Computation

Computational Challenges

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

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

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

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?

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

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

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

- 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

f-Learning

f-Learning

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

f-Learning

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

$$f(u) = \sup_t (tu - f^*(t))$$

f-Learning

f-divergence

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

variational representation

$$= \sup_T \left[\mathbb{E}_{X\sim P} T(X) - \mathbb{E}_{X\sim Q} f^*(T(X))\right]$$

f-Learning

f-divergence

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

variational representation

$$= \sup_T \left[\mathbb{E}_{X\sim P} T(X) - \mathbb{E}_{X\sim Q} f^*(T(X))\right]$$

optimal T

$$T(x) = f'\left(\frac{p(x)}{q(x)}\right)$$

f-Learning

f-divergence

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

variational representation

$$= \sup_T \left[\mathbb{E}_{X\sim P} T(X) - \mathbb{E}_{X\sim Q} f^*(T(X))\right]$$

$$= \sup_{\tilde{Q}} \left\{\mathbb{E}_{X\sim P} f'\left(\frac{d\tilde{Q}(X)}{dQ(X)}\right) - \mathbb{E}_{X\sim Q} f^*\left(f'\left(\frac{d\tilde{Q}(X)}{dQ(X)}\right)\right)\right\}$$

f-Learning

$$\max_{T \in \mathcal{T}} \left\{ \frac{1}{n} \sum_{i=1}^{n} T(X_i) - \int f^* \left(T \right) dQ \right\}$$

$$\max_{\tilde{Q} \in \tilde{\mathcal{Q}}} \left\{ \frac{1}{n} \sum_{i=1}^{n} f' \left(\frac{\tilde{q}(X_i)}{q(X_i)} \right) - \int f^* \left(f' \left(\frac{\tilde{q}}{q} \right) \right) dQ \right\}$$

f-Learning

$$\min_{Q \in \mathcal{Q}} \max_{T \in \mathcal{T}} \left\{ \frac{1}{n} \sum_{i=1}^{n} T(X_i) - \int f^* \left(T \right) dQ \right\}$$

$$\min_{Q \in \mathcal{Q}} \max_{\tilde{Q} \in \tilde{\mathcal{Q}}} \left\{ \frac{1}{n} \sum_{i=1}^{n} f' \left(\frac{\tilde{q}(X_i)}{q(X_i)} \right) - \int f^* \left(f' \left(\frac{\tilde{q}}{q} \right) \right) dQ \right\}$$

f-Learning

f-GAN

$$\min_{Q \in \mathcal{Q}} \max_{T \in \mathcal{T}} \left\{ \frac{1}{n} \sum_{i=1}^{n} T(X_i) - \int f^* \left(T \right) dQ \right\}$$

f-Learning

$$\min_{Q \in \mathcal{Q}} \max_{\tilde{Q} \in \tilde{\mathcal{Q}}} \left\{ \frac{1}{n} \sum_{i=1}^{n} f' \left(\frac{\tilde{q}(X_i)}{q(X_i)} \right) - \int f^* \left(f' \left(\frac{\tilde{q}}{q} \right) \right) dQ \right\}$$

f-Learning

f-GAN

$$\min_{Q\in\mathcal{Q}} \max_{T\in\mathcal{T}} \left\{ \frac{1}{n}\sum_{i=1}^{n} T(X_i) - \int f^*\left(T\right) dQ \right\}$$

f-Learning

$$\min_{Q\in\mathcal{Q}} \max_{\tilde{Q}\in\tilde{\mathcal{Q}}} \left\{ \frac{1}{n}\sum_{i=1}^{n} f'\left(\frac{\tilde{q}(X_i)}{q(X_i)}\right) - \int f^*\left(f'\left(\frac{\tilde{q}}{q}\right)\right) dQ \right\}$$

[Nowozin, Cseke, Tomioka]

f-Learning

f-Learning

Jensen-Shannon	$f(x) = x \log x - (x+1)\log(x+1)$	**GAN**

[Goodfellow et al.

f-Learning

Jensen-Shannon	$f(x) = x \log x - (x+1)\log(x+1)$	**GAN**
Kullback-Leibler	$f(x) = x \log x$	**MLE**

[Goodfellow et al.

