Margin-Based Generalization
Lower Bounds for Boosted Classifiers

Allan Grønlund, Lior Kamma, Kasper Green Larsen,
Alexander Mathiasen and Jelani Nelson
Boosting Algorithms

- Construct strong classifiers out of weak ones.

Accurate

Slightly better than guessing
Boosting Algorithms

- Construct strong classifiers out of weak ones.

By combining them into a powerful “ensemble”
Boosting Algorithms

- Construct strong classifiers out of weak ones.

- Intuition: Train many weak classifiers, each “focusing” on a different part of the input space.

  Achieved by re-weighing the input sample.
Example: Axis Aligned Lines
Example: Axis Aligned Lines

Margin-Based Generalization Lower Bounds for Boosted Classifiers
Surprising phenomenon: Even though the strong classifier gets more complicated, it does not overfit.
Boosting Algorithms and Margins

Surprising phenomenon: Even though the strong classifier gets more complicated, it does not overfit.

Observed in experiments by Schapire et al.
Surprising phenomenon: Even though the strong classifier gets more complicated, it does not overfit.

That is, more weak classifiers are involved.
Boosting Algorithms and Margins

- Surprising phenomenon: Even though the strong classifier gets more complicated, it does not overfit.

- Prominent explanation: Margin Theory

Loosely speaking, the “confidence” of the classifier on a point.
Margin Theory

- Formally, let $\mathcal{H} \subseteq \mathcal{X} \rightarrow \{-1,1\}$ be the space of weak classifiers, and $S = \{(x_j, y_j)\}_{j=1}^m$ is the sample used to train a strong classifier $f = \sum_{h \in \mathcal{H}} \alpha_h h$.

- The margin of $f$ on the $j$th sample point is defined as $\theta_j := y_j f(x_j)$.
Margin Theory

Formally, let $\mathcal{H} \subseteq X \rightarrow \{-1,1\}$ be the space of weak classifiers, and $S = \{(x_j, y_j)\}_{j=1}^m$ is the sample used to train a strong classifier $f = \sum_{h \in \mathcal{H}} \alpha_h h$.

The margin of $f$ on the $j$th sample point is defined as $\theta_j = y_j f(x_j)$.
Formally, let $\mathcal{H} \subseteq \mathcal{X} \to \{-1,1\}$ be the space of weak classifiers, and $S = \{(x_j, y_j)\}_{j=1}^m$ is the sample used to train a strong classifier $f = \sum_{h \in \mathcal{H}} \alpha_h h$. The margin of $f$ on the $j$th sample point is defined as $\theta_j := y_j f(x_j)$.
Margin Theory

- Formally, let $\mathcal{H} \subseteq \mathcal{X} \rightarrow \{-1,1\}$ be the space of weak classifiers, and $S = \{(x_j, y_j)\}_{j=1}^m$ is the sample used to train $f = \sum_{h \in \mathcal{H}} \alpha_h h$.

  - If $\theta_j$ is positive, then $\text{sign}(f)$ classifies $(x_j, y_j)$ correctly.

- The margin of $f$ on the $j^{th}$ sample point is defined as $\theta_j := y_j f(x_j)$.
Margin Theory

Formally, let $\mathcal{H} \subseteq \mathcal{X} \rightarrow \{-1,1\}$ be the space of weak classifiers, and $S = \{(x_j, y_j)\}_{j=1}^{m}$ is the sample used to train $f = \sum_{h \in \mathcal{H}} \alpha_h h$.

Intuitively, the closer $\theta_j$ is to 1, the more “confident” $f$ is.

The margin of $f$ on the $j^{th}$ sample point is defined as $\theta_j := y_j f(x_j)$.
Schapire et al. (1998) showed the following bound on the error probability of voting classifiers.

\[
\Pr_{(x,y) \sim D} [yf(x) \leq 0] 
\leq \Pr_{(x,y) \sim S} [yf(x) \leq \theta] + O\left(\sqrt{\frac{\ln|H| \ln m}{m\theta^2}}\right)
\]
Schapire et al. (1998) showed the following bound on the error probability of voting classifiers.

\[
\Pr_{(x, y) \sim \mathcal{D}} [y f(x) \leq 0] \leq \frac{|\mathcal{H}| \ln m}{m \theta^2} + O(\ln |\mathcal{H}|)
\]

The error probability of \( f \) with respect to the unknown distribution \( \mathcal{D} \) over \( \mathcal{X} \times \{-1, 1\} \).
Schapire et al. (1998) showed the following bound on the error probability of voting classifiers.

\[
\Pr_{(x,y) \sim \mathcal{D}} [y f(x) \leq \theta] \leq \Pr_{(x,y) \sim \mathcal{S}} [y f(x) \leq \theta] + O \left( \sqrt{\frac{\ln |\mathcal{H}| \ln m}{m \theta^2}} \right)
\]

