Low-Complexity Nonparametric Bayesian Online Prediction with Universal Guarantees

amadeus

Alix Lhéritier   Frédéric Cazals

alix.lheritier@amadeus.com   frederic.cazals@inria.fr

Inria

Online prediction with side information

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

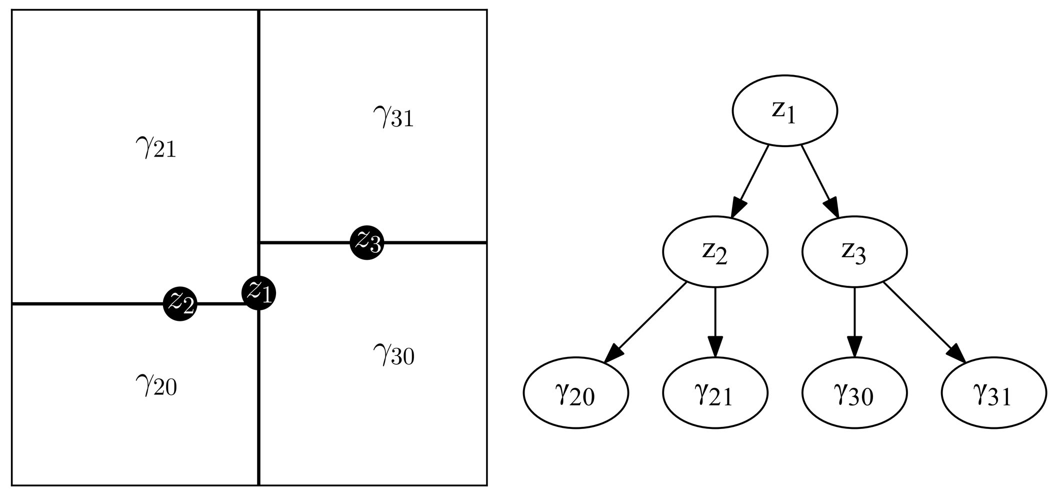
Labels $\in \mathcal{L}$ (discrete) $\quad l^n \equiv \boxed{l_1, \ldots, l_n}$ $\quad$ predict: $P(l_n | l^{n-1}, z^n)$

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

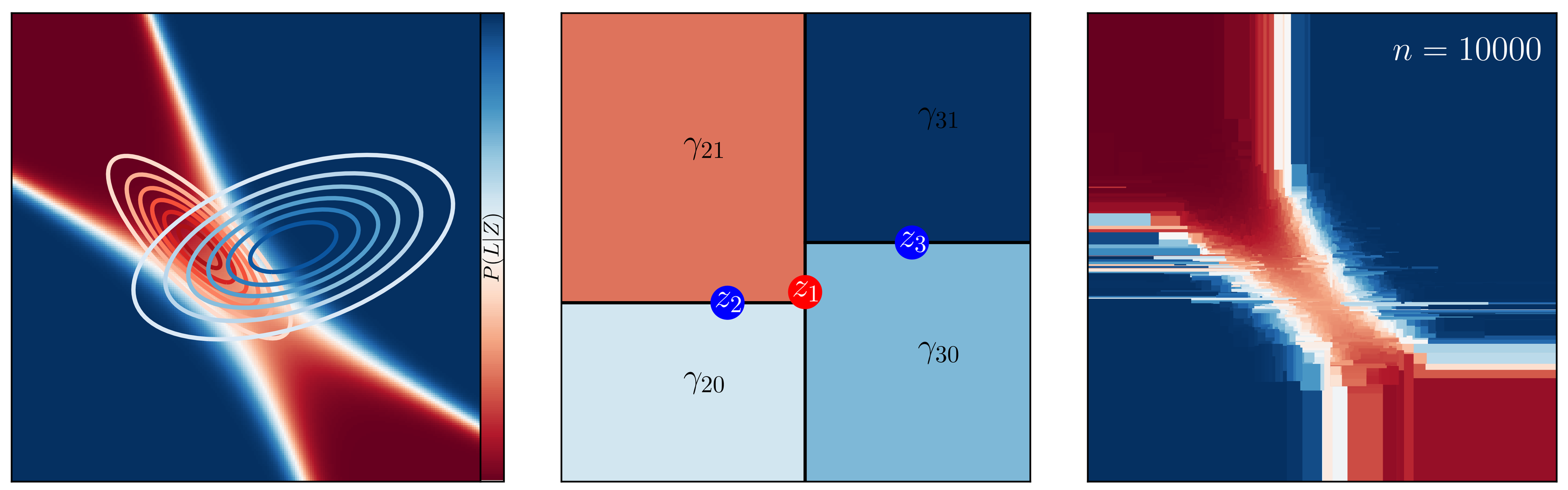
Online data-driven discretization of the feature space

