A Stochastic Approximation View of Boosting

C. Andy Tsao   Yuan-chin Ivan Chang*

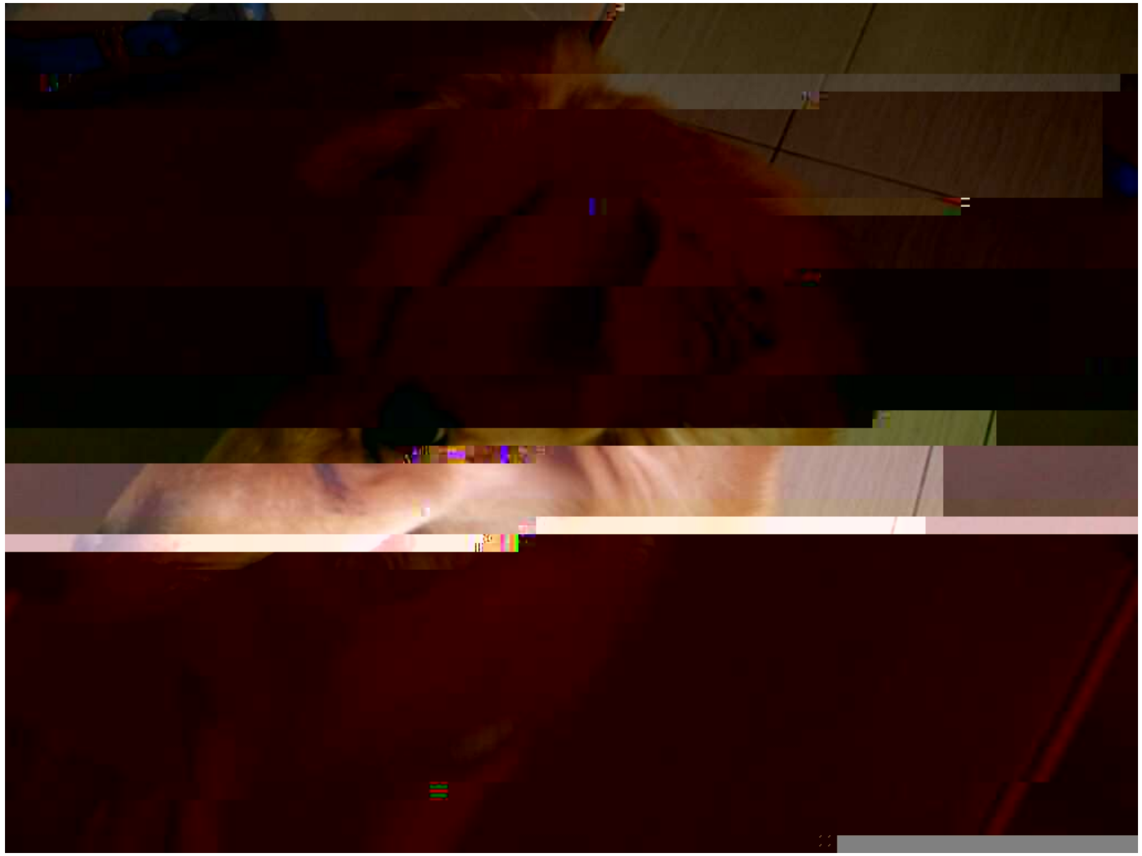
Department of Applied Math, National Dong Hwa University

Institute of Statistical Science, Academia Sinica*

Outline

- Introduction
- FHT's Interpretation
- Stochastic Approximation Viewpoint
- Results
- Conclusion and Discussion

Murray at work



A not-so-Lean not-so-Mean Stat Machine

Intro: Supervised Learning

- Training data $(\mathbf{x}_i, \mathbf{y}_i)_1^N$; $\mathbf{x} \in \mathbf{X}, \mathbf{y} \in \mathbf{Y} = \{\pm 1\}$.

  Find *Machine* (Classifier) $\mathbf{H}_f : \mathbf{X} \to \mathbf{Y}$

- Testing Data $(\mathbf{x}'_j, \mathbf{y}'_j)_1^M$.

