

# Machine Learning Theory 2021

## Lecture 12

**Wouter M. Koolen**

Download these slides now from [elo.mastermath.nl](http://elo.mastermath.nl)!

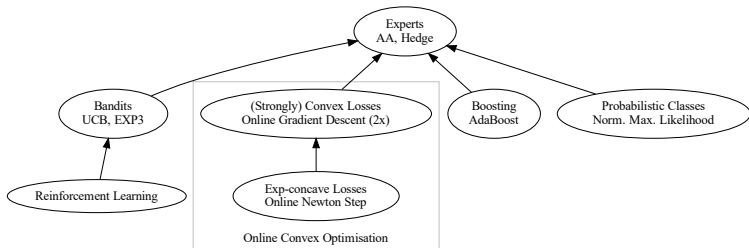
- ▶ Boosting:
  - ▶ Weak and strong PAC learning
  - ▶ Boosting by Online Learning
  - ▶ AdaBoost algorithm
  - ▶ Analysis
  - ▶ VC dimension results



homework roulette  
in the break

# Recap

# Overview of Second Half of Course



Material: course notes and selection of sources on MLT website.

# Outlook

Today: application of **online learning** to good effect in **statistical learning**

Main point: Boosting gets the training error down (to 0).

With: Bound on VC dimension

Get: PAC learning guarantee

# Weak Learning

# Weak Learning

Consider a hypothesis class  $\mathcal{H} \subseteq \{\pm 1\}^{\mathcal{X}}$  for binary classification.  
 $\mathcal{D}$  is  $\mathcal{H}$ -realisable if there is  $h \in \mathcal{H}$  such that  $\mathbb{P}_{X, Y \sim \mathcal{D}}[h(X) = Y] = 1$ .

## Definition (Strong Learnability)

Algorithm  $\mathcal{A}$  **PAC learns**  $\mathcal{H}$  with sample complexity  $m_{\mathcal{H}} : (0, 1)^2 \rightarrow \mathbb{N}$  if for any  $\mathcal{H}$ -realisable  $\mathcal{D}$ , any  $(\epsilon, \delta) \in (0, 1)^2$  and any  $m \geq m_{\mathcal{H}}(\epsilon, \delta)$

$$\mathbb{P}_{S^m \stackrel{\text{iid}}{\sim} \mathcal{D}} \left\{ L_{\mathcal{D}}(h_{\mathcal{A}, S}) \leq \epsilon \right\} \geq 1 - \delta.$$

## Definition ( $\gamma$ -Weak Learnability)

Algorithm  $\mathcal{A}$   **$\gamma$ -weakly learns**  $\mathcal{H}$  with sample complexity  $m_{\mathcal{H}} : (0, 1) \rightarrow \mathbb{N}$  if for any  $\mathcal{H}$ -realisable  $\mathcal{D}$ , any  $\delta \in (0, 1)$  and any  $m \geq m_{\mathcal{H}}(\delta)$

$$\mathbb{P}_{S^m \stackrel{\text{iid}}{\sim} \mathcal{D}} \left\{ L_{\mathcal{D}}(h_{\mathcal{A}, S}) \leq \frac{1}{2} - \gamma \right\} \geq 1 - \delta.$$

# Question

Is a weakly learnable class always PAC learnable?

- ▶ If NO  $\Rightarrow$  perhaps should focus on weak learnability?
- ▶ If YES  $\Rightarrow$  how? efficiency?

# What we know

## Proposition

$\mathcal{H}$  is PAC learnable iff it is weak learnable.

## Proof.

- ▶ If  $\text{VCdim}(\mathcal{H}) < \infty$  then  $\mathcal{H}$  is PAC learnable and hence weak learnable.
- ▶ If  $\text{VCdim}(\mathcal{H}) = \infty$ , then by the Fundamental Theorem the sample complexity at  $(\epsilon, \delta)$  is at least of order

$$\geq \frac{\text{VCdim}(\mathcal{H}) + \ln \frac{1}{\delta}}{\epsilon}$$

which is infinite even for  $\epsilon = \frac{1}{2} - \gamma$ .





# What we know

Idea: perhaps ERM for  $\mathcal{B} \subseteq \mathcal{H}$  is a weak learner for  $\mathcal{H}$ .

