$\text{sign}(w_t h_t(x))$

The output of Adaboost is a composition of half space over the predictions of the T weak hypotheses.

Assume that the weak hypotheses belong to the base hypothesis class, B.

The hypothesis produced by Adaboost belongs to a hypothesis class of linear half spaces defined over predictions by T hypotheses h_t $t \in [1, T]$ where $h_t \in B$.

Base Hypothesis class

stages

We denote the hypothesis class of Adaboost as $L(B, T)$.

Linear Halfspace

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

To learn a hypothesis $\boxed{h \in L(B,T)}$, the AdaBoost learner needs to learn $w \in \mathbb{R}^T$ and $h_t \in B$ $t \in [1, T]$.

During the inference phase, the learned h will classify a given test instance x by constructing $\varphi(x) = (h_1(x), h_2(x) \ldots h_T(x))$ and then applying the half space defined by w on $\varphi(x)$.

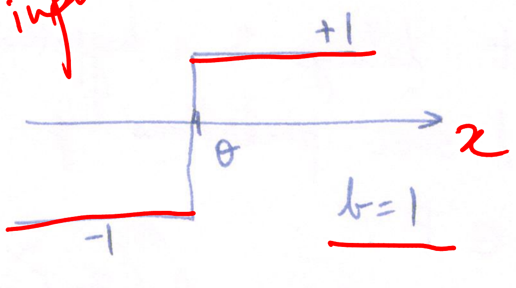
VC dimension of Adaboost Hypothesis Class $L(B, T)$ is bounded by T times the VC dimension of the base hypothesis class B.

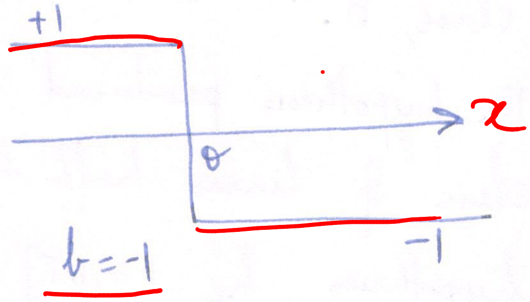
The empirical risk of Adaboost decreases with T.

Consider a base hypothesis class of Decision Stumps.

$$\mathcal{H}_{DS} = \left\{ x \longmapsto \text{sign}(x-\theta)\cdot b \;:\; \begin{matrix}\theta \in \mathbb{R} \\ b \in \{\pm 1\}\end{matrix} \right\}.$$

scalar input



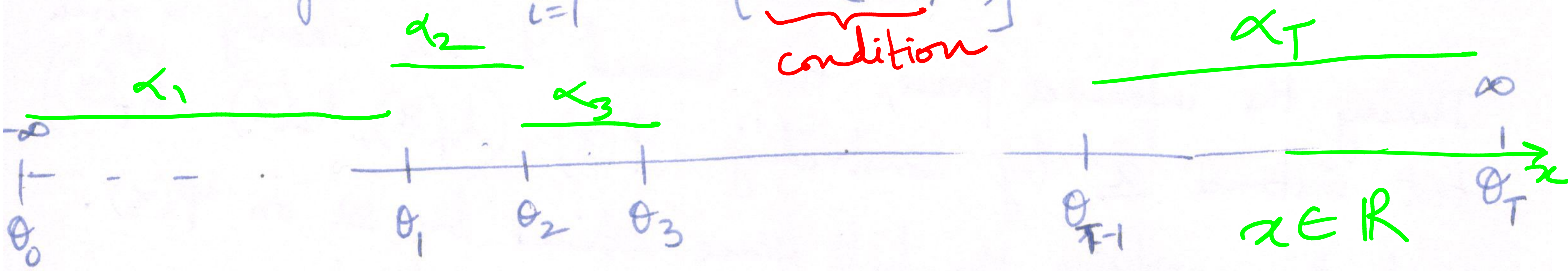


We show that Adaboost can use $B \equiv \mathcal{H}_{DS}$ and construct a rather complex class of piecewise constant functions.

A piece wise constant function over T steps can be defined as

$$g_T(x) = \sum_{i=1}^{T} \alpha_i \, \mathbb{1}_{[\underbrace{x \in (\theta_{i-1}, \theta_i)}_{\text{condition}}]} \qquad \forall i \quad \alpha_i \in \{\pm 1\}$$

indicator



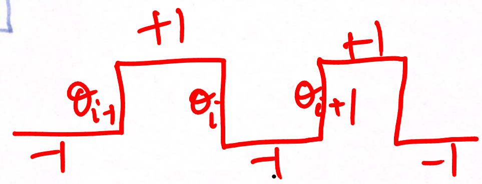
There are T pieces in $g_T(x)$.

The class of half spaces over T decision stumps yields all the piece-wise constant classifiers with at most T pieces.

