CSIE5002 Prediction, learning, and games

Lecture 11: Second-order bound & adaptive regret

Yen-Huan Li (yenhuan.li@csie.ntu.edu.tw)

03.06.2019

Department of Computer Science and Information Engineering
National Taiwan University

Abstract

The standard definition of the regret is with respect to the worst-case data and the best fixed action/expert.

- In practice, we may not encounter the worst case and the regret can be smaller. Can we develop an algorithm whose regret is small with *easy data*?
- In some situations, we would like an online learning algorithm to compete with *time-varying* actions/experts. Can we develop such an algorithm?

Recommended reading

- S. de Rooij *et al.* Follow the leader if you can, hedge if you must. 2014.
- D. Adamskiy *et al.* A closer look at adaptive regret. 2016.
- N. Cesa-Bianchi *et al.* Improved second-order bounds for prediction with expert advice. 2007.
- E. Hazan & C. Seshadhri. Efficient learning algorithms for changing environments. 2009.

Table of contents

Preluede: Hedge with decreasing learning rates

Recap: Decision-theoretic online learning

Protocol. (DTOL) Let $T \in \mathbb{N}$. Let $\mathcal{A} = \{ 1, \ldots, K \}$ for some $K \in \mathbb{N}$. For every $1 \leq t \leq T$, the following happen sequentially.

- $\textsc{Learner}$ announces $\gamma_t \in \Delta(\mathcal{A})$.
- $\textsc{Reality}$ announces $\omega_t \in [0,1]^K$.
- $\textsc{Learner}$ suffers the loss

$$\lambda(\omega_t, \gamma_t) \coloneqq \langle \omega_t, \gamma_t \rangle .$$

Regret. The regret is given by

$$R_T(i) \coloneqq \sum_{t=1}^{T} \lambda(\omega_t, \gamma_t) - \sum_{t=1}^{T} \lambda(\omega_t, \gamma_t(i)), \quad \forall 1 \leq i \leq K.$$

Weak aggregating algorithm

Algorithm. (Weak aggregating algorithm) Let γ_1 be the uniform distribution on $\mathcal{A}$. For every $1 \leq t \leq T$, announce $\gamma_t \in \Delta(\mathcal{A})$ such that

$$\gamma_t(i) \propto \gamma_1(i) \mathrm{e}^{-\eta_t \sum_{\tau=1}^{t-1} \omega_\tau(i)}, \quad \forall 1 \leq i \leq K.$$

Theorem 1. The weak aggregating algorithm with

$$\eta_t = 2\sqrt{\frac{\log K}{t}}$$

achieves

$$R_T(i) := \sqrt{T \log K}.$$

Yu. Kalnishkan & M. V. Vyugin. The weak aggregating algorithm and weak mixability. 2008.

Proof of Theorem 1 (1/4)

Lemma 1. Define the mix loss

$$m(\omega, \gamma; \eta) \coloneqq \frac{-1}{\eta} \log \sum_{i=1}^{K} \gamma(i) \mathrm{e}^{-\eta \omega(i)}.$$

Then, the mix loss is non-increasing in η.

Proof. Let $\eta_1 > \eta_2 > 0$. Let $\xi = \omega(i)$ be a r.v., where i follows the probability distribution γ. By Jensen's inequality, we write

$$\begin{aligned} \frac{-1}{\eta_1} \log \mathsf{E}\, \mathrm{e}^{-\eta_1 \xi} &= \frac{-1}{\eta_1} \log \mathsf{E} \left[\left(\mathrm{e}^{-\eta_2 \xi} \right)^{\eta_1/\eta_2} \right] \\ &\leq \frac{-1}{\eta_1} \log \left[\left(\mathsf{E}\, \mathrm{e}^{-\eta_2 \xi} \right)^{\eta_1/\eta_2} \right] \\ &= \frac{-1}{\eta_2} \log \mathsf{E}\, \mathrm{e}^{-\eta_2 \xi}. \end{aligned}$$

Proof of Theorem 1 (2/4)

Define the cumulative mix loss of the standard hedge as

$$M_t^{(\eta)} := -\frac{1}{\eta} \log \sum_{i=1}^{K} \frac{1}{K} \mathrm{e}^{-\eta \sum_{\tau=1}^{t} \omega_\tau(i)}$$

Lemma 2. It holds that

$$M_T \coloneqq \sum_{t=1}^{T} m(\omega_t, \gamma_t; \eta_t) \leq M_T^{(\eta_T)}.$$

