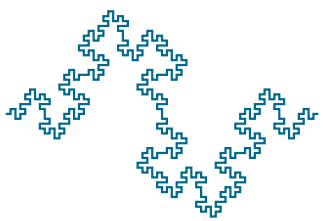
On Characterizing Optimal Wasserstein GAN Solutions for Non-Gaussian Data

Yu-Jui Huang
University of Colorado, Boulder

Joint work with
Shih-Chun Lin (NTU, Taiwan), Yu-Chih Huang (NYCU, Taiwan),
Guan-Huei Lyu (NTU, Taiwan), Hsin-Hua Shen (NTU, Taiwan),
Wan-Yi Lin (Bosch Center for AI, USA)



2023 IEEE ISIT
June 27, 2023

BACKGROUND

Unsupervised Learning:

- $\mathcal{P}(\mathbb{R}^d)$: the set of density functions on $\mathbb{R}^d$.
- $\mu \in \mathcal{P}(\mathbb{R}^d)$: the (unknown) data distribution.
- *Goal:* Find $\rho \in \mathcal{P}(\mathbb{R}^d)$ closest to μ, i.e.,

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

  where $d(\cdot, \cdot)$ is a metric on $\mathcal{P}(\mathbb{R}^d)$.
- **Traditional method:** Parametrize μ.
  - Parameter fitting by maximum likelihood estimations.

BACKGROUND

- **Generative adversarial network (GAN):**

  A min-max game between *generator* & *discriminator*.

  - Proposed by Goodfellow et al. (2014).
  - The generator approximates μ under *Jensen Shannon Divergence* (JSD), i.e.,

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

  - **Advantages:**
    - No constraints on the form of μ.
    - *"Generative"* in nature.
  - **Drawback:** algorithms *don't* converge so easily...
    - Zhu, Jiao, & Tse (2020): The two-player game *doesn't* have a value: min-max game $\neq$ max-min game.

How to fix the instability of GANs?

1) Replace JSD by other metrics on $\mathcal{P}(\mathbb{R}^d)$
   - Arjovsky, Chintala, & Bottou (2017): Wasserstein GAN

2) Replace "*min-max game*" perspective by "*gradient descent*"
   - Huang & Zhang (2023+): The training of GANs is equivalent to gradient descent in $\mathcal{P}(\mathbb{R}^d)$.
     - **Algorithm:** Simulate a distribution-dependent ODE (ordinary differential equation).
     - Many numerical techniques for ODEs can be imported...

3) $\cdots$

$\vdots$

WASSERSTEIN GAN

- Consider (q^{th}-order) Wasserstein distance:

$$\mathcal{W}_q(\mu, \rho) := \inf_{X\sim\mu, Y\sim\rho} \left(\mathbb{E}[\|X - Y\|^q]\right)^{1/q}.$$

- Wasserstein GAN:

$$\min_{\theta} \mathcal{W}_q\left(\mu, \rho^{G_\theta(Z)}\right).$$

  - Z: a simple random variable (e.g., uniform, Gaussian).
  - $G_\theta(\cdot)$: generator neural network (NN), with parameter θ.
  - $\rho^{G_\theta(Z)} \in \mathcal{P}(\mathbb{R}^d)$: density of $G_\theta(Z)$.
  - **Performance:**
    - Easier to converge than vanilla (original) GAN.
    - Can be *computationally costly*...

      Analytical properties of optimal θ would substantially reduce search complexity.

- **Few analytical properties of WGAN were known**
  - Even $\mathcal{W}_q(\cdot,\cdot)$ is typically numerically approximated; Peyre & Cuturi (2019).
- **A special case: the LQG setting**
  - Feizi, Farnia, Ginart, & Tse (2020): WGAN can be solved analytically, with

    Linear generator + **Q**uadratic Wasserstein distance ($q = 2$) + **G**aussian data distribution
  - Reshetova, Bai, Wu, & Ozgur (2021): LQG setting with additional entropic and Sinkhorn regularizers.

Question: Can we analytically solve WGAN beyond LQG case?

OUR CONTRIBUTIONS

- **Analytically solve WGANs with non-Gaussian data:**
  - Allow for *general data distribution*, not limited to Gaussian or uniform as assumed in Bailey & Telgarsky (2018), Feizi, Farnia, Ginart, & Tse (2020), Reshetova, Bai, Wu, & Ozgur (2021).
- **When data is one-dimensional**:
  - some *nonlinear generators* can be considered.
- **Extensions to higher dimensions:**
  - made possible by considering *sliced* WGAN.

