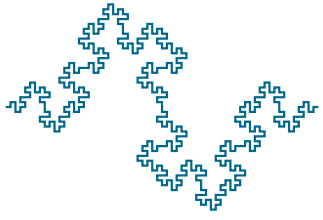
Generative Modeling by Minimizing the Wasserstein-2 Loss

Yu-Jui Huang
University of Colorado, Boulder

Joint work with
Zachariah Malik (*University of Colorado, Boulder*)



National Yang Ming Chiao Tung University
August 5, 2024

UNSUPERVISED LEARNING

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

$$\min_{\mu \in \mathcal{P}(\mathbb{R}^d)} d(\mu, \mu_{\mathrm{d}}),$$

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

- **Traditional method (Statistics):** Parametrize μ_{d}.
  - Parameter fitting by maximum likelihood estimations.
  - *What do we get?* A formula of μ_{d}.
- **Generative modeling:**
  - *What do we get?* A random variable X whose law is μ_{d}.
  - Usually, $X = G(Z)$, where
    - Z is a simple r.v. (e.g., Gaussian);
    - G is a complicated function (process) to be learned.
  - How to learn G?
    - Generative adversarial network, GAN
      (a min-max game between *generator* & *discriminator*.)
    - Diffusion probabilistic model
      (forward SDE to add noise, reverse-time SDE to denoise).

    ⋮

- **In this talk,**

$$\min_{\mu \in \mathcal{P}_2(\mathbb{R}^d)} W_2^2(\mu, \mu_{\mathrm{d}}) \tag{1}$$

  - $\mathcal{P}_2(\mathbb{R}^d) := \left\{\mu \in \mathcal{P}(\mathbb{R}^d) : \int_{\mathbb{R}^d} |y|^2 d\mu(y) < \infty\right\}$.
  - *Second-order Wasserstein distance:* for any $\mu, \nu \in \mathcal{P}_2(\mathbb{R}^d)$,

$$W_2(\mu, \nu) := \inf_{\gamma \in \Gamma(\mu,\nu)} \left(\int_{\mathbb{R}^d \times \mathbb{R}^d} |x - y|^2 \, d\gamma(x, y) \right)^{1/2} \tag{2}$$

$$= \sup_{\phi \in L^1(\mathbb{R}^d,\mu), (\phi^c)^c = \phi} \left\{ \int_{\mathbb{R}^d} \phi(x) \, d\mu(x) + \int_{\mathbb{R}^d} \phi^c(y) \, d\nu(y) \right\} \tag{3}$$

    - $\Gamma(\mu, \nu)$: the set of $\gamma \in \mathcal{P}(\mathbb{R}^d \times \mathbb{R}^d)$ with marginals μ and ν.
    - $\phi^c(y) := \inf_{x \in \mathbb{R}^d} \{\frac{1}{2}|x - y|^2 - \phi(x)\}, y \in \mathbb{R}^d$ [c-transform of ϕ]
  - **Optimizers exist!**
    - $\Gamma_0(\mu, \nu)$: the set of $\gamma \in \Gamma(\mu, \nu)$ that minimizes (2)
    - *Kantorovich potential* ϕ_μ^ν: a maximizer of (3).

- Why take W_2^2 in (1) (instead of W_2)?
  - $\mu \mapsto W_2^2(\mu, \mu_{\mathrm{d}})$ is *strictly convex* on $\mathcal{P}_2(\mathbb{R}^d)$.
- Recall: 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

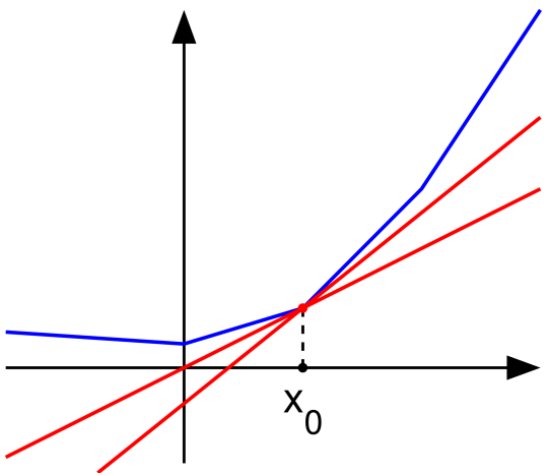
$$J(\cdot) := \frac{1}{2} W_2^2(\cdot, \mu_{\mathrm{d}}) : \mathcal{P}_2(\mathbb{R}^d) \to \mathbb{R}, \tag{5}$$

can we also do **gradient descent** to find $\mu_{\mathrm{d}} \in \mathcal{P}_2(\mathbb{R}^d)$?

SUBDIFFERENTIAL CALCULUS

- For a *convex* $f : \mathbb{R}^d \to \mathbb{R}$,
  - $\xi \in \mathbb{R}^d$ is called a **subgradient** of f at $x_0 \in \mathbb{R}^d$ if

$$f(y) \geq f(x_0) + \xi \cdot (y - x_0) \quad \forall y \in \mathbb{R}^d.$$



- The **subdifferential** of f at $x_0 \in \mathbb{R}^d$ (written $\partial f(x_0)$): the set of all such $\xi \in \mathbb{R}^d$

Consider a *convex* $G : \mathcal{P}_2(\mathbb{R}^d) \to \mathbb{R}$.

Definition

$\xi : \mathbb{R}^d \to \mathbb{R}^d$ is called a **subgradient** of G at $\mu \in \mathcal{P}_2(\mathbb{R}^d)$ if $\xi \in L^2(\mathbb{R}^d, \mu)$ and

$$G(\nu) \geq G(\mu) + \inf_{\gamma \in \Gamma_0(\mu,\nu)} \int_{\mathbb{R}^d} \xi(x) \cdot (y - x)\, d\gamma(x, y) + o(W_2(\mu, \nu)), \quad \forall \nu \in \mathcal{P}_2(\mathbb{R}^d).$$

- The **subdifferential** of G at $\mu \in \mathcal{P}_2(\mathbb{R}^d)$ (written $\partial G(\mu)$): the set of all such $\xi \in \mathbb{R}^d \to \mathbb{R}^d$.

For $J(\cdot) := \frac{1}{2}W_2^2(\cdot, \mu_{\mathrm{d}}) : \mathcal{P}_2(\mathbb{R}^d) \to \mathbb{R}$,

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

$$\frac{dY_t}{dt} \in -\partial J(\mu^{Y_t}), \quad \rho^{Y_0} = \mu_0 \in \mathcal{P}_2(\mathbb{R}^d). \tag{6}$$

  - μ^{Y_t} : the law of r.v. Y_t.
  - AMBROSIO ET AL. (2008): $\xi \in \partial J(\mu)$ generally exists, but...

    not unique and *no* explicit construction.
  - **Special case:** if $\mu \ll \mathcal{L}^d$, $\partial J(\mu) = \{\nabla \phi_\mu^{\mu_{\mathrm{d}}}\}$ is a singleton!

