Two Views of Boosting:

Margin vs. Convex Loss Minimization

Liwei Wang

Peking University

Background

- Learning and Classification:
  - Training examples
    $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n) \qquad x_i \in X, \quad y_i \in Y$
    i.i.d. from an underlying joint distribution P
  - Classifier: $C : X \to Y$
  - Generalization Error: $P(C(x) \neq y)$

The Boosting algorithm

Input: $S = (x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$
where $x_i \in X$, $y_i \in \{-1, 1\}$.
Initialization: $D_1(i) = 1/n$.
for $t = 1$ **to** T **do**

1. Train base learner using distribution D_t.
2. Get base classifier $h_t : X \to \{-1, 1\}$.
3. Choose α_t.
4. Update:

$$D_{t+1}(i) = \frac{D_t(i) \exp(-\alpha_t y_i h_t(x_i))}{Z_t},$$

where Z_t is a normalization factor chosen so that D_{t+1} will be a distribution.

end
Output: The final Classifier

$$H(x) = \operatorname{sgn}\left(\sum_{t=1}^{T} \alpha_t h_t(x)\right).$$

Background

- Empirical observation:
  - AdaBoost + Decision trees + Calibration =

    the best classification algorithm (Caruana, 2006, Breiman, 1998).

  - AdaBoost often resists to overfitting:
    - The test error of the combined classifier usually keeps decreasing as its size becomes very large, and even after the training error is zero, which seems contradicts the Occam’s razor!

Background

- We need a theoretical explanation of the Boosting algorithm:
  - Understanding the “mysteries”.
  - Develop more efficient algorithms.

Background

- A complete theoretical explanation should answer two questions:
  - Why AdaBoost often has good performance?
  - Why AdaBoost is often (though not always) immune to overfitting?

Outline

- The Margin Explanation
- The Convex Loss Explanation
- Margin vs. Convex Loss
- Open Problems

The Margin Explanation

The Concept of Margin

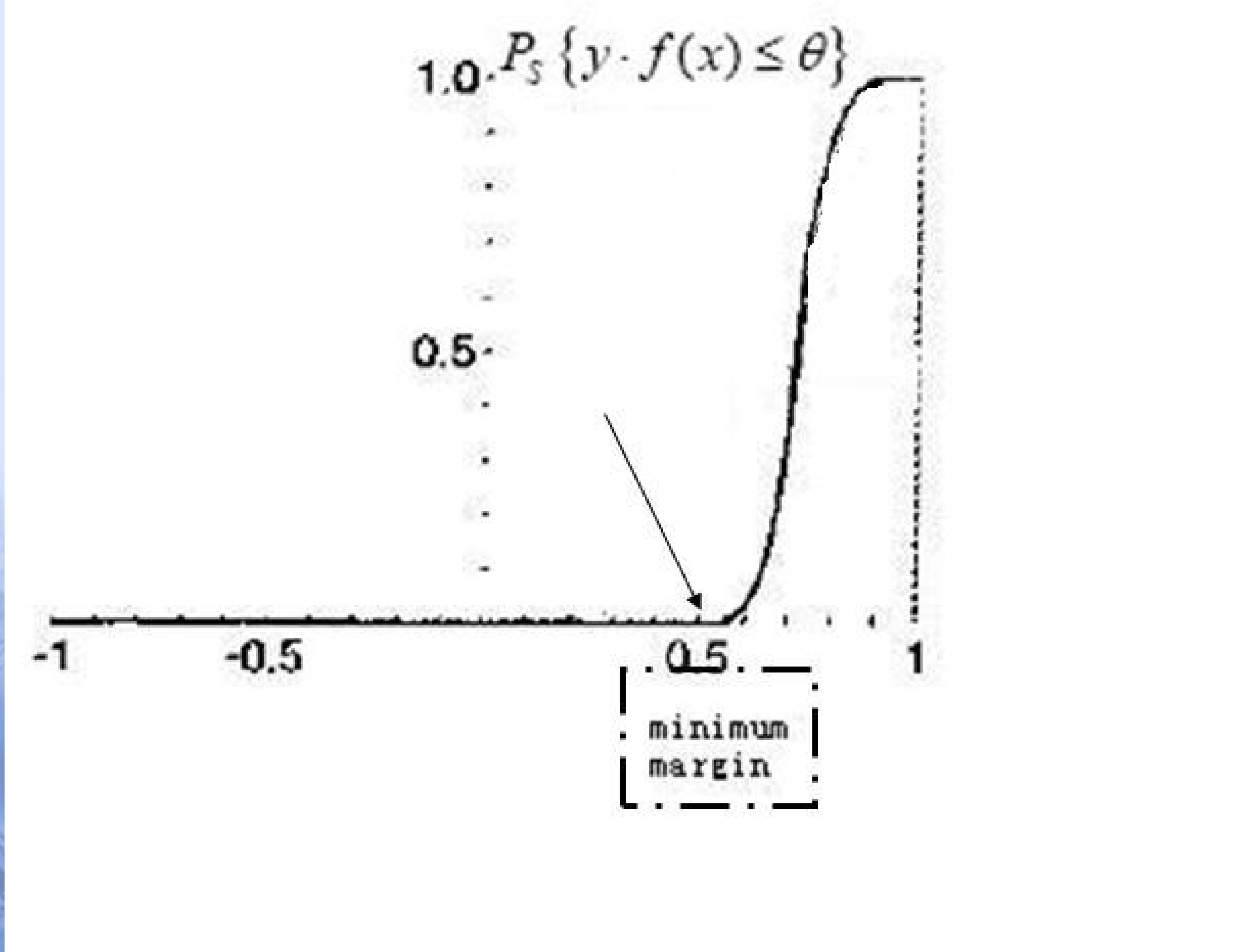
- Margins in Boosting:
  - The combined (voting) classifier produced by the ensemble learning algorithms could be written as:

