

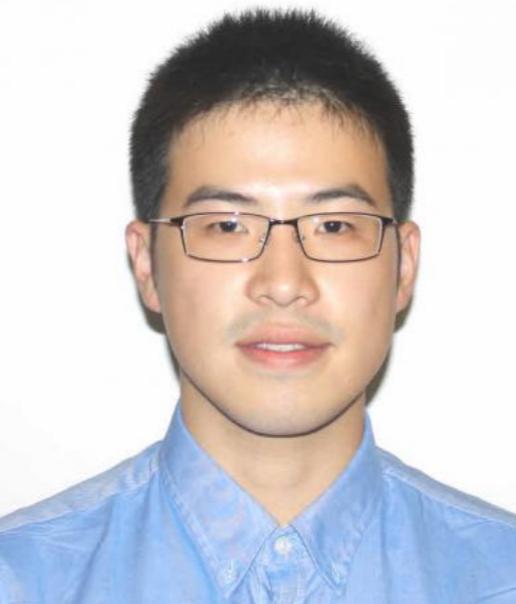
Robust Statistics and Generative Adversarial Networks

Yuan YAO
HKUST





Chao GAO
U. Chicago



Jiyi LIU
Yale U.



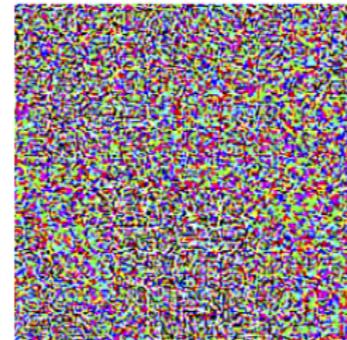
Weizhi ZHU
HKUST

Deep Learning is Notoriously Not Robust!



\mathbf{x}
“panda”
57.7% confidence

$+ .007 \times$



$\text{sign}(\nabla_{\mathbf{x}} J(\theta, \mathbf{x}, y))$
“nematode”
8.2% confidence

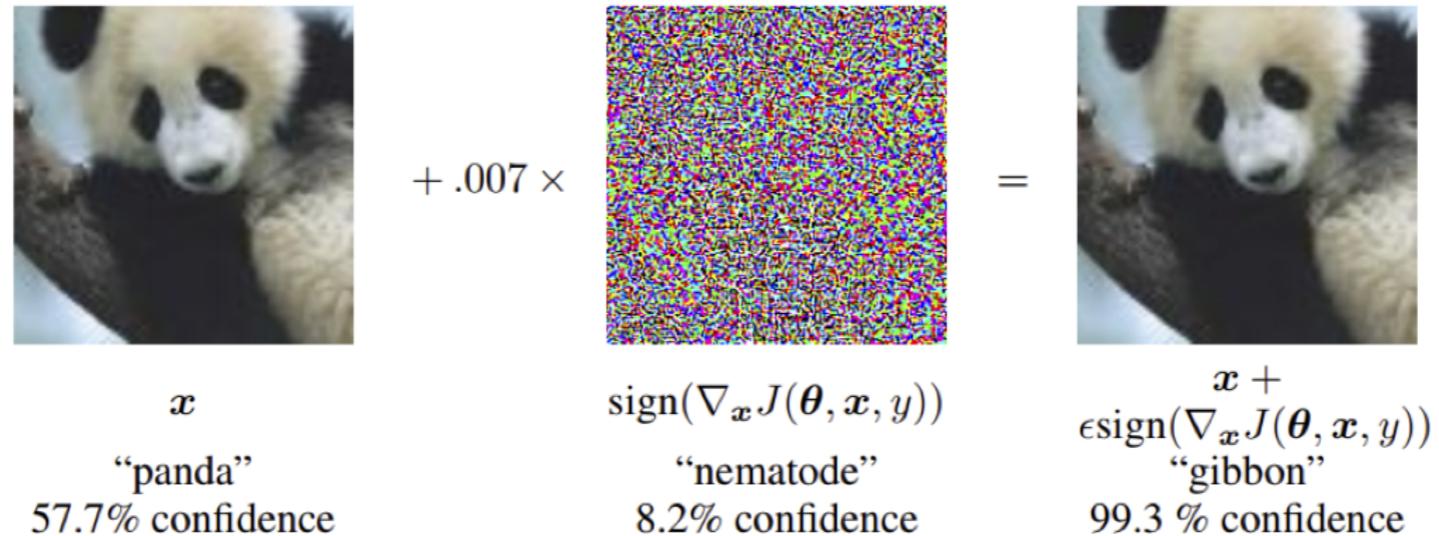
=



$\mathbf{x} +$
 $\epsilon \text{sign}(\nabla_{\mathbf{x}} J(\theta, \mathbf{x}, y))$
“gibbon”
99.3 % confidence

[Goodfellow et al., 2014]

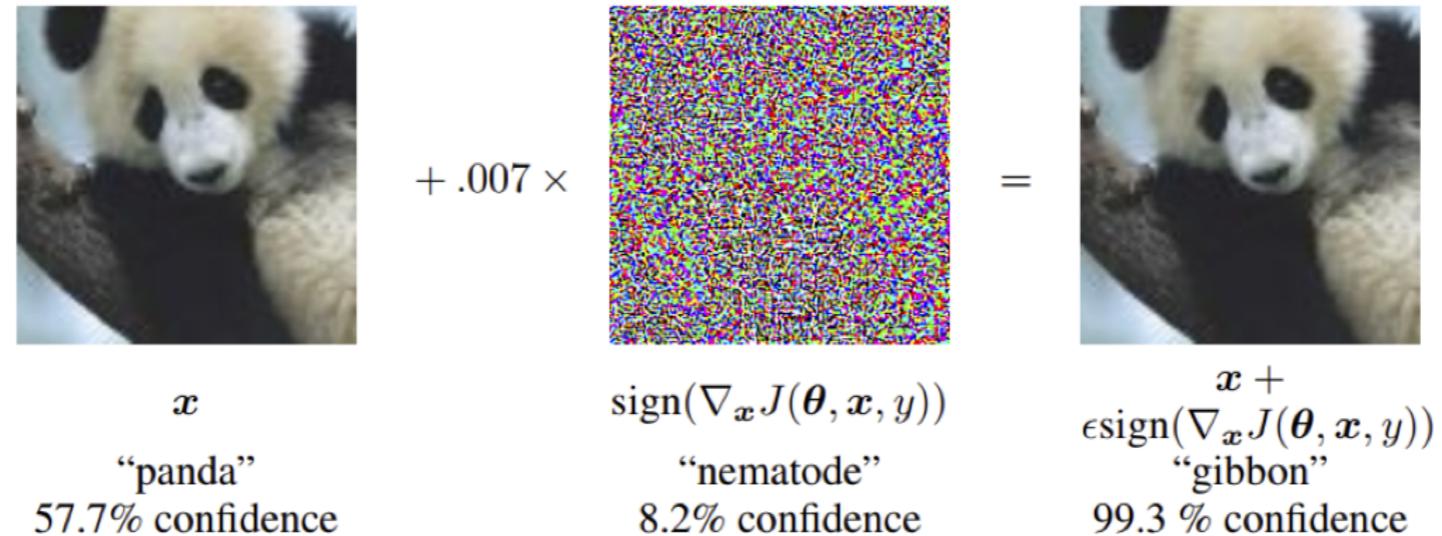
Deep Learning is Notoriously Not Robust!



[Goodfellow et al., 2014]

- Imperceptible adversarial examples are ubiquitous to fail neural networks

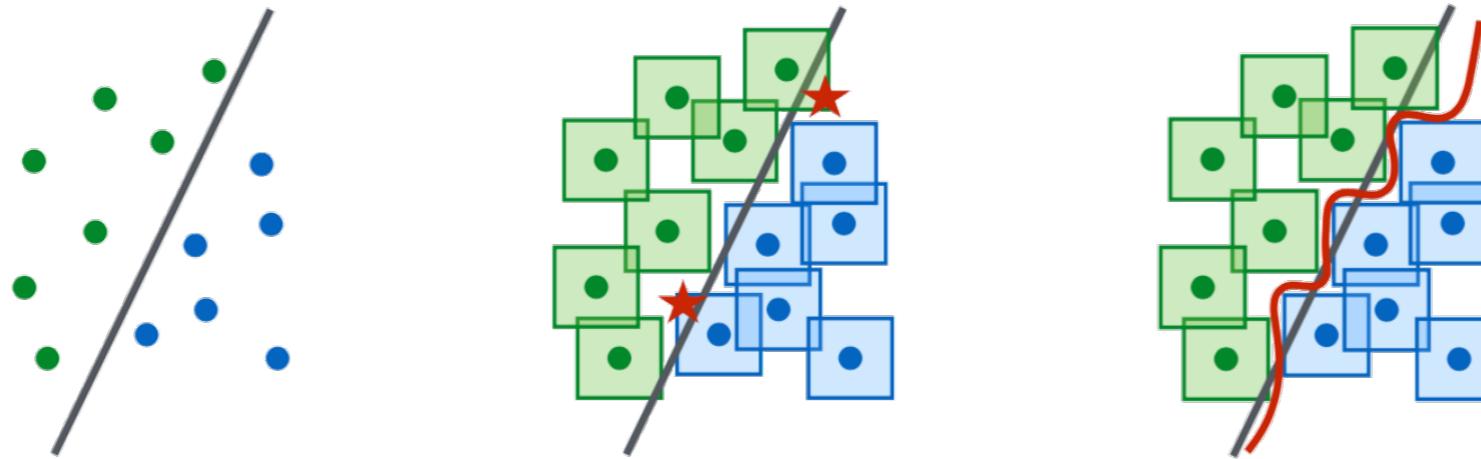
Deep Learning is Notoriously Not Robust!



[Goodfellow et al., 2014]

- Imperceptible adversarial examples are ubiquitous to fail neural networks
- How can one achieve **robustness**?

