Wassernstein GAN

Martin Arjovsky et al.

2017/1

1 Introduction

この論文が関係している問題は教師なし学習のそれである．主に問題は，確率分布を学習するということはどういうことなのか？である．古典的な回答は，確率密度を学ぶ事ということである．これはよく，パラメトリックな分布族 (P_θ) を定義し，そしてデータにおいて尤度を最大にした θ を見つけることによってなされる．

もし，データ $\{x^{(i)}\}_{i=1}^m$ がある場合，以下の式を解くことでパラメータ θ を求める．

$$\underset{\theta\in\mathbb{R}^d}{\operatorname{argmax}} \frac{1}{m}\sum_{i=1}^{m} \log P_\theta(x^{(i)}) \tag{1}$$

もし，実データ分布 $\mathbb{P}_r$ が密度分布（連続）であるとすることができ，パラメータ化された確率密度 P_θ を $\mathbb{P}_\theta$ とすると，漸近的に上式は KL ダイバージェンス $KL(\mathbb{P}_r|\mathbb{P}_r)$ の最小化となる．

存在しないであろう密度分布 $\mathbb{P}_r$ を推定するのではなく，私たちは固定された分布 $p(z)$ に従うランダムな変数 Z を定義し，直接ある分布 $\mathbb{P}_\theta$ に従うサンプルを生成するパラメトリックな関数 $g_\theta : Z \longmapsto \chi$（主に何らかのニューラルネットワー）に通すことができる．θ を様々に変化させることによって，$\mathbb{P}_\theta$ を変化させデータの分布 $\mathbb{P}_r$ と近づけることができる．これは次の２つにおいて役に立つ．まず最初に，densities と違い，この方法は低次元多様体に制限された分布を表すことができる．次に簡単にデータを生成できる能力は分布の数値的な値を知ることよりも役に立つ[*1]．一般には，任意の高次元密度分布と仮定し，サンプルを生成することは難しい [1].

GAN や VAE はこのアプローチのよく知られている例である．GAN は目的関数の定義において柔軟であるが学習が難しいと知られている．

この論文では私たちはモデルの分布 $\mathbb{P}_\theta$ とデータの分布 $\mathbb{P}_r$ がどの程度近いかを測る様々な方法や，同様に距離またはダイバージェンス $\rho(\mathbb{P}_\theta, \mathbb{P}_r)$ を定義する様々な方法に注意を向ける．

この論文の主な貢献は以下の３つである．

1. 第二章では，Earth Mover (EM) 距離が learning distribution で使われるポピュラーな距離，ダイバージェンスと比べて，どのように振る舞うのかの理論的な解析を行う．
2. 第 3 章では，Wasserstein-GAN と呼ばれ，妥当であり，効率の良い EM distance の近似を最小化する GAN の形式を定義する．そして，対応する最適化問題が妥当，安定していることを理論的に証明する．
3. 第 4 章では，経験的に WGAN が GAN の主な学習問題を解決することを示す．具体的には，WGAN

[*1] 例えば，超解像やセマンティックセグメンテーションの領域において入力が与えられて出力の条件付き分布を考える時に役に立つ

の学習は Discriminator と Generator の学習において慎重にバランスを取ることを必要とせず，ニューラルネットワークの構造の慎重な設計も必要としない．GAN に特有の mode dropping(collapse) 現象もまた劇的に減少させる．もっとも注目を惹きつける WGAN の実用的な利点の一つは Discriminator を最適化することによって，連続的に (絶え間なく)EM 距離を推定することができる能力である．これらの学習曲線をプロットすることはデバックやハイパーパラメータの探索に役に立つだけでなく，学習曲線は観測されたサンプルの質と強い相関がある．

