Low-Complexity Nonparametric Bayesian Online Prediction with Universal Guarantees

Alix Lhéritier Frédéric Cazals

amadeus Inria informatics mathematics

Online prediction with side information

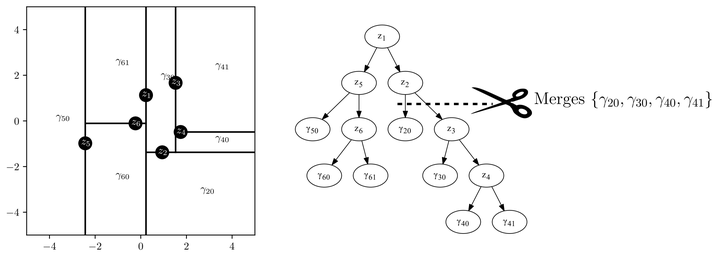
Features $\in \mathbb{R}^d$ $\quad z^n \equiv z_1, \ldots, z_n$

Labels $\in \mathcal{L}$ (discrete) $\quad l^n \equiv l_1, \ldots, l_n$ predict: $P(l_n | l^{n-1}, z^n)$

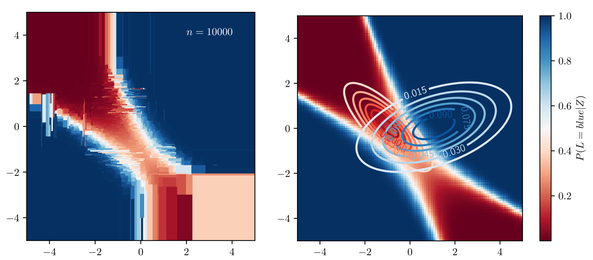
- **Goal:** $-\frac{1}{n} \log P(l^n | z^n)$ **asymptotically optimal**.
- **Probabilistic setting:** (z_i, l_i) are i.i.d. realizations of RV (Z, L).
  $\Rightarrow$ Optimum: $H(L|Z)$.
- **Previous work:** scale-hyperparameter dependence and high complexity
  $\Rightarrow$ k-nn: needs $k(n)$, $O(n^2)$. Gaussian Processes: need kernel width, $O(n^4)$.

Online data-driven discretization of the feature space

- **Full-fledged k-d trees:** given z^n, feature space is recursively partitioned by hyperplanes passing by the projection of z_i onto a perpendicular random axis.
- Depth is $O(\log n)$ in probability with respect to $\mathbb{P}_{Z^n} \Rightarrow$ **low-complexity**.
- $\mathcal{P}_n$ is the set of all the partitions obtained by pruning the partitioning tree.



- Partitioning rule π_n s.t. $H(L|\pi_n(Z|Z^n)) \xrightarrow{\text{a.s.}} H(L|Z) \Rightarrow$ **nonparametric**.



Bayesian label prediction with context tree models

- **Context Tree Weighting:** a Bayesian mixture on piecewise multinomial distributions over a nested $\mathcal{P}$, with a *natural encoding prior* ($P(\text{split}) = 1/2$)

$$P_{\text{CTW}}(l^n) \equiv \sum_{C \in \mathcal{P}} w_{\text{nat}}(C) \prod_{\gamma \in C} P_{\text{Jeffreys}}(\gamma(l^n))$$

  where $\gamma(l^n)$ is the subsequence falling in γ.
- **Context Tree Switching** is a mixture over sequences of these distributions.
- For any partition $C \in \mathcal{P}$, the context $C(L)$ is the node γ_i where L falls.

  $-\lim_{n\to\infty} \frac{1}{n} \log P_{\text{CT}.}(L^n) \leq H(L|C(L))$ a.s. $\Rightarrow$ **universal**.

The kd-switch distribution

- For a given node γ split into γ_1 and γ_2 by a k-d tree on z^n,

$$P^{\gamma}_{\text{kds}}(l^n|z^n) \equiv \sum_{i^n \in \{a,b\}^n} w_\gamma(i^n) \prod_{k=1}^{n} \left[\mathbb{1}_{\{i_k=a\}} \phi_a(l_k|l^{k-1}) + \mathbb{1}_{\{i_k=b\}} \phi^{\gamma}_b(l_k|l^{k-1}, z^k) \right]$$

$$\phi_a(l_k|l^{k-1}) \equiv P_{\text{kt}}\left(l_k|l^{k-1}\right)$$

$$\phi^{\gamma}_b(l_k|l^{k-1}, z^k) \equiv \begin{cases} P_{\text{kt}}\left(l_k|l^{k-1}\right) & \text{if } k < \tau_k(\gamma) \\ \dfrac{P^{\gamma_j}_{\text{kds}}\left(\gamma_j\left(l^k\right)|\gamma_j\left(z^k\right)\right)}{P^{\gamma_j}_{\text{kds}}\left(\gamma_j\left(l^k\right)^{-1}|\gamma_j\left(z^k\right)^{-1}\right)} & \text{with } j : z_k \in \gamma_j, \text{ otherwise} \end{cases}$$

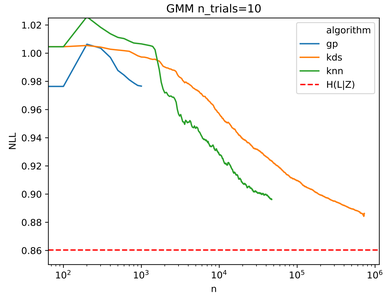
where $w_\gamma(\cdot)$ is a prior over $\{a,b\}^n$ and $\tau_n(\gamma)$ is the splitting index in $\gamma(z^n)$.

