Robust Estimation and GANs

Selikhanovych Daniil

Project for the course
«Theoretical methods of Deep Learning»

December 18, 2020

Problem statement

Huber's ε-contamination problem

- Given i.i.d. observations

$$X_1, \dots, X_n \sim (1-\varepsilon)P_\theta + \varepsilon Q$$

  the task is to estimate the model parameter θ for $\varepsilon \in [0, 1)$.
- Meaning: each observation has a $1-\varepsilon$ probabilty to be drawn from P_θ and the other ε probability to be drawn from the unknown contamination distribution Q.
- Example: normal mean estimation problem with $P_\theta = \mathcal{N}(\theta, I_p)$.
- Motivation: robust statistics and theoretical computer science needs to find both statistically optimal and computationally feasible procedures.

Previous results

Minimax rates of estimation

- It has been shown that the minimax rate $R(\varepsilon)$ of estimating θ under Huber's ε-contamination problem takes the form of

$$R(\varepsilon) \asymp \max\{R(0), \omega(\varepsilon, \Theta)\},$$

  where $\omega(\varepsilon, \Theta)$ is the modulus of continuity between the loss function and the TV distance with respect to the space $\theta \in \Theta$.
- For the normal mean estimation problem, the minimax rate with respect to the squared ℓ_2 loss scales like $\max\{\frac{p}{n}, \varepsilon^2\}$, and is achieved by Tukey's median:

$$\widehat{\theta} = \operatorname*{argsup}_{\eta \in \mathbb{R}^p} \inf_{\|u\|=1} \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}\left\{u^T (X_i - \eta) \geq 0\right\}$$

Proposed method, idea

Computation problems

- Despite the statistical optimality of Tukey's median, its computation is not tractable. In fact, even an approximate algorithm takes $O(\exp(Cp))$ in time.
- How to achieve minimax optimal robust estimation and develop good computational strategies?
- Authors proposed the method based on minimizers of variational lower bounds of the total variation distance between the empirical measure and the model distribution.
- Variational lower bounds are computed through neural network approximations.

f-divergence

Variational lower bound

- Given a strictly convex function f that satisfies $f(1) = 0$, the f-divergence between two probability distributions P and Q with densities p and q respectively is defined by

$$D_f(P\|Q) = \int f\left(\frac{p}{q}\right) dQ.$$

- Examples: $f(x) = x\log(x)$ and $f(x) = ReLU(x-1)$ gives KL divergence and TV distance respectively.
- Convex conjugate of f: $f^*(t) \stackrel{\text{def}}{=} \sup_{u\in\text{dom}_f}(ut - f(u))$.
- It's easy to prove for any function class $\mathcal{T}$:

$$D_f(P\|Q) \geq \sup_{T\in\mathcal{T}} [E_P T(X) - E_Q f^*(T(X))]$$

f-GAN

Robust Estimation with f-GAN

- With i.i.d. observations $X_1, \ldots, X_n \sim P$, this variational lower bound naturally leads to the following learning method:

$$\widehat{P} = \operatorname*{arginf}_{Q \in \mathcal{Q}} \sup_{T \in \mathcal{T}} \left[\frac{1}{n} \sum_{i=1}^{n} T(X_i) - E_Q f^*(T(X)) \right].$$

- The idea of f-GAN is to find a $\widehat{P}$ so that the best discriminator T in the class $\mathcal{T}$ cannot tell the difference between $\widehat{P}$ and the empirical distribution $\frac{1}{n}\sum_{i=1}^{n} \delta_{X_i}$.
- How to choose the function f that leads to robust learning procedures which are easy to optimize? How to specify the discriminator class to learn the parameter of interest with minimax rate under Huber's ε-contamination model?

