

Multivariate density estimation and its applications

Wing Hung Wong
Stanford University

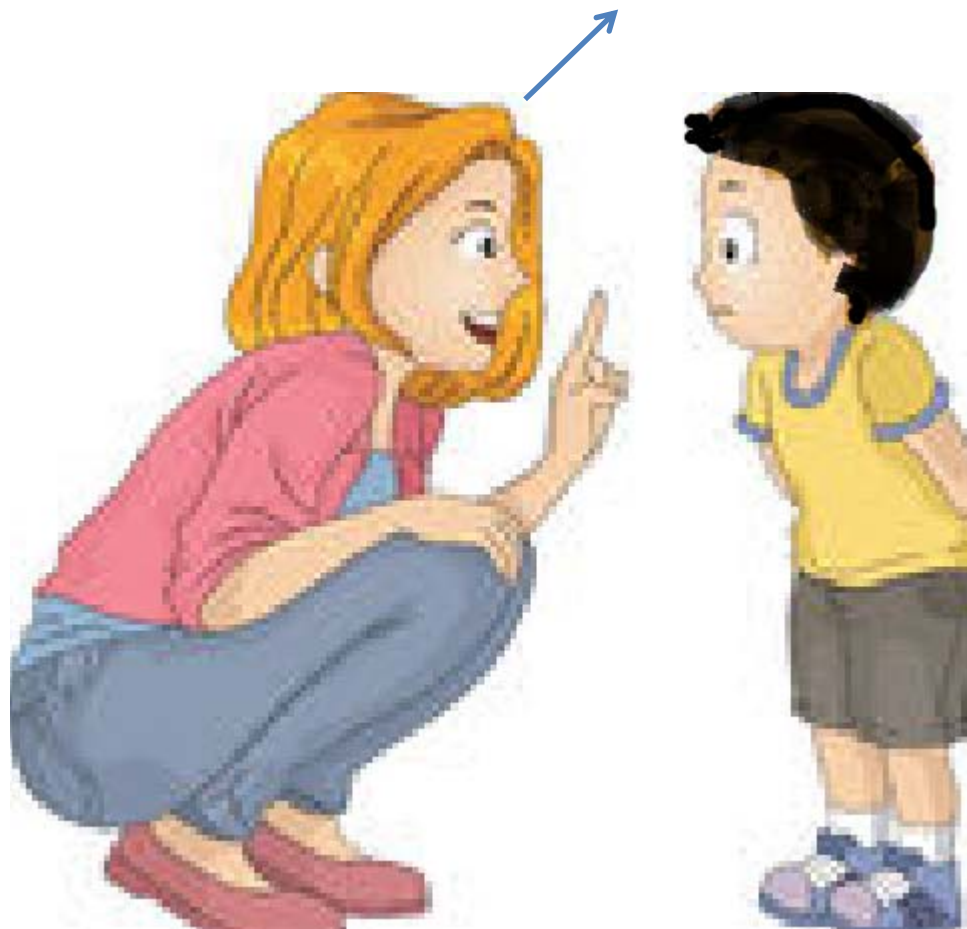
June 2014, Madison, Wisconsin. Conference in honor
of the 80th birthday of Professor Grace Wahba

Lessons I learned in graduate school:

Estimate the function...

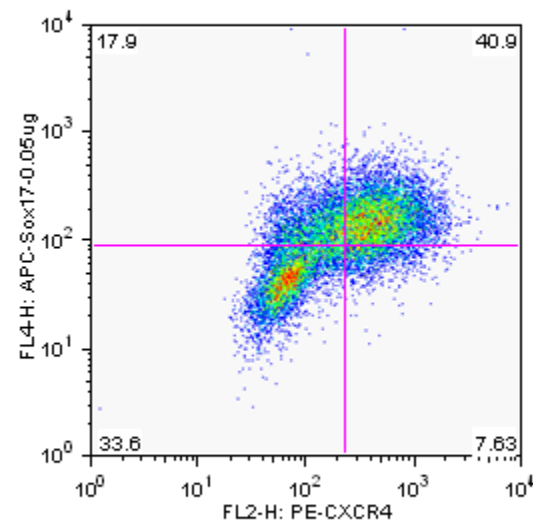
Put prior on function space...

Use RKHS, splines, approximation theory....

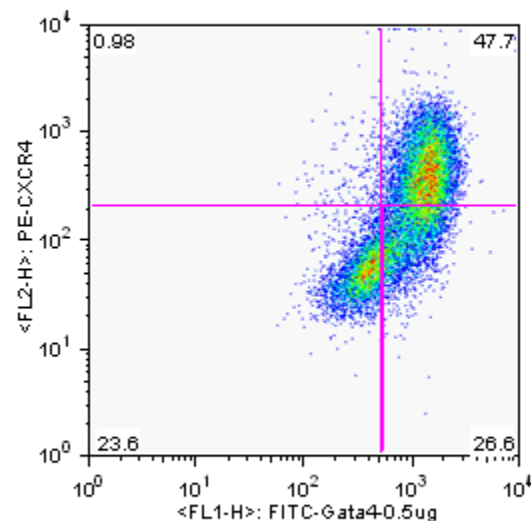


A 3-dimensional density function from flow cytometry

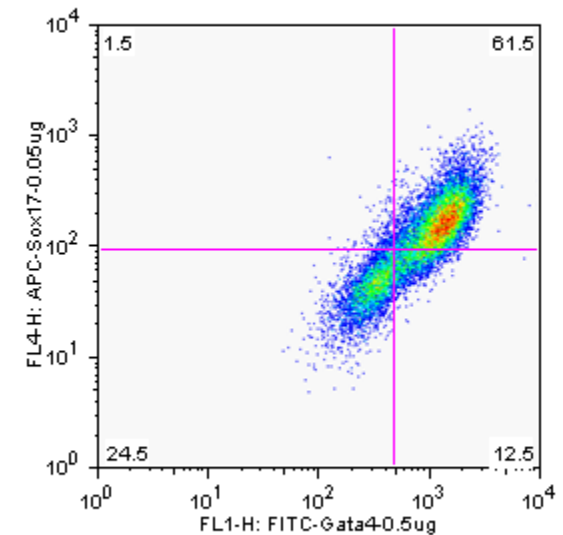
D 5 activin



FSC-H, SSC-H subset
CXCR4 Sox17-P43 d5 A 40108
Event Count: 17005



FSC-H, SSC-H subset
CXCR4 Gata4 d5 A 40108
Event Count: 17462

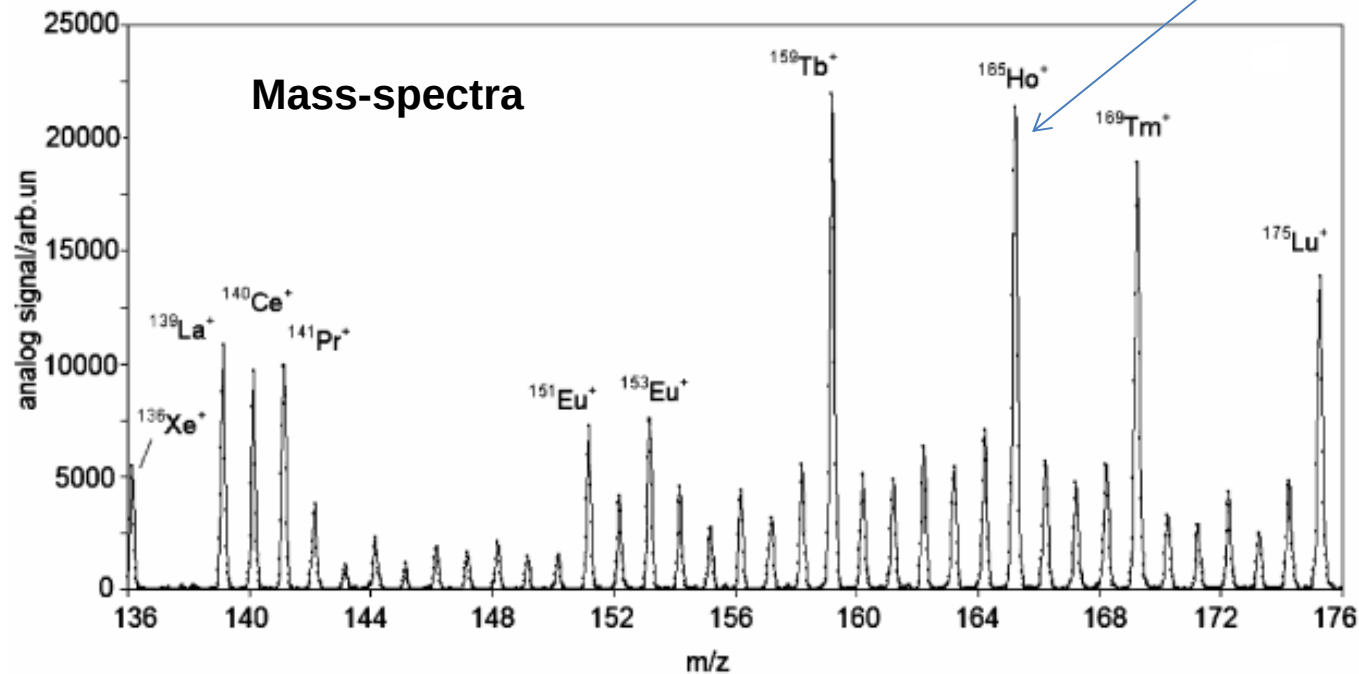


FSC-H, SSC-H subset
Gata4 Sox17-P43 d5 A 40108
Event Count: 16637

Mass cytometry: replace fluorescent labels with elemental labels

surface markers	tag isotopes	surface markers	tag isotopes
CD15	^{139}La	CD45	^{159}Tb
CD33	^{141}Pr	CD5	^{164}Dy
CD4	^{142}Nd	CD38	^{165}Ho
CD13	^{144}Nd	CD36	^{166}Er
CD14	^{145}Nd	CD57	^{170}Er
CD117	^{146}Nd	CD34	^{169}Tm
HLA-DR	^{147}Sm	CD19	^{171}Yb
CD3	^{152}Sm	CD56	^{174}Yb
CD45RA	^{151}Eu	CD28	^{176}Yb
CD64	^{153}Eu	CR	^{127}I -contrast reagent
CD20	^{156}Gd	Ir	$^{191,193}\text{Ir}$ -intercalator

