Finding Mixed Nash Equilibria of Generative Adversarial Networks

Ya-Ping Hsieh
ya-ping.hsieh@epfl.ch

Laboratory for Information and Inference Systems (LIONS)
École Polytechnique Fédérale de Lausanne (EPFL)
Switzerland

ICML

[June 12, 2019]

Joint work with
Chen Liu and Volkan Cevher @ LIONS

Learning distributions

◦ A balancing act between data, models, and computation

  ▷ upshots: data generation, compression, domain transfer, and recognition

  ▷ trends: from simple parametric models to super expressive neural networks

  ▷ challenges: computational costs as well as the difficulty of training

Learning distributions

◦ A balancing act between data, models, and computation

▷ upshots: data generation, compression, domain transfer, and recognition

▷ trends: from simple parametric models to super expressive neural networks

▷ challenges: computational costs as well as the difficulty of training

◦ Highlight: Generative Adversarial Networks (GANs) [Goodfellow et al., 2014]

▷ train a **generator** neural net, generating "fake" data

▷ train a **discriminator** neural net, authenticating this data based on "real" samples

▷ setup a minimax game between the two

Learning distributions

- A balancing act between data, models, and computation
  - ▷ upshots: data generation, compression, domain transfer, and recognition
  - ▷ trends: from simple parametric models to super expressive neural networks
  - ▷ challenges: computational costs as well as the difficulty of training
- Highlight: Generative Adversarial Networks (GANs) [Goodfellow et al., 2014]
  - ▷ train a **generator** neural net, generating "fake" data
  - ▷ train a **discriminator** neural net, authenticating this data based on "real" samples
  - ▷ setup a minimax game between the two
- Several variants exist [Karras et al., 2017, Brock et al., 2018]
  - ▷ running example: Wasserstein GANs [Arjovsky et al., 2017]

Wasserstein GANs

◦ A natural *pure* strategy-based minimax objective

$$\min_{\theta \in \Theta} \max_{w \in \mathcal{W}} \mathbb{E}_{X \sim P_{\text{real}}} [D_w(X)] - \mathbb{E}_{X \sim P^{\theta}_{\text{fake}}} [D_w(X)] .$$

▷ θ: a **generator** neural net

▷ w: a **discriminator** neural net

▷ D_w: output of discriminator at w, *highly* non-convex/non-concave

Wasserstein GANs

◦ A natural *pure* strategy-based minimax objective

$$\min_{\theta \in \Theta} \max_{w \in \mathcal{W}} \mathbb{E}_{X \sim P_{\text{real}}} \left[D_w(X) \right] - \mathbb{E}_{X \sim P^{\theta}_{\text{fake}}} \left[D_w(X) \right].$$

- ▷ θ: a **generator** neural net
- ▷ w: a **discriminator** neural net
- ▷ D_w: output of discriminator at w, *highly* non-convex/non-concave

◦ Theoretical challenges

- ▷ a saddle point might NOT exist [Dasgupta and Maskin, 1986]
- ▷ no provably convergent algorithm

Wasserstein GANs

- A natural *pure* strategy-based minimax objective

$$\min_{\theta\in\Theta}\max_{w\in\mathcal{W}} \mathbb{E}_{X\sim P_{\text{real}}}\left[D_w(X)\right] - \mathbb{E}_{X\sim P^{\theta}_{\text{fake}}}\left[D_w(X)\right].$$

  - ▷ θ: a **generator** neural net
  - ▷ w: a **discriminator** neural net
  - ▷ D_w: output of discriminator at w, *highly* non-convex/non-concave

- Theoretical challenges
  - ▷ a saddle point might NOT exist [Dasgupta and Maskin, 1986]
  - ▷ no provably convergent algorithm

- Practical challenges
  - ▷ the simple (alternating) SGD does NOT work well in practice...
  - ▷ adaptive methods (Adam, RMSProp,...) highly unstable, heavy tuning...

