Generalization of Neural Network

Qingru Zhang

Apex Data & Knowledge Management Lab
Shanghai Jiao Tong University

Content

- **Generalization Bound**
- **PAC-Bayesian Framework**
- **Randomization Test**
- **Overparameterization**
- **Conclusion**

Generalization Bound

Definition:

$\ell(y, y')$ - loss function

$\mathcal{H}$ - hypothesis space

$h(x)$ - prediction of hypothesis $h \in \mathcal{H}$ on sample x

$L(h) = \mathbb{E}_{(x,y)\sim\mathcal{D}}[\ell(y, h(x))]$ - expected loss of h

$\hat{L}(h) = \frac{1}{m}\sum_{i=1}^{m} \ell(y_i, h(x_i))$ - empirical loss of h

generalization error $= L(h) - \hat{L}(h)$

Generalization Bound

Generalization Bound Overview:

With probability $1-\delta$

$$\hat{L}(h) \le L(h) \le \hat{L}(h) + \Omega(m, R, \delta)$$

E_{train} $\quad$ E_{test} $\quad$ **Smaller δ, larger Ω**

m is the number of training data $\Rightarrow$ **Larger m, smaller Ω**

R is the "capacity" of your model $\Rightarrow$ **Larger R, larger Ω**

("size" of the function set)

Generalization Bound Framework:

- **VC-dimension**
- **PAC-Bayesian**
- **Rademacher complexity** $\hat{\mathfrak{R}}_n(\mathcal{H}) = \mathbb{E}_\sigma \left[\sup_{h \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^{n} \sigma_i h(x_i) \right]$

Generalization Bound

Current Framework:

- **VC-dimension**

$$\text{VC-dim} = \tilde{O}(d * \dim(\mathbf{w}))$$

- **PAC-Bayesian Framework**

- **Rademacher complexity**

$$\hat{\mathfrak{R}}_n(\mathcal{H}) = \mathbb{E}_\sigma \left[\sup_{h \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^{n} \sigma_i h(x_i) \right]$$

General PAC-Bayesian Theorem

Preliminaries:

- ***P*** **is a prior distribution on** ***h***
- ***Q*** **is a posterior distribution on** ***h***
- **Take the loss expectation on** ***Q*****.**

Gibbs Risk / Linear Loss

The stochastic Gibbs classifier $G_Q(x)$ draws $h' \in \mathcal{H}$ according to Q and output $h'(x)$.

$$\begin{aligned} R_D(G_Q) &= \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim D} \mathop{\mathbf{E}}_{h\sim Q} \mathrm{I}\Big[h(x) \neq y\Big] \\ &= \mathop{\mathbf{E}}_{h\sim Q} \mathcal{L}_D^{\ell_{01}}(h)\,, \end{aligned}$$

where $\ell_{01}(h,x,y) = \mathrm{I}\big[h(x) \neq y\big]$.

General PAC-Bayesian Theorem

Δ-function: "distance" between $\widehat{R}_S(G_Q)$ et $R_D(G_Q)$

Convex function $\Delta : [0,1] \times [0,1] \to \mathbb{R}$.

General theorem (*Bégin et al. (2014b, 2016); Germain (2015)*)

For any distribution D on $\mathcal{X} \times \mathcal{Y}$, for any set $\mathcal{H}$ of voters, for any distribution P on $\mathcal{H}$, for any $\delta \in (0,1]$, and for any Δ-function, we have, with probability at least $1-\delta$ over the choice of $S \sim D^m$,

$$\forall\, Q \text{ on } \mathcal{H}: \quad \Delta\Big(\widehat{R}_S(G_Q), R_D(G_Q)\Big) \;\leq\; \frac{1}{m}\left[\mathrm{KL}(Q\|P) + \ln \frac{\mathcal{I}_\Delta(m)}{\delta}\right],$$

where

$$\mathcal{I}_\Delta(m) \;=\; \sup_{r\in[0,1]} \left[\, \sum_{k=0}^{m} \underbrace{\tbinom{m}{k} r^k (1-r)^{m-k}}_{\mathbf{Bin}\left(k;m,r\right)} e^{m\Delta\left(\frac{k}{m},\, r\right)} \right].$$

