Results on the Convergence of Boosting Algorithms

Cynthia Rudin, PhD
NSF Postdoctoral Research Fellow
New York University

Joint work with Rob Schapire and Ingrid Daubechies
Dedicated to Leo Breiman
Motivation

• Interesting history and analysis
  • Margin theory, coming up
  • AdaBoost has great dynamics!

• Boosting algorithms are useful
  • In fact, really useful (Caruana and Niculescu-Mizil ICML 05)
  • Freund and Schapire, Gödel prize for AdaBoost
The edge \( r_t \) measures how well the weak learning algorithm performs at iteration \( t \).
"the edge"
Outline

- History of the Margin Theory

Results:

- “The Case of Bounded Edges” for AdaBoost

- A Convergence Rate for arc-gv

- (AdaBoost is a Ranking Algorithm)
  (joint work with Corinna Cortes, Mehryar Mohri, and Rob Schapire, COLT 05)
History of the Margin Theory for AdaBoost

• Freund and Schapire design AdaBoost in 1996.

• AdaBoost often tends not to overfit, even when training error is zero. (Breiman 96, Cortes and Drucker 97)

• Margin theory for boosting (Schapire, Freund, Bartlett and Lee 98) (Note: boosting margin not quite the same as svm margin.) (Note: margin = “distance” to decision boundary.)

• Remember, AdaBoost was invented before the margin theory.

• Does AdaBoost maximize the margin? Can we understand AdaBoost’s convergence?
History of the Margin Theory for AdaBoost

• Does AdaBoost maximize the margin? Can we understand AdaBoost’s convergence?

Empirical results:

• Yes: AdaBoost seemed to maximize the margin in the limit. (Grove and Schuurmans 98, Rätsch and Warmuth 02, and others)

• No: Breiman disagrees. (Breiman 96)
  • His “proof” is arc-gv.
  • arc-gv achieves larger margins, but worse error than AdaBoost.
  • (Somewhat indecipherable) proof shows arc-gv maximizes the margin... asymptotically. (see Meir and Rätsch 02)
  • Neither AdaBoost nor arc-gv’s convergence really understood.

• In the meantime...
History of the Margin Theory for AdaBoost

• AdaBoost generates a margin that is at least $\rho / 2$, where $\rho$ is the maximum margin. (Schapire et al. 98)
History of the Margin Theory for AdaBoost

- AdaBoost generates a margin that is at least $\rho / 2$, where $\rho$ is the maximum margin. (Schapire et al. 98)

- AdaBoost generates a margin that is at least $Y(\rho) \geq \rho / 2$. (Rätsch and Warmuth 02)

$$Y(\rho) := -\ln(1 - \rho^2) / \ln\left(1 + \frac{\rho}{1 - \rho}\right)$$

“Gap in Theory”

margin AdaBoost achieves

optimal margin

(Schapire et al. 98)
History of the Margin Theory for AdaBoost

• The questions remain:

A) Does AdaBoost maximize the margin?

B) Should we maximize the margin?
i.e., which is better, AdaBoost or Breiman’s arc-gv? (Breiman says AdaBoost.)

C) If arc-gv is useful, can we understand its convergence?
Outline

- History of the Margin Theory

- “The Case of Bounded Edges” for AdaBoost

  A) Does AdaBoost maximize the margin?

  B) Should we maximize the margin?
  i.e., which is better, AdaBoost or Breiman’s arc-gv?
  (Breiman says AdaBoost.)

- A Convergence Rate for arc-gv

  C) If arc-gv is useful, can we understand its convergence?
A) Does AdaBoost maximize the margin?
No: AdaBoost may converge to a margin that is significantly below maximum. (Rudin, Daubechies, Schapire 04)
A) Does AdaBoost maximize the margin?
No: AdaBoost may converge to a margin that is significantly below maximum. (Rudin, Daubechies, Schapire 04)

• How bad is it?
Main Result 1

Theorem ("The Case of Bounded Edges")

The bound of (Rätsch and Warmuth 02) is tight, i.e., AdaBoost will converge to a margin of $Y(\rho)$ whenever $\lim r_t = \rho$.