Holmium
164.93031 amu



The Bayesian nonparametric problem

- $x_1, x_2, \dots, x_n$ are independent r.v. on a space Ω
- Their distribution Q is unknown but assumed to be drawn from a prior distribution π .
- Our tasks: Prior construction, posterior computation

$$\pi(Q) \quad Q(X) \rightarrow \Pr(Q, X) \rightarrow \Pr(Q | X) \rightarrow \Pr(g(Q))$$

- Want this to work well when Ω is of moderately high dimension, e.g. 5-50

Ferguson's criteria (1973)

- Support of the prior should be large
- The posterior should be tractable

Dirichlet process prior satisfies Ferguson's conditions.

However, under this prior the random distribution Q does not possess a density

Density is useful in many applications

- Anomaly detection
- Classification
- Compression
- Probabilistic networks
- Image analysis

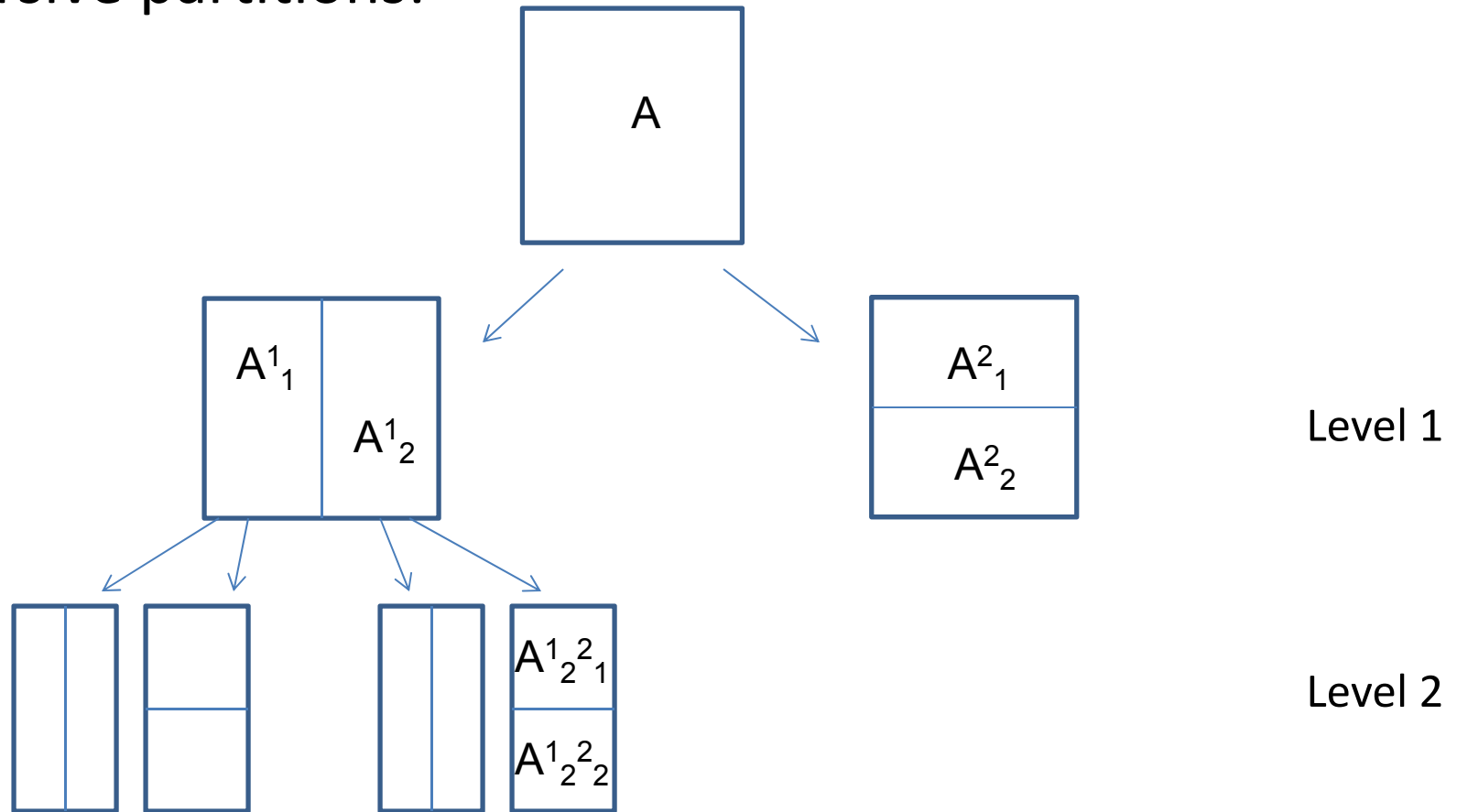
and more

We want to define a prior on the space of simple density functions:

$$f(x) = \sum c_i I_{A_i}(x)$$

To reduce complexity of this space, assume that $\Omega = A_1 \cup A_2 \dots \cup A_m$ is a recursive partition

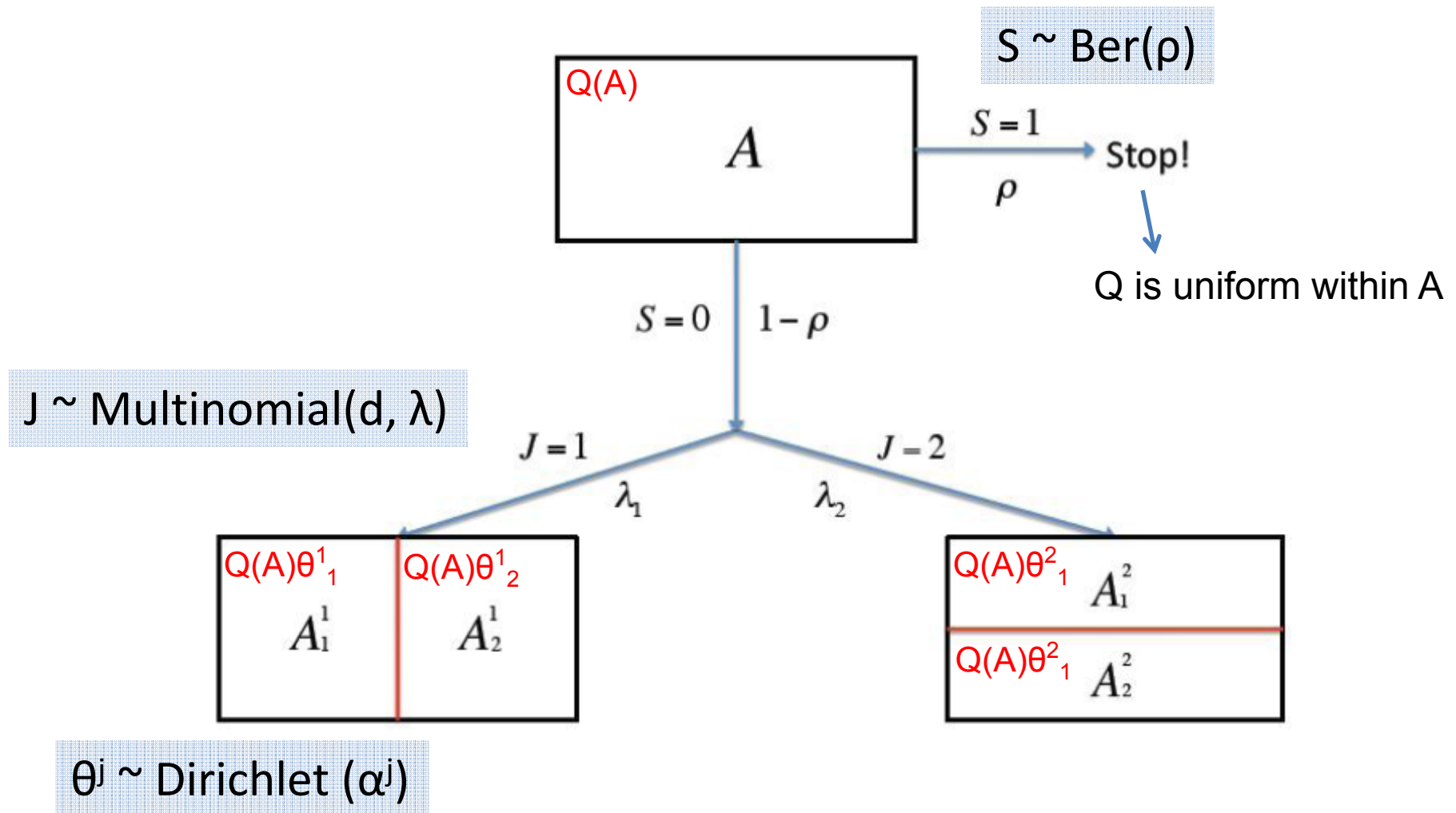
Recursive partitions:



In general, $A_k^j = k^{\text{th}}$ part of the j^{th} way to partition A

Recursive definition of random simple density function:

Suppose $Q(A)$ is known, define how $Q(\cdot)$ is distributed within A



Density on partitions of finite depth

- Suppose we have drawn $Q^{(k)}$ supported on a partition composing of regions up to level k
- For each region not yet stopped, repeat the partitioning process
- This gives a random distribution $Q^{(k+1)}$ with a density $q^{(k+1)}$ that is piecewise constant on a partition with regions up to level $k+1$

Optional Polya Tree (OPT) (Wong & Li, 2010)

Theorem: If the stopping probabilities are uniformly positive, then $Q^{(k)}$ converges almost surely in variational distance to an absolutely continuous distribution Q .

$$P \left(\int |q^{(k)} - q| dx \rightarrow 0 \text{ for some density } q \right) = 1$$

Q is said to have an OPT distribution with parameters

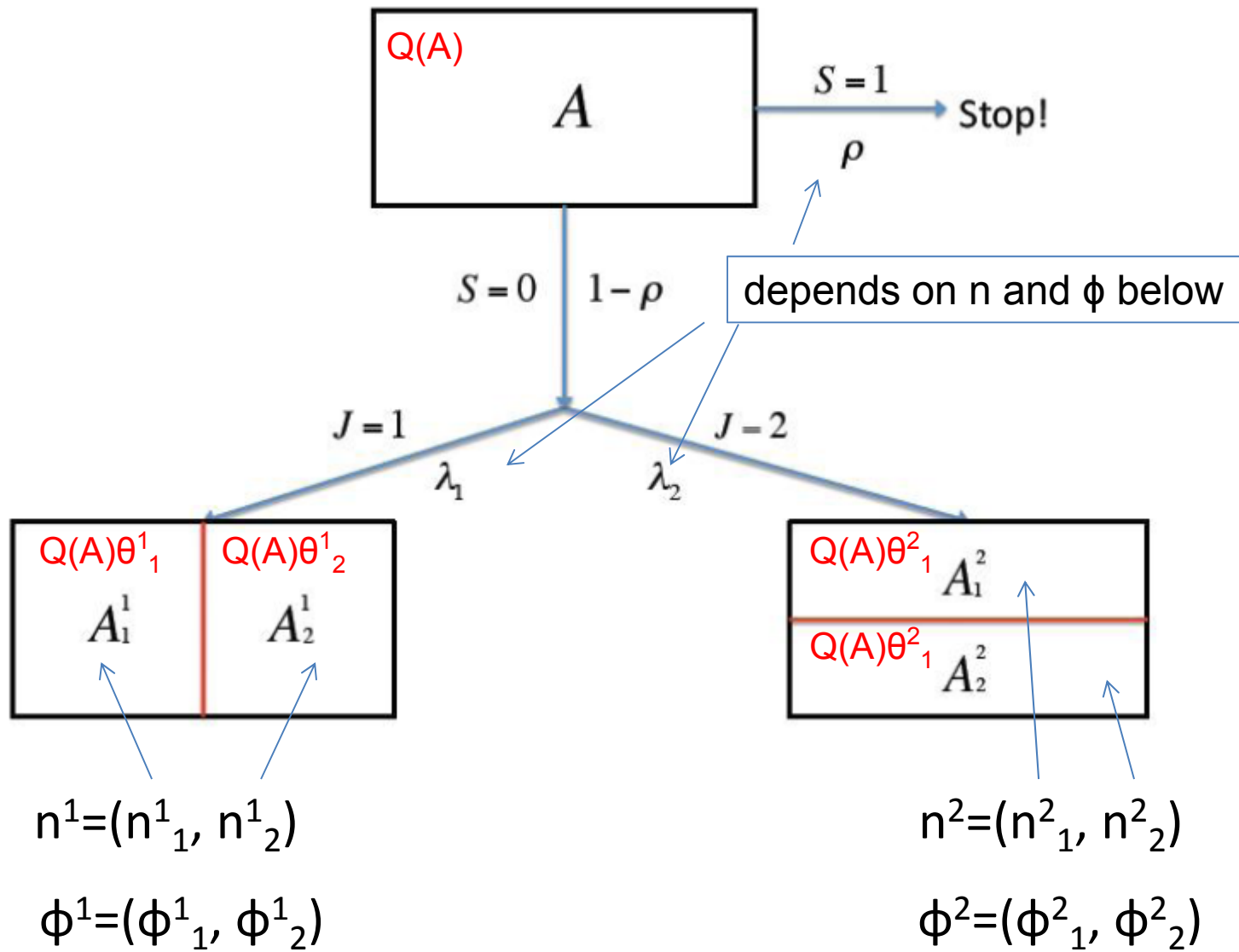
ρ (stopping rule)

λ (selection probabilities)

α (probability assignment weights)

OPT prior satisfies Ferguson's criteria

- 1) Any L_1 ball has positive prior probability
- 2) $\pi(Q \mid x_1, \dots, x_n)$ is also OPT with parameters
 $\rho(x_1, \dots, x_n), \lambda(x_1, \dots, x_n), \alpha(x_1, \dots, x_n),$
computable in finite time



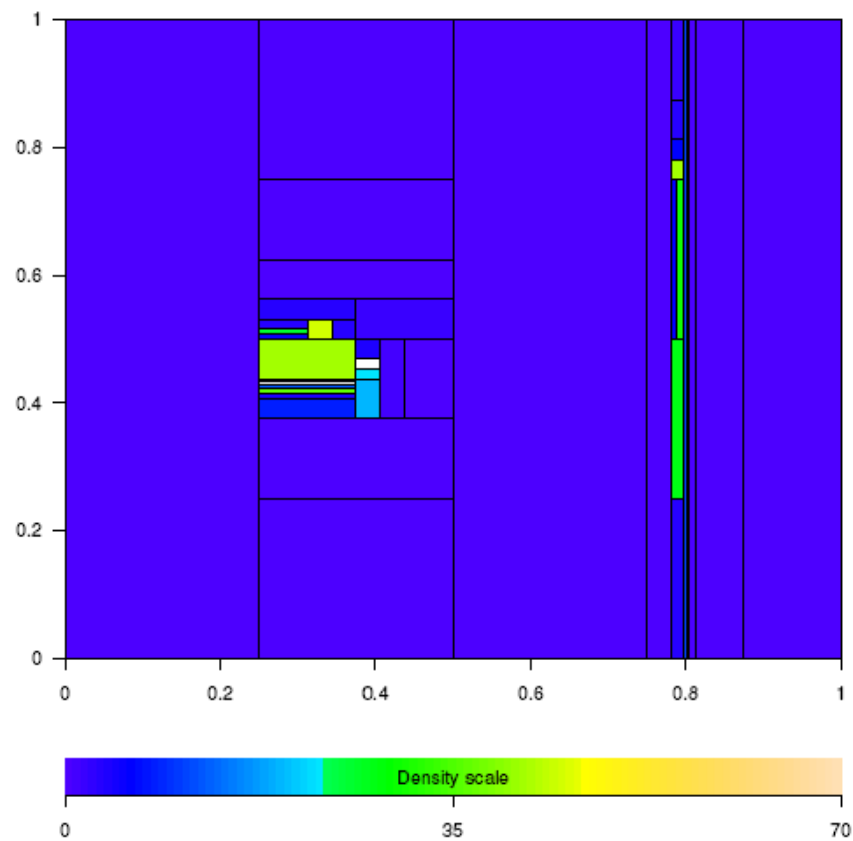
Computation of $\Phi(A)$ by recursion

$$\Phi(A) = \rho\Phi_0(A) + (1 - \rho) \sum_{j=1}^M \lambda_j \frac{D(\boldsymbol{n}^j + \boldsymbol{\alpha}^j)}{D(\boldsymbol{\alpha}^j)} \prod_{i=1}^{K^j} \Phi(A_i^j)$$

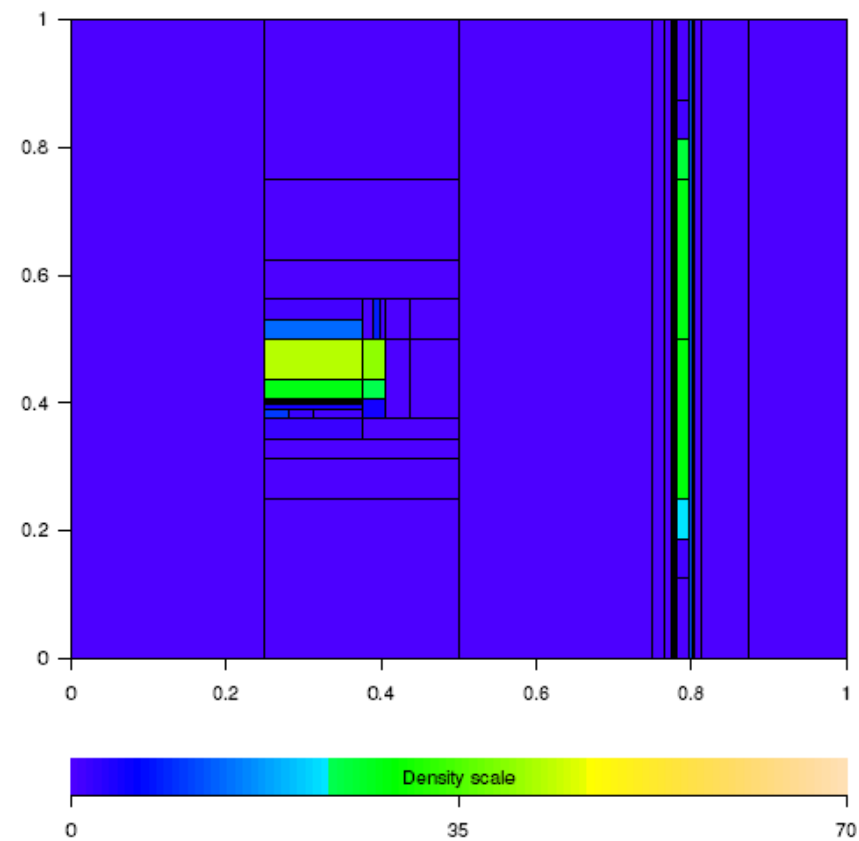
where $D(\boldsymbol{t}) = \Gamma(t_1) \dots \Gamma(t_k) / \Gamma(t_1 + \dots + t_k)$

Recursion ends when A has either 0 or 1 data points.