General PAC-Bayesian Theorem

$\Pr_{S\sim D^m}\left(\forall Q \text{ on } \mathcal{H}:\ \Delta\left(\widehat{R}_S(G_Q), R_D(G_Q)\right) \leq \frac{1}{m}\left[\mathrm{KL}(Q\|P) + \ln\frac{\mathcal{I}_\Delta(m)}{\delta}\right]\right) \geq 1-\delta$. **Proof.**

$$m\cdot\Delta\left(\mathop{\mathbf{E}}_{h\sim Q}\widehat{\mathcal{L}}^{\ell}_S(h),\ \mathop{\mathbf{E}}_{h\sim Q}\mathcal{L}^{\ell}_D(h)\right)$$

Jensen's Inequality

$$\leq \quad \mathop{\mathbf{E}}_{h\sim Q} m\cdot\Delta\left(\widehat{\mathcal{L}}^{\ell}_S(h), \mathcal{L}^{\ell}_D(h)\right)$$

Change of measure

$$\leq \quad \mathrm{KL}(Q\|P) + \ln \mathop{\mathbf{E}}_{h\sim P} e^{m\Delta\left(\widehat{\mathcal{L}}^{\ell}_S(h), \mathcal{L}^{\ell}_D(h)\right)}$$

Markov's Inequality

$$\leq_{1-\delta} \quad \mathrm{KL}(Q\|P) + \ln\frac{1}{\delta}\mathop{\mathbf{E}}_{S'\sim D^m}\mathop{\mathbf{E}}_{h\sim P} e^{m\cdot\Delta(\widehat{\mathcal{L}}^{\ell}_{S'}(h), \mathcal{L}^{\ell}_D(h))}$$

Expectation swap

$$= \quad \mathrm{KL}(Q\|P) + \ln\frac{1}{\delta}\mathop{\mathbf{E}}_{h\sim P}\mathop{\mathbf{E}}_{S'\sim D^m} e^{m\cdot\Delta(\widehat{\mathcal{L}}^{\ell}_{S'}(h), \mathcal{L}^{\ell}_D(h))}$$

Binomial law

$$= \quad \mathrm{KL}(Q\|P) + \ln\frac{1}{\delta}\mathop{\mathbf{E}}_{h\sim P}\sum_{k=0}^{m}\mathbf{Bin}\left(k; m, \mathcal{L}^{\ell}_D(h)\right) e^{m\cdot\Delta(\frac{k}{m}, \mathcal{L}^{\ell}_D(h))}$$

Supremum over risk

$$\leq \quad \mathrm{KL}(Q\|P) + \ln\frac{1}{\delta}\sup_{r\in[0,1]}\left[\sum_{k=0}^{m}\mathbf{Bin}\left(k; m, r\right) e^{m\Delta(\frac{k}{m}, r)}\right]$$

$$= \quad \mathrm{KL}(Q\|P) + \ln\frac{1}{\delta}\mathcal{I}_\Delta(m)\,. \qquad \square$$

Norm-based PAC-Bayesian Bound

Margin loss

$$L_\gamma(f_{\mathbf{w}}) = \mathbb{P}_{(\mathbf{x},y)\sim\mathcal{D}}\left[f_{\mathbf{w}}(\mathbf{x})[y] \le \gamma + \max_{j\ne y} f_{\mathbf{w}}(\mathbf{x})[j]\right]$$

PAC-Bayesian Lemma

$$L_0(f_{\mathbf{w}}) \le \widehat{L}_\gamma(f_{\mathbf{w}}) + 4\sqrt{\frac{KL\left(\mathbf{w}+\mathbf{u}\|P\right) + \ln\frac{6m}{\delta}}{m-1}}.$$

Perturbation Bound

$$|f_{\mathbf{w}+\mathbf{u}}(\mathbf{x}) - f_{\mathbf{w}}(\mathbf{x})|_2 \le eB\left(\prod_{i=1}^{d}\|W_i\|_2\right)\sum_{i=1}^{d}\frac{\|U_i\|_2}{\|W_i\|_2}.$$

