Adaptive Hedge

Tim van Erven — VU University Amsterdam
Peter Grünwald — CWI Centrum Wiskunde & Informatica
Wouter M. Koolen — Royal Holloway University of London
Steven de Rooij — CWI Centrum Wiskunde & Informatica

Summary

AdaHedge is a new online learning algorithm that adapts to the difficulty of the data

	Difficulty	Regret
	Worst-case data	$O\left(\sqrt{L_T^* \ln(K)}\right)$
NEW	Easy data	constant: $O(K)$

Key Ideas

- Bounds on the *mixability gap* (see top-right panel) play a crucial role in previous analyses of the Hedge algorithm.
- We only bound the mixability gap in the analysis, but not in the algorithm!
- On easy data, the probabilities output by Hedge converge on a single action. In this case we improve the standard bounds.
- Example: if one action is always better than all others.

Online Learning Setting

Decision Theoretic Online Learning

In rounds $t = 1, \ldots, T$:

1. Assign probabilities $\boldsymbol{w}_t = (w_t^1, \ldots, w_t^K)$ to K actions
2. Actions get losses $\boldsymbol{\ell}_t \in [0,1]^K$
3. Our loss: $\boldsymbol{w}_t \cdot \boldsymbol{\ell}_t$

Aim to minimize the *regret*

$$R(T) = \sum_{t=1}^{T} \boldsymbol{w}_t \cdot \boldsymbol{\ell}_t - L_T^*,$$

where $L_T^* = \min_k \sum_{t=1}^T \ell_t^k$ is the loss of the best action in hindsight.

Mixability Gap

The *mixability gap* is

$$\delta_t(\eta) = \boldsymbol{w}_t \cdot \boldsymbol{\ell}_t - \Big(-\tfrac{1}{\eta} \ln(\boldsymbol{w}_t \cdot e^{-\eta \boldsymbol{\ell}_t}) \Big).$$

- In Prediction with Expert Advice terms: $\delta_t(\eta)$ measures the difference with a mixable loss function.
- In Bayesian terms: $\delta_t(\eta)$ measures the difference between randomizing according to the posterior and mixing according to the posterior.

Hedge

- Hedge predicts with exponential weights:

$$w_t^k \propto \exp\Big(-\eta \sum_{s=1}^{t-1} \ell_s^k \Big).$$

- Its performance depends strongly on the *learning rate* $\eta > 0$.

AdaHedge

- Tune η optimally for a budget $b(\eta)$ on the cumulative mixability gap $\Delta_T(\eta) = \sum_{t=1}^T \delta_t(\eta)$
- Increase the budget using the doubling trick.

Algorithm

1. Start with $\eta = 1$
2. Run a new instance of Hedge with learning rate η until $\Delta_T(\eta)$ exceeds budget
$$b(\eta) = \Big(\frac{1}{\eta} + \frac{1}{e-1}\Big) \ln(K).$$
3. Set $\eta \leftarrow \eta/2$ and goto 2.

Theoretical Results

AdaHedge is worst-case optimal...

Theorem 1 *The regret of AdaHedge is bounded by*

$$R(T) \le 5.1 \sqrt{L_T^* \ln(K)} + O\Big(\ln(L_T^* + 2) \ln(K) \Big).$$

...and has strong theoretical guarantees on 'easy' data

Theorem 2 *Suppose the loss vectors $\boldsymbol{\ell}_t$ are independent random variables and there exists a k^* such that*

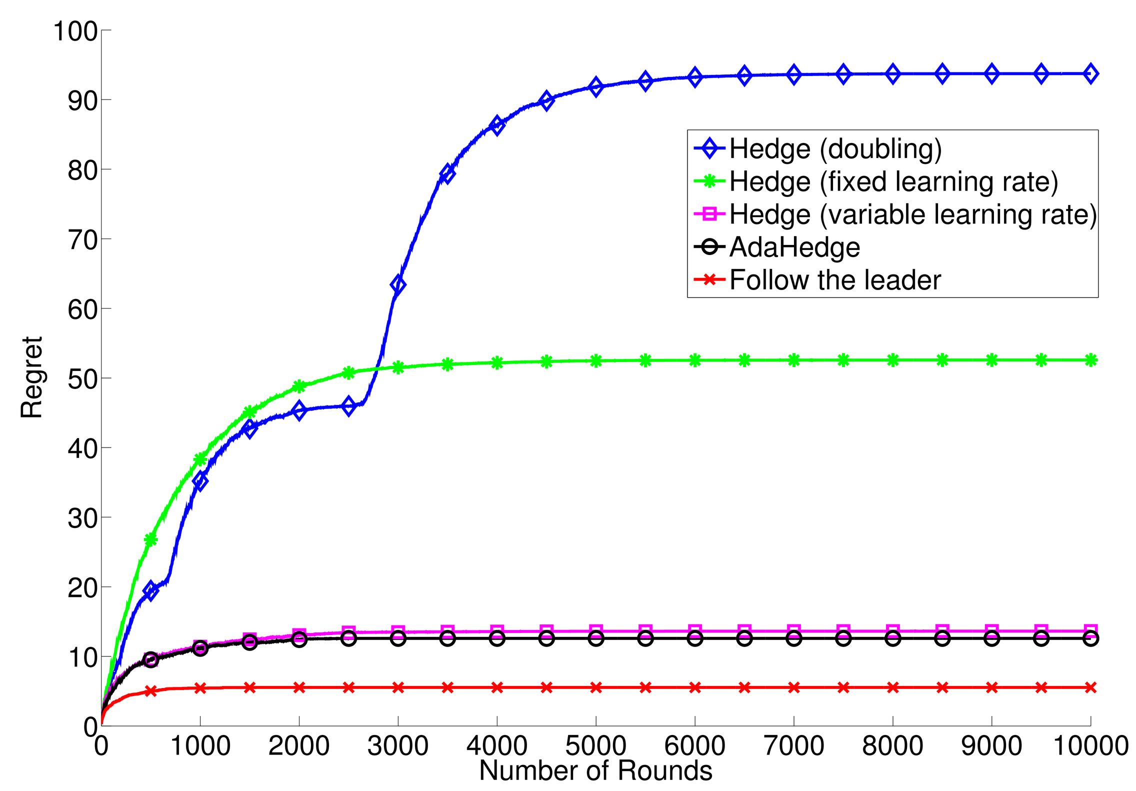
$$\min_{k \neq k^*} \mathbb{E}[\ell_t^k - \ell_t^{k^*}] > 0 \qquad \textit{for all } t \in \mathbb{Z}^+.$$

