Random Classification Noise Defeats All Convex Potential Boosters

Phil Long
Google

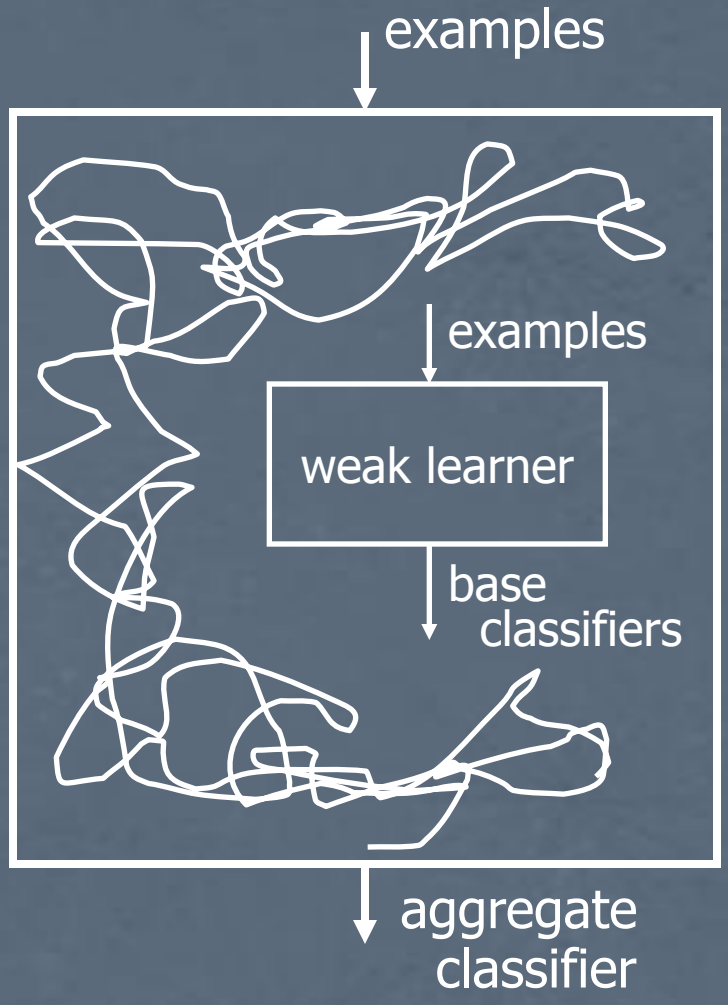
Rocco Servedio
Columbia

Boosting [Sch89,FS95]

Combine rough "base classifiers" into accurate aggregate classifier

Theoretical framework

- a target concept c: $X \rightarrow \{-1,1\}$
- random examples:
  - x drawn from distribution D
  - class designation c(x)
- a "weak learner"
  - gets examples:
    - distributions D' possibly different from D
    - classification still according to c
  - outputs base classifier $h: X \rightarrow \{-1,1\}$
  - advantage $E_{x \sim D'}[h(x)c(x)] \geq \gamma$
- goal: aggregate classifier error < ε



