Report for CSE 5339 2018 — (OTMLSA)
Optimal Transport in Machine Learning and Shape Analysis

From OT to generative modeling: the VEGAN cookbook.

Cheng Xin

04/24/1018

1 Introduction

Recently the optimal transport techniques have raised great attention in the field of machine learning and data analysis. This idea is used in the problem of domain adaption and generative modeling.

The Kantorovich's formulation [1] of this problem is given by:

$$W_c(P,Q) := \inf_{\Gamma \in \mathcal{P}(\mathcal{X}\sim P, \mathcal{Y}\sim Q)} \mathbb{E}_{(X,Y)\sim\Gamma}[c(X,Y)], \tag{1}$$

where $c(x,y) : \mathcal{X} \times \mathcal{X} \to \mathcal{R}_+$ is any measurable cost function and $\mathcal{P}(X \sim P, Y \sim Q)$ is a set of all joint distribution of X, Y with marginals P and Q respectively. An element in the set $\mathcal{P}(X \sim P, Y \sim Q)$ is also called a coupling. When we choose $(\mathcal{X}, d)$ as a metric space and $c(x,y) = d^p(x,y)$ for $p \geqslant 1$, the distance W_p is called the *p-Wasserstein distance*.

The Kantorovich-Rubinstein theorem gives a duality for the 1-Wasserstein distance, which holds under mild assumptions on P and Q:

$$W_1(P,Q) := \sup_{f \in \mathcal{F}_L} \mid \mathbb{E}_{X\sim P}[f(X)] - \mathbb{E}_{Y\sim Q}[f(Y)] \mid, \tag{2}$$

where $\mathcal{F}$ is the class of all bounded 1-Lipschitz functions on the matric space (X, d).

This work [2] is to study the problem of constructing what are called latent variable model P_G to mimic the true but unknown data distribution P_X. Previously, there are two main ideas. One is called autoencoder, which is based on latent encoding and decoding functions. The other one is called adversarial generative network which does not rely on an explicit encoding function. It optimizes an objective function which is an lower bound of JS-divergence. A later varient version called wasserstein generative adversial network [3] applies the idea of optimal transport to measuring the divergence of the generative model distribution and the underlying data distribution and use it as a more efficient objective w.r.t. the convergence of the optimization algorithm.

This work tried to show an explicit connection between these two ideas. They proposed an idea to factorize the coupling between the generative distribution and the data distribution by the encoding function which satisfies some good probability. They also proposed a relaxed version with some convex penalty function to make their objective of the problem computationally tractable. Later in their paper, they also discussed some connection between their objective function and other previous works.

2 Background

The motivation of this work is the following problem, given a sampling from an unknown distribution, how can we find a reasonable generative model whose sampling has the similar distribution as the unknown distribution. Since it is quite hard to build a good generative model w.r.t maximal likelihood estimation in high dimensional space, one way to do that is to find some latent function, called generator, to generate some lower dimensional distribution which we understand it better and can be sampled easily. First we introduction some notation:

- sets: calligraphic letters, $\mathcal{X}$
- random variable: capital letters, X
- random variable values: lower case letters, x
- probability distributions: capital letters functions, $P(X)$ or P_X
- lower case letters functions, $p(x)$

The problem is like the following:

From a code in a (latent) lower dimensional space $Z \sim P_Z$ for some fixed distribution P_Z which we know how to sample efficiently, we want to find a generator $G : \mathcal{Z} \to \mathcal{X}$ which is good in the sense of that the induced distribution P_G of the image $G(Z)$ in the space of the original dataset $\mathcal{X}$ is closed to the underlying but unknown distribution P_X, w.r.t. some distance function.

For the **original generative adversarial network** (GAN) [4], the distance function is:

$$D_{GAN}(P_X, P_G) = \sup_{T \in \mathcal{T}} \mathbb{E}_{X \sim P_X}[\log T(X)] + \mathbb{E}_{Z \sim P_Z}[\log(1 - T(G(Z)))] \tag{3}$$

where T is a collection of non-parametric choice functions.

A recent variation of called **Wasserstein GAN**(WGAN) [3] choose the distance function as:

$$D_{WGAN}(P_X, P_G) = \sup_{T \in \mathcal{W}} mathdsE_{X \sim P_X}[T(X)] - \mathbb{E}_{Z \sim P_Z}[T(G(Z))] \tag{4}$$

3 Penalized OT

To build the connection between OT problem in generative model and the encoding function in AE problem, the author proposed a technique to factorize couplings. The intuition is like the following. For some coupling $\gamma \in \mathcal{P}(X \sim P_X, y \sim P_G)$, it can be factorized as the following:

$$\gamma(x, y) = \int_{\mathcal{Z}} p_G(y|z) q(z|x) p_X(x) dz, \tag{5}$$

conditional distribution $Q(Z|X)$ satisfies $q_Z(z) := \int_{\mathcal{X}} q(z|x) p_X(x) dx = p_Z(z)$ for all $z \in \mathcal{Z}$. This means that the the latent coding space Z is fully determined by the data distribution P_Z. Then, it is argued in the paper that a part of the search space (couplings set) of the OT problem can be reduced into a smaller space with respect to the *probabilistic encoders* $Q(Z|X)$.

