Fast Non-Parametric Bayesian Inference on Infinite Trees

Marcus Hutter

Istituto Dalle Molle di Studi sull'Intelligenza Artificiale
IDSIA, Galleria 2, CH-6928 Manno-Lugano, Switzerland
marcus@idsia.ch, http://www.idsia.ch/~marcus

AISTATS-2005, January 6 – 8

Abstract

Given i.i.d. data from an unknown distribution, we consider the problem of predicting future items. An adaptive way to estimate the probability density is to recursively subdivide the domain to an appropriate data-dependent granularity. A Bayesian would assign a data-independent prior probability to "subdivide", which leads to a prior over infinite(ly many) trees. We derive an exact, fast, and simple inference algorithm for such a prior, for the data evidence, the predictive distribution, the effective model dimension, and other quantities. We illustrate the behavior of our model on some prototypical functions.

Keywords

Bayesian density estimation, exact linear time algorithm, non-parametric inference, adaptive infinite tree, Polya tree, scale invariance

Table of Contents

Inference

Given: i.i.d. data D sampled from unknown distribution q.

Goal: Infer/estimate probability density q from data D

$\Longrightarrow$ All other quantities of interest can be derived

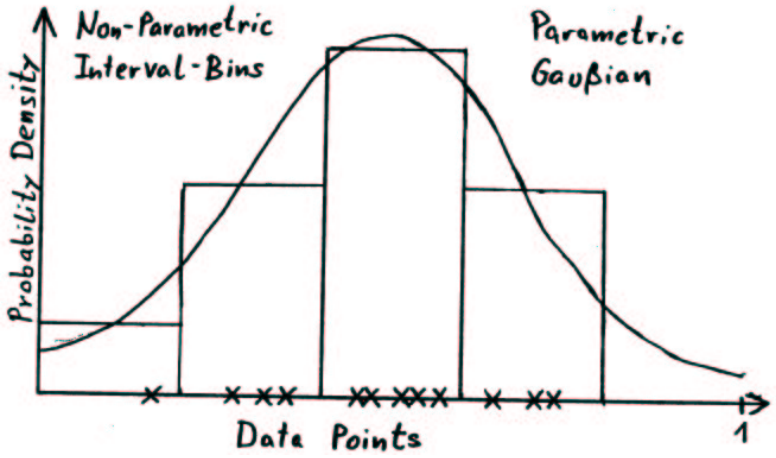
Many methods ...

(Non)Parametric Estimation

- Density q assumed to belong to an (in)finite-dimensional family. (e.g. family of Gaussians parameterized by mean and (co)variance)
- Maximum Likelihood (ML) estimate (can overfit if the family is large)
- Penalize complex distributions by assigning a prior (2nd order) probability to the densities q.
- Maximize the model posterior (MAP $\approx$ MDL $\approx$ MML)
- Bayesians keep the complete posterior for inference (MAP can fail while Bayes works [PH'04])
- How to choose the prior?

Interval-bins: Frequency Estimate

Most simple non-parametric model class



Drawbacks:

- Distributions are discontinuous
- Restricted to one (or low) dimension
- Uniform (or fixed or heuristic) discretization
- Heuristic choice of the number of bins.

We present a full Bayesian solution/improvement to most these problems

Setup and Basic Quantities of Interest

- Given: i.i.d. data $D = (x^1, ..., x^n) \in \Gamma^n$ from domain Γ, sampled from unknown probability density $q : \Gamma \to I\!R$.
- Standard inference: Estimate q from D -or- predict next $x^{n+1} \in \Gamma$.
- Data likelihood under model q is $p(D|q) \equiv q(x_1) \cdot ... \cdot q(x_n)$
- Assume prior $p(q)$ over models $q \in Q$
- Data evidence: $p(D) = \int_Q p(D|q)p(q)dq$

  $\Rightarrow$ posterior: $p(q|D) = p(D|q)p(q)/p(D)$ from Bayes' rule

  $\Rightarrow$ Predictive distribution: $p(x|D) = p(D,x)/p(D)$

  $\Rightarrow$ Expected q-prob. of x: $E[q(x)|D] := \int q(x)p(q|D)dq = p(x|D)$

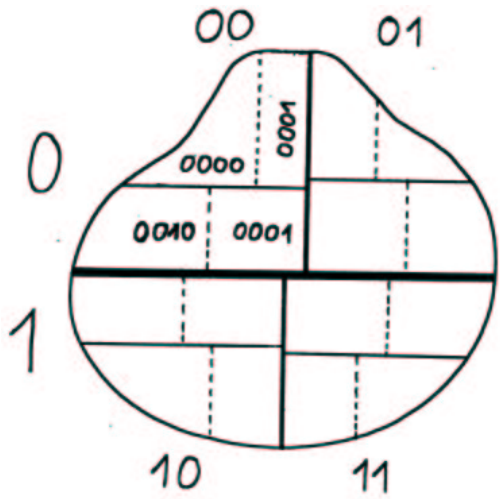
  $\Rightarrow$ Similarly for (co)variances

Hierarchical Tree Partitioning

Recursively (sub)partition $\Gamma_z = \Gamma_{z0} \dot\cup \Gamma_{z1}$ for $z \in \mathbb{B}^*$

$\Gamma_\epsilon = \Gamma$, where ϵ is the empty string.

Examples for Γ: Interval $[0,1)$, tree, volume,
finite strings $\mathbb{B}^*$ or infinite sequences $\mathbb{B}^\infty$, ...



Classification: Γ = class+feature-space

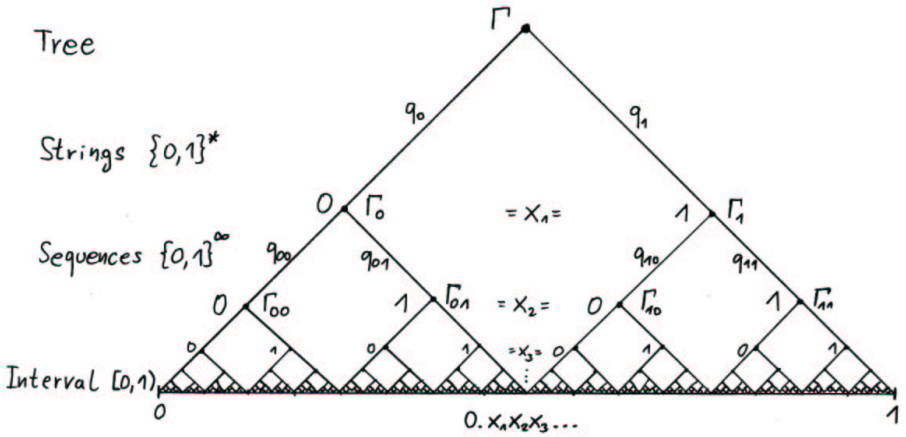
$z_1 \in \mathbb{B}$ is class-label

$z_2 \in \mathbb{B}$ is most important feature

$z_3 \in \mathbb{B}$ is 2nd most important feature

...

or independent real-valued features.



