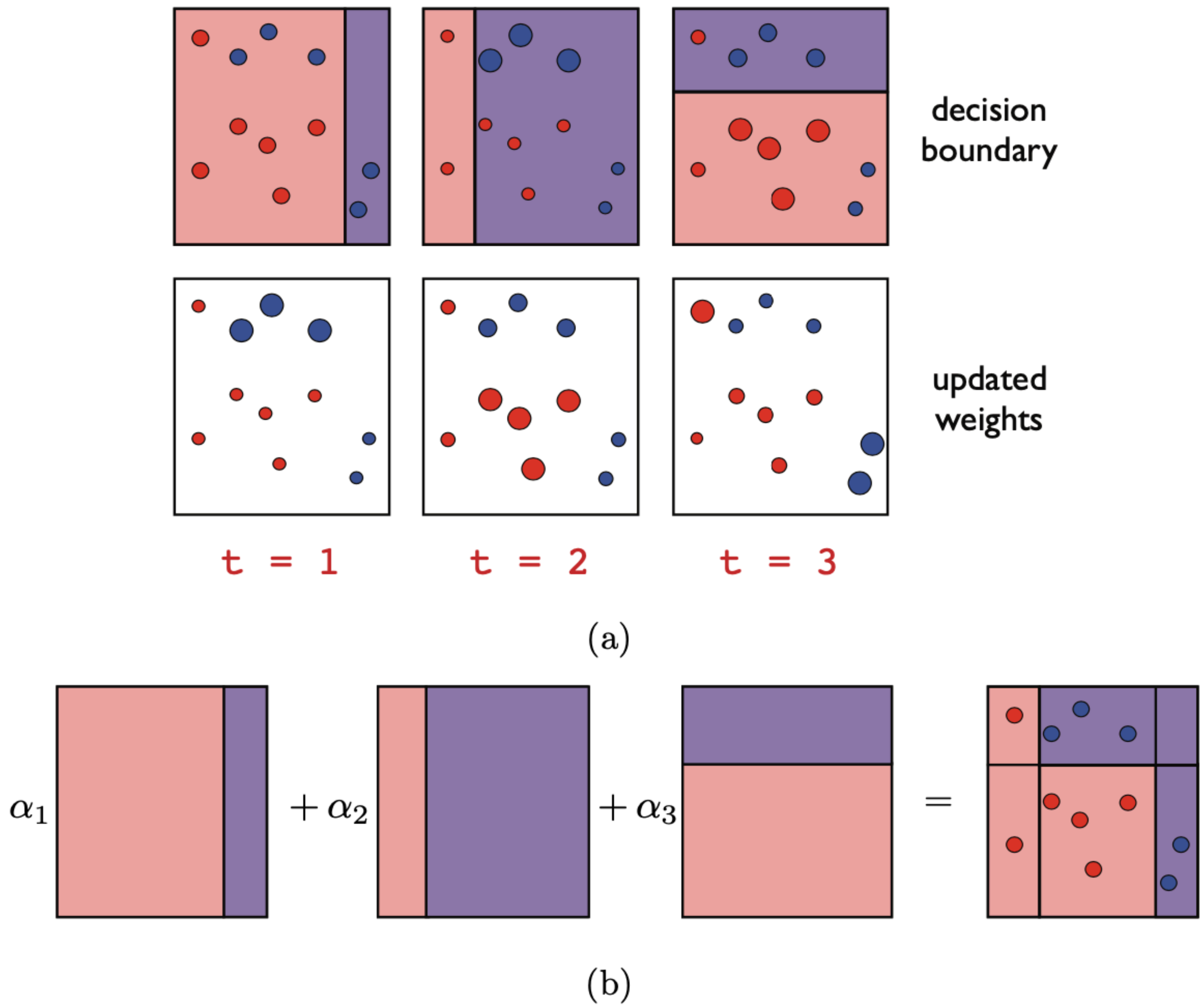
