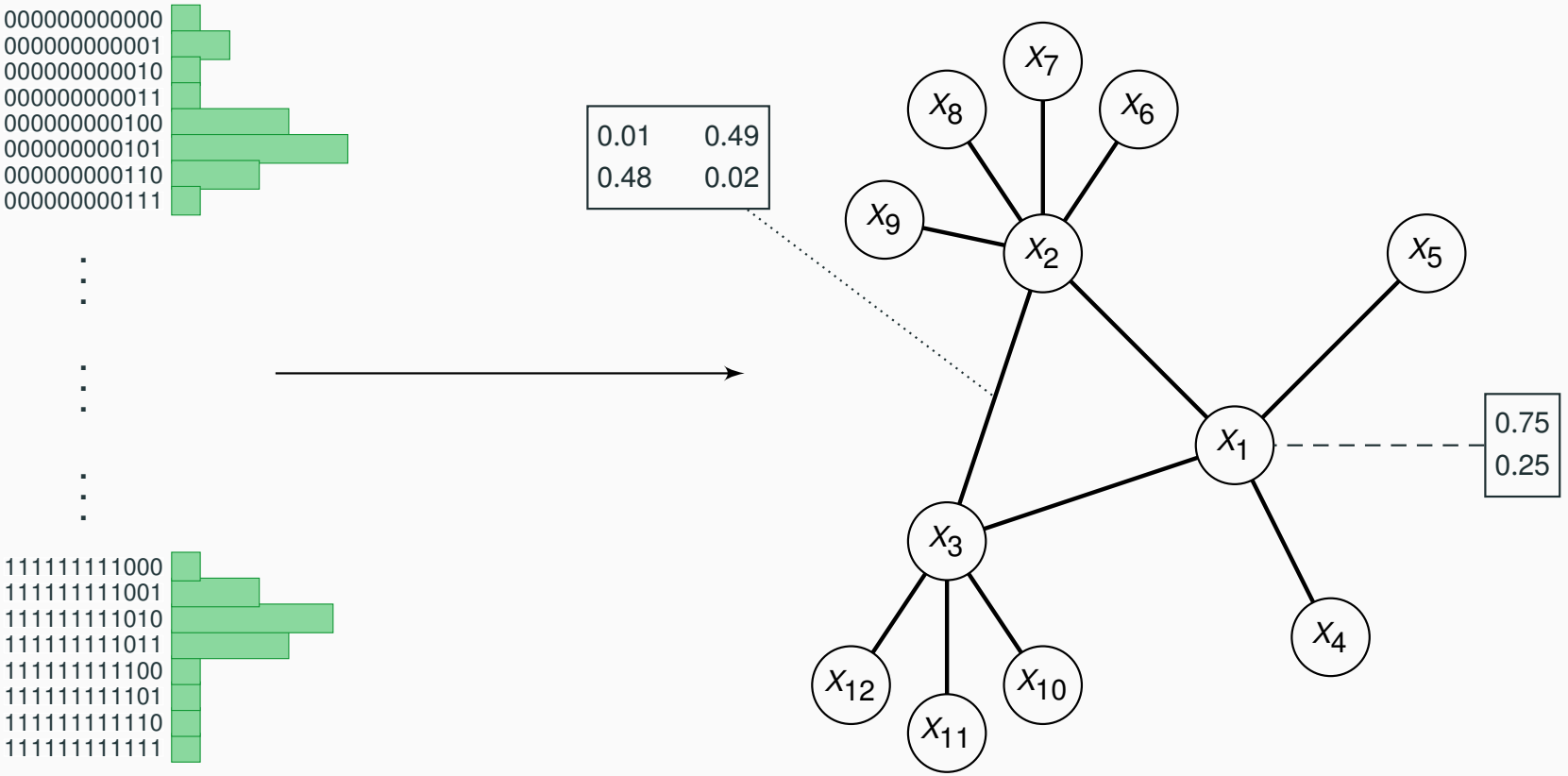
Online Forest Density Estimation

Frédéric Koriche
CRIL - CNRS UMR 8188, Univ. Artois
koriche@cril.fr

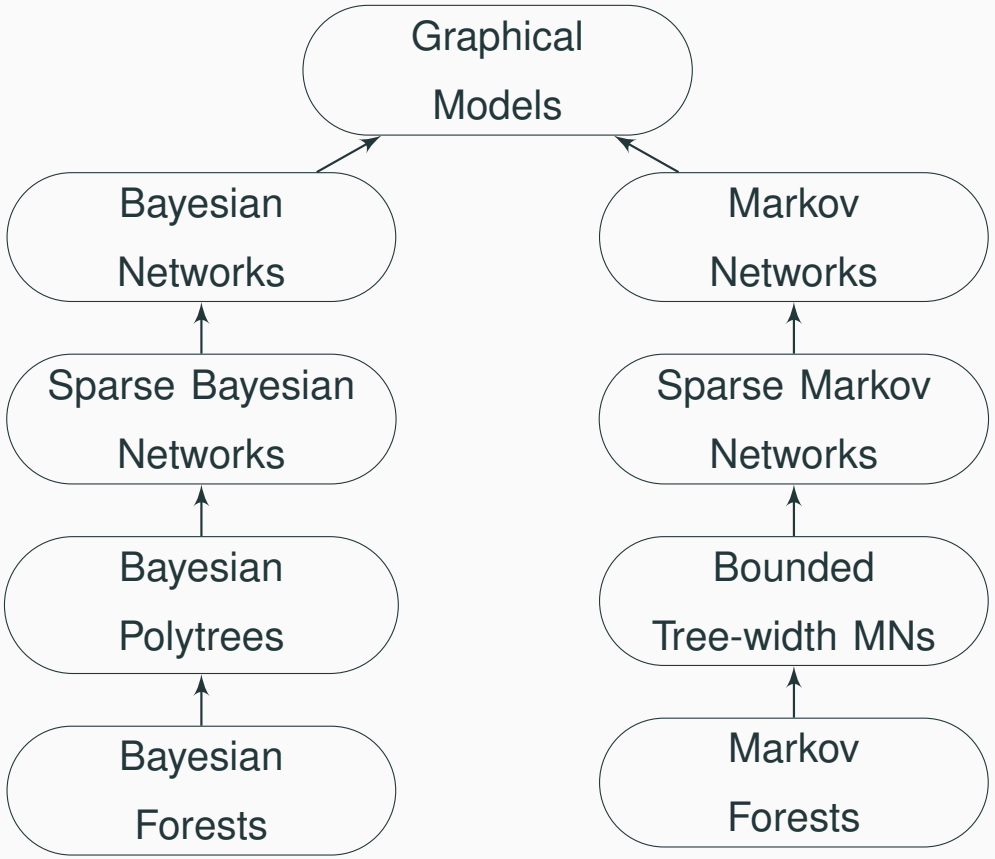
UAI'16

Outline



Graphical models encode high-dimensional distributions in a compact and intuitive way:

- Qualitative uncertainty (interdependencies) is captured by the structure
- Quantitative uncertainty (probabilities) is captured by the parameters