$$f(x) = \sum \alpha_i h_i(x), \qquad \sum \alpha_i = 1, \qquad \alpha_i \geq 0.$$

The Concept of Margin

- For binary classification, $y \in \{-1, +1\}$. The quantity $yf(x)$ is called the **margin** of the example (x, y) with respect to the classifier f.
- Margin is a confidence measure (like in SVM).
- The **minimum margin** is the smallest margin over the set of training examples.

- Margin distribution:



Margin Theory

- Margin theory is essentially upper bounds on the generalization error of the voting classifier, in terms of various **margin** notions.

The Margin Distribution Bound

- Theorem 1 (Schapire et al. 1998):
  - For any $\delta > 0$, with probability at least $1-\delta$ over the random choice of the training set S of n examples, every voting classifier satisfies the following bound:

$$P_D\left(yf(x) \le 0\right) \le \inf_{\theta \in (0,1]} \left[P_S\left(yf(x) \le \theta\right) + O\left(\frac{1}{\sqrt{n}} \left(\frac{\log n \log |H|}{\theta^2} + \log \frac{1}{\delta} \right)^{1/2} \right) \right]$$

where H is the set which the base classifiers are chosen from.

Margin Explanation

- Schapire et al. also demonstrated theoretically and empirically that AdaBoost can generate good margin distribution.
- The margin distribution keeps improving even after the training error is zero. This accounts for AdaBoost's resistance to overfitting.

- Breiman's doubt and the Arc-gv algorithm:
  - Arc-gv provably generate the largest possible minimum margin among all boosting type algorithms.

