

Implicit Distribution Matching

Arsalan Zafar

Visual Loss Department
Deep Render Ltd

Overview

1. Background
2. Introduction
3. Class probability estimation
4. ϕ -Divergence methods
5. Bergman Divergences
6. Integral probability metric (IPM)
7. Final words

Background

- One way to view the visual loss is to match p_x to $p_{\hat{x}}$

Background

- One way to view the visual loss is to match p_x to $p_{\hat{x}}$
- In our entropy case, we can compute q , so minimise KL is possible

Background

- One way to view the visual loss is to match p_x to $p_{\hat{x}}$
- In our entropy case, we can compute q , so minimise KL is possible
- What if we want to match p_x and $p_{\hat{x}}$ but we can only sample from them?

Background

- One way to view the visual loss is to match p_x to $p_{\hat{x}}$
- In our entropy case, we can compute q , so minimise KL is possible
- What if we want to match p_x and $p_{\hat{x}}$ but we can only sample from them?
- Can we evaluate some divergence implicitly with $x \sim p_x$ and $x \sim p_{\hat{x}}$?

Introduction

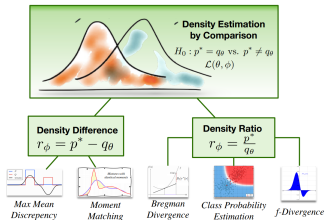


Figure: Learning in Implicit Generative Models

Introduction

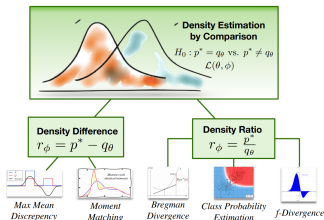


Figure: Learning in Implicit Generative Models

Ratio methods

1. Class probability estimation
2. ϕ -Divergence
3. Bregman Divergence

Introduction

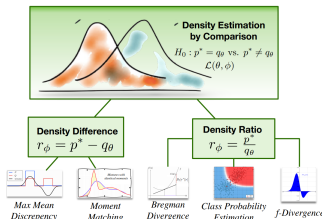


Figure: Learning in Implicit Generative Models

Ratio methods

1. Class probability estimation
2. ϕ -Divergence
3. Bregman Divergence

Difference methods (IPM)

1. Wasserstine distance
2. TV
3. MMD

Class probability estimation

Theorem (Density ratio — Bayes' rule)

$$\text{Density Ratio: } \frac{p(x)}{q(x)}$$

Class probability estimation

Theorem (Density ratio — Bayes' rule)

Density Ratio: $\frac{p(x)}{q(x)}$

Bayes' Rule: $p(y|x) = \frac{p(x|y)p(y)}{p(x)}$

Class probability estimation

Theorem (Density ratio — Bayes' rule)

$$\text{Density Ratio: } \frac{p(x)}{q(x)} \quad \text{Bayes' Rule: } p(y|x) = \frac{p(x|y)p(y)}{p(x)}$$

$$x \sim p \text{ label } y = 1 \quad \text{and} \quad \hat{x} \sim q \text{ label } y = 0 \quad (1)$$

Class probability estimation

Theorem (Density ratio — Bayes' rule)

$$\text{Density Ratio: } \frac{p(x)}{q(x)} \quad \text{Bayes' Rule: } p(y|x) = \frac{p(x|y)p(y)}{p(x)}$$

$$x \sim p \text{ label } y = 1 \quad \text{and} \quad \hat{x} \sim q \text{ label } y = 0 \quad (1)$$

We can represent $p(x) = p(x|y = 1)$ and $q(x) = p(x|y = 0)$.

Class probability estimation

Theorem (Density ratio — Bayes' rule)

$$\text{Density Ratio: } \frac{p(x)}{q(x)} \quad \text{Bayes' Rule: } p(y|x) = \frac{p(x|y)p(y)}{p(x)}$$

$$x \sim p \text{ label } y = 1 \quad \text{and} \quad \hat{x} \sim q \text{ label } y = 0 \quad (1)$$

We can represent $p(x) = p(x|y = 1)$ and $q(x) = p(x|y = 0)$.

$$\frac{p(x)}{q(x)}$$

Class probability estimation

Theorem (Density ratio — Bayes' rule)

$$\text{Density Ratio: } \frac{p(x)}{q(x)} \quad \text{Bayes' Rule: } p(y|x) = \frac{p(x|y)p(y)}{p(x)}$$

$$x \sim p \text{ label } y = 1 \quad \text{and} \quad \hat{x} \sim q \text{ label } y = 0 \quad (1)$$

We can represent $p(x) = p(x|y = 1)$ and $q(x) = p(x|y = 0)$.

$$\frac{p(x)}{q(x)} = \frac{p(x|y = 1)}{p(x|y = 0)}$$

Class probability estimation

Theorem (Density ratio — Bayes' rule)

$$\text{Density Ratio: } \frac{p(x)}{q(x)} \quad \text{Bayes' Rule: } p(y|x) = \frac{p(x|y)p(y)}{p(x)}$$

$$x \sim p \text{ label } y = 1 \quad \text{and} \quad \hat{x} \sim q \text{ label } y = 0 \quad (1)$$

We can represent $p(x) = p(x|y = 1)$ and $q(x) = p(x|y = 0)$.

$$\frac{p(x)}{q(x)} = \frac{p(x|y = 1)}{p(x|y = 0)} = \frac{\frac{p(y=1|x)p(x)}{p(y=1)}}{\frac{p(y=0|x)p(x)}{p(y=0)}}$$

Class probability estimation

Theorem (Density ratio — Bayes' rule)

$$\text{Density Ratio: } \frac{p(x)}{q(x)} \quad \text{Bayes' Rule: } p(y|x) = \frac{p(x|y)p(y)}{p(x)}$$

$$x \sim p \text{ label } y = 1 \quad \text{and} \quad \hat{x} \sim q \text{ label } y = 0 \quad (1)$$

We can represent $p(x) = p(x|y = 1)$ and $q(x) = p(x|y = 0)$.

$$\frac{p(x)}{q(x)} = \frac{p(x|y = 1)}{p(x|y = 0)} = \frac{\frac{p(y=1|x)p(x)}{p(y=1)}}{\frac{p(y=0|x)p(x)}{p(y=0)}} = \frac{p(y = 1|x)}{p(y = 0|x)}$$

Class probability estimation

Theorem (Density ratio — Bayes' rule)

$$\text{Density Ratio: } \frac{p(x)}{q(x)} \quad \text{Bayes' Rule: } p(y|x) = \frac{p(x|y)p(y)}{p(x)}$$

$$x \sim p \text{ label } y = 1 \quad \text{and} \quad \hat{x} \sim q \text{ label } y = 0 \quad (1)$$

We can represent $p(x) = p(x|y = 1)$ and $q(x) = p(x|y = 0)$.