We demonstrate this by taking a specific example that

$$g_T(x) = \sum_{i=1}^{T} (-1)^i \, \mathbb{1}_{[x \in (\theta_{i-1}, \theta_i)]}$$

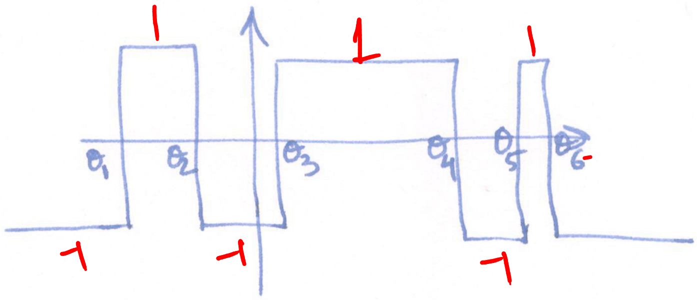
$\alpha_i = (-1)^i$ i.e. $\alpha_i = (-1)^i$



Such a function $g_T(x) = \sum_{i=1}^{T} (-1)^{i} \mathbb{1}_{[x \in (\theta_{i-1}, \theta_i)]}$

A given x would belong to only one interval.

would look like a square wave



The hypothesis produced by the Adaboost classifier

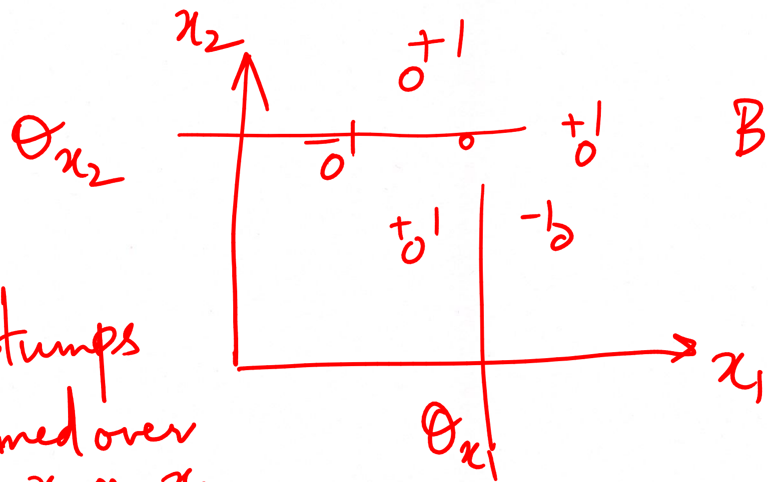
$$h(x) = \text{sign}\left(\sum_{t=1}^{T} w_t \, \text{sign}(x - \theta_{t-1})\right)$$

This will give an output same as $g_T(x)$

Try it

with $\omega_1 = -0.5$
$\omega_t = (-1)^t$

will behave exactly in the same way as $g_T(x)$



$B = H_{DS}$.

$\underline{x} \in \mathbb{R}^2$

Decision stumps can be defined over x_1 or x_2

supervisory input

We have seen construction of supervised machines when the training examples S. were in the form of $\{(x, y)\}$ target.

We used learning rules such as ERM, Hard margin, soft margin, etc.

$x \longrightarrow y$

Now we consider training data where examples are represented using features x but there is no target label.

$\boxed{\text{Unsupervised learning.}}$

We would like to learn the structure of data / distribution of the features $P(X)$. density function $p(x)$. continuous valued feature

$P(X)$ is a data generating distribution.

We can generate observations x by sampling from $P(X)$ $x \sim P(X)$.

Generative model can generate observations x

Now there are two questions.

1) What parameterized model would best describe the data generating distribution?

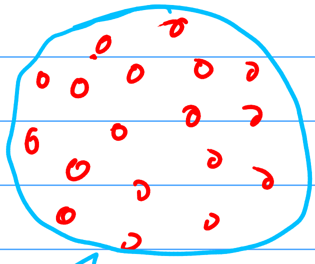
$x \in \mathbb{R}^d$ d-dim space

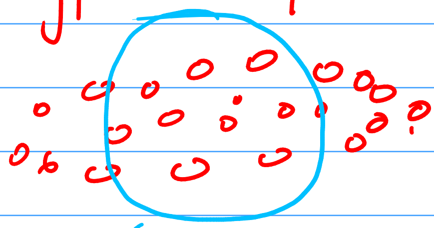
2) How do we estimate the parameters of the data / model chosen for generating distribution?

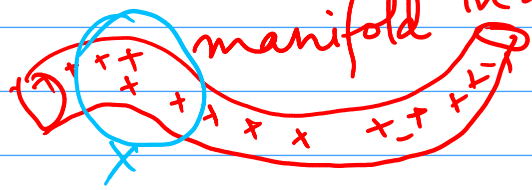
We use maximum likelihood estimation (MLE) for estimating the model parameters.

We formulate a likelihood function which gives the likelihood (probability) of observing / generating all the examples x_i in a training set S

multivariate Gaussian with $C = I$
covariance matrix







Choose

Decide the model → parameterized model (shape of the distribution)

fit the model → parameters are calculated.
MLE (estimation of parameters)