Summary of PAC bounds for finite model classes

With probability ≥ 1-δ,

1) For all h $\in$ H s.t. $error_{train}(h) = 0$,

$$error_{true}(h) \leq \varepsilon = \frac{\ln |H| + \ln \frac{1}{\delta}}{m}$$

Haussler's bound

2) For all h $\in$ H

$$|error_{true}(h) - error_{train}(h)| \leq \varepsilon = \sqrt{\frac{\ln |H| + \ln \frac{2}{\delta}}{2m}}$$

Hoeffding's bound

PAC bound and Bias-Variance tradeoff

$$P\left(|\text{error}_{true}(h) - \text{error}_{train}(h)| \geq \epsilon\right) \leq 2|H|e^{-2m\epsilon^2} \leq \delta$$

- Equivalently, with probability $\geq 1 - \delta$

$$\text{error}_{true}(h) \leq \text{error}_{train}(h) + \sqrt{\frac{\ln|H| + \ln\frac{2}{\delta}}{2m}}$$

- Fixed m

Model class		
complex	small	large
simple	large	small

What about the size of the model class?

$$2|H|e^{-2m\epsilon^2} \leq \delta$$

- Sample complexity

$$m \geq \frac{1}{2\epsilon^2}\left(\ln|H| + \ln\frac{2}{\delta}\right)$$

- How large is the model class?

Binary classification

Number of decision trees of depth k

Recursive solution:

$$m \geq \frac{1}{2\epsilon^2}\left(\ln|H| + \ln\frac{2}{\delta}\right)$$

$\ln|H| \sim 2^k$

$\Rightarrow |H| \sim 2^{2^k}$

Given *n* **binary** attributes

H_k = Number of **binary** decision trees of depth k

$H_0 = 2$

H_k = (#choices of root attribute)

*(# possible left subtrees)

*(# possible right subtrees) = $n * H_{k-1} * H_{k-1}$

Write $L_k = \log_2 H_k$ → $L_k = \log_2 n + 2\log_2 H_{k-1}$, $\log_2 H_{k-1} = L_{k-1}$

$L_0 = 1$

$L_k = \log_2 n + 2L_{k-1} = \log_2 n + 2(\log_2 n + 2L_{k-2})$

$= \log_2 n + 2\log_2 n + 2^2\log_2 n + \ldots + 2^{k-1}(\log_2 n + 2L_0)$

So $L_k = (2^k-1)(1+\log_2 n) + 1$

PAC bound for decision trees of depth k

$$m \geq \frac{\ln 2}{2\epsilon^2}\left((2^k - 1)(1 + \log_2 n) + 1 + \log_2 \frac{2}{\delta}\right)$$

- Bad!!!
  - Number of points is exponential in depth k!

- But, for *m* data points, decision tree can't get too big…

Number of leaves never more than number data points

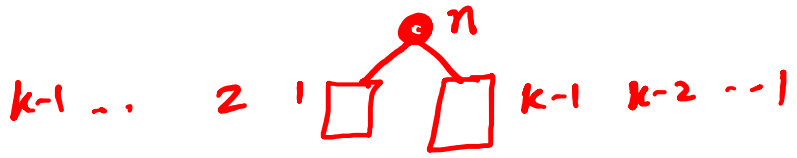
Number of decision trees with k leaves

$$m \geq \frac{1}{2\epsilon^2}\left(\ln|H| + \ln\frac{2}{\delta}\right)$$

H_k = Number of binary decision trees with k leaves

H_1 =2

→ H_k = (#choices of root attribute) *



[(# left subtrees wth 1 leaf)*(# right subtrees wth k-1 leaves)

+ (# left subtrees wth 2 leaves)*(# right subtrees wth k-2 leaves)

+ …

+ (# left subtrees wth k-1 leaves)*(# right subtrees wth 1 leaf)]

$$H_k = n\sum_{i=1}^{k-1} H_i H_{k-i}$$ = n^{k-1} C_{k-1} (C_{k-1} : Catalan Number)

Loose bound (using Sterling's approximation):

$$H_k \leq n^{k-1}2^{2k-1}$$

Number of decision trees

- With k leaves

$$m \geq \frac{1}{2\epsilon^2}\left(\ln|H| + \ln\frac{2}{\delta}\right)$$

$$\log_2 H_k \leq (k-1)\log_2 n + 2k - 1$$ linear in k

number of points m is linear in #leaves

- With depth k

$\log_2 H_k = (2^k\text{-}1)(1+\log_2 n) + 1$ exponential in k

number of points m is exponential in depth

PAC bound for decision trees with k leaves – Bias-Variance revisited

With prob ≥ 1-δ $\text{error}_{true}(h) \leq \text{error}_{train}(h) + \sqrt{\frac{\ln |H| + \ln \frac{2}{\delta}}{2m}}$

With $H_k \leq n^{k-1} 2^{2k-1}$, we get

$$\text{error}_{true}(h) \leq \text{error}_{train}(h) + \sqrt{\frac{(k-1)\ln n + (2k-1)\ln 2 + \ln \frac{2}{\delta}}{2m}}$$

→	k = m	0	large (~ > ½)
→	k < m	>0	small (~ <½)

What did we learn from decision trees?