f-Learning

Jensen-Shannon	$f(x) = x \log x - (x+1)\log(x+1)$	**GAN**
Kullback-Leibler	$f(x) = x \log x$	**MLE**
Hellinger Squared	$f(x) = 2 - 2\sqrt{x}$	**rho**

[Goodfellow et al., Baraud and Birge]

f-Learning

Jensen-Shannon	$f(x) = x \log x - (x+1)\log(x+1)$	**GAN**
Kullback-Leibler	$f(x) = x \log x$	**MLE**
Hellinger Squared	$f(x) = 2 - 2\sqrt{x}$	**rho**
Total Variation	$f(x) = (x-1)_+$	**depth**

[Goodfellow et al., Baraud and Birge]

TV-Learning

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

TV-Learning

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

$$\mathcal{Q} = \left\{ N(\theta, I_p) : \theta \in \mathbb{R}^p \right\} \qquad \tilde{\mathcal{Q}} = \left\{ N(\tilde{\theta}, I_p) : \tilde{\theta} \in \mathcal{N}_r(\theta) \right\}$$

TV-Learning

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

$$\mathcal{Q} = \left\{ N(\theta, I_p) : \theta \in \mathbb{R}^p \right\} \qquad \tilde{\mathcal{Q}} = \left\{ N(\tilde{\theta}, I_p) : \tilde{\theta} \in \mathcal{N}_r(\theta) \right\}$$

$r \to 0$

TV-Learning

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

$$\mathcal{Q} = \left\{ N(\theta, I_p) : \theta \in \mathbb{R}^p \right\} \qquad \tilde{\mathcal{Q}} = \left\{ N(\tilde{\theta}, I_p) : \tilde{\theta} \in \mathcal{N}_r(\theta) \right\}$$

$r \to 0$

Tukey depth $\max_{\theta\in\mathbb{R}} \min_{\|u\|=1} \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}\left\{ u^T X_i \geq u^T \theta \right\}$

TV-Learning

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

TV-Learning

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

$$\mathcal{Q} = \left\{ N(0, \Sigma) : \Sigma \in \mathbb{R}^{p \times p} \right\} \quad \tilde{\mathcal{Q}} = \left\{ N(0, \tilde{\Sigma}) : \tilde{\Sigma} = \Sigma + r u u^T, \|u\| = 1 \right\}$$

TV-Learning

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

$$\mathcal{Q} = \left\{ N(0, \Sigma) : \Sigma \in \mathbb{R}^{p \times p} \right\} \qquad \tilde{\mathcal{Q}} = \left\{ N(0, \tilde{\Sigma}) : \tilde{\Sigma} = \Sigma + r u u^T, \|u\| = 1 \right\}$$

$r \to 0$

TV-Learning

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

$$\mathcal{Q} = \left\{ N(0,\Sigma) : \Sigma \in \mathbb{R}^{p\times p} \right\} \qquad \tilde{\mathcal{Q}} = \left\{ N(0,\tilde{\Sigma}) : \tilde{\Sigma} = \Sigma + ruu^T, \|u\| = 1 \right\}$$

$r \to 0$

(related to) matrix depth

$$\max_{\Sigma} \min_{\|u\|=1} \left[\left(\frac{1}{n}\sum_{i=1}^{n} \mathbb{I}\{|u^T X_i|^2 \leq u^T\Sigma u\} - \mathbb{P}(\chi_1^2 \leq 1) \right) \wedge \left(\frac{1}{n}\sum_{i=1}^{n} \mathbb{I}\{|u^T X_i|^2 > u^T\Sigma u\} - \mathbb{P}(\chi_1^2 > 1) \right) \right]$$