Proof. We write

$$\begin{aligned} M_T = \sum_{t=1}^{T} m(\omega_t, \gamma_t; \eta_t) &= \sum_{t=1}^{T} \left(M_t^{(\eta_t)} - M_{t-1}^{(\eta_t)} \right) \\ &\leq \sum_{t=1}^{T} \left(M_t^{(\eta_t)} - M_{t-1}^{\eta_{(t-1)}} \right) \leq M_T^{(\eta_T)}. \end{aligned}$$

S. de Rooij *et al.* Follow the leader if you can, hedge if you must. 2014.

Proof of Theorem 1 (3/4)

Proof. (Theorem 1) We write

$$\sum_{t=1}^{T} \lambda(\omega_t, \gamma_t) = M_T + \sum_{t=1}^{T} \delta_t,$$

where δ_t denote the mixability gaps

$$\delta_t \coloneqq \lambda(\omega_t, \gamma_t) - m(\omega_t, \gamma_t; \eta_t).$$

By Lemma 2 and Hoeffding's lemma, we write

$$\sum_{t=1}^{T} \lambda(\omega_t, \gamma_t) \leq \frac{-1}{\eta_T} \log \sum_{i=1}^{K} \frac{1}{K} \mathrm{e}^{-\eta_T \sum_{t=1}^{T} \omega_t(i)} + \frac{1}{8} \sum_{t=1}^{T} \eta_t.$$

Proof of Theorem 1 (4/4)

Proof continued. (Theorem 1) The first term on the RHS is bounded as

$$\frac{-1}{\eta_T} \log \sum_{i=1}^{K} \frac{1}{K} \mathrm{e}^{-\eta_T \sum_{t=1}^{T} \omega_t(i)} \leq \sum_{t=1}^{T} \omega_t(i) + \frac{1}{\eta_T} \log K.$$

The second term on the RHS is bounded as

$$\frac{1}{8} \sum_{t=1}^{T} \eta_t \leq \frac{1}{8} \int_0^T \frac{2\sqrt{\log K}}{\sqrt{t}} = \frac{\sqrt{T \log K}}{2}.$$

Summing up the two bounds leads to the theorem.

AdaHedge & second-order bounds

Possibility of a second-order bound (1/4)

Consider the hedge, which iterates as

$$\gamma_{t+1}(i) \propto \gamma_t(i)\mathrm{e}^{-\eta\omega_t(i)}, \quad \forall t \in \mathbb{N}, 1 \le i \le K.$$

Recall we have

$$R_T(i) \le \frac{1}{\eta}\log K + \sum_{t=1}^{T}\delta_t, \quad \forall 1 \le i \le K,$$

where δ_t denotes the mixability gap

$$\delta_t := \lambda(\omega_t, \gamma_t) + \frac{1}{\eta}\log\sum_{i=1}^{K}\pi_t(i)\mathrm{e}^{-\eta\lambda(\omega_t,\gamma_t(i))}, \quad \forall 1 \le t \le T.$$

Possibility of a second-order bound (2/4)

Previously, we bounded the mixability gap from above by Hoeffding's lemma. There are other choices.

Theorem 2. (Bennett's inequality) Let ξ be a random variable taking values in $]-\infty, 1]$. Then, we have

$$\log \mathsf{E}\, \mathrm{e}^{\eta(\xi - \mathsf{E}\,\xi)} \leq \mathsf{var}\,(\xi)\,(\mathrm{e}^{\eta} - \eta - 1)\,, \quad \forall \eta > 0.$$

S. Boucheron *et al. Concentration Inequalities: A Nonasymptotic Theory of Independence*. 2013.

Possibility of a second-order bound (3/4)

Lemma 3. The mixability gap is bounded as

$$\delta_t \leq \frac{\mathrm{e}^\eta - \eta - 1}{\eta} v_t, \quad 1 \leq t \leq T,$$

where v_t denotes the variance of $\omega_t(i)$ whose probability distribution is induced by π_t.

Proof. Define the r.v. $\xi_t := 1 - \omega_t(i)$, where i follows the probability distribution γ_t. Then, we have $\xi_t \leq 1$. By Bennett's inequality, we write

$$\log \mathsf{E}\, \mathrm{e}^{\eta(\xi_t - \mathsf{E}\,\xi_t)} = \eta\delta_t \leq \mathsf{var}\left(\xi_t\right) \left(\mathrm{e}^\eta - \eta - 1\right).$$

It remains to notice $\mathsf{var}(\xi_t) = v_t$.