Wasserstein GANs: From pure to mixed Nash Equilibrium

◦ Objective of Wasserstein GANs is a pure strategy formulation:

$$\min_{\theta\in\Theta}\max_{w\in\mathcal{W}} \mathbb{E}_{X\sim P_{\mathsf{real}}}\left[D_w(X)\right] - \mathbb{E}_{X\sim P_{\mathsf{fake}}}\left[D_w(X)\right].$$

Wasserstein GANs: From pure to mixed Nash Equilibrium

◦ Objective of Wasserstein GANs is a pure strategy formulation:

$$\min_{\theta \in \Theta} \max_{w \in \mathcal{W}} \mathbb{E}_{X \sim P_{\text{real}}} \left[D_w(X) \right] - \mathbb{E}_{X \sim P_{\text{fake}}} \left[D_w(X) \right].$$

◦ A new objective of Wasserstein GANs: Our **mixed** strategy proposal via game theory

$$\min_{\nu \in \mathcal{M}(\Theta)} \max_{\mu \in \mathcal{M}(\mathcal{W})} \mathbb{E}_{w \sim \mu} \mathbb{E}_{X \sim P_{\text{real}}} \left[D_w(X) \right] - \mathbb{E}_{w \sim \mu} \mathbb{E}_{\theta \sim \nu} \mathbb{E}_{X \sim P_{\text{fake}}^{\theta}} \left[D_w(X) \right].$$

where $\mathcal{M}(\mathcal{Z}) := \{$all (regular) probability measures on $\mathcal{Z}\}$.

Wasserstein GANs: From pure to mixed Nash Equilibrium

◦ Objective of Wasserstein GANs is a pure strategy formulation:

$$\min_{\theta \in \Theta} \max_{w \in \mathcal{W}} \mathbb{E}_{X \sim P_{\text{real}}} [D_w(X)] - \mathbb{E}_{X \sim P_{\text{fake}}} [D_w(X)].$$

◦ A new objective of Wasserstein GANs: Our **mixed** strategy proposal via game theory

$$\min_{\nu \in \mathcal{M}(\Theta)} \max_{\mu \in \mathcal{M}(\mathcal{W})} \mathbb{E}_{w \sim \mu} \mathbb{E}_{X \sim P_{\text{real}}} [D_w(X)] - \mathbb{E}_{w \sim \mu} \mathbb{E}_{\theta \sim \nu} \mathbb{E}_{X \sim P^{\theta}_{\text{fake}}} [D_w(X)].$$

where $\mathcal{M}(\mathcal{Z}) := \{\text{all (regular) probability measures on } \mathcal{Z}\}$.

◦ Existence of NE (ν^*, μ^*): Glicksberg's existence theorem [Glicksberg, 1952].

A re-thinking of GANs via the mixed Nash equilibrium

◦ **Upshot:** Our mixed Nash Equilibrium proposal $\equiv$ bi-affine matrix games

$$\min_{\nu \in \mathcal{M}(\Theta)} \max_{\mu \in \mathcal{M}(\mathcal{W})} \mathbb{E}_{w\sim\mu} \mathbb{E}_{X\sim P_{\text{real}}} [D_w(X)] - \mathbb{E}_{w\sim\mu} \mathbb{E}_{\theta\sim\nu} \mathbb{E}_{X\sim P^{\theta}_{\text{fake}}} [D_w(X)]$$

$$\Updownarrow$$

$$\min_{\nu \in \mathcal{M}(\Theta)} \max_{\mu \in \mathcal{M}(\mathcal{W})} \langle \mu, g \rangle - \langle \mu, G\nu \rangle$$

A re-thinking of GANs via the mixed Nash equilibrium

◦ **Upshot:** Our mixed Nash Equilibrium proposal $\equiv$ bi-affine matrix games

$$\min_{\nu \in \mathcal{M}(\Theta)} \max_{\mu \in \mathcal{M}(\mathcal{W})} \mathbb{E}_{w\sim\mu}\mathbb{E}_{X\sim P_{\text{real}}}[D_w(X)] - \mathbb{E}_{w\sim\mu}\mathbb{E}_{\theta\sim\nu}\mathbb{E}_{X\sim P^{\theta}_{\text{fake}}}[D_w(X)]$$

$$\Updownarrow$$

$$\min_{\nu \in \mathcal{M}(\Theta)} \max_{\mu \in \mathcal{M}(\mathcal{W})} \langle \mu, g\rangle - \langle \mu, G\nu\rangle$$