The New Tree Mixture Model

- Probability of branching left in tree-node Γ_z is $q_{z0} := P[\Gamma_{z0}|\Gamma_z, q]$
- All branching probabilities: $\vec{q}_{z*} := (q_{zy} : y \in \mathbb{B}^* \setminus \{\epsilon\})$
- The prior $p(q)$ follows from specifying a prior over $\vec{q}_*$, since $q(x) \propto q_{x_1} \cdot q_{x_1x_2} \cdot q_{x_1x_2x_3} ...$ (chain rule)

a) With probability $\frac{1}{2}$, choose q uniform on Γ_z .

b) With probability $\frac{1}{2}$, split Γ_z into two the parts Γ_{z0} and Γ_{z1}, and assign recursively a prior to each part, i.e. in each part again either uniform or split, etc.

$$p(\vec{q}_{z*}) = \overbrace{\tfrac{1}{2} \prod_{y \in \mathbb{B}^* \setminus \{\epsilon\}} \delta(q_{zy} - \tfrac{1}{2})}^{\text{uniform}} + \overbrace{\tfrac{1}{2}\delta(q_{z0}+q_{z1}-1) \underbrace{p(\vec{q}_{z0*})p(\vec{q}_{z1*})}_{\text{recursion}}}^{\text{split}}$$

Properties of the Tree Mixture Model

- Any probability measure q can be represented
- Scale invariance
- Symmetry
- Continuous predictive probability density for $n \to \infty$
- No tunable parameters (or at most two)

None of these desirable properties would satisfied for a finite tree model!

The Evidence Recursion

Notation: $D_z := \{x \in D : x \in \Gamma_z\}$, and $n_z := |D_z|$,
$p_z \propto p$ restricted to Γ_z, and $\Delta_z := \frac{n_{z0}}{n_z} - \frac{1}{2}$.

$$p_z(D_z) = \int p_z(D_z|\vec{q}_{z*})p(\vec{q}_{z*})d\vec{q}_{z*} \stackrel{\cdots}{=} \underbrace{\tfrac{1}{2}\big[1}_{\text{uniform}} + \underbrace{\frac{p_{z0}(D_{z0})p_{z1}(D_{z1})}{w_{n_z}(\Delta_z)}\big]}_{\text{split}}$$

$$w_{n_z}(\Delta_z) = 2^{-n_z}\frac{(n_z+1)!}{n_{z0}!n_{z1}!} \approx \begin{cases} \Theta(\sqrt{n_z}) \stackrel{n_z\to\infty}{\longrightarrow} \infty & \text{if} \quad \dot{q}_{z0} = \dot{q}_{z1}, \\ \mathrm{e}^{-\Theta(n_z)} \stackrel{n_z\to\infty}{\longrightarrow} 0 & \text{if} \quad \dot{q}_{z0} \neq \dot{q}_{z1}. \end{cases}$$

Weight w_z is large/small for uniform/non-uniform $q_z()$,
correctly causing uniform/split.

Asymptotic convergence/consistency ($n \to \infty$)

- posterior $p_z(\vec{q}_{z*}|D)$ concentrates around the true distribution $\dot{\vec{q}}_{z*}$ for $n \to \infty$.

  $\Rightarrow$ posterior $p_z(x|D_z) \to \dot{q}_z(x)$ for all $x \in \Gamma_z$.

- Evidence $p_z(D_z) \to const.$ for uniform $\dot{q}_z()$,
  and increases exponentially with n_z for non-uniform $\dot{q}_z()$.

Model Dimension and Number of Bins

Effective model dimension $N_{\vec{q}_{z*}} = \#\{q \in \vec{q}_{z*} : q \neq \frac{1}{2}\}$ of $\vec{q}_{z*}$ can be given recursively as

$$N_{\vec{q}_{z*}} = \begin{cases} 0 & \text{if} \quad q_{z0} = \frac{1}{2} \\ 1 + N_{\vec{q}_{z0*}} + N_{\vec{q}_{z1*}} & \text{if} \quad q_{z0} \neq \frac{1}{2} \end{cases}$$

$$P_z[N_{\vec{q}_{z*}} = k+1|D_z] = g_z(D_z) \sum_{i=0}^{k} P_{z0}[N_{\vec{q}_{z0*}} = i|D_{z0}] \cdot P_{z1}[N_{\vec{q}_{z1*}} = k-i|D_{z1}],$$

Splitting probability: $g_z(D_z) := \frac{1}{2} \frac{p_{z0}(D_{z0}) p_{z1}(D_{z1})}{p_z(D_z) w(n_{z0}, n_{z1})} = 1 - \frac{1}{2 p_z(D_z)}$

Interpretation: The probability that Γ_z has dimension $k+1$ equals

- the posterior probability $g_z(D_z)$ of splitting Γ_z,
- times the probability that left subtree has dimension i,
- times the probability that right subtree has dimension $k-i$,
- summed over all possible i.

Number of bins $\equiv 1 +$ model dimension (due to probability constraint)

Tree Height and Bin Size

Effective tree height $\vec{q}_{z*}$ at $x \in \Gamma_z$ is given recursively as

$$h_{\vec{q}_{z*}}(x) \;=\; \begin{cases} 0 & \text{if} \quad q_{z0} = \frac{1}{2} \\ 1 + h_{\vec{q}_{zx_{l+1}*}}(x) & \text{if} \quad q_{z0} \neq \frac{1}{2} \end{cases}$$