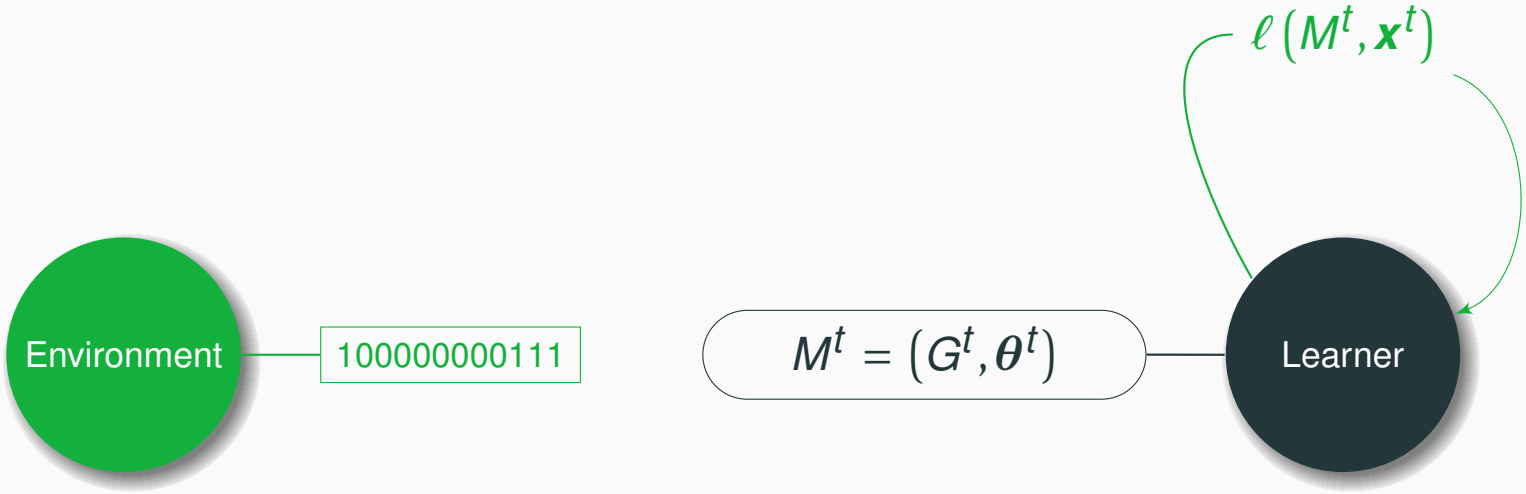


For an outcome space $\mathscr{X} \subseteq \mathbb{R}^n$, a class of graphical models is a pair $\mathscr{M} = \boldsymbol{G} \times \boldsymbol{\Theta}$, where $\boldsymbol{G}$ is space of n-dimensional graphs, and $\boldsymbol{\Theta}$ is a space of d-dimensional vectors.

- $\boldsymbol{G}$ captures structural constraints (directed vs. undirected, sparse vs. dense, etc.)
- $\boldsymbol{\Theta}$ captures parametric constraints (binomial, multinomial, Gaussian, etc.)

Outline

Online Density Estimation



Repeated game between the learner and its environment. During each trial $t = 1, \cdots, T$,

- the learner chooses (the structure and the parameters of) a model $M^t \in \mathscr{M}$;
- the environment responds by an outcome $\boldsymbol{x}^t \in \mathscr{X}$, and the learner incurs the log-loss

$$\ell(M^t, \boldsymbol{x}^t) = -\ln \mathbb{P}_{M^t}(\boldsymbol{x}^t)$$

Online density estimation is particularly suited to:

* Adaptive environments, where the target distribution can change over time;
* Streaming applications, where all the data is not available in advance;
* Large-scale datasets, by processing only one outcome at a time.

In the literature of online density estimation (universal coding):

- uni-dimensional models (binomial, multinomial, exponential families) have been extensively studied

  Xie and Barron (2000); Takimoto and Warmuth (2000); Kotłowski and Grünwald (2011),...

- much less is known, however, about multi-dimensional models, especially graphical models, where both the structure and the parameters are updated at each iteration!

The performance of an online learning algorithm A is measured according to two metrics:

Minimax Regret

Defined by the maximum, over every sequence of outcomes $\boldsymbol{x}^{1:T} = (x^1, \cdots, x^T)$, of the cumulative relative loss between A and the best model in $\mathscr{M}$:

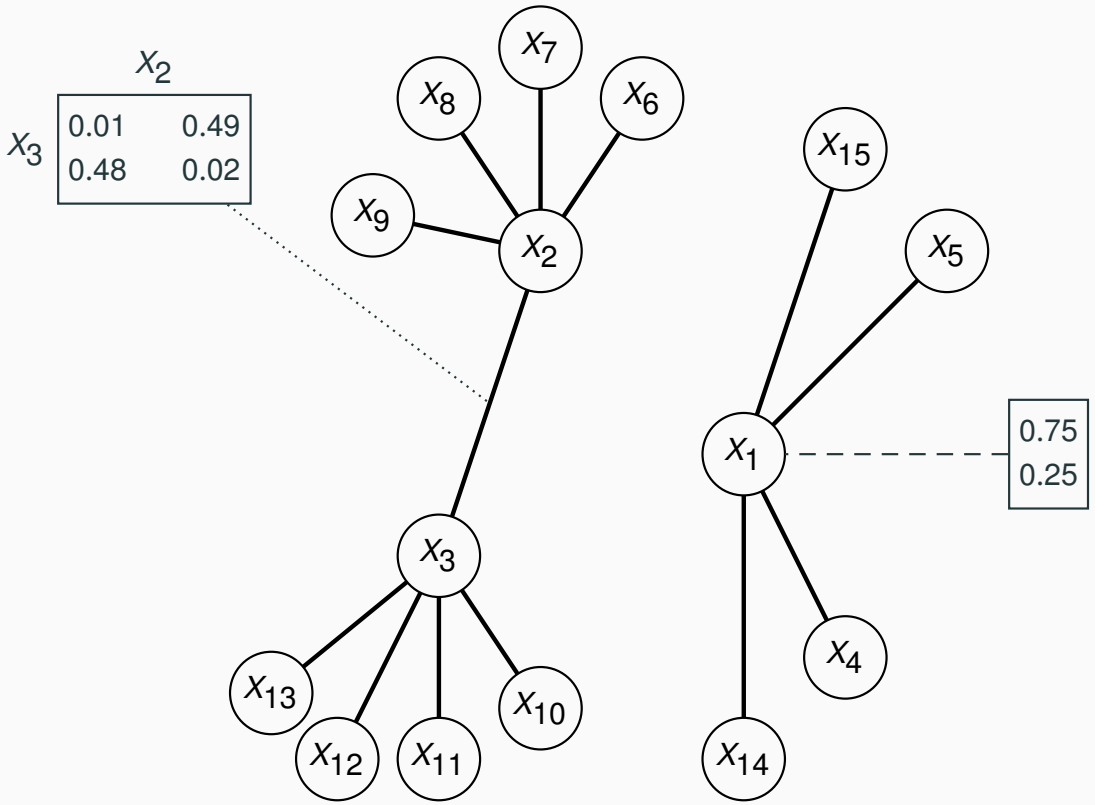
$$R(A,T) = \max_{\boldsymbol{x}^{1:T} \in \mathscr{X}^T} \left[\sum_{t=1}^{T} \ell(M^t, \boldsymbol{x}^t) - \min_{M \in \mathscr{M}} \sum_{t=1}^{T} \ell(M, \boldsymbol{x}^t) \right]$$

Per-round complexity

Given by the amount of computational resources spent by A at each trial t, for choosing a model M^t in $\mathscr{M}$, and evaluating its log-loss $\ell\left(M^t, \boldsymbol{x}^t\right) = -\ln \mathbb{P}_{M^t}(\boldsymbol{x}^t)$.

Outline

Markov Forests



For a set of n random variables defined over the discrete domain $\{0, 1, \cdots, m-1\}$, the class of (m-ary n-dimensional) Markov Forests is given by the product $\mathscr{F}_{m,n} = \boldsymbol{F}_n \times \boldsymbol{\Theta}_{m,n}$, where

- $\boldsymbol{F}_n$ is the space of all acyclic graphs of order n;
- $\boldsymbol{\Theta}_{m,n}$ is the space of all parameter vectors mapping
  - a probability table $\boldsymbol{\theta}_i \subseteq [0,1]^m$ to each candidate node i, and
  - a probability table $\boldsymbol{\theta}_{ij} \subseteq [0,1]^{m \times m}$ to each candidate edge (i, j).

Two key properties

For the class of Markov forests,

The probability distribution associated with a Markov forest $M = (f, \boldsymbol{\theta})$ can be factorized into a closed-form:

$$\mathbb{P}_M(\boldsymbol{x}) = \prod_{i=1}^{n} \theta_i(x_i) \prod_{(i,j)\in\binom{n}{2}} \left(\frac{\theta_{ij}(x_i, x_j)}{\theta_i(x_i)\theta_j(x_j)} \right)^{f_{ij}}$$

