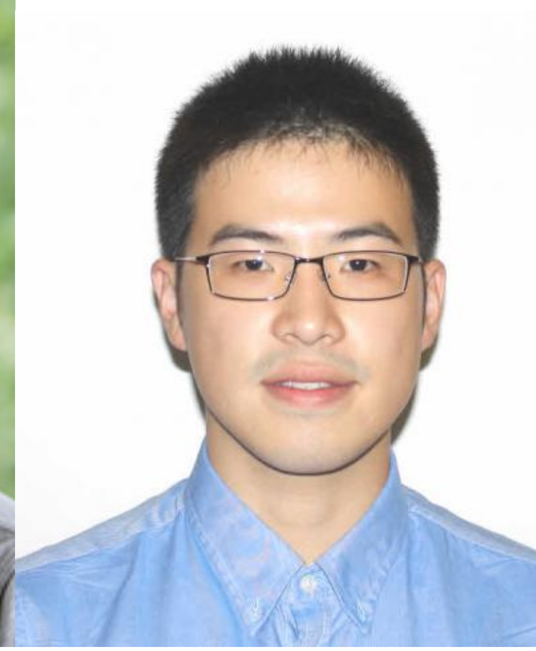
Robust Statistical Learning and Generative Adversarial Networks

Yuan YAO
HKUST

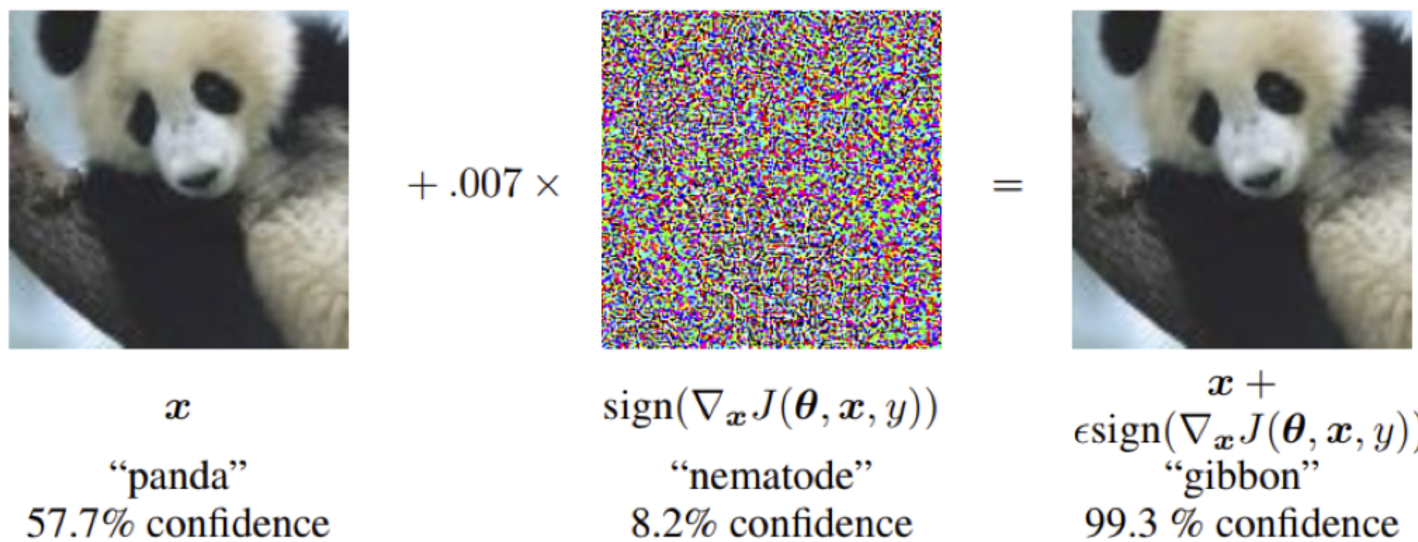


Chao GAO
U. Chicago

Jiyi LIU
Yale U.

Weizhi ZHU
HKUST

Deep Learning is Notoriously Not Robust!

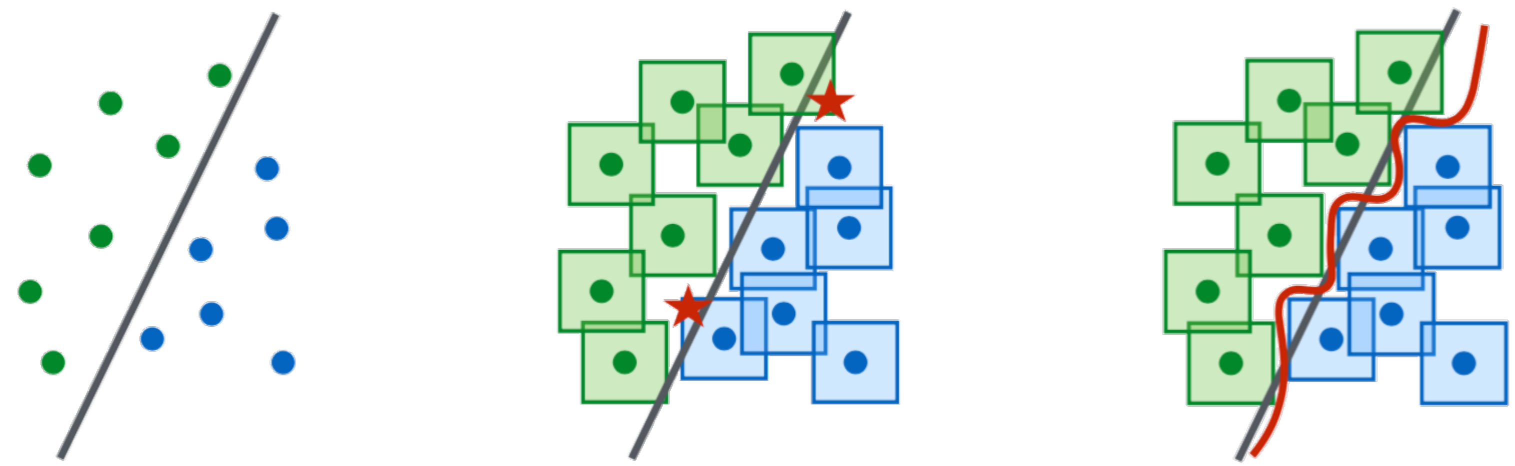


[Goodfellow et al., 2014]

- Imperceivable adversarial examples are ubiquitous to fail neural networks

- How can one achieve robustness?

Robust Optimization



- 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 (Madry et al. ICLR'2018):

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

  - robust to any distributions, yet computationally hard

Distributionally Robust Optimization (DRO)

- Distributional Robust Optimization:

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

  - $\mathcal{D}$ 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 Distributionally Robust Optimization

Wasserstein-DRO:

$$\min_{\theta} \max_{P_\epsilon : W_p(\mathbb{P}_\epsilon, \mathbb{P}_n) \leq \epsilon} \mathbb{E}_{z \sim \mathbb{P}_\epsilon \in \mathcal{D}}[\ell_\theta(z)]$$

where

$$\mathcal{W}_p(\mathbb{P}, \mathbb{Q}) = \begin{cases} \left(\min_{\gamma \in \Gamma(\mathbb{P},\mathbb{Q})} \left\{ \int_{\mathcal{Z} \times \mathcal{Z}} \mathrm{d}^p(z, z') \, \gamma(dz, dz') \right\}\right)^{1/p}, & \text{if } 1 \leq p < \infty, \\ \inf_{\gamma \in \Gamma(\mathbb{P},\mathbb{Q})} \{\gamma - \mathrm{esssup}_{\mathcal{Z} \times \mathcal{Z}} \, \mathrm{d}(z, z')\}, & \text{if } p = \infty, \end{cases}$$

- For a broad class of loss functions, Wasserstein-DRO is asymptotically equivalent to the following regularization problem

$$\min_{\theta} \mathbb{E}_{(x,y) \sim \mathbb{P}_n} [\ell_\theta(x, y)] + \alpha \cdot \left\| \nabla_{(x,y)} \ell_\theta \right\|_{\mathbb{P}_n, p_*}$$

where $p_* = \frac{p}{p-1}$ and the penalty term $\left\| \nabla_{(x,y)} \ell_\beta \right\|_{\mathbb{P}_n, p_*}$ represents the empirical dual p_*-norm of the gradient of the loss function with respect to the data (input Lipschitz).