- **Assuming** $\mu^{Y_t} \ll \mathcal{L}^d$ for all $t \geq 0$, (6) becomes

$$dY_t = -\nabla \phi_{\mu^{Y_t}}^{\mu_{\mathrm{d}}}(Y_t)dt, \quad \mu^{Y_0} = \mu_0 \in \mathcal{P}_2^r(\mathbb{R}^d),$$

where $\mathcal{P}_2^r(\mathbb{R}^d) := \{\mu \in \mathcal{P}_2(\mathbb{R}^d) : \mu \ll \mathcal{L}^d\}$.

Our **gradient-descent** ODE:

$$dY_t = -\nabla \phi^{\mu_d}_{\mu^{Y_t}}(Y_t)dt, \quad \mu^{Y_0} = \mu_0 \in \mathcal{P}_2^r(\mathbb{R}^d). \tag{7}$$

- This ODE is *distribution-dependent*.
  - Y_t is a random variable, whose law is $\mu^{Y_t} \in \mathcal{P}_2(\mathbb{R}^d)$.
  - "$-\nabla \phi^{\mu_d}_{\mu^{Y_t}}(\cdot) : \mathbb{R}^d \to \mathbb{R}^d$" is a *vector field* at time t:
    it dictates how each $y \in \mathbb{R}^d$, a sample of Y_t, moves forward.

Our Goal:

There exists a unique solution Y to ODE (7). Moreover,

$$W_2(\mu^{Y_t}, \mu_d) \to 0, \quad \text{as } t \to \infty.$$

- **Challenge:** How to find a solution Y to (7)?
  - The regularity of $(y, \mu) \mapsto \nabla \phi_\mu^{\mu_d}(y)$ is obscure
    $\Longrightarrow$ results for McKean-Vlasov SDEs cannot be applied.
  - The solution Y needs to fulfill $\boxed{\mu^{Y_t} \ll \mathcal{L}^d \text{ for all } t \geq 0}$.

Our Strategy:

Don't work with ODE (7) directly.

Work with its **Fokker-Planck equation**.

If Y is a solution to ODE (7), $\boldsymbol{\mu}_t := \mu^{Y_t} \in \mathcal{P}_2(\mathbb{R}^d)$ heuristically satisfies **(nonlinear) Fokker-Planck equation**

$$\partial_t \boldsymbol{\mu}_t - \mathrm{Div}(\nabla \phi^{\mu_\mathrm{d}}_{\boldsymbol{\mu}_t} \boldsymbol{\mu}_t) = 0, \quad \boldsymbol{\mu}_0 = \mu_0 \in \mathcal{P}_2^r(\mathbb{R}^d) \tag{8}$$

in the sense of distributions, i.e., $\forall \varphi \in C_c^\infty((0,\infty) \times \mathbb{R}^d)$,

$$\int_0^\infty \int_{\mathbb{R}^d} \Big(\partial_t \varphi(t,x) - \nabla \phi^{\mu_\mathrm{d}}_{\boldsymbol{\mu}_t}(x) \cdot \nabla \varphi(t,x) \Big) d\boldsymbol{\mu}_t(x) dt = 0.$$

- **Our Plan:**
  1) Find a solution $\{\boldsymbol{\mu}_t\}_{t \geq 0}$ to (8) with $\boldsymbol{\mu}_t \ll \mathcal{L}^d \ \forall t \geq 0$.
  2) Construct a solution Y to ODE (7) with $\mu^{Y_t} = \boldsymbol{\mu}_t \ \forall t \geq 0$.

  - This plan, introduced in BARBU & RÖCKNER (2020), was also used in HUANG & ZHANG (2023, JMLR).

OPTIMAL TRANSPORT MAP

- Given $f : \mathbb{R}^d \to \mathbb{R}^d$ and $\mu \in \mathcal{P}(\mathbb{R}^d)$, define $f_{\#}\mu \in \mathcal{P}(\mathbb{R}^d)$ by

$$f_{\#}\mu(B) := \mu(f^{-1}(B)), \quad \forall \text{ Borel } B \subseteq \mathbb{R}^d.$$

- Given $\mu, \nu \in \mathcal{P}_2(\mathbb{R}^d)$,
  - $\boldsymbol{t} : \mathbb{R}^d \to \mathbb{R}^d$ is a *transport map from* μ *to* ν if $\boldsymbol{t}_{\#}\mu = \nu$. Note

$$W_2(\mu, \nu) = \inf_{\boldsymbol{t}:\mathbb{R}^d \to \mathbb{R}^d, \boldsymbol{t}_{\#}\mu = \nu} \int_{\mathbb{R}^d} |x - \boldsymbol{t}(x)|^2 d\mu(x).$$

  - A transport map $\boldsymbol{t}_\mu^\nu$ is *optimal* if it attains the infimum.
  - AMBROSIO ET AL. (2008): if $\mu \ll \mathcal{L}^d$, then

$$\boldsymbol{t}_\mu^\nu(x) = x - \nabla\phi_\mu^\nu(x) = (\boldsymbol{i} - \nabla\phi_\mu^\nu)(x) \quad \text{for } \mu\text{-a.e. } x \in \mathbb{R}^d, \quad (9)$$

    where $\boldsymbol{i} : \mathbb{R}^d \to \mathbb{R}^d$ is the identity map.

CONSTANT-SPEED GEODESIC

Given $\mu_0 \in \mathcal{P}_2^r(\mathbb{R}^d)$ and $\mu_{\mathrm{d}} \in \mathcal{P}_2(\mathbb{R}^d)$, define

$$\boldsymbol{\beta}_s := \left((1-s)\boldsymbol{i} + s\boldsymbol{t}_{\mu_0}^{\mu_{\mathrm{d}}}\right)_{\#}\mu_0 \in \mathcal{P}_2(\mathbb{R}^d), \quad s \in [0,1]. \tag{10}$$

- $\boldsymbol{\beta}_0 = \boldsymbol{i}_{\#}\mu_0 = \mu_0,\ \boldsymbol{\beta}_1 = \boldsymbol{t}_{\mu_0}^{\mu_{\mathrm{d}}}{}_{\#}\mu_0 = \mu_{\mathrm{d}}$.
- THEOREM 7.2.2, AMBROSIO ET AL. (2008):

  $\{\boldsymbol{\beta}_s\}_{s\in[0,1]}$ is a **constant-speed geodesic**, i.e.,

$$W_2(\boldsymbol{\beta}_{s_1}, \boldsymbol{\beta}_{s_2}) = (s_2 - s_1)W_2(\mu_0, \mu_{\mathrm{d}}), \quad \forall 0 \le s_1 \le s_2 \le 1. \tag{11}$$

- $\boxed{\beta_s \ll \mathcal{L}^d \text{ for all } s \in [0,1).}$

Proof Sketch:

- For any $s \in (0,1)$, by (10) and (9),

$$\begin{aligned}\beta_s &= (\boldsymbol{i} + s(\boldsymbol{t}_{\mu_0}^{\mu_d} - \boldsymbol{i}))_{\#}\mu_0 \\ &= (\boldsymbol{i} - s\nabla\phi_{\mu_0}^{\mu_d})_{\#}\mu_0 \\ &= (\nabla f_s)_{\#}\mu_0, \qquad \text{with} \quad f_s(x) := \tfrac{1}{2}|x|^2 - s\phi_{\mu_0}^{\mu_d}(x). \end{aligned} \tag{12}$$

- LEMMA 5.5.3, AMBROSIO ET AL. (2008):

$$\begin{cases} \mu \ll \mathcal{L}^d \\ g : \mathbb{R}^d \to \mathbb{R} \textit{ uniformly convex} \end{cases} \implies (\nabla g)_{\#}\mu \ll \mathcal{L}^d.$$

- Recall from (3) that $((\phi_{\mu_0}^{\mu_d})^c)^c = \phi_{\mu_0}^{\mu_d}$

$\implies f(x) := \frac{1}{2}|x|^2 - \phi_{\mu_0}^{\mu_d}(x)$ is convex

$\implies f_s(x) = \frac{1}{2}|x|^2 - s\phi_{\mu_0}^{\mu_d}(x)$ is uniformly convex.

