Generalization Theory for GANs

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

Overview

Notations

Symbol	Meaning
$\mathcal{F}$	Function class
D/G	Discriminator / Generator
$\mathcal{F}_D/\mathcal{F}_G$	Discriminator / Generator class
μ, ν	Probability distributions
μ^m, ν^m	Empirical distributions over m samples
$\mathcal{D}_{real}$	Real distribution
$\mathcal{D}_G$	Distribution of data from G
d	Metric between two distributions

Objectives

In theoretical analysis:

$$\inf_{G \in \mathcal{F}_G} d(\mathcal{D}_{real}, \mathcal{D}_G)$$

In practice: something different

$$\min_{G \in \mathcal{F}_G} d(\mathcal{D}_{real}^m, \mathcal{D}_G)$$

$$\min_{G \in \mathcal{F}_G} \mathbb{E}_{\mathcal{D}_G^m} d(\mathcal{D}_{real}^m, \mathcal{D}_G^m)$$

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.

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

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

f divergence [Nowozin et al. (2016)]

f divergence:

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

GAN objective ($d = d_f$):

$$\inf_{G \in \mathcal{F}_G} d_f(\mathcal{D}_{real}, \mathcal{D}_G)$$

$\Rightarrow$

$$\inf_{G \in \mathcal{F}_G} \mathbb{E}_{x \sim \mathcal{D}_G} f\left(\frac{\mathcal{D}_{real}}{\mathcal{D}_G}\right)$$

There is no $\mathcal{F}_D$!

Examples

$f(t)$	d_f	Output
$t\log t-(t+1)\log(t+1)$	JS divergence	Vanilla-GAN
$t\log t$	KL divergence	KL-GAN
$-\log t$	Rev-KL divergence	
$\lvert t-1\rvert/2$	Total Variation	EB-GAN
$(\sqrt{t}-1)^2$	Hellinger distance	
$(t-1)^2$	χ^2 divergence	LS-GAN

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

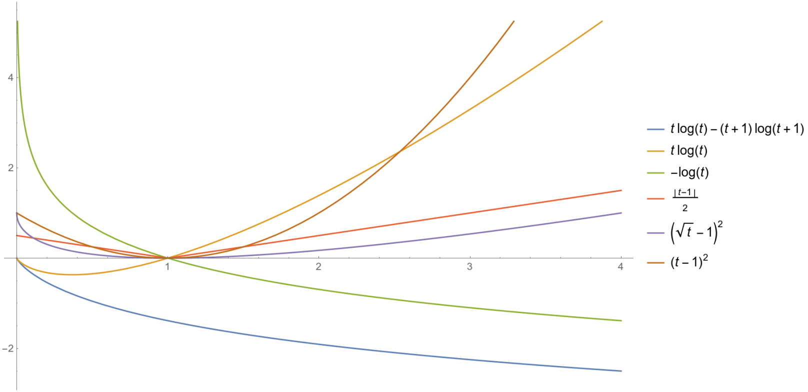

Figure: f divergences

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 $\mu = \mathcal{D}_{real}, \nu = \mathcal{D}_G$, and the optimal discriminator

$$D^* = \frac{\mathcal{D}_{real}}{\mathcal{D}_{real} + \mathcal{D}_G} = \frac{\mu}{\mu+\nu}$$

$$\begin{aligned}
\Rightarrow \inf_{G\in\mathcal{F}_G} d_f(\mathcal{D}_{real},\mathcal{D}_G) &= \inf_{G\in\mathcal{F}_G} \mathbb{E}_{\mathcal{D}_{real}} \log D^* + \mathbb{E}_{\mathcal{D}_G}\log(1-D^*) \\
&= \inf_{G\in\mathcal{F}_G} \sup_D \mathbb{E}_{\mathcal{D}_{real}} \log D + \mathbb{E}_{\mathcal{D}_G}\log(1-D) \\
&\Rightarrow \text{Vanilla-GAN}
\end{aligned}$$

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

IPM:

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

GAN objective ($d = d_{\mathcal{F}_D}$):

$$\inf_{G \in \mathcal{F}_G} d_{\mathcal{F}_D}(\mathcal{D}_{real}, \mathcal{D}_G)$$

$\Rightarrow$

$$\inf_{G \in \mathcal{F}_G} \sup_{D \in \mathcal{F}_D} \mathbb{E}_{x \sim \mathcal{D}_{real}} D(x) - \mathbb{E}_{x \sim \mathcal{D}_G} D(x)$$

Examples

$\mathcal{F}_{\mathcal{D}}$	Metric	Output
$\{f : \|f\|_{Lip} \leq 1\}$	d_W	W-GAN
$\{f : \|f\|_{\mathcal{H}_k} \leq 1\}$	MMD	MMD-GAN
$\{f : \|f\|_{\infty} \leq 1\}$	Total Variation	EB-GAN
$\{f : \|f\|_{BL} \leq 1\}$	Dudley metric	Dudley-GAN
More complicated	Fisher-IPM	Fisher-GAN
More complicated	Sobolev-IPM	Sobolev-GAN

d_W = 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 ϵ 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. $\mathcal{F}_D = \mathcal{F} = \{$neural networks that map $\mathbb{R}^d$ to $[0,1]\}$. Let ϕ be a concave measuring function, then the neural network distance (neural ϕ 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 μ 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 $\forall \mu, \nu \in \mathcal{F}_G$,

