

Variational Inference: A Unified Framework of Generative Models and Some Revelations

Jianlin Su

School of Mathematics, Sun Yat-sen University

Abstract

We reinterpreting the variational inference in a new perspective. Via this way, we can easily prove that EM algorithm, VAE, GAN, AAE, ALI(BiGAN) are all special cases of variational inference. The proof also reveals the loss of standard GAN is incomplete and it explains why we need to train GAN cautiously. From that, we find out a regularization term to improve stability of GAN training.

In recent years, deep generative models, especially Generative Adversarial Networks (GANs) (Goodfellow et al. 2014), have achieved impressive success. We can find dozens of different variants of GANs. However, most of them are achieved empirically, rarely have completely theoretical guidance.

This paper aims to establish a unified framework of these various generative models by variational inference. Firstly, we present a new formulation of variational inference, which can derive EM algorithm and Variational Autoencoders (VAEs) (Kingma and Welling 2013) in only several lines. Then we re-derive GAN by our new variational inference and find the loss of standard GAN is not complete, which lacks of a regularization term. Without this term, we need to adjust hyperparameters carefully to make GAN converge.

In fact, the original purpose of our work is to incorporate GAN into the variational inference framework. It seems we are successful to accomplish it. The new regularization term is an unexpected result. Fortunately, we are glad to see it actually work in our experiment.

Variational Inference

Suppose x is an explicit variable, z is a latent variable and $\tilde{p}(x)$ is evidence distribution of x . We let

$$q_\theta(x) = \int q_\theta(x, z) dz \quad (1)$$

and we hope $q_\theta(x)$ will be a good approximation of $\tilde{p}(x)$. In general cases, we want to maximize log likelihood function

$$\theta = \arg \max_{\theta} \int \tilde{p}(x) \log q(x) dx \quad (2)$$

which is equivalent to minimizing $KL(\tilde{p}(x)\|q(x))$:

$$KL(\tilde{p}(x)\|q(x)) = \int \tilde{p}(x) \log \frac{\tilde{p}(x)}{q(x)} dx \quad (3)$$

But if we can not calculate the integral analytically, we can not maximize log likelihood or minimize KL-divergence directly.

The variational inference changes objective function: rather minimizing KL-divergence of marginal distributions $KL(\tilde{p}(x)\|q(x))$, we can minimize the KL-divergence of joint distribution $KL(p(x, z)\|q(x, z))$ or $KL(q(x, z)\|p(x, z))$. We have

$$\begin{aligned} & KL(p(x, z)\|q(x, z)) \\ &= KL(p(x)\|q(x)) + \int p(x) KL(p(z|x)\|q(z|x)) dx \quad (4) \\ &\geq KL(p(x)\|q(x)), \end{aligned}$$

which suggests $KL(p(x, z)\|q(x, z))$ is an upper bound of $KL(\tilde{p}(x)\|q(x))$. In many cases, joint KL-divergence easier to calculate than marginal KL-divergence. Therefore, variational inference provides a computable solution. If it works, we have $q(x, z) \rightarrow p(x, z)$ which means $q(x) = \int q(x, z) dz \rightarrow \int p(x, z) dz = \tilde{p}(x)$. Namely, $q(x)$ becomes an approximation of the real distribution $\tilde{p}(x)$.

VAE and EM algorithm

Due to our new insight of variational inference, VAE and EM algorithm can be derived in a very simple way.

In VAE, we let $q(x, z) = q(x|z)q(z)$, $p(x, z) = \tilde{p}(x)p(z|x)$, while $q(x|z)$, $p(z|x)$ are Gaussian distributions with unknown parameters and $q(z)$ is standard Gaussian distribution. The loss we need to minimize is

$$\begin{aligned} & KL(p(x, z)\|q(x, z)) \\ &= \iint \tilde{p}(x)p(z|x) \log \frac{\tilde{p}(x)p(z|x)}{q(x|z)q(z)} dx dz \quad (5) \end{aligned}$$

while $\log \tilde{p}(x)$ does not contain any parameters, it does not change final result. So loss can be transed into

$$\mathbb{E}_{x \sim \tilde{p}(x)} \left[- \int p(z|x) \log q(x|z) dz + KL(p(z|x)\|q(z)) \right] \quad (6)$$