Generalization Bound

$$L_0(f_{\mathbf{w}}) \le \widehat{L}_\gamma(f_{\mathbf{w}}) + \mathcal{O}\left(\sqrt{\frac{B^2d^2h\ln(dh)\Pi_{i=1}^{d}\|W_i\|_2^2\sum_{i=1}^{d}\frac{\|W_i\|_F^2}{\|W_i\|_2^2} + \ln\frac{dm}{\delta}}{\gamma^2 m}}\right).$$

https://arxiv.org/abs/1707.09564

APEX 数据和知识管理实验室
DATA & KNOWLEDGE MANAGEMENT LAB

No-vacuous PAC-Bayesian Bound

A compression approach:

- **different network has different compressibility**
- **Construct prior P according to compressibility**

$$\pi_c(h) = \frac{1}{Z} m(|h|_c) 2^{-|h|_c}, \text{ where } Z = \sum_{h \in \mathcal{H}_c} m(|h|_c) 2^{-|h|_c}.$$

- **Give larger probability to more compressible network**
- **Reduce KL-distance between P and Q**

$$\mathrm{KL}(\rho_{S,C,Q}, \pi) \leq (k\lceil \log r \rceil + |S|_c + |C|_c) \log 2 - \log m(k\lceil \log r \rceil + |S|_c + |C|_c) + \sum_{i=1}^{k} \mathrm{KL}\Big(\mathrm{Normal}(c_{q_i}, \sigma^2), \sum_{j=1}^{r} \mathrm{Normal}(c_j, \tau^2)\Big).$$

https://arxiv.org/abs/1804.05862

No-vacuous PAC-Bayesian Bound

Bound:

$$\mathrm{KL}(\rho_{S,C,Q},\pi) \leq (k\lceil\log r\rceil + |S|_c + |C|_c)\log 2 - \log m(k\lceil\log r\rceil + |S|_c + |C|_c) + \sum_{i=1}^{k}\mathrm{KL}\Big(\mathrm{Normal}(c_{q_i},\sigma^2), \sum_{j=1}^{r}\mathrm{Normal}(c_j,\tau^2)\Big).$$

Experiments:

Dataset	Orig. size	Comp. size	Robust. Adj.	Eff. Size	Error Bound Top 1	Error Bound Top 5
MNIST	168.4 KiB	8.1 KiB	1.88 KiB	6.23 KiB	$< 46\,\%$	NA
ImageNet	5.93 MiB	452 KiB	102 KiB	350 KiB	$< 96.5\,\%$	$< 89\,\%$

