Adaboost

Aaditya Ramdas

Dept. of Statistics and Data Science
Machine Learning Dept.
Carnegie Mellon University

Outline

1. Weak learning implies strong learning (1/2 class)

2. ***Adaboost (1/2 class)***

Recap

A "mixed" strategy is a distribution over actions.

Expected payoff is $\mathbb{E}_{r\sim p, c\sim q}[M(r,c)] = p^T M q = \sum_{r\in[R]} \sum_{c\in C} M(r,c) p_r q_c$

Theorem: $\min_{q\in\Delta_C} \max_{p\in\Delta_R} p^T M q = \max_{p\in\Delta_R} \min_{q\in\Delta_C} p^T M q = v^*$

Value of the game

e_j is the canonical basis vector [0,…,1,…0]

Implications: $\exists p \in \Delta_R\ \forall q \in \Delta_C\ p^T M q \geq v^*$

$\forall q \in \Delta_C \exists i \in [R]\ \ e_i^T M q \geq v^*$

$\exists q \in \Delta_C \forall p \in \Delta_R\ p^T M q \leq v^*$

$\forall p \in \Delta_R \exists j \in [C]\ p^T M e_j \leq v^*$

For "mixed strategies", order does not matter!

Zero-sum games [edit]

The minimax theorem was first proven and published in 1928 by John von Neumann,[3] who is quoted as saying "*As far as I can see, there could be no theory of games … without that theorem … I thought there was nothing worth publishing until the Minimax Theorem was proved*".[4]

Formally, von Neumann's minimax theorem states:

Let $X \subset \mathbb{R}^n$ and $Y \subset \mathbb{R}^m$ be compact convex sets. If $f : X \times Y \to \mathbb{R}$ is a continuous function that is concave-convex, i.e.

$f(\cdot, y) : X \to \mathbb{R}$ is concave for fixed y, and
$f(x, \cdot) : Y \to \mathbb{R}$ is convex for fixed x.

Then we have that

$$\max_{x \in X} \min_{y \in Y} f(x, y) = \min_{y \in Y} \max_{x \in X} f(x, y).$$

Sion's minimax theorem

From Wikipedia, the free encyclopedia

In mathematics, and in particular game theory, **Sion's minimax theorem** is a generalization of John von Neumann's minimax theorem, named after Maurice Sion.

It states:

Let X be a compact convex subset of a linear topological space and Y a convex subset of a linear topological space. If f is a real-valued function on $X \times Y$ with

$f(x, \cdot)$ upper semicontinuous and quasi-concave on $Y, \forall x \in X$, and
$f(\cdot, y)$ lower semicontinuous and quasi-convex on $X, \forall y \in Y$

then,

$$\min_{x \in X} \sup_{y \in Y} f(x, y) = \sup_{y \in Y} \min_{x \in X} f(x, y).$$

Notation

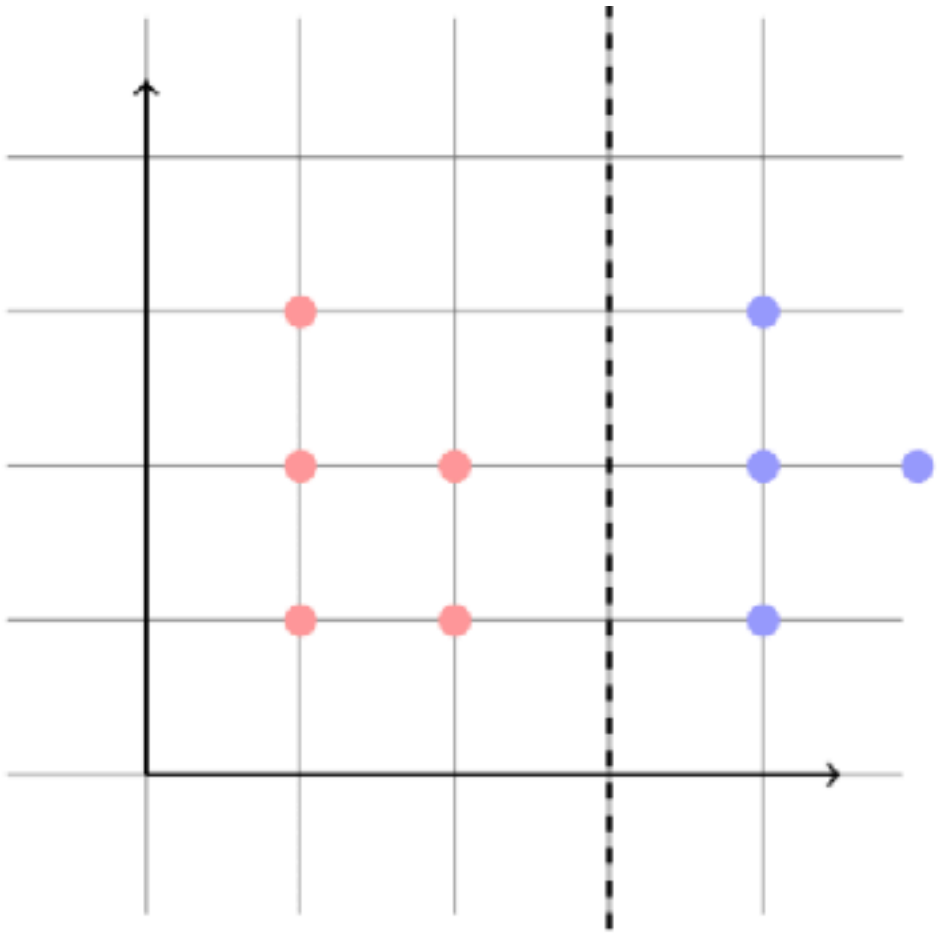
- **Last lectures** (binary) classifier outputs 0/1. In this case the Bayes classifier has form $\mathbb{I}\{\mathbb{E}[Y|X] > 1/2\}$.
- **This lecture** (binary) classifier outputs -1/1. In this case the Bayes classifier has form $\mathbb{I}\{\mathbb{E}[Y|X] > 0\} = \text{sign}(\mathbb{E}[Y|X])$.