Example

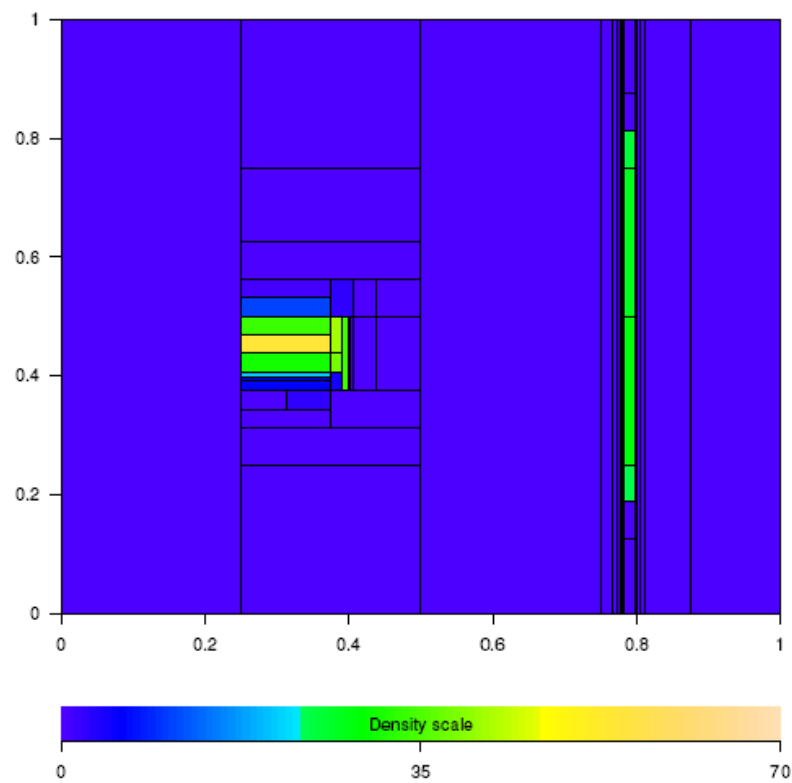


(a) Sample size = 100

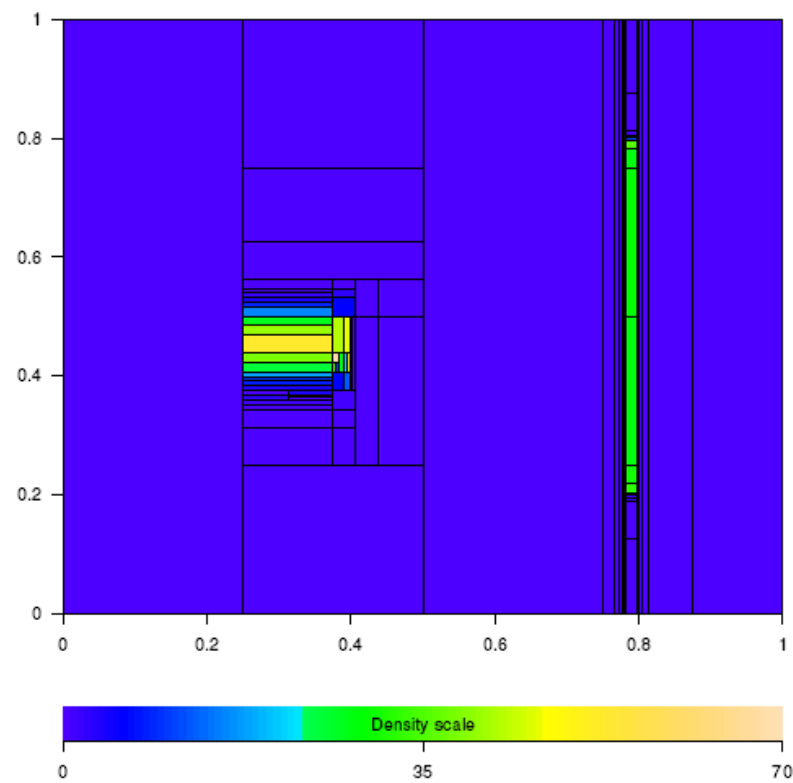


(b) Sample size = 500

Example (continued)