- ▷ $\langle \mu, h\rangle \coloneqq \int h d\mu$ for a measure μ and function h (Riesz representation)
- ▷ the g-function $g(w) \coloneqq \mathbb{E}_{X\sim P_{\text{real}}}[D_w(X)]$
- ▷ the linear operator G and its adjoint $G^\dagger$:

$$G : \mathcal{M}(\Theta) \to \text{ a function on } \mathcal{W}, \quad G^\dagger : \mathcal{M}(\mathcal{W}) \to \text{ a function on } \Theta,$$
$$(G\nu)(w) \coloneqq \mathbb{E}_{\theta\sim\nu}\mathbb{E}_{X\sim P^{\theta}_{\text{fake}}}[D_w(X)],$$
$$(G^\dagger\mu)(\theta) \coloneqq \mathbb{E}_{w\sim\mu}\mathbb{E}_{X\sim P^{\theta}_{\text{fake}}}[D_w(X)]$$

A re-thinking of GANs via the mixed Nash equilibrium

- **Upshot:** Our mixed Nash Equilibrium proposal $\equiv$ bi-affine matrix games

$$\min_{\nu \in \mathcal{M}(\Theta)} \max_{\mu \in \mathcal{M}(\mathcal{W})} \mathbb{E}_{w\sim\mu}\mathbb{E}_{X\sim P_{\text{real}}}[D_w(X)] - \mathbb{E}_{w\sim\mu}\mathbb{E}_{\theta\sim\nu}\mathbb{E}_{X\sim P^{\theta}_{\text{fake}}}[D_w(X)]$$

$$\Updownarrow$$

$$\min_{\nu \in \mathcal{M}(\Theta)} \max_{\mu \in \mathcal{M}(\mathcal{W})} \langle \mu, g\rangle - \langle \mu, G\nu\rangle$$

  - ▷ $\langle \mu, h\rangle := \int h d\mu$ for a measure μ and function h (Riesz representation)
  - ▷ the g-function $g(w) := \mathbb{E}_{X\sim P_{\text{real}}}[D_w(X)]$
  - ▷ the linear operator G and its adjoint $G^\dagger$:

$$G : \mathcal{M}(\Theta) \to \text{ a function on } \mathcal{W}, \quad G^\dagger : \mathcal{M}(\mathcal{W}) \to \text{ a function on } \Theta,$$

$$(G\nu)(w) := \mathbb{E}_{\theta\sim\nu}\mathbb{E}_{X\sim P^{\theta}_{\text{fake}}}[D_w(X)],$$

$$(G^\dagger\mu)(\theta) := \mathbb{E}_{w\sim\mu}\mathbb{E}_{X\sim P^{\theta}_{\text{fake}}}[D_w(X)]$$

- Caveat: **Infinite dimensions!!!**

A re-thinking of GANs via the mixed Nash equilibrium

- **Upshot:** Our mixed Nash Equilibrium proposal $\equiv$ bi-affine matrix games

$$\min_{\nu\in\mathcal{M}(\Theta)} \max_{\mu\in\mathcal{M}(\mathcal{W})} \mathbb{E}_{w\sim\mu}\mathbb{E}_{X\sim P_{\text{real}}}[D_w(X)] - \mathbb{E}_{w\sim\mu}\mathbb{E}_{\theta\sim\nu}\mathbb{E}_{X\sim P^{\theta}_{\text{fake}}}[D_w(X)]$$

$$\Updownarrow$$

$$\min_{\nu\in\mathcal{M}(\Theta)} \max_{\mu\in\mathcal{M}(\mathcal{W})} \langle \mu, g\rangle - \langle \mu, G\nu\rangle$$

  - ▷ $\langle \mu, h\rangle := \int h d\mu$ for a measure μ and function h (Riesz representation)
  - ▷ the g-function $g(w) := \mathbb{E}_{X\sim P_{\text{real}}}[D_w(X)]$
  - ▷ the linear operator G and its adjoint $G^{\dagger}$:

$$G : \mathcal{M}(\Theta) \to \text{ a function on } \mathcal{W}, \quad G^{\dagger} : \mathcal{M}(\mathcal{W}) \to \text{ a function on } \Theta,$$
$$(G\nu)(w) := \mathbb{E}_{\theta\sim\nu}\mathbb{E}_{X\sim P^{\theta}_{\text{fake}}}[D_w(X)],$$
$$(G^{\dagger}\mu)(\theta) := \mathbb{E}_{w\sim\mu}\mathbb{E}_{X\sim P^{\theta}_{\text{fake}}}[D_w(X)]$$

- Caveat: **Infinite dimensions!!!**
  - ▷ Ideas for finite-dimensional games apply: Mirror Descent.