- Gao, Kleywegt (2016); Gao, Chen, Kleywegt (2017); Blanchet, Kang, Murphy (2016), et al.

Certified Robustness of Lasso

Theorem (Blanchet, Kang, Murphy (2016))

Consider the cost for $z = (x, y)$:

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

For $p = \infty$ and linear regression, Wasserstein-DRO is equivalent to SQRT-Lasso:

$$\min_{\beta} \max_{P: W_c(P, P_n) \leq \delta} E_P\left(\left(Y - \beta^T X\right)^2\right)$$
$$= \min_{\beta} \left\{ E_{P_n}^{1/2}\left[\left(Y - \beta^T X\right)^2\right] + \sqrt{\delta}\|\beta\|_{p_*} \right\}^2$$

where $p_* = \frac{p}{p-1} = 1$.

- Generalized to asymmetric cost (Bregman divergence, Asymmetric Mahalannobis) by Blanchet, Glynn, Hui, Xie (2022)

TV-neighborhood

- Now 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

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

contamination proportion

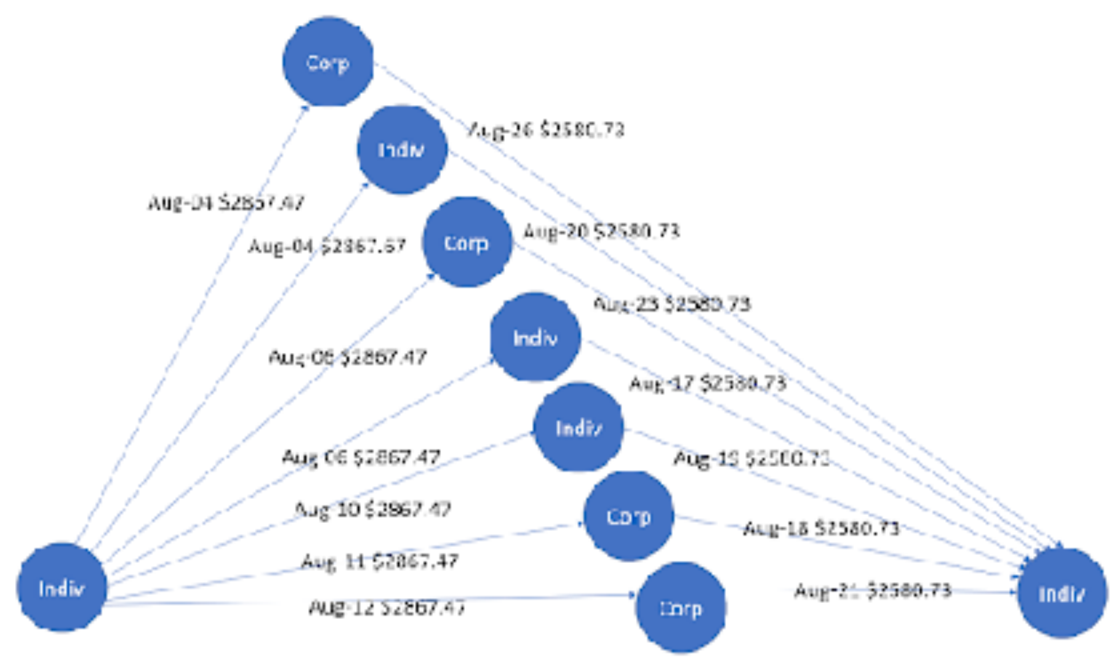
arbitrary contamination

parameter of interest

[Huber 1964]

Example: Financial Fraud

- *P* represent normal transactions
- *Q* represent fraudulent transactions, e.g. money laundering, which is sparse and arbitrarily close to *P*
- Finding *P* and its dual problem in finding *Q*?



An Example

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

how to estimate ?

Medians

1. **Coordinatewise median**

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

2. **Tukey's median (1975)**

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

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)	$\frac{p}{n} + p\epsilon^2$	$\frac{p}{n} + \epsilon^2$: **optimal** [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]

Depth and Statistical Properties

Multivariate Location Depth

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

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

[Tukey, 1975]

Regression Depth

model $$y|X \sim N(X^T\beta, \sigma^2)$$

embedding $$Xy|X \sim N(XX^T\beta, \sigma^2 XX^T)$$

projection $$u^T Xy|X \sim N(u^T XX^T\beta, \sigma^2 u^T XX^T u)$$

$$\left\{ \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}\{u^T X_i(y_i - X_i^T\eta) > 0\} \wedge \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}\{u^T X_i(y_i - X_i^T\eta) \leq 0\} \right\}$$

[Rousseeuw & Hubert, 1999]

Tukey's depth is not a special case of regression depth.

Multi-task Regression Depth

$$(X, Y) \in \mathbb{R}^p \times \mathbb{R}^m \sim \mathbb{P}$$

$$B \in \mathbb{R}^{p \times m}$$

population version:

$$\mathcal{D}_{\mathcal{U}}(B, \mathbb{P}) = \inf_{U \in \mathcal{U}} \mathbb{P}\left\{\left\langle U^T X, Y - B^T X \right\rangle \geq 0\right\}$$

empirical version:

$$\mathcal{D}_{\mathcal{U}}(B, \{(X_i, Y_i)\}_{i=1}^n) = \inf_{U \in \mathcal{U}} \frac{1}{n} \sum_{i=1}^n \mathbb{I}\left\{\left\langle U^T X_i, Y_i - B^T X_i \right\rangle \geq 0\right\}$$

[Mizera, 2002]

Multi-task Regression Depth

$$\mathcal{D}_{\mathcal{U}}(B, \mathbb{P}) = \inf_{U \in \mathcal{U}} \mathbb{P}\left\{\left\langle U^T X, Y - B^T X \right\rangle \geq 0\right\}$$

$p = 1, X = 1 \in \mathbb{R},$

$$\mathcal{D}_{\mathcal{U}}(b, \mathbb{P}) = \inf_{u \in \mathcal{U}} \mathbb{P}\left\{u^T (Y - b) \geq 0\right\}$$

$m = 1,$

$$\mathcal{D}_{\mathcal{U}}(\beta, \mathbb{P}) = \inf_{U \in \mathcal{U}} \mathbb{P}\left\{u^T X (y - \beta^T X) \geq 0\right\}$$

Statistical Errors of Multi-task Regression Depth

Estimation Error. For any $\delta > 0$,

$$\sup_{B \in \mathbb{R}^{p \times m}} |\mathcal{D}(B, \mathbb{P}_n) - \mathcal{D}(B, \mathbb{P})| \leq C\sqrt{\frac{pm}{n}} + \sqrt{\frac{\log(1/\delta)}{2n}},$$

with probability at least $1 - 2\delta$.