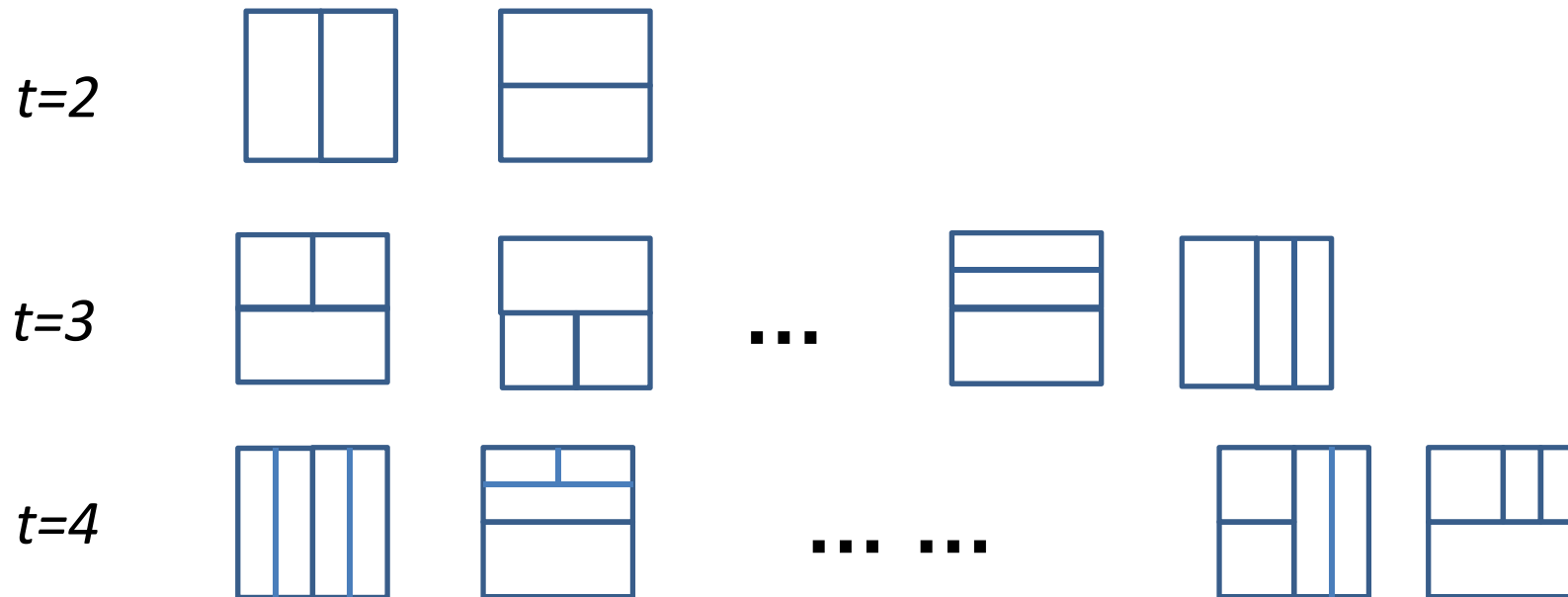


(c) Sample size = 1000



(d) Sample size = 5000

2nd approach: build up partition sequentially



Given t , want to define a posterior score for the partition directly

Assume Dirichlet (α) allocation of probabilities given the partition, then

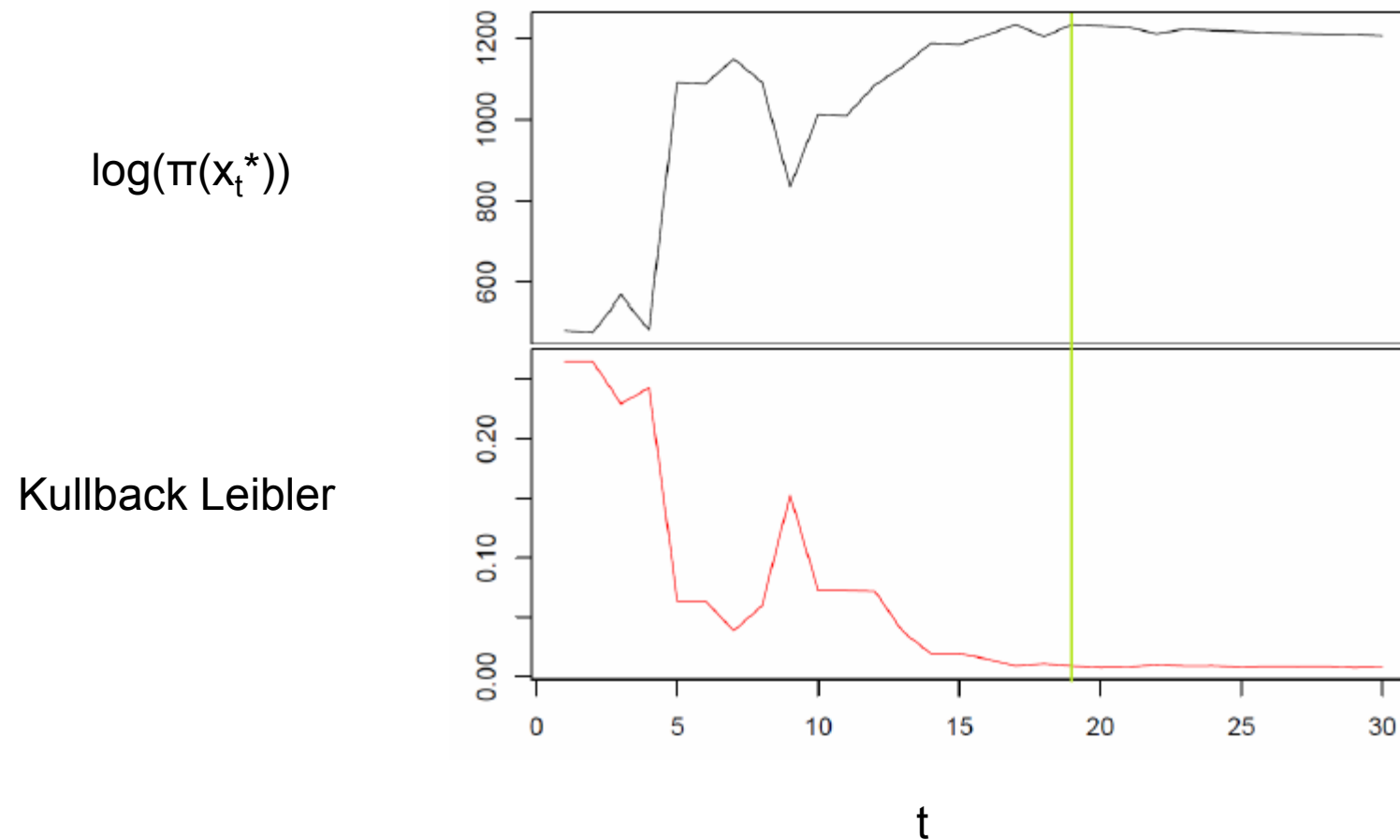
$$\begin{aligned}
 &P(D|A_1, \dots, A_t) \\
 &= \int_{\alpha_1, \dots, \alpha_t} P(D|\alpha_1, \dots, \alpha_t, A_1, \dots, A_t) P(\alpha_1, \dots, \alpha_t|A_1, \dots, A_t) d\alpha_1 \dots d\alpha_t \\
 &= \prod_{k=1}^t \frac{1}{|A_k|^{n_k}} \int_{\alpha_1, \dots, \alpha_t} \prod_{k=1}^t \alpha_k^{n_k} \frac{\prod_{k=1}^t \alpha_k^{-0.5}}{D(0.5, \dots, 0.5)} d\alpha_1 \dots d\alpha_t \\
 &= \prod_{k=1}^t \frac{1}{|A_k|^{n_k}} \frac{D(n_1 + 0.5, \dots, n_t + 0.5)}{D(0.5, \dots, 0.5)}.
 \end{aligned}$$

Bayesian score of a partition X of size t :

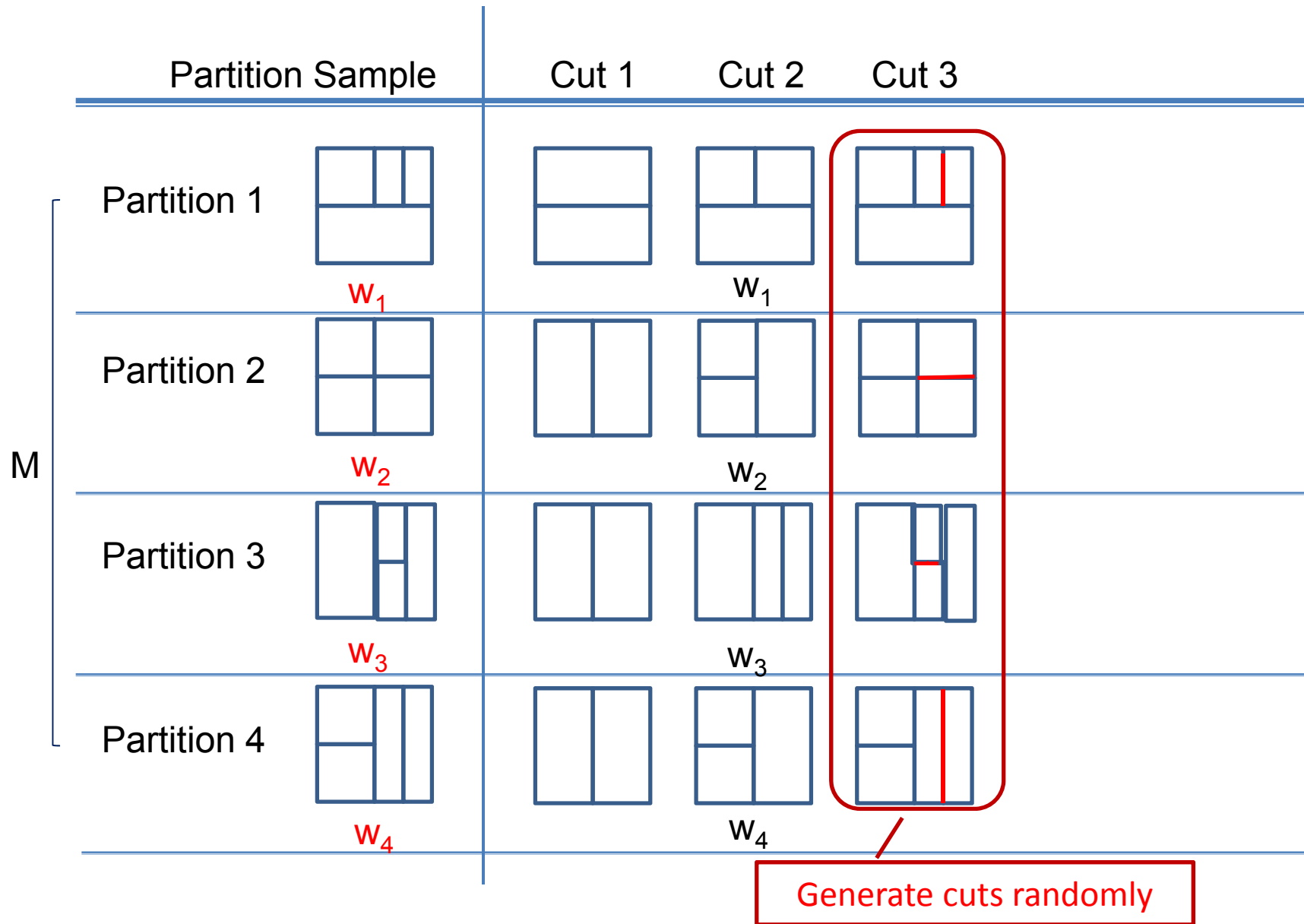
$$\pi(X) = e^{-t} \frac{D(n_1+0.5, \dots, n_t+0.5)}{D(0.5, \dots, 0.5)} \prod_{k=1}^t \frac{1}{|A_k|^{n_k}}$$

Asymptotically, the score tracks distance from true density

x_t^* = best scoring partition of size t



Sequential importance sampling



How to choose the proposal density?

$$\pi_t(\hat{y}) \propto e^{-t} \frac{D(\frac{1}{2} + n_1, \dots, \frac{1}{2} + n_t)}{D(\frac{1}{2}, \dots, \frac{1}{2})} \prod_{k=1}^t \left(\frac{1}{|A_k|} \right)^{n_k}$$

