Generalization Theory for GANs

Zhifeng Kong
University of California San Diego
z4kong@eng.ucsd.edu
Overview

Notations

Metrics

\( f \) divergence
Integral Probability Metric (IPM)

Generalization Theory

Definition I
Results I
Definition II
Results II

Discussion

Zhifeng Kong

Generalization Theory for GANs
## Notations

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Meaning</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\mathcal{F}$</td>
<td>Function class</td>
</tr>
<tr>
<td>$D/G$</td>
<td>Discriminator / Generator</td>
</tr>
<tr>
<td>$\mathcal{F}_D / \mathcal{F}_G$</td>
<td>Discriminator / Generator class</td>
</tr>
<tr>
<td>$\mu, \nu$</td>
<td>Probability distributions</td>
</tr>
<tr>
<td>$\mu^m, \nu^m$</td>
<td>Empirical distributions over $m$ samples</td>
</tr>
<tr>
<td>$\mathcal{D}_{\text{real}}$</td>
<td>Real distribution</td>
</tr>
<tr>
<td>$\mathcal{D}_G$</td>
<td>Distribution of data from $G$</td>
</tr>
<tr>
<td>$d$</td>
<td>Metric between two distributions</td>
</tr>
</tbody>
</table>
Objectives

In theoretical analysis:

\[
\inf_{G \in \mathcal{F}_G} d(D_{\text{real}}, D_G)
\]

In practice: something different

\[
\min_{G \in \mathcal{F}_G} d(D^m_{\text{real}}, D_G)
\]

\[
\min_{G \in \mathcal{F}_G} \mathbb{E}_{D^m_G} d(D^m_{\text{real}}, D^m_G)
\]
Example

If $\mathcal{F}_G$ is the set of Mix of Gaussian distributions and $d = KL$, then we obtain the Maximum Likelihood Estimate for Gaussian Mixture Model.

$$\arg \min_{\nu} KL(\mu, \nu) = \arg \min_{\nu} \int \mu(x) \log \frac{\mu(x)}{\nu(x)} dx$$
$$= \arg \max_{\nu} \int \mu(x) \log \nu(x) dx$$
$$= \arg \max_{\nu} \mathbb{E}_{x \sim \mu} \log \nu(x)$$
$$\Rightarrow \text{Maximum Likelihood Estimation}$$

In GANs [Goodfellow et al. (2014)], there are two types of metrics: IPM and $f$ divergence.
\( f \) divergence [Nowozin et al. (2016)]

\[
d_f(\mu, \nu) = \mathbb{E}_{x \sim \nu} f \left( \frac{\mu(x)}{\nu(x)} \right) = \int f \left( \frac{\mu(x)}{\nu(x)} \right) \nu(x) \, dx
\]

GAN objective (\( d = d_f \)):

\[
\inf_{G \in \mathcal{F}_G} d_f(D_{\text{real}}, D_G)
\]

\( \Rightarrow \)

\[
\inf_{G \in \mathcal{F}_G} \mathbb{E}_{x \sim D_G} f \left( \frac{D_{\text{real}}}{D_G} \right)
\]

There is no \( \mathcal{F}_D \)!
### Examples

<table>
<thead>
<tr>
<th>$f(t)$</th>
<th>$d_f$</th>
<th>Output</th>
</tr>
</thead>
<tbody>
<tr>
<td>$t \log t - (t + 1) \log(t + 1)$</td>
<td>JS divergence</td>
<td>Vanilla-GAN</td>
</tr>
<tr>
<td>$t \log t$</td>
<td>KL divergence</td>
<td>KL-GAN</td>
</tr>
<tr>
<td>$- \log t$</td>
<td>Rev-KL divergence</td>
<td></td>
</tr>
<tr>
<td>$</td>
<td>t - 1</td>
<td>/2$</td>
</tr>
<tr>
<td>$(\sqrt{t} - 1)^2$</td>
<td>Hellinger distance</td>
<td></td>
</tr>
<tr>
<td>$(t - 1)^2$</td>
<td>$\chi^2$ divergence</td>
<td>LS-GAN</td>
</tr>
</tbody>
</table>

Note: A constant factor in the JS divergence is removed.