robust
statistics
community
deep
learning
community


robust
statistics
community
f-Learning
f-GAN
deep
learning
community

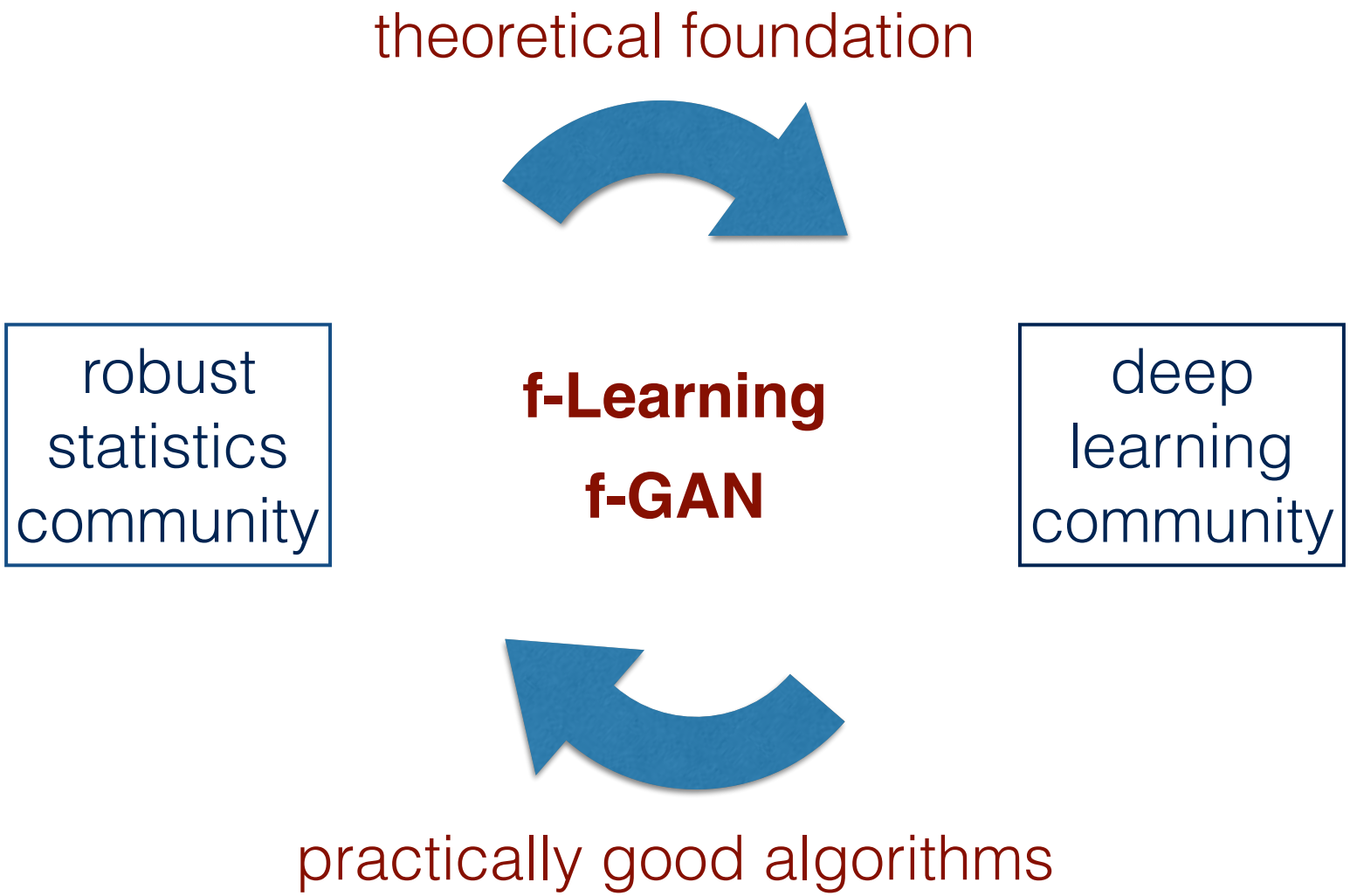
robust statistics community

f-Learning

f-GAN

deep learning community

practically good algorithms


theoretical foundation
robust statistics community
f-Learning
f-GAN
deep learning community
practically good algorithms

TV-GAN

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

TV-GAN

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

$$N(\eta, I_p)$$

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

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

very hard to optimize!

