Statistical regular sub-pavings in multivariate density estimation

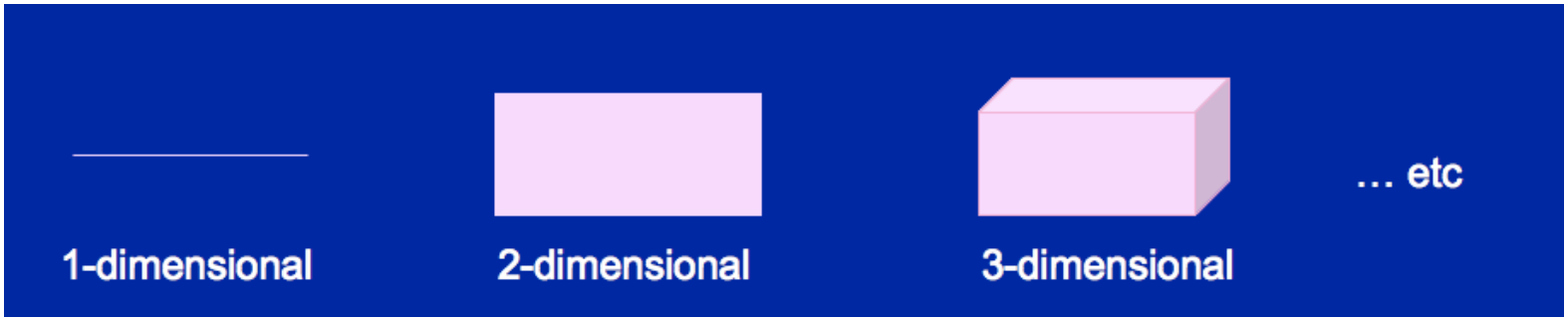
Jennifer Harlow, Dominic Lee, **Raaz**esh Sainudiin and Gloria Teng

Department of Mathematics and Statistics, University of Canterbury, Christchurch, New Zealand

Uppsala, Sweden, December 2009

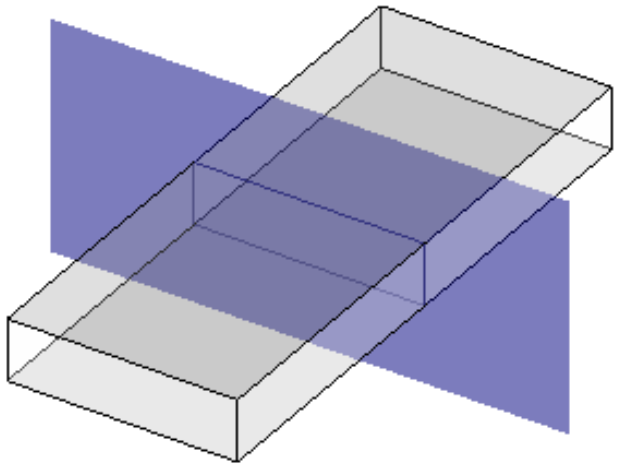
Intervals, Boxes and Regular Sub-pavings

Intervals and **Boxes** as interval vectors. Interval analysis allows us to perform arithmetic with boxes.

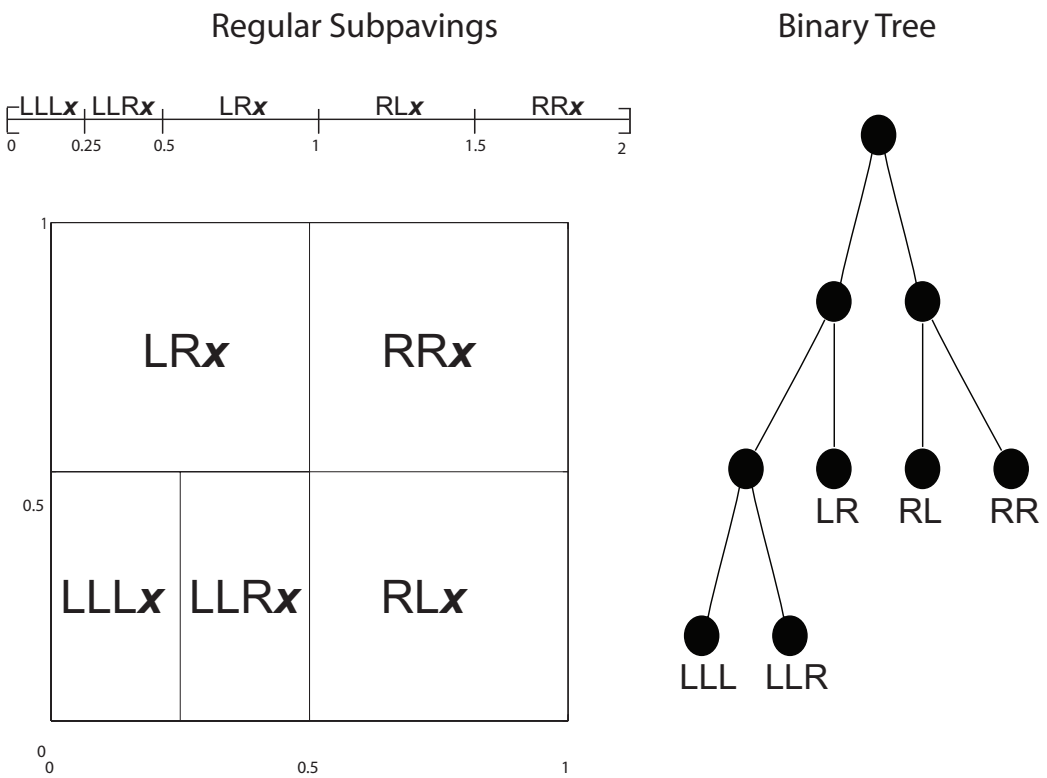


Intervals, Boxes and Regular Sub-pavings

Regular Subpavings (RSP): A regular subpaving is formed by successive bisections of boxes along the first widest side, starting with the root box **x**. Some of these boxes are selected while others are discarded in a RSP. This class of subsets or partitions of a box **x** is closed under unions, intersections and as m_∞-close as desired to any subset in the box **x** (Jaulin, Kieffer, Didrit and Walter, 2001)



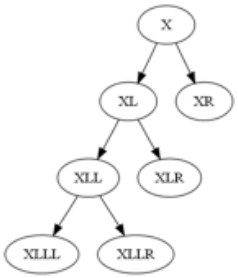
Regular Sub-pavings, Partitions and Binary Trees



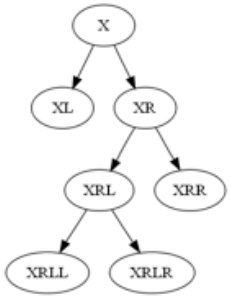
Two regular subpavings (RSPs) in 1D and 2D respectively. The corresponding binary tree of both RSPs is shown on the right.

State Space of Regular Sub-pavings

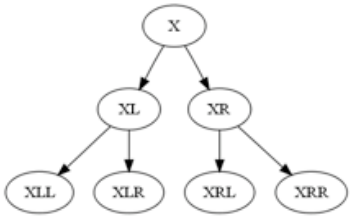
Leaf-depth encoded RSPs



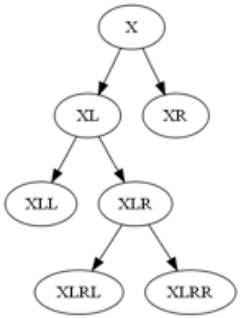
(3, 3, 2, 1)



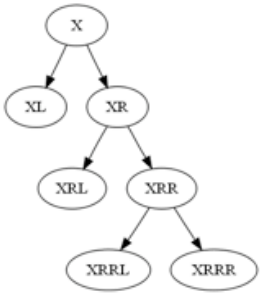
(1, 3, 3, 2)



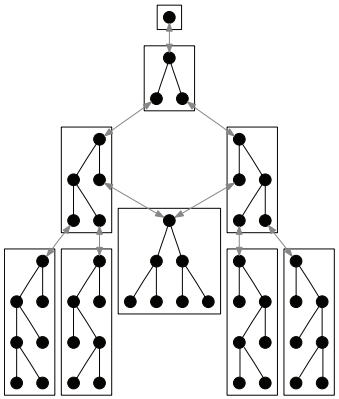
(2, 2, 2, 2)



(2, 3, 3, 1)

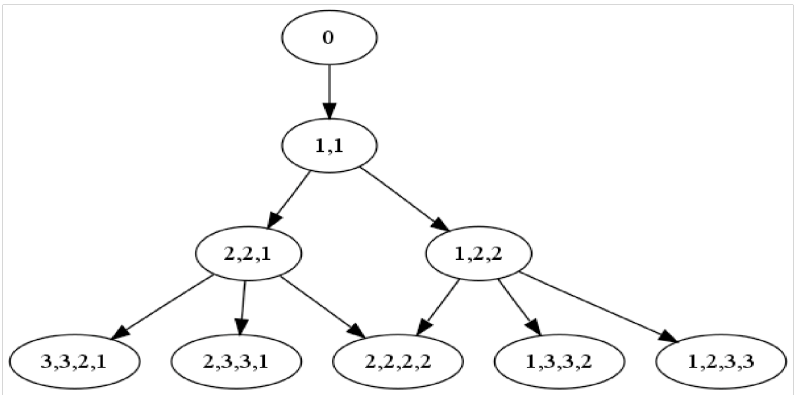


(1, 2, 3, 3)



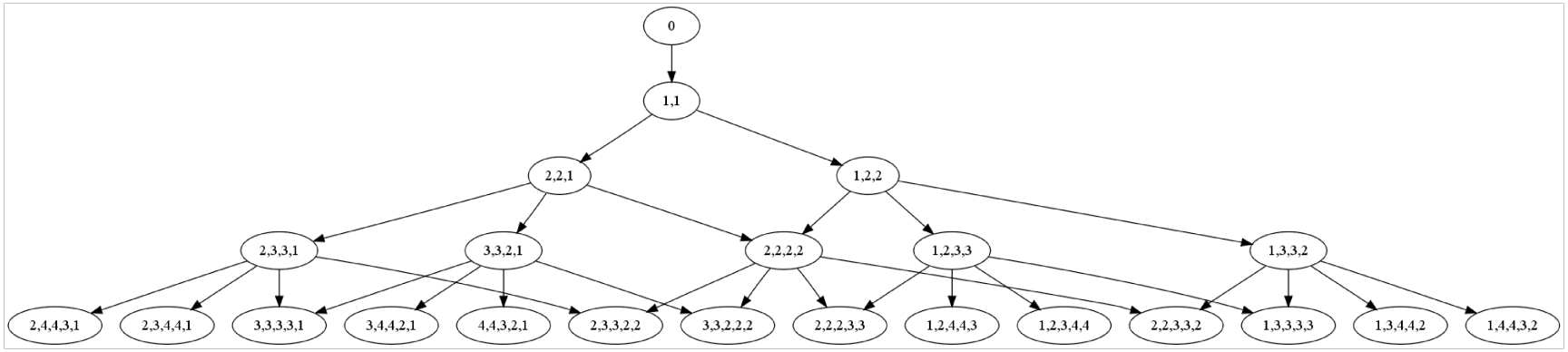
C_k RSPs wit k splits

$$
\begin{aligned}
C_0 &= 1 \\
C_1 &= 1 \\
C_2 &= 3 \\
C_3 &= 5 \\
C_4 &= 14 \\
C_5 &= 429 \\
C_k &= \frac{2k!}{(k+1)!k!}
\end{aligned}
$$



State Transition Diagram of Regular Sub-pavings

State Transition Diagram of RSPs with 0, 1, 2, 3 and 4 splits.



1. The above state space is denoted by $\mathcal{RSP}_{0:4}$
2. Number of RSPs with k splits is the Catalan number C_k
3. There is more than one way to reach a RSP by k splits
4. Randomized algorithms of interest are often Markov chains on $\mathcal{RSP}_{0:\infty}$

Adaptive Multidim. Data-structure for Massive Data

1. Massive Data $X_{1:n} := X_1, X_2, \ldots, X_n \sim f^* : \mathbb{R}^d \to \mathbb{R}$

Adaptive Multidim. Data-structure for Massive Data

1. Massive Data $X_{1:n} := X_1, X_2, \ldots, X_n \sim f^* : \mathbb{R}^d \to \mathbb{R}$
2. Here n is so large that $X_{1:n}$ cannot be represented in RAM

Adaptive Multidim. Data-structure for Massive Data

1. Massive Data $X_{1:n} := X_1, X_2, \ldots, X_n \sim f^* : \mathbb{R}^d \to \mathbb{R}$
2. Here n is so large that $X_{1:n}$ cannot be represented in RAM
3. We want an adaptive multidimensional metric data-structure for $X_{1:n}$ on $\mathcal{RSP}_{0:\infty}$ that caches certain statistics of $X_{1:n}$ and easily represented in RAM

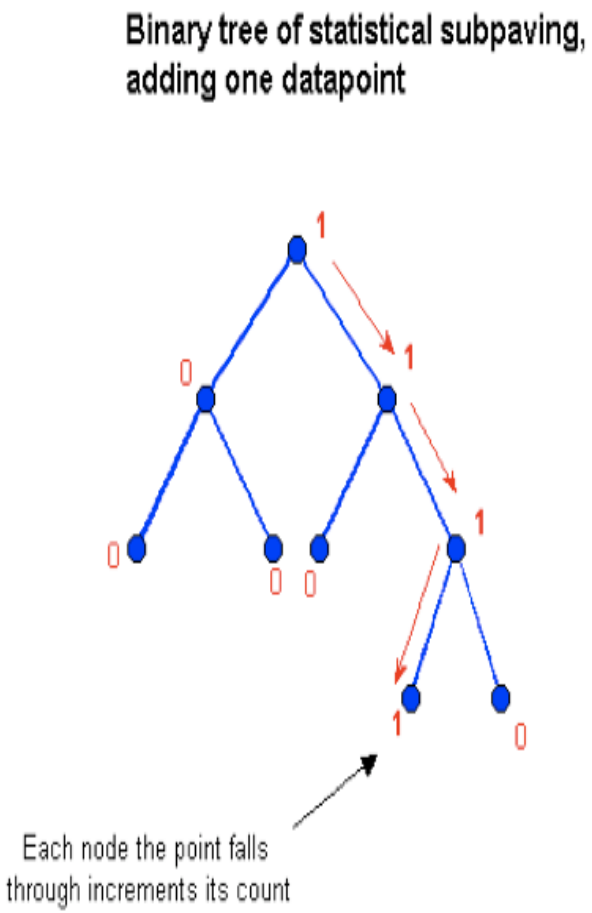
Adaptive Multidim. Data-structure for Massive Data

1. Massive Data $X_{1:n} := X_1, X_2, \ldots, X_n \sim f^* : \mathbb{R}^d \to \mathbb{R}$
2. Here n is so large that $X_{1:n}$ cannot be represented in RAM
3. We want an adaptive multidimensional metric data-structure for $X_{1:n}$ on $\mathcal{RSP}_{0:\infty}$ that caches certain statistics of $X_{1:n}$ and easily represented in RAM
4. Statistical Regular Sub-paving (SRSP) is one such data-structure

Adaptive Multidim. Data-structure for Massive Data

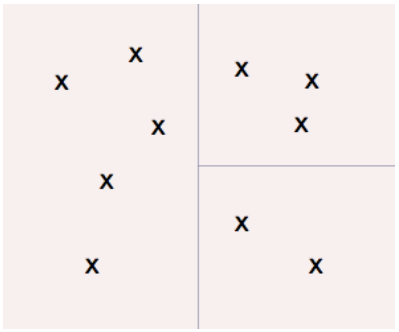
1. Massive Data $X_{1:n} := X_1, X_2, \ldots, X_n \sim f^* : \mathbb{R}^d \to \mathbb{R}$
2. Here n is so large that $X_{1:n}$ cannot be represented in RAM
3. We want an adaptive multidimensional metric data-structure for $X_{1:n}$ on $\mathcal{RSP}_{0:\infty}$ that caches certain statistics of $X_{1:n}$ and easily represented in RAM
4. Statistical Regular Sub-paving (SRSP) is one such data-structure
5. We can extend arithmetic to SRSPs

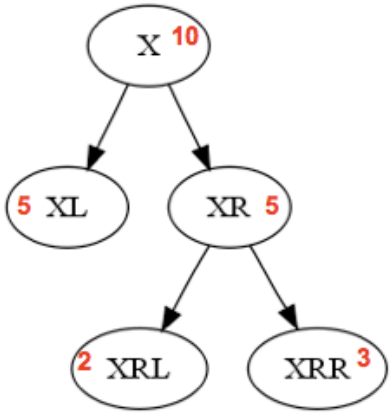
Caching Recursively Computable Statistics

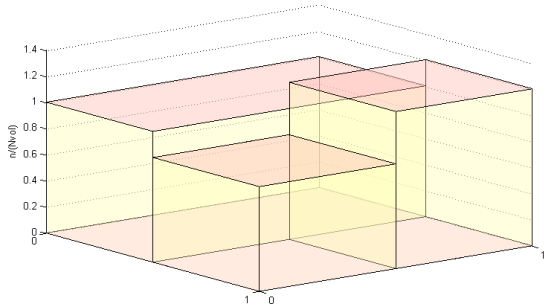


1. As Data $X_{1:n}$ falls through the SRSP
2. SRSP caches recursively computable statistics at each node or box:
   - sample count
   - sample mean vector
   - sample variance-covariance matrix
   - volume of the box
3. These cached statistics are sufficient for certain decision problems or their enclosures.

Statistical Regular Sub-pavings for Density Estimation

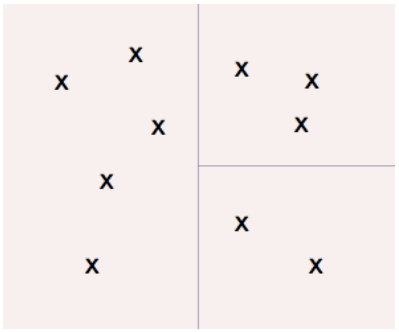


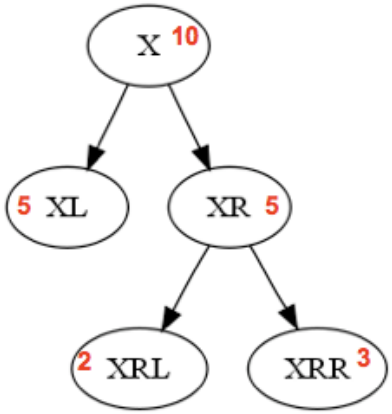


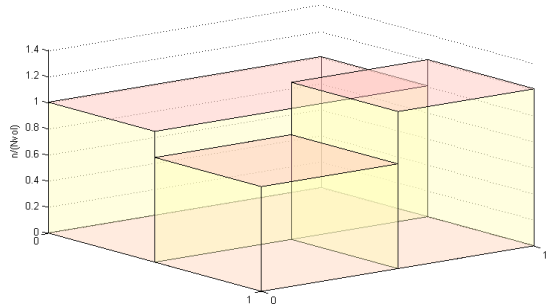


- SRSP represents data $X_1, X_2, \ldots, X_n \sim f^* : \mathbb{R}^d \to \mathbb{R}$.

Statistical Regular Sub-pavings for Density Estimation

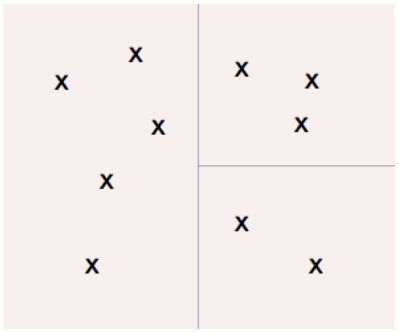


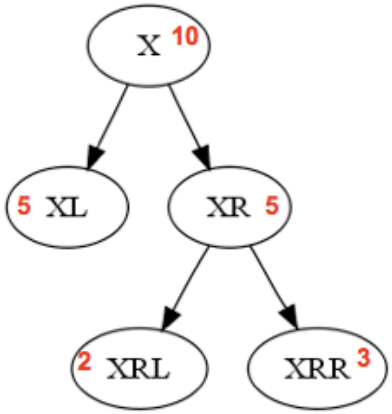


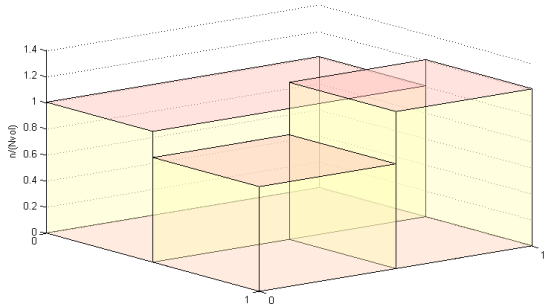


- SRSP represents data $X_1, X_2, \ldots, X_n \sim f^* : \mathbb{R}^d \to \mathbb{R}$.
- Each data point will be contained in a leaf-box of SRSP

Statistical Regular Sub-pavings for Density Estimation

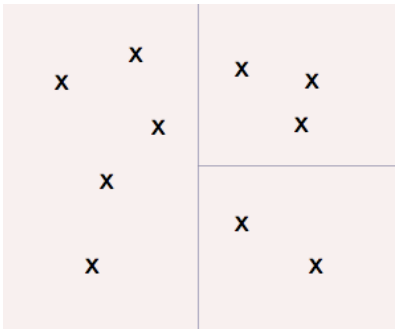


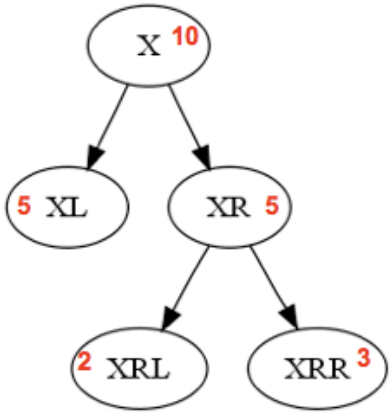


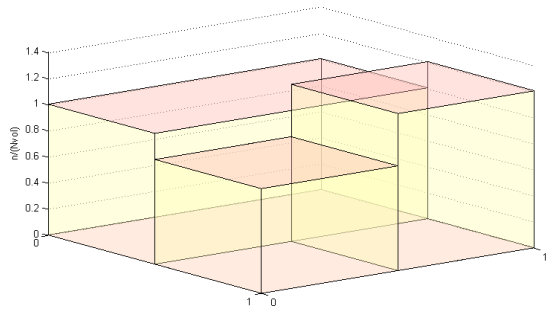


- SRSP represents data $X_1, X_2, \ldots, X_n \sim f^* : \mathbb{R}^d \to \mathbb{R}$.
- Each data point will be contained in a leaf-box of SRSP
- Update the recursively computable count statistic at each node as data passes through

Statistical Regular Sub-pavings for Density Estimation

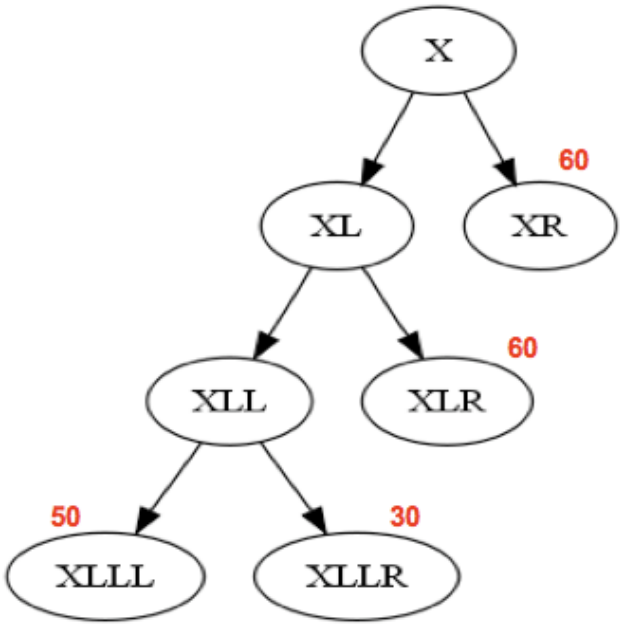


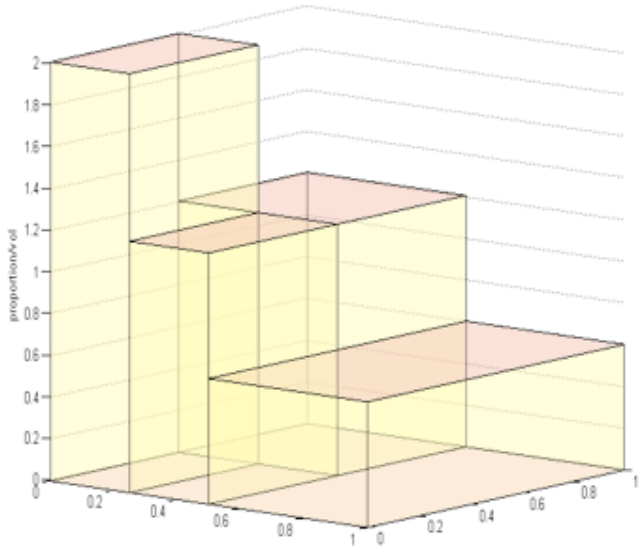




- SRSP represents data $X_1, X_2, \ldots, X_n \sim f^* : \mathbb{R}^d \to \mathbb{R}$.
- Each data point will be contained in a leaf-box of SRSP
- Update the recursively computable count statistic at each node as data passes through
- Adaptively grow/prune the SRSP tree, i.e. split/merge boxes

