Generative Adversarial Networks: Game and Control Perspectives

Xin Guo

University of California at Berkeley

International Seminar on SDEs and Related Topics
Feb 25, 2022
Based on joint work with Haoyang Cao of Ecole Polytechnique and Othmane Mounjid of Amazon

Roadmap

GANs (Goodfellow et. al. (2014))

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

- Generator network G
- Discriminator network D

A generator network G

- 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 the true distribution $\mathbb{P}_r$
- G is implemented through an NN

A discriminator network D

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

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
- Portfolio and risk management
- Simulation of financial time-series data
- Fraud detection
- Computing mean-field games

(Cao & G. (2021) and Eckerli & Osterrieder (2021) for reviews)

GANs have many challenges...

- Vanishing gradient/imbalance between G and D training
  Berard (2020)
- Convergence issue
  Mescheder, Geiger, and Nowozin (2018), Cao and G. (2020)
- Mode collapsing/gradient exploding

GANs and divergence

- 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 minimax 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

Improper divergence function

- **Example:** Given $\theta \in [0,1]$, assume that $\mathbb{P}$ and $\mathbb{Q}$ satisfy

$$\forall (X,Y) \sim \mathbb{P},\ X = 0,\ \ Y \sim \text{Uniform}(0,1),$$
$$\forall (X,Y) \sim \mathbb{Q},\ X = \theta,\ \ Y \sim \text{Uniform}(0,1)$$

- As $\theta \neq 0$,

$$KL(\mathbb{P},\mathbb{Q}) = KL(\mathbb{Q},\mathbb{P}) = +\infty,\ \ JS(\mathbb{P},\mathbb{Q}) = \log(2)$$

- As $\theta = 0$,

$$KL(\mathbb{P},\mathbb{Q}) = KL(\mathbb{Q},\mathbb{P}) = JS(\mathbb{P},\mathbb{Q}) = 0$$

GANs and divergence

- KL is **infinite** when two distributions are disjoint
- JS has sudden jump, **discontinuous** at $\theta = 0$
- W_1 is continuous and relatively smooth
- Wasserstein L^1 divergence outperforms KL and JS divergences

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))

Theoretical Studies of GANs

- Connecting GANs with mean-field games
  Cao, G., and Laurière (2020), Lin, Fung, Li, Nurbekyan, and Osher (2020)
- Connecting GANs with reinforcement learning actor-critic
  Pfau and Vinyals (2016)
- Connecting GANs with optimal transport
  Cao, G., and Laurière (2020), Xu, Wenliang, Munn, Acciaio (2020)

GANs training via SGA

- 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

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

with

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

- Minimax games between the generator network G_θ and the discriminator network D_ω

- The parametrized version of vanilla GANs training is to find

$$\min_{\theta}\max_{\omega}\mathbb{E}_{X\sim\mathbb{P}_r}[\log D_\omega(X)]+\mathbb{E}_{Z\sim\mathbb{P}_z}[\log(1-D_\omega(G_\theta(Z)))]$$

- The parametrized version of general GANs training is to find

$$\min_{\theta}\max_{\omega}\mathbb{E}_{X\sim\mathbb{P}_r}[f_1(D_\omega(X))]+\mathbb{E}_{Z\sim\mathbb{P}_z}[f_2(D_\omega(G_\theta(Z)))]$$

where f_1, f_2 are some quasi-concave functions chosen to address the stability issues of GANs game, including WGANs.

Sion's theorem (1958)

Assuming

- ω and θ chosen from compact and convex sets
- g is upper continuous and quasi-convex in θ and lower continuous and quasi-concave in ω

then minimax problem has no duality gap, i.e.,

$$\min_{\theta} \max_{\omega} g(\omega, \theta) = \max_{\omega} \min_{\theta} g(\omega, \theta)$$

(Generalized minimax theorem of John von Neumann (1959))

Example with convexity/concavity issue

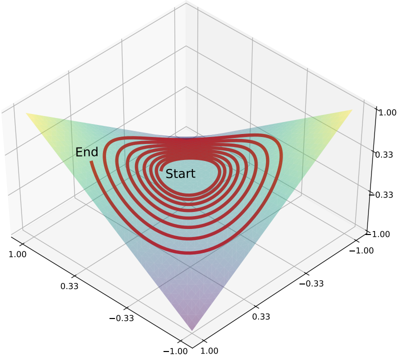

Figure: SGA to solve $\min_y \max_x xy$ with $(0, 0)$ the unique Nash equilibrium

Improper parametrization for GANs training

Alert: Many existing works for GANs lack of proper concavity and convexity properties!