Usually classifiers have form $h(x) = \text{sign}(H(x))$. Examples include classification based on logistic regression, k-nearest-neighbors, boosting.

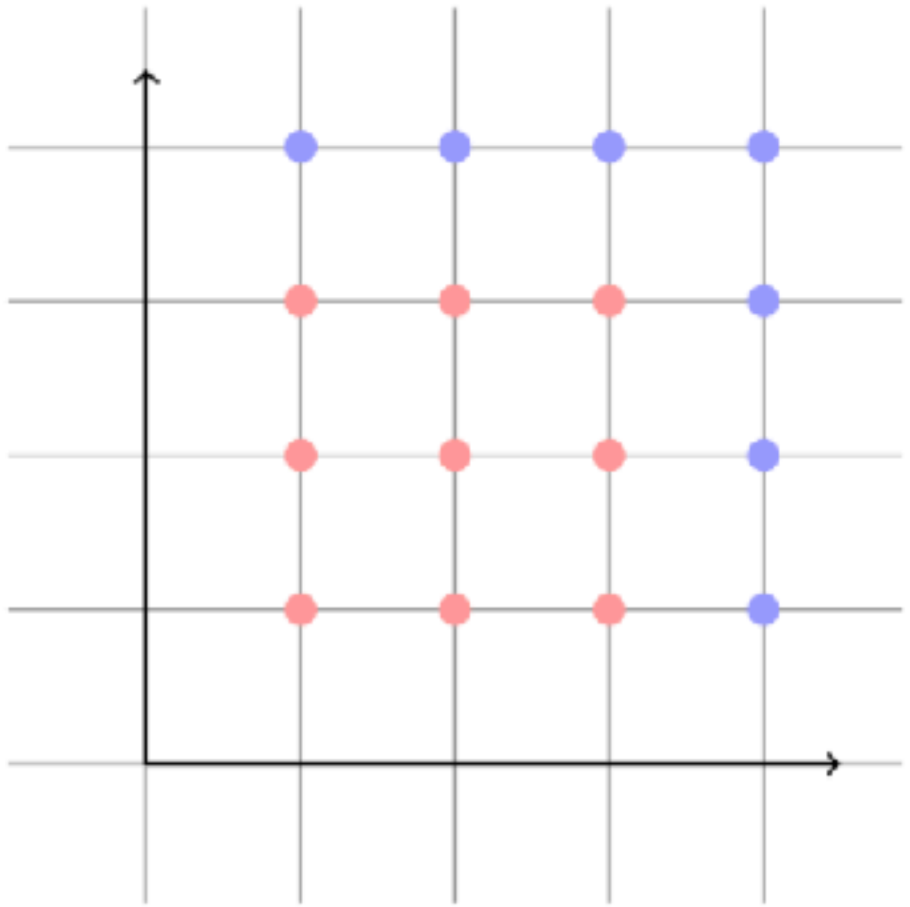
Decision stump: $h(X_i) = 2\mathbf{1}(e_j^T X_i \geq c) - 1$, for some j, c

Decision list: a sequence of if/else decision stumps

Decision tree: a tree of if/else decision stumps



(a) Decision stump performs well.



(b) Decision stump fails. However, decision lists does well

Algorithm 1 Decision list example

if $e_1^\top x_i > 3.5$ **then**
  Predict $+1$
else if $e_2^\top x_i > 3.5$ **then**
  Predict $+1$
else
  Predict -1
end if

