topics about $f$-divergence

Presented by Liqun Chen
Mar 16th, 2018
Outline

1. $f$-GAN: Training Generative Neural Samplers using Variational Divergence Minimization
   - Introduction
   - Method
   - Experiments

2. $f$-GANs in an Information Geometric Nutshell
   - Introduction
   - Method
   - Experiments
Outline

1 \(f\)-GAN: Training Generative Neural Samplers using Variational Divergence Minimization
   - Introduction
   - Method
   - Experiments

2 \(f\)-GANs in an Information Geometric Nutshell
   - Introduction
   - Method
   - Experiments
Introduction

Given a generative model $Q$ from a class $\mathcal{Q}$ of possible models we are generally interested in performing one or multiple of the following operations:

- **Sampling.** Produce a sample from $Q$. By inspecting samples or calculating a function on a set of samples we can obtain important insight into the distribution or solve decision problems.

- **Estimation.** Given a set of iid samples $\{x_1, x_2, \ldots, x_n\}$ from an unknown true distribution $P$, find $Q \in \mathcal{Q}$ that best describes the true distribution.

- **Point-wise likelihood evaluation.** Given a sample $x$, evaluate the likelihood $Q(x)$.

Generative adversarial networks (GAN) allows sampling and estimation.
In original GAN paper, we use symmetric *Jensen-Shannon divergence*:

\[
D_{JS}(P\|Q) = \frac{1}{2}D_{KL}(P\|\frac{1}{2}(P+Q)) + \frac{1}{2}D_{KL}(Q\|\frac{1}{2}(P+Q)),
\]

(1)

where \(D_{KL}\) denotes the Kullback-Leibler divergence.

In this work the authors show that the principle of GANs is more general and we can extend the variational divergence estimation framework to recover the GAN training objective and generalize it to arbitrary \(f\)-divergences.
Introduction

Contributions of this paper:

- they derived the GAN training objectives for all $f$-divergences and provide as example additional divergence functions, including the Kullback-Leibler and Pearson divergences.
- they simplified the saddle-point optimization procedure of the original GAN paper and provide a theoretical justification.
- they provided experimental insight into which divergence function is suitable for estimating generative neural samplers for natural images.
Given two distributions $P$ and $Q$ that possess, respectively, an absolutely continuous density function $p$ and $q$ with respect to a base measure $dx$ defined on the domain $\mathcal{X}$, we define the $f$-divergence,

$$D_f(P\|Q) = \int_{\mathcal{X}} q(x) f \left( \frac{p(x)}{q(x)} \right) \, dx,$$

where the generator function $f : \mathbb{R}_+ \rightarrow \mathbb{R}$ is a convex, lower-semicontinuous function satisfying $f(1) = 0$. 
Variational Estimation of $f$-divergence

Every convex, lower-semicontinuous function $f$ has a *convex conjugate* function $f^*$, also known as *Fenchel conjugate*. This function is defined as

$$f^*(t) = \sup_{u \in \text{dom} f} \{ut - f(u)\}.$$  \hfill (3)

The function $f^*$ is again convex and lower-semicontinuous and the pair $(f, f^*)$ is dual to another in the sense that $f^{**} = f$. Therefore, we can also represent $f$ as $f(u) = \sup_{t \in \text{dom} f^*} \{tu - f^*(t)\}$. 
Variational Estimation of $f$-divergence

Using the conjugate function, we can rewrite $f$-divergence as this:

$$D_f(P\|Q) = \int_{\mathcal{X}} q(x) \sup_{t \in \text{dom}f^*} \left\{ t \frac{p(x)}{q(x)} - f^*(t) \right\} \, dx$$

$$\geq \sup_{T \in \mathcal{T}} \left( \int_{\mathcal{X}} p(x) T(x) \, dx - \int_{\mathcal{X}} q(x) f^*(T(x)) \, dx \right)$$

$$= \sup_{T \in \mathcal{T}} \left( \mathbb{E}_{x \sim P} [T(x)] - \mathbb{E}_{x \sim Q} [f^*(T(x))] \right), \quad (4)$$

where $\mathcal{T}$ is an arbitrary class of functions $T : \mathcal{X} \to \mathbb{R}$.

