) \rightarrow v_\infty \equiv \text{constant}$.
- As $\int_{\mathbb{R}^d} v(t_n) d\mu_d = \int_{\mathbb{R}^d} u(t_n) dy = 1$, we have $\int_{\mathbb{R}^d} v_\infty d\mu_d = 1$.
- So, $v_\infty \equiv 1$. Then, $\rho^{Y_{t_n}} = u(t_n) = \rho_d v(t_n) \rightarrow \rho_d$.

Theorem

Let Y be the unique solution to (5). Then,

$$\rho^{Y_t} \rightarrow \rho_d \text{ in } L^1(\mathbb{R}^d) \text{ as } t \rightarrow \infty.$$

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

SIMULATION OF ODE (5)

Challenges:

- ▶ The dynamics involves *unknown* ρ_d .
- ▶ “Interacting particle system” does not easily work.

SIMULATION OF ODE (5)

1. Approximate Y_t by $G(Z)$

- ▶ Z is a simple r.v. (e.g., Gaussian), fixed over time.
- ▶ $G : \mathbb{R}^d \rightarrow \mathbb{R}^d$ is complicated and updated over time.

2. Substituting $\rho^{G(Z)}$ for ρ^{Y_t} in ODE (5) \implies

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

▶ Goodfellow et al. (2014):

- ▶ With G given, $D : \mathbb{R}^d \rightarrow [0, 1]$ is the unique maximizer of

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

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

Algorithm 2 Simulating ODE (5)

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

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

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

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

- 7: • Update $G : \mathbb{R}^d \rightarrow \mathbb{R}^d$ by descending along

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

- 8: **end for**
-

Proposition

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

- ▶ This is because

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

GAN algorithm performs simulation of ODE (5)!

IMPLICATIONS

► A new theoretic framework for GANs

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

min max



Cooperative game
(*navigator with calibrator*)

min

IMPLICATIONS

- ▶ **(Theoretic) Convergence of GANs**
 - ▶ Convergence to ρ_d established rigorously (complements Section 4.2 of Goodfellow et al. (2014)).
 - ▶ Likely the *fastest* possible convergence to ρ_d ...

THANK YOU!!

Q & A

Preprint available @ arXiv: 2205.02910
“GANs as Gradient Flows that Converge”