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

Ruyue Han

August 31, 2018

1. Introduction

Probabilistic generative models describe a probability distribution over a given domain $\mathcal{X}$.Given a generative model Q,we are generally interested in performing one or multiple of the following operations:

- *Sampling.*Produce a sample from Q.
- *Estimation.*Given a set of samples $\{x_1, x_2, \ldots, x_n\}$ from an unknown true distribution P , find Q that best describes the true distribution.
- *Point-wise likelihood evaluation.*Given a sample x, evaluate the likelihood Q(x).

GANs proposed by [2] are an expressive class of generative models that allow exact sampling and approximate estimation.In the original GAN paper the authors show that it is possible to estimate neural samplers by approximate minimization of the symmetric Jensen-Shannon divergence,

$$D_{JS}(P||Q) = \frac{1}{2}D_{KL}(P||\frac{P+Q}{2}) + \frac{1}{2}D_{KL}(Q||\frac{P+Q}{2}) \tag{1}$$

where D_{KL} denotes the Kullback-Leibler divergence.The key technique used in the GAN training is that of introducing a second "discriminator" neural networks which is optimized simultaneously.Because $D_{JS}(P||Q)$ is a proper divergence measure between distributions this implies that the true distribution P can be approximated well in case where samples are enough.

In this work author shows that the principle of GANs is more general and extend the variational divergence estimation framework proposed by Nguyen et al. [5] to recover the GAN training objective and generalize it to arbitrary *f-divergences.*

2. Method

2.1. The f-divergence Family

A large class of different divergences are the so called *f* -divergences [1, 4].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 $\mathcal{X}$, the define the *f* -divergence

$$D_f(P||Q) = \int_{\mathcal{X}} q(x) f(\frac{p(x)}{q(x)}) dx \tag{2}$$

where the generator function f is a convex, lower-semicontinuous function satisfying f (1) = 0.Different choices of f recover popular divergences as special cases in function(2).Author illustrates common choices in Fig (1).

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

Figure 1. List of f -divergences D f (P kQ) together with generator functions and the optimal variational functions.

2.2. Variational Estimation of f-divergences

Every convex, lower-semicontinuous function f has a brother function,*convex conjugate* function f^*,also known as Fenchel conjugate [3].This function is defined as

$$f^*(t)=\sup_{x\in dom_f}\{xt-f(x)\} \tag{3}$$

you can understand this function like this

$$f^*(t)=\max_{x\in dom_f}\{xt-f(x)\} \tag{4}$$

$xt-f(x)$ is a group linear equation,like $at-b,(a,b\in\mathbb{R})$.The function f^* is again convex and lower-semicontinuous and $f^{**}=f$ .So function f can be described as

$$f(x)=\sup_{t\in dom_{f^*}}\{tx-f^*(t)\} \tag{5}$$

we use $\frac{p(x)}{q(x)}$ to replace the x in equation (5),then we can get function (6)

$$f(\frac{p(x)}{q(x)})=\sup_{t\in dom_{f^*}}\{t\frac{p(x)}{q(x)}-f^*(t)\} \tag{6}$$

and then put function (6) into function (2),we get function

$$D_f(P||Q)=\int_{\mathcal{X}}q(x)\sup_{t\in dom_{f^*}}\{t\frac{p(x)}{q(x)}-f^*(t)\}dx \tag{7}$$

if we define a funtion D,the input is x and the out put is t,the function is $t=D(x),(x\in\mathcal{X})$,this may casuse a question that is the range of t maybe smaller than befor $t\in dom_{f^*}$.so putting function D into (7) we get equation(8)

$$D_f(P||Q)\geqslant\sup_D\int_{\mathcal{X}}p(x)D(x)-q(x)f^*(D(x))dx=\sup_D\{\mathbb{E}_{x\backsim P}\left[D(x)\right]-E_{x\backsim Q}\left[f^*(D(x))\right]\} \tag{8}$$

If we choose $f(u)=u\log u-(u+1)\log(u+1)$,which function is choosed by GAN,and then get f's brother convex conjugate $f^*(t)=-\log(1-e^t)$ from Fig 2,repalcing $D(x)$ by $\log(D(x))$.Then we can get

$$D_f(P||Q)\geqslant\sup_D\int_{\mathcal{X}}p(x)\log(D(x))+q(x)log(1-D(x))dx=\sup_D\{\underline{\mathbb{E}_{x\backsim P}\left[D(x)\right]+E_{x\backsim Q}\left[log(1-D(x))\right]}\} \tag{9}$$

If we define $V(Q,D)=\mathbb{E}_{x\backsim P}\left[D(x)\right]+E_{x\backsim Q}\left[log(1-D(x))\right]$.we can see this function is the loss function of GAN,and also know why have to minmax function $V(G,D)$.

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 2. List of f -divergences D f (P kQ) together with generator functions and the optimal variational functions.

References

[1] I. Csiszr and P. C. Shields. Information theory and statistics: A tutorial. *Foundations and Trends in Communications and Information Theory*, 1:417–528, 2004. 1

[2] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. *In NIPS*, pages 2672–2680, 2014. 1

[3] J. B. Hiriart-Urruty and C. Lemarchal. Fundamentals of convex analysis. *Springer*, 2012. 2

[4] F. Liese and I. Vajda. On divergences and informations in statistics and information theory. *Information Theory, IEEE*, 52(10):4394–4412, 2006. 1

[5] X. Nguyen, M. J. Wainwright, and M. I. Jordan. Estimating divergence functionals and the likelihood ratio by convex risk minimization. *Information Theory, IEEE*, 56(11):5847–5861, 2010. 1