TV-GAN

Robust Estimation with TV-GAN

- For $f(x) = ReLU(x-1)$ we have $f^*(t) = t\mathbb{I}(0 \le t \le 1)$.
- For discriminator $D(x) = T(x) \in [0,1]$ with Gaussian location family $\mathcal{Q} = \{\mathcal{N}(\eta, I_p) : \eta \in \mathbb{R}^p\}$ we obtain

$$\widehat{\theta} = \underset{\eta \in \mathbb{R}^p}{\operatorname{arginf}} \sup_{D \in \mathcal{D}} \left[\frac{1}{n} \sum_{i=1}^{n} D(X_i) - E_{N(\eta, I_p)} D(X) \right].$$

- Need to specify the class of discriminators $\mathcal{D}$ to solve the classification problem between $\mathcal{N}(\eta, I_p)$ and $\frac{1}{n}\sum_{i=1}^{n} \delta_{X_i}$.
- One of the simplest discriminator classes is the logistic regression:

$$\mathcal{D} = \left\{ D(x) = \operatorname{sigmoid}\left(w^T x + b\right) : w \in \mathbb{R}^p, b \in \mathbb{R} \right\}$$

TV-GAN optimality

TV-GAN optimality for Huber's ε-contamination problem

Assume $\frac{p}{n} + \varepsilon^2 \leq c$ for some sufficiently small constant $c > 0$. With i.i.d. observations $X_1, \ldots, X_n \sim (1-\varepsilon)P_\theta + \varepsilon Q$, the estimator

$$\widehat{\theta} = \underset{\eta \in \mathbb{R}^p}{\operatorname{arginf}} \sup_{D \in \mathcal{D}} \left[\frac{1}{n} \sum_{i=1}^{n} D\left(X_i\right) - E_{N(\eta, I_p)} D(X) \right]$$

with logistic regression discriminators family $\mathcal{D}$ satisfies

$$\|\widehat{\theta} - \theta\|^2 \leq C \cdot \max\left\{\frac{p}{n}, \epsilon^2\right\}$$

with probability at least $1 - \exp\left(-C^{'}(p + n\varepsilon^2)\right)$ uniformly over all $\theta \in \mathbb{R}^p$ and all Q. The constants $C, C^{'} > 0$ are universal.

TV-GAN optimization

Algorithm `TV-GAN optimization`$(S, D_w, G_\eta, \gamma_d, \gamma_w, m, K, T)$:

input: $S = \{X_i\}_{i=1}^m \in \mathbb{R}^p$ - observation set; $D_w(x), G_\eta(z) = z + \eta$ - discriminator/generator networks.

for $t = 1, \ldots, T$ **do**

  for $k = 1, \ldots, K$ **do**

    Sample mini-batch $\{X_i\}_{i=1}^m$ from S;

    Sample $\{Z_i\}_{i=1}^m$ from $\mathcal{N}(0_p, I_p)$;

    $$g_w \leftarrow \nabla_w \left[\frac{1}{m} \sum_{i=1}^m D_w(X_i) - \frac{1}{m} \sum_{i=1}^m D_w(G_\eta(Z_i)) \right];$$

    $w \leftarrow w + \gamma_d g_w$;

  end

  Sample $\{Z_i\}_{i=1}^m$ from $\mathcal{N}(0_p, I_p)$;

  $$g_\eta \leftarrow \nabla_\eta \left[-\frac{1}{m} \sum_{i=1}^m D_w(G_\eta(Z_i)) \right], \ \eta \leftarrow \eta - \gamma_g g_\eta;$$

end

end

TV-GAN

Numerical optimization details

- Initialization of w: $w \sim \mathcal{N}(0, 0.05)$ independently on each element or Xavier. Initialization of b: zero.
- Initialization of η: coordinatewise median of S.
- Though TV-GAN can achieve the minimax rate, it may suffer from optimization difficulties especially when the distributions Q and $\mathcal{N}(\theta, I_p)$ are far away from each other.
- The main obstacle is, with optimization based on gradient, the discriminator may be stuck in a local maximum which will then pass wrong signals to the generator.