See [Goodfellow et al. (2014), Nowozin et al. (2016), Zhao et al. (2016), Mao et al. (2017), Beran et al. (1977)].
Plots

Figure: $f$ divergences
Example: Vanilla-GAN

Let $f(t) = t \log t - (t + 1) \log(t + 1)$. Then,

$$d_f(\mu, \nu) = \mathbb{E}_\mu \log \left( \frac{\mu}{\mu + \nu} \right) + \mathbb{E}_\nu \log \left( \frac{\nu}{\mu + \nu} \right)$$

Let $\mu = D_{\text{real}}, \nu = D_G$, and the optimal discriminator

$$D^* = \frac{D_{\text{real}}}{D_{\text{real}} + D_G} = \frac{\mu}{\mu + \nu}$$

$$\Rightarrow \inf_{G \in \mathcal{F}_G} d_f(D_{\text{real}}, D_G) = \inf_{G \in \mathcal{F}_G} \mathbb{E}_{D_{\text{real}}} \log D^* + \mathbb{E}_{D_G} \log(1 - D^*)$$

$$= \inf_{G \in \mathcal{F}_G} \sup_{D} \mathbb{E}_{D_{\text{real}}} \log D + \mathbb{E}_{D_G} \log(1 - D)$$

$$\Rightarrow \text{Vanilla-GAN}$$

Zhifeng Kong

Generalization Theory for GANs

UC San Diego
Integral Probability Metric (IPM) [Sriperumbudur et al. (2009)]

IPM:
\[
d_F(\mu, \nu) = \sup_{D \in \mathcal{F}} \mathbb{E}_{x \sim \mu} D(x) - \mathbb{E}_{x \sim \nu} D(x)
\]

GAN objective \((d = d_{\mathcal{F}_D})\):
\[
\inf_{G \in \mathcal{F}_G} d_{\mathcal{F}_D}(\mathcal{D}_{\text{real}}, \mathcal{D}_G)
\]
\[
\Rightarrow \quad \inf_{G \in \mathcal{F}_G} \sup_{D \in \mathcal{F}_D} \mathbb{E}_{x \sim \mathcal{D}_{\text{real}}} D(x) - \mathbb{E}_{x \sim \mathcal{D}_G} D(x)
\]
### Examples

<table>
<thead>
<tr>
<th>$\mathcal{F}_D$</th>
<th>Metric</th>
<th>Output</th>
</tr>
</thead>
<tbody>
<tr>
<td>${ f : | f |_{Lip} \leq 1 }$</td>
<td>$d_W$</td>
<td>W-GAN</td>
</tr>
<tr>
<td>${ f : | f |_{H_k} \leq 1 }$</td>
<td>MMD</td>
<td>MMD-GAN</td>
</tr>
<tr>
<td>${ f : | f |_{\infty} \leq 1 }$</td>
<td>Total Variation</td>
<td>EB-GAN</td>
</tr>
<tr>
<td>${ f : | f |_{BL} \leq 1 }$</td>
<td>Dudley metric</td>
<td>Dudley-GAN</td>
</tr>
<tr>
<td>More complicated</td>
<td>Fisher-IPM</td>
<td>Fisher-GAN</td>
</tr>
<tr>
<td>More complicated</td>
<td>Sobolev-IPM</td>
<td>Sobolev-GAN</td>
</tr>
</tbody>
</table>

$d_W = \text{Wasserstein distance}; \| f \|_{BL} = \| f \|_{Lip} + \| f \|_{\infty};$

MMD = Maximum Mean Discrepancy.