(Note: this is a specific case of a more general theorem.)

$$Y(\rho) := -\ln(1-\rho^2) / \ln\left(\frac{1+\rho}{1-\rho}\right)$$

![Graph showing $Y(\rho)$ and the optimal margin for different values of $\rho$.](image)

- $(3/8, 1/3)$
- $\rho$
- $Y(\rho)$ (Rätsch & Warmuth 02)
- $\rho / 2$
- (Schapire et al. 98)

margin AdaBoost achieves

optimal margin
Approaching Main Result 1:
AdaBoost as a coordinate descent algorithm

\[ F(\lambda) := \sum_{i=1}^{m} \exp[-(M\lambda)_i] \]

(Breiman 99, Friedman et al. 00, Rätsch et al. 01, Duffy and Helmbold 99, Mason et al. 00)
Approaching Main Result 1:
AdaBoost as a coordinate descent algorithm

\[ F(\lambda) := \sum_{i=1}^{m} \exp[-(M\lambda)_i] \]

\[ \{(x_i, y_i)\}_{i=1,...,m}, \text{where } (x_i, y_i) \in X \times \{-1,1\} \]

\[ h_j : X \rightarrow \{-1,1\}, \quad j = 1,...,n \]

• features, or “weak classifiers”
• can be produced using a “weak learning algorithm”

\[ M \in \{-1,1\}^{m \times n}, M_{ij} = y_i h_j(x_i) \]

• \(M\) is the “input” to AdaBoost

\(\lambda \in R^n\)
Approaching Main Result 1:
AdaBoost as a coordinate descent algorithm

\[ F(\lambda) := \sum_{i=1}^{m} \exp[-(M\lambda)_i] \]

At iteration \( t \): choose direction \( j_t \) and distance determined by \( r_t \).
AdaBoost:

for $t=1\ldots T_{final}$

calculate edge $r_t$

choose coordinate $j_t$

calculate distance $a_t$

dependent for
AdaBoost:

$\lambda_1 = 0$, $M$ given

for $t=1 \ldots T_{\text{final}}$

\(d_{t,i} = \frac{1}{Z} e^{-(M\lambda_t)_i}\) for all $i$, where $Z$ normalizes

\[j_t \in \begin{cases} \arg \max_j (d_t^T M)_j & \text{"optimal case"} \\ \{ j : (d_t^T M)_j \geq \rho \} & \text{"non-optimal case"} \end{cases}\]

\[r_t = (d_t^T M)_{j_t}\]

\[\alpha_t = \frac{1}{2} \ln \left( \frac{1 + r_t}{1 - r_t} \right)\]

\[\lambda_{t+1} = \lambda_t + \alpha_t e_{j_t}\]

end for
arc-gv:

\( \lambda_1 = 0, \ M \) given

for \( t=1...T_{final} \)

\[ d_{t,i} = \frac{1}{Z} e^{-(M\lambda_t)_i} \] for all \( i \), where \( Z \) normalizes

\[ j_t \in \begin{cases} \arg \max_j (d_t^T M)_j & \text{"optimal case"} \\ \{ j : (d_t^T M)_j \geq \rho \} & \text{"non-optimal case"} \end{cases} \]

\[ r_t = (d_t^T M)_{j_t} \]

\[ \alpha_t = \frac{1}{2} \ln \left( \frac{1+r_t}{1-r_t} \right) - \frac{1}{2} \ln \left( \frac{1+\mu_t}{1-\mu_t} \right) \]

where \( \mu_t = \min_i (M\lambda_t)_i \)

\[ \lambda_{t+1} = \lambda_t + \alpha_t e_{j_t} \]

end for
$r_t$ “the edge”:

- $(1-r_t)/2 = \text{Prob. Error (with respect to } d_t) \text{ between } y_i \text{'s and } h_{jt}.$ (measures accuracy of weak classifier)

- Directional derivative along the $j_t^{th}$ direction.
Two cases pinpointed by Rätsch and Warmuth for direction.