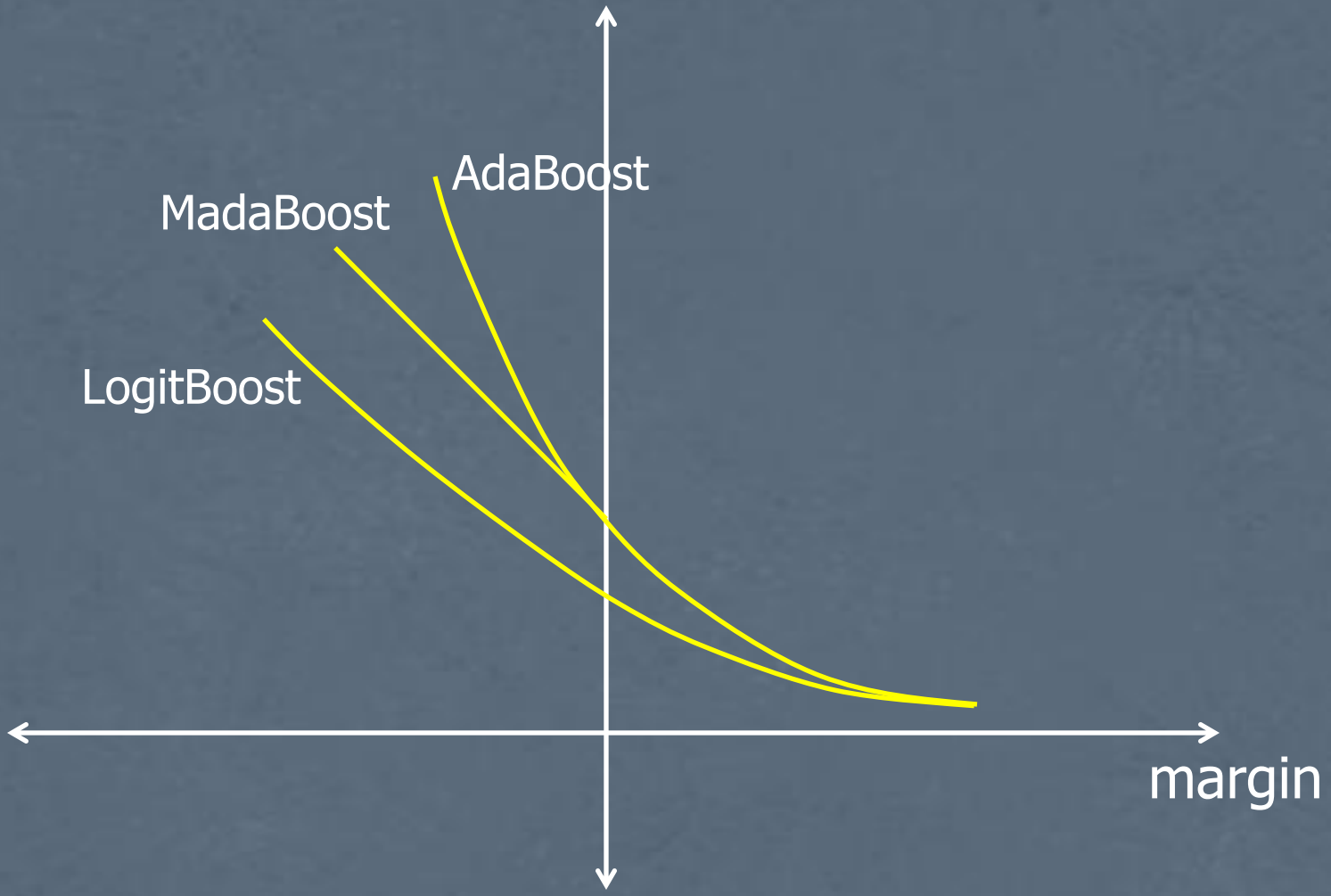
AdaBoost [FS95]

- Works by
  - repeatedly
    - assigning more weight where many base classifiers wrong
    - training new base classifier
  - combining results by weighted voting
- Practical success
- But sensitive to noise [MO97,Die00]

Convex potential boosters [MBBF99,DH99,CSS00]

- Construct classifiers that for
  - base classifiers $h_1,\ldots,h_n$
  - voting weights $w_1,\ldots,w_n$
  - predict $\mathrm{sign}(w_1 h_1(x) + \ldots + w_n h_n(x))$
- Do coordinate-wise gradient descent on
  - $E_{(x,y)\sim D}(\Phi(y\,(w_1 h_1(x) + \ldots + w_n h_n(x))))$
  - where $y\,(w_1 h_1(x) + \ldots + w_n h_n(x))$ is margin for (x, y), and
  - Convex $\Phi(\cdot)$ penalizes large-margin errors
- Examples: AdaBoost, MadaBoost, LogitBoost