See [Gulrajani et al. (2017), Li et al. (2017), Zhao et al. (2016), Anonymous (2018), Mroueh et al. (2017a), Mroueh et al. (2017b)].
Definition for Generalization I [Arora et al. (2017)]

**Definition.** $\mathcal{D}_G$ generalizes under distance $d$ with generalization error $\epsilon$ if with high probability,

$$|d(\mathcal{D}_{real}, \mathcal{D}_G) - d(\mathcal{D}_{real}^m, \mathcal{D}_G^m)| \leq \epsilon$$

Advantage: Over-fitting is avoided.

Drawback: No bound on $d(\mathcal{D}_{real}^m, \mathcal{D}_G^m)$; $\mathcal{D}_G$ could be arbitrary.

See Definition 1 in [Arora et al. (2017)].
Metrics that do not generalize

**Theorem.** Let $\mu, \nu \sim \mathcal{N}(0, \frac{1}{d} I)$, then with probability at least $1 - m^2 \exp(-\Omega(d))$,

$$d_{JS}(\mu^m, \nu^m) = \log 2$$

$$d_{W}(\mu^m, \nu^m) \geq 1.1$$

See Lemma 1 in [Arora et al. (2017)].
Neural Network distance generalizes

**Definition.** $\mathcal{F}_D = \mathcal{F} = \{\text{neural networks that map } \mathbb{R}^d \text{ to } [0, 1]\}$. Let $\phi$ be a concave measuring function, then the neural network distance (neural $\phi$ divergence) is defined as

$$d_{\mathcal{F}, \phi}(\mu, \nu) = \sup_{D \in \mathcal{F}} \mathbb{E}_{x \sim \mu} \phi(D(x)) + \mathbb{E}_{x \sim \nu} \phi(1 - D(x)) - 2\phi(1/2)$$

**Theorem.** $\exists$ constant $c$ such that when $m \geq \frac{cp\Delta^2 \log(LL_{\Phi}p/\epsilon)}{\epsilon^2}$, we have with probability as least $1 - \exp(-p)$,

$$|d_{\mathcal{F}, \phi}(\mu, \nu) - d_{\mathcal{F}, \phi}(\mu^m, \nu^m)| \leq \epsilon$$

See Definition 2 and Theorem 3.1 in [Arora et al. (2017)].
Rademacher Complexity based Results

**Definition.** Rademacher complexity of function class $\mathcal{F}$ on distribution $\mu$ is defined by

$$\mathcal{R}_m(\mathcal{F}, \mu) = \mathbb{E}_{\epsilon, x} \sup_{f \in \mathcal{F}} \left| \frac{1}{m} \sum_{i=1}^{m} \epsilon_i f(x_i) \right|$$

where $\epsilon_i \overset{i.i.d.}{\sim} Unif\{-1, +1\}$, $x_i \overset{i.i.d.}{\sim} \mu$. Define

$$\mathcal{R}_m(\mathcal{F}, \mathcal{G}) = \sup_{\mu \in \mathcal{G}} \mathcal{R}_m(\mathcal{F}, \mu)$$

See [Balcan (2011), Bai et al. (2018)].
Rademacher Complexity based Results

**Theorem.** For neural network distance ($\mathcal{F}_D = \mathcal{F} =$neural nets, $\phi(t) = t$), we have $\forall \mu, \nu \in \mathcal{F}_G$,

$$\E_{\mu^m, \nu^m} |d_\mathcal{F}(\mu, \nu) - d_\mathcal{F}(\mu^m, \nu^m)| \leq 4\mathcal{R}_m(\mathcal{F}_D, \mathcal{F}_G)$$

**Lemma.** Let $\epsilon > 0$. Suppose $\mathcal{F}$ satisfies that $\forall \nu \in \mathcal{F}_G$, $\exists f \in \mathcal{F}$ such that $\|f - \log \mu + \log \nu\|_\infty \leq \epsilon$, and $\forall f \in \mathcal{F}$, $f$ is $L$-Lipschitz, then we have

$$KL(\mu, \nu) + KL(\nu, \mu) - \epsilon \leq d_\mathcal{F}(\mu, \nu) \leq L \cdot d_W(\mu, \nu)$$