Input: $S = (x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$
where $x_i \in X$, $y_i \in \{-1, 1\}$.
Initialization: $D_1(i) = 1/n$.
for $t = 1$ **to** T **do**

1. Train base learner using distribution D_t.
2. Get base classifier $h_t : X \to \{-1, 1\}$.
3. Choose α_t.
4. Update:

$$D_{t+1}(i) = \frac{D_t(i)\exp(-\alpha_t y_i h_t(x_i))}{Z_t},$$

where Z_t is a normalization factor chosen so that D_{t+1} will be a distribution.

end
Output: The final Classifier

$$H(x) = \operatorname{sgn}\left(\sum_{t=1}^{T} \alpha_t h_t(x)\right).$$

Breiman's Doubt

- The minimum margin bound (Breiman1999):

$$\forall_{\delta} \quad P_D\big(yf(x) \le 0\big) \le R\left(\log 2n + \log\frac{1}{R} + 1\right) + \frac{1}{n}\log\left(\frac{|H|}{\delta}\right).$$

where

$$R = \frac{32 \log 2 |H|}{n\theta_0^2}$$

and θ_0 is the minimum margin.

Breiman's Doubt

- Breiman's argument:
  - The minimum margin bound is sharper than the margin distribution bound.

$$O\left(\frac{\log n}{n}\right) \text{ vs. } O\left(\sqrt{\frac{\log n}{n}}\right)$$

If the bound of Schapire et al. implies that the margin distribution is the key to the generalization error, his bound implies more stronly that the minimum margin governs the generalization error.

Breiman’s Doubt

- Breiman conducted experiments, and found that arc-gv performs **consistently worse** than AdaBoost although it always generates larger minimum margins!
- Arc-gv even generates uniformly better margin distribution than AdaBoost.
- Breiman concluded that neither the margin distribution nor the minimum margin is the right explanation!

Recent Discovery

- An important discovery (Reyzin and Schapire 2006):
  - In the margin bounds, the generalization error depends not only on the margin, but also the complexity of the set of base classifiers.

$$P_D\left(yf(x) \le 0\right) \le \inf_{\theta \in (0,1]} \left[P_S\left(yf(x) \le \theta\right) + O\left(\frac{1}{\sqrt{n}} \left(\frac{\log n \log |H|}{\theta^2} + \log \frac{1}{\delta} \right)^{1/2} \right) \right]$$

  - To study how margin affects the generalization, one has to keep other factors fixed.

Recent Discovery

- Breiman’s experiment:
  - Base classifiers: Using decision trees of a fixed number of leaves.
- Reyzin and Schapire’s discovery:
  - Trees generated by arc-gv are much deeper than those generated by AdaBoost!
  - Deeper trees are more complex even though the number of leaves are the same!
  - Breiman’s experiment is not a fair comparison.

Recent Discovery

- A fair comparison:
  - Base classifier: decision stump.
  - Results:
    - AdaBoost has better performance than arc-gv.
    - Arc-gv has larger minimum margins than AdaBoost.
    - The margin distribution generated by AdaBoost is "better" than arc-gv.

Two Problems Left

- Has Breiman's doubt been fully answered?
  - Arc-gv generates larger minimum margin yet has worse performance. Contradict to the (sharper) minimum margin bound!
  - What does it mean a "better" margin distribution?

The EMargin Explanation

- Main results of EMargin :
  - A bound for the generalization error of voting classifiers in terms of a new margin notion——Equilibrium Margin (Emargin). This bound is uniformly sharper than the minimum margin bound.
  - We show that a large Emargin implies a smaller generalization error.

- Bernoulli Relative Entropy:

$$D(q \| p) = q \log \frac{q}{p} + (1-q) \log \frac{1-q}{1-p}, \qquad 0 \le p, q \le 1.$$

  - For fixed q , D is a monotone increasing function of p for $q \le p \le 1$.