Convex potential functions

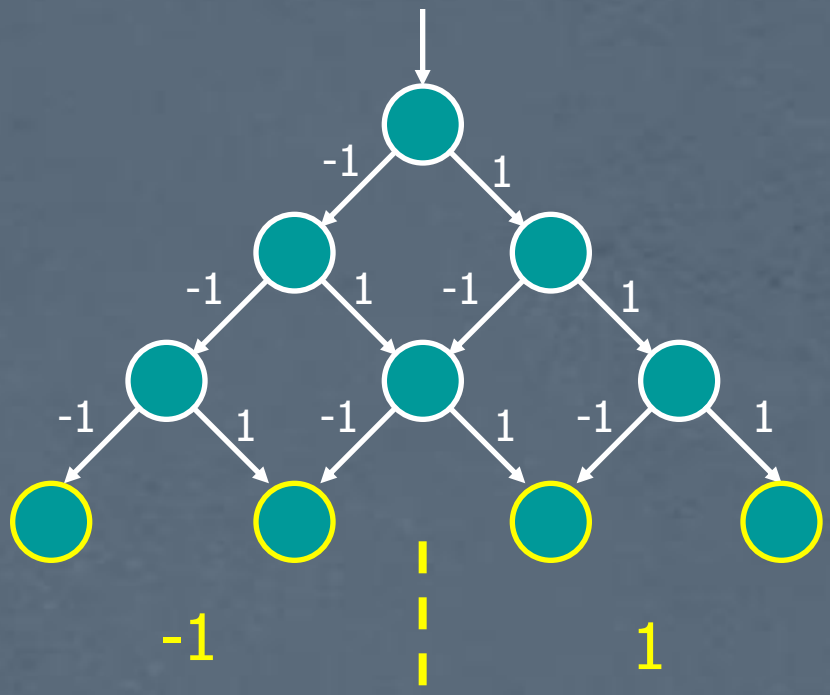


Boosting with noise – theorems

- **Definition [KS03]**: A **noise-tolerant weak learner:** $E[h(x)\, c(x)] > \gamma$ even when random classification noise rate η.
- **Theorem [KS03,LS05]**: Given (γ, η)-noise tolerant weak learner, if random classification noise rate η, can achieve error $\eta + \tau$ in poly$(1/\tau, 1/(1/2 - \eta), 1/\gamma)$ time.
- **Theorem [KS03]** (paraphrased): Achieving error better η than impossible.

Boosting with noise - algorithms

- Output branching programs (not weighted votes)
- Not potential boosters
- Question: can convex potential boosters tolerate noise?