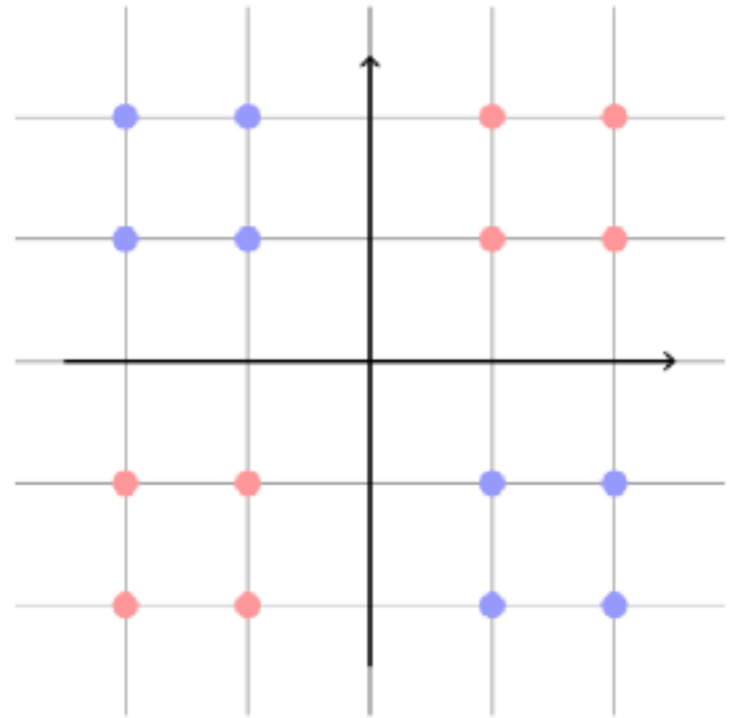


Figure 5.2: Decision list perform poorly. However, decision tree performs well

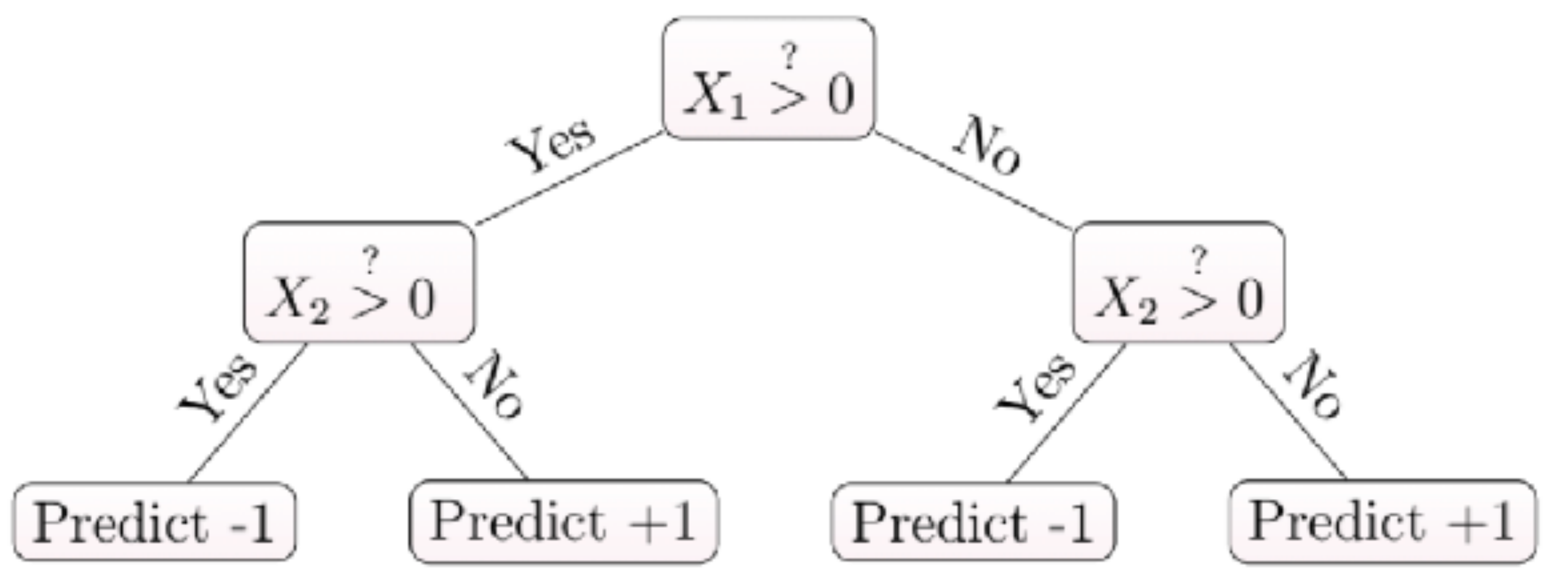


Figure 5.3: Example of a decision tree

Edge or "margin", weighted edge

"Edge" of a classifier Edge of classifier $h \in \mathcal{H}$ is defined as $\frac{1}{n}\sum_{i=1}^{n} Y_i h(X_i)$. Edge provides a way to describe how much better than chance the classifier is:

- Assume that there is a perfect classifier $h^\star : h^\star(X_i) = Y_i, \forall i \in \{1, \dots, n\}$. Then its edge is simply equal to 1.
- Consider, on the contrary, a random guess classifier. It is trivial to show that with high probability its edge concentrates around 0.

Error of a classifier in this case can be viewed as: $\frac{1}{2} \cdot (1 - \text{edge})$. For further analysis we define a "weighted" edge as $\sum_{i=1}^{n} w_i Y_i h(X_i)$ where weights satisfy:

$$\sum_{i=1}^{n} w_i = 1, \qquad w_i > 0,\ \forall i \in \{1, \dots, n\}$$

In the previous definition each data point is equally weighted with weight $1/n$.

Weak learning hypothesis

Weak learning hypothesis: $\exists \gamma > 0$, such that for any set of weights w, there is a classifier $h \in \mathscr{H}$ with weighted edge at least γ.

Let $M(r,i) = h_r(X_i)Y_i$ $\quad \forall w \in \Delta_n \ \exists h \in [H] \ \ e_h^T M w \geq \gamma \qquad (|\mathscr{H}| = H)$

Strong learning: $\exists$a classifier in span$(\mathscr{H})$ with zero training error

$$\exists p \in \Delta_H \ \forall w \in \Delta_n \ p^T M w \geq \gamma$$

$$\exists p \in \Delta_H \qquad \underbrace{\forall i \in [n] \ p^T M e_i \geq \gamma}_{\text{every element of } p^T M \text{ is positive}}$$

$f(X_i) :=$ sign$(p^T M e_i)$ has zero training error, for some $p \in \Delta_H$

The breakthrough

Weak learning implies strong learning!

But how do we find this mixture $p \in \Delta_H$ of classifiers?