Because $q(z)$, $p(z|x)$ are both Gaussian, we can get the analytic expression of $KL(p(z|x)\|q(z))$. And with the reparametrization trick, the first term can be approximate as $\log q(x|z)$. Consequently, the final loss for VAE is

$$\mathbb{E}_{x \sim \tilde{p}(x)} \left[- \log q(x|z) + KL(p(z|x)\|q(z)) \right] \quad (7)$$

The assumption of EM algorithm is like VAE, excluding supposing $p(z|x)$ is Gaussian. In EM algorithm, the loss is still (5), but we treat entire $p(z|x)$ as training parameters. Rather than minimizing the loss directly, here we use an alternate training way. Firstly, we fix $p(z|x)$ and just optimize $q(x|z)$. Removing the "constant" term, the loss of $q(x|z)$ is

$$q(x|z) = \arg \max_{q(x|z)} \mathbb{E}_{x \sim \tilde{p}(x)} \left[\int p(z|x) \log q(x, z) dz \right] \quad (8)$$

Secondly, we fix $q(x|z)$ and optimize $p(z|x)$. We define $q(z|x)$ by $q(x|z)$

$$q(x) = \int q(x|z) q(z) dz, \quad q(z|x) = \frac{q(x|z) q(z)}{q(x)} \quad (9)$$

now we have

$$\begin{aligned} p(z|x) &= \arg \min_{p(z|x)} \mathbb{E}_{x \sim \tilde{p}(x)} \left[\int p(z|x) \log \frac{p(z|x)}{q(z|x) q(x)} dz \right] \\ &= \arg \min_{p(z|x)} \mathbb{E}_{x \sim \tilde{p}(x)} [KL(p(z|x) \| q(z|x)) - \log q(x)] \\ &= \arg \min_{p(z|x)} \mathbb{E}_{x \sim \tilde{p}(x)} [KL(p(z|x) \| q(z|x))] \end{aligned} \quad (10)$$

Because we don't make any assumptions about the form of $p(z|x)$, we can let $p(z|x) = q(z|x)$ make loss equal zero, which is an optimal solution of $p(z|x)$. In other words, the optimal $p(z|x)$ is

$$p(z|x) = \frac{q(x|z) q(z)}{\int q(x|z) q(z) dz} \quad (11)$$

EM algorithm is just to perform (8), (11) alternately

GAN within Variational Inference

In this section, we describe a general approach to incorporate GAN into the variational inference, which leads a new insight to GAN and results a effective regularization for GAN.

General Framework

As same as VAE, GAN also want to achieve a generative model $q(x|z)$, which can transform $z \sim q(z) = N(z; 0, I)$ to the evidence distribution $x \sim \tilde{p}(x)$. Different from Gaussian assumption in VAE, GAN let $q(x|z)$ be a Dirac delta function

$$q(x|z) = \delta(x - G(z)), \quad q(x) = \int q(x|z) q(z) dz \quad (12)$$

whose $G(z)$ is a neural network of generative model, called generator.

Generally, we considered z is a random latent variable in generative model. However, it is well-known that Dirac delta function is non-zero at only one point, so the mapping from z to x in GAN is almost one to one. The variable z is not "random" enough, so we do not treat it as a latent variable (that means we need not to consider posterior distribution $p(z|x)$). In fact, we just consider the binary random variable y as a random latent variable in GAN:

$$q(x, y) = \begin{cases} \tilde{p}(x)p_1, & y = 1 \\ q(x)p_0, & y = 0 \end{cases} \quad (13)$$

here $p_1 = 1 - p_0$ describing a Bernoulli distribution. For simpler we set $p_1 = p_0 = 1/2$.

On the other hand, we let $p(x, y) = p(y|x)\tilde{p}(x)$, while $p(y|x)$ is a conditional Bernoulli distribution. Distinct from VAE, GAN choose another direction of KL-divergence as optimal objective:

$$\begin{aligned} &KL(q(x, y) \| p(x, y)) \\ &= \int \left(\tilde{p}(x)p_1 \log \frac{\tilde{p}(x)p_1}{p(1|x)\tilde{p}(x)} + q(x)p_0 \log \frac{q(x)p_0}{p(0|x)\tilde{p}(x)} \right) dx \\ &\sim \int \tilde{p}(x) \log \frac{1}{p(1|x)} dx + \int q(x) \log \frac{q(x)}{p(0|x)\tilde{p}(x)} dx \end{aligned} \quad (14)$$

Once succeed, we have $q(x, y) \rightarrow p(x, y)$, means

$$\begin{aligned} &p_1\tilde{p}(x) + p_0q(x) \\ &= \sum_y q(x, y) \rightarrow \sum_y p(x, y) = \tilde{p}(x) \end{aligned} \quad (15)$$

consequently $q(x) \rightarrow \tilde{p}(x)$.