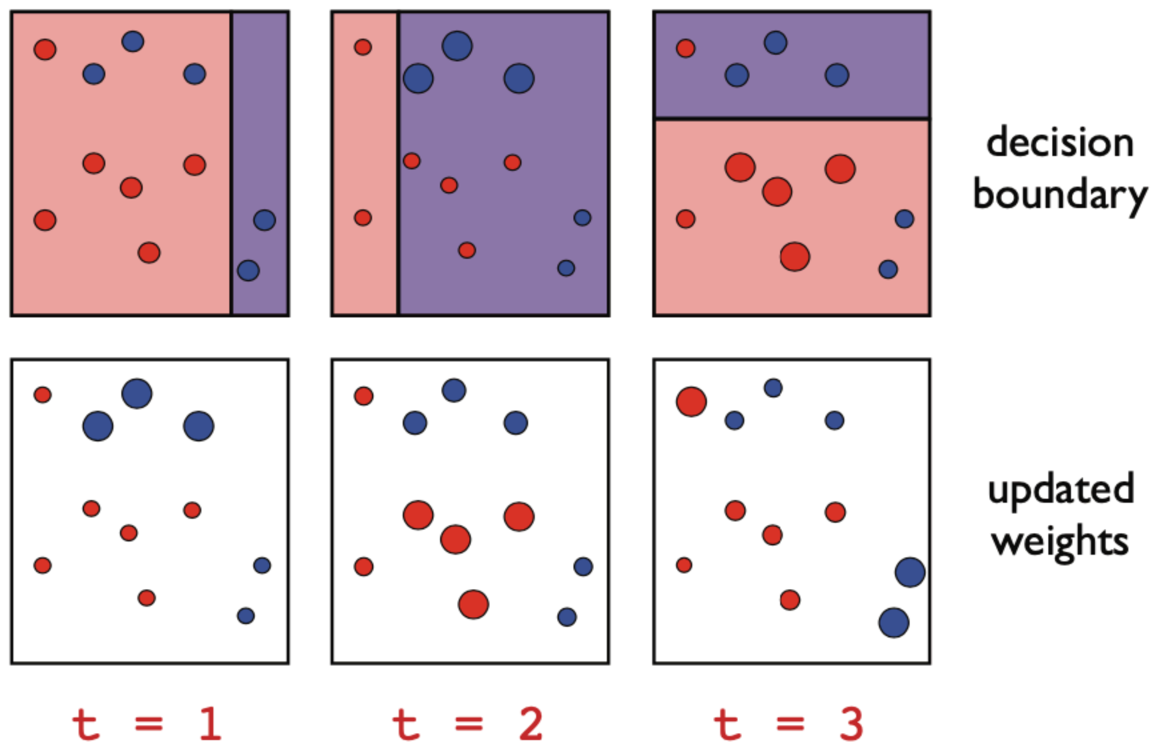
Boosting, Margins and Perceptron



