Robust Estimation and Generative Adversarial Networks

Weizhi ZHU

Hong Kong University of Science and Technology

wzhuai@ust.hk

April 3, 2019

Robust Estimation and Generative Adversarial Nets [GLYZ18]
Generative Adversarial Nets for Robust Scatter Estimation: A Proper Scoring Rule Perspective [GYZ19]

Huber's Contamination Model

Huber's contamination model [*Huber, 1964*],

$$P = (1-\epsilon)P_\theta + \epsilon Q.$$

Strong contamination model [*Diakonikolas et al., 2016a*],

$$TV(P, P_\theta) \leq \epsilon.$$

Can we recover θ by data drawn from P with arbitrary unknown contamination (ϵ, Q)?

Example: Robust Mean Estimation

Let's firstly consider the robust estimation of location parameter θ in normal distribution,

$$X_1, \ldots, X_n \sim (1-\epsilon)\mathcal{N}(\theta, I_p) + \epsilon Q$$

- Coordinate-wise median.
- Tukey median *[Tukey, 1978]*.

$$\widehat{\theta} = \operatorname*{argmax}_{\eta \in \mathbb{R}^p} \min_{\|u\|_2 = 1} \sum_{i=1}^{n} 1\left\{u^T X_i > u^T \eta\right\} \wedge \sum_{i=1}^{n} 1\left\{u^T X_i \leq u^T \eta\right\}$$

Comparison

	Median	Tukey Median
statistical convergence rate (no contamination)	$\frac{p}{n}$	$\frac{p}{n}$
statistical convergence rate (Huber's ϵ contamination)	$\frac{p}{n} \vee p\epsilon^2$	$\frac{p}{n} \vee \epsilon^2$, [minimax]
computational complexity	Polynomial	NP-Hard

Example: Robust Covariance Estimation

We can also estimate the covariance matrix Σ in normal distribution,

$$X_1, \ldots, X_n \sim (1-\epsilon)\mathcal{N}(0, \Sigma) + \epsilon Q$$

- Covariance depth *[Chen-Gao-Ren, 2017]*.

$$\widehat{\Gamma} = \operatorname*{argmax}_{\Gamma > 0} \min_{\|u\|_2 = 1} \sum_{i=1}^{n} 1\left\{|u^T X_i|^2 > u^T \Gamma u\right\} \wedge \sum_{i=1}^{n} 1\left\{|u^T X_i|^2 \le u^T \Gamma u\right\}, \tag{1}$$

$$\widehat{\Sigma} = \frac{\widehat{\Gamma}}{\beta}, \mathbb{P}\left(\mathcal{N}(0,1) < \sqrt{\beta}\right) = \frac{3}{4}.$$

- $\|\widehat{\Sigma} - \Sigma\|_{op} \le C(\frac{p}{n} + \epsilon^2)$ with high probability uniformly over Σ and Q.

Computational Complexity

- Polynomial algorithms are proposed *[Lai et al., 2016; Diakonikolas et al., 2018]* of nearly minimax optimal statistical precision.
  - Prior knowledge on ϵ.
  - Needs some moment constraints.
- Advantages of the depth estimation.
  - Does not need prior knowledge on ϵ.
  - Adaptive to any elliptical distributions.
  - A well defined objective function.
  - Any feasible algorithms in practice?

f-divergence

Given a convex function f satisfying $f(1) = 0$, the f-divergence of P from Q is defined as,

$$D_f(P\|Q) = \int f\left(\frac{dP}{dQ}\right) dQ \tag{2}$$

Let f^* be the convex conjugate of f, then a variational lower bound of (2) is given by,

$$\begin{aligned} D_f(P\|Q) &= \int q(x) \sup_{t \in \mathrm{dom}_{f^*}} \left\{ t\frac{p(x)}{q(x)} - f^*(t) \right\} dx, \\ &\geq \sup_{T \in \mathcal{T}} \mathbb{E}_{x\sim P}\left[T(x)\right] - \mathbb{E}_{x\sim Q}\left[f^*\left(T(x)\right)\right]. \end{aligned} \tag{3}$$