- Inverse Relative Entropy Function:

$$D^{-1}(q,u): \quad D(q \| D^{-1}(q,u)) = u. \qquad u \ge 0.$$

- The Emargin Bound Theorem:

$$\forall_\delta \quad P_D\big(yf(x)\le 0\big)\le \frac{\log|H|}{n}+\min_{q\in\{0,\frac{1}{n},\frac{2}{n},\ldots,1\}} D^{-1}\big(q,u(\theta)\big).$$

where $$P_S\big(yf(x)<\theta\big)=q,$$

$$u(\theta)=\frac{1}{n}\left(\frac{8}{\theta^2}\log\left(\frac{2n^2}{\log|H|}\right)\log|H|+\log|H|+\log\frac{n}{\delta}\right).$$

Let q^* and θ^* be the optimal q, θ in the Emargin bound

$$\forall_\delta \quad P_D\big(yf(x)\le 0\big)\le \frac{\log|H|}{n}+D^{-1}\big(q^*,u(\theta^*)\big).$$

θ^* is referred to as Emargin.

- Explanation of Emargin bound:
  - The Emargin bound has a similar flavor to the margin distribution bound. The Emargin and Emargin error depend, in a complicated way on the whole margin distribution.
  - The minimum margin is only a special case when the optimal q^* is zero.

- Theorem:
  - The Emargin bound is uniformly sharper than the minimum margin bound.

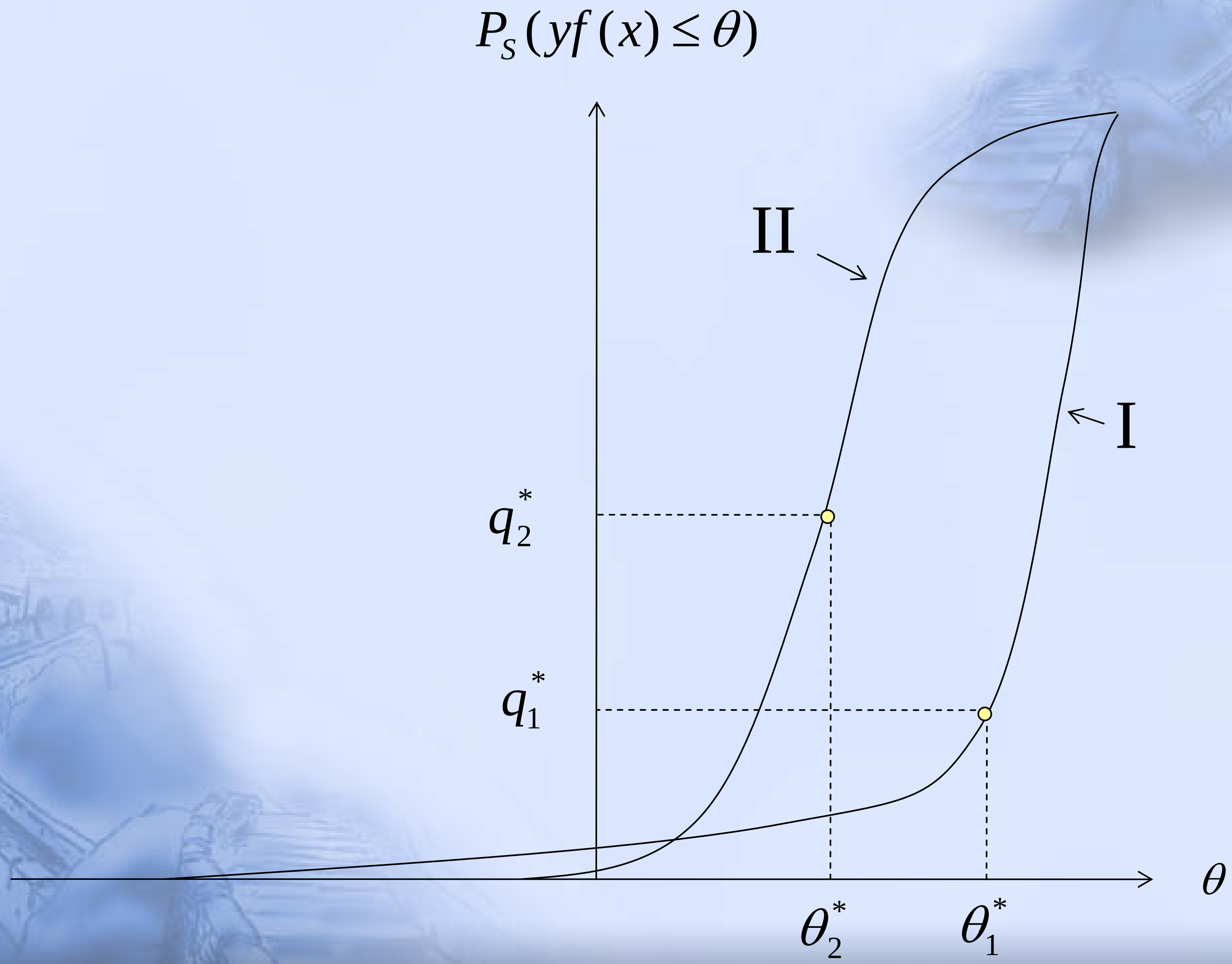
- Minimum margin is not crucial for the generalization error. Arc-gv does not necessarily have better performance than AdaBoost.