The fraction of sample points with margin at most \(\theta\).
Schapire et al. (1998) showed the following bound on the generalization error of voting classifiers:

\[
\Pr_{(x,y) \sim D} [yf(x) \leq 0] 
\leq \Pr_{(x,y) \sim S} [yf(x) \leq \theta] + O\left(\sqrt{\frac{\ln|\mathcal{H}| \ln m}{m\theta^2}}\right)
\]

Holds for all voting classifiers \(f\) and margins \(\theta \in (0,1]\).
Schapire et al. (1998) showed the following bound on the error probability of voting classifiers.

\[
\Pr_{x,y \sim S} [y f(x) \leq \theta] + O \left( \frac{\ln |\mathcal{H}| \ln m}{\sqrt{m \theta^2}} \right)
\]

This holds with high probability over the choice of the \(m\) sample points.
Schapire et al. (1998) showed the following bound on the error probability of voting classifiers.

$$\Pr_{(x,y) \sim D} [y f(x) \leq 0] \leq \Pr_{(x,y) \sim S} [y f(x) \leq \theta] + O\left(\sqrt{\frac{\ln|\mathcal{H}| \ln m}{m \theta^2}}\right)$$

The result gave rise to boosting algorithms that intentionally aim to optimize margins.
Breimann (1999) showed the following bound on the error probability of voting classifiers.

\[
\Pr_{(x,y) \sim D} [yf(x) \leq 0] \leq O \left( \frac{\ln |\mathcal{H}| \ln m}{m \hat{\theta}^2} \right)
\]

Holds for all voting classifiers \( f \) where \( \hat{\theta} \) is the minimum margin.
Breimann (1999) showed the following bound on the error probability of voting classifiers.

\[ \Pr_{(x,y) \sim D} [y f(x) \leq 0] \leq O \left( \frac{\ln |\mathcal{H}| \ln m}{m \hat{\theta}^2} \right) \]

This holds with high probability over the choice of the \( m \) sample points.

Holds for all voting classifiers \( f \) where \( \hat{\theta} \) is the minimum margin.
State-of-the-Art bounds were given by Gao and Zhou (2013)

\[
\mathbb{P}_{(x,y) \sim \mathcal{D}} [yf(x) \leq 0] \leq \mathbb{P}_{(x,y) \sim \mathcal{S}} [yf(x) \leq \theta] + O\left(\frac{\ln|\mathcal{H}| \ln m}{m\theta^2} + \sqrt{\frac{\ln|\mathcal{H}| \ln m}{m\theta^2}} \mathbb{P}_{(x,y) \sim \mathcal{S}} [yf(x) \leq \theta]\right)
\]
Margin-Based Upper Bounds

This holds with high probability over the choice of the $m$ sample points.

\[
\Pr_{(x,y) \sim \mathcal{D}}[y f(x) \leq 0] \leq \Pr_{(x,y) \sim \mathcal{S}}[y f(x) \leq \theta] + O \left( \frac{\ln|\mathcal{H}| \ln m}{m \theta^2} + \sqrt{\frac{\ln|\mathcal{H}| \ln m}{m \theta^2} \Pr_{(x,y) \sim \mathcal{S}}[y f(x) \leq \theta]} \right)
\]

Holds for all voting classifiers $f$ and margins $\theta \in (0,1]$.
Despite being studied for over two decades, the tightness of margin-based generalization bounds was not settled.

In fact, no margin-based lower bounds were known.
Our main result shows that any algorithm $\mathcal{A}$ optimizing margins cannot do much better than the known upper bounds.
Margin-Based Lower Bounds

Formally, for all $N, \theta, \tau$ there exist a set $X$ and a hypothesis set $H$ such that for every large enough $m$ and algorithm $A$ that optimizes margins there exists a distribution $D$ for which

$$\Pr_{(x,y) \sim D} [yf_A(x) \leq 0] \geq \Pr_{(x,y) \sim S} [yf_A(x) \leq \theta]$$

$$+ O \left( \frac{\ln |H|}{\theta^2} + \sqrt{\frac{\ln |H|}{\theta^2}} \Pr_{(x,y) \sim S} [yf_A(x) \leq \theta] \right)$$
Margin-Based Lower Bounds

Formally, \( \forall N, \theta, \tau \) there exist a set \( X \) and a hypothesis set \( \mathcal{H} \) such that for every large enough \( m \) and algorithm \( \mathcal{A} \) that optimizes margins there exists a distribution \( \mathcal{D} \) for which

\[
\Pr_{(x,y) \sim \mathcal{D}} [yf_{\mathcal{A}}(x) \leq 0] \geq \Pr_{(x,y) \sim \mathcal{S}} [yf_{\mathcal{A}}(x) \leq \theta]
\]

\[
+ O \left( \frac{\ln |\mathcal{H}|}{\theta^2} + \sqrt{\frac{\ln |\mathcal{H}|}{\theta^2} \Pr_{(x,y) \sim \mathcal{S}} [yf_{\mathcal{A}}(x) \leq \theta]} \right)
\]