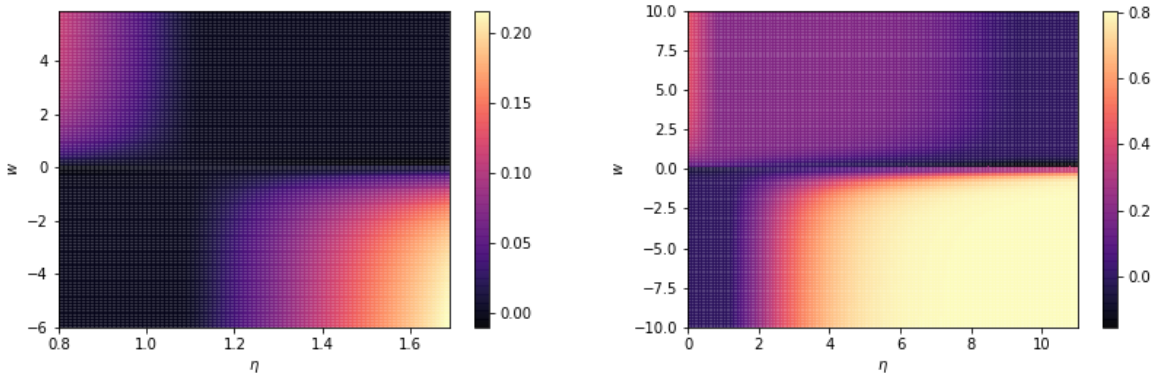


Figure: Heatmaps of the landscape of $F(\eta, w) = \sup_b [E_P \text{sigmoid}(wX + b) - E_{N(\eta,1)} \text{sigmoid}(wX + b)]$, where b is maximized out for visualization. Left: samples are drawn from $P = (1 - \epsilon)N(1, 1) + \epsilon N(1.5, 1)$ with $\epsilon = 0.2$. Right: samples are drawn from $P = (1 - \epsilon)N(1, 1) + \epsilon N(10, 1)$ 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 $f(x) = x \log x - (x+1) \log \frac{x+1}{2}$,

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

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:

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

JS-GAN

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

JS-GAN

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

numerical experiment

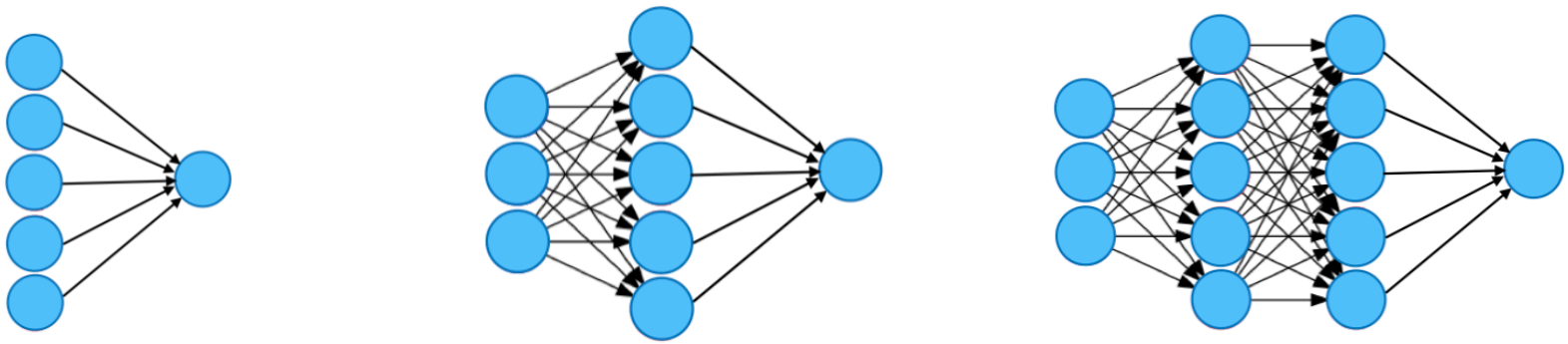
$$X_1, ..., X_n \sim (1 - \epsilon) N(\theta, I_p) + \epsilon N(\widetilde{\theta}, I_p)$$

