Optimal Sampling Strategies for tree-based models

Rolf Riedi

Vinay Ribeiro

R. Baraniuk

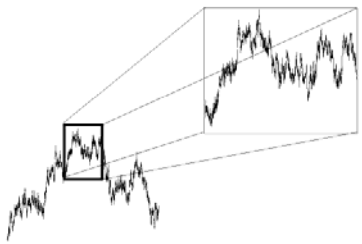
Los Alamos, May 2005

Statistical Models with Scaling

A tree based model for Brownian motion

Brownian Motion (BM)

- Brownian motion B(s):
  - Gaussian process
  - E[B(s)B(t)]= min(s,t)
  - B(0)=0 a.s.



- Increments
  - $Y_k^{(\Delta)} = B(k\Delta) - B((k-1)\Delta)$
  - i.i.d. (white noise): $\mathcal{N}(0, \Delta)$

$$\begin{aligned} E[Y_0 Y_k] &= E[B(\Delta)B(k\Delta)] - E[B(\Delta)B((k-1)\Delta)] \\ &= \min(\Delta, k\Delta) - \min(\Delta, (k-1)\Delta) = 0 \end{aligned}$$

  - Self-similar

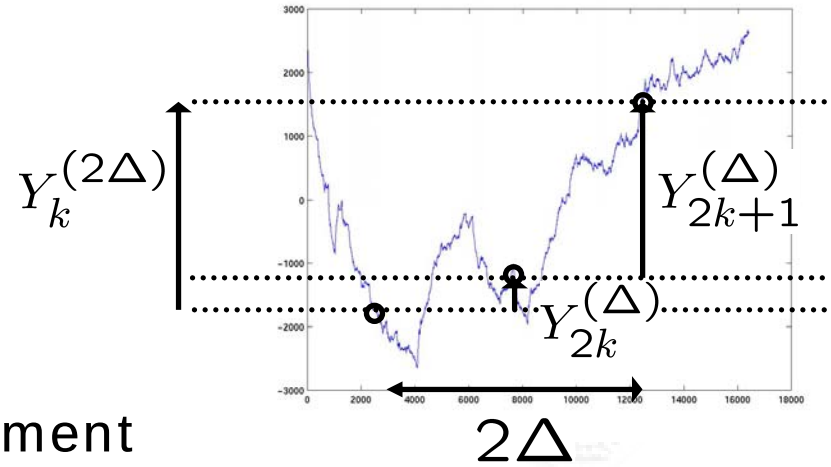
$$\frac{1}{\sqrt{\Delta}}(Y_1^{(\Delta)}, Y_2^{(\Delta)}, \ldots) \stackrel{\text{law}}{=} (Y_1^{(1)}, Y_2^{(1)}, \ldots)$$

Multi-scale approach to BM

- By definition:

$$Y_k^{(2\Delta)} = Y_{2k}^{(\Delta)} + Y_{2k+1}^{(\Delta)}$$

with $E[Y_{2k}^{(\Delta)} Y_{2k+1}^{(\Delta)}] = 0$ and $Y_m^{(\Delta)} \sim \mathcal{N}(0, \Delta)$

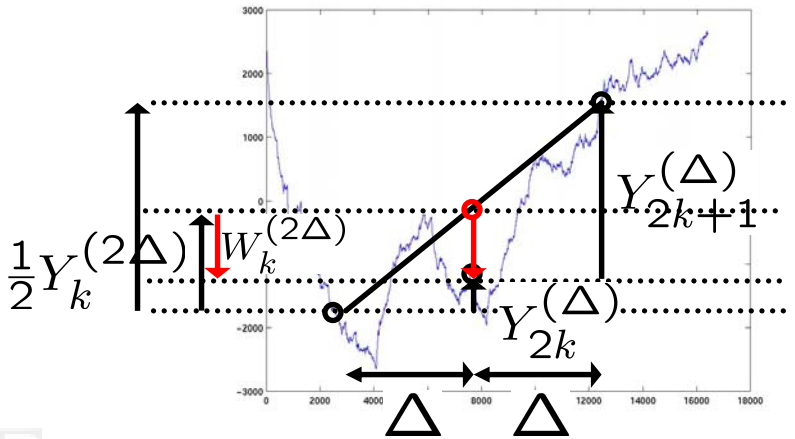


- Alternative writing:
  - Celebrated mid-point displacement

$$Y_{2k}^{(\Delta)} = \tfrac{1}{2}(Y_k^{(2\Delta)} + W_k^{(2\Delta)})$$

$$Y_{2k+1}^{(\Delta)} = \tfrac{1}{2}(Y_k^{(2\Delta)} - W_k^{(2\Delta)})$$

Note: $E[Y_k^{(2\Delta)} W_k^{(2\Delta)}] = 0$ and $W_k^{(2\Delta)} \sim \mathcal{N}(0, 2\Delta)$
implies $E[Y_{2k}^{(\Delta)} Y_{2k+1}^{(\Delta)}] = 0$ and $Y_m^{(\Delta)} \sim \mathcal{N}(0, \Delta)$



- Compare: Haar wavelet decomposition
- Alternative view:
  - BM given by Tree of innovations