$$\frac{p(x)}{q(x)} = \frac{p(x|y = 1)}{p(x|y = 0)} = \frac{\frac{p(y=1|x)p(x)}{p(y=1)}}{\frac{p(y=0|x)p(x)}{p(y=0)}} = \frac{p(y = 1|x)}{p(y = 0|x)} = \frac{\mathcal{D}(x; \theta)}{1 - \mathcal{D}(x; \theta)} \quad (2)$$

Class probability estimation

Theorem (Density ratio — Bayes' rule)

$$\text{Density Ratio: } \frac{p(x)}{q(x)} \quad \text{Bayes' Rule: } p(y|x) = \frac{p(x|y)p(y)}{p(x)}$$

$$x \sim p \text{ label } y = 1 \quad \text{and} \quad \hat{x} \sim q \text{ label } y = 0 \quad (1)$$

We can represent $p(x) = p(x|y = 1)$ and $q(x) = p(x|y = 0)$.

$$\frac{p(x)}{q(x)} = \frac{p(x|y = 1)}{p(x|y = 0)} = \frac{\frac{p(y=1|x)p(x)}{p(y=1)}}{\frac{p(y=0|x)p(x)}{p(y=0)}} = \frac{p(y = 1|x)}{p(y = 0|x)} = \frac{\mathcal{D}(x; \theta)}{1 - \mathcal{D}(x; \theta)} \quad (2)$$

Where $\mathcal{D}(x; \theta) = p(y = 1|x)$ as is bounded in $[0,1]$, we can use $\frac{1}{1+e^{-a}}$ where a is the output of our discriminator. Now $\mathcal{D}(x; \theta)$ is our scoring function, and we can specify a rule too, we can use BCE. This is exactly what is done in logistic regression.

Class probability estimation

Loss	Objective Function ($\mathcal{D} := \mathcal{D}(\mathbf{x}; \phi)$)
Bernoulli loss	$\pi \mathbb{E}_{p^*(\mathbf{x})}[-\log \mathcal{D}] + (1 - \pi) \mathbb{E}_{q_\theta(\mathbf{x})}[-\log(1 - \mathcal{D})]$
Brier score	$\pi \mathbb{E}_{p^*(\mathbf{x})}[(1 - \mathcal{D})^2] + (1 - \pi) \mathbb{E}_{q_\theta(\mathbf{x})}[\mathcal{D}^2]$
Exponential loss	$\pi \mathbb{E}_{p^*(\mathbf{x})} \left[\left(\frac{1 - \mathcal{D}}{\mathcal{D}} \right)^{\frac{1}{2}} \right] + (1 - \pi) \mathbb{E}_{q_\theta(\mathbf{x})} \left[\left(\frac{\mathcal{D}}{1 - \mathcal{D}} \right)^{\frac{1}{2}} \right]$
Misclassification	$\pi \mathbb{E}_{p^*(\mathbf{x})}[\mathbb{I}[\mathcal{D} \leq 0.5]] + (1 - \pi) \mathbb{E}_{q_\theta(\mathbf{x})}[\mathbb{I}[\mathcal{D} > 0.5]]$
Hinge loss	$\pi \mathbb{E}_{p^*(\mathbf{x})} \left[\max \left(0, 1 - \log \frac{\mathcal{D}}{1 - \mathcal{D}} \right) \right] + (1 - \pi) \mathbb{E}_{q_\theta(\mathbf{x})} \left[\max \left(0, 1 + \log \frac{\mathcal{D}}{1 - \mathcal{D}} \right) \right]$
Spherical	$\pi \mathbb{E}_{p^*(\mathbf{x})}[-\alpha \mathcal{D}] + (1 - \pi) \mathbb{E}_{q_\theta(\mathbf{x})}[-\alpha(1 - \mathcal{D})]; \quad \alpha = (1 - 2\mathcal{D} + 2\mathcal{D}^2)^{-1/2}$

Figure: Learning in Implicit Generative Models

Class probability estimation

BCE objective:

$$L(\phi, \theta) = E_{x \sim p(x)}[-\log(\mathcal{D}(x; \phi))] + E_{z \sim q(z)}[-\log(1 - \mathcal{D}(\mathcal{G}(z; \theta); \phi))] \quad (3)$$

Class probability estimation

BCE objective:

$$L(\phi, \theta) = E_{x \sim p(x)}[-\log(\mathcal{D}(x; \phi))] + E_{z \sim q(z)}[-\log(1 - \mathcal{D}(\mathcal{G}(z; \theta); \phi))] \quad (3)$$

We have the **ratio loss** (this is the min of NLL):

$$\min_{\phi} E_{x \sim p(x)}[-\log(\mathcal{D}(x; \phi))] + E_{z \sim q(z)}[-\log(1 - \mathcal{D}(\mathcal{G}(z; \theta); \phi))] \quad (4)$$

Class probability estimation

BCE objective:

$$L(\phi, \theta) = E_{x \sim p(x)}[-\log(\mathcal{D}(x; \phi))] + E_{z \sim q(z)}[-\log(1 - \mathcal{D}(\mathcal{G}(z; \theta); \phi))] \quad (3)$$

We have the **ratio loss** (this is the min of NLL):

$$\min_{\phi} E_{x \sim p(x)}[-\log(\mathcal{D}(x; \phi))] + E_{z \sim q(z)}[-\log(1 - \mathcal{D}(\mathcal{G}(z; \theta); \phi))] \quad (4)$$

We have the **generator loss** (min neg-class posterior probability)

$$\min_{\theta} E_{z \sim q(z)}[\log(1 - \mathcal{D}(\mathcal{G}(z; \theta); \phi))] \quad (5)$$

Class probability estimation

BCE objective:

$$L(\phi, \theta) = E_{x \sim p(x)}[-\log(\mathcal{D}(x; \phi))] + E_{z \sim q(z)}[-\log(1 - \mathcal{D}(\mathcal{G}(z; \theta); \phi))] \quad (3)$$

We have the **ratio loss** (this is the min of NLL):

$$\min_{\phi} E_{x \sim p(x)}[-\log(\mathcal{D}(x; \phi))] + E_{z \sim q(z)}[-\log(1 - \mathcal{D}(\mathcal{G}(z; \theta); \phi))] \quad (4)$$

We have the **generator loss** (min neg-class posterior probability)

$$\min_{\theta} E_{z \sim q(z)}[\log(1 - \mathcal{D}(\mathcal{G}(z; \theta); \phi))] \quad (5)$$

Eq 4 and 5 form our bi-level optimisation, where our D objective is just learning a logistic classifier, by minimizing the NLL of binary classification and our G is trying to minimise posterior probability of negative class, driving ratio to 1.[2]

ϕ -Divergence

Given two distributions, p and q , with and continuous density function w.r.t a base measure dx defined on the domain $\mathcal{X}$, the ϕ -Divergence is defined as:

ϕ -Divergence

Given two distributions, p and q , with continuous density function w.r.t a base measure dx defined on the domain $\mathcal{X}$, the ϕ -Divergence is defined as:

Theorem (ϕ -Divergence)

$$D(p||q)_f = E_{x \sim q}[f(\frac{p(x)}{q(x)})] \quad (6)$$

ϕ -Divergence

Given two distributions, p and q , with continuous density function w.r.t a base measure dx defined on the domain $\mathcal{X}$, the ϕ -Divergence is defined as:

Theorem (ϕ -Divergence)

$$D(p||q)_f = E_{x \sim q}[f(\frac{p(x)}{q(x)})] \quad (6)$$

where f is a convex function evaluation $f(1)=0$. If we select $f(u)$ to be $u \log u$, we get forward KL:

ϕ -Divergence

Given two distributions, p and q , with continuous density function w.r.t a base measure dx defined on the domain $\mathcal{X}$, the ϕ -Divergence is defined as:

Theorem (ϕ -Divergence)

$$D(p||q)_f = E_{x \sim q}[f(\frac{p(x)}{q(x)})] \quad (6)$$

where f is a convex function evaluation $f(1)=0$. If we select $f(u)$ to be $u \log u$, we get forward KL:

$$D(p||q)_{KL} = E_{x \sim p}[\log(\frac{p(x)}{q(x)})] \quad (7)$$

ϕ -Divergence

Given two distributions, p and q , with and continuous density function w.r.t a base measure dx defined on the domain $\mathcal{X}$, the ϕ -Divergence is defined as:

Theorem (ϕ -Divergence)

$$D(p||q)_f = E_{x \sim q}[f(\frac{p(x)}{q(x)})] \quad (6)$$

where f is a convex function evaluation $f(1)=0$. If we select $f(u)$ to be $u \log u$, we get forward KL:

$$D(p||q)_{KL} = E_{x \sim p}[\log(\frac{p(x)}{q(x)})] \quad (7)$$

If we select f to be $-\log u$, we get reverse KL:

$$D(p||q)_{KL} = E_{x \sim q}[\log(\frac{q(x)}{p(x)})] \quad (8)$$

ϕ -Divergence

We can't estimate $p(x)$ or $q(x)$, but it is shown that we can get a variational lower bound to it given only samples from p and q , using the convex conjugate [1]:

ϕ -Divergence

We can't estimate $p(x)$ or $q(x)$, but it is shown that we can get a variational lower bound to it given only samples from p and q , using the convex conjugate [1]:

$$f(u) = \sup_{t \in \text{dom}_{f^*}} \{tu - f^*(t)\} \quad (9)$$

ϕ -Divergence

We can't estimate $p(x)$ or $q(x)$, but it is shown that we can get a variational lower bound to it given only samples from p and q , using the convex conjugate [1]:

$$f(u) = \sup_{t \in \text{dom}_{f^*}} \{tu - f^*(t)\} \quad (9)$$

If we substitute [9] into [6]:

$$D(p||q)_f = E_{x \sim q} [\sup_{t \in \text{dom}_{f^*}} \{t \frac{p}{q} - f^*(t)\}]$$

ϕ -Divergence

We can't estimate $p(x)$ or $q(x)$, but it is shown that we can get a variational lower bound to it given only samples from p and q , using the convex conjugate [1]:

$$f(u) = \sup_{t \in \text{dom}_{f^*}} \{tu - f^*(t)\} \quad (9)$$

If we substitute [9] into [6]:

$$\begin{aligned} D(p||q)_f &= E_{x \sim q} [\sup_{t \in \text{dom}_{f^*}} \{t \frac{p}{q} - f^*(t)\}] \\ &\geq \sup_{T \in \mathcal{T}} [E_{x \sim q} (T \frac{p}{q}) - E_{x \sim q} (f^*(T))] \end{aligned}$$

ϕ -Divergence

We can't estimate $p(x)$ or $q(x)$, but it is shown that we can get a variational lower bound to it given only samples from p and q , using the convex conjugate [1]:

$$f(u) = \sup_{t \in \text{dom}_{f^*}} \{tu - f^*(t)\} \quad (9)$$

If we substitute [9] into [6]:

$$\begin{aligned} D(p||q)_f &= E_{x \sim q} [\sup_{t \in \text{dom}_{f^*}} \{t \frac{p}{q} - f^*(t)\}] \\ &\geq \sup_{T \in \mathcal{T}} [E_{x \sim q} (T \frac{p}{q}) - E_{x \sim q} (f^*(T))] \\ &= \sup_{T \in \mathcal{T}} [E_{x \sim p} (T(x)) - E_{x \sim q} (f^*(T(x)))] \end{aligned} \quad (10)$$

ϕ -Divergence

We can't estimate $p(x)$ or $q(x)$, but it is shown that we can get a variational lower bound to it given only samples from p and q , using the convex conjugate [1]:

$$f(u) = \sup_{t \in \text{dom}_{f^*}} \{tu - f^*(t)\} \quad (9)$$

If we substitute [9] into [6]:

$$\begin{aligned} D(p||q)_f &= E_{x \sim q} [\sup_{t \in \text{dom}_{f^*}} \{t \frac{p}{q} - f^*(t)\}] \\ &\geq \sup_{T \in \mathcal{T}} [E_{x \sim q} (T \frac{p}{q}) - E_{x \sim q} (f^*(T))] \\ &= \sup_{T \in \mathcal{T}} [E_{x \sim p} (T(x)) - E_{x \sim q} (f^*(T(x)))] \end{aligned} \quad (10)$$

where $\mathcal{T}$ is an arbitrary class of functions $T : \mathcal{X} \rightarrow \mathbb{R}$. This is a lower bound since we swap the Expectation and supremum, and the class of functions $\mathcal{T}$ is only a subset of all functions. We can show that optimal T , $T^* = f'(\frac{p}{q})$, so discriminator is learning a ratio, and the supremum is attained when predicted ratio = True ratio. [1].

ϕ to KL and RKL

We need to respect the domain of f^* , and in order to do that, we constrain output of variational function T , using some function g . We replace T with $g(V)$. Our loss then becomes:

ϕ to KL and RKL

We need to respect the domain of f^* , and in order to do that, we constrain output of variational function T , using some function g . We replace T with $g(V)$. Our loss then becomes:

$$L(\theta, w) = E_{x \sim p}(g(V_w(x))) - E_{x \sim q_\theta}(f^*(g(V_w(x)))) \quad (11)$$

where $V_w : \mathcal{X} \rightarrow \mathbb{R}$ is unbounded and $g_f : \mathbb{R} \rightarrow \text{dom}_{f^*}$.

