Comprehensive Review: Generalization in Boosting

Master's Level Data Science

Contents

1 Introduction 1

2 AdaBoost Recap and Training Error Bound 1

3 Empirical Observations on Overfitting 2

4 Margins and Generalization 2

5 Exponential Loss and Coordinate Descent View 3

6 Geometric Illustration 3

7 Algorithm Summary 3

8 Interpretation & Guidelines 4

9 Future Directions / Extensions 4

1 Introduction

This review synthesizes the lecture slides (`ensemble-3.pdf`) and audio transcript (`GeneralizationBoosting.txt`) on the surprising generalization behavior of AdaBoost. We cover:

- Recap of AdaBoost and its training error bound.
- Empirical observations on overfitting.
- The notion of *margin* and its role in generalization.
- Exponential loss interpretation and coordinate-descent view.
- Practical insights and extensions.

2 AdaBoost Recap and Training Error Bound

Given data $\{(x_i, y_i)\}_{i=1}^{n}$, $y_i \in \{-1, +1\}$, AdaBoost initializes weights

$$D_1(i) = \frac{1}{n},$$

and for $t = 1, \ldots, T$:

1. Train weak learner on D_t to obtain $h_t : X \to \{-1, +1\}$.
2. Compute weighted error
$$\varepsilon_t = \sum_{i=1}^{n} D_t(i)\, \mathbf{1}[h_t(x_i) \neq y_i].$$
3. Set classifier weight
$$\alpha_t = \tfrac{1}{2} \ln\Big(\frac{1 - \varepsilon_t}{\varepsilon_t}\Big).$$
4. Update and normalize:
$$D_{t+1}(i) \propto D_t(i)\, \exp\big(-\alpha_t\, y_i\, h_t(x_i)\big).$$

The final classifier is
$$H(x) = sign\Big(\sum_{t=1}^{T} \alpha_t\, h_t(x)\Big).$$

If each weak learner has edge $\gamma_t = \frac{1}{2} - \varepsilon_t > 0$, then the training error satisfies
$$\frac{1}{n} \sum_{i=1}^{n} \mathbf{1}[H(x_i) \neq y_i] \;\leq\; \exp\Big(-2 \sum_{t=1}^{T} \gamma_t^2\Big) \;\leq\; \exp\big(-2T\gamma^2\big),$$
where $\gamma = \min_t \gamma_t$. Thus training error decays *exponentially* in T.

3 Empirical Observations on Overfitting

Experiments on the UCI "letter" dataset by Freund and Schapire showed:

# Rounds	5	100	1000
Train error (%)	0.0	0.0	0.0
Test error (%)	8.4	3.3	3.1

Despite growing to over two million nodes (1000 trees), test error *kept decreasing*, contradicting the usual overfitting expectation.

4 Margins and Generalization

Define the *normalized score*
$$f(x) = \frac{\sum_{t=1}^{T} \alpha_t\, h_t(x)}{\sum_{t=1}^{T} \alpha_t},$$
and the *margin* on (x_i, y_i) by
$$\text{margin}_i = y_i\, f(x_i) \in [-1, 1].$$

A positive margin implies correct classification; larger margins reflect stronger confidence. Empirically:

- At 5 rounds: minimum margin ≈ 0.14.
- At 100 rounds: minimum margin ≈ 0.52.
- At 1000 rounds: minimum margin ≈ 0.55.

Even after zero training error, AdaBoost *continues to increase margins*, which correlates with improved test performance.

5 Exponential Loss and Coordinate Descent View

Boosting can be viewed as minimizing the *exponential loss*

$$L(f) \; = \; \frac{1}{n}\sum_{i=1}^{n} e^{-y_i f(x_i)},$$

an upper bound on the 0-1 loss. In the (potentially infinite) feature mapping

$$\phi(x) = \big(h(x)\big)_{h\in\mathcal{H}},$$

AdaBoost learns a linear classifier $f(x) = w \cdot \phi(x)$ by *coordinate descent*, each round adjusting one coordinate (the new weak classifier) to descend the exponential loss.

6 Geometric Illustration

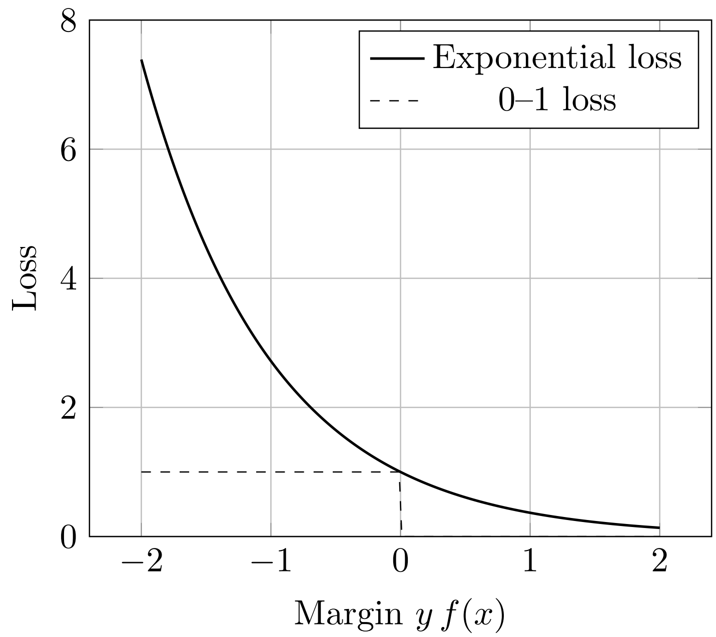

Figure 1: Exponential loss (solid) vs. 0–1 loss (dashed) as a function of margin.

7 Algorithm Summary

1. Initialize uniform weights $D_1(i) = 1/n$.
2. For $t = 1, \ldots, T$:
   (a) Train weak learner under D_t to get h_t.
   (b) Compute ε_t, set $\alpha_t = \frac{1}{2}\ln((1-\varepsilon_t)/\varepsilon_t)$.
   (c) Update and renormalize $D_{t+1}(i) \propto D_t(i)\exp(-\alpha_t y_i h_t(x_i))$.
3. Output $H(x) = sign(\sum_t \alpha_t h_t(x))$.

8 Interpretation & Guidelines

- **Margin maximization** drives generalization even after zero training error.
- **No overfitting** often observed: boosting focuses on increasing margins rather than merely fitting labels.
- **Monitoring**: one may stop when margin gains plateau rather than when training error hits zero.
- **Extensions**: adding shrinkage or early stopping can further control complexity.

9 Future Directions / Extensions

- **Gradient Boosting**: adapt boosting to arbitrary differentiable losses.
- **Regularized Boosting**: incorporate shrinkage (learning rate), subsampling (stochastic GBM).
- **Multi-class Boosting**: SAMME, multinomial boosting approaches.
- **Kernelized Boosting**: combine boosting with kernel methods for richer feature mappings.