Catching Up Faster in Bayesian Model Selection and Model Averaging (T61)

Tim van Erven Peter Grünwald Steven de Rooij

1. Introduction & Summary

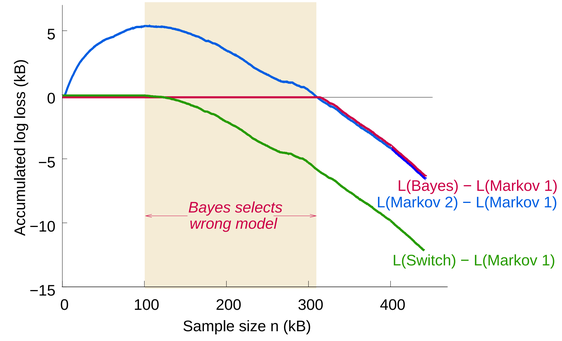
- AIC-BIC Dilemma: Bayes factor model averaging, model selection and their approximations such as BIC are generally statistically consistent, but sometimes achieve slower rates of convergence than other methods such as AIC and leave-one-out cross-validation. On the other hand, these other methods can be inconsistent.

	Consistent	Fast convergence
BIC, Bayes, MDL	✓	✗
AIC, LOO cross-validation	✗	✓
New method: Switch-Distribution	✓	✓

- Catch-Up Phenomenon: our novel explanation for the slow convergence of Bayesian methods.
- Switch Distribution: A modification of the Bayesian marginal distribution, suggested by analysis of the catch-up phenomenon.
- Theoretical Optimality Results: We prove that model selection and prediction based on the switch-distribution is typically both consistent and achieves optimal convergence rates, thereby resolving the AIC-BIC dilemma.
- Practical Use: The method is practical; we give an efficient implementation.

2. The Catch-up Phenomenon

Example: Log loss on the first n characters in "The Picture of Dorian Gray" as a function of n



- L(Markov 1) and L(Markov 2): Log loss (in bits) on $x^n = x_1, \ldots, x_n$ of the Bayesian marginal distribution for the first-order and second-order Markov chains with uniform prior $w_j(\theta) \equiv 1$:

$$\text{L(Markov } j) = -\log P_j(x^n) \;\; ; \;\; P_j(x^n) = \int P_\theta(x^n) w_j(\theta)\, d\theta$$

- L(Bayes): Log loss of Bayesian model averaging over first- and second-order Markov chains
- The catch-up phenomenon: After $100\,000$ characters, Markov 2 is $40\,000$ bits behind on Markov 1, and $210\,000$ characters further into the novel Markov 2 manages to catch up.

3. The Catch-up Phenomenon Makes Bayes Suboptimal

The Marginal Distribution is a Sequential Prediction Strategy

- By the chain rule, any (Bayesian) marginal distribution on x^n may be written as the product of sequential predictions of the next character given all previous characters:

$$P(x^n) = \prod_{i=1}^{n} P(x_i \mid x^{i-1}) \tag{1}$$

- Taking the negative logarithm of this expression shows that the log loss on x^n may be viewed as the accumulated log loss that is incurred in sequentially predicting the characters one by one.

Suboptimality of Bayesian Model Averaging

- Hence, the fact that Markov 2 is catching up with Markov 1 in the shaded region of the Example figure, means that Markov 2 is making better predictions than Markov 1.
- We see, however, that the Bayesian loss follows that of Markov 1, which has smallest *accumulated* log loss: In the shaded region, the Bayesian predictions follow the wrong model!

Theoretical Example: Histogram Density Estimation, Regression

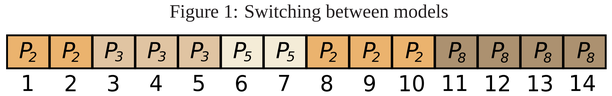
- The catch-up phenomenon occurs because different models are best at different sample sizes.
- This also causes Bayes to achieve a suboptimal rate of convergence when selecting the number bins in histogram density estimation and regression (with e.g., polynomials). (It wastes an $O(\log n)$ factor.)

4. Solution: The Switch-Distribution

A Single Switching Sequence

- By viewing marginal distributions as sequential prediction strategies, it becomes possible to switch between them.
- For example, given an infinite number of models $\mathcal{M}_1, \mathcal{M}_2, \ldots$ with respective marginal distributions $P_1, P_2, \ldots$, Figure 1 illustrates switching from P_2 to P_3, P_5, P_2 again, and finally to P_8.