ϕ to KL and RKL

We need to respect the domain of f^* , and in order to do that, we constrain output of variational function T , using some function g . We replace T with $g(V)$. Our loss then becomes:

$$L(\theta, w) = E_{x \sim p}(g(V_w(x))) - E_{x \sim q_\theta}(f^*(g(V_w(x)))) \quad (11)$$

where $V_w : \mathcal{X} \rightarrow \mathbb{R}$ is unbounded and $g_f : \mathbb{R} \rightarrow \text{dom}_{f^*}$. For the KL where the $f(u) = u \log u$, $f^*(u) = e^{(u-1)}$ and $g_f = v$, the loss term becomes:

ϕ to KL and RKL

We need to respect the domain of f^* , and in order to do that, we constrain output of variational function T , using some function g . We replace T with $g(V)$. Our loss then becomes:

$$L(\theta, w) = E_{x \sim p}(g(V_w(x))) - E_{x \sim q_\theta}(f^*(g(V_w(x)))) \quad (11)$$

where $V_w : \mathcal{X} \rightarrow \mathbb{R}$ is unbounded and $g_f : \mathbb{R} \rightarrow \text{dom}_{f^*}$. For the KL where the $f(u) = u \log u$, $f^*(u) = e^{(u-1)}$ and $g_f = v$, the loss term becomes:

$$L(\theta, w) = E_{x \sim p}[V_w(x)] - E_{x \sim q_\theta}[e^{(V_w(x)-1)}] \quad (12)$$

ϕ to KL and RKL

We need to respect the domain of f^* , and in order to do that, we constrain output of variational function T , using some function g . We replace T with $g(V)$. Our loss then becomes:

$$L(\theta, w) = E_{x \sim p}(g(V_w(x))) - E_{x \sim q_\theta}(f^*(g(V_w(x)))) \quad (11)$$

where $V_w : \mathcal{X} \rightarrow \mathbb{R}$ is unbounded and $g_f : \mathbb{R} \rightarrow \text{dom}_{f^*}$. For the KL where the $f(u) = u \log u$, $f^*(u) = e^{(u-1)}$ and $g_f = v$, the loss term becomes:

$$L(\theta, w) = E_{x \sim p}[V_w(x)] - E_{x \sim q_\theta}[e^{(V_w(x)-1)}] \quad (12)$$

For the reverse KL where the $f(u) = -\log u$, $f^*(u) = -1 - \log(-u)$ and $g_f = e^{\frac{1}{-v}}$, the loss term becomes:

ϕ to KL and RKL

We need to respect the domain of f^* , and in order to do that, we constrain output of variational function T , using some function g . We replace T with $g(V)$. Our loss then becomes:

$$L(\theta, w) = E_{x \sim p}(g(V_w(x))) - E_{x \sim q_\theta}(f^*(g(V_w(x)))) \quad (11)$$

where $V_w : \mathcal{X} \rightarrow \mathbb{R}$ is unbounded and $g_f : \mathbb{R} \rightarrow \text{dom}_{f^*}$. For the KL where the $f(u) = u \log u$, $f^*(u) = e^{(u-1)}$ and $g_f = v$, the loss term becomes:

$$L(\theta, w) = E_{x \sim p}[V_w(x)] - E_{x \sim q_\theta}[e^{(V_w(x)-1)}] \quad (12)$$

For the reverse KL where the $f(u) = -\log u$, $f^*(u) = -1 - \log(-u)$ and $g_f = e^{\frac{1}{-v}}$, the loss term becomes:

$$L(\theta, w) = E_{x \sim p}[-e^{V_w(x)}] + E_{x \sim q_\theta}[1 + (V_w(x))] \quad (13)$$

BCE

We can also show that choosing f to be the BCE, we recover the GAN objective. The domain we need to keep our variational function is the 0-1, so g_f can be sigmoid. The conjugate of the BCE is $\log(1 - e^u)$, and the loss becomes:

BCE

We can also show that choosing f to be the BCE, we recover the GAN objective. The domain we need to keep our variational function is the 0-1, so g_f can be sigmoid. The conjugate of the BCE is $\log(1 - e^u)$, and the loss becomes:

$$L(\theta, w) = E_{x \sim p}(\log(\frac{1}{1 + e^{-v}})) - E_{x \sim q_\theta}(\log(1 + e^{\log(\frac{1}{1 + e^{-v}})}))$$

BCE

We can also show that choosing f to be the BCE, we recover the GAN objective. The domain we need to keep our variational function is the 0-1, so g_f can be sigmoid. The conjugate of the BCE is $\log(1 - e^u)$, and the loss becomes:

$$\begin{aligned} L(\theta, w) &= E_{x \sim p} \left(\log \left(\frac{1}{1 + e^{-v}} \right) \right) - E_{x \sim q_\theta} \left(\log \left(1 + e^{\log \left(\frac{1}{1 + e^{-v}} \right)} \right) \right) \\ &= E_{x \sim p} \log(D(x)) + E_{x \sim q_\theta} \log(1 - D(x)) \end{aligned} \tag{14}$$

BCE

We can also show that choosing f to be the BCE, we recover the GAN objective. The domain we need to keep our variational function is the 0-1, so g_f can be sigmoid. The conjugate of the BCE is $\log(1 - e^u)$, and the loss becomes:

$$\begin{aligned} L(\theta, w) &= E_{x \sim p}(\log(\frac{1}{1 + e^{-v}})) - E_{x \sim q_\theta}(\log(1 + e^{\log(\frac{1}{1 + e^{-v}})})) \\ &= E_{x \sim p} \log(D(x)) + E_{x \sim q_\theta} \log(1 - D(x)) \end{aligned} \tag{14}$$

This demonstrates ϕ -Divergences are a generalisation to what is proposed in the original paper. We can use the frame work of variational divergence minimisation to learn a generative model. Using our general form (flip the sign for consistency):

We can also show that choosing f to be the BCE, we recover the GAN objective. The domain we need to keep our variational function is the 0-1, so g_f can be sigmoid. The conjugate of the BCE is $\log(1 - e^u)$, and the loss becomes:

$$\begin{aligned} L(\theta, w) &= E_{x \sim p}(\log(\frac{1}{1 + e^{-v}})) - E_{x \sim q_\theta}(\log(1 + e^{\log(\frac{1}{1 + e^{-v}})})) \\ &= E_{x \sim p} \log(D(x)) + E_{x \sim q_\theta} \log(1 - D(x)) \end{aligned} \tag{14}$$

This demonstrates ϕ -Divergences are a generalisation to what is proposed in the original paper. We can use the frame work of variational divergence minimisation to learn a generative model. Using our general form (flip the sign for consistency):

$$L(\theta, w) = E_{x \sim p}(-g(V_w(x))) + E_{x \sim q_\theta}(f^*(g(V_w(x)))) \tag{15}$$

We can also show that choosing f to be the BCE, we recover the GAN objective. The domain we need to keep our variational function is the 0-1, so g_f can be sigmoid. The conjugate of the BCE is $\log(1 - e^u)$, and the loss becomes:

$$\begin{aligned} L(\theta, w) &= E_{x \sim p} \left(\log \left(\frac{1}{1 + e^{-v}} \right) \right) - E_{x \sim q_\theta} \left(\log \left(1 + e^{\log \left(\frac{1}{1 + e^{-v}} \right)} \right) \right) \\ &= E_{x \sim p} \log(D(x)) + E_{x \sim q_\theta} \log(1 - D(x)) \end{aligned} \quad (14)$$

