CSIE5002 Prediction, learning, and games

Lecture 6: Multiplicative weight update II

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

01.04.2019

Department of Computer Science and Information Engineering
National Taiwan University

Abstract

- What is the optimal regret value achievable in the individual sequence prediction problem?
- Can the optimal value be achieved?
- Can the optimal value be achieved efficiently?

This lecture addresses the questions above.

Recommended reading

- N. Cesa-Bianchi and G. Lugosi. 2006. *Prediction, Learning, and Games.* Chapter 9.

- Y. Freund. 2003. Predicting a binary sequence almost as well as the optimal biased coin.

- E. Takimoto and M. K. Warmuth. 2000. The last-step minimax algorithm.

- F. Hedayati and P. Bartlett. 2017. Exchangeability characterizes optimality of sequential normalized maximum likelihood and Bayesian prediction.

Table of contents

Normalized maximum likelihood

Individual sequence prediction problem

Let $\mathcal{A}$ be a finite alphabet. Let $\mathcal{H}$ be a class of hypotheses $h : \mathcal{A}^* \to \Delta \subset \mathbb{R}^{|\mathcal{A}|}$. We do not require $\mathcal{H}$ to be countable.

Protocol. (Individual sequence prediction) For $t = 1, \ldots, T$, the following happen in order.

1. Learner announces $\gamma_t = (\gamma_t(a))_{a \in \mathcal{A}} \in \Delta$.
2. Reality announces $\omega_t \in \mathcal{A}$.
3. Learner observes the loss

$$\lambda(\omega, \gamma) := -\log \gamma_t(\omega_t), \quad \forall (\omega, \gamma) \in \{1, \ldots, m\} \times \Delta.$$

Minimax regret

Notice there is an equivalence relation between a sequence $(\gamma_t)_{1\leq t\leq T}$ and a probability distribution $\hat{p}$ on $\mathcal{A}^T$, with

$$\gamma_t(a) = \frac{\hat{p}(\omega_{1:t-1}a)}{\hat{p}(\omega_{1:t-1})}, \quad \forall a \in \mathcal{A},$$
$$\hat{p}(a_{1:T}) = \Pi_{t=1}^{T}\gamma_t(a_t), \quad \forall a_{1:T} \in \mathcal{A}^T.$$

Definition. (Minimax regret) The *minimax regret* is defined as

$$R_T^\star := \min_{\hat{p}} \max_{a_{1:T}} \left\{ -\log \hat{p}(a_{1:T}) - \inf_{h\in\mathcal{H}} \sum_{t=1}^{T} [-\log h(a_t|a_{1:t-1})] \right\},$$

where $a_{1:0}$ corresponds to the empty string.

Normalized maximum likelihood

Definition. (Equalizer) We call a probability distribution on $\mathcal{A}^T$ an *equalizer*, if and only if it induces the same regret value on every $a_{1:T} \in \mathcal{A}^T$.

Theorem 1. (Shtarkov) Every equalizer on $\mathcal{A}^T$ achieves the minimax regret $R_T^\star$, and equals the *normalized maximum-likelihood (NML) distribution*

$$p^\star(a_{1:T}) = \frac{\sup_{h \in \mathcal{H}} \Pi_{t=1}^T h(a_t|a_{1:t-1})}{\sum_{a_{1:T} \in \mathcal{A}^T} \sup_{h \in \mathcal{H}} \Pi_{t=1}^T h(a_t|a_{1:t-1})}, \quad \forall a_{1:T} \in \mathcal{A}^T.$$

Yu. M. Shtarkov. 1987. Universal sequential coding of single messages.
F. Hedayati and P. Bartlett. 2012. Exchangeability characterizes optimality of sequential normalized maximum likelihood and Bayesian prediction with Jeffreys prior.

Proof of Theorem 1 (1/2)

For every probability distribution p on $\mathcal{A}^T$ and $a_{1:T} \in \mathcal{A}^T$, define

$$R(p; a_{1:T}) := -\log p(a_{1:T}) - \inf_{h \in \mathcal{H}} \sum_{t=1}^{T} \left[-\log h(a_t | a_{1:t-1}) \right].$$

Lemma 1. The NML distribution $p^\star$ is the unique equalizer.