疑問点

1. For this to make sense, we need the model density P_θ to exist. This is not the case in the rather common situation where we are dealing with distributions supported by low dimensional manifolds. It is then unlikely that the model manifold and the true distribution's support have a non-negligible intersection (see [2]), and this means that the KL distance is not defined (or simply infinite).
   なぜ，モデルの多様体と真の分布のサポートが無視できない交点を持つと KL は発散するのか．確かに 2 章の Example1 は低次元多様体であり交点を持たない．その場合，KL は発散している．
2. First of all, unlike densities, this approach can represent distributions confined to a low dimensional manifold.
   なぜ，低次元多様体に制限された分布を表現できるのか．また，低次元多様体の分布を表現しなければならないのは何故か．
3. A sequence of distributions $(\mathbb{P}_t)_{t\in\mathbb{N}}$ converges if and only if there is a distribution $\mathbb{P}_\infty$ such that $\rho(\mathbb{P}_t, \mathbb{P}_\infty)$ tends to zero, something that depends on how exactly the distance ρ is defined. Informally, a distance ρ induces a weaker topology when it makes it easier for a sequence of distribution to converge.
   A sequence of distributions とは? また，収束しやすいとき ρ は弱位相を含むとはどういうことなのか．
4. The weaker this distance, the easier it is to define a continuous mapping from θ-space to $\mathbb{P}_\theta$-space, since it's easier for the distributions to converge.
   距離が弱いとはどういうことなのか．また，分布が収束しやすいため，距離が弱いと連続な θ-space から $\mathbb{P}_\theta$-space への写像が定義しやすいとはどういうことなのか．
5. The main reason we care about the mapping $\theta \longmapsto \mathbb{P}_\theta$ to be continuous is as follows. If ρ is our notion of distance between two distributions, we would like to have a loss function $\theta \longmapsto \rho(\mathbb{P}_\theta, \mathbb{P}_r)$ that is continuous, and this is equivalent to having the mapping $\theta \longmapsto \mathbb{P}_\theta$ be continuous when using the distance between distributions ρ.
   なぜ $\theta \longmapsto \mathbb{P}_\theta$ が連続ならば $\theta \longmapsto \rho(\mathbb{P}_\theta, \mathbb{P}_r)$ も連続になるのか．

2 Different Distance

3 Standard GAN の問題点

4 WGAN の改善点，利点

5 実装

著者の実装について

1. z は opt.resize_(opt.batchSize, nz, 1, 1).normal_(0, 1) として Generater に与えている．
2. 前処理: 画像をスケーリング (64 × 64 等に) し CenterCrop() したのち，Normalize((0.5, 0.5, 0.5), (0.5,0.5, 0.5)) で正規化している．つまり各チャンネルごとに以下の数式で正規化している．

$$input[channel] = (input[channel] - mean[channel])/std[channel] \tag{2}$$

   おそらく入力画像の画素値の取りうる範囲を $[-1, 1]$ にしている．
3. 初期値: Conv 系 (Convolution, Transpose) は平均 0 , 分散 0.02 の正規分布にて設定される．また，Batch Normalization の γ は平均 1.0, 分散 0.02 の正規分布で初期化され，β は 0 で初期化される．
4. Generator は ConvTranspose2d と BatchNormalization, ReLU を繰り返す構造をしている．詳しくは notebook 参照．
5. 論文では書いていないが，著者は

   > The only addition to the code (that we forgot, and will add, on the paper) are the lines 163-166 of main.py. These lines act only on the first 25 generator iterations or very sporadically (once every 500 generator iterations). In such a case, they set the number of iterations on the critic to 100 instead of the default 5. This helps to start with the critic at optimum even in the first iterations. There shouldn't be a major difference in performance, but it can help, especially when visualizing learning curves (since otherwise you'd see the loss going up until the critic is properly trained). This is also why the first 25 iterations take significantly longer than the rest of the training as well.

   と言っているのに注意
6. また，コード z は平均 0, 標準偏差 1 の正規分布から取ってきている．

疑問点

1. Lipschitz 関数であるために重み $w \in \mathcal{W}$ に対して $[-0.01, 0.01]$ 等に weight clipping を行うが，この場合 Batch Normalization の β, γ はどうすれば良いのか? こちらもまた weight clipping するべき? なにやら，Clitic（Discriminator）に weight clipping を施すと学習が崩壊することもあるらしい．
2. $[-0.01, 0.01]$ を超えたものは clip されるため、初期値の分散が大きいと全て-0.01 か 0.01 になる．よって重みの初期値をどう設定しているかが重要になりそう．

参考文献

[1] Radford M. Neal. Annealed importance sampling. Statistics and Computing, 11(2):125 − 139, April 2001.

[2] Martin Arjovsky and L ´ eon Bottou. Towards principled methods for training generative adversarial networks. I
https://arxiv.org/abs/1701.04862