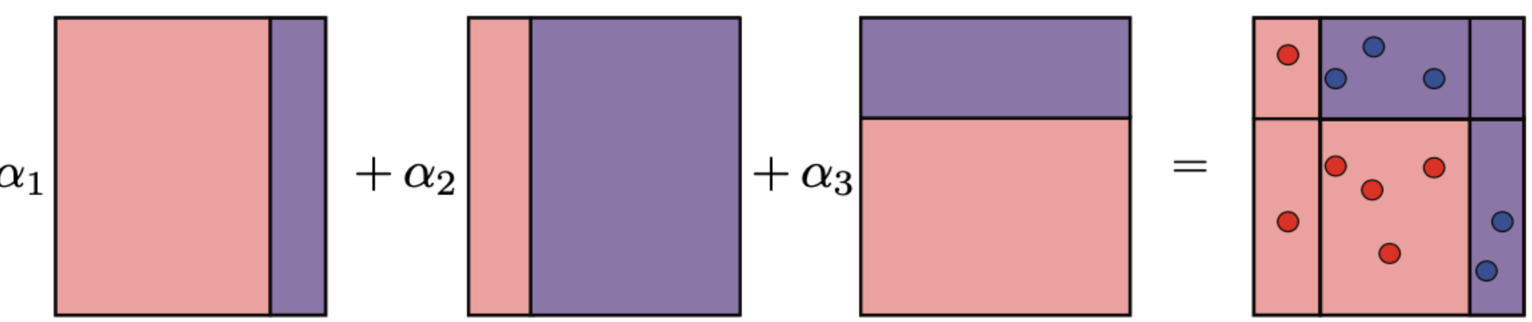
Aaditya Ramdas

Carnegie Mellon University

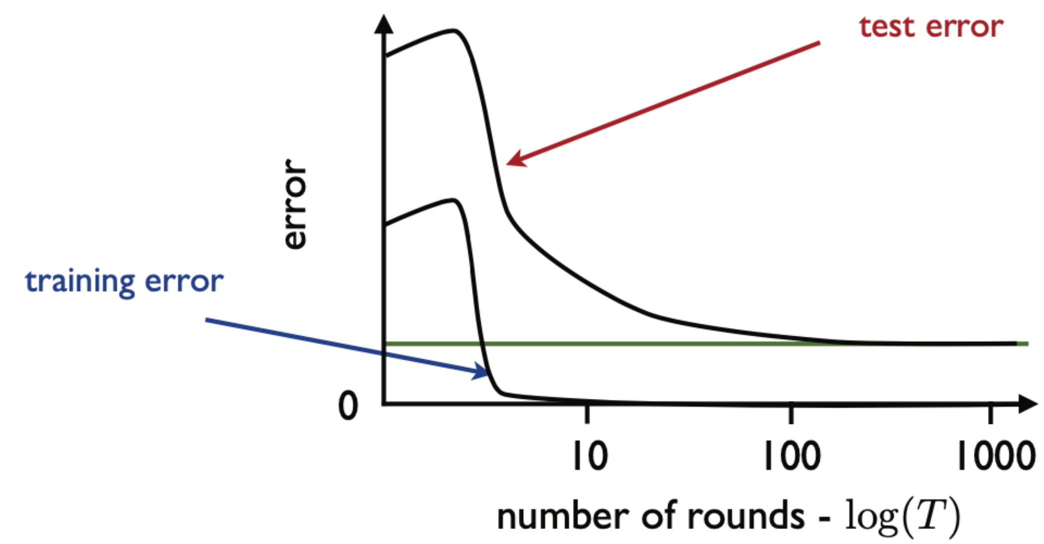
Boosting: VC unsatisfactory



(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*

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

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

Outline

1. *Perceptron (1958-62: Rosenblatt, Block, Novikoff) (1/3 class)*

2. *Margins (1/3 class)*

3. *Boosting (1/3 class)*

Perceptron Learning Algorithm

$k \leftarrow 1; \mathbf{w}_k \leftarrow \mathbf{0}$.
While there exists $i \in \{1, 2, \ldots, n\}$ such that $y^i(\mathbf{w}_k \cdot \mathbf{x}^i) \leqslant 0$:
    Pick an arbitrary $j \in \{1, 2, \ldots, n\}$ such that $y^j(\mathbf{w}_k \cdot \mathbf{x}^j) \leqslant 0$.
    $\mathbf{w}_{k+1} \leftarrow \mathbf{w}_k + y^j x^j$.
    $k \leftarrow k + 1$.
Return $\mathbf{w}_k$.

Assumption 1 (Linear Separability). *There exists some $\mathbf{w}^\star \in \mathbb{R}^d$ such that $||\mathbf{w}^\star|| = 1$ and for some $\gamma > 0$, for all $i \in \{1, 2, \ldots, n\}$,*

$$y^i(\mathbf{w}^\star \cdot \mathbf{x}^i) > \gamma.$$

Assumption 2 (Bounded coordinates). *There exists $R \in \mathbb{R}$ such that for $i \in \{1, 2, \ldots, n\}$,*

$$||\mathbf{x}^i|| \leqslant R.$$

Theorem 3 (Perceptron convergence). *The Perceptron Learning Algorithm makes at most $\frac{R^2}{\gamma^2}$ updates (after which it returns a separating hyperplane).*

Theorem 3 (Perceptron convergence). *The Perceptron Learning Algorithm makes at most $\frac{R^2}{\gamma^2}$ updates (after which it returns a separating hyperplane).*

Note that $\mathbf{w}^1 = \mathbf{0}$, and for $k \geqslant 1$, note that if $\mathbf{x}^j$ is the misclasssified point during iteration k, we have

$$\begin{aligned}
\mathbf{w}^{k+1} \cdot \mathbf{w}^\star &= (\mathbf{w}^k + y^j \mathbf{x}^j) \cdot \mathbf{w}^\star \\
&= \mathbf{w}^k \cdot \mathbf{w}^\star + y^j (\mathbf{x}^j \cdot \mathbf{w}^\star) \\
&> \mathbf{w}^k \cdot \mathbf{w}^\star + \gamma.
\end{aligned}$$