Robust Optimization



Madry's Adversarial Training

- Traditional training:

$$\min_{\theta} J_n(\theta, \mathbf{z} = (x_i, y_i)_{i=1}^n)$$

- e.g. square or cross-entropy loss as negative log-likelihood of logit models

- Robust optimization:

$$\min_{\theta} \max_{\|\epsilon_i\| \leq \delta} J_n(\theta, \mathbf{z} = (x_i + \epsilon_i, y_i)_{i=1}^n)$$

- robust to any distributions, yet perhaps too conservative

Distributionally Robust Optimization (DRO)

- Distributional Robust Optimization:

$$\min_{\theta} \max_{\epsilon} \mathbb{E}_{z \sim P_\epsilon \in \mathcal{D}} [J_n(\theta, z)]$$

- \mathcal{D} is a set of ambiguous distributions, e.g. Wasserstein ambiguity set

$$\mathcal{D} = \{P_\epsilon : W_2(P_\epsilon, \text{uniform distribution}) \leq \epsilon\}$$

where DRO may be reduced to **regularized maximum likelihood** estimates ([Shafieezadeh-Abadeh, Esfahani, Kuhn, NIPS'2015](#)) that are convex optimizations and tractable

Wasserstein DRO and Sqrt-Lasso (Jose Blanchet et al.'2016)

Theorem (B., Kang, Murthy (2016)) Suppose that

$$c((x, y), (x', y')) = \begin{cases} \|x - x'\|_q^2 & \text{if } y = y' \\ \infty & \text{if } y \neq y' \end{cases}.$$

Then, if $1/p + 1/q = 1$

$$\max_{P: D_c(P, P_n) \leq \delta} E_P^{1/2} \left((Y - \beta^\top X)^2 \right) = E_{P_n}^{1/2} \left[(Y - \beta^\top X)^2 \right] + \sqrt{\delta} \|\beta\|_p.$$

Remark 1: This is sqrt-Lasso (Belloni et al. (2011)).

Remark 2: Uses RoPA duality theorem & "**judicious choice of $c(\cdot)$** "

Certified Robustness of Lasso

Take $q = \infty$ and $p = 1$, with

$$c((x, y), (x', y')) = \begin{cases} \|x - x'\|_\infty^2 & \text{if } y = y' \\ \infty & \text{if } y \neq y' \end{cases}$$

Then for

$$P'_n = \frac{1}{n} \sum_i \delta_{x'_i}$$

with $\|x_i - x'_i\|_\infty \leq \delta$,

$$D_c(P'_n, P_n) = \int \pi((x, y), (x', y')) c((x, y), (x', y')) \leq \delta,$$

for small enough δ and well-separated x 's. Sqrt-Lasso

$$\begin{aligned} & \min_{\beta} \left\{ E_{P_n}^{1/2} \left[(Y - \beta^T X)^2 \right] + \sqrt{\delta} \|\beta\|_1 \right\}^2 \\ &= \min_{\beta} \max_{P: D_c(P, P_n) \leq \delta} E_P \left((Y - \beta^T X)^2 \right) \end{aligned}$$

provides a certified robust estimate in terms of Madry's adversarial training, using a convex Wasserstein relaxation.

TV-neighborhood

- How about the TV-uncertainty set?

$$\mathcal{D} = \{P_\epsilon : TV(P_\epsilon, \text{uniform distribution}) \leq \epsilon\}?$$

TV-neighborhood

- How about the TV-uncertainty set?

$$\mathcal{D} = \{P_\epsilon : TV(P_\epsilon, \text{uniform distribution}) \leq \epsilon\}?$$

TV-neighborhood

- How about the TV-uncertainty set?

$$\mathcal{D} = \{P_\epsilon : TV(P_\epsilon, \text{uniform distribution}) \leq \epsilon\}?$$

- an example from *robust statistics* ...

Huber's Model

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

[Huber 1964]

Huber's Model

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

parameter of interest

[Huber 1964]

Huber's Model

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

contamination proportion

parameter of interest

[Huber 1964]

Huber's Model

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

contamination proportion

parameter of interest

arbitrary contamination

[Huber 1964]

An Example

$$X_1, \dots, X_n \sim (1 - \epsilon)N(\theta, I_p) + \epsilon Q.$$

An Example

$$X_1, \dots, X_n \sim (1 - \epsilon)N(\theta, I_p) + \epsilon Q.$$

how to estimate ?

Robust Maximum-Likelihood

Does not work!

$$X_1, \dots, X_n \sim (1 - \epsilon)N(\theta, I_p) + \epsilon Q.$$

Robust Maximum-Likelihood

Does not work!

$$X_1, \dots, X_n \sim (1 - \epsilon)N(\theta, I_p) + \epsilon Q.$$

$$\begin{aligned}\ell(\theta, Q) &= \text{negative log-likelihood} = \sum_{i=1}^n (\theta - X_i)^2 \\ &\sim (1 - \epsilon)\mathbb{E}_{N(\theta)}(\theta - X)^2 + \epsilon\mathbb{E}_Q(\theta - X)^2\end{aligned}$$

the sample mean

$$\hat{\theta}_{mean} = \frac{1}{n} \sum_{i=1}^n X_i = \arg \min_{\theta} \ell(\theta, Q)$$

$$\min_{\theta} \max_Q \ell(\theta, Q) \geq \max_Q \min_{\theta} \ell(\theta, Q) = \max_Q \ell(\hat{\theta}_{mean}, Q) = \infty$$

Medians

1. Coordinatewise median

$\hat{\theta} = (\hat{\theta}_j)$, where $\hat{\theta}_j = \text{Median}(\{X_{ij}\}_{i=1}^n)$;

Medians

1. Coordinatewise median

$\hat{\theta} = (\hat{\theta}_j)$, where $\hat{\theta}_j = \text{Median}(\{X_{ij}\}_{i=1}^n)$;

2. Tukey's median

$$\hat{\theta} = \arg \max_{\eta \in \mathbb{R}^p} \min_{\|u\|=1} \frac{1}{n} \sum_{i=1}^n \mathbb{I}\{u^T X_i > u^T \eta\}.$$

Multivariate Location Depth

$$\text{Definition: } \left\{ \frac{1}{n} \sum_{i=1}^n \mathbb{I}\{u^T X_i > u^T \eta\} \wedge \frac{1}{n} \sum_{i=1}^n \mathbb{I}\{u^T X_i \leq u^T \eta\} \right\}$$

Multivariate Location Depth

$$\min_{\|u\|=1} \left\{ \frac{1}{n} \sum_{i=1}^n \mathbb{I}\{u^T X_i > u^T \eta\} \wedge \frac{1}{n} \sum_{i=1}^n \mathbb{I}\{u^T X_i \leq u^T \eta\} \right\}$$

Multivariate Location Depth

$$\hat{\theta} = \arg \max_{\eta \in \mathbb{R}^p} \min_{\|u\|=1} \left\{ \frac{1}{n} \sum_{i=1}^n \mathbb{I}\{u^T X_i > u^T \eta\} \wedge \frac{1}{n} \sum_{i=1}^n \mathbb{I}\{u^T X_i \leq u^T \eta\} \right\}$$

Multivariate Location Depth

$$\hat{\theta} = \arg \max_{\eta \in \mathbb{R}^p} \min_{\|u\|=1} \left\{ \frac{1}{n} \sum_{i=1}^n \mathbb{I}\{u^T X_i > u^T \eta\} \wedge \frac{1}{n} \sum_{i=1}^n \mathbb{I}\{u^T X_i \leq u^T \eta\} \right\}$$

$$= \arg \max_{\eta \in \mathbb{R}^p} \min_{\|u\|=1} \frac{1}{n} \sum_{i=1}^n \mathbb{I}\{u^T X_i > u^T \eta\}.$$

[Tukey, 1975]

Comparisons

	Coordinatewise Median	Tukey's Median
breakdown point	1/2	1/3
statistical precision (no contamination)	$\frac{p}{n}$	$\frac{p}{n}$
statistical precision (with contamination)	$\frac{p}{n} + p\epsilon^2$	$\frac{p}{n} + \epsilon^2$: minimax [Chen-Gao-Ren'15]
computational complexity	Polynomial	NP-hard [Amenta et al. '00]

Note: R-package for Tukey median can not deal with more than **10** dimensions!

[<https://github.com/ChenMengjie/DepthDescent>]

Computation

Computational Challenges

$$X_1, \dots, X_n \sim (1 - \epsilon)N(\theta, I_p) + \epsilon Q.$$

Computational Challenges

$$X_1, \dots, X_n \sim (1 - \epsilon)N(\theta, I_p) + \epsilon Q.$$

Lai, Rao, Vempala
Diakonikolas, Kamath, Kane, Li, Moitra, Stewart
Balakrishnan, Du, Singh

Computational Challenges

$$X_1, \dots, X_n \sim (1 - \epsilon)N(\theta, I_p) + \epsilon Q.$$

Lai, Rao, Vempala
Diakonikolas, Kamath, Kane, Li, Moitra, Stewart
Balakrishnan, Du, Singh

- Polynomial algorithms are proposed [[Diakonikolas et al.'16](#), [Lai et al. 16](#)] of minimax optimal statistical precision
 - needs information on second or higher order of moments
 - some priori knowledge about ϵ

Advantages of Tukey Median

-

Advantages of Tukey Median

- A well-defined objective function

Advantages of Tukey Median

- **A well-defined objective function**
- **Adaptive to ϵ and Σ**

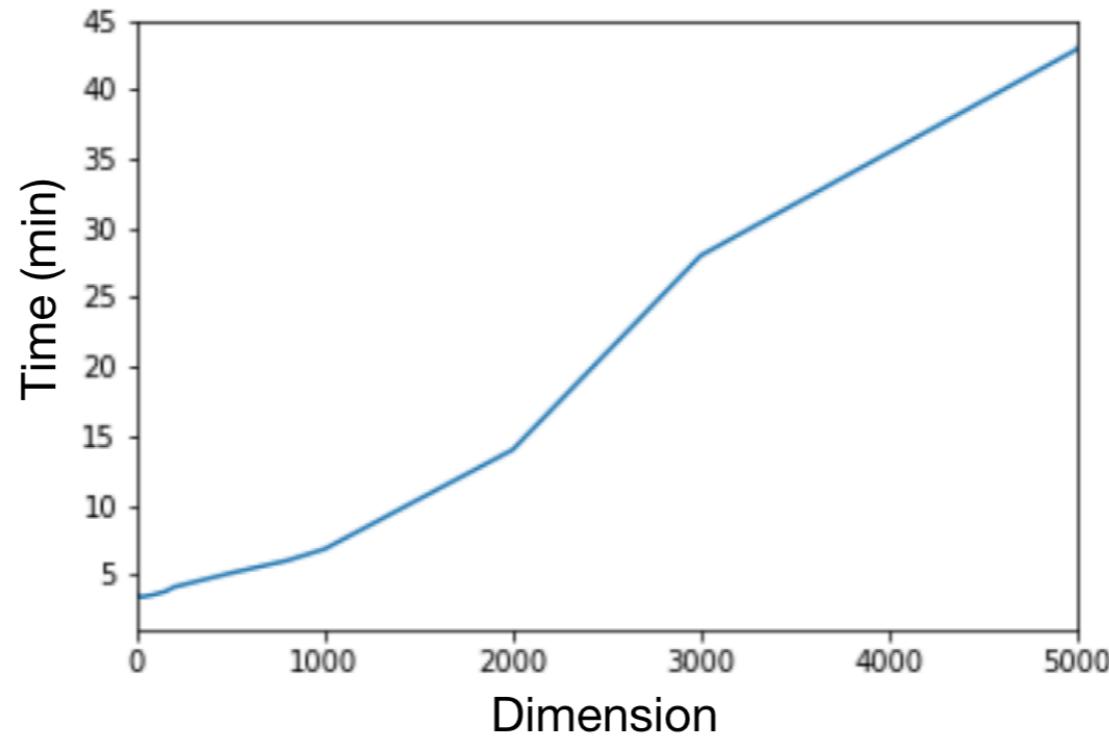
Advantages of Tukey Median

- A well-defined objective function
- Adaptive to ϵ and Σ
- Optimal for any elliptical distribution

A practically good algorithm?