Tree Models

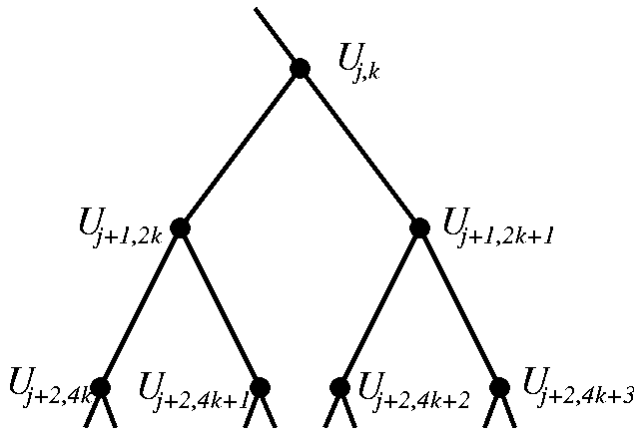
Algorithmically efficient

Versatile

Ex: Network traffic modeling

Trees

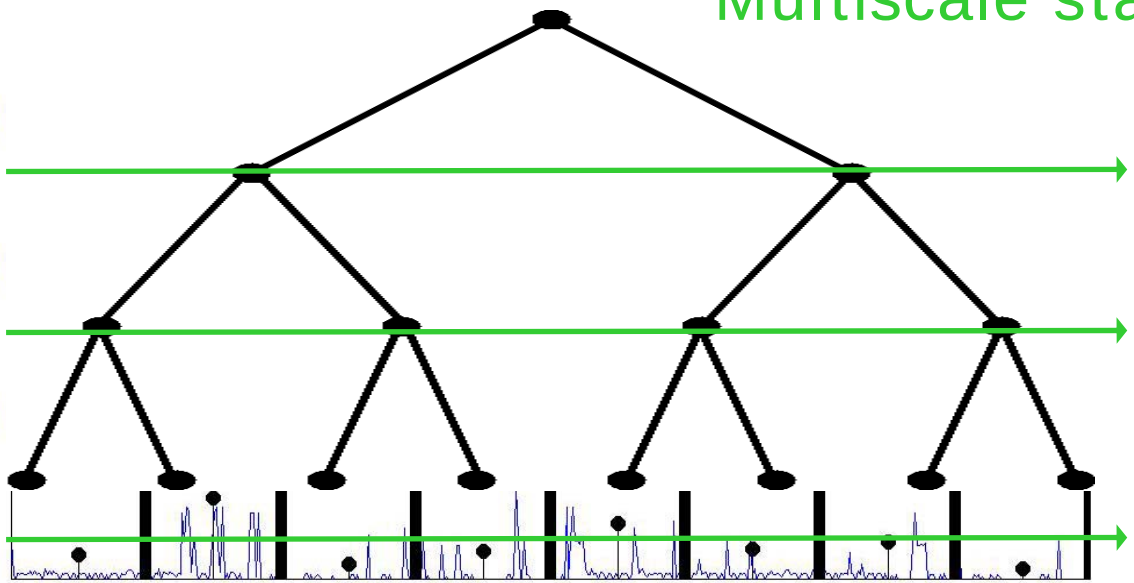
- Simple graphical model
- Symbiosis of probability theory and graph theory
- Parsimonious representation of complexity
- Examples
  - Time series at multiple scales
  - Wavelet scaling tree
  - Sensor placement
  - Bandwidth estimation
  - Multi-resolution image