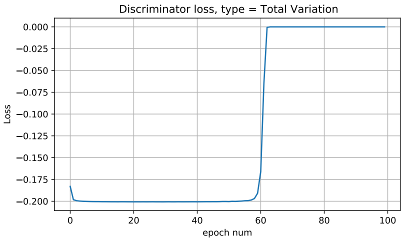
1D Example

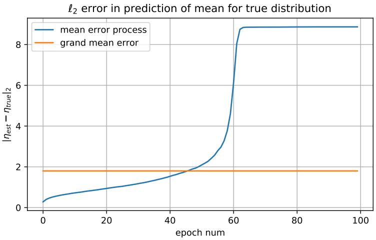
Example

- Let's consider the case with Gaussian contamination distribution.
- The first case $P = (1-\varepsilon)\mathcal{N}(1,1) + \varepsilon\mathcal{N}(10,1)$, when Q and $\mathcal{N}(\theta, I_p)$ are far away.
- The second case $P = (1-\varepsilon)\mathcal{N}(1,1) + \varepsilon\mathcal{N}(1.5,1)$, when Q and $\mathcal{N}(\theta, I_p)$ are close.
- The second case is hard, which is well predicted by the minimax theory of robust estimation.
- We will use 50000 examples of P with batch size $= 500$ for training.

TV-GAN optimization

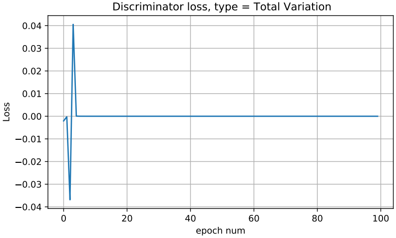
TV-GAN may suffer from optimization difficulties for $\eta_{cont} = 10$.
Figure: training of η estimation of TV-GAN for true initialization $\eta_0 = \eta_{true}$.
Orange curve - ℓ_2-error for grand mean $\eta_{est} = (1-\varepsilon)\eta_{true} + \varepsilon\eta_{cont}$.

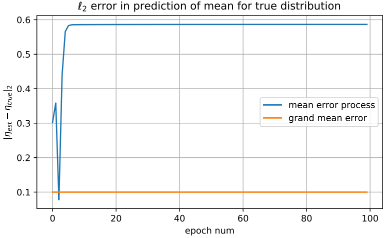




TV-GAN optimization

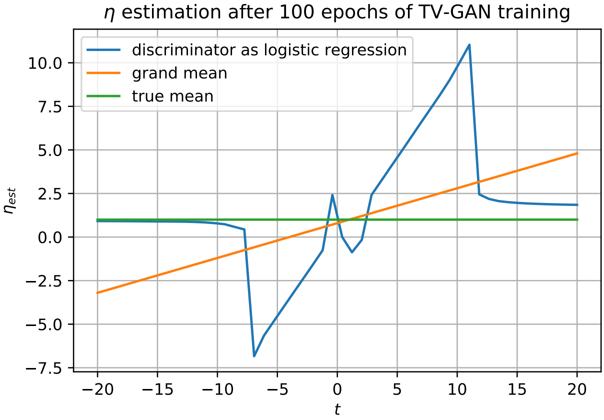
TV-GAN may suffer from optimization difficulties for $\eta_{cont} = 1.5$.
Figure: training of η estimation of TV-GAN for true initialization $\eta_0 = \eta_{true}$.
Orange curve - ℓ_2-error for grand mean $\eta_{est} = (1 - \varepsilon)\eta_{true} + \varepsilon\eta_{cont}$.





TV-GAN optimization

What about very large η_{cont}?
Figure: estimation of η after training TV-GAN for different $\eta_{cont} = t$ with 100 epochs. Orange curve - $\eta_{est}(t) = (1 - \varepsilon)\eta_{true} + \varepsilon t$.