See Theorem 2.1, Lemma 4.1 and Theorem 4.3 in [Bai et al. (2018)].
Definition for Generalization II [Zhang et al. (2018)]

Assume

\[ d(D^m_{\text{real}}, D_G) \leq \inf_{\nu \in F_G} d(D^m_{\text{real}}, \nu) + \epsilon \]

Can we bound

\[ d(D_{\text{real}}, D_G) - \inf_{\nu \in F_G} d(D_{\text{real}}, \nu) \]

Advantage: \( D_G \) is near optimal, thus making more sense.
Drawback: \( d(D^m_{\text{real}}, D_G) \) is tractable for IPMs but intractable for many \( f \) divergences.
Theorem. Suppose $\mathcal{F}_D$ is even ($f \in \mathcal{F}_D$ implies $-f \in \mathcal{F}_D$) and $
abla = \sup_{f \in \mathcal{F}_D} \|f\|_\infty$, then with probability at least $1 - \delta$,

$$d_{\mathcal{F}}(D_{\text{real}}, D_G) - \inf_{\nu \in \mathcal{F}_G} d_{\mathcal{F}}(D_{\text{real}}, \nu) \leq 2d_{\mathcal{F}}(D_{\text{real}}, D_{\text{real}}^m) + \epsilon \\
\leq 4R_m(\mathcal{F}_D, D_{\text{real}}) + 2\mathcal{O}\left(\Delta \sqrt{\frac{\log 1/\delta}{m}}\right) + \epsilon$$

See Theorem 3.1 in [Zhang et al. (2018)].
Results for Neural $\phi$ divergence ($d = d_{\mathcal{F},\phi}$)

**Theorem.** With probability at least $1 - 2\delta$ ($\mathcal{F}_D$ does not need to be even),

$$d_{\mathcal{F},\phi}(D_{\text{real}}, D_G) - \inf_{\nu \in \mathcal{F}_G} d_{\mathcal{F},\phi}(D_{\text{real}}, \nu) \leq 4\mathcal{R}_m(\mathcal{F}_D, D_{\text{real}}) + 2O \left( \Delta \sqrt{\frac{\log 1/\delta}{m}} \right) + \epsilon$$

See Theorem B.3 in [Zhang et al. (2018)].
Results under spectrum control

**Theorem.** (Informal) Under some Lipschitz conditions and spectrum norm constraints (Assumption 1 in [Jiang et al. (2019)]), with probability at least $1 - \delta$

\[
d_{\mathcal{F},\phi}(\mathcal{D}_{\text{real}}, \mathcal{D}_G) - \inf_{\nu \in \mathcal{F}_G} d_{\mathcal{F},\phi}(\mathcal{D}_{\text{real}}, \nu) \leq \tilde{O}\left(\sqrt{\frac{d^2L}{m}}\right)
\]

See Theorem 2 in [Jiang et al. (2019)].
Discussion

- Neither of the definition is perfect.

- It is possible to combine the two definitions: bound

\[ d(D_{real}, D_G) - \inf_{\nu \in \mathcal{F}_G} d(D_{real}, \nu) \]

assuming that

\[ d(D^m_{real}, D^m_G) \leq \inf_{\nu \in \mathcal{F}_G} d(D^m_{real}, \nu^m) + \epsilon \]
Relationship to previous definitions

- Local version of the first definition: we only bound
  
  $$|d(D_{\text{real}}, D_G) - d(D_{\text{real}}, D_{G}^m)|$$

  for $D_{G}^m$ closed to $D_{\text{real}}^m$.

- Revised version of the second definition: the intractability of
  
  $d(D_{\text{real}}^m, D_G)$ is resolved.
References I


References III


THANK YOU