Now we have to solve $p(y|x)$ and $G(x)$. For simpler we set $p(1|x) = D(x)$, called discriminator. Like EM algorithm, we use a alternately training strategy. Firstly, we fix $G(z)$, so $q(x)$ does. Ignoring constants for $G(z)$, we get:

$$\begin{aligned} D &= \arg \min_D - \mathbb{E}_{x \sim \tilde{p}(x)} [\log D(x)] \\ &\quad - \mathbb{E}_{x \sim q(x)} [\log(1 - D(x))] \end{aligned} \quad (16)$$

Then we fix $D(x)$ for optimizing $G(x)$. Ignoring constants for $D(x)$, we get the pure loss:

$$G = \arg \min_G \int q(x) \log \frac{q(x)}{(1 - D(x))\tilde{p}(x)} dx \quad (17)$$

For minimizing this loss, we need the formula of $\tilde{p}(x)$, which is always impossible. For the same reason as (11), if $D(x)$ has enough fitting ability, the optimal $D(x)$ is

$$D(x) = \frac{\tilde{p}(x)}{\tilde{p}(x) + q^o(x)} \quad (18)$$

$q^o(x)$ is $q(x)$ at previous stage. We can solve $\tilde{p}(x)$ from it and replace $\tilde{p}(x)$ in (17):

$$\begin{aligned} &\int q(x) \log \frac{q(x)}{D(x)q^o(x)} dx \\ &= -\mathbb{E}_{x \sim q(x)} [\log D(x)] + KL(q(x) \| q^o(x)) \\ &= -\mathbb{E}_{z \sim q(z)} [\log D(G(z))] + KL(q(x) \| q^o(x)) \end{aligned} \quad (19)$$

Basic Analysis

It is obviously that the fisrt term is one of the standard G losses of GAN:

$$-\mathbb{E}_{z \sim q(z)} \log D(G(z)) \quad (20)$$

The second extra item describes the distance between the new distribution and the old distribution. Two terms are adversarial. $KL(q(x) \| q^o(x))$ try to make the two distributions more similar, while $-\log D(x)$ will be very large because $D(x)$ will be very small for $x \sim q^o(x)$ if discriminator is

trained fully (all of them will be considered as negative samples), and vice versa. Thus, minimizing entire loss requires model to inherit the old distribution $q^o(x)$ and explore the new world $p(1|y)$.

As we know, the generator's loss in current standard GAN has no the second term, which is actually an incomplete loss. Suppose there is an omnipotent optimizer which can identify the global optimum in very short time and $G(z)$ has enough fitting ability, then $G(z)$ can only generate just one sample which make $D(x)$ largest. In other words, the global optimal solution of $G(z)$ is $G(z) = x_0$, while $x_0 = \arg \max_x D(x)$. That is called Model Collapse, which will occur certainly in theory.

So, what enlightenment can $KL(q(x)\|q^o(x))$ give for us? We let

$$q^o(x) = q_{\theta-\Delta\theta}(x), \quad q(x) = q_\theta(x) \quad (21)$$

that means the updates of parameters of $G(z)$ in this iteration is $\Delta\theta$. Using Taylor series to expand $q_{\theta-\Delta\theta}(x)$ to second order, we get

$$\begin{aligned} KL(q(x)\|q^o(x)) &\approx \int \frac{(\Delta\theta \cdot \nabla_\theta q_\theta(x))^2}{2q_\theta(x)} dx \\ &\approx (\Delta\theta \cdot c)^2 \end{aligned} \quad (22)$$

We have already indicated that a complete loss should contain $KL(q(x)\|q^o(x))$. If not, we should keep it small during training by other way. The above approximation shows the extra loss is about $(\Delta\theta \cdot c)^2$, which can not be too large, meaning $\Delta\theta$ can not be too large because c can be regarded as a constant during one iteration.