- Take $X \sim N(m, \sigma^2), \quad Z \sim N(0,1)$, with $(m, \sigma) \in \mathbb{R} \times \mathbb{R}_+$.
- Consider the parametrization of the discriminator and the generator networks:

$$\begin{cases} D_w(x) = D_{(w_1,w_2,w_3)}(x) = \dfrac{1}{1 + e^{-(w_3/2 \cdot x^2 + w_2 x + w_1)}}, \\ G_\theta(z) = G_{(\theta_1,\theta_2)}(z) = \theta_2 z + \theta_1, \end{cases}$$

where $w = (w_1, w_2, w_3) \in \mathbb{R}^3$, and $\theta = (\theta_1, \theta_2) \in \mathbb{R} \times \mathbb{R}_+$.

Example of improper learning rate

- Consider

$$f(x) = (a/2)\, x^2 + b\, x, \qquad \forall x \in \mathbb{R},$$

where $(a, b) \in \mathbb{R}_+ \times \mathbb{R}$.

- Finding the minimum $x^* = -(b/a)$ of f via the gradient algorithm goes as follows:

$$x_{n+1} = x_n - \eta(ax_n + b), \qquad \forall n \geq 0,$$

with $x_0 \in \mathbb{R}$ given and η the learning rate.

Consider the error $e_n = |x_n - x^*|^2$:

$$e_{n+1} = |x_{n+1} - x^*|^2 = \big(1 - \eta a(2 - \eta a)\big)|x_n - x^*|^2$$

Thus,

$$e_{n+1} = r\, e_n \underset{n\to\infty}{\longrightarrow} +\infty$$

as $r = \big(1 - \eta a(2 - \eta a)\big) > 1$ when $\eta > 2/a$.

Optimal controls for GANs training

Three key parameters for fine tuning

- Learning rate: on how far to move along the gradient direction
- Batch size: the number of training samples used in the gradient estimation
- Time scale: the number of updates of the variables θ and ω

Remark

- Smaller learning rate and larger minibatch size reduce error and oscillation (Cao, G. and Laurière (2020))
- Optimal control of time scale can be shown to be equivalent to optimal control of learning rate (G. and Mounjid (2021))

Optimal learning rate: mathematical formulation

- Starting from initial guess (w_0, θ_0)
- SGA updating:

$$
\begin{aligned}
w_{t+1} &= w_t + \eta_t^w g_w(w_t, \theta_t), \\
\theta_{t+1} &= \theta_t - \eta_t^\theta g_\theta(w_t, \theta_t)
\end{aligned}
$$

with $g_w = \nabla_w g$, $g_\theta = \nabla_\theta g$, and $(\eta_t^w, \eta_t^\theta) \in \mathbb{R}_+^2$ the learning rate

Coupled SDEs approximation (Cao and G. (2020))

$$\begin{cases} dw(t) & = g_w(q(t))dt + \sqrt{\eta}\sigma_w(q(t))dW^1(t), \\ d\theta(t) & = -g_\theta(q(t))dt + \sqrt{\eta}\sigma_\theta(q(t))dW^2(t) \end{cases}$$

- $q(t) = (w(t), \theta(t))$
- $\sigma_w : \mathbb{R}^M \times \mathbb{R}^N \to \mathcal{M}_{\mathbb{R}}(M)$ and $\sigma_\theta : \mathbb{R}^M \times \mathbb{R}^N \to \mathcal{M}_{\mathbb{R}}(N)$ are approximated by the covariance of g_w and g_θ
- Brownian motions W^1 and W^2 are independent
- Learning rates (η, η) are fixed constants for the generator and the discriminator

Adaptive learning rate process $\eta(t)$

A learning rate $\eta(t)$ at time t

$$\begin{aligned}\eta(t) = \left(\eta^w(t), \eta^\theta(t)\right) &= \left(u^w(t) \times \bar{\eta}^w(t), u^\theta(t) \times \bar{\eta}^\theta(t)\right) \\ &= u(t) \bullet \bar{\eta}(t) \quad \forall t \geq 0\end{aligned}$$