CHANGE OF SPEED

Consider *time-changed* constant-speed geodesic

$$\boldsymbol{\mu}_t^* = \boldsymbol{\beta}_{1-e^{-t}} = \left(e^{-t}\boldsymbol{i} + (1-e^{-t})\boldsymbol{t}_{\mu_0}^{\mu_\mathrm{d}}\right)_{\#}\mu_0, \quad t \in [0,\infty). \tag{13}$$

- $\boldsymbol{\mu}_0^* = \boldsymbol{\beta}_0 = \mu_0$, $\lim_{t\to\infty} \boldsymbol{\mu}_t^* = \boldsymbol{\beta}_1 = \mu_\mathrm{d}$.
- $\boxed{\boldsymbol{\mu}_t^* \ll \mathcal{L}^d \text{ for all } t \geq 0}$.
- $\boldsymbol{\mu}^* \in AC^p((0,\infty); \mathcal{P}_2(\mathbb{R}^d))$ for all $p \geq 1$.
- **Exponential convergence to μ_d:**

$$W_2(\boldsymbol{\mu}_t^*, \mu_\mathrm{d}) = W_2(\boldsymbol{\beta}_{1-e^{-t}}, \boldsymbol{\beta}_1) = e^{-t} W_2(\mu_0, \mu_\mathrm{d}), \quad t \geq 0. \tag{14}$$

Question: Does $\boldsymbol{\mu}^*$ satisfy **Fokker-Planck equation** (8)?

- For any $\boldsymbol{\mu} = \{\boldsymbol{\mu}_t\}_{t\geq 0} \in AC^2_{\text{loc}}((0,\infty); \mathcal{P}_2(\mathbb{R}^d))$,

$$\begin{aligned}\left|\frac{d}{dt}J(\boldsymbol{\mu}_t)\right| &= \lim_{h\to 0}\left|\frac{J(\boldsymbol{\mu}_{t+h}) - J(\boldsymbol{\mu}_t)}{h}\right| \\ &= \lim_{h\to 0}\left|\frac{J(\boldsymbol{\mu}_{t+h}) - J(\boldsymbol{\mu}_t)}{W_2(\boldsymbol{\mu}_{t+h}, \boldsymbol{\mu}_t)}\frac{W_2(\boldsymbol{\mu}_{t+h}, \boldsymbol{\mu}_t)}{h}\right| \\ &\leq \limsup_{\nu\to\boldsymbol{\mu}_t}\left|\frac{J(\nu) - J(\boldsymbol{\mu}_t)}{W_2(\nu, \boldsymbol{\mu}_t)}\right| \lim_{h\to 0}\frac{W_2(\boldsymbol{\mu}_{t+h}, \boldsymbol{\mu}_t)}{h} := M(t) < \infty.\end{aligned}$$

- By (11)-(13), $\frac{d}{dt}J(\boldsymbol{\mu}_t^*) = -M(t)\ \forall t \geq 0$. [*steepest descent!!*]

- **Idea:** (a) $\implies$ (b), where
  (a) Steepest descent of $t \mapsto J(\boldsymbol{\mu}_t^*)$
  (b) $\boldsymbol{\mu}_t^*$ evolves along "negative gradient of J at $\boldsymbol{\mu}_t^*$"

  [cf. THEOREM 11.1.3, AMBROSIO ET AL. (2008)]

- **Conclusion:** the *vector field* of $\boldsymbol{\mu}^*$ is "$-\nabla\phi_{\boldsymbol{\mu}_t^*}^{\mu_{\text{d}}}$".

Proposition

$\{\boldsymbol{\mu}_t^*\}_{t\geq 0}$ in (13) is a solution to the **F-P equation** (8), i.e.,

$$\partial_t \boldsymbol{\mu}_t^* - \mathrm{Div}(\nabla \phi_{\boldsymbol{\mu}_t^*}^{\mu_\mathrm{d}} \boldsymbol{\mu}_t^*) = 0, \quad \boldsymbol{\mu}_0^* = \mu_0 \in \mathcal{P}_2^r(\mathbb{R}^d). \tag{15}$$

Moreover, for $\mathcal{L}^1$-a.e. $t > 0$,

$$\|\nabla \phi_{\boldsymbol{\mu}_t^*}^{\mu_\mathrm{d}}\|_{L^2(\mathbb{R}^d, \boldsymbol{\mu}_t^*)} = \lim_{s\to t} \frac{W_2(\boldsymbol{\mu}_s^*, \boldsymbol{\mu}_t^*)}{|s-t|} < \infty. \tag{16}$$

If $\boldsymbol{\mu}$ is a solution to (8) s.t. $\boldsymbol{\mu} \in AC^2_{\mathrm{loc}}((0,\infty), \mathcal{P}_2(R^d))$ and (16) holds, then

$$W_2(\boldsymbol{\mu}_t, \boldsymbol{\mu}_t^*) = 0, \quad \forall t \geq 0.$$

- **Note:** By (11), (16) becomes

$$\|\nabla \phi_{\boldsymbol{\mu}_t^*}^{\mu_\mathrm{d}}\|_{L^2(\mathbb{R}^d, \boldsymbol{\mu}_t^*)} = e^{-t} W_2(\mu_0, \mu_\mathrm{d}). \tag{17}$$

SOLVING ODE (7)

- Replace μ^{Y_t} in ODE (7) by $\boldsymbol{\mu}_t^*$ in (13)

$$dY_t = -\nabla \phi_{\boldsymbol{\mu}_t^*}^{\mu_{\mathrm{d}}}(Y_t)dt, \quad \mu^{Y_0} = \mu_0 \in \mathcal{P}_2^r(\mathbb{R}^d). \tag{18}$$

  - An ODE of standard form—*no* distribution dependence!