Generative Adversarial Networks

[Goodfellow et al. 2014]



Note: R-package for Tukey median can not deal with more than 10 dimensions [<https://github.com/ChenMengjie/DepthDescent>]

Robust Learning of Cauchy Distributions

Table 4: 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.

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))

- *Dimension Halving:* [Lai et al.'16]
<https://github.com/kal2000/AgnosticMeanAndCovarianceCode>.
- *Iterative Filtering:* [Diakonikolas et al.'17]
<https://github.com/hoonose/robust-filter>.

f-GAN

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

$$D_f(P\|Q) = \int f\left(\frac{p}{q}\right) dQ. \quad (8)$$

Let f^* be the convex conjugate of f . A variational lower bound of (8) is

$$D_f(P\|Q) \geq \sup_{T \in \mathcal{T}} [\mathbb{E}_P T(X) - \mathbb{E}_Q f^*(T(X))]. \quad (9)$$

where equality holds whenever the class \mathcal{T} contains the function $f'(p/q)$.

[Nowozin-Cseke-Tomioka'16] f-GAN minimizes the variational lower bound (9)

$$\hat{P} = \arg \min_{Q \in \mathcal{Q}} \sup_{T \in \mathcal{T}} \left[\frac{1}{n} \sum_{i=1}^n T(X_i) - \mathbb{E}_Q f^*(T(X)) \right]. \quad (10)$$

with i.i.d. observations $X_1, \dots, X_n \sim P$.

From f-GAN to Tukey's Median: f-learning

Consider the special case

$$\mathcal{T} = \left\{ f' \left(\frac{\tilde{q}}{q} \right) : \tilde{q} \in \tilde{\mathcal{Q}} \right\}. \quad (11)$$

which is tight if $P \in \tilde{\mathcal{Q}}$. The sample version leads to the following f -learning

$$\hat{P} = \arg \min_{Q \in \mathcal{Q}} \sup_{\tilde{Q} \in \tilde{\mathcal{Q}}} \left[\frac{1}{n} \sum_{i=1}^n f' \left(\frac{\tilde{q}(X_i)}{q(X_i)} \right) - \mathbb{E}_Q f^* \left(f' \left(\frac{\tilde{q}(X)}{q(X)} \right) \right) \right]. \quad (12)$$

- If $f(x) = x \log x$, $\mathcal{Q} = \tilde{\mathcal{Q}}$, (12) \Rightarrow Maximum Likelihood Estimate
- If $f(x) = (x - 1) +$, then $D_f(P||Q) = \frac{1}{2} \int |p - q|$ is the TV-distance,
 $f^*(t) = t \mathbb{I}\{0 \leq t \leq 1\}$, f-GAN \Rightarrow TV-GAN
- $\mathcal{Q} = \{N(\eta, I_p) : \eta \in \mathbb{R}^p\}$ and $\tilde{\mathcal{Q}} = \{N(\tilde{\eta}, I_p) : \|\tilde{\eta} - \eta\| \leq r\}$, (12) $\stackrel{r \rightarrow 0}{\Rightarrow}$ Tukey's Median

TV-GAN

$$\widehat{\theta} = \operatorname{argmin}_{\eta} \sup_{w,b} \left[\frac{1}{n} \sum_{i=1}^n \frac{1}{1 + e^{-w^T X_i - b}} - E_{\eta} \frac{1}{1 + e^{-w^T X - b}} \right]$$

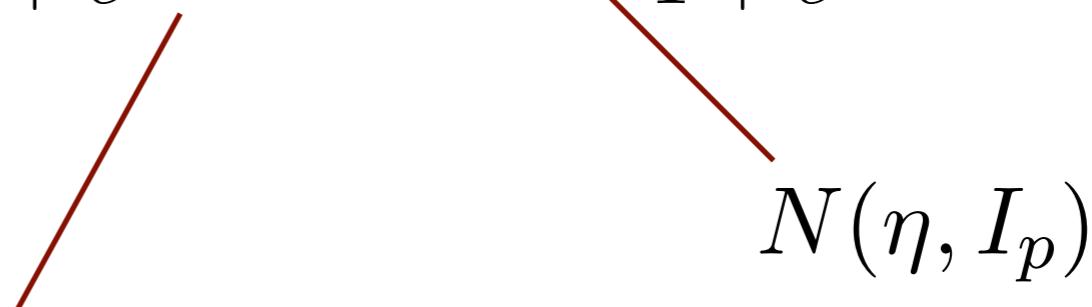
TV-GAN

$$\widehat{\theta} = \operatorname{argmin}_{\eta} \sup_{w,b} \left[\frac{1}{n} \sum_{i=1}^n \frac{1}{1 + e^{-w^T X_i - b}} - E_{\eta} \frac{1}{1 + e^{-w^T X - b}} \right]$$

$N(\eta, I_p)$

TV-GAN

$$\hat{\theta} = \operatorname{argmin}_{\eta} \sup_{w,b} \left[\frac{1}{n} \sum_{i=1}^n \frac{1}{1 + e^{-w^T X_i - b}} - E_{\eta} \frac{1}{1 + e^{-w^T X - b}} \right]$$


$$N(\eta, I_p)$$

logistic regression classifier

TV-GAN

$$\hat{\theta} = \operatorname{argmin}_{\eta} \sup_{w,b} \left[\frac{1}{n} \sum_{i=1}^n \frac{1}{1 + e^{-w^T X_i - b}} - E_{\eta} \frac{1}{1 + e^{-w^T X - b}} \right]$$


logistic regression classifier


 $N(\eta, I_p)$

Theorem [GLYZ18]. For some $C > 0$,

$$\|\hat{\theta} - \theta\|^2 \leq C \left(\frac{p}{n} \vee \epsilon^2 \right)$$

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

Yet TV-GAN has a rugged landscape!

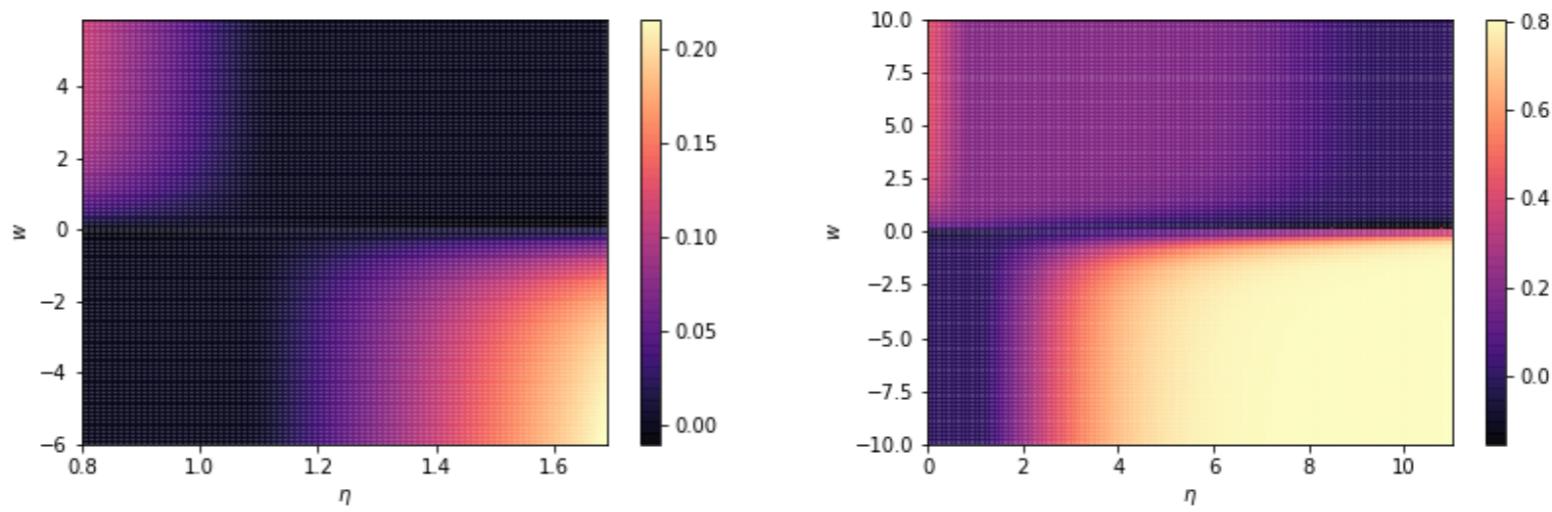


Figure: Heatmaps of the landscape of $F(\eta, w) = \sup_b [E_P \text{sigmoid}(wX + b) - E_{N(\eta, 1)} \text{sigmoid}(wX + b)]$, where b is maximized out for visualization. Left: samples are drawn from $P = (1 - \epsilon)N(1, 1) + \epsilon N(1.5, 1)$ with $\epsilon = 0.2$. Right: samples are drawn from $P = (1 - \epsilon)N(1, 1) + \epsilon N(10, 1)$ with $\epsilon = 0.2$. Left: the landscape is good in the sense that no matter whether we start from the left-top area or the right-bottom area of the heatmap, gradient ascent on η does not consistently increase or decrease the value of η . This is because the signal becomes weak when it is close to the saddle point around $\eta = 1$. Right: it is clear that $\tilde{F}(w) = F(\eta, w)$ has two local maxima for a given η , achieved at $w = +\infty$ and $w = -\infty$. In fact, the global maximum for $\tilde{F}(w)$ has a phase transition from $w = +\infty$ to $w = -\infty$ as η grows. For example, the maximum is achieved at $w = +\infty$ when $\eta = 1$ (blue solid) and is achieved at $w = -\infty$ when $\eta = 5$ (red solid). Unfortunately, even if we initialize with $\eta_0 = 1$ and $w_0 > 0$, gradient ascents on η will only increase the value of η (green dash), and thus as long as the discriminator cannot reach the global maximizer, w will be stuck in the positive half space $\{w : w > 0\}$ and further increase the value of η .