- Predefined base learning rate $\bar{\eta}_t = (\bar{\eta}^w(t), \bar{\eta}^\theta(t))$ fixed by the controller
- An adapted learning rate $u_t = (u^w(t), u^\theta(t))$, adjusted around $\bar{\eta}_t$ and adaptive to the training process
- $u^w(t)$ and $u^\theta(t)$ assumed bounded by a fixed constant $u^{\max}$
- Clipping parameter $u^{\max} \geq 0$ introduced to handle the convexity and explosion issue

State dynamics

With the adaptive learning rate $\eta(t)$, the corresponding SDE for GANs training becomes

$$\begin{cases} dw(t) = u^w(t) g_w(q(t))dt + \left(u^w \sqrt{\bar{\eta}^w}\right)(t) \sigma_w(q(t)) dW^1(t), \\ d\theta(t) = -u^\theta(t) g_\theta(q(t))dt + \left(u^\theta \sqrt{\bar{\eta}^\theta}\right)(t) \sigma_\theta(q(t)) dW^2(t) \end{cases}$$

Control problem

- $T < \infty$ a finite time horizon
- Reward function

$$J(T, t, q; u) = \mathbb{E}[g(q(T))|q(t) = q, u]$$

- $\mathcal{U}^w$ and $\mathcal{U}^\theta$ respective admissible controls set for u^w and u^θ
- Objective

$$v(t, q) = \min_{u^\theta \in \mathcal{U}^\theta} \max_{u^w \in \mathcal{U}^w} J(T, t, q; u)$$

for any $(t, q) \in [0, T] \times \mathbb{R}^M \times \mathbb{R}^N$

Analysis of optimal learning rate control problem

Value function (G. and Mounjid (2021))

Under proper regularity assumptions, the value function v is a solution (classical or viscosity) to the following equation:

$$\begin{cases} v_t + \max\min_{(u^w, u^\theta \in [0, u^{\max}])} \quad \Big\{ \big(u^w g_w^\top v_w - u^\theta g_\theta^\top v_\theta \big) \\ \qquad + \frac{1}{2} \Big[(u^w)^2 (\bar{\Sigma}^w : v_{ww}) + (u^\theta)^2 (\bar{\Sigma}^\theta : v_{\theta\theta}) \Big] \Big\} = 0, \\ v(T, \cdot) = g(\cdot) \end{cases}$$

Optimal learning rate (G. Mounjid (2021))

Under simple regularity conditions, the optimal adaptive learning rate $\bar{u}^w$ and $\bar{u}^\theta$ are given by

$$\bar{u}^w = \begin{cases} 0 \vee \left(\frac{-g_w^\top v_w}{\bar{\Sigma}^w : v_{ww}} \wedge u^{\max}\right), & \text{if } \bar{\Sigma}^w : v_{ww} < 0, \\ u^{\max}, & \text{otherwise} \end{cases}$$

$$\bar{u}^\theta = \begin{cases} 0 \vee \left(\frac{g_\theta^\top v_\theta}{\bar{\Sigma}^\theta : v_{\theta\theta}} \wedge u^{\max}\right), & \text{if } \bar{\Sigma}^\theta : v_{\theta\theta} > 0, \\ u^{\max}, & \text{otherwise} \end{cases}$$

Here the matrices $\bar{\Sigma}^w$ and $\bar{\Sigma}^\theta$ satisfy

$$\begin{cases} \bar{\Sigma}^w(t,q) = \{\bar{\sigma}_t^w(\bar{\sigma}_t^w)^\top\}(q), & \bar{\Sigma}^\theta(t,q) = \{\bar{\sigma}_t^\theta(\bar{\sigma}^\theta)_t^\top\}(q), \\ \bar{\sigma}_t^w(q) = \sqrt{\bar{\eta}^w(t)}\sigma^w(q), & \bar{\sigma}_t^\theta(q) = \sqrt{\bar{\eta}^\theta(t)}\sigma^\theta(q) \end{cases}$$

for any $t \in \mathbb{R}_+$, and $q = (w,\theta) \in \mathbb{R}^M \times \mathbb{R}^N$,
$A : B = \mathrm{Tr}[A^\top B]$ for any real matrices A and B