This demonstrates ϕ -Divergences are a generalisation to what is proposed in the original paper. We can use the frame work of variational divergence minimisation to learn a generative model. Using our general form (flip the sign for consistency):

$$L(\theta, w) = E_{x \sim p} (-g(V_w(x))) + E_{x \sim q_\theta} (f^*(g(V_w(x)))) \quad (15)$$

We want to min L w.r.t w , since we want to max the variational lower bound to close the gap; we want to minimise $E_{x \sim q_\theta} (-f^*(g(V_w(x))))$, the generative loss w.r.t θ to drive q_θ to p

ϕ -Divergence

Name	$D_f(P\ Q)$	Generator $f(u)$	$T^*(x)$
Total variation	$\frac{1}{2} \int p(x) - q(x) \, dx$	$\frac{1}{2} u - 1 $	$\frac{1}{2} \text{sign}(\frac{p(x)}{q(x)} - 1)$
Kullback-Leibler	$\int p(x) \log \frac{p(x)}{q(x)} \, dx$	$u \log u$	$1 + \log \frac{p(x)}{q(x)}$
Reverse Kullback-Leibler	$\int q(x) \log \frac{q(x)}{p(x)} \, dx$	$-\log u$	$-\frac{q(x)}{p(x)}$
Pearson χ^2	$\int \frac{(q(x) - p(x))^2}{p(x)} \, dx$	$(u - 1)^2$	$2(\frac{p(x)}{q(x)} - 1)$
Neyman χ^2	$\int \frac{(p(x) - q(x))^2}{q(x)} \, dx$	$\frac{(1-u)^2}{u}$	$1 - [\frac{q(x)}{p(x)}]^2$
Squared Hellinger	$\int \left(\sqrt{p(x)} - \sqrt{q(x)} \right)^2 \, dx$	$(\sqrt{u} - 1)^2$	$(\sqrt{\frac{p(x)}{q(x)}} - 1) \cdot \sqrt{\frac{q(x)}{p(x)}}$
Jeffrey	$\int (p(x) - q(x)) \log \left(\frac{p(x)}{q(x)} \right) \, dx$	$(u - 1) \log u$	$1 + \log \frac{p(x)}{q(x)} - \frac{q(x)}{p(x)}$
Jensen-Shannon	$\frac{1}{2} \int p(x) \log \frac{2p(x)}{p(x)+q(x)} + q(x) \log \frac{2q(x)}{p(x)+q(x)} \, dx$	$-(u + 1) \log \frac{1+u}{2} + u \log u$	$\log \frac{2p(x)}{p(x)+q(x)}$
Jensen-Shannon-weighted	$\int p(x) \pi \log \frac{p(x)}{\pi p(x) + (1-\pi)q(x)} + (1-\pi)q(x) \log \frac{q(x)}{\pi p(x) + (1-\pi)q(x)} \, dx$	$\pi u \log u - (1 - \pi + \pi u) \log(1 - \pi + \pi u)$	$\pi \log \frac{p(x)}{(1-\pi)q(x) + \pi p(x)}$
GAN	$\int p(x) \log \frac{2p(x)}{p(x)+q(x)} + q(x) \log \frac{2q(x)}{p(x)+q(x)} \, dx - \log(4)$	$u \log u - (u + 1) \log(u + 1)$	$\log \frac{p(x)}{p(x)+q(x)}$
α -divergence ($\alpha \notin \{0, 1\}$)	$\frac{1}{\alpha(\alpha-1)} \int \left(p(x) \left[\left(\frac{q(x)}{p(x)} \right)^\alpha - 1 \right] - \alpha(q(x) - p(x)) \right) \, dx$	$\frac{1}{\alpha(\alpha-1)} (u^\alpha - 1 - \alpha(u - 1))$	$\frac{1}{\alpha-1} \left[\left(\frac{p(x)}{q(x)} \right)^{\alpha-1} - 1 \right]$

Figure: f-GAN: Training Generative Neural Samplers using Variational Divergence Minimization

ϕ -Divergence

Name	Output activation g_f	dom_{f^*}	Conjugate $f^*(t)$	$f'(1)$
Total variation	$\frac{1}{2} \tanh(v)$	$-\frac{1}{2} \leq t \leq \frac{1}{2}$	t	0
Kullback-Leibler (KL)	v	$\mathbb{R}$	$\exp(t - 1)$	1
Reverse KL	$-\exp(v)$	$\mathbb{R}_-$	$-1 - \log(-t)$	-1
Pearson χ^2	v	$\mathbb{R}$	$\frac{1}{4}t^2 + t$	0
Neyman χ^2	$1 - \exp(v)$	$t < 1$	$2 - 2\sqrt{1-t}$	0
Squared Hellinger	$1 - \exp(v)$	$t < 1$	$\frac{t}{1-t}$	0
Jeffrey	v	$\mathbb{R}$	$W(e^{1-t}) + \frac{1}{W(e^{1-t})} + t - 2$	0
Jensen-Shannon	$\log(2) - \log(1 + \exp(-v))$	$t < \log(2)$	$-\log(2 - \exp(t))$	0
Jensen-Shannon-weighted	$-\pi \log \pi - \log(1 + \exp(-v))$	$t < -\pi \log \pi$	$(1 - \pi) \log \frac{1-\pi}{1-\pi e^{t/\pi}}$	0
GAN	$-\log(1 + \exp(-v))$	$\mathbb{R}_-$	$-\log(1 - \exp(t))$	$-\log(2)$
α -div. ($\alpha < 1, \alpha \neq 0$)	$\frac{1}{1-\alpha} - \log(1 + \exp(-v))$	$t < \frac{1}{1-\alpha}$	$\frac{1}{\alpha} (t(\alpha - 1) + 1)^{\frac{\alpha}{\alpha-1}} - \frac{1}{\alpha}$	0
α -div. ($\alpha > 1$)	v	$\mathbb{R}$	$\frac{1}{\alpha} (t(\alpha - 1) + 1)^{\frac{\alpha}{\alpha-1}} - \frac{1}{\alpha}$	0

Figure: f-GAN: Training Generative Neural Samplers using Variational Divergence Minimization

$$\text{BCE} = \text{JSD}$$

What value do we want the probabilistic classifier to output? We know that:

BCE = JSD

What value do we want the probabilistic classifier to output? We know that:

$$\frac{p(x)}{q(x)}$$

What value do we want the probabilistic classifier to output? We know that:

$$\frac{p(x)}{q(x)} = \frac{\mathcal{D}(x; \phi)}{1 - \mathcal{D}(x; \phi)} \quad (16)$$

BCE = JSD

What value do we want the probabilistic classifier to output? We know that:

$$\frac{p(x)}{q(x)} = \frac{\mathcal{D}(x; \phi)}{1 - \mathcal{D}(x; \phi)} \quad (16)$$

Some algebra gets us to:

$$\mathcal{D}(x; \phi) = \frac{p}{p + q} \quad (17)$$

and when our generator is optimal, $q = p$, and $D = D^* = \frac{1}{2}$

$$\text{BCE} = \text{JSD}$$

If the generator is optimal and the classifier is optimal, what does our BCE loss become?