- Moral of the story:

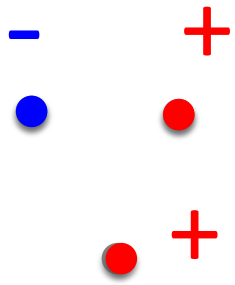
  Complexity of learning not measured in terms of size of model space, but in maximum *number of points* that can be classified using a classifier from this model space

Rademacher Complexity

- Instead of all possible labelings, measure complexity by how accurately a model space can match a random labeling of the data.

For each data point i, draw random label

σ_i s.t. $P(\sigma_i = +1) = ½ = P(\sigma_i = -1)$



Then empirical Rademacher complexity of H is

$$\widehat{R}_m(H) = \mathbb{E}_\sigma \left[\sup_{h \in H} \left(\frac{1}{m} \sum_{i=1}^{m} \sigma_i h(X_i) \right) \right]$$

Finite model class

- Rademacher complexity can be upper bounded in terms of model class size |H|:

$$\widehat{R}_m(H) \leq \sqrt{\frac{2 \ln |H|}{m}}$$

- Often Rademacher bounds are significantly better, e.g. …

Linear models with bounded norm

- Consider $h(X_i) = \langle w, X_i \rangle$ with fixed $\|w\|$, $\|X_i\| \leq R$

$$\widehat{R}_m(H) = \mathbb{E}_\sigma \left[\sup_{h \in H} \left(\frac{1}{m} \sum_{i=1}^{m} \sigma_i h(X_i) \right) \right]$$

$$\vdots$$

$$\leq \frac{\|w\| R}{\sqrt{m}}$$

Other norm bounds – HW5!

Complexity increases with number of parameters d and norm of weights

$$\tilde{R}_m(H) = E_\sigma \left[\sup_{w: \|w\|_2 \le B} \frac{1}{m} \sum_{i=1}^m \sigma_i w^T x_i \right] \qquad h^{(x_i)} = w^T x$$

$$= E_\sigma \left[\sup_{w: \|w\|_2 \le B} w^T \frac{1}{m} \sum_{i=1}^m \sigma_i x_i \right]$$

$$\le \frac{B}{m} E_\sigma \left[\left\| \sum_{i=1}^m \sigma_i x_i \right\| \right]$$

Cauchy-Schwarz

$w^T z \le \|w\|_2 \|z\|_2$

$\|w\|$ — p ; $\|w\|_*$ — q

$\frac{1}{p} + \frac{1}{q} = 1$

$$\le \frac{B}{m} \sqrt{E_\sigma \left[\left\| \sum_{i=1}^m \sigma_i x_i \right\|^2 \right]}$$

Jensen's inequality

$Ez \le \sqrt{Ez^2}$

$$= \frac{B}{m} \sqrt{E_\sigma \left[\sum_{i=1}^m \sigma_i^2 \|x_i\|^2 + \sum_{i \ne j} \sigma_i \sigma_j x_i x_j \right]}$$

($E[\sigma_i \sigma_j] = 0$)

$\|x_i\| \le R$

$$= \frac{B}{m} \sqrt{\sum_i \|x_i\|^2} \le \frac{B}{m} \sqrt{mR^2} = \frac{BR}{\sqrt{m}}$$

$K(x_i, x_i) = \phi(x_i)^T \phi(x_i)$ ⟵ $x_i^T x_i \le R^2$

$\|x_i\| \le R$

Kernel SVMs

- Consider $h(X_i) = \langle w, \phi(X_i) \rangle$ with fixed $\|w\|, K_{x,x} \le R^2$

$$\widehat{R}_m(H) = \mathbb{E}_\sigma \left[\sup_{h \in H} \left(\frac{1}{m} \sum_{i=1}^{m} \sigma_i h(X_i) \right) \right]$$

$$\vdots$$

$$\le \frac{\|w\| R}{\sqrt{m}}$$

Independent of dim d if kernel is bounded!