Contamination Error.

$$\sup_{B,Q} |\mathcal{D}(B, (1 - \epsilon P_{B^*}) + \epsilon Q) - \mathcal{D}(B, P_{B^*})| \leq \epsilon$$

Statistical Optimality of Multi-task Regression Depth

$$(X, Y) \sim P_B$$

$$(X_1, Y_1), ..., (X_n, Y_n) \sim (1-\epsilon)P_B + \epsilon Q$$

Theorem [G17]. For some $C > 0$,

$$\mathsf{Tr}((\widehat{B} - B)^T \Sigma (\widehat{B} - B)) \leq C\sigma^2 \left(\frac{pm}{n} \vee \epsilon^2\right),$$

$$\|\widehat{B} - B\|_{\mathrm{F}}^2 \leq C\frac{\sigma^2}{\kappa^2} \left(\frac{pm}{n} \vee \epsilon^2\right),$$

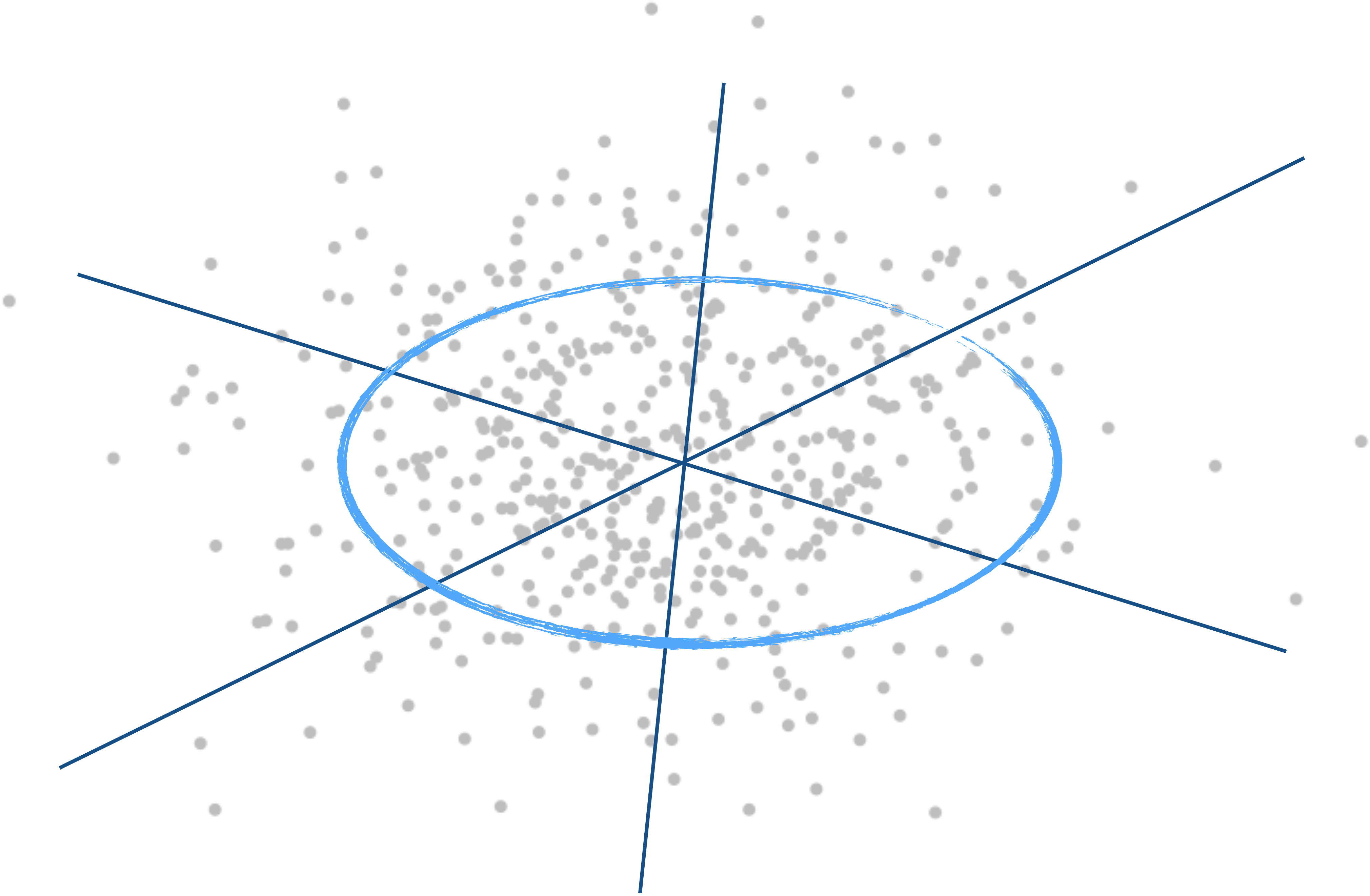
with high probability uniformly over B, Q.

Covariance Matrix

$$X_1, ..., X_n \sim (1-\epsilon)N(0,\Sigma) + \epsilon Q.$$

how to estimate ?

Covariance Matrix



Covariance Matrix

$$\mathcal{D}(\Gamma, \{X_i\}_{i=1}^n) = \min_{\|u\|=1} \min \left\{ \frac{1}{n}\sum_{i=1}^n \mathbb{I}\{|u^T X_i|^2 \geq u^T \Gamma u\}, \frac{1}{n}\sum_{i=1}^n \mathbb{I}\{|u^T X_i|^2 < u^T \Gamma u\} \right\}$$

$$\hat{\Gamma} = \arg\max_{\Gamma \succeq 0} \mathcal{D}(\Gamma, \{X_i\}_{i=1}^n) \qquad \hat{\Sigma} = \hat{\Gamma}/\beta$$

Theorem [CGR15]. For some $C > 0$,

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

with high probability uniformly over Σ, Q.

Summary

mean	$\|\cdot\|^2$	$\frac{p}{n} \vee \epsilon^2$
reduced rank regression	$\|\cdot\|_F^2$	$\frac{\sigma^2}{\kappa^2}\frac{r(p+m)}{n} \vee \frac{\sigma^2}{\kappa^2}\epsilon^2$
Gaussian graphical model	$\|\cdot\|_{\ell_1}^2$	$\frac{s^2\log(ep/s)}{n} \vee s\epsilon^2$
covariance matrix	$\|\cdot\|_{\mathrm{op}}^2$	$\frac{p}{n} \vee \epsilon^2$
sparse PCA	$\|\cdot\|_F^2$	$\frac{s\log(ep/s)}{n\lambda^2} \vee \frac{\epsilon^2}{\lambda^2}$

Computation

Computational Challenges

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

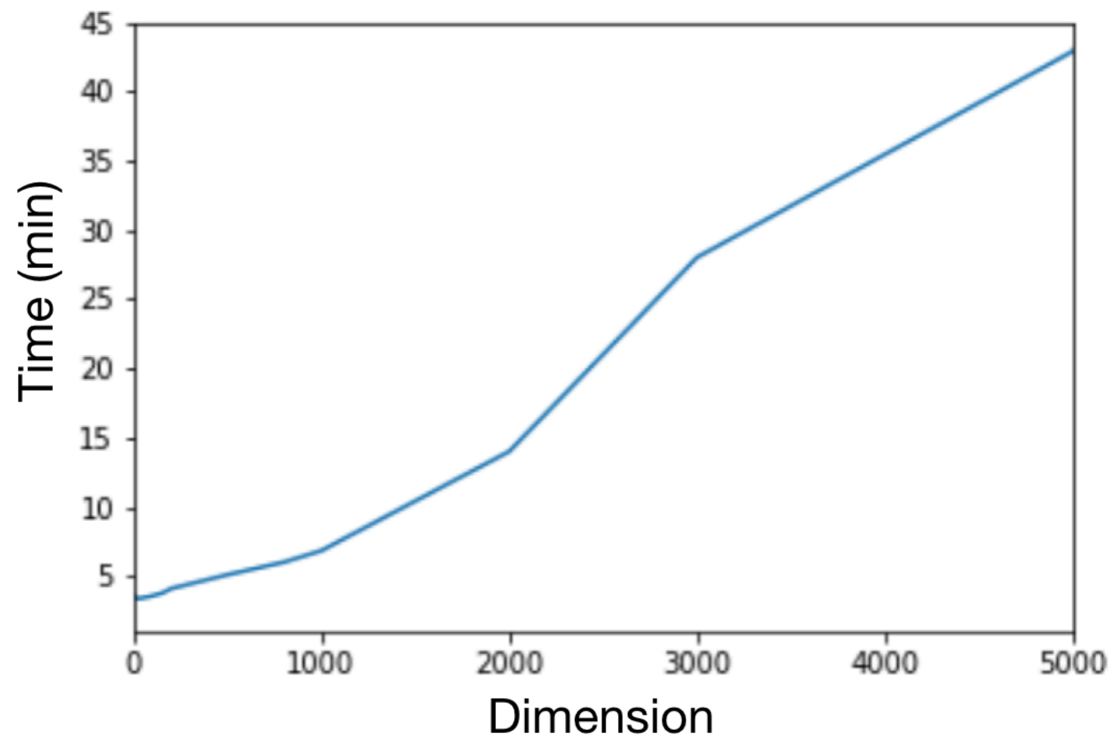
Lai, Rao, Vempala
Diakonikolas, Kamath, Kane, Li, Moitra, Stewart
Balakrishnan, Du, Singh