It follows by induction that $\mathbf{w}^{k+1} \cdot \mathbf{w}^\star > k\gamma$. Since $\mathbf{w}^{k+1} \cdot \mathbf{w}^\star \leqslant \|\mathbf{w}^{k+1}\| \|\mathbf{w}^\star\| = \|\mathbf{w}^{k+1}\|$, we get

$$\|\mathbf{w}^{k+1}\| > k\gamma. \tag{1}$$

To obtain an upper bound, we argue that

$$\begin{aligned}
\|\mathbf{w}^{k+1}\|^2 &= \|\mathbf{w}^k + y^j \mathbf{x}^j\|^2 \\
&= \|\mathbf{w}^k\|^2 + \|y^j \mathbf{x}^j\|^2 + 2(\mathbf{w}^k \cdot \mathbf{x}^j) y^j \\
&= \|\mathbf{w}^k\|^2 + \|\mathbf{x}^j\|^2 + 2(\mathbf{w}^k \cdot \mathbf{x}^j) y^j \\
&\leqslant \|\mathbf{w}^k\|^2 + \|\mathbf{x}^j\|^2 \\
&\leqslant \|\mathbf{w}^k\|^2 + R^2,
\end{aligned}$$

from which it follows by induction that

$$\|\mathbf{w}^{k+1}\|^2 \leqslant kR^2. \tag{2}$$

Outline

1. *Perceptron (1/3 class)*

2. *Margins (1/3 class)*

3. *Boosting (1/3 class)*

$$y_i x_i \equiv a_i$$

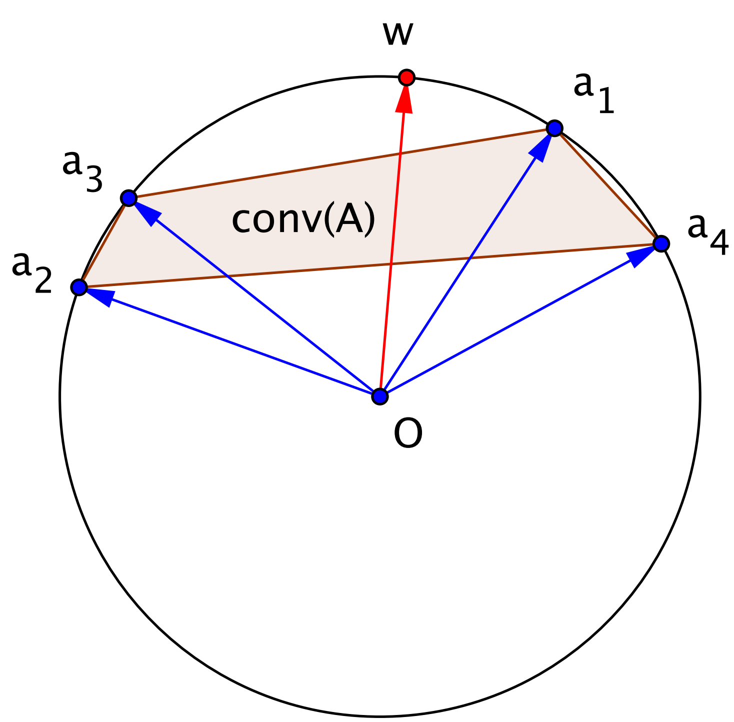
Given $A \in \mathbb{R}^{n \times d}$: n points in d dimensions,
all normalised to unit ℓ_2 norm, i.e. $\|a_i\| = 1$