Can we **boost** an **efficient** weak learner for  $\mathcal{H}$  to an **efficient** strong learner for  $\mathcal{H}$ ?

## Example

Say

$$\mathcal{H} = \{\text{Three-piece classifiers}\}$$

and

$$\mathcal{B} = \{\text{Two-piece classifiers}\}$$

For every  $\mathcal{H}$ -realisable  $\mathcal{D}$  there is a hypothesis  $f_{\mathcal{B}}^* \in \mathcal{B}$  with  $L_{\mathcal{D}}(f_{\mathcal{B}}^*) \leq \frac{1}{3}$ .

As  $\text{VCdim}(\mathcal{B}) = 2$ , we can agnostic(!) PAC learn  $\mathcal{B}$  to accuracy  $\epsilon = \frac{1}{12}$  with sample size of order  $\epsilon^{-2} \ln \frac{1}{\delta}$ .

With probability  $1 - \delta$ , get

$$L_{\mathcal{D}}(h_S) \leq L_{\mathcal{D}}(f_{\mathcal{B}}^*) + \frac{1}{12} \leq \frac{1}{3} + \frac{1}{12} = \frac{1}{2} - \frac{1}{12}$$

So we can  $\gamma$ -weak learn  $\mathcal{H}$  for  $\gamma = \frac{1}{12}$ .

# Boosting

# Boosting Cartoon

Start with sample  $S = (x_i, y_i)_{i=1}^m$ .

Maintain a hypothesis  $f_t$ . In round  $t$ ,

- ▶ Create distribution  $P_t$  with weight  $w_t^i$  on  $(x_i, y_i)$ .  
**Put more weight on examples misclassified by  $f_t$**
- ▶ Ask Weak Learner for new hypothesis  $h_t$  making  $\leq 1/2 - \gamma$  mistakes on  $P_t$ .  
 **$h_t$  gets right what  $f_t$  gets wrong**
- ▶ Obtain improved  $f_{t+1}$  by incorporating  $h_t$  into  $f_t$ .

# Boosting by Online Learning (BOL)

Fix

- ▶ A  $\gamma$ -weak learner  $\mathcal{W}$  for  $\mathcal{H}$ .
- ▶ A sample  $S = (x_i, y_i)_{i=1}^m$ .
- ▶ A learner  $\mathcal{A}$  for bounded losses on the simplex  $\Delta_m$ .  $\Rightarrow$  Hedge.

## Definition

For  $t = 1, 2, \dots, T$

- ▶ Get  $w_t$  from  $\mathcal{A}$
- ▶ Get  $h_t$  from  $\mathcal{W}$  applied to  $\mathbb{P}^t (X = x, Y = y) = \sum_{i: x=x_i, y=y_i} w_t^i$
- ▶ Set  $\ell_t^i = \mathbf{1} \{h_t(x_i) \neq y_i\}$ .
- ▶ Send  $\ell_t$  to  $\mathcal{A}$ .

Output  $h_S(x) = \text{sign} \left( \sum_{t=1}^T h_t(x) \right)$ .

Duality: Experts  $\Leftrightarrow$  data points    Rounds  $\Leftrightarrow$  hypotheses.

# BOL Analysis

Let  $R_T$  be a regret (bound) for  $\mathcal{A}$ .

## Theorem (Zero training loss)

Consider BOL run for  $T$  rounds such that  $\frac{R_T}{T} \leq \frac{\gamma}{2}$ , with the weak learner error probability set to  $\delta = \frac{\delta}{T}$ . Then

$$L_S(h_S) = 0$$

with probability  $1 - \delta$  (over the possibly randomised weak learner)

# BOL Analysis

Let  $R_T$  be a regret (bound) for  $\mathcal{A}$ .

## Theorem (Zero training loss)

Consider BOL run for  $T$  rounds such that  $\frac{R_T}{T} \leq \frac{\gamma}{2}$ , with the weak learner error probability set to  $\delta = \frac{\delta}{T}$ . Then

$$L_S(h_S) = 0$$

with probability  $1 - \delta$  (over the possibly randomised weak learner)

For the typical case  $R_T = \sqrt{T \ln m}$  we find zero training loss after  $T \geq \frac{4 \ln m}{\gamma^2}$  rounds.