Optimal parameters for WGAN with *one-dimensional* data

- Consider a nonlinear generator

$$G_\theta(Z) := \theta_1 + \theta_2 h(Z),$$

  - $\theta = (\theta_1, \theta_2) \in \mathbb{R}^2$.
  - $h : \mathbb{R} \to \mathbb{R}$ can be nonlinear (e.g., sigmoid, ReLU).

- Write $\min_\theta \mathcal{W}_q\left(\mu, \rho^{G_\theta(Z)}\right)$ as

$$\min\left(\min_{\theta_1\in\mathbb{R},\theta_2\geq 0} \mathcal{W}_q\left(\mu, \rho^{G_\theta(Z)}\right), \min_{\theta_1\in\mathbb{R},\theta_2\leq 0} \mathcal{W}_q\left(\mu, \rho^{G_\theta(Z)}\right)\right). \tag{1}$$

  - For *any* μ with continuous CDF, necessary condition for minimizers (i.e., KKT condition) of each sub-problem can be explicitly written down.
  - Relying on 1-D optimal transport techniques.

Theorem (1-D, $q = 2$)

A unique minimizer (θ_1^*, θ_2^*) exists for (1).

- **Case I:** $\mathrm{Cov}\left(X, \Psi^{-1}(F_\mu(X)) + \Psi^{-1}(1 - F_\mu(X))\right) \geq 0$:

$$\begin{aligned} \theta_2^* &= \frac{\mathrm{Cov}\left(X, \Psi^{-1}(F_\mu(X))\right)}{\mathrm{Var}(h(Z))} \geq 0, \\ \theta_1^* &= \mathbb{E}_\mu[X] - \theta_2^* \mathbb{E}\left[h(Z)\right]; \end{aligned} \tag{2}$$

- **Case II:** $\mathrm{Cov}\left(X, \Psi^{-1}(F_\mu(X)) + \Psi^{-1}(1 - F_\mu(X))\right) < 0$:

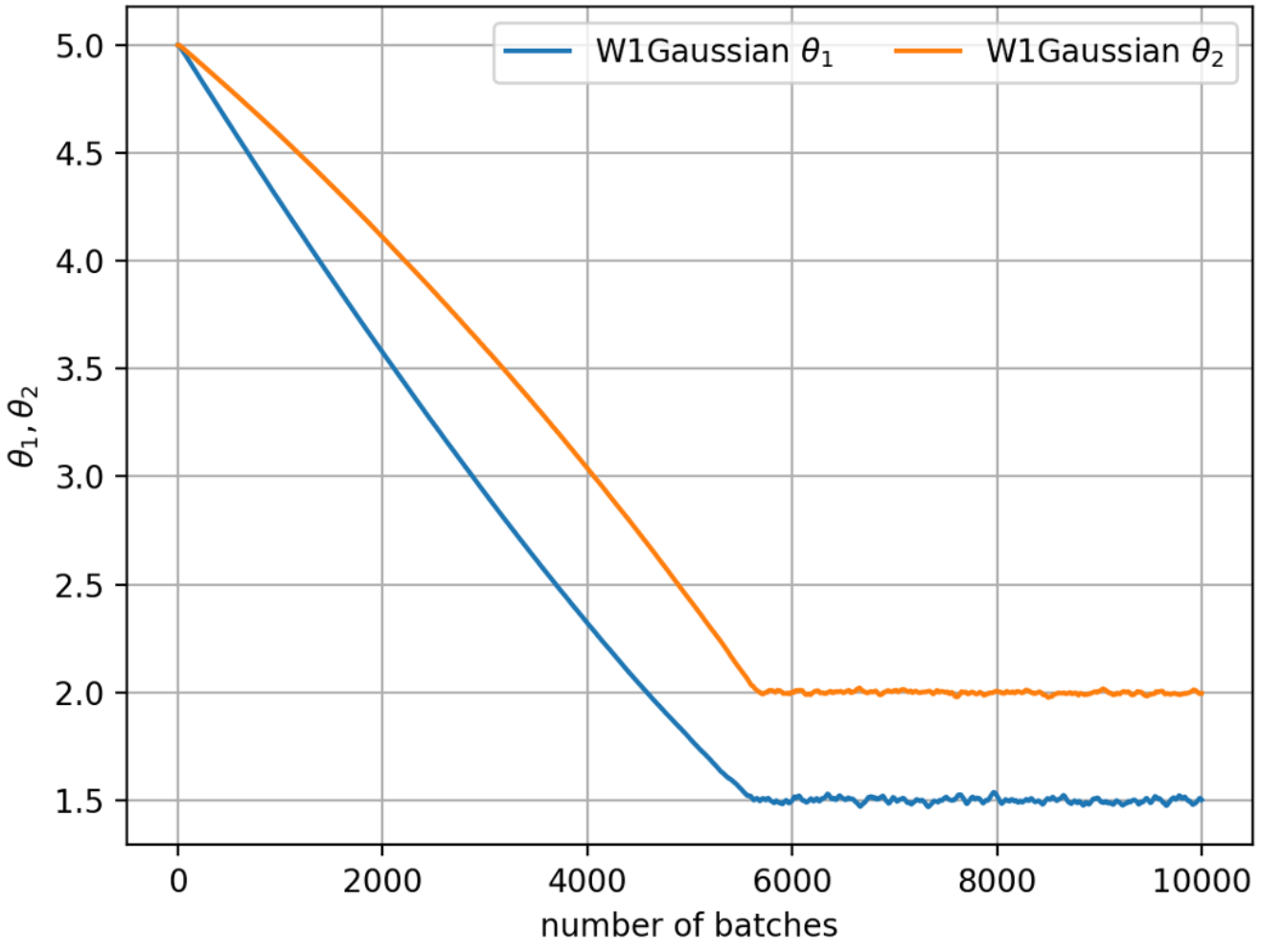
$$\begin{aligned} \theta_2^* &= \frac{\mathrm{Cov}\left(X, \Psi^{-1}(1 - F_\mu(X))\right)}{\mathrm{Var}(h(Z))} \leq 0, \\ \theta_1^* &= \mathbb{E}_\mu[X] - \theta_2^* \mathbb{E}\left[h(Z)\right]; \end{aligned} \tag{3}$$