- The equality holds in (3) if $f'\left(\frac{p}{q}\right) \in \mathcal{T}$.

$$D_f(P\|Q) \geq \max_{\tilde{Q} \in \tilde{\mathcal{Q}}} \frac{1}{n}\sum_{i=1}^{n} f'\left(\frac{\tilde{q}(X_i)}{q(X_i)}\right) - \mathbb{E}_{X\sim Q}\left[f^*\left(f'\left(\frac{\tilde{q}(X_i)}{q(X_i)}\right)\right)\right]. \tag{4}$$

f-GAN and f-Learning

- f-Learning. Let $\tilde{\mathcal{Q}}$ be a distribution family,

$$\widehat{P} = \underset{Q \in \mathcal{Q}}{\operatorname{argmin}} \max_{\tilde{Q} \in \tilde{\mathcal{Q}}} \frac{1}{n} \sum_{i=1}^{n} f'\left(\frac{\tilde{q}(X_i)}{q(X_i)}\right) - \mathbb{E}_{X \sim Q}\left[f^*\left(f'\left(\frac{\tilde{q}(X_i)}{q(X_i)}\right)\right)\right].$$

- f-GAN [Nowozin et al., 2016],

$$\widehat{P} = \underset{Q \in \mathcal{Q}}{\operatorname{argmin}} \max_{T \in \mathcal{T}} \frac{1}{n} \sum_{i=1}^{n} T(X_i) - \mathbb{E}_{X \sim Q}\left[f^*(T(x))\right],$$

where $\mathcal{T}$ is usually parametrized by a neural network.

- f-GAN can smooth f-Learning's objective function.
- f-divergence is robust.
- There exist practical efficient algorithms to solve.

Example

- f(x) $= x \log x$ (KL-divergence), $p \in \tilde{\mathcal{Q}}$ (or $f'(p/q) \in \mathcal{T}$), then KL-Learning (or KL-GAN) becomes maximal likelihood estimate.
- f(x) $= x \log x - (x+1) \log \frac{1+x}{2}$ (JS-divergence), which leads to the original JS-GAN [Goodfellow et al., 2014],

$$\widehat{P} = \operatorname*{argmin}_{Q \in \mathcal{Q}} \max_{T \in \mathcal{T}} \frac{1}{n} \sum_{i=1}^{n} \log\left(\text{sigmoid}\left(T(X_i)\right)\right) + \mathbb{E}_{x \sim Q} \log\left(1 - \text{sigmoid}\left(T(x)\right)\right).$$

Example (Continued)

- f(x) $= (x-1)_+$ (TV-divergence) and $f^*(t) = t, 0 \leq t \leq 1$.
  - When taking $\mathcal{Q} = \{\mathcal{N}(\theta, I_p) : \theta \in \mathbb{R}^p\}$, $\tilde{Q}(\theta, r) = \{\mathcal{N}(\tilde{\theta}, I_p) : \|\tilde{\theta} - \theta\|_2 \leq r\}$. TV-Learning is defined as,

$$\min_{Q \in \mathcal{Q}} \max_{\tilde{Q} \in \tilde{\mathcal{Q}}(\theta, r)} \frac{1}{n} \sum_{i=1}^{n} 1\left\{\frac{\tilde{q}(X_i)}{q(X_i)} \geq 1\right\} - Q\left(\frac{\tilde{q}}{q} \geq 1\right)$$