Outline

1. Weak learning implies strong learning (1/2 class)

2. ***Adaboost (1/2 class)***

Algorithm 1 AdaBoost algorithm

for $m = 1, \ldots M$ **do**

(1) Compute weighted error:

$$\varepsilon(h) = \sum_{i=1}^{n} w_i \mathbb{I}\{Y_i \neq h(X_i)\}$$

Find a classifier h_m:

$$h_m = \arg\min_{h \in \mathcal{H}} \varepsilon(h)$$

or pick any h with nontrivial edge

(2) Compute:

$$\alpha_m = \frac{1}{2} \log\left(\frac{1-\varepsilon_m}{\varepsilon_m}\right)$$

$$\epsilon_m = \epsilon(h_m)$$

(3) Update weights as:

$$w_i \leftarrow \frac{w_i e^{-\alpha_m Y_i h_m(X_i)}}{Z_m}$$

where Z is a normalization

end for

Output the classifier:

$$\begin{aligned}
Z_m &= \sum_{i=1}^{n} w_m(i) \exp(-\alpha_m y_i h(x_i)) \\
&= \sum_{i: y_i h_m(x) = 1} w_m(i) \exp(-\alpha_m) + \sum_{i: y_i h_m(x) = -1} w_m(i) \exp(\alpha_m) \\
&= (1-\varepsilon_m) \exp(-\alpha_m) + \varepsilon_m \exp(\alpha_m) \\
&= 2\sqrt{\varepsilon_m (1-\varepsilon_m)}
\end{aligned}$$

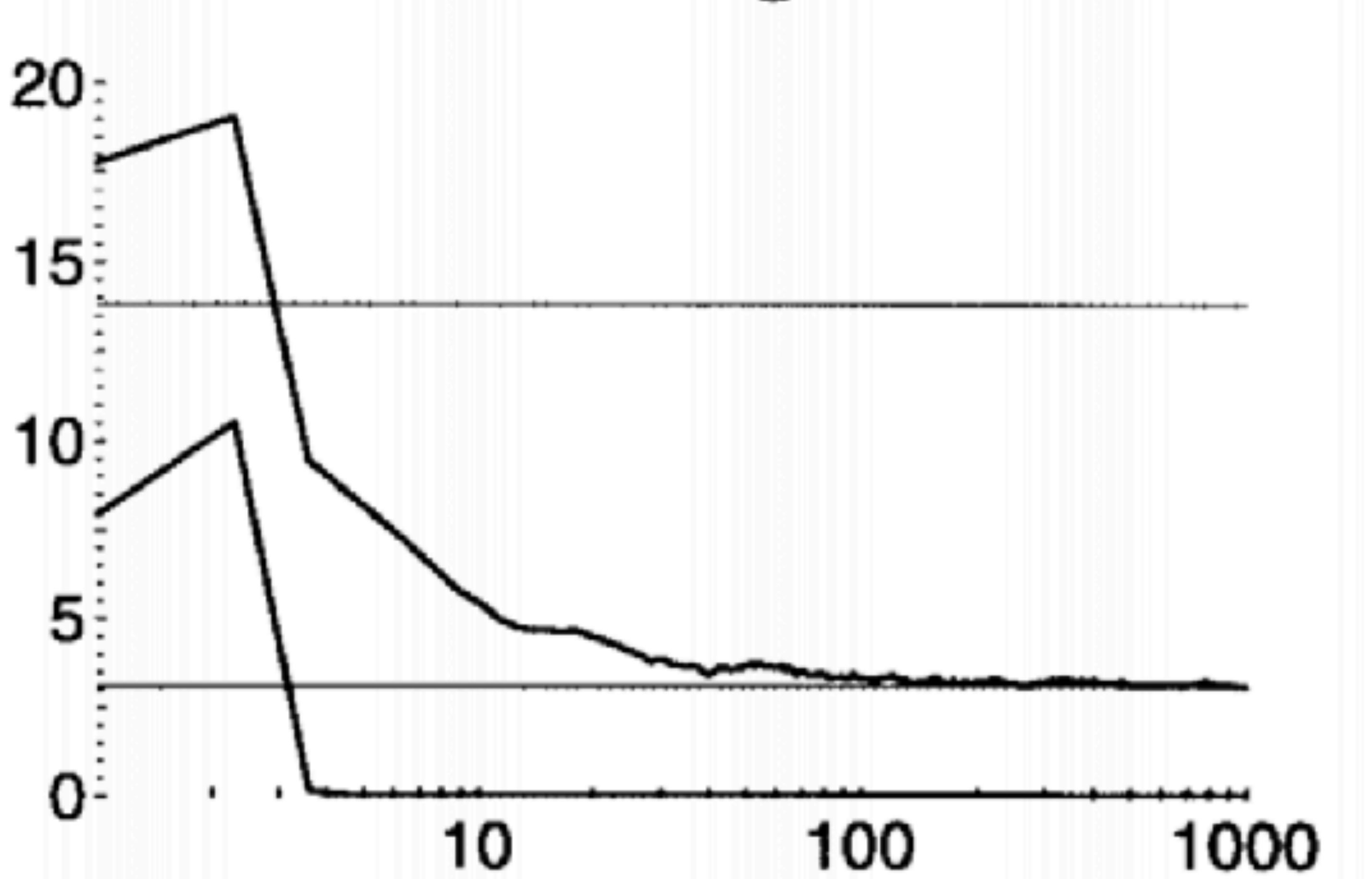


Figure 5.4: This fugure is taken from [2]: each learning curve shows the training and test error curves (lower and upper curves, respectively) of the combined classifier as a function of the number of classifiers combined.

Theorem 2 (Convergence of empirical risk of Adaboost) *The empirical risk of the output of Adaboost algorithm 1 $\hat{R}(f)$ satisfies:*

$$\hat{R}(f) \leq \exp(-2\sum_{m=1}^{M}(\frac{1}{2} - \varepsilon_m)^2) \tag{6.3}$$

$\leq \exp(-2M\gamma^2)$ if weak learning hypothesis is true

If $M > \frac{\log n}{2\gamma^2}$, then $\hat{R}(f) < 1/n$, and hence $\hat{R}(f) = 0$