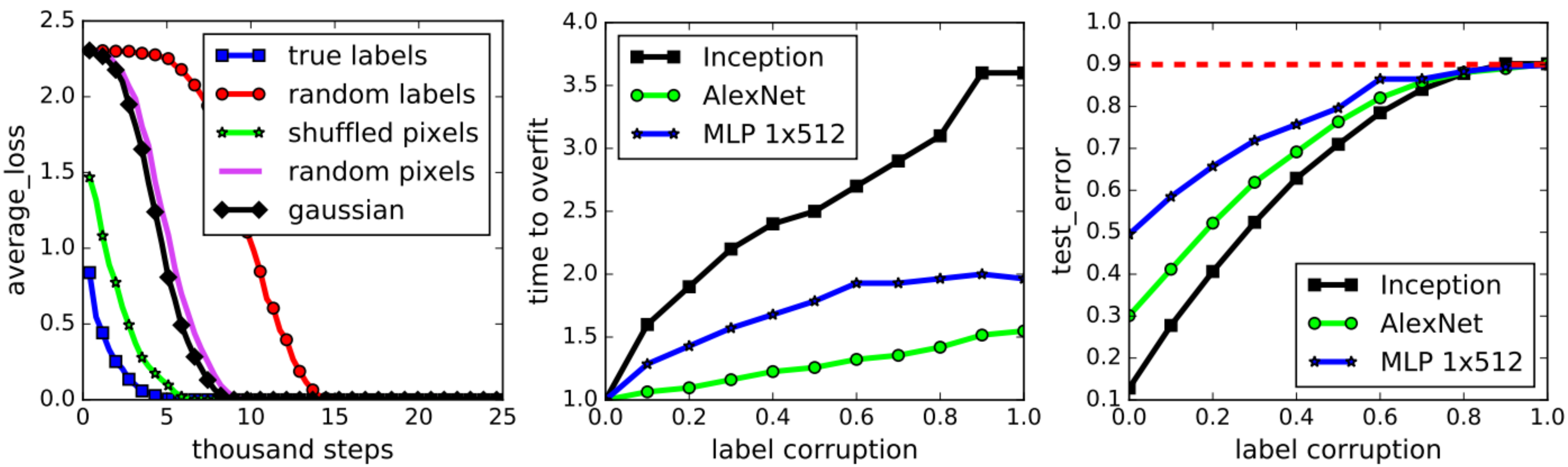
https://arxiv.org/abs/1804.05862

Randomization tests

Test Insight: randomizing labels alone to force the generalization error jumping up without changing the model.

Result:

Deep neural networks easily fit random labels.



(a) learning curves (b) convergence slowdown (c) generalization error growth

APEX 数据和知识管理实验室
DATA & KNOWLEDGE MANAGEMENT LAB

Randomization tests

Expectation:

- Since initially the label is uncorrelated
- Large predictions errors are back-propagated
- Make large gradients for parameter updates

Observation:

- No need to change the learning rate schedule
- Once the fitting starts, it converges quickly
- it converges to (over) fit the training set perfectly.

Challenge for Complexity Measures

Rademacher complexity

$$\hat{\mathfrak{R}}_n(\mathcal{H}) = \mathbb{E}_\sigma \left[\sup_{h \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^{n} \sigma_i h(x_i) \right]$$

- **Since many neural networks fit the training set with random labels perfectly, we expect that R ≈ 1, which is a trivial upper bound.**

VC-dimension

$$\text{VC-dim} = \tilde{O}(d * \dim(\mathbf{w}))$$

- **When the number of parameters is more than the number of samples, it becomes too weak.**

The role of regularization

Regularization technique:

- **Data augmentation, Weight Decay, Dropout.**

Result:

model	# params	random crop	weight decay	train accuracy	test accuracy
Inception	1,649,402	yes	yes	100.0	89.05
		yes	no	100.0	89.31
		no	yes	100.0	86.03
		no	no	100.0	85.75
(fitting random labels)		no	no	100.0	9.78
Inception w/o BatchNorm	1,649,402	no	yes	100.0	83.00
		no	no	100.0	82.00
(fitting random labels)		no	no	100.0	10.12
Alexnet	1,387,786	yes	yes	99.90	81.22
		yes	no	99.82	79.66
		no	yes	100.0	77.36
		no	no	100.0	76.07
(fitting random labels)		no	no	99.82	9.86
MLP 3x512	1,735,178	no	yes	100.0	53.35
		no	no	100.0	52.39
(fitting random labels)		no	no	100.0	10.48
MLP 1x512	1,209,866	no	yes	99.80	50.39
		no	no	100.0	50.51
(fitting random labels)		no	no	99.34	10.61

The role of regularization

Conclusion:

Both regularization techniques help to improve the generalization performance

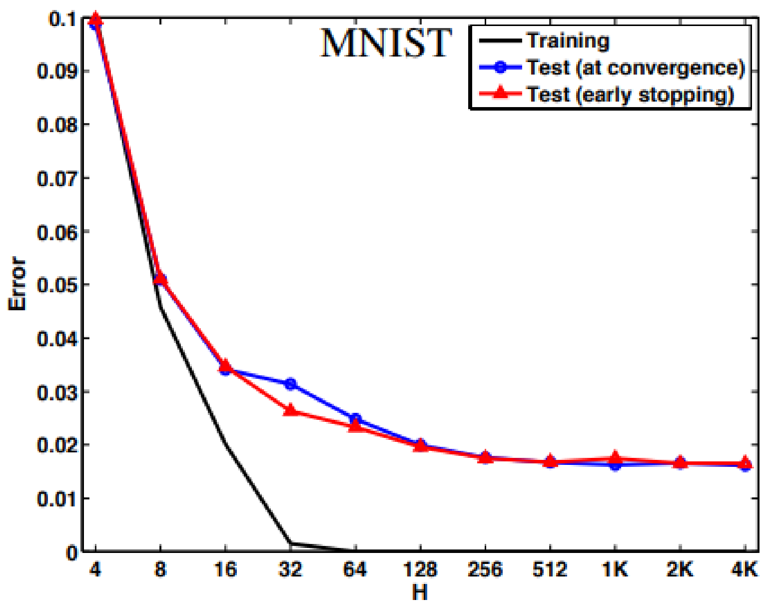
But turned off all regularizers, all of the models still generalize very well.