The Original JS-GAN

[Goodfellow et al. 2014] For $f(x) = x \log x - (x + 1) \log \frac{x+1}{2}$,

$$\hat{\theta} = \arg \min_{\eta \in \mathbb{R}^p} \max_{D \in \mathcal{D}} \left[\frac{1}{n} \sum_{i=1}^n \log D(X_i) + \mathbb{E}_{\mathcal{N}(\eta, I_p)} \log(1 - D(X)) \right] + \log 4. \quad (15)$$

What are \mathcal{D} , the class of discriminators?

- Single layer (no hidden layer):

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

- One-hidden or Multiple layer:

$$\mathcal{D} = \left\{ D(x) = \text{sigmoid}(w^T g(X)) \right\}$$

JS-GAN

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



JS-GAN

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

numerical experiment

$$X_1, \dots, X_n \sim (1 - \epsilon)N(\theta, I_p) + \epsilon N(\tilde{\theta}, I_p)$$

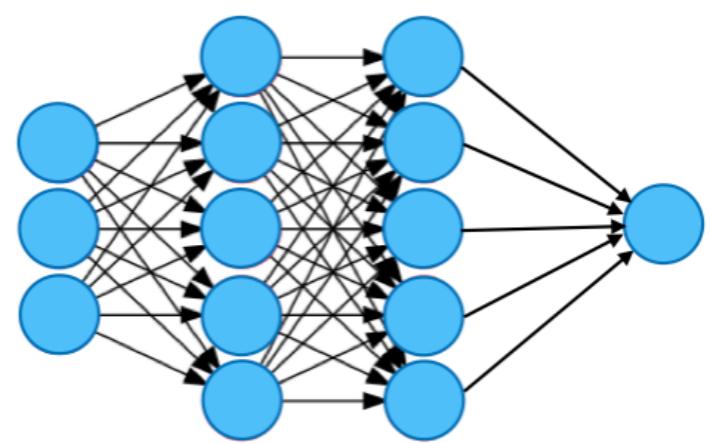
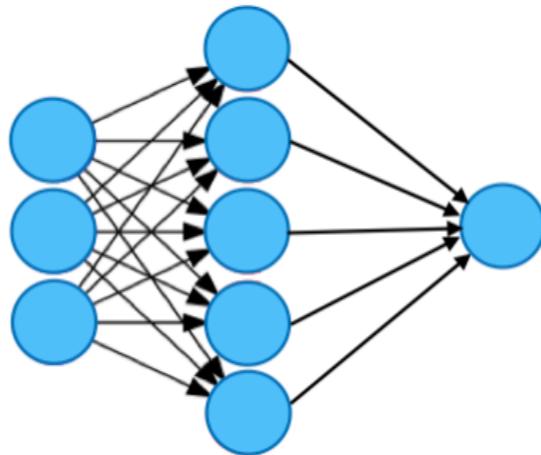
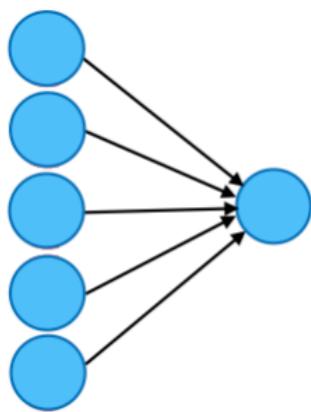


JS-GAN

$$\hat{\theta} = \operatorname{argmin}_{\eta \in \mathbb{R}^p} \max_{T \in \mathcal{T}} \left[\frac{1}{n} \sum_{i=1}^n \log T(X_i) + E_\eta \log(1 - T(X)) \right] + \log 4$$

numerical experiment

$$X_1, \dots, X_n \sim (1 - \epsilon)N(\theta, I_p) + \epsilon N(\tilde{\theta}, I_p)$$

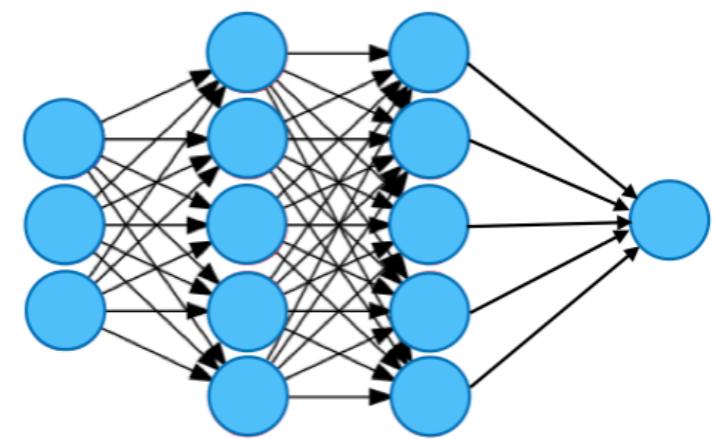
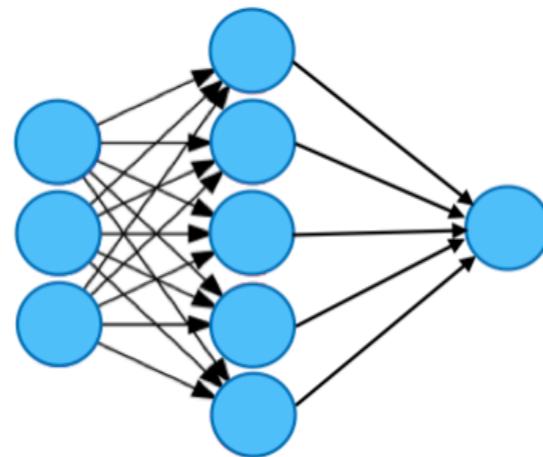
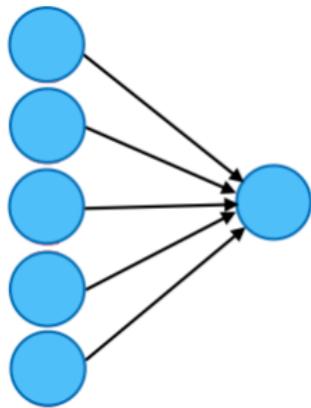


JS-GAN

$$\hat{\theta} = \operatorname{argmin}_{\eta \in \mathbb{R}^p} \max_{T \in \mathcal{T}} \left[\frac{1}{n} \sum_{i=1}^n \log T(X_i) + E_\eta \log(1 - T(X)) \right] + \log 4$$

numerical experiment

$$X_1, \dots, X_n \sim (1 - \epsilon)N(\theta, I_p) + \epsilon N(\tilde{\theta}, I_p)$$



$$\hat{\theta} \approx (1 - \epsilon)\theta + \epsilon\tilde{\theta}$$

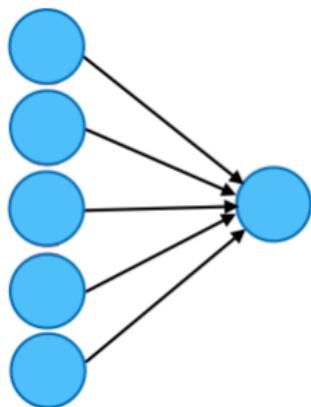


JS-GAN

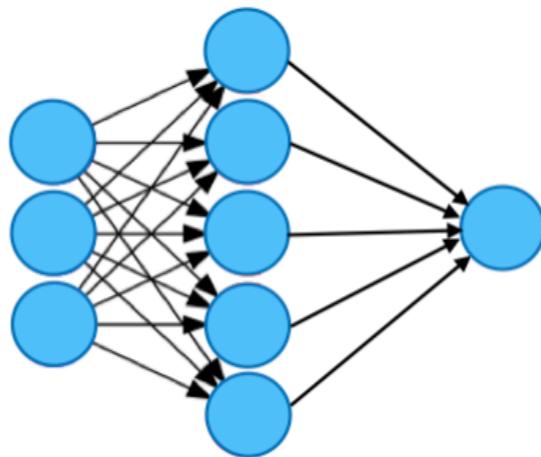
$$\widehat{\theta} = \operatorname{argmin}_{\eta \in \mathbb{R}^p} \max_{T \in \mathcal{T}} \left[\frac{1}{n} \sum_{i=1}^n \log T(X_i) + E_\eta \log(1 - T(X)) \right] + \log 4$$

numerical experiment

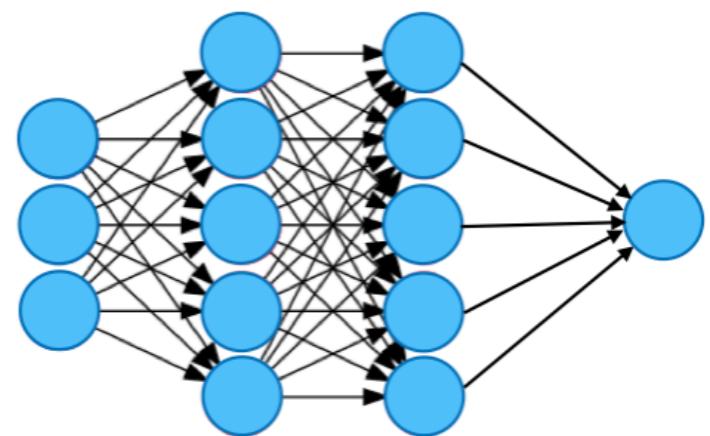
$$X_1, \dots, X_n \sim (1 - \epsilon)N(\theta, I_p) + \epsilon N(\tilde{\theta}, I_p)$$



$$\widehat{\theta} \approx (1 - \epsilon)\theta + \epsilon\widetilde{\theta}$$



$$\hat{\theta} \approx \theta$$



$$\widehat{\theta} \approx \theta$$

JS-GAN

A classifier with hidden layers leads to robustness. Why?