- **Full-fledged k-d trees:** online recursive data-driven partitioning of $\mathbb{R}^d$.

Each z_i induces a hyperplane, perpendicular to a random axis, cutting the cell containing z_i.



- Depth is $O(\log n)$ in prob. w.r.t. $\mathbb{P}_{Z^n}$ $\Rightarrow$ **low-complexity**.
- Partitioning rule π_n s.t. $H(L|\pi_n(Z|Z^n)) \xrightarrow{\text{a.s.}} H(L|Z)$ $\Rightarrow$ **nonparametric**.

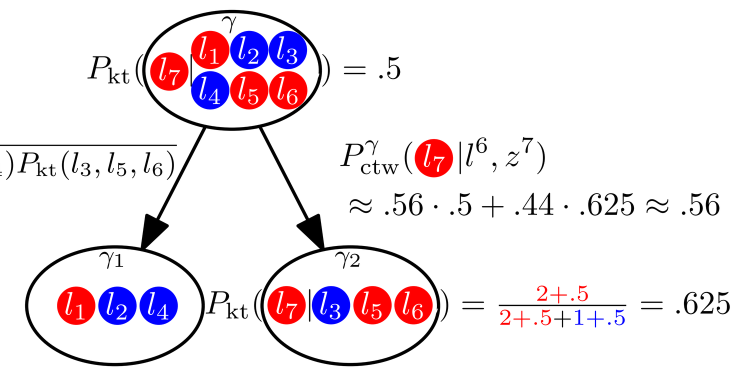


Bayesian label prediction with context tree models

- Mixture of Bernoullis in each cell: $P_{\text{kt}}(l^n) = \int_\theta \mathcal{B}(l^n|\theta) w_{\text{Jeffreys}}(\theta) d\theta$
- Recursive mixture at each internal node γ of a given partitioning tree T:

$$P^{\gamma}_{\text{ctw}}(l^n|z^n) = \tfrac{1}{2} P_{\text{kt}}(l^n) + \tfrac{1}{2} P^{\gamma_1}_{\text{ctw}}(\gamma_1(l^n)|\gamma_1(z^n)) P^{\gamma_2}_{\text{ctw}}(\gamma_2(l^n)|\gamma_2(z^n))$$

$$w_\gamma(\neg\text{split}|l^6, z^6) = \frac{P_{\text{kt}}(l^6)}{P_{\text{kt}}(l^6) + P_{\text{kt}}(l_1, l_2, l_4) P_{\text{kt}}(l_3, l_5, l_6)} \approx \frac{.0049}{.0049 + .0039} \approx .56$$



Switching:
Even if true $\mathbb{P}$ is complex, simpler models perform better when few samples are available.
Switch prior $w_\gamma(\rho_i \neq \rho_{i-1}) = \alpha_i$ allows time-dependent models ρ_i.

- **For any partition** A (with any number of cells) defined by a pruning of T,

$$-\lim_{n\to\infty} \frac{1}{n} \log P_{\text{ct}\cdot}(L^n|Z^n) \leq H(L|A(Z)) \text{ a.s.} \Rightarrow \textbf{automatic scale learning}.$$

The kd-switch distribution

- Given l^n s.t $z_i \in \gamma$ split into γ_1 and γ_2 at index τ_γ by a k-d tree on z^n,

$$P^{\gamma}_{\text{kds}}(l^n|z^n) \equiv \sum_{\rho^n \in \{P_{\text{kt}}, P_{\text{rec}}\}^n} w_\gamma(\rho^n) \prod_{k=1}^{n} \rho_k(l_k|l^{k-1}, z^k)$$

