Harvard John A. Paulson School of Engineering and Applied Sciences

Robustness Implies Generalization via Data-Dependent Generalization Bounds

Zhun Deng*

Harvard University

* Joining Columbia University in fall as a postdoc.

Collaborators

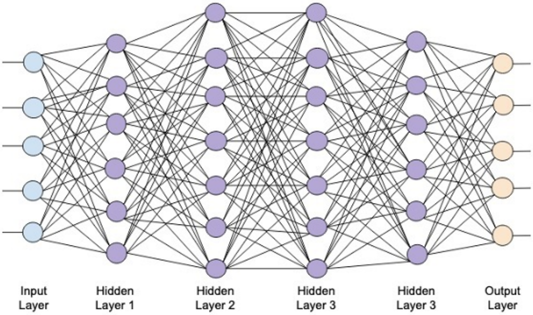
Kenji Kawaguchi

Kyle Luh

Jiaoyang Huang

New Generalization Bounds are Needed in Modern Learning

Mysteries of Modern Machine Learning

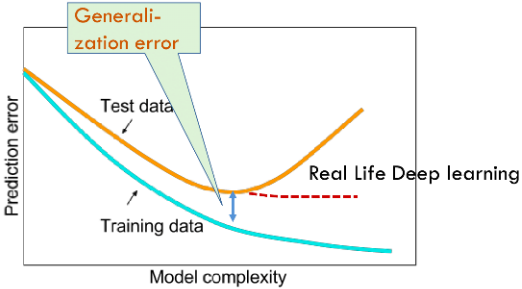


Neural networks:

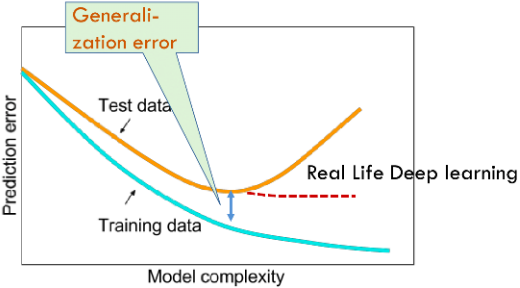
Severely over-parameterized.

Still generalize well?!

Mysteries of Modern Machine Learning

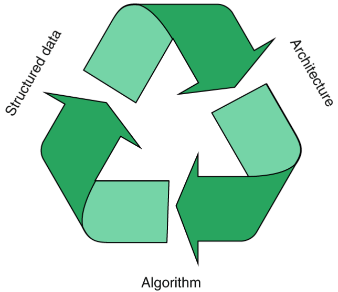


Mysteries of Modern Machine Learning



Traditional generalization bounds can no longer work!

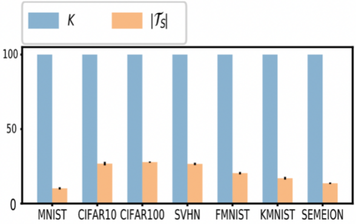
Mysteries of Modern Machine Learning

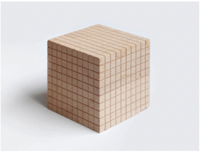


Three keys to demystify — "Understanding Deep Learning is Also a Job for Physicists" by Lenka Zdeborov*á*

Data-dependent Generalization Bounds are Essential in Modern Learning

Modern Dataset Structure

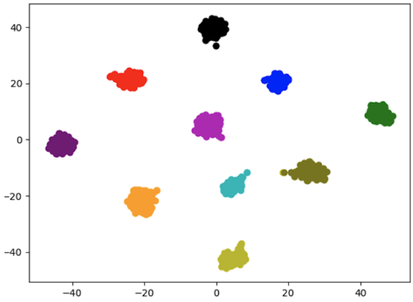




K is a partition of input space of real data sets; $|\mathcal{T}_S|$ is the number of partitions with non-zero data points.

Popolar modern datasets are very sparse! Actually, the datasets are spares after projection, so they live on low dimensional manifolds.

Modern Dataset Structure



2-D visualization of Cifar-10's representation embeddings after projection.