Outline

1. Adaboost surrogate loss, coordinate descent, etc

2. Margins, history

Algorithm 1 AdaBoost algorithm

for $m = 1, \dots M$ **do**

(1) Compute weighted error:

$$\varepsilon(h) = \sum_{i=1}^{n} w_i \mathbb{I}\{Y_i \neq h(X_i)\}$$

Find a classifier h_m:

$$h_m = \arg\min_{h \in \mathcal{H}} \varepsilon(h)$$

or pick any h with nontrivial edge

(2) Compute:

$$\alpha_m = \frac{1}{2} \log\left(\frac{1 - \varepsilon_m}{\varepsilon_m}\right)$$

$$\epsilon_m = \epsilon(h_m)$$

(3) Update weights as:

$$w_i \leftarrow \frac{w_i e^{-\alpha_m Y_i h_m(X_i)}}{Z_m}$$

where Z is a normalization

end for

Output the classifier:

$$\begin{aligned} Z_m &= \sum_{i=1}^{n} w_m(i) \exp(-\alpha_m y_i h(x_i)) \\ &= \sum_{i: y_i h_m(x) = 1} w_m(i) \exp(-\alpha_m) + \sum_{i: y_i h_m(x) = -1} w_m(i) \exp(\alpha_m) \\ &= (1 - \varepsilon_m) \exp(-\alpha_m) + \varepsilon_m \exp(\alpha_m) \\ &= 2\sqrt{\varepsilon_m(1 - \varepsilon_m)} \end{aligned}$$

Theorem 2 (Convergence of empirical risk of Adaboost) *The empirical risk of the output of Adaboost algorithm 1 $\hat{R}(f)$ satisfies:*

$$\hat{R}(f) \leq \exp(-2\sum_{m=1}^{M}(\frac{1}{2}-\varepsilon_m)^2) \tag{6.3}$$

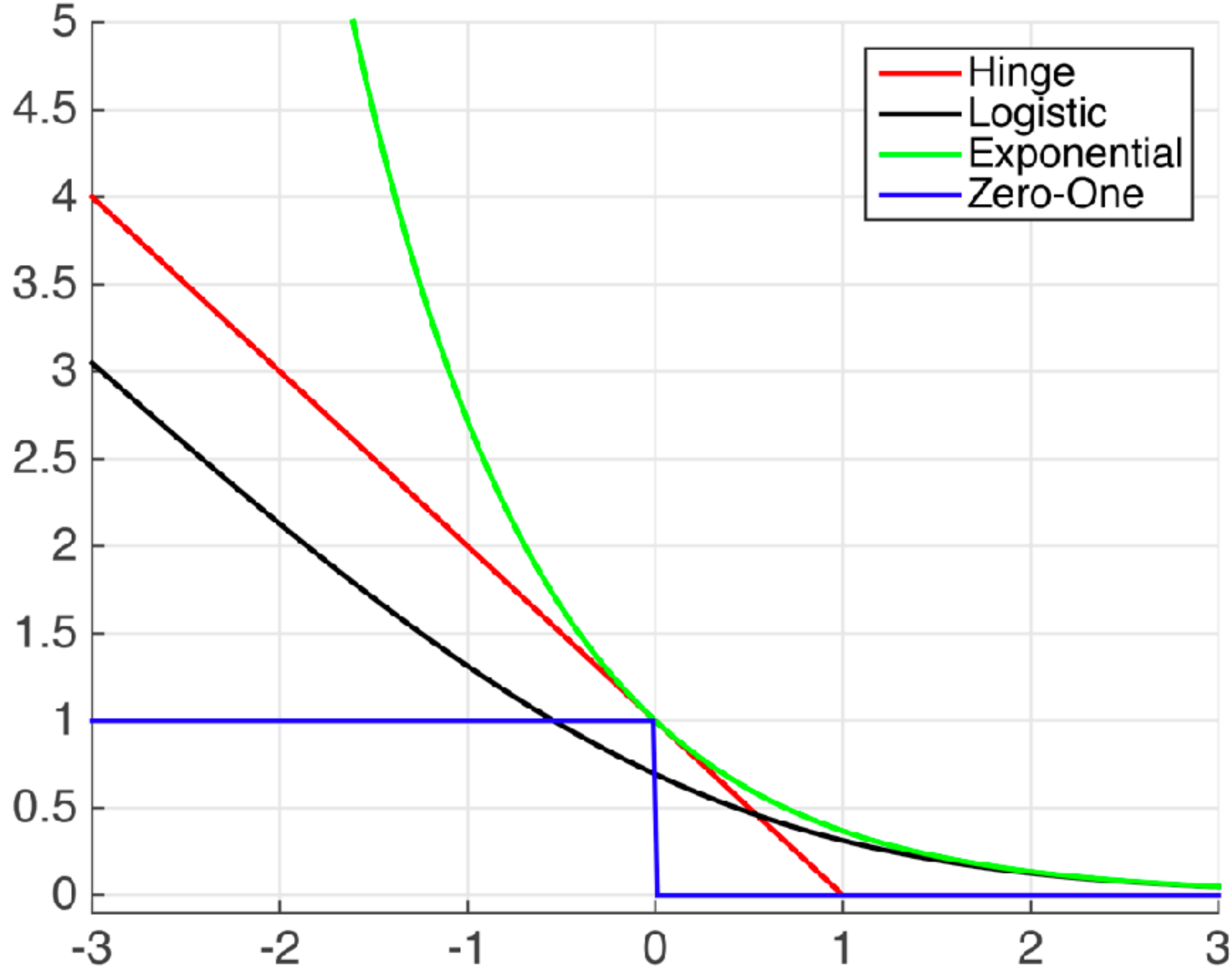
$$\begin{aligned}
\hat{R}(f) =& \frac{1}{n}\sum_{i=1}^{n}\mathbb{I}(y_i \neq f(x_i)) \\
\leq& \frac{1}{n}\sum_{i=1}^{n}\exp(-y_i f(x_i)) \\
=& \frac{1}{n}\sum_{i=1}^{n}[n\prod_{m=1}^{M} Z_m] w_{M+1}(i) \\
=& \prod_{m=1}^{M} Z_m
\end{aligned}$$