- 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

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
$\text{Cauchy}(1.5 * 1_p, I_p)$	**0.0664 (0.0065)**	0.0743 (0.0103)	0.3529 (0.0543)	0.1244 (0.0114)
$\text{Cauchy}(5.0 * 1_p, I_p)$	**0.0480 (0.0058)**	0.0540 (0.0064)	0.4855 (0.0616)	0.1687 (0.0310)
$\text{Cauchy}(1.5 * 1_p, 5 * I_p)$	0.0754 (0.0135)	**0.0742 (0.0111)**	0.3726 (0.0530)	0.1220 (0.0112)
$\text{Normal}(1.5 * 1_p, 5 * I_p)$	**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

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

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

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

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)

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

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

From f-GAN to Tukey's Median: f-learning (GLYZ'18)

Consider the special case

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

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

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

- If $f(x) = x \log x$, $\mathcal{Q} = \widetilde{\mathcal{Q}}$, (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, $f^*(t) = t\mathbb{I}\{0 \le t \le 1\}$, f-GAN $\Rightarrow$ TV-GAN
  - $\mathcal{Q} = \{N(\eta, I_p) : \eta \in \mathbb{R}^p\}$ and $\widetilde{\mathcal{Q}} = \{N(\widetilde{\eta}, I_p) : \|\widetilde{\eta} - \eta\| \le r\}$, (12) $\overset{r \to 0}{\Rightarrow}$ Tukey's Median

TV-GAN

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

$N(\eta, I_p)$

logistic regression classifier

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

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

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

TV-GAN

rugged landscape!

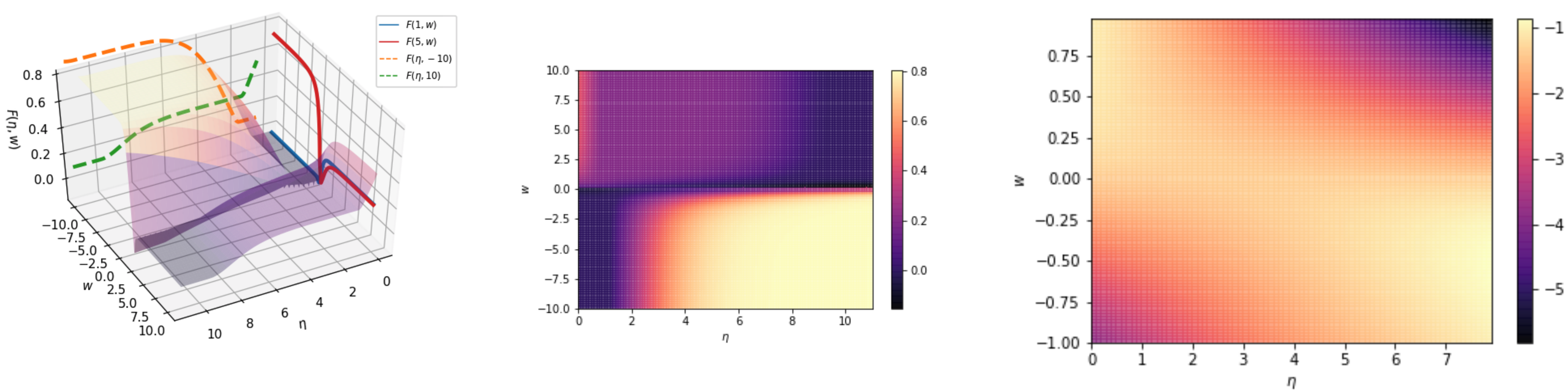


Figure 1: Landscape of TV-GAN objective function $F(\eta, w) = \sup_b[E_P \text{sigmoid}(wX + b) - E_{N(\eta,1)} \text{sigmoid}(wX + b)]$, where b is maximized out for visualization. Samples are drawn from $P = (1-\epsilon)N(1,1) + \epsilon N(10,1)$ with $\epsilon = 0.2$. Left: a surface plot of $F(\eta, w)$. The solid curves are marginal functions for fixed η's: $F(1, w)$ (red) and $F(5, w)$ (blue), and the dash curves are marginal functions for fixed w's: $F(\eta, -10)$ (orange) and $F(\eta, 10)$ (green). Right: a heatmap of $F(\eta, w)$. 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.

JS-GAN

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

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

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:

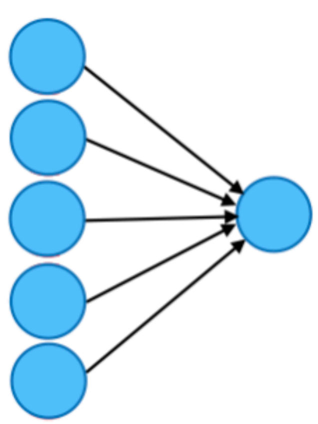
$$\mathcal{D} = \left\{ D(x) = \text{sigmoid}(w^T g(X)) \right\}$$

Deep JS-GAN

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

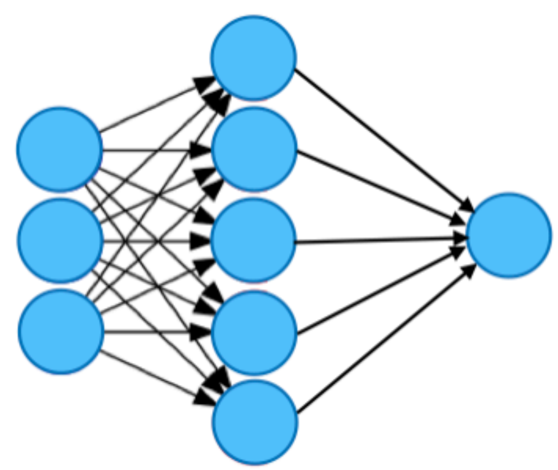
numerical experiment

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



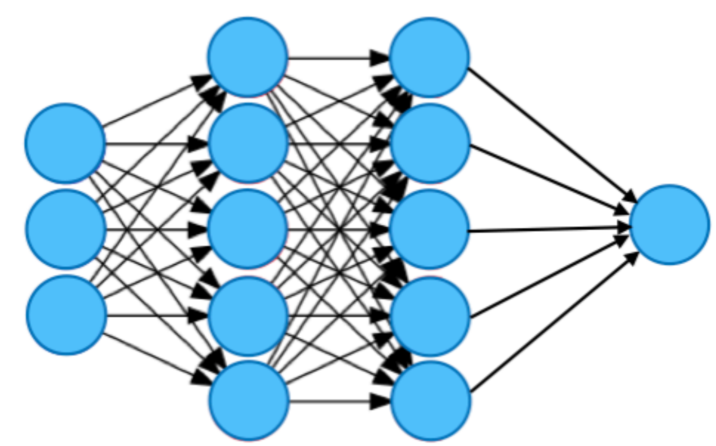
$\widehat{\theta} \approx (1-\epsilon)\theta + \epsilon\widetilde{\theta}$

✗



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

✓



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

✓

JS-GAN

A classifier with hidden layers leads to robustness. Why?

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

Proposition.

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

JS-GAN

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

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 $\theta \in \mathbb{R}^p, Q$.

JS-GAN: Adaptation to Unknown Covariance

unknown covariance?