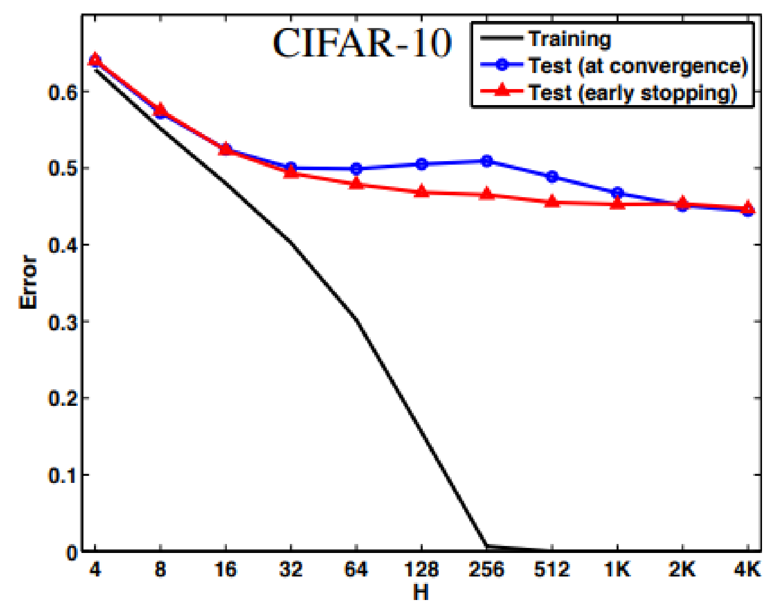
Some Extension:

- **SGD acts as an implicit regularizer**
- **For linear models, SGD always converges to a solution with small norm.**
- **What the properties are inherited by models that were trained using SGD?**

Overparameterized Network

- **Overparameterized Network still can generalize**
- ***IN SEARCH OF THE REAL INDUCTIVE BIAS*** **point out implicit regularization of SGD**