Then with probability at least $1-\delta$ the regret of AdaHedge is bounded by a constant:

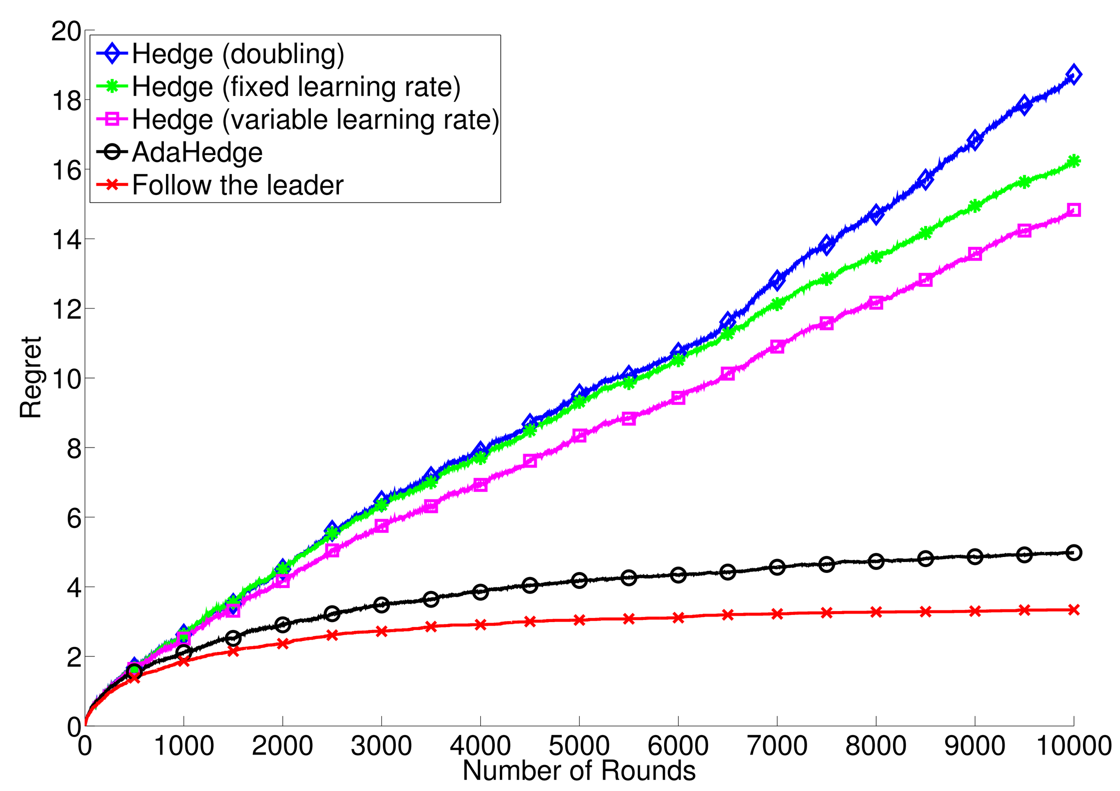
$$R(T) = O\Big(K + \log(1/\delta) \Big).$$

Experiments

Simulation Study on 'Easy' Data



I.I.D. losses



Correlated losses

AdaHedge has excellent practical performance

N.B. Follow-the-leader does very well here, but gets *linear* regret $\ge T/2 - 1$ in the worst case!

Proof Techniques

Everyone bounds the mixability gap δ_t.

Standard Analysis

- Optimize η after bounding $\delta_t(\eta) \le \eta/8$.

Our Approach

- Optimize η before bounding!
- If the posterior probabilities $\boldsymbol{w}_t$ converge on a single action, the mixability gap goes to 0!

$$\delta_t(\eta) \le (e-2)\eta\big(1 - \max_k w_t^k\big) \qquad (0 < \eta \le 1)$$

Current Work

Avoid the Doubling Trick

- Better performance in practice
- Still very clean analysis
- Improved the worst-case bound to

$$R(T) \le 2\sqrt{\frac{L_T^*(T - L_T^*)}{T} \ln(K)} + \tfrac{8}{3}\ln(K) + 2.$$

Weaker Conditions for Easy Data

- Guarantee regret bounded by the best *regret* of AdaHedge and Follow-the-Leader, up to a small constant factor.

Funding

PASCAL2 Network of Excellence grant IST-2007-216886 and NWO Rubicon grant 680-50-1010