BCE = JSD

If the generator is optimal and the classifier is optimal, what does our BCE loss become?
Let us sub optimal values $q(x) = p(x)$ and $D^* = D = \frac{1}{2}$ into 6:

BCE = JSD

If the generator is optimal and the classifier is optimal, what does our BCE loss become?

Let us sub optimal values $q(x) = p(x)$ and $D^* = D = \frac{1}{2}$ into 6:

$$\max_D L = E_{x \sim p(x)}[\log(\frac{1}{2})] + E_{x \sim q(x)}[\log(1 - \frac{1}{2})]$$

BCE = JSD

If the generator is optimal and the classifier is optimal, what does our BCE loss become?

Let us sub optimal values $q(x) = p(x)$ and $D^* = D = \frac{1}{2}$ into 6:

$$\begin{aligned}\max_D L &= E_{x \sim p(x)}[\log(\frac{1}{2})] + E_{x \sim q(x)}[\log(1 - \frac{1}{2})] \\ &= -2\log 2\end{aligned}\tag{18}$$

BCE = JSD

If the generator is optimal and the classifier is optimal, what does our BCE loss become?

Let us sub optimal values $q(x) = p(x)$ and $D^* = D = \frac{1}{2}$ into 6:

$$\begin{aligned}\max_D L &= E_{x \sim p(x)}[\log(\frac{1}{2})] + E_{x \sim q(x)}[\log(1 - \frac{1}{2})] \\ &= -2\log 2\end{aligned}\tag{18}$$

The Jensen-Shannon Divergence between $p(x)$ and $q(x)$ is defined as:

BCE = JSD

If the generator is optimal and the classifier is optimal, what does our BCE loss become?

Let us sub optimal values $q(x) = p(x)$ and $D^* = D = \frac{1}{2}$ into 6:

$$\begin{aligned}\max_D L &= E_{x \sim p(x)}[\log(\frac{1}{2})] + E_{x \sim q(x)}[\log(1 - \frac{1}{2})] \\ &= -2\log 2\end{aligned}\tag{18}$$

The Jensen-Shannon Divergence between $p(x)$ and $q(x)$ is defined as:

Theorem (Jensen Shannon Divergence)

$$D_{JS}(p||q) = \frac{1}{2}D_{KL}(p||\frac{p+q}{2}) + \frac{1}{2}D_{KL}(q||\frac{p+q}{2})$$

BCE = JSD

If the generator is optimal and the classifier is optimal, what does our BCE loss become?

Let us sub optimal values $q(x) = p(x)$ and $D_* = D = \frac{1}{2}$ into 6:

$$\begin{aligned}\max_D L &= E_{x \sim p(x)}[\log(\frac{1}{2})] + E_{x \sim q(x)}[\log(1 - \frac{1}{2})] \\ &= -2\log 2\end{aligned}\tag{18}$$

The Jensen-Shannon Divergence between $p(x)$ and $q(x)$ is defined as:

Theorem (Jensen Shannon Divergence)

$$\begin{aligned}D_{JS}(p||q) &= \frac{1}{2}D_{KL}(p||\frac{p+q}{2}) + \frac{1}{2}D_{KL}(q||\frac{p+q}{2}) \\ &= \frac{1}{2}[\log 2 + E_{x \sim p}[\log(\frac{p}{p+q})]] + \log 2 + E_{x \sim q}[\log(\frac{q}{p+q})]\end{aligned}$$

BCE = JSD

If the generator is optimal and the classifier is optimal, what does our BCE loss become?

Let us sub optimal values $q(x) = p(x)$ and $D^* = D = \frac{1}{2}$ into 6:

$$\begin{aligned}\max_D L &= E_{x \sim p(x)}[\log(\frac{1}{2})] + E_{x \sim q(x)}[\log(1 - \frac{1}{2})] \\ &= -2\log 2\end{aligned}\tag{18}$$

The Jensen-Shannon Divergence between $p(x)$ and $q(x)$ is defined as:

Theorem (Jensen Shannon Divergence)

$$\begin{aligned}D_{JS}(p||q) &= \frac{1}{2}D_{KL}(p||\frac{p+q}{2}) + \frac{1}{2}D_{KL}(q||\frac{p+q}{2}) \\ &= \frac{1}{2}[\log 2 + E_{x \sim p}[\log(\frac{p}{p+q})]] + \log 2 + E_{x \sim q}[\log(\frac{q}{p+q})]] \\ &= \frac{1}{2}[\log 4 + L(G, D)]\end{aligned}\tag{19}$$

BCE = JSD

Therefore, we are minimising the JSD. When JSD falls to zero by the generator becoming optimal, we end up at $-2\log 2$, the optimum value of BCE.

BCE = JSD

Therefore, we are minimising the JSD. When JSD falls to zero by the generator becoming optimal, we end up at $-2\log 2$, the optimum value of BCE.

$$L(G, D) = 2D_{JS}(p||q) - 2\log 2$$

BCE = JSD

Therefore, we are minimising the JSD. When JSD falls to zero by the generator becoming optimal, we end up at $-2\log 2$, the optimum value of BCE.

$$\begin{aligned}L(G, D) &= 2D_{JS}(p||q) - 2\log 2 \\L(G^*, D^*) &= -2\log 2\end{aligned}\tag{20}$$

Bergman divergences

Can we directly try to match the density ratios? If we define the true ratio as $r^*(x) = \frac{p(x)}{q_\theta(x)}$ and its approximation as r_ϕ , we can directly match the ratio and try to drive our learning.

Bergman divergences

Can we directly try to match the density ratios? If we define the true ratio as $r^*(x) = \frac{p(x)}{q_\theta(x)}$ and its approximation as r_ϕ , we can directly match the ratio and try to drive our learning. A Bergman divergence is defined as:

Theorem (Bergman Divergence)

$$\mathcal{B}_f(r^*(x)||r_\phi(x)) = E_{q_\theta} \left(f(r^*(x)) - f(r_\phi(x)) - f'(r_\phi(x))[r^*(x) - r_\phi(x)] \right)$$

Bergman divergences

Can we directly try to match the density ratios? If we define the true ratio as $r^*(x) = \frac{p(x)}{q_\theta(x)}$ and its approximation as r_ϕ , we can directly match the ratio and try to drive our learning. A Bergman divergence is defined as:

Theorem (Bergman Divergence)

$$\begin{aligned}\mathcal{B}_f(r^*(x)||r_\phi(x)) &= E_{q_\theta} \left(f(r^*(x)) - f(r_\phi(x)) - f'(r_\phi(x))[r^*(x) - r_\phi(x)] \right) \\ &= E_{q_\theta} f(r^*(x)) - E_{q_\theta} f(r_\phi(x)) + E_{q_\theta} f'(r_\phi(x))r_\phi(x) - E_{q_\theta} f'(r_\phi(x))r^*(x)\end{aligned}$$

