Understanding GANs from MFGs and SDEs Approximations

Xin Guo

University of California, Berkeley

BFS One World Seminar
July 2nd, 2020

Based on the following works

- H. Y. Cao, X. Guo, and M. Lauriére (2020). Connecting MFGs and GANs. Arxiv 2002.04112.
- H. Y. Cao and X. Guo (2020). Approximation and convergence of GANs training: an SDE approach. Arxiv 2006.02047.

Roadmap

Which one is real face?

https://research.nvidia.com/sites/default/files/pubs/2017-10_Progressive-Growing-of/karras2018iclr-paper.pdf

GANs (Goodfellow et. al. (2014))

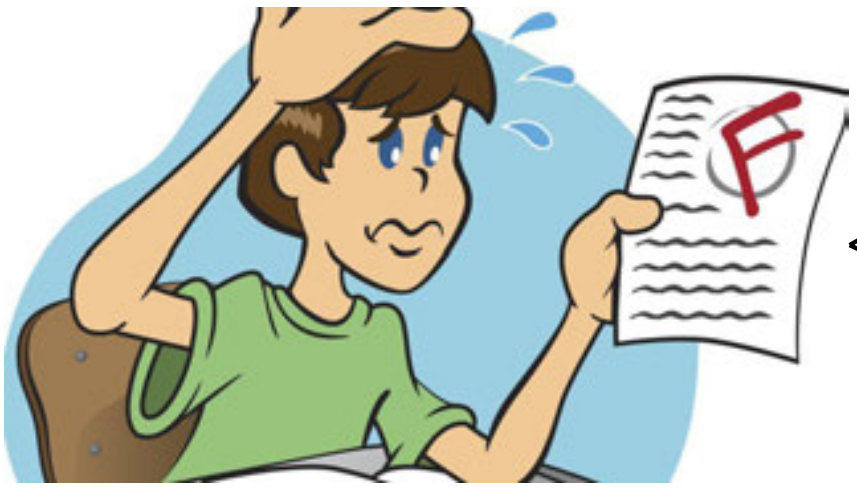
GANs, as generative models, via the game of two neural networks

- A generator network G
- A discriminator network D

Generative Model

Generator G produces (fake) samples $\sim \mathbb{P}_G$

- The true distribution $\mathbb{P}_r$ for the sample data is "inaccessible"
- Generator mimics the sample generated from $\mathbb{P}_r$



"inaccessible"

generate

Generator Network

- Takes a random variable Z with a fixed $\mathbb{P}_z$, and maps it through a parametric function G
- $\mathbb{P}_G$ is the probability distribution of $G(Z)$
- Optimizes G so that $\mathbb{P}_G$ can best resemble $\mathbb{P}_r$
- G is implemented through an NN

How to Make Generative Model Better?