Possibility of a second-order bound (4/4)

Proposition 1. Set

$$\eta = -W_{-1}\left(-\mathrm{e}^{-\frac{\log K}{V_t}-1}\right) - \frac{\log K}{V_t} - 1,$$

where W_{-1} denotes the lower branch of the Lambert W function. Define $V_T \coloneqq v_1 + \cdots + v_T$. If

$$V_T \geq \frac{\log K}{18\left(1-\beta\right)^2},$$

for some $\beta \in]0,1[$, then the hedge achieves

$$R_T \coloneqq \max_{1\leq i\leq K} R_T(i) \leq \frac{1}{\beta}\sqrt{2V_T \log K}, \quad \forall 1 \leq i \leq K.$$

R. M. Corless *et al.* On the Lambert W function. 1996.

Proof of Proposition 1 (1/2)

Lemma 4. It holds that

$$W_{-1}\left(-\mathrm{e}^{-x-1}\right) < -1 - \sqrt{2x} - \frac{2}{3}x, \quad \forall x > 0.$$

Proof. (Proposition 1) By Lemma 3, we write

$$R_T \leq \frac{1}{\eta}\log K + \sum_{t=1}^{T}\delta_t \leq \frac{1}{\eta}\log K + \frac{\mathrm{e}^{\eta} - \eta - 1}{\eta}V_t.$$

Optimizing over the learning rate, we obtain the expression for the optimal η.

I. Chatzigeorgiou. Bounds on the Lambert function and their application to the outage analysis of user cooperation. 2013.

Proof of Proposition 1 (2/2)

Proof continued. (Proposition 1) Define $u := (\log K)/V_T$. By assumption, we have

$$\sqrt{2u} - \frac{u}{3} \geq \frac{\sqrt{2u}}{\alpha},$$

where $\alpha := 1/\beta$. By Lemma 4, we then write

$$R_T \leq \frac{2\log K}{\sqrt{2u} - \frac{u}{3}} \leq \frac{2\alpha \log K}{\sqrt{2u}},$$

which completes the proof.

Remark. The difficulty lies in *turning the learning rate η*.

AdaHedge

Algorithm. (AdaHedge) Let γ_1 be the uniform distribution on $\{ 1, \ldots, K \}$. For every $2 \leq t \leq T$, define

$$\eta_t := \frac{\log K}{\sum_{\tau=1}^{t-1} \delta_\tau},$$

and announce γ_t such that

$$\gamma_t(i) \propto \gamma_1(i) \mathrm{e}^{-\eta_t \sum_{\tau=1}^{t-1} \omega_\tau(i)}, \quad \forall 1 \leq i \leq K.$$

Theorem 3. AdaHedge achieves

$$R_T := \max_{1 \leq i \leq K} R_T(i) \leq 2\sqrt{V_T \log K} + \frac{4}{3} \log K + 2.$$

---

de Rooij *et al.* Follow the leader if you can, hedge if you must. 2014.

Proof of Theorem 3 (1/3)

Proof. (Theorem 3) Define the mix and cumulative mix losses

$$m(\omega, \gamma; \eta) := \frac{-1}{\eta} \log \sum_{i=1}^{K} \gamma(i) \mathrm{e}^{-\eta \omega(i)}, \quad M_t := \sum_{\tau=1}^{t} m(\omega_\tau, \gamma_\tau; \eta_\tau).$$

Define the mixability and cumulative mixability gaps

$$\delta_t := \lambda(\omega_t, \gamma_t) - m_t(\omega_t, \gamma_t; \eta_t), \quad \Delta_t := \sum_{\tau=1}^{t} \delta_\tau.$$

Then, we have

$$R_T = (M_T - L_T^\star) + \Delta_T,$$

where $L_T^\star$ is the cumulative loss of the best action/expert.

Proof of Theorem 3 (2/3)

Proof continued. (Theorem 3) By Lemma 2 and the definition of AdaHedge, we have

$$M_T \leq M_T^{(\eta_T)} \leq L_T^\star + \frac{\log K}{\eta_T} = L_T^\star + \Delta_{T-1}.$$

Then, we obtain

$$R_T \leq \Delta_{T-1} + \Delta_T \leq 2\Delta_T.$$

Applying Bennett's inequality, we write

$$R_T \leq 2 \sum_{t=1}^{T} \frac{\mathrm{e}^{\eta_t} - \eta_t - 1}{\eta_t} v_t,$$

where v_t denotes the variance of $\omega_t(i)$ following the probability distribution γ_t.

Proof of Theorem 3 (3/3)

Proof continued. (Theorem 3) The theorem follows from the following inequality. See de Rooij *et al.*, 2014 for the details.

