Redundancy Capacity Theorem for On-Line Learning Under a Certain Form of Hypotheses Class

Shachar Shayovitz and Meir Feder
Tel Aviv University

Information Theory Workshop 2018
Guangzhou, China

November 27, 2018

Universal Prediction

General Framework

- Predict y_t based on y^{t-1} where $y^{t-1} = \{y_1, y_2, ..., y_{t-1}\}$
- The model of y^n is unknown
- There are primarily two settings for this problem: individual and stochastic
- In the stochastic setting, the sequence is generated by some source $P_\theta(y^n)$ from the hypotheses class.

Objective

Sequentially predict y_t as if the statistical model was known!

Stochastic Setting

Assumptions

- Assume that there is a parameterized family of distributions $P_\theta(y^n)$ (Hypotheses Class)
- Nature chooses θ

Probabilistic Predictors

We concentrate on predictors of the form: $0 \leq q\left(\cdot|y^{t-1}\right) \leq 1$ where $\sum_{y_t} q(y_t|y^{t-1}) = 1$

Cost Function - Log-Loss

Log-loss is a commonly used cost function in many applications such as classification, data compression and more :

$$-\ln q\left(y_t|y^{t-1}\right)$$

Repeated Games - Regret & Redundancy

Regret (Comparing against the best)

$$Reg(\theta, y^n) = \sum_{t=1}^{n} \left(\ln P_\theta \left(y_t | y^{t-1} \right) - \ln q \left(y_t | y^{t-1} \right) \right)$$

Expected Regret - Redundancy

- Average over y (average case VS worst case)

$$R(q_1, q_2, ..., q_n, \theta) = E_{y^n} \left\{ Reg(\theta, y^n) \right\}$$

  where $q_t = q(y_t | y^{t-1})$
- Had θ been known, then the logloss optimal predictor is $P_\theta \left(y_t | y^{t-1} \right)$

Minimax Redundancy

Objective

$$R_{minimax} = \min_{q_1, q_2, \ldots, q_n} \max_{\theta} R(q_1, q_2, \ldots, q_n, \theta)$$

Relaxed Objective

$$R_{minimax} = \min_{q_1, q_2, \ldots, q_n} \max_{\pi(\theta)} E_{\pi(\theta)}\{R(q_1, q_2, \ldots, q_n, \theta)\}$$

Objective

- Find the universal predictor that minimizes the redundancy for the worst possible prior $\pi(\theta)$
- If the minimax redundancy grows sub-linearly, then redundancy rate goes to zero - universal predictor performs asymptotically as if θ had been known in hindsight.

Capacity Redundancy Theorem

Capacity Redundancy Theorem *(Gallager 79, Davisson & Leon-Garcia 80 and Rybako 79)*

$$R_{minimax} = \max_{\pi(\theta)} I(\theta; y^n)$$

Optimal Predictor - Mixture

$$q(y^n) = \sum_{\theta} \pi^*(\theta) p_\theta(y^n)$$

Sequential Form

$$q(y_t|y^{t-1}) = \sum_{\theta} w_t(\theta) P_\theta(y_t|y^{t-1})$$

On-Line Learning

Universal Prediction with Side Information

- Consider on-line learning with log-loss
- The goal is to predict the label (y_t) of a feature (x_t) based on past features and associated labels (x^{t-1}, y^{t-1}).
- The features may be considered as side information
- The hypotheses class: $P_\theta(y^n|x^n)$
- The predictor is now defined as $q\left(y_t|y^{t-1}, x^{t-1}, x_t\right)$
- Redundancy and Regret are changed accordingly

Hypotheses Class

Certain Form of Hypotheses Class

$$p^{\underline{\theta}}\left(y^t|x^t\right) = \Pi_{j=1}^{K} p^{\theta_j}\left(\underline{y}_j^t\right)$$

where

$$\underline{\theta} = [\theta_0, \theta_1, \theta_2, ..., \theta_{K-1}], \theta_i \in \Theta$$

and

$$\underline{y}_j^t = \{y_i, 0 \leq i \leq t | x_i = j\}$$

Conditional Probability

$$p^{\underline{\theta}}\left(y_t|y^{t-1}, x^t\right) = p^{\theta_{x_t}}\left(y_t|\underline{y}_{x_t}^{t-1}\right)$$