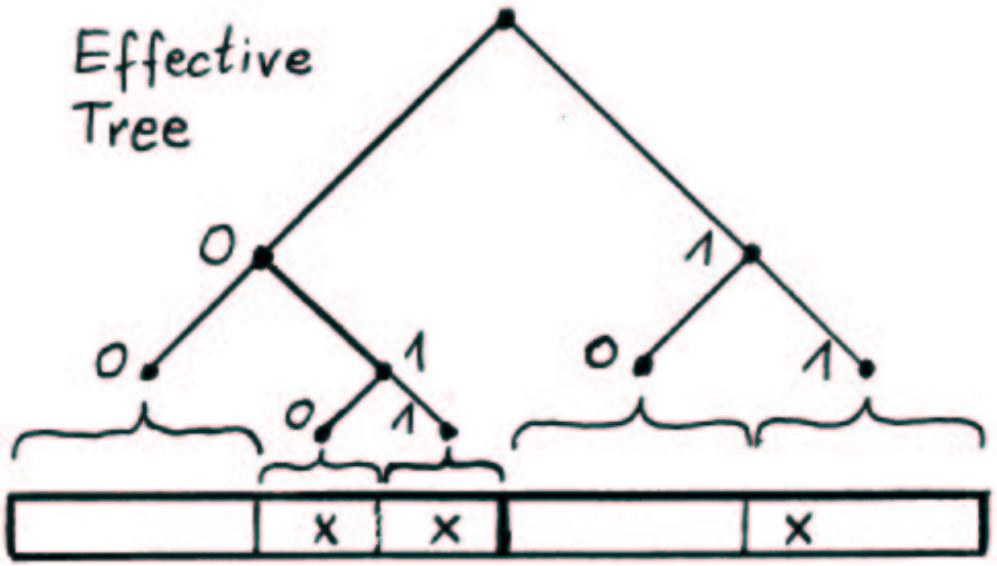
$$E_z[h_{\vec{q}_{z*}}(x)|D_z] \;=\; g_z(D_z)\Big[1 + E_{zx_{l+1}}[h_{\vec{q}_{zx_{l+1}*}}(x)|D_{zx_{l+1}}]$$

Average cell/bin size or volume: $v_{\vec{q}_*} = 2^{-\bar{h}_{\vec{q}_*}}$

Fast BayesTree Algorithm

- Recurse down the tree until $D_z = \phi$ is empty or $D_z = (x_i) \in \Gamma_z$ is a singleton (data separation level)
- Subdividing empty or singleton cells further, remain empty or singletons.

  $\Rightarrow$ Solve quadratic self-consistent equation
- which can be done analytically .



Solution of Self-Consistency Equations

- Evidence $p_z(\phi) = p_z(x_i) = 1$ (trivial)
- Effective tree height $E_z[h_{\vec{q}_{z*}}(x)|\phi \text{ or } x] = 1$ (easy)
- Effective model dimension $P_z[N_{\vec{q}_{z*}} = k|\phi \text{ or } x] = a_k$ with

$$a_{k+1} = \tfrac{1}{2}\sum_{i=0}^{k} a_i \cdot a_{k-i} \quad \text{with} \quad a_0 = \tfrac{1}{2}$$

$$a_k = \frac{1}{2(k+1)4^k}\binom{2k}{k} \sim \frac{1}{2\sqrt{\pi}}\, k^{-3/2}$$

- This is exactly how a proper non-informative prior on $\mathbb{N}$ should look like: as uniform as possible, i.e. slowly decreasing.
- Conclusion: Finite $O(n =$data size$)$ procedure for exactly computing all quantities of interest in infinite BayesTree model.

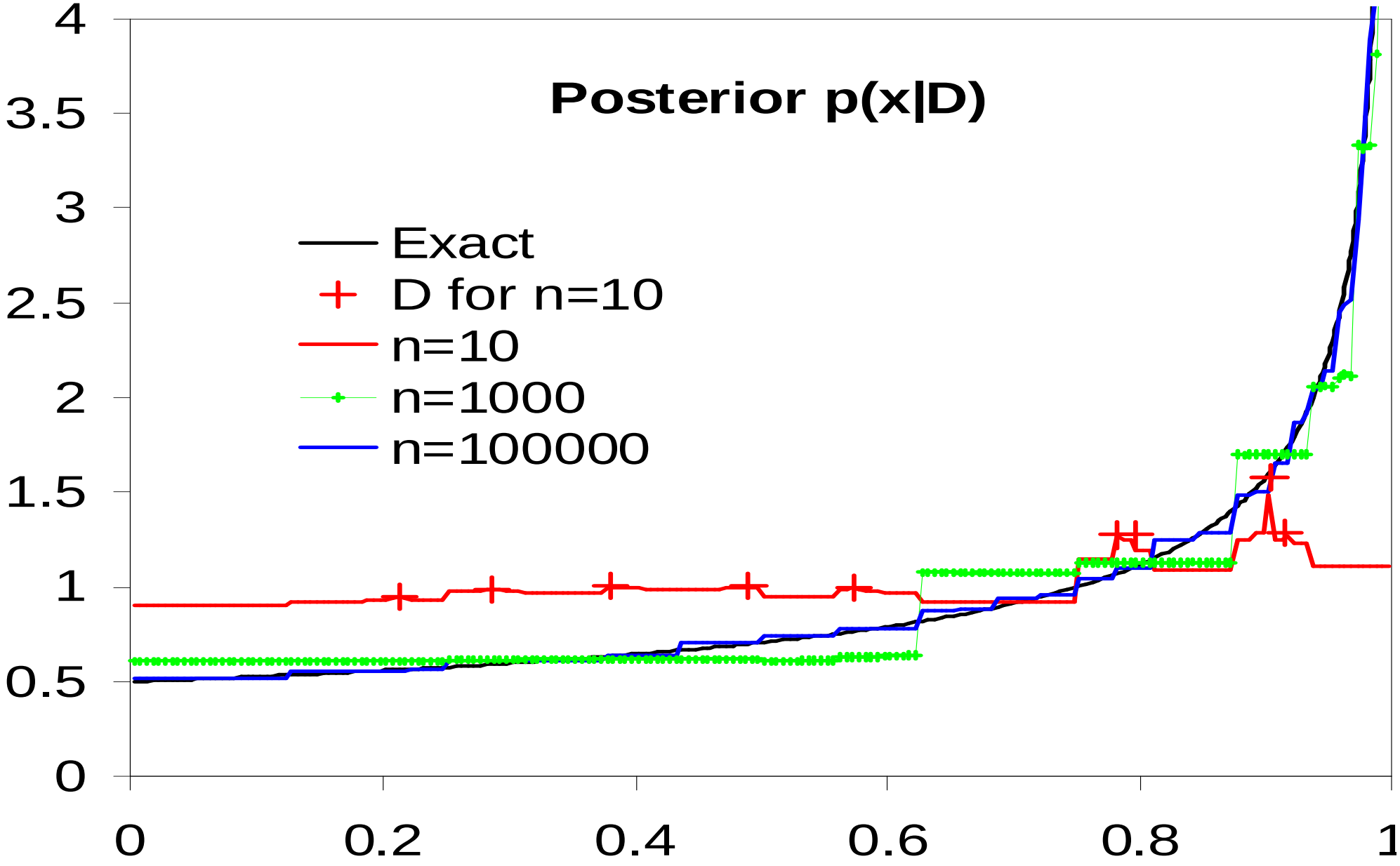
BayesTree($D[], n, x$) – Algorithm in Pseudo Code