- **optimal case**: stumps
- **non-optimal case**: decision trees, neural networks

$$j_t \in \begin{cases} \arg \max_j (d_t^T M)_j & \text{"optimal case"} \\ \{j : (d_t^T M)_j \geq \rho \} & \text{"non-optimal case"} \end{cases}$$

$$r_t = (d_t^T M)_{j_t}$$

Edge obeys $r_t = \rho$, where $\rho = \max_\lambda \mu(f_{\lambda_t})$ and

$$\mu(f_{\lambda_t}) := \min_i \frac{(M \lambda_t)_i}{\|\lambda_t\|_1}.$$
Theorem ("The Case of Bounded Edges") (General version)

- If AdaBoost’s edge $r_t$ is within $[\rho', \rho'+\sigma]$, its asymptotic margin will be within $[Y(\rho'), Y(\rho'+\sigma)]$.
- If $\lim_{t \to \infty} r_t = \rho$, then AdaBoost will converge to a margin of $Y(\rho)$.
- If the edges are bounded within a small interval, so is the margin!
- The bound of (Rätsch and Warmuth 02) is exactly tight.
1 trial AdaBoost, largest edge
8 trials AdaBoost with pre-determined edge

As pre-specified edge increases, asymptotic margin increases.

\[ Y(r) := -\frac{\ln(1 - r^2)}{\ln(1 + \frac{r}{\sqrt{1 - r}})} \]
Theorem (“The Case of Bounded Edges”) (General version)

- If AdaBoost’s edge $r_t$ is within $[\rho', \rho'+\sigma]$, its asymptotic margin will be within $[Y(\rho'), Y(\rho'+\sigma)]$. 

- If $\lim r_t = \rho$, then AdaBoost will converge to a margin of $Y(\rho)$. 

- If the edges are bounded within a small interval, so is the margin! 
  - This implies the bound of (Rätsch and Warmuth 02) is exactly tight. 
  - We can “coerce” AdaBoost to converge to any margin we’d like!

Theorem (bound of previous Theorem is non-vacuous)

Given $\rho'$ and $\sigma$, there is some matrix $M$ such that AdaBoost may choose an infinite sequence of weak classifiers with edge values in $[\rho', \rho'+\sigma]$.
• Case of Bounded Edges:
  We understand AdaBoost’s convergence!

Now an insightful experiment.
Outline

- History of the Margin Theory

- “The Case of Bounded Edges” for AdaBoost

  A) Does AdaBoost maximize the margin? No! And we know how bad it fails.

  B) Should we maximize the margin? i.e., which is better, AdaBoost or Breiman’s arc-gv? (Breiman says AdaBoost.)

- A Convergence Rate for arc-gv

  C) If arc-gv is useful, can we understand its convergence?
1 trial AdaBoost, largest edge
8 trials AdaBoost with pre-determined edge
As the margin increases, error decreases!
Outline

- History of the Margin Theory
- “The Case of Bounded Edges” for AdaBoost

A) Does AdaBoost maximize the margin? No! And we know how bad it fails.

B) Should we maximize the margin? i.e., which is better, AdaBoost or Breiman’s arc-gv? (Breiman says AdaBoost.)

Grove and Schuurmans agree...
Cynthia doesn’t...
- controlled experiment
  - AdaBoost can get “stuck” in a degenerate cycle
Reyzin and Schapire - ICML 2006 - provide insight...
Outline

- History of the Margin Theory
- “The Case of Bounded Edges” for AdaBoost
  A) Does AdaBoost maximize the margin? No! And we know how bad it fails.
  B) Should we maximize the margin? i.e., which is better, AdaBoost or Breiman’s arc-gv? (Breiman says AdaBoost.)
- A Convergence Rate for arc-gv
  C) If arc-gv is useful, can we understand its convergence?
• AdaBoost Sparkler
  • rum, or perhaps bourbon or vermouth
  • grenadine, or was that orange?
  
  • you never know exactly what… ingredients the bartender will converge to.