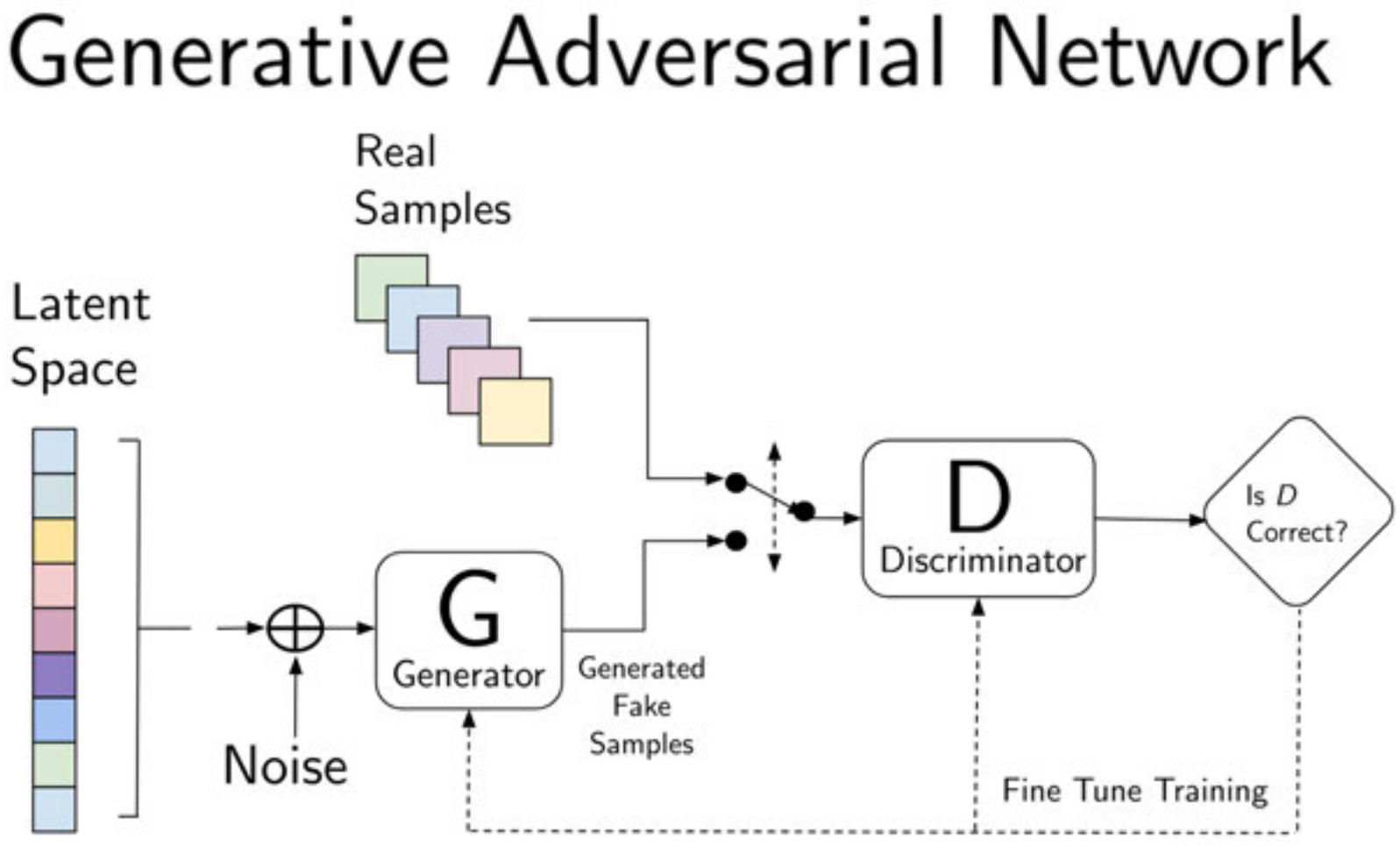
Discriminator, as a knowledgeable mentor

- checks via another NN whether the samples are fake or real
- assigns a score between 0 (fake) and 1 (real)

pass!

fail again!

GANs structure



GANs are popular in ML

- high resolution image generation
- image inpainting
- visual manipulation
- text-to-image synthesis
- video generation
- style transfer

GANs attract attention in MF

- Deep Learning for asset pricing (Chen, Pelger, Zhou (2019))
- Simulation of financial time-series data
  (Wiese, Bai, Wood, Buehler (2019))
  (Wiese, Knobloch, Korn, and Kretschmer (2019))
  (Takahashi, Chen and, Tanaka-Ishii (2019))
  (Zhang et. al. (2019))

GANs as minimax games

- GANs as minimax games between G and D

$$\min_G \max_D \{\mathbb{E}_{X\sim\mathbb{P}_r}[\log D(X)] + \mathbb{E}_{Z\sim\mathbb{P}_z}[\log(1 - D(G(Z)))]\}$$

- Fix G and optimize for D, then the optimal discriminator is

$$D^*_G(x) = \frac{p_r(x)}{p_r(x) + p_g(x)}$$

with p_r and p_g the density functions of $\mathbb{P}_r$ and $\mathbb{P}_G$ respectively

- Therefore, the min-max game becomes

$$\min_G \left\{ \mathbb{E}_{X\sim\mathbb{P}_r}\left[\log \frac{p_r(X)}{p_r(X) + p_g(X)}\right] + \mathbb{E}_{X\sim\mathbb{P}_G}\left[\log \frac{p_g(X)}{p_r(X) + p_g(X)}\right]\right\}$$
$$= -\log 4 + 2JS(\mathbb{P}_r, \mathbb{P}_G),$$