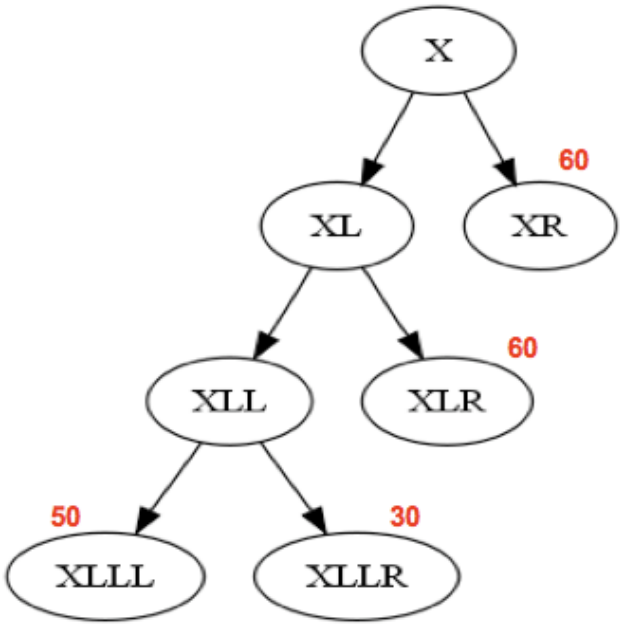
Adaptive Histograms from Statistical Sub-pavings

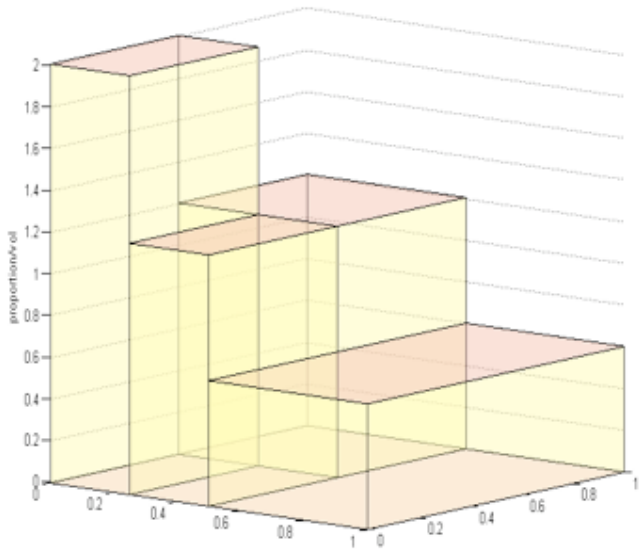




- Consider the leaf nodes as histogram bins $b_1, b_2, \ldots, b_L$

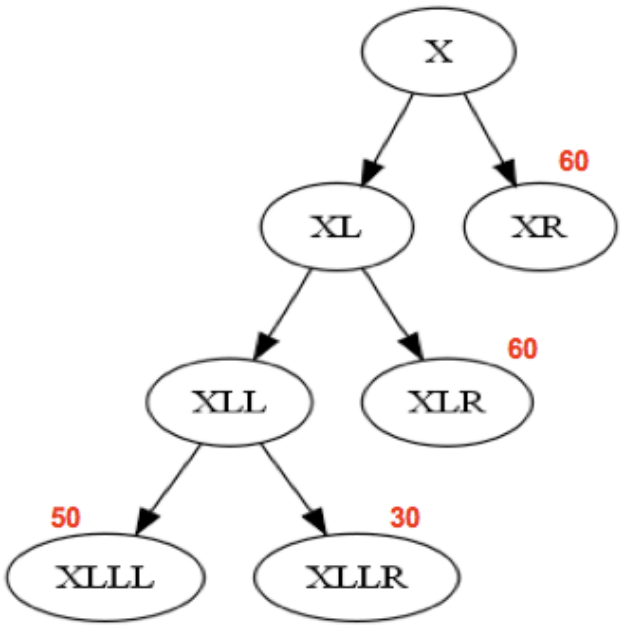
Adaptive Histograms from Statistical Sub-pavings

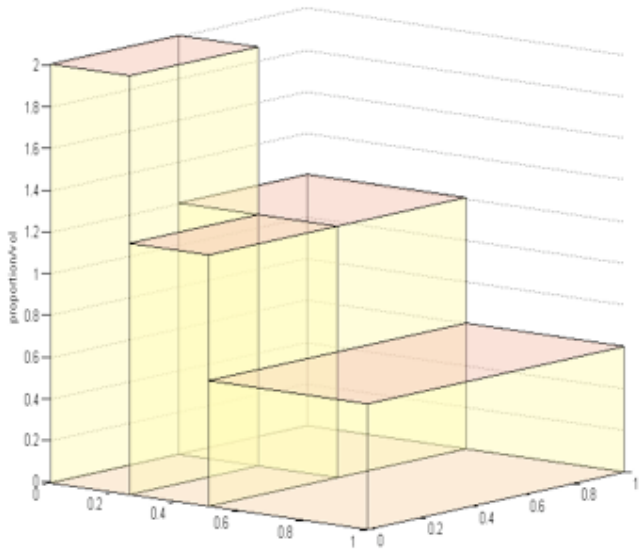




- Consider the leaf nodes as histogram bins $b_1, b_2, \ldots, b_L$
- Let $n_j :=$ # of data points in b_j and $v_j =$ volume of bin j

Adaptive Histograms from Statistical Sub-pavings

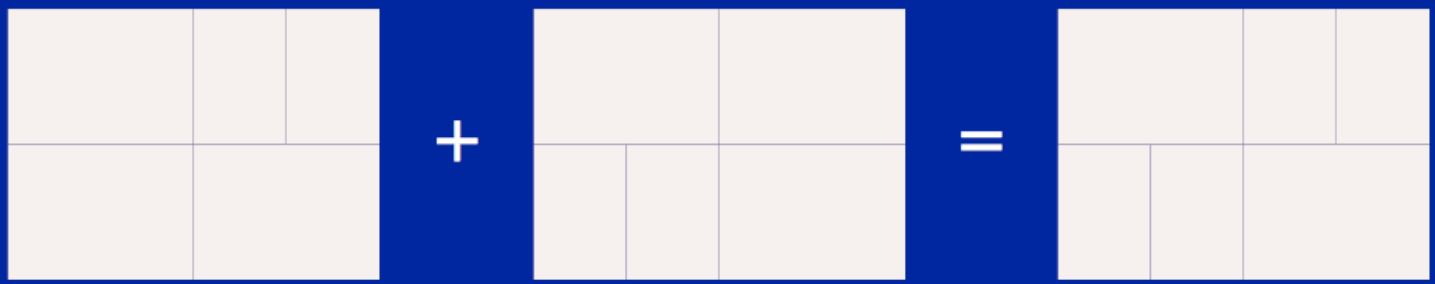




- Consider the leaf nodes as histogram bins $b_1, b_2, \ldots, b_L$
- Let $n_j :=$ # of data points in b_j and $v_j =$ volume of bin j
- Histogram Estimate: $\hat{f}(x; X_1, X_2, \ldots, X_n) = \sum_{j=1}^{L} \mathbf{1}_{b_j}(x) \frac{n_j}{n\ v_j}$

Averaging Two Histograms

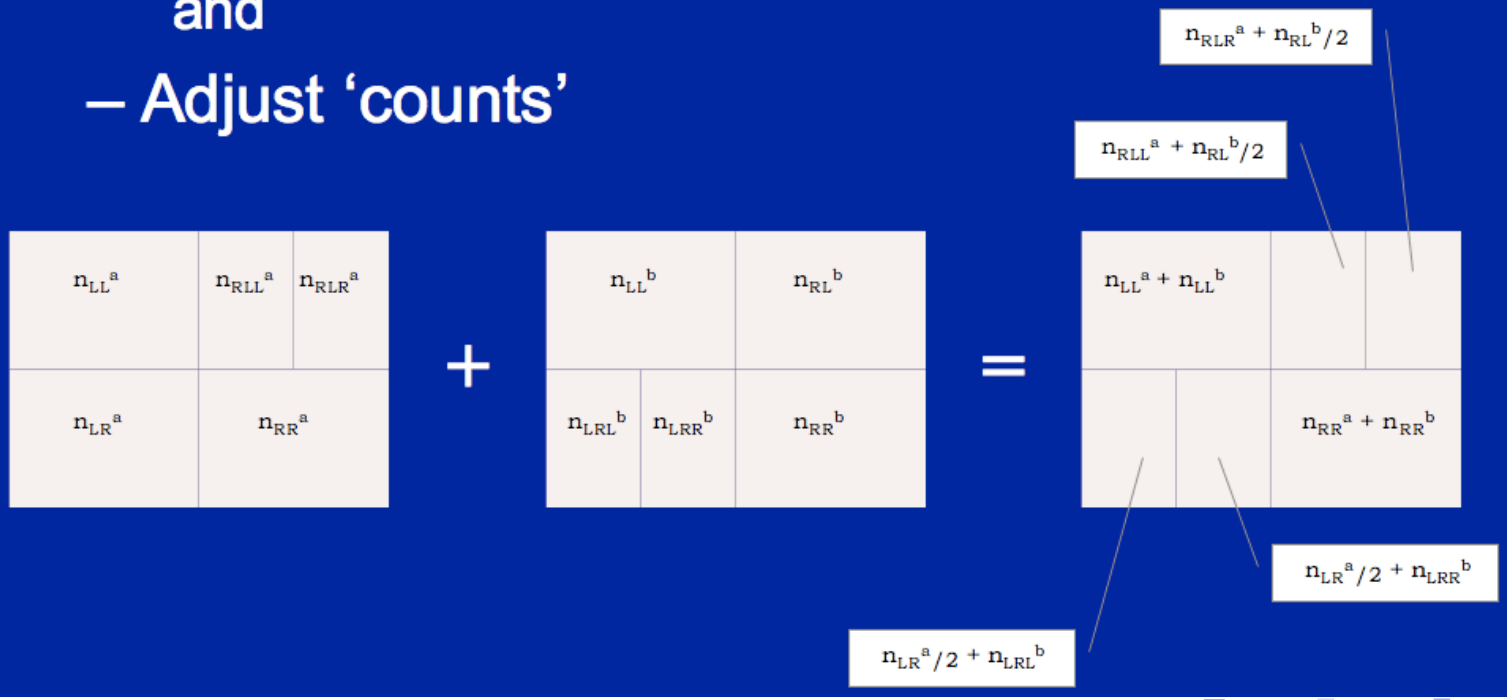
- Adding plain vanilla subpavings
  - Addition ~ non-minimal union



Averaging Two Histograms

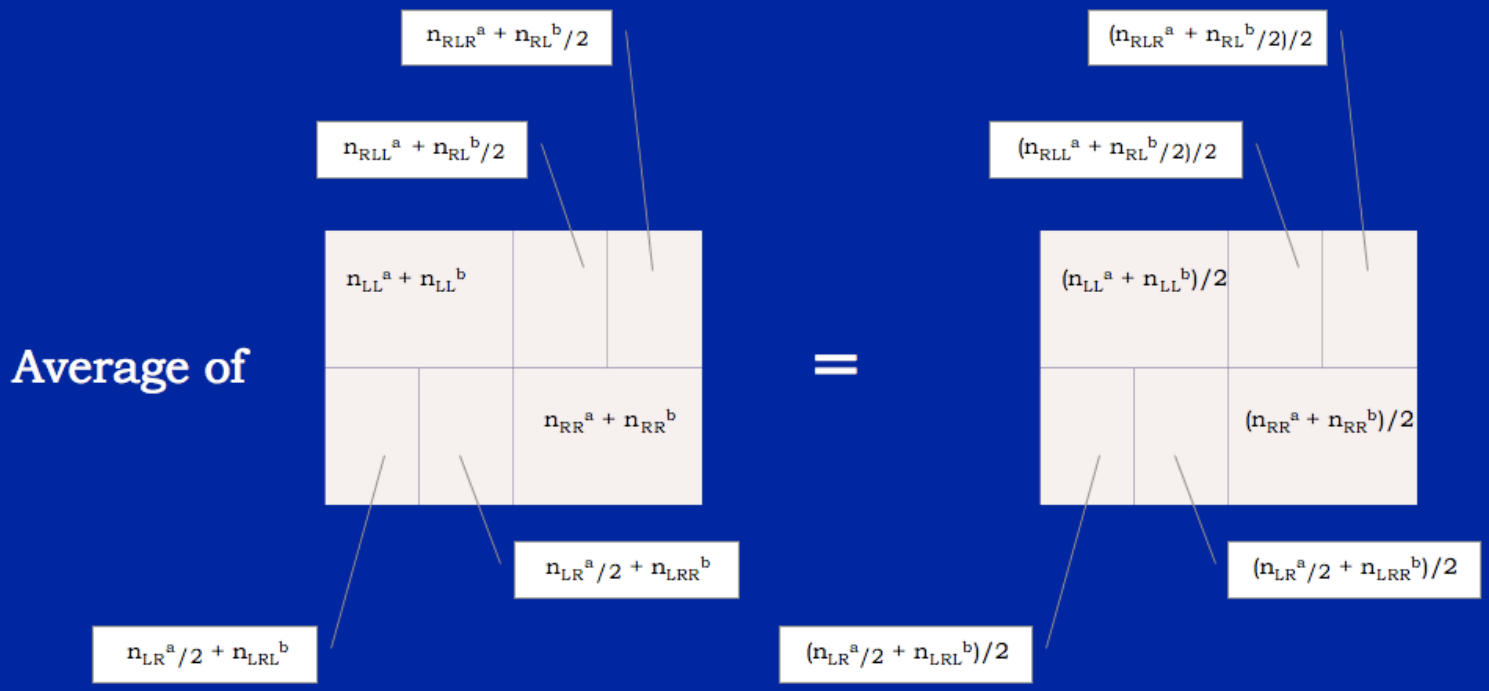
- Adding statistical subpavings
  - Plain vanilla addition

    and
  - Adjust 'counts'



Averaging Two Histograms

- Average the adjusted 'counts' over the number of histograms



Posterior Distribution over Histograms in $\mathcal{RSP}_{0:\infty}$

- Let m be a histogram with partition $\Pi(m)$ given by the leaves of a RSP with k splits and $k+1$ leaves in $\mathcal{RSP}_k$

Posterior Distribution over Histograms in $\mathcal{RSP}_{0:\infty}$

- Let m be a histogram with partition $\Pi(m)$ given by the leaves of a RSP with k splits and $k+1$ leaves in $\mathcal{RSP}_k$
- Then for this partition, the most likely histogram estimate is

$$\hat{f}(x; \text{data}) = \hat{f}(x; X_{1:n}) = \sum_{j=1}^{k+1} \mathbf{1}_{b_j}(x) \frac{n_j}{n\, v_j}$$

Posterior Distribution over Histograms in $\mathcal{RSP}_{0:\infty}$

- Let m be a histogram with partition $\Pi(m)$ given by the leaves of a RSP with k splits and $k+1$ leaves in $\mathcal{RSP}_k$
- Then for this partition, the most likely histogram estimate is

$$\hat{f}(x; \text{data}) = \hat{f}(x; X_{1:n}) = \sum_{j=1}^{k+1} \mathbf{1}_{b_j}(x) \frac{n_j}{n\, v_j}$$

- Let the prior probability be $P(m) \propto \frac{1}{C_k^2}$, $\Pi(m) \in \mathcal{RSP}_{0:\infty}$

Posterior Distribution over Histograms in $\mathcal{RSP}_{0:\infty}$

- Let m be a histogram with partition $\Pi(m)$ given by the leaves of a RSP with k splits and $k+1$ leaves in $\mathcal{RSP}_k$
- Then for this partition, the most likely histogram estimate is

$$\hat{f}(x; \text{data}) = \hat{f}(x; X_{1:n}) = \sum_{j=1}^{k+1} \mathbf{1}_{b_j}(x) \frac{n_j}{n\, v_j}$$

- Let the prior probability be $P(m) \propto \frac{1}{C_k^2}$, $\Pi(m) \in \mathcal{RSP}_{0:\infty}$
- Then the posterior density of histogram m with k splits is

$$P(m|X_{1:n}) \propto P(X_{1:n}|m)P(m) = \prod_{j=1}^{k+1} \left(\frac{n_j}{n\, v_j}\right)^{n_j} \frac{1}{C_k^2}$$

Metropolis-Hastings Algorithm

Use a proposal density $Q(m'; m^t)$ which depends on the current state m^t, to generate a new proposed state m'

Draw $u \in U(0, 1)$

Accept the proposal if $u < \frac{P(m')Q(m^t; m')}{P(m^t)Q(m'; m^t)}$

If the proposal is accepted, $m^{t+1} \leftarrow m'$

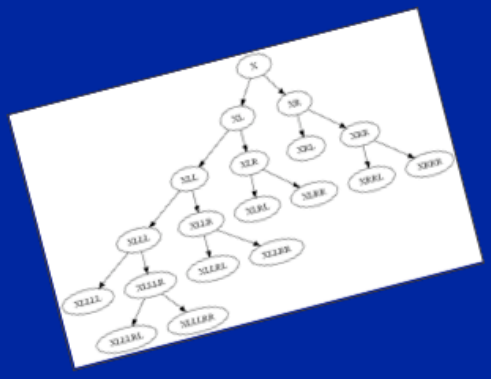
otherwise $m^{t+1} \leftarrow m^t$

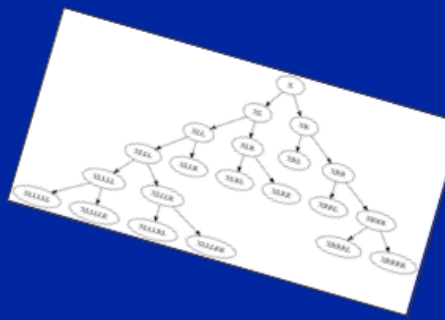
And then start again with a new proposal $m' \ldots$

Metropolis-Hastings Algorithm

- Start from some initial state m^0

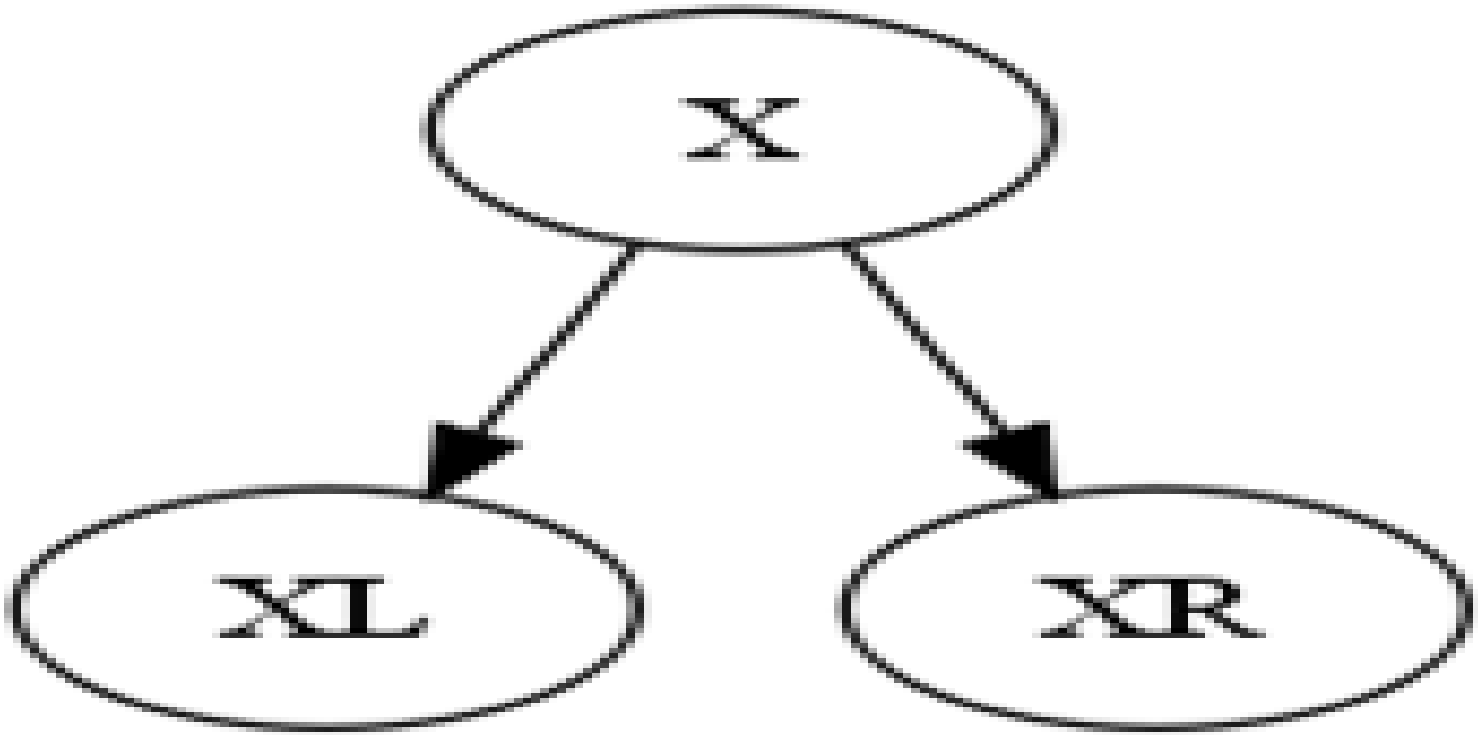
- Burn-in: run until initial state is 'forgotten'
- States after burn-in are sample histograms



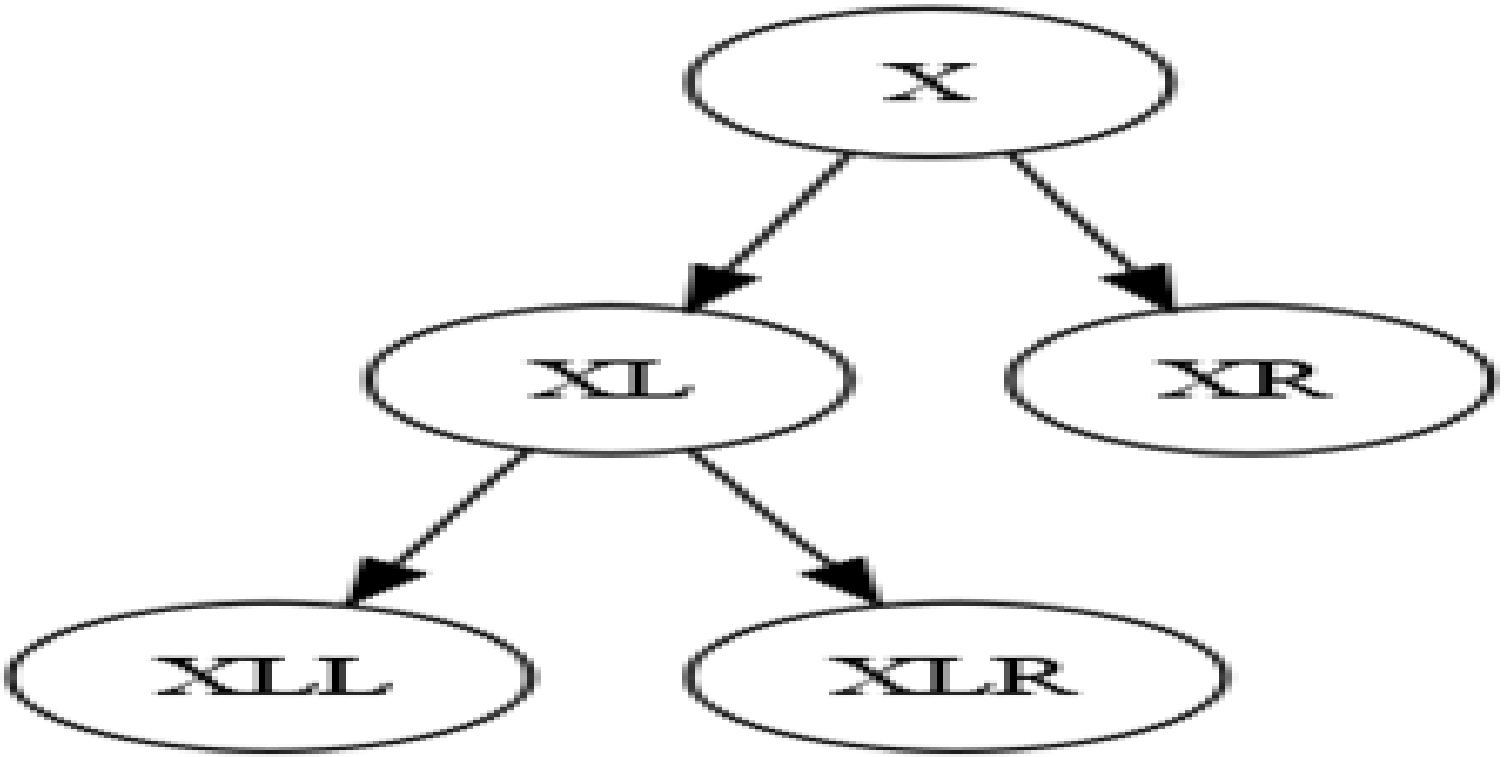


etc...