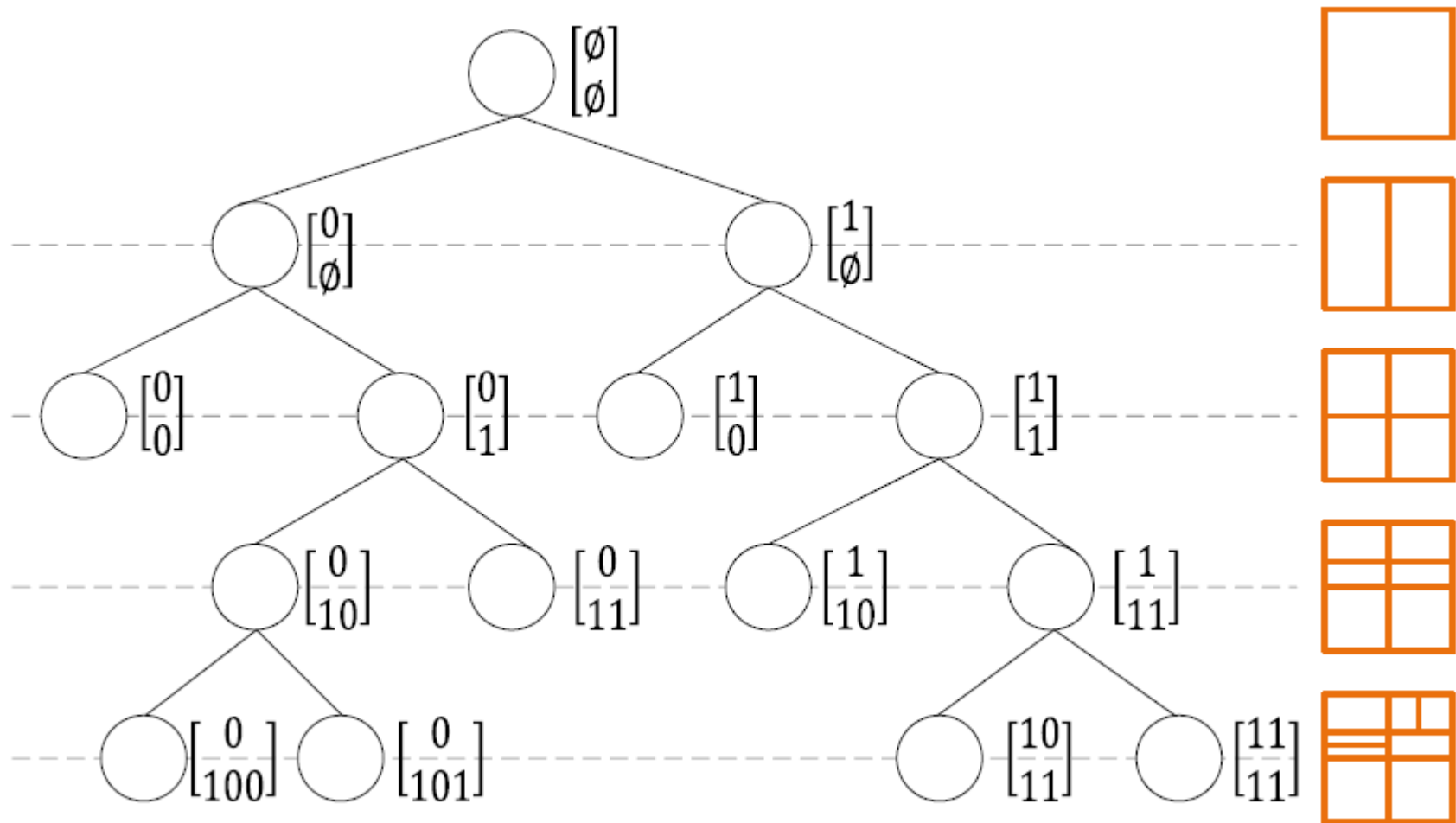
$$\pi_t(y_1, y_2, \dots, y_t) = \pi_t(y_1) \pi_t(y_2 | y_1) \dots \pi_t(y_t | y_{t-1})$$

infeasible

feasible

$$q_t(y_1, y_2, \dots, y_t) = \pi_1(y_1) \pi_2(y_2 | y_1) \dots \pi_t(y_t | y_{t-1})$$

Data structure for region counts



Counting can be accelerated by hardware

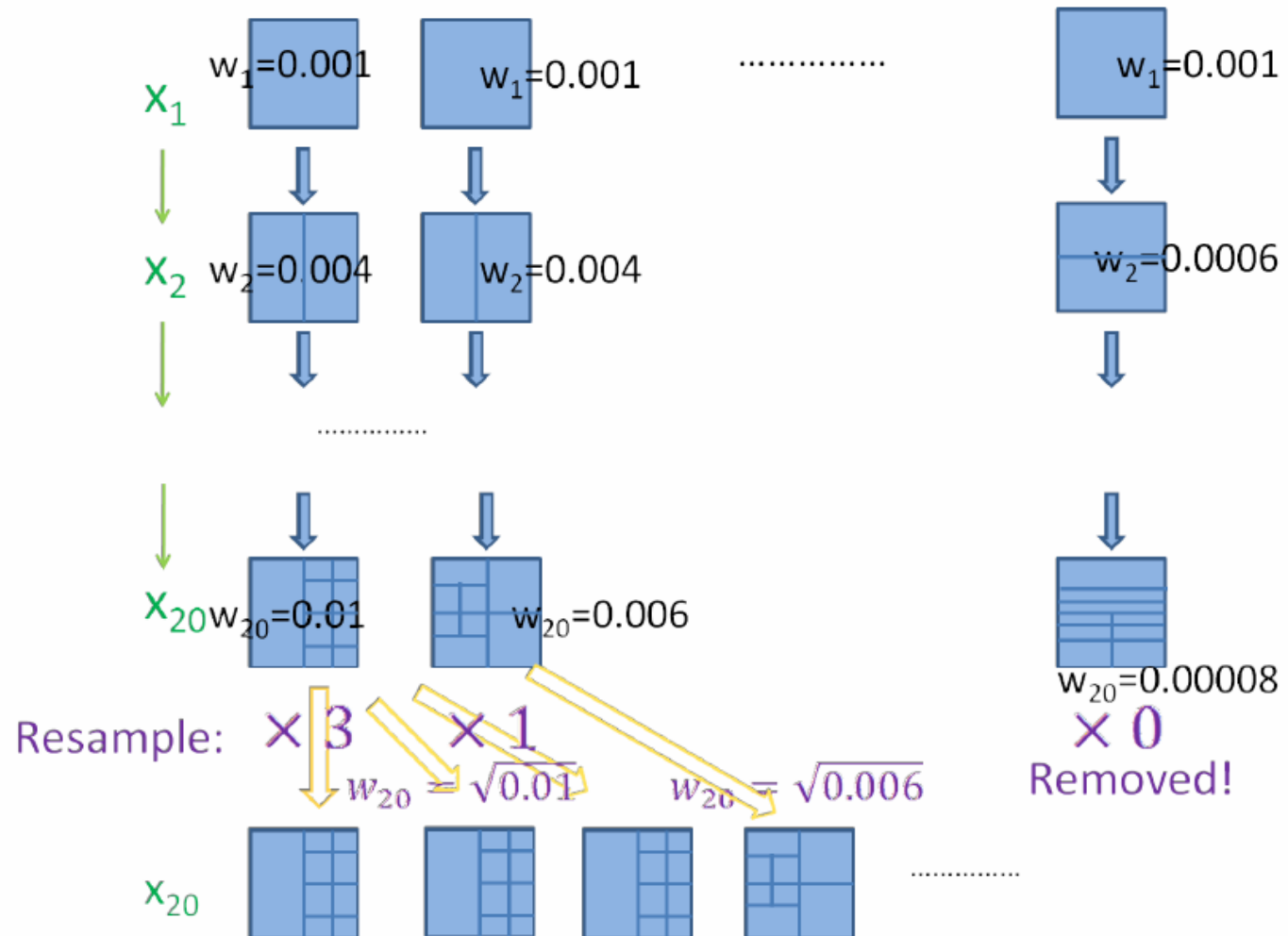
- Intel(R) Xeon(R) CPU E5640 @ 2.67GHz
- GeForce GTX 680
 - 1536 CUDA cores (8×192) @ 1.08GHz
 - 512KB L2
 - 2G RAM
 - Memory clock rate: 3GHz
 - Memory bus width: 256-bit
 - Bandwidth: 150GB/s

Experiment result (CountEngine)

- Partition = 300, Cut = 1000

Dim_# of data	CPU	GPU	Speedup
32_10^5	33375.40	18.56	1798.54
32_10^6	316794.00	28.82	10992.20
64_10^5	113327.00	24.59	4609.34
64_10^6	1086660.00	39.38	27593.51
128_10^5	438960.00	34.19	12837.45
128_10^6	NA	57.50	NA

Resampling:



For $j' = 1, \dots, 1000$,
 let $x_t^{(j')}$ be $x_t^{(j)}$ independently with prob. proportional to
 $(\sqrt{0.01}, \sqrt{0.006}, \dots, \sqrt{0.00008})$.