⌈ if ($n \leq 1$ and ($n == 0$ or $D[0] == x$ or $x \notin [0,1)$))
  ⌈ if ($x \in [0,1)$) then $h = 1$; else $h = 0$;
  ⌊ $p = 1$; for($k = 0, .., N_{max}$) $\tilde{p}[k] = a_k$;
  else
  ⌈ $n0 = n1 = 0$;
    for($i = 0, .., n-1$)
    ⌈ if ($D[i] < \frac{1}{2}$) then[$D_0[n_0] = 2D[i]$; $n_0 = n_0 + 1$;]
    ⌊ else [$D_1[n_1] = 2D[i] - 1$; $n_1 = n_1 + 1$;]
    $(p_0, h_0, \tilde{p}_0[])$=BayesTree($D_0[], n_0, 2x$);
    $(p_1, h_1, \tilde{p}_1[])$=BayesTree($D_1[], n_1, 2x-1$);
    $p = \frac{1}{2}[1 + p_0 \cdot p_1 / \ln w(n_0, n_1)]$;
    $g = 1 - 1/2p$;
    if ($x \in [0,1)$) then $h = g \cdot (1 + h_0 + h_1)$; else $h = 0$;
    $\tilde{p}[0] = 1 - g$;
  ⌊ for($k = 0, .., N_{max}$) $\tilde{p}[k+1] = g \cdot \sum_{i=0}^{k} \tilde{p}_0[i] \cdot \tilde{p}_1[k-i]$;
⌊ **return $(p, h, \tilde{p}[])$;**

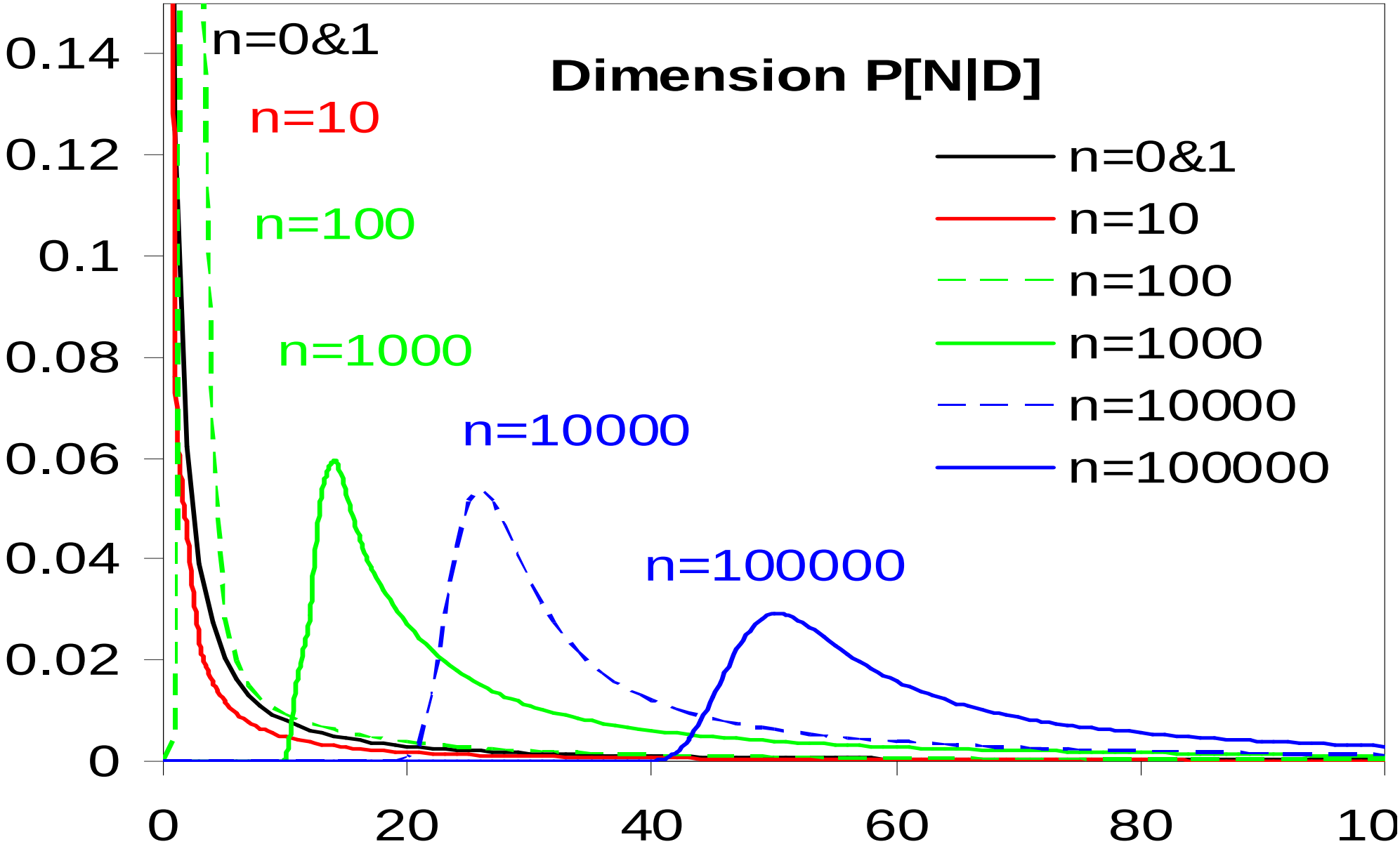
Numerical Examples: General observations

- The posteriors $p(x|D)$ clearly converge for $n \to \infty$ to the true distribution $\dot{q}()$,
- accompanied by a (necessary) moderate growth of the effective dimension (except for Jump-at-1/2).
- For $n = 10$ we show the data points. It is visible how each data point pulls the posterior up, as it should be ("one sample seldom comes alone").
- Optimal bin-number $O(n^{1/3})$ is nicely consistent with the BayesTree model dimension.
- The expected tree height $E[h(x)|D]$ at x correctly reflects the local needs for (non)splits.