- Ψ denotes CDF of $h(Z)$.

- For 1-D, $q = 1$,
  - necessary condition for minimizers of WGAN is

$$\begin{aligned} &\mathbb{E}_\mu\Big[\operatorname{sign}\Big(\theta_1^* + \theta_2^* \Psi^{-1}(F_\mu(X)) - X\Big)\Big] = 0, \\ &\mathbb{E}_\mu\Big[\operatorname{sign}\Big(\theta_1^* + \theta_2^* \Psi^{-1}(F_\mu(X)) - X\Big)\Psi^{-1}(F_\mu(X))\Big] = 0. \end{aligned} \tag{4}$$

  - **Numerical example:**
    - Synthetic data $X \sim N(\mu = 1.5, \sigma^2 = 4)$.
    - Linear generator, i.e., $h(Z) = Z$.
    - Use kernel density estimation for $F_\mu(\cdot)$.
    - Estimate (θ_1^*, θ_2^*) from (4), via stochastic gradient decent with momentum.
    - **Result:** (θ_1, θ_2) converges to $(1.4957, 2.0905)$, close to true minimizer $(\mu, \sigma) = (1.5, 2)$.


W1Gaussian θ_1
W1Gaussian θ_2
5.0
4.5
4.0
3.5
3.0
2.5
2.0
1.5
θ_1, θ_2
0
2000
4000
6000
8000
10000
number of batches

Extensions to *high dimensions*

- **Sliced** Wasserstein distance:

$$\mathcal{SW}_q(\mu,\rho) := \left(\int_\Omega \inf_{X\sim\mu, Y\sim\rho} \mathbb{E}[|\omega^T X - \omega^T Y|^q] d\omega\right)^{\frac{1}{q}}$$

  - $\Omega := \{\omega \in \mathbb{R}^d : \|\omega\| = 1\}$ (unit sphere in $\mathbb{R}^d$).
  - $\mathcal{SW}_q$ equivalent to $\mathcal{W}_q$.

- Sliced Wasserstein GAN:

$$\min_\theta \left(\int_\Omega \inf_{X\sim\mu, Z\sim\rho^Z} \mathbb{E}[|\omega^T X - \omega^T G_\theta(Z)|^q] d\omega\right)^{\frac{1}{q}}.$$

  - Deshpande, Zhang, & Schwing (2018), Kolouri et al. (2019).

GAUSSIAN RANDOM PROJECTION

- **Gaussian** sliced Wasserstein distance:

$$\widehat{\mathcal{SW}}_q(\mu,\nu) := \left(\int_{\mathbb{R}^d} \inf_{X\sim\mu, Y\sim\nu} \mathbb{E}[|\omega^T X - \omega^T Y|^q] d\gamma(\omega)\right)^{1/q}.$$

  - $\omega \sim \gamma = N(0, (1/d)\mathbf{I}_d)$.
  - $\widehat{\mathcal{SW}}_2$ equivalent to $\mathcal{SW}_2$, thus to $\mathcal{W}_2$ (Nadjahi et al. (2021)).