JS-GAN

Robust Estimation with JS-GAN

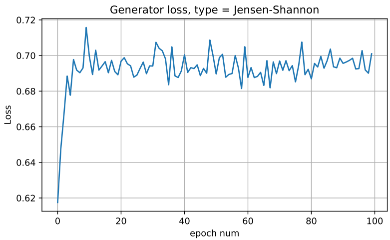
- For $f(x) = x\log(x) - (x+1)\log\left(\frac{x+1}{2}\right)$ we have $f^*(t) = -\log(1-t)$.
- For discriminator $D(x) = T(x) \in [0,1]$ with Gaussian location family $\mathcal{Q} = \{\mathcal{N}(\eta, I_p) : \eta \in \mathbb{R}^p\}$ we obtain

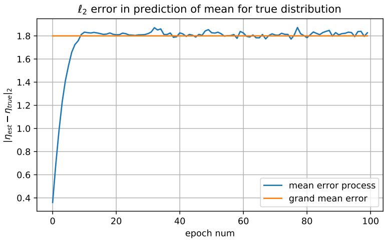
$$\widehat{\theta} = \operatorname*{arginf}_{\eta\in\mathbb{R}^p} \sup_{D\in\mathcal{D}} \left[\frac{1}{n}\sum_{i=1}^{n} \log\left(D\left(X_i\right)\right) + E_{N(\eta,I_p)} \log\left(1 - D(X)\right)\right].$$

- Need to specify the class of discriminators $\mathcal{D}$ to solve the classification problem between $\mathcal{N}(\eta, I_p)$ and $\frac{1}{n}\sum_{i=1}^{n}\delta_{X_i}$.
- We will see numerically that logistic regression in that case doesn't lead to robust estimation, but hidden layer built into the neural nets works well.

JS-GAN with logistic regression

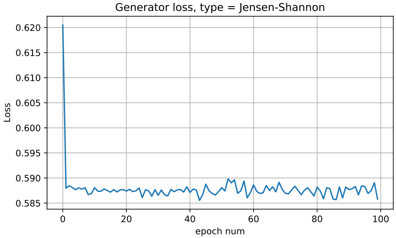
Training JS-GAN with logistic regression as discriminator for $\eta_{cont} = 10$ with Adam and $\gamma_d = 0.2, \gamma_g = 0.02$.
Left figure - loss for generator, right figure - ℓ_2-error in prediction of η_{true}.
Orange curve - ℓ_2-error for grand mean $\eta_{est} = (1 - \varepsilon)\eta_{true} + \varepsilon\eta_{cont}$.

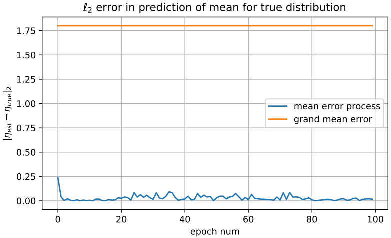




JS-GAN with 1 hidden layer NN

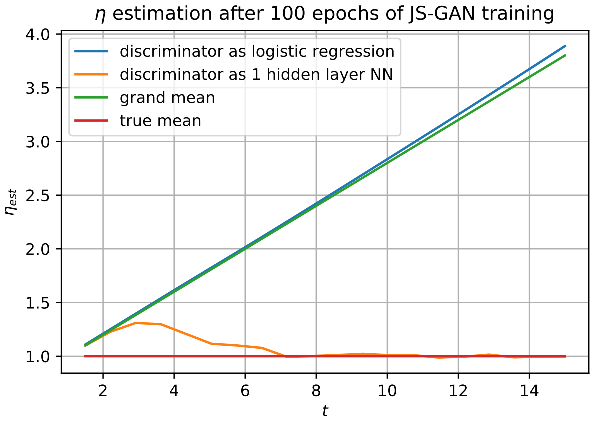
Training JS-GAN with 1 hidden layer NN (20 neurons) as discriminator for $\eta_{cont} = 10$ with Adam and $\gamma_d = 0.2, \gamma_g = 0.02$. Left figure - loss for generator, right figure - ℓ_2-error in prediction of η_{true}.
Orange curve - ℓ_2-error for grand mean $\eta_{est} = (1 - \varepsilon)\eta_{true} + \varepsilon\eta_{cont}$.