Bergman divergences

Can we directly try to match the density ratios? If we define the true ratio as $r^*(x) = \frac{p(x)}{q_\theta(x)}$ and its approximation as r_ϕ , we can directly match the ratio and try to drive our learning. A Bergman divergence is defined as:

Theorem (Bergman Divergence)

$$\begin{aligned}\mathcal{B}_f(r^*(x)||r_\phi(x)) &= E_{q_\theta} \left(f(r^*(x)) - f(r_\phi(x)) - f'(r_\phi(x))[r^*(x) - r_\phi(x)] \right) \\ &= E_{q_\theta} f(r^*(x)) - E_{q_\theta} f(r_\phi(x)) + E_{q_\theta} f'(r_\phi(x))r_\phi(x) - E_{q_\theta} f'(r_\phi(x))r^*(x) \quad (21) \\ &= D_f[p(x)||q_{\theta(x)}] + E_{q_\theta}[f'(r_\phi(x))r_\phi(x) - f(r_\phi(x))] - E_p f'(r_\phi(x))\end{aligned}$$

Bergman divergences

Can we directly try to match the density ratios? If we define the true ratio as $r^*(x) = \frac{p(x)}{q_\theta(x)}$ and its approximation as r_ϕ , we can directly match the ratio and try to drive our learning. A Bergman divergence is defined as:

Theorem (Bergman Divergence)

$$\begin{aligned}\mathcal{B}_f(r^*(x)||r_\phi(x)) &= E_{q_\theta} \left(f(r^*(x)) - f(r_\phi(x)) - f'(r_\phi(x))[r^*(x) - r_\phi(x)] \right) \\ &= E_{q_\theta} f(r^*(x)) - E_{q_\theta} f(r_\phi(x)) + E_{q_\theta} f'(r_\phi(x))r_\phi(x) - E_{q_\theta} f'(r_\phi(x))r^*(x) \quad (21) \\ &= D_f[p(x)||q_\theta(x)] + E_{q_\theta}[f'(r_\phi(x))r_\phi(x) - f(r_\phi(x))] - E_p f'(r_\phi(x))\end{aligned}$$

Extracting all the terms related to r_ϕ gives us the ratio loss we need to minimise, which will minimise the distance between r^* and r_ϕ where $r_\phi : \mathcal{X} \rightarrow \mathbb{R}$ (this is our discriminator loss):

Bergman divergences

Can we directly try to match the density ratios? If we define the true ratio as $r^*(x) = \frac{p(x)}{q_\theta(x)}$ and its approximation as r_ϕ , we can directly match the ratio and try to drive our learning. A Bergman divergence is defined as:

Theorem (Bergman Divergence)

$$\begin{aligned}\mathcal{B}_f(r^*(x)||r_\phi(x)) &= E_{q_\theta} \left(f(r^*(x)) - f(r_\phi(x)) - f'(r_\phi(x))[r^*(x) - r_\phi(x)] \right) \\ &= E_{q_\theta} f(r^*(x)) - E_{q_\theta} f(r_\phi(x)) + E_{q_\theta} f'(r_\phi(x))r_\phi(x) - E_{q_\theta} f'(r_\phi(x))r^*(x) \quad (21) \\ &= D_f[p(x)||q_\theta(x)] + E_{q_\theta}[f'(r_\phi(x))r_\phi(x) - f(r_\phi(x))] - E_p f'(r_\phi(x))\end{aligned}$$

Extracting all the terms related to r_ϕ gives us the ratio loss we need to minimise, which will minimise the distance between r^* and r_ϕ where $r_\phi : \mathcal{X} \rightarrow \mathbb{R}$ (this is our discriminator loss):

$$L(r_\phi) = E_{q_\theta}[f'(r_\phi(x))r_\phi(x) - f(r_\phi(x))] - E_p f'(r_\phi(x)) \quad (22)$$

Bregman divergences

For the generative loss, we gather the terms dependant on q_θ from 21 and get:

Bregman divergences

For the generative loss, we gather the terms dependant on q_θ from 21 and get:

$$L(q_\theta) = D_f[p(x)||q_{\theta(x)}] + E_{q_\theta}[f'(r_\phi(x))r_\phi(x) - f(r_\phi(x))] \quad (23)$$

Bregman divergences

For the generative loss, we gather the terms dependant on q_θ from 21 and get:

$$L(q_\theta) = D_f[p(x)||q_{\theta(x)}] + E_{q_\theta}[f'(r_\phi(x))r_\phi(x) - f(r_\phi(x))] \quad (23)$$

however, we can't use this as a generative loss, since we have to evaluate q_θ which is not known. An approximation suggested by [3] is to use $p = r_\phi q_\theta$, which simplifies the D_f term to:

Bregman divergences

For the generative loss, we gather the terms dependant on q_θ from 21 and get:

$$L(q_\theta) = D_f[p(x)||q_{\theta(x)}] + E_{q_\theta}[f'(r_\phi(x))r_\phi(x) - f(r_\phi(x))] \quad (23)$$

however, we can't use this as a generative loss, since we have to evaluate q_θ which is not known. An approximation suggested by [3] is to use $p = r_\phi q_\theta$, which simplifies the D_f term to:

$$D_f[p(x)||q_{\theta(x)}] =$$

Bregman divergences

For the generative loss, we gather the terms dependant on q_θ from 21 and get:

$$L(q_\theta) = D_f[p(x)||q_{\theta(x)}] + E_{q_\theta}[f'(r_\phi(x))r_\phi(x) - f(r_\phi(x))] \quad (23)$$

however, we can't use this as a generative loss, since we have to evaluate q_θ which is not known. An approximation suggested by [3] is to use $p = r_\phi q_\theta$, which simplifies the D_f term to:

$$D_f[p(x)||q_{\theta(x)}] = E_{q_\theta}[f(\frac{p}{q_\theta(x)})] \approx$$

Bregman divergences

For the generative loss, we gather the terms dependant on q_θ from 21 and get:

$$L(q_\theta) = D_f[p(x)||q_{\theta(x)}] + E_{q_\theta}[f'(r_\phi(x))r_\phi(x) - f(r_\phi(x))] \quad (23)$$

however, we can't use this as a generative loss, since we have to evaluate q_θ which is not known. An approximation suggested by [3] is to use $p = r_\phi q_\theta$, which simplifies the D_f term to:

$$D_f[p(x)||q_{\theta(x)}] = E_{q_\theta}[f(\frac{p}{q_{\theta(x)}})] \approx E_{q_\theta}[f(\frac{r_\phi q_\theta}{q_{\theta(x)}})] =$$