Proof. A direct calculation of $R(p^\star; a_{1:T})$ shows that $p^\star$ is an equalizer. For any $q \neq p^\star$, there must exist some $a_{1:T}$ and $a'_{1:T}$ in $\mathcal{A}^T$ such that

$$q(a'_{1:T}) < p^\star(a'_{1:T}), \quad q(a_{1:T}) > p^\star(a_{1:T}),$$

showing that an equalizer is uniquely defined.

Proof of Theorem 1 (2/2)

Proof. (Theorem 1) Let $q \neq p^\star$ be a distribution on $\mathcal{A}^T$. Then, there must exist some $a'_{1:T} \in \mathcal{A}^T$, such that

$$p^\star(a'_{1:T}) > q(a'_{1:T}).$$

Then, we write

$$\begin{aligned}\max_{a_{1:T} \in \mathcal{A}^T} R(q; a_{1:T}) &\geq R(q; a'_{1:T}) \\ &> R(p^\star; a'_{1:T}) = \max_{a_{1:T} \in \mathcal{A}^T} R(p^\star; a_{1:T}).\end{aligned}$$

The above proves optimality of $p^\star$ and sub-optimality of q.

F. Hedayati and P. Bartlett. 2012. Exchangeability characterizes optimality of sequential normalized maximum likelihood and Bayesian prediction with Jeffreys prior.

Implications of Theorem 1

Corollary 1. The minimax regret is given *exactly* by

$$R_T^\star = \log \sum_{a_{1:T} \in \mathcal{A}^T} \sup_{h \in \mathcal{H}} \left[\prod_{t=1}^{T} h(a_t | a_{1:t-1}) \right].$$

Remark. It is rare that we can get an (almost) closed-form expression of the minimax regret.

Universal coding. For universal coding, Theorem 1 gives the optimal coding scheme. However, the computational complexity of the optimal scheme can be high. (Why?) Therefore, we use an *online algorithm*, e.g., the mixture forecaster, for universal coding.

Predicting a binary sequence with static hypotheses

Minimax regret

Theorem 2. Consider the case where $\mathcal{A} = \{ 0, 1 \}$. Consider the class $\mathcal{H} = \{ h_q \mid q \in [0, 1] \}$ of *static hypotheses*

$$h_q(a_{1:t}) \coloneqq (h_q(0), h_q(1)) \coloneqq (1 - q, q), \quad \forall q \in [0, 1].$$

Then, we have

$$R_T^\star = \frac{1}{2} \log(T + 1) + \frac{1}{2} \log \frac{\pi}{2} - O\left(\frac{1}{\sqrt{T}} \right).$$

Remark. A static hypothesis resembles a sequence of i.i.d. r.v.'s.

Yu. M. Shtarkov. 1987. Universal sequential coding of single messages.
Y. Freund. 1996. Predicting a binary sequence almost as well as the optimal biased coin.

Proof of Theorem 2

Sketch of proof. (Theorem 2) Denote the number of 0's in $a_{1:T}$ by t_0, and that of 1's by t_1. Then, we write

$$R_T^\star = \log \sum_{a_{1:T} \in \mathcal{A}^T} \left[\sup_{q \in [0,1]} (1-q)^{t_0} q^{t_1} \right].$$

It is easily checked that

$$\sup_{q \in [0,1]} (1-q)^{t_0} q^{t_1} = \left(\frac{t_0}{T} \right)^{t_0} \left(\frac{t_1}{T} \right)^{t_1}.$$

By Corollary 1, we write

$$R_T^\star = \log \sum_{t_0=0}^{T} \binom{T}{t_0} \left(\frac{t_0}{T} \right)^{t_0} \left(\frac{t_1}{T} \right)^{t_1}.$$

Y. Freund. 2003. Predicting a binary sequence almost as well as the optimal biased coin.

Lapalce mixture (1/3)

Algorithm. (Laplace mixture) The *Laplace mixture* forecaster is the mixture forecaster with π being the uniform distribution on $[0,1]$ (parameter space for q); that is,

$$\hat{p}_{\mathsf{L}}(a_{1:T}) = \int_0^1 (1-q)^{t_0} q^{t_1} \,\mathrm{d}q, \quad \forall a_{1:T} \in \mathcal{A}^T,$$

where n_0 and n_1 denote the numbers of 0's and 1's in $a_{1:T}$, respectively.