Now we can explain why we need to adjust hyperparameters carefully to make GAN converge(Salimans et al. 2016). The most common optimizers we use are all based on gradient descent, so $\Delta\theta$ is proportional to gradients. We need to keep $\Delta\theta$ small, which is equivalent to keep gradients small. Consequently, we apply gradients clipping, Batch Normalization in GAN because they can make gradients steady. For the same reason, we always use Adam rather than SGD+Momentum. Meanwhile, the iterations of $G(z)$ can not be too many, while $\Delta\theta$ will be large if $G(z)$ updates a lot.

Regularization Term

Here we focus on getting something really useful and practical. We try to calculate $KL(q(x)\|q^o(x))$ directly, for obtain an regularization term to add in generator's loss. Because of difficulty of directly calculating, we estimate it via calculating $KL(q(x,z)\|\tilde{q}(x,z))$ (maybe inspired by variational inference):

$$\begin{aligned} &KL(q(x,z)\|\tilde{q}(x,z)) \\ &= \iint q(x|z)q(z) \log \frac{q(x|z)q(z)}{\tilde{q}(x|z)q(z)} dxdz \\ &= \iint \delta(x - G(z))q(z) \log \frac{\delta(x - G(z))}{\delta(x - G^o(z))} dxdz \\ &= \int q(z) \log \frac{\delta(0)}{\delta(G(z) - G^o(z))} dz \end{aligned} \quad (23)$$

we have a limitation

$$\delta(x) = \lim_{\sigma \rightarrow 0} \frac{1}{(2\pi\sigma^2)^{d/2}} \exp\left(-\frac{x^2}{2\sigma^2}\right) \quad (24)$$

which means $\delta(x)$ can be replaced with a Gaussian distribution of small variance. So we have

$$KL(q(x)\|q^o(x)) \sim \lambda \int q(z) \|G(z) - G^o(z)\|^2 dz \quad (25)$$

(19) becomes

$$\mathbb{E}_{z \sim q(z)} [-\log D(G(z)) + \lambda \|G(z) - G^o(z)\|^2] \quad (26)$$

In other words, we can use the distance between samples from old and new generator as a regularization term, to guarantee the new generator has little deviation from old generator.

Experiment 1 and 2 on CelebA datasets¹ shows this regularization works well.

Models Related to GAN

Adversarial Autoencoders (AAE)(Makhzani et al. 2015) and Adversarially Learned Inference (ALI)(Dumoulin et al. 2016) are two variants of GAN, they also can be incorporated into variational inference. Of course, with the preparation above, it is just like two homework questions.

AAE under GAN framework

Actually, for obtaining AAE, the only thing we need to do is exchanging x, z in standard GAN. In detail, AAE wants to train an encoder $p(z|x)$ to map the distribution of real data $\tilde{q}(x)$ to the standard Gaussian distribution $q(z) = N(z; 0, I)$, while

$$p(z|x) = \delta(z - E(x)), \quad p(z) = \int p(z|x)\tilde{q}(x)dx \quad (27)$$

whose $E(x)$ is a neural network of encoder.

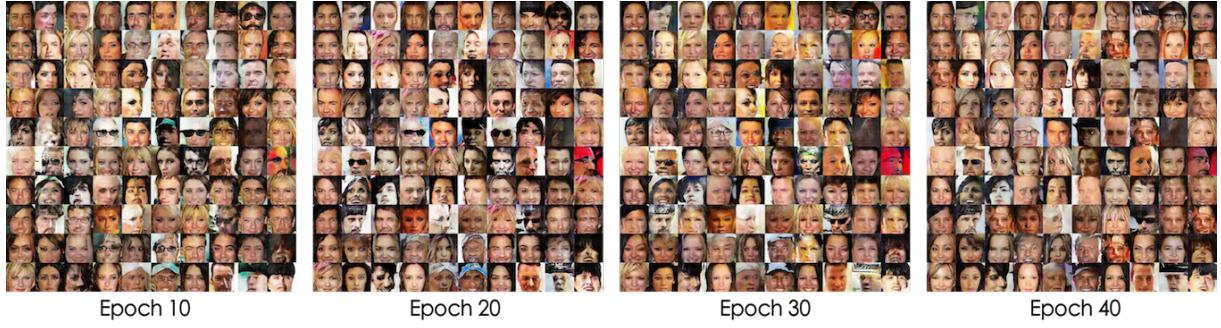
Like GAN, AAE needs a binary random latent variable y , and

$$p(z,y) = \begin{cases} p(z)p_1, & y = 1 \\ q(z)p_0, & y = 0 \end{cases} \quad (28)$$