Bregman divergences

For the generative loss, we gather the terms dependant on q_θ from 21 and get:

$$L(q_\theta) = D_f[p(x)||q_{\theta(x)}] + E_{q_\theta}[f'(r_\phi(x))r_\phi(x) - f(r_\phi(x))] \quad (23)$$

however, we can't use this as a generative loss, since we have to evaluate q_θ which is not known. An approximation suggested by [3] is to use $p = r_\phi q_\theta$, which simplifies the D_f term to:

$$D_f[p(x)||q_{\theta(x)}] = E_{q_\theta}[f(\frac{p}{q_\theta(x)})] \approx E_{q_\theta}[f(\frac{r_\phi q_\theta}{q_\theta(x)})] = E_{q_\theta}[f[r_\phi(x)]] \quad (24)$$

Bregman divergences

For the generative loss, we gather the terms dependant on q_θ from 21 and get:

$$L(q_\theta) = D_f[p(x)||q_{\theta(x)}] + E_{q_\theta}[f'(r_\phi(x))r_\phi(x) - f(r_\phi(x))] \quad (23)$$

however, we can't use this as a generative loss, since we have to evaluate q_θ which is not known. An approximation suggested by [3] is to use $p = r_\phi q_\theta$, which simplifies the D_f term to:

$$D_f[p(x)||q_{\theta(x)}] = E_{q_\theta}[f(\frac{p}{q_{\theta(x)}})] \approx E_{q_\theta}[f(\frac{r_\phi q_\theta}{q_{\theta(x)}})] = E_{q_\theta}[f[r_\phi(x)]] \quad (24)$$

This simplifies our generative loss $L(q_\theta)$ to:

$$L(q_\theta) = E_{q_\theta}[f'(r_\phi(x))r_\phi(x)] \quad (25)$$

Bergman divergence

Optimisation using Bergman loss. Our discriminator now predicts the ratio r_ϕ [3].

Algorithm 1: b-GAN

for number of training iterations **do**

sample $\hat{X} = \{x_1, \dots, x_B\}$ from $p(x)$ and $\hat{Z} = \{z_1, \dots, z_B\}$ from an uniform distribution.

D-step: Update θ_D :

$$\theta_D^{t+1} = \theta_D^t - \nabla_{\theta_D} \left(\frac{1}{B} \sum_{i=1}^B f'(r_{\theta_D}(G(z_i))) r_{\theta_D}(G(z_i)) - f(r_{\theta_D}(G(z_i))) - f'(r_{\theta_D}(x_i)) \right).$$

G-step: Update θ_G :

$$\theta_G^{t+1} = \theta_G^t - \nabla_{\theta_G} \left(\frac{1}{B} \sum_{i=1}^B f(r(G_{\theta_G}(z_i))) \right).$$

end for

Figure: Generative Adversarial Nets from a Density Ratio Estimation Perspective

We choose f similar to D_f methods. We can express all methods, class probability, f -divergence and moment matching as Bergman divergences and show they are all closely related.

Integral probability metrics

IPM is defined as:

Theorem (Integral Probability metric)

$$IPM(p, q) = \sup_{f \in \mathcal{F}} |E_{x \sim p} f(x) - E_{x \sim q} f(x)| \quad (26)$$

Integral probability metrics

IPM is defined as:

Theorem (Integral Probability metric)

$$IPM(p, q) = \sup_{f \in \mathcal{F}} |E_{x \sim p} f(x) - E_{x \sim q} f(x)| \quad (26)$$

Depending on the conditions we enforce on f , we can get different metrics [4]:

Integral probability metrics

IPM is defined as:

Theorem (Integral Probability metric)

$$IPM(p, q) = \sup_{f \in \mathcal{F}} |E_{x \sim p} f(x) - E_{x \sim q_\theta} f(x)| \quad (26)$$

Depending on the conditions we enforce on f , we can get different metrics [4]:

- $\mathcal{F} = \{f: f \text{ is 1-Lipschitz}\}$: Wasserstein GAN [5]

Integral probability metrics

IPM is defined as:

Theorem (Integral Probability metric)

$$IPM(p, q) = \sup_{f \in \mathcal{F}} |E_{x \sim p} f(x) - E_{x \sim q_\theta} f(x)| \quad (26)$$

Depending on the conditions we enforce on f , we can get different metrics [4]:

- $\mathcal{F} = \{f: f \text{ is 1-Lipschitz}\}$: Wasserstein GAN [5]
- $\mathcal{F} = \{f : \|f\|_\infty \leq 1\}$: TV used in EBGAN [6]

Integral probability metrics

IPM is defined as:

Theorem (Integral Probability metric)

$$IPM(p, q) = \sup_{f \in \mathcal{F}} |E_{x \sim p} f(x) - E_{x \sim q_\theta} f(x)| \quad (26)$$

Depending on the conditions we enforce on f , we can get different metrics [4]:

- $\mathcal{F} = \{f: f \text{ is 1-Lipschitz}\}$: Wasserstein GAN [5]
- $\mathcal{F} = \{f : \|f\|_\infty \leq 1\}$: TV used in EBGAN [6]
- $\mathcal{F} = \{f : \|f\|_H \leq 1\}$: f is RKHS, MMD, GMMN [7]

Integral probability metrics

IPM is defined as:

Theorem (Integral Probability metric)

$$IPM(p, q) = \sup_{f \in \mathcal{F}} |E_{x \sim p} f(x) - E_{x \sim q_\theta} f(x)| \quad (26)$$

Depending on the conditions we enforce on f , we can get different metrics [4]:

- $\mathcal{F} = \{f: f \text{ is 1-Lipschitz}\}$: Wasserstein GAN [5]
- $\mathcal{F} = \{f : \|f\|_\infty \leq 1\}$: TV used in EBGAN [6]
- $\mathcal{F} = \{f : \|f\|_H \leq 1\}$: f is RKHS, MMD, GMMN [7]

Final words

- BINARY CLASSIFICATION == ϕ -Divergence == IPM == $\mathcal{B}_f$
- BCE learns $p/(p+q)$, f-div learns $f'(p/q)$, and Bergman learns p/q , so these methods all learn ratios and max them over G
- The equivalence to divergences and metrics is only possible under optimal classifiers, assuming our discriminator is optimal (which is never the case)
- Since we don't actually have a perfect disc, we have a learnt divergence/lower bound
- We optimise some variations approximation of the true divergence
- Issue here is with the gradients, supports are disjoint, KL/Reverse KL are ∞ JS is constant, but this is okay since we are optimising lower bounds to them

References



f-GAN: Training Generative Neural Samplers using Variational Divergence Minimization



Learning in Implicit Generative Models



Generative Adversarial Nets from a Density Ratio Estimation Perspective



On Integral Probability Metrics, ϕ -Divergences and Binary Classification



Wasserstein Generative Adversarial Networks



ENERGY-BASED GENERATIVE ADVERSARIAL NETWORKS



GENERATIVE MOMENT MATCHING NETWORKS

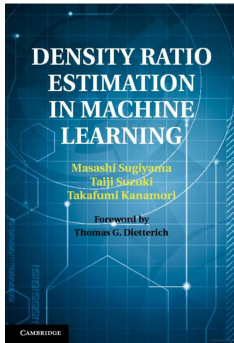


Figure: Density Ratio Estimation in Machine Learning

The End