The formal statement is the following theorem:

Theorem If $P_G(Y|Z = z) = \delta_{G(z)} \forall z \in \mathcal{Z}$, where $G : \mathcal{Z} \to \mathcal{X}$, we have:

$$W_c(P_X, P_G) = W_c^\dagger = \inf_{P \in \mathcal{P}(\mathcal{X} \sim \mathcal{P}_\mathcal{X}, \mathcal{Z} \sim \mathcal{P}_\mathcal{Z})} \mathbb{E}_{(X,Y) \sim P}[c(X, G(Z))] = \inf_{Q : Q_Z = P_Z} \mathbb{E}_{P_X} \mathbb{E}_{Q(Z|X)}[c(X, G(Z))] \tag{6}$$

where Q_Z is the marginal distribution of Z when $X \sim P_X$ and $Z \sim Q(Z|X)$.

The Dirac measure condition makes sure that we have a well-defined determinstic generator $G : \mathcal{Z} \to \mathcal{X}$, the resulting model P_G is just the push-forward distribution of P_Z through generator function G.

The condition on the r.h.s of the above equation is a bit restrict since intuitively, it says that the encoding probability $Q(Z|X)$ should be good enough w.r.t. the induced marginal distribution $Q_Z(z) = \int_X Q(z|x)P(x)dx$ coincides with the marginal (prior) distribution P_Z. This is not a simple problem to solve. So the author proposed a relaxation for this problem. The idea is quite classical. Replace the strong constrain $P_Z = Q_Z$ with some *panelty* $F : Q \to \mathcal{R}_+$, such that $F(Q) = 0$ iff $P_Z = Q_Z$, and for any $\lambda > 0$, construct the following relaxed version of $W_c^\dagger(P_X, P_G)$:

$$W_c^\lambda(P_X, P_G) := \inf_{Q(Z|X)} \mathbb{E}_{P_X}\mathbb{E}_{Q(Z|X)}[c(X, G(Z))] + \lambda F(Q) \tag{7}$$

There are a lot of choices of convex penalty functions F. The author argued that some of them like D_{JS}, D_{KL} or other f-divergence family can result in intractable F. So they choose adversarial approximation $D_{GAN}(Q_Z, P_Z)$, which becomes tight in the nonparametric limit. This give the object function which they called *penalized optimal transport(POT)*:

$$D_{POT}(P_X, P_G) := \inf_{Q(Z|X)\in\mathcal{Q}} \mathbb{E}_{P_X}\mathbb{E}_{Q(Z|X)}[c(X, G(Z))] + \lambda D_{GAN}(Q_Z, P_Z) \tag{8}$$

where $\mathcal{Q}$ is any nonparametric set of conditional distributions.

With a differential cost function c, this problem can be solved with SGD algorithm similarly to AAE, by iteratively updating Q and G and adversarial discriminator of D_{GAN}. Later, the author also shew that D_{AEE} can be viewed as a special case when choosing the squared Euclidean as the cost c and the $P_G(Y|Z)$ being the Gaussian.

When the cost function is chosen to be $c(x, y) = ||x-y||^2$ and $P_G(Y|Z) = \mathcal{N}(Y; G(Z), \delta^2 \cdot I)$, the authors compared their objective with objectives in other previous models, VAE, AVB, AAE.

3.1 relations to VAE, AVB

Variational autoencoder (VAE) [5] is a method of generative modeling with the following objective

$$D_{VAE}(P_X, P_G) = \inf_{Q(Z|X)\in\mathcal{Q}} \mathbb{E}_{P_X}[D_{KL}(Q(Z|X), P_Z) - \mathbb{E}_{Q(Z|X)}[\log p_G(X|Z)]]. \tag{9}$$

w.r.t. *generator(decoder) mapping* $P_G(X|Z)$. If the set $\mathcal{Q}$ is rich enough to contain all conditional probability distribution $Q(Z|X)$, the objective D_{VAE} concides with the negative marginal log-likelihood $D_{VAE}(P_X, P_G) = -\mathbb{E}_{P_X}[\log P_G(X)]$. In practice, $\mathcal{Q}$ is a class of Gaussian distributions. So the optimal solution minize an upper bound on the negative log-likelihood or, equivalently, on the KL-divergence $D_KL(P_X, P_G)$.

