Machine Learning Theory 2024
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**: $\ell(h, \boldsymbol{X}, Y)$

Rademacher complexity:

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

where *Rademacher random variables*

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

are i.i.d. with

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

Concentration and Symmetrization for Any Bounded Loss

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

1. **Concentration:**

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

2. **Symmetrization:**

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

Proved last week:

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

Concentration and Symmetrization for Any Bounded Loss

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

1. **Concentration:**

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

2. **Symmetrization:**

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

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

Uniform Convergence via Rademacher Complexity

Lemma

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

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

Uniform Convergence via Rademacher Complexity

Lemma

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

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

Proof:

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

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

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

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 $\ell(h, \boldsymbol{X}, Y) \in [0,1]$. Then

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

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 $\ell(h, \boldsymbol{X}, Y) \in [0,1]$. Then

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

Is this bound tight?

Uniform Convergence via Rademacher Complexity

Lemma

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

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

Is this bound tight? **YES!**
Rademacher complexity sandwiches uniform convergence
for bounded losses!

Lemma (Converse Bound*)

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

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

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

Converse Bound

Lemma (Converse Bound*)

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

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

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

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

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

2. **Desymmetrization:**

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

Desymmetrization

Lemma:

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

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

$$\mathbb{E}_{S,S'} \left[\sup_{h\in\mathcal{H}} \frac{1}{m} \sum_{i=1}^m z_i(h) - z_i'(h) \right] = \mathbb{E}_{S,S'} \left[\sup_{h\in\mathcal{H}} \frac{1}{m} \sum_{i=1}^m \sigma_i (z_i(h) - z_i'(h)) \right]$$

where $S' = \begin{pmatrix} y_1' \\ x_1' \end{pmatrix}, \ldots, \begin{pmatrix} y_m' \\ x_m' \end{pmatrix}$

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

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

$$\mathrm{I} = \mathbb{E}_{S,\sigma}\left[\sup_{h\in\mathcal{H}} \mathbb{E}_{S'}\left[\frac{1}{m}\sum_{i=1}^{m} \sigma_i (z_i(h) - z_i'(h))\right]\right]$$

$$\leq \mathbb{E}_{\sigma}\mathbb{E}_{S,S'}\left[\sup_{h\in\mathcal{H}} \frac{1}{m}\sum_{i=1}^{m} \sigma_i (z_i(h) - z_i'(h))\right]$$

$$\overset{\text{by } (*)}{=} \mathbb{E}_{S,S'}\left[\sup_{h\in\mathcal{H}} \frac{1}{m}\sum_{i=1}^{m} (z_i(h) - z_i'(h))\right]$$

$$= \mathbb{E}_{S,S'}\left[\sup_{h\in\mathcal{H}} \frac{1}{m}\sum_{i=1}^{m} (z_i(h) - L_D(h) + L_D(h) - z_i'(h))\right]$$

$$\leq \mathbb{E}_{S}\left[\sup_{h\in\mathcal{H}} \frac{1}{m}\sum_{i=1}^{m}\left(Z_i(h) - L_D(h)\right)\right]$$

$$+ \mathbb{E}_{S'}\left[\sup_{h\in\mathcal{H}} \frac{1}{m}\sum_{i=1}^{m}\left(L_D(h) - Z_i'(h)\right)\right]$$

$$\leq 2\,\mathbb{E}_{S}\left[\sup_{h\in\mathcal{H}} |L_D(h) - L_S(h)|\right]$$

$$\mathrm{II} \leq \mathbb{E}_{\sigma}\left[\max\left\{0\cdot\frac{1}{m}\sum_{i=1}^{m}\sigma_i,\ 1\cdot\frac{1}{m}\sum_{i=1}^{m}\sigma_i\right\}\right]$$

$$= \mathbb{E}_{\sigma}\left[\max_{a\in\left\{\begin{pmatrix}0\\ \vdots\\ 0\end{pmatrix},\begin{pmatrix}1\\ \vdots\\ 1\end{pmatrix}\right\}} \frac{1}{m}\sum_{i=1}^{m}\sigma_i\cdot a_i\right]$$

$$\leq \sqrt{\frac{2\ln 2}{m}} \qquad \text{(by Massart's Lemma)}$$

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

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

Rademacher Calculus 1

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

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

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

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

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

Rademacher Calculus 2

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

Enlarging the Class:
$\mathcal{R}(\mathcal{A}) \leq \mathcal{R}(\mathcal{B}) \qquad$ for $\mathcal{A} \subset \mathcal{B}$

Rademacher Calculus 2

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

Enlarging the Class:
$\mathcal{R}(\mathcal{A}) \leq \mathcal{R}(\mathcal{B})$ for $\mathcal{A} \subset \mathcal{B}$

Affine Transformations:
$\mathcal{R}(\{c\boldsymbol{a} + \boldsymbol{a}_0 : \boldsymbol{a} \in \mathcal{A}\}) = |c|\mathcal{R}(\mathcal{A})$ for any $c \in \mathbb{R}$, $\boldsymbol{a}_0 \in \mathbb{R}^m$

- 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

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

Enlarging the Class:
$\mathcal{R}(\mathcal{A}) \leq \mathcal{R}(\mathcal{B}) \qquad$ for $\mathcal{A} \subset \mathcal{B}$

Affine Transformations:
$\mathcal{R}(\{c\boldsymbol{a} + \boldsymbol{a}_0 : \boldsymbol{a} \in \mathcal{A}\}) = |c|\mathcal{R}(\mathcal{A}) \qquad$ for any $c \in \mathbb{R}$, $\boldsymbol{a}_0 \in \mathbb{R}^m$

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

Convex Hull:
$\mathcal{R}(\mathcal{A}) = \mathcal{R}(\mathrm{conv}(\mathcal{A}))$

Rademacher Calculus 3: Advanced Properties

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

If $\phi_i : \mathbb{R} \to \mathbb{R}$ 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 $\phi \circ \mathcal{A} = \Big\{\big(\phi_1(a_1), \ldots, \phi_m(a_m)\big) : \boldsymbol{a} \in \mathcal{A}\Big\}$.

If $\phi_i : \mathbb{R} \to \mathbb{R}$ 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 $\phi_i(z) = |Y_i - z|$:

$$\mathcal{R}\Big(\big\{(|Y_1 - h(\boldsymbol{X}_1)|, \ldots, |Y_m - h(\boldsymbol{X}_m)|) : h \in \mathcal{H}\big\}\Big) \leq \mathcal{R}\Big(\big\{(h(\boldsymbol{X}_1), \ldots, h(\boldsymbol{X}_m)) : h \in \mathcal{H}\big\}\Big)$$

Rademacher Calculus 3: Advanced Properties

Contraction: Let $\phi \circ \mathcal{A} = \left\{ \left(\phi_1(a_1), \ldots, \phi_m(a_m) \right) : \boldsymbol{a} \in \mathcal{A} \right\}$.

If $\phi_i : \mathbb{R} \to \mathbb{R}$ 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 $\phi_i(z) = |Y_i - z|$:

$$\mathcal{R}\Big(\big\{ (|Y_1 - h(\boldsymbol{X}_1)|, \ldots, |Y_m - h(\boldsymbol{X}_m)|) : h \in \mathcal{H} \big\} \Big) \leq \mathcal{R}\Big(\big\{ (h(\boldsymbol{X}_1), \ldots, h(\boldsymbol{X}_m)) : h \in \mathcal{H} \big\} \Big)$$

Massart's Lemma:
Suppose $|\mathcal{A}| = N$ is finite. Then $\mathcal{R}(\mathcal{A}) \leq \max_{\boldsymbol{a} \in \mathcal{A}} \|\boldsymbol{a}\| \frac{\sqrt{2 \ln N}}{m}$.

Corollary: If $\boldsymbol{a} \in [-1, +1]^m$ for all $\boldsymbol{a} \in \mathcal{A}$, 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 $\phi \circ \mathcal{A} = \left\{ \left(\phi_1(a_1), \ldots, \phi_m(a_m) \right) : \boldsymbol{a} \in \mathcal{A} \right\}$.

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

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

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

$$\mathcal{R}\Big(\big\{ (|Y_1 - h(\boldsymbol{X}_1)|, \ldots, |Y_m - h(\boldsymbol{X}_m)|) : h \in \mathcal{H} \big\} \Big) \le \mathcal{R}\Big(\big\{ (h(\boldsymbol{X}_1), \ldots, h(\boldsymbol{X}_m)) : h \in \mathcal{H} \big\} \Big)$$

Massart's Lemma:
Suppose $|\mathcal{A}| = N$ is finite. Then $\mathcal{R}(\mathcal{A}) \le \max_{\boldsymbol{a} \in \mathcal{A}} \|\boldsymbol{a}\| \frac{\sqrt{2 \ln N}}{m}$.

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

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

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

Example: Bounded Regression with Lasso

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

Theorem (Lasso Estimator)

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

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

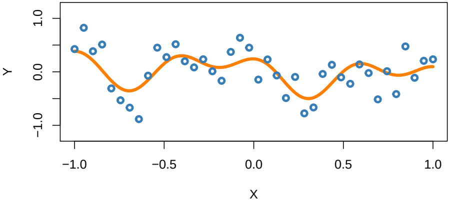
Theorem (Lasso Estimator)

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

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

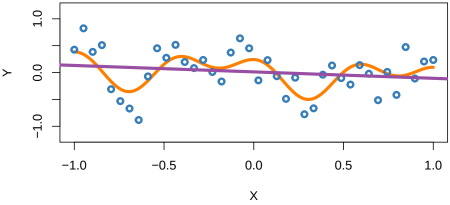
Theorem (Lasso Estimator)

Consider linear regression with $\ell(h, \boldsymbol{X}, Y) = \frac{1}{2}(Y - \langle \boldsymbol{w}, \boldsymbol{X} \rangle)^2$ *for* $\boldsymbol{X} \in [-1, +1]^d$, $Y \in [-1, +1]$.
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 $Y = h(\boldsymbol{X}) + \text{noise}$ for any $h \in \mathcal{H}$!

Example: Bounded Regression with Lasso

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

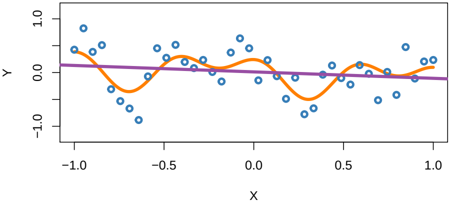
Theorem (Lasso Estimator)

Consider linear regression with $\ell(h, \boldsymbol{X}, Y) = \frac{1}{2}(Y - \langle \boldsymbol{w}, \boldsymbol{X} \rangle)^2$ *for* $\boldsymbol{X} \in [-1, +1]^d$, $Y \in [-1, +1]$.
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 all the tools from this lecture.

NB Do not assume that $Y = h(\boldsymbol{X}) + \text{noise}$ 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

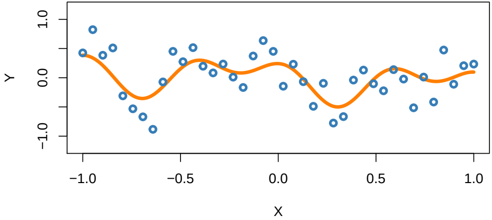
- **all distributions** $\mathcal{D}$, and
- **all hypotheses** $h \in \mathcal{H}$.

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:



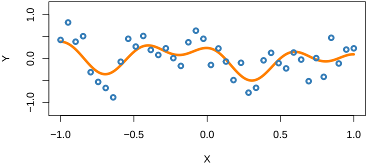
- Distributions still restricted via $(\mathcal{X}, \mathcal{Y})$, e.g. bounded regression
- Uniform convergence not possible in unbounded regression...
- ... unless we **restrict class of possible distributions**

PAC-Learning Guarantees are Very Strong

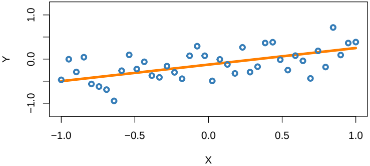
Requires learning with the **same sample complexity** $m(\epsilon, \delta)$ for

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

Hypotheses:



Complex function h



Simple function h

- **Non-uniform learnability**: allow sample complexity $m^{\mathrm{NUL}}(\epsilon, \delta, h)$ 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 $m_{\mathcal{H}}(\epsilon, \delta)$,

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

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

$$L_{\mathcal{D}}(h_S) \le L_{\mathcal{D}}(h) + \epsilon \qquad \text{with probability} \ge 1 - \delta,$$

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

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