$JS(\cdot,\cdot)$ denoting the Jensen-Shannon divergence

GANs and divergence

- f-GANs: f-divergence (Nock et. al. (2017))
- LSGANs: Least square loss (Mao et. al (2017))
- DRAGANs: Regret minimization (Kodali et. al. (2017))
- CGANs: Conditional extension (Mirza and Osindero (2014))
- WGANs: Wasserstein-1 distance
  (Arjovsky, Chintala, and Bottou (2017)),
  (Gulrajani et. al. (2017))
- RWGANs: Relaxed Wasserstein divergence
  (G., Hong, Lin, Yang (2017))
- GANs with scaled Bregman:
  (Srivastava, Greenewald, and Mirzazadeh (2019))

MFGs

Mean-field games (MFGs) are

- Originated from physics on weakly interacting particles
- Theoretical works pioneered by Lasry and Lions (2007) and Huang, Malhamé, and Caines (2006)
- Stochastic games with very large population of small interacting individuals
- About small interacting individuals, with each player choosing optimal strategy in view of the macroscopic information (mean field)

PDE/Control Approach of General MFGs

MFGs can be analyzed via

- the backward HJB equation for the value function of the underlying control problem
- the forward Fokker-Planck (FP) equation for the controlled dynamics

Connecting MFGs with GANs [Cao, G., and Laurière, 2019]

Idea originated from

- Minimax structure of GANs
- Minimax representation for a class of MFGs (Cirant and Nurbekyan (2018))

MFG as GAN

	GANs	**MFGs**
Generator G	NN for $G : \mathcal{Z} \mapsto \mathcal{X}$	NN for solving HJB
Characterization of $\mathbb{P}_r$	Sample data	FP equation for consistency
Discriminator D	NN for divergence between $\mathbb{P}_G$ and $\mathbb{P}_r$	NN for measuring differential residual from the FP equation

Numerical example

Consider an ergodic MFG on one-dimensional torus $\mathbb{T} = [0, 1]$ characterized by the following system of the PDE's [Almulla, Ferreira and Gomes (2015)]

$$\begin{cases} \dfrac{[\partial_x u(x)]^2}{2} + \sin(2\pi x) = \ln(m(x)) + \bar{H}, \\ -\partial_x [m(x) \cdot \partial_x u(x)] = 0, \end{cases}$$

with

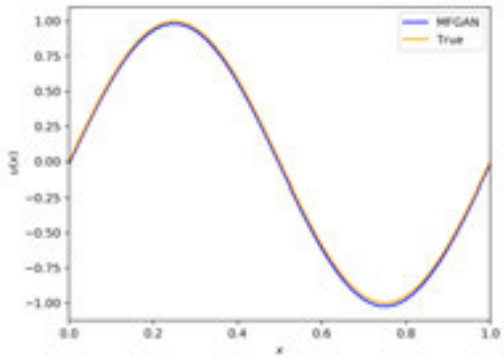
$$\int_{\mathbb{T}} u(x)dx = 0,\ m > 0,\ \int_{\mathbb{T}} m(x)dx = 1.$$

Here, the unknowns are the periodic value function $u(x)$, the periodic density function $m(x)$, and a real number $\bar{H}$.

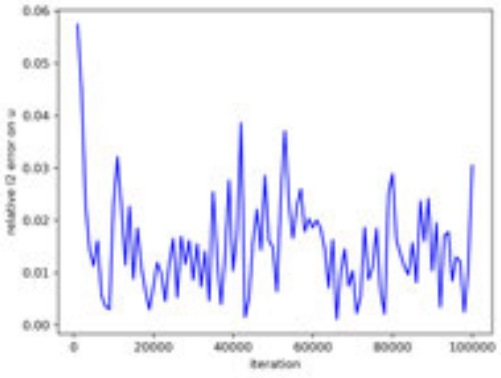
Implementation

- Neural network architecture as in the Deep Galerkin Method (Sirignano and Spiliopoulos (2018))
- $u = u_\theta$, $m = m_\omega = \frac{\exp\{f_\omega\}}{\hat{Z}}$ (Finn, Levine and Abbeel (2016))
- Generator loss $L_G = L_{\text{HJB}} + \beta L_{\text{zero u integration}}$
- Discriminator loss $L_D = L_{\text{FP}}$