  - TV-Learning $\overset{r \to 0}{\to}$ Tukey median, $\max_{\eta \in \mathbb{R}^p} \min_{\|u\|_2 = 1} \sum_{i=1}^{n} 1\left\{u^T X_i > u^T \eta\right\}$.
  - With $\mathcal{T}$ parameterized by the class of neural networks, TV-GAN is defined as,

$$\widehat{P} = \operatorname*{argmin}_{Q \in \mathcal{Q}} \max_{T \in \mathcal{T}} \frac{1}{n} \sum_{i=1}^{n} \text{sigmoid}\left(T(X_i)\right) - \mathbb{E}_{x \sim Q}\left[\text{sigmoid}\left(T(x)\right)\right].$$

Proper Scoring Rule

- $\{S(\cdot,1), S(\cdot,0)\}$ is the forecaster's reward if a player quotes t when event 1 or 0 occurs.
- $S(t;p) = pS(t,1) + (1-p)S(t,0)$ is the expected reward when the event occurs with probability p.
- $\{S(\cdot,1), S(\cdot,0)\}$ is a proper scoring rule if

$$S(p;p) \geq S(t;p), \forall t \in [0,1].$$

- (Savage representation) S is proper iff there exists a convex function $G(\cdot)$ such that,

$$\begin{cases} S(t,1) = G(t) + (1-t)G'(t), \\ S(t,0) = G(t) - tG'(t). \end{cases}$$

Proper Scoring Rule and f-divergence

We consider a natural cost function with assumption $X|y = 1 \sim P$ and $X|y = 0 \sim Q$ with prior $\mathbb{P}(y = 1) = 1/2$, that is,

$$\mathbb{E}_{X\sim P}\frac{1}{2}S(T(X), 1) + \mathbb{E}_{X\sim Q}\frac{1}{2}S(T(X), 0).$$

Then one can find a good classification rule $T(\cdot)$ by maximizing the above objective over $T \in \mathcal{T}$,

$$D_{\mathcal{T}}(P, Q) = \max_{T\in\mathcal{T}}\left[\frac{1}{2}\mathbb{E}_{X\sim P}S(T(X), 1) + \frac{1}{2}\mathbb{E}_{X\sim Q}S(T(X), 0) - G(\frac{1}{2})\right]$$

- Log Score (JS-divergence). $S(t, 1) = \log t, S(t, 0) = \log(1 - t)$
- Zero-One Score (TV-divergence). $S(t, 1) = \mathbb{I}\{t \geq 1/2\}$, $S(t, 0) = \mathbb{I}\{t < 1/2\}$.

(Multi-layers) JS-GAN is Statistical Optimal

$$\widehat{\theta} = \mathop{\mathrm{argmin}}_{\eta \in \mathbb{R}^p} \max_{T \in \mathcal{T}} \left[\frac{1}{n} \sum_{i=1}^{n} \log T(X_i) + E_{\mathcal{N}(\eta, I_p)} \log(1 - T(X_i)) \right] + \log 4,$$

Theorem (Gao-Liu-Yao-Zhu' 2018)

With i.i.d. observations $X_1, \ldots, X_n \sim (1-\epsilon)N(\theta, I_p) + \epsilon Q$ and some regularizations on weight matrix, we have

$$\|\widehat{\theta} - \theta\|^2 \lesssim \begin{cases} \frac{p}{n} \vee \epsilon^2, & \text{at least one bounded activation} \\ \frac{p \log p}{n} \vee \epsilon^2, & \text{ReLU} \end{cases}$$

with high probability uniformly over all $\theta \in \mathbb{R}^p$ and all Q.

- It can be generalized to elliptical distribution $\mu + \Sigma^{1/2} \xi U$ and the strong contamination model.
- Covariance and mean can be estimated simultaneously.