L. D. Davisson. 1973. Universal noiseless coding.

Laplace mixture (2/3)

Proposition 1. The Laplace mixture forecaster has the closed-form expression

$$\hat{p}_{\mathsf{L}}(a_{1:T}) = \frac{1}{(T+1)\binom{T}{t_1}}.$$

Proof. Direct calculation.

Remark. The calculation is not trivial; see the textbook by Cesa-Bianchi & Lugosi for the details.

L. D. Davisson. 1973. Universal noiseless coding.

Laplace mixture (3/3)

Proposition 2. (Laplace's *rule of succession*) The Laplace mixture forecaster has the closed-form expression

$$\hat{p}_{\mathsf{L}}(0|a_{1:t-1}) = \frac{t_0(a_{1:t-1}) + 1}{t+1},$$
$$\hat{p}_{\mathsf{L}}(1|a_{1:t-1}) = \frac{t_1(a_{1:t-1}) + 1}{t+1},$$

where $t_0(a_{1:t-1})$ and $t_1(a_{1:t-1})$ denote the numbers of 0's and 1's in $a_{1:t-1}$, respectively.

P.-S. Laplace. 1814. *Essai philosophique sur les probabilités*.

Regret of the Laplace mixture

Theorem 3. The regret of the Laplace mixture forecaster satisfies

$$\sup_{a_{1:T} \in \mathcal{A}^T} R(\hat{p}_{\mathsf{L}}, a_{1:T}) = \log(T+1).$$

Remark. Recall the minimax regret is

$$R^\star = \frac{1}{2}\log(T+1) + \frac{1}{2}\log\frac{\pi}{2} - O\left(\frac{1}{\sqrt{T}}\right).$$

Therefore, the Laplace mixture forecaster is sub-optimal.

Proof of Theorem 3

Proof. (Theorem 3) By Proposition 1, we have

$$
\begin{aligned}
&- \log \hat{p}_{\mathsf{L}}(a_{1:T}) - \inf_{q \in [0,1]} \left\{ - \log \left[(1-q)^{t_0} q^{t_1} \right] \right\} \\
&\quad = - \log \frac{1}{(T+1)\binom{T}{t_1}} + \log \left[\left(\frac{t_0}{T} \right)^{t_0} \left(\frac{t_1}{T} \right)^{t_1} \right] \\
&\quad = \log \left[(T+1) \binom{T}{t_1} \left(1 - \frac{t_1}{T} \right)^{T-t_1} \left(\frac{t_1}{T} \right)^{t_1} \right] \\
&\quad \leq \log(T+1).
\end{aligned}
$$

The upper bound is achieved when $a_t = 1$ for all t.

Laplace & the sunrise problem

Problem. (Sunrise problem) How likely will the sun rise tomorrow?

Exercise. Google "sunrise problem" and "Good-Turing frequency estimation."

P.-S. Laplace. 1814. *Essai philosophique sur les probabilités*.
A. Orlitsky *et al.* 2003. Always Good Turing: Asymptotically optimal probability estimation.