To decrease the gap, Adversarial variational Bayes (AVB) [6] is proposed to improve it by enlarging the class $\mathcal{Q}$. The idea is similar to what we have mentioned in the first chapter, instead of parameterize the conditional distribution directly, we can parameterize transformation functions $e : \mathcal{X} \times \mathcal{R} \to \mathcal{Z}$. A random variable $e(x, \epsilon)$ implicitely defines a conditional distribution $Q_e(Z|X = x)$. Instead of only a collection of Gaussian distributions, AVB allows $\mathcal{Q}$ to be a collection of all distributions induced by these transformation functions which should be differential. Also, it replaces the intractable term $D_{KL}(Q_e(Z|X), P_Z)$ by $D_{f,GAN}(Q_e(Z|X), P_Z)$, which results the following objective:

$$D_{AVB}(P_X, P_G) = \inf_{Q_e(Z|X)\in\mathcal{Q}} \mathbb{E}_{P_X}[D_{f,GAN}(Q_e(Z|X), P_Z) - \mathbb{E}_{Q(Z|X)}[\log p_G(X|Z)]]. \tag{10}$$

Compared with VAE and AVB, they proposed the following proposition:

Proposition Let $\mathcal{X} = \mathcal{R}^d$ and assume $c(x, y) = ||x - y||^2, P_G(Y|Z) = \mathcal{N}(Y; G(Z), \delta^2 \cdot I)$ with any function $G : \mathcal{X} \to \mathcal{R}$. If $\delta^2 > 0$, then the function G^*_δ and $G^\dagger$ minimizing $W_c(P_X, P^\delta_G)$ and $W^\dagger_c(P_X, P^\delta_G)$ respectively are different: G^*_δ depends on δ^2, while $G^\dagger$ does not. The function $G^\dagger$ is also a minimizer of $W_c(P_X, P^0_G)$.

They argued that this is an adventage of their method since first it gets rid of an parameter, and second, more importantly, it means their algorithm is stable w.r.t any $\delta > 0$, and third, when $\delta = 0$, the optimal $G^\dagger$ actually is an minimizer of $W_c(P_X, P^0_G)$.

3.2 Relation to AAE

Adversarial auto-encoders (AAE) is also a quite nice generative model first raised in the work [6]. It reported good empirical results from this model on generative problem.

The objective AAE is defined as the following.:

$$D_{AAE}(P_X, P_G) = \inf_{Q(Z|X) \in \mathcal{Q}} D_{GAN}(Q_Z, P_Z) - \mathbb{E}_{P_X} \mathbb{E}_{Q(Z|X)}[\log p_G(X|Z)], \tag{11}$$

where Q_Z is the marginal distribution of Z which is induced by P_X and $Q(Z|X)$, which is also called *aggregated posterior [6]*. One way to understand the AAE objective is that it replaces the D_{KL} term in equation 9 with another regularizer $D_{GAN}(Q_Z, P_Z)$. The author argued that, similarly to AVB, there is no clear link to log-likelihood since they shew that $D_{AAE} \leqslant D_{AVB}$. So another way to understand the effectiveness of AAE is to utilize their D_{POT} method. Since it is easy to see that D_{AAE} is a special case of D_{POT} with the cost function $c(X, G(Z))$ chosen to be $\log p_G(x|z)$. By their analysis, they suggested that AAE is infact attempting to minize 2-Wasserstein distance between P_X and P^δ_B.

4 Conclusion

This paper propose a divergence function called penalized optimal transport objective D_{POT}. The connection between the adversarial generative model and the autoencoder was further studied. The author proposed a way to factorize the coupling with a good choice encoding probability when the generative probability is a deterministic, Dirac measure. It is also studied in this paper that the connection of the objective function in AAE and the D_{POT} they proposed. They argued that their POT objective can be viewed as a generalization of AAE. This also justifies the effectiveness of the AAE model in a theorical level. So it will also be intereting to see if some other variance of D_{POT} can be useful besides AAE.

References

[1] L Kantorovich. On the transfer of masses (in russian). In *Doklady Akademii Nauk*, volume 37, pages 227–229, 1942.

[2] Olivier Bousquet, Sylvain Gelly, Ilya Tolstikhin, Carl-Johann Simon-Gabriel, and Bernhard Schoelkopf. From optimal transport to generative modeling: the vegan cookbook. *arXiv preprint arXiv:1705.07642*, 2017.

[3] Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein gan. *arXiv preprint arXiv:1701.07875*, 2017.

[4] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014.

[5] D. P Kingma and M. Welling. Auto-Encoding Variational Bayes. *ArXiv e-prints*, December 2013.

[6] A. Makhzani, J. Shlens, N. Jaitly, I. Goodfellow, and B. Frey. Adversarial Autoencoders. *ArXiv e-prints*, November 2015.