JS-GAN

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

numerical experiment

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

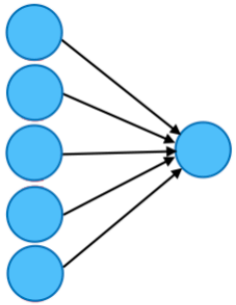


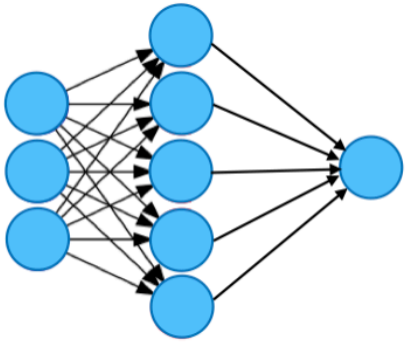
JS-GAN

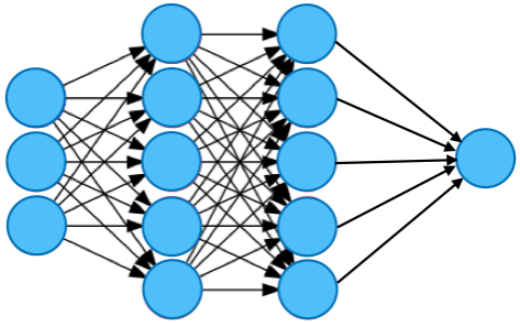
$$\widehat{\theta} = \underset{\eta \in \mathbb{R}^p}{\operatorname{argmin}} \max_{T \in \mathcal{T}} \left[\frac{1}{n} \sum_{i=1}^{n} \log T(X_i) + E_\eta \log(1 - T(X)) \right] + \log 4$$

numerical experiment

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







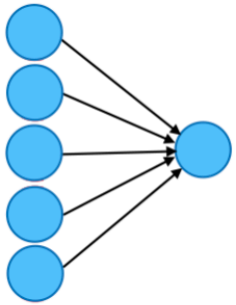
$$\widehat{\theta} \approx (1-\epsilon)\theta + \epsilon\widetilde{\theta}$$

JS-GAN

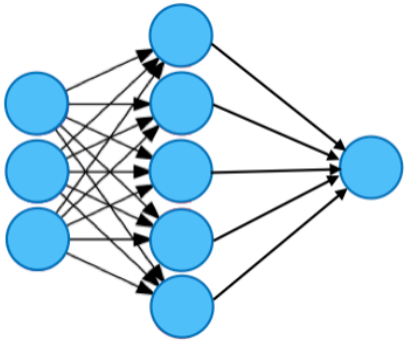
$$\widehat{\theta} = \underset{\eta \in \mathbb{R}^p}{\operatorname{argmin}} \max_{T \in \mathcal{T}} \left[\frac{1}{n} \sum_{i=1}^{n} \log T(X_i) + E_\eta \log(1 - T(X)) \right] + \log 4$$

numerical experiment

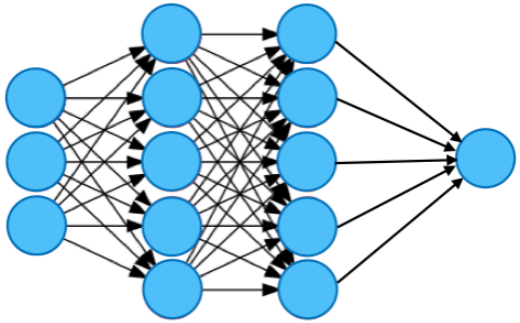
$$X_1, ..., X_n \sim (1-\epsilon) N(\theta, I_p) + \epsilon N(\widetilde{\theta}, I_p)$$



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



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



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

JS-GAN

A classifier with hidden layers leads to robustness. Why?

JS-GAN

A classifier with hidden layers leads to robustness. Why?

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

JS-GAN

A classifier with hidden layers leads to robustness. Why?

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

Proposition.

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

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

$$\|\widehat{\theta} - \theta\|^2 \lesssim \begin{cases} \dfrac{p}{n} + \epsilon^2 & \text{(indicator/sigmoid/ramp)} \\ \dfrac{p \log p}{n} + \epsilon^2 & \text{(ReLU after top two layers)} \end{cases}$$

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

JS-GAN

unknown covariance?

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

JS-GAN

unknown covariance?