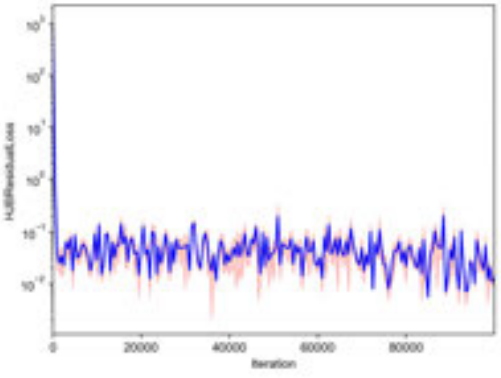
Input of $d = 1$



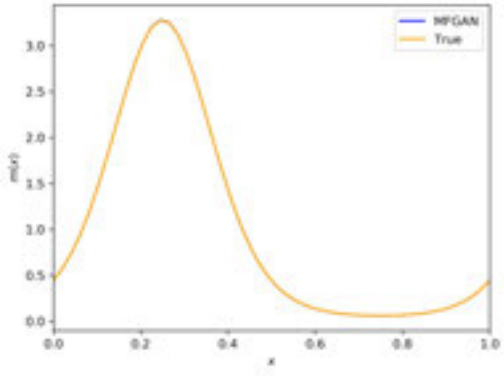
(a) Value function u.



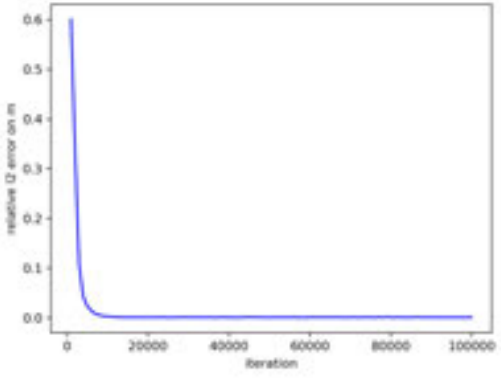
(b) Relative l_2 error u.



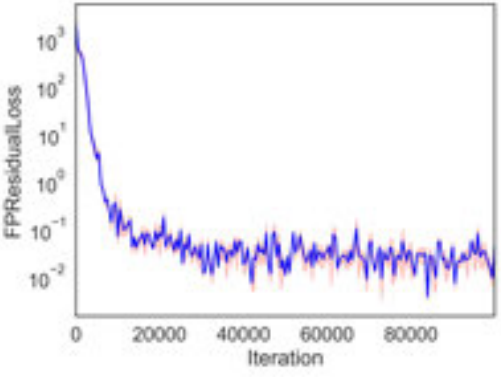
(c) HJB residual loss.



(d) Density function m.



(e) Relative l_2 error m.

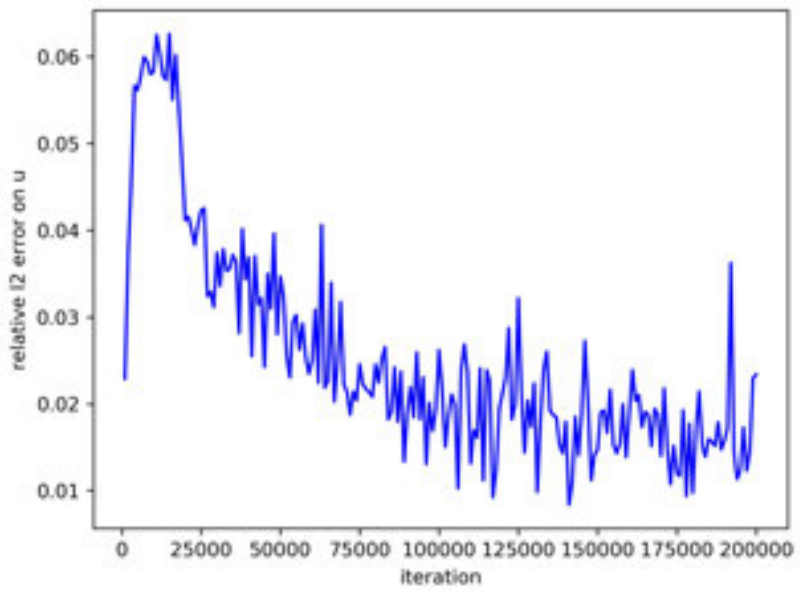


(f) FP residual loss.