The above derivation yields a lower bound for two reasons:

- because of Jensen’s inequality when swapping the integration and supremum operations.
- the class of functions $\mathcal{T}$ may contain only a subset of all possible functions.
Variational Estimation of $f$-divergence

By taking the variation of the lower bound in (4) w.r.t. $T$, we find that under mild conditions on $f$, the bound is tight for

$$T^*(x) = f' \left( \frac{p(x)}{q(x)} \right),$$

(5)

For example, the popular reverse Kullback-Leibler divergence corresponds to $f(u) = -\log u$ resulting in $T^*(x) = -q(x)/p(x)$. 
Variational Divergence Minimization

we follow the generative-adversarial: using two neural networks, $Q$ and $T$.

- $Q$ is the generative model (generator), parametrized $Q$ through a vector $\theta$ and write $Q_\theta$.
- $T$ is the variational function (discriminator), taking as input a sample and returning a scalar, parametrized $T$ using a vector $\omega$ and write $T_\omega$.

We can learn a generative model $Q_\theta$ by finding a saddle-point of the following $f$-GAN objective function, where we minimize with respect to $\theta$ and maximize with respect to $\omega$,

$$F(\theta, \omega) = \mathbb{E}_{x \sim P} [T_\omega(x)] - \mathbb{E}_{x \sim Q_\theta} [f^*(T_\omega(x))].$$ (6)
Variational Divergence Minimization

To apply the variational objective (6) for different $f$-divergences, we need to respect the domain $\text{dom}_{f^*}$ of the conjugate functions $f^*$. The authors assume that variational function $T_\omega$ is represented in the form $T_\omega(x) = g_f(V_\omega(x))$ and rewrite the saddle objective (6) as follows:

$$F(\theta, \omega) = \mathbb{E}_{x \sim P}[g_f(V_\omega(x))] + \mathbb{E}_{x \sim Q_\theta}[-f^*(g_f(V_\omega(x)))] ,$$

where $V_\omega: \mathcal{X} \rightarrow \mathbb{R}$ without any range constraints on the output, and $g_f: \mathbb{R} \rightarrow \text{dom}_{f^*}$ is an output activation function specific to the $f$-divergence used.
### Table 2: Recommended final layer activation functions and critical variational function level defined by $f'(1)$. The critical value $f'(1)$ can be interpreted as a classification threshold applied to $T(x)$ to distinguish between true and generated samples.
### Experiments

**MNIST digits:**

<table>
<thead>
<tr>
<th>Training divergence</th>
<th>KDE $\langle LL \rangle$ (nats)</th>
<th>$\pm$ SEM</th>
</tr>
</thead>
<tbody>
<tr>
<td>Kullback-Leibler</td>
<td>416</td>
<td>5.62</td>
</tr>
<tr>
<td>Reverse Kullback-Leibler</td>
<td>319</td>
<td>8.36</td>
</tr>
<tr>
<td>Pearson $\chi^2$</td>
<td>429</td>
<td>5.53</td>
</tr>
<tr>
<td>Neyman $\chi^2$</td>
<td>300</td>
<td>8.33</td>
</tr>
<tr>
<td>Squared Hellinger</td>
<td>-708</td>
<td>18.1</td>
</tr>
<tr>
<td>Jeffrey</td>
<td>-2101</td>
<td>29.9</td>
</tr>
<tr>
<td>Jensen-Shannon</td>
<td>367</td>
<td>8.19</td>
</tr>
<tr>
<td>GAN</td>
<td>305</td>
<td>8.97</td>
</tr>
<tr>
<td>Variational Autoencoder</td>
<td>445</td>
<td>5.36</td>
</tr>
<tr>
<td>KDE MNIST train (60k)</td>
<td>502</td>
<td>5.99</td>
</tr>
</tbody>
</table>
Outline

1. \(f\)-GAN: Training Generative Neural Samplers using Variational Divergence Minimization
   - Introduction
   - Method
   - Experiments

2. \(f\)-GANs in an Information Geometric Nutshell
   - Introduction
   - Method
   - Experiments
**Introduction**

We define two important classes of distortion measure.