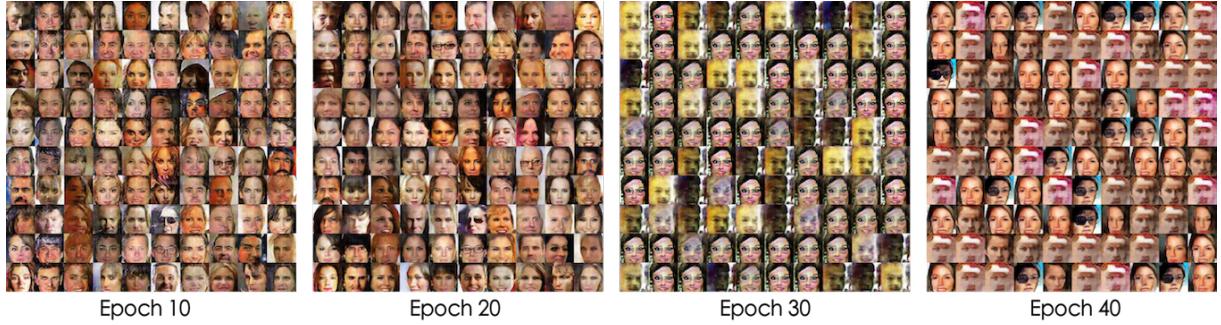
we also let $p_1 = p_0 = 1/2$. On the other hand, we set $q(z,y) = q(y|z)q(z)$, whose posterior distribution $p(y|z)$ is conditional Bernoulli distribution taking z as input. Now we minimize $KL(p(z,y)\|q(z,y))$:

$$\begin{aligned} &KL(p(z,y)\|q(z,y)) \\ &= \int p(z)p_1 \log \frac{p(z)p_1}{q(1|z)q(z)} dz + \int q(z)p_0 \log \frac{q(z)p_0}{q(0|z)q(z)} dz \\ &\sim \int p(z) \log \frac{p(z)}{q(1|z)q(z)} dz + \int q(z) \log \frac{1}{q(0|z)} dz \end{aligned} \quad (29)$$

¹the code is modified from <https://github.com/LynnHo/DCGAN-LSGAN-WGAN-WGAN-GP-Tensorflow>, now available at <https://github.com/bojone/gan/tree/master/vgan>.

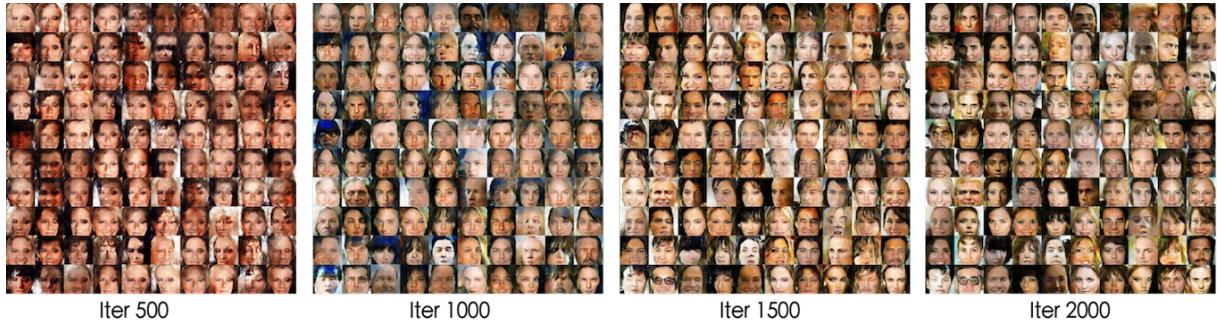


(a) With our regularization, the model keeps steady.