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

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

no need to change the discriminator class

Generalization

Strong Contamination model:

$$X_1, ..., X_n \overset{iid}{\sim} P \quad \text{for some } P \text{ satisfying } \mathsf{TV}(P, E(\theta, \Sigma, H)) \leq \epsilon$$

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

- We are going to replace the log likelihoods in JS-GAN by some scoring functions

$$\log t \mapsto S(t, 1) : [0, 1] \to \mathbb{R}$$

$$\log(1 - t) \mapsto S(t, 0) : [0, 1] \to \mathbb{R}$$

that map the probability (likelihood) to some real numbers.

Fisher Consistency: Proper Scoring Rule

- With a Bernoulli experiment of probability p observing 1, define the expected score

$$S(t, p) = pS(t, 1) + (1 - p)S(t, 0)$$

- Like likelihood functions, as a function of t, we hope that $S(t, p)$ is maximized at $t = p$

$$\max_t S(t, p) = S(p, p) =: G(p)$$

- Such a score is called **Proper Scoring Rule.**

Savage Representation of Proper Scoring Rule

Lemma (Savage representation)

- For a proper scoring rule $S(t, p)$:
  - $G(t) = S(t, t)$ is convex
  - $S(t, 0) = G(t) - tG'(t)$
  - $S(t, 1) = G(t) + (1 - t)G'(t)$
  - $S(t, p) = pS(t, 1) + (1 - p)S(t, 0) = G(t) + G'(t)(p - t)$

Divergence

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

Proposition 1 *Given any regular proper scoring rule* $\{S(\cdot,1), S(\cdot,0)\}$ *and any class* $\mathcal{T} \ni \{\frac{1}{2}\}$, $D_{\mathcal{T}}(P,Q)$ *is a divergence function, and*

$$D_{\mathcal{T}}(P,Q) \leq D_f\Big(P \Big\| \frac{1}{2}P + \frac{1}{2}Q\Big), \tag{4}$$

where $f(t) = G(t/2) - G(1/2)$. *Moreover, whenever* $\mathcal{T} \ni \frac{dP}{dP+dQ}$, *the inequality above becomes an equality.*

- 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 $S(1,0) = -\infty$.

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

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

   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

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

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 $\alpha = \beta \to \infty$. Moreover, it also leads to asymmetric scoring rules with $\alpha \neq \beta$.

Smooth Proper Scores

Assumption (Smooth Proper Scoring Rules)

We assume that

- $G^{(2)}(1/2) > 0$ and $G^{(3)}(t)$ is continuous at $t = 1/2$;
- Moreover, there is a universal constant $c_0 > 0$, such that $2G^{(2)}(1/2) \geq G^{(3)}(1/2) + c_0$.
  - The condition $2G^{(2)}(1/2) \geq G^{(3)}(1/2) + c_0$ is automatically satisfied by a symmetric scoring rule, because $S(t,1) = S(1-t,0)$ immediately implies that $G^{(3)}(1/2) = 0$.
  - For the Beta score 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 $\alpha, \beta > -1$, it is easy to check that such a c_0 (only depending on α, β) exists as long as $|\alpha - \beta| < 1$.

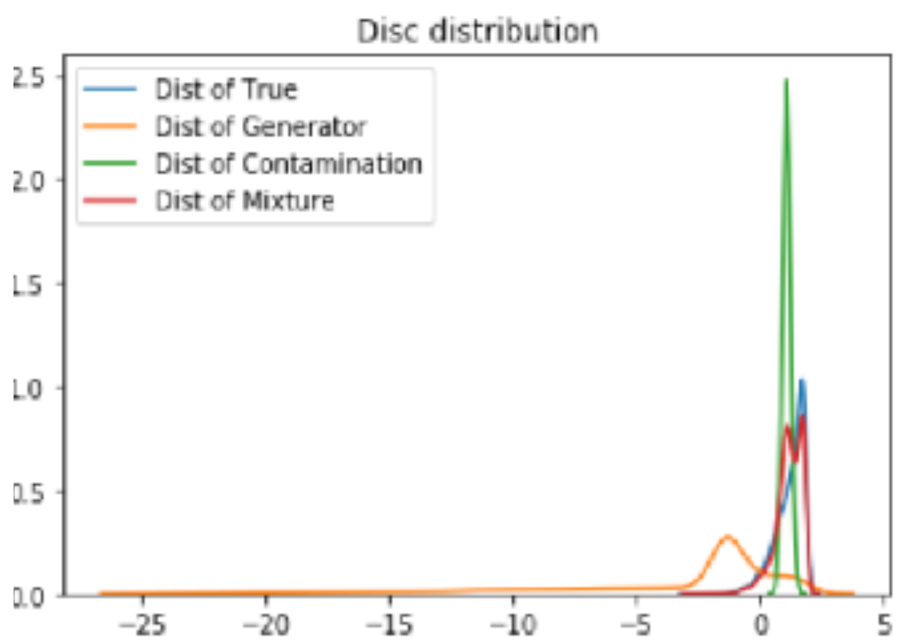
Statistical Optimality

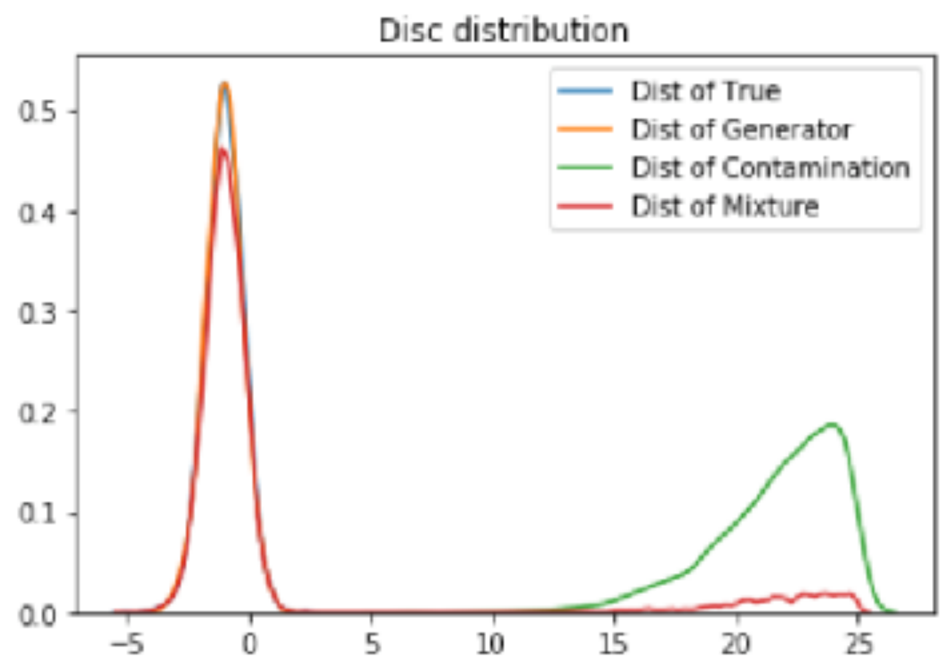
Theorem [GYZ19]. For a neural network class $\mathcal{T}$ with at least one hidden layer and appropriate regularization, we have

$$\begin{aligned} \|\widehat{\theta}-\theta\|^2 &\leq C\left(\frac{p}{n}\vee\epsilon^2\right), \\ \|\widehat{\Sigma}-\Sigma\|_{\mathrm{op}}^2 &\leq C\left(\frac{p}{n}\vee\epsilon^2\right), \end{aligned}$$