Lemma 5. It holds that

$$\Delta_T^2 \leq V_T \log K + \left(1 + \frac{2}{3} \log K\right) \Delta_T.$$

Remark. Notice this lemma is not general but specific to AdaHedge.

S. de Rooij *et al.* Follow the leader if you can, hedge if you must. 2014.

Prelude: Specialist AA

Learning with specialist experts

Protocol. (Learning with specialist experts) Let $T \in \mathbb{N}$. For every $1 \le t \le T$, the following happen sequentially.

1. Reality announces $\mathcal{A}_t \subseteq \mathcal{A}$.
2. Learner announces $\gamma_t \in \Delta(\mathcal{A})$.
3. Reality announces $\omega_t \in]-\infty, +\infty]^K$.
4. Learner suffers the *mix loss*

$$\lambda(\omega_t, \gamma_t) \coloneqq -\log \sum_{i \in \mathcal{A}} \gamma_t(i) \mathrm{e}^{-\omega_t(i)}.$$

A. Blum. Empirical support for Winnow and weighted-majority algorithms: Results on a calendar scheduling domain. 1997.

Y. Freund *et al.* Using and combining predictors that specialize. 1997.

D. Adamskiy *et al.* A closer look at adaptive regret. 2016.

Algorithm. (Specialist AA) Let $w_1 = (1, \dots, 1) \in \mathbb{R}^K$ be the vector of initial weights. For every $1 \le t \le T$, announce γ_t such that

$$\gamma_t(i) = \begin{cases} \frac{w_t(i)}{\sum_{i \in \mathcal{A}_t} w_t(i)} & , \forall i \in \mathcal{A}_t, \\ 0 & \text{, otherwise,} \end{cases}$$

and compute w_{t+1} such that

$$w_{t+1}(i) = \begin{cases} w_t(i) \mathrm{e}^{\lambda(\omega_t, \gamma_t) - \omega_t(i)} & , \forall i \in \mathcal{A}_t, \\ w_t(i) & \text{, otherwise.} \end{cases}$$

A. Chernov & V. Vovk. Prediction with expert evaluator's advice. 2009.

Regret of the specialist AA

Proposition 2. The specialist AA achieves

$$\sum_{1 \le t \le T, \mathcal{A}_t \ni i} \lambda(\omega_t, \gamma_t) - \sum_{1 \le t \le T, \mathcal{A}_t \ni i} \omega_t(i) \le \log K, \quad \forall i \in \mathcal{A}.$$

Proof. The specialist AA is simply AA with the loss vectors $\tilde{\omega}_t$ defined as

$$\tilde{\omega}_t(i) := \begin{cases} \omega_t(i) & , \forall i \in \mathcal{A}_t, \\ \lambda(\omega_t, \gamma_t) & , \text{otherwise.} \end{cases}$$

A. Chernov & V. Vovk. Prediction with expert evaluator's advice. 2009.

Fixed Share & adaptive regret

Adaptive regret

Recall the regret bounds for DTOL we have seen are for the standard definition:

$$R_T := \sum_{t=1}^{T} \lambda(\omega_t, \gamma_t) - \min_{1 \leq i \leq K} \sum_{t=1}^{T} \omega_T(i).$$

Definition. (Adaptive regret) The *adaptive regret* is defined as

$$R_{t_1,t_2} := \sum_{t=t_1}^{t_2} \lambda(\omega_t, \gamma_t) - \min_{1 \leq i \leq K} \sum_{t=t_1}^{t_2} \omega_t(i).$$

N. Littlestone & M. K. Warmuth. The weighted majority algorithm. 1994.

E. Hazan & C. Seshadhri. Efficient learning algorithms for changing environments. 2009.

Possibility of an adaptive regret bound (1/2)

Let $\mathcal{A}$ be a given set of experts. For each $1 \le t \le T$ and $i \in \mathcal{A}$, we consider a virtual expert who sleeps during the first $(t-1)$ rounds, and predicts as $i \in \mathcal{A}$ afterwards. Then, the specialist AA achieves, for every $1 \le t \le T$,

$$\sum_{\tau=t}^{T} \lambda(\omega_t, \gamma_t) - \min_{i \in \mathcal{A}} \sum_{\tau=t}^{T} \omega_t(i) \le \log K + \log T.$$

Question. Is it possible to compete with the best expert w.r.t. any interval $[t_1, t_2]$?

Remark. Notice we will consider the *mix loss* game afterwards.