Figure 1: Switching between models



Then the distribution on x^{14} is: $P(x^{14}) = P_2(x^2)P_3(x^5|x^2)P_5(x^7|x^5)P_2(x^{10}|x^7)P_8(x^{14}|x^{10})$

The Switch-Distribution

- The switch-distribution P_{sw} is defined by putting a prior distribution, π, on all possible switching sequences.
- In the Example, we see that it closely follows the best-predicting model at all sample sizes. E.g. we may take

$$\pi(\text{switch at sample sizes } t_1, \ldots, t_m \text{ to } \mathcal{M}_{k_1}, \ldots, \mathcal{M}_{k_m}) \quad \propto \quad 2^{-m} \prod_{i=1}^{m} k_i^{-2} t_i^{-2} \tag{2}$$

5. Theorems

Let $X^\infty = X_1, X_2, \ldots$ be a sequence of random variables, and let $X_a^b := X_a, X_{a+1}, \ldots, X_b$. Suppose that $P_1, P_2, \ldots$ are the Bayesian marginal distributions corresponding to parametric models $\mathcal{M}_1, \mathcal{M}_2, \ldots$ with respective parameter spaces $\Theta_1, \Theta_2, \ldots$ and priors $w_1, w_2, \ldots$ Finally, suppose that P_{sw} is the switch-distribution for these models with prior π.

Theorem 1 (Consistency of the Switch-Distribution)**.** *Suppose π is as in (2). Suppose also that for every k, k', $k \neq k'$ and any n initial outcomes x^n, the conditional distributions $P_k(X_{n+1}^\infty \mid x^n)$ and $P_{k'}(X_{n+1}^\infty \mid x^n)$ are mutually singular. Then, for all $k^* \in \mathbb{Z}^+$, for all $\theta^* \in \Theta_{k^*}$ except for a subset of Θ_{k^*} of w_{k^*}-measure 0, the posterior distribution on models satisfies*

$$\pi(k^* \mid X_1^n) \overset{n\to\infty}{\longrightarrow} 1 \qquad \textit{with } P_{\theta^*}\textit{-probability } 1. \tag{3}$$

- For $\mathcal{M} = \bigcup_{k\geq 1} \mathcal{M}_k$, define the information closure as $\langle \mathcal{M} \rangle = \{P^* \mid \inf_{P\in\mathcal{M}} D(P^*\|P) = 0\}$, the set of distributions for X^∞ that can be arbitrarily well approximated by elements of $\mathcal{M}$.
- The risk at sample size $n \geq 1$ of an estimator P relative to P^* is defined as

$$R_n(P^*, P) = E_{X^{n-1}\sim P^*}[D(P^*(X_n = \cdot \mid X^{n-1}) \| P(X_n = \cdot \mid X^{n-1}))],$$

- For each k, let P_k be an estimator associated with model $\mathcal{M}_k$ (e.g. ML estimator or Bayes predictive distribution). For $\mathcal{M}^* \subset \langle \mathcal{M} \rangle$, let $h(n)$ be the minimax convergence rate:

$$h(n) = \inf_{\delta:\mathcal{X}^n\to\{1,2,\ldots,n\}} \sup_{P^*\in\mathcal{M}^*} \sup_{n'>n} R_{n'}(P^*, P_\delta). \tag{4}$$

Theorem 2 (Optimal Rates of Convergence of the Switch-Distribution)**.** *Suppose π is as in (2). Let $\mathcal{M}^*$ be any subset of $\langle \mathcal{M} \rangle$ with minimax rate h such that $nh(n)$ is increasing, and $nh(n)/(\log n)^2 \to \infty$. Then*

$$\limsup_{n\to\infty} \frac{\sup_{P^*\in\mathcal{M}^*} \sum_{i=1}^n R_i(P^*, P_{\text{sw}})}{\sum_{i=1}^n h(i)} \leq 1. \tag{5}$$

6. Efficient Algorithm

- In the paper we give an algorithm to compute the probability $P_{\text{sw}}(x^n)$, which has running time that is linear in both the sample size and the number of models.
- The algorithm is similar to the FIXED-SHARE algorithm for tracking the best expert, see M. Herbster and M. K. Warmuth. *Tracking the best expert.* Machine Learning, 32:151–178, 1998

Acknowledgements

This work was generously supported by the PASCAL Network of Excellence.

Contact information

E-mail: {Tim.van.Erven, Peter.Grunwald, Steven.de.Rooij}@cwi.nl
Centrum voor Wiskunde en Informatica (CWI)
Kruislaan 413, P.O. Box 94079
1090 GB Amsterdam, The Netherlands

PASCAL
www.pascal-network.org

CWI