Complexity of kernel SVMs measured by $\|w\| \propto 1/\text{margin}$ whereas |H|, VC(dim) = ∞ (for Gaussian kernels)

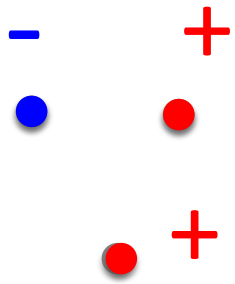
Rademacher Complexity

- Define for any function class f - empirical Rademacher complexity of *F* is

$\sigma_i \sim \pm 1$

$$\widehat{R}_m(\mathcal{F}) = \mathbb{E}_\sigma \left[\sup_{f \in \mathcal{F}} \left(\frac{1}{m} \sum_{i=1}^{m} \sigma_i f(X_i) \right) \right]$$

$\{X_i\}_{i=1}^n$



Max correlation possible with random labels

→ Distribution and data dependent

- True Rademacher Complexity – distribution dependent

$$R_m(\mathcal{F}) = \mathbb{E}_X \left[\widehat{R}_m(\mathcal{F}) \right]$$

Rademacher Concentration Bounds

- With probability ≥ 1-δ, bound deviation of true and empirical averages for any f in *F* using true and empirical Rademacher complexity of *F*

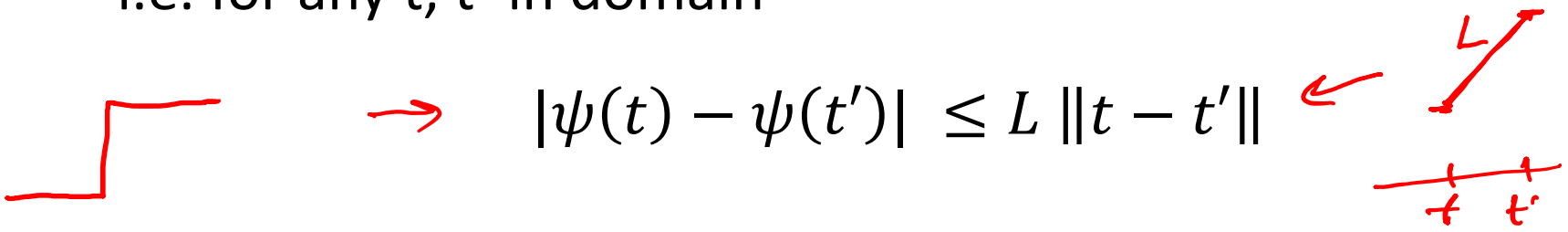
$$\sup_{f \in \mathcal{F}} \mathbb{E}_Z[f(Z)] - \frac{1}{m}\sum_{i=1}^{m} f(Z_i) \leq 2R_m(\mathcal{F}) + \sqrt{\frac{\ln 1/\delta}{m}}$$

$$\sup_{f \in \mathcal{F}} \mathbb{E}_Z[f(Z)] - \frac{1}{m}\sum_{i=1}^{m} f(Z_i) \leq 2\widehat{R}_m(\mathcal{F}) + 3\sqrt{\frac{2\ln 1/\delta}{m}}$$

loss(h(Zi)) = loss ∘ h

Contraction Property of Rademacher Complexity

- Talagrand Lemma/Lipschitz composition: If ψ is L-Lipschitz, i.e. for any t, t' in domain



$$|\psi(t) - \psi(t')| \leq L \|t - t'\|$$

then

$$\widehat{R_m}(\psi \circ H) \leq L\widehat{R_m}(H)$$

- Allows us to map between F (loss of model class) and H (model class) for Lipschitz losses

  f(z) = loss(h(z))

  $f = \text{loss} \circ h$

Other properties – HW5!

Rademacher Error Bounds

- For Lipschitz losses: With probability ≥ 1-δ,

$$\text{error}_{true}(h) \leq \text{error}_{train}(h) + 2L\widehat{R}_m(H) + 3\sqrt{\frac{\log(2/\delta)}{m}}$$

where empirical Rademacher complexity of H

$$\widehat{R}_m(H) = \mathbb{E}_\sigma \left[\sup_{h \in H} \left(\frac{1}{m} \sum_{i=1}^{m} \sigma_i h(X_i) \right) \right]$$

is purely data-dependent.

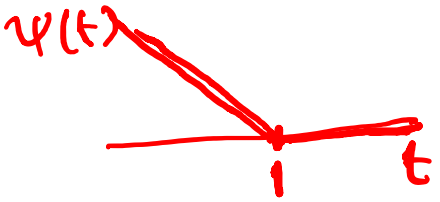
Lipschitz losses

$(1 - \langle w, x_i \rangle Y_i)_+$

- Hinge loss

$$\begin{aligned} \text{loss}(h(Z_i)) &= (1 - h(X_i)Y_i)_+ \\ &= \max(0, 1 - h(X_i)Y_i) \end{aligned}$$