Proof Sketch

- $\sup_{D\in\mathcal{D}} |E_{\mathbb{P}_n} D(X) - E_P D(X)| \le C\left(\sqrt{\frac{p}{n}} + \sqrt{\frac{\log(1/\delta)}{n}}\right).$
- $\sup_{D\in\mathcal{D}} |E_{P_\theta} D(X) - E_{P_{\hat\theta}}(D(X))| \le 2C\left(\sqrt{\frac{p}{n}} + \sqrt{\frac{\log(1/\delta)}{n}}\right) + 2\epsilon.$
- $|f(t) - f(0)| \ge c'|t|, |t| < \tau$ for some $\tau > 0$, where $f(t) = E_{N(0,1)}\,(\text{sigmoid}(z-t))$ satisfies,
  $E_{P_\theta} D(X) \xlongequal{\|w\|_2=1, b=-w^T\theta} f(0), E_{P_{\hat\theta}} D(X) = f(w^T(\theta - \hat\theta)).$

Covariance Matrix Estimation: Improper Network Structure

$$\mathcal{T}_1 = \left\{ T(x) = \text{sigmoid}\left(\sum_{j \geq 1} w_j \text{sigmoid}(u_j^T x) \right) : \sum_{j \geq 1} |w_j| \leq \kappa, u_j \in \mathbb{R}^p \right\}.$$

$$\mathcal{T}_2 = \left\{ T(x) = \text{sigmoid}\left(\sum_{j \geq 1} w_j \text{ReLU}(u_j^T x) \right) : \sum_{j \geq 1} |w_j| \leq \kappa, \|u_j\| \leq 1 \right\}.$$

Covariance Matrix Estimation: Proper Network Structure

$$\mathcal{T}_3 = \left\{ T(x) = \text{sigmoid}\left(\sum_{j \geq 1} w_j \text{sigmoid}(u_j^T x + b_j) \right) : \sum_{j \geq 1} |w_j| \leq \kappa, u_j \in \mathbb{R}^p, b_j \in \mathbb{R} \right\}.$$

$$\mathcal{T}_4 = \left\{ T(x) = \text{sigmoid}\left(\sum_{j \geq 1} w_j \text{sigmoid}\left(\sum_{l=1}^{H} v_{jl} \text{ReLU}(u_l^T x) \right) \right) : \sum_{j \geq 1} |w_j| \leq \kappa_1, \sum_{l=1}^{H} |v_{jl}| \leq \kappa_2, \|u_l\| \leq 1 \right\}.$$

$$\widehat{\Sigma} = \operatorname*{argmin}_{\Gamma \in \mathcal{E}_p(M)} \max_{T \in \mathcal{T}} \left[\frac{1}{n} \sum_{i=1}^{n} S(T(X_i), 1) + \mathbb{E}_{X \sim N(0,\Gamma)} S(T(X), 0) \right]$$

Theorem (Gao-Yao-Zhu' 2019)

With i.i.d. observations $X_1, ..., X_n \sim (1-\epsilon)N(0,\Sigma) + \epsilon Q$ and some regularizations on network weight matrix, we have

$$\|\widehat{\Sigma} - \Sigma\|_{\mathrm{op}}^2 \lesssim \frac{p}{n} \vee \epsilon^2$$

with high probability uniformly over all $\|\Sigma\|_{\mathrm{op}} \leq M = O(1)$ and all Q.

Experiments: Comparison

Q	n	p	ϵ	TV-GAN	JS-GAN	Dimension Halving	Iterative Filtering
$N(0.5 * 1_p, I_p)$	50,000	100	.2	**0.0953 (0.0064)**	0.1144 (0.0154)	0.3247 (0.0058)	0.1472 (0.0071)
$N(0.5 * 1_p, I_p)$	5,000	100	.2	**0.1941 (0.0173)**	0.2182 (0.0527)	0.3568 (0.0197)	0.2285 (0.0103)
$N(0.5 * 1_p, I_p)$	50,000	200	.2	**0.1108 (0.0093)**	0.1573 (0.0815)	0.3251 (0.0078)	0.1525 (0.0045)
$N(0.5 * 1_p, I_p)$	50,000	100	.05	0.0913 (0.0527)	0.1390 (0.0050)	0.0814 (0.0056)	**0.0530 (0.0052)**
$N(5 * 1_p, I_p)$	50,000	100	.2	2.7721 (0.1285)	**0.0534 (0.0041)**	0.3229 (0.0087)	0.1471 (0.0059)
$N(0.5 * 1_p, \Sigma)$	50,000	100	.2	0.1189 (0.0195)	**0.1148 (0.0234)**	0.3241 (0.0088)	0.1426 (0.0113)
Cauchy$(0.5 * 1_p)$	50,000	100	.2	0.0738 (0.0053)	**0.0525 (0.0029)**	0.1045 (0.0071)	0.0633 (0.0042)