**Definition**

For any two distributions \( P \) and \( Q \) having respective densities \( P \) and \( Q \) absolutely continuous with respect to a base measure \( \mu \), the \( f \)-divergence between \( P \) and \( Q \), where \( f : \mathbb{R}_+ \rightarrow \mathbb{R} \) is convex with \( f(1) = 0 \), is

\[
I_f(P \| Q) = \mathbb{E}_{X \sim Q} \left[ f \left( \frac{P(X)}{Q(X)} \right) \right] = \int_{\mathcal{X}} Q(x) \cdot f \left( \frac{P(x)}{Q(x)} \right) \, d\mu(x) .
\]  

(8)

For any convex differentiable \( \varphi : \mathbb{R}^d \rightarrow \mathbb{R} \), the \((\varphi\text{-})Bregman divergence between \( \theta \) and \( \varrho \) is:

\[
D_{\varphi}(\theta \| \varrho) = \varphi(\theta) - \varphi(\varrho) - (\theta - \varrho)^\top \nabla \varphi(\varrho) ,
\]  

(9)

where \( \varphi \) is called the generator of the Bregman divergence.
The contributions of this paper:

- $f$-GAN($P, \text{escort}(Q)) = D(\theta \| \vartheta) + \text{Penalty}(Q)$ for $P$ and $Q$ (with respective parameters $\theta$ and $\vartheta$) which happen to lie in a superset of exponential families called \textit{deformed exponential families}
- they developed a novel min-max game interpretation of eq. above.
\( \chi \)-logarithm

define generalizations of exponential families. Let \( \chi : \mathbb{R}_+ \rightarrow \mathbb{R}_+ \) be non-decreasing. They define the \( \chi \)-logarithm, \( \log_\chi \), as

\[
\log_\chi(z) = \int_1^z \frac{1}{\chi(t)} \, dt
\]  \hspace{1cm} (10)

The \( \chi \)-exponential is

\[
\exp_\chi(z) = 1 + \int_0^z \lambda(t) \, dt
\]  \hspace{1cm} (11)

where \( \lambda \) is defined by \( \lambda(\log_\chi(z)) = \chi(z) \).
Deformed exponential family

Definition

A distribution $P$ from a $\chi$-exponential family (or deformed exponential family, $\chi$ being implicit) with convex cumulant $C : \Theta \rightarrow \mathbb{R}$ and sufficient statistics $\phi : \mathcal{X} \rightarrow \mathbb{R}^d$ has density given by:

$$P_{\chi,C}(x|\theta, \phi) = \exp \chi(\phi(x)^\top \theta - C(\theta)) \ ,$$

(12)

with respect to a dominating measure $\mu$. Here, $\Theta$ is a convex open set and $\theta$ is called the coordinate of $P$. The escort density (or $\chi$-escort) of $P_{\chi,C}$ is

$$\hat{P}_{\chi,C} = \frac{1}{Z} \cdot \chi(P_{\chi,C}) \ ,$$

(13)

where

$$Z = \int_{\mathcal{X}} \chi(P_{\chi,C}(x|\theta, \phi))d\mu(x)$$

(14)

is the escort’s normalization constant.
Theorem

for any two $\chi$-exponential distributions $P$ and $Q$ with respective densities $P_{\chi,C}, Q_{\chi,C}$ and coordinates $\theta, \vartheta$,

$$D_C(\theta \parallel \vartheta) = \mathbb{E}_{X \sim \hat{Q}}[\log_\chi(Q_{\chi,C}(X)) - \log_\chi(P_{\chi,C}(X))] . \quad (15)$$

Definition

For any $\chi$-logarithm and distributions $P, Q$ having respective densities $P$ and $Q$ absolutely continuous with respect to base measure $\mu$, the $KL_\chi$ divergence between $P$ and $Q$ is defined as:

$$KL_\chi(P \parallel Q) = \mathbb{E}_{X \sim P} \left[ - \log_\chi \left( \frac{Q(X)}{P(X)} \right) \right] . \quad (16)$$

Since $\chi$ is non-decreasing, $-\log_\chi$ is convex ans so any $KL_\chi$ divergence is $f$-divergence.
Theorem

Letting distributions $P = P_{\chi,C}$ and $Q = Q_{\chi,C}$ for short in last theorem, we have

$$\mathbb{E}_{X \sim \hat{Q}}[\log \chi(Q_{\chi,C}(X)) - \log \chi(P_{\chi,C}(X))] = KL_{\chi\hat{Q}}(\hat{Q}\|P) - J(Q),$$

with

$$J(Q) = KL_{\chi\hat{Q}}(\hat{Q}\|Q).$$

Theorem

$KL_{\chi Q}(Q\|P)$ admits the variational formulation

$$KL_{\chi\hat{Q}}(\hat{Q}\|P) = \sup_{T \in \mathbb{R}_{++}^{\infty}} \left\{ \mathbb{E}_{X \sim P}[T(X)] - \mathbb{E}_{X \sim \hat{Q}}[(- \log \chi_{\hat{Q}})^*(T(X))] \right\},$$