- Training Error

$$\mathbf{TE} = \frac{1}{N} \sum_i \mathbf{1}_{[y_i \neq H(x_i)]}$$

- Generalization Error

$$\mathbf{GE} = 1$$

Intro: Boosting

- *Weak (base) learner*
- Sequentially applying it to reweighted version of the training data
  - Higher weights on the previous misclassified data
  - Boosting iteration: **T**
- Weighted majority vote

Schapire (1990), Freund and Schapire (1997), Friedman, Hastie and Tibshirani (2000).

Brieman (2004), Jiang (2004), etc.

Intro: Discrete AdaBoost

1. Start with weights $\mathbf{D_t(i)} = 1\mathbf{=N; i} = 1$ to $\mathbf{N:}$
2. Repeat for $\mathbf{t} = 1$ to $\mathbf{T}$
   - Obtain $\mathbf{h_t(x)}$ from weak learner $\mathbf{h}$ using weighted training data wrt $\mathbf{D_t}$
   - Compute $\ _{\mathbf{t}} = \mathbf{E}_{\mathbf{D}_t} 1_{[\mathbf{y} \neq \mathbf{h}_t(\mathbf{x})]}$; $\ _{\mathbf{t}} = \log \frac{1- \ _t}{\ _t}$:
   - Update $\mathbf{i} = 1$ to $\mathbf{N}$,

$$\mathbf{D}_{\mathbf{t}+1}(\mathbf{i}) = \frac{1}{\mathbf{Z_t}} \mathbf{D_t(i)} \exp[\ _{\mathbf{t}} 1_{[\mathbf{y}_i \neq \mathbf{h}_t(\mathbf{x}_i)]}];$$

   where $\mathbf{Z_t}$ is the normalizer.

3. Output the classifi er $\mathbf{sgn}\left[\sum_{\mathbf{t}=1}^{\mathbf{T}} \ _{\mathbf{t}} \mathbf{h_t(x)}\right]$

Freund and Schapire (1997)

Intro: Theories and Observation-1

- Under *weak base hypothesis assumption*, TE goes to zero exponentially as **T ! 1**
- PAC bound for GE: Let **d**: VC dim of the weak base hypothesis space.

  $$\mathbf{GE} \quad \mathbf{TE} + \mathbf{O}(\sqrt{\frac{Td}{N}})$$

- Margin. Schapire, et al (1998). $\mathbf{m}(\mathbf{x}_i; \mathbf{y}_i) = \frac{y_i \quad _t a_t h_t(x_i)}{_t a_t}$.

  $\mathbf{GE} \quad \hat{\mathbf{P}}[\mathbf{m}(\mathbf{x}; \mathbf{y}) \quad] + \mathbf{O}(\sqrt{\frac{d}{N\theta^2}})$ for any **>** 0 with high probability.

Intro: Theories and

FHT's Interpretation

Friedman, Hastie and Tibishirani (2000).

Result 1 *The Discrete AdaBoost (population version) builds an additive logistic regression model via Newton-like updates for minimizing* $\mathbf{E}(e^{-YF(X)})$:

SWOT

Strength

- Easier for statisticians than MLs6 -0 1190.R143 Do Q 0 (

- For $c > 0$, $f(\mathbf{x}) = \mathrm{sgn}(E_w(y|\mathbf{x}))$ minimizes (1),
- Given $f(\mathbf{x}) = \pm 1$,

$$c = \arg\min_c J(F + cf) = \arg\min_c E_w e^{-cyf(x)}$$

$$= \frac{1}{2}\log\left(\frac{1 - err}{err}\right); \qquad err = E_w 1_{[yf(x)<0]}$$

Alternative expression

$$\mathbf{F}_{t+1}(\mathbf{x}) = \mathbf{F}_t(\mathbf{x}) + {}_t\ \mathbf{sgn}(\mathbf{E}_{w_t}(\mathbf{Y}|\mathbf{x}))$$

where

$${}_t = \log(\frac{1 \qquad {}_t}{}$$

Remarks on FHT

- $f(\mathbf{x})$ is fixed only when $\mathbf{x}$ is given.
- $E_{X,Y}(e^{-YF(X)}) = E_X\left[E_{Y|X}(e^{-YF(X)})\right]$
- Convergence of the Newton-like updates via minimizing

$$(F) = E_{Y|x} L(Y; F(\mathbf{x}))?$$

- $L(Y;F) = e^{-YF}$

Friedman (2001).