- On regularity conditions: v belongs to $\mathcal{C}^{1,2}([0,T],\mathbb{R}^M \times \mathbb{R}^N)$. For instance, when g and $\bar{\sigma}$ are Lipschitz continuous.
- The clipping parameter $u^{\max}$ is closely related to the convexity issue discussed for GANs minimax games.
- When the convexity condition $\bar{\Sigma}^w : v_{ww} < 0$ is violated, explosion in GANs training can be prevented by fixing an upper bound $\bar{u}^{\max}$ for the learning rate.
- The control $(\bar{u}^w, \bar{u}^\theta)$ is closely related to the standard Newton algorithm.

Numerical experiment

- Vanila GANs setup
- $X \sim N(3, 1), Z \sim N(0, 1)$
- Discriminator accuracy expected to be 0.5
- Epoch: the number of gradient updates needed to pass the entire training dataset

ADAM with base andadaptive learning rate

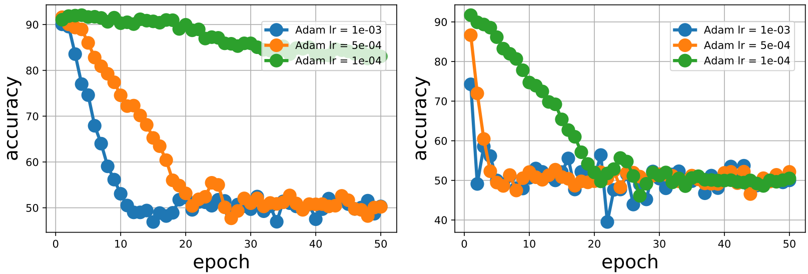

Figure: Left: discriminator accuracy for ADAM with base learning rate; Right: ADAM with an additional adaptive learning rate component

Monge's formulation of optimal transport

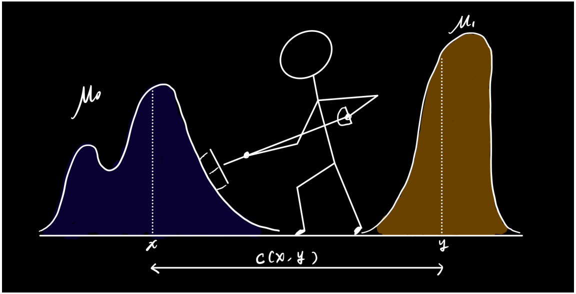

Figure: Earth mover problem

$$\inf_T \left\{ \int_{\mathcal{X}} c(x, T(x))\mu_0(dx) \middle| T\#\mu_0 = \mu_1 \right\}$$

Primal and dual formulation of optimal transport

Kantorovich's (primal) formulation of optimal transport

$$\inf_{\pi \in \Pi(\mu,\nu)} \int_{\mathcal{X}\times\mathcal{Y}} c(x,y)\pi(dx,dy)$$

with $\Pi(\mu,\nu)$ the collection of couplings of μ and ν

Kantorovich-Rubinstein Duality [Villani, 2009]

Under proper conditions on the transport cost c, the primal and dual problems are equivalent.

WGANs and OT

Proposition [Cao, G. and Laurière, 2019]

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

Earlier geometric view of connecting GANs and optimal transport in (Lei, Su, Cui, Yau, and Gu (2017))

- 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

Key idea

- 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,

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

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

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

Remark: this connection can be generalized to any GANs assuming that the corresponding OT problem has a dual presentation.

Discussion

- Optimal controls of parameter fine tuning will improve the performance of GANs: more applications?
- Applying connection between GANs with mean-field games and optimal transport for finance problems beyond financial data simulation: high dimensional MFGs, MFCs, FBSDEs?

This talk is based on

- H. Y. Cao and X. Guo. (2021). GANs, some analytical perspectives. Handbook of Machine Learning and Applications to Mathematical Finance.
- H. Y. Cao, X. Guo, and M. Lauriére (2020). Connecting GANs and MFGs. Under review.
- H. Y. Cao and X. Guo (2020). Approximation and convergence of GANs training: an SDE approach. Under review.
- X. Guo and O. Mounjid (2021). GANs training: a stochastic control and game framework. Under review.

Questions?
Thank you!