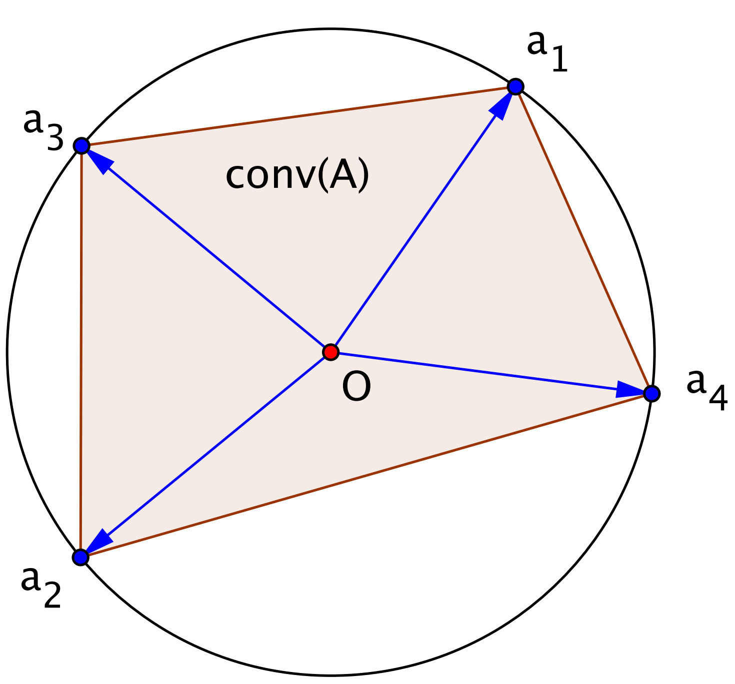
Primal Problem

$$?\exists w : \;\; A^T w > 0 \qquad (P)$$

Dual Problem

$$?\exists p \geq 0 : \;\; Ap = 0 \qquad (D)$$





Margin

$$\rho_A := \sup_{w \in \bigcirc} \inf_{p \in \triangle} w^T A p$$
$$= \sup_{w \in \bigcirc} \inf_{i} w^T a_i .$$

Gordan's theorem

$$\textit{1. Either } \exists w \in \bigcirc \quad s.t. \;\; A^T w > \mathbf{0},$$
$$\textit{or } \exists p \in \triangle \;\; s.t. \;\; Ap = \mathbf{0}.$$

$$\textit{2. Either } \exists w \in \bigcirc \quad s.t. \;\; A^T w > \gamma,$$
$$\textit{or } \exists p \in \triangle \;\; s.t. \;\; \|Ap\| \le \gamma.$$

(for any $\gamma > 0$)

Normalized Perceptron for (P)

$$\begin{aligned} a_i &= \arg\min_{a_i}\{w_{t-1}^T a_i\} \\ w_t &\leftarrow \left(1 - \tfrac{1}{t}\right) w_{t-1} + \left(\tfrac{1}{t}\right) a_i \end{aligned}$$

Similar to Perceptron by Rosenblatt (1958), analyzed by Block (1962), Novikoff (1962).

1. When (P) is feasible, Normalized Perceptron finds satisfying w in $1/\rho_A^{+2}$ steps.
2. Normalized Perceptron is a subgradient method for: $\min_w \max_i\{-w^T a_i\} + \frac{1}{2}\|w\|^2$
3. When (D) is feasible, Normalized Perceptron finds ϵ-certificate in $16/\epsilon^2$ steps!
4. Normalized Perceptron is a margin maximizer! If $\rho_t = \min_i w_t^T a_i$, then $\rho^* - \rho_t \le 8/\rho_A^+\sqrt{t}$
5. Normalized Perceptron finds ρ_A^+ approximately! $\rho_A^+ \;\le\; \|w_t\| \;\le\; \rho_A^+ + 4/\sqrt{t}$

Outline

1. *Perceptron (1/3 class)*
2. *Margins (1/3 class)*
3. *Boosting (1/3 class)*

Back to Boosting

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*

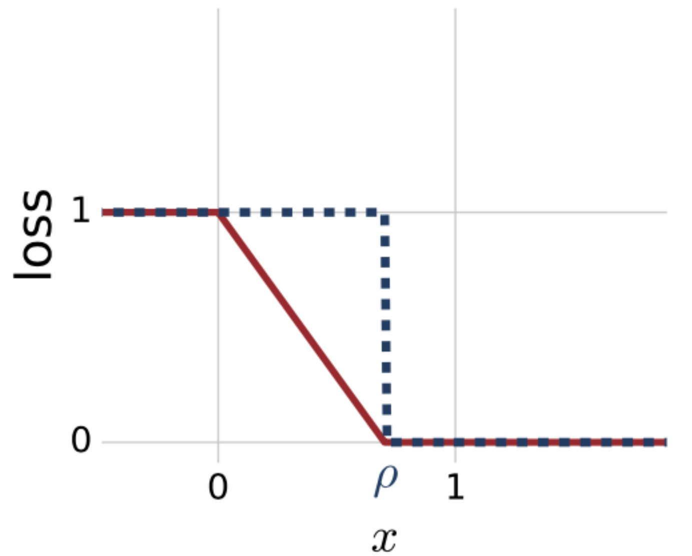
$$\rho_f(x) = \frac{|f(x)|}{\|\boldsymbol{\alpha}\|_1} = \frac{\left|\sum_{t=1}^{T} \alpha_t h_t(x)\right|}{\|\boldsymbol{\alpha}\|_1} = \frac{|\boldsymbol{\alpha} \cdot \mathbf{h}(x)|}{\|\boldsymbol{\alpha}\|_1}. \tag{7.10}$$

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:

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

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*

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



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

Back to Boosting

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 $h \in \operatorname{conv}(\mathcal{H})$:*

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

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

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 $h \in \operatorname{conv}(\mathcal{H})$:*

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

(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.