(17)

with $\mathbb{R}_{++}^{\infty} = \mathbb{R} \setminus \mathbb{R}_{++}$. Furthermore, letting $Z$ denoting the normalization constant of the $\chi$-escort of $Q$, the optimum $T^* : \mathcal{X} \rightarrow \mathbb{R}_{++}$ to eq. (17) is

$$T^*(x) = -\frac{1}{Z} \cdot \frac{\chi(Q(x))}{\chi(P(x))}.$$  

(18)
Theorem

From the theorems above, we can have

\[
\sup_{T \in \mathbb{R}^{++}} \left\{ \mathbb{E}_{X \sim P}[T(X)] - \mathbb{E}_{X \sim \hat{Q}}[(- \log_{\hat{Q}})^*(T(X))] \right\} = D_C(\theta || \vartheta) + J(Q), \tag{19}
\]

where \(\theta\) (resp. \(\vartheta\)) is the coordinate of \(P\) (resp. \(Q\)).
### Method

<table>
<thead>
<tr>
<th>Name</th>
<th>$v(z)$</th>
<th>$\chi(z)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>ReLU ($\dagger$)</td>
<td>$\max{0, z}$</td>
<td>$\frac{1}{z &gt; 0}$</td>
</tr>
<tr>
<td>Leaky-ReLU ($\dagger$)</td>
<td>$\begin{cases} z \text{ if } z &gt; 0 \ \epsilon z \text{ if } z \leq 0 \end{cases}$</td>
<td>$\begin{cases} 1 \text{ if } z &gt; -\delta \ \frac{1}{\epsilon} \text{ if } z \leq -\delta \end{cases}$</td>
</tr>
<tr>
<td>$(\alpha, \beta)$-ELU ($\heartsuit$)</td>
<td>$\begin{cases} \frac{\beta z}{\exp(z) - 1} \text{ if } z &gt; 0 \ \alpha \text{ if } z \leq 0 \end{cases}$</td>
<td>$\begin{cases} \beta \text{ if } z &gt; \alpha \ \frac{z}{\beta} \text{ if } z \leq \alpha \end{cases}$</td>
</tr>
<tr>
<td>prop-$\tau$ ($\clubsuit$)</td>
<td>$k + \frac{\tau^<em>(z)}{\tau^</em>(0)}$</td>
<td>$\frac{\tau'(-\tau^<em>)^{-1}(\tau^</em>(0)z)}{\tau^*(0)}$</td>
</tr>
<tr>
<td>Softplus ($\bigcirc$)</td>
<td>$k + \log_2(1 + \exp(z))$</td>
<td>$\frac{1}{\log 2} \cdot (1 - 2^{-z})$</td>
</tr>
<tr>
<td>$\mu$-ReLU ($\heartsuit$)</td>
<td>$k + \frac{z + \sqrt{(1-\mu)^2 + z^2}}{2}$</td>
<td>$\frac{4z^2}{(1-\mu)^2 + 4z^2}$</td>
</tr>
<tr>
<td>LSU</td>
<td>$k + \begin{cases} 0 \text{ if } z &lt; -1 \ (1 + z)^2 \text{ if } z \in [-1, 1] \ 4z \text{ if } z &gt; 1 \end{cases}$</td>
<td>$\begin{cases} 2\sqrt{z} \text{ if } z &lt; 4 \ 4 \text{ if } z &gt; 4 \end{cases}$</td>
</tr>
</tbody>
</table>

Table 1: Some strongly/weakly admissible couples $(v, \chi)$. ($\dagger$): is the indicator function; ($\dagger$): $\delta \leq 0, 0 < \epsilon \leq 1$ and dom$(v) = [\delta/\epsilon, +\infty)$. ($\heartsuit$): $\beta \geq \alpha > 0$; ($\bigcirc$): $\ast$ is Legendre conjugate; ($\clubsuit$): $\mu \in [0, 1)$. Shaded: prop-$\tau$ activations; $k$ is a constant (e.g. such that $v(0) = 0$) (see text).
MNIST

Figure 1: Summary of our results on MNIST, on experiment A (left+center) and B (right). *Left:* comparison of different values of $\mu$ for the $\mu$-ReLU activation in the generator (ReLU = 1-ReLU, see text). Thicker horizontal dashed lines present the ReLU average baseline: for each color, points above the baselines represent values of $\mu$ for which ReLU is beaten on average. *Center:* comparison of different activations in the generator, for the same architectures as in the left plot. *Right:* comparison of different link function in the discriminator (see text, best viewed in color).