Traditional Generalization Bounds

Rademacher complexity bounds (Bartlett & Mendelson, 2002)

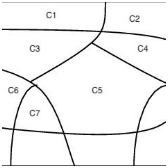
Uniform stability bounds (Bousquet & Elisseef, 2002)

Robust generalization bounds (Xu & Mannor, 2012)

All those bounds cannot directly take advantage of the input data structure!

Definition ($(K, \epsilon(\cdot))$-robust)

1. Algorithm $\mathcal{A} : \mathcal{Z}^n \to \mathbb{R}$;
2. The input space $\mathcal{Z}$ can be partitioned into K disjoint sets – $\{\mathcal{C}_k\}_{k=1}^{K}$;
3. if $s, z \in \mathcal{C}_k,$ then $|\ell(\mathcal{A}_S, s) - \ell(\mathcal{A}_S, z)| \leq \epsilon(S)$.



Robust Generalization Bounds

Proposition (Xu & Mannor, 2012)

1. *$\ell(h, z) \leq B$;*
2. *$\mathcal{A}$ is $(K, \epsilon(\cdot))$-robust (with $\{\mathcal{C}_k\}_{k=1}^{K}$);*

with probability at least $1 - \delta$,

$$\mathbb{E}_z[\ell(\mathcal{A}_S, z)] \leq \frac{1}{n}\sum_{i=1}^{n} \ell(\mathcal{A}_S, z_i) + \epsilon(S) + B\sqrt{\frac{2K \ln 2 + 2\ln(1/\delta)}{n}}.$$

Why Robust Generalization Bounds?

Example (Xu & Mannor, 2012 (Lasso))

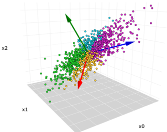
$\mathcal{Z}$ is compact, and loss function $\ell(\mathcal{A}_S, z) = |z^{(y)} - \mathcal{A}_S(z^{(x)})|$.

Lasso can be formulated as:

$$\underset{w}{\text{minimize}} : \frac{1}{n}\sum_{i=1}^{n}(s_i^{(y)} - w^\top s_i^{(x)})^2 + c\|w\|_1.$$

This algorithm is $(\mathcal{N}(\nu/2, \mathcal{Z}, \|\cdot\|_\infty), \nu(\frac{1}{n}\sum_{i=1}^{n}(s_i^{(y)})^2)/c + \nu)$-robust for all $\nu > 0$.

Why Robust Generalization Bounds?



Example (Xu & Mannor, 2012 (PCA))

For $\mathcal{Z} \subset \mathbb{R}^m$, a set with the maximum ℓ_2 norm bounded by B, with loss function

$$\ell((w_1, \ldots, w_d), z) = \sum_{j=1}^{d} (w_j^\top z)^2,$$

then finding the first d principal components via the optimization problem:

$$\text{Maximize: } \sum_{i=1}^{n} \sum_{j=1}^{d} (w_j^\top s_i)^2$$

with the constraint that $\|w_j\|_2 = 1$ and $w_i^\top w_j = 0$ for $i \neq j$ is $(\mathcal{N}(\gamma/2, \mathcal{Z}, \|\cdot\|_2), 2d\gamma B)$-robust, for all $\gamma > 0$.

Our Bounds

Theorem

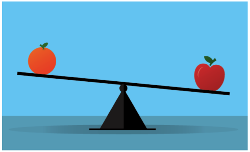
1. $\ell(h, z) \leq B$;
2. $\mathcal{A}$ *is* $(K, \epsilon(\cdot))$*-robust (with* $\{\mathcal{C}_k\}_{k=1}^K$*);*

with probability at least $1 - \delta$*, the following holds:*

$$\mathbb{E}_z[\ell(\mathcal{A}_S, z)] \leq \frac{1}{n}\sum_{i=1}^{n} \ell(\mathcal{A}_S, z_i) + \epsilon(S) + \zeta(\mathcal{A}_S)\left((\sqrt{2}+1)\sqrt{\frac{|\mathcal{T}_S| \ln(2K/\delta)}{n}} + \frac{2|\mathcal{T}_S| \ln(2K/\delta)}{n}\right),$$

where $\mathcal{I}_k^S := \{i \in [n] : z_i \in \mathcal{C}_k\}$, $\zeta(\mathcal{A}_S) := \max_{z \in \mathcal{Z}}\{\ell(\mathcal{A}_S, z)\}$, *and* $\mathcal{T}_S := \{k \in [K] : |\mathcal{I}_k^S| \geq 1\}$.