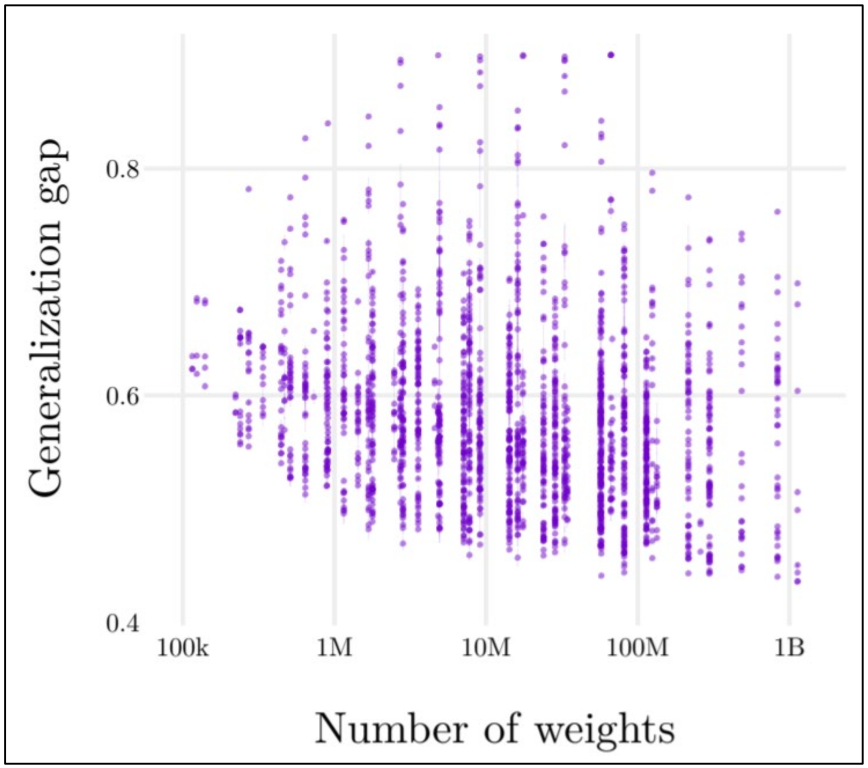




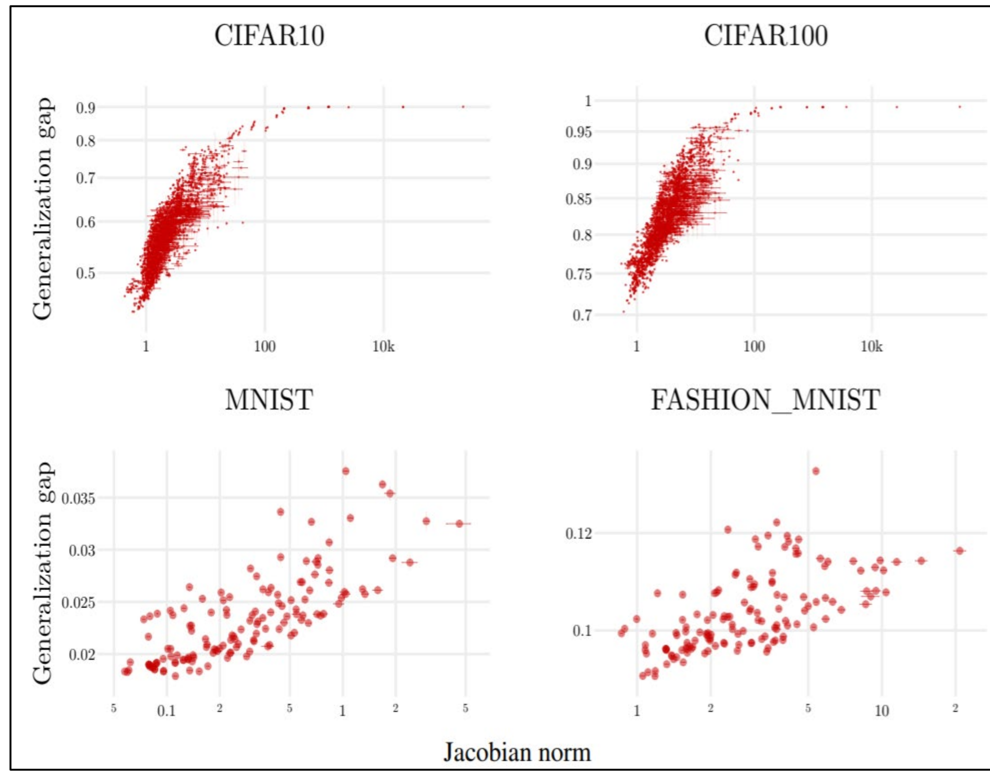
数据和知识管理实验室
DATA & KNOWLEDGE MANAGEMENT LAB

Overparameterized Network

- **Overparameterized Network still can generalize**
- ***SENSITIVITY AND GENERALIZATION IN NEURAL NETWORKS* focus on norm of Jocabian**



CIFIR-10 100% training accuracy



F-norm

APEX 数据和知识管理实验室
DATA & KNOWLEDGE MANAGEMENT LAB

Overparameterized Network

Conclusion

- Overparameterized Network's capacity is large
- But it does not overfit and still can generalize
- The reason is unclear yet.
    - SGD's implicit regularization
    - Network regularizes itself
    - ……

Conclusion

- Overparameterized network still generalize
- Neural network can fit any dataset, random label or true label
- Theoretic analysis should consider dataset structure.
- Neural network owns some properties:
  - robust to perturbation of parameters
  - compressibility
- State-of-art theoretic bounds are still far from being effective

Q & A

Thanks for Listening!