Table: Comparison of various robust mean estimation methods. Samples $X_1, \ldots, X_n$ are drawn from $(1-\epsilon)\mathcal{N}(0, I_p) + \epsilon Q$ with (ϵ, Q) to be specified. Net structure: One-hidden layer network with 20 hidden units when $n = 50,000$ and 2 hidden units when $n = 5,000$. The number in each cell is the average of ℓ_2 error $\|\widehat{\theta} - \theta\|$ with standard deviation in parenthesis estimated from 10 repeated experiments and the smallest error among four methods is highlighted in bold.

- Dimension Halving *[Lai et al., 2016]*.
- Iterative Filtering *[Diakonikolas et al., 2018]*.

Experiments: Deeper May be Better in High-Dimension

p	200-100-20-1	200-200-100-1	200-100-1	200-20-1
200	0.0910 (0.0056)	**0.0790 (0.0026)**	0.3064 (0.0077)	0.1573 (0.0815)
p	400-200-100-50-20-1	400-200-100-20-1	400-200-20-1	400-200-1
400	0.1477 (0.0053)	0.1732 (0.0397)	**0.1393 (0.0090)**	0.3604 (0.0990)

Table: The samples are drawn independently from $(1-\epsilon)N(0_p, I_p) + \epsilon N(0.5 * 1_p, I_p)$ with $\epsilon = 0.2$, $p \in \{200, 400\}$ and $n = 50,000$.

Experiments: Generalization to Elliptical Distribution

- Elliptical distribution, $X \overset{d}{=} \theta + \xi A U$.
- Modifications on the Generator,
  - $G_1(\xi, U) = g_\omega(\xi) U + \theta$.
  - $G_2(\xi, U) = g_\omega(\xi) A U + \theta$.

Contamination Q	JS-GAN (G_1)	JS-GAN (G_2)	Dimension Halving	Iterative Filtering
Cauchy$(1.5 * 1_p, I_p)$	0.0664 (0.0065)	0.0743 (0.0103)	0.3529 (0.0543)	0.1244 (0.0114)
Cauchy$(5.0 * 1_p, I_p)$	0.0480 (0.0058)	0.0540 (0.0064)	0.4855 (0.0616)	0.1687 (0.0310)
Cauchy$(1.5 * 1_p, 5 * I_p)$	0.0754 (0.0135)	0.0742 (0.0111)	0.3726 (0.0530)	0.1220 (0.0112)
Normal$(1.5 * 1_p, 5 * I_p)$	0.0702 (0.0064)	0.0713 (0.0088)	0.3915 (0.0232)	0.1048 (0.0288))

Table: Comparison of various methods of robust location estimation under Cauchy distributions. Samples are drawn from $(1-\epsilon)\text{Cauchy}(0_p, I_p) + \epsilon Q$ with $\epsilon = 0.2, p = 50$ and various choices of Q. Sample size: 50,000. Discriminator net structure: 50-50-25-1. Generator $g_\omega(\xi)$ structure: 48-48-32-24-12-1 with absolute value activation function in the output layer.