JS-GAN

A classifier with hidden layers leads to robustness. Why?

$$\text{JS}_g(\mathbb{P}, \mathbb{Q}) = \max_{w \in \mathbb{R}^d} \left[\mathbb{P} \log \frac{1}{1 + e^{-w^T g(X)}} + \mathbb{Q} \log \frac{1}{1 + e^{w^T g(X)}} \right] + \log 4.$$

JS-GAN

A classifier with hidden layers leads to robustness. Why?

$$\text{JS}_g(\mathbb{P}, \mathbb{Q}) = \max_{w \in \mathbb{R}^d} \left[\mathbb{P} \log \frac{1}{1 + e^{-w^T g(X)}} + \mathbb{Q} \log \frac{1}{1 + e^{w^T g(X)}} \right] + \log 4.$$

Proposition.

$$\text{JS}_g(\mathbb{P}, \mathbb{Q}) = 0 \iff \mathbb{P}g(X) = \mathbb{Q}g(X)$$

JS-GAN

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

Theorem [GLYZ18]. For a neural network class \mathcal{T} with at least one hidden layer and appropriate regularization, we have

$$\|\widehat{\theta} - \theta\|^2 \lesssim \begin{cases} \frac{p}{n} + \epsilon^2 & \text{(indicator/sigmoid/ramp)} \\ \frac{p \log p}{n} + \epsilon^2 & \text{(ReLUs+sigmoid features)} \end{cases}$$

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

Discriminator identifies outliers

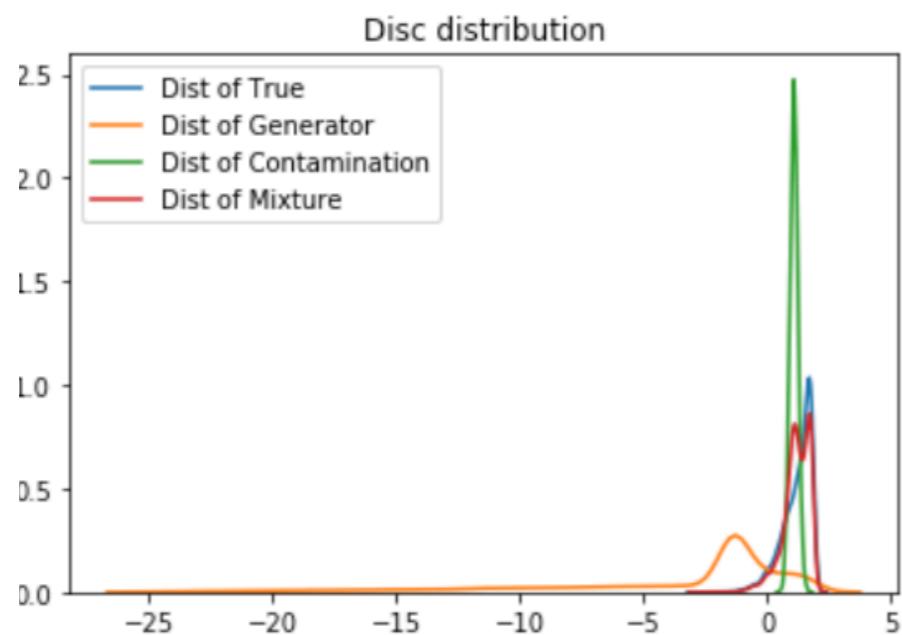
$$(1 - \epsilon)N(0_p, I_p) + \epsilon Q$$

Discriminator identifies outliers

$$(1 - \epsilon)N(0_p, I_p) + \epsilon Q \xleftarrow{N(5 * 1_p, I_p)}$$

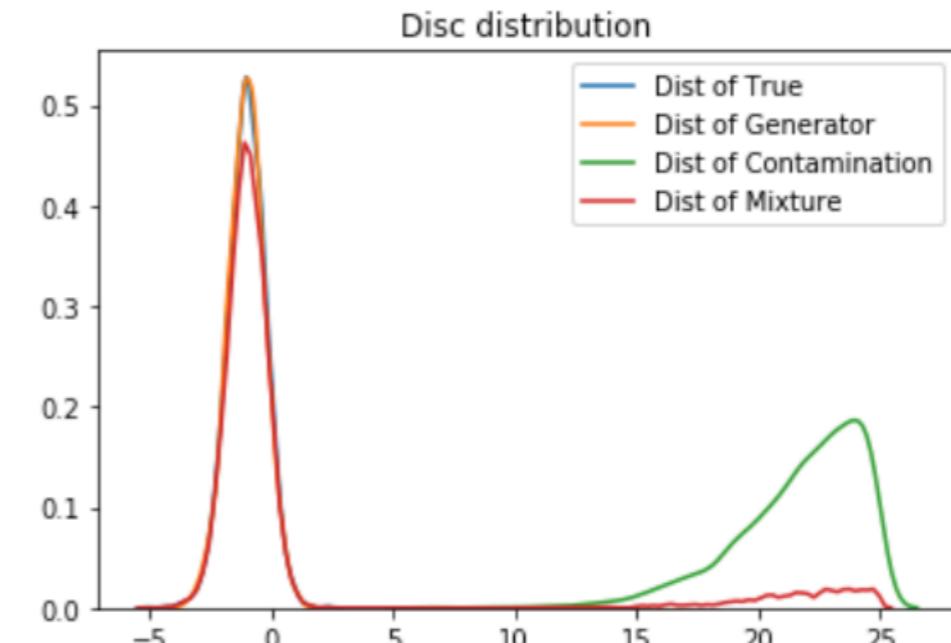
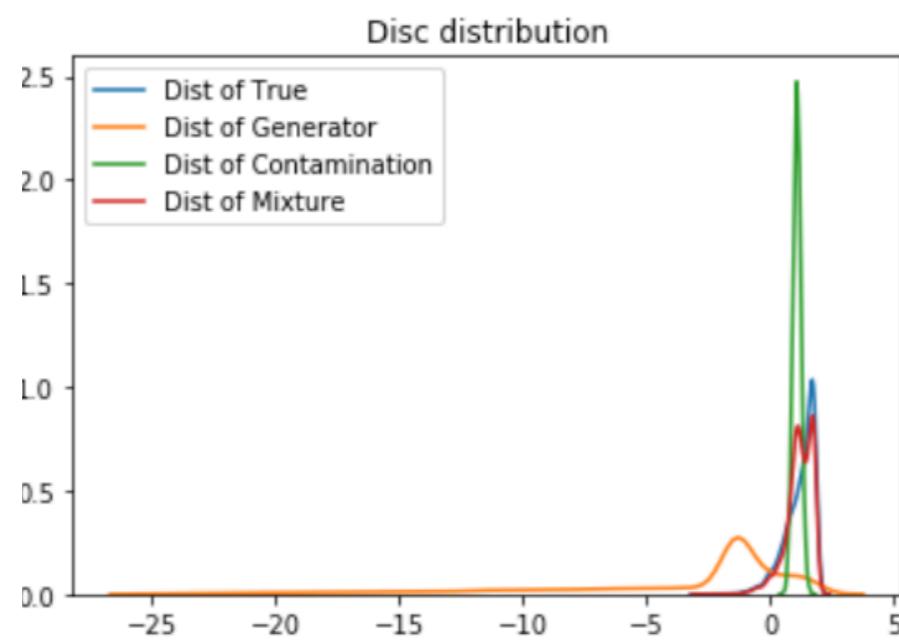
Discriminator identifies outliers

$$(1 - \epsilon)N(0_p, I_p) + \epsilon Q \xrightarrow{N(5 * 1_p, I_p)}$$



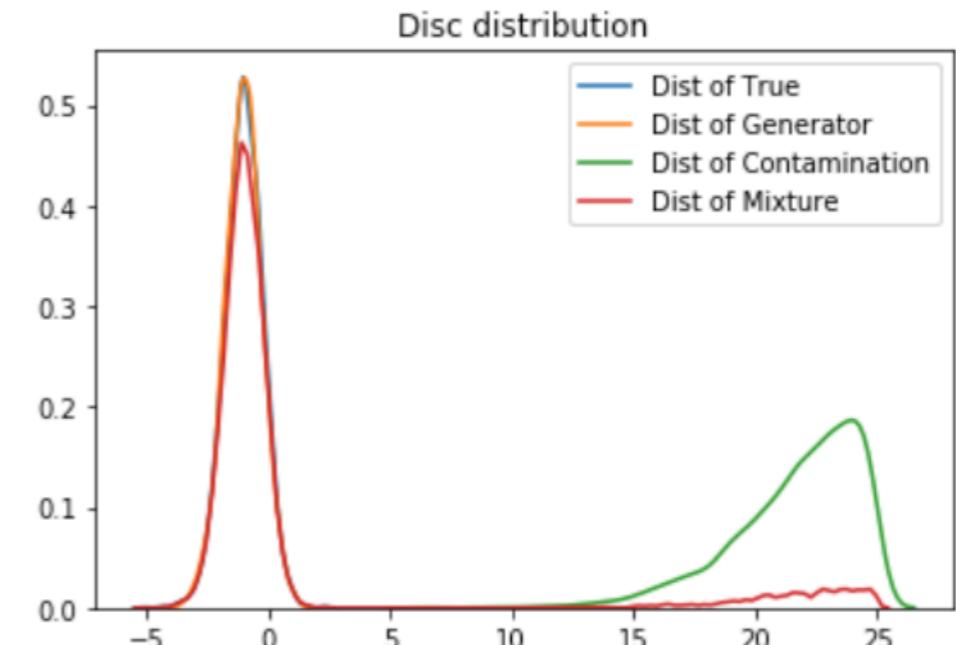
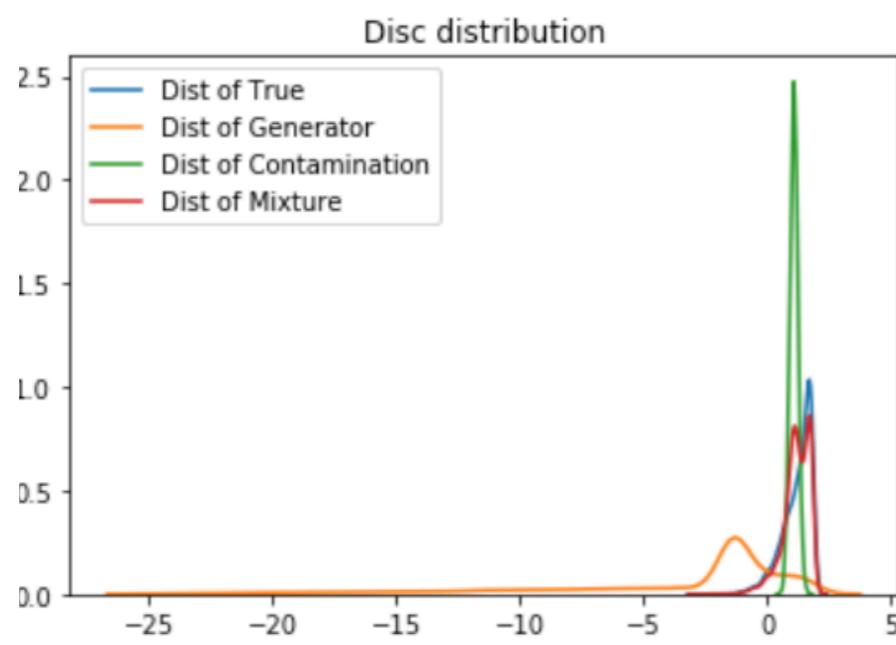
Discriminator identifies outliers

$$(1 - \epsilon)N(0_p, I_p) + \epsilon Q \xleftarrow{N(5 * 1_p, I_p)}$$



Discriminator identifies outliers

$$(1 - \epsilon)N(0_p, I_p) + \epsilon Q \xleftarrow{N(5 * 1_p, I_p)}$$



- Discriminator helps identify outliers or contaminated samples
- Generator fits uncontaminated portion of true samples