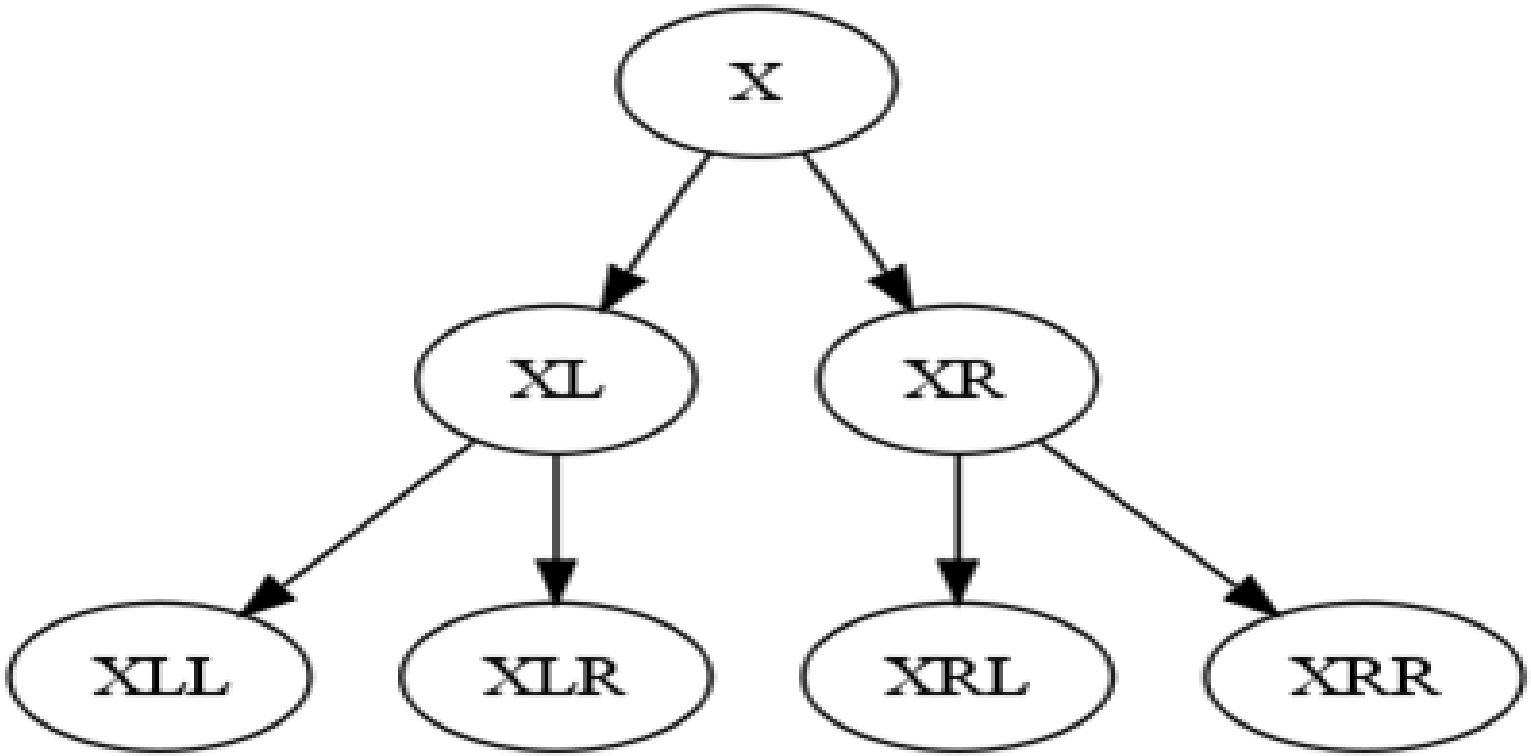
Monte Carlo Markov Chain over Histograms in $\mathcal{RSP}_{0:\infty}$



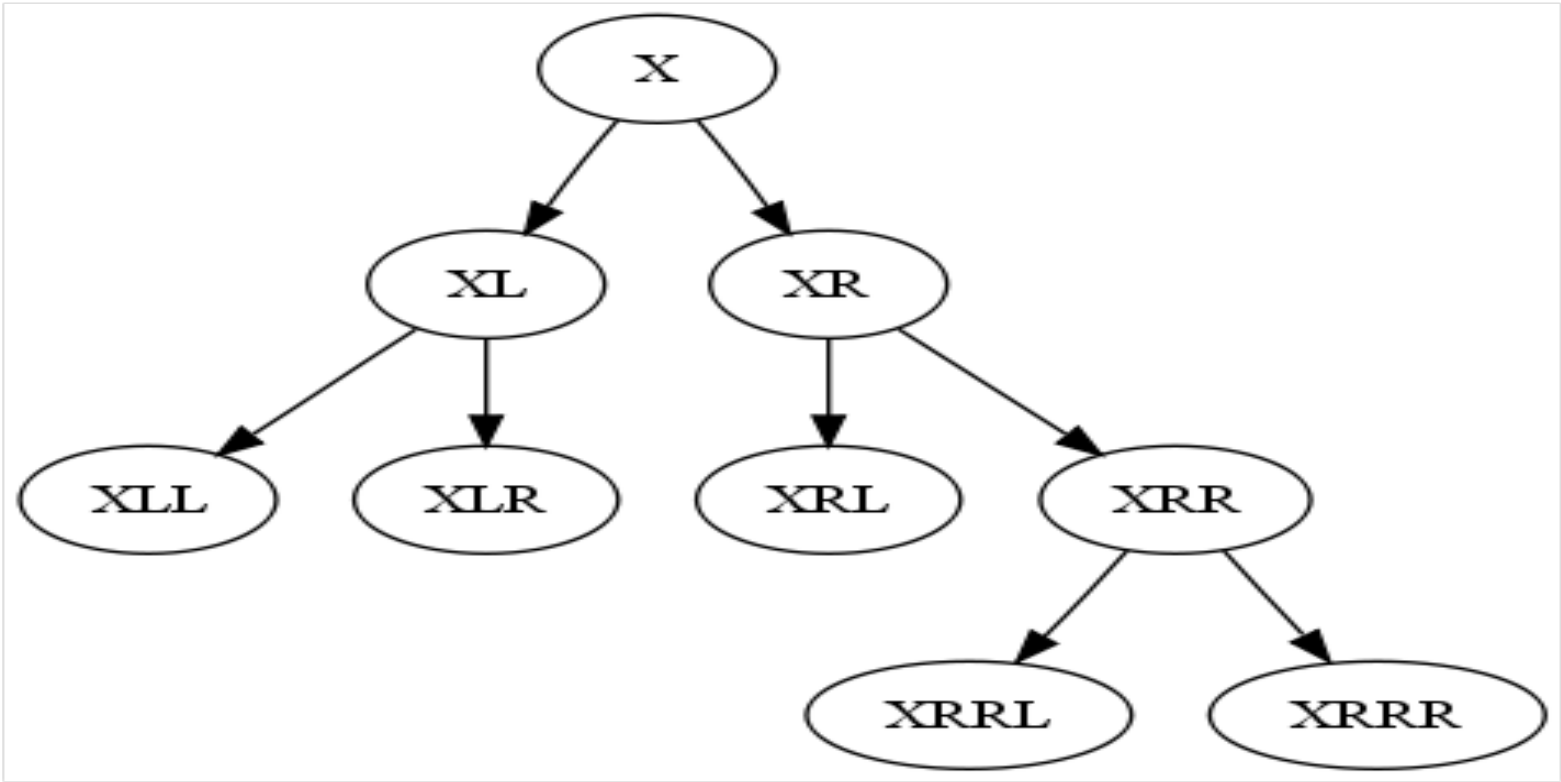
Monte Carlo Markov Chain over Histograms in $\mathcal{RSP}_{0:\infty}$



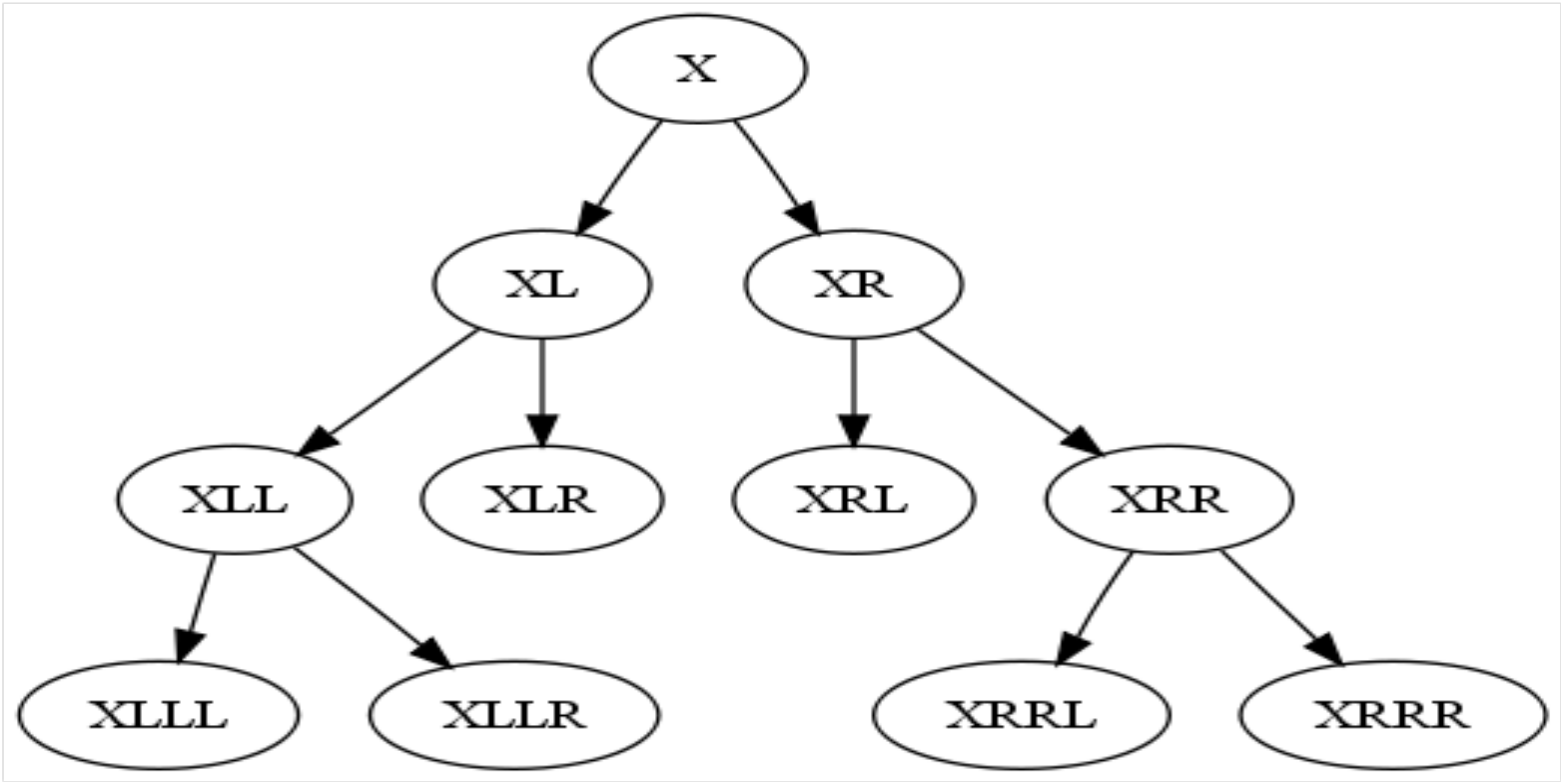
Monte Carlo Markov Chain over Histograms in $\mathcal{RSP}_{0:\infty}$



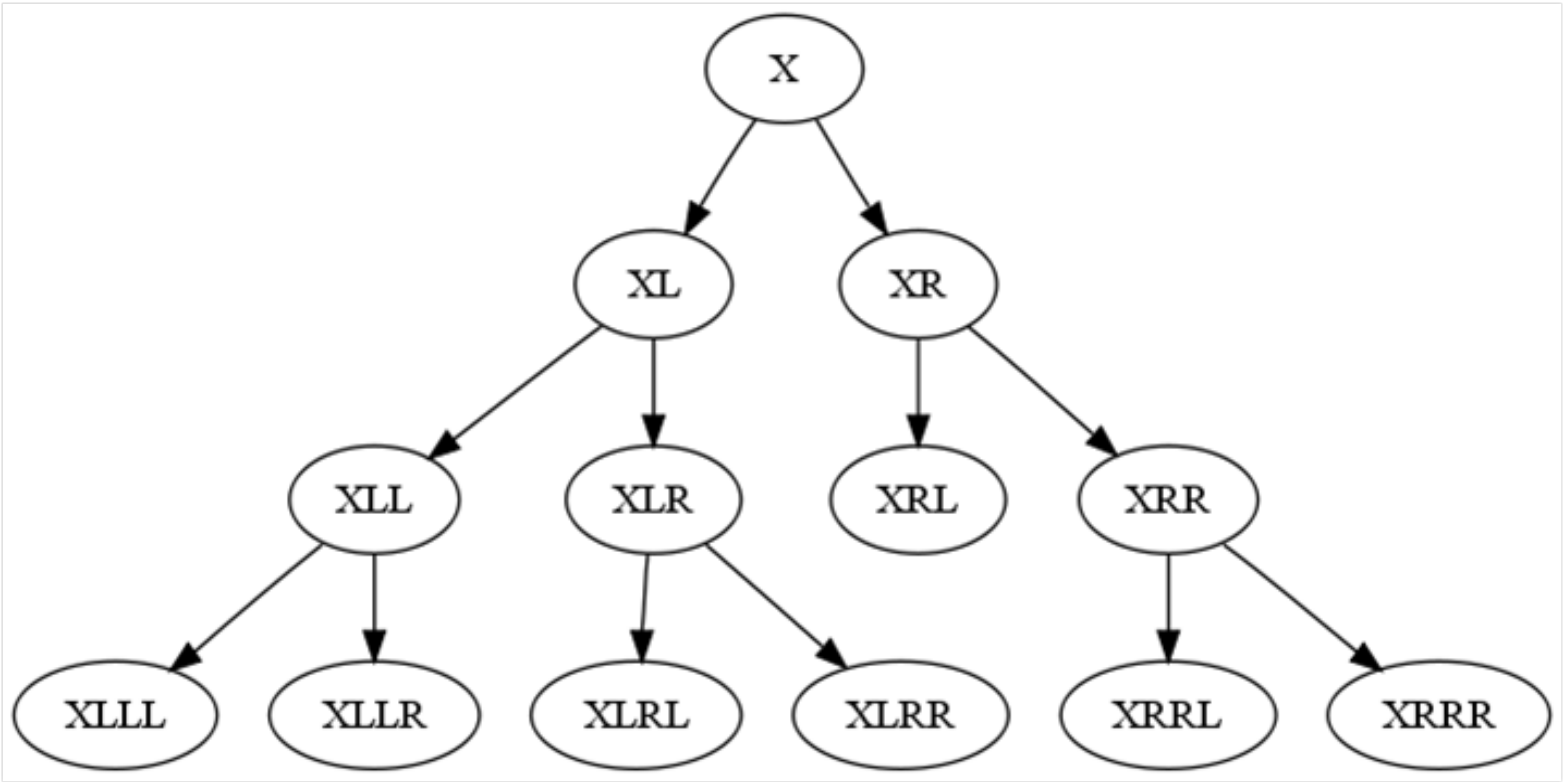
Monte Carlo Markov Chain over Histograms in $\mathcal{RSP}_{0:\infty}$



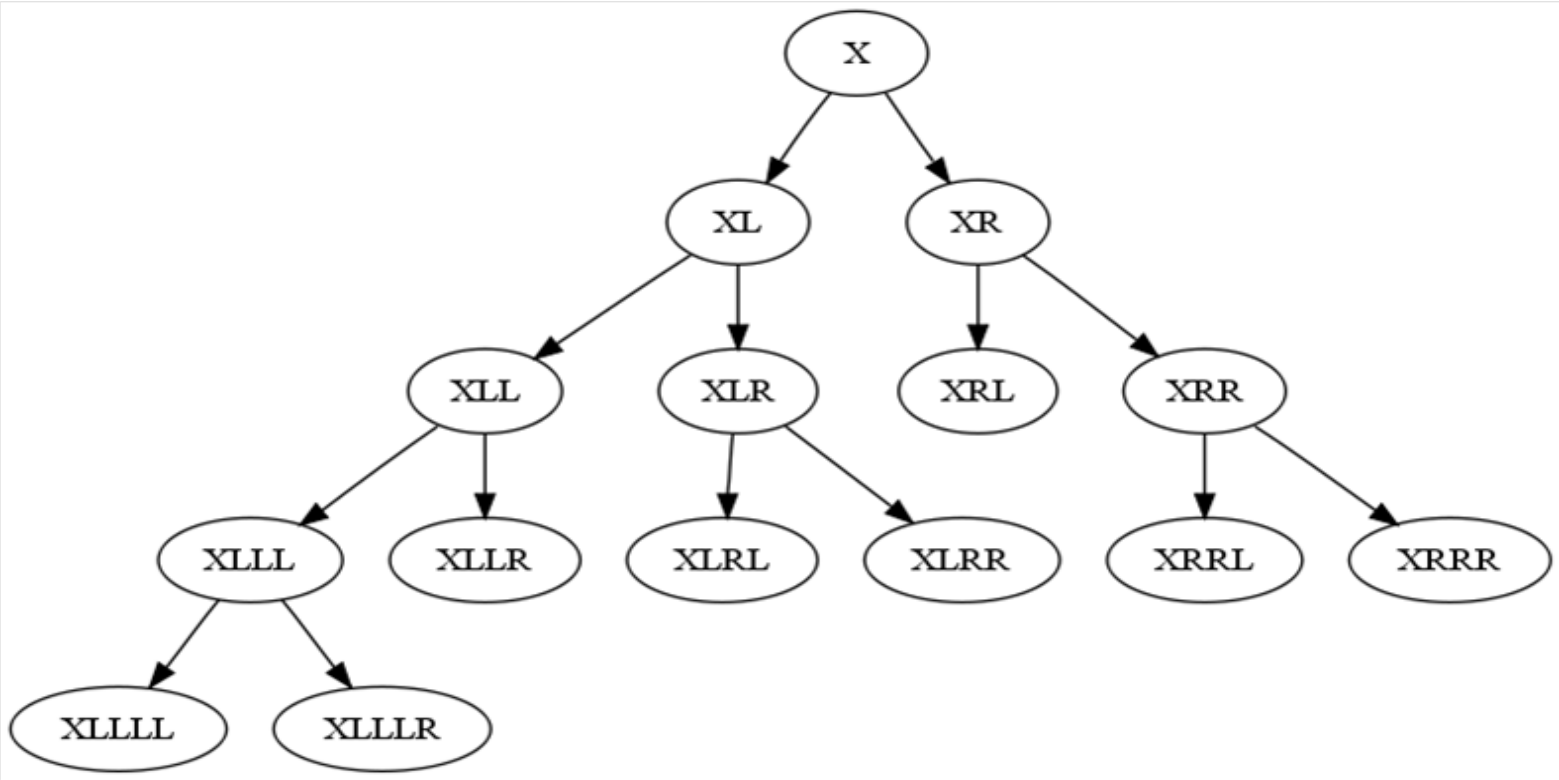
Monte Carlo Markov Chain over Histograms in $\mathcal{RSP}_{0:\infty}$



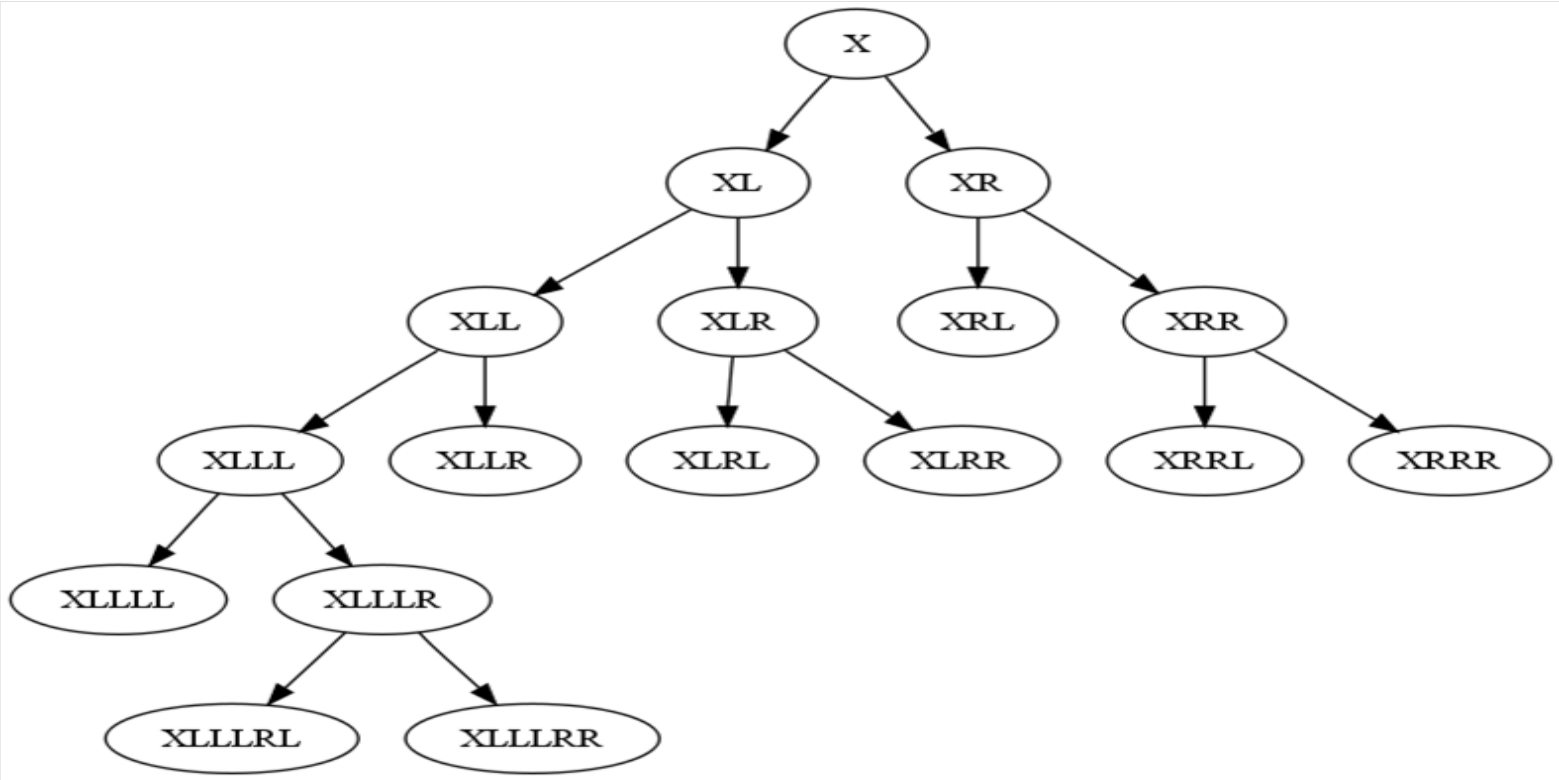
Monte Carlo Markov Chain over Histograms in $\mathcal{RSP}_{0:\infty}$



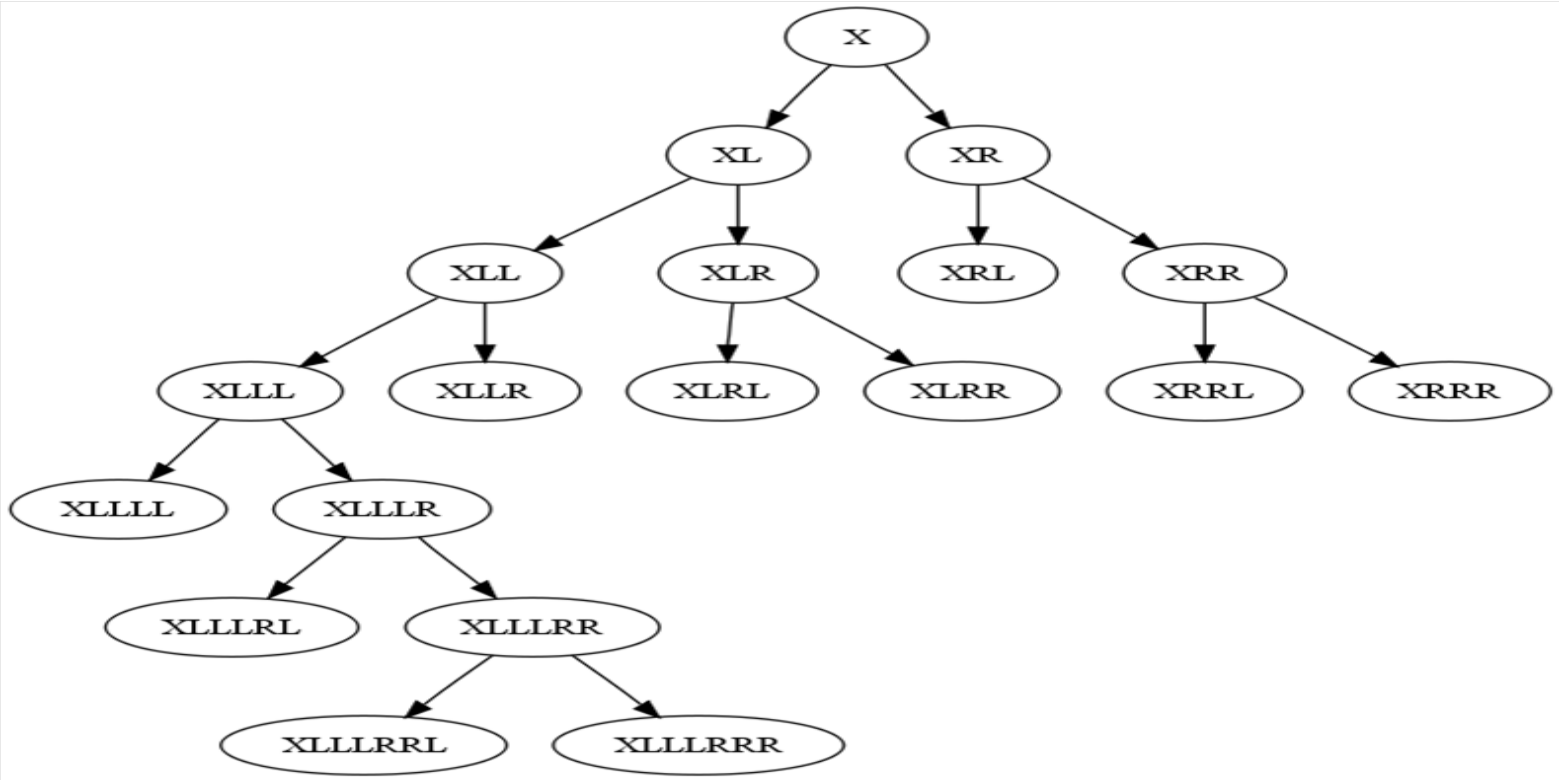
Monte Carlo Markov Chain over Histograms in $\mathcal{RSP}_{0:\infty}$