SA: Robbins-Monro Awgorithm

Goal: Find

$$\mathbf{F}_* = \mathbf{arg}\min_F \quad (\mathbf{F}) \quad [= \mathbf{E}_{Y|x}\mathbf{L}(\mathbf{Y};\mathbf{F}(\mathbf{x}))]$$

$$\mathbf{F}_* \quad ! \quad {}'(\mathbf{F}_*) = \mathbf{E}_{Y|x}\mathbf{L}'(\mathbf{Y};\mathbf{F}_*) = 0$$

where

$$\mathbf{L}'(\mathbf{Y};\mathbf{F}) = \frac{@}{@\mathbf{F}}\mathbf{L}(\mathbf{Y};\mathbf{F})$$

Duflo (1997).

Robbins and Monro (1951), Kiefer and Wolfowitz (1952).

RM-D Algorithm

- Choose $\mathbf{F}_0$ arbitrarily.
- Iterate $\mathbf{t} = 0; 1; :::$

$$\mathbf{g}_{t+1} = \mathbf{E}\,\mathbf{L}^{'}(\mathbf{Y}_{t+1}; \mathbf{F}_t)$$

$$\mathbf{F}_{t+1} = \mathbf{F}_t \quad {}_t\, \mathbf{g}_{t+1}; \qquad {}_t \quad 0$$

RM-D with Exponential Criterion

- Choose $\mathbf{F}_0$ arbitrarily.
- Iterate **t** = 0; 1; :::; $\mathbf{Y}_t \quad ^{iid}$ **Y jx**

$$
\begin{aligned}
\mathbf{g} \quad &= \quad \mathbf{EY}_{t+1}\,\mathrm{exp}(\quad \mathbf{Y}_{t+1}\mathbf{F}_t) \qquad\qquad t{+}1 \\
\mathbf{F} \quad &= \mathbf{F}_t \qquad {}_t\mathbf{g}_{t+1} \qquad\qquad\qquad\qquad\qquad t{+}1 \\
&= \mathbf{F}_t + \quad {}_t\mathbf{E}_{w_t}(\mathbf{Y}_{t+1}\mathbf{jx}); \qquad\qquad t \quad 0
\end{aligned}
$$

where $\mathbf{w}_t(\mathbf{x};\mathbf{y}) = \mathrm{exp}(\quad \mathbf{yF}_t(\mathbf{x}))$:

RM-D convergence

Proposition 1 *If is a conti. real ftn such that* $(\mathbf{F}_*) =$ *and for all* **F**

$(\ (\mathbf{F}) \quad)(\mathbf{F} \quad \mathbf{F}_*) > 0$ *and* $\mathbf{j} \ (\mathbf{F})\mathbf{j} \quad \mathbf{K}(1 + \mathbf{jFj});$ *for some* **K:**

Suppose $\mathbf{f} \ _t\mathbf{g}; \mathbf{f} \ _t\mathbf{g}$ *are seq's of reals and* $_t \quad 0.$ *Define*

$$\mathbf{F}_{t+1} = \mathbf{F}_t \quad _t (\ (\mathbf{F}_t) + \ _{t+1}):$$

If $_t$ **&** 0 *s.t.*

$$\sum_t \ _t = \mathbf{1}; \qquad \sum_t \ _t \ _t < \mathbf{1};$$

then $\mathbf{F}_t \ \mathbf{!} \ \mathbf{F}_*$ *for any initial* $\mathbf{F}_0$

- $(F) = E_{Y|x} L'(Y; F)$.
- t "step-size".
- Common choice

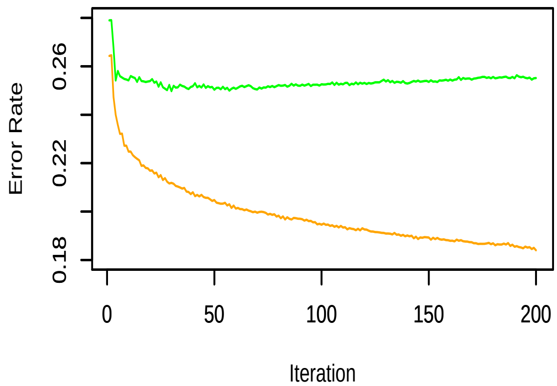
Table 1: Testing errors of SABoost, AdaBoost with Decision Stumps and RBF-Network as weak base learners.