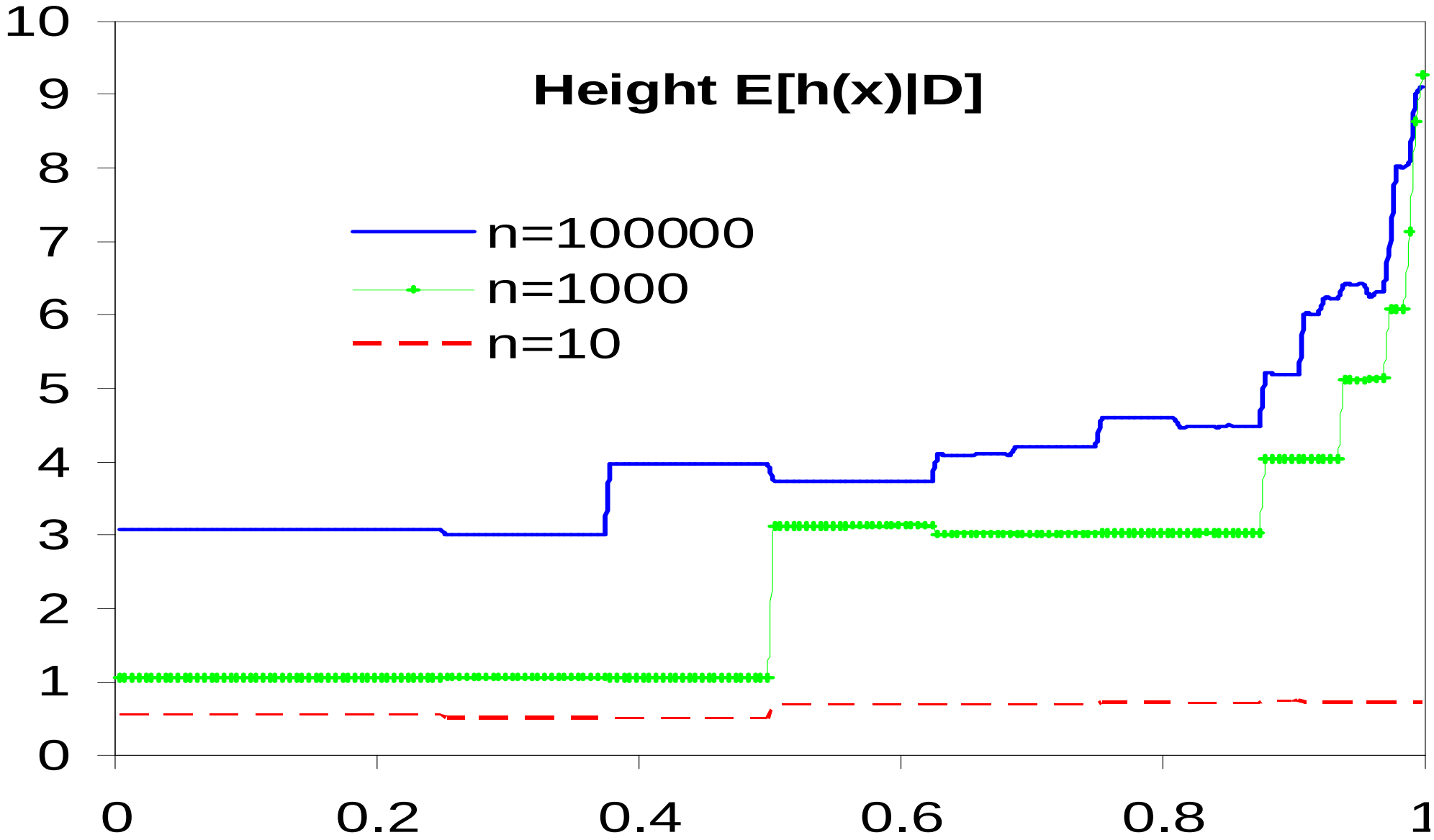
Singular Distribution $\dot{q}(x) = 2/\sqrt{1-x}$