Experiments

Discriminator identifies outliers



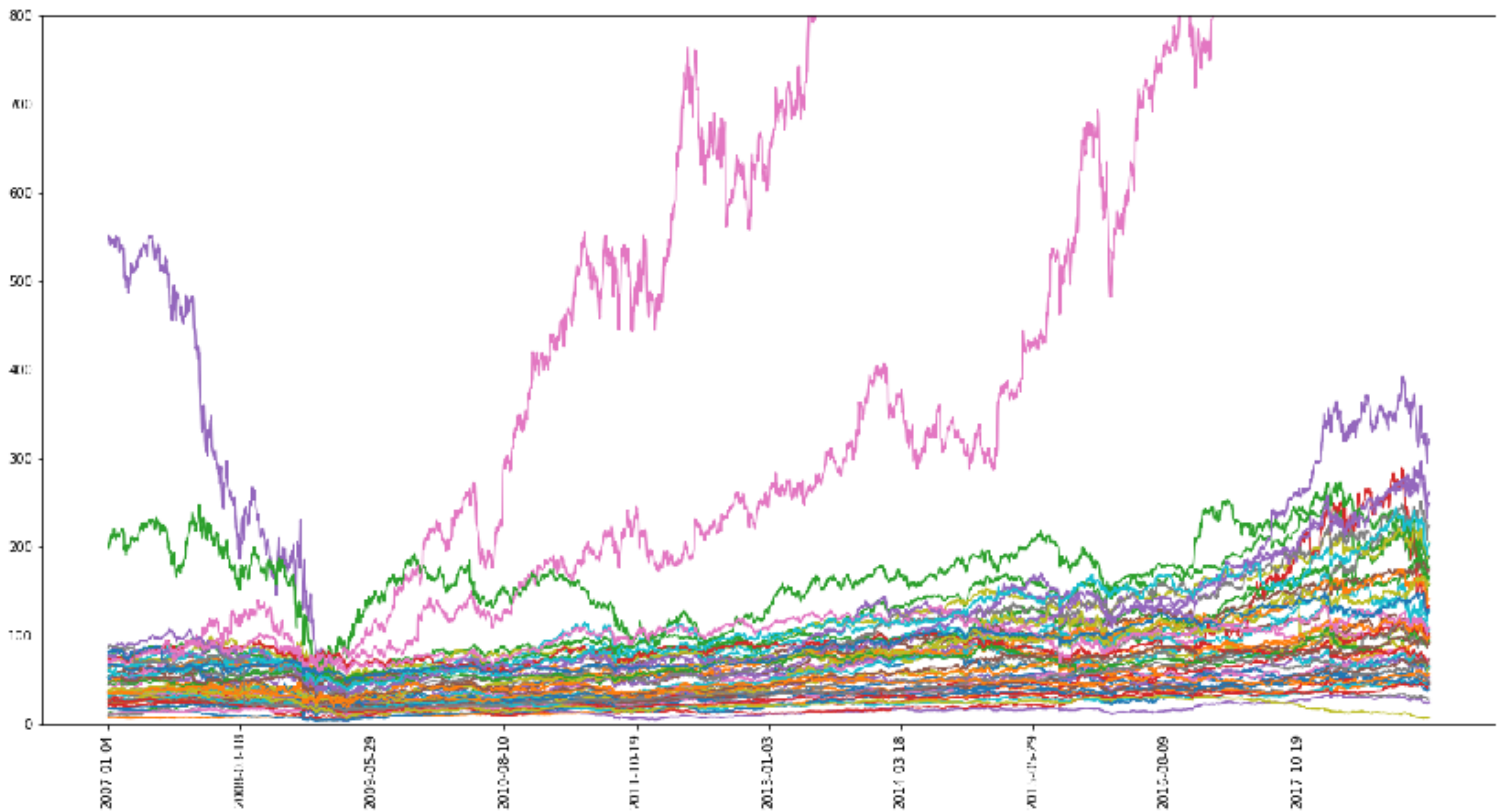


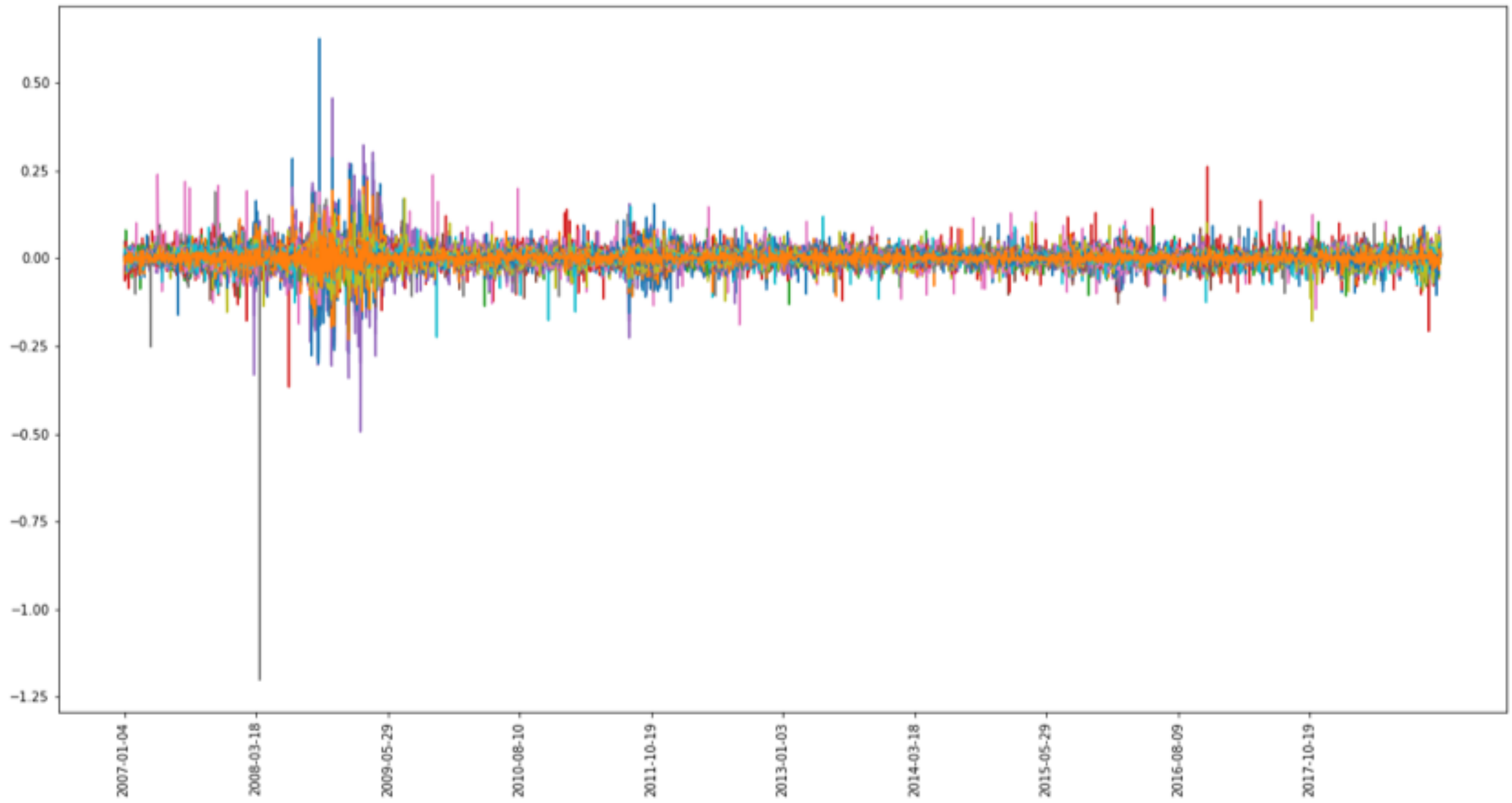
$$(1-\epsilon)N(0_p, I_p) + \epsilon Q$$