$$P_D\left(yf(x) \le 0\right) \le \frac{\log |H|}{n} + D^{-1}\left(q^*, u(\theta^*)\right).$$

- The Emargin bound implies that it is the Emargin and the Emargin error affect the performance of the classifier.

- How do Emargin and Emargin error affect the generalization error?

- The Comparison Theorem:
  - For two voting classifiers f_1, f_2 , if f_1 has a larger Emargin and a smaller Emargin error than f_2 , then the Emargin bound of f_1 is smaller than f_2.



$P_S(yf(x) \le \theta)$
II
I
q_2^*
q_1^*
θ
θ_2^*
θ_1^*

- Further explanation of the Emargin bound:
  - By using simple upper bounds of the inverse relative entropy function $D^{-1}(q,u)$, we can recover previous bounds and obtain new bound in simpler forms.

$$\inf_q D^{-1}\left(q,u\left(\hat{\theta}(q)\right)\right) \le D^{-1}\left(0,u\left(\hat{\theta}(0)\right)\right) \le u\left(\hat{\theta}(0)\right).$$ $\longrightarrow$ minimum margin bound

$$\inf_q D^{-1}\left(q,u\left(\hat{\theta}(q)\right)\right) \le \inf_q \left(q + \left(\frac{u\left(\hat{\theta}(q)\right)}{2} \right)^{1/2} \right)$$ $\longrightarrow$ margin distribution bound

$$\inf_q D^{-1}\left(q,u\left(\hat{\theta}(q)\right)\right) \le \inf_{q \le Cu\left(\hat{\theta}(q)\right)} D^{-1}\left(q,u\left(\hat{\theta}(q)\right)\right)$$
$$\le \inf_{q \le Cu\left(\hat{\theta}(q)\right)} C'u\left(\hat{\theta}(q)\right).$$

$\longrightarrow$ a new $O(\frac{\log n}{n})$ bound for the nonzero error case.

Experiments

- Setting:
  - UCI and USPS datasets.
  - Five-fold CV.
  - Binary classification.
  - Finite base classifiers.
  - Comparison of AdaBoost and Arc-gv on their EMargin, EMargin error, test error and minimum margin.

		Emargin	Emargin Error	Test Error	Minimum margin
Breast	AdaBoost	**0.313**	**0.803**	**0.052**	0.005
	arc-gv	0.281	0.909	0.057	0.008
Diabetes	AdaBoost	**0.110**	**0.748**	**0.255**	-0.064
	arc-gv	0.049	0.759	0.256	-0.017
German	AdaBoost	0.157	0.824	0.258	-0.118
	arc-gv	0.034	0.780	0.261	-0.026
Image	AdaBoost	**0.196**	**0.610**	0.023	-0.009
	arc-gv	0.195	0.705	**0.021**	-0.003
Ionosphere	AdaBoost	0.323	0.800	0.100	0.084
	arc-gv	0.131	0.577	0.106	0.061
Letter	AdaBoost	**0.078**	**0.645**	**0.174**	-0.165
	arc-gv	0.063	0.958	0.178	-0.034
Satimage	AdaBoost	**0.133**	**0.521**	**0.053**	-0.054
	arc-gv	0.133	0.956	0.057	-0.019
USPS	AdaBoost	**0.108**	**0.972**	**0.450**	-0.142
	arc-gv	0.053	0.990	0.460	-0.024
Vehicle	AdaBoost	**0.105**	**0.698**	**0.201**	-0.024
	arc-gv	0.063	0.720	0.205	-0.009
Wdbc	AdaBoost	**0.350**	**0.581**	**0.035**	-0.130
	arc-gv	0.350	0.710	**0.035**	-0.100