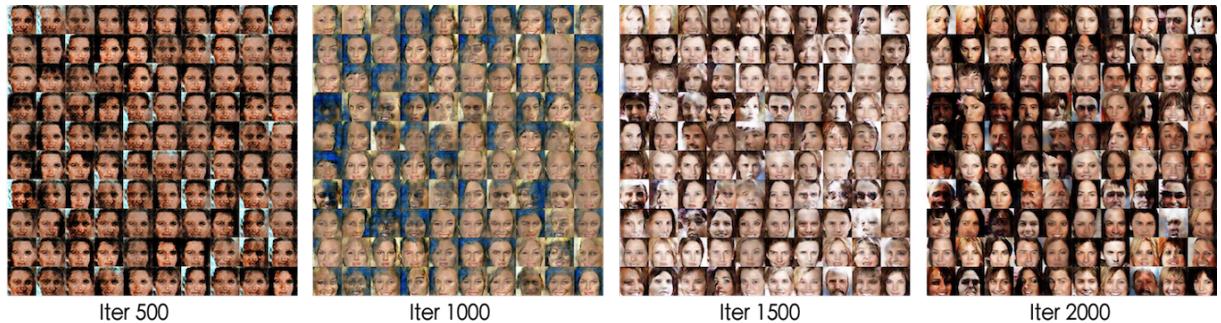


(b) Without regularization, the model collapses after 25 epochs.

Figure 1: An ordinary DCGAN model on CelebA, train discriminator and generator one iteration per period.



(a) With our regularization, the model has a faster convergence.



(b) Without regularization, the model need more iterations for convergence.

Figure 2: An ordinary DCGAN model removing BN, train discriminator and generator 5 iterations per period.

Now we have to solve $q(y|z)$ and $E(x)$. we set $p(0|z) = D(z)$ and still train it alternately. Firstly, we fix $E(x)$, so $p(z)$ does. Ignoring constants for $E(z)$, we get:

$$D = \arg \min_D - \mathbb{E}_{z \sim p(z)} [\log(1 - D(z))] - \mathbb{E}_{z \sim q(z)} [\log D(z)] \quad (30)$$

Then we fix $D(z)$ for optimizing $E(z)$. Ignoring constants for $D(z)$, we get the pure loss:

$$E = \arg \min_E \int p(z) \log \frac{p(z)}{(1 - D(z))q(z)} dz \quad (31)$$

Use the theoretical solution $D(z) = q(z)/[p^o(z) + q(z)]$ and replace $q(z)$:

$$\mathbb{E}_{x \sim \tilde{p}(x)} [-\log D(E(x))] + KL(p(z) \| p^o(z)) \quad (32)$$

On the one hand, like standard GAN, if we train carefully, we may remove the second term and have

$$\mathbb{E}_{x \sim \tilde{p}(x)} [-\log D(E(x))] \quad (33)$$

on the other hand, we can train a decoder $G(z)$ after finishing the adversarial training. However, if our $E(x), G(z)$ has strong enough modeling ability, then we can add a reconstruction error into encoder's loss, which will not interfere the original adversarial optimization of encoder. Therefore, we get a joint loss:

$$G, E = \arg \min_{G, E} \mathbb{E}_{x \sim \tilde{p}(x)} [-\log D(E(x)) + \lambda \|x - G(E(x))\|^2] \quad (34)$$

Our Version of ALI

ALI is like a fusion of GAN and AAE. And there is an almost identical version called Bidirectional GAN (BiGAN)(Donahue, Krhenbhl, and Darrell 2017). Compared with GAN, they treats z as a latent variable, so it needs a posterior distribution $p(z|x)$. Concretely, in ALI we have

$$q(x, z, y) = \begin{cases} p(z|x)\tilde{p}(x)p_1, & y = 1 \\ q(x|z)q(z)p_0, & y = 0 \end{cases} \quad (35)$$

and $p(x, z, y) = p(y|x, z)p(z|x)\tilde{p}(x)$, then we minimize $KL(q(x, z, y) \| p(x, z, y))$:

$$\begin{aligned} & \iint p(z|x)\tilde{p}(x) \log \frac{p(z|x)\tilde{p}(x)p_1}{p(1|x, z)p(z|x)\tilde{p}(x)} dx dz \\ & + \iint q(x|z)q(z) \log \frac{q(x|z)q(z)p_0}{p(0|x, z)p(z|x)\tilde{p}(x)} dx dz \end{aligned} \quad (36)$$

which is equivalent to minimize

$$\begin{aligned} & \iint p(z|x)\tilde{p}(x) \log \frac{1}{p(1|x, z)} dx dz \\ & + \iint q(x|z)q(z) \log \frac{q(x|z)q(z)}{p(0|x, z)p(z|x)\tilde{p}(x)} dx dz \end{aligned} \quad (37)$$