So, the log-loss extended to $\operatorname{conv} \boldsymbol{F}_n \times \boldsymbol{\Theta}_{m,n}$ is an affine function of the structure:

$$\ell(\boldsymbol{p}, \boldsymbol{\theta}, \boldsymbol{x}) = \psi(\boldsymbol{x}) + \langle \boldsymbol{p}, \phi(\boldsymbol{x}) \rangle, \text{where } \psi(\boldsymbol{x}) = \sum_{i\in[n]} \ln \frac{1}{\theta_i(x_i)} \text{ and } \phi_{ij}(x_i, x_j) = \ln \left(\frac{\theta_i(x_i)\theta_j(x_j)}{\theta_{ij}(x_i, x_j)} \right)$$

The space $\boldsymbol{F}_n$ of forest structures is a matroid; minimizing a linear function over $\boldsymbol{F}_n$ can be done in quadratic time using the matroid greedy algorithm.

The Algorithm

set $\boldsymbol{\theta}^1 = \mathscr{U}_{m,n}$

set $\boldsymbol{p}^1 = \mathbf{0}$

} Initialization

for each trial $t = 1, \cdots, T$

play $M^t = (\boldsymbol{f}^t, \theta^t)$, where $\boldsymbol{f}^t = \textsc{Swap}_1(\boldsymbol{p}^t)$

receive $\boldsymbol{x}^t$

set $\theta_i^{t+1}(u) = \dfrac{t_u + 1/2}{t + m/2}$ for all n nodes

set $\theta_{ij}^{t+1}(u,v) = \dfrac{t_{uv} + 1/2}{t + m^2/2}$ for all $\binom{n}{2}$ possible edges

} Parameter Update (Jeffreys rule)

draw $\boldsymbol{r}_t$ in $\left[0, \frac{1}{\beta_t}\right]^{\binom{n}{2}}$ uniformly at random

set $\boldsymbol{f}^{t+\frac{1}{2}} = \operatorname{argmin}_{\boldsymbol{f} \in \boldsymbol{F}_n} \langle \boldsymbol{f}, \boldsymbol{r}^t + \sum_{s=1}^t \phi^s(\boldsymbol{x}^s) \rangle$

set $\boldsymbol{p}^{t+\frac{1}{2}} = \alpha_t \boldsymbol{p}^t + (1 - \alpha_t) \boldsymbol{f}^{t+\frac{1}{2}}$

set $\boldsymbol{p}^{t+1} = \textsc{Swap}_k\left(\boldsymbol{p}^{t+\frac{1}{2}}\right)$

} Structure Update (Mixture of perturbed leaders)

$\textsc{Swap}_k$ uses random base exchanges to derive a convex mixture with at most k components.

Analysis

Based on the closed-form expression of Markov forests, parameter updates and structure updates can be analyzed in an independent way:

$$R\left(M^{1:T}, \boldsymbol{x}^{1:T}\right) = R\left(\boldsymbol{p}^{1:T}, \boldsymbol{x}^{1:T}\right) + R\left(\boldsymbol{\theta}^{1:T}, \boldsymbol{x}^{1:T}\right)$$

where

$$R\left(\boldsymbol{p}^{1:T}, x^{1:T}\right) = \sum_{t=1}^{T} \ell(\boldsymbol{p}^t, \boldsymbol{\theta}^t, \boldsymbol{x}^t) - \ell(\boldsymbol{p}^*, \boldsymbol{\theta}^t, \boldsymbol{x}^t) \qquad \text{(Structural Regret)}$$

$$R\left(\boldsymbol{\theta}^{1:T}, x^{1:T}\right) = \sum_{t=1}^{T} \ell(\boldsymbol{p}^*, \boldsymbol{\theta}^t, \boldsymbol{x}^t) - \ell(\boldsymbol{p}^*, \boldsymbol{\theta}^*, \boldsymbol{x}^t) \qquad \text{(Parametric Regret)}$$