Surrogate loss

$$\boxed{w_{M+1}(i) = \frac{w_M(i)e^{-\alpha_m Y_i h_m(X_i)}}{Z_M} = \frac{e^{-Y_i\sum_m \alpha_m h_m(X_i)}}{n\prod_{m=1}^{M} Z_m}}$$

α_m minimizes weighted loss

$$\begin{aligned}
Z_m =& \sum_{i=1}^{n} w_m(i)\exp(-\alpha_m y_i h(x_i)) \\
=& \sum_{i:y_i h_m(x)=1} w_m(i)\exp(-\alpha_m) + \sum_{i:y_i h_m(x)=-1} w_m(i)\exp(\alpha_m) \\
=& (1-\varepsilon_m)\exp(-\alpha_m) + \varepsilon_m\exp(\alpha_m) \\
=& 2\sqrt{\varepsilon_m(1-\varepsilon_m)} \ \boxed{\leq \exp(-2(\frac{1}{2}-\varepsilon_m)^2)}
\end{aligned}$$



Hinge
Logistic
Exponential
Zero-One
5
4.5
4
3.5
3
2.5
2
1.5
1
0.5
0
-3
-2
-1
0
1
2
3

Convex surrogate loss minimization by coordinate descent

$$\hat{R}(\beta) = \frac{1}{n}\sum_{i=1}^{n} \exp\left(-y_i \sum_{j=1}^{H} \beta_j h_j(x_i)\right)$$

Adaboost solves $\min_{\beta \in \mathbb{R}_+^H} \hat{R}(\beta) = \min_{f \in \text{span}(\mathscr{H})} \hat{R}(f)$ by "coordinate descent".

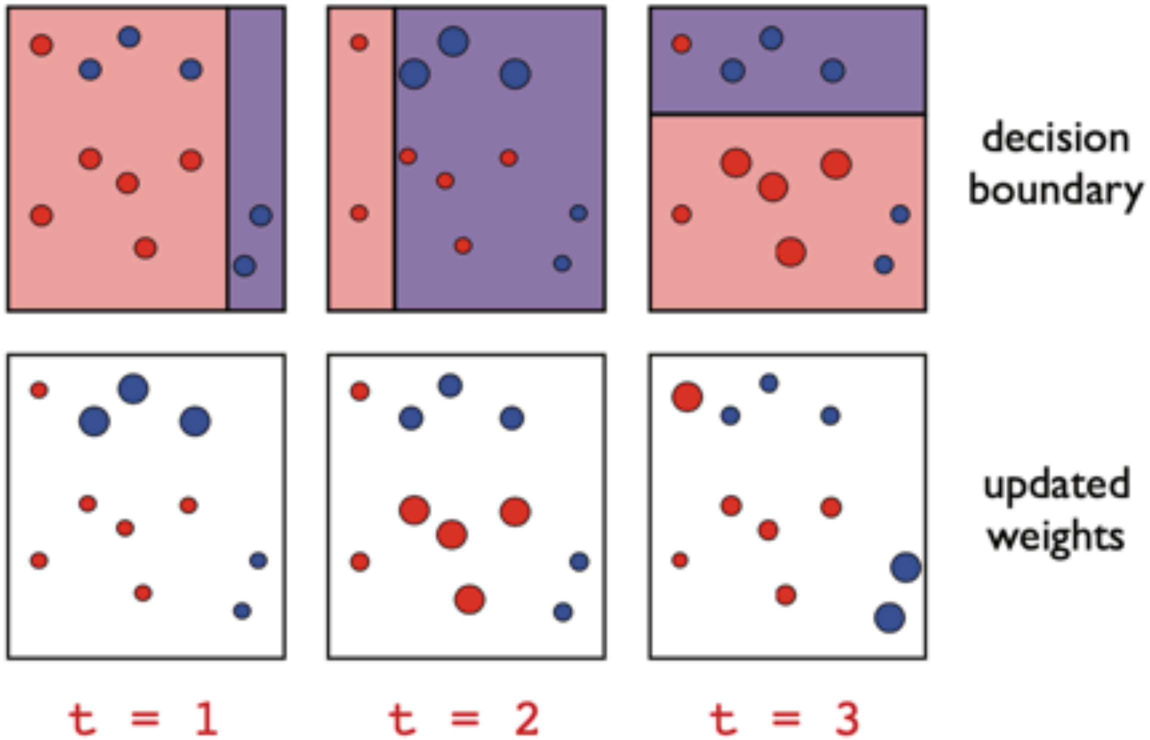
1. Begin at $\beta^{(0)} = [0,0,\ldots,0]$
2. At step t, pick direction $e_t \in \{e_j\}_{j \in J}$ and stepsize $\alpha_t \geq 0$ to minimize $\hat{R}(\beta^{(t-1)} + \alpha_t e_t)$
3. Gradient with respect to coordinate j is $\hat{R}'(\beta^{t-1})_j \propto (2\epsilon_{t,j} - 1)\prod_{s=1}^{t-1} Z_s$, where $\epsilon_{t,j}$ is weighted error of h_j
4. h_t is chosen to minimize the weighted error, optimal stepsize happens to equal $\log(\frac{1-\epsilon_t}{\epsilon_t})$

(History)