- Look for a solution Y to (18) such that

$$\mu^{Y_t} = \boldsymbol{\mu}_t^* \quad \forall t \geq 0.$$

Consider an SDE

$$dX_t = b(t, X_t)dt + \sigma(t, X_t)dW_t. \tag{19}$$

Superposition Principle [TREVISAN (2016)]:

- If $\{\nu_t\}_{t\geq 0}$ is a solution to Fokker-Planck eqn. associated with (19) s.t.

$$\int_0^T \int_{\mathbb{R}^d} (|b(t,x)| + |\sigma\sigma^T(t,x)|)d\nu_t dt < \infty \quad T > 0, \tag{20}$$

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 (19) (i.e., $X_t(\omega) := \omega(t)$, $t \geq 0$, satisfies (19) under $\mathbb{P}$);

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

To check condition (20) in our case,

$$\begin{aligned}\int_0^T \int_{\mathbb{R}^d} \left|\nabla \phi^{\mu_{\mathrm{d}}}_{\boldsymbol{\mu}^*_t}(y)\right| \, d\boldsymbol{\mu}^*_t(y)dt &= \int_0^T \|\nabla \phi^{\mu_{\mathrm{d}}}_{\boldsymbol{\mu}^*_t}\|_{L^1(\mathbb{R}^d,\boldsymbol{\mu}^*_t)}dt \\ &\leq \int_0^T \|\nabla \phi^{\mu_{\mathrm{d}}}_{\boldsymbol{\mu}^*_t}\|_{L^2(\mathbb{R}^d,\boldsymbol{\mu}^*_t)}dt \\ &= \int_0^T e^{-t} W_2(\mu_0,\mu_{\mathrm{d}})dt < \infty \qquad \forall T > 0,\end{aligned}$$

where the last equality follows from (17).

MAIN RESULTS

Theorem

There is a solution Y to ODE (7) s.t. $\mu^{Y_t} = \boldsymbol{\mu}_t^*$ for all $t \geq 0$. Thus,

$$W_2(\mu^{Y_t}, \mu_{\mathrm{d}}) = W_2(\boldsymbol{\mu}_t^*, \mu_{\mathrm{d}}) = e^{-t} W_2(\mu_0, \mu_{\mathrm{d}}), \quad \forall t \geq 0.$$

- The last equality is due to (14).

Proposition

Let Y be a solution to ODE (7) s.t. $\boldsymbol{\mu}_t := \mu^{Y_t} \in \mathcal{P}_2^r(\mathbb{R}^d)$ satisfies $\boldsymbol{\mu} \in AC((0,\infty); \mathcal{P}_2(\mathbb{R}^d)) \cap AC^2_{\mathrm{loc}}((0,\infty); \mathcal{P}_2(\mathbb{R}^d))$ and (16). Then,

$$W_2(\boldsymbol{\mu}_t, \boldsymbol{\mu}_t^*) = 0 \quad \forall t \geq 0.$$

FORWARD EULER SCHEME

Fix a time step $0 < \varepsilon < 1$. Let's discretize ODE

$$dY_t = -\nabla \phi^{\mu_{\mathrm{d}}}_{\mu^{Y_t}}(Y_t)dt, \quad \mu^{Y_0} = \mu_0 \in \mathcal{P}^r_2(\mathbb{R}^d).$$

- Time 0: Simulate r.v. Y^ε_0 following μ_0, so that

$$\mu^{Y^\varepsilon_0} = \mu_{0,\varepsilon} := \mu_0 \in \mathcal{P}^r_2(\mathbb{R}^d)$$

- Time ε: Simulate r.v. $Y^\varepsilon_1 := Y^\varepsilon_0 - \varepsilon \nabla \phi^{\mu_{\mathrm{d}}}_{\mu_0}(Y^\varepsilon_0)$, so that

$$\mu^{Y^\varepsilon_1} = \mu_{1,\varepsilon} := (\boldsymbol{i} - \varepsilon \nabla \phi^{\mu_{\mathrm{d}}}_{\mu_0})_{\#}\mu_0.$$

- Time 2ε: Simulate r.v. $Y^\varepsilon_2 := Y^\varepsilon_1 - \varepsilon \nabla \phi^{\mu_{\mathrm{d}}}_{\mu_{1,\varepsilon}}(Y^\varepsilon_1)$, so that

$$\mu^{Y^\varepsilon_2} = \mu_{2,\varepsilon} := (\boldsymbol{i} - \varepsilon \nabla \phi^{\mu_{\mathrm{d}}}_{\mu_{1,\varepsilon}})_{\#}\mu_{1,\varepsilon}.$$

- That is, we simulate recursively the r.v.'s

$$Y_n^\varepsilon := Y_{n-1}^\varepsilon - \varepsilon \nabla \phi^{\mu_{\mathrm{d}}}_{\mu_{n-1,\varepsilon}}(Y_{n-1}^\varepsilon) \quad n \in \mathbb{N},$$

  which have the distributions

$$\mu^{Y_n^\varepsilon} = \mu_{n,\varepsilon} := (\boldsymbol{i} - \varepsilon \nabla \phi^{\mu_{\mathrm{d}}}_{\mu_{n-1,\varepsilon}})_{\#}\mu_{n-1,\varepsilon} \quad n \in \mathbb{N}.$$

- This yields flow of measures $\boldsymbol{\mu}^\varepsilon : [0,\infty) \to \mathcal{P}_2(\mathbb{R}^d)$ given by

$$\boldsymbol{\mu}_t^\varepsilon := \mu_{n,\varepsilon} \quad \text{if} \quad t \in [n\varepsilon, (n+1)\varepsilon), \qquad \forall n \in \mathbb{N} \cup \{0\}.$$

Theorem

For any $t \in [0,\infty)$, $\lim_{\varepsilon \downarrow 0} W_2(\boldsymbol{\mu}_t^\varepsilon, \boldsymbol{\mu}_t^*) = 0$.

- Euler scheme does converge to the right limit!

SIMULATION OF ODE (7)

Now, we set out to simulate

$$dY_t = -\nabla \phi^{\mu_{\mathrm{d}}}_{\mu^{Y_t}}(Y_t)dt, \quad \mu^{Y_0} = \mu_0 \in \mathcal{P}_2^r(\mathbb{R}^d).$$

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

SIMULATION OF ODE (7)

1. **Approximate Y_t by $G_\theta(Z)$**
   - Z is a simple r.v. (e.g., Gaussian), *fixed* over time.
   - $G_\theta : \mathbb{R}^d \to \mathbb{R}^d$ is complicated and *updated* over time.
2. **Substituting $\mu^{G_\theta(Z)}$ for μ^{Y_t} in ODE (7)** $\Longrightarrow$

$$dY_t = -\nabla \phi^{\mu_d}_{\mu^{G_\theta(Z)}}(Y_t)dt, \quad \mu^{Y_0} = \mu_0 \in \mathcal{P}_2^r(\mathbb{R}^d). \tag{21}$$

   - With G_θ given (i.e., $G_\theta(Z)$ represents Y_t),
     - estimate $\left(\phi^{\mu_d}_{\mu^{G_\theta(Z)}}, (\phi^{\mu_d}_{\mu^{G_\theta(Z)}})^c\right)$ by two neural networks (ϕ_w, ψ_v), following ALGORITHM 1, SEGUY ET AL. (2018).
   - Once $\phi^{\mu_d}_{\mu^{G_\theta(Z)}}$ is known, move along ODE (21) to time $t + \varepsilon$.
   - Update G_θ such that $G_\theta(Z)$ can suitably represent $Y_{t+\varepsilon}$.