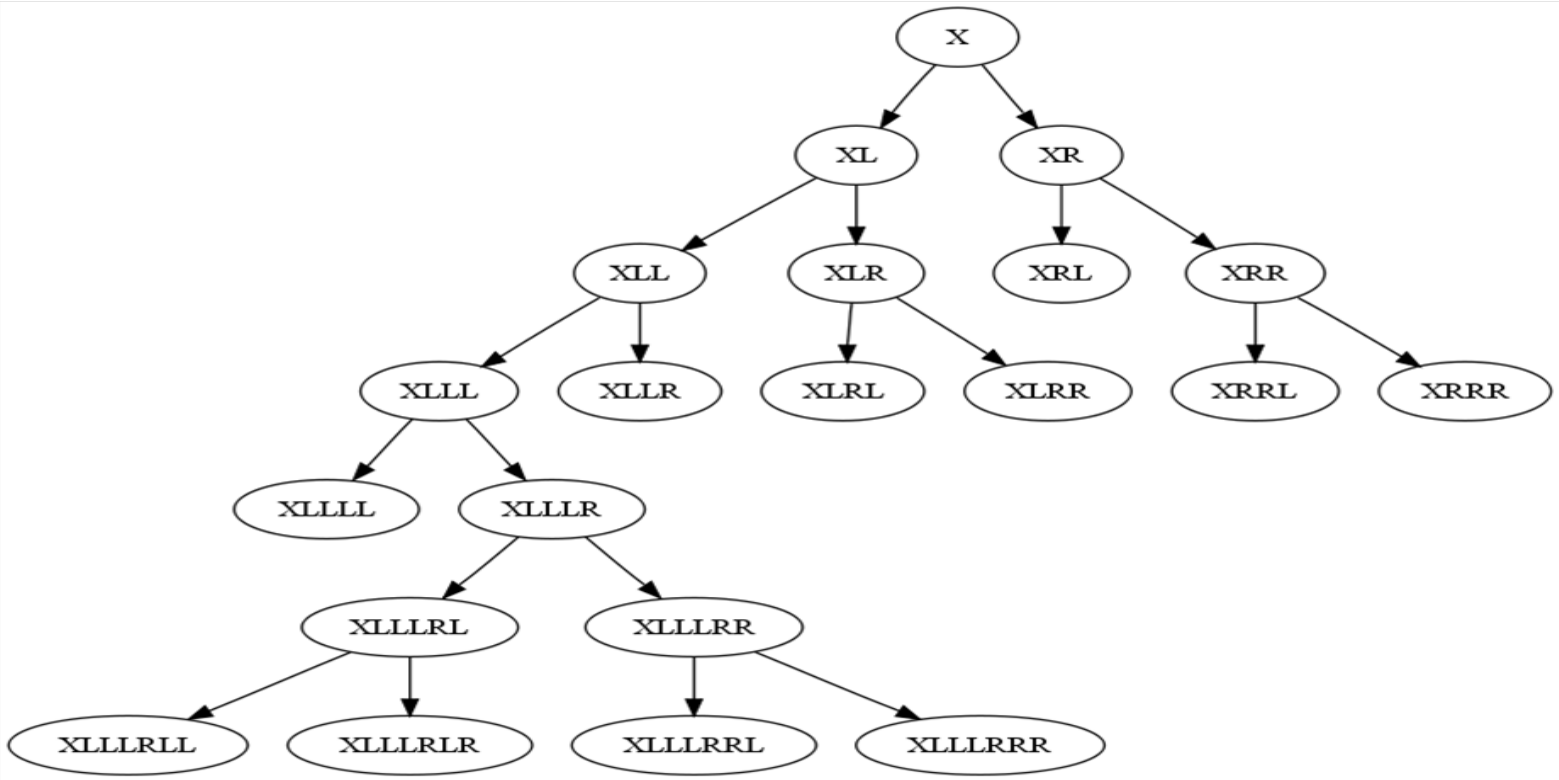
Monte Carlo Markov Chain over Histograms in $\mathcal{RSP}_{0:\infty}$



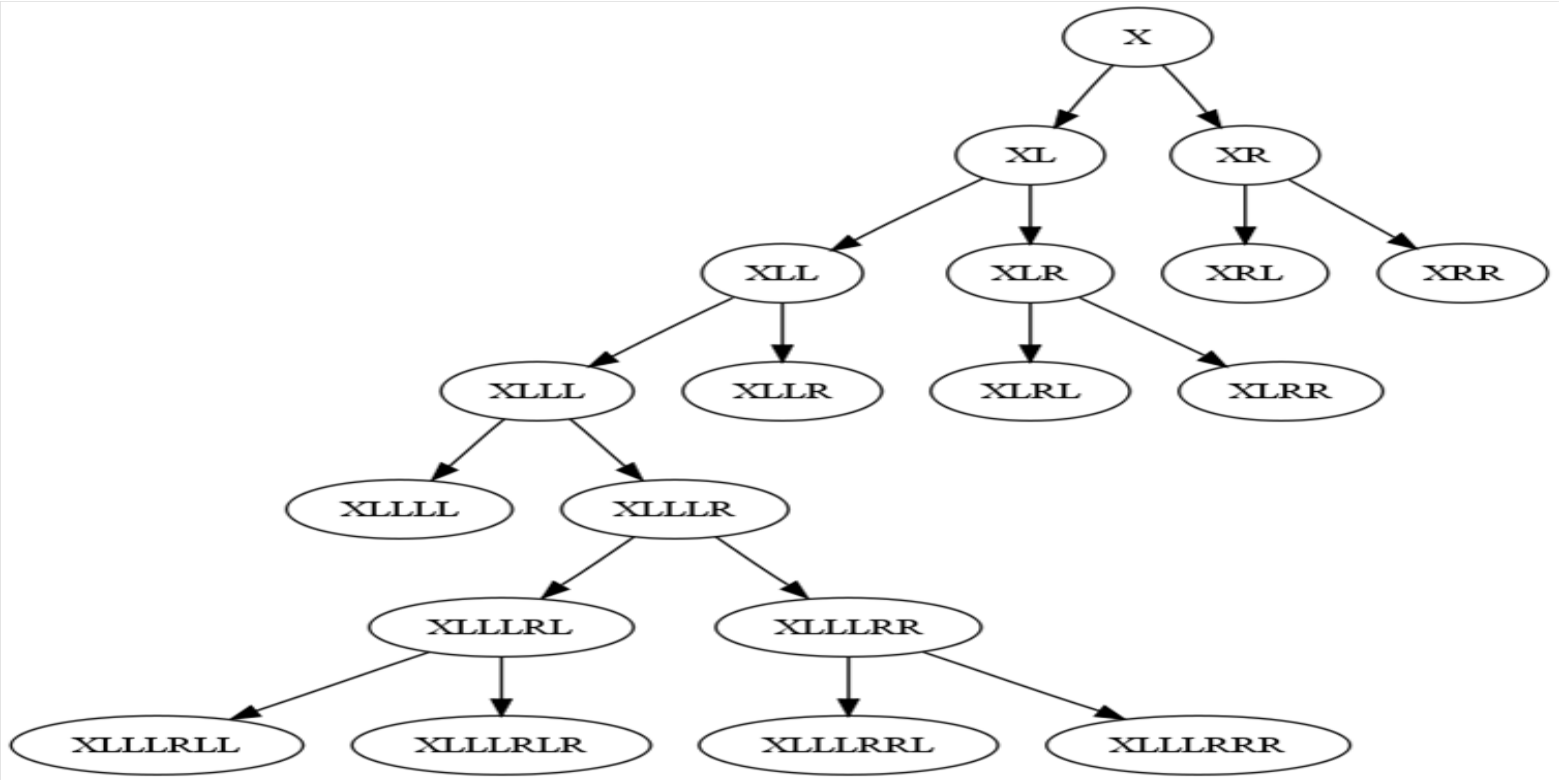
Monte Carlo Markov Chain over Histograms in $\mathcal{RSP}_{0:\infty}$



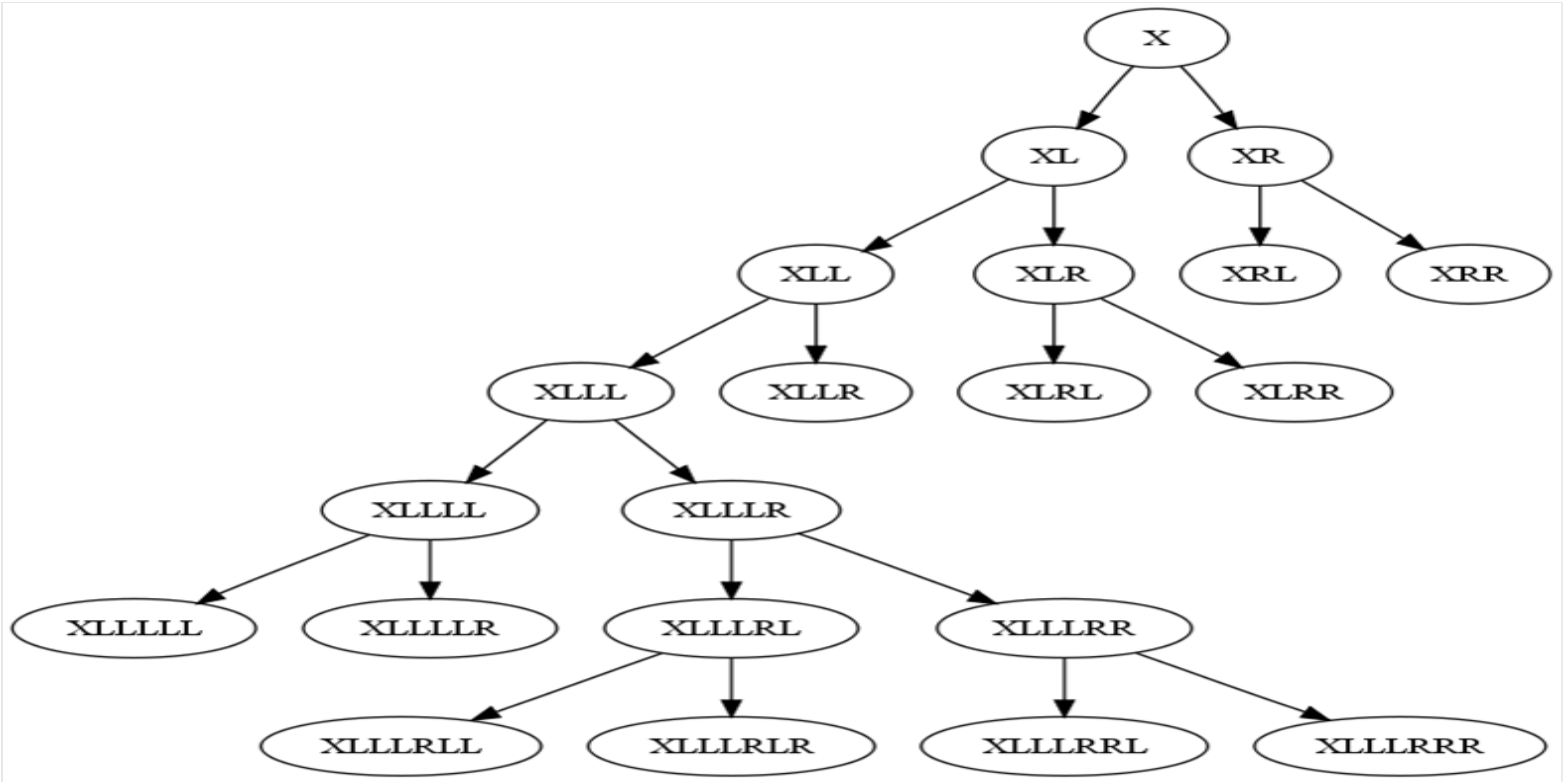
Monte Carlo Markov Chain over Histograms in $\mathcal{RSP}_{0:\infty}$



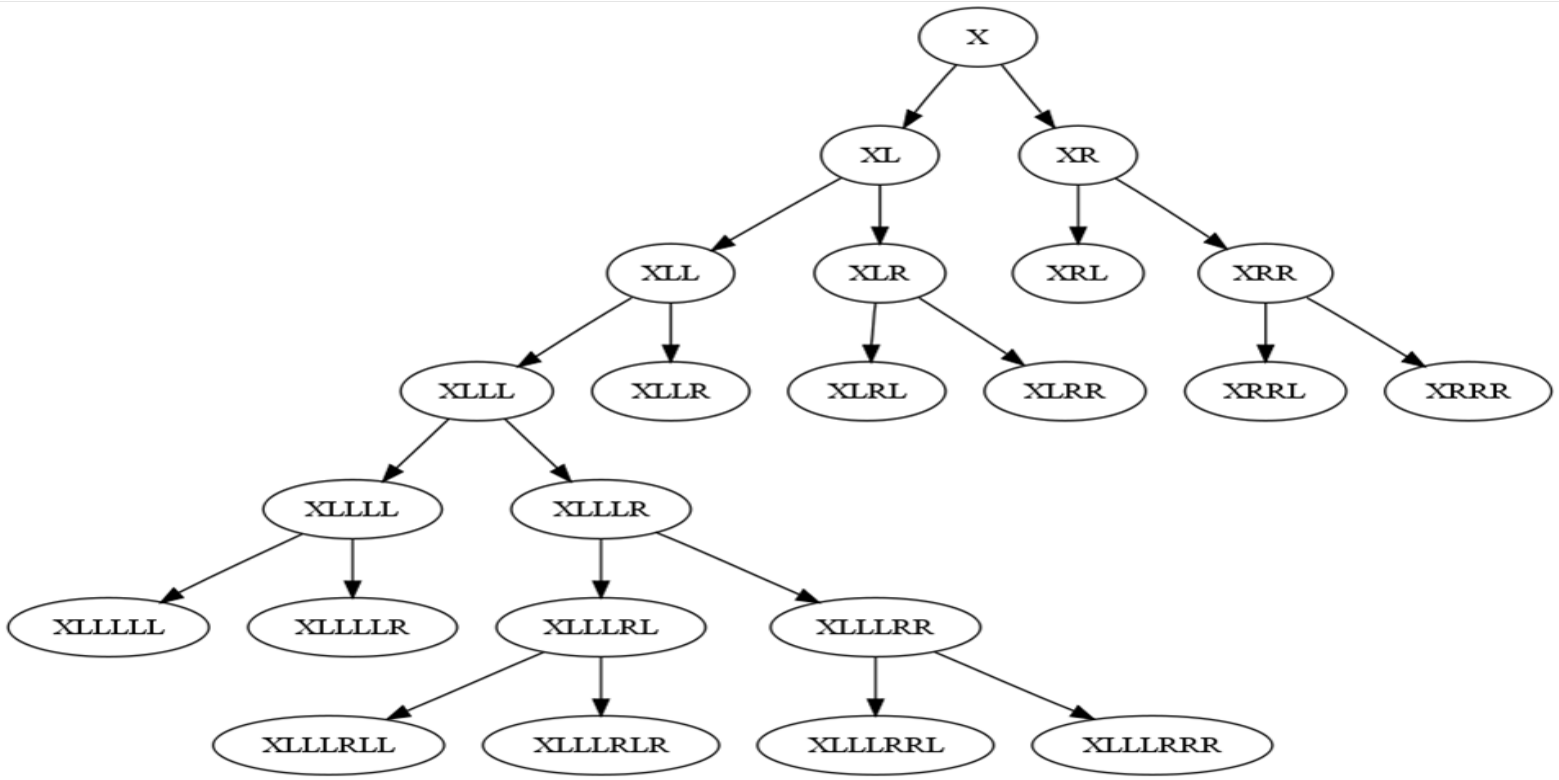
Monte Carlo Markov Chain over Histograms in $\mathcal{RSP}_{0:\infty}$



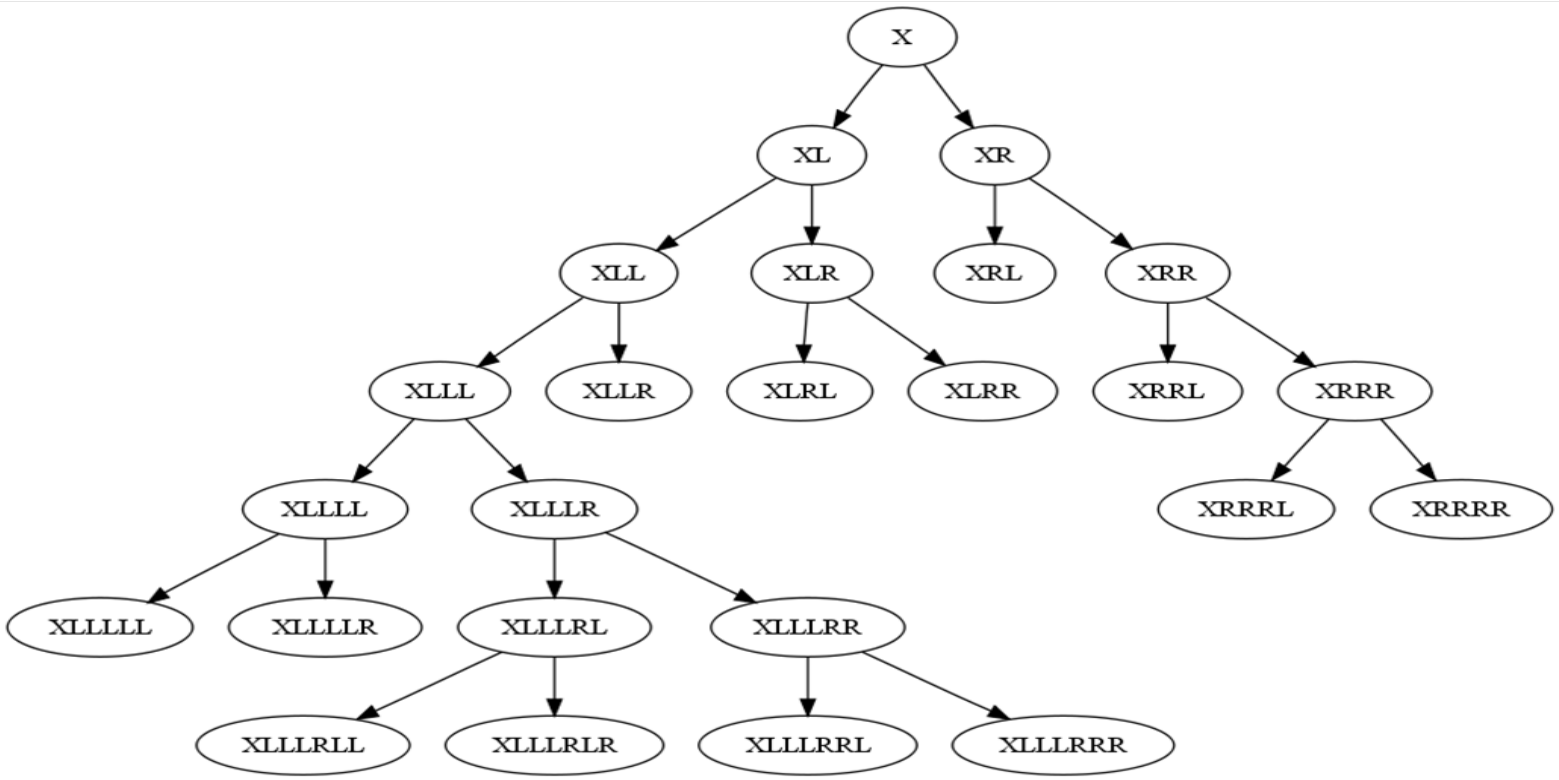
Monte Carlo Markov Chain over Histograms in $\mathcal{RSP}_{0:\infty}$



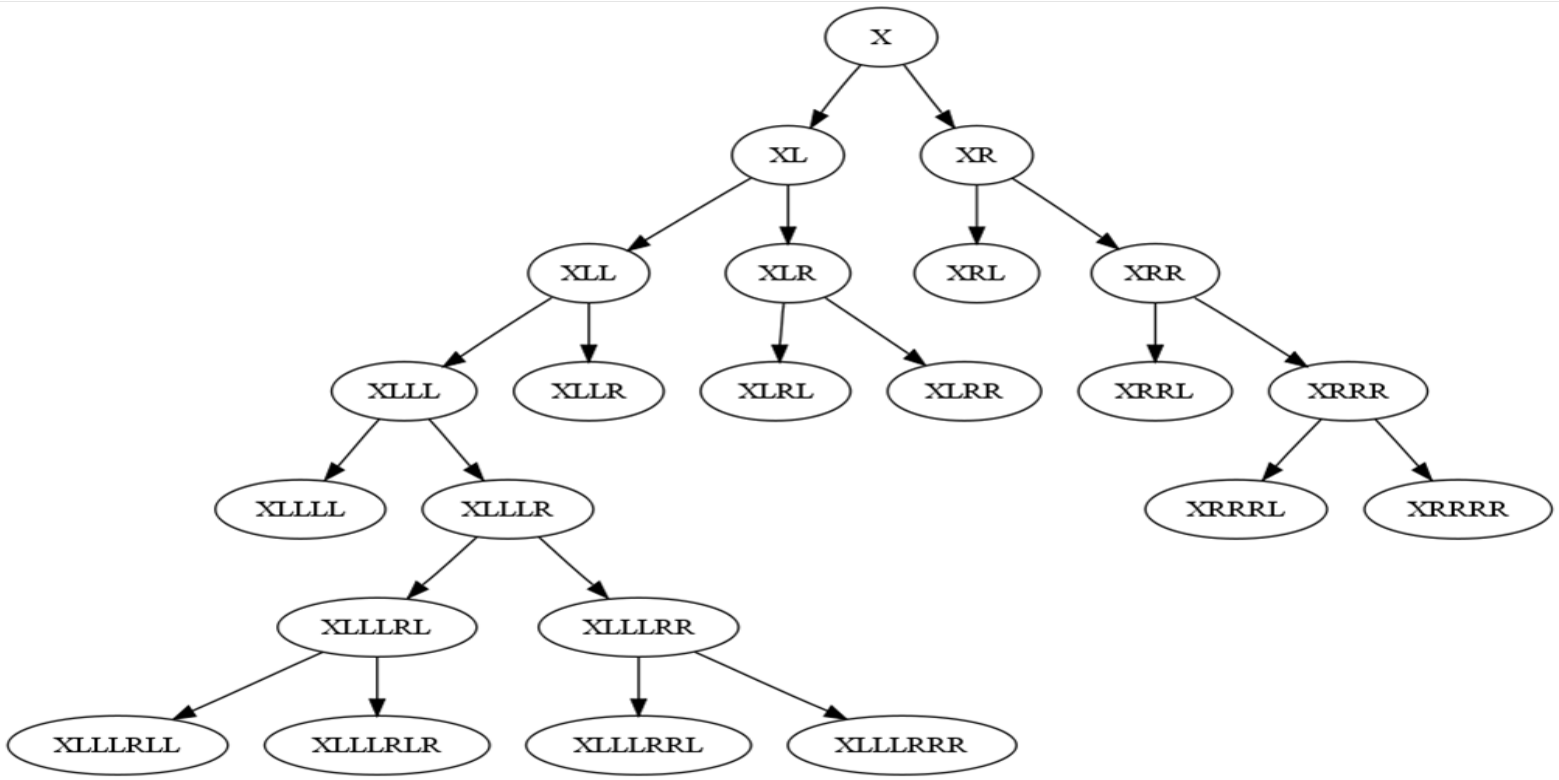
Monte Carlo Markov Chain over Histograms in $\mathcal{RSP}_{0:\infty}$