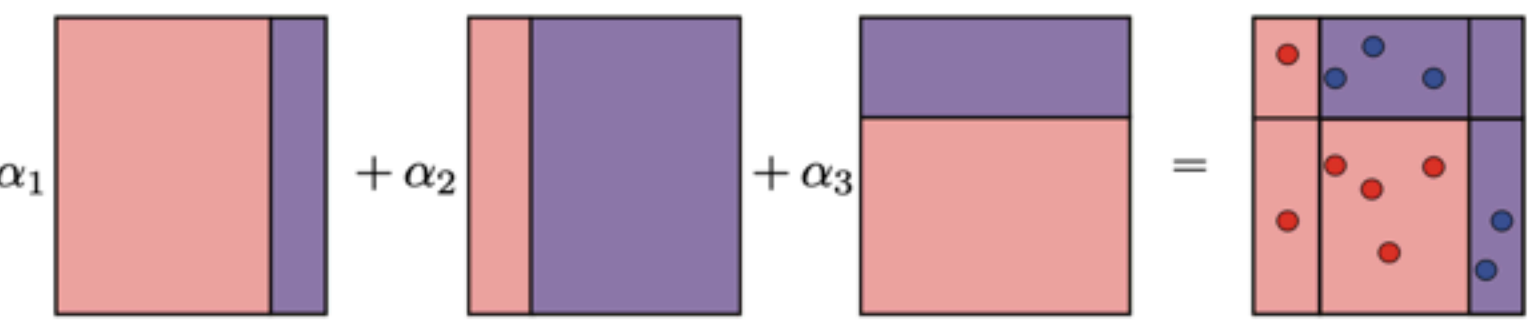
The question of whether a weak learning algorithm could be *boosted* to derive a strong learning algorithm was first posed by Kearns and Valiant [1988, 1994], who also gave a negative proof of this result for a distribution-dependent setting. The first positive proof of this result in a distribution-independent setting was given by Schapire [1990], and later by Freund [1990].

These early boosting algorithms, boosting by filtering [Schapire, 1990] or boosting by majority [Freund, 1990, 1995] were not practical. The AdaBoost algorithm introduced by Freund and Schapire [1997] solved several of these practical issues. Freund and Schapire [1997] further gave a detailed presentation and analysis of the algorithm including the bound on its empirical error, a VC-dimension analysis, and its applications to multi-class classification and regression.

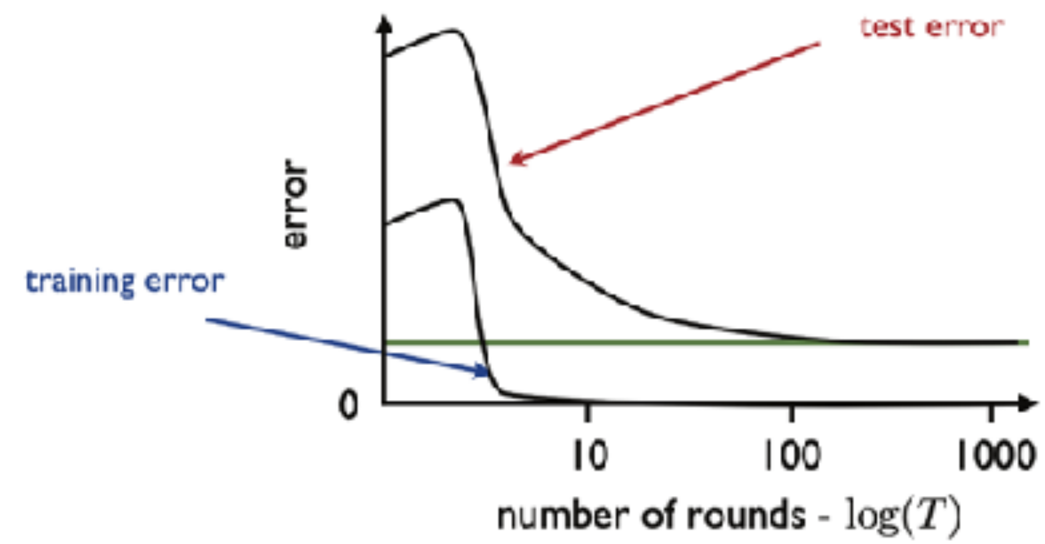
Boosting: VC unsatisfactory



(a)



(b)



Why?

$$\mathcal{F}_T = \left\{ \operatorname{sgn}\left(\sum_{t=1}^{T} \alpha_t h_t \right) : \alpha_t \geq 0, h_t \in \mathcal{H}, t \in [T] \right\}.$$

$$\operatorname{VCdim}(\mathcal{F}_T) \leq 2(d+1)(T+1)\log_2((T+1)e)\,.$$

(Boosting material from Mohri, Rostamizadeh, Talwalkar)

Empirical Rademacher complexity

$$\operatorname{conv}(\mathcal{H}) = \left\{ \sum_{k=1}^{p} \mu_k h_k : p \geq 1, \forall k \in [p], \mu_k \geq 0, h_k \in \mathcal{H}, \sum_{k=1}^{p} \mu_k \leq 1 \right\}. \tag{7.12}$$

The following lemma shows that, remarkably, the empirical Rademacher complexity of $\operatorname{conv}(\mathcal{H})$, which in general is a strictly larger set including $\mathcal{H}$, coincides with that of $\mathcal{H}$.

