The Many Faces of Optimal Weak-to-Strong Learning

Mikael Møller Høgsgaard – Aarhus University

Kasper Green Larsen – Aarhus University

Markus Engelund Mathiasen – Aarhus University

NeurIPS 2024

Weak-to-Strong Learning

Binary classification

Weak Learner $\mathcal{W}$:

For any distribution $\mathcal{D}$

Samples: $S \sim \mathcal{D}^{m_0}$

Let $h_S \leftarrow \mathcal{W}(S)$

Then with prob. $1 - \delta_0$:

$$\mathrm{er}_{\mathcal{D}}(h) \leq \frac{1}{2} - \gamma$$

Weak-to-Strong Learning

Binary classification

Weak Learner $\mathcal{W}$:

For any distribution $\mathcal{D}$

Samples: $S \sim \mathcal{D}^{m_0}$

Let $h_S \leftarrow \mathcal{W}(S)$

Then with prob. $1 - \delta_0$:

$$\mathrm{er}_{\mathcal{D}}(h) \leq \frac{1}{2} - \gamma$$

Strong Learner $\mathcal{A}$:

For any distribution $\mathcal{D}$ and parameters $0 < \varepsilon, \delta < 1$

Samples: $S \sim \mathcal{D}^{m(\varepsilon,\delta)}$

$m(\varepsilon, \delta)$ is sample-complexity of $\mathcal{A}$.

Let $\mathrm{h}_S \leftarrow \mathcal{A}(S)$

Then with prob. $1 - \delta$:

$$\mathrm{er}_{\mathcal{D}}(h) \leq \varepsilon$$

Weak-to-Strong Learning

Binary classification

Weak Learner $\mathcal{W}$:

For any distribution $\mathcal{D}$

Samples: $S \sim \mathcal{D}^{m_0}$

Let $h_S \leftarrow \mathcal{W}(S)$

Then with prob. $1 - \delta_0$:

$$\mathrm{er}_{\mathcal{D}}(h) \leq \frac{1}{2} - \gamma$$

→

Strong Learner $\mathcal{A}$:

For any distribution $\mathcal{D}$ and parameters $0 < \varepsilon, \delta < 1$

Samples: $S \sim \mathcal{D}^{m(\varepsilon,\delta)}$

$m(\varepsilon, \delta)$ is sample-complexity of $\mathcal{A}$.

Let $\mathrm{h}_S \leftarrow \mathcal{A}(S)$

Then with prob. $1 - \delta$:

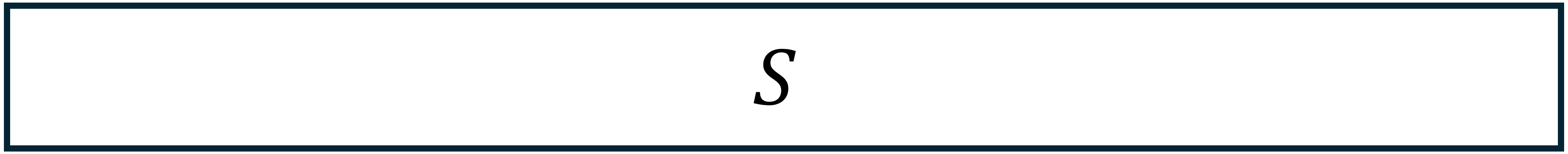
$$\mathrm{er}_{\mathcal{D}}(h) \leq \varepsilon$$

Majority of 5

Algorithm 1: MAJORITY-OF-5$(S, \mathcal{W})$

Input: Training set $S = (x_1, y_1), \ldots, (x_m, y_m)$. Weak learner $\mathcal{W}$.
Result: Hypothesis $g : \mathcal{X} \to \{-1, 1\}$.