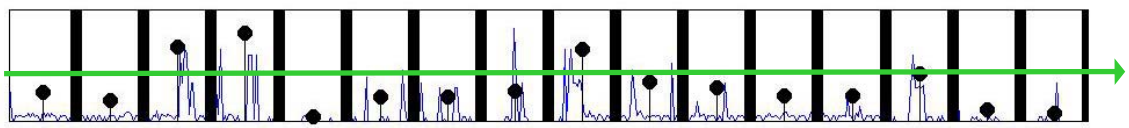
Multiscale Modeling

Time $\longrightarrow$

Scale $\downarrow$



Analysis: flow up the tree by adding



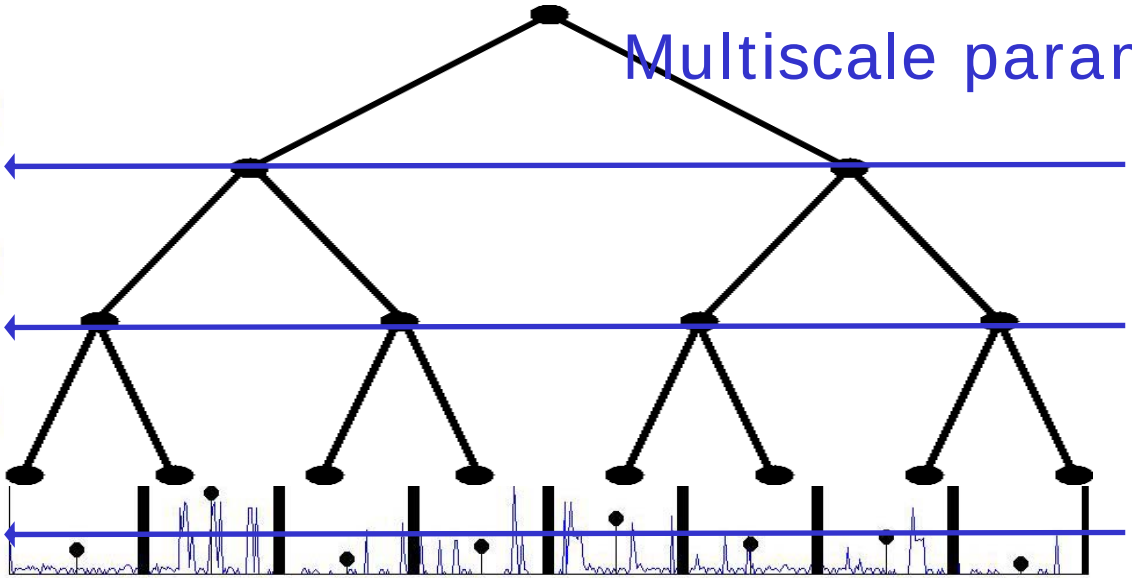
Start at bottom with trace itself

Multiscale Modeling

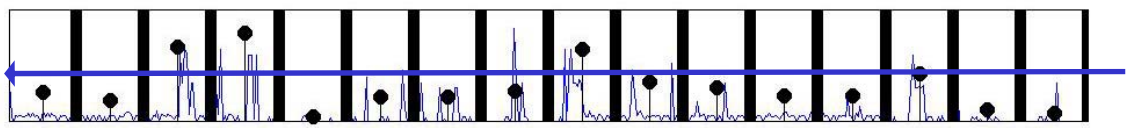
Time $\longrightarrow$

Start at top with total arrival

Scale $\downarrow$

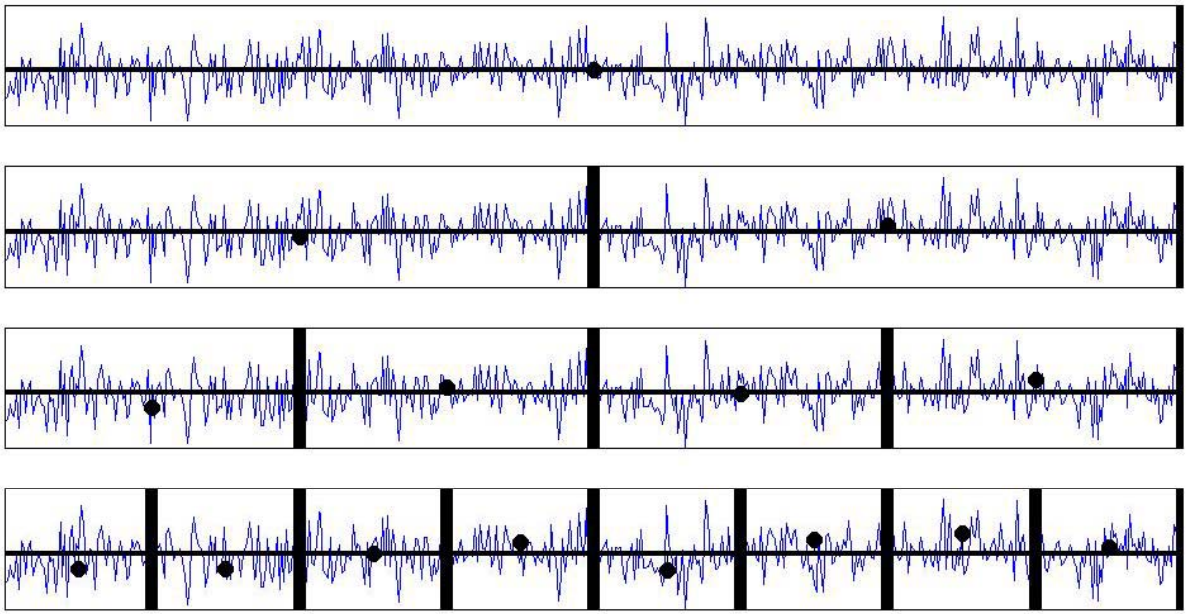


Synthesis: flow down via innovations

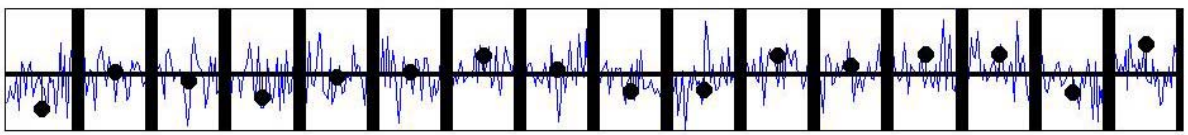


Signal: bottom nodes

Additive innovations: Linear Processes



Match variances on all dyadic scales
CLT: asymptotically Gaussian

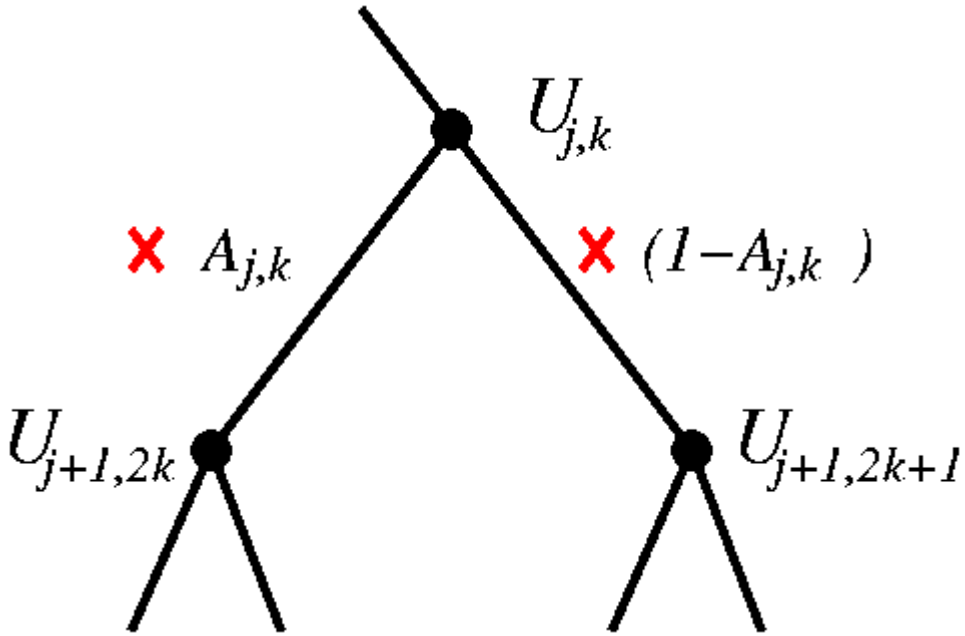


Additive Innovations $W_{jk} \sim \mathcal{N}(0, \sigma^2 2^{-j(2H-1)})$: Model for $B_H(t)$

Multiplicative innovations Multifractal Wavelet Model (MWM)

- Random *multiplicative innovations* $A_{j,k}$ on [0,1]

  eg: beta



- Parsimonious modeling (one parameter per scale)