Experiments: Tail Dependence

degrees of freedom v	$G_1(Z;A) = AZ$	$G_2(U,z;A,w_g) = g_{w_g}(z)AU$	Dimension Halving	Tyler's M-estimator	Kendall's τ	MVE
1	0.2808 (0.0440)	0.3350 (0.0681)	-	372.9637 (582.3385)	52.5653 (0.6361)	50.2995 (0.6259)
2	0.3450 (0.0157)	0.4059 (0.0254)	-	55.5152 (1.1901)	64.7625 (0.4798)	20.1941 (1.8645)
4	0.2751 (0.0147)	0.2775 (0.0456)	1.2834 (0.0512)	38.7569 (0.2740)	72.8037 (0.3369)	0.1920 (0.0299)
8	0.2131 (0.0162)	0.2113 (0.0306)	0.8902 (0.0728)	39.0265 (0.2014)	77.2117 (0.3486)	0.1753 (0.0218)
16	0.1764 (0.0120)	0.2076 (0.0210)	0.8354 (0.0926)	39.1167 (0.3200)	79.2252 (0.2728)	0.1683 (0.0136)
32	0.1576 (0.0067)	0.2056 (0.0202)	0.8572 (0.0687)	39.1985 (0.2153)	80.2075 (0.1706)	0.1493 (0.0085)

Table: Simulation results with $n = 50,000, p = 100, \epsilon = 0.2$ and $v \in \{1, 2, 4, 8, 16, 32\}$. We show the average error $\|\widehat{\Sigma} - \Sigma\|_{\mathrm{op}}$ in each cell with standard deviation in parenthesis from 10 repeated experiments.

Simultaneous Estimation

(P,Q)	$G_1(z;A)=Az$	$G_3(z;A,\mu)=Az+\mu$		$G_2(u,z;A,w_g)=g_{w_g}(z)Au$	$G_4(u,z;A,w_g,\mu)=g_{w_g}(z)Au+\mu$	
	$\|\widehat{\Sigma}-\Sigma\|_{\mathrm{op}}$	$\|\widehat{\Sigma}-\Sigma\|_{\mathrm{op}}$	$\|\widehat{\theta}-\theta\|$	$\|\widehat{\Sigma}-\Sigma\|_{\mathrm{op}}$	$\|\widehat{\Sigma}-\Sigma\|_{\mathrm{op}}$	$\|\widehat{\theta}-\theta\|$
$(N(0,I_p),N(5,5I_p))$	0.1615 (0.0134)	0.1537 (0.0155)	0.0508 (0.0054)	0.1624 (0.0141)	0.1694 (0.0105)	0.0519 (0.0048)
$(N(0,\Sigma_{\mathrm{ar}}),\delta_{4\mathcal{I}_p})$	0.1530 (0.0059)	0.1640 (0.0106)	0.0547 (0.0039)	0.1557 (0.0142)	0.1880 (0.0134)	0.0544 (0.0073)
$(T_1(0,\Sigma_{\mathrm{ar}}),T_1(5,5I_p))$	0.2808 (0.0440)	0.2512 (0.0479)	0.0656 (0.0065)	0.3350 (0.0681)	0.4678 (0.0498)	0.0575 (0.0048)
$(T_2(0,\Sigma_{\mathrm{ar}}),T_2(5,5I_p))$	0.3450 (0.0157)	0.3743 (0.0097)	0.0640 (0.0056)	0.4059 (0.0254)	0.4704 (0.0299)	0.0642 (0.0040)

Table: Simulation results with i.i.d. observations generated from $(1-\epsilon)P+\epsilon Q$, where $n=50,000$, $p=100$ and $\epsilon=0.2$. We show the average errors $\|\widehat{\Sigma}-\Sigma\|_{\mathrm{op}}$ and $\|\widehat{\theta}-\theta\|$ in each cell with standard deviation in parenthesis from 10 repeated experiments.

Future directions

- Provable robust GANs for regression.
- Application: Low rank recover, volatility matrix estimation, etc.
- Does it lead to an alternative approach against adversarial examples in neural networks?
- Does it lead to an explanation on mode collapse in GANs training?


Thank you!