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

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

$$d(\mathcal{D}_{real}^m, \mathcal{D}_G) \le \inf_{\nu \in \mathcal{F}_G} d(\mathcal{D}_{real}^m, \nu) + \epsilon$$

Can we bound

$$d(\mathcal{D}_{real}, \mathcal{D}_G) - \inf_{\nu \in \mathcal{F}_G} d(\mathcal{D}_{real}, \nu)$$

Advantage: $\mathcal{D}_G$ is near optimal, thus making more sense.
Drawback: $d(\mathcal{D}_{real}^m, \mathcal{D}_G)$ is tractable for IPMs but intractable for many f divergences.

Results for IPMs ($d = d_{\mathcal{F}}$)

Theorem. Suppose $\mathcal{F}_D$ is even ($f \in \mathcal{F}_D$ implies $-f \in \mathcal{F}_D$) and $\Delta = \sup_{f\in\mathcal{F}_D} \|f\|_\infty$, then with probability at least $1-\delta$,

$$\begin{aligned} & d_{\mathcal{F}}(\mathcal{D}_{real}, \mathcal{D}_G) - \inf_{\nu\in\mathcal{F}_G} d_{\mathcal{F}}(\mathcal{D}_{real}, \nu) \\ & \quad \leq 2d_{\mathcal{F}}(\mathcal{D}_{real}, \mathcal{D}_{real}^m) + \epsilon \\ & \quad \leq 4\mathcal{R}_m(\mathcal{F}_D, \mathcal{D}_{real}) + 2\mathcal{O}\left(\Delta\sqrt{\frac{\log 1/\delta}{m}}\right) + \epsilon \end{aligned}$$

See Theorem 3.1 in [Zhang et al. (2018)].

Results for Neural ϕ divergence ($d = d_{\mathcal{F},\phi}$)

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

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

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 $1-\delta$

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

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

Discussion

- Neither of the definition is perfect.
- It is possible to combine the two definitions: bound

$$d(\mathcal{D}_{real}, \mathcal{D}_G) - \inf_{\nu \in \mathcal{F}_G} d(\mathcal{D}_{real}, \nu)$$

assuming that

$$d(\mathcal{D}^m_{real}, \mathcal{D}^{m}_G) \leq \inf_{\nu \in \mathcal{F}_G} d(\mathcal{D}^m_{real}, \nu^{m}) + \epsilon$$

Relationship to previous definitions

- Local version of the first definition: we only bound

$$|d(\mathcal{D}_{real}, \mathcal{D}_G) - d(\mathcal{D}^m_{real}, \mathcal{D}^m_G)|$$

  for $\mathcal{D}^m_G$ closed to $\mathcal{D}^m_{real}$.

- Revised version of the second definition: the intractability of $d(\mathcal{D}^m_{real}, \mathcal{D}_G)$ is resolved.

References I

Anonymous. Deep lipschitz networks and dudley gans. 2018.

S. Arora et al. Generalization and equilibrium in generative adversarial nets (gans). In *ICML*, pages 224–232. JMLR.org, 2017.

Y. Bai et al. Approximability of discriminators implies diversity in gans. *arXiv preprint arXiv:1806.10586*, 2018.

M. F. Balcan. Rademacher complexity, November 2011.

R. Beran et al. Minimum hellinger distance estimates for parametric models. *The annals of Statistics*, 5(3):445–463, 1977.

I. Goodfellow et al. Generative adversarial nets. In *NIPS*, pages 2672–2680, 2014.

I. Gulrajani et al. Improved training of wasserstein gans. In *NIPS*, pages 5767–5777, 2017.

References II

H. Jiang et al. On computation and generalization of generative adversarial networks under spectrum control. In *ICLR*, 2019.

C.-L. Li et al. Mmd gan: Towards deeper understanding of moment matching network. In *NIPS*, pages 2203–2213, 2017.

X. Mao et al. Least squares generative adversarial networks. In *IEEE ICCV*, pages 2794–2802, 2017.

Y. Mroueh et al. Fisher gan. In *NIPS*, pages 2513–2523, 2017a.

Y. Mroueh et al. Sobolev gan. *arXiv preprint arXiv:1711.04894*, 2017b.

S. Nowozin et al. f-gan: Training generative neural samplers using variational divergence minimization. In *NIPS*, pages 271–279, 2016.

References III

B. K. Sriperumbudur et al. On integral probability metrics, ϕ-divergences and binary classification. *arXiv preprint arXiv:0901.2698*, 2009.

P. Zhang et al. On the discrimination-generalization tradeoff in gans. In *ICLR*, 2018.

J. Zhao et al. Energy-based generative adversarial network. *arXiv preprint arXiv:1609.03126*, 2016.

THANK YOU