JS-GAN comparison

JS-GAN using discriminators without hidden layers always gives an estimator close to $0.2 + 0.8t$, , while the JS-GAN using discriminators with one hidden layer leads to robust estimation.



1 hidden layer NNs with bounded weights

Let's consider the following class of discriminators:

$$\mathcal{D} = \left\{ D(x) = \text{sigmoid}\left(\sum_{j \geq 1} w_j \sigma\left(u_j^T x + b_j\right) \right) : \sum_{j \geq 1} |w_j| \leq \varkappa, u_j \in \mathbb{R}^p, b_j \in \mathbb{R} \right\}$$

While the dimension of the input layer is p, the dimension of the hidden layer can be arbitrary, as long as the weights have a bounded ℓ_1 norm.

JS-GAN optimality

JS-GAN optimality for Huber's ε-contamination problem

Assume $\frac{p}{n} + \varepsilon^2 \le c$ for some sufficiently small constant $c > 0$ and set $\varkappa = O\left(\sqrt{\frac{p}{n}} + \varepsilon\right)$. With i.i.d. observations $X_1, \ldots, X_n \sim (1-\varepsilon)P_\theta + \varepsilon Q$, the estimator

$$\widehat{\theta} = \operatorname*{arginf}_{\eta \in \mathbb{R}^p} \sup_{D \in \mathcal{D}} \left[\frac{1}{n} \sum_{i=1}^{n} \log D\left(X_i\right) + E_{N(\eta, I_p)} \left(1 - \log D(X)\right) \right]$$

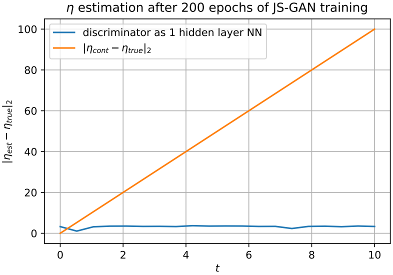
with 1 hidden layer NNs discriminators family $\mathcal{D}$ satisfies

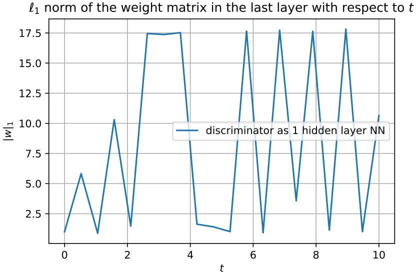
$$\|\widehat{\theta} - \theta\|^2 \le C \cdot \max\left\{\frac{p}{n}, \epsilon^2\right\}$$

with probability at least $1 - \exp\left(-C^{'}(p + n\varepsilon^2)\right)$ uniformly over all $\theta \in \mathbb{R}^p$ and all Q. The constants $C, C^{'} > 0$ are universal.

JS-GAN in big dimensionality

JS-GAN using discriminators with one hidden layer leads to robust estimation even in big dimensionality $p = 100$.





Conclusion

- Initialization in GANs plays significant role for the result.
- Needs for stopping criteria.
- Theory is very hard: Rademacher complexity, Dudley's integral entropy bound, VC-dimension for sigmoids.

Contribution

- Study the paper and technical proofs.
- Work with authors code and fix some mistakes.
- Source code and plots for the project can be found here `https://github.com/Daniil-Selikhanovych/f-gan`

Acknowledgment

I thank Maxim Panov for useful discussions and problem formulation!

References

Gao, Chao and Liu, Jiyi and Yao, Yuan and Zhu, Weizhi (2019). Robust Estimation and Generative Adversarial Nets. arXiv preprint arXiv:1810.02030.

Goodfellow, Ian and Pouget-Abadie, Jean and Mirza, Mehdi and Xu, Bing and Warde-Farley, David and Ozair, Sherjil and Courville, Aaron and Bengio, Yoshua (2014). Generative adversarial nets. Advances in neural information processing systems, pp. 2672-2680.

Huber, Peter J (1992). Robust estimation of a location parameter. Breakthroughs in statistics, pp. 492–518.

Thank you for attention!