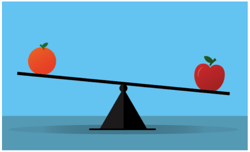
Comparisons



1. (Xu & Mannor, 2012) $B\sqrt{\frac{2K\ln 2+2\ln(1/\delta)}{n}}$
2. (Ours) $\zeta(\mathcal{A}_S)\left((\sqrt{2}+1)\sqrt{\frac{|\mathcal{T}_S|\ln(2K/\delta)}{n}}+\frac{2|\mathcal{T}_S|\ln(2K/\delta)}{n}\right)$

where $\mathcal{I}_k^S := \{i \in [n] : z_i \in \mathcal{C}_k\}$, $\zeta(\mathcal{A}_S) := \max_{z\in\mathcal{Z}}\{\ell(\mathcal{A}_S, z)\}$, and $\mathcal{T}_S := \{k \in [K] : |\mathcal{I}_k^S| \geq 1\}$.

Comparisons



1. (Xu & Mannor, 2012) $B\sqrt{\frac{2K\ln 2+2\ln(1/\delta)}{n}}$
2. (Ours) $\zeta(\mathcal{A}_S)\left((\sqrt{2}+1)\sqrt{\frac{|\mathcal{T}_S|\ln(2K/\delta)}{n}}+\frac{2|\mathcal{T}_S|\ln(2K/\delta)}{n}\right)$

where $\mathcal{I}_k^S := \{i \in [n] : z_i \in \mathcal{C}_k\}$, $\zeta(\mathcal{A}_S) := \max_{z\in\mathcal{Z}}\{\ell(\mathcal{A}_S, z)\}$, and $\mathcal{T}_S := \{k \in [K] : |\mathcal{I}_k^S| \geq 1\}$.

$$K \text{ v.s. } |\mathcal{T}_S|$$

Comparisons

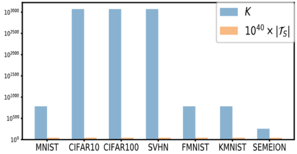


Figure 3. The values of K versus $|\mathcal{T}_S|$ with real-world data and the ϵ-covering. The values of $|\mathcal{T}_S|$ are extremely small compared to those of K in all datasets.

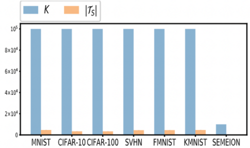


Figure 4. The values of K versus $|\mathcal{T}_S|$ with real-world data and the clustering using unlabeled data. With clustering to reduce K, we still have $|\mathcal{T}_S| < K$. Here, $|\mathcal{T}_S|$ was close to zero for Semeion.

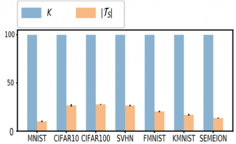


Figure 5. The values of K versus $|\mathcal{T}_S|$ with real-world data and random projection. With random projection to reduce K, we still have $|\mathcal{T}_S| < 30 < K = 100 < n \approx 60,000$ for the real-life datasets. Here, n is the full train data size of each dataset: e.g., $n = 60,000$ for MNIST.

$$K \gg |\mathcal{T}_S|$$

Theoretical Comparisons

Proposition

1. $p_k = \mathbb{P}(z \in \mathcal{C}_k)$ *where* $p_1 \geq p_2 \geq \cdots \geq p_K$;
2. p_k *decays as* $p_k \leq Ce^{-(k/\beta)^\alpha}$;

with probability at least $1 - \delta$,

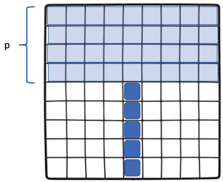
$$|\mathcal{T}_S| \leq \beta(\ln n)^{1/\alpha} + C(e-1)\frac{\beta}{\alpha} + \log(1/\delta).$$

Example (Lasso)

1. Recall that Lasso is $(\mathcal{N}(\nu/2, \mathcal{Z}, \|\cdot\|_\infty), \nu(\frac{1}{n}\sum_{i=1}^n (s_i^{(y)})^2)/c + \nu)$-robust for all $\nu > 0$.
2. Consider $z^{(y)} \in \mathbb{R}$ and $z^{(x)} \in \mathbb{R}^d$. Given any $\nu > 0$, let z follow a distribution $\mathcal{D}_z$, such that $z^{(x)} = ({x^{(1)}}^\top, {x^{(2)}}^\top)^\top$, where $x^{(1)} \sim N(0, I_p)|_{[-1,1]^p}$. $x^{(2)} \sim N(\mu, \sigma^2 I_r)|_{[-1,1]^r}$, and $r = d - p$, $z^{(y)} = {w^*}^\top z^{(x)}$, where $\|w^*\|_1 \leq 1$.

Example (Lasso)

1. Recall that Lasso is $(\mathcal{N}(\nu/2, \mathcal{Z}, \|\cdot\|_\infty), \nu(\frac{1}{n}\sum_{i=1}^{n}(s_i^{(y)})^2)/c + \nu)$-robust for all $\nu > 0$.
2. Consider $z^{(y)} \in \mathbb{R}$ and $z^{(x)} \in \mathbb{R}^d$. Given any $\nu > 0$, let z follow a distribution $\mathcal{D}_z$, such that $z^{(x)} = (x^{(1)\top}, x^{(2)\top})^\top$, where $x^{(1)} \sim N(0, I_p)|_{[-1,1]^p}$. $x^{(2)} \sim N(\mu, \sigma^2 I_r)|_{[-1,1]^r}$, and $r = d - p$, $z^{(y)} = w^{*\top} z^{(x)}$, where $\|w^*\|_1 \leq 1$.



Theoretical Comparisons

Example (Lasso)

1. Recall that Lasso is $(\mathcal{N}(\nu/2, \mathcal{Z}, \|\cdot\|_\infty), \nu(\frac{1}{n}\sum_{i=1}^{n}(s_i^{(y)})^2)/c + \nu)$-robust for all $\nu > 0$.
2. Consider $z^{(y)} \in \mathbb{R}$ and $z^{(x)} \in \mathbb{R}^d$. Given any $\nu > 0$, let z follow a distribution $\mathcal{D}_z$, such that $z^{(x)} = (x^{(1)^\top}, x^{(2)^\top})^\top$, where $x^{(1)} \sim N(0, I_p)|_{[-1,1]^p}$. $x^{(2)} \sim N(\mu, \sigma^2 I_r)|_{[-1,1]^r}$, and $r = d - p$, $z^{(y)} = w^{*\top} z^{(x)}$, where $\|w^*\|_1 \leq 1$.

Theoretical Comparisons

Example (Lasso)

1. Recall that Lasso is $(\mathcal{N}(\nu/2, \mathcal{Z}, \|\cdot\|_\infty), \nu(\frac{1}{n}\sum_{i=1}^n (s_i^{(y)})^2)/c + \nu)$-robust for all $\nu > 0$.
2. Consider $z^{(y)} \in \mathbb{R}$ and $z^{(x)} \in \mathbb{R}^d$. Given any $\nu > 0$, let z follow a distribution $\mathcal{D}_z$, such that $z^{(x)} = ({x^{(1)}}^\top, {x^{(2)}}^\top)^\top$, where $x^{(1)} \sim N(0, I_p)|_{[-1,1]^p}$. $x^{(2)} \sim N(\mu, \sigma^2 I_r)|_{[-1,1]^r}$, and $r = d - p$, $z^{(y)} = {w^*}^\top z^{(x)}$, where $\|w^*\|_1 \leq 1$.

There exists parameters such that our bound is much tighter than that in Proposition in Xu & Mannor as
$|\mathcal{T}_\mathcal{S}| = \Theta((2/\nu)^{p+1}) \ll \Theta((2/\nu)^{d+1}) = \mathcal{N}(\nu/2, \mathcal{Z}, \|\cdot\|_\infty)$.

Algorithmic Stability

Training sample: $\mathcal{S}$.

Algorithmic Stability

Training sample: $\mathcal{S}$.

A learning algorithm $\mathscr{A} : \mathcal{Z}^m \to \mathcal{F}$ outputs a function $\mathcal{A}_S \in \mathcal{F}$.

Training sample: $\mathcal{S}$.

A learning algorithm $\mathscr{A} : \mathcal{Z}^m \to \mathcal{F}$ outputs a function $\mathcal{A}_S \in \mathcal{F}$.

Definition (Uniform Stability(Bousquet & Elisseeff, 2002))

An algorithm $\mathscr{A}$ has uniform stability β_m^{U} with respect to the loss function l if

$$|l(\mathcal{A}_S, z) - l(\mathcal{A}_{S^{\setminus i}}, z)| \leq \beta_m^{\mathsf{U}}$$

holds for all $S \in \mathcal{Z}^m, 1 \leq i \leq m$, and $z \in \mathcal{Z}$.

Theoretical Comparisons with Uniform Stability

Example (Regularized least square regression)

1. Let $z^{(y)} \in [0, B]$ and $z^{(x)} \in [0, 1]$. The regularized least squares regression is defined as $\mathcal{A}_S = \operatorname{argmin}_w \frac{1}{n} \sum_{i=1}^{n} \ell(w, z_i) + \lambda |w|^2$, where $\ell(w, z) = (w \cdot z^{(x)} - z^{(y)})^2$ and $w \in \mathbb{R}$.
2. Uniform stability: $\beta \leq \frac{2B^2}{\lambda n}$.
3. Consider z: $z^{(y)} = w^* \cdot z^{(x)} + \epsilon \mathbf{1}(|\epsilon| < B)$. In addition, $z^{(x)}$ follows a continuous distribution on $[0, 1]$.

With a probability of at least $1 - \delta$, $|\mathcal{T}_S| = \Theta(2\nu)$. Thus, if $B^2/\lambda \gg 2/\nu$, then our bound is a **far more** precise bound than that obtained via uniform stability.

New Techniques

The key technical hurdle: to avoid an explicit $\sqrt{K}$ dependence for the following form:

$$\sum_{i=1}^{K} a_i(X) \left(p_i - \frac{X_i}{n} \right),$$

where a_i is an arbitrary function with $a_i(X) \geq 0$ for all $i \in \{1, \dots, K\}$.

New Techniques

Lemma

For any $\delta > 0$, with probability at least $1 - \delta$,

$$\sum_{i=1}^{K} a_i(X) \left(p_i - \frac{X_i}{n} \right) \leq \left(\sum_{i=1}^{K} a_i(X) \sqrt{p_i} \right) \sqrt{\frac{2 \ln(K/\delta)}{n}}.$$

1. Lemma holds with $a_i(X) = \mathrm{sign}(p_i - \frac{X_i}{n})$, where $\mathrm{sign}(q)$ is the sign of q.
2. If $p_i = 1/K$, recovers Bretagnolle-Huber-Carol inequality.
3. $p_1 \approx 1$, other p_i's are ≈ 0, then $\sum \sqrt{p_i} \approx 1$.

New Techniques

Lemma

For any $\delta > 0$, with probability at least $1-\delta$,

$$\sum_{i=1}^{K} a_i(X)\left(p_i - \frac{X_i}{n}\right) \leq \left(\sum_{i=1}^{K} a_i(X)\sqrt{p_i}\right)\sqrt{\frac{2\ln(K/\delta)}{n}}.$$

1. Lemma holds with $a_i(X) = \mathrm{sign}(p_i - \frac{X_i}{n})$, where $\mathrm{sign}(q)$ is the sign of q.
2. If $p_i = 1/K$, recovers Bretagnolle-Huber-Carol inequality.
3. $p_1 \approx 1$, other p_i's are ≈ 0, then $\sum \sqrt{p_i} \approx 1$.

Our result interpolates between these cases!

End

Thanks for listening!