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．

2 Standard GAN の問題点

3 WGAN の改善点，利点

4 実装

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．

参考文献

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