JS-GAN: Adaptation to Unknown Covariance

**unknown
covariance?**

$$X_1, \dots, X_n \sim (1 - \epsilon)N(\theta, \Sigma) + \epsilon Q$$

JS-GAN: Adaptation to Unknown Covariance

unknown covariance?

$$X_1, \dots, X_n \sim (1 - \epsilon)N(\theta, \Sigma) + \epsilon Q$$

$$(\hat{\theta}, \hat{\Sigma}) = \operatorname{argmin}_{\eta, \Gamma} \max_{T \in \mathcal{T}} \left[\frac{1}{n} \sum_{i=1}^n \log T(X_i) + \mathbb{E}_{X \sim N(\eta, \Gamma)} \log(1 - T(X)) \right]$$

JS-GAN: Adaptation to Unknown Covariance

unknown covariance?

$$X_1, \dots, X_n \sim (1 - \epsilon)N(\theta, \Sigma) + \epsilon Q$$

$$(\hat{\theta}, \hat{\Sigma}) = \operatorname{argmin}_{\eta, \Gamma} \max_{T \in \mathcal{T}} \left[\frac{1}{n} \sum_{i=1}^n \log T(X_i) + \mathbb{E}_{X \sim N(\eta, \Gamma)} \log(1 - T(X)) \right]$$

no need to change the discriminator class

Generalization

Strong Contamination model:

$X_1, \dots, X_n \stackrel{iid}{\sim} P$ for some P satisfying $\text{TV}(P, E(\theta, \Sigma, H)) \leq \epsilon$

Generalization

Strong Contamination model:

$X_1, \dots, X_n \stackrel{iid}{\sim} P$ for some P satisfying $\text{TV}(P, E(\theta, \Sigma, H)) \leq \epsilon$

$$(\hat{\theta}, \hat{\Sigma}, \hat{H}) = \underset{\eta \in \mathbb{R}^p, \Gamma \in \mathcal{E}_p(M), H \in \mathcal{H}(M')}{\operatorname{argmin}} \max_{T \in \mathcal{T}} \left[\frac{1}{n} \sum_{i=1}^n S(T(X_i), 1) + \mathbb{E}_{X \sim E(\eta, \Gamma, G)} S(T(X), 0) \right]$$

Generalization

Strong Contamination model:

$X_1, \dots, X_n \stackrel{iid}{\sim} P$ for some P satisfying $\text{TV}(P, E(\theta, \Sigma, H)) \leq \epsilon$

$$(\hat{\theta}, \hat{\Sigma}, \hat{H}) = \underset{\eta \in \mathbb{R}^p, \Gamma \in \mathcal{E}_p(M), H \in \mathcal{H}(M')}{\operatorname{argmin}} \max_{T \in \mathcal{T}} \left[\frac{1}{n} \sum_{i=1}^n S(T(X_i), 1) + \mathbb{E}_{X \sim E(\eta, \Gamma, G)} S(T(X), 0) \right]$$

A scoring rule S is *regular* if both $S(\cdot, 0)$ and $S(\cdot, 1)$ are real-valued, except possibly that $S(0, 1) = -\infty$ or $S(1, 0) = -\infty$. The celebrated Savage representation [50] asserts that a regular scoring rule S is proper if and only if there is a convex function $G(\cdot)$, such that

$$\begin{cases} S(t, 1) = G(t) + (1-t)G'(t), \\ S(t, 0) = G(t) - tG'(t). \end{cases} \quad (10)$$

Here, $G'(t)$ is a subgradient of G at the point t . Moreover, the statement also holds for strictly proper scoring rules when convex is replaced by strictly convex.

Statistical Optimality

Theorem [GYZ19]. For a neural network class \mathcal{T} with at least one hidden layer and appropriate regularization, we have

$$\|\hat{\theta} - \theta\|^2 \leq C \left(\frac{p}{n} \vee \epsilon^2 \right),$$

$$\|\hat{\Sigma} - \Sigma\|_{\text{op}}^2 \leq C \left(\frac{p}{n} \vee \epsilon^2 \right),$$

Example 1: Log Score and JS-GAN

1. *Log Score.* The log score is perhaps the most commonly used rule because of its various intriguing properties [31]. The scoring rule with $S(t, 1) = \log t$ and $S(t, 0) = \log(1 - t)$ is regular and strictly proper. Its Savage representation is given by the convex function $G(t) = t \log t + (1 - t) \log(1 - t)$, which is interpreted as the negative Shannon entropy of Bernoulli(t). The corresponding divergence function $D_{\mathcal{T}}(P, Q)$, according to Proposition 3.1, is a variational lower bound of the Jensen-Shannon divergence

$$\text{JS}(P, Q) = \frac{1}{2} \int \log \left(\frac{dP}{dP + dQ} \right) dP + \frac{1}{2} \int \log \left(\frac{dQ}{dP + dQ} \right) dQ + \log 2.$$

Its sample version (13) is the original GAN proposed by [25] that is widely used in learning distributions of images.

Example 2: Zero-One Score and TV-GAN

2. *Zero-One Score.* The zero-one score $S(t, 1) = 2\mathbb{I}\{t \geq 1/2\}$ and $S(t, 0) = 2\mathbb{I}\{t < 1/2\}$ is also known as the misclassification loss. This is a regular proper scoring rule but not strictly proper. The induced divergence function $D_{\mathcal{T}}(P, Q)$ is a variational lower bound of the total variation distance

$$\text{TV}(P, Q) = P \left(\frac{dP}{dQ} \geq 1 \right) - Q \left(\frac{dP}{dQ} \geq 1 \right) = \frac{1}{2} \int |dP - dQ|.$$

The sample version (13) is recognized as the TV-GAN that is extensively studied by [21] in the context of robust estimation.

Example 3: Quadratic Score and LS-GAN

3. *Quadratic Score.* Also known as the Brier score [6], the definition is given by $S(t, 1) = -(1 - t)^2$ and $S(t, 0) = -t^2$. The corresponding convex function in the Savage representation is given by $G(t) = -t(1 - t)$. By Proposition 2.1, the divergence function (3) induced by this regular strictly proper scoring rule is a variational lower bound of the following divergence function,

$$\Delta(P, Q) = \frac{1}{8} \int \frac{(dP - dQ)^2}{dP + dQ},$$

known as the triangular discrimination. The sample version (5) belongs to the family of least-squares GANs proposed by [39].

Example 4: Boosting Score

4. *Boosting Score.* The boosting score was introduced by [7] with $S(t, 1) = -\left(\frac{1-t}{t}\right)^{1/2}$ and $S(t, 0) = -\left(\frac{t}{1-t}\right)^{1/2}$ and has an connection to the AdaBoost algorithm. The corresponding convex function in the Savage representation is given by $G(t) = -2\sqrt{t(1-t)}$. The induced divergence function $D_{\mathcal{T}}(P, Q)$ is thus a variational lower bound of the squared Hellinger distance

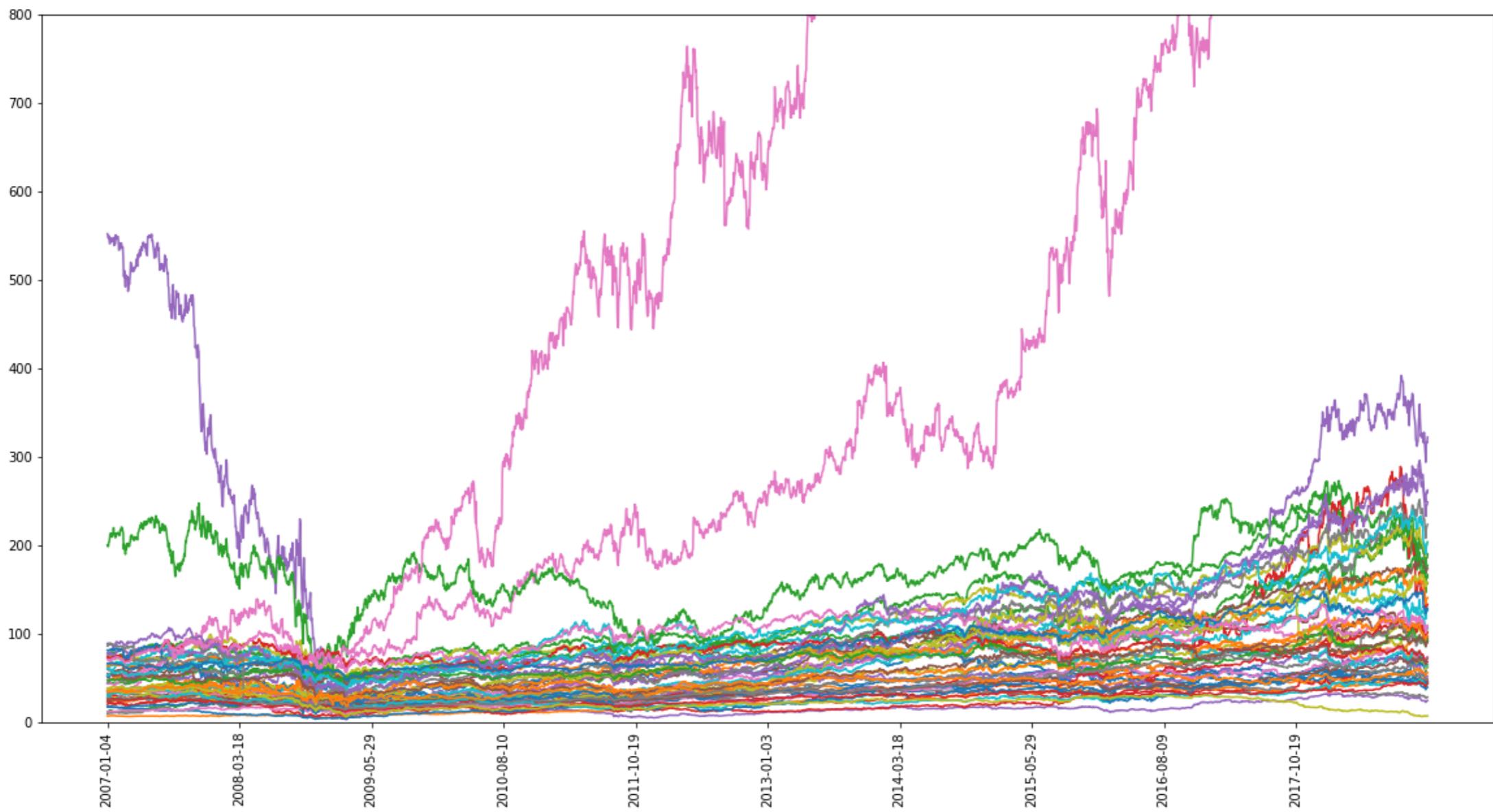
$$H^2(P, Q) = \frac{1}{2} \int \left(\sqrt{dP} - \sqrt{dQ} \right)^2.$$