- Strong ties with rich theory of *multifractals*

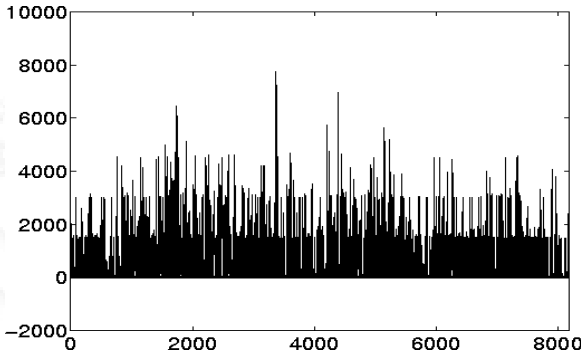
Multiscale Traffic Trace Matching

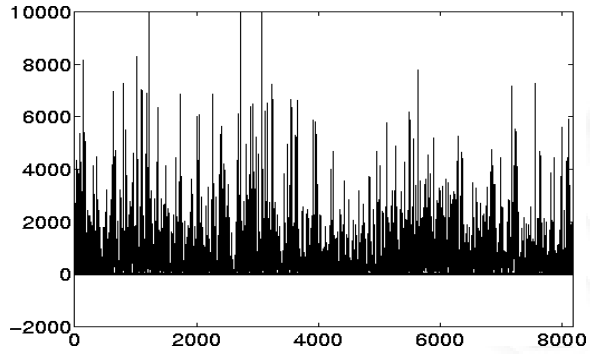
scale

Auckland 2000

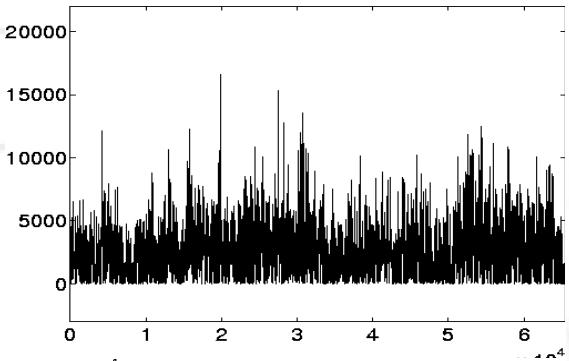
MWM match

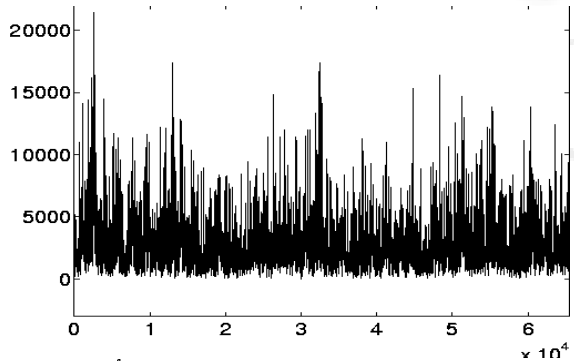
4ms



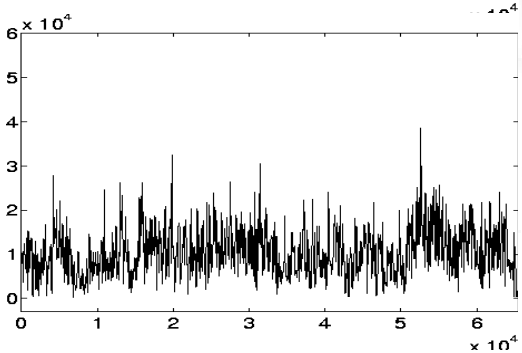


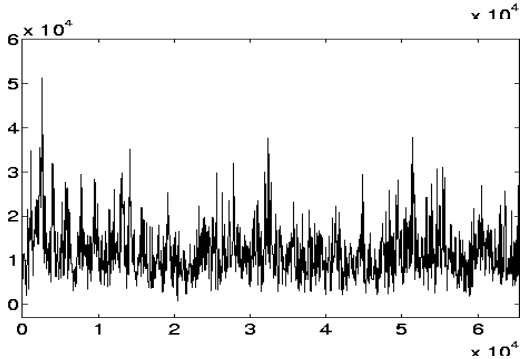
16ms



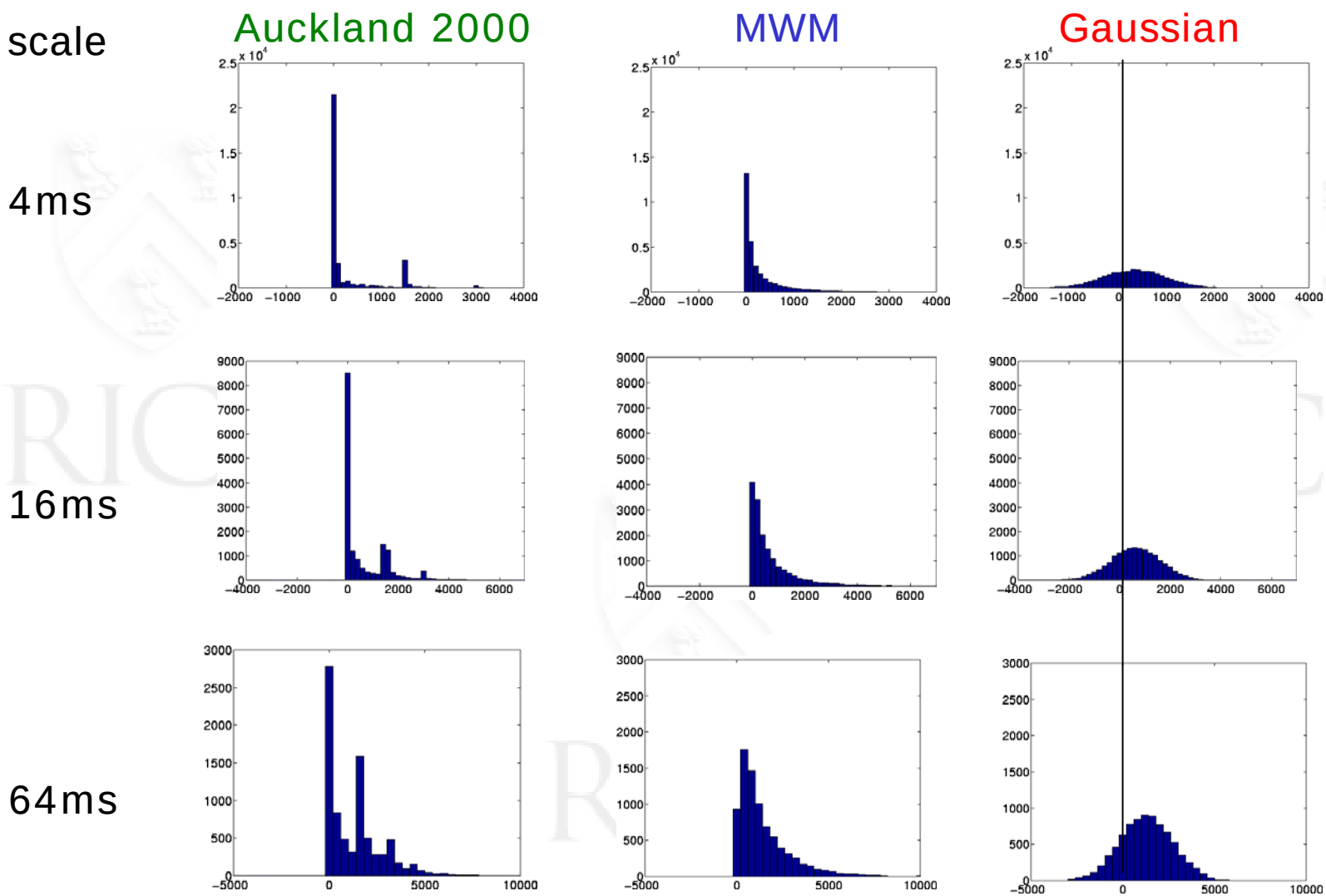


64ms

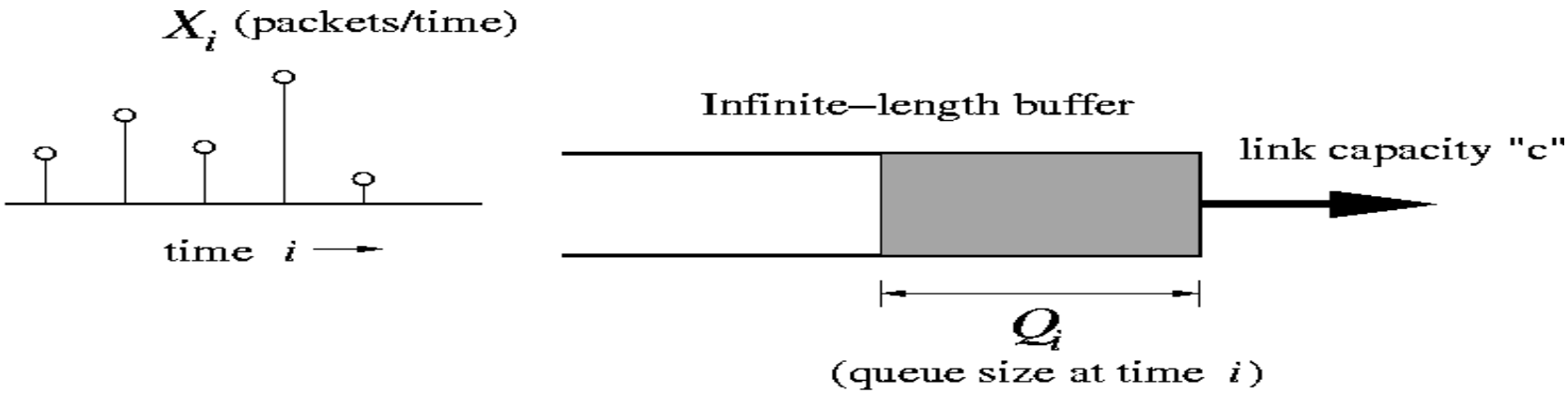




Marginal Matching



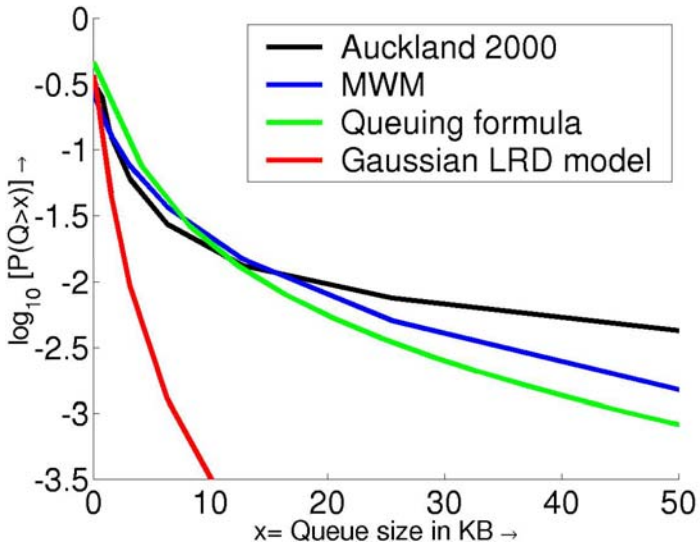
Multiscale Queuing



Summary:

- Tree models can accurately capture salient features of time series, such as Queuing of network traffic

- Multiplicative tree models superior to additive ones for network traffic