Algorithm 1 W2-FE

for number of training iterations **do**
  for number of updates **do**
    • Sample $\{z_i\}_{i=1}^m$ from μ^Z, and $\{x_i\}_{i=1}^m$ from μ_{d}.
    • $w \leftarrow w + \gamma_d \nabla_w L_D$ and $v \leftarrow v + \gamma_d \nabla_v L_D$, with

$$L_D = \frac{1}{m} \sum_{i=1}^{m} \Big[\phi_w(G_\theta(z_i)) + \psi_v(x_i) - \lambda \big(\phi_w(G_\theta(z_i)) + \psi_v(x_i) - ||G_\theta(z_i) - x_i||_2^2/2 \big)_+ \Big].$$

  end for
  • Sample $\{z_i\}_{i=1}^m$ from μ^Z.
  • $y_i \leftarrow G_\theta(z_i)$ [Samples of Y_t]
  • $\zeta_i \leftarrow y_i - \varepsilon \nabla \phi_w(y_i)$. [Samples of $Y_{t+\varepsilon}$]
  for K generator updates **do**
    • $\theta \leftarrow \theta - \frac{\gamma_g}{m} \nabla_\theta \sum_{i=1}^m |\zeta_i - G_\theta(z_i)|^2$
  end for
end for

PERSISTENT TRAINING

To update the "*generator*" G_θ,

- Reduce **mean squared error** between $\{G_\theta(z_i)\}$ and $\{\zeta_i\}$.
  - **Idea:** Want $G_\theta(Z)$ to *learn* the distribution of $Y_{t+\varepsilon}$.
- Do stochastic gradient descent (SGD) $K \in \mathbb{N}$ times *with same minibatch*!
  - $K = 1$: standard SGD.
  - $K > 1$: **persistent training** (FISCHETTI ET AL. (2018)).
    - *Conceptually*
      —ideal for our differential equation approach.
    - *Numerically*
      —speeds up training and achieves better results.

CONNECTIONS TO GANS

Generative adversarial networks (GANs) also solves

$$\min_{\mu \in \mathcal{P}(\mathbb{R}^d)} d(\mu, \mu_{\mathrm{d}}),$$

by a *min-max* game (between *generator* and *discriminator*).

- GOODFELLOW ET AL. (2014) (**vanilla GAN**):
  - $d =$ Jensen-Shannon divergence
  - highly unstable...
- ARJOVSKY ET AL. (2017) (**WGAN**):
  - $d =$ Wasserstein-1 distance
  - Stable, one of the most popular GANs.
- LEYGONIE ET AL. (2019) (**W2GAN**):
  - $d =$ Wasserstein-2 distance
  - Performs similarly to WGAN, receives little attention...

Proposition

Algorithm 1 (with $K = 1$) is equivalent to **W2GAN algorithm** (LEYGONIE ET AL. (2019)), up to adjustment of learning rates.

- This is because

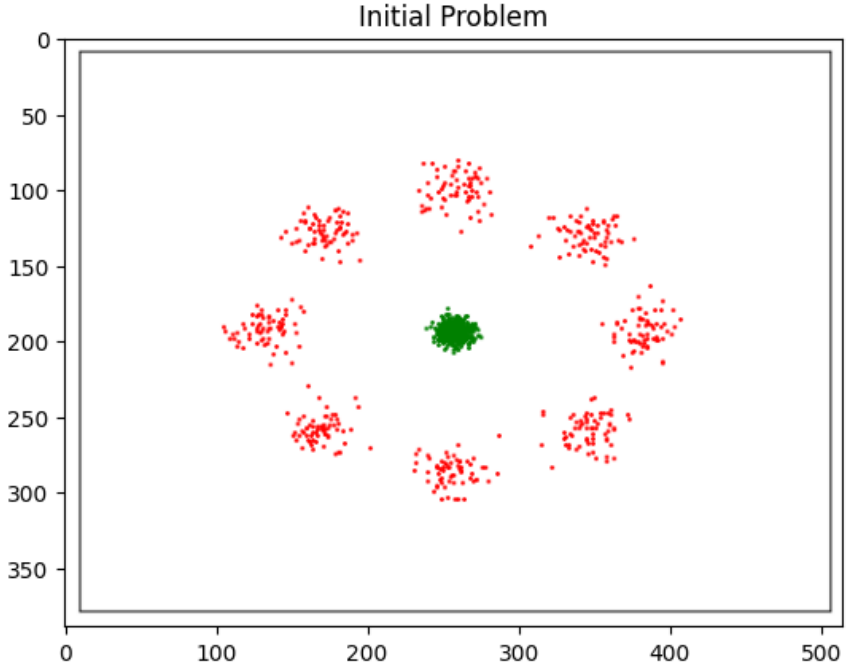
$$\begin{aligned}\nabla_\theta \frac{1}{m}\sum_{i=1}^{m} |G_\theta(z_i) - \zeta_i|^2 &= \frac{2}{m}\sum_{i=1}^{m} (G_\theta(z_i) - \zeta_i) \cdot \nabla_\theta G_\theta(z_i) \\ &= \frac{2}{m}\sum_{i=1}^{m} \varepsilon \nabla\phi\left(G_\theta(z_i)\right) \cdot \nabla_\theta G_\theta(z_i) \\ &= \varepsilon \nabla_\theta \frac{1}{m}\sum_{i=1}^{m} \phi\left(G_\theta(z_i)\right)\end{aligned}$$

- Equivalence *fails* for $K > 1$!