• arc-gv Martini (designed by Breiman)
  • 2/3 gin, 1/3 crème de menthe, superfine sugar, mint leaves.
  
  • Breiman would perhaps have liked the first one better, whatever it is.
arc-gv:

\( \lambda_1 = 0, \ M \) given

for \( t=1\ldots T_{\text{final}} \)

\[
d_{t,i} = \frac{1}{Z} e^{-\langle M\lambda_t \rangle_i} \quad \text{for all } i, \text{ where } Z \text{ normalizes}
\]

\[
\hat{j}_i \in \begin{cases} \arg \max_j \langle d_t^T M \rangle_j & \text{"optimal case"} \\ \{ j : \langle d_t^T M \rangle_j \geq \rho \} & \text{"non-optimal case"} \end{cases}
\]

\[
r_t = \langle d_t^T M \rangle_{\hat{j}_i}
\]

\[
\alpha_t = \frac{1}{2} \ln \left( \frac{1 + r_t}{1 - r_t} \right) - \frac{1}{2} \ln \left( \frac{1 + \mu_t}{1 - \mu_t} \right)
\]

\[
\lambda_{t+1} = \lambda_t + \alpha_t e_{\hat{j}_i} \quad \text{where } \mu_t = \min_i \langle M\lambda_t \rangle_i
\]

end for

- arc-gv is not coordinate descent
Towards a convergence rate for arc-gv

\[ F(\lambda) := \sum_{i=1}^{m} \exp[-(M\lambda)_i] \quad \text{AdaBoost’s objective} \]

\[ G(\lambda) := \frac{-\ln F(\lambda)}{\|\lambda\|_1} \quad \text{Smooth Margin} \]

\[ \mu(\lambda) := \min_i \frac{(M\lambda)_i}{\|\lambda\|_1} \quad \text{Margin} \]

• Hard to measure progress wrt margin
• Possible: measure wrt smooth margin
• Use Recursive relation:

\[ \|\lambda_{t+1}\|_1 G(\lambda_{t+1}) - \|\lambda_t\|_1 G(\lambda_t) = \int \tanh u \, du \]

\[ \text{arctanh}(r_t) \quad \text{arctanh}(r_t)-\text{stepsize} \]
Theorem (Convergence Rate for arc-gv)

arc-gv will have achieved a margin in $[\rho - \varepsilon, \rho]$ within at most

$$C_1 + C_2 \varepsilon \left[\frac{-(3-\rho)}{(1-\rho)}\right]$$

iterations, where $\rho$ is the maximum margin.
Theorem (Convergence Rate for arc-gv)

Let $\tilde{1}$ be the iteration at which $G$ becomes positive. Then
$$\max_{\ell=\tilde{1},\ldots,t} \mu(\lambda_{\ell})$$
will be in $[\rho - \varepsilon, \rho]$ within at most
$$\tilde{1} + (\|\lambda_{\tilde{1}}\|_1 + \ln 2)\varepsilon^{-(3-\rho)/(1-\rho)}$$
iterations, where $\rho$ is the maximum margin.

Main Result 2
• So we really do understand arc-gv now!

• But which one is better… AdaBoost or arc-gv?

• Problem Domain
  • Theoretical: arc-gv
  • Experimental: AdaBoost

• Pruning / Complexity Control (see Reyzin and Schapire 06)
• Margin Distribution

• Non-asymptotic regime, strength of weak learning algorithm

• Plenty of open questions! Plenty of other algorithms!
1 trial AdaBoost, largest edge
8 trials AdaBoost with pre-determined edge
Outline

- History of the Margin Theory

- “The Case of Bounded Edges” for AdaBoost
  
  AdaBoost does not maximize the margin, and we know how bad it fails.

- A Convergence Rate for arc-gv
  
  arc-gv does maximize the margin, and we have a new rate of convergence.

- And there are plenty of beautiful open questions.
Thank you!

Special thanks to: Eero Simoncelli, the NSF, and the organizers: Joe Verducci, Xiaotong Shen, John Lafferty