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

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

JS-GAN

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

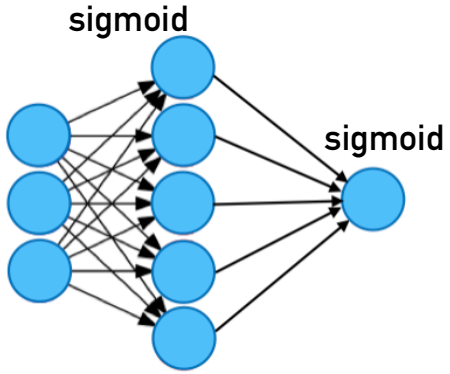
Covariance Matrix

JS-GAN

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

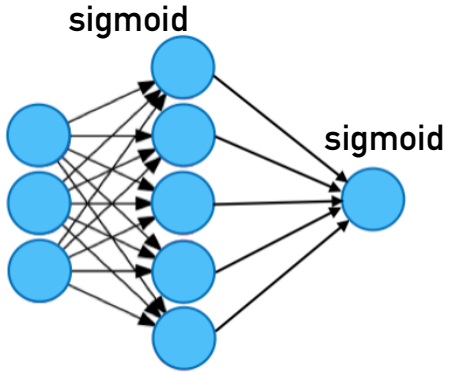
JS-GAN

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



JS-GAN

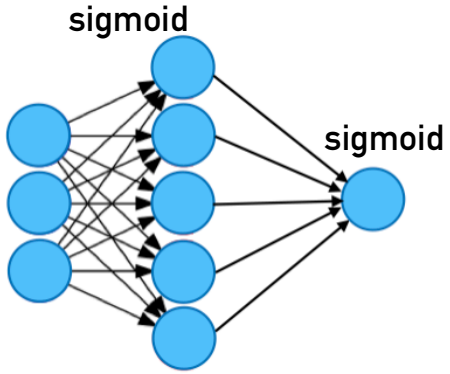
$$\widehat{\Sigma} = \underset{\Gamma}{\operatorname{argmin}} \max_{T \in \mathcal{T}} \left[\frac{1}{n} \sum_{i=1}^{n} \log T(X_i) + \mathbb{E}_{X \sim N(0,\Gamma)} \log(1 - T(X)) \right]$$



optimal for mean estimation

JS-GAN

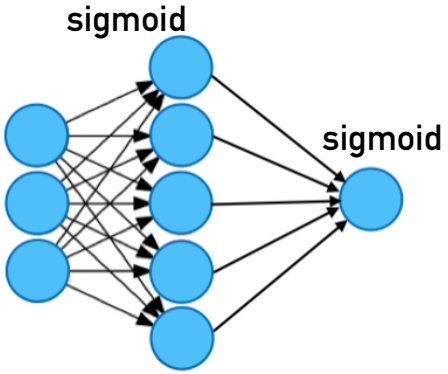
$$\widehat{\Sigma} = \underset{\Gamma}{\operatorname{argmin}} \max_{T \in \mathcal{T}} \left[\frac{1}{n} \sum_{i=1}^{n} \log T(X_i) + \mathbb{E}_{X \sim N(0,\Gamma)} \log(1 - T(X)) \right]$$

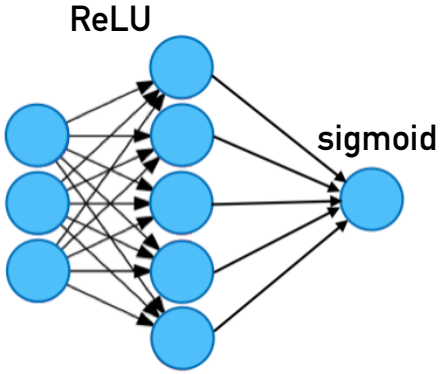


optimal for mean estimation
but **inconsistent** for
covariance estimation

JS-GAN

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

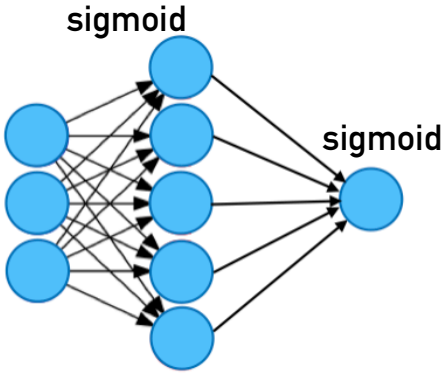




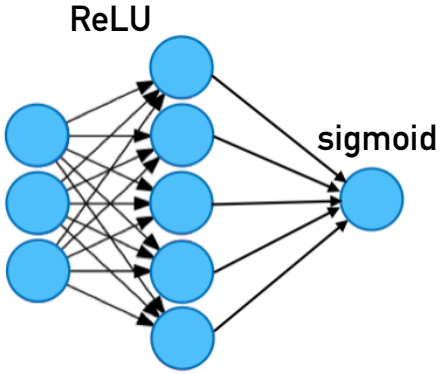
optimal for mean estimation
but **inconsistent** for
covariance estimation