Krichevsky-Trofimov mixture (1/2)

Algorithm. (Krichevsky-Trofimov mixture forecaster) The *Krichevsky-Trofimov mixture forecaster* is the mixture forecaster with π being the beta distribution of parameter $(1/2, 1/2)$; that is,

$$\hat{p}_{\mathsf{KT}}(a_{1:T}) = \int_0^1 (1-q)^{t_0} q^{t_1} \frac{1}{Z\sqrt{q(1-q)}} \,\mathrm{d}q, \quad \forall a_{1:T} \in \mathcal{A}^T,$$

where n_0 and n_1 denote the numbers of 0's and 1's in $a_{1:T}$, respectively, and Z denotes the normalizing constant

$$Z := \int_0^1 \frac{1}{\sqrt{q(1-q)}} \,\mathrm{d}q.$$

R. E. Krichevsky and V. K. Trofimov. 1981. The performance of universal encoding.

Krichevsky-Trofimov mixture (2/2)

Proposition 3. The KT mixture forecaster has the closed-form expression

$$\hat{p}_{\mathsf{KT}}(0|a_{1:t-1}) = \frac{t_0(a_{1:t-1}) + (1/2)}{t},$$
$$\hat{p}_{\mathsf{KT}}(1|a_{1:t-1}) = \frac{t_1(a_{1:t-1}) + (1/2)}{t},$$

where $t_0(a_{1:t-1})$ and $t_1(a_{1:t-1})$ denote the numbers of 0's and 1's in $a_{1:t-1}$, respectively.

Proof. Non-trivial calculation.

R. E. Krichevsky and V. K. Trofimov. 1981. The performance of universal encoding.

Regret of the KT mixture

Theorem 4. The regret of the KT mixture forecaster satisfies

$$\sup_{a_{1:T} \in \mathcal{A}^T} R(\hat{p}_{\mathsf{KT}}, a_{1:T}) \leq \frac{1}{2} \log(T+1) + \frac{1}{2} \log \pi.$$

Proof. Non-trivial calculations. See the references below.

Remark. Recall the minimax regret is

$$R^\star = \frac{1}{2} \log(T+1) + \frac{1}{2} \log \frac{\pi}{2} - O\left(\frac{1}{\sqrt{T}}\right).$$

Therefore, the KT mixture forecaster is close to being optimal.

Y. Freund. 2003. Predicting a binary sequence almost as well as the optimal biased coin.

K. S. Azoury and M. K. Warmuth. 2001. Relative loss bounds for on-line density estimation with the exponential family of distributions.

KT mixture and non-informative prior (1/3)

Problem. Let $\mathcal{P}_\Theta = \{ p_\theta \mid \theta \in \Theta \}$ be a parametric class of probability density/mass functions. Consider the problem of estimating $\theta^\natural \in \Theta$, given a random variable y following $p^\star_\theta$. Suppose we do not have any *a priori* information about $\theta^\natural$, and we want to adopt the Bayesian approach. What prior distribution on Θ should we use?

Remark. This is still an open problem! Even the necessity and formulation of the problem can be questioned.

H. Jeffreys. 1946. An invariant form for the prior probability in estimation problems.
E. T. Jaynes. 1995. *Probability Theory: The Logic of Science*.
J. O. Berger, J. M. Bernardo and D. Sun. 2009. The formal definition of reference priors.

KT mixture and non-informative prior (2/3)

Definition. (Jeffreys prior) Define the *Fisher information*

$$I(\theta) := -\mathsf{E}_{p_\theta}\left[\frac{\mathrm{d}^2 \log p(\theta)}{\mathrm{d}\theta^2}\right], \quad \forall \theta \in \Theta.$$

The *Jeffreys prior* is given as

$$\pi_{\mathsf{J}}(\theta) \propto \sqrt{I(\theta)}, \quad \forall \theta \in \Theta.$$

Remark. The Jeffreys prior is invariant under reparametrization.

Remark. The Jeffreys prior may not be *proper*.

H. Jeffreys. 1946. An invariant form for the prior probability in estimation problems.

KT mixture and non-informative prior (3/3)

Proposition 4. Consider the Bernoulli trial model, where $\Theta = [0, 1]$ and the class of probability mass functions

$$p_\theta(a_{1:T}) = (1-\theta)^{t_0(a_{1:T})}\theta^{t_1(a_{1:T})}, \quad \forall a_{1:T} \in \{\,0,1\,\}^T\,.$$

Then, the Jeffreys prior is the beta distribution of parameter $(1/2, 1/2)$.