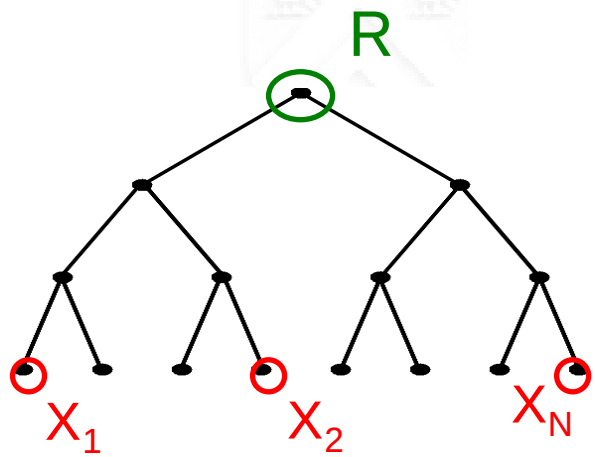


Optimal Estimation

Formulating the problem

Estimation Problem

- Find optimal choice of N leaf node samples to estimate tree root
- ...using the LMMSE

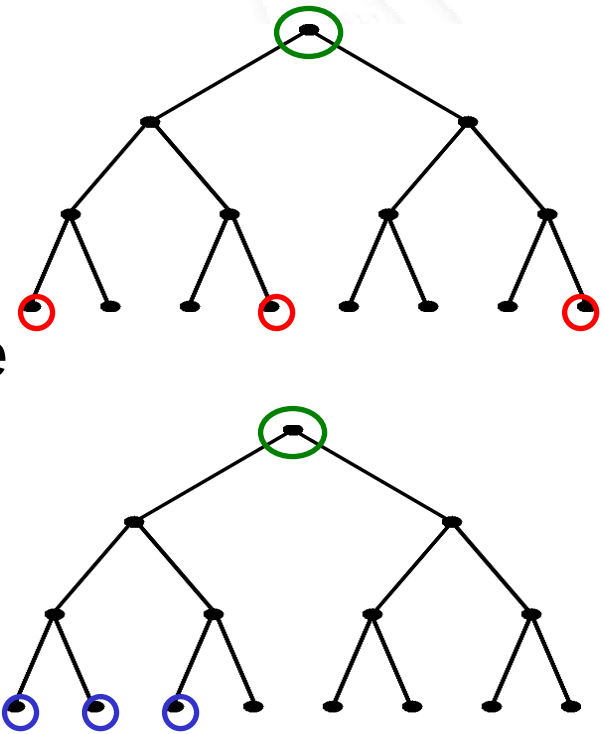


- Applications:
  - Time-series: Root gives the average over a large interval or square which may not be observable
  - Probing for traffic volume in Internet
  - Prediction from partial observation
  - Sensor Networks: average climate in a region

Estimation Problem

- Find optimal choice of N leaf node samples to estimate tree root
- ...using the LMMSE

- Intuition:
  - Positive correlation:
    Space samples as far as possible
    But how?
  - Negative correlation:
    Pull samples close together
    Next to each other?