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

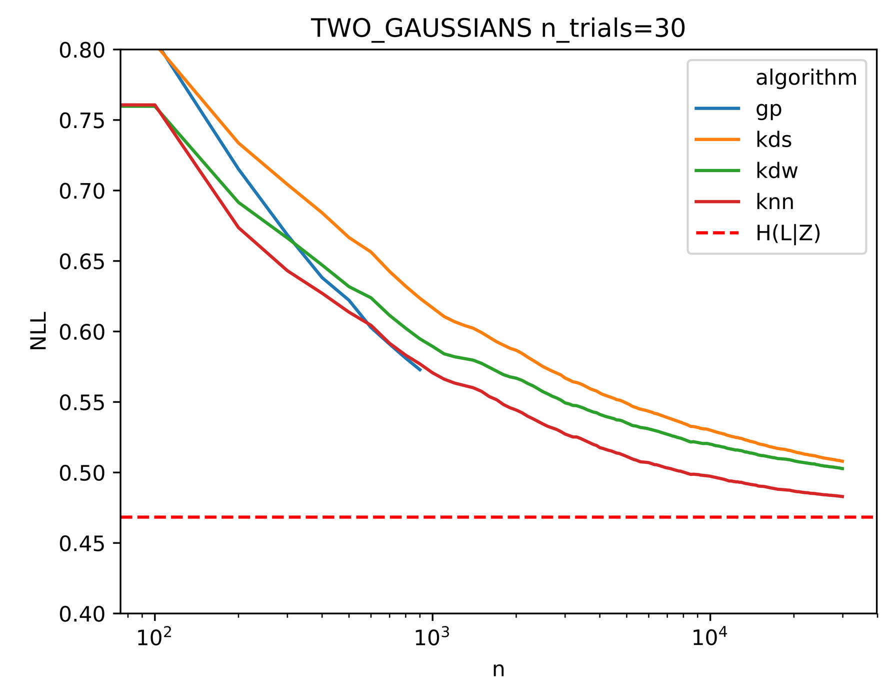
where $\cdot^{-1}$ removes the last symbol.

Main result: pointwise universality

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

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

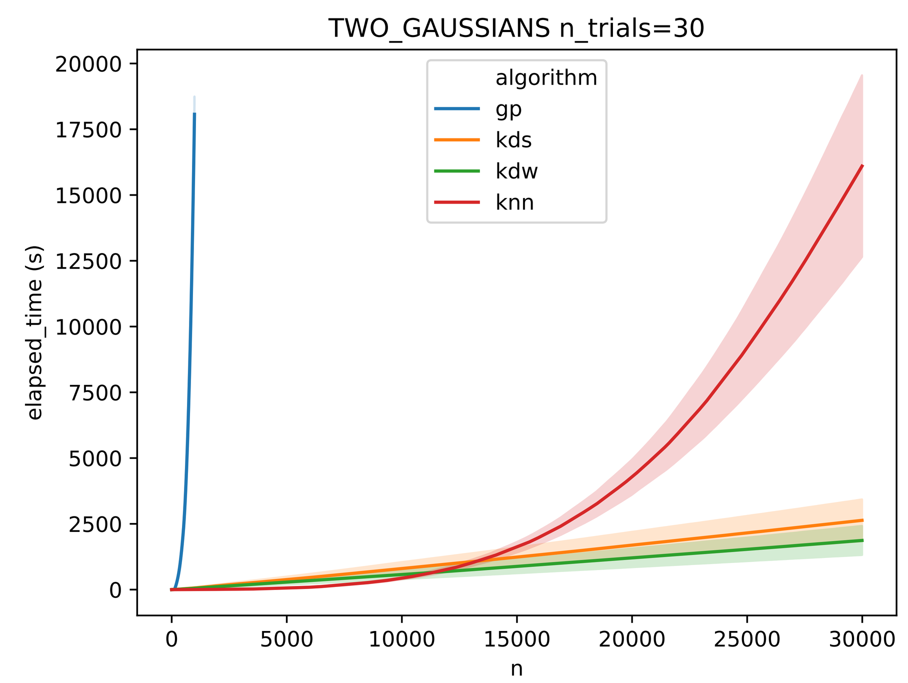
for any $\mathbb{P}$ generating the samples s.t. $\mathbb{P}_{Z|L} \ll \lambda$.