METHOD	SABOOST	ADABOOST	ADABOOST
DATA SET	$\gamma = 1.0$	DS	RBF-N
IONOSPHERE	10.13±4.77	13.98±5.46	—
BUPA	28.92±7.00	32.56±8.10	—
PIMA	23.90±5.51	25.60±4.52	—
WDBC	4.18±2.66	3.01±2.25	—
DIABETES	25.06±1.90	25.52±1.99	26.47±2.29
GERMAN	26.50±2.33	26.81±2.47	27.45±2.50
HEART	18.16±3.23	20.01±3.76	20.29±3.44
SPLICE	13.06±0.83	14.22±1.09	10.14±0.51
TWONORM	9.59±1.18	5.85±0.57	3.03±0.28
WAVEFORM	14.73±2.43	13.18±5.69	10.84±0.58
BREAST	29.87±4.98	30.03±5.05	30.36±4.73

There are 50 replications and 1000 iterations for Ionosphere, Bupa liver-disorder, Pima Indian-Diabetes, and WDBC breast-cancer using random sampling. For Breast, Diabetes, German, Heart, Splice Twonorm, and Waveform, we follow the original partitions in IDA Benchmark Repository. Thus, there are 100 replications and 200 iterations for each of them except Splice. There are only 20 combinations in Splice.

Table 2: SA Boosting with different step sizes.

DATA SET	$\gamma = 0.5$	$\gamma = 1.5$	HYBRID ($\gamma = 0.5$)
IONOSPHERE	13.03±1.53	10.97±4.86	13.80±4.76
BUPA	29.47±7.12	29.04±6.95	30.58±8.60
PIMA	24.61±4.77	24.24±5.47	24.81±4.40
WDBC	4.77±2.89	3.97±2.59	2.80±2.09
DIABETIS	25.14±1.65	25.38±1.95	25.19±1.84
GERMAN	25.94±2.31	26.73±2.43	26.14±2.40
HEART	17.58±3.37	18.64±3.32	18.76±3.19
SPLICE	13.04±0.83	13.57±0.89	13.67±0.85
TWONORM	6.06±0.42	6.96±0.41	5.64±0.47
WAVEFORM	15.73±0.89	13.86±2.43	13.04±0.56
BREAST	29.94±4.93	30.07±5.11	30.03±5.05



Error Rate
0.18
0.22
0.26
0
50
100
150
200
Iteration

Table 4: Error rates (in %) of AdaBoost, SABoost ($\gamma = 0.5, 1$) with different base learner (SVM with Linear Kernel Polynomial Kernel with degree=3 and coef=1 and RBF kernel default parameter). Testing error(GE) and Training Error (TE).

Base Learner:	SVM-Linear	
Methods	GE (TE)	$\|TE - GE\|$
AdaBoost	34.7 (22.1)	0.122±0.20
SABoost $\gamma = 0.5$	31.8 (25.9)	0.059±0.06
SABoost $\gamma = 1$	35.5 (26.3)	0.091±0.09
SVM-Linear GE	46.64±4.84	
Base Learner:	SVM-Polynomial	
AdaBoost	23.06 (18.29)	3.31±0.86
SABoost $\gamma = 0.5$	23.81 (21.99)	2.06±0.39
SABoost $\gamma = 1$	22.96 (21.35)	1.74±0.39
SVM-Poly GE	22.49±1.86	
Base Learner:	SVM-RBF	
AdaBoost	23.63 (2. 41	3.34±0.76

SABoost $\gamma =$

SABoost $\gamma =$.

Conclusion/Discussion

- SA provides new insight/tools to boosting-like algorithms.
- Understandings
  - WBHA might not help.
  - “Theoretical" explanation for shrinkage
  - “Optimal" w