Independent Innovation Tree

- Each child node is obtained from its parent node by adding an independent innovation
- Formally:

Given tree root R,

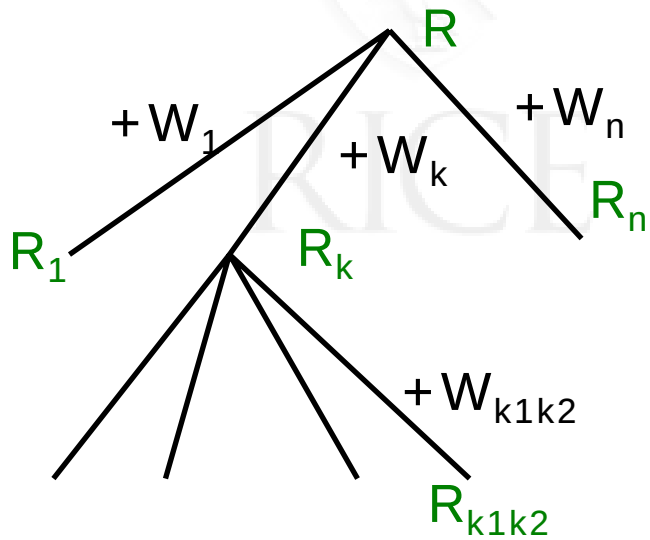
and independent innovations

W_k, $W_{k_1k_2}$, $W_{k_1k_2k_3}$

Define child nodes iteratively by

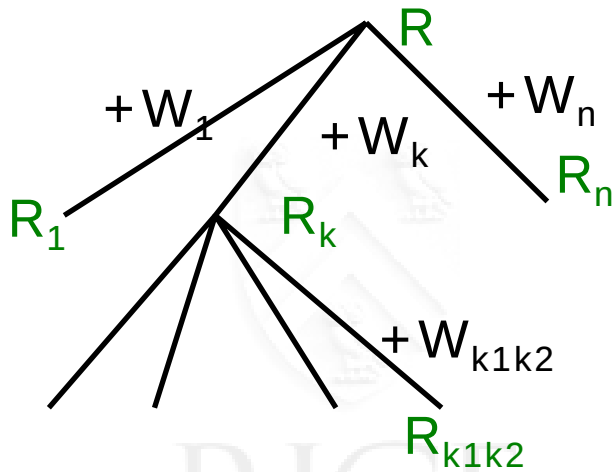
$R_k = R + W_k$

$R_{k_1k_2} = R_k + W_{k_1k_2}$



Independent Innovations (2)

Given tree root R,
and independent innovations
W_k, $W_{k_1k_2}$, $W_{k_1k_2k_3}$



- Simplified Correlation Structure:

$$R_{k_1\dots k_p} = R + W_k + W_{k_1k_2} + \dots + W_{k_1\dots k_p}$$

Thus

$$\mathrm{Cov}(R_{k_1\dots k_p}, R_{l_1\dots l_q}) = \mathrm{Var}(R_{k_1\dots k_m})$$

where $(k_1 \dots k_m) = (l_1 \dots l_m)$ and $k_{m+1} \neq l_{m+1}$.

Water-filling

An optimization method

and

Vital ingredient

A useful lemma

- Optimization of sum of concave functions
  - $\psi_k : \mathbb{N} \mapsto \mathbb{R}$ *concave* functions, i.e.,
    $\psi(n+1) - \psi(n) \geq \psi(n+2) - \psi(n+1)$
    and $\psi_k(0) = 0$
  - Wanted: $h(N) := \max \left\{ \sum_{k=1}^{K} \psi_k(n_k) \, : \, n_1 + \ldots + n_K = N \right\}.$
- Solution, iterative in N:
  - Lemma: *If* $h(N) = \sum_{k=1}^{K} \psi_k(g_k)$ *then* $h(N+1) = \sum_{k=1}^{K} \psi_k(\tilde{g}_k)$ *where*

$$\tilde{g}_k = \begin{cases} g_k + 1, & k = m \\ g_k, & k \neq m \end{cases}$$

and $\psi_m(g_m + 1) - \psi_m(g_m) \geq \psi_k(g_k + 1) - \psi_k(g_k)$ *for all* k

  - In other words, “Greedy waterfilling” solves the problem:

$$h(N+1) - h(N) \;=\; \max_k (\psi_k(g_k + 1) - \psi_k(g_k))$$

A useful lemma: anchor

- Optimization of sum of concave functions

$$h(N) := \max \left\{ \sum_{k=1}^{K} \psi_k(n_k) \,:\, n_1 + \ldots + n_K = N \right\}.$$

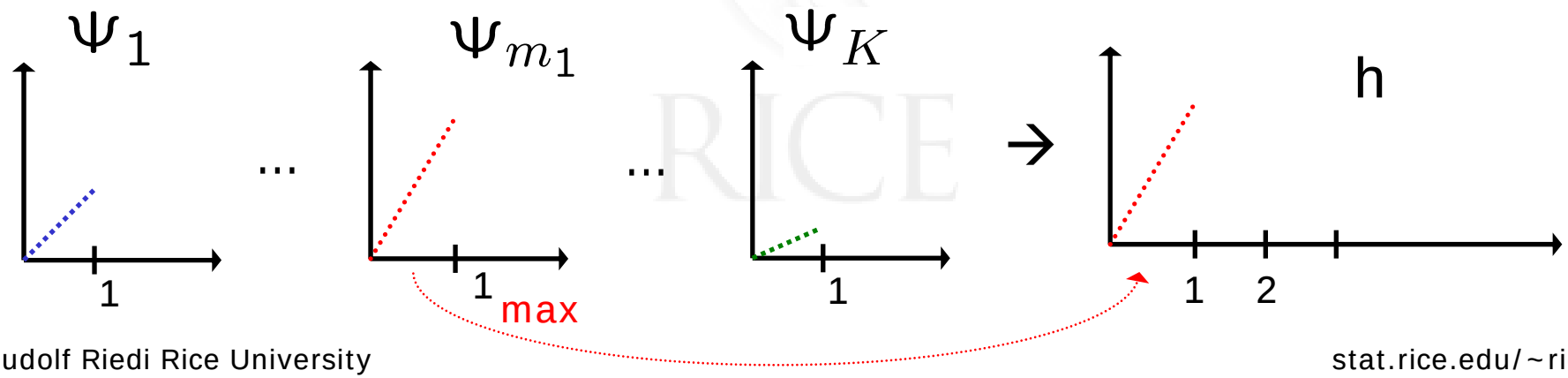
  - Lemma: *If* $h(N) = \sum_{k=1}^{K} \psi_k(g_k)$ *then*

$$h(N+1) - h(N) \;=\; \max_k (\psi_k(g_k + 1) - \psi_k(g_k))$$

  - Anchor:

$$h(0) \;=\; 0$$

$$h(1) \;=\; \max_k \psi_k(1) =: \psi_{m_1}(1)$$



A useful lemma: induction

- Optimization of sum of concave functions

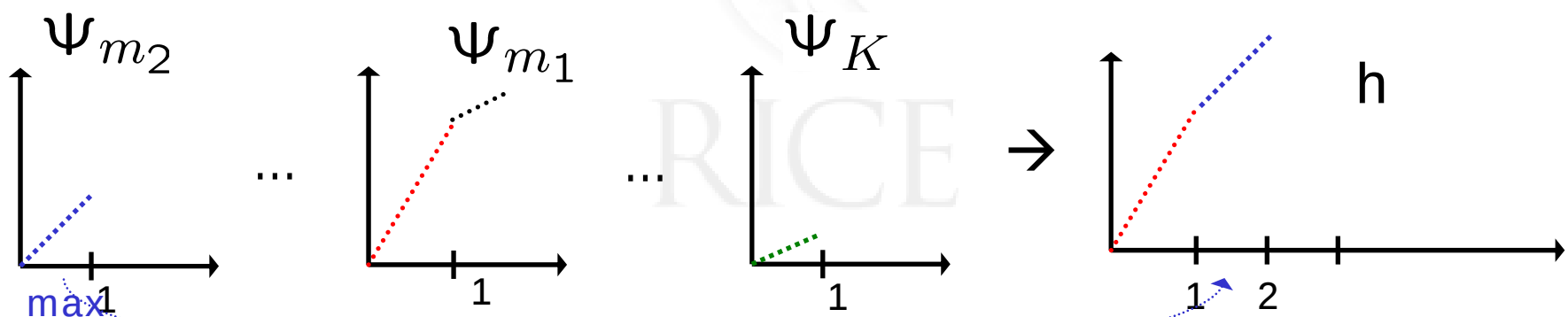
$$h(N) := \max \left\{ \sum_{k=1}^{K} \psi_k(n_k) \,:\, n_1 + \ldots + n_K = N \right\}.$$

  - Lemma: *If* $h(N) = \sum_{k=1}^{K} \psi_k(g_k)$ *then*

$$h(N+1) - h(N) \;=\; \max_k (\psi_k(g_k + 1) - \psi_k(g_k))$$

  - Induction :

$$\begin{aligned} h(2) &= h(1) + \max(\psi_{m_1}(2) - \psi_{m_1}(1), \max_{k \neq m_1} \psi_k(1)) \\ &= h(1) + \max_k \psi_k(g_k + 1) - \psi_k(g_k)) \end{aligned}$$



A useful lemma: alternative view

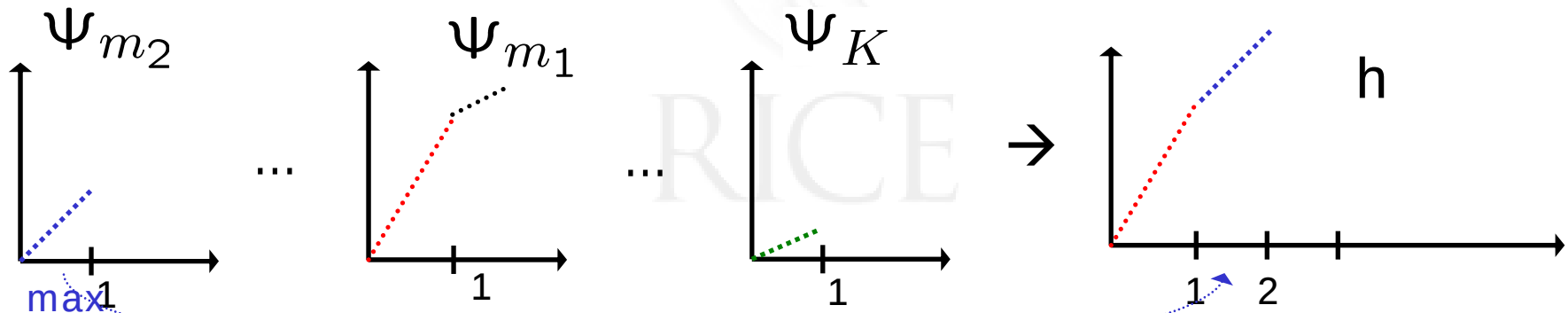
- Optimization of sum of concave functions

$$h(N) := \max \left\{ \sum_{k=1}^{K} \psi_k(n_k) \,:\, n_1 + \ldots + n_K = N \right\}.$$

  - Lemma: *If* $h(N) = \sum_{k=1}^{K} \psi_k(g_k)$ *then*

$$h(N+1) - h(N) \;=\; \max_k (\psi_k(g_k + 1) - \psi_k(g_k))$$

  - Alternative view: order increments $\{\psi_k(n+1) - \psi_k(n)\}_{k,n}$ according to size. Due to concavity they come in increasing variable for each ψ_k



Optimal Tree Estimation

Iterative solution

Recall: problem setting

- Given N
  - N = number of leaf nodes available for estimation
- Minimize Var(root | leaves) over Λ_N,
  - Λ_N = collection of all sets of N leaf nodes L

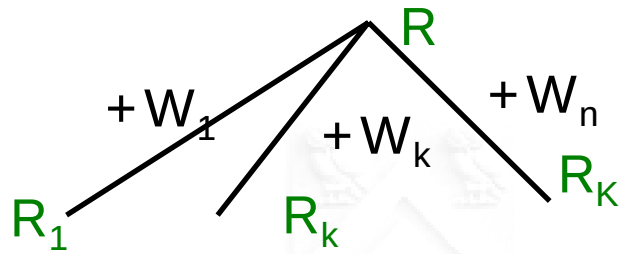
$$\min_{L \in \Lambda_N} \text{var}(R|L)$$

- Simple correlation structure on trees
  - Correlations depend only on common ancestors
  - Optimal configuration depends only on the number of samples per sub-tree

Subtrees

- Key-lemma [Willsky '02]:

$$\frac{1}{\mathrm{var}(R|L)} + \frac{K-1}{\mathrm{var}(R)} = \sum_{k=1}^{K} \frac{1}{\mathrm{var}(R|R_k \cap L)}$$



R_k: subtree, root of subtree

- Consequence:

$$\min_{L \in \Lambda_N} \mathrm{var}(R|L) \Leftrightarrow \max_{L \in \Lambda_N} \sum_{k=1}^{K} \frac{1}{\mathrm{var}(R|R_k \cap L)}$$

- Divide and Conquer:
  - Optimal configuration ←→ number of samples per sub-tree