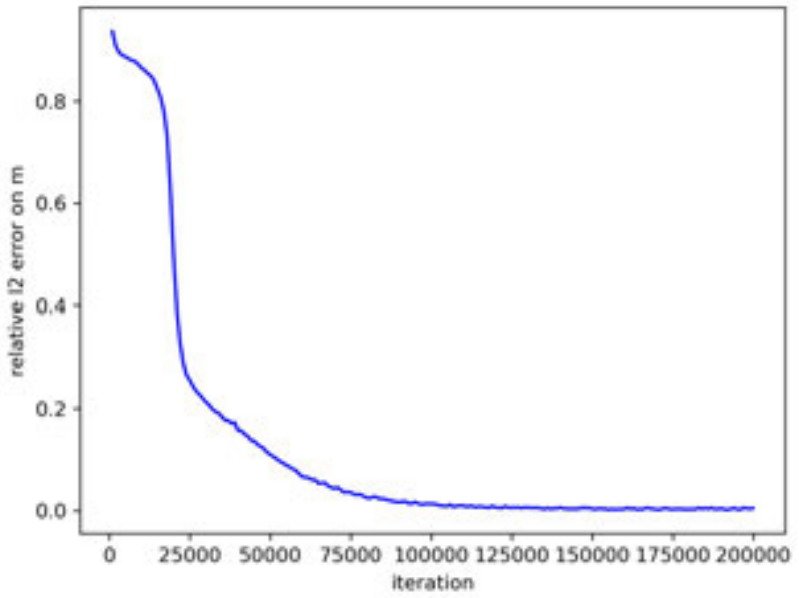
Ablation study

- Relative l_2 errors (for both u and m) and residual loss (for both HJB and FP) decrease as learning rate decreases, with less oscillation
- Relative l_2 errors (for both u and m) and residual loss (for both HJB and FP) decrease as minibatch size increases, with less oscillation.

Input of $d = 4$



(g) Relative l_2 error u.



(h) Relative l_2 error m

Total number of iterations is 2×10^5.
Similar experiment in (Carmona and Laurière (2019)) needed 10^6 iterations for same level of accuracy .

Remark

GANs as a computational tool can be applied beyond MFGs, as long as there is appropriate variational structure.
Example: FBSDE
(Cao, G, Han, Lauriére (2020))

GANs as MFGs

Theorem [Cao, G. and Laurière, 2019]

The GAN in Goodfellow et. al. (2014) is an MFG under Pareto Optimality criterion.

GANs as MFGs

Generators G

- N indistinguishable players, initial belief distributed as $\mathbb{P}_z$ across the population
- As $N \to \infty$, $\frac{1}{N}\sum_{k=1}^{N} \delta_{G(Z_k)} \Rightarrow \mathbb{P}_G$, the mean-field information

Discriminator D

- Players collaborate to fool the best discriminator among $\mathcal{D}$

$$\max_{D\in\mathcal{D}} \frac{1}{N}\frac{1}{M}\sum_{k=1}^{N}\sum_{j=1}^{M} \log\left[D(X_j)\left(1-D(G(Z_k))\right)\right]$$

- As $M, N \to \infty$, this becomes

$$\max_{D\in\mathcal{D}} \mathbb{E}_{\mathbb{P}}\left[\log D(X)\right] + \mathbb{E}_{\mathbb{P}_z}\left[\log\left(1-D(G(Z))\right)\right]$$

WGANs and OT

Proposition [Cao, G. and Laurière, 2019]

For a given G, WGAN is an optimal transport problem.

Remark: Earlier geometric view of connecting GANs and optimal transport in (Lei et. al. (2017)).

WGANs and Optimal Transport (OT)

- WGANs as a minmax game of

$$\min_G \max_D \mathbb{E}_{X\sim\mathbb{P}_r}[\log D(X)] - \mathbb{E}_{Z\sim\mathbb{P}_z}[\log D(G(Z))]$$

- If $f = \log \circ D$, assume f to be 1-Lipschitz, by Kantorovich-Rubinstein duality,

$$\sup_{f \text{ s.t. } \|f\|_L \le 1} \mathbb{E}_{X\sim\mathbb{P}_r}[f(X)] - \mathbb{E}_{Z\sim\mathbb{P}_z}[f(G(Z))] = W_1(\mathbb{P}_r, \mathbb{P}_G)$$

$$:= \inf_{\gamma\in\Pi(\mathbb{P}_r,\mathbb{P}_G)} \int_{\Omega\times\Omega} |x-y|\gamma(dx,dy),$$

with $\Pi(\mathbb{P}_r, \mathbb{P}_G)$ the collection of couplings of $\mathbb{P}_r$ and $\mathbb{P}_G$