Comparison with Kernel Density Estimate

Sample n points from a D-dim distribution:

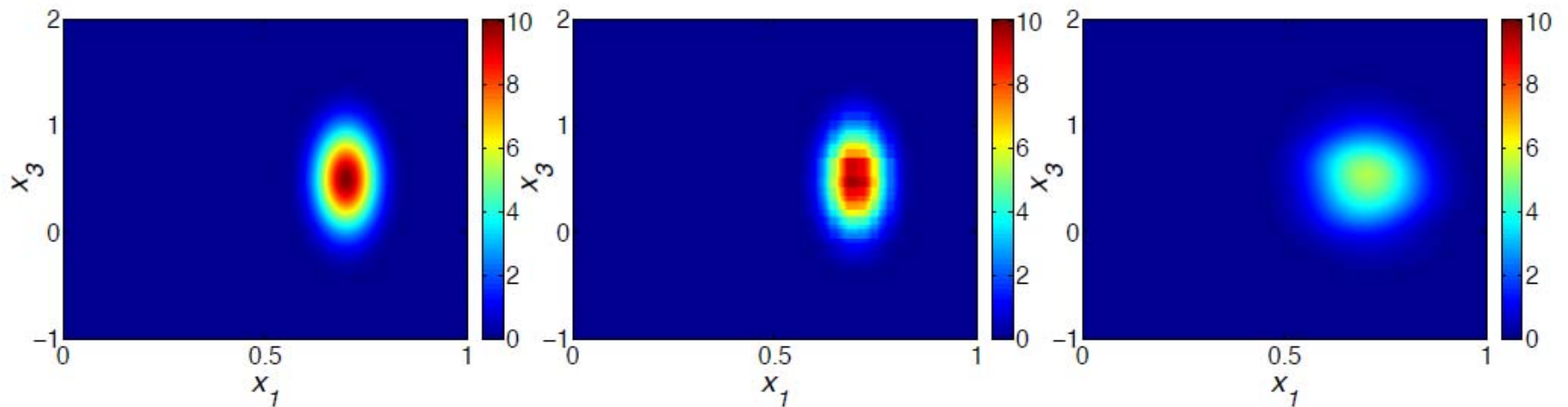
$$\begin{aligned} \begin{pmatrix} X_1 \\ X_2 \end{pmatrix} &\sim \frac{2}{5}N\left(\begin{pmatrix} 0.3 \\ 0.6 \end{pmatrix}, 0.05^2 I\right) + \frac{3}{5}N\left(\begin{pmatrix} 0.7 \\ 0.4 \end{pmatrix}, 0.05^2 I\right) \\ X_3, X_4 &\sim N(0.5, 0.1) \\ X_5, \dots, X_D &\sim \frac{1}{2}N(0.35, 0.1) + \frac{1}{2}N(0.6, 0.5) \end{aligned}$$

Results:

DISTRIBUTION		KL DIVERGENCE		
D	n	BSP	KDE	
			ROT	MSP
5	10^4	0.0388	0.3067	0.3958
5	5×10^4	0.0145	0.1996	0.2668
30	5×10^4	0.06749	8.9081	8.6062

Estimate of conditional density in 7 dimension

$$p(x_1, x_3 | x_2 = 0.3, x_4 = 0.5, x_5 = 0.45, x_6 = 0.5, x_7 = 0.65)$$



(a) Ground Truth

(b) BSP

(c) KDE

- Case 1: $p(x_1|x_2 = 0.45, x_3 = 0.5, x_4 = x_5 = x_6 = x_7 = 0.5)$
- Case 2: $p(x_1, x_3|x_2 = 0.3, x_4 = 0.5, x_5 = 0.45, x_6 = 0.5, x_7 = 0.65)$
- Case 3: $p(x_5, x_7|x_1 = 0.3, x_2 = 0.5, x_3 = 0.45, x_4 = 0.5, x_6 = 0.65)$
- Case 4: $p(x_1, x_2, x_5|x_3 = 0.45, x_4 = 0.55, x_6 = 0.6, x_7 = 0.7)$

Case	BSP		KDE	
	KL	Time [sec]	KL	Time [sec]
1	0.0065	22.51	0.2719	932.36
2	0.0052	21.38	0.2671	1311.15
3	0.0038	21.37	0.0271	1380.22
4	0.0105	23.45	0.3128	1252.07

Density estimation is a building block for other statistical methods.

Classification:

- 1) Estimate class density $f_i(x)$ for classes $1, \dots, k$
- 2) Use Bayes classifier: $p(i|x) \sim \alpha_i f_i(x)$

MAGIC data: 10 dimension, 12,000 cases, 7,000 controls

Letter data: 16 dimension, 26 classes, n=16,000 within class

	LDA	QDA	KNN	SVM	CART	ADA	KDE	BSP
MAGIC	0.785	0.786	0.813	0.871	0.839	0.871	0.832	0.824
Letter	0.689	0.875	0.959	0.943	0.818	0.774	-	0.937

Data Compression: the estimated density can guide the design of optimal compression scheme

A sequencer yield 1 billion reads in about 1 day.

Quality scores associated with the base calls take up too much disk space

```
@SEQ_ID
GATTGTTGGGGTTCAAAGCAGTATCGATCAAATAGTAAATCCATTGTTCACAGTTT
+
!''*((( (***) ) 000000++ ) (000000) .1***-+*'') ) **55CCF>>>>>CCCCCCC65
```

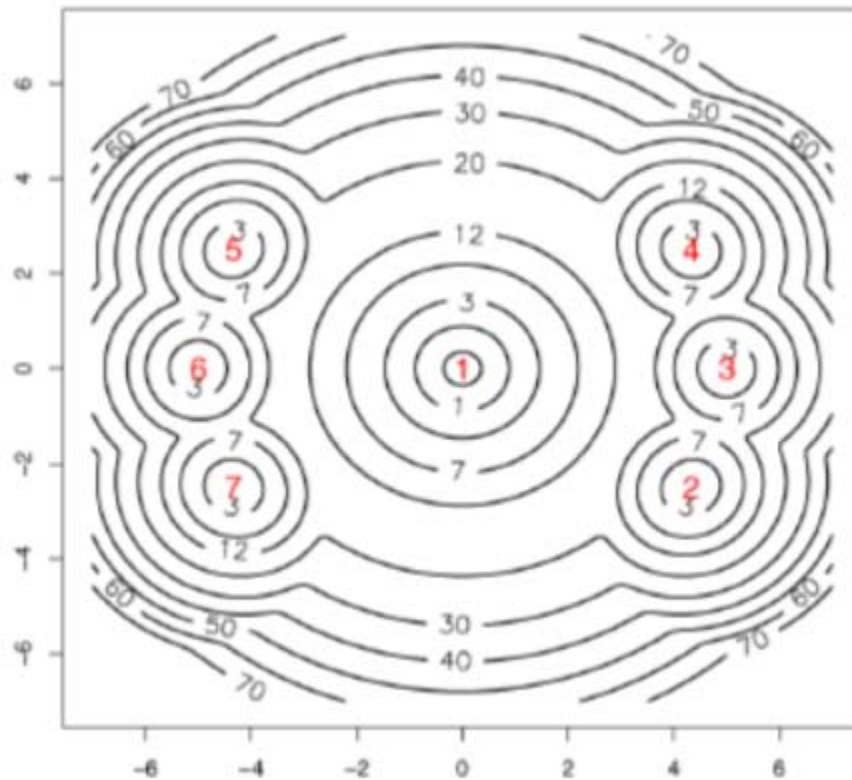
Test: n=1,940,271 quality score vectors (100 dimensional, divided into 20 sub-vectors)

Result:

Our method uses 206 bits per read for lossless compression

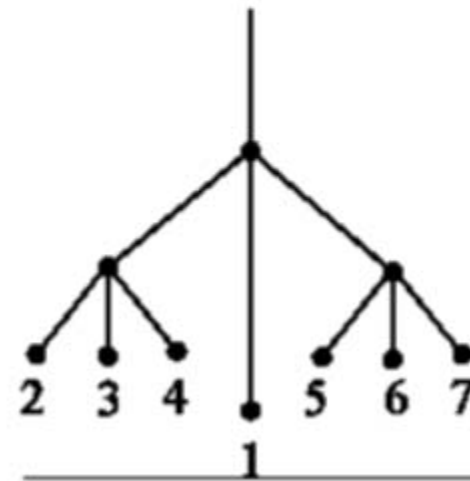
In comparison, 7-zip uses 213 bits per read

Visualization of information



(A)

Contour plot of the energy function
of a 2D density with seven modes



(B)

Sub-level tree of energy
(log-density)

Sublevel tree of bone marrow data (n=2,000,000)