JS-GAN

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



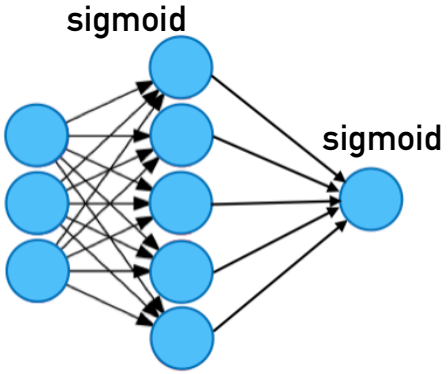
optimal for mean estimation but **inconsistent** for covariance estimation



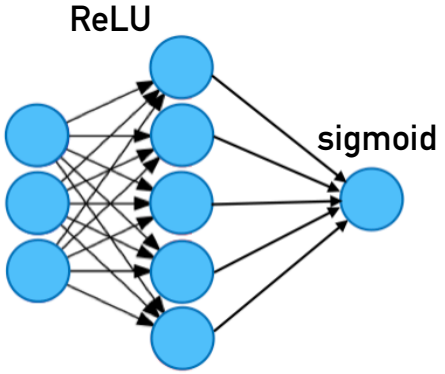
optimal without contamination

JS-GAN

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



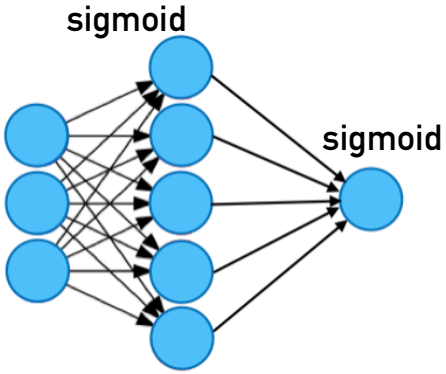
optimal for mean estimation but **inconsistent** for covariance estimation

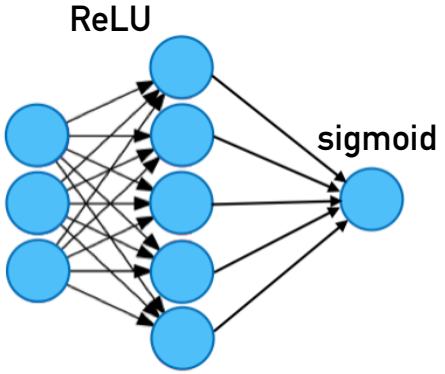


optimal without contamination but **not robust**

JS-GAN

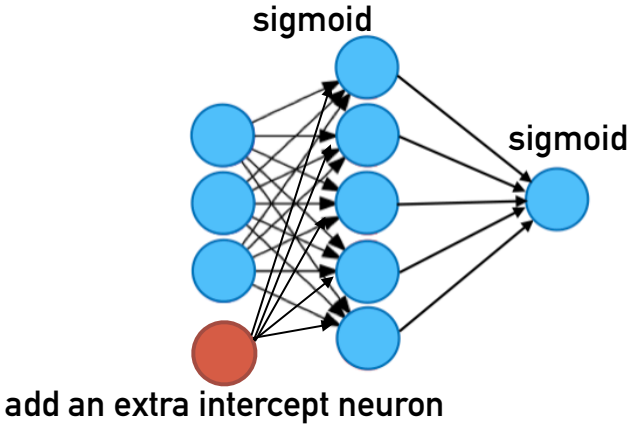
$$\widehat{\Sigma} = \underset{\Gamma}{\operatorname{argmin}} \max_{T \in \mathcal{T}} \left[\frac{1}{n} \sum_{i=1}^{n} \log T(X_i) + \mathbb{E}_{X \sim N(0,\Gamma)} \log(1 - T(X)) \right]$$