Partition S into 5 disjoint pieces $S_1, \ldots, S_5$ of size $m/5$.
for $t = 1, \ldots, 5$ **do**
  Run AdaBoost on S_t with $\mathcal{W}$ to obtain $f_t : \mathcal{X} \to \{-1, 1\}$.
$g \leftarrow \text{sign}(\sum_t f_t)$.
return g

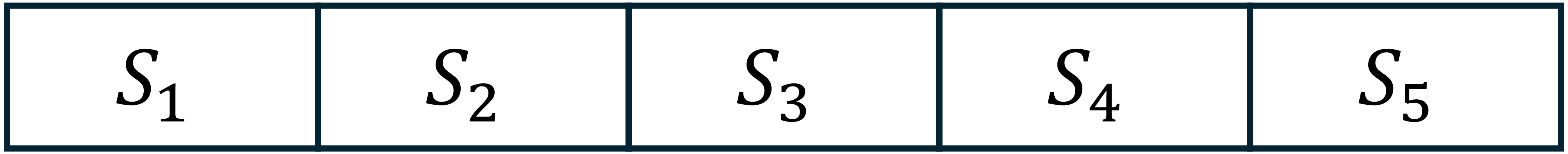


Majority of 5

Algorithm 1: MAJORITY-OF-5($S, \mathcal{W}$)

Input: Training set $S = (x_1, y_1), \ldots, (x_m, y_m)$. Weak learner $\mathcal{W}$.
Result: Hypothesis $g : \mathcal{X} \to \{-1, 1\}$.

```
Partition S into 5 disjoint pieces S_1, ..., S_5 of size m/5.
for t = 1, ..., 5 do
  | Run AdaBoost on S_t with W to obtain f_t : X → {-1, 1}.
g ← sign(∑_t f_t).
return g
```

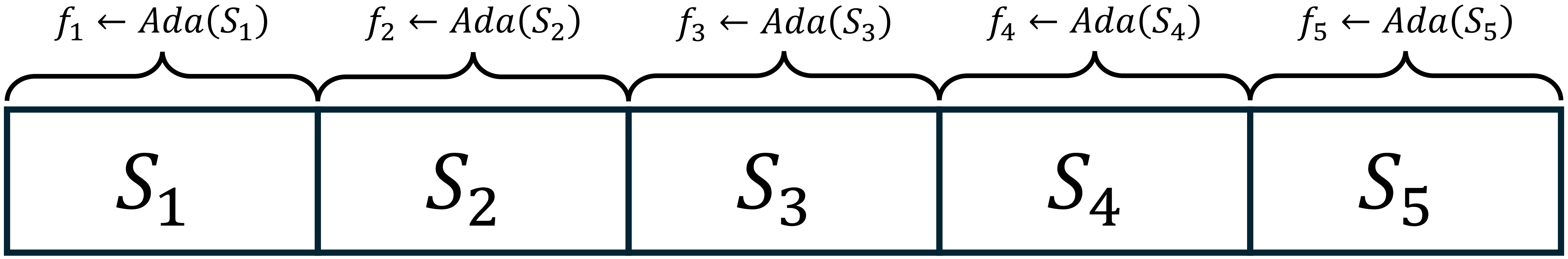


Majority of 5

Algorithm 1: MAJORITY-OF-5$(S, \mathcal{W})$

Input: Training set $S = (x_1, y_1), \ldots, (x_m, y_m)$. Weak learner $\mathcal{W}$.
Result: Hypothesis $g : \mathcal{X} \to \{-1, 1\}$.

Partition S into 5 disjoint pieces $S_1, \ldots, S_5$ of size $m/5$.
for $t = 1, \ldots, 5$ **do**
    Run AdaBoost on S_t with $\mathcal{W}$ to obtain $f_t : \mathcal{X} \to \{-1, 1\}$.
$g \leftarrow \text{sign}(\sum_t f_t)$.
return g