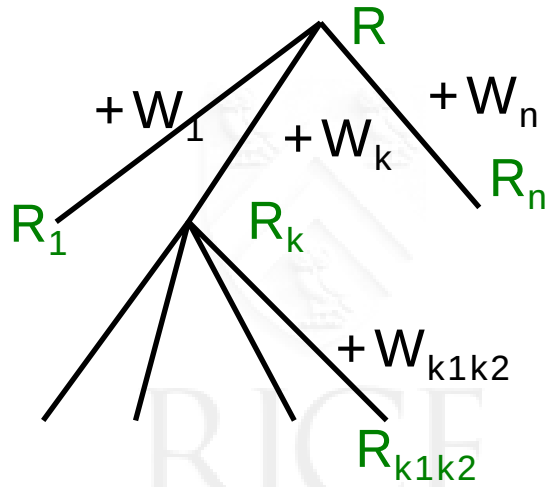
$$\max_{L \in \Lambda_N} \sum_{k=1}^{K} \frac{1}{\mathrm{var}(R|R_k \cap L)} = \max_{n_1 + \ldots + n_K = N} \sum_{k=1}^{K} \max_{L \in \Lambda_{n_k}} \frac{1}{\mathrm{var}(R|R_k \cap L)}$$

Recursion and water-filling

- Recall:

$$\min_{L \in \Lambda_N} \mathrm{var}(R|L) \Leftrightarrow \max_{L \in \Lambda_N} \sum_{k=1}^{K} \frac{1}{\mathrm{var}(R|R_k \cap L)}$$

$$\Leftrightarrow \max_{n_1+\ldots+n_K=N} \sum_{k=1}^{K} \underbrace{\max_{L \in \Lambda_{n_k}} \frac{1}{\mathrm{var}(R|R_k \cap L)}}_{\psi_k(n_k)}$$



- Technical lemma:
  - (a) $\psi_k(\cdot)$ is concave
  - (b) $\psi_k(n) = \max_{L \in \Lambda_n} \frac{1}{\mathrm{var}(R_k|R_k \cap L)}$

- Meaning:
  - (a): Water-filling applies
  - (b): Recursively find $\psi_k(n)$ as optimal LMMS error

Optimal estimation

- Optimal N leaf nodes for the LSMME of root
  - can be found recursively via subtrees.
  - and iteratively with respect to N

- Numerically efficient search: Water-filling
  - Given solution for N leaf nodes
  - Computation: start with leaves and compute improvement for each subtree for adding one node to it; work up to root
  - Placement: On each level add the new leaf node to one subtree with largest improvement;
  - Overall cost: N*total number of leaves
  - Avoids searching all possible placements (exponential cost)

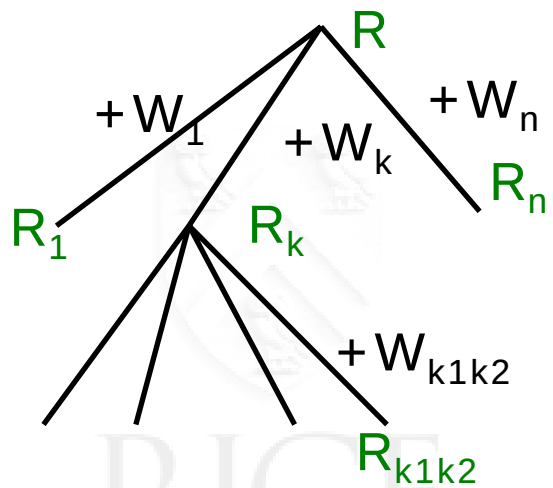
Corollaries

- Special case:
  - Assume: Var of innovation depends only on depth
  - Then: All Ψ_k are identical and
    Optimal placement means homogeneous, or well-balanced

- Generalization to covariance trees:
  - $\mathrm{Cov}(R_{k_1 \ldots k_n}, R_{l_1 \ldots l_n}) = \mathrm{Var}(R_{k_1 \ldots k_m}) =: \phi(m)$
  - If ϕ is increasing in m, then:
    - Homogeneous placement is optimal
    - Next-neighbor samples are provably worst
  - if ϕ is decreasing in m, then
    - Homogeneous provably worst.

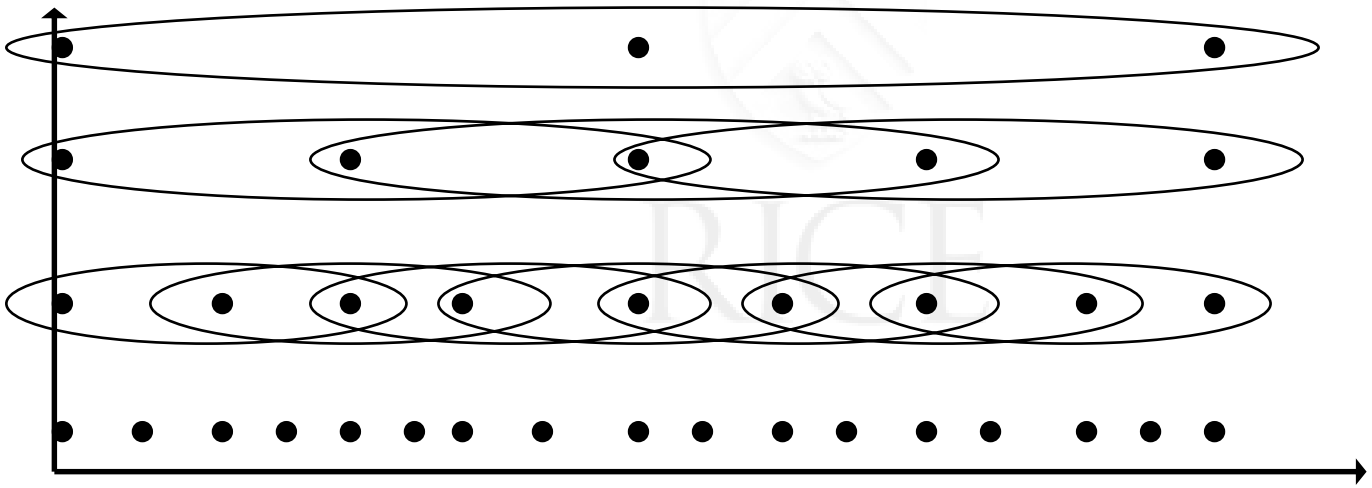
Growing the tree

- Refining the “resolution”
  - add subtrees at leaves
  - Recompute improvements with the new leaves which are one level further out
  - Since the solution is recursive by subtrees, we only need to assign the leaf node to the most effective place in the new subtrees.



Summary

- Exploit tree structure to develop intuitive and simple strategies
- Homogeneous sampling is optimal for a positively correlated statistical tree
- Future: Generalize to vector-trees
  - Towards overcoming non-stationarities



Willsky Tree of Triples