Singular Distribution $\dot{q}(x) = 2/\sqrt{1-x}$

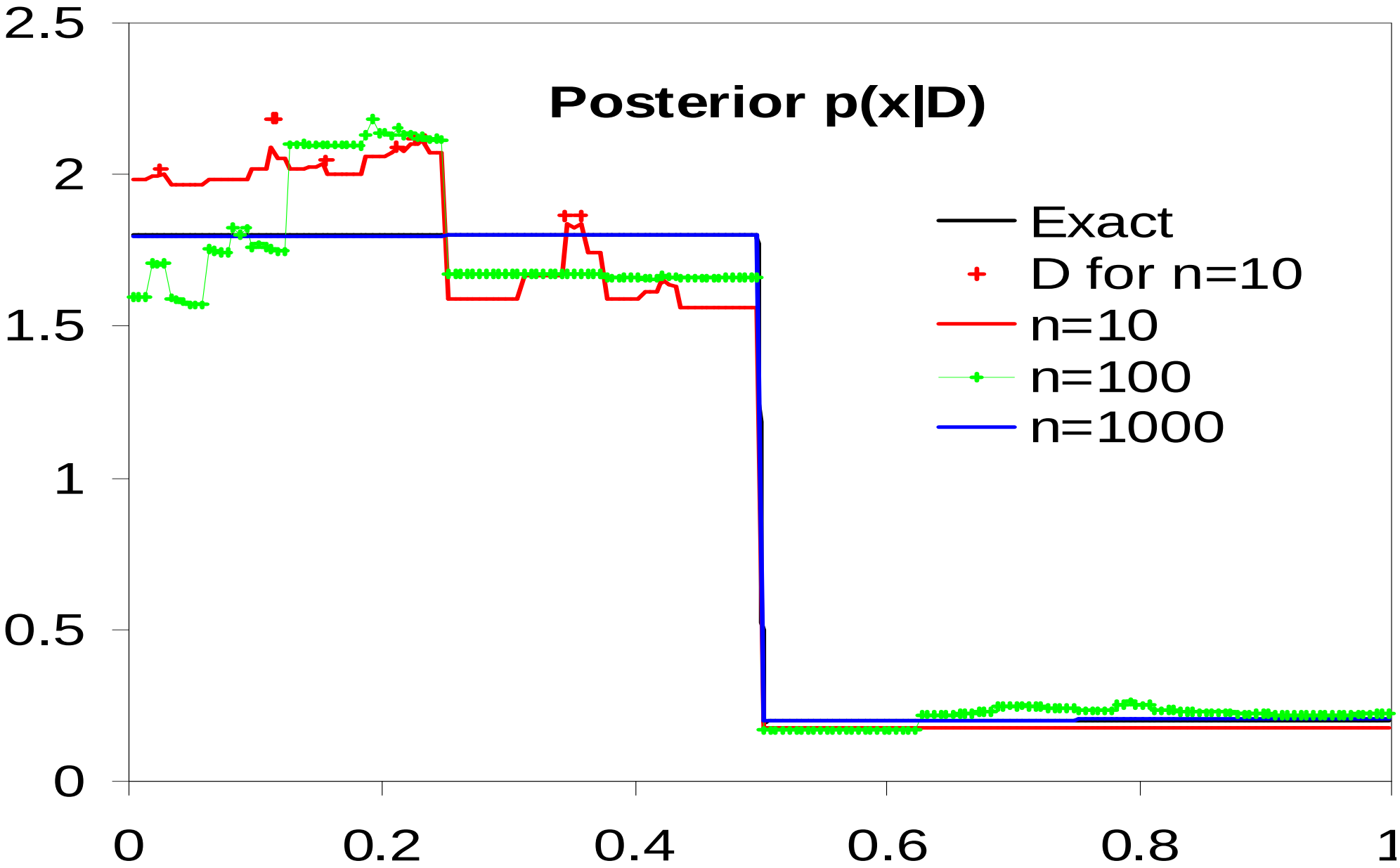


Singular Distribution $\dot{q}(x) = 2/\sqrt{1-x}$

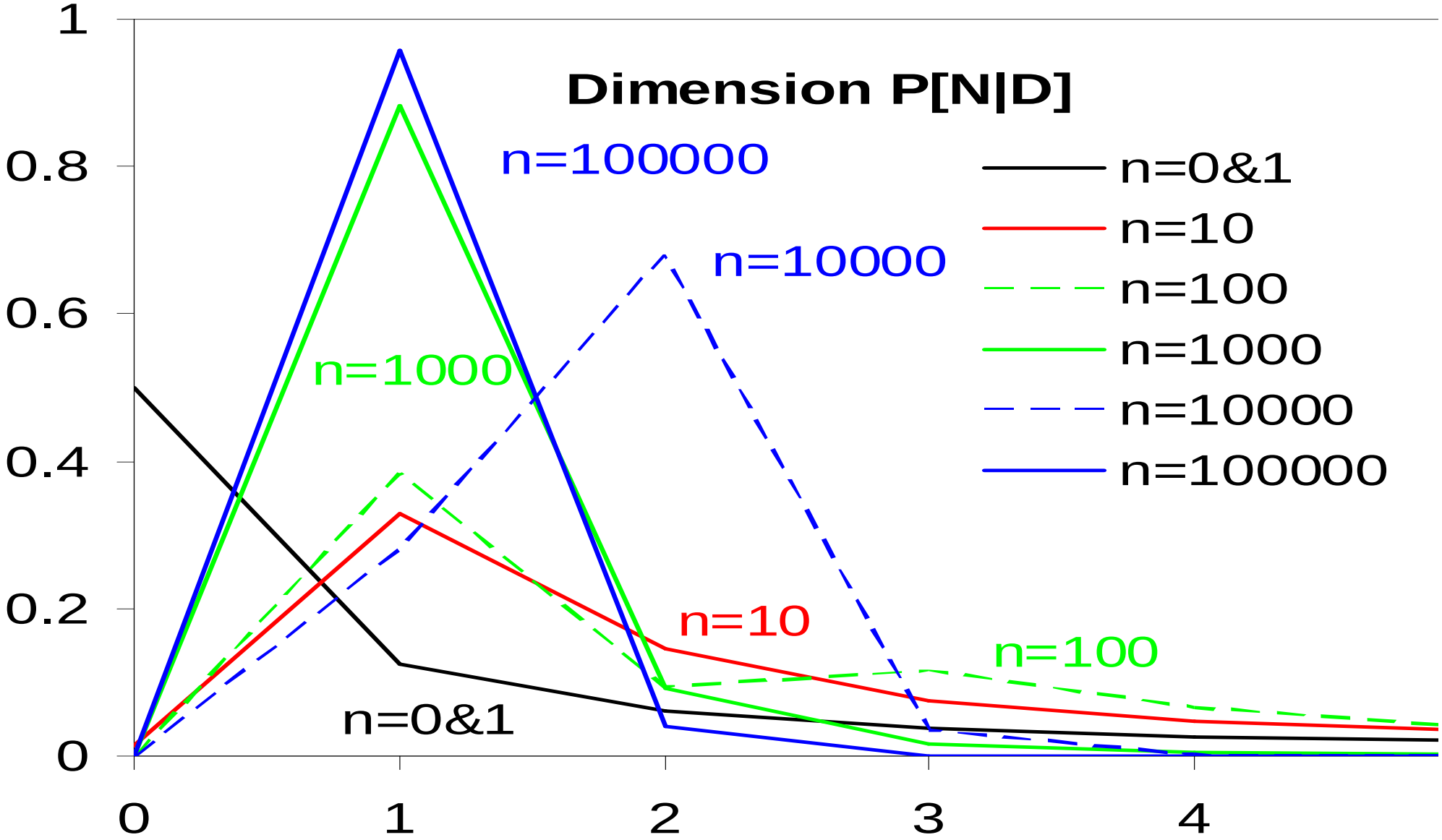


Tree height is necessarily larger near the singularity at $x = 1$.