Parametric Regret

Decomposable into local regrets, which can be bounded using universal coding techniques:

$$R\left(\boldsymbol{\theta}^{1:T}, x^{1:T}\right) = \sum_{i=1}^{n} \ln \frac{\theta_i^*(x_i^{1:T})}{\theta_i^{1:T}(x_i^{1:T})} \qquad \text{Univariate estimators}$$

$$+ \sum_{(i,j)\in F} \ln \frac{\theta_{ij}^*(x_{ij}^{1:T})}{\theta_{ij}^{1:T}(x_{ij}^{1:T})} \qquad \text{Bivariate estimators}$$

$$+ \sum_{(i,j)\in F} \ln \frac{\theta_i^{1:T}(x_i^{1:T})}{\theta_i^*(x_i^{1:T})} \frac{\theta_j^{1:T}(x_j^{1:T})}{\theta_j^*(x_j^{1:T})} \qquad \text{Bivariate compensation}$$

Using symmetric Dirichlet mixtures for the parametric estimators,

$$\theta^{1:T}(x^{1:T}) = \int \prod_{t=1}^{T} \mathbb{P}_{\boldsymbol{\lambda}}(x^t) p_\mu(\boldsymbol{\lambda}) d\boldsymbol{\lambda} = \frac{\Gamma(m\mu)}{\Gamma(\mu)^m} \frac{\prod_{v=1}^{m} \Gamma(t_v + \mu)}{\Gamma(t + m\mu)}$$

the parametric regret for $\mu = \frac{1}{2}$ (Jeffreys mixture) is in $O(\ln T)$. The per-round time complexity for parameter updates is in $O(m^2 n^2)$.