Proof. Direct calculation.

Remark. The KT mixture corresponds to the mixture forecaster with the Jeffreys prior.

Sequential NML (1/2)

Algorithm. (Sequential normalized maximum likelihood) The *sequential normalized maximum likelihood (SNML) forecaster* outputs

$$\hat{p}_{\mathsf{SNML}}(a|a_{1:t-1}) = \frac{\hat{p}_{\mathsf{SNML}}(a_{1:t-1}a)}{\sum_{a' \in \mathcal{A}} \hat{p}_{\mathsf{SNML}}(a_{1:t-1}a')}, \quad \forall a_{1:t-1}a \in \mathcal{A}^t,$$

where

$$\hat{p}_{\mathsf{SNML}}(a_{1:t}) := \arg\inf_{p} \max_{a_{1:t} \in \mathcal{A}^t} R(p|a'_{1:t}).$$

E. Takimoto and M. K. Warmuth. 2000. The last-step minimax algorithm.

J. Rissanen and T. Roos. 2007. Conditional NML universal models.

W. Kotłowski and P. Grünwald. 2011. Maximum likelihood vs. sequential normalized maximum likelihood in on-line density estimation.

Sequential NML (2/2)

Remark. That is, the SNML forecaster coincides with running the NML forecaster for each round. By Theorem 1, we have

$$\hat{p}_{\mathsf{SNML}}(a_{1:t}) = \frac{\sup_{h\in\mathcal{H}} \prod_{\tau=1}^{t} h(a_\tau|a_{1:\tau-1})}{\sum_{a_{1:t}\in\mathcal{A}^t} \sup_{h\in\mathcal{H}} \prod_{\tau=1}^{t} h(a_\tau|a_{1:\tau-1})}, \quad \forall a_{1:t} \in \mathcal{A}^t.$$

Proposition 5. For the binary alphabet case, we have

$$\hat{p}_{\mathsf{SNML}}(1|a_{1:t-1}) = \frac{k_0^{k_0}(k_1+1)^{k_1+1}}{k_0^{k_0}(k_1+1)^{k_1+1} + (k_0+1)^{k_0+1}k_1^{k_1}},$$

where $k_0 := t_0(a_{1:t-1})$ and $k_1 := t_1(a_{1:t-1})$.

Regret of SNML

Theorem 5. The regret of the SNML forecaster satisfies

$$\sup_{a_{1:T} \in \mathcal{A}^T} R(\hat{p}_{\mathsf{SNML}}, a_{1:T}) \leq \frac{1}{2} \log(T+1) + \frac{1}{2}.$$

Remark. Recall the minimax regret is

$$R^\star = \frac{1}{2} \log(T+1) + \frac{1}{2} \log \frac{\pi}{2} - O\left(\frac{1}{\sqrt{T}}\right),$$

and the regret bound for the KT mixture is

$$\sup_{a_{1:T} \in \mathcal{A}^T} R(\hat{p}_{\mathsf{KT}}, a_{1:T}) \leq \frac{1}{2} \log(T+1) + \frac{1}{2} \log \pi.$$

E. Takimoto and M. K. Warmuth. 2000. The last-step minimax algorithm.

Proof of Theorem 5

Sketch of proof. (Theorem 5) Define

$$\delta_t := -\log \hat{p}_{\mathsf{SNML}}(a_t|a_{1:t-1}) - [-\log \hat{p}_{\mathsf{ML}}(a_{1:t})] + [-\log \hat{p}_{\mathsf{ML}}(a_{1:t-1})]\,,$$

where

$$\hat{p}_{\mathsf{ML}}(a_{1:t}) := \left(\frac{t_0(a_{1:t})}{t}\right)^{t_0(a_{1:t})} \left(\frac{t_1(a_{1:t})}{t}\right)^{t_1(a_{1:t})}.$$

Then, we have

$$R(\hat{p}_{\mathsf{SNML}}; a_{1:T}) = \sum_{t=1}^{T} \delta_t.$$

E. Takimoto and M. K. Warmuth. 2000. The last-step minimax algorithm.

What is the issue?