Jump-at-1/2 Distribution: Finite Bayes Tree

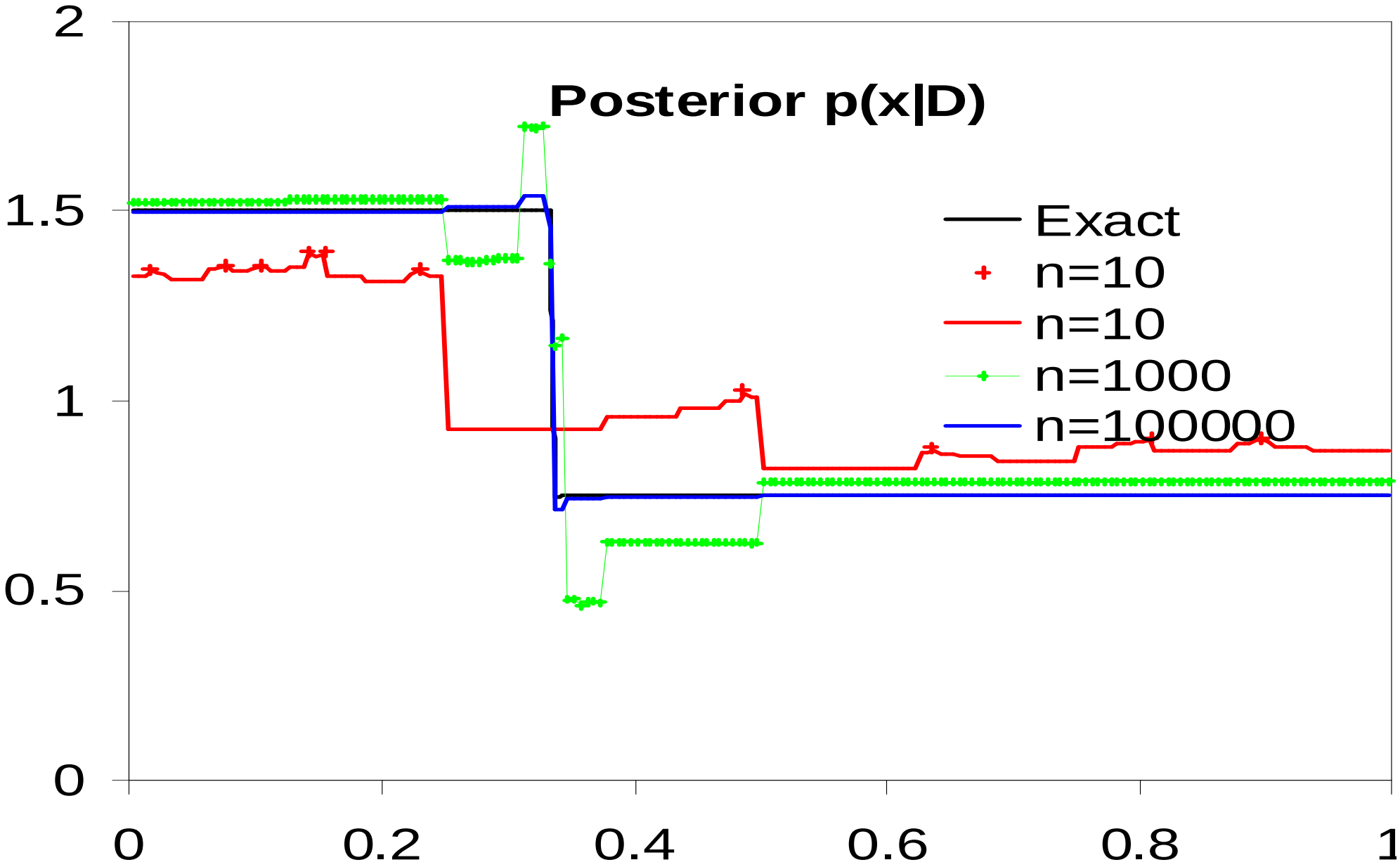


Jump-at-1/2 Distribution: Finite Bayes Tree

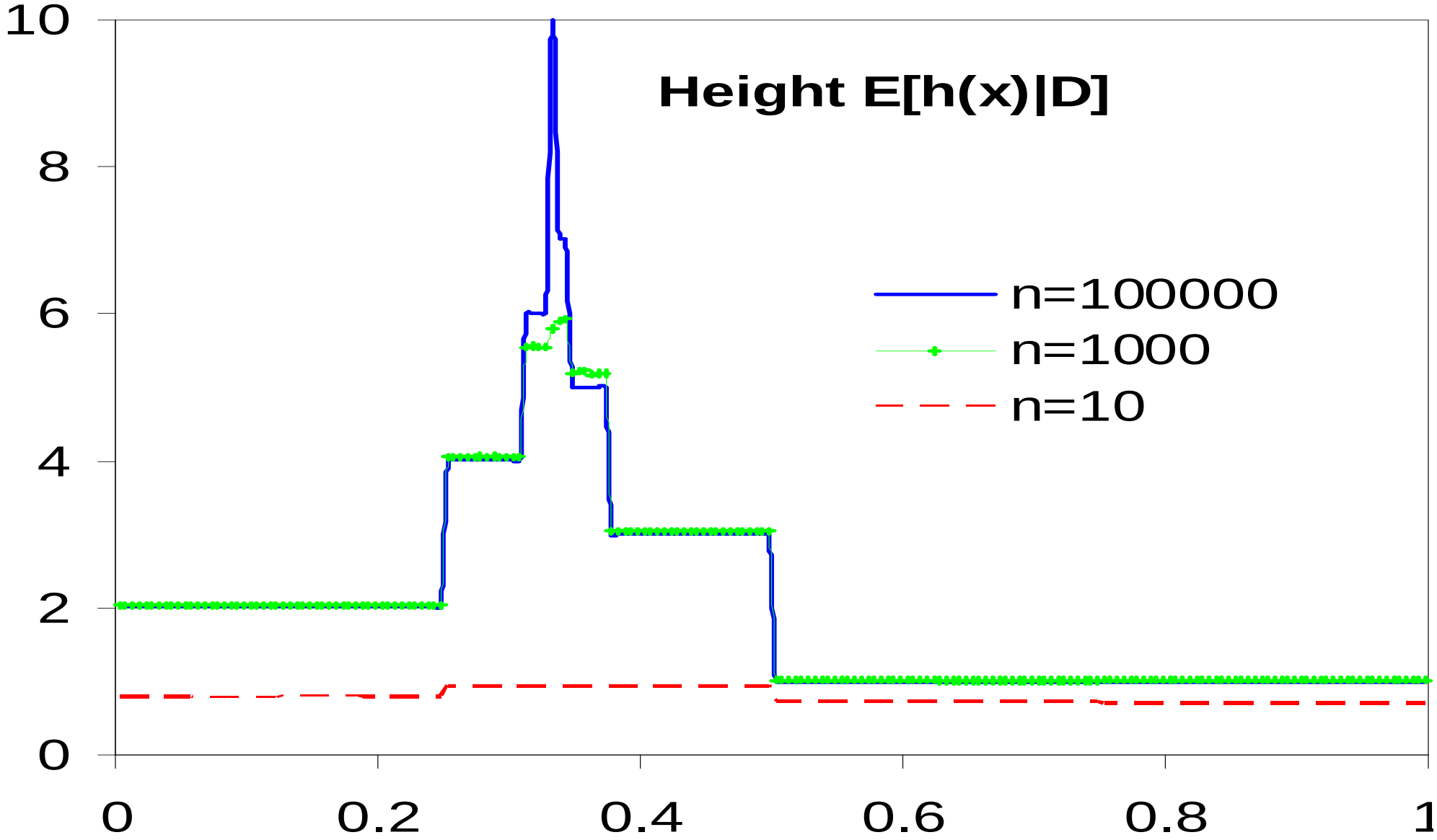


Finite model dimension and tree height. All quantities converge rapidly.