# BOL Analysis I

Suppose  $h_S$  misclassifies sample  $(x_i, y_i)$ . Then

$$h_S(x_i) = \text{sign} \left( \sum_{t=1}^T h_t(x_i) \right) \neq y_i \quad \text{so that} \quad \sum_{t=1}^T \mathbf{1} \{h_t(x_i) = y_i\} \leq \frac{T}{2}$$

This means that

$$\min_j \sum_{t=1}^T \ell_t^j \leq \sum_{t=1}^T \mathbf{1} \{h_t(x_i) = y_i\} \leq \frac{T}{2}$$

and hence by the regret bound for  $\mathcal{A}$ ,

$$\sum_{t=1}^T \mathbf{w}_t^\top \ell_t \leq \frac{T}{2} + R_T \leq T \left( \frac{1}{2} + \frac{\gamma}{2} \right)$$



## BOL Analysis II

Moreover

$$\sum_{t=1}^T \mathbf{w}_t^\top \ell_t = \sum_{t=1}^T \sum_{j=1}^m w_t^j \mathbf{1}\{h_t(x_j) = y_j\}$$

In each round, we have

$$\mathbf{w}_t^\top \ell_t = \sum_{j=1}^m w_t^j \mathbf{1}\{h_t(x_j) = y_j\} = 1 - L_{S, \mathbf{w}_t}(h_t)$$

The weak learner, with probability  $\frac{\delta}{T}$  guarantees in each round

$$L_{S, \mathbf{w}_t}(h_t) \leq \frac{1}{2} - \gamma$$

Overall, with probability  $\geq 1 - \delta$ , we have

$$\sum_{t=1}^T \mathbf{w}_t^\top \ell_t \geq T(1 - (\frac{1}{2} - \gamma)) = T(\frac{1}{2} + \gamma)$$

## BOL Analysis III

But then we obtain the contradiction

$$T\left(\frac{1}{2} + \gamma\right) \leq T\left(\frac{1}{2} + \frac{\gamma}{2}\right)$$

So after all,  $h_S$  must be **perfect** on  $S$ .

# AdaBoost

# AdaBoost

In particular, do not want to assume knowledge of  $\gamma$  up front.

AdaBoost instead **computes** the empirical error:

$$\epsilon_t = \sum_{i=1}^m w_t^i \mathbf{1} \{h_t(x_i) \neq y_i\}$$

Fancy online learning method  $\Rightarrow$  fancy boosting.

# AdaBoost

Fix

- ▶ Aggregating Algorithm
- ▶ A  $\gamma$ -weak learner  $\mathcal{W}$  for  $\mathcal{H}$ .
- ▶ A sample  $S = (x_i, y_i)_{i=1}^m$ .

## Definition

For  $t = 1, 2, \dots, T$

- ▶ Get  $w_t$  from AA.
- ▶ Get  $h_t$  from  $\mathcal{W}$  applied to  $\mathbb{P}^t (X = x, Y = y) = \sum_{i:x=x_i, y=y_i} w_t^i$
- ▶ Compute error  $\epsilon_t = \sum_{i=1}^m w_t^i \mathbf{1} \{h_t(x_i) \neq y_i\}$
- ▶ Sets the round-coefficient to  $\alpha_t = \frac{1}{2} \ln \left( \frac{1}{\epsilon_t} - 1 \right)$
- ▶ Set  $\ell_t^i = \alpha_t y_i h_t(x_i)$ .
- ▶ Send  $\ell_t$  to AA.

Output  $h_S(x) = \text{sign} \left( \sum_{t=1}^T \alpha_t h_t(x) \right)$ .

# AdaBoost Result

## Theorem

Suppose  $\mathcal{W}$   $\gamma$ -weak learns  $\mathcal{H}$ , i.e.  $\epsilon_t \leq \frac{1}{2} - \gamma$ . Then the training error after  $T$  rounds of AdaBoost is at most

$$L_S(h_S) \leq e^{-2\gamma^2 T}.$$

# AdaBoost I

Get  $w_t$  by running AA on losses  $\ell_t^i = \alpha_t y_i h_t(x_i)$ . The mix loss in round  $t$  is