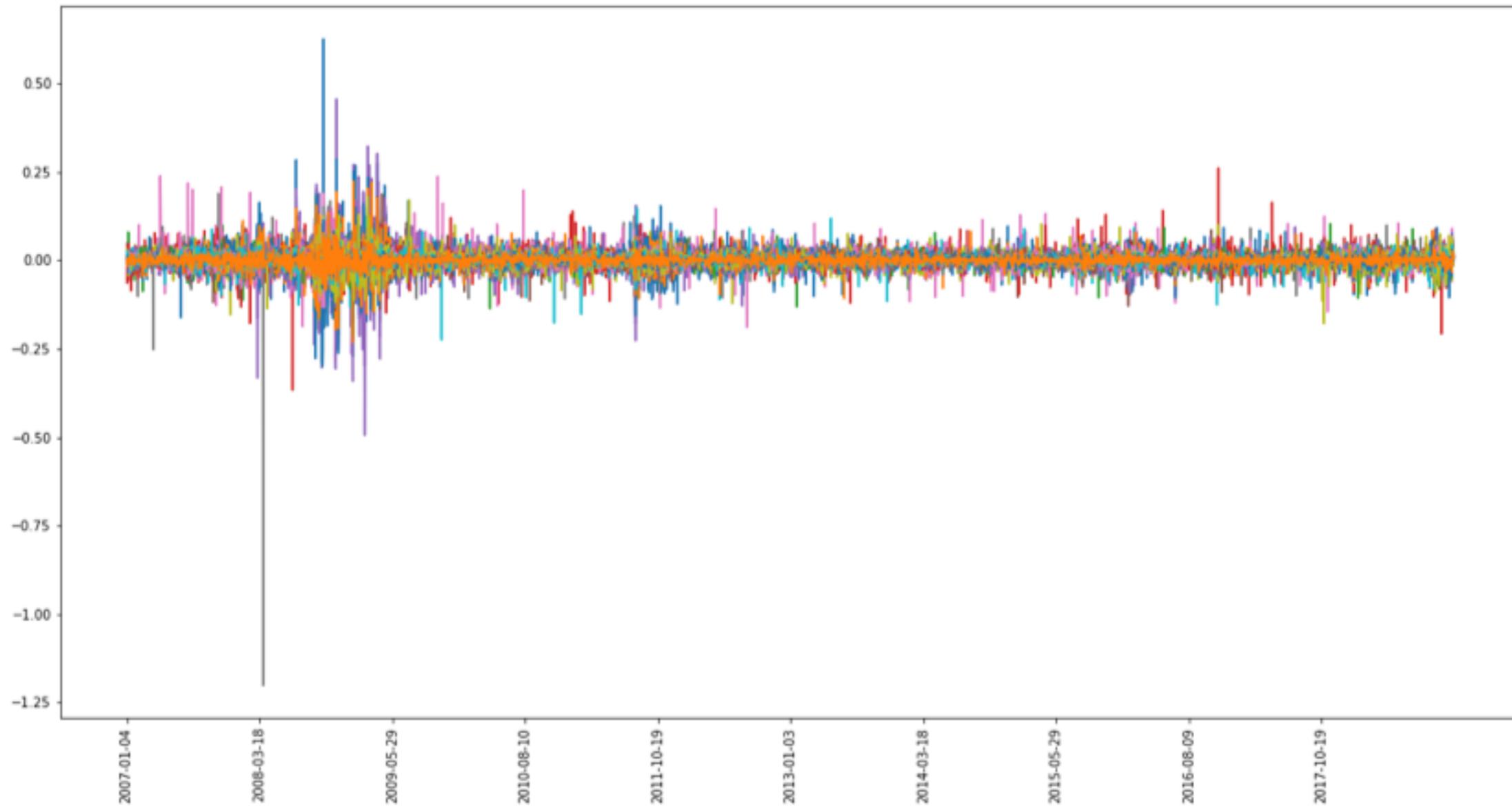
Example 5: Beta Score and new GANs

5. *Beta Score.* A general Beta family of proper scoring rules was introduced by [7] with $S(t, 1) = - \int_t^1 c^{\alpha-1} (1-c)^{\beta} dc$ and $S(t, 0) = - \int_0^t c^{\alpha} (1-c)^{\beta-1} dc$ for any $\alpha, \beta > -1$. The log score, the quadratic score and the boosting score are special cases of the Beta score with $\alpha = \beta = 0$, $\alpha = \beta = 1$, $\alpha = \beta = -1/2$. The zero-one score is a limiting case of the Beta score by letting $\alpha = \beta \rightarrow \infty$. Moreover, it also leads to asymmetric scoring rules with $\alpha \neq \beta$.

Application: Robust PCA

Price of 50 stocks from 2007/01 to 2018/12 Corps are selected by ranking in market capitalization





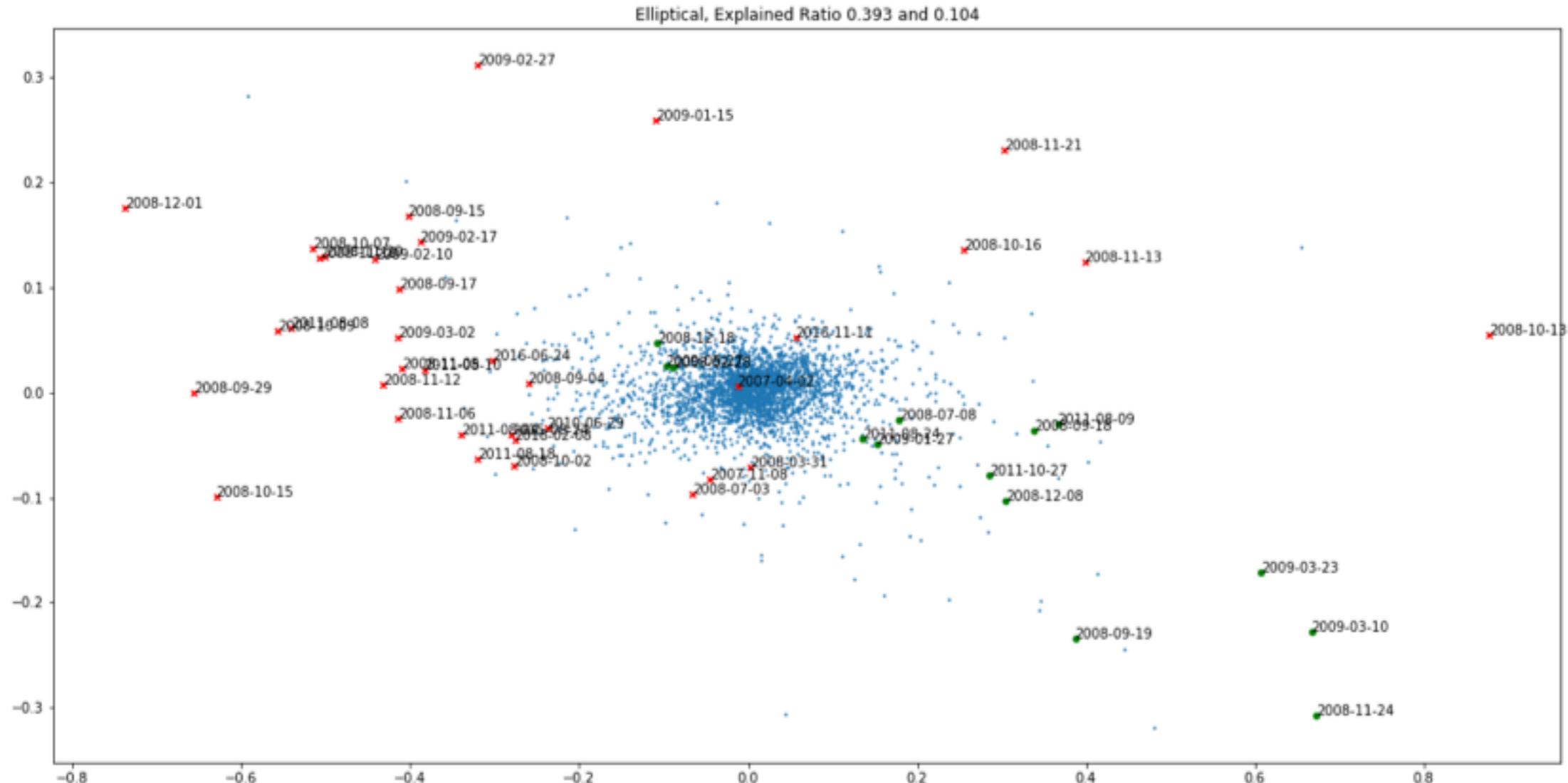
Log-return. $y[i] = \log(\text{price}_{\{i+1\}}/\text{price}_{\{i\}})$

Fit data by Elliptica-GAN.