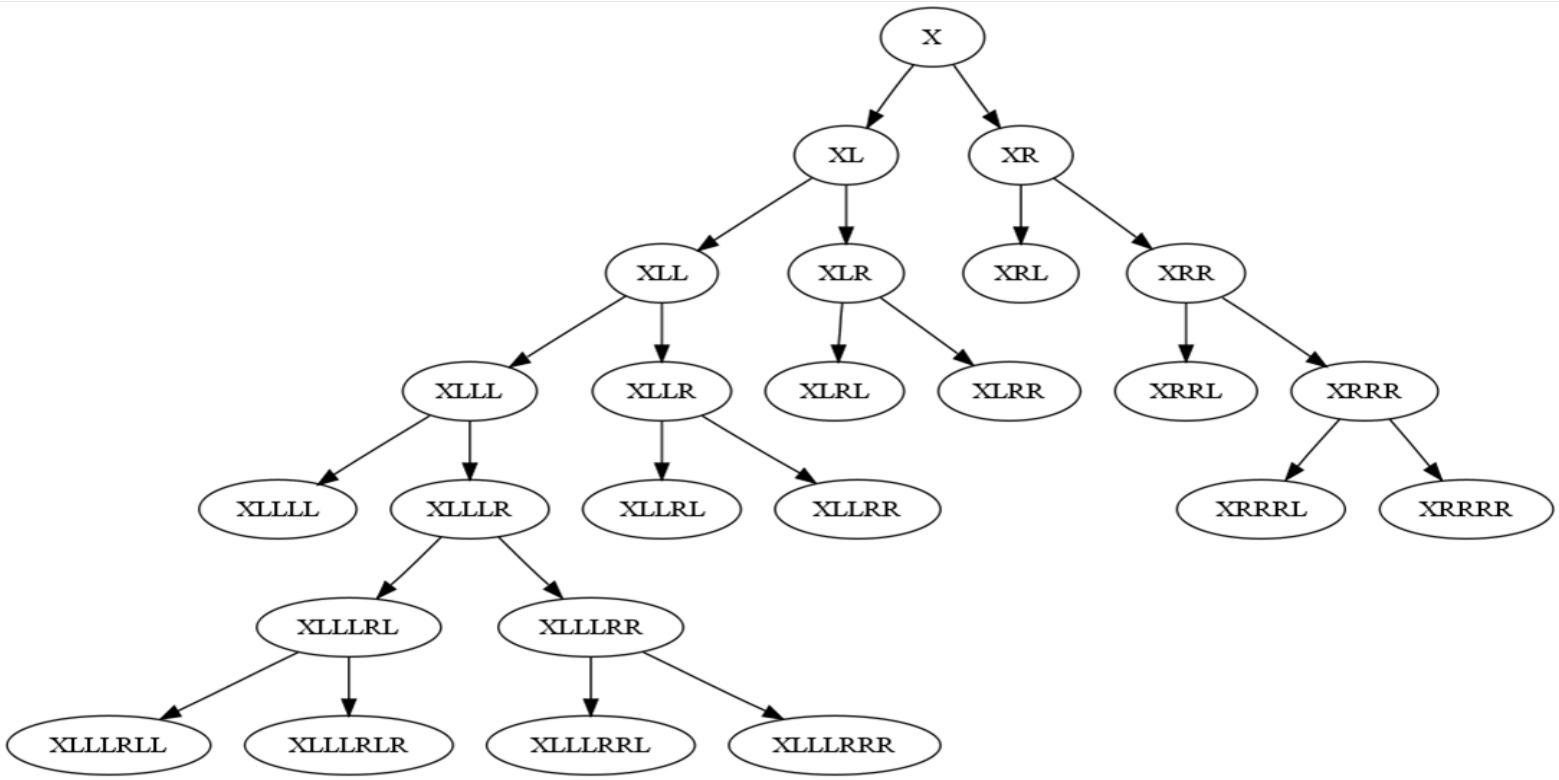
Monte Carlo Markov Chain over Histograms in $\mathcal{RSP}_{0:\infty}$



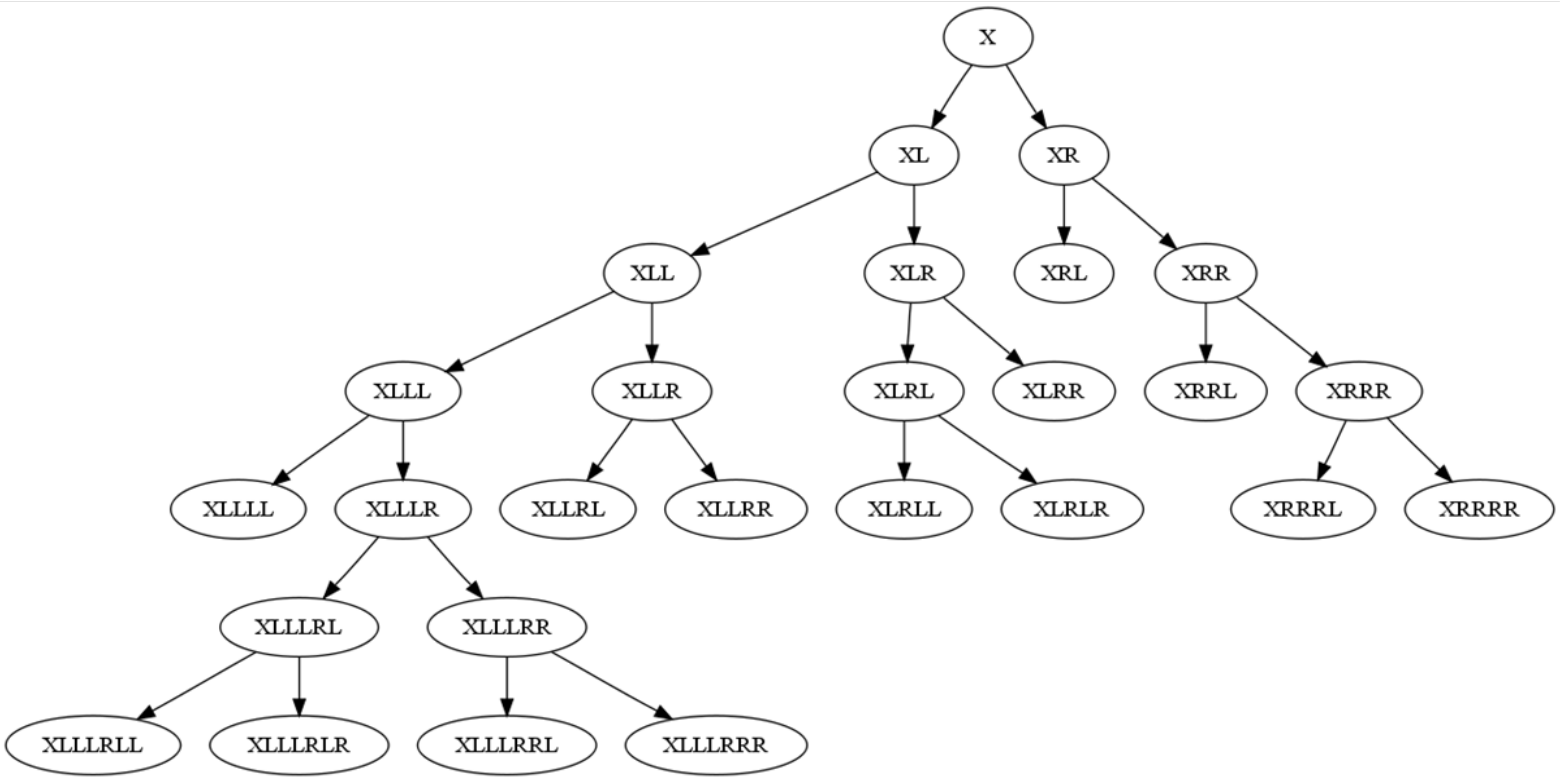
Monte Carlo Markov Chain over Histograms in $\mathcal{RSP}_{0:\infty}$



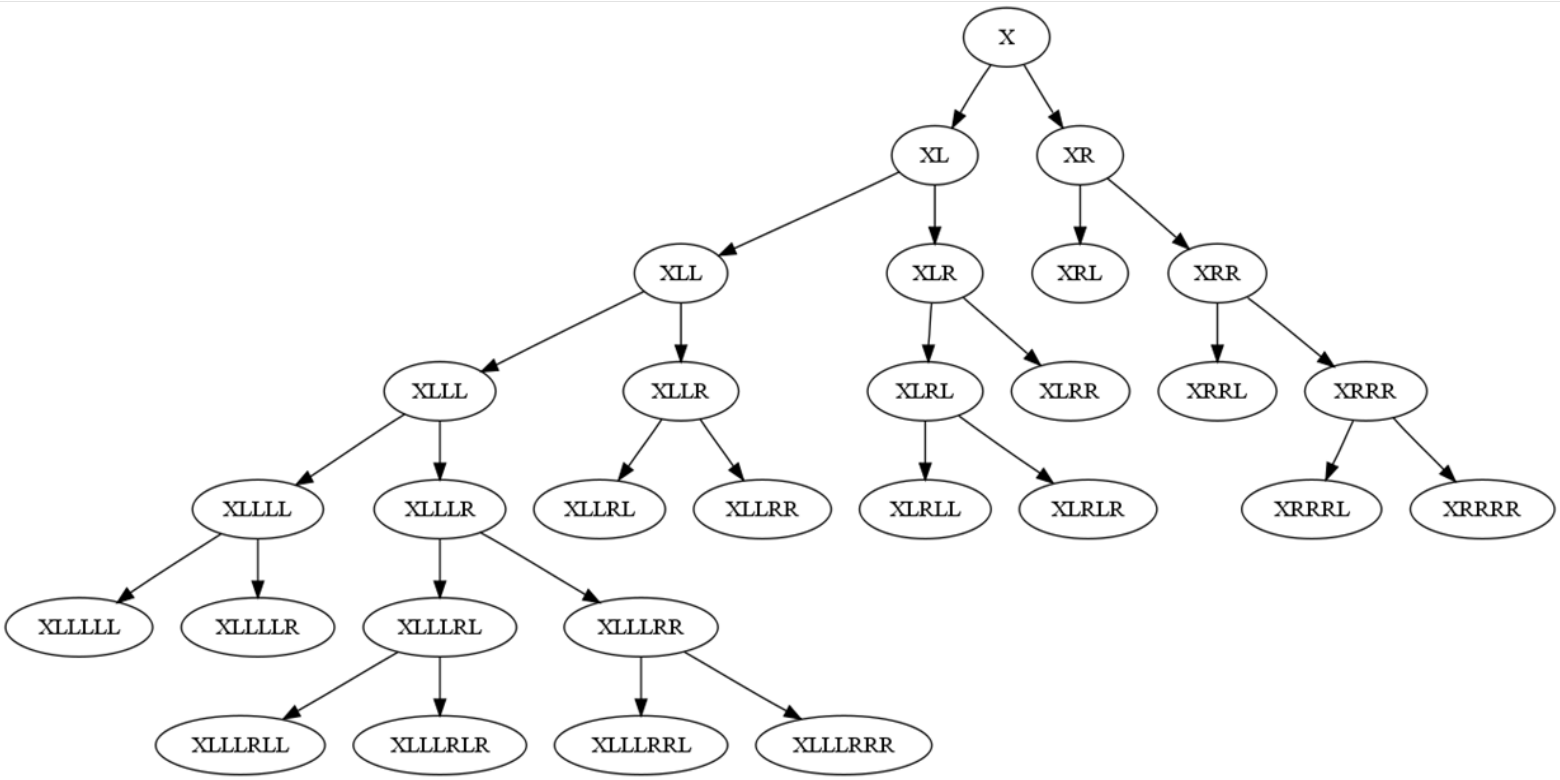
Monte Carlo Markov Chain over Histograms in $\mathcal{RSP}_{0:\infty}$



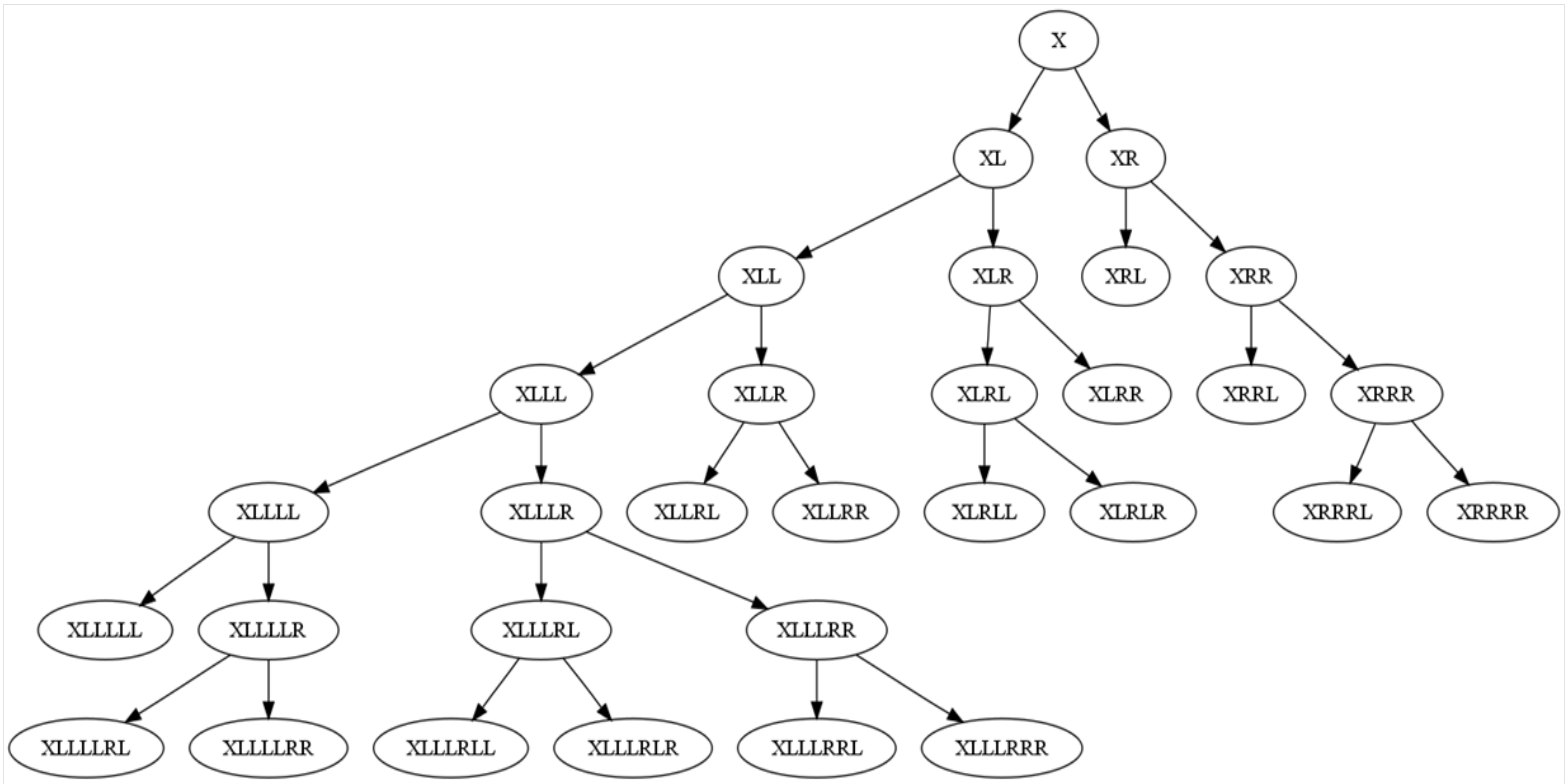
Monte Carlo Markov Chain over Histograms in $\mathcal{RSP}_{0:\infty}$



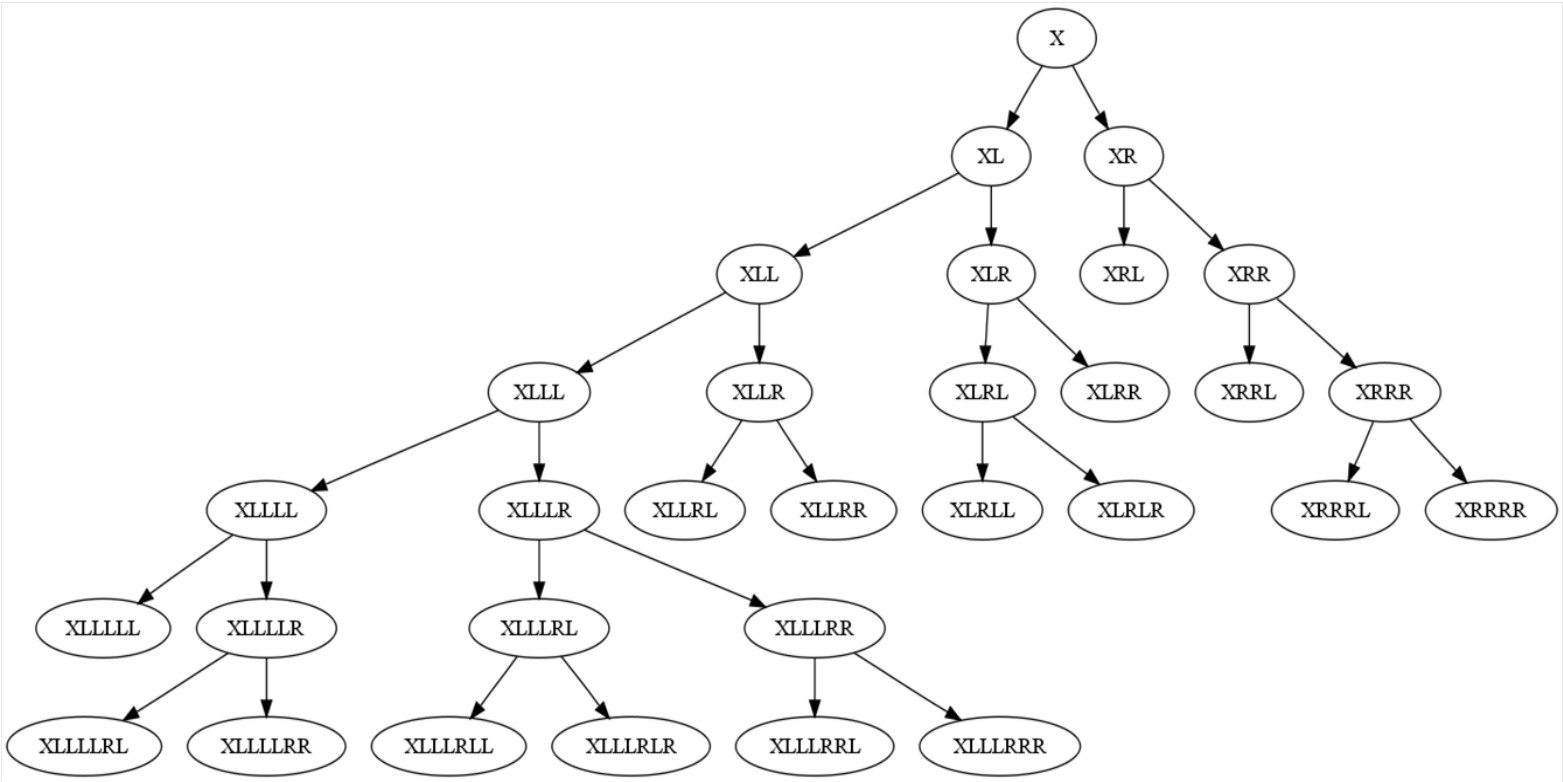
Monte Carlo Markov Chain over Histograms in $\mathcal{RSP}_{0:\infty}$



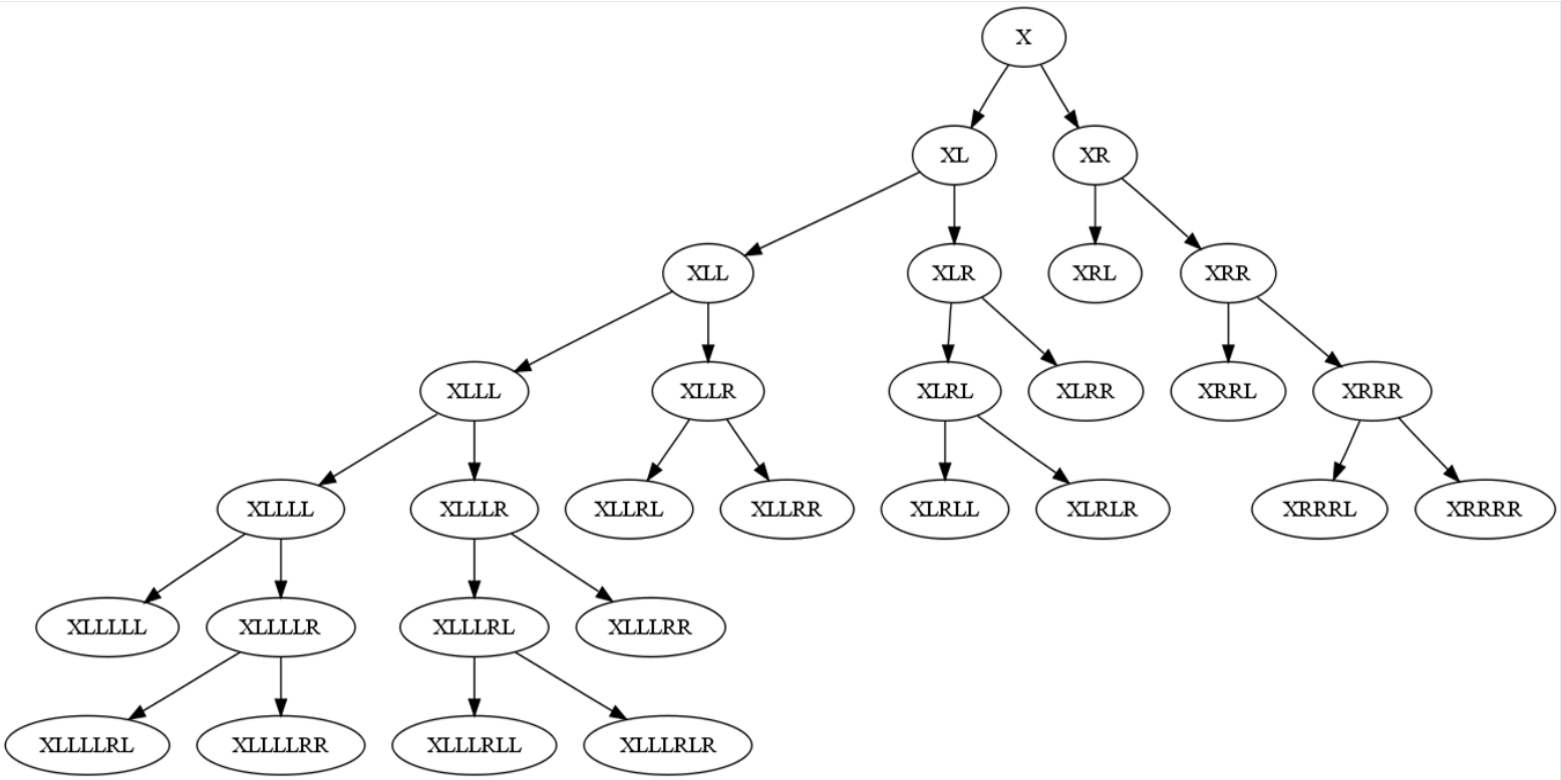
Monte Carlo Markov Chain over Histograms in $\mathcal{RSP}_{0:\infty}$