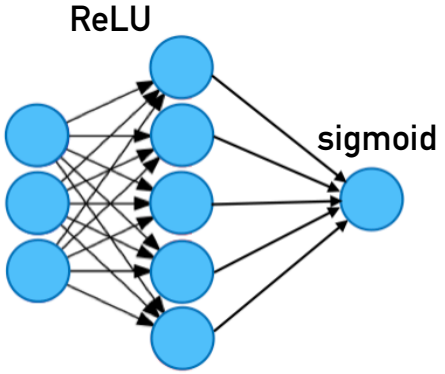




JS-GAN

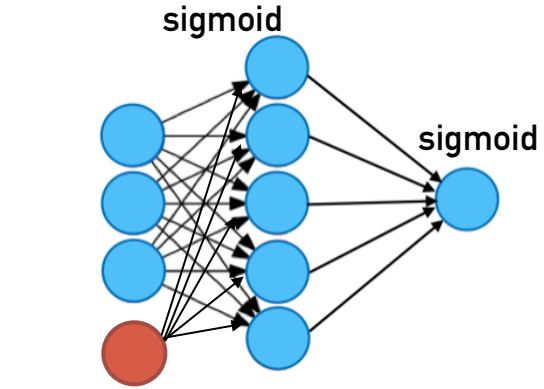
$$\widehat{\Sigma} = \underset{\Gamma}{\operatorname{argmin}} \max_{T \in \mathcal{T}} \left[\frac{1}{n} \sum_{i=1}^{n} \log T(X_i) + \mathbb{E}_{X \sim N(0,\Gamma)} \log(1 - T(X)) \right]$$



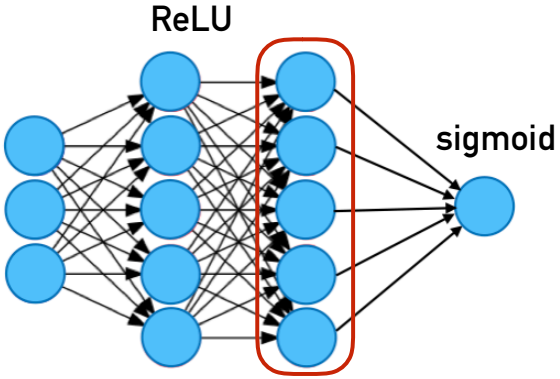


JS-GAN

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



add an extra intercept neuron



add an extra sigmoid layer

JS-GAN

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

Theorem [GYZ18+]. For the above two neural network classes, we have

$$\|\widehat{\Sigma} - \Sigma\|_{\mathrm{op}}^2 \lesssim \begin{cases} \dfrac{p}{n} + \epsilon^2 & \text{(2-layer sigmoid with intercept)} \\ \dfrac{p \log p}{n} + \epsilon^2 & \text{(3-layer ReLU)} \end{cases}$$

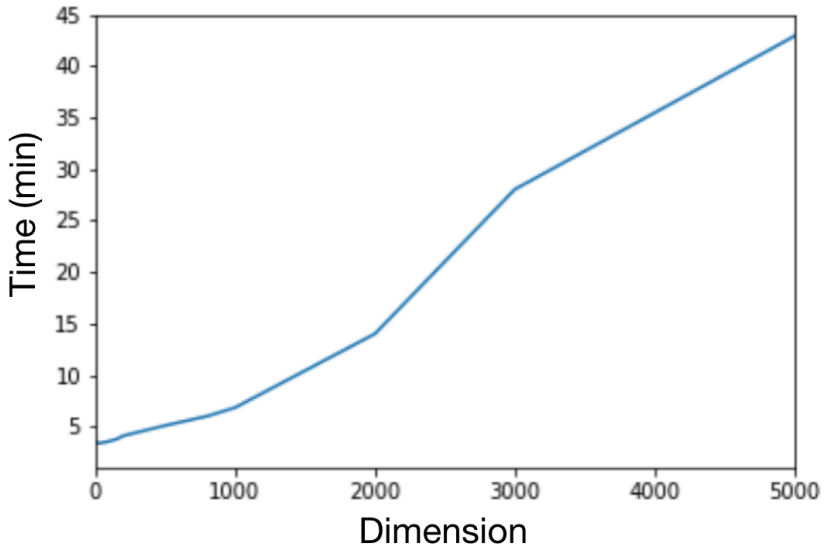
with high probability uniformly over Σ, Q .

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. The smallest error of each case is highlighted in bold.

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

JS-GAN


Time (min)
45
40
35
30
25
20
15
10
5
0
1000
2000
3000
4000
5000
Dimension

Summary

Thank You