On-Line Learning

Example - Horse Race with Side Information

- The labels indicate the winning horse in each race - y_t
- The side information indicates the weather (sunny or rainy) in each race - x_t is binary
- The probability of winning can change based on the weather:

$$p^{\underline{\theta}}(y^n|x^n) = p^{\theta_{rainy}}(y_0^n)p^{\theta_{sunny}}(y_1^n)$$

Insight

- Notice that there is no assumption on $\pi(\theta)$ and in the extreme $\theta_{rainy} = \theta_{sunny}$
- The sequential predictor is $q(y_t|y^{t-1}, x^t)$.
- Can something be gained by looking at all the labels and not only on those in the same partition?

Minimax Redundancy for On Line Learning

Minimax Redundancy

$$R(x^n) = E_{\pi(\underline{\theta}|x^n),p^{\underline{\theta}}(y^n|x^n)}\left(\sum_{t=1}^{n}\left(\ln p^{\theta_{x_t}}\left(y_t|\underline{y}_{x_t}^{t-1}\right) - \ln q\left(y_t|y^{t-1},x^t\right)\right)\right)$$

Objective

$$R_{minimax}(x^n) = \min_{q_1,q_2,\ldots,q_n} \max_{\pi(\underline{\theta}|x^n)} R(x^n)$$

where $q_t = q(y_t|y^{t-1},x^t)$ is a universal predictor for y_t.

Related Results

Xie and Barron (2000)

- Hypotheses class was modeled as a multiplication of several i.i.d sources, determined by the side information.
- Proposed predictor is a multiplication of mixtures
- Achieves the *asymptotic minimax regret*.

Cover and Ordentlich (1996)

- A closely related problem of universal portfolio with side information was considered.
- Portofolios are compared to the best state constant rebalanced portfolios, where the side information determines the state.
- Optimal portfolio is a multiplication of mixtures of portfolios.
- Attains *asymptotic growth rate*

Related Results

Bottom Line

- The above universal predictors with side information only attain the asymptotic minimax regret
- It is unclear if these predictors are optimal in the non asymptotic minimax redundancy sense
- Does a Capacity Redundancy equivalence exists in these scenarios?

On-Line Learning

Main Question

Can the Capacity Redundancy theorem be extended to on-line learning?

On-Line Learning

Main Question

Can the Capacity Redundancy theorem be extended to on-line learning?

Answer

Still open (probably cannot be extended in general) but for a certain form of hypotheses class it holds

Capacity Redundancy Theorem for On-Line Learning

Theorem

There is an equivalence between the minimax redundancy and the sum of channel capacities.

$$R_{minimax}(x^n) = \sum_{j=1}^{K} C_j(x^n)$$

where,

$$C_j(x^n) = \max_{\pi(\theta_j|x^n)} I\left(\theta_j; \underline{y}_j^n | x^n\right)$$

the capacity of the channel between θ_j and $\underline{y}_j^n$ given x^n.

Capacity Redundancy Theorem for On-Line Learning

Theorem

The minimax redundancy problem for the hypotheses class is attained by the following on-line learner

$$q_t(y_t|y^{t-1},x^t) = \sum_{\theta_{x_t}} w(\theta_{x_t}) p^{\theta_{x_t}}\left(y_t|\underline{y}_{x_t}^{t-1}\right)$$

where,

$$w(\theta_{x_t}) = \frac{\pi(\theta_{x_t}|x^n) p^{\theta_{x_t}}\left(\underline{y}_{x_t}^{t-1}\right)}{\sum_{\theta_{x_t}} \pi(\theta_{x_t}|x^n) p^{\theta_{x_t}}\left(\underline{y}_{x_t}^{t-1}\right)}$$

Proof Outline - Maximin Solution

It turns out that the maximin problem can be written as,

$$R_{maximin}(x^n) = \max_{\pi(\underline{\theta}|x^n)} \sum_{\underline{\theta}} \pi(\underline{\theta}|x^n) D_{KL}\left(p^{\underline{\theta}}\left(y^n|x^n\right) || q^*\left(y^n|x^n\right)\right)$$

where,

$$q^*(y^n|x^n) = \sum_{\underline{\theta}} \pi(\underline{\theta}|x^n) p^{\underline{\theta}}\left(y^n|x^n\right)$$