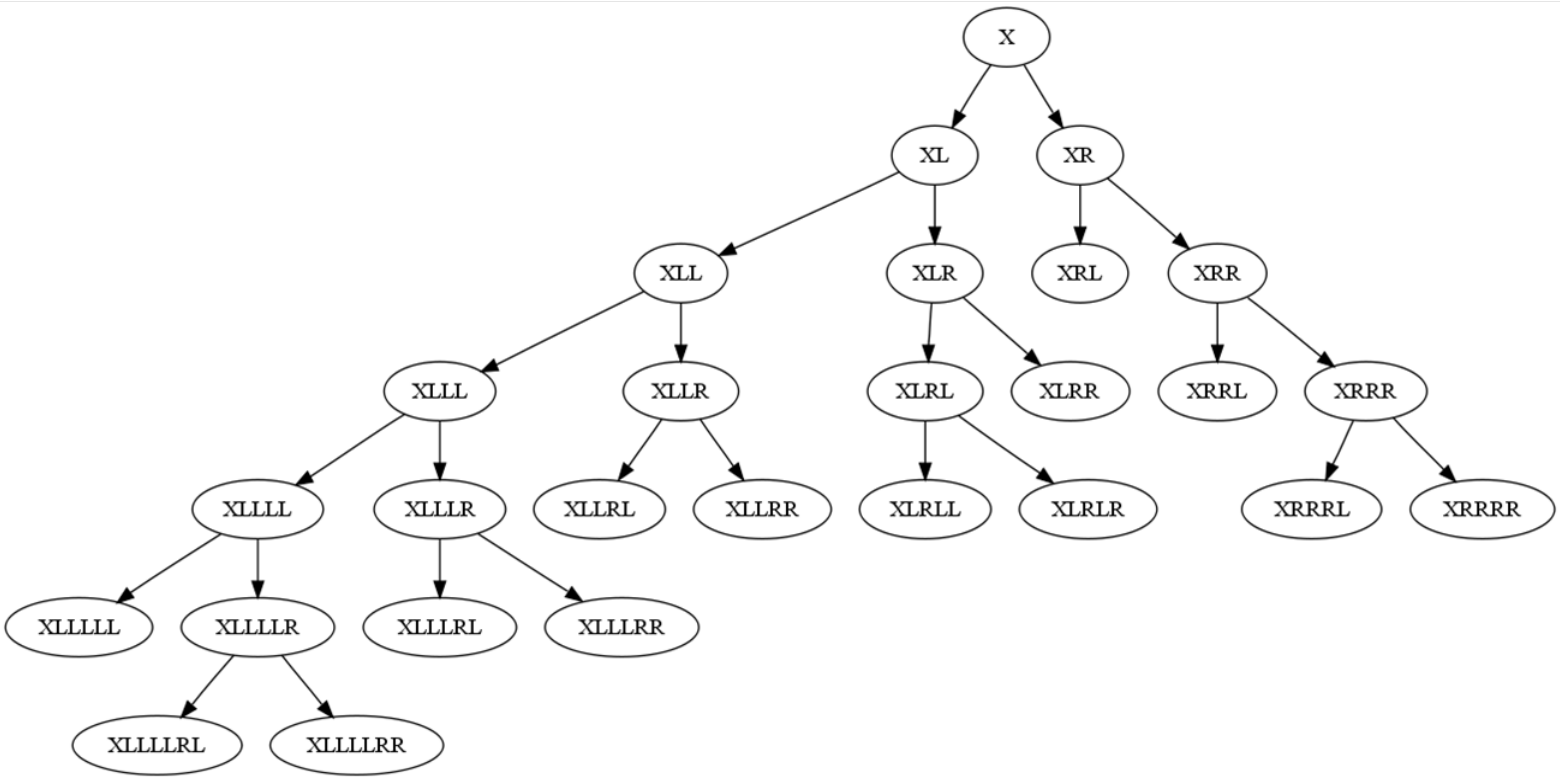
Monte Carlo Markov Chain over Histograms in $\mathcal{RSP}_{0:\infty}$



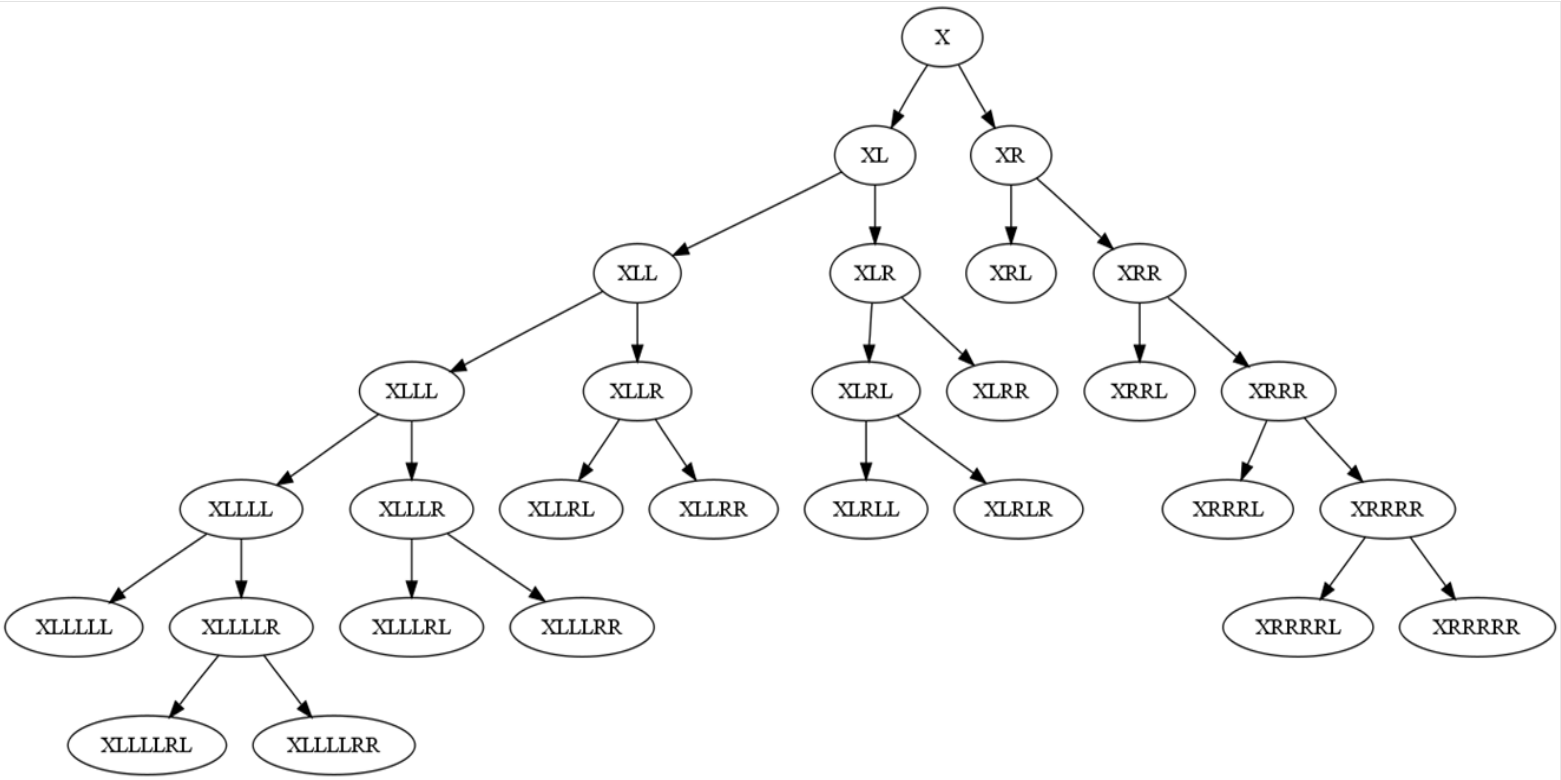
Monte Carlo Markov Chain over Histograms in $\mathcal{RSP}_{0:\infty}$



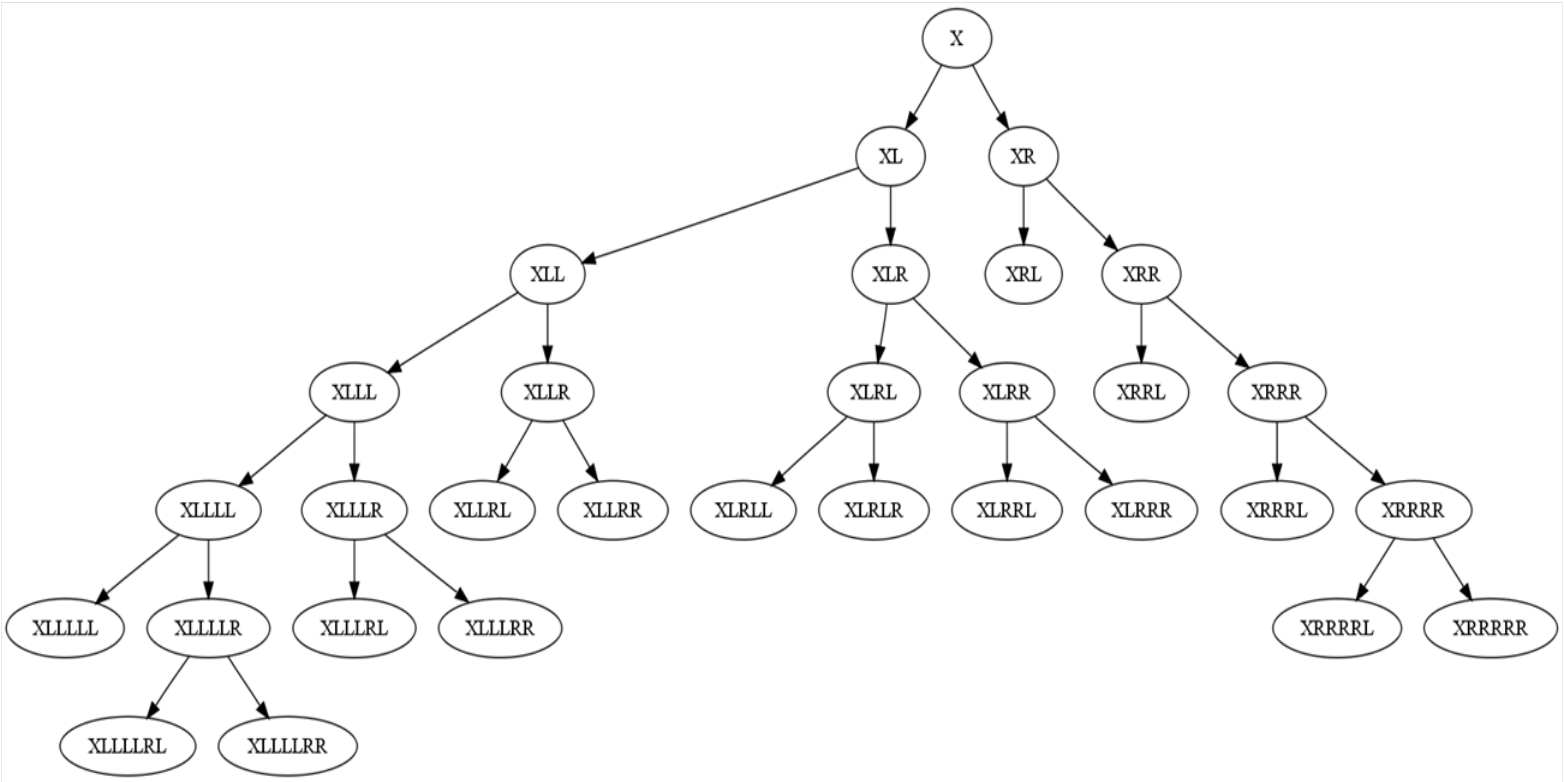
Monte Carlo Markov Chain over Histograms in $\mathcal{RSP}_{0:\infty}$



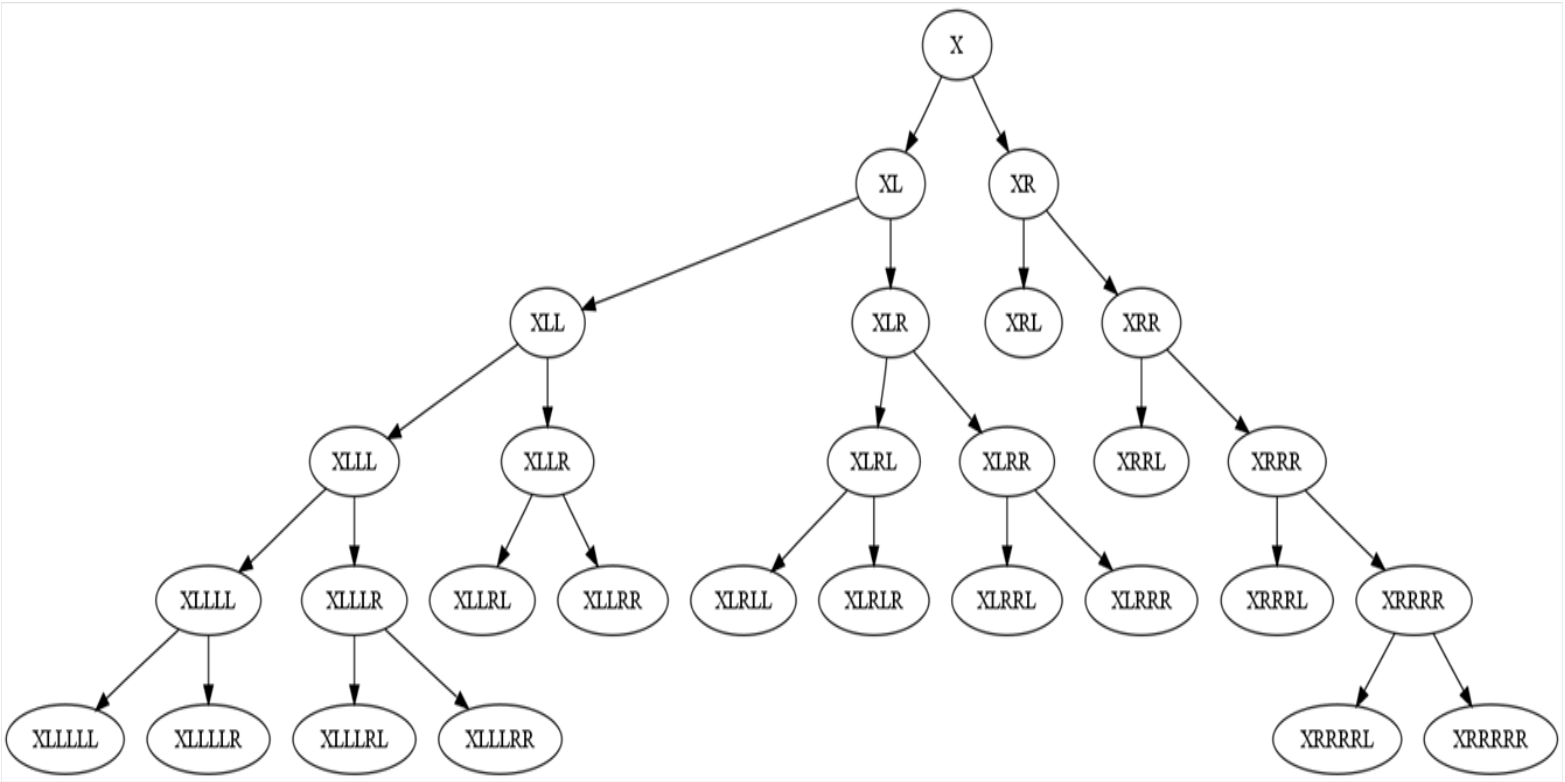
Monte Carlo Markov Chain over Histograms in $\mathcal{RSP}_{0:\infty}$



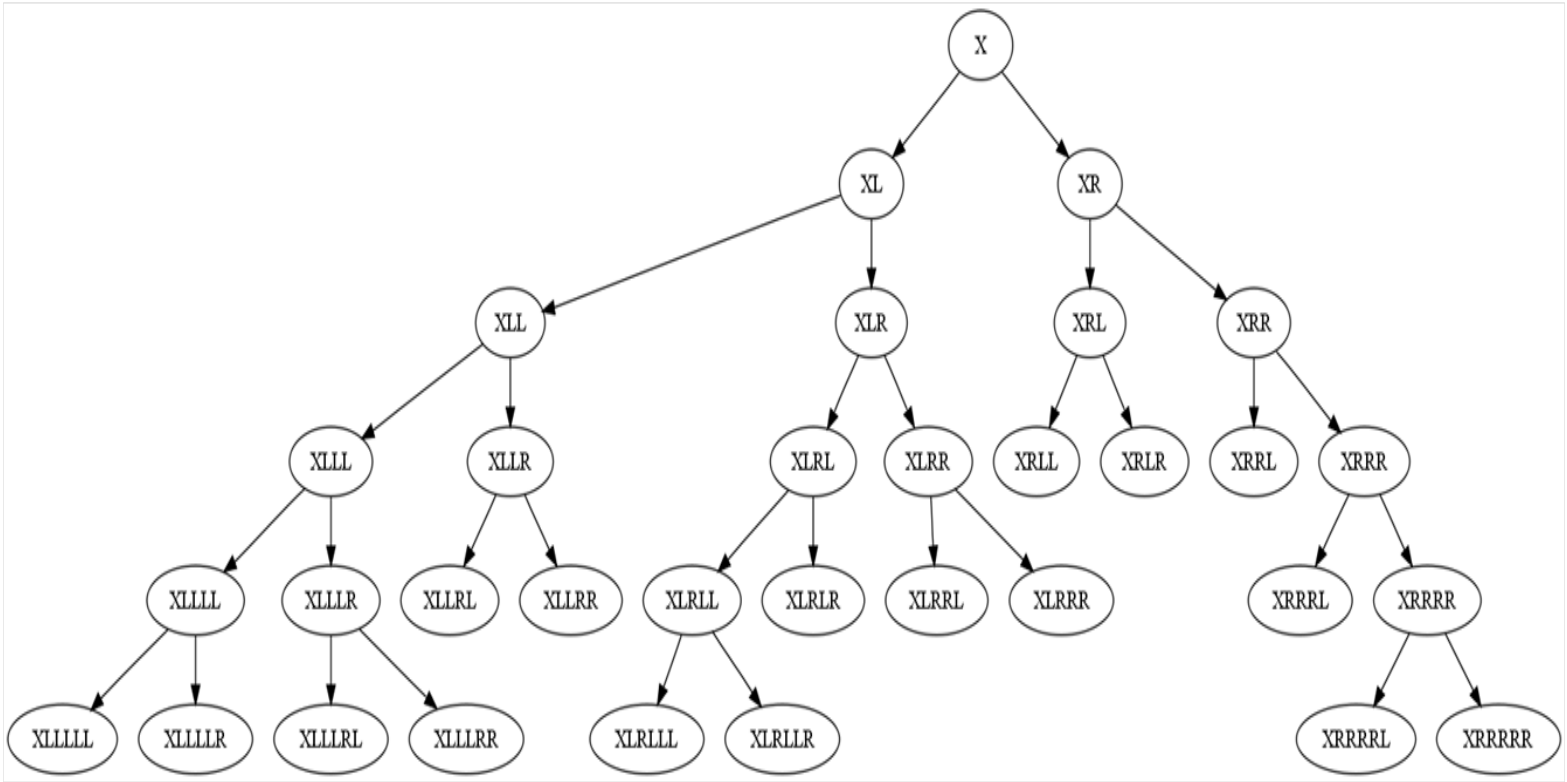
Monte Carlo Markov Chain over Histograms in $\mathcal{RSP}_{0:\infty}$



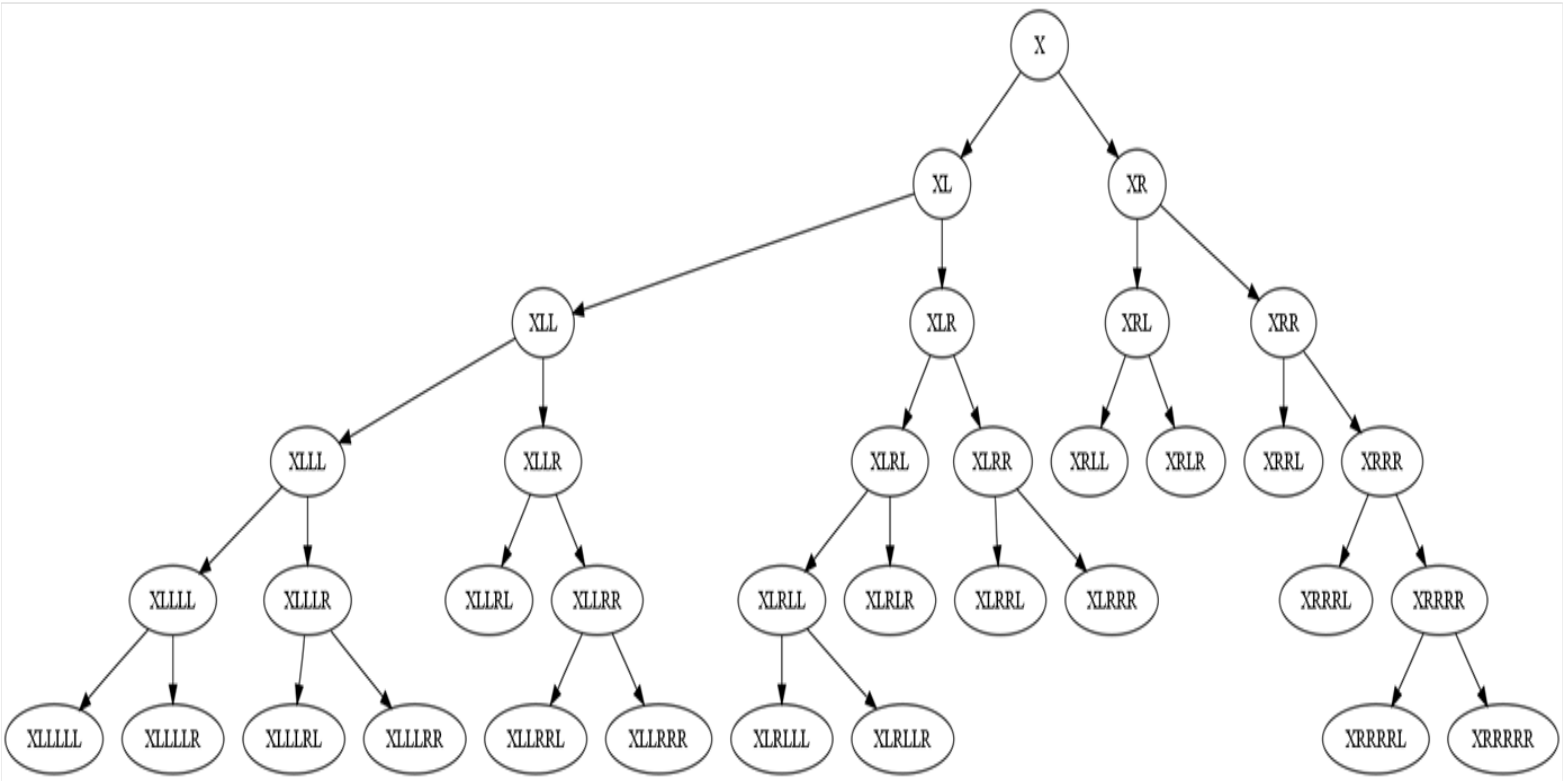
Monte Carlo Markov Chain over Histograms in $\mathcal{RSP}_{0:\infty}$



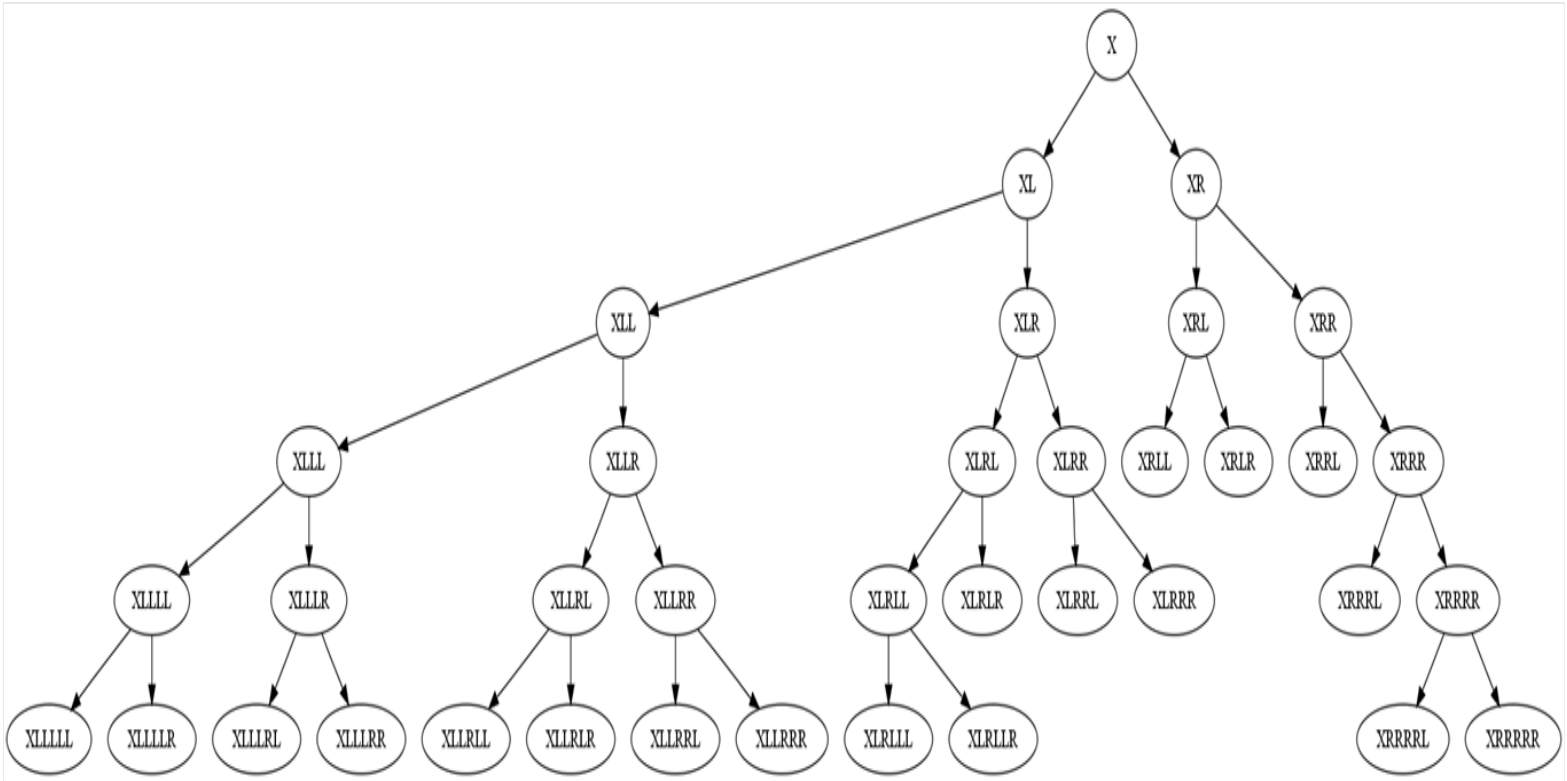
Monte Carlo Markov Chain over Histograms in $\mathcal{RSP}_{0:\infty}$



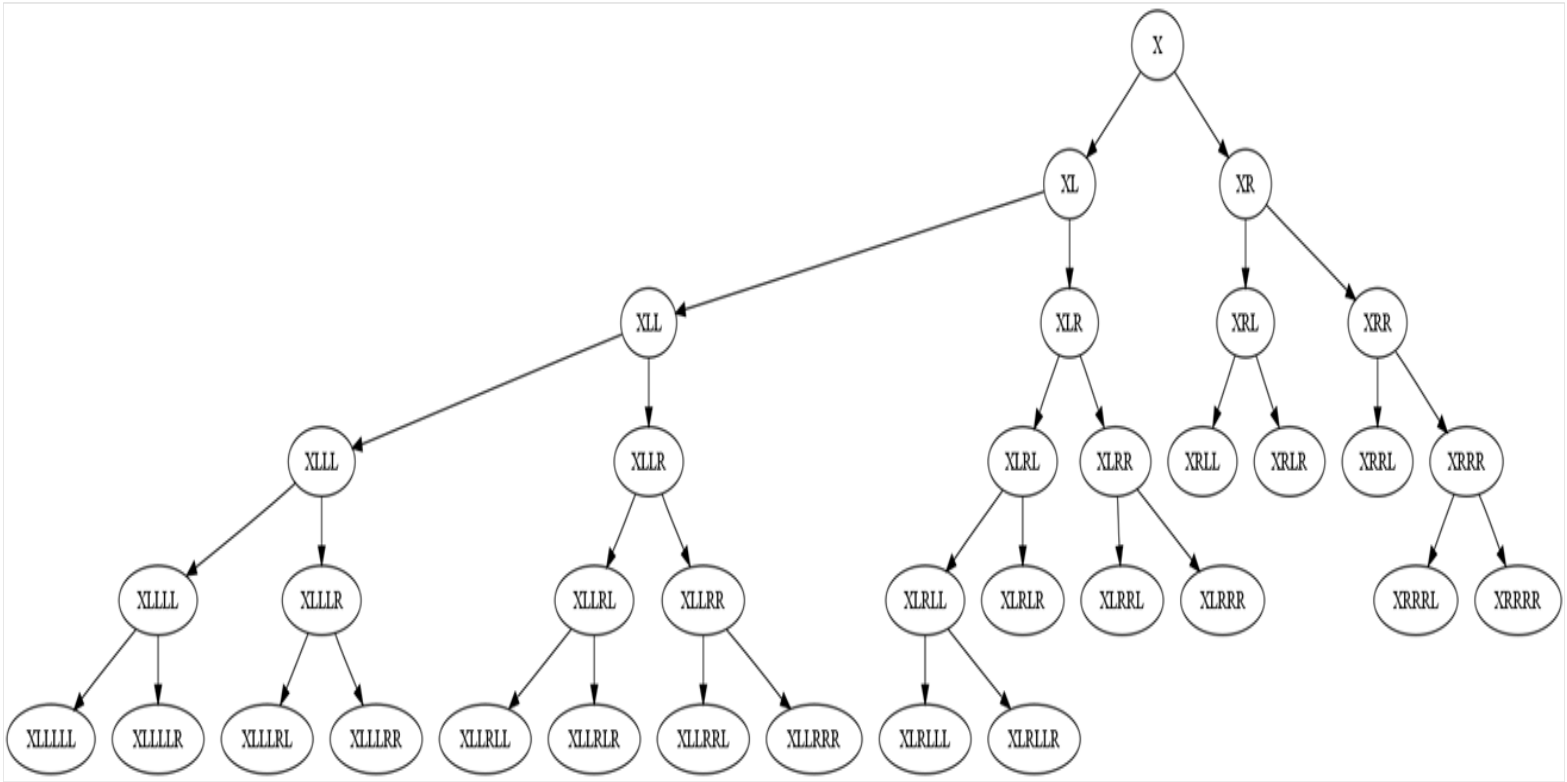
Monte Carlo Markov Chain over Histograms in $\mathcal{RSP}_{0:\infty}$



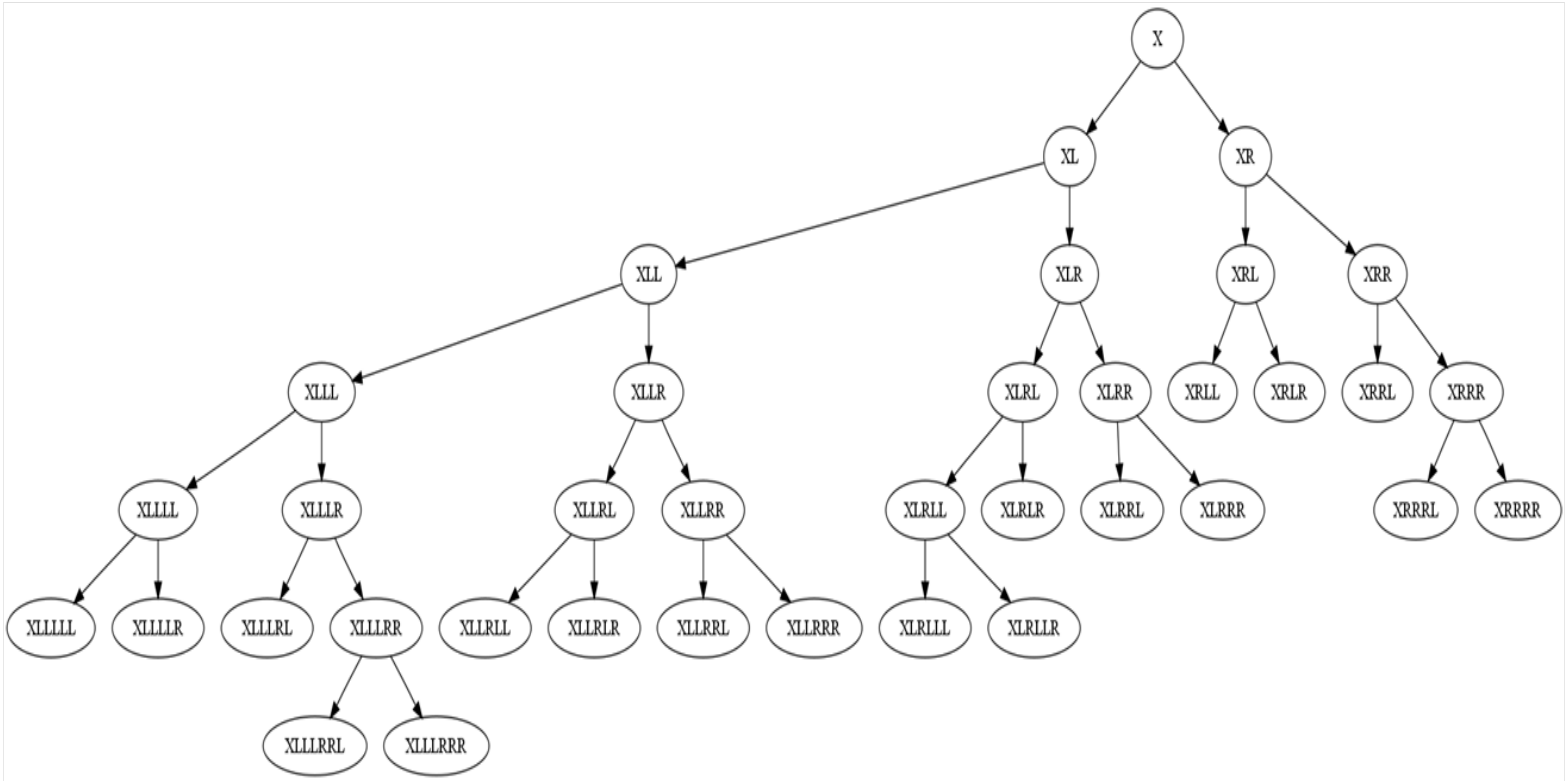
Monte Carlo Markov Chain over Histograms in $\mathcal{RSP}_{0:\infty}$



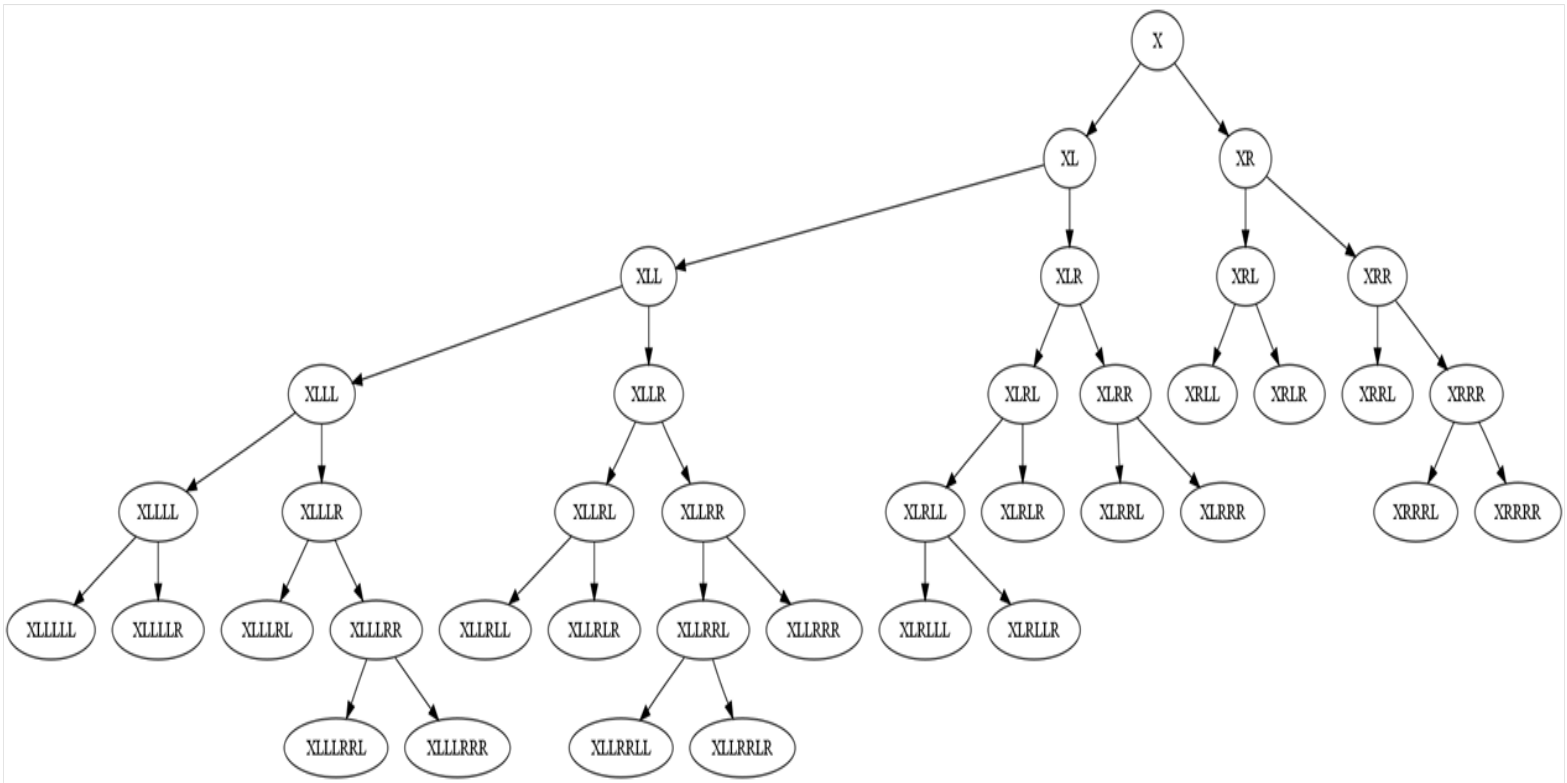
Monte Carlo Markov Chain over Histograms in $\mathcal{RSP}_{0:\infty}$



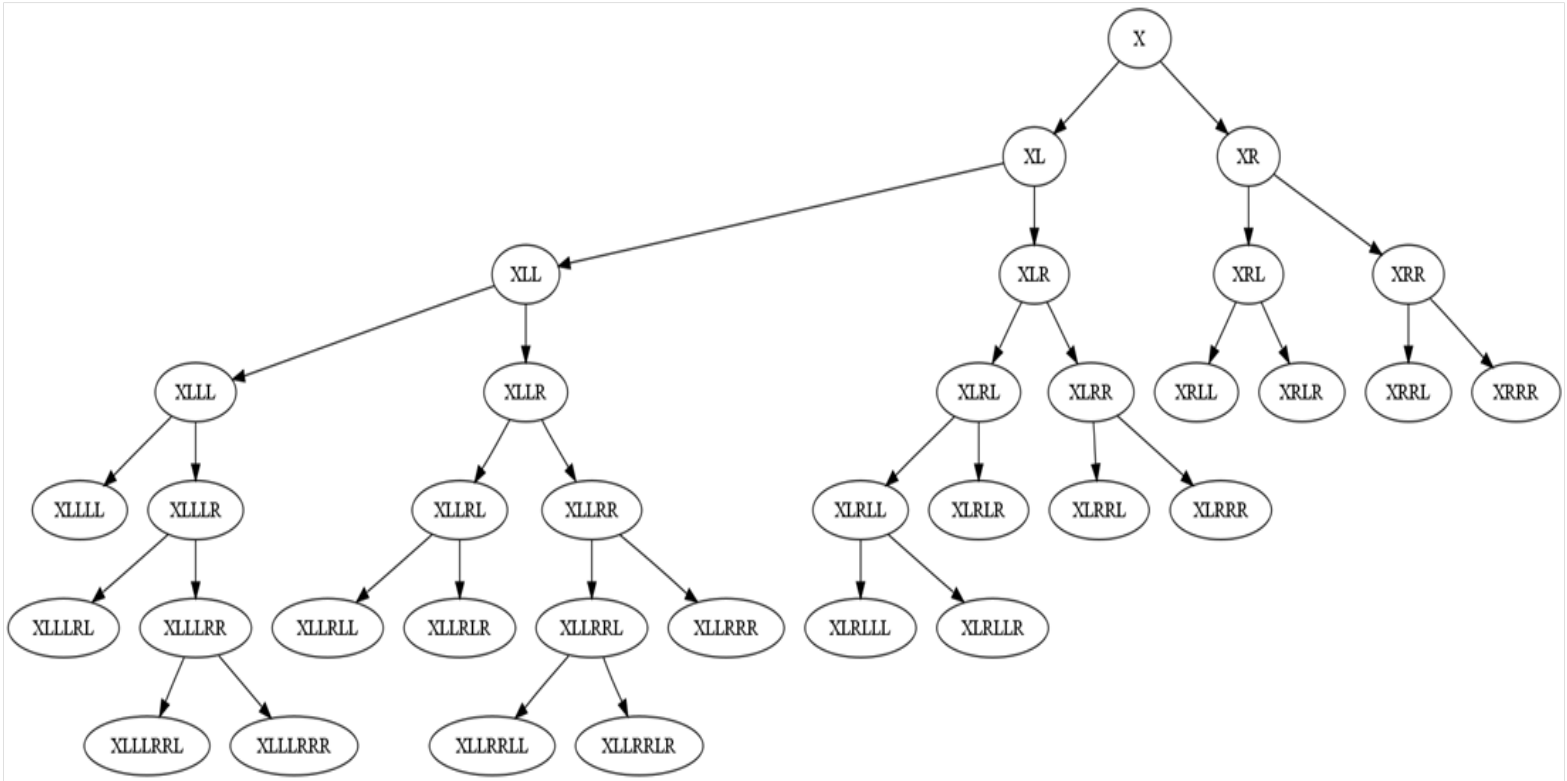
Monte Carlo Markov Chain over Histograms in $\mathcal{RSP}_{0:\infty}$



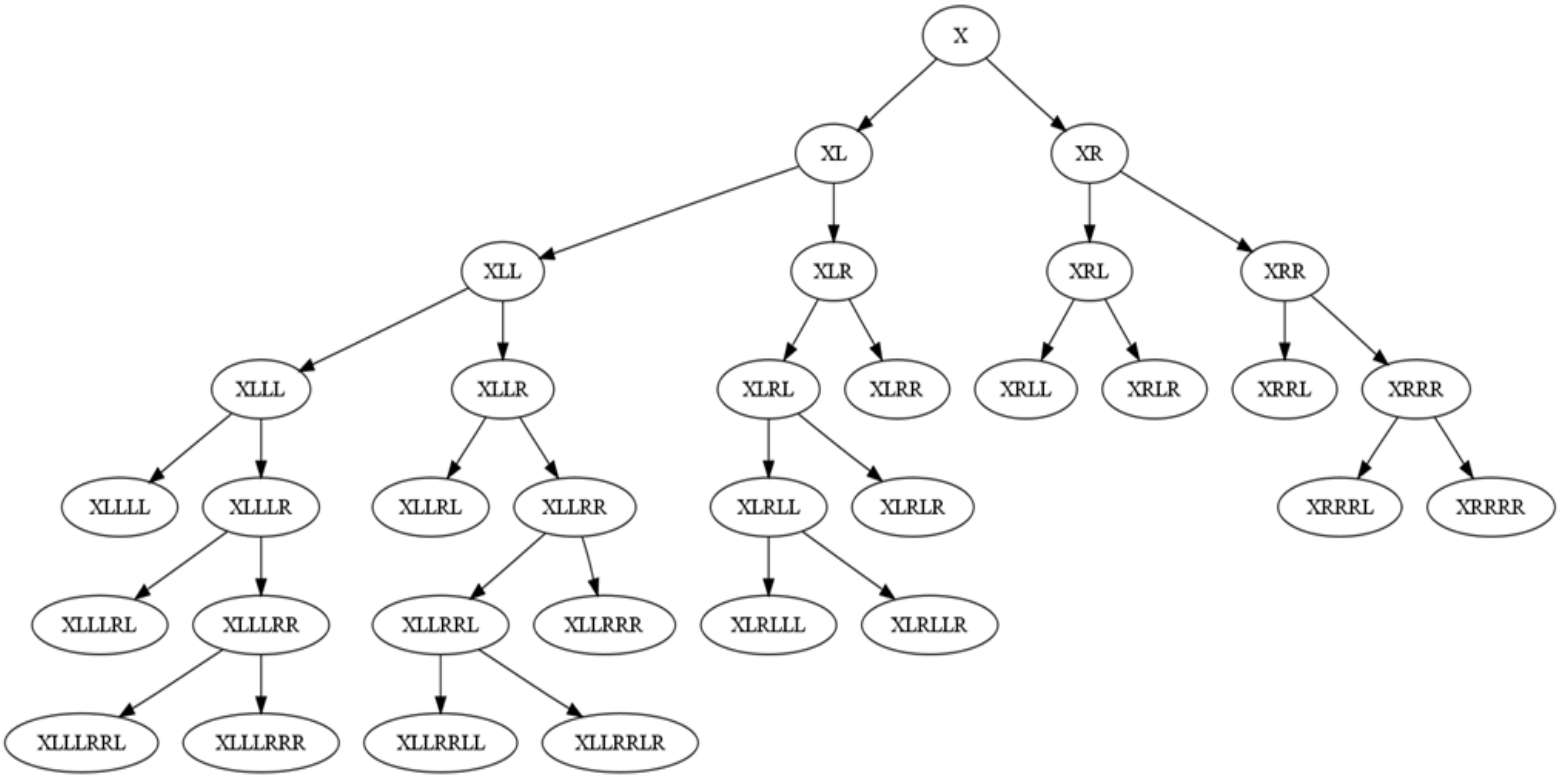
Monte Carlo Markov Chain over Histograms in $\mathcal{RSP}_{0:\infty}$



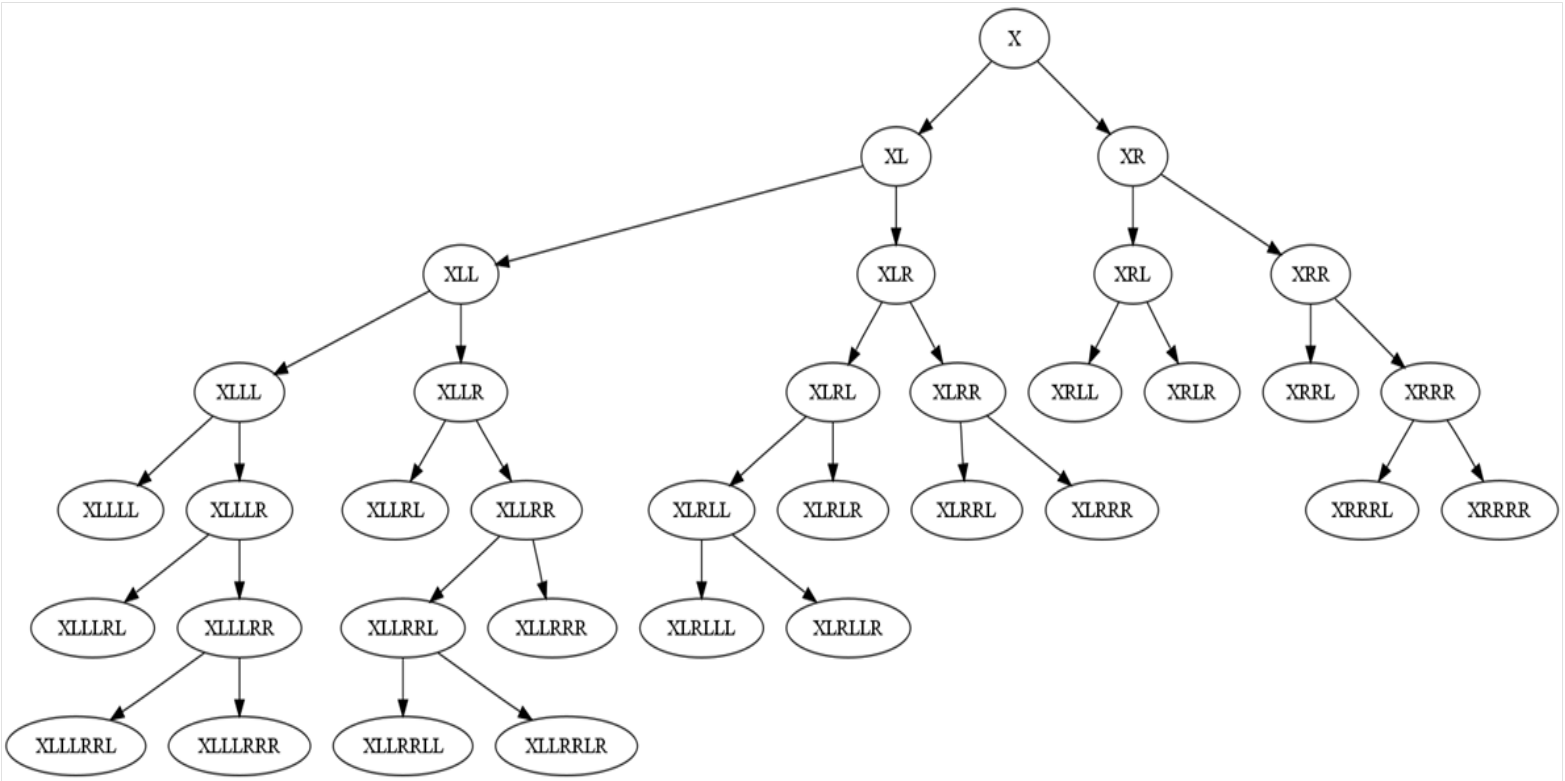
Monte Carlo Markov Chain over Histograms in $\mathcal{RSP}_{0:\infty}$



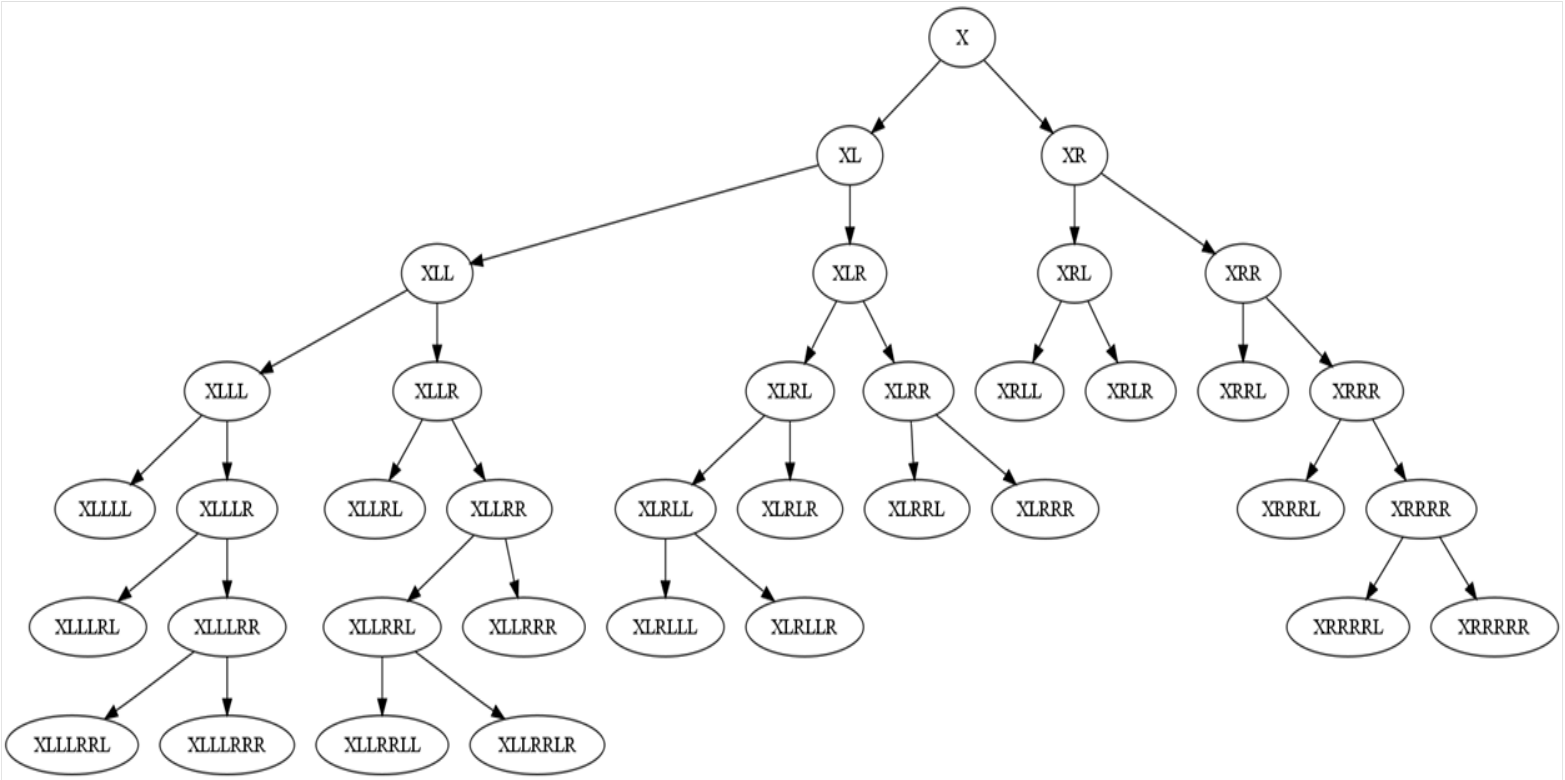
Monte Carlo Markov Chain over Histograms in $\mathcal{RSP}_{0:\infty}$



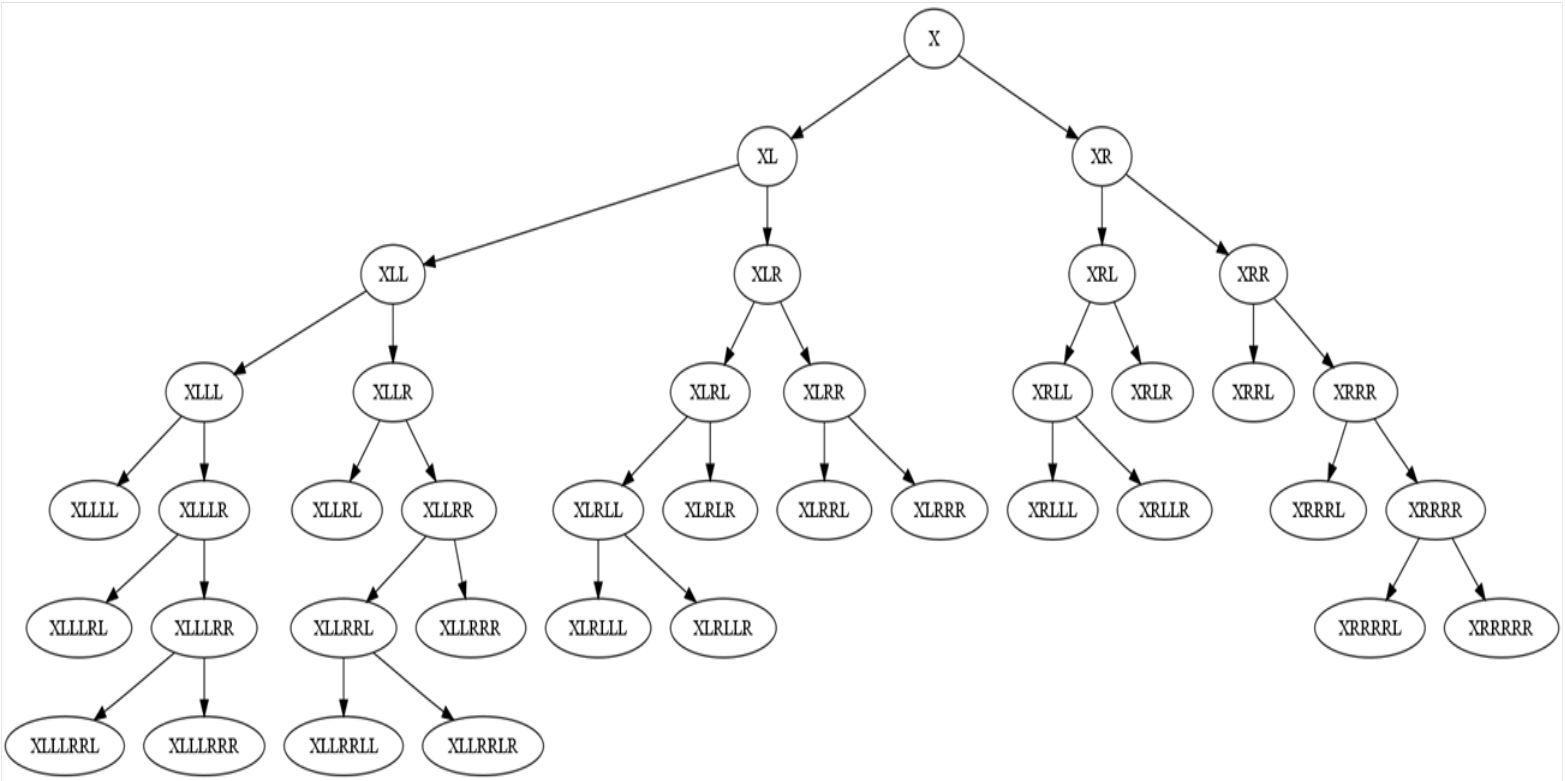
Monte Carlo Markov Chain over Histograms in $\mathcal{RSP}_{0:\infty}$



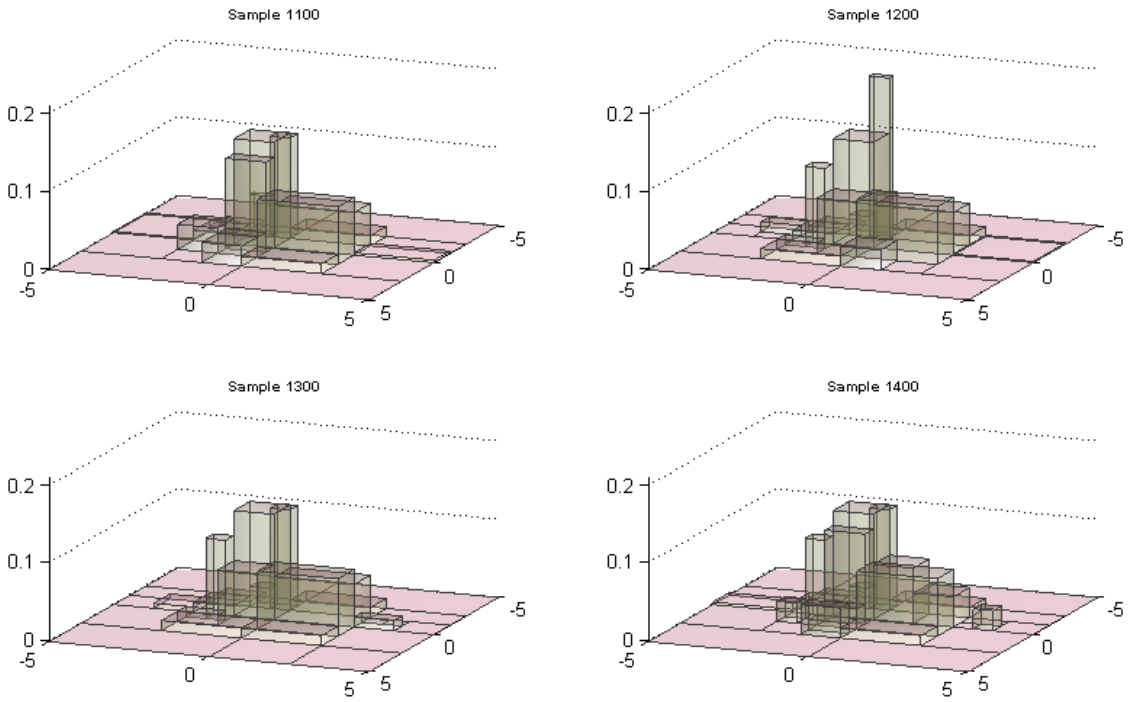
Monte Carlo Markov Chain over Histograms in $\mathcal{RSP}_{0:\infty}$



Monte Carlo Markov Chain over Histograms in $\mathcal{RSP}_{0:\infty}$

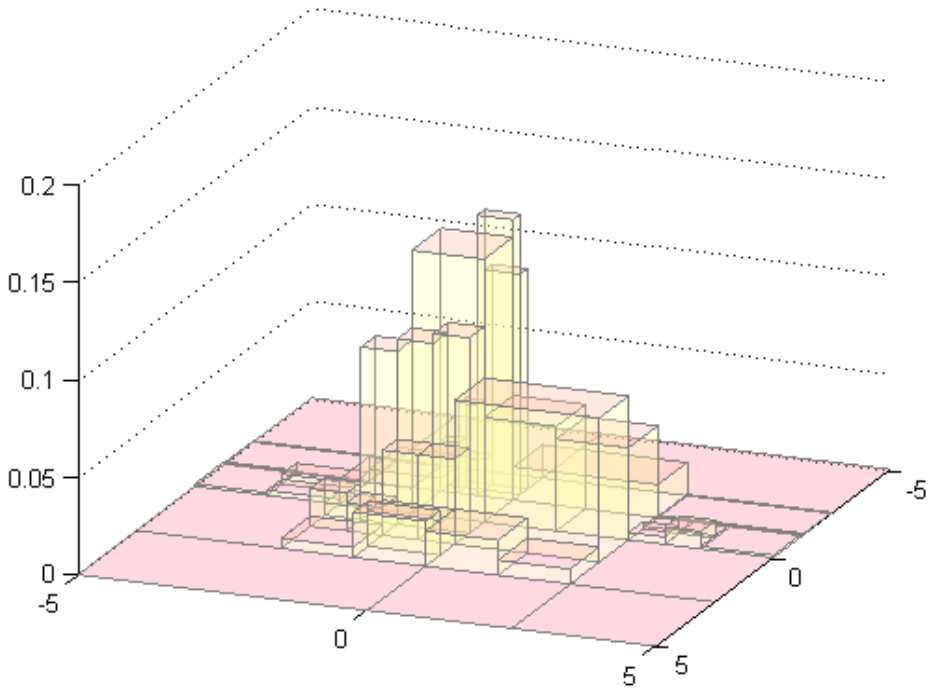


Histogram Estimates - Standard Bivariate Gaussian



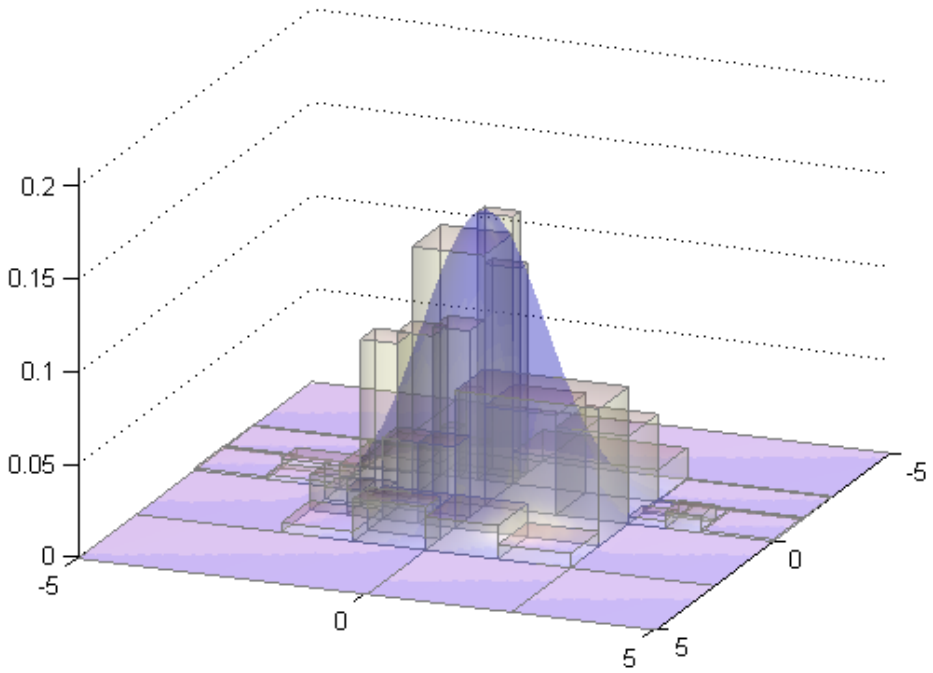
Four sample histograms

Histogram Estimates - Standard Bivariate Gaussian



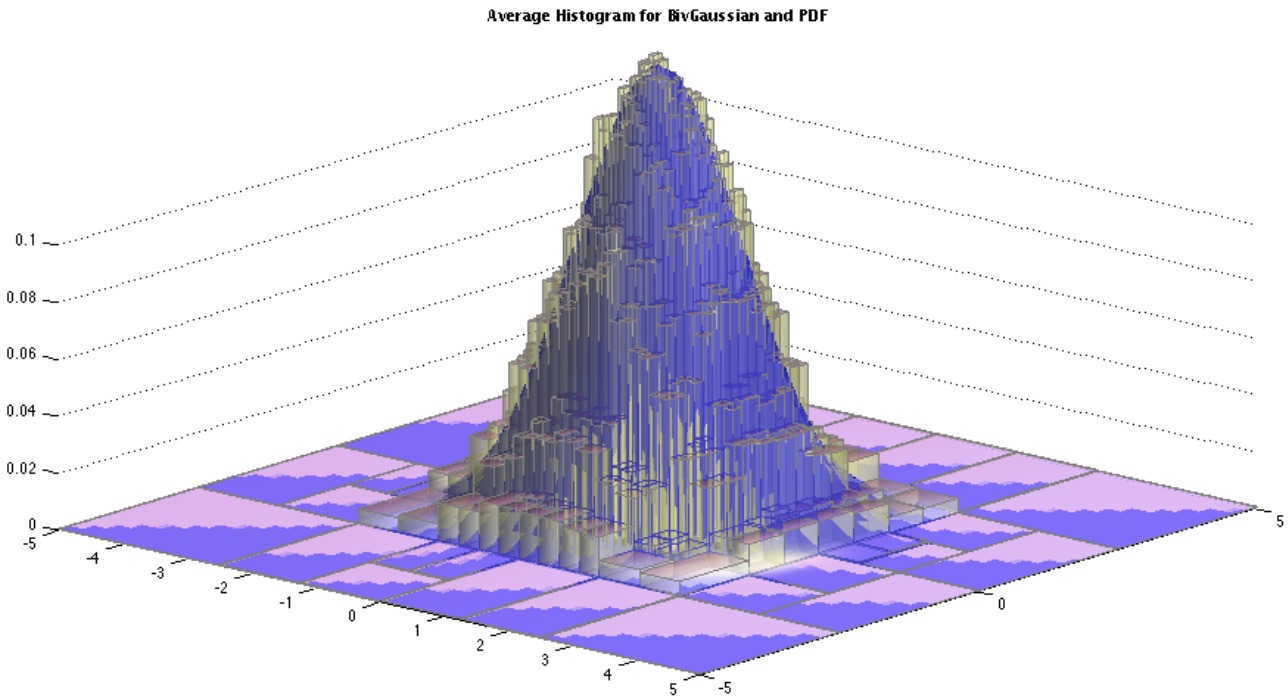
Average of the four sampled histograms

Histogram Estimates - Standard Bivariate Gaussian



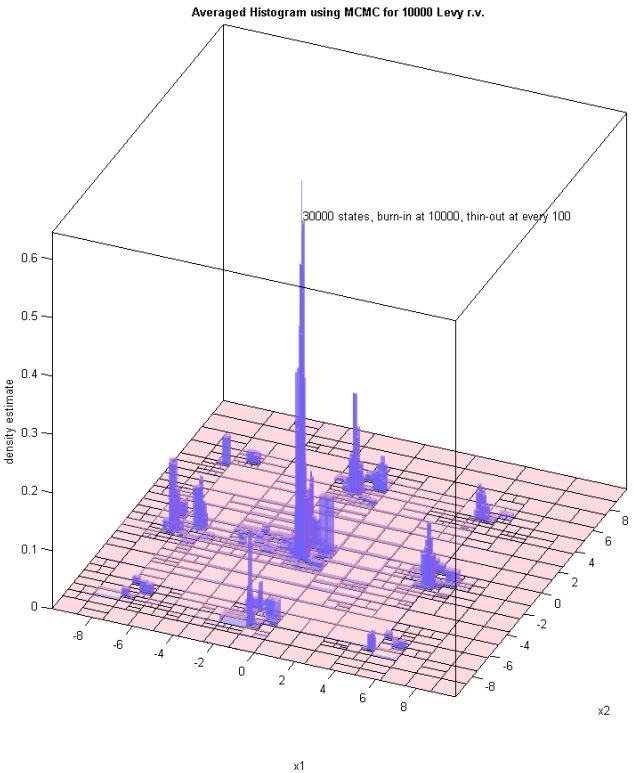
Average of the four sampled histograms with Gaussian PDF

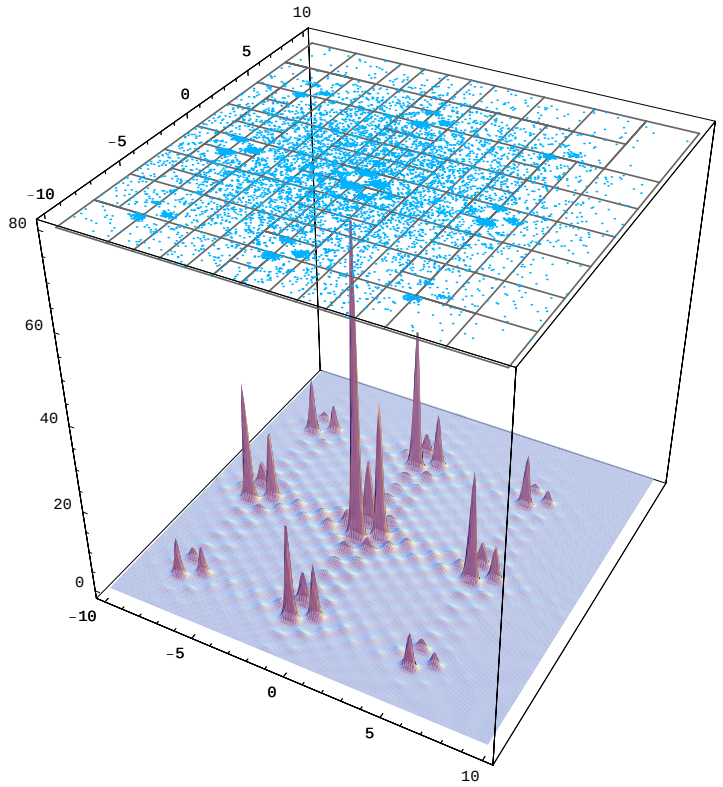
Histogram Estimates - Standard Bivariate Gaussian



A much better estimate

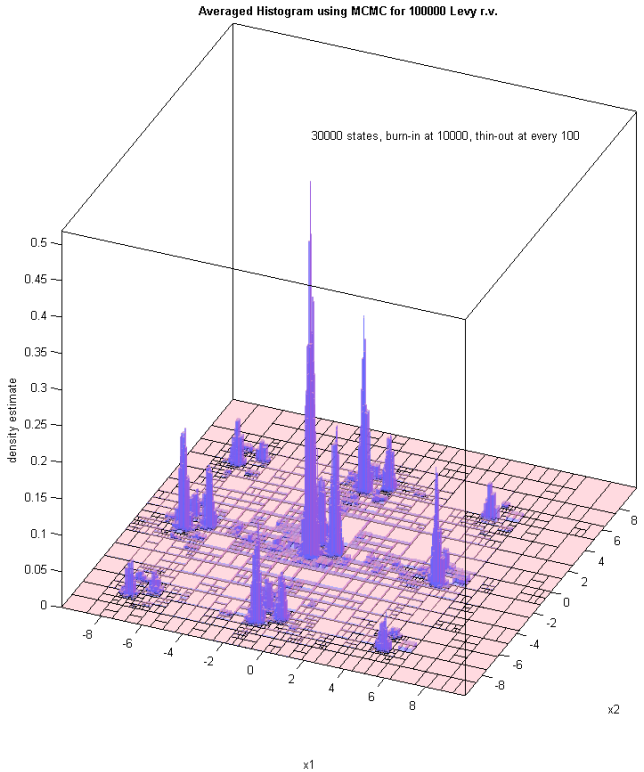
Histogram Estimates - Bivariate Levy Density

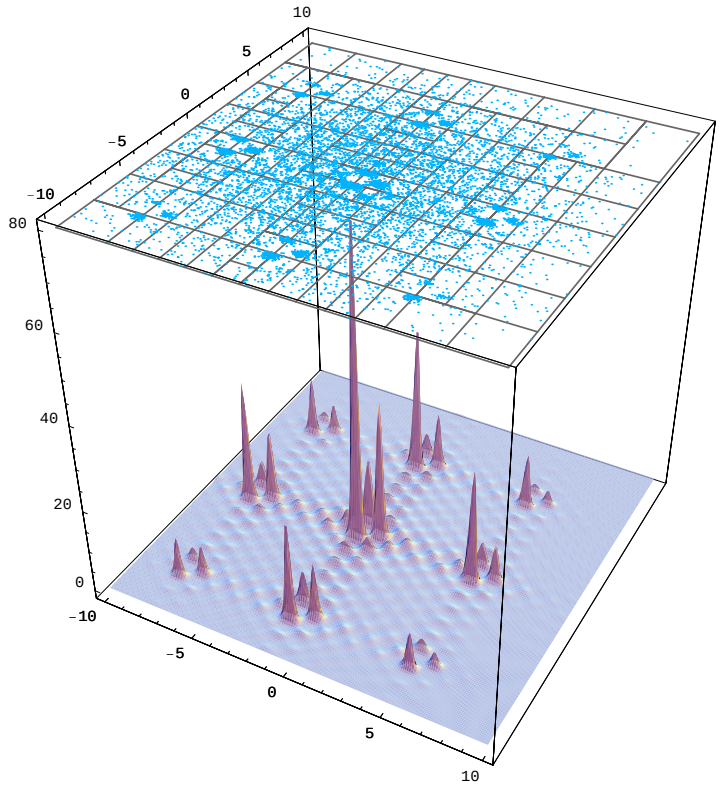




Data points = 10000, Number of states = 30000, Burn-in = 10000, Thin-out = 100, Averaged over 201 states, Time taken = 14.16s

Histogram Estimates - Bivariate Levy Density





Data points = 100000, Number of states = 30000, Burn-in = 10000, Thin-out = 100, Averaged over 201 states, Time taken = 50.59s

Conclusions

- Statistical Regular Sub-paving (SRSP) is a sufficient statistical data-structure for density estimation (massive data)

Conclusions

- Statistical Regular Sub-paving (SRSP) is a sufficient statistical data-structure for density estimation (massive data)
- We can grow or prune the SRSP tree data-adaptively

Conclusions

- Statistical Regular Sub-paving (SRSP) is a sufficient statistical data-structure for density estimation (massive data)
- We can grow or prune the SRSP tree data-adaptively
- Arithmetic can be efficiently extended to SRSPs - averaging histograms (also possible to extend to several SRSPs)

Conclusions

- Statistical Regular Sub-paving (SRSP) is a sufficient statistical data-structure for density estimation (massive data)
- We can grow or prune the SRSP tree data-adaptively
- Arithmetic can be efficiently extended to SRSPs - averaging histograms (also possible to extend to several SRSPs)
- This allows us to obtain posterior mean from MCMC samples on the space of adaptive multi-variate histograms with partitions in $\mathcal{RSP}_{0:\infty}$. MCMC convergence issues!

Conclusions

- Statistical Regular Sub-paving (SRSP) is a sufficient statistical data-structure for density estimation (massive data)
- We can grow or prune the SRSP tree data-adaptively
- Arithmetic can be efficiently extended to SRSPs - averaging histograms (also possible to extend to several SRSPs)
- This allows us to obtain posterior mean from MCMC samples on the space of adaptive multi-variate histograms with partitions in $\mathcal{RSP}_{0:\infty}$. MCMC convergence issues!
- Higher dimensional densities can be estimated with the approach