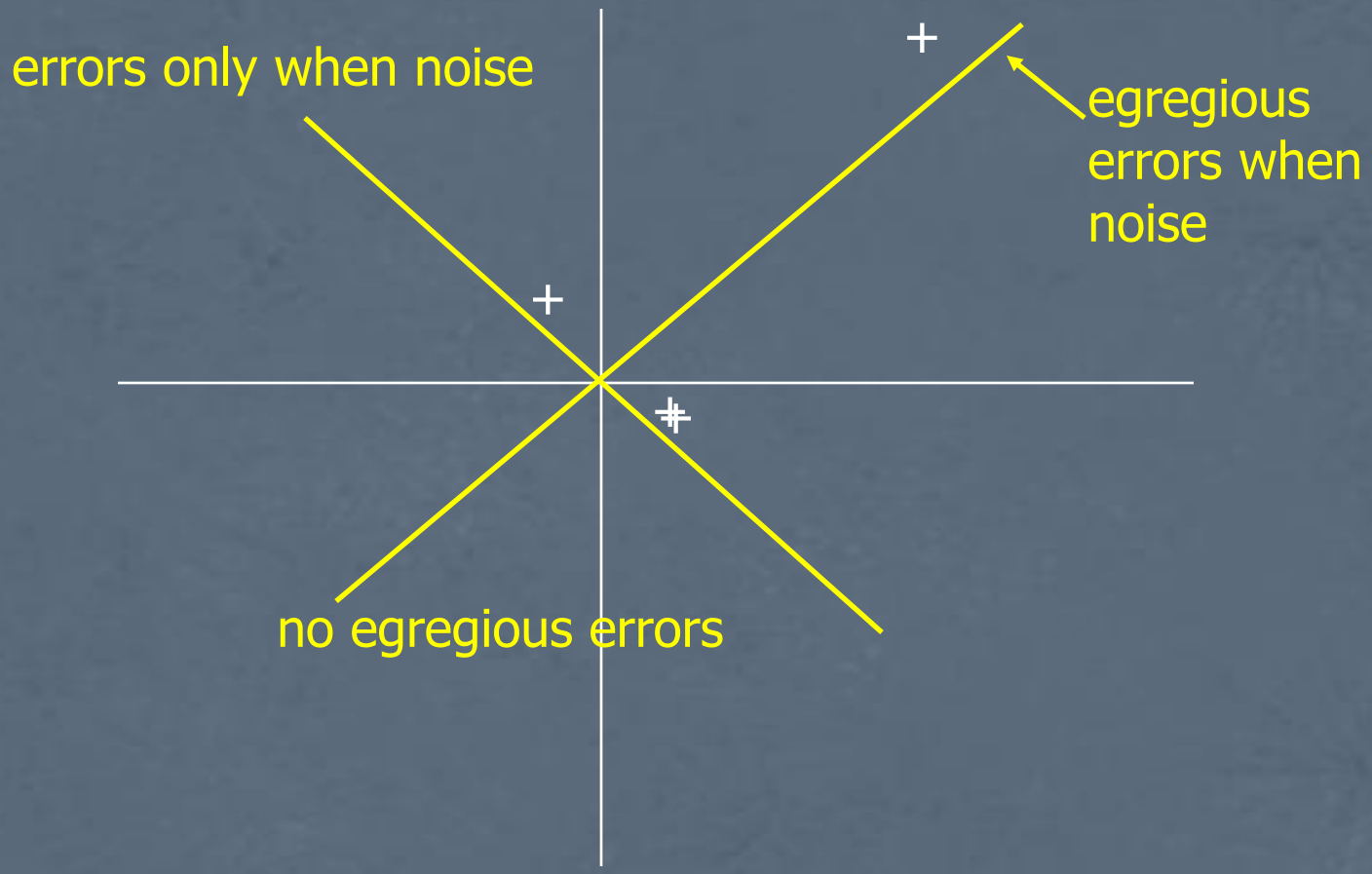
Main Result

- **Thm** (paraphrased): No "convex potential booster" can tolerate noise
- If
  - noise rate $\eta > 0$,
  - convex potential booster's error is $\frac{1}{2}$
  - (noise-tolerant booster's error near η)
- Conditions on Φ:
  - convex
  - nonincreasing
  - $\Phi'(0) < 0$
  - asymptotes to 0 as margin approaches infinity
  - has a continuous first derivative

Why convex potential boosters don't tolerate noise

- Convex potential boosters assign voting weights
- Potential means errors "bad" and "egregious"
- "Egregious": wrong, and vote not even close
- Construction: give algorithm choice between
  - many errors
  - relatively few "egregious" errors

Why convex potential boosters cannot tolerate noise



“Early Stopping” cannot help

- Can rotate construction so that optimum on a coordinate axis
- Convergence in one iteration

But isn't AdaBoost consistent? [BT06]

- Condition required for proof of consistency not satisfied by construction
- Condition (paraphrased): for minimizing potential $E[\Phi(y\ f(h_1(x),\ldots,h_n(x)))]$, some linear f as good as any f
- With linear constraint, can force bad vs. egregious choice
- If any f allowed, free to set confidence independent of position

Experiments with the binary case

- Examples: class label y and features $x_1,\ldots,x_n$
  - 1000 "large margin examples": $x_1 = \ldots = x_n = y$
  - 1000 "pullers":
    - $x_1 = \ldots = x_{11} = y$
    - $x_{12} = \ldots = x_{21} = -y$
  - 2000 "penalizers":
    - random 5 of first 11 features equal to y
    - random 6 of the last 10 features equal to y
- Vote over all 21 features
  - always correct
  - egregiously incorrect on noisy large-margin examples
- Vote over $x_1,\ldots,x_6,-x_7,\ldots,-x_{11}$
  - correct on large-margin examples, pullers
  - usually incorrect on penalizers
  - but (almost) never incorrect by a large margin

Experimental Results

With noise rate 10%, training error for

- AdaBoost is 33%,
- LogitBoost 30%,
- MadaBoost 27%

Future work

- Weaken conditions on Φ (hinge loss?)
- Work out corresponding analysis for halfspaces
  - poly-time noise-tolerant algorithms known [BFKV96,Coh97], but
  - don't minimize convex potential functions
- BrownBoost [Fre99]
  - "gives up" on some examples,
  - thus is "non-convex";
  - is something like it provably noise-tolerant?