Main result: pointwise universality

Thm: The **kd-switch** distribution is pointwise universal, i.e.

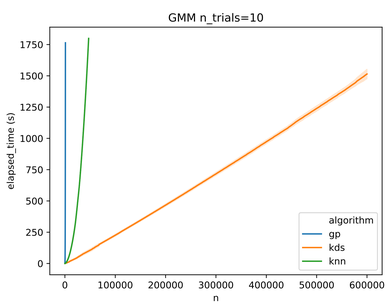
$$-\lim_{n\to\infty} \frac{1}{n} \log P_{\text{kds}}(L^n|Z^n) \leq H(L|Z) \text{ a.s.}$$

for any probability measure $\mathbb{P}$ generating the samples such that $\mathbb{P}_{Z|L}$ are absolutely continuous with respect to the Lebesgue measure.



Low-complexity online algorithm

The **kd-switch** distribution can be computed online in $O(n \log n)$ time.



Application: Sequential two-sample testing [5]

- The two-sample problem:

$$\begin{cases} \mathtt{H}_0 : \mathbb{P}_{Z|L=0} = \mathbb{P}_{Z|L=1} \text{ a.e.} \\ \mathtt{H}_1 : \neg \mathtt{H}_0 \end{cases}.$$

- Thm: Given an online predictor P, a test rejecting $\mathtt{H}_0$ at **any index** n s.t.

$$\frac{\mathbb{P}(l^n)}{P(l^n|z^n)} \leq \alpha$$

  has a Type I error probability less or equal than α, i.e.

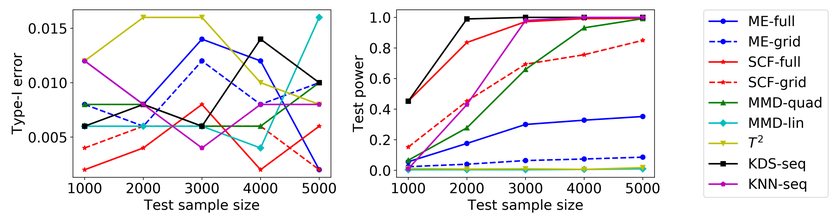
$$\mathbb{P}_{\mathtt{H}_0}\left(\exists n : \frac{\mathbb{P}(L^n)}{P(L^n|Z^n)} \leq \alpha \right) \leq \alpha.$$

- Thm: A **pointwise universal** P yields a **consistent** two-sample test:

$$\mathbb{P}(\text{Type II error}) \to 0.$$

Sequential vs training-set-optimized two-sample tests

- KNN-seq: k-nn based sequential test [5].
- SCF-full and ME-full [1, 3]:
  - Kernel based consistent two-sample tests.
  - Train/test paradigm to optimize revealing locations and kernel width.
  - **Valid if n is fixed in advance.**



References

[1] K. P. Chwialkowski, A. Ramdas, D. Sejdinovic, and A. Gretton.
Fast two-sample testing with analytic representations of probability measures.
In *Advances in Neural Information Processing Systems*, pages 1972–1980, 2015.

[2] L. Devroye, L. Györfi, and G. Lugosi.
A probabilistic theory of pattern recognition, volume 31.
Springer Verlag, 1996.

[3] W. Jitkrittum, Z. Szabó, K. P. Chwialkowski, and A. Gretton.
Interpretable distribution features with maximum testing power.
In *Advances in Neural Information Processing Systems*, pages 181–189, 2016.

[4] R. Krichevsky and V. Trofimov.
The performance of universal encoding.
Information Theory, IEEE Transactions on, 27(2):199–207, 1981.

[5] A. Lhéritier and F. Cazals.
A sequential non-parametric multivariate two-sample test.
IEEE Transactions on Information Theory, 64(5):3361–3370, 2018.

[6] J. Veness, K. S. Ng, M. Hutter, and M. Bowling.
Context tree switching.
In *Data Compression Conference (DCC), 2012*, pages 327–336. IEEE, 2012.

[7] F. M. J. Willems, Y. M. Shtarkov, and T. J. Tjalkens.
The context-tree weighting method: Basic properties.
Information Theory, IEEE Transactions on, 41(3):653–664, 1995.