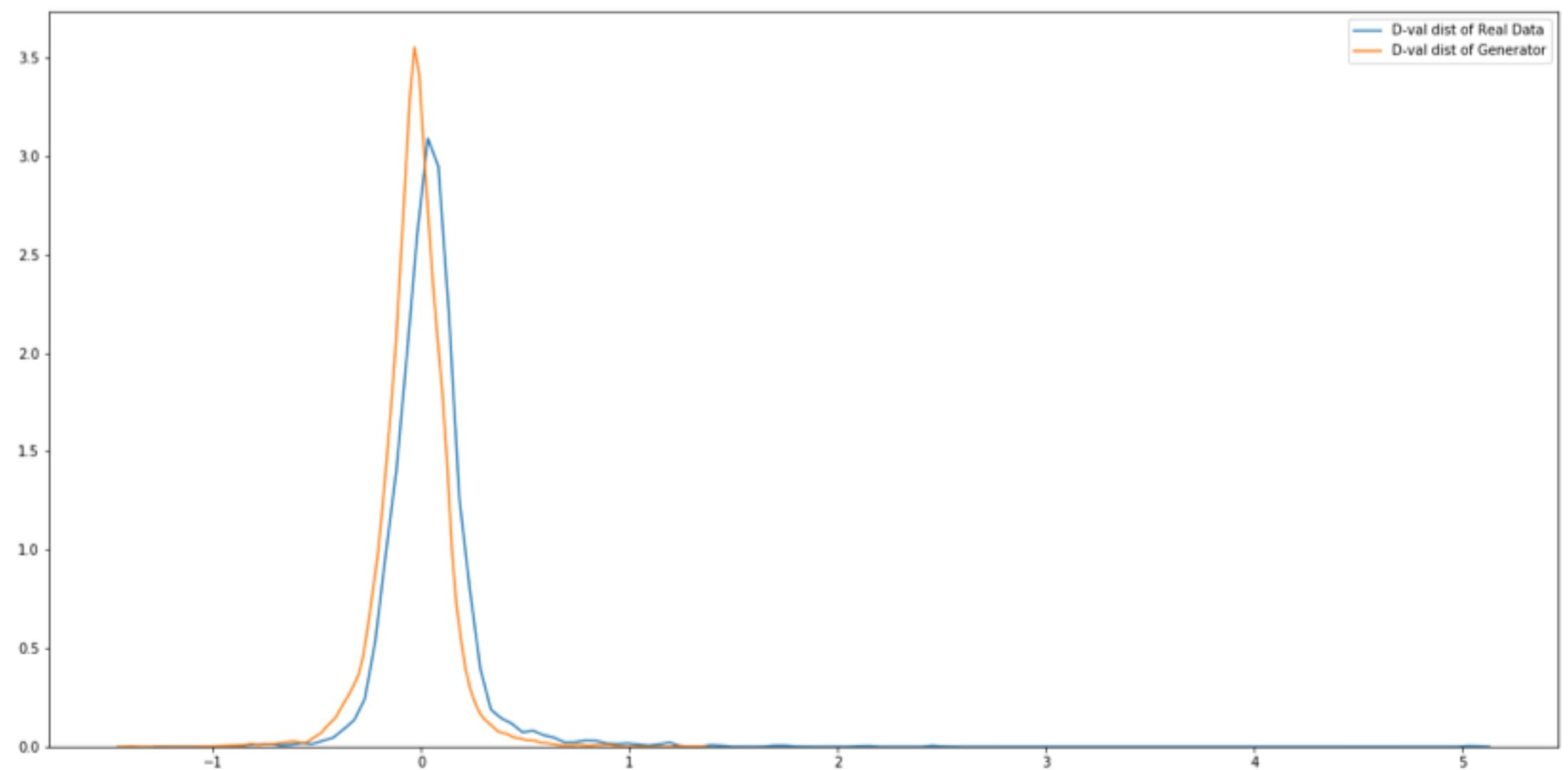
Apply SVD on scatter.

Dimension reduction on R².

outlier x and o are selected from Discriminator value distribution.

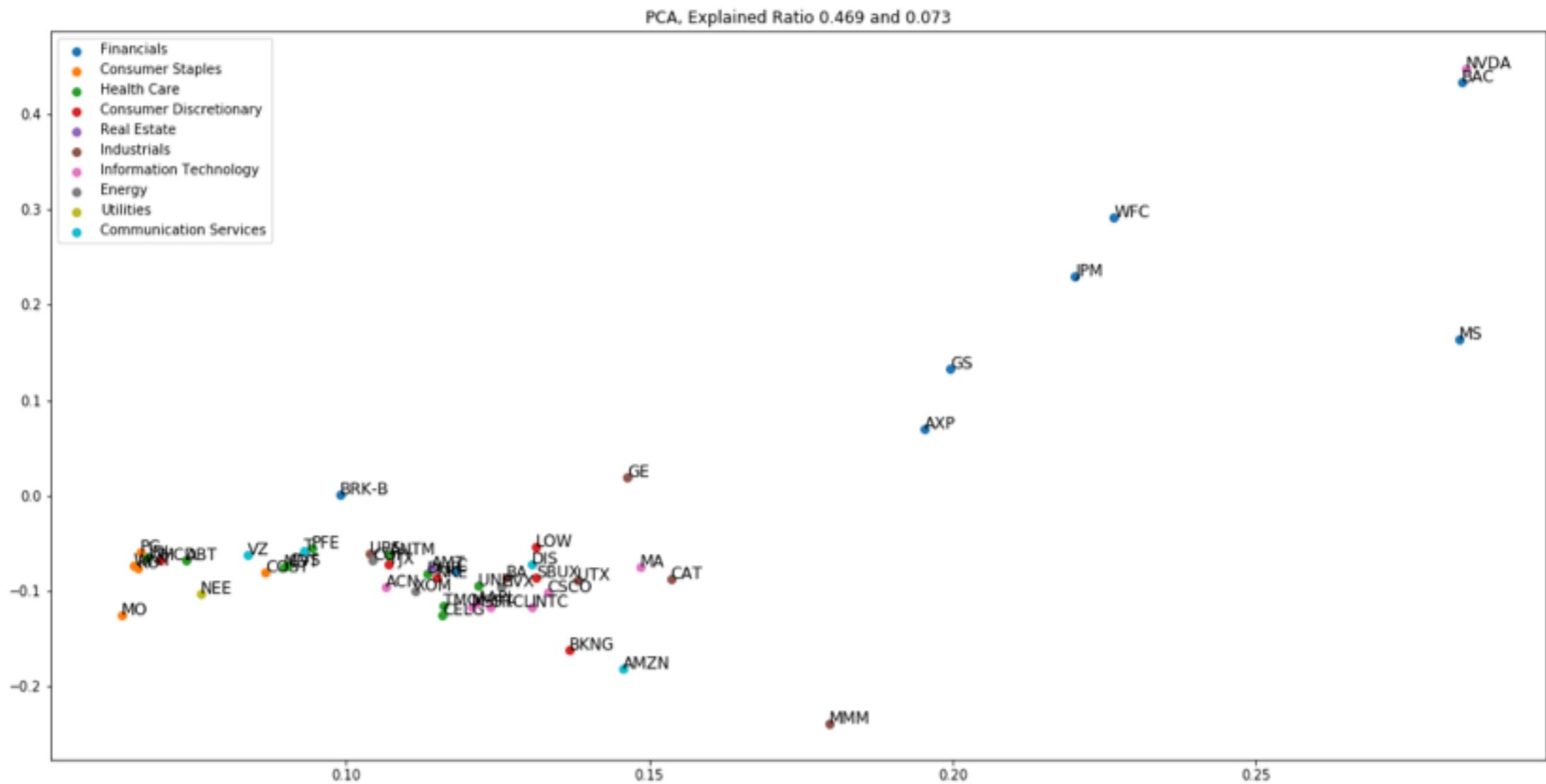


Discriminator value distribution from (Elliptical) Generator and real samples. Outliers are chosen from samples larger/ lower than a chosen percentile of Generator distribution

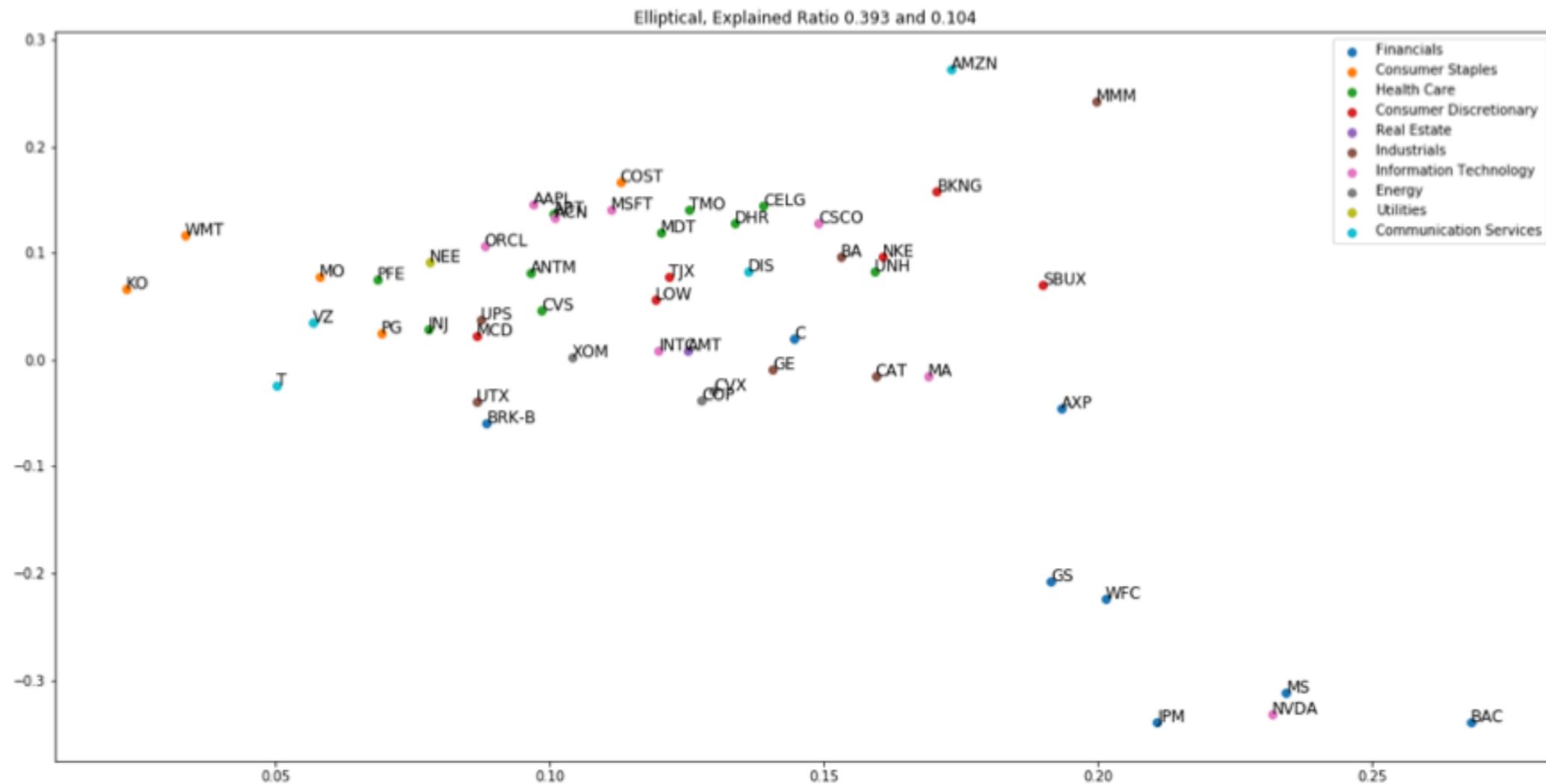


Loading of PCA.

First two direction are dominated by few corps → not robust



Loading of Elliptical Scatter: Comparing with PCA, it's more robust in the sense that it does not totally dominate by Financial company (JPM, GS)



Reference

- Gao, Liu, Yao, Zhu, Robust Estimation and Generative Adversarial Networks, *ICLR 2019*,
<https://arxiv.org/abs/1810.02030>
- Gao, Yao, Zhu, Generative Adversarial Networks for Robust Scatter Estimation: A Proper Scoring Rule Perspective, <https://arxiv.org/abs/1903.01944>

Thank You