EXPERIMENT 1

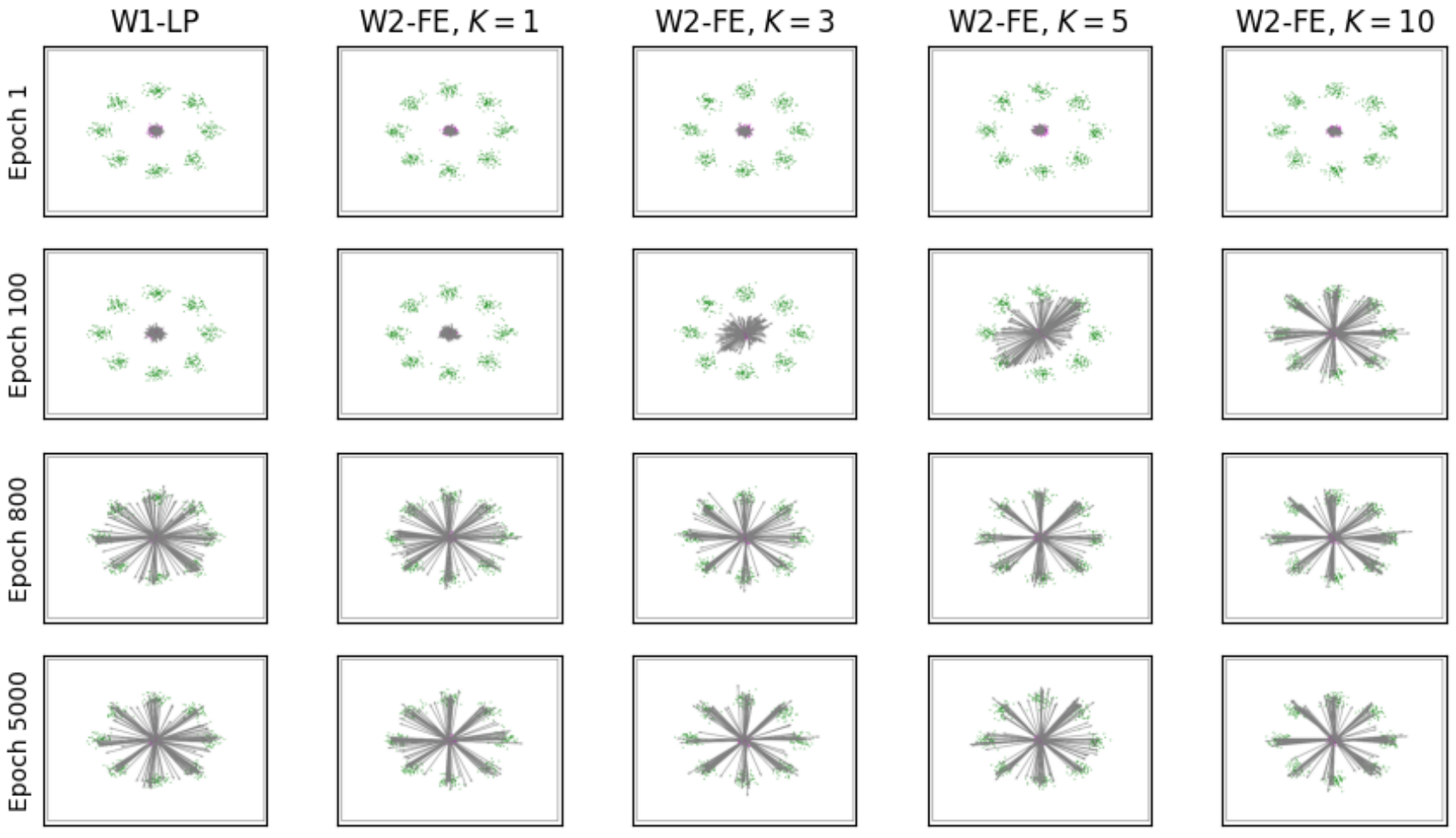
The Ring Example (METZ ET AL. (2017, ICLR)):

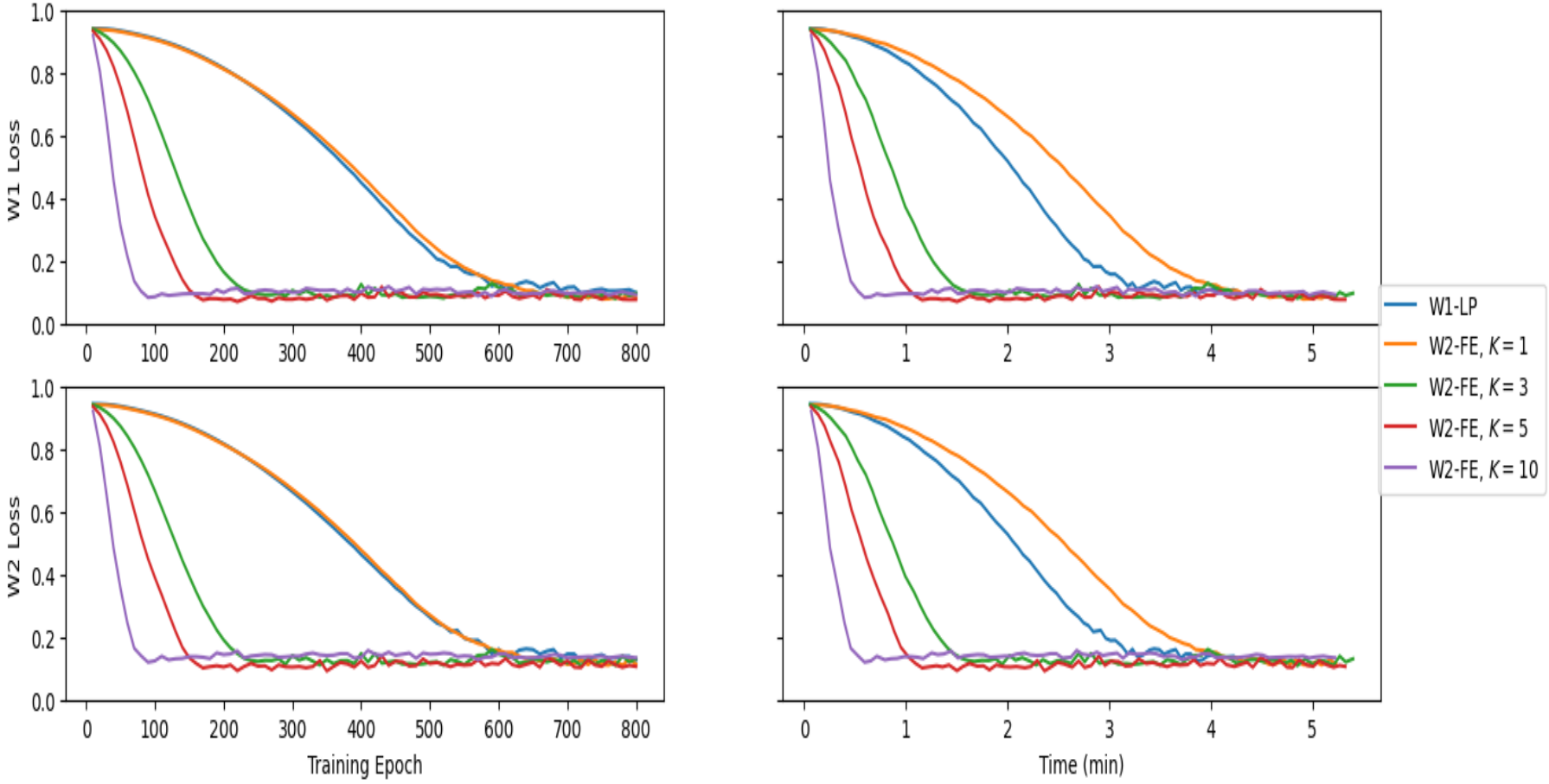
- True data distribution: 8 Gaussians arranged on a circle
- Initial distribution: 1 Gaussian at the center



- In the literature:
  - **(Vanilla) GAN:** Extreme mode collapse!!
  - **WGAN:** No obvious mode collapse, but low resolution
  - **Refined WGAN (W1-LP)** in PETZKA ET AL. (2018): No mode collapse, high resolution.

- **What we will do:**
  - Implement **Algorithm 1 (W2-FE)** with varying K values.
  - Compare it performance with **W1-LP**.