Performance is boosted by a Bayesian mixture over $J = 50$ trees.

Low-complexity online algorithm

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



Application: sequential two-sample testing [4]

- The two-sample problem:

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

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

$$\frac{\mathbb{P}(l^n)}{P(l^n|z^n)} \leq \alpha \quad (\mathbb{P}_L \text{ is known})$$

has a **Type I error probability**

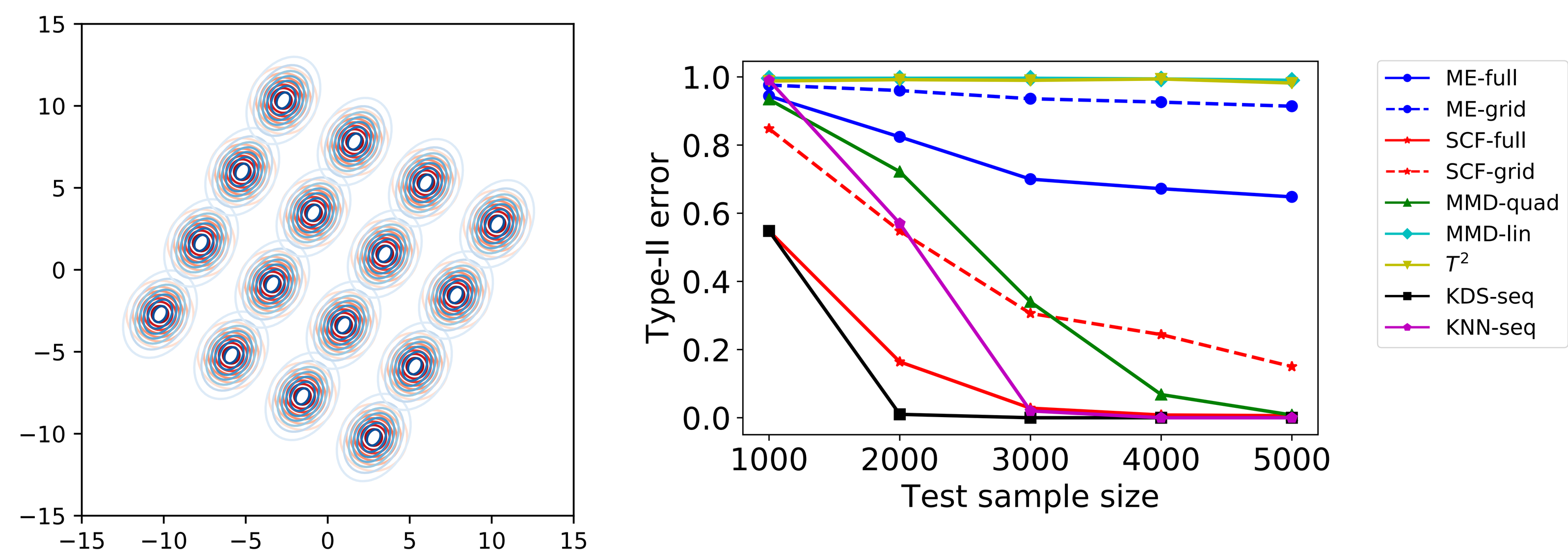
$$\mathbb{P}_{\mathtt{H}_0}\left(\exists n : \frac{\mathbb{P}(L^n)}{P(L^n|Z^n)} \leq \alpha\right) \leq \alpha.$$

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

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

Sequential vs training-set-optimized two-sample tests

- KNN-seq: k-nn based sequential test [4].
- MMD, SCF and ME [2]:
  - Kernel based consistent tests.
  - Train/test paradigm to optimize revealing locations and kernel width.
  - **Valid if n is fixed in advance.**



$\mathtt{H}_1$: Two Gaussian mixtures with small scale differences. $\alpha = .01$, #trials $= 500$, $J = 50$.

References

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

[2] 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.

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

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

[5] T. van Erven, P. Grünwald, and S. de Rooij.
Catching up faster by switching sooner: a predictive approach to adaptive estimation with an application to the AIC–BIC dilemma.
Journal of the Royal Statistical Society: Series B (Statistical Methodology), 74(3):361–417, 2012.

[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.