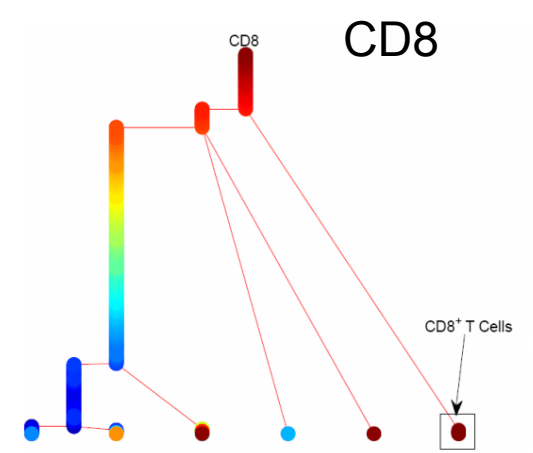
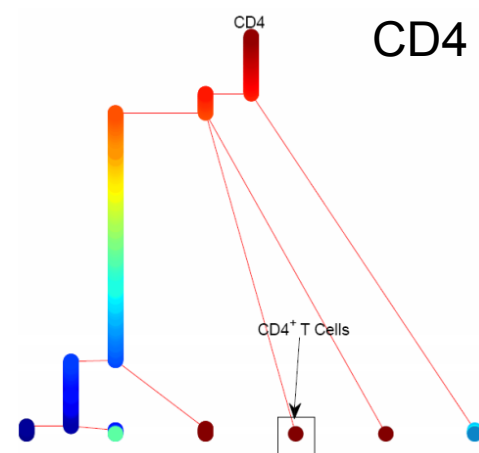
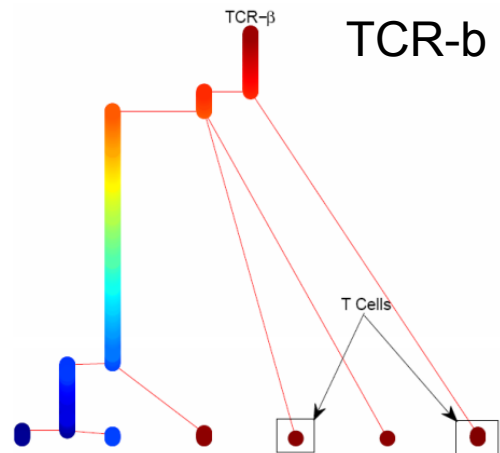
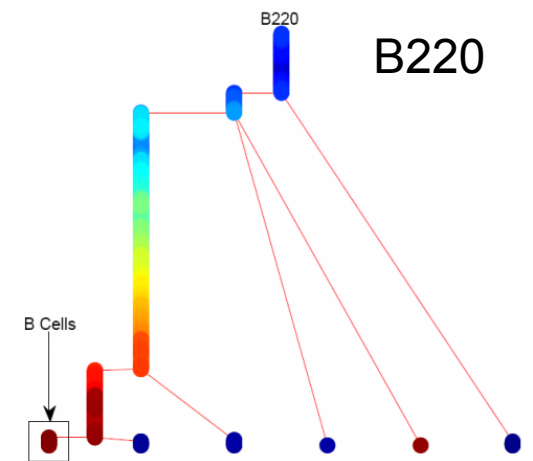
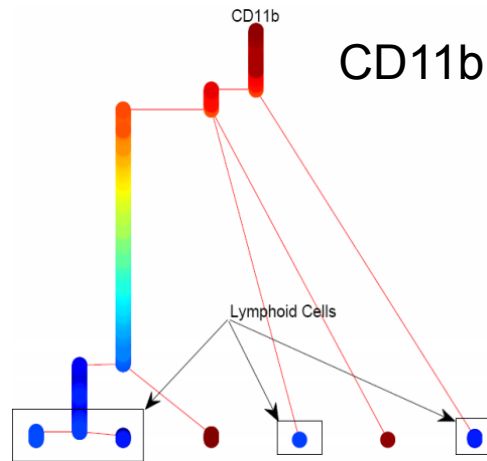
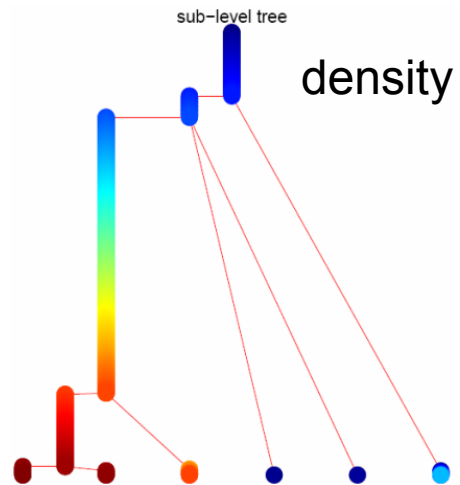


Image segmentation

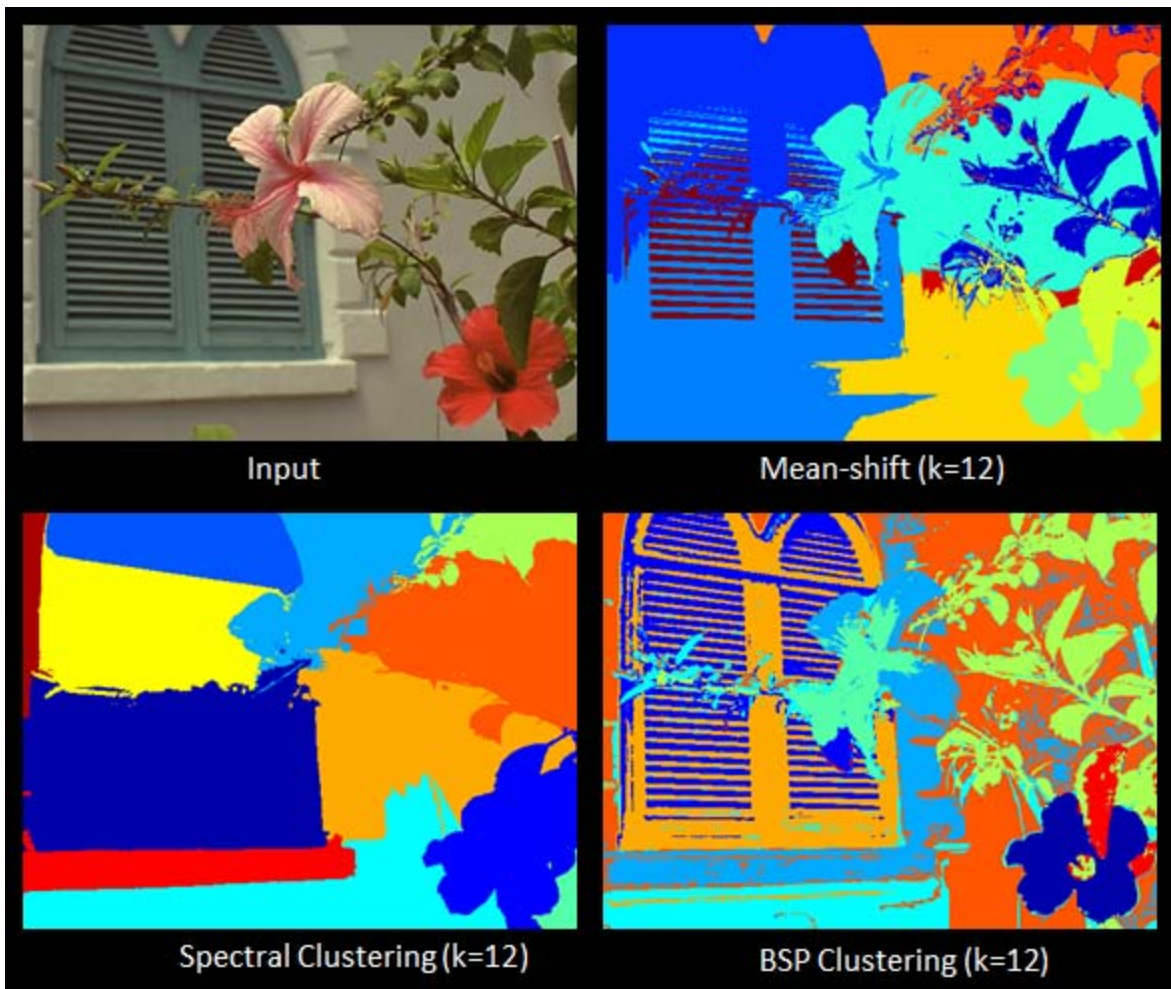


Image enhancement



(a) Input image



(b) Modified smoothed image



(c) Output image

Convergence rates (sieve MLE case)

1. Class of simple functions on BPs of size l :

$$\Theta_l = \{ f(\cdot) : f(\cdot) = \sum \beta_i I_{A_i}(\cdot), \beta_i \geq 0, \sum \beta_i \mu(A_i) = 1, \text{ and } A_i, i=1, \dots, l \text{ form a binary partition} \}$$

2. Log-likelihood of f :

$$L_n(f) = \sum_{\{j=1, \dots, n\}} \log f(y_j) = \sum_{\{i=1, \dots, l\}} n_i \log(\beta_i)$$

3. MLE based on Θ_l :

$$\hat{f}_{n,l} = \operatorname{argmax} \{ L_n(f) : f \in \Theta_l \}$$

4. Sieve MLE based on $\{ \Theta_l, l=1, 2, \dots \}$:

$$\hat{f}_n = \hat{f}_{n,l(n)}, \text{ where } l=l(n) \uparrow \infty$$

Classical result (Stone 1980):

rate $\sim n^{-\alpha}$, $\alpha = p/(2p+d)$

p = # of bounded derivatives of f_0

d = dimension of Ω

We want a new type of result:

If f_0 can be approximated by Θ_l at rate l^{-r}

then for some suitable choices of $l(n)$,

$\hat{f}_n$ achieves a rate $n^{-\alpha(r)}$ with $\alpha(r) \uparrow 1/2$

The key is to remove dependency of α on d .

A relevant result was given in Wong & Shen (1995):

Let H_l be the bracketing Hellinger entropy of Θ_l , and

δ_l be the approximation rate of Θ_l to f_0 ,

Then (with ρ denoting the Hellinger distance),

$$P(\rho(\hat{f}_n, f_0) > \varepsilon_n) < \exp\{-kn\varepsilon_n^2\},$$

whenever $\bar{\delta} = \delta_{l(n)}$, $\varepsilon = \varepsilon_n$, $H = H_{l(n)}$ satisfy

- i) $\bar{\delta} = o(\varepsilon^2)$ and
- ii) $\int_{[\varepsilon^2, \varepsilon]} H(u)^{1/2} du = O(n^{1/2} \varepsilon^2)$

However δ_l was required to be much stronger than ρ

Result (Linxi Liu & WW)

With $r > 1/2$ and $l(n)$ chosen to be $\left(\frac{n}{\log n}\right)^{\frac{1}{2r} + \xi}$

our sieve MLE has a rate upper bounded by

$$\left(\frac{\log n}{n}\right)^{\frac{1}{2} - \frac{1}{4r} - \frac{\xi}{2}} (\log n)^{1/2}$$

This result can be used to establish
spatial adaptation and variable selection

Acknowledgments

BSP: Luo Lu

OPT: Li Ma

Clustering & image: TY Wu, CY Tseng

Flow-cytometry: Michael Yang

Compression: Luo Lu, John Mu

Convergence rate: Linxi Liu