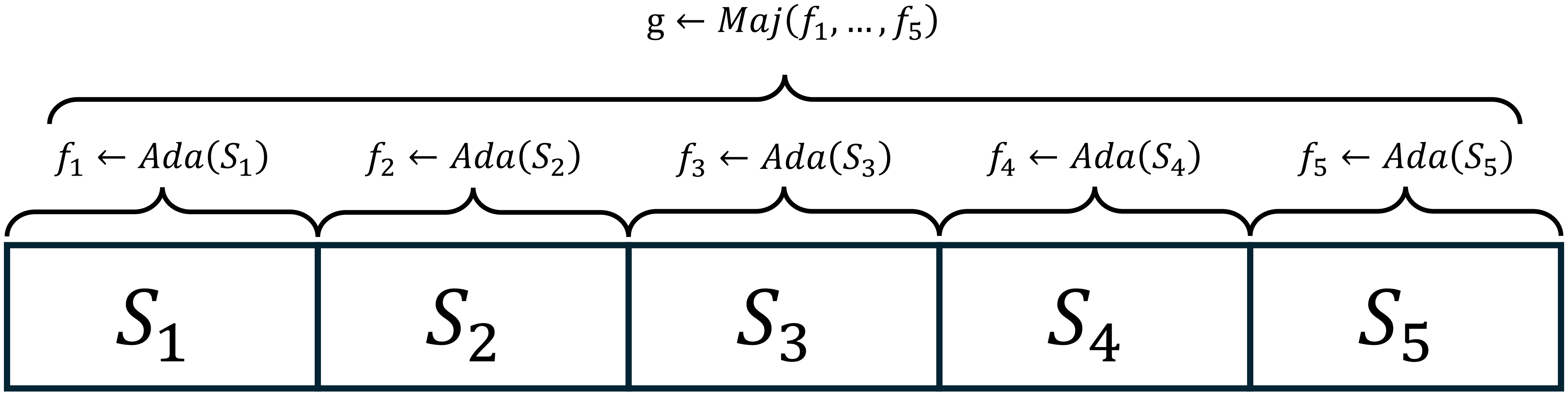


Majority of 5

Algorithm 1: MAJORITY-OF-5$(S, \mathcal{W})$

Input: Training set $S = (x_1, y_1), \ldots, (x_m, y_m)$. Weak learner $\mathcal{W}$.
Result: Hypothesis $g : \mathcal{X} \to \{-1, 1\}$.

Partition S into 5 disjoint pieces $S_1, \ldots, S_5$ of size $m/5$.
for $t = 1, \ldots, 5$ **do**
  Run AdaBoost on S_t with $\mathcal{W}$ to obtain $f_t : \mathcal{X} \to \{-1, 1\}$.
$g \leftarrow \text{sign}(\sum_t f_t)$.
return g



Majority of 5 - Guarantee

Theorem 1

$$\mathbb{E}_{S\sim\mathcal{D}^m}[\mathrm{er}_{\mathcal{D}}(g)] = O\left(\frac{d}{\gamma^2 m}\right)$$

Algorithm 1: MAJORITY-OF-5$(S, \mathcal{W})$

Input: Training set $S = (x_1, y_1), \ldots, (x_m, y_m)$. Weak learner $\mathcal{W}$.
Result: Hypothesis $g : \mathcal{X} \to \{-1, 1\}$.

Partition S into 5 disjoint pieces $S_1, \ldots, S_5$ of size $m/5$.
for $t = 1, \ldots, 5$ **do**
  Run AdaBoost on S_t with $\mathcal{W}$ to obtain $f_t : \mathcal{X} \to \{-1, 1\}$.
$g \leftarrow \mathrm{sign}(\sum_t f_t)$.
return g

Sample-Optimal Weak-to-Strong Learners

Algorithm	Error	Invocations of Weak Learner
AdaBoost	$O\left(\frac{d\ln^2(m)}{\gamma^2 m}\right)$	$O\left(\frac{\ln(m)}{\gamma^2}\right)$
LarsenRitzert[1]	$O\left(\frac{d}{\gamma^2 m}\right)$	$O\left(\frac{m^{0.8}}{\gamma^2}\right)$
Bagged AdaBoost[2]	$O\left(\frac{d}{\gamma^2 m}\right)$	$O\left(\frac{\ln^2(m)}{\gamma^2}\right)$
Majority of 5	$O\left(\frac{d}{\gamma^2 m}\right)$	$O\left(\frac{\ln(m)}{\gamma^2}\right)$

1) Green Larsen, K., & Ritzert, M. (2022). Optimal weak to strong learning. *Advances in Neural Information Processing Systems*, *35*, 32830-32841.

2) Larsen, K. G. (2023, July). Bagging is an optimal PAC learner. In *The Thirty Sixth Annual Conference on Learning Theory* (pp. 450-468). PMLR.

Open Questions

- Is Majority of 5 optimal in high probability setting?
- Is Majority of 3 sufficient?