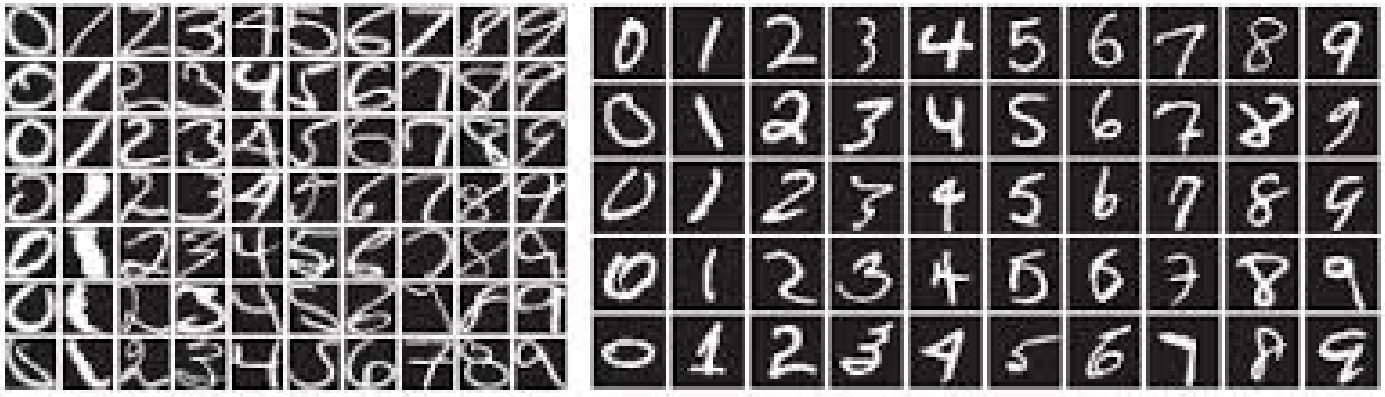

W1-LP
W2-FE, $K = 1$
W2-FE, $K = 3$
W2-FE, $K = 5$
W2-FE, $K = 10$
Epoch 1
Epoch 100
Epoch 800
Epoch 5000


W1 Loss
W2 Loss
Training Epoch
Time (min)
W1-LP
W2-FE, K = 1
W2-FE, K = 3
W2-FE, K = 5
W2-FE, K = 10

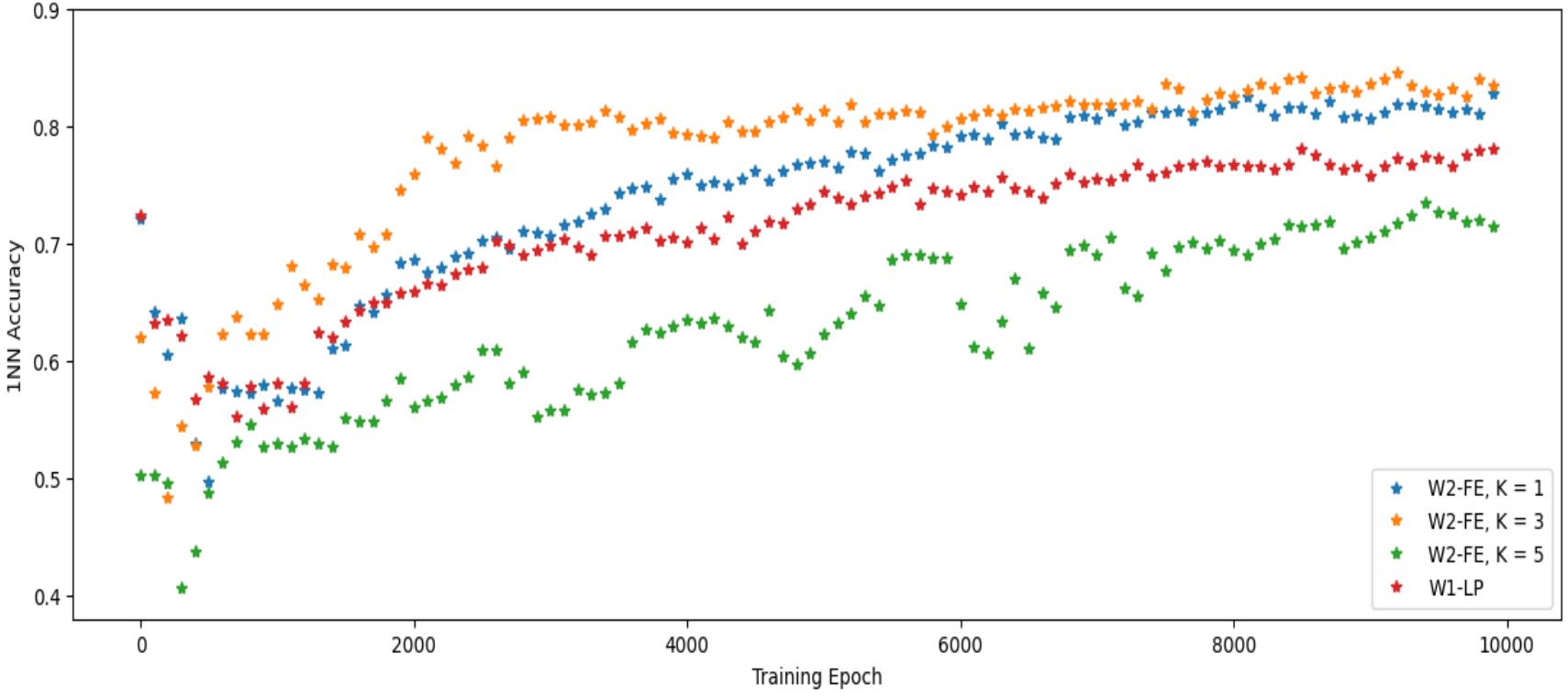
EXPERIMENT 2

Domain Adaptation

- True data distribution: MNIST dataset
- Initial distribution: USPS dataset



(a) USPS (b) MNIST



- **Accuracy measure:** 1-nearest neighbor (1-NN) classifier.
  - evaluated every 100 epochs.
- Performance deteriorates at $K = 5$ (due to overfitting).

SUMMARY

- We solve the problem

$$\min_{\mu \in \mathcal{P}_2(\mathbb{R}^d)} W_2^2(\mu, \mu_{\mathrm{d}})$$

  through a *distribution-dependent* ODE

$$dY_t = -\nabla \phi_{\mu^{Y_t}}^{\mu_{\mathrm{d}}}(Y_t)dt, \quad \mu^{Y_0} = \mu_0 \in \mathcal{P}_2^r(\mathbb{R}^d).$$

- A **generative model** is built by simulating the ODE.
  It combines:
  - A forward Euler scheme for the ODE,
  - *Persistent training* to learn μ^{Y_t}.
- Our algorithm outperforms typical Wasserstein GAN algorithms.

EXTENSION TO W_1

- Consider

$$\min_{\mu \in \mathcal{P}_1(\mathbb{R}^d)} W_1(\mu, \mu_{\mathrm{d}}) \tag{22}$$

  - $\mathcal{P}_1(\mathbb{R}^d) := \{\mu \in \mathcal{P}(\mathbb{R}^d) : \int_{\mathbb{R}^d} |y| d\mu(y) < \infty\}$.
  - W_1 is the *first-order Wasserstein distance*

$$W_1(\mu, \nu) := \inf_{\gamma \in \Gamma(\mu,\nu)} \int_{\mathbb{R}^d \times \mathbb{R}^d} |x - y| \, d\gamma(x, y)$$

$$= \sup_{\phi:\mathbb{R}^d \to \mathbb{R},\ \|\phi\|_{Lips} \leq 1} \left(\int_{\mathbb{R}^d} \phi(x) d\mu(x) - \int_{\mathbb{R}^d} \phi(y) d\nu(y) \right).$$

  - Maximizer ϕ_μ^ν exists!
    (also called *Kantorovich potential*).

