

Machine Learning Theory 2025

Lecture 6

Tim van Erven

- ▶ Rademacher complexity controls uniform convergence (for any bounded loss)
- ▶ Rademacher calculus
- ▶ Beyond PAC-Learning

Rademacher Complexity in General

Consider any supervised learning task:

- ▶ **Hypothesis class** $\mathcal{H}$: some set of functions h from $\mathcal{X}$ to $\mathcal{Y}$
- ▶ **Loss**: $\ell(h, \mathbf{X}, Y)$

Rademacher complexity:

$$\mathcal{R}(\ell, \mathcal{H}, S) = \frac{1}{m} \mathbb{E}_{\sigma} \left[\sup_{h \in \mathcal{H}} \sum_{i=1}^m \sigma_i \ell(h, \mathbf{X}_i, Y_i) \right]$$

where *Rademacher random variables*

$$\sigma = (\sigma_1, \dots, \sigma_m) \in \{-1, +1\}^m$$

are i.i.d. with

$$\Pr(\sigma_i = -1) = \Pr(\sigma_i = +1) = 1/2.$$

Concentration and Symmetrization for Any Bounded Loss

For $\ell(h, \mathbf{X}, Y) \in [0, 1]$, abbreviate $Z = \sup_{h \in \mathcal{H}} L_{\mathcal{D}}(h) - L_S(h)$.

1. Concentration:

$$\mathbb{E}_S[Z] - \sqrt{\frac{\ln(2/\delta)}{2m}} \leq Z \leq \mathbb{E}_S[Z] + \sqrt{\frac{\ln(2/\delta)}{2m}} \quad \text{w.p.} \geq 1 - \delta.$$

2. Symmetrization:

$$\mathbb{E}_S[Z] \leq 2 \mathbb{E}_S[\mathcal{R}(\ell, \mathcal{H}, S)]$$

Proved last week:

1. McDiarmid's Inequality
2. Symmetrization by a 'ghost' sample S'
 - NB This step does not require $\ell(h, \mathbf{X}, Y) \in [0, 1]$

Concentration and Symmetrization for Any Bounded Loss

For $\ell(h, \mathbf{X}, Y) \in [0, 1]$, abbreviate $Z = \sup_{h \in \mathcal{H}} L_{\mathcal{D}}(h) - L_S(h)$.

1. Concentration:

$$\mathbb{E}_S[Z] - \sqrt{\frac{\ln(2/\delta)}{2m}} \leq Z \leq \mathbb{E}_S[Z] + \sqrt{\frac{\ln(2/\delta)}{2m}} \quad \text{w.p.} \geq 1 - \delta.$$

2. Symmetrization:

$$\mathbb{E}_S[Z] \leq 2 \mathbb{E}_S[\mathcal{R}(\ell, \mathcal{H}, S)]$$

$$\sup_{h \in \mathcal{H}} L_{\mathcal{D}}(h) - L_S(h) \leq 2 \mathbb{E}_S[\mathcal{R}(\ell, \mathcal{H}, S)] + \sqrt{\frac{\ln(2/\delta)}{2m}} \quad \text{w.p.} \geq 1 - \delta.$$

Uniform Convergence via Rademacher Complexity

Lemma

Consider any supervised learning task with $\ell(h, \mathbf{X}, Y) \in [0, 1]$. Then

$$\sup_{h \in \mathcal{H}} |L_{\mathcal{D}}(h) - L_S(h)| \leq 2 \mathbb{E}[\mathcal{R}(\ell, \mathcal{H}, S)] + \sqrt{\frac{\ln(4/\delta)}{2m}} \quad w.p. \geq 1 - \delta.$$

Uniform Convergence via Rademacher Complexity

Lemma

Consider any supervised learning task with $\ell(h, \mathbf{X}, Y) \in [0, 1]$. Then

$$\sup_{h \in \mathcal{H}} |L_{\mathcal{D}}(h) - L_S(h)| \leq 2 \mathbb{E}[\mathcal{R}(\ell, \mathcal{H}, S)] + \sqrt{\frac{\ln(4/\delta)}{2m}} \quad \text{w.p.} \geq 1 - \delta.$$

Proof:

$$\sup_{h \in \mathcal{H}} L_{\mathcal{D}}(h) - L_S(h) \leq 2 \mathbb{E}_S[\mathcal{R}(\ell, \mathcal{H}, S)] + \sqrt{\frac{\ln(2/\delta)}{2m}} \quad \text{w.p.} \geq 1 - \delta.$$

1. Apply with ℓ and $\ell' = 1 - \ell$ + union bound. Then, w.p. $\geq 1 - \delta$,

$$\sup_{h \in \mathcal{H}} |L_{\mathcal{D}}(h) - L_S(h)| \leq 2 \max \left\{ \mathbb{E}_S[\mathcal{R}(\ell, \mathcal{H}, S)], \mathbb{E}_S[\mathcal{R}(1 - \ell, \mathcal{H}, S)] \right\} + \sqrt{\frac{\ln(4/\delta)}{2m}}.$$

2. $\mathcal{R}(1 - \ell, \mathcal{H}, S) = \mathcal{R}(\ell, \mathcal{H}, S)$ by **Rademacher calculus**

Uniform Convergence via Rademacher Complexity

Lemma

Consider any supervised learning task with $\ell(h, \mathbf{X}, Y) \in [0, 1]$. Then

$$\sup_{h \in \mathcal{H}} |L_{\mathcal{D}}(h) - L_S(h)| \leq 2 \mathbb{E}[\mathcal{R}(\ell, \mathcal{H}, S)] + \sqrt{\frac{\ln(4/\delta)}{2m}} \quad \text{w.p.} \geq 1 - \delta.$$

Recall that uniform convergence is **sufficient** for agnostic PAC-learnability:

If $h_S \in \arg \min_{h \in \mathcal{H}} L_S(h)$ is **ERM hypothesis**, then

$$L_{\mathcal{D}}(h_S) - \inf_{h \in \mathcal{H}} L_{\mathcal{D}}(h) \leq 2 \sup_{h \in \mathcal{H}} |L_{\mathcal{D}}(h) - L_S(h)|$$

But for learning tasks other than binary classification, uniform convergence may **not** be a **necessary** requirement.

Uniform Convergence via Rademacher Complexity

Lemma

Consider any supervised learning task with $\ell(h, \mathbf{X}, Y) \in [0, 1]$. Then

$$\sup_{h \in \mathcal{H}} |L_{\mathcal{D}}(h) - L_S(h)| \leq 2 \mathbb{E}[\mathcal{R}(\ell, \mathcal{H}, S)] + \sqrt{\frac{\ln(4/\delta)}{2m}} \quad w.p. \geq 1 - \delta.$$

Is this bound tight?

Uniform Convergence via Rademacher Complexity

Lemma

Consider any supervised learning task with $\ell(h, \mathbf{X}, Y) \in [0, 1]$. Then

$$\sup_{h \in \mathcal{H}} |L_{\mathcal{D}}(h) - L_S(h)| \leq 2 \mathbb{E}[\mathcal{R}(\ell, \mathcal{H}, S)] + \sqrt{\frac{\ln(4/\delta)}{2m}} \quad w.p. \geq 1 - \delta.$$

Is this bound tight? **YES!**

Rademacher complexity sandwiches uniform convergence
for bounded losses!

Lemma (Converse Bound*)

Consider any supervised learning task with $\ell(h, \mathbf{X}, Y) \in [0, 1]$. Then

$$\sup_{h \in \mathcal{H}} |L_{\mathcal{D}}(h) - L_S(h)| \geq \frac{1}{2} \mathbb{E}[\mathcal{R}(\ell, \mathcal{H}, S)] - \sqrt{\frac{2 \ln(2/\delta)}{m}} \quad w.p. \geq 1 - \delta.$$

*Converse bound is bonus, will not be on the exam.

Converse Bound

Lemma (Converse Bound*)

Consider any supervised learning task with $\ell(h, \mathbf{X}, Y) \in [0, 1]$. Then

$$\sup_{h \in \mathcal{H}} |L_{\mathcal{D}}(h) - L_S(h)| \geq \frac{1}{2} \mathbb{E}[\mathcal{R}(\ell, \mathcal{H}, S)] - \sqrt{\frac{2 \ln(2/\delta)}{m}} \quad \text{w.p.} \geq 1 - \delta.$$

Proof: Let $Z = \sup_{h \in \mathcal{H}} |L_{\mathcal{D}}(h) - L_S(h)|$.

1. McDiarmid: Z satisfies bounded differences with $b = 1/m$, so

$$Z \geq \mathbb{E}_S[Z] - \sqrt{\frac{\ln(2/\delta)}{2m}} \quad \text{w.p.} \geq 1 - \delta.$$

2. **Desymmetrization:**

$$\mathbb{E}_S[Z] \geq \frac{1}{2} \mathbb{E}_S[\mathcal{R}(\ell, \mathcal{H}, S)] - \sqrt{\frac{\ln 2}{2m}}.$$

Desymmetrization

Lemma:

$$\mathbb{E} \left[\sup_{h \in \mathcal{H}} |L_0(h) - L_S(h)| \right] \geq \frac{1}{2} \mathbb{E} \left[R(\mathcal{L}, \mathcal{H}, S) \right] - \sqrt{\frac{\ln 2}{2m}}$$

Proof: Abbreviate $Z_i(h) = \ell(h, X_i, Y_i)$, $Z'_i(h) = \ell(h, X'_i, Y'_i)$

Key symmetrization identity (*):

for all $\sigma \in \{-1, +1\}^m$:

$$\mathbb{E} \left[\sup_{s, s'} \frac{1}{m} \sum_{i=1}^m Z_i(h) - Z'_i(h) \right] = \mathbb{E} \left[\sup_{s, s'} \frac{1}{m} \sum_{i=1}^m \sigma_i (Z_i(h) - Z'_i(h)) \right]$$

where $s' = \begin{pmatrix} y'_1 \\ x'_1 \end{pmatrix}, \dots, \begin{pmatrix} y'_m \\ x'_m \end{pmatrix}$

$$\mathbb{E}[\mathcal{R}(\mathcal{L}, \mathcal{H}, S)] = \mathbb{E} \left[\sup_{S, \sigma} \frac{1}{n} \sum_{i=1}^n \sigma_i z_i(h) \right]$$

$$\leq \underbrace{\mathbb{E} \left[\sup_{S, \sigma} \frac{1}{n} \sum_{i=1}^n \sigma_i (z_i(h) - L_D(h)) \right]}_I + \underbrace{\mathbb{E} \left[\sup_{\sigma} \frac{1}{n} \sum_{i=1}^n \sigma_i L_D(h) \right]}_{II}$$

$$I = \mathbb{E} \left[\sup_{S, \sigma} \mathbb{E}_{S'} \left[\frac{1}{n} \sum_{i=1}^n \sigma_i (z_i(h) - z'_i(h)) \right] \right]$$

$$\leq \mathbb{E} \mathbb{E} \left[\sup_{\sigma, S, S'} \frac{1}{n} \sum_{i=1}^n \sigma_i (z_i(h) - z'_i(h)) \right]$$

$$\stackrel{\text{by (*)}}{=} \mathbb{E} \left[\sup_{S, S'} \frac{1}{n} \sum_{i=1}^n (z_i(h) - z'_i(h)) \right]$$

$$= \mathbb{E} \left[\sup_{S, S'} \frac{1}{n} \sum_{i=1}^n (z_i(h) - L_D(h) + L_D(h) - z'_i(h)) \right]$$

$$\begin{aligned}
&\leq \mathbb{E} \left[\sup_S \frac{1}{m} \sum_{i=1}^m (Z_i(L) - L_D(h)) \right] \\
&\quad + \mathbb{E} \left[\sup_{S'} \frac{1}{m} \sum_{i=1}^m (L_D(h) - Z_i'(h)) \right] \\
&\leq 2 \mathbb{E} \left[\sup_S |L_D(h) - L_S(h)| \right]
\end{aligned}$$

$$\begin{aligned}
II &\leq \mathbb{E}_{\sigma} \left[\max \left\{ 0 \cdot \frac{1}{m} \sum_{i=1}^m \sigma_i, 1 \cdot \frac{1}{m} \sum_{i=1}^m \sigma_i \right\} \right] \\
&= \mathbb{E}_{\sigma} \left[\max_{a \in \left\{ \begin{pmatrix} 0 \\ 1 \end{pmatrix}, \begin{pmatrix} 1 \\ 1 \end{pmatrix} \right\}} \frac{1}{m} \sum_{i=1}^m \sigma_i \cdot a_i \right] \\
&\leq \sqrt{\frac{2 \ln 2}{m}} \quad (\text{by Massart's Lemma})
\end{aligned}$$

$$\mathbb{E} [R(\mathcal{L}, \mathcal{H}, S)] \leq 2 \mathbb{E} \left[\sup_S |L_D(h) - L_S(h)| \right] + \sqrt{\frac{2 \ln 2}{m}}$$

$$\mathbb{E} \left[\sup_S |L_D(h) - L_S(h)| \right] \geq \frac{1}{2} \mathbb{E} [R(\mathcal{L}, \mathcal{H}, S)] - \sqrt{\frac{\ln 2}{2m}} \quad \square$$

Rademacher Calculus 1

More abstractly, Rademacher complexity depends on a **set of vectors**:

$$\mathcal{A} = \left\{ \left(\ell(h, \mathbf{X}_1, Y_1), \dots, \ell(h, \mathbf{X}_m, Y_m) \right) : h \in \mathcal{H} \right\} \subset \mathbb{R}^m$$

$$\mathcal{R}(\ell, \mathcal{H}, S) = \frac{1}{m} \mathbb{E}_{\sigma} \left[\sup_{h \in \mathcal{H}} \sum_{i=1}^m \sigma_i \ell(h, \mathbf{X}_i, Y_i) \right]$$
$$\mathcal{R}(\mathcal{A}) = \frac{1}{m} \mathbb{E}_{\sigma} \left[\sup_{\mathbf{a} \in \mathcal{A}} \sum_{i=1}^m \sigma_i a_i \right]$$

Rademacher complexity behaves very nicely
under **certain operations on $\mathcal{A}$** !

Rademacher Calculus 2

$$\mathcal{R}(\mathcal{A}) = \frac{1}{m} \mathbb{E}_{\sigma} \left[\sup_{a \in \mathcal{A}} \sum_{i=1}^m \sigma_i a_i \right]$$

Enlarging the Class:

$$\mathcal{R}(\mathcal{A}) \leq \mathcal{R}(\mathcal{B}) \quad \text{for } \mathcal{A} \subset \mathcal{B}$$

Rademacher Calculus 2

$$\mathcal{R}(\mathcal{A}) = \frac{1}{m} \mathbb{E}_{\sigma} \left[\sup_{\mathbf{a} \in \mathcal{A}} \sum_{i=1}^m \sigma_i a_i \right]$$

Enlarging the Class:

$$\mathcal{R}(\mathcal{A}) \leq \mathcal{R}(\mathcal{B}) \quad \text{for } \mathcal{A} \subset \mathcal{B}$$

Affine Transformations:

$$\mathcal{R}(\{c\mathbf{a} + \mathbf{a}_0 : \mathbf{a} \in \mathcal{A}\}) = |c| \mathcal{R}(\mathcal{A}) \quad \text{for any } c \in \mathbb{R}, \mathbf{a}_0 \in \mathbb{R}^m$$

► E.g. $\mathcal{R}(1 - \ell, \mathcal{H}, S) = |-1| \mathcal{R}(\ell, \mathcal{H}, S) = \mathcal{R}(\ell, \mathcal{H}, S)$

Rademacher Calculus 2

$$\mathcal{R}(\mathcal{A}) = \frac{1}{m} \mathbb{E}_{\sigma} \left[\sup_{\mathbf{a} \in \mathcal{A}} \sum_{i=1}^m \sigma_i a_i \right]$$

Enlarging the Class:

$$\mathcal{R}(\mathcal{A}) \leq \mathcal{R}(\mathcal{B}) \quad \text{for } \mathcal{A} \subset \mathcal{B}$$

Affine Transformations:

$$\mathcal{R}(\{c\mathbf{a} + \mathbf{a}_0 : \mathbf{a} \in \mathcal{A}\}) = |c| \mathcal{R}(\mathcal{A}) \quad \text{for any } c \in \mathbb{R}, \mathbf{a}_0 \in \mathbb{R}^m$$

► E.g. $\mathcal{R}(1 - \ell, \mathcal{H}, S) = |-1| \mathcal{R}(\ell, \mathcal{H}, S) = \mathcal{R}(\ell, \mathcal{H}, S)$

Convex Hull:

$$\mathcal{R}(\mathcal{A}) = \mathcal{R}(\text{conv}(\mathcal{A}))$$

Rademacher Calculus 3: Advanced Properties

Contraction: Let $\phi \circ \mathcal{A} = \{(\phi_1(a_1), \dots, \phi_m(a_m)) : \mathbf{a} \in \mathcal{A}\}$.

If $\phi_i : \mathbb{R} \rightarrow \mathbb{R}$ is **L-Lipschitz**: $|\phi_i(\alpha) - \phi_i(\beta)| \leq L|\alpha - \beta|$, then:

$$\mathcal{R}(\phi \circ \mathcal{A}) \leq L\mathcal{R}(\mathcal{A})$$

Rademacher Calculus 3: Advanced Properties

Contraction: Let $\phi \circ \mathcal{A} = \{(\phi_1(a_1), \dots, \phi_m(a_m)) : \mathbf{a} \in \mathcal{A}\}$.

If $\phi_i : \mathbb{R} \rightarrow \mathbb{R}$ is **L-Lipschitz**: $|\phi_i(\alpha) - \phi_i(\beta)| \leq L|\alpha - \beta|$, then:

$$\mathcal{R}(\phi \circ \mathcal{A}) \leq L\mathcal{R}(\mathcal{A})$$

Example: Get rid of Lipschitz loss function using $\phi_i(z) = |Y_i - z|$:

$$\mathcal{R}\left(\{(|Y_1 - h(\mathbf{X}_1)|, \dots, |Y_m - h(\mathbf{X}_m)|) : h \in \mathcal{H}\}\right) \leq \mathcal{R}\left(\{(h(\mathbf{X}_1), \dots, h(\mathbf{X}_m)) : h \in \mathcal{H}\}\right)$$

Rademacher Calculus 3: Advanced Properties

Contraction: Let $\phi \circ \mathcal{A} = \{(\phi_1(a_1), \dots, \phi_m(a_m)) : \mathbf{a} \in \mathcal{A}\}$.

If $\phi_i : \mathbb{R} \rightarrow \mathbb{R}$ is **L-Lipschitz**: $|\phi_i(\alpha) - \phi_i(\beta)| \leq L|\alpha - \beta|$, then:

$$\mathcal{R}(\phi \circ \mathcal{A}) \leq L\mathcal{R}(\mathcal{A})$$

Example: Get rid of Lipschitz loss function using $\phi_i(z) = |Y_i - z|$:

$$\mathcal{R}\left(\{(|Y_1 - h(\mathbf{X}_1)|, \dots, |Y_m - h(\mathbf{X}_m)|) : h \in \mathcal{H}\}\right) \leq \mathcal{R}\left(\{(h(\mathbf{X}_1), \dots, h(\mathbf{X}_m)) : h \in \mathcal{H}\}\right)$$

Massart's Lemma:

Suppose $|\mathcal{A}| = N$ is finite. Then $\mathcal{R}(\mathcal{A}) \leq \max_{\mathbf{a} \in \mathcal{A}} \|\mathbf{a}\| \sqrt{\frac{2 \ln N}{m}}$.

Corollary: If $\mathbf{a} \in [-1, +1]^m$ for all $\mathbf{a} \in \mathcal{A}$, then $\mathcal{R}(\mathcal{A}) \leq \sqrt{\frac{2 \ln N}{m}}$.

► E.g. $\mathcal{R}\left(\left\{\begin{pmatrix} 0 \\ \vdots \\ 0 \end{pmatrix}, \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix}\right\}\right) \leq \sqrt{\frac{2 \ln 2}{m}}$, as used in desymmetrization proof

Rademacher Calculus 3: Advanced Properties

Contraction: Let $\phi \circ \mathcal{A} = \{(\phi_1(a_1), \dots, \phi_m(a_m)) : \mathbf{a} \in \mathcal{A}\}$.

If $\phi_i : \mathbb{R} \rightarrow \mathbb{R}$ is **L-Lipschitz**: $|\phi_i(\alpha) - \phi_i(\beta)| \leq L|\alpha - \beta|$, then:

$$\mathcal{R}(\phi \circ \mathcal{A}) \leq L\mathcal{R}(\mathcal{A})$$

Example: Get rid of Lipschitz loss function using $\phi_i(z) = |Y_i - z|$:

$$\mathcal{R}\left(\{(|Y_1 - h(\mathbf{X}_1)|, \dots, |Y_m - h(\mathbf{X}_m)|) : h \in \mathcal{H}\}\right) \leq \mathcal{R}\left(\{(h(\mathbf{X}_1), \dots, h(\mathbf{X}_m)) : h \in \mathcal{H}\}\right)$$

Massart's Lemma:

Suppose $|\mathcal{A}| = N$ is finite. Then $\mathcal{R}(\mathcal{A}) \leq \max_{\mathbf{a} \in \mathcal{A}} \|\mathbf{a}\| \sqrt{\frac{2 \ln N}{m}}$.

Corollary: If $\mathbf{a} \in [-1, +1]^m$ for all $\mathbf{a} \in \mathcal{A}$, then $\mathcal{R}(\mathcal{A}) \leq \sqrt{\frac{2 \ln N}{m}}$.

► E.g. $\mathcal{R}\left(\left\{\begin{pmatrix} 0 \\ \vdots \\ 0 \end{pmatrix}, \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix}\right\}\right) \leq \sqrt{\frac{2 \ln 2}{m}}$, as used in desymmetrization proof

Remark: For binary classification we proved that: $\mathcal{R}(\ell, \mathcal{H}, S) = \mathcal{R}(\ell, \mathcal{H}_S, S) \leq \sqrt{\frac{2 \ln |\mathcal{H}_S|}{m}}$.
General proof goes along the same lines.

Example: Bounded Regression with Lasso

$$\mathcal{H}_1^B = \{h_{\mathbf{w}}(\mathbf{X}) = \langle \mathbf{w}, \mathbf{X} \rangle : \mathbf{w} \in \mathbb{R}^d, \|\mathbf{w}\|_1 \leq B\}.$$

Theorem (Lasso Estimator)

Consider linear regression with $\ell(h, \mathbf{X}, Y) = \frac{1}{2}(Y - \langle \mathbf{w}, \mathbf{X} \rangle)^2$ for $\mathbf{X} \in [-1, +1]^d$, $Y \in [-1, +1]$.

Then $\mathcal{H}_1^B$ is agnostically PAC-learnable by ERM with sample complexity

$$m(\epsilon, \delta) \leq c_B \frac{\ln(2d) + \ln(4/\delta)}{\epsilon^2}$$

for some constant $c_B > 0$ that depends only on B .

Example: Bounded Regression with Lasso

$$\mathcal{H}_1^B = \{h_w(\mathbf{X}) = \langle \mathbf{w}, \mathbf{X} \rangle : \mathbf{w} \in \mathbb{R}^d, \|\mathbf{w}\|_1 \leq B\}.$$

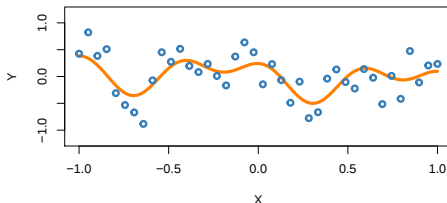
Theorem (Lasso Estimator)

Consider linear regression with $\ell(h, \mathbf{X}, Y) = \frac{1}{2}(Y - \langle \mathbf{w}, \mathbf{X} \rangle)^2$ for $\mathbf{X} \in [-1, +1]^d$, $Y \in [-1, +1]$.

Then $\mathcal{H}_1^B$ is agnostically PAC-learnable by ERM with sample complexity

$$m(\epsilon, \delta) \leq c_B \frac{\ln(2d) + \ln(4/\delta)}{\epsilon^2}$$

for some constant $c_B > 0$ that depends only on B .



Possible $\mathcal{D}$:

Example: Bounded Regression with Lasso

$$\mathcal{H}_1^B = \{h_w(\mathbf{X}) = \langle \mathbf{w}, \mathbf{X} \rangle : \mathbf{w} \in \mathbb{R}^d, \|\mathbf{w}\|_1 \leq B\}.$$

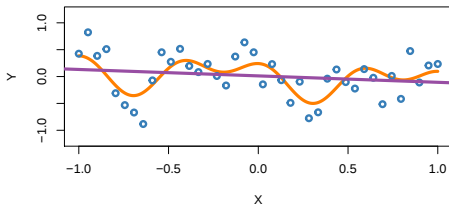
Theorem (Lasso Estimator)

Consider linear regression with $\ell(h, \mathbf{X}, Y) = \frac{1}{2}(Y - \langle \mathbf{w}, \mathbf{X} \rangle)^2$ for $\mathbf{X} \in [-1, +1]^d$, $Y \in [-1, +1]$.

Then $\mathcal{H}_1^B$ is agnostically PAC-learnable by ERM with sample complexity

$$m(\epsilon, \delta) \leq c_B \frac{\ln(2d) + \ln(4/\delta)}{\epsilon^2}$$

for some constant $c_B > 0$ that depends only on B .



Possible $\mathcal{D}$:

NB Do not assume that $Y = h(\mathbf{X}) + \text{noise}$ for any $h \in \mathcal{H}$!

Example: Bounded Regression with Lasso

$$\mathcal{H}_1^B = \{h_w(\mathbf{X}) = \langle \mathbf{w}, \mathbf{X} \rangle : \mathbf{w} \in \mathbb{R}^d, \|\mathbf{w}\|_1 \leq B\}.$$

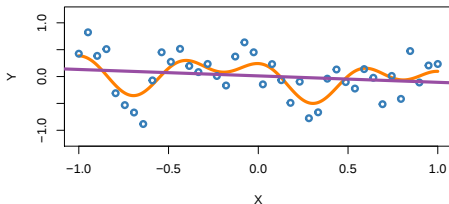
Theorem (Lasso Estimator)

Consider linear regression with $\ell(h, \mathbf{X}, Y) = \frac{1}{2}(Y - \langle \mathbf{w}, \mathbf{X} \rangle)^2$ for $\mathbf{X} \in [-1, +1]^d$, $Y \in [-1, +1]$.

Then $\mathcal{H}_1^B$ is agnostically PAC-learnable by ERM with sample complexity

$$m(\epsilon, \delta) \leq c_B \frac{\ln(2d) + \ln(4/\delta)}{\epsilon^2}$$

for some constant $c_B > 0$ that depends only on B .



Possible $\mathcal{D}$:

Proof: Homework

► Hint: apply the tools from this lecture

NB Do not assume that $Y = h(\mathbf{X}) + \text{noise}$ for any $h \in \mathcal{H}$!

Beyond PAC-Learning

PAC-Learning Guarantees are Very Strong

Requires learning with the **same sample complexity** $m(\epsilon, \delta)$ for

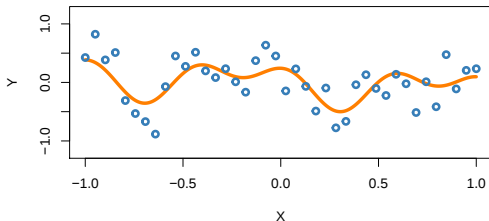
- ▶ all distributions $\mathcal{D}$, and
- ▶ all hypotheses $h \in \mathcal{H}$.

PAC-Learning Guarantees are Very Strong

Requires learning with the **same sample complexity** $m(\epsilon, \delta)$ for

- ▶ all distributions $\mathcal{D}$, and
- ▶ all hypotheses $h \in \mathcal{H}$.

Distributions:



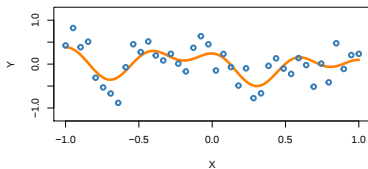
- ▶ Distributions still restricted via $(\mathcal{X}, \mathcal{Y})$, e.g. bounded regression
- ▶ Uniform convergence not possible in unbounded regression...
- ▶ ... unless we **restrict class of possible distributions**

PAC-Learning Guarantees are Very Strong

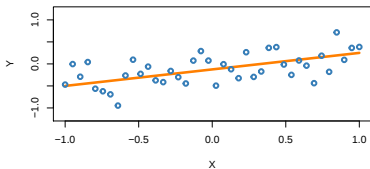
Requires learning with the **same sample complexity** $m(\epsilon, \delta)$ for

- ▶ all distributions $\mathcal{D}$, and
- ▶ all hypotheses $h \in \mathcal{H}$.

Hypotheses:



Complex function h



Simple function h

- ▶ **Non-uniform learnability**: allow sample complexity $m^{\text{NUL}}(\epsilon, \delta, h)$ to depend on (complexity of) h

Non-uniform Learning

$\mathcal{H}$ is **agnostically PAC-learnable**:

Exists learner (selecting $h_S \in \mathcal{H}$) that achieves, for finite $m_{\mathcal{H}}(\epsilon, \delta)$,

$$L_{\mathcal{D}}(h_S) \leq \inf_{h \in \mathcal{H}} L_{\mathcal{D}}(h) + \epsilon \quad \text{with probability } \geq 1 - \delta,$$

whenever $m \geq m_{\mathcal{H}}(\epsilon, \delta)$,

for all $\mathcal{D}, \epsilon, \delta$.

Non-uniform Learning

$\mathcal{H}$ is **non-uniform learnable**:

Exists learner (selecting $h_S \in \mathcal{H}$) that achieves, for **all** $h \in \mathcal{H}$, finite $m_{\mathcal{H}}(\epsilon, \delta, h)$,

$$L_{\mathcal{D}}(h_S) \leq \text{---} L_{\mathcal{D}}(h) + \epsilon \quad \text{with probability } \geq 1 - \delta,$$

whenever $m \geq m_{\mathcal{H}}(\epsilon, \delta, h)$,

for all $\mathcal{D}, \epsilon, \delta$.