D. Adamskiy *et al.* A closer look at adaptive regret. 2016.

Possibility of an adaptive regret bound (2/2)

The specialist AA can be expressed in the following equivalent form.

Algorithm. Define

$$\alpha_t = \frac{K-1}{Kt}, \quad \forall t \in \mathbb{N}.$$

Let γ_1 be the uniform distribution on $\mathcal{A}$. For every $1 \leq t \leq T-1$, announce γ_{t+1} given by

$$\gamma_{t+1}(i) = \frac{\alpha_{t+1}}{K-1} + \left(1 - \frac{K}{K-1}\alpha_{t+1}\right) \frac{\gamma_t(i)\mathrm{e}^{-\omega_t(i)}}{\sum_{i \in \mathcal{A}} \gamma_t(i)\mathrm{e}^{-\omega_t(i)}}, \quad \forall i \in \mathcal{A}.$$

D. Adamskiy *et al.* A closer look at adaptive regret. 2016.

Fixed Share

Algorithm. (Fixed Share) Let $\alpha_1, \ldots, \alpha_T$ be the *switching rates* in $[0, 1]$. Let γ_1 be the uniform distribution on $\mathcal{A}$. For every $1 \leq t \leq T - 1$, announce γ_{t+1} given by

$$\gamma_{t+1}(i) = \frac{\alpha_{t+1}}{K-1} + \left(1 - \frac{K}{K-1}\alpha_{t+1}\right) \frac{\gamma_t(i)\mathrm{e}^{-\omega_t(i)}}{\sum_{i \in \mathcal{A}} \gamma_t(i)\mathrm{e}^{-\omega_t(i)}}, \quad \forall i \in \mathcal{A}.$$

Theorem 4. The worst-case adaptive regret of Fixed Share with switching rates $\alpha_t \in [0, (K-1)/K]$ achieves

$$\max_{(\omega_t)_{t \in \mathbb{N}}} R_{t_1, t_2} = -\log \left[\frac{\alpha_{t_1}}{K-1} \prod_{t=t_1+1}^{t_2} (1 - \alpha_t) \right].$$

M. Herbster & M. K. Warmuth. Tracking the best expert. 1998.

D. Adamskiy *et al.* A closer look at adaptive regret. 2016.

Proof of Theorem 4

Proof sketch. (Theorem 4) The proof consists of three lemmas. Suppose $t_1 > 1$. Suppose the best expert for the time interval $[t_1, t_2]$ is $i \in \mathcal{A}$. Consider the adaptive regret R_{t_1,t_2}.

1. The adaptive regret is maximized when $\omega_{t_2}(j) = +\infty$ for all $j \neq i$.
2. Let $t_1 \leq t \leq t_2$. Suppose $\omega_\tau(j) = +\infty$ for all $t < \tau \leq t_2$ and $j \neq i$. Then, the adaptive regret is maximized when $\omega_t(j) = +\infty$ for all $j \neq i$.
3. Suppose that $\omega_t(j) = +\infty$ for all $t_1 \leq t \leq t_2$ and $j \neq i$. Then, the adaptive regret is maximized when $\omega_{t_1-1}(i) = +\infty$.

D. Adamskiy *et al.* A closer look at adaptive regret. 2016.

Regret of Fixed Share

Corollary 1. Fixed Share with switching rates $\alpha_t = 1/t$ (except for $\alpha_1 = (K-1)/K$) achieves

$$\max_{(\omega_t)_{t\in\mathbb{N}}} R_{t_1,t_2} = \begin{cases} \log(K-1) + \log t_2 & \text{, if } t_1 > 1, \\ \log K + \log t_2 & \text{, if } t_1 = 1. \end{cases}$$

Corollary 2. Let $(i_t)_{1\le t\le T} \in \mathcal{A}^T$ be a *piecewise-constant* sequence of m blocks. Then, Fixed Share with the switching rates above achieves the *shifting regret* bound:

$$\sum_{t=1}^{T} \lambda(\omega_t, \gamma_t) - \sum_{t=1}^{T} \omega_t(i_t) \le \log K + (m-1)\log(K-1) + m \log T.$$

D. Adamskiy *et al.* A closer look at adaptive regret. 2016.

Conclusions

Conclusions

- The decreasing-learning-rate case can be handled by a cumulative mix loss bound.
- The key in deriving a second-order regret bound is to use a second-order mixability gap bound, e.g., Bennett's inequality.
- Fixed Share can be viewed as a variant of specialist AA.
- The adaptive regret analysis of fixed share is obtained by directly examining the worst-case data.