WGANs as OT

Under any fixed generator G, define

- cost function as $c(x, y) = |x - y|$ on $\mathcal{X} \times \mathcal{X}$, $\mathcal{X}$ the sample space
- set of all possible couplings Π_G between $Law(G(Z)) = \mathbb{P}_G$ and $\mathbb{P}_r$, where $Z \sim \mathbb{P}_z$

WGANs as OT

- Discriminator is to locate the best coupling among Π_G under a given G and Π_G
- Generator is to refine the set of possible couplings Π_G so that the infimum becomes 0 eventually

GANs and OT

- GANs minimize some proper divergence W between $\mathbb{P}_G$ and $\mathbb{P}_r$, where
$$W : \mathcal{P}(\mathcal{X}) \times \mathcal{P}(\mathcal{X}) \mapsto \mathbb{R}^+.$$
- $W = W_c$ as the optimal cost of an OT problem with an appropriate cost function c
- OT problem has a dual presentation
- GANs consists of two sub-problems:
  1. Fixed generator G, under set of possible couplings Π_G, solving the dual problem of OT
  2. Discriminator, as the price function, minimizing over Π_G through G

Summary

- MFGs as GANs: leverage the empirical success of GANs and design algorithms to solve MFGs and problems with variational/minimax structures
- GANs as MFGs/OT: build the mathematical foundation for GANs

However ...

GANs have many issues...

- Vanishing gradient: imbalance between G and D training
  — gradient vector field of loss function by Berard (2020)
- Challenges of GANs convergence
  — regularization by Mescheder, Geiger, and Nowozin (2018)
- Mode collapsing/gradient exploding
- ....

Recall GANs

- Zero-sum two-player games between the generator network G_θ and the discriminator network D_ω
- Training over a dataset $\mathcal{D} = \{(z_i, x_j)\}_{1 \le i \le N, 1 \le j \le M}$, with $\{z_i\}_{i=1}^N \sim \mathbb{P}_G$ and $\{x_j\}_{j=1}^M \sim \mathbb{P}_r$
- the minimax problem

$$\min_{\theta \in \mathbb{R}^{d_\theta}} \max_{\omega \in \mathbb{R}^{d_\omega}} \Phi(\theta, \omega),$$

with

$$\Phi(\theta, \omega) = \frac{\sum_{i=1}^N \sum_{j=1}^M F(D_\omega(x_j), D_\omega(G_\theta(z_i)))}{N \cdot M}$$

Stochastic Gradient Algorithms (SGAs)

- Ascent of ω along g_ω

$$\omega_{t+1} = \omega_t + \eta g_\omega^{\mathcal{B}}(\theta_t, \omega_t)$$

- Descent of θ along g_θ

$$\theta_{t+1} = \theta_t - \eta g_\theta^{\mathcal{B}}(\theta_t, \omega_{t+1})$$

with

- $g_\theta = \nabla_\theta \Phi = \frac{\sum_i \sum_j g_\theta^{i,j}}{N \cdot M}$
  $g_\omega = \nabla_\omega \Phi = \frac{\sum_i \sum_j g_\omega^{i,j}}{N \cdot M}$, with
  $g_\theta^{i,j}(\theta, \omega) = \nabla_\theta F(D_\omega(x_j), D_\omega(G_\theta(z_i)))$
  $g_\omega^{i,j}(\theta, \omega) = \nabla_\omega F(D_\omega(x_j), D_\omega(G_\theta(z_i)))$
- $g_\theta^{\mathcal{B}} = \frac{\sum_k g_\theta^{I_k, J_k}}{B}$ $g_\omega^{\mathcal{B}} = \frac{\sum_k g_\omega^{I_k, J_k}}{B}$
  estimated from i.i.d samples of batch size B
- Learning rate η