$$\begin{aligned} -\ln \sum_i w_t^i e^{-\ell_t^i} &= -\ln \left( \sum_i w_t^i e^{-\alpha_t y_i h_t(x_i)} \right) \\ &= -\ln \left( e^{-\alpha_t} \sum_{i: h_t(x_i)=y_i} w_t^i + e^{\alpha_t} \sum_{i: h_t(x_i) \neq y_i} w_t^i \right) \\ &= -\ln (e^{-\alpha_t} (1 - \epsilon_t) + e^{\alpha_t} \epsilon_t) \\ &\stackrel{\max_{\alpha_t}}{=} -\frac{1}{2} \ln (4\epsilon_t (1 - \epsilon_t)) \\ &\stackrel{\text{assn.}}{\geq} -\frac{1}{2} \ln (1 - 4\gamma^2) \\ &\geq 2\gamma^2 \end{aligned}$$

## AdaBoost II

Moreover, observe that

$$e^{-yf(x)} \geq \mathbf{1}\{yf(x) \leq 0\} = \mathbf{1}\{\text{sign}(f(x)) \neq y\}$$

Then by the AA telescope

$$\begin{aligned} T2\gamma^2 &\leq -\ln\left(\sum_i \frac{1}{m} e^{-y_i \sum_{t=1}^T \alpha_t h_t(x_i)}\right) \\ &\leq -\ln\left(\sum_i \frac{1}{m} \mathbf{1}\left\{\text{sign}\left(\sum_{t=1}^T \alpha_t h_t(x_i)\right) \neq y_i\right\}\right) \\ &= -\ln\left(\sum_i \frac{1}{m} \mathbf{1}\{h_S(x_i) \neq y_i\}\right) \\ &= -\ln(L_S(h_S)) \end{aligned}$$



# AdaBoost Conclusion

Training error  $< \frac{1}{m}$  means training error = 0.

We have  $e^{-2\gamma^2 T} \leq \frac{1}{m}$  for

$$T \geq \frac{\ln m}{2\gamma^2}.$$

# AdaBoost Conclusion

Training error  $< \frac{1}{m}$  means training error = 0.

We have  $e^{-2\gamma^2 T} \leq \frac{1}{m}$  for

$$T \geq \frac{\ln m}{2\gamma^2}.$$

Is 0 training error useful?

**Risk**

# Is zero training loss good?

Should we worry about over-fitting?

# The VC story

AdaBoost outputs

$$h_S(x) = \text{sign} \left( \sum_{t=1}^T \alpha_t h_t(x) \right)$$

A **half-space** classifier applied to **features**  $(h_t(x))_{t=1}^T$ .

## Definition

Consider the class of all size  $T$  half-spaces over  $\mathcal{B}$

$$L(\mathcal{B}, T) = \left\{ x \mapsto \text{sign} \left( \sum_{t=1}^T w_t h_t(x) \right) \mid w \in \mathbb{R}^T \text{ and } h_t \in \mathcal{B} \right\}$$

# Capacity control

Boosting is safe.

## Theorem

*Let  $d = \text{VCdim}(\mathcal{B})$ . Then*

$$\text{VCdim}(L(\mathcal{B}, T)) \leq 2(d+1)T \log_2(2(d+1)T)$$

## Capacity control, Analysis

Let  $C$  be shattered by  $L(\mathcal{B}, T)$ .

Each labelling of  $C$  consists of a halfspace  $w$  applied to  $h_1, \dots, h_T \in \mathcal{B}$ . By Sauer's Lemma,  $C$  can only be labeled in  $(em/d)^d$  ways by  $\mathcal{B}$  where  $d = \text{VCdim}(\mathcal{B})$ .

Picking  $T$  elements of  $\mathcal{B}$  gives  $(em/d)^{Td}$  possible 0/1 vectors. There are at most  $(em/T)^T$  linear predictors. All in all

$$(em/T)^T (em/d)^{Td} \leq m^{(d+1)T}$$

Shattering  $C$  means  $2^m \leq m^{(d+1)T}$ .

Taking the log and solving (using the tangent bound) implies

$$m \leq 2(d+1)T \log_2(2(d+1)T)$$

# Conclusion



# Conclusion

- ▶ Can boost weak learner to strong learner **efficiently**.
- ▶ Can hence compute ERM on big class from ERM on small class.
- ▶ Useful technique in theory/practice.
- ▶ Relation to margin theory (Chapter 15).