Where \( \theta \in \left( \frac{1}{N}, \frac{1}{40} \right) \) and \( \tau \in \left[ 0, \frac{49}{100} \right] \) are not too large.
Margin-Based **Lower Bounds**

Formally, \( \forall N, \theta, \tau \) There exist a set \( \mathcal{X} \) and a hypothesis set \( \mathcal{H} \) such that for every large enough \( m \) and algorithm \( \mathcal{A} \) that optimizes margins there exists a distribution \( \mathcal{D} \) for which

\[
\Pr_{(x,y) \sim \mathcal{D}}[yf_{\mathcal{A}}(x) \leq 0] \geq \Pr_{(x,y) \sim \mathcal{S}}[yf_{\mathcal{A}}(x) \leq \theta] + O\left(\frac{\ln|\mathcal{H}|}{\theta^2} + \sqrt{\frac{\ln|\mathcal{H}|}{\theta^2}} \Pr_{(x,y) \sim \mathcal{S}}[yf_{\mathcal{A}}(x) \leq \theta]\right)
\]

Small set of weak classifiers, 
\[
\ln |\mathcal{H}| = \Theta(\ln N)
\]

Margin-Based Generalization Lower Bounds for Boosted Classifiers
Margin-Based *Lower* Bounds

Formally, \( \forall N, \theta, \tau \) There exist a set \( X \) and hypothesis set \( \mathcal{H} \) such that for every large enough \( m \) and algorithm \( \mathcal{A} \) that optimizes margins there exists a distribution \( D \) for which

\[
Pr_{(x,y) \sim D} [yf_\mathcal{A}(x) \leq 0] \geq Pr_{(x,y) \sim S} [yf_\mathcal{A}(x) \leq \theta]
\]

\[
+ O \left( \frac{\ln |\mathcal{H}|}{\theta^2} + \sqrt{\frac{\ln |\mathcal{H}|}{\theta^2}} Pr_{(x,y) \sim S} [yf_\mathcal{A}(x) \leq \theta] \right)
\]
Margin-Based Lower Bounds

Formally, \( \forall N, \theta, \tau \) There exist a set \( \mathcal{X} \) and a hypothesis set \( \mathcal{H} \) such that for every large enough \( m \) and algorithm \( \mathcal{A} \) that optimizes margins there exists a distribution \( \mathcal{D} \) for which

\[
\Pr_{(x,y) \sim \mathcal{D}} [yf_{\mathcal{A}}(x) \leq 0] \geq \Pr_{(x,y) \sim \mathcal{S}} [yf_{\mathcal{A}}(x) \leq \theta]
\]

\[
+ O \left( \frac{\ln|\mathcal{H}|}{\theta^2} + \sqrt{\frac{\ln|\mathcal{H}|}{\theta^2}} \Pr_{(x,y) \sim \mathcal{S}} [yf_{\mathcal{A}}(x) \leq \theta] \right)
\]

The classifier returned by \( \mathcal{A} \).
Margin-Based Lower Bounds

Formally, \( \forall N, \theta, \tau \) there exist a set \( \mathcal{X} \) and a hypothesis set \( \mathcal{H} \) such that for every large enough \( m \) and algorithm \( \mathcal{A} \) that optimizes margins there exists a distribution \( \mathcal{D} \) for which

\[
\Pr_{(x, y) \sim \mathcal{D}} [y \cdot f_{\mathcal{A}}(x) \leq 0] \geq \Pr_{(x, y) \sim \mathcal{S}} [y \cdot f_{\mathcal{A}}(x) \leq \theta]
\]

assuming this is at most \( \tau \).

\[
+ O \left( \frac{\ln|\mathcal{H}|}{\theta^2} + \sqrt{\frac{\ln|\mathcal{H}|}{\theta^2}} \Pr_{(x, y) \sim \mathcal{S}} [y \cdot f_{\mathcal{A}}(x) \leq \theta] \right)
\]
Margin-Based Lower Bounds

Formally, \( \forall N, \theta, \tau \) There exist a set \( X \) and a hypothesis set \( \mathcal{H} \) such that for every large enough \( m \) and algorithm \( \mathcal{A} \) that optimizes margins there exists a distribution \( \mathcal{D} \) for which

\[
\Pr_{(x,y) \sim \mathcal{D}} [yf_\mathcal{A}(x) \leq 0] \geq \Pr_{(x,y) \sim \mathcal{S}} [yf_\mathcal{A}(x) \leq \theta] + O\left(\frac{\ln|\mathcal{H}|}{\theta^2} + \sqrt{\frac{\ln|\mathcal{H}|}{\theta^2}} \Pr_{(x,y) \sim \mathcal{S}} [yf_\mathcal{A}(x) \leq \theta]\right)
\]
Summary

- We show margin-based generalization lower bounds which almost match the best known upper bounds.
- These bounds essentially complete the theory of generalization bounds based on margins alone.
- Open Question: Are there parameters other than margin that can be used to better explain the practical properties of voting classifiers?