Heuristics

Suppose $\eta \sim \Delta t$, as the batch size gets sufficiently large, by CLT

$$\omega_{t+1} = \omega_t + \eta g_\omega^{\mathcal{B}}(\theta_t, \omega_t) \approx \omega_t + \eta g_\omega(\theta_t, \omega_t) + \frac{\eta}{\sqrt{B}} \Sigma_\omega^{\frac{1}{2}}(\theta_t, \omega_t) Z_t^1,$$

$$\theta_{t+1} = \theta_t - \eta g_\theta^{\mathcal{B}}(\theta_t, \omega_{t+1}) \approx \theta_t - \eta g_\theta(\theta_t, \omega_{t+1}) + \frac{\eta}{\sqrt{B}} \Sigma_\theta^{\frac{1}{2}}(\theta_t, \omega_{t+1}) Z_t^2,$$

with $Z_t^i \overset{i.i.d.}{\sim} N(0, I)$ for $i = 1, 2$ and any iteration t.
Question: is the following SDE a proper approximation,

$$d \begin{pmatrix} \Theta_t \\ \mathcal{W}_t \end{pmatrix} = \begin{pmatrix} -g_\theta(\Theta_t, \mathcal{W}_t) \\ g_\omega(\Theta_t, \mathcal{W}_t) \end{pmatrix} dt + \sqrt{2\beta^{-1} \begin{pmatrix} \Sigma_\theta(\Theta_t, \mathcal{W}_t) & 0 \\ 0 & \Sigma_\omega(\Theta_t, \mathcal{W}_t) \end{pmatrix}} dW_t?$$

with $\beta = \frac{2B}{\eta}$.

The SDE

- The correct SDE takes the form of

$$d\begin{pmatrix}\Theta_t\\ \mathcal{W}_t\end{pmatrix} = \left[\begin{pmatrix}-g_\theta(\Theta_t,\mathcal{W}_t)\\ g_\omega(\Theta_t,\mathcal{W}_t)\end{pmatrix} + \frac{\eta}{2}\begin{pmatrix}\nabla_\theta g_\theta(\Theta_t,\mathcal{W}_t) & -\nabla_\omega g_\theta(\Theta_t,\mathcal{W}_t)\\ -\nabla_\theta g_\omega(\Theta_t,\mathcal{W}_t) & -\nabla_\omega g_\omega(\Theta_t,\mathcal{W}_t)\end{pmatrix}\begin{pmatrix}-g_\theta(\Theta_t,\mathcal{W}_t)\\ g_\omega(\Theta_t,\mathcal{W}_t)\end{pmatrix}\right]dt + \sqrt{2\beta^{-1}}\begin{pmatrix}\Sigma_\theta(\Theta_t,\mathcal{W}_t)^{\frac{1}{2}} & 0\\ 0 & \Sigma_\omega(\Theta_t,\mathcal{W}_t)^{\frac{1}{2}}\end{pmatrix}dW_t$$

- The extra term in the drift highlights the interaction of the generator and the discriminator

Approximation and Error Bound

Theorem [Cao and G., 2020]

Fix an arbitrary time horizon $\mathcal{T} > 0$ and take the learning rate $\eta \in (0, 1 \wedge \mathcal{T})$ and the number of iterations $N = \left\lfloor \frac{\mathcal{T}}{\eta} \right\rfloor$. Assume appropriate regularity conditions for g, Φ. Then, given any initialization $\theta_{\mathbf{0}} = \theta$ and $\omega_{\mathbf{0}} = \omega$, for any sufficiently smooth test function f, the following weak approximation holds:

$$\max_{t=\mathbf{1},\ldots,N} \left| \mathbb{E} f(\theta_t, \omega_t) - \mathbb{E} f(\Theta_{t\eta}, \mathcal{W}_{t\eta}) \right| \leq C\eta^{\mathbf{2}}$$

for constant $C \geq 0$, where (θ_t, ω_t) and $(\Theta_{t\eta}, \mathcal{W}_{t\eta})$ are given by the update scheme and its SDE approximation, respectively.

SDE approximation for GANs

SDEs approximation of GANs enables analyzing the evolution of GANs parameters,

- The extra term in the drift highlights the interaction of the generator and the discriminator
- The form of SDE prescribes the ratio between the batch size and the learning rate in order to modulate the fluctuations of SGAs in GANs training
- Regularity conditions for the coefficients of the SDE guides the training of GANs away from the explosive gradients, and confirms mathematically the importance of appropriate choices of network depth and introduction of gradient clipping and gradient penalty.

Analysis of SDEs via invariant measures and by Itô's formula for various functional of the parameters

- Conditions for existence of invariant measures suggest the conditions for convergence of GANs training
- The dynamics of training loss and its subsequent fluctuation-dissipation-relation reveals the trade-off of the loss landscape between the generator and the discriminator

Questions?
Thank you!