Horizon-Independent Optimal Prediction with Log-Loss in Exponential Families

Peter Bartlett, Peter Grünwald, Peter Harremoës, Fares Hedayati, Wojciech Kotłowski

University of California at Berkeley
Centrum Wiskunde & Informatica
Copenhagen Business College
University of California at Berkeley
Poznań University of Technology

Player-Adversary's Game

- Players predict the outcomes of an event in an online fashion consulting with a class experts.
- At t player reveals belief about Y_t in form of $p_t(Y_t|y^{t-1})$, player can consult with i.i.d $p_\theta(\cdot)$, where $\theta \in \Theta$
- Adversary reveals y_t , the value of Y_t
- Player suffers $-\log p_t(y_t|y^{t-1})$
- Cumulative loss over n rounds is : $\sum_{t=1}^{n} -\log p_t(y_t|y^{t-1})$
- Cumulative loss if listened to $p_\theta(\cdot)$: $\sum_{t=1}^{n} -\log p_\theta(y_t)$

Online Learning with Logarithmic Loss

- GOAL: minimize the difference between player's cumulative loss and the loss of the best distribution (REGRET), over sequences of p_t and y_t :

$$R^{\Theta}(y^n, q^{(n)}) =$$

$$\sum_{t=1}^{n} -\log p_t(y_t|y^{t-1}) - \min_{\theta\in\Theta} \sum_{t=1}^{n} -\log p_\theta(y_t)$$

$$= \log \frac{\sup_\theta \; p_\theta(y^n)}{p^{(n)}(y^n)}$$

Sequential Probability Assignment Equivalent to Joint Distribution

Note that any sequential probability assignment of length n defines a joint distribution on the n outcomes and vice versa. This is because

$$\sum_{y^n} \prod_{t=1}^{n} p_t(y_t|y^{t-1}) = 1$$

And given a joint probability $p^{(n)}(\cdot)$, the conditional at time t is :

$$p_t(y_t|y^{t-1}) = \frac{p^{(n)}(y^t)}{p^{(n)}(y^{t-1})}$$

Normalized Maximum Likelihood

$$p^{(n)}_{nml}(y^n) \;\propto\; \sup_{\theta \in \Theta} \; p_\theta(y^n)$$

Theorem

NML achieves the minimax bound,that is,

$$p^{(n)}_{nml} = \text{argmin}_{q^{(n)}} \max_{y^n} R^{\Theta}(y^n, q^{(n)})$$

Bayesian Strategies

- Prior $\pi(\theta)$ on distributions $p_\theta(\,\cdot\,)$.
- Initially the strategy is a mixture of experts with prior π. As more y_t are observed we update the posterior and mix. The joint will be: $p_\pi(y^n) = \int_{\theta\in\Theta} p_\theta(y^n)\pi(\theta)\,d\theta$
- Conditionals will be :

$$p_\pi(Y_t = y_t \mid y^{t-1}) = \int_{\theta\in\Theta} p_\theta(y_t)\pi(\theta|y^{t-1})\,d\theta$$

- Jeffreys prior proportional to $\sqrt{I(\theta)}$ is asymptotically optimal (under some conditions called ineccsi) .

SNML

- Sequential normalized maximum likelihood.

$$p_{snml}(Y_t = y_t \mid y^{t-1}) \;\propto\; \sup_{\theta \in \Theta} \; p_\theta(y^{t-1}, y_t)$$

- One-step ahead lookup, following the advice of the maximum likelihood probability distribution of history concatenated with one observation in the future.

- Naturally defined in terms of conditionals.

- The regret is a constant away from the minimax regret.

Exponential Families

Suppose the parametric family of i.i.d distributions are a class of exponential distributions. $p_\theta(y) = h(y)e^{\theta^\intercal y - A(\theta)}$.
Hedayati and Bartlett showed that:
SNML and Bayesian with Jeffreys and NML are either all equivalent or are all different from each other. They are the same if and only if SNML is exchangeable.

Implications of Exchangeability

- NML becomes horizon–independent
- At time t instead of marginalizing $n - t$ random variables out, NML can just marginalize the next variable out as SNML does.
- NML becomes an infinite process, Bayesian updating.
- SNML and Bayesian with Jeffreys become optimal.

SNML-exchangeable Exponential Families

The only SNML- exchangeable one–dimensional exponential families are *Gaussian*, *gamma*, *Tweedie*$(\frac{3}{2})$ and any one–to–one transformation of them.