Entropic Mirror Descent Iterates in Infinite Dimension

- Negative Shannon entropy and its Fenchel dual: (dz :=Lebesgue)
  - ▷ $\Phi(\mu) = \int \mu \log \frac{d\mu}{dz}$.
  - ▷ $\Phi^\star(h) = \log \int e^h$.
  - ▷ $d\Phi$ and $d\Phi^\star$: Fréchet derivatives.[1]

Theorem (Infinite-Dimensional Mirror Descent, informal)

For a learning rate η, a probability measure μ, and an arbitrary function h, we can equivalently define

$$\mu_+ = \mathsf{MD}_\eta(\mu, h) \quad \equiv \quad \mu_+ = d\Phi^\star\left(d\Phi(\mu) - \eta h\right) \equiv \quad d\mu_+ = \frac{e^{-\eta h} d\mu}{\int e^{-\eta h} d\mu}.$$

Moreover, the convergence rates are the same as in finite dimension.

- Continuous analog of the entropic mirror descent [Beck and Teboulle, 2003]
  - ▷ Mirror-prox also possible [Nemirovski, 2004]

[1] Under mild regularity conditions on the measure/function.

A Practical Algorithm

Algorithm 1: INFINITE-DIMENSIONAL ENTROPIC MD

Input: Initial distributions μ_1, ν_1, learning rate η

for $t = 1, 2, \ldots, T-1$ **do**

$\quad \nu_{t+1} = \mathrm{MD}_\eta\left(\nu_t, -G^\dagger \mu_t\right), \quad \mu_{t+1} = \mathrm{MD}_\eta\left(\mu_t, -g + G\nu_t\right);$

return $\bar{\nu}_T = \frac{1}{T}\sum_{t=1}^{T} \nu_t$ and $\bar{\mu}_T = \frac{1}{T}\sum_{t=1}^{T} \mu_t$.

A Practical Algorithm

Algorithm 1: INFINITE-DIMENSIONAL ENTROPIC MD

Input: Initial distributions μ_1, ν_1, learning rate η

for $t = 1, 2, \ldots, T-1$ **do**

$\quad \nu_{t+1} = \mathrm{MD}_\eta\left(\nu_t, -G^\dagger \mu_t\right), \quad \mu_{t+1} = \mathrm{MD}_\eta\left(\mu_t, -g + G\nu_t\right);$

return $\bar{\nu}_T = \frac{1}{T}\sum_{t=1}^T \nu_t$ and $\bar{\mu}_T = \frac{1}{T}\sum_{t=1}^T \mu_t$.

- How do we run it?
  - ▷ Cannot update probability measures.

A Practical Algorithm

Algorithm 1: INFINITE-DIMENSIONAL ENTROPIC MD

Input: Initial distributions μ_1, ν_1, learning rate η
for $t = 1, 2, \ldots, T-1$ **do**
  $\nu_{t+1} = \mathrm{MD}_\eta\left(\nu_t, -G^\dagger \mu_t\right), \quad \mu_{t+1} = \mathrm{MD}_\eta\left(\mu_t, -g + G\nu_t\right);$
return $\bar{\nu}_T = \frac{1}{T}\sum_{t=1}^{T} \nu_t$ and $\bar{\mu}_T = \frac{1}{T}\sum_{t=1}^{T} \mu_t$.

- How do we run it?
  - ▷ Cannot update probability measures.

- Key idea: Can take **samples** using **SGLD** [Welling and Teh, 2011]!!
  - ▷ Leading to updates as cheap as SGD.
  - ▷ For more details as well as numerical evidence, please visit our poster.

Thanks!!