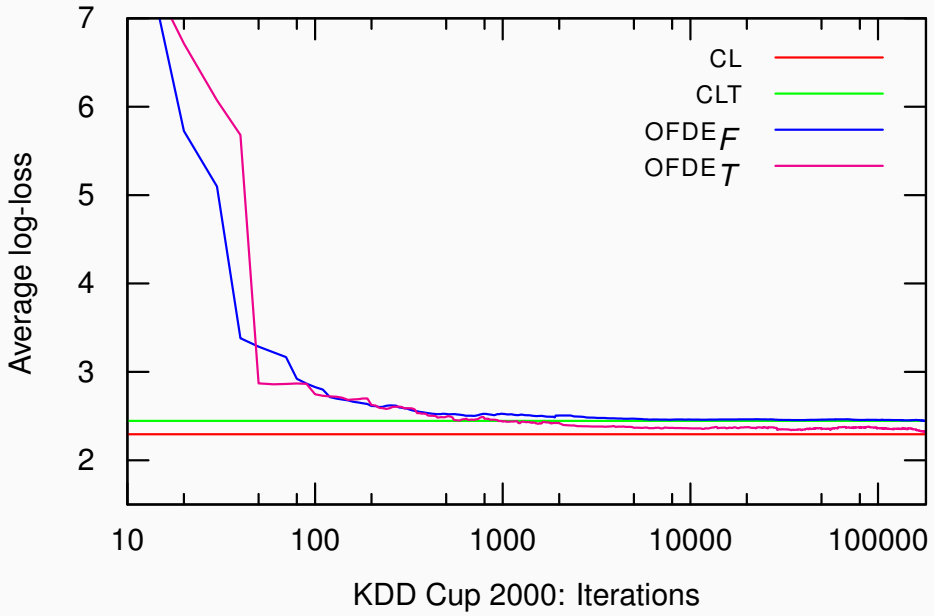
Structural Regret

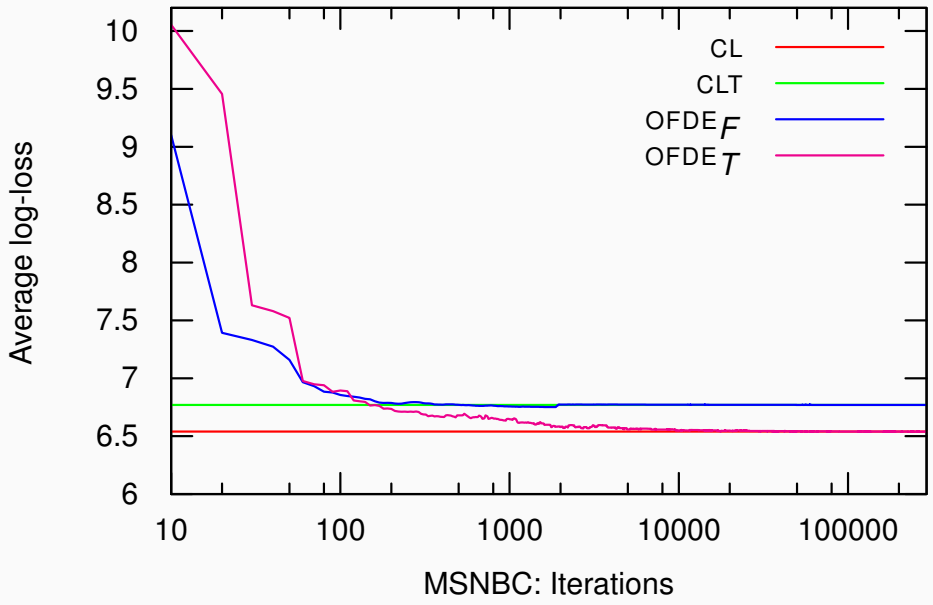
Based on the telescopic decomposition (and using $\ell^t = \phi^t(\boldsymbol{x}^t)$),

$$
\begin{aligned}
R\left(\boldsymbol{p}^{1:T}, x^{1:T}\right) = \sum_{t=1}^{T} \langle \boldsymbol{p}^t, \ell^t \rangle - \langle \boldsymbol{p}^{t+\frac{1}{2}}, \ell^t \rangle & \qquad \leq 0 \\
+ \sum_{t=1}^{T} \langle \boldsymbol{p}^{t+\frac{1}{2}}, \ell^t \rangle - \langle \boldsymbol{f}^{t+\frac{1}{2}}, \ell^t \rangle & \qquad \text{Convex mixture} \\
+ \sum_{t=1}^{T} \langle \boldsymbol{f}^{t+\frac{1}{2}}, \ell^t \rangle - \langle \boldsymbol{p}^{*}, \ell^t \rangle & \qquad \text{Follow the Perturbed Leader (Kalai and Vempala, 2005)}
\end{aligned}
$$

the structural regret is in $O(\sqrt{T} \ln T)$. The per-round time complexity for structure updates is in in $O(n^2 \log n + kn^2)$.

Preliminary Experiments





The OFDE algorithm (with F for forests, and T for trees, $k = \ln n$) was compared to batch algorithms (Chow-Liu (1968) for trees, and Chow-Liu with Thresholding (2011) for forests), which had the benefit of hindsight for the train set.

The average log-loss was measured on the test set at the end of each iteration.

Outline

Conclusions

Online density estimation has very attractive properties

- Designed for adversarial environments
- Naturally suited to streaming applications
- Can be applied to large-scale applications with massive amounts of data

Online graphical density estimation is challenging

We are faced with a tradeoff between minimax optimality and computational complexity:

- Minimax optimality often requires super-exponential time.
- Online approximation algorithms (Kakade et al., 2009) look promising for handling more expressive graphical models (ex: polytrees, bounded treewidth networks).

Thank You!

(and thanks to the anonymous reviewers for helping to improve this paper!)

References

C. K. Chow and C. N. Liu. Approximating discrete probability distributions with dependence trees. *IEEE Transactions on Information Theory*, 14(3):462–467, 1968.

S. Kakade, A. Kalai, and K. Ligett. Playing games with approximation algorithms. *SIAM J. Comput.*, 39(3): 1088–1106, 2009.

A. Kalai and S. Vempala. Efficient algorithms for online decision problems. *Journal of Computer and System Sciences*, 71(3):291–307, 2005.

W. Kotłowski and P. Grünwald. Maximum likelihood vs. sequential normalized maximum likelihood in on-line density estimation. In *Proc. of COLT*, pages 457–476, 2011.

E. Takimoto and M. Warmuth. The last-step minimax algorithm. In *Proc. of ALT*, pages 279–290, 2000.

V. Tan, A. Anandkumar, and A. Willsky. Learning high-dimensional Markov forest distributions: Analysis of error rates. *Journal of Machine Learning Research*, 12:1617–1653, 2011.

Q. Xie and A. Barron. Asymptotic minimax regret for data compression, gambling, and prediction. *IEEE Transactions on Information Theory*, 46(2):431–445, 2000.