Lemma 7.4 *Let $\mathcal{H}$ be a set of functions mapping from $\mathcal{X}$ to $\mathbb{R}$. Then, for any sample S, we have*

$$\widehat{\mathfrak{R}}_S\big(\operatorname{conv}(\mathcal{H})\big) = \widehat{\mathfrak{R}}_S(\mathcal{H})\,.$$

$$\begin{aligned}
\widehat{\mathfrak{R}}_S\big(\operatorname{conv}(\mathcal{H})\big) &= \frac{1}{m} \mathop{\mathbb{E}}_{\sigma}\left[\sup_{h_1,\ldots,h_p \in \mathcal{H}, \boldsymbol{\mu} \geq 0, \|\boldsymbol{\mu}\|_1 \leq 1} \sum_{i=1}^{m} \sigma_i \sum_{k=1}^{p} \mu_k h_k(x_i) \right] \\
&= \frac{1}{m} \mathop{\mathbb{E}}_{\sigma}\left[\sup_{h_1,\ldots,h_p \in \mathcal{H}} \sup_{\boldsymbol{\mu} \geq 0, \|\boldsymbol{\mu}\|_1 \leq 1} \sum_{k=1}^{p} \mu_k \sum_{i=1}^{m} \sigma_i h_k(x_i) \right] \\
&= \frac{1}{m} \mathop{\mathbb{E}}_{\sigma}\left[\sup_{h_1,\ldots,h_p \in \mathcal{H}} \max_{k \in [p]} \sum_{i=1}^{m} \sigma_i h_k(x_i) \right] \\
&= \frac{1}{m} \mathop{\mathbb{E}}_{\sigma}\left[\sup_{h \in \mathcal{H}} \sum_{i=1}^{m} \sigma_i h(x_i) \right] = \widehat{\mathfrak{R}}_S(\mathcal{H}),
\end{aligned}$$

L_1 margin

Definition 7.3 (L_1-geometric margin) *The L_1-geometric margin $\rho_f(x)$ of a linear function $f = \sum_{t=1}^{T} \alpha_t h_t$ with $\boldsymbol{\alpha} \neq 0$ at a point $x \in \mathcal{X}$ is defined by*

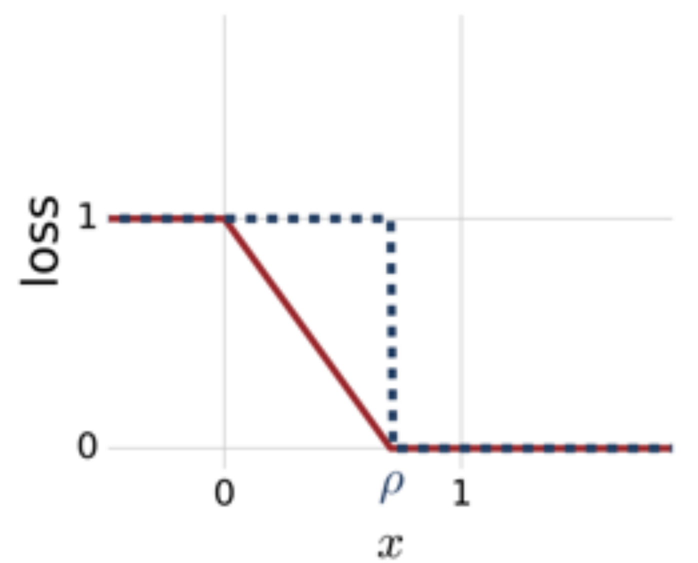
$$\rho_f(x) = \frac{|f(x)|}{\|\boldsymbol{\alpha}\|_1} = \frac{|\sum_{t=1}^{T} \alpha_t h_t(x)|}{\|\boldsymbol{\alpha}\|_1} = \frac{\big|\boldsymbol{\alpha} \cdot \mathbf{h}(x)\big|}{\|\boldsymbol{\alpha}\|_1}. \tag{7.10}$$

The L_1-margin of f over a sample $S = (x_1, \ldots, x_m)$ is its minimum margin at the points in that sample:

$$\rho_f = \min_{i \in [m]} \rho_f(x_i) = \min_{i \in [m]} \frac{\big|\boldsymbol{\alpha} \cdot \mathbf{h}(x_i)\big|}{\|\boldsymbol{\alpha}\|_1}. \tag{7.11}$$

Definition 5.6 (Empirical margin loss) *Given a sample $S = (x_1, \ldots, x_m)$ and a hypothesis h, the empirical margin loss is defined by*

$$\widehat{R}_{S,\rho}(h) = \frac{1}{m} \sum_{i=1}^{m} \Phi_\rho(y_i h(x_i)). \tag{5.37}$$



$$\Phi_\rho(x) = \min\left(1, \max\left(0, 1 - \frac{x}{\rho}\right)\right) = \begin{cases} 1 & \text{if } x \leq 0 \\ 1 - \frac{x}{\rho} & \text{if } 0 \leq x \leq \rho \\ 0 & \text{if } \rho \leq x. \end{cases}$$

Empirical Rademacher complexity, margin loss

Corollary 7.5 (Ensemble Rademacher margin bound) *Let $\mathcal{H}$ denote a set of real-valued functions. Fix $\rho > 0$. Then, for any $\delta > 0$, with probability at least $1 - \delta$, each of the following holds for all $h \in \mathrm{conv}(\mathcal{H})$:*

$$R(h) \leq \widehat{R}_{S,\rho}(h) + \frac{2}{\rho}\mathfrak{R}_m(\mathcal{H}) + \sqrt{\frac{\log \frac{1}{\delta}}{2m}} \tag{7.13}$$

$$R(h) \leq \widehat{R}_{S,\rho}(h) + \frac{2}{\rho}\widehat{\mathfrak{R}}_S(\mathcal{H}) + 3\sqrt{\frac{\log \frac{2}{\delta}}{2m}}. \tag{7.14}$$

Corollary 7.6 (Ensemble VC-Dimension margin bound) *Let $\mathcal{H}$ be a family of functions taking values in $\{+1, -1\}$ with VC-dimension d. Fix $\rho > 0$. Then, for any $\delta > 0$, with probability at least $1 - \delta$, the following holds for all $h \in \mathrm{conv}(\mathcal{H})$:*

$$R(h) \leq \widehat{R}_{S,\rho}(h) + \frac{2}{\rho}\sqrt{\frac{2d \log \frac{em}{d}}{m}} + \sqrt{\frac{\log \frac{1}{\delta}}{2m}}. \tag{7.15}$$