Recall the definition of the SNML forecaster:

$$\hat{p}_{\mathsf{SNML}}(1|a_{1:t-1}) = \frac{k_0^{k_0}(k_1+1)^{k_1+1}}{k_0^{k_0}(k_1+1)^{k_1+1} + (k_0+1)^{k_0+1}k_1^{k_1}}.$$

Definition. (Conditional regret) The *conditional regret* of a forecasting strategy $\hat{p}$ on $a_{T_0:T}$ on $a_{1:T_0-1}$ is given by

$$R(\hat{p}; a_{T_0:T}|a_{1:T_0-1}) := \sum_{t=T_0}^{T} \left[-\log \hat{p}(a_t|a_{1:t-1})\right] - \inf_{h\in\mathcal{H}} \sum_{t=1}^{T} \left[-\log h(a_t|a_{1:t-1})\right].$$

F. Hedayati and P. Bartlett. 2012. Exchangeability characterizes optimality of sequential normalized maximum likelihood and Bayesian prediction.

NML, SNML, and exchangeability (1/3)

Definition. (Exchangeability) A stochastic process $\{ \xi_t \mid t \in \mathbb{N} \}$ is called *exchangeable*, if and only if for every $T \in \mathbb{N}$ and permutation ϖ on $\{ 1, \ldots, T \}$, the joint probability distributions of $\xi_1, \ldots, \xi_T$ and $\xi_{\varpi(1)}, \ldots, \xi_{\varpi(T)}$ are the same.

Theorem. (de Finetti) Let $\{ \xi_t \mid t \in \mathbb{N} \}$ be an exchangeable collection of $\{ 0, 1 \}$-valued random variables. Then, there exists a probability distribution π on $[0, 1]$, such that for every $T \in \mathbb{N}$,

$$\mathsf{P}\left(\xi_{1:T} = x_{1:T}\right) = \int_0^1 (1-q)^{t_0(x_{1:T})} q^{t_1(x_{1:T})} \pi(\mathrm{d}q).$$

J. F. C. Kingman. 1978. Uses of exchangeability.

NML, SNML, and exchangeability (2/3)

Example. (Polya's urn.) Consider an urn in which there are n_0 white balls and n_1 black balls. Repeat the following procedures:

1. Draw a ball at random.
2. Note its color.
3. Put the ball back, together with a new ball of the same color.

Exercise. Show that the sequence of colors is an exchangeable stochastic process, while the outcomes of sequential drawings are statistically dependent.

NML, SNML, and exchangeability (3/3)

Definition. (Finite exchangeability) A stochastic process $\{ \xi_1, \ldots, \xi_T \}$ is called *exchangeable*, if and only if for every permutation ϖ on $\{ 1, \ldots, T \}$, the joint probability distributions of $\xi_1, \ldots, \xi_T$ and $\xi_{\varpi(1)}, \ldots, \xi_{\varpi(T)}$ are the same.

Theorem. (Hedayati-Bartlett) Consider the conditional regret such that the conditional $\hat{p}_{\mathsf{SNML}}$ is well-defined. Then, the SNML and NML forecasters are equivalent, if and only if $\hat{p}_{\mathsf{SNML}}$ is exchangeable.

Theorem. (Hedayati-Bartlett) For binary sequence prediction, the SNML forecaster is not a mixture forecaster, and does not achieve the minimax regret.

F. Hedayati and P. Bartlett. 2017. Exchangeability characterizes optimality of sequential normalized maximum likelihood and Bayesian prediction.

Conclusions

Conclusions

- The NML forecaster achieves the minimax regret.
- The Laplace and Krichevsky-Trofimov mixture forecasters achieve regret values close to the minimax value (with respect to static hypotheses).
- The KT mixture adopts the Jeffreys prior.
- The SNML forecaster may not be a mixture forecaster, while can yield satisfactory regret values.

Next lecture

- Aggregating algorithm.