Jump-at-1/3 Distribution

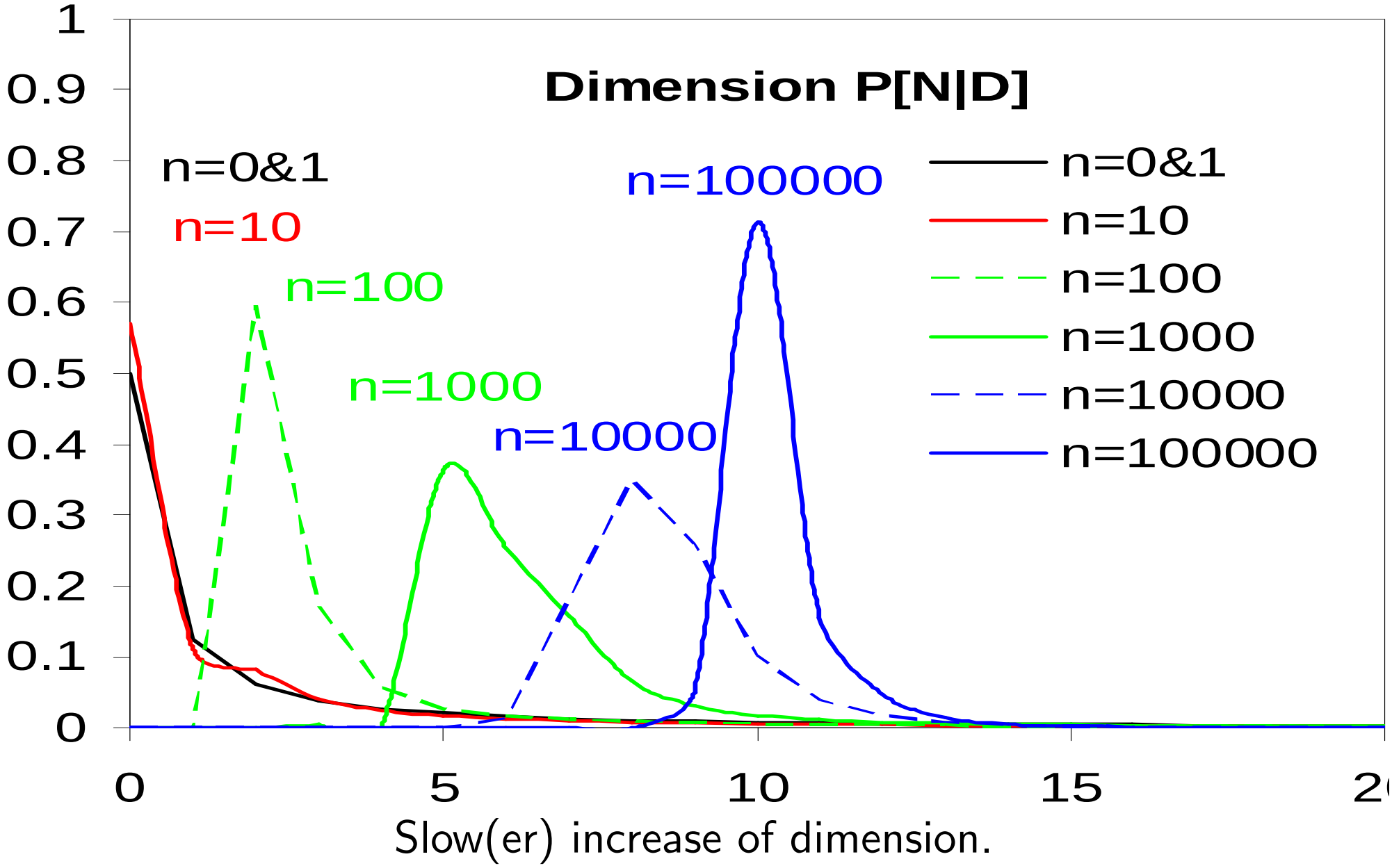


Jump-at-1/3 Distribution



Only one branch of the tree has to grow to infinity. Singularity at 1/3.

Jump-at-1/3 Distribution



Extensions

- Multi-Points ($x_i \equiv x_j$): Important for higher moments
  $\Rightarrow$ New interesting phenomena!
- Splitting probability $\neq \frac{1}{2} \Rightarrow$ New interesting phenomena!
- Uniform prior over branching mass q_{z0} could be generalized to a Dirichlet distribution $\Rightarrow$ Allows informative prior.
- Expected entropy can be computed similar to [WW'96,H'01]
- A sort of maximum a posteriori (MAP) tree skeleton can also easily be extracted.

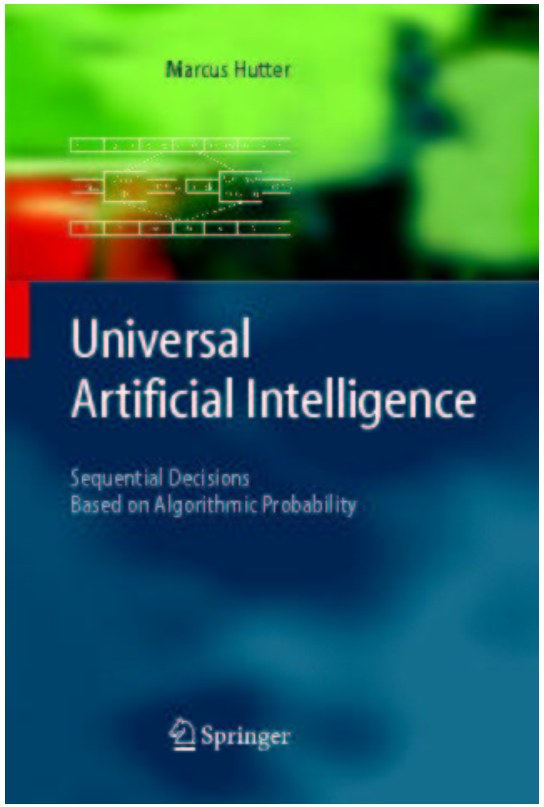
Summary

- We presented a Bayesian model on infinite trees, where we split a node into two subtrees with prior probability $\frac{1}{2}$, and uniform choice of the probability assigned to each subtree.
- We devised closed form expressions for various inferential quantities of interest at the data separation level, which led to an exact algorithm with runtime essentially linear in the data size.
- The theoretical and numerical model behavior was very reasonable, e.g. consistency (no underfitting) and finite effective dimension (no overfitting).
- A challenge is to generalize the model from piecewise constant to piecewise linear continuous functions

Thanks! Questions? Details:

Jobs: PostDoc and PhD positions at IDSIA, Switzerland

Projects at http://www.idsia.ch/~marcus



A Unified View of Artificial Intelligence

$$
\begin{array}{ccc}
= & & = \\
\text{Decision Theory} & = & \text{Probability} + \text{Utility Theory} \\
+ & & + \\
\text{Universal Induction} & = & \text{Ockham} + \text{Bayes} + \text{Turing}
\end{array}
$$

Open research problems at www.idsia.ch/~marcus/ai/uaibook.htm