- Gaussian Sliced Wasserstein GAN:

$$\min_\theta \left(\int_{\mathbb{R}^d} \inf_{X\sim\mu, Z\sim\rho^Z} \mathbb{E}[|\omega^T X - \omega^T G_\theta(Z)|^2] d\gamma(\omega)\right)^{1/2}. \quad (5)$$

  - **Benefit:** $\omega^T X$ is approximately Gaussian, for any data distribution μ; Reeves (2017).

- Approximate (5) by

$$\min_{\theta} \left(\int_{\mathbb{R}^d} \inf_{Y \sim N(0,\sigma^2), Z \sim \rho^Z} \mathbb{E}[|Y - \omega^T G_\theta(Z)|^2] d\gamma(\omega) \right)^{1/2}. \quad (6)$$

  - $Y \sim N(0, \sigma^2)$ indep. of ω, with $\sigma^2 = \mathbb{E}_\mu[\|X\|^2]/d$.
  - **Idea:** With Gaussian data Y and $q = 2$, LQG result in Feizi et al. (2020) suggests that it is enough to take

$$G_\theta(Z) = \Theta Z, \quad \forall Z \in \mathbb{R}^r,$$

    for some $\Theta \in \mathbb{R}^{d \times r}$.

- Ultimately, we solve

$$\boxed{\min_{\Theta} \left(\int_{\mathbb{R}^d} \inf_{Y \sim N(0,\sigma^2), Z \sim \rho^Z} \mathbb{E}[|Y - \omega^T \Theta Z|^2] d\gamma(\omega) \right)^{1/2}.}$$

With data points in $\mathbb{R}^d$ $(d > 1)$ and $q = 2$,

$$\begin{aligned}
& \textbf{WGAN} \\
\iff & \textbf{Sliced WGAN} \\
\iff & \textbf{Gaussian Sliced WGAN} \\
\approx & \boxed{\min_{\Theta} \left(\int_{\mathbb{R}^d} \inf_{Y \sim N(0,\sigma^2), Z \sim \rho^Z} \mathbb{E}[|Y - \omega^T \Theta Z|^2] d\gamma(\omega) \right)^{1/2}.}
\end{aligned}$$

Theorem

The minimizer $\Theta^* \in \mathbb{R}^{d\times r}$ *of*

$$\min_{\Theta}\left(\int_{\mathbb{R}^d}\inf_{Y\sim N(0,\sigma^2),Z\sim\rho^Z}\mathbb{E}[|Y-\omega^T\Theta Z|^2]d\gamma(\omega)\right)^{1/2}$$

is the minimizer of

$$\frac{\mathrm{Tr}(\Theta\Theta^T)}{d}+\frac{2\sigma}{\Gamma(1/2)}\int_0^\infty z^{-1/2}\frac{\partial}{\partial z}\left|\mathbf{I}+\frac{2z}{d}\Theta\Theta^T\right|^{-1/2}dz.$$

- The derivative can be computed by matrix calculus:

$$\frac{\partial}{\partial z}\left|\mathbf{I}+\frac{2z}{d}\Theta\Theta^T\right|^{-1/2}=-\frac{1}{2}\left|\mathbf{I}+\frac{2z}{d}\Theta\Theta^T\right|^{-3/2}\mathrm{Tr}\left[\mathrm{adj}(\frac{2z}{d}\Theta\Theta^T)\frac{2}{d}\Theta\Theta^T\right].$$

Theorem (Continued)

The gap between

$$\boxed{\min_{\Theta}\left(\int_{\mathbb{R}^d}\inf_{Y\sim N(0,\sigma^2),Z\sim\rho^Z}\mathbb{E}[|Y-\omega^T\Theta Z|^2]d\gamma(\omega)\right)^{1/2}} \tag{7}$$

and

$$\min_{\Theta}\left(\int_{\mathbb{R}^d}\inf_{X\sim\mu,Z\sim\rho^Z}\mathbb{E}[|\omega^T X-\omega^T\Theta Z|^2]d\gamma(\omega)\right)^{1/2}$$

is bounded by a function $f^{\mu}(d)=O(d^{-1/8})$.

- The gap between (7) & **Gaussian Sliced WGAN** (5) =??

Conclusions

CONCLUSIONS

- **We analytically solved WGANs beyond LQG setting**
  - Particularly, for *non-Gaussian data*.
- **For one-dimensional data:**
  - $q = 2$ (quadratic case):
    Closed-form formula of optimal parameters, for *nonlinear generator* and *non-Gaussian data*.
  - $q = 1$:
    Numerical algorithm converges to optimal parameters, for *linear generator* and *Gaussian data*.
- **For high-dimensional data:**
  - $q = 1, 2$: Good approximate of WGAN that can be solved more explicitly.

THANK YOU!!

Q & A

E-mail: sclin2@ntu.edu.tw