Experiments

- Conclusion from the experiments:
  - Usually AdaBoost has a larger EMargin and a smaller EMargin error than arc-gv. This accounts for AdaBoost’s superior performances.

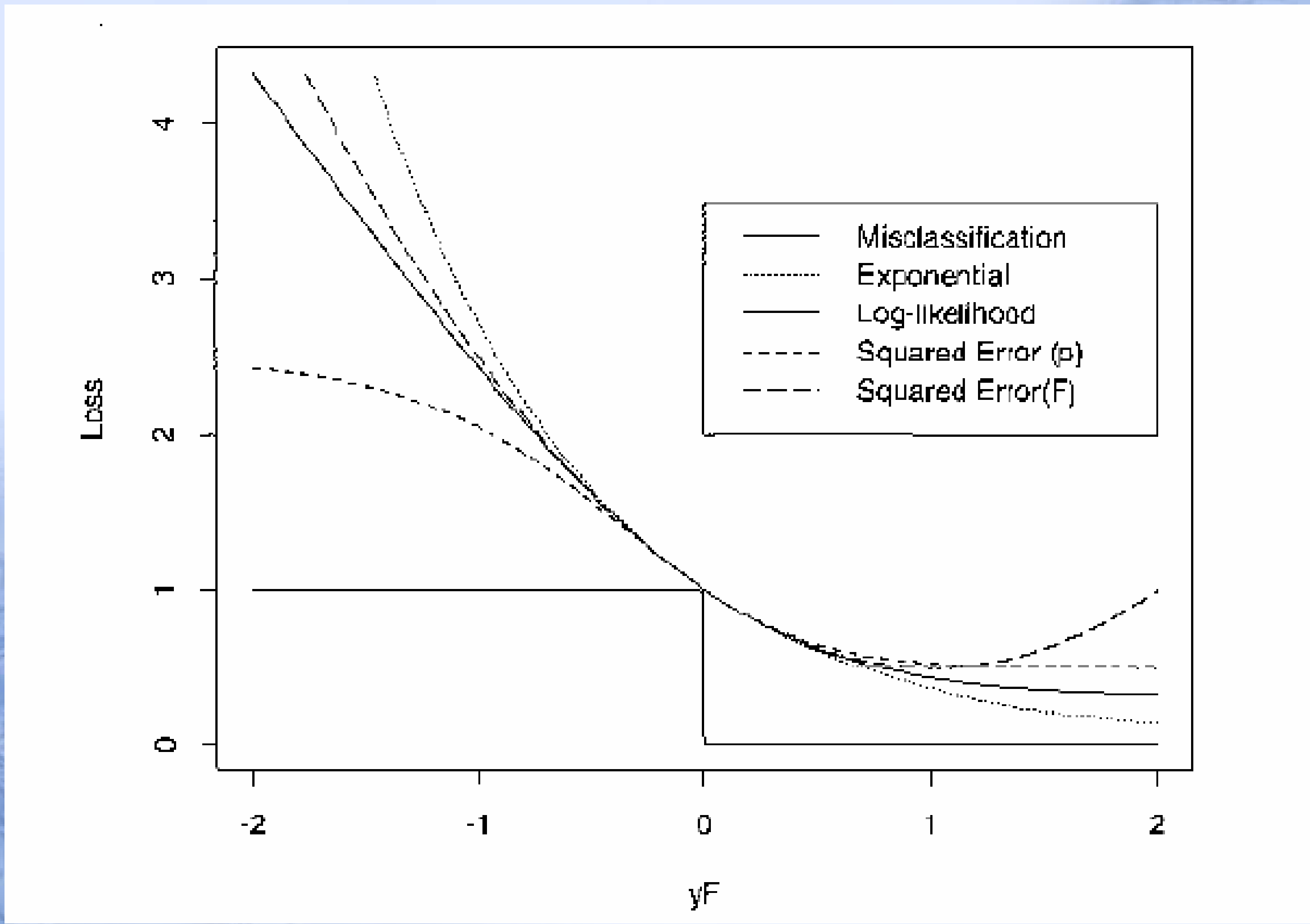
Convex Loss Minimization

Convex Loss Minimization

- Breiman discovered that AdaBoost was a down-the-gradient method for minimizing the exponential loss.

$$\text{exp loss} = \frac{1}{n}\sum_{i=1}^{n}\exp\left(-y_i f(x_i)\right) = \frac{1}{n}\sum_{i=1}^{n}\exp\left(-y_i \sum_{t=1}^{T}\alpha_t h_t(x_i)\right).$$

$$\text{0-1 loss} = \frac{1}{n}\sum_{i=1}^{n} I\left(-y_i f(x_i)\right) = \frac{1}{n}\sum_{i=1}^{n} I\left(-y_i \sum_{t=1}^{T}\alpha_t h_t(x_i)\right).$$


Loss
0
1
2
3
4
-2
-1
0
1
2
yF
Misclassification
Exponential
Log-likelihood
Squared Error (p)
Squared Error(F)

Convex Loss Minimization

- A natural question:
  - To what extent solving the approximated convex surrogate minimization is equivalent to minimizing the generalization error?

Convex Loss Minimization

- The statistical consequences of minimizing a surrogate:
  - Bayes Consistency:
    - Minimizing the convex loss (boosting) is NOT consistent.
    - With regularization, boosting is consistent:
      - Early stopping
      - L1 regularization
  - Rate of Convergence (dimension independent):

$$n^{-1/4} \sim n^{-1/2}$$

Convex Loss Minimization

- Large margin vs. convex loss minimization:
  - They are complementary explanations.
  - Convex loss minimization:
    - Asymptotic results, compare to the Bayes risk.
    - Depends on precise algorithms
  - Margin:
    - Nonasymptotic uniform bounds, gives confidence interval of the generalization error.
    - Algorithm independent.

Convex Loss Minimization

- Limitation of the convex loss minimization:
  - All based on an important assumption:
    - The linear span of the base classifiers is dense in the space of all measurable functions. Or at least the global minimizer of the convex loss is contained in the linear span.
  - If the base classifiers are decision stumps or other simple models, this assumption does not hold.

Convex Loss Minimization

- What is the consequence of boosting in the misspecified setting:
  - Bayes consistent is impossible.
  - Consistency to the best classifier in the model?
  - Empirically, boosting decision stumps often yields good performance.

Convex Loss Minimization

- A surprising result (Long & Servedio 2008):

  There are learning problems such that

  - Bayes error is slightly larger than zero.
  - The Bayes classifier is within the model class.
  - Minimizing the convex loss returns in a classifier whose performance is the same as random guess, even if there are infinitely many training examples!
  - Early stopping and L1 regularization do not help!

Convex Loss Minimization

- Convex loss minimization can not explain these results.
- Margin theory can predict the performance by giving the upper bound of the generalization error.

Summary and Future Work

Summary and Future Work

- The EMargin bound is an answer to Breiman's doubt of the margin explanation.
- Margin and convex loss minimization are complementary explanations of boosting.
- Is it possible to optimize the margin distribution (Emargin)?

Thanks