Thus,

$$R_{maximin}(x^n) = \max_{\pi(\underline{\theta}|X^n=x^n)} I\left(\underline{\theta}; Y^n|X^n = x^n\right)$$

Proof Outline - Maximin Solution

Given x^n, we basically have K independent channels. Therefore, the distribution $\pi(\underline{\theta}|x^n)$ which maximizes the corresponding mutual information induces independence,

$$\pi(\underline{\theta}|x^n) = \Pi_{j=1}^{K}\pi(\theta_j|x^n)$$

Plugging in,

$$q^*(y^n|x^n) = \sum_{\underline{\theta}} \Pi_{j=1}^{K}\pi(\theta_j|x^n)p^{\theta_j}\left(\underline{y}_j^n\right)$$

Thus,

$$q^*(y^n|x^n) = \Pi_{j=1}^{K} \sum_{\theta_j} \pi(\theta_j|x^n)p^{\theta_j}\left(\underline{y}_j^n\right)$$

Then the maximin optimal universal predictor is in fact a multiplication of mixtures

Proof Outline - Maximin Solution

Also,

$$q_t^*(y_t|y^{t-1},x^t) = \sum_{\theta_{x_t}} w(\theta_{x_t}) p^{\theta_{x_t}} \left(y_t | \underline{y}_{x_t}^{t-1} \right)$$

where the weights $w(\theta_{x_t})$

$$w(\theta_{x_t}) = \frac{\pi(\theta_{x_t}|x^n) p^{\theta_{x_t}} \left(\underline{y}_{x_t}^{t-1} \right)}{\sum_{\theta_{x_t}} \pi(\theta_{x_t}|x^n) p^{\theta_{x_t}} \left(\underline{y}_{x_t}^{t-1} \right)}$$

Finally,

$$R_{maximin}(x^n) = \sum_{j=1}^{K} \max_{\pi(\theta_j|x^n)} I\left(\theta_j ; \underline{y}_j^n | x^n \right)$$

Proof Outline - Upper Bound on Minimax

We propose to minimize over a smaller set, Q, of universal predictors, each of the following form,

$$q(y^n|x^n) = \Pi_{j=1}^{K} q_j\left(\underline{y}_j^n\right)$$

where $\underline{y}_j^n = \{y_i, 0 \leq i \leq n | x_i = j\}$.
Plugging in,

$$\tilde{R}_{minimax}(x^n) = \min_{q(y^n|x^n)\in Q} \max_{\pi(\underline{\theta}|x^n)} E_{p(\underline{y},\underline{\theta}|x^n)} \left(\ln \frac{\Pi_{j=1}^{K} p^{\theta_j}\left(\underline{y}_j^n\right)}{\Pi_{j=1}^{K} q_j\left(\underline{y}_j{}^n\right)} \right)$$

where $R_{minimax}(x^n) \leq \tilde{R}_{minimax}(x^n)$

Proof Outline - Upper Bound on Minimax

After simple manipulations,

$$\tilde{R}_{minimax}(x^n) = \sum_{j=1}^{K} \min_{q(\underline{y}_j^n|\underline{x}_j^n)} \max_{\pi(\theta_j|x^n)} \sum_{\theta_j} \pi(\theta_j|x^n) D_{KL}\left(p^{\theta_j}\left(\underline{y}_j^n\right) || q_j\left(\underline{y}_j^n\right)\right)$$

Using minimax theorem K times we get,

$$\tilde{R}_{minimax}(x^n) = \sum_{j=1}^{K} \max_{\pi(\theta_j|x^n)} I\left(\theta_j; \underline{y}_j^n | x^n\right)$$

with the same multiplication of mixtures predictor

Summary

- Capacity Redundancy theorem for on-line learning was proven, under a certain form of hypotheses class
- On-line universal predictor that can achieve the minimax redundancy for an appropriate choice of prior distribution.
- Moreover, the universal predictors proposed by Xie & Barron and Cover & Ordentlich are in-fact special cases of ou proposed predictor, which according to our proof achieves the minimax redundancy even non asymptotically.