EXTENSION TO W_1

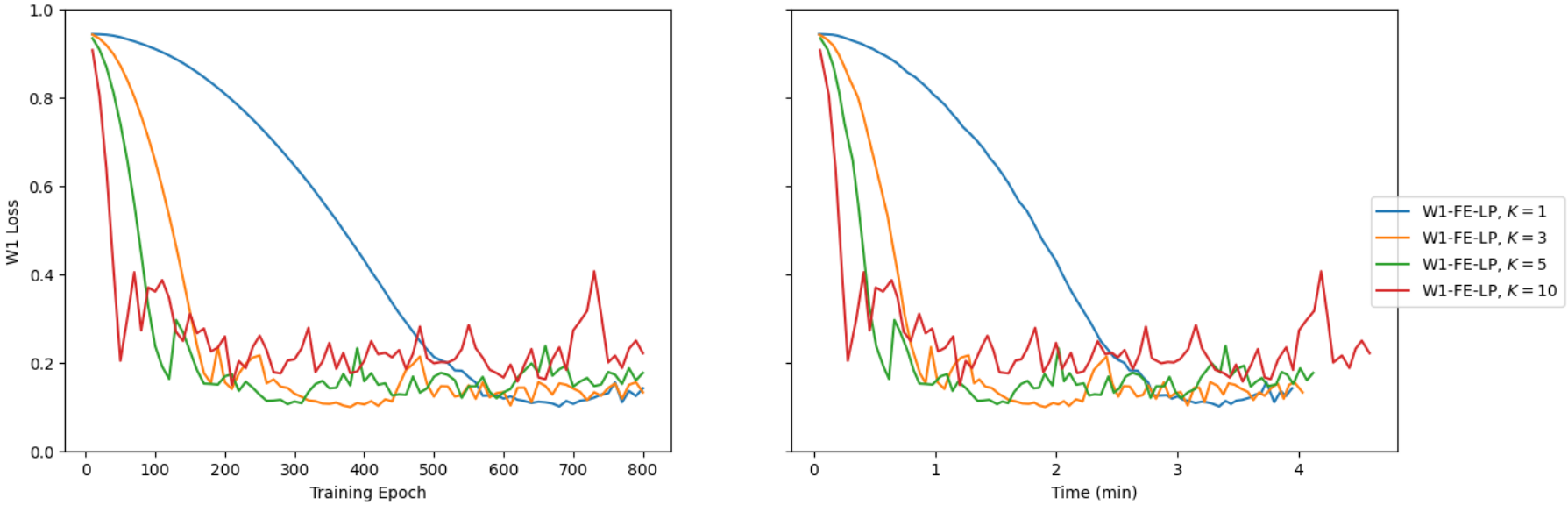
- *Tempting* to solve (22) through

$$dY_t = -\nabla \phi^{\mu_d}_{\mu^{Y_t}}(Y_t)dt, \quad \mu^{Y_0} = \mu_0 \in \mathcal{P}_1^r(\mathbb{R}^d).$$

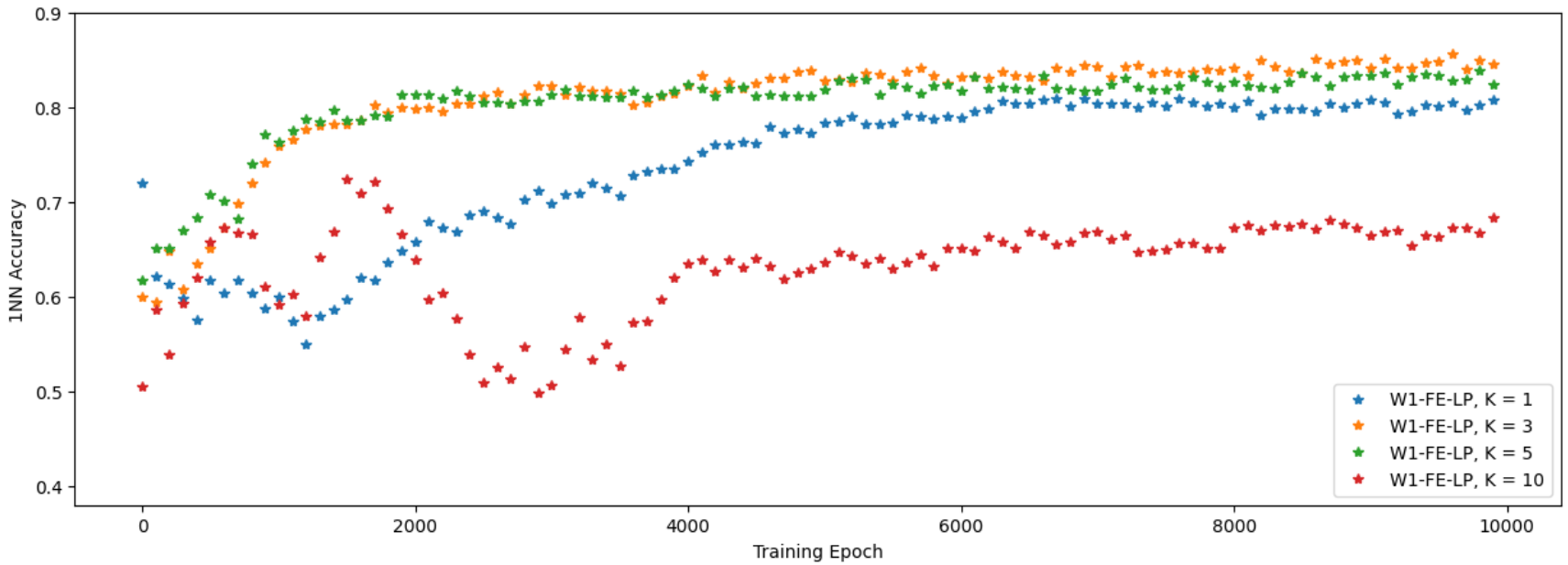
- **Challenges:**
  1) "Gradient of $G : \mathcal{P}_1(\mathbb{R}^d) \to \mathbb{R}$" is *not* well-defined.
     - Subdifferential calculus in AMBROSIO ET AL. (2008) excludes $\mathcal{P}_1(\mathbb{R}^d)$!
     - MALIK & HUANG (2024) points to a possible way around....
  2) In $\mathcal{P}_1(\mathbb{R}^d)$, $\mu_0 \ll \mathcal{L}^d \implies \mu^{Y_t} \ll \mathcal{L}^d$.
     - Arguments below (12) no longer holds, by loss of convexity.

- Despite all this, can do numerical experiments first!

EXPERIMENT 1



EXPERIMENT 2



THANK YOU!!

Q & A
Preprint available

@ arXiv: 2406.13619
"Generative Modeling by Minimizing the Wasserstein-2 Loss"

@ arXiv: 2405.16351
"A Differential Equation Approach for Wasserstein GANs and Beyond"