Now we have to solve $p(y|x, z), p(z|x), q(x|z)$. we set $p(1|x, z) = D(x, z)$, while $p(z|x)$ is a Gaussian distribution including an encoder $E(x)$ and $q(x|z)$ is an another

Gaussian distribution including an generator $G(z)$. Still alternately train it. Firstly we fix E, G , the loss related to D is

$$D = \arg \min_D - \mathbb{E}_{x \sim \tilde{p}(x), z \sim p(z|x)} \log D(x, z) - \mathbb{E}_{z \sim q(z), x \sim q(x|z)} \log(1 - D(x, z)) \quad (38)$$

As same as VAE, the expectation of $p(z|x)$ and $q(x|z)$ can be done using the reparametrization trick. Now fix D for optimizing G, E , and because of cooccurrence of E, G , loss (37) can not be simplified. But using the theoretical solution of D

$$D(x, z) = \frac{p^o(z|x)\tilde{p}(x)}{p^o(z|x)\tilde{p}(x) + q^o(x|z)q(z)} \quad (39)$$

can transform it to

$$\begin{aligned} & - \iint p(z|x)\tilde{p}(x) \log D(x, z) dx dz \\ & - \iint q(x|z)q(z) \log D(x, z) dx dz \\ & + \int q(z)KL(q(x|z) \| q^o(x|z)) dz \\ & + \iint q(x|z)q(z) \log \frac{p^o(z|x)}{p(z|x)} dx dz \end{aligned} \quad (40)$$

Due to Gaussianity of $q(x|z)$ and $p(x|z)$, we can calculate last two term analytically, or ignore them while optimizing it carefully, leading

$$\begin{aligned} & - \iint p(z|x)\tilde{p}(x) \log D(x, z) dx dz \\ & - \iint q(x|z)q(z) \log D(x, z) dx dz \end{aligned} \quad (41)$$

That is our version of ALI, which has little different from the standard ALI. The current popular view is to treat ALI (includes GAN) as a min-max problem. From that, the loss of encoder and generator is

$$\begin{aligned} & \iint p(z|x)\tilde{p}(x) \log D(x, z) dx dz \\ & + \iint q(x|z)q(z) \log(1 - D(x, z)) dx dz \end{aligned} \quad (42)$$

or

$$\begin{aligned} & - \iint p(z|x)\tilde{p}(x) \log(1 - D(x, z)) dx dz \\ & - \iint q(x|z)q(z) \log D(x, z) dx dz \end{aligned} \quad (43)$$

both of which are not like (41). Our experiment shows (41) has the same performance as (42) and (43). That means treating adversarial networks as a min-max problem is not the only one approach. Variational inference may give us some new insight sometimes.

Conclusion

Our results prove that variational inference is a general framework to derivate and explain many generative models, including VAE and GAN. We also describe how variational

inference do that by introducing a new interpretation of variational inference. This interpretation is powerful, which can lead to VAE and EM algorithm in severral lines and deduce GAN in clearly.

An related work is (Hu et al. 2018), which also attemps to link VAE and GAN with variational inference. However, their processing is not clear enough. we made up for this deficiency, trying to give an simpler view on GAN under variational inference.

It seems the potential of variational inference is waitting to be mined.

References

- [Donahue, Krhenbhl, and Darrell 2017] Donahue, J.; Krhenbhl, P.; and Darrell, T. 2017. Adversarial feature learning.
- [Dumoulin et al. 2016] Dumoulin, V.; Belghazi, I.; Poole, B.; Mastropietro, O.; Lamb, A.; Arjovsky, M.; and Courville, A. 2016. Adversarially learned inference.
- [Goodfellow et al. 2014] Goodfellow, I. J.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; and Bengio, Y. 2014. Generative adversarial nets. In *International Conference on Neural Information Processing Systems*, 2672–2680.
- [Hu et al. 2018] Hu, Z.; Yang, Z.; Salakhutdinov, R.; and Xing, E. P. 2018. On unifying deep generative models.
- [Kingma and Welling 2013] Kingma, D. P., and Welling, M. 2013. Auto-encoding variational bayes.
- [Makhzani et al. 2015] Makhzani, A.; Shlens, J.; Jaitly, N.; and Goodfellow, I. 2015. Adversarial autoencoders. *Computer Science*.
- [Salimans et al. 2016] Salimans, T.; Goodfellow, I.; Zaremba, W.; Cheung, V.; Radford, A.; and Chen, X. 2016. Improved techniques for training gans.