$$N(5 * 1_p, I_p)$$

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

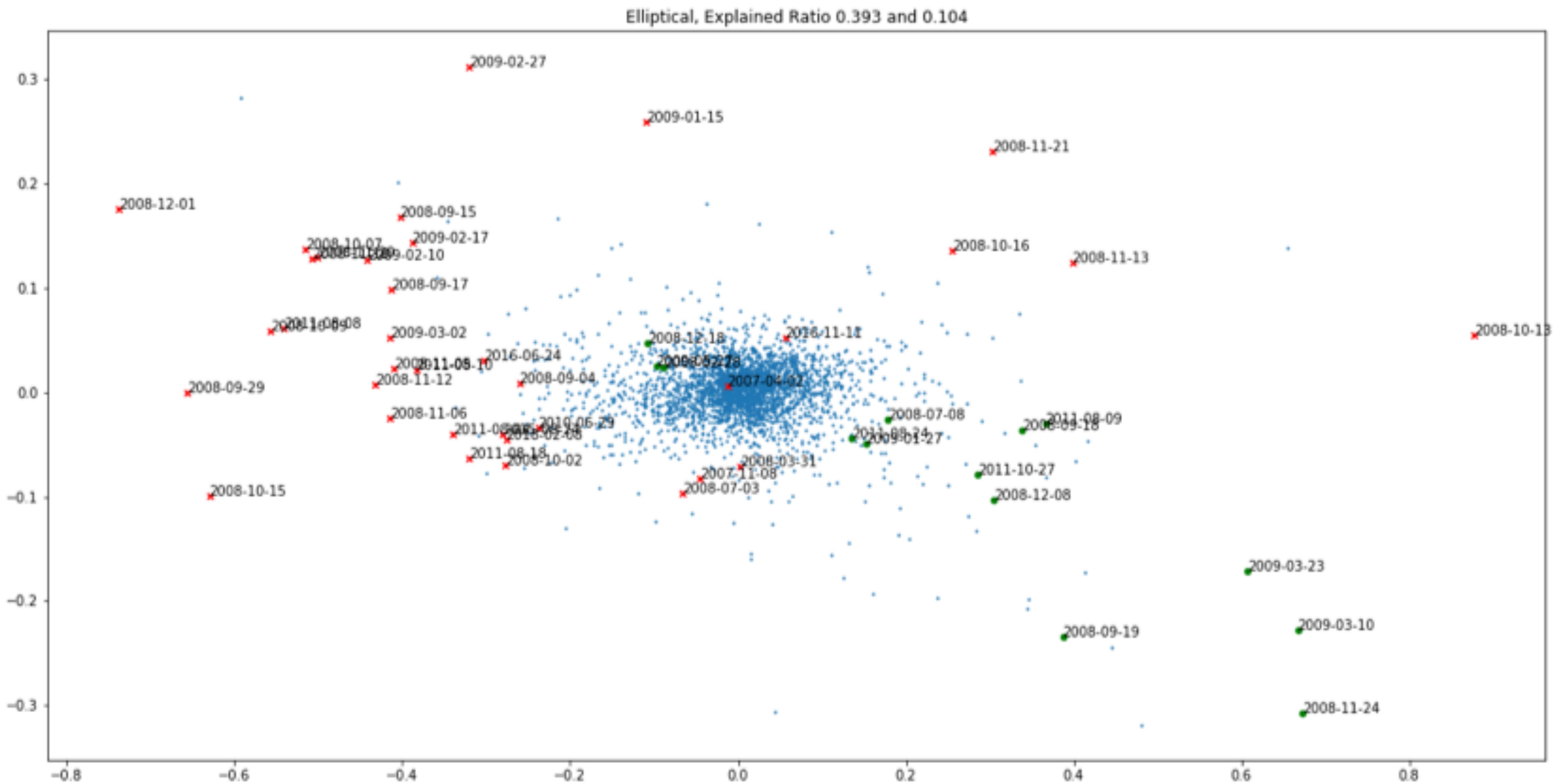
Application: Price of 50 stocks from 2007/01 to 2018/12
Corps are selected by ranking in market capitalization



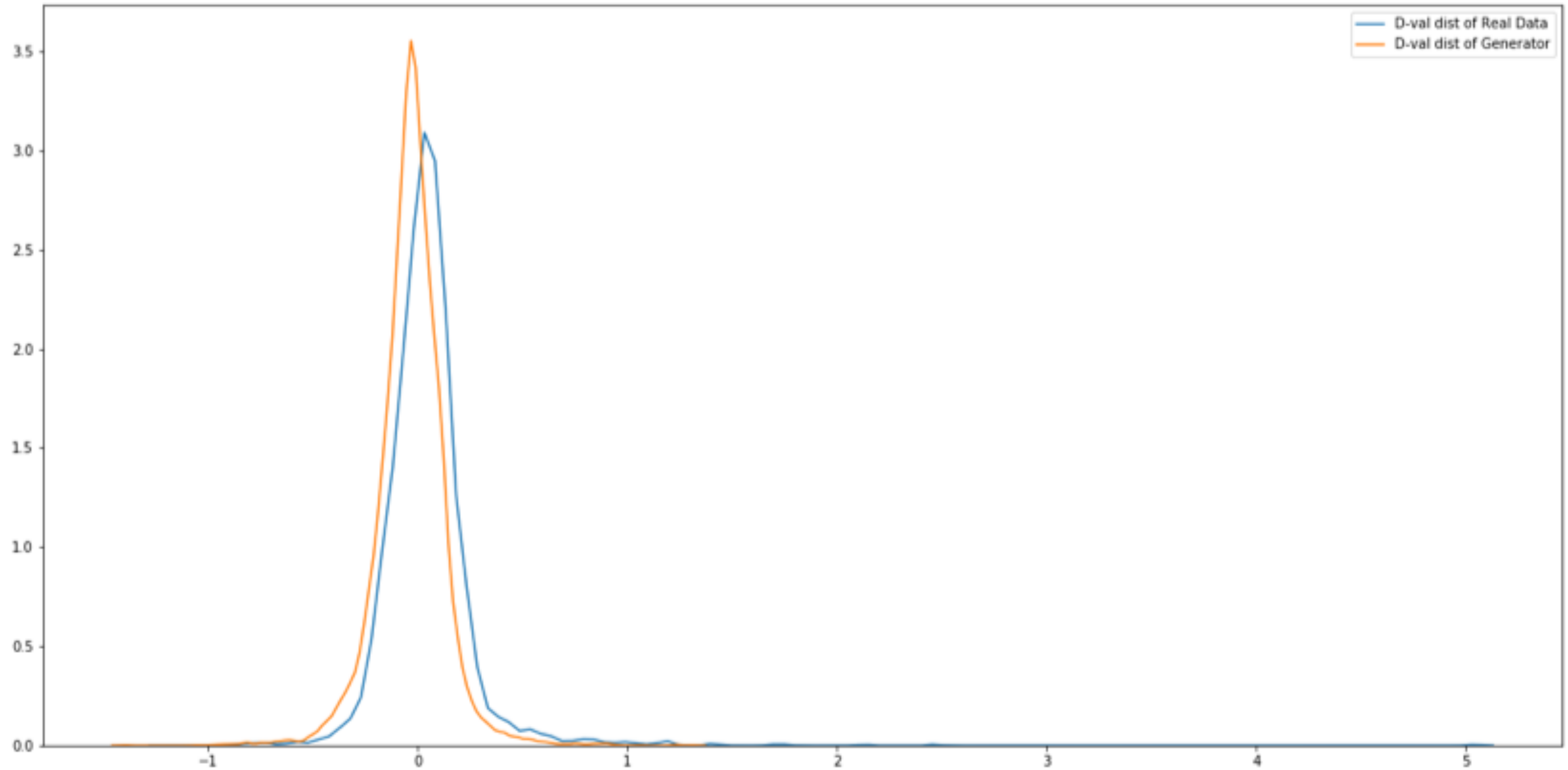


Log-return. y[i] = log(price_{i+1}/price_{i})

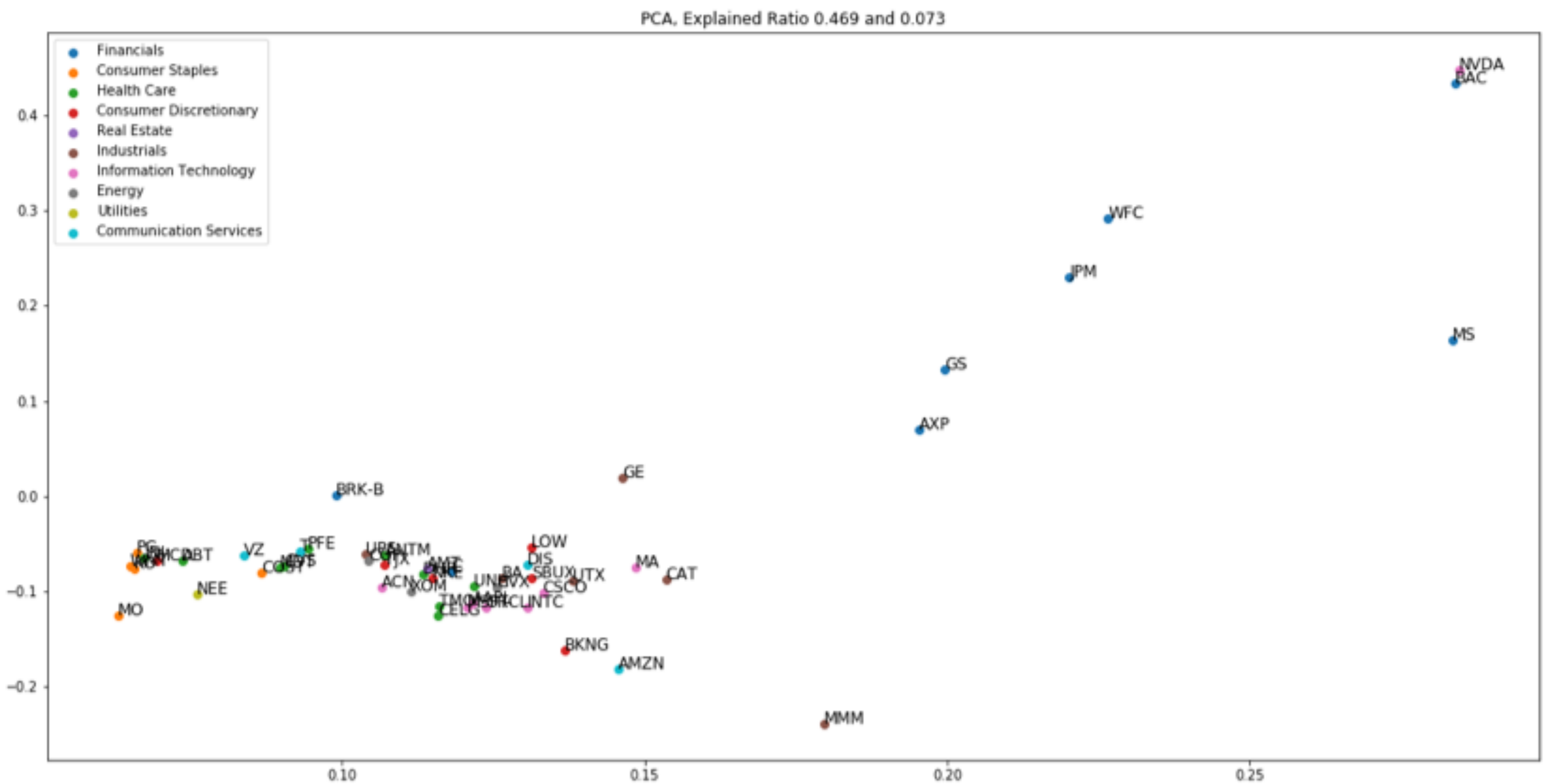
Fit data by Elliptical-GAN.
Apply SVD on scatter.
Dimension reduction on R^2.
outlier x and o are selected from Discriminator value distribution.



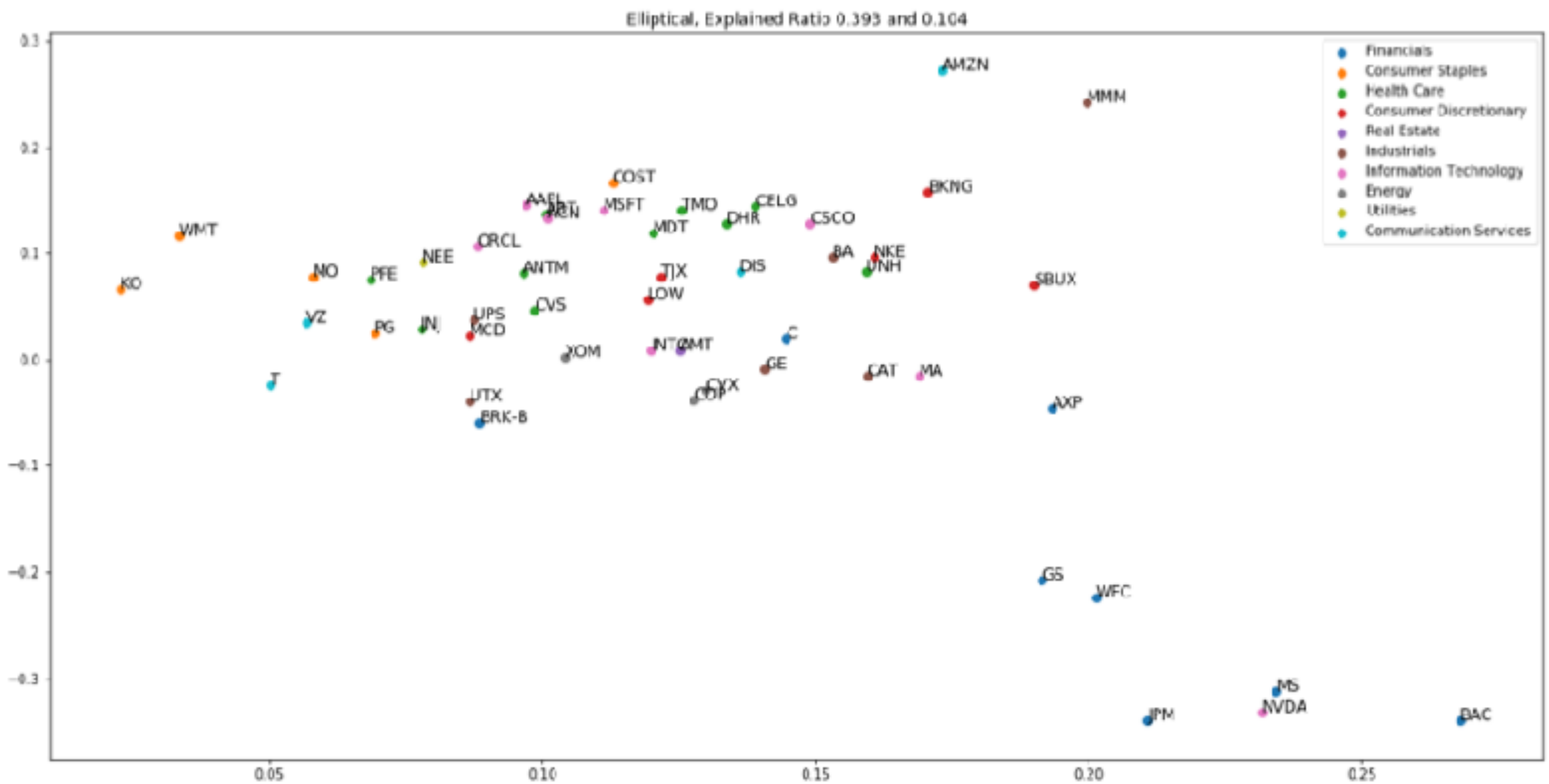
Discriminator value distribution from (Elliptical) Generator and real samples. Outliers are chosen from samples larger/ lower than a chosen percentile of Generator distribution



Standard (non-robust) PCA:
First two direction are dominated by few corps —> not robust



Robust PCA: Loadings 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, *Journal of Machine Learning Research*, 21(160):1-48, 2020. https://arxiv.org/abs/1903.01944

Thank You