→ $\psi(t) = \max(0, 1 - t)$ is 1-Lipschitz



since $|\psi(t) - \psi(t')| \leq L\|t - t'\|$

$|\max(0,1-t) - \max(0,1-t')| \leq L\|t-t'\|$

$L=1$

$L=1.$

$$\widehat{R}_m(\mathcal{F}) \leq \widehat{R}_m(H)$$

$\hat{R}_m(F) \leq L\, \hat{R}_m(H)$

$f \in F \quad f = \text{loss}_{\text{hinge}} \circ h$

Binary 0/1 loss

- Binary (0/1) loss (in general, not Lipschitz over real domain but Lipschitz over its finite domain)

$$\text{loss}(h(Z_i)) = (1 - h(X_i)Y_i)/2$$

$h(X_i) \in \{-1, 1\}$

$Z_i = (X_i, Y_i)$

L

→ $\psi(t) = (1 - t)/2$ is 1/2-Lipschitz

since $|\psi(t) - \psi(t')| \leq \frac{1}{2}\|t - t'\|$

$$\widehat{R}_m(\mathcal{F}) \leq \frac{1}{2}\widehat{R}_m(H)$$

Rademacher Error Bounds

- **Hinge Error:** With probability ≥ 1-δ,

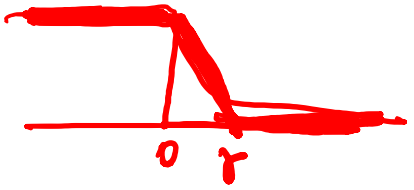
$$\text{error}^{\text{hinge}}_{true}(h) \leq \text{error}^{\text{hinge}}_{train}(h) + 2\widehat{R}_m(H) + 3\sqrt{\frac{\log(2/\delta)}{m}}$$

- **0/1 Error:** With probability ≥ 1-δ,

$$\text{error}_{true}(h) \leq \text{error}_{train}(h) + \widehat{R}_m(H) + 3\sqrt{\frac{\log(2/\delta)}{m}}$$

γ-margin loss

- γ-margin loss (upper bounds 0/1 loss)



$$\psi(t) = \begin{cases} 0 & t \geq \gamma \\ 1 & t \leq 0 \\ 1 - t/\gamma & otherwise \end{cases}$$ is $1/\gamma$-Lipschitz

since $|\psi(t) - \psi(t')| \leq \frac{1}{\gamma}\|t - t'\|$

$\frac{1}{\gamma} = L$

$f = loss \circ h$

$$\widehat{R}_m(\mathcal{F}) \leq \frac{1}{\gamma}\widehat{R}_m(H)$$

Kernel SVM Margin Bound

With probability ≥ 1-δ,

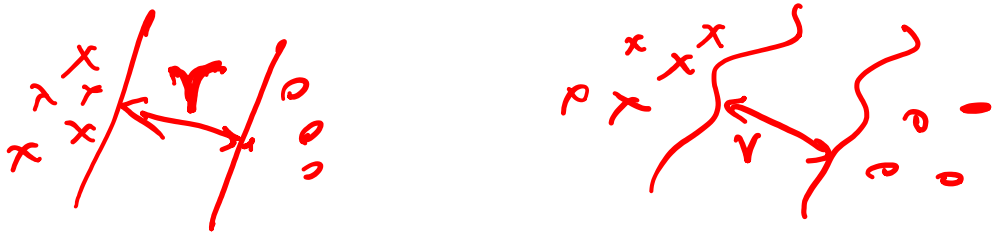
$$\text{error}^{0/1}_{true}(h) \leq \text{error}^{\gamma}_{train}(h) + \frac{\|w\| R}{\gamma\sqrt{m}} + 3\sqrt{\frac{\log(2/\delta)}{m}}$$

$\text{error}^{0/1}_{true} \leq \text{error}^{\gamma}_{true} \leq \text{error}^{\gamma}_{train} + \frac{1}{\gamma}\hat{R}_m(H) + \sqrt{\ }$

$= 0$

small

- If $\text{error}^{\gamma}_{train}$ is small for large γ (f has large margin on data), then true 0/1 error is small.



Summary of PAC bounds

With probability ≥ 1-δ,

1) for all h $\in$ H s.t. $error_{train}(h) = 0$,

$$error_{true}(h) \leq \varepsilon = \frac{\ln |H| + \ln \frac{1}{\delta}}{m}$$

2) for all h $\in$ H,

$$|error_{true}(h) - error_{train}(h)| \leq \varepsilon = \sqrt{\frac{\ln |H| + \ln \frac{2}{\delta}}{2m}}$$

Finite hypothesis space

3) For all h $\in$ H,

$$|error_{true}(h) - error_{train}(h)| \leq \varepsilon = \widehat{R}_m(